# Complete Project Documentation

# Introduction

This report focuses on cleaning, exploring, and analyzing a dataset to extract meaningful insights and answer specific questions using Python. The process involves data preprocessing steps such as handling missing values, correcting data inconsistencies, and transforming variables for better usability. Exploratory Data Analysis (EDA) is performed to understand the relationships between variables and identify trends and patterns within the data. Additionally, visualizations are employed to answer specific questions and communicate insights effectively. The objective is to leverage statistical and graphical techniques to draw actionable conclusions from the data.

# Data

The dataset contains information about students, including their demographic details, academic background, and other personal attributes. Key columns in the dataset include Age, Height, Weight(lbs), Credit Amount, and Marital Status, among others. The data also provides information on students' preferences, such as the type of car they own and their favorite apps.

Before analysis, the data required significant preprocessing to handle missing values, split combined columns, and standardize inconsistent formats. For instance, columns such as `Height` were converted from feet and inches to centimeters, and `personal\_status` was split into `Gender` and `Marital Status`. This cleaned dataset serves as the foundation for uncovering meaningful insights through statistical and visual exploration.

A screenshot of a computer

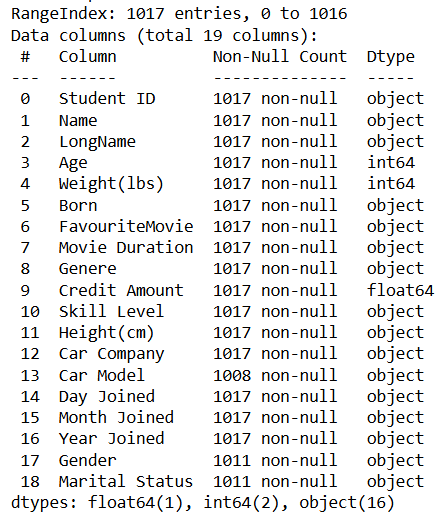
Description automatically generated

A screenshot of a computer

Description automatically generated

# Task 1: Data Cleaning

Data cleaning was performed to prepare the dataset for analysis by addressing missing values, correcting data types, and standardizing formats. Missing values were handled using appropriate replacements, such as the mode for categorical columns. Inconsistent formats, like heights in feet and inches, were converted to centimeters, and columns such as `personal\_status` were split into `Gender` and `Marital Status`. Data types were corrected, and anomalies or inconsistencies were resolved to ensure the dataset was accurate and ready for exploration. After cleaning this is how our data looks:



# Task 2: Data Exploration

1. Summary statistics for numeric columns

A screenshot of a computer

Description automatically generated

2. Number of males and females in the data from column Gender

A black text on a white background

Description automatically generated

3. Calculating mean for 'Credit Amount' by grouping data by 'Gender'

A number of numbers and letters

Description automatically generated with medium confidence

4. Average credit amount by 'Car Company' type in descending order

A screenshot of a computer

Description automatically generated

5. Calculating range for 'Age' column



# Task 4: Data Visualisation

Q1: Does age correlate with credit amount, and if so, how?

A graph of blue dots

Description automatically generated

The correlation coefficient between Age and Credit Amount: 0.10856961609925535

Since the correlation is positive, it suggests that as Age increases, Credit Amount tends to increase slightly, but the relationship is weak, as it is almost 0, meaning age does not strongly influence credit amounts in this dataset. Also, the scatter plot is randomly scattered so, it suggests that there is no correlation between these two.

Q 2: Identify the top 10 female students with the highest weights(lbs).

Here are the top 10 female students with highest weight.

Top of the list is Anbu Ezhil who is above 200lbs.

A graph of different colored bars

Description automatically generated

Q3: What is the average weight across different age groups?

A graph of weight across age groups

Description automatically generated

The age group with the highest weight is 46-70 in the dataset.

Q4: How does marital status relate to credit amount?

A graph of a credit amount by martial status

Description automatically generated

The boxplot shows that individuals with a single marital status tend to have higher credit amounts with several outliers reaching over 15,000. People who are divorced/separated or widowed have relatively lower credit amounts, and their distributions are more concentrated within the lower ranges. The divorced/separated group shows some higher credit amounts, while married individuals have a slightly higher median credit amount than the other groups. There are some outliers in all categories, especially in the single and divorced/separated groups.

Q5: What is the distribution of student ages?

A graph of a distribution of student ages

Description automatically generated

Here, the age distribution of the students can be seen starting from around 10 it goes up until it reaches the top with around 33 and then gradually comes down till 55. Hence, most of the population density is around 21 to 40.

# Conclusion:

This report successfully cleaned, explored, and analyzed the dataset to uncover meaningful insights. Through data preprocessing, inconsistencies were addressed, and the dataset was transformed into a reliable format. Exploratory Data Analysis and visualizations provided a deeper understanding of relationships and trends, such as the weak correlation between age and credit amount and the weight distribution among students. The analysis highlights the importance of data cleaning and visualization in extracting valuable information, offering actionable insights for further decision-making.