SMDM PROJECT REPORT (CODED)

DSBA

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.

Objective :-

They want to analyze the data to get a fair idea about the demand of customers which will help them in enhancing their customer experience. Suppose you are a Data Scientist at the company and the Data Science team has shared some of the key questions that need to be answered. Perform the data analysis to find answers to these questions that will help the company to improve the business.

Data Overview

- The Dataset is austo-automobile.csv.
- The dataset contains 1581 rows and 14 columns.
- It has 1 float, 5 integer and 8 object datatypes.

```
df.info()
```

```
<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
```

- We can notice that there are null values in Gender and Partner_salary.
- We will check the data in the categorical columns:

```
In [229]: df['Gender'].unique()
Out[229]: array(['Male', 'Femal', 'Female', nan, 'Femle'], dtype=object)

In [230]: df['Profession'].unique()
Out[230]: array(['Business', 'Salaried'], dtype=object)

In [231]: df['Marital_status'].unique()
Out[231]: array(['Married', 'Single'], dtype=object)

In [232]: df['Education'].unique()
Out[232]: array(['Post Graduate', 'Graduate'], dtype=object)

In [233]: df['Personal_loan'].unique()
Out[233]: array(['No', 'Yes'], dtype=object)

In [234]: df['House_loan'].unique()
Out[234]: array(['No', 'Yes'], dtype=object)

In [235]: df['Partner_working'].unique()
Out[235]: array(['Yes', 'No'], dtype=object)

In [236]: df['Make'].unique()
Out[236]: array(['SUV', 'Sedan', 'Hatchback'], dtype=object)
```

- There are misspelt values in Gender which can be fixed by the following

```
In [237]: df['Gender'].replace('Femal','Female',inplace=True)

In [238]: df['Gender'].replace('Femle','Female',inplace=True)
```

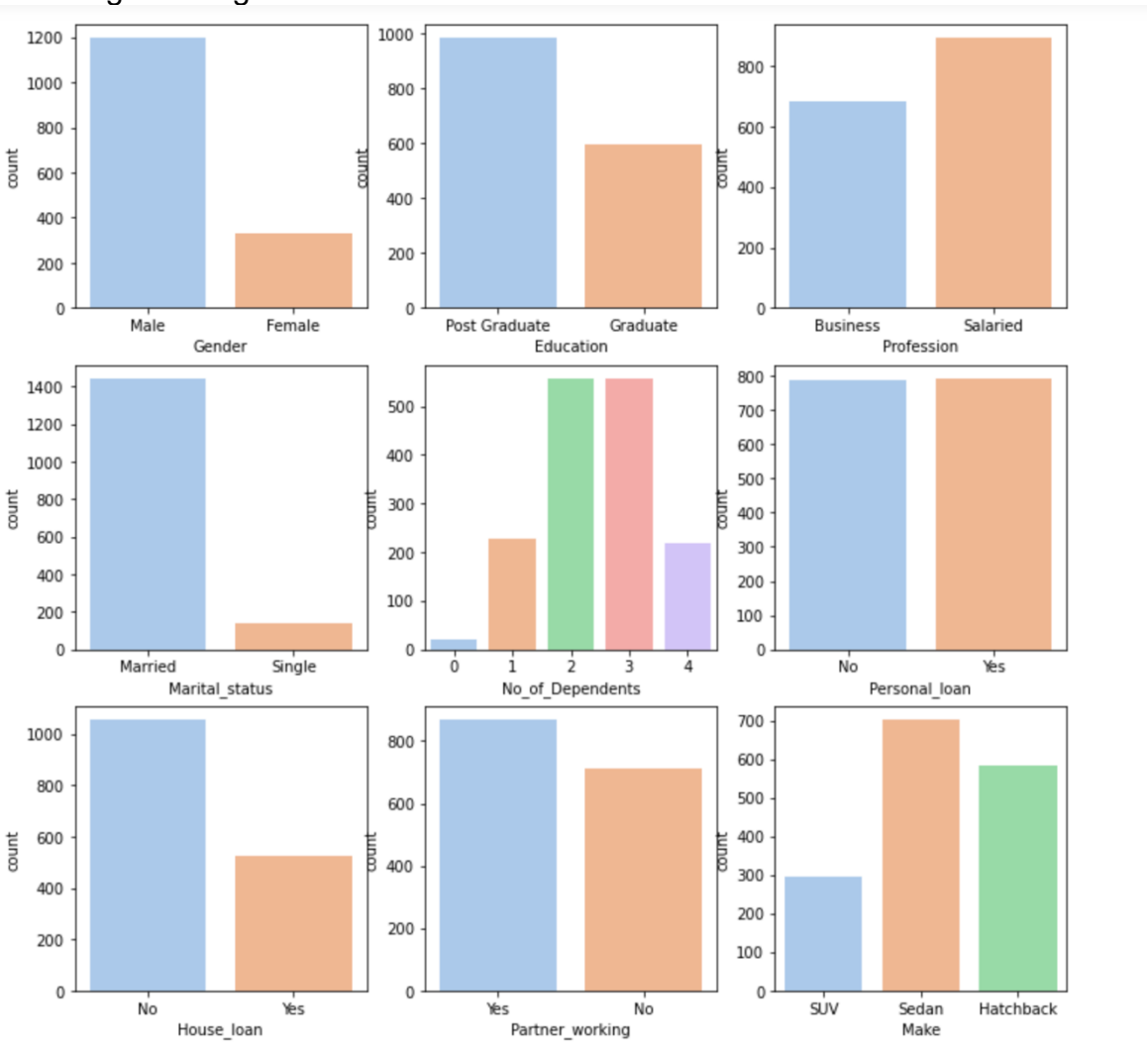
- The above also shows a few null values

```
In [239]: df.isnull().sum()

Out[239]: 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
```

- The null values in Partner_salary can be replaced by 0.0 while the gender can be left blank

Univariate Analysis:-

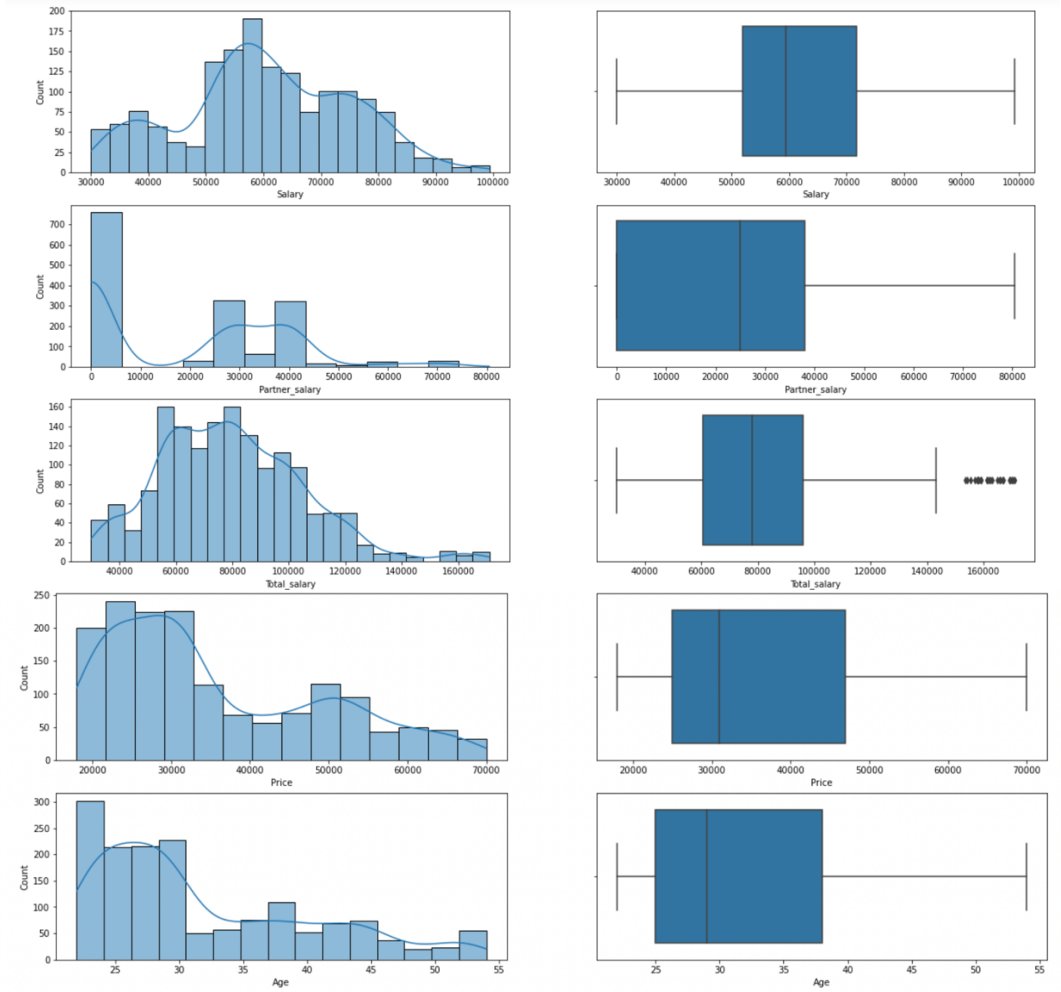
- Considering all categorical values:



The results are as follow:

- The data shows that there are more male customers than female.
- Maximum number of customers are married while less than 200 are single
- There are more post graduates than graduates.
- Salaried buyers are more than business professionals
- More than 500 have 2 and 3 dependents, while around a little over 200 have 1 or 4 dependents while around 20 people have no dependents
- The no of customers with and without personal loan are almost equal , just 3 more people have personal loan
- There is a huge variation in house loan as around 500 have house loan while more than 1000 do have a house loan.
- More than 800 customers have working partner while around 700 don't have working partner.
- The Sedan is most purchased with a total of 702 sales followed by hatchback with 582 and then suv in the least with just 297 sales

- Considering all numerical values:



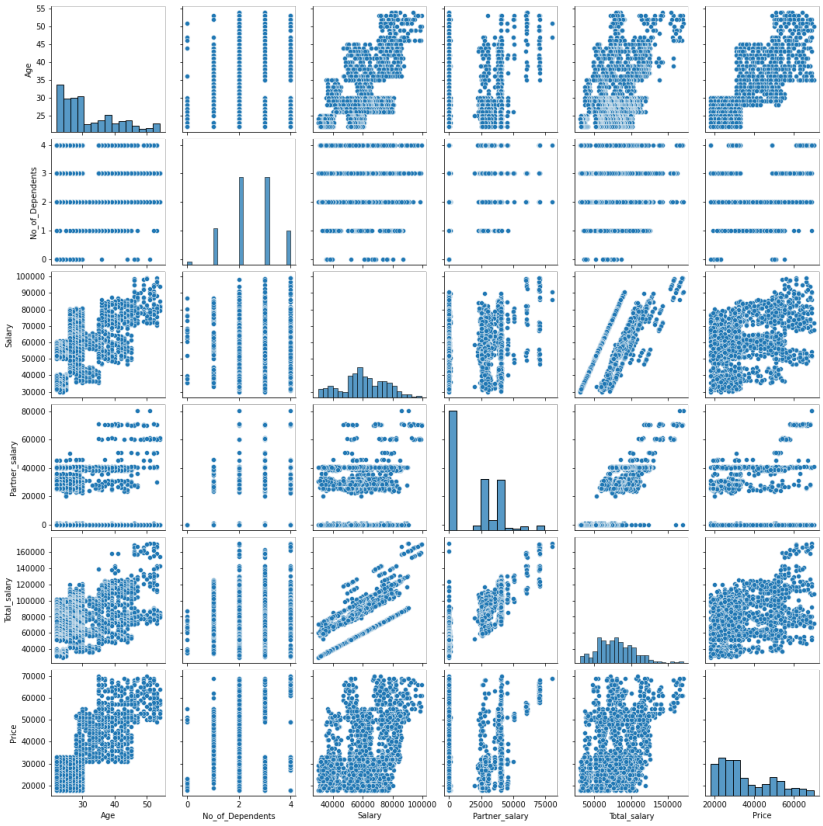
	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1581.0	18869.512966	19570.644035	0.0	0.0	24900.0	38000.0	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

- The results are as follow:
- In age the average age is 31 and maximum buyers are of the age 22 to 30. The boxplot shows a right skewed graph
- Maximum cars are priced between 20000 to 30000. The graph is right skewed with a mean of 35597 rs.
- In partner salary majority have 0 to less that 10000 and hence the graph is left skewed
- Salary has an almost symmetrical distribution that is shows a very little right skewed graph
- Majority of totale salary is between 60000 and 100000 and it also consists of a lot of outliners at 160000

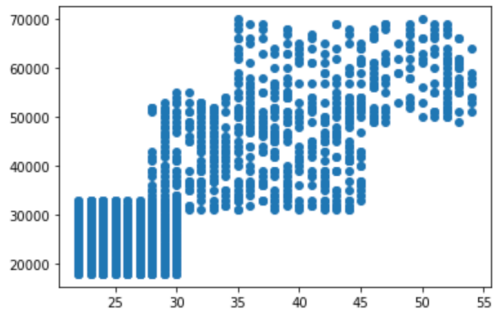
Bivariate Analysis:-

- First will check the correlation between all the numerical values

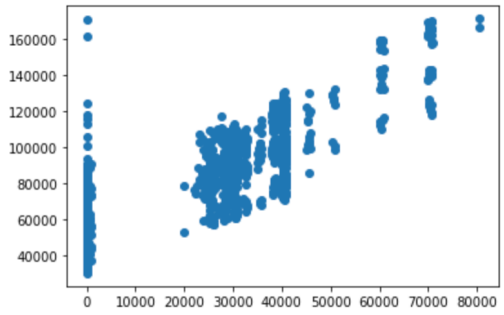
- As seen in the Heat map there is a good correlation between price and age and partner_salary and total_salary



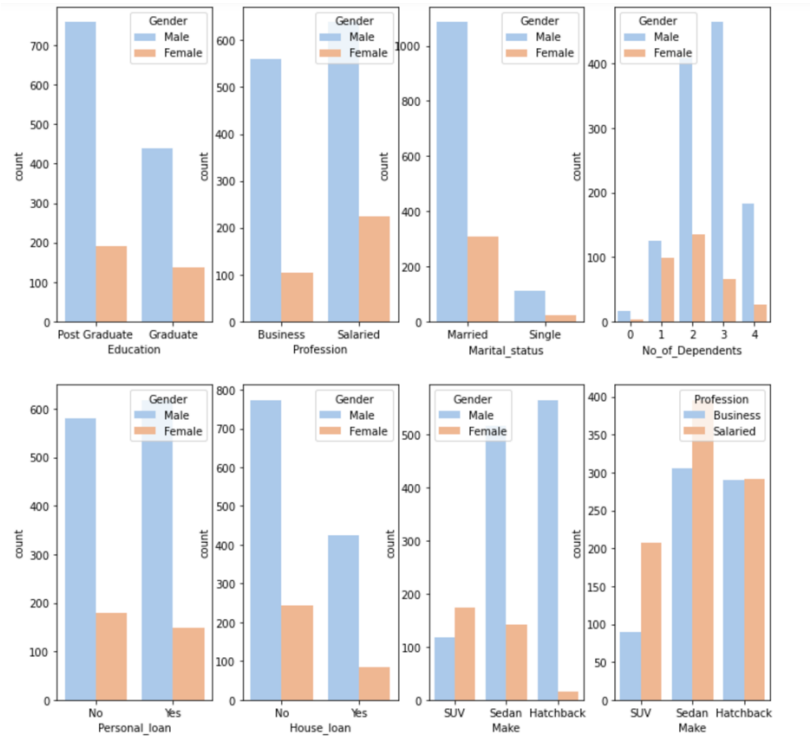
```
In [35]: plt.scatter(df['Age'],df['Price']);
```




```
In [36]: plt.scatter(df['Partner_salary'],df['Total_salary']);
```



- Taking a closer look at price vs age scatterplot shows that there is appositive correlation between the two. That is the price increases with age
- In partner_salary vs total_salary it does show a positive correlation but with some exception cases(outliners)



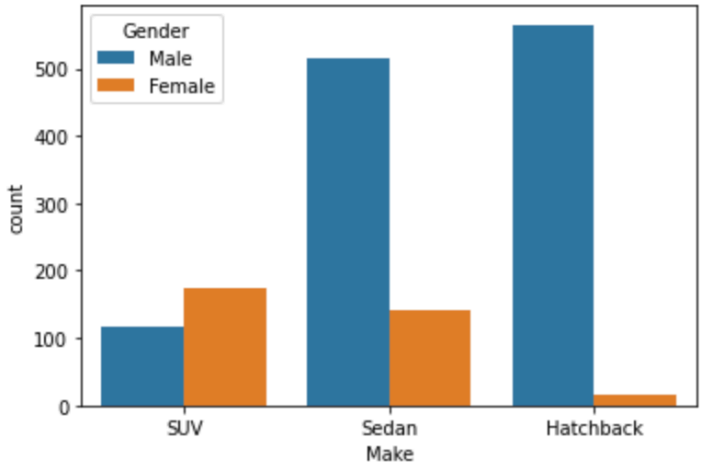
- There are more post graduate male.
- In the customers the highest are married male and lowest of single female
- Salaried men are more buyers than business professionals
- Men have more dependents than female.
- The highest are male with 2 or 3 dependents while in female it is 1 or 2 dependents.
- While in male more have personal loans while more female do not have personal loans
- Maximum men buy hatchback followed by sedan and then SUV , but in case of female buyers it is opposite. That is more prefer to buy SUV followed by sedan and then hatchback
- professionally salaried prefer sedan as first followed by hatchback and then SUV. Although it is same with business professionals but the difference between sedan and hatchback is very less

```
Profession  Gender  Marital_status  Make
Salaried    Male    Married         Sedan        266
                                    Hatchback    245
Business    Male    Married         Hatchback    239
                                    Sedan        227
Salaried    Female  Married         SUV          113
                                    Sedan         84
            Male    Married         SUV           79
Business    Female  Married         SUV           53
            Male    Single          Hatchback     50
            Female  Married         Sedan         43
            Male    Married         SUV           32
Salaried    Male    Single          Hatchback     31
            Female  Married         Hatchback     14
            Male    Single          Sedan         13
Business    Male    Single          Sedan         10
Salaried    Female  Single          Sedan          7
Business    Female  Single          Sedan          7
Salaried    Male    Single          SUV            6
            Female  Single          SUV            5
Business    Female  Single          SUV            2
            Male    Single          SUV            1
Salaried    Female  Single          Hatchback      1
dtype: int64
```

- The highest buyers are salaried married male buying sedan followed by hatchback.
- The lowest is salaried single female buying hatchback.
- While married male have highest purchase of sedan and hatchback, married female have highest purchase of suv

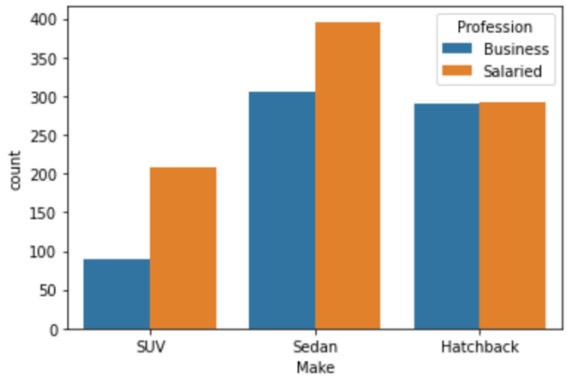
Key Questions:-

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



As shown in the graph, men do not tend to prefer SUVs when compared to women. They more over prefer hatchback or sedan

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



Profession	Business	Salaried	All
Make			
Hatchback	18.342821	18.469323	36.812144
SUV	5.629349	13.156230	18.785579
Sedan	19.354839	25.047438	44.402277
All	43.327008	56.672992	100.000000

The bar graph clearly shows that 25% of salaried prefer sedan which is the highest.

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  Gender  Make
Business    Male    Hatchback    18.913613
Salaried    Male    Sedan        18.259162
                    Hatchback    18.062827
Business    Male    Sedan        15.510471
Salaried    Female  SUV           7.722513
                    Sedan         5.955497
            Male    SUV           5.562827
Business    Female  SUV           3.599476
                    Sedan         3.272251
            Male    SUV           2.159686
Salaried    Female  Hatchback     0.981675
dtype: float64
```

It is very clear that salaried men prefer sedan and hatchback over SUVs.
As shown 18.25% prefer sedan and 18.06% prefer hatch back while only 5.5% prefer SUVs

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

```
Make       Gender
Hatchback  Female      412000
           Male      14959000
SUV        Female     9252000
           Male       7031000
Sedan      Female     6031000
           Male      17358000
Name: Price, dtype: int64
```

The highest amount paid is by men to buy sedans followed by hatchback, while the lowest is paid by women to buy hatchback. Women have paid highest to buy SUV

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

```
df.groupby(['Personal_loan'])['Price'].sum()
```

```
Personal_loan
No     28990000
Yes    27290000
Name: Price, dtype: int64
```

```
df.groupby(['Personal_loan','Make'])['Price'].sum()
```

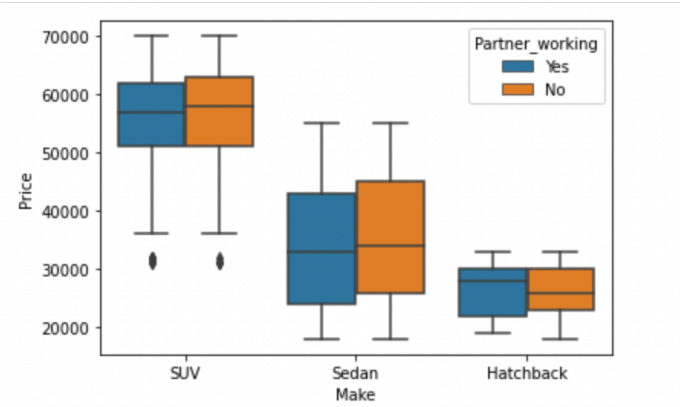
```
Personal_loan  Make
No             Hatchback     7765000
               SUV          10373000
               Sedan        10852000
Yes            Hatchback     7643000
               SUV           6207000
               Sedan        13440000
Name: Price, dtype: int64
```

Maximum spent buy customers with no personal loan.
To be more specific maximum spent on sedan followed by SUV
Customers having personal_loan spent highest on sedan followed by hatchback

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

		count	mean	std	min	25%	50%	75%	max
Partner_working	**Make**								
No	**Hatchback**	281.0	26323.843416	4147.947104	18000.0	23000.0	26000.0	30000.0	33000.0
	SUV	144.0	56173.611111	8873.099017	31000.0	51000.0	58000.0	63000.0	70000.0
	Sedan	288.0	35354.166667	11016.597393	18000.0	26000.0	34000.0	45000.0	55000.0
Yes	**Hatchback**	301.0	26614.617940	4421.649647	19000.0	22000.0	28000.0	30000.0	33000.0
	SUV	153.0	55496.732026	9549.127945	31000.0	51000.0	57000.0	62000.0	70000.0
	Sedan	414.0	34082.125604	11229.200059	18000.0	24000.0	33000.0	43000.0	55000.0

Has seen in the table, people with working partners helps in increase of cars being bought.



In hatchback the median of price with working partners is more while in sedan and SUV it is slightly less but the number of people that buy re significantly more
The table shows that working partners does influence buying higher priced cars .as the percentage of cars bought are more with working partners

Partner_working	No	Yes	All
Price			
18000	1.012018	1.328273	2.340291
19000	1.075269	1.771031	2.846300
20000	2.024035	2.972802	4.996837
21000	1.012018	1.454775	2.466793
22000	1.518027	1.960784	3.478811
23000	2.403542	2.277040	4.680582
24000	1.834282	2.213789	4.048071
25000	1.707780	1.265022	2.972802
26000	1.391524	2.024035	3.415560
27000	1.138520	2.087287	3.225806
28000	1.897533	2.340291	4.237824
29000	1.391524	1.897533	3.289058
30000	1.834282	2.024035	3.858318
31000	2.340291	2.530044	4.870335
32000	2.530044	3.036053	5.566097
33000	1.454775	2.024035	3.478811
34000	0.632511	0.569260	1.201771
35000	0.632511	0.506009	1.138520
36000	0.695762	0.695762	1.391524
37000	0.379507	0.822264	1.201771
38000	0.253004	0.506009	0.759013
39000	0.569260	1.075269	1.644529
47000	0.695762	0.506009	1.201771
48000	0.506009	0.822264	1.328273
49000	0.885515	0.759013	1.644529
50000	0.948767	1.265022	2.213789
51000	1.138520	1.012018	2.150538
52000	0.695762	1.012018	1.707780
53000	0.759013	0.885515	1.644529
54000	0.632511	0.759013	1.391524
55000	0.379507	0.885515	1.265022
56000	0.126502	0.316256	0.442758
57000	0.695762	0.695762	1.391524
58000	0.569260	0.316256	0.885515
59000	0.442758	0.253004	0.695762
60000	0.316256	0.253004	0.569260
61000	0.442758	0.695762	1.138520
62000	0.379507	0.379507	0.759013
63000	0.442758	0.316256	0.759013
64000	0.506009	0.316256	0.822264
65000	0.316256	0.189753	0.506009
66000	0.316256	0.442758	0.759013
67000	0.253004	0.253004	0.506009
68000	0.316256	0.379507	0.695762
69000	0.253004	0.442758	0.695762
70000	0.063251	0.063251	0.126502
All	45.098039	54.901961	100.000000

Actionable Insights & Recommendations:-

1. The data shows that married men prefer sedans and hatchbacks has they are more cheaper than SUV. But SUV can be made available for them as it is more spacious for family by introducing discounts and offers
2. Single customers prefer hatchback, hence marketing hatchback should focus on single men
3. Women prefer SUVs and sedans over hatchback and hence marketing the two should be focused on women
4. Professionally salaried buy more cars and the overall ads should focus on business individuals needing to own a car
5. If customer has working partner tend to buy costlier car and putting light on this matter can also increase in sales not just in terms of number of cars sold but in terms of revenue generated.

Problem 2 :–

Context

A bank generates revenue through interest, transaction fees, and financial advice, with interest charged on customer loans being a significant source of profits. GODIGT Bank, a mid-sized private bank, offers various banking products and cross-sells asset products to existing customers through different communication methods. However, the bank is facing high credit card attrition, leading them to re-evaluate their credit card policy to ensure customers receive the right card for higher spending and intent, resulting in profitable relationships.

Objective

As a Data Scientist at the company and the Data Science team has shared some data. You are supposed to find the key variables that have a vital impact on the analysis which will help the company to improve the business.

Framing Analytics Problem:-

The dataset has the following:-

```
<class 'pandas.core.frame.DataFrame'>
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
```

The dataset has 8 objects ,1 date and time, and 19 int datatype

	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

The data shows the max and min in all the numerical fields

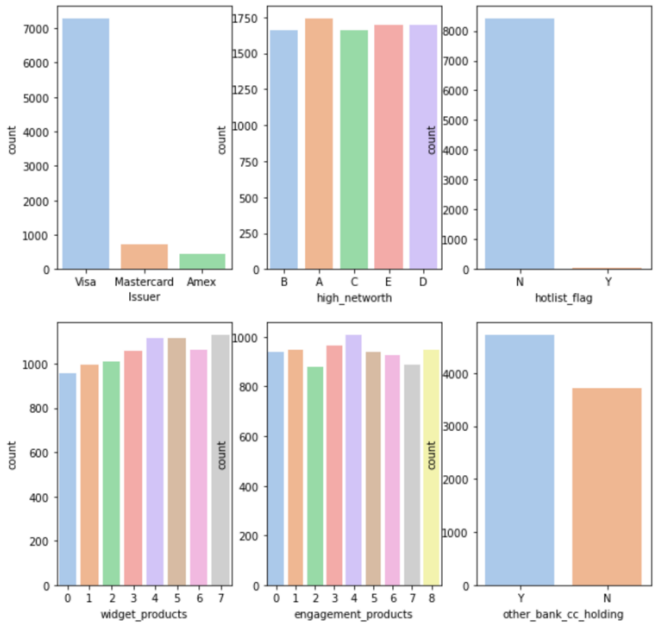
After checking the unique values of all categorical values Occupation_at_source had a discrepancy where data was 0 and hence I have replaced it with Null

```
In [145]: df1['card_no'].unique()
Out[145]: array(['4384 39XX XXXX XXXX', '4377 48XX XXXX XXXX',
                 '4258 06XX XXXX XXXX', '5241 78XX XXXX XXXX',
                 '4055 33XX XXXX XXXX', '4375 51XX XXXX XXXX',
                 '4386 28XX XXXX XXXX', '4262 41XX XXXX XXXX', '37694 5XXXX XXXXX',
                 '4477 47XX XXXX XXXX', '37691 6XXXX XXXXX'], dtype=object)

In [146]: df1['Issuer'].unique()
Out[146]: array(['Visa', 'Mastercard', 'Amex'], dtype=object)

In [147]: df1['card_type'].unique()
Out[147]: array(['edge', 'prosperity', 'rewards', 'indianoil', 'cashback',
                 'shoprite', 'chartered', 'aura', 'gold', 'smartearn', 'prime',
                 'pulse', 'platinum', 'centurion', 'elite'], dtype=object)

In [148]: df1['high_networth'].unique()
Out[148]: array(['B', 'A', 'C', 'E', 'D'], dtype=object)

In [149]: df1['hotlist_flag'].unique()
Out[149]: array(['N', 'Y'], dtype=object)

In [150]: df1['other_bank_cc_holding'].unique()
Out[150]: array(['Y', 'N'], dtype=object)

In [151]: df1['Transactor_revolver'].unique()
Out[151]: array(['T', 'R', nan], dtype=object)

In [159]: df1['Occupation_at_source'].unique()
Out[159]: array(['Self Employed', None, 'Student', 'Salaried', 'Retired',
                 'Housewife'], dtype=object)
```

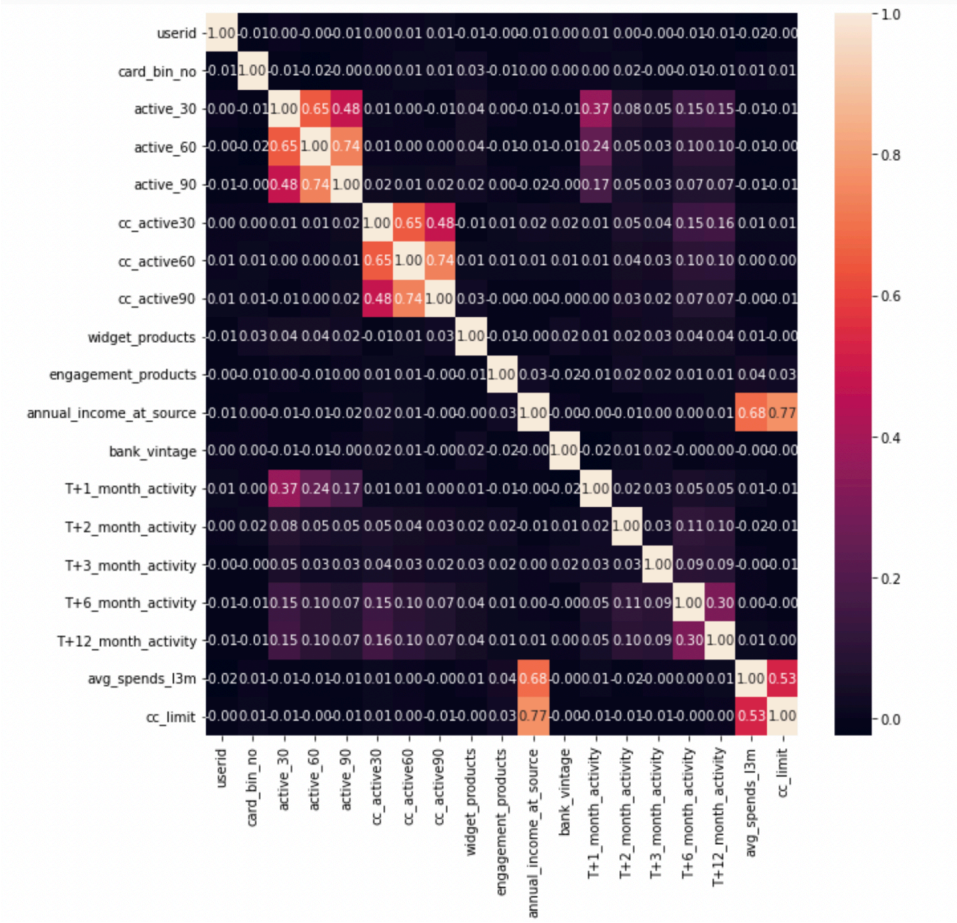
```
userid                        0
card_no                       0
card_bin_no                   0
Issuer                        0
card_type                     0
card_source_date              0
high_networth                 0
active_30                     0
active_60                     0
active_90                     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        261
cc_limit                      0
dtype: int64
```

Transactor_revolver and Occupation_at_source have null values as shown above

Analysis are as follow:

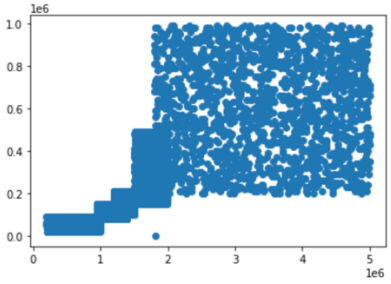


Maximum users have Visa Issuer and Not Hotlisted
Maximum no of customers and having A as high_networth followed by E,D,B, and C

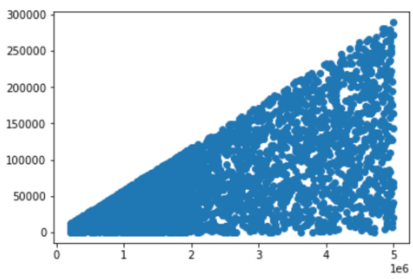


The Heatmap shows the realtion between all the numericl values.
As seen there is correlation between annual_income_at_source and cc_limit.
And also a small correlation between annual_income_at_source and avg_spends_l3m
The rest are not correlated

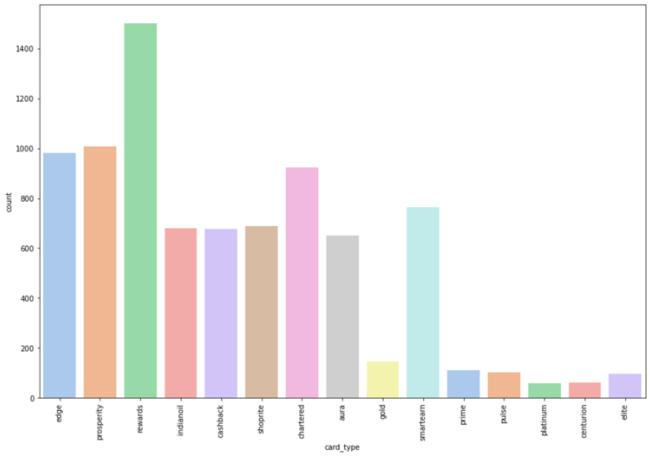
```
In [227]: plt.scatter(df1['annual_income_at_source'],df1['cc_limit']);
```

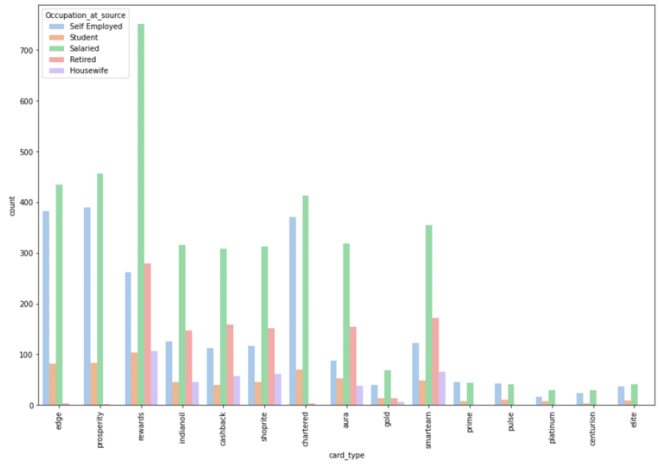


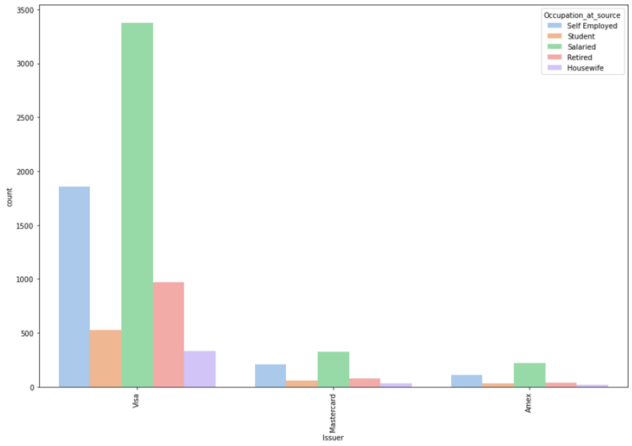

```
In [228]: plt.scatter(df1['annual_income_at_source'],df1['avg_spends_l3m']);
```

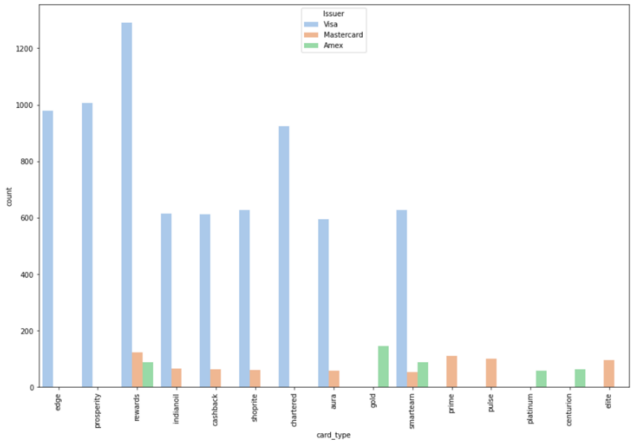


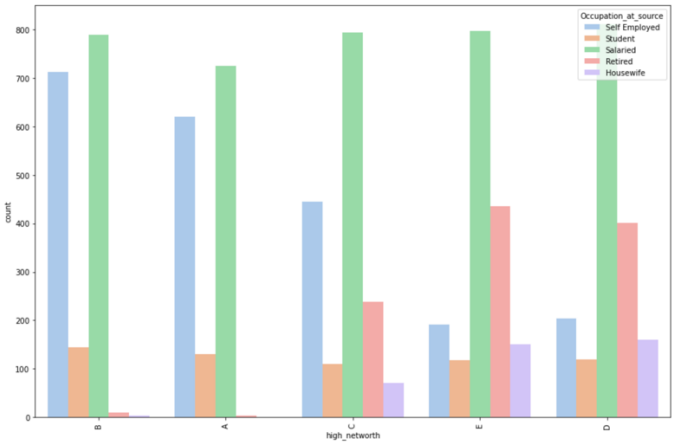
Both the graph are positively correlated


count
card_type
edge
prosperity
rewards
indianoil
cashback
shoprite
chartered
aura
gold
smartearn
prime
pulse
platinum
centurion
elite


Occupation_at_source
Self Employed
Student
Salaried
Retired
Housewife
count
card_type

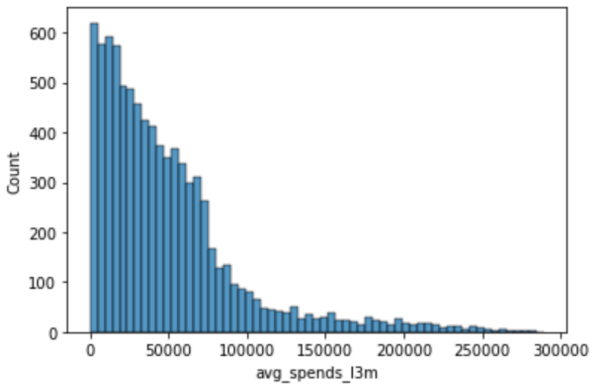

Occupation_at_source
Self Employed
Student
Salaried
Retired
Housewife
count
Visa
MasterCard
Amex
Issuer





The above graphs show the relation between the other categorical fields:

1. The most issued card type is rewards and the least is centurion
2. Salaried people have the highest no of credit cards issued
3. Housewife have the lest no of credit cards issued
4. Visa issues the highest no of cards to salaried people
5. Amex is the least issued
6. Visa gives particular card type similar to amex and mastercard
7. Highest networth is of salaried employees with all almost equal
8. Least in the high networth are the housewifes



The mean expenditure is `49527.36553030303` `the average expenditure is between 0 and 750000`

Actionable Insights & Recommendations:-

The questions to answer are:

1. Which is the highest used card type?
2. Which are the highest issuers?
3. Which occupation uses the highest card?
4. Average expenditures?
5. Which card type is issued by which issuer?