SMDM PROJECT REPORT

DSBA

Contents

Problem 1

Analysts are required to explore data and reflect on the insights. Clear writing skill is an integral part of a good report. Note that the explanations must be such that readers with minimum knowledge of analytics is able to grasp the insight.

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**.**

You as an analyst have been tasked with performing a thorough analysis of the data and coming up with insights to improve the marketing 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)

Firstly load the set and check for info() function found that 1581 rows are there and 14 columns .

```
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
```

Secondly for seeing that the set is being loaded we need to always check head() and tail() function .

head() function

1 to 5 of 5 entries Filter

index	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

tail function()

1 to 5 of 5 entries Filter

index	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary	Price	Mak
1576	22	Male	Salaried	Single	Graduate	2	No	Yes	No	33300	0.0	33300	27000	Hatchb
1577	22	Male	Business	Married	Graduate	4	No	No	No	32000	NaN	32000	31000	Hatchb
1578	22	Male	Business	Single	Graduate	2	No	Yes	No	32900	0.0	32900	30000	Hatchb
1579	22	Male	Business	Married	Graduate	3	Yes	Yes	No	32200	NaN	32200	24000	Hatchb
1580	22	Male	Salaried	Married	Graduate	4	No	No	No	31600	0.0	31600	31000	Hatchb

Show 25 per page

When we used info function() we found that there are no duplicates and there are 8 categorical variables and 6 numerical variables

Categorical variables are ::

- Gender
- Profession
- Marital_status
- Education
- Personal_loan
- House_loan
- Partner_working
- Make

Numeric Variables are::

- Age
- No_of_Dependents
- Salary
- Partner_salary
- Total_salary
- Price

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.

Firstly, we have found the null values i.e. is there any null values present in dataset.

```
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
```

We found that in Gender and Partner_salary there are null values i.e. 53 anf 106 respectively.

For handling nulls we need to see that if null values are more than 60% then it should be removed basically the column should be dropped and otherwise we need to impute that null values with different techniques

In both Gender and Partner_salary is rows contains small number of missing values so we

cant drop it we need to impute the values

For gender we impute it with the majority i.e Males basically mode .

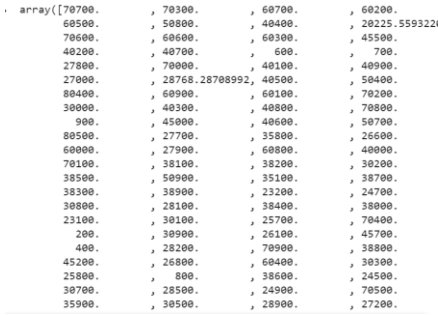
For Partner_salary we need to impute it with mean and median for imputing .

We have related parameters to salary

- Partner_salary
- Total_salary
- Salary

.

For Partner_Salary we also checked the unique values so that we can also find any anonymous values present or not in that .



```
array([70700.        , 70300.        , 60700.        , 60200.        ,
       60500.        , 50800.        , 40400.        , 20225.55932203,
       70600.        , 60600.        , 60300.        , 45500.        ,
       40200.        , 40700.        ,   600.        ,   700.        ,
       27800.        , 70000.        , 40100.        , 40900.        ,
       27000.        , 28768.28708992, 40500.        , 50400.        ,
       80400.        , 60900.        , 60100.        , 70200.        ,
       30000.        , 40300.        , 40800.        , 70800.        ,
         900.        , 45000.        , 40600.        , 50700.        ,
       80500.        , 27700.        , 35800.        , 26600.        ,
       60000.        , 27900.        , 60800.        , 40000.        ,
       70100.        , 38100.        , 38200.        , 30200.        ,
       38500.        , 50900.        , 35100.        , 38700.        ,
       38300.        , 38900.        , 23200.        , 24700.        ,
       30800.        , 28100.        , 38400.        , 38000.        ,
       23100.        , 30100.        , 25700.        , 70400.        ,
         200.        , 30900.        , 26100.        , 45700.        ,
         400.        , 28200.        , 70900.        , 38800.        ,
       45200.        , 26800.        , 60400.        , 30300.        ,
       25800.        ,   800.        , 38600.        , 24500.        ,
       30700.        , 28500.        , 24900.        , 70500.        ,
       35900.        , 30500.        , 28900.        , 27200.        ,
       23100.        , 30100.        , 25700.        , 70400.        ,
         200.        , 30900.        , 26100.        , 45700.        ,
         400.        , 28200.        , 70900.        , 38800.        ,
       45200.        , 26800.        , 60400.        , 30300.        ,
       25800.        ,   800.        , 38600.        , 24500.        ,
       30700.        , 28500.        , 24900.        , 70500.        ,
       35900.        , 30500.        , 28900.        , 27200.        ,
       45900.        , 25300.        , 35600.        , 25000.        ,
       25200.        , 26300.        , 35700.        , 23800.        ,
       32700.        , 30600.        , 45600.        , 28000.        ,
       32600.        , 28600.        , 25100.        , 26700.        ,
       24200.        , 25900.        , 22100.        , 25400.        ,
       28400.        , 28300.        , 25500.        , 24000.        ,
       32400.        , 28800.        , 32300.        , 32900.        ,
       27600.        , 35500.        , 23500.        , 45400.        ,
       28700.        , 32800.        , 23700.        , 32500.        ,
       35300.        , 45800.        , 27300.        , 22900.        ,
       29800.        , 30400.        , 35400.        , 24300.        ,
       29200.        , 24600.        ,   100.        , 50300.        ,
       32000.        , 32200.        , 24400.        , 32100.        ,
       22600.        , 26900.        ,   300.        , 20000.        ,
       26200.        , 22300.        , 25600.        ,   500.        ,
       35200.        , 35000.        ])
```

After imputing the null values there are no null values present. We have checked the same

```
Age 0
Gender 0
Profession 0
Marital_status 0
Education 0
No_of_Dependents 0
Personal_loan 0
House_loan 0
Partner_working 0
Salary 0
Partner_salary 0
Total_salary 0
Price 0
Make 0
```

Checking we need to find that is there any anonymous values present so we found that in Gender there are anonymous values present.

We found that Female is spelled as Femal and Femle at 2 places in the dataset.

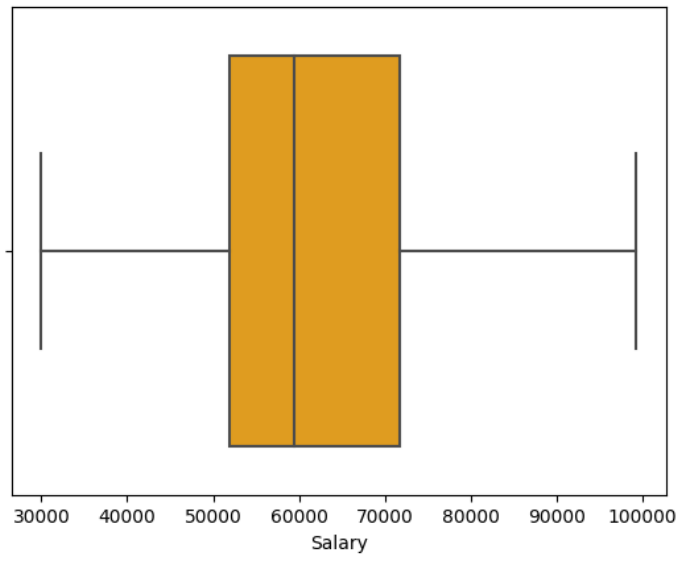
```
Male      1199
Female     327
Femal        1
Femle        1
Name: Gender, dtype: int64
```

For treating the same we need to change both Femal and Femle to Female so that the data doesnot have any anonymous values present.
We replaced both the anonymous values with Female as shown below
So now all null values along with anonymous values of Gender has been treated well.

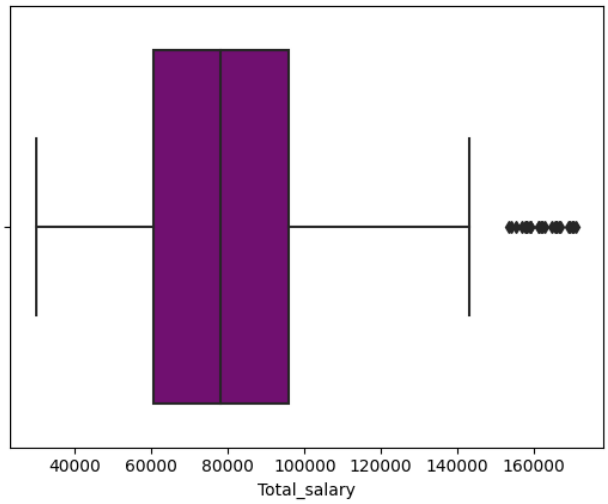
```
Male      1252
Female     329
Name: Gender, dtype: int64
```

We need to check the outliers for all the fields given so that we can treat the same.
For checking the same we used boxplot of the numeric fields.

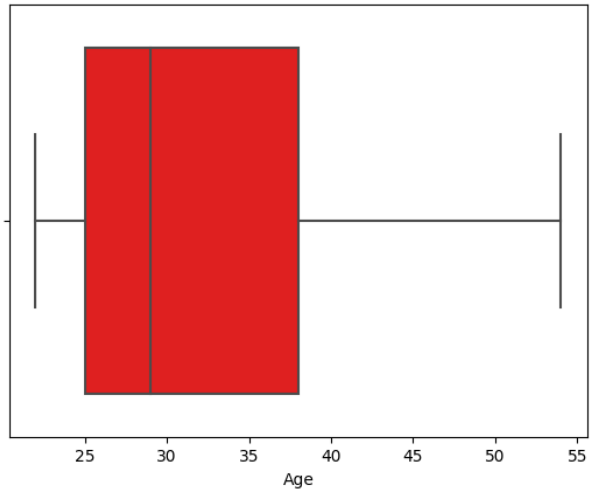
- For Salary: No outliers are found.



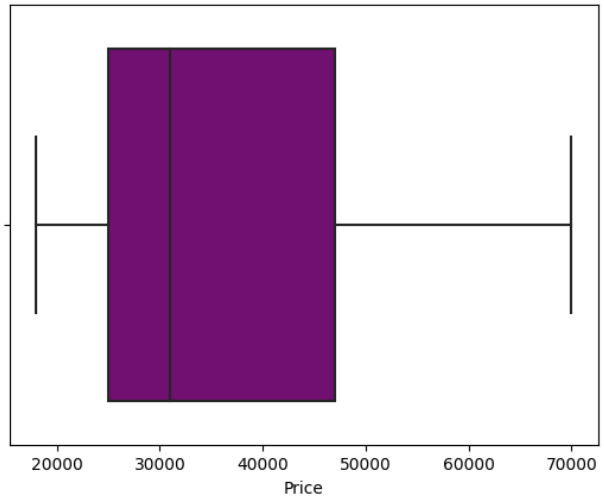
- For Total_salary:: We found out ouliers are present.



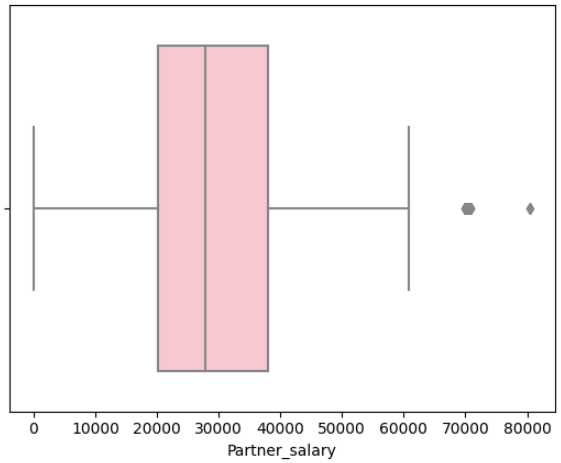
- For Age:: No outlier is present.



- For Price : No outlier is present



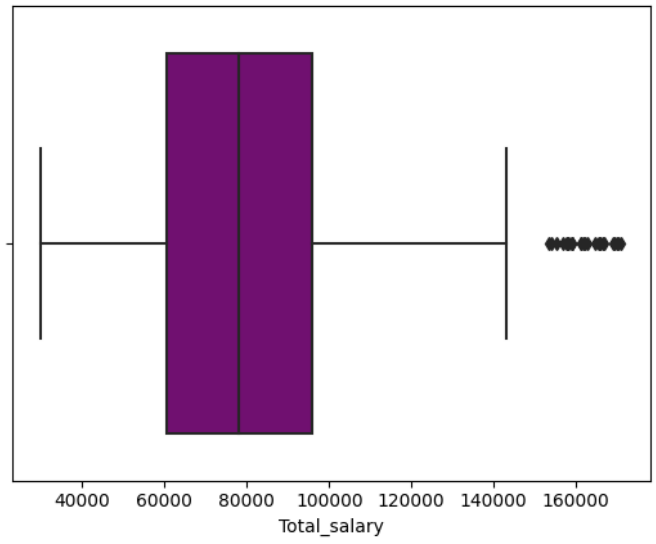
- For Partner_salary : outlier is present



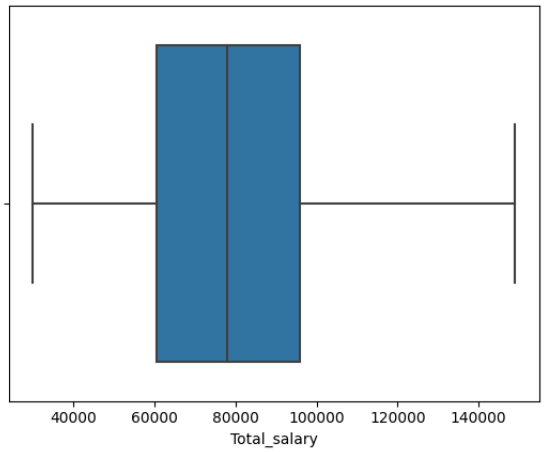
- We need to treat these outliers for Total_salary by using the outlier treatment i.e. IQR method . After treating with this method we found that the outliers has been removed from the parameter Total_salary.

- We find the lr,ur i.e. lower_range=Q1-(1.5*IQR)
- upper_range=Q3+(1.5*IQR) and then do the analysis accordingly with IQR method

Earlier when outlier is present:



- After treating the outliers in Total_salary with IQR treatment of outliers we didn't found any outliers now



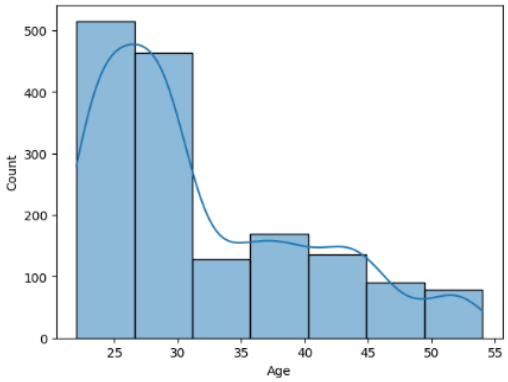
C. Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.

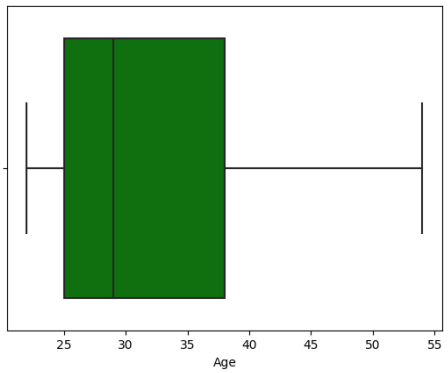
We need to use the describe function to see how many numerical variables are present and what's the numerical calculations present for the numeric variables. We have transposed the table to have a better understanding.

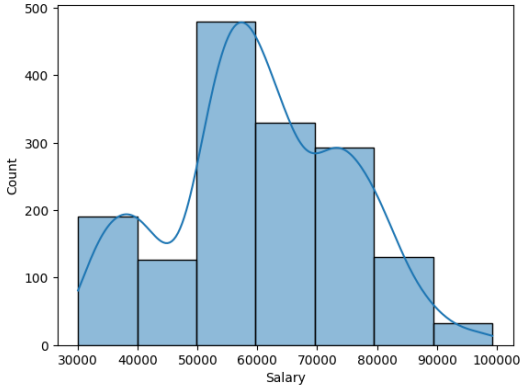
	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.000000	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.000000	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.000000	59500.0	71800.0	99300.0
Partner_salary	1581.0	28768.287090	11312.756609	100.0	20225.559322	27900.0	38000.0	80500.0
Total_salary	1581.0	79398.545225	24849.147996	30000.0	60500.000000	78000.0	95900.0	149000.0
Price	1581.0	35597.722960	13633.636545	18000.0	25000.000000	31000.0	47000.0	70000.0

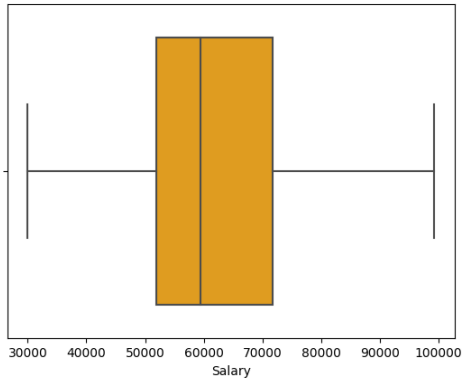
Doing Univariate analysis on Numeric Variables:

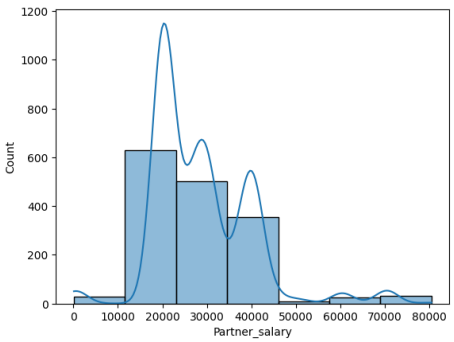
In univariate analysis we always or you can saw mostly uses Boxplot and histplot .
These analysis has been done after all treatment of the data i.e. removing null values ,imputation etc.

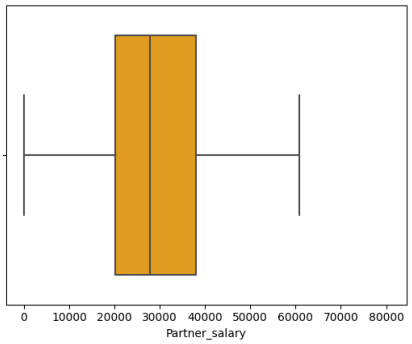

Count
Age

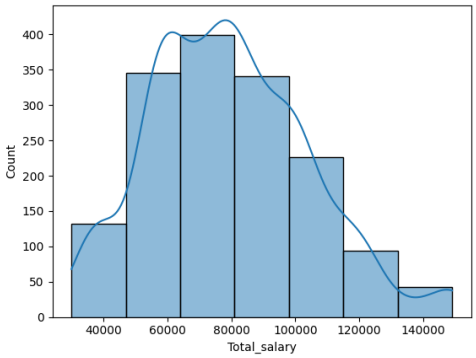

Age

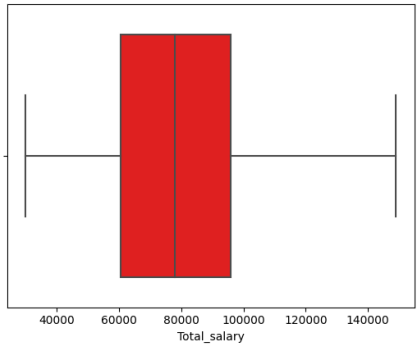

Count
Salary

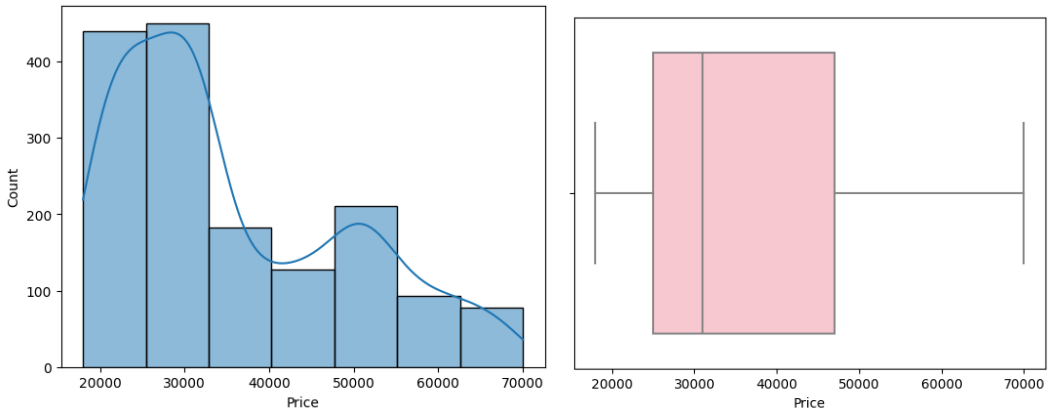

Salary


Count
Partner_salary


Partner_salary

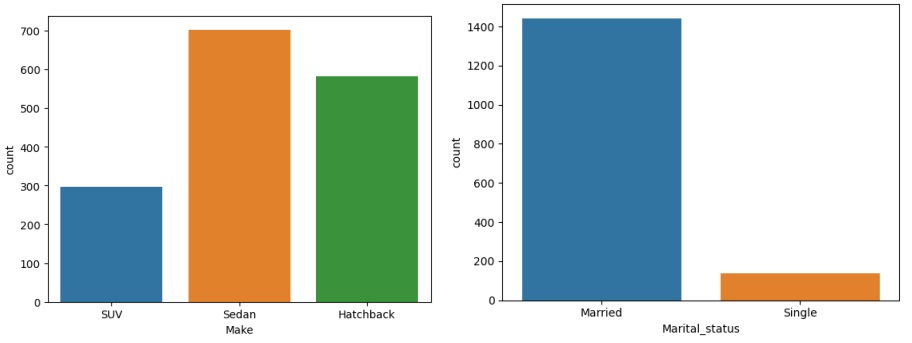

Count
Total_salary

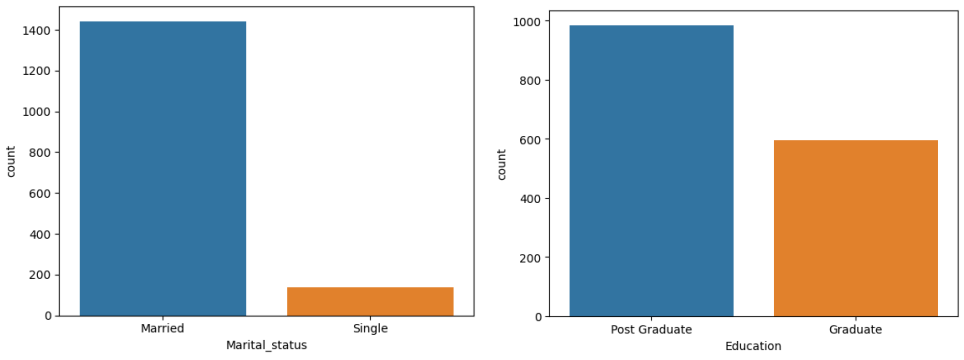

Total_salary

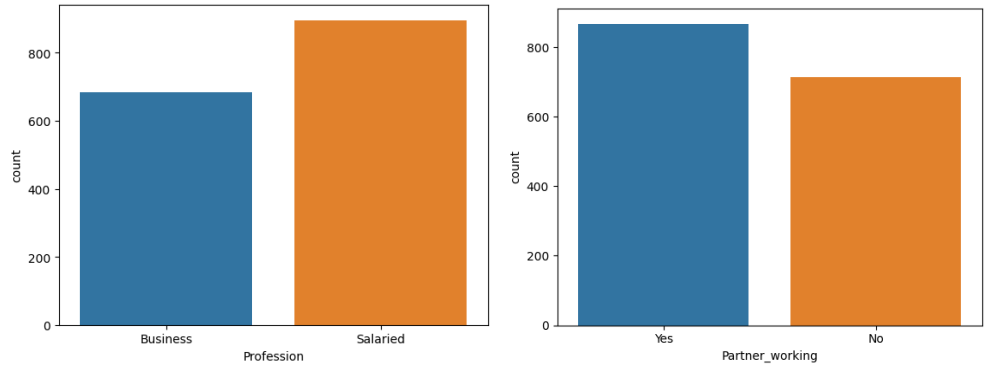


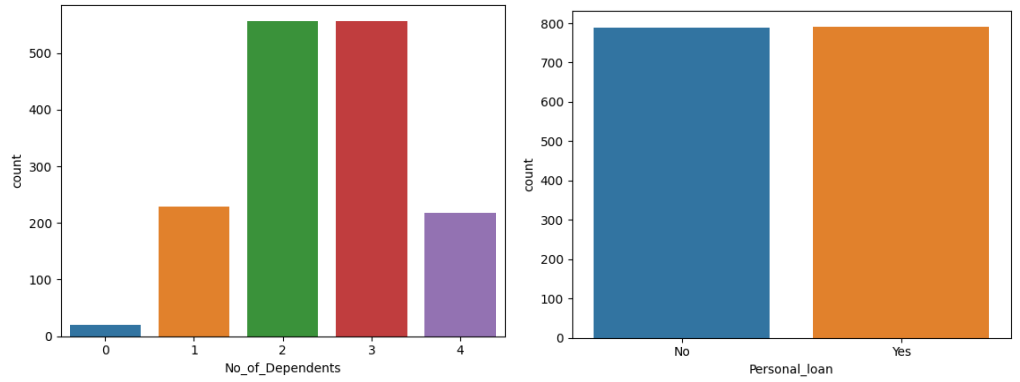
1) Salary has a multimodal distribution and data points in the range 50K to 70K.
2) Price seems to have a Bi-modal distribution a positive skew of 0.74.
3) Skewness of Total_salary has reduced significant post outlier treatment.
4) All the variables have some skewness present, thus none of them follow a Normal distribution. Total_salary can be considered Near-Normal distribution with fair bit of approximation.

Univariate analysis of Categorical variables







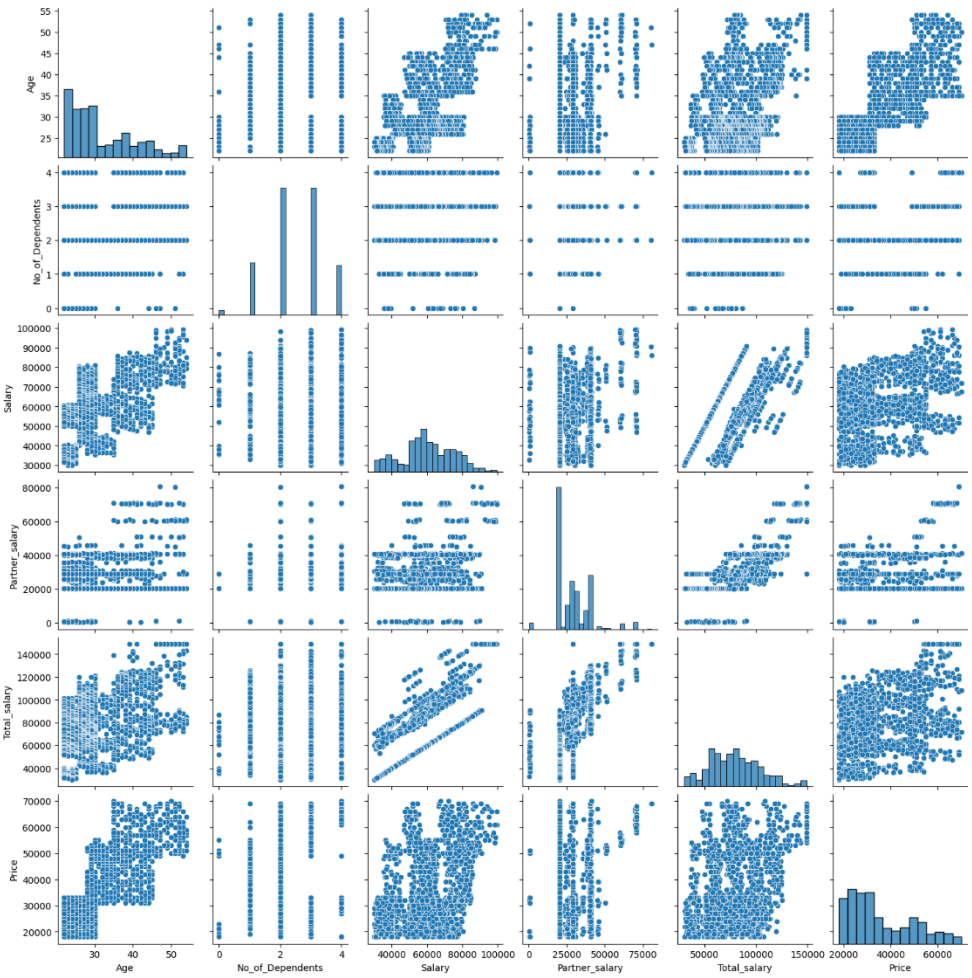


- Count of salaried men/women are more than the Business ones.
- Very less has 0 dependents mostly has 2 and 3 dependents.
- Partner_working is more as compared to partner not working.
- Most of the customer are post graduates.
- Married customers are more as compared to single customers.

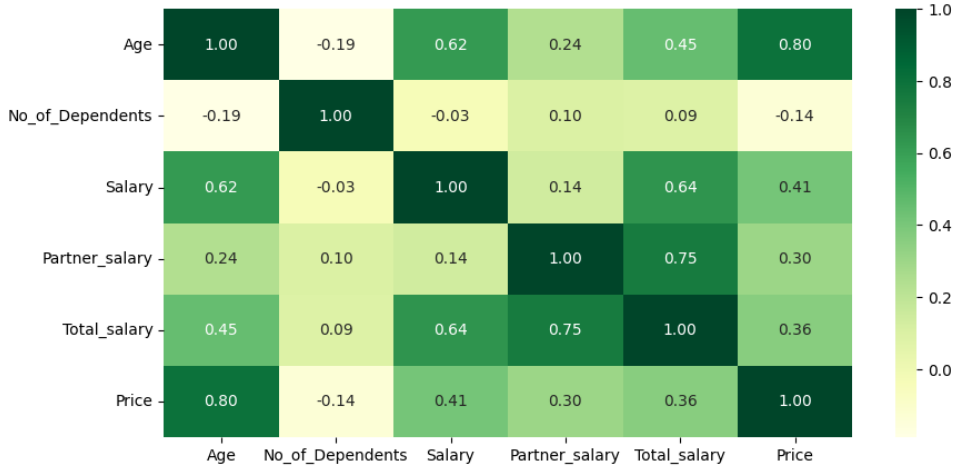
D. Understanding the relationships among the variables in the dataset is crucial for every analytical project. Perform analysis on the data fields to gain deeper insights. Comment on your understanding of the data. D. Understanding the relationships among the variables in the dataset is crucial for every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

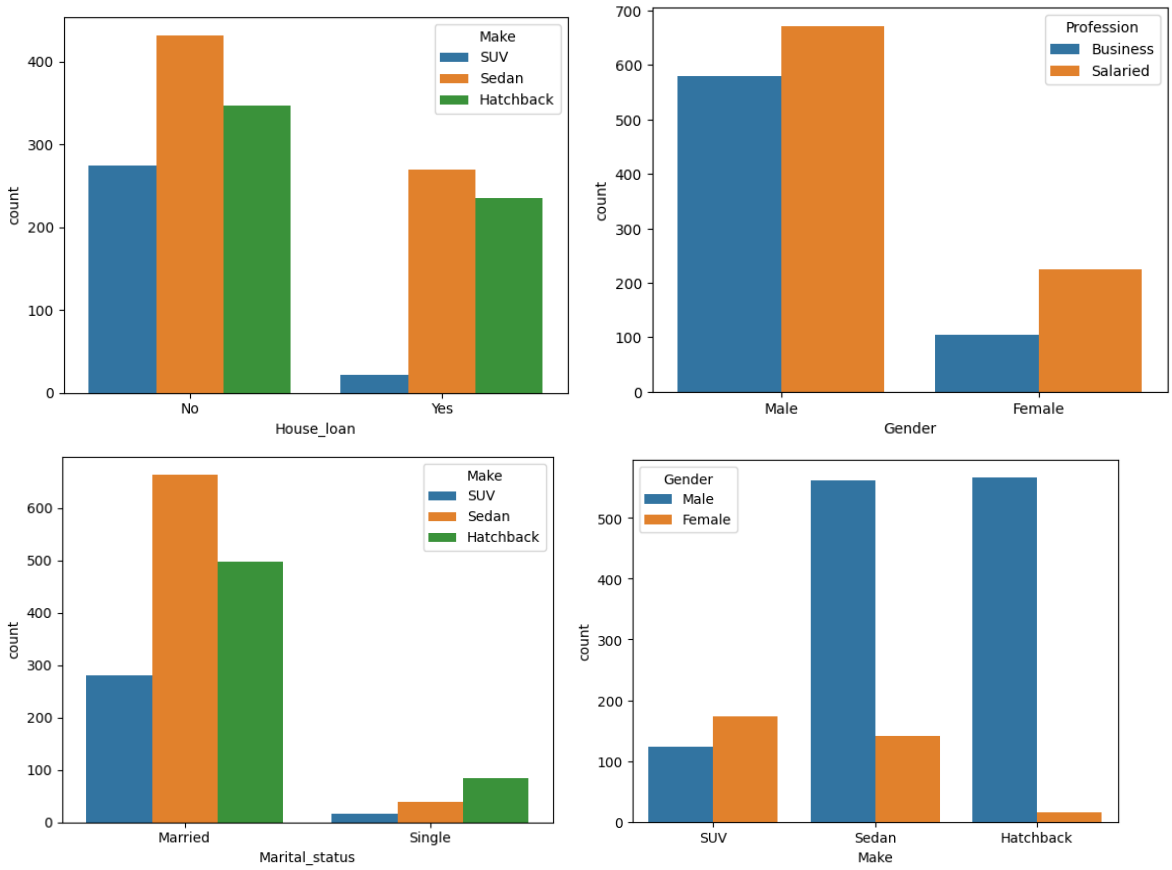
- <u>Bivariate analysis of Numerical Variables :</u>

We have done using the pairplot.



Hardly we found any relationships between all the functions in pairplot.



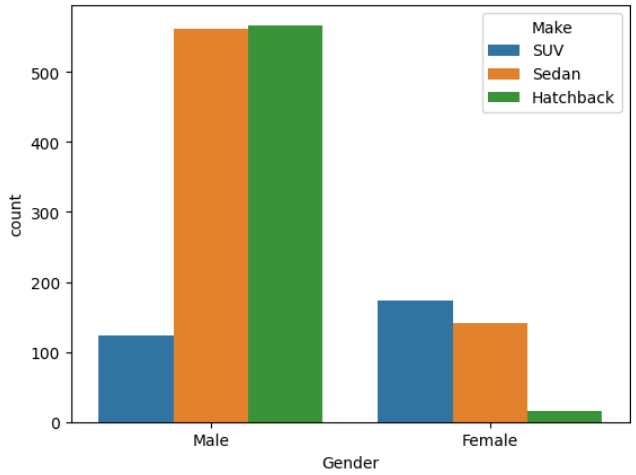


- Married customers prefer Sedan and single customer prefer Hatchback more however mostly customers are married in the dataset given
- Male prefer Hatchback more and female prefer SUV more
- Mostly male are salaried and mostly female are also salaried.

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.

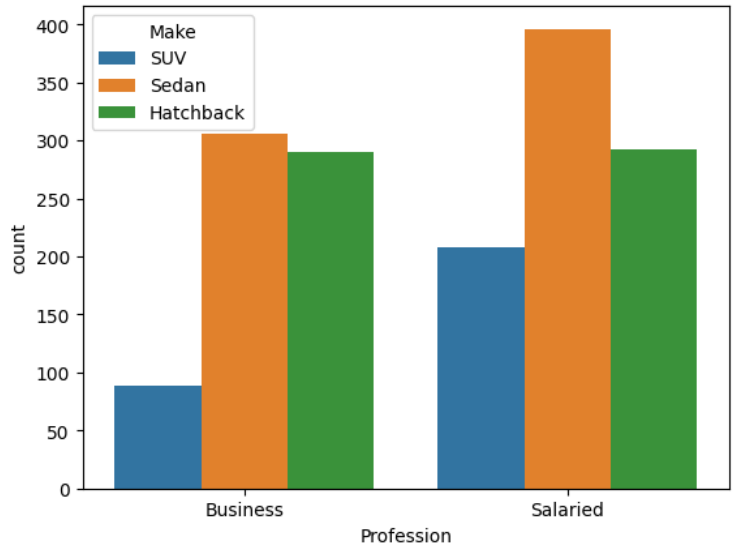
E1) Steve Roger says “Men prefer SUV by a large margin, compared to the women”

No, the statement is not correct as told by Steve Roger. The below complete analysis with graph has been done.

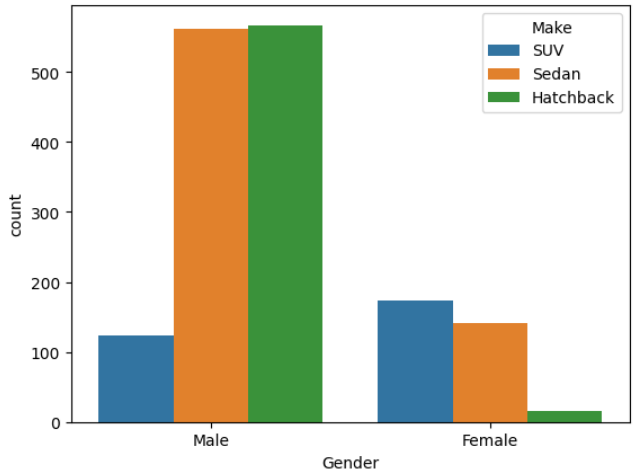


E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.

By analyzing the graph we can say that the salaried person is more likely to buy Sedan . Hence the statement 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.



Make	Hatchback		SUV		Sedan		All
Profession	Business	Salaried	Business	Salaried	Business	Salaried	
Gender							
Female	0	15	55	118	50	91	329
Male	290	277	34	90	256	305	1252
All	290	292	89	208	306	396	1581

Salaried male prefer Sedan more than the SUV so the given statement made by Sheldon 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

Mean of female and male is 47705 and 32416 resp.

```
Gender
Female    47705.167173
Male      32416.134185
Name: Price, dtype: float64
```

Median of female and male is 49000 and 29000 resp.

```
Gender
Female    49000.0
Male      29000.0
Name: Price, dtype: float64
```

Both mean and median price for female is more i.e spend of female is more . Female are more likely to buy SUV .

F2) Personal_loan

Mean

```
Personal_loan
No      36742.712294
Yes     34457.070707
Name: Price, dtype: float64
```

Median

```
Personal_loan
No      32000.0
Yes     31000.0
Name: Price, dtype: float64
```

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

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

Mean

```
Partner_working
No      36000.000000
Yes     35267.281106
Name: Price, dtype: float64
```

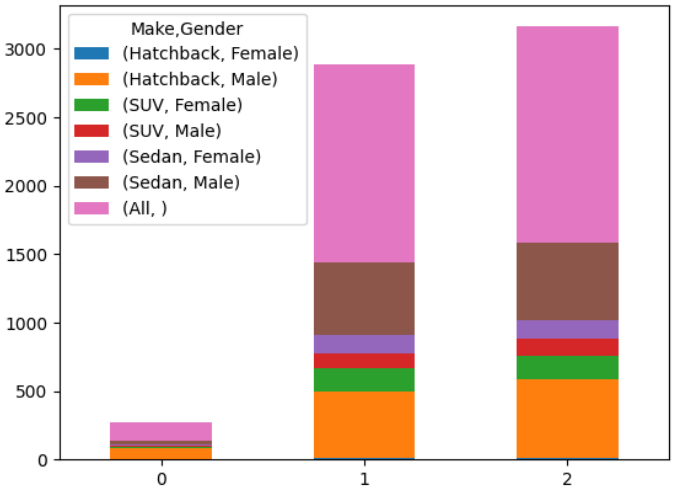
Median

```
Partner_working
No      31000.0
Yes     31000.0
Name: Price, dtype: float64
```

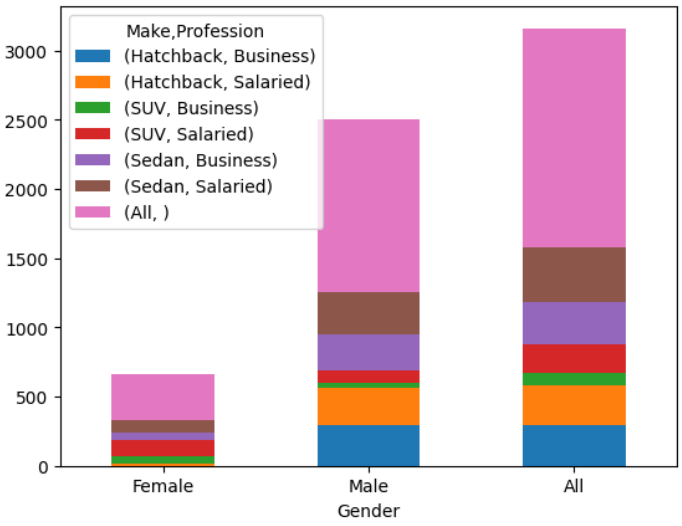
Median is same and for partner not working is slighly more however it will not affect the purchasing of cars. So partner working or not will not affect the sale of the car

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.

Make	Hatchback		SUV		Sedan		All
Gender	Female	Male	Female	Male	Female	Male	
0	1	83	7	9	14	24	138
1	14	484	166	115	127	537	1443
2	15	567	173	124	141	561	1581



From above analysis we can say that married female prefer SUV more , Single female prefer sedan more , Married male prefer Sedan more and single male prefer Hatchback more.



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.)

Problem 2 Question: (Analyze the dataset and list down the top 5 important variables, along with the business justifications.

We have loaded the dataset given and started analysis of the data.
Firstly we have done info() to check the parameters and found 28 columns and 8447 rows

```
RangeIndex: 8448 entries, 0 to 8447
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
```

Done the head() and tail () function to the dataset to analyse the data.

- head()

[60]

	userid	card_no	card_bin_no	Issuer	card_type	card_source_date	high_networth	active_30	active_60	active_90	...	bank_vintage	T+1_month_activ
0	1	4384 39XX XXXX XXXX	438439	Visa	edge	2019-09-29	B	0	1	1	...	27	
1	2	4377 48XX XXXX XXXX	437748	Visa	prosperity	2002-10-30	A	1	1	1	...	52	
2	3	4377 48XX XXXX XXXX	437748	Visa	rewards	2013-10-05	C	0	0	0	...	23	
3	4	4258 06XX XXXX XXXX	425806	Visa	indianoil	1999-06-01	E	0	1	1	...	49	
4	5	4377 48XX XXXX XXXX	437748	Visa	edge	2006-06-13	B	1	1	1	...	21	

- tail()

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	wi
8443	8444	4262 41XX XXXX XXXX	426241	Visa	chartered	2010-01-10 00:00:00	A	1	1	1	0	0	0	N	
8444	8445	37691 6XXXX XXXXX	376916	Amex	centurion	2006-07-19 00:00:00	A	0	1	1	0	1	1	N	
8445	8446	4375 51XX XXXX XXXX	437551	Visa	rewards	2006-02-15 00:00:00	D	0	1	1	0	0	0	N	
8446	8447	4477 47XX XXXX XXXX	447747	Visa	indianoil	2003-11-06 00:00:00	C	0	0	0	0	0	0	N	
8447	8448	4262 41XX XXXX XXXX	426241	Visa	rewards	2013-01-13 00:00:00	D	0	1	1	0	0	0	N	

- Describe()

	count	mean	std	min	25%	50%	75%	max
userid	8448.0	4224.500000	2438.871870	1.0	2112.75	4224.5	6336.25	8448.0
card_bin_no	8448.0	436747.044508	30489.752417	376916.0	426241.00	437551.0	438439.00	524178.0
active_30	8448.0	0.292377	0.454881	0.0	0.00	0.0	1.00	1.0
active_60	8448.0	0.494792	0.500002	0.0	0.00	0.0	1.00	1.0
active_90	8448.0	0.642045	0.479427	0.0	0.00	1.0	1.00	1.0
...	...	...	...	...	...	...	...	...
T+3_month_activity	8448.0	0.080374	0.271888	0.0	0.00	0.0	0.00	1.0
T+6_month_activity	8448.0	0.008878	0.093809	0.0	0.00	0.0	0.00	1.0
T+12_month_activity	8448.0	0.009470	0.096856	0.0	0.00	0.0	0.00	1.0
avg_spends_l3m	8448.0	49527.365530	46244.954836	0.0	17110.00	37943.0	66095.75	289292.0
cc_limit	8448.0	251706.912879	229114.856385	0.0	90000.00	150000.0	350000.00	990000.0

19 rows × 8 columns

- Checking the null values :: No null values found

```
for i in cf_excel.columns:
  print(i,cf_excel[i].isnull().sum())
```

```
userid 0
card_no 0
card_bin_no 0
Issuer 0
card_type 0
card_source_date 0
high_networth 0
active_30 0
active_60 0
active_90 0
cc_active30 0
cc_active60 0
cc_active90 0
hotlist_flag 0
widget_products 0
engagement_products 0
annual_income_at_source 0
other_bank_cc_holding 0
bank_vintage 0
T+1_month_activity 0
T+2_month_activity 0
T+3_month_activity 0
```

5 Important variables :

- **cc_limit::** A cash credit limit, often known as a CC limit, is a type of current account that includes a chequebook. Small medium companies (SMEs) are granted a cash credit limit, or CC limit, by the bank to meet their working capital needs. The bank accepts stock and debtors as principal security from CC limit holders.

- **annual_income_at_source**::Income can used by banks to take different decisions such as campaign , loan limit etc. Annual income recoreded in credit card application

- **avg_spends_l3m::** Average of credit card spend in 3 months can be useful if the bank do provide to increase the limit of credit card that can help the customer and bank also by making customer more frequent use of credit card and can provide offers also on that.

- **T+1_month_activity::** Provide offers to increase the use of credit card so that customer use it more often.

- **Cc_active30-** Can be useful for the bank to understand how frequent customer uses the credit card if the account is dormant.