Statistical Methods for Decision Making

PROJECT REPORT

Gowdham S
PGPDSA.O.DEC23.A

Table of Contents

List of Figures

List of Equations

List of Tables

Data Dictionary of Problem 1

Column Name	Description	Data Type
age	The age of the individual in years.	int64
gender	The gender of the individual, categorized as male or female.	object
profession	The occupation or profession of the individual.	object
marital_status	The marital status of the individual, such as married &, single	object
education	The educational qualification of the individual Graduate and Post Graduate	object
no_of_dependents	The number of dependents (e.g., children, elderly parents) that the individual supports financially.	int64
personal_loan	A binary variable indicating whether the individual has taken a personal loan "Yes" or "No"	object
house_loan	A binary variable indicating whether the individual has taken a housing loan "Yes" or "No"	object
partner_working	A binary variable indicating whether the individual's partner is employed "Yes" or "No"	object
salary	The individual's salary or income.	int64
partner_salary	The salary or income of the individual's partner, if applicable.	float64
Total_salary	The total combined salary of the individual and their partner (if applicable).	int64
price	The price of a product or service.	int64
make	The type of automobile	object

Data Dictionary of Problem 2

Column Name	Description	Data Type
userid	Unique bank customer-id	int64
card_no	Masked credit card number	object
card_bin_no	Credit card IIN number	int64
Issuer	Card network issuer	object
card_type	Credit card type	object
card_source_data	Credit card sourcing date	datetime64
high_networth	Customer category based on their net-worth value (A: High to E: Low)	object
active_30	Savings/Current/Salary etc. account activity in last 30 days	int64
active_60	Savings/Current/Salary etc. account activity in last 60 days	int64
active_90	Savings/Current/Salary etc. account activity in last 90 days	int64
cc_active30	Credit Card activity in the last 30 days	int64
cc_active60	Credit Card activity in the last 60 days	int64
cc_active90	Credit Card activity in the last 90 days	int64
hotlist_flag	Whether card is hot-listed(Any problem noted on the card)	object
widget_products	Number of convenience products customer holds (dc, cc, net-banking active, mobile banking active, wallet active, etc.)	int64
engagement_products	Number of investment/loan products the customer holds (FD, RD, Personal loan, auto loan)	int64
annual_income_at_source	Annual income recorded in the credit card application	int64
other_bank_cc_holding	Whether the customer holds another bank credit card	object
bank_vintage	Vintage with the bank (in months) as on Tthmonth	int64
T+1_month_activity	Whether customer uses credit card in T+1 month (future)	int64
T+2_month_activity	Whether customer uses credit card in T+2 month (future)	int64
T+3_month_activity	Whether customer uses credit card in T+3 month (future)	int64
T+6_month_activity	Whether customer uses credit card in T+6 month (future)	int64
T+12_month_activity	Whether customer uses credit card in T+12 month (future)	int64
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.	object
avg_spends_l3m	Average credit card spends in last 3 months	int64
Occupation_at_source	Occupation recorded at the time of credit card application	object
cc_limit	Current credit card limit	int64

Problem 1

Context

Austo Motor Company, a prominent automobile manufacturer known for its SUV, Sedan, and Hatchback models, addressed concerns about the effectiveness of its current marketing campaign during its latest board meeting.

Objective

To examine the data to gain a comprehensive understanding of customer demand, ultimately enhancing their customer experience. Conduct data analysis to uncover insights that will assist the company in enhancing its business operations.

Problem 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.

Solution of Data Overview:

1. We have imported all the required libraries such as,
   - Numpy
   - Pandas
   - Matplotlib
   - Seaborn
   - Scipy and
   - Warnings - to ignore the warning messages

2. After importing all the necessary modules, we then proceed with reading the dataset. The dataset is provided in the form of .csv file. So we've used pandas **read_csv()** method to read the dataset and assigned it to the variable 'df'.
3. We used **shape** attribute of pandas library [df.shape] to determine the number of rows and columns of the dataframe df.

Shape	Description
(1581, 14)	We have 1581 rows and 14 columns in the dataset

Table 1 Shape of df

4. To check the types of data, we used **info()** method of pandas library to list the basic information of the data such as rows count, columns count and datatype of the columns.

Column	Non-Null Count	Data Type
Age	1581 non-null	int64
Gender	1528 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	1475 non-null	float64
Total_salary	1581 non-null	int64
Price	1581 non-null	int64
Make	1581 non-null	object

Table 2 Basic Information of df

- From the above table, We have **8** object, **5** int and **1** float data types in the dataset. Also we can see that there are **53** and **106** null values in Gender and Partner_salary columns respectively.

Columns	Data Type	No of Missing Values
Gender	object	53
Partner_salary	float64	106

Table 3 Null values count

5. The data type of Gender column is object and Partner_salary is float. So we will impute the missing values of Gender with **mode** as Gender is a **categorical** column.

- In the Partner_salary column we have 106 null values, out of which 90 null values does not have their partner working and 16 values have their partner working. Then we compared Total_salary and salary of those 90 values and found all the 90 values are equal. So we can proceed with imputing Total_salary - salary for those 106 columns
- We can notice that in column “Gender” the spelling of Female is incorrect resulting in 2 more unique values i.e.“Femal” & “Femle”.

Column	Unique Values
Gender	Male, Femal, Female, Femle, nan
Profession	Business, Salaried
Marital_status	Married, Single
Education	Post, Graduate, Graduate
No_of_Dependents	4, 3, 2, 1, 0
Personal_loan	No, Yes
House_loan	No, Yes
Partner_working	Yes, No
Make	SUV, Sedan, Hatchback

Table 4 Unique values of Categorical Columns

- We replaced the incorrect spelling in Gender column with **replace()** method. After replacing the incorrect values, we have **329 Female** values and **1199 Male** values and **53 missing** values.
- Now, We can find the mode of Gender using mode() method.

Columns	Type	Mode
Gender	Categorical	Male

Table 5 Mode Value of Gender

- After finding the mode and Total_salary - Salary values, we can impute the missing values with these values for Gender and Partner_salary using **where()** method. After this imputation, we confirmed that there are no missing values.

6. Once the above is done, we found the statistical summary of the dataset using **describe** method().

Index	Count	Top	Mean	Std	Min	25%	50%	75%	Max
Age	1581	NaN	31.9222	8.425978	22	25	29	38	54
Gender	1581	Male	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Profession	1581	Salaried	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Marital_status	1581	Married	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Education	1581	Post Graduate	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_Dependents	1581	NaN	2.457938	0.943483	0	2	2	3	4
Personal_loan	1581	Yes	NaN	NaN	NaN	NaN	NaN	NaN	NaN
House_loan	1581	No	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Partner_working	1581	Yes	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	1581	NaN	60392.22	14674.83	30000	51900	59500	71800	99300
Partner_salary	1581	NaN	19233.78	19670.39	0	0	25100	38100	80500
Total_salary	1581	NaN	79626	25545.86	30000	60500	78000	95900	171000
Price	1581	NaN	35597.72	13633.64	18000	25000	31000	47000	70000
Make	1581	Sedan	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 6 Statistical Summary

7. As discussed in Point 5, We had discrepancy in Gender column and treated them with correct values using replace method. Also the data type of Partner_salary is float64 but, for Salary and Total Salary it is int64. So we are converting data type of Partner_salary from float64 to int64 using astype('int64').

Column	Unique Values
Gender	Male, Female

Table 7 Unique Values of Gender after Treatment

Observations and Insights:

Based on the data cleaning and exploration we've conducted, here are some key observations and insights

- **Gender Discrepancy and Correction:** We identified and corrected discrepancies in the "Gender" column, standardizing the values to 'Male', 'Female', and handled missing values. This correction ensures accurate gender representation in the dataset.
- **Missing Values:** The dataset contained missing values in "Gender" and "Partner_salary" columns. We handled the missing values using mode and Total_salary - Salary imputation for "Gender" and "Partner_Salary" respectively.
- **Data Consistency:** Other columns such as 'Profession', 'Marital_status', 'Education', 'Personal_loan', 'House_loan', 'Partner_working', and 'Make' showed consistent and expected unique values, indicating good data quality in these fields.
- **Data Types:** The dataset primarily consists of categorical data ('Gender', 'Profession', 'Marital_status', 'Education', Personal_loan', 'House_loan', 'Partner_working', 'Make') and numerical data ('Age', 'No_of_Dependents', 'Salary', 'Partner_salary', 'Price', 'Total Salary'). This diversity in data types suggests that analysis could include both statistical summaries and categorical data analysis.
- **Sedan** is the most sold type of car and **Salaried Married males** are the top customer. The median age of customers is **29** and the median salary of customers is **59500**. The median price of cars is **31000**

These insights provide a foundation for further analysis, such as exploring relationships between gender and salary, the impact of marital status on car preferences, or the correlation between education level and loan status. Further statistical analysis could uncover more nuanced insights within this dataset.

Problem 1 - Univariate Analysis

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

Solution of Univariate Analysis:

1. First, we selected the numerical and categorical columns from the dataframe and assigned them to the varables '**numerical_cols**' and '**categorical_cols**' using select_dtypes().

Variable name	Columns
numerical_cols	Age, No_of_Dependents, Salary, Partner_salary, Price, Total Salary
categorical_cols	Gender, Profession, Marital_status, Education, Personal_loan, House_loan, Partner_working, Make

Table 8 Numerical & Categorical Columns

2. After separating the numerical and categorical columns, we created boxplots for each of the numerical columns using searborn library

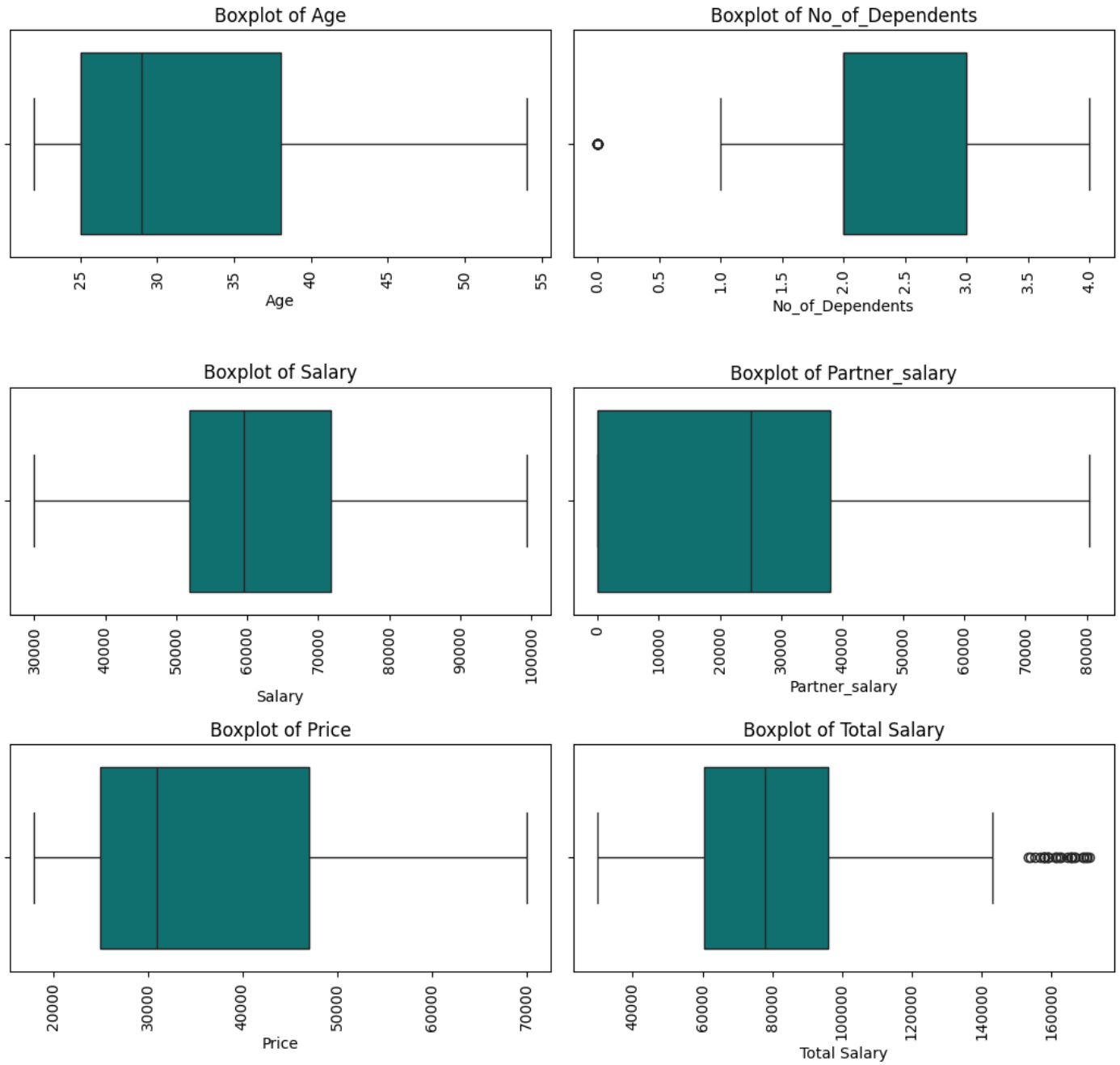

Figure 1 Box Plot of Numerical Columns

Observations:

- From the above Box Plots, we can see that there are **outliers** in No_of_Dependents and Total Salary columns.
- We can also see the five-number summary(Min, Q1, Median, Q3, Max) for the numerical columns from the box plots and .

Column	Min	25%	50%	75%	Max
Age	22	25	29	38	54
No_of_Dependents	0	2	2	3	4
Salary	30000	51900	59500	71800	99300
Partner_salary	0	0	25100	38100	80500
Total_salary	30000	60500	78000	95900	171000
Price	18000	25000	31000	47000	70000

Table 9 Five Number Summary

3. After creating Box Plots for numerical columns, we plotted the distribution of numerical columns using histogram with **KDE** value as True from the seaborn library.

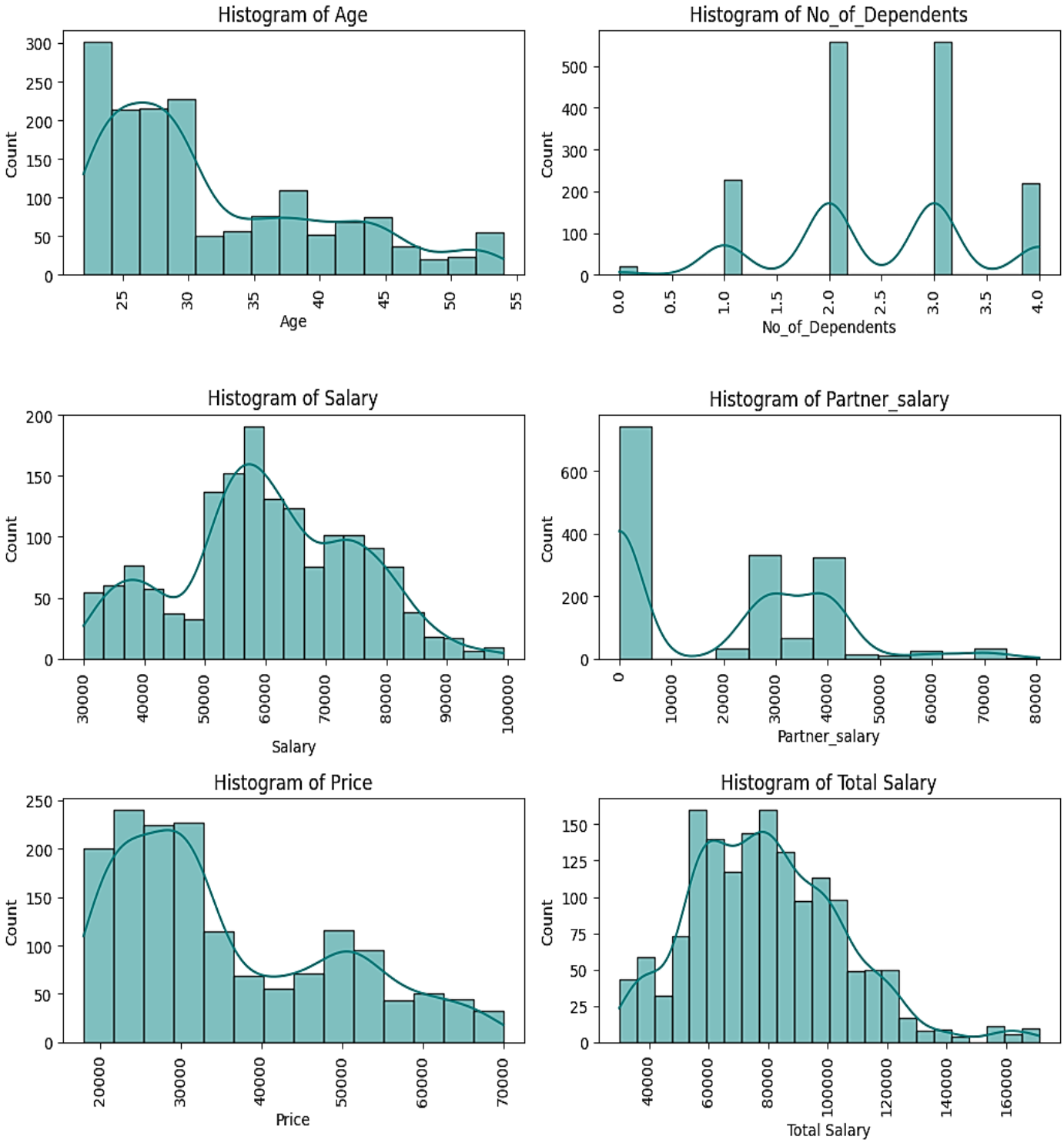

Figure 2 Histogram of Numerical Columns

Observations:

- We can see the distribution of values of the numerical columns from the Histogram.
- The distribution is somewhat normal in the Total Salary column, but it is also skewed at right.
- In case of Distribution of Price, it has similar distribution to Bimodal distribution as we have 2 bumps.

4. After Plotting the Histogram of numerical columns, we then proceeded with plotting Count plot of Categorical columns which shows us the frequency of categorical variables within the columns.

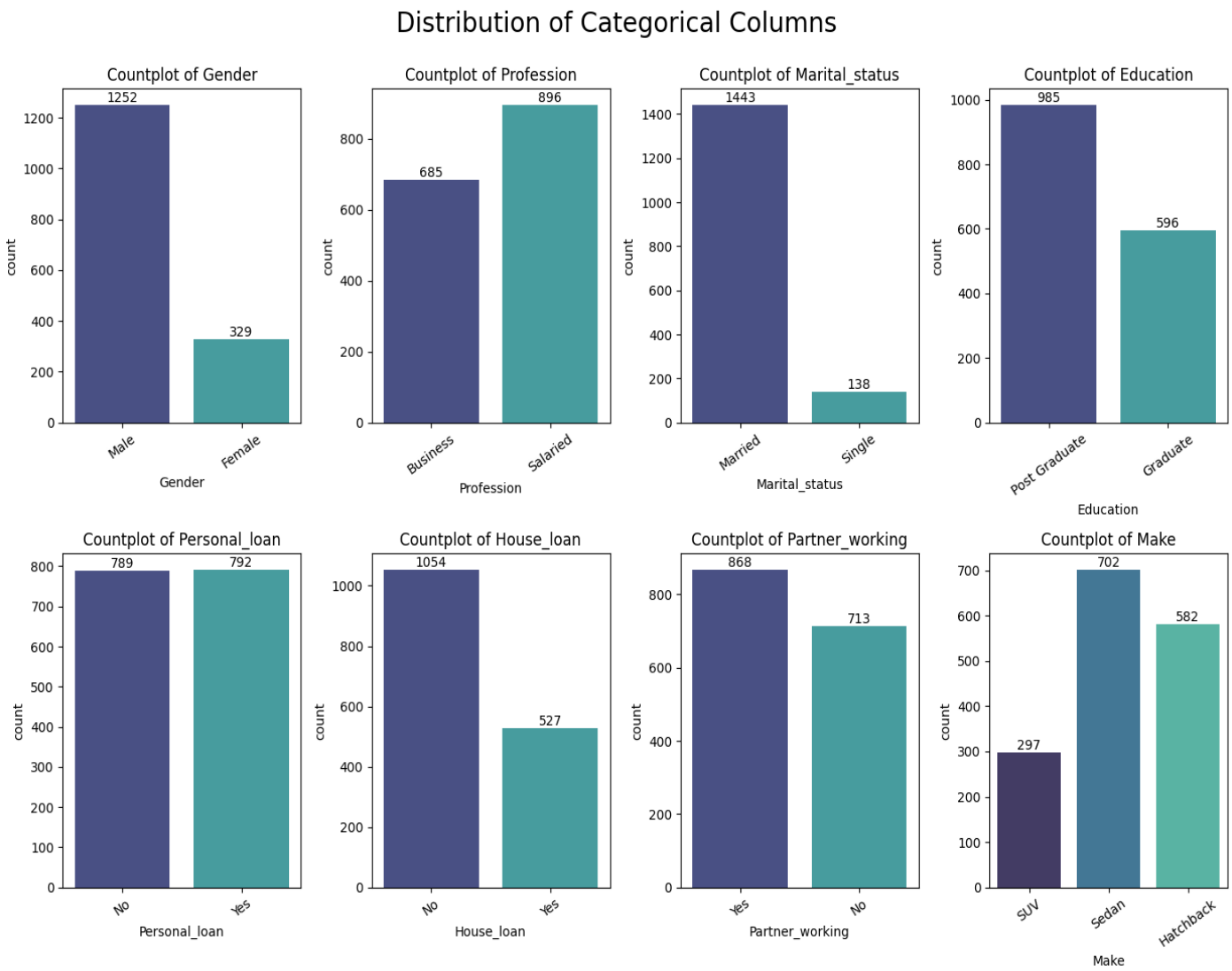

Figure 3 Count Plot of Categorical Columns

Observations:

- There are more male customers than female customers.
- The count of salaried professionals exceeds that of business professionals.
- Married customers outnumber single customers.
- The majority of customers have a post-graduation educational qualification.
- The number of customers who opted for a personal loan is approximately equal to those who did not.
- There are more customers without a home loan compared to those with one.
- Customers with working partners made the most car purchases.
- Sedans are the top-selling make, followed by hatchbacks and SUVs.

5. From Figure 1, we can see some outliers in No_of_Dependents and Total Salary columns. A box plot might visually highlight outliers, but it doesn't necessarily mean it should be treated as an anomaly. Outliers are more relevant when they represent data points significantly deviating from the normal and might be indicative of errors or anomalies in the data. However, a value of 0 for the number of dependents (No_of_Dependents) is a legitimate and reasonable observation.

   Since No_of_Dependents can be zero, we are considering 0 as relevant value and **not outlier**. In case of Total Salary column, the **outlier values are valuable and it is essential for the analysis of high paying customers**. So we are **not treating the outliers here**. But we can treat the outliers with the below formula,

```
Q1 = df['Total Salary'].quantile(0.25)
Q3 = df['Total Salary'].quantile(0.75)
IQR = Q3 - Q1

Lower Bound = Q1 - 1.5 * IQR
Upper Bound = Q3 + 1.5 * IQR
```

Equation 1 Calculation of Upper & Lower bound values

Observations and Insights:

The exploration of variables in the dataset reveals the following insights:

- **Numerical Variables:** The dataset contains six numerical columns: 'Age', 'No_of_Dependents', 'Salary', 'Partner_salary', 'Total Salary', and 'Price'. These variables are key for any quantitative analysis, such as understanding salary distributions or price ranges of automobiles.
- **Categorical Variables:** There are eight categorical columns: 'Gender', 'Profession', 'Marital_status', 'Education', 'Personal_loan', 'House_loan', 'Partner_working', and 'Make'. These variables are crucial for segmenting the data and analyzing trends across different groups.
- **Outliers Detection:** Boxplots were generated for each numerical variable to visually inspect for outliers. Outliers can significantly impact statistical analysis, so it's important to decide whether to keep, remove, or adjust these values based on the analysis of data.
- **Decision Making:** The company should prioritize crafting marketing strategies that enhance the number of female and unmarried customers. Additionally, it should concentrate on boosting the sales of SUVs.

Problem 1 - Bivariate Analysis

Explore the relationship between all numerical variables - Explore the correlation between all numerical variables - Explore the relationship between categorical vs numerical variables

Solution of Bivariate Analysis:

1. We plotted a PairGrid Figure 4 from seaborn library, in which we used KDE plot for the upper corner and Scatter plot for the lower corner. The KDE plot shows the spread of data points between the axis and the Scatter plot shows us the relation between the axis. The diagonal shows us the histogram of the columns.

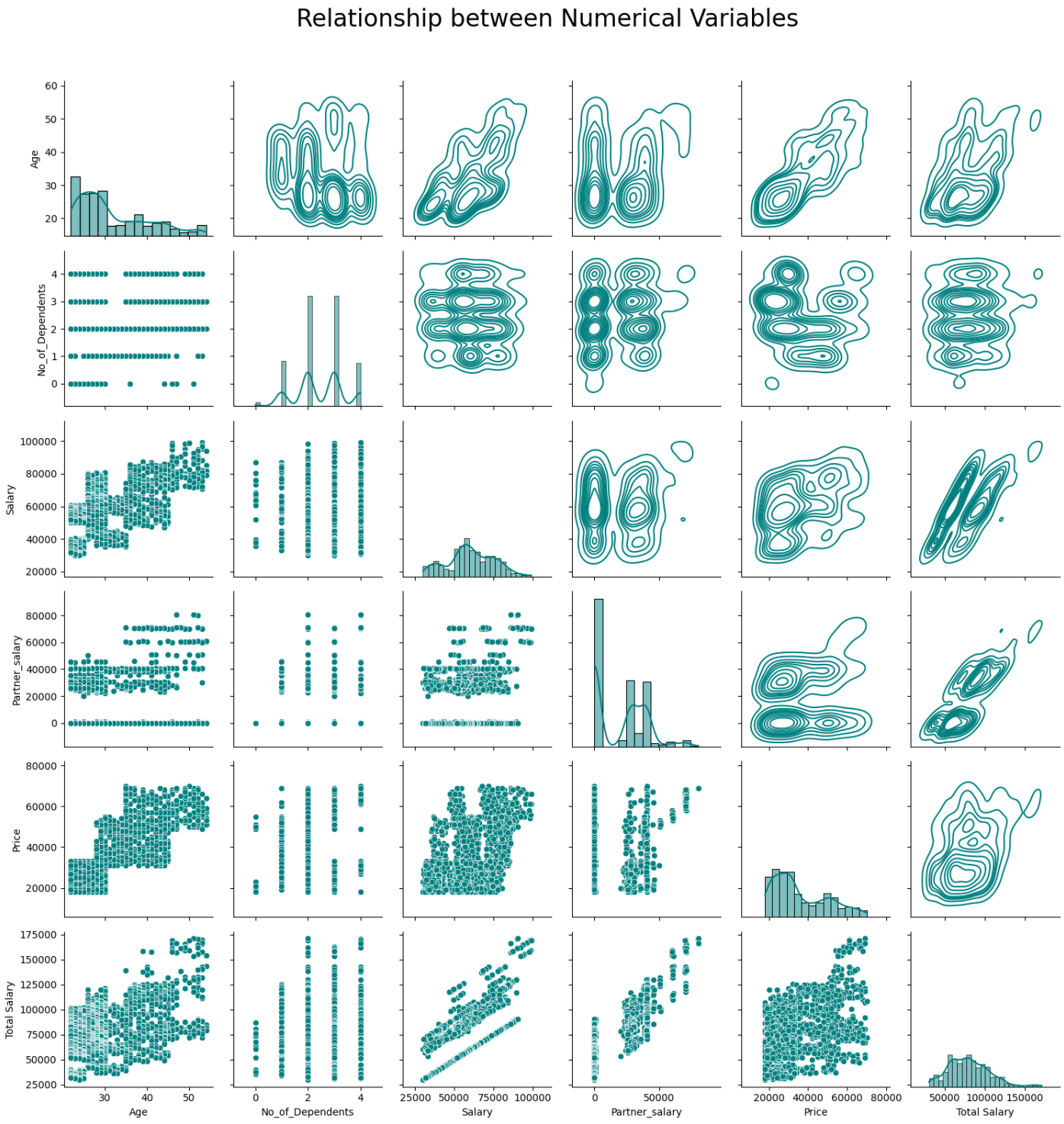

Figure 4 Pairgrid showing Relationship between Numerical Variables

Observations:

- Age and Salary have positive correlation i.e., With increase in Age, the salary also increases.
- The variables Salary and Price do have positive correlation by do not show a clear linear relationship.
- The variables Total Salary and Price do not show a clear linear relationship.
- The KDE plot between Total Salary and Partner_salary shows that there are some extreme data points.
- The variables Salary and Partner_salary do not show a clear linear relationship.

2. After plotting the Pairgrid of numerical variables, we calculated the correlation between the numerical variables and then plotted them on the Heatmap Figure 5.
   - We can see the correlation between the numerical variables from the below table Table 10 CorrelationTable

index	Age	No_of_Dependents	Salary	Partner_salary	Price	Total Salary
Age	1	-0.19	0.62	0.14	0.8	0.46
No_of_Dependents	-0.19	1	-0.03	0.14	-0.14	0.09
Salary	0.62	-0.03	1	0.09	0.41	0.64
Partner_salary	0.14	0.14	0.09	1	0.17	0.82
Price	0.8	-0.14	0.41	0.17	1	0.37
Total Salary	0.46	0.09	0.64	0.82	0.37	1

Table 10 CorrelationTable of Numerical variables

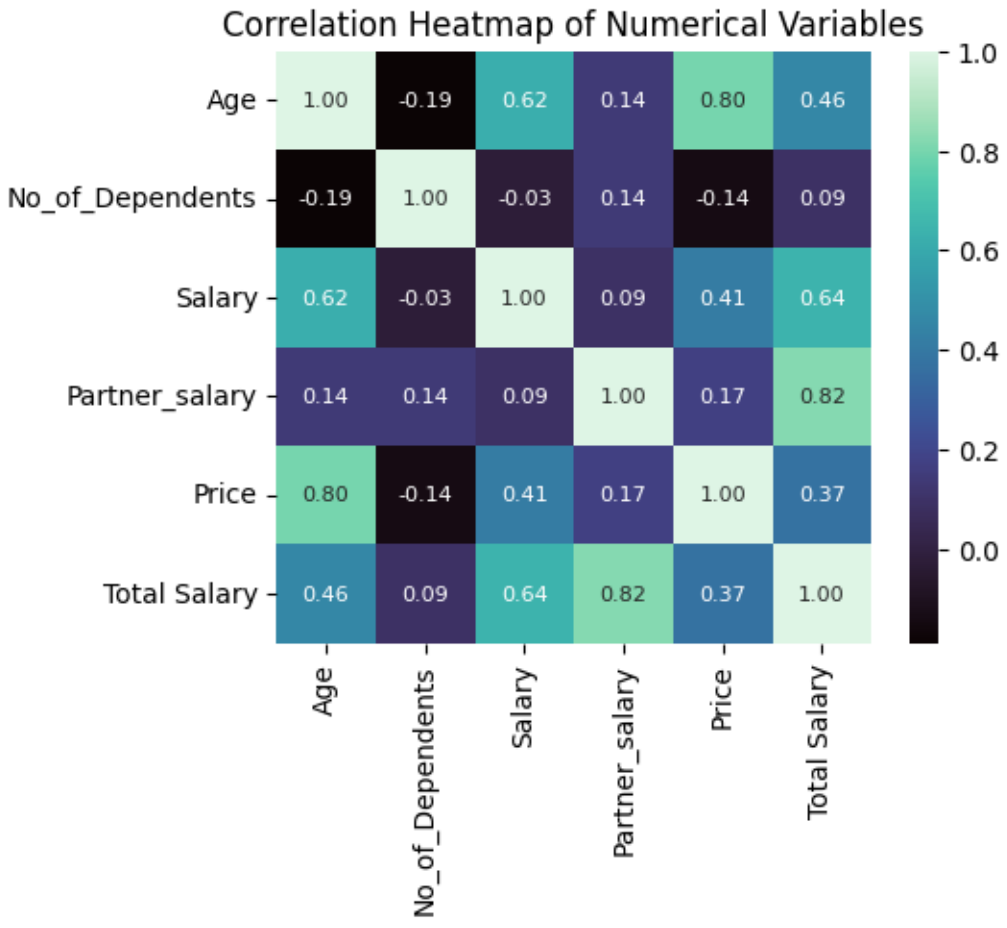

Figure 5 Correlation Heatmap of Numerical Columns

3. After generating the correlation heatmap, we proceeded with plotting categorical columns against numerical columns for bivariate analysis. Below are the violin plots illustrating the relationship between each categorical column and all the numerical columns.

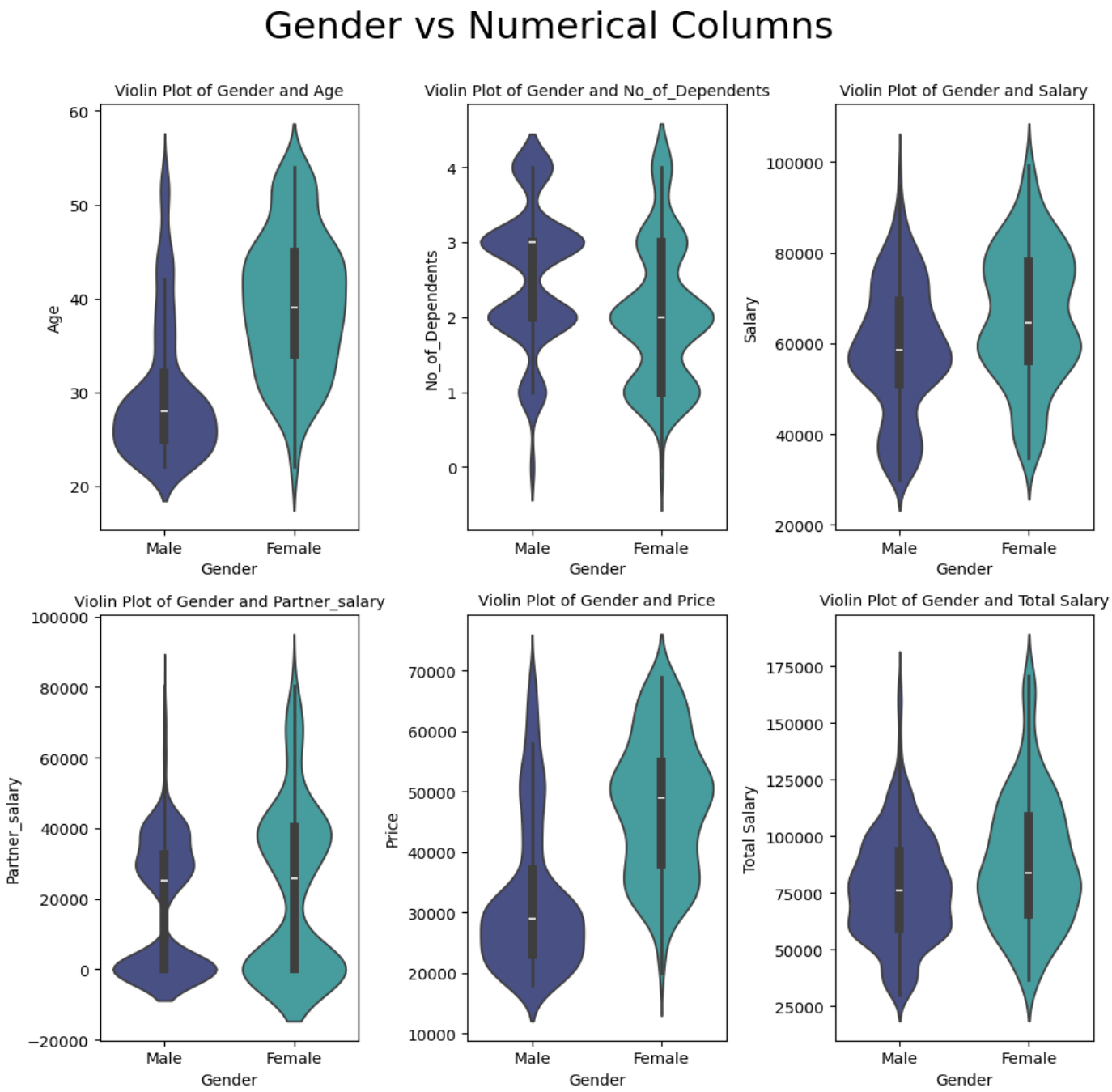

Figure 6 Violinplots showing Gender vs Numerical Columns

Observations:

- The median age of males is below 30, while the majority of females fall within the age range of 30 to 40.
- Most male customers have three dependents, whereas female customers typically have two dependents.
- The median salary of males is below 60,000, while females have a median salary just above 60,000.
- Male customers typically purchase cars with a median price below 30,000, whereas female customers tend to buy cars with a median price just under 50,000.

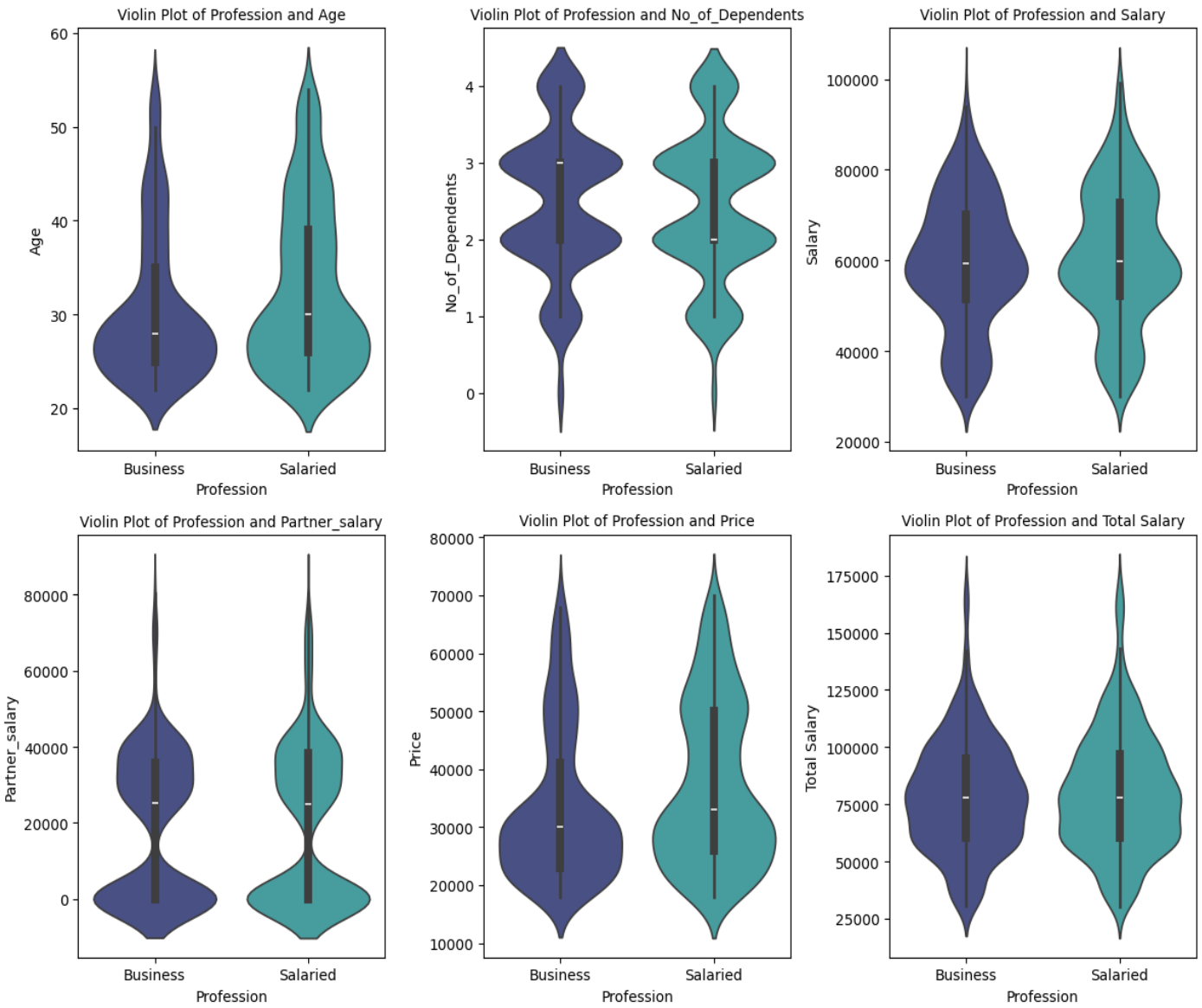

Figure 7 Violinplots showing Profession vs Numerical Columns

Observations:

- The median age of business customers is below 30, while salaried customers are primarily around 30 years old.
- Business customers commonly have three dependents, whereas salaried customers typically have two dependents.
- The median salary of both business and salaried customers is nearly identical.
- Similarly, the median partner salary of both business and salaried customers is also nearly equal.
- Business customers tend to purchase cars with a median price around 30,000, while salaried customers generally buy cars with a median price higher than that of business customers.

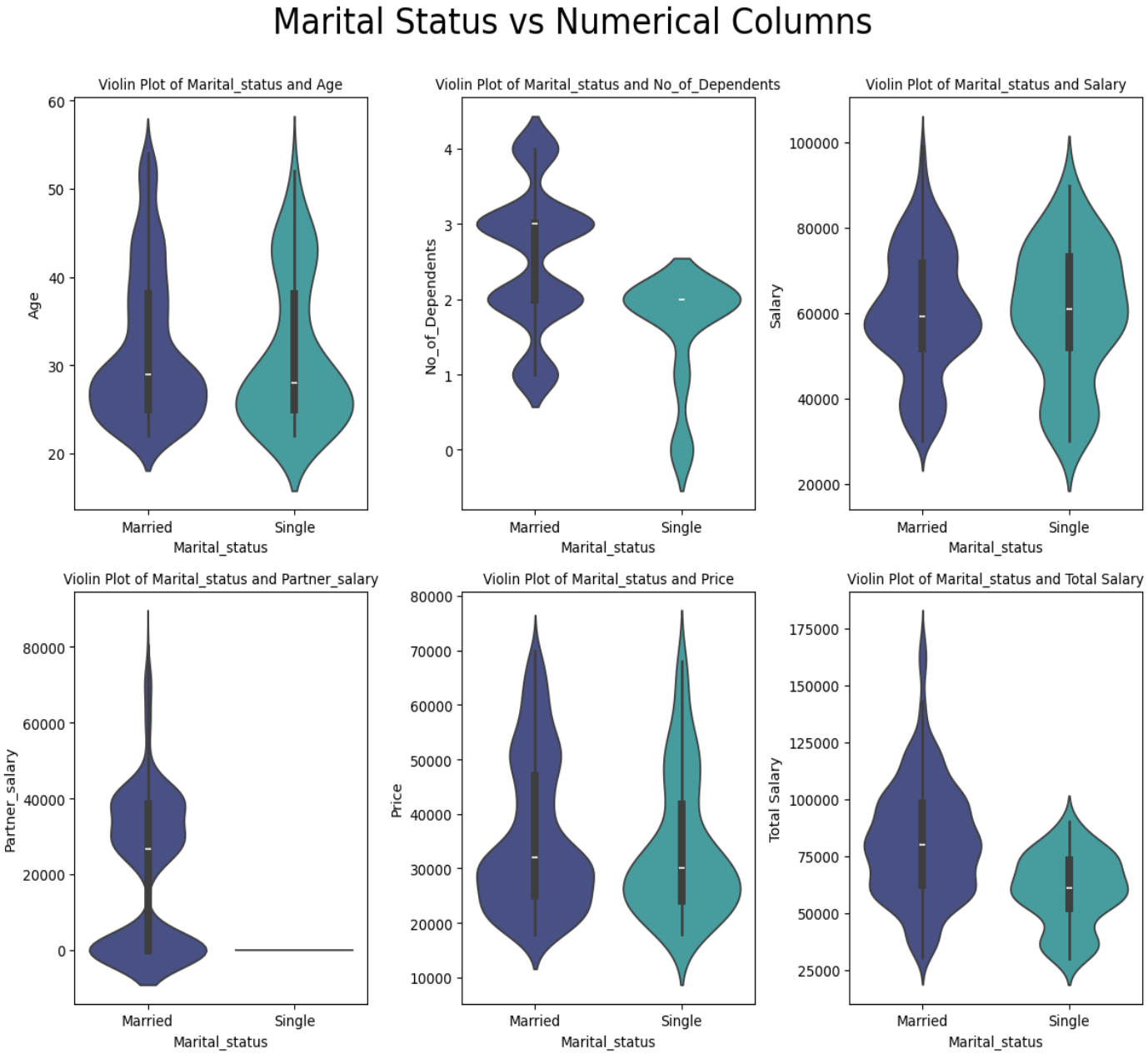

Figure 8 Violinplots showing Marital Status vs Numerical Columns

Observations:

- The median age of both married and single customers falls below 30 years.
- Married customers tend to have three dependents, while single customers typically have two dependents.
- The median salary of both married and single customers is nearly equal, hovering around 60,000.
- The median partner salary of married customers typically ranges around 30,000.
- Both married and single customers have allocated a similar amount towards the purchase price.

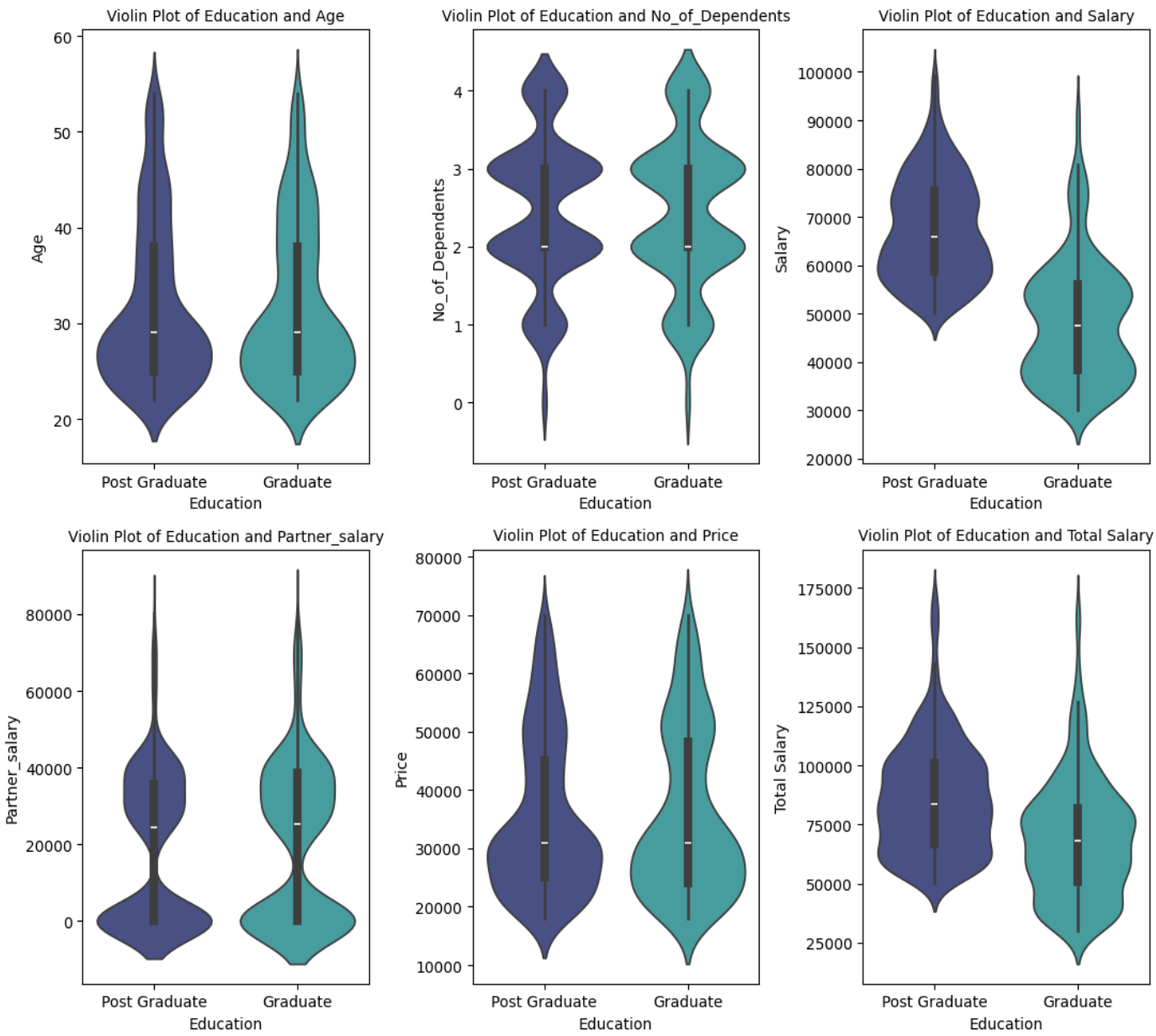

Figure 9 Violinplots showing Education vs Numerical Columns

Observations:

- The median age of both graduate and post-graduate customers is below 30 years.
- Both graduate and post-graduate customers typically have two dependents.
- The median salary of post-graduate customers is approximately 65,000, whereas the median salary of graduate customers is just below 50,000.
- The median partner salary of both graduate and post-graduate customers typically ranges between 20,000 to 30,000.
- Both graduate and post-graduate customers have spent a similar amount on the purchase price.

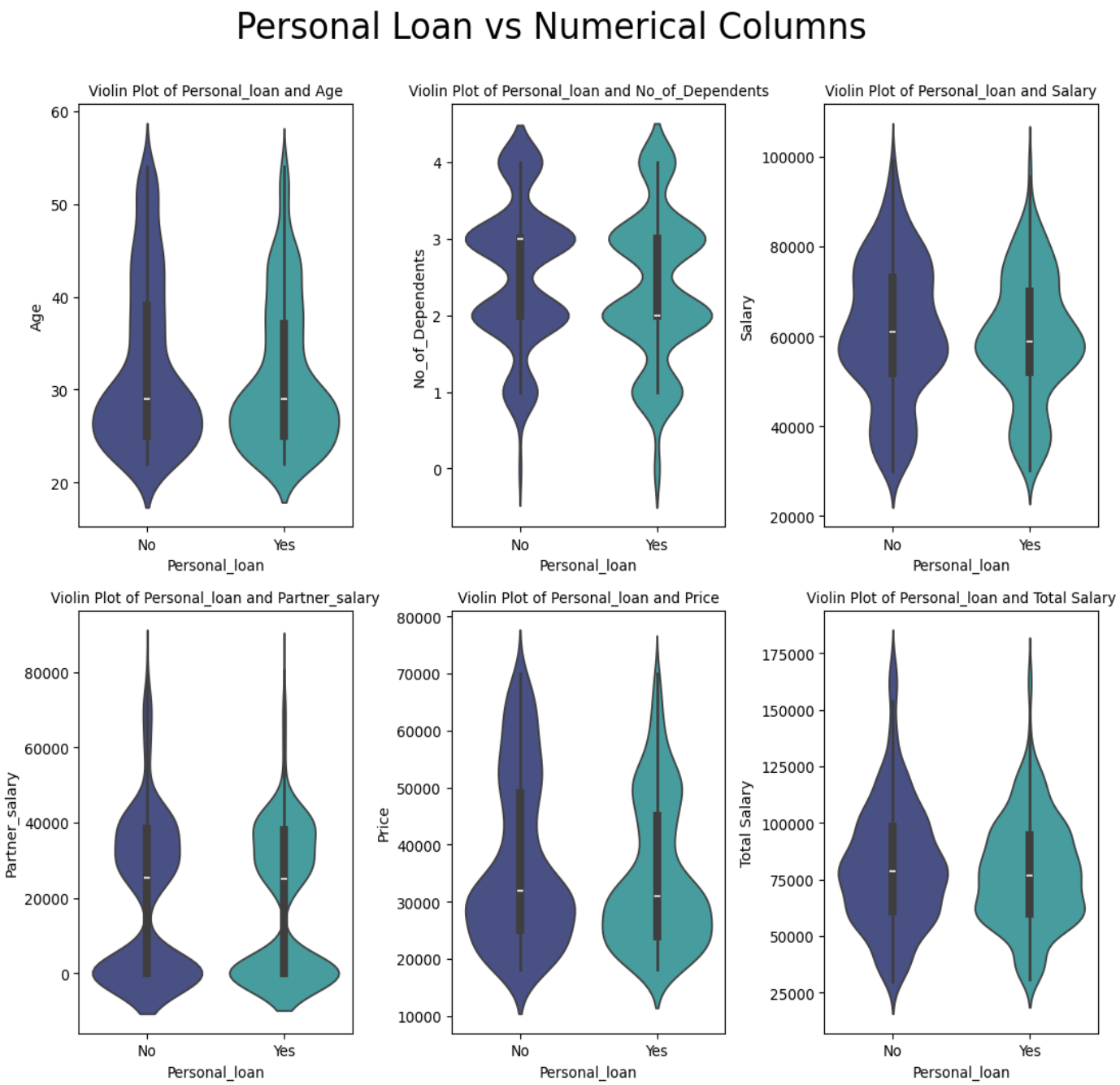

Figure 10 Violinplots showing Personal Loan vs Numerical Columns

Observations:

- The median age of both customers who opted for a personal loan and those who did not falls below 30 years.
- Customers who opted for a personal loan tend to have three dependents, while those who did not typically have two dependents.
- The median salary of customers who opted for a personal loan and those who did not is nearly equal, around 80,000.
- The median partner salary of customers who opted for a personal loan and those who did not typically falls below 30,000.
- Both customers who opted for a personal loan and those who did not have allocated a similar amount towards the purchase price.

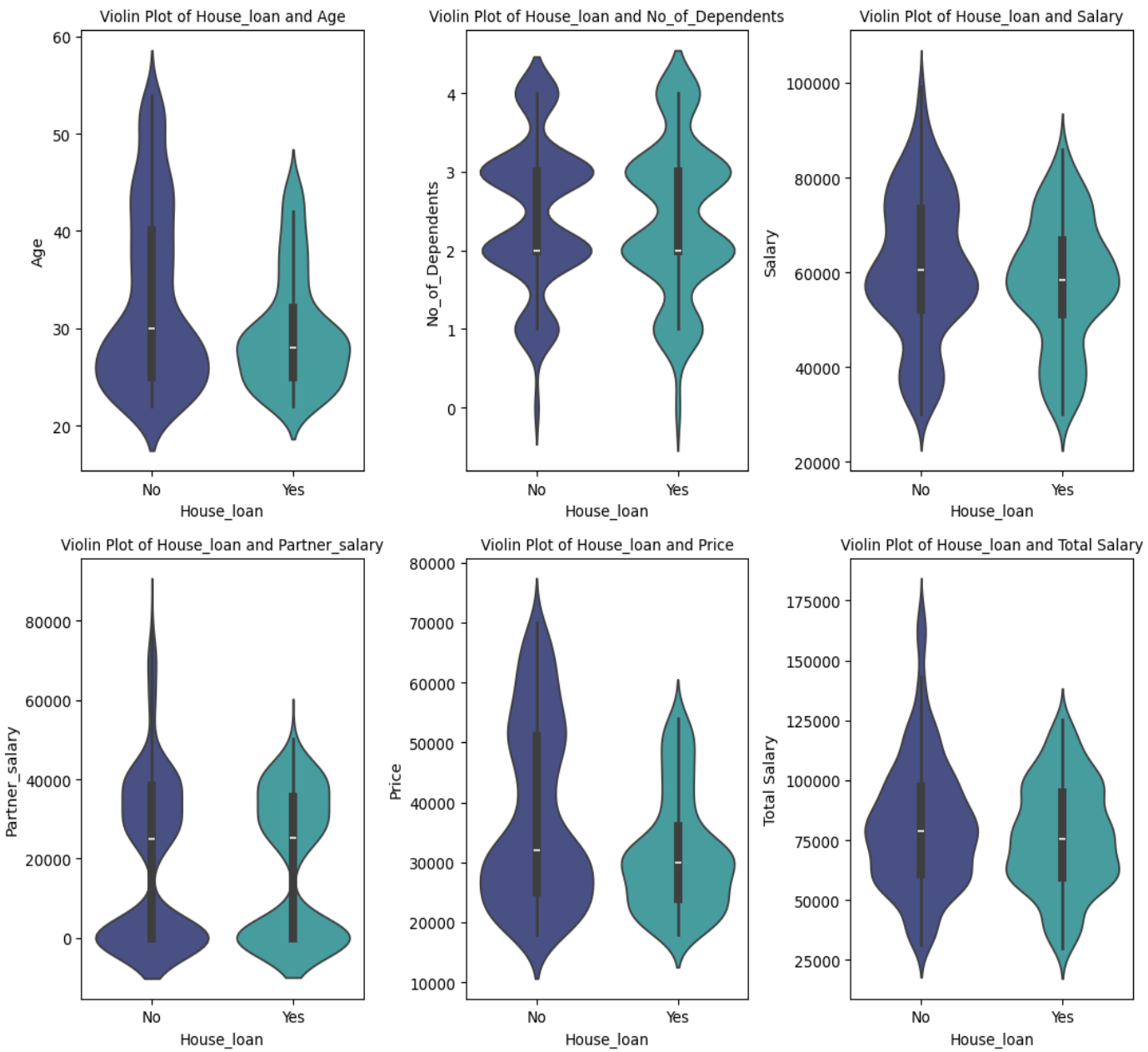

Figure 11 Violinplots showing House Loan vs Numerical Columns

Observations:

- The median age of customers, whether they opted for a house loan or not, is 30 years or younger.
- Customers who opted for a house loan and those who did not usually have two dependents.
- The median salary of customers who opted for a house loan is nearly the same, approximately 60,000.
- The median partner salary for customers who opted for a house loan and those who did not typically falls within the range of 25,000 to 30,000.
- Customers who opted for a house loan and those who did not have allocated a similar amount for the purchase price.

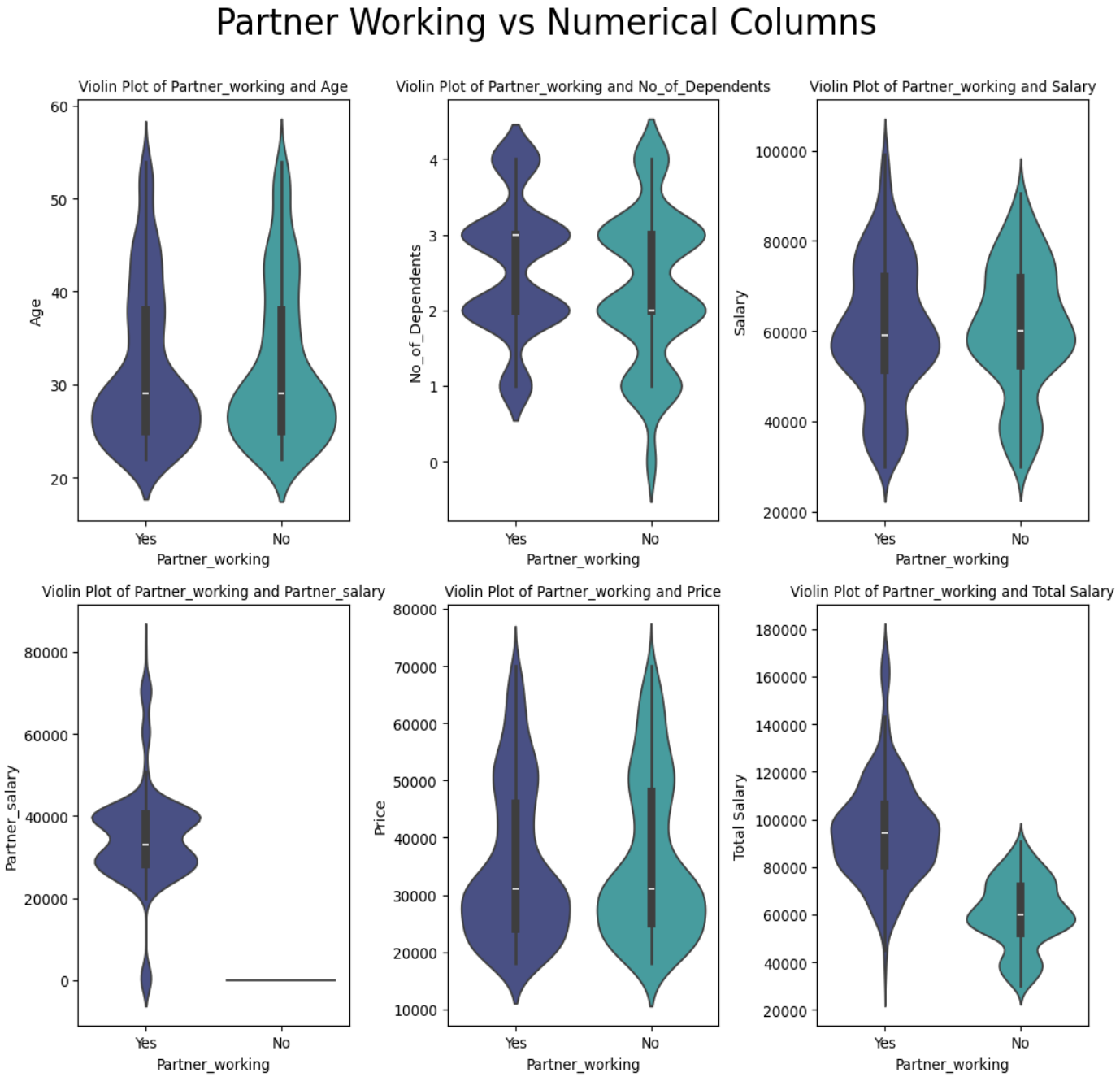

Figure 12 Violinplots showing Partner Working vs Numerical Columns

Observations:

- Customers, whether their partners work or not, typically have a median age below 30 years.
- Customers with working partners tend to have three dependents, while those with non-working partners usually have two dependents.
- The median salary for both customers with working partners and those without is approximately 60,000.
- The median partner salary generally falls within the range of 35,000.
- Customers, regardless of whether their partners work or not, allocate a similar amount towards the purchase price.

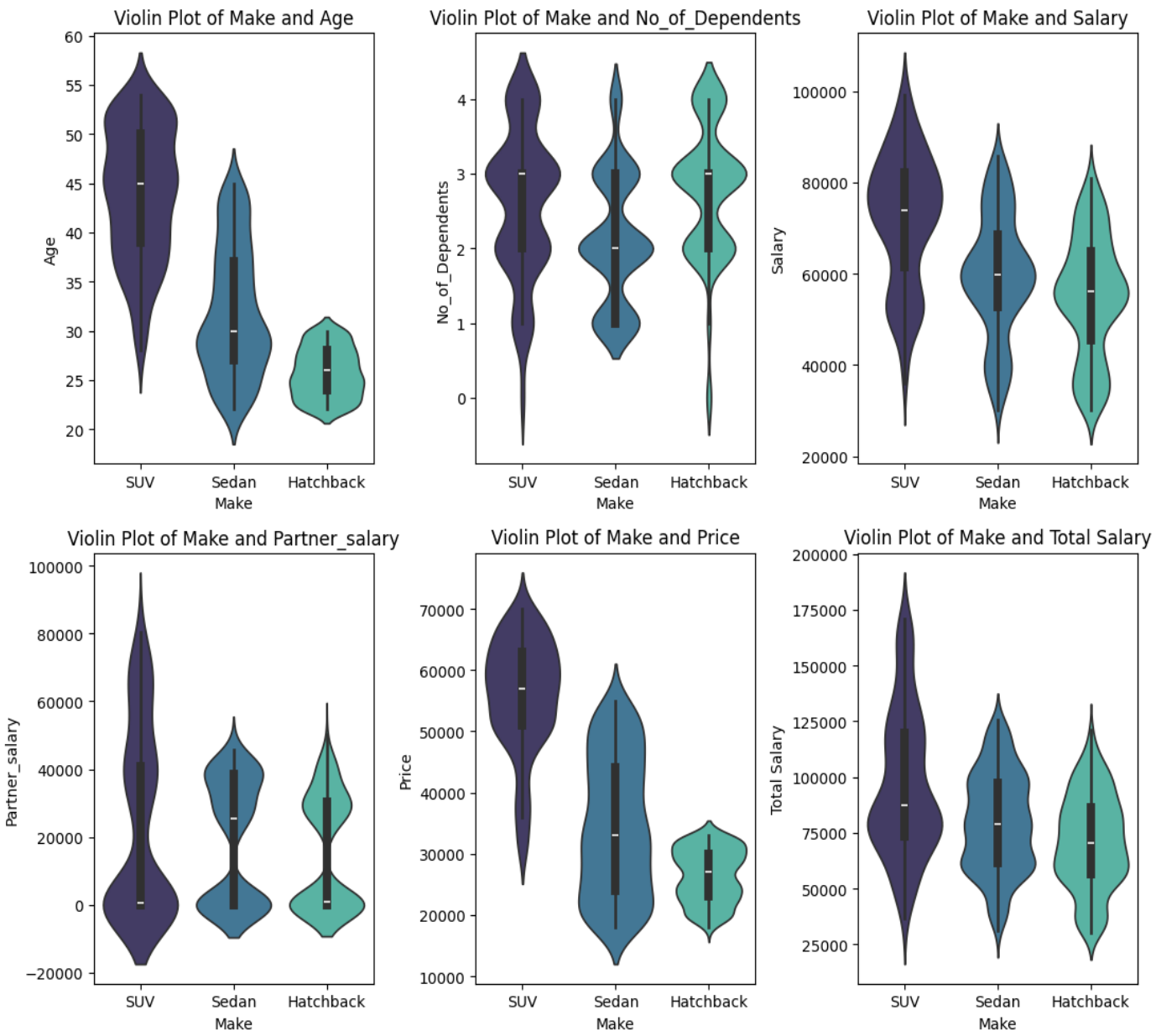

Figure 13 Violinplots showing Make vs Numerical Columns

Observations:

- The median age of customers who purchased SUVs, Sedans, and Hatchbacks is 45, 30, and 26, respectively.
- Customers who bought SUVs and Hatchbacks tend to have three dependents, while those who bought Sedans typically have two dependents.
- The median price for customers who bought SUVs is around 55,000 to 60,000, for Sedans around 30,000 to 35,000, and for Hatchbacks under 30,000.

Problem 1 - Key Questions

Explore the data to answer the following key questions:

1. Do men tend to prefer SUVs more compared to women?
2. What is the likelihood of a salaried person buying a Sedan?
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?
4. How does the amount spent on purchasing automobiles vary by gender?
5. How much money was spent on purchasing automobiles by individuals who took a personal loan?
6. How does having a working partner influence the purchase of higher-priced cars?

Solution of Key Questions:

1. Do men tend to prefer SUVs more compared to women?
    - We plotted a Countplot for the Gender column with Make column as hue.

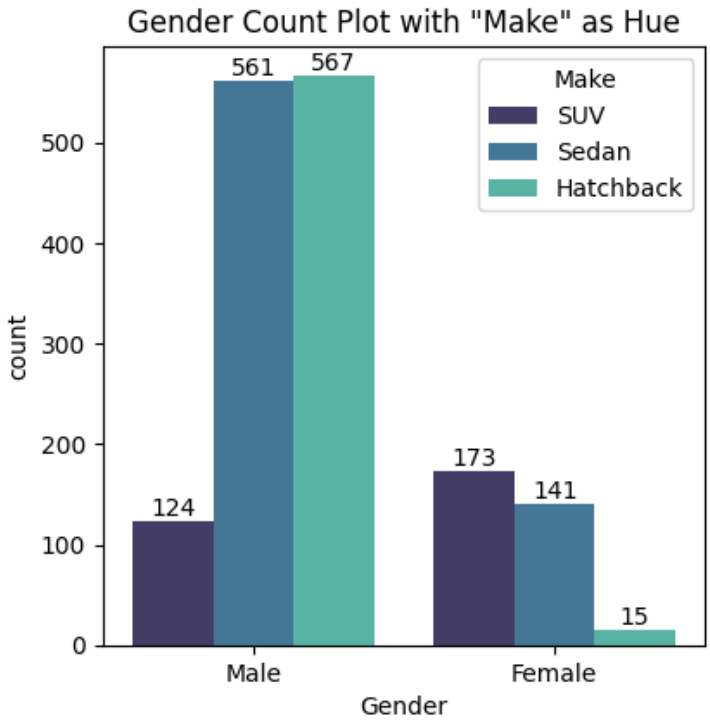

Figure 14 Gender Countplot with Make as Hue

- Based on the countplot above, we observe that **173** female customers prefer SUVs, while **124** male customers prefer SUVs.
- Therefore, we can conclude that men do not prefer SUVs more than women.

2. What is the likelihood of a salaried person buying a Sedan?
    - We generated a countplot for the Profession column with Make as the hue parameter.
    - From Figure 15 Profession Countplot with Make as Hue, it's evident that salaried customers prefer sedans over other makes of cars. Similarly, business customers also show a preference for sedans.

- Based on this analysis, we can conclude that sedan is the most commonly preferred make among salaried customers compared to others.

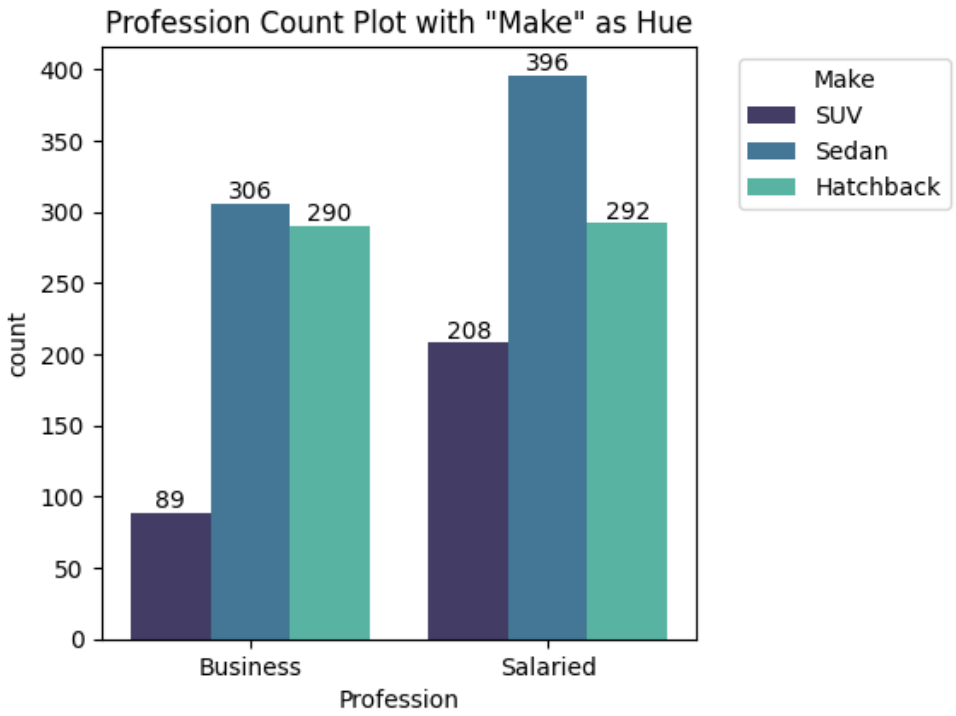

Figure 15 Profession Countplot with Make as Hue

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?
    - We plotted a Catplot, where Gender as column(col) parameter, Profession on x-axis and Make as hue parameter.

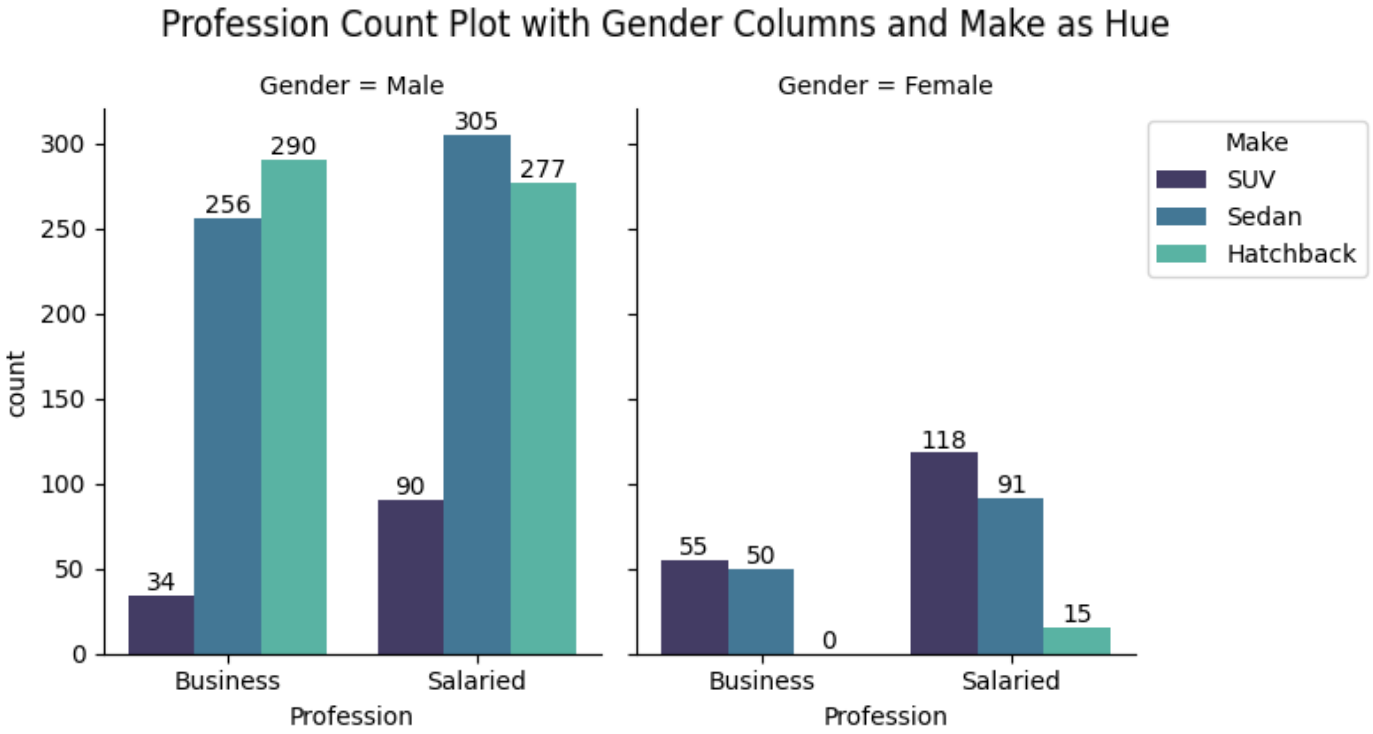

Figure 16 Profession Catplot with Gender Columns and Make as Hue

- We can clearly see from the above catplot that Salaried males mostly prefer Sedan type of cars over SUVs. So, from this we can conclude that Sheldon's claim is **incorrect.**

4. How does the amount spent on purchasing automobiles vary by gender?
   - We grouped the Price column by Gender and Make columns using the groupby() method and calculated the mean, median and sum values.

Gender	Make	Median	Mean	Sum of Price	Total Sum of Price
Female	Hatchback	28000	27466.66	412000	15695000
	SUV	55000	53479.76	9252000	
	Sedan	43000	42773.04	6031000	
Male	Hatchback	26000	26447.97	14996000	40585000
	SUV	59000	59096.77	7328000	
	Sedan	31000	32550.8	18261000	

Table 11 Median, Mean and Sum of Price by Gender

- We then plotted a Box plot of Gender and Price with Make as hue parameter.

Figure 17 Box plot of Gender and Price with Make as Hue

- From Table 11 and Figure 17, it's apparent that the total purchase price associated with males (40,585,000) is roughly 2.5 times greater than that of females (15,695,000). However, this difference is influenced by the respective numbers of male and female customers. Interestingly, the median price of Hatchback and Sedan among female customers surpasses that of males. Conversely, the median price of SUVs among male customers exceeds that of females.

5. How much money was spent on purchasing automobiles by individuals who took a personal loan?
    - We grouped the Price column by personal loan and make columns using the groupby() method and calculated the sum values.

Personal Loan	Make	Sum of Price	Total Sum of Price
No	Hatchback	7765000	28990000
	SUV	10373000	
	Sedan	10852000	
Yes	Hatchback	7643000	27290000
	SUV	6207000	
	Sedan	13440000	

Table 12 Sum of Price by Personal Loan

- From the above Table 12, we can observe that,
    - Buyers without a personal loan tend to spend more on SUVs.
    - Buyers with a personal loan tend to spend more on Sedans.
    - Buyers with personal loans tend to spend less on SUVs and Hatchbacks.

- So, The total amount spent on purchasing automobiles by individuals who took a personal loan is **27290000.**

6. How does having a working partner influence the purchase of higher-priced cars?
    - We have calculated the median, max price and sum of price of Price column gouped by the Partner_working column.

Partner working	Make	Median	Max Price	Max Price Count	Sum of Price	Total Sum of Price
No	Hatchback	26000	33000	281	7397000	25668000
	SUV	58000	70000	144	8089000	
	Sedan	34000	55000	288	10182000	
Yes	Hatchback	28000	33000	301	8011000	30612000
	SUV	57000	70000	153	8491000	
	Sedan	33000	55000	414	14110000	

Table 13 Median, Max and Sum of Price by Partner Working

- We can see that count of max price of cars bought by customer having their partner working is higher than those whose partners are not working
- The total amount spent on purchasing automobiles by customers with their partner working is **30612000** and customers with their partner not working is **25668000**.

Problem 1 - Actionable Insights & Recommendations

- The scatter plot shows that as total salary increases, people tend to spend more on cars. This suggests that those with higher incomes often buy pricier vehicles.
- The number of dependents can influence how customers manage their money.
- Men are the main customers, whether they are married or single.
- Businesswomen don't seem interested in Hatchbacks. The company should take note of this and adjust its strategies accordingly.
- Married men tend to prefer Sedans and Hatchbacks, probably because of financial obligations like loans. Offering discounts on SUVs could appeal to them as they are more spacious and may better suit their family needs.
- Use the insights from salary, gender, and car make to tailor marketing approaches for potential customers.
- Focus marketing efforts on attracting more female and unmarried customers, and boost SUV sales.
- Customers with higher salaries often consider buying more expensive items.
- Stay updated on market trends and competitor activities to stay competitive in the industry.

Problem 2

Context

A bank earns income from interest, transaction fees, and financial advice, with interest charged on customer loans being a major source of profits. GODIGT Bank, a mid-sized private bank, provides a range of banking products and promotes asset products to current customers using various communication channels. However, the bank is experiencing significant credit card attrition, prompting a review of its credit card policy to ensure customers receive the most suitable card for their spending habits and intentions, fostering profitable relationships.

Objective

To find the key variables that have a vital impact on the analysis which will help the company to improve the business.

Problem 2 - Framing Analytics Problem

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

Solution of Framing Analytics Problem:

1. We first conducted initial analysis on the dataset. The analysis involved the following steps,
   - Imported the dataset and stored it in a variable named 'bank_data' using the pandas read_excel method.
   - Explored the basic information about the dataset using the info() method.
   - Examined the statistical summary of the dataset using the describe() method.
   - Verified the shape of the dataset, revealing 8448 rows and 28 columns.
   - Conducted a check for null values, identifying 38 missing values in the 'Transactor_revolver' column. As 'Transactor_revolver' is a categorical column, we decided to impute the missing values with the mode value.
   - Determined the mode value ('T') and replaced the null values with 'T'.
   - Ensured there were no duplicates present in the dataset.
   - Noticed an inconsistency in the 'Occupation_at_source' column where '0' values were present. Since 'Occupation' is a categorical variable, we decided to replace the '0' values with the mode of 'Occupation', which is 'Salaried'.

2. Once the above analysis was done, listed below are some of the questions that can be raised by overlooking the dataset.
   A. What is the relationship between Annual income and CC limit?
   B. What is the occupation of the customers using credit cards?
   C. How much does customers from each occupation spent on last 3 months?
   D. How much credit cards issued with respect to card category and occupation?
   E. Which type of card has been issued the most?

   Solution A:
   - We plotted a scatterplot to find the relation between Annual income and CC limit.

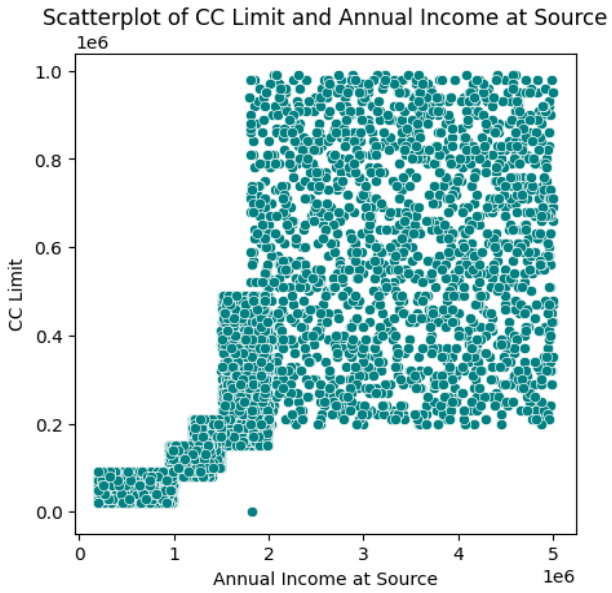

Figure 18 Relationship Between Annual Income and CC Limit

- From the above scatterplot, we can see that annual income and cc limit are positively correlated i.e., the cc limit increases with increase in annual income.

Solution B:

- We plotted a countplot for the Occupation_at_source column to show us the count of customers grouped by their occupation.

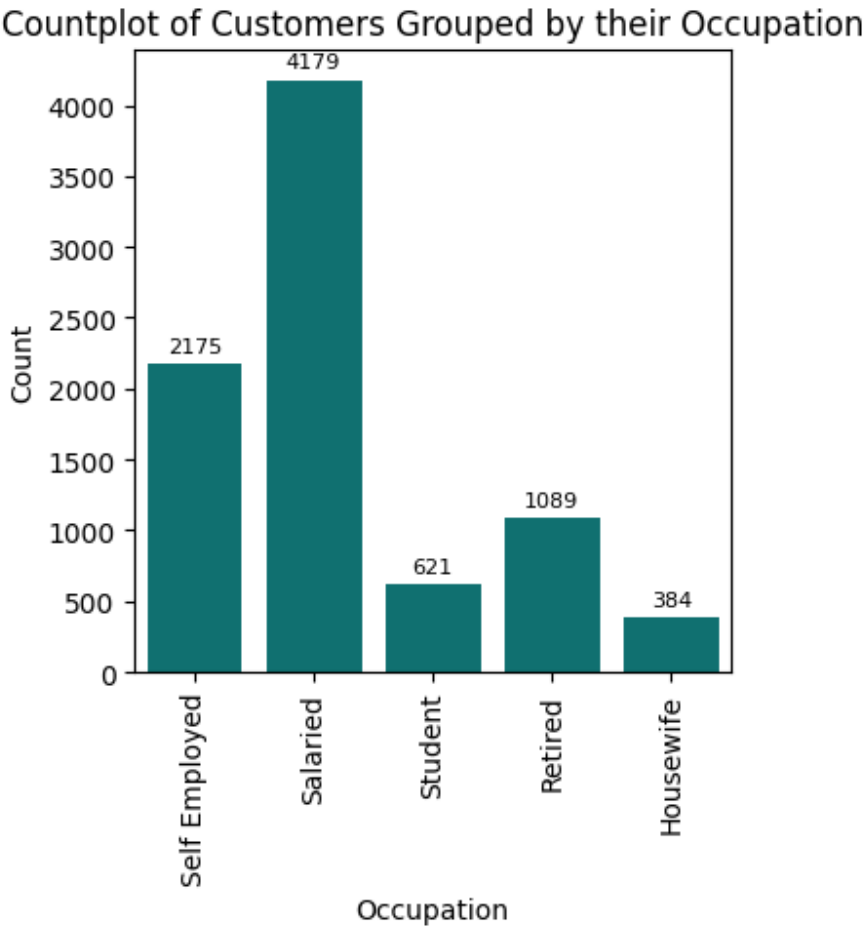

Figure 19 Countplot of Customers Grouped by their Occupation

- From above plot, we can conclude that Salaried individuals are the predominant credit card holders followed by Self Employed individuals. Also we can see that Housewives holds the least amount of credit cards compared to other.

Solution C:

- For this, we first created a temporary dataframe, where we grouped the average spend of last 3 months by their occupation. On that we calculated the sum of values and sorted them in descending order.
- Then we plotted a Barplot with occupation at x-axis and average spend of last 3 months at y-axis and the above temporary dataframe as the data.

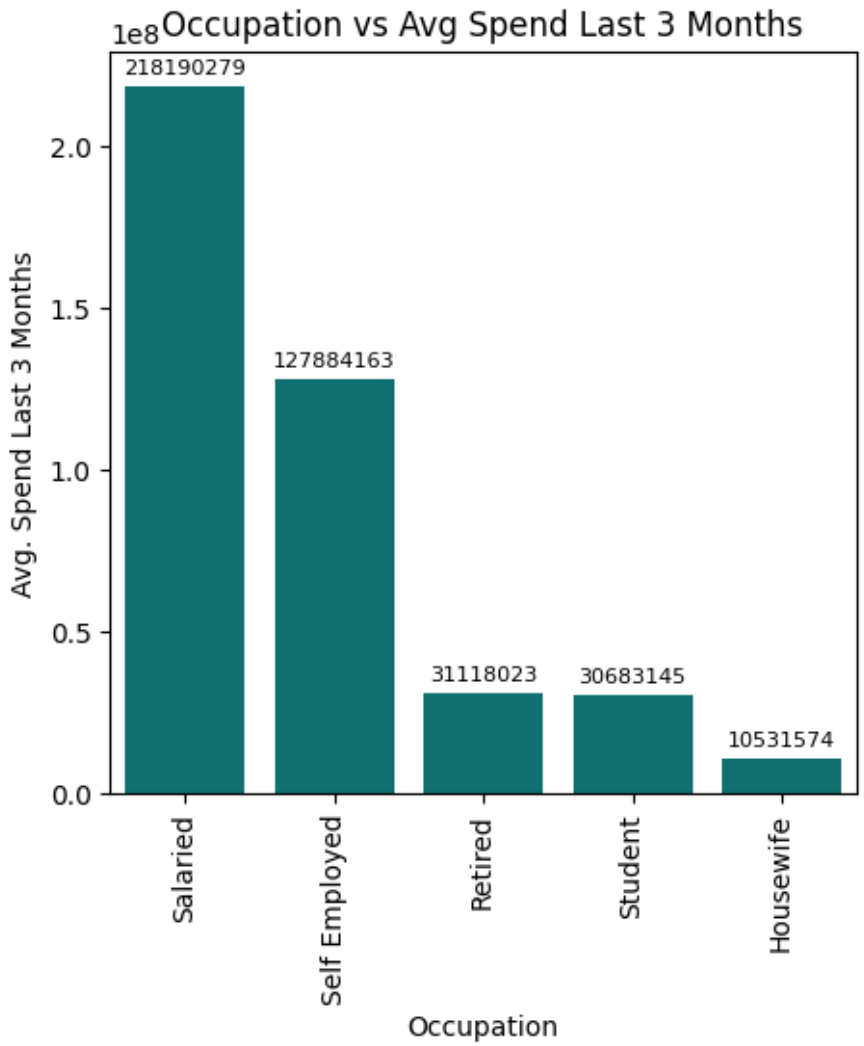

Figure 20 Occupation vs Avg. Spend Last 3 Months

- From the above plot, it is evident that Salaried individuals spent the most in the last 3 months followed by Self Employed Individuals. Similarly Housewives are the least spent among others.

Solution D:

- We created a crosstab for Card type and Occupation and derived the following result.

Occupation	Aura	Cashback	Centurion	Chartered	Edge	Elite	Gold	Indian Oil	Platinum	Prime	Prosperity	Pulse	Rewards	Shoprite	Smartearn
Housewife	38	57	1	0	0	0	6	46	0	0	1	1	106	62	66
Retired	155	158	0	4	4	1	13	147	0	1	2	1	279	152	172
Salaried	319	309	35	478	512	50	73	315	35	58	531	45	752	312	355
Self Employed	87	112	23	371	382	36	40	126	17	46	390	43	262	117	123
Student	53	40	3	70	82	9	13	46	7	7	83	11	103	45	49

Table 14 Crosstab of Occupation and Card Type

- From the above crosstab, we observed that Rewards type of credit card is mostly used among Salaried people and Self Employed individuals prefer Chartered, Edge and Prosperity type of cards.

Solution E:

- We created a countplot for the card type column to check the count of each card type issued.

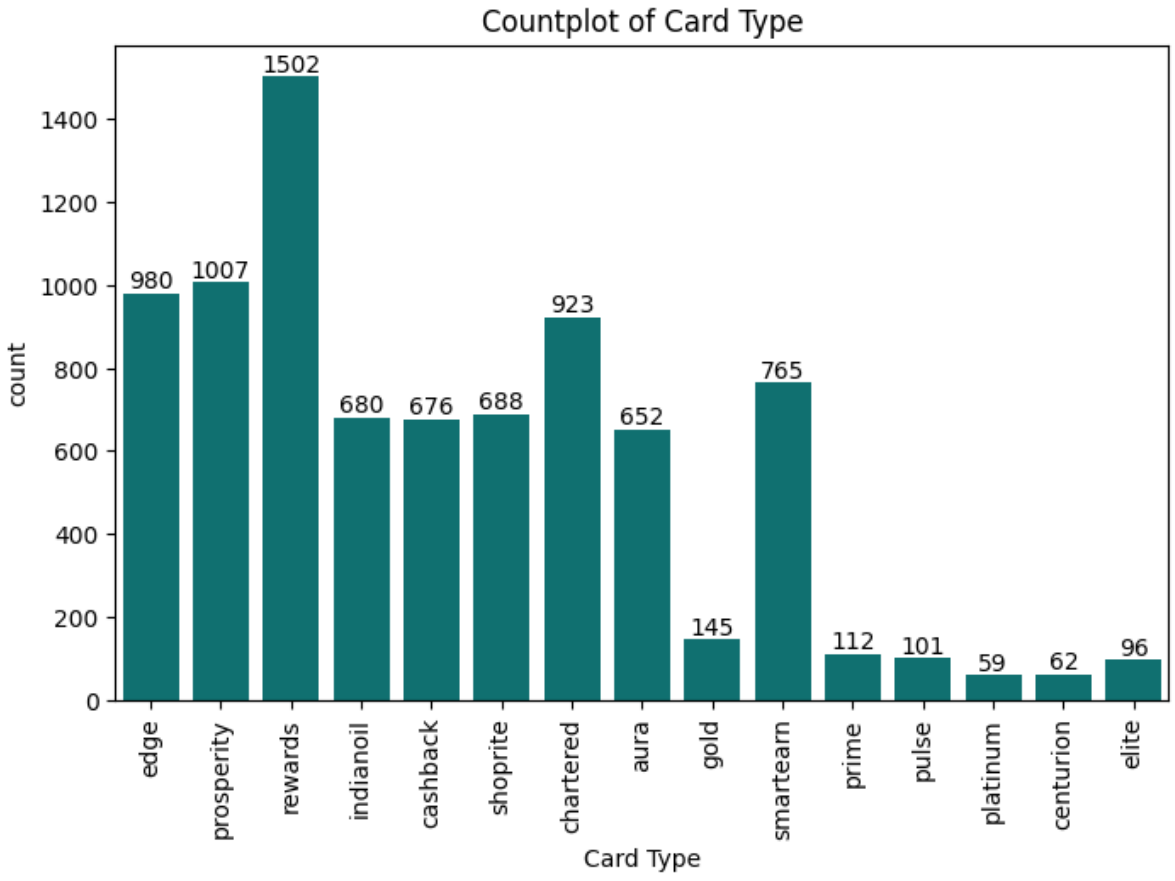

Figure 21 Countplot of Card Type

- It is evident from the above plot that Rewards is the most issued card type followed by prosperity. Platinum is the least issued type of credit card.

3. Further we generated a heatmap table to check how the variables are correlated with each other.

Index	Widget Products	Engagement Products	Annual Income	Bank Vintage	Avg Spend L3M	CC Limit
Widget Products	1	-0.01	0	0.02	0.01	0
Engagement Products	-0.01	1	0.03	-0.02	0.04	0.03
Annual Income	0	0.03	1	0	0.68	0.77
Bank Vintage	0.02	-0.02	0	1	0	0
Avg Spend L3M	0.01	0.04	0.68	0	1	0.53
CC Limit	0	0.03	0.77	0	0.53	1

Table 15 Correlation Table

- From the above correlation table, we can see that there is a positive correlation between annual_income_at_source and cc_limit, annual_income_at_source and avg_spends_l3m and avg_spends_l3m and cc_limit.

And finally, we plotted a scatterplot to find the relationship between annual_income_at_source and avg_spends_l3m and found that with increase in annual income the average spend also increases.

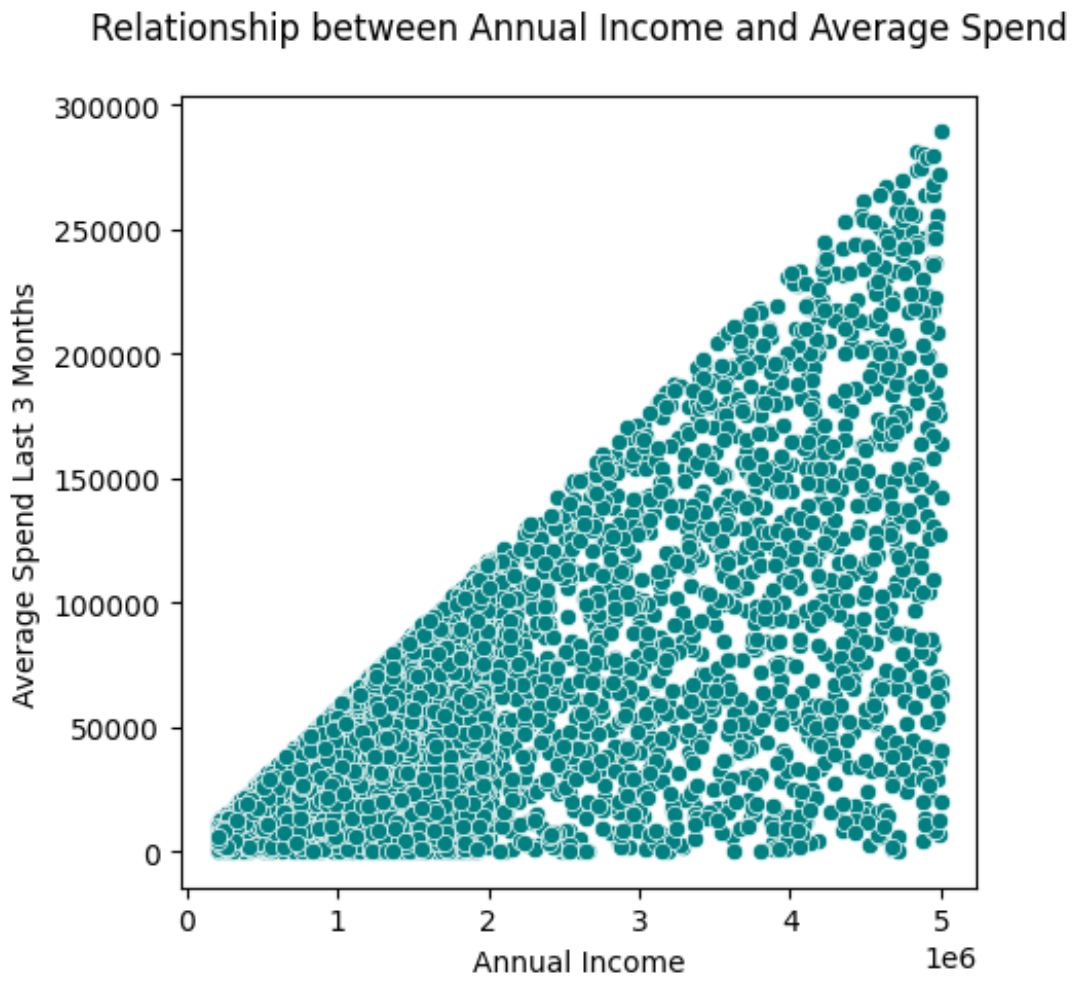

Figure 22 Relationship Between Annual Income and Avg. Spend of Last 3 months

Based on the above analysis, we came up with the below variables which plays a vital role for in-depth analysis,

- annual_income_at_source
- avg_spends_l3m
- cc_limit
- Occupation_at_source
- card_type

The following are five focal points for the bank to address in order to mitigate attrition:

- The bank should prioritize the promotion and enhancement of engagement products. These initiatives can effectively boost customer retention rates.
- Understanding the annual income of customers is crucial. This knowledge enables the customization of financial products and services to better suit individual needs, ultimately reducing attrition.
- Tailoring offerings and benefits to specific occupational groups is essential. Personalized approaches can foster stronger customer loyalty and diminish attrition rates.
- Providing appropriate credit card limits based on customer profiles is pivotal. This approach enhances satisfaction levels and diminishes attrition risk.
- Analyzing the spending behaviours of transactor revolver customers is insightful. Utilizing these insights, the bank can devise targeted retention strategies, thereby mitigating attrition.