A
Project Report
On
"Data Science Internship with CRM"

(CS453 - Software Project Major)

Prepared by

Parth Niteshkumar Patel (19DCS098)

Under the Supervision of

Prof. Dipak Ramoliya

Submitted to

Charotar University of Science & Technology (CHARUSAT)
for the Partial Fulfillment of the Requirements for the
Degree of Bachelor of Technology (B.Tech.)
in Computer Science & Engineering (CSE)
for 8th semester B.Tech

Submitted at

Accredited with Grade A+ by NAAC
Accredited with Grade A by KCG

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Faculty of Technology & Engineering (FTE), CHARUSAT
At: Changa, Dist: Anand, Pin: 388421.
April, 2023

DECLARATION BY THE CANDIDATE

I hereby declare that the project report entitled "**Data Science Internship with CRM**" submitted by me to Devang Patel Institute of Advance Technology and Research, Changa in partial fulfilment of the requirement for the award of the degree of **B.Tech** in Computer Science & Engineering, from Department of Computer Science & Engineering, DEPSTAR-FTE, CHARUSAT, is a record of bonafide CS453 Software Project Major (project work) carried out by me under the guidance of **Prof. Dipak Ramoliya**. I further declare that the work carried out and documented in this project report has not been submitted anywhere else either in part or in full and it is the original work, for the award of any other degree or diploma in this institute or any other institute or university.

NOTE: During the course of their internship, the intern has been involved in multiple projects, working on small chunks at a time and not on a single project. This approach has enabled the intern to employ a variety of techniques in different projects. This report highlights some of the techniques that the intern has utilized during their time at the company.

AVM COMPUTER VADODARA

AVM CLUTCH.CA COMPUTERS

Parth Niteshkumar Patel (19DCS098)

Prof. Dipak Ramoliya
Asst. Professor
Department of Computer Science & Engineering,
DEPSTAR-FTE, CHARUSAT-Changa.

clutch

AVM COMPUTERS PVT LTD (NOW CLUTCH)

Email: manager.intern@avmcomputer.co.in

Date: 19-04-2023

TO WHOM IT MAY CONCERN

This is to certify **PARTH NITESHKUMAR PATEL**, a student of Charotar University of Science and Technology (CHARUSAT), Changa, have successfully completed their internship tenure at AVM Computers PVT LTD. (now merged into Clutch)
Following are the internship details:

POSITION: **DATA SCIENCE INTERN (WITH CRM)**

DURATION: **4 MONTHS (APPROX. 16-18 WEEKS)**

START DATE: **19 DECEMBER 2022**

END DATE: **19 APRIL 2023**

STIPEND: **INR 20,000**

During this period, Parth worked on various projects as he had opted for the portfolio internship.
During the entire tenure, he was one of the top performers. He never missed any deadline, he was always punctual, always on time. He is a great problem solver, a great team player and a good manager.
The intern was successful in satisfying the minimum criteria for the internship completion.

Authorized Signatory

Ajay Shah

Company Stamp:

AVM COMPUTER • VADODARA •

AVM CLUTCH.CA COMPUTERS

CHARUSAT
CHAROTAR UNIVERSITY OF SCIENCE AND TECHNOLOGY

Accredited with Grade A+ by NAAC
Accredited with Grade A by KCG

CERTIFICATE

This is to certify that the report entitled “**Data Science Internship with CRM**” is a bonafied work carried out by **Parth Niteshkumar Patel(19DCS098)** under the guidance and supervision of **Prof. Dipak Ramoliya** & **Mr. Ajay Shah** for the subject **Software Project Major (CS453)** of 8th Semester of Bachelor of Technology in **Computer Science & Engineering** at Devang Patel Institute of Advance Technology and Research (DEPSTAR), Faculty of Technology & Engineering (FTE) – CHARUSAT, Gujarat.

To the best of my knowledge and belief, this work embodies the work of candidate himself, has duly been completed, and fulfills the requirement of the ordinance relating to the B.Tech. Degree of the University and is up to the standard in respect of content, presentation and language for being referred by the examiner(s).

Under the supervision of,

Prof. Dipak Ramoliya
Asst. Professor
Department of Computer Science & Engineering.
DEPSTAR-FTE, CHARUSAT, Changa, Gujarat

Mr. Ajay Shah
Team Leader
AVM Computers PVT LTD./ CLUTCH

Dr. (Prof.) Chirag Patel,
I/c. Head- Department of Computer Science & Engineering, DEPSTAR-FTE,
CHARUSAT, Changa, Gujarat

Dr. (Prof.) Amit Nayak
I/c. Principal-DEPSTAR,
CHARUSAT, Changa, Gujarat.

Devang Patel Institute of Advance Technology and Research (DEPSTAR)
Faculty of Technology & Engineering (FTE), CHARUSAT

At: Changa, Ta. Petlad, Dist. Anand, Pin:388421. Gujarat

ABSTRACT

This report outlines my experience as a data science and analytics intern, with a minor focus on customer relationship management (CRM), where I worked on multiple projects. The internship provided me with the opportunity to gain practical experience in various data science techniques and tools, as well as insights into the role of data analytics in enhancing customer engagement.

The report begins with an introduction to the company and the internship program, followed by an overview of the data science and analytics projects that I worked on during the internship. The projects included data cleaning and pre-processing, exploratory data analysis, statistical modelling, machine learning, and data visualization. In each project, I applied different techniques and tools to analyze data, draw insights, and develop actionable recommendations for stakeholders.

The report also discusses the minor focus on customer relationship management, highlighting the strategies and techniques used to analyze and improve customer engagement. This included customer segmentation analysis, churn prediction modelling, and personalized marketing campaign development. I learned how data analytics can play a critical role in understanding customer behaviour and preferences, enabling companies to enhance customer satisfaction and loyalty.

Throughout the internship, I worked closely with my supervisors, colleagues, and my internal guide who provided valuable guidance and support. They shared their expertise in data science and analytics, enabling me to develop my technical skills and gain practical experience in real-life scenarios. I also gained valuable insights into the role of teamwork, communication, and collaboration in delivering successful data science projects.

The report concludes with a summary of the key lessons learned during the internship, including the importance of effective data management, attention to detail, and a customer-centric approach to data analysis. Overall, the internship provided me with a solid foundation in data science and analytics, as well as an understanding of the role of data analytics in enhancing customer engagement.

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to everyone who has made my data science and analytics internship, with a minor focus on customer relationship management (CRM), a valuable and enriching experience.

Firstly, I would like to thank the management team at the company for providing me with the opportunity to intern with them. I am grateful for the guidance and support that I have received from my supervisors, who have shared their expertise and provided valuable feedback throughout my internship.

I would also like to extend my appreciation to my colleagues, who have made my internship experience engaging and enjoyable. They have provided a collaborative and inclusive environment, and I have learned a lot from their expertise in various aspects of data science, analytics, and customer relationship management.

I would like to thank the data science and analytics teams at the company, who have shared their knowledge and skills with me, enabling me to learn new techniques and approaches to data analysis. Their mentorship has been invaluable, and I am grateful for the insights and tips that they have shared.

I would also like to express my appreciation to the clients of the company, who have provided me with the opportunity to apply my data science and analytics skills in real-life scenarios. Their trust and confidence in my abilities have been a source of inspiration throughout my internship.

Finally, I would like to thank my internal guide (Prof. Dipak Ramoliya) for his unwavering support throughout my internship. His encouragement has kept me motivated and focused on achieving my goals.

Thank you all for your support and guidance throughout my data science and analytics internship, where I worked on multiple projects with a minor focus on customer relationship management.

TABLE OF CONTENTS

ABSTRACT .... **IV**

ACKNOWLEGEMENT .... **V**

CHAPTER 1: INTRODUCTION **1**

1.1 INTERNSHIP SUMMARY 1

1.2 PURPOSE OF INTERNSHIP 2

1.3 OBJECTIVE OF INTERNSHIP 3

1.4 SCOPE OF INTERNSHIP 4

1.5 INTERNSHIP TECHNOLOGY & LITERATURE REVIEW 6

1.6 DISCLAIMER FROM THE ORGANIZATION 8

CHAPTER 2: PROJECT MANAGEMENT **9**

2.1 PROJECT PLANNING 9

2.1.1 GENERAL APPROACH USED IN THE PROJECTS 10

2.1.2 PROJECT EFFORTS, TIME & COST ESTIMATION 11

2.1.3 ROLES & RESPONSIBILITIES 14

2.2 PROJECT SCEDULING 16

CHAPTER 3: SYSTEM REQUIREMENTS STUDY **18**

3.1 SRS IN DATA ANALYTICS & SCIENCE PROJECTS 18

3.2 SOFTWARE & HARDWARE REQUIREMENTS 20

3.3 ASSUMPTIONS & DEPENDENCIES 21

CHAPTER 4: SYSTEM ANALYSIS **22**

4.1 STUDY OF CURRENT SYSTEM 22

4.2 FEASIBILITY STUDY 23

4.2.1 FEASIBILITY STUDY OF WEB SCRAPPING PROJECT 24

4.2.2 FEASIBILITY STUDY OF ANALYTICS PROJECT 25

4.2.3 FEASIBILITY STUDY OF TABLEAU PROJECTS 26

4.2.4 FEASIBILITY STUDY OF PANCARD TAMPERING 27

4.2.5 FEASIBILITY STUDY OF TEXT EXTRACTION 28

4.3 OO CONCEPTS IN DATA ANALYTICS & SCIENCE PROJECTS 29

4.4 E-R DIAGRAMS 30

4.5 UML DIAGRAMS 31

CHAPTER 5: MINI-PROJECTS IN DETAIL....32
5.1 DATA VISUALIZATION WITH TABLEAU 32
5.1.1 MINI-PROJECT-1....32
5.1.2 MINI-PROJECT-2....33
5.1.3 MINI-PROJECT-3....34
5.1.4 MINI-PROJECT-4....35
5.2 INSIGHTS PROJECTS 37
5.2.1 AMAZON JOB ANALYSIS 37
5.2.2 START-UP FUNDING PROBLEM 43
5.3 TMDB API PROJECT 46
5.4 REVIEW SCRAPPER MINI-PROJECT 50
5.5 PANCARD TAMPERING MINI-PROJECT....51
5.6 TEXT EXTRACTION MINI-PROJECT 55
5.7 CRM DUTY 59
5.8 INTERN MANAGEMENT....60

CHAPTER 6: TESTING....61
6.1 CODING STANDARDS....61
6.2 TESTING & TEST CASES....63
6.2.1 TEST CASES & SCENARIOS FOR PROJECT-1....65
6.2.2 TEST CASES & SCENARIOS FOR PROJECT-2....67
6.2.3 TEST CASES & SCENARIOS FOR PROJECT-3....68

CHAPTER 7: CONCLUSION 69
7.1 SELF-ANALYSIS & DISCUSSION 69
7.2 PROBLEMS ENCOUNTERED DURING THE INTERNSHIP 70
7.3 INTERNSHIP SUMMARY 73

BIBLIOGRAPHY....74

LIST OF FIGURES

Fig. 2.1 Agile Model Phases 9
Fig. 2.2 Gantt Chart for PROJ-011-T-0011.1 INTERNSHIP SUMMARY 16
Fig. 2.4 Gantt Chart for PROJ-027-T-001-003 16
Fig. 2.3 Gantt Chart for PROJ-012-T-001 17
Fig. 5.1 Mini-Project-1 Dashboard 32
Fig. 5.2 Mini-Project-2 Dashboard 33
Fig. 5.3 Mini-Project-3 Dashboard 34
Fig. 5.4 Mini-Project-4 Dashboard 35
Fig. 5.5 Insight-1 Code 37
Fig. 5.6 Output of above Question 38
Fig. 5.7 Code to find the answer of Question-2 39
Fig. 5.8 Output of the Question 39
Fig. 5.9 Output of the Question-2 40
Fig. 5.10 Code for Question-3 40
Fig. 5.11 Output of Question-3 41
Fig. 5.12 Code for Question-4 42
Fig. 5.13 Output for Question-5 42
Fig. 5.14 Output of Question-1 43
Fig. 5.15 Output of Question-2 44
Fig. 5.16 Code & Output for Question-1 46
Fig. 5.17 Code & Output for Question-2 46
Fig. 5.18 Code & Output for Question-3 46
Fig. 5.19 Code & Output for Question-4 47
Fig. 5.20 Code & Output for Question-5 47
Fig. 5.21 Code & Output for Question-5 48
Fig. 5.22 Code & Output for Question-6 48
Fig. 5.23 Code & Output for Question-7 48
Fig. 5.24 Code & Output for Question-8 49
Fig. 5.25 Code & Output for Question-9 50
Fig. 5.26 Search Page for the Product Review Scrapping 50
Fig. 5.27 Output after Scrapping is Completed 50
Fig. 5.28 Initial UI of the application 53
Fig. 5.29 Upload Section to upload the Pan Card Image 53
Fig. 5.30 Result of Tampered Pan Card 54

Fig. 5.31 Result of an original Pan Card 54
Fig. 5.32 Initial Page of the application..57
Fig. 5.33 Upload Section for the extraction...57
Fig. 5.34 Image Uploaded 58
Fig. 5.35 Text Extracted 58

1 INTRODUCTION

1.1 INTERNSHIP SUMMARY

During my internship, I had the opportunity to work on multiple projects, which provided me with a diverse range of experiences and skills. My work involved various stages of the data analysis pipeline, from data collection and pre-processing to analytics, visualization, and consulting. Additionally, I had the opportunity to contribute to intern management and customer relationship management (CRM).

I worked with a range of tools and technologies, including Python, SQL, Excel, Tableau, and Google Colab, VS Code , to complete my projects. I used Python to perform data cleaning, data wrangling, and statistical analysis. I used SQL to query and join large datasets, and Excel for data formatting and analysis. For data visualization, I used Tableau to create interactive dashboards and visualizations. I also had the opportunity to develop my consulting skills, where I worked closely with clients to identify their business requirements and provide them with data-driven insights and recommendations.

During the internship, I also had the opportunity to contribute to intern management by leading and mentoring fellow interns. I learned how to effectively communicate complex technical concepts to non-technical stakeholders and how to lead and manage projects in a team environment.

In addition, I was involved in minor customer relationship management, where I applied data analytics techniques to understand customer behaviour and preferences. This included customer segmentation analysis, churn prediction modelling, and personalized marketing campaign development. I learned how to effectively communicate the results of these analyses to stakeholders, and provide them with actionable insights to improve customer engagement. Overall, my internship provided me with a comprehensive understanding of the data analysis process, from data collection to insights delivery. I gained experience in using a range of tools and technologies to solve real-world business problems. I also developed my consulting, management, and communication skills.

1.2 PURPOSE OF INTERNSHIP

The purpose of my internship was to gain practical experience in data analytics and data science, with a focus on the entire data analysis pipeline from data collection to insights delivery. My work involved multiple projects, each of which required the use of a diverse range of skills and technologies, including data collection, data pre-processing, data analytics, data visualization, data consulting, minor customer relationship management, intern management, and relevant tools and technology.

I aimed to develop my technical skills in data analysis, using various programming languages such as Python to perform statistical analysis, data cleaning, and data visualization. Additionally, I sought to gain experience in data management, working with large datasets and database management systems such as SQL. Through my projects, I also aimed to improve my consulting skills, collaborating closely with clients to identify their requirements and provide them with data-driven insights and recommendations.

In addition to developing my technical skills, I also aimed to develop my managerial skills, such as project management, intern management, and customer relationship management. Through intern management, I learned how to effectively communicate complex technical concepts to non-technical stakeholders and mentor other interns. Through minor customer relationship management, I learned how to apply data analytics techniques to understand customer behavior and preferences, enabling me to provide actionable insights to improve customer engagement.

Overall, my internship aimed to provide me with a comprehensive understanding of the data analysis process, from data collection to insights delivery, and to develop both my technical and managerial skills in the field of data analytics.

1.3 OBJECTIVE OF THE INTERNSHIP

My objective for the internship was to gain practical experience in data science and analytics, with a focus on the entire data analysis pipeline, including data collection, data pre-processing, data analytics, data visualization, data consulting, minor customer relationship management, intern management, and relevant tools and technology. I aimed to work on multiple projects, each of which would require the use of a diverse range of skills and technologies.

Through the internship, I aimed to develop my technical skills in data analysis, using programming languages such as Python to perform statistical analysis, data cleaning, and data visualization. I also aimed to gain experience in data management, working with large datasets and database management systems such as SQL. Additionally, I aimed to improve my consulting skills, working closely with clients to identify their requirements and provide them with data-driven insights and recommendations.

In addition to developing my technical skills, I also aimed to develop my managerial skills, such as project management, intern management, and customer relationship management. Through intern management, I aimed to learn how to effectively communicate complex technical concepts to non-technical stakeholders and mentor other interns. Through minor customer relationship management, I aimed to learn how to apply data analytics techniques to understand customer behaviour and preferences, enabling me to provide actionable insights to improve customer engagement.

Overall, my objective for the internship was to gain a comprehensive understanding of the data analysis process, from data collection to insights delivery, and to develop both my technical and managerial skills in the field of data analytics. I aimed to use my experiences to prepare myself for a career in data analytics, where I can effectively apply data analytics and managerial skills to solve real-world business problems.

1.4 THE SCOPE OF THE INTERNSHIP

The scope of my internship involved working on multiple projects that required the use of data science and analytics skills, as well as managerial skills. Specifically, I was involved in data collection, data pre-processing, data analytics, data visualization, data consulting, minor customer relationship management, intern management, and the use of relevant tools and technology.

In terms of data collection, I was responsible for identifying and collecting relevant data from various sources, such as databases, APIs, and web scraping. I also had to ensure the accuracy and completeness of the data collected.

In data pre-processing, I was responsible for cleaning and transforming the data to ensure that it was ready for analysis. This involved techniques such as data normalization, imputation of missing values, and feature engineering.

Data analytics was a major component of my internship, where I had to apply techniques to analyze the data and derive insights. I used programming languages such as Python to perform data analysis, and used various libraries and frameworks such as NumPy, Pandas, and Scikit-learn to facilitate the analysis.

Data visualization was also a crucial part of my internship, where I had to create visualizations and dashboards to communicate insights to stakeholders. I used tools such as Tableau,Matplotlib to create these visualizations.

In data consulting, I had to work closely with clients to understand their business requirements and provide them with data-driven insights and recommendations. I had to communicate complex technical concepts to non-technical stakeholders in a clear and concise manner.

Minor customer relationship management involved using data analytics techniques to understand customer behavior and preferences, enabling me to provide actionable insights to improve customer engagement.

Intern management involved mentoring other interns and providing them with guidance and support throughout their projects. I also had to communicate with the project manager to ensure that the projects were on track and completed within the given timelines.

The scope of my internship did not involve making final business decisions, but rather providing recommendations based on data analysis. Additionally, my role did not involve software development or coding for production systems.

Overall, the scope of my internship was to gain practical experience in data science and analytics, and to develop my managerial skills. I was able to use a diverse range of skills and technologies to complete multiple projects and gain valuable experience in the field of data analytics.

1.5 INTERNSHIP TECHNOLOGY AND LITERATURE REVIEW

During my internship, I worked on multiple projects that required the use of various tools and technologies to perform data collection, pre-processing, analysis, visualization, and consulting. Here is an overview of the technologies and literature I used during my internship:

Data Collection:

- Databases: MySQL, MongoDB
- APIs
- Web Scraping: BeautifulSoup, Selenium

Data Pre-processing:

- Data Cleaning: Pandas, NumPy
- Data Transformation: Scikit-learn

Data Analytics:

- Statistical Analysis: Python
- Machine Learning: Scikit-learn, Teseract, and many more

Data Visualization:

- Tableau
- Matplotlib
- Seaborn

Customer Relationship Management:

- Salesforce

IDEs:

- Google Colab
- Tableau Public
- Tableau Server
- VS Code
- Gretl

Literature Review:

- Python for Data Analysis, 2nd Edition by Wes McKinney
- Data Science from Scratch, 2nd Edition by Joel Grus

The literature review involved reading and analyzing books and research papers related to data science and analytics. These resources provided me with a theoretical foundation for the techniques and technologies I used during my internship. I also consulted online forums and communities, such as Stack Overflow and Kaggle, for guidance and support.

Overall, the technologies and literature I used during my internship enabled me to perform data science and analytics tasks efficiently and effectively. I was able to apply my knowledge of these tools and technologies to complete multiple projects and develop my skills in the field of data science and analytics.

1.6 DISCLAIMER FROM THE ORGANIZATION

We, AVM Computers PVT LTD (now merged into the CLUTCH Canada) ,would like to state that the intern **Parth Niteshkumar Patel (I-003)** worked on real-time projects during their internship with us. However, in order to maintain the confidentiality, privacy, and security of both our clients and our organization, we strictly prohibit the intern from sharing any project codes or environments with any third party.

We understand that you may be interested in seeing the work completed by the intern, but we must prioritize the security of our clients' data and our organization's intellectual property. Therefore, we would like to inform you that the intern is not authorized to disclose any information about our clients, their data, or the projects they worked on.

However, we encourage the intern to verbally communicate about the strategies and methods they applied during the projects. Also, the intern is permitted to show dummy projects, which are similar to the ones they worked on but do not contain any sensitive information.

We apologize for any inconvenience this may cause, but we hope you understand our commitment to protecting the privacy and security of our clients' data and our organization's intellectual property.

Sincerely,

AVM COMPUTER • VADODARA •

A V M
CLUTCH.CA
COMPUTERS

2 PROJECT MANAGEMENT

2.1 PROJECT PLANNING

2.1.1 General Approach Used in Projects

AGILE MODEL

The Agile model is a project management methodology that emphasizes flexibility and adaptability. It is commonly used in software development projects, but can also be applied to data science and analytics projects.

The Agile model is based on the Agile Manifesto, which values individuals and interactions, working software, customer collaboration, and responding to change over following a rigid plan. It encourages iterative and incremental development, with frequent releases and continuous feedback from stakeholders.

In an Agile project, the team works in short sprints, typically lasting two to four weeks. Each sprint focuses on delivering a working product or feature, and the team conducts daily stand-up meetings to discuss progress and identify any obstacles. The team also engages in continuous testing and feedback to ensure that the product is meeting the needs of stakeholders.

One of the key benefits of the Agile model is its flexibility and ability to adapt to changing requirements or priorities. It allows for a more collaborative and responsive approach to project management, with frequent opportunities for stakeholder feedback and course correction. However, it also requires a high level of communication and coordination within the team, as well as a willingness to embrace change and uncertainty.

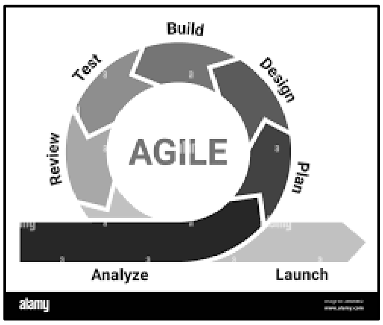


Fig. 2.1 Agile Model Phases

HOW AGILE MODEL IS USED IN DATA SCIENCE AND ANALYTICS PROJECT

Planning Phase: During the planning phase, the team defines the project goals and objectives, identifies the stakeholders, and breaks down the project into smaller tasks or features. The team then prioritizes these tasks based on their importance to stakeholders, and creates a backlog of features to be worked on during the project.

Sprint Phase: The sprint phase typically lasts two to four weeks and involves the team working on a subset of features from the backlog. During this phase, the team conducts daily stand-up meetings to discuss progress, identify any obstacles, and adjust priorities as needed. The team works on data collection, data cleaning, data modeling, data visualization, and other tasks, depending on the project requirements.

Review Phase: At the end of each sprint, the team conducts a review to demonstrate the work completed during the sprint and gather feedback from stakeholders. Based on this feedback, the team can make adjustments to the project's scope, timeline, or priorities for future sprints.

Retrospective Phase: The retrospective phase involves the team reflecting on the previous sprint and identifying areas for improvement. The team discusses what worked well and what did not work well, and identifies any changes to make for the next sprint.

Delivery Phase: The delivery phase involves delivering the completed project to the stakeholders. This may involve presenting insights and findings to stakeholders, delivering a dashboard or report, or integrating the results into a larger system.

Maintenance Phase: After the project is delivered, the team may need to provide ongoing maintenance and support to ensure the insights and findings remain relevant and useful over time.

Overall, Agile methodology provides a flexible and iterative approach to data analytics project management that can help teams to be more responsive to changing requirements and stakeholder needs.

2.1.2 Project Effort and Time, Cost Estimation

Project effort and time, as well as cost estimation, are not always calculated in data analytics projects because the scope and requirements of such projects can often change rapidly and unpredictably as new data becomes available or insights are gained. This means that traditional project management approaches, which rely on detailed planning and fixed timelines, may not be suitable for data analytics projects.

Additionally, data analytics projects often require a high degree of experimentation and iteration in order to identify meaningful insights and patterns in the data. This can involve trying out different algorithms, visualizations, and statistical models, which can be difficult to predict in terms of time and resource requirements.

Instead, data analytics projects may be managed using agile or other flexible methodologies that prioritize collaboration, rapid prototyping, and continuous improvement. These approaches allow data analysts and stakeholders to work together closely and make adjustments as needed throughout the project, rather than relying on a fixed plan that may quickly become outdated.

Per the university's mandate, I have provided a tentative COCOMO Model estimation. It is important to note that this is only a preliminary estimation and the actual estimations are confidential and cannot be shared with anyone outside the organization.

Basic COCOMO Model:

Effort = 2.4 * (KLOC)^1.05 * (1.01^SF) , where

SF (scale factor) = 1.0 for non-stringent deadline
SF = 0.91 for team size of 9
SF = 1.46 for 6 experienced employees and 3 inexperienced interns
SF = 1.17 for a complexity of statistical models

Putting the values, we get:

Effort = 2.4 * (500)^1.05 * (1.01^1.17) * 0.91 * 1.46 = **454 person-months**

Duration = 2.5 * (Effort^0.38) = **14.5 months**

Intermediate COCOMO Model:

Effort = 3.0 * (KLOC)^1.12 * (0.9 + 0.01 * SF), where

SF (scale factor) = 1.0 for non-stringent deadline

SF = 0.91 for team size of 9

SF = 1.46 for 6 experienced employees and 3 inexperienced interns

SF = 1.17 for a complexity of statistical models

Putting the values, we get:

Effort = 3.0 * (500)^1.12 * (0.9 + 0.01 * 1.17) * 0.91 * 1.46 = **736 person-months**

Duration = 2.5 * (Effort^0.35) = **18.6 months**

Detailed COCOMO Model:

Effort = (2.94 * (KLOC^1.08)) * (0.95 + 0.01 * ΣSF), where

ΣSF (sum of scale factors) = 1.0 for non-stringent deadline

ΣSF = 0.91 for team size of 9

ΣSF = 1.46 for 6 experienced employees and 3 inexperienced interns

ΣSF = 1.17 for a complexity of statistical models

Putting the values, we get:

Effort = (2.94 * (500^1.08)) * (0.95 + 0.01 * (0.91 + 1.46 + 1.17)) = **1045 person-months**

Duration = 2.5 * (Effort/SE)^(0.38 + 0.2*(E-B)) = **23.2 months**

where SE (software efficiency) = 3.0

B (basic programming experience) = 1.12

E (level of programming experience) = 1.20

Given that the cost per person-month is **INR 25,000** we can calculate the cost for each model as follows:

For **Basic COCOMO Model**:

Cost = Effort * Cost Per Person-Month = 454 * 25000 = **INR 1,13,50,000**

For **Intermediate COCOMO Model**:

Cost = Effort * Cost Per Person-Month = 736 * 25000 = **INR 1,84,00,000**

For **Detailed COCOMO Model**:

Cost = Effort * Cost Per Person-Month = 1045 * 25000 = **INR 2,61,25,000**

Therefore, the cost for each model is INR 1,13,50,000 for Basic COCOMO, INR 1,84,00,000 for Intermediate COCOMO, and INR 2,61,25,000 for Detailed COCOMO.

2.1.3 Roles and Responsibilities

PHASE-1 (From December 19 2023 till February 20 2023):

Team Lead:

- Responsible for the overall management & Synchronization

Team-member-002 & Team-member-003:

- Data Gathering and pre-processing

Team-member-004:

- Dashboard Development

Team-member-005:

- Communication and Data Management and Security

Team Member-003-I:

- Assist all the team members (except for team-member-005) and handle CRM

PHASE-2 (From February 23 2023 till April 21 2023):

Team Lead:

- Responsible for Overall Management and Synchronization of the Project

Team Members:

Team Members 002 and 003:

- Responsible for Data Gathering and Pre-Processing

Team Members 004 and 005:

- Responsible for Dashboard Development

Team Members 006 and 007:

- Responsible for Communication, Data Management, and Security

Team Members 008, 009, and 015-I:

- Responsible for Software Development and Testing

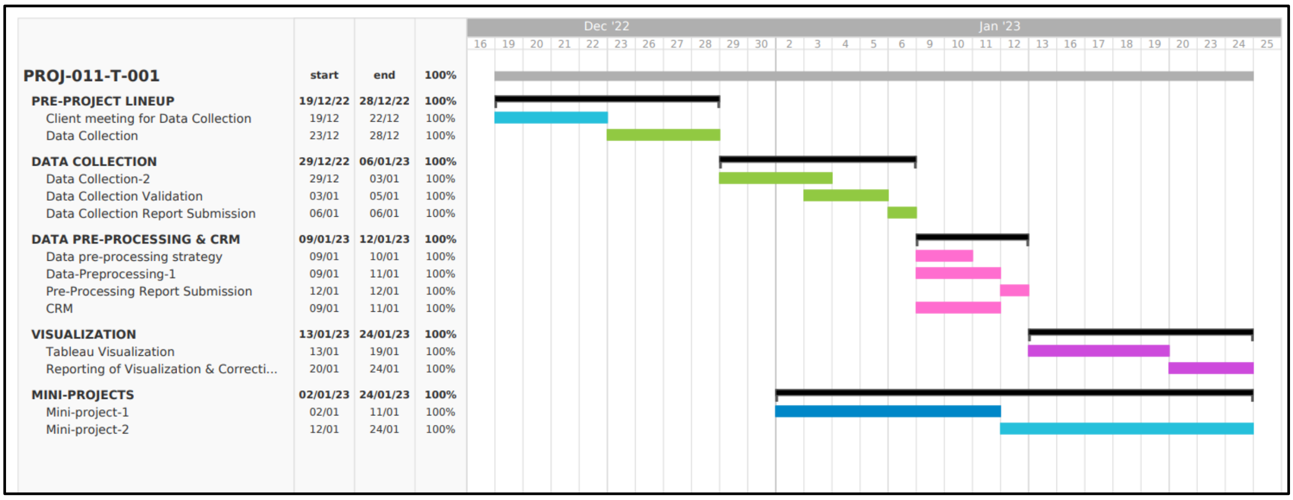
Team Member 009-I:

- Responsible for Document Auditing and Synchronization with other Departments

Team Member 003-I:

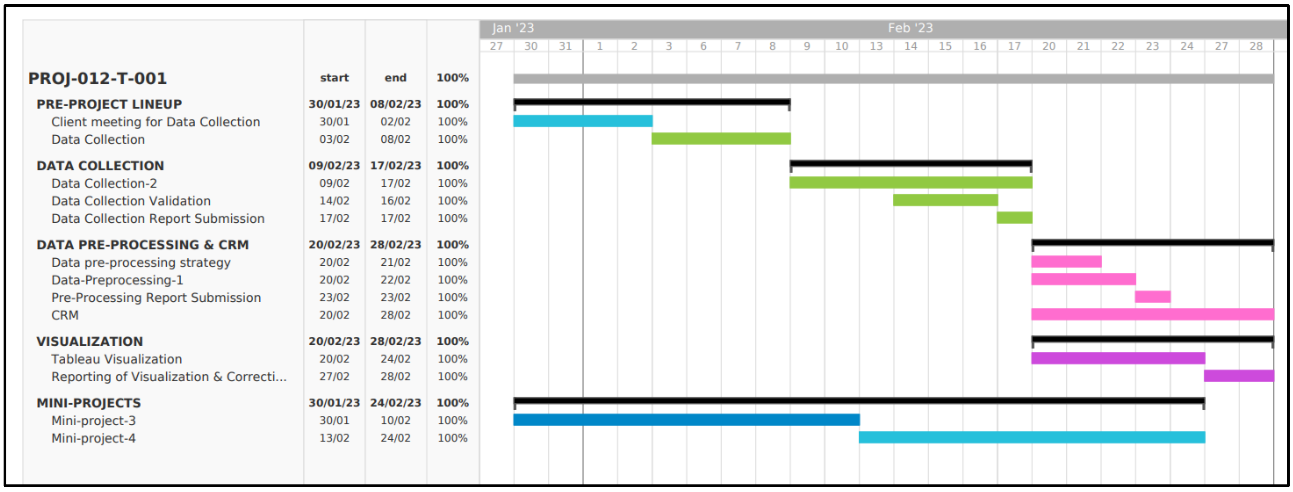
- Responsible for Assisting all Team Members (except Team Members 006, 007, 008, and 009) and Handling CRM

2.2 PROJECT SCHEDULING



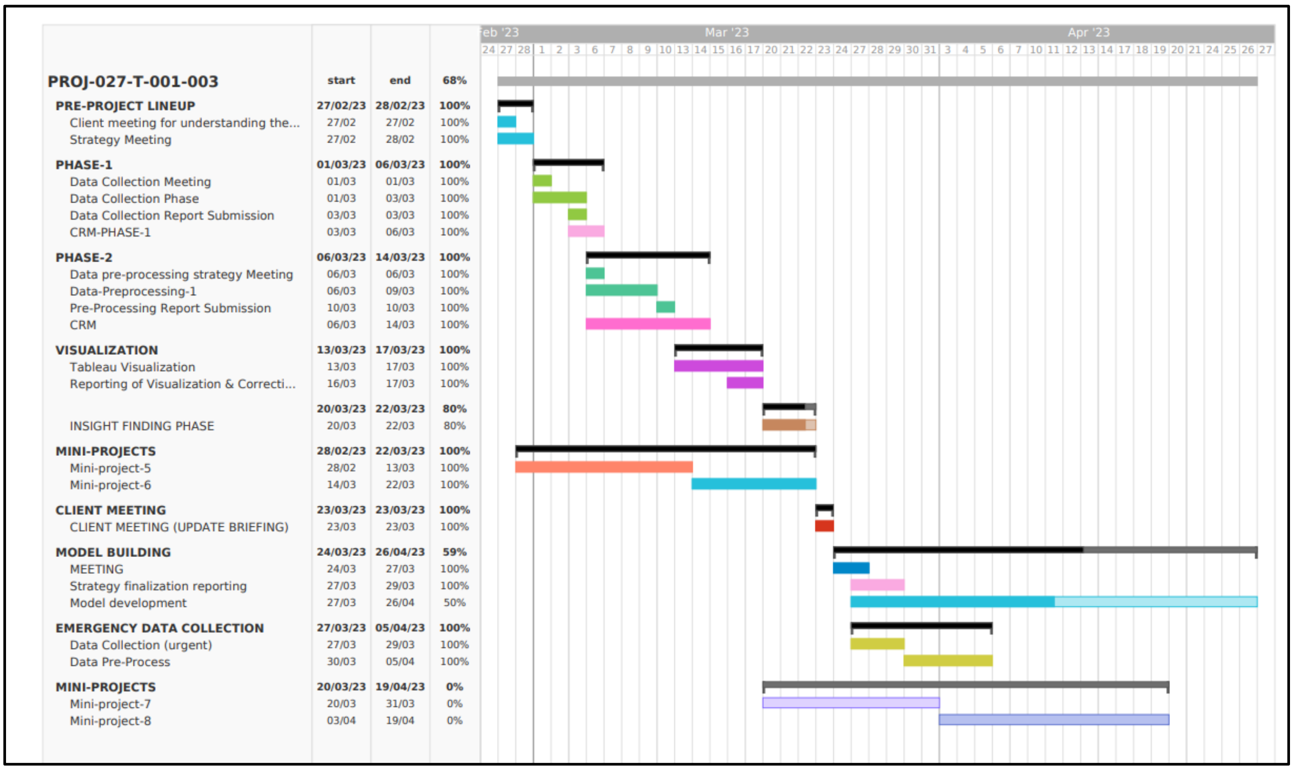
PROJ-011-T-001	start	end	100%
PRE-PROJECT LINEUP	19/12/22	28/12/22	100%
Client meeting for Data Collection	19/12	22/12	100%
Data Collection	23/12	28/12	100%
DATA COLLECTION	29/12/22	06/01/23	100%
Data Collection-2	29/12	03/01	100%
Data Collection Validation	03/01	05/01	100%
Data Collection Report Submission	06/01	06/01	100%
DATA PRE-PROCESSING & CRM	09/01/23	12/01/23	100%
Data pre-processing strategy	09/01	10/01	100%
Data-Preprocessing-1	09/01	11/01	100%
Pre-Processing Report Submission	12/01	12/01	100%
CRM	09/01	11/01	100%
VISUALIZATION	13/01/23	24/01/23	100%
Tableau Visualization	13/01	19/01	100%
Reporting of Visualization & Correcti...	20/01	24/01	100%
MINI-PROJECTS	02/01/23	24/01/23	100%
Mini-project-1	02/01	11/01	100%
Mini-project-2	12/01	24/01	100%

Fig. 2.2 Gantt Chart for PROJ-011-T-001



PROJ-012-T-001	start	end	100%
PRE-PROJECT LINEUP	30/01/23	08/02/23	100%
Client meeting for Data Collection	30/01	02/02	100%
Data Collection	03/02	08/02	100%
DATA COLLECTION	09/02/23	17/02/23	100%
Data Collection-2	09/02	17/02	100%
Data Collection Validation	14/02	16/02	100%
Data Collection Report Submission	17/02	17/02	100%
DATA PRE-PROCESSING & CRM	20/02/23	28/02/23	100%
Data pre-processing strategy	20/02	21/02	100%
Data-Preprocessing-1	20/02	22/02	100%
Pre-Processing Report Submission	23/02	23/02	100%
CRM	20/02	28/02	100%
VISUALIZATION	20/02/23	28/02/23	100%
Tableau Visualization	20/02	24/02	100%
Reporting of Visualization & Correcti...	27/02	28/02	100%
MINI-PROJECTS	30/01/23	24/02/23	100%
Mini-project-3	30/01	10/02	100%
Mini-project-4	13/02	24/02	100%

Fig. 2.3 Gantt Chart for PROJ-012-T-001



PROJ-027-T-001-003	start	end	68%
PRE-PROJECT LINEUP	**27/02/23**	**28/02/23**	**100%**
Client meeting for understanding the...	27/02	27/02	100%
Strategy Meeting	27/02	28/02	100%
PHASE-1	**01/03/23**	**06/03/23**	**100%**
Data Collection Meeting	01/03	01/03	100%
Data Collection Phase	01/03	03/03	100%
Data Collection Report Submission	03/03	03/03	100%
CRM-PHASE-1	03/03	06/03	100%
PHASE-2	**06/03/23**	**14/03/23**	**100%**
Data pre-processing strategy Meeting	06/03	06/03	100%
Data-Preprocessing-1	06/03	09/03	100%
Pre-Processing Report Submission	10/03	10/03	100%
CRM	06/03	14/03	100%
VISUALIZATION	**13/03/23**	**17/03/23**	**100%**
Tableau Visualization	13/03	17/03	100%
Reporting of Visualization & Correcti...	16/03	17/03	100%
	20/03/23	**22/03/23**	**80%**
INSIGHT FINDING PHASE	20/03	22/03	80%
MINI-PROJECTS	**28/02/23**	**22/03/23**	**100%**
Mini-project-5	28/02	13/03	100%
Mini-project-6	14/03	22/03	100%
CLIENT MEETING	**23/03/23**	**23/03/23**	**100%**
CLIENT MEETING (UPDATE BRIEFING)	23/03	23/03	100%
MODEL BUILDING	**24/03/23**	**26/04/23**	**59%**
MEETING	24/03	27/03	100%
Strategy finalization reporting	27/03	29/03	100%
Model development	27/03	26/04	50%
EMERGENCY DATA COLLECTION	**27/03/23**	**05/04/23**	**100%**
Data Collection (urgent)	27/03	29/03	100%
Data Pre-Process	30/03	05/04	100%
MINI-PROJECTS	**20/03/23**	**19/04/23**	**0%**
Mini-project-7	20/03	31/03	0%
Mini-project-8	03/04	19/04	0%

Fig. 2.4 Gantt Chart for PROJ-027-T-001-003

3 SYSTEM REQUIREMENTS STUDY

3.1 SRS IN DATA ANALYTICS AND SCIENCE PROJECTS

Software Requirements Specification (SRS) is an important document that outlines the requirements and specifications for a software project. In the case of data analytics and data science projects, the SRS should cover the following areas:

Introduction:
The introduction section should provide a brief overview of the project, including the purpose, scope, and objectives.

Functional requirements:
This section should describe the functional requirements of the system. It should include details on the following:

- **Data collection:**
    - How the data will be collected, stored, and processed.
- **Data pre-processing:**
    - How the data will be cleaned, transformed, and prepared for analysis.
- **Data analytics:**
    - How the data will be analyzed and what techniques and tools will be used.
- **Data visualization:**
    - How the data will be presented to stakeholders in a clear and concise way.
- **Data consulting:**
    - How the project team will provide insights and recommendations to stakeholders based on the data analysis.
- **Minor customer relationship management:**
    - How the project team will manage interactions with stakeholders, including communication, feedback, and issue resolution.

Non-functional requirements:

This section should describe the non-functional requirements of the system. It should include details on the following:

- **Performance:**
  - The system should be able to handle large volumes of data and provide fast and responsive analytics.
- **Scalability:**
  - The system should be able to scale up or down depending on the size and complexity of the data.
- **Security:**
  - The system should be secure and protect the confidentiality, integrity, and availability of the data.
- **Usability:**
  - The system should be easy to use and navigate, with clear instructions and user-friendly interfaces.
- **Compatibility:**
  - The system should be compatible with various platforms, tools, and technologies.

Overall, the SRS for data analytics and data science projects should provide a clear and comprehensive description of the project requirements, design, and implementation. It should serve as a guide for the project team and stakeholders throughout the project lifecycle.

3.2 SOFTWARE AND HARDWARE REQUIRMENTS

As an individual who has worked on multiple projects in the field of data analytics and data science, I understand the importance of having the right hardware and software for these projects. As I have worked on multiple projects in chunks, as a result, I am mentioning the generic requirements which will be common for all the projects. The following are some general requirements for such projects:

Hardware:

- A reliable and fast computer with a multi-core processor and a minimum of 8GB RAM to handle the data processing and analysis workload effectively.
- Adequate storage space to store the data and software used for analysis.
- A high-speed internet connection to access online resources and cloud-based services.

Software:

- Data analytics and visualization tools such as R, Python, and Tableau to process, analyze, and visualize the data.
- Database management systems like MySQL, MongoDB, and PostgreSQL to manage large volumes of data efficiently.
- Statistical analysis tools like SAS and SPSS to conduct advanced statistical analysis and modeling.
- Machine learning and deep learning frameworks like TensorFlow, Keras, and PyTorch to develop and train machine learning models.
- Cloud-based services like AWS, Google Cloud, and Azure for scalable computing power and storage.

It is important to note that the specific hardware and software requirements for data analytics and data science projects may vary depending on the project's objectives, scope, and available resources. Therefore, a thorough analysis of the project requirements should be conducted to choose the appropriate software and hardware for the project.

3.3 ASSUMPTIONS AND DEPENDENCIES

Assumptions:

- The data collected is accurate and relevant to the project goals.
- The data preprocessing techniques used are appropriate for the data being analyzed.
- The chosen data analytics and visualization tools are suitable for the specific project requirements.
- The customer relationship management activities will not significantly impact the data analysis timeline and deliverables.
- The intern management activities will not significantly impact the intern's ability to complete project tasks.

Dependencies:

- Availability of the necessary hardware and software resources for the project.
- Availability and access to the relevant data sources required for the project.
- Access to necessary tools and technology required for data analysis, visualization, and consulting.
- Effective communication and collaboration with clients and stakeholders for gathering project requirements, discussing progress, and sharing results.
- Timely feedback from clients and stakeholders to ensure the project is meeting their needs and requirements.

4 SYSTEM ANALYSIS

4.1 STUDY OF CURRENT SYSTEM

It is important to note that a comprehensive study of the current system would require a thorough understanding of the project objectives, stakeholders, data sources, software tools, and processes. Without access to these critical components, it would be difficult to conduct a meaningful analysis of the system. As a data analyst, it is crucial to have a clear understanding of the underlying data and systems in order to make informed decisions and provide valuable insights to stakeholders.

However, in the absence of complete information, it is necessary to exercise caution and avoid making assumptions or drawing conclusions that may be inaccurate or misleading.

4.2 FEASIBILITY STUDY

Feasibility study is an important aspect of data analytics projects. It involves evaluating the viability and feasibility of the project by assessing various factors such as technical feasibility, operational feasibility, economic feasibility, legal feasibility, and scheduling feasibility.

Technical feasibility involves assessing the technology requirements, resources, and skills needed to implement the project successfully. Operational feasibility involves analyzing the impact of the project on the organization's operations and processes. Economic feasibility involves evaluating the costs and benefits of the project and determining its financial feasibility. Legal feasibility involves assessing the project's compliance with legal and regulatory requirements. Scheduling feasibility involves evaluating the project's timeline and ensuring it is achievable within the given time constraints.

Conducting a feasibility study helps in identifying potential risks, challenges, and limitations of the project and determining whether the project is worth pursuing or not. It helps in making informed decisions about the project and ensures that the project is aligned with the organization's goals and objectives.

4.2.1 Feasibility Study on Web Scrapping

Web scraping is the process of extracting data from websites. Before starting a web scraping project, it is important to conduct a feasibility study to determine whether the project is viable and feasible.

Technical feasibility:

- The technical feasibility of web scraping depends on the structure of the website and the tools and technologies available for web scraping. It is important to evaluate whether the website can be scraped and whether the tools and technologies available are sufficient to perform the required scraping.

Legal feasibility:

- Web scraping can be subject to legal restrictions, such as copyright and privacy laws. It is important to evaluate the legal feasibility of web scraping and determine whether the scraping is legal and ethical.

Operational feasibility:

- Web scraping can require significant resources, including time, personnel, and technology. It is important to evaluate whether the required resources are available and whether the project is feasible given the available resources.

Economic feasibility:

- Web scraping can have significant costs, including hardware and software costs, personnel costs, and legal costs. It is important to evaluate whether the costs are justified by the benefits of the project and whether the project is economically feasible.

4.2.2 Feasibility Study of Analytics Project

Technical Feasibility:

- The project can be technically feasible if the necessary hardware and software resources are available. The project will require a computer with a multi-core processor, sufficient amount of RAM, and adequate storage space. Data analytics tools such as R, Python, and Tableau will also be required. These resources should be available for the project.

Financial Feasibility:

- The project can be financially feasible if the costs associated with the project can be justified by the benefits obtained from the analysis. This will require a cost-benefit analysis to determine if the project is financially feasible.

Operational Feasibility:

- The project can be operationally feasible if the necessary resources, including personnel, are available. This will require a review of the current staff and their capabilities to determine if they can perform the necessary tasks.

Legal and Ethical Feasibility:

- The project must adhere to legal and ethical standards. This will require a review of any potential legal or ethical issues that may arise during the project.

4.2.3 Feasibility Study on Tableau Projects (Visualization)

Technical Feasibility:

- Tableau is a powerful and versatile data visualization tool with a user-friendly interface, making it easy to create interactive dashboards and reports. The software provides a wide range of features and functionalities to manipulate, analyze, and visualize data. It supports a variety of data sources, including Excel spreadsheets, databases, cloud-based services, and web-based APIs. Moreover, Tableau offers extensive documentation, tutorials, and community forums, making it easy to learn and use. Therefore, the project is technically feasible, and Tableau is a suitable tool for data visualization.

Economic Feasibility:

- Tableau offers a range of pricing plans, including free and paid versions, to suit various needs and budgets. The free version, Tableau Public, allows users to create and share visualizations on the web. However, it has limited functionality and requires public data sources. The paid version, Tableau Desktop, provides more advanced features and supports various data sources, including cloud-based and on-premises databases. The cost of Tableau Desktop starts at $70 per user per month, making it a relatively expensive option for small businesses or individuals. Therefore, the project's economic feasibility depends on the available budget and the project's objectives.

Operational Feasibility:

- The success of the project depends on the availability and quality of the data, the skills and expertise of the project team, and the project's objectives and scope. To ensure the operational feasibility of the project, it is essential to establish clear requirements, goals, and timelines. Moreover, it is necessary to have a skilled team with expertise in data analysis, visualization, and design. The team should also have a thorough understanding of the project's data sources, limitations, and potential biases. Furthermore, the project should have adequate resources, including hardware, software, and data storage, to support the project's objectives and scope.

4.2.4 Feasibility Study of Pan Card Tampering

The Pan Card Tampering Project is a software application that aims to detect tampering in Pan Cards using image processing techniques. The application uses Python as the programming language and OpenCV library for image processing. Pytesseract is used for Optical Character Recognition (OCR) to extract text from the images.

The first step in the feasibility study is to analyze the technical feasibility of the project. The Pan Card Tampering Project uses advanced image processing techniques that require specialized knowledge and expertise. The project also requires the use of advanced programming languages and libraries such as Python, OpenCV, and Pytesseract. Therefore, the technical feasibility of the project is high, and the required resources are readily available.

The second step is to analyze the economic feasibility of the project. The Pan Card Tampering Project is a cost-effective solution for detecting tampering in Pan Cards. The software application can be developed using open-source libraries and programming languages, which reduces the cost of development. The project also has the potential to generate revenue by providing the service to individuals and organizations that need to verify the authenticity of Pan Cards.

The third step is to analyze the operational feasibility of the project. The Pan Card Tampering Project requires minimal human intervention, and the application can be run on any device that supports Python and OpenCV. The application can also be integrated with other systems that require verification of Pan Cards. Therefore, the operational feasibility of the project is high, and the application can be easily integrated with existing systems.

The fourth step is to analyze the legal feasibility of the project. The Pan Card Tampering Project complies with all the legal requirements for verifying the authenticity of Pan Cards. The application does not violate any data privacy laws, and it can be used in accordance with the regulations set by the Government of India.

4.2.5 Feasibility Study on Text Extraction

Technical Feasibility:

One of the critical factors to consider when evaluating the technical feasibility of a text extraction project is the availability of appropriate tools and technologies to achieve the desired results. For example, the project would require the use of OCR (Optical Character Recognition) technology, image processing, machine learning algorithms, and natural language processing (NLP) tools. Therefore, it is essential to assess whether these tools and technologies are readily available and accessible for the project's successful execution. Additionally, the project's technical feasibility would also involve evaluating the data sources and their compatibility with the selected technologies and tools.

Economic Feasibility:

Another significant factor to consider is the economic feasibility of the project. It is crucial to evaluate the cost-benefit analysis of the project, including the estimated costs associated with acquiring necessary tools and technologies, hiring skilled personnel, and maintaining and upgrading the system. The feasibility study should consider the potential return on investment (ROI) of the project and assess whether it justifies the investment.

Operational Feasibility:

The operational feasibility of the project involves assessing the ease of implementing and maintaining the system. It is essential to evaluate whether the project aligns with the company's goals and whether the organization has the necessary resources to implement and maintain the system. Additionally, the study would assess the impact of the project on the organization's operations, processes, and workflow.

Legal Feasibility:

The legal feasibility of the project involves assessing the compliance of the project with legal and regulatory requirements. For example, the project would need to comply with data privacy laws and regulations that govern data handling and processing. Therefore, it is essential to evaluate whether the project complies with relevant laws and regulations to avoid legal liabilities and ensure data protection.

4.3 OO CONCEPTS IN DATA ANALYTICS & SCIENCE PROJECTS

Object-Oriented Programming (OOP) is a programming paradigm that is widely used in software development. However, in data analytics and science projects, OOP is not as commonly used. There are several reasons why OOP is not always followed in these types of projects:

Data analytics and science projects are typically more focused on declarative programming, where the emphasis is on specifying what you want to happen, rather than how it should happen. This approach is more focused on the data and the desired outcomes, rather than the implementation details.

In traditional software development, there is a focus on creating reusable code modules that can be used in multiple applications. In data analytics and science projects, the focus is more on creating specific functionality for a single project or analysis, rather than on creating reusable code.

Data analytics and science projects often rely on specific tools and libraries, such as R, Python, and SQL, which have their own built-in functionality and do not necessarily require OOP implementation.

OOP can add complexity to code and make it more difficult to understand, which can be counterproductive in data analytics and science projects where the focus is on data understanding and insights.

Overall, while OOP has its advantages, it is not always necessary or practical in data analytics and science projects. The focus in these projects is on the data and the insights that can be gained from it, rather than on creating reusable code or implementing complex object-oriented programming paradigms.

4.4 E-R DIAGRAMS

Data modelling diagrams such as Entity-Relationship (E-R) diagrams are not as useful in data analytics and data science projects as they are in traditional software development projects. This is because in data analytics and data science projects, the primary focus is on analyzing and visualizing data, rather than designing a database schema or defining relationships between entities.

In data analytics and data science projects, the focus is on understanding and processing the data itself, rather than on how it is stored or organized. Therefore, data modeling diagrams are not as useful because the structure of the data is often flexible and can change frequently. The focus is on the insights gained from the data, rather than the structure of the data itself.

Additionally, data modelling diagrams may not accurately represent the complexity and variability of real-world data. The data in data analytics and data science projects often comes from various sources and may not fit neatly into a predefined schema. Therefore, using a rigid data modelling approach may not be effective in capturing the full complexity of the data.

Instead, data analytics and data science projects may use other methods for organizing and understanding data, such as data exploration and visualization tools. These tools can help identify patterns and relationships in the data without the need for a predefined data model.

4.5 UML DIAGRAMS

Unified Modelling Language (UML) is a standardized modelling language used in software engineering to visualize, design, and document software systems. It includes various diagrams such as Use Case, Class, Activity, Sequence, and others that help to represent the different aspects of a software system. However, these diagrams are not commonly used in data analytics and data science projects for the following reasons:

Data analytics and data science projects are not primarily software engineering projects: While software engineering projects involve designing and building software systems, data analytics and data science projects are focused on extracting insights from data. The main goal of data analytics and data science projects is to solve business problems using data, not to build software systems. Therefore, UML diagrams are not as relevant to these projects as they are to software engineering projects.

Data analytics and data science projects are more exploratory in nature. Data analytics and data science projects involve a lot of exploration and experimentation with the data. This means that the requirements and design of the project are often unclear and may change frequently as the project progresses. UML diagrams are not well-suited for this kind of exploratory work as they are more rigid and require a lot of upfront planning.

Data analytics and data science projects involve a wide range of data sources and tools. Data analytics and data science projects involve working with a wide range of data sources and tools, such as databases, spreadsheets, statistical software, and visualization tools. UML diagrams are primarily used for modelling software systems and are not as relevant for representing data sources and tools.

In summary, while UML diagrams are useful in software engineering projects, they are not commonly used in data analytics and data science projects. These projects have different goals and requirements, and their focus is on extracting insights from data, rather than building software systems.

5 MINI PROJECTS IN DETAIL

5.1 DATA VISUALIZATION WITH TABLEAU

5.1.1 Mini-Project-1

Business Problem:

The client wants to keep an eye on their overall sales and profit across the Europe on yearly basis, along with minute details of the customer.

Dashboard:

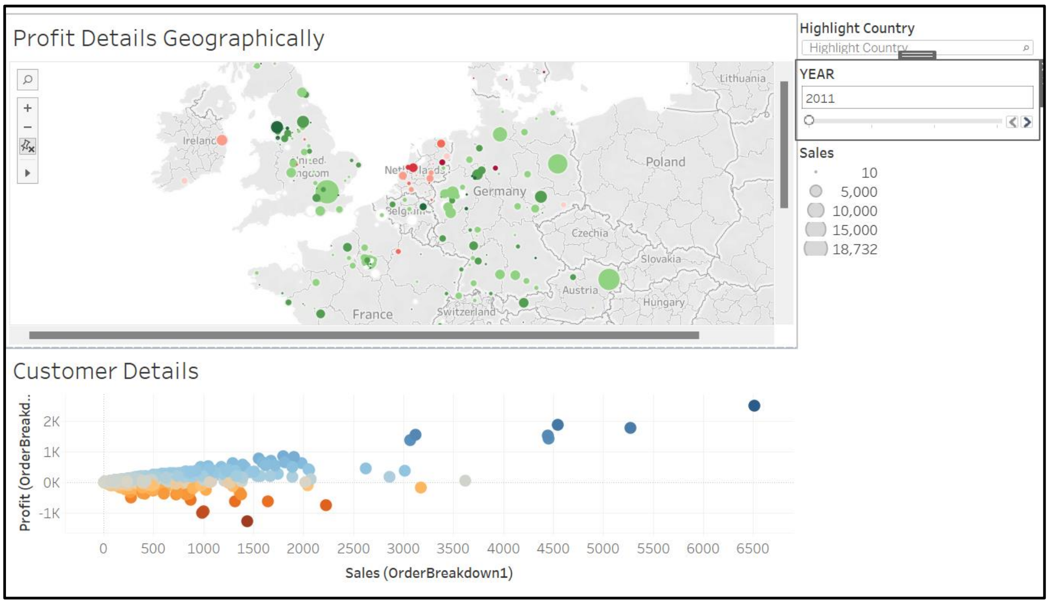


Fig. 5.1 Mini-Project-1 Dashboard

Change in granularity: 1. Country wise and 2. Yearly basis.

Reason behind such design:

- The customer wanted simple color combination.
- The geographical map showcases yearly profits of the areas.
- Towards red means loss, and profit for going towards green.
- The lower chart depicts the sales and profit for every customer.
- The charts are interactive and filter is applicable to both.

5.1.2 Mini-Project-2

Business Problem:

The client wants to give yearly bonus to top-performing employee from each region. Help them in finding out the one who deserves to get the bonus.

Dashboard:

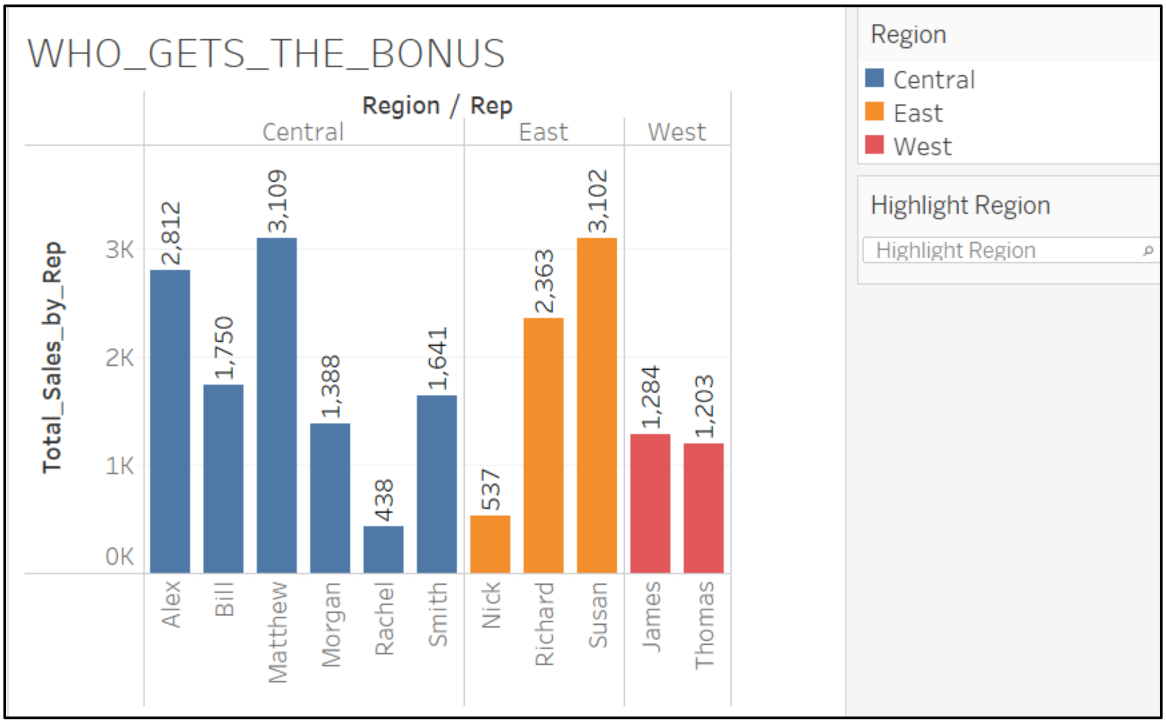


Fig. 5.2 Mini-Project-2 Dashboard

My Insights:

- Firstly, from the data provided, there were no direct indications for the sales and profits made by every employee. Thus, I created a calculated field which calculates the overall sales and profit made by all the employees.
- There are 3 regions: Central, West and East.
- Thus, 3 employees in total will get the bonus.
- So, it is evident that Matthew from Central, Susan from East and James from West will get the bonus.

5.1.3 Mini-Project-3

Business Problem:

The client wants city-wise summarized details of how their stores are performing?

Dashboard:

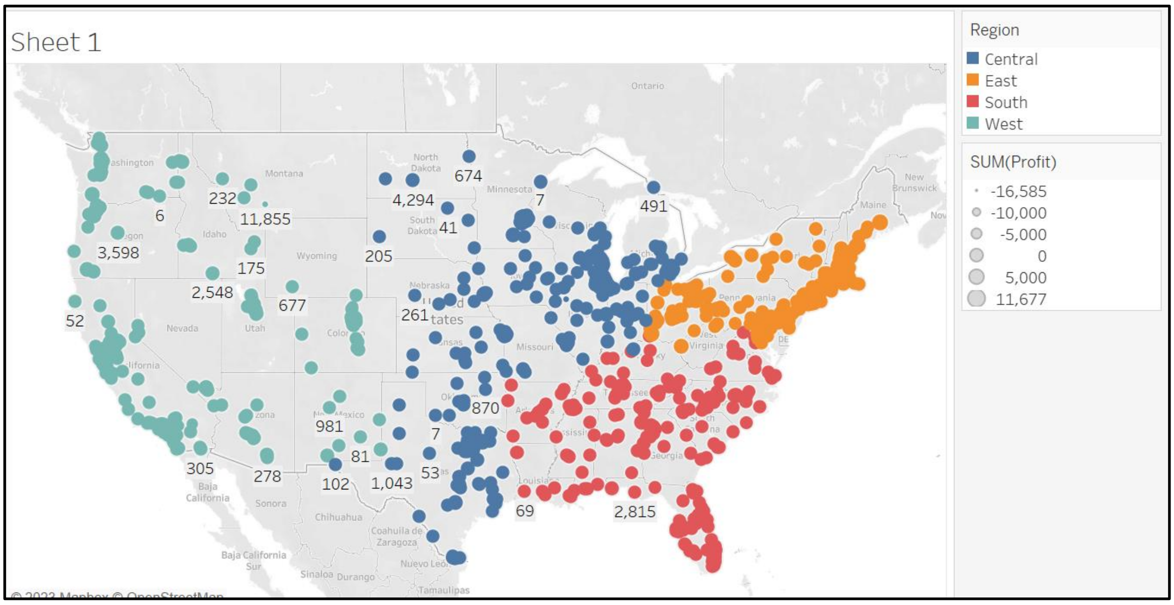


Fig. 5.3 Mini-Project-3 Dashboard

5.1.4 Mini Project-4

Business Problem:

There is a bank in United Kingdom. Now, the bank wants to retain their customers and to add new one. The board of directors have agreed to develop multiple schemes for different sections of people and regions. They want you to design a dashboard through which they can play with it to determine the new policies. Also, list out the important parameters that they should consider for development of new policies.

Dashboard:

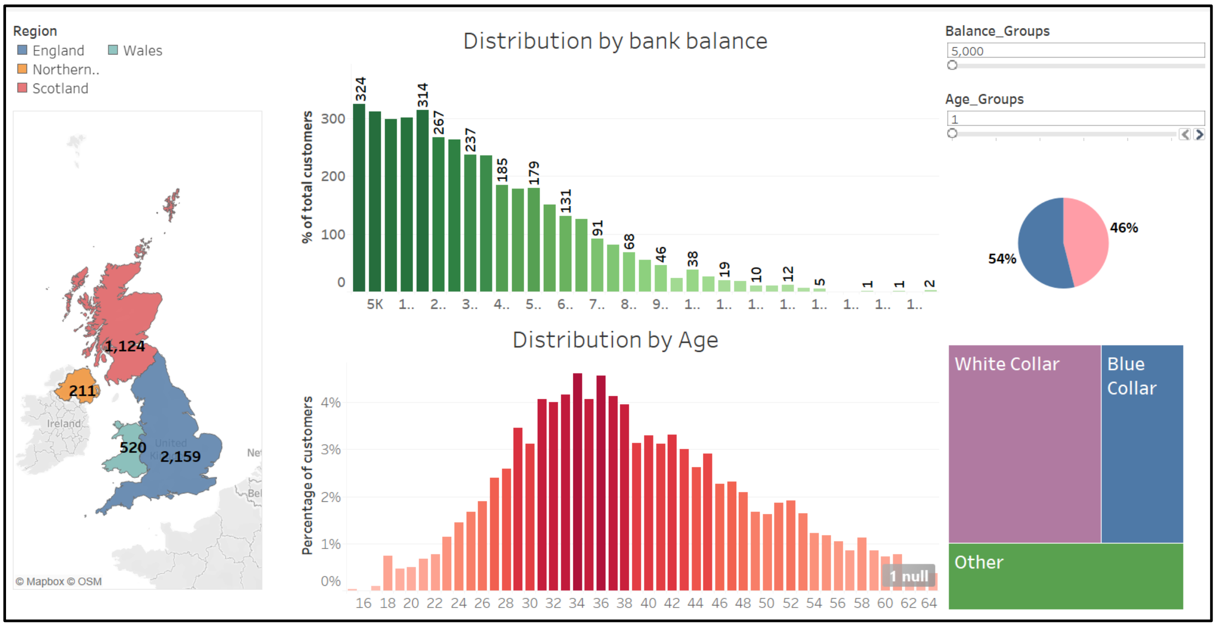


Fig. 5.4 Mini-Project-4 Dashboard

Filter: Applicable to all

Individual Granularity: Limited to Age Distribution and Bank Balance.

My Insights:

There are 4 regions in the UK where bank have presence. Out of the 4, the England have largest customer base. Thus, it will be a good idea to implement your new policies and incentives from England.

Secondly, most of the customers are having white collar jobs, and it is obvious as the economy of the UK is more of technology oriented where white collar jobs form a major workforce. Thus, new incentives should be focused more on benefitting white collar job holders.

Here, though male constitutes a whopping 56% customer base, the bank should focus on bringing more and more women customers for a simple reason that, the world is changing and so, more and more females are contributing in the economy. Hence, lucrative offers for women can be a master stroke. Remember, the bank need to grab the first mover advantage.

Considering the bank balance of the customers, majority of them falls in middle class category. Thus, the committee needs to have a dual strategy here. The one for adding more salaried middle-class customer and second to lure upper class people or business tycoons.

Considering the age distribution, it is quite opposite to the national trend. While UK has more aging population, however, the customer base of the bank is relatively younger. Thus, considering the pro-immigrant stance of the government, the bank should consider to target younger immigrants and citizens for sustainable growth.

5.2 INSIGHTS PROJECT

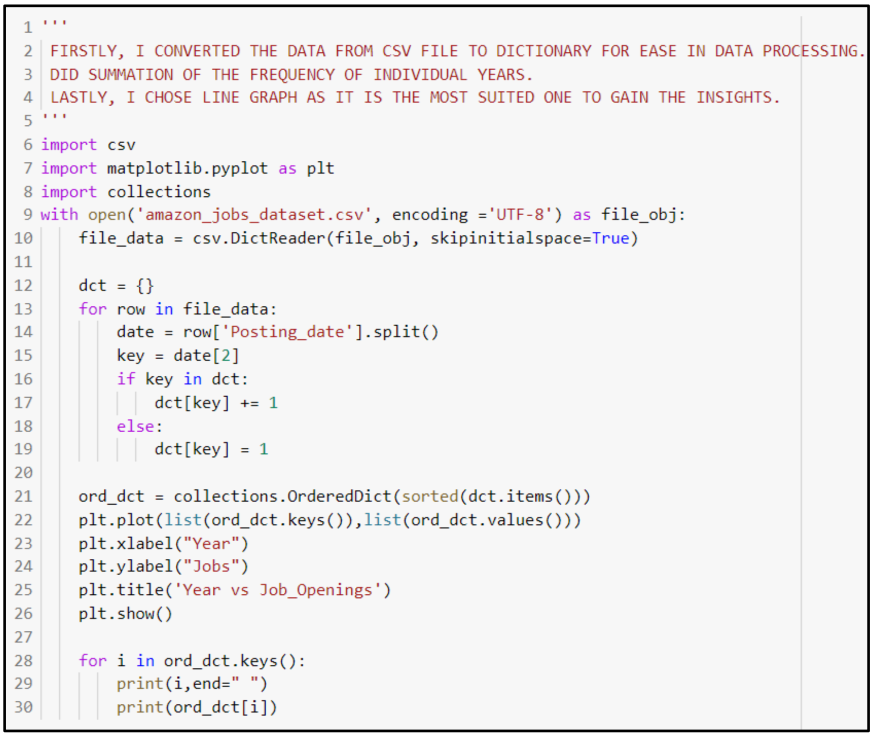
5.2.1 Amazon Job Analysis Mini-Project

Data-Set Description:

It is a dataset including information on amazon job opening around the world from June 2011 to March 2018. This dataset is collected using Selenium and BeautifulSoup by scraping all of the jobs for Amazon job site.

Question-1:

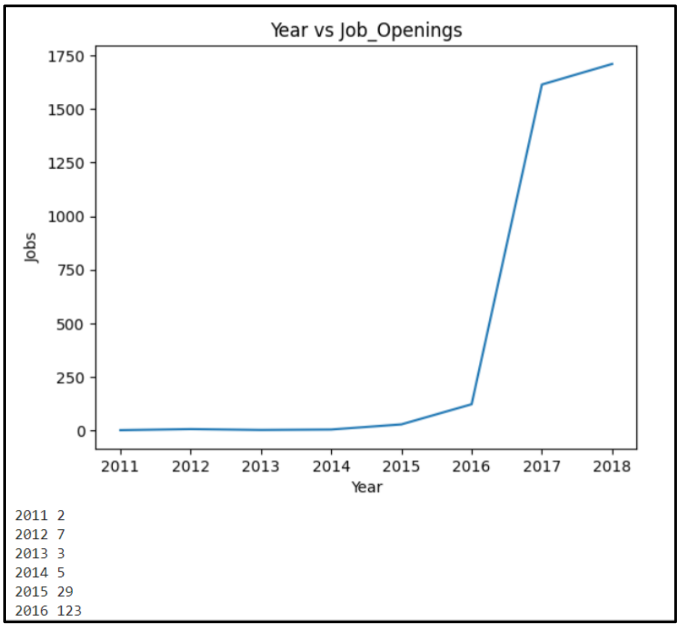
Plot the line graph between no. of Job postings with respect to year. Print the year and the number of job posting as integer value.



```
'''
 FIRSTLY, I CONVERTED THE DATA FROM CSV FILE TO DICTIONARY FOR EASE IN DATA PROCESSING.
 DID SUMMATION OF THE FREQUENCY OF INDIVIDUAL YEARS.
 LASTLY, I CHOSE LINE GRAPH AS IT IS THE MOST SUITED ONE TO GAIN THE INSIGHTS.
'''
import csv
import matplotlib.pyplot as plt
import collections
with open('amazon_jobs_dataset.csv', encoding ='UTF-8') as file_obj:
    file_data = csv.DictReader(file_obj, skipinitialspace=True)

    dct = {}
    for row in file_data:
        date = row['Posting_date'].split()
        key = date[2]
        if key in dct:
            dct[key] += 1
        else:
            dct[key] = 1

    ord_dct = collections.OrderedDict(sorted(dct.items()))
    plt.plot(list(ord_dct.keys()),list(ord_dct.values()))
    plt.xlabel("Year")
    plt.ylabel("Jobs")
    plt.title('Year vs Job_Openings')
    plt.show()

    for i in ord_dct.keys():
        print(i,end=" ")
        print(ord_dct[i])
```

Fig. 5.5 Insight-1 Code

Output-1:



```
2011 2
2012 7
2013 3
2014 5
2015 29
2016 123
```

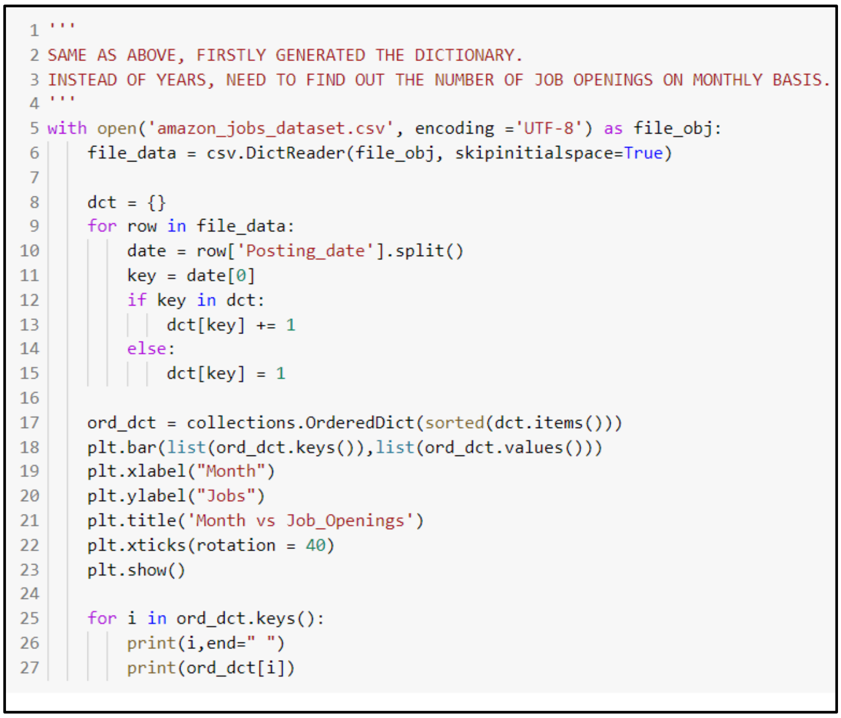
Fig. 5.6 Output of above Question

Insights:

Based on the analysis of the data, it is evident that there has been a significant increase in the number of job openings starting from the year 2011. The most substantial growth occurred in the years 2015 and 2016. Prior to 2015, the number of job openings remained relatively constant from 2011 to 2014. However, this trend shifted when Amazon began establishing offices in India in 2015, resulting in a significant increase in the number of job openings.

Question-2:

Plot the Bar graph between Month vs Job Openings. Print the month name and the number of job posting as integer value. Order of months doesn't matter.



```python
'''
SAME AS ABOVE, FIRSTLY GENERATED THE DICTIONARY.
INSTEAD OF YEARS, NEED TO FIND OUT THE NUMBER OF JOB OPENINGS ON MONTHLY BASIS.
'''
with open('amazon_jobs_dataset.csv', encoding ='UTF-8') as file_obj:
    file_data = csv.DictReader(file_obj, skipinitialspace=True)

    dct = {}
    for row in file_data:
        date = row['Posting_date'].split()
        key = date[0]
        if key in dct:
            dct[key] += 1
        else:
            dct[key] = 1

    ord_dct = collections.OrderedDict(sorted(dct.items()))
    plt.bar(list(ord_dct.keys()),list(ord_dct.values()))
    plt.xlabel("Month")
    plt.ylabel("Jobs")
    plt.title('Month vs Job_Openings')
    plt.xticks(rotation = 40)
    plt.show()

    for i in ord_dct.keys():
        print(i,end=" ")
        print(ord_dct[i])
```

Fig. 5.7 Code to find the answer of Question-2

Output-2:

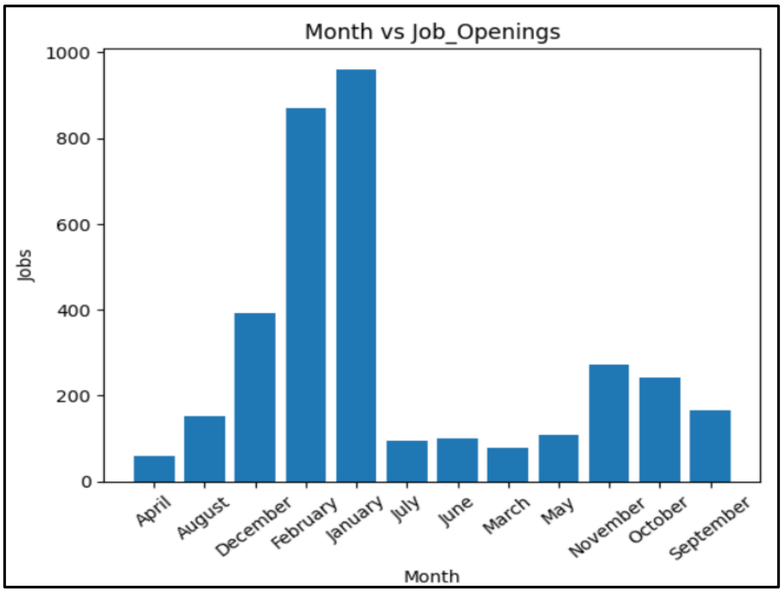
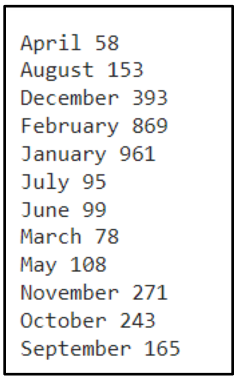


Fig. 5.8 Output of the Question-2



```
April 58
August 153
December 393
February 869
January 961
July 95
June 99
March 78
May 108
November 271
October 243
September 165
```

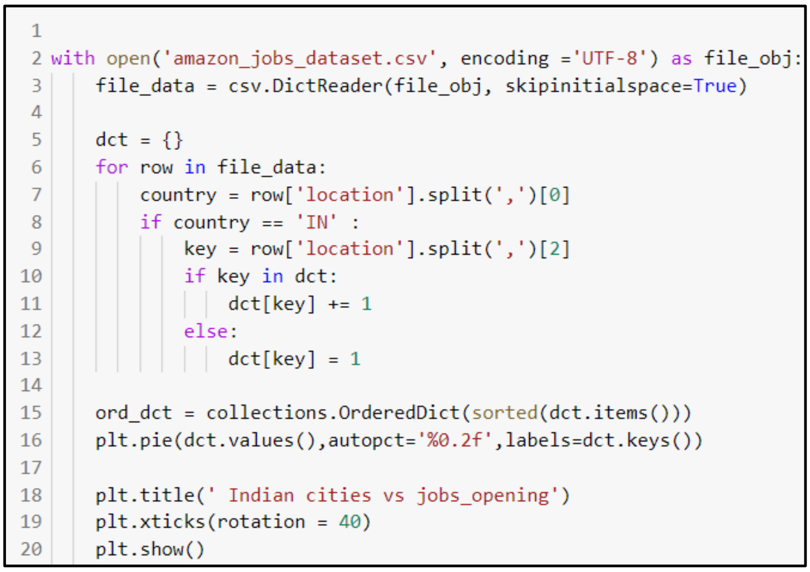
Fig. 5.9 Output of the Question-2

INSIGHTS:

Based on our analysis, it is apparent that the number of job openings consistently increased from October to February, regardless of the year. This trend can be attributed to several factors. In the United States, the fiscal year begins in October, and in India, the fiscal year ends in March. Therefore, this period marks a time of significant shuffling and attrition in the job market, resulting in an increase in job openings. Additionally, many students in India begin their job search around October, which also contributes to the rise in job openings during this period. Therefore, it is crucial to consider these factors when analyzing the job market trends during this time of the year.

Question-3:

Plot the Pie chart between Indian cities vs No. of job openings.



```python

with open('amazon_jobs_dataset.csv', encoding ='UTF-8') as file_obj:
    file_data = csv.DictReader(file_obj, skipinitialspace=True)

    dct = {}
    for row in file_data:
        country = row['location'].split(',')[0]
        if country == 'IN' :
            key = row['location'].split(',')[2]
            if key in dct:
                dct[key] += 1
            else:
                dct[key] = 1

    ord_dct = collections.OrderedDict(sorted(dct.items()))
    plt.pie(dct.values(),autopct='%0.2f',labels=dct.keys())

    plt.title(' Indian cities vs jobs_opening')
    plt.xticks(rotation = 40)
    plt.show()
```

Fig. 5.10 Code for Question-3

Output:

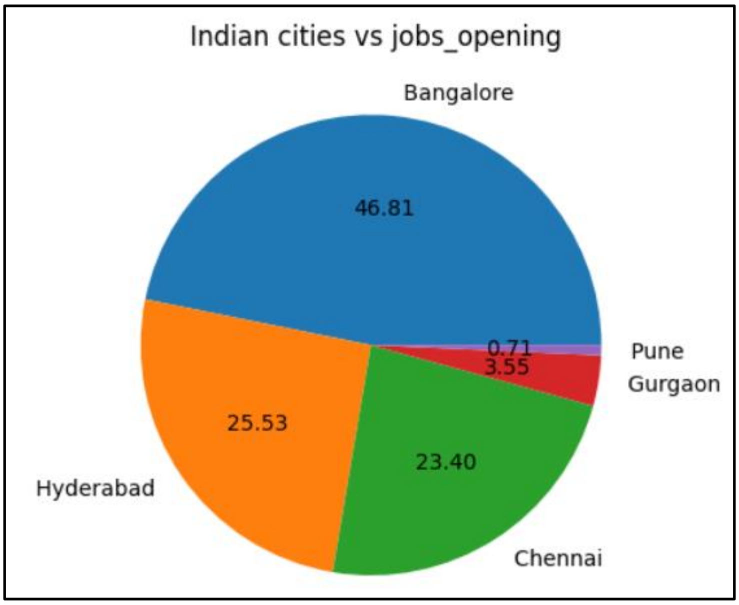


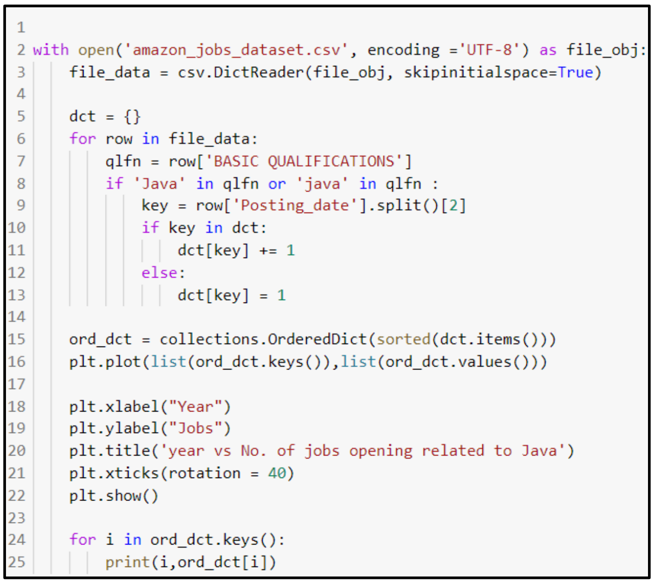
Fig. 5.11 Output of Question-3

Insights:

Based on our analysis, it is apparent that the majority of job openings, specifically more than 45 percent, are from the Bangalore office. This comes as no surprise as Bangalore is not only the headquarters of the company in India but also known as the Silicon Valley of India, attracting a larger talent pool. The second highest number of job openings is from Hyderabad, which is also an emerging Silicon Valley in India. The remaining offices have a minor contribution to the overall job openings.

Question-4:

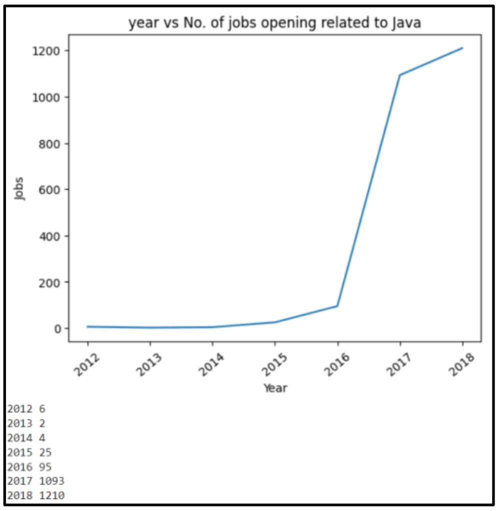
Plot the scatter graph between year vs No. of jobs opening related to Java. Print the year and number of Jobs opening in Java Profile.



```python

with open('amazon_jobs_dataset.csv', encoding ='UTF-8') as file_obj:
    file_data = csv.DictReader(file_obj, skipinitialspace=True)

    dct = {}
    for row in file_data:
        qlfn = row['BASIC QUALIFICATIONS']
        if 'Java' in qlfn or 'java' in qlfn :
            key = row['Posting_date'].split()[2]
            if key in dct:
                dct[key] += 1
            else:
                dct[key] = 1

    ord_dct = collections.OrderedDict(sorted(dct.items()))
    plt.plot(list(ord_dct.keys()),list(ord_dct.values()))

    plt.xlabel("Year")
    plt.ylabel("Jobs")
    plt.title('year vs No. of jobs opening related to Java')
    plt.xticks(rotation = 40)
    plt.show()

    for i in ord_dct.keys():
        print(i,ord_dct[i])
```

Fig. 5.12 Code for Question-4

Output-4:



```
2012 6
2013 2
2014 4
2015 25
2016 95
2017 1093
2018 1210
```

Fig.5.13 Output for Question-5

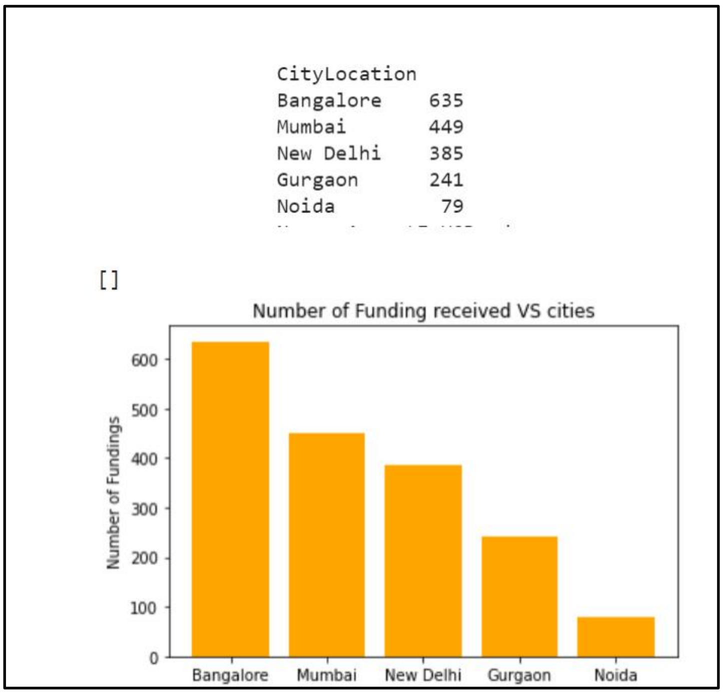
Insights:

Upon careful analysis, it is evident that as the company expands and grows, the number of job openings is likely to increase. Furthermore, it is observed that jobs requiring Java skills are in high demand in India, which is not surprising given the importance of Java in the tech industry.

5.2.2 Start-Up Funding Problem

Question-1:

Your Friend has developed the Product and he wants to establish the product startup and he is searching for a perfect location where getting the investment has a high chance. But due to its financial restriction, he can choose only between three locations - Bangalore, Mumbai, and NCR. As a friend, you want to help your friend deciding the location. NCR include Gurgaon, Noida and New Delhi. Find the location where the greatest number of funding is done. That means, find the location where startups have received funding maximum number of times. Plot the bar graph between location and number of funding. For few startups multiple locations are given, one Indian and one Foreign. Consider the startup if any one of the city lies in given locations.



```
CityLocation
Bangalore    635
Mumbai       449
New Delhi    385
Gurgaon      241
Noida         79
```

```
[]
```

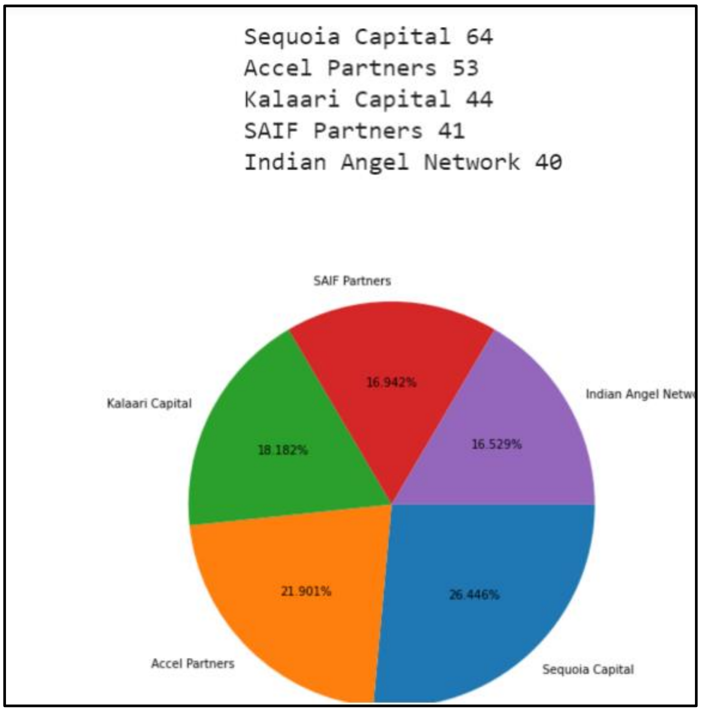
Fig. 5.14 Output of Question-1

Insights:

Based on the analysis of funding data, it is evident that Bangalore has received the highest amount of funding compared to other cities. However, it is important to note that if we combine the total fundings received in Delhi, Gurgaon, and Noida (collectively known as NCR), the sum amounts to 705. Hence, when we compare the individual cities, Bangalore has received the highest funding, but when we compare Bangalore, Mumbai, and NCR as regions, NCR emerges as the clear winner. Therefore, I recommend my friend to consider NCR as it will provide more opportunities to explore as it comprises three cities. This can lead to better networking and collaboration opportunities, ultimately leading to more significant growth and development.

Question-2:

Even after trying for so many times, your friend's startup could not find the investment. So you decided to take this matter in your hand and try to find the list of investors who probably can invest in your friend's startup. Your list will increase the chance of your friend startup getting some initial investment by contacting these investors. Find the top 5 investors who have invested maximum number of times (consider repeat investments in one company also). In a startup, multiple investors might have invested. So, consider each investor for that startup. Ignore undisclosed investors.



```
Sequoia Capital 64
Accel Partners 53
Kalaari Capital 44
SAIF Partners 41
Indian Angel Network 40
```

Fig. 5.15 Output of Question-2

Insights:

After analyzing the data, it has become evident that there are certain investors who have heavily invested in multiple startups. Therefore, I would highly recommend my friend to approach these top 5 investors in order to seek funding. It is important to note that this recommendation is based on the data analysis conducted on the available information regarding the investors.

I must highlight that due to the undisclosed nature of some investments, there may be other investors who have invested heavily but are not represented in the data. Nonetheless, based on the available data, the top 5 investors who have invested in multiple startups are listed below:

- Sequoia Capital
- Accel Partners
- Kalaari Capital
- SAIF Partners
- Nexus Venture Partners

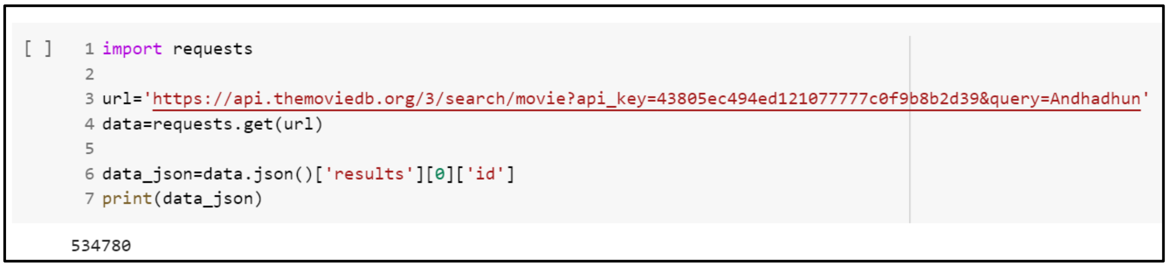
These investors have a track record of investing in startups across various sectors and have a reputation of providing strategic support to their portfolio companies. Therefore, approaching them for funding may prove to be highly beneficial for my friend's startup.

5.3 TMDB API MINI-PROJECT

Question-1:

TMDb API enables you to find out the latest information about TV Shows, Movies and the biggest names in entertainment sector for a marvellous and fun TV/Movie watching experience.

Find the 'id' of the movie "Andhadhun" using TMDb API.

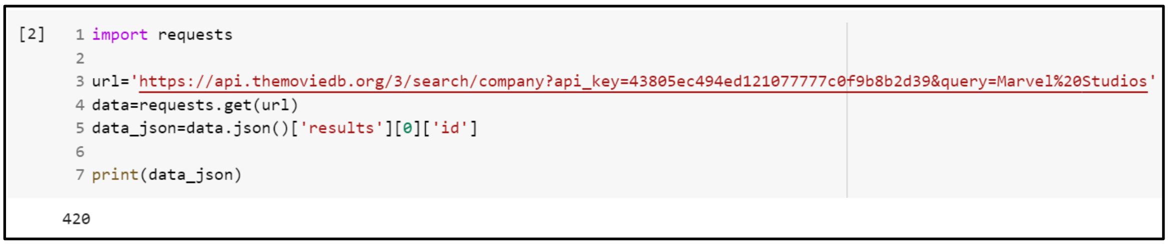


```
[ ]  import requests

     url='https://api.themoviedb.org/3/search/movie?api_key=43805ec494ed121077777c0f9b8b2d39&query=Andhadhun'
     data=requests.get(url)

     data_json=data.json()['results'][0]['id']
     print(data_json)

534780
```

Fig. 5.16 Code & Output for Question-1

Question-2:

Fetch the company id company 'Marvel Studios' using TMDb. Print the id.

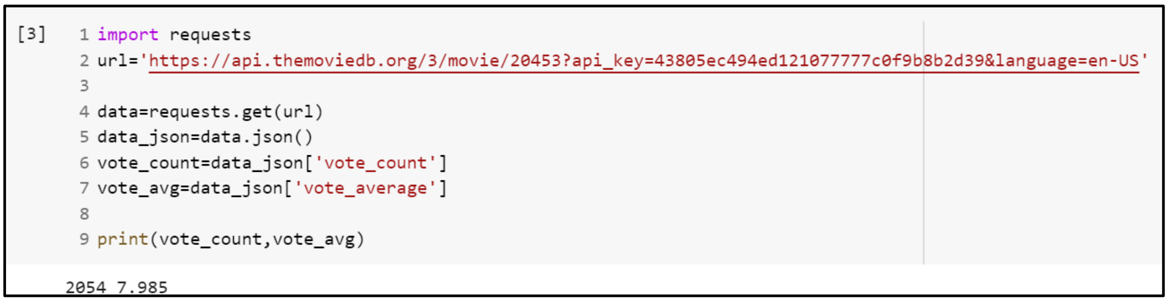


```
[2]  import requests

     url='https://api.themoviedb.org/3/search/company?api_key=43805ec494ed121077777c0f9b8b2d39&query=Marvel%20Studios'
     data=requests.get(url)
     data_json=data.json()['results'][0]['id']

     print(data_json)

420
```

Fig. 5.17 Code & Output for Question-2

Question-3:

Find the vote count and vote average of the movie "3 Idiots" using the TMDb API

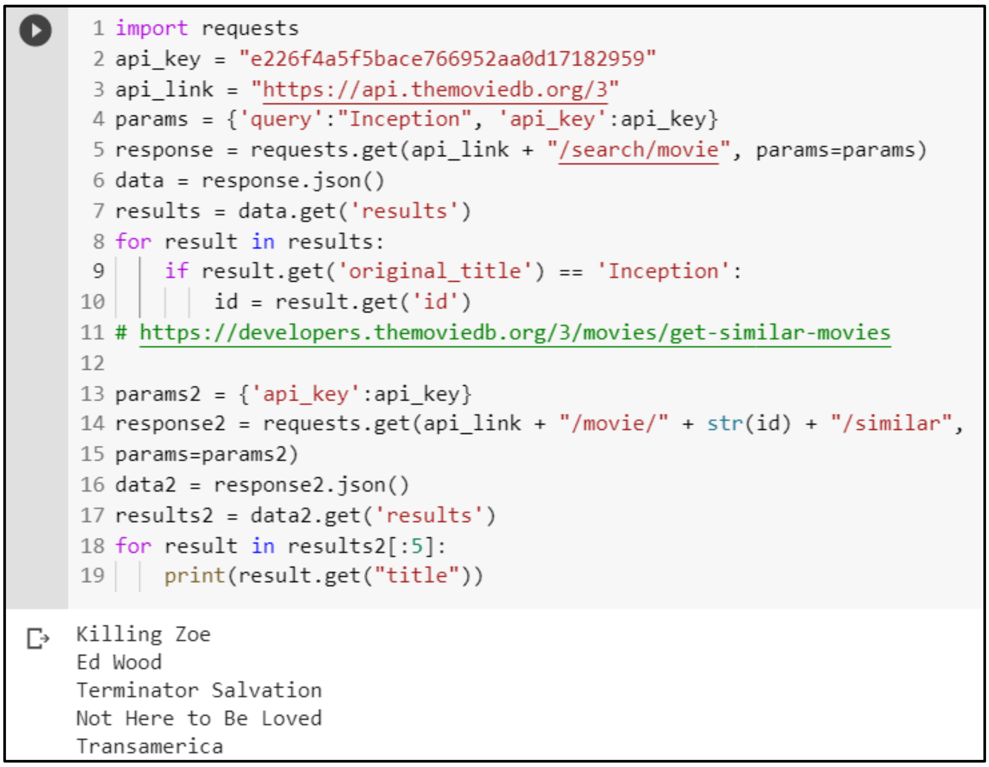


```
[3]  import requests
     url='https://api.themoviedb.org/3/movie/20453?api_key=43805ec494ed121077777c0f9b8b2d39&language=en-US'

     data=requests.get(url)
     data_json=data.json()
     vote_count=data_json['vote_count']
     vote_avg=data_json['vote_average']

     print(vote_count,vote_avg)

2054 7.985
```

Fig. 5.18 Code & Output for Question-3

Question-4:

Fetch the names of top 5 similar movies to 'Inception' from the TMDb API.



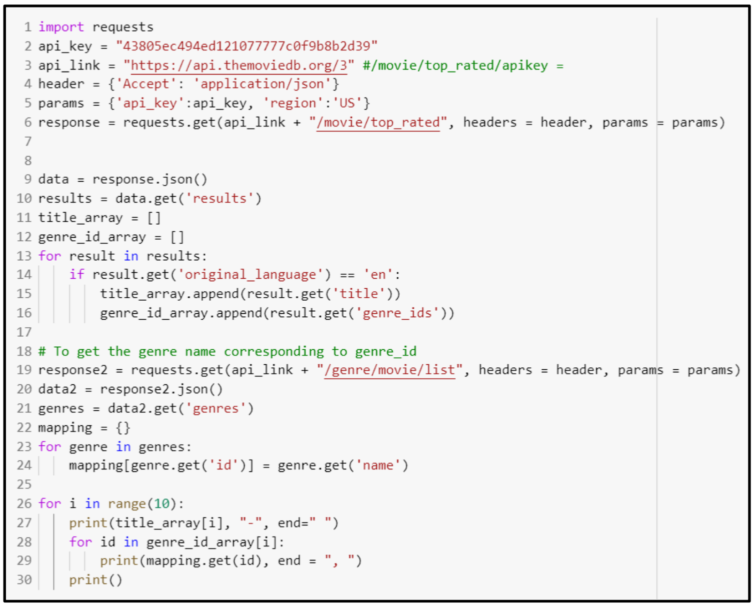
```python
import requests
api_key = "e226f4a5f5bace766952aa0d17182959"
api_link = "https://api.themoviedb.org/3"
params = {'query':"Inception", 'api_key':api_key}
response = requests.get(api_link + "/search/movie", params=params)
data = response.json()
results = data.get('results')
for result in results:
    if result.get('original_title') == 'Inception':
        id = result.get('id')
# https://developers.themoviedb.org/3/movies/get-similar-movies

params2 = {'api_key':api_key}
response2 = requests.get(api_link + "/movie/" + str(id) + "/similar",
params=params2)
data2 = response2.json()
results2 = data2.get('results')
for result in results2[:5]:
    print(result.get("title"))
```

```
Killing Zoe
Ed Wood
Terminator Salvation
Not Here to Be Loved
Transamerica
```

Fig. 5.19 Code & Output for Question-4

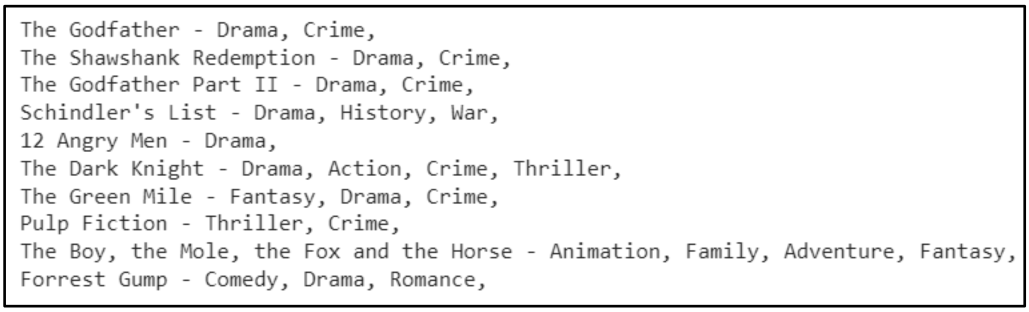
Question-5:

Fetch the top-rated English movies in the US region using the TMDb API. From the result, print the first 10 movies which have original language as English. Also print their genres.



```python
import requests
api_key = "43805ec494ed121077777c0f9b8b2d39"
api_link = "https://api.themoviedb.org/3" #/movie/top_rated/apikey =
header = {'Accept': 'application/json'}
params = {'api_key':api_key, 'region':'US'}
response = requests.get(api_link + "/movie/top_rated", headers = header, params = params)


data = response.json()
results = data.get('results')
title_array = []
genre_id_array = []
for result in results:
    if result.get('original_language') == 'en':
        title_array.append(result.get('title'))
        genre_id_array.append(result.get('genre_ids'))

# To get the genre name corresponding to genre_id
response2 = requests.get(api_link + "/genre/movie/list", headers = header, params = params)
data2 = response2.json()
genres = data2.get('genres')
mapping = {}
for genre in genres:
    mapping[genre.get('id')] = genre.get('name')

for i in range(10):
    print(title_array[i], "-", end=" ")
    for id in genre_id_array[i]:
        print(mapping.get(id), end = ", ")
    print()
```

Fig. 5.20 Code & Output for Question-5

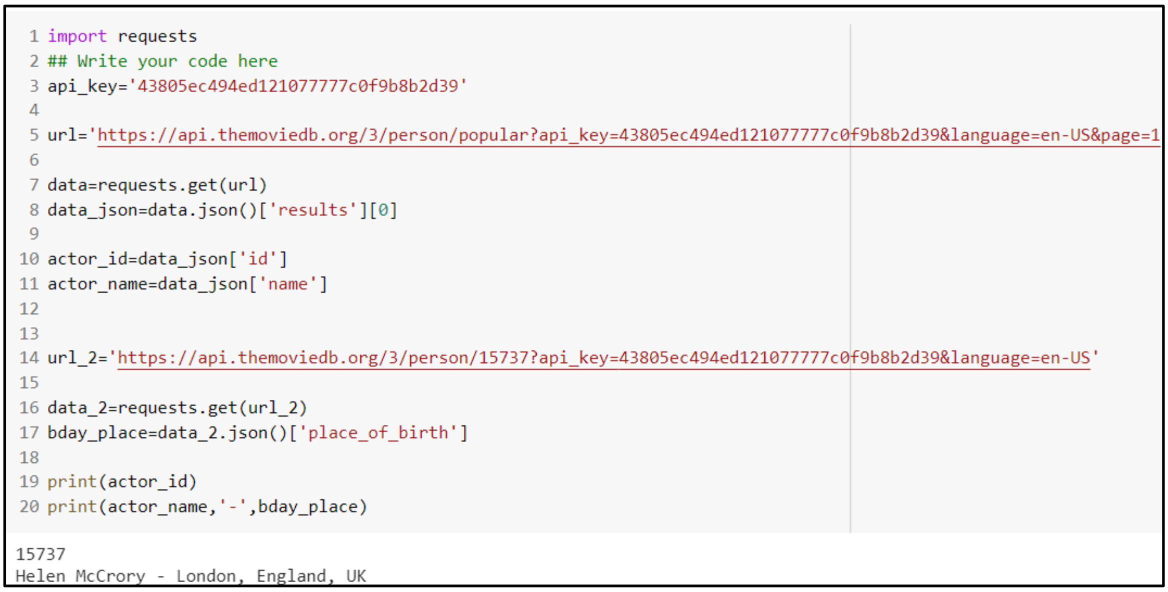


```
The Godfather - Drama, Crime,
The Shawshank Redemption - Drama, Crime,
The Godfather Part II - Drama, Crime,
Schindler's List - Drama, History, War,
12 Angry Men - Drama,
The Dark Knight - Drama, Action, Crime, Thriller,
The Green Mile - Fantasy, Drama, Crime,
Pulp Fiction - Thriller, Crime,
The Boy, the Mole, the Fox and the Horse - Animation, Family, Adventure, Fantasy,
Forrest Gump - Comedy, Drama, Romance,
```

Fig. 5.21 Code & Output for Question-5

Question-6:

Find the name and birthplace of the present most popular person according to TMDb API.



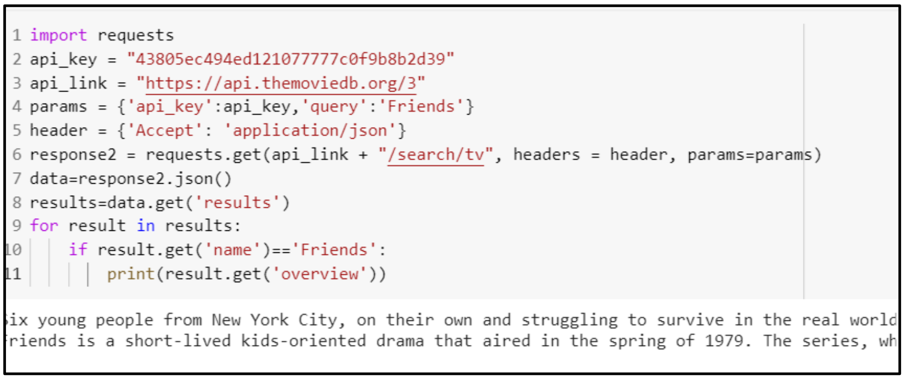
```
import requests
## Write your code here
api_key='43805ec494ed121077777c0f9b8b2d39'

url='https://api.themoviedb.org/3/person/popular?api_key=43805ec494ed121077777c0f9b8b2d39&language=en-US&page=1

data=requests.get(url)
data_json=data.json()['results'][0]

actor_id=data_json['id']
actor_name=data_json['name']


url_2='https://api.themoviedb.org/3/person/15737?api_key=43805ec494ed121077777c0f9b8b2d39&language=en-US'

data_2=requests.get(url_2)
bday_place=data_2.json()['place_of_birth']

print(actor_id)
print(actor_name,'-',bday_place)
```

```
15737
Helen McCrory - London, England, UK
```

Fig. 5.22 Code & Output for Question-6

Question-7:

Fetch the overview of the TV Show "FRIENDS" using TMDb API.



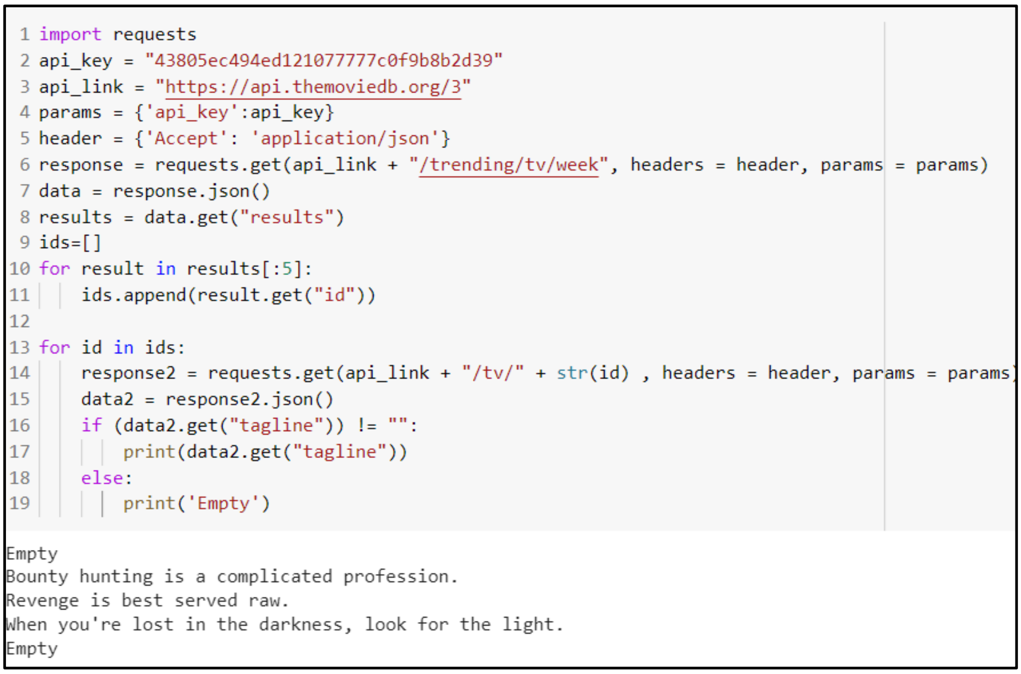
```
import requests
api_key = "43805ec494ed121077777c0f9b8b2d39"
api_link = "https://api.themoviedb.org/3"
params = {'api_key':api_key,'query':'Friends'}
header = {'Accept': 'application/json'}
response2 = requests.get(api_link + "/search/tv", headers = header, params=params)
data=response2.json()
results=data.get('results')
for result in results:
    if result.get('name')=='Friends':
        print(result.get('overview'))
```

```
ix young people from New York City, on their own and struggling to survive in the real world
riends is a short-lived kids-oriented drama that aired in the spring of 1979. The series, wh
```

Fig. 5.23 Code & Output for Question-7

Question-8:

Fetch the trending TV Shows for the week from the TMDb API and print the taglines of the top 5 shows. If there is no tagline, print 'Empty' instead



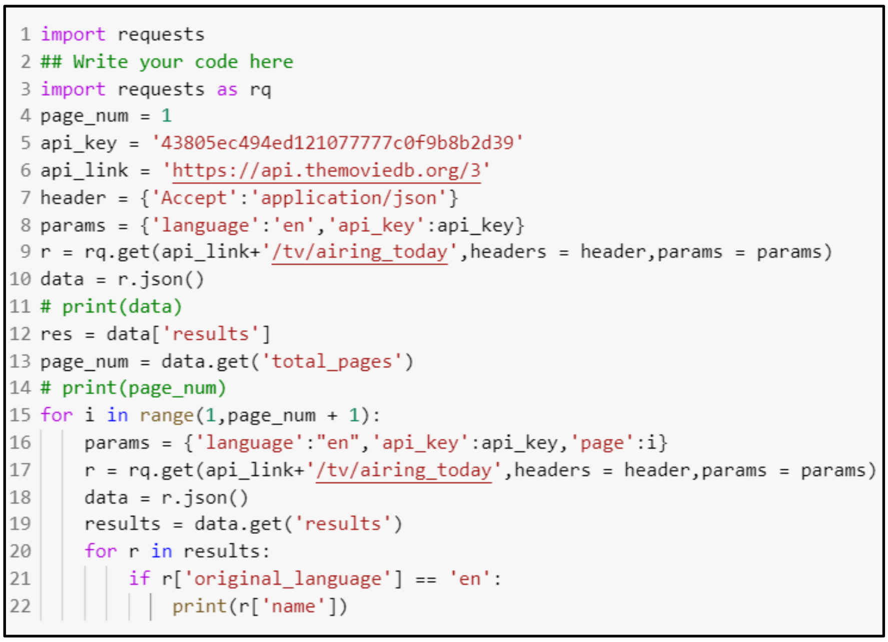
```python
import requests
api_key = "43805ec494ed121077777c0f9b8b2d39"
api_link = "https://api.themoviedb.org/3"
params = {'api_key':api_key}
header = {'Accept': 'application/json'}
response = requests.get(api_link + "/trending/tv/week", headers = header, params = params)
data = response.json()
results = data.get("results")
ids=[]
for result in results[:5]:
    ids.append(result.get("id"))

for id in ids:
    response2 = requests.get(api_link + "/tv/" + str(id) , headers = header, params = params)
    data2 = response2.json()
    if (data2.get("tagline")) != "":
        print(data2.get("tagline"))
    else:
        print('Empty')
```

```
Empty
Bounty hunting is a complicated profession.
Revenge is best served raw.
When you're lost in the darkness, look for the light.
Empty
```

Fig. 5.24 Code & Output for Question-8

Question-9:

Print the names of all the TV shows to be aired today whose original language is english.



```python
import requests
## Write your code here
import requests as rq
page_num = 1
api_key = '43805ec494ed121077777c0f9b8b2d39'
api_link = 'https://api.themoviedb.org/3'
header = {'Accept':'application/json'}
params = {'language':'en','api_key':api_key}
r = rq.get(api_link+'/tv/airing_today',headers = header,params = params)
data = r.json()
# print(data)
res = data['results']
page_num = data.get('total_pages')
# print(page_num)
for i in range(1,page_num + 1):
    params = {'language':"en",'api_key':api_key,'page':i}
    r = rq.get(api_link+'/tv/airing_today',headers = header,params = params)
    data = r.json()
    results = data.get('results')
    for r in results:
        if r['original_language'] == 'en':
            print(r['name'])
```

Fig. 5.25 Code & Output for Question-9

5.4 REVIEW SCRAPPER MINI-PROJECT

The purpose of this project is to extract customer reviews from Flipkart's website and analyze the sentiments expressed by customers towards various products. The project aims to provide insights on customer behaviour, their preferences, and identify areas of improvement for the company.

The web scraping process will involve extracting data from the Flipkart website using Python and Beautiful Soup. The data will include customer reviews, ratings, product names, and other relevant information. The data will be stored in a structured format such as CSV or Excel for further analysis.

The benefits of this project include providing insights on customer behavior, their preferences, and identifying areas of improvement for the company. The project will help in improving the overall customer experience by addressing customer complaints and providing better products and services. The project will also help in providing a competitive advantage to the company by identifying trends in the market and providing better solutions to customers.

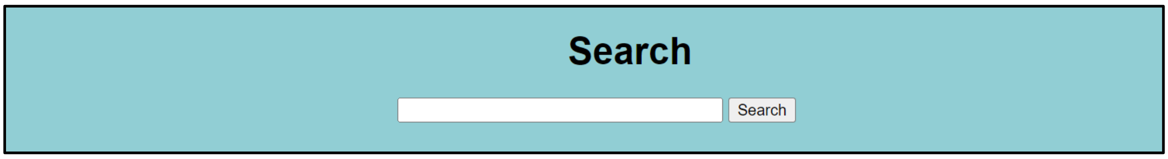
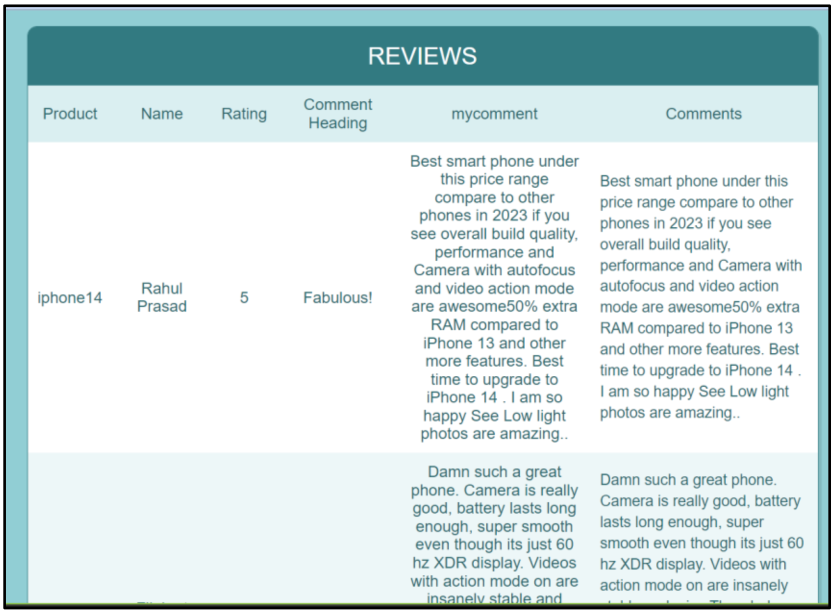


Fig. 5.26 Search Page for the Product Review Scrapping


REVIEWS

Product	Name	Rating	Comment Heading	mycomment	Comments
iphone14	Rahul Prasad	5	Fabulous!	Best smart phone under this price range compare to other phones in 2023 if you see overall build quality, performance and Camera with autofocus and video action mode are awesome50% extra RAM compared to iPhone 13 and other more features. Best time to upgrade to iPhone 14 . I am so happy See Low light photos are amazing..	Best smart phone under this price range compare to other phones in 2023 if you see overall build quality, performance and Camera with autofocus and video action mode are awesome50% extra RAM compared to iPhone 13 and other more features. Best time to upgrade to iPhone 14 . I am so happy See Low light photos are amazing..
				Damn such a great phone. Camera is really good, battery lasts long enough, super smooth even though its just 60 hz XDR display. Videos with action mode on are insanely stable and	Damn such a great phone. Camera is really good, battery lasts long enough, super smooth even though its just 60 hz XDR display. Videos with action mode on are insanely

Fig. 5.27 Output after Scrapping is Completed.

5.5 PAN CARD TAMPERING MINI-PROJECT

Introduction:

The purpose of this internship mini-project is to develop a system that can detect tampered PAN cards using Python, tesseract, openCV, and other relevant technologies. This project aims to address the issue of fraudulent practices in obtaining PAN cards through tampering of personal information such as name, date of birth, and photograph. The project involves building a program that can recognize and compare the content of a scanned PAN card with a standard PAN card template to detect any discrepancies.

Overview of the Project:

The project involves developing a program that can identify tampered PAN cards through the use of image processing and optical character recognition (OCR) techniques. The system will require the user to input a scanned copy of the PAN card, and the program will then analyze the image to determine whether any tampering has taken place. The project will be implemented using Python as the primary programming language, with the use of tesseract and openCV libraries for OCR and image processing, respectively.

Project Objectives:

The primary objective of the project is to develop a system that can detect tampered PAN cards accurately and efficiently. The specific objectives of the project include:

Developing a program that can read and extract text from scanned PAN card images using OCR technology.

Designing a standard PAN card template for the system to compare the scanned PAN card with and identify any differences.

Integrating the image processing techniques to detect any tampering in the scanned image such as photo manipulation or alterations in the personal details of the cardholder.

Providing a user-friendly interface for users to input the scanned PAN card and display the results of the tampering detection analysis.

Methodology:

The methodology for the PAN Card Tampering Detection project involves a series of steps that are necessary for the system to work correctly. The steps are:

Image Preprocessing: The system will use openCV libraries to preprocess the scanned PAN card image, which includes resizing, cropping, and color conversion to improve the image quality.

Text Extraction: The tesseract OCR library will be used to extract the text content of the PAN card image.

Template Matching: The program will use the standard PAN card template as a reference to compare with the scanned image to identify any differences.

Image Comparison: The system will utilize image processing techniques such as histogram equalization and edge detection to identify any discrepancies in the scanned PAN card image compared to the template.

Tampering Detection: The program will analyze the text and image data of the scanned PAN card to identify any tampering or alterations made to the card.

User Interface: The program will provide a user-friendly interface for users to input the scanned PAN card and display the results of the tampering detection analysis.

Conclusion:

The PAN Card Tampering Detection project is a significant initiative in addressing the issue of fraudulent practices in obtaining PAN cards through tampering of personal information. The system developed using Python, tesseract, and openCV technologies can accurately and efficiently detect tampered PAN cards, providing a reliable solution to the problem. The project's successful implementation will provide a valuable contribution to the government's efforts in ensuring the security and integrity of PAN cards issued to citizens.

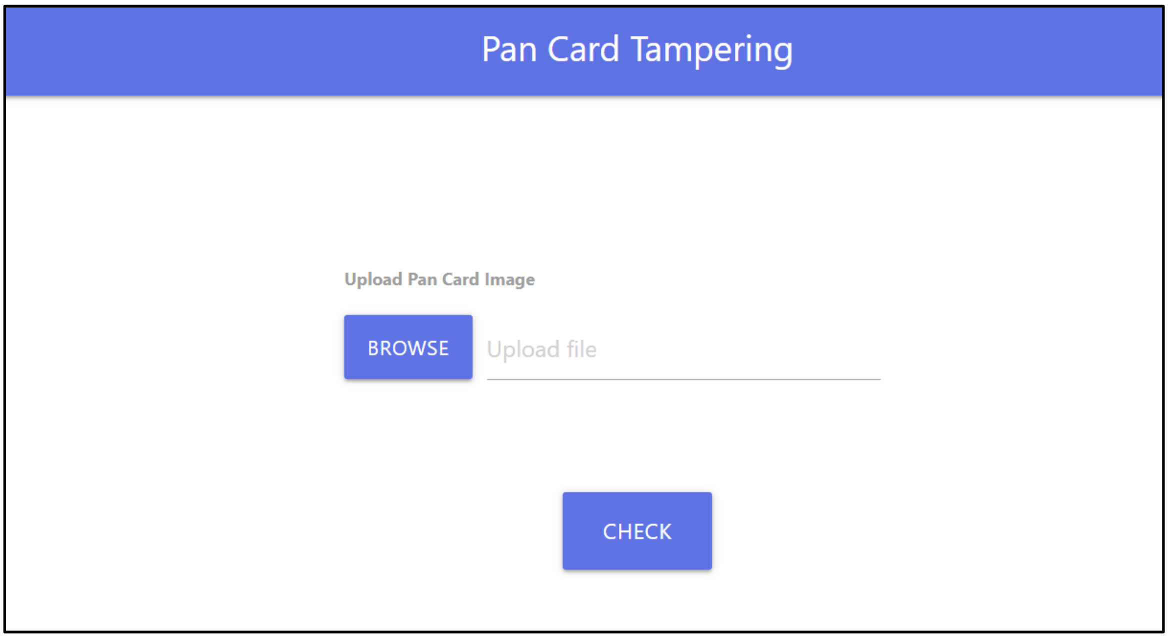


Fig. 5.28 Initial UI of the application

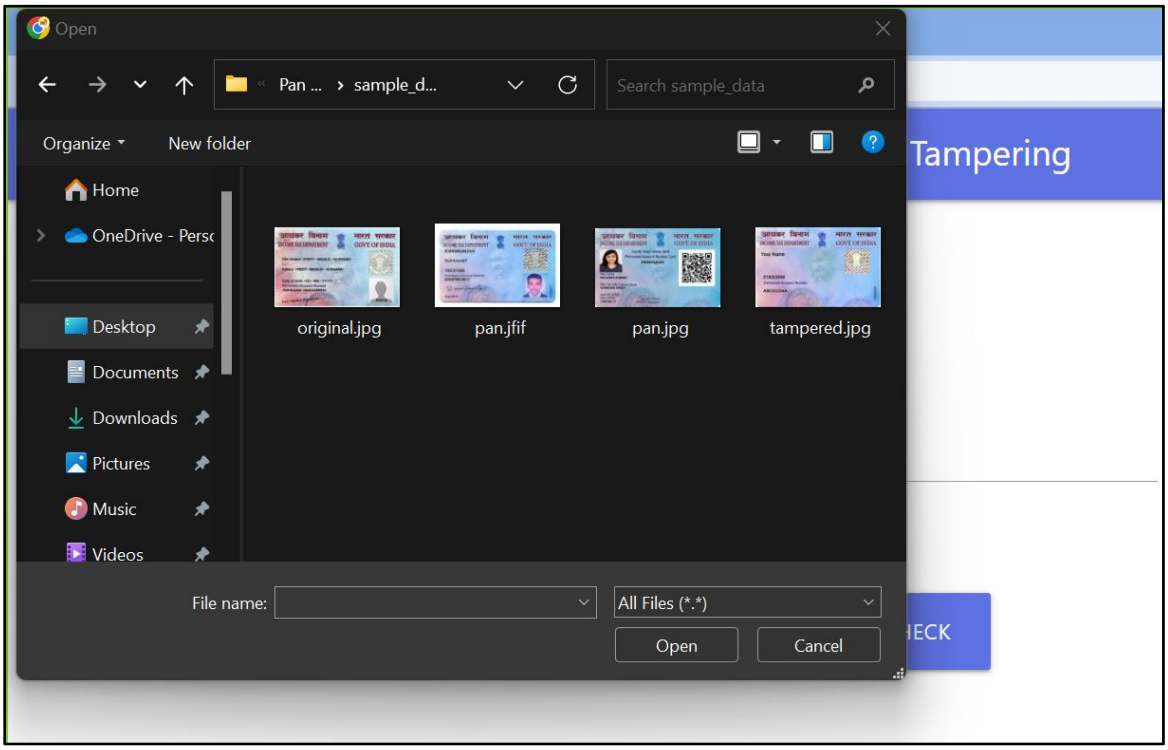
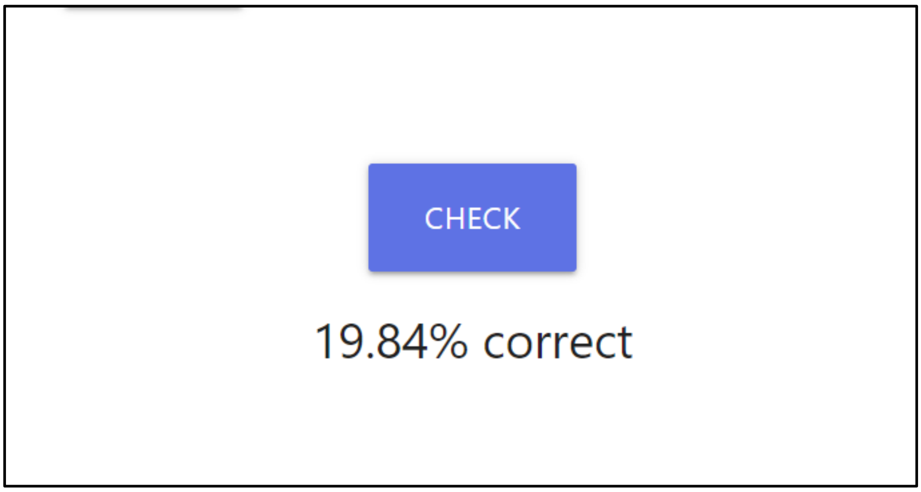
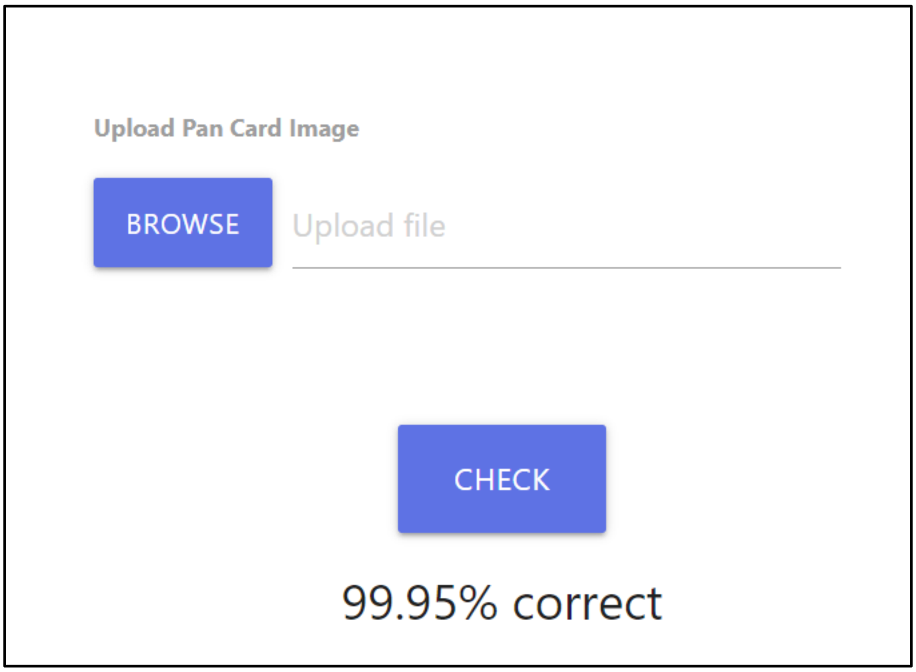


Fig. 5.29 Upload Section to upload the Pan Card Image



19.84% correct

Fig. 5.30 Result of Tampered Pan Card



99.95% correct

Fig. 5.31 Result of an original Pan Card

5.6 TEXT EXTRACTION PROJECT

Introduction:

Text extraction is an essential task in various fields such as document management, data analysis, and information retrieval. With the advancement in technology, automated text extraction has become a need for many businesses to save time and reduce errors. In this mini-project, we aim to develop a text extraction system using Python, OpenCV, Tesseract, and other relevant libraries.

Overview:

The text extraction system developed in this project extracts text from images and converts it into machine-readable text. The system can handle images with various types of text, including handwritten text, printed text, and text with different font styles and sizes. The extracted text is then saved in a file for further processing or analysis.

Method:

The text extraction system developed in this project follows the following steps:

Image Acquisition: The first step is to acquire the image containing the text. The image can be acquired from various sources, such as a scanner, camera, or downloaded image.

Preprocessing: The acquired image is preprocessed to enhance the quality of the text. The preprocessing steps include noise reduction, thresholding, and image smoothing.

Text Detection: The preprocessed image is then processed for text detection. The system uses OpenCV's text detection algorithm to identify the location of the text in the image.

Text Recognition: Once the text location is identified, the system uses Tesseract OCR to recognize the text. Tesseract OCR is an open-source OCR engine that can recognize various fonts and text sizes.

Post-processing: The recognized text may contain errors due to noise, low-quality image, or complex text layout. To minimize these errors, the recognized text is post-processed by applying text normalization techniques such as spelling correction, punctuation removal, and stop-word removal.

Output: The final step is to save the extracted text into a file format such as txt, doc, or pdf.

Relevant Libraries:

The following libraries are used in this project:

OpenCV: OpenCV is an open-source computer vision library used for image and video processing.

Tesseract OCR: Tesseract OCR is an open-source OCR engine that can recognize various fonts and text sizes.

Pytesseract: Pytesseract is a Python wrapper for Tesseract OCR, making it easier to use in Python-based projects.

Numpy: Numpy is a numerical computing library in Python used for array operations.

Pandas: Pandas is a data manipulation library in Python used for data analysis.

Conclusion:

The text extraction system developed in this mini-project can extract text from various types of images and convert it into machine-readable text. The system uses OpenCV's text detection algorithm and Tesseract OCR for text recognition. The extracted text can be saved in a file format for further processing or analysis. The use of Python and relevant libraries has made the development of this system efficient and effective. The system can be further improved by incorporating deep learning techniques for better accuracy and performance.

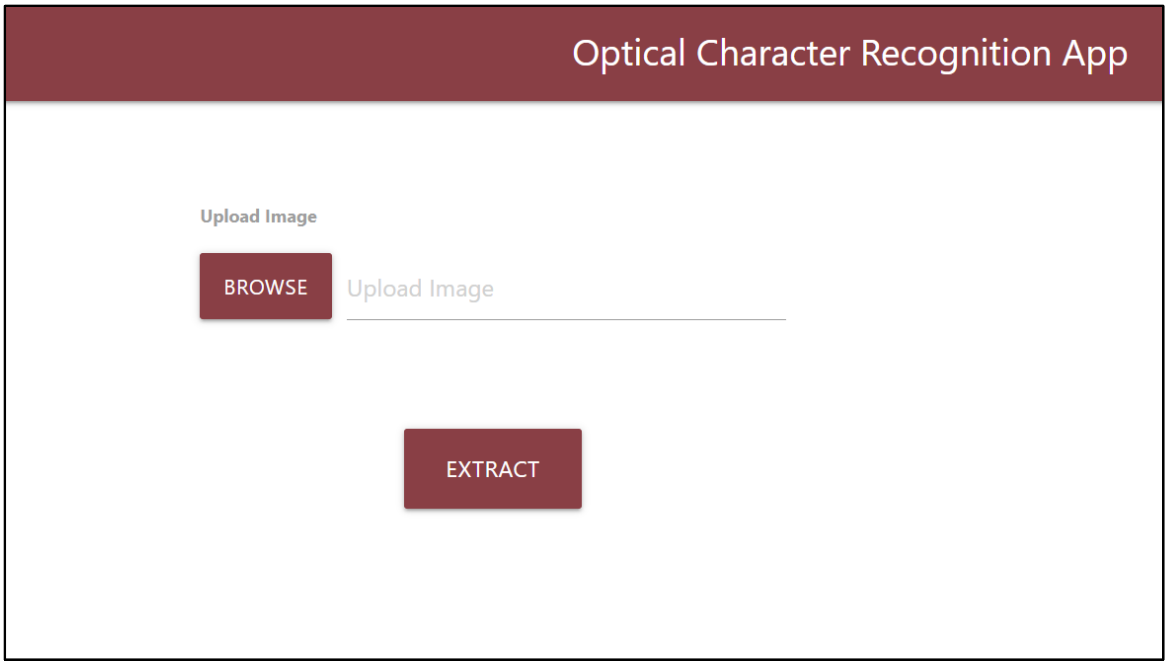


Fig. 5.32 Initial Page of the application

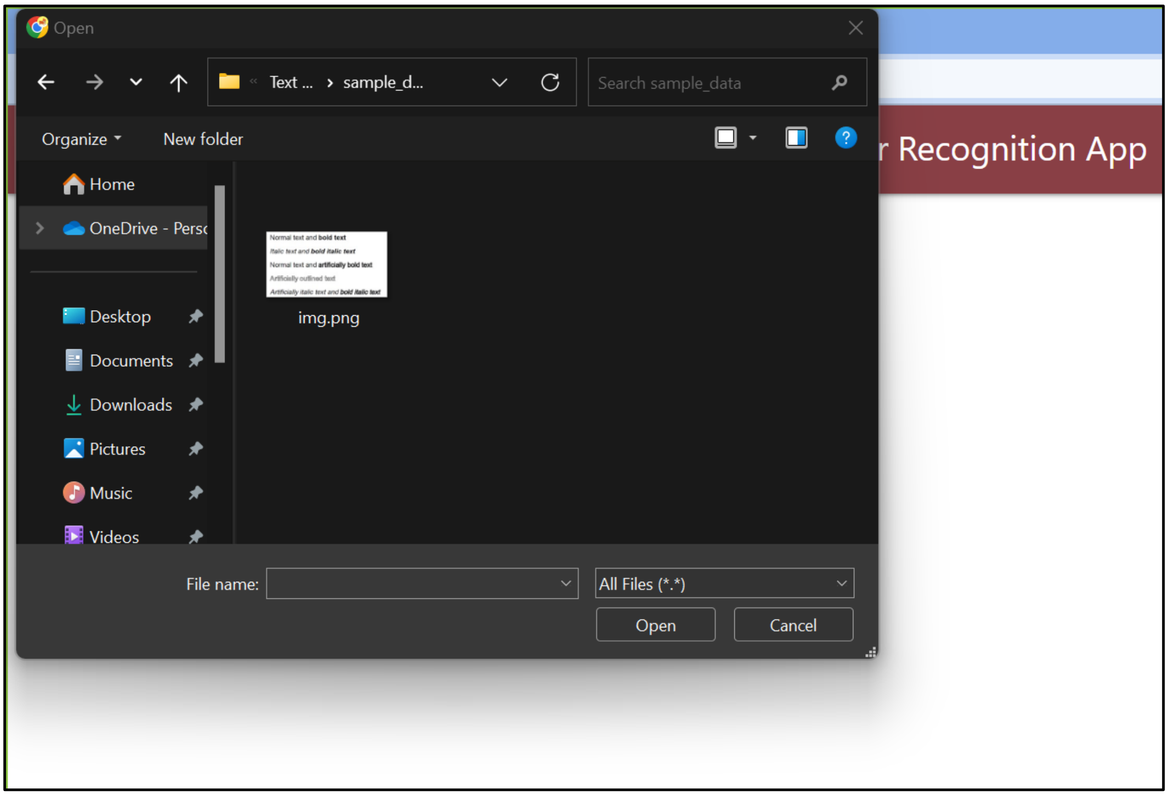
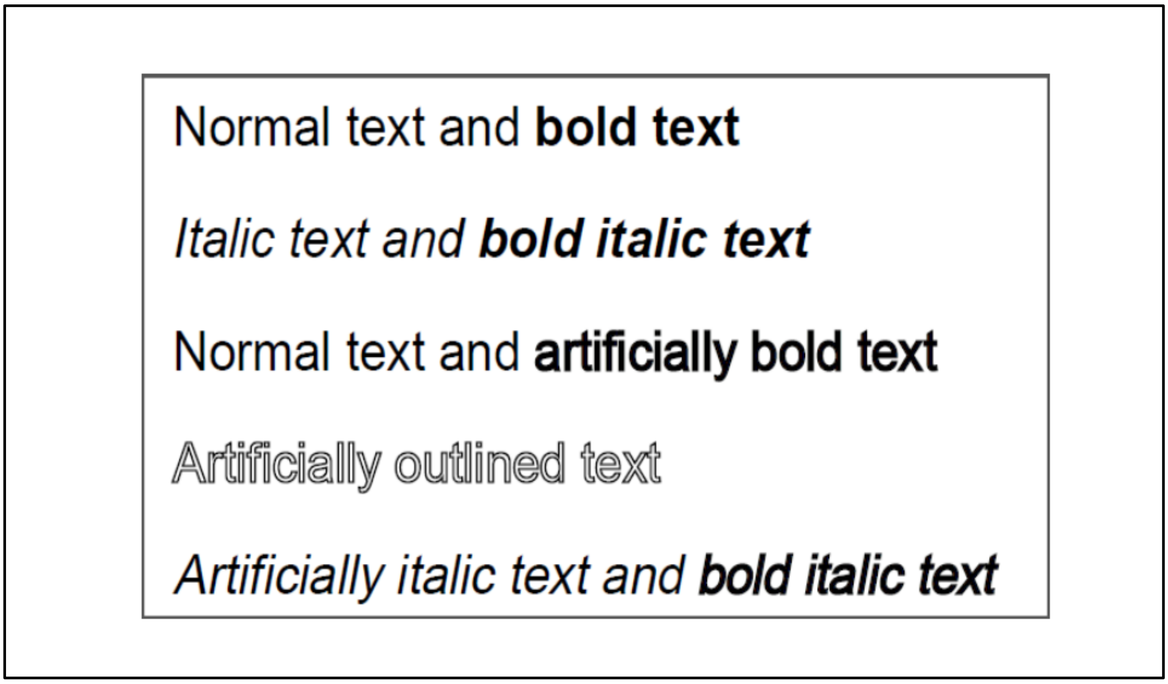


Fig. 5.33 Upload Section for the extraction



Normal text and **bold text**

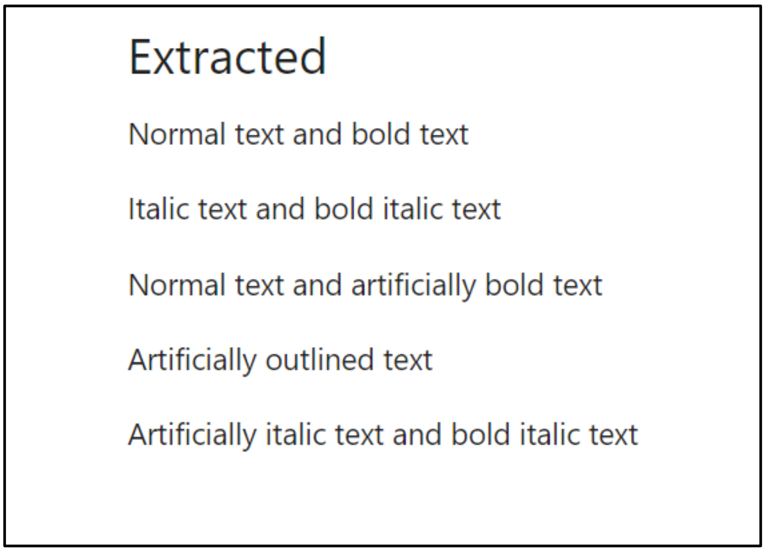
*Italic text and **bold italic text***

Normal text and **artificially bold text**

Artificially outlined text

*Artificially italic text and **bold italic text***

Fig. 5.34 Image Uploaded



Extracted

Normal text and bold text

Italic text and bold italic text

Normal text and artificially bold text

Artificially outlined text

Artificially italic text and bold italic text

Fig. 5.35 Text Extracted

5.7 CRM DUTY

As an intern, working on customer relationship management (CRM) is an essential aspect of building professional skills and gaining experience in the industry. In CRM, an intern will be responsible for building and maintaining strong relationships with customers by understanding their needs, addressing their concerns, and providing high-quality support.

One of the key duties of an intern in CRM is to ensure customer satisfaction. This involves listening to customers' queries, feedback, and complaints, and taking appropriate actions to resolve any issues they may have. This can include answering questions, providing technical support, or offering alternative solutions. Another important responsibility of an intern in CRM is to maintain accurate customer records. This includes collecting and organizing data about customers, such as their contact information, purchase history, and preferences. Accurate customer records help to ensure that customers receive timely and personalized support and that the company can make informed decisions about its products and services. In addition to customer support, an intern in CRM may also be responsible for developing and implementing customer engagement strategies. This can involve creating targeted marketing campaigns, organizing customer events or webinars, and developing social media content to engage with customers.

Effective communication is crucial in CRM, and as an intern, one should have excellent communication skills to ensure that customers feel heard and valued. One should also be able to work collaboratively with other team members, such as sales and marketing teams, to ensure that customer needs are met and that the company's goals are aligned.

In conclusion, working on customer relationship management as an intern is an excellent opportunity to develop interpersonal skills, gain experience in the industry, and contribute to the success of a company. It involves building and maintaining strong relationships with customers, ensuring their satisfaction, maintaining accurate records, and developing effective customer engagement strategies. Effective communication and collaboration are essential in this role, and by mastering these skills, an intern can make a valuable contribution to any organization.

5.8 INTERN MANAGEMENT

As an intern, one of my responsibilities was intern management, where I was responsible for overseeing and assisting other interns in the team. This role was critical in ensuring that the project ran smoothly and all deliverables were met on time.

One of the key aspects of intern management was being available to assist other interns in solving any problems they faced during their work. This required good communication skills and the ability to work collaboratively with others. I made sure to set aside time each day to check in with the other interns and ask if they needed any assistance or had any questions.

In addition to assisting with technical issues, I also helped to resolve any interpersonal issues that arose between the interns. I found that being a good listener and mediator was important in this role, as it allowed me to understand the root cause of the problem and come up with effective solutions.

Another important aspect of intern management was assisting our manager in monitoring the progress of the project. I was responsible for keeping track of each intern's work and reporting any issues or concerns to our manager. This required good organizational skills and attention to detail, as I needed to ensure that all tasks were being completed on time and to the required standard.

Overall, intern management was a challenging but rewarding aspect of my internship. It allowed me to develop my leadership skills and work collaboratively with others. By taking on this responsibility, I was able to contribute to the success of the project and gain valuable experience in project management.

6 TESTING

6.1 CODING STANDARDS

Coding standards play a vital role in software development as they ensure consistency, readability, and maintainability of the code. Data analytics, data visualization, and data science projects are no different. In this document, we will discuss the coding standards that should be followed while developing code for such projects.

Naming Conventions:

- Variable names should be in lowercase, and if a variable name contains multiple words, they should be separated by underscores.
    - Example: total_sales, customer_name

- Function names should be in lowercase and use underscores to separate words.
    - Example: calculate_total_sales()

- Class names should be in CamelCase.
    - Example: SalesReportGenerator

- Module names should be in lowercase and should not contain underscores.
    - Example: sales_report.py

- Commenting:
    - Code should be well-documented with comments explaining the functionality of the code.
    - Comments should be clear, concise, and easy to understand.
    - Use comments to explain the logic behind the code, not what the code is doing.
    - Avoid unnecessary comments that do not add value to the code.

- Formatting:
    - Use four spaces for indentation, not tabs.
    - Lines should not exceed 79 characters.
    - Blank lines should be used to separate code blocks logically.

- Functions:
    - Functions should be designed to perform a single task and should not be too long.
    - Use descriptive names for functions.
    - Functions should have a clear purpose and should be well-documented.
    - Use default arguments to make functions more flexible.

- Libraries:
    - Import libraries at the top of the script.
    - Use short, descriptive aliases for libraries.
    - Use specific imports instead of importing everything from a library.

- Error Handling:
    - Use try-except blocks to handle errors.
    - Use specific exception types instead of catching all exceptions.
    - Use meaningful error messages to help users understand what went wrong.

These are some of the coding standards that should be followed while developing code for data analytics, data visualization, and data science projects. By following these standards, we can ensure that the code is consistent, easy to read, and maintainable.

6.2 TESTING AND TEST CASES

In the field of data analytics, data visualization, and data science, testing is not considered as important as it is in software development. The main reason behind this is the nature of these projects. Unlike software development, where there is a set of predefined requirements that the product must meet, data science projects are exploratory in nature. Data science projects are more focused on uncovering insights and patterns in data that can be used to inform business decisions, rather than on building a product that meets specific requirements. In data science projects, the data itself is the main focus, and the algorithms and techniques used to extract insights from the data are often complex and difficult to understand. This means that testing becomes a challenging task, as it is difficult to predict the expected outcomes of data processing and analysis. Additionally, the data used in these projects is often large and complex, making it difficult to create a comprehensive test suite that covers all possible scenarios.

Another reason why testing is not as important in data science projects is that the results of these projects are often subjective. The insights and patterns uncovered in the data are not always black and white, and there is often room for interpretation. This means that even if a test suite is created, it may not provide a clear indication of the accuracy of the results.

However, this does not mean that testing is completely irrelevant in data science projects. While traditional testing methods may not be suitable, there are other methods that can be used to ensure the accuracy of the results. One such method is to use a validation set, which is a subset of the data that is not used in the analysis process but is reserved for testing the accuracy of the model.

Another important aspect of ensuring the accuracy of data science projects is to use good data management practices. This includes ensuring the data is clean, complete, and free from errors, and ensuring that any preprocessing steps are documented and reproducible.

In conclusion, testing is not as important in data analytics, data visualization, and data science projects as it is in software development. However, this does not mean that accuracy is not important. Rather, the focus is on using alternative methods to ensure the accuracy of the results, such as using validation sets and good data management practices.

Testing and test cases are important parts of software development, but they are not always applicable in data analytics, data visualization, and data science projects. Here are some reasons why:

Data analytics and visualization projects often involve exploratory analysis and visual inspection of data, which is not always deterministic in nature. Unlike software development, where the inputs and outputs are well-defined, the output of data analytics and visualization projects can be subjective and dependent on the analyst's interpretation of the data. As a result, it is not always possible to define clear test cases.

Data analytics and visualization projects are often iterative in nature, with the analyst exploring different data sets, visualization techniques, and analysis methods to identify patterns and insights. In this type of environment, it can be difficult to create comprehensive test cases that cover all possible scenarios.

Data analytics and visualization projects often involve data sets that are too large to test comprehensively. In many cases, analysts use random sampling techniques to select a representative subset of the data for analysis. It is not always possible or practical to test the entire data set.

Data analytics and visualization projects often involve multiple tools and techniques, which can make it difficult to create comprehensive test cases. For example, an analyst may use R for data analysis, Tableau for visualization, and Excel for data cleaning. Testing all of these tools and their interactions can be time-consuming and resource-intensive.

While testing and test cases may not be relevant for all data analytics, data visualization, and data science projects, it is still important to ensure the accuracy and integrity of the data being analyzed. This can be achieved through data validation techniques such as cross-validation, outlier detection, and data profiling. Additionally, analysts should document their analysis methodology and provide transparency around their data sources and assumptions to enable others to replicate their results.

As per the university mandate, I would like to provide an overview of the testing process for the project, even though testing may not be considered mandatory for data science and analytics projects. While the project has not yet reached the testing phase, I would like to explain a few test cases that could be used to ensure the quality and reliability of the project.

I understand that test suites have not been designed, and the organization has prohibited their sharing. However, I believe it is important to highlight the potential testing requirements for this project, which could include unit testing, integration testing, and system testing. These tests can help ensure that the project meets the requirements and specifications and performs as intended.

Unit testing could involve testing individual functions and methods within the project to ensure that they behave as expected and provide the correct output. Integration testing could involve testing how different modules or components within the project work together, while system testing could involve testing the project as a whole to ensure that it meets the functional and non-functional requirements.

While testing may not be a mandatory requirement for data science and analytics projects, it can play an important role in ensuring the quality, reliability, and accuracy of the project.

6.2.2 Test Cases and scenarios for data visualization projects

Here are some test cases that can be used for testing data visualization projects using Tableau:

Data Accuracy Test:

- Verify that the data is accurately reflected in the visualization.
- Ensure that the data is up-to-date and the same as that in the data source.
- Verify that the aggregated data is correct.

Functionality Test:

- Test the interactive functionality of the visualization, such as filtering, sorting, and zooming in/out.
- Check the tooltip text, drill-down features, and other interactive elements.

Performance Test:

- Test the loading time of the visualization.
- Check the rendering time of the visualization when interacting with different features.
- Test the response time of the visualization when changing the data source.

Compatibility Test:

- Ensure that the visualization works on different web browsers and versions.
- Test the visualization on different operating systems, such as Windows, Mac, and Linux.
- Check the visualization on different mobile devices, such as tablets and smartphones.

Usability Test:

- Verify that the visualization is easy to understand and navigate.
- Check the labelling of axes, legends, and color schemes.
- Test the accessibility of the visualization, such as screen reader compatibility.

Security Test:

- Ensure that the data is secure and not exposed to unauthorized users.
- Test the access controls of the visualization and ensure that only authorized users can access it.
- Check for any data leakage or vulnerability in the visualization.

Integration Test:

- Verify that the visualization integrates well with other tools and applications, such as data sources and reporting tools.
- Test the compatibility of the visualization with different file formats, such as PDF, CSV, and Excel.
- These are some of the test cases that can be used for testing data visualization projects using Tableau

6.2.3 Test Cases and Scenario for Pan Card Tampering Detection Project

Here are some test cases that could be designed for a pan card tampering detection project:

Valid Pan Card:

- Test the system's ability to identify a valid, unaltered pan card.

Altered Pan Card:

- Test the system's ability to identify a pan card that has been altered in some way, such as a changed photo or name.

Fake Pan Card:

- Test the system's ability to identify a fake pan card that has been created using fraudulent information.

Blurred or Low-Quality Pan Card Image:

- Test the system's ability to identify tampering in pan card images that are of poor quality or low resolution.

Different Light Conditions:

- Test the system's ability to identify tampering in pan card images taken under different lighting conditions.

Partially Covered Pan Card:

- Test the system's ability to identify tampering when part of the pan card is covered or obscured.

Different Languages:

- Test the system's ability to identify tampering in pan card images that contain text in different languages.

Multiple Tampering Techniques:

- Test the system's ability to identify tampering when multiple tampering techniques have been used on the same pan card image.

Multi-pan card:

- Test the system's ability to identify tampering in multiple pan cards simultaneously.

Mobile Camera Images:

- Test the system's ability to identify tampering in pan card images taken through mobile cameras.

6.2.4 Test Cases and Scenarios for Text Extraction Project

Here are some test cases for a text extraction project from an image:

Input image file format:

- Ensure that the system can accept the input image file in different formats such as PNG, JPEG, TIFF, etc.

Image resolution:

- Test the system's ability to extract text from images with different resolutions.

Image quality:

- Test the system's ability to extract text from images with different levels of quality, such as blurred or pixelated images.

Language support:

- Test the system's ability to extract text from images in different languages.

Text recognition accuracy:

- Check the system's ability to accurately extract text from an image. Test the accuracy with images with varying degrees of text complexity, such as font size, style, color, orientation, and background.

Text processing speed:

- Test the time taken by the system to extract text from the image. Ensure that the system can handle a large number of images without slowing down.

Handling multiple text regions:

- Test the system's ability to extract text from multiple regions of an image. Ensure that the system can handle different regions and extract text accurately from each region.

Error handling:

- Test the system's ability to handle errors, such as if the image file is corrupted, or if there is no text present in the image.

Integration with other systems:

- Test the system's ability to integrate with other systems, such as database or cloud storage. Ensure that the extracted text can be easily transferred to other systems for further processing

7 CONCLUSION

7.1 SELF ANALYSIS & DISCUSSION

During my 18-week internship in data analytics, visualization, and data science, I was able to gain a wealth of experience and skills that have prepared me for a successful career in the field. Throughout the internship, I had the opportunity to work on real-time projects for clients, develop mini-projects, and participate in intern and customer relationship management. Working on real-time projects was an invaluable experience that helped me to develop my technical skills and learn how to effectively communicate with clients. One of the most notable projects I worked on was for a healthcare industry client, where I was responsible for analyzing large datasets and creating visualizations to identify trends and patterns. Through this project, I learned how to manage expectations, deliver high-quality work, and operate in a fast-paced environment.

Along with real-time projects, I developed several mini-projects to deepen my knowledge of various data analysis and visualization techniques. For example, I worked on a project that used machine learning algorithms to predict customer churn for a telecommunications company. I also developed a project that used natural language processing to analyze customer feedback and identify common themes and sentiments. In addition to technical skills, I also gained valuable interpersonal skills during my internship. I had the opportunity to communicate with different clients, team members, and stakeholders, which helped me learn how to collaborate effectively, manage conflicts, and present ideas in a clear and concise manner. Participating in intern and customer relationship management also gave me valuable experience in leadership and teamwork. I worked with other interns to ensure that we met our deliverables and exceeded expectations, and I worked with the customer success team to ensure that our projects aligned with the customers' goals.

Overall, my internship in data analytics, visualization, and data science was an incredible experience that allowed me to develop a wide range of skills and knowledge. I am excited to continue growing and learning in this field and apply my skills and experience to make a positive impact.

7.2 PROBLEMS ENCOUNTERED DURING THE INTERNSHIP

During my internship in data analytics, visualization, and data science, I encountered both common and hard problems that required creative problem-solving skills and perseverance. Here are some of the most common and challenging problems I faced:

Data Quality Issues: One of the most common problems I encountered during my internship was data quality issues. This included missing data, incorrect data, and inconsistent data. To overcome this problem, I had to use various data cleaning techniques such as imputation, outlier detection, and data normalization. I also had to work closely with the client to understand their data collection processes and identify ways to improve data quality.

Data Security: Another common problem was ensuring data security. As we worked with sensitive data, we had to ensure that the data was protected from unauthorized access and that the data was not compromised during the analysis process. We had to ensure that we adhered to strict data security protocols and obtain proper approvals from the client before accessing any data.

Technical Challenges: During the development of mini-projects, I encountered various technical challenges such as integrating different software, working with different data formats, and optimizing code for efficiency. To overcome these challenges, I had to conduct research, experiment with different techniques, and consult with my mentors and team members.

Managing Expectations: When working on real-time projects, it was important to manage expectations with the client and stakeholders. This included setting realistic deadlines, providing regular progress updates, and ensuring that we met the client's requirements. We had to communicate regularly with the client to ensure that we were meeting their expectations and deliverables.

Communication and Collaboration: Communication and collaboration with team members, clients, and stakeholders were critical for the success of our projects. We had to work closely with the client to understand their requirements and expectations and communicate any issues or challenges we encountered during the project. We also had to collaborate with our team members to ensure that we were meeting our deadlines and delivering high-quality work.

Data Visualization: One of the most challenging aspects of the internship was creating effective data visualizations. We had to ensure that the visualizations effectively communicated the insights we had uncovered from the data while also being aesthetically pleasing and easy to understand. We had to experiment with different chart types and visualization techniques to identify the best way to present the data.

Overall, my internship in data analytics, visualization, and data science presented various challenges that required me to use my problem-solving skills and work closely with my team members and clients to overcome them. These challenges provided me with valuable experience and prepared me for future work in the field.

7.3 INTERNSHIP SUMMARY

In conclusion, my internship experience in data analytics, visualization, and data science was extremely valuable and rewarding. I had the opportunity to work on real-time projects and develop mini-projects, gaining practical experience and in-depth knowledge of various tools and technologies like Python, Google Colab, Pandas, numpy, matplotlib, seaborn, beautifulsoup, data analytics, text extraction, model training, feature engineering, visualization using tableau, and management, communication.

Throughout the internship, I faced various challenges and obstacles, including data quality issues, feature engineering problems, and communication challenges with clients and team members. However, through these challenges, I was able to develop my problem-solving skills and learn how to effectively communicate and collaborate with others to achieve our goals.

In addition to the technical skills I gained, I also developed my interpersonal and management skills through intern management and customer relationship management. I learned how to effectively collaborate with others, manage conflicts, and present my ideas in a clear and concise manner. These skills will undoubtedly be valuable in my future career in the data analytics and data science field.

Overall, my internship experience provided me with a wealth of knowledge and experience that will undoubtedly be valuable in my future career. I am grateful for the opportunity to work on real-time projects, develop mini-projects, and learn from experienced professionals in the field. I look forward to applying the skills and knowledge I gained during my internship to future projects and continuing to grow and develop in this field.

BIBLIOGRAPHY

1. *"Python for Data Analysis"* by Wes McKinney

This book is a comprehensive guide to data analysis in Python. It covers the use of libraries such as Pandas, NumPy, and Matplotlib, which are essential for data analytics and visualization.

2. *"Data Visualization with Tableau"* by Ben Jones

This book provides a comprehensive guide to data visualization using Tableau. It covers various visualization techniques and how to use Tableau to create effective visualizations.

3. *"Python Data Science Handbook"* by Jake VanderPlas

This book is a complete guide to data science in Python. It covers various aspects of data science such as machine learning, data visualization, and data manipulation using libraries like Pandas, NumPy, and Scikit-learn.

4. *"Python Machine Learning"* by Sebastian Raschka and Vahid Mirjalili

This book is a comprehensive guide to machine learning in Python. It covers various algorithms and techniques used in machine learning, such as decision trees, random forests, and neural networks.

5. *"Customer Relationship Management: Concepts and Technologies"* by Francis Buttle

This book provides an overview of customer relationship management and its various concepts and technologies. It covers topics such as customer retention, customer loyalty, and customer satisfaction.

6. *"The Lean Startup"* by Eric Ries

This book provides insights into the startup world and how to build a successful startup. It covers various topics such as customer development, lean manufacturing, and agile development.

7. *"Intern Management for Dummies"* by Eric Woodard

This book provides practical advice on how to manage interns effectively. It covers topics such as recruiting interns, managing their work, and evaluating their performance.

8. *"Storytelling with Data"* by Cole Nussbaumer Knaflic

This book provides guidance on how to create effective data visualizations that tell a story. It covers various techniques for creating effective visualizations and how to communicate data effectively.

9. *"Data Science from Scratch"* by Joel Grus

This book provides a comprehensive introduction to data science. It covers various topics such as data cleaning, data manipulation, and data visualization using Python.

10. *"Data Visualization: A Practical Introduction"* by Kieran Healy

This book provides an introduction to data visualization and its various techniques. It covers topics such as bar charts, line charts, and scatter plots using libraries such as Matplotlib and Seaborn.