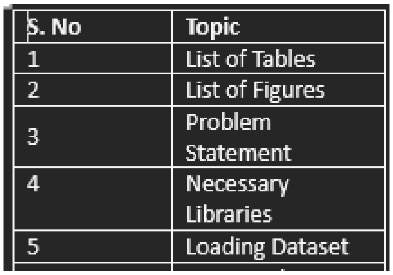
SUMMER BOOTCAMP PROJECT 2024

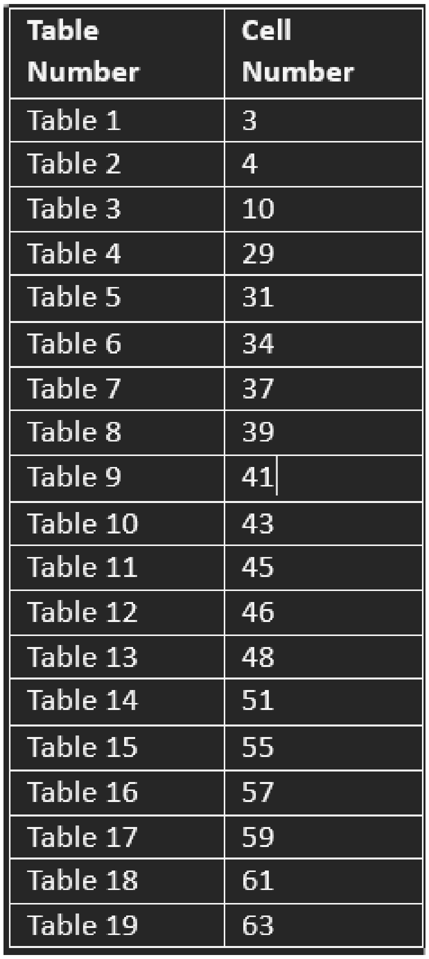
MRIDUL HEMRAJANI

2022511920

INDEX



LIST OF TABLES



LIST OF FIGURES

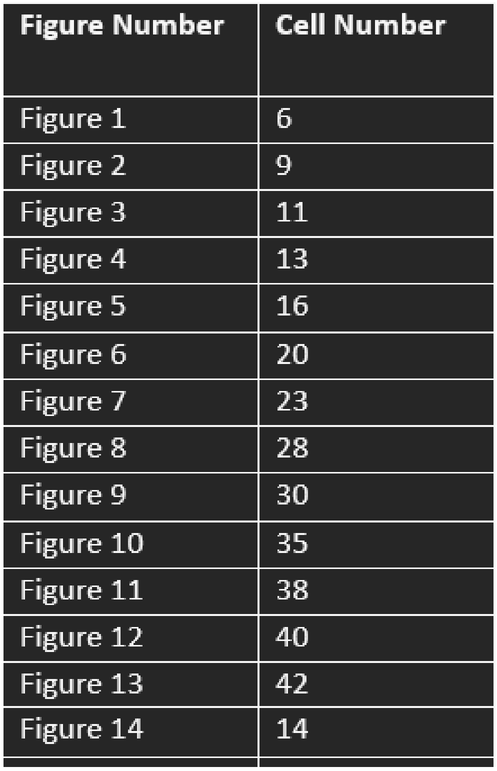


Figure Number	Cell Number
Figure 1	6
Figure 2	9
Figure 3	11
Figure 4	13
Figure 5	16
Figure 6	20
Figure 7	23
Figure 8	28
Figure 9	30
Figure 10	35
Figure 11	38
Figure 12	40
Figure 13	42
Figure 14	14

**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.

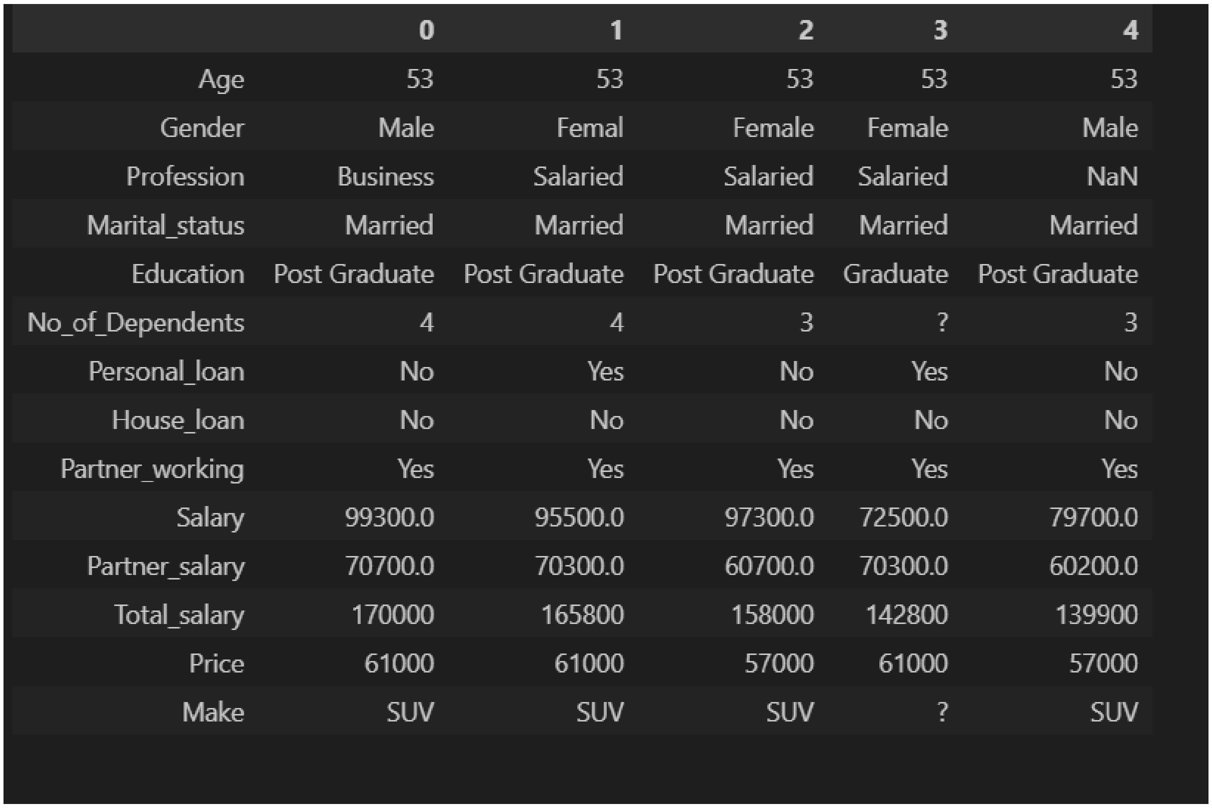
IMPORTING THE NECESSARY LIBRARIES

LOADING THE DATASET

BASIC EXPLORATION

1. FIRST 5 ROWS

Table 1



	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

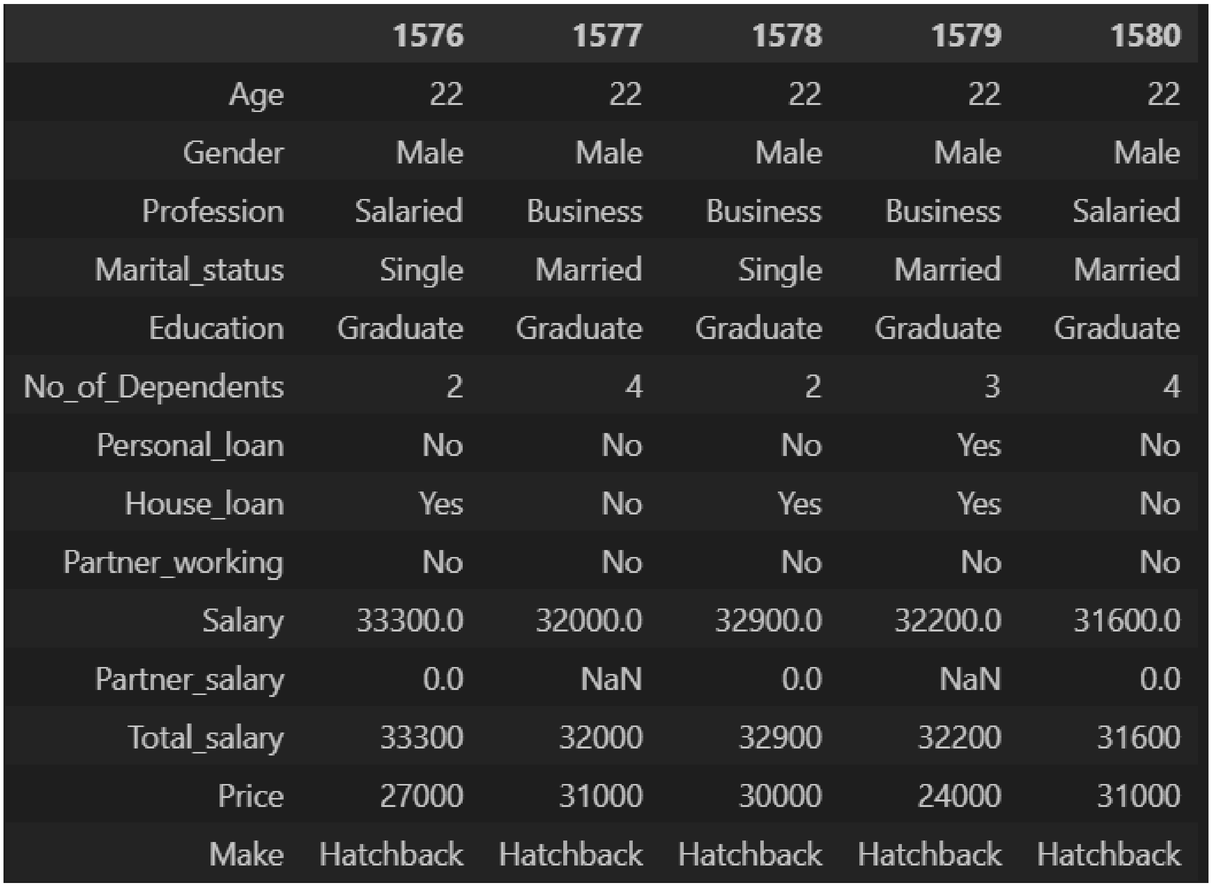
OBSERVATION

'No_of_Dependants' and 'Make' has value '?'. Need to check that.

'Profession' is NULL at one place.

2. LAST 5 ROWS

Table 2



	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

OBSERVATIONS

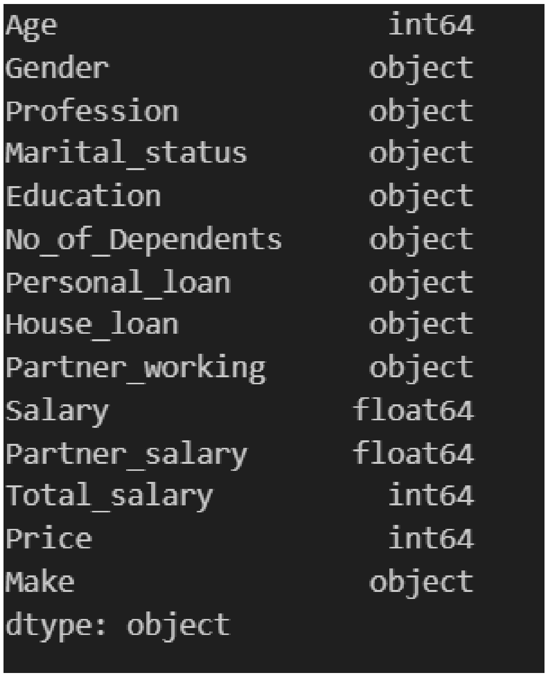
'Partner_Salary' is NULL at places.

3. Shape

There are 1581 rows and 14 columns in the given dataset.

4.Datatypes

Figure 1



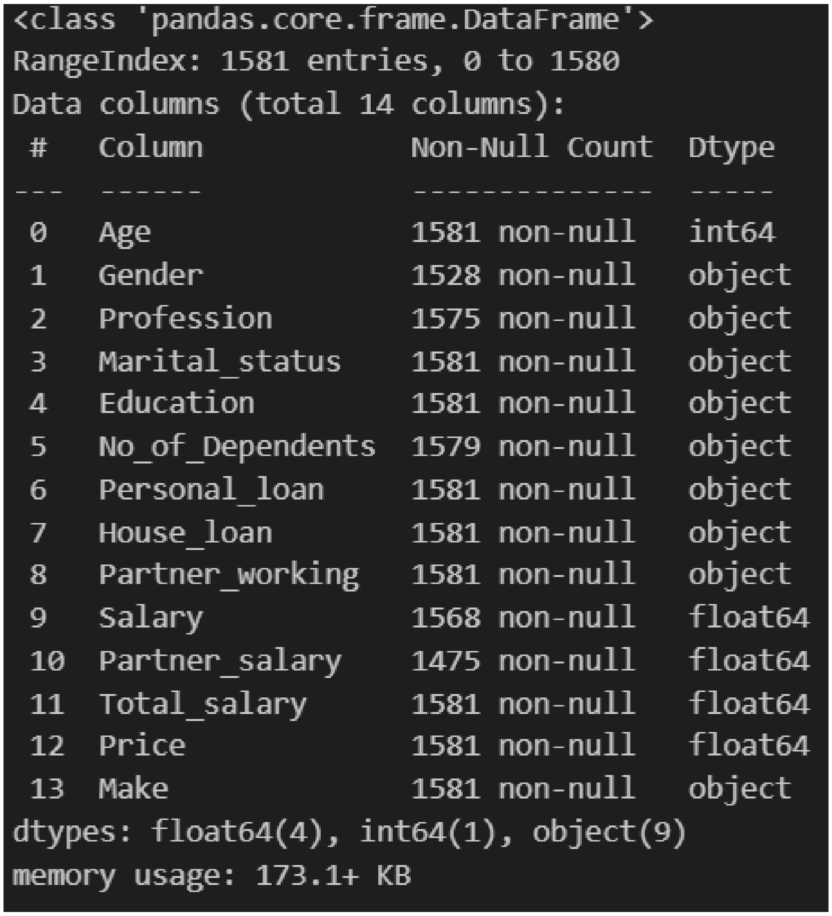
```
Age                   int64
Gender               object
Profession           object
Marital_status       object
Education            object
No_of_Dependents     object
Personal_loan        object
House_loan           object
Partner_working      object
Salary              float64
Partner_salary      float64
Total_salary          int64
Price                 int64
Make                 object
dtype: object
```

OBSERVATIONS

Data Types of 'Total_Salary' and 'Price' is int, it should be float.

Data Type of 'No_Of_Dependants' is object.

Figure 2



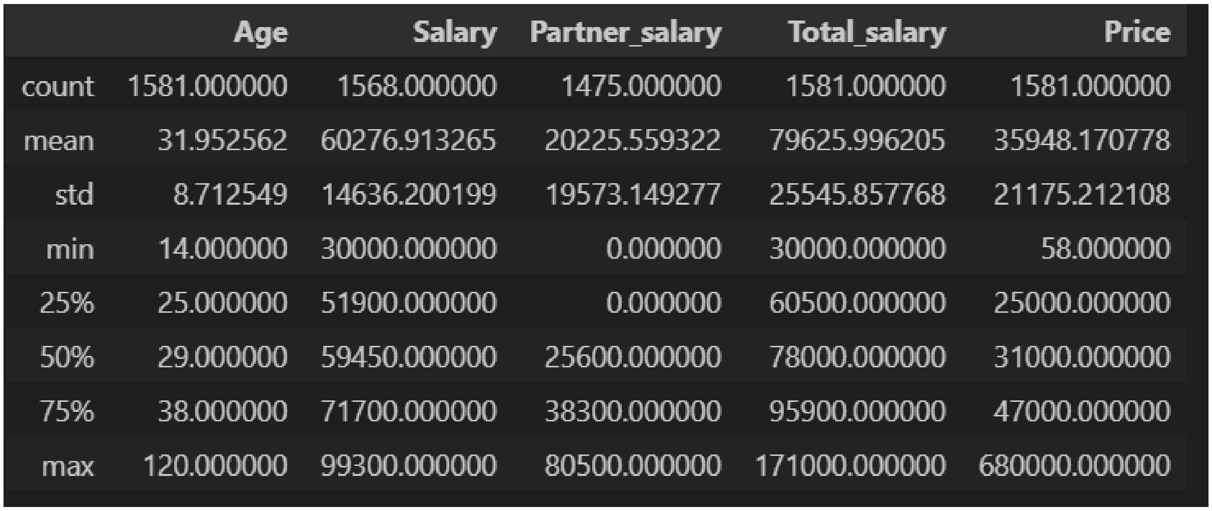
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1581 entries, 0 to 1580
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Age               1581 non-null   int64
 1   Gender            1528 non-null   object
 2   Profession        1575 non-null   object
 3   Marital_status    1581 non-null   object
 4   Education         1581 non-null   object
 5   No_of_Dependents  1579 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   float64
 12  Price             1581 non-null   float64
 13  Make              1581 non-null   object
dtypes: float64(4), int64(1), object(9)
memory usage: 173.1+ KB
```

Observations

NULL values in several columns.

5. Statistical Summary

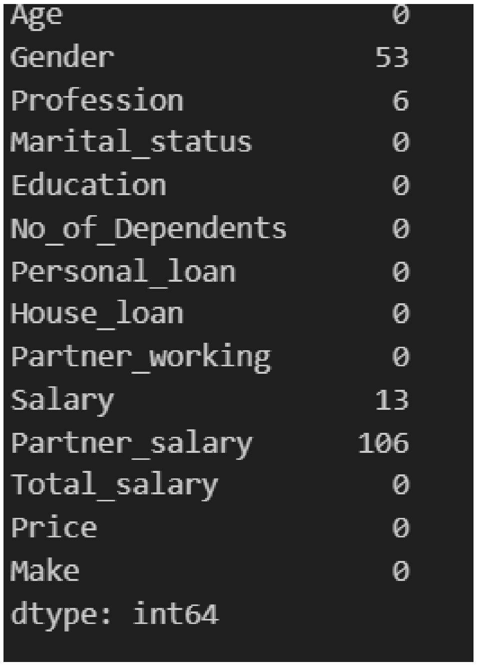
Table 3



	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

6. NULL Values

Figure 3



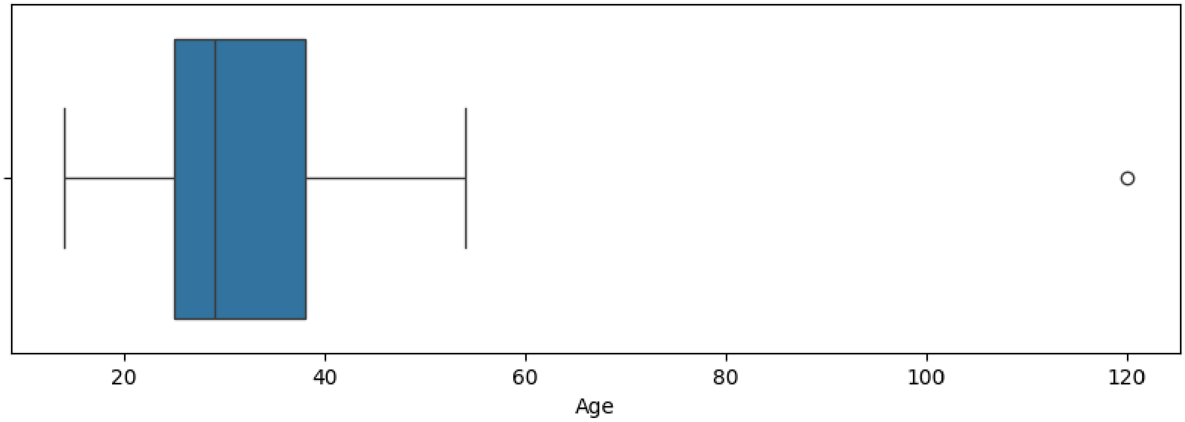
```
Age                 0
Gender             53
Profession          6
Marital_status      0
Education           0
No_of_Dependents    0
Personal_loan       0
House_loan          0
Partner_working     0
Salary             13
Partner_salary    106
Total_salary        0
Price               0
Make                0
dtype: int64
```

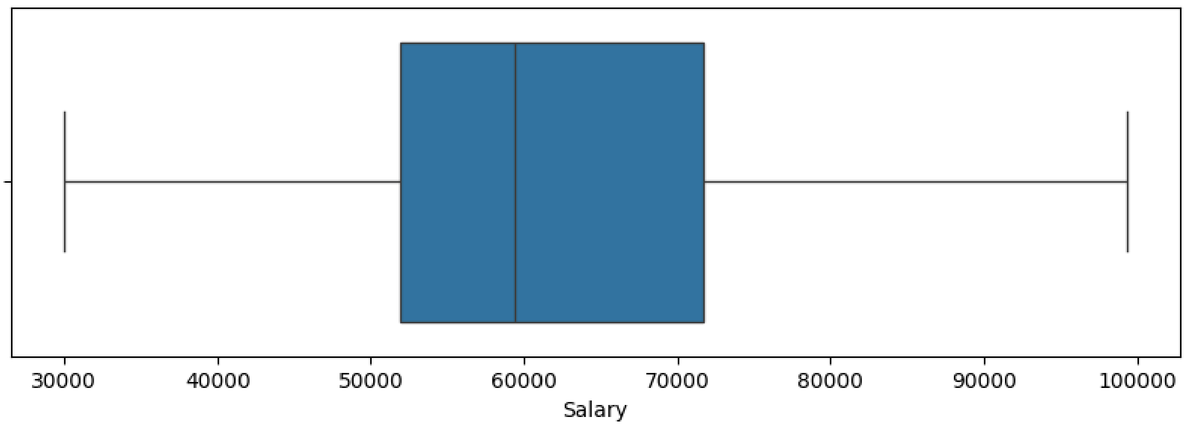
7. Duplicated Values

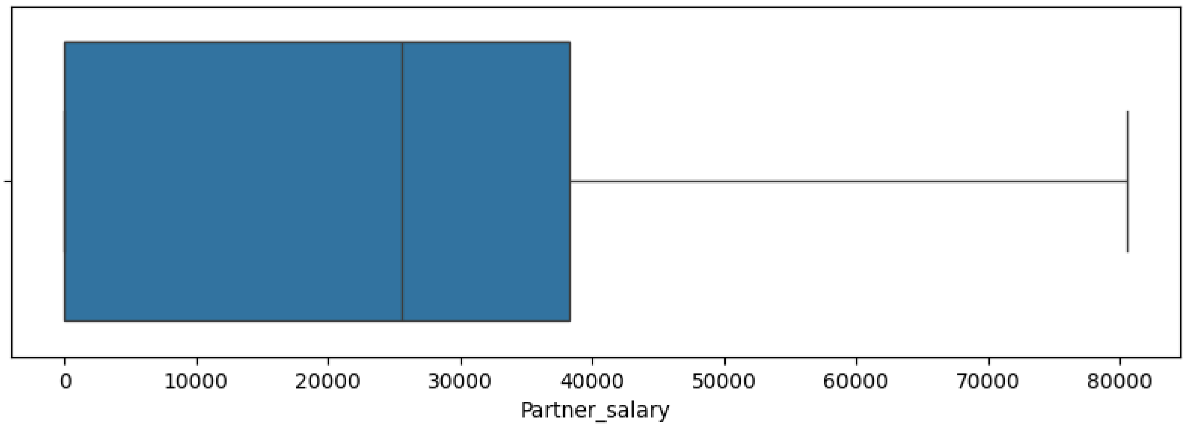
0 duplicated values in the dataset.

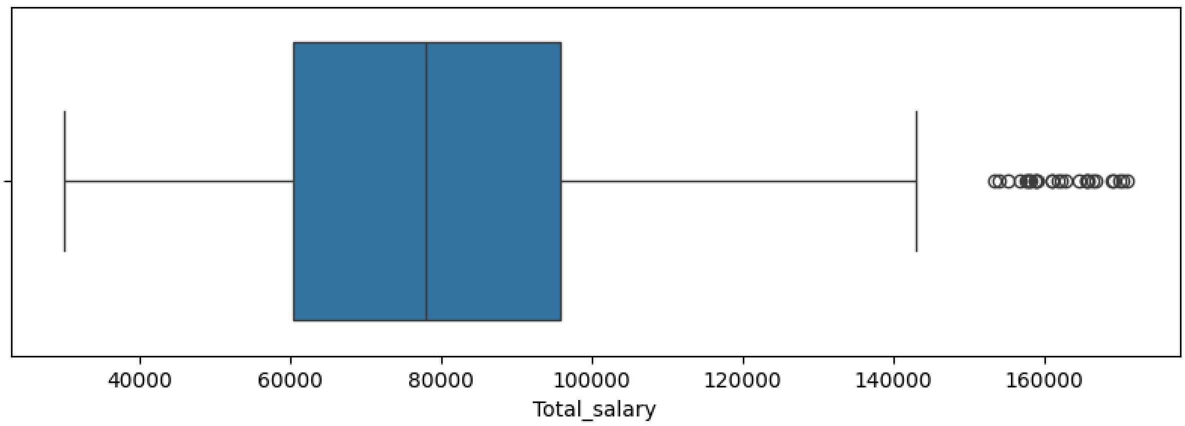
8. Outliers

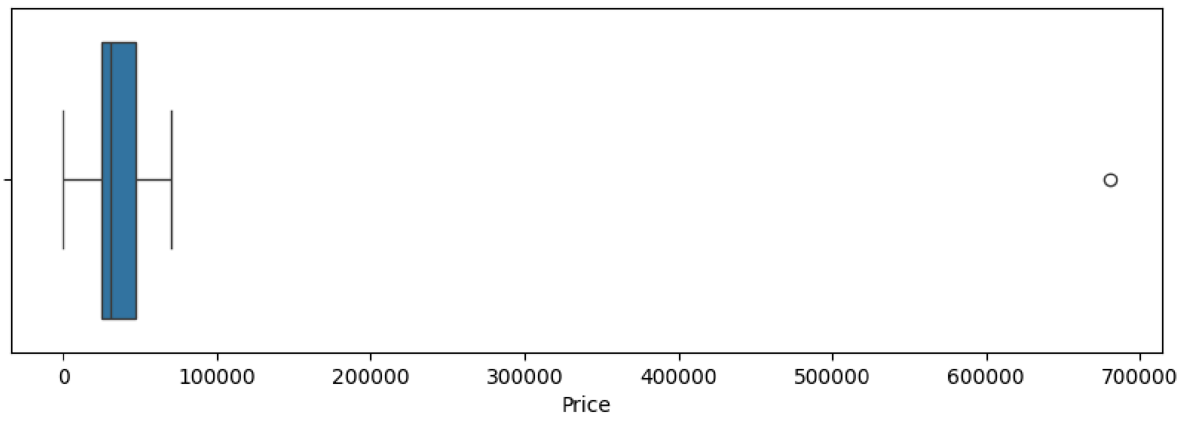
Figure 4




30000
40000
50000
60000
70000
80000
90000
100000
Salary


0
10000
20000
30000
40000
50000
60000
70000
80000
Partner_salary

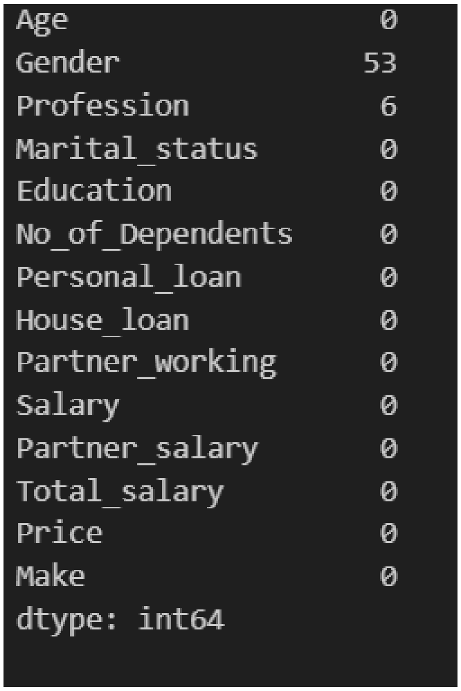

40000
60000
80000
100000
120000
140000
160000
Total_salary


0
100000
200000
300000
400000
500000
600000
700000
Price

9. Data Cleaning

Filling NULL values in numerical columns with medians due to the presence of outliers.

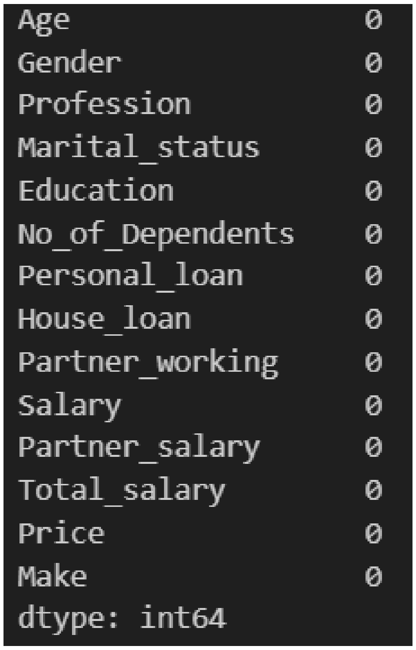
Figure 5



```
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               0
Partner_salary       0
Total_salary         0
Price                0
Make                 0
dtype: int64
```

Treating anomalies present in the 'Gender','No_of_Dependents' and 'Make' columns of the dataset and dealing with their NULL values.

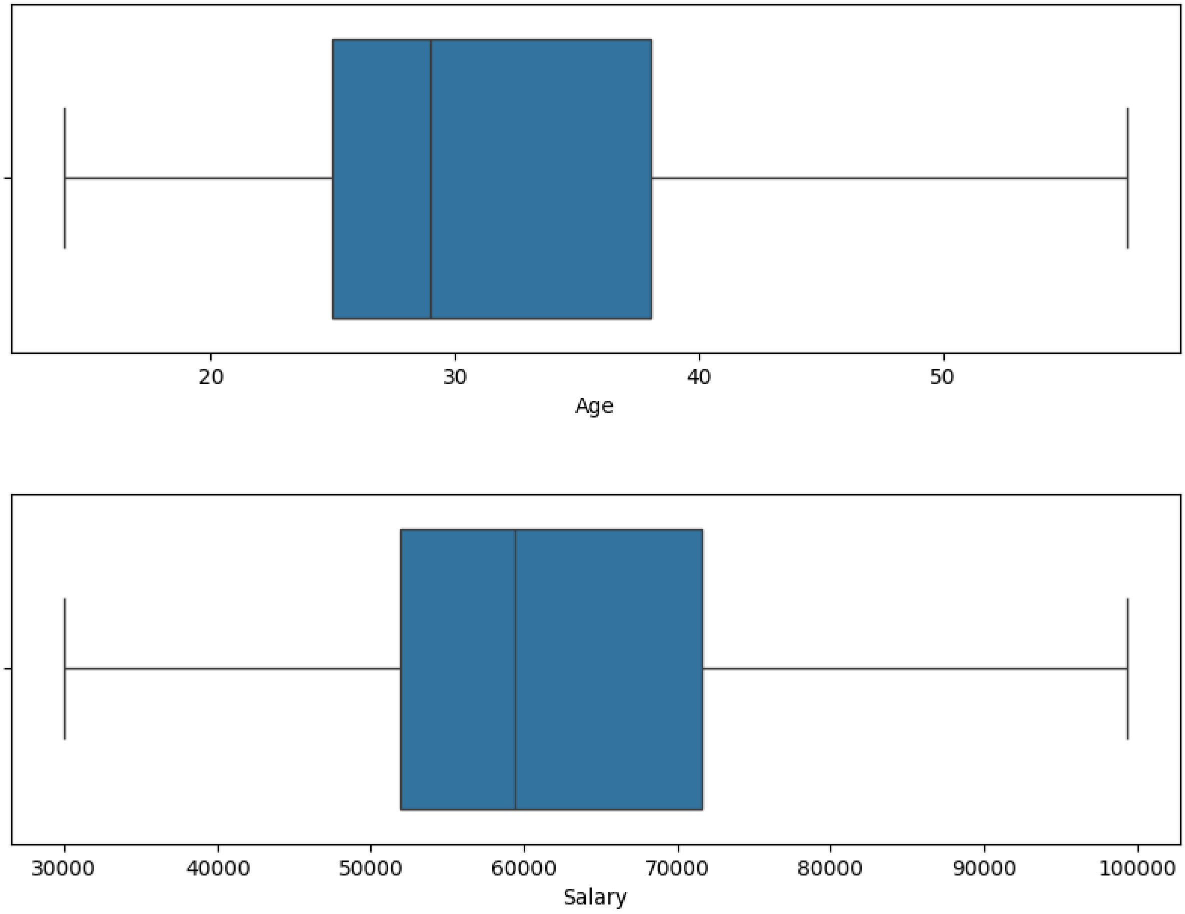
Figure 6

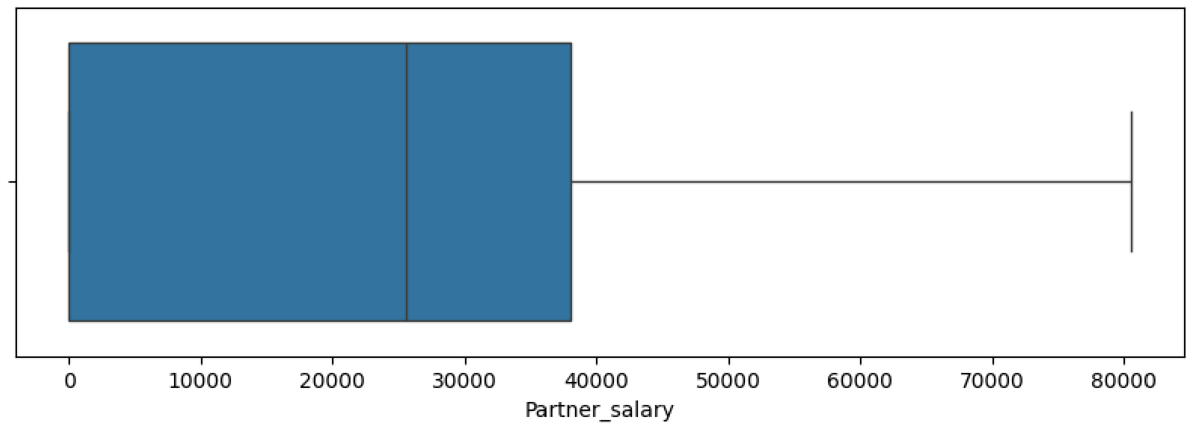


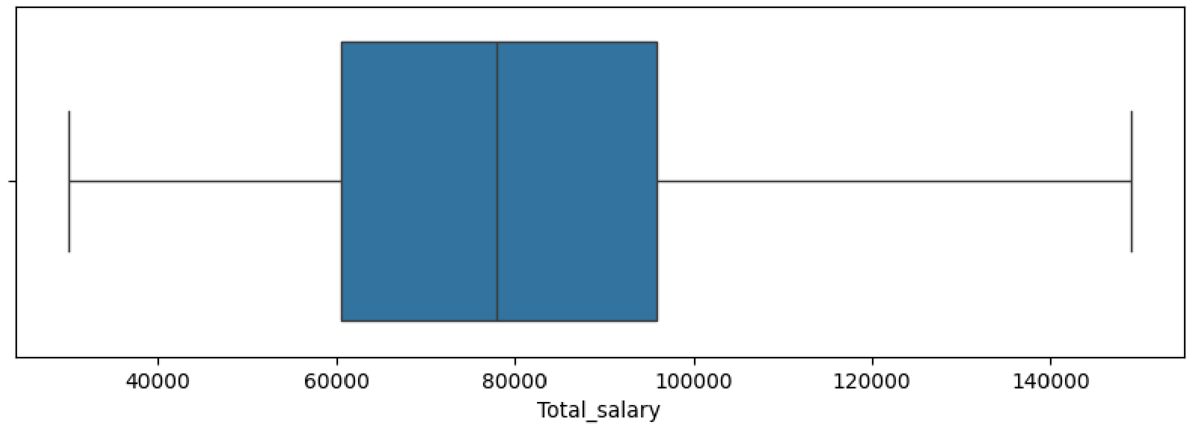
```
Age                 0
Gender              0
Profession          0
Marital_status      0
Education           0
No_of_Dependents    0
Personal_loan       0
House_loan          0
Partner_working     0
Salary              0
Partner_salary      0
Total_salary        0
Price               0
Make      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

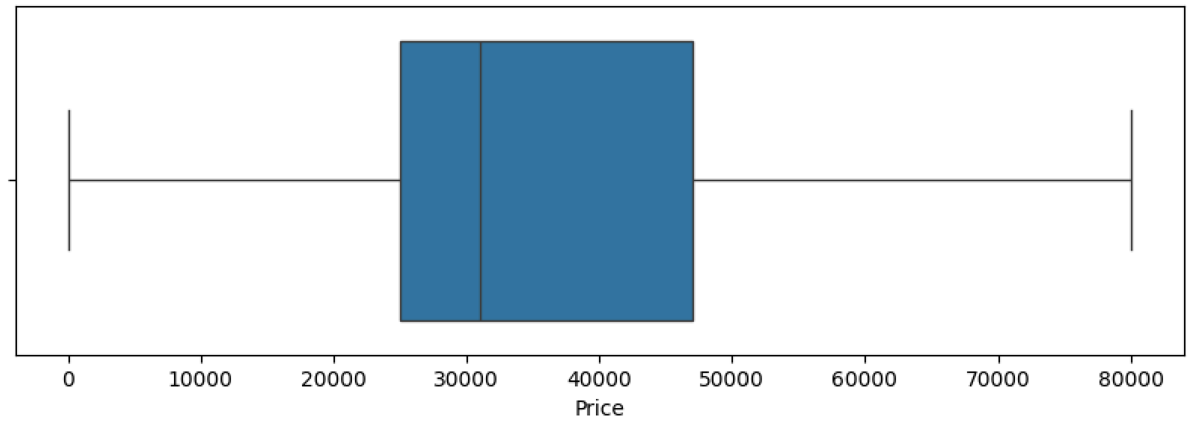
Dealing with outliers present in the data.

Figure 7









Descriptive Statistics

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

The mean is 31.91302972802024

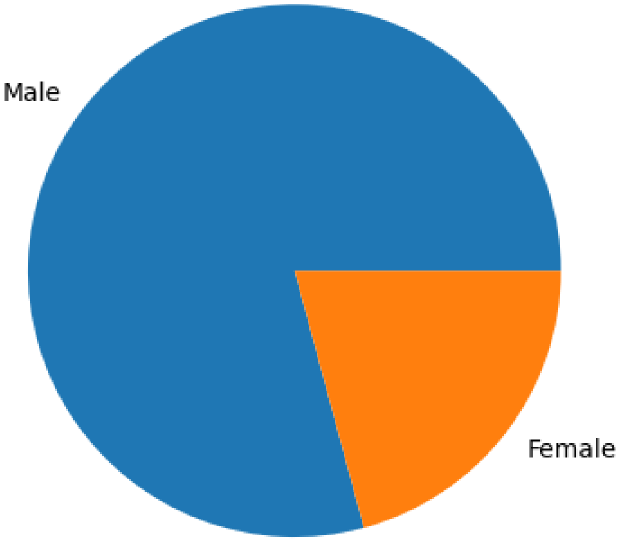
The median is 29.0

The standard deviation is 8.450649424059444

• Data Distribution

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

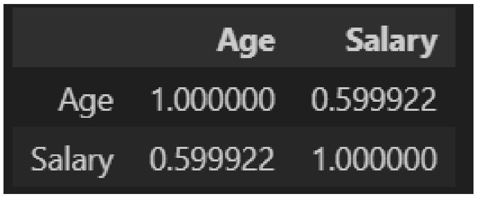
Figure 8



Correlation Analysis

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

Table 4



	Age	Salary
Age	1.000000	0.599922
Salary	0.599922	1.000000

The correlation coefficient is 0.599922.

This means that Age and Salary are not related to each other significantly.

Salary Analysis

o What is the average salary for individuals based on their educational qualifications (Graduate vs. Post

Figure 9

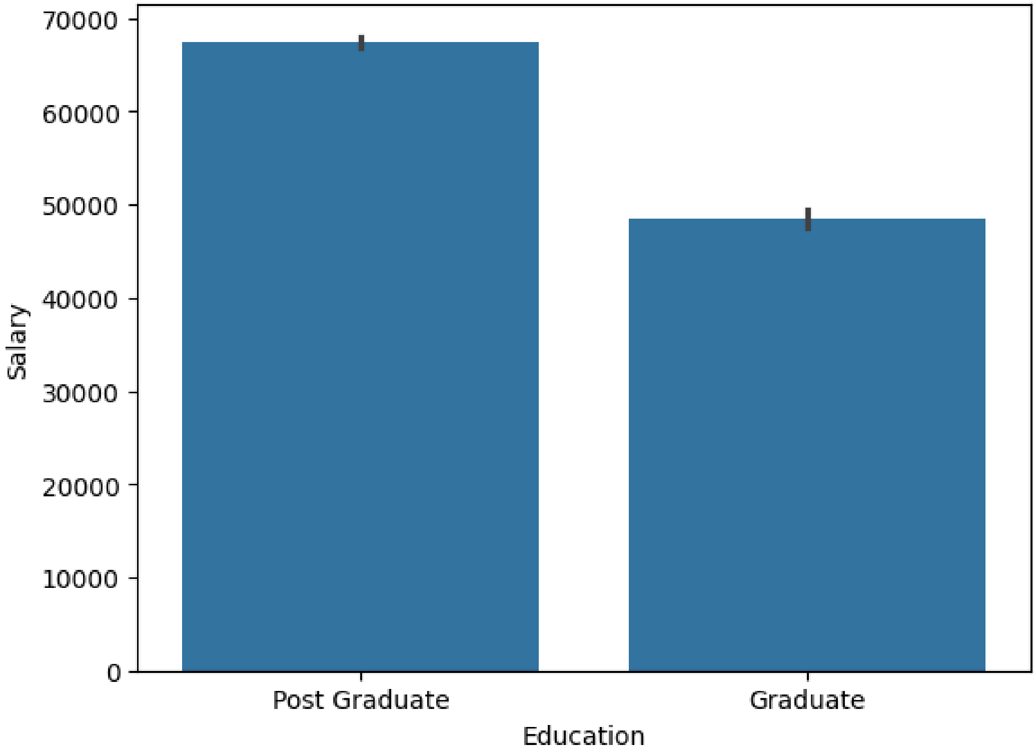
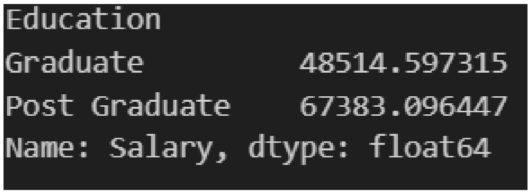


Table 5



```
Education
Graduate         48514.597315
Post Graduate    67383.096447
Name: Salary, dtype: float64
```

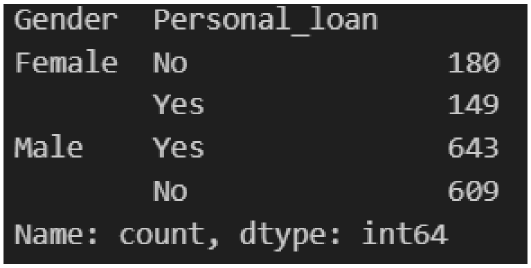
The average salary of Post Graduates is higher than that of Graduates.

Loan Status

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

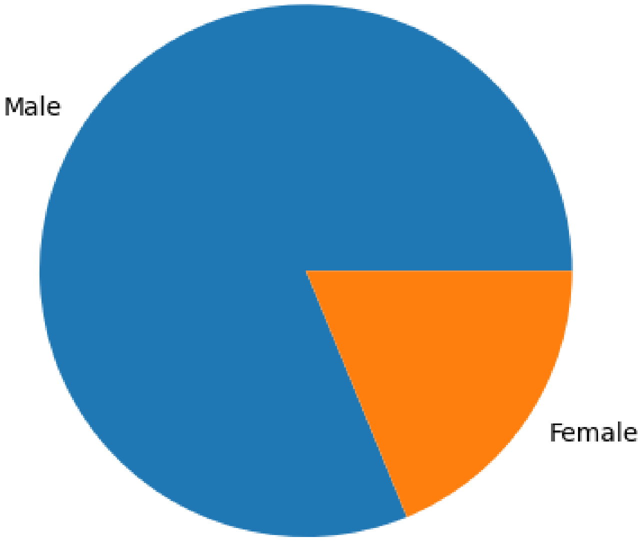
The percentage of people who have taken a personal loan is 50.094876660341555

Table 6



```
Gender  Personal_loan
Female  No              180
        Yes             149
Male    Yes             643
        No              609
Name: count, dtype: int64
```

Figure 10

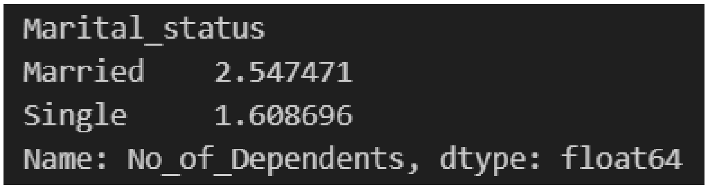


Males have taken more Personal Loans than Females.

Marital Status and Dependents

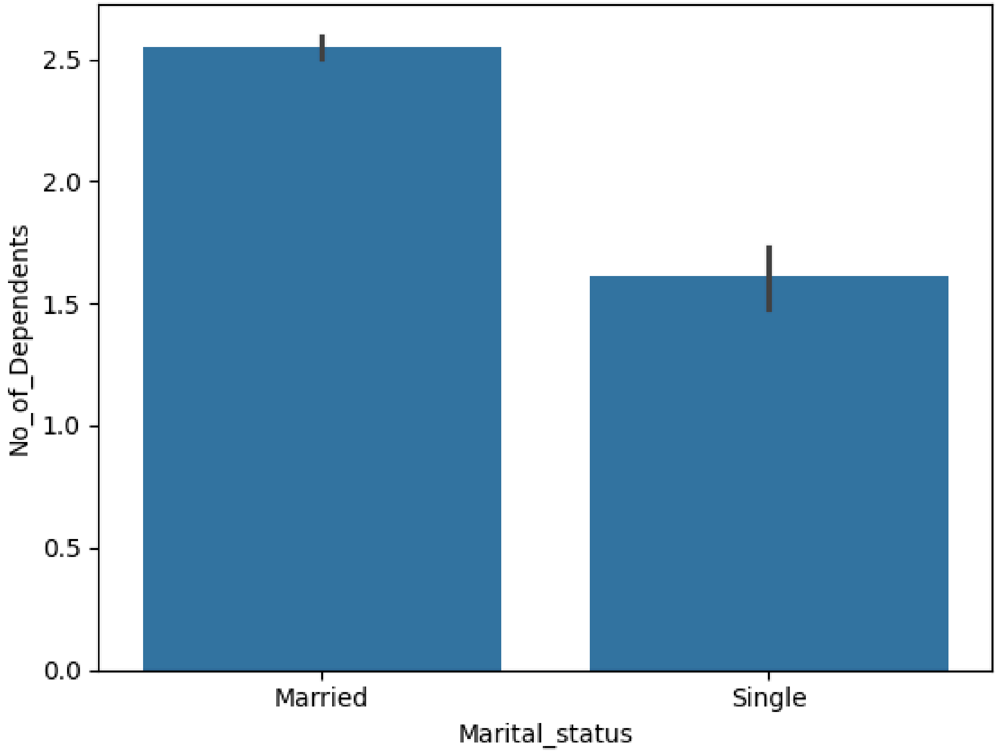
o What is the average number of dependents for married individuals versus single individuals

Table 7



```
Marital_status
Married    2.547471
Single     1.608696
Name: No_of_Dependents, dtype: float64
```

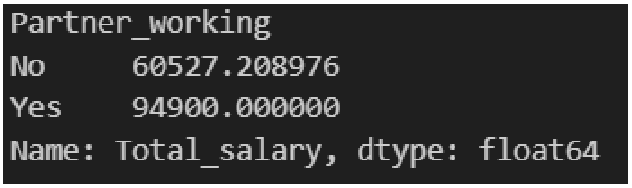
Figure 11



Partner Employment

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

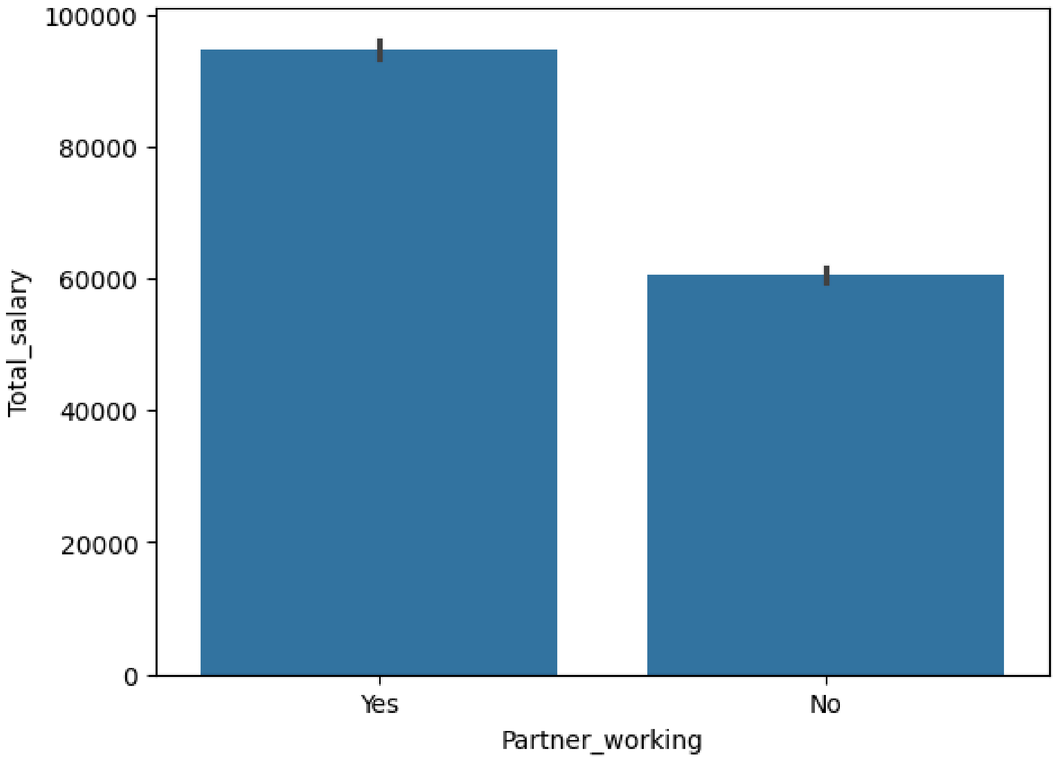
Table 8



```
Partner_working
No     60527.208976
Yes    94900.000000
Name: Total_salary, dtype: float64
```

The total salary is more in cases where the partner is working.

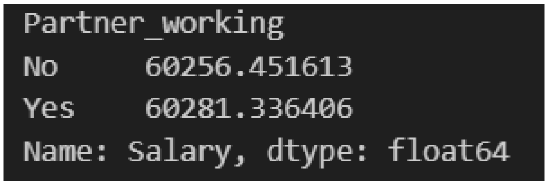
Figure 12



Salary Comparison

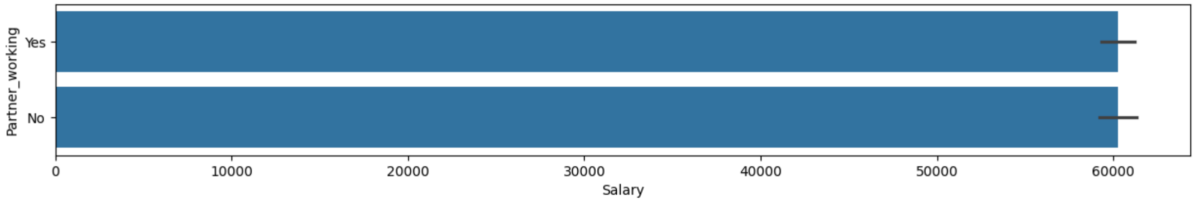
o Compare the average salary of individuals whose partners are working versus those whose partners are not working.

Table 9



```
Partner_working
No      60256.451613
Yes     60281.336406
Name: Salary, dtype: float64
```

Figure 13

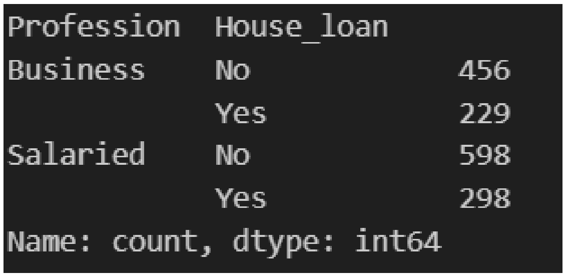


Average salary is slightly more in the case where the partner is working.

House Loan Analysis

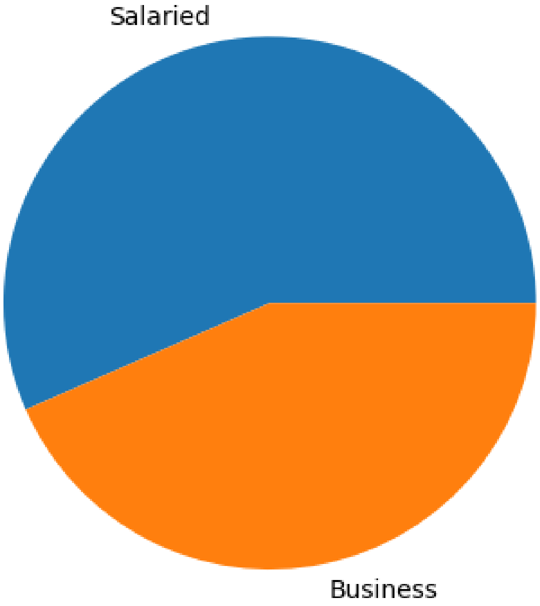
o What is the proportion of individuals with house loans based on their profession?

Table 10



```
Profession  House_loan
Business    No            456
            Yes           229
Salaried    No            598
            Yes           298
Name: count, dtype: int64
```

Figure 14

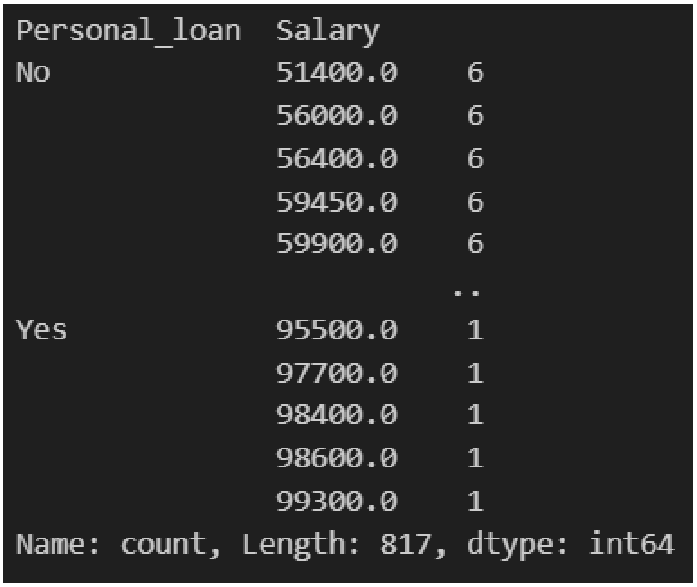


People who are salaried have taken more House Loans than people who are salaried.

Salary Distribution

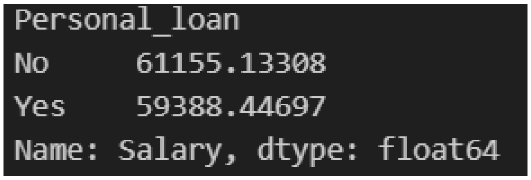
o What is the distribution of salaries for individuals with personal loans versus those without personal loans? Represent it using a box plot.

Table 11



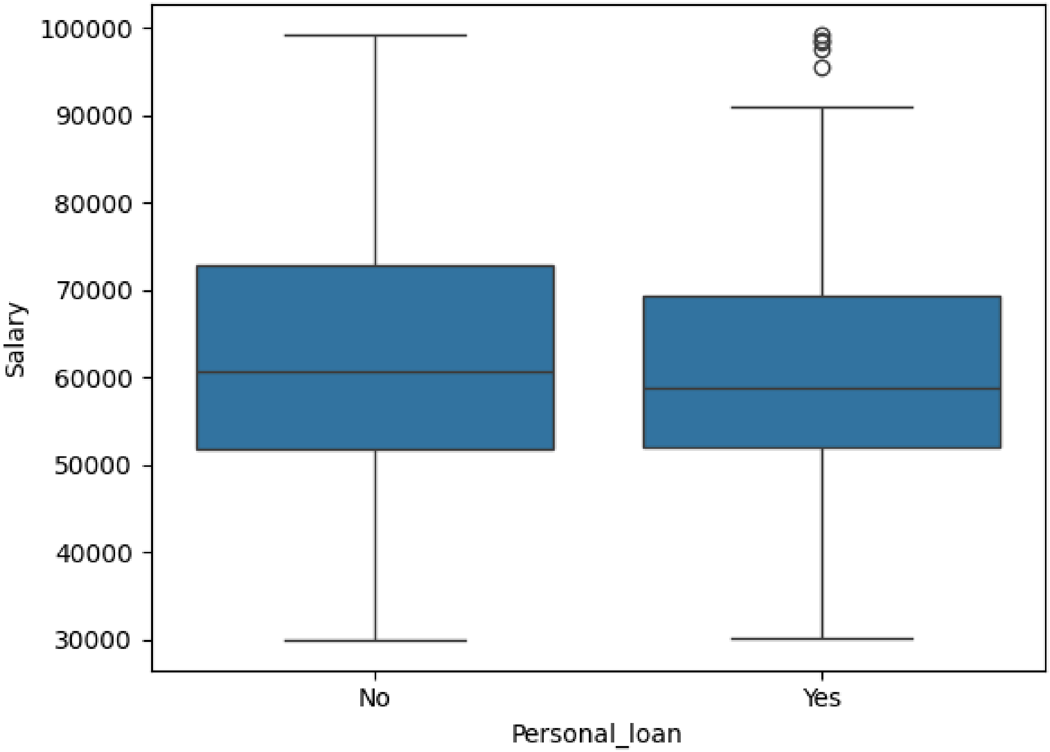
```
Personal_loan  Salary
No             51400.0    6
               56000.0    6
               56400.0    6
               59450.0    6
               59900.0    6
                         ..
Yes            95500.0    1
               97700.0    1
               98400.0    1
               98600.0    1
               99300.0    1
Name: count, Length: 817, dtype: int64
```

Table 12



```
Personal_loan
No     61155.13308
Yes    59388.44697
Name: Salary, dtype: float64
```

Figure 15

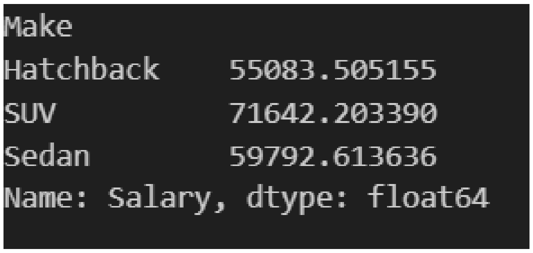


Salary of people with no personal loan is on average greater than those with personal loans. Although there are a few people with high salaries who have a personal loan.

Automobile Make Analysis:

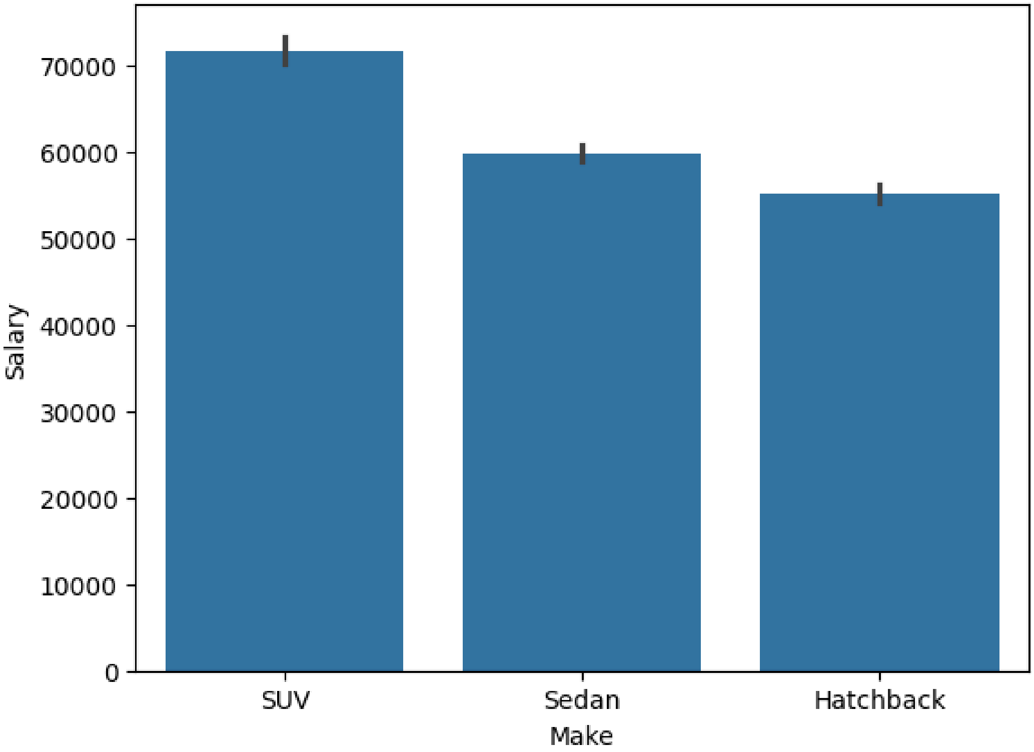
o How does the type of automobile relate to the salary of the individuals? Provide insights based on the make of the automobile.

Table 13



```
Make
Hatchback    55083.505155
SUV          71642.203390
Sedan        59792.613636
Name: Salary, dtype: float64
```

Figure 16



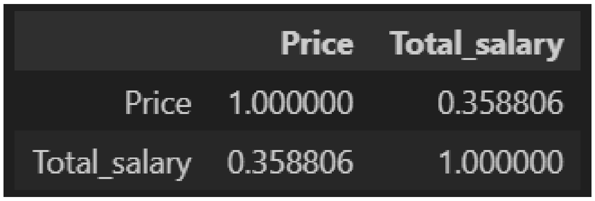
People who have SUVs have the highest salaries on average. Sedan Makes are a second while Hatchbacks are the lowest.

Price Analysis

o What is the average price of the product/service in the dataset? How does this price vary based on the

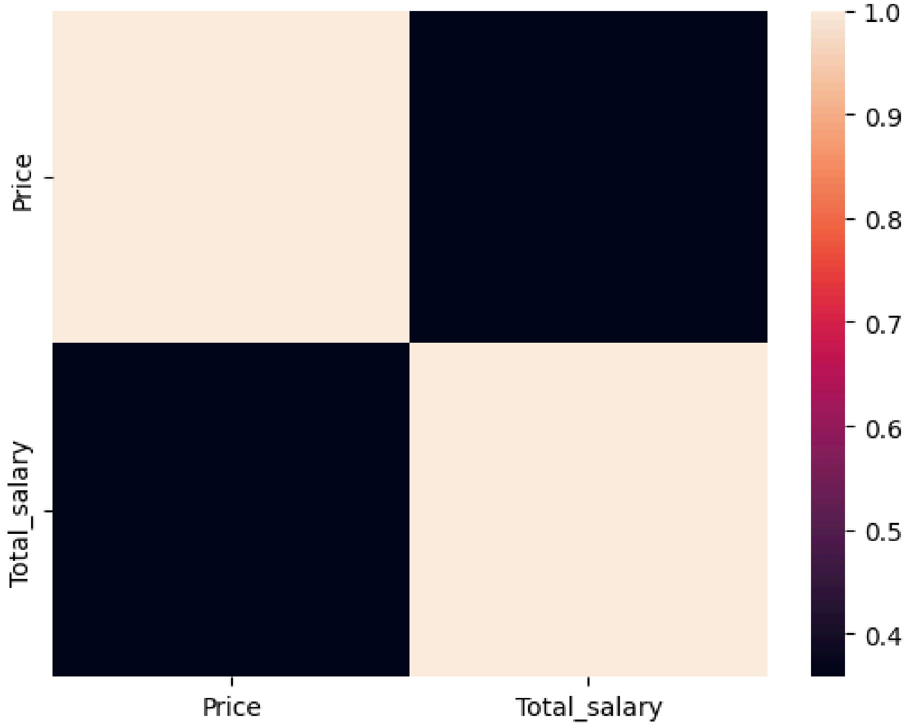
The average price of the products is 35568.66413662239

Table 14



	Price	Total_salary
Price	1.000000	0.358806
Total_salary	0.358806	1.000000

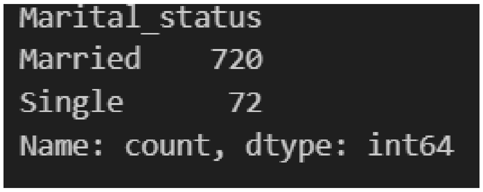
Figure 17



The price of the products do not vary much(close to almost none) with the salaries of the individual. The correlation coefficient between them is 0.358806

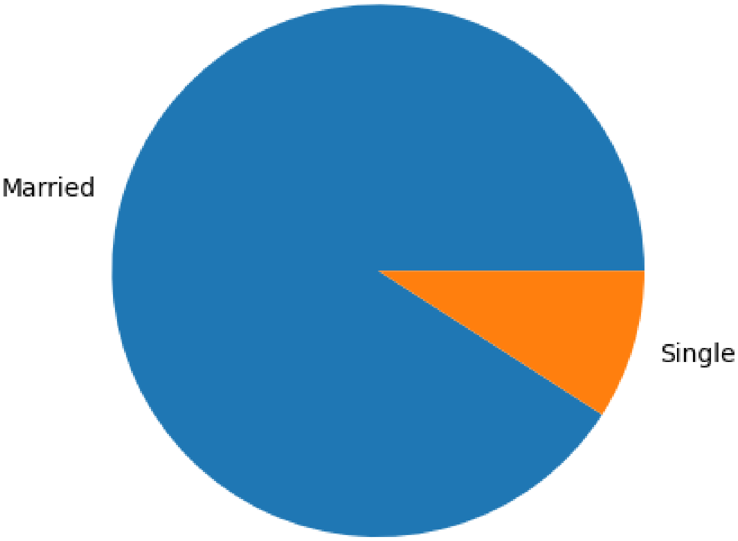
Marital Status and Loans

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



```
Marital_status
Married     720
Single       72
Name: count, dtype: int64
```

Figure 18

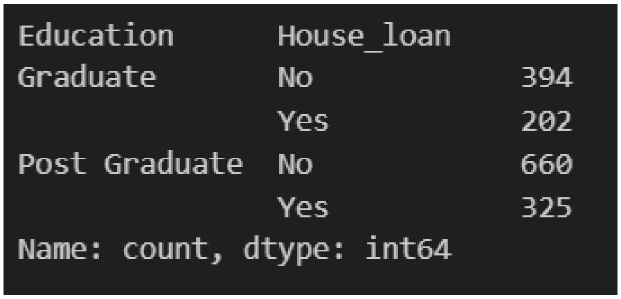


There is a significant difference. People who have married have taken more personal loans than people who are single.

Educational Qualification Impact

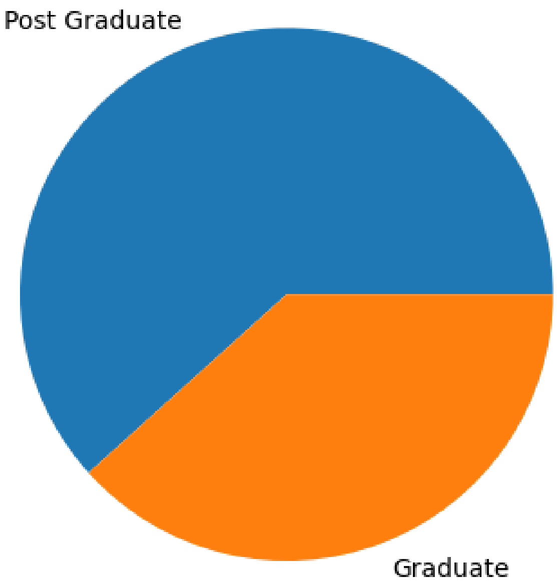
o How does educational qualification impact the likelihood of taking a house loan?

Table 15



```
Education      House_loan
Graduate       No            394
               Yes           202
Post Graduate  No            660
               Yes           325
Name: count, dtype: int64
```

Figure 19

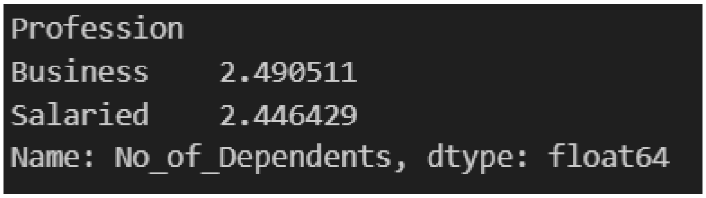


Post Graduates are more likely to take a house loan.

Dependent Count Analysis

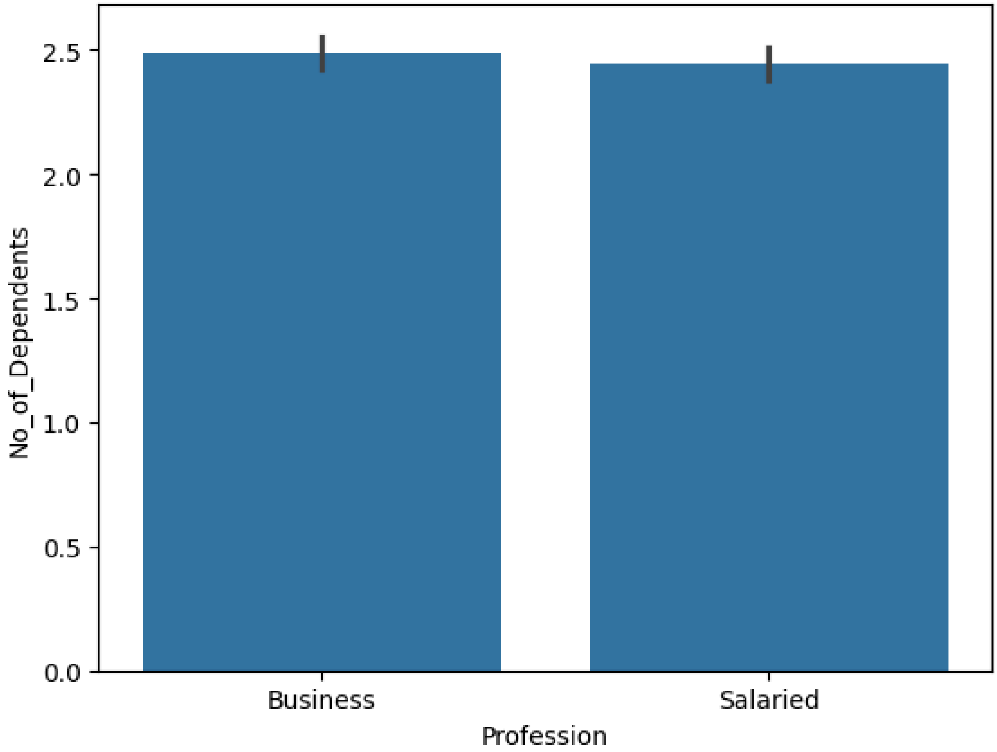
o Analyze the number of dependents based on the profession of the individual. Which profession has the highest average number of dependents?

Table 16



```
Profession
Business    2.490511
Salaried    2.446429
Name: No_of_Dependents, dtype: float64
```

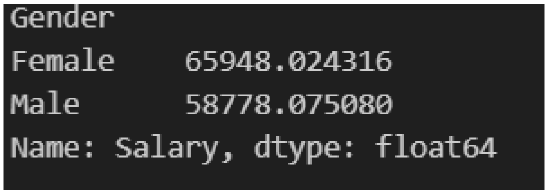
Figure 20



Gender and Salary

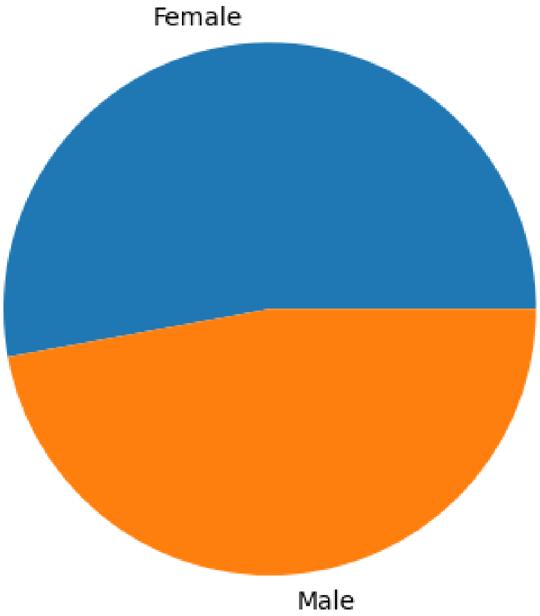
o Is there a significant difference in salaries between males and females? Provide statistical evidence.*

Table 17



```
Gender
Female    65948.024316
Male      58778.075080
Name: Salary, dtype: float64
```

Figure 21

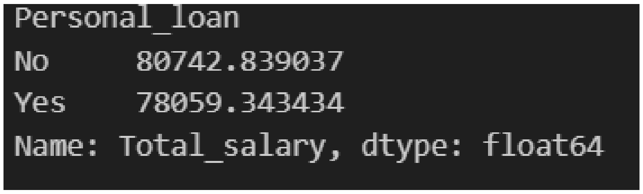


Females have a higher average salary than males, although it is not very significant.

Loan Status Impact

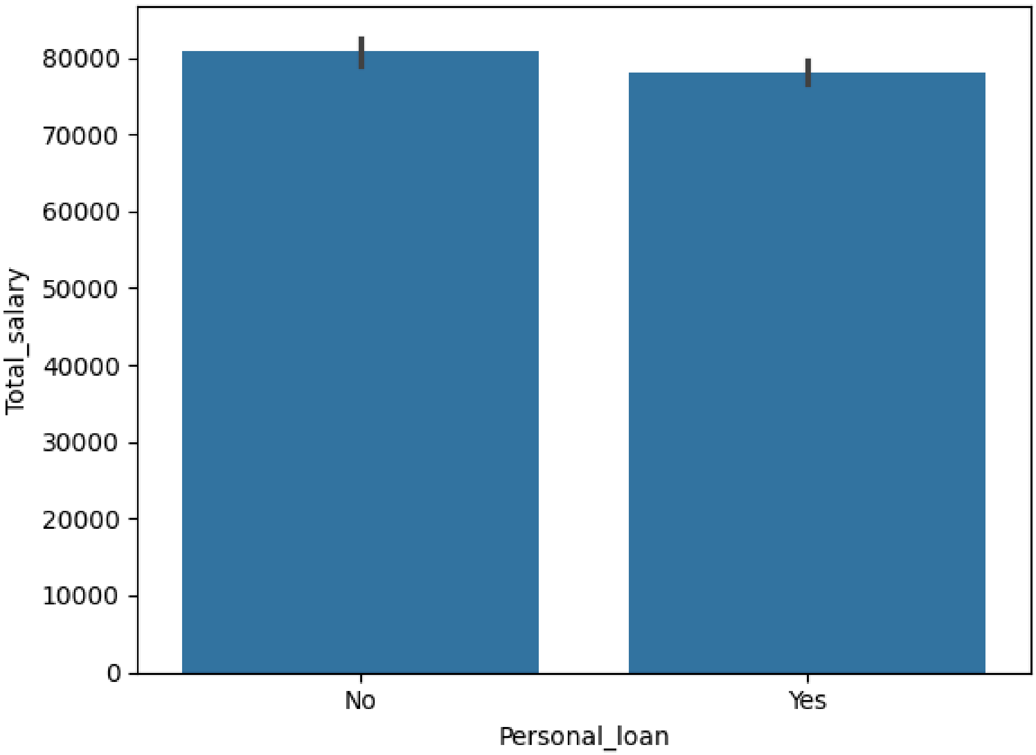
o How does having a personal loan affect the total combined salary of the individual and their partner?

Table 18



```
Personal_loan
No     80742.839037
Yes    78059.343434
Name: Total_salary, dtype: float64
```

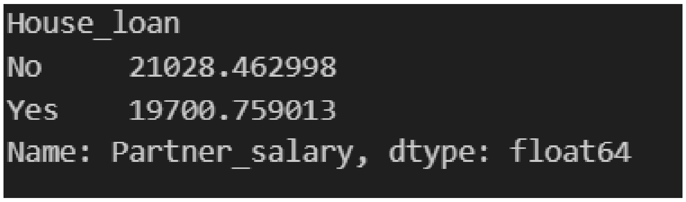
Figure 22



Partner's Salary Contribution

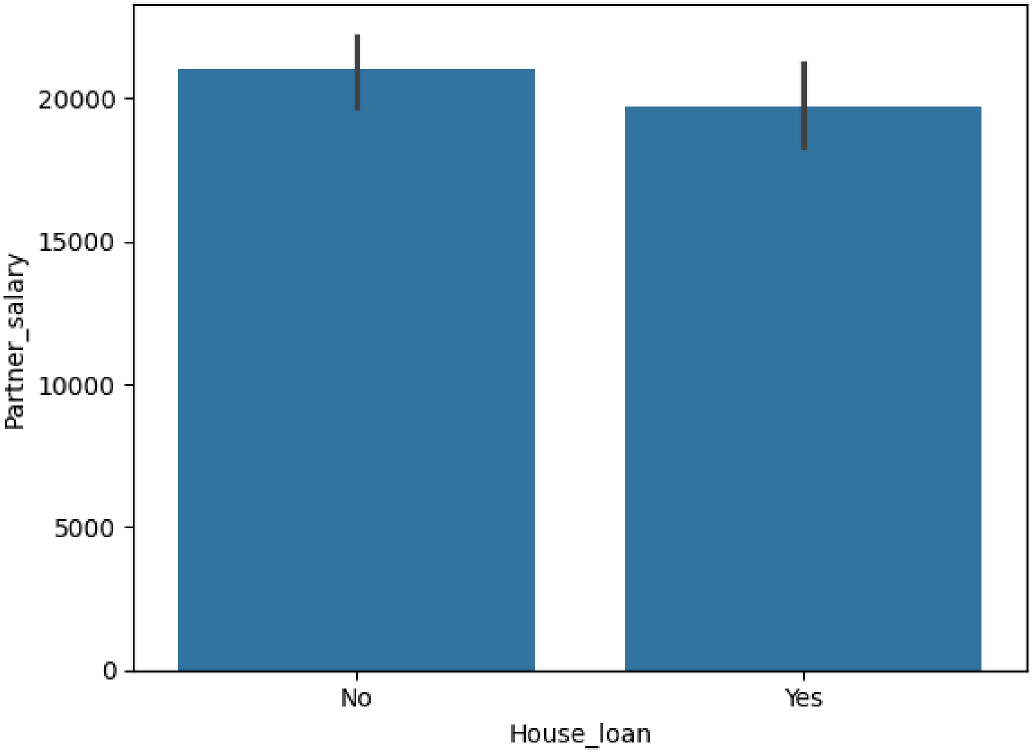
o What is the average partner's salary for individuals with and without house loans?

Table 19



```
House_loan
No     21028.462998
Yes    19700.759013
Name: Partner_salary, dtype: float64
```

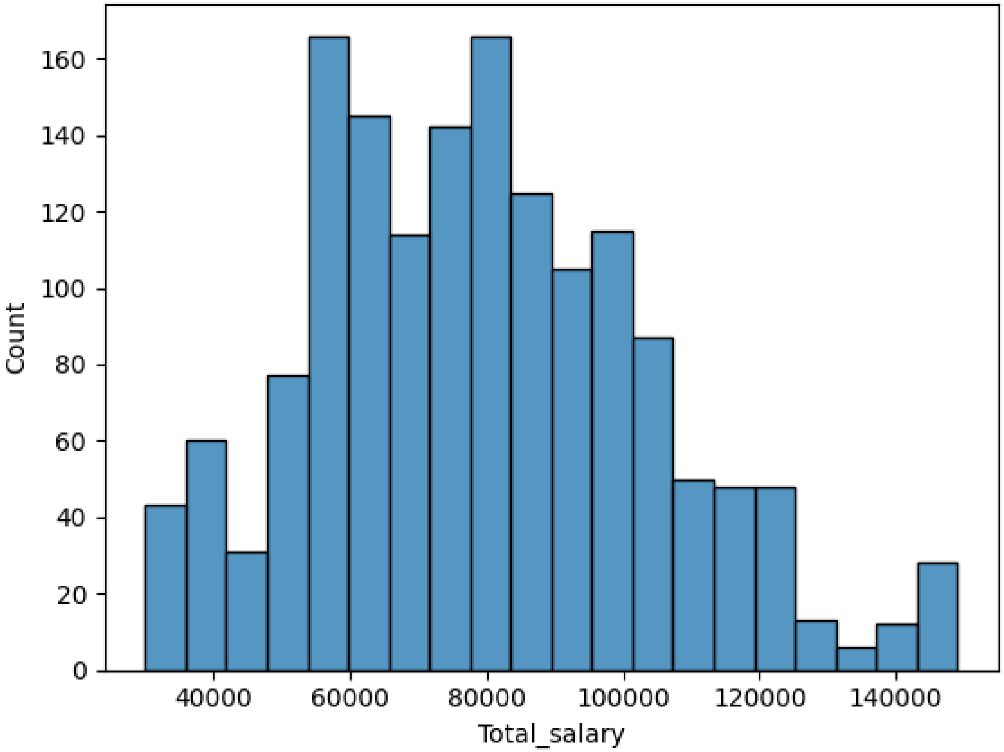
Figure 23



Total Salary Distribution

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

Figure 24


Count
160
140
120
100
80
60
40
20
0
40000
60000
80000
100000
120000
140000
Total_salary