# Bright Motor Company Data Analysis


## Index

# List of Tables



# List of Figures



## Problem Statement

Bright Motor Company 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 Dictionary

- **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.

## Importing the necessary Libraries

## Loading the Dataset

# Basic Exploration

### 1. - Displaying the top 5 rows

[7]:

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| **Age** | 53 | 53 | 53 | 53 | 53 |
| **Gender** | Male | Femal | Female | Female | Male |
| **Profession** | Business | Salaried | Salaried | Salaried | NaN |
| **Marital_status** | Married | Married | Married | Married | Married |
| **Education** | Post Graduate | Post Graduate | Post Graduate | Graduate | Post Graduate |
| **No_of_Dependents** | 4 | 4 | 3 | ? | 3 |
| **Personal_loan** | No | Yes | No | Yes | No |
| **House_loan** | No | No | No | No | No |
| **Partner_working** | Yes | Yes | Yes | Yes | Yes |
| **Salary** | 99300.0 | 95500.0 | 97300.0 | 72500.0 | 79700.0 |
| **Partner_salary** | 70700.0 | 70300.0 | 60700.0 | 70300.0 | 60200.0 |
| **Total_salary** | 170000 | 165800 | 158000 | 142800 | 139900 |
| **Price** | 61000 | 61000 | 57000 | 61000 | 57000 |
| **Make** | SUV | SUV | SUV | ? | SUV |

- Table 1: Top 5 rows

# Observations

### From the head we can infer that:

- "No_of_Dependents", "Make" columns might be having wrong entries like "?"
- There are null values in the dataset

## 2. - Displaying the last 5 Rows

| | 1576 | 1577 | 1578 | 1579 | 1580 |
|---|---|---|---|---|---|
| **Age** | 22 | 22 | 22 | 22 | 22 |
| **Gender** | Male | Male | Male | Male | Male |
| **Profession** | Salaried | Business | Business | Business | Salaried |
| **Marital_status** | Single | Married | Single | Married | Married |
| **Education** | Graduate | Graduate | Graduate | Graduate | Graduate |
| **No_of_Dependents** | 2 | 4 | 2 | 3 | 4 |
| **Personal_loan** | No | No | No | Yes | No |
| **House_loan** | Yes | No | Yes | Yes | No |
| **Partner_working** | No | No | No | No | No |
| **Salary** | 33300.0 | 32000.0 | 32900.0 | 32200.0 | 31600.0 |
| **Partner_salary** | 0.0 | NaN | 0.0 | NaN | 0.0 |
| **Total_salary** | 33300 | 32000 | 32900 | 32200 | 31600 |
| **Price** | 27000 | 31000 | 30000 | 24000 | 31000 |
| **Make** | Hatchback | Hatchback | Hatchback | Hatchback | Hatchback |

- Table 2: Bottom 5 Rows

## 3. - Getting the shape of the dataset i.e to get the number of rows and columns

## Observations

### We can see that in our dataset:

- There are 1581 rows.
- There are 14 columns.

## 4. - Check the datatypes of each feature.

```
<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        1575 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            1568 non-null   float64
 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(2), int64(3), object(9)
memory usage: 173.1+ KB
```

- Figure 1: displaying the datatypes of columns

## Observations

### We can see that in our dataset we have:

- 3 columns which are of Int datatype.
- 2 columns of float datatype.
- Remaining 9 columns are of object datatype.type.
- No_of_Dependents column should be numerical datatype but a as object type data which is incorrect and needs to be corrected.ppearing as object

## 5. - Check the Statistical summary

| | Age | Salary | Partner_salary | Total_salary | Price |
|---|---|---|---|---|---|
| **count** | 1581.000000 | 1568.000000 | 1475.000000 | 1581.000000 | 1581.000000 |
| **mean** | 31.952562 | 60276.913265 | 20225.559322 | 79625.996205 | 35948.170778 |
| **std** | 8.712549 | 14636.200199 | 19573.149277 | 25545.857768 | 21175.212108 |
| **min** | 14.000000 | 30000.000000 | 0.000000 | 30000.000000 | 58.000000 |
| **25%** | 25.000000 | 51900.000000 | 0.000000 | 60500.000000 | 25000.000000 |
| **50%** | 29.000000 | 59450.000000 | 25600.000000 | 78000.000000 | 31000.000000 |
| **75%** | 38.000000 | 71700.000000 | 38300.000000 | 95900.000000 | 47000.000000 |
| **max** | 120.000000 | 99300.000000 | 80500.000000 | 171000.000000 | 680000.000000 |

- Table 3: Displaying statistical summary

### Observations

- The maximum age is given as 120
- the minimum age is given as 14 this needed to be checked as a 14 year cannot drive a car

# Data cleaning

### 6. - Check the null values

```
Age                  0
Gender              53
Profession           6
Marital_status       0
Education            0
No_of_Dependents     0
Personal_loan        0
House_loan           0
Partner_working      0
Salary              13
Partner_salary     106
Total_salary         0
Price                0
Make                 0
```

- Table 4: Displaying the null values

## Observations

### We can see that in our dataset:

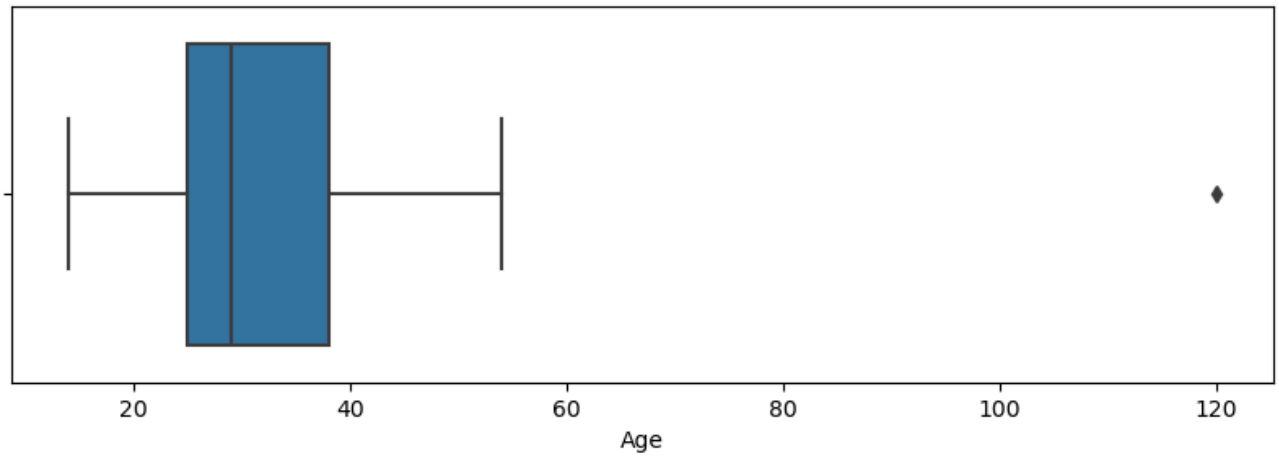
- only 4 columns (Gender, Profession, Salary, Partner_salary) have null values.
- Gender have 53 null values.
- Profession have 6 null values.
- Salary have 13 null values.
- Partner_salary have 106 null values.

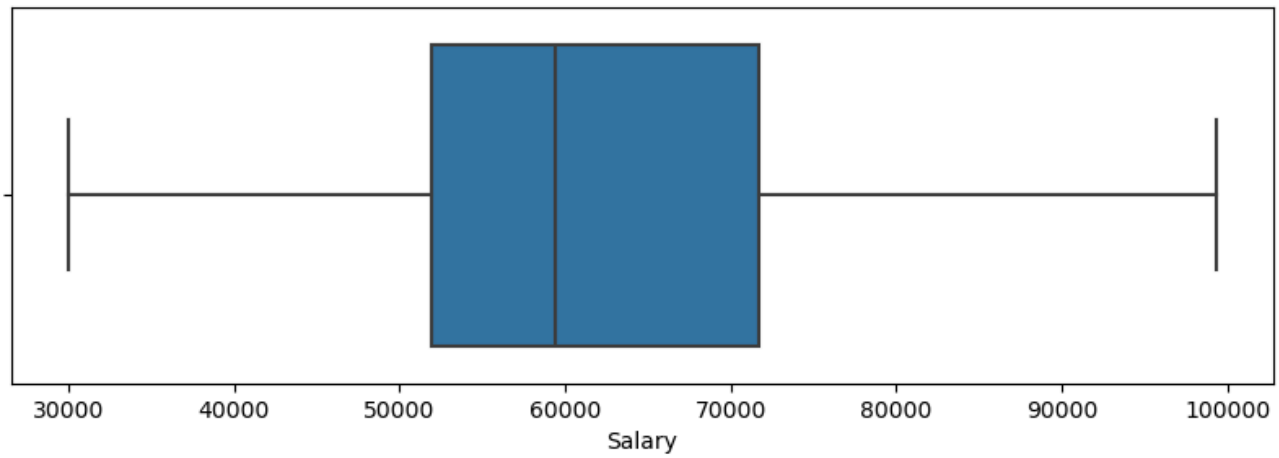
## 7. - Check the duplicate values

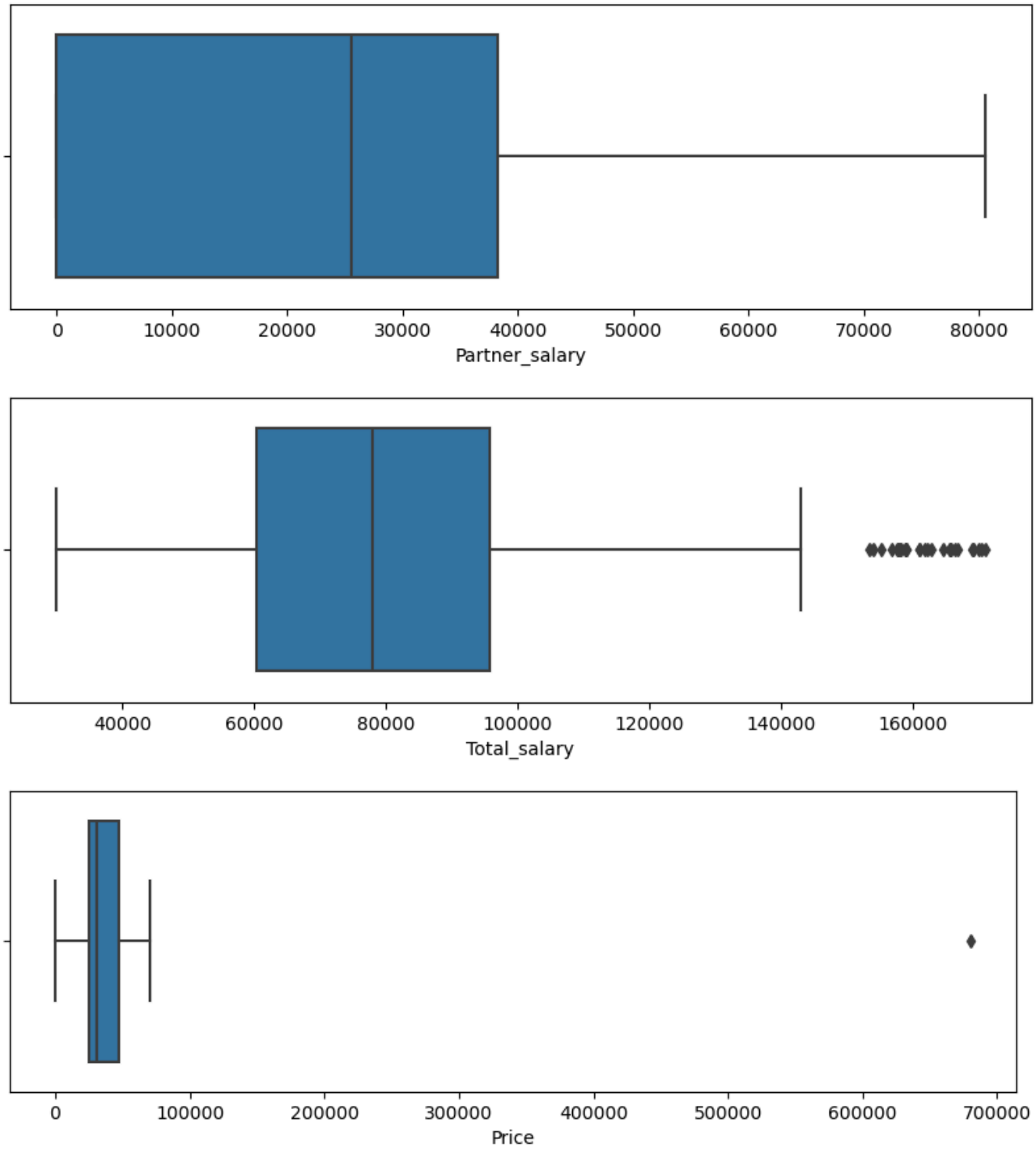
### We can see that all the rows are unique and we have no duplicated rows

## 8. - Check the outliers and their authenticity.

In [9]:






- Figure 2: Displaying the outliers

## Observations

### we can see that:

- Column 'Age','Total Salary and 'Price' has outliers which needed to be seen

### we can see that in the Total_salary column we have outliers but they are correct and genuine values as the outliers are showing when the total salary includes the partner working as 'Yes', so when the partner work the total salary will be more than when partner is not workings o they need not to be replaced

## 9. - Removing Anomalies

## Observation:

## In 2 columns we have anomalies as they contain invalid values:

- In No_of_Dependents we have '?' as some value which are wrong and needed to be replaced
  - The '?' values replaced with nan value are then replaced with the median of the column.
- In Make column we have '?' as some value which are wrong and needed to be replaced.
  - The '?' values replaced with nan values which are replaced with the mode of column.

Removing the data where age is 14 as it is an incorrect value as a 14 year old cannot use an automobile.

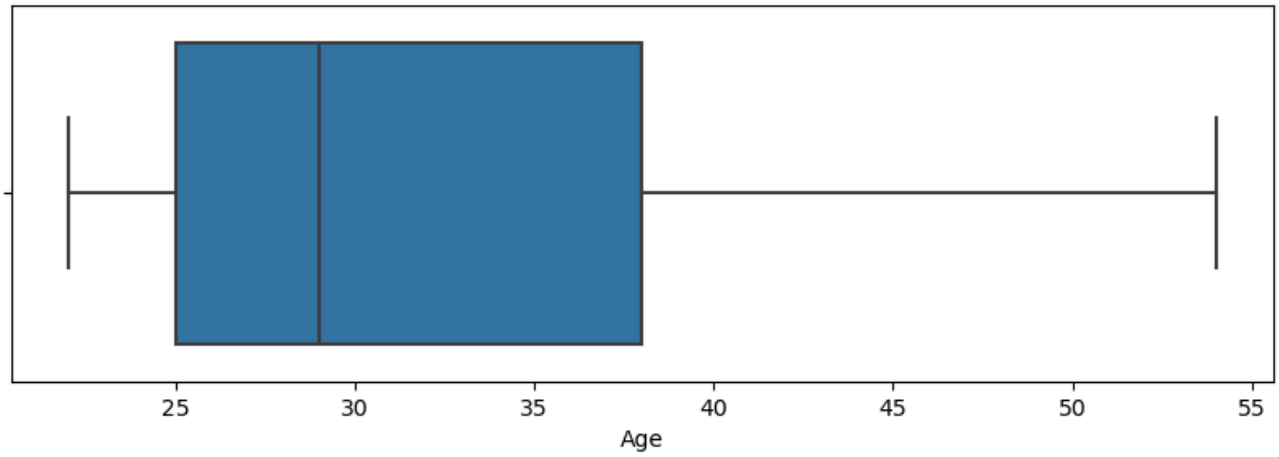
Removing the data where age is 120 as it appears to be a anomaly.

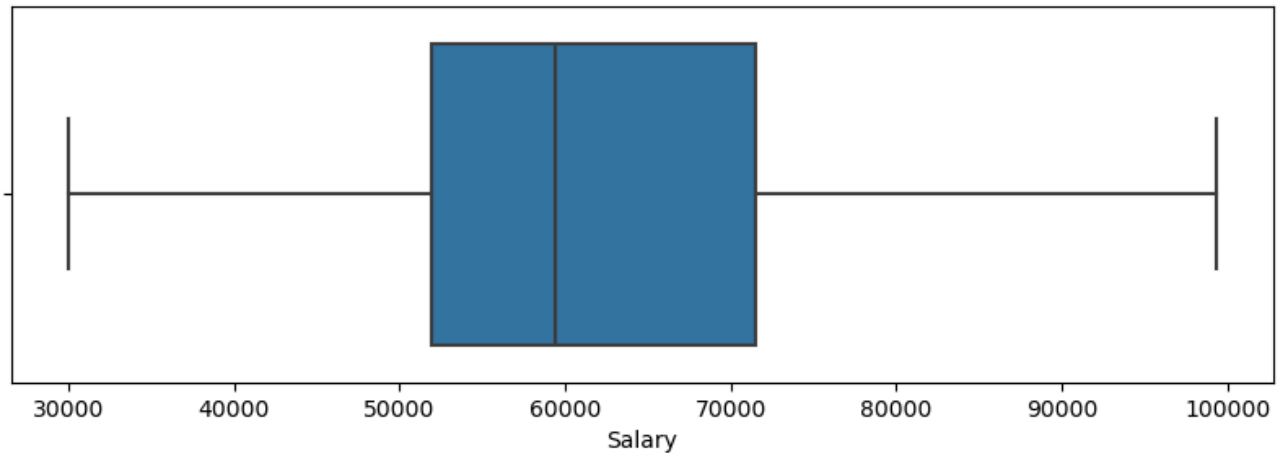
## 10. - Null value imputation and outlier correction

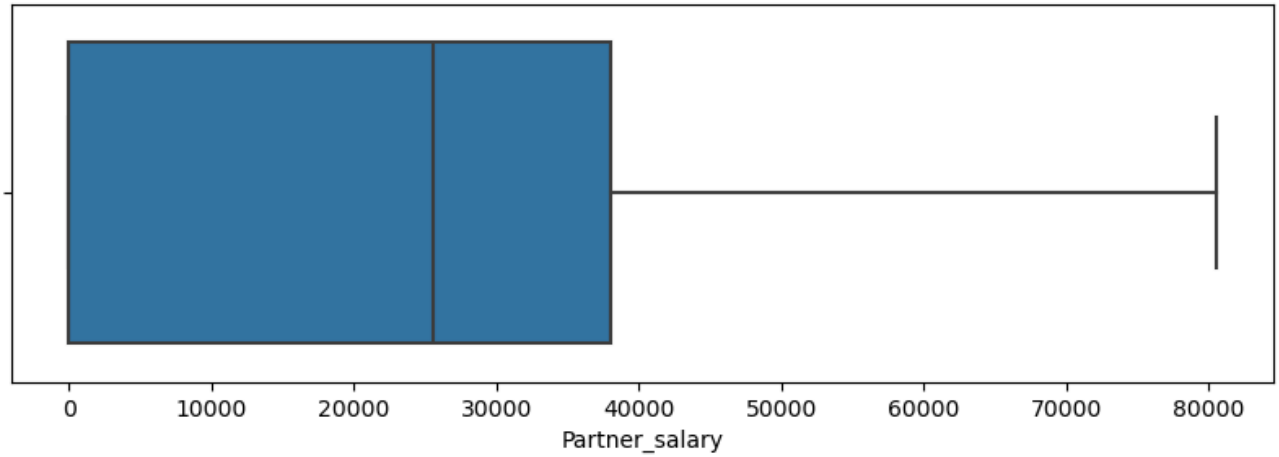
| | |
|---|---|
| Age | 0 |
| Gender | 0 |
| Profession | 0 |
| Marital_status | 0 |
| Education | 0 |
| No_of_Dependents | 0 |
| Personal_loan | 0 |
| House_loan | 0 |
| Partner_working | 0 |
| Salary | 0 |
| Partner_salary | 0 |
| Total_salary | 0 |
| Price | 0 |
| Make | 0 |

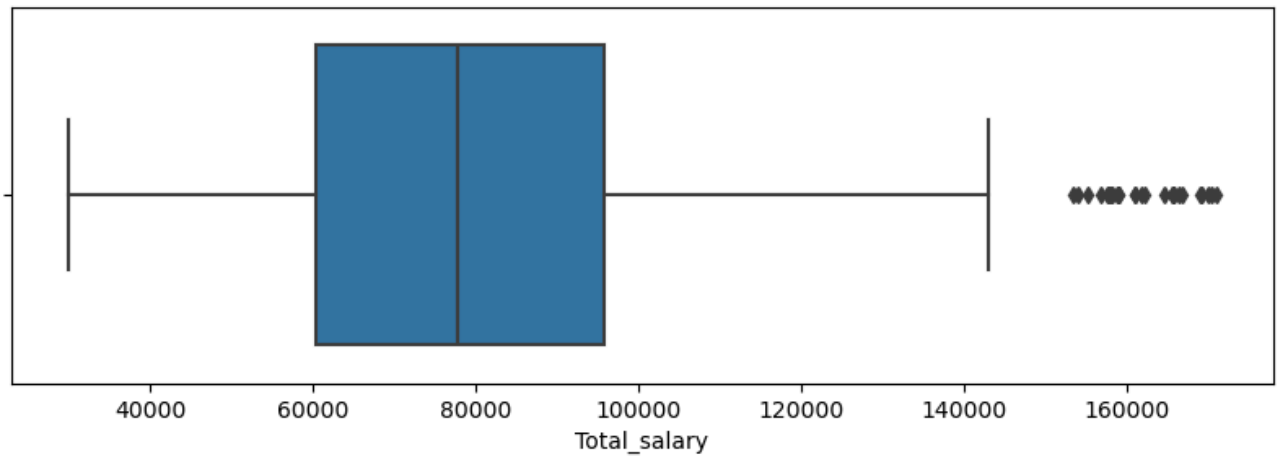
- Table 5: New table displaying no null value

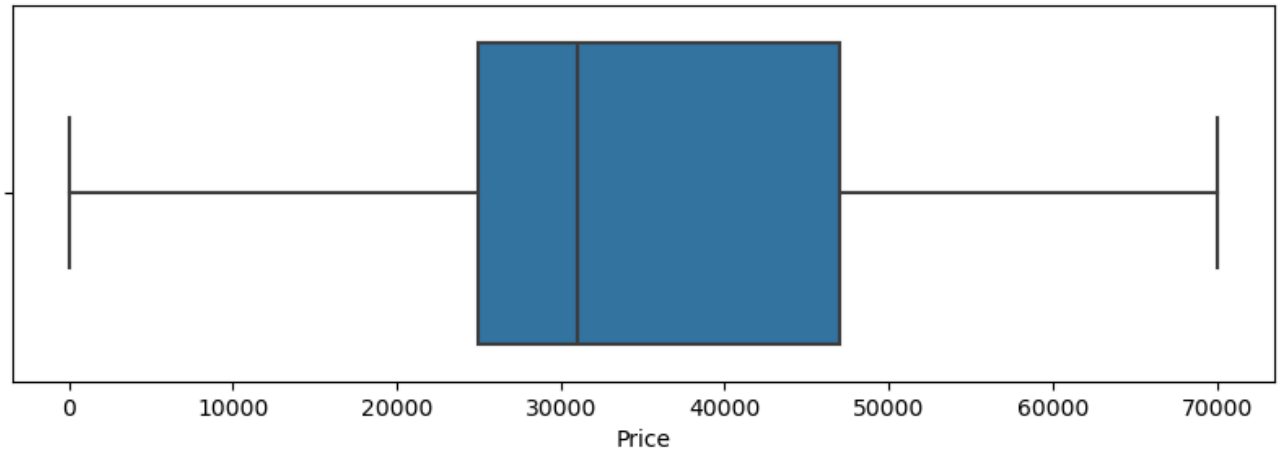
In [41]:


25
30
35
40
45
50
55
Age


30000
40000
50000
60000
70000
80000
90000
100000
Salary


0
10000
20000
30000
40000
50000
60000
70000
80000
Partner_salary


40000
60000
80000
100000
120000
140000
160000
Total_salary




- Figure 3: Figure displaying no outliers present

### Observations

All the Null values are now removed.

We can see that now all the Outliers have been removed

# Questions

## 1. Descriptive Statistics:

- What are the mean, median, and standard deviation of the ages of individuals in the dataset?

### Observation

We can infer that:

- Mean Age is: 31.9 years
- Median Age is: 29 years
- Standard Deviation of Age is: 8.4 years

## 2. Data Distribution:

- What is the distribution of gender in the dataset? Represent it using a pie chart.

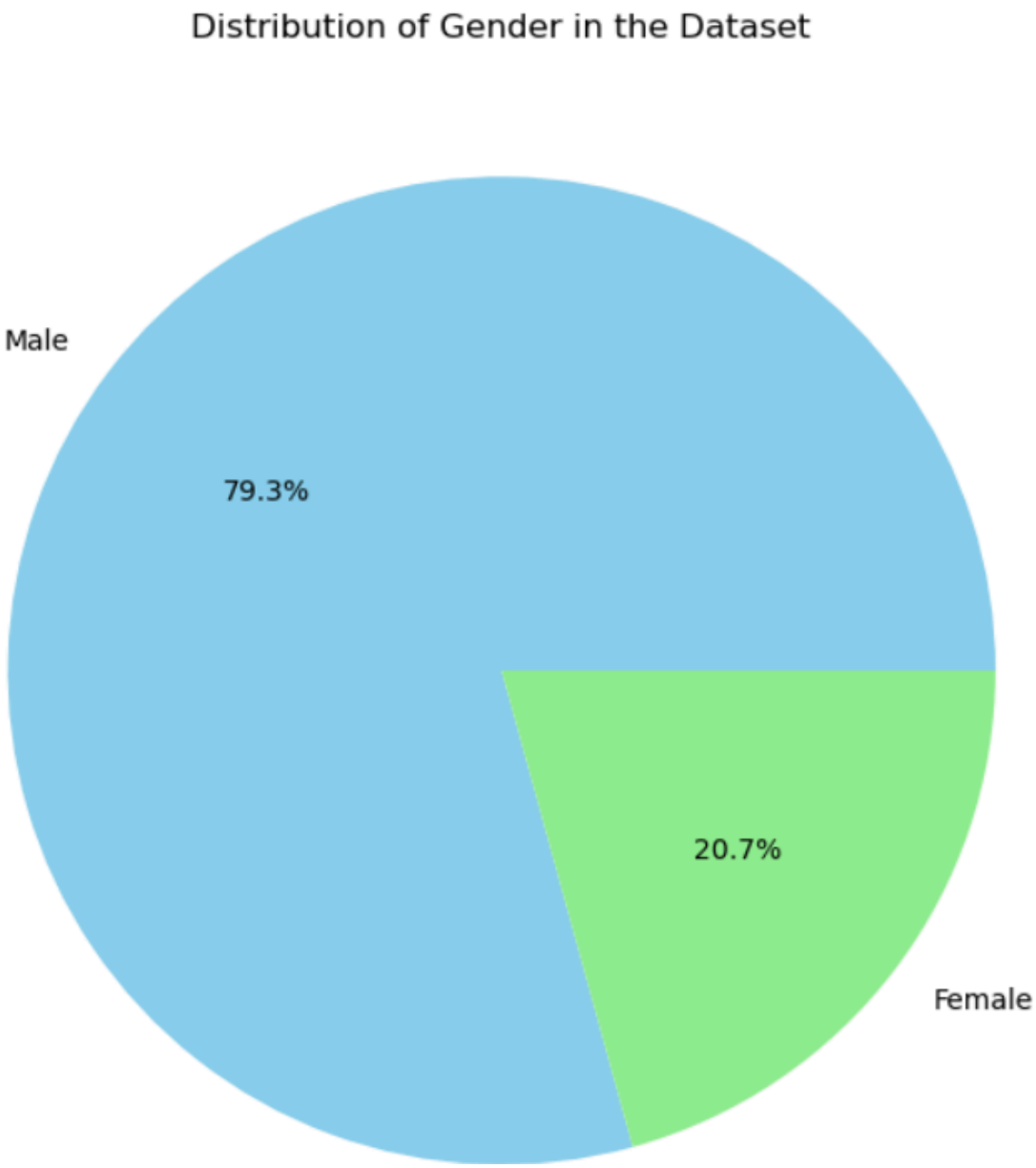
### Obseravtion

We can see that we have 2 enteries as:

- Femal

- Femle

These are wrong entries and needed to be corrected




- Figure 4: Distribution of Gender

## Observation

From the Pie Chart we can infer that:

- There are 79.2% Male.
- There are 20.8% Female.

# 3. Correlation Analysis:

- Is there a correlation between age and salary? Provide the correlation coefficient and interpret the result.

```
Correlation coefficient between Age and Salary: 0.6007704272305254
There is a positive linear relationship between Age and Salary.
```

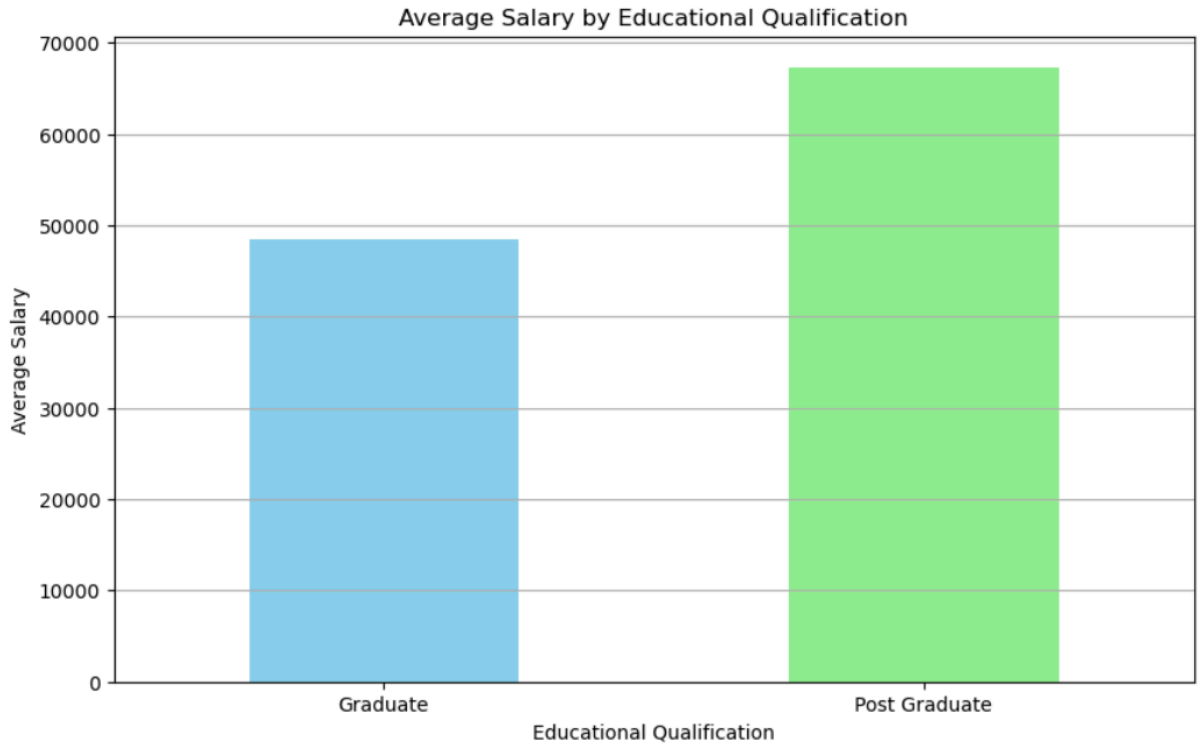
## Observation

### From the data we can infer that:

- The correlation between Age and Salary is coming out to be - 0.591
- There is a strong positive correlation, meaning as age increases, salary tends to increase.

# 4. Salary Analysis:

- ## What is the average salary for individuals based on their educational qualifications (Graduate vs. Post Graduate)?




- Figure 5: Average salary by education qualification

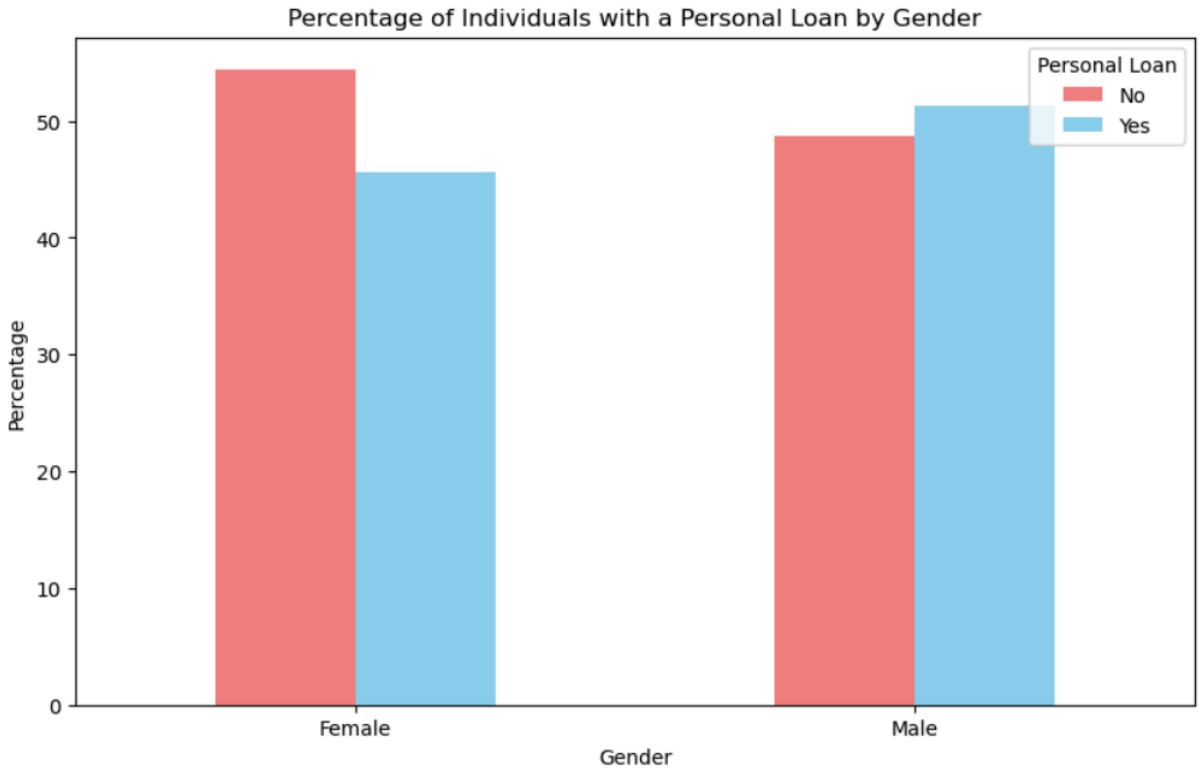
## Obseravtion

### From this we can infer that:

- Average Salary for Graduate person is: Rs 48514.59
- Average Salary for Post Graduate person is: Rs 67383.09
- The person with Post Graduate qualification have more average salary as compared to to the person with Graduate qualification

# 5. Loan Status:

- What percentage of individuals have taken a personal loan? How does this compare between males and females?




- Figure 6: percentage of individuals with personal loan by gender

### Observation

We can infer that The percentage of population with a personal is approx 50%

In case of Male and Female :

- 45% of Females have a personal loan
- 51% of Men have a personal loan

## 6. Marital Status and Dependents:

- What is the average number of dependents for married individuals versus single individuals?

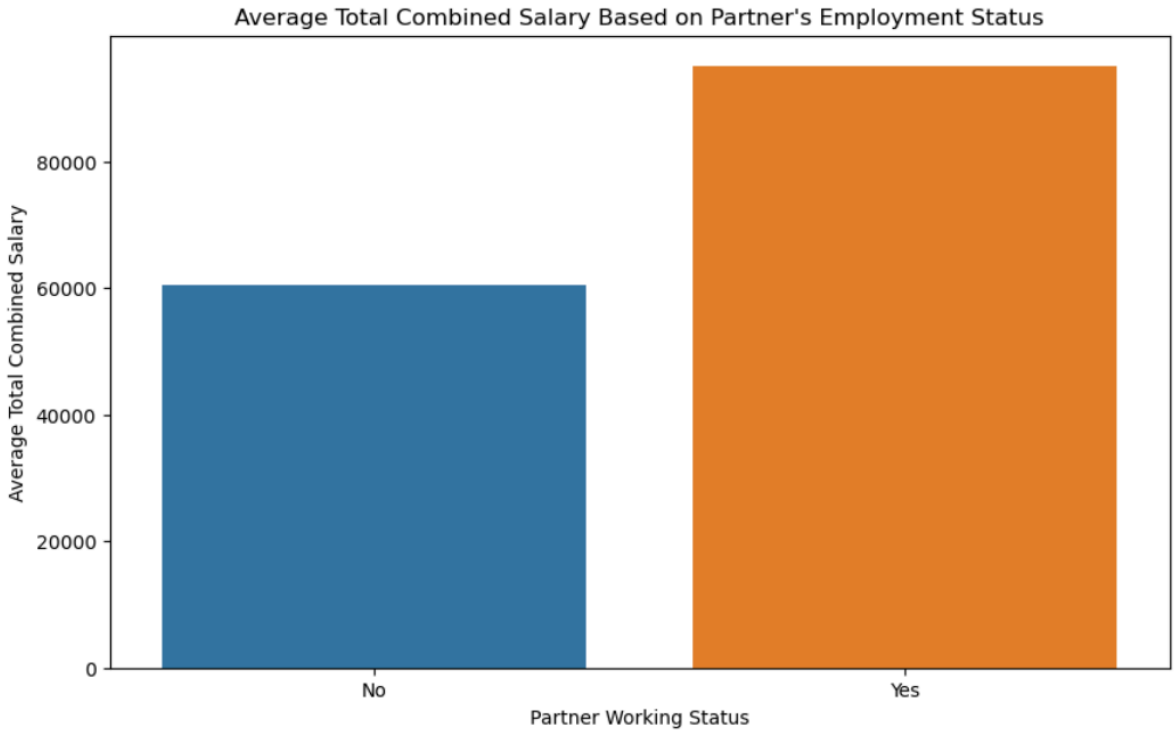
### Observation

we can see that the average number of dependents:

- for married dependents - 3
- for single individuals - 2

## 7. Partner Employment:

- ## How does the employment status of a partner affect the total combined salary?




- Figure 7: Average total combined salary based on partner employement

### Observation

We can infer that the average total combined salary on the basis of partner employement sataus is:

- for those whose partner are not working: Rs 60527.20
- for those whose partner are working: Rs 95314.28
- This means that if both the partners are working the average combined salary is more

# 8. Salary Comparison:

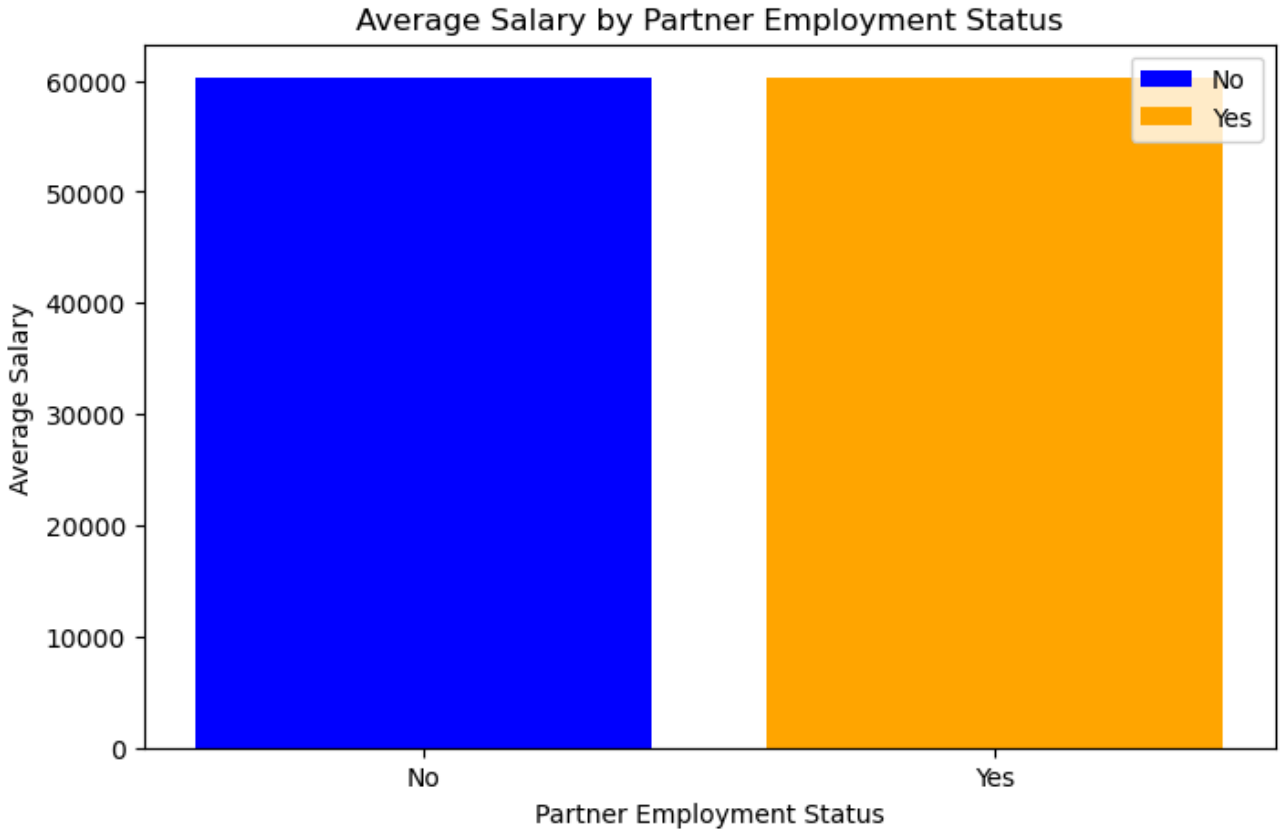
- ## Compare the average salary of individuals whose partners are working versus those whose partners are not working.

In [55]:
```
average_salary_partner_working = data.groupby('Partner_working')['Salary'].mean()

print("Average salary based on partner's employment status:")
print(average_salary_partner_working)

plt.figure(figsize=(8, 5))
bars = plt.bar(average_salary_partner_working.index, average_salary_partner_working.value
plt.xlabel('Partner Employment Status')
plt.ylabel('Average Salary')
plt.title('Average Salary by Partner Employment Status')
plt.legend(bars, average_salary_partner_working.index)
plt.show()
```

```
Average salary based on partner's employment status:
Partner_working
No     60255.539972
Yes    60195.491329
Name: Salary, dtype: float64
```




- Figure 8: Average salary by partner employement

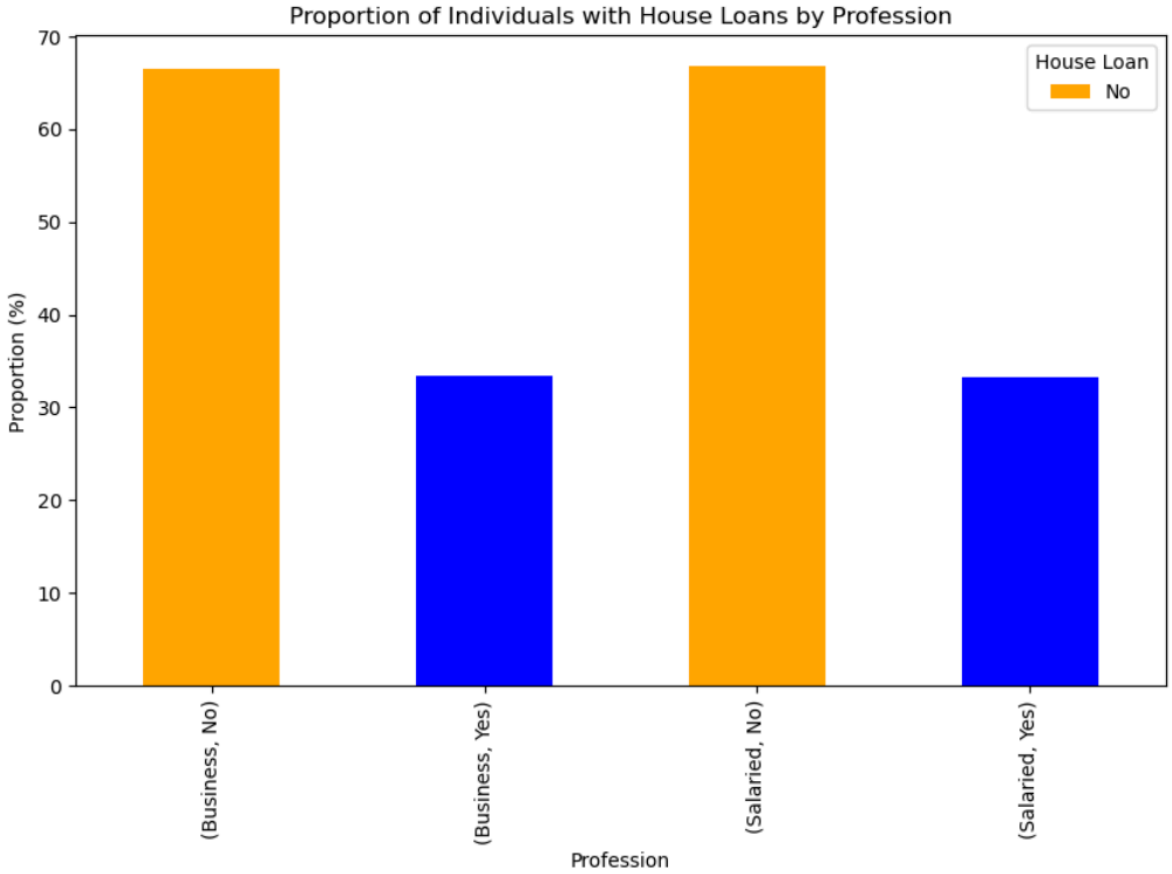
## Observation

- Average salary when:
  - when partner works is Rs. 60195.49
  - when partner does not work is Rs. 60255.53

This shoes that there is not much difference in the average salary on the baiss of partner employement status.

# 9. House Loan Analysis:

- What is the proportion of individuals with house loans based on their profession?




- Figure 9: House Loan proportion by profession

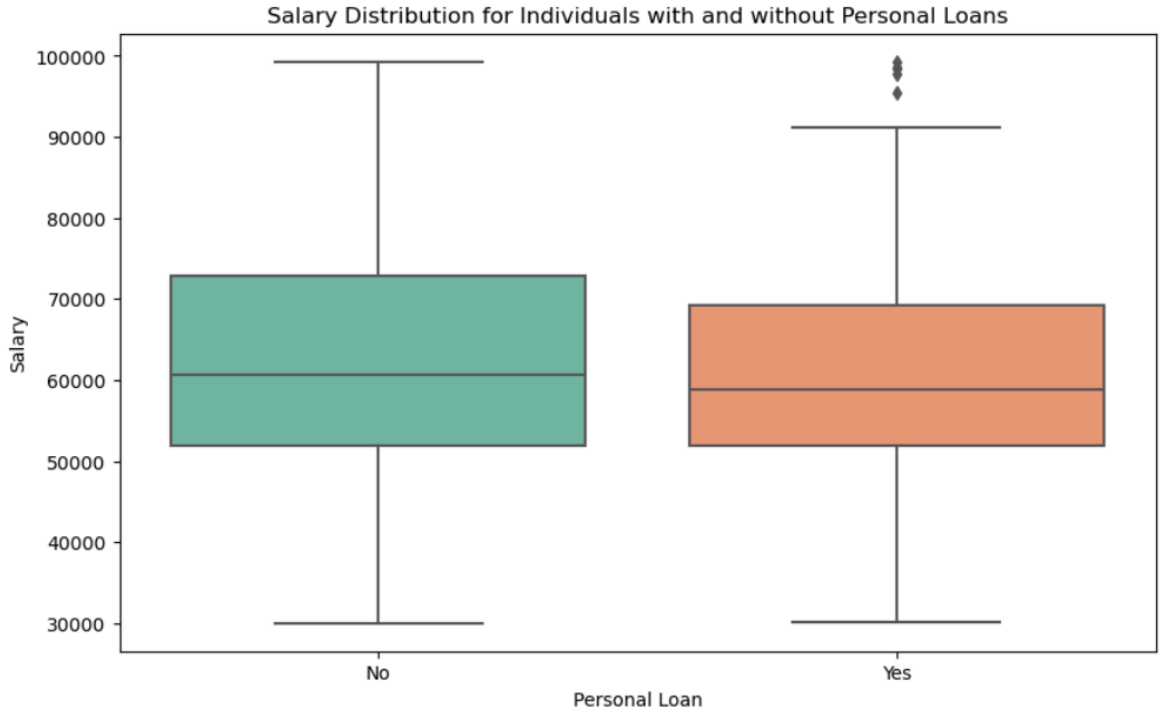
## Observation

### We can infer that:

- There are 33.43% people with Business as their profession who have house loans.
- There are 33.25% people with Salaried as their profession who have house loans.

# 10. Salary Distribution:

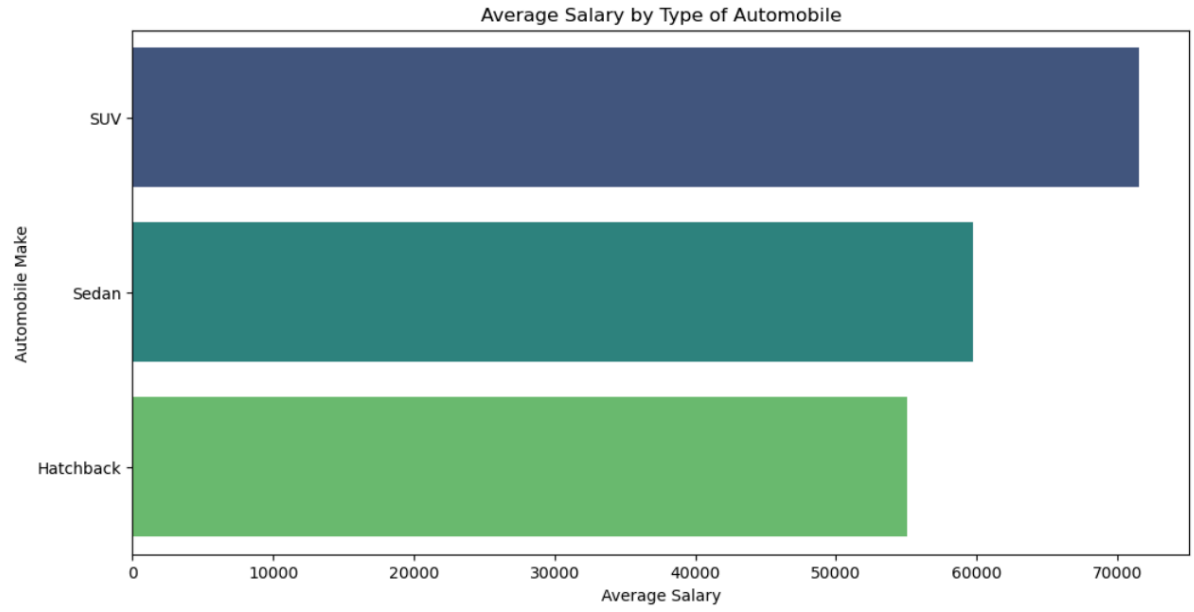
- **What is the distribution of salaries for individuals with personal loans versus those without personal loans? Represent it using a box plot.**




- Figure 10: Salary distribution by personal loans

# 11. Automobile Make Analysis:

## How does the type of automobile relate to the salary of the individuals? Provide insights based on the make of the automobile.




- Figure 11: Average salary by Automobile make

## Observation

### We can see that:

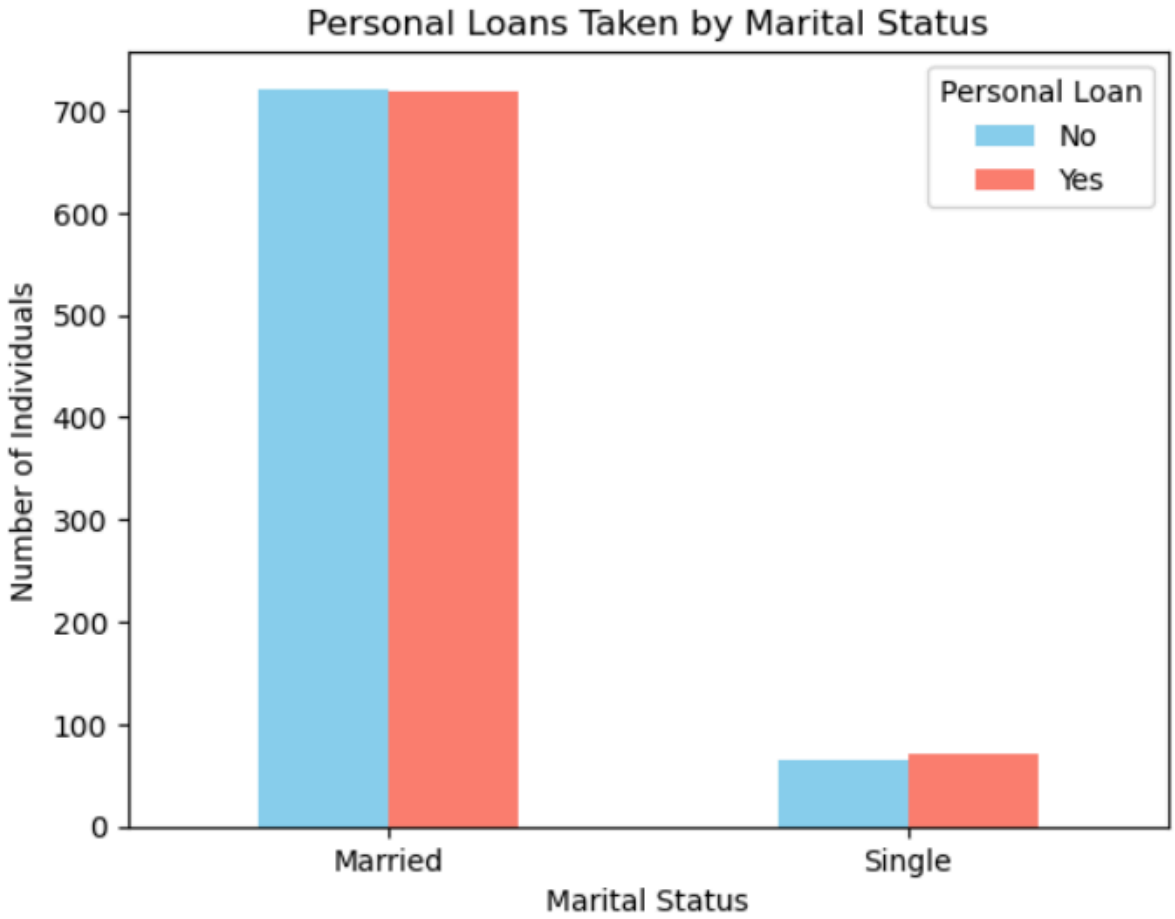
- The make of the Automobile is 'SUV' when the average salary is Rs. 71533.76
- The make of the Automobile is 'SEDAN' when the average salary is Rs. 59762.87
- The make of the Automobile is 'HATCHBACK' when the average salary is Rs. 55083.50

When the average salary is larger the automobile make is also the costlier one

As the salary decreases the make of automobile becomes the cheaper one

# 12. Marital Status and Loans:

## Is there a significant difference in the number of personal loans taken by married individuals compared to single individuals?




- Figure 12: Personal loan by marital status

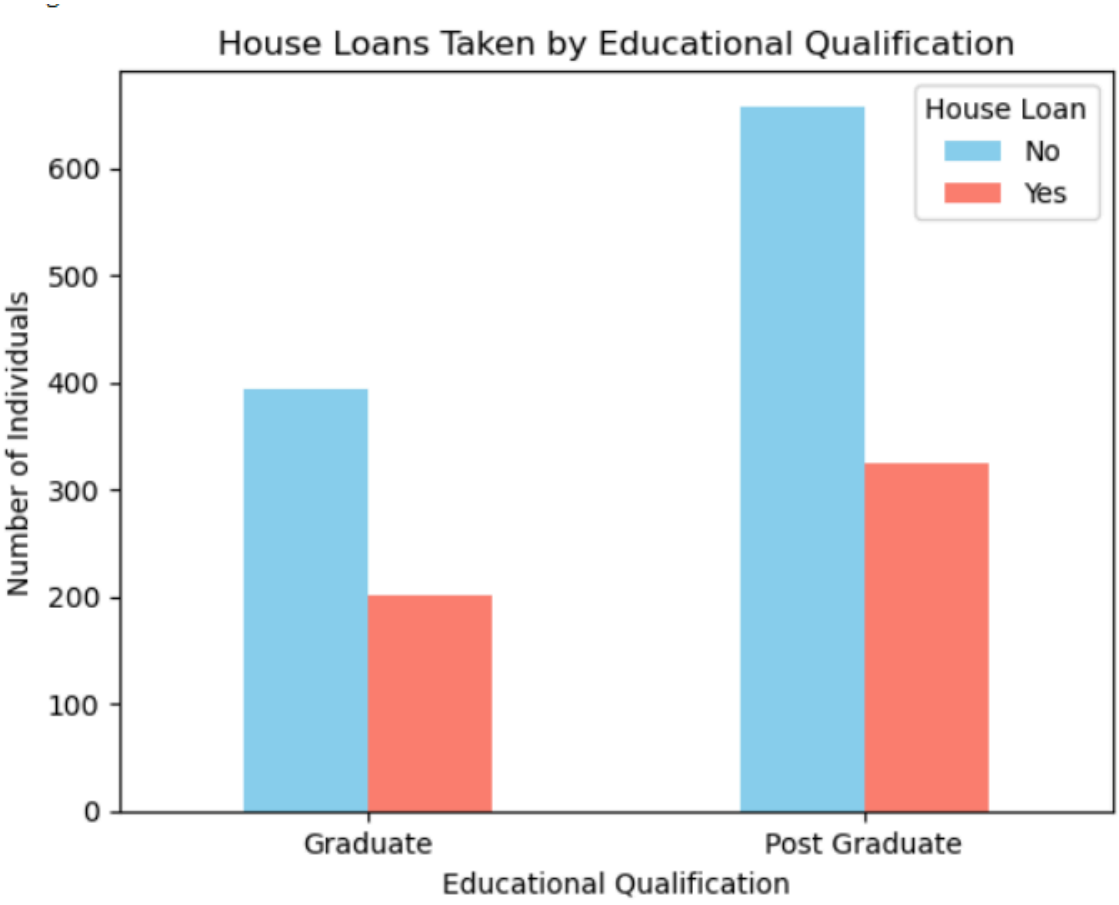
## Observation

### We can infer that:

- In case of married status there are 720 individuals who have personal loan and 723 individuals who do not have personal loan.
- In case of single status there are 72 individuals who have personal loan and 66 individuals who do not have personal loan.

# 13. Educational Qualification Impact:

## How does educational qualification impact the likelihood of taking a house loan?




- Figure 13: House Loan by Educational qualification

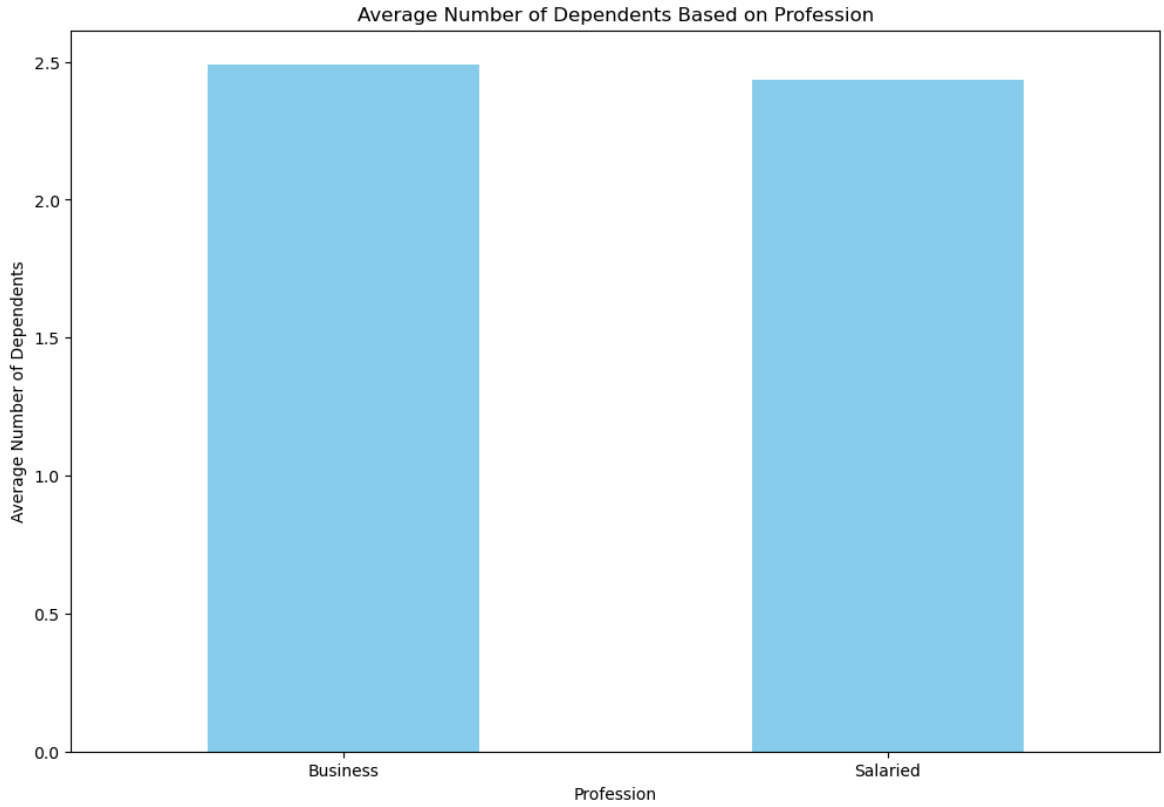
## Obseravtion

### From this we can infer that:

- In case of Graduate qualification the individuals with house loan are 202
- In case of Post Graduate qualification the individuals with house loan are 324
- This shows that individuals with Post Graduate Education take more house loan as compared to the individuals with Graduate education qualification

# 14. Dependent Count Analysis:

## Analyze the number of dependents based on the profession of the individual. Which profession has the highest average number of dependents?




- Figure 14: Average dependents by profession

## Observation

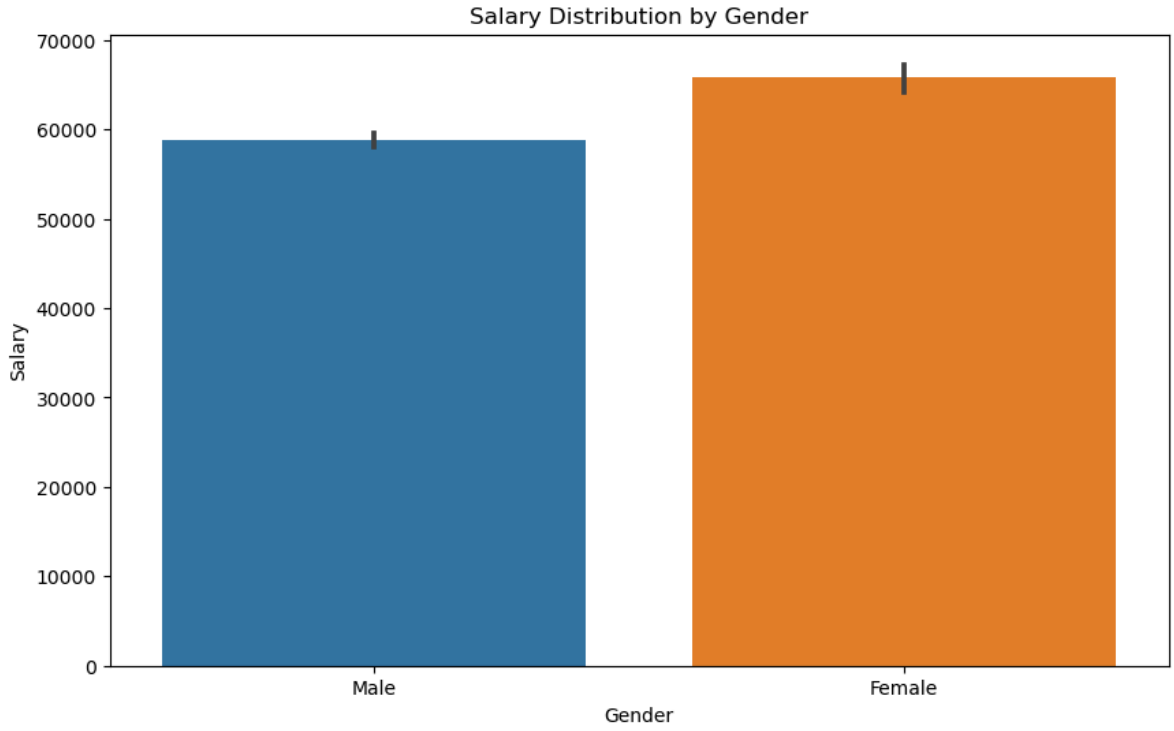
Individuals with Business profession has the highest number of dependents

# 15. Gender and Salary:

## Is there a significant difference in salaries between males and females? Provide statistical evidence.

```
Descriptive statistics of salaries based on gender:
         count          mean           std      min      25%      50%  \
Gender
Female   327.0  65816.513761  14380.357985  34800.0  56700.0  63900.0
Male    1251.0  58760.431655  14239.046917  30000.0  51200.0  58500.0

            75%      max
Gender
Female  77300.0  99300.0
Male    68950.0  99300.0
```




- Figure 15: Salary distribution by gender

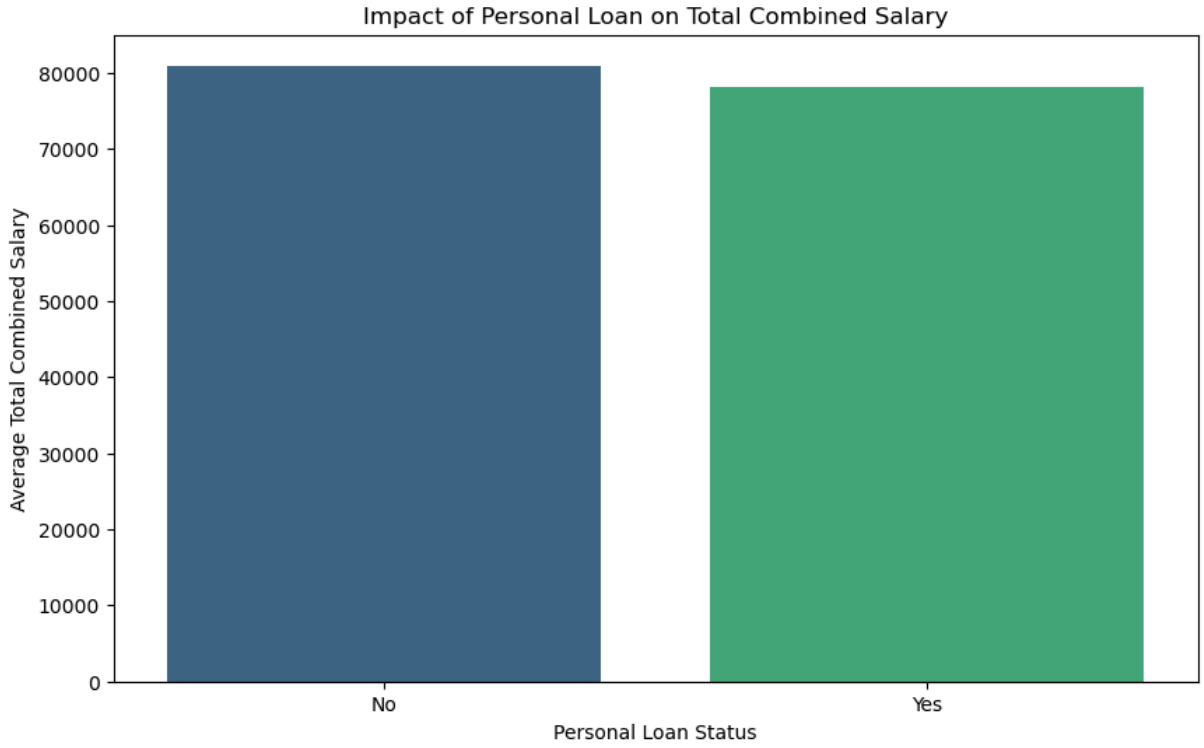
## Observation

## According to the statistical analysis:

- count of Male is 1251, whereas for Female is 327
- Mean salary for male is Rs. 58760.43, whereas for Female it is Rs. 65816.51
- Minimum salary for Male is Rs. 30000, whereas for Female it is Rs. 34800
- Maximum salary for both Male and Female is Rs. 99300

# 16. Loan Status Impact:

- How does having a personal loan affect the total combined salary of the individual and their partner?




- Figure 16: personal loan impact by total combined salary

## Observation

Having a personal Loan does not much affect the total combined salary of the individual and their partners

# 17. Partner's Salary Contribution:

## What is the average partner's salary for individuals with and without house loans?

```
Average partner's salary for individuals with house loans: 19661.216730038024
Average partner's salary for individuals without house loans: 20975.190114068442
```

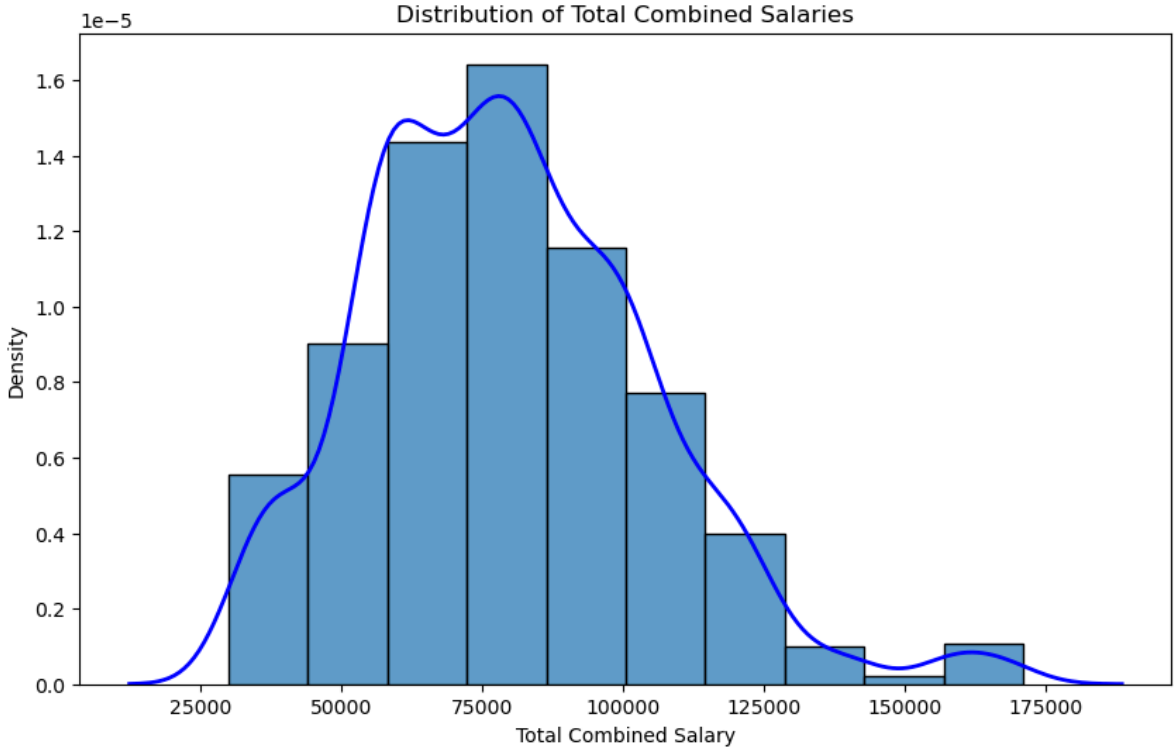
## Observation

Average salary of partner for individuals with house loan are: Rs 19661.21

Average salary of partner for individuals without house loan are: Rs 20975.19

# 18. Total Salary Distribution:

## Create a histogram showing the distribution of total combined salaries. Identify and discuss any skewness or outliers in the data.




- **Figure 17: Distribution of total combined salary**

# THANK YOU

In [ ]: