SMDM Project Report

Contents

List of Figures

List of Tables

SMDM Project

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 is able to grasp the insight.

I. Austo Motor Company is a leading car manufacturer and specializes in SUV, Sedan and Hatchback models. In its recent board meeting, concerns were raised by the members on the efficiency of the marketing campaign currently being used. The board decides to rope in an analytics professional to improve the existing campaign.
You as an analyst have been tasked with performing a thorough analysis of the data and to come up with insights to improve the marketing campaign.

The instructions below are given to help you complete the project –

A) What is the important technical information about the dataset that a database administrator would be interested in? (Hint: Information about the size of the dataset and the nature of the variables) (2 Marks)

Load the required packages, set the working directory and load the data file.

Dataset has 1581 rows and 14 columns

It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function.

Table 1: Top five rows of the dataset

	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

Table 2: Basic Information of the dataset

```
<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: 173.0+ KB
```

A quick look at the dataset information tells us that there are 6 are numerical and 8 categorical variables. There are few Null records present in two variables: Gender and Partner_salary, which will be analyzed in detail in the next section.

There are no duplicate records in the dataset.

B) Take a critical look at the data and do a preliminary analysis of the variables. Do a quality check of the data so that the variables are consistent. Are there any discrepancies present in the data?

Are there any extreme values in the variables – Age, Salary, Partner_salary, Total_salary, Price? Comment whether the extreme values may be considered as anomalies or not from a business perspective? (8 Marks)

(Hints: Quantitative summarization is the first step of understanding data. But is that enough to check for internal data consistency? Extremely low or high values may be considered as anomalies if it is impossible to obtain such a value in normal circumstances)

- Inspecting Null Values -
  There are Nulls in Gender and Partner_salary variables.

  Gender - total 53 Nulls

  Partner_salary - Total 106 Nulls

- Handling Nulls -

Nulls are usually handled by the following techniques –

a) If the proportion of Null values is more than 60 % of the total number of records in a column, then drop the column. Here you assume that the column is uninformative.

b) If any row is missing a large amount of records across columns then that row may also be dropped.

c) Otherwise, the missing values may be imputed.

For the given data, neither (a) nor (b) is applicable since the proportion of null values in any column is small and no row contains a large number of missing observations.

Simple rules for imputation:

a) For categorical variables we can impute the Nulls with the majority class.

   For the current dataset, Null values in 'Gender' field are imputed with 'Male' (Male being the majority class).

b) For continuous variables it is possible to impute the Null values with mean/median of the variable depending upon the nature of the distribution. However, more efficient imputation is possible if variables are internally related.

   The three variables on salary are related to one another as:

Salary + Partner_salary = Total_salary

Also, non-null values in Partner_salary field is possible only if the Binary variable Partner_working is YES. Hence for this data we do a rule based imputation instead of the mean/median imputation –

If Partner_working = 'No' then Partner_salary = 0

If Partner_working = 'Yes' then Partner_salary = Total_salary - Salary

- Inspecting the Summary Statistics of the Dataset (Numerical fields)

Table 3: Numerical summarization of the dataset

	Age	No_of_Dependents	Salary	Partner_salary	Total_salary	Price
count	1581.00	1581.00	1581.00	1475.00	1581.00	1581.00
mean	31.92	2.46	60392.22	20225.56	79626.00	35597.72
std	8.43	0.94	14674.83	19573.15	25545.86	13633.64
min	22.00	0.00	30000.00	0.00	30000.00	18000.00
25%	25.00	2.00	51900.00	0.00	60500.00	25000.00
50%	29.00	2.00	59500.00	25600.00	78000.00	31000.00
75%	38.00	3.00	71800.00	38300.00	95900.00	47000.00
max	54.00	4.00	99300.00	80500.00	171000.00	70000.00

Table 4: Skewness of variables

Variable	Skew
Age	0.89
No_of_Dependents	-0.13
Salary	-0.01
Partner_salary	0.34
Total_salary	0.61
Price	0.74

1) The customers are between 22 and 54 years old. It may be said that they belong to working age group. Mean age is 31.92 while median age is 29 years, indicating age distribution is positively skew. The value of skewness is 0.89.
2) The salary of the customers ranges between 30K and 99.3K and the distribution is symmetric. The mean and the median values are very close and skewness is very close to 0.
3) Total_salary ranges between 30K and 171K and does not show a high degree of skewness.
4) The minimum price of the purchased automobile is 18K, whereas max is 70K. Price has a skewness of + 0.74 indicating moderate skewness. This indicates a small number of high priced purchases were made.

Checking for anomalous values in categorical variables

Determining the unique values for each categorical variable to check if any junk/garbage values are present. This check can also help us to identify if any data entry issues are present.

Table 5: Value Counts of the Categorical variables

```
GENDER :  4
Femle         1
Femal         1
Female      327
Male       1252
Name: Gender, dtype: int64
```

```
EDUCATION :  2
Graduate         596
Post Graduate    985
Name: Education, dtype: int64
```

```
PROFESSION :  2
Business     685
Salaried     896
Name: Profession, dtype: int64
```

```
PERSONAL_LOAN :  2
No      789
Yes     792
Name: Personal_loan, dtype: int64
```

```
PARTNER_WORKING :  2
No      713
Yes     868
Name: Partner_working, dtype: int64
```

```
MARITAL_STATUS :  2
Single       138
Married     1443
Name: Marital_status, dtype: int64
```

```
HOUSE_LOAN :  2
Yes      527
No      1054
Name: House_loan, dtype: int64
```

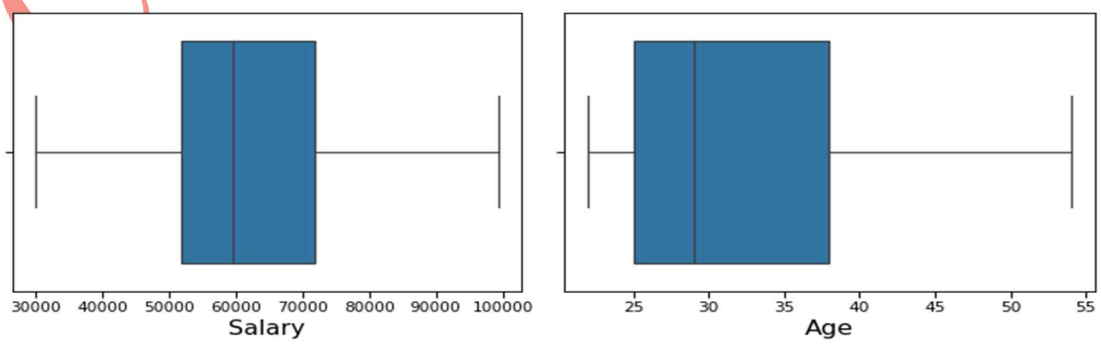
```
MAKE :  3
SUV          297
Hatchback    582
Sedan        702
Name: Make, dtype: int64
```

From the value counts of the Gender variable, we find that there are two instances of possible data entry issue. The word Female has been misspelt as **'Femle'** and **'Femal'**.

For the current dataset we are confident that category Female has been misspelt, so we can go ahead and impute these records with the correct spelling i.e. 'Female'. However, in real time data the issues might not be this straightforward all the time, it might need thorough inspection and domain knowledge to rectify such issues.

Rest of the categorical fields seem to be free from any such issues.

Inspecting continuous fields for anomalies/extreme values –



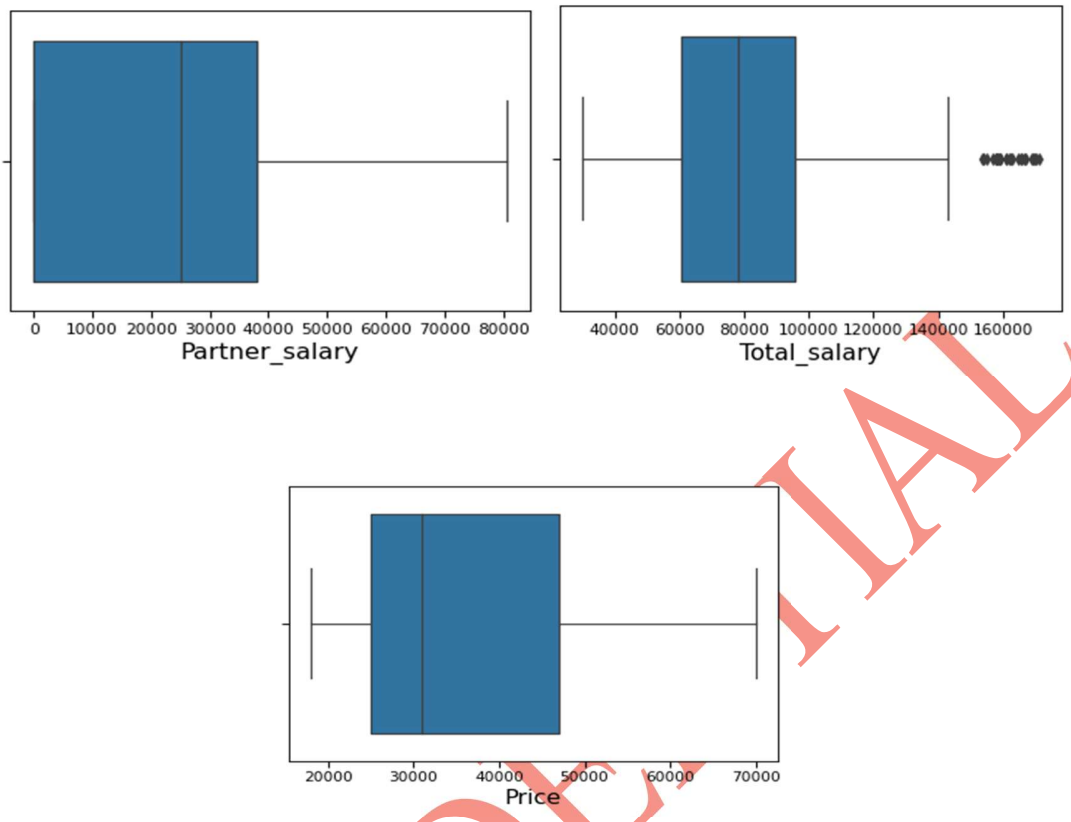


Figure 1: Boxplots of the numerical variables

There are no negative values present in the numerical fields.

From the boxplots we can observe outlier values are present in Total_salary variables. (Using the 1.5 * IQR rule)

Note – If more than 25-30% of the total records lie beyond the range defined by the 1.5 times IQR rule, then to avoid a significant loss of data 3 times IQR rule is considered.

Total_Salary- A total of 27 outlier values are present in the variable.

To handle the outliers in Total_salary, we can choose any of the following two approaches -

1) We can treat the outlier values using Winsorization. However, this may lead to loss of valuable information hence should be used with caution.

2) We do not treat the outlier values, and see if analysing them separately can give us some more insights.

For the current study, we will implement both the approach and will analyse how results/ inferences have changed due to the Outlier Treatment.

Creating two datasets to create solution for the two approach –

df- dataset without outlier treatment

data_out_treat – dataset with outlier treatment

Outlier Treatment in dataset ***data_out_treat*** **-**

Outliers were treated by using Winsorization, i.e. bringing the larger outliers (Data points above the Q3 + 1.5 * IQR value) to the upper whisker value and bringing the smaller outliers (Data points below the Q1 - 1.5 * IQR value) to the lower whisker.

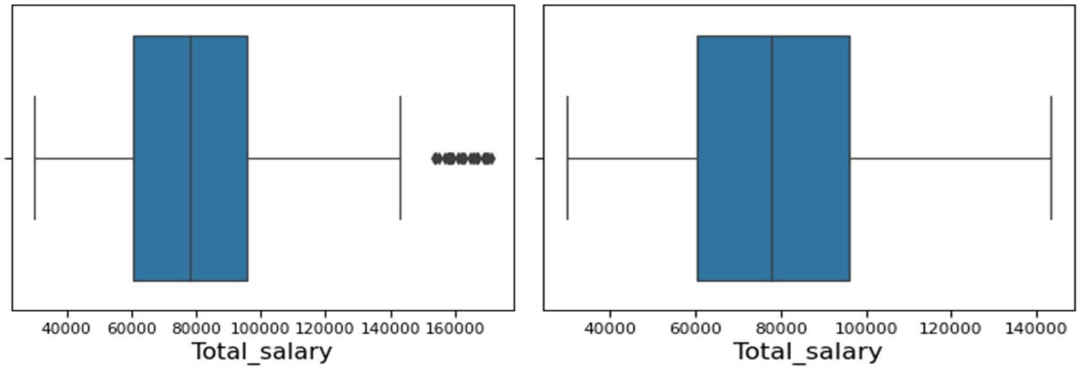


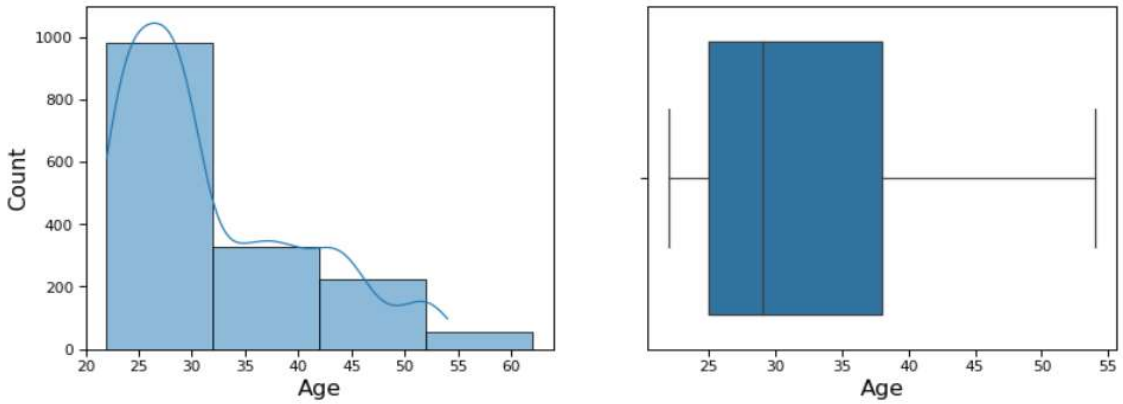
Figure 2: Effect of outlier treatment on Total_salary variable

C) Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business. (8 Marks)

(Hint: Perform Univariate analysis of the variables, choose the visualization as per the data type of the field)

Univariate Analysis of Numerical fields –

For performing Univariate analysis we will take a look at the Boxplots and Histograms to get better understanding of the distributions. Note that these plots have been produced after all data pre-processing (Null value imputations and Winsorization) have been done



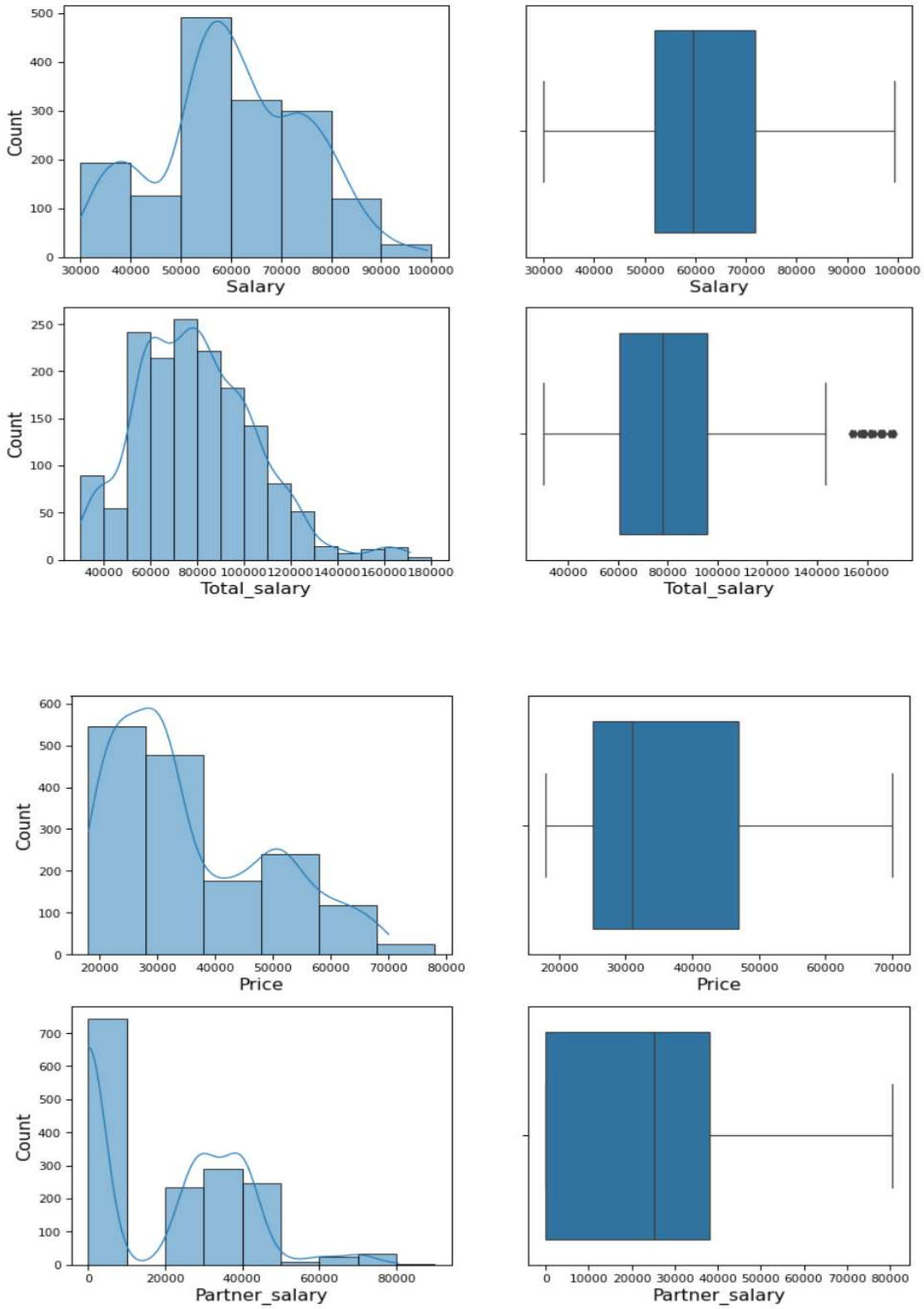


Figure 3: Univariate analysis of numerical variables

Total_salary after outlier treatment –

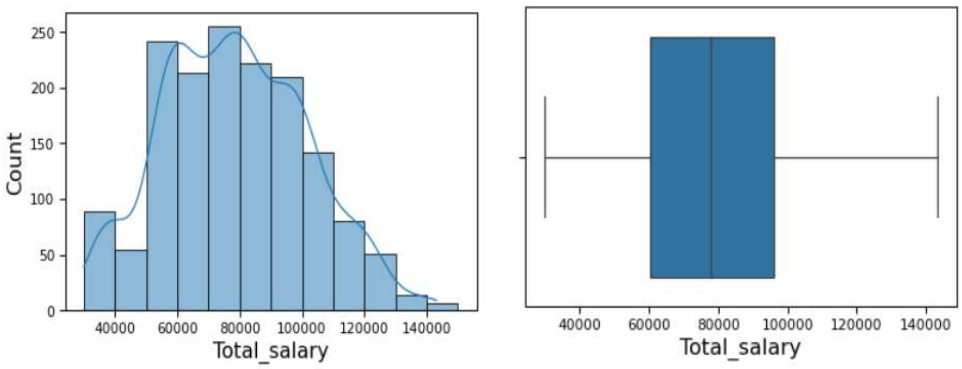
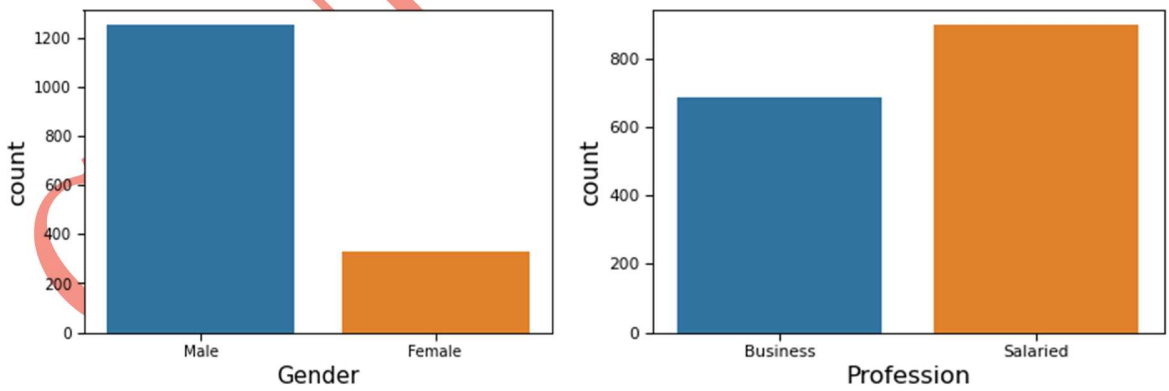


Figure 4: Univariate analysis of Total_salary post outlier treatment

Inferences –

1) Salary has a multimodal distribution, with bulk of data points in the range 50K to 70K.
2) Price seems to have a Bi-modal distribution, and a positive skew of 0.74.
3) Age seems to have a multimodal distribution, and has the highest positive skew of 1.14 among all the fields.
4) Skewness of Total_salary has reduced significant post outlier treatment. The distribution seems to be multimodal, with bulk of data points in the range of 60K to 100K.
5) Almost all the variables have some skewness present, thus none of them follow a Normal distribution. Total_salary can be considered Near-Normal distribution with fair bit of approximation.

Univariate analysis of Categorical variables –



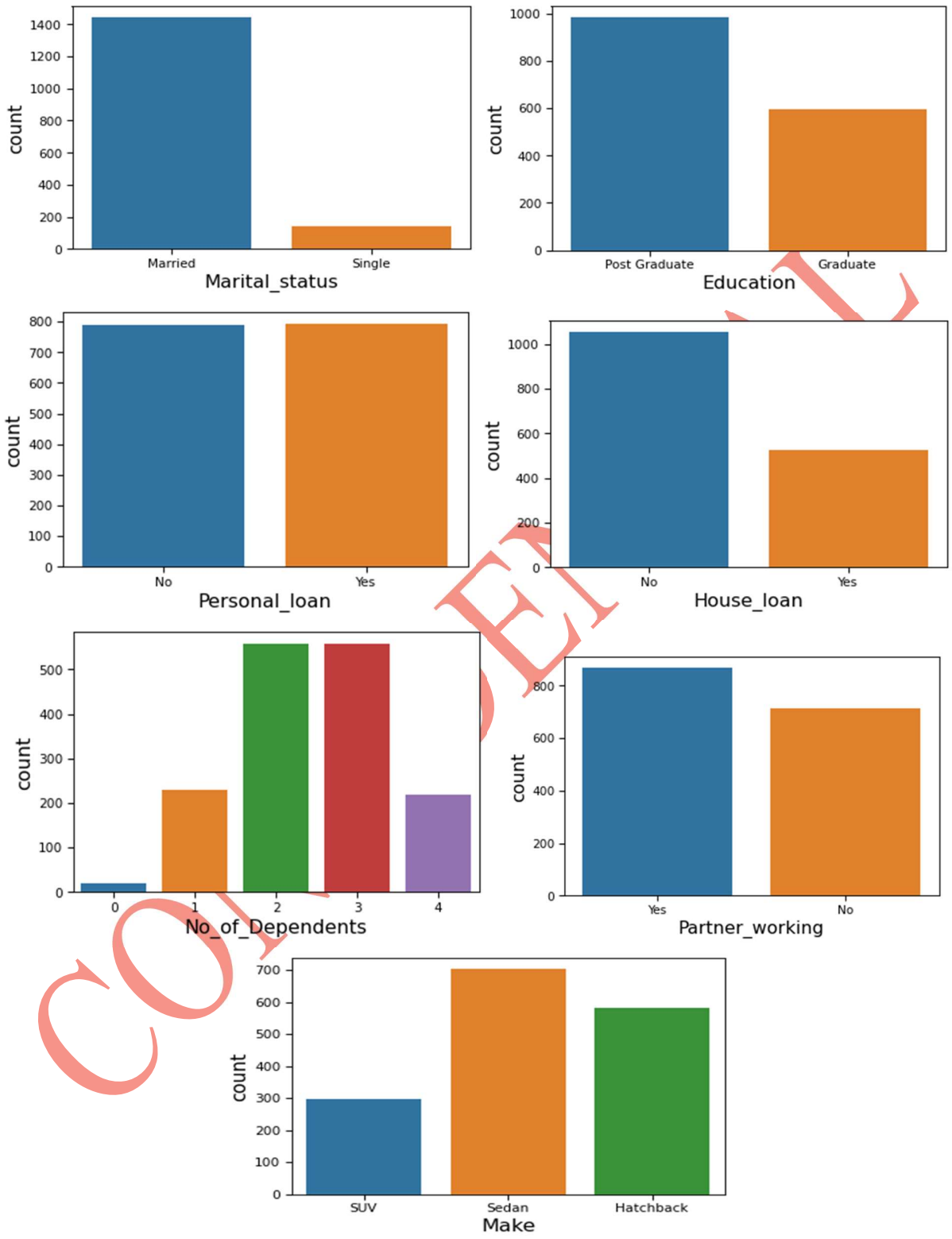


Figure 5: Univariate analysis of categorical fields

Inferences –

1) Sedan is the most preferred purchase, followed by Hatchback and SUV.
2) The number of customers having a working partner are slightly higher than customers with non-working partner or singles. There are a total of 713 customers with Partner_working variable as 'No', out of which 138 customers are 'Single'.
3) Number of Customers who did not take a House Loan is almost double the customers who took a House Loan.
4) The data consists of very small proportion of Single customers when compared to married customers.
5) Count of Salaried customers is slightly higher than that of Business customers.
6) Majority of the customers in the dataset are Post Graduate.
7) From the Barplot of No_of_dependnts variable we can infer that majority of the customers have either 2 or 3 dependents, followed by 1 or 4 dependents. Very few customers have zero no of dependents.

D) Understanding the relationships among the variables in the dataset is crucial for every analytical project. Perform analysis on the data fields to gain deeper insights. Comment on your understanding of the data. (8 Marks) (Hint: How can you explore joint relationship among two or more variables using appropriate visualization?)

Bivariate analysis of Numerical variables:

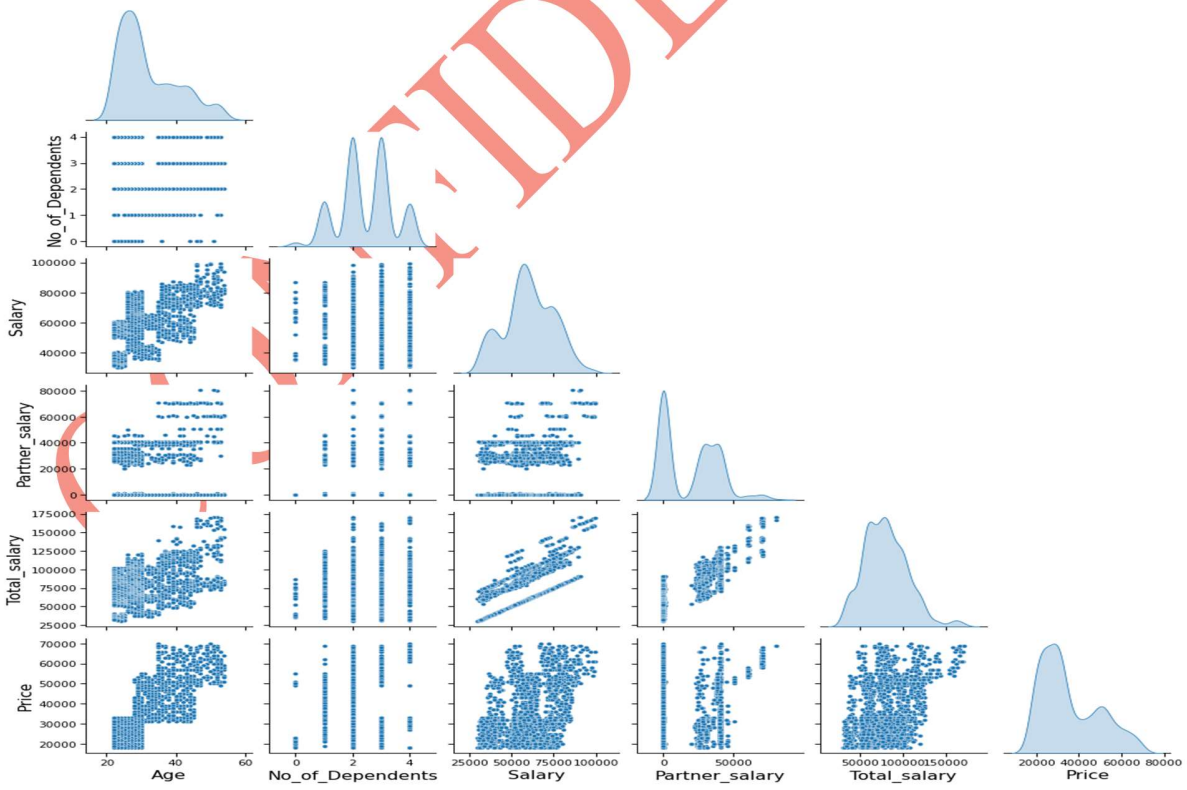


Figure 6: Pair plot on the Data set without treating outliers in Total_salary

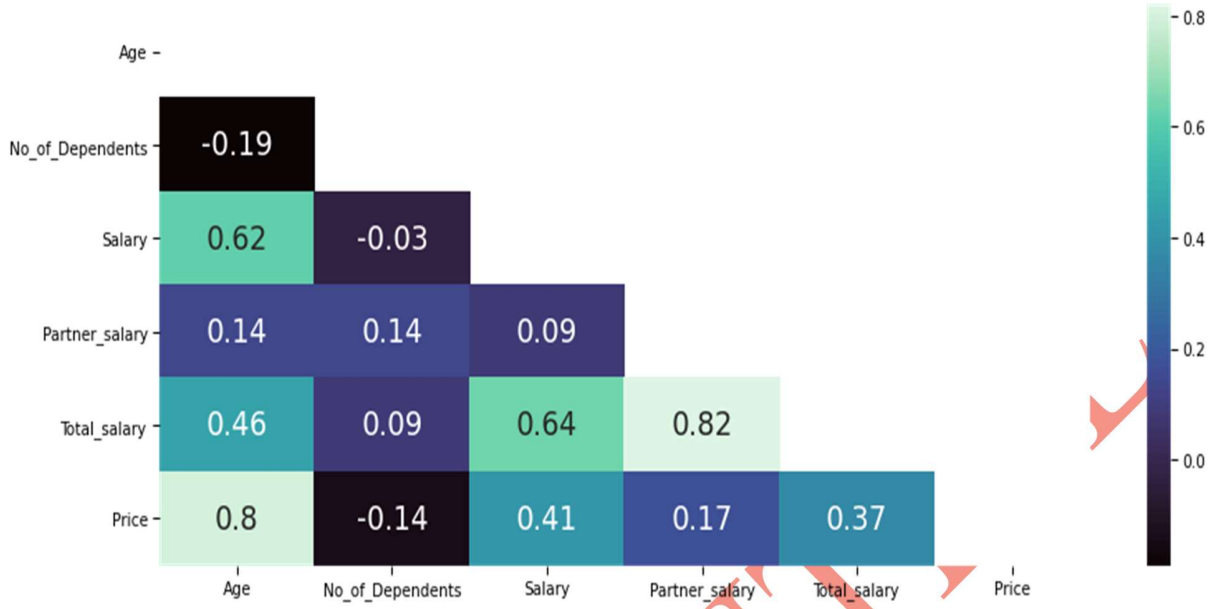


Figure 7: Correlation Heatmap without treating outliers in Total_salary

Post outlier treatment –

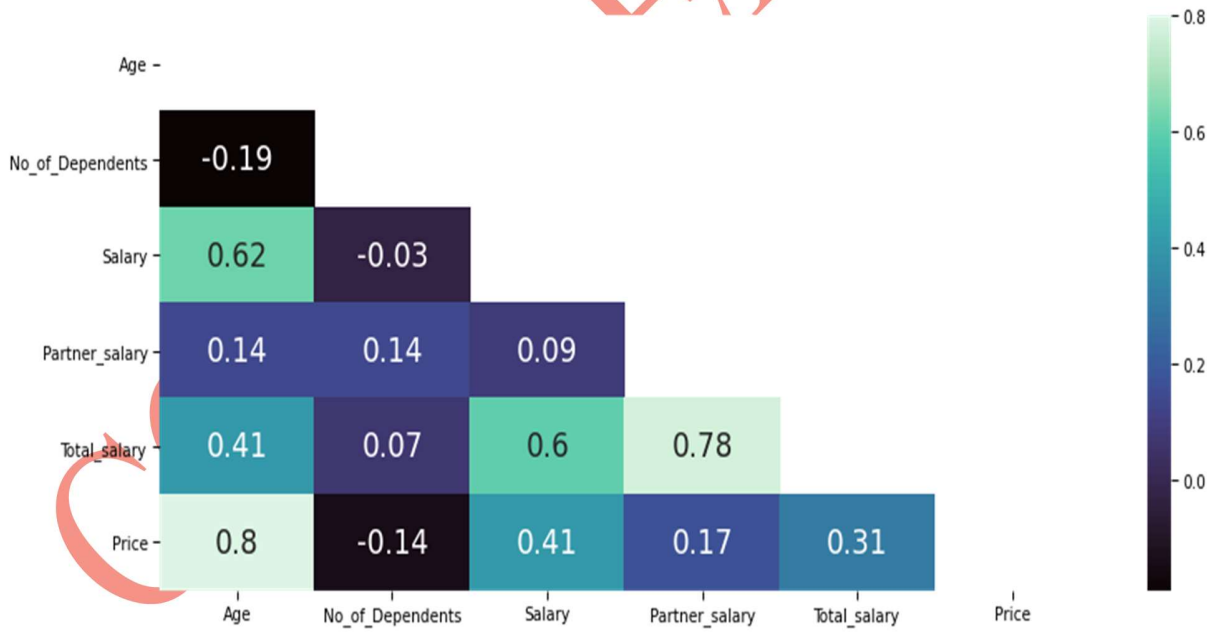


Figure 8: Correlation Heatmap post outlier treatment in Total_salary

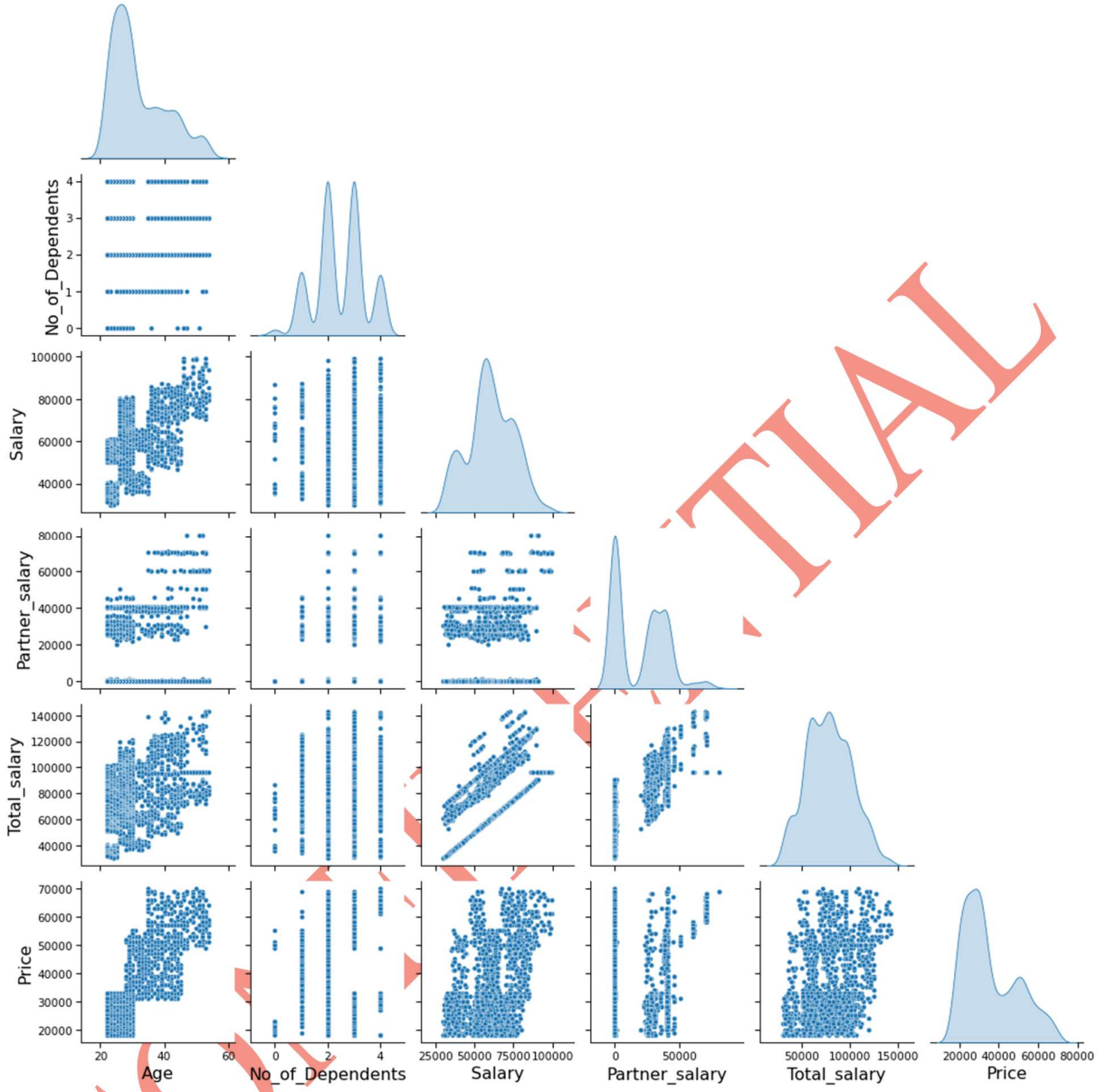


Figure 9: Pair plot on the dataset post outlier treatment in Total_salary

Inferences –

1) Hardly any linear relationships present among the fields.
2) Highest positive correlation exists between Price and Age (Post Outlier Treatment), and Total_salary and Partner_salary (without outlier treatment)

Bi- Variate analysis of Categorical vs Categorical variables –

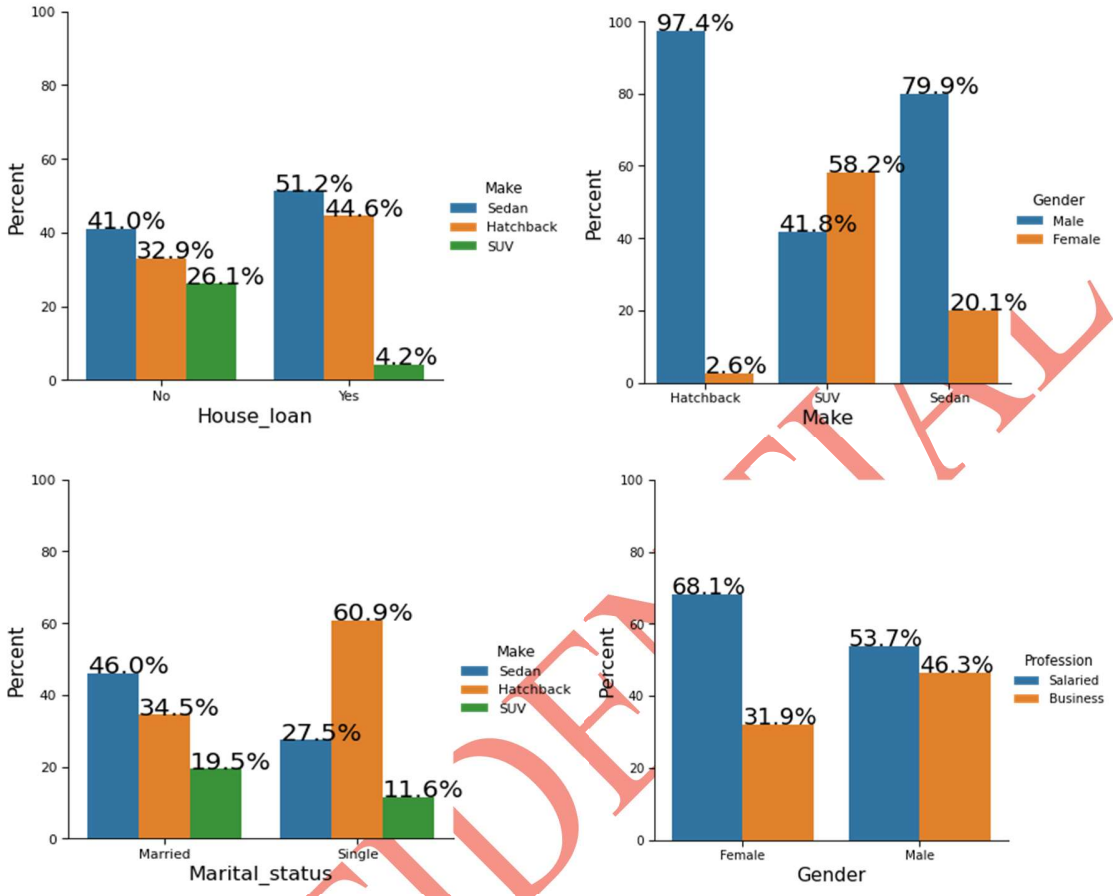


Figure 10: Proportion plots for Categorical vs Categorical fields

Inferences –

1) Customers who have a house loan are not likely to buy an SUV (which is the most costly make among the three). Sedan is most preferred across both the categories.
2) Females prefer SUV and are least likely to buy a Hatchback, whereas Male prefer Sedan or hatchback. SUV is least preferable among males
3) Married customers prefer Sedan whereas single customers prefer Hatchbacks.

E) Employees working on the existing marketing campaign have made the following remarks.

Based on the data and your analysis state whether you agree or disagree with their observations. Justify your answer Based on the data available. (6 Marks)

E1) Steve Roger says "Men prefer SUV by a large margin, compared to the women"

Analysing the ratio of SUV purchases for both the Genders, we get:

Proportion of females buying SUV = 0.52 (Number of females who bought SUV / Total number of females)

Proportion of Males buying SUV = 0.09 (Number of males who bought SUV / Total number of males)

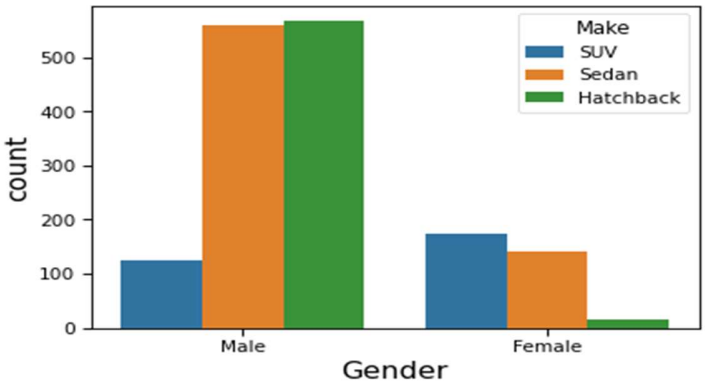


Figure 11: Count Plot of Gender vs Make

Hence the statement made by Steve Rogers is incorrect.

E2) Ned Stark believes that a salaried person is more likely to buy a Sedan.

Analysing the Proportion of Car Make purchases for salaried customers, we get:

Proportion of Hatchbacks purchased = 0.32 (Total Hatchbacks bought by salaried / Total Cars purchased by salaried)

Proportion of SUV purchased = 0.23 (Total SUVs bought by salaried / Total Cars purchased by salaried)

Proportion of Sedan purchased = 0.44 (Total Sedans bought by salaried / Total Cars purchased by salaried)

Using Visualization to arrive at the conclusion, we plot a count plot of Profession as x , while Make as Hue parameter.

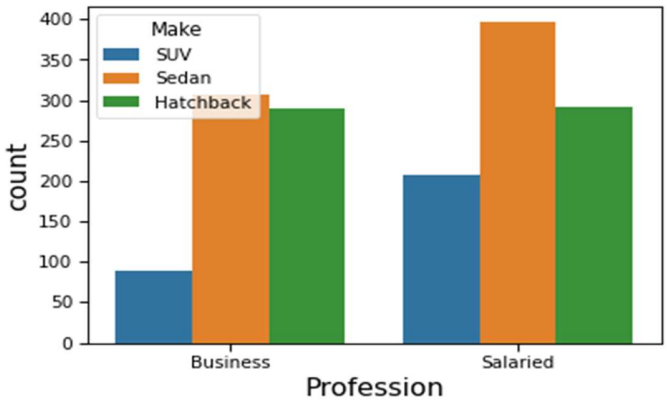


Figure 12: Count Plot of Profession vs Make

From the above results and chart, it is evident that salaried person is more likely to buy a Sedan.

Hence the statement made by Ned Stark is correct.

E3) Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for a SUV sale over a Sedan Sale.

Calculating Total number of Cars purchased by Salaried Male Customers for each Make, we get -

Proportion of Hatchback = 277/ 672 = 0.41 (Total Hatchbacks purchased / Total Cars purchased)

Proportion of SUV = 90/672 = 0.13 (Total SUV purchased / Total Cars purchased)

Proportion of Sedan = 305/672 = 0.45 (Total Sedans purchased / Total Cars purchased)

Using Visualization to arrive at the conclusion, we plot a count plot of Profession as x , while Make as Hue parameter for the Male customers.

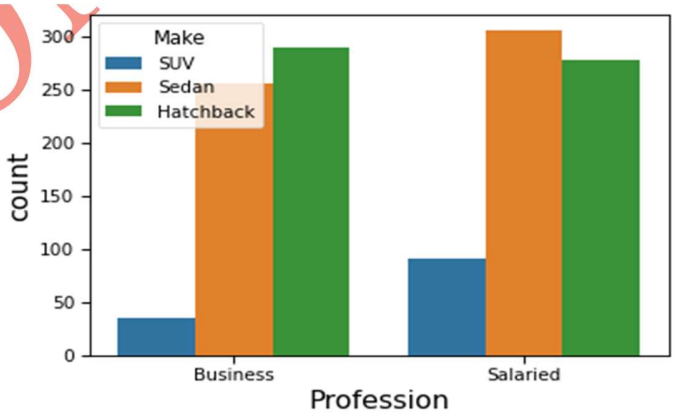


Figure 13: Count Plot of Profession vs Make (For Male customers)

From the above results and chart, it is evident that Salaried male prefers Sedan over SUV.

Hence the statement made by Sheldon Cooper is incorrect.

F) From the given data, comment on the amount spent on purchasing automobile across the following categories –

Comment on how a Business can utilize the results from this exercise.

Give justification along with presenting metrics/charts used for arriving at the conclusions. (4 Marks)

F1) Gender,

Females are more likely to buy SUVs and on an average spend more on cars than males 47705 Units against 32416 Units.

Mean of Price across Gender:

Female = 47705

Male = 32416

Median of Price across Gender:

Female = 49000

Male = 29000

Mean and Median Price for Female customers is higher than Male customers.

F2) Personal loan

Mean of Price across Personal Loan:

Personal Loan: No= 36742

Personal Loan: Yes= 34457

Median of Price across Personal Loan:

Personal Loan: No= 32000

Personal Loan: Yes= 31000

Mean and Median of Price for purchase made by customers without a Personal loan is slightly higher than customers who have a Personal Loan.

To ensure increased spend of customers with Personal loans, the business can look to make the interest rate cheaper (for Automobile purchase) or ease down the repayment terms.

G) From the current data set comment if having a working partner leads to purchase of a higher priced car. (2 Marks)

(Hint: Analyse variables Partner_working and Price)

Mean of Price across Partner_working:

Partner_working: No = 36000

Partner_working: Yes = 35267

Median of Price across Partner_working:

Partner_working: No = 31000

Partner_working: Yes = 31000

The Mean and Median price of the purchased automobile is almost similar across the Partner_working category, thus indicating that partner working or not has no effect on the Purchase made by the customer.

H) The main objective of this analysis is to devise an improved marketing strategy to send targeted information to different groups of potential buyers present in the data. For the current analysis use Gender and Marital_status - fields to arrive at groups with similar purchase history. (6Marks)

To arrive at an efficient marketing strategy, we can use the historic data and determine the type of Car that was most frequently purchased by the group of buyers based on the two variables Gender & Marital_status.

Computing the mode (most frequently occurring value) of the Make variable, grouped on Gender and Marital_status we get the following results –

Table 6: Customer Groups and Target Car Make

Group	Gender	Marital_status	Mode of Make
1	Female	Married	SUV
2	Female	Single	Sedan
3	Male	Married	Sedan
4	Male	Single	Hatchback

Also, we can plot two visualizations to get info on the purchase history of these groups -

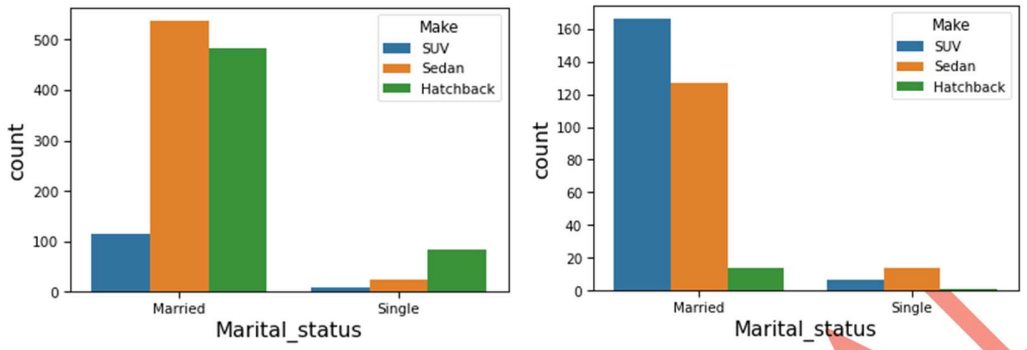


Figure 14: Marital Status vs Make for Male & Marital Status vs Make for Female

From the above diagrams and tables , we can make four groups and assign a Car make to each group basis the most frequently occurring value in the past –

A. Married Female – SUV
B. Married Male - Sedan
C. Single Female – Sedan
D. Single Male – Hatchback

Extra: Analyzing Records with Outlier Values in Total_salary Variable

When we detected outliers in the Total_salary, we decided to keep the outliers in one of the dataset. Let us analyze those records separately to see if we can unearth additional info, and further improve the marketing campaign -

Creating a subset of data for the outlier values (1.5 IQR). We get a total of 27 records in the dataset.

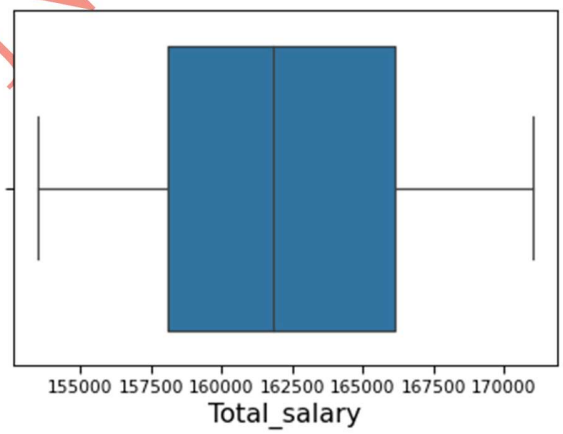


Figure 15: Boxplot of Total_salary for the Extreme Values subset

Finding Most frequently purchased Car make grouped on Marital_Status and Gender, we find :

Female – Married: SUV

Male – Married: SUV

(Note: In the Outlier subset data there were only Married customers)

Analyzing the mean Price of purchased car across the Marital_status and Gender, we find:

Mean Price for purchases made by Married Females = 62857

Mean Price for purchases made by Married Males = 60692

Having a quick look at the Mode of the Car make for Gender and Marital_status fields, we observe that we get only two groups Male Married & Female Married with both the groups preferring SUV.

Similarly analyzing the Mean of Price, we can see that Mean for Male Married is approx. 60K while it is 62K for Female Married.

Additional insight from extreme value dataset-

1. All the Male Married Customers with Total Salary greater than 149 K purchased SUV. Whereas Married male with lower Total_salary preferred Sedan.

FRAMING AN ANALYTICS PROBLEM

- QUESTION - **Analyse the dataset and list down the top 5 important variables, along with the business justifications. (10 Points) (Context and the Data Dictionary is as below)**

CONTEXT:

A bank can generate revenue in a variety of ways, such as charging interest, transaction fees and financial advice. Interest charged on the capital that the bank lends out to customers has historically been the most significant method of revenue generation. The bank earns profits from the difference between the interest rates it pays on deposits and other sources of funds, and the interest rates it charges on the loans it gives out.

GODIGT Bank is a mid-sized private bank that deals in all kinds of banking products, such as savings accounts, current accounts, investment products, etc. among other offerings. The bank also cross-sells asset products to its existing customers through personal loans, auto loans, business loans, etc., and to do so they use various communication methods including cold calling, e-mails, recommendations on net banking, mobile banking, etc.

GODIGT Bank also has a set of customers who were given credit cards based on risk policy and customer category class but due to huge competition in the credit card market, the bank is observing high attrition in credit card spending. The bank makes money only if customers spend more on the credit cards. Given the attrition, the Bank wants to revisit its credit card policy and make sure that the card given to the customer is the right credit card. The bank will make profit only through the customers that show higher intent towards a recommended credit card. (Higher intent means consumers would want to use the card and hence not be attrite.)

DATA SET DESCRITPION

Credit Card Data for GODIGIT Bank

DATA DICTIONARY:

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 last 30 days
cc_active60	Credit Card activity in last 60 days
cc_active90	Credit Card activity in 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 customer holds (FD, RD, Personal loan, auto loan)
annual_income_at_source	Annual income recorded in credit card application
other_bank_cc_holding	Whether customer holds another bank credit card
bank_vintage	Vintage with the bank (in months) as on Tth month
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

**All above data has been recorded as on T^{th} month excluding T+1_month_activity, T+2_month_activity, T+3_month_activity, T+6_month_activity, T+12_month_activity*

Solution –

Below are the Top 5 important variables from the given dataset with justification.

1) **annual_income_at_source -**

   Annual income plays a big role in the purchasing power of an individual hence is a vital piece of info. Income can be used by the banks to make better decisions in areas such as risk profiling, targeted ads, campaigns, offers, loan limits etc.

2) **cc_limit –**

   Defining Credit Card limit for customers basis their attributes (such as income, CIBIL Score, etc.) is part of the Risk Management practice wherein the banks try to minimize the number of defaulters. The banks seek a quantifiable answer to the query "How much is too much?"

3) **cc_active30 –**

   Flag variables such as cc_active30, cc_active60 can be used to get an understanding over how frequently does the customer use the credit card, if the account is dormant or if the customer is experiencing any issues leading to reduced usage of the card etc.

4) **T+1_month_activity –**

   Flag variables such as T+1_month_activity can be used to plan out campaigns and promotional offers so as to increase activity in the credit card.

5) **avg_spends_l3m –**

   The avg_spends_l3m variable can give important insights on the customer spending behavior. It can be used to identify whether the credit card is primary or secondary card of customer, i.e. high spend indicates primary account whereas lower spend would mean secondary account. Campaigns can be rolled out on the basis of the customer preference, customized offers can be given to lure customers into using the credit account more frequently.

Few variables which are unimportant from an analysis point of view, and are merely customer/account identifiers–

1) userid
2) card_no
3) card_bin_no