Python for Data Science
Project Report

Contents

List of Tables

List of Figures

Problem Definition

Context

In the 21st century, cars are an important mode of transportation that provides us the opportunity for personal control and autonomy. In day-to-day life, people use cars for commuting to work, shopping, visiting family and friends, etc. Research shows that more than 76% of people prevent themselves from traveling somewhere if they don't have a car. Most people tend to buy different types of cars based on their day-to-day necessities and preferences. So, it is essential for automobile companies to analyze the preference of their customers before launching a car model into the market. Austo, a UK-based automobile company aspires to grow its business into the US market after successfully establishing its footprints in the European market.

In order to be familiar with the types of cars preferred by the customers and factors influencing the car purchase behavior in the US market, Austo has contracted a consulting firm. Based on various market surveys, the consulting firm has created a dataset of 3 major types of cars that are extensively used across the US market. They have collected various details of the car owners which can be analyzed to understand the automobile market of the US.

Objective

Austo's management team wants to understand the demand of the buyers and trends in the US market. They want to build customer profiles based on the analysis to identify new purchase opportunities so that they can manipulate the business strategy and production to meet certain demand levels. Further, the analysis will be a good way for management to understand the dynamics of a new market. Suppose you are a Data Scientist working at a consulting firm that has been contracted by Austo. You are given the task to create buyer profiles for different types of cars with the available data as well as a set of recommendations for Austo. Perform the data analysis to generate useful insights that will help the automobile company to grow its business.

Data Description

austo_automobile.csv: The dataset contains buyer's data corresponding to different types of products(cars).

Data Dictionary

- Age: Age of the customer
- Gender: Gender of the customer
- Profession: Indicates whether the customer is a salaried or business person
- Marital_status: Marital status of the customer
- Education: Refers to the highest level of education completed by the customer
- No_of_dependents: Number of dependents(partner/children/spouse) of the customer

- Personal_loan: Indicates whether the customer availed a personal loan or not
- House_loan: Indicates whether the customer availed house loan or not
- Partner_working: Indicates whether the customer's partner is working or not
- Salary: Annual Salary of the customer
- Partner_salary: Annual Salary of the customer's partner
- Total_salary: Annual household income (Salary + Partner_salary) of the customer's family
- Price: Price of the car
- Make: Car type (Hatchback/Sedan/SUV)

Data Overview

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

The dataset has 1581 rows and 14 columns. It is always a good practice to view a sample of the rows. A simple way to do that is to use head() function.

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

Table 1: Top five rows of the dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1581 entries, 0 to 1580
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Age              1581 non-null   int64
 1   Gender           1528 non-null   object
 2   Profession       1581 non-null   object
 3   Marital_status   1581 non-null   object
 4   Education        1581 non-null   object
 5   No_of_Dependents 1581 non-null   int64
 6   Personal_loan    1581 non-null   object
 7   House_loan       1581 non-null   object
 8   Partner_working  1581 non-null   object
 9   Salary           1581 non-null   int64
 10  Partner_salary   1475 non-null   float64
 11  Total_salary     1581 non-null   int64
 12  Price            1581 non-null   int64
 13  Make             1581 non-null   object
dtypes: float64(1), int64(5), object(8)
memory usage: 173.0+ KB
```

Table 2: Basic Information of the Dataset

A quick look at the dataset information tells us that there are 6 are numerical and 8 categorical variables. There are few Null records present in two variables: Gender and Partner_salary, which will be analyzed in detail in the next section. There are no duplicate records in the dataset.

Missing value treatment

Inspecting Null Values -

There are Nulls in Gender and Partner_salary variables.

- Gender - total 53 Nulls
- Partner_salary - Total 106 Nulls

Handling Nulls -

Nulls are usually handled by the following techniques –

- If the proportion of Null values is more than 60 % of the total number of records in a column, then drop the column. Here you assume that the column is uninformative.
- If any row is missing a large amount of records across columns then that row may also be dropped.
- Otherwise, the missing values may be imputed.

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

Simple rules for imputation:

- For categorical variables, we can impute the Nulls with the majority class. For the current dataset, Null values in the 'Gender' field are imputed with 'Male' (Male being the majority class).

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

The three variables on salary are related to one another:

Salary + Partner_salary = Total_salary

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

- If Partner_working = 'No' then Partner_salary = 0
- If Partner_working = 'Yes' then Partner_salary = Total_salary - Salary

Statistical Summary

Inspecting the Summary Statistics of the Dataset (Numerical fields)

	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

Table 3: Numerical summarization of the dataset

Observations:

- The average age of the customers is around 32 years. 75% of the customers are below 38 years and the minimum age of the customer is 22. This indicates that buyers in the age group 22-38 purchase new cars.
- 50% of the customers have at least 2 dependents.
- The salary of the customer lies between 30,000 to 90,000, with an average of around 60,000 and a standard deviation of 14,278. The mean salary is almost equal to the median, this suggests that salary distribution is symmetrical.
- At least 25% of the customer's partners are not working. The average partner's salary of the customer is around 20000. The mean salary is less than the median, this suggests that salary distribution will be left-skewed.
- The average household salary of the customer is around 80000, with a standard deviation of around 25000. The mean salary is approximately equal to the median, this suggests that salary distribution is symmetrical.
- The price of the car lies in the range of 18000 to 70000 with an average of around 36000. The mean salary is greater than the

median, this suggests that salary distribution will be a bit right-skewed.

Checking for anomalous values in categorical variables

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

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

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

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

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

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

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

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

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

Table 4: Value Counts of the Categorical Variables

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

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

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

Univariate Analysis

For performing Univariate analysis we will take a look at the Boxplots and Histograms to get a better understanding of the distributions.

Numerical variables

- **Observations on Age**

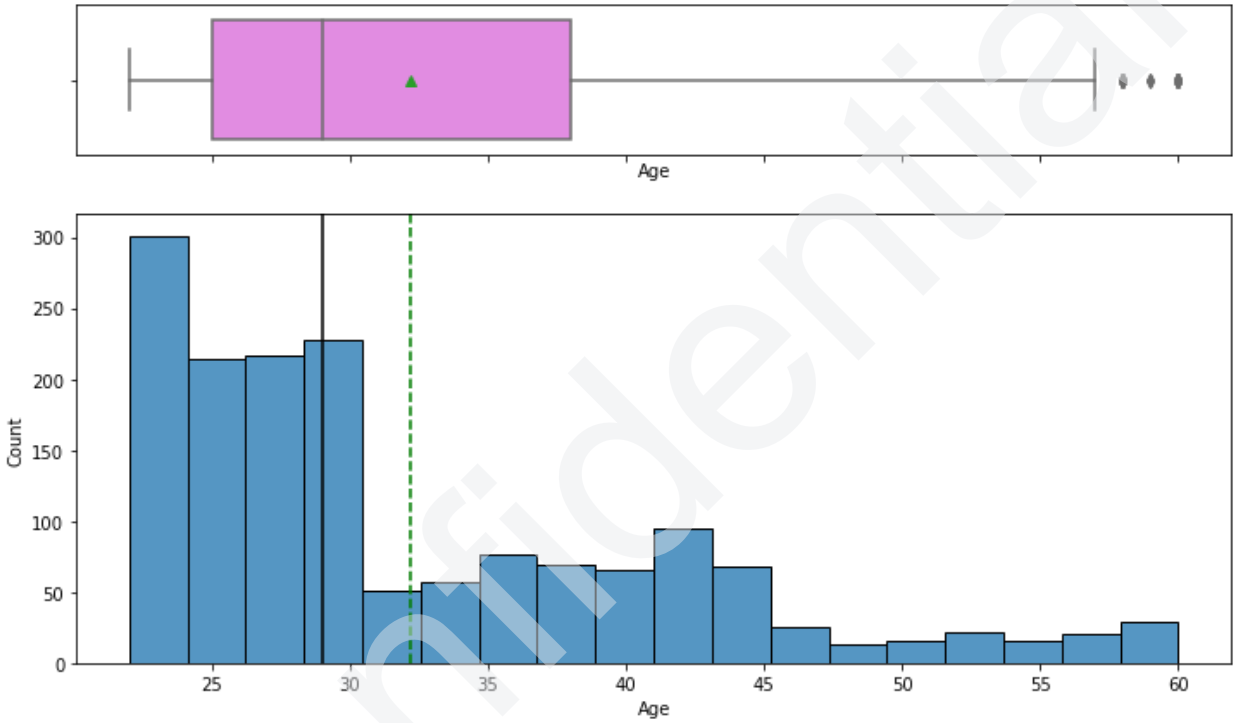


Figure 1: Univariate analysis of Age

Observations:

- The distribution of Age is right skewed.
- From the boxplot we can see that the second quartile(Q2) is less than 30 which means more than 50% of customers in the dataset are below the age of 30.
- There are a few outliers in this variable.

- **Observations on Salary**

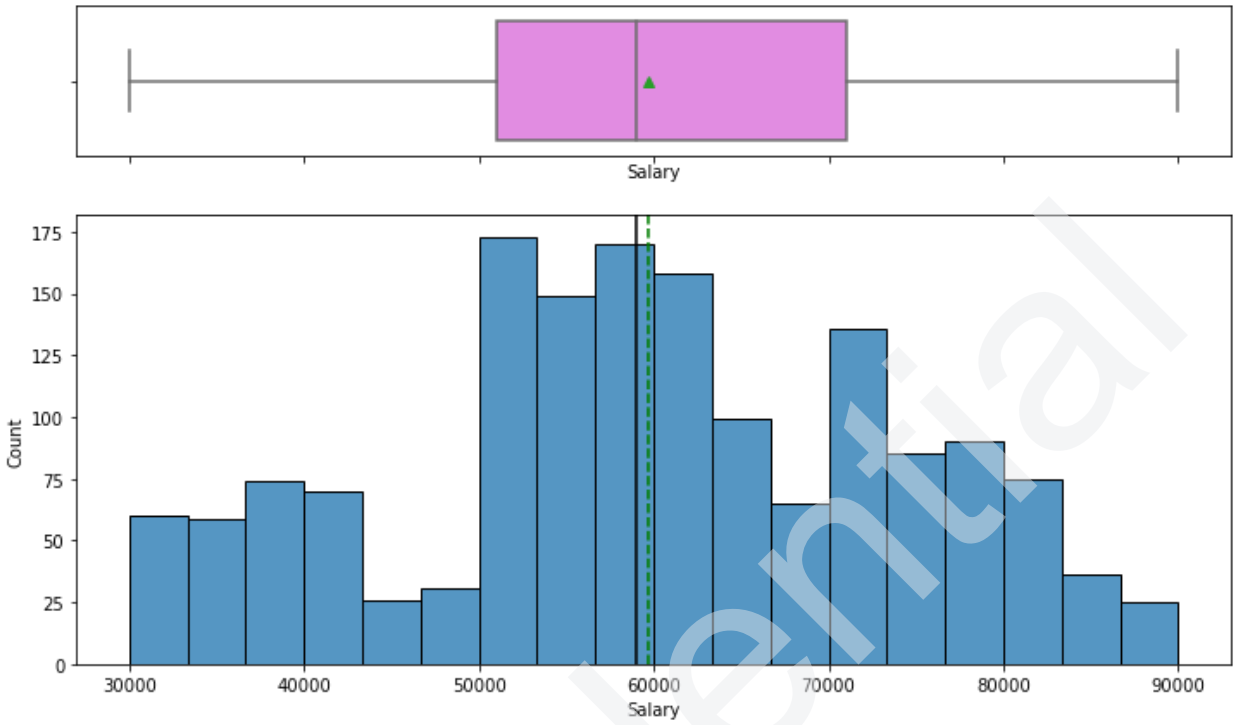


Figure 2: Univariate analysis of Salary

Observations:

- The salary of the customer lies between 30,000 to 90,000, with an average of around 60,000.
- The mean salary is almost equal to the median.

- **Observations on Partner's salary**

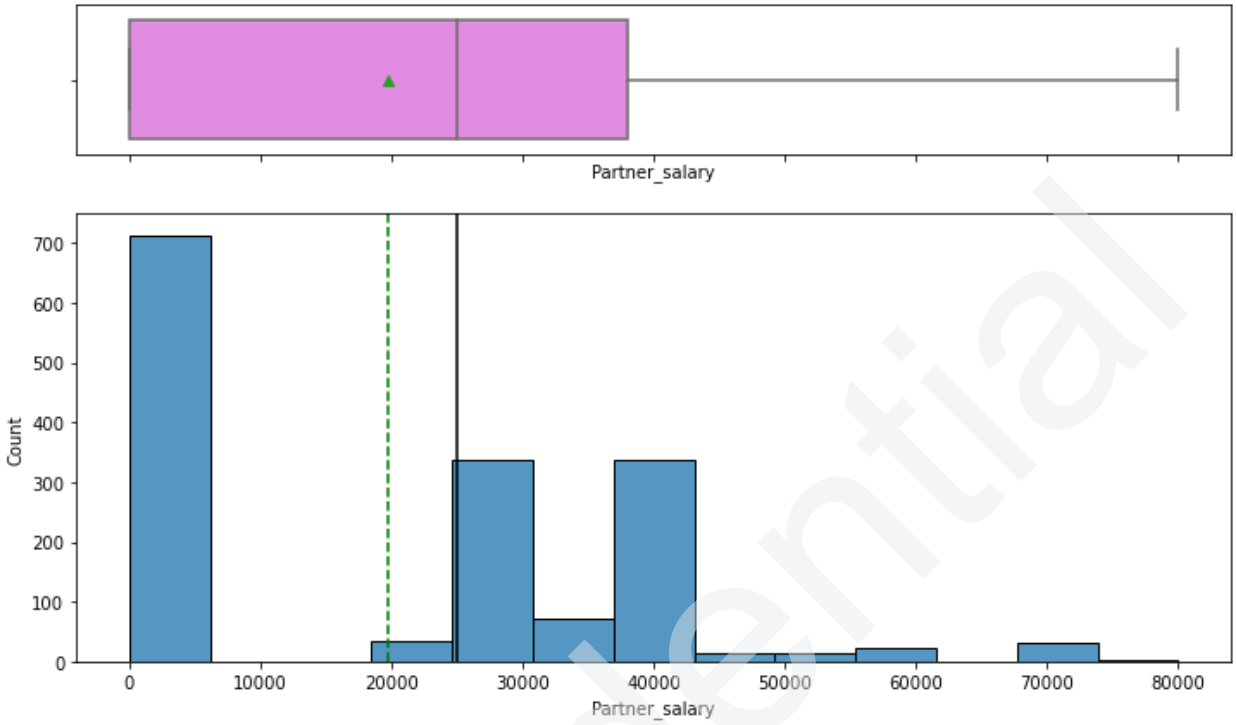


Figure 3: Univariate analysis of Partner_salary

Observations:

- Around 45% of the customer's partners do not work. Hence, their salary is 0.
- Most of the working partners earn in the range of 20000-60000.

- **Observations on Total salary**

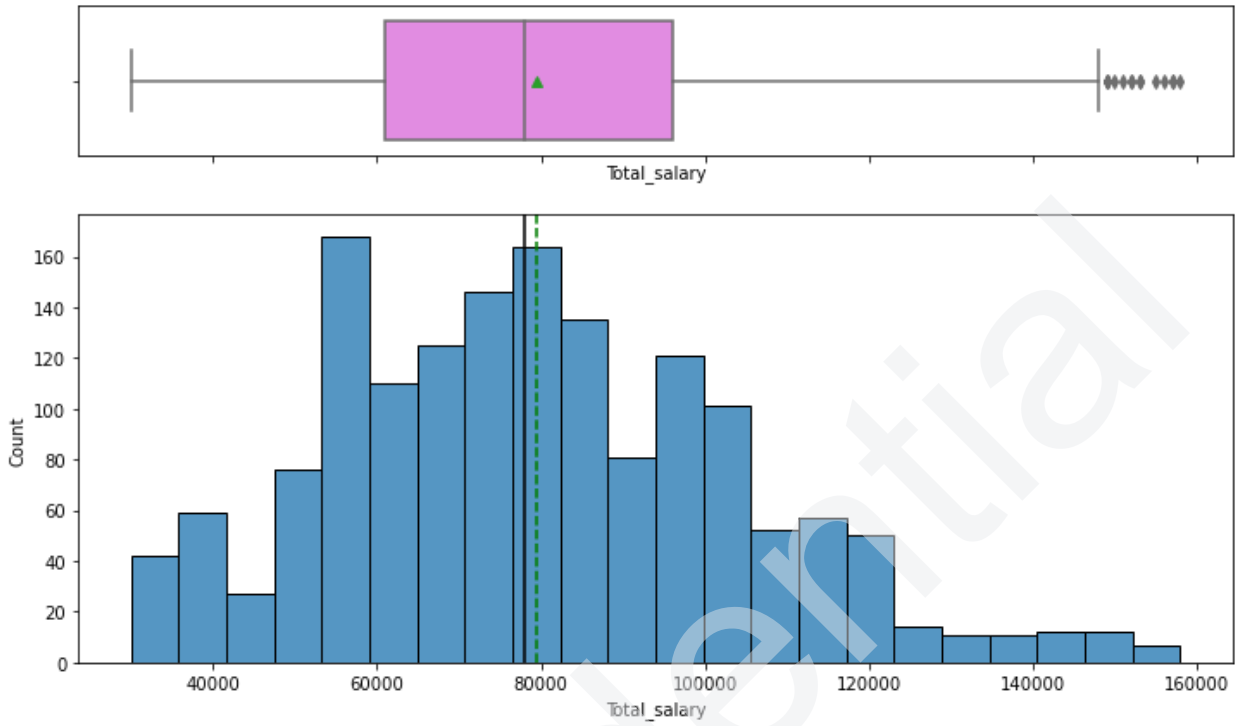


Figure 4: Univariate analysis of Total_salary

Observations:

- The total salary of the customer's household follows a normal distribution, with an average of around 80,000.
- The mean salary is almost equal to the median.
- There are a few outliers in this variable. However, we will not treat them as if they are proper values.

- **Observations on Price**

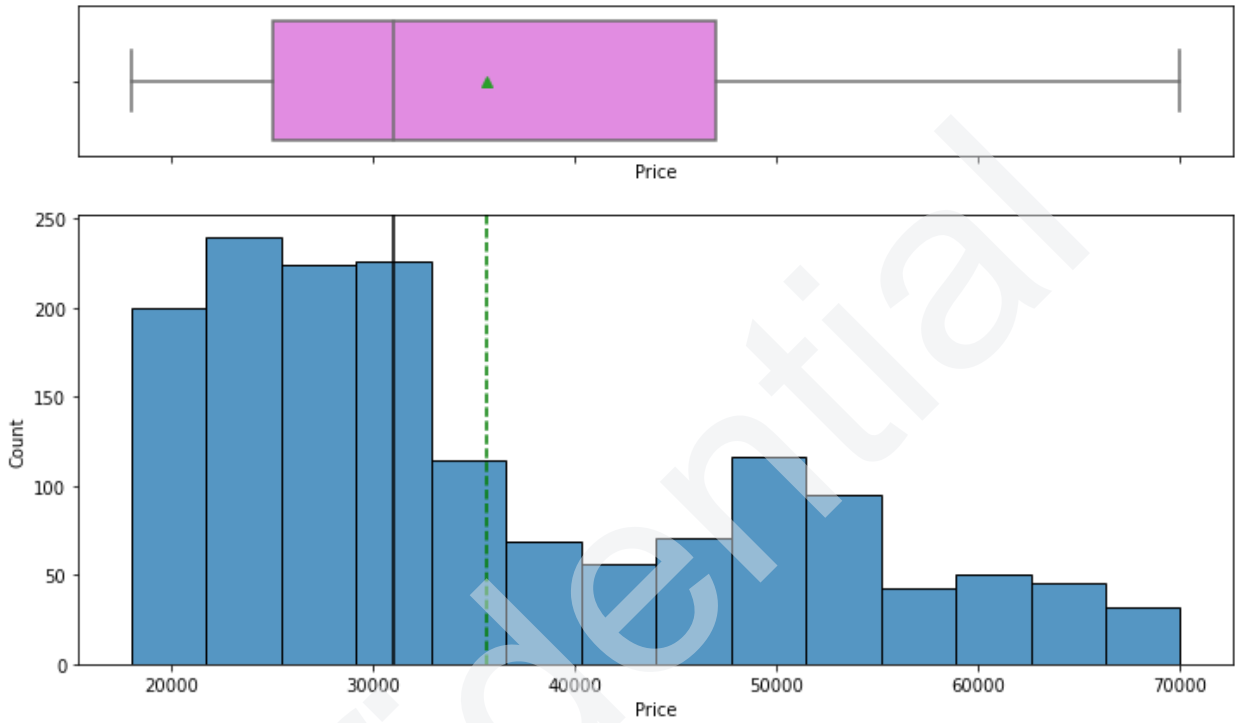


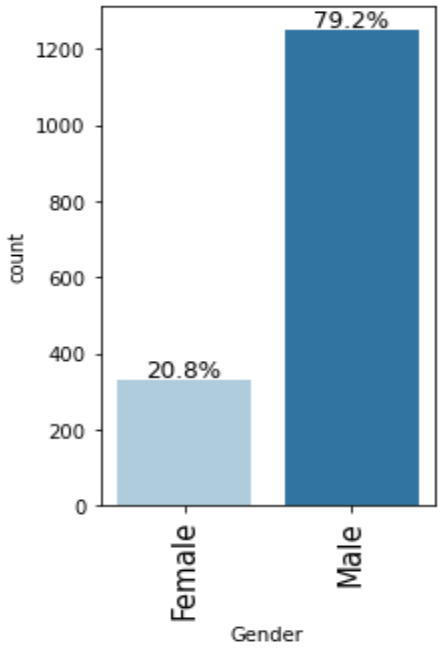
Figure 5: Univariate analysis of Price

Observations:

- Most of the cars cost in the range 20000-40000.
- The mean price of the cars is greater than the median. This indicates that the car price is right-skewed.

Categorical variables

- **Observations on Gender**

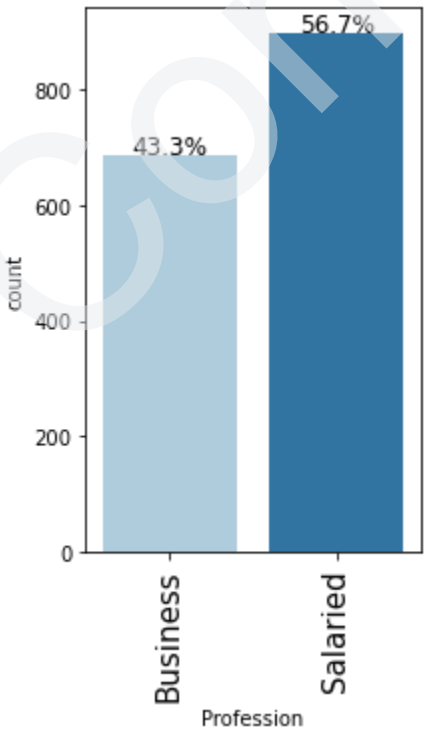


Observations:

- There are more male customers(around 79%) than females(around 21%).

Figure 6: Univariate analysis of Gender

- **Observations on Profession**

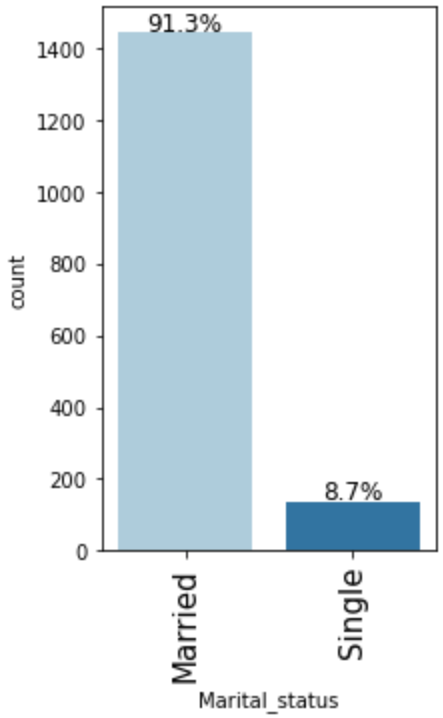


Observations:

- There are more salaried customers(around 57%) than business persons(around 43%).

Figure 7: Univariate analysis of Profession

- **Observations on Marital Status**



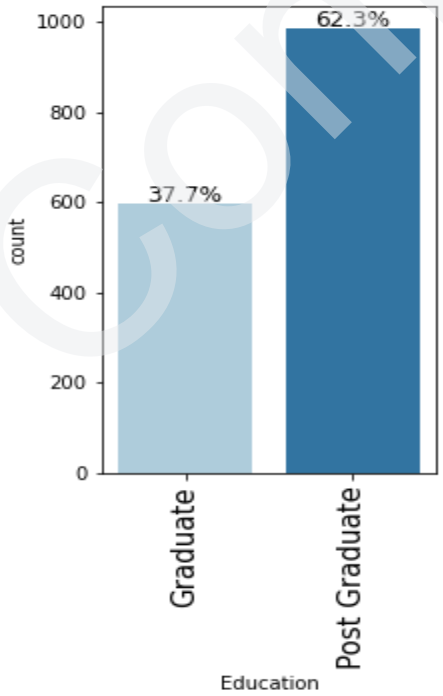
Observations:

- 91.3% customers are married.

Only 8.7% customers are single.

Figure 8: Univariate analysis of Marital Status

- **Observations on Education**

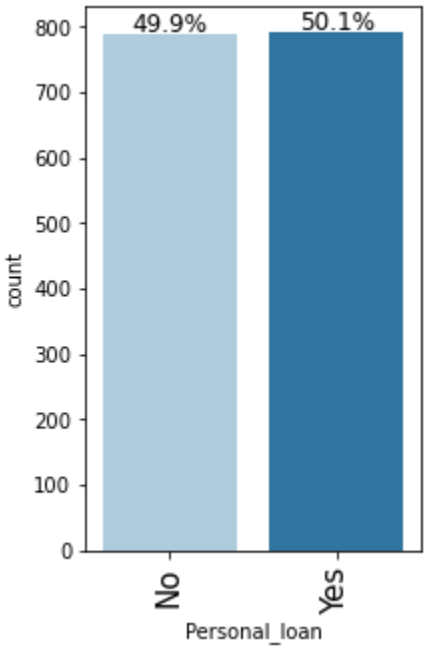


Observations:

- Around 38% customers are graduate; whereas 62% have completed their post graduation.

Figure 9: Univariate analysis of Education

- **Observations on Personal Loan**

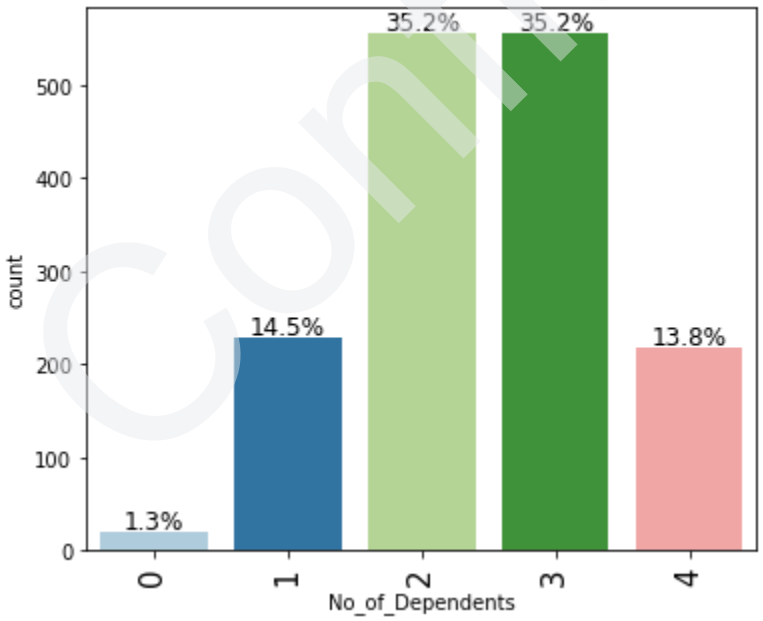


Observations:

- Around 50% of the customers have a personal loan.

Figure 10: Univariate analysis of Personal Loan

- **Observations on Number of dependents**

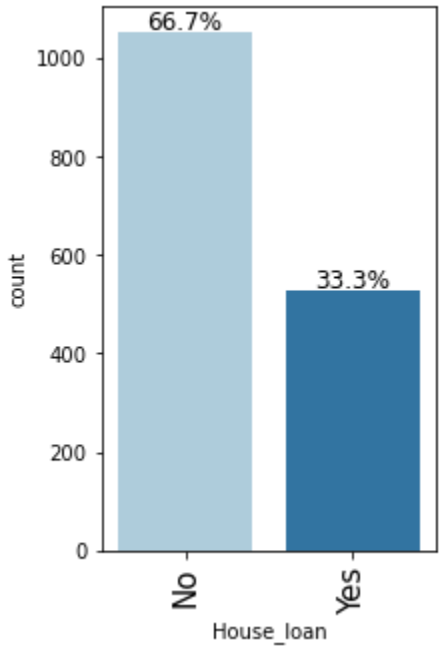


Observations:

- Around 84% of the customers have at least 2 dependents.

Figure 11: Univariate analysis of Number of dependents

- **Observations on House loan**

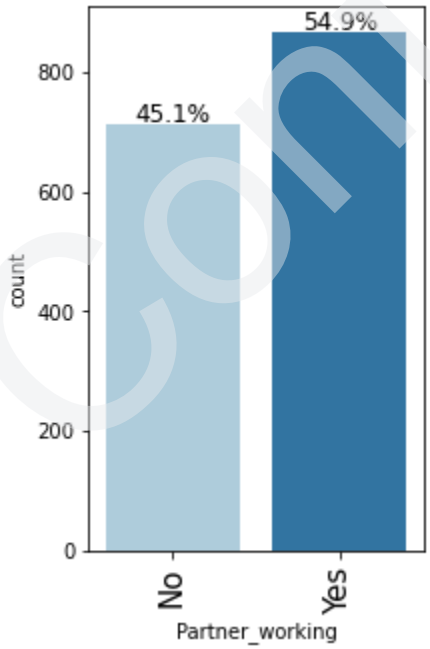


Observations:

- Around 33% of the customers have a house loan.

Figure 12: Univariate analysis of House Loan

- **Observations on Partner working**

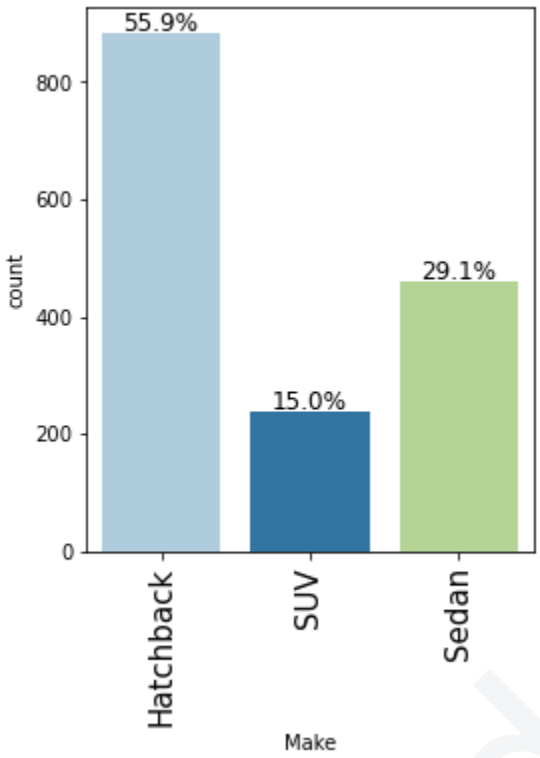


Observations:

- Around 55% of the customers have working partners.

Figure 13: Univariate analysis of Partner working

- **Observations on Make**



Observations:

- The Sale of the 'Hatchback' type car is more compared to SUV and Sedan.
- Only 15% of the customers buy SUVs.

Figure 14: Univariate analysis of Make

Insights

- Sedan is the most preferred purchase, followed by Hatchback and SUV.
- The number of customers having a working partner are slightly higher than customers with nonworking partner or singles. There are a total of 713 customers with Partner_working variable as 'No', out of which 138 customers are 'Single'.
- Number of Customers who did not take a House Loan is almost double the customers who took a House Loan.
- The data consists of very small proportion of Single customers when compared to married customers.
- Count of Salaried customers is slightly higher than that of Business customers.

- Majority of the customers in the dataset are Post Graduate.
- From the Barplot of No_of_dependnts variable we can infer that majority of the customers have either 2 or 3 dependents, followed by 1 or 4 dependents. Very few customers have zero no of dependents.

Multivariate Analysis

- **Correlation of Numerical Variables**

Figure 15: Correlation of numerical variables

Observations:

- Age is moderately correlated with the customer's salary. This is expected as the salary of the customers in the higher age group will be more compared to the lower ones.
- Age is highly correlated with the price of the car. It is possible that higher age group customers tend to buy costly cars.

- **Show the relationship between numerical variables**

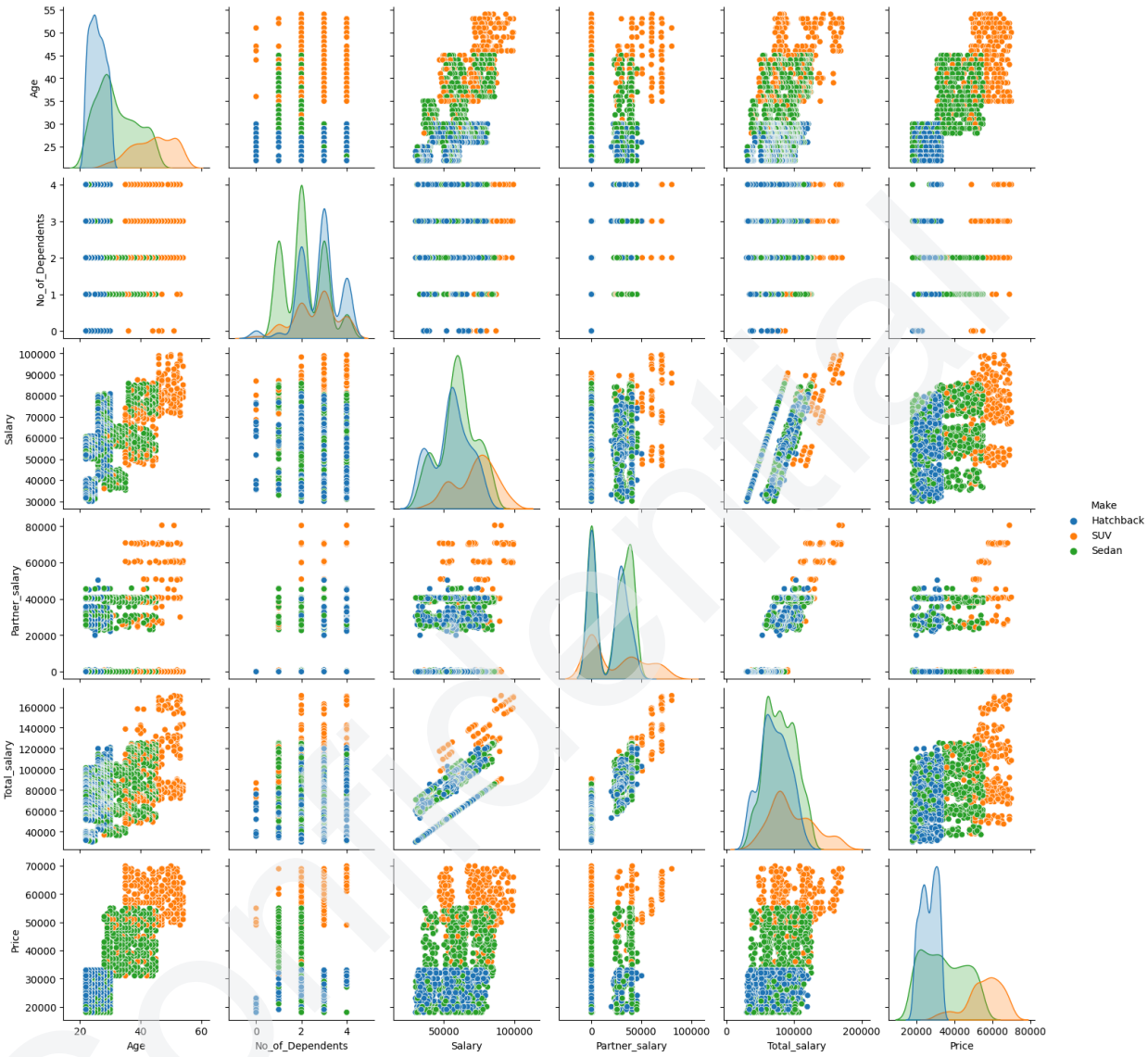


Figure 16: Relationship between numerical variables

Observations:

- Customers with higher household salaries prefer SUVs and sedans; whereas customers with lower household salaries prefer Hatchback cars.
- Customers in the higher age group prefer SUVs; whereas young customers prefer hatchbacks.
- Let's analyze it further to get more insights.

Find relationship between Numerical and Categorical variables

- **Make vs Age**

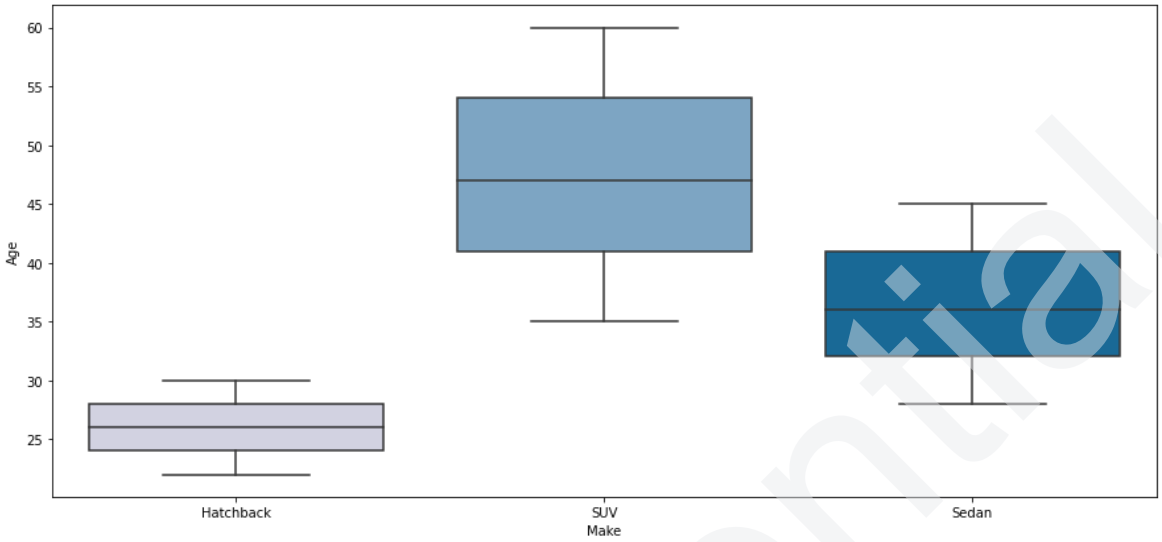


Figure 17: Make vs Age Plot

Observations:

- SUV is preferred by customers in the age group 35-60.
- Sedan is preferred by customers in the age group 30-45.
- Hatchback is preferred by the younger customers in the age group 22-30.

- **Make vs Price**

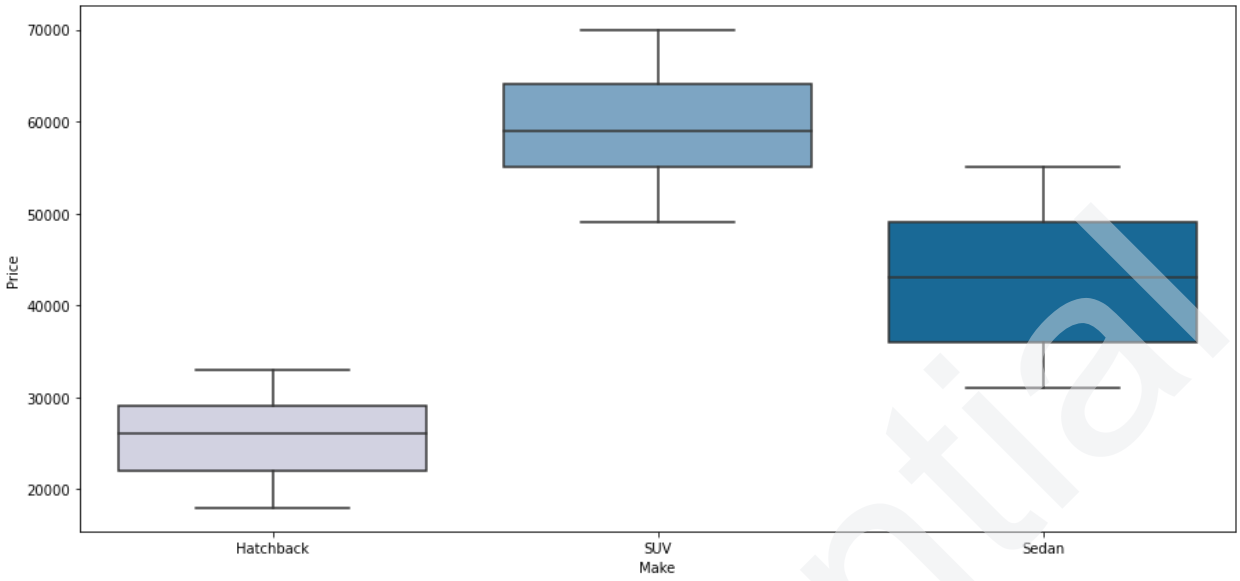


Figure 18: Make vs Price

Observations:

- SUV is the costliest type of car among the three car types. The price range of the SUVs is 50000-70000.
- Sedan is costlier compared to hatchback type cars.
- Hatchback is the most affordable car ranging between 15000-35000.

- **Make vs Salary**

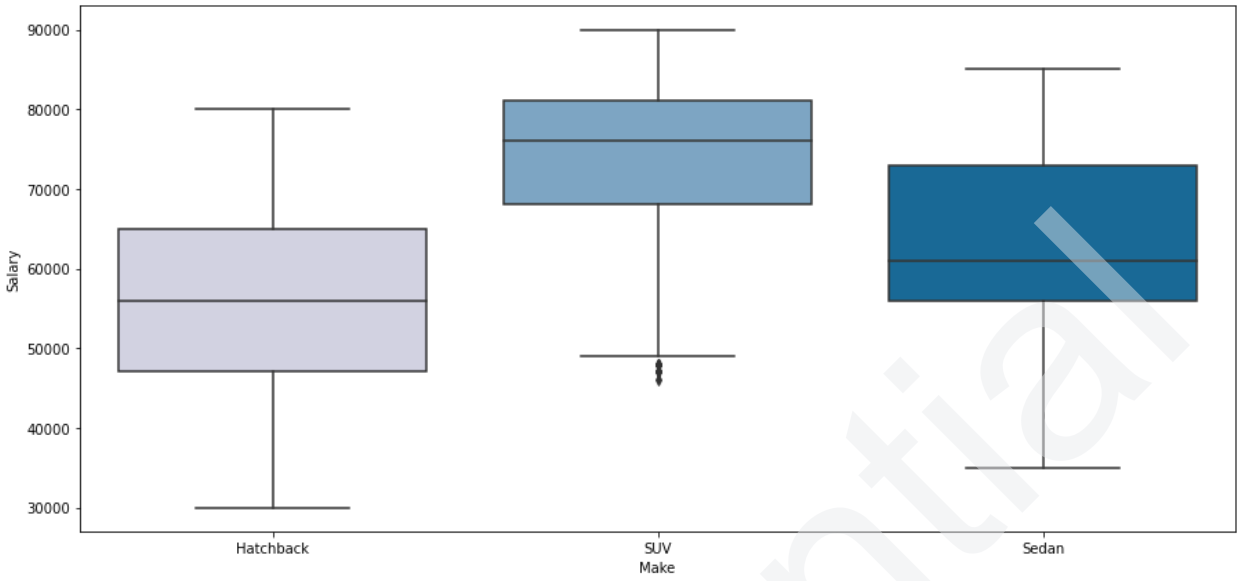


Figure 19: Make vs Salary

Observations:

- SUV is the costliest type of car among the three car types. Hence, customers with higher household incomes prefer to buy SUVs.

- **Make vs Education**

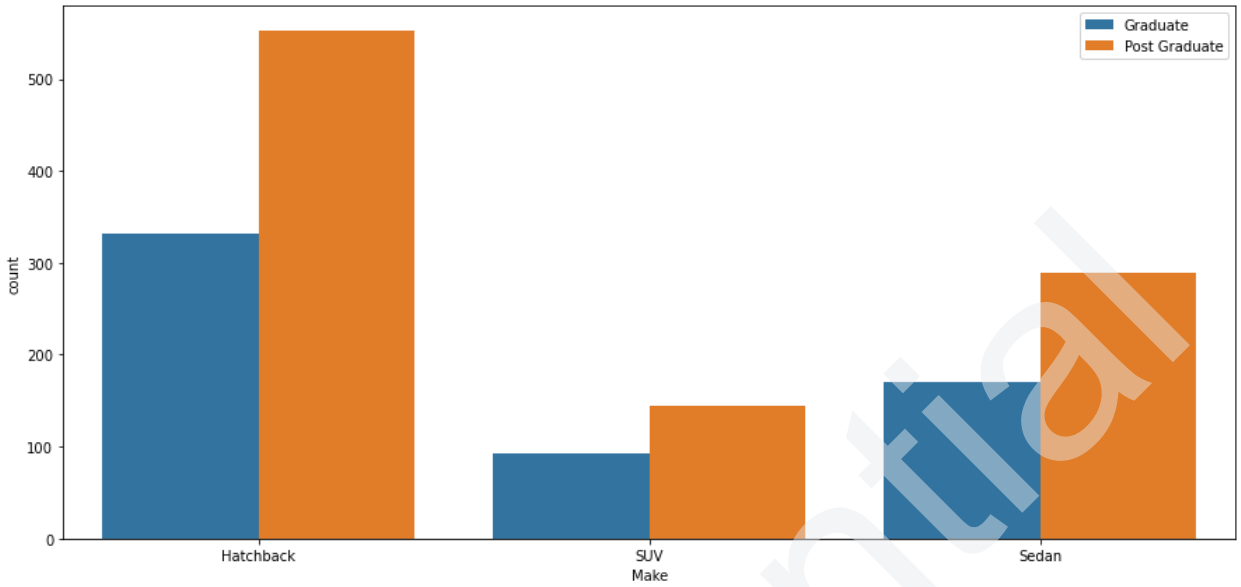


Figure 20: Make vs Education

Observations:

- Customers with higher education are more tend to buy cars. As observed Post Graduates have purchased more cars of all types.

- **Make vs Number of dependents**

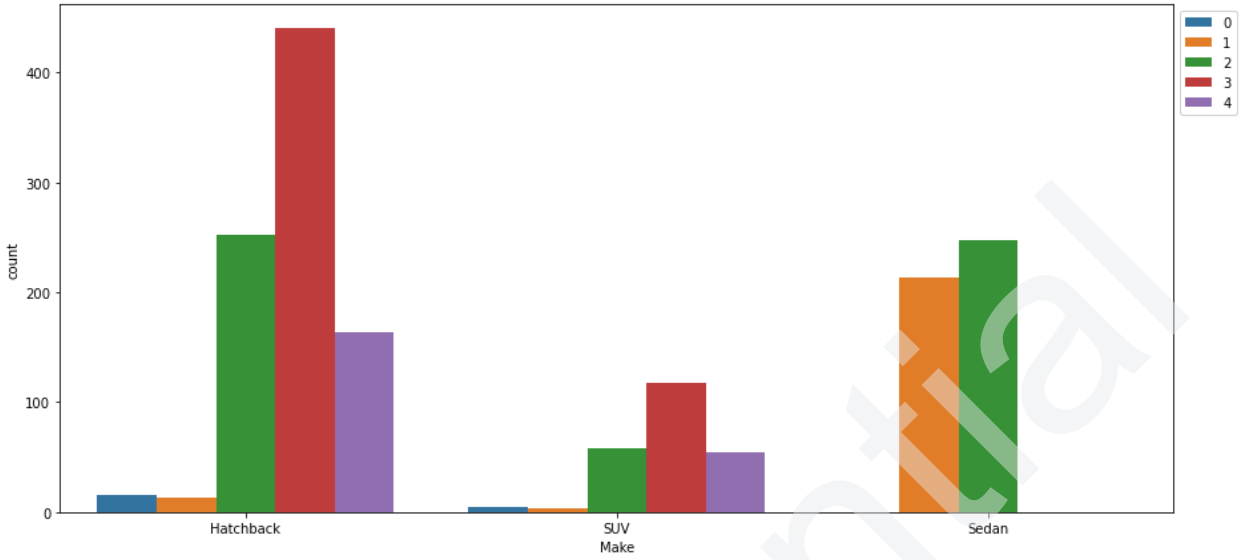


Figure 21: Make vs Number of dependents

Observations:

- Customers with 3 or more number of dependents are more likely to buy a Hatchback or SUV.
- Sedan cars are purchased by customers with 1 or 2 dependents.

- **Make vs Profession**

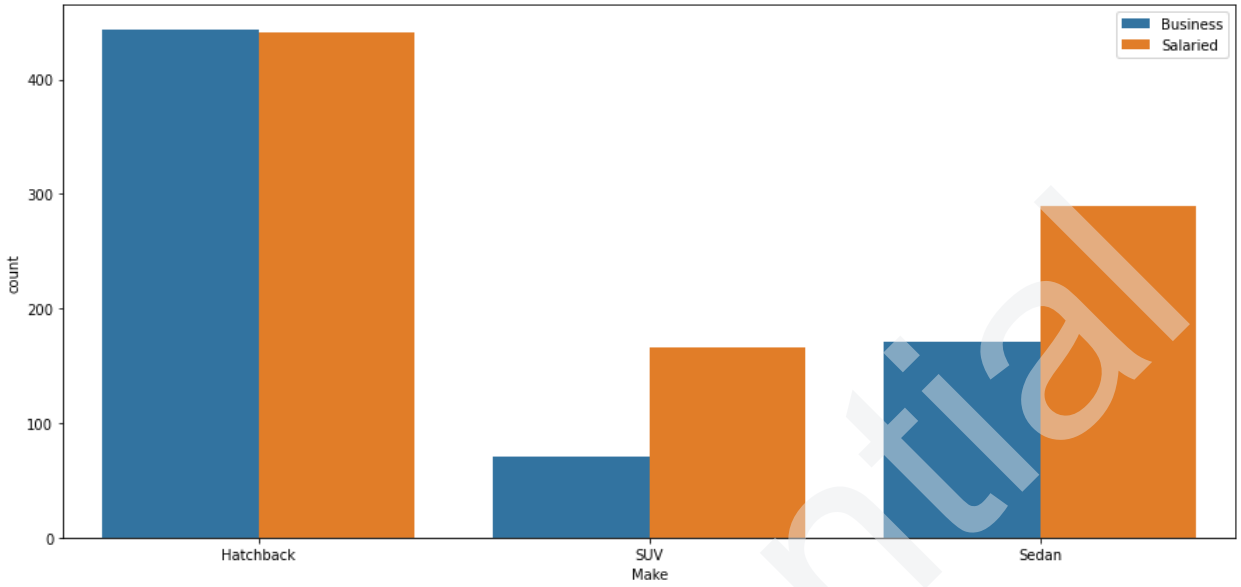


Figure 22:Make vs Profession

Observations:

- Customers with salaries buy more cars compared to customers who own their businesses.
- Sales of the hatchback are almost the same for Business and Salaried individuals. Hatchback is more popular in both the professions.

- **Make vs Personal loan**

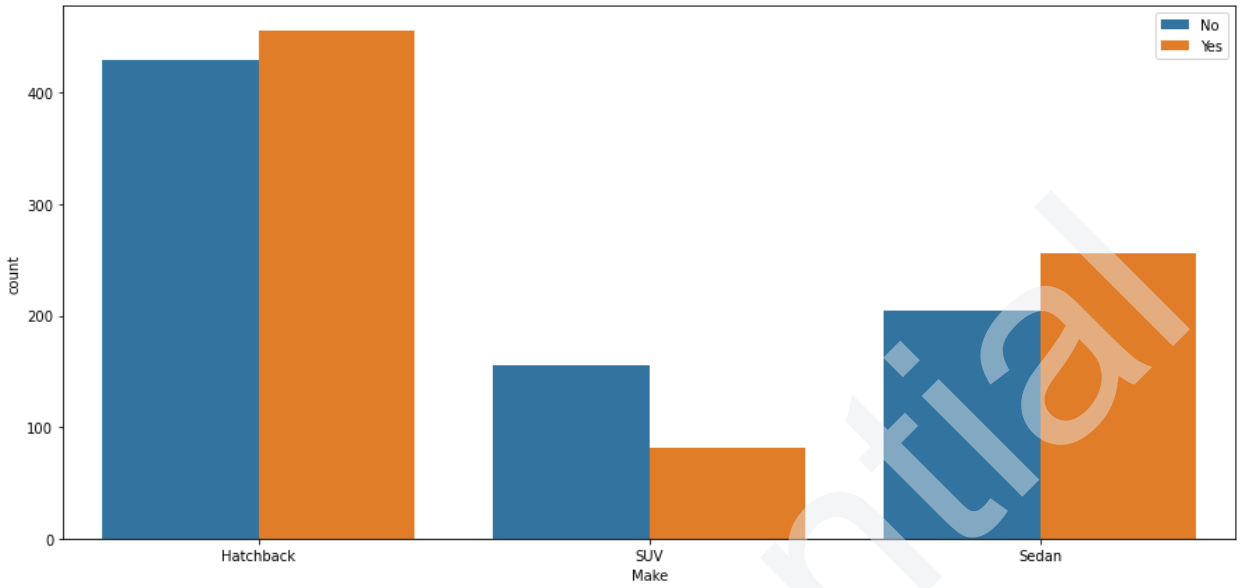


Figure 23: Make vs Personal loan

Observations:

- Few SUV customers have personal loans on them.
- For Hatchback and Sedan, there is equal distribution of personal loans among the customers.

- **Make vs House loan**

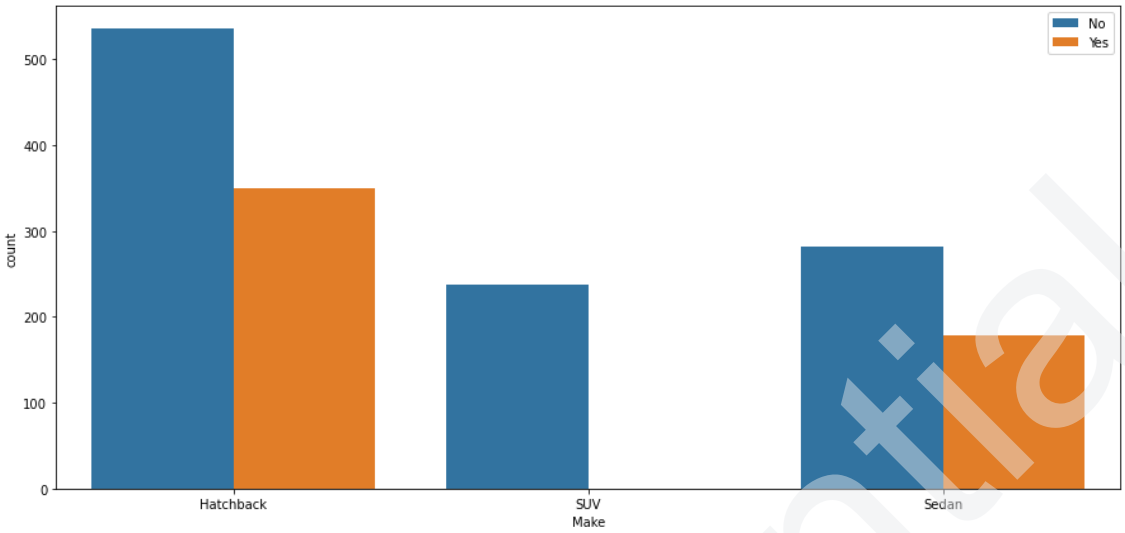


Figure 24: Make vs House loan

Observations:

- SUV customers do not have a house loan.
- More of the hatchback customers have a house loan compared to the Sedan customers.

- **Make vs Gender**

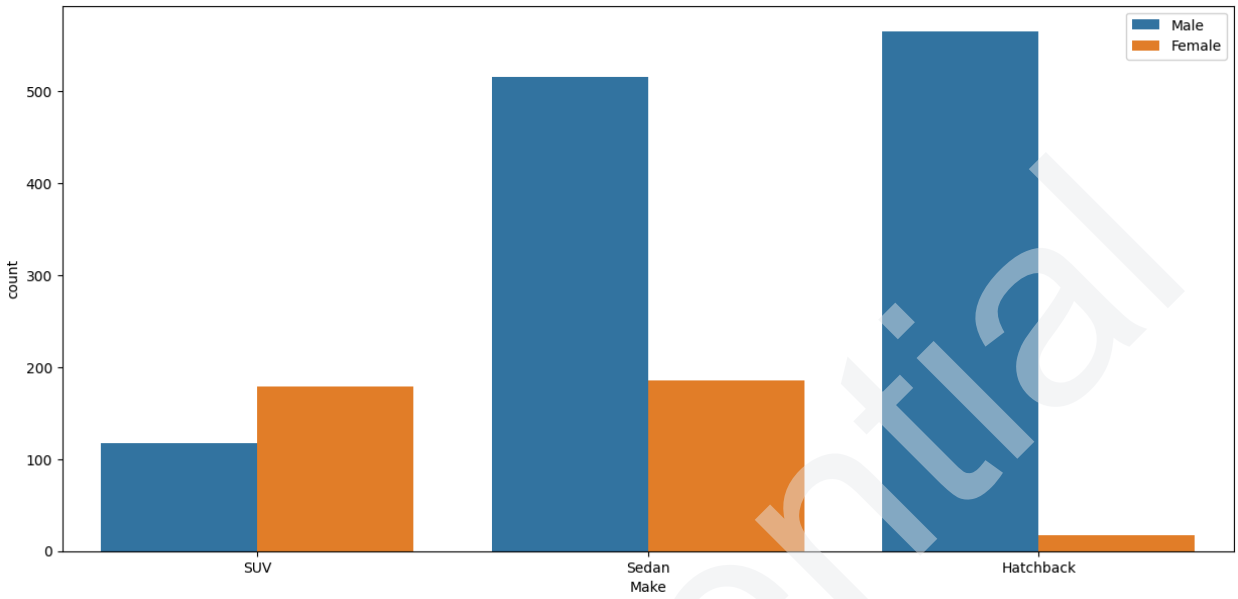


Figure 25: Make vs Gender

Observations:

- Females prefer SUV and are least likely to buy a Hatchback
- Males prefer Sedan or hatchback
- SUV is least preferable among males

- **Make vs Marital status**

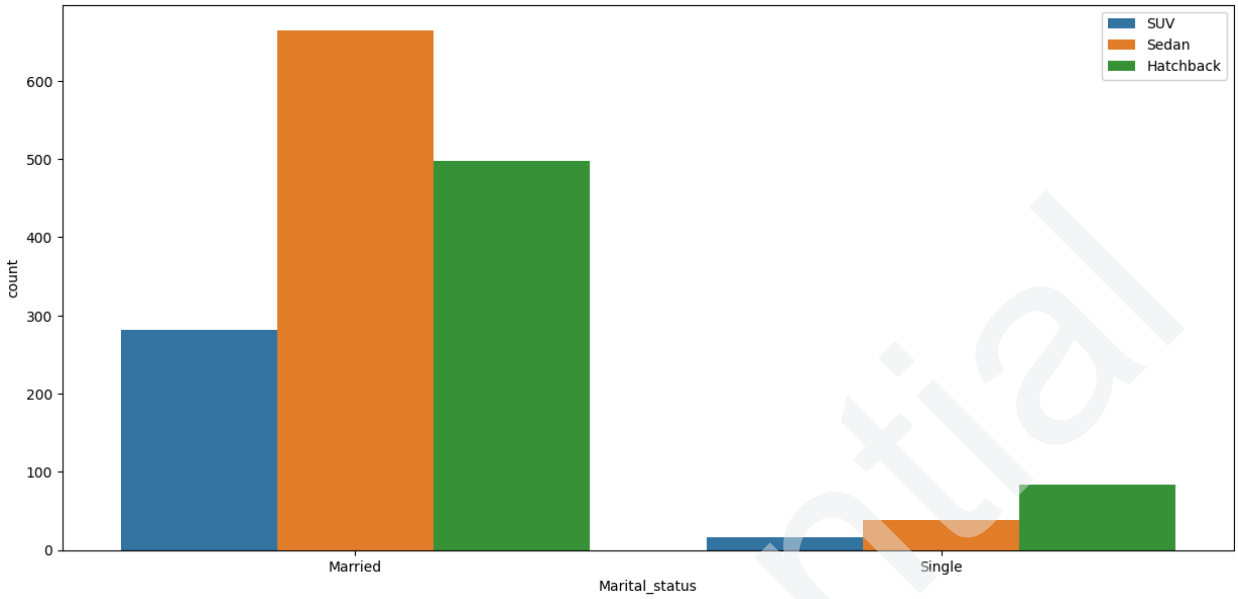


Figure 26: Make vs Marital status

Observations:

- Married person is most likely prefers a sedan and hatchback. Whereas, a single person most likely prefers a hatchback

Answer Key Questions

- **Do men tend to prefer SUVs more compared to women?**

Analyzing the ratio of SUV purchases for both Genders, we get

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

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

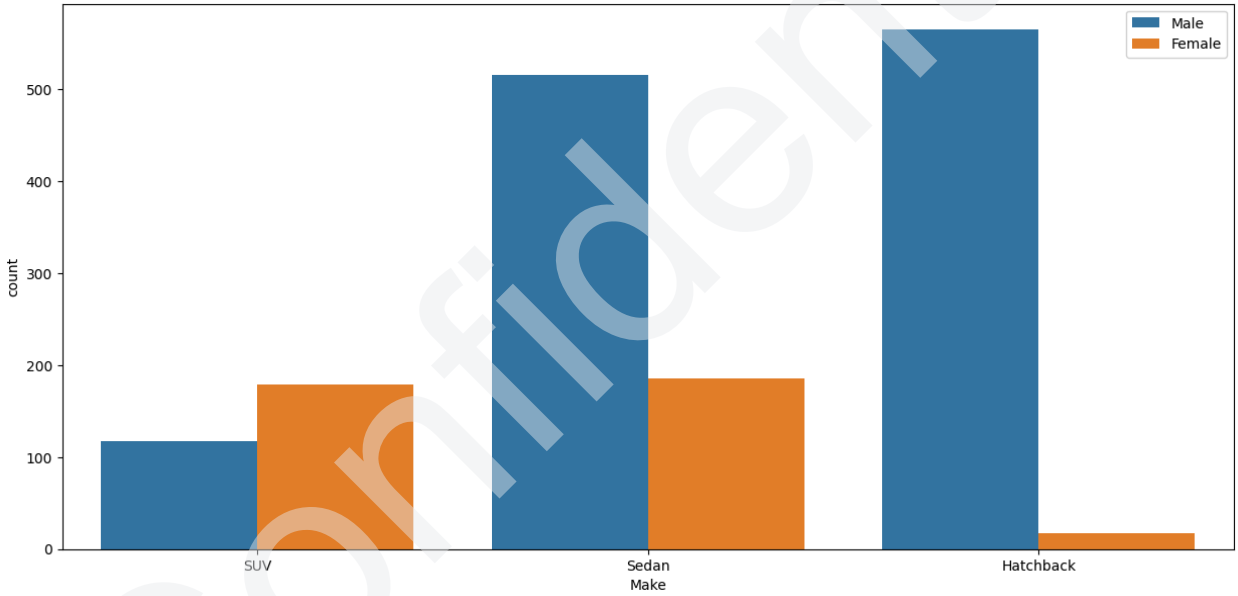


Figure 27: Make vs Gender

Hence the statement is incorrect.

- **What is the likelihood of a salaried person buying a Sedan?**

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

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

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

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

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

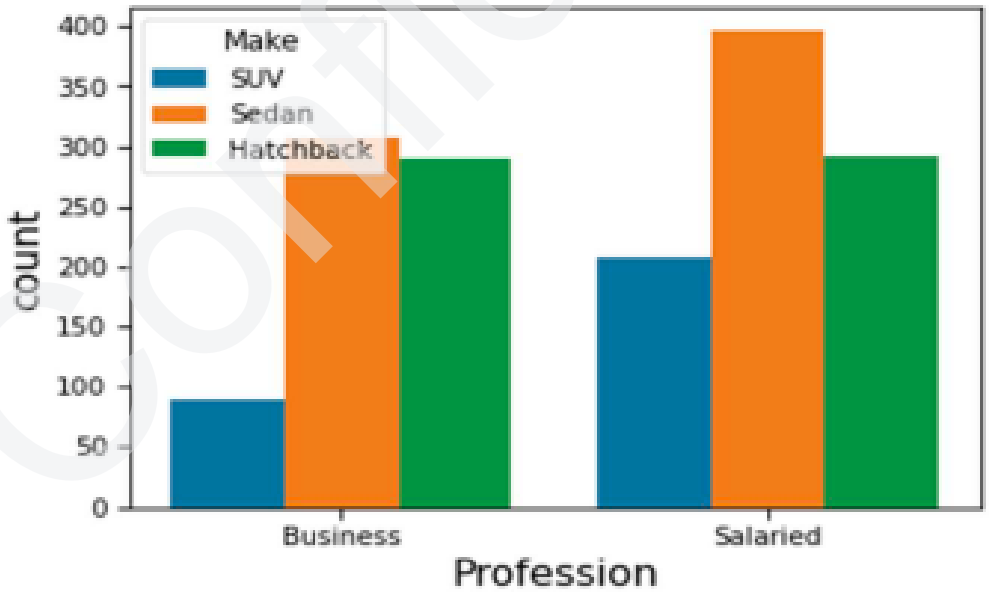


Figure 28: Make vs Profession

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

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

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

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

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

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

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

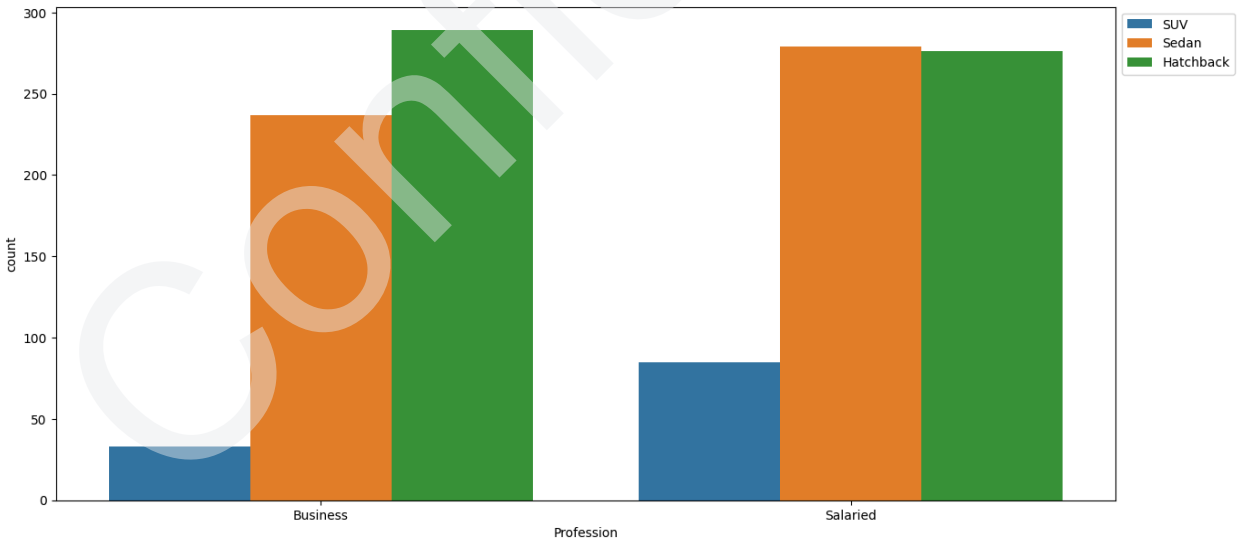


Figure 29: Make vs Profession (Male)

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

- **How does the amount spent on purchasing automobiles across gender?**

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

The mean of Price across Gender:

Female = 47705
Male = 32416

Median Price across Gender:

Female = 49000
Male = 29000

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

- **Is there a noticeable difference in the prices of purchases between customers without a Personal loan and those with a Personal Loan?**

The mean of Price across Personal Loan:

Personal Loan: No= 36742
Personal Loan: Yes= 34457

Median of Price across Personal Loan:

Personal Loan: No= 32000
Personal Loan: Yes= 31000

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

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

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

Mean of Price across Partner_working:

Partner_working: No = 36000
Partner_working: Yes = 35267

Median of Price across Partner_working:

Partner_working: No = 31000
Partner_working: Yes = 31000

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

Actionable Insights and Recommendations

Actionable Insights

Hatchback:

- An affordable and general-purpose car that can be used by a wide range of users.
- It can be considered as an entry-level car generally targeted at the younger population with an average income of 55k.

Sedan:

- Slightly costlier compared to hatchback-type cars
- The product also generally targets customers in their 30's who have a slightly higher income.
- The product is suitable for single customers.

SUV:

- A costly car that will excite the car-lovers
- It has a higher price point and is more suitable for customers who do not have any kind of loans on them.
- The buyers in this segment are elder and salaried individuals.

Business Recommendations

- Austo should first launch the affordable Hatchback model in the US market targeting the younger population. This car type can be the flagship product that brings in profits for the company as most of the young USA customers prefer this model.
- Then, Austo should launch a good and affordable Sedan model. The company needs to engage in more marketing for this model and

should try to lure the younger age group customers into buying this model.

- After the successful launch of these models, the company can launch the SUV model with a competitive pricing strategy to gain more profits from the US automobile market. SUVs can be targeted to people from the age group of 35 -60. As most of the customers for SUVs are in this age range.