

**SMDM**

**Business Report**

**By: Dhruv Dosad**

|  |  |
| --- | --- |
| **Content** | **Page** |
| **Problem 1** | 03-11 |
| A. 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 |
| B. 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. | 3 |
| C. Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business. | 5 |
| 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. | 7 |
| 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. | 9 |
| 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. | 10 |
| G. From the current data set comment if having a working partner leads to the purchase of a higher-priced car. | 10 |
| 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. | 10 |
| **Problem 2** | 11-13 |
| Analyze the dataset and list down the top 5 important variables, along with the business justifications. | 11 |

**Problem-1 : Austo Motors**

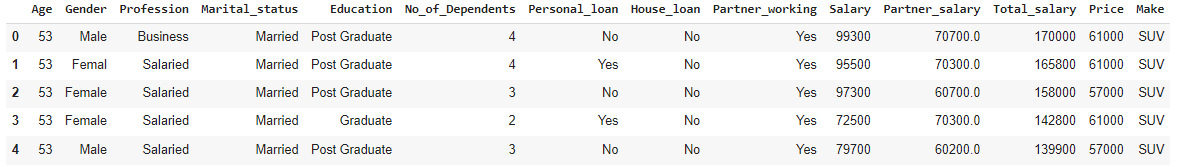
**Austo Motor Company is a leading car manufacturer specializing in SUV, Sedan, and Hatchback models. In its recent board meeting, concerns were raised by the members on the efficiency of the marketing campaign currently being used. The board decides to rope in an analytics professional to improve the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

 Dataset - [Link](https://olympus.mygreatlearning.com/courses/87091/files/7283021/download?verifier=D7dUkEA55E9I7udmucuVDJwWyscmRlirUlqSRaeI&wrap=1)

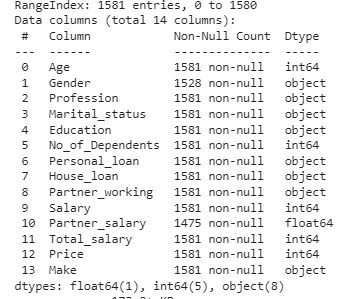
**A. What is the important technical information about the dataset that a database administrator would be interested in? (Hint: Information about the size of the dataset and the nature of the variables)**

**Size of Dataset**: Dataset has 1581 rows and 14 columns.

**Data headers** : PFB the data headers present in the dataset for quick reference

****

**Dataset Information:** There are 6 numerical and 8 categorical variables. PFB the details of each: -



**B. 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.**

**Inspecting Missing Values:** There are Null records present in two variables : Gender and Partner\_salary.

Gender - total 53 Nulls, Partner\_salary - Total 106 Nulls

Table

Description automatically generated

**Treating the Null values: -**

1. **Gender:** Null values in Gender field can be imputed with ‘**Male’** having as the **mode** (maximum value in the dataset)
2. **Partner**\_**salary** : Non-null values in Partner\_salary field is possible only if the variable Partner\_working is YES. Hence for this data we do a rule based imputation instead of the mean/median imputation – If Partner\_working = ‘No’ then Partner\_salary = 0

If Partner\_working = ‘Yes’ then Partner\_salary = Total\_salary – Salary

**Duplicate Values**: There are **no duplicate** records in the dataset.

**Bad Values**: Bad values are present in Gender as **Femal** or **Femle** . Rest of the categorical fields seem to be fine.

Table

Description automatically generated with low confidence

We will be treating the above by **replacing** the values Femal or Femle with **Female**

**Inspecting the Summary Statistics of the Dataset**

**Table

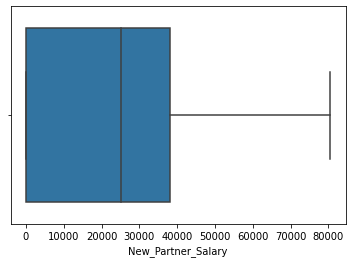
Description automatically generated**

1. The customer’s age is between **22** and **54** years old i.e. majority might belong to **working age** group. **Mean age is 31.92** while **median age is 29** years, indicating age distribution is **positively skewed.**
2. The **Salary** of the customers **ranges between 30K and 99.3K** and the **distribution is** **symmetric**. The close mean and the median shows **skewness** is near to 0.
3. **Total\_salary** ranges between **30K and 171K** and does not show a high degree of skewness.
4. The **minimum price** of the **purchased automobile is 18K**, whereas **max is 70K.** Skewness indicates a small number of high priced purchases were made.

**Checking Outliers in the numerical variable**

**Chart

Description automatically generated Chart

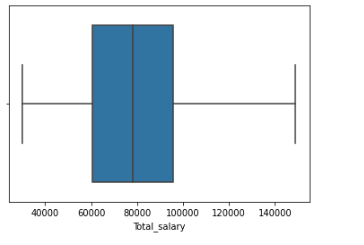
Description automatically generated Chart

Description automatically generated Chart

Description automatically generated**

* There are **no negative** values present in the numerical fields.
* From the boxplots we can observe **outlier values are present in Total\_salary** variables.
* Outliers are treated by using **Winsorization**, i.e. bringing the larger outliers (Data points above the Q3 + 1.5 \*IQR value) to the upper whisker

**Boxplot after Outlier treatment**



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

**Univariate Analysis of Numerical fields**

**Chart, histogram

Description automatically generated**

**Univariate Analysis of Categorical fieldsChart, bar chart

Description automatically generated**

**Inferences**

- **Majority** of the customers in the dataset are **Post Graduate**.

-  **Sedan** is the **most preferred** **purchase**, followed by **Hatchback** and **SUV**

- **Salaried customer** **count** is **slightly higher** than that of **Business customers**.

- The number of **customers** having a **working partner** are **slightly higher** than customers with **non-working partner or singles**.

- **Majority** of the customers have **either 2 or 3 dependents**, followed by **1 or 4** dependents. Very few customers have zero no of dependents.

**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.**

**Bivariate analysis of Numerical variables**

Pair plot on the Data set: -

**Diagram, engineering drawing

Description automatically generated**

**Chart

Description automatically generated**

Inferences –

1. Hardly any linear relationships present among the fields.
2. **Positive correlation** between **Price and Age**, and **Total\_salary and New\_Partner\_salary**

**Bi- Variate analysis of Categorical vs Categorical variables –**

**Chart, bar chart

Description automatically generated**

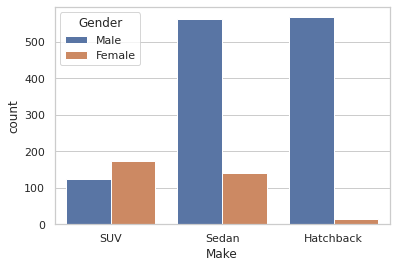
**Inferences –**

1. Customers who have a **house loan** are **not likely to buy an SUV** (which is the costliest make among the three).
2. **Females prefer SUV** and are **least likely to buy a Hatchback**, whereas **Male prefer Sedan or hatchback**. **SUV** is **least preferable** among **males.**
3. **Married** customers **prefer Sedan** whereas **single** customers **prefer Hatchbacks**.

**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”**

Analyzing the ratio of SUV purchases for both the Genders, we get:

 Table

Description automatically generated

From above data, we can **conclude** that the **statement** made by **Steve Rogers is Incorrect**

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

Analyzing the ratio of **Sedan** purchases against **profession**, we get:

**Chart, bar chart

Description automatically generated** **Table

Description automatically generated**

From above data, we can **conclude** that the **statement** made by **Ned Stark is Correct**

**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.**

Analyzing the ratio of **SUV** purchases against **profession**, we get:

**Chart, bar chart

Description automatically generated** **Table

Description automatically generated**

From above data, we can **conclude** that the **statement** made by **Sheldon Cooper is Incorrect**

**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**

Females are more likely to buy SUVs and on an average spend more on cars than males 47705 Units against32416 Units.

Mean of Price across Gender:

Female = 47705

Male = 32416

Median of Price across Gender:

Female = 49000

Male = 29000

**Mean and Median** **Price** for **Female** customers is **higher than Male** **customers.**

**F2) Personal\_loan**

Mean of Price across Personal Loan:

Personal Loan: No= 36742

Personal Loan: Yes= 34457

Median of Price across Personal Loan:

Personal Loan: No= 32000

Personal Loan: Yes= 31000

**Mean and Median of Price** for purchase made by customers **without a Personal loan** is **slightly higher** than **customers** **who have a Personal Loan**.

**To ensure increased spend** of customers with Personal loans, the **business can look at cheaper interest rate s** (for Automobile purchase) or **easy the repayment** **terms**.

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

**Mean of Price across Partner\_working:**

Partner\_working: No = 36000

Partner\_working: Yes = 35267

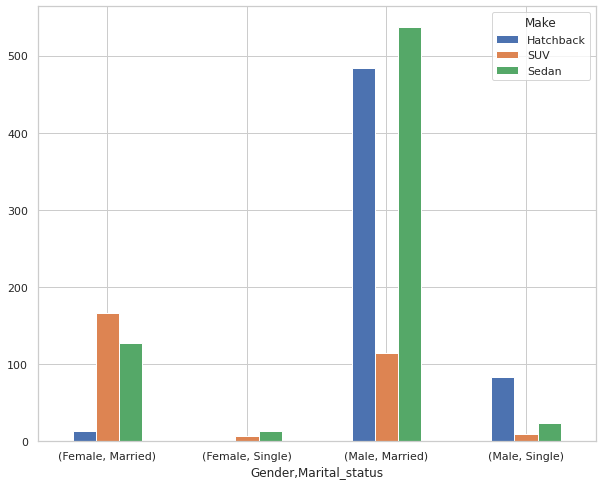
**Median of Price across Partner\_working:**

Partner\_working: No = 31000

Partner\_working: Yes = 31000

**The Mean and Median price** of the purchased automobile is **almost similar across the Partner\_working category**, thus indicating whether **partner is working or not,** it has **no impact** on the **Purchase made by the customer.**

**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.**

** Graphical user interface

Description automatically generated with low confidence**

**Most frequently purchased** Car make grouped on Marital\_Status and Gender, we find :

Female – Married: SUV

Male – Married: SUV

Analyzing the **mean Price of purchased car** across the Marital\_status and Gender, we find:

Mean Price for purchases made by Married Females = 62857

Mean Price for purchases made by Married Males = 60692

* **Mode** of the Car make for Gender and Marital\_status fields shows that **both the married groups preferring SUV**.
* Similarly, the **Mean of Price** for **Male Married is approx. 60K** while it is **62K for Female Married**.
* All the **Male Married Customers** with **Total Salary greater than 149 K** **purchased SUV**. Whereas **Married male** with **lower Total\_salary preferred Sedan**

**Problem-2 GODIGT Bank**

**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.**

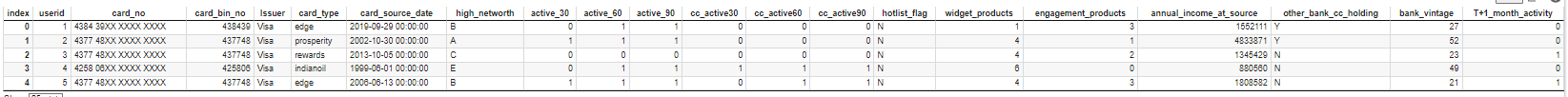
**GODIGT Bank is a mid-sized private bank that deals in all kinds of banking products, such as savings accounts, current accounts, investment products, etc. among other offerings. The bank also cross-sells asset products to its existing customers through personal loans, auto loans, business loans, etc., 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.)**

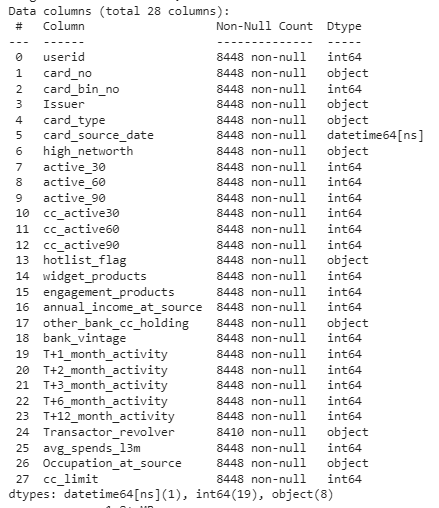
**Framing An Analytics Problem - Analyze the dataset and list down the top 5 important variables, along with the business justifications.**

**- Size of Dataset**: Dataset has 8448 rows and 28 columns.

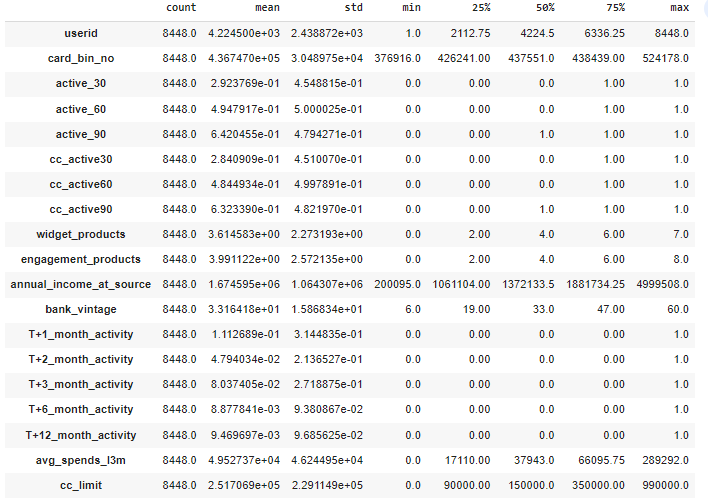
**- Data headers** : PFB the data headers present in the dataset for quick reference

****

**- Dataset Information:** There are 19 numerical and 8 categorical variables. PFB the details of each: -



**Five Points Summary:**

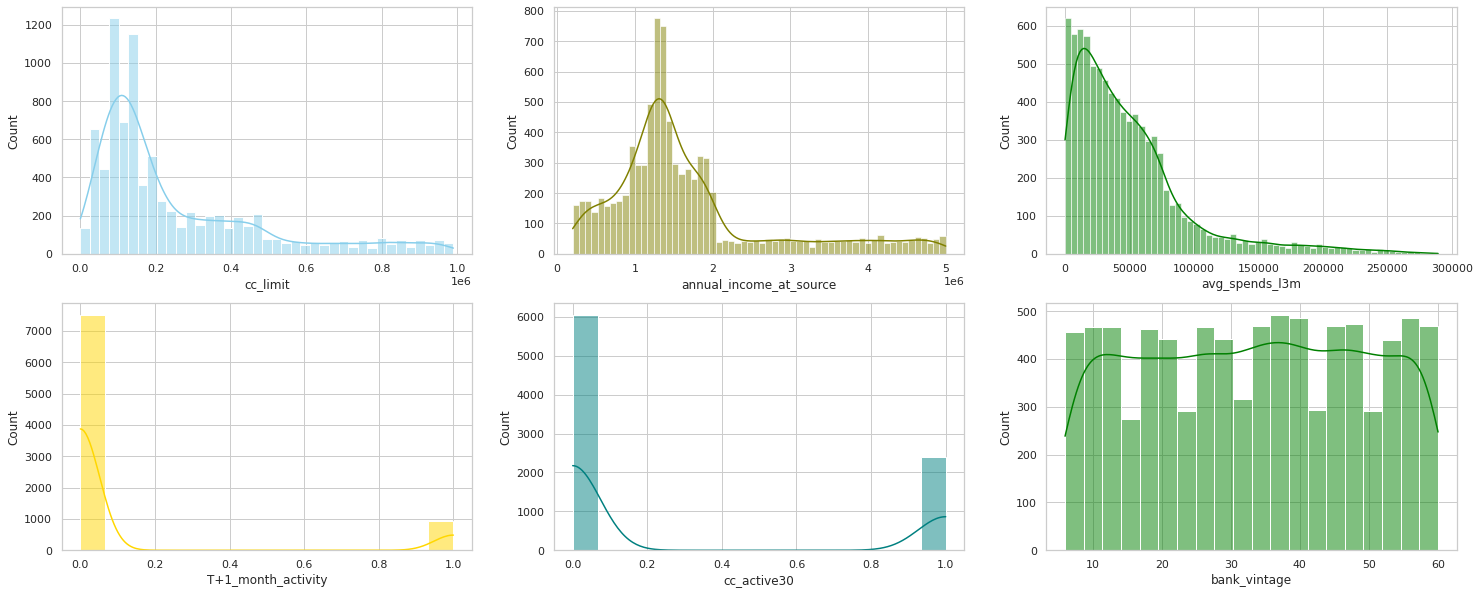


**Exploring the data: -**

**- No duplicate values** found

- In **Transactor\_revolver** we found **38 null values**

**Histograph for the numerical values**



**Below are the Top 5 important variables from the given dataset with justification.**

**1) Annual Income at source-**

Annual income plays a big role in the purchasing power of an individual hence is a vital piece of info. Income can be used by the banks to make better decisions in areas such as risk profiling, targeted ads, campaigns, offers, loan limits etc.

**2) CC\_limit -**

Defining Credit Card limit for customers basis their attributes (such as income, CIBIL Score, etc.) is part of the Risk Management practice wherein the banks try to minimize the number of defaulters. The banks seek a quantifiable answer to the query “How much is too much?”

**3) CC\_active30 –**

Flag variables such as cc\_active30, cc\_active60 can be used to get an understanding over how frequently does the customer use the credit card, if the account is dormant or if the customer is experiencing any issues leading to reduced usage of the card etc.

**4) T+1\_month\_activity-**

Flag variables such as T+1\_month\_activity can be used to plan out campaigns and promotional offers so as to increase activity in the credit card.

**5) avg\_spends\_13m-**

The avg\_spends\_l3m variable can give important insights on the customer spending behavior. It can be used to identify whether the credit card is primary or secondary card of customer, i.e. high spend indicates primary account whereas lower spend would mean secondary account. Campaigns can be rolled out on the basis of the customer preference, customized offers can begiven to lure customers into using the credit account more frequently.

-------------------------End of Report------------------------