SMDM
Business Report

By: Dhruv Dosad

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

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

	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

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

```
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)
```

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

```
Age                   0
Gender               53
Profession            0
Marital_status        0
Education             0
No_of_Dependents      0
Personal_loan         0
House_loan            0
Partner_working       0
Salary                0
Partner_salary      106
Total_salary          0
Price                 0
Make                  0
dtype: int64
```

Treating the Null values:-

a. **Gender:** Null values in Gender field can be imputed with '**Male'** having as the **mode** (maximum value in the dataset)

b. **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.

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

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

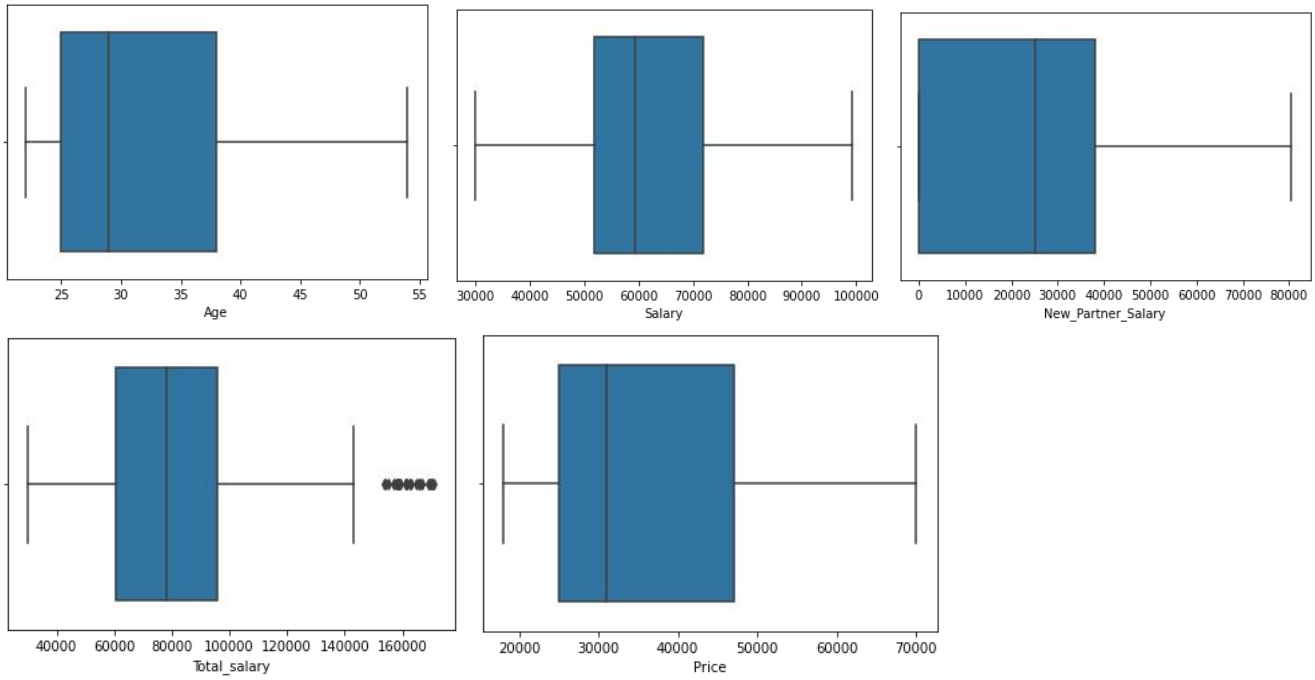
Inspecting the Summary Statistics of the Dataset

	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	20585.895003	79398.545225	35597.722960
std	8.425978	0.943483	14674.825044	18952.938643	24849.147996	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	25600.000000	78000.000000	31000.000000
75%	38.000000	3.000000	71800.000000	38000.000000	95900.000000	47000.000000
max	54.000000	4.000000	99300.000000	80500.000000	149000.000000	70000.000000

1) The customers 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.

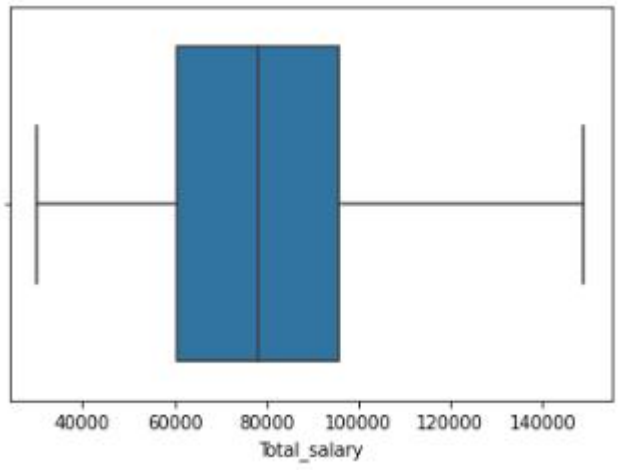
4) The **minimum price** of the **purchased automobile is 18K**, whereas **max is 70K.** Skewness indicates a small number of high priced purchases were made.

Checking Outliers in the numerical variable



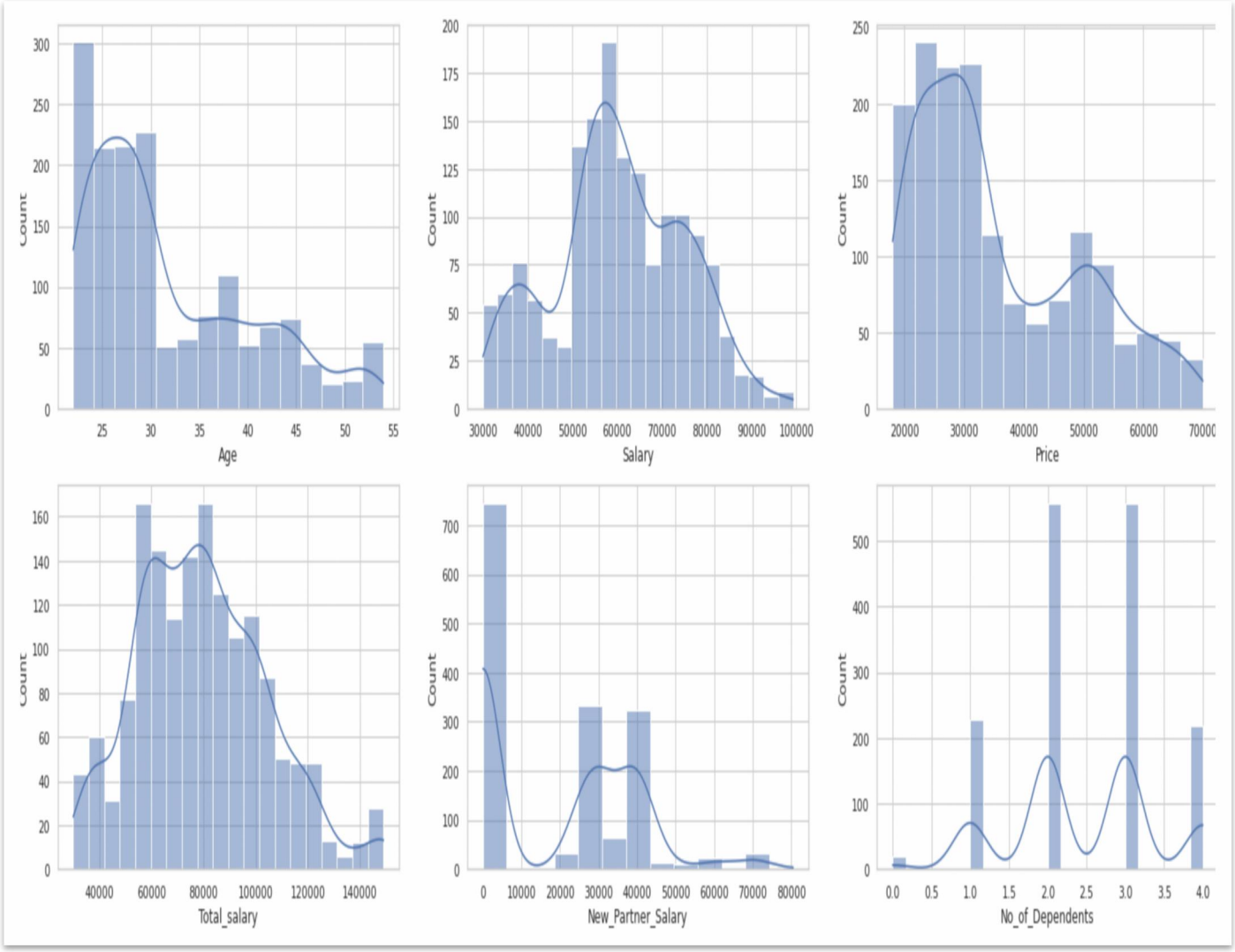
- There are **no negative** values present in the numerical fields.
- From the boxplots we can observe **outlier values are present in Total_salary** variables.
- Outliers are treated by using **Winsorization**, i.e. bringing the larger outliers (Data points above the Q3 + 1.5 *IQR value) to the upper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

Boxplot after Outlier treatment

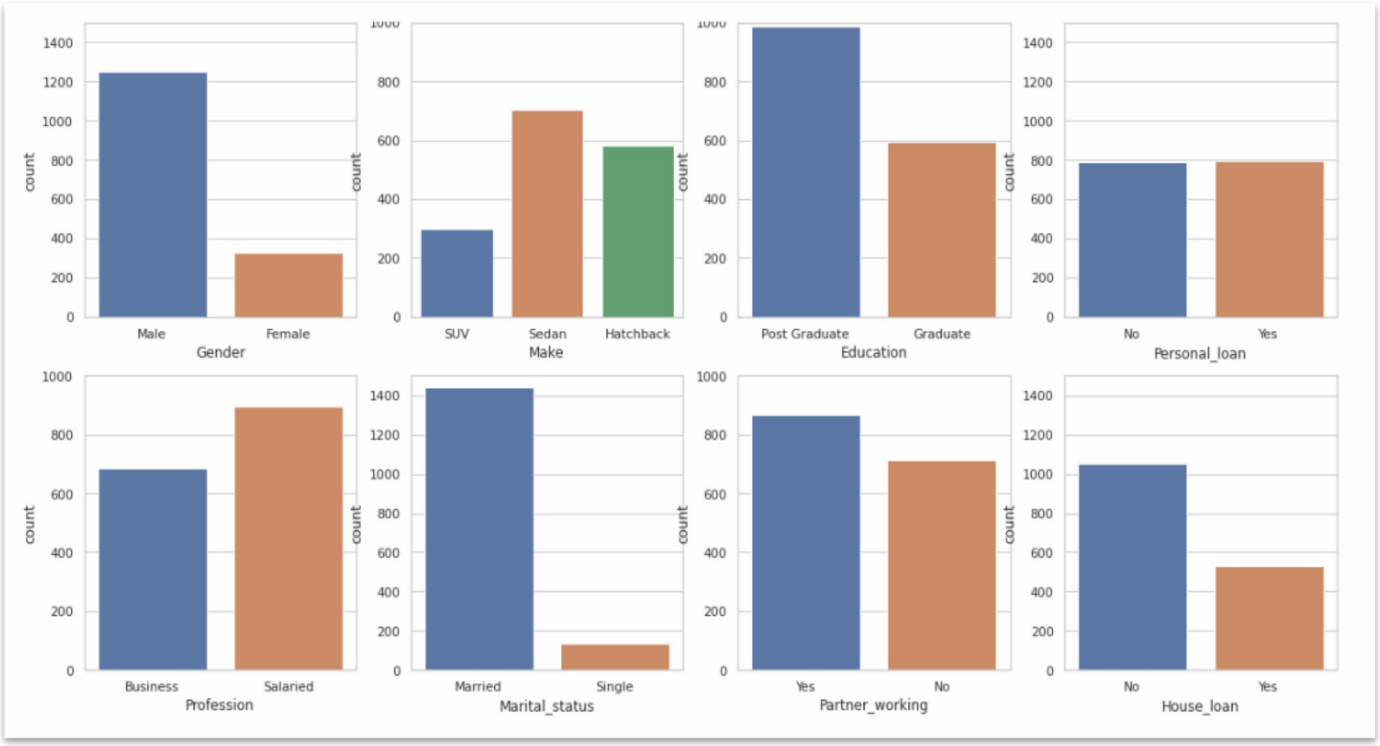


C. Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.

Univariate Analysis of Numerical fields



Univariate Analysis of Categorical fields



Inferences

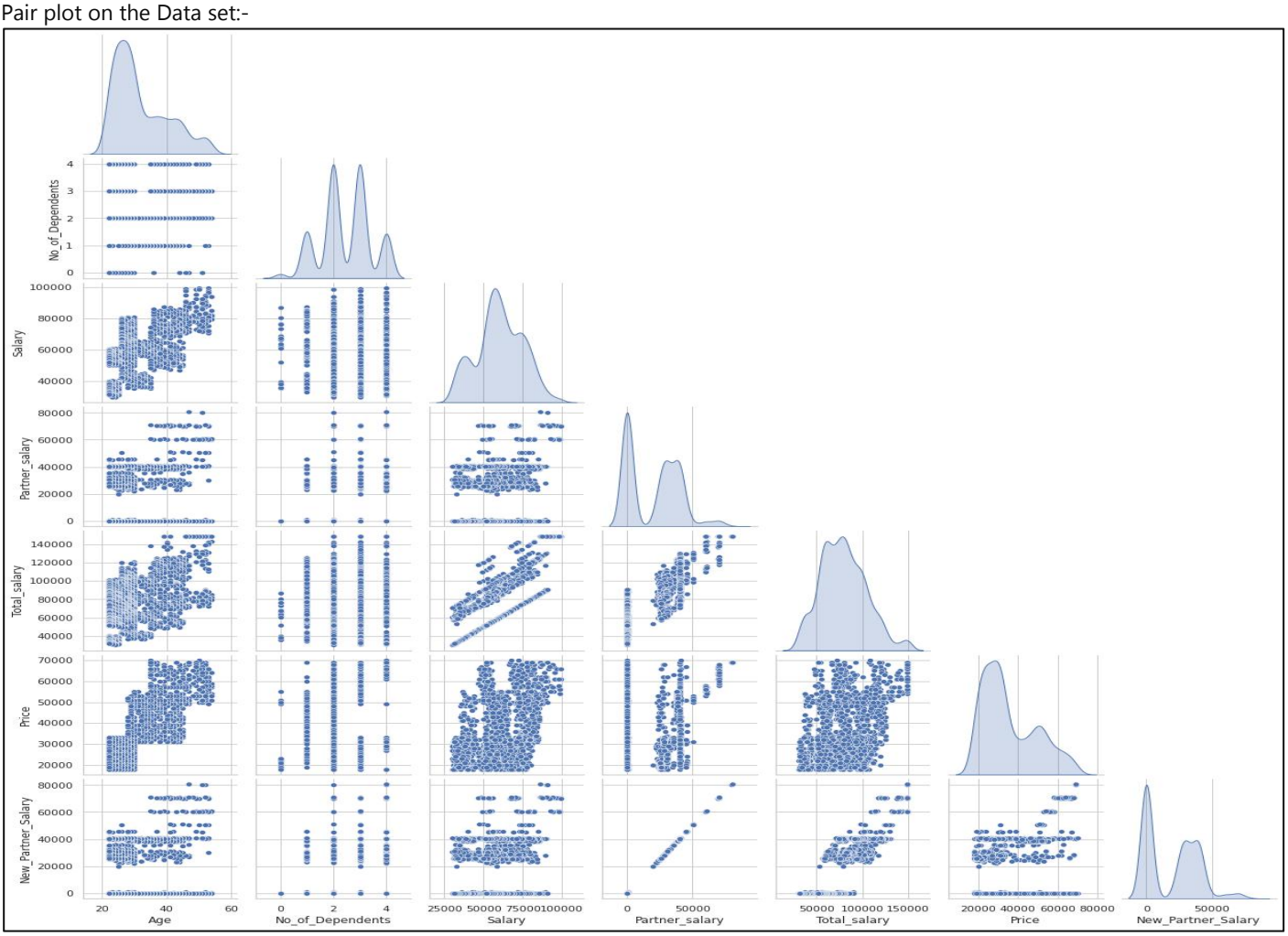
- **Majority** of the customers in the dataset are **Post Graduate**.

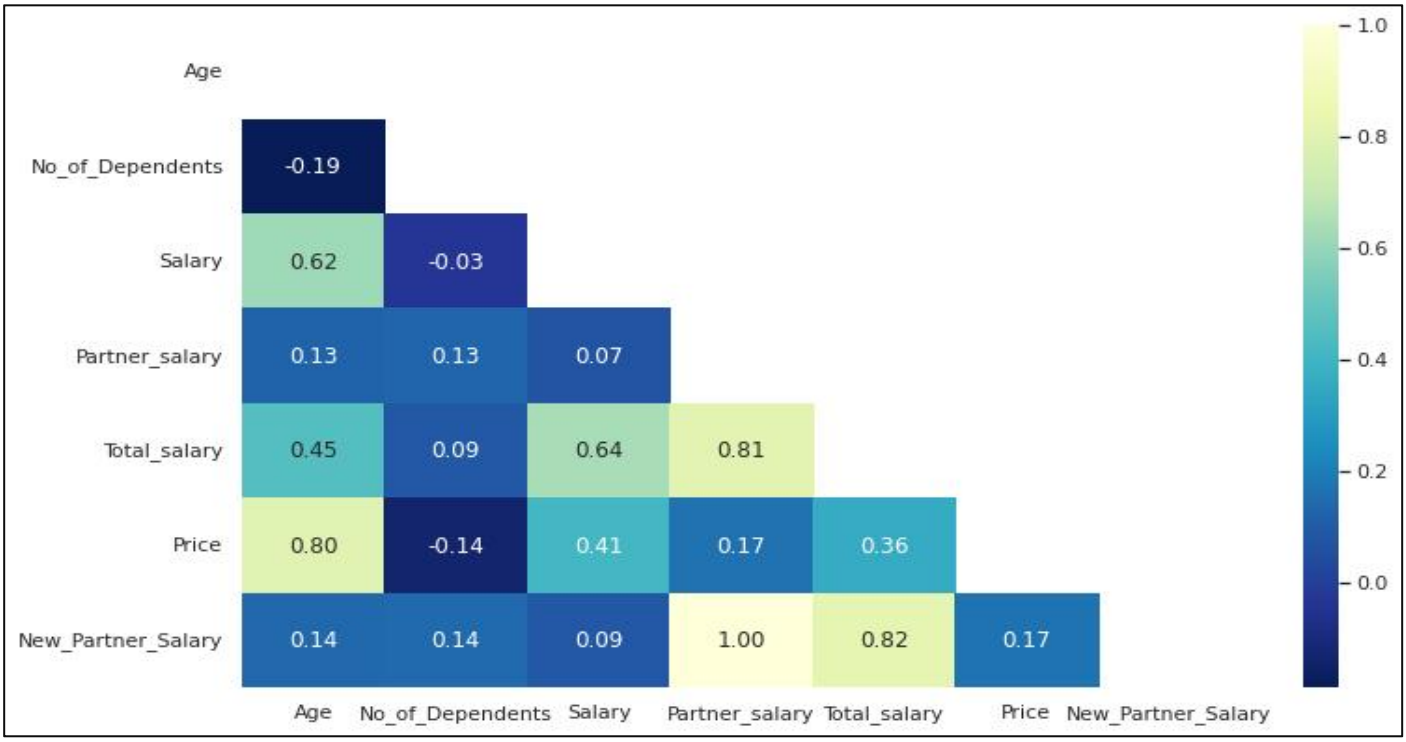
- **Sedan** is the **most preferred purchase**, followed by **Hatchback** and **SUV**
- **Salaried customer count** is **slightly higher** than that of **Business customers**.
- The number of **customers** having a **working partner** are **slightly higher** than customers with **non-working partner 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:-

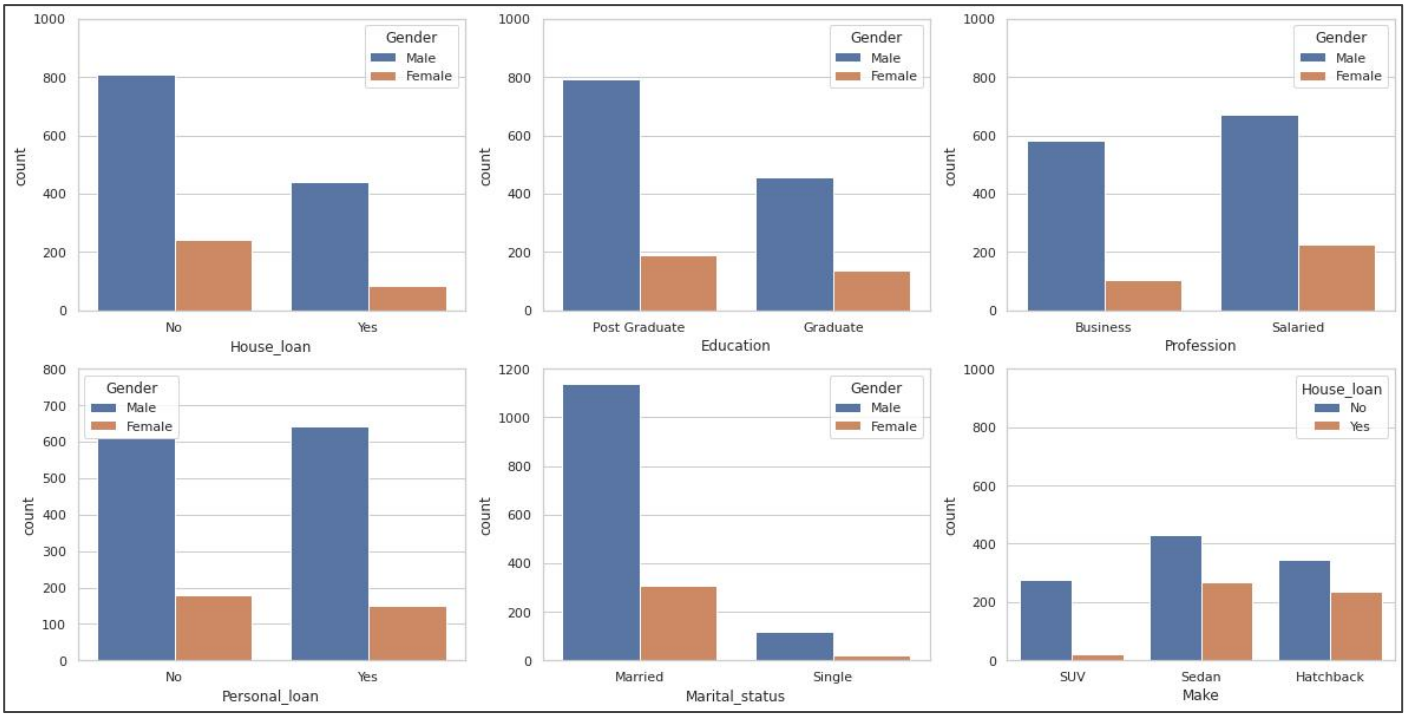




Inferences –

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 –

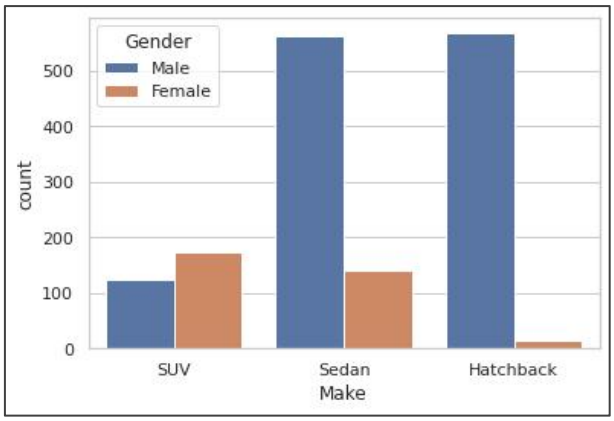


Inferences –

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.

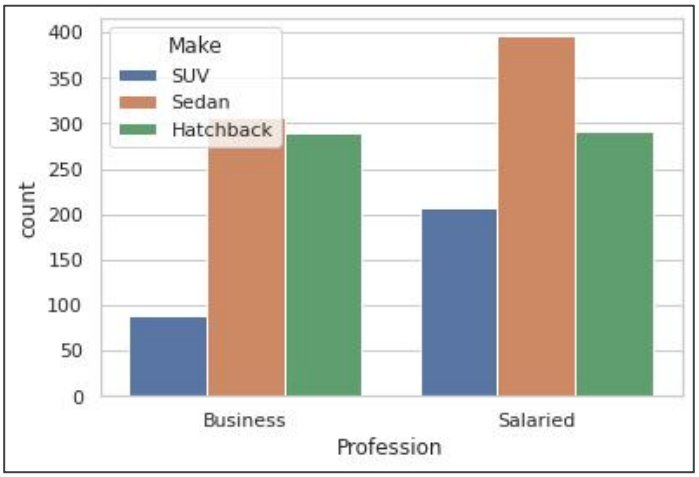
E1) Steve Roger says "Men prefer SUV by a large margin, compared to the women"
Analyzing the ratio of SUV purchases for both the Genders, we get:



Make	Hatchback	SUV	Sedan	All
Gender				
Female	15	173	141	329
Male	567	124	561	1252
All	582	297	702	1581

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

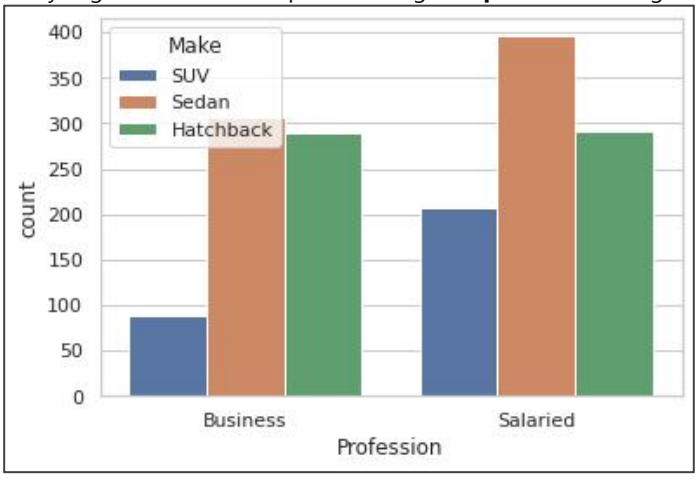
E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.
Analyzing the ratio of **Sedan** purchases against **profession**, we get:



Make	Hatchback	SUV	Sedan	All
Profession				
Business	290	89	306	685
Salaried	292	208	396	896
All	582	297	702	1581

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:



Make	Hatchback	SUV	Sedan	All
Profession				
Business	290	89	306	685
Salaried	292	208	396	896
All	582	297	702	1581

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**.

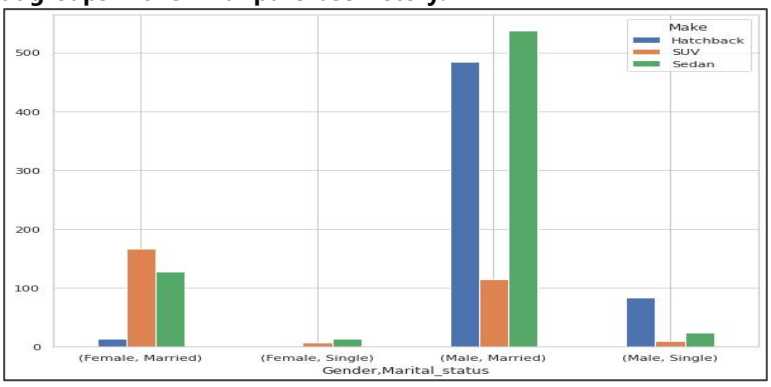
G. From the current data set comment if having a working partner leads to the purchase of a higher-priced car.

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.



Gender	Marital_status			
Female	Married	14	166	127
	Single	1	7	14
Male	Married	484	115	537
	Single	83	9	24

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

index	userid	card_no	card_bin_no	Issuer	card_type	card_source_date	high_networth	active_30	active_60	active_90	cc_active30	cc_active60	cc_active90	hotlist_flag	widget_products	engagement_products	annual_income_at_source	other_bank_cc_holding	bank_vintage	T+1_month_activity
0	1	4384 38XX XXXX XXXX	438408	Visa	edge	2019-09-29 00:00:00	B	0	1	1	0	0	0	N	1	3	1662111	Y	27	0
1	2	4377 48XX XXXX XXXX	437748	Visa	prosperity	2002-10-30 00:00:00	A	1	1	1	0	0	0	N	4	1	4833871	Y	62	0
2	3	4377 48XX XXXX XXXX	437748	Visa	rewards	2013-10-05 00:00:00	C	0	0	0	0	0	0	N	4	2	1345429	N	23	1
3	4	4268 06XX XXXX XXXX	426806	Visa	indianoil	1999-08-01 00:00:00	E	0	1	1	1	1	1	N	8	0	880660	N	49	0
4	5	4377 48XX XXXX XXXX	437748	Visa	edge	2006-08-13 00:00:00	B	1	1	1	0	1	1	N	4	3	1808682	N	21	1

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

```
Data columns (total 28 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   userid                    8448 non-null   int64
 1   card_no                   8448 non-null   object
 2   card_bin_no               8448 non-null   int64
 3   Issuer                    8448 non-null   object
 4   card_type                 8448 non-null   object
 5   card_source_date          8448 non-null   datetime64[ns]
 6   high_networth             8448 non-null   object
 7   active_30                 8448 non-null   int64
 8   active_60                 8448 non-null   int64
 9   active_90                 8448 non-null   int64
 10  cc_active30               8448 non-null   int64
 11  cc_active60               8448 non-null   int64
 12  cc_active90               8448 non-null   int64
 13  hotlist_flag              8448 non-null   object
 14  widget_products           8448 non-null   int64
 15  engagement_products       8448 non-null   int64
 16  annual_income_at_source   8448 non-null   int64
 17  other_bank_cc_holding     8448 non-null   object
 18  bank_vintage              8448 non-null   int64
 19  T+1_month_activity        8448 non-null   int64
 20  T+2_month_activity        8448 non-null   int64
 21  T+3_month_activity        8448 non-null   int64
 22  T+6_month_activity        8448 non-null   int64
 23  T+12_month_activity       8448 non-null   int64
 24  Transactor_revolver       8410 non-null   object
 25  avg_spends_l3m            8448 non-null   int64
 26  Occupation_at_source      8448 non-null   object
 27  cc_limit                  8448 non-null   int64
dtypes: datetime64[ns](1), int64(19), object(8)
```

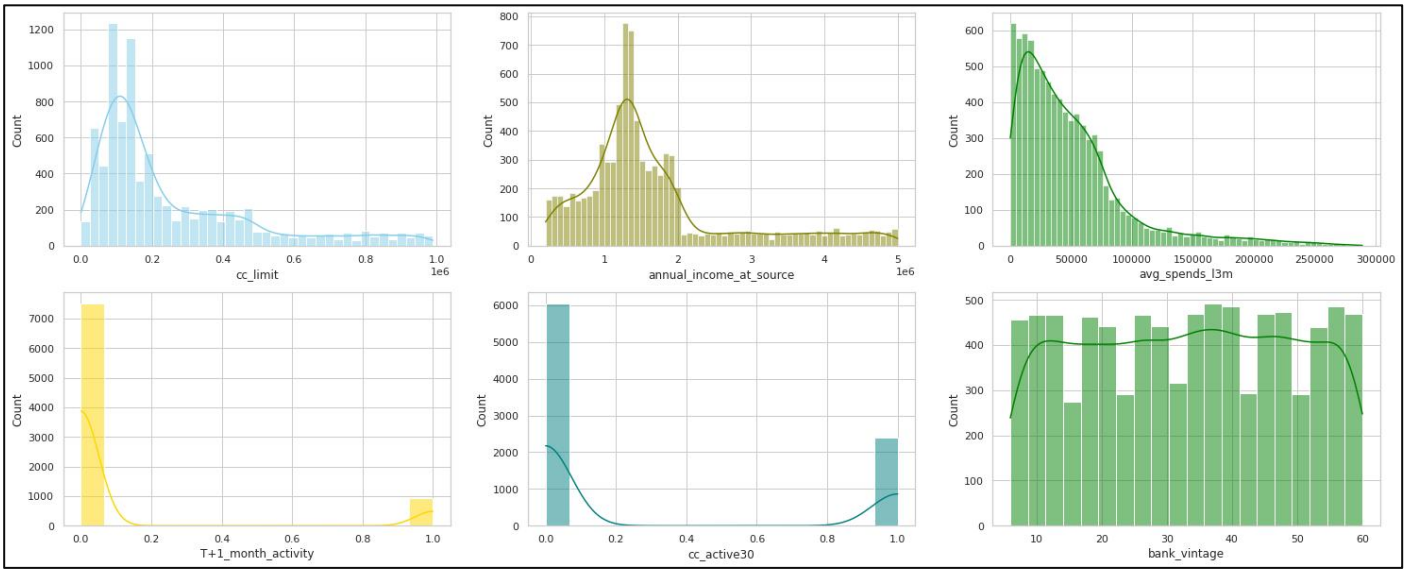
Five Points Summary:

	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

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 behaviour. 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----------