Project Report

STATISTICAL METHODS AND DECISION MAKING

By – Suraj Mishra

Contents

A. Data overview

B. Univariate analysis

C. Bivariate analysis

D. Key questions.

1. Do men tend to prefer SUVs more compared to women?
2. What is the likelihood of a salaried person buying a Sedan?
3. What evidence or data supports Sheldon Cooper's claim that a salaried male is an easier target for a SUV sale over a Sedan sale?
4. How does the amount spent on purchasing automobiles vary by gender?
5. How much money was spent on purchasing automobiles by individuals who took a personal loan?
6. How does having a working partner influence the purchase of higher-priced cars?

Problem 1 - Actionable Insights & Recommendations - Actionable Insights - Business Recommendations

Problem 2 - Framing Analytics Problem

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

Problem 1

A)

Data Frame Provided is = Austo Motor Company

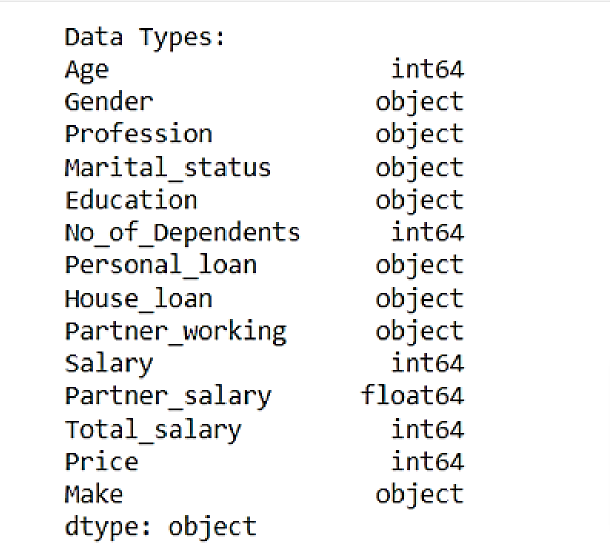
Number of Rows = 1581

Number of Columns = 14

Float Type Data = 1

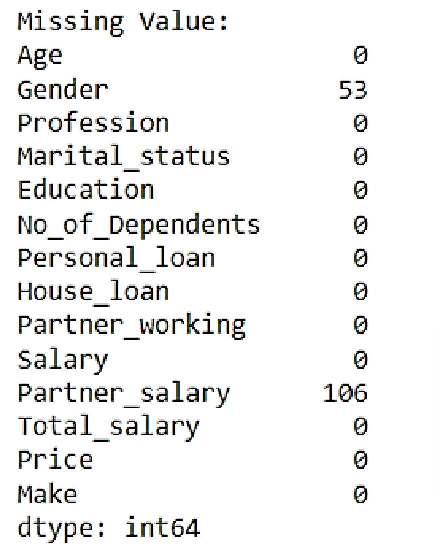
Integer Type Data = 5

Object Type Data = 8



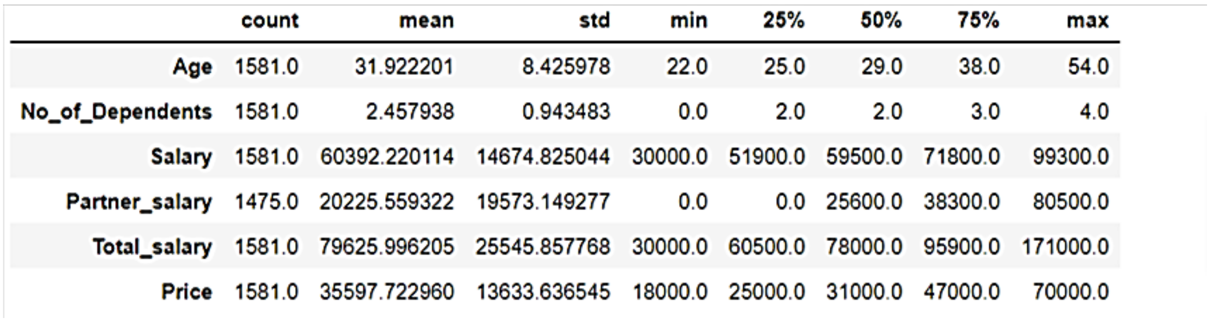
```
Data Types:
Age                   int64
Gender               object
Profession           object
Marital_status       object
Education            object
No_of_Dependents      int64
Personal_loan        object
House_loan           object
Partner_working      object
Salary                int64
Partner_salary      float64
Total_salary          int64
Price                 int64
Make                 object
dtype: object
```

Column 'Gender' have 53 Null Value and 'Partner_salary' have 106 null values



```
Missing Value:
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
```

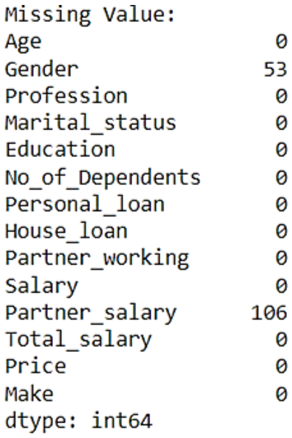
Statstical analysis of the data below



	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1475.0	20225.559322	19573.149277	0.0	0.0	25600.0	38300.0	80500.0
Total_salary	1581.0	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
Price	1581.0	35597.722960	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0

- From the datasets we are described all the mean,median,standard,max,min value,here we get the whole summary of the data
- The Gender Age between 22 to 54 are belong to working people and median age is 29
- The Overall data of Salary given people ranges from 33000 to 99300
- The Total salary is ranges from 30000 to 171000
- The Minimum purchase of the car is 18000,where maximum car purchased 70000.

Missing Values



```
Missing Value:
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
```

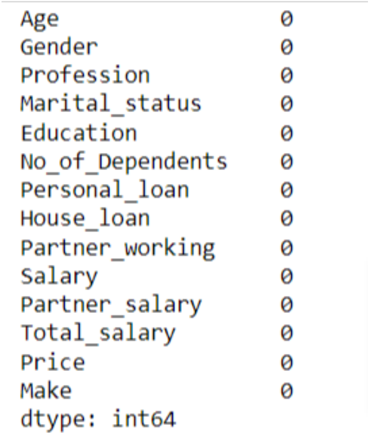
Below Unique number has been identified in the Column ‘Gender’

array(['Male', 'Femal', 'Female', nan, 'Femle'], dtype=object)

- **There are two spelling error found in the column ’Gender’ (Female & Femle)**
- **This spelling error has been corrected and replaced ,below is the result of the correction in the data value .**

array(['Male', 'Female', nan], dtype=object)

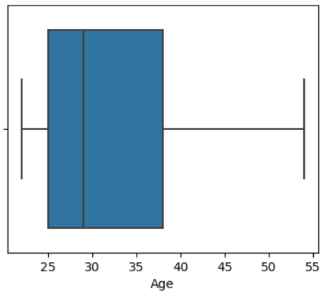
Replaced NaN Value with Partner_salary mean value in the ‘Partner_salary’Column

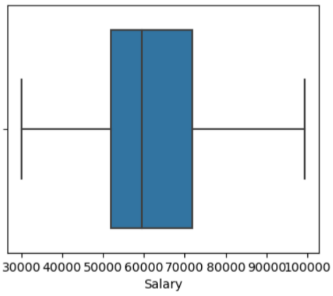


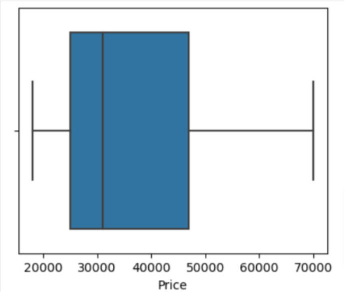
```
Age                 0
Gender              0
Profession          0
Marital_status      0
Education           0
No_of_Dependents    0
Personal_loan       0
House_loan          0
Partner_working     0
Salary              0
Partner_salary      0
Total_salary        0
Price               0
Make      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

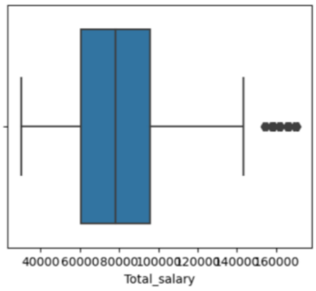
- Now there are no null values in the data set after treatment of ‘Gender’ Column and ‘Partner’Salary’

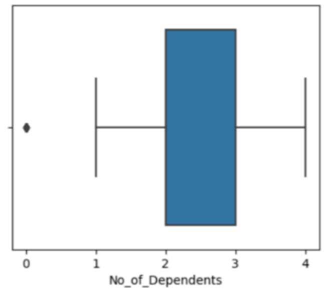
B) Checking of Outliers in the Data

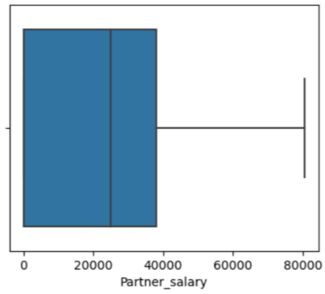






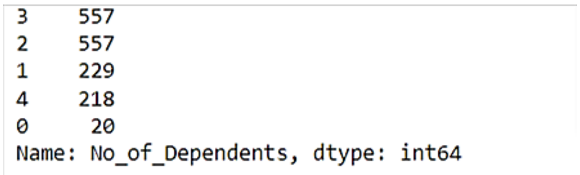






- There are outliers in the ‘No_of_Dependents’column as well as ‘Total_salary as per above boxplot
- I will proceed to treat the outlier for the ‘Total_salary’ only because there is probability of having 0 dependent value and treating dependent could led the mislead the analysis ,

- Kindly refer to below table there are 20 values with 0 de,so I will continue treating only ‘Total_salary’.

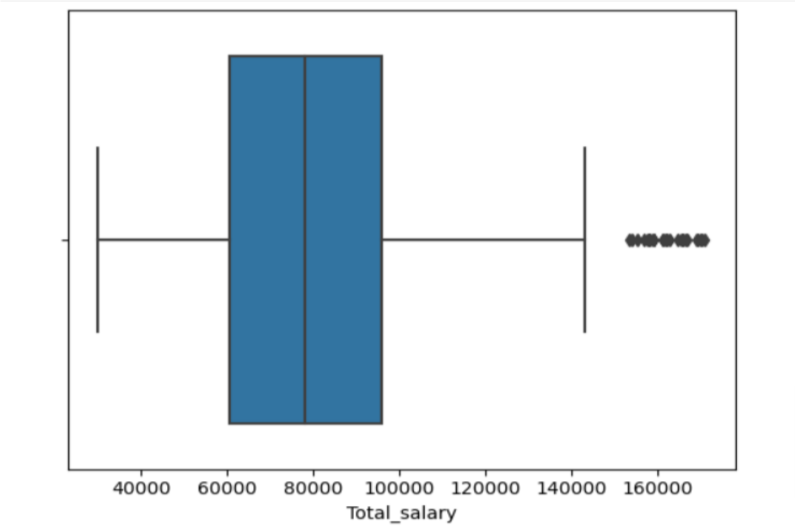


```
3    557
2    557
1    229
4    218
0     20
Name: No_of_Dependents, dtype: int64
```

Also taking mean for the Total Salary in order to avoid creating any manipulative analysis and mean will

provide us overall correct representation of data:

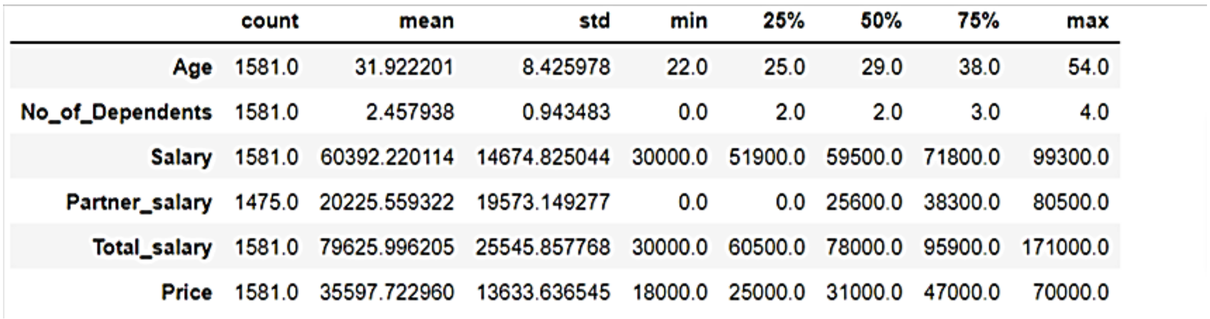
- Mean of the Total Salary is 79625.996
- Treating Outlier(Total Salary)
- Upper Range = 149000
- Lower Range = 7400
- Q1 = 25%
- Q3 = 75%
- Formula to be used = IQR = Q3-Q1
- Lower range =Q1-(1.5*IQR)
- Upper range = Q3+(1.5*IQR)



As we can see from the above plot that outliers has been treated, now there is no outlier for the Total_salary

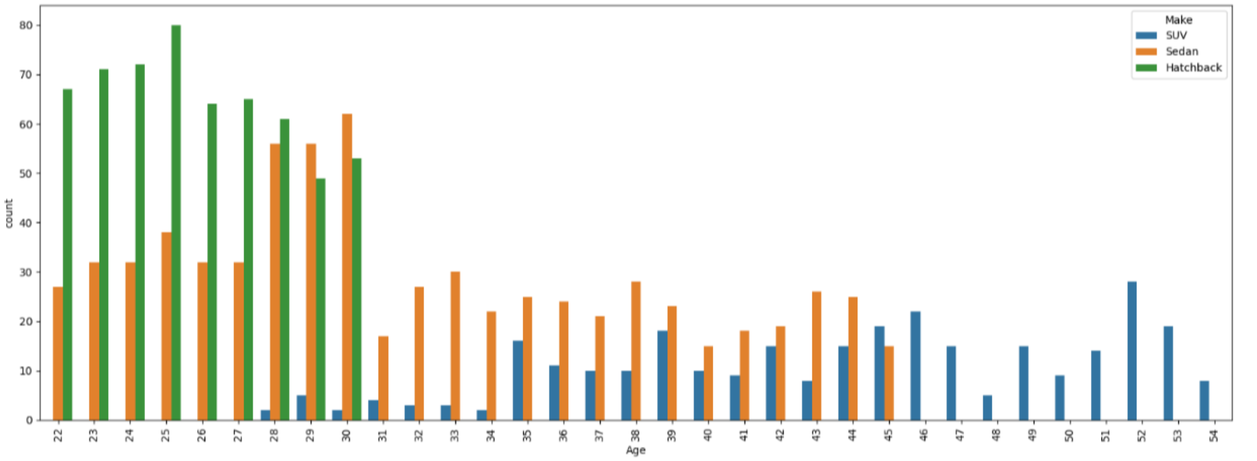
Exploring all the features of the Data separately by using appropriate visualizations and draw insights that can be utilized by the business.

- **Statistical analysis of the data which helps to summarize the data.**



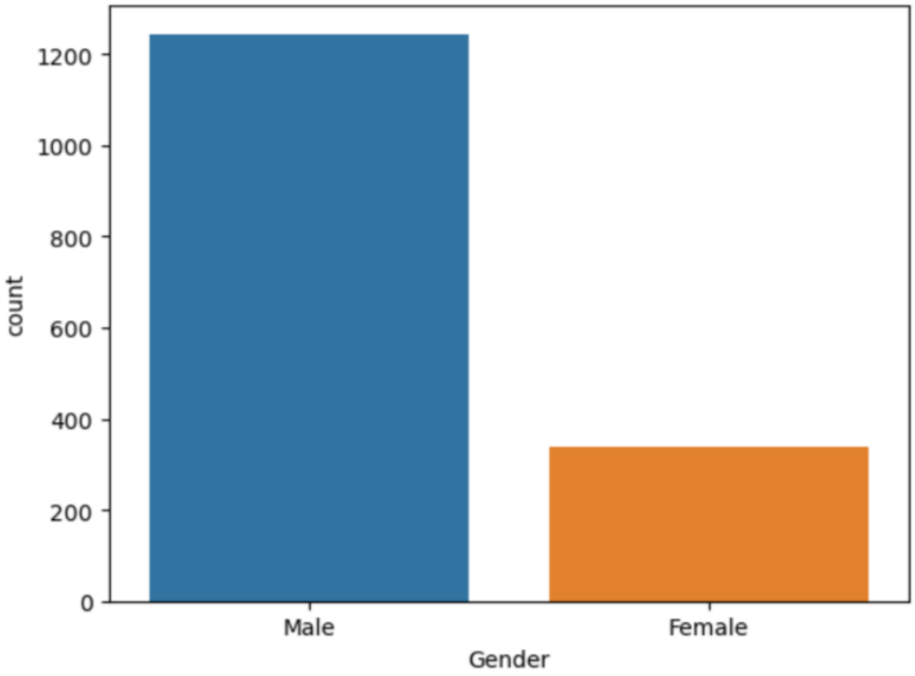
	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1475.0	20225.559322	19573.149277	0.0	0.0	25600.0	38300.0	80500.0
Total_salary	1581.0	79625.996205	25545.857768	30000.0	60500.0	78000.0	95900.0	171000.0
Price	1581.0	35597.722960	13633.636545	18000.0	25000.0	31000.0	47000.0	70000.0

Analyzing the Age Variable



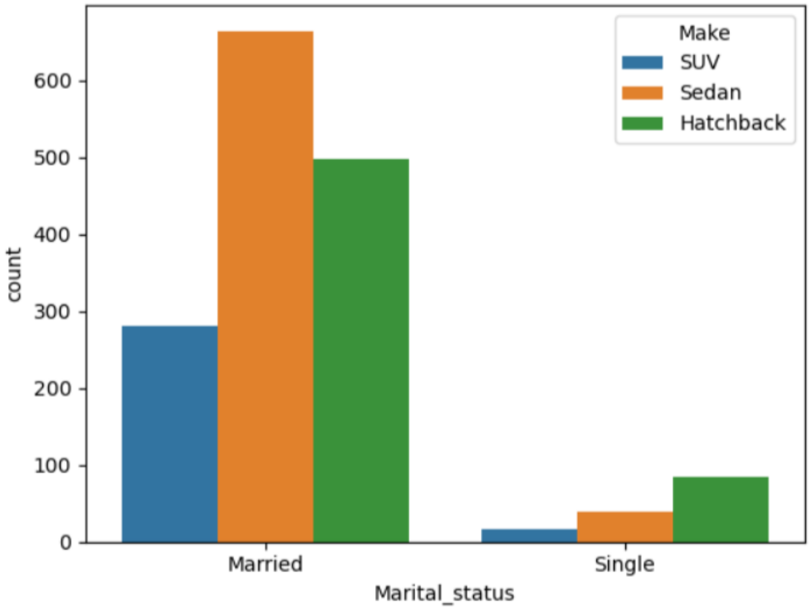
- With reference to above ‘Age’ vs ‘Make’ graph we can conclude that
- Younger age group (20-30) they mostly by Hatchback and Sedan as compared to SUV
- Mid age group (31-45) they mostly buy Sedan and SUV (few sales) with no sales in hatchback to this age group
- Older Age group (46-55) only buy SUV with no sales of Sedan or Hatchback
- There is zero sales of Hatchback from the age group 31 to 46

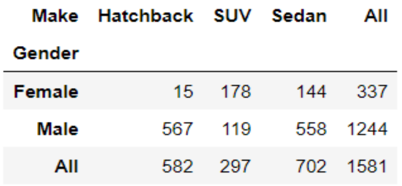
Analysing ‘Gender’ Variables



From the above chart we can say that Male buy more cars than Females

Analysing Gender vs Make Variables



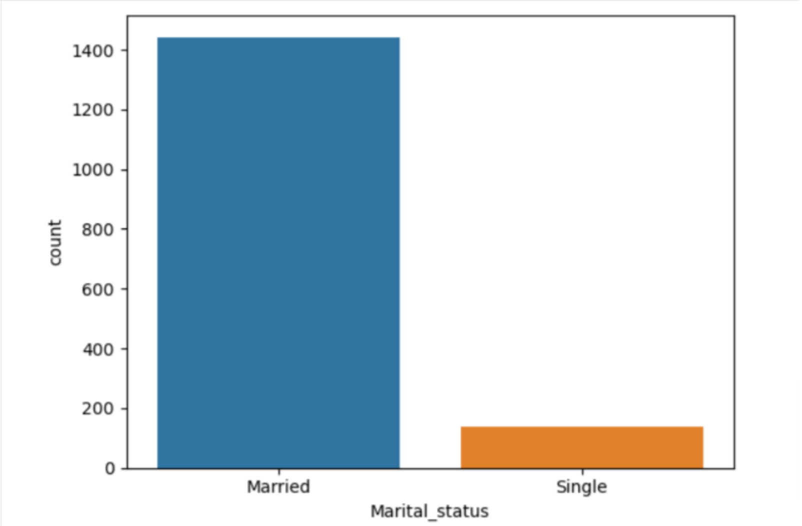


Make	Hatchback	SUV	Sedan	All
Gender				
Female	15	178	144	337
Male	567	119	558	1244
All	582	297	702	1581

With the above data we can conclude that

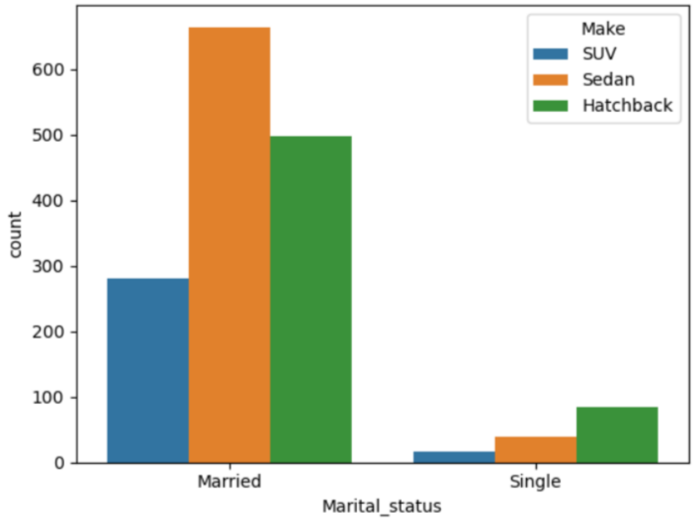
- Females preferred to buy SUV and Sedan and very few Hatchback.
- Males preferred to buy Sedan and Hatchback as compared to SUV.
- 44 Percent of the Customers tend to buy Sedan.

Analysing Marital Status Variable



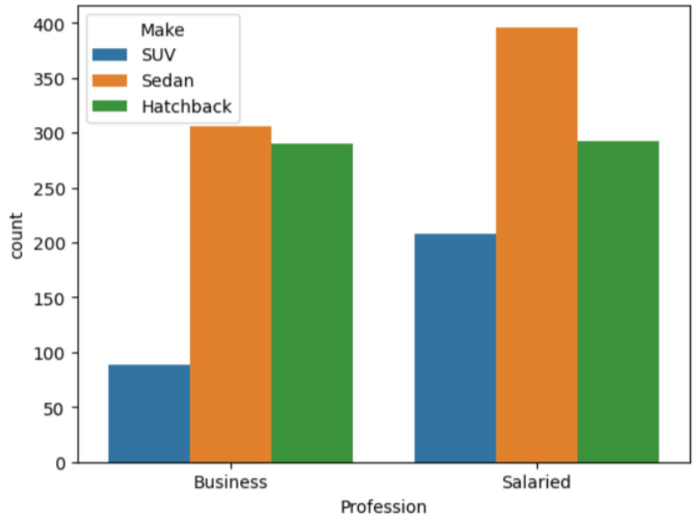
- As per above graph we can say married couple tends to buy more cars as compared to the Singles

Analysing Marital Status and Make Variables



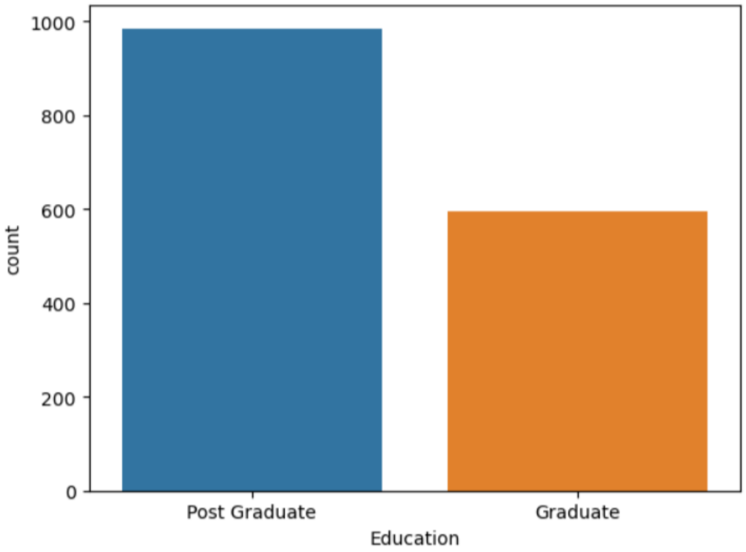
- Married Couples preferred to buy Sedan and Hatchback as compared to SUV's.
- Single preferred to buy Hatchback as well as very little preference to Sedan and SUV.

Analysing Profession vs Make Variables



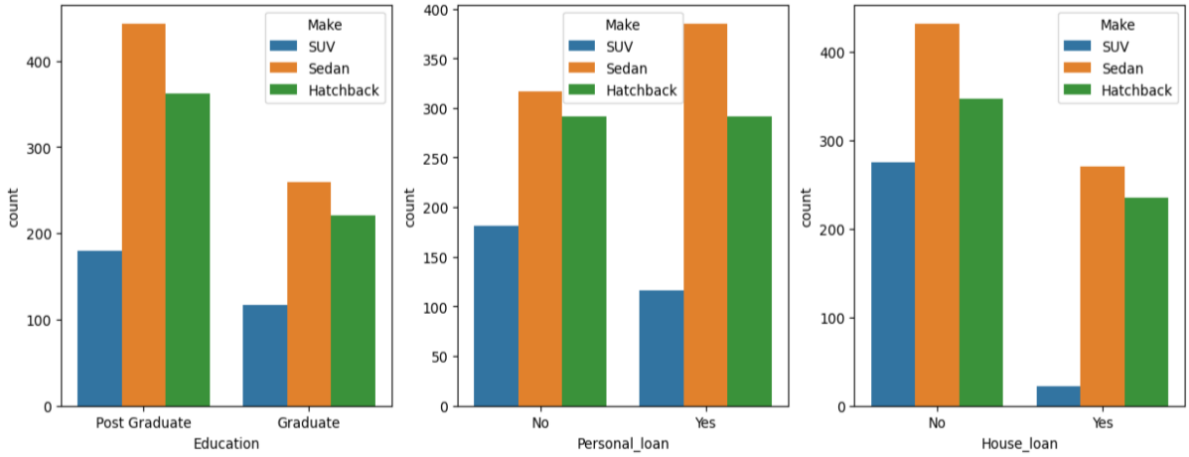
- From the above graphs we can conclude that Salaried people buy more cars as compared to the business profession group
- Salaried prefers to buy Sedan as compared to Hatchback with little preference to SUV.
- Whereas Business professionals prefer to buy Sedan and Hatchback as compared to SUV.
- SUV is not much preferred by the Business people.

Analysing Education Variables



As Per the above graph we can say that Post Graduate buys more Cars than Graduates.

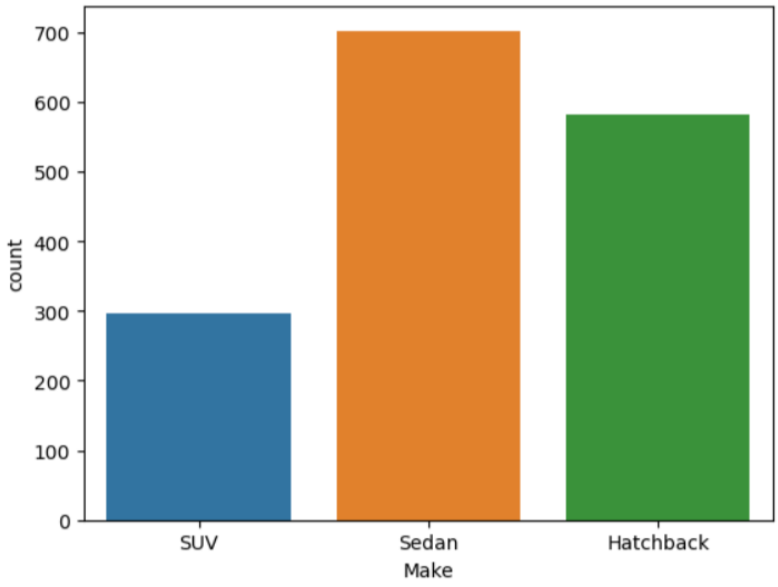
Analysing the Education, Personal Loan and House loan Variables.



As per above data and graph we can conclude that

- Postgraduates student mostly buy Sedan and Hatchback with lesser preferences to SUV as compared to Make
- Graduates also prefers to buy Sedan and Hatchback as compared to SUV.
- Customers with and without Personal_loan buys more Sedan and Hatchback as compared to SUV.
- Customers with and without House_Loan buys more Sedan and Hatchback with lesser preference given to SUV.
- Customers without house loan buys more SUV as compared to customer with House_loan.

Analysing Make Variable

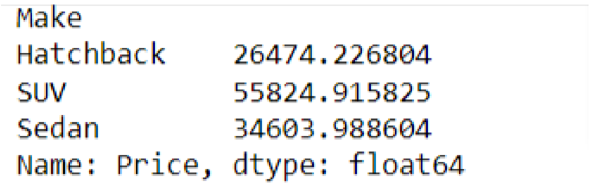


- **With above plot we can say Company is making more Sedan and Hatchback as compared to SUV.**
- **Hatchback manufacturing takes second place after Sedan.**

With above all the data we can concludes that

- **Average age for buying cars is 29**
- **Average price to buy car is 35597.722**

Average price for the Make is as below



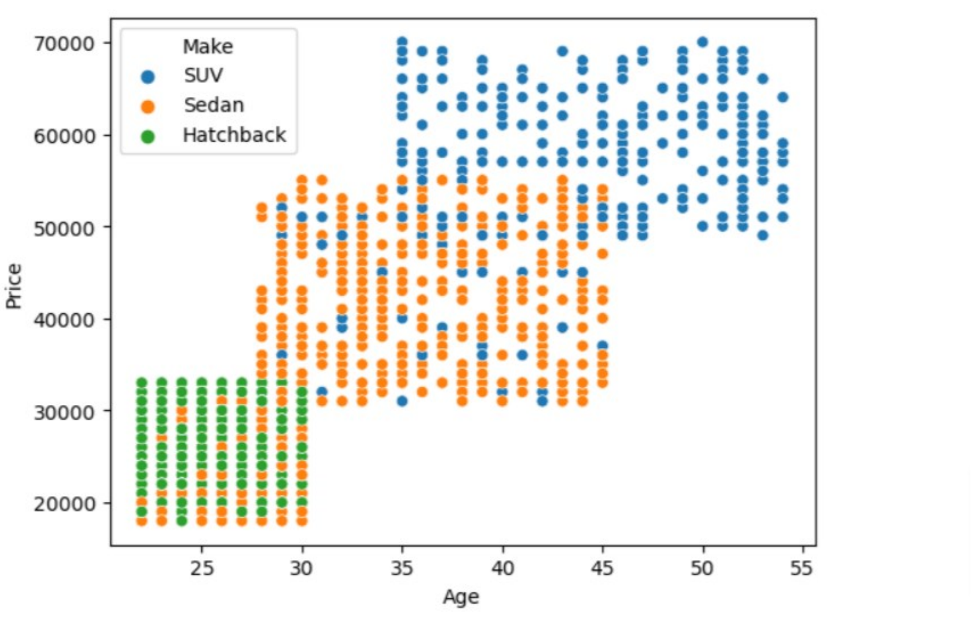
```
Make
Hatchback     26474.226804
SUV           55824.915825
Sedan         34603.988604
Name: Price, dtype: float64
```

C) Bivariate Analysis-

Explore the relationship between all numerical variables - Explore the correlation between all numerical variables - Explore the relationship between categorical vs numerical variables

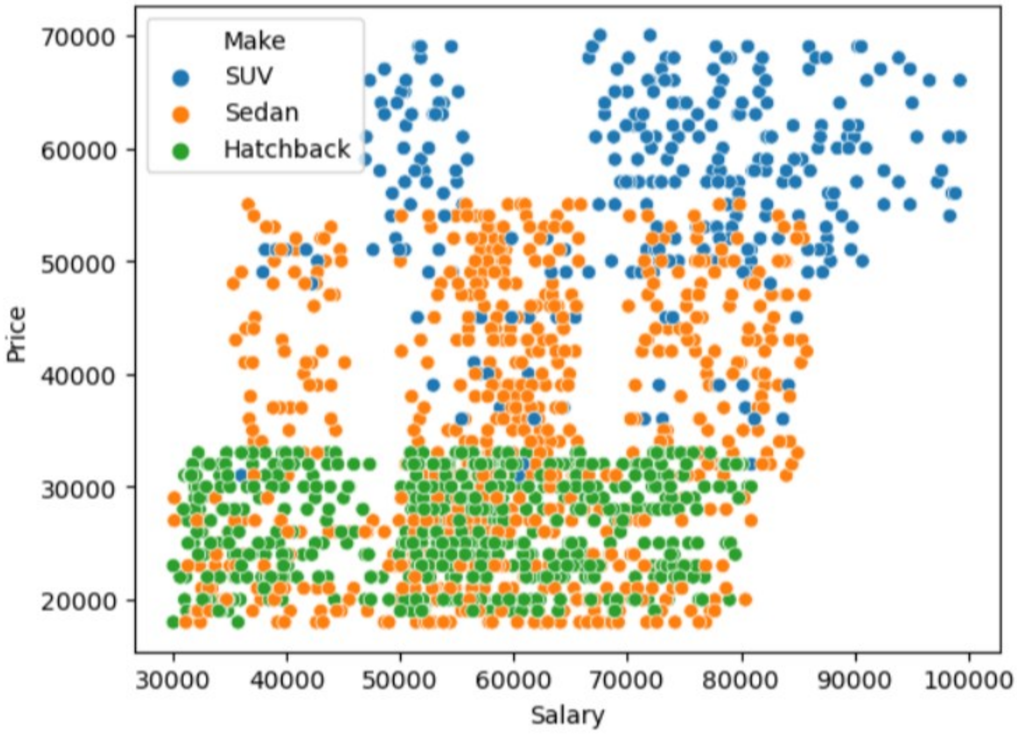
Understanding the relationship between among the variables in the dataset to gain deeper insights

Establishing co-relation between Age & Price.



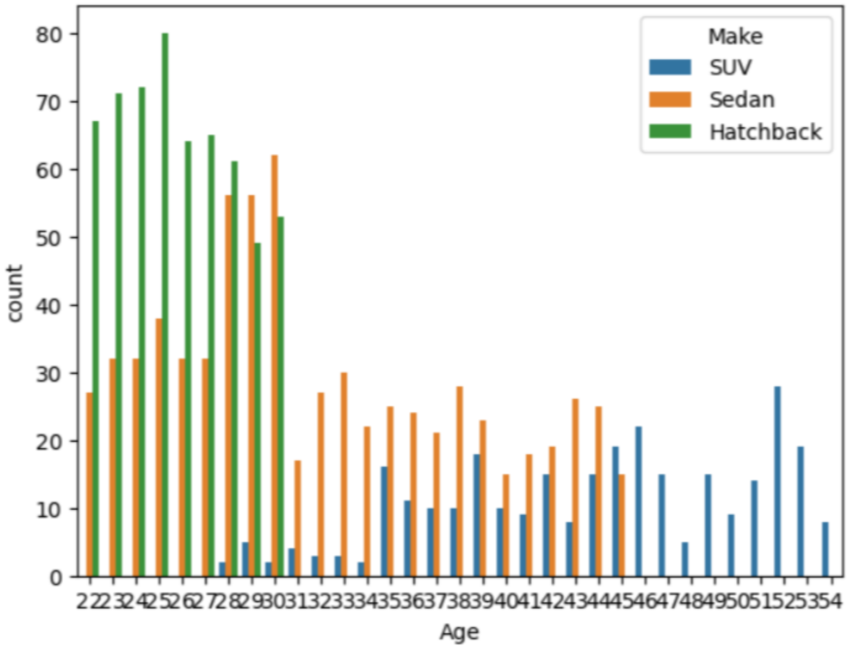
- With above figure as reference ,there is positive correlation between age of Customer and amount of money spent on the buying cars and as the customer age increases they tends to buy more expensive cars, this clearly shown in the above scatterplot
- As the age of the Customer increases the amount of money spent on this automotive sector also increases ,Age and Price are positively correlated .

Establishing co-relation between Salary & Price



Insights for the above scatter plot reveals that as the Salary of Individual increases then price of the cars is also increasing, Hence, Price and Salary are Positively Correlated.

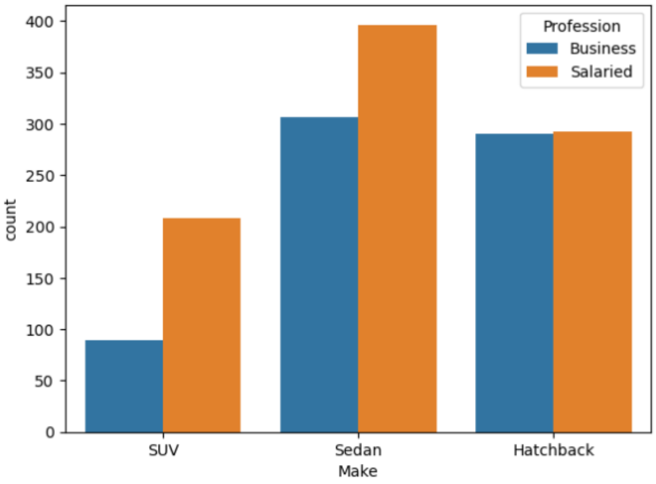
Establishing co-relation between Age & Make



With reference to above 'Age' bs 'Make' graph we can conclude that

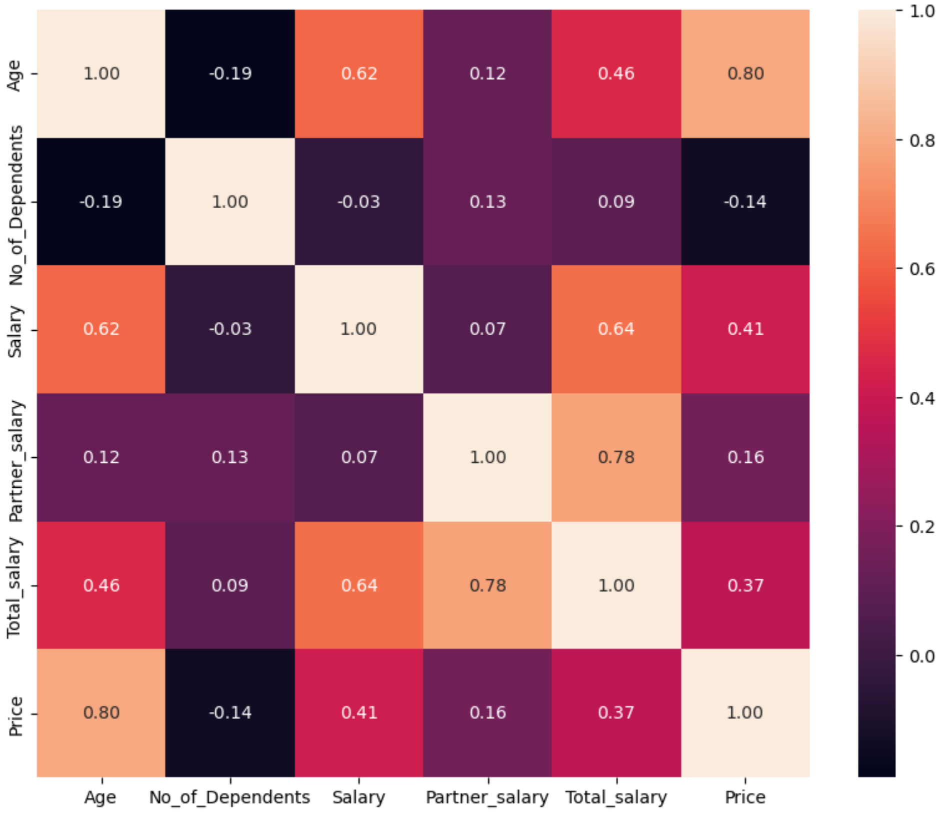
- **Younger age group (20-30) they mostly buy Hatchback and Sedan as compared to SUV**
- **Mid age group (31-45) they mostly buy Sedan and SUV (Few sales only) with no Sales of hatchback to this age group.**
- **Older Age group (46-55) only buy SUV with no sales of Sedan or Hatchback.**
- **There is zero sales of Hatchback from the age group 31 to 46.**

Establishing co-relation between Professional & Gender



- Male Business professional first choice is Hatch Back and Second is Sedan and comparatively less preferred is SUV.
- Whereas female Business Professions prefer to buy Sedan as well as SUV with Similar interest in Make
- Salaried Female Customer first choice is SUV whereas Second choice is Sedan and with fewer sales of hatchback amongst them.
- Salaried Male Customer first choice is either Sedan OR Hatchback as compared to SUV,SUV is comparatively less demanding amongst them .

Establishing co-relation with Heatmap with Price,Age,No_of Dependants Salary, Partner salary & Total salary



1 = Perfect Correlation

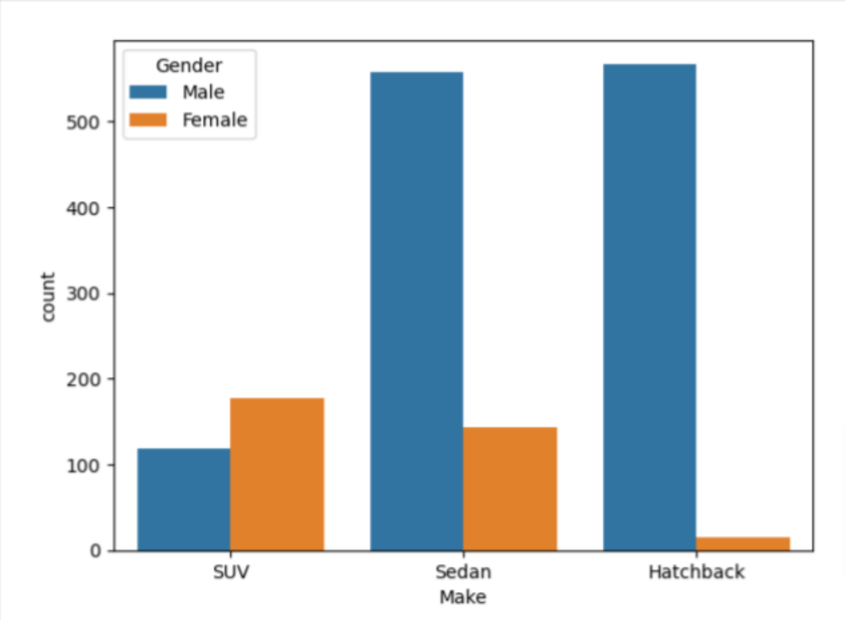
-1 to 0 = Negative Correlations

0 to 1 = Positive Correlation

- There is strong Correlation between Age and Price
- There is strong Correlation between Partner Salary and Total Salary
- There is medium correlation between Salary and Price
- There is medium correlation between Total Salary and Price

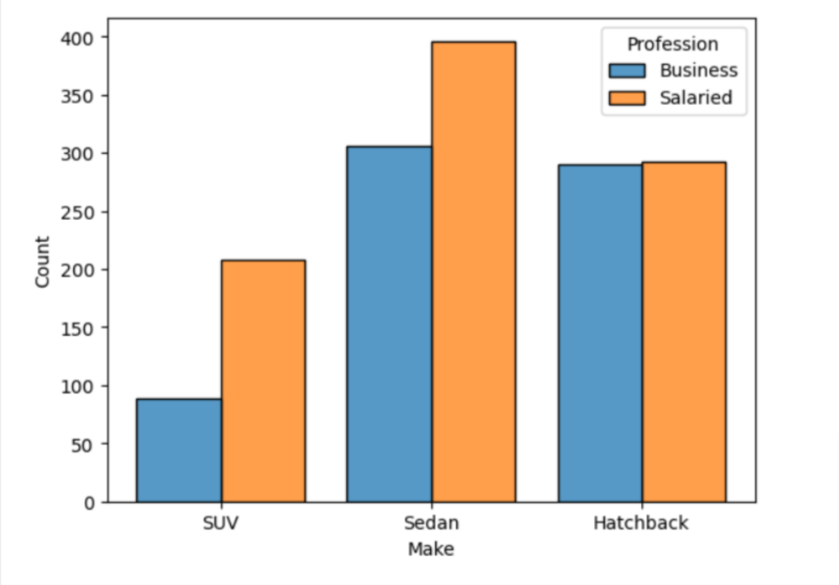
D.) Exploring the Data (Key Questions)

1. Do men tend to prefer SUVs more compared to Women?



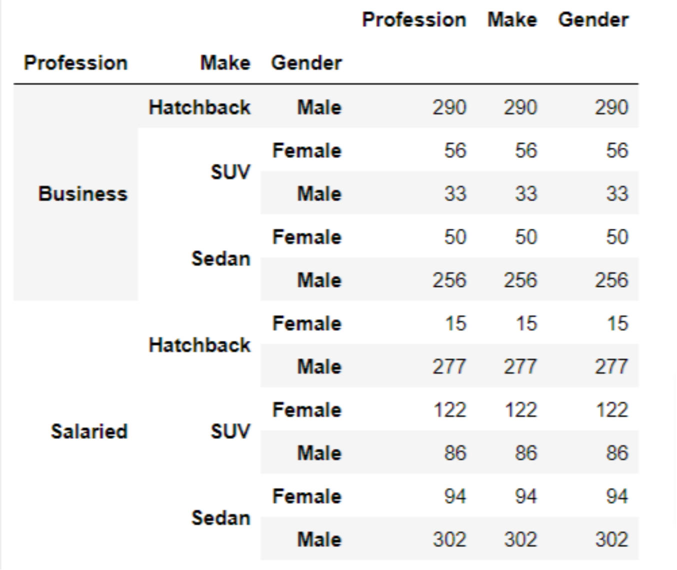
= Both Male and female preferred to buy SUV, but Women prefer SUV by a larger margin then a Men.

2. What is the likelihood of a salaried person buying a Sedan?



= Based on the Histogram, Salaried person buys more Sedan than Business Professionals.

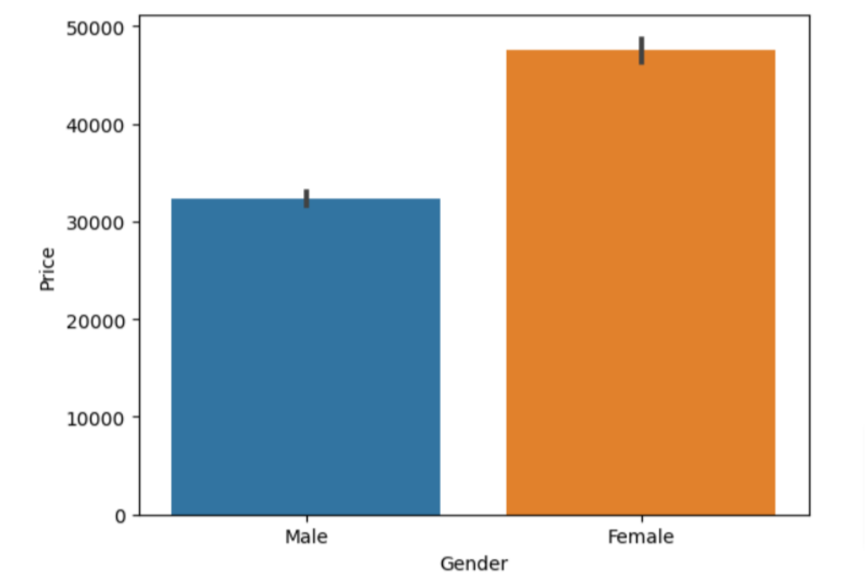
3. What evidence or data supports Sheldon Cooper's claim that a salaried male is an easier target for a SUV sale over a Sedan sale?



			Profession	Make	Gender
Profession	Make	Gender			
Business	Hatchback	Male	290	290	290
	SUV	Female	56	56	56
		Male	33	33	33
	Sedan	Female	50	50	50
		Male	256	256	256
Salaried	Hatchback	Female	15	15	15
		Male	277	277	277
	SUV	Female	122	122	122
		Male	86	86	86
	Sedan	Female	94	94	94
		Male	302	302	302

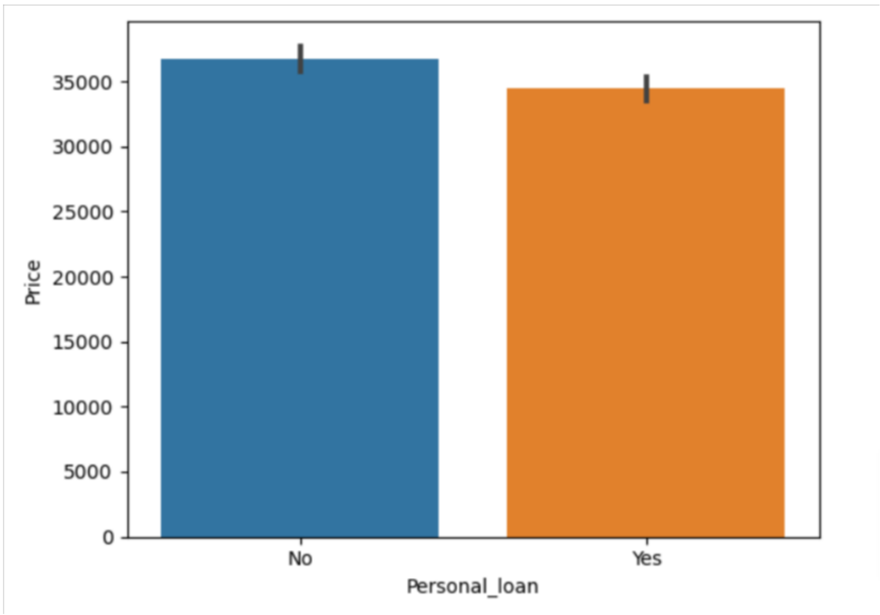
= Based on the data above, Sheldon Cooper is wrong, people prefer Hatchbacks and Sedans rather than SUVs. SUVs are not as popular with salaried Male as they were with Female.

4. How does the the amount spent on purchasing automobiles vary by gender?



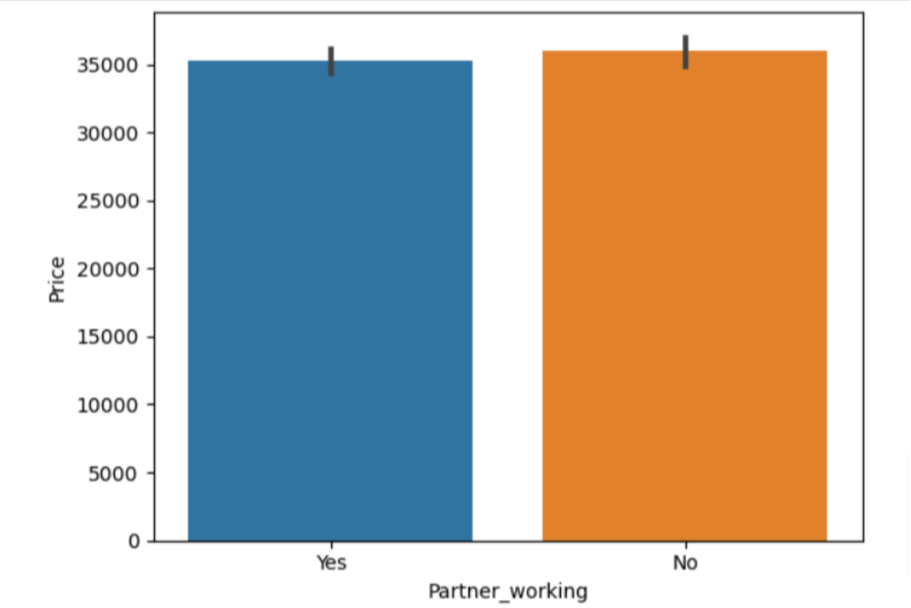
- A woman has overall purchased a more expensive car than a man based on the above calculation Gender vs Price (Mean/Average).
- According to the above plot, customers who don't take loans buy more expensive cars than those who do

5. How much money was spent on purchasing automobiles by individuals who took a personal loan?



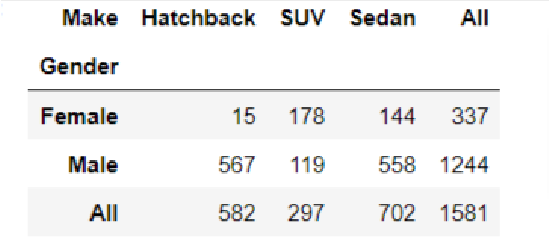
Using the plot above, we can see that customers who do not take out personal loans buy more expensive cars.

6. How does having a working partner influence the purchase of higher-priced cars?



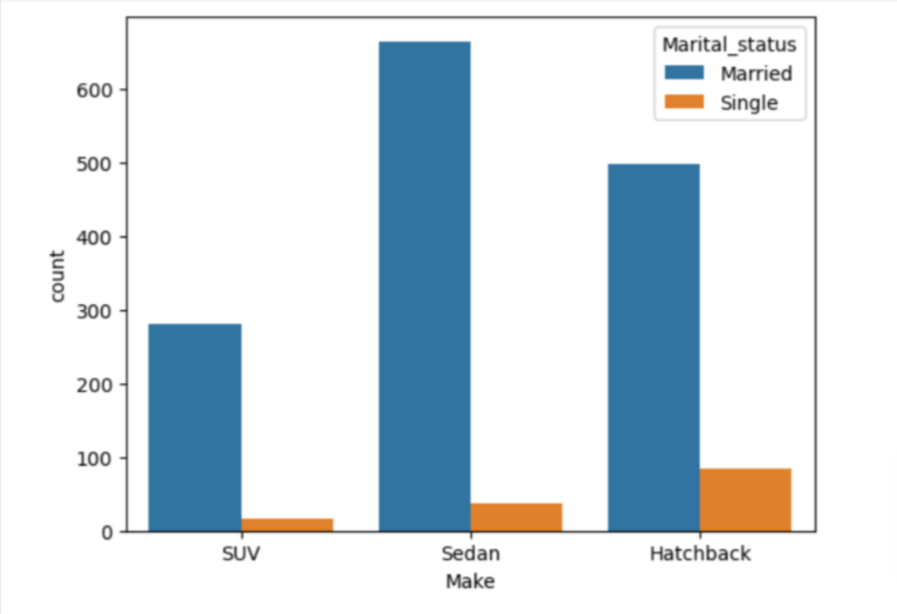
There is only a marginal difference between them, showing that customers whose partners are not working tend to buy more expensive cars. It does not matter if your partner is working or not working.

Problem -Actionable Insights & Recommendations



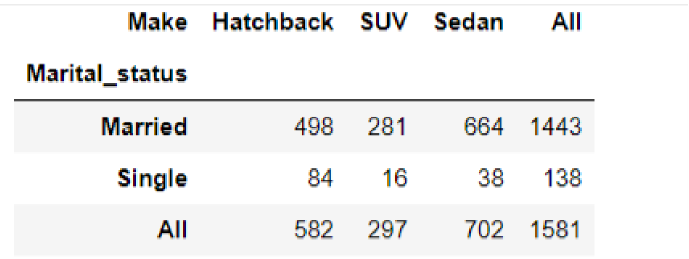
Make	Hatchback	SUV	Sedan	All
Gender				
Female	15	178	144	337
Male	567	119	558	1244
All	582	297	702	1581

- As per below crosstab information with Gender aspect, we can conclude that total Male customer buys more cars with highest number of Hatchback followed by Sedan on second number and SUV takes third position in buying preference.
- For Female customer, they buy more SUV as compared to Sedan, Hatchback takes last position in the buying preference list for Females.



Insights for the above chart ;

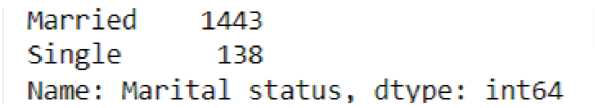
- Less sales amongst females with business professionals for the Hatchback. Their First preference is to buy an SUV and Second choice is a Sedan
- The first choice of salaried women is also SUVs, followed by sedans, and few choose hatchbacks.
- Hatchbacks are the most popular choice among male business professionals, followed by sedans and fewer SUVs.
- Among salaried males, the most preferred vehicle is a sedan, followed by a hatchback, and the most preferred vehicle is an SUV.
- In comparison to hatchbacks, married customers are more likely to buy sedans, and SUVs are the last choice for married customers.
- A single customer buys more Hatchbacks than Sedans, while SUVs take last place in the choice list



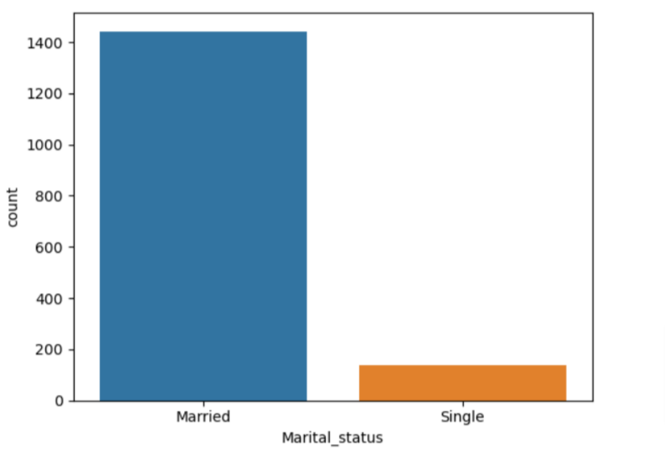
Make	Hatchback	SUV	Sedan	All
Marital_status				
Married	498	281	664	1443
Single	84	16	38	138
All	582	297	702	1581

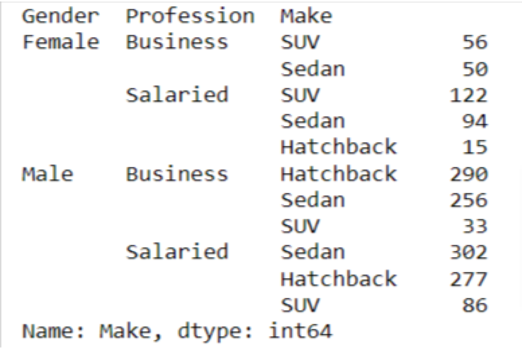
Insights can be derived from the above table .

- There are total 1443 married and 138 Singles ,henceforth there are more married customers in the company record .
- Married business professional they prefer to buy Sedan followed by Hatchback and SUV became last choice for them .
- Single business professional tends to buy more Hatchback than Sedan with very few prefers to buy SUV.
- Salaried and Married prefers to buy more Sedan than Hatchback and SUV is the last choice for them .
- Salaried and single prefers to buy Hatchback followed by Sedan with fewer choice for SUV.



```
Married     1443
Single       138
Name: Marital status, dtype: int64
```



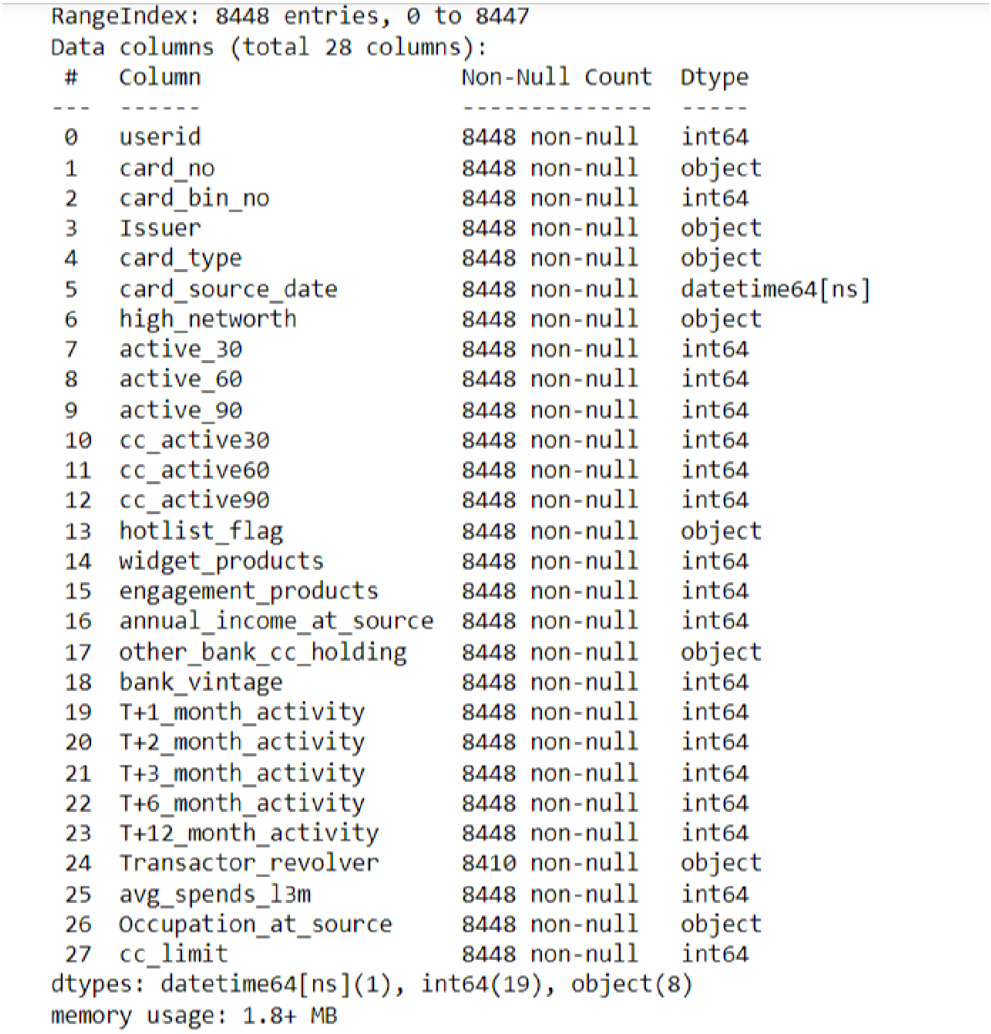


```
Gender  Profession  Make
Female  Business    SUV           56
                    Sedan         50
        Salaried    SUV          122
                    Sedan         94
                    Hatchback     15
Male    Business    Hatchback    290
                    Sedan        256
                    SUV           33
        Salaried    Sedan        302
                    Hatchback    277
                    SUV           86
Name: Make, dtype: int64
```

Problem 2 – Framing Analytics Problem

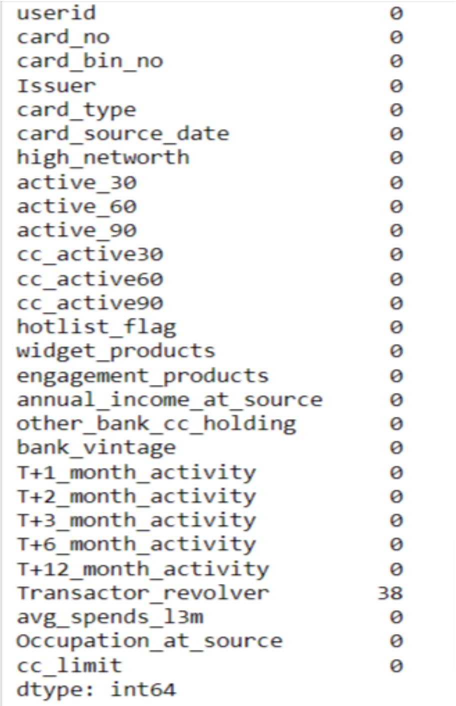
Analyse the dataset and list down the top 5 important Variables along with the business justifications .

- **Data Set is =** godigit_cc_data
- **There are 8 object type,19 integer type and data time type variable in the data set**
- **There are no duplicates in the data**
- **There are Data :** Rows 8448 / Columns 28



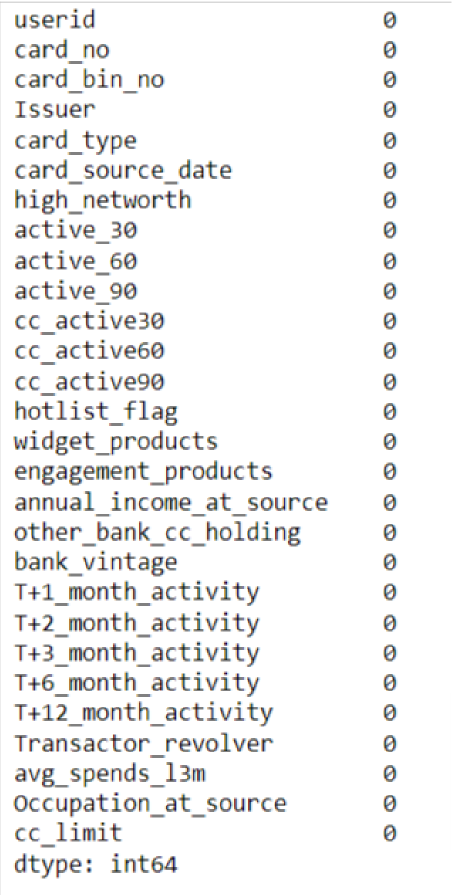
```
RangeIndex: 8448 entries, 0 to 8447
Data columns (total 28 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   userid                  8448 non-null   int64
 1   card_no                 8448 non-null   object
 2   card_bin_no             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
 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)
memory usage: 1.8+ MB
```

There are total 38 missing values in Transactor_revolver, See the table below



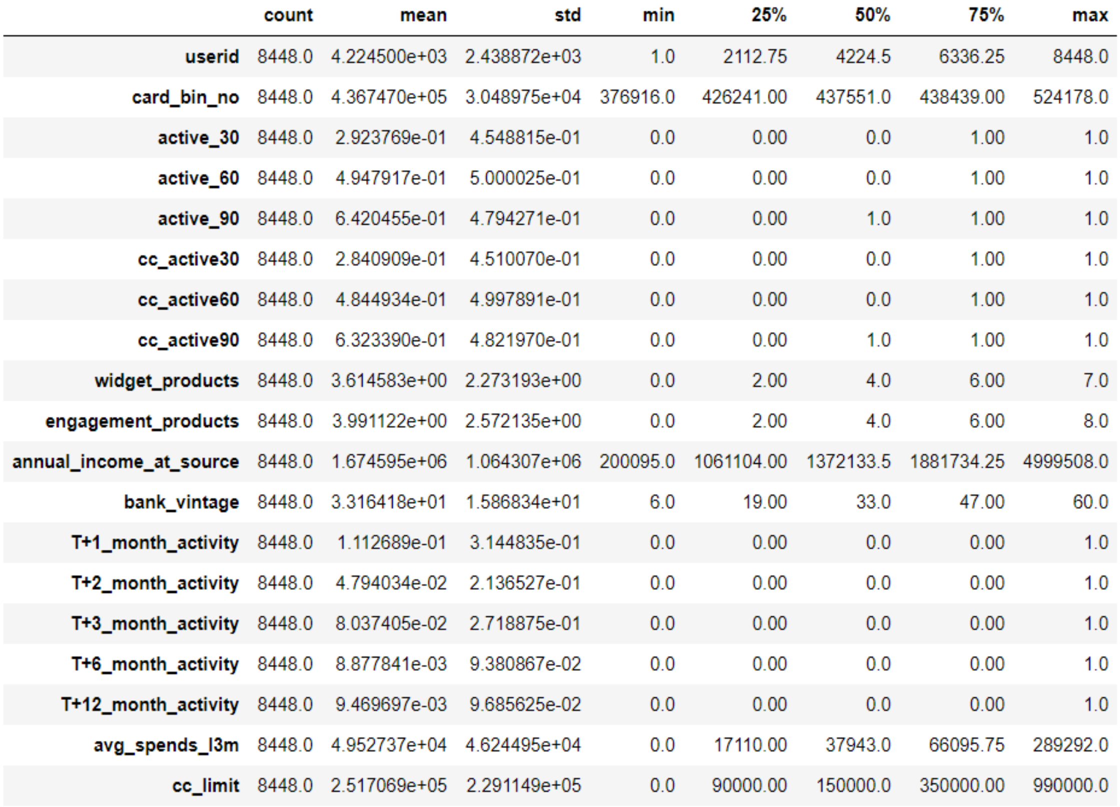
```
userid                     0
card_no                    0
card_bin_no                0
Issuer                     0
card_type                  0
card_source_date           0
high_networth              0
active_30                  0
active_60                  0
active_90                  0
cc_active30                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
T+6_month_activity         0
T+12_month_activity        0
Transactor_revolver       38
avg_spends_l3m             0
Occupation_at_source       0
cc_limit                   0
dtype: int64
```

- Treating Transactor_revolver,this will be done by replacing the missing values with the mode of ‘Transactor_revolver’
- The mode of Transactor_revolver is ‘T’ , Replacing the missing values with T.
- After treatment there is no null values.



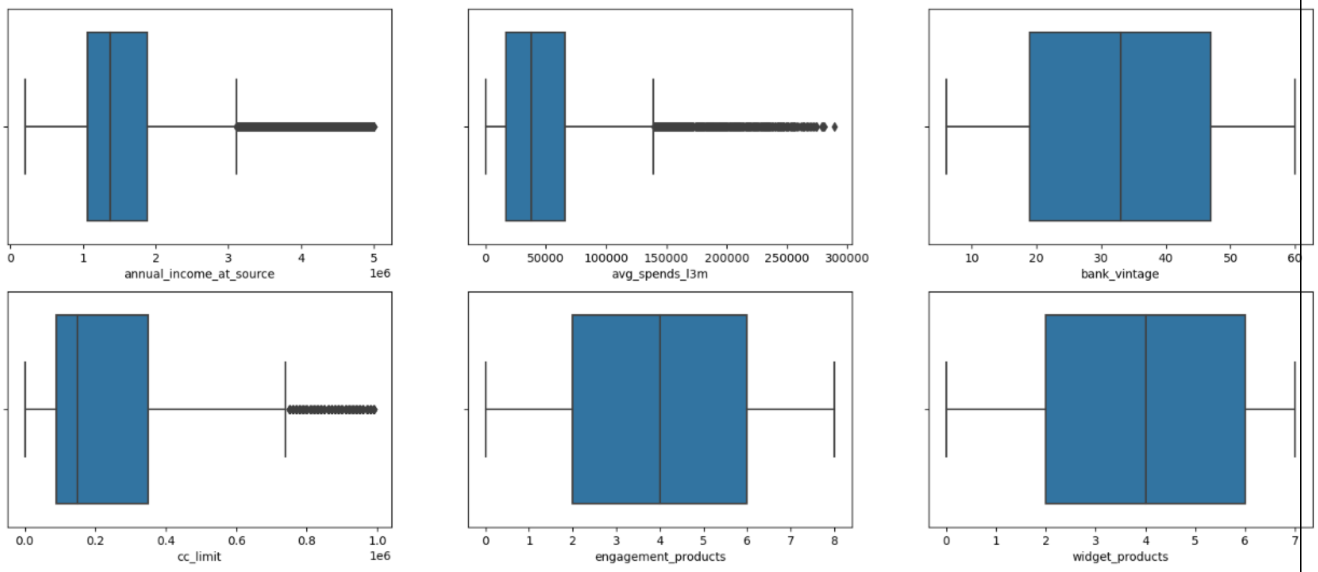
```
userid                   0
card_no                  0
card_bin_no              0
Issuer                   0
card_type                0
card_source_date         0
high_networth            0
active_30                0
active_60                0
active_90                0
cc_active30              0
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
T+6_month_activity       0
T+12_month_activity      0
Transactor_revolver      0
avg_spends_l3m           0
Occupation_at_source     0
cc_limit                 0
dtype: int64
```

Statistical Analysis of the Data



	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

Checking Outliers of the Data



- There are outliers in ‘annual income at source’, avg_spends_13m’,’cc_limit’
- Outliers will not be treated as it will impact the analysis.

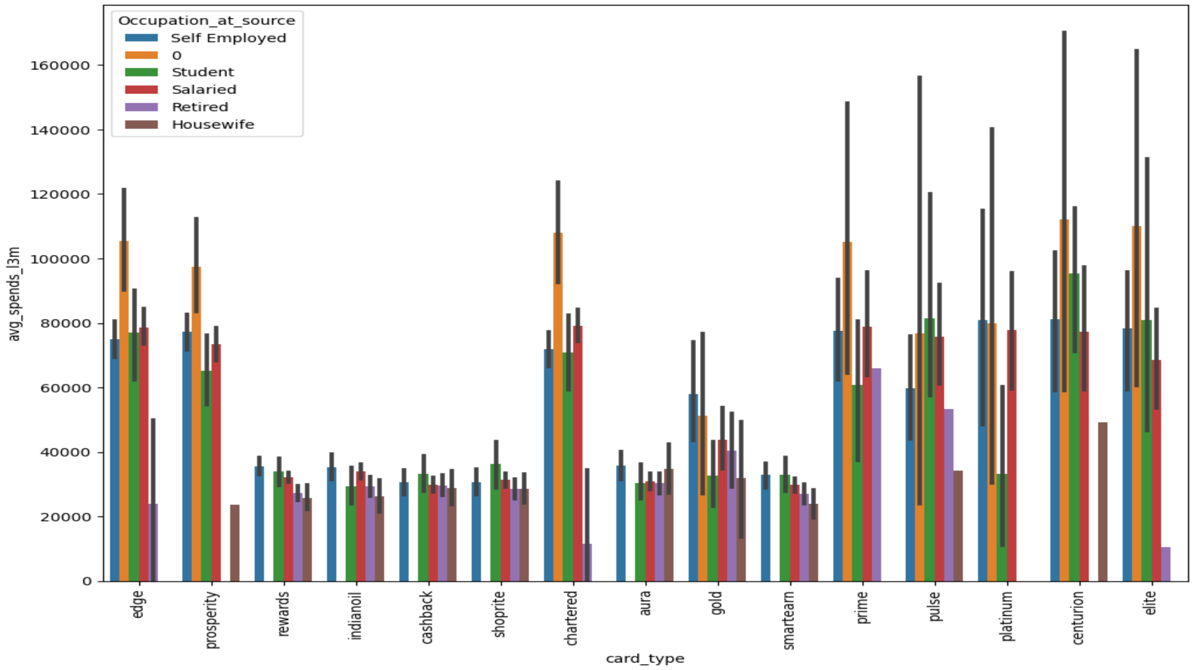
Framing few Analytics problem which can include in their Strategies.

1. What is the relationship between active user ID and total spend amount?
2. Analysis of spending by customers with multiple bank credit cards to determine which bank cards are most preferred and why?
3. Does credit card limit restrict customer spending and are there any correlations.?
4. What is the relationship between customers' average spending and their annual income?
5. Does spending increase as annual income increases? What is the spending pattern of different customers with different occupations?
6. According to the type of customer, which card types should be incorporated? and it can be done by analysing and evaluating past records.

Most important Variables from the data

1) Card Type

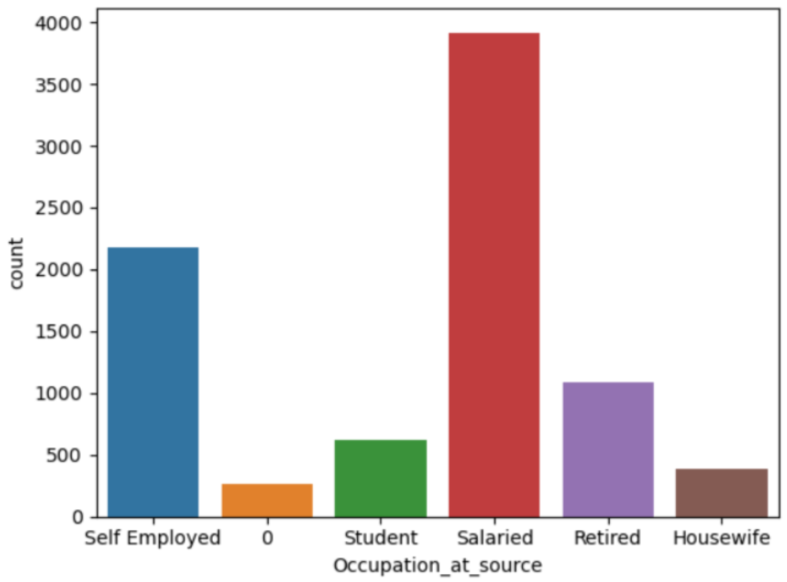
   Card Type are the most important Variable ,Based on above data Salaried customer and high profile customer prefers to have more reward cards,Card_Types helps to increase the avg spent.



Cards usage based on occupations

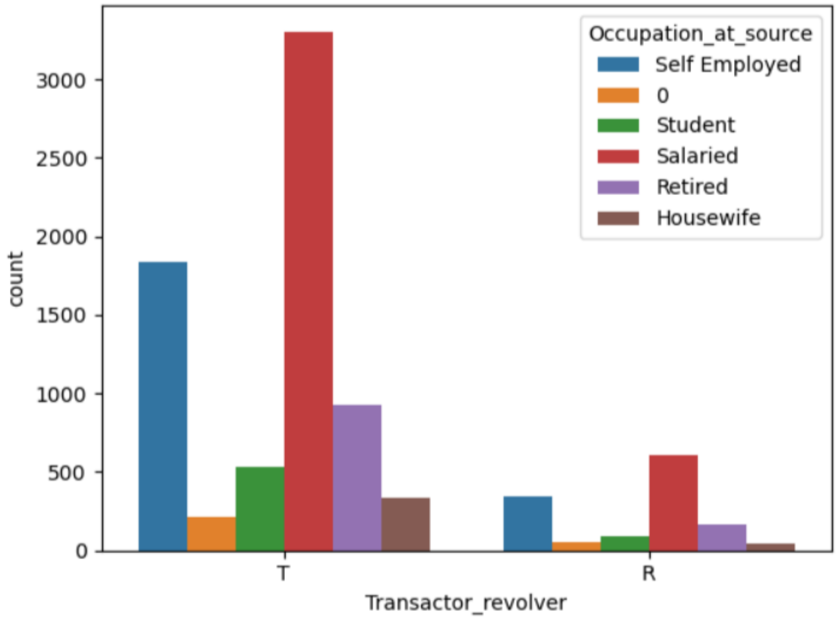
- Housewife – 384
- Retired – 1089
- Salaried – 417
- Self Employed- 2175
- Students - 621

2) Occupation at Source



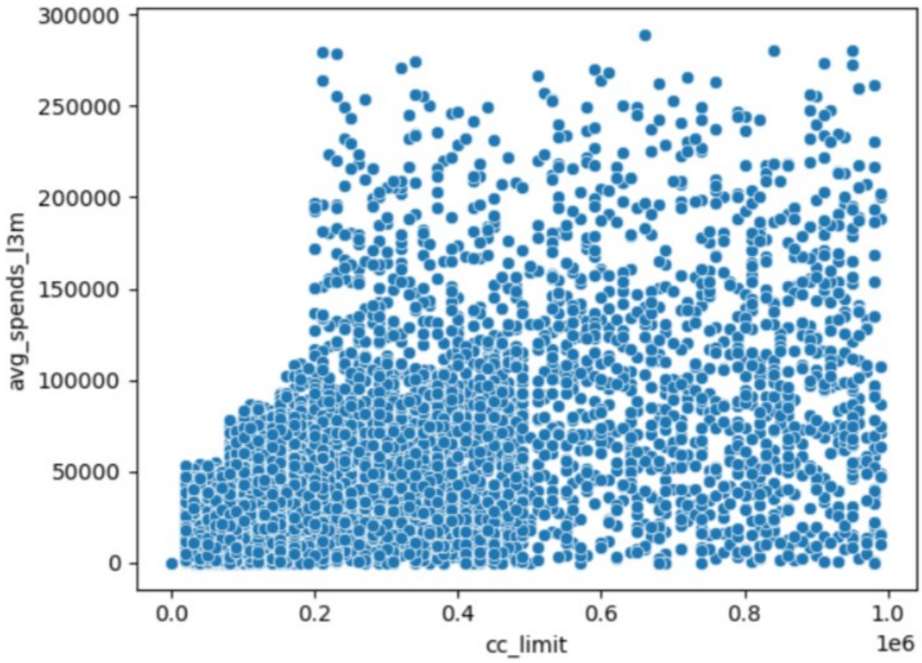
- Bank Should Strategize to sell the correct card_type as per the occupations which will increase the avg_spends and will eventually increase the profits.
- Highest CC Customers are self Employed and salaried and their avg spent is high .

3) Transactor_revolver



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

4) Annual Income at source



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

5) Avg_Spends_13m

- Avg Spends is very important variable as this will give window to the bank for generating profits.
- These are directly proportional to the card_Type ,occupation at source and Transactor_revolver.
- This is also proportional to the interest received ,more spending means more interest and thus becomes the mode for the revenue generation to the bank .
- Rise in income has positive relation with average spends,with increasing salary people tends to spend more.