## 

| horizontal linePlaceholder image  *PREDICT CREDIT CARD ACCEPTANCE* 27.12.2020 ─ |
| --- |

# **Acknowledgement**

# **Table of Contents**

[**Acknowledgement**](#_ybq20ms8lfxl) **2**

[**Table of Contents**](#_fm1rwyk7foes) **2**

[**Project Objective & Scope**](#_iek4gev4bbz5) **3**

[Objective](#_eac19f74iu10) 3

[Scope](#_u6a93mk0x7b4) 4

[**Methodology**](#_o51w0yaz66v8) **4**

[**Data Acquisition**](#_dtirm5dlqgyu) **5**

[**Loading the Dataset in a Jupyter Environment**](#_n9w8mz652cn9) **5**

[**Data Description**](#_mhk7em9bng7y) **5**

[Searching for Null Values](#_f1jiyt2mjxu) 6

[**Exploratory Data Analysis**](#_wb0yitoxllt6) **8**

[Heatmap for correlation of attributes](#_akp9t20sle0) 8

[Count Plots](#_o01a7gpc7g61) 9

[Scatter Plot and LM plot](#_trartajhdpw1) 11

[Distplot and Histogram](#_sgy5odzgmlhp) 13

[Box Plots](#_z02xvox5lub6) 15

[**Data Preprocessing**](#_i91446iqv7ur) **18**

[Outliers](#_28mmr9d86phe) 18

[Dealing with Outliers](#_ys4426hq62d7) 18

# Project Objective & Scope

## Objective

* In this project we have taken a small credit card dataset for simple econometric analysis which is taken originally from William Greene’s book of ‘**Econometric analysis**’
* Our objective here is to first study the dataset by plotting some graphs and plots and then pre-process the given dataset to make it suitable for training the models. Then, we would train 4 models namely ‘**Linear regression model**’, ‘**Decision Tree model**’, ‘**Naive Bayes model**’, ‘**K-NN model**’. After training the above mentioned models, we would then find out the score, accuracy, classification report and then plot the confusion matrix and ‘**Receiver Operating Characteristic (ROC)**’ graph and find the ‘**Area under the Curve (AUC)**’. We would then choose the best model out of the 4 models based on accuracy and score.
* Our goal is to predict whether a credit card application will be accepted based upon various data about the applicant.

## **Scope**

* The scope of this project is to determine whether the new applicants are likely to receive the card or not.
* To find out the optimum algorithm for this data out of the 4 algorithms used here.
* Customers who intend to receive credit cards can use these trained models to check whether their application will be accepted or not.

# **Methodology**

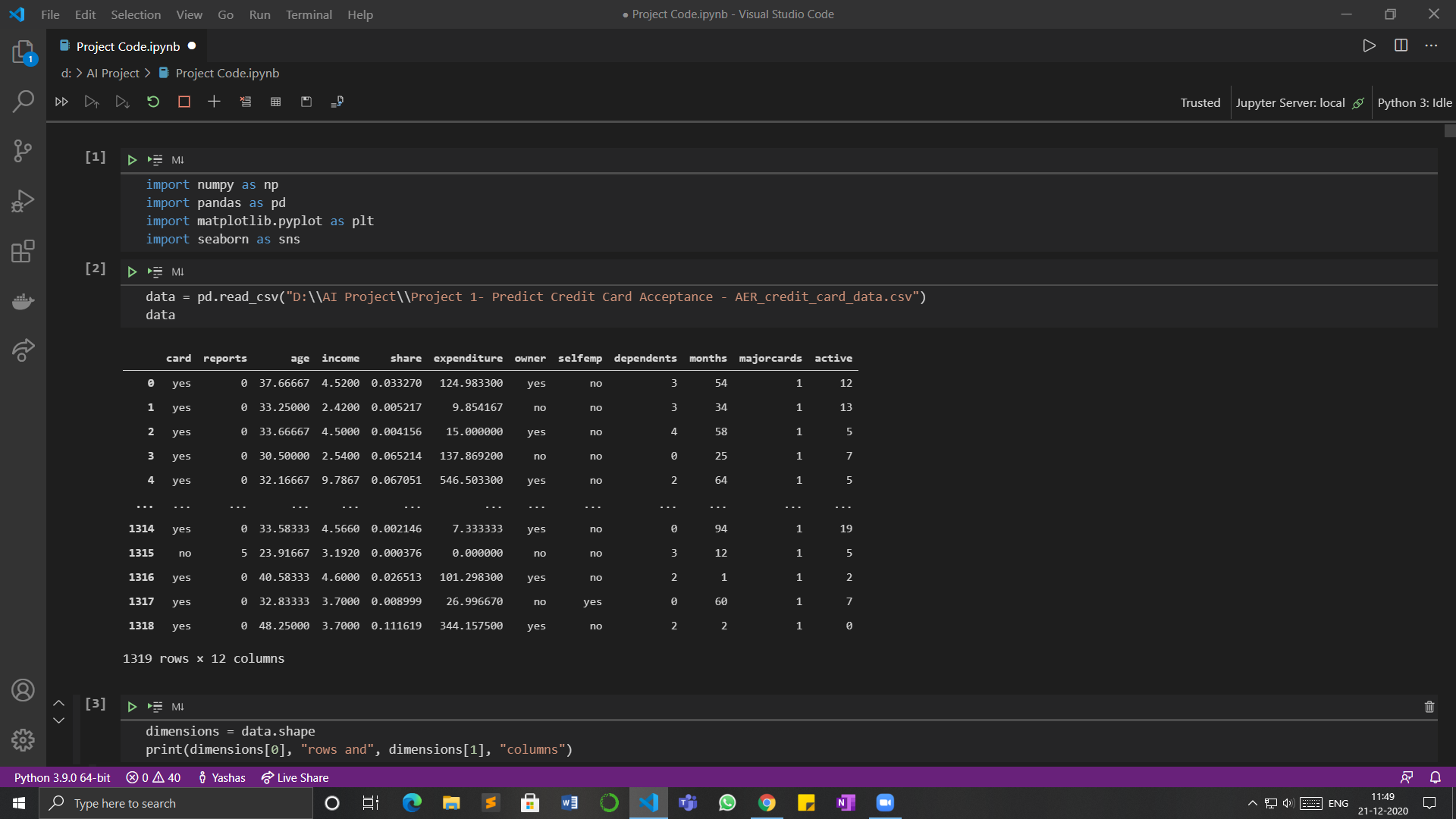
* Data Acquisition
* Loading the Dataset in a Jupyter Environment
* Data Description
* Exploratory Data Analysis
  + Data visualization using plots and graphs
  + Studying the Data
* Data preprocessing
  + Removing outliers using median
* Modelling
  + Training, Testing, Prediction and Analysis of the 4 models
* Comparing the Accuracy of the 4 models
* Choosing the best model on the basis of Accuracy

# **Data Acquisition**

Source of the Data: [GitHub](https://github.com/Anubhav2324/Credit-Card-Acceptance/blob/master/credit_card%20data.csv)

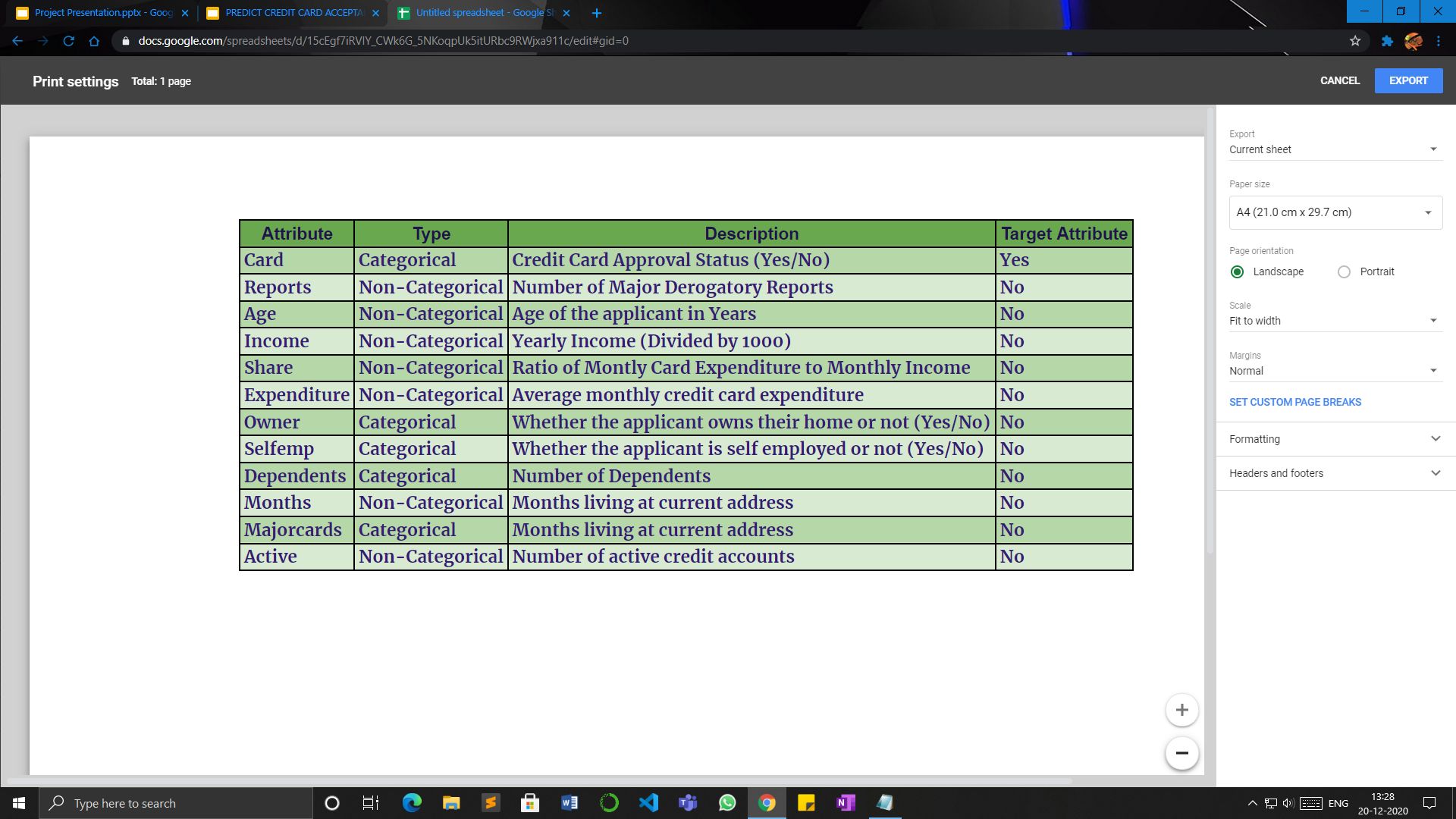
# **Loading the Dataset in a Jupyter Environment**

We loaded the dataset in a jupyter environment using the ‘read\_csv’ method in ‘pandas’ module

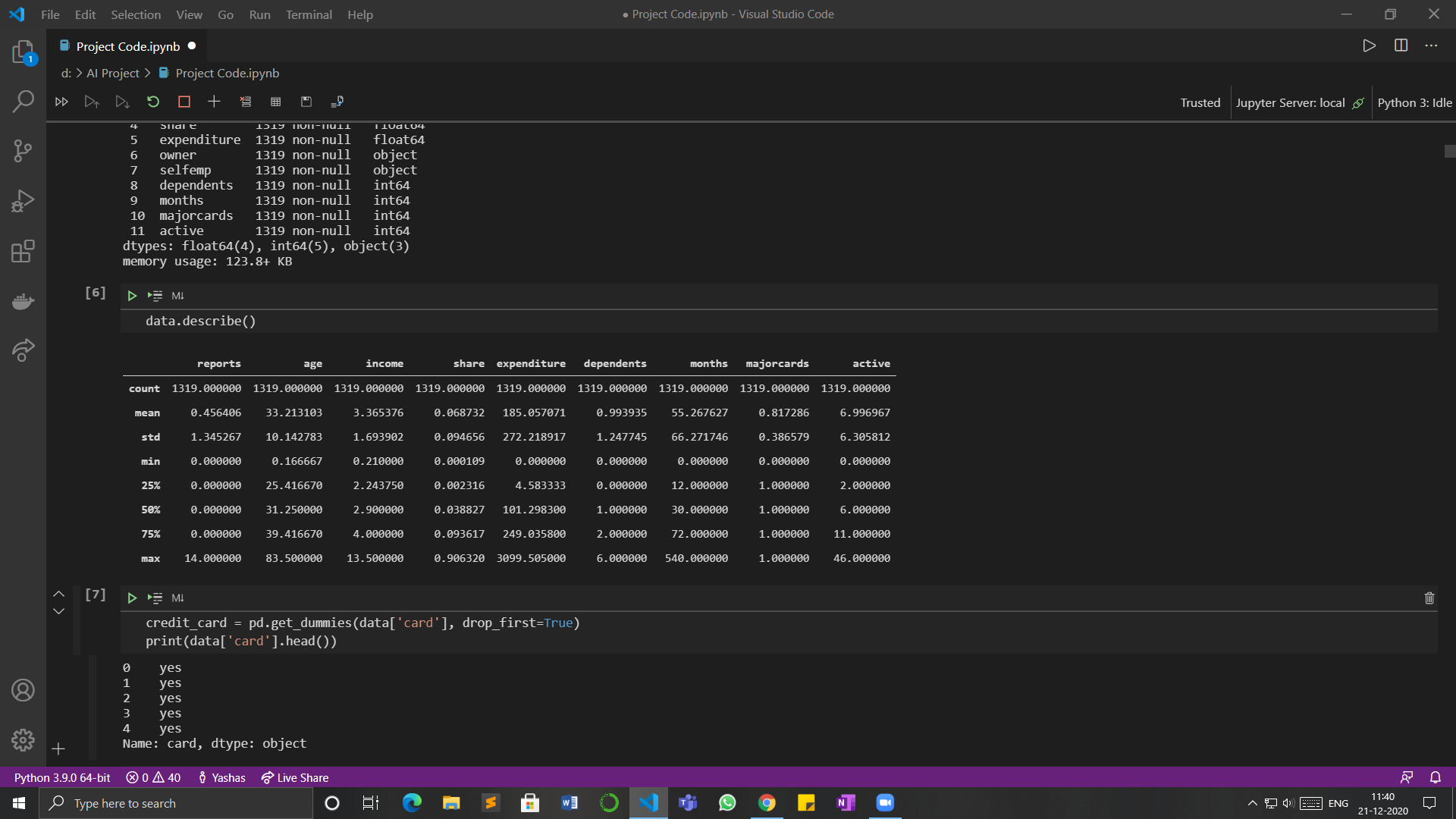


# **Data Description**

**Data Description:** The given dataset has **1319 rows** and **12 columns**.

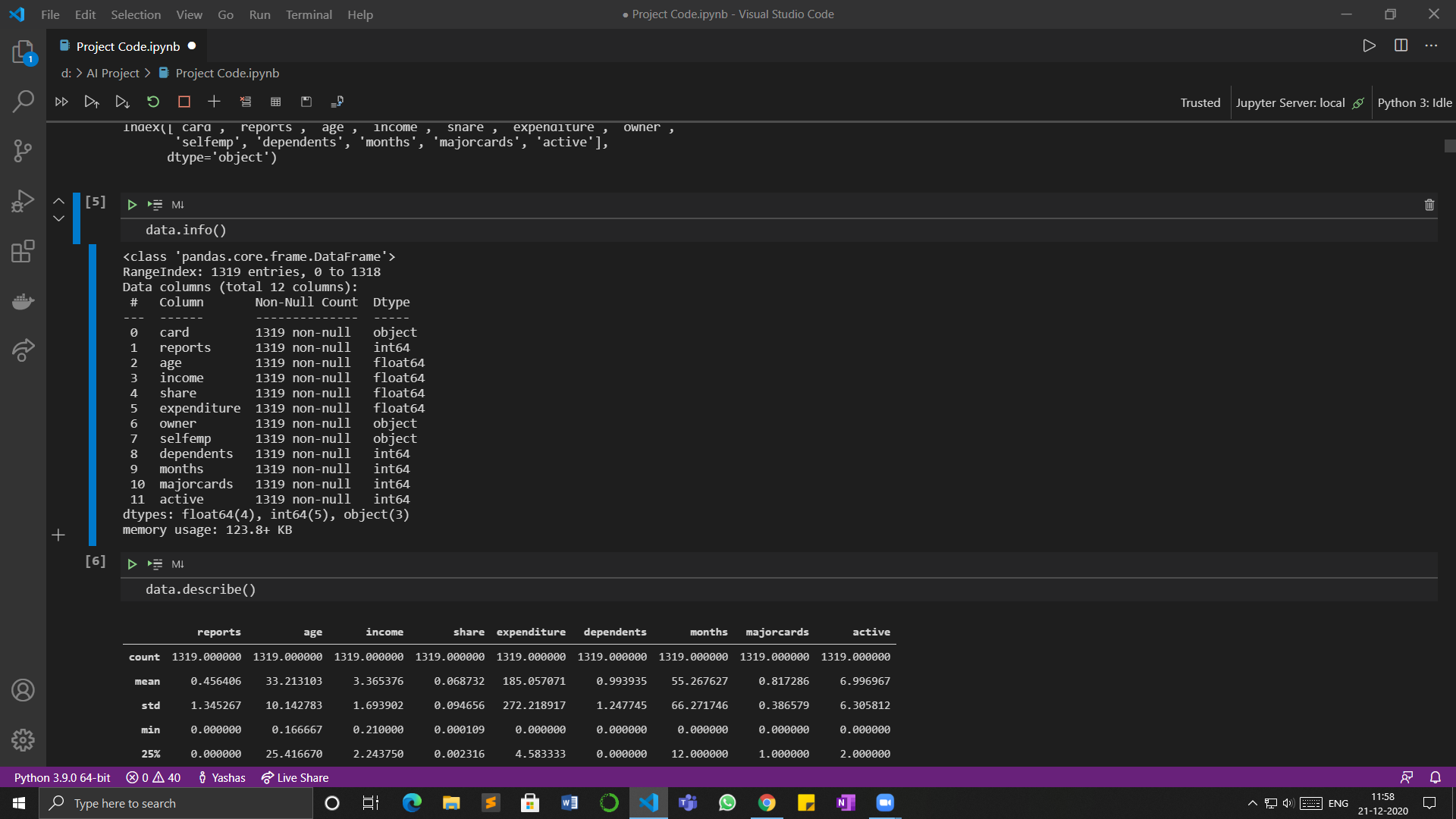


The following table shows the statistics of all the attributes in the dataset :-



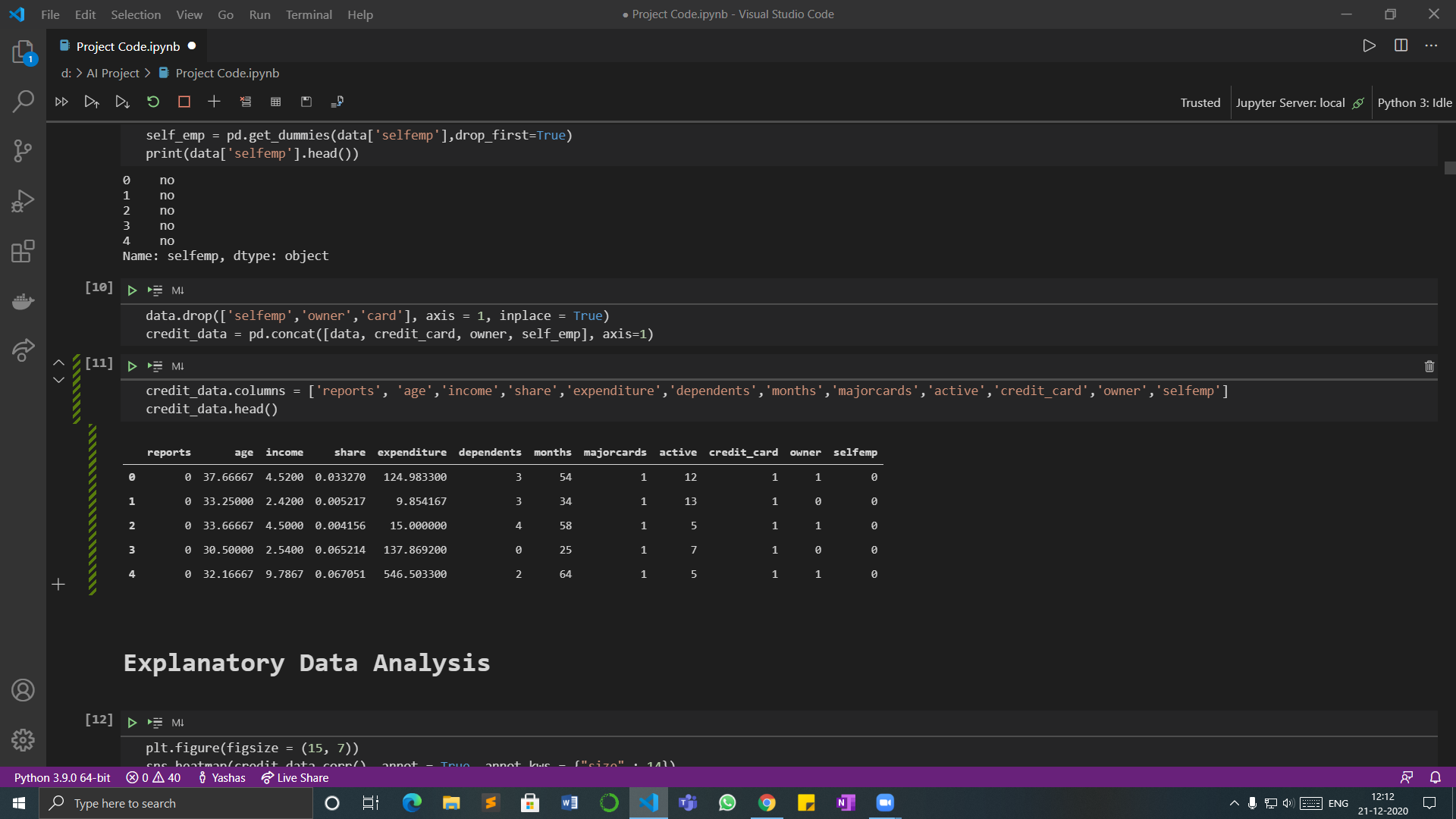
## **Searching for Null Values**

We searched for null values in all the attributes of the dataset using ‘**data.info()**’ method



* Hence, There are no null values present in the data.
* The dataset contains **Categorical** and **Non-Categorical** variables mixed.
* The **card** attribute being the target attribute, the owner attribute and the **selfemp** attribute have values in 2 classes that is, Yes or No. So we have converted Yes value into 1 and No value into 0 using ‘get\_dummies’ method

The following is the first 5 rows of the dataset after converting into 1’s and 0’s :-

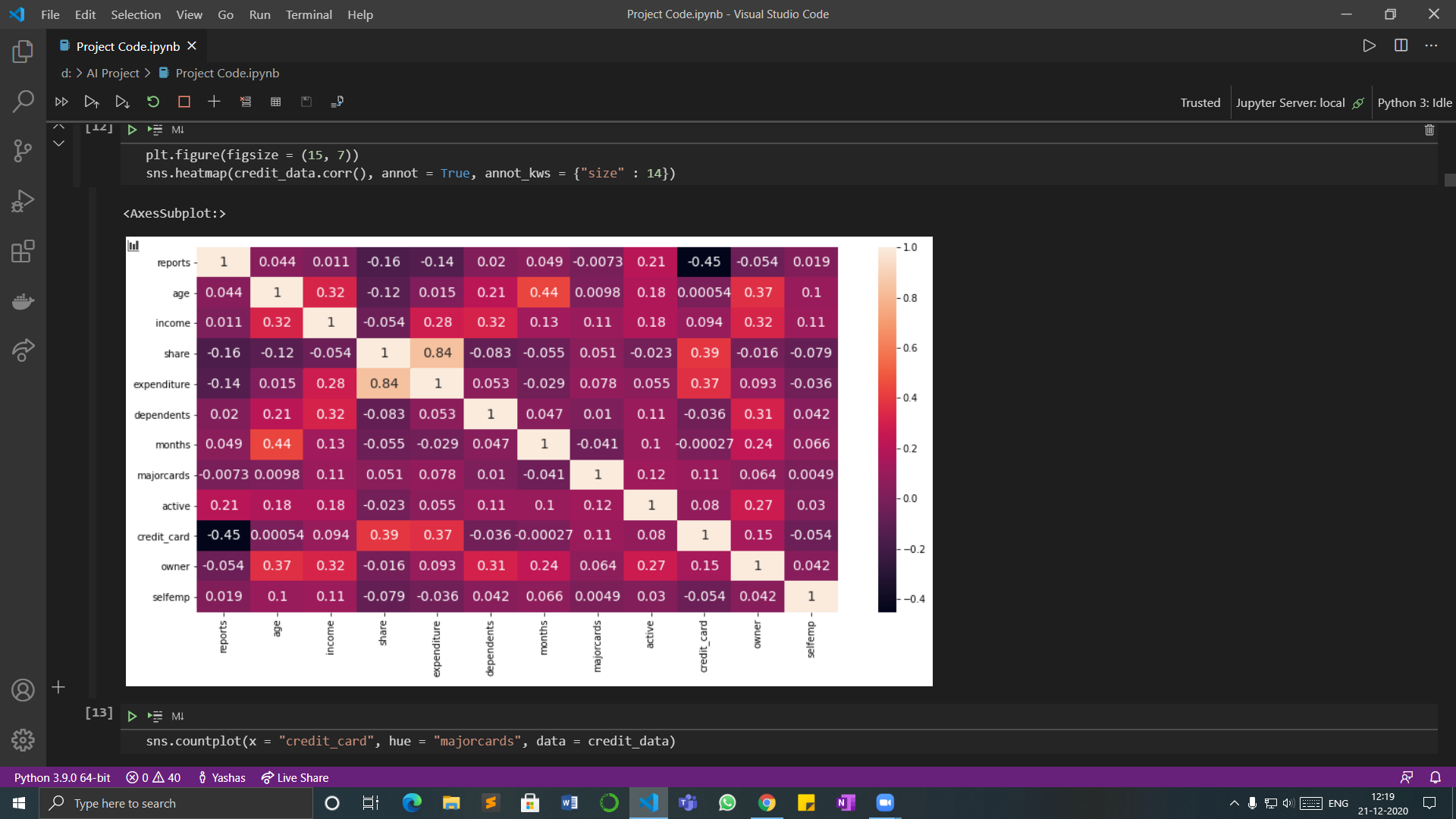


# **Exploratory Data Analysis**

Next, we have plotted some graphs for visualisation, analysis and studying the data

## **Heatmap for correlation of attributes**

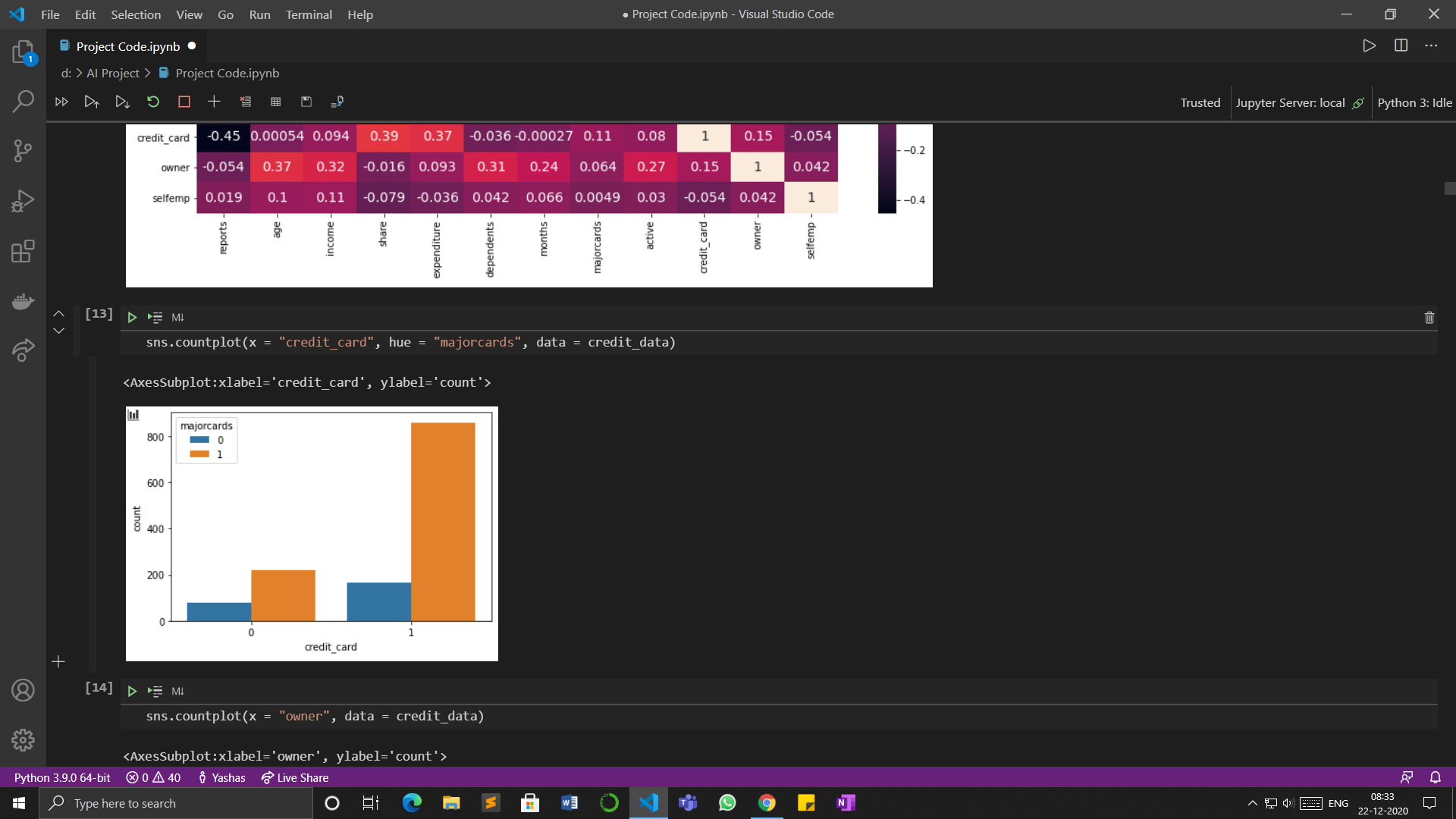
* A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors
* Below is the heatmap on the basis of correlation of each attribute to understand which attributes are more related to each, primarily which attribute is the target variable more related to, and which attributes are least related to each other.
* The value of correlation lies between -1 to 1, -1 being the least related and 1 being the most related

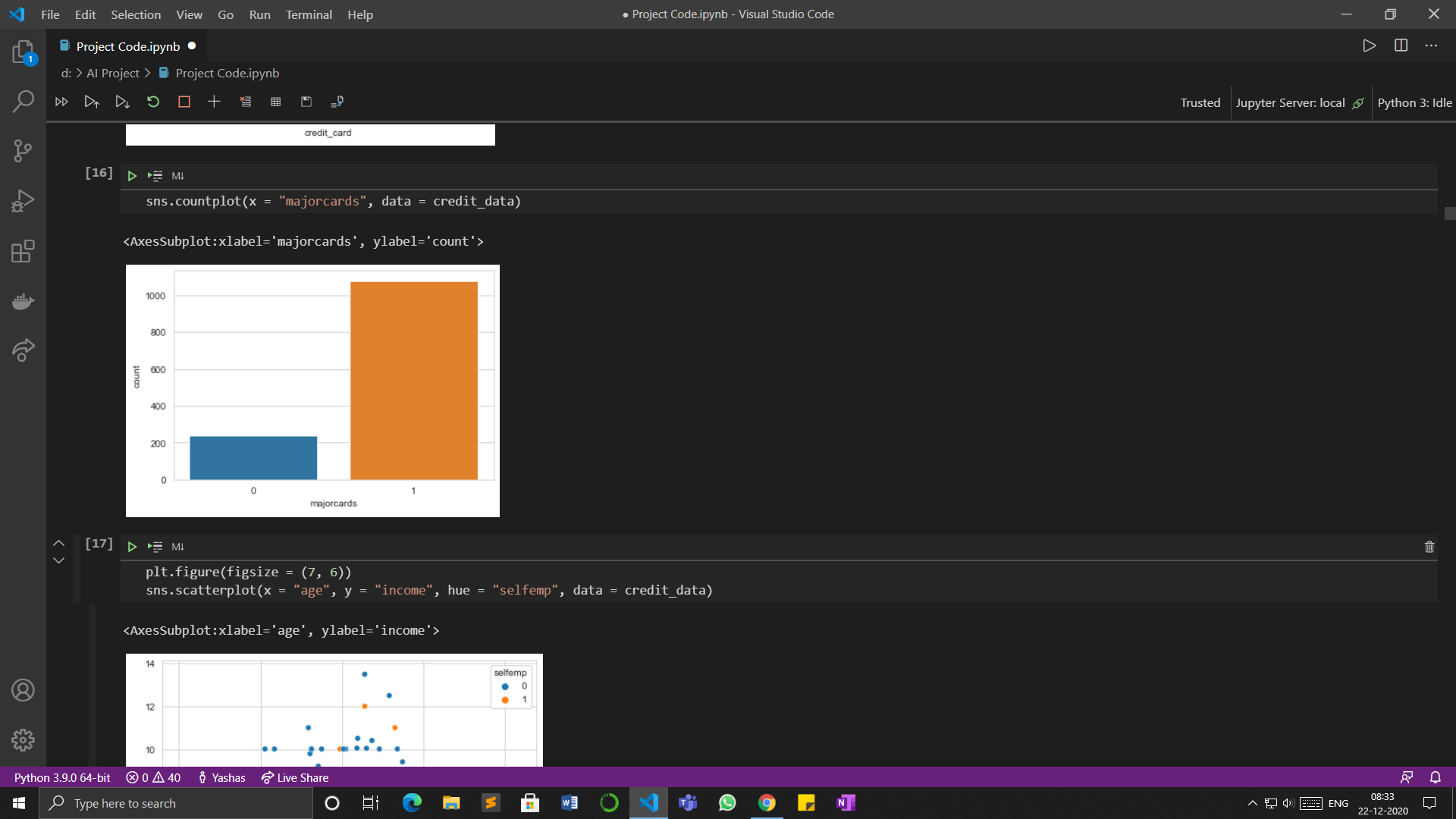
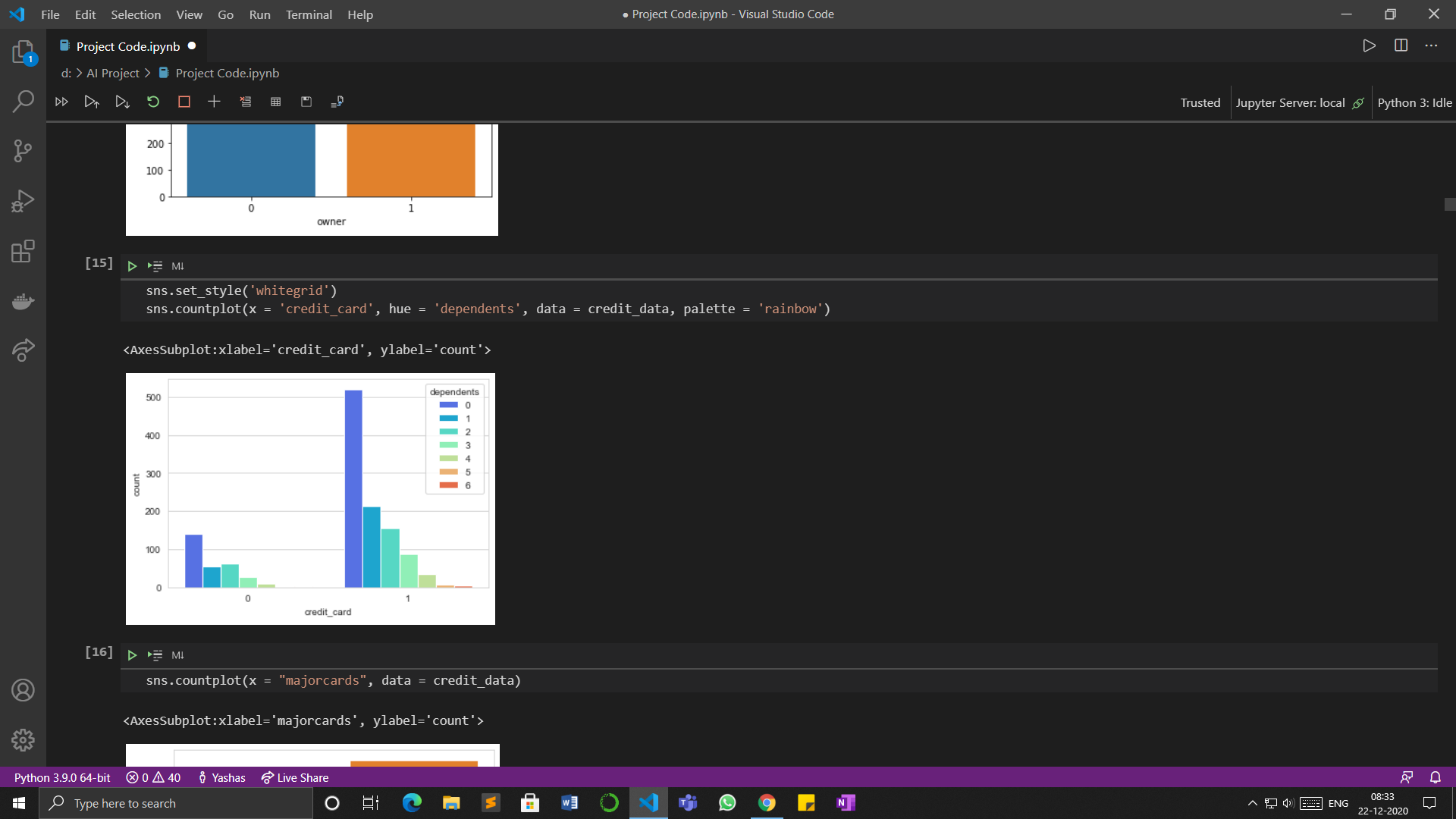
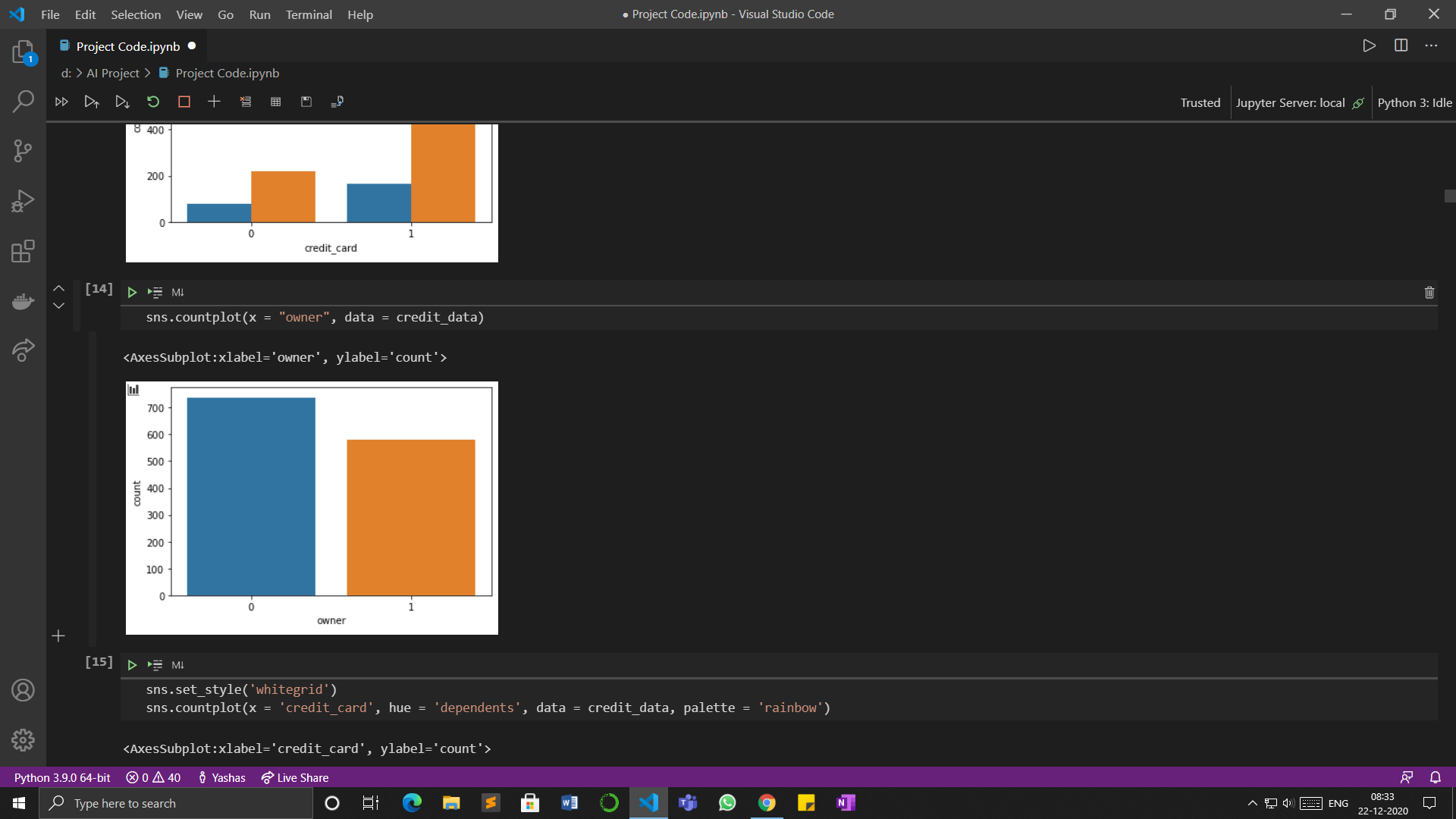


* From this plot, we can conclude that the target attribute card or credit\_card is more related to share attribute and the expenditure attribute and least related to reports attribute
* The expenditure attribute and share attribute are highly related to each other
* We can also conclude that age attribute, income attribute and owner attribute are comparatively more prominent or more related to other attributes

## **Count Plots**

We have plotted **count plots** for categorical variables to find out how many values are present in each class.

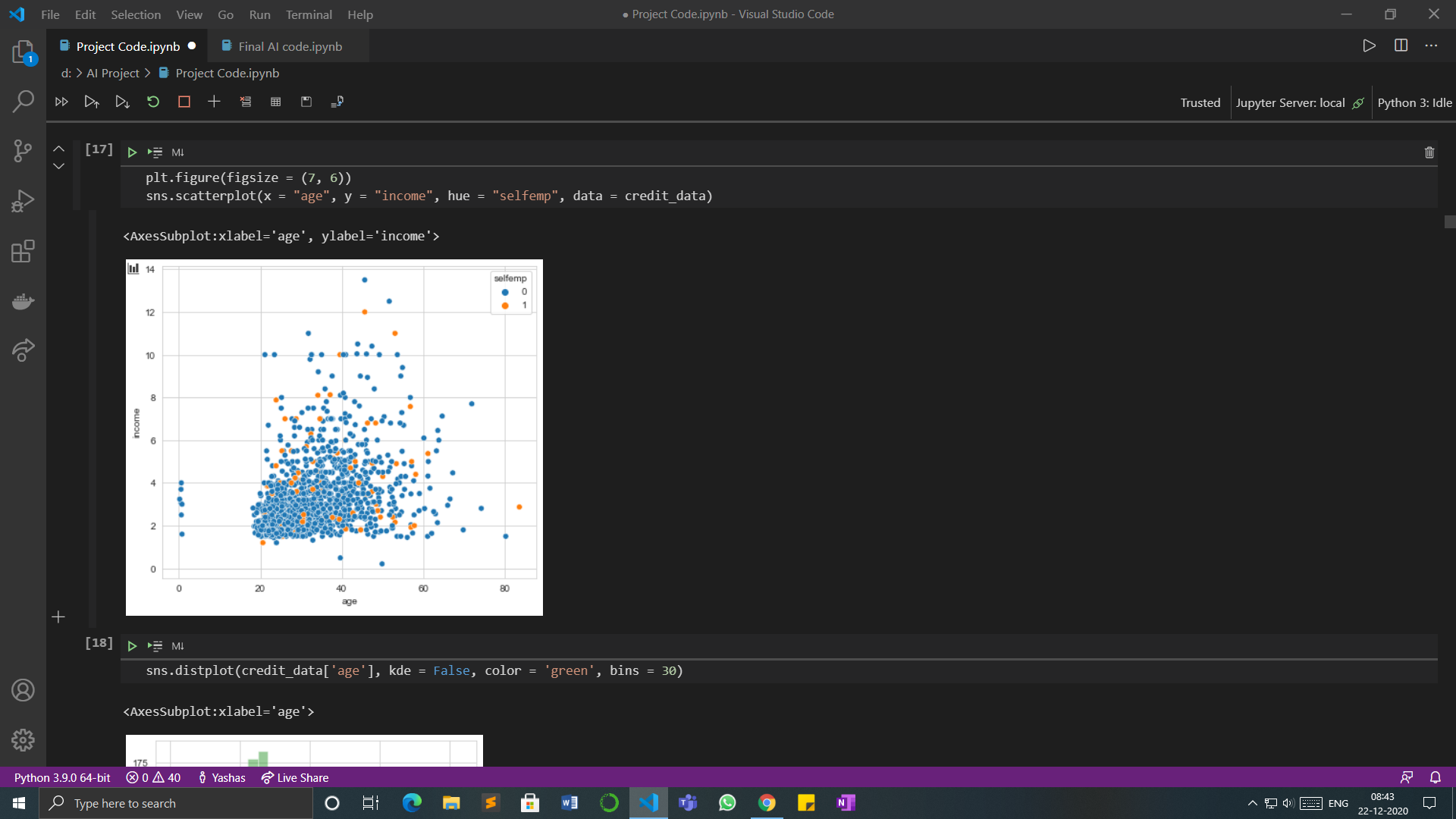




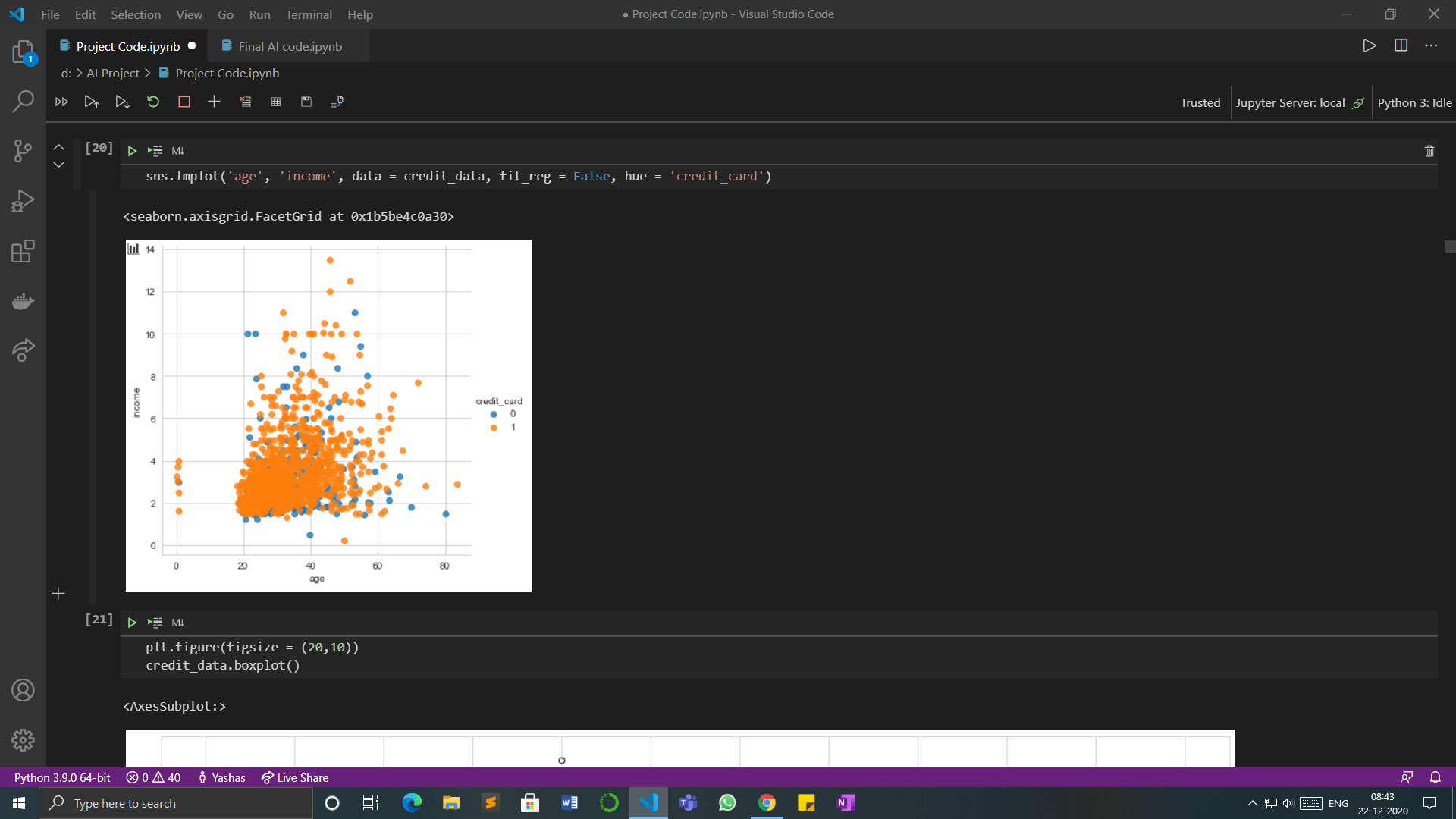
## **Scatter Plot and LM plot**

Below are the scatterplot and lmplot for age and income based on **selfemp** and **credit\_card**

* A **scatter plot** is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.

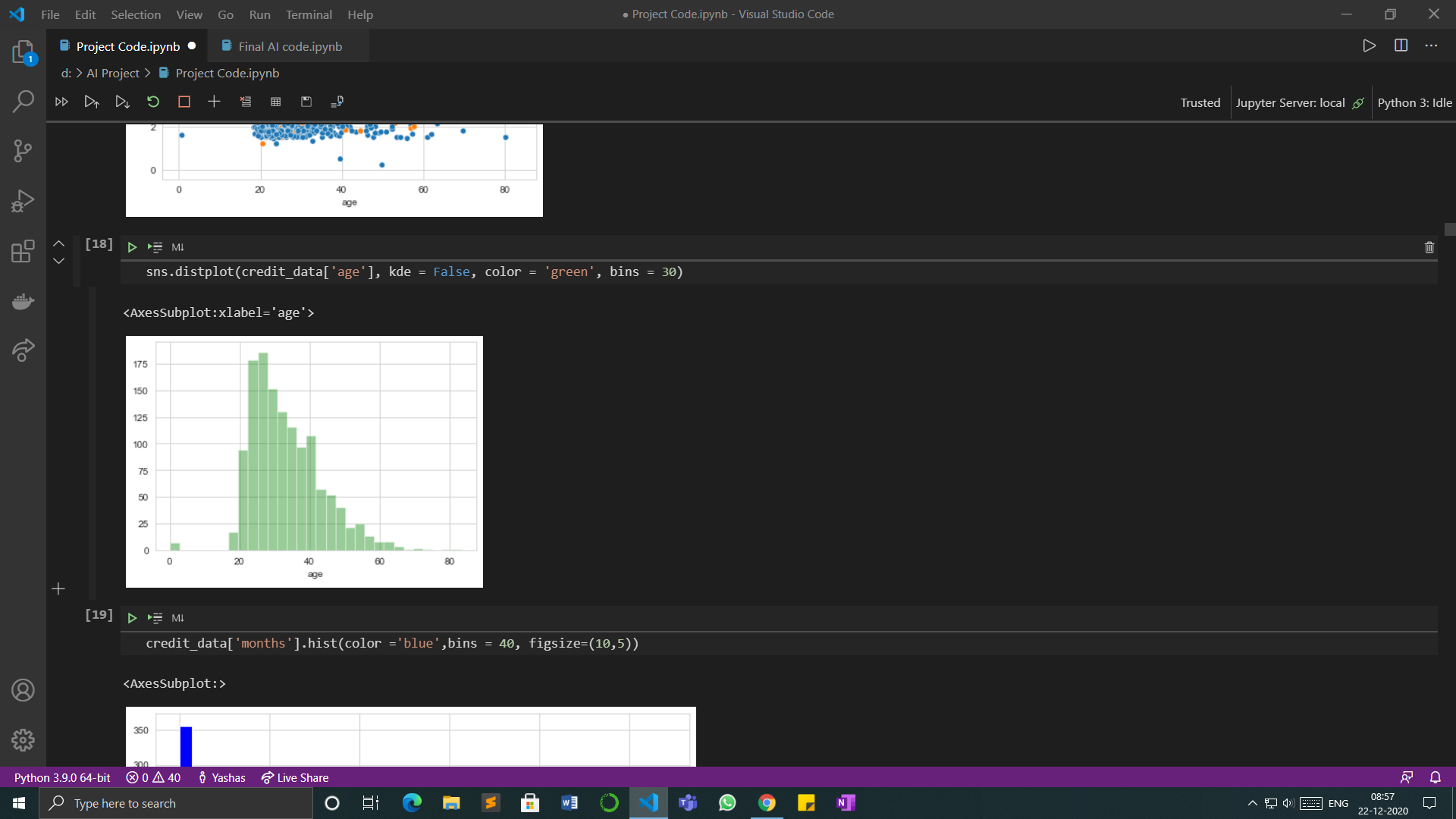


* A **lmplot** is a 2D scatterplot with an optional overlaid regression line.

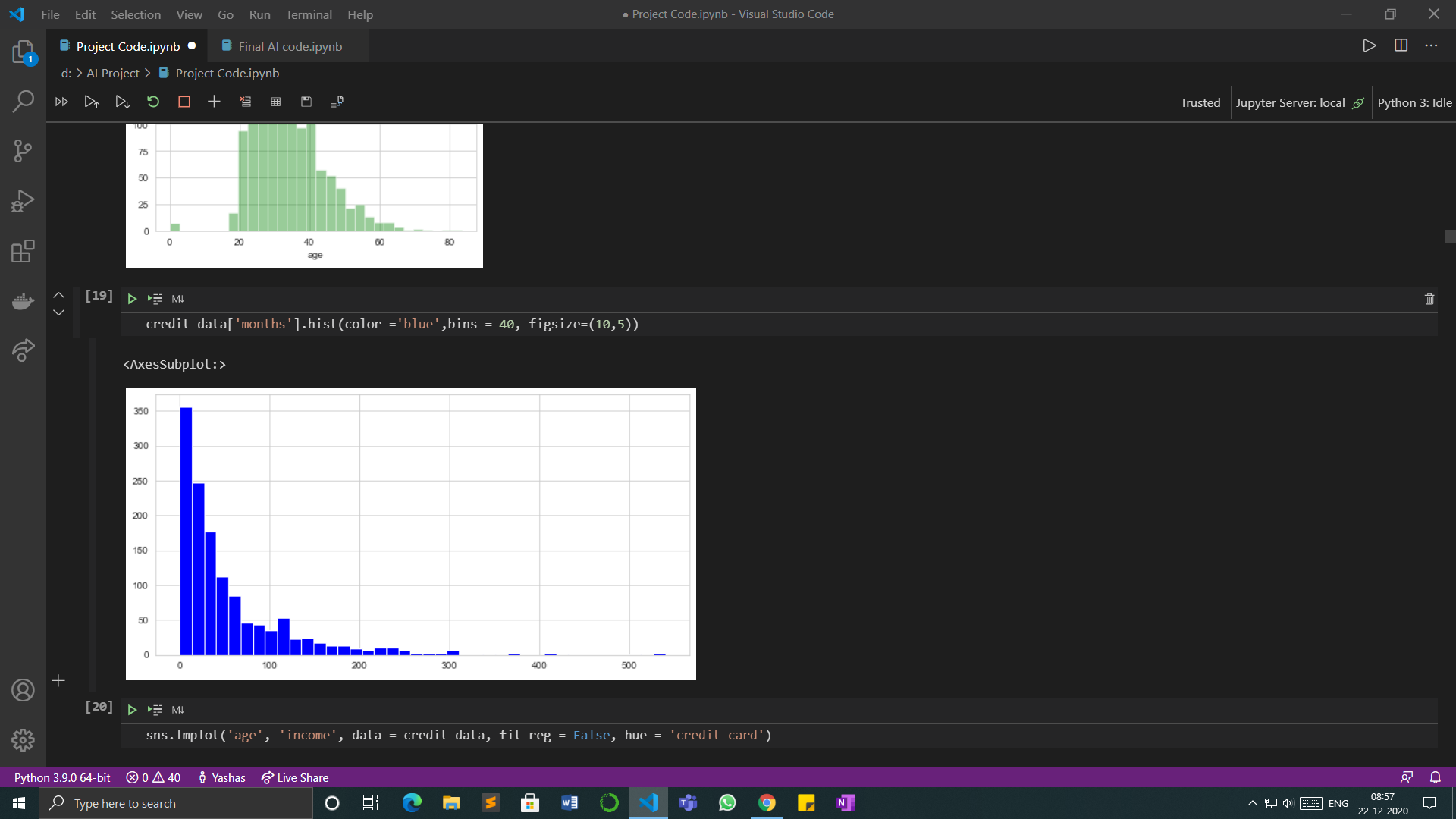


## **Distplot and Histogram**

* A Distplot or distribution plot, depicts the variation in the data distribution.



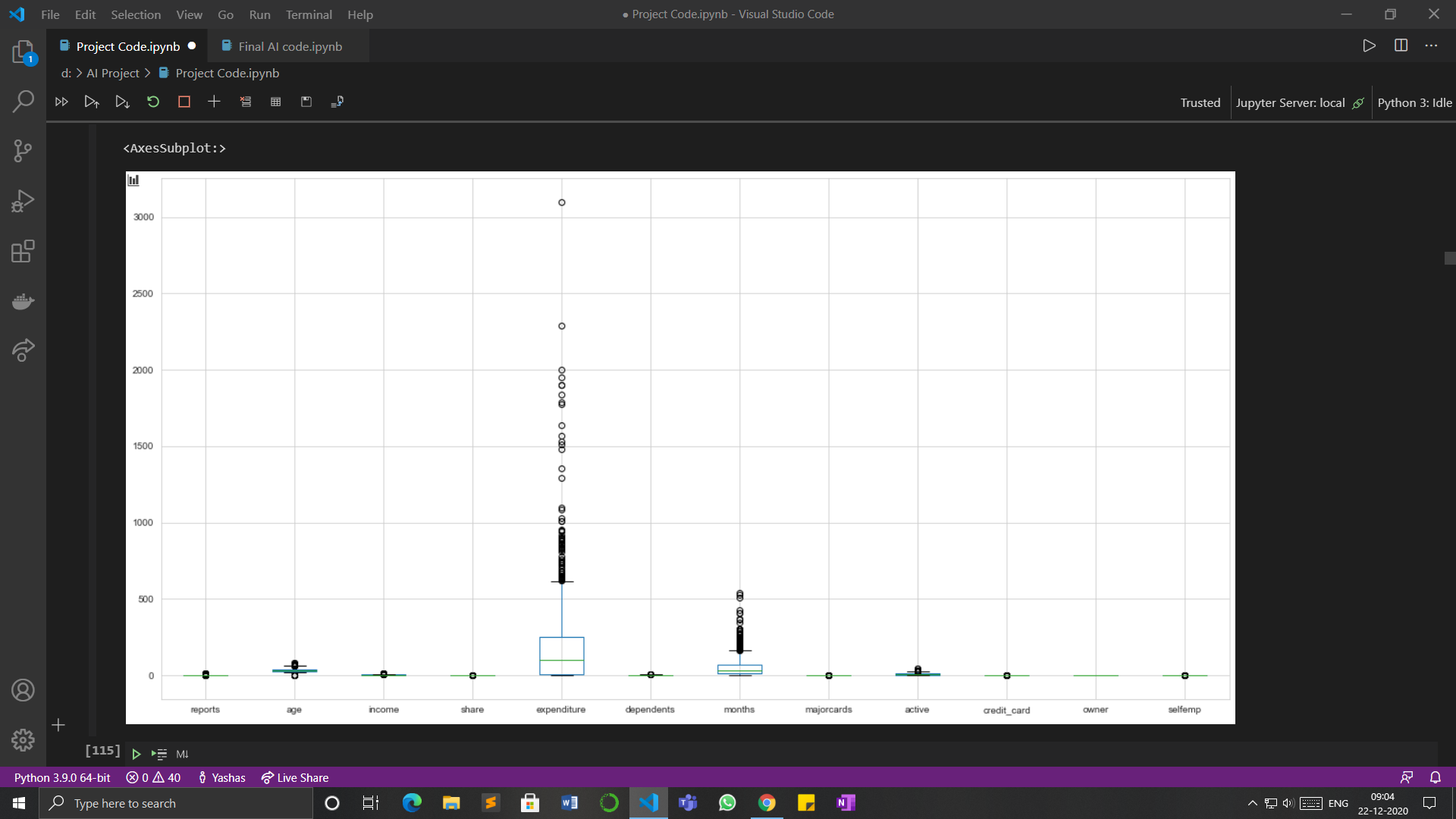
* A **Histogram** is a graphical display of data using bars of different heights. It is similar to a Bar Chart, but a histogram groups numbers into ranges. The height of each bar shows how many fall into each range.

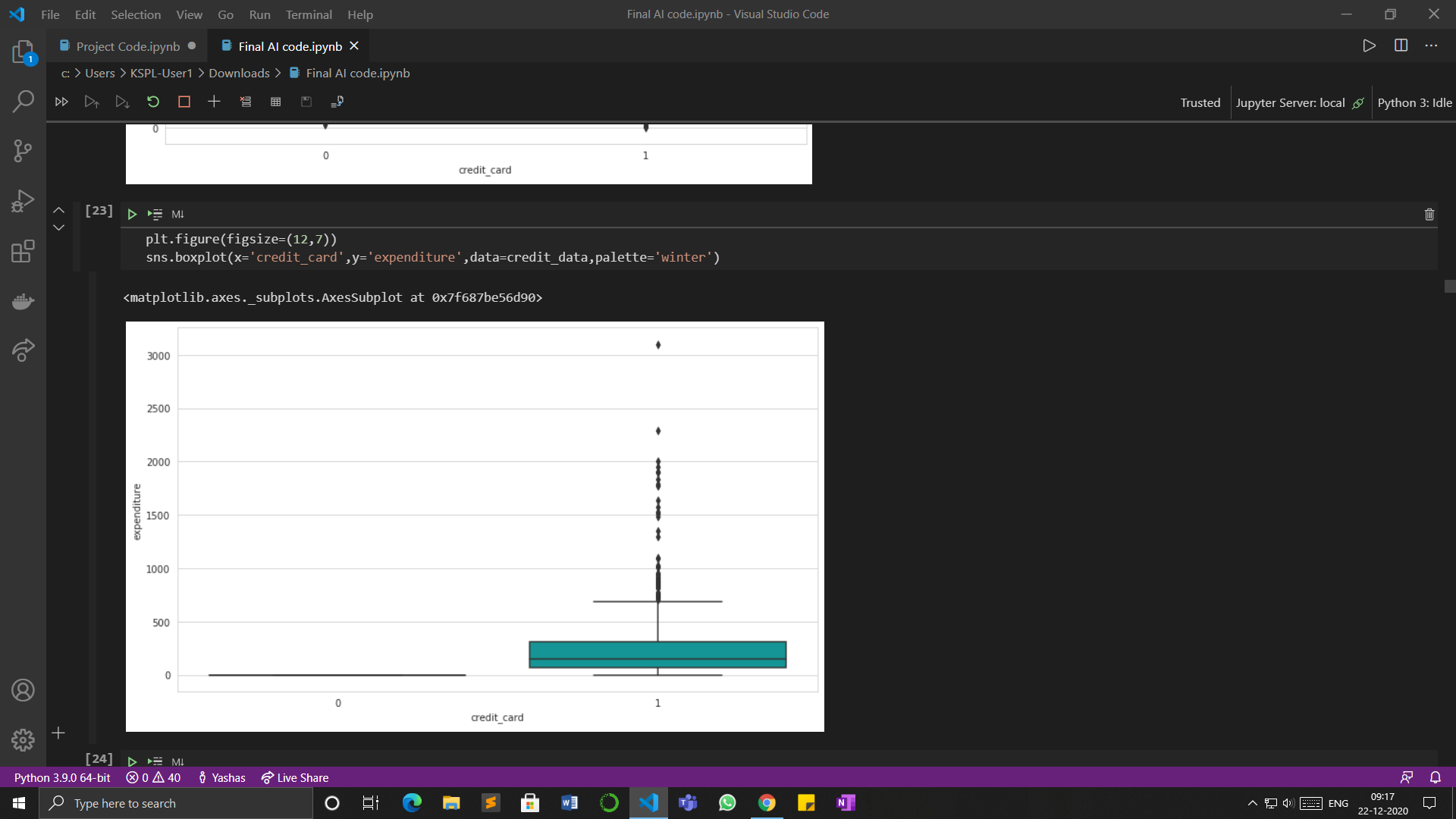


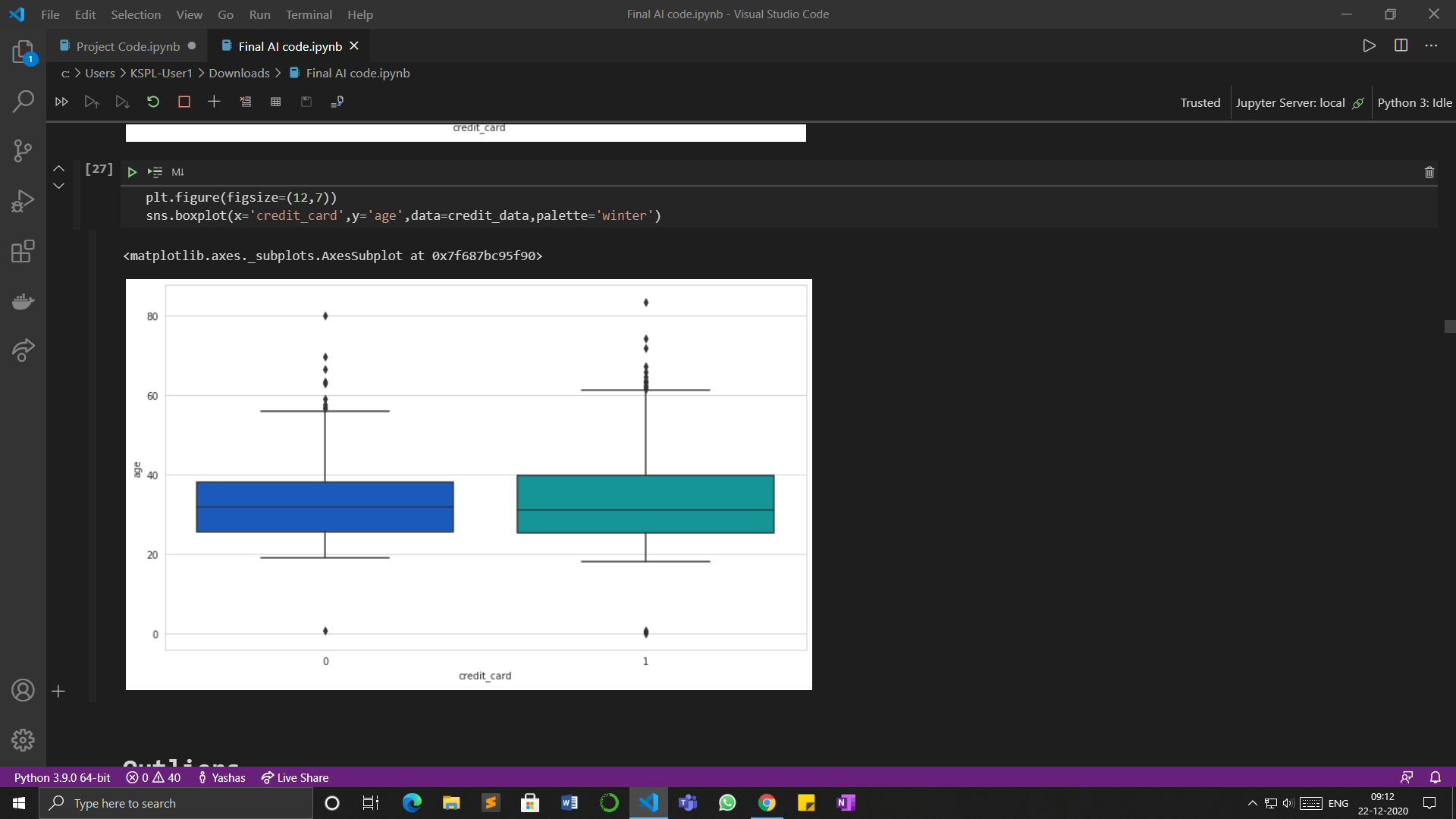
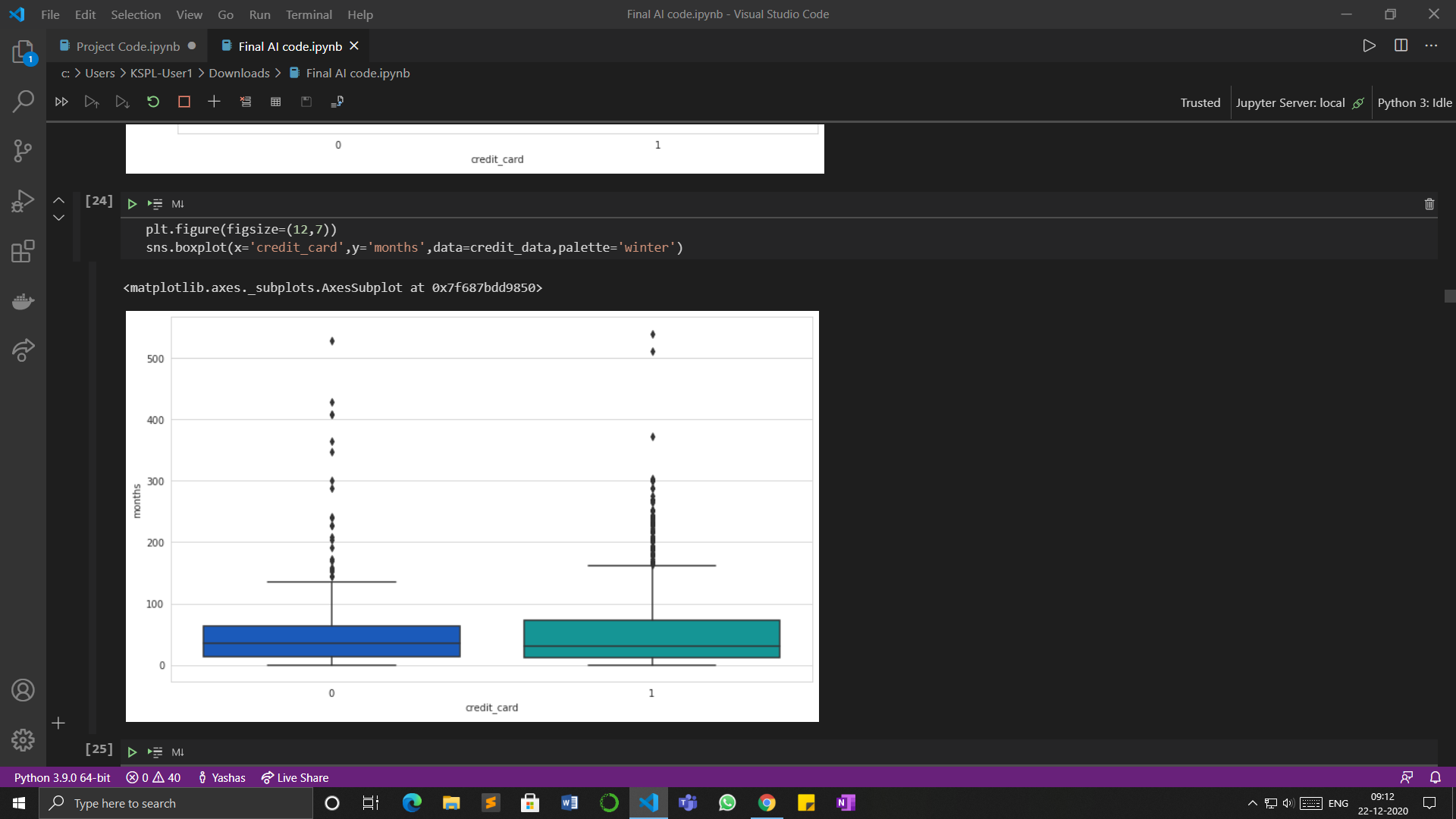
## **Box Plots**

* Boxplot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending from the boxes indicating variability outside the upper and lower quartiles. Here, those lines depict the existence of outliers in the data

Below is the boxplot of all the attributes in the dataset combined.







* From the above plots we can conclude that the expenditure, months and age have outliers.

# **Data Preprocessing**

## **Outliers**

* An **outlier** is an observation that lies an abnormal distance from other values in a random sample from a population. Examination of the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers.
* The **box plot** is a useful graphical display for describing the behavior of the data in the middle as well as at the ends of the distributions. The box plot uses the median and the lower and upper quartiles (defined as the 25th and 75th percentiles). If the lower quartile is Q1 and the upper quartile is Q3, then the difference (Q3 - Q1) is called the interquartile range or IQ.
* **Outliers** increase the variability in your data, which decreases statistical power. Consequently, excluding outliers can cause your results to become statistically significant which improves the model efficiency

## **Dealing with Outliers**