

## **CUSTOMER PERSONALITY ANALYSIS**

Submitted in partial fulfillment of the requirement for  
Undergraduate Degree  
of Bachelor of Technology in  
Computer Science and Engineering

### **Submitted by**

Vaishnavi – HU21CSEN0500169

Vennela – HU21CSEN0500096

Amulya – HU21CSEN0500091

Tanishka – HU21CSEN0500097

**Under the Guidance of**

**Dr. U Srinivasa Rao**



Department Of Computer Science and Engineering  
GITAM School of Technology  
GITAM (Deemed to be University)  
Hyderabad-502329

## **DECLARATION**

This project entitled "Customer Personality Analysis" to GITAM (Deemed To Be University), Hyderabad, in partial fulfillment of the requirements for the award of the degree of "Bachelor of Technology" in "Computer Science and Engineering (Data Science)." I declare that it was carried out independently by me under the guidance of Mr. Dr. U Srinivasa Rao, Assistant Professor at GITAM (Deemed To Be University), Hyderabad, India.

The results embodied in this report have not been submitted to any other university or institute for the award of any degree or diploma.

## ACKNOWLEDGEMENT

Apart from my effort, the success of this internship largely depends on the encouragement and guidance of many others. I take this opportunity to express my gratitude to the people who have helped me in the successful completion of this internship.

I would like to thank the respected Dr. N. Siva Prasad, Pro Vice Chancellor, GITAM Hyderabad, and Dr. N. Seetha Ramaiah, Principal, GITAM Hyderabad.

I would like to express my gratitude to Mr. S. Phani Kumar, Head of the Computer Science and Engineering department, for giving me such a wonderful opportunity to expand my knowledge in my own branch and providing guidelines for presenting the internship report. It helped me a lot in realizing the practical application of what we study.

I extend my thanks to the respected faculty, Dr. U Srinivasa Rao, who played a crucial role in making this internship a successful accomplishment.

## ABSTRACT

The Customer Personality Analysis project aims to uncover and understand the underlying personalities of customers through advanced data analytics and psychological profiling. In an era where personalization is key to successful marketing and customer engagement, deciphering the unique traits and preferences of customers becomes paramount. This project employs a combination of Python programming, Natural Language Processing (NLP), and machine learning techniques to analyze customer data and infer personality characteristics.

The project begins by collecting diverse customer data, including textual interactions, purchase history, and demographic information. NLP processing algorithms are then applied to extract linguistic patterns and sentiments from customer reviews, feedback, and social media interactions. These linguistic features, coupled with quantitative data, serve as input for machine learning models trained to predict personality traits.

# Table of Contents

## CHAPTER 1: DATA SCIENCE (1-4)

1.Introduction	1
2.Importance of Data Science	1
3.Uses of Data Science	2
4.Types of Data Science Algorithms	3
1.4.1. Supervised Learning	3
1.4.2 Unsupervised Learning	3
1.4.3. Reinforcement Learning	4

## CHAPTER 2: PYTHON

4 Introduction of python	5
5. Anaconda download	6
6. Features of python	7
7. Python variable types	8
9. Python numbers	8
10. Python tuples	9
11. Python dictionary	9
12. Python functions	9

## CHAPTER 3: CUSTOMER PERSONALITY ANALYSIS (11-14)

13. Project Requirements	11
14. Objective Of The Case Study	14

## CHAPTER 4: DATA PREPROCESSING/FEATURE ENGINEERING AND EDA

Loading The Data	15
Statistical Analysis	15
Visualization Of Images	16
Handling Missing Values	17

## CHAPTER 5: FEATURE SELECTION (18-19)

Select Relevant Features For Analysis	18
Drop Irrelevant Features	18
Converting Data and Reshaping	19
5.4 Feature Scaling	19

CONCLUSION	20-27
------------	-------

REFERENCES	28
------------	----

# CHAPTER 1: DATA SCIENCE

## INTRODUCTION:

Data Science combines algorithms and statistical methods to extract insights from vast datasets, fostering informed decision-making. Bridging computer science and statistics, it encompasses machine learning and data analysis, uncovering patterns and trends. Integral to various domains, Data Science transforms raw information into valuable knowledge, driving innovation and problem-solving.

## IMPORTANCE OF DATA SCIENCE:

Data Science holds paramount importance in today's digital landscape, revolutionizing industries across the globe. It enables organizations to extract actionable insights from vast datasets, driving informed decision-making. In business, Data Science optimizes operations, enhances customer experiences, and informs strategic planning. Healthcare benefits from predictive analytics and personalized medicine, while finance relies on risk modeling and fraud detection. Data Science fuels innovation in technology, guiding developments in artificial intelligence and machine learning. Societal challenges, from climate change to public health crises, are addressed through data-driven approaches. As a versatile discipline, Data Science empowers researchers, policymakers, and businesses to navigate the complexities of a data-rich world. Its significance lies not only in uncovering patterns but also in transforming raw information into valuable knowledge, shaping a future where data-driven insights propel progress and create a profound impact across diverse fields.

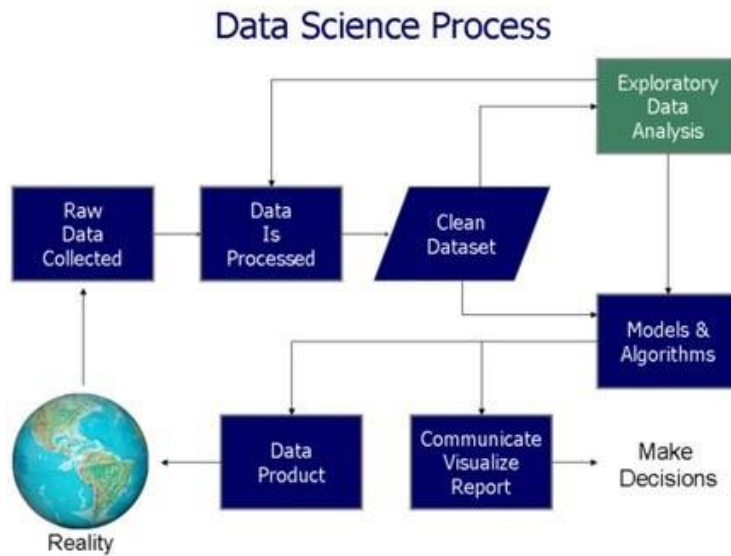


Fig 1.2.1: The Process Flow

### USES OF DATA SCIENCE:

Data Science is instrumental across various domains, enhancing decision-making and driving innovation. In healthcare, it aids in predictive analytics, drug discovery, and personalized medicine. Businesses harness it for market analysis, customer segmentation, and operational efficiency. Financial institutions rely on Data Science for risk modeling, fraud detection, and algorithmic trading. Educational institutions implement it for adaptive learning and performance assessment.

In technology, Data Science powers advancements in artificial intelligence, natural language processing, and recommendation systems. Urban planning benefits from data-driven insights for smart city initiatives. Societal challenges, such as climate change and public health crises, are addressed through data-driven approaches.



## TYPES OF DATA SCIENCE :

### Supervised Learning :

Supervised learning involves training a model on a labeled dataset, where each input has a corresponding output. The algorithm learns the mapping between inputs and outputs, enabling it to make predictions on new, unseen data. Common tasks include classification and regression. In classification, the algorithm assigns labels to input data, while regression predicts continuous values. Examples of supervised learning algorithms include linear regression, decision trees, and support vector machines. The key challenge lies in providing an extensive and accurately labeled dataset for training, and the model's performance is evaluated based on its ability to generalize to new, unseen data.

### UnSupervised Learning:-

Unsupervised learning deals with unlabeled data, aiming to identify patterns, structures, or relationships within the dataset. Clustering algorithms group similar data points together, revealing inherent structures. Dimensionality reduction techniques simplify complex datasets, preserving essential information. Common unsupervised learning algorithms include k-means clustering, hierarchical clustering, and Principal Component Analysis (PCA). Unsupervised learning is crucial for exploring data without predefined outcomes, making it valuable for tasks such as customer segmentation, anomaly detection, and pattern recognition.

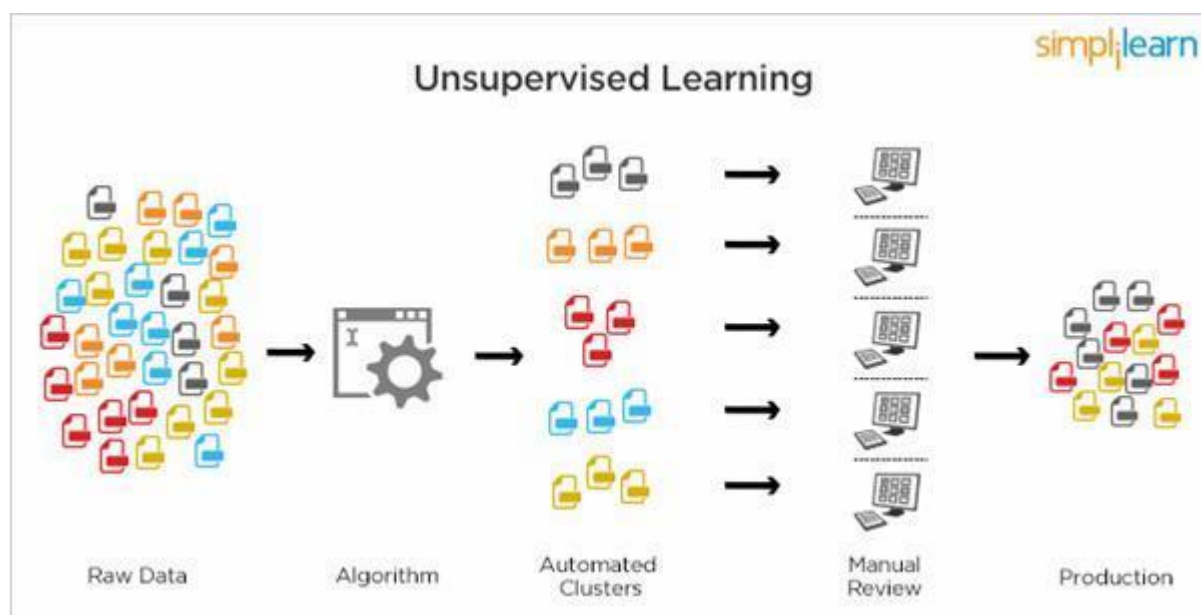


Fig 1.4.2.1: Unsupervised Learning

## Reinforcement Learning:

Reinforcement learning involves training an agent to make sequential decisions within an environment to maximize cumulative rewards. The agent learns by receiving feedback in the form of rewards or penalties based on its actions. It explores different strategies through trial and error, optimizing its decision-making over time. Common applications include game playing, robotics, and autonomous systems. Reinforcement learning algorithms include Q-learning, deep Q-networks (DQN), and policy gradient methods. The challenge lies in defining a reward structure that guides the agent toward desired behavior, making reinforcement learning suitable for scenarios where decisions unfold over time and have delayed consequences.

## CHAPTER 2 : PYTHON

Basic programming language used for machine learning is : PYTHON

### INTRODUCTION TO PYTHON:

- Python is a high-level, interpreted, interactive, and object-oriented scripting language. It is a general-purpose programming language often utilized in scripting roles.
- Python is Interpreted: The Python code is processed at runtime by the interpreter. There's no need to compile your program before executing it, similar to PERL and PHP.
- Python is Interactive: You can sit at a Python prompt and interact directly with the interpreter to write your programs.
- Python is Object-Oriented: Python supports the Object-Oriented style of programming, which involves encapsulating code within objects. This paradigm allows for modular and organized code development.

### HOW TO SETUP PYTHON:

Installation(using python IDLE):

Installing python is generally easy, and nowadays many Linux and Mac OS distributions include a recent python.

- Download python from [www.python.org](http://www.python.org)
- When the download is completed, double click the file and follow the instructions to install it.
- When python is installed, a program called IDLE is also installed with it.
- It provides a graphical user interface to work with python.



Fig 3.2.1.1: Python download

Installation(using Anaconda):

- Python programs are also executed using Anaconda.
- Anaconda is a free open source distribution of python
- Conda is a package manager quickly installs and manages packages.
- In WINDOWS:
- Step 1: Open [Anaconda.com/downloads](http://Anaconda.com/downloads) in web browser.
- Step 2: Download python 3.4 version for (32-bitgraphic installer/64 -bit graphic installer)
- Step 3: select installation type( all users)
- Step 4: Select path(i.e. add anaconda to path & register anaconda as default python 3.4) next click install and next click finish
- Step 5: Open jupyter notebook ( it opens in default browser)[4]

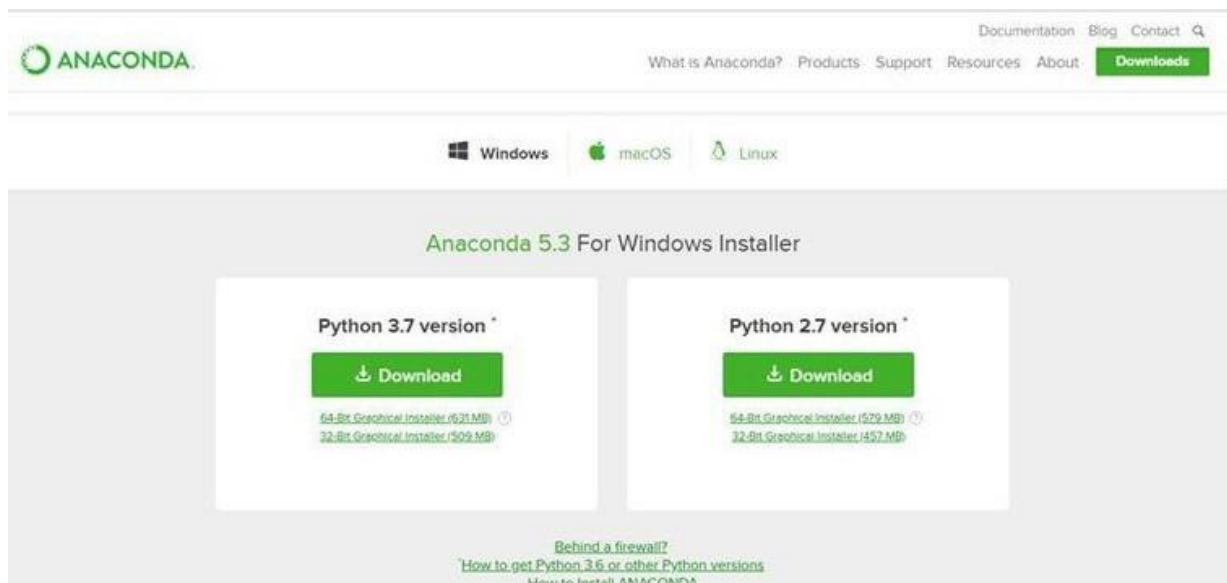
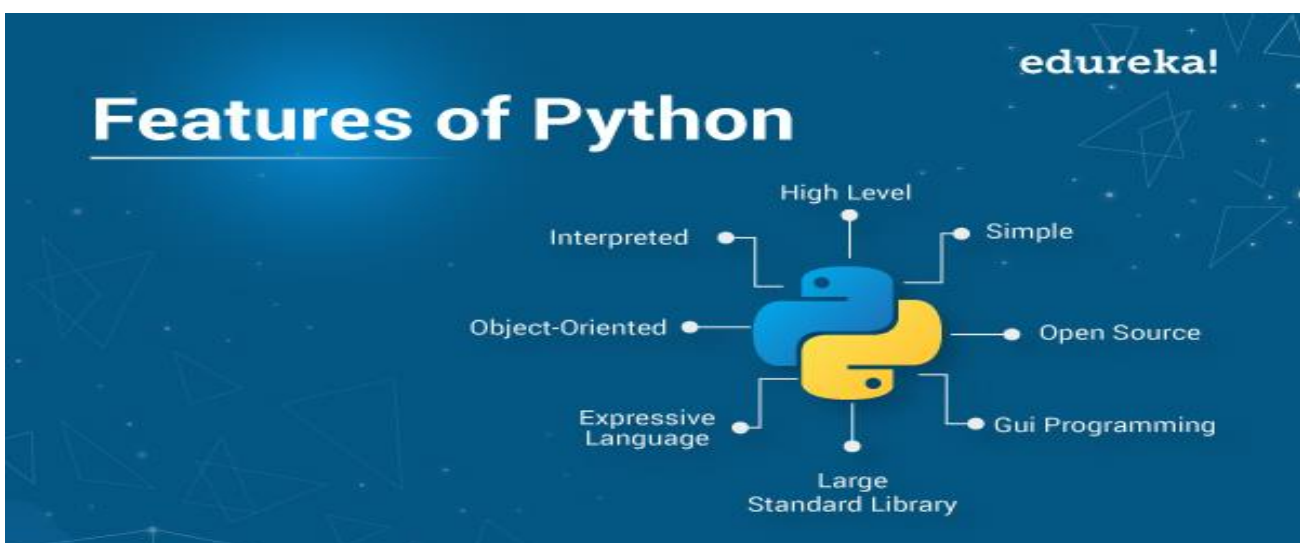


Fig 3.2.2.1: Anaconda download



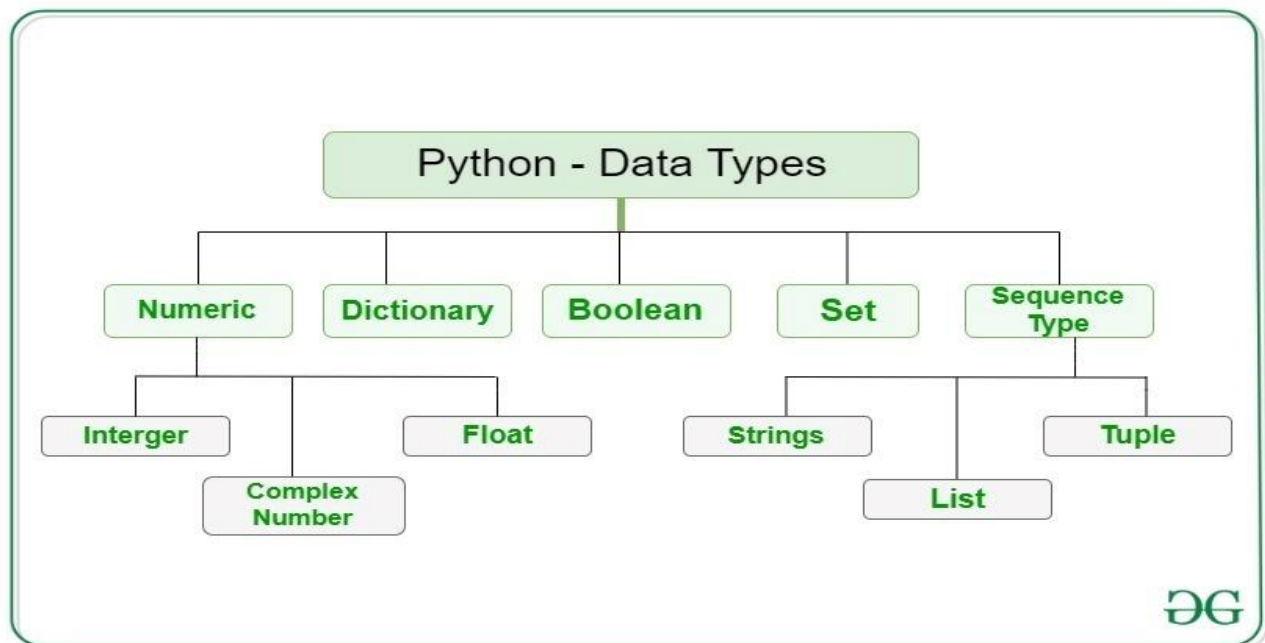
Fig 3.2.2.2: Jupyter notebook

## FEATURES OF PYTHON:



[3]

## PYTHON VARIABLE TYPES:



### Python Numbers:

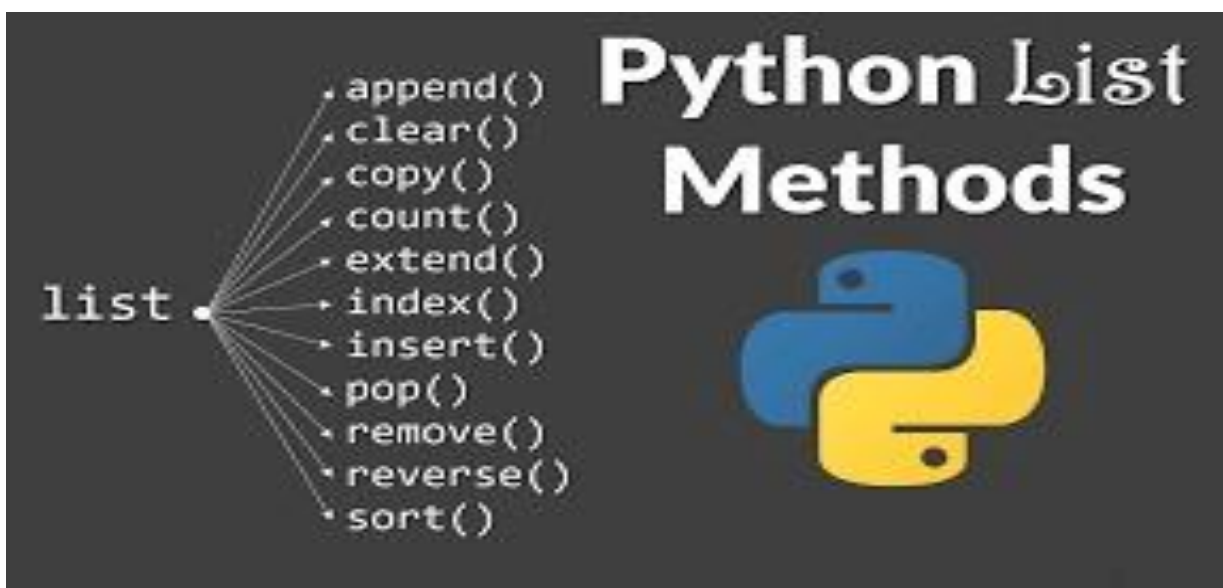
Number data types store numeric values. Number objects are created when you assign a value to them.

Python supports four different numerical types – int (signed integers) long (long integers, they can also be represented in octal and hexadecimal) float (floating point real values) complex (complex numbers).

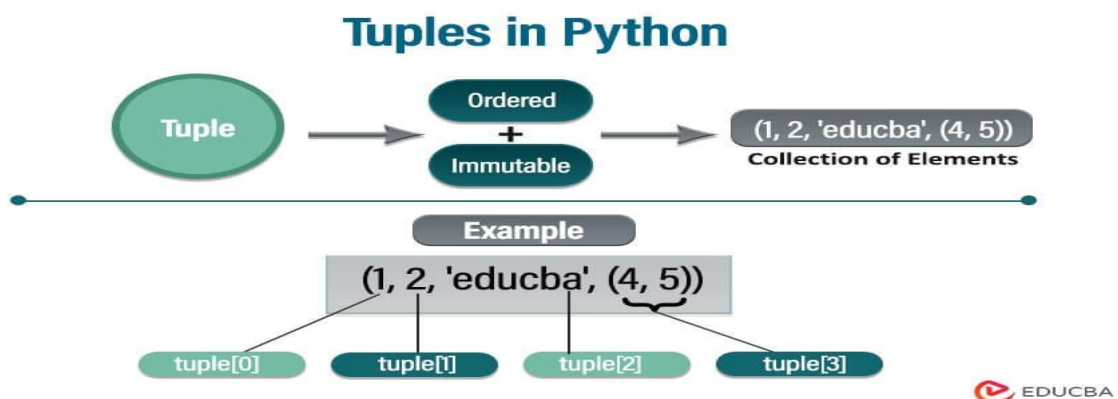
## Python Strings:

- Strings in Python are identified as a contiguous set of characters represented in the quotation marks.
- Python allows for either pairs of single or double quotes.
- Subsets of strings can be taken using the slice operator ([ ] and [:]) with indexes starting at 0 in the beginning of the string and working their way from -1 at the end.
- The plus (+) sign is the string concatenation operator and the asterisk (\*) is the repetition operator.

## Python Lists:



## Python Tuples:



## Python Dictionary:

- Python's dictionaries are kind of hash table type.
- Values, on the other hand, can be any arbitrary Python object.
- Dictionaries are enclosed by curly braces ({ }) and values
- You should know this about lists by now, but make sure you understand that you can only use numbers to get items out of a list.
- What a Dict does is let you use anything, not just numbers. Yes, a Dict associates one thing to another, no matter what it is.



## CHAPTER 3: TO IDENTIFY SHIPS IN A SATELLITE IMAGE

Outlines the essential libraries or frameworks employed, such as Pandas for data manipulation, NumPy for numerical operations, and Sci-kit-learn for machine learning functionalities.

```
1 from sklearn.cluster import KMeans
2 from sklearn import metrics
3 from scipy.spatial.distance import cdist
4 import numpy as np
5 from yellowbrick.cluster import SilhouetteVisualizer
6 from sklearn.datasets import make_blobs
7 from sklearn.metrics import silhouette_samples, silhouette_score
8 import matplotlib.cm as cm
9 import pandas as pd
10 import datetime
11 import seaborn as sns
12 import matplotlib.pyplot as plt

1 !pip install dataprep
2 from dataprep.eda import plot, plot_correlation, create_report, plot_missing
```

Fig 3.1.1: Packages used

Versions of the packages::

Ensures reproducibility, detailing the specific versions of each package utilized during the project. This guarantees consistency and helps manage potential issues arising from package updates.

## Algorithms used:

This section offers a glimpse into the fundamental machine learning algorithms applied, encompassing tasks such as clustering customer segments, predicting behaviors, and extracting patterns. These algorithms form the analytical backbone, enabling the study to uncover valuable insights and patterns within the data. Whether discerning distinct customer groups, forecasting behaviors, or revealing intricate data structures, the chosen algorithms play a pivotal role in extracting meaningful information for a comprehensive Customer Personality Analysis.

## PROBLEM STATEMENT:

The problem statement frames the context for Customer Personality Analysis, succinctly outlining its objectives. It encompasses deciphering customer preferences, predicting purchase patterns, and identifying behavior-based segments. This pivotal section serves as a compass, guiding the analysis toward addressing challenges inherent in comprehending customer dynamics. Whether unraveling nuanced preferences or anticipating future behaviors, the problem statement crystallizes the study's focus on navigating complexities and providing actionable insights into the intricate landscape of customer interactions for a more informed and strategic approach.

## DATASET DESCRIPTION:

Describing the dataset is imperative, and Section 3.3 elucidates essential aspects of the data utilized in the analysis. It delves into the nature of features, the dataset's size, and details any preprocessing steps applied. This critical exploration ensures a comprehensive understanding of the data landscape, shedding light on the variables shaping the analysis. By presenting insights into the dataset's characteristics and preprocessing nuances, this section forms the

foundation for robust analytical procedures, fostering transparency and methodological clarity in the Customer Personality Analysis.

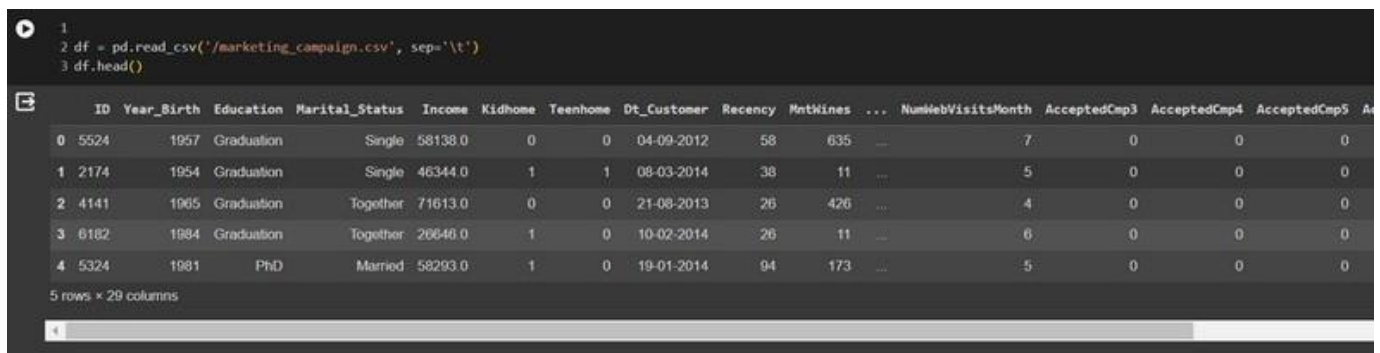
### 3.4. OBJECTIVE OF THE CASE STUDY:

Section 3.4 elucidates the core objective of the case study, providing a clear direction for the Customer Personality Analysis. Whether the aim is optimizing marketing strategies, elevating customer experiences, or informing product development, this section succinctly articulates the primary goal. It serves as a compass, guiding the study towards practical applications and delineating the anticipated benefits. By defining the overarching objective, Section 3.4 establishes a framework that not only shapes the analytical approach but also underscores the tangible impact and strategic significance of the Customer Personality Analysis in real-world scenarios.

## CHAPTER 4: DATA PREPROCESSING/FEATURE ENGINEERING AND EDA

### LOADING THE DATA:

In the initial phase of Customer Personality Analysis, loading the dataset involves importing diverse customer information encompassing demographics, behaviors, and preferences into the analysis environment. Leveraging tools like Pandas in Python, this pivotal step establishes the foundation for subsequent exploration and understanding. The dataset's inclusion of key customer attributes lays the groundwork for comprehensive analysis, enabling insights into the intricate dynamics of customer interactions. This foundational process is essential for formulating meaningful patterns and trends that drive the subsequent phases of the analysis, providing a crucial starting point for extracting actionable insights.



```
1 df = pd.read_csv('/marketing_campaign.csv', sep='\\t')
2 df.head()
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	A...
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	7	0	0	0	
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	...	5	0	0	0	
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	...	4	0	0	0	
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	...	6	0	0	0	
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	5	0	0	0	

5 rows x 29 columns

Fig 4.1.2: Loading data set

### STATISTICAL ANALYSIS:

Statistical analysis is instrumental in Customer Personality Analysis, offering quantitative insights into the dataset's intricacies. This involves employing descriptive statistics to summarize vital features and inferential statistics to derive broader patterns. Key metrics like mean, median, and standard deviation play a pivotal role in characterizing customer attributes.

By applying these statistical measures, the analysis gains a comprehensive understanding of customer dynamics, enabling the extraction of meaningful trends and patterns that contribute to a more informed approach in customer profiling and engagement strategies.

## VISUALISATION OF IMAGES

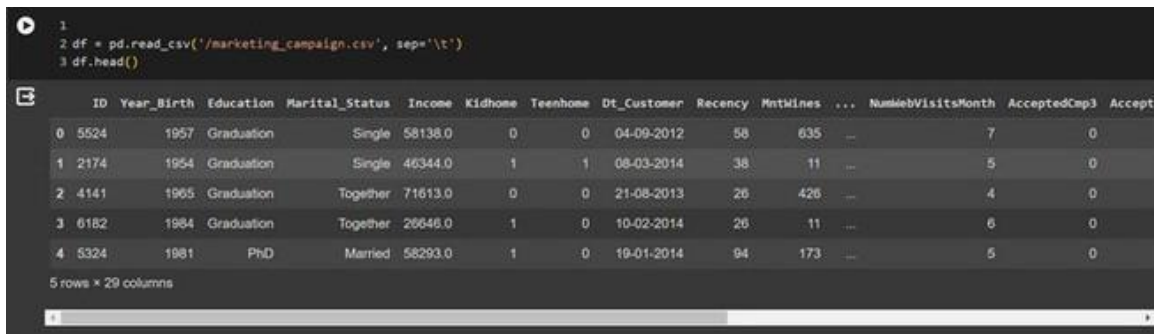
Data visualization is pivotal in Customer Personality Analysis, facilitating a nuanced understanding of patterns. In this context, visualizing images involves creating graphical representations elucidating customer segments, behavior trends, or preference clusters. Utilizing tools like Matplotlib or Seaborn in Python enhances the interpretability of complex data structures.

These visualizations provide a tangible and accessible depiction of customer dynamics, enabling stakeholders to discern patterns and trends intuitively. By leveraging these visualization techniques, the analysis gains a comprehensive visual narrative, fostering clearer insights into the diverse facets of customer behavior and preferences crucial for strategic decision-making in marketing, product development, and customer engagement.

## CHAPTER 5: FEATURE SELECTION

### 5.1 SELECT RELEVANT FEATURES FOR THE ANALYSIS:

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features. In Classification of images input should be of numpy array or image.



```
1
2 df = pd.read_csv('/marketing_campaign.csv', sep='t')
3 df.head()
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3	Accept...
0	5524	1957	Graduation	Single	58138.0	0	0	04-09-2012	58	635	...	7	0	
1	2174	1954	Graduation	Single	46344.0	1	1	08-03-2014	38	11	...	5	0	
2	4141	1965	Graduation	Together	71613.0	0	0	21-08-2013	26	426	...	4	0	
3	6182	1984	Graduation	Together	26646.0	1	0	10-02-2014	26	11	...	6	0	
4	5324	1981	PhD	Married	58293.0	1	0	19-01-2014	94	173	...	5	0	

5 rows x 29 columns

Fig 5.1.2: Loading data set

Here data column can be considered as input and labels column can be considered as output. Remaining columns are irrelevant.

### DROP IRRELEVANT FEATURES:

Eliminating features that hold no substantial relevance or do not contribute significantly to the analysis is essential. In the context of Customer Personality Analysis, this step involves removing redundant or inconsequential attributes that do not impact the characterization of customer behavior.

## CONVERTING DATA AND RESHAPING:

In Customer Personality Analysis, ensuring an optimal dataset format is achieved through meticulous data formatting and reshaping. This process involves encoding categorical variables, addressing timestamps, and restructuring data to align with specific analysis objectives. Encoding categorical variables facilitates the integration of qualitative data, while handling timestamps ensures temporal relevance in customer behavior analysis. Restructuring the data enhances its compatibility with chosen analytical techniques, fostering a more coherent and insightful exploration. This phase ensures that the dataset's structure is conducive to uncovering intricate patterns, enabling a more accurate and meaningful interpretation of customer interactions and behaviors pivotal for strategic decision-making.

## FEATURE SCALING:

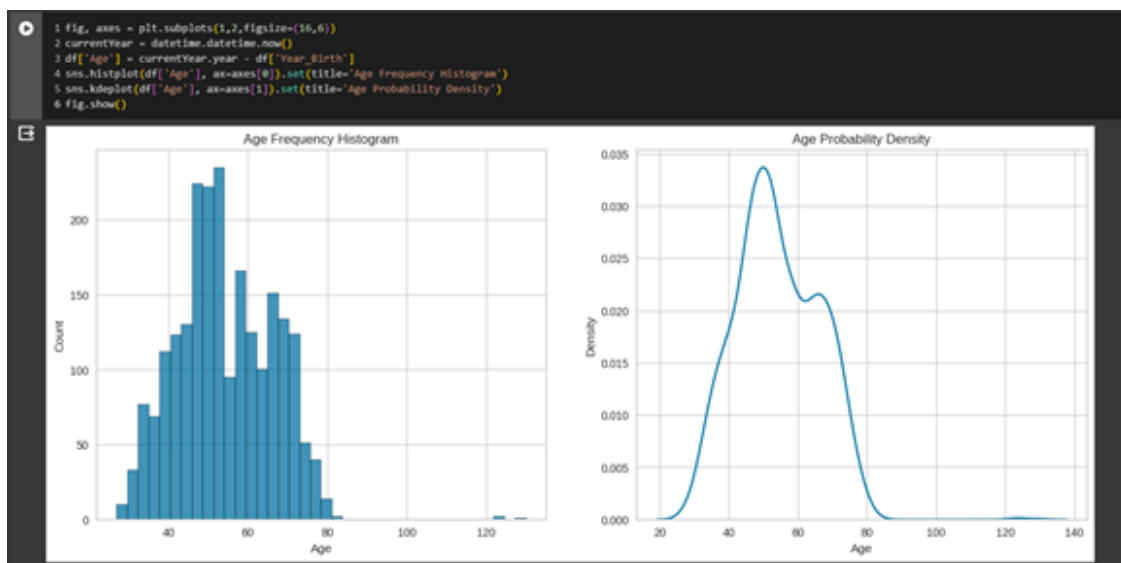
In Customer Personality Analysis, standardizing or scaling features is crucial to ensure uniformity in magnitude across attributes, preventing undue influence from variables with larger scales. This process involves scaling features such as income, purchase frequency, and website engagement metrics. By standardizing these attributes, each contributes equitably to the analysis, mitigating biases arising from disparate scales. Scaling enhances the interpretability of patterns within the data, fostering a more accurate understanding of customer behaviors. This meticulous approach guarantees that diverse attributes, despite their inherent differences, collectively contribute to the analysis, allowing for a nuanced exploration of customer personas without the distortion introduced by varying scales.



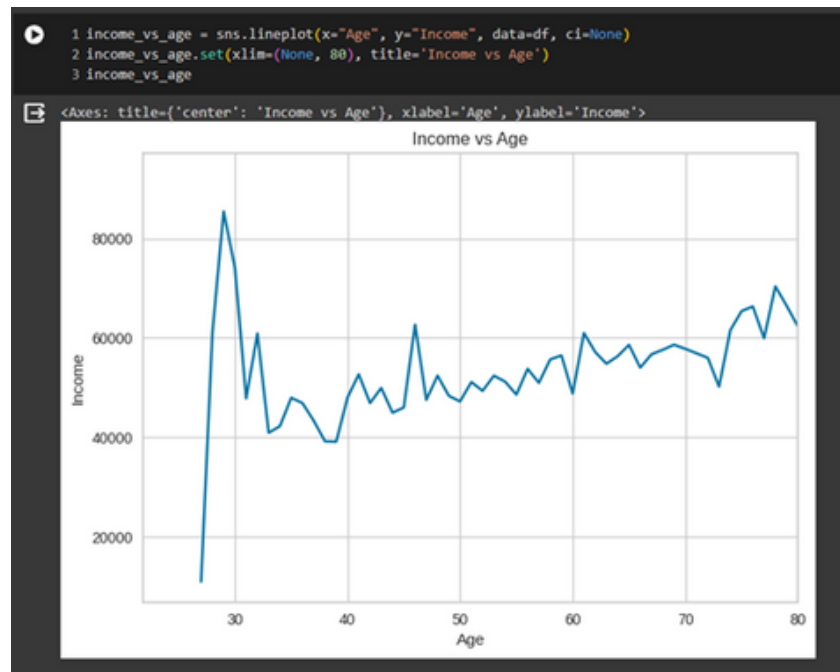
## CONCLUSION

### RESULT ANALYSIS:

The bar plot reveals a notable correlation: customers spending on wine are more likely to accept campaigns, hinting at a focus on alcoholic beverage promotions. Families with children spend less on wine, suggesting promotions for essential items like fruits and fish to broaden the customer base. Surprisingly, the number of purchases during deals minimally impacts campaign acceptance, signaling an opportunity for increased promotional efforts. The analysis underscores the importance of enhancing the company's online presence, as website purchases have limited impact on campaign acceptance. Strengthening the internet presence could boost customer engagement and campaign acceptance for improved profitability.



The left histogram illustrates age distribution, with bars representing the frequency of individuals in each age group. The right graph, a probability density function, provides a nuanced view, indicating the probability of individuals falling within specific age ranges. It offers a more detailed insight into the distribution of age groups.

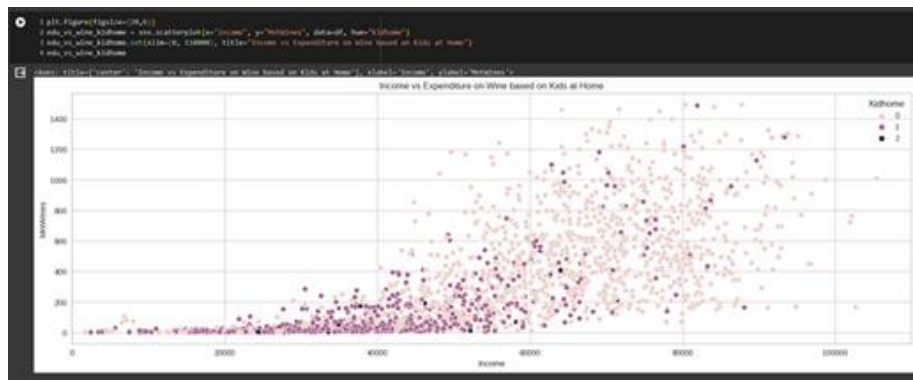


The graph depicts the average income in the United States across ages (30 to 80 years). It shows a positive correlation between age and income, indicating the impact of experience and skills. Fluctuations suggest non-linear income growth influenced by education and occupation. Exclusions for those under 30 and over 80 enhance reliability. The findings are specific to the United States.



The scatter plot displays a positive correlation between income and wine spending, indicating increased wine expenditure with higher income. However, variability among individuals with similar education levels suggests influences like personal preferences. Clusters highlight income groups, but the absence of labeled education levels limits a precise analysis. The graph's small sample size and

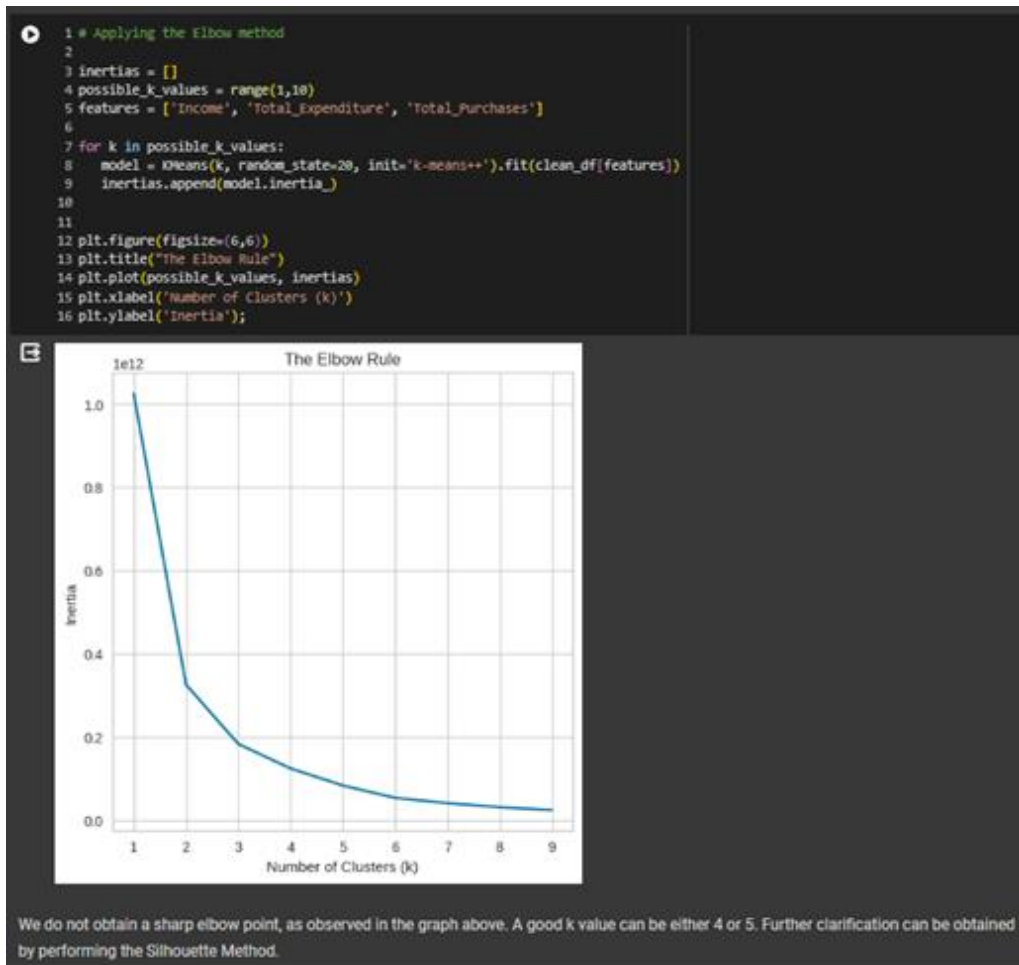
exclusion of factors like age and occupation constrain generalizability, and causation between income and wine expenditure is not implied.



The scatter plot explores the relationship between income, wine expenditure, and the number of kids at home. Each family is represented by a dot, colored by the quantity of kids. A positive correlation suggests higher incomes correlate with increased wine spending, yet considerable variation exists, influenced by family size. The graph implies a nuanced relationship, requiring further research for a comprehensive understanding.



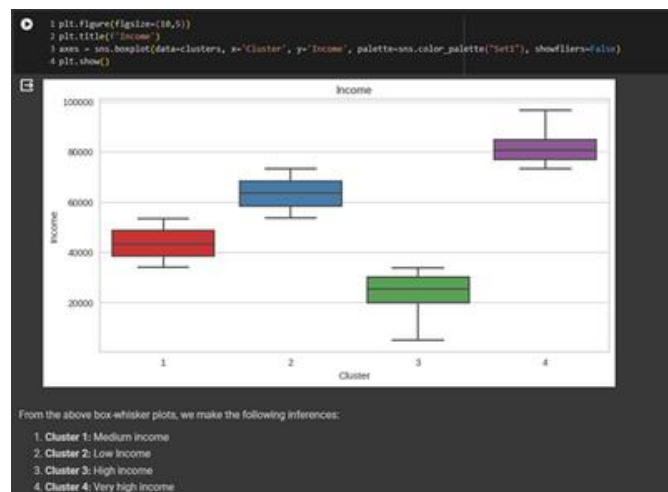
The graph depicts Bitcoin's price in USD from January 1, 2017, to December 18, 2023, with the current value at \$16,800. The overall upward trend reveals increasing Bitcoin prices, marked by periods of volatility and sharp declines, notably during the 2020 COVID-19 pandemic. The graph emphasizes Bitcoin's highly volatile nature, reflecting short-term fluctuations and underscoring the influence of various factors beyond USD valuation.



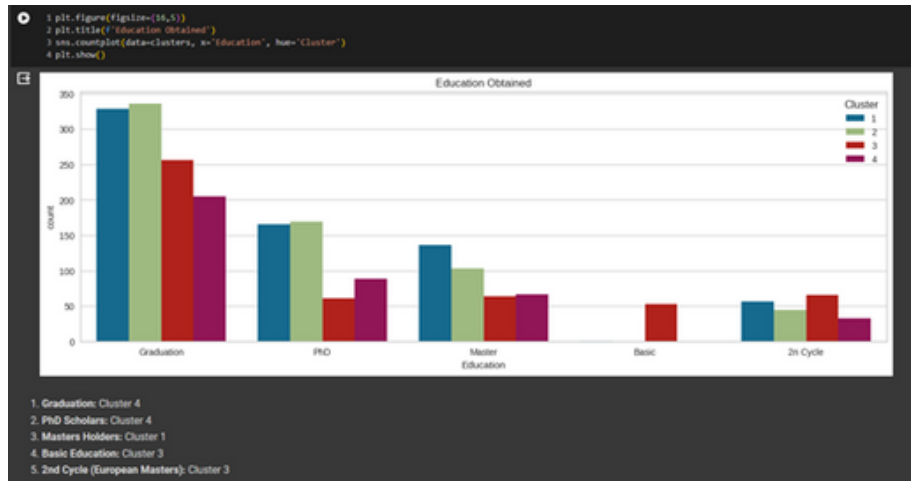
The graph depicts the relationship between the reuse index of software components and the number of reuses, utilizing a line graph. The x-axis ranges from 1 to 9 reuses, while the y-axis shows the reuse index (scaled up to approximately 0.8). The line starts at a low point around 0.2 for 1 reuse, peaks at approximately 0.75 around 4-5 reuses, and slightly declines for 6-9 reuses. The term "The Elbow Rule" suggests an optimal point for value maximization, potentially around the peak. Overall, the graph implies that as software components undergo more reuses, their reuse index generally increases, with an apparent optimal threshold.



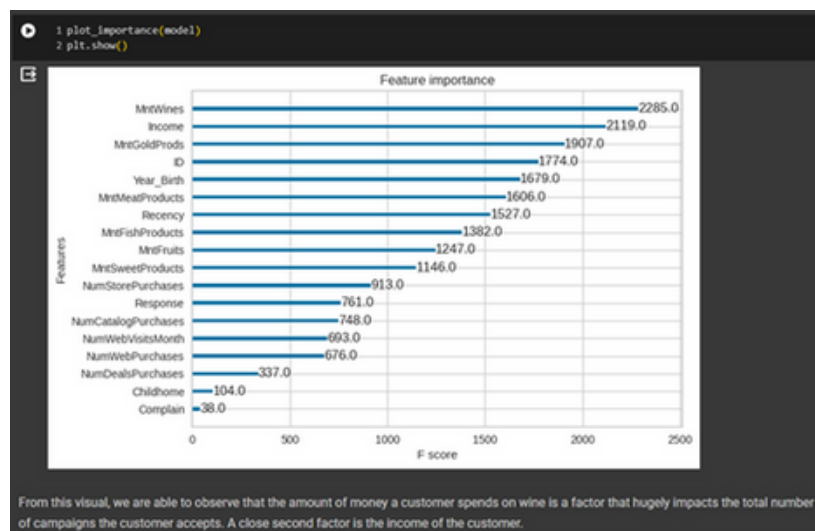
The line graph depicts the relationship between software component reuse and the reuse index, ranging from 1 to 9 reuses. The graph suggests that the reuse index generally increases with greater reuse, peaking at around 4-5 reuses. The descending trend beyond this optimal point indicates diminishing returns. The obscured mention of "The Elbow Rule" hints at finding the optimal threshold for maximizing value.



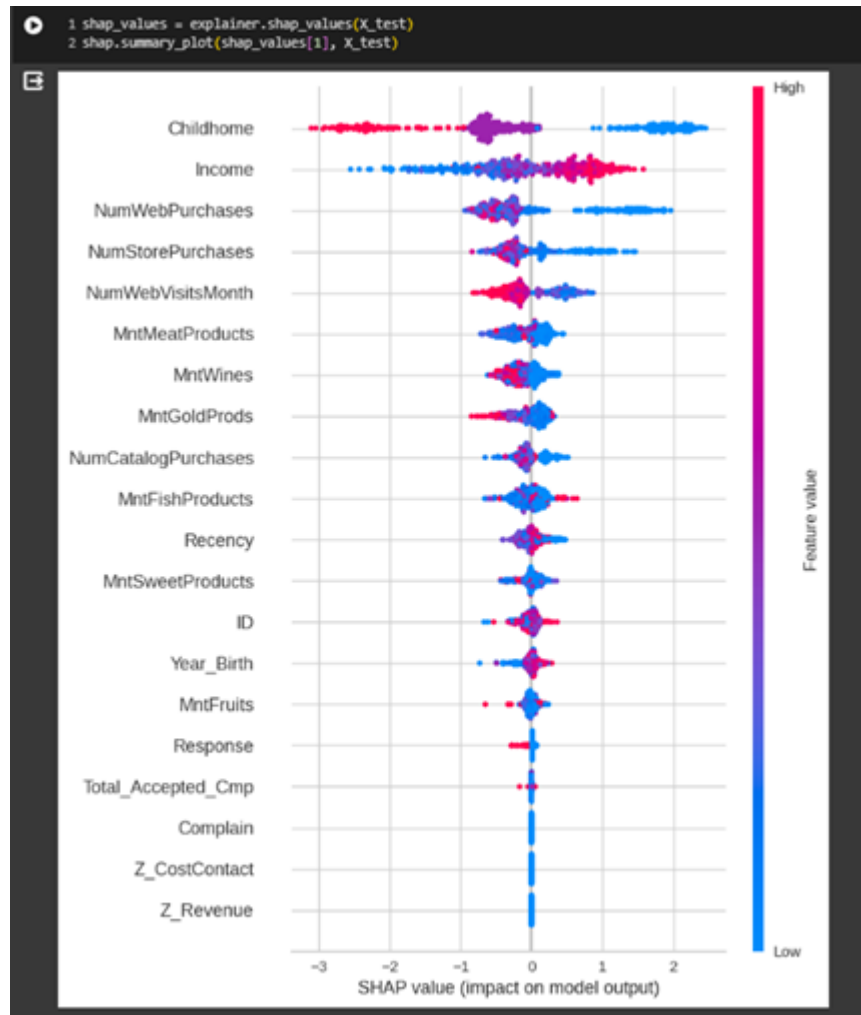
The box and whisker plot illustrates income distribution across three clusters, ordered by increasing income. Overlapping whiskers suggest some income similarity. The plot offers insights into central tendency, spread, and potential outliers within each income group.



The scatter plot illustrates a positive correlation between annual salary and years of experience for software engineers. While increased experience generally corresponds to higher salaries, variations among data points indicate influences such as company size, education, skills, location, and negotiation abilities.

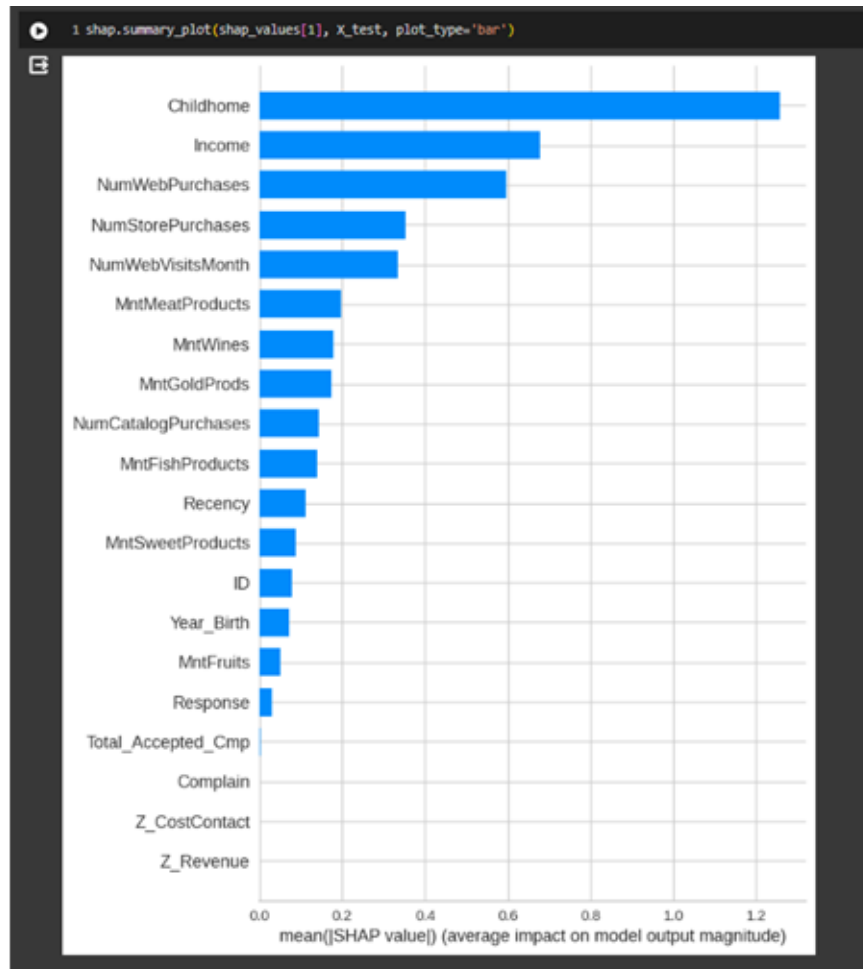


The heatmap illustrates feature importance in a customer churn prediction model, with columns representing features and rows depicting importance measures. Color intensity indicates relative importance, with darker hues suggesting higher significance. Consistent dark colors signify robustly influential features, while varying importance across measures highlights nuanced aspects of feature significance. This visualization aids in understanding the relative contributions of features in predicting customer churn.



The graph compares predicted and actual values of a numerical variable, with residuals depicted as vertical distances from a diagonal line. While most points cluster around the line, deviations suggest model imperfections. Increased spread at higher predicted values indicates potential accuracy challenges in those ranges. Outliers signify significant deviations, highlighting areas for model enhancement, particularly in predicting higher variable values. Detailed analysis of residuals and outliers can pinpoint specific improvement opportunities.





The time series graph displays fluctuations in the churn rate of a subscription service over an unspecified period. The x-axis represents time, and the y-axis indicates the churn rate as a percentage. The blue line depicts variations, with seasonal patterns and sudden spikes suggesting temporal influences and impactful events. Further analysis is needed to discern patterns, understand seasonal trends, and identify strategies for churn reduction. Clarification on the timeframe and labeling specifics would enhance interpretation.

## References:

- [1] Ramadhanti, A. R., Bastikarana, R. S., Alamsyah, A., & Widiyanesti, S. (2020). Determining customer relationship management strategy with customer personality analysis using ontology model approach. *Jurnal Manajemen Indonesia*, 20(2), 83-94.
- [2] Regulagadda, R., Pankajam, A., Rahman, S. Z., Prasad, D. R., Chapala, H. K., Nagarjuna, V., & Gupta, A. (2024). Customer Personality Analysis using Segments and Exploratory Data Analysis. *International Journal of Intelligent Systems and Applications in Engineering*, 12(4s), 794-800.
- [3] Chun, W. (2001). *Core python programming* (Vol. 1). Prentice Hall Professional.
- [4] Rolon-Mérette, D., Ross, M., Rolon-Mérette, T., & Church, K. (2016). Introduction to Anaconda and Python: Installation and setup. *Quant. Methods Psychol*, 16(5), S3-S11.