

## Problem Statement - SMDM Project - Coded

### Problem Statement 1:

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

### Objective:

They want to analyze the data to get a fair idea about the demand of customers which will help them in enhancing their customer experience. Suppose you are a Data Scientist at the company and the Data Science team has shared some of the key questions that need to be answered. Perform the data analysis to find answers to these questions that will help the company to improve the business.

### Data Description:

- **age:** The age of the individual in years.
- **gender:** The gender of the individual, categorized as male or female.
- **profession:** The occupation or profession of the individual.
- **marital\_status:** The marital status of the individual, such as married &, single
- **education:** The educational qualification of the individual Graduate and Post Graduate
- **no\_of\_dependents:** The number of dependents (e.g., children, elderly parents) that the individual supports financially.
- **personal\_loan:** A binary variable indicating whether the individual has taken a personal loan "Yes" or "No"
- **house\_loan:** A binary variable indicating whether the individual has taken a housing loan "Yes" or "No"
- **partner\_working:** A binary variable indicating whether the individual's partner is employed "Yes" or "No"
- **salary:** The individual's salary or income.
- **partner\_salary:** The salary or income of the individual's partner, if applicable.
- **Total\_salary:** The total combined salary of the individual and their partner (if applicable).
- **price:** The price of a product or service.
- **make:** The type of automobile

### Problem 1 - Data Overview

#### Observations and Insights:

- **Structure of the data:** The provided dataset has 1581 rows with 14 columns.
- **The types of the data are as follows:**

```

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

```

- **Missing values:** There are missing values for two columns i.e., **gender** and **partner\_salary** and these cannot be treated as can't predict gender and partner salary could be empty if they are not working or no fixed salary to be mentioned and total salary could be based on the spouse/partner salary.

```

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

```

~

Missing Values After Treatment:

```

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

```

### Statistical summary & Observations and Insights:

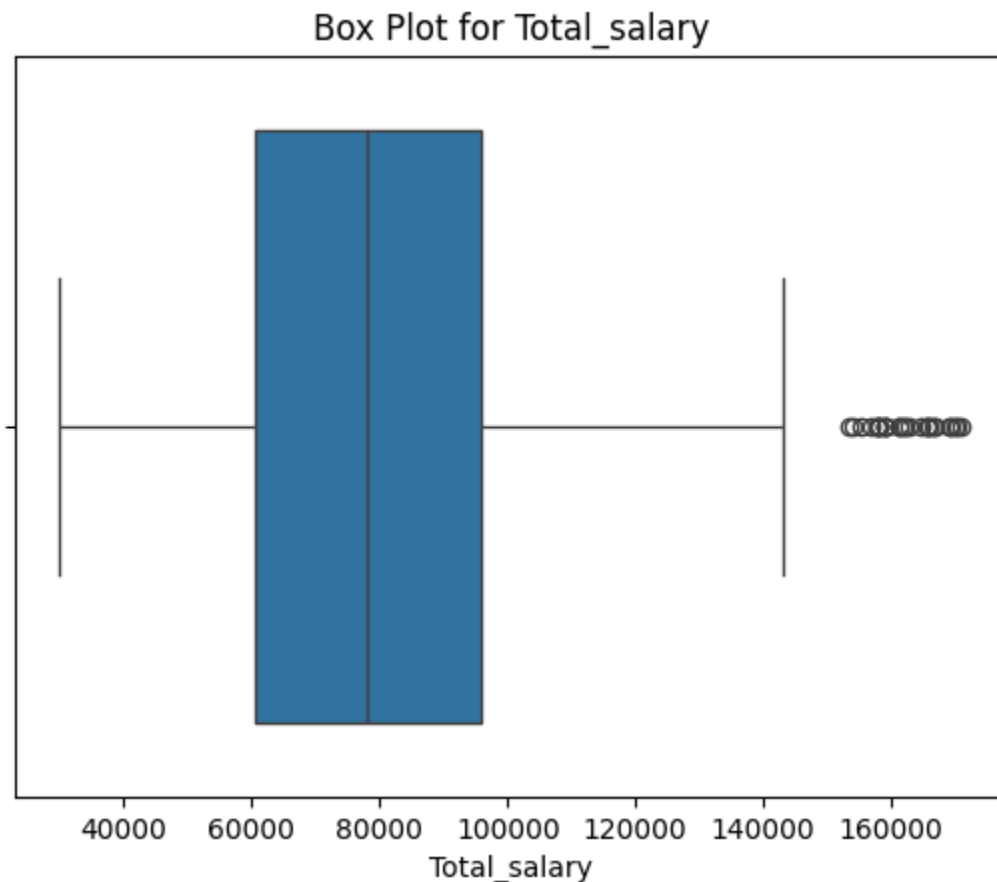
- **Age:** The youngest person is 22 years old, and the oldest is 54 years old.
- **No\_of\_Dependents:** On average, there are around 2.46 dependents & max being 4.
- **Salary:** The average salary is roughly 60,392.22 & The minimum salary is 30,000, and the maximum is 99,300.
- **Partner\_salary:** total 106 missing values found which is considered as min i.e., 0 & max is 80,500.
- **Total\_salary:** The average total salary is around 62,625.99.
- **Price:** Average product/ service price is : 35,597.72, min: 18,000 & max : 70,000

### Problem 1 - Univariate Analysis

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

### Numerical columns ['Age', 'No of Dependents', 'Salary', 'Total salary', 'Price']:

- No other numeric/ int64 has outliers other than Total\_Salary



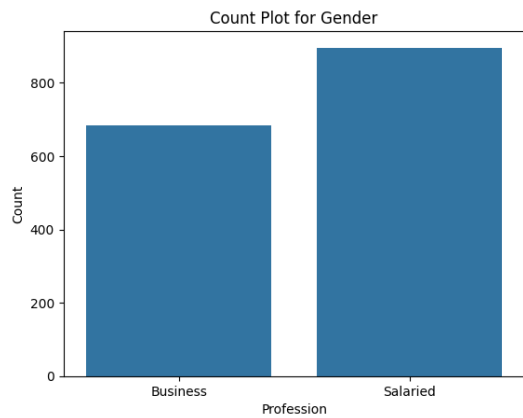
### Gender:

- Some incorrect gender values identified hence treated them by using fillna method and mean().
- Most of the customers i.e., **~76% from Male** gender category

```
Male      1252
Female     329
Name: Gender, dtype: int64
```

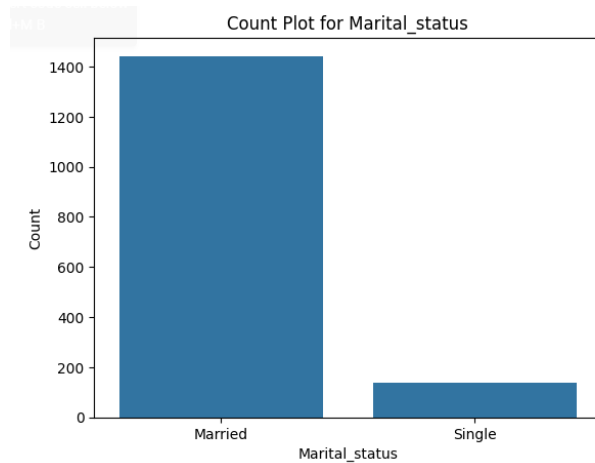
### **Profession:**

- 56.7% are salaried and the remaining are into business.



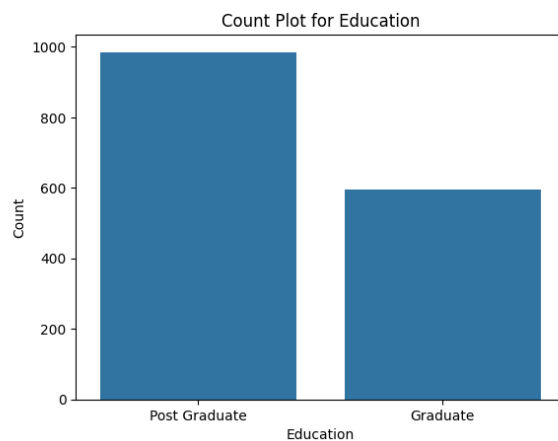
### **Marital status:**

- 91% are Married and the remaining states single.



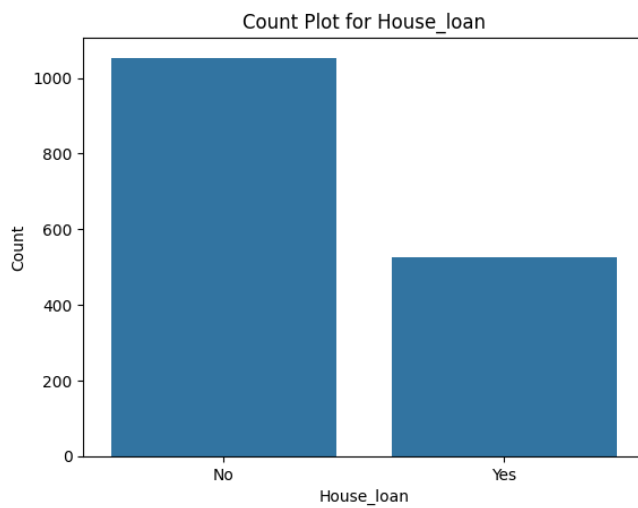
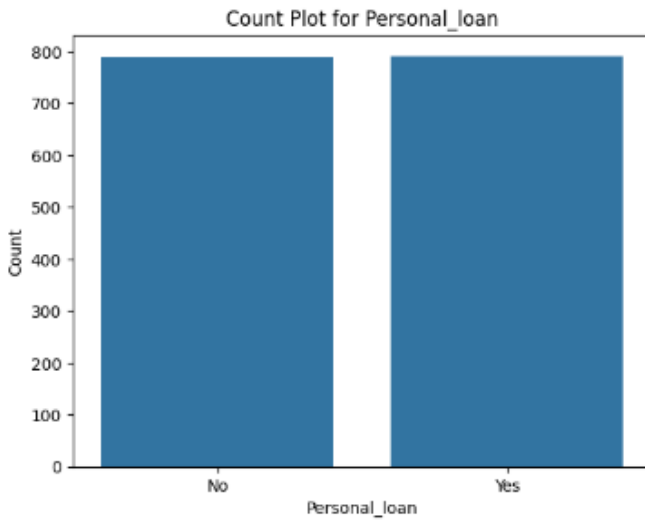
### **Education:**

- All the customers are educated in which 62% are post graduates and 37% garduates.

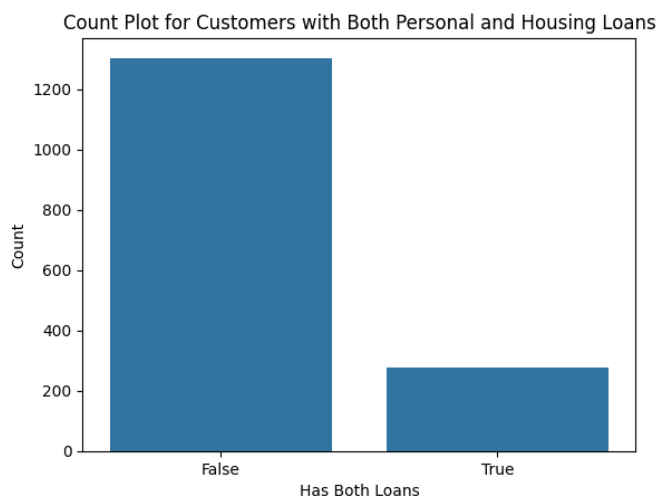


### **Personal loan & House loan:**

- 50% of the customers are having personal loan and 66% are having a housing loan

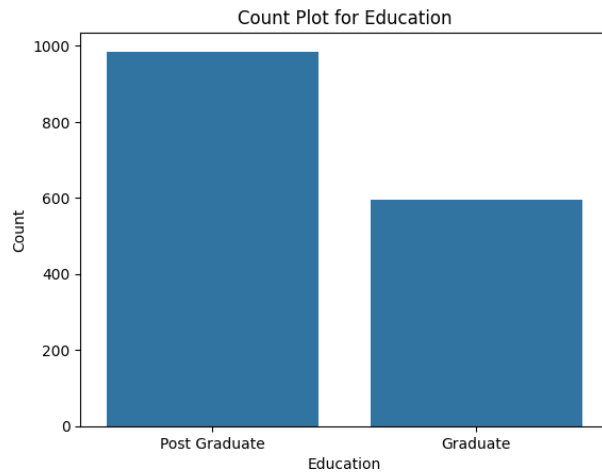


- However, ~18% of the customers are having both the loans which is 278 customers



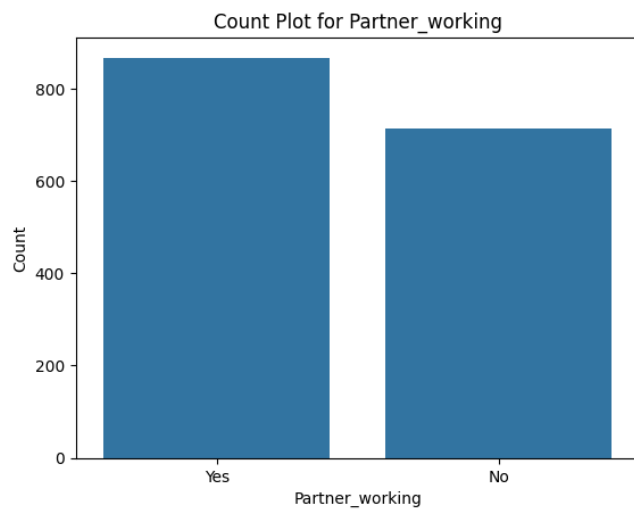
### **Education:**

- All the customers are educated in which 62% are post graduates and 37% graduates.



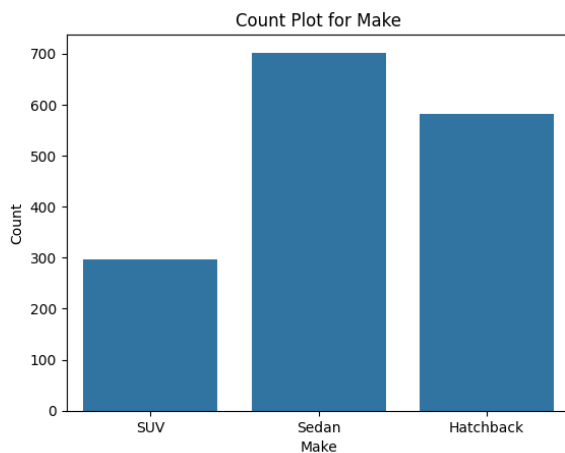
**Partner working:**

- 54% of the customers partners are working i.e., 868.



**Make:**

- There are 3 category car Types: Sedan, Hatchback & SUV and proportion is as follows:



```
Sedan      702
Hatchback  582
SUV        297
Name: Make, dtype: int64
Sedan      44.402277
Hatchback  36.812144
SUV        18.785579
Name: Make, dtype: float64
```

## 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

### - Explore the relationship between all numerical variables:

#### Age vs. No of Dependents:

- The scatter plots show a uniform/diverse distribution, and no clear linear trend is observed between Age and No of Dependents.

#### Age vs. Salary:

- There is a positive correlation between Age and Salary, suggesting that, in general, as age increases, the Salary also increase.

#### Salary vs. Partner salary:

- No positive correlation is seen.

#### Total salary vs. Price:

- It's a left-skewed scatter plot between Total salary & Price and the tail of the distribution extends towards the left.

#### No of Dependents vs. Total salary:

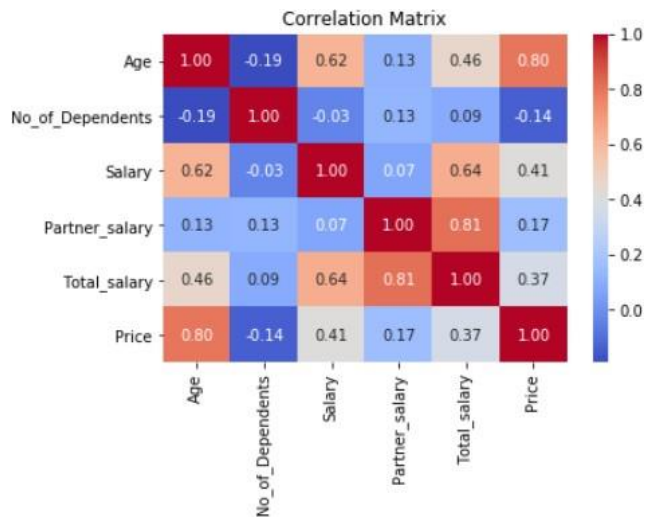
- We can see a somewhat uniform relationship between these two variables.



### No of Dependents vs. Price:

- The scatter plots show varying/equally distributed patterns, and no conclusive relationship is evident between the number of dependents and the price of items.

### - Explore the correlation between all numerical variables:



### Insights from the above correlation Matrix:

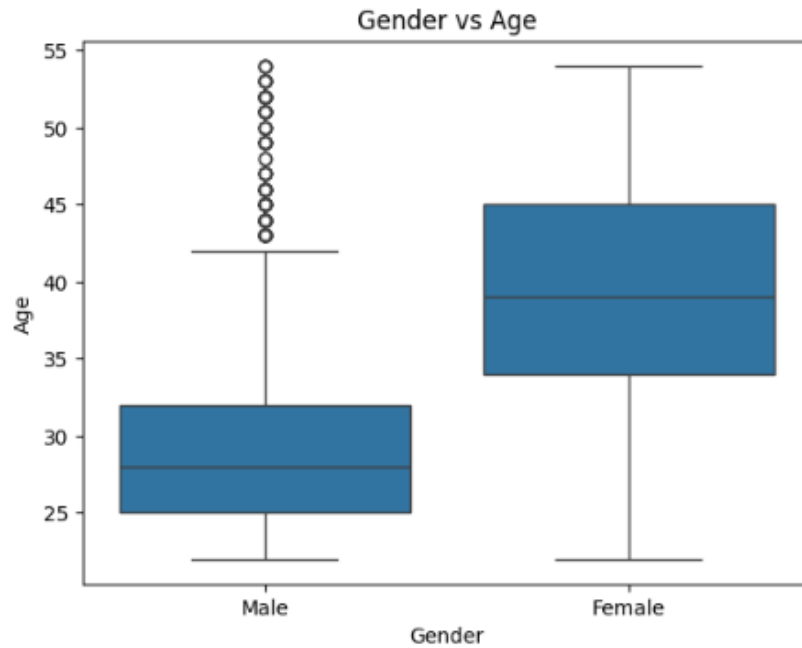
- We can see a Strong/highest positive correlation between 'Age' & 'Price' i.e., **0.80** and 'Total salary' & 'Partner Salary' i.e., **0.81**
- 'Total salary' and 'Price' have a moderate positive correlation i.e., **0.37**.
- Highest negative correlation is seen between No of Dependents & Age i.e., **-0.19** and No of Dependents & Price i.e., **-0.14**
- A weak or no linear correlation is seen between Salary & No of Dependents i.e., **-0.03**

### - Explore the relationship between categorical vs numerical variables:

### Insights about the Relationship between Categorical and Numerical Variables

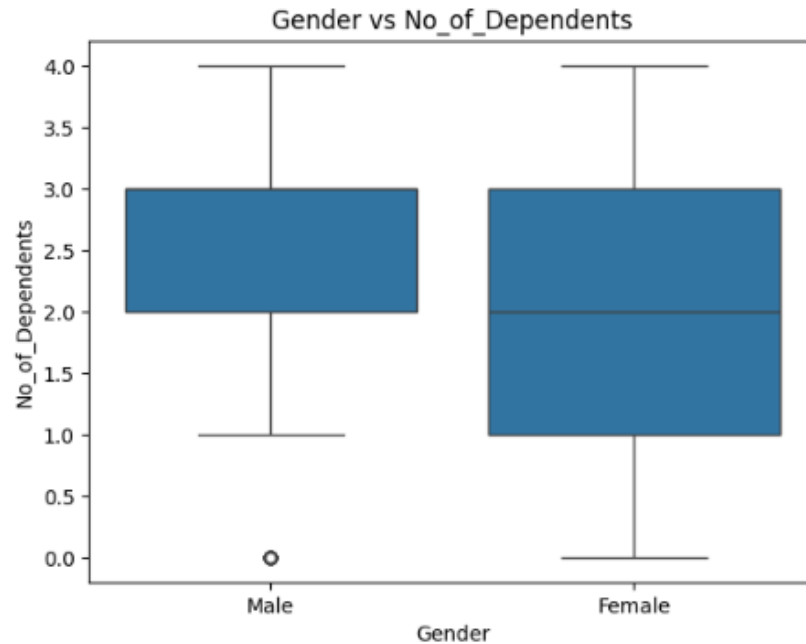
#### Gender vs. Age:

- Females are having a slightly higher median age than males.
- Male median age is around 27 whereas for females it is between 37 & 40.
- Outliers are seen in male age distribution.



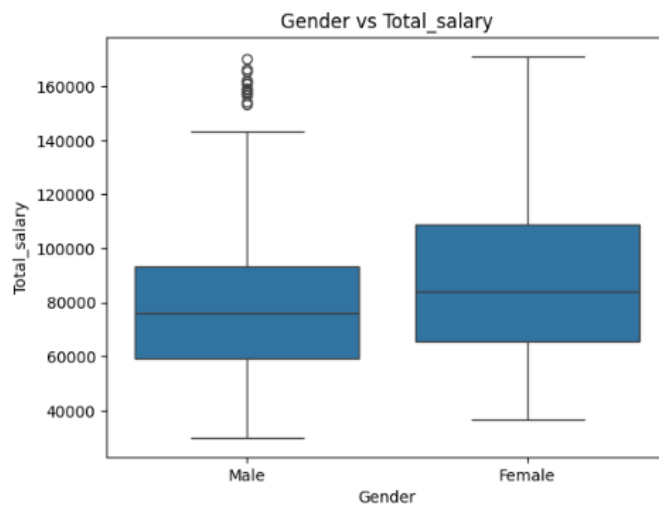
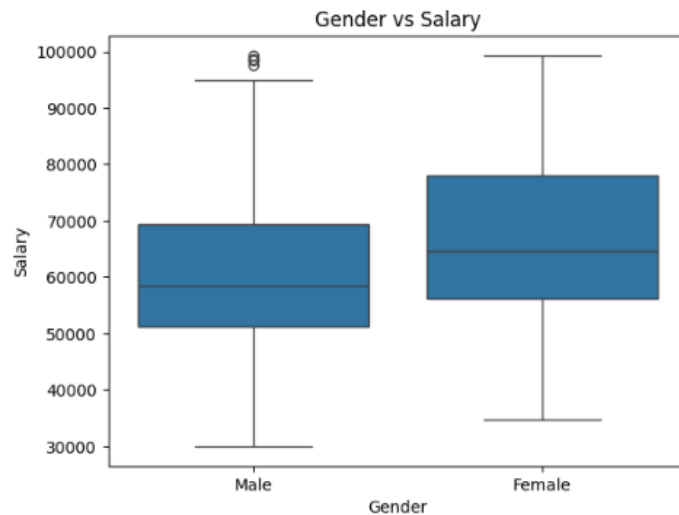
#### Gender vs. No of Dependents:

- Male and female's maximum no of dependents count are exactly same. And some female's min no of dependents are 0



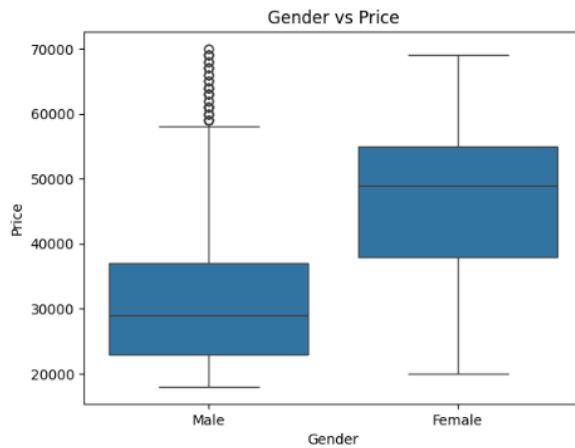
### Gender vs. Salary/ Total Salary:

- Outliers are seen in male category vs Salary/ Total Salary. Overall male salaries are lower than female salaries.
- So, we can prefer targeting female customers for any service or product campaigns if applicable.



### Gender vs. Price:

- Outliers are seen in male category.
- Male median price is below 30K and female's is close to 50K. It proves that females are more likely spending huge amounts on product or services.



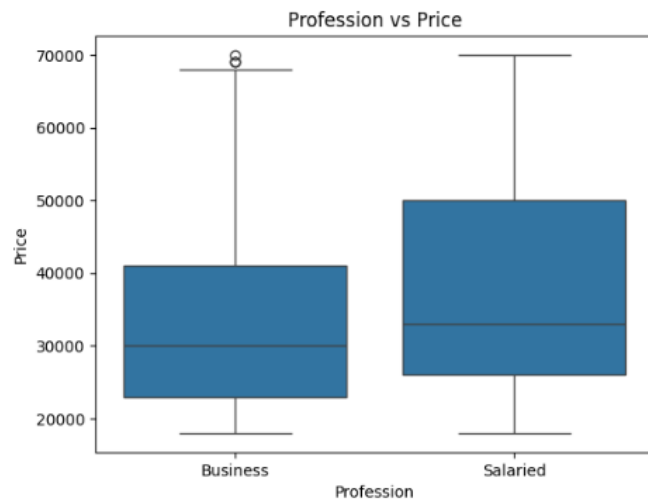
### Profession vs. partner salary:

- Irrespective of the professions, salaries are equally aligned no much difference is seen. So, our target customers can be an of the profession category.



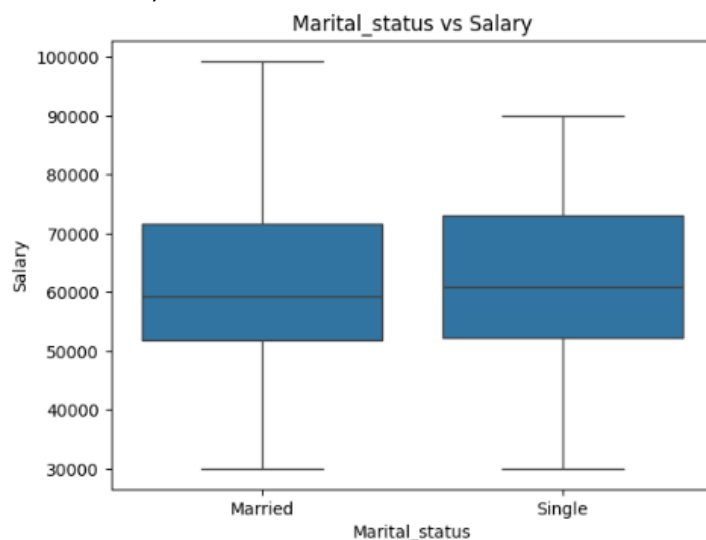
### Profession vs. Price:

- Salaried customers are more likely to spend more compared to business. Slight outlier part is in business category.



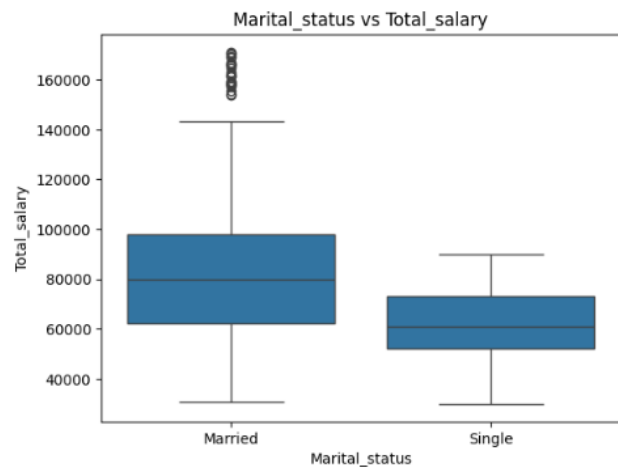
### Marital Status vs. Salary:

- Maximum salary of married customers is close to 100,000 which is higher than singles i.e., around 85K.



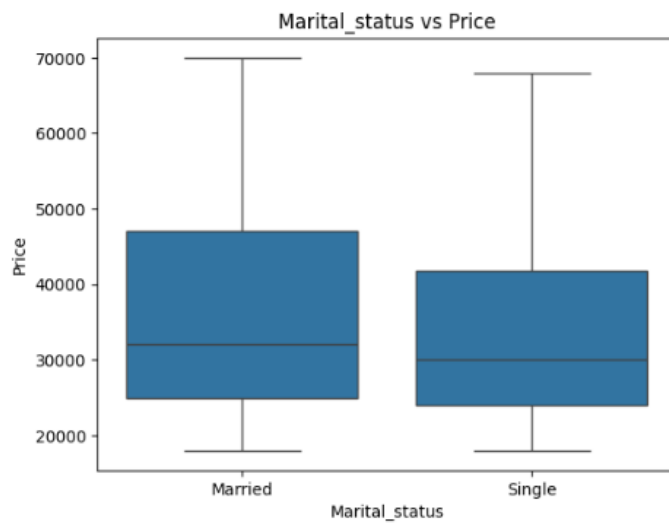
### Marital status vs. Total salary/Price:

- Married individuals have a different salary distribution compared to those who are single as married total salary is very high (max: 1,40,000) compared to singles (Max: ~90,000) (Outliers are seen in married total salary data).
- Due to the above reason, there more likely to spend more and same is seen in the below boxplot.



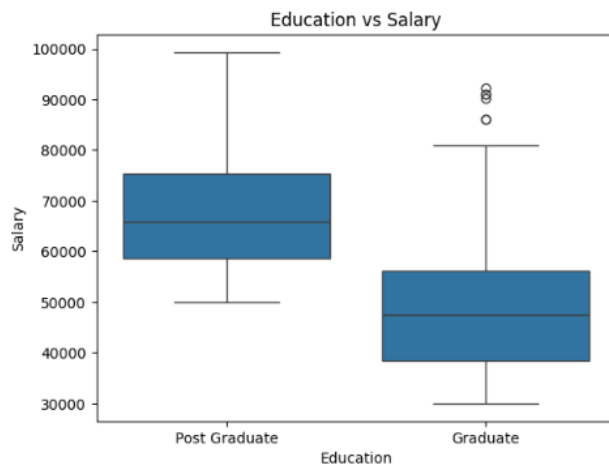
### Marital status vs. Price:

- Married individuals have a different salary distribution compared to those who are single as married total salary is very high (max: 1,40,000) compared to singles (Max: ~90,000)

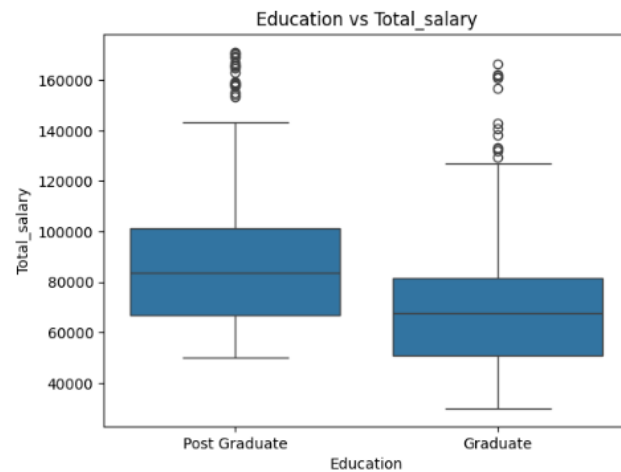


### Education vs. Salary/ Total Salary:

- Postgraduates (~65,000) are holding higher salaries compared to graduates (~45,000) total salaries, with individuals with higher education levels generally having higher total salaries.

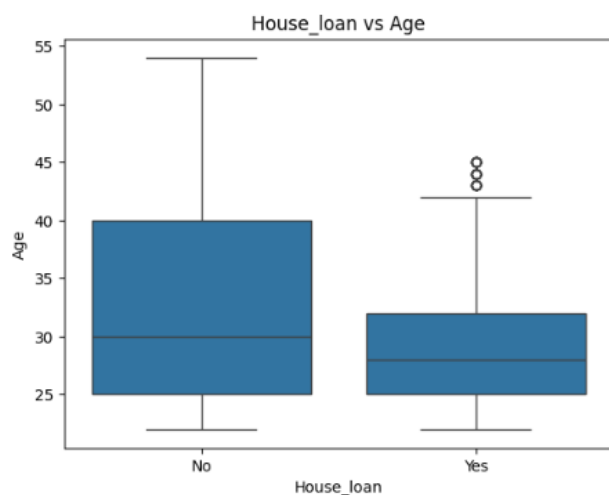


- Similar trend is seen in education vs. total salary with outliers in both education categories.



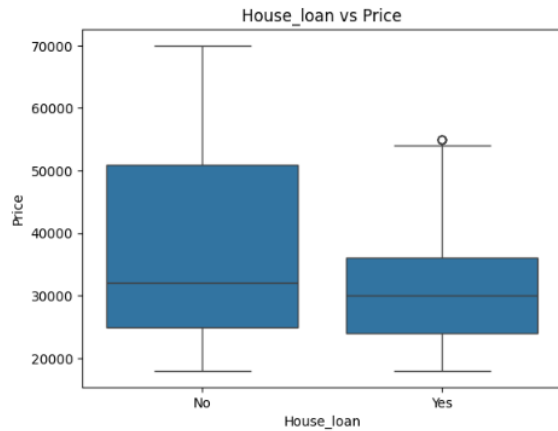
### House Loan vs. Age:

- Looks like younger customers are having the housing loan compared to age 43 and above customers.



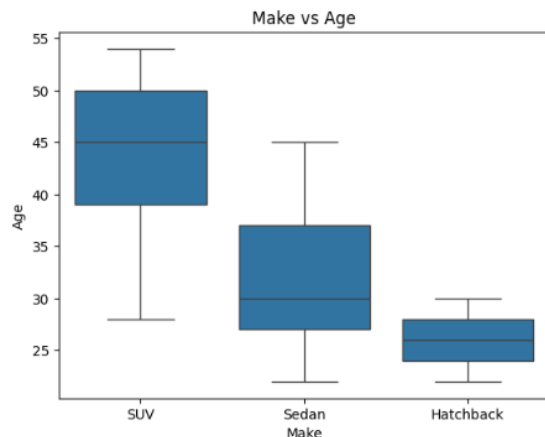
### House/personal loan vs. Price:

- Customers who are not having housing/personal loans are spending more on our products/services compared to those who doesn't have loans (as they may have different spending patterns compared to those without loans).



### Make vs. Age:

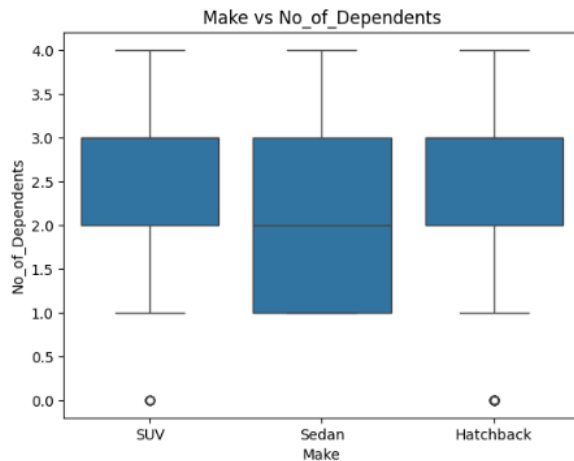
- Looks like age group between 25 & 30 are preferring hatchbacks, who are 30's and 40's preferring Sedan's, the higher age group i.e., 40 and above are more preferable in SUV model.





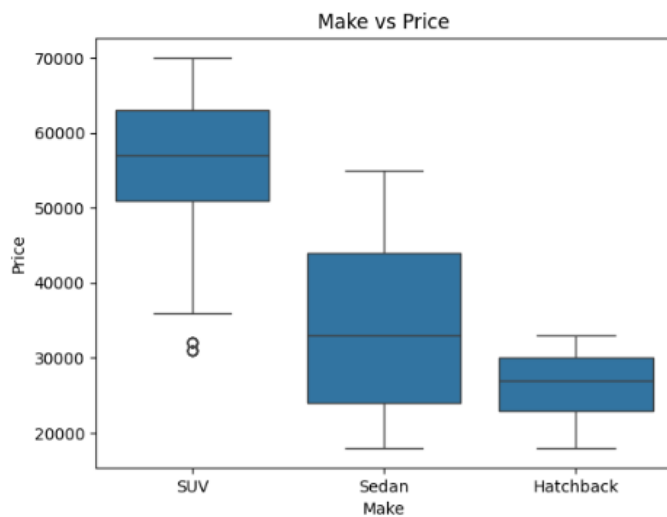
### Make vs. No of Dependents:

- As overall dataset has maximum no of dependents as 4, In general, all our make models (SUV, Sedan, Hatchbacks) are quite good for 4 number families.
- And boxplot talks about the same.



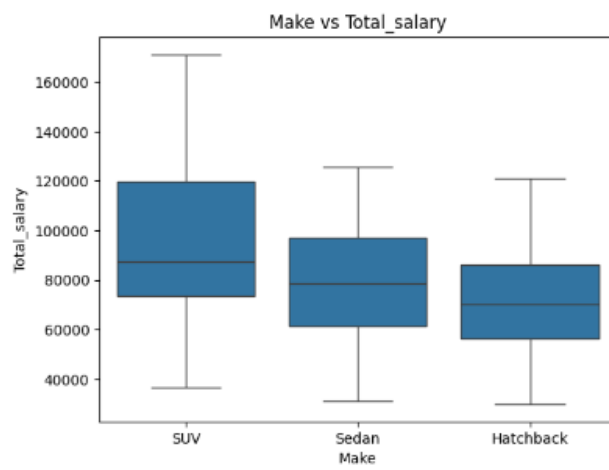
### Make vs. Price:

- SUV model holders are spending max: 70,000 on services, sedan holders are spending around 55,000 and hatch back holders are spending around 30,000.



### Make vs. Salary/Total salary:

- High salaried (**Med: ~75,000**) customers are preferred SUV model and medium salaried (Med: ~ 60,000) are preferring Sedan and rest of the customers likely less salaried (Med: ~55,000) are preferring hatchbacks as budget friendly.

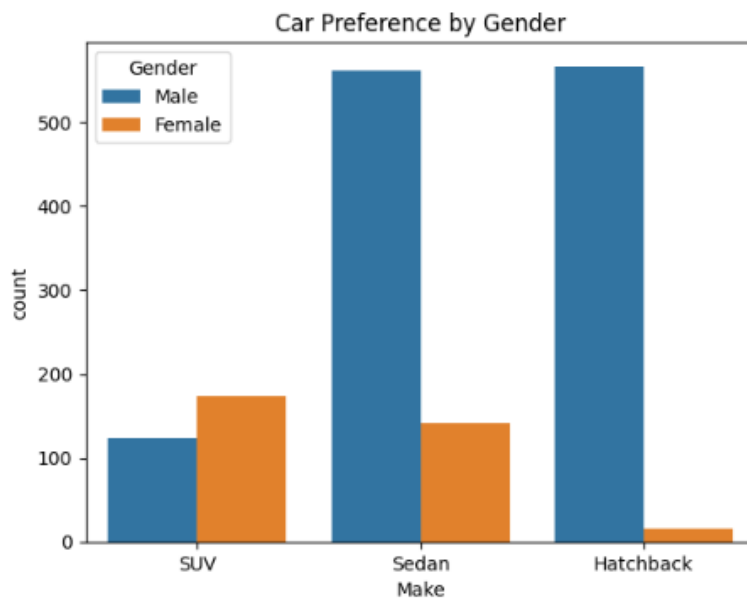


## Problem 1 - Key Questions

Explore the data to answer the following key questions:

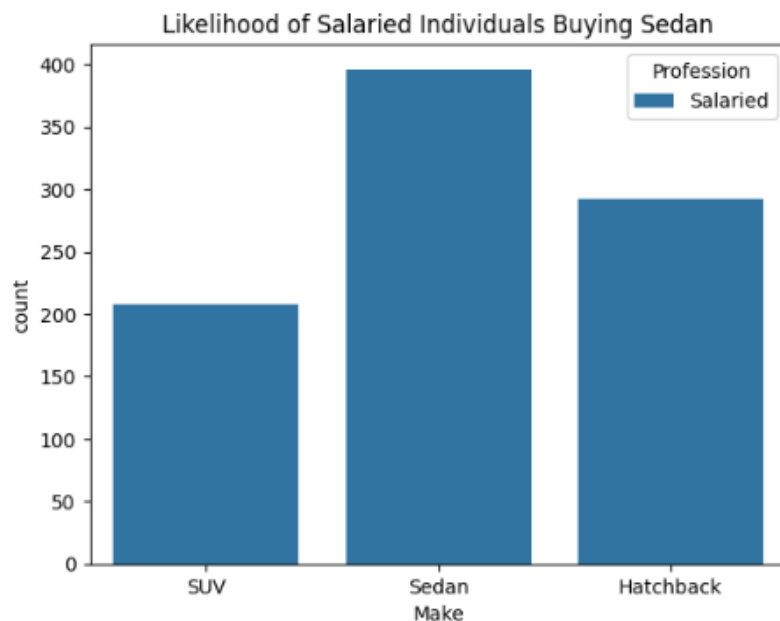
### 1. Do men tend to prefer SUVs more compared to women?

- No, as we see in the below plot Women are more interested in SUV compared to Men.



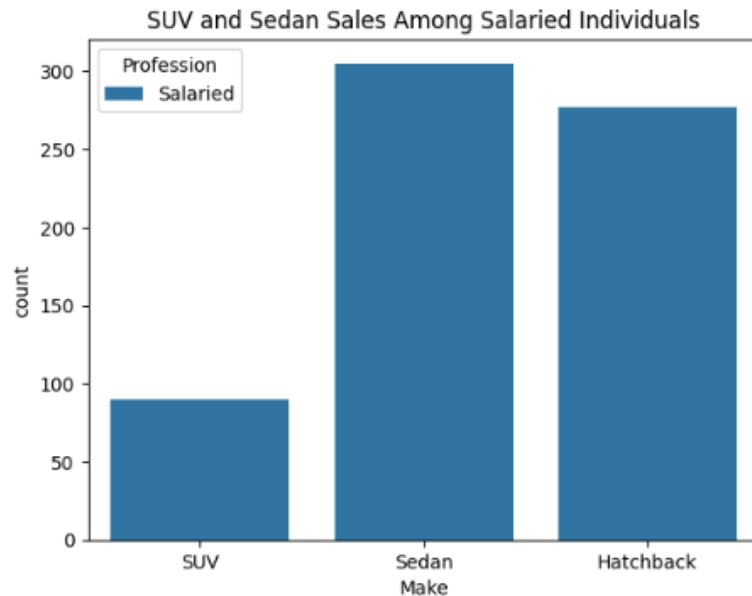
### 2. What is the likelihood of a salaried person buying a Sedan?

- More Salaried customers (~ 400 customers) are preferring sedan model compared to other category models.



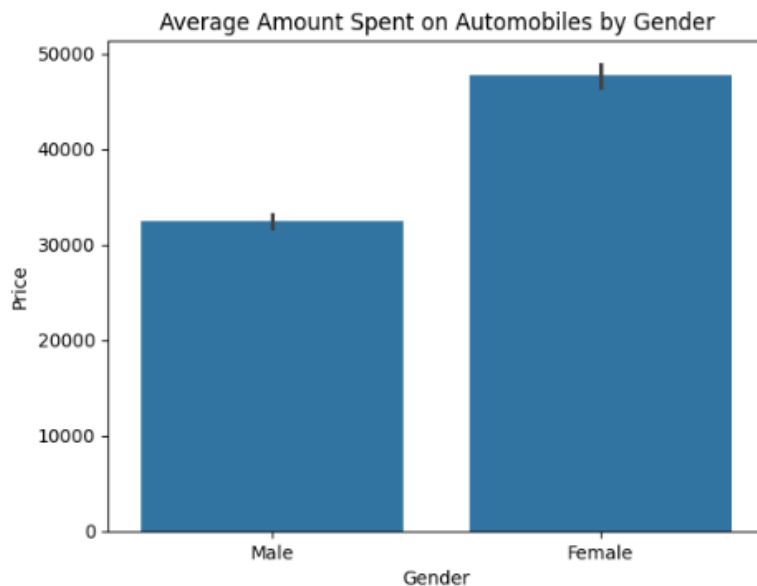
### 3. What evidence or data supports Sheldon Cooper's claim that a salaried male is an easier target for a SUV sale over a Sedan sale?

- Examining the distribution of male salaried customers between SUVs and Sedans (Fig mentioned below) reveals that over 250 individuals prefer Sedans, while approximately 80 customers opt for SUVs.
- These numbers suggest that Sheldon Cooper's claim aligns with the observed data.



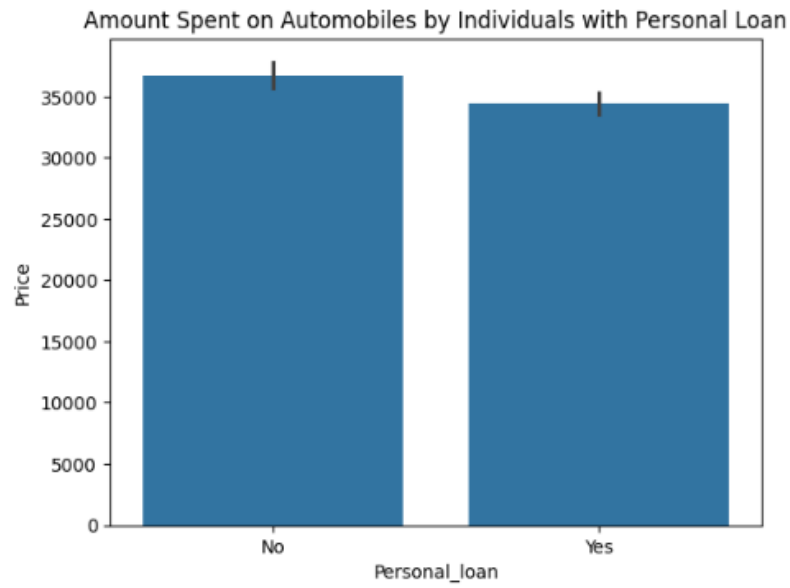
#### 4. How does the amount spent on purchasing automobiles vary by gender?

- As per the given data, women purchasing behavior (**Max spend: ~50K**) on automobiles is much higher compared to men (max spend: 30K+).



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

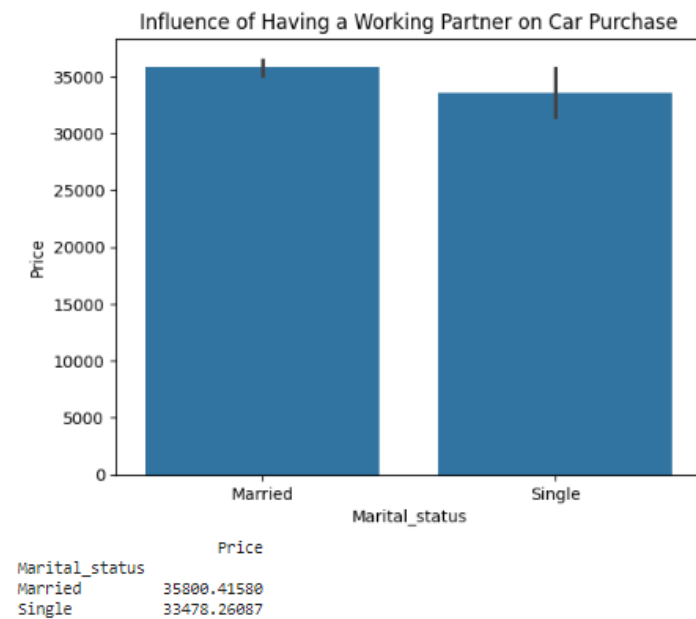
- Amount **34,457** was spent on automobiles by individuals who took a personal loan.



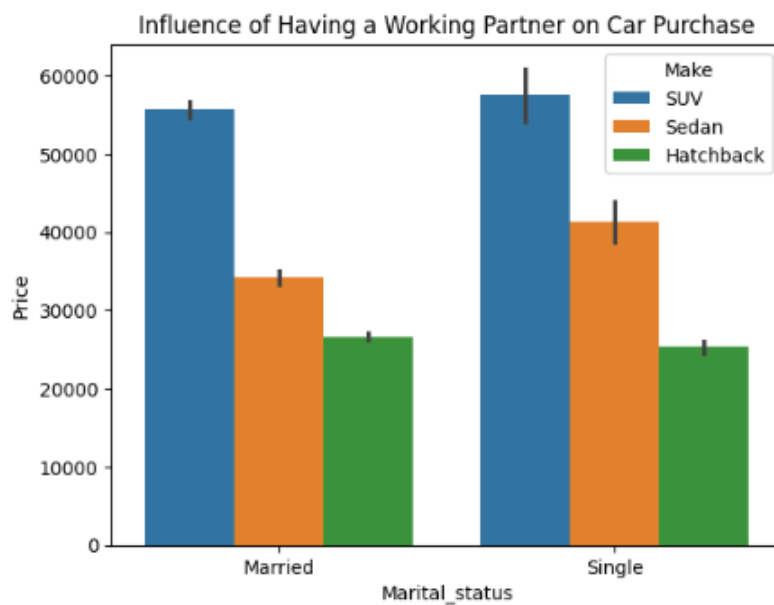
Personal_loan	Price
No	36742.712294
Yes	34457.070707

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

- Although there isn't a significant difference observed, married customers tend to spend slightly more than singles.



- However, when considering that SUVs are the top-end models compared to hatchbacks and sedans, singles are more likely to purchase higher-priced cars.



## Problem 1 - Actionable Insights & Recommendations

### - Actionable Insights - Business Recommendations

**1. Correct Gender Entries:**

Validated and corrected gender entries which has spelling mistakes/incorrect entries, but we need to ensure accurate data categorization by filling 53 missing values Nan.

**2. Target Females Strategically:**

Design targeted campaigns for females, capitalizing on their higher spending tendencies.

**3. Highlight Sedans for Salaried Customers:**

Tailor marketing to showcase Sedans, preferred by salaried individuals, especially females.

**4. Promote SUVs to Salaried Males:**

Support Sheldon Cooper's claim by focusing SUV promotions on male salaried customers.

**5. Understand Personal Loan Taker Preferences:**

Explore preferences of customers with personal loans to customize marketing and financing options.

**6. Attract Singles to Higher-Priced Cars:**

Develop campaigns for singles, emphasizing features of higher-priced models, particularly SUVs.

**7. Optimize Marketing for Married Customers:**

Target married individuals with family-oriented promotions and potential package deals.

**8. Promote SUVs to High-Salaried Individuals:**

Emphasize SUV features for high-salaried customers, leveraging their preferences.

**9. Monitor and Address Outliers:**

Regularly investigate outliers in salary and total salary data to ensure accuracy.

## 10. Enhance Customer Segmentation:

Differentiate marketing strategies based on demographics to maximize engagement and conversion rates.

### Problem 2 Context

A bank generates revenue through interest, transaction fees, and financial advice, with interest charged on customer loans being a significant source of profits. GODIGT Bank, a mid-sized private bank, offers various banking products and cross-sells asset products to existing customers through different communication methods. However, the bank is facing high credit card attrition, leading them to reevaluate their credit card policy to ensure customers receive the right card for higher spending and intent, resulting in profitable relationships.

### Objective

As a Data Scientist at the company and the Data Science team has shared some data. You are supposed to find the key variables that have a vital impact on the analysis which will help the company to improve the business.

### Data Description

- **userid** - Unique bank customer-id
- card\_no** - Masked credit card number
- card\_bin\_no** - Credit card IIN number
- Issuer** - Card network issuer
- card\_type** - Credit card type
- card\_source\_data** - Credit card sourcing date
- high\_networth** - Customer category based on their net-worth value (A: High to E: Low)
- active\_30** - Savings/Current/Salary etc. account activity in last 30 days
- active\_60** - Savings/Current/Salary etc. account activity in last 60 days
- active\_90** - Savings/Current/Salary etc. account activity in last 90 days
- cc\_active30** - Credit Card activity in the last 30 days
- cc\_active60** - Credit Card activity in the last 60 days
- cc\_active90** - Credit Card activity in the last 90 days
- hotlist\_flag** - Whether card is hot-listed(Any problem noted on the card)
- widget\_products** - Number of convenience products customer holds (dc, cc, net-banking active, mobile banking active, wallet active, etc.)
- engagement\_products** - Number of investment/loan products the customer holds (FD, RD, Personal loan, auto loan)
- annual\_income\_at\_source** - Annual income recorded in the credit card application
- other\_bank\_cc\_holding** - Whether the customer holds another bank credit card
- bank\_vintage** - Vintage with the bank (in months) as on Tthmonth



**T+1\_month\_activity** - Whether customer uses credit card in T+1 month (future)  
**T+2\_month\_activity** - Whether customer uses credit card in T+2 month (future)  
**T+3\_month\_activity** - Whether customer uses credit card in T+3 month (future)  
**T+6\_month\_activity** - Whether customer uses credit card in T+6 month (future)  
**T+12\_month\_activity** - Whether customer uses credit card in T+12 month (future)  
**Transactor\_revolver** - Revolver: Customer who carries balances over from one month to the next. Transactor: Customer who pays off their balances in full every month.  
**avg\_spends\_l3m** - Average credit card spends in last 3 months  
**Occupation\_at\_source** - Occupation recorded at the time of credit card application  
**cc\_limit** - Current credit card limit

## Problem 2 - Framing Analytics Problem

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

```

8448
Index(['userid', 'card_no', 'card_bin_no', 'Issuer', 'card_type',
      'card_source_date', 'high_networth', 'active_30', 'active_60',
      'active_90', 'cc_active30', 'cc_active60', 'cc_active90',
      'hotlist_flag', 'widget_products', 'engagement_products',
      'annual_income_at_source', 'other_bank_cc_holding', 'bank_vintage',
      'T+1_month_activity', 'T+2_month_activity', 'T+3_month_activity',
      'T+6_month_activity', 'T+12_month_activity', 'Transactor_revolver',
      'avg_spends_l3m', 'Occupation_at_source', 'cc_limit'],
      dtype='object')
userid                int64
card_no              object
card_bin_no          int64
Issuer               object
card_type            object
card_source_date      datetime64[ns]
high_networth         object
active_30             int64
active_60             int64
active_90             int64
cc_active30           int64
cc_active60           int64
cc_active90           int64
hotlist_flag          object
widget_products       int64
engagement_products   int64
annual_income_at_source int64
other_bank_cc_holding object
bank_vintage          int64
T+1_month_activity    int64
T+2_month_activity    int64
T+3_month_activity    int64
T+6_month_activity    int64
T+12_month_activity   int64
Transactor_revolver   object
avg_spends_l3m        int64
Occupation_at_source  object
cc_limit              int64
dtype: object

```

	T+1_month_activity	T+2_month_activity	T+3_month_activity	T+6_month_activity	T+12_month_activity
high_networkth	-0.007266	-0.008875	-0.000815	0.009303	0.002106
cc_active30	0.009978	0.053988	0.044509	0.150242	0.155215
cc_active60	0.009472	0.040777	0.033131	0.097626	0.100857
cc_active90	0.004373	0.033213	0.024057	0.072167	0.074556
Transactor_revolver	0.002869	-0.002519	0.013973	-0.005085	-0.005706
avg_spends_l3m	0.008703	-0.015421	-0.000437	0.003185	0.013882
cc_limit	-0.006859	-0.010475	-0.005167	-0.002523	0.001032

The Top 5 important variables are: -




Based on the correlation analysis between key variables and future credit card activity (T+1 to T+12-month activity), the top 5 important variables that could impact credit card attrition and usage are:

- 1) **cc\_active30 (Credit Card activity in the last 30 days)**: This variable shows a significant correlation with future activity, especially in the T+6 and T+12-month indicators. This suggests that recent activity is a strong predictor of continued engagement and can be a focus for reducing attrition. Higher activity in the recent past is associated with a higher likelihood of future use.
- 2) **cc\_active60 and cc\_active90 (Credit Card activity in the last 60 and 90 days)**: While slightly less correlated than cc\_active30, these variables still show a meaningful relationship with future activity, indicating that sustained engagement over a longer period is crucial for retention.
- 3) **Avg\_spends\_l3m (Average spends in the last 3 months)**: Although the correlations are generally lower, this variable's positive correlation with T+12-month activity suggests that customers who spend more on average are more likely to continue using their credit card, highlighting the importance of encouraging higher spending patterns.
- 4) **Transactor\_revolver**: This variable shows a slight correlation with future activity, implying that the nature of credit card use (paying off balances vs. carrying over) may influence long-term engagement, though its impact is not as strong as the activity metrics.

- 5) **cc\_limit (Current credit card limit)**: The correlation is relatively low with future activities, suggesting that while important, the credit limit itself may not directly influence future activity as strongly as other factors. However, it's essential to consider that a higher credit limit might enable higher spending levels, indirectly affecting engagement and retention.

### Business Justifications:

- Focusing on increasing credit card activity (**cc\_active30, cc\_active60, cc\_active90**) through personalized offers, rewards, or promotions can significantly enhance customer retention and engagement.
- Encouraging higher spending (**avg\_spends\_l3m**) through targeted marketing campaigns can not only increase immediate revenue but also bolster long-term card usage.
- Understanding the behavior of Transactors vs. Revolvers can help tailor financial advice, offers, and credit limits to suit different customer profiles, potentially increasing satisfaction and reducing attrition rates.
- Adjusting credit limits (**cc\_limit**) strategically, based on customer spending habits and engagement levels, could further personalize the banking experience, encouraging higher spending and loyalty.
- These insights suggest that strategies focusing on increasing credit card activity and spending, along with personalized credit management, could be key to improving business outcomes for **GODIGT Bank**.

Highest Correlation with Future Activity		
Variable		
cc_active30	0.155215	
cc_active60	0.100857	
cc_active90	0.074556	
avg_spends_l3m	0.013882	
Transactor_revolver	0.013973	