SMDM PROJECT– BUSINESS REPORT

HARIHARASUDHAN VENUGOPALAN

02-05-2024

Contents

Summary

The document contains the business report for the Statistical Methods for Decision Making (SMDM) project containing Problem 1 and Problem 2. Problem 1 is based on the Austo Motor company's marketing campaign and Problem 2 is based on the GODIGT Bank data for identifying key variables.

The business report explains output from the python as table and graphs which are explained in detail.

Problem 1

Context

Analysts are required to explore data and reflect on the insights. Clear writing skill is an integral part of a good report. Note that the explanations must be such that readers with minimum knowledge of analytics are able to grasp the insight.

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 analytics professional to improve the existing campaign.

Objective

They 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 Postgraduate
- **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 Data Overview

- Import the libraries - Load the data - Check the structure of the data - Check the types of the data - Check for and treat (if needed) missing values - Check the statistical summary - Check for and treat (if needed) data irregularities - Observations and Insights

1.1.A Import the Libraries

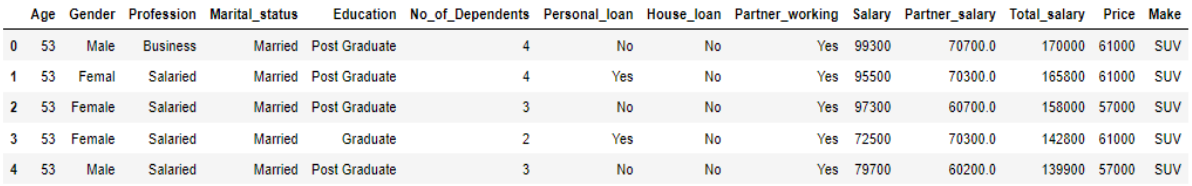
Import all the necessary libraries that are required for the execution of the scripts and plots such as pandas, numpy, matplotlib and seaborn.

1.1.B Load the data

Depending on the working directory set the right parameters to import the austo_automobile data that was provided.

1.1.C Check the Structure of the data.

Data set has 1581 rows and 14 columns. It's a good idea to check the sample or the top 5 rows in the dataset using head function.

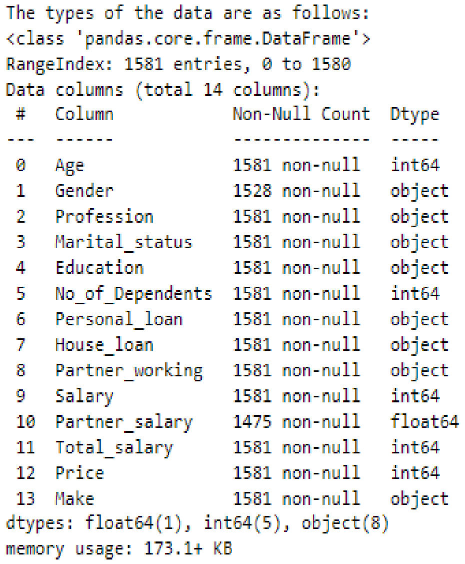


	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary	Price	Make
0	53	Male	Business	Married	Post Graduate	4	No	No	Yes	99300	70700.0	170000	61000	SUV
1	53	Femal	Salaried	Married	Post Graduate	4	Yes	No	Yes	95500	70300.0	165800	61000	SUV
2	53	Female	Salaried	Married	Post Graduate	3	No	No	Yes	97300	60700.0	158000	57000	SUV
3	53	Female	Salaried	Married	Graduate	2	Yes	No	Yes	72500	70300.0	142800	61000	SUV
4	53	Male	Salaried	Married	Post Graduate	3	No	No	Yes	79700	60200.0	139900	57000	SUV

Fig1: Sample dataset output using head function.

1.1.D Check the types of the data.

Following figure shows the types of data. The data set contains 6 numerical variables and 8 variables with Object datatype. There are null values in Gender and Partner_salary fields.

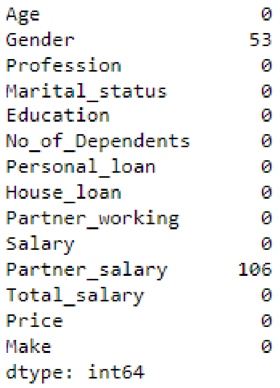


```
The types of the data are as follows:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1581 entries, 0 to 1580
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   Age               1581 non-null   int64  
 1   Gender            1528 non-null   object 
 2   Profession        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)
memory usage: 173.1+ KB
```

Fig2: Dataset information on the types of data

1.1.E Check for and treat (if needed) missing values.

Step 1: Check for missing values.



```
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
```

Fig3: Missing values information in the data

There are 53 missing values in the Gender field and 106 missing values in the Partner_salary field.

Since the total number of missing values in Gender is only 3% of the population and Partner salary is only 6% of the population, it is required to impute the values.

Step 2: Handle missing values in Gender using mode as it is a categorical variable. Use the mode function to impute the values as **Male** as it’s the most occurring value.

Step 3: Handle missing values in Partner_working using conditions and correlation.

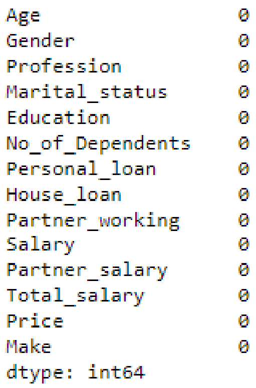
On checking the sample data, the following analysis was made:

Total_salary = Salary + Partner_salary

Also, the variable Partner_working can be used as a condition to determine the value of Partner_working where it is missing:

1. If the partner is working (Partner_working = ‘Yes’) then the Partner_salary = Total_salary – salary
2. If the partner is not working (Partner_working = ‘No’) then Partner_salary = 0

The above logic was used to impute the missing values in the Partner_salary field. Compare fig 3 above and fig 4 below to compare the missing values imputation.

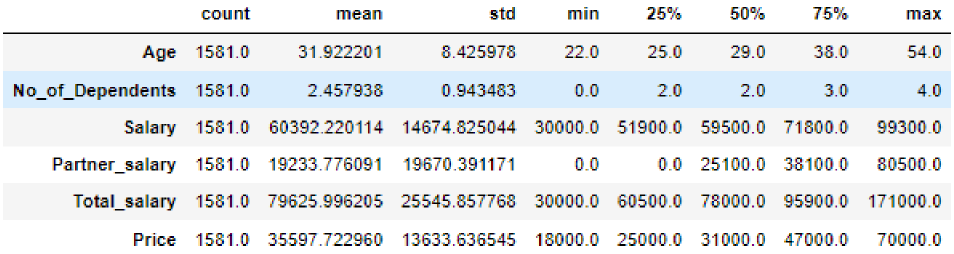


```
Age                 0
Gender              0
Profession          0
Marital_status      0
Education           0
No_of_Dependents    0
Personal_loan       0
House_loan          0
Partner_working     0
Salary              0
Partner_salary      0
Total_salary        0
Price               0
Make      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

Fig4: Missing values imputed in the data.

1.1.F Check the statistical summary.

Using the describe function, check for the statistical summary of the data set.



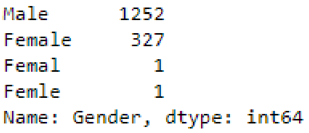
	count	mean	std	min	25%	50%	75%	max
Age	1581.0	31.922201	8.425978	22.0	25.0	29.0	38.0	54.0
No_of_Dependents	1581.0	2.457938	0.943483	0.0	2.0	2.0	3.0	4.0
Salary	1581.0	60392.220114	14674.825044	30000.0	51900.0	59500.0	71800.0	99300.0
Partner_salary	1581.0	19233.776091	19670.391171	0.0	0.0	25100.0	38100.0	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

Fig5: statistical summary of the data.

1.1.G Check for and treat (if needed) data irregularities.

Categorical Variables treatment:

To treat the data irregularities, one needs to identify the irregularities in the data. Using value_counts() function on all the categorical variables, it was determined that the Gender field contains irregularities. Other categorical variables did not show any irregularities.

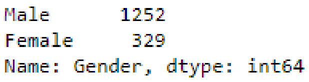


```
Male      1252
Female     327
Femal        1
Femle        1
Name: Gender, dtype: int64
```

Fig6: value counts for Gender Variable

The word Female is misspelled as 'Femal' and 'Femle'. This needs to be treated by replacing the misspelled words appropriately.

Post treatment the value counts of the field Gender changed as shown below.



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

Numerical variables treatment:

To identify the irregularities in the numerical variables, box plots were created for the numerical fields as shown below.

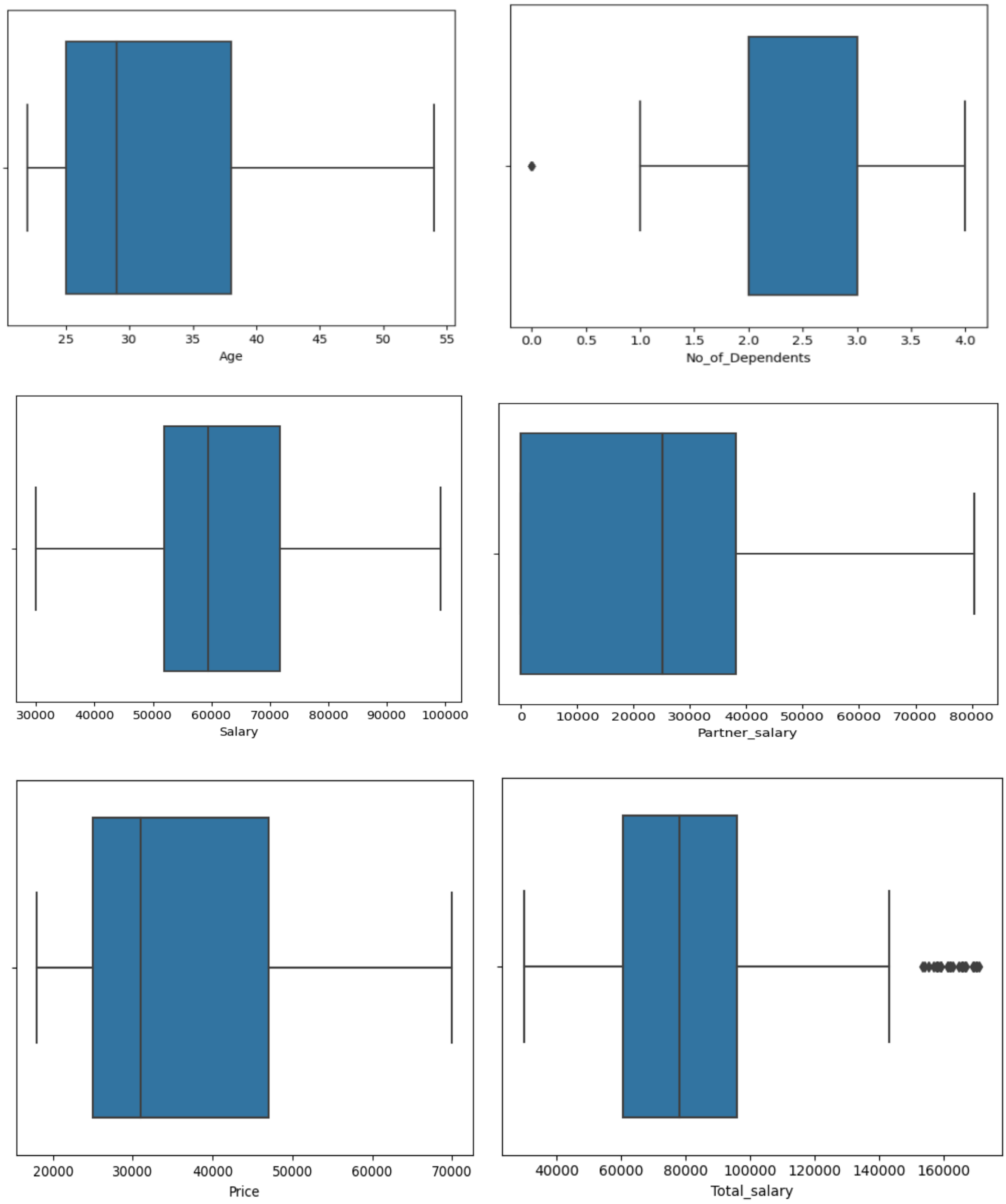


Fig7: Box plots for numerical fields

Based on the above charts, the Total_salary variable contains outliers, and all the other variables does not show any outliers.

To treat the outliers, present in the Total_salary variable, lets create a copy of the dataset (df_new) to see the pre and post outlier treatments.

Create a function to detect outliers with the following logic:

lower_range = Q1-(1.5*IQR)

upper_range = Q3+(1.5*IQR)

based on this the lower range and upper range is calculated as:

7400.0 and upper range 149000.0

Here the upper range is imputed into the outlier values.

Post Transformation, the box plot is shown as below for Total_salary.

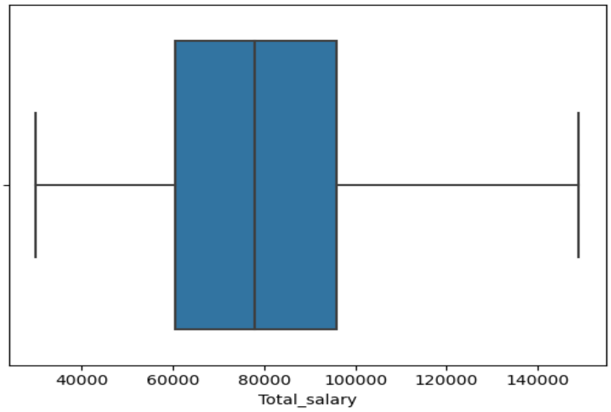
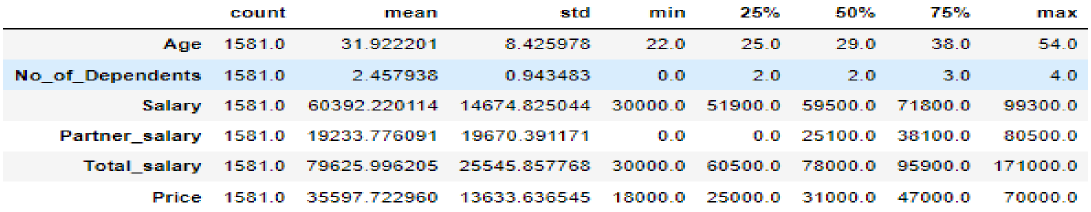


Fig7: Box plots for Total_salary post outlier treatment.

1.1.H Observations and Insights.



	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	19233.776091	19670.391171	0.0	0.0	25100.0	38100.0	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 of the customers are between 22 and 54.
- Mean salary of the customers is 60392 and Median salary is 59500 and do not have much difference.
- Total salaries are ranging between a minimum of 30000 and 171000, whereas the starting price of the cars are at 18000.
- Most of the variable are skewed and do not follow normal distribution.

1.2 Univariate Analysis

- Explore all the variables (categorical and numerical) in the data - Check for and treat (if needed) outliers
- Observations and Insights

1.2.A Explore all the variables (categorical and numerical) in the data.

The dataset df_new is used for all further analysis as it has been treated for outliers and discrepancies.

1.2.A.1 Numerical variables – Univariate analysis:

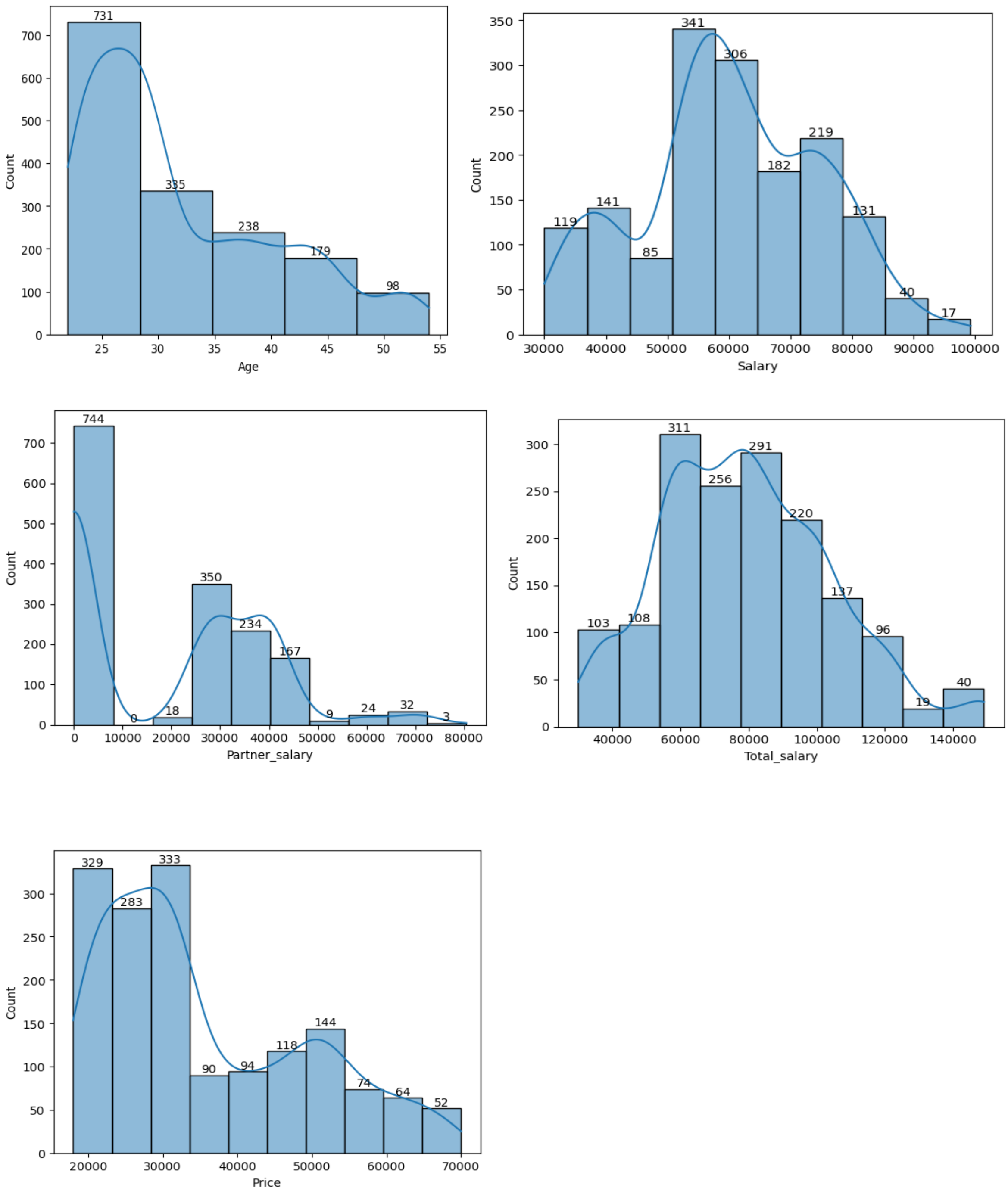
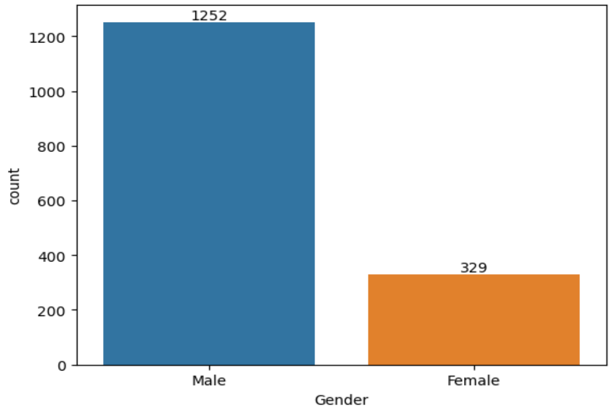


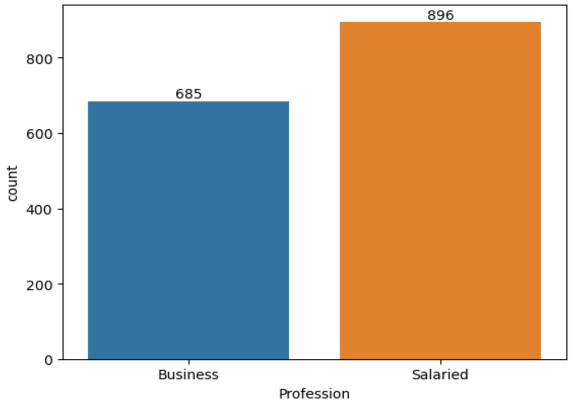
Fig8: Univariate analysis of numerical variables

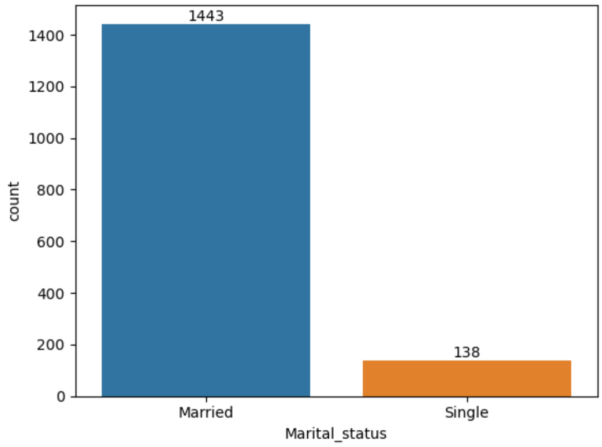
<u>Observations and Insights</u>:

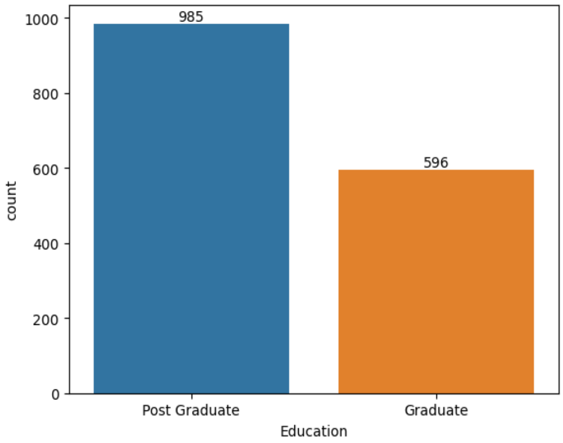
1. Age is positively skewed with lower the age the capacity of buying a car is higher.
2. Salary has a multimodal distribution, with more people in the range 50000 to 70000.
3. Price is positively skewed as lower the price, larger the number of car sales.
4. Skewness of Total_salary has reduced significant post outlier treatment.
5. Most of the variable are skewed and do not follow normal distribution.

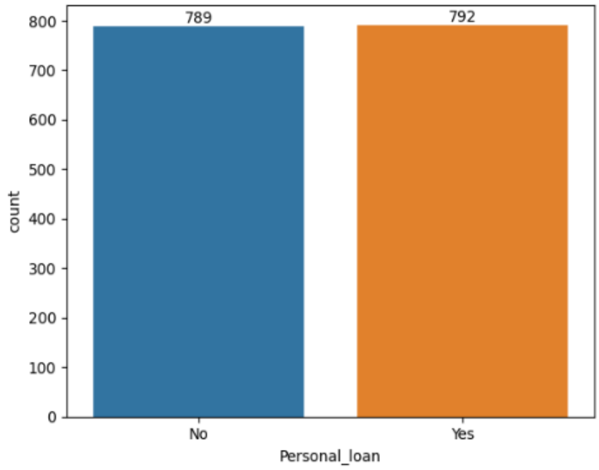
1.2.A.2 Categorical variables – Univariate analysis:

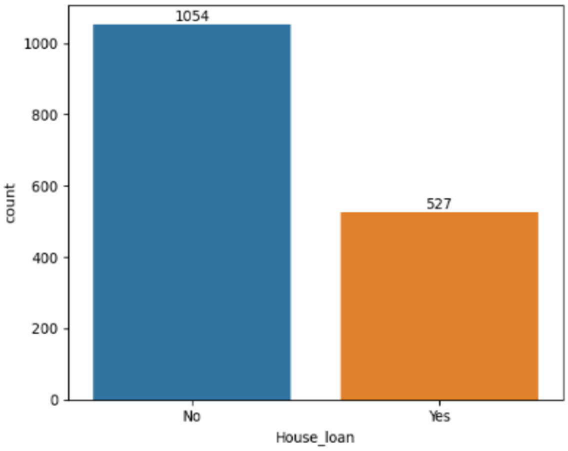












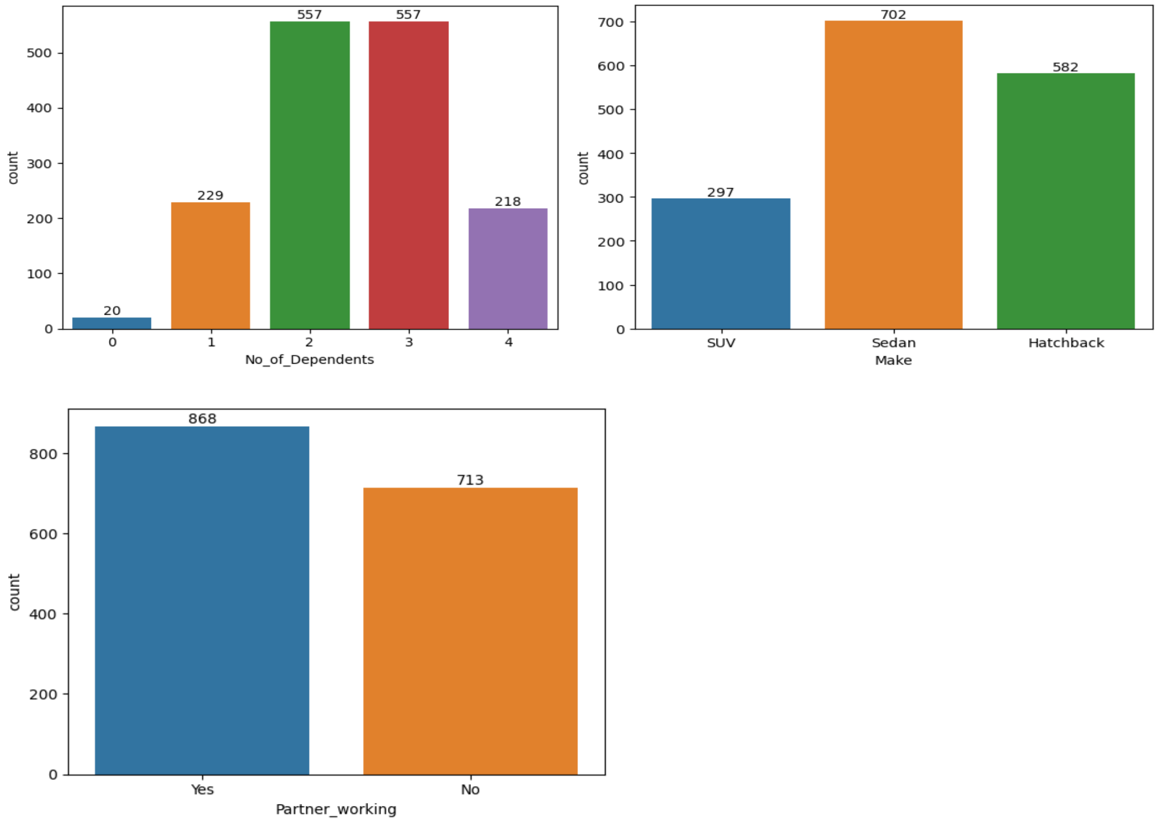


Fig9: Univariate analysis of categorical variables

Observations and Insights:

1. Number of male owners are more than the females.
2. Salaried people are a better target than business customers as the count is higher.
3. Data contains more married customers compared to single customers.
4. Customers having home loan is half the customers not having a home loan.
5. Customers prefer Sedan the most as compared to SUV and hatchback.
6. The count of customers with or without a personal loan is almost the same.
7. Postgraduates have a higher count for buying cars as compared to Graduates.
8. Working partners are slightly higher than the non-working partners.
9. Most customer have 2 or 3 dependents. Very few have no dependents.

1.2.B Check for and treat (if needed) outliers.

Since we have already treated the outliers in the previous step, there is no treatment required in this step.

1.2.C Observations and Insights.

Observations and insights are added in 1.2.A below the plots individually.

1.3 Bivariate Analysis

1.3.A Explore the relationship between all numerical variables.

To explore the relationship between all numerical variables we can use a pair plot.

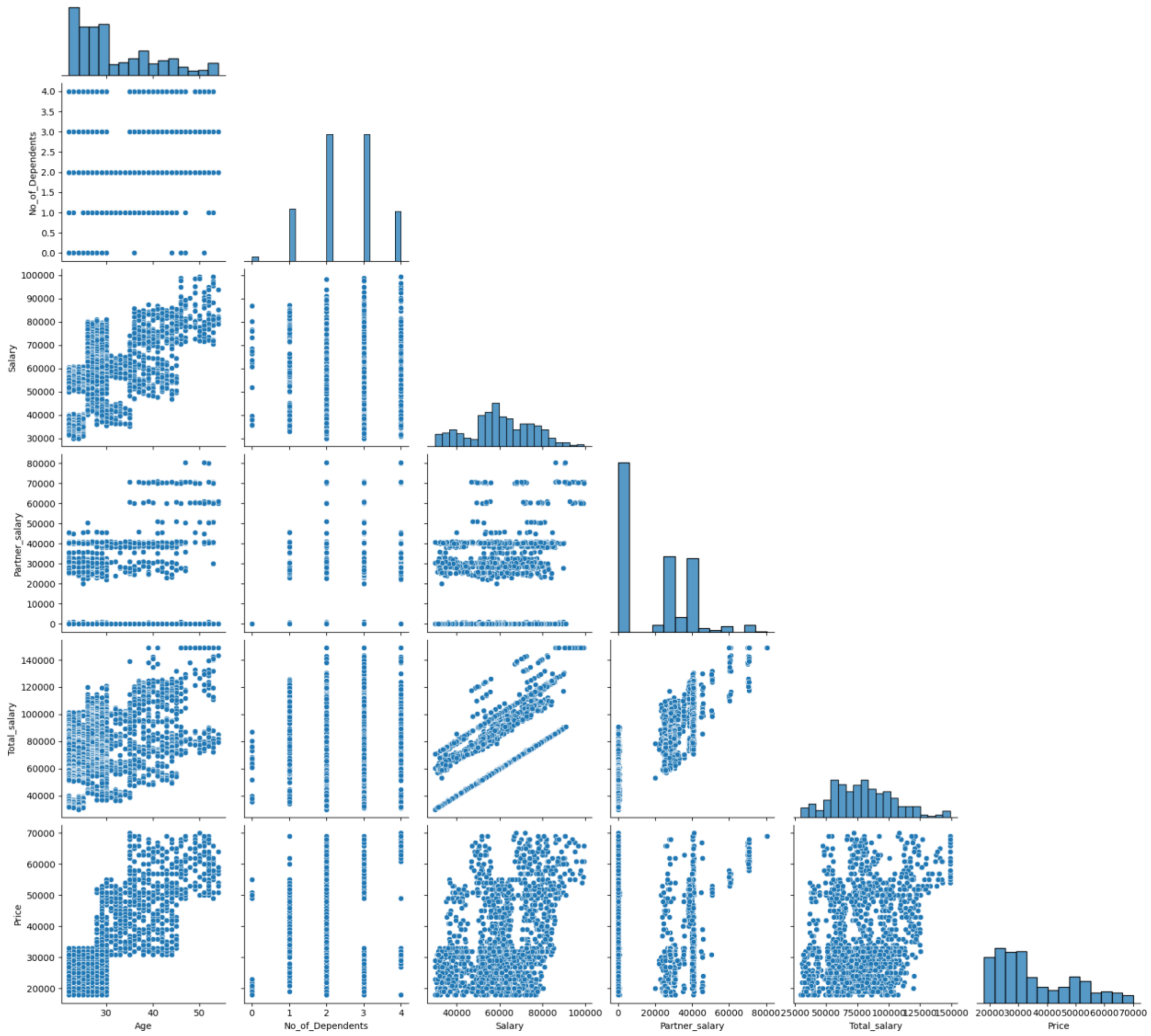


Fig10: Bivariate analysis of numerical variables with outlier treatment in Total_salary

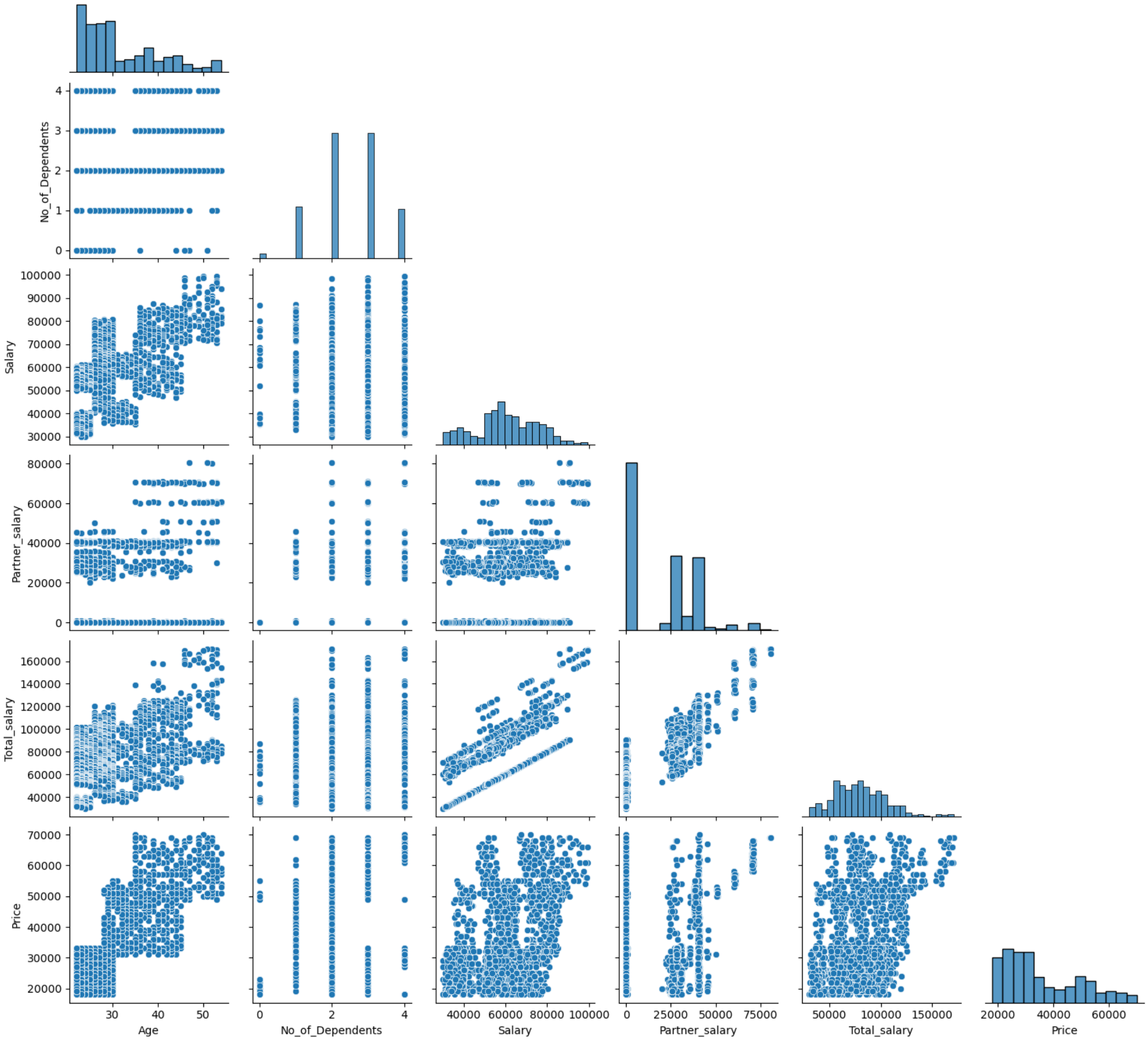


Fig11: Bivariate analysis of numerical variables without outlier treatment in Total_salary

1.3.B Explore the correlation between all numerical variables.

Without outlier treatment:

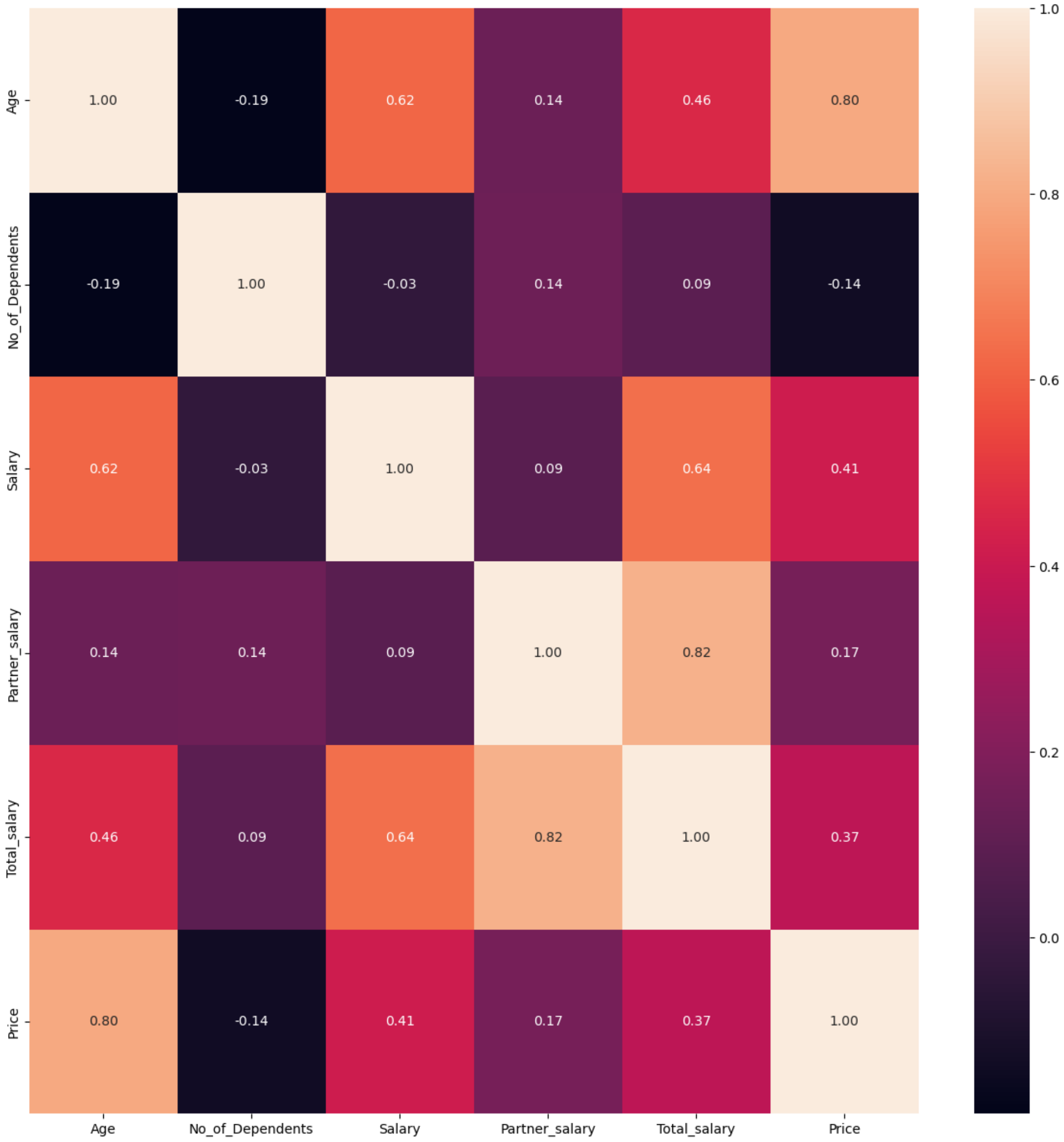


Fig12: correlation between numerical variables without outlier treatment in Total_salary

Post outlier treatment:

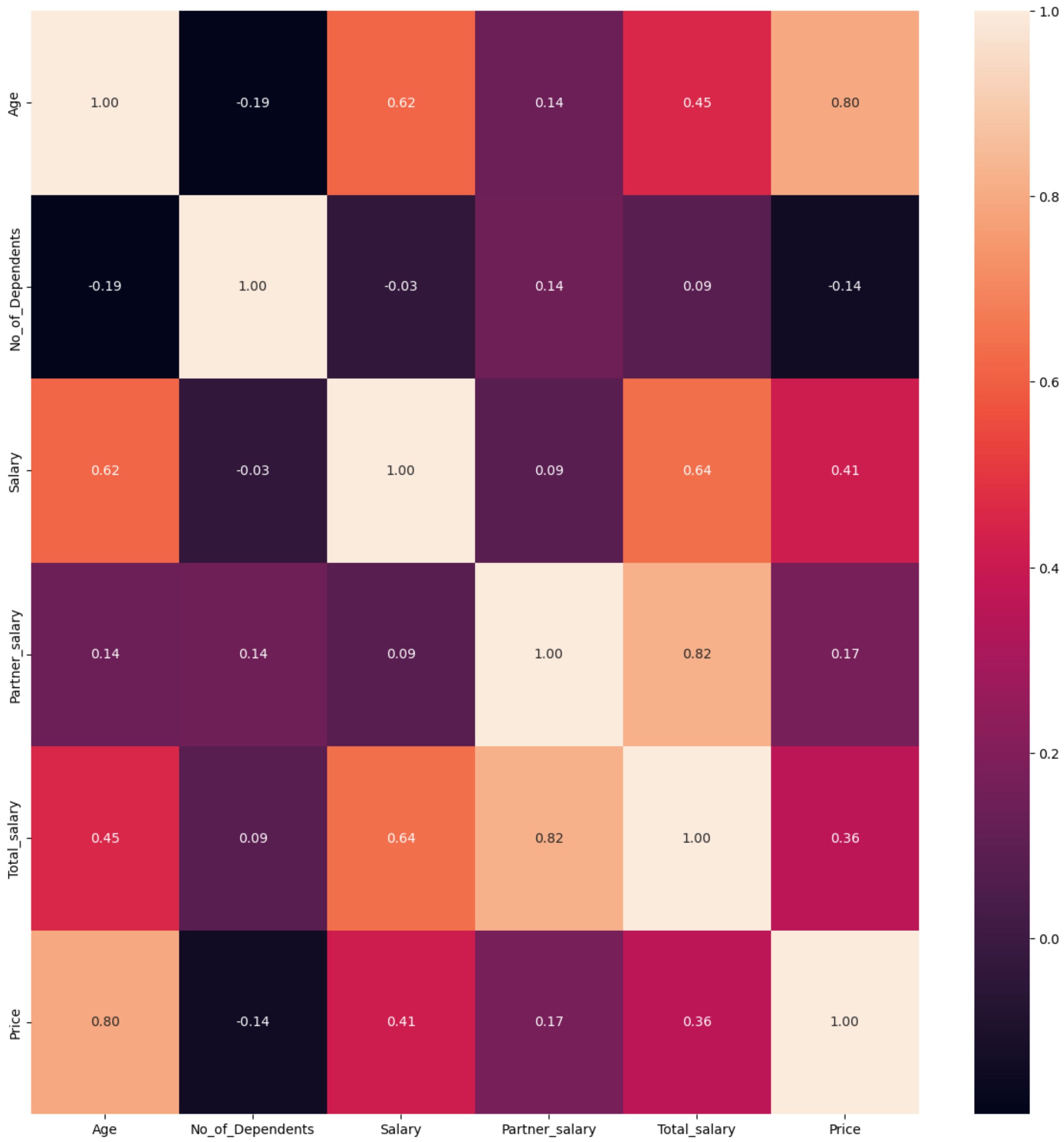
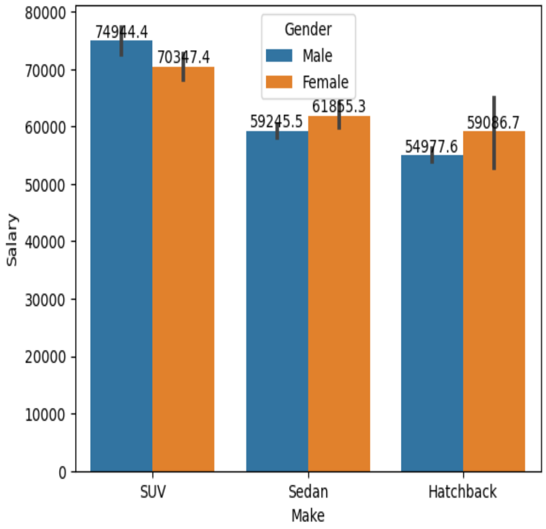


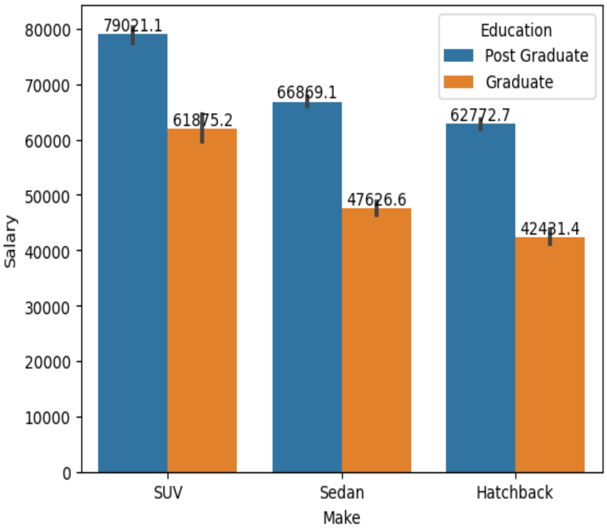
Fig12: correlation between numerical variables with outlier treatment in Total_salary

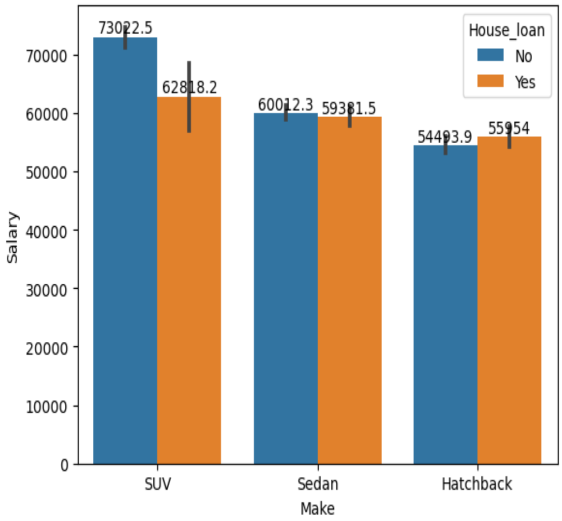
Insights:

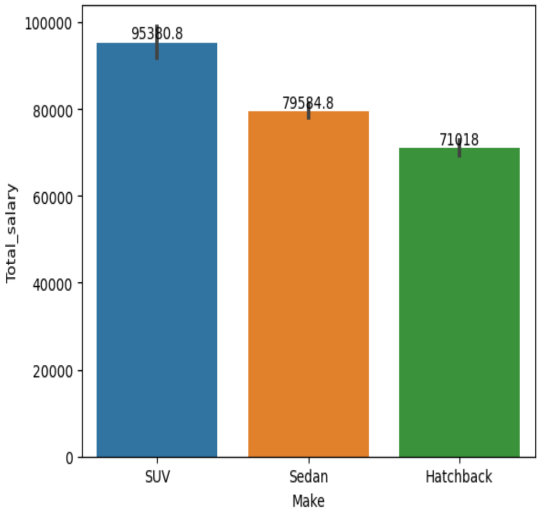
1. Highest correlation is between Total_salary and Partner_salary at 0.82.
2. Second highest correlation is between Price and Age at 0.80.

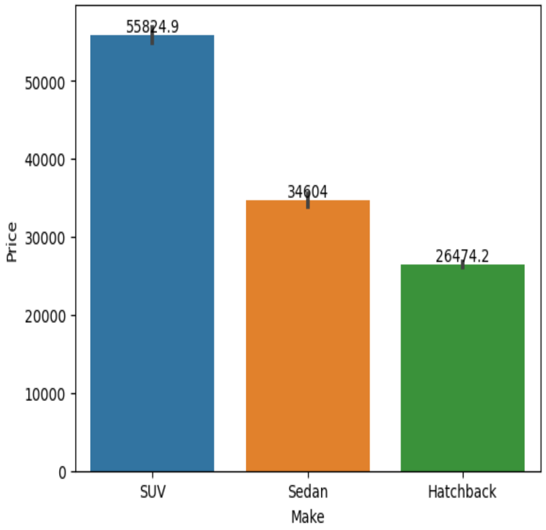
1.3.B Explore the relationship between categorical vs numerical variables.

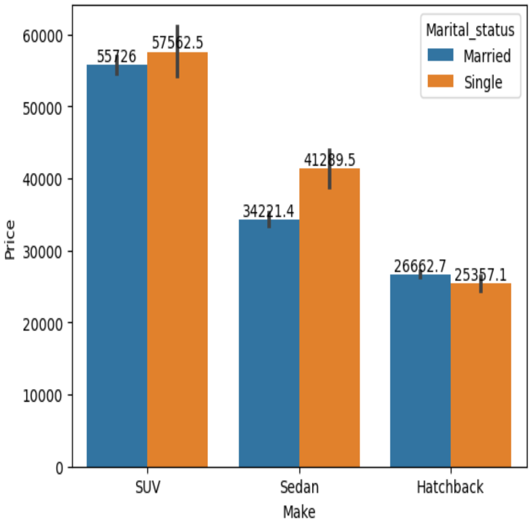












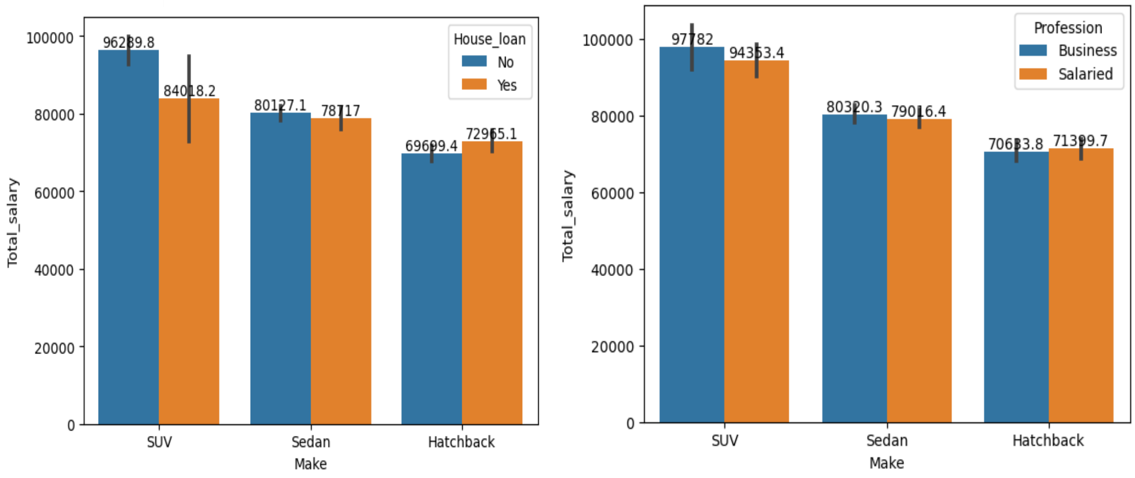


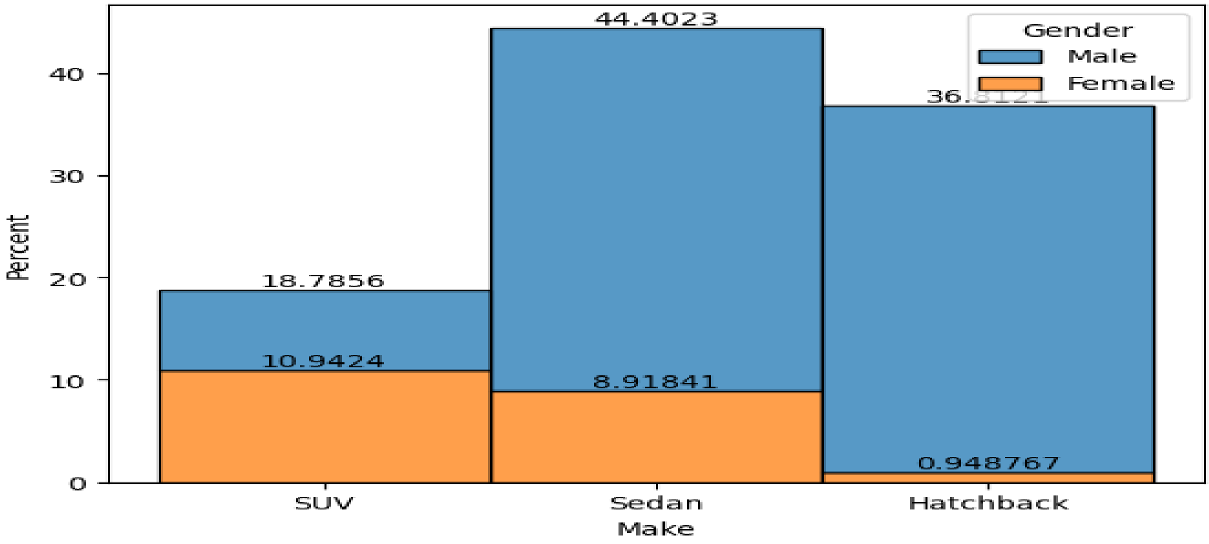
Fig12: Relation between numerical variables and categorical variables

Inferences:

- People with higher Total_salary tend to buy SUV, Sedan, and Hatchbacks in the respective order.
- People with no House loan has a higher purchasing power for SUV.
- SUVs are the highest priced cars comparted to Sedan and Hatchback
- Single people buy SUV and Sedan more than married people whereas married people buy more hatchback compared to single customers.
- Female customers prefer Sedan and hatchback compared to male customers.

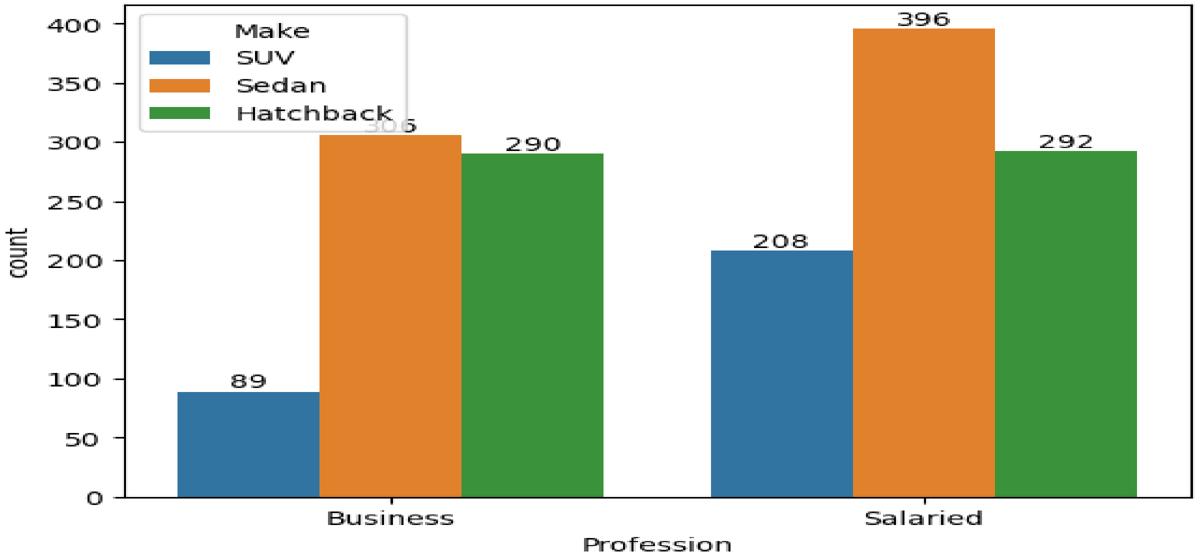
1.4 Key Questions

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



Answer: The percentage of number of SUV's brought by men (18.78%) are higher than women (10.94%)

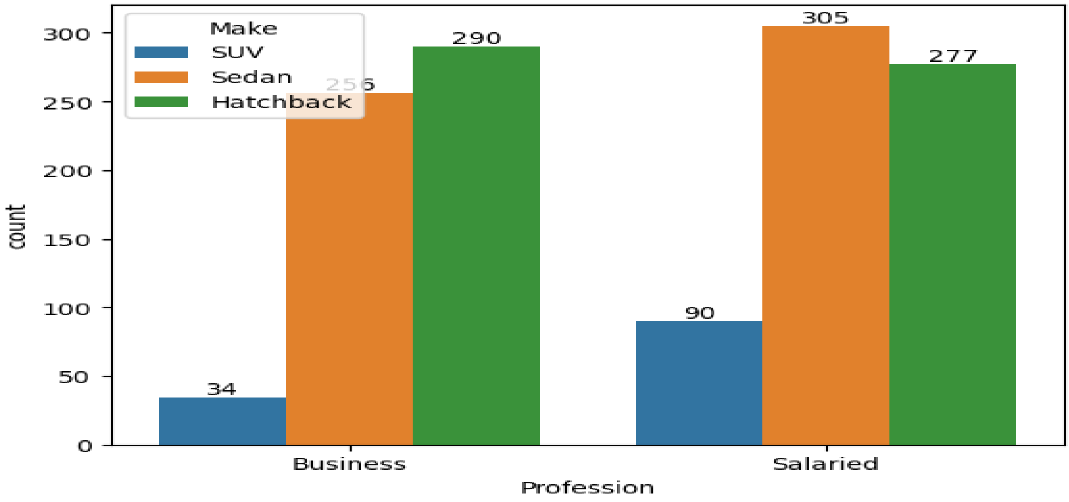
1.4.2 What is the likelihood of a salaried person buying a Sedan?



Likelihood of Salaried person buying SUV is 23.214285714285715
Likelihood of Salaried person buying Sedan is 44.19642857142857
Likelihood of Salaried person buying Hatchback is 32.589285714285715

Answer: From the above chart and analysis, the salaried person is more likely to buy a sedan.

1.4.3 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?

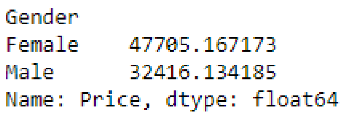


Likelihood of Salaried person buying SUV is 13.392857142857142
Likelihood of Salaried person buying Sedan is 45.38690476190476
Likelihood of Salaried person buying Hatchback is 41.220238095238095

Answer: From the above results Salaried Males prefer Sedan and not SUV which is not as per Sheldon Cooper's claim

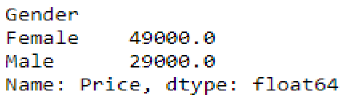
1.4.4 How does the amount spent on purchasing automobiles vary by gender?

Mean:



```
Gender
Female    47705.167173
Male      32416.134185
Name: Price, dtype: float64
```

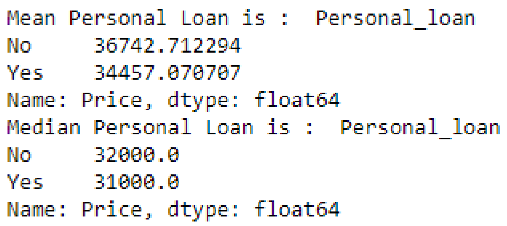
Median:



```
Gender
Female    49000.0
Male      29000.0
Name: Price, dtype: float64
```

Answer: The amount spent by females is more compared to males as the mean and median price is higher.

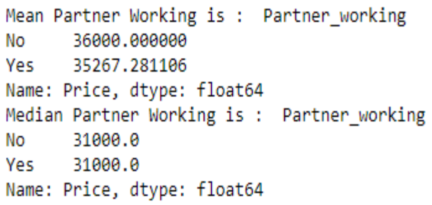
1.4.5 How much money was spent on purchasing automobiles by individuals who took a personal loan?



```
Mean Personal Loan is :  Personal_loan
No     36742.712294
Yes    34457.070707
Name: Price, dtype: float64
Median Personal Loan is :  Personal_loan
No     32000.0
Yes    31000.0
Name: Price, dtype: float64
```

Answer: The mean and median are higher for people who brought a car without a personal loan

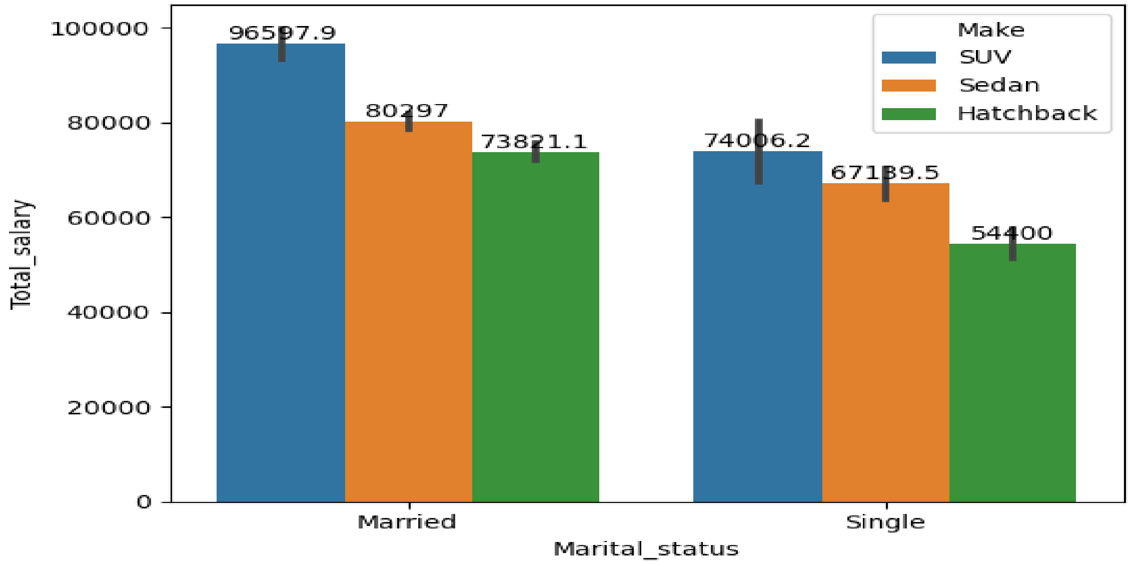
1.4.6 How does having a working partner influence the purchase of higher-priced cars?

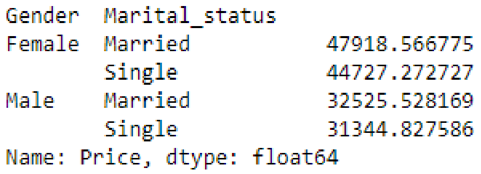


```
Mean Partner Working is :  Partner_working
No     36000.000000
Yes    35267.281106
Name: Price, dtype: float64
Median Partner Working is :  Partner_working
No     31000.0
Yes    31000.0
Name: Price, dtype: float64
```

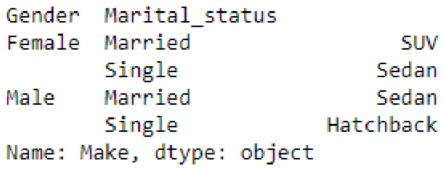
Answer: There is no impact on purchase made by a customer if the partner is working or not as the mean and median values looks almost the same.

1.5 Actionable Insights & Recommendation





```
Gender  Marital_status
Female  Married           47918.566775
        Single            44727.272727
Male    Married           32525.528169
        Single            31344.827586
Name: Price, dtype: float64
```



```
Gender  Marital_status
Female  Married                 SUV
        Single                Sedan
Male    Married               Sedan
        Single            Hatchback
Name: Make, dtype: object
```

From the above diagrams and outputs, the following observations can be made:

a) Married Female prefer SUV.
b) Single Female prefer Sedan.
c) Married Male prefers Sedan.
d) Single males prefer Hatchback.

Problem 2

Context

A bank generates revenue through interest, transaction fees, and financial advice, with interest charged on customer loans being a significant source of profits. GODIGT Bank, a mid-sized private bank, offers various banking products and cross-sells asset products to existing customers through different communication methods. However, the bank is facing high credit card attrition, leading them to reevaluate their credit card policy to ensure customers receive the right card for higher spending and intent, resulting in profitable relationships.

Objective

As a Data Scientist at the company and the Data Science team has shared some data. You are supposed to find the key variables that have a vital impact on the analysis which will help the company to improve the business.

Data Description

userid - Unique bank customer-id
card_no - Masked credit card number
card_bin_no - Credit card IIN number
Issuer - Card network issuer
card_type - Credit card type
card_source_date - Credit card sourcing date
high_networth - Customer category based on their net-worth value (A: High to E: Low)
active_30 - Savings/Current/Salary etc. account activity in last 30 days
active_60 - Savings/Current/Salary etc. account activity in last 60 days
active_90 - Savings/Current/Salary etc. account activity in last 90 days
cc_active30 - Credit Card activity in the last 30 days
cc_active60 - Credit Card activity in the last 60 days
cc_active90 - Credit Card activity in the last 90 days
hotlist_flag - Whether card is hot-listed(Any problem noted on the card)
widget_products - Number of convenience products customer holds (dc, cc, net-banking active, mobile banking active, wallet active, etc.)
engagement_products - Number of investment/loan products the customer holds (FD, RD, Personal loan, auto loan)
annual_income_at_source - Annual income recorded in the credit card application
other_bank_cc_holding - Whether the customer holds another bank credit card
bank_vintage - Vintage with the bank (in months) as on Tthmonth
T+1_month_activity - Whether customer uses credit card in T+1 month (future)
T+2_month_activity - Whether customer uses credit card in T+2 month (future)
T+3_month_activity - Whether customer uses credit card in T+3 month (future)
T+6_month_activity - Whether customer uses credit card in T+6 month (future)
T+12_month_activity - Whether customer uses credit card in T+12 month (future)
Transactor_revolver - Revolver: Customer who carries balances over from one month to the next. Transactor: Customer who pays off their balances in full every month.
avg_spends_l3m - Average credit card spends in last 3 months
Occupation_at_source - Occupation recorded at the time of credit card application
cc_limit - Current credit card limit

2.1 Analyse the dataset and list down the top 5 important variables, along with the business justifications.

Below are the Top 5 important variables:

cc_limit

Credit Card limit for customers depends on salary/income, CIBIL / Credit Score, Other loans, Education etc. and is used to identify risk and advise the banks try to provide only allowable limits of credits to a customer.

avg_spends_l3m

The avg_spends_l3m variable can give information on spending behaviour. Campaigns can be used to target the customer based on their preferences and more customized offers to use the card often.

cc_active30, cc_active60, cc_active90

Flag variables such as cc_active30, cc_active60, cc_active90 can be used understand the frequency of usage of the credit card. This can help identify if there are any issues with the card or why it is not being used.

T+1_month_activity, T+2_month_activity , T+3_month_activity

This can be used by marketing team for promotional offers during a particular period to maximize the spending on the card.

annual_income_at_source

This field shows the ability of the individual to spend and make repayment on time. This can also be used for CIBIL score, Risks involved, ads, campaigns , loans, offers etc.