smdm project

**Problem 1**

Austo Motor Company is a leading car manufacturer specializing in SUV, Sedan, and Hatchback models. In its recent board meeting, concerns were raised by the members on the efficiency of the marketing campaign currently being used. The board decides to rope in an analytics professional to improve the existing campaign.

**1.1. Information about the size of the dataset and the nature of the variables.**

**Nature of the dataset shows as:**

Text

Description automatically generated

Table 1.1

**The data has 1581 observations with 14 entries. 5 integer type as numerical variables, 1 float type as numerical variables and 8 object type as categorical variables.**

**1.2 Information about the size of the dataset :  
**Table 1.2

**The data set contains 1581 observations of data and 14 variables.**   
  
  
**2. Take a critical look at the data and do a preliminary analysis of the variables. Do a quality check of the data so that the variables are consistent. Are there any discrepancies present in the data?**  
**2.1 Checking for missing value :**

Text

Description automatically generated with medium confidence

Table 2.1

### There are 53 null values in Gender and 106 null values in Partner\_Salary dataset.

### 2.2 Treating NAN values: Table 2.2.1

### But when we checked the Gender and Partner\_salary we found some discrepancies and data is not consistent. Let’s fix it. Also when we check unique values of gender column, we see another 2 extra attributes like Femle , Femal , nan as mentioned so may be the backend agent would have mistyped it to feed the data. Let’s fix it.

### Chart Description automatically generated with medium confidence

### Table 2.2.2

### 2.3 Treating Missing values: So after fixing the discrepancies, such as fixing the null values and adding the count value of Femle and Femal to Female attribute. Then the data looks like:

### Text Description automatically generated

### Table 2.3.1

### Now we can see only 2 attributes such as Male and Female and 53 missing values in Gender column has been added to Male attribute by taking the Mode value of Gender column we see the output shows as 0 Male hence we filled all the missing values in gender column as Male.

### Since Partner\_Salary attribute had 106 missing values so we imputed all missing values using mean since Partner\_Salary has no outliers as mentioned below. Chart Description automatically generated

### Fig 2.3.2

### From the above plot, now we can see missing values in Gender and Partner\_salary has been already treated.

### 

### Fig 2.3.3

### Text Description automatically generated

### Table 2.3.4

### 2.4 Checking for duplicates :

### Fig 2.4.1 As we can see no duplicates records found.

### 2.5 Check for summary of the dataset

### Table Description automatically generated

### Table 2.5.1

### A screenshot of a computer Description automatically generated with medium confidence

### Table 2.5.2

### The above picture depicts the five-point summary of the continuous attributes. Analyzing at the age column, we can see that the distribution of the adult population is between the minimum of age 22 years and with maximum age of 54 years. So 25% and 50% of the people having age of 25 and 29 years has number of dependents as value 2 and 75% of the people having age of 38 years has number of dependents as value 3. And number of dependents shows as 0 for minimum age of 22 years and value 4 being the highest number of dependents for people having age with 54 years.

### 3. Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.

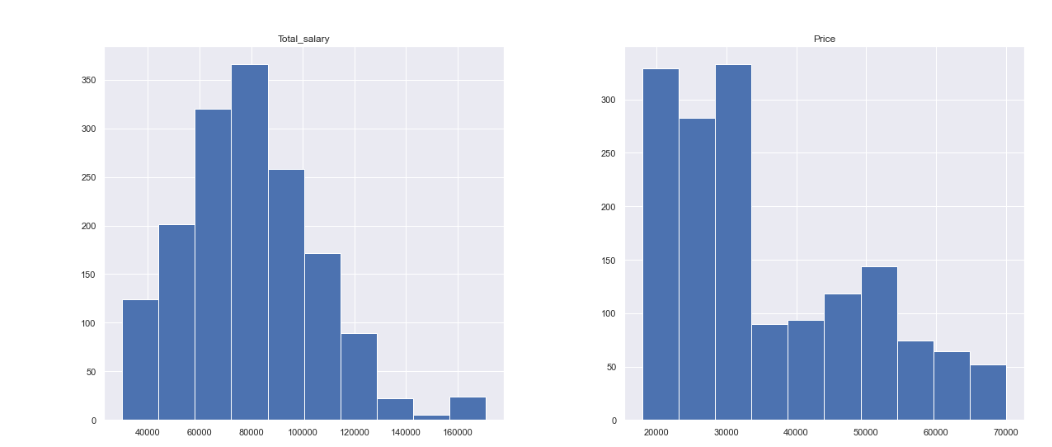
### 3.1 Let’s plot the histogram to see the distribution of the continuous features continuously.

**Chart, histogram

Description automatically generated**   
Fig 3.1.1

Chart, histogram

Description automatically generated  
Fig 3.1.2

****Fig 3.1.3 **Looking at the above histograms, we can see that age (left-skewed) , no\_of\_dependents is not uniformly distributed (right-skewed) , salary is uniformly distributed , partner\_salary is not uniformly distributed (left-skewed) , total\_salary (left-skewed) and price is left-skewed.   
  
  
3.2 Now we shall look at how the variables are distributed with the help of countplot.**

Chart, bar chart

Description automatically generated   
Fig 3.2.1  
  
We can see Gender count of Male is higher than Female.   
Chart, bar chart

Description automatically generated   
Fig 3.2.2  
  
We can see professional wise – Salaried people are more than business professionals.

Chart, bar chart

Description automatically generated  
Fig 3.2.3  
  
  
From the above plot, we can draw conclusions that married people are more than the singled one’s.   
  
Chart, bar chart

Description automatically generated  
Fig 3.2.4  
  
  
We can see most people background education shows as post graduate and compartively people have also pursued graduate degree education.   
  
  
  
Chart, bar chart

Description automatically generated  
Fig 3.2.5

As depicted, we can see personal\_loan status shows as same for both the values : Yes and No.

Chart, bar chart

Description automatically generated  
Fig 3.2.6   
  
Here, we can assume that more number of people didn’t take house\_loan but half of the count of the people have taken house\_loan.

Chart, bar chart

Description automatically generated   
Fig 3.2.7  
  
  
In this we can see, that partner\_working counts (Married and partner working) shows as little high than those of for partner\_working which are not working. (Married but partner not working).   
Chart, bar chart

Description automatically generated  
Fig 3.2.8  
  
  
The above plot depicts that Brand ‘Sedan’ is the most purchased followed by ‘Hatchback’ and the least is ‘SUV’.

### Chart, box and whisker chart Description automatically generated Fig 3.2.9

### We can see from the above plot that Gender Male has more extreme values of Age when compared to Female.

### Table Description automatically generated Table 3.2.1.1

### So here we can see that most of the Female Gender uses SUV automobile and Gender Male uses more Hatchback automobile.

### Table Description automatically generated Table 3.2.1.2

### Another analysis, we can do here is that we can see more Married people are purchasing Sedan automobile while Single’s are preferring more Hatchback automobile.

### Chart, histogram Description automatically generated Fig 3.2.10

### Here using histplot, we can see the Age is rightly-skewed distribution and with the help of boxplot, we can see no outliers found in Age variable.

### Chart, histogram Description automatically generated Fig 3.2.11

### In the same manner, we can analyze that even Partner\_Salary has no outliers which we tried to check using boxplot and through the help of histplot, we see the distribution of Partner\_Salary is again right-skewed distribution. 3.3 Bi-variate distribution of every possible attribute pair Diagram Description automatically generated with medium confidence Fig 3.3.1

By using the pairplot we can see the bivariate distribution. As age gets increased even the salary is getting increased and looks like people with more age are spending more price on the purchases.

**3.4 Correlation :**

Now let’s have a look at correlation with the help of heatmap.

Chart

Description automatically generated **Fig 3.4.1**

**4. Understanding the relationships among the variables in the dataset is crucial for every analytical project. Perform analysis on the data fields to gain deeper insights. Comment on your understanding of the data.**

**4.1 Let’s start implementing Univariate Analysis**

**Variables used : Age, No\_of\_Dependents (Numeric with numeric)**

Table

Description automatically generated   
Table 4.1  
  
  
We can see Age attribute has a minimum of 22 years and maximum of 54 years. From the above data, we can see that 50% of the people has age 29.

Minimum no-of-Dependents is 0 and maximum is 4. Here, we can see 25% and 50% of the age group ranging between 25 to 29 years has no-of-Dependents as 2 and 3 for 75% of the people having age 38.

Chart, histogram

Description automatically generated  
Fig 4.1.1

We can observe from the above 2 plots, there is no outliers for Age but we can see outliers present in no-of-Dependents.

**Now we shall take another 2 numeric variables to draw some more insights.**

**Variables used : Total\_salary and Partner\_Salary.**

Table

Description automatically generated  
Table 4.1.1.1  
  
We can conclude that minimum salary of Total\_salary is 30,000 with maximum salary as 1.71la. In Partner\_salary we can see minimum salary as 0 and 25% of the partner who are working also do not contribute anything. So the maximum salary from Partner\_salary shows as 80500.

Chart, histogram

Description automatically generated  
Fig 4.1.2  
We can see more outliers found in Total\_Salary which we will treat it later. But no outliers found in Partner\_salary.

**Categorical variable :**

Text

Description automatically generated  
Table 4.1.1.2

So the Gender categorical variable we have displayed the results in percentage form which contributes 0.80% as Male and %0.20 as Female.

Chart, bar chart

Description automatically generated  
Fig 4.1.3   
  
From the above plot, we can depict that count of Male gender is more when compared to Female gender.

**Let’s draw some more insights with another categorical variable.**

**Variable used : Profession.**

Text

Description automatically generated with low confidence  
  
Table 4.1.1.3

Chart, bar chart, treemap chart

Description automatically generated  
Fig 4.1.4

So we can conclude that 0.57% constitutes Salaried profession and remaining 0.43% belongs to Business profession.

**4.2 Let’s do Bivariate Analysis**

**So here we take 2 numeric variables. Variables used are : Salary , Price.**

Chart, scatter chart

Description automatically generated  
Fig 4.2.1

From the above plot, we can see as gradually salary keeps on increasing even the amount of price spent also gets increased.

**So lets do some more bivariate analysis on 2 categorical variables**.

**Here variables used are : Marital\_Status and Partner\_working**

Chart, bar chart

Description automatically generated   
Fig 4.2.2  
  
  
Table

Description automatically generated  
Table 4.2.1.1

So from both the data and plot, we can assume that 0.087% are single and 0.91% constitutes to married people which includes partner\_working.

**Now let’s see bivariate analysis on categorical variable with numeric variable.**

**Variables used : Salary , Profession**.

Chart, box and whisker chart

Description automatically generated

Fig 4.2.3   
  
So when we see the median salary of both business and salaried profession people it looks like same salary.

**Let’s analyze with some more variables.**

**Variables used : Total\_salary , Personal\_loan**

Chart, box and whisker chart

Description automatically generated  
Fig 4.2.4  
  
We can observe here that median value of total\_salary who has taken is little less but almost nearer to the median value of the total\_salary who haven’t opted for personal loan.

**4.3 Multivariate Analysis**

**Variables used : Marital\_status , Age , Gender.**

Chart, box and whisker chart

Description automatically generated  
Fig 4.3.1  
  
From the above plot, we can see that Marital\_Status of Gender Female who are single has higher median value when compared with Male.

**Now let’s draw some more insights with the help of more than 2 variables.**

So with the help of FacetGrid, we can do some analysis.

Chart, scatter chart

Description automatically generated  
Fig 4.3.2.  
  
Hereby, we can overall see that most of the profession who are salaried has more number of eductaion status as post graduation and graduate when compared to business profession.

**4.4 Skewness :**Let’s measure the skewness of the required columns

Graphical user interface

Description automatically generated with low confidenceFig 4.4.1Skewness of Age , Total\_salary , Price and Partner\_salary seems to be moderately skewed distribution wherein no\_of\_dependents and Salary are approximately symmetric distribution.

**4.5 Checking for Outliers :**We shall make use of boxplot to check for outliers.   
Chart, box and whisker chart

Description automatically generatedFig 4.5.1

We can see lot of extreme outliers present in **Total\_salary** as it is highly skewed and only 1 outlier in **no-of\_dependents** but no outliers present in other variables.

**Removing the outliers :**

As we had seen outliers in **Total\_salary** and **No\_of\_dependents.** Let’s remove it by replacing the outlier value using IQR.

So once outliers removed then we can see the following attributes does not contain outliers now.   
  
Chart, box and whisker chart

Description automatically generated   
Fig 4.5.2

**4.6 Encoding**

There are 2 types. But we shall use **Label Encoding** to see how it works.

Text

Description automatically generated  
Table 4.6.1

Here, we see label as male and female so let’s replace male and female with values 0 and 1.

A picture containing logo

Description automatically generated  
Table 4.6.2

Now, we can see Male label has been assigned as value 0 and Female label has been assigned as 1.

**5. Employees working on the existing marketing campaign have made the following remarks. Based on the data and your analysis state whether you agree or disagree with their observations. Justify your answer Based on the data available.**

**E1) Steve Roger says “Men prefer SUV by a large margin, compared to the women”**   
  
Text

Description automatically generated with medium confidence  
Table E1

Chart, bar chart

Description automatically generated  
Fig E1  
  
From the above analysis, we can see count of Female gender who prefer SUV is 173 but Male Gender who prefers SUV count is 118 hence disagree with Steve Roger statement and can say that Female prefer SUV by a medium margin when compared to Men. **So “Men prefer SUV by a large margin, compared to the women” is wrong and false.**

**E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.**  
Text

Description automatically generated  
Table E2

Chart, bar chart

Description automatically generated  
Fig E2

**Yes, correct.** We can see the count of Salaried people buying Sedan is 396 whereas the Business profession people buying Sedan count is 306 hence we can agree to Ned Stark statement of “Salaried person is more likely to buy a Sedan”.

**E3) Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for a SUV sale over a Sedan Sale.**

Chart, bar chart

Description automatically generated  
Fig E3

As per the plot, we can see that the count of salaried Male people buys SUV more than Sedan when compared to Female salaried people. **Hence the Sheldon Cooper statement is true stating that a salaried male is an easier target for a SUV sale over a Sedan sale.**

**F. From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a Business can utilize the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions.**

**Give justification along with presenting metrics/charts used for arriving at the conclusions.**

**F1) Gender**

**F2) Personal\_loan**

Chart, bar chart

Description automatically generated  
Fig F

So gender wise, if we can see Female are spending more than Male.

With regards to personal\_loan analysis, we can conclude that without having any personal\_loan Female are spending more on purchasing automobiles when compared to Male but Female gender count taking personal\_loan is almost near close to the counts of Female not opting for Personal\_loan.

**G. From the current data set comment if having a working partner leads to the purchase of a higher-priced car.** Chart, bar chart, funnel chart

Description automatically generatedFig G

We can see that partner who are not working seems to be slightly more edged than partner who are working with regards to purchasing in high priced car. But partner who are working seems to be very close near to that of partner who are not working. **So, conclusion is, if having a working partner leads to the purchase of a higher-priced car – it’s not agreed, it’s no and statement is proved to be wrong with the analysis.**

**H. The main objective of this analysis is to devise an improved marketing strategy to send targeted information to different groups of potential buyers present in the data. For the current analysis use the Gender and Marital\_status - fields to arrive at groups with similar purchase history.**   
  
Chart, box and whisker chart

Description automatically generated  
Fig H.1

With respect to Marital\_Status wise, we can observe that Male Gender who are single and married seems to be very closer to the median value and purchase history also shows close to similar values. But for Female Gender we can see the purchase history seems nearly close with that of same married and single female type.

Chart, box and whisker chart

Description automatically generated  
Fig H.2

With regards to Gender wise, from the above plot we can assume that purchase history of SUV automobile seems to be similar in both Male and Female. But for other models in both the Genders, we can see no traces of similar purchase history and are completely not uniformly distributes.

Chart, scatter chart

Description automatically generated  
Fig H.3

We can also analyze by combining all the variables together so here we can see that Married Male and Married Female have similar purchase history in all the models but Males who are Single and Females who are single we can conclude that their purchase history patterns are scattered not consistent and are not similar.

**OVERALL ANALYSIS :**

**So in Problem 1, through all these analysis we can conclude that Female are spending more on automobiles on SUV model and that too married Females who have been working. Hence we can see the Price of the SUV model which is compatible to be purchase so that’s the most of the reason Females are choosing SUV brand. We should check the other 2 models features compared to SUV model and price analysis so that the other 2 models can be bought into competition so that Female and Gender category can purchase it based on the customization of the features , price , etc which is found in SUV model.**

**PROBLEM 2**

A bank can generate revenue in a variety of ways, such as charging interest, transaction fees and financial advice. Interest charged on the capital that the bank lends out to customers has historically been the most significant method of revenue generation. The bank earns profits from the difference between the interest rates it pays on deposits and other sources of funds, and the interest rates it charges on the loans it gives out.

GODIGT Bank is a mid-sized private bank that deals in all kinds of banking products, such as savings accounts, current accounts, investment products, etc. among other offerings. The bank also cross-sells asset products to its existing customers through personal loans, auto loans, business loans, etc., and to do so they use various communication methods including cold calling, e-mails, recommendations on the net banking, mobile banking, etc.

GODIGT Bank also has a set of customers who were given credit cards based on risk policy and customer category class but due to huge competition in the credit card market, the bank is observing high attrition in credit card spending. The bank makes money only if customers spend more on credit cards. Given the attrition, the Bank wants to revisit its credit card policy and make sure that the card given to the customer is the right credit card. The bank will make a profit only through the customers that show higher intent towards a recommended credit card. (Higher intent means consumers would want to use the card and hence not be attrite.)

**2. Framing An Analytics Problem. Analyze the dataset and list down the top 5 important variables, along with the business justifications.**

**2.1 So let’s see the nature of the variables first :**

Table

Description automatically generatedTable 2.1

**The above screenshot clearly shows that there are 21 integer data type variables , 6 object data type variables and 1 datetime type variable.**

Table 2.1.2

**The size of the data : We can observe total of 8448 observations with 28 entries.**

**2.1.3 Data Summary :**

Graphical user interface, application, table

Description automatically generated

From the above screenshot, we can see that overall 5 summary of the dataframe. So based on unique variable we can check the values listed such as count , mean , standard deviation , min and max value , 25% , 50% , 75% value. Data which is displayed seems to be reasonable as we can see 5 types of different category of customers , customers having maximum of 7 widget products and 8 engagement products.

**Variable high\_networth has mean value as : 1.99 and median values as : 2.0**

**cc\_limit variable has mean : 251706.91 , median : 150000.0**

**other\_bank\_cc\_holding variable has mean : 0.44 , median : 0.0**

**avg\_spends\_l3m variable has mean : 49527.36 , median : 37943.0**

**active\_90 variable has mean : 0.64 , median : 1.0**

**2.1.4 Now we shall check any null values present or not:**

Text

Description automatically generated  
Table 2.1.4

**So only we have 38 null values in Transactor\_revolver variable which has to**

**be fixed.**

**2.1.5 Then let’s analyze whether any duplicate values exist or not :**

Graphical user interface, text, application

Description automatically generated  
Fig 2.1.5

**So we can see no duplicate values found in the dataset.**

**2.2 Univariate Analysis :**

Let’s see the distribution of each and every variable separately. So we have taken only 5 top most variables for review.

**Variables: high\_networth**

Chart, bar chart

Description automatically generated  
Fig 2.2.1

Graphical user interface

Description automatically generated with low confidence  
Table 2.2.1.1

So here, we can see 5 different types of customers ranging based on their networth value and 1740 customers being categorized as A group customer , B group with 1660 customers , C group has 1659 customers , D group has 1696 customers , E group has 1693 customers. Hence we can conclude A group has the high graded customers who are having high networth value while C group being the low networth value customers.

**Variable : cc\_limit**

Chart, histogram

Description automatically generated  
Fig 2.2.2

  
Table 2.2.2.1

The above data shows the current credit card limit of the customer with maximum credit given value shows as 990000 and min is 0.

**Variable : other\_bank\_cc\_holding**

Chart, bar chart

Description automatically generated  
Fig 2.2.3

Logo

Description automatically generated with medium confidence  
Table 2.2.3.1

So here we can see 4728 customers are using almost other bank credit cards and 3720 customers are not using any other bank credit card nor using other bank credit cards.   
  
  
**Variable : avg\_spends\_l3m**

Chart, histogram

Description automatically generated  
Fig 2.2.4

  
Table 2.2.4.1

So here we can see the average credit card being spent in the last 3 months by the customer. Hence we can assume that maximum credit spent is 289292 and minimum value is 0.

**Variable : active\_90**

Chart, bar chart

Description automatically generated  
Fig 2.2.5

A picture containing text

Description automatically generated  
Table 2.2.5.1

active\_90 shows the customer’s account activity in the last 90 days. Here we can see that 3024 customers are not active in the last 90 days means not performing any transactions and 5424 customers are actively making the transactions.

**2.3 Bivariate Analysis of the top 5 most variables :**

Chart

Description automatically generated  
Fig 2.3

So from the above plot, we can depict that variable avg\_spends\_l3m and cc\_limit are related with each other. So as the cc\_limit gets increased then the avg\_spends\_l3m values shows in a increased manner.

**2.4 Now let us check the correlation between each variables :**

Chart, treemap chart

Description automatically generated  
Fig 2.4

So the heatmap shows there is a strong relationship between variable avg\_spends\_l3m and cc\_limit and we saw this already in pairplot also.

**2.5 Outliers :**

So we have to check whether any extreme values being found so outliers has to be checked.

Chart

Description automatically generated  
Fig 2.5.1

Chart, box and whisker chart

Description automatically generated  
Fig 2.5.2

A blue rectangle with white text

Description automatically generated with low confidence  
Fig 2.5.3

Chart, box and whisker chart

Description automatically generated  
Fig 2.5.4

Chart, shape

Description automatically generated  
Fig 2.5.5

We can see more extreme values found in cc\_limit and avg\_spends\_l3m variables which should be treated and high\_networth variable has no outliers found.

**OVERALL ANALYSIS :**

**Here the top 5 most variables are high\_networth , cc\_limit , other\_bank\_cc\_holding , avg\_spends\_ls3m , active\_90 so we have to first categorized the customer properly based on their income and the credit limit which they are looking into. So based on that we can offer the different types of credit cards if they are eligible and already existing customers records of last 3 months credit usage can be checked for more analysis whether they are using other bank credit cards or not using our credit card or how much they have been spending and how much is the limit given so based on the account status in the last 90 days of the customer by using all these 5 top variables we can segregate the customers to improve the credit card supply rather than simply offering to customer who are rarely using it.**