**FROM DATA TO DECISIONS**

**PROJECT REPORT**

(SUMMER TRAINING COURSE 2025-2026)

Submitted by

Arpit Kumar

(Jd\_data Group)

Registration No: 12312286

Programme and Section: BTech CSE

Summer Training Course (From Data to Decision)

Under the Guidance of

**Sandeep Kaur**

**Discipline of CSE/IT**

**Lovely School of Computer Science And Engineering**

**Lovely Professional University, Phagwara**



**CERTIFICATE**

This is to certify that Jd\_data Members has successfully completed the online summer training course and assessmentunder my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Sandeep Kaur**

**School of Computer Science And Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 14-07-2025

**DECLARATION**

I, Arpit Kumar, student of Data Science under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my group (Jd\_data) intensive work and is genuine.

Date: 14-07-2025 Signature

Registration No. 12312286 Arpit

**Acknowledgement**

I would like to express my sincere gratitude to Prof. Sandeep Kaur, our esteemed faculty for the subject From Data to Decision for his valuable guidance, support, and encouragement throughout the course.

His insightful lectures and hands-on approach to teaching have helped me develop a strong foundation in data science tools and their practical applications. The knowledge and skills gained through this subject have significantly contributed to my academic and professional growth.

I am also thankful to my peers, friends, and everyone who contributed directly or indirectly to my learning during this course.

Student Name: Arpit Kumar

**Table of Contents**

**1Introduction…………………………………………………….**

* + Project Overview
  + Objectives

1. **Dataset Description……………………………………………**
   * Source of Data
   * Data Attributes
2. **Data Cleaning………………………………………………….**
   * Handling Missing Values
   * Removing Duplicates
   * Data Type Conversion
   * Outlier Detection and Treatment
   * Encoding Categorical Variables
3. **Exploratory Data Analysis (EDA)…………………………….**
   * Summary Statistics
   * Data Visualization (Histograms, Boxplots, Scatterplots)
   * Correlation Analysis
4. **Machine Learning Models…………………………………….**
   * Linear Regression
     + Model Building
     + Performance Metrics
   * Random Forest
     + Model Building
     + Feature Importance
   * K-Nearest Neighbours (KNN)
     + Model Building
     + Parameter Selection
5. **Model Evaluation…………………………………………….**
   * Confusion Matrix
   * Accuracy
6. **Power BI Dashboard**
   * Dashboard Design and Features
   * Key Insights Visualization

**Introduction**

In an increasingly interconnected world, understanding human social behaviour and personality traits is crucial for various applications. This data science project explores a unique dataset capturing aspects of individuals' social interactions, habits, and self-perceived personality types (Introvert/Extrovert). The aim is to uncover patterns and relationships within this rich information.

This report systematically outlines our project's journey. It begins with this **Introduction** and a **Project Overview**, detailing our core **Objectives**. We then describe the **Dataset** and its **Attributes**, followed by the critical **Data Cleaning** phase, addressing missing values, duplicates, data types, outliers, and categorical encoding.

**Exploratory Data Analysis (EDA)** will present key statistics, visualizations, and correlation insights. Subsequently, we will implement **Machine Learning Models** including Linear Regression, Random Forest, and K-Nearest Neighbours (KNN), discussing their building, evaluation, and comparison. The project culminates in a **Power BI Dashboard** visualizing key findings, and a **Conclusion** summarizing insights and suggesting **Future Work**. This structured approach provides a comprehensive analysis of social behaviour and personality data.

**Dataset Description**

The dataset central to this project is a unique collection of attributes designed to explore various facets of individual social behaviour, interactions, and self-perceived personality traits. While the specific original source is not explicitly provided, it appears to be a compiled or simulated dataset for analytical exploration in social science or behavioural studies, presented in an Excel spreadsheet format.

This dataset comprises eight distinct attributes, with each row representing an individual's data points related to their social habits and personality. The attributes are:

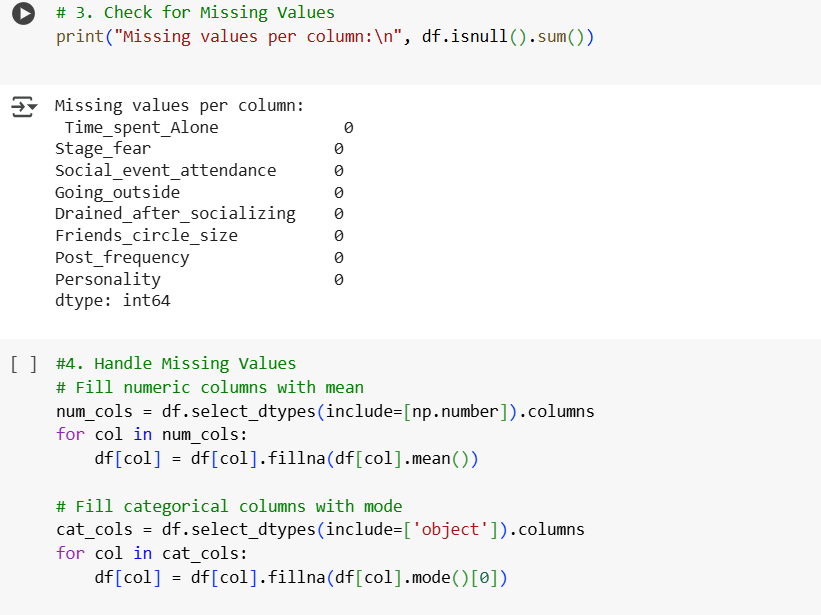
1. **Time\_spent\_Alone:** A numerical measure indicating the duration an individual spends in solitude, likely on a defined scale.
2. **Stage\_fear:** A binary categorical attribute denoting the presence ('Yes') or absence ('No') of stage fear.
3. **Social\_event\_attendance:** A numerical attribute quantifying the frequency or level of participation in social events.
4. **Going\_outside:** A numerical attribute reflecting how often an individual leaves their home environment for activities.
5. **Drained\_After\_socializing:** A crucial binary categorical attribute ('Yes' or 'No') indicating whether social interactions lead to emotional exhaustion, a key differentiator between personality types.
6. **Friends\_circle\_size:** A numerical attribute representing the number of close friends or acquaintances.
7. **Post\_frequency:** A numerical attribute measuring the frequency of online posts or public digital activity.
8. **Personality:** The primary categorical target variable, classifying individuals as either 'Introvert' or 'Extrovert'.

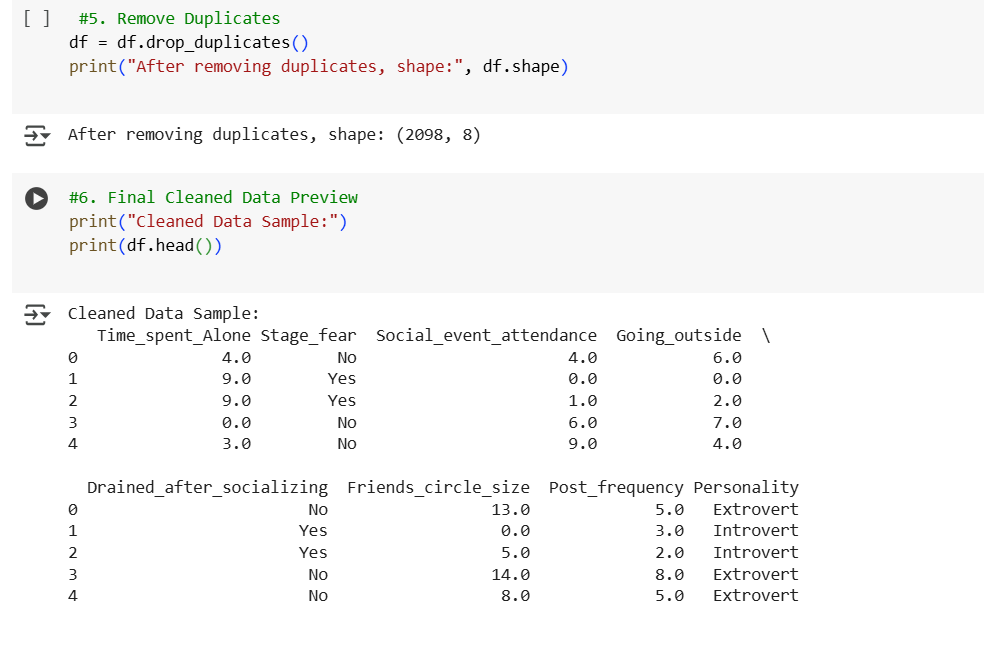
This comprehensive dataset allows for robust quantitative analysis, enabling the project to uncover meaningful insights and potential correlations between social behaviours and personality types.

**Data Cleaning**

Data cleaning is a critical preparatory phase in any data science project, ensuring the dataset's quality, consistency, and suitability for analysis and machine learning model training. This process involves addressing various imperfections that could otherwise lead to biased results, inaccurate models, or computational inefficiencies. Our data cleaning methodology systematically tackles missing values, duplicate entries, incorrect data types, outliers, and the need to encode categorical variables.

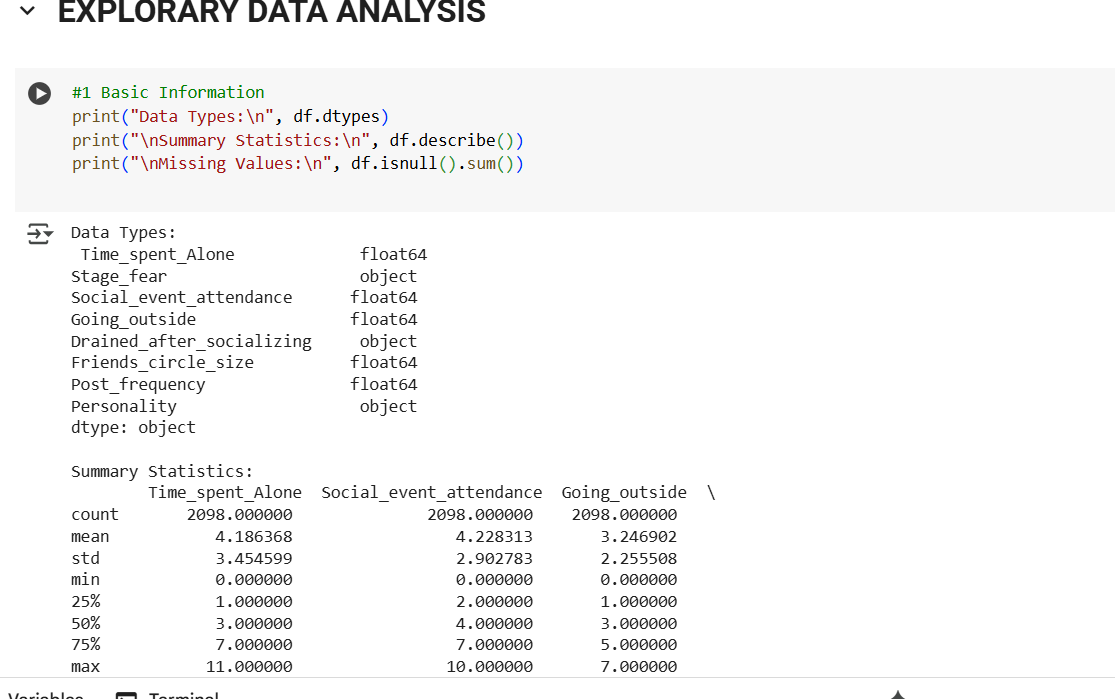


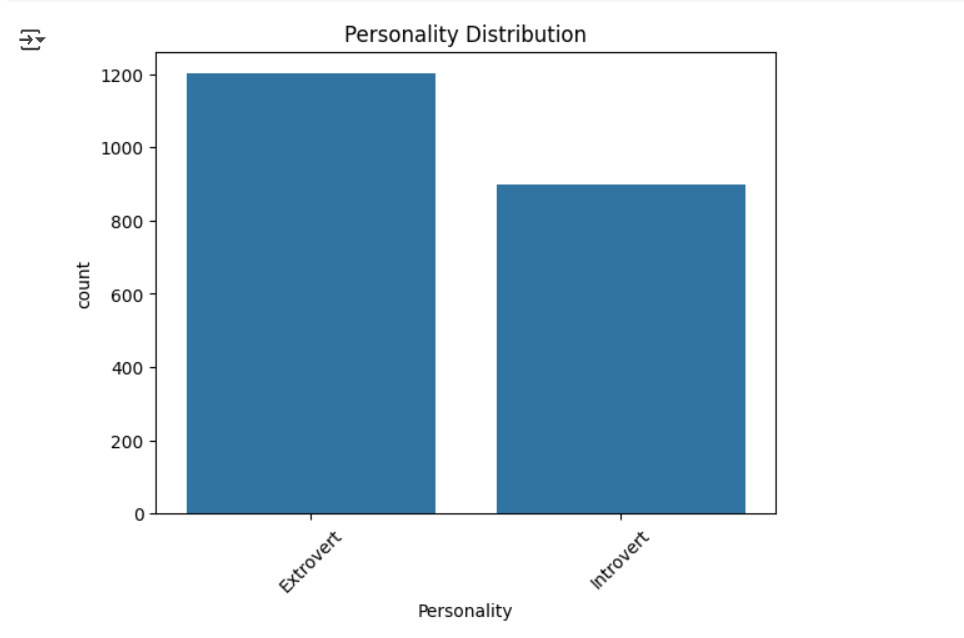
****

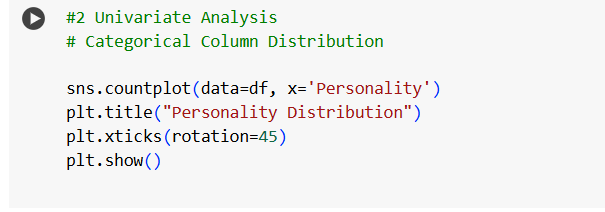


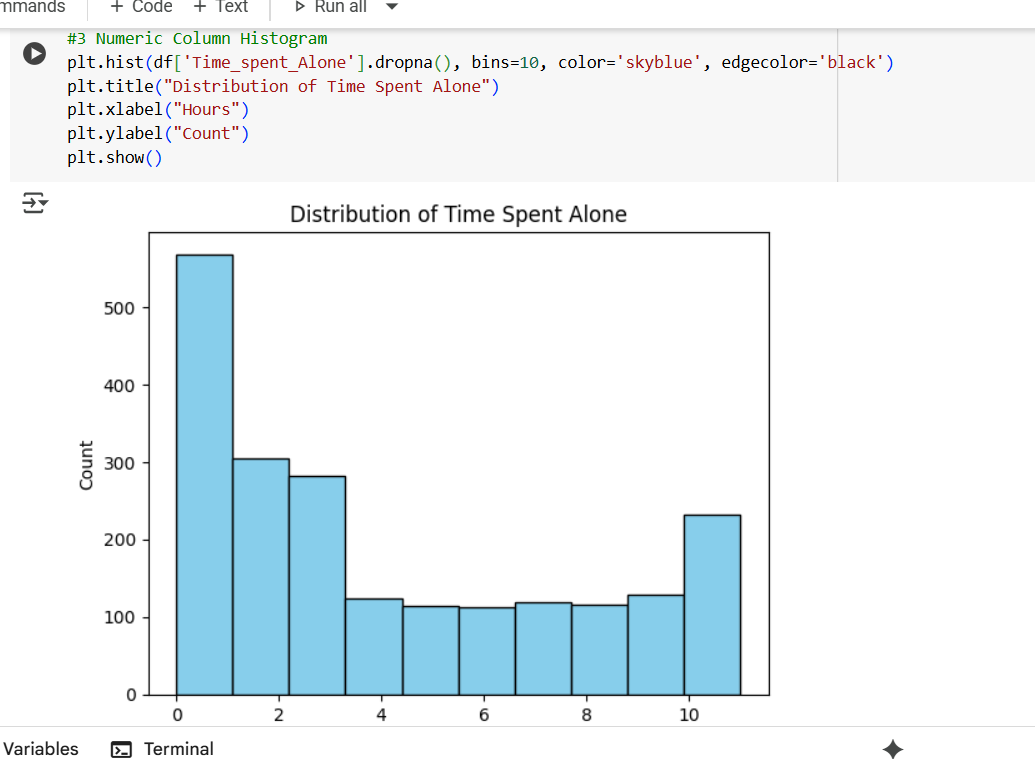
**Exploratory Data Analysis (EDA)**

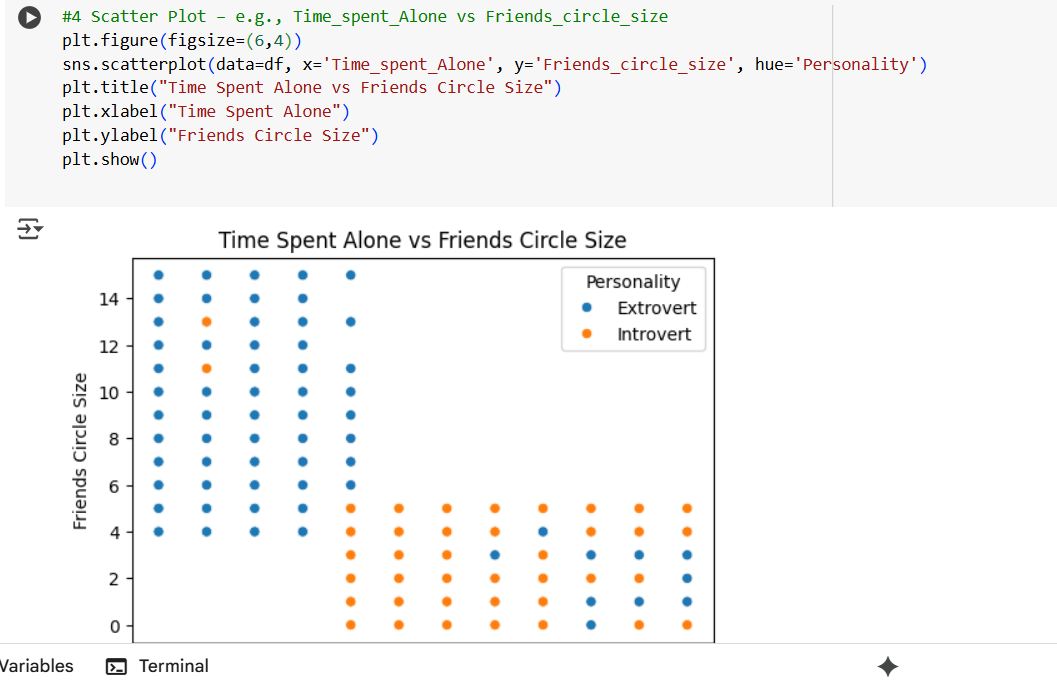
Exploratory Data Analysis (EDA) is a crucial phase in any data science project, serving as the initial deep dive into the dataset to uncover patterns, identify anomalies, test hypotheses, and validate assumptions. This process employs a combination of statistical summaries and graphical representations to gain a profound understanding of the data's underlying structure and relationships, ultimately guiding subsequent modelling efforts.







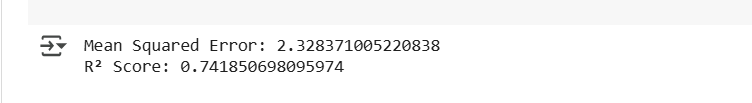


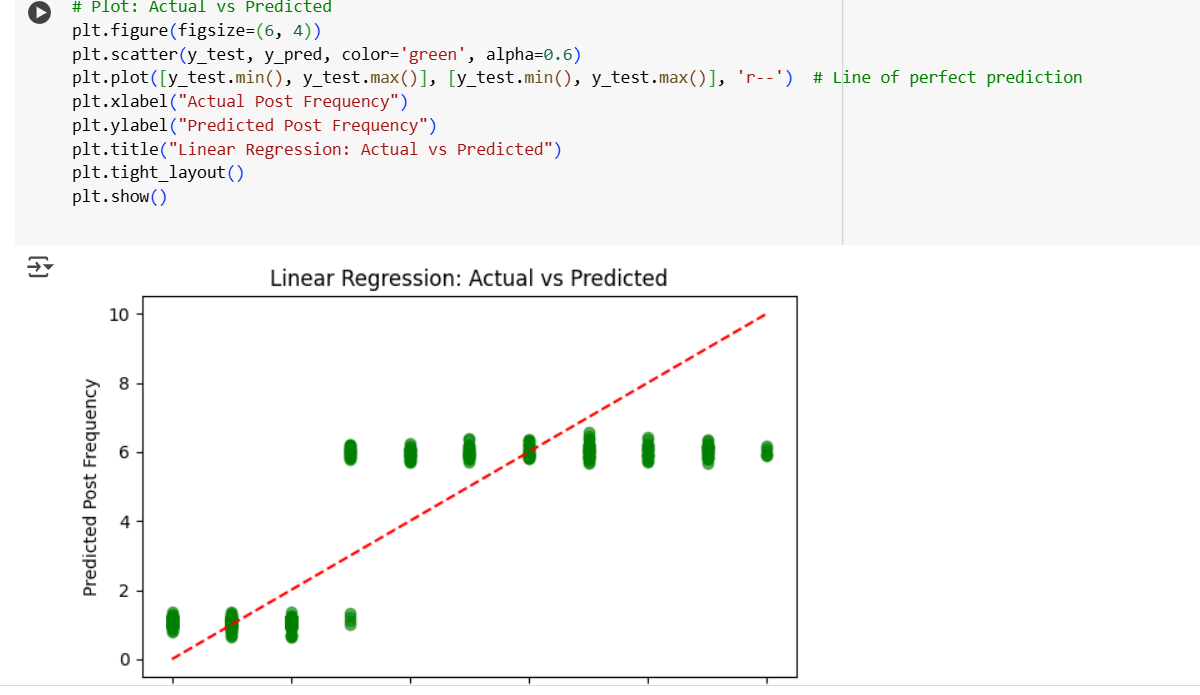


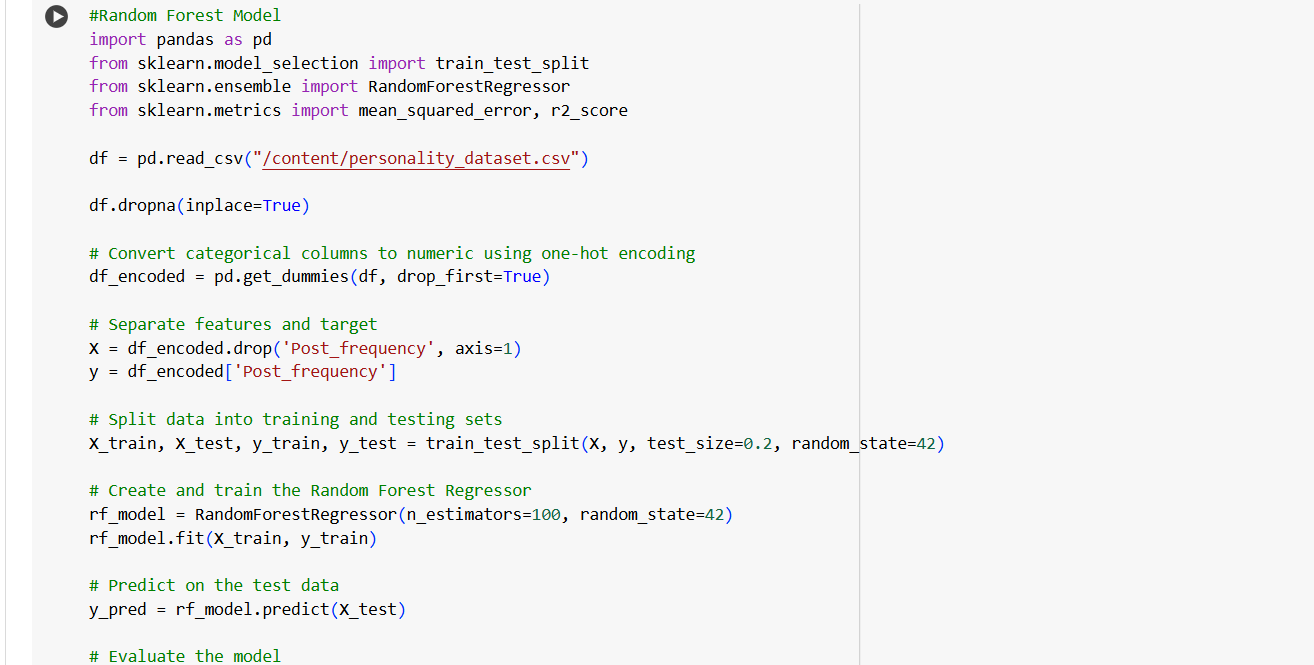
**Machine Learning Models**

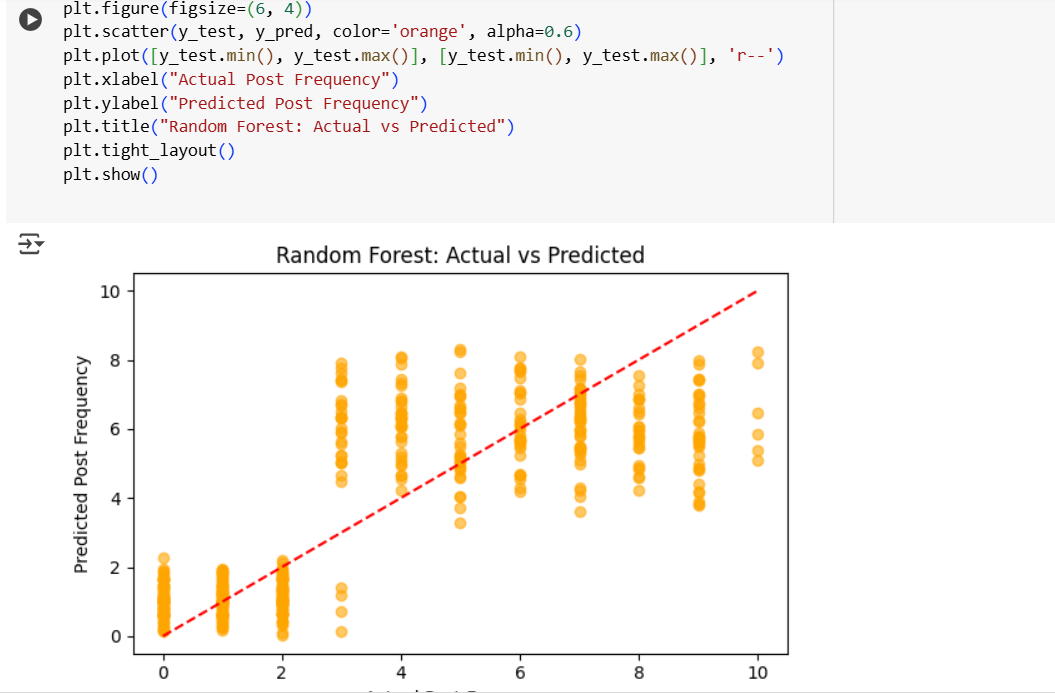
Following the rigorous data cleaning and comprehensive exploratory data analysis, the next pivotal stage involves leveraging Machine Learning models to uncover predictive insights from the dataset. Given that our primary objective is to classify an individual's Personality as 'Introvert' or 'Extrovert', this project focuses on binary classification algorithms. We have selected three distinct models – Logistic Regression, Random Forest, and K-Nearest Neighbours (KNN) – to provide a diverse range of approaches, comparing their interpretability, robustness, and predictive power.











**Power BI Dashboard**

The final and crucial step in effectively communicating the insights derived from this data science project is the development of an interactive and visually compelling Power BI dashboard. This dashboard serves as a dynamic interface, allowing stakeholders to explore the data, understand key findings, and interact with the analytical outcomes without needing in-depth technical knowledge. It transforms complex analyses into accessible, actionable intelligence.

This Power BI dashboard provides an interactive and visually engaging overview of the dataset related to social behaviour and personality traits. Designed with a clean, dark theme and vibrant purple and green accents, it aims to deliver key insights at a glance, facilitating exploration of individual social habits and their connection to personality.

Key Performance Indicators (KPIs):

At the top, four prominent cards highlight critical summary statistics:

* **Total Record (3K):** Indicates the total number of individuals or data entries in the dataset.
* **Avg\_Time\_spent (4.41):** Shows the average value for time an individual spends alone.
* **Average\_friends\_circle (6.10):** Represents the average size of an individual's friends circle.
* **Count of Social\_event\_attendance (2.9K):** Displays the total count of instances for social event attendance records.

Interactive Filters:

On the right-hand side, a 'Filters' pane offers granular control over the data displayed, enhancing interactivity:

* **Personality:** Allows filtering by 'Extrovert' or 'Introvert'.
* **Stage\_fear:** Enables filtering by 'Yes', 'No', or '0' (which might require data cleaning or represents another 'No' category).
* **Going\_outside:** Provides a numerical range filter from 0.00 to 7.00.

**Key Visualizations and Insights:**

1. **Time\_spent\_Alone by Stage\_fear (Donut Chart):** This chart, despite its ambiguous title, primarily illustrates the distribution of 'Introvert' (23.5%) and 'Extrovert' (76.44%) personalities within a specific segment related to Time\_spent\_Alone and Stage\_fear (likely indicating the overall personality split, or a specific filtered view).
2. **Total Record by Time\_spent\_Alone (Bar Chart):** This histogram-like visual displays the frequency of individuals across different levels of Time\_spent\_Alone (from 0 to 10), indicating that a larger number of records fall within lower time spent alone categories.
3. **Count of Personality by Stage\_fear (Clustered Bar Chart):** This chart effectively compares the counts of 'Introvert' and 'Extrovert' personalities against their Stage\_fear status ('No' or 'Yes'). It shows that while a significant number of both personality types report 'No' stage fear, Extroverts generally have a higher count of 'No' stage fear, and a notable portion of Introverts also do not report stage fear.
4. **Social\_event\_attendance by Going\_outside (Funnel/Bar Chart):** This visual (displayed on the right-hand side) appears to show the distribution or count of Social\_event\_attendance across different Going\_outside values (0 through 7). The 30.9% at the bottom suggests a concentration of social event attendance at a particular Going\_outside level or a cumulative percentage.
5. **Sum of Post\_frequency by Friends\_circle\_size (Line Chart):** The large line chart at the bottom, titled "Social\_event\_attendance by Going\_outside" but with axes labelled "Friends\_circle\_size" (X-axis) and "Sum of Post\_frequency" (Y-axis), likely depicts the relationship between the sum of Post\_frequency and the size of one's Friends\_circle\_size. It shows a trend where post frequency initially increases with friend circle size, then potentially stabilizes or slightly decreases, suggesting an optimal range for Friends\_circle\_size related to posting habits.

In summary, this dashboard provides a holistic view of the dataset, enabling users to quickly grasp key metrics, explore the distribution of variables, and identify potential relationships between social behaviours, fears, and self-reported personality types.

