DSBA Business report

Contents

Problem 1

Context

Analysts are required to explore data and reflect on the insights. Clear writing skill is an integral part of a good report. Note that the explanations must be such that readers with minimum knowledge of analytics is able to grasp the insight.

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).

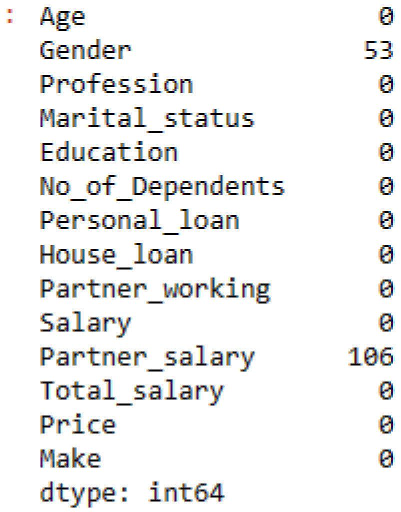
price: The price of a product or service.

make: The type of automobile

Analysis:

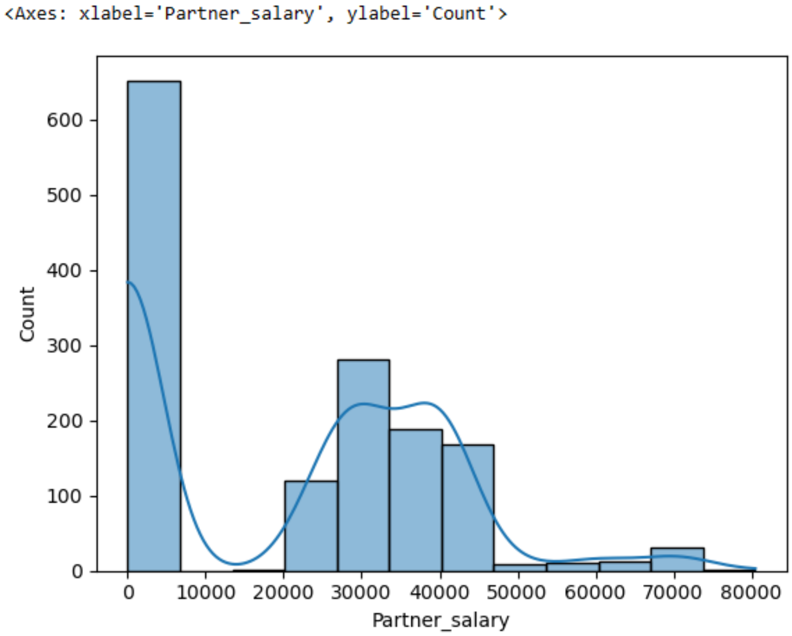
Austo_Automobile.csv file contains data which has 1581 rows and 14 columns. It contains both numerical and categorical data.

The Dataset contained some null values and incorrect values. For instance, Gender and Partner_Salary has incorrect data.



```
: 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
```

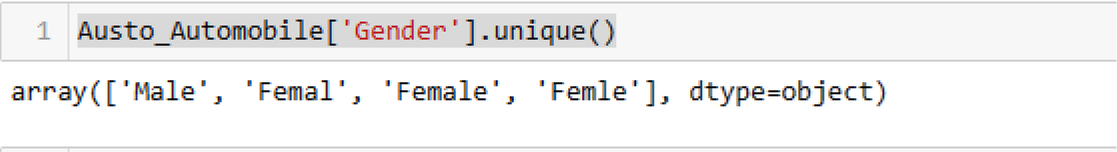
Treating the Incorrect data of Partner_Salary required replacing the nan with median.
The reason for choosing median was: The graph is not symmetrical and could use mean as an option for that.



```
<Axes: xlabel='Partner_salary', ylabel='Count'>
```

The Gender had two ways of incorrect values.

The spelling in the values of the Gender was given as Femal and Femle.



```
Austo_Automobile['Gender'].unique()
```

```
array(['Male', 'Femal', 'Female', 'Femle'], dtype=object)
```

Justification: The Mode is the right treatment for bad data for categorical variable.

The Gender had some null values.

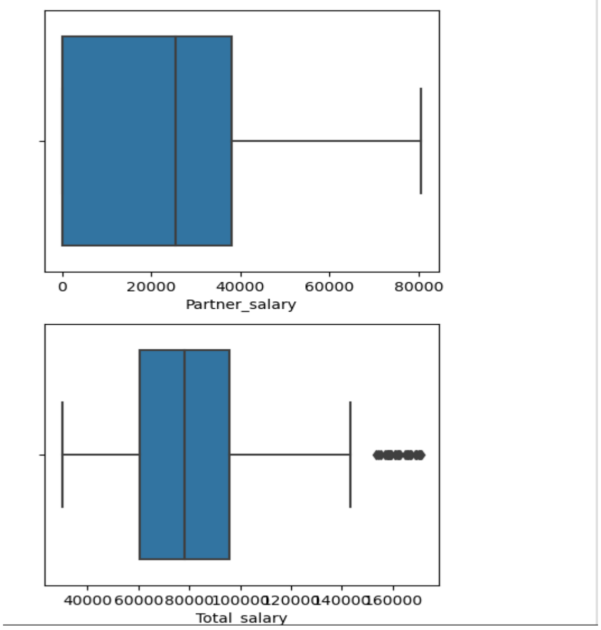
Justification: Replaced with mode.

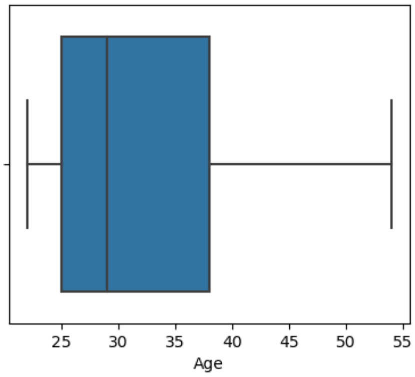
Univariant analysis:

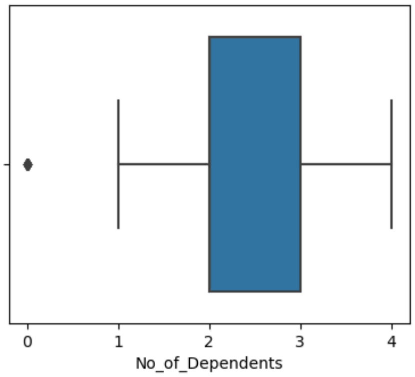
On Numerical:

No of dependents and Total_Salary have outliners. Price is Right Skewed. Total_Salary is normally distributed.Partner salary does not have q1 value. Its value starts from 0 i.e min.

The median is not having the correct sized data. Its not normally distributed.

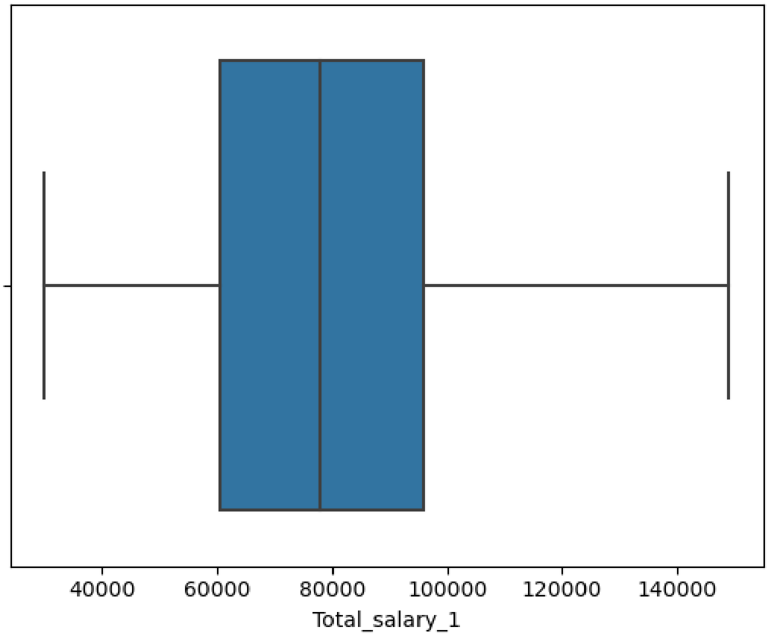




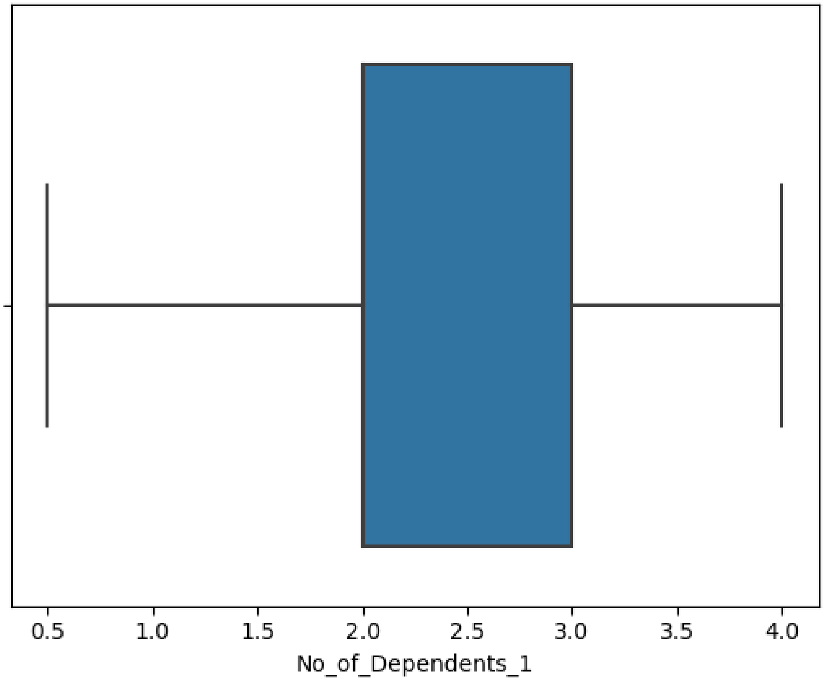


Treatment of Outliners is done by Zscore and other method used is box plot type. After treatment the graph looks like below.

Total_Salary



No of dependents

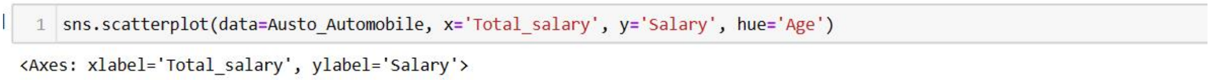


Categorical variable Analysis:

1200+ are Male categorical variable. People prefer Sedan cars compared to Hatchback and SUV. Around 50% of people have house loans. There are equal number of people who have opted for personal loans and not opted for personal loans. 868 people have working partner and more than 100% are married. More number of people are post graduates and salaried.

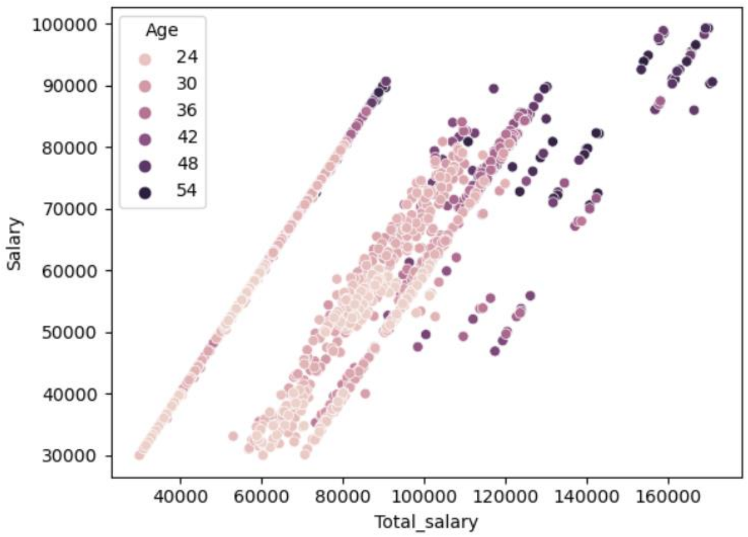
Bivariant analysis:

The Scatterplot is used to depict the analysis and insights of the Numeric variable analysis.

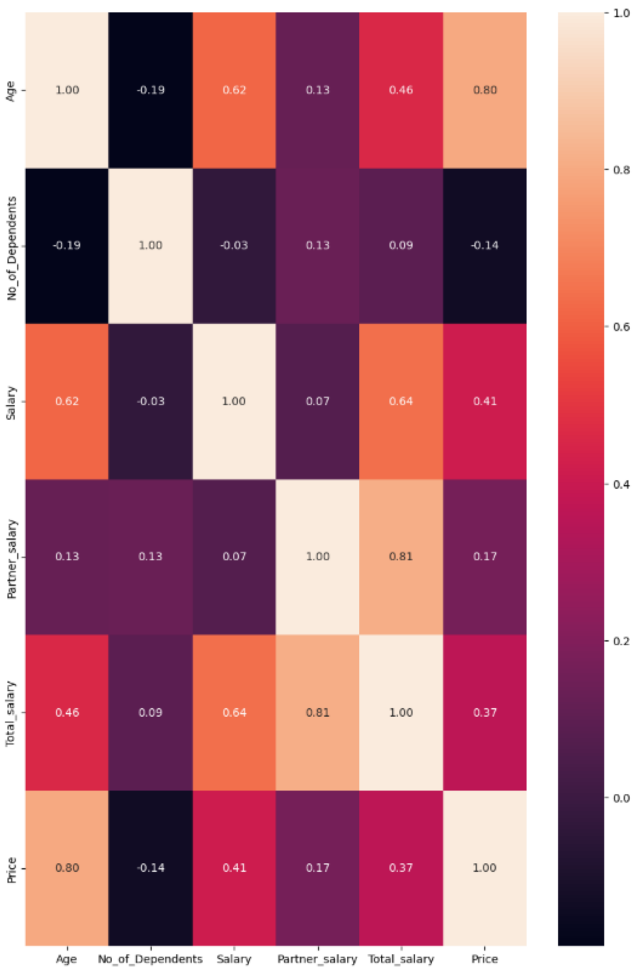


```
sns.scatterplot(data=Austo_Automobile, x='Total_salary', y='Salary', hue='Age')
```

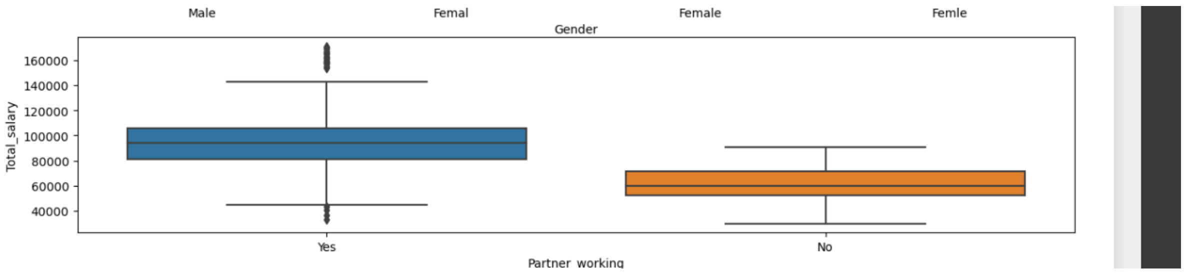
```
<Axes: xlabel='Total_salary', ylabel='Salary'>
```



Identified correlation between different categorical variables. Heatmap is used in identifying so.



The categorical bivariant analysis as shown. Here we could see the partner working has more salary compared to non-working partner.



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

2.1. For this data, construct the following contingency tables (Keep userid as row variable):

Important variables.

Userid, Interest

card_bin_no

annual_income_at_source

Occupation_at_source

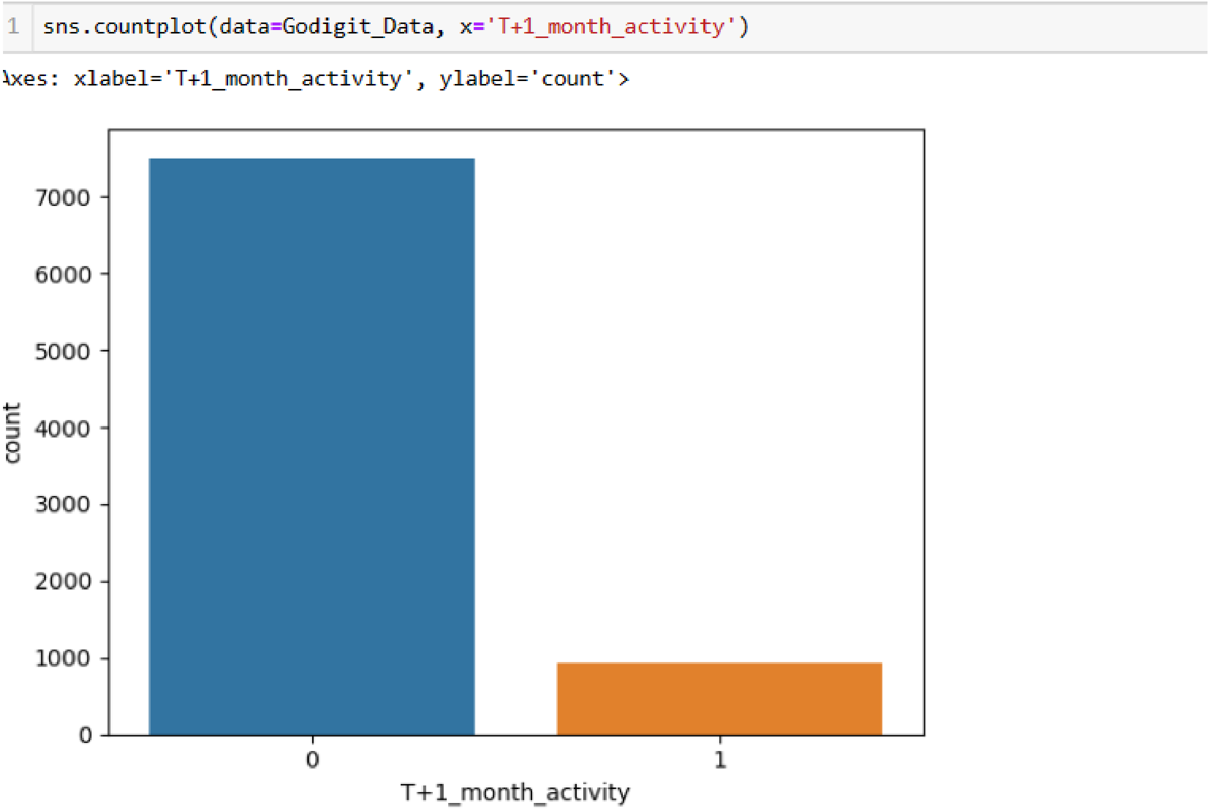
cc_limit

Justification:

Bank always consider the Occupation of a person to give the credit card and also the income of the person. More the income more the credit limit. Every user will be issued with one card which will have card no and card bin no which will be unique number.

Based on the data:

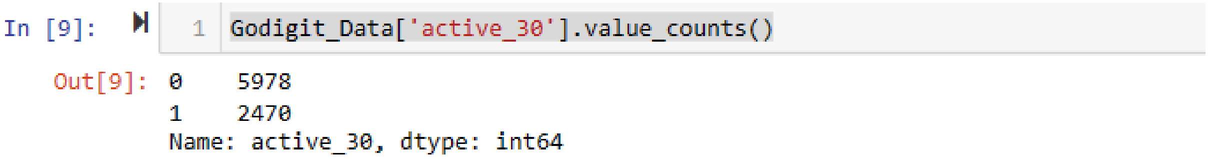
What is the probability of Whether customer uses credit card in T+1 month?



```
sns.countplot(data=Godigit_Data, x='T+1_month_activity')
```

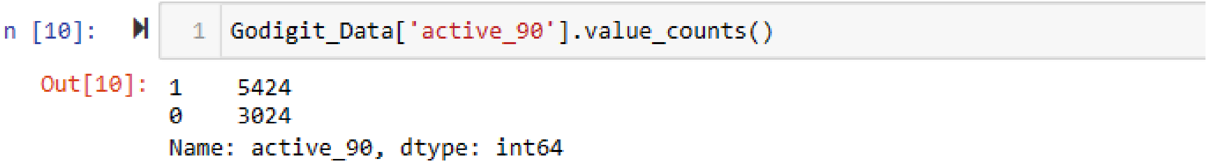
```
\xes: xlabel='T+1_month_activity', ylabel='count'>
```

What is the data exists to show that CC activity is less in 30 days compared to 90 days?



```
In [9]:    Godigit_Data['active_30'].value_counts()

Out[9]: 0    5978
        1    2470
        Name: active_30, dtype: int64
```

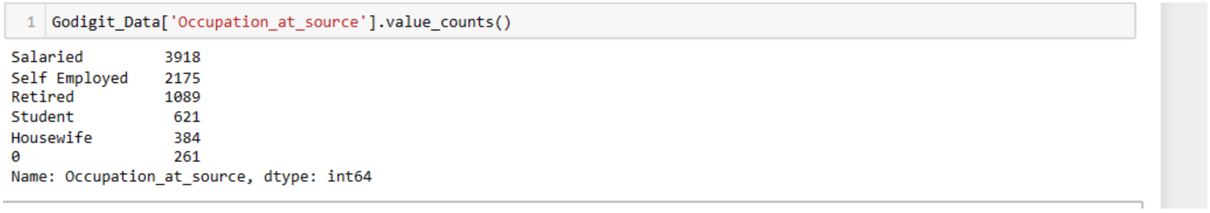


```
n [10]:    Godigit_Data['active_90'].value_counts()

Out[10]: 1    5424
         0    3024
         Name: active_90, dtype: int64
```

As seen the active 1 is less in 30 days as compared to 90 days.

What are the chances that Occupation recorded at time of credit card application was more of Salaried?

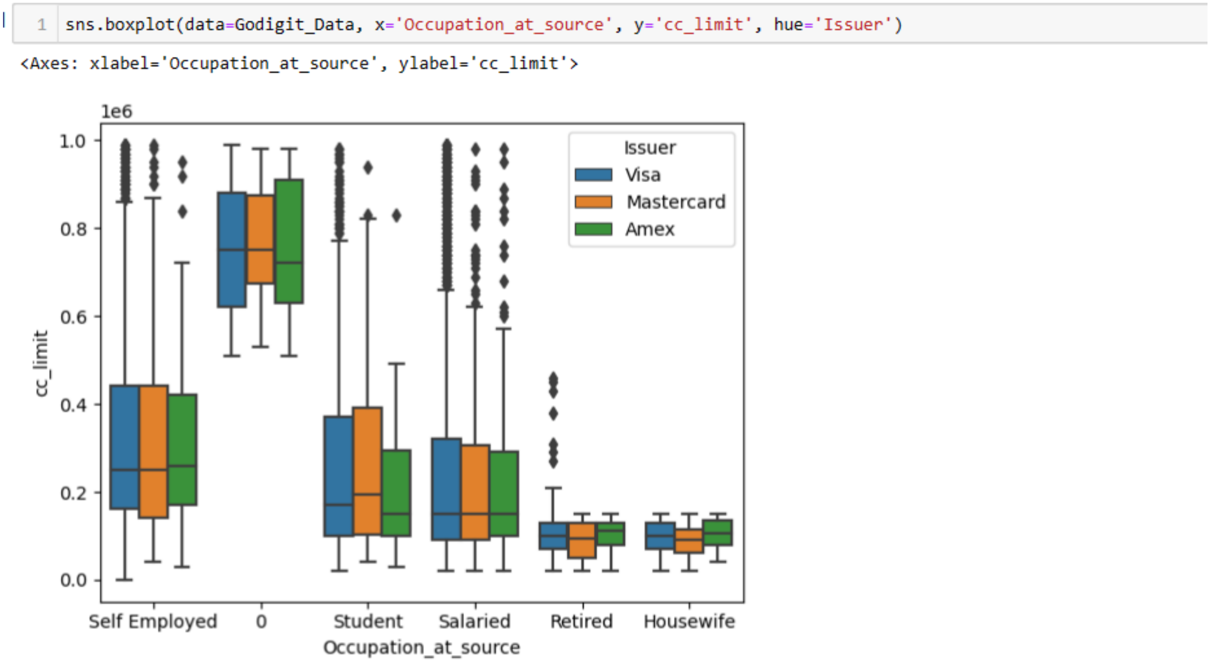


```
Godigit_Data['Occupation_at_source'].value_counts()
```

```
Salaried         3918
Self Employed    2175
Retired          1089
Student           621
Housewife         384
0                 261
Name: Occupation_at_source, dtype: int64
```

Salaried employees were given more preference as compared to other employement.

Based on Occupation is that credit card limit is issued?

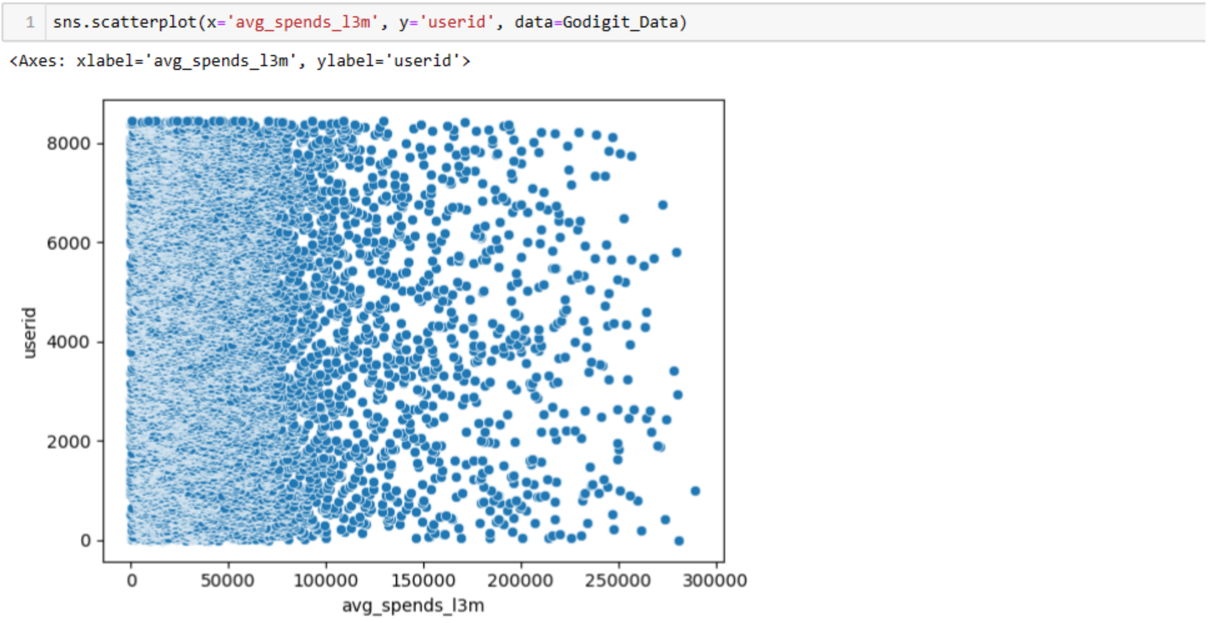


```
sns.boxplot(data=Godigit_Data, x='Occupation_at_source', y='cc_limit', hue='Issuer')
```

```
<Axes: xlabel='Occupation_at_source', ylabel='cc_limit'>
```

From the graph it looks like student and salaried both almost have same limit compare to other occupation.

What is the average spent of credit card for all credit card holders?

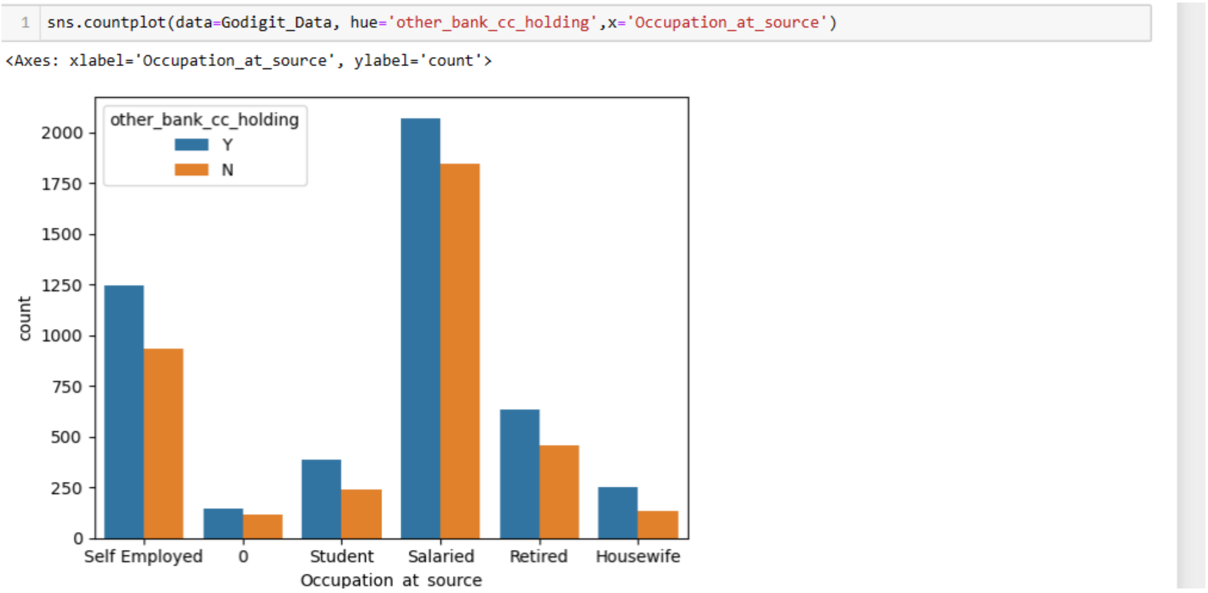


```
sns.scatterplot(x='avg_spends_l3m', y='userid', data=Godigit_Data)
```

```
<Axes: xlabel='avg_spends_l3m', ylabel='userid'>
```

Its positive correlation when compared to all userid.

Whether the customer holds another bank credit card?

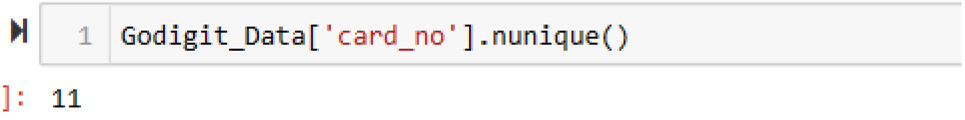


```
sns.countplot(data=Godigit_Data, hue='other_bank_cc_holding',x='Occupation_at_source')
```

```
<Axes: xlabel='Occupation_at_source', ylabel='count'>
```

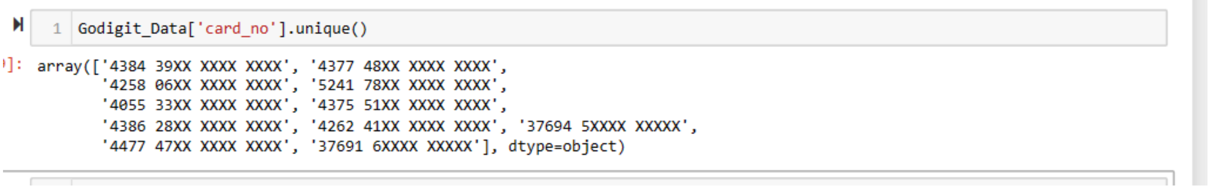
Salaried person tends to keep more than one credit card. And other occupation as well except for students.

Show Every card number is unique and different from other card holder?



```
Godigit_Data['card_no'].nunique()
```

```
]:  11
```



```
Godigit_Data['card_no'].unique()
```

```
]:  array(['4384 39XX XXXX XXXX', '4377 48XX XXXX XXXX',
           '4258 06XX XXXX XXXX', '5241 78XX XXXX XXXX',
           '4055 33XX XXXX XXXX', '4375 51XX XXXX XXXX',
           '4386 28XX XXXX XXXX', '4262 41XX XXXX XXXX', '37694 5XXXX XXXXX',
           '4477 47XX XXXX XXXX', '37691 6XXXX XXXXX'], dtype=object)
```

It is not unique and few have repeated values. Ideally this should not be the case and every card must be unique.