Statistical Methods and Decision Making

Project Report

By Ruchita Parulekar

Contents

|  |  |
| --- | --- |
| **Question** | **Page** |
| What is the important technical information about the dataset that a database administrator would be interested in? (Hint: Information about the size of the dataset and the nature of the variables) | 3 |
| 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? If yes, perform preliminary treatment of data. | 4 |
| Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business. | 5 |
| 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. | 10 |
| 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” \*\*\*E2) Ned Stark believes that a 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. | 12 |
| 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 | 13 |
| From the current data set comment if having a working partner leads to purchase of a higher priced car. | 15 |
| 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 Gender and Marital\_status - fields to arrive at groups with similar purchase history. | 15 |
| \*\*\*Framing An Analytics Problem\*\*\* Analyse the dataset and list down the top 5 important variables, along with the business justifications. | 18 |

# 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 analytics professional to improve the existing campaign.**

1. ***What is the important technical information about the dataset that a database administrator would be interested in? (Hint: Information about the size of the dataset and the nature of the variables)***

*SOLUTION:*

*Information about the dataset –*

*Shape – (1581,14)  
Total Rows – 1581*

*Total Columns ­­­­– 14  
Float DataTypes – 1*

*Object DataTypes – 8*

*Int DataTypes – 5*

*There are missing values in 'Gender' and 'Partner\_salary'. 'Gender' is categorised as object type.*

*A screenshot of a computer

Description automatically generated*

1. **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? If yes, perform preliminary treatment of data.**

*SOLUTION-*

*Below unique names have been identified in column “Gender”*

*There were two spelling error found in the column “Gender” – ‘Femal & Femle’*

**

*The above anomaly is corrected by creating a function, and below are the results.*

**

*The null values in “Gender” are removed by using the fillna method.*

*A screenshot of a computer screen

Description automatically generated*

*“Partner\_salary” has missing values, but the values are missing where there is no partner working , so there’s no need to treat the missing values as it will affect the analysis.*

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

***SOLUTION –***

*Analyzing Age Variable -*

*Younger age group (Range 20- 30) tends to buy more cars as compared to the middle aged (Range 31-45) and older age group (range from 46-55). Also there is fluctuation in the buying pattern for the age group between 35-40, sales for the cars between this age group is slightly better after young age group and compared to rest of the age group.*

***A graph of age and age

Description automatically generated***

***Analyzing Make wrt Profession -*** *We can observe that Salaried Professionals are likely to buy SUV's and Sedan's by a large margin than a Business Professionals.*

*We can also observe that Hatchback Cars are equally bought by Salaried and Business Professionals.*

***A graph of different colored bars

Description automatically generated***

***Analyzing Make wrt Gender -*** *Large number of Males prefer to buy Sedan and Hatchback than SUV. There are also very few Female Buyers*

***A graph of a number of people

Description automatically generated***

*Total\_salary has multiple outliers in upper values*

***A group of blue rectangular objects

Description automatically generated***

***Analyzing Total\_salary and Price*** *-People with Total salary between 60000 to 100000 are the greatest number of buyers and also they buy in the lowest category 20000 to 40000 category.*

***A graph of different colored squares

Description automatically generated***

***Analyzing Make and Total\_salary wrt Gender*** *- SUV's are more likely to be bought by Female than Male population and Sedan and Hatchback are bought more by Male than Female population.*

*The very little population which has high salary prefer to buy SUV than Sedan and Hatchback and population with Medium to low salary prefer to buy Sedan and Hatchback.*

***A group of blue and orange dots

Description automatically generated***

***D. 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.***

*SOLUTION –*

*Younger age group (20-30 ) they mostly buy Hatchback and Sedan as compared to SUV*

*Mid age group (31-45) they mostly buy Sedan and SUV (few sales only) with no sales of hatchback to this age group.Older Age group (46-55) only buy SUV with no sales of Sedan or Hatchback. There is zero sales of Hatchback from the age group 31 to 46.*

*A graph of different colored bars

Description automatically generated*

1 = Perfect correlation

-1 to 0 = Negative correlations

0 to 1 = Positive correlation

There is a Strong correlation between Age and Price.

There is a strong correlation between Partner salary and Total salary.

There is medium correlation between Salary and Price

There is medium correlation between Total salary and Price.

*A colorful squares with numbers

Description automatically generated*

***E. 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”***

***E2) Ned Stark believes that a 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.***

*SOLUTION –*

*Answer E1) Disagree. Men and women both prefer SUV. But Women prefer SUV by a larger margin than Men.*

*Answer E2) Agree. Based on the histogram plot from C), a salaried person is more likely to buy a sedan.*

***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***

***A screenshot of a computer

Description automatically generated***

***A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated***

As per above calculation females buy more cars than Male.

As per above calculation Gender vs Price Mean/Average, we can say Female has bought altogether more expensive car than Male .

**F2) Personal\_loan**

*SOLUTION –*

*With reference to below plot and calculations, customer who don’t take loan buy more expensive cars than who avails Personal loan*

*A screenshot of a graph

Description automatically generated*

*A screenshot of a graph

Description automatically generated*

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

*SOLUTION -*

We can conclude that it doesn’t matter if partner is working or not working Customer will buy their preferred car, there is although a marginal difference between them which slightly shows that customers whose partner is not working buy more expensive cars.

***A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated***

***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***

*SOLUTION –*

* *Using the crosstab function, we observe that Male customers buy more Hatchback and Sedan car, whereas as for SUV the female customers are higher than the Male Customers.*

*A screenshot of a computer

Description automatically generated*

* *According to the below calculations, there are zero sales of Hatchback cars amongst female customers with Business Profession.*
* *Sales for sedan cars amongst Female customers are on the lesser side than among Male Customers.*
* *Sales for SUV cars amongst Female customers are on the higher side than among Male Customers.*
* *As for the Salaried Professionals, the sales for Hatchback cars amongst Male customers are high than the female customers by a large scale.*
* *Female Salaried Customers tend to buy SUV cars most than Sedan or Hatchback.*

*A screenshot of a computer

Description automatically generated*

* *According to the below calculations, we can observe that, in general Married customers tend to buy Cars more than Single customers.*

*A screenshot of a computer

Description automatically generated*

###### **Problem 2**

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

*SOLUTION –*

*A screenshot of a computer program

Description automatically generated*

*Information about the dataset –*

*Shape – (8448,29)  
Total Rows – 8448*

*Total Columns ­­­­– 29  
Float Data Types – 1*

*Object Data Types – 8*

*Int Data Types – 19*

*DateTime Data Types – 1*

*There is missing data in the columns, “Transactor\_revolver”. There is also a unnamed column with no data in it at all.*

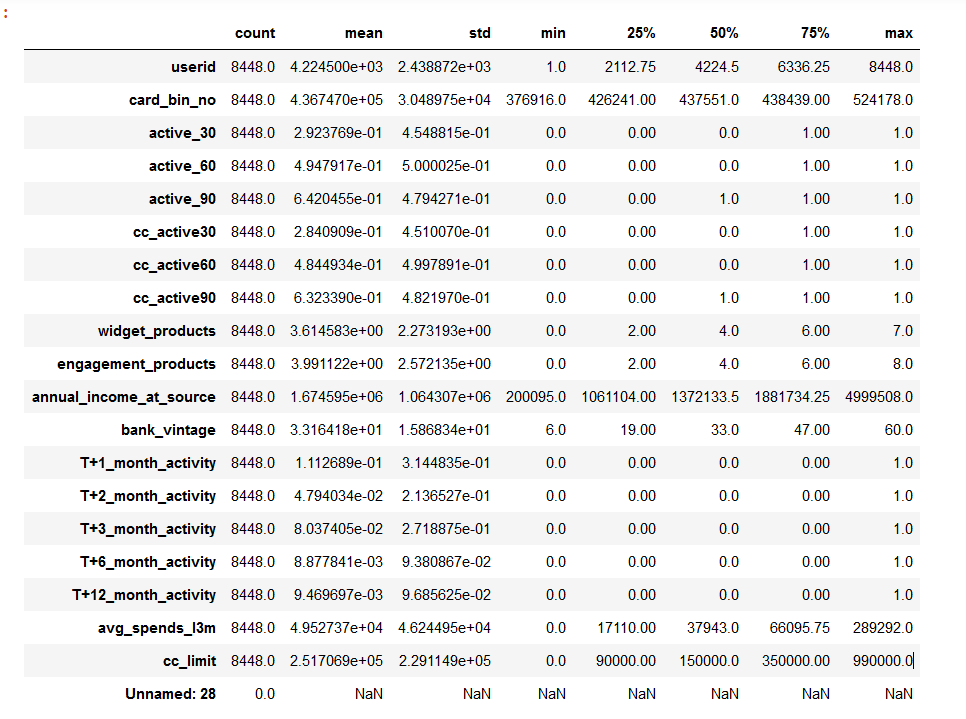
*Using fillna method to fill the missing values of the column, “Transactor\_revolver”. We can observe there are no null values in the dataset. The unnamed column has also been dropped from the dataframe.*

*A screenshot of a computer program

Description automatically generated*

*We are not going to do anything about outliers as it will affect the analysis.*

***Statistical Analysis of the data.***

**

* ***Most Important Variables in the dataset : -***

1. *card\_type:*

* Card \_types are the most important variable.
* Based on below data Salaried customer and high profile customer prefers to have more reward cards, card\_types helps to increase the Avg Spent .

A graph of different colored lines

Description automatically generated

1. *avg\_spends\_l3m:*

* *Avg spends is very important variable as this will give window to the bank for generating profits.*
* *These are directly proportional to the card\_type,occupation\_at\_source and Transactor\_revolver.*
* *This is also proportional to the interest received, more spending means more interest and thus becomes the mode for the revenue generation to the bank.*
* *Rise in income has positive relation with average spends, with increasing salary people tends to spend more.*

1. *Occupation\_at\_source:*

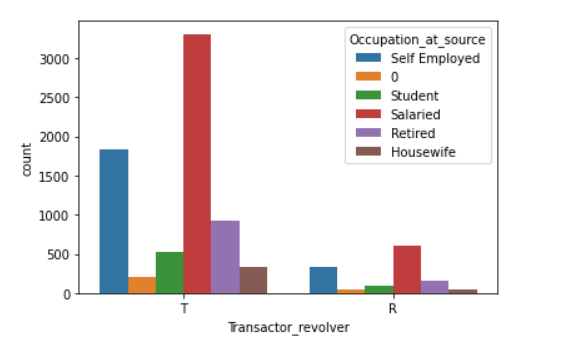
* *Bank should strategize to sell the correct card\_type as per the occupations which will increase the avg spends and will eventually increase the profits.*
* *Highest CC Customers are self Employed and salaried and their avg spent is high.*

*A graph of a number of people

Description automatically generated*

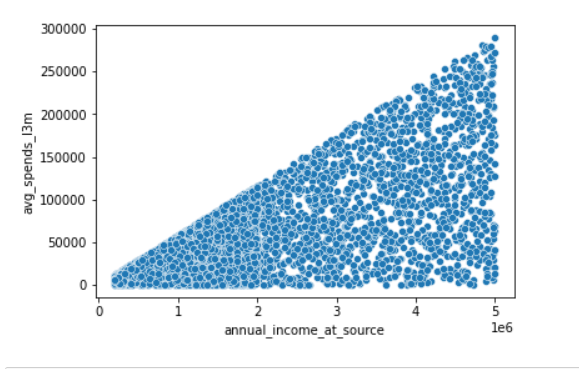
1. *Transactor\_revolver:*

* *High transaction customers’ need to be targeted for the revolver payments as customer can easily pay the amount and not being the defaulters.*
* *From the data given we can say Self-employed, salaried and students opt for revolver payment.*
* *This variable is important as high transactor, avg spending is more and uses revolver to do the payment which is beneficial for the bank to generate the profits.*

**

1. *annual\_income\_at\_source:*

* *Rise in income has positive relation with average spends, with increasing salary people tends to spend more.*

**