BUSINESS REPORT

G S Jaigururam
Data Scientist

CONTENTS

Problem I

Data Overview

We have Data file that need to be analyzed had been load and we have data description as follow: -

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

Imported necessary Python libraries for data analysis, including pandas, numpy, seaborn and matplotlib. Successfully loaded the dataset into a pandas DataFrame for further analysis.

Structure of the Data:

Data had 15 columns and 1581 rows.
First 5 rows are:

Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary	Pr
53	Male	Business	Married	Post Graduate	4	No	No	Yes	99300	70700.0	170000	61
53	Femal	Salaried	Married	Post Graduate	4	Yes	No	Yes	95500	70300.0	165800	61
53	Female	Salaried	Married	Post Graduate	3	No	No	Yes	97300	60700.0	158000	57
53	Female	Salaried	Married	Graduate	2	Yes	No	Yes	72500	70300.0	142800	61
53	Male	Salaried	Married	Post Graduate	3	No	No	Yes	79700	60200.0	139900	57

Column Details:

Name of columns and their data types are given below:

```
Column            Non-Null Count  Dtype
------            --------------  -----
Age               1581 non-null   int64
Gender            1581 non-null   object
Profession        1581 non-null   object
Marital_status    1581 non-null   object
Education         1581 non-null   object
No_of_Dependents  1581 non-null   int64
Personal_loan     1581 non-null   object
House_loan        1581 non-null   object
Partner_working   1581 non-null   object
Salary            1581 non-null   int64
Partner_salary    1581 non-null   float64
Total_salary      1581 non-null   int64
Price             1581 non-null   int64
Make              1581 non-null   object
```

4. Types of Data:

Reviewed the data types of each column to ensure consistency and accuracy.

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

5. Missing Values Handling:

Detected few missing values as shown below:

```
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
```

Addressed missing values for numerical columns (Partner salary) we are going to Impute missing values with the mean value and for categorical columns (gender) we are going to Impute missing values with the Mode value.

```
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
```

Ensured data completeness and integrity by handling missing values effectively.

6. Statistical Summary:

Generated descriptive statistics to summarize the central tendency, dispersion, and shape of the dataset. Calculated key statistical metrics such as mean, median, standard deviation, and quartiles.

	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	20225.559322	18905.183912	0.0	0.0	24900.0	38000.0	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

7. Data Irregularities:

There was no Duplicate data in the file.

8. Observations and Insights:

1.The dataset comprises a total of 1581 rows, indicating a substantial amount of data available for analysis.

2.The dataset consists of 6 numerical and 8 categorical variables, providing a diverse range of information for analysis.

3.Some missing values were identified within the dataset and appropriately treated using suitable techniques to ensure data completeness and accuracy.

4.No duplicate rows were found within the dataset.

Univariate Analysis

Numerical data analysis

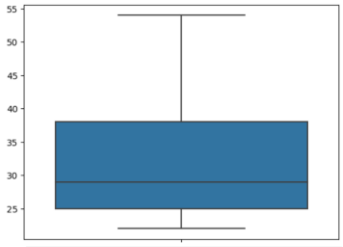
Data Name

Visualization

Age

Observation:

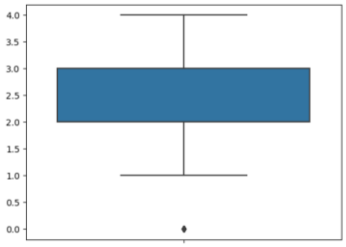
1. The median age is around 29 years.
2.The interquartile range (IQR) is between 25 and 38 years, 3. indicating that 50% of the data lies within this range.
4. There are a No outliers.
5. The distribution of ages is slightly skewed to the right.



No_of_Dependents

Observations:

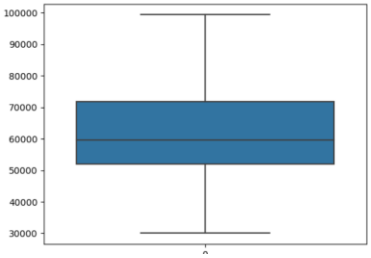
1.The median number of dependents is 0.
2.The interquartile range (IQR) is between 2 and 3 dependents.
3. There are a few outliers with less than 1 dependents.



Salary

Observations:

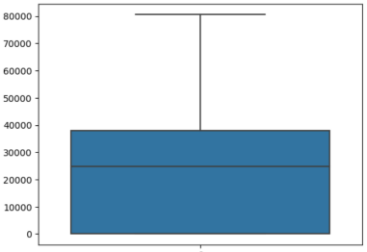
1.The median salary is around 59500.
2.The interquartile range (IQR) is between 51900 and 71800, indicating that 50% of the data lies within this range.
3.There are a No outliers.
4.The distribution of salary is slightly skewed to the right.



Partner_salary

Observations:

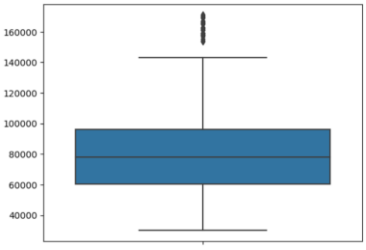
1.The median partner salary is around 24900.
2.The interquartile range (IQR) is between 0 and 38000, indicating that 50% of the data lies within this range.
3.There are a No outliers.
4.The distribution of partner salary is slightly skewed to the left.



Total_salary

Observations:

1.The median Total salary is around 78000.
2.The interquartile range (IQR) is between 60500 and 95900, indicating that 50% of the data lies within this range.
3.There are outliers.
4.The distribution of Total salary is very slightly skewed to the right.



Price

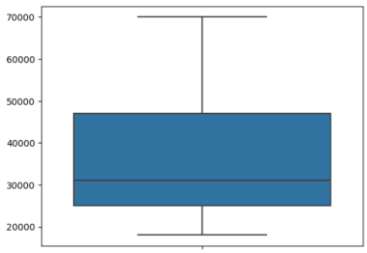
Observations:

1.The median Price is around 31000.

2.The interquartile range (IQR) is between 25000 and 47000, indicating that 50% of the data lies within this range.

#There are a No outliers.

#The distribution of Price is slightly skewed to the right.



Categorical data analysis

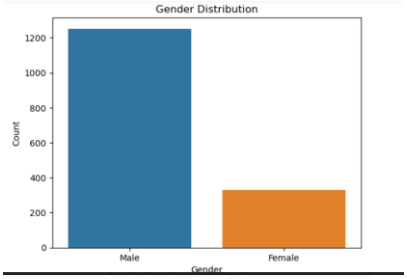
Data Name

Gender

Observations:

1. We saw that Gender column had few data with spelling mistake so we corrected it.
2. Data has more male then female

Visualization

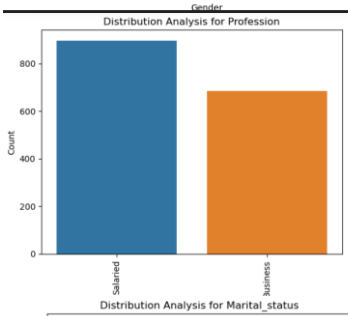


Profession

Observations:

1.No Change or data manipulation required.

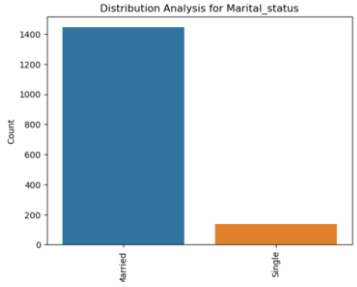
2**.** Data seem to have more salaried personnel then business personnel



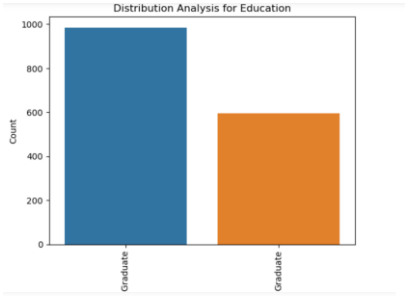
Marital_status

Observations:

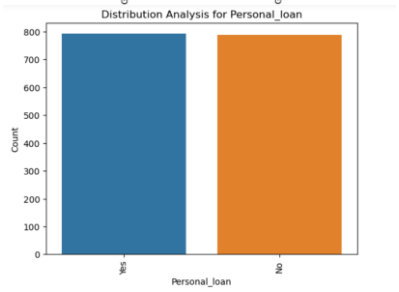
1. No Change or data manipulation required.
2. Data seem to have more married personnel then single personnel



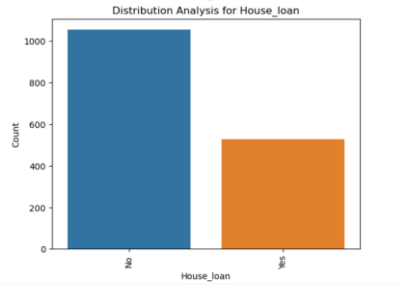
Education
Observations:
1. No Change or data manipulation required.
2. Data seems to have more post graduate personnel then graduate personnel



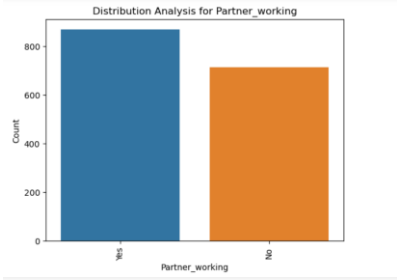
Personal_loan
Observations:
1. No Change or data manipulation required.
2. 50% of data has Personal and other 50% do not have personal loan.



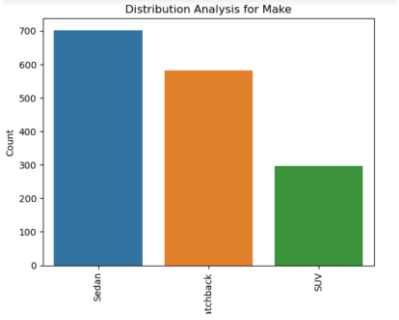
House_loan
Observations:
1. No Change or data manipulation required.
2. Data has more personnel without house loan



Partner_working
Observations:
1. No Change or data manipulation required.
2. Data has more personnel with their partner working



Make
Observations:
1. No Change or data manipulation required.
2. Data has more personnel with sedan type automobile followed by hatchbacks and then SUVs.



Overall observations

1. Age:
The median age is approximately 29 years.
The interquartile range (IQR) spans from 25 to 38 years, indicating that 50% of the data falls within this range.
There are no outliers.
The age distribution is slightly skewed to the right.

2. Number of Dependents:
The median number of dependents is 0.
The interquartile range (IQR) ranges from 2 to 3 dependents.
There are a few outliers with less than 1 dependent. But no need to treat as they are valid values.

3. Salary:
The median salary is around $59,500.
The interquartile range (IQR) is from $51,900 to $71,800, encompassing 50% of the data.
There are no outliers.
The salary distribution is slightly skewed to the right.

4. Partner Salary:
The median partner salary is approximately $24,900.
The interquartile range (IQR) spans from $0 to $38,000.
There are no outliers.
The partner salary distribution is slightly skewed to the left.

5. Total Salary:
The median total salary is around $78,000.
The interquartile range (IQR) ranges from $60,500 to $95,900.
There are outliers present. But no need to treat as they are valid values.
The distribution of total salary is very slightly skewed to the right.

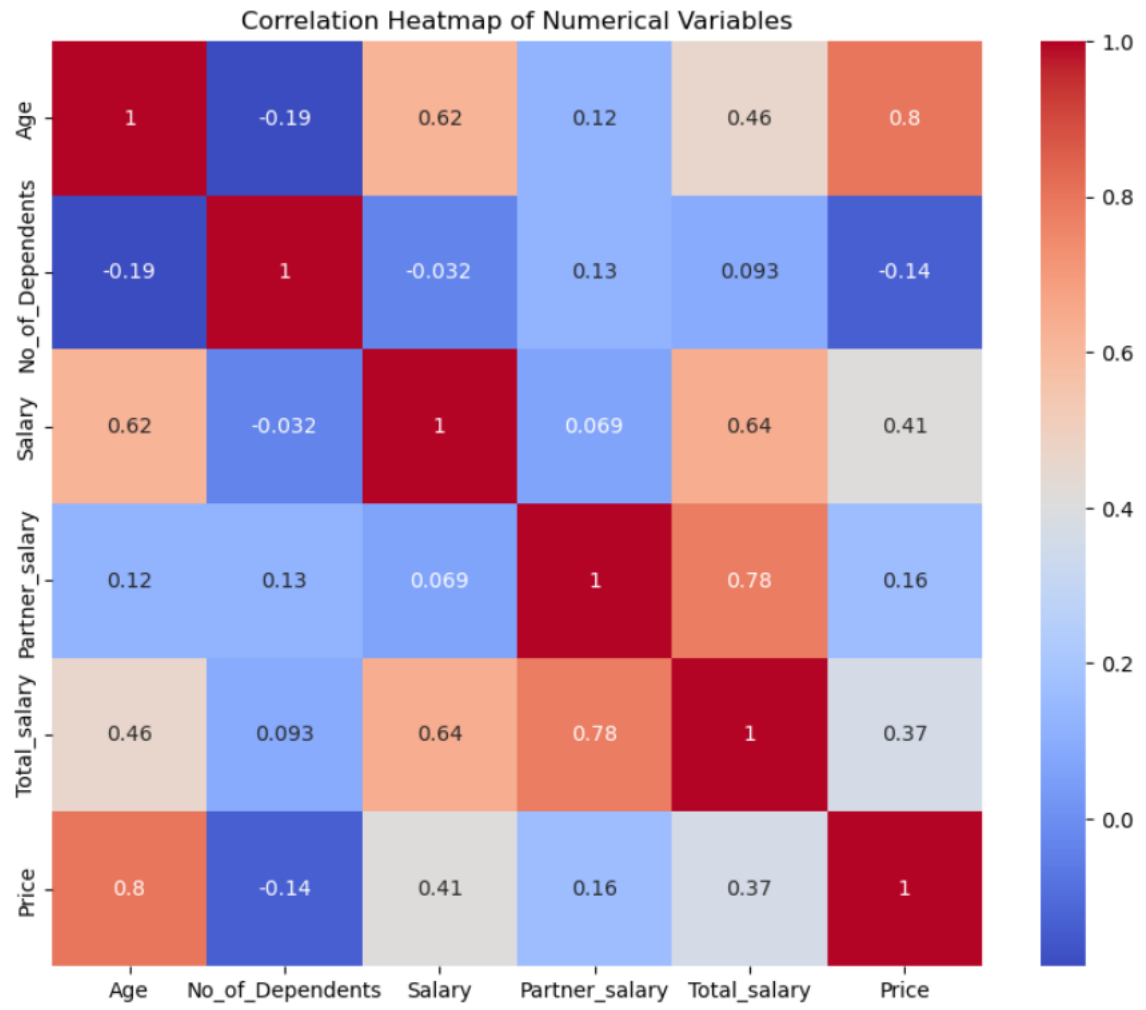
6. Price:
The median price is approximately $31,000.
The interquartile range (IQR) is from $25,000 to $47,000.
There are no outliers.
The price distribution is slightly skewed to the right.

7. Categorical data Observations:
Some data entries in the "Gender" variable contain spelling errors, which have been corrected.
The data predominantly consists of males.
There are more salaried personnel than business personnel.
The dataset contains more married individuals than single individuals.
Post-graduate individuals out number graduate individuals.
Approximately 50% of the dataset has personal loans.
The majority of individuals do not have a house loan.
Most individuals have a working partner.
Sedan-type automobiles are more prevalent than hatchbacks and SUVs

Bivariate Analysis

Exploring the correlation between all numerical variables.

We plot a heatmap that will show correlations between the numerical variable.



Observations:
1.Strong Positive Correlations:
Total salary is strongly positively correlated with Salary and Partner salary.
This indicates that as Salary and Partner salary increase, Total_salary also tends to increase.

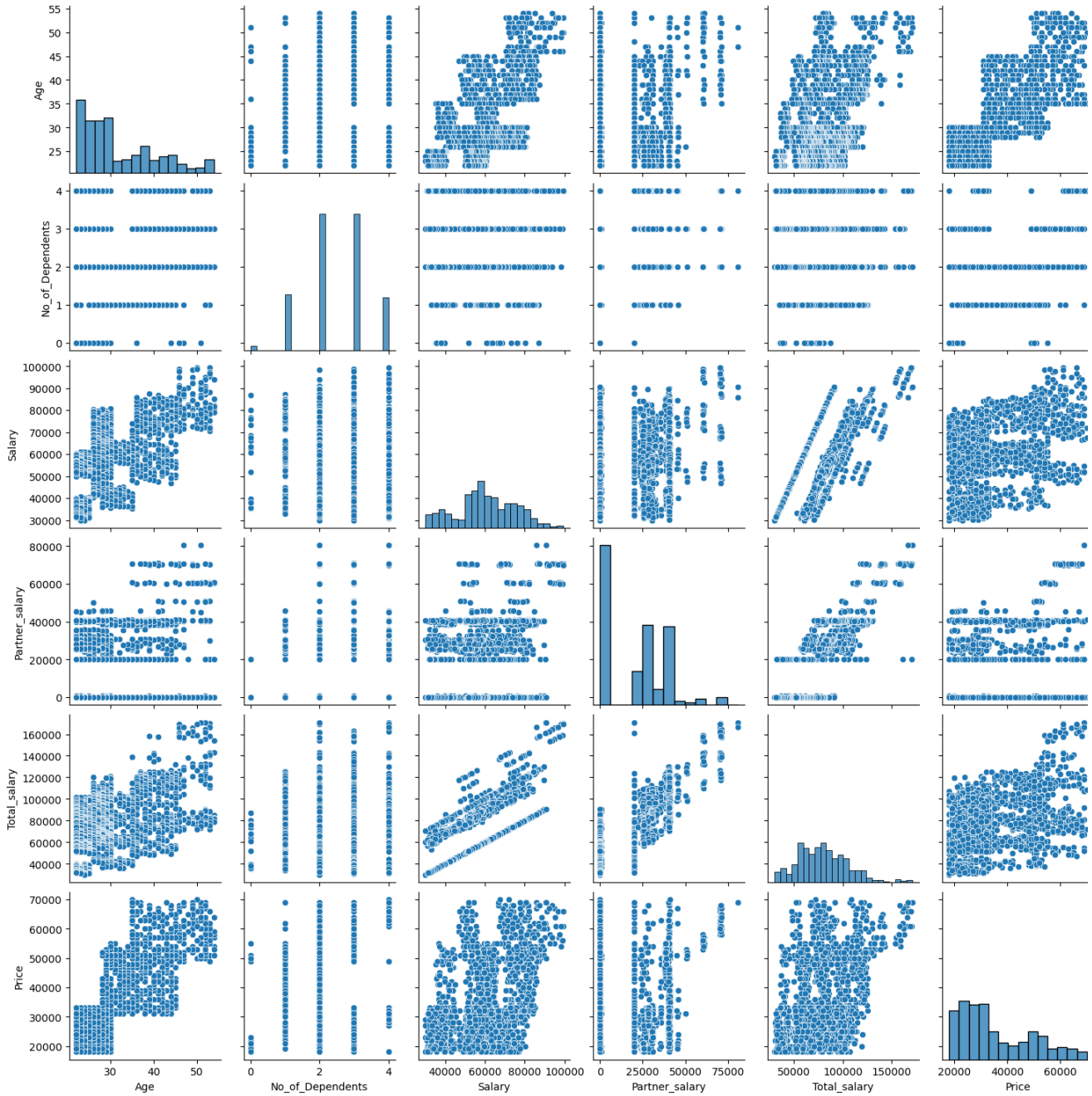
2.Negative Correlations:
Total salary is negatively correlated with Age and No of Dependents.
This suggests that as Age and No of Dependents increase, Total salary tends to decrease.

3.Weak Correlations:
The correlations between Total salary and Gender and between Total_salary and Education are relatively weak.

This implies that these factors have a less significant influence on Total salary.

Exploring the relationship between all numerical variables

We plot a Pair plot that will show relationship between the numerical variable



Observations:
1.Positive Linear Relationship:
The scatter plot between Total salary and Salary shows a positive linear relationship. As Salary increases, Total salary also tends to increase.

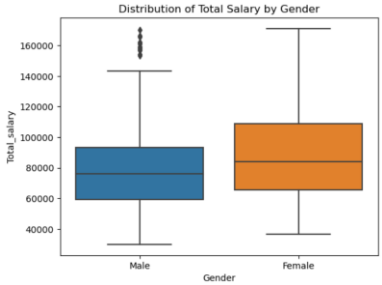
2.Outliers:
There are a few data points that appear as outliers in the scatter plots involving Total_salary. These outliers may represent individuals with unusually high or low salaries compared to the rest of the population.

3.Skewness:
The histograms for Age and No of Dependents indicate that these variables are skewed. This means that the distribution of values is not symmetrical.

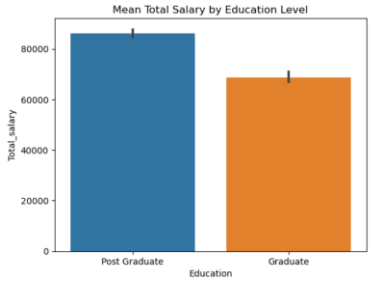
Exploring the relationship between categorical vs numerical variables

Data Name
Total salary VS Gender
Observations:
1.The median 'Total_salary' for males is lower than that of females.
2.There is a wider range of 'Total_salary' values for females compared to males.
3.There are outliers.

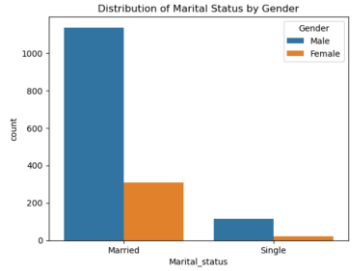
Visualization



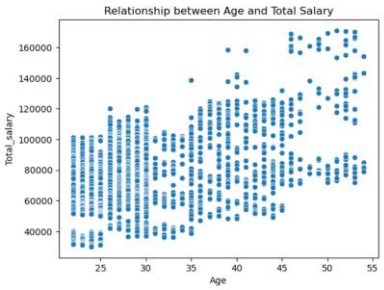
Total salary VS Education
Observations:
1.Individuals with a higher level of education tend to have a higher 'Total_salary'.



Marital status VS Gender
Observations:
1.The majority of individuals are married.
2.Among married individuals, there are slightly more males than females.
3.Among single individuals, there are more females than males.



Total salary VS Age
Observations:
1.There is a positive correlation between 'Age' and 'Total_salary'.
2.Younger individuals tend to have lower Total salary, while older individuals tend to have higher Total salary.



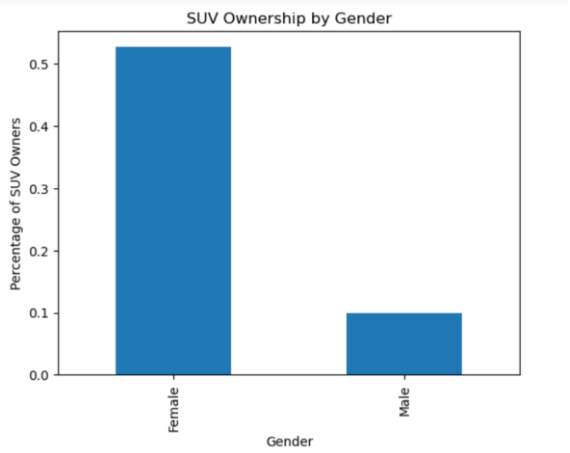
Observation

1. Gender Analysis: The median Total salary for males is lower than that for females, and there is a wider range of Total salary values among females.
2. Education Impact: Individuals with higher levels of education tend to command higher Total salary.
3. Marital Status: The majority of individuals are married, with a slightly higher proportion of males among married individuals. Single individuals show a higher proportion of females compared to males.
4. Age Influence: Younger individuals tend to have lower Total salary, while older individuals tend to have higher Total salary.

Key Questions

Q1. Do men tend to prefer SUVs more compared to women?

A1: No Male do not prefer SUV compared to female. From below plot we can see that Female Owns SUV more compare to Male.



Q2. What is the likelihood of a salaried person buying a Sedan?

A2: The likelihood of a salaried person buying a Sedan is: 44.2%

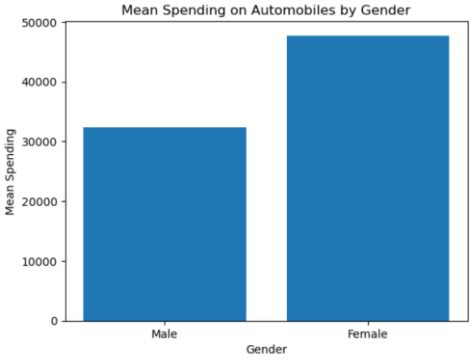
Q3. What evidence or data supports Sheldon Cooper's claim that a salaried male is an easier target for a SUV sale over a Sedan sale?

A3: The data does not support Sheldon Cooper's claim. As Percentage of salaried female owning SUV was more compared to male

- Percentage of salaried males who own an SUV: 13.39%
- Percentage of salaried females who own an SUV: 52.67%

Q4. How does the amount spent on purchasing automobiles vary by gender?

A4: On average, females tend to spend more on automobiles compared to males. From Below plot we can see that.

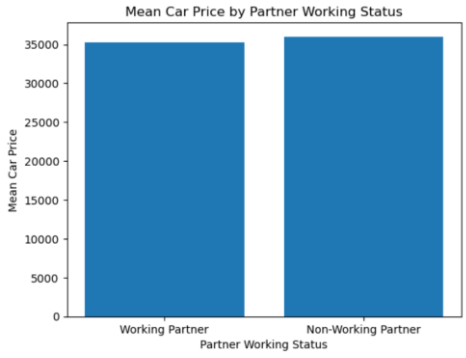


Q5. How much money was spent on purchasing automobiles by individuals who took a personal loan?

A5: Total amount spent on automobiles by individuals who took a personal loan: 27290000 dollars.

Q6. How does having a working partner influence the purchase of higher-priced cars?

A6: The mean price of cars purchased by individuals with working partners is less compared to those without working partners. We can observe that, on average, individuals without working partners tend to purchase cars at a slightly higher price compared to those with working partners. However, the difference in mean prices between the two groups is relatively small. It appears that having a working partner might not significantly influence the purchase of higher priced cars.



Actionable Insights & Recommendations

Actionable Insights:

1. Target Females for SUV Sales: The analysis shows that females are more likely to own SUVs compared to males. Sales and marketing efforts for SUVs should focus on male consumers.

2. Promote Sedans to Salaried Individuals: Salaried individuals are more likely to buy Sedans compared to other professions. Sales representatives should prioritize promoting Sedans to salaried individuals.

3. Tailor Marketing Strategies by Gender: Males and females have different preferences when it comes to automobiles. Marketing strategies should be tailored to the specific preferences of each gender. Male Perfer Sedan & hatcbacks more than SUV compare to Female.

4. Consider Loans for Higher Sales: Individuals who have taken loans tend to spend more on automobiles. Offering loan options can encourage customers to make higher value purchases.

Business Recommendations

1. Target High Income Earners: Focus marketing efforts on individuals with higher incomes, as they are more likely to have the financial means to purchase expensive automobiles.

2. Offer Incentives for Sedan Purchases: Provide attractive incentives, such as discounts or extended warranties, to encourage the purchase of Sedans among individuals.

3. Personalize Marketing Messages: Leverage customer data to personalize marketing messages based on factors such as gender, income, and profession. This can enhance the effectiveness of marketing campaigns.

4. Enhance Customer Experience: Prioritize customer satisfaction by providing excellent customer service, offering flexible financing options, and ensuring a seamless purchasing experience.

5. Expand Product Offerings: Consider expanding the product line to include a wider range of vehicle models and price points to cater to diverse customer needs and preferences.

6. Conduct Market Research: Regularly conduct market research to stay updated on changing customer preferences, emerging trends, and competitive dynamics.

7. Analyze Customer Feedback: Actively collect and analyze customer feedback to identify areas for improvement and make necessary adjustments to marketing strategies and product offerings.

Problem 2

Data Overview

We have Data file that need to be analyzed had been load and we have data description as follow: -

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_data - 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

Structure of the Data:

Data had 28 columns and 8448 rows.
First 5 rows are:

userid	card_no	card_bin_no	Issuer	card_type	card_source_date	high_networth	active_30	active_60	active_90	...	bank_vintage	T+1_month_activity	T+2
1	4384 39XX XXXX XXXX	438439	Visa	edge	2019-09-29	B	0	1	1	...	27	0	
2	4377 48XX XXXX XXXX	437748	Visa	prosperity	2002-10-30	A	1	1	1	...	52	0	
3	4377 48XX XXXX XXXX	437748	Visa	rewards	2013-10-05	C	0	0	0	...	23	1	
4	4258 06XX XXXX XXXX	425806	Visa	indianoil	1999-06-01	E	0	1	1	...	49	0	
5	4377 48XX XXXX XXXX	437748	Visa	edge	2006-06-13	B	1	1	1	...	21	1	

Framing Analytics Problem

1. Problem Statement: Given a dataset containing information about customers of a credit card company, develop a predictive model to identify card payments.
Business Objective: The goal is to develop a model that can accurately predict the likelihood of a customer defaulting on their credit card payments. This information can then be used to take proactive measures to prevent defaults, such as offering early intervention programs or adjusting credit limits.

2. Problem Statement: Identify the key factors that influence customer spending on credit cards and develop targeted marketing strategies to increase revenue.
Business Objective: The goal is to identify the key variable that drive the credit card spending. Develop predictive models to estimate customer spending. Segment customers into distinct groups based on spending behavior. Design targeted marketing campaigns for each customer segment.

Question - Analyse the dataset and list down the top 5 important variables, along with the business justifications.
Answer - Top 5 import Variable will be:-

1. active_90: This variable represents whether the user has been active in the last 90 days. It is crucial for understanding user engagement and retention. High values indicate active users, which are vital for the company's business success and revenue generation.
2. annual_income_at_source: This variable denotes the annual income of the user. It is essential for segmenting customers based on their income levels and targeting them with appropriate products and services. Understanding the income distribution helps in tailoring marketing strategies and product offerings to different customer segments.
3. cc_active90: This variable indicates whether the user's credit card has been active in the last 90 days. It provides insights into the usage patterns of credit cards among users. High values suggest active credit card users, which can be valuable for promoting credit card related products and services.
4. avg_spends_l3m: This variable represents the average spending of the user over the last 3 months. It is critical for assessing the spending behavior and financial health of users. High spending levels indicate potential high value customers who may be interested in premium products or services.
5. Occupation_at_source: This variable captures the occupation of the user. Understanding the occupation profile of customers helps in creating targeted marketing campaigns and personalized offerings. It enables the company to tailor its products and services to meet the specific needs and preferences of different occupational groups.