Aditya Raj

Bright Automative Company Analysis

SUMMER BOOTCAMP PROJECT 2024

List of Tables

List of Figures

Problem Statement/Objectives:

Data

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 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

Loading The necessary libraries.

1. Display the top 5 rows.

	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
- Based on above results we can see that the following columns of 'No_of_Dependents' and 'Make' have some wrong values as:
- 'No_of_Dependents' have value as '?' on index 3.
- 'Make' have vaue as '?' on index 3.
- 'Profession have value as 'NaN' on index 4.

2. Display 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	2.0	4.0	2.0	3.0	4.0
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	20225.559322	0.0	20225.559322	0.0
Total_salary	33300	32000	32900	32200	31600
Price	27000	31000	30000	24000	31000
Make	Hatchback	Hatchback	Hatchback	Hatchback	Hatchback

- Table 2: Last 5 rows.
- Based on above results we can see that the following columns of **'Partner_salary'** have some wrong values as:
- **'Partner_salary'** have value as **'NaN'** on index no. **1577** and **1579**.

3. Check the shape of dataset.

- It shows that the dataset is having **1581** rows and **14** columns.

3. Check the shape of dataset.

The datasets have following types of data in given column.

- **1. Age have integer type of data.**
- **2. Gender have object type of data.**
- **3. Proffesion have object type of data.**
- **4. Marital_Status have object type of data.**
- **5. Education have object type of data.**
- **6. No_of_Dependents have object type of data.**
- **7. Personal_loan has object type of data.**
- **8. House_loan has object type of data.**
- **9. Partner_working has object type of data.**
- **10. Salary has float64 type of data.**

- **11. Partner_salary has float64 type of data.**
- **12. Total_salary has integer type of data.**
- **13. Price has integer type of data.**
- **14. Make has object type of data.**

5. Check the Statistical summary.

Results of checking the staistical summary is:

- Age have count of 1568, avg of 31.95, Standard Deviation of 8.71, Minimum (min) of 14, 25percent of 25, Median of 29, 75percent of 38, Maximum of 120.
- Salary column has a count of 1568, an average salary of ₹60,276.91, a standard deviation of ₹14,636.20, a minimum salary of ₹30,000, a 25th percentile at ₹51,900, a median salary of ₹59,450, a 75th percentile at ₹71,700, and a maximum salary of ₹99,300.
- Partner_salary column has a count of 1475, an average partner salary of ₹20,225.56, a standard deviation of ₹19,573.15, a minimum partner salary of ₹0, a 25th percentile at ₹0, a median partner salary of ₹25,600, a 75th percentile at ₹38,300, and a maximum partner salary of ₹80,500.
- Total_salary column has a count of 1581, an average total salary of ₹79,626.00, a standard deviation of ₹25,545.86, a minimum total salary of ₹30,000, a 25th percentile at ₹60,500, a median total salary of ₹78,000, a 75th percentile at ₹95,900, and a maximum total salary of ₹171,000.
- Price column has a count of 1581, an average price of ₹35,948.17, a standard deviation of ₹21,175.21, a minimum price of ₹58, a 25th percentile at ₹25,000, a median price of ₹31,000, a 75th percentile at ₹47,000, and a maximum price of ₹680,000.

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
dtype: int64
```

In the given dataset we have null values in following columns only:

- Gender column has 53 missing values.
- Profession column has 6 missing values.

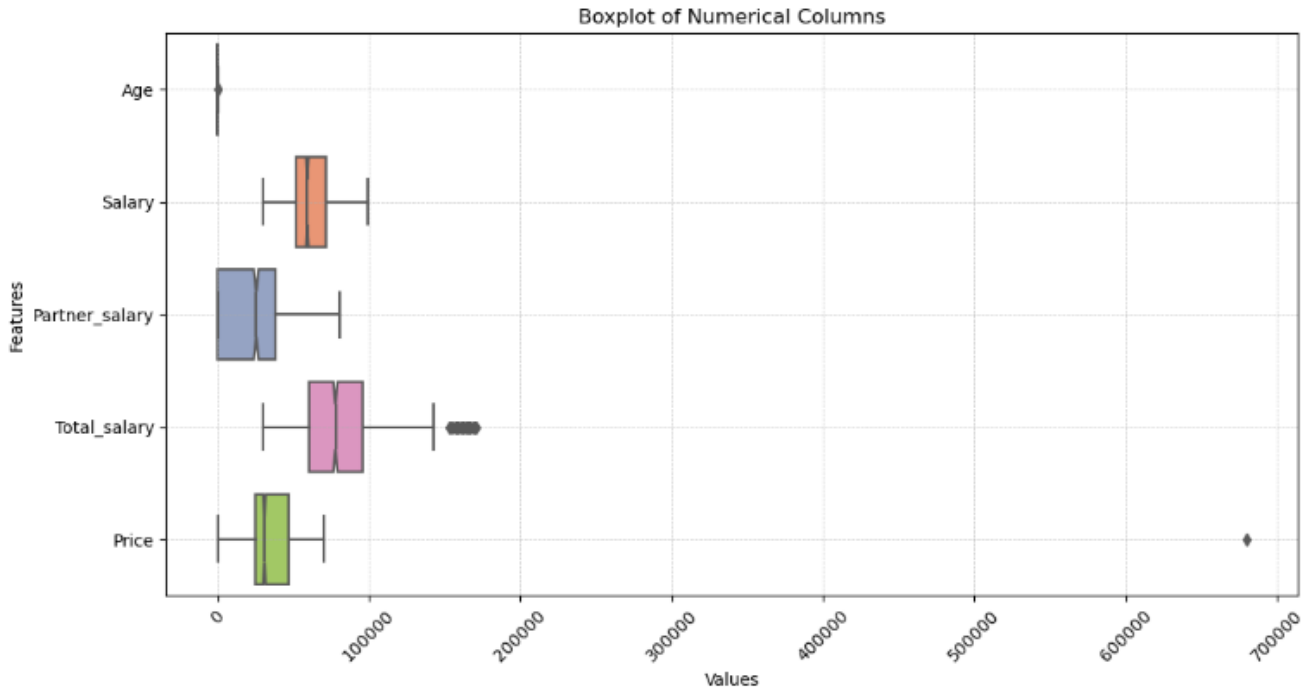
- Salary column has 13 missing values.
- Partner_salary column has 106 missing values.

7. Check the duplicate values

In the given dataset there are 0 duplicate values.

8. Check the anomalies or wrong entries.

- **We will use IQR(Interquartile Range) for detecting outlier in our dataset.**




```
Outliers detected using IQR method:
Age                1
Salary             0
Partner_salary     0
Total_salary      27
Price              1
dtype: int64
```

The detected outliers for the following columns are:

- Age have 1 outlier.
- Salary have 0 outlier.
- Partner_salary have 0 outlier.
- Total_salary have 27 outlier.
- Price have 1 outlier.

10. Do the necessary data cleaning steps like dropping duplicates, unnecessary columns, null value imputation, outliers treatment etc.

i. Dropping the duplicate values in case if there is any.

- There are 0 dupliccate values.

ii. Unecessary columns

- All the columns are required in this datasets.

iii. Null Value imputation

- After analysing dataset we can see that there is a wrong entry in Gender column as 'Femal', 'Femle'.
- We need to fix this mistake by replacing it with 'Female'.
- We can se that there is ? as value in some columns so we will replace it with nan.
- We will impute the null values of **Gender, Profession, Make** using the mode.

```
Age                  0
Gender               0
Profession           0
Marital_status       0
Education            0
No_of_Dependents     2
Personal_loan        0
House_loan           0
Partner_working      0
Salary              13
Partner_salary     106
Total_salary         0
Price                0
Make                 0
dtype: int64
```

- As we can see that the null values of above mentioned columns are replaced with mode.
- After analysing the data we can see that the No_Of_Dependents is filled with numerical value but denoted as object-type (user defined) so we will change it in numerical format.
- **Now we will replace the missing values for numerical columns mentioned in dataset using mean.**

```
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
```

Now there is no more Null values in dataset.

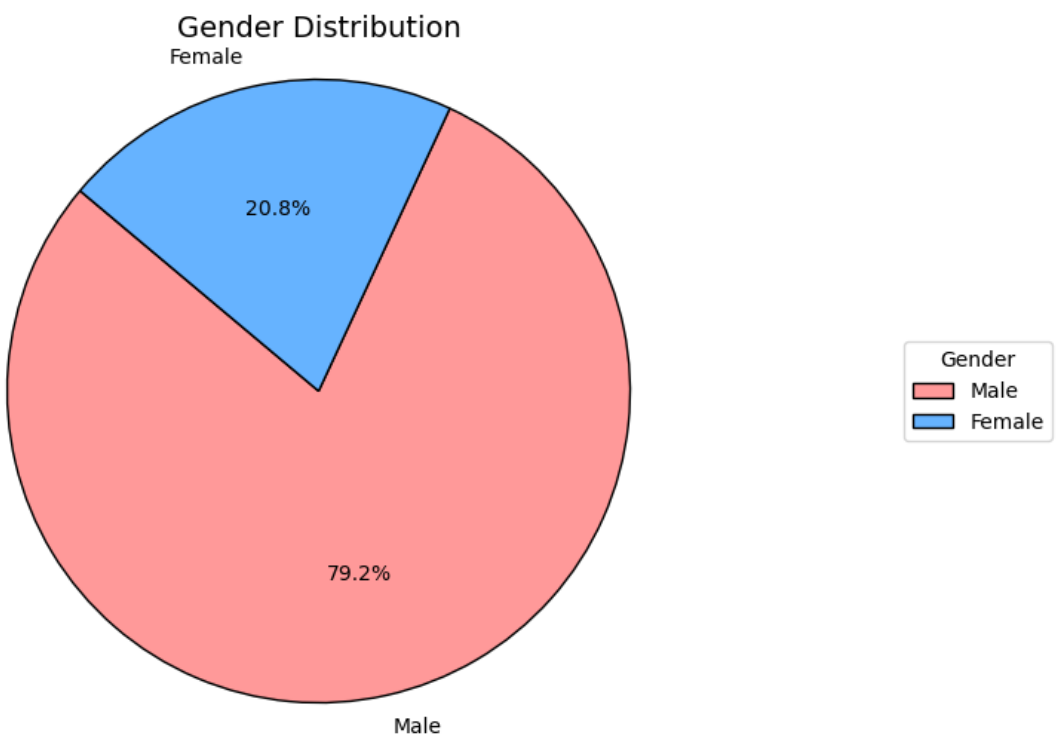
1. Descriptive Statistics:

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

- Mean age of the individual person is **31.952561**
- Median age of the individual person is **29.0**
- Standard deviation of the age for individual person is **8.71254886**

2. Data Distribution:

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



- The distribution of **Gender** in dataset is:
- Distribution of **Male** is: **79.2%**
- Distribution of **Female** is: **20.8%**

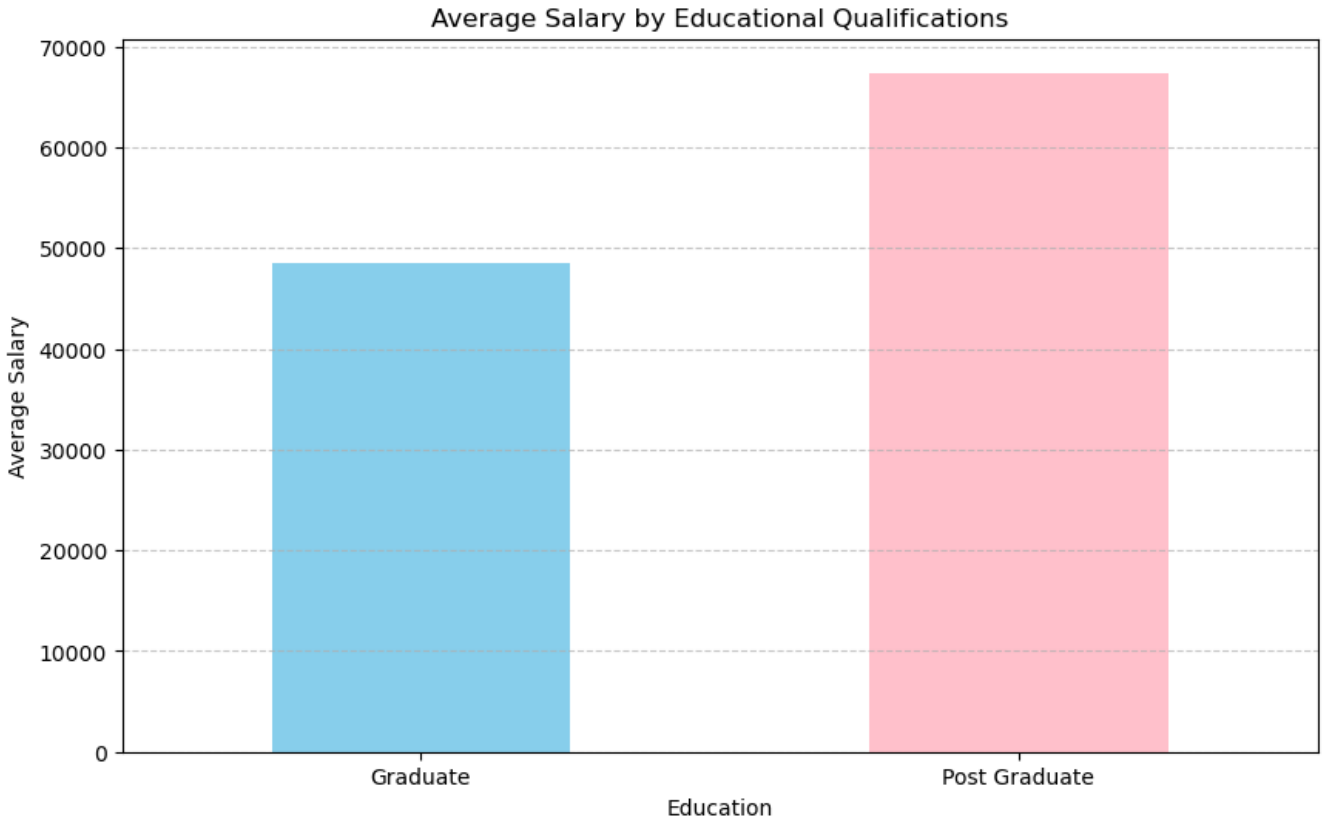
3. Correlation Analysis

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

- The correlation coefficient cames out to be **0.5928552429412259**.
- It show there is a very strong **positive** correlation between **Age** and **Salary** of the person.

4. Salary Analysis:

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



- The average salary based on thier **Education** qualifications is:
- **Graduate** is having an **Average Salary** of **48520.147069**
- **Post Graduate** is having an **Average Salary** of **67390.651999**

5. Loan Status

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

The percentage of individual with personal loans:

- person having personal loan are **50.094877%**.
- person not having personal loan are **49.905123%**.

Loans

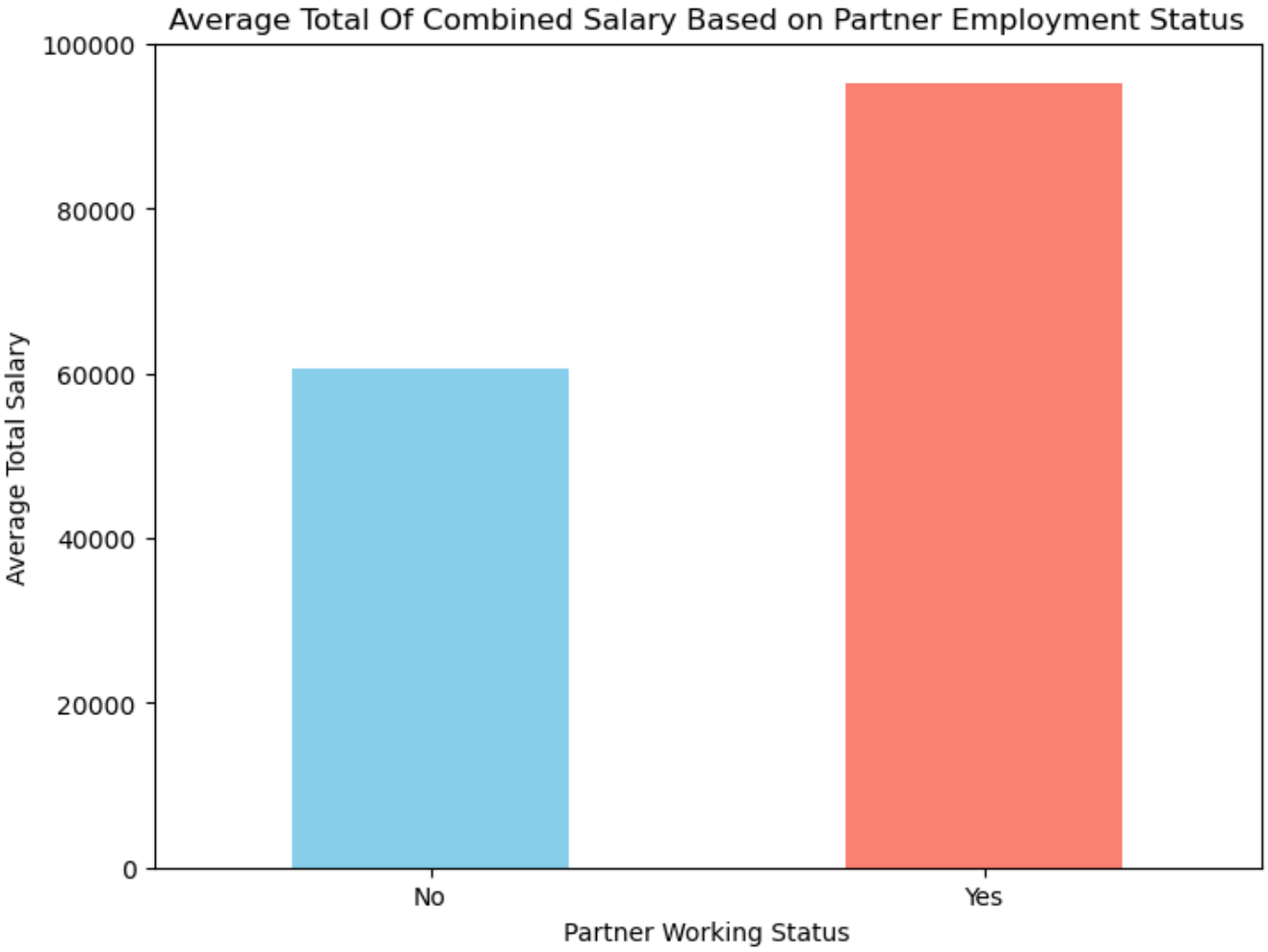
- **Males** having loans are **51.357827%**.
- **Males** having no loans are **48.642173%**.
- **Females** having loans are **45.288754%**.
- **Females** having no loans are **54.711246%**.

6. Marital Status and Dependents

- Average number of dependents on **Married** individuals are **2.538402**.
- Average number of dependents on **Single** individuals are **1.608696**.

7. Partner Employment:

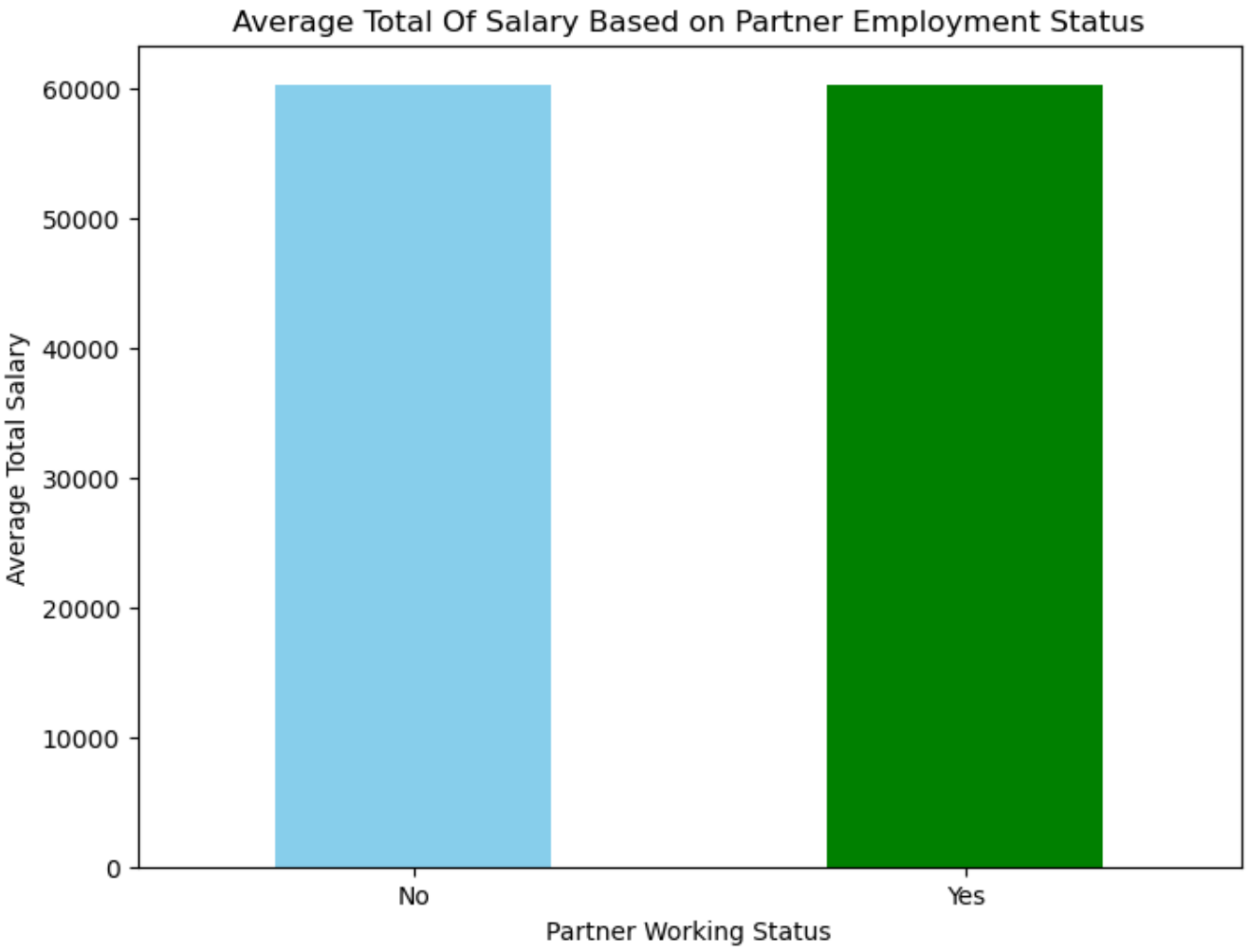
- **How does the employment status of a partner affect the total combined salary?**



- Based on the baove analysis we can say that the employment of partner directly depends on the total combined salary as:
- Partner working have **average total salary** of **95314.285714**.
- Partner who does not work have **average total salary** of **60527.208976**.

8. Salary Comparison:

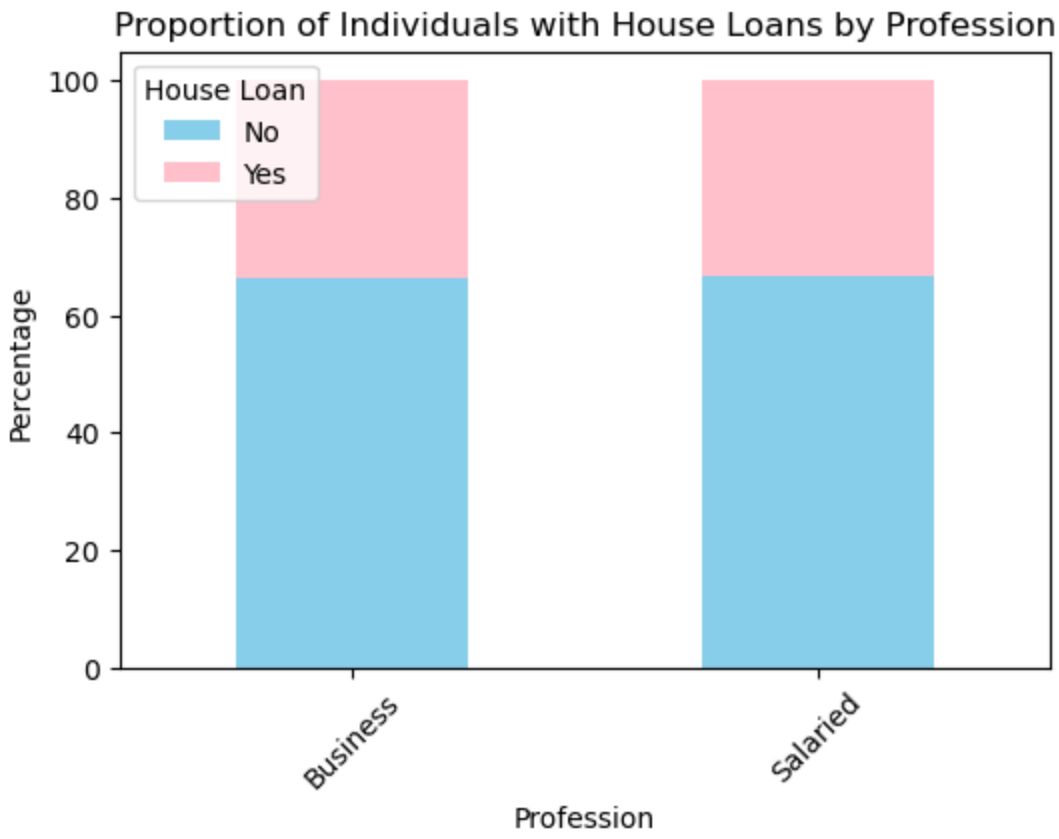
- **Compare the average salary of individuals whose partners are working versus those whose partners are not working.**



- **Based on above analysis we can say the following things about the given arguements in problem statement:**
- Individual whose partners are not employed have average **Salary** of **60271.528573**.
- Individuals whose partners are employed have average **Salary** of **60281.336406**.
- So from the baove data we can say that there is very negligible difference in average salary of both individuals irresepect of thier partners are working or not.

9. House Loan Analysis:

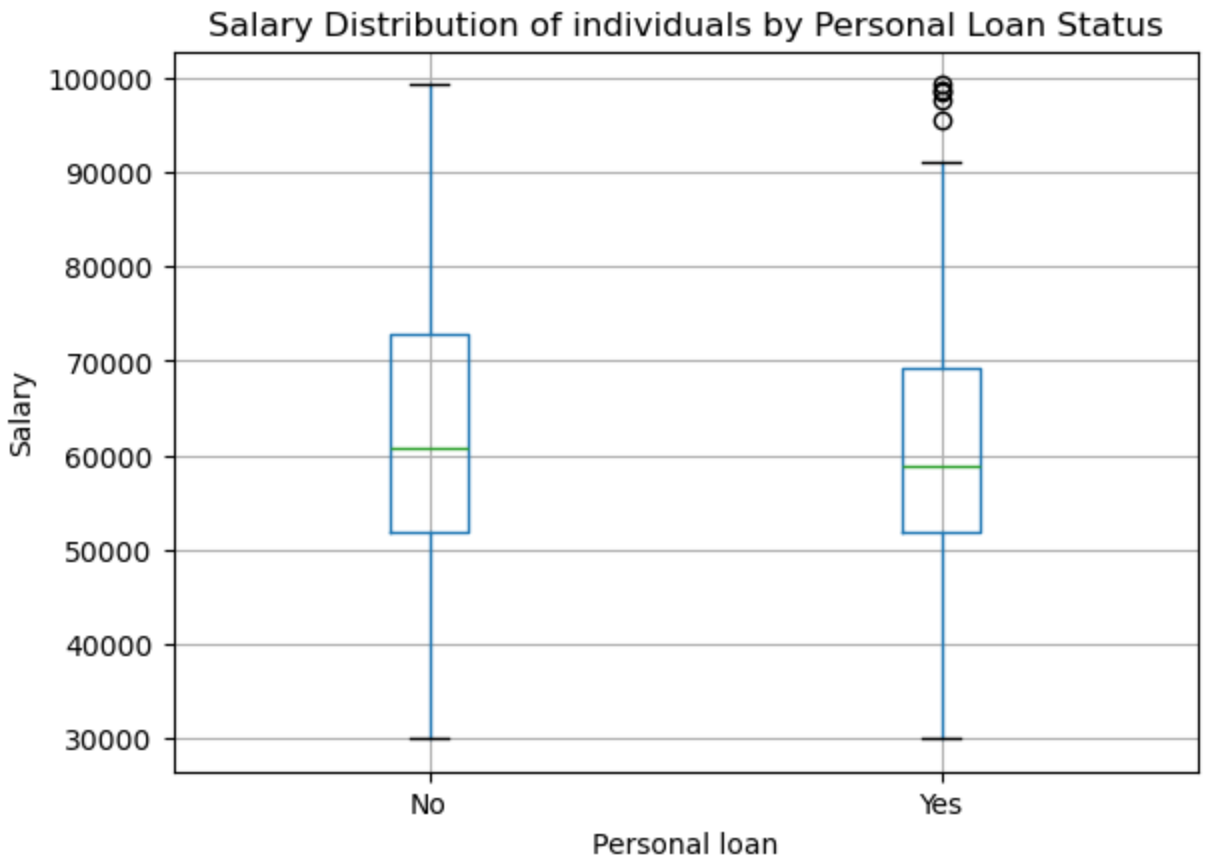
- **What is the proportion of individuals with house loans based on their profession?**



- **Based on above visualisation we can say that the following individuals with followiing profession are having house loans:**
- Business man taking house loans are **33.430657%** and Business man taking no House Loans are **66.569343%**.
- Salaried class individuals taking House Loans are **33.258929%** and Salaried class individuals not taking House Loans are **33.258929%**.
- **Based on above results we can say that the Business class are taking slightly more House Loans than the salaried class.**

10. Salary Distribution:

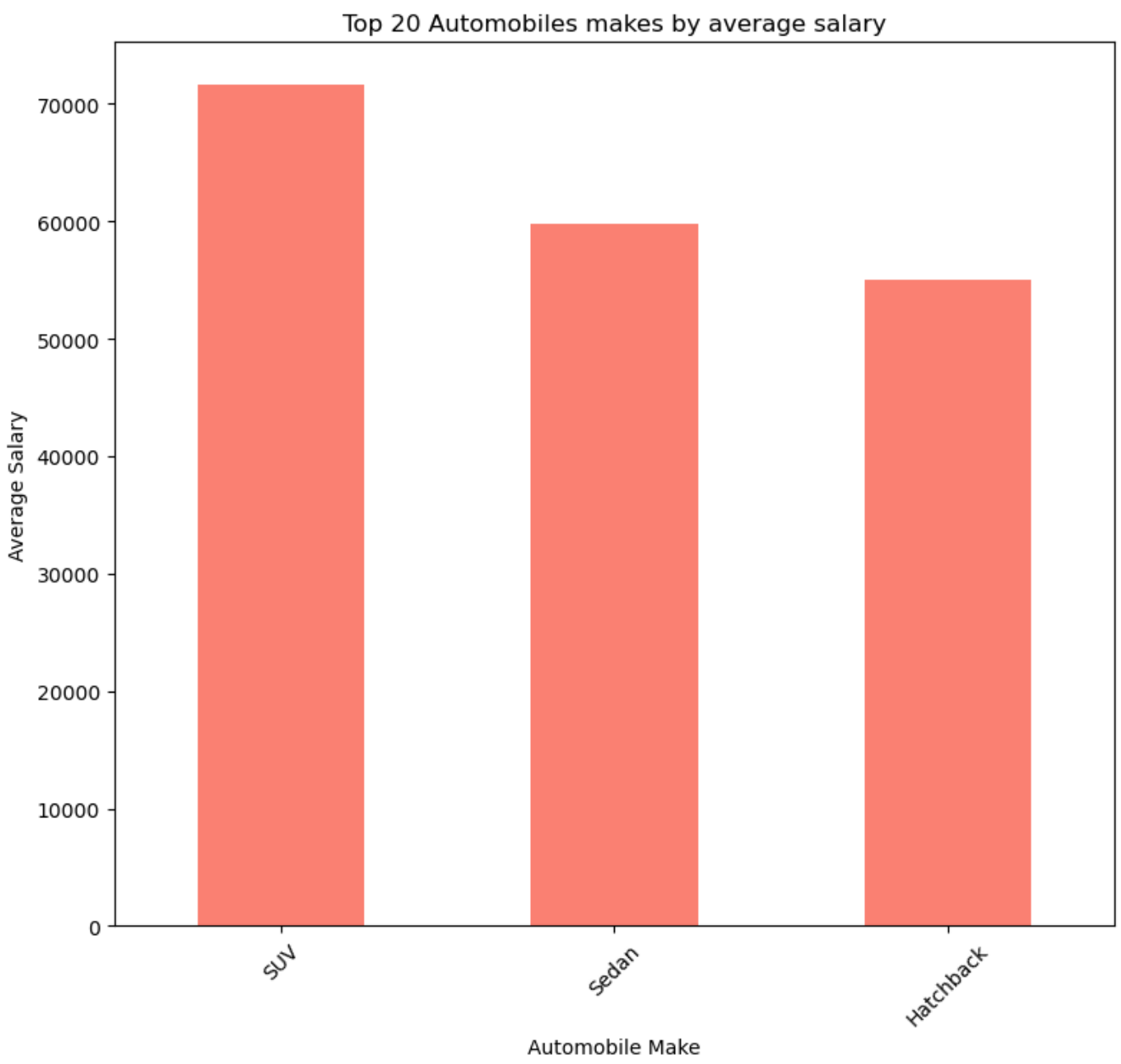
- **What is the distribution of salaries for individuals with personal loans versus those without personal loans? Represent it using a box plot.**



- To interpret the result from box plot we will perform t test on boxplot.
- A p-value of 0.01599892337904925 is less than the common significance level of 0.05, suggesting that there is a statistically significant difference in the salaries between individuals with personal loans and those without personal loans.
- Since the p-value is less than 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference in the salaries of individuals with personal loans compared to those without personal loans.
- Both class have diffrent salaries.

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.**



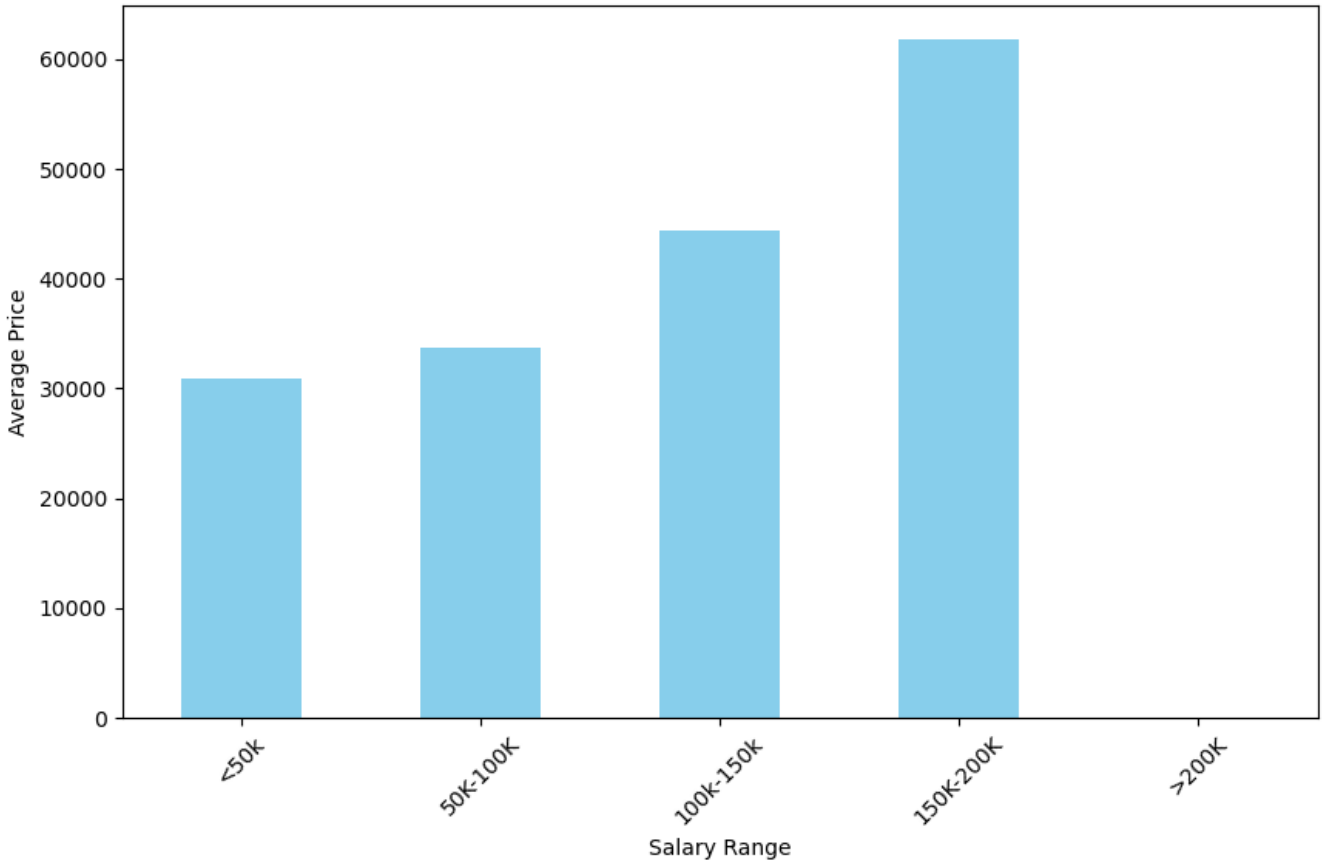
- **Based on the above results we can say that the:**
- Person having average salary of **71673.037444** makes **SUV**.
- Person having average salary of **59794.962822** makes **Sedan**.
- Person having average salary of **55083.505155** makes **Hatchback**.

- It further describes that the **SUV** is owned by rich individuals.

- **Sedan** is owned by moderate individuals.
- **Hatchback** is owned by lower-moderate individuals.

12. Price Analysis:

- **What is the average price of the product/service in the dataset? How does this price vary based on the individual's total salary?**



- **Average price** of the **product/service**: **$35948.17**

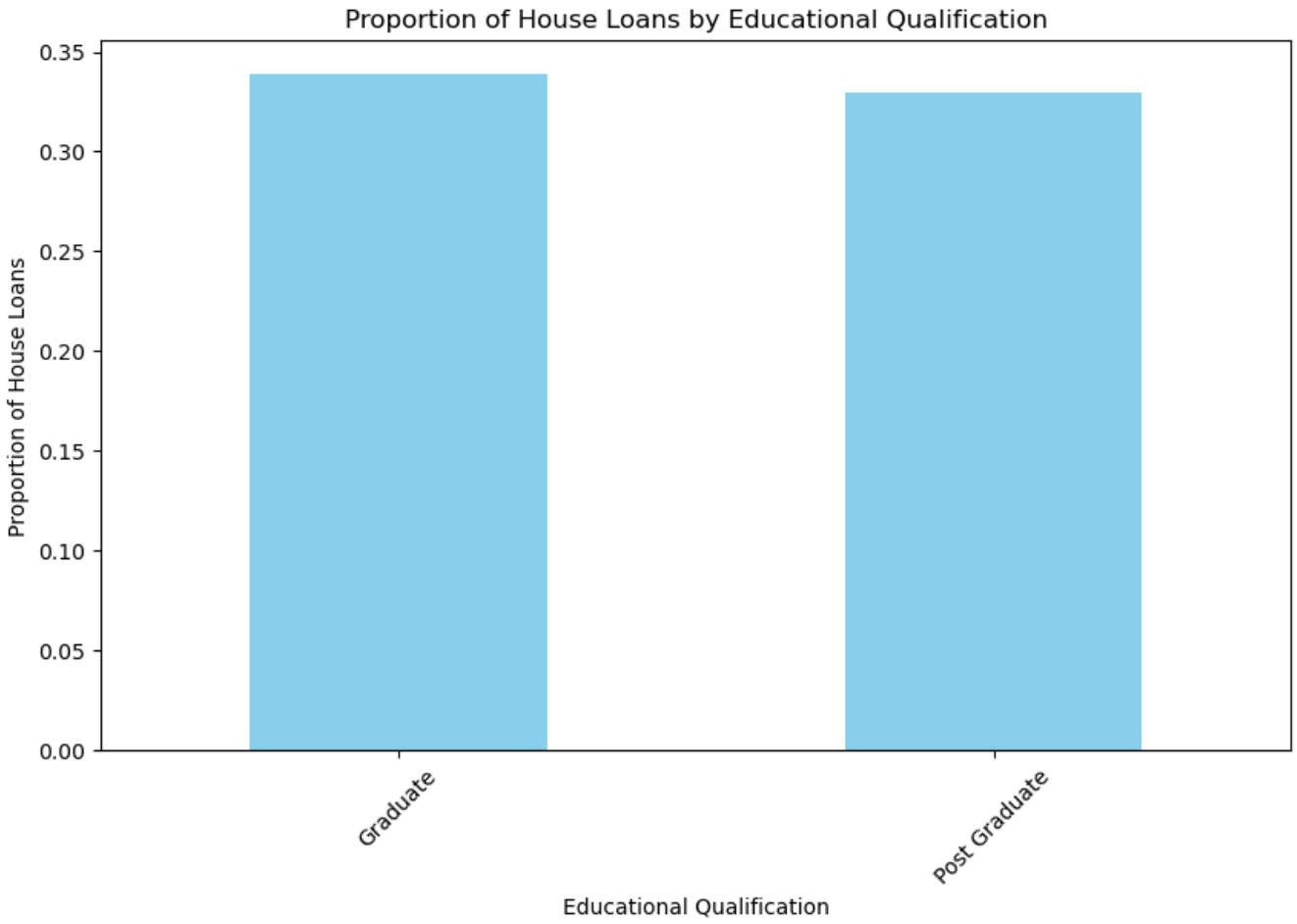
13. Martial Status and Loans

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

```
Personal_loan    No  Yes
Marital_status
Married         723  720
Single           66   72
Chi-square Test statistics: 0.1782394652161357
P-value: 0.6728906021290224
There is no significant difference in the number of personal loans taken by married individuals compared to single individuals.
```

14. Educational Qualification Impact:

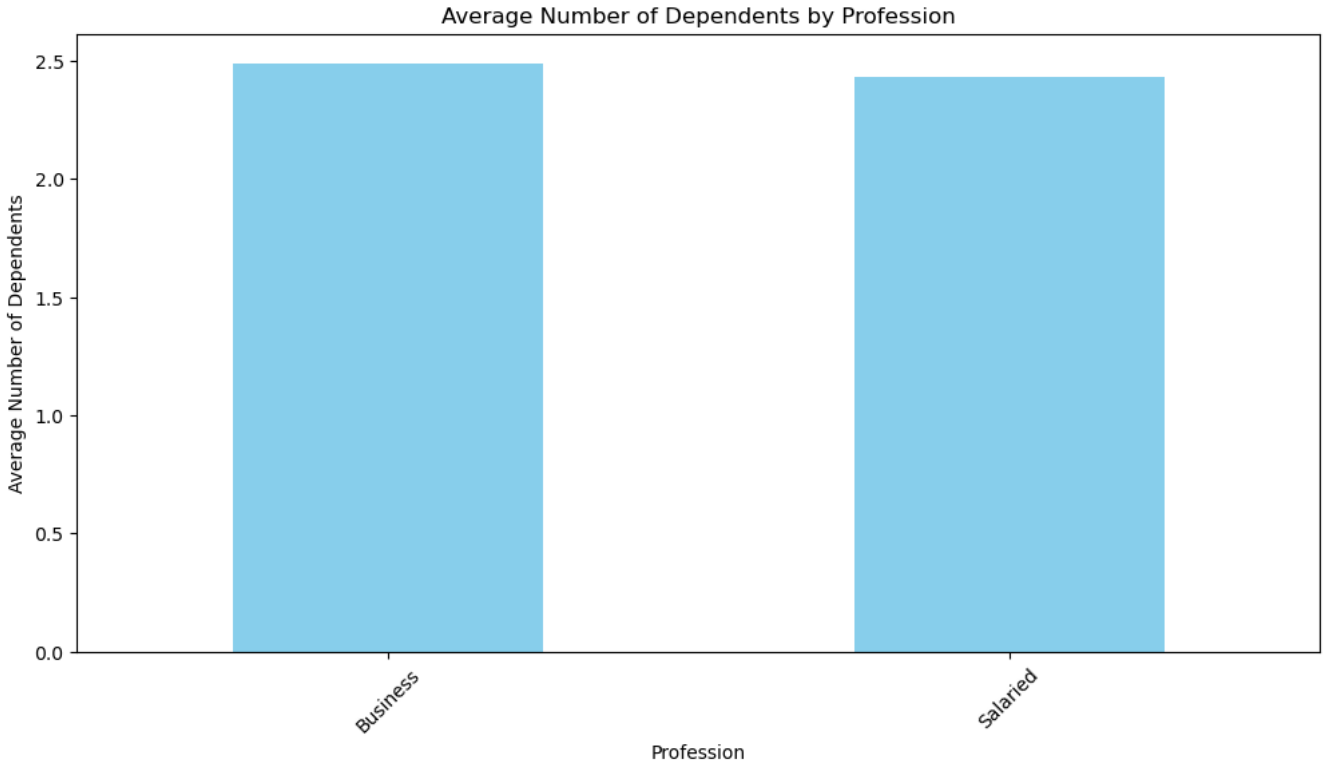
- **How does educational qualification impact the likelihood of taking a house loan?**



Educational qualification does not appear to strongly influence house loan decisions in this dataset. Further analysis could explore other factors affecting loan approval.

15. Dependent Count Analysis:

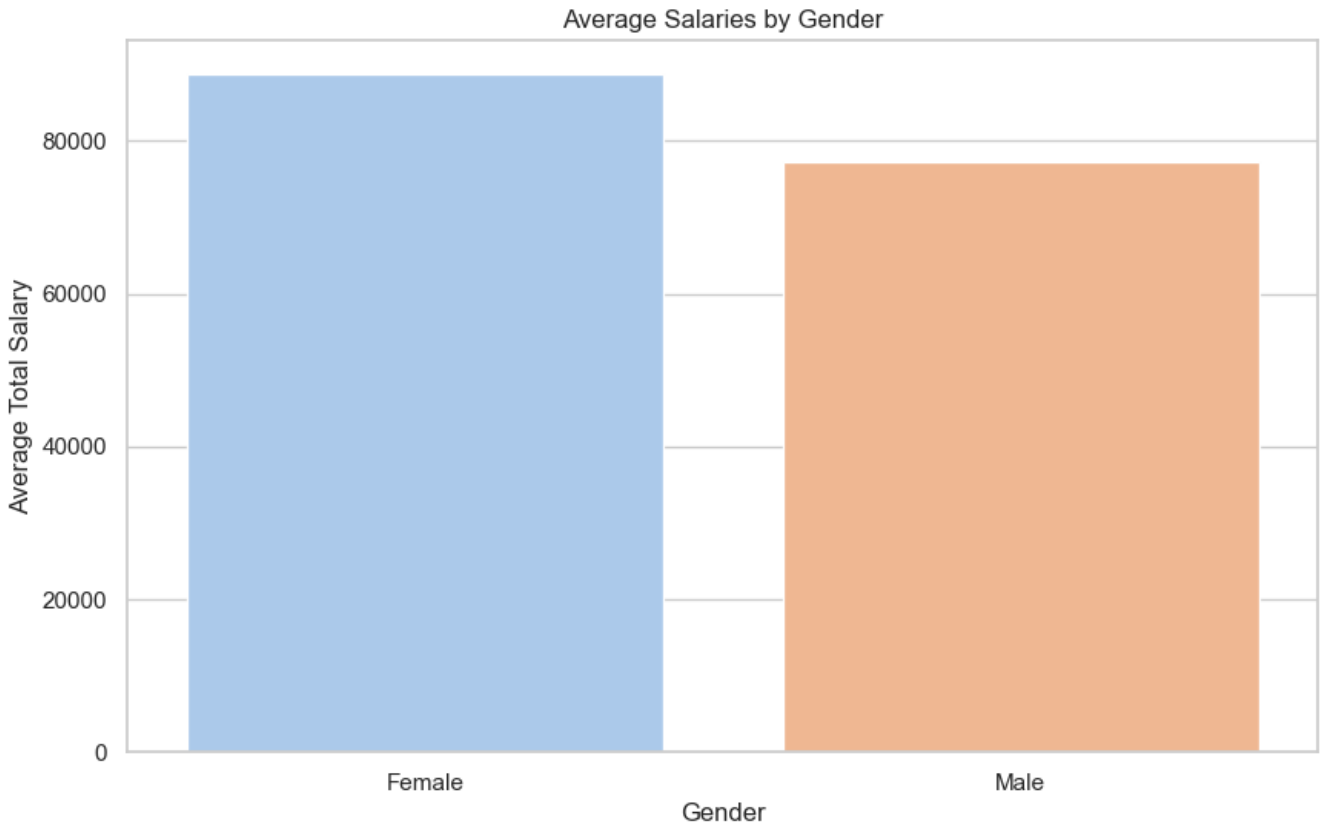
- **Analyze the number of dependents based on the profession of the individual. Which profession has the highest average number of dependents?**



- **The profession with the highest average number of dependents is: Business Average number of dependents in this profession: 2.49**

16. Gender and Salary:

- **Is there a significant difference in salaies between males and females? Provide statistical evidence.**



- **Male** earns lesser than the female with thier **Total_salary** as **77244.329073**, while **Female** earns an average **Total_salary** as **88689.361702**.

17. Regression Analysis

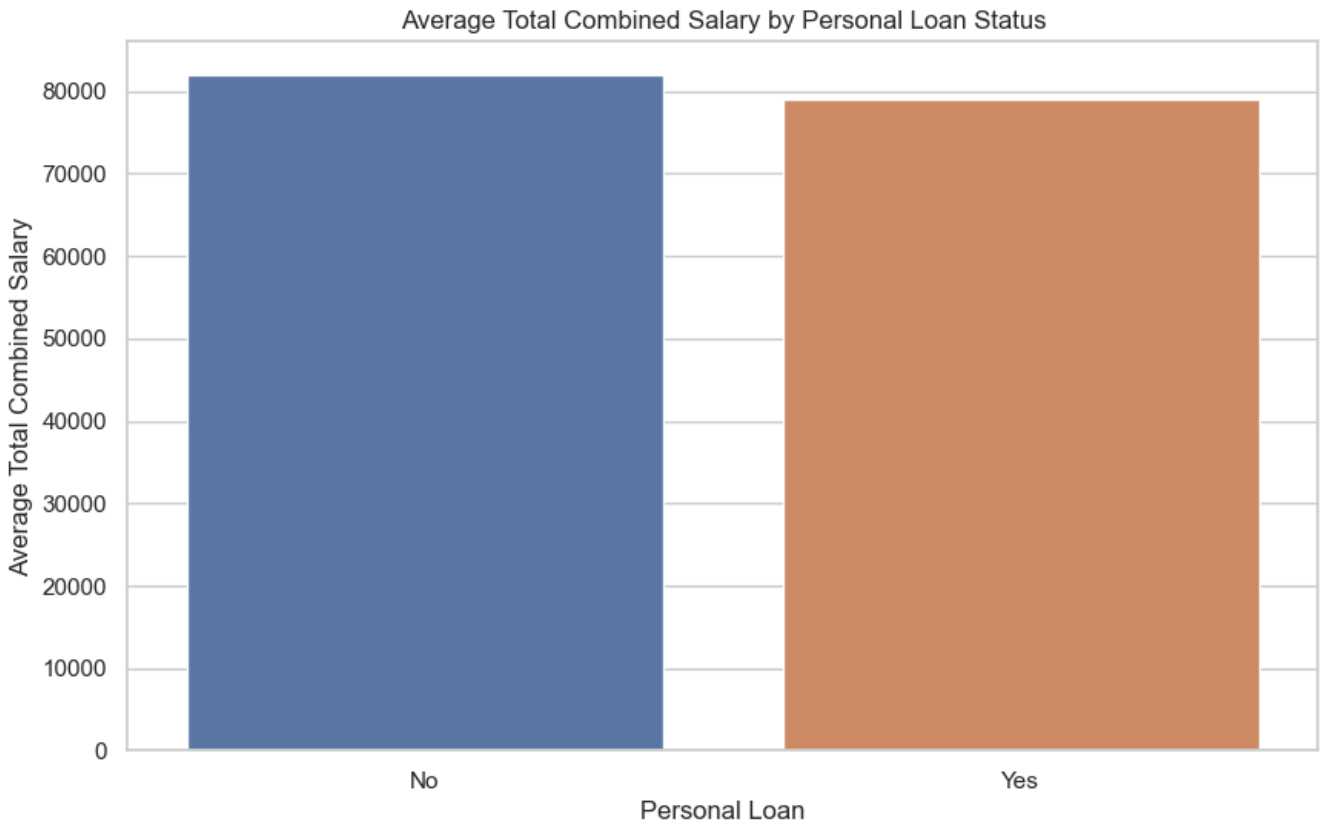
- **Build a regression model to predict an individual's salary based on age, education, and number of dependents. Discuss the model's accuracy and significance.**

- **Mean Absolute Error (MAE): 5528.008469466143**
- **R-squared (R^2): 0.7784914637819996**

- **Feature Coefficients:**
- Age 981.33125
- No_of_Dependents 890.018480
- ducation_Post Graduate 18600.409028

18. Loan Status Impact:

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

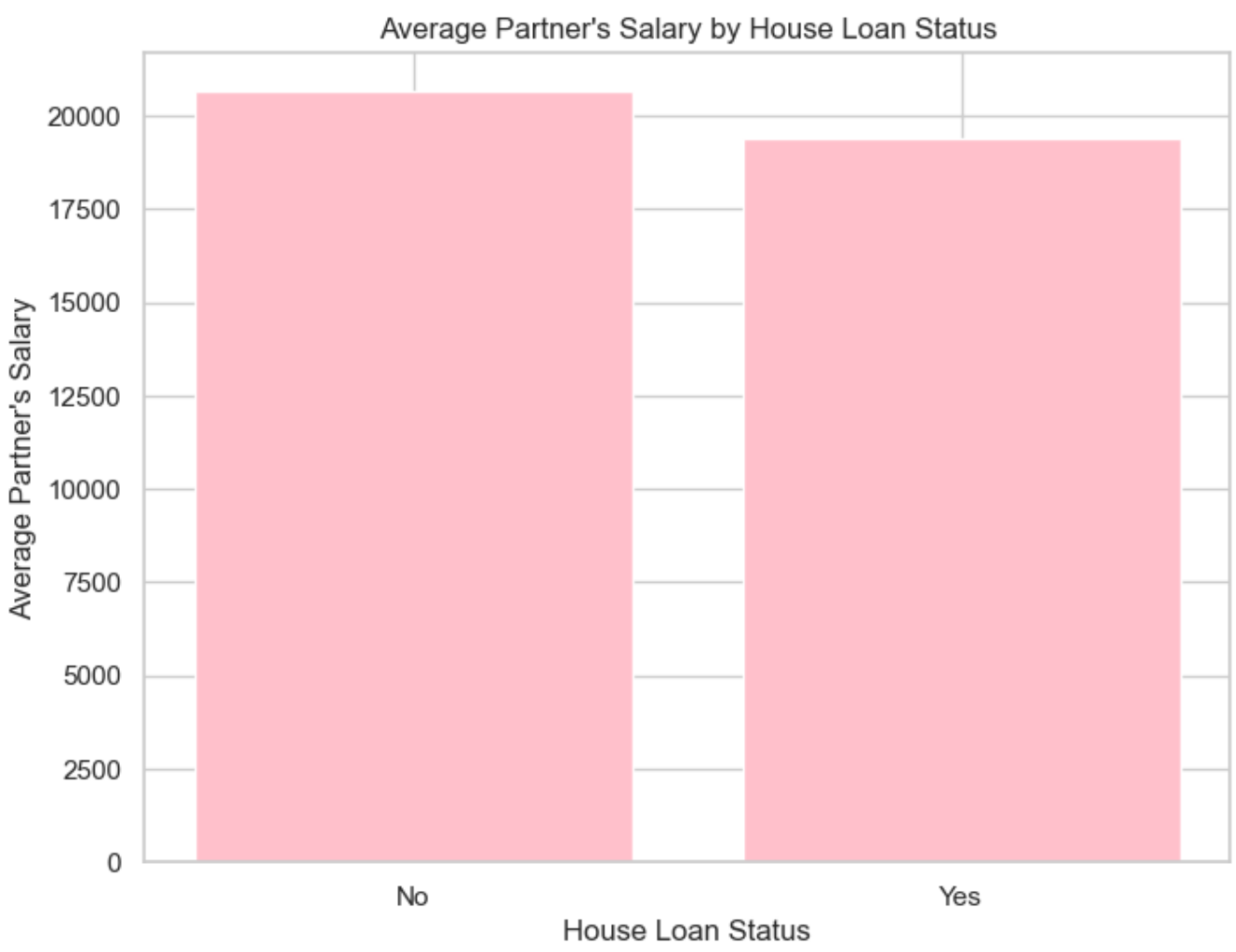


Average Total Combined Salary by Personal Loan Status

	Personal_loan	Total_combined_salary
0	No	81963.61
1	Yes	79046.87

19. Partner's Salary Contribution:

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

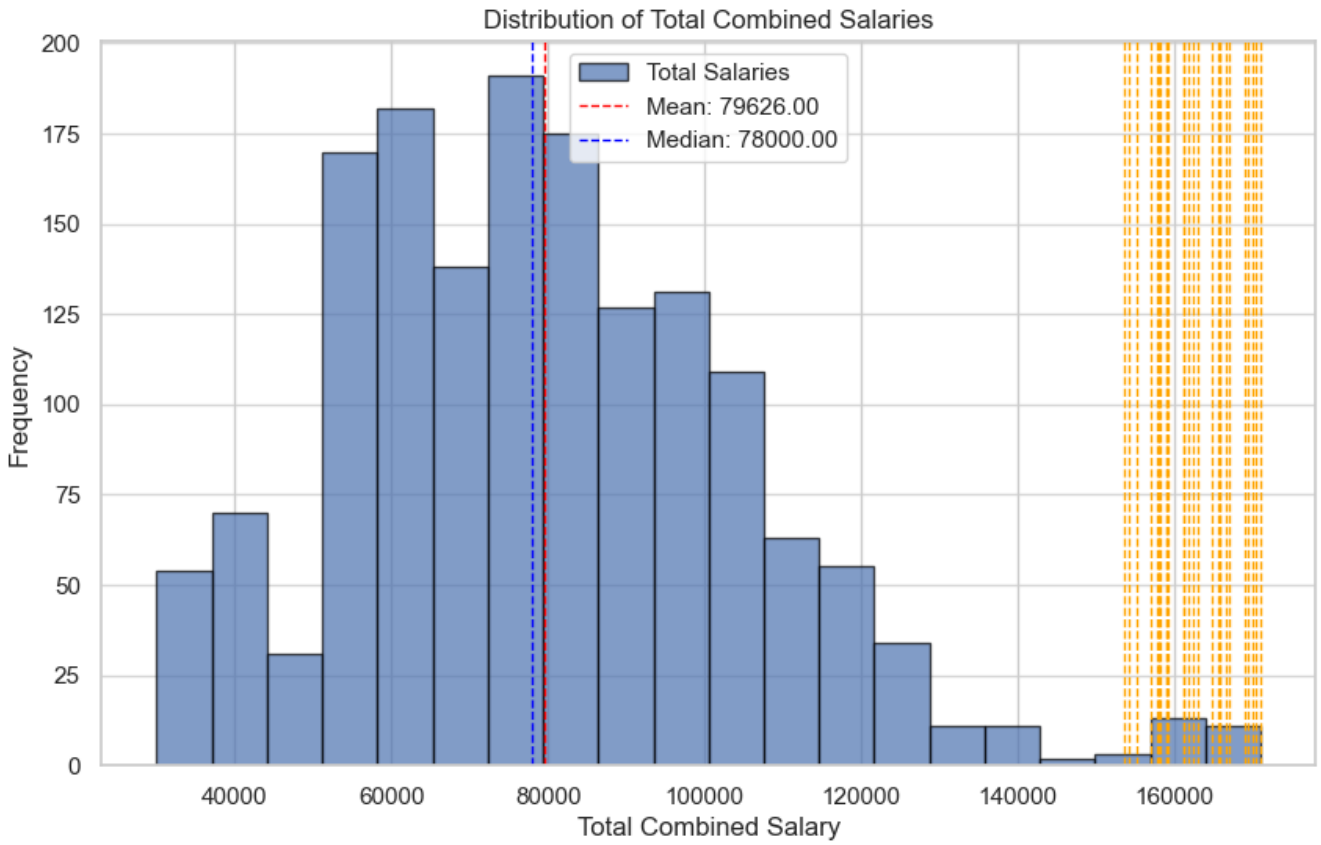


Average Partner's Salary by House Loan Status

	House_loan	Partner_salary
0	No	20646.03
1	Yes	19384.62

20. Total Salary Distribution:

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



- **Mean Salary 79625.996205**
- **Median Salary 78000**
- **Standard Deviation 25545.857768**
- **Outliers: 170000 165800 158000 165700 162900 159000 169000 165600 161100 166900 155200 170400 171000 154100 164700 161800 153500 169300 159100 162300 161100 166500 156900 158900 157700 157900 158200**.

Thank You For Reviewing My Project!!!.