greatlearning
Learning for Life

SMDM PROJECT REPORT

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.

Data Description

- **Age**: The age of the individual in years.
- **Gender**: The gender of the individual, categorized as male or female.
- **Profession**: The occupation or profession of the individual.
- **Marital_status**: The marital status of the individual, such as married &, single
- **Education**: The educational qualification of the individual Graduate and Post Graduate
- **No_of_Dependents**: The number of dependents (e.g., children, elderly parents) that the individual supports financially.
- **Personal_loan**: A binary variable indicating whether the individual has taken a personal loan "Yes" or "No"
- **House_loan**: A binary variable indicating whether the individual has taken a housing loan "Yes" or "No"
- **Partner_working**: A binary variable indicating whether the individual's partner is employed "Yes" or "No"
- **Salary**: The individual's salary or income.
- **Partner_salary**: The salary or income of the individual's partner, if applicable.
- **Total_salary**: The total combined salary of the individual and their partner (if applicable).
- **Price**: The price of a product or service.
- **Make**: The type of automobile

1.1. Information about the size of the dataset and the nature of the variables

Nature of dataset:

```
 #   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)
```

The dataset comprises 1581 observations, each containing 14 entries. Specifically, it includes 5 integer-type variables representing numerical data, 1 float-type variable representing numerical data, and 8 object-type variables representing categorical data.

1.2. Size of the Dataset

```
Size of the dataset: 1581 rows, 14 columns
```

The data set contains 1581 observations of data and 14 variables.

2.1 Checking for missing values

```
Missing values/blanks in the dataset:
Age                   0
Gender               53
Profession            0
Marital_status        0
Education             0
No_of_Dependents      0
Personal_loan         0
House_loan            0
Partner_working       0
Salary                0
Partner_salary      106
Total_salary          0
Price                 0
Make                  0
dtype: int64
```

There are 53 null values in Gender and 106 null values in Partner_Salary dataset.

After Treatment:

```
Missing values/blanks in the dataset:
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
```

2.2 Data discrepancies and treatment

We found some discrepancies on the data

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

After Treatment:

```
Gender
Male      1199
Female     329
Name: count, dtype: int64
```

2.3 Checking on duplicate Vaules

```
Number of duplicate rows = 0
```

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

No Duplicates Found

2.4 Summary of the Dataset

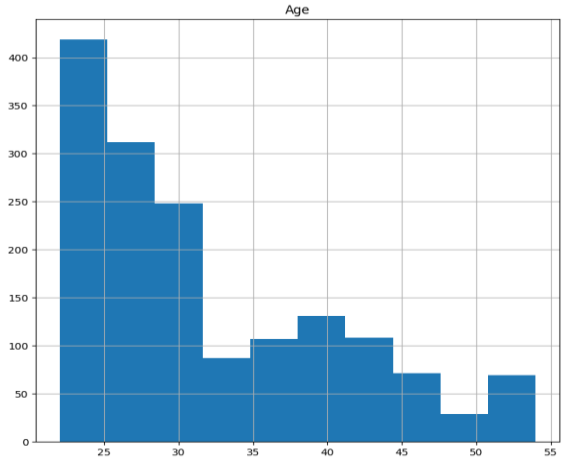
Statistical summary of numerical columns:

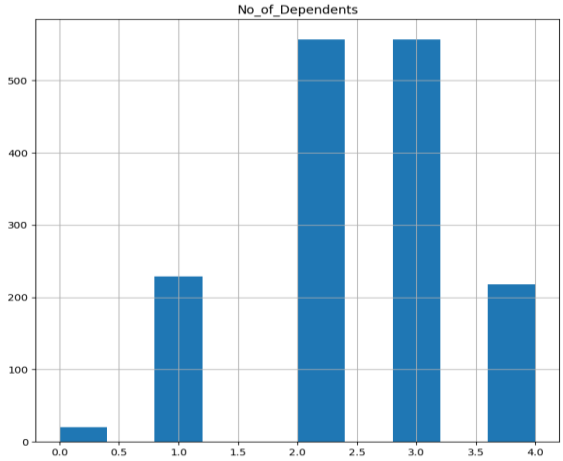
	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1581.0	20230.655880	18909.850652	0.0	0.0	24900.0	38000.0	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

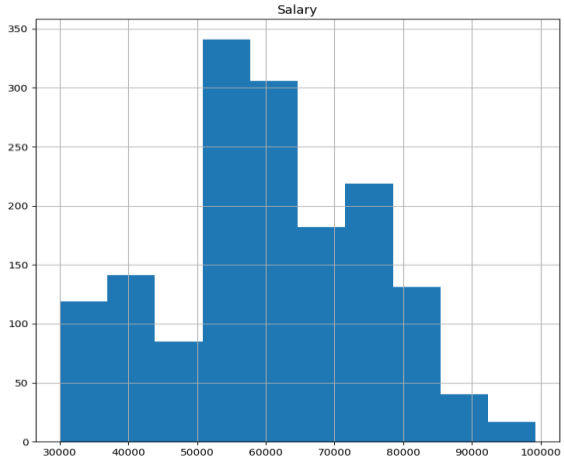
	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
count	1581.000000	1581.000000	1581.000000	1581.000000	1581.000000	1581.000000
mean	31.922201	2.457938	60392.220114	20230.655880	79625.996205	35597.722960
std	8.425978	0.943483	14674.825044	18909.850652	25545.857768	13633.636545
min	22.000000	0.000000	30000.000000	0.000000	30000.000000	18000.000000
25%	25.000000	2.000000	51900.000000	0.000000	60500.000000	25000.000000
50%	29.000000	2.000000	59500.000000	24900.000000	78000.000000	31000.000000
75%	38.000000	3.000000	71800.000000	38000.000000	95900.000000	47000.000000
max	54.000000	4.000000	99300.000000	80500.000000	171000.000000	70000.000000

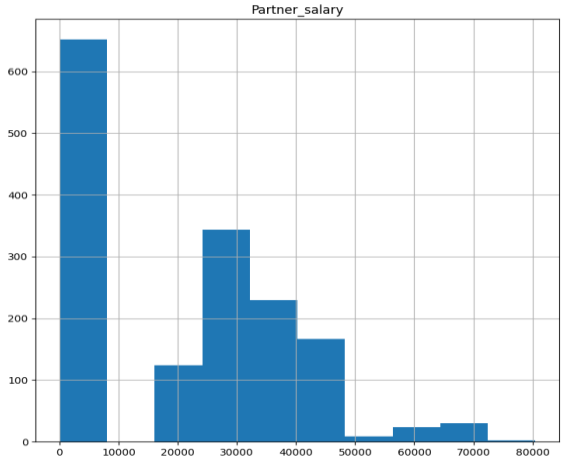
The provided illustration portrays the five-point summary of the continuous attributes. Upon analyzing the age column, it is evident that the distribution of the adult population spans from a minimum age of 22 years to a maximum age of 54 years. Specifically, 25% and 50% of individuals aged between 25 and 29 years have a number of dependents value of 2, while 75% of those aged 38 years have a number of dependents value of 3. Furthermore, the number of dependents is recorded as 0 for individuals aged 22 years (the minimum age) and as 4 for those aged 54 years (the maximum age).

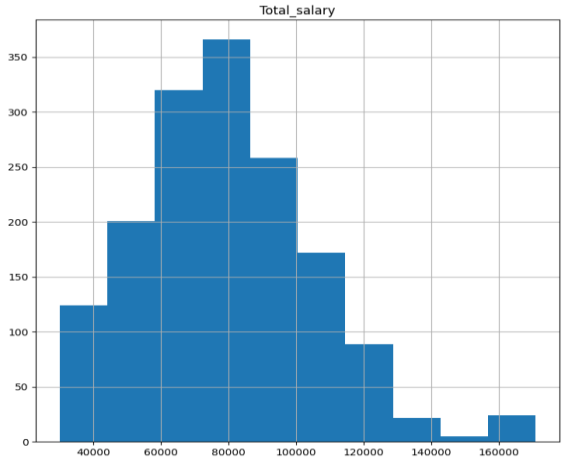
3.1 Let's plot the histogram to see the distribution of the continuous features continuously

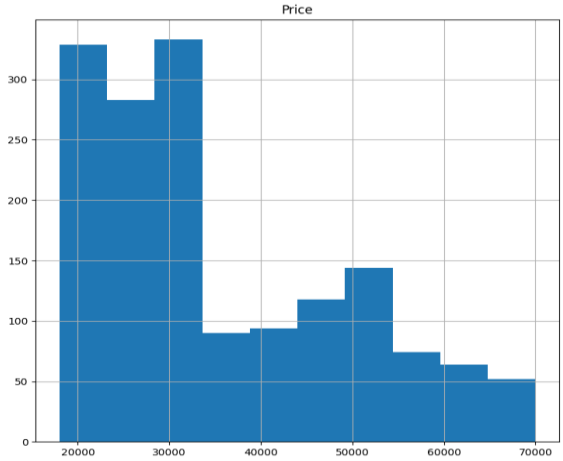






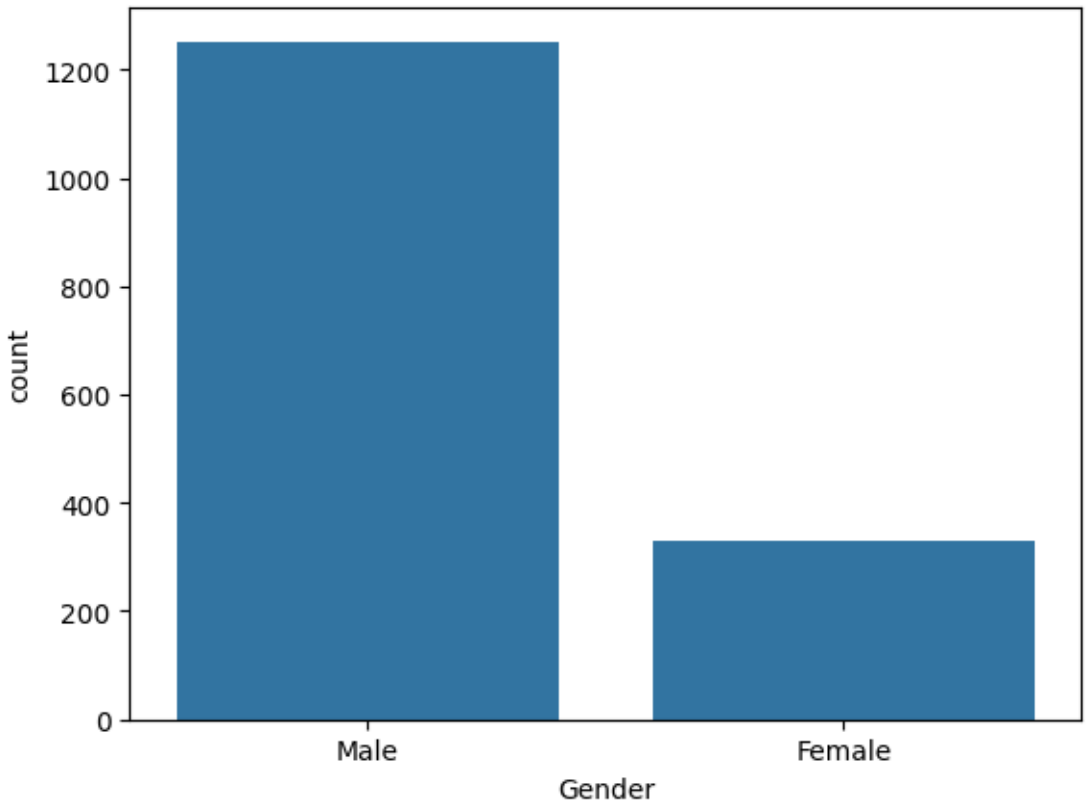




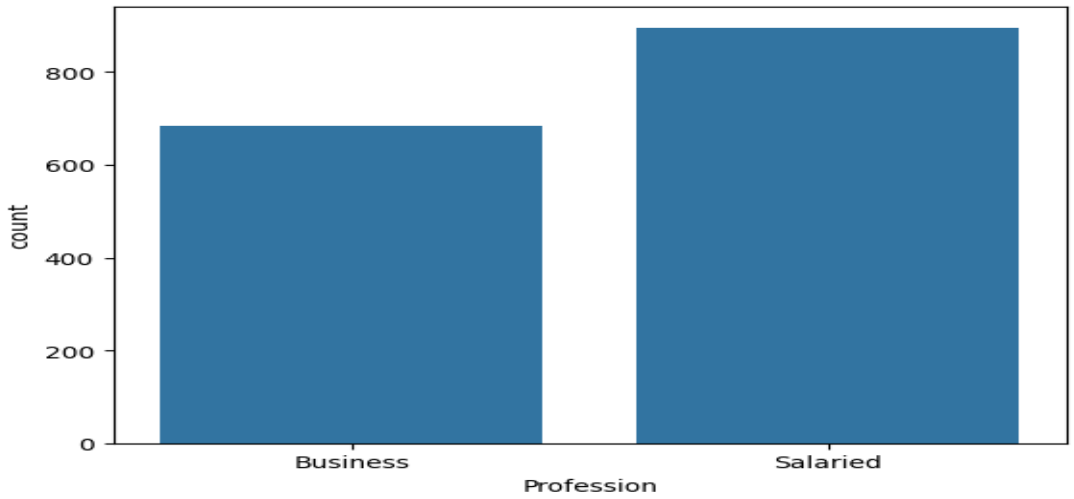


Based on the histograms above, it is evident that the age distribution is left-skewed, while the distribution of no_of_dependents is not uniform and appears right-skewed. The distribution of salary appears to be uniform, whereas that of partner_salary is non-uniform and exhibits a left-skewed pattern. Similarly, the distribution of total_salary and price also show left-skewed tendencies.

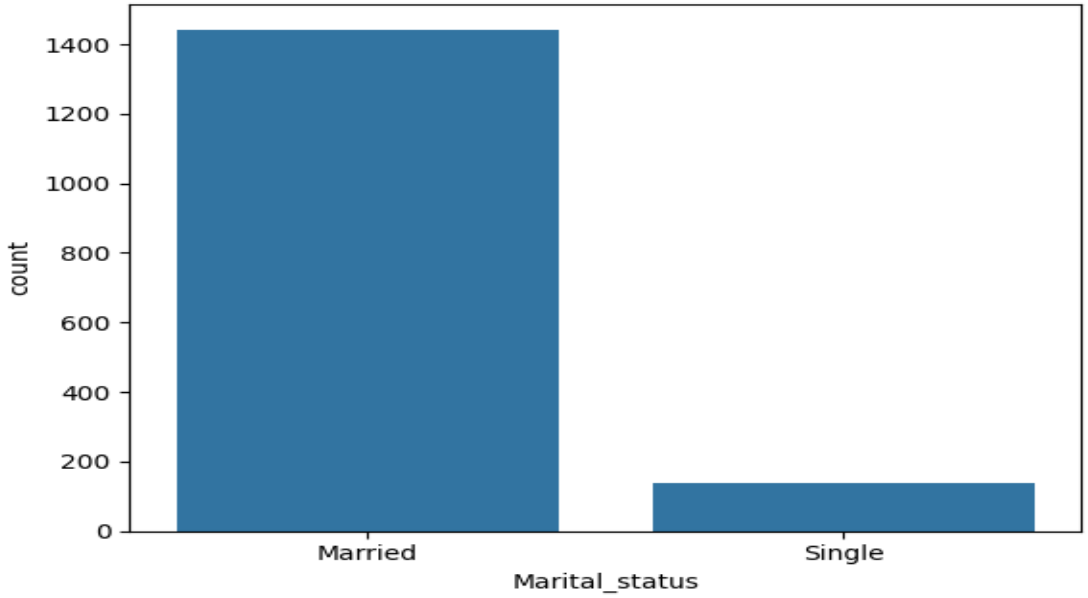
3.2 Now we shall look at how the variables are distributed with the help of countplot.



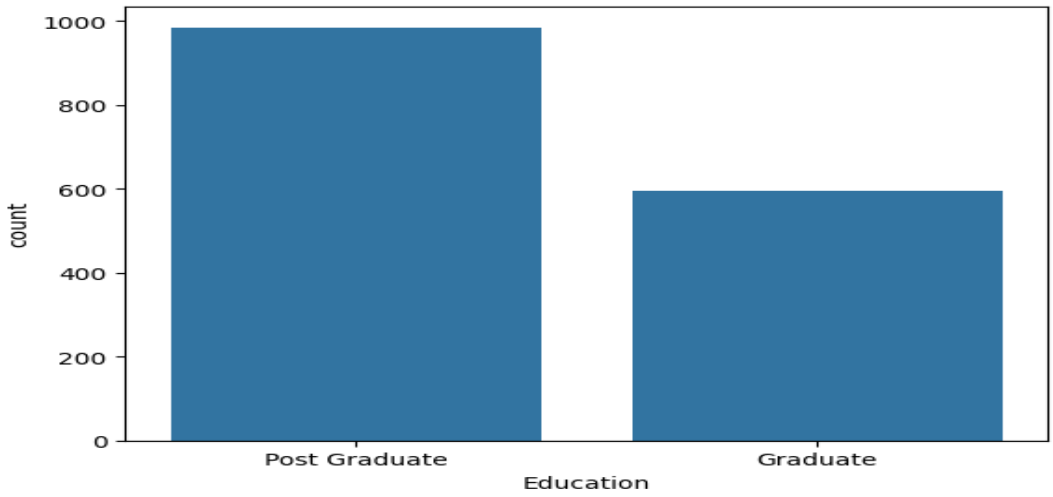
We can see Gender count of Male is higher than Female



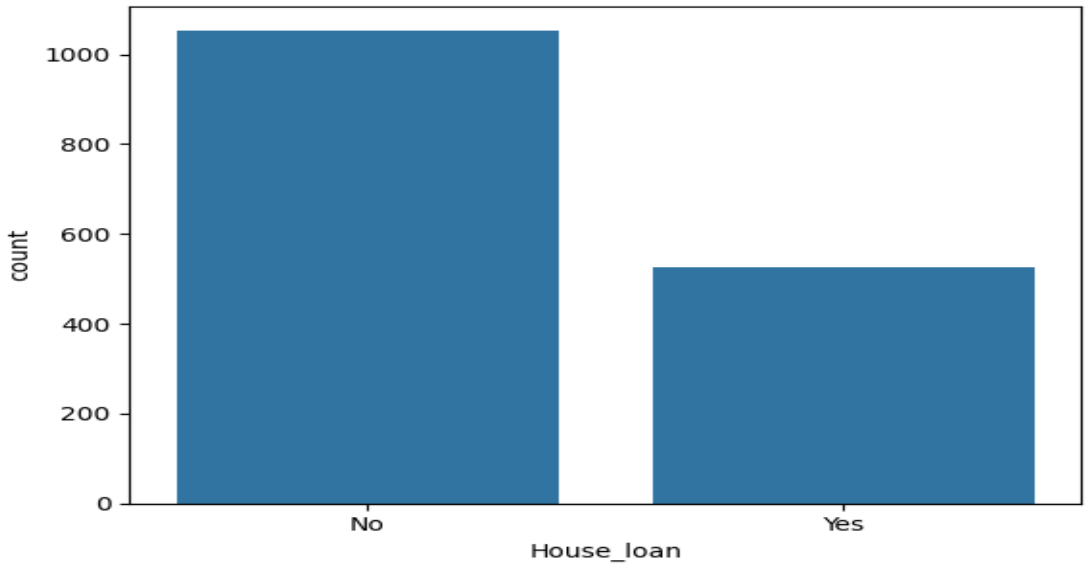
We can see professional wise – Salaried people are more than business professionals.



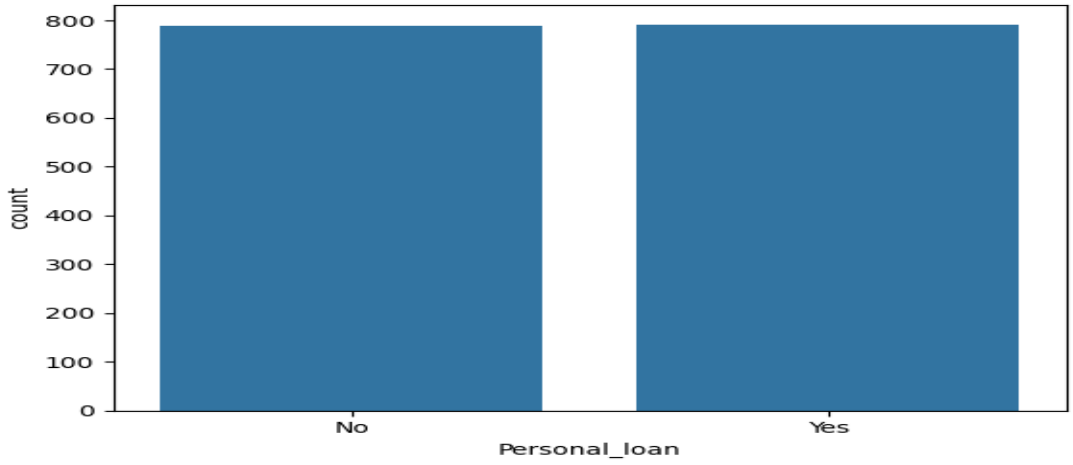
married people are more than the singled one's.



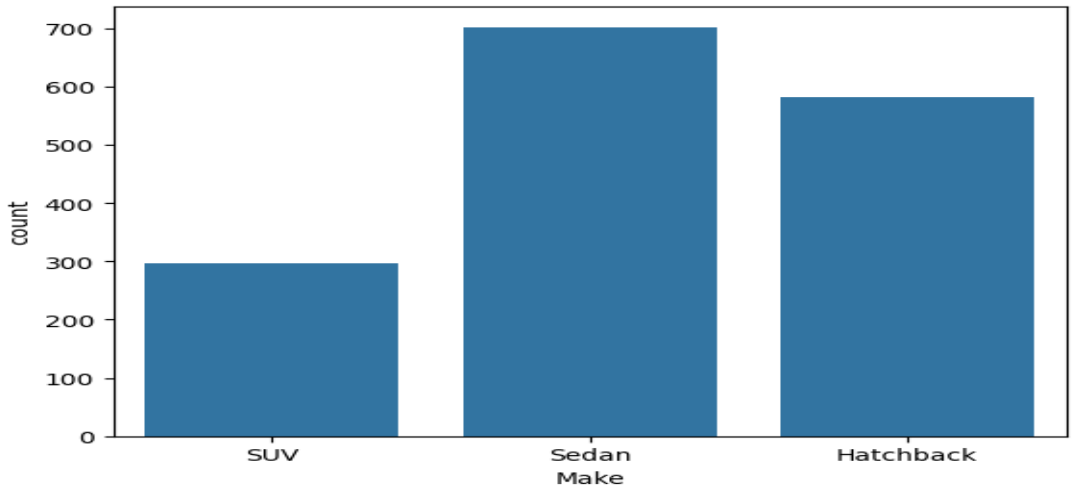
We can see most people background education shows as post graduate and compartively people have also pursued graduate degree education.

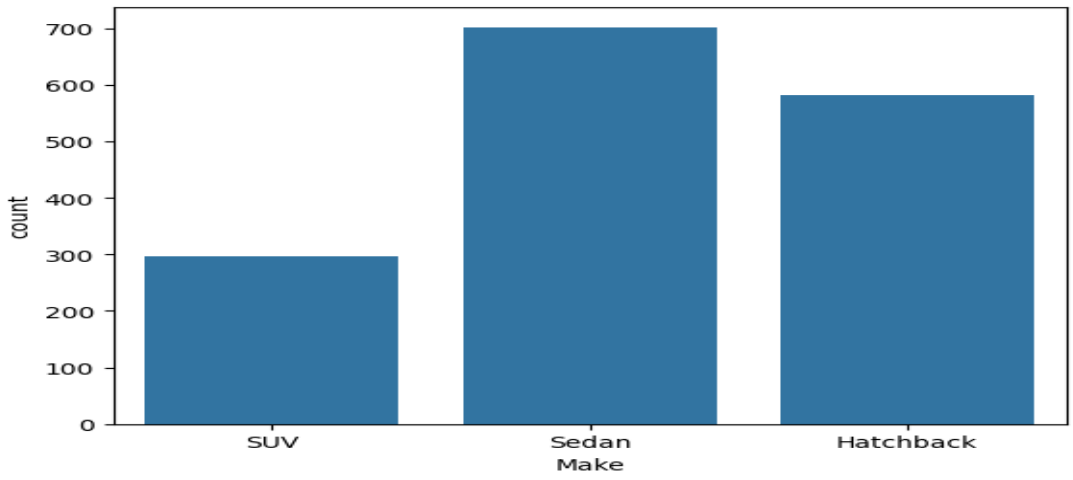


more number of people didn't take house_loan but half of the count of the people have taken house_loan.

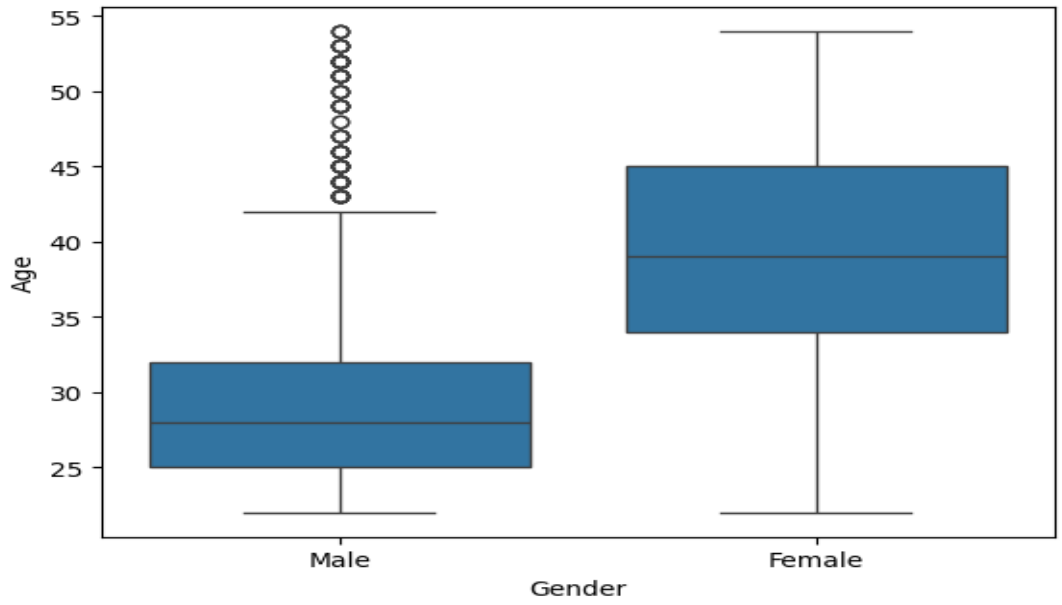


personal_loan status shows as same





The above plot depicts that Brand ‘Sedan’ is the most purchased followed by ‘Hatchback’ and the least is ‘SUV’



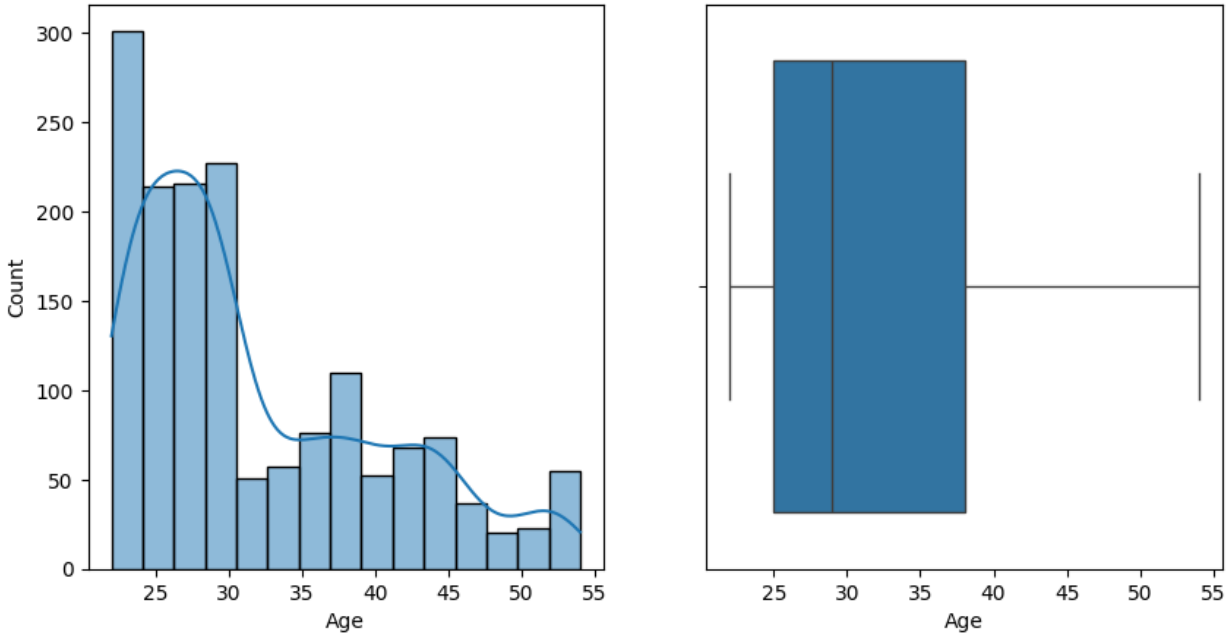
From the plot above, it is apparent that the male gender exhibits a higher frequency of extreme age values compared to the female gender.

Gender	Female	Male
Make		
Hatchback	15	567
SUV	173	124
Sedan	141	561

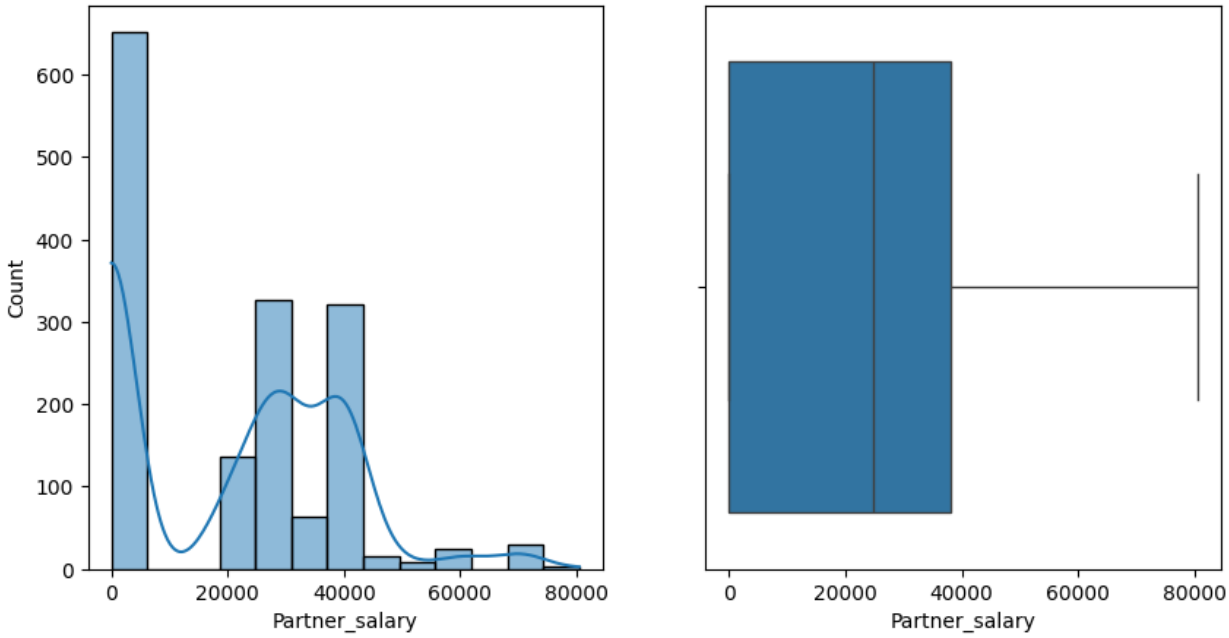
Here, the data suggests a preference among females for SUV automobiles, while males tend to favor hatchback automobiles.

Marital_status	Married	Single
Make		
Hatchback	498	84
SUV	281	16
Sedan	664	38

Another observation to note is that there appears to be a higher proportion of married individuals opting for sedan automobiles, whereas singles tend to show a preference for hatchback automobiles.

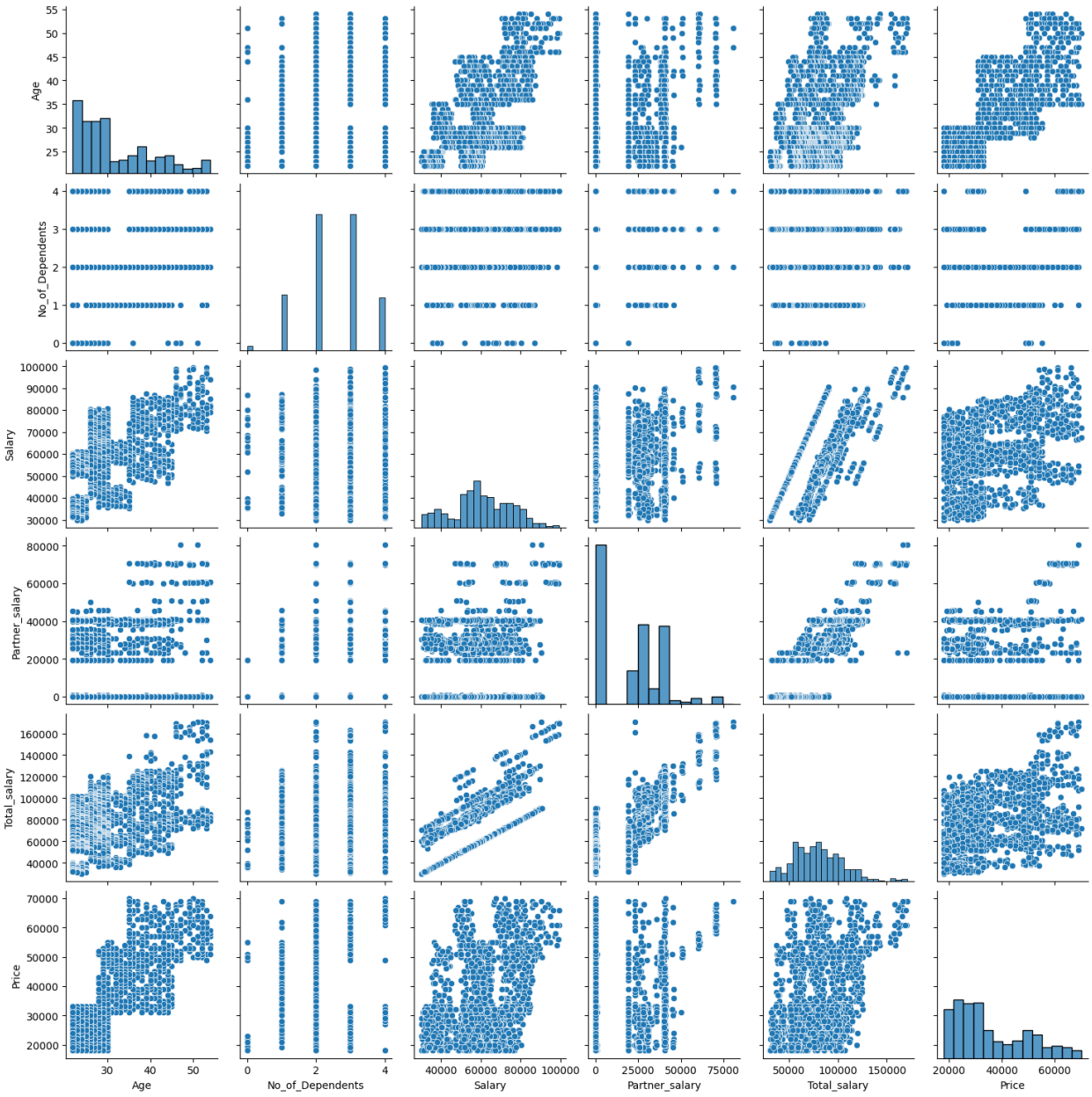


Using a histogram plot, we observe a right-skewed distribution in the variable "Age." Additionally, employing a boxplot reveals the absence of outliers within the "Age" variable.



Similarly, employing a boxplot analysis indicates the absence of outliers in the "Partner_Salary" variable. Furthermore, utilizing a histogram plot reveals a right-skewed distribution pattern within the "Partner_Salary" variable.

3.3 **Bi-variate distribution**



Utilizing a pairplot enables us to visualize the bivariate distribution. It appears that as age increases, salary also tends to increase. Additionally, there seems to be a positive correlation between age and the amount spent on purchases, indicating that individuals with higher ages tend to spend more.

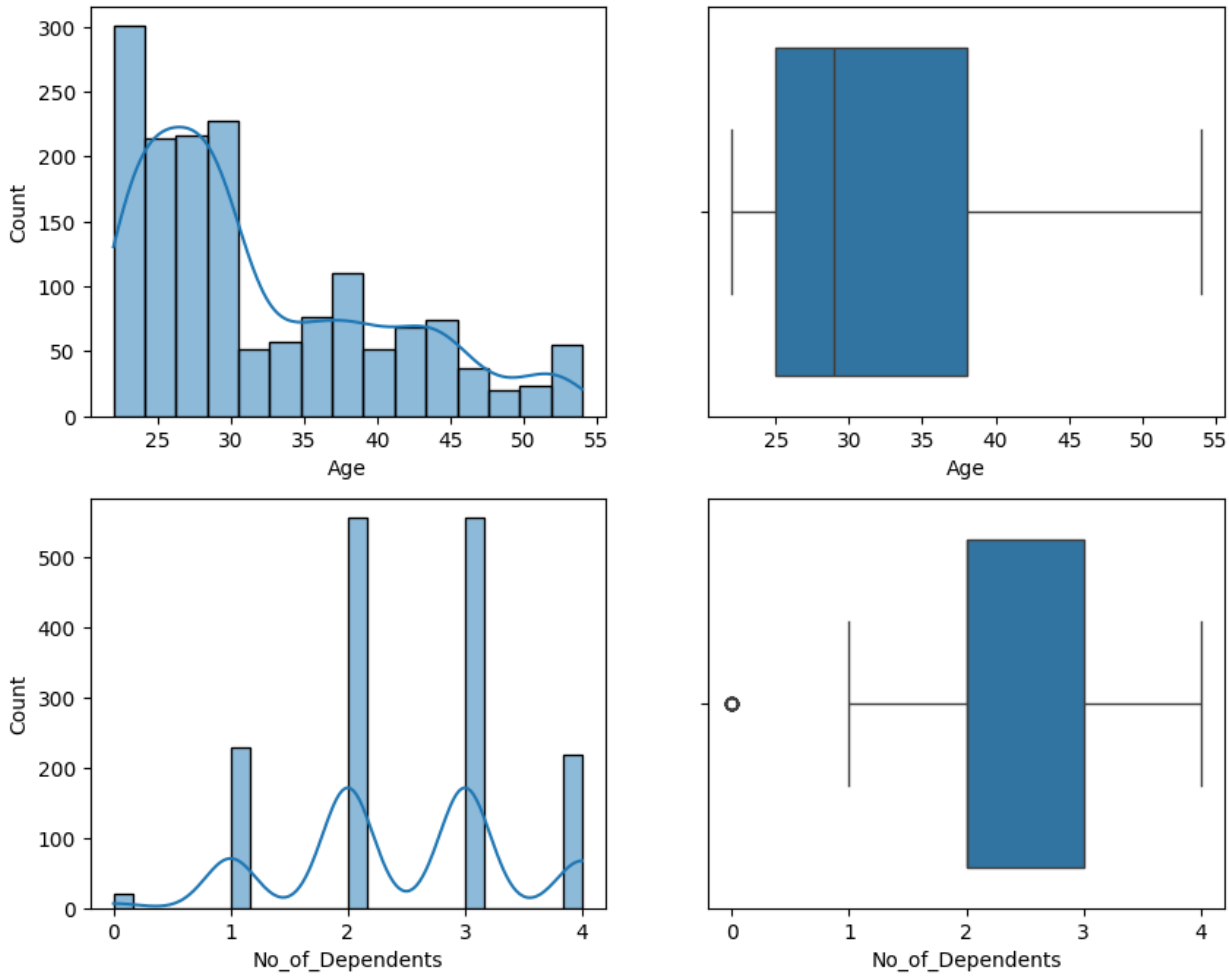
3.4 Correlation:

4.1 Univariate analysis using Age and no_of_dependents

	Age	No_of_Dependents
count	1581.000000	1581.000000
mean	31.922201	2.457938
std	8.425978	0.943483
min	22.000000	0.000000
25%	25.000000	2.000000
50%	29.000000	2.000000
75%	38.000000	3.000000
max	54.000000	4.000000

The "Age" attribute ranges from a minimum of 22 years to a maximum of 54 years. Upon analysis, it is observed that 50% of the individuals have an age of 29 years.

Regarding the "No_of_Dependents" attribute, the minimum value is 0, and the maximum is 4. Further examination reveals that 25% and 50% of the age group between 25 to 29 years have 2 dependents, while 75% of individuals aged 38 have 3 dependents.

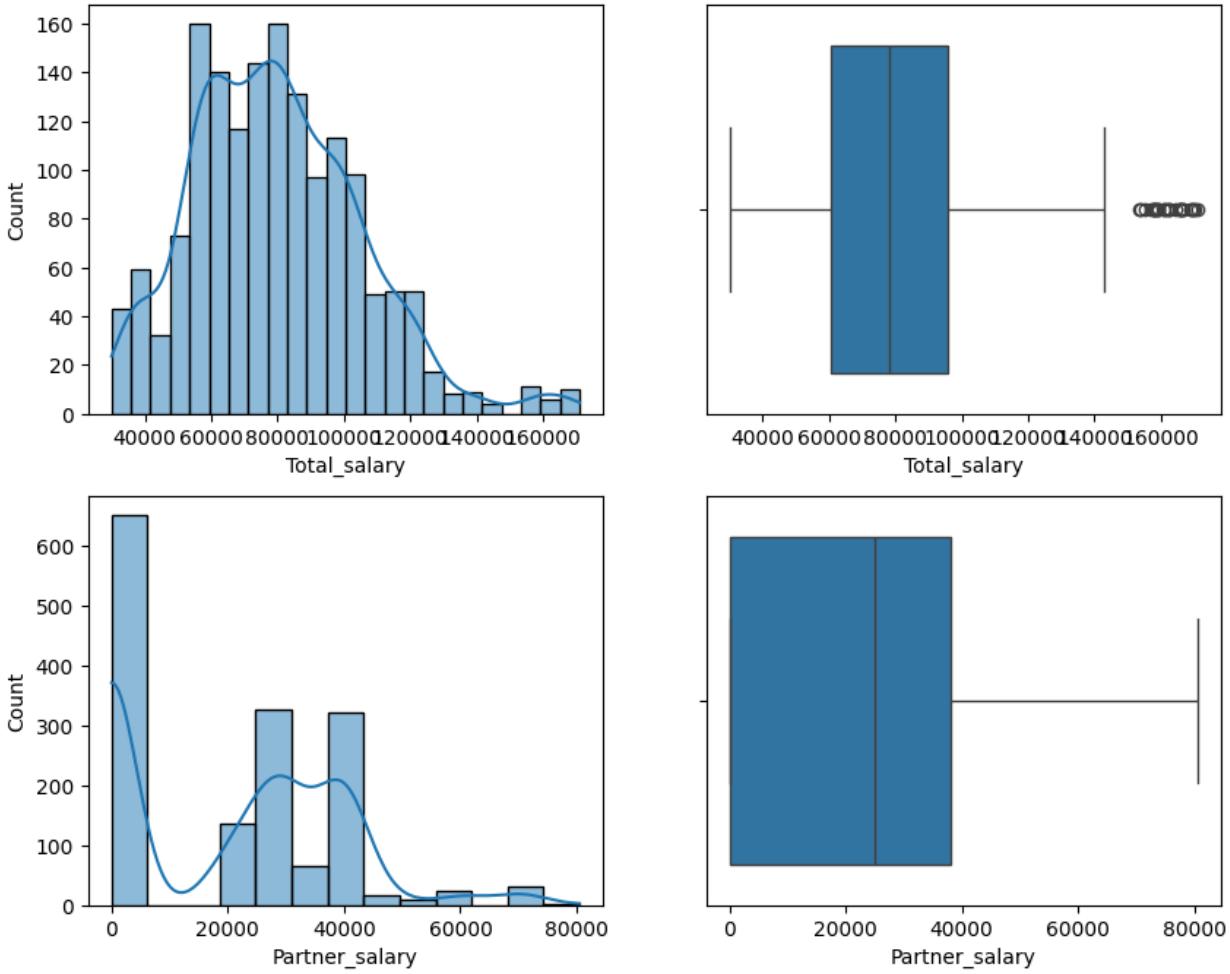


From the two plots above, it's evident that there are no outliers present in the "Age" variable. However, outliers are apparent in the "No_of_Dependents" variable.

4.2 Analysis on Total_salary and Partner_salary

	Total_salary	Partner_salary
count	1581.000000	1581.000000
mean	79625.996205	20230.655880
std	25545.857768	18909.850652
min	30000.000000	0.000000
25%	60500.000000	0.000000
50%	78000.000000	24900.000000
75%	95900.000000	38000.000000
max	171000.000000	80500.000000

From the data, we can conclude that the minimum salary for "Total_salary" is 30,000, with a maximum salary of 171,000. In the case of "Partner_salary," the minimum salary is 0, indicating that 25% of partners who are working do not contribute financially. The maximum salary observed in the "Partner_salary" column is 80,500.

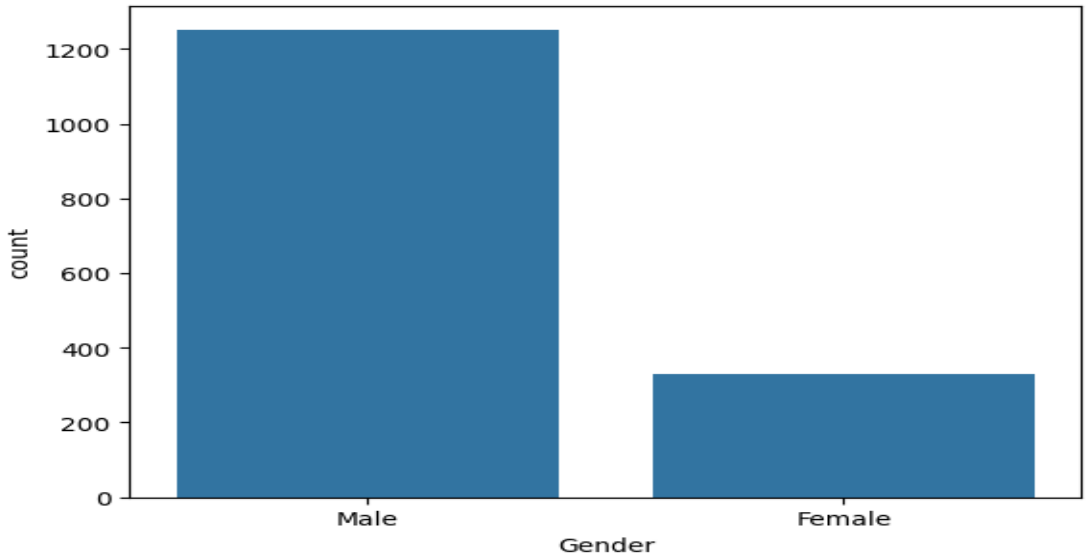


We can see more outliers found in Total_Salary which we will treat it later. But no outliers found in Partner_salary.

4.3 Categorical variable using Gender

```
Gender
Male      0.791904
Female    0.208096
Name: proportion, dtype: float64
```

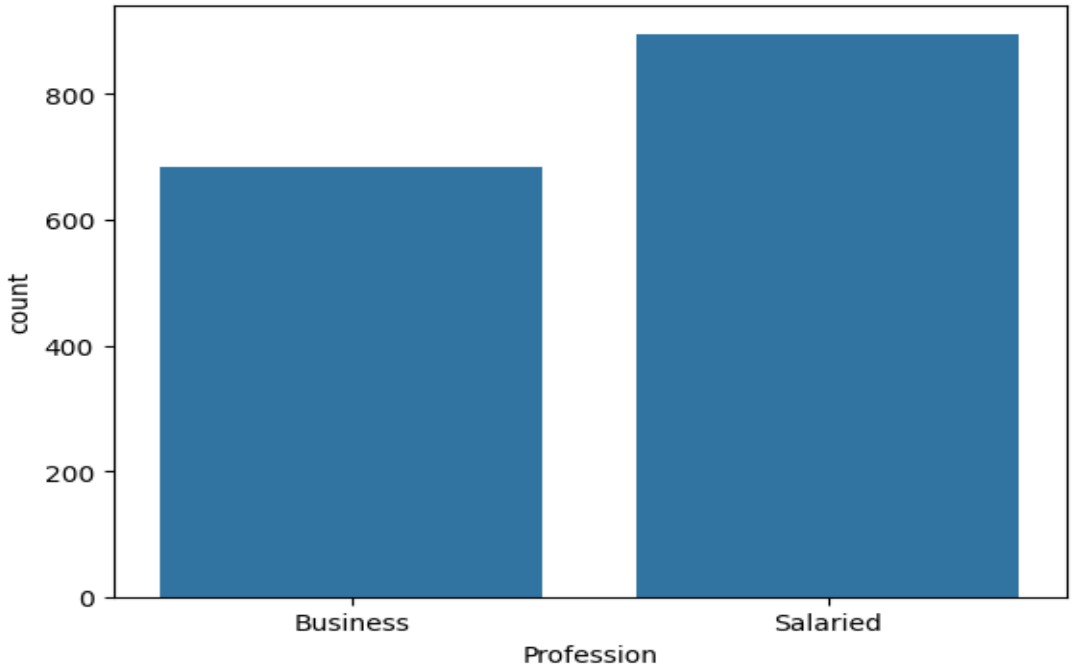
So the Gender categorical variable we have displayed the results in percentage form which contributes 0.80% as Male and %0.20 as Female



From the above plot, we can depict that count of Male gender is more when compared to Female gender.

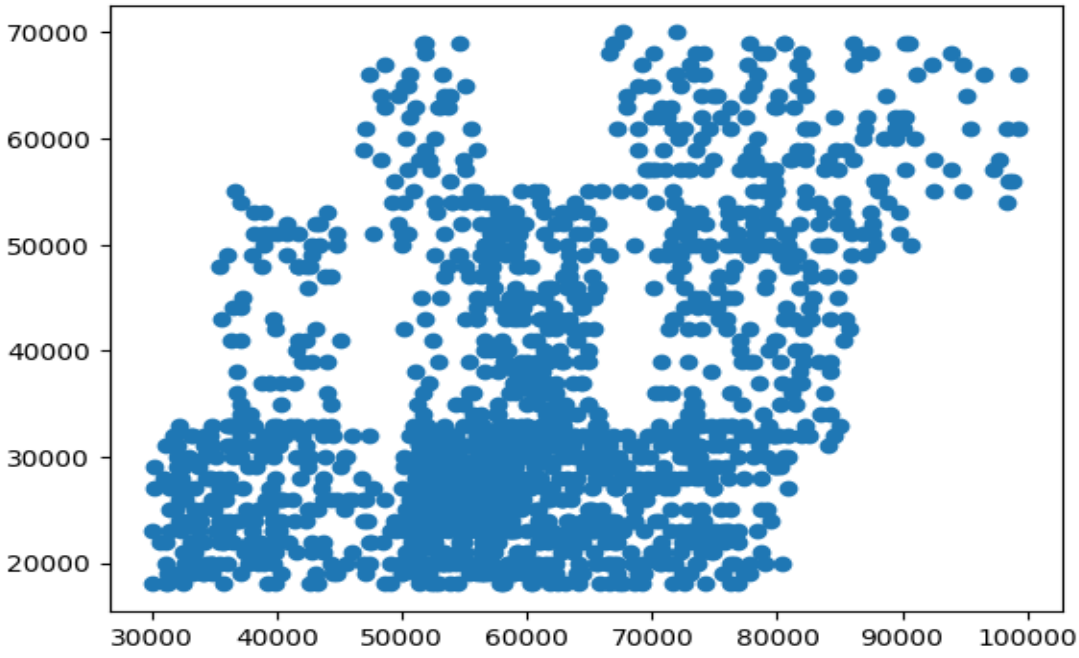
4.4 Categorical variable using Profession

```
Profession
Salaried    0.56673
Business    0.43327
Name: proportion, dtype: float64
```



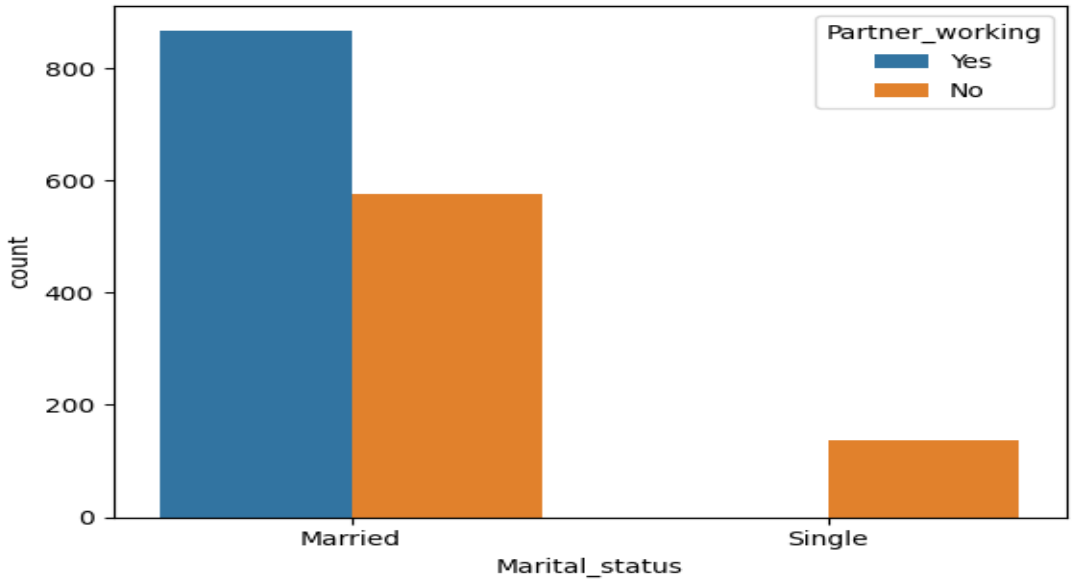
So we can conclude that 0.57% constitutes Salaried profession and remaining 0.43% belongs to Business profession

4.5 Bivariate Analysis using 2 numeric variables such as Salary and Price



we can see as gradually salary keeps on increasing even the amount of price spent also gets increased.

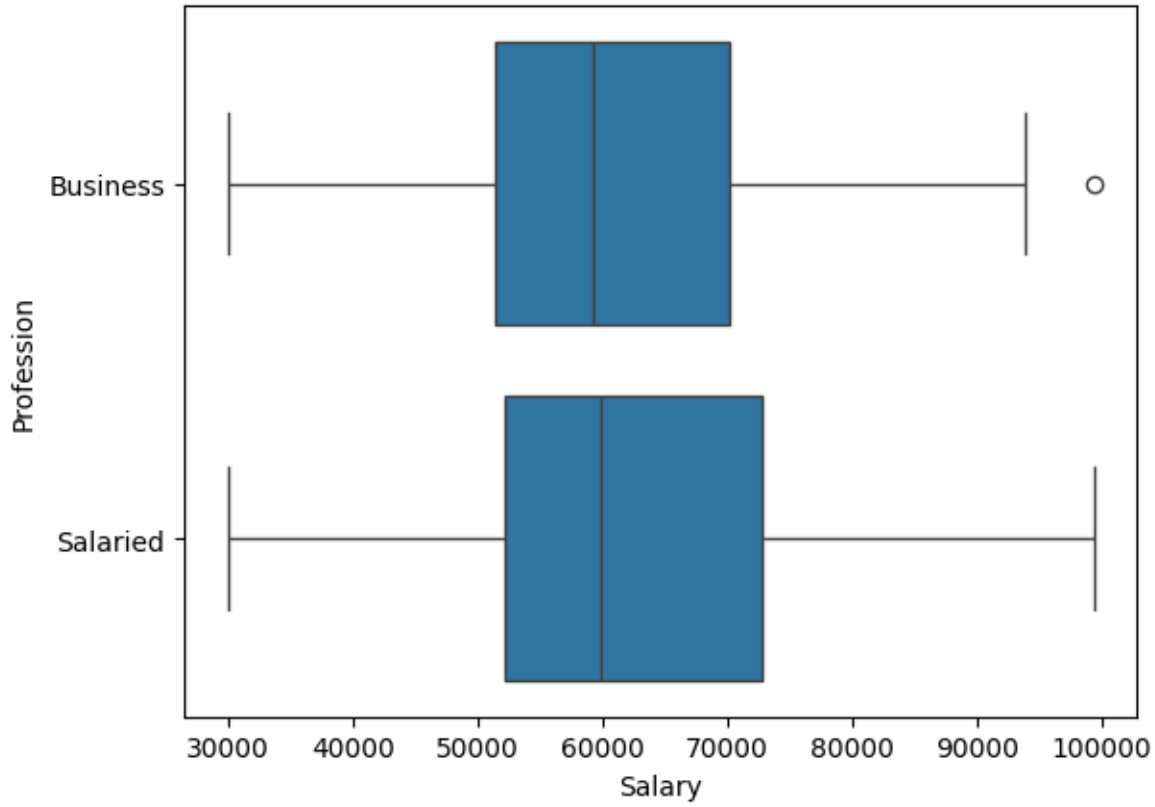
4.6 categorical variables Marital_status and Partner_working



Partner_working	No	Yes	All
Marital_status			
Married	0.363694	0.54902	0.912713
Single	0.087287	0.00000	0.087287
All	0.450980	0.54902	1.000000

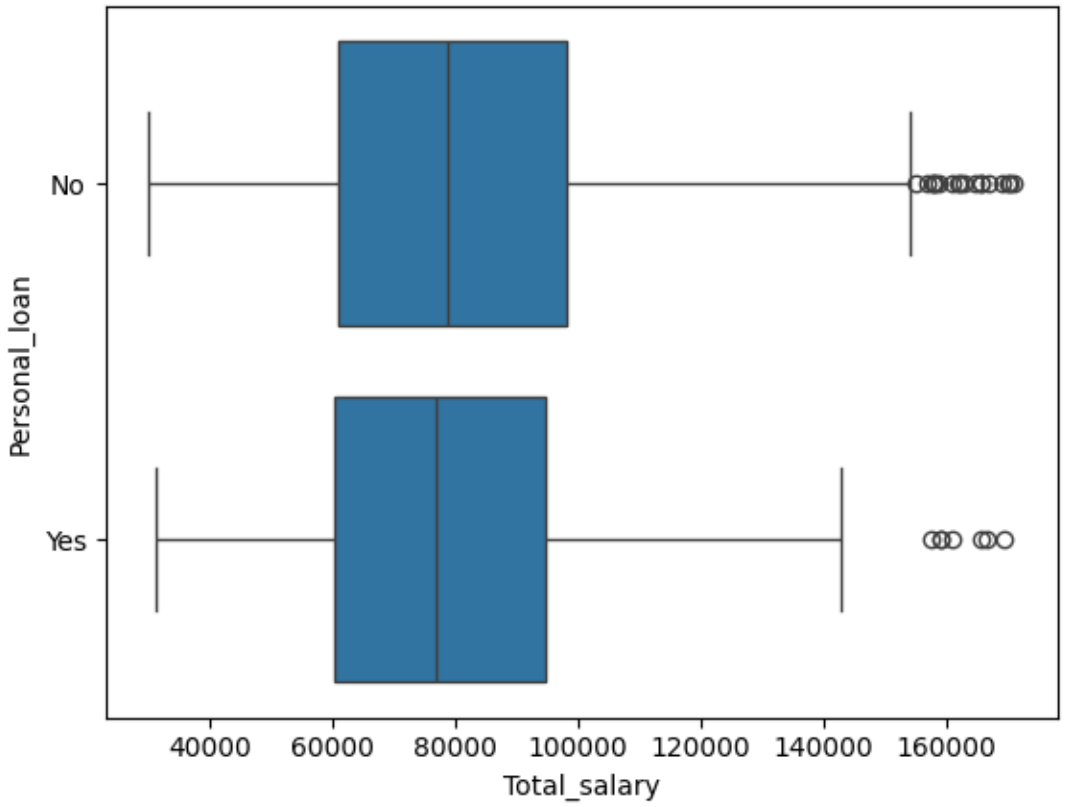
we can assume that 0.087% are single and 0.91% constitutes to married people which includes partner_working

4.7 Categorical & Numerical value Salary and Profession



So when we see the median salary of both business and salaried profession people it looks like same salary.

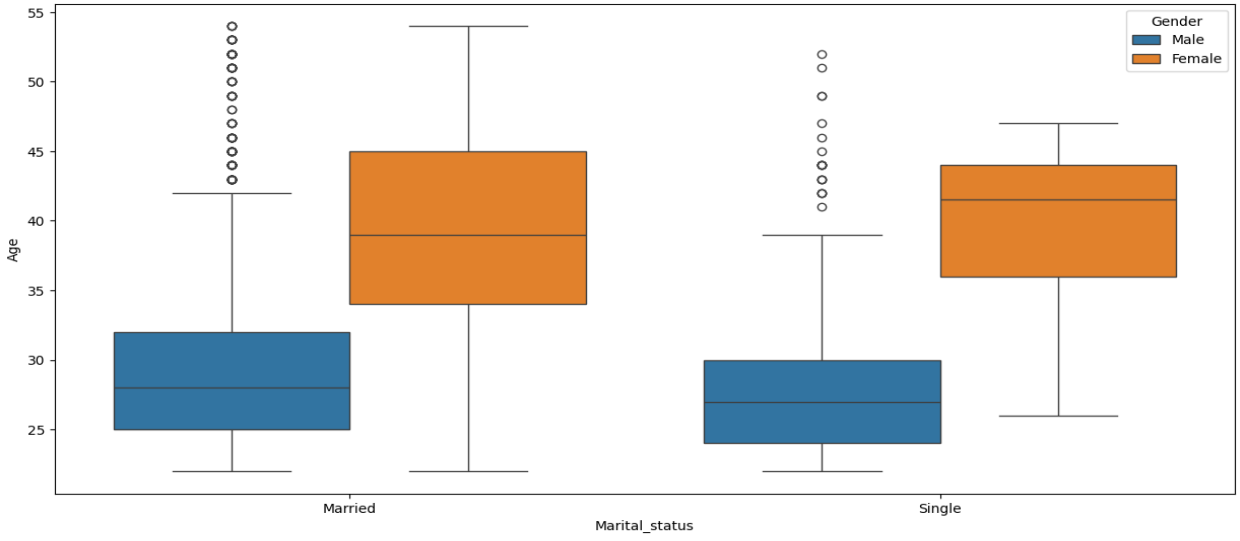
4.8 Total_salary and Personal_loan



Here, it's noticeable that the median value of "total_salary" for individuals who have taken a personal loan is slightly lower, but still quite close to the median value of "total_salary" for those who haven't opted for a personal loan.

5.1 Multivariate Analysis

Multivariate Analysis using Marital_status , Age , Gender



From the provided plot, it's apparent that among females, those who are single exhibit a higher median value in terms of marital status when compared to males.

5.2 For 2 or more variales using Facegrid

Hereby, we can overall see that most of the profession who are salaried has more number of eductaion status as post graduation and graduate when compared to business profession.

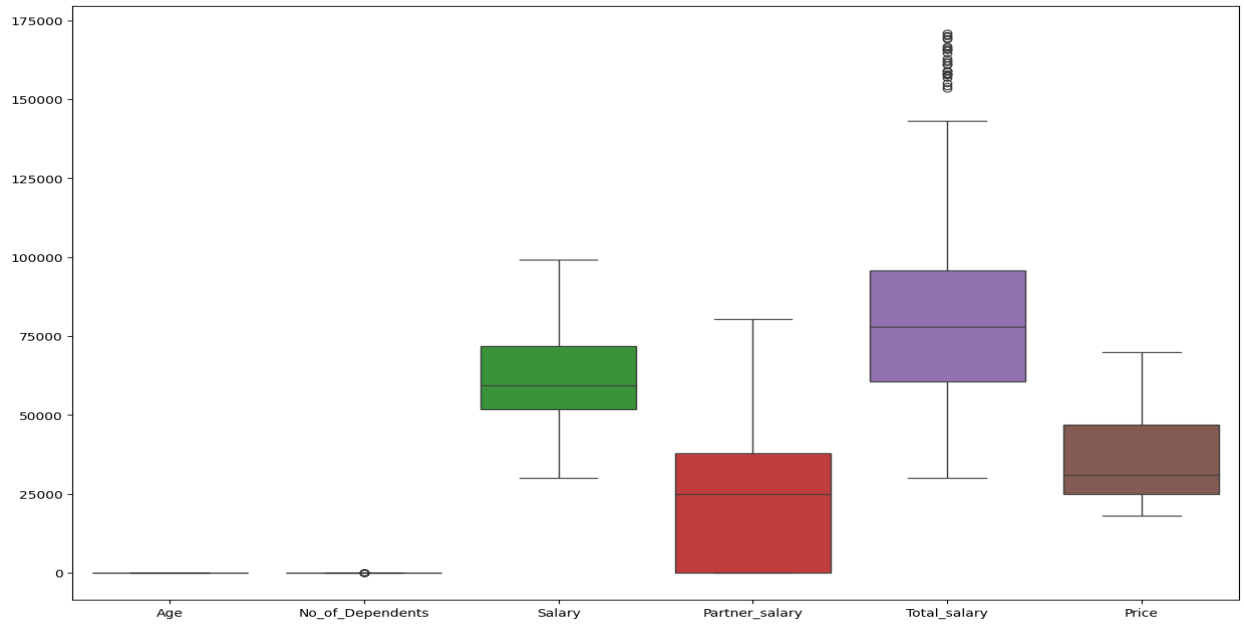
5.3 Skewness:

Let’s measure the skewness of the required columns

	Skewness
Age	0.892240
No_of_Dependents	-0.129685
Salary	-0.011560
Total_salary	0.609127
Price	0.740171
Partner_salary	0.348835

It appears that the distributions of "Age," "Total_salary," "Price," and "Partner_salary" exhibit moderately skewed characteristics, while "No_of_Dependents" and "Salary" appear to approximate symmetric distributions.

5.4 Checking for outliners

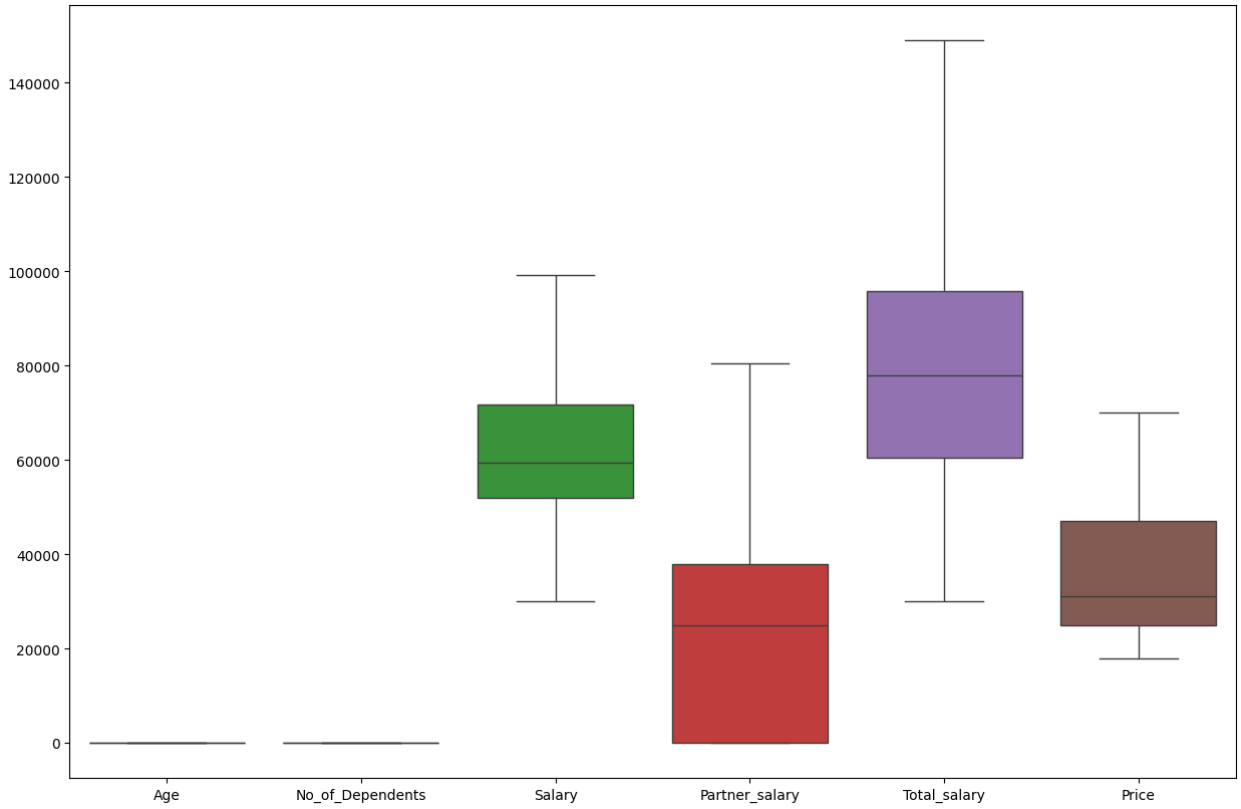


We observe numerous extreme outliers in "Total_salary" due to its highly skewed distribution. Conversely, only one outlier is evident in "No_of_Dependents," while no outliers are present in the other variables.

5.5 After Removing Outliners

As we had seen outliers in **Total_salary** and **No_of_dependents.** Let's remove it by replacing the outlier value using IQR.

So once outliers removed then we can see the following attributes does not contain outliers now.



5.6 Encoding

There are 2 types. But we shall use **Label Encoding** to see how it works.

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

Here, we see label as male and female so let's replace male and female with values 0 and 1.

```
0    1252
1     329
Name: Gender, dtype: int64
```

Now, we can see Male label has been assigned as value 0 and Female label has been assigned as 1.

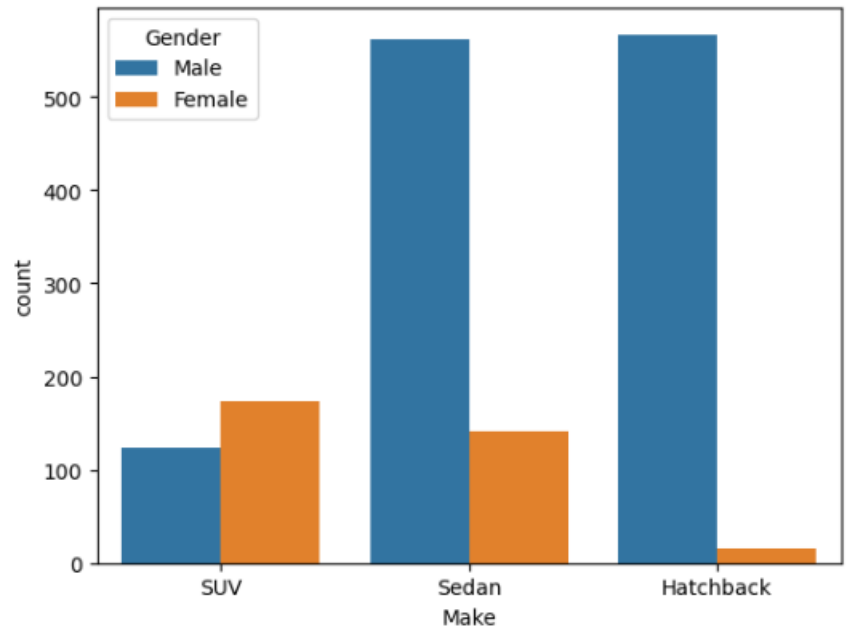
Questions

6. Explore the data to answer the following key questions

6.1 Do men tend to prefer SUVs more compared to women?

```
Gender  Make
Female  SUV          173
        Sedan        141
        Hatchback     15
Male    Hatchback    567
        Sedan        561
        SUV          124
Name: count, dtype: int64
```

```
<Axes: xlabel='Make', ylabel='count'>
```

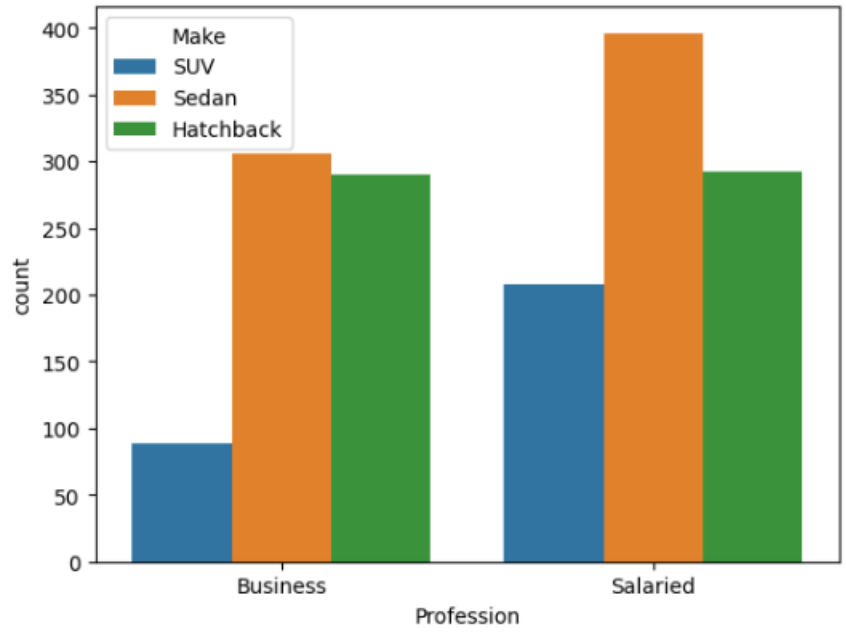


Based on the analysis provided, it seems that there are 173 females who prefer SUVs compared to 118 males. Therefore, it contradicts the statement that "Men prefer SUVs by a large margin compared to women," as the data suggests that more females prefer SUVs.

6.2 What is the likelihood of a salaried person buying a Sedan?

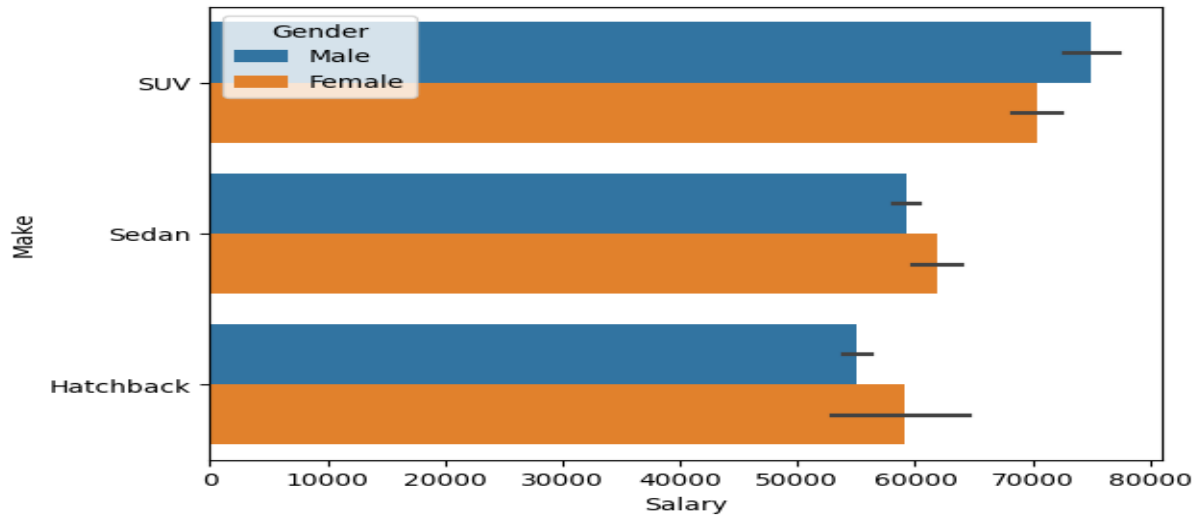
```
Profession  Make
Business    Sedan        306
            Hatchback    290
            SUV           89
Salaried    Sedan        396
            Hatchback    292
            SUV          208
Name: count, dtype: int64
```

```
<Axes: xlabel='Profession', ylabel='count'>
```



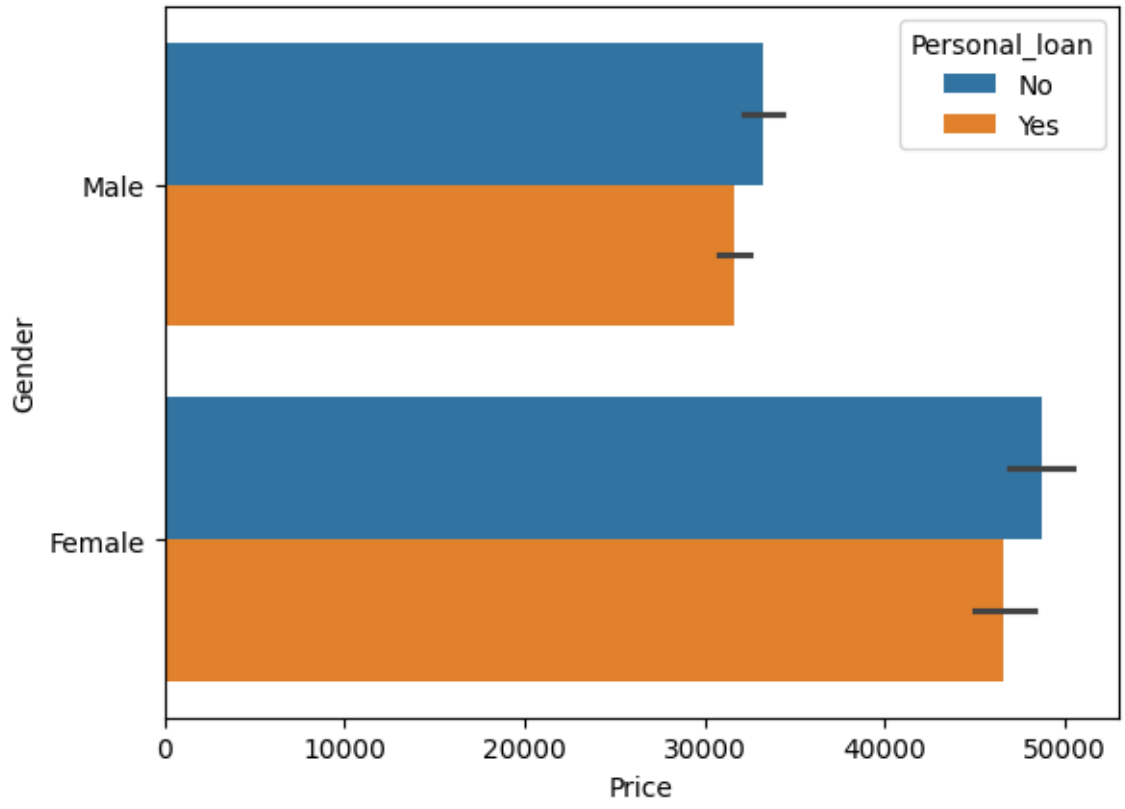
Indeed, the data indicates that there are 396 instances of salaried individuals purchasing sedans, while only 306 instances of individuals in the business profession buying sedans. Hence, we can agree with Ned Stark's statement that "A salaried person is more likely to buy a sedan.

6.3 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?



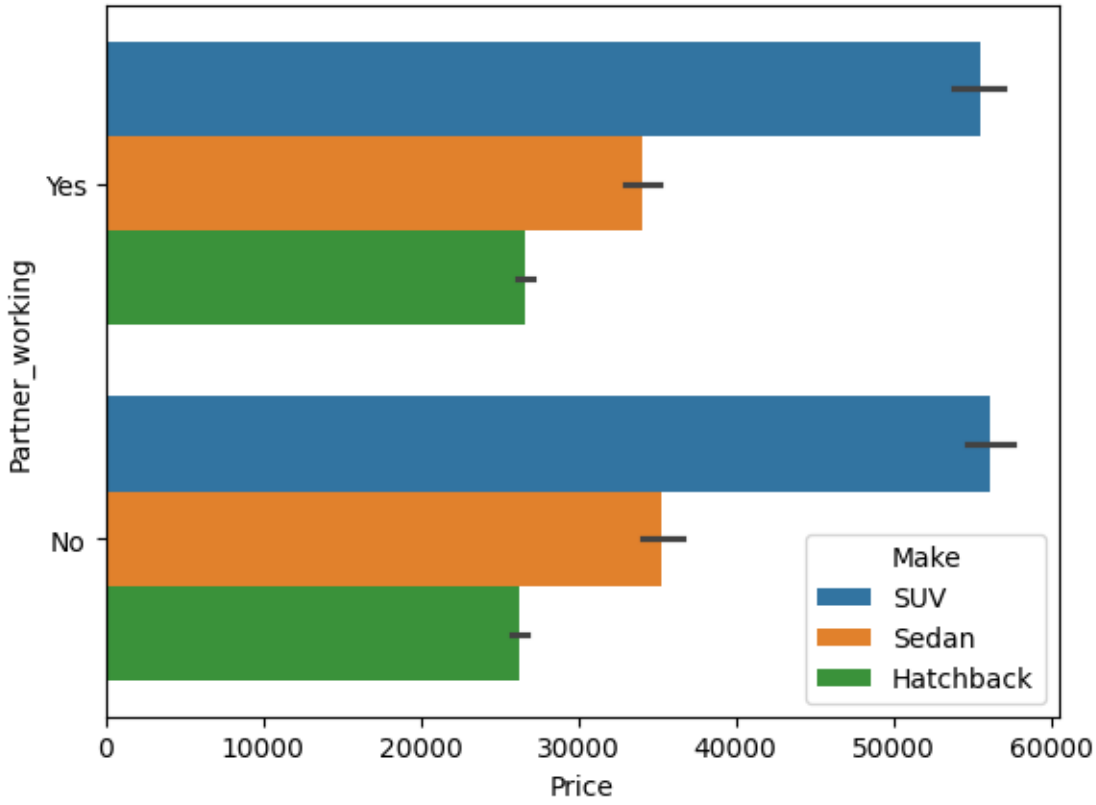
Based on the plot, it appears that the count of salaried males purchasing SUVs exceeds that of sedans, compared to salaried females. Therefore, it aligns with Sheldon Cooper's statement that "a salaried male is an easier target for an SUV sale over a sedan sale."

6.4 How does the the amount spent on purchasing automobiles vary by gender? & 6.5. How much money was spent on purchasing automobiles by individuals who took a personal loan?



Based on the plot, it appears that the count of salaried males purchasing SUVs exceeds that of sedans, compared to salaried females. Therefore, it aligns with Sheldon Cooper's statement that "a salaried male is an easier target for an SUV sale over a sedan sale."

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



Based on the analysis, it appears that individuals with non-working partners are slightly more inclined towards purchasing higher-priced cars compared to those with working partners. However, the difference between the two groups is not substantial, with individuals having working partners also showing a propensity for purchasing higher-priced cars. Therefore, it cannot be conclusively stated that having a working partner leads to the purchase of a higher-priced car, disproving the statement.

7.1 Actionable Insight:

Diversify Product Offerings: Develop and introduce new car models or variants that cater to the preferences of married females, especially those who are employed. Consider designing vehicles with features and specifications that appeal to this demographic segment.

Competitive Pricing Strategy: Conduct a thorough analysis of the features and pricing of the other two models in comparison to the SUV model preferred by married, employed females. Implement competitive pricing strategies to ensure that the new models offer value for money and are attractive alternatives to SUVs.

Feature Customization: Offer customization options for car features to cater to the diverse preferences and needs of married, employed females. Allow customers to personalize their vehicles according to their lifestyle, preferences, and budget constraints.

Tailored Marketing Campaigns: Develop targeted marketing campaigns specifically tailored to appeal to married, employed females. Highlight the unique features, benefits, and value propositions of the new car models, emphasizing how they meet the needs and preferences of this demographic group.

Enhanced Customer Experience: Focus on providing an exceptional customer experience throughout the purchasing journey, from pre-sales inquiries to post-sales support. Train sales representatives to understand the specific requirements of married, employed females and to provide personalized assistance and recommendations.

Continuous Monitoring and Adaptation: Continuously monitor customer feedback, sales performance, and market trends to identify areas for improvement and adaptation. Stay agile and responsive to changes in consumer preferences, technological advancements, and competitive dynamics.

By implementing these actionable insights, the company can effectively capitalize on the preferences and purchasing behavior of married, employed females, thereby expanding its market share and enhancing its competitive position in the automotive industry.

7.2 Business Recommendation:

Based on the comprehensive analysis, it is evident that married females, particularly those who are employed, are spending more on automobiles, particularly the SUV model. Therefore, it's crucial to assess the features and pricing of the other two models in comparison to the SUV model. By identifying and highlighting unique features and competitive pricing, we can potentially attract female customers and expand the market share beyond SUVs. Customization options and tailored marketing strategies can also be implemented to cater to the preferences and needs of different demographic groups within the female gender category.