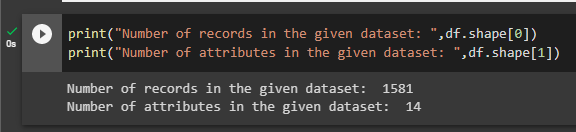
**Problem 1:**

**A. What is the important technical information about the dataset that a database administrator would be interested in? (Hint: Information about the size of the dataset and the nature of the variables)**

The given dataset is of volume 1581\*14 i.e., 1581 records and 14 columns/attributes.



Out of 14 columns, below are quantitative variables in which only No\_of\_Dependents is decrete, rest 5 are continuous

| **Age** | **No\_of\_Dependents** | **Salary** | **Partner\_salary** | **Total\_salary** | **Price** |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |

Remaining all 8 variables(Gender, Profession, Marital\_status, Education, Personal\_loan, House\_loan, Partner\_working, Make) are categorical.

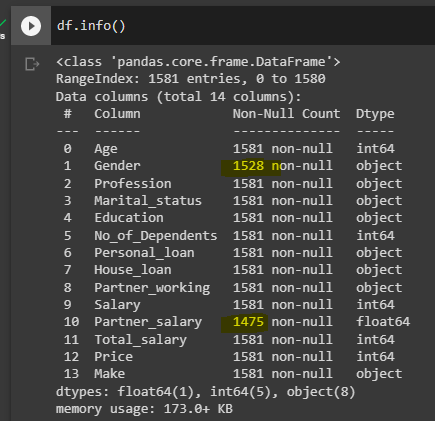
|  | **Gender** | **Profession** | **Marital\_status** | **Education** |  | **Personal\_loan** | **House\_loan** | **Partner\_working** |  |  |  |  | **Make** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

(1581, 14)

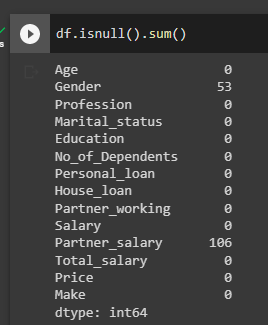
**B. Take a critical look at the data and do a preliminary analysis of the variables. Do a quality check of the data so that the variables are consistent. Are there any discrepancies present in the data? If yes, perform preliminary treatment of data.**

In the given dataset, columns Gender and Partner\_salary has some nulls as shown below

At this point, we don’t know its impact yet, going forward if we feel these nulls gonna impact our analysis then we can remove these.

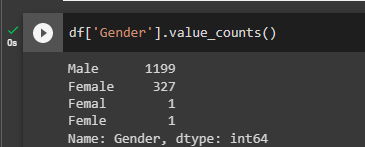


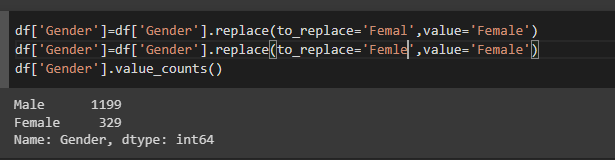
Below are the null counts



Also there are few values in Gender columns which are incorrect, lets fix this by this below loc

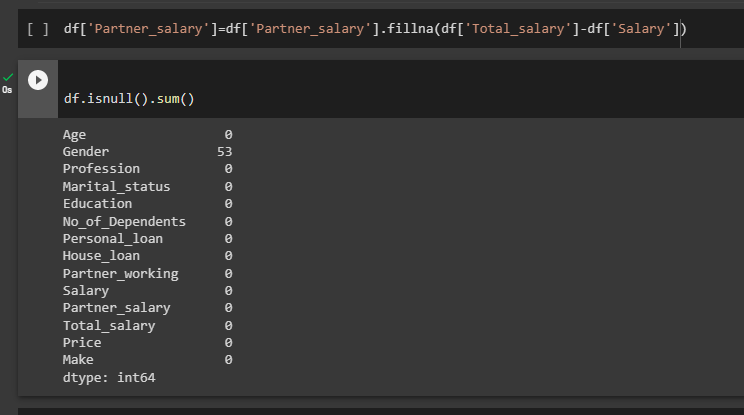
Please note, we are not making any modification to null entries in gender, because there is no way to identify the gender, even if we impute with mode then the given dataset would become more male dominant. Hence leaving the nulls as is. However these nulls wont impact our graphs below.





Similarly lets impute null values of total\_salary with some values using below formula

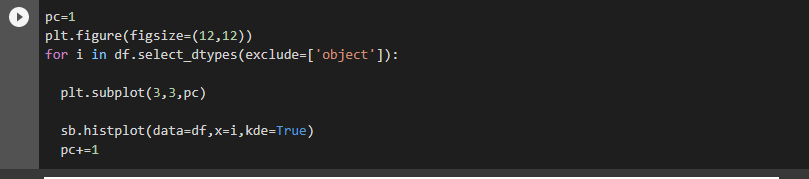
Partner\_salary=Total\_salary-Salary



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

Lets explore all the features of the data and check how each variable is spread.

First, lets see how quantitative variables are spread



I have looped over all the numerical variables and plotted histplots on each of these variables.

If we see the below graphs, we can infer the following information,

1. **Age:**

As clearly seen below, data distribution of age is somewhat right skewed. That means, majority of the purchases were done by people with relatively lower ages. Hence we can analyse more on higher age people and conduct necessary marketing campaigns to bring more elder customers too. Also understand why there are very less number of car owners in higher age group. Identify the reasons and device solutions for it.

1. **No\_of\_Dependents:**

Maximum purchases were done by people with dependents 2 and 3 and minimum were made by people with 0 dependents. Here we can do some interesting thinking on how we can increase purchases by people with dependents further also can we increase sales to independent people by creating ads targeting those independent people.

1. **Salary:**

People with salary of around 60000 are purchasing more as per the data. After certain salary increase, there is a reduction in purchases.

What is it that’s drawing more people around this salary bracket(55000 to 62000 approx.) to buy more, can we apply the same to the people in other salary brackets.

1. **Partner\_salary:**

This attribute has minor impact on purchases as per the given data but still there is some impact.

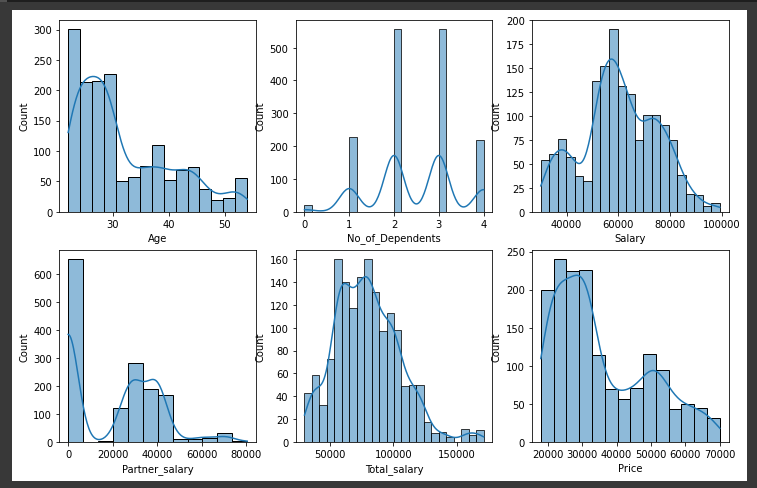
But people with less partner salary has more contribution to purchases, then it took a dip and again increased at around 20000 to 45000 salary bracket. Could be taken interesting decisions on this.

1. **Total\_salary:**

The distribution of this attribute is almost similar to Salary distribution but a bit more right skewed due to addition of partner\_salary.

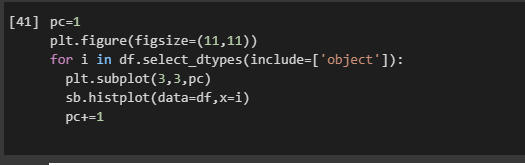
1. **Price:**

More purchases made for the cars with price that falls below 35000 and purchases dropped significantly from there. Can we create any campaigns to sell high end cars to people with higher salary or income by creating ads promoting luxury.



Now lets explore categorical variables to understand each of the variable’s dominance

For categorical variables, I have looped over all the object type attributes and plotted countplot from seaborn for each of these attribute.



1. **Gender:**

There are more Male buyers than female drivers. Can we increase Female customers by creating campaigns promoting women drivers or creating cult of female car owners.

1. **Profession:**

As per this graph, Salaried people tends to have more purchases than Business people.

1. **Marital\_status:**

Married people have more contribution to car purchases which is in line with dependent graph. Interestingly both single and independent people(as seen from quantitative variable analysis above) have very less contribution to purchases, there should be some ad/campaign to draw these customers. May be same campaign targeting both categories.

1. **Education:**

Post Graduates have more purchases. Highly educated people might have higher earning which lead to more spends on car purchase.

1. **Personal\_loan:**

Buys with loans and without loans are almost similar

1. **House\_loan:**

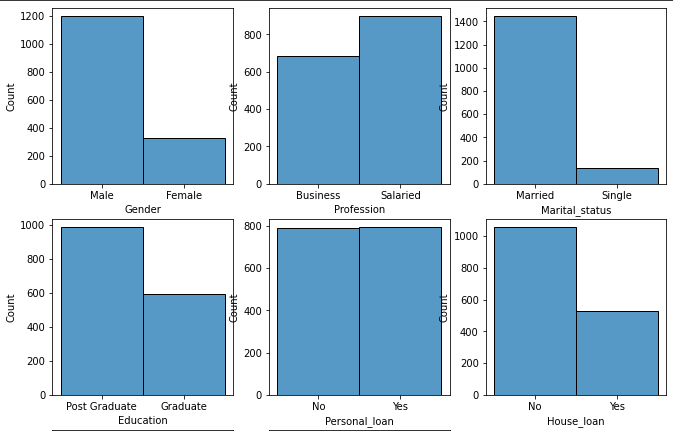
Buyers without home loans are more than those with home loans. To those who are already having loans, can we target them with low cost cars.

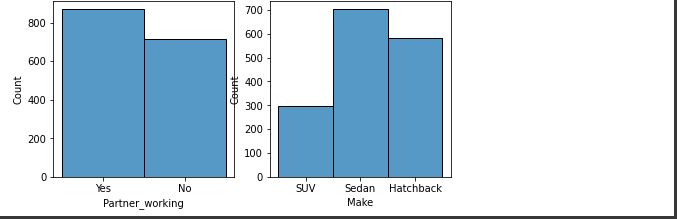
1. **Partner\_working:**

People with working partners have more purchases.

1. **Make:**

Of all the makes, Sedan is the most sold followed by Hatchback and then SUV. More sedan sales could be due to more purchases by dependent people. Can we create specific ads promoting Hatchbacks to single people with home loans or can we create ads to promote SUV’s to single independent people. All this thinking can be applied.



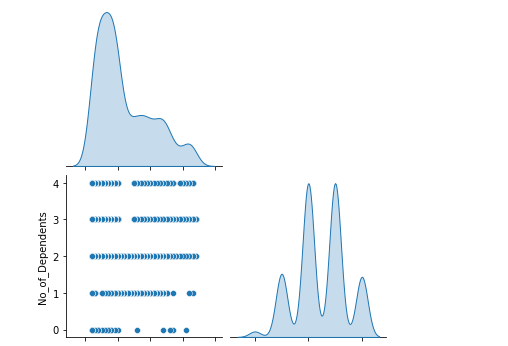


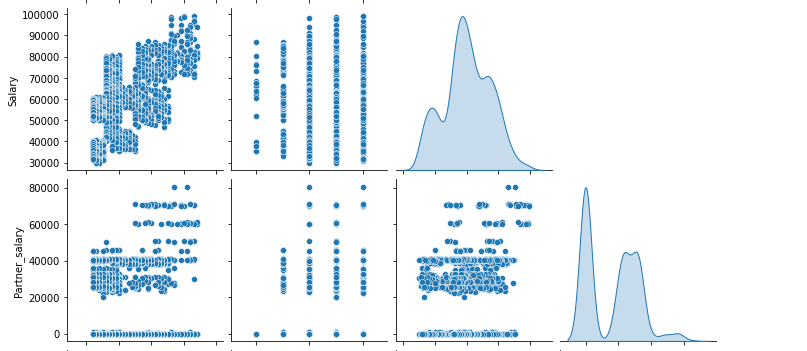
**D. Understanding the relationships among the variables in the dataset is crucial for every analytical project. Perform analysis on the data fields to gain deeper insights. Comment on your understanding of the data.**

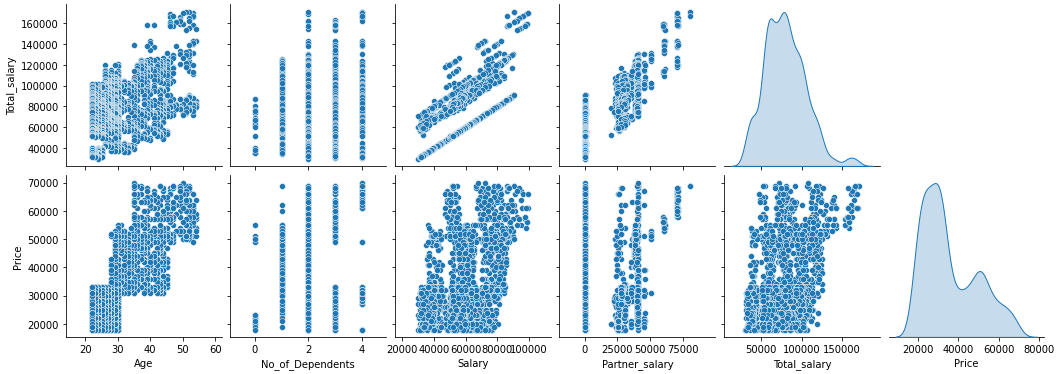
First, Here is the relationship between quantitative variables,

Lets create a pairplot across all the numerical variables in the dataset to understand how each variable is varying with the other variable.

sb.pairplot(data=df,vars=df.select\_dtypes(exclude='object'),diag\_kind='kde',kind='scatter',markers=["0"],corner=True);



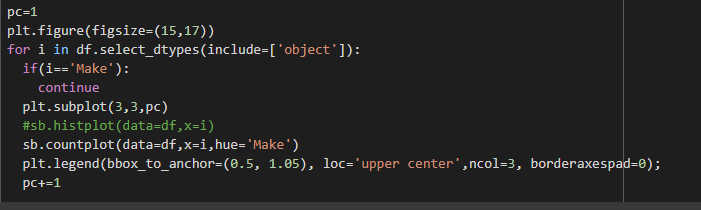


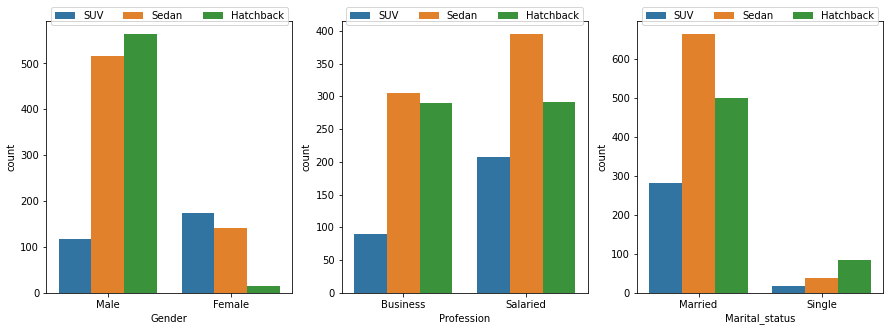


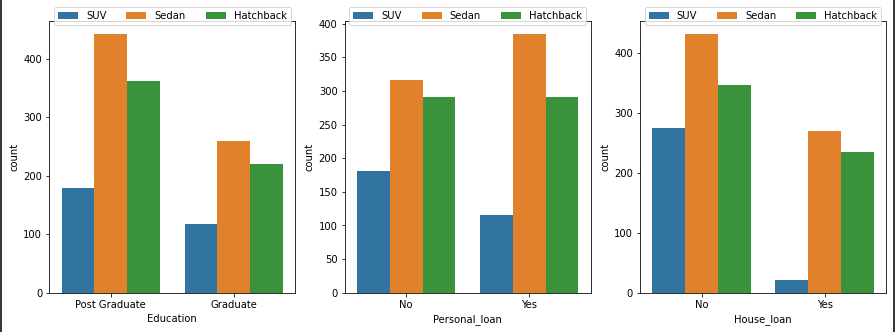
Based on above graph, there seems to be a slight correlation between “Salary-Age”, “Price-Age”

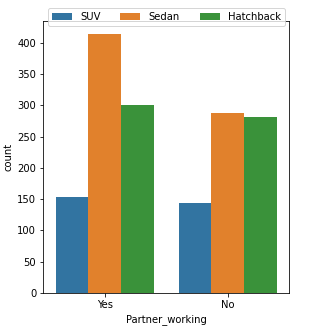
Similarly, here is the relationship between categorical variables,

For this visualization, I have used countplot on all object type attributes with hue on “Make” for each graph.









Insights:

Males have higher dominance in buying cars in the given data. Also males are mostly preferring hatchback more then sedan and then suv. Interestingly females are preferring SUV’s more followed by sedan and hatchback.

Overall, more salaried professionals own cars than business professionals and sedan type is the highest picked car among both professions.

Married people are the major car owners than that of single.

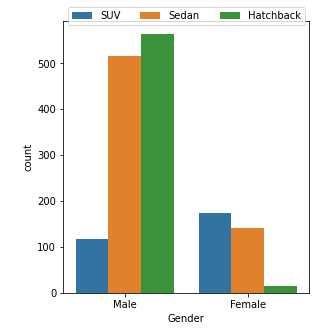
Post graduates have higher purchase count than that of graduates.

Most of the car owners doesn’t have home loan.

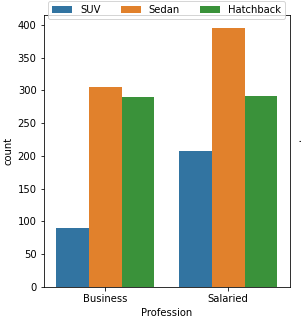
People with working partners are preferring sedan followed by hatchback and suv, same is the case with non working partners.

**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.**

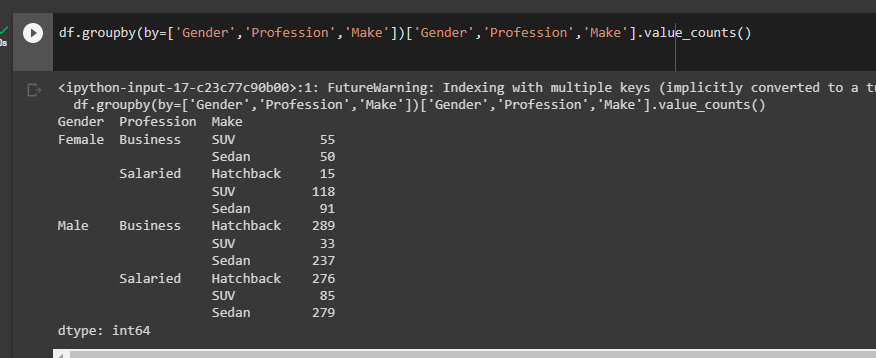
E1) Steve Roger says “Men prefer SUV by a large margin, compared to the women” **– Disagree, As per below graph Female prefer SUV by a large margin than Male.**



E2) Ned Stark believes that a salaried person is more likely to buy a Sedan – **Agree, As per below graph, salaried person is more likely to buy a sedan.**



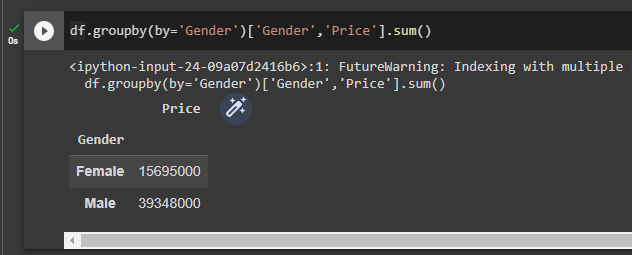
E3) Sheldon Cooper does not believe any of them; he claims that a salaried male is an easier target for a SUV sale over a Sedan Sale – **Disagree, As per below data it clearly shows salaried males prefer sedan more than SUV.**



**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.**

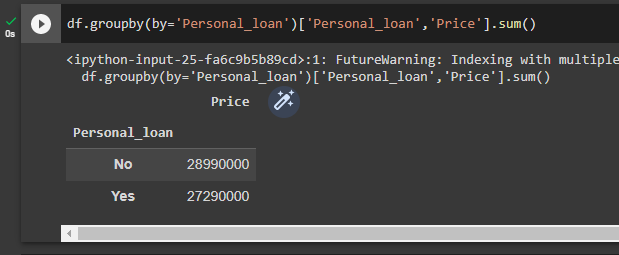
**F1) Gender**



As per above data, Males spent twice as much as money when compared to females. This is due to, the given data has more females than males.

So if somehow we can give more offers or some incentives to females then their expenditure on automobile may increase.

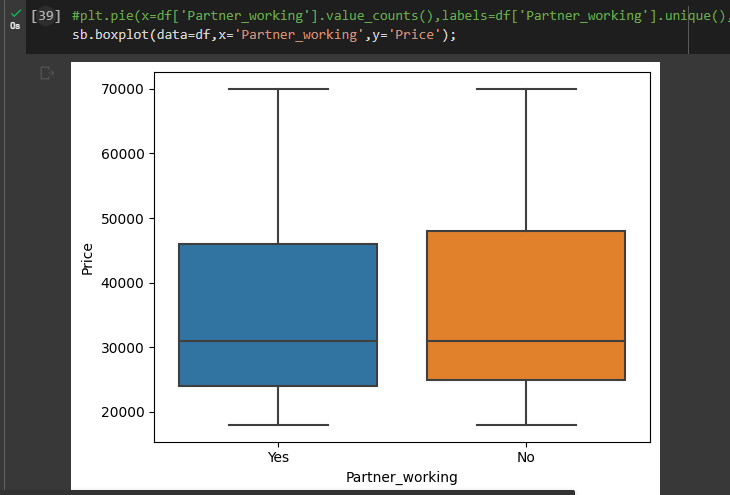
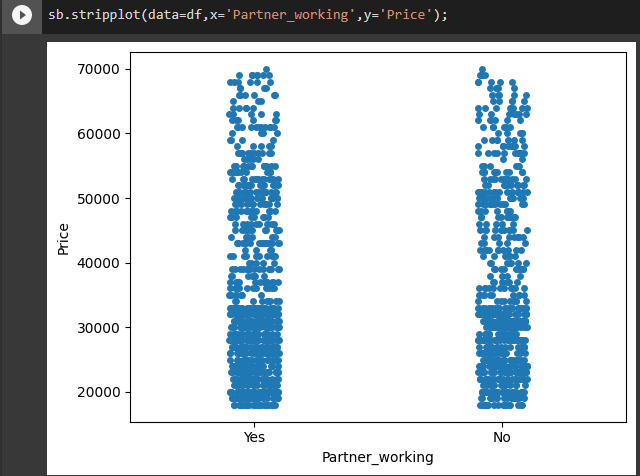
**F2) Personal\_loan**

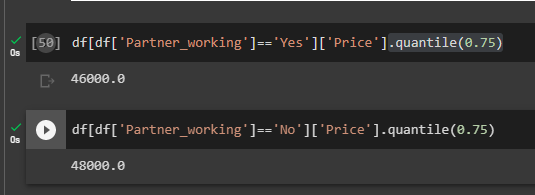


People with Presonal\_loan spent a bit less amount on automobiles when compared to people with no personal\_loan

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

As It can be seen below, price spent by both Partner\_working and not working is almost same. But when we see the 75th percentile of the price spent by both categories, people with non working partners spent more on high-priced car.

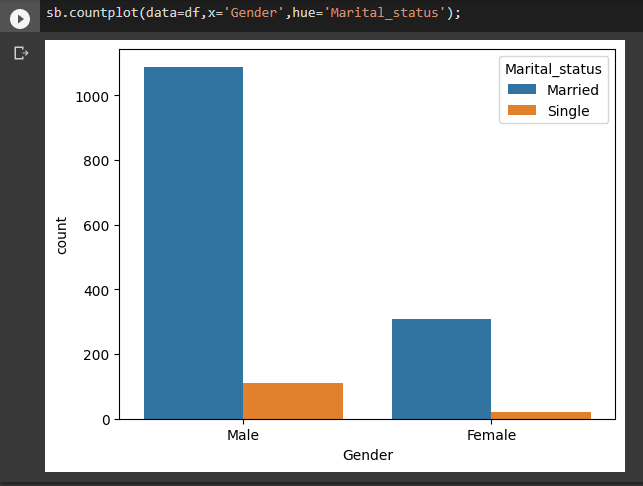


**H. The main objective of this analysis is to devise an improved marketing strategy to send targeted information to different groups of potential buyers present in the data. For the current analysis use the Gender and Marital\_status - fields to arrive at groups with similar purchase history.**

As per below data, In both the genders, married are more likely to buy cars than that of single.

We can either target single males and females by giving incentives or offers to them to increase sales to singles.

We can also increase more sales to marred Females by giving some offers.



-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Problem 2:**

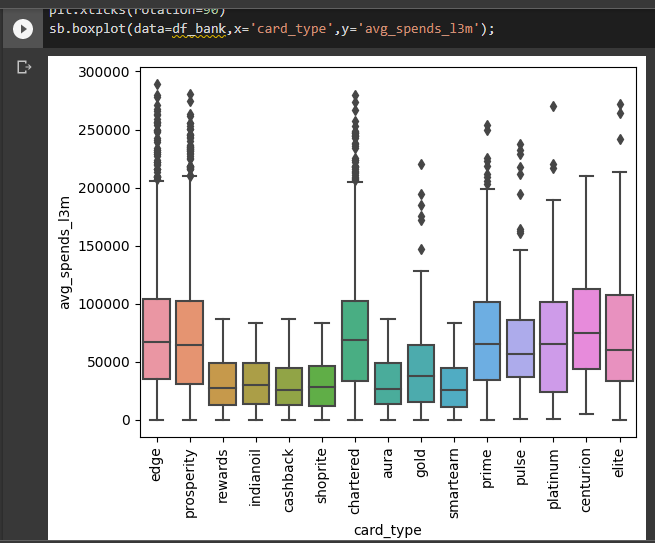
Below are the 5 attributes that are very important to this problem statement(Revisit the credit card policy and make sure the card given to the customer is the right credit card)

1. **Card\_type**
2. **Other\_bank\_cc\_holding**
3. **Avg\_spends\_l3m**
4. **Annual\_income\_at\_source**
5. **Occupation\_at\_source**

Below is my justification for each of the attributes.

**card\_type** is very important to consider here so that we can analyze which card type is contributing to more avg spends as shown below.

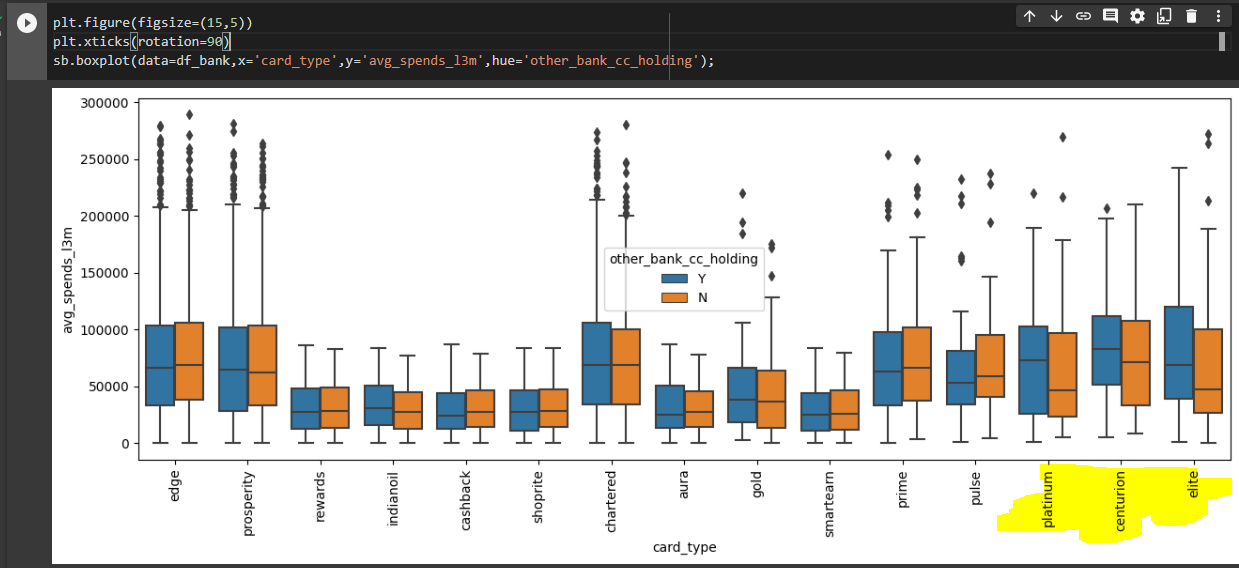
*Have used boxplot across each card\_type against avg\_spends\_l3m*



**other\_bank\_cc\_holding**

This attribute is also important as we can see some interesting pattern for the cards – “platinum”,”centurion” and “elite”. For these cards, even though they have cards with other banks but still they have more spends with our bank. Why is it so, we can check more on this and try to bring their spends on other cards to our bank. Also may be we can apply the same to other card\_types as well.

*Have used boxplot across each card\_type against avg\_spends\_l3m with hue on other\_bank\_cc\_holding.*

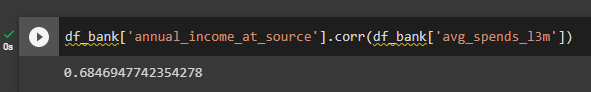


**avg\_spends\_l3m**

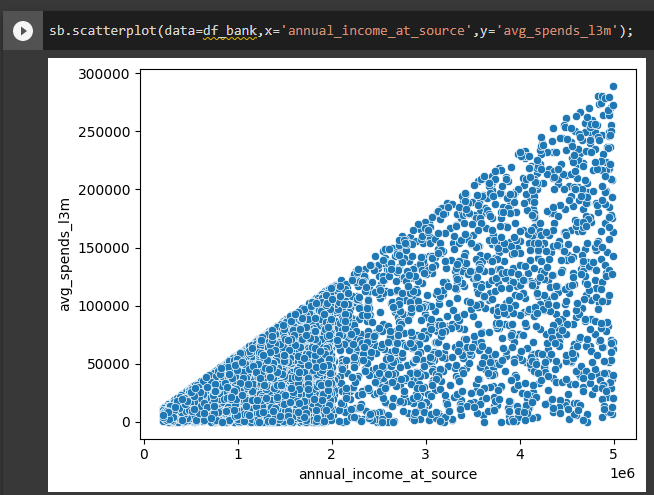
average spends is the feature which we use to validate/examine our analysis against. This is the variable we use are reference on how much spends a customer has done using credit card.

**annual\_income\_at\_source**

Here we have to give attention to correlation between annual\_income\_at\_source and avg\_spends\_l3m attributes.



*Have used scatterplot from seaborn between annual\_income\_at\_source and avg\_spends\_l3m*



Occupation\_at\_source could be a important feature to consider as well, as shown below, some categories of occupation\_at\_source have higher spends when compared with other card types. Can we give the highest spent card types to these categories and all we can check using this feature.

*Here i have used boxplot across each card\_type against avg\_spends\_l3m with hue on Occupation\_at\_source.*

