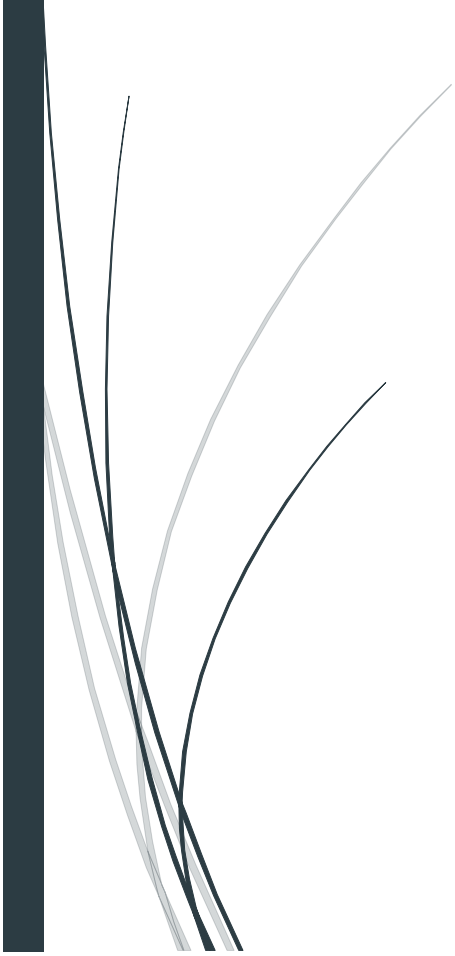
Business Report

Statistical Methods for Decision Making



Ayush Sharma

Table of Contents

Problem 1

Problem 2

Problem 1

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)

Ans: As observed from the info above, the dataset consists of **1581 rows and 15 columns** respectively. There is a total of **6 numeric variables** while **8 categorical variables**.

```
No of Rows:1581
No of Columns:15
```

```
<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
```

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.

Ans: The dataset requires certain pre-processing in order to treat the null or discrepant values along with treating the outliers.

- DISCREPANT and NULL VALUES exist in the columns **Gender and Partner_salary.** These values need to be treated by imputation:
    - The **mode of observations** has been used for imputation in the Gender column
    - The Partner Salary can be calculated by **subtracting the salary from total salary** column
- OUTLIERS have been observed in the **Total_salary** column. These outliers can be imputed by using the IQR method:
    - Allotting the upper limit to the values more than Q3 +1.5IQR
    - Allotting the lower limit to the values less than Q1 - 1.5IQR

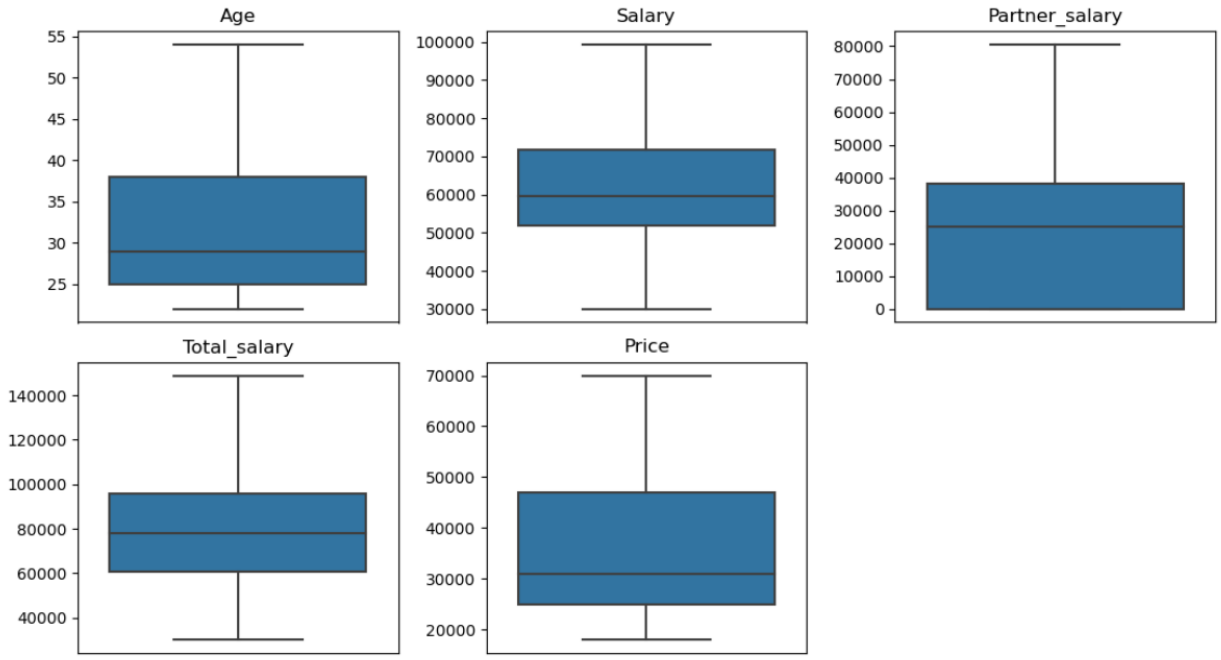
The null, discrepant and outliers have now been treated and the dataset can now be used to perform EDA.

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

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

Ans: A univariate analysis can be done on the dataset to derive initial insights. The analysis has been done separately for the continuous and categorical variables. Following are the visualizations along with the respective insights:

- **Continuous variables** - Boxplots have been used for the visualization of the continuous variables.



  - **Age**
    - The age of individuals ranges from **22 to 55 years**
    - A major proportion of individuals lie between the age of **25 to 38 years** which can be considered as the **key target audience** for the automobile company
    - The median age is of **29 years** while the average age of a customer is approximately **32 years**

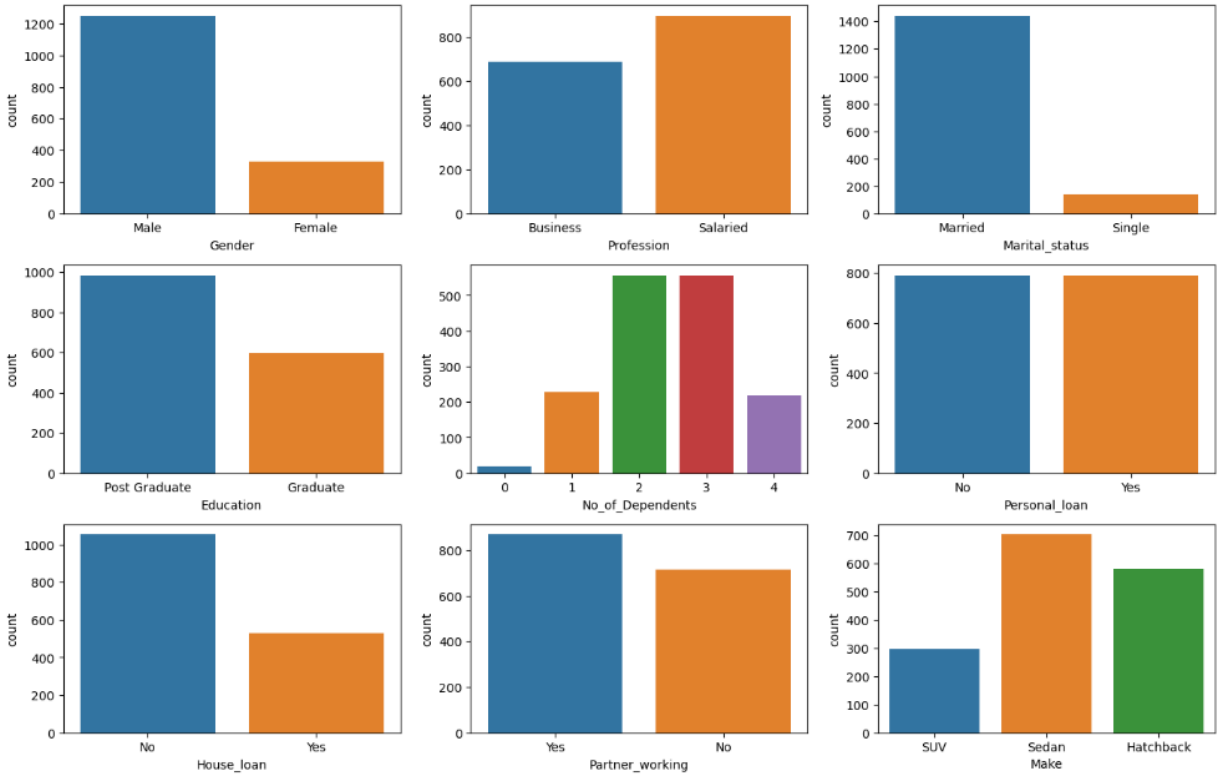
  - **Salary**
    - The salary of individuals ranges from **$30,000 to $1,00,000**
    - A major proportion of individuals have their salaries between the range of **$51,000 to $71,000**
    - The median and average salary is of approximately **$60,000** for the customers

  - **Partner Salary**
    - The partner salaries range from **$0 to $81,000**
    - A major proportion of partner salaries lie in the range of **$0 to $38,000**
    - The median partner salary is of **$25,000** while the average partner salary is of approximately **$19,000**

- **Total Salary**
  - The total salary ranges from **$2000 to $1,45,000**
  - A major proportion of individuals have their total salaries between the range of **$60,500 to $96,000**
  - The median and average total salary is close to **$79,000** for the customers

- **Price**
  - The price of vehicles ranges from **$20,000 to $70,000**
  - A majority of people incline towards the purchase of vehicles in the price range of **$25,000 to $47,000.**
  - The median price is of **$31,000** while the mean price lies close to **$36,000**
  - These observations are very vital as they provide the automobile company with an estimate of the ideal price range they can allocate their vehicles as per the price range of their key target audience

- **Categorical Variables** - Count plots have been used for the visualization of the categorical variables.



  - **Gender**
    - The **men dominate the women** by a large margin
    - There are approximately **1250 men** and **300 women**
    - Hence the male population is the majority buyer constituting approximately **80%** of the total population

  - **Profession**
    - **Salaried employees** are more in numbers with respect to the customers owning a business
    - Approximately **900 customers are salaried** by profession and close to **700 are business owners**

  - **Marital Status**
    - A major proportion of the customers are **married**
    - Approximately **1400 customers are married** while close to **130 customers are single**
    - This indicates that **married individuals** can be categorized amongst the **key target audience** for the automobile company

  - **Education**
    - Approximately **900 customers are post-graduates** and **600 people are under-graduates**
  - **No of Dependents**
    - It can be observed that a major proportion of the customers accounting for approximately **550 customers each** has either **2 or 3 number of dependents**

- The count of the customers having 1 or 4 number of dependents stands at approximately **220 each**
- The customers with no dependents constitute a very small proportion of the population at a count of **20 customers**
- This indicates that **more than 95%** of the customers have a dependent such as their children or parents

- **Personal loan**
  - Approximately **half of the customers are opting for a personal loan** while the **other half is not opting for a personal loan**
  - The impact of this categorical variable on the dataset will be determined by conducting bi-variate and multi-variate analysis on the data

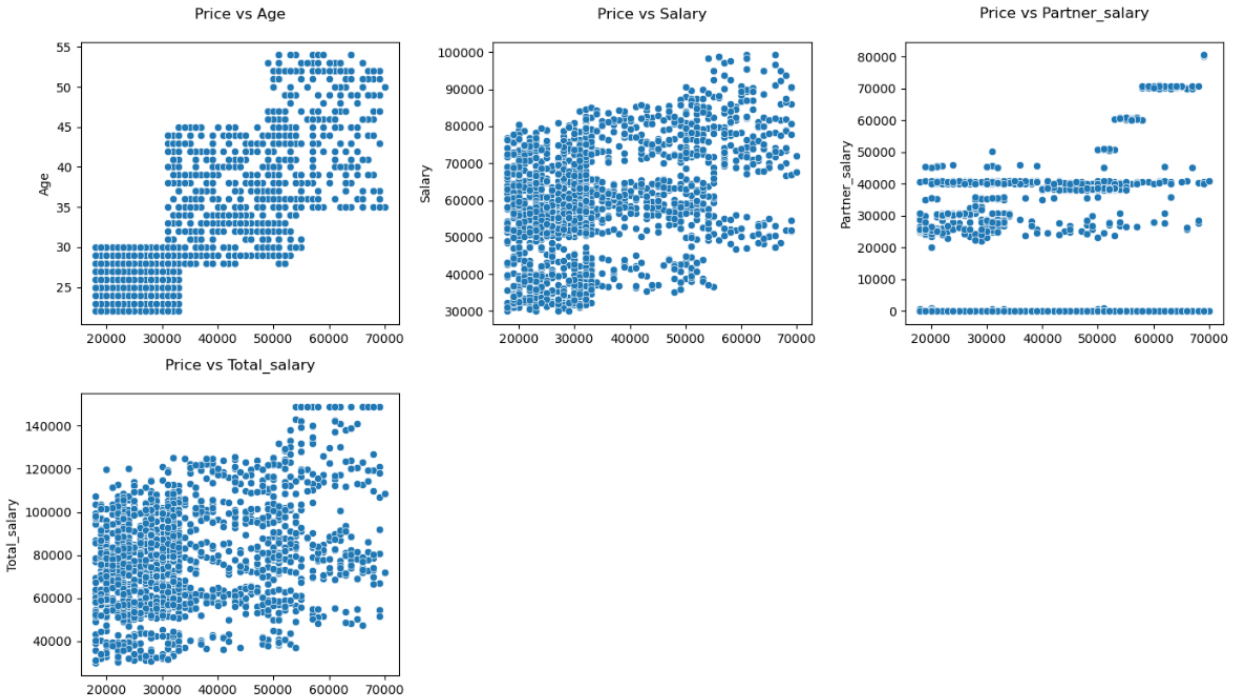
- **House Loan**
  - The vehicles purchased by the customers without a house loan is approximately 50% higher than the customers with a house loan indicating that **customers with no house loan are more likely to purchase a vehicle than the customers with a house loan**

- **Partner Working**
  - The number of **customers with a working partner stands at approximately 850** while the number of **customers without a working partner stands at 700**

- **Make**
  - **Sedans** dominate the rest of its counterparts having a count of approximately **700 customers**
  - **Hatchbacks** and **SUVs** are respectively **second and third in demand** with a count of approximately **580 and 300 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.

Ans: The univariate analysis as per the data visualizations has already been performed above providing certain useful insights about the dataset. We can now proceed by conducting bi-variate and multi-variate analysis in order to deduce different insights by observing the different relationships of the variables with each other.



Different scatter plots have been created between the Price column and the other numeric columns in order to understand the respective trends of these columns with each other. The following insights can be drawn on the basis of the above visualizations:

- **Age and Price**
  - The age of the customers purchasing a car between the price range of **$18,000 to $33,000** lies between **22 to 30 years**
  - The age of the majority of the customers purchasing a car between the price range of **$33,000 to $55,000** lies between **28 to 45 years**
  - The age of the customers purchasing a car between the price range of **$55,000 to $70,000** lies between **35 to 53 years**
  - The plot density gradually decreases as the price and age increases indicating that **younger people are likely to purchase a cheaper car** while **older people are more likely to purchase an expensive car**
  - It can also be seen that the number of cars purchased decreases with age indicating that **younger people purchase a greater number of cars as compared to the older people**

- **Salary and Price**
  - The salary of a considerable proportion of the buyers purchasing a car between the price range of **$18,000 to $33,000** lies between **$30,000 to $80,000**
  - The proportion of the buyers purchasing a car between the price range of **$33,000 to $55,000 is relatively lesser** and lies between **$38,000 to $83,000**
  - The proportion of the buyers purchasing a car between the price range of **$55,000 to $70,000 is even lesser** and lies between **$48,000 to $98,000**

- It can thus be inferred that **despite their higher salaries**, a major proportion of customers are **purchasing a car within the price range of $18,000 to $55,000**

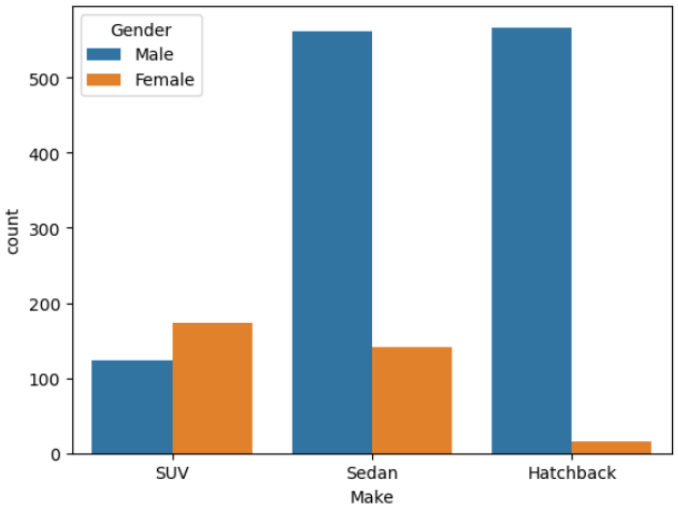
- **Partner Salary and Price**
  - It can be observed that the major proportion of the partner salaries range from **$20,000 to $45,000**
  - People with partner salaries between **$2,000 to $3,000** are more likely to purchase a car within the price range of **$18,000 to $35,000**
  - People with partner salaries between **$3,500 to $4,000** purchase relatively more cars within the price range of **$40,000 to $55,000**

- **Total Salary and Price**
  - The trends in the total salary are similar to the salary and price visualization indicating that a majority of the **customers are purchasing a car within the price range of $18,000 to $55,000**

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.

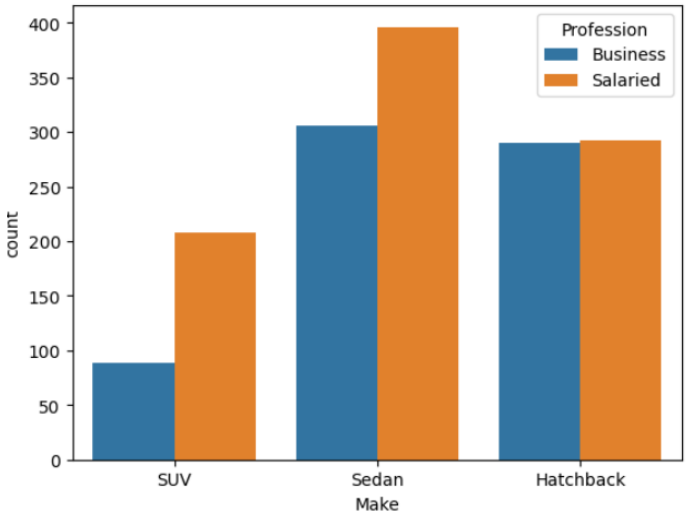
E1) Steve Roger says "Men prefer SUV by a large margin, compared to the women"

Ans: As per the below visualization between the make and the gender columns, it can be inferred that the above observation is incorrect and **women prefer SUV by a small margin when compared to men**.



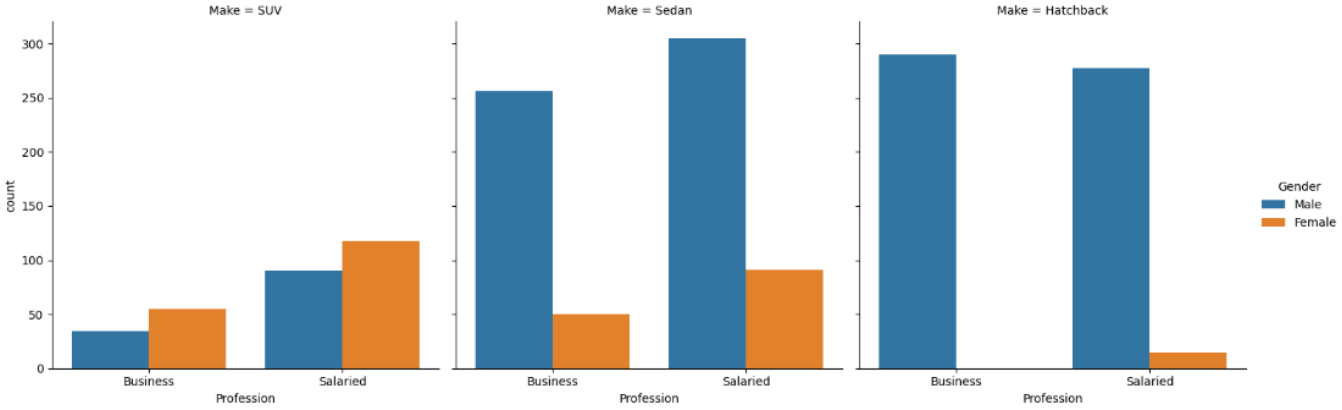
E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.

Ans: As per the below visualization between the make and the profession columns, it can be inferred that the above observation is correct and **a salaried person is indeed more likely to buy a Sedan**.



E3) Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for a SUV sale over a Sedan Sale.

Ans: As per the below visualization between the make, profession and gender columns, it can be inferred that the above observation is incorrect and on the contrary, **salaried men purchase the greatest number of Sedans when compared to the rest of the models.**



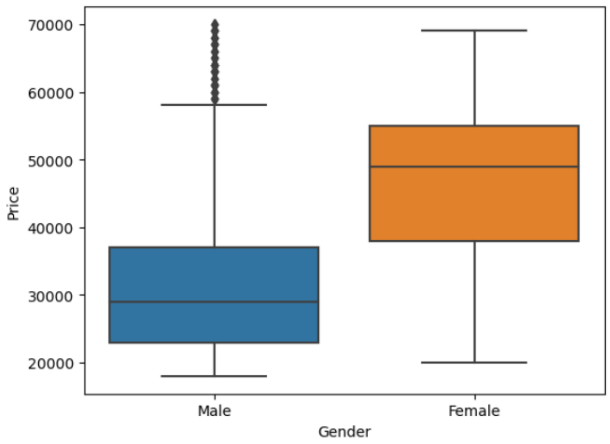
F. From the given data, comment on the amount spent on purchasing automobiles across the following categories. Comment on how a Business can utilize the results from this exercise. Give justification along with presenting metrics/charts used for arriving at the conclusions.

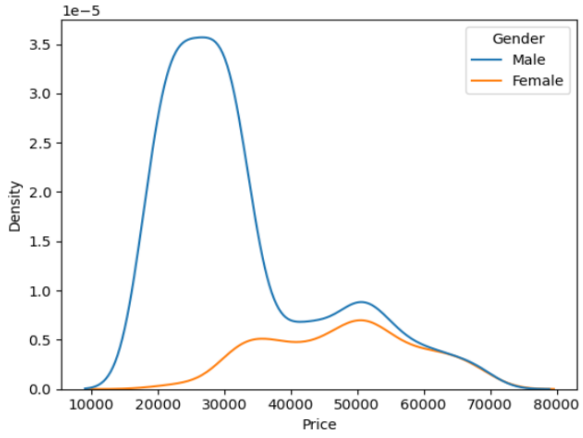
F1) Gender

F2) Personal_loan

Ans:

F1) **Gender**





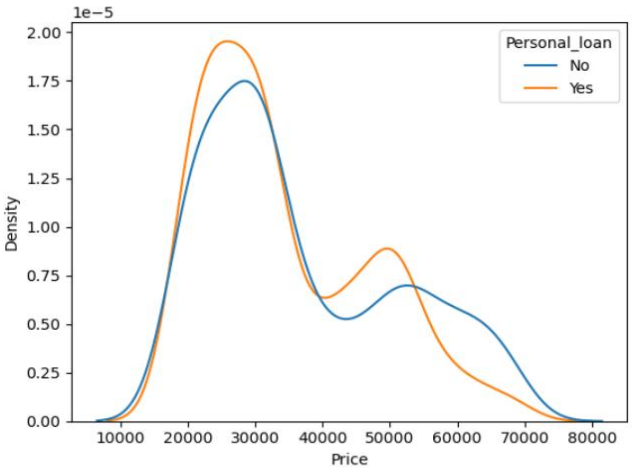

```
Gender
Female    15695000
Male      40585000
Name: Price, dtype: int64
```

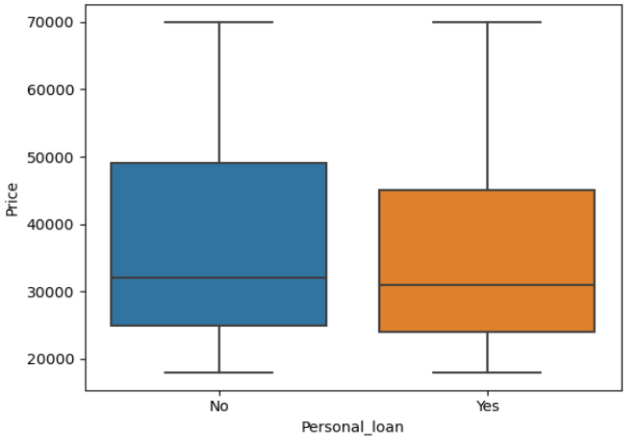
```
Female revenue percentage: 27.88734896943852%
Male revenue percentage: 72.11265103056148%
```

The following insights can be drawn from the above visualizations:

- The total revenue generated by men is of **$40,585,000/- (72.11%)** and the total revenue generated by women is of **$15,695,000/- (27.88%)**.
- The majority of the male population purchases a car in the price range of **$23,000 to $37,000** with the **median and mean prices at $32,000 and $29,000** respectively.
- The majority of the female population purchases a car in the price range of **$38,000 to $55,000** with the **median and mean prices at $49,000 and $48,000** respectively.
- It can be suggested to the automobile company that more **vehicles in the price range of $20,000 to $40,000 can be targeted towards the men**
- It can be suggested to the automobile company that more **vehicles in the price range of $40,000 to $55,000 can be targeted towards the women**

F2) **Personal_loan**






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

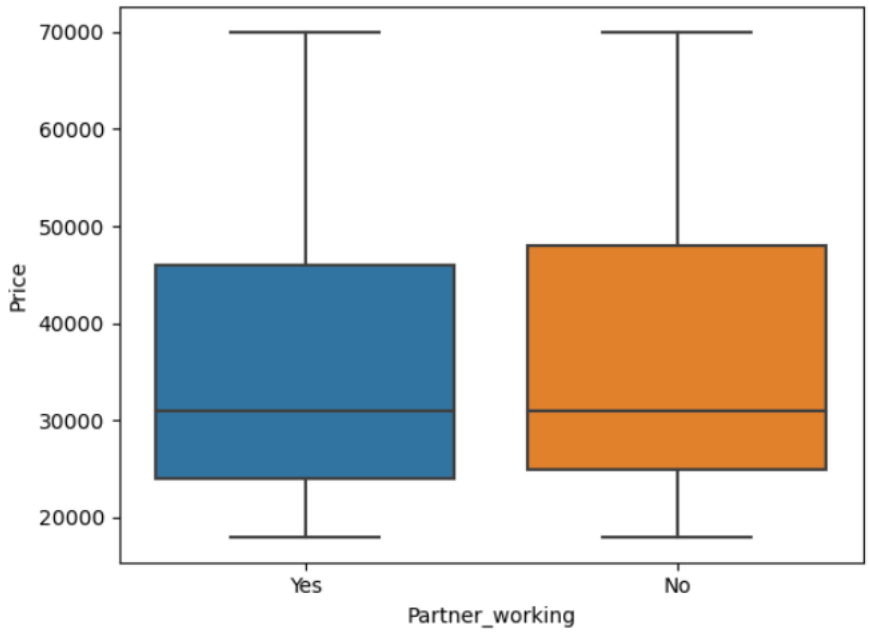
```
Personal loan percentage: 48.49%
No personal loan percentage: 51.51%
```

The following insights can be drawn from the above visualizations:

- The total revenue generated by customers having a personal loan is of **$27,290,000/- (48.48%)** and the total revenue generated by customers not having a personal loan is of **$28,990,000/- (51.51%)**.
- It can be observed that the **trends of customers whether or not they've applied for a personal loan are quite similar**.
- The majority of the customers who applied for a personal loan purchase a car in the price range of **$24,000 to $45,000** with the **median and mean prices at $31,000 and $34,000** respectively.
- The majority of the customers who did not apply for a personal loan purchase a car in the price range of **$25,000 to $49,000** with the **median and mean prices at $32,000 and $37,000** respectively.
- Hence it can be observed that the **affordability of the customers not having a personal loan is slightly more between the range of $45,000 to $49,000** when compared to the customers having a personal loan
- It can thus be suggested to the automobile company that **customers with no personal loan can be considered as potential buyers for vehicles in the price range of $45,000 to $50,000**

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

Ans:



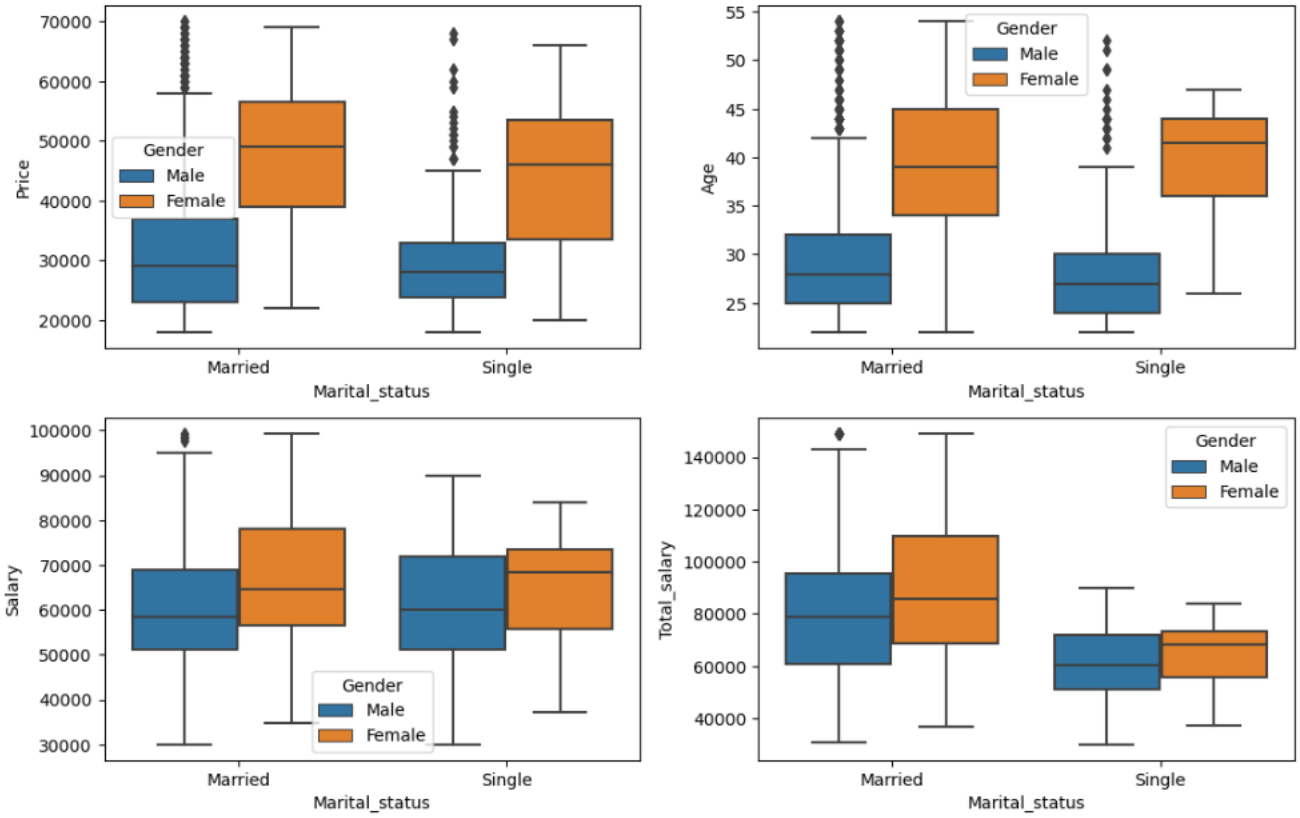
Partner_working	count	mean	std	min	25%	50%	75%	max
No	713.0	36000.000000	13817.734086	18000.0	25000.0	31000.0	48000.0	70000.0
Yes	868.0	35267.281106	13479.532555	18000.0	24000.0	31000.0	46000.0	70000.0

The following insights can be drawn as per the above visualization and statistics:

- The mean prices of the vehicles regardless of whether or not a customer has a working partner is quite similar at about **$35,000 and $36,000** respectively.
- The median prices of the vehicles regardless of whether or not a customer has a working partner is similar at **$31,000** respectively.
- It can thus be said that **customers having a working partner do not purchase a higher priced vehicle**

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.

Ans:

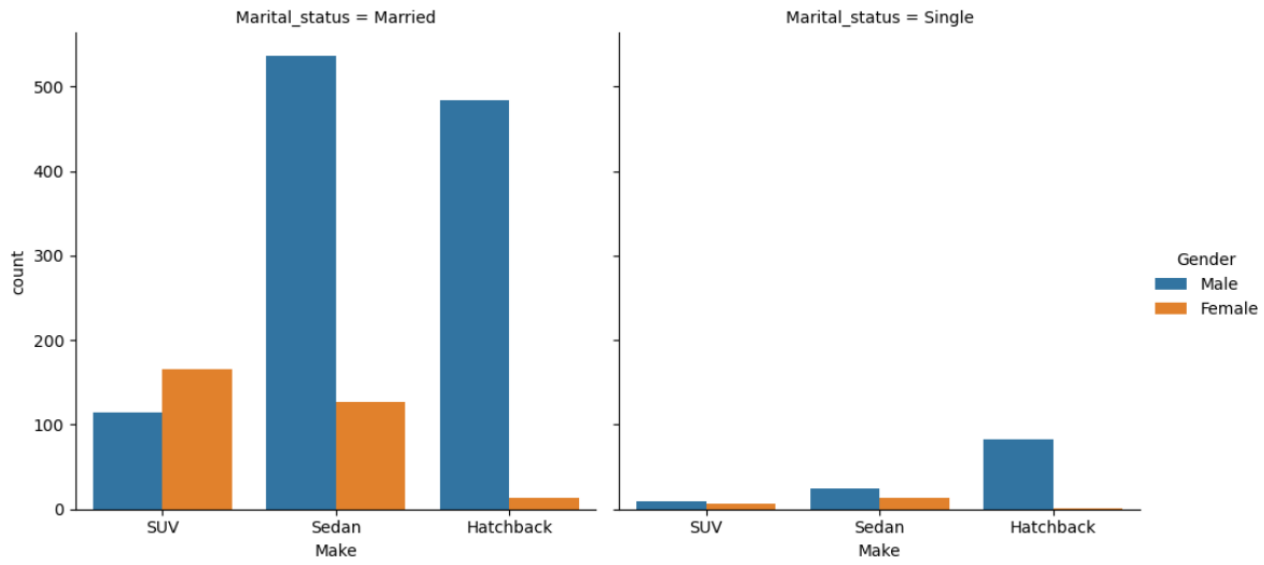


The following observations can be drawn from the above visualizations based on the **Gender and Marital Status** of the customers when compared to the other continuous variables:

- **Price**
  - Married men are more likely to purchase a vehicle in the **price range of $23,000 to $37,000** while married women are more likely to purchase a vehicle in the **price range of $39,000 to $56,000.**
  - Single men are more likely to purchase a vehicle in the **price range of $24,000 to $33,000** while single women are more likely to purchase a vehicle in the **price range of $33,000 to $53,000.**

- **Age**
  - Married men are more likely to purchase a vehicle in the **age of 25 to 32 years** while married women are more likely to purchase a vehicle in the **age of 34 to 45 years.**
  - Single men are more likely to purchase a vehicle in the **age of 24 to 30 years** while single women are more likely to purchase a vehicle in the **age of 36 to 44 years.**

- **Salary**
  - Married men have their **salaries in the range of $50,000 to $69,000** while married women have their **salaries in the range of $56,000 to $78,000**
  - Single men have their **salaries in the range of $51,000 to $72,000** while single women have their **salaries in the range of $56,000 to $73,000**



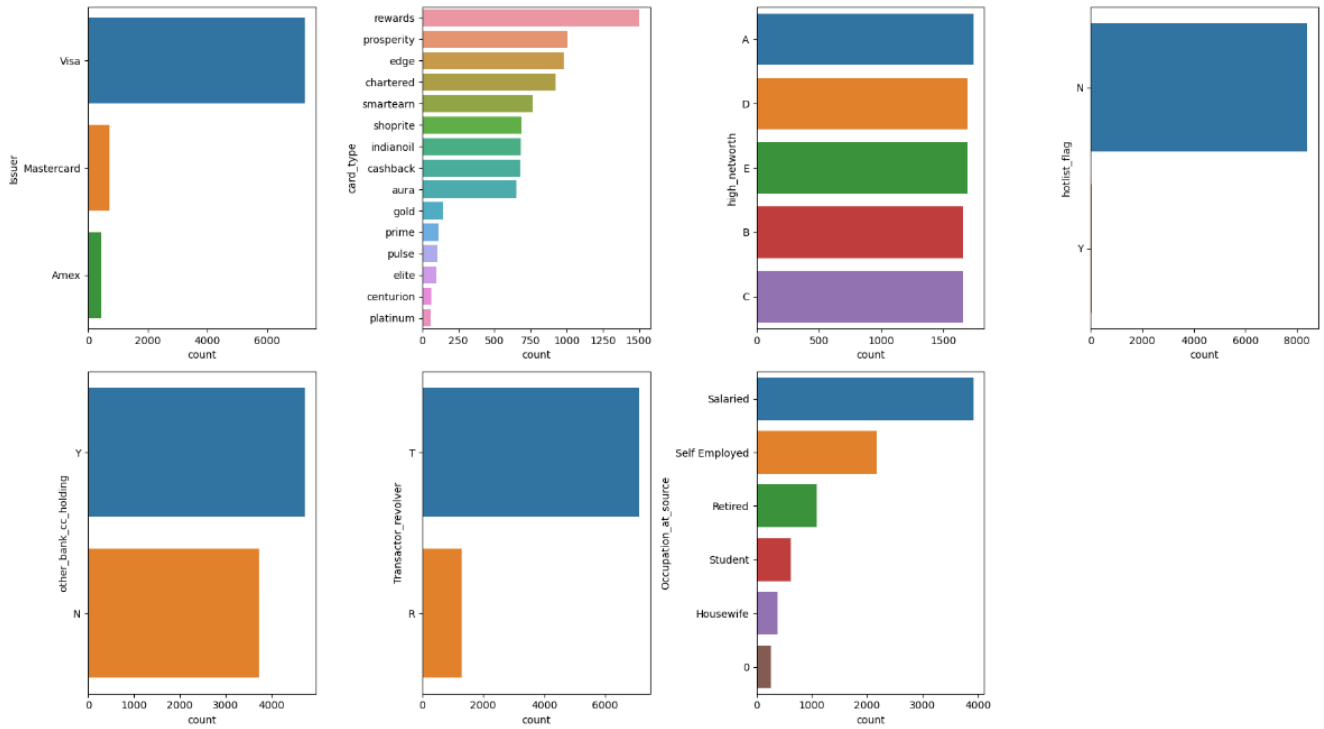
The following observations can be drawn from the above visualizations based on the **Gender and Marital Status** of the customers when compared to the **Make** of the vehicles:

- **Married men** have purchased **the greatest number of Sedans** followed by **Hatchbacks and SUVs**
- **Married women** purchased **SUVs the most** followed by **Sedans and Hatchbacks**
- **Single men** gave the most preference to **Hatchbacks** followed by **Sedans and SUVs**
- **Single women** purchased **Sedans the most** followed by **SUVs and Hatchbacks**

Problem 2

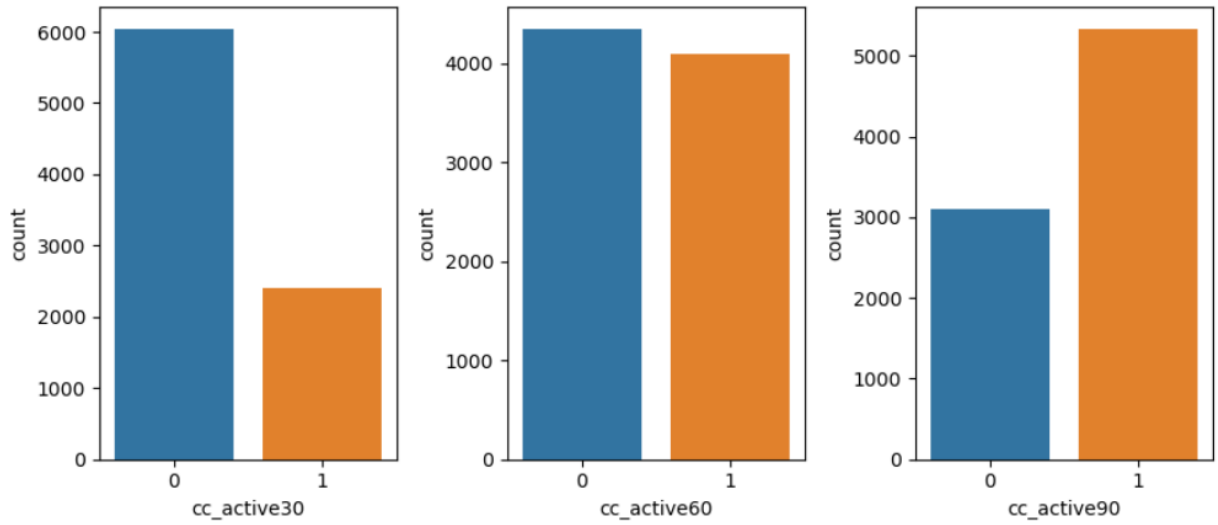
Analyse the dataset and list down the top 5 important variables, along with the business justifications.

Ans:



The following insights can be drawn from the above visualizations:

- **Visa Credit Cards** are the most widely used cards among all the other card categories
- **Salaried employees** have the greatest count in terms of owning a credit card
- Most of the customers **(85%) pay off their bill in full** while **the rest (15%) use the revolver credit facility**



The following insights can be drawn from the above visualizations regarding the credit card activity of the customers:

- It can be observed that the credit card activity was **low in the 1st month** with the usage restricted to **2400 people**
- The card activity was significantly **higher in the first 2 months** with **over 4000 active users**
- The card activity for a **consecutive 3 months** was higher indicating a gradual raise in the usage of the cards

It can thus be said that the following 5 variables are the most important in this dataset:

- Occupation
- Card Issuer
- Credit Card Activity
- Average Spends
- Annual Income at source