

**Hochschule Merseburg
Department of Computer Science**

Bachelor Thesis

<h1>E-Learning – Analytics using Jupyter Notebooks</h1>
--

Author: Jigneshkumar Jitubhai Sondagar
Born on: 23.10.1993 in Surat
Matriculation No.: 23602

First Examiners: Prof. Dr. rer. nat. habil, Eckhard Liebscher
Hochschule Merseburg (FH)
Second Examiners: Dr. Benjamin Wacker
Hochschule Merseburg (FH)

Place and Date: Merseburg, 29.07.2021

Aufgabenstellung
für die Bachelorarbeit (B. Eng.)
von Herrn Jigneshkumar Jitubhai Sondagar
(BAIN 23602)

Thema: E-Learning-Analysen mit Jupyter Notebooks
(E-Learning-Analytics using Jupyter Notebooks)

Betreuer: Prof. Dr. Eckhard Liebscher
Dr. Benjamin Wacker

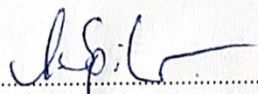
Aufgabenstellung

Unter Verwendung von Jupiter Notebooks sind Software-Tools zur Analyse von Assessmentdaten aus dem E-Learning-System Ilias zu erstellen.

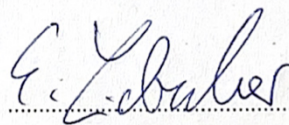
Schwerpunkte

1. Deskriptive Statistiken der erzielten Punkte
2. Detektion von Fehlerquellen bei den Lernenden und bei Lehrern
3. Clusteranalyse der Resultate der Lernenden

abzugebende Exemplare: 2 + PDF-Datei



Prof. Dr. Spillner
Vorsitzender des Prüfungsausschusses



Prof. Dr. Liebscher
Themenstellender Hochschullehrer

Contents

1	INTRODUCTION	4
1.1	THESIS OBJECT	6
1.2	THESIS STRUCTURE.....	6
2	THEORETICAL BACKGROUND	7
2.1	E-LEARNING FUNDAMENTAL.....	7
2.2	PYTHON AND JUPYTER NOTEBOOK	10
2.3	STATISTICS	21
3	BIG DATA	25
3.1	BIG DATA CHARACTERISTICS	25
3.2	DATA MINING	26
3.3	K-MEANS CLUSTERING	27
4	CONCEPT	33
4.1	DESCRIPTIVE STATISTICS OF THE POINTS SCORED	33
4.2	DETECTION OF SOURCE OF ERROR AMONG LEARNERS AND TEACHERS	33
4.3	CLUSTER ANALYSIS OF THE RESULTS OF THE LEARNERS.....	34
5	IMPLEMENTATION.....	35
5.1	DATA CLEANING	35
5.2	DESCRIPTIVE STATISTICS OF THE POINTS SCORED	38
5.3	BEST ATTEMPT	41
5.4	FAILURE ATTEMPT.....	42
5.5	CLUSTER.....	43
6	ANALYSIS EXAMPLE	46
6.1	BAR CHART	46
6.2	STATISTICAL ANALYSIS.....	54
6.3	CLUSTER ANALYSIS	56
7	SUMMARY AND OUTLOOK.....	58
8	ABBREVIATION	59
9	LIST OF FIGURE	60
10	LIST OF TABLES	61
11	REFERENCES	61
12	HOCHSCHULE LIBRARY AND ONLINE LIBRARY	64
	THANKSGIVING	65
	EIDESSTÄTTLICHE ERKLÄRUNG.....	66

1 Introduction

This chapter introduces the context of the bachelor thesis and briefly details on areas such as thesis background, objectives, and structure.

This thesis is written as part of my bachelor thesis at Hochschule Merseburg, Merseburg under Pro.Dr. Eckhard Liebscher. The E-learning management systems (LMS) in the education area is extensively used and it grows nowadays. It generates numbers of and several of data from the students activities using the online content. It produces a large amount of helpful and unhelpful data and these issues cannot be processed and supported by the traditional learning analytics. The cultured technology of Big Data and also became a recent trend now is huge evolved and relates to data-driven decision making. It is recognized as big data analytics and importance to apply for e-Learning in higher education system. This thesis study and evaluation the use of Big Data in e-Learning as the basic of Big Data and e-Learning and its power. Big Data Analytics has been established as successful path for learning data mining and learning analytics.

The e-Learning Analytics is set a techniques goals to work out useful information from available online education datasets.

The e-learning analytics categories:

- ⇒ Education: - Target to boost online education effect and students performance, like
 - Reducing students' dropouts
 - Improving students' understanding and learning
 - Deciding which content has relevancy for a given user
 - Improving training materials
 - Enhancing tutoring capabilities
- ⇒ Business: Target to boost the return of investment (ROI) of educational initiatives, like
 - Helping on marketing courses among the proper target market
 - Reducing tutoring costs

E-Learning analytics combined the employment of knowledge science techniques over data coming from various sources. in a very typical online education environment, the foremost way of knowledge sources are:

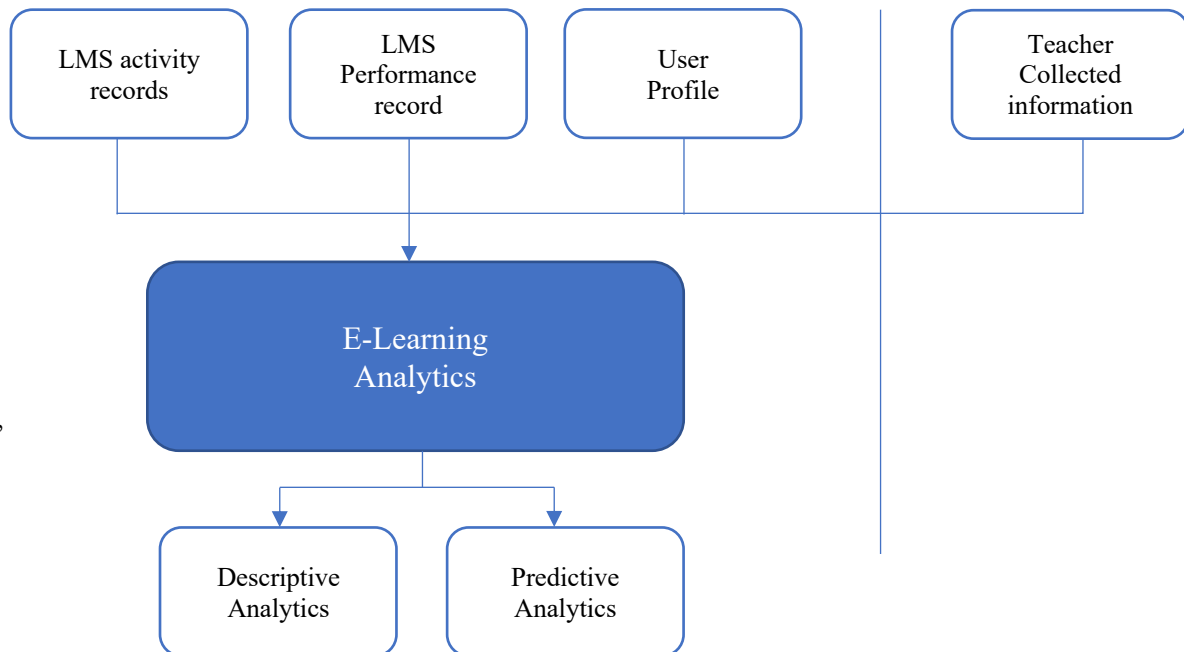
- ⇒ Learning Management System (LMS) activity records describing the users' interaction with the web platform and training content
- ⇒ User profile information, specially those characteristics that would impact the way students learn.

In environments where online content is employed as a supporting tool and student interaction also happens outside the training Management System, teachers may collect data that might be used as an extra input to the analysis. during this case, educators engagement is vital to form the process relevant.

Collected data is processed and analyzed, using advanced data science techniques, and a collection of relevant analytics are obtained as a result. These analytics provide insights into the educational process, that ought to be accustomed perform whatever actions are required to fulfill the academic or business goals.

There are two main sorts of analytics:

- ⇒ Descriptive: Provide insights about the past and permit to form decisions aimed to impact future learning processes.
- ⇒ Predictive: Perform predictions about elements and variables that might impact ongoing learning processes. These style of analytics allow educators to require proactive actions. (Omedes, 2020,What are e-Learning analytics about?)



(Omedes, 2020, What are e-Learning analytics about?)

1.1 Thesis Object

It is certain that the technological E-learning tools have come to play a vital role within the dissemination of data. The LMS allows you to require explicit data through the activity drifting by students from curriculum. The object of this thesis is software tools for the analysis of assessment data from the e-learning system Ilias are to be created using Jupyter Notebooks. This study selected the examination of E-Learning data and analysis of 76 student marks according to different assignment marks in 182 attempts. The object of this thesis is dividing in to three parts. The first part, Find Descriptive statistics of the points scored, which find the best attempt of the student, which means how many attempts students need to achieve maximum marks and represent into a graphical view and also find mean value of every single assignment which represent in a bar chart. The second object of this thesis is the detection of the source of error among learners and teachers. How many times students do not achieve maximum marks? That means to find a number of a failed attempts as well as the percentage of failure and represent them in the bar chart. And the final part is cluster analysis of the results of students. The objective of the analysis is that assigned the number of attempts and total marks in the group, clustering a group like in how many attempts are achieve 90 % marks.

1.2 Thesis Structure

This thesis is structured into 7 chapters. the primary chapter provides an introduction, thesis objectives of the thesis. and also, the structure of the thesis.

The first chapter of the thesis is that the introduction of LMS¹ and data analytic. The second section gives details about E-Learning Fundamental, Python and Jupyter Notebook and Statistics. The subsequent section is about big data and K-means Clustering. Section fourth is described the effect of Concepts of thesis. Next section figures out the Implementation . Section sixth is about the analysis. The last one is that the conclusion for this thesis.

¹ Learning Management System

2 Theoretical background

2.1 E-Learning Fundamental

Electronic learning or e-learning can be a deep word accustomed to describe computer-enhanced learning. An outsized range of terminologies is used for e-learning within the past, making it to difficult specify a general definition. Usually used e-learning terms include networked learning, virtual learning, Internet learning, distributed learning, computer-assisted learning, Web-based learning, and distance learning. All of these terminologies imply that the learner or trainee is at a physical distance from the tutor or instructor. The learner uses various types of technology to access the tutorial materials, using the system to interact with the tutor or instructor and learners, which some form of support is provided to learners. (Z.H.Tatli, 2009 Computer based education: Online Learning and teaching facilities)

E-learning has proved to be the most effective means within the corporate sector, especially when training programs are conducted by MNC²s for professionals across the world and employees are able to acquire important skills while sitting in an exceedingly board room, or by having seminars, which are conducted for workers of the identical or the various organizations under one roof. The colleges which use E-learning technologies are a step previous those which still have the standard approach towards learning. No doubt, it is equally important to require forward the concept of non-electronic teaching with the assistance of books and lectures, but the importance and effectiveness of technology-based learning can not be taken lightly or ignored completely. It is believed that the human brain can easily remember and relate to what is seen and heard via moving pictures or videos. It is also been found that visuals, except for holding the eye of the scholar, are retained by the brain for extended periods. Various sectors, including agriculture, medicine, education, services, business, and government setups are adapting to the concept of E-learning which helps within the progress of a nation. (Economic, 2021 ,Times The Economic)

2.1.1 What is value of E-Learning?

Online learning has numerous advantages over traditional learning methods. A number of these include the likelihood for college kids to form use of self-paced learning and to decide on their own learning environments. Additionally, e-learning is both cost-effective and cost-efficient, because it removes the geographical obstacles often related to traditional classrooms and education.

With that being said, it must be noted that e-learning is not perfect. Conducting any of the assorted sorts of e-learning through the net means sacrifices in a way or another. Increased risk of cheating during

² Multi National Company

assessments, social isolation, and lack of communicational skill development in online students are some of the challenges of e-learning which require to be addressed. (Tamm, 2020, Defining what is e-learning is not as easy as it might first appear.)

2.1.2 E-Learning and Learning Management System definition

Applying Learning Management Systems (LMSs³) in educational environments has help the communication between students and teachers, and built new challenges more well. The aim of this thesis is to analyses and analyses the role of LMS within the learning and teaching processes from students and teachers' perspectives.

A learning system help formalised teaching but with the assistance of electronic resources is understood as E-learning. While teaching are often based in or out of the school rooms, the utilization of computers and also the Internet forms the most important component of E-learning. E-learning may be termed as a network enabled transfer of skills and knowledge, and also the delivery of education is formed to an oversized number of recipients at the identical or different times. Earlier, it absolutely was not accepted wholeheartedly because it was assumed that this technique lacked the human element required in learning. However, with the rapid progress in technology and also the advancement in learning systems, it is now embraced by the masses. The introduction of computers was the concept of this revolution and with the passage of it slow, as we get hooked to smartphones, tablets, etc, these devices now have an importance place within the lecture rooms for learning. Books are getting replaced by electronic educational devices like optical discs or pen drives. Knowledge may additionally be shared via the net, which is accessible 24/7, anywhere, anytime. (Economic, 2021 ,Times The Economic)

2.1.3 Advantages, disadvantages and misconceptions

One of the leading advantages that e-learning provides to users is flexibility. They will access the course from virtually anywhere and anytime(recorded) with internet access. Since the courses are asynchronous, each learner can adjust when and the way long they require to participate, reckoning on their daily commitments. this enables learners to be enrolled in desired courses and still maintain their regular hours at work. Additionally, the learner would save time and transportation costs that occur when traveling to the campus. Learners can use the net to access up-to-date and relevant learning materials, and might communicate with experts within the field within which they are studying. Another key feature is that the convenience that e-learning provides. The learner can repeat each lesson as

³ Learning Management System

repeatedly as he or she wants. Videos and other media will be played as often as required. Also, using resources as media provides greater variation within the learning experience and may supply greater adaptability to the learners needs. Learners can easily interact with one another, irrespective of their physical location. For the trainer, tutoring may be done at any time and from anywhere. Online materials are updated, and learners are ready to see the changes without delay. There are great benefits within the business: although implementation cost may be on the high side, training costs go dramatically lower as more users use the system. Courses may be taken multiple times, with no extra cost to the host company.

E-learning may have drawbacks: the dearth of face-to-face communication, not only with the college members but also student-to-student interaction, leaves the learner with a way of isolation. There are tools available to attenuate this effect (such as chats and forums), the overall sentiment remains constant. The trainee motivation is additionally an affair. Since the learner is not bound by a hard and fast schedule sort of a real class would have. No progress is completed if he/she is not well-disciplined and motivated.

Marc Rosenberg⁴ lists the common misconceptions and misbeliefs that individuals have when brooding about E-learning. so as to successfully implement an E-learning system these points have to be taken into consideration then cleared out. The misconceptions are:

- Everyone Understands What E-learning Is: - This is often simply not true. Without an agreed-upon definition and a typical framework for thinking and talking about e-learning, confusion reigns. (2015)
- E-Learning is Easy: - Building and deploying great e-learning that are both effective and efficient takes real effort, discipline, and Knowledge within the fields like design, information design, communications, psychology, project management, and psychometrics, to not mention a healthy consideration of needs assessment and evaluation. (Rosenberg, 2005, - Approaches and Technologies to Enhance Organizational Knowledge, Learning, and Performance)
- It is boring by nature: - Many only think about e-learning as a video display with a protracted text. There are many ways to form the courses engaging and interesting by using different resources like media and interactive elements like a bunch of discussions.
- Success is getting E-Learning to work: - For too many, this implies getting the technology to figure. It is way more than that. Course flow and layout, generating focused and well-explained media is among many other elements that are required for a successful system.

⁴ Rosenberg, M. Beyond E-Learning

- Only certain content will be taught online: - This case, often extended by sponsors as an aim to not support e-learning, is wrong. With the proper instructional design approach, almost any form of knowledge or skill will be developed and delivered online.
- E-Learning value proposition relies on lowering the value of coaching delivered: - What should really be the main target is that the substantial benefits e-learning can generate in worker efficiency, speed of education deployment, and shortened times to competence.

2.2 Python and Jupyter Notebook

2.2.1 Python Programming language

Python could be a high-level, general-purpose, and really popular programming language.

Python artificial language (latest Python 3) is being employed in web development, Machine Learning applications, together with all cutting-edge technology in Software Industry. Python artificial language is extremely compatible for Beginners, also for skilled programmers with other programming languages like C++ and Java.

Introduction

Python may be a widely-used all-purpose, high-level programming language. It had been created by Guido van Rossum in 1991 and further developed by the Python Software Foundation. It absolutely was designed with stress on code readability, and its syntax allows programmers to specify their concepts in fewer lines of code.

Python might be a programming language that facilitates you to work quickly and integrate systems more efficiently. Python has two main versions: Python 2 and Python 3. Both are fairly different. (Wickramarathna, 2020, Python lesson 1(Introduction to python))

Advantages

1. Improved Productivity

Python is a very productive programming language. Reason of the simplicity, programmer can focus on solving error or problem. Developer do not need to spend more time in understanding the syntax or behavior of the language. Write less code and get more things done in Python.

2. Extensive support libraries

The standard library of Python is bulky and vast, where you can find most of the functions needed for your program. So, you do not need to depend on external libraries. But whether or

not, you are doing, a Python package manager (pip) makes things easier to import other great packages from the Python package index (PyPi). It hold over 200,000 packages.

3. Open source and community development

Python comes under the OSI approved open-source license. This makes it totally free of cost to use and distribute. You can download the source file, modify it and even distribute your version of Python. This is often useful for organizations that want to change some specific behavior and use their version for development.

4. High-level language

It has English like syntax, it is made code easily redable and understable. It is really easy to pick up and learn for beginner. You need less line code of code to execute same code in compare to other major language like C, C++ or JAVA.

5. Dynamically typed language

Python does not know the kind of variable until we run the code. It automatically assigns the data type during execution. The programmer does not have to worry about declaring variables and their data types.

6. Portable and Interactive

In other programming languages like C/C++ or Java, need to change code to run the program on different platforms. But here in Python don't need to change any code only write a code and run it anywhere.

7. Interpreter Language

Python is an interpreted language, that means Python can directly executes the code line by line. If in case code has any error, it stops further execution and reports back the error. It shows only one error even if the program has multiple errors. This makes debugging easier. (Joshi, 2019, Learning-Object-Oriented-Python)

Application

1. GUI⁵ based desktop applications (Games, Scientific Applications)

Python use for desktop applications Python has a Tkinter library, which can be used to make a graphical user interface. Python is also using for interactive games development. The PySoy library supporting to develop the 3D games in python3. For Example, Disney's Toontown Online, Vega Strike have been built help of Python.

⁵ Graphic User Interphase

2. Web frameworks and applications

Nowadays python rapidly using to develop web-application, because of the frameworks Python uses to develop this application.

This making framework help of common backend logic and many libraries that can help combine protocols such as HTTPS⁶, FTP⁷, SSL⁸.

3. Data Science and Data Visualization

Data is money if you recognize a way to extract relevant information which might help to take a calculated risk and increase profit. Python has such Pandas, Numpy libraries which help in extracting data. As well as python has some visualization libraries like Matplotlib, Seaborn, which are supporting to plotting graphs and much more.

4. Machine Learning and Artificial Intelligence

Machine Learning and Artificial Intelligence are the most demandable careers for the future. Learning is any process by which a system improves performance from experience and creates an algorithm that makes the computer learn itself. It is python.

5. Audio and Video Application

Python is using to develop multitasking application and also output media. The TimPlayer, Cplay python libraries have been used to make Video and audio application.

6. Embedded Application

Python is based on C language. It means python can be used for embedded application to create Embedded C. This helping to perform higher-level application on small device. The very famous embedded application is Raspberry Pi.

Organizations using Python

1. Google

Now a days Python is Google official server-side languages. C++, Java and Go that are three languages allowed to be deployed to production.

⁶ Hypertext Transfer Protocol secure

⁷ File Transfer Protocol

⁸ Secure Sockets Layer

2. Facebook

Facebook production engineers are exceptionally keen on python, making it the third preferred language at the social media giant. (just behind C++ and PHP dialect, Hack)

3. NetFlix

Netflix uses Python during a very similar manner to Spotify, looking forward to the language to power its data analysis on the server side. It does not just stop there, however. Netflix allows their software engineers to decide on what language to code in, and have noticed an oversized upsurge within the number of Python application. When surveyed, Netflix engineers cite the quality library, the extremely active development community, and also the rich sort of the third-party libraries available to resolve nearly any given problem. Additionally, Python is really easy to develop, it has become a key-point in many of Netflix's other services.

4. Spotify

This Music Streaming giant could be a huge proponent of Python, using the language primarily for data analysis and backside services. On the other end, there are an outsized number of services that each one communicates over ZeroMQ, an open-source networking library and framework that is written in Python and C++. The one only reason Spotify are written in python is that event Pipeline.

5. Dropbox

Dropbox is that the home for all of your docs, files, photos and videos. The whole code stack of Dropbox was written in Python. Numbers of corporate company's libraries are not open source, so it is difficult the level of Dropbox's dependence on Python. On the opposite hand, it released an API quoted in Python, and thus we will believe that a large amount of server-side coding written in Python.

6. Instagram

Instagram is one of most popular social media applications which share photo and video. It is one amongst the foremost companies, who use Python3 with Django in recent time. Instagram has 'Stories' features is used every day approx. 500 million active users as reported in January 2019. (Mistry, 2021-Top 7 Companies That Use Python Are Making A Mark In 2021)

Mode of Python Programming

Python has two basic modes: script and interactive.

Interactive mode

The Interactive mode is that the mode where the scripted and finished `.py` files are run within the Python interpreter. The Interactive mode could also be a command-line shell that offers instant feedback for every declaration while running previously fed statements inactive memory. As the latest lines are fed into the interpreter, the fed program is calculated both in part and in whole.

Interactive mode is a decent way to performance around and try differences on syntax.

On macOS or Linux, open a terminal and easily type "python". On Windows, say the command prompt and type "py", or start an interactive Python session by selecting "Python (command line)", "IDLE", or an analogous program from the taskbar/app menu. IDLE may be a GUI ⁹that features both an interactive mode and options to edit and run files.

Python should print approximately like this:

```
$ python
Python 3.0b3 (r30b3:66303, Sep  8 2008, 14:01:02) [MSC v.1500 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

Figure 1 Python IDLE

The `>>>` is Python's method of telling you that you just are in interactive mode. In interactive mode what you type is instantly run. Try typing `1+1` in. Python will respond with `2`. Interactive mode grants you to test out and see what Python will do. If you ever feel the requirement to play with new Python statements, get into interactive mode and take a look at them out.

A sample interactive session:

```
>>> 5
5
>>> print(5*7)
35
>>> "hello" * 4
'hellohellohellohello'
>>> "hello"._class_
<type 'str'>
```

Figure 2 Python Interactive session

In its place of Python exiting when the program is over, you can use the `-i` flag to start a communicating session. This can be very valuable for debugging and prototyping. (DannyS712, 2020)


```
python -i hello.py
```

Figure 3 debugging and prototyping

Script Mode

In script mode, well, Python does not automatically display results. In order to work out output from a Python script, We will introduce the print statement. This statement takes a listing of values and prints their string representation on the quality computer file. The quality output is often directed to the Terminal window. We will revisit this thorough within the print Statement.

```
print "PI = ", 355.0/113.0
```

Application scripts may be of any size or complexity. For the subsequent examples, We will create an easy, two-line script, called example1.py.

Example: - example1.py

```
print 65, "F"
```

```
print (65 - 32) * 5 / 9, "C"
```

There are several ways we will start the Python interpreter and have it an evaluate our script file.

- Implicitly from the program line. during this case, We will either use the GNU/Linux shell comment (sharp-bang marker) or We will depend upon the file association in Windows.
- Manually from within IDLE. it is important for newbies to recollect that IDLE should not be a part of the ultimate delivery of a working application. However, this is often a good thanks to start development of a program.

Running Python scripts from the command-line applies to any or all operating systems. It is the core of delivering final applications. We may add an icon for launching the applying, but under the hood, a programmer is actually a command-line start of the Python interpreter. (linuxtopia, 2021, Script Mode)

Python Libraries

Python Libraries are a collection of useful functions that eliminate the necessity for writing codes from scratch. There are over 137,000 python libraries present today. Python libraries play a significant role in developing machine learning, data science, data visualization, image and data manipulation application, and more.

What is Library?

A library could be a set of pre-prepaid codes that may be used attentively to scale back the time to put in writing to code. They are particularly useful for accessing the pre-written usually used codes, rather

than writing code from scratch every single time. Same on to the physical libraries, these are a set of reusable resources, which suggest every library encompasses a root source. This is often the bottom behind the amount of open-source libraries available in Python.

Most Useful Libraries.

Scikit-learn: - It is free of cost software ml¹⁰ library for the Python programming language and may be productively used for a range of applications which include classification, regression, clustering, model selection, naiveBayes, grade boosting, K-means, and preprocessing.

Scikit-learn requires:

- Python (≥ 2.7 or ≥ 3.3),
- NumPy ($\geq 1.8.2$),
- SciPy ($\geq 0.13.3$).

Spotify uses Scikit-learn for it is music recommendations and Evernote for building it is classifiers. If you have already got a working installation of numPy and Scipy, the best way to install Scikit-learn is using pip.

NumPy: - When it comes to scientific computing, NumPy is one in every of the essential packages for Python given support for big multidimensional arrays and matrices together with a group of high-level mathematical functions to execute these functions swiftly. NumPy built on BLAS¹¹ and LAPACK¹² for efficient linear algebra computations. NumPy may be used as an efficient multi-dimensional container of generic data.

TensorFlow: - The foremost popular dl¹³ framework, TensorFlow is an open-source software library for high-performance numerical computation. It is a perfect math library and is additionally used for machine learning and deep learning algorithms. TensorFlow was developed by the researchers at the Google Brain team within Google AI¹⁴ organisation, and today it is getting used by researchers for machine learning algorithms, and by physicists for complex mathematical computations. The subsequent operating systems support TensorFlow: macOS¹⁵ 10.12.6 (Sierra) or later; Ubuntu 16.04 or later; Windows 7 or above; Raspbian 9.0 or later.

¹⁰ Machine learning

¹¹ Basic Linear Algebra Subprograms

¹² Linear Algebra Package

¹³ deep learning

¹⁴ artificial intelligent

¹⁵ Mac operating system

Pandas: - It is an open-source, BSD¹⁶ licensed library. Pandas enable the availability of easy data structure and quicker data analysis for Python. For operations like data analysis and modelling, Pandas makes it possible to hold these out with no need to change to more domain-specific language like R. The simplest way to install Pandas is by Conda installation.

Scipy: - This can be yet one more open-source software used for scientific computing in Python. Except that, Scipy is additionally used for Data Computation, productivity, and high-performance computing and quality assurance. The assorted installation packages will be found here. The core Scipy packages are Sympy, Numpy, Matplotlib, SciPy library, Pandas, and IPython¹⁷.

Matplotlib: - All the libraries that we have discussed are capable of a gamut of numeric operations but when it involves dimensional plotting, Matplotlib steals the show. This open-source library in Python is widely used for publication of quality figures during a form of text formats and interactive environments across platforms. You will design charts, graphs, pie charts, scatterplots, histograms, error charts, etc. with just some lines of code.

Seaborn: - When it involves visualisation of statistical models like heat maps, Seaborn is among the reliable sources. This Python library come from Matplotlib and closely integrated with Pandas data structures. Visit the installation page to determine how this package may be installed. (Advani, 2020)

2.2.2 Jupyter Notebook

One of the foremost common questions people ask is which IDE¹⁸/environment/tool to use while performing on your data science/data analytics projects. As you'd expect, there's no dearth of options available – from language-specific IDEs like R Studio, PyCharm to editors like Sublime Text or Atom – the selection is intimidating for a beginner.

If there is one tool that each and every data scientist/data analytics should use or must be comfortable with, it is Jupyter Notebooks (previously called iPython¹⁹ notebooks). Jupyter Notebooks are powerful, versatile, shareable, and supply the power to perform data visualization within the same environment.

Jupyter Notebooks allow data scientists/data analytics to make and share their documents, from codes to full-blown reports. they assist data scientists/data analytics to streamline their work and enable more

¹⁶ Berkeley Source Distribution

¹⁷ Interactive Python

¹⁸ Integrated Development Environment

¹⁹ Interactive Python

productivity and simple collaboration. thanks to these and several other reasons you will see below, Jupyter Notebooks are one of all the foremost popular tools among data scientists/data analytics.



Figure 4 Jupyter Notebook

What is Jupyter Notebook?

Jupyter Notebook is an open-source web application that grants us to create and share codes and documents.

It provides an environment, where you will document your code, run it, examine the result, visualize data and see the results without leaving the environment. This lead to it a handy tool for operating end-to-end data science workflows – data cleaning, building and training machine learning models, visualizing data, statistical modeling, and many, many other uses.

Jupyter Notebooks really shines after you are still within the prototyping phase. This is often because your code is written in independent cells, which are executed individually. This permits the user to check a particular block of code in an exceedingly project without having to execute the code from the beginning of the script. Many other IDE environments (like RStudio) also try this in several ways, but I have got personally found Jupyter’s individual cells structure to be the most effective of the lot.

As you may see during this article, these Notebooks are incredibly flexible, interactive, and powerful tools within the hands of a knowledge scientist. They even permit you to run other languages besides Python, like R, SQL²⁰, etc.

Since they are more interactive than an IDE platform, they have widely accustomed to display codes in an exceedingly more pedagogical manner. (Chavan, 2018, Top 5: Online Notebook(ipynb) and other cloud services)

²⁰ Structured Query Language

How to install Jupyter Notebook

Before install Jupyter Notebook, we need to install Python in our machine first. Either Python Version 2.7 or Python Version 3.3 is sufficient.

Anaconda: -

Anaconda distribution to install both python and the Jupyter notebook. Anacondas installs both these tools and include quite a lot packages commonly used in the data science and machine learning community.

Below the figure we can see that Anaconda Navigator can launch various IDE such JupyterLab, Jupyter Notebook, PyCharm, Qt Console, Spyder, VS²¹ Code, Glueviz, Orange 3.

Once we click on Jupyter Notebook – Lunch button, it will open in default browser with the below URL: <http://localhost:8888/tree>.

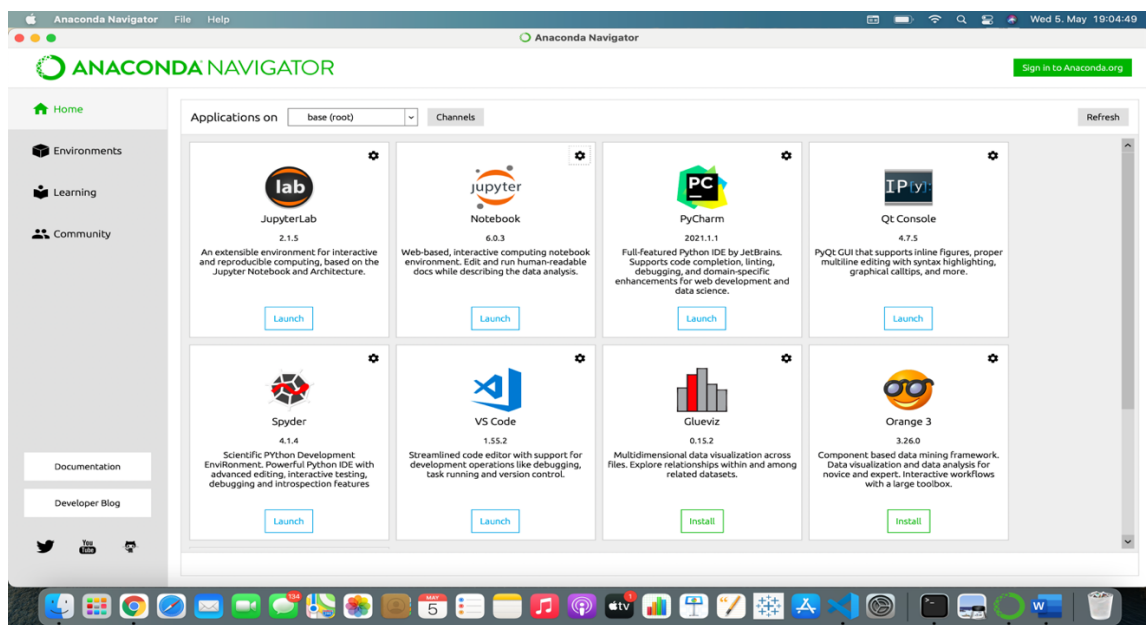


Figure 5 Anaconda Navigator

Sometimes, it may be not open automatically. A URL²² will be generated in the terminal/command prompt with the token key. You will need to copy paste this entire URL, including the token key, into your browser when you are opening a Notebook.

²¹ Visual Studio

²² Uniform Resource Locator

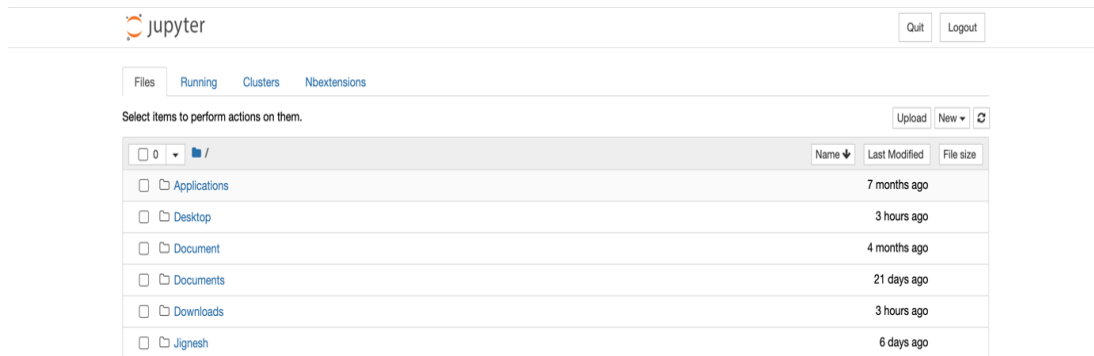


Figure 6 Dashboard of Jupyter Notebook

Once the Notebook is opened, you will see three tabs at the top: Files, Running and Clusters. Files basically lists all the files, running shows you the terminals and notebooks you currently have open, and Clusters is provided by IPython parallel.

To open a replacement Jupyter notebook, click on the 'New' option on the right-hand side of the page. Here, you get four options to decide on from:

- Python3
- Text File
- Folder
- Terminal

In a document, you are given a blank slate. Add whatever alphabets, words and numbers you want. It basically works as a text editor (similar to the appliance on Ubuntu). You furthermore may get the choice to settle on a language (there are a plethora of them given to you) so you will be able to write a script therein. You furthermore may have the power to seek out and replace words within the file.

In the Folder option, it does what the name suggests. you'll be able to create a brand-new folder to place your documents in, rename it and delete it, whatever your requirement.

The Terminal works exactly just like the terminal on your Mac or Linux machine (cmd on Windows). It does employment of supporting terminal sessions within your application program. Type python during this terminal and voila! Your python script is prepared to be written.

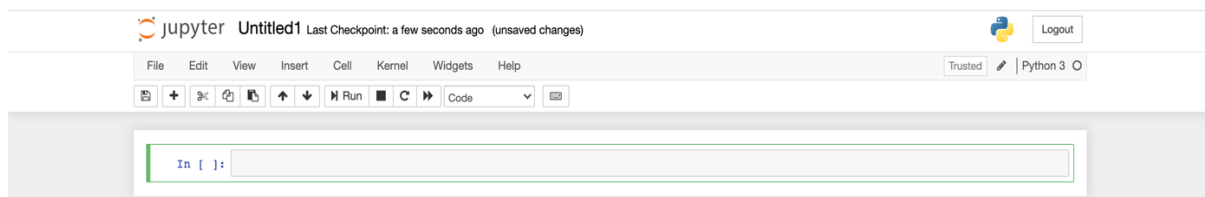


Figure 7 Command line

You can then start things off by importing the foremost common Python libraries: pandas and NumPy. within the menu just above the code, you have got options to manipulate with the cells: add, edit, cut, move cells up and down, run the code within the cell, stop the code, save your work and restart the kernel. (Dar, 2018, Comprehensive Beginner's Guide to Jupyter Notebooks for Data Science & Machine Learning)

2.3 Statistics

2.3.1 Central Tendency

Central Tendency is defined as the summary statistics that represent the center or middle point of the dataset. These measures show where highest values in a distribution fall and are also mentioned to as the central location of a distribution.

Now, we have to calculate statistics or the central tendency of such datasets to make inferences. The values that come under these are mostly mean, median and mode.

Mean

Mean is defined as the average of all the values inside the dataset. In mathematical terms. It is the sum of all elements divided by the total number of elements. We can calculate the mean for the only the quantitative dataset. It is also represented as “ μ ”.

In most cases, it gives the answer in the middle of the dataset.

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Where, n = number of elements
X= elements
 \bar{X} = Sum of all elements

Example:

Series = [10,11,10,11,10,11,10,10,20,30,40,25]

$$\text{mean}(\hat{\mu}) = \frac{10 + 11 + 10 + 11 + 10 + 11 + 10 + 10 + 20 + 30 + 40 + 25}{12}$$

$$\text{mean}(\mu) = 16.5$$

Median

Median, as the name defines middle, is defined as the value that occurs in the middle of the dataset, if sorted in ascending order. In mathematical terms, it is the middle of the sorted elements. We can calculate the mean for only quantitative dataset.

In most cases, it gives the answer in the middle of the dataset, near to the mean.

Step 1 – Sorting the dataset in ascending order

Step 2 – Find the position of the middle element position using one of the below methods.

$$m_x = \begin{cases} X_{(N)} & \text{If number of elements (n) is odd,} \\ \frac{1}{2} X_{(L)} + \frac{1}{2} X_{(L+1)} & \text{If number of elements (n) is even,} \end{cases} \quad N = \frac{n+1}{2}, L = \frac{n}{2}$$

Step 3 – Find the elements at the median position.

Sometimes, you will find two median values or single median value. We can also take the average of both median values in case n is even.

Example: Series = [56,67,54,34,78,,43,23]
Sorted series = [23,34,43,54,56,67,78]
n = 7 and Seven is odd number, so

$$median = \frac{7 + 1}{2}$$

median = 4th observation

Median = 54

Example: Series = [65,67,24,34,78,43]
Sorted series = [24,34,43,50,67,78]
n = 6 = even, so

$$median = \frac{6}{2} \text{ and } = \frac{6}{2} + 1$$

median = 3rd and 4th observation

median = 43 and 50

median = average of (43,50) = 46.5

Mode

Mode is defined as an element having the most frequency or occurrence. It may or not lie in the central value of the dataset. It has few drawbacks. Firstly, it will be the element occurring the most so it can lie on the extreme left or extreme right of the histogram, giving no information. Secondly, the dataset having unique values has no common element and has no mode. Although we can have median and mean.

Example: Series = [10,11,10,11,10,11,10,10,20,30,40,25]

Mode = 10

Example: Series = [8,9,7,5,8,9,7,5,9,8,7,3,1,0]

Mode = NA (Note Available)

2.3.2 Descriptive Statistics

Apart from central tendency we also have to understand the nature of the data. The basic questions that popped in the data analytics world, is how far a record from the mean or how data is spread in feature space.

Quartiles and Percentiles (Vedantu, 2021, Quartile Deviation)

The quartile calculates the spread of values above and below the mean by dividing the distribution into four groups. It splits the dataset into three points, a lower quartile, median and upper quartile to form four groups of the data set. Quartiles are used to measure the interquartile range, which is a measure of variability around the median.

Relationship between quintile, quartile, percentile.

Quintile	Quartile	Percentile
0	0	0
1	0.25	25
2	0.50	50
3	0.75	75
4	1	100

Table 1 Relation between quintile, quartile and percentile

$$\hat{q}_\alpha = \begin{cases} X_{(N)} & \text{für } \alpha n \notin \mathbb{N} \\ \frac{1}{2} X_{(\alpha n)} + \frac{1}{2} X_{(\alpha n + 1)} & \text{für } \alpha n \in \mathbb{N} \end{cases} \quad (\text{Pro.Dr.Liebscher, 2021})$$

once you get the position value, we calculate the percentile using below formula,

Percentile value = lower location value + location decimal (upper value – lower value)

Inter Quartile Range ²³(IQR) is defined as the region between the 3rd quartile and 2nd quartile. It is helpful in calculating the region for maximum data points.

$$IQR^{24} = Q_3 - Q_1 = 75 \text{ percentile value} - 25 \text{ percentile values}$$

Q2 is also called the median value of the dataset.

IQR has an application in making boxplots for the datasets.

Example: We have a score card for a student having 10 subjects. We have to find all percentiles.

Score	70	80	75	65	90	95	85	90	70	85
--------------	----	----	----	----	----	----	----	----	----	----

Table 2 Example of Inter Quartile Range

Solve:

Sorted Data:

Score	65	70	70	75	80	85	85	90	90	95
--------------	----	----	----	----	----	----	----	----	----	----

Table 3 Sorted data

$$L_p = \frac{p(n+1)}{100}$$

Where, n = number of elements in sequence

L_p = Location of sorted data

p = Percentile you are looking for

$$L_{25} = \frac{25(10+1)}{100} = \frac{25(11)}{100} = 2.75$$

$$\begin{aligned} 25 \text{ percentiles} &= 2^{\text{nd}} \text{ position value} + 0.25 (3^{\text{rd}} \text{ position value} - 2^{\text{nd}} \text{ position value}) \\ &= 70 + 0.25(70-70) = 70 \end{aligned}$$

$$L_{50} = \frac{50(10+1)}{100} = \frac{50(11)}{100} = 5.50$$

$$\begin{aligned} 50 \text{ percentiles} &= 5^{\text{th}} \text{ position value} + 0.50 (6^{\text{th}} \text{ position value} - 5^{\text{th}} \text{ position value}) \\ &= 80 + 0.50(85-80) = 82.5 \end{aligned}$$

$$L_{75} = \frac{75(10+1)}{100} = \frac{75(11)}{100} = 8.25$$

$$\begin{aligned} 75 \text{ percentiles} &= 8^{\text{th}} \text{ position value} + 0.75 (9^{\text{th}} \text{ position value} - 8^{\text{th}} \text{ position value}) \\ &= 90 + 0.75(90-90) = 90 \end{aligned}$$

$$IQR = Q_3 - Q_1 = 90 - 70 = 20$$

²³ & ²⁰ Inter Quartile Range

3 Big Data

Data has always played an important role in decision-making. With the coming of technology, data is being created in an exponential growth. There is a wave of data seen in every area universally. Digital data is available in every section of the company, educational institutions being no exception. Availability and sharing of data over the digital network, it is led to a huge increase in data quantity. Social networking, smartphones, and wireless sensor networks (WSN) are few means through which this large data is being produced. This huge data is called “Big Data” as these datasets are on the other side of the ability of traditional database tools to capture, search, storage, transfer, manage, visualization, sharing, querying, analyze, updating, and information privacy.

The definition of Big Data can differ from the group by group depending on the tools used and the average size of datasets associated with the sector. The present growth rate of data collected is enormous. It is a big challenge to handle and analyze the data productively. Big Data is defined as three-dimensional which are volume, variety, and velocity. It has been described as a “New generation of technologies and architecture designed to extract value for large datasets of wide variety by high-velocity capture, discovery, and analysis”.

3.1 Big Data Characteristics

A. Volume : - Data is being produced in terms of thousands of terabytes, petabytes, or zettabytes. The area of the data is the data generated or available to education. The Limit of data in LMS²⁵ or called as the records become bigger. The Problem is focused on capacity for data processing and also the system capability to process it.

B. Velocity : - The percentage of creation of data is defined as the velocity of data. It is seen that data is being generated at an exponential rate. The digital being generated is a number of and is overwhelming. It is relating to the characteristic of the area, bigger more time needed to process it. The speed is growing when the new data generated and moves.

C. Variety : - This is a number of different data representations like text, audio, video, images. These various data are structured, semi-structured, and unstructured forms data. Apart from the above mentioned about three dimensions. There have been other dimensions like value and variability being considered by different data scientists. These are working with the characteristic of volume, velocity, and variety for the accuracy and potential value of Big Data.

²⁵ Learning Management System

D. Variability : - Data apart from being varied, voluminous is also highly inconsistent inflow. Data flow will be event- triggered, seasonal among other reasons. These variable data loads are challenging. Unstructured and variable load of information makes the task of analyzing more complex.

E. Value : - This” fifth V “is the key element that understands the character or pattern generated by the data. This construct path for a predictive model generation. (Tulasi, 2014, Learning Analytics and Big Data in higher Education)

3.2 Data Mining

To analyze this large amount of information is beginning to use, progressively, two treatments or processes are also known as Data Mining and Big Data. Data Mining also called KDD²⁶ which is a process that allows the reveal the of hidden information in a bigger size of data (large amount). The processes in data subsets; finding and working with similar patterns of behavior and predictive model enables the useful processed data can be utilized.

The methods of data mining for data analytics . The tools which support the method of data mining or statistical analysis are appropriate for the large data sets. This is one of the characteristic in Big Data. In fact, the main idea respecting the data analytics in data mining is the larger of data sets, the statistics become more authentic. The main methods for learning analytics are defined in detail below. Initially, it is used for economic aims; the various possibilities have enabled us to expand the usage of it in education. The main methods used in the data mining of e-learning and the applications are as follows:

- **Prediction:** - From the data mining process, the emulation of student behavior as an example able to predict future activities. Other than that, it can draw a fitted regression curve for the prediction of the outcomes/ profit for e-learning.
- **Clustering:** - The data grouped to the useful cluster with the same elements or similar characteristics using the rules of partition (grouping the data).it is related to the patterns of the students in the same group with the same characteristic. It is able to use for prediction as well.
- **Relationship:** - This is a technology and technique for recognizing and tracking patterns within data and establishes the relationship to solve any issues related to e-learning or specifically the learning process between the learners and the educators. It is related to the associations with the sorting and sequencing the useful data.

²⁶ Knowledge Discovery Databases

3.3 K-Means Clustering

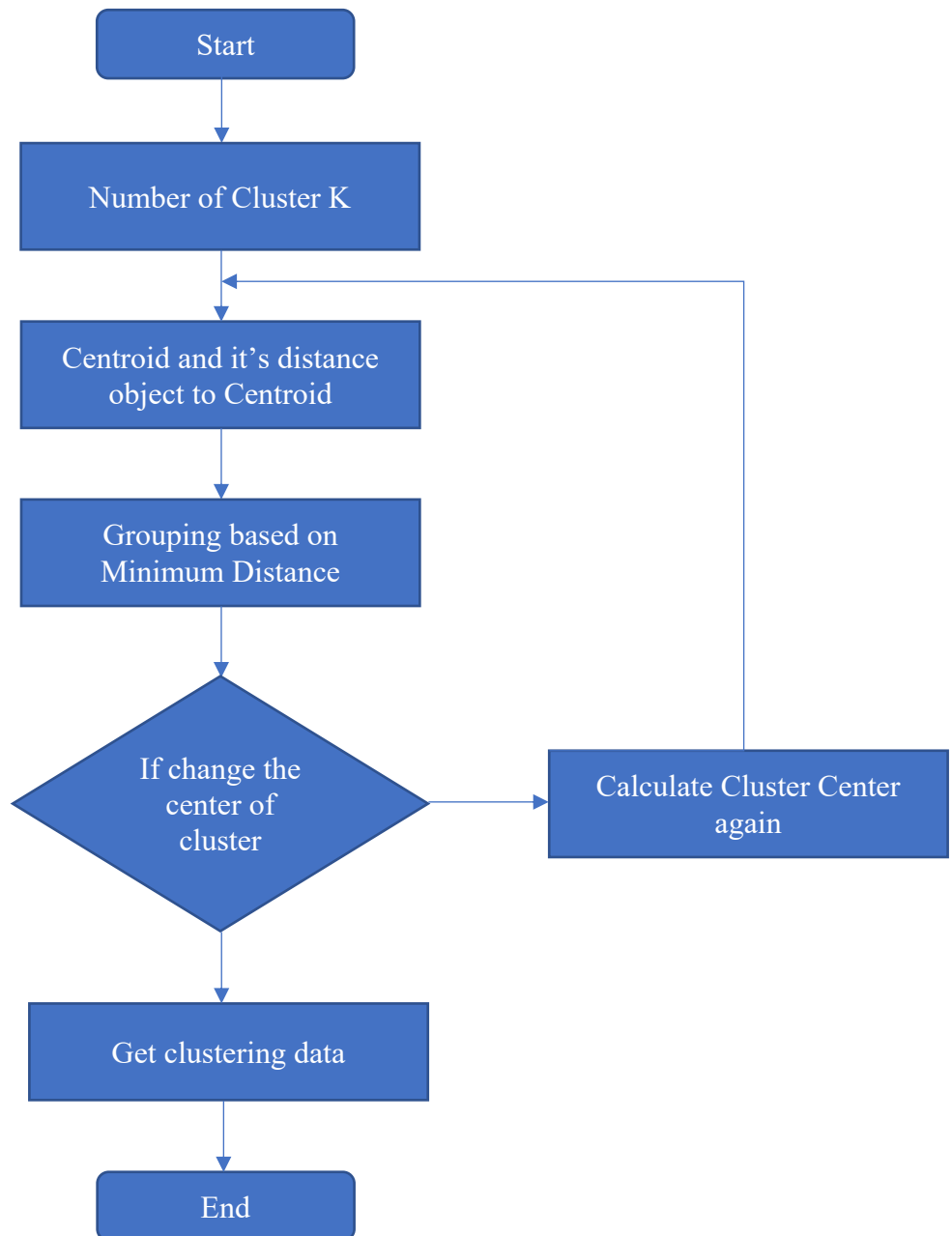
K-means is one of the methods, which is part of unsupervised learning. During this method, the developers don't give any data labels to the dataset, which means they do not range the whole data into specified categories or classes. The K within the algorithm name points to the k number of groups during which the algorithm is meant to divide data. This algorithm's basic purpose is to arrange data into different categories and the number of groups expressed by k, which could be any number. The users do it by evaluating the data features and working on them by keeping in view their similarities.

Cluster means a group of data, so this algorithm distributes n given data items into k clusters. The center points of each cluster are known means or cluster centroid. These points select which item belongs to which cluster. Each data item is to figure out and assigned to the cluster whose mean is closest to the item. Since the number of groups is not pre-defined in this algorithm, it is up to developers to choose the K numbers.

K-means clustering algorithm performs on the iterative process where k is various in every iteration to search optimal results. This algorithm takes k and the given data set, which consists of data points having various features. At the beginning, it uses evaluation for k centroids, which may be opted randomly from the data itself. These k centroids make k number of clusters by collect nearest points. The developers use the method of squared Euclidean distance in defining which data point belongs to which cluster. When the algorithm has made the basic clusters, it then updates the centroids. The following step is to process the data points in each cluster and determine their mean to update the new values of centroids, after which it repeats the analysis.

This analysis process finally stops when it is found the optimal clusters. The algorithm makes the stopping decision by using facts like the maximum number of iterations are reached, or the data points no longer change their respective clusters. The values of k are tested over an outlined range to seek out the most effective solution. Therefore, the developers repeat the method several times to search the optimal value of k.

3.3.1 K-means-Algorithms



1. Select number of cluster K.
2. Select randomly the centroids K point.
3. Choose each data point to the nearest centroid → that from K cluster.
4. Figure out and place the new centroid of each cluster.
5. Reassign every data point to the new nearest centroid.

If any resignation?

Revise step number 4 otherwise, the model is ready.

3.3.2 A 5-Step Implementation

The all-cluster algorithm implementing process is as simple because of it gets lesser human decision-making in the process and parameter toward compared to another machine learning algorithm. In this section I am expanding a 5-step of K-means clustering implementation in Python environment use of different libraries like Sklearn, seaborn and pandas.

Step 1. Install dependencies

Generally, need three libraries: pandas to deal with data, seaborn for visualization and Input pre-processing and modeling for Sklearn.

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import preprocessing
from sklearn.cluster import KMeans
```

Figure 8 Import Libraries

Step 2. Data

I am heir using 'iris' dataset because of the cluster can be easily visualize and separated in a scatter plot. In many datasets after importing data, may need to do some data cleaning such as replace or remove NaN values and etc.

```
df = pd.read_csv('/Users/jignesh/Jignesh/Python/Data/csv/iris.csv')
df.head()
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Figure 9 Read data

Step 3. Prepare model inputs

After selecting dataset, next step is pre-processing/formatting input in such way that the prepare model can use it. Two things can be happening in this step: Normalization of selected data and convert the dataframe into Numpy arrays.

```
df = preprocessing.scale(df)
df = pd.DataFrame(df)

x = df.iloc[:, [0, 1]].values
```

Figure 10 Prepare Model

Step 4: Determine number of clusters

How many clusters need in K-means algorithm? Determine the number of clusters that minimizing the sum of squared errors and it is called “elbow method”.

The k-means algorithm aims to decide on centroids that minimise the inertia, or within-cluster sum-of-squares criterion:

$$SD_{ges} = \sum_{k=1}^r \sum_{i \in g_k} \bar{d}_{ik} \text{ for } G = (g_1, g_2, \dots, g_r)$$

Where, \bar{d}_{ik} = distance y_i to centroid

g_k = element of group k

Step 5: Model Implementation

Once, visualize the number of cluster and made a determination on the required parameter, it is perfect to fit the model, plot in a two-dimensional and do further analysis.

```
k_range = range(1,10)
sse = []
for k in k_range:
    km = KMeans(n_clusters=k)
    km.fit(X)
    sse.append(km.inertia_)

plt.xlabel("K")
plt.ylabel("Sum of squared errors")
plt.plot(k_range, sse, marker='o')
plt.show()
```

Figure 11 K_range

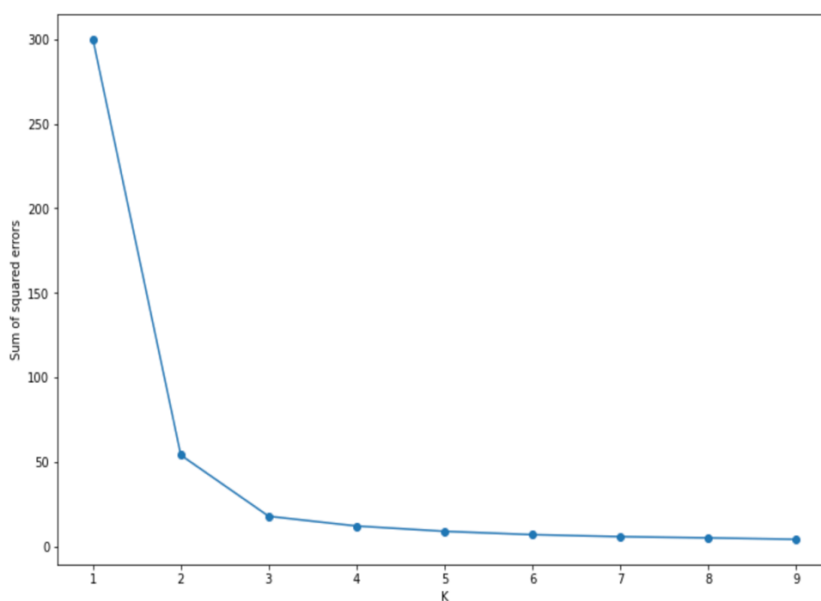


Figure 12 Elbow Methode

```

km = KMeans(n_clusters = 3)
y_km = km.fit_predict(X)

plt.scatter(
    X[y_km == 0, 0], X[y_km == 0,1],
    s=50, c='lightgreen',
    marker='s', edgecolors='black',
    label = 'cluster 1'
)

plt.scatter(
    X[y_km == 1, 0], X[y_km == 1,1],
    s=50, c='orange',
    marker='o', edgecolors='black',
    label = 'cluster 2'
)

plt.scatter(
    X[y_km == 2, 0], X[y_km == 2,1],
    s=50, c='lightblue',
    marker='v', edgecolors='black',
    label = 'cluster 3'
)
plt.show()

```

Figure 13 Model Implimentation

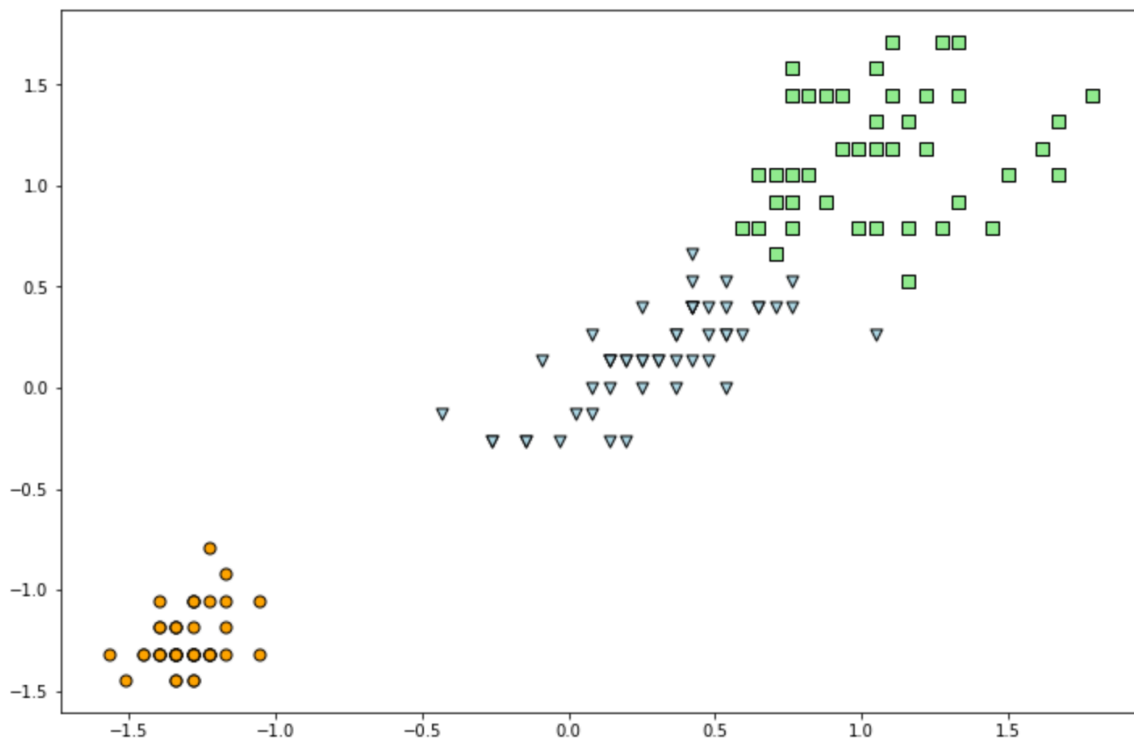


Figure 14 Clustering

3.3.3 Advantages of K-means Clustering

K-means is an iterative algorithm accustomed to finding the proper number of groups the data is supposed to classify. This algorithm proves helpful in business dealings by determining groups from the unsupervised complex data. Some advantages are the following:

Fast Computation : - K-means clustering convinces ways rapidly than hierarchical clustering when there are huge numbers of variables involved. In this study, keeping the number of groups small proves time-efficient with faster computations.

Tightly Bound Clusters : - Other Side to hierarchical clustering, k-means clustering helps in generating clusters that are closer to each other. This clustering is necessary in the case of working with globular clusters.

Simpler Implementation : - It is easy and smooth to implement. The choice of defining and updating the number of groups gives the advantage to developers to use it the way they find it helpful.

Easy Adaption : - It support to work with vast datasets and ensures the convergence of data. This clustering works well for the prepare dataset and correctly analyzes the newly-added examples.

Generalization : - K-means clustering discover to the clusters that possess various types of shapes and sizes easily.

3.3.4 Disadvantages of K-means Clustering

K-means having many advantages of the algorithm, it also has some shortcomings. Following are some disadvantages of the k-means clustering:

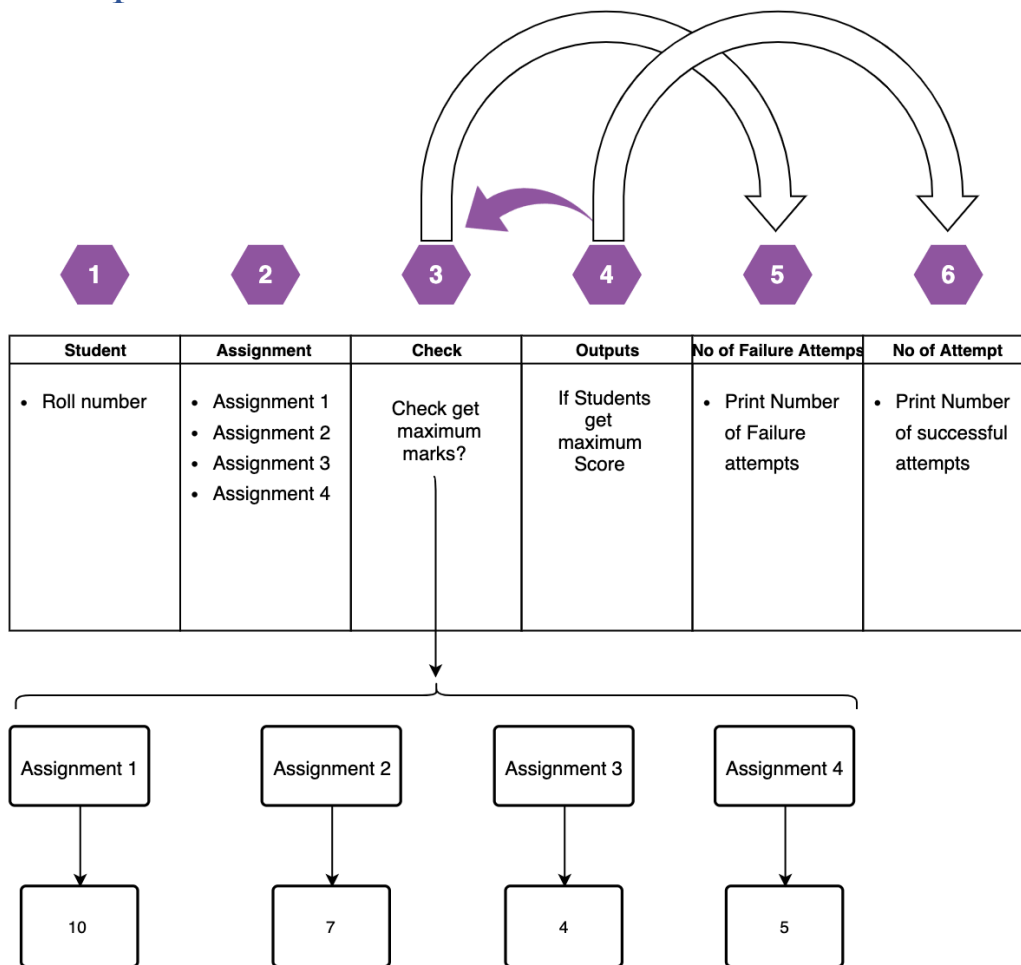
Difficulty with the K value : - Choosing k values manually in the algorithm can create problems in providing accurate solutions. Guess the correct value of k is a tough task.

Dependency on Initial Values : - The solution present by the algorithm mainly depends on the originally selected values, which can distract the algorithm from providing suitable clusters. The various selection of original partitions often results in different final clusters. Therefore, the result is to change the values of k several times and find the best solution.

Cluster of Outliers : - Since this algorithm works on classification by using similarities in data items, the outliers may be clustered separately from the opposite data. Outliers can also disturb centroids. The possible solution in this plot is to remove outliers from the data before clustering.

Dimensionality Problem : - The curse of dimensionality occurs in k-means clustering. Since each dimension is responsible for representing various attributes, the algorithm needs data transformation before analysis. Measurements of different attributes may require various types of scales, which can distort cluster analysis. Therefore, dimension reduction is required in such cases to stop the results from being misleading.

4 Concept



4.1 Descriptive statistics of the points scored

The purpose of this research is the examination of E-Learning data and analysis of 182 total attempts according to different assignment. The study group of this research raw data collected by 76 students curriculum activity. There are three commonly reported descriptive statistics called measures of central tendency. Find into the given data Descriptive statistics information and make it convert in to in graphic.

- Define, construct, and interpret visual descriptions of data: frequency distribution and histogram.
- Define, calculate, and interpret descriptive statistics concepts: mean, median, mode, and range.

4.2 Detection of source of error among learners and teachers

The examination of E-Learning data and analysis of 182 attempts according to different assignments find missing value and analysis among them.

In this task try to find how many attempts needed students to achieve maximum marks. Here we have 4 different tasks and their different maximum marks. After finding a successful attempt to reach maximum marks, we should calculate how many attempts they took to complete the task.

For Example, roll number 37 Student has a complete task in 6 attempts that means a successful attempt is 6. They got maximum marks (90%) on the 6th attempt, which means 5 try fails. That is called detection of a failure in the task.

4.3 Cluster analysis of the results of the learners

This study selected the examination of E-Learning data and analysis of 182 total attempts according to different assignment marks of 76 students. Data mining and find right the patterns and explore the main factors impact on the E-Learning performance. The problem of taking a set of data and separating it into subgroups where the elements of each subgroup are more similar to each other.

This method can separate students total marks and no of assignment by common traits in their answers, without any prior knowledge of what form those groups would take (unbiased classification). In this paper we start from a detailed analysis of the data coding needed in Cluster Analysis, in order to discuss the meaning and the limits of the interpretation of quantitative results. Then one method commonly used in Cluster Analysis are described and the variables and parameters involved are outlined and criticized.

5 Implementation

5.1 Data Cleaning

This chapter describes the implementation of E-Learning Data analytics. Primarily, the general structure of the project will be explained. Furthermore, the above Chapters and Subchapters briefly explain the theory and Statical. And finally, of the analysis the data will be discussed and explained how we get hidden information.

Row Data: - Row data Given by Prof. Dr. Eckhard Liebscher. Given Data has multiple column such V1, V2, V3, to V23 and 364 Rows. Data represent that submitted assignment by student in number of attempts and different information like total marks, students are pass or not and Highest marks in exam. Below We can see given row data in excel format.

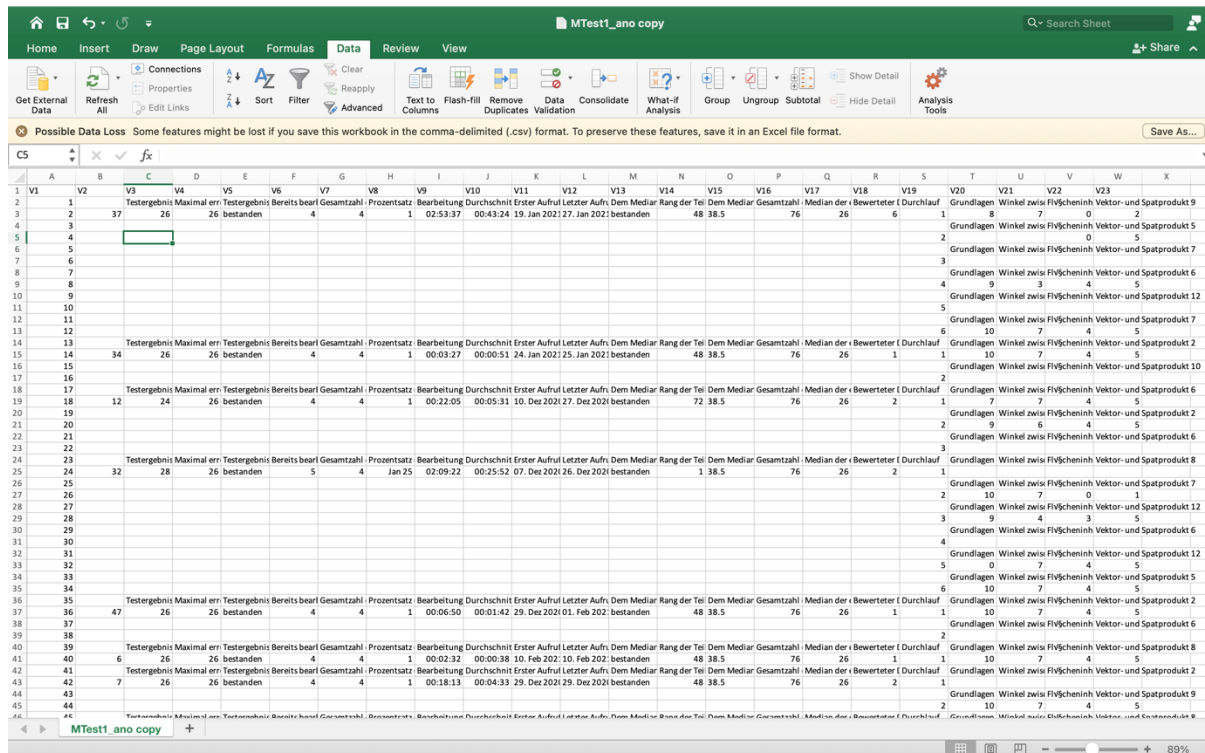


Figure 15 Row Data

Load a data into Jupyter Notebook, we need to Open anaconda navigator and lunch Jupyter Notebook and load some important libraries.

In python. We use the **import** keyword to make code in one module available in another. Imports in Python are import for structuring your code effectively. The import system is powerful and using imports. Properly will make you more productive, allowing you to reuse code while keeping thesis maintain.

Pandas: - Pandas is a package usually used to deal with data analysis. It simplifies the loading of data from external source such databases, as well as providing ways of analysing and manipulating data once it is loaded in computer.

Hire, I am importing pandas as pd

NumPy: - The NumPy library contains a multidimensional array and matrix data structures. NumPy can be used to perform a wide variety of mathematical operations on an array. I am using NumPy as np.

Matplotlib.pyplot: - Pyplot is a collection of functions in the popular visualization package matplotlib. Its functions use elements of a figure, such as creating a figure, creating a plotting area plotting lines, adding plot labels, etc. Normally matplotlib.pyplot using as plt.

```
In [1]: # Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Figure 16 Import Libraries

Read data from file ‘Mtest1_ano.csv’.

The primary process of loading data from the CSV²⁷ file into pandas Data Frame is accomplished using the “read_csv” function in Pandas. (The same directory must be same, where your python program is based)

When loading data with Pandas, the read_csv function is used for reading any define text file, and by changing the delimiter using the sep²⁸ parameter.

The head() function is inform about the first 5 rows in the DataFrame. This function returns the first n²⁹ rows for the object based on position. It is useful for immediately testing if your object has the right type of data in it.

²⁷ Comma separate value

²⁸ Separater

²⁹ Unknown number

```

In [2]: # Upload data
Data = pd.read_csv('/Users/jignesh/Jignesh/Python/Data/csv/MTest1_ano.csv', sep=';')

In [5]: # Print data
Data.head()

Out[5]:

```

Unnamed: 0	V2	V3	V4	V5	V6	V7	V8	V9	V10	...	V14	V15	
0	1	NaN	Testergebnis in Punkten	Maximal erreichbare Punktezahl	Testergebnis als Note	Bereits bearbeitete Fragen	Gesamtzahl der Fragen	Prozentsatz (vollständig bearbeitet)	Bearbeitungsdauer	Durchschnittliche Bearbeitungsdauer	...	Rang der Teilnehmerin	Dem Mediar zugeordnete Rang
1	2	37.0	26	26	bestanden	4	4	1	02:53:37	00:43:24	...	48	38.5
2	3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
3	4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN
4	5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN

5 rows x 23 columns

Figure 17 Read data and print first five data

Data Cleaning: - Very important part of every data analytics because of that data has many helpful and unhelpful information. Without cleaning data analytics will be miss guide.

Replace Nan to 0 value: - NaN means for Not A Number and is one of the basic ways to represent the missing value in the data. It is an extraordinary floating-point value and cannot be translated to any other type than float. NaN value is one of the key problems in Data Analysis. It is very basic need and very important to deal with NaN in order to get the desired results.

```
fillna() → Data.fillna(0, inplace=True)
```

Changing/ Removing/ the index of DataFrame: - Dropping the index column of pandas.DataFrame removes the index column from DataFrame and replaces it with the standard sequential indexing.

```
df = Data.iloc[1: , : ]
```

Dropping unnecessary columns in a DataFrame: - Remove columns or Rows by listing label names and equivalent axis, or by listing directly index or column names. When using a multi-index, labels can be removed by point out the level.

```
df_1 = df.iloc[ : : 2]
```

Clean data: - Dealing with messy data is inevitable. Data cleaning is just part of the process of a thesis. In this section, I tried over some ways to detect, summarized, and replace missing values. Finally got clean data.

```
In [24]: # print first 15 data
df_1.head(15)
```

```
Out[24]:
```

	Unnamed: 0	V2	V3	V4	V5	V6	V7	V8	V9	V10	...	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23
1	2	37.0	26	26	bestanden	4	4	1	02:53:37	00:43:24	...	48	38.5	76	26	6	1	8	7	0	2
3	4	0.0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	2	0	0	0	5
5	6	0.0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	3	0	0	0	0
7	8	0.0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	4	9	3	4	5
9	10	0.0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	5	0	0	0	0
11	12	0.0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	6	10	7	4	5
13	14	34.0	26	26	bestanden	4	4	1	00:03:27	00:00:51	...	48	38.5	76	26	1	1	10	7	4	5
15	16	0.0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	2	0	0	0	0
17	18	12.0	24	26	bestanden	4	4	1	00:22:05	00:05:31	...	72	38.5	76	26	2	1	7	7	4	5
19	20	0.0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	2	9	6	4	5
21	22	0.0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	3	0	0	0	0
23	24	32.0	28	26	bestanden	5	4	1.25	02:09:22	00:25:52	...	1	38.5	76	26	2	1	0	0	0	0
25	26	0.0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	2	10	7	0	1
27	28	0.0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	3	9	4	3	5
29	30	0.0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	4	0	0	0	0

15 rows x 23 columns

Figure 18 Clean and useful dataset

Create another data frame as df2

Hire, I am extracting some useful column, it will be store in dataframe as df2.

Extracting column: - [V2, V3, V5, V20, V21, V22, V23]

```
In [10]: # create another dataframe
df2 = pd.DataFrame(df_1, columns=["V2", "V3", "V5", "V20", "V21", "V22", "V23"])
df2.head()
```

```
Out[10]:
```

	V2	V3	V5	V20	V21	V22	V23
1	37.0	26	bestanden	8	7	0	2
3	0.0	0	0	0	0	0	5
5	0.0	0	0	0	0	0	0
7	0.0	0	0	9	3	4	5
9	0.0	0	0	0	0	0	0

Figure 19 Create DataFrame 2

5.2 Descriptive statistics of the points scored

we need to know, what is statistical information about the different attempts of assignment? That means the Mean value of the First attempts, the Mean value of the second try, etc.

let's deep drive into the dataset. In the below image, we can see I try to explain using different colors.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	...	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23
1	2	37.0	26	26	bestanden	4	4	1	02:53:37	00:43:24	...	48	38.5	76	26	6	1	8	7	0	2
3	4	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	2	0	0	0	5
5	6	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	3	0	0	0	0
7	8	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	4	9	3	4	5
9	10	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	5	0	0	0	0
11	12	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	6	10	7	4	5
13	14	34.0	26	26	bestanden	4	4	1	00:03:27	00:00:51	...	48	38.5	76	26	2	1	10	7	4	5
15	16	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	2	0	0	0	0
17	18	12.0	24	26	bestanden	4	4	1	00:22:05	00:05:31	...	72	38.5	76	26	3	1	7	7	4	5
19	20	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	2	9	6	4	5
21	22	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	3	0	0	0	0
23	24	32.0	26	26	bestanden	5	4	1.25	02:09:22	00:25:52	...	1	38.5	76	26	6	1	0	0	0	0
25	26	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	2	10	7	0	1
27	28	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	3	9	4	3	5
29	30	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	4	0	0	0	0
31	32	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	5	0	7	4	5
33	34	0.0	0	0		0	0	0	0	0	...	0	0	0	0	0	6	10	7	4	5

Figure 20 Extract data

The above image shows a data frame that has row and columns, Row has V1 to V23, and columns have odd number data like 1,3, 5, so on. Row number two is V2 and shows the Roll number of the student as well as Row number V19 shows the number of attempts that means how many attempts students have taken to complete assignments. so, let's try to extract data like Roll number 37 has taken 6 tries to complete an assignment (show in image purple color) and roll number 34 has taken 2 tries (green color) to mean while roll number 12 had done in 3 tries (orange color).

```
df_First_try = df_1.loc[(df_1['V19']=='1')]
df_First_try = df_First_try[['V19','V20','V21','V22','V23']].astype(int)
df_First_try.head(10)
```

Figure 21 Code of extract First try

Now we need to know what is the mean value of the First try? which means, I have to extract only the First try a value of every student. for example, roll number 37 has 8,7,0 and 2 marks on the First try and roll number 34 has 10,7,4, and 5 on the first attempt, like every student collection of data store in df_First_try Dataframe.

	V19	V20	V21	V22	V23
1	1	8	7	0	2
13	1	10	7	4	5
17	1	7	7	4	5
23	1	0	0	0	0
35	1	10	7	4	5
39	1	10	7	4	5
41	1	0	0	0	0
45	1	4	6	4	0
55	1	10	7	4	5
57	1	10	6	4	5

Figure 22 Data collection of first try

Now same above process for second attempt and collect value of every student. For example, roll number 37 has 0,0,0 and 5 marks on second try and roll number 34 has 0,0,0, and 0 on second attempt, like every student collection of data store in df_second_try Dataframe.

```
df_Second_try = df_1.loc[(df_1['V19']=='2')]
df_Second_try = df_Second_try[['V19','V20','V21','V22','V23']].astype(int)
df_Second_try.head(50)
```

Figure 23 Code of extract second try

The below image shows data collection of second attempt.

	V19	V20	V21	V22	V23
3	2	0	0	0	5
15	2	0	0	0	0
19	2	9	6	4	5
25	2	10	7	0	1
37	2	0	0	0	0
43	2	10	7	4	5
47	2	7	7	2	1
67	2	10	7	4	5
73	2	8	7	4	5
77	2	10	0	0	0

Figure 24 Data collection of second try

To continue I try to collect all attempts data like third try, fourth try till ninth try because many students has done in 9th attempt. So far, I extract data to help of above code and get desire output can see in below image.

V19	V20	V21	V22	V23	V19	V20	V21	V22	V23	V19	V20	V21	V22	V23			
5	3	0	0	0	0	7	4	9	3	4	5	9	5	0	0	0	0
21	3	0	0	0	0	29	4	0	0	0	0	31	5	0	7	4	5
27	3	9	4	3	5	51	4	10	7	1	5	53	5	10	7	3	5
49	3	9	4	1	0	81	4	10	7	4	5	95	5	10	7	4	5
79	3	0	0	0	0	93	4	10	6	2	5	141	5	10	6	4	5
91	3	10	5	1	5	131	4	10	7	4	5	175	5	10	4	0	0
109	3	10	7	4	3	139	4	10	7	4	1	209	5	10	7	4	5
129	3	10	3	3	0	173	4	0	0	0	0	257	5	10	7	1	0
137	3	9	6	4	0	207	4	8	3	2	5	289	5	0	0	0	0
147	3	10	7	4	5	255	4	10	7	1	0	331	5	8	0	0	0
V19	V20	V21	V22	V23	V19	V20	V21	V22	V23	V19	V20	V21	V22	V23			
11	6	10	7	4	5	261	7	0	0	0	0	263	8	10	7	4	5
33	6	10	7	4	5	293	7	10	7	4	5	359	8	0	0	0	0
177	6	10	7	3	5	335	7	10	7	4	5						
259	6	9	3	4	5	357	7	0	0	0	0						
291	6	7	7	4	5												
333	6	10	0	0	0												
355	6	0	0	0	0												
V19	V20	V21	V22	V23	V19	V20	V21	V22	V23								
361	9	10	7	4	5												

Figure 25 Data collection of all attempts

5.3 Best Attempt

In this task try to find how many attempts needed students to achieve maximum marks. Here we have 4 different tasks and their different maximum marks. The different task is listing name like V20, V21, V22 and V23 and their maximum marks like indexing 10, 7, 4 and 5. After finding a successful attempt to reach maximum marks, we should calculate how many attempts they took to complete the task.

For Example, roll number 37 Student has a complete task in 6 attempts that means a successful attempt is 6. They got maximum marks (90%) on the 6th attempt, which means 5 try fails. Well take second example, roll number 34 student has complete task in first attempt. As well as roll number 12 has taken 2 tries to complete all task or reach max marks.

```
Best = {'Versucht': [1,2,3,4,5,6,7,8,9],
        'V20': [41,26,14,10,6,5,2,1,1],
        'V21': [44,23,6,6,5,4,2,1,1],
        'V22': [50,26,11,6,4,4,2,1,1],
        'V23': [49,26,10,8,5,5,2,1,1]}

df_Best_attempt = pd.DataFrame(Best)
df_Best_attempt = df_Best_attempt.set_index('Versucht')

df_Best_attempt['V20 %'] = round(df_Best_attempt['V20'] / 182 * 100,2)
df_Best_attempt['V21 %'] = round(df_Best_attempt['V21'] / 182 * 100,2)
df_Best_attempt['V22 %'] = round(df_Best_attempt['V22'] / 182 * 100,2)
df_Best_attempt['V23 %'] = round(df_Best_attempt['V23'] / 182 * 100,2)
df_Best_attempt
```

	V20	V21	V22	V23	V20 %	V21 %	V22 %	V23 %
Versucht								
1	41	44	50	49	22.53	24.18	27.47	26.92
2	26	23	26	26	14.29	12.64	14.29	14.29
3	14	6	11	10	7.69	3.30	6.04	5.49
4	10	6	6	8	5.49	3.30	3.30	4.40
5	6	5	4	5	3.30	2.75	2.20	2.75
6	5	4	4	5	2.75	2.20	2.20	2.75
7	2	2	2	2	1.10	1.10	1.10	1.10
8	1	1	1	1	0.55	0.55	0.55	0.55
9	1	1	1	1	0.55	0.55	0.55	0.55

Figure 26 Best attempts Results

5.4 Failure Attempt

The examination of E-Learning data and analysis of 182 attempts according to different assignments find missing value and analysis among them.

In this task try to find how many attempts needed students to achieve maximum marks. Here we have 4 different tasks and their different maximum marks. After finding a successful attempt to reach maximum marks, we should calculate how many attempts they took to complete the task.

Below figure number 20 we try to extract data, which shown that how many attempts student take to reach maximum marks. We decided maximum marks criteria is 90% marks of maximum. For example, V20 Column has 10 marks is highest marks. Like V21, V22 and V23 are indexing 7, 4, and 5 marks are highest.

The Following code help to find who has achieved all requirement in which attempts. As you can see in below figure number 20, roll number 37 student has been done all tasks in 6 attempts. As well as roll number 34 has been done in only 1 attempt and roll number 12 has taken 2 try to reach maximum marks.

```
df_2=df_1[['V2','V19','V20','V21','V22','V23']].loc[((df_1['V20']==9) | (df_1['V20']==10)) &
          ((df_1['V21']==6) | (df_1['V21']==7)) &
          ((df_1['V22']==3) | (df_1['V22']==4)) &
          ((df_1['V23']==4) | (df_1['V23']==5))
          ]
```

```
df_2['Fehler'] = df_1['V19'] - 1
df_2
```

	V2	V19	V20	V21	V22	V23	Fehler
11	37	6	10	7	4	5	5
13	34	1	10	7	4	5	0
19	12	2	9	6	4	5	1
33	32	6	10	7	4	5	5
35	47	1	10	7	4	5	0
...
321	72	2	9	7	4	5	1
335	76	7	10	7	4	5	6
339	85	1	10	7	4	5	0
361	9	9	10	7	4	5	8
363	48	1	10	7	4	5	0

Figure 27 Extract data from dataframe_1

Above figure shown extract data. Now find how many failures students have done to reach maximum marks. Here I have to apply simple mathematic logic to find failure.

Let's deep drive to find. First of all, V19 column shown information about the total attempt that means how many attempts needed to reach maximum marks. The first data show in V19 is 6 tries/attempt, which means 5-time students face failure and 6th time they got desire result those maximum marks.

Here I make one extra 'Fehler' column, it extracts data which successful attempt minus 1. So 'Fehler' column represents the number of Failure and Extracting data put in a new data frame which easily represent the bar chart in the analysis part.

	Fehler	FehlerPercentage
Versucht		
1	31	17.03
2	20	10.99
3	4	2.20
4	3	1.65
5	4	2.20
6	3	1.65
7	2	1.10
8	1	0.55
9	1	0.55

Figure 28 Create Dataframe for Failure/Fehler attempts

5.5 Cluster

The K-Means algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion called the inertia or within-cluster sum-of-squares (see below). This algorithm requires the quantity of clusters to be specified. It scales well to sizable amount of samples and has been used across an outsized range of application areas in many various fields.

Import libraries:

Firstly, we import the pandas, pylab, sklearn libraries, matplotlib. Pandas is for the purpose of importing the dataset in csv format, pylab is the graphing library used in this example, and sklearn is used to devise the clustering algorithm and matplotlib for plotting a graph and visualisation.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler
import seaborn as sns
%matplotlib inline
import sklearn.cluster as cluster
```

Figure 29 Import Library

Scatter Plot

Then, the 'df_V3_V18' dataframe is imported, with our Y variable defined as 'Number of attempt(V18)' and X variable defined as 'Marks (V3)'.

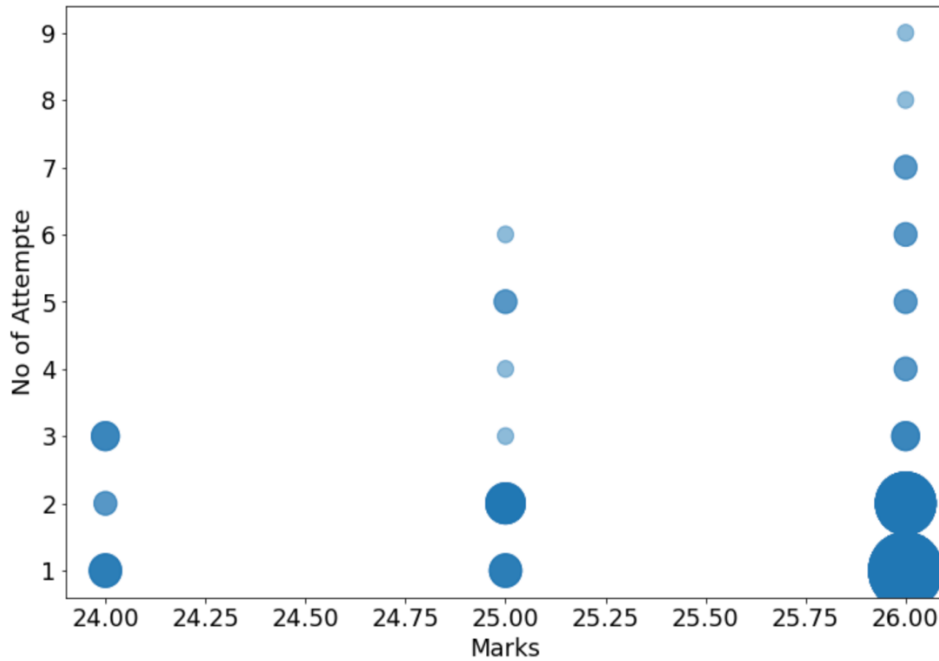


Figure 30 Scatter Plot

The K-means algorithm aims to decide on centroids that minimise the inertia, or within-cluster sum-of-squares criterion: This method is to run k-means clustering on the dataset for a range of values of k (say, k from 1 to 11 in the examples below), and for each value of k calculate the sum of squared errors (SSE). Like this:

$$SD_{ges} = \sum_{k=1}^r \sum_{i \in g_k} \bar{d}_{ik} \text{ für } G = (g_1, g_2, \dots, g_r)$$

Where, \bar{d}_{ik} = distance y_i to centroid

g_k = element of group k

```

k_rng = range(1,11)
sse = []
for k in k_rng:
    km = KMeans(n_clusters=k)
    km.fit(df_v3_v18[['v3', 'v18']])
    sse.append(km.inertia_)

```

```

sse
[289.52777777777777,
93.3694581280788,
64.83796992481206,
42.22365003417633,
28.918088235294128,
22.68236714975845,
15.982492997198873,
12.956349206349207,
10.843604108309993,
8.090476190476188]

```

Figure 31 Sum of Squire Error

The number of clusters is user-defined and also the algorithm will try and group the information whether or not this number isn't optimal for the precise case. Therefore we've got to return up with the best way that somehow will help us decide how many clusters we should always use for the K-Means model.

The Elbow method is also an extremely talked-about technique and also the concept is to run k-means clustering for a spread of clusters k (let's say from 1 to 10) and for every value, we are calculating the sum of squared distances from each point to its assigned center (distortions). When the distortions are plotted and also the plot feels like an arm then the "elbow" (the point of inflection on the curve) is that the best value of k .

Note Dataset `df_V3_V18` on the left. At the top we see a number line plotting each point in the dataset, and below we see an elbow chart showing the SSE^{30} after running k-means clustering for k going from 1 to 10. We see a pretty clear elbow at $k = 2$, indicating that 2 is the best number of clusters.

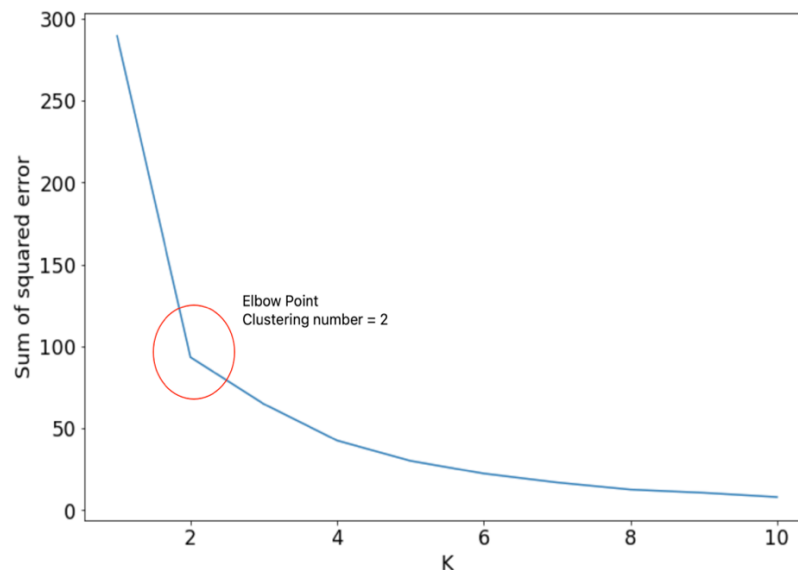


Figure 32 Elbow Method

Every clustering algorithm has two variants: First is class, that implements the `fit` method to learn the clusters on train data, and Second is function, that, given train data, returns an array of integer labels respective to the different clusters. For the class, the labels over the training data will be found in the `labels_` attribute.

```
km = KMeans(n_clusters=2)
y_predict = km.fit_predict(df_V3_V18[['V3', 'V18']])

df_cl['Cluster'] = y_predict
```

Figure 33 Cluster Prediction

Cluster centroid is that the middle of a cluster. A centroid could be a vector that contains one number for every variable, where each number is that the mean of a variable for the observations therein cluster. The centroid may be thought of because the multi-dimensional average of the cluster.

```
km.cluster_centers_
array([[25.5, 1.62068966],
       [25.71428571, 5.78571429]])
```

Figure 34 Cluster Center

³⁰ Sum of Squared Error

6 Analysis example

The analysis is an interpretive process that pulls conclusions from a collection of facts.

For our purposes, the subsequent definitions will apply:

- A plot summary could be a brief informative report of what happens in an exceedingly written material (the facts of the story).
- An interpretation could be a logical analytical conclusion a few work supported the facts of the story.
- A literary analysis could be a careful examination of the mechanism of a piece of writing and a discussion of how that mechanism functions to reveal meaning.

6.1 Bar Chart

A bar chart is a way of summarizing a collection of categorical data (continuous data will be made categorical by auto-binning). The bar graph displays data employing a number of bars, each representing a specific category. The peak of every bar is proportional to a particular aggregation. The categories can be something like an cohort or a geographical location. It is also possible to paint or split each bar into another categorical column within the data, which enables you to determine the contribution from different categories to every bar or group of bars within the bar graph.

6.1.1 Best attempts

Interpreting data requires you to understand the information you are getting from the graph and able to say what it means.

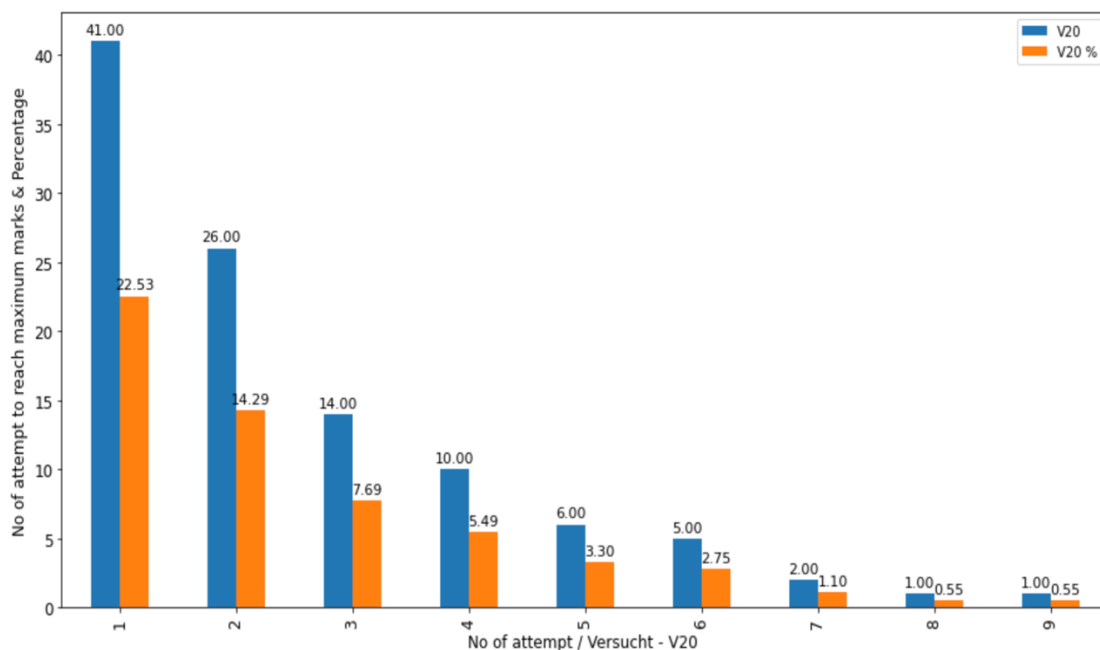


Figure 35 Best attempts of V20

Explanation V20: - The figure number 28 shows best attempt of V20 column, the graph is left skewed, the question is that how many attempts needed to achieve maximum marks? First try give information that 41-attempt needed to reach mark in different assignment, 22.53% student has achieved maximum marks in first attempts. Second try need 26 attempt and 14.29% students needed to achieve maximum mark. As following graph represent more information in graphical view.

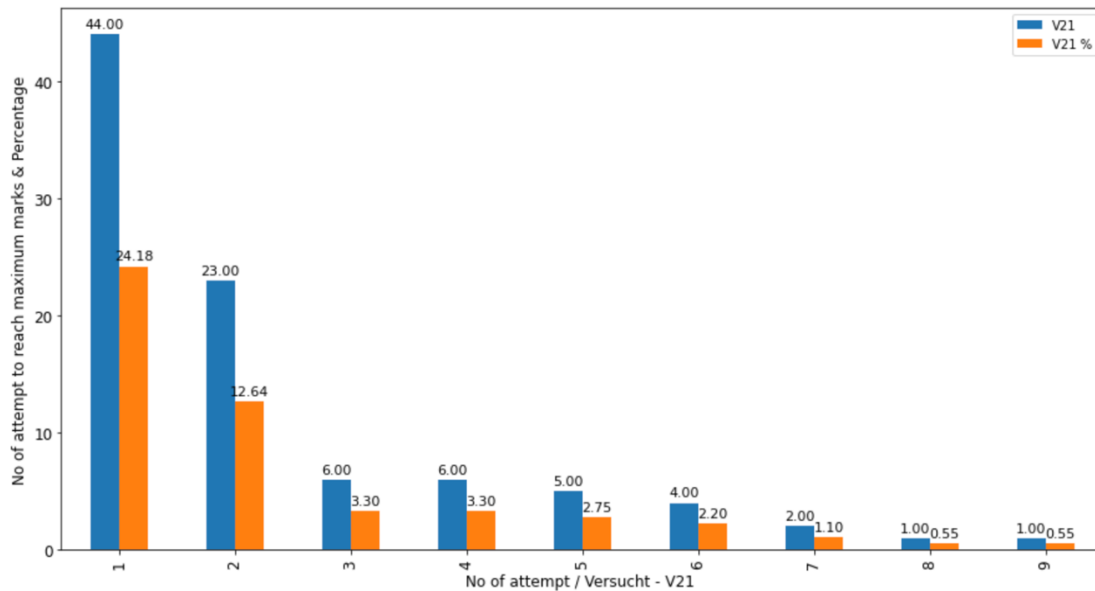


Figure 36 Best attempts of V21

Explanation V21: - You might be asked to find the successful no of attempt to reach maximum marks in V21? According to the chart, there were upward trends in 'No of attempt' on the first try and second try. The most of students have completed their tasks in 44 and 23 attempts. On the other hand, 'No of tries/attempts' increasing so does 'No of attempts to reach maximum marks and percentage' decreasing.

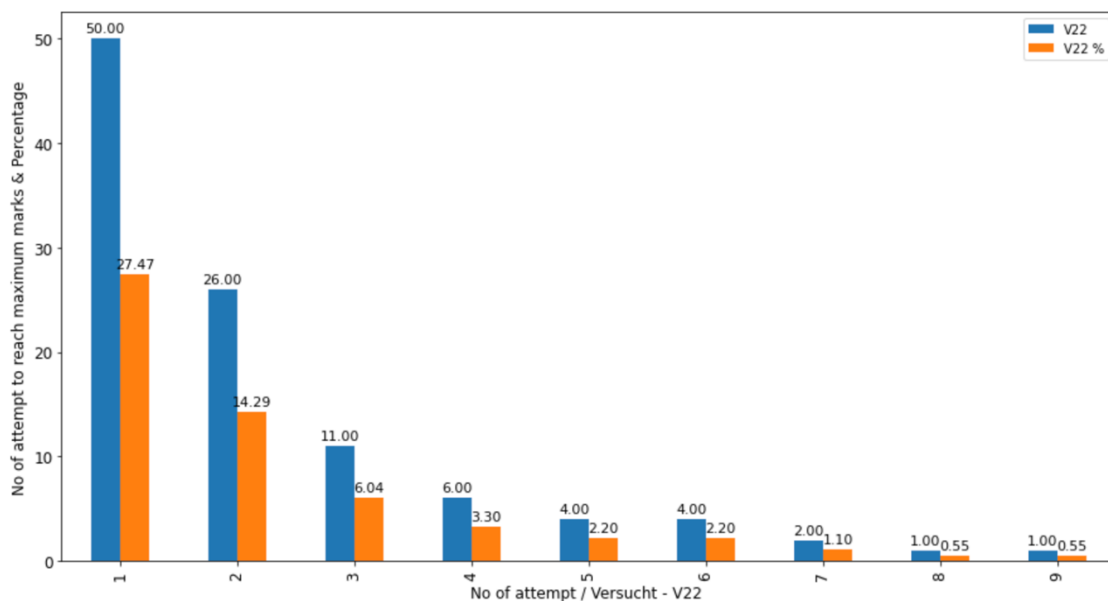


Figure 37 Best attempts of V22

Explanation V22: - You might be asked to find out who has good calculation and who has bad calculation, which means who takes a long time to calculate assignments and who is done quickly. The below figure represents more detail than more than 50 ‘No of attempts to reach to maximum marks’ on the first try and 26 attempts need on the second try. In this graph, we can see 1 attempt also need to complete the task in the 8th and 9th tries. As we can see the graph is left-skewed.

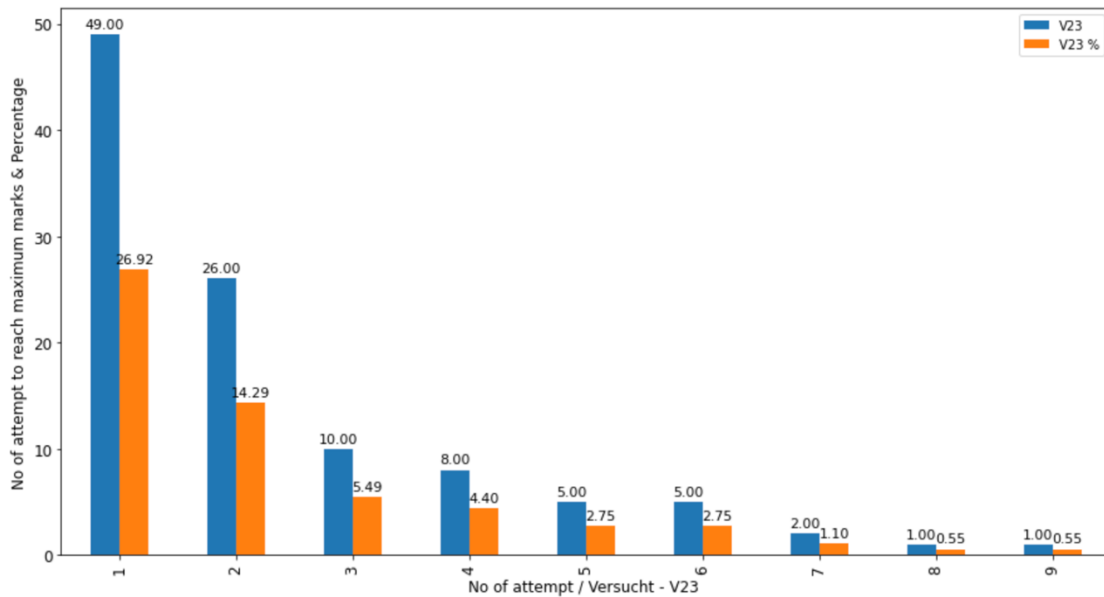


Figure 38 Best attempts of V23

Explanation V23: - ‘No of attempt to reach maximum marks’ in V23 was is than 50 attempts in first try, and it decreased in second try, when about third try bar in V23 ended with a 10 attempt. However, the figure experienced a steady decreased to the last try.

6.1.2 Total attempts

Creating a Bar Chart will provide a representation that is visual in nature of given data set or the data distribution. Bar Chart display the frequency of the data values and large amount of data. The Bar Chart helps in determining the median and the distribution of the given dataset. Also, this can display any outliers in the given set of data.

In Matplotlib, we use the bar as kind in plot function to create Bar Chart. This function will use an array of numbers to create a Bar Chart, the array is use function as an argument. Below all Bar Chart gives information about students achieved marks in different assignment. X-axis number of marks and Y-axis number of attempts.

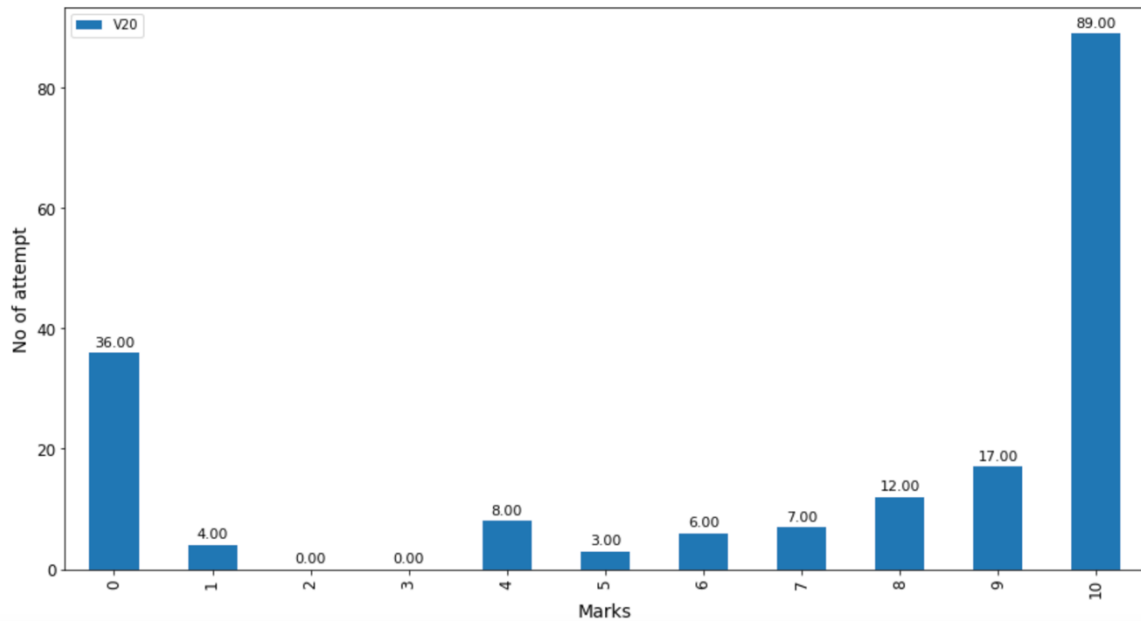


Figure 39 Number of attempt at V20

V20: - Use Bar Chart to understand the center of the data. In the Bar Chart above, you can see that the mean value is 7.01. Most values in the dataset will be close to 8 to 10 marks. The distribution is roughly symmetric and the values fall between approximately 4 to 7 marks.

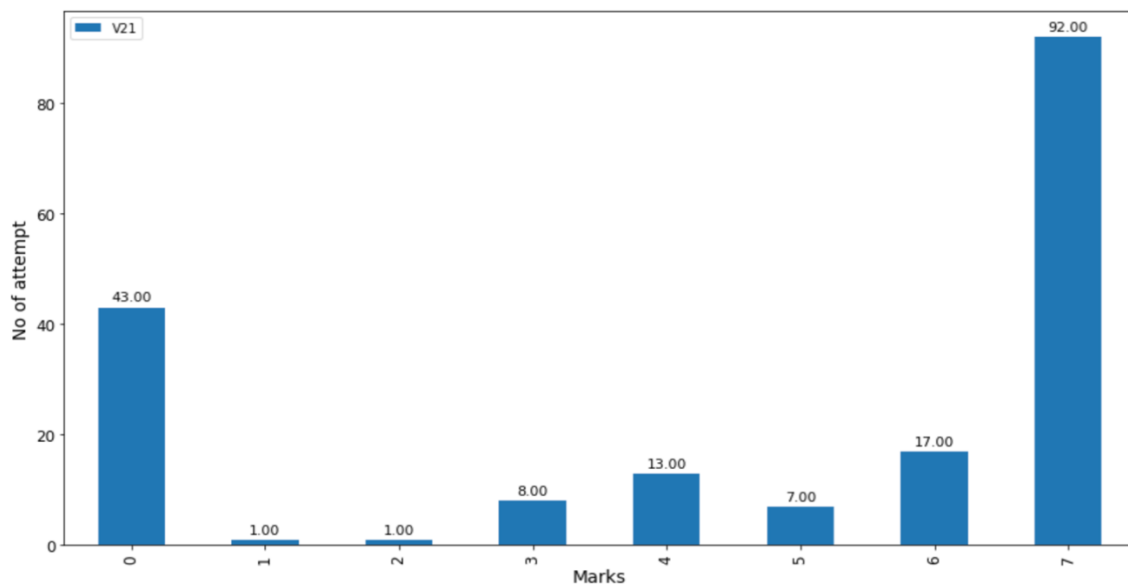


Figure 40 Number of attempt at V21

V21: - V21 Bar Chart is right-skewed distributions, the long tail extends to the right while most values cluster on the left, as shown below. The 92 students have 7 marks in second assignment. The mean value is 2.63.

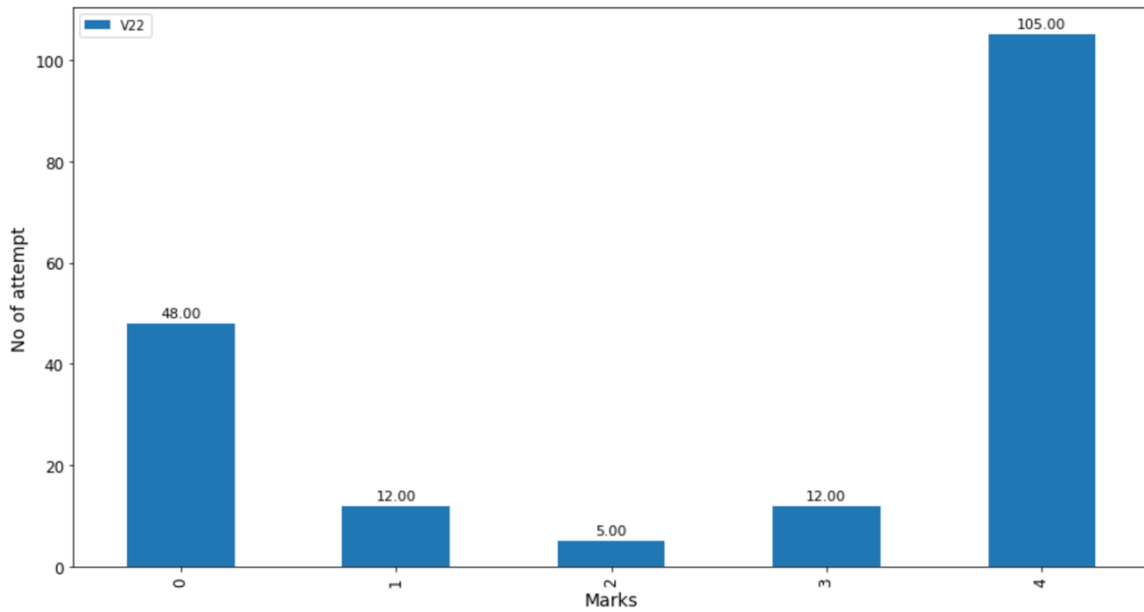


Figure 41 Number of attempt at V22

V22: - This Bar Chart shows information about V22 column of dataset. Marks between 0 to 4 were quite frequent. Finally, due to the atypical large value, the Bar Chart is slightly skewed to the right, or positively skewed. Without this value, the Bar Chart would be reasonably symmetric.

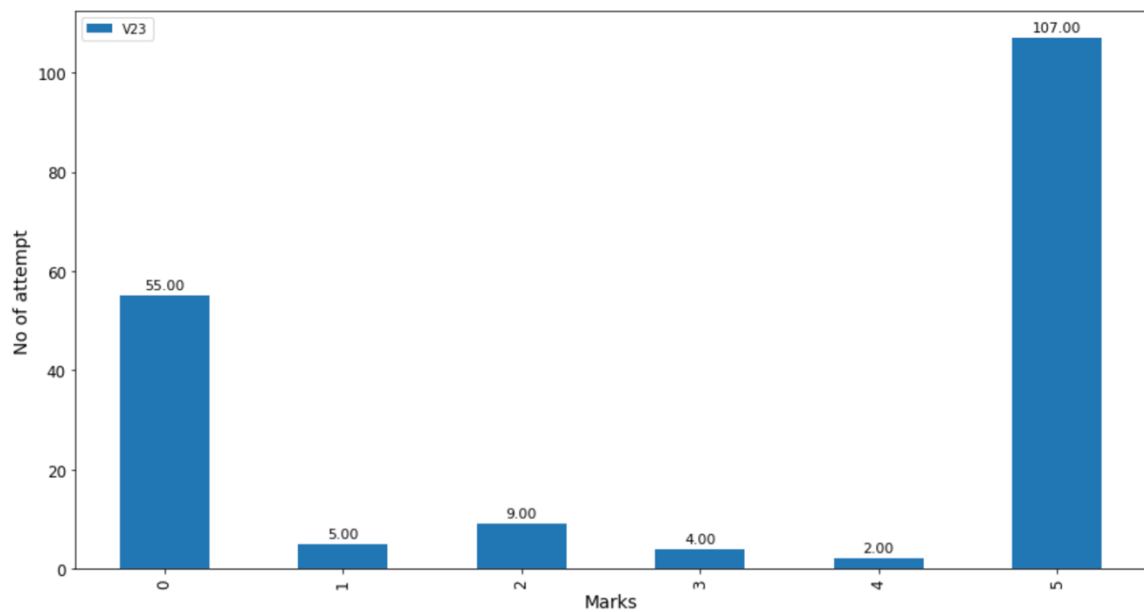


Figure 42 Number of attempt at V23

V23: - The Most of student are got 5 marks in assignment 4 as known as column V23. The range of this start with 0 and end with 5. The graph is right skewed.

```

def get_summary_statistics(df2):
    mean = np.round(np.mean(df2), 2)
    median = np.round(np.median(df2), 2)
    min_value = np.round(df2.min(), 2)
    max_value = np.round(df2.max(), 2)
    quartile_1 = np.round(df2.quantile(0.25), 2)
    quartile_3 = np.round(df2.quantile(0.75), 2)
    # Interquartile range
    iqr = np.round(quartile_3 - quartile_1, 2)
    print('Min: %s' % min_value)
    print('Mean: %s' % mean)
    print('Max: %s' % max_value)
    print('25th percentile: %s' % quartile_1)
    print('Median: %s' % median)
    print('75th percentile: %s' % quartile_3)
    print('Interquartile range (IQR): %s' % iqr)

print('\n\nV20 summary statistics \n')
get_summary_statistics(df2.V20)

print('\n\nV21 summary statistics \n')
get_summary_statistics(df2.V21)

print('\n\nV22 summary statistics \n')
get_summary_statistics(df2.V22)

print('\n\nV23 summary statistics \n')
get_summary_statistics(df2.V23)

```

Figure 43 statistic summary

<p>V20 summary statistics</p> <p>Min: 0 Mean: 7.01 Max: 10 25th percentile: 4.0 Median: 9.0 75th percentile: 10.0 Interquartile range (IQR): 6.0</p>	<p>V22 summary statistics</p> <p>Min: 0 Mean: 2.63 Max: 4 25th percentile: 0.0 Median: 4.0 75th percentile: 4.0 Interquartile range (IQR): 4.0</p>
<p>V21 summary statistics</p> <p>Min: 0 Mean: 4.73 Max: 7 25th percentile: 3.0 Median: 7.0 75th percentile: 7.0 Interquartile range (IQR): 4.0</p>	<p>V23 summary statistics</p> <p>Min: 0 Mean: 3.18 Max: 5 25th percentile: 0.0 Median: 5.0 75th percentile: 5.0 Interquartile range (IQR): 5.0</p>

Figure 44 Summary statistics

6.1.3 Failure attempt

The Below Bar Chart Shows the Failure attempts. Refer to the scale range of the y-axis to determine the total attempts and the x-axis determine the No of students.

The Following Bar Chart is Left skewed and it's represented information about How many students have done their task for the first time and how many students take a long time to complete the task.

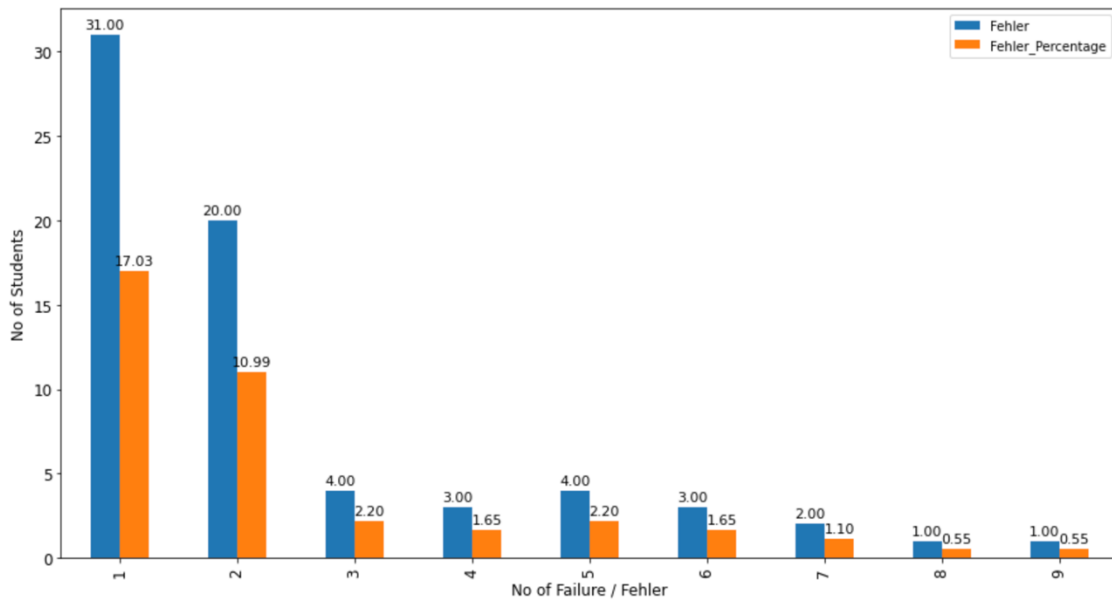


Figure 45 Bar Chart of number of students vs number of Failure

Explanation: -

As we can see in Bar Chart first bar, 31 students have done the task on the first tries which means they done the task with 0 Failure on the hand 17.03% Students has been complete task in first try. The second bar has 20 values it shows 20 students have done the task a second time with 1 failure.

The two groups of students have achieved the same bar chart. The third bar has a complete task with 2 Failures and the fifth bar has done in 4 failures and both bars have 4 student values. As mention above one student has taken 9 tries to achieve maximum marks so we can say that they did with 8 failures.

In conclusion, Number of Failure is independent variable and Number of students is dependent variable, Number of Failure increasing compared to Number of students decreasing.

6.1.4 Mean Value Chart

A bar Chart is a represent of a distribution of data. A common way of visualizing the distribution of a single numerical variable is by Bar Chart. A bar chart a divides the value within a numerical variable into “bins” and counts the number of observations that in to each bin.

	V20	V21	V22	V23
Versucht				
1	6.96	5.17	2.86	3.53
2	6.93	4.65	2.57	3.07
3	7.48	4.33	2.76	2.76
4	7.57	4.14	2.14	2.93
5	6.18	4.09	1.82	2.27
6	8.00	4.43	2.71	3.57
7	5.00	3.50	2.00	2.50
8	5.00	3.50	2.00	2.50
9	10.00	7.00	4.00	5.00

Figure 46 All attempt mean value

Figure number 46 Shows all attempt (one to nine attempt) mean value respect to different assignment and its store in one dataframe.

We have created a bar chart using four different assignments mean value in x-axis and number of marks in y-axis. As shown above, we have created a histogram with 9 bins with each has sub 4 bin. The horizontal axis shows total 36 bin. We can see in the graph 4 assignment which V20 indicate by blue bin, V21 shows in orange bin, V22 for green bin and V23 represent by red bin.

The data in the below graph are Random distribution.

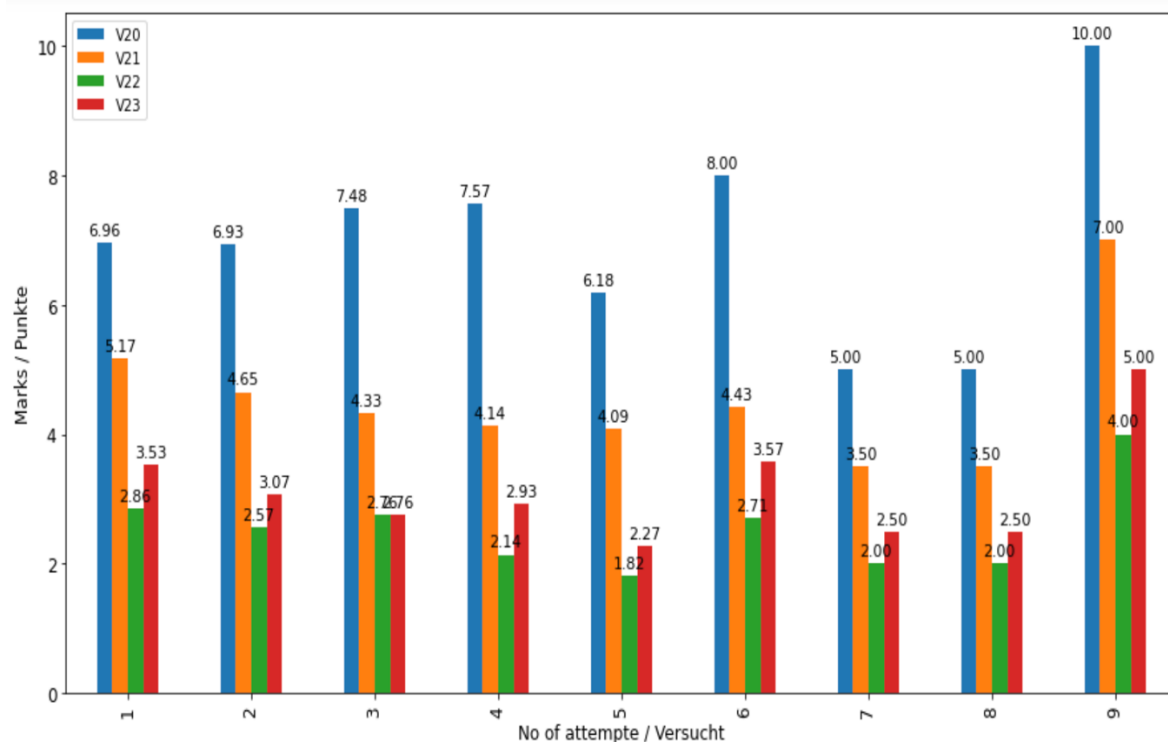


Figure 47 Bar chart of all attempts

Explanation of V20: - We can see in the above graph the different 9 mean value of the first assignment. The graph is a random distribution, that means the value of numbers of attempt is always fluctuating. The mean value of 1st attempt is 6.96 and the mean value of 2nd attempt is 6.93. Compare between the first attempt and second attempt values little bit goes down. Let's see the results of the 7th and 8th attempts, which mean value is 5. In the last bin 9th attempt, students got a 10.0 mean value.

Explanation of V21: - The above graph shows the random distribution, but we can divide it into three parts. The first part is right-skewed distribution, it starts to 1st attempt till the end of the 5th attempt. After a little bit of increase and one more time till 9th attempt decrease, and in the last end of 7 mean value at 9th attempt.

Explanation of V22: - The histogram gives information about the mean value of the third attempt by the student. We can see in the graph green bins show random distribution, it is the range between 2.86 to 4.00, and the lowest mean value in the graph it is 1.82.

Explanation of V23: - The last attempt of assignment shows in a graph with help of the red bin and it is also random distribution. It is range between 3.53 to 5.00 mean value.

6.2 Statistical Analysis

Investigating new trends, patterns, and relationships by quantitative data it is called Statistical analysis. Descriptive statistics include those that summarize the central tendency, dispersion and sharp of a dataset distribution, excluding NaN values. For numeric data, the result index will include count, mean, std, min, max further as lower, 50 and upper percentiles. By default, the lower percentile is 25 and therefore the upper percentile is 75. The 50 percentile is that the same because the median.

For object data (e.g., strings or timestamps), the result index will include count, unique, top, and freq. The highest is that the commonest value. The freq is that the most typical value frequency. Timestamps also include the primary and last items. If multiple object values have the best count, then the count and top results are going to be arbitrarily chosen from among those with the very best count.

	V19	V20	V21	V22	V23
count	76.00	76.00	76.00	76.00	76.00
mean	1.00	6.96	5.17	2.86	3.53
std	0.00	3.87	2.72	1.72	2.13
min	1.00	0.00	0.00	0.00	0.00
25%	1.00	4.00	4.00	1.00	2.00
50%	1.00	9.00	7.00	4.00	5.00
75%	1.00	10.00	7.00	4.00	5.00
max	1.00	10.00	7.00	4.00	5.00

Figure 48 Descriptive statistic of First attempt

The mean is the usual average: it shows in image number 28 with blue rectangle. The largest value in the list is 10,7,4 and 5 indexes of V20, V21, V22 and V23 and smallest value is 0 for all index.

	V19	V20	V21	V22	V23		V19	V20	V21	V22	V23		V19	V20	V21	V22	V23
count	46.00	46.00	46.00	46.00	46.00	count	21.00	21.00	21.00	21.00	21.00	count	14.00	14.00	14.00	14.00	14.00
mean	2.00	6.93	4.65	2.57	3.07	mean	3.00	7.48	4.33	2.76	2.76	mean	4.00	7.57	4.14	2.14	2.93
std	0.00	3.93	2.98	1.80	2.33	std	0.00	3.91	2.65	1.64	2.36	std	0.00	4.15	3.08	1.79	2.50
min	2.00	0.00	0.00	0.00	0.00	min	3.00	0.00	0.00	0.00	0.00	min	4.00	0.00	0.00	0.00	0.00
25%	2.00	4.25	0.75	0.00	0.00	25%	3.00	6.00	3.00	1.00	0.00	25%	4.00	8.25	0.75	0.25	0.00
50%	2.00	9.00	6.50	4.00	5.00	50%	3.00	10.00	5.00	4.00	3.00	50%	4.00	10.00	5.00	2.00	5.00
75%	2.00	10.00	7.00	4.00	5.00	75%	3.00	10.00	7.00	4.00	5.00	75%	4.00	10.00	7.00	4.00	5.00
max	2.00	10.00	7.00	4.00	5.00	max	3.00	10.00	7.00	4.00	5.00	max	4.00	10.00	7.00	4.00	5.00

Figure 49 Descriptive statistic of Second, Third and Fourth attempt

Pandas Describe does exactly what it seems like, describe your data. Describe will return a series of descriptive information. This Series will tell you:

- The count of values
- The number of unique values
- The top (most frequent) value
- The frequency of your top value
- The mean, variance, min and max values
- The percentiles of your data: 25%, 50%, 75% by default

	V19	V20	V21	V22	V23		V19	V20	V21	V22	V23		V19	V20	V21	V22	V23
count	11.00	11.00	11.00	11.00	11.00	count	7.00	7.00	7.00	7.00	7.00	count	4.00	4.00	4.00	4.00	4.00
mean	5.00	6.18	4.09	1.82	2.27	mean	6.00	8.00	4.43	2.71	3.57	mean	7.00	5.00	3.50	2.00	2.50
std	0.00	4.94	3.36	1.94	2.61	std	0.00	3.70	3.36	1.89	2.44	std	0.00	5.77	4.04	2.31	2.89
min	5.00	0.00	0.00	0.00	0.00	min	6.00	0.00	0.00	0.00	0.00	min	7.00	0.00	0.00	0.00	0.00
25%	5.00	0.00	0.00	0.00	0.00	25%	6.00	8.00	1.50	1.50	2.50	25%	7.00	0.00	0.00	0.00	0.00
50%	5.00	10.00	6.00	1.00	0.00	50%	6.00	10.00	7.00	4.00	5.00	50%	7.00	5.00	3.50	2.00	2.50
75%	5.00	10.00	7.00	4.00	5.00	75%	6.00	10.00	7.00	4.00	5.00	75%	7.00	10.00	7.00	4.00	5.00
max	5.00	10.00	7.00	4.00	5.00	max	6.00	10.00	7.00	4.00	5.00	max	7.00	10.00	7.00	4.00	5.00

Figure 50 Descriptive statistic of Fifth, Sixth and Seventh attempt

	V19	V20	V21	V22	V23		V19	V20	V21	V22	V23
count	2.00	2.00	2.00	2.00	2.00	count	1.00	1.00	1.00	1.00	1.00
mean	8.00	5.00	3.50	2.00	2.50	mean	9.00	10.00	7.00	4.00	5.00
std	0.00	7.07	4.95	2.83	3.54	std	nan	nan	nan	nan	nan
min	8.00	0.00	0.00	0.00	0.00	min	9.00	10.00	7.00	4.00	5.00
25%	8.00	2.50	1.75	1.00	1.25	25%	9.00	10.00	7.00	4.00	5.00
50%	8.00	5.00	3.50	2.00	2.50	50%	9.00	10.00	7.00	4.00	5.00
75%	8.00	7.50	5.25	3.00	3.75	75%	9.00	10.00	7.00	4.00	5.00
max	8.00	10.00	7.00	4.00	5.00	max	9.00	10.00	7.00	4.00	5.00

Figure 51 Descriptive statistic of Eighth and Ninth attempt

In the below image create one dataset and store all the mean value. After collecting all mean value try to convert into graphical view (Histogram). The Histogram and analysis find in 6.2 section.

6.3 Cluster Analysis

Interpretation

Use the cluster centroid as a general measure of cluster location and to assist interpret each cluster. Each centroid is seen as representing the "average observation" within a cluster across all the variables within the analysis. Minitab calculates the distances between the centroids of the clusters that are included within the final partition. For every cluster, Minitab also calculates various distance measures between the cluster centroid and therefore the observations within the cluster. For more information, see the subject for every distance measure.

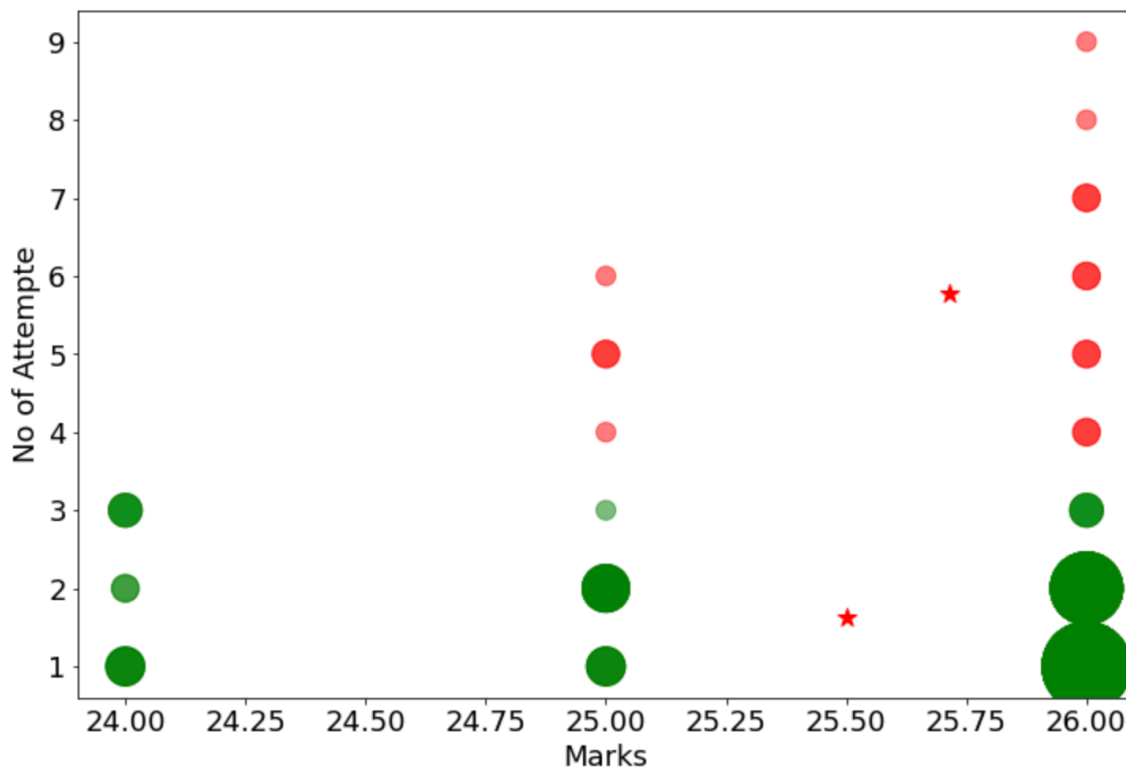


Figure 52 Cluster Analysis

From the above, we see that the clustering algorithm demonstrates an overall positive correlation between Number of attempts and Marks, implying that Most of Student have achieved 26 and 25 marks in first and second try. While this is a more simplistic example and could be modelled through linear regression analysis, there are many instances where relationships between data will not be linear and k-means can serve as a valuable tool in understanding the data through clustering methods.

Let's take an example. We have data about student, who has complete task in several attempt. for instance, Student number 37 has all the task in 6 attempts complete and Student number 34 has

completed all task in exactly 2 try. There you may see different varieties of number of attempts. The one thing you will notice there that the marks are going to be arranged in a very group of their types. Like all the 24 are going to be kept in one place, 25 are going to be kept with their kinds then on. If you may notice here then you will find that they are forming a bunch or cluster, where each of the marks is kept within their quite group forming the clusters.

Now, take a look at the above two figures. Scatter Plot and scatter plot with cluster. What did you observe? allow us to discuss the primary figure. the primary figure shows the information before applying the k-means clustering algorithm. Here all two different categories are messed up. After you will see such data within the globe, you may not capable to work out the various categories.

Now, examine the second figure (fig 2). This shows the information after applying the K-means clustering algorithm. you'll see that everyone two different items are classified into two different categories which are called clusters.

7 Summary and Outlook

In conclusion, The examination of E-Learning data analysis of 76 students of curriculum activity which total 182 attempts students had taken to complete the different assignment. According to data visualization, most of the students has done assignment in 1st and 2nd tries. As well as the failure ratio is also increased in the 1st and 2nd tries and later on it decrease because of students has complete assignment. The final summary about cluster analysis is that students need more time to achieve maximum marks, in detailed students took nine attempts to achieve 26 marks compared to 25- and 24-marks index 6, and only 3 attempts were needed to achieve those marks.

There are still plenty of ideas for this thesis that can be implemented. Multiple models are available to improve e-Learning data analysis like the VAK model, Kolb's model, and Felder-Soloman Model. All the models are similar to each other, but I also recommended Felder-Soloman Model because it is a bit more related to this thesis.

The Felder-Soloman Method is the Felder-Soloman Learning Style Model. The Model is based on How we process, Perceive, Receive and understand information. This model Classify categories into students/learners based on the different levels of the learning process. In this model, instructors know which category a student belongs to. The instructors deliver the contents as what information students need to complete the task. For example, active students need a little bit less attention compared to passive students. So, instructors should give hire more attention to passive students, who are taking a long time to complete tasks. This turn can speed to improve student activity.

Summary of thesis of data cleaning, transforming and modeling data to find useful information for improve education decision-making. In conclusion, it must be said that the project was a great enrichment for me and that it significantly broadened my horizons and my skills.

8 Abbreviation

API	Application program interface
GUI	Graphics User Interphase
BLAS	Basic Linear Algebra Subprograms
LAPACK	Linear Algebra Package
BSD	Berkeley Source Distribution
IPYTHON	Interactive Python
CLI	Command-line interface
IDE	Integrated Development Environment
HTML	HyperText Markup Language
HTTP	Hyper Text Transfer protocol
JSON	JavaScript Object Notation
SQL	Structured Query Language
VS	Visual Studio
WSN	Wireless Sensor Networks
OOP	Object-Oriented Programming
REST	Representational state transfer
URL	Uniform Resource Locator
XML	Extensible Markup Language
ANOVA	Analysis of Variance
kNN	k-Nearest Neighbors
MSE	Mean Squared Error
R^2	R-Squared
KDD	Knowledge Discovery Databases
CSV	Comma Separate Value
SSE	Sum of Squared Error
HTTPS	Hypertext Transfer Protocol Secure
FTP	File Transfer Protocol
SSL	Secure Sockets Layer

9 List of Figure

Figure 1 Python IDLE	14
Figure 2 Python Interactive session	14
Figure 3 debugging and prototyping	15
Figure 4 Jupyter Notebook.....	18
Figure 5 Anaconda Navigator	19
Figure 6 Dashboard of Jupyter Notebook	20
Figure 7 Command line.....	20
Figure 8 Import Libraries	29
Figure 9 Read data.....	29
Figure 10 Prepare Model.....	29
Figure 11 K_range.....	30
Figure 12 Elbow Method.....	30
Figure 13 Model Implimentation	31
Figure 14 Clustering.....	31
Figure 15 Row Data	35
Figure 16 Import Libraries	36
Figure 17 Read data and print first five data.....	37
Figure 18 Clean and useful dataset	38
Figure 19 Create DataFrame 2	38
Figure 20 Extract data	39
Figure 21 Code of extract First try.....	39
Figure 22 Data collection of first try.....	39
Figure 23 Code of extract second try	40
Figure 24 Data collection of second try	40
Figure 25 Data collection of all attempts	40
Figure 26 Best attempts Results	41
Figure 27 Extract data from dataframe_1	42
Figure 28 Create Dataframe for Failure/Fehler attempts	43
Figure 29 Import Library.....	43
Figure 30 Scatter Plot.....	44
Figure 31 Sum of Squire Error.....	44
Figure 32 Elbow Method.....	45
Figure 33 Cluster Prediction	45
Figure 34 Cluster Center	45
Figure 35 Best attempts of V20.....	46

Figure 36 Best attempts of V21.....	47
Figure 37 Best attempts of V22.....	47
Figure 38 Best attempts of V23.....	48
Figure 39 Number of attempt at V20	49
Figure 40 Number of attempt at V21	49
Figure 41 Number of attempt at V22	50
Figure 42 Number of attempt at V23	50
Figure 43 statistic summary	51
Figure 44 Summary statistics	51
Figure 45 Bar Chart of number of students vs number of Failure	52
Figure 46 All attempt mean value	52
Figure 47 Histogram of all attempts.....	53
Figure 48 Descriptive statistic of First attempt	54
Figure 49 Descriptive statistic of Second, Third and Fourth attempt	55
Figure 50 Descriptive statistic of Fifth, Sixth and Seventh attempt.....	55
Figure 51 Descriptive statistic of Eighth and Nineth attempt	55
Figure 52 Cluster Analysis.....	56

10 List of Tables

Table 1 Relation between quintile, quartile and percentile.....	23
Table 2 Example of Inter Quartile Range	24
Table 3 Sorted data.....	24

11 References

- [1] **Manideep Paduchuri** GeeksforGeek [Online] // GeeksforGeek. - 20 April 2021. - <https://www.geeksforgeeks.org/python-language-advantages-applications/?ref=lbp>.
Python Programming/Interactive mode [Online] // WIKIBOOKS. - Creative Commons Attribution-ShareAlike License . - 20 April 2021. - https://en.wikibooks.org/wiki/Python_Programming/Interactive_mode.
- [2] **Advani Vaishali** 34 Open-Source Python Libraries You Should Know About [Online] // Great Learning. - 11 September 2020. - 05 May 2021. - <https://www.mygreatlearning.com/blog/open-source-python-libraries/>.

Studying the common uses of E-Learning information technology essay [Online] // essays.pw. - 30 November 2015. - 05 May 2021. - <https://essays.pw/essay/studying-the-common-uses-of-e-learning-information-technology-essay-206664>.

[3] **DannyS712** Python Programming/Interactive mode [Online] // Wikibooks. - May 2020. - 05 May 2021. - https://en.m.wikibooks.org/wiki/Python_Programming/Interactive_mode.

[4] **Frost Jim** Introduction to Statistics [Book].

[5] **Vedantu** Quartile Deviation [Online] // Vedantu Learn Live Online. - Vedantu Innovation Pvt. Ltd, 06 May 2021. - 06 May 2021. - <https://www.vedantu.com/maths/quartile-deviation>.

[6] **Z.H.Tatli** Computer based education: Online Learning and teaching facilities [Online] // Research gate. - January 2009 Computer based education: Online Learning and teaching facilities. - 05 May 2021. - https://www.researchgate.net/publication/288156487_Computer_based_education_Online_learning_and_teaching_facilities.

[7] **Rosenberg Marc J.** Beyond E-Learning [Online] // Google book. - John Wiley & Sons, 13 December 2005, - Approaches and Technologies to Enhance Organizational Knowledge, Learning, and Performance. - 05 May 2021. - https://books.google.de/books?id=M5REo6Qj74gC&redir_esc=y.

[8] **Wickramarathna Nirodha** Python lesson 1(Introduction to python) [Online] // MAT TECH. - May 2020, Python lesson 1(Introduction to python). - 05 May 2021. - <https://www.maztars.com/python-lesson-1-introduction-to-python/>.

[9] **Joshi Harsh** Learning-Object-Oriented-Python [Online] // medium.com. - 05 October 2019, Learning-Object-Oriented-Python. - 05 May 2021. - https://medium.com/@joshharsh/learning-object-oriented-python-3aa8dd07750f?id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6IjY5ZWQ1N2Y0MjQ0OTEyODJhMTgwMjBmZDU4NTk1NGI3MGJiNDVhZTAiLCJ0eXAiOiJKV1QiLCJ0eXciOiJpc3MiOiJodHRwczovL2FjY291bnRzLmdvb2dsZS5jb20iLCJuYmYiOiJlMjMjAyMDk5O.

[10] **Chavan Siddhesh** Top 5: Online Notebook(ipynb) and other cloud services [Online] // medium.com. - 21 August 2018, Top 5: Online Notebook(ipynb) and other cloud services. - 05 May 2021. - <https://medium.com/@siddesh.001/top-5-online-free-notebook-ipynb-and-other-cloud-services-dbf9580d99e3>.

[11] **Dar Pranav** Comprehensive Beginner's Guide to Jupyter Notebooks for Data Science & Machine Learning [Online] // Analytics Vidhya. - 24 May 2018, Comprehensive Beginner's Guide to Jupyter Notebooks for Data Science & Machine Learning. - 05 May 2021. - <https://www.analyticsvidhya.com/blog/2018/05/starters-guide-jupyter-notebook/>.

[12] **Tulasi B** Learning Analytics and Big Data in higher Education [Online] // ijert.org. - IJERT, 01 February 2014, Learning Analytics and Big Data in higher Education. - 06 May 2021. - <https://www.ijert.org/learning-analytics-and-big-data-in-higher-education>.

[13] **Economic Times** The Economic Times [Online]. - 01 July 2021 ,Times The Economic. - 01 July 2021. - <https://economictimes.indiatimes.com/definition/e-learning>.

[14] **Tamm Sander** What is E-Learning? [Online]. - 21 December 2020, Defining what is e-learning is not as easy as it might first appear.. - 01 July 2021. - <https://e-student.org/what-is-e-learning/>.

[15] **linuxtopia** Script Mode [Online] // Linuxtopia. - 05 May 2021, Script Mode . - 01 July 2021. - https://www.linuxtopia.org/online_books/programming_books/python_programming/python_ch03s03.html.

[16] **Pro.Dr.Liebscher** Vorherige Links der Breadcrumb Computergestützte Datenanalyse [Online]. - 15 July 2021. - 15 July 2021. - https://homeportal.hsmerseburg.de/sendfile.php?type=0&file_id=a5d8376e87cface89664d5ae6e8a89f1&file_name=Vlcdans.pdf.

[17] **Omedes Jose** What are e-Learning analytics about? [Online]. - 10 September 2020, What are e-Learning analytics about?. - 16 July 2021. - <https://www.iadlearning.com/e-learning-analytics/>.

[18] **Mistry Jiger** 7 World-class Companies that use Python [Online]. - 15 April 2021-Top 7 Companies That Use Python Are Making A Mark In 2021. - 17 July 2021. - <https://www.monocubed.com/companies-that-use-python/>.

12 Hochschule Library and Online Library

1. KGTorrent: A Dataset of Python Jupyter Notebooks from Kaggle
Authors: - Quaranta, Luigi
Calefato, Fabio
Lanubile, Filippo
Quelle: - 2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)
MSR Mining Software Repositories (MSR), 2021 IEEE/ACM 18th International Conference on.
:550-554 May, 2021
2. Title: - Learning Management System
Authors: - Micro Lang
3. Primer on Process Mining: Practical Skill with Python
Authors: - Diogo R. Ferreira
4. Statistische und maschinelles lernen: gängige Verfahren im Überblick
Authors: - Stefan Rich
5. Datenanalyse mit SPSS
Authors: - Peter P. Eckstein
6. Computergeschützt Datenanalysis
Link: - <https://www.hs-merseburg.de/liebscher-eckhardprof-dr-rer-nat-habil/lehre>
Authors: - Pro. Do. Eckhard Liebscher
7. Machine Learning, Cryptography
Material: - Home Portal and Study Material
Authors: - Pro. Dr. Michael Schenke

Thanksgiving

At this point I would like to thank everyone who made this bachelor thesis possible through their professional and personal support. I would like to thank Prof. Dr. rer. nat. habil. Eckhard Liebscher for looking after my topic and for all the helpful advice on my bachelor thesis. I would also like to thank Dr. Benjamin Wacker for the pleasant supervision as a second reviewer.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich diese Bachelorarbeit selbstständig und ohne unzulässige Hilfsmittel verfasst habe. Alle verwendeten Quellen sind im Literaturverzeichnis dieser Arbeit angegeben. Diese Arbeit wurde bisher noch nicht an anderer Stelle zur Begutachtung eingereicht.

Ort Datum Unterschrift