PYTHON FOR DATA SCIENCE - GRADED PROJECT

Contents:

Problem Statement:

Austo Motor Company is a leading car manufacturer specialising 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.

Introduction to Data Set:

The given data set has 1581 rows and 14 columns. Of 14 columns, six are numerical columns and 8 are categorical columns. Below is the snapshot of the data.

	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

Numerical columns: Age, No_of_Dependents, Salary, Partner_salary, Total_salary and Price.
Categorical columns: Gender, Profession, Marital_status, Education, Personal_loan, House_loan, Partner_working and Make.

Below are the summary statistics for numerical variables.

	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

First I checked if there were any duplicate rows and found out that there were no duplicates.

In the gender column, there were two misspelled words. After replacing them with the correct word. Then, I checked for null values in the dataset. I found null values in two columns, namely, Gender and Partner_salary.

Treating Missing Values:

Gender:
As gender is a categorical column I replaced nan values with the mode of that column. In our case, there are more Males than Females. So all the missing values are now filled with Male. Now, the Gender column is treated.

Partner_salary:
There were 106 missing values in Partner_salary.
Case1: Partner is working but partner salary column is empty
Here I have done a subtraction between Total salary and salary and got the partner salary value
Case2: Partner is not working but partner salary is present
Here for all the columns where the Partner is not working but the partner salary column has some value, I changed them to zero.

By doing this I made sure that there were no missing values and data was good to go for analysis.

Exploratory Data Analysis:

Univariate Analysis:
Univariate data visualization plot helps us comprehend the descriptive summary of the particular data variable. These plots help in understanding its distribution and dispersion.
We can check the distributions of observations by plotting Histograms and Boxplots.

Histogram:
A histogram takes as input a numeric variable only. The variable is cut into several bins, and the number of observations per bin is represented by the height of the bar.

Boxplot:
A boxplot gives a summary of one or several numeric variables. The line that divides the box into 2 parts represents the median of the data. The end of the box shows the upper and lower quartiles. The extreme lines show the highest and lowest values excluding outliers.

I have used histograms and boxplots to understand the distribution of data for numerical variables.

If we consider **age**, we can see that the age is right skewed.

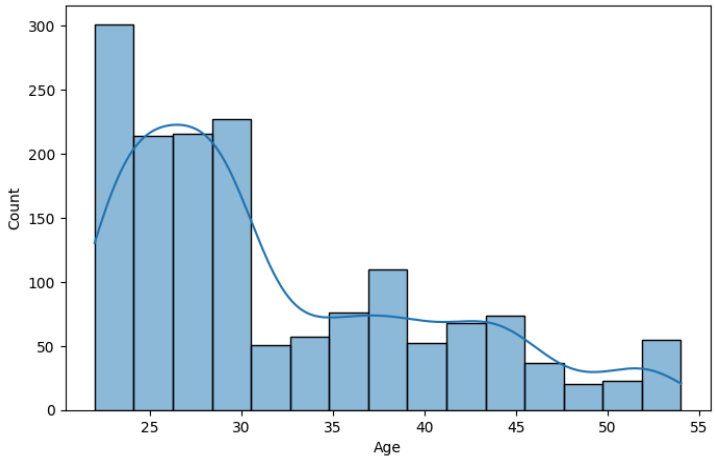


Fig: Histogram

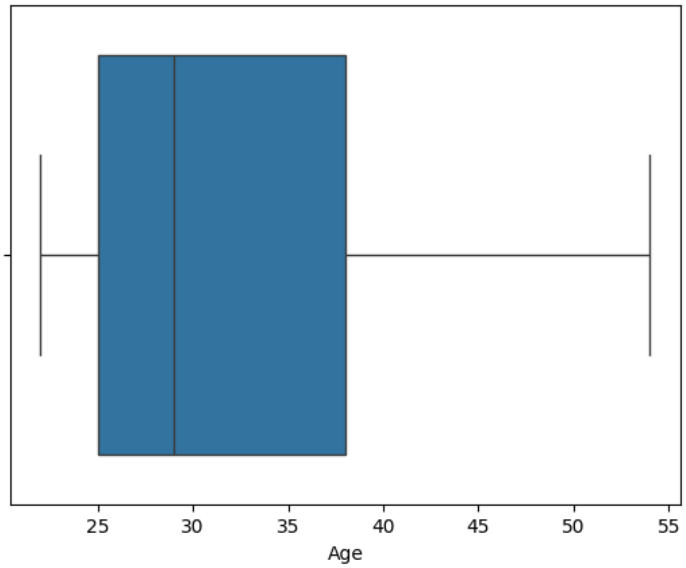


Fig: Boxplot

Observations:

- Most of the people in our data are young and under 40 years of age.
- In our dataset, the minimum age is 22, and the maximum age is 54.

If we consider **salary**, it is not perfectly normally distributed, it shows right skewness as there is a rise around salary of 50,000 - 60,000.

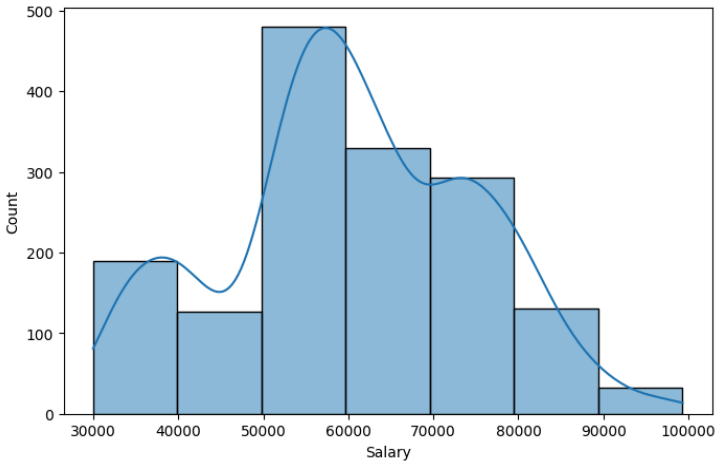


Fig: Histogram

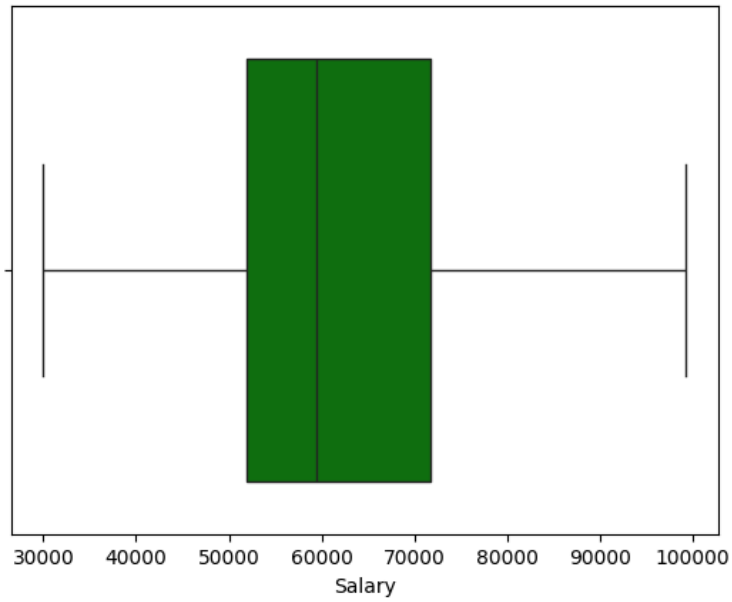


Fig: Boxplot

Observations:

- If we see the boxplot we could say that salary is almost normally distributed. It is not perfectly symmetrical because the median is towards the lower whisker.
- The minimum salary is 30,000 and the maximum salary is 99300 with a median salary of 59500.

If we consider total_salary, the histogram plot and boxplot are as follows:

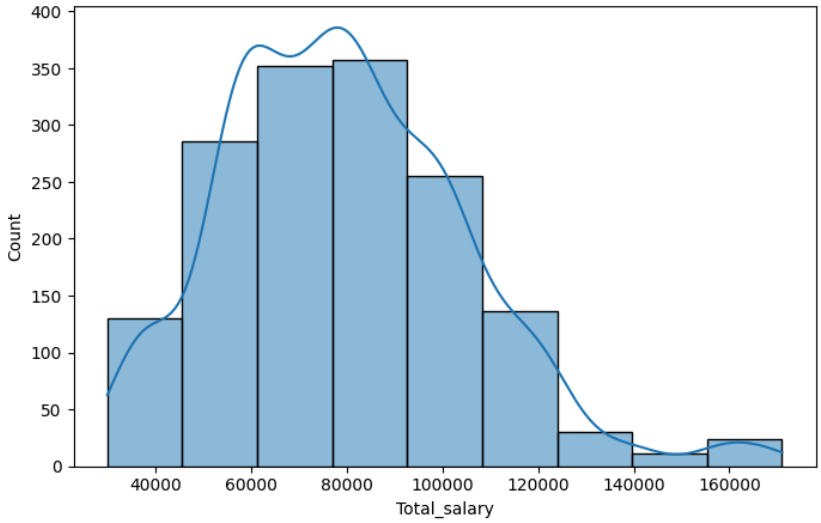


Fig: Histogram

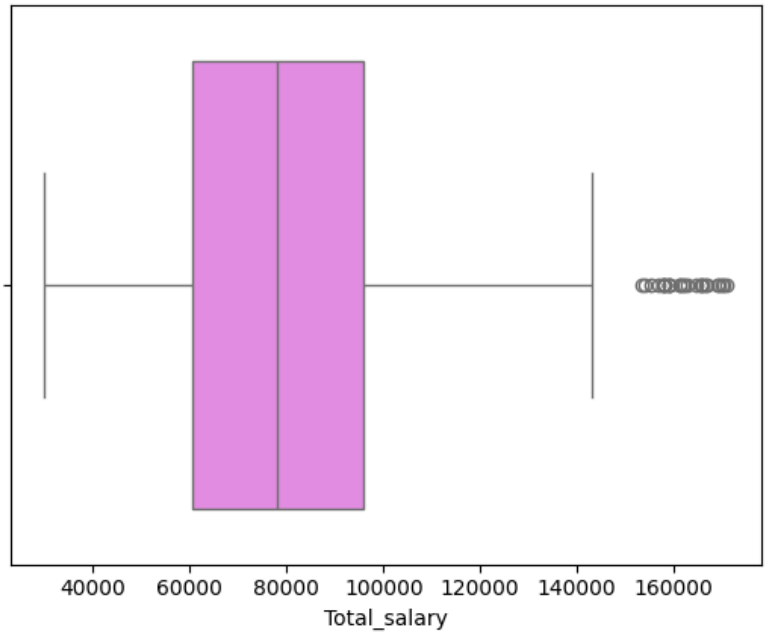
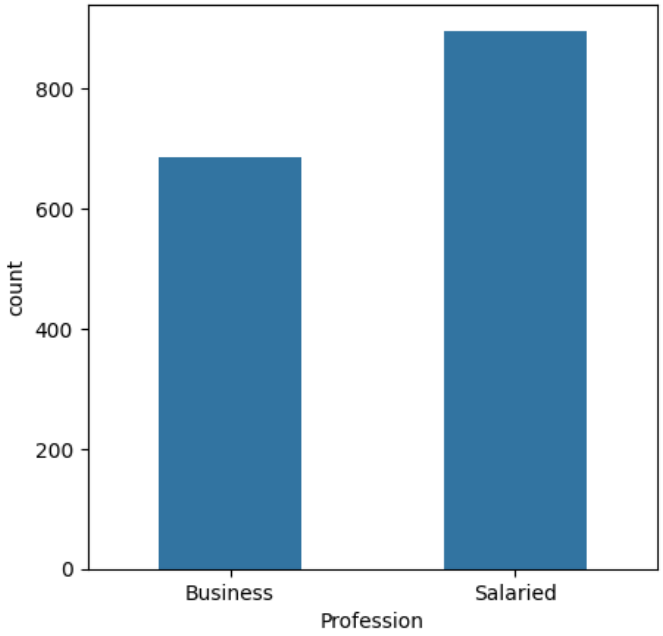


Fig: Boxplot

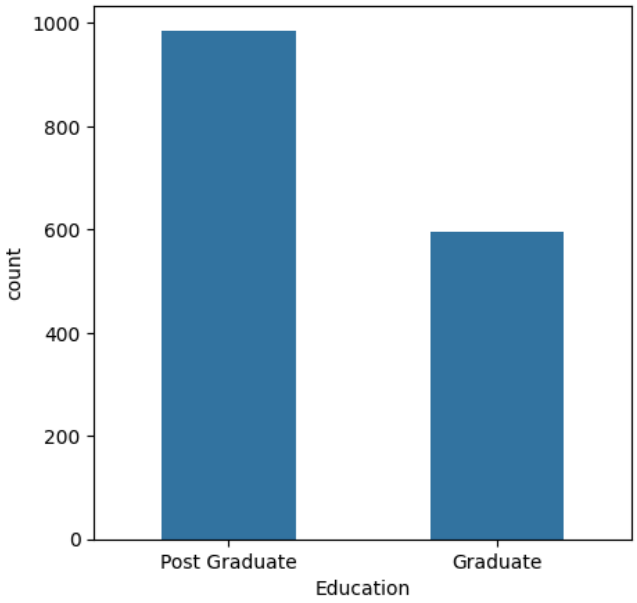
Observations:

- We could see outliers in the boxplot.
- The reason is that there are few people who earn a lot of money when compared to other people and it is absolutely normal in real life.
- Assuming the situation I have not treated the outliers here.

Let’s explore a few categorical variables now:



Observation: There are around 56% of salaried people and 46% of business people.



Observation: There are more than 62% of people with post-graduation and around 38% of people are graduates.

Bivariate analysis:
Bi means two and variate means variable, so there are two variables. The analysis is related to the relationship between two variables.

Different types of Bivariate analysis that can be done are:

- Bivariate analysis of two numerical variables
- Bivariate analysis of two categorical variables
- Bivariate analysis of one numerical variable and one categorical variable.

Correlation by Heatmap:
A heatmap is a graphical representation of data as a colour-encoded matrix. It is a great way of representing the correlation for each pair of columns in the data.

Fig: Correlation matrix of numerical variables with their respective correlation values

Observations:

- Salary and age are positively correlated with a correlation score of 0.62. In simple terms, we can say that as age increases the salary of a person will also increase.
- We could also say that Partner_salary and Total_salary are positively correlated with a correlation of 0.82. If the partner's salary is high, the total salary will also be higher.
- No_of_Dependents is negatively correlated with Salary, Age and Price with a correlation of -0.032, -0.19 and -0.14 respectively.
- From the initial analysis, we could say that dropping the No_of_Dependents column will not cause any problems as we move forward in our analysis.

Considering two numerical variables, **Age and Price**. To understand the trend or how these variables are related we can use a line plot for visualisation.

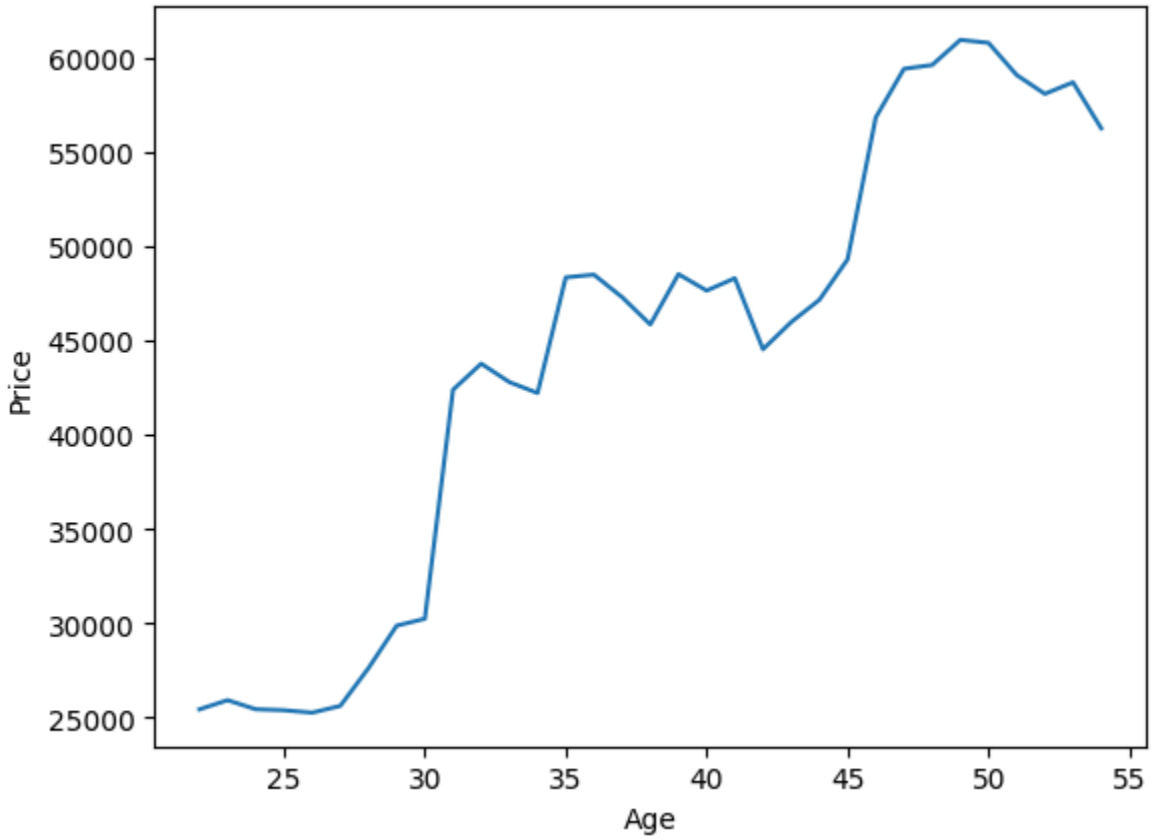


Fig: Age vs Price

Observation: Overall, as we grow older the money spent on cars is also increasing.

Considering two numerical variables, **Age and Total_salary**. We use a line plot to understand the trend between these variables.

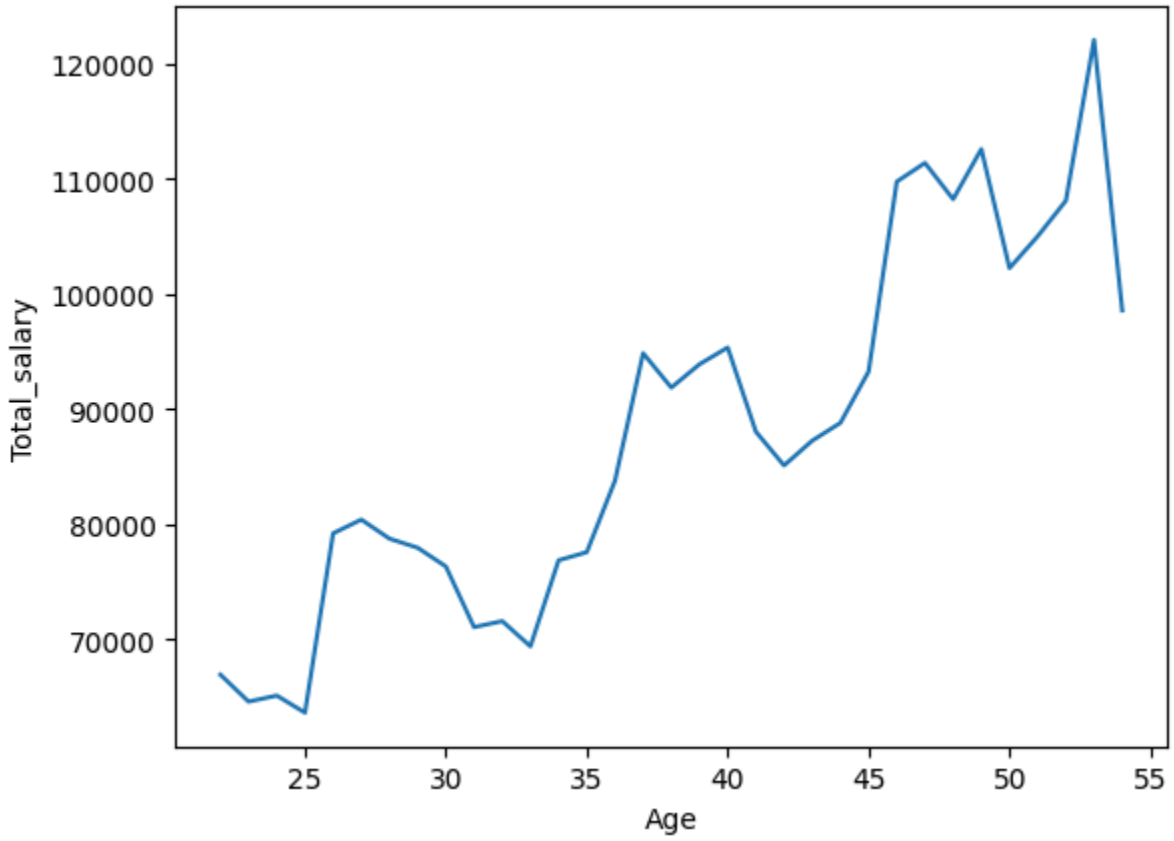


Fig: Age vs Total_salary

Observation: Despite a few dips in the Total_salary at certain ages, the general trend is that as we grow older the Total_salary also increases.

Considering a categorical variable and a numerical variable, **Gender and Price**. I used a boxplot to understand the purchasing trends between males and females.

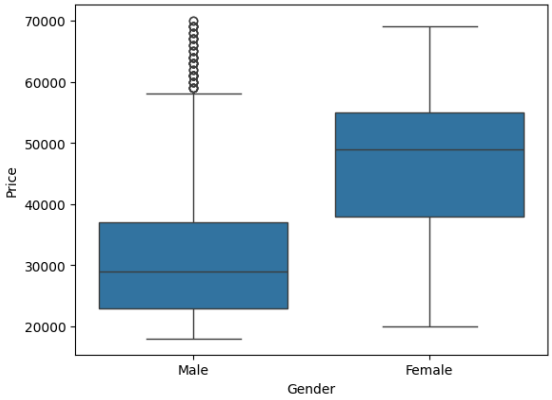


Fig: Gender vs Price

Observations:

- Overall, we could see that money spent on cars by females is higher as their median is higher when compared to males.
- If we consider males, there are few cases where people have spent a lot of money on cars.
- If we see purchasing behaviour, females tend to have a wider range of choices in terms of buying cars. Few spent less money whereas few spent more money.
- If we consider males, we can see most of them buying cars that fall under a particular range of prices.

Considering two categorical variables and one numerical variable, **Marital_status, Profession and Price**. We are now gonna see how purchasing behaviour changes as we consider both marital status and the profession they are in. We used a boxplot to understand the pattern among these three variables.

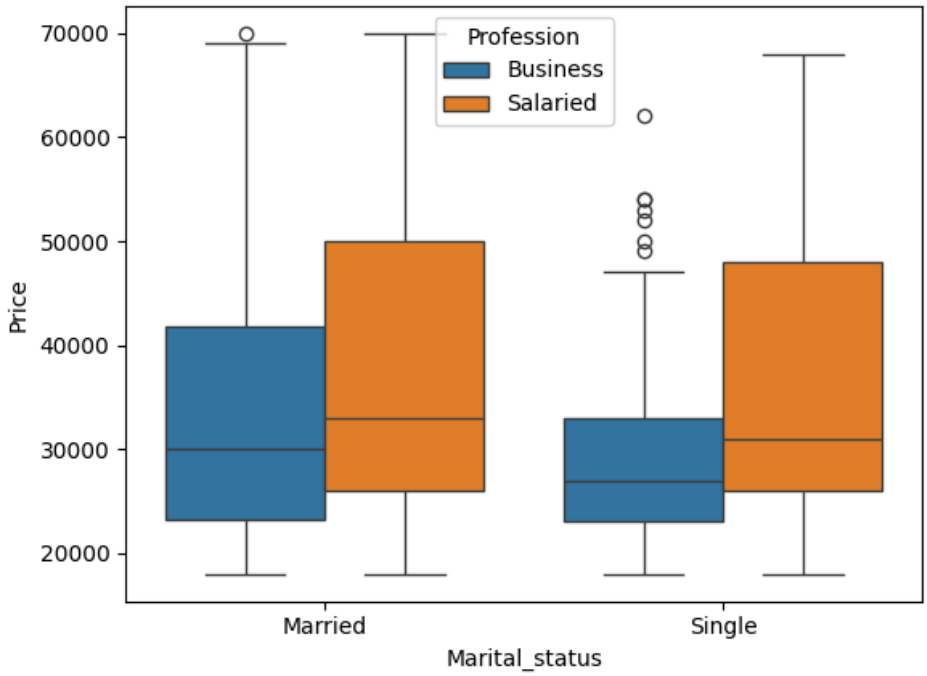
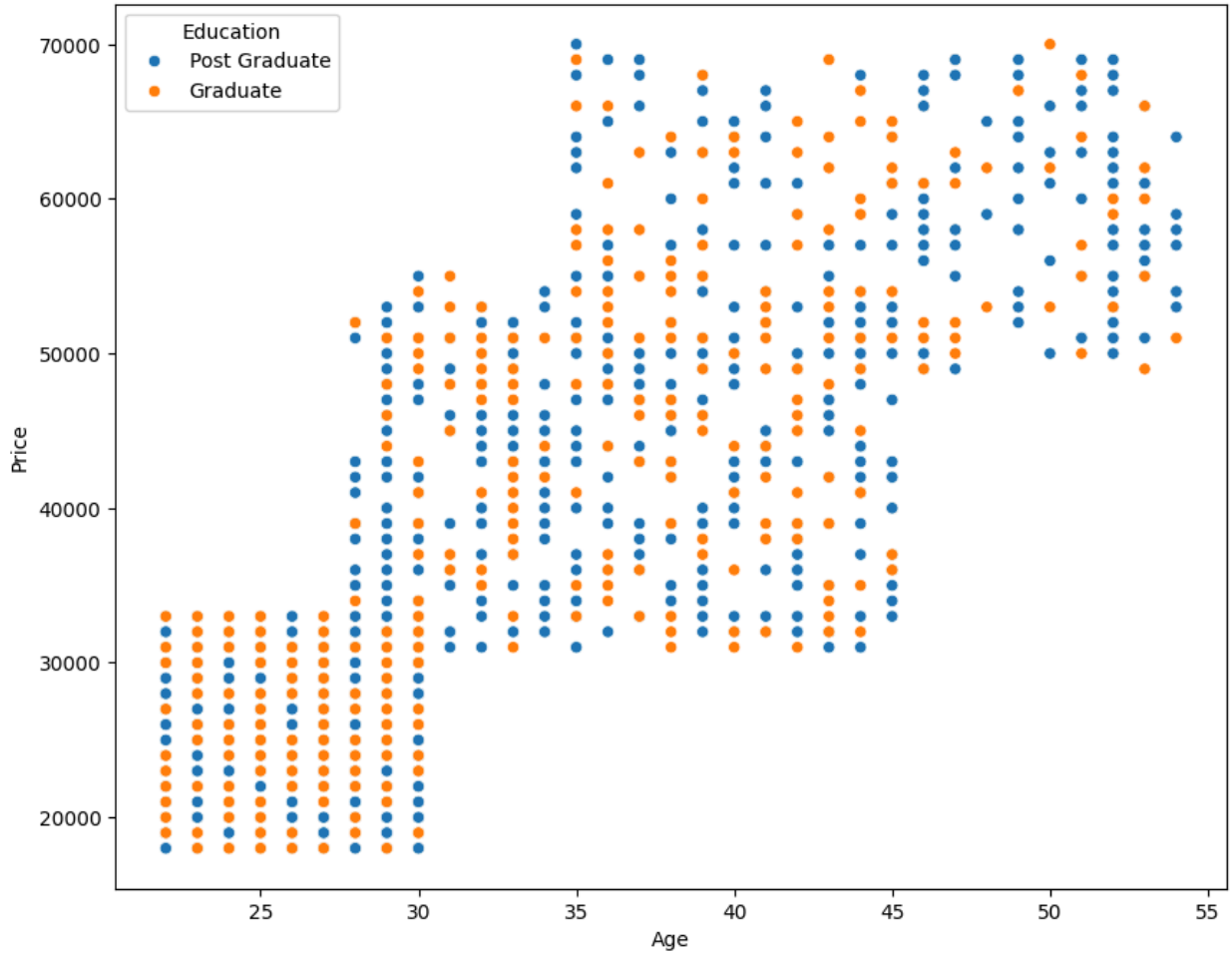


Fig: Impact on purchasing when considering marital status and profession

Observations:

- We could see all the boxplots are right-skewed.
- Overall, we can see that married people tend to spend more on cars than single people.
- The median amount spent on buying cars ranges from 28,000 to 33,000 for all groups of people.
- This shows that many people are interested in buying affordable cars.
- We can see that there are few single people who are into business and have spent huge money on cars.

Considering two numerical variables and one categorical variable, **Age, Price and Education**. We understand the purchasing behaviour of people with age and their education qualifications. I used a **scatterplot** to understand the relationship among these variables.



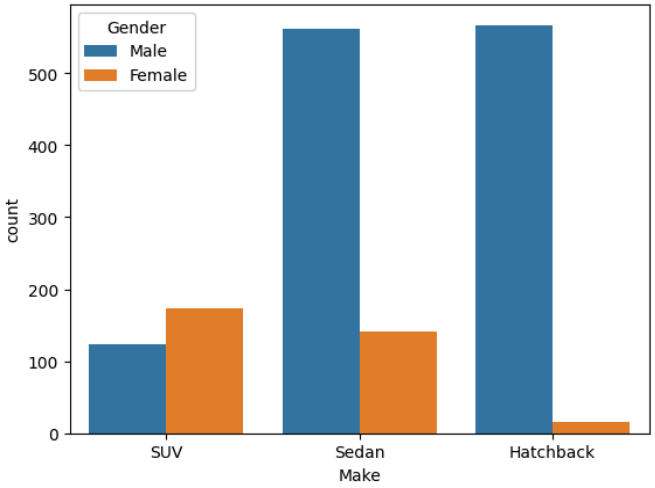
Observations:

- All the people under 30 years of age and are graduates showed keen interest in buying affordable cars.
- As the age increases we can see an increasing trend of money spent on cars by both Graduates and Post Graduates.
- We can take note that for ages 40 to 55 there is a tendency to buy expensive cars among which most of them are postgraduates.
- The overall trend is that as age increases the amount of money spent on cars also increases.

Key Questions:

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

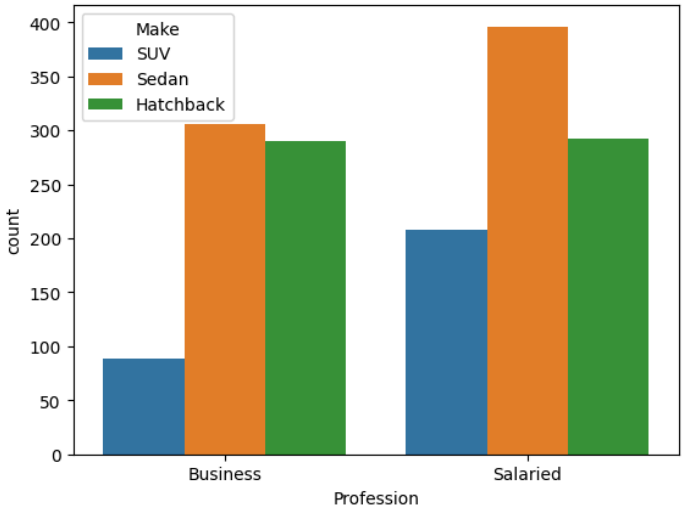
To answer this question I have used a count plot. I have taken x as Make and hue as Gender. Below is the count plot.



Females tend to prefer SUVs more than males.

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

I have first grouped all the rows where the profession is salaried. Then I have further grouped in the salaried rows where the Make is Sedan.
By doing this we got a count of salaried people and also salaried people who have a sedan.
The likelihood of a salaried person buying a sedan is number of salaried rows who bought a sedan divided by the total number of salaried rows.
By doing this, the likelihood of a salaried person buying a Sedan is 0.44.

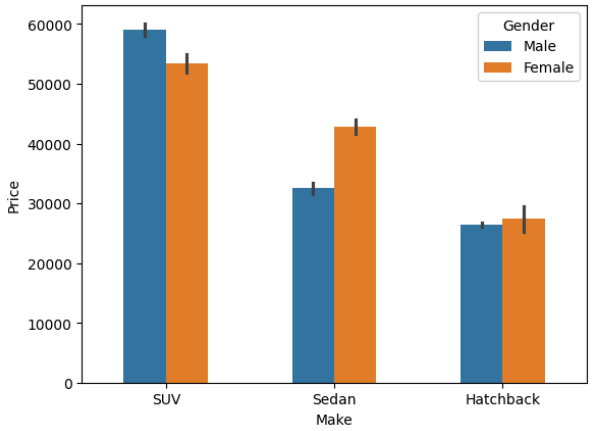


3) What evidence or data supports Sheldon Cooper's claim that a salaried male is an easier target for an SUV sale over a Sedan sale?

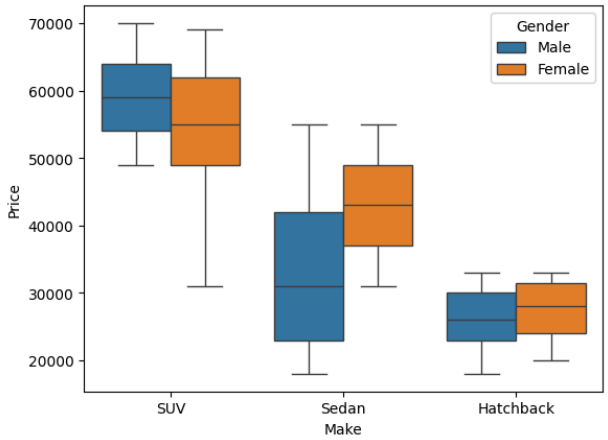
I have grouped all the rows where Gender is Male, Profession is Salaried and Make is SUV into salaried_male_SUV. Then I used shape to know the count of rows.
Then, I have grouped all the rows where Gender is Male, Profession is Salaried and Make is Sedan into salaried_male_Sedan. Then I used shape to know the count of rows.
From that, I have come to the conclusion that Sheldon Cooper's claim that a salaried male is an easier target for an SUV sale over a Sedan sale is wrong.

4) How does the amount spent on purchasing automobiles vary by Gender?

I have used barplot to know the average money spent on all the type of cars by males and females.



I have also used a boxplot to know the distribution of data and understand the patterns.

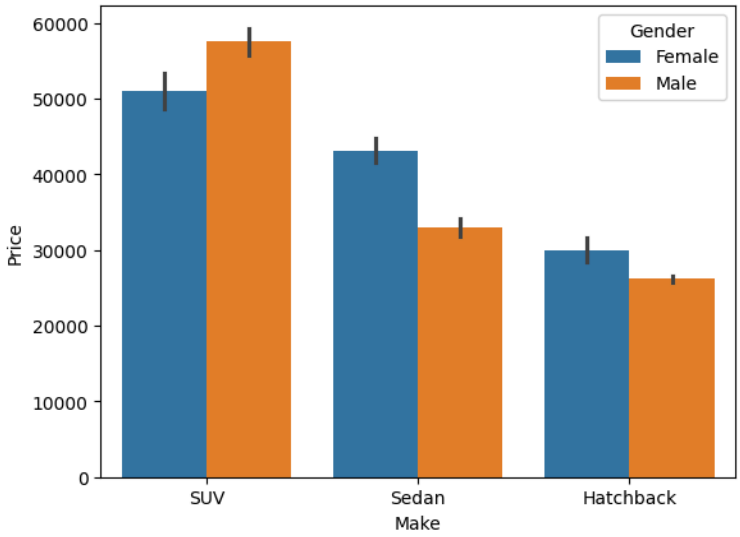


Observations:

- Overall males spend more money on SUVs than Females.
- Overall females spend more money on Sedans and Hatchbacks than males.
- If we consider Sedans then we could observe that the spread of data is more for males.

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

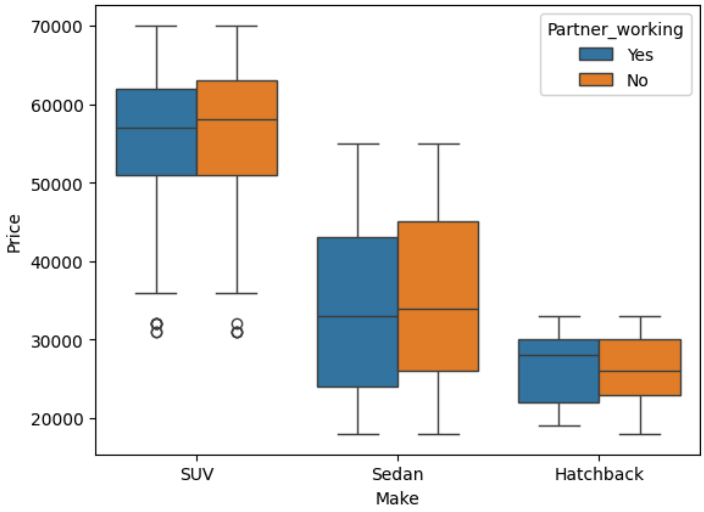
First I have grouped the data where Personal_loan is Yes to people_PL data frame. Then I used a bar plot to know the average money spent on different types of cars.



Observations:

- Females spent an average of 50,000 on SUVs.
- Males spent an average of 57,000 on SUVs.
- Females spent an average of 42,000 on Sedan.
- Males spent an average of 32,000 on Sedan.
- Females spent an average of 30,000 on Hatchback.
- Males spent an average of 28,000 on Hatchback.

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



Observations:

- We could see that despite not having a partner there are people who have purchased SUVs.
- The sedan is more affordable and highly chosen by people. Sedan prices are spread out significantly.
- Hatchbacks are the cheapest.
- In general, we can say that Partner_working status does not have much impact on the purchasing behaviour of people.

Final Outcomes:

- 75% of the population falls under the age of 38 which means that most buyers are young.
- 79% of the population is male in the dataset.
- Most buyers are married and have a post-graduation degree. (i.e. 62% of people)
- Business and salaried individuals are almost the same but salaried people have a wider range of salaries.
- Graduates buy affordable cars whereas Postgraduates buy expensive cars.
- The median purchase price is higher for salaried people than for people involved in business.
- Males spend the most money on SUVs, while females are more into buying Sedans and Hatchbacks.
- SUVs are the most expensive, followed by Sedans. Hatchbacks are the cheapest.
- We could say that income grows with age.
- High-income people avoid taking loans.
- Married individuals spend more money on cars compared to singles.
- The average price range for most buyers is 28,000 to 33,000.
- The sedan is the most chosen car type by individuals.

Recommendations for a better marketing campaign:

Digital Advertising

- As most people fall under the category of youth, using social media platforms like Instagram and Facebook can help generate more leads.
- Customising the ads based on the segment of people present.
- Since young people are in majority collaborating with influencers will also help a lot.

Instant cashback and other perks

- For every new car release, if you book within 30 days of the car launch, you will get a cashback of 10,000 at the time of delivery of your car.
- Attracting young buyers by giving an EMI break of 6 months from the date of buying the car.
- Additional discounts on weekday purchases.

Ease of loan approval and other important things for faster and higher conversion of sales

- Partnering with top banks for easy loan approvals and processing.
- Having highly skilled salesperson for good customer interactions.
- Offering low down payment schemes for people who already have taken either personal or home loans.
- Offering free insurance for people buying high-priced cars.

Gender-based marketing

- For Male buyers, promote the highlights of SUVs and talk more about the features.
- As female buyers are more into spending money on sedans and Hatchbacks, make them aware of the safety and ease of driving those cars.
- For both male and female buyers try giving complimentary things like installing car covers for free or giving them important car accessories which will be needed for them.