Contents

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.

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)

Import some libraries like Numpy, Pandas, Seaborn, Matplotlib, etc. After that, load our data set, austo.csv, and use the head() function to view the data. Using the shape function, we can determine that there are 1581 rows and 14 columns. Find out the characteristics of the columns using the info() method. The datatypes for the float64(1), int64(5), and object(8) columns are present. There are some null values, but there aren't any duplicate values.

- **head()** it given by default top five data

	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary	Price	Make
0	53	Male	Business	Married	Post Graduate	4	No	No	Yes	99300	70700.0	170000	61000	SUV
1	53	Femal	Salaried	Married	Post Graduate	4	Yes	No	Yes	95500	70300.0	165800	61000	SUV
2	53	Female	Salaried	Married	Post Graduate	3	No	No	Yes	97300	60700.0	158000	57000	SUV
3	53	Female	Salaried	Married	Graduate	2	Yes	No	Yes	72500	70300.0	142800	61000	SUV
4	53	Male	Salaried	Married	Post Graduate	3	No	No	Yes	79700	60200.0	139900	57000	SUV

Table 1: Top five data of dataset austo

- **shape** it tells numbers of rows and columns in given dataset.

```
austo.shape
```

```
(1581, 14)
```

Table 2: 1581 rows & 14 columns

- **info()** it tells a concise summary of a DataFrame

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1581 entries, 0 to 1580
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Age               1581 non-null   int64
 1   Gender            1528 non-null   object
 2   Profession        1581 non-null   object
 3   Marital_status    1581 non-null   object
 4   Education         1581 non-null   object
 5   No_of_Dependents  1581 non-null   int64
 6   Personal_loan     1581 non-null   object
 7   House_loan        1581 non-null   object
 8   Partner_working   1581 non-null   object
 9   Salary            1581 non-null   int64
 10  Partner_salary    1475 non-null   float64
 11  Total_salary      1581 non-null   int64
 12  Price             1581 non-null   int64
 13  Make              1581 non-null   object
dtypes: float64(1), int64(5), object(8)
memory usage: 173.0+ KB
```

Table 3: Information about the austo structure and content.

- 8 Categorical Variables (5 are Binary and 3 are Multi-level).
- 4 Continuous Variables
- 2 Discrete Variables.
- Gender and Partner_salary have Null Values.
- No Duplicates Values.

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.

Yes, there are discrepancies in the Gender and Partner_salary columns. Both columns have null values, so, imputed the values.

```
Male      1199
Female     327
Femal        1
Femle        1
Name: Gender, dtype: int64
```

Table 4: Counts the Gender Column values

- There is spelling incorrect for Female, so correct it first, and then replace null values with Male because the accuracy of Male is maximum.

- According to data know, Total_Salary=Salary+Partner_Salary, so Partner_Salary=Total_Salary-salary based on Partner_working status is Yes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1581 entries, 0 to 1580
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Age               1581 non-null   int64
 1   Gender            1581 non-null   object
 2   Profession        1581 non-null   object
 3   Marital_status    1581 non-null   object
 4   Education         1581 non-null   object
 5   No_of_Dependents  1581 non-null   int64
 6   Personal_loan     1581 non-null   object
 7   House_loan        1581 non-null   object
 8   Partner_working   1581 non-null   object
 9   Salary            1581 non-null   int64
 10  Partner_salary    1581 non-null   float64
 11  Total_salary      1581 non-null   int64
 12  Price             1581 non-null   int64
 13  Make              1581 non-null   object
dtypes: float64(1), int64(5), object(8)
memory usage: 173.0+ KB
```

Table 5: Information about the austo structure and content after Null values treatment

- Now there are no discrepancies in the dataset

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

- **Describe()** it tells summary of the central tendency, dispersion, and shape of the distribution of the data.

	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
count	1581.000000	1581.000000	1581.000000	1581.000000	1581.000000	1581.000000
mean	31.922201	2.457938	60392.220114	19233.776091	79625.996205	35597.722960
std	8.425978	0.943483	14674.825044	19670.391171	25545.857768	13633.636545
min	22.000000	0.000000	30000.000000	0.000000	30000.000000	18000.000000
25%	25.000000	2.000000	51900.000000	0.000000	60500.000000	25000.000000
50%	29.000000	2.000000	59500.000000	25100.000000	78000.000000	31000.000000
75%	38.000000	3.000000	71800.000000	38100.000000	95900.000000	47000.000000
max	54.000000	4.000000	99300.000000	80500.000000	171000.000000	70000.000000

Table 6: Descriptive statistics of an austo

- Age groups between 22-54 are working customers. Age group between 22-25 (Aporx): partners are non-working.
- Salary increases positively between 30000-99300 for age groups between 22-54.
- Car prices range from 18000–70000.

➢ **Univariate Analysis with Numerical Variables** it tells valuable insights into the characteristics and behaviour of individual variables in a dataset.

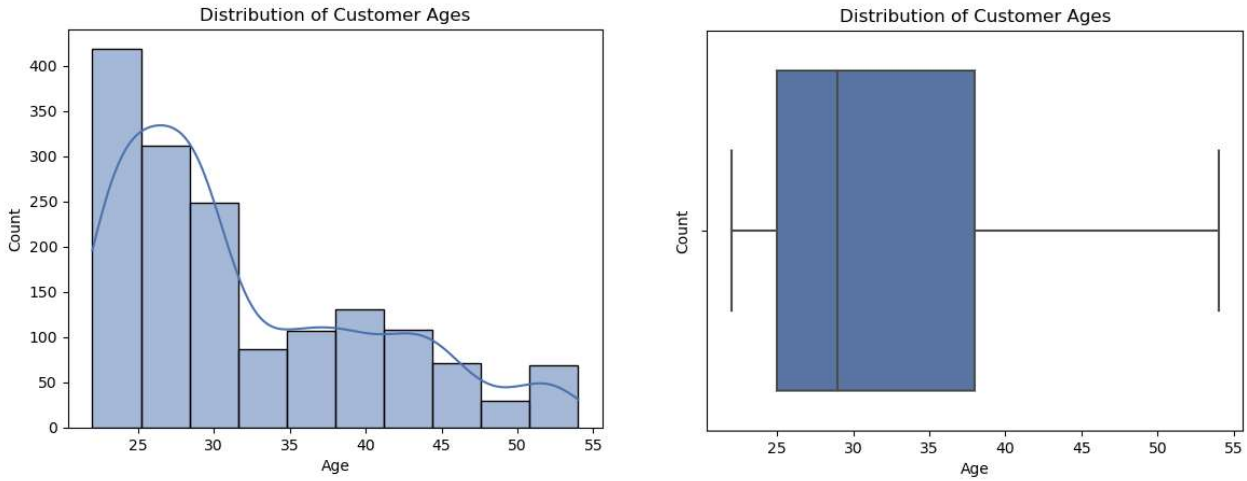

Fig 1: Histogram and Boxplot of Distribution of customer's age

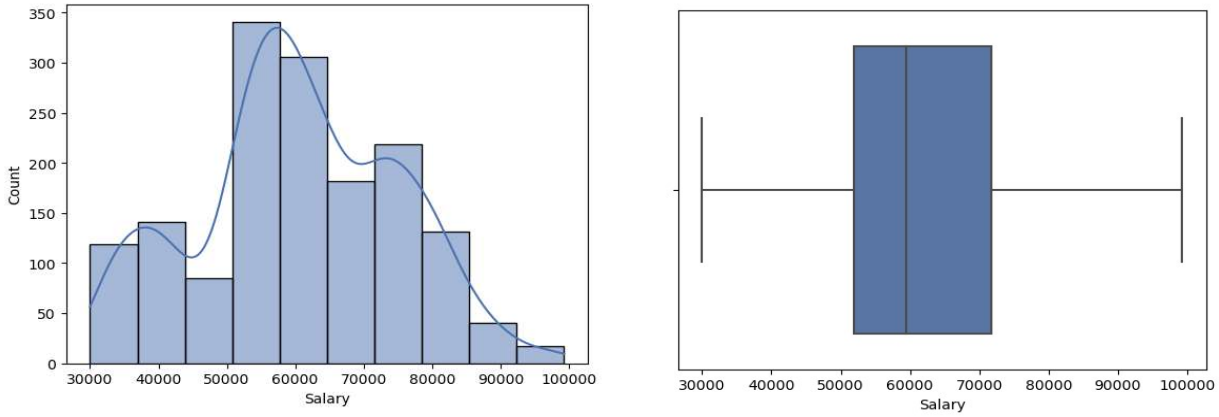

Fig 2: Histogram and Boxplot of Distribution of customer's salary

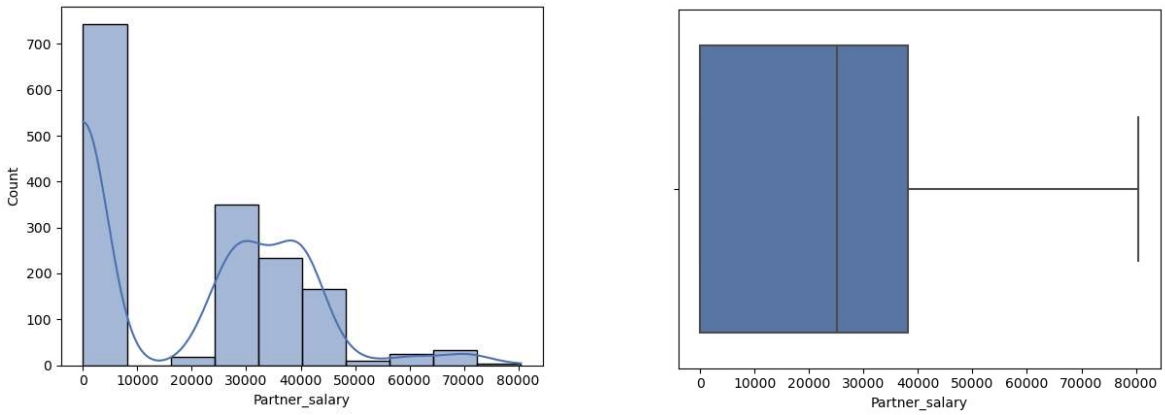

Fig 3: Histogram and Boxplot of Distribution of customer's Partner salary

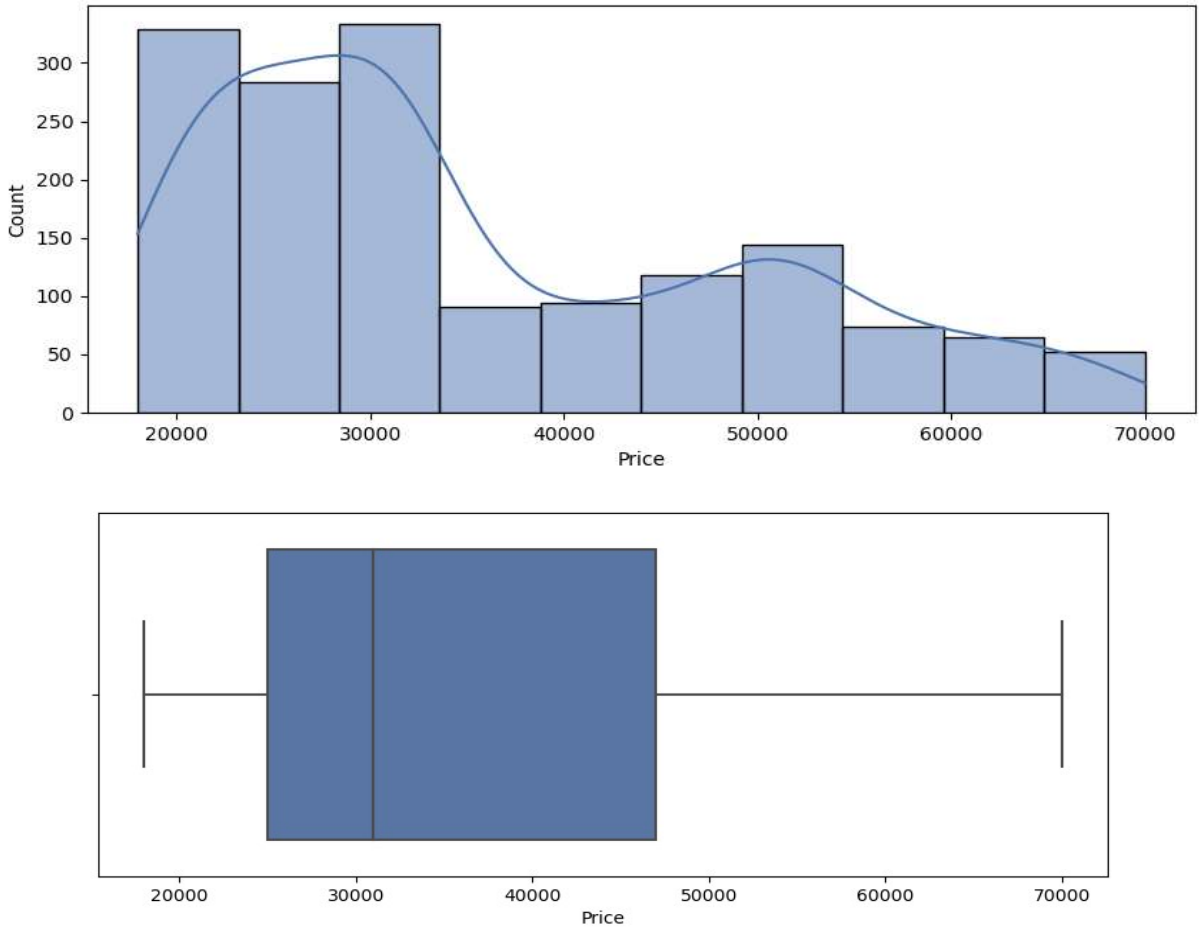

Fig 4: Histogram and Boxplot of Distribution of customer's Price

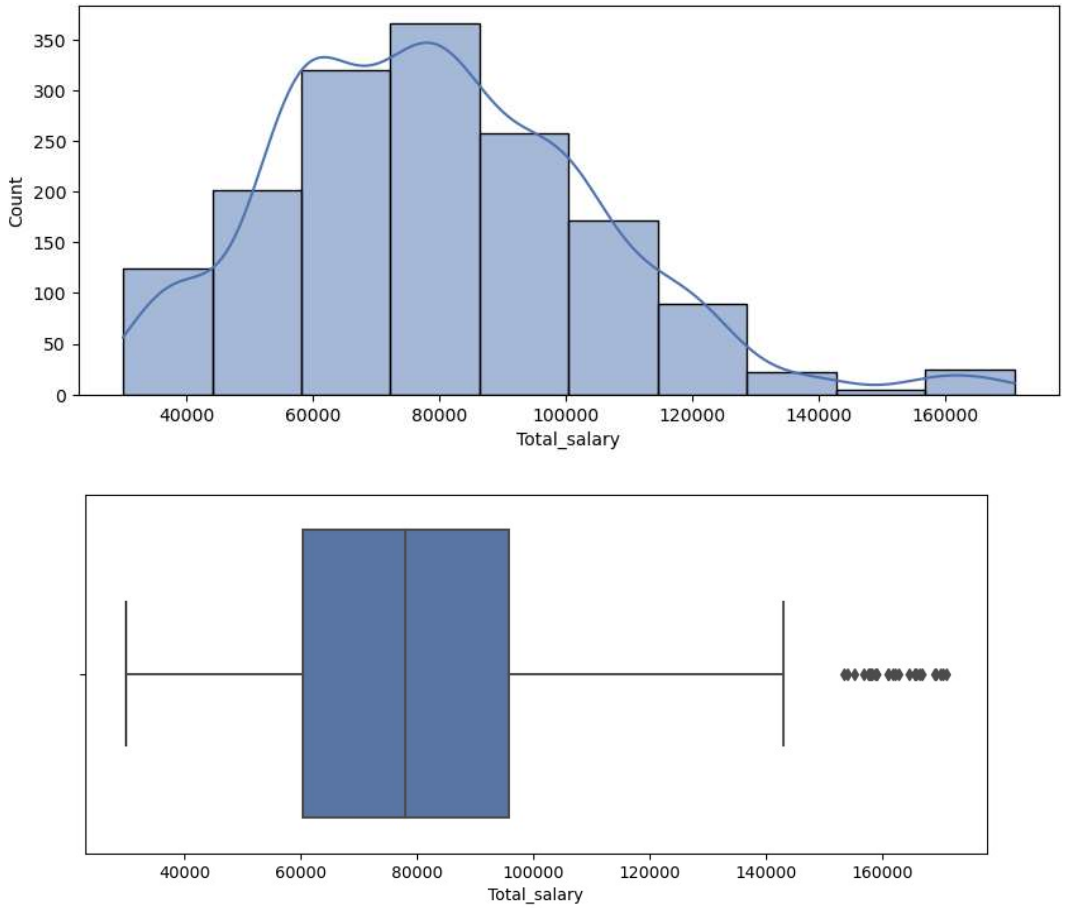

Fig 5: Histogram and Boxplot of Distribution of customer's Total salary with outliers

- No negative values present in Numerical variables.
- There are 27 outliers parents in Total salary.
- For outlier treatment apply IQR rule. (3 times IQR rule is taken into consideration to prevent a substantial loss of data if more than 25-30 % of the records are outside the 1.5 times IQR rule's defined range.)

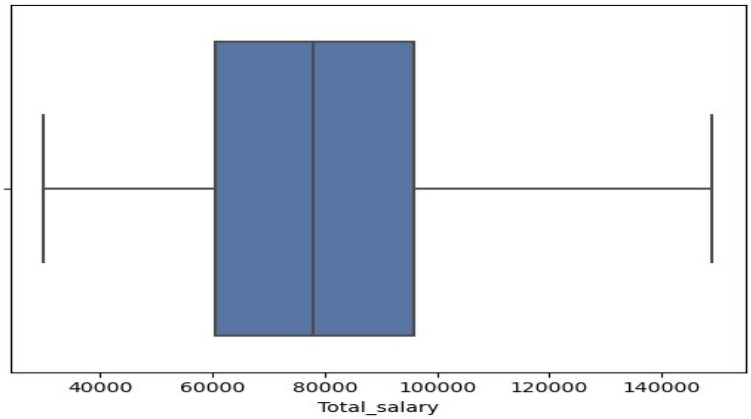

Fig 6: Boxplot of Distribution of customer's Total salary without outliers

➢ **Univariate Analysis with Categorical Variables** it tells valuable insights into the characteristics and behaviour of individual variables in a dataset.

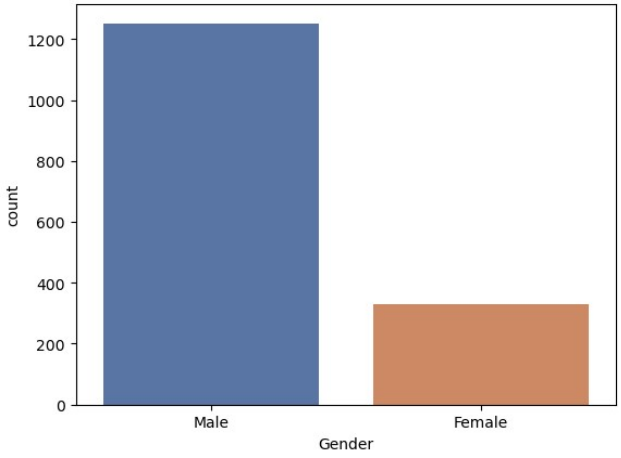

Fig 7: Count plot of Gender

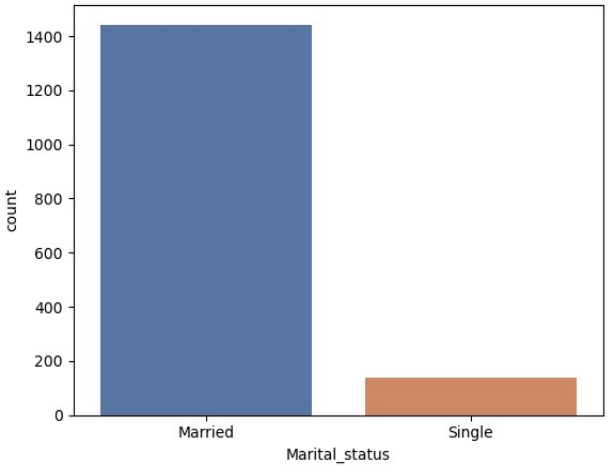

Fig 8: Count plot of Marital status

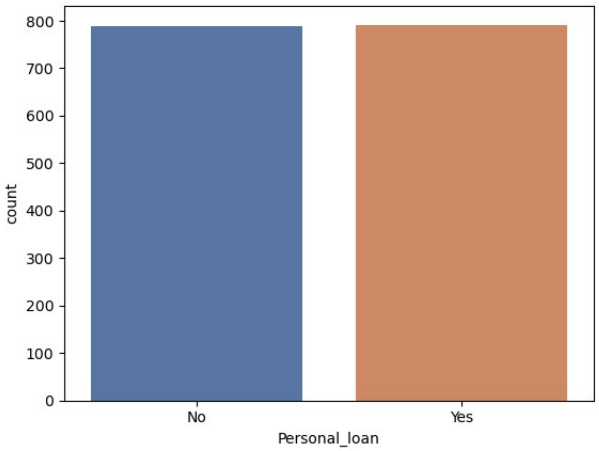

Fig 9: Count plot of Personal loan

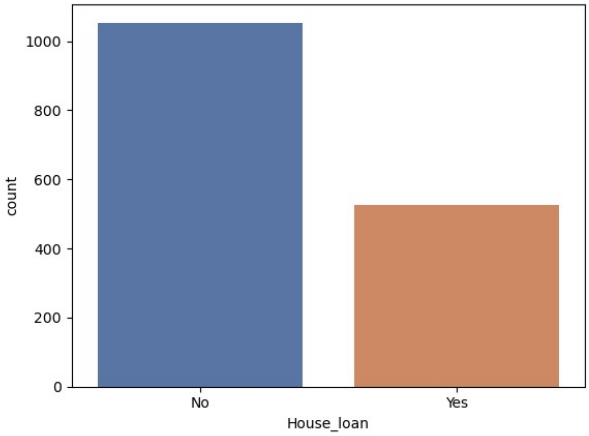

Fig 10: Count plot of House loan

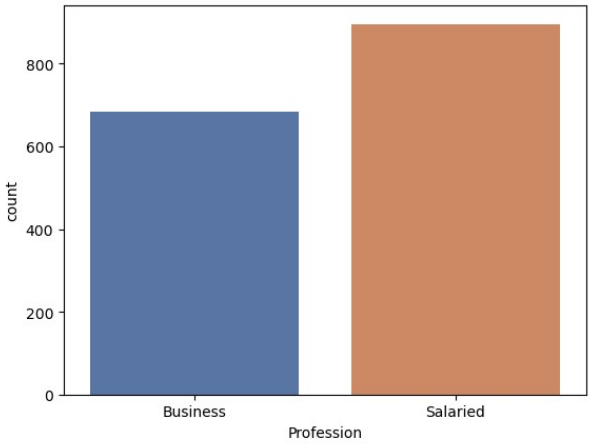

Fig 11: Count plot of Profession

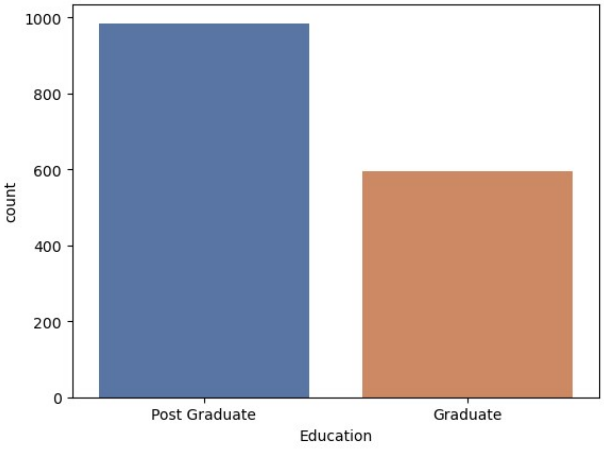

Fig 12: Count plot of Education

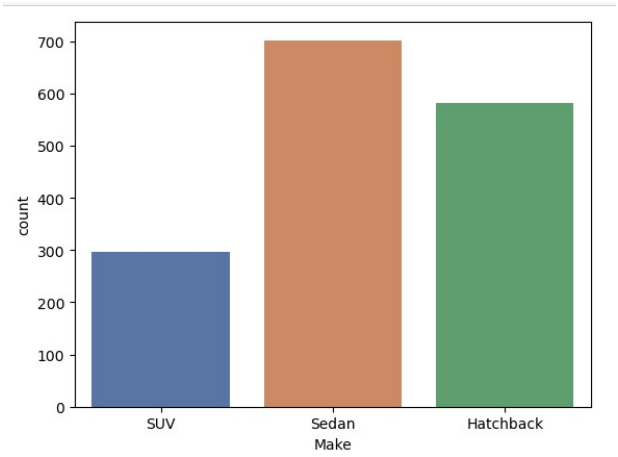

Fig 13: Count plot of Profession

- Number of Male customers are more than Female customers.
- Married customers are more than Single customers.
- Approx. 500 customers have a house loan or almost double the number of customers who do not have a house loan.
- Salaried customers count are slightly higher than the business customers
- Postgraduate customers' majority is higher in the dataset.
- A Sedan is the most preferred purchase car rather than a hatchback or SUV.

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.

- **Perform Bivariate Analysis Categorical Variables** it tell the relationship between two categorical variables and identifies any patterns or dependencies.

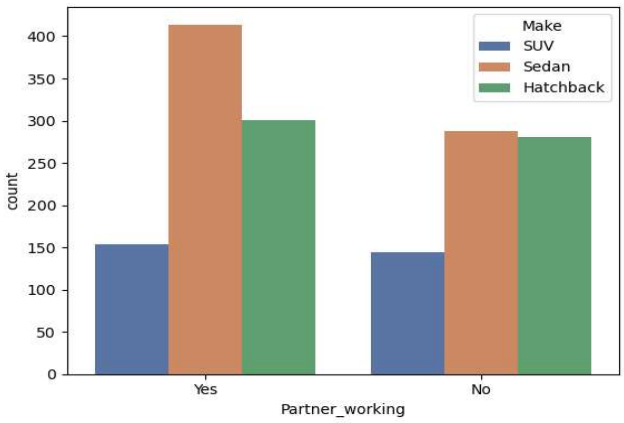

Fig 14: Count plot of Partner working

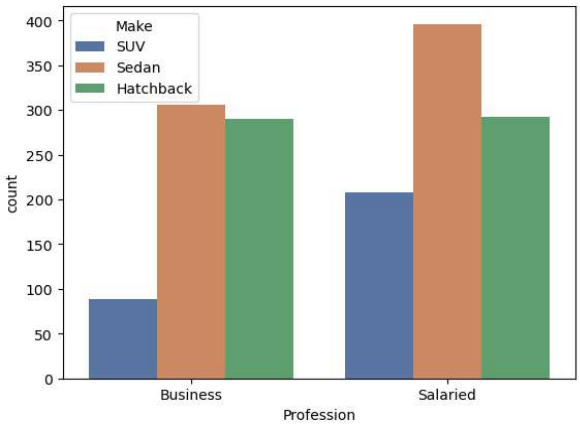

Fig 15: Count plot of Profession

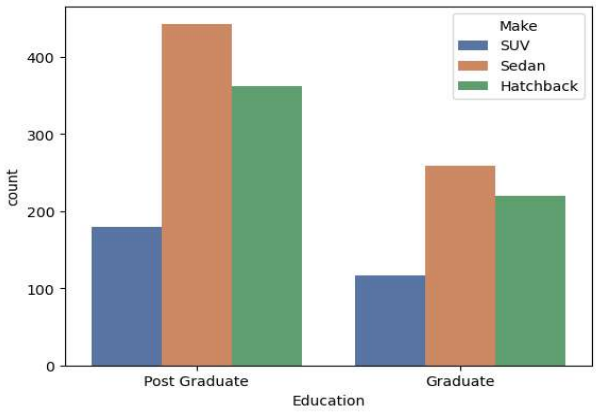

Fig 16: Count plot of Education

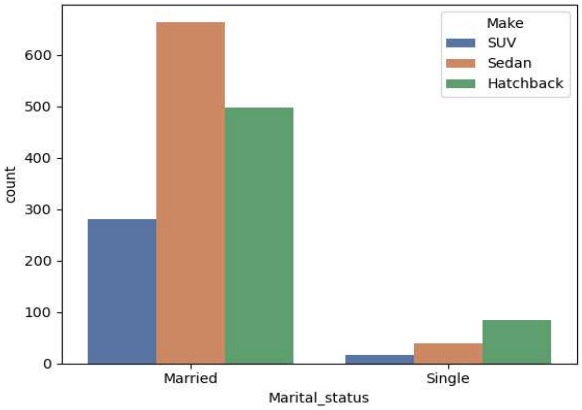

Fig 17: Count plot of Marital status

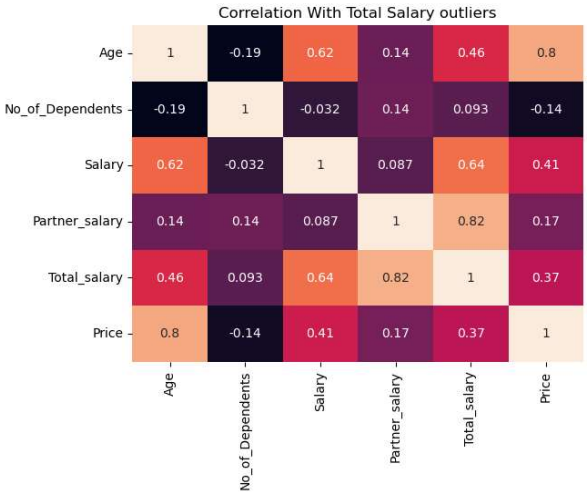

Fig 18: Hitmap of Distribution of customer's Total salary with outliers

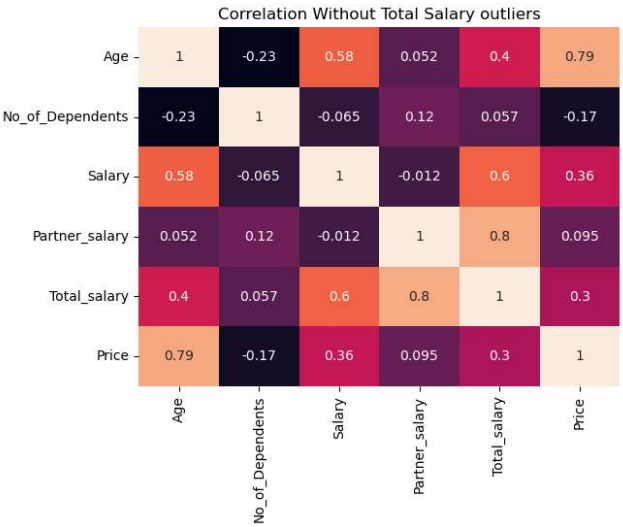

Fig 19: Hitmap of Distribution of customer's Total salary without outliers

- Customers who have loans are less likely to choose an SUV—the most expensive model of the three. In both categories, a sedan is highly favoured.
- Males tend to choose sedans or hatchbacks, whereas females favour SUVs and are least likely to purchase one. SUVs outsell sedans in favour of men.
- Customers who are married favour sedans while those who are single favour hatchbacks.
- The fields hardly have any linear relationships with one another.
- Price and age (with outliers) and total salary and partner salary (without outliers) had the highest positive association.

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.

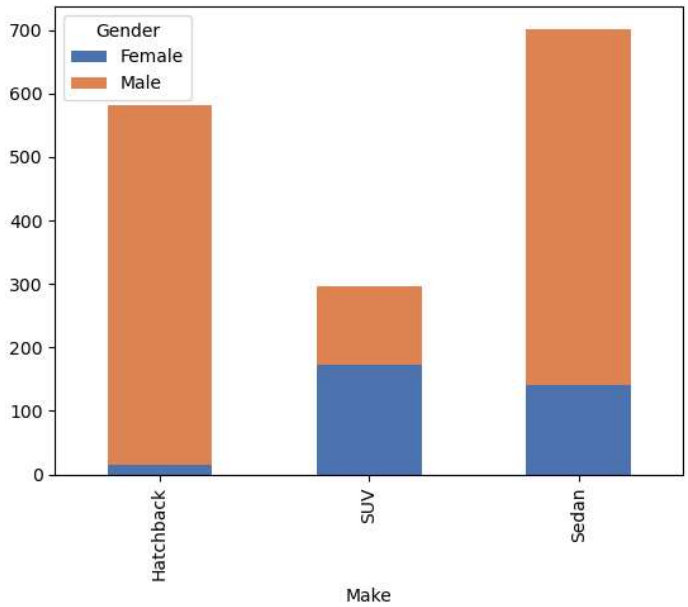

Fig 20: Bivariate Analysis between Gender and Make

- Steve Roger's statement is wrong because, Women prefer SUVs by a large margin, compared to Men.
- SUV make are the preferred cars by Females, followed by Sedans and then Hatchback.

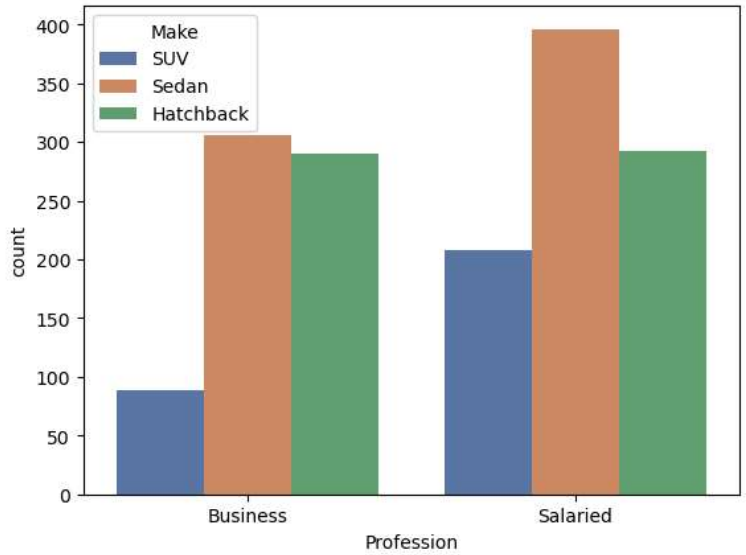

Fig 21: Bivariate Analysis between Profession and Make

- Ned Stark's believe is true because salaried person is more likely to buy a Sedan

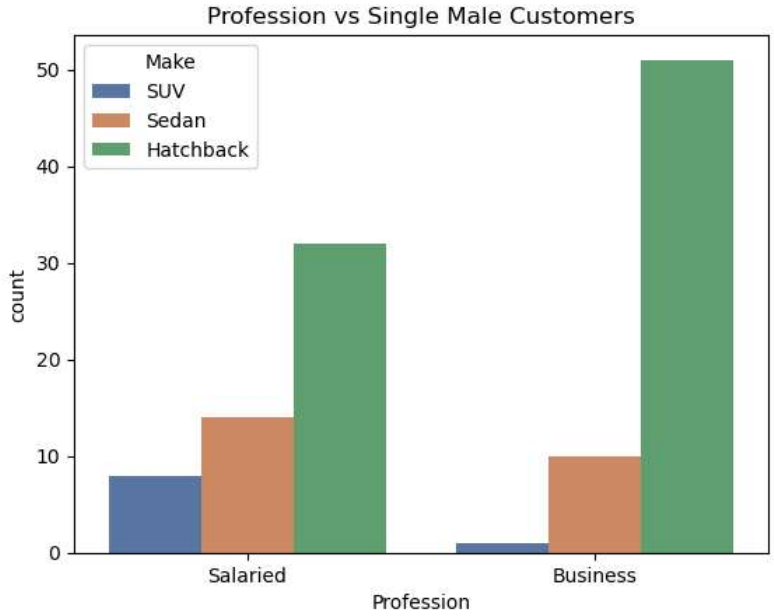

Fig 22: Bivariate Analysis Profession $^{V}/_{S}$ Single Male Customers

- Sheldon Cooper's claim is wrong because Salaried male is an easier target for a Sedan sale over an SUV 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

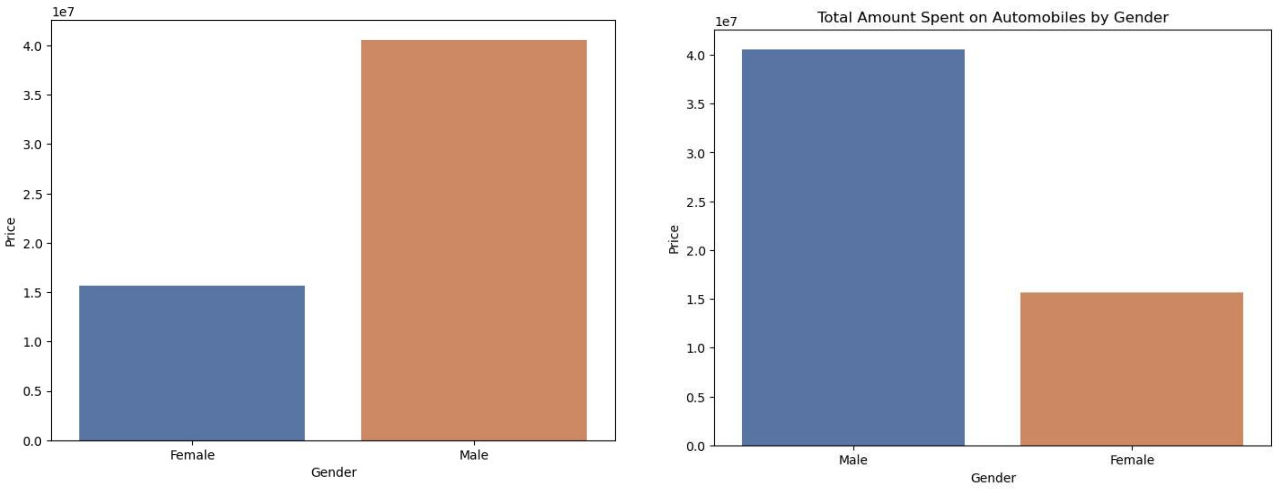

Fig 23: Barplot of Price and Gender

- Male spent more amounts in purchasing car than Female.
- However, when we break down spending by gender, women spend more than men do. Both those with and without personal loans spend almost the same amount on automobile purchases.

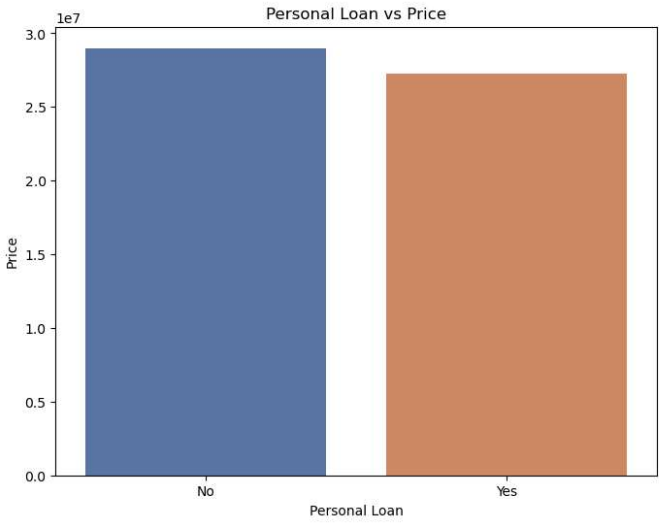

Fig 24: Barplot of Price and Personal Loan

- Both those with and without personal loans spend almost the same amount on automotive purchases.

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

- Average price of cars for customers with a working partner: 35267.28110599078
- Average price of cars for customers without a working partner: 36000.0
- Based on the current dataset, having a working partner is not necessarily associated with the pur chase of a higher-priced car.

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.

	Gender	Marital_status	Make
0	Female	Married	SUV
1	Female	Single	Sedan
2	Male	Married	Sedan
3	Male	Single	Hatchback

Table 7: Group the 'Gender" and "Marital Status" and find the preferred car.

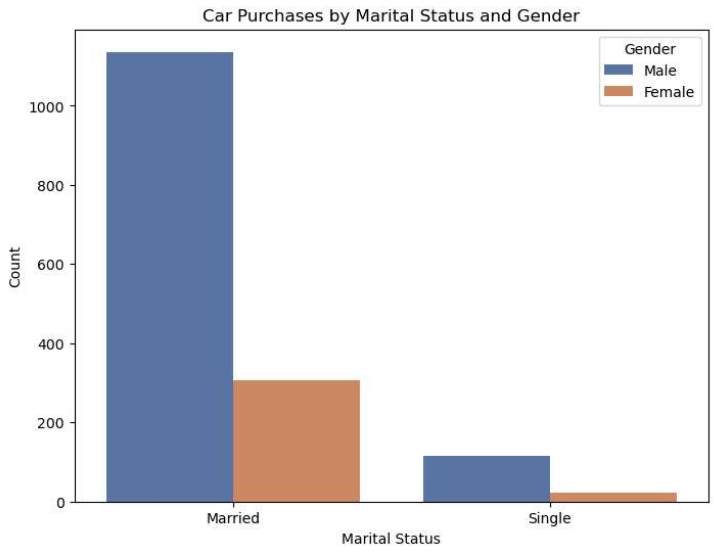

Fig 25: Countplot of Marital Status and Gender

- Based on the past numbers that have occurred the most frequently, we may divide the data into four groups and give each group a different automobile make.
    1. Married Male Preferred Sedan
    2. Married Female Preferred SUV
    3. Single Female Preferred Sedan
    4. Single Male Preferred Hatchback

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.)

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

userid	Unique bank customer id
card_no	Masked credit card number
card_bin_no	Credit card IIN number
Issuer	Card network issuer
card_type	Credit card type
card_source_date	Credit card sourcing date
high_networth	Customer category basis their networth value (A: High to E: Low)
active_30	Savings/Current/Salary etc account activity in last 30 days
active_60	Savings/Current/Salary etc account activity in last 60 days
active_90	Savings/Current/Salary etc account activity in last 90 days
cc_active30	CC activity in last 30 days
cc_active60	CC activity in last 60 days
cc_active90	CC activity in last 90 days
hotlist_flag	Whether card is hotlisted
widget_products	Number of convenient product customer holds (dc, cc, netbanking active, mobile banking active, wallet active etc)
engagement_products	Number of investment/loan product customer holds (FD, RD, Personal loan, auto loan etc)
annual_income_at_source	Annual income recoreded in credit card application
other_bank_cc_holding	Hold other bank credit card
bank_vintage	Vintage with the bank (in months) as on Tth month
T+1_month_activity	Customer spends next (T) month using credit card
T+2_month_activity	Customer spends in T+2 month using credit card
T+3_month_activity	Customer spends next month using credit card
T+6_month_activity	Customer spends next month using credit card
T+12_month_activity	Customer spends next month using credit card
Transactor_revolver	Revolver: Customer who carries balances over from one month to the next. Transactor: Customer who pays off their balances in full every month.
avg_spends_l3m	Average credit card spends in last 3 months
Occupation_at_source	Occupation recorded at the time of credit card application
cc_limit	Current credit card limit

**All above data has been recorded as on T^{th} month excluding T+1_month_activity, T+2_month_activity, T+3_month_activity, T+6_month_activity, T+12_month_activity*

Fig 26: DataSet Description

	count	mean	std	min	25%	50%	75%	max
userid	8448.0	4.224500e+03	2.438872e+03	1.0	2112.75	4224.5	6336.25	8448.0
card_bin_no	8448.0	4.367470e+05	3.048975e+04	376916.0	426241.00	437551.0	438439.00	524178.0
active_30	8448.0	2.923769e-01	4.548815e-01	0.0	0.00	0.0	1.00	1.0
active_60	8448.0	4.947917e-01	5.000025e-01	0.0	0.00	0.0	1.00	1.0
active_90	8448.0	6.420455e-01	4.794271e-01	0.0	0.00	1.0	1.00	1.0
cc_active30	8448.0	2.840909e-01	4.510070e-01	0.0	0.00	0.0	1.00	1.0
cc_active60	8448.0	4.844934e-01	4.997891e-01	0.0	0.00	0.0	1.00	1.0
cc_active90	8448.0	6.323390e-01	4.821970e-01	0.0	0.00	1.0	1.00	1.0
widget_products	8448.0	3.614583e+00	2.273193e+00	0.0	2.00	4.0	6.00	7.0
engagement_products	8448.0	3.991122e+00	2.572135e+00	0.0	2.00	4.0	6.00	8.0
annual_income_at_source	8448.0	1.674595e+06	1.064307e+06	200095.0	1061104.00	1372133.5	1881734.25	4999508.0
bank_vintage	8448.0	3.316418e+01	1.586834e+01	6.0	19.00	33.0	47.00	60.0
T+1_month_activity	8448.0	1.112689e-01	3.144835e-01	0.0	0.00	0.0	0.00	1.0
T+2_month_activity	8448.0	4.794034e-02	2.136527e-01	0.0	0.00	0.0	0.00	1.0
T+3_month_activity	8448.0	8.037405e-02	2.718875e-01	0.0	0.00	0.0	0.00	1.0
T+6_month_activity	8448.0	8.877841e-03	9.380867e-02	0.0	0.00	0.0	0.00	1.0
T+12_month_activity	8448.0	9.469697e-03	9.685625e-02	0.0	0.00	0.0	0.00	1.0
avg_spends_l3m	8448.0	4.952737e+04	4.624495e+04	0.0	17110.00	37943.0	66095.75	289292.0
cc_limit	8448.0	2.517069e+05	2.291149e+05	0.0	90000.00	150000.0	350000.00	990000.0

Table 8: Statistics description of GODIGT dataframe

1. **annual_income_at_source** The yearly income of consumers at the source is represented by this variable. Its mean value of 1674595 represents the typical yearly income of its consumers. The correlation between income levels and potential credit card use and spending capacity makes this variable crucial.

2. **avg_spends_l3m** The average customer spending over the previous three months is represented by this variable. Its mean value of 49527 represents the typical amount spent by clients over that time period. Higher average spending indicates higher potential revenue for the bank, hence this variable is crucial.

3. **cc_limit** The credit limit imposed on clients' credit cards is represented by this variable. Higher credit limits provide clients with more freedom and spending power, which boosts the likelihood that they will use credit cards. The bank may make sure that consumers have access to the right credit card with a sufficient credit limit by evaluating their creditworthiness and allocating appropriate credit limits, enhancing their intent to use the card.

4. **T+1_month_activity** Used to target customers specifically and create tailored marketing campaigns to keep them.

5. **cc_active30** This may be used to monitor credit card usage and find out how frequently customers use their cards. This data may be used by GODIGT to remarket to their clients and find potential explanations for declining credit card usage.

- The bank may create strategies to keep consumers, promote credit card use, and increase profitability by concentrating on these factors.
- This might involve better client engagement programs, customised offers, targeted marketing efforts, and changes to credit limits.