Contents

List of Figures

PROBLEM STATEMENT

Austo Motor Company is a leading car manufacturer specializing in SUV, Sedan, and Hatchback models. In its recent board meeting, the members raised concerns about the efficiency of the marketing campaign currently being used. The board decides to rope in an analytics professional to improve the existing campaign.

As an analyst hired by the company, you have been tasked with performing a thorough analysis of the data and coming up with insights to improve the marketing campaign.

A. What is the important technical information about the dataset that a database administrator would be interested in?

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? If yes, perform preliminary treatment of data.

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

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.

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.

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

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

    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.

F. From the given data, comment on the amount spent on purchasing automobiles 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. Give justification along with presenting metrics/charts used for arriving at the conclusions. ***F1) Gender ***F2) Personal_loan

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

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.

EXECUTIVE SUMMARY

The purpose of this report is to analyze the current market trends for Austo Motor Company and provide insights to enhance the efficiency of its marketing campaign. The analysis is based on the available dataset and aims to identify the potential target audience to increase the effectiveness of the marketing campaign and ultimately boost sales.

1. A. What is the important technical information about the dataset that a database administrator would be interested in?

The dataset utilized for this analysis comprises of 1581 rows and 14 columns. Among these columns, there are 6 numerical features, consisting of 5 columns with the int64 data type and 1 column with the float64 data type. Additionally, there are 8 categorical columns of object data type.

Figure 1 : Number of Rows and Number of Columns

```
No of Rows : 1581
No of Columns : 14
```

Figure 2 : Data Information

```
<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
```

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? If yes, perform preliminary treatment of data.

Head of the Data shows there is a spelling error in one of the fields of Gender column – "Femal", which needs to be corrected.

Figure 3 : Head of the Data

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

The tail of the data shows a few missing values in the Partner_salary column, which will be corrected based on the values from the Salary and Total_salary columns.

Figure 4 : Tail of the Data

	Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_sala
1576	22	Male	Salaried	Single	Graduate	2	No	Yes	No	33300	0.0	333(
1577	22	Male	Business	Married	Graduate	4	No	No	No	32000	NaN	320(
1578	22	Male	Business	Single	Graduate	2	No	Yes	No	32900	0.0	329(
1579	22	Male	Business	Married	Graduate	3	Yes	Yes	No	32200	NaN	322(
1580	22	Male	Salaried	Married	Graduate	4	No	No	No	31600	0.0	316(

Figure 2: Data Information image revealed that there are few missing values in Gender and Partner_salary columns. Gender column has 53 missing values and Partner_salary has 106 missing values.

Figure 5 : Number of missing values

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

Below images shows the Statistical Information for Numerical and Categorical columns, with which we can understand the general summary of the data.

Figure 6 : Information on Numerical Variables

	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	1475.0	20225.559322	19573.149277	0.0	0.0	25600.0	38300.0	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

Figure 7 : Information on Categorical Variables

	Gender	Profession	Marital_status	Education	Personal_loan	House_loan	Partner_working	Make
count	1528	1581	1581	1581	1581	1581	1581	1581
unique	4	2	2	2	2	2	2	3
top	Male	Salaried	Married	Post Graduate	Yes	No	Yes	Sedan
freq	1199	896	1443	985	792	1054	868	702

As per the below image, it becomes evident that the spelling error is present only in the Gender column, where "Femal" and "Femle" are incorrectly recorded. To ensure data accuracy and consistency, these entries must be corrected to "Female" in order to maintain uniformity across the dataset.

Figure 8 : Data Quality Check of Categorical Columns

```
Male      1199
Female     327
Femal        1
Femle        1
Name: Gender, dtype: int64
Salaried    896
Business    685
Name: Profession, dtype: int64
Married    1443
Single      138
Name: Marital_status, dtype: int64
Post Graduate    985
Graduate         596
Name: Education, dtype: int64
```

```
Yes    792
No     789
Name: Personal_loan, dtype: int64
No     1054
Yes     527
Name: House_loan, dtype: int64
Yes    868
No     713
Name: Partner_working, dtype: int64
Sedan        702
Hatchback    582
SUV          297
Name: Make, dtype: int64
```

HANDLING MISSING VALUES

1. As per Figure 5: Number of missing values, “Gender” column has 53 missing values which constitute to 3% of the total entries in the column. For Categorical fields, it is recommended to impute the missing values with the mode of the column.
   Since the percentage of missing value is very low in “Gender” column, we can impute the missing values as “Male” which is the mode of Gender column as per the below image.

Figure 9 : Mode of Gender

```
Male      1199
Female     329
Name: Gender, dtype: int64
```

2. As per Figure 5: Number of missing values, “Partner_salary” column has 106 missing values which constitute to ~7% of the total entries in the column. Missing values are imputed using the following calculation -

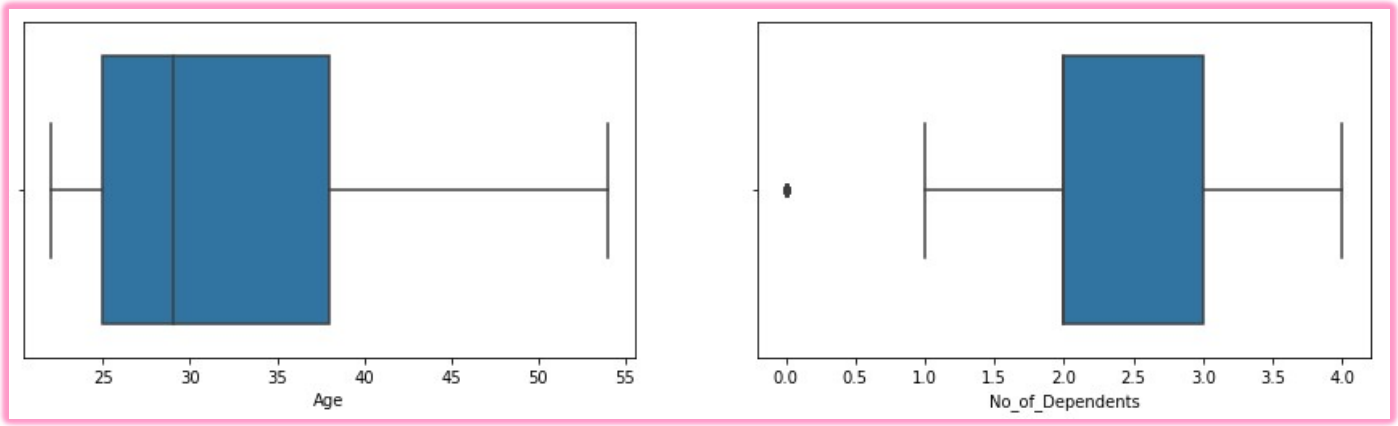
   Partner Salary = Total Salary – Salary

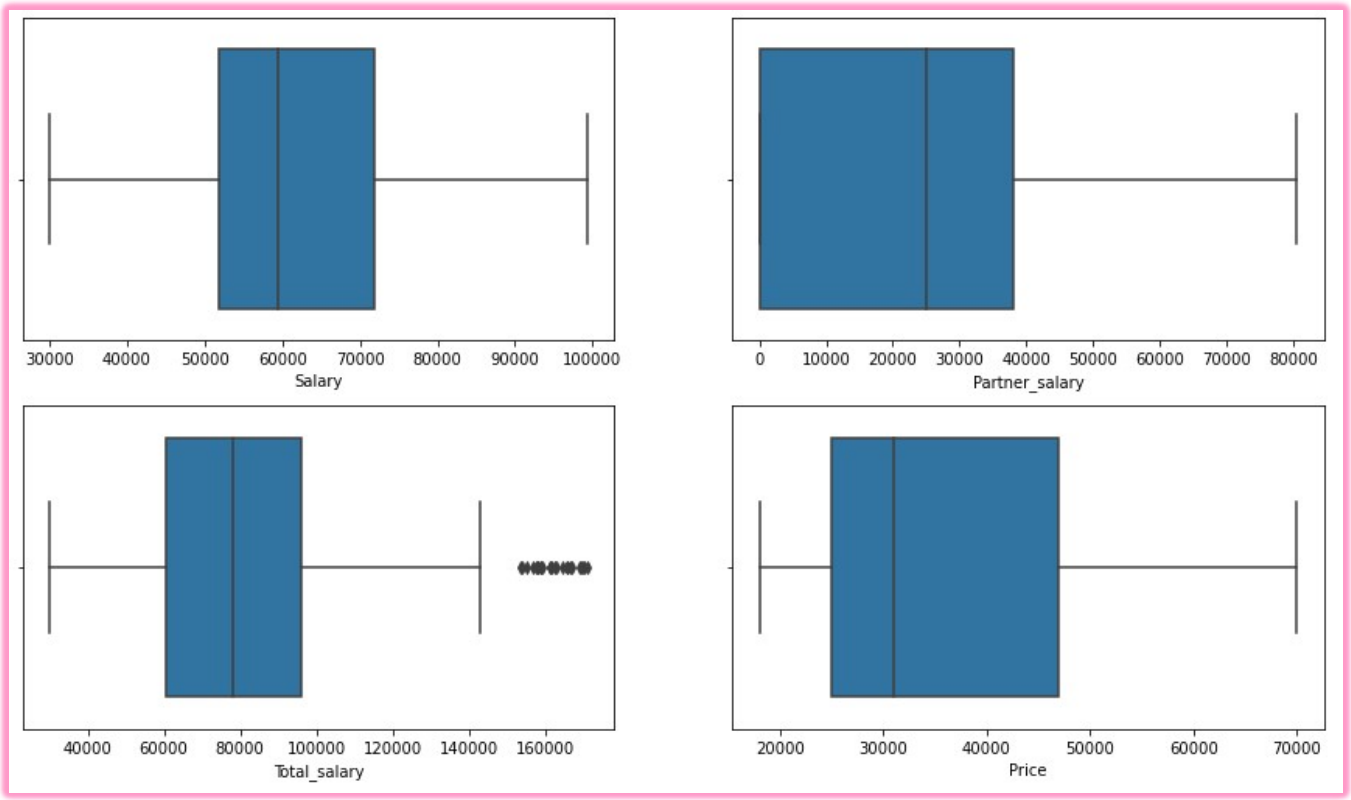
After correcting the spelling errors for "Female" and imputing missing values in the Gender and Partner_salary columns using appropriate techniques, an examination of the dataset reveals the absence of any duplicated rows.

Figure 10 : Duplicate rows check

Age	Gender	Profession	Marital_status	Education	No_of_Dependents	Personal_loan	House_loan	Partner_working	Salary	Partner_salary	Total_salary

Figure 11 : Box-plots for Numerical columns

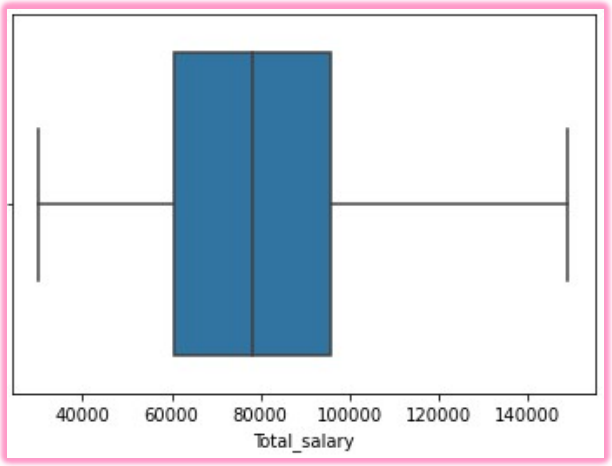




1. Based on the above box plots, outliers are observed in both the "No of Dependents" and "Total Salary" columns. It is worth noting that the "No of Dependents" column does not have a wide range and can be considered categorical, dividing the group into 5 parts. Altering the data in this column may lead to a misleading analysis, as customers with 0 dependents could have specific car preferences. Further analysis will be conducted to explore these preferences in detail.
2. There are 27 Outliers present in the Total Salary column. We utilized the Inter Quartile Range (IQR) method to identify and eliminate the outlier. By performing the calculation below, we determined the upper range and replaced all values exceeding it with $149,000, which represents the upper boundary.

Upper range = Q3 + (1.5 * IQR)

Figure 12 : Total Salary – Post Outlier Treatment

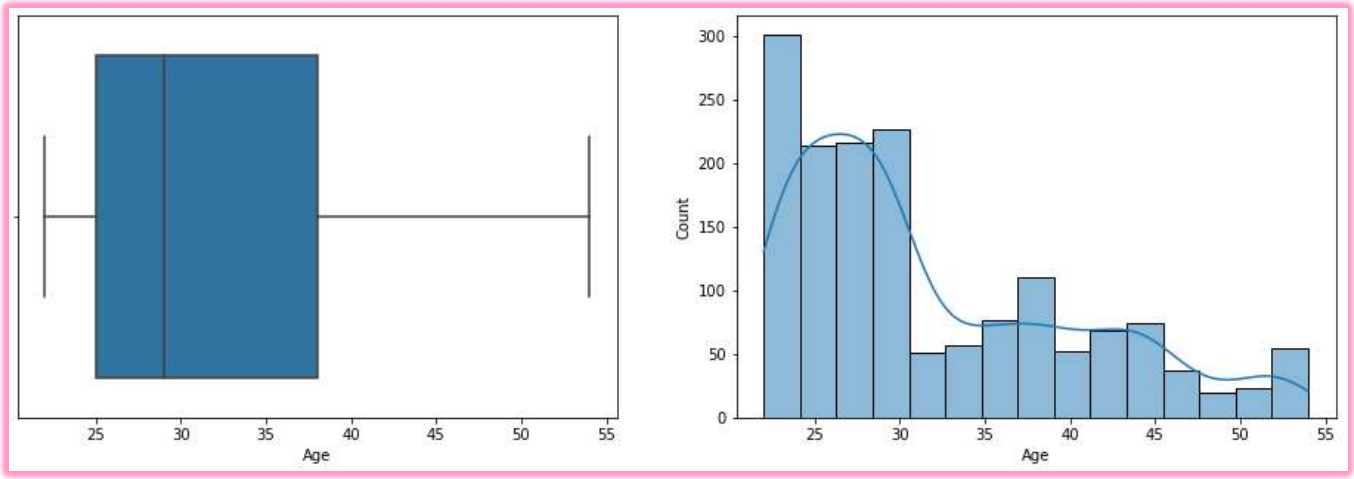


C. Explore all the features of the data separately by using appropriate visualizations and draw insights that can be utilized by the business.

Univariate Analysis for Numerical Columns

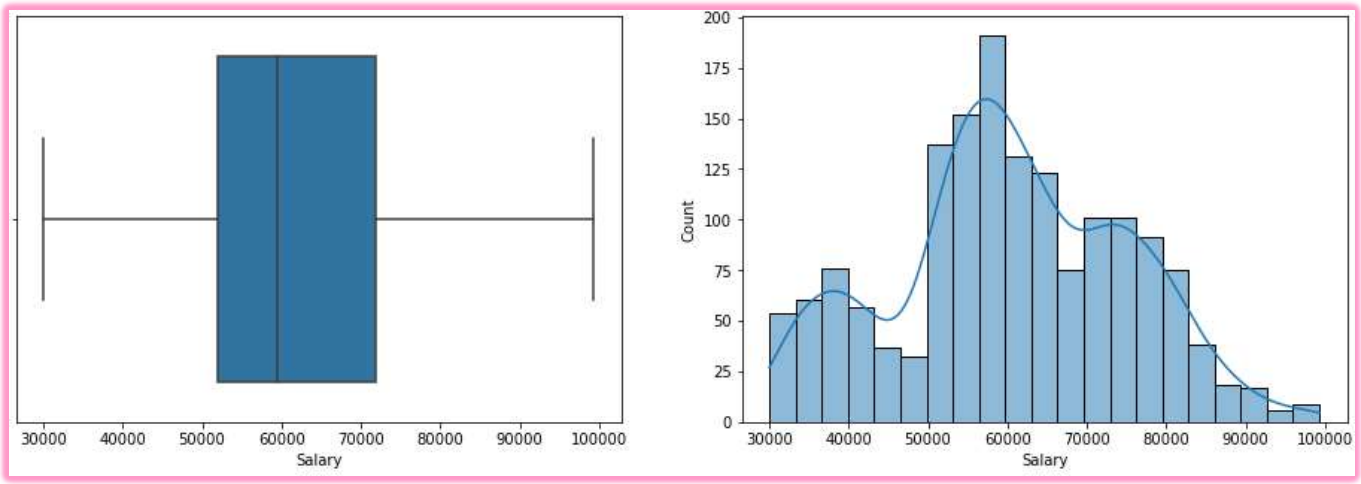
In order to conduct a thorough univariate analysis of the numerical variable, we will utilize two effective visualization techniques: boxplots and histograms. By employing these methods, we can gain valuable insights into the distribution characteristics of the variable, enabling a comprehensive understanding of its underlying patterns and trends.

Figure 13 : Visualization for Age



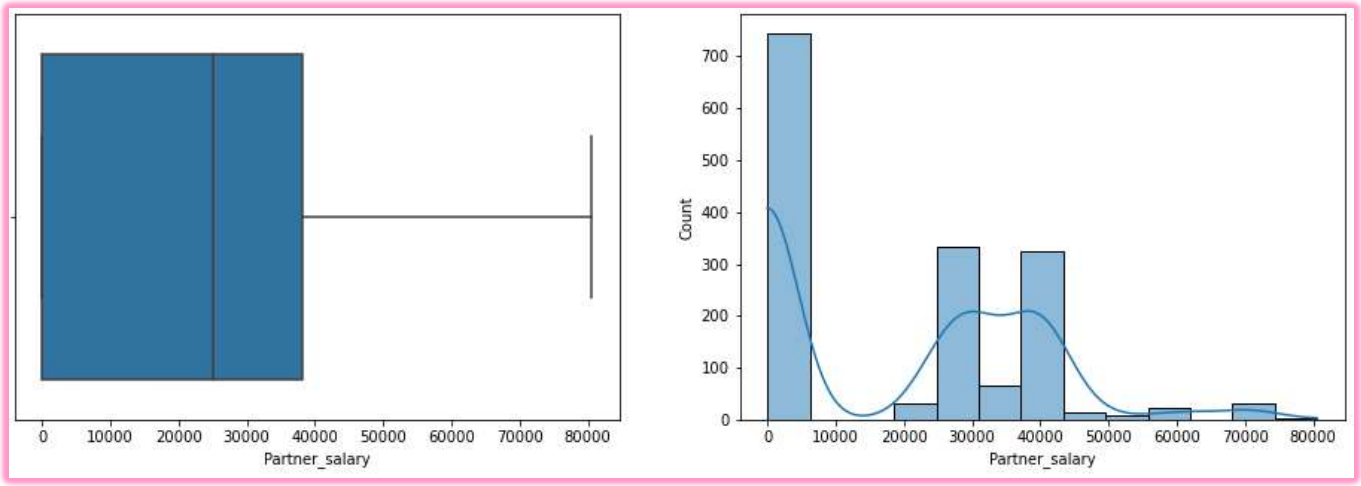
Insight 1: Based on the above plots for the Age column, we can witness all customers fall within the age range of 22 to 54, suggesting that they are all from working class. Moreover, we can also understand the buying pattern with respect to age group, as it becomes evident that individuals prefer to buy cars at a younger age of between 20-30 when compared to other age groups.

Figure 14 : Visualization for Salary



Insight 2: Based on the above plots for the Salary column, we can witness that majority of car purchase is happening amongst individuals earning between $50,000 to $70,000.

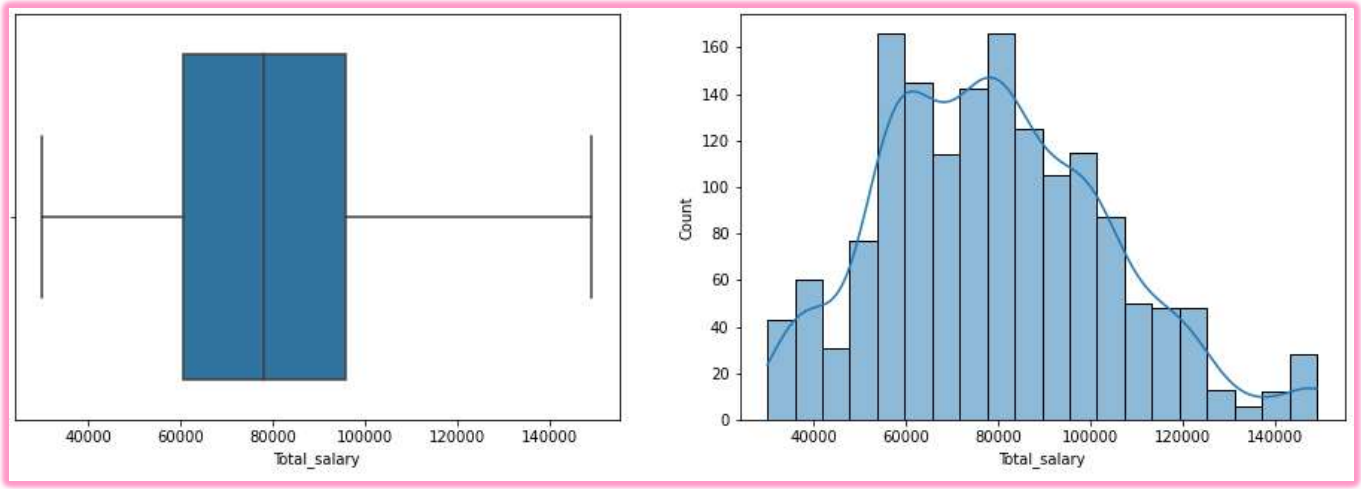
Figure 15 : Visualization for Partner Salary



Analyzing the provided plots for the Partner Salary column, it is evident that the majority of partners fall into the category of non-working individuals, as indicated by the starting range of 0. Additionally, we observe a concentration of partner salaries within the range of $25,000 to $40,000 for those who are earning.

Insight 3: Most Car purchases are done by people having non-earning partners

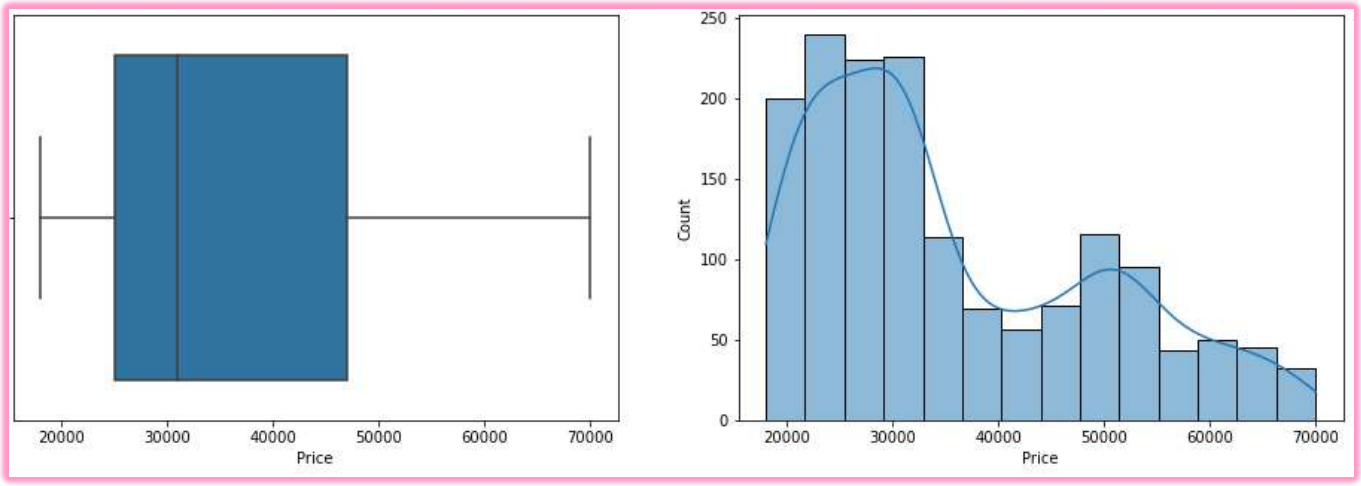
Figure 16 : Visualization for Total Salary



The plots provided showcase the household salary, which represents the combined sum of both customers and their partners salaries. After removing outliers using the IQR method and replacing them with the upper range value, we can observe from the aforementioned plots that the majority of household salaries fall within the range of $60,000 to $100,000.

Insight 4: Most Car purchases has happened with people having $50k to $ 100k

Figure 17 : Visualization for Price

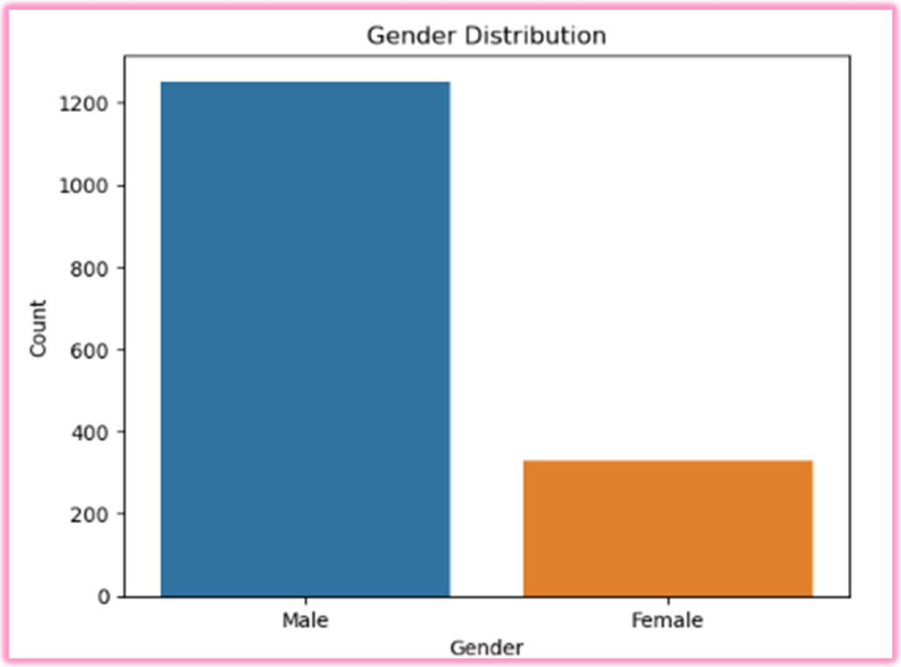


Based on the price distribution depicted above for cars manufactured by Austo Motor Company, it is evident that the prices of all cars fall within the range of $18,000 to $70,000.

Insight 5: Furthermore, a significant proportion of the cars sold are priced in the range of $20,000 to $30,000, suggesting that customers exhibit a sense of price consciousness in their purchasing decisions.

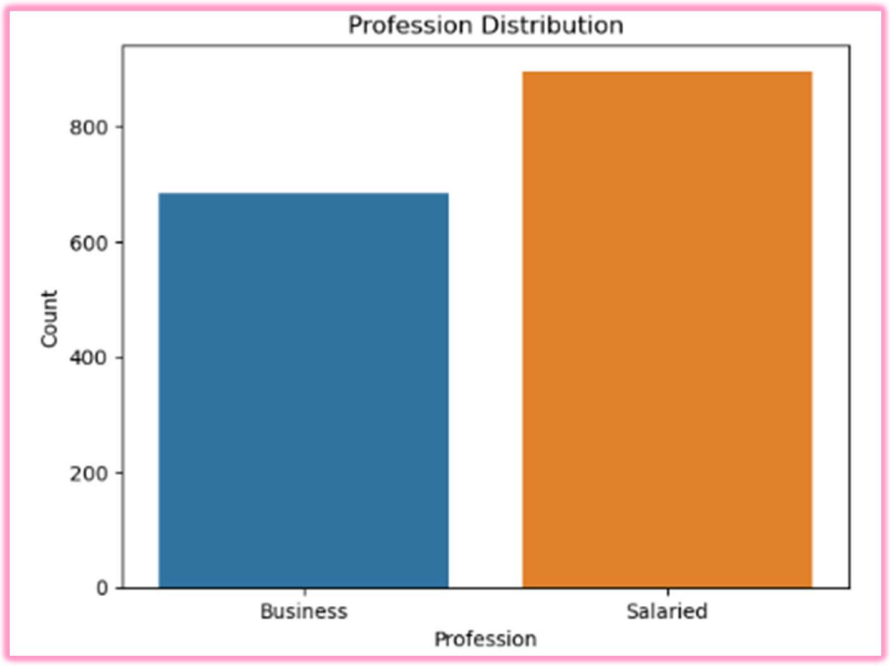
Univariate Analysis for Categorical Columns

Figure 18 : Visualization for Gender



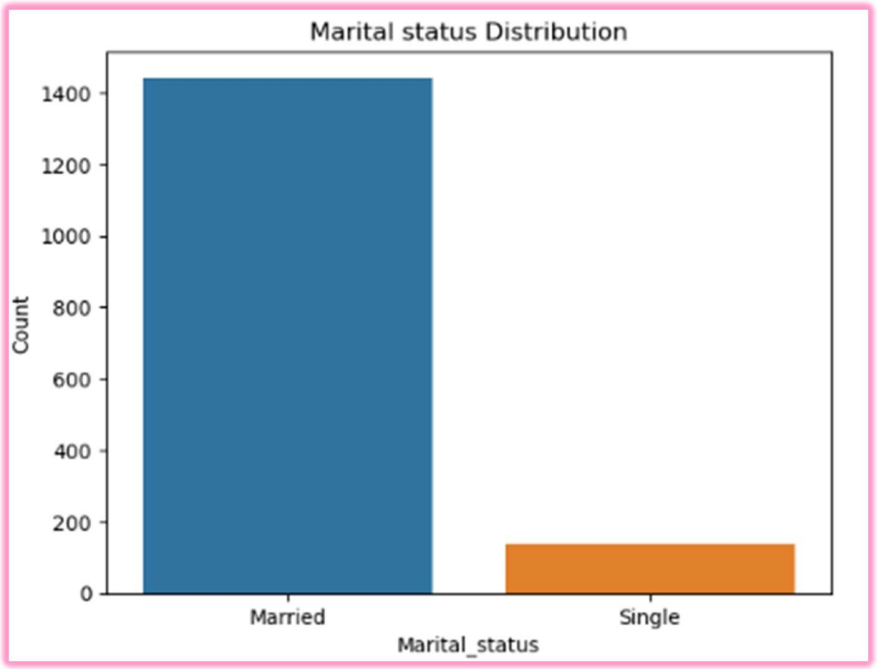
Insight 6: The Count Plot for Gender reveals a gender bias in the data, with males showing a higher preference for purchasing cars compared to females.

Figure 19 : Visualization for Profession



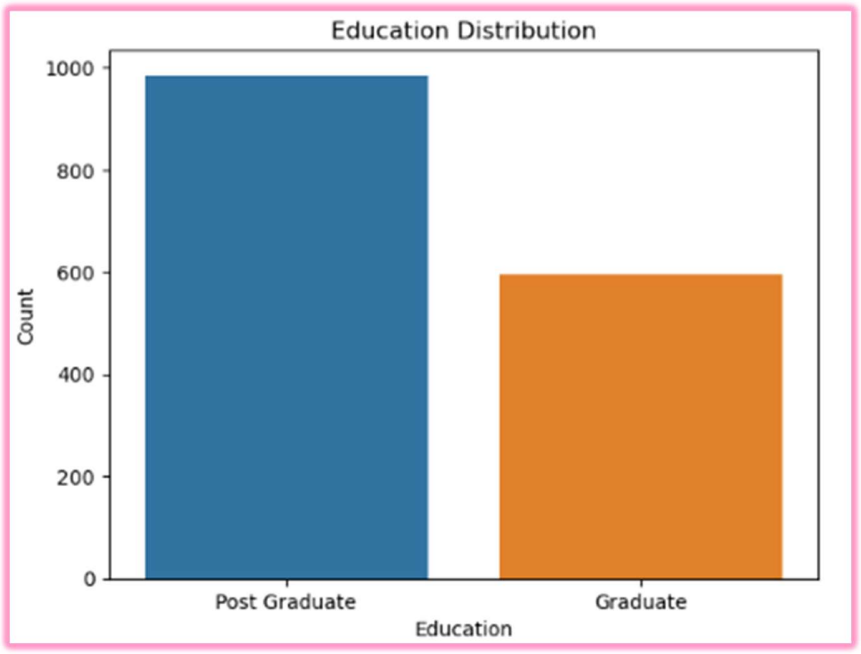
Insight 7: The Count Plot for the "Profession" category indicates a slightly higher number of car purchases among Salaried individuals compared to Business professionals.

Figure 20 : Visualization for Marital status



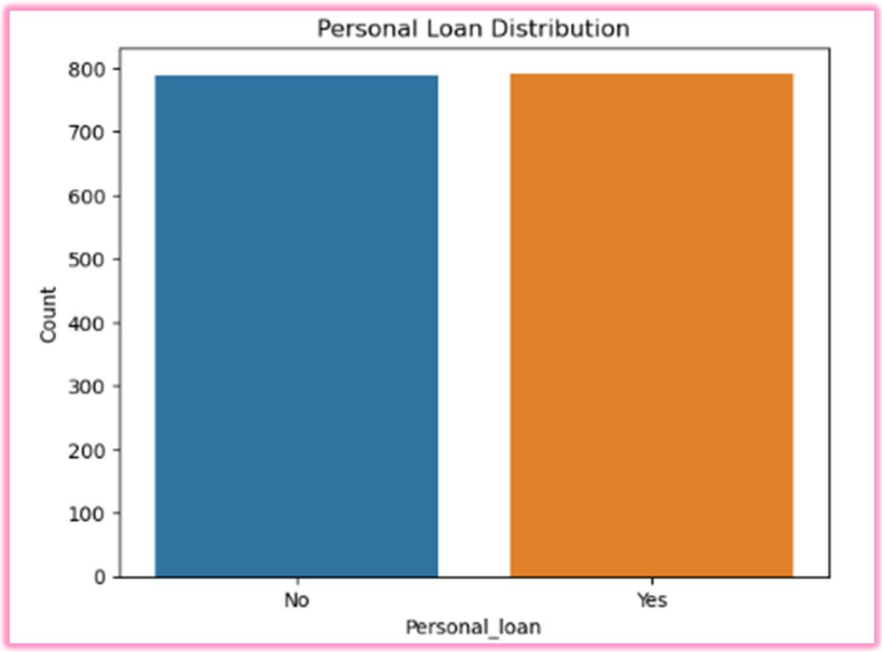
Insight 8: Analyzing the marital status distribution, it is evident that a greater number of married people are likely to buy cars than single people. Thus, the marketing strategy can be focused more on married people for better sales.

Figure 21 : Visualization for Education



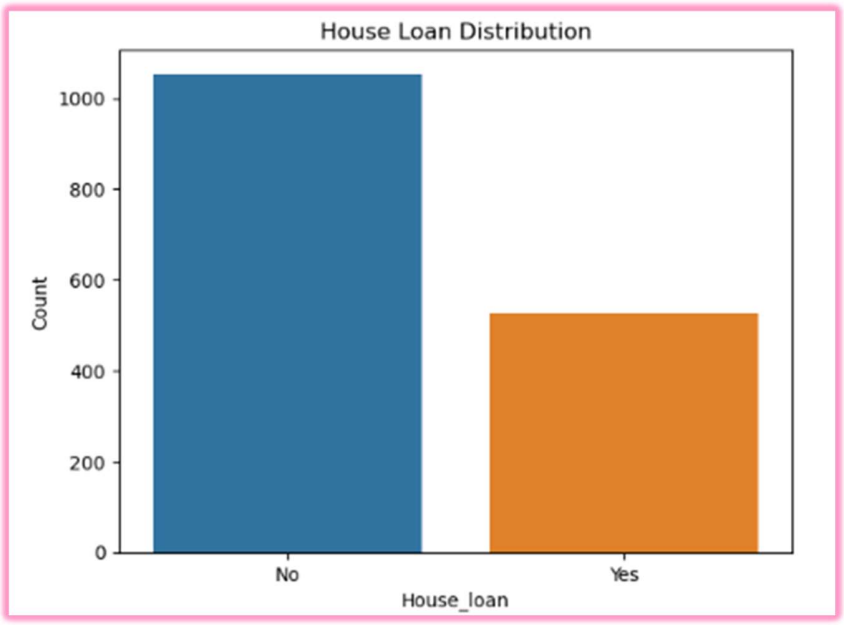
Insight 9: As per the above Count plot, in the dataset majority of customers purchased car is possessing a postgraduate education level.

Figure 22 : Visualization for Personal loan



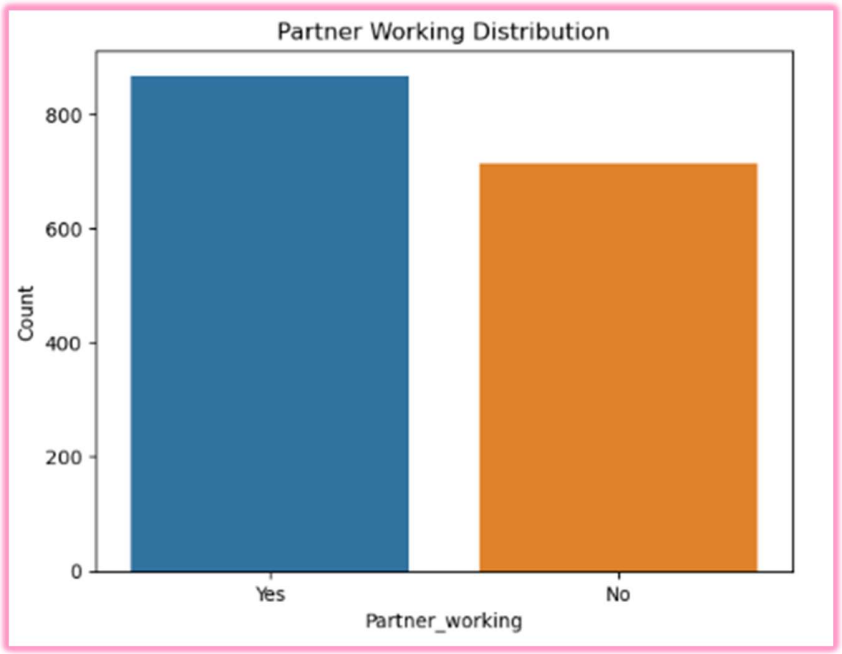
Insight 10: The Count Plot suggests that the presence of a personal loan does not significantly impact the decision to purchase a car, as the counts of customers with and without personal loans are similar. Thus, this variable need not be taken into account for marketing strategies.

Figure 23 : Visualization for house loan



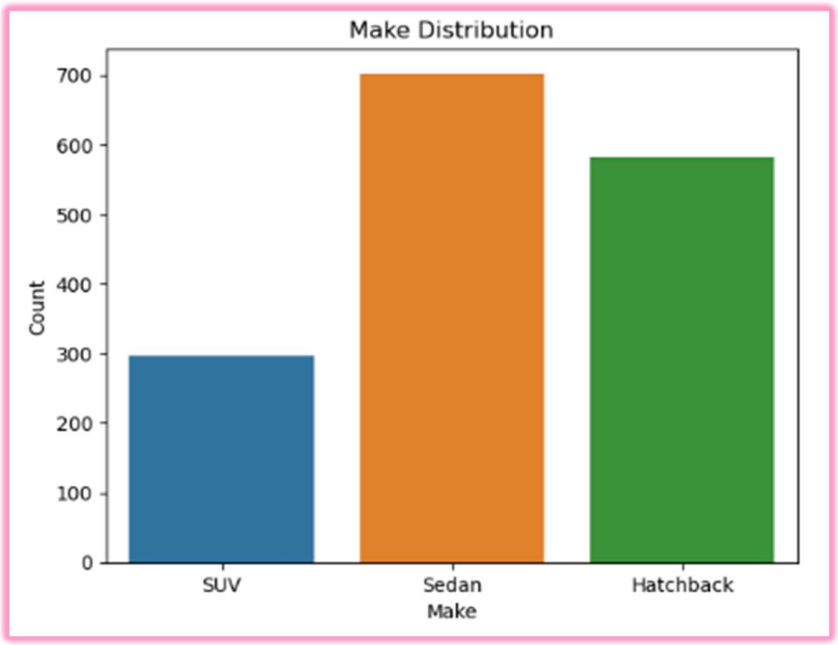
Insight 11: The number of customers without a house loan is approximately twice the number of customers with a house loan, implying that prior commitments such as mortgages could influence car buying decisions.

Figure 24 : Visualization for partner working



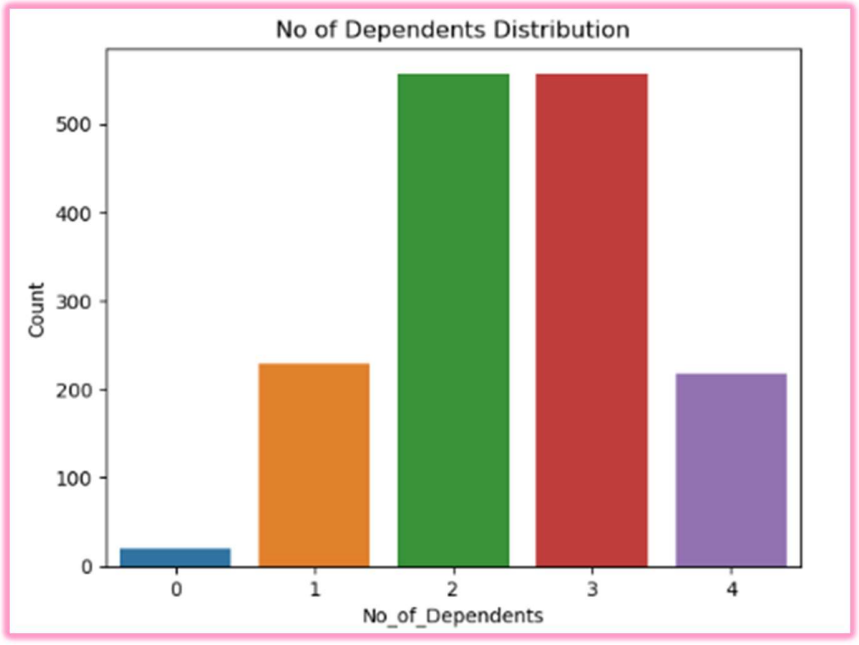
Insight 12: There is a higher count of customers with working partners compared to customers with partners who do not work. In the dataset approx. 55% of the customers have working partners. Having a working partner do influence the buying decision of the car

Figure 25 : Visualization for Make



Insight 13: From the above plot we can infer that among the three car models, Sedan emerges as the top-selling car, accounting for 44% of the total sales, indicating its popularity.

Figure 26 : Visualization for No of Dependents



Insight 14: As previously discussed, the number of dependents is better suited as a categorical variable. The Count Plot analysis reveals that a majority of customers (approximately 70%) have 2 or 3 dependents

People with 2 or 3 dependents are more prone to buy car than lesser number 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.

To gain a comprehensive understanding of the relationships between numerical variables, we will conduct an analysis of their correlation. This will be accomplished through the utilization of visual tools such as a pairplot and/or heatmap. These techniques will enable us to visualize and assess the strength and direction of the correlations, identifying any significant patterns or associations among the variables.

Figure 27 : Pair plot

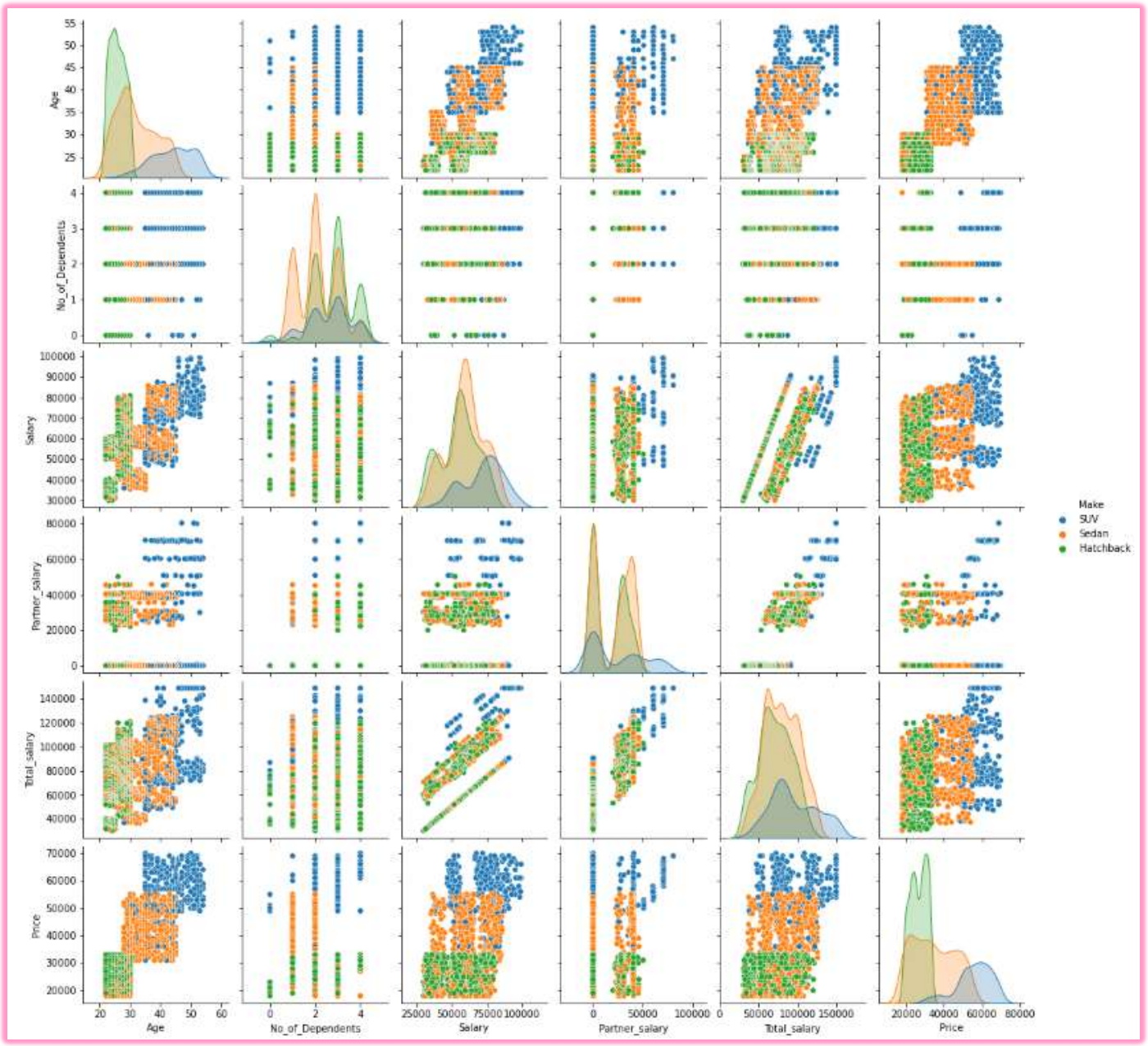
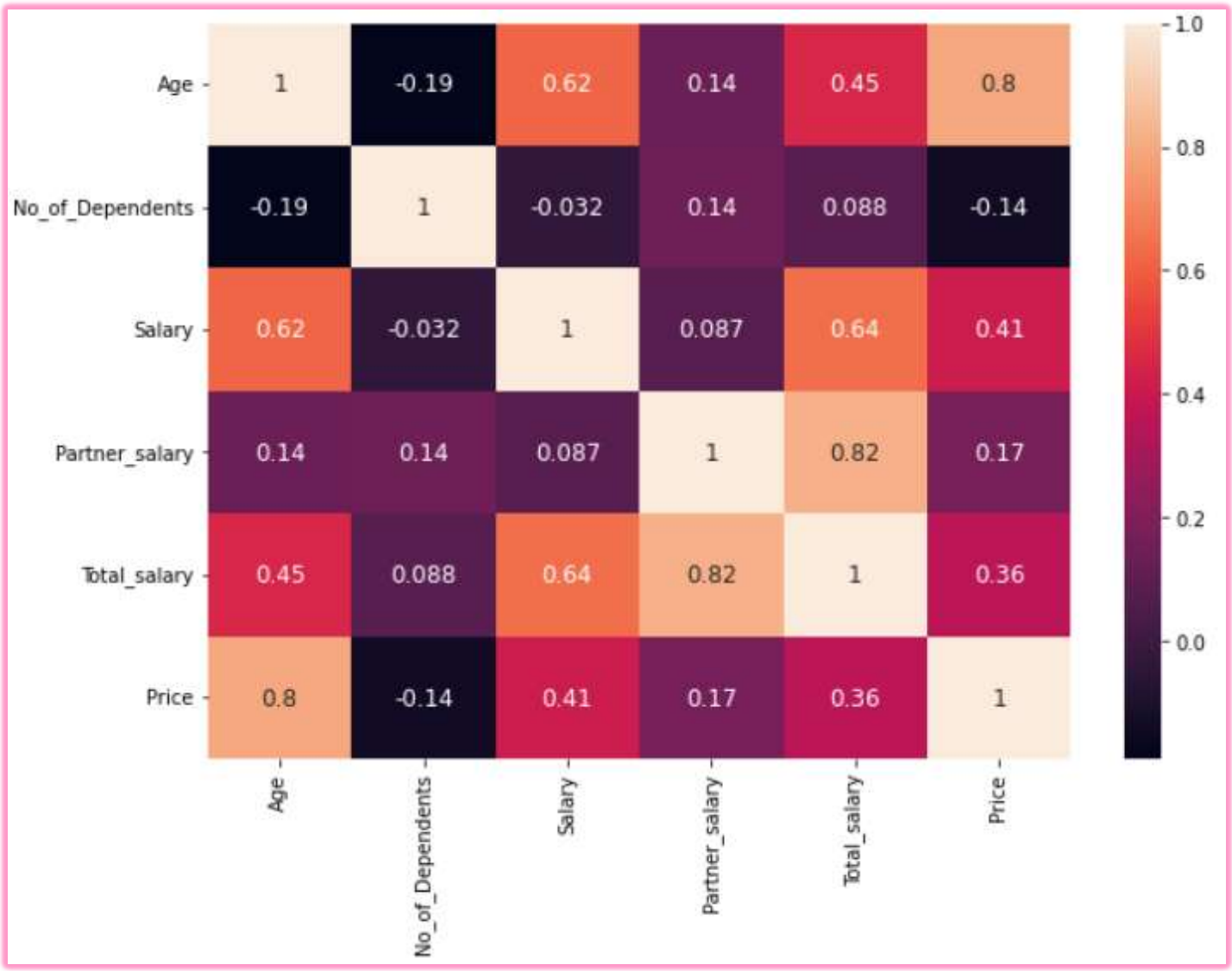


Figure 28 : Heatmap

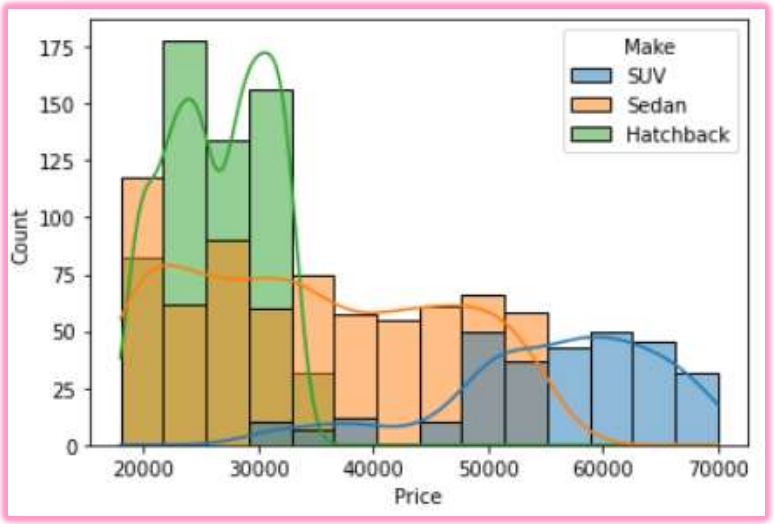


Based on the analysis of the Pair plot and Heatmap, it is evident that most of the variables in the dataset exhibit weak or no correlation with each other. However, there are a few meaningful correlations worth noting:

- Age and Price demonstrate a high correlation of 0.80, indicating a strong relationship between age and the price of the product.
- Age and Salary exhibit a medium correlation of 0.62, suggesting a moderate association between a person's age and their salary.
- Salary and Price, as well as Total salary and Price, show weak correlations of 0.41 and 0.36, respectively. This indicates a slight relationship between salary-related factors and the price of the product

These observations allow us to make the following conclusions: With increasing age, customers tend to prefer more expensive cars. Additionally, there is a tendency for the salary to increase with age, but this increase in salary does not necessarily translate into the purchase of more expensive cars.

Figure 29 : Price vs Make

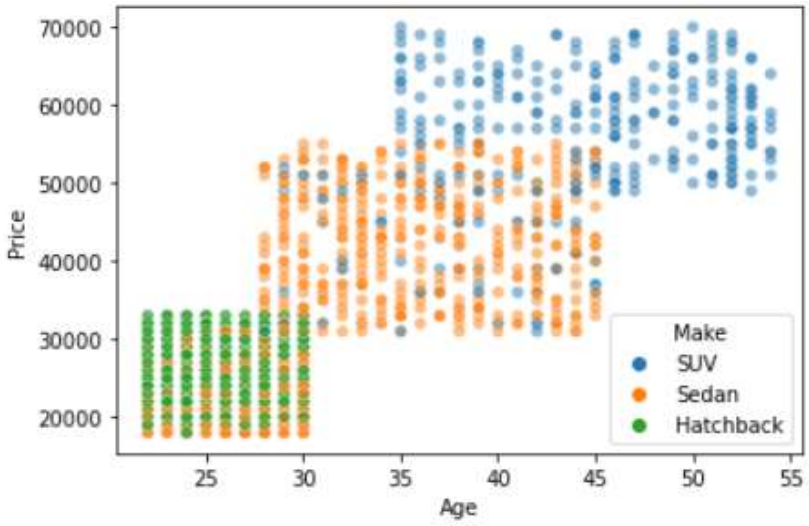


Upon analyzing the Price distribution plot, we can draw the following conclusions:

- Hatchback cars are primarily priced between $18K and $35K. It is reasonable to categorize them as affordable options, especially suitable for price-conscious customers.
- Sedan cars exhibit a wider price range, typically falling between $18K and $55K. This model category can be considered moderate in terms of pricing, catering to both price-conscious customers and those willing to invest slightly more for additional features and comfort when compared to Hatchback.
- SUV cars, on the other hand, have a smaller representation in the affordable price range. The majority of SUV models are priced between $50K and $70K, indicating their positioning in the premium segment.

Based on these observations, we can conclude that Hatchback cars are affordable, Sedan cars fall into the moderate price range, and SUV cars are primarily positioned as premium vehicles in terms of pricing.

Figure 30 : Price vs Age

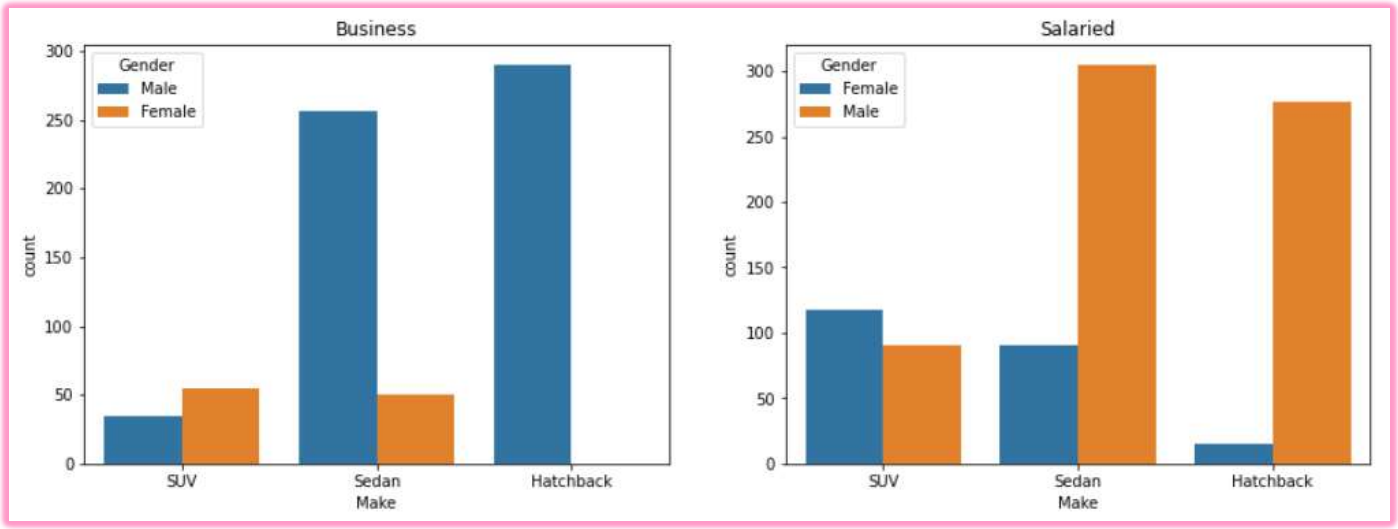


The Price vs Age plot from the Pair plot analysis reveals some interesting observations, especially considering the high correlation between these two variables. Here are the key insights:

- Customers between the ages of 22 and 30 tend to prefer cheaper cars in the price range of approximately $18K to $35K. This preference aligns with the affordability aspect, and as a result, they mostly opt for Hatchback cars, which are the most budget-friendly option. However, there are also a few customers in this age group who choose cheaper models of Sedan cars.
- Customers in the age range of 30 to 45 show a preference for cars priced between approximately $30K and $55K. This indicates a higher willingness to invest in vehicles with a slightly higher price tag. As a result, Sedan cars become a popular choice among this age group, likely due to the balance between affordability and additional features offered by Sedans.
- Customers above the age of 45 display a preference for more expensive cars priced above $55K. This demographic tends to lean towards SUV cars, which belong to the premium segment.

In summary, the Price vs Age plot allows us to infer that customers in different age ranges have distinct preferences when it comes to car prices. Younger customers prioritize affordability, while those in the middle age group are willing to spend slightly more, and older customers gravitate towards higher-priced vehicles in the premium segment.

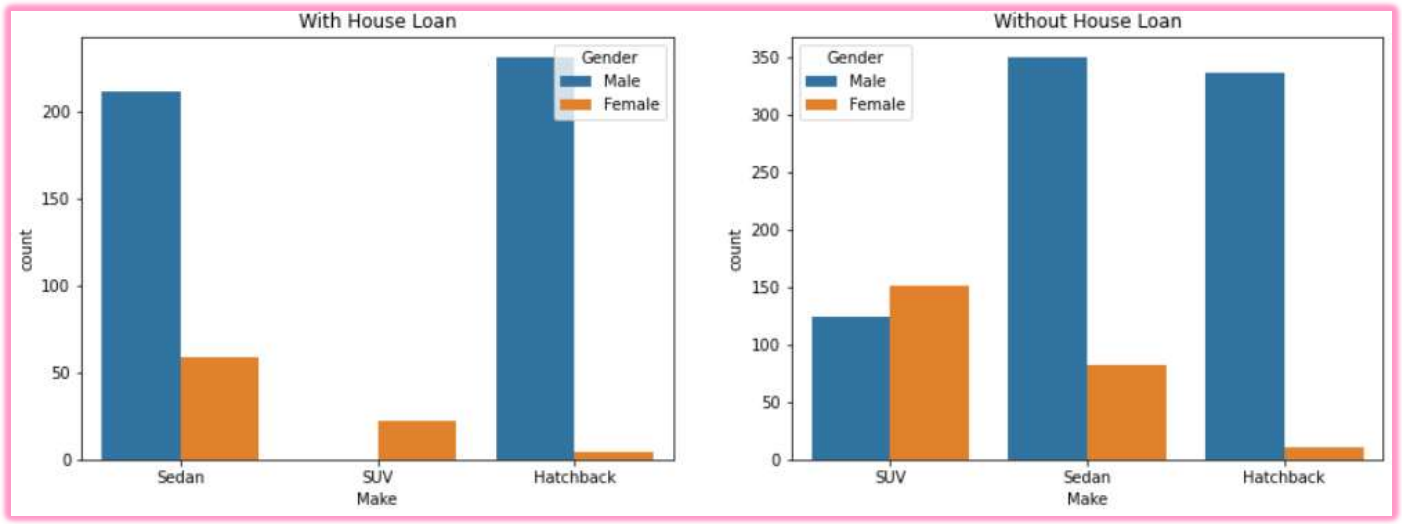
Figure 31 : Profession vs Make vs Gender



The above plots provide insights into the car preferences of different gender and profession groups.

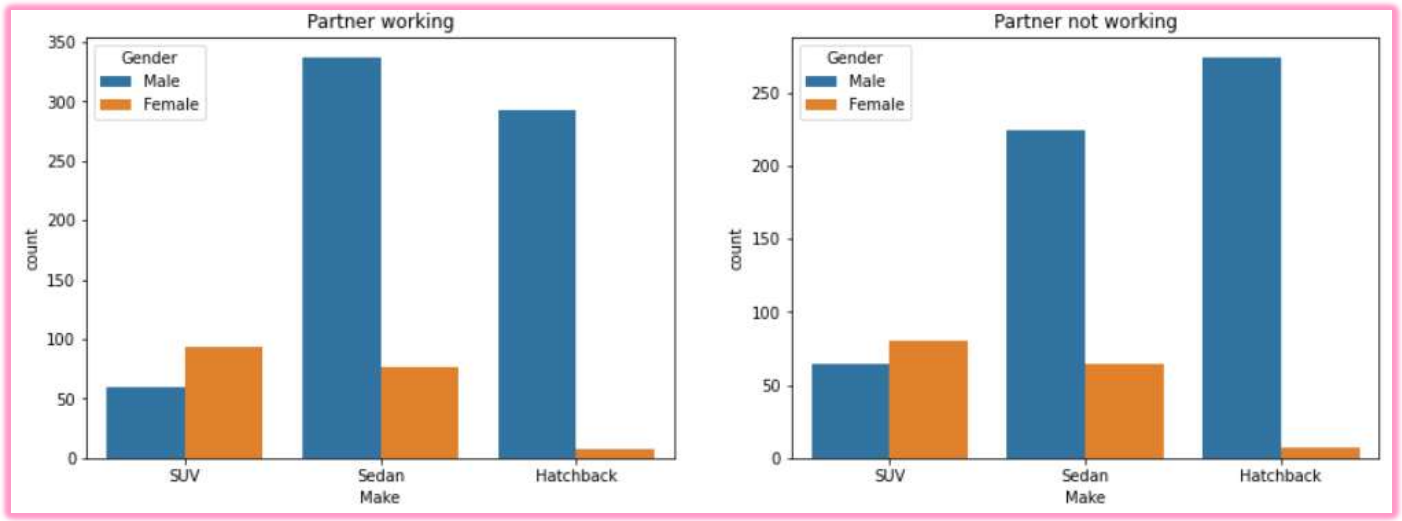
- Business women do not prefer Hatchback cars at all. they have a more even preference between Sedan and SUV cars, while Business men prefer Hatchback cars the most.
- Salaried women show a higher preference for SUVs, while Salaried men prefer Sedans the most, followed by Hatchbacks.

Figure 32 : House Loan vs Make vs Gender



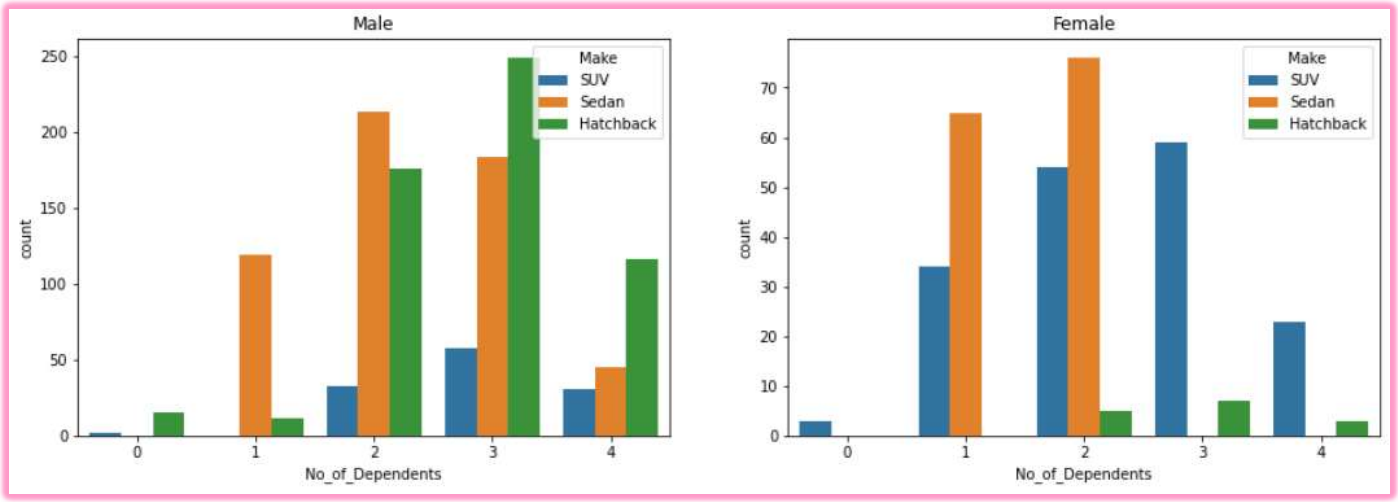
The above plot reveals that the presence of a house loan can have an impact on car preferences, particularly for males. Males with a house loan tend to favor Sedan and Hatchback cars, showing little to no preference for SUVs. In contrast, females with a house loan still exhibit some interest in SUVs, but their top choice remains Sedan, likely due to financial considerations and affordability. For males, their car preference is more balanced regardless of having a house loan, with a similar inclination towards Sedan and Hatchback models. However, SUVs are preferred by males only when they do not have a house loan. On the other hand, females without a house loan prioritize SUVs as their top choice, with Sedan being their secondary preference.

Figure 33 : Partner working vs Make vs Gender



Based on the analysis of the above plots, there doesn't seem to be a significant change in car preferences for females based on whether they have a working partner or not. Their preference remains consistent across Hatchback, Sedan, and SUV models. For males, there is a slight increase in the preference for Sedan cars when they have a working partner, as compared to when they don't. However, the overall difference is relatively small, indicating that the presence of a working partner has minimal impact on their car preference.

Figure 34 : No of Dependents vs Make vs Gender

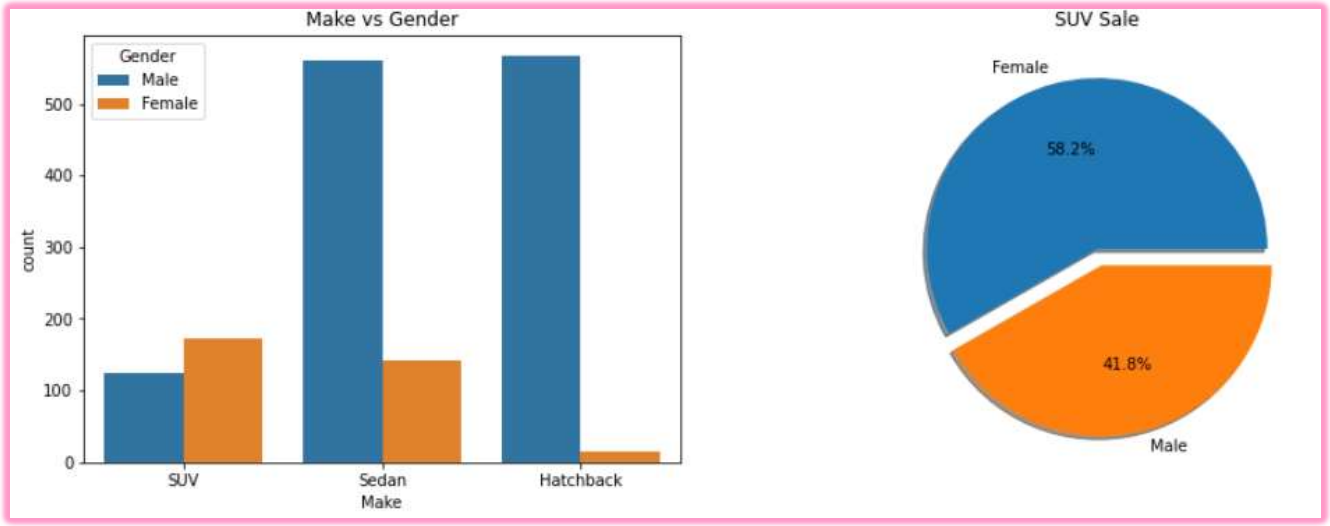


Based on the analysis of the above plot, we can infer that males tend to prefer Hatchback cars when they have a larger family size. On the other hand, females show a preference for SUVs. For families with an average family size, both males and females exhibit a preference for Sedan cars.

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.

E1) STEVE ROGER SAYS "MEN PREFER SUV BY A LARGE MARGIN, COMPARED TO THE WOMEN"

Figure 35 : Analyzing SUV preference based on Gender



Based on the above plots, it can be observed that a majority of SUVs (58%) were purchased by females, while males accounted for a smaller proportion (42%) of SUV purchases.

Figure 36 : Percentage of SUV sale based on Gender

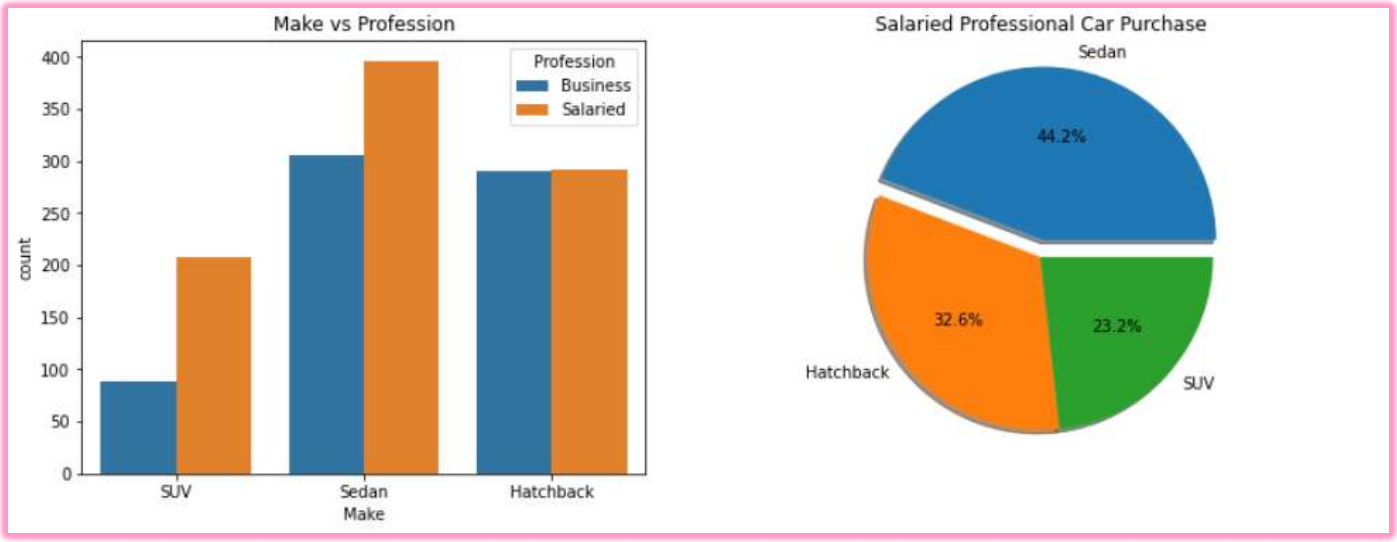
```
Female    52.583587
Male       9.904153
Name: Gender, dtype: float64
```

Based on above data, it can be observed that out of all the male customers, only 10% purchased an SUV, while among the female customers, 52% opted for an SUV. This stark difference in percentages highlights that SUVs are more popular among females compared to males.

Hence, we disagree with the statement made by Steve Rogers since it is not supported by the data.

E2) NED STARK BELIEVES THAT A SALARIED PERSON IS MORE LIKELY TO BUY A SEDAN.

Figure 37 : Analyzing Car preference for Salaried professionals

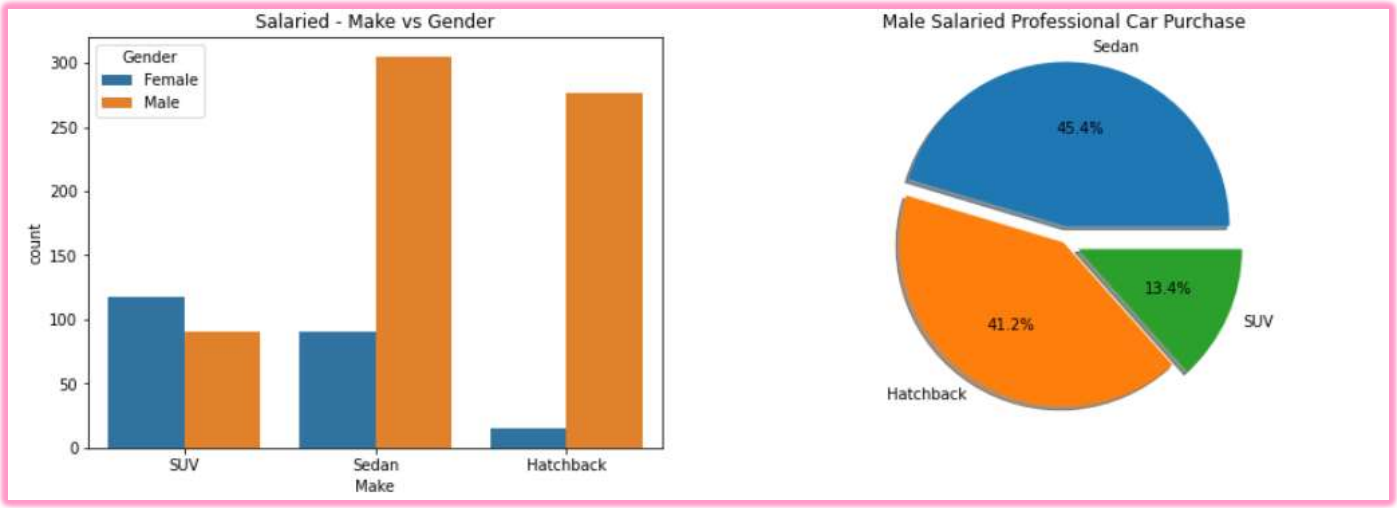


Based on the above plots, it is evident that a significant majority of Salaried Professionals (44.2%) prefer to purchase Sedan cars, while 32.6% opt for Hatchback and 23.2% choose SUV.

Therefore, we agree with Ned Stark that Salaried Person is more likely to buy Sedan cars.

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.

Figure 38 : Analyzing Male Salaried Professional Car Preference



Based on the above plots, it is evident that a significant majority of Male Salaried Professionals (45.4%) prefer to purchase Sedan cars, while 41.2% opt for Hatchback and 13.4% choose SUV, making SUV the least preferred car by Male Salaried Professionals.

Therefore, we disagree with Sheldon Cooper's claim that a Salaried Male is an Easier Target for SUV Sale over Sedan sale since it is not supported by the data.

F. From the given data, comment on the amount spent on purchasing automobiles 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. Give justification along with presenting metrics/charts used for arriving at the conclusions.

F1) GENDER

Figure 39 : Analyzing Purchase based on Gender

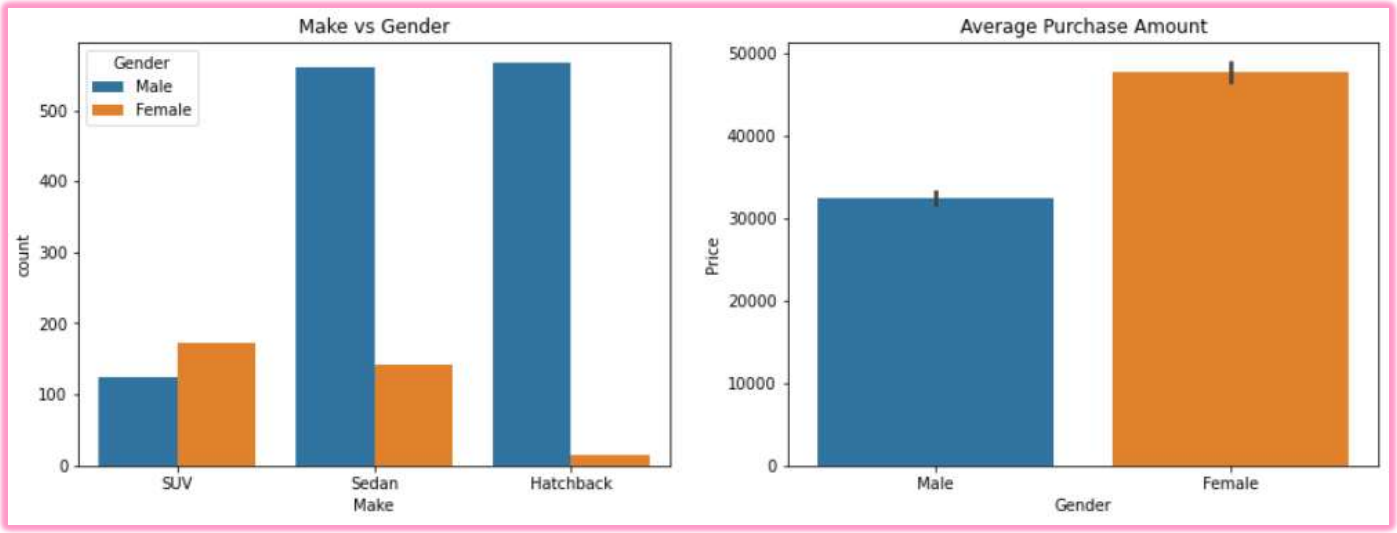


Figure 40 : Statistical data based on Gender

Gender	count	mean	std	min	25%	50%	75%	max
Female	329.0	47705.167173	11244.836378	20000.0	38000.0	49000.0	55000.0	69000.0
Male	1252.0	32416.134185	12366.253107	18000.0	23000.0	29000.0	37000.0	70000.0

Based on the above plot, it is evident that females show a preference for purchasing SUVs. Additionally, referring to **Figure 29: Price vs Make**, we have already established that SUVs belong to the premium segment and are generally more expensive than other car models. This is reflected in the average purchase amount, where females have an average purchase amount of $47,705 compared to males with an average purchase amount of only $32,416.

Considering that females constitute only 21% of the total population in the dataset, it is advisable for the company to focus their marketing campaign specifically on females when promoting SUV cars. Since it is evident that SUVs are generating higher revenue for the company. By targeting their marketing efforts towards promoting SUVs, the company can capitalize on the higher sales potential and maximize their profitability.

F2) PERSONAL_LOAN

Figure 41 : Analyzing Purchase based on Personal Loan

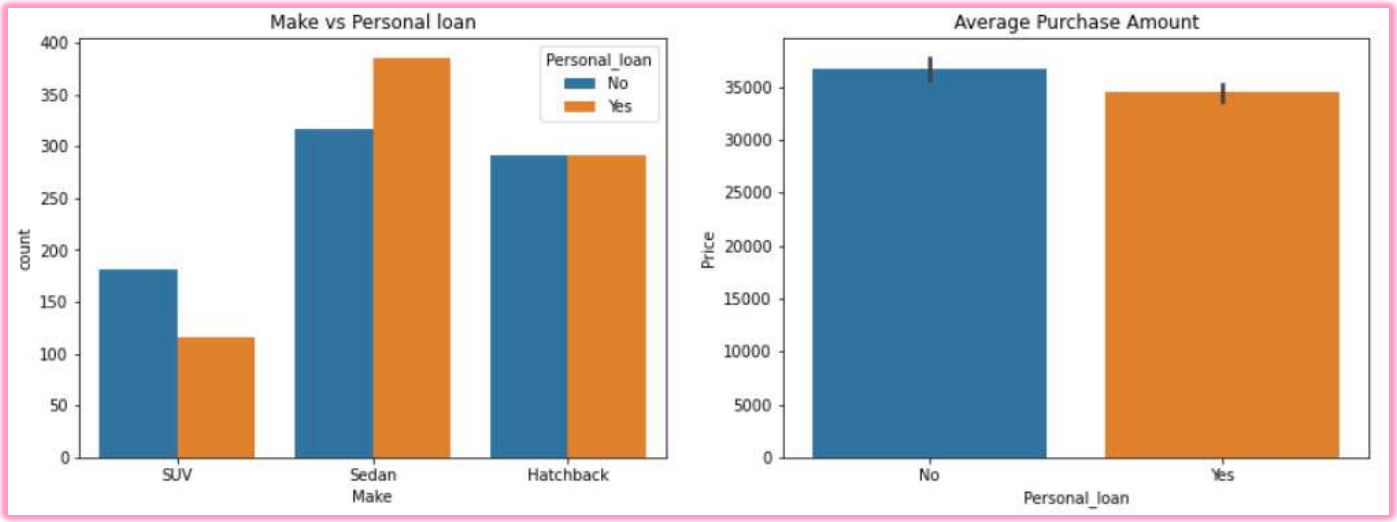


Figure 42 : Statistical data based on Personal Loan

Personal_loan	count	mean	std	min	25%	50%	75%	max
No	789.0	36742.712294	14534.344526	18000.0	25000.0	32000.0	49000.0	70000.0
Yes	792.0	34457.070707	12578.780338	18000.0	24000.0	31000.0	45000.0	70000.0

From the analysis of the above plot, it can be observed that customers without a personal loan have a slightly higher average purchase amount of $36,742 compared to customers with a personal loan, who have an average purchase amount of $34,457. Although there is a difference, it is not significantly large, indicating that the presence of a personal loan does not have a major impact on total sales.

Further analysis reveals that the presence of a personal loan does not have a significant effect on the sales of Sedan and Hatchback cars. However, the sale of SUVs is lower for customers with a personal loan. This can be attributed to the fact that SUVs tend to be more expensive compared to other car models. To address this, the company can consider providing more favorable offers or incentives to customers with a personal loan, specifically targeting SUV purchases. By offering attractive financing options or discounts, the company can potentially encourage customers with a personal loan to consider purchasing an SUV, thus boosting sales and maximizing profit.

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

Figure 43 : Analyzing purchase of higher priced

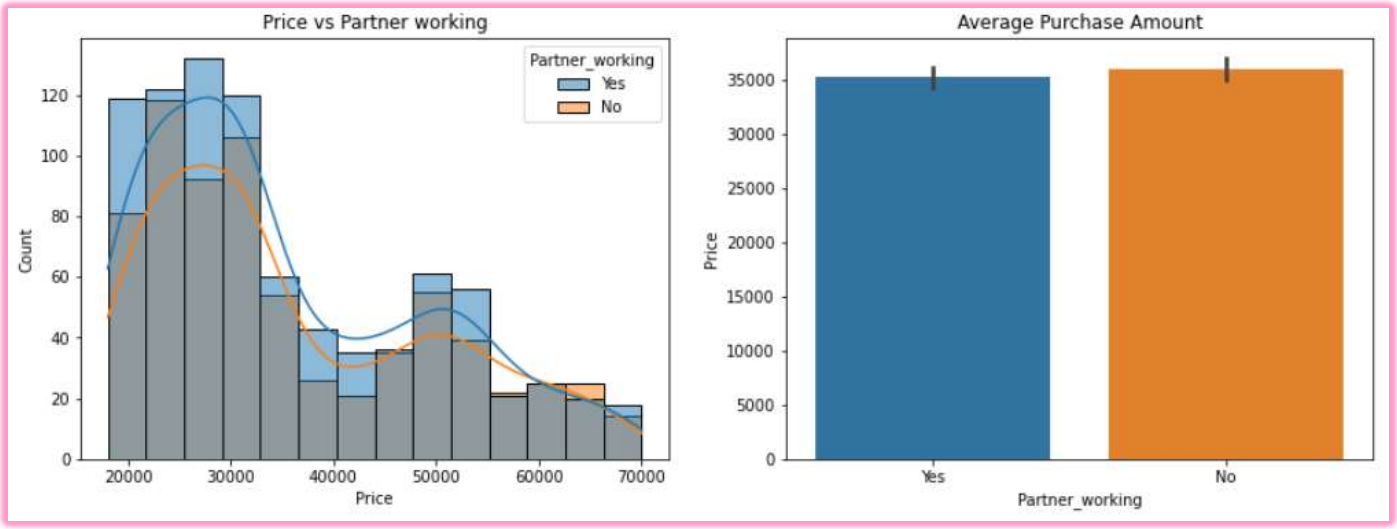


Figure 44 : Statistical data based on Partner working

Partner_working	count	mean	std	min	25%	50%	75%	max
No	713.0	36000.000000	13817.734086	18000.0	25000.0	31000.0	48000.0	70000.0
Yes	868.0	35267.281106	13479.532555	18000.0	24000.0	31000.0	46000.0	70000.0

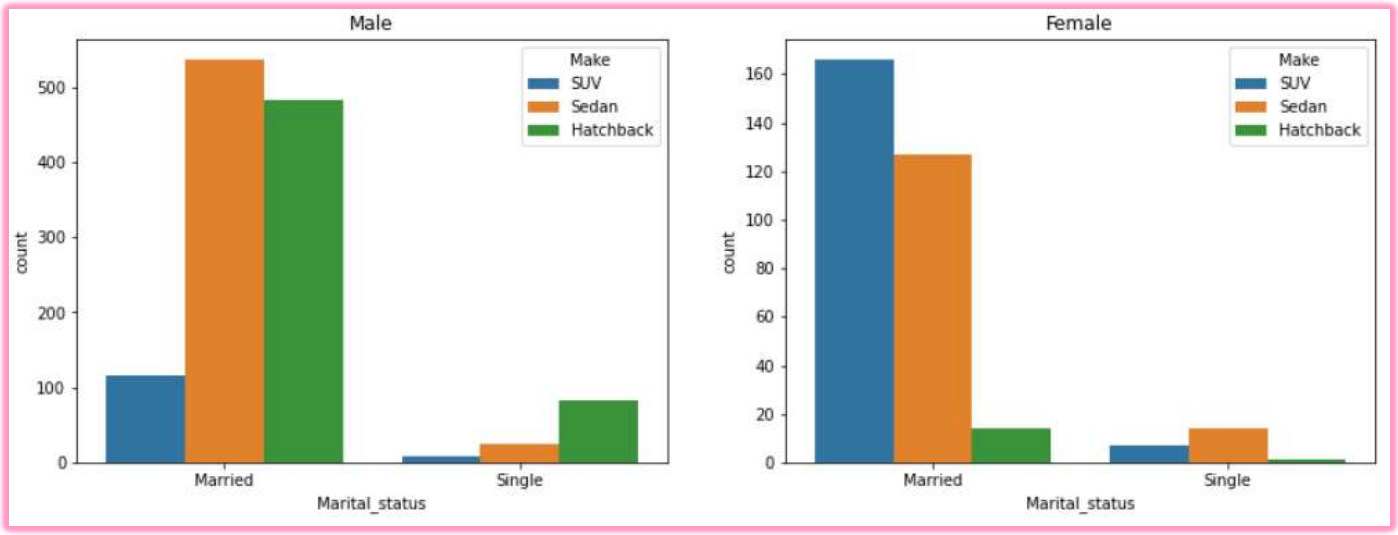
Based on the analysis of the above plots, it can be inferred that the average purchase amount for customers with working partners ($35,267) is almost equal to the average purchase amount for customers without working partners ($36,000). This suggests that having a working partner does not necessarily lead to purchasing a higher-priced car.

Furthermore, when comparing the 75th percentile values, it can be observed that customers with working partners have a 75th percentile purchase amount of $48,000, while customers without working partners have a 75th percentile purchase amount of $46,000. This indicates that there is a slight difference in the higher range of purchase amounts, but it is not substantial enough to conclude that having a working partner significantly influences the decision to purchase a high-priced car.

Based on this data, it can be concluded that the presence of a working partner does not necessarily result in a higher purchase amount or a preference for high-priced cars.

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.

Figure 45 : Marital Status vs Gender vs Make



Based on the analysis of the above plot, the following observations can be made:

- Married Men: Married men show a preference for Sedan cars, followed by Hatchback. They have a lower preference for SUVs, which are higher-priced cars. This suggests that married men are more price-conscious and may prioritize affordability when making car purchase decisions.
- Single Men: Single men prefer Hatchback cars the most, followed by Sedan. This preference can be attributed to the fact that single men, typically in the age group of 20-30, may have lower incomes compared to married men who are generally older. Hatchback cars tend to be more affordable and cater to their budgetary constraints.
- Married Women: Married women show a preference for SUVs, followed by Sedan. Hatchback cars are the least preferred option for them. This indicates that women, in general, may not be as price conscious as men and are more inclined to purchase expensive cars like SUVs.
- Single Women: Single women prefer Sedan cars the most, followed by SUVs. This suggests that although single women may prefer expensive cars, they tend to prioritize affordability compared to married women. Sedans are relatively more affordable than SUVs, making them a preferred choice for single women.

These observations provide insights into the car preferences of different demographic groups based on marital status and gender.

Marketing Strategy based on the Observations:

1. Married Men: Provide periodic promotions, discounts, and special offers on Sedan and Hatchback cars to attract price-conscious married men. Highlight the affordability and value for money these models offer. Offer flexible financing options, such as low-interest rates or extended repayment periods, to make car ownership more feasible for married men with varying budgets.

2. Single Men:
    - Special offers for newcomers: Create special offers and incentives targeted at single men who are new to their careers or the workforce. This can include discounted prices, exclusive financing options, or additional benefits to encourage them to consider Hatchback cars as an affordable and practical choice.
    - Youth-oriented events: Organize events or partnerships that cater to the interests of the younger generation, such as car showcases at office campuses. This can help create awareness and attract the attention of single men, increasing traffic and potential conversions.

3. Married Women: This segment of existing Customers can be moving Brand ambassadors. Conduct events such Drive through the Wilderness, explore unexplored territories & certain additional women centric activities during Women's day. Sponsor some of their kitty parties and utilize their influences to carry out demo's at Club's and other social events to generate lead.

4. Single Women: Provide exclusive offers and incentives targeted specifically at single women, such as discounted prices, cashback offers, or complimentary maintenance packages. This can create a sense of value and affordability for Sedan cars, making them more appealing.

2. A physiotherapist with a male football team is interested in studying the relationship between foot injuries and the positions at which the players play from the data collected.

	Striker	Forward	Attacking Midfielder	Winger	**Total**
Players Injured	45	56	24	20	**145**
Players Not Injured	32	38	11	9	**90**
Total	**77**	**94**	**35**	**29**	**235**

2.1 WHAT IS THE PROBABILITY THAT A RANDOMLY CHOSEN PLAYER WOULD SUFFER AN INJURY?

P (Injury) = 145/235

The probability of randomly chosen player would suffer an injury is **61.70%**

2.2 WHAT IS THE PROBABILITY THAT A PLAYER IS A FORWARD OR A WINGER?

P (Forward U Winger) = P(Forward) + P(Winger)
P (Forward U Winger) = (94/235) +(29/235)

Probability that a Player is a forward or a winger is **52.34%**

2.3 WHAT IS THE PROBABILITY THAT A RANDOMLY CHOSEN PLAYER PLAYS IN A STRIKER POSITION AND HAS A FOOT INJURY?

P (Striker ∩ Injury) = 45/235

Probability that a randomly chosen player plays in a striker position and has a foot injury is **19.15%**

2.4 WHAT IS THE PROBABILITY THAT A RANDOMLY CHOSEN INJURED PLAYER IS A STRIKER?

P (Striker | Injured) = 45/145

Probability that a randomly chosen injured player is a striker is **31.03%**

2.5 WHAT IS THE PROBABILITY THAT A RANDOMLY CHOSEN INJURED PLAYER IS EITHER A FORWARD OR AN ATTACKING MIDFIELDER?

P (Forward U Midfielder) = P(Forward) + P(Midfielder)
P (Forward U Midfielder) = (56/145) + (24/145)

Probability that a randomly chosen injured player is either a forward or an attacking midfielder is **55.17%**

3. An independent research organization is trying to estimate the probability that an accident at a nuclear power plant will result in radiation leakage. The types of accidents possible at the plant are, fire hazards, mechanical failure, or human error. The research organization also knows that two or more types of accidents cannot occur simultaneously.

According to the studies carried out by the organization, the probability of a radiation leak in case of a fire is 20%, the probability of a radiation leak in case of a mechanical 50%, and the probability of a radiation leak in case of a human error is 10%. The studies also showed the following:

- The probability of a radiation leak occurring simultaneously with a fire is 0.1%.
- The probability of a radiation leak occurring simultaneously with a mechanical failure is 0.15%.
- The probability of a radiation leak occurring simultaneously with a human error is 0.12%.

On the basis of the information available, answer the questions below:

Given,

P (Radiation | Fire) = **0.2**

P (Radiation | Mechanical failure) = **0.5**

P (Radiation | Human error) = **0.1**

P (Fire ∩ Radiation) = **0.001**

P (Mechanical Failure ∩ Radiation) = **0.0015**

P (Human error ∩ Radiation) = **0.0012**

3.1 WHAT ARE THE PROBABILITIES OF A FIRE, A MECHANICAL FAILURE, AND A HUMAN ERROR RESPECTIVELY?

- P (Fire) = P (Fire ∩ Radiation) / P (Radiation | Fire)

  P (Fire) = 0.001/0.2

  Probability of Fire is **0.50%**

- P (Mechanical failure) = P (Mechanical failure ∩ Radiation) / P (Radiation | Mechanical failure)

  P (Mechanical failure) = 0.0015/0.5

  Probability of Mechanical failure is **0.30%**

- P (Human error) = P (Human error ∩ Radiation) / P (Radiation | Human error)

  P (Human error) = 0.0012/0.1

  Probability of Human error is **1.20%**

3.2 WHAT IS THE PROBABILITY OF A RADIATION LEAK?

P (Radiation) = P (Fire ∩ Radiation) + P (Mechanical failure ∩ Radiation) + P (Human error ∩ Radiation)

P(Radiation) = 0.001 + 0.0015+ 0.0012

Probability of Radiation Leak is - **0.37%**

3.3 SUPPOSE THERE HAS BEEN A RADIATION LEAK IN THE REACTOR FOR WHICH THE DEFINITE CAUSE IS NOT KNOWN. WHAT IS THE PROBABILITY THAT IT HAS BEEN CAUSED BY:

- A Fire

  P (Fire | Radiation) = P (Fire ∩ Radiation)/P (Radiation)

  P (Fire | Radiation) = 0.001/0.0037

  Probability of Fire given Radiation Leak is **27.03%**

- A Mechanical Failure

  P (Mechanical Failure | Radiation) = P (Mechanical Failure ∩ Radiation)/P (Radiation)

  P (Mechanical Failure | Radiation) = 0.0015/0.0037

  Probability of Fire given Radiation Leak is **40.54%**

- A Human Error

  P (Human Error | Radiation) = P (Human Error and Radiation)/P (Radiation)

  P (Human Error | Radiation) = 0.0012/0.0037

  Probability of Fire given Radiation Leak is **32.43%**

4. The breaking strength of gunny bags used for packaging cement is normally distributed with a mean of 5 kg per sq. centimetre and a standard deviation of 1.5 kg per sq. centimetre. The quality team of the cement company wants to know the following about the packaging material to better understand wastage or pilferage within the supply chain; Answer the questions below based on the given information; **(Provide an appropriate visual representation of your answers, without which marks will be deducted)**

4.1 WHAT PROPORTION OF THE GUNNY BAGS HAVE A BREAKING STRENGTH LESS THAN 3.17 KG PER SQ CM?

Given,

Mean = 5

Standard deviation (SD) = 1.5

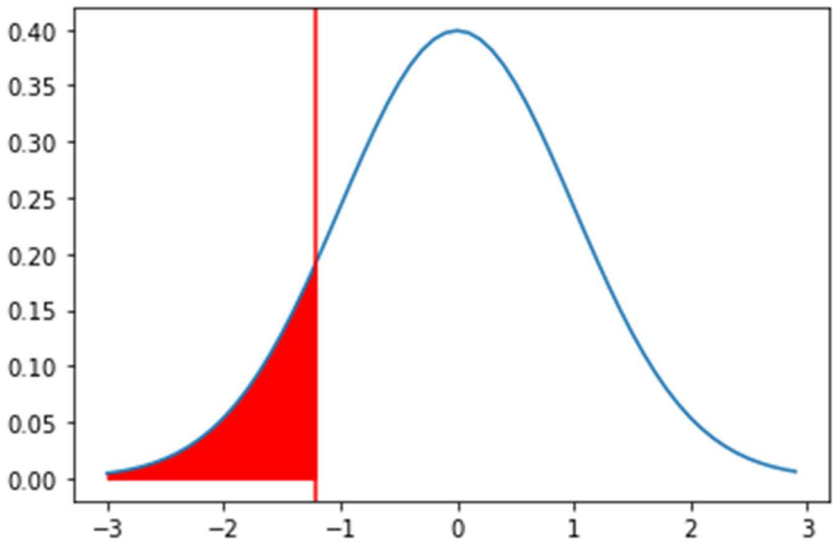
And we know that Z score = (X – Mean)/SD

z1 = (3.17 - 5) / 1.5 = -1.22

We will use the **stats.norm.cdf(-1.22)** function to calculate the area to the left of the distribution (shaded region in plot)

Proportion of the gunny bags having a breaking strength less than 3.17 kg per sq. cm is **11.12%**

Figure 46 : Normal Distribution - 1



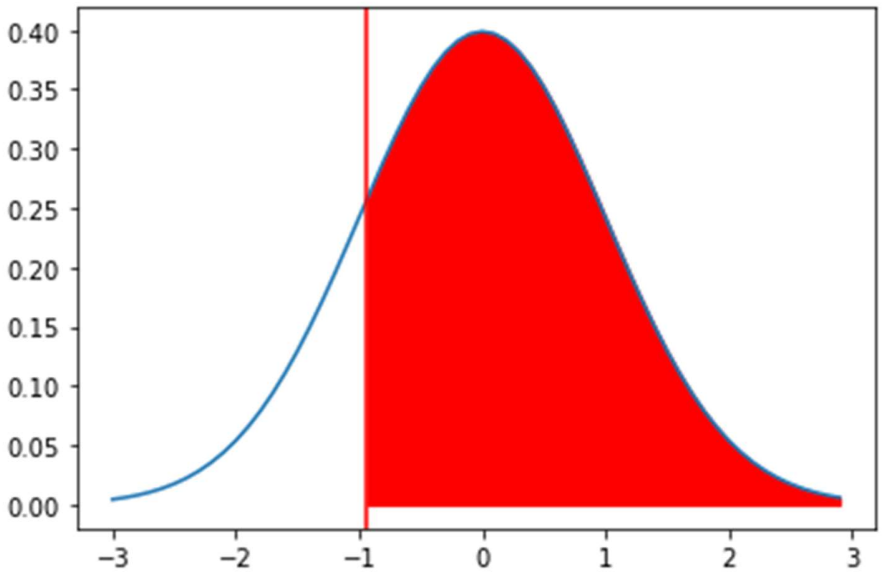
4.2 WHAT PROPORTION OF THE GUNNY BAGS HAVE A BREAKING STRENGTH AT LEAST 3.6 KG PER SQ CM?

z2 = (3.6 - 5) / 1.5 = -0.93

We will use the **stats.norm.sf(-0.93)** function to calculate the area to the right of the distribution (shaded region in plot)

Proportion of the gunny bags have a breaking strength at least 3.6 kg per sq. cm is **82.47%**

Figure 47 : Normal Distribution - 2



4.3 WHAT PROPORTION OF THE GUNNY BAGS HAVE A BREAKING STRENGTH BETWEEN 5 AND 5.5 KG PER SQ CM?

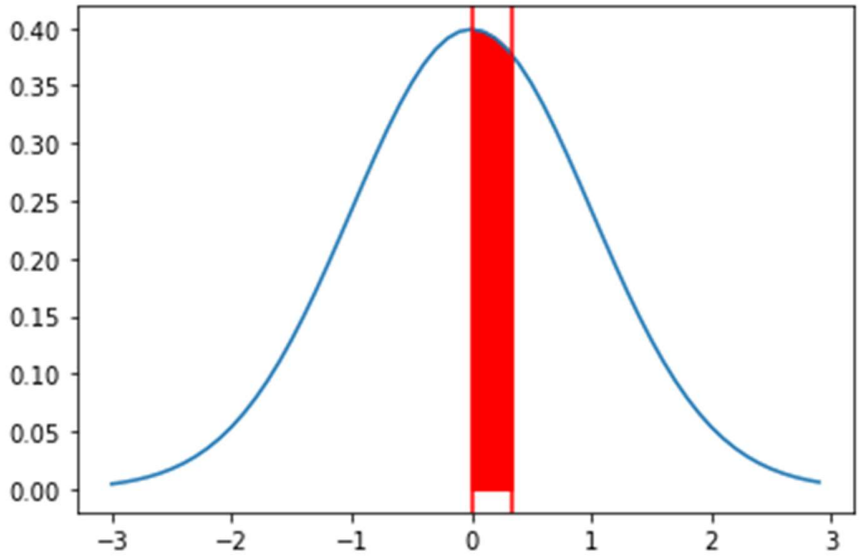
z1 = (5.5 - 5) / 1.5 = 0.33

z2 = (5 - 5) / 1.5 = 0

We will use the **stats.norm.cdf(0.33) - stats.norm.cdf(0)** to calculate the area of the shaded region in below distribution plot.

Proportion of the gunny bags having a breaking strength between 5 and 5.5 kg per sq. cm is **13.06%**

Figure 48 : Normal Distribution - 3



4.4 WHAT PROPORTION OF THE GUNNY BAGS HAVE A BREAKING STRENGTH NOT BETWEEN 3 AND 7.5 KG PER SQ CM?

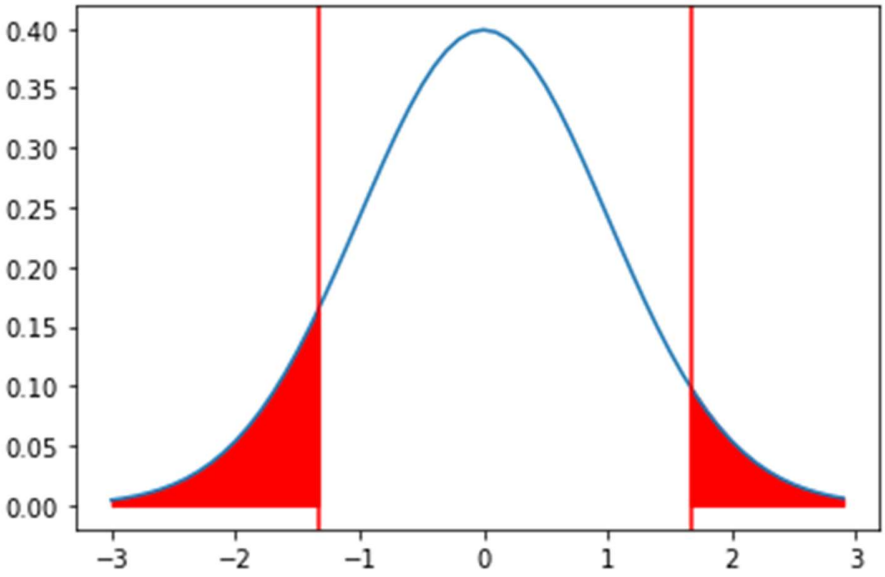
z1 = (3 - 5) / 1.5 = -1.33

z2 = (7.5 - 5) / 1.5 = 1.67

We will use the **stats.norm.cdf(-1.33)** + **stats.norm.sf(1.67)** to calculate the area of the shaded region in below distribution plot.

Proportion of the gunny bags have a breaking strength NOT between 3 and 7.5 kg per sq cm is **13.90%**

Figure 49 : Normal Distribution - 4



5. Grades of the final examination in a training course are found to be normally distributed, with a mean of 77 and a standard deviation of 8.5. Based on the given information answer the questions below.

5.1 WHAT IS THE PROBABILITY THAT A RANDOMLY CHOSEN STUDENT GETS A GRADE BELOW 85 ON THIS EXAM?

Given,

Mean = 77

Standard deviation (SD) = 8.5

And we know that Z score = (X – Mean)/SD

z1 = (85-77)/8.5 = 0.94

We will use the **stats.norm.cdf(0.94)** to calculate the area to the left of the distribution.

Probability that a randomly chosen student gets a grade below 85 on this exam is **82.67%**

5.2 WHAT IS THE PROBABILITY THAT A RANDOMLY SELECTED STUDENT SCORES BETWEEN 65 AND 87?

z1 = (87-77)8.5 = 1.17

z2 = (65-77)/8.5 = -1.41

We will use the **stats.norm.cdf(1.17) - stats.norm.cdf(-1.41)** to calculate the area between these two points

Probability that a randomly selected student scores between 65 and 87 is **80.13%**

5.3 WHAT SHOULD BE THE PASSING CUT-OFF SO THAT 75% OF THE STUDENTS CLEAR THE EXAM?

Loc = 77

Scale=8.5

We will use the **stats.norm.ppf(0.25, loc = 77, scale = 8.5)** function to get the cutoff score

Passing cut off score so that 75% of the students clear the exam is **71.27**