**Intermediate Customer Data Analysis Report**

**GROUP 2**

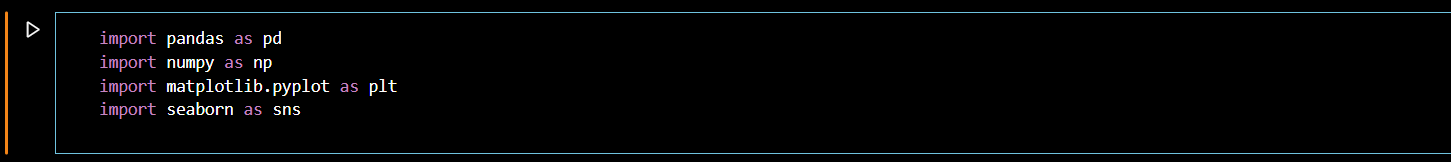
Nwachukwu Zinachidi

Adegoke Maryam

Olajide Abdullateef (Team Lead)

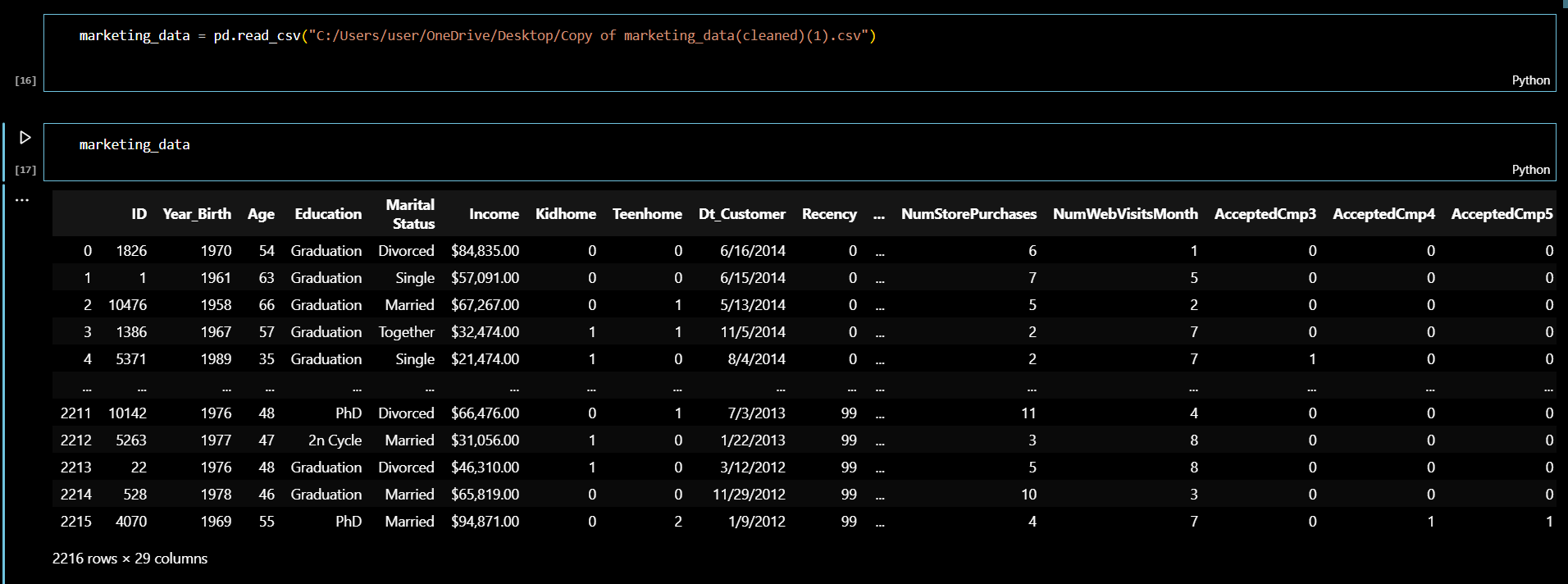
Tasks:

1. Data Import & Cleaning



*Figure 1*

The first step to carry out for this python task is to import the marketing dataset and all the libraries we will use to carry out the data analysis using python. In Figure 1, This code snippet imports several libraries commonly used in data science and machine learning. pandas (imported as pd) is used for data manipulation and analysis, providing data structures like DataFrames. numpy (imported as np) is essential for numerical operations, enabling efficient array handling and mathematical functions. matplotlib.pyplot (imported as plt) is a plotting library that allows for the creation of static, animated, and interactive visualizations. seaborn (imported as sns) is a statistical data visualization library based on matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics. These libraries together facilitate data handling, numerical analysis, and visualization, making them fundamental tools for data analysis and scientific computing.

**

*Figure 2*

In Figure 2, the entire marketing dataset is imported and shows a total of 2216 rows and 29 columns.

A screenshot of a computer

Description automatically generated

*Figure 3*

This code snippet in Figure 3 checks for missing values in the marketing\_data DataFrame and prints the total count of missing values for each column. The isnull() method returns a DataFrame of the same shape as marketing\_data, where each element is True if the corresponding element in marketing\_data is NaN (missing) and False otherwise. The sum() method then adds up the True values (which are treated as 1) for each column, providing the total number of missing values in each column. The print() function outputs these counts to the console, allowing you to quickly assess which columns have missing data and how many missing values each column contains.

A screenshot of a computer

Description automatically generated

*Figure 4*

This code snippet checks and prints the data types (dtypes) of each column in the marketing\_data DataFrame. The dtypes attribute of a DataFrame returns a Series with the data type of each column. By calling print(marketing\_data.dtypes), the code outputs a list of columns along with their corresponding data types (e.g., int64, float64, object), helping you understand the kind of data stored in each column.

A computer screen with text on it

Description automatically generated

*Figure 5*

In Figure 5, the code snippet adds a new column named Age to the marketing\_data DataFrame, which calculates the age of individuals based on their year of birth. It first imports the datetime module to access the current date and time. The current year is retrieved using date time. datetime.now().year. Then, a new column Age is created in the marketing\_data DataFrame by subtracting the values in the Year\_Birth column from the current\_year. This operation is performed using the loc accessor to ensure the new column is added correctly. Finally, the print (marketing\_data.head()) statement displays the first few rows of the updated DataFrame, showing the newly added Age column along with the existing data.

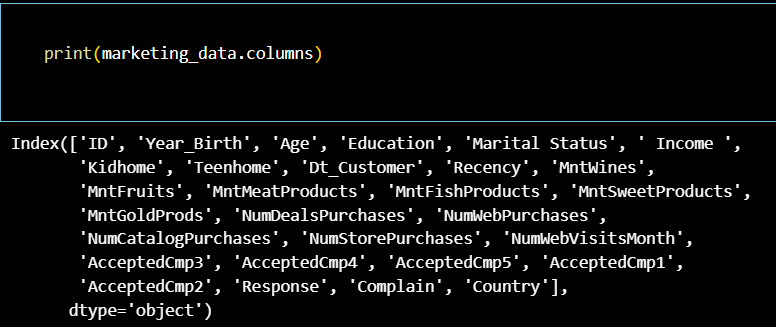
Output:

*A screenshot of a computer screen

Description automatically generated*

*Figure 6*

Figure 6 shows the output of first 5 columns of the original dataset with the added age column that was calculated.



*Figure 7*

The code snippet in figure 7 prints out the final columns of the marketing dataset

1. Exploratory Data Analysis

A screen shot of a computer code

Description automatically generated

*Figure 8*

This code snippet generates and displays descriptive statistics for the marketing\_data DataFrame. The describe() method is used to compute summary statistics for each numerical column in the DataFrame, including count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile or Q2), 75th percentile (Q3), and maximum. These statistics provide a quick overview of the distribution and spread of the data. The result is stored in the variable descriptive\_stats. The subsequent print statements output a title ("Descriptive Statistics:") followed by the descriptive statistics, allowing you to easily review key statistical measures for each numerical column in the dataset.

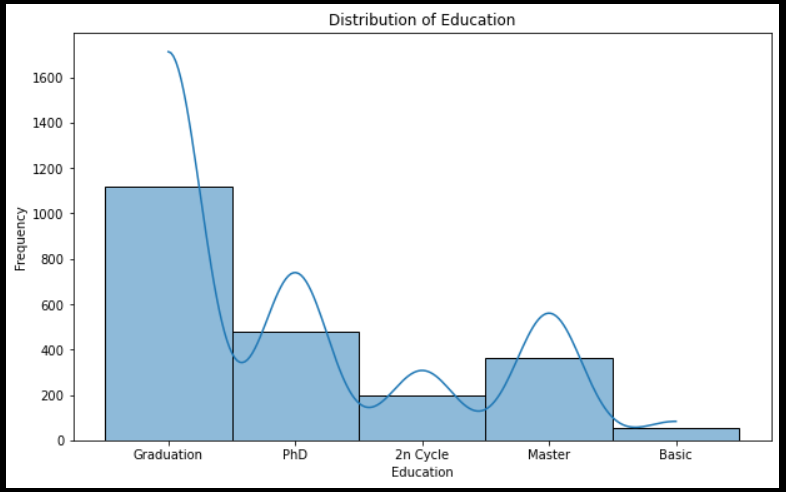
Output:

*A screenshot of a computer screen

Description automatically generated*

Figure 9

Using a Histogram



*Figure 10*

**Insights:**

Graduation: 1000-1100 people

PhD: 400-600 people

2n Cycle: 200 people

Master: 200-400 people

Basic: 0-200 people

From the above figure, we can see the different education levels. We can derive the most dominant education level from our dataset are people with who graduated with a count of more than a thousand people, the second most dominant education level to be those with a PhD, the third education level to be the people who attained a master’s degree, the fourth to be people with 2n Cycle and the lowest education level to be people who have a basic education with a count of approximately 90 people.

Using a scatterplot

A graph with blue dots

Description automatically generated

*Figure 11*

**Insights:**

This scatterplot explores the relationship between education level and income. We can derive that the people who have education levels of graduation, PhD, 2n Cycle and a master’s degree have an income of 100,000 and above while the people with a basic education level do not have an income of 100,000. This shows that the people with a higher level of education will make more money than the people who just have a basic level of education

Using a heat mapA chart of a heatmap

Description automatically generated

*Figure 12*

Color Intensity:

* The color gradient, from light yellow to dark blue, indicates the frequency of individuals within each age-education combination.
* Lighter colors (yellow) represent lower frequencies, while darker colors (blue) represent higher frequencies.

**Insights**

Most individuals have a 'Graduation' education level, as evidenced by the darker blue vertical stripe in the 'Graduation' column.

The 'Master' education level also shows significant frequencies, particularly around ages 40-60.

The other education levels ('2n Cycle', 'Basic', 'PhD') show relatively lower frequencies, as indicated by the lighter colors.

This heatmap provides a clear visual representation of how education levels are distributed across different ages, highlighting patterns and trends in the marketing dataset.

Using a boxplot

A chart with colorful boxes

Description automatically generated with medium confidence

*Figure 13*

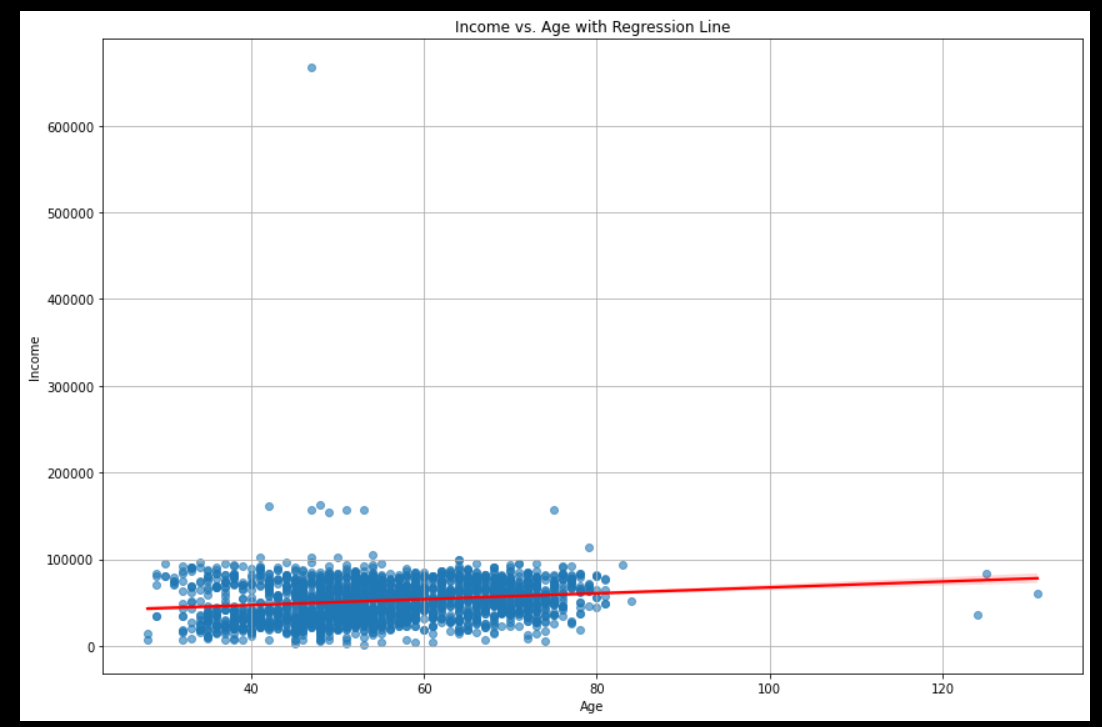
**Insights:**

This box plot helps in understanding the variability and central tendency of income across different education levels, highlighting how higher education generally correlates with higher income, though with considerable individual variation.

* Graduation: The income distribution for individuals with a 'Graduation' level education shows a median around 50,000, with some outliers above 100,000 and even one above 600,000.
* PhD: Like 'Graduation', the median income is around 50,000, with outliers extending up to 100,000.
* 2n Cycle: The median income is slightly lower than 'Graduation' and 'PhD', with fewer outliers.
* Master: This education level has a slightly higher median income compared to '2n Cycle', with some notable outliers.
* Basic: The lowest median income among all education levels, with the distribution tightly clustered and no significant outliers.

Overall higher education levels like 'PhD' and 'Master' generally show higher median incomes compared to 'Basic' and '2n Cycle'

Using another scatterplot:



*Figure 14*

**Insights:**

This scatter plot visualizes the relationship between Income and Age in the marketing dataset, with a regression line overlaid to show the trend. Each blue dot represents an individual from the dataset, with their age on the x-axis and their income on the y-axis. Most of the data points are clustered between ages 30 and 80, with incomes generally below 100,000. There are a few outliers with exceptionally high incomes.

The plot shows a wide variation in income at different ages, with many data points scattered around the regression line. There are a few outliers with very high incomes that are significantly above the main cluster of data points. Overall, this plot helps to visualize the relationship between age and income, showing that while there is a slight positive trend, income varies widely within age groups.

1. Purchasing Behaviour Analysis

A comparison of bar graph

Description automatically generated

*Figure 15*

**Insights:**

A screenshot of a computer

Description automatically generated

**From the table above table, MntWines** leads by a significant margin, indicating a strong preference for wine among the customer base. This is followed by **MntMeatProducts,** which also has a substantial amount of spending. **MntGoldProds,** despite being a luxury item, ranks third, suggesting a market for premium products.

Low Spending Segment: With an average of about 13.96 on wines and smaller amounts on other things, this group spends the least money overall. There isn't much interaction with more expensive products like MntGoldProds in this area.

Medium spending segment: In comparison to the Low group, there is a rise in spending in all categories. The noticeable increase in spending for MntWines and MntMeatProducts indicates a moderate preference towards these categories.

High spending segment: Spending has increased significantly in this category, particularly in MntWines and MntMeatProducts. The sector also reveals a significant rise in MntGoldProd spending, suggesting a growing demand for luxury items.

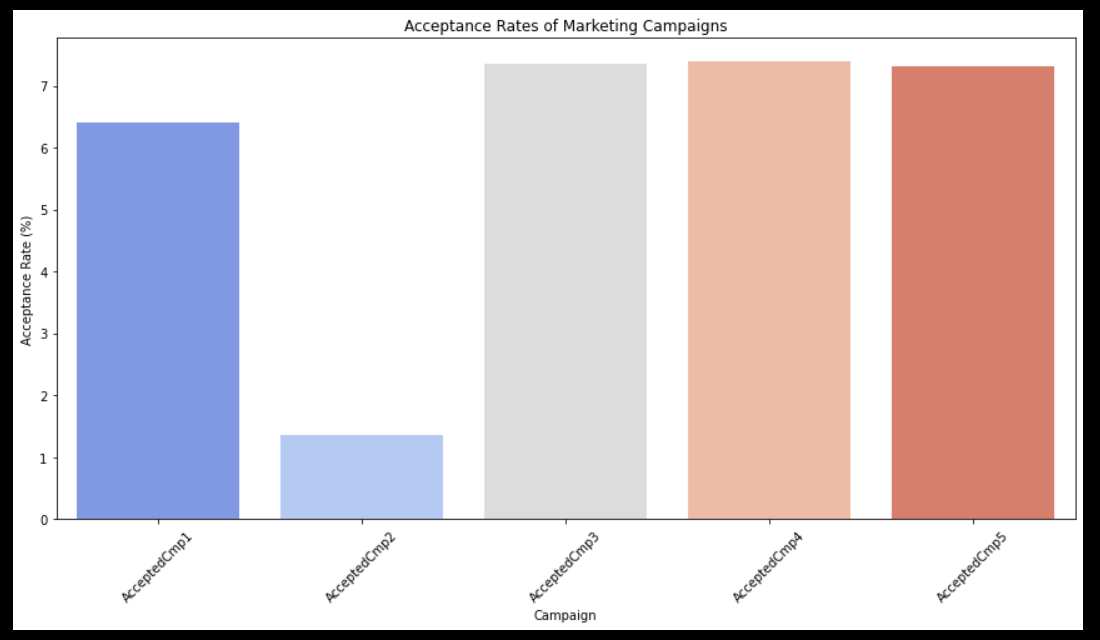
Very high spending segment: highest spending in every category, with MntMeatProducts and MntWines standing out especially. Significant expenditure is also observed at MntGoldProds, supporting the trend that affluent consumers have a preference for luxury items.

Recommendations

Discounts, special offers, or emphasizing the distinctive qualities of these products are some techniques that could assist boost customer interest in the less priority categories and raise the amount of money spent on them.

1. Marketing Campaign Effectiveness:

Using a bar plot



*Figure 15*

**Insights:**

Acceptance Rates for Each Campaign:

AcceptedCmp1 6.407942

AcceptedCmp2 1.353791

AcceptedCmp3 7.355596

AcceptedCmp4 7.400722

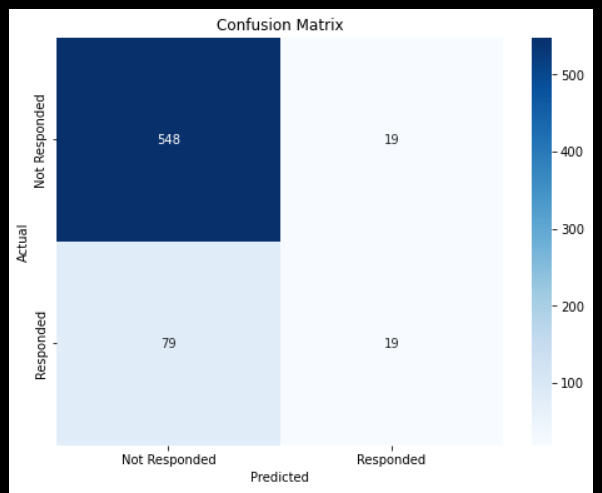
AcceptedCmp5 7.310469

From the bar chart we can see that the 3rd, 4th and 5th acceptance campaigns had a higher acceptance rate than the 1st and the 2nd. The campaign with the lowest acceptance rate is the second one

Recommendation

In light of the results, make the required changes in the campaign in order to steer clear of same mistakes in the future. This can entail modifying the timing, enhancing the targeting parameters, or altering the offer. Aim to target certain client segments with your campaigns. Enhance the probability of acceptance by using consumer data for communications and offers.

1. Predictive Analysis



*Figure 16*

Accuracy: 0.85

Precision: 0.50

Recall: 0.19

F1 Score: 0.28

From the value above, it shows that the accuracy is 0.85 which the model correctly

predicted the outcome 85% of the time which is reasonable.

The positive predictive value(precision) of 0.50 indicates that when the model predict a positive outcome, it is correct 50% of the time.

The Recall which calculates how many real positive cases the model properly detected. The model successfully recognized 19% of the real positive cases with a 0.19. this indicates that it is low and the model is missing a large number of true positive cases (false negatives).

An F1 score of 0.28 in this case indicates a relatively poor balance between precision and recall.

The performance metrics suggest a model with decent overall accuracy but Significant issues with detecting positive cases. The precision indicates that when the model does predict a positive case, it is correct half of the time. However, the very

Low recall and F1 score highlight a major problem: the model is failing to identify a

Large portion of true positive cases, which could be critical depending on the

Application’s context.

Recommendation

I will recommend for improvement that there should be focus more on improving one statistic than the other, depending on how important precision is compared to recall in your application.

More visualization

Using a heatmap of the correlation matrix to see if Age, Education, and Income are correlated.

A red and blue squares with white text

Description automatically generated

*Figure 17*

Insight:

From the above correlation matrix, it will be observing that the matrix is symmetric, meaning the correlation between Age and Income is the same as between Income and Age which is slightly correlated.

Given that education is a categorical variable and that correlation matrices comprehend the connections among various numberical variables, it is not possible to display the correlation graph.

Analysing average income by education level to see if higher education correlates with higher income.

A graph of a graph showing the average income by education level

Description automatically generated

**Insights:**

This bar plot illustrates the correlation between individuals education level and the income they make. Individuals with a higher level of education like ‘Graduated’, ‘Master’ and ‘PhD’ tend to make a higher income than the individuals with a lower education level.