**LAB 1**

1. Examine each attribute in the data file (vTargetMailCustomer.csv ) to select a set of Features (Attributes) that would affect to predict future bike buyers. Remove all the unnecessary attributes from the data file VTargetBuyerMailList to Create a new file VTargetBuyers with the selected Features only. Include ID and the Class column (BikeBuyer) in your file. However, the class attribute BikeBuyer should NOT be included for the rest of the data processing and the calculation below.

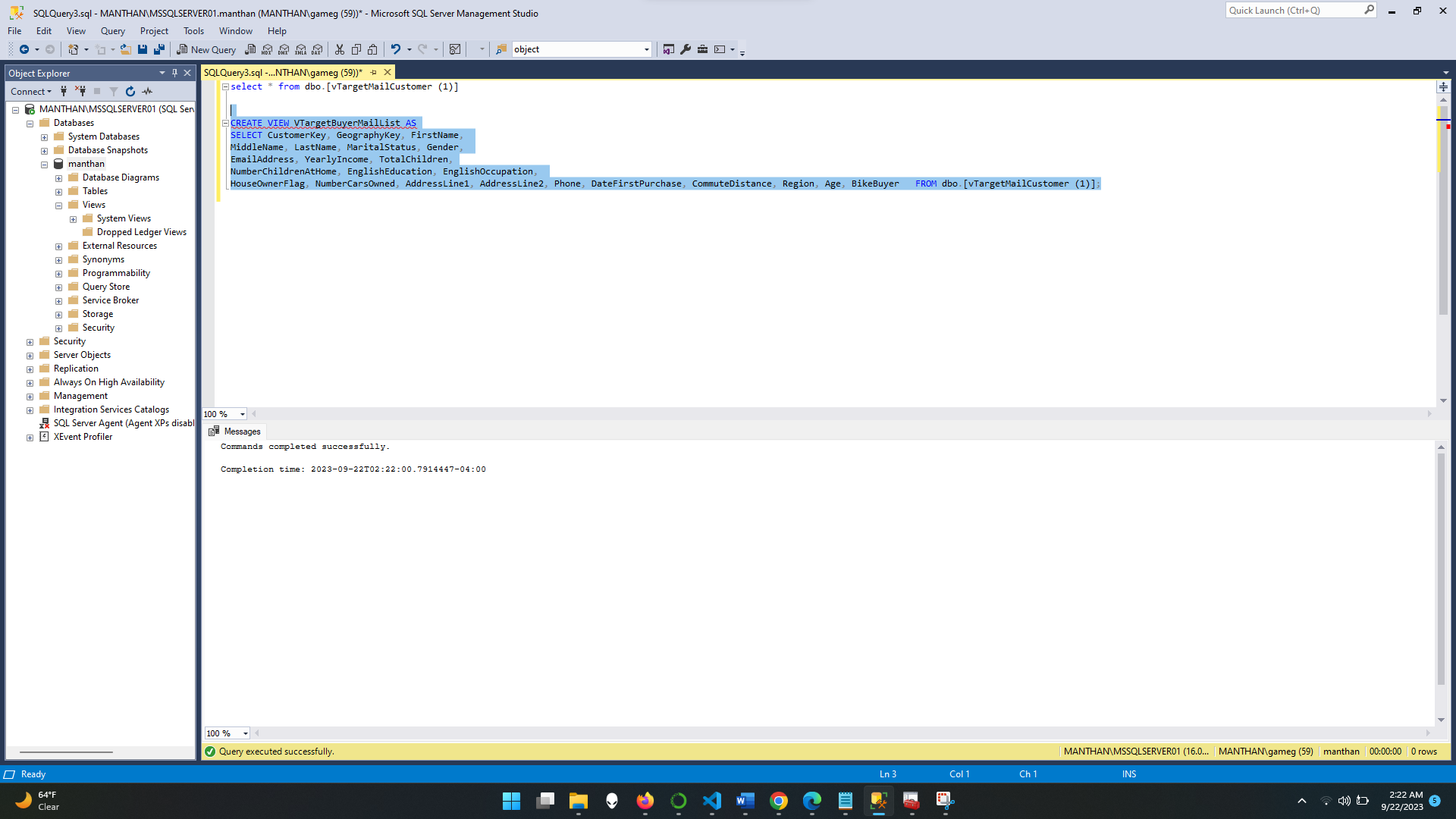
Part 1:

* First, I install ssms in my system to perform lab 1
* Now I created database on object explorer
* I added the CSV file in the database
* After that running this query, I get the data from csv file

A screenshot of a computer

Description automatically generated

* Over here I created a new view and created new CSV file of vTargetBuyer to perform all the tasks I want on the CSV file with the Jupiter notebook in my visual studio and I am going to use Python language for that.



A computer screen with text

Description automatically generated

A computer screen with text boxes

Description automatically generated

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

2. . For each selected feature in your selection, determine the Properties of data covered in class as follow:

1. All the Properties Whether it is Nominal, Ordinal, Interval, or Ratio
2. Data Type whether it is Discrete, or Continuous

|  |  |
| --- | --- |
| Attributes | Types Of Attributes |
| CustomerKey | Continuous |
| MaritalStatus | Binary / Discrete |
| Gender | Binary / Discrete |
| YearlyIncome | Continuous |
| TotalChildren | Discrete |
| NumberChildrenAtHome | Discrete |
| HouseOwnerFlag | Binary / Discrete |
| NumberCarsOwned | Discrete |
| DateFirstPurchase | Discrete |
| CommuteDistance | Discrete |
| Region | Nominal / Discrete |
| Age | Discrete |

**Data Preprocessing and Transformation**

Use all the data rows (~= 18000 rows) with the selected features as an input file to apply the tasks below, do not perform each task on the smaller data set that you got from your random sampling.

Platform - vscode

Python editer – Jupiter in vscode

* Here I am going to perform all the tasks on the data file name as (vTargetBuyers.csv)
* First, I import libraries to work on this data
* The I stored the data in the variable called vTargetBuyers to work on that in the future.
* Now I use pandas command head() to get the information of data but over here I want 18600 rows to work with it, so I write that head(18600)

A screenshot of a computer

Description automatically generated

**Handling Null Values to Replace**

* Here I want null values replacement so first I used null to find null values so it will give answer in true and falls so it’s easy to find nulls and replace that on the next step.

A screenshot of a computer

Description automatically generated

**• Perform Random Sampling, Bootstrap Sampling with Replacement**

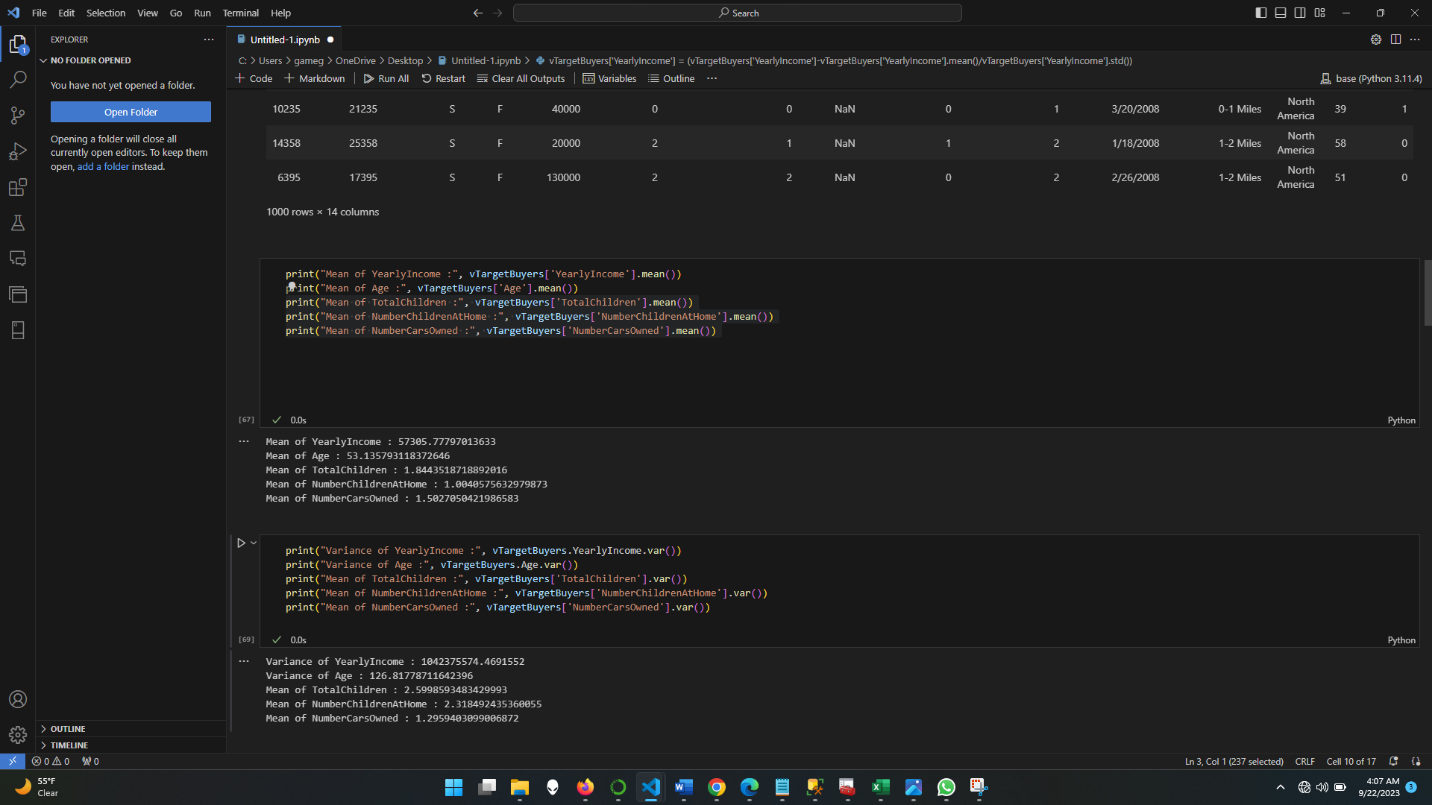
Where I take the random data of 100 lines from the code and use NumPy random function to choose random lines from to understand that data because we discussed in the class about this topic

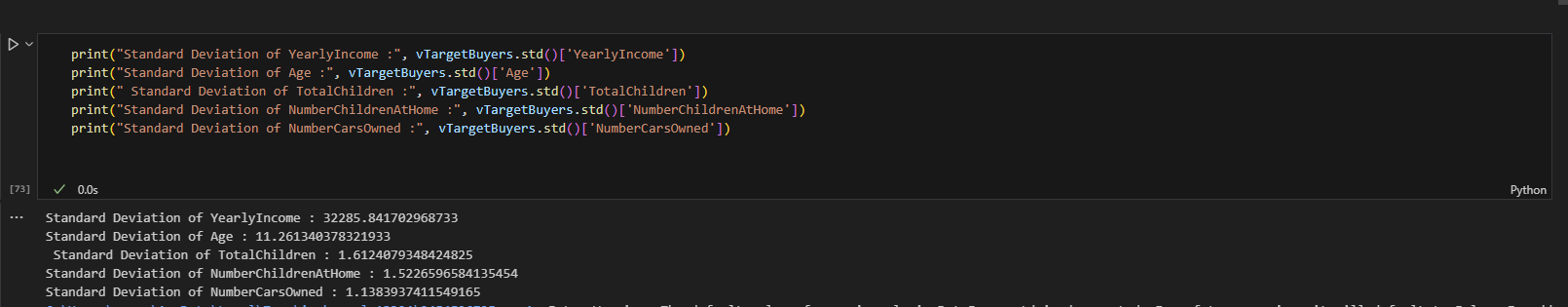
A screenshot of a computer

Description automatically generated

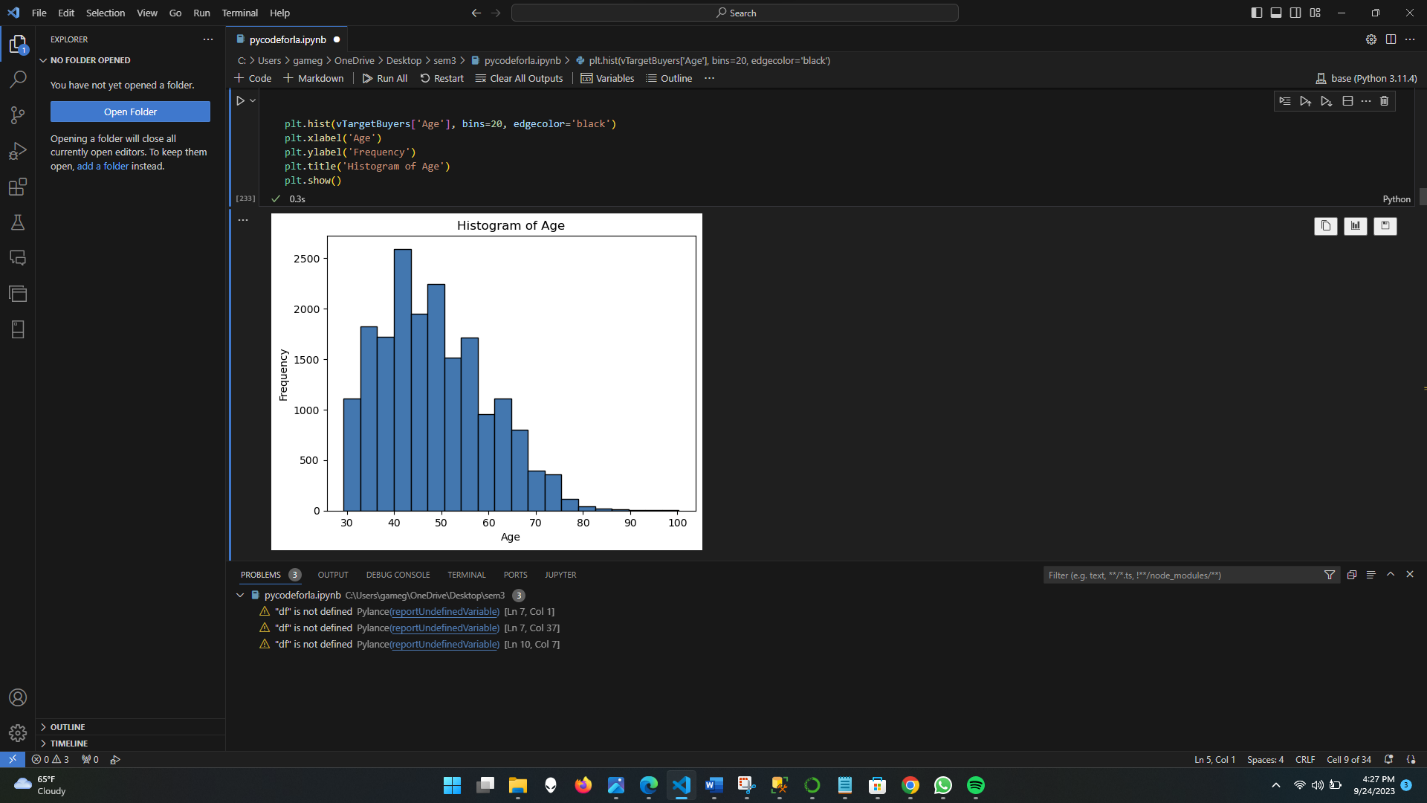
**• Mean/Variance/Standard Deviation for Ordinal Numeric attributes.**

So, I can better understand this data and better and give the overall range of this information to get achieve future goals. Also, it helps to find outliers and reducing the distraction in the data to achieve desire insides and narrow the search, so it is process of data cleaning





**• Identify Data Dispersion (Distribution) for each feature using one of the methods: Box plot, Histogram, or z-score visualization.**



The majority of customers are most importantly 30-60 where we can focus and the majority of that people have the experience and this market could be on intrest and need base becose if someone has the lower salary and that person's age is under 50 that could be a good target who can purchase or other market where person has good salary and everything but they like to ride then that could be a potential customer

A screenshot of a computer screen

Description automatically generated

It represents the number of children and it shows that the majority of people have 0 and the second heist has 1-2 so those could be the potential customer for kid bikes.



This shows the people who have children 1-2 but most of the child is not at home so those could be a key point for the purchasing as parents and this feature shows that focusing on people who have children could be a bad idea

A screenshot of a graph

Description automatically generated

This graph shoes that around 40% people have the 1-2 cars and 20-25% people have the more then 2 cars so those could be customers but not that much important on this section

A screenshot of a computer

Description automatically generated

Over here as per yearly salary data I can understand that the majority of people have the 35 to 50k salary on average also this rage along with the 60 to 75k salary early so this gonna be the most number of people and I thing 60 -75k is easy to convince if they have pervious purchase and if they have the interest on riding because they can spent money on that

Example of understanding and problem

If we can get the product to target this group then that would be a biggest market also this market is more varsity becouse of money and it could be more challenging due to affordable option for company production dn marketing the main reason is they can take too many profits or no extra money to spent on marketing

**• Identify Outliers with 1.5IQR Method to Replace**

In this section I get the IQR value of data where it is the highest lowest and mid of the data also it is found high and low are averages of that data where I get the information about Q1,Q2, Q3 data

A screenshot of a computer program

Description automatically generated

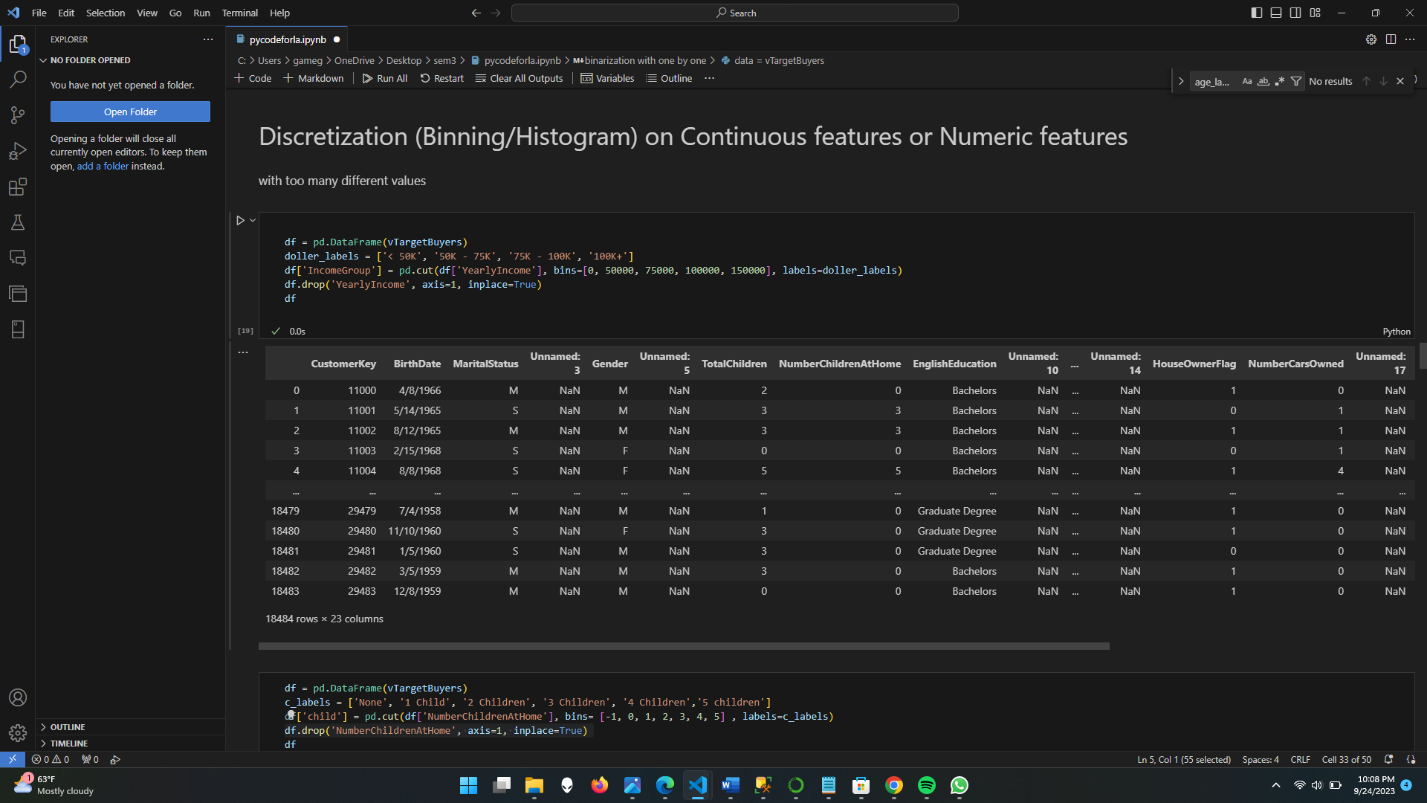
A screenshot of a computer

Description automatically generated

**Discretization (Binning/Histogram) on Continuous features or Numeric features with too many different values**

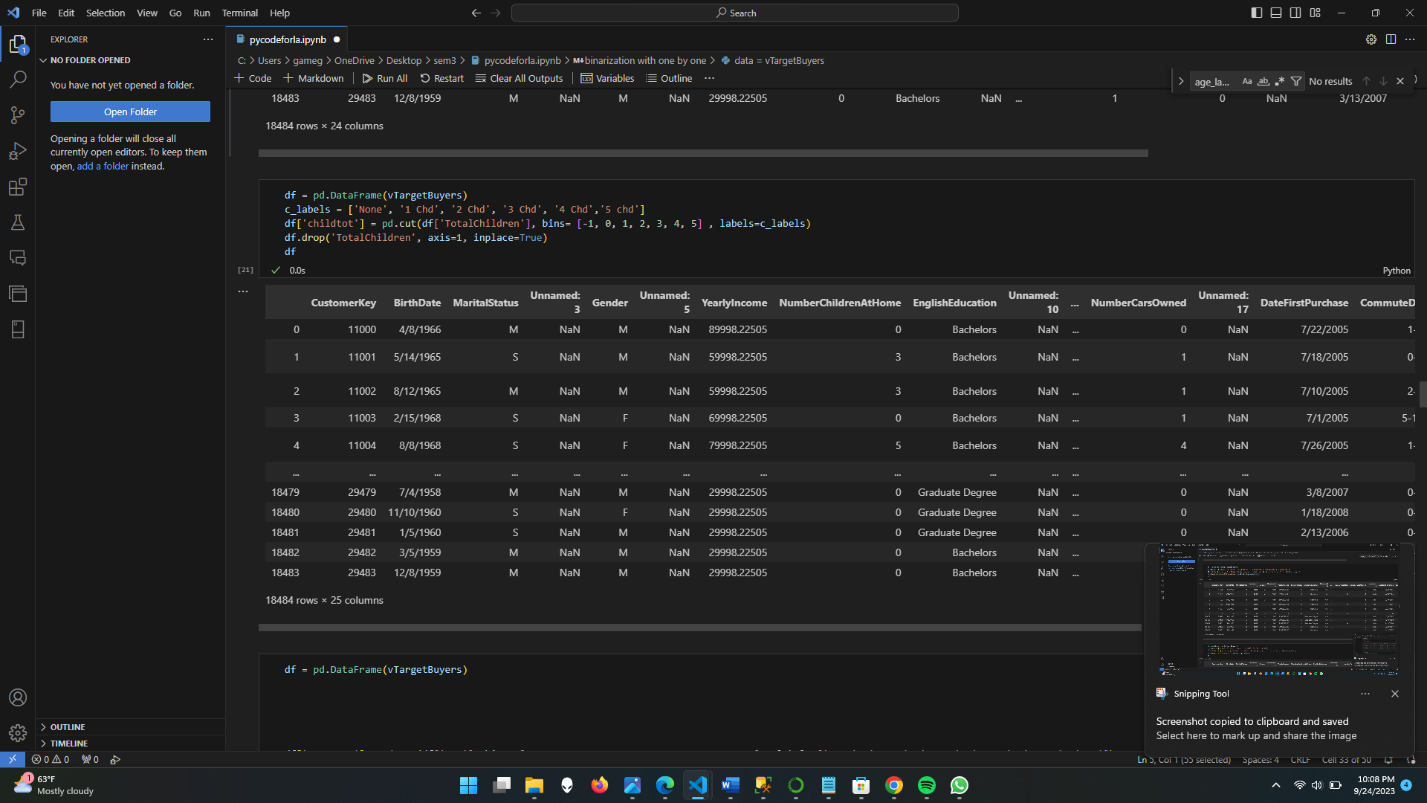
Over here I created a group for getting data to minimize information to understand the data and also give you overall information about data what’s going on and drop that data in the columns to use that data again.

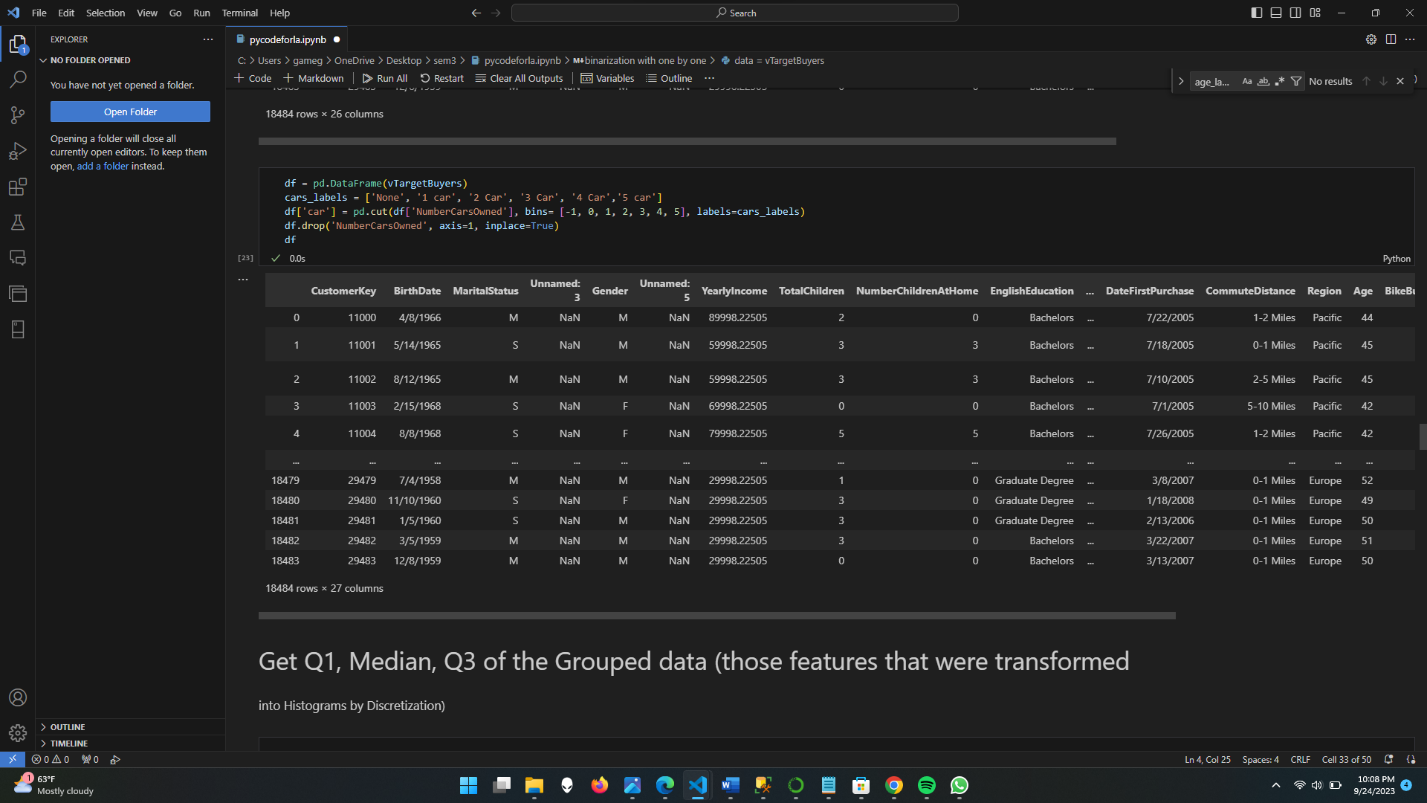
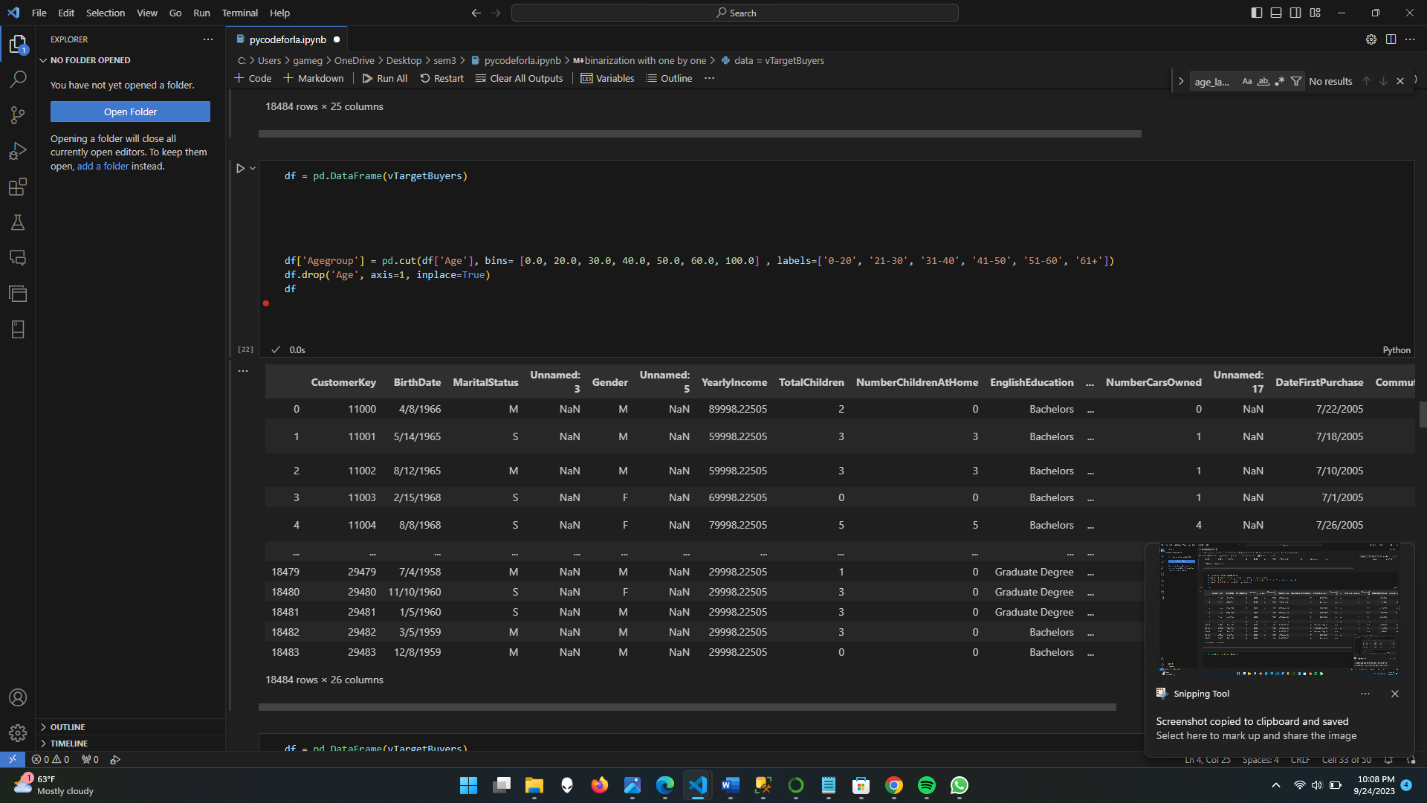
Create a grouping and adding the that values in range and grouping



A screenshot of a computer

Description automatically generated

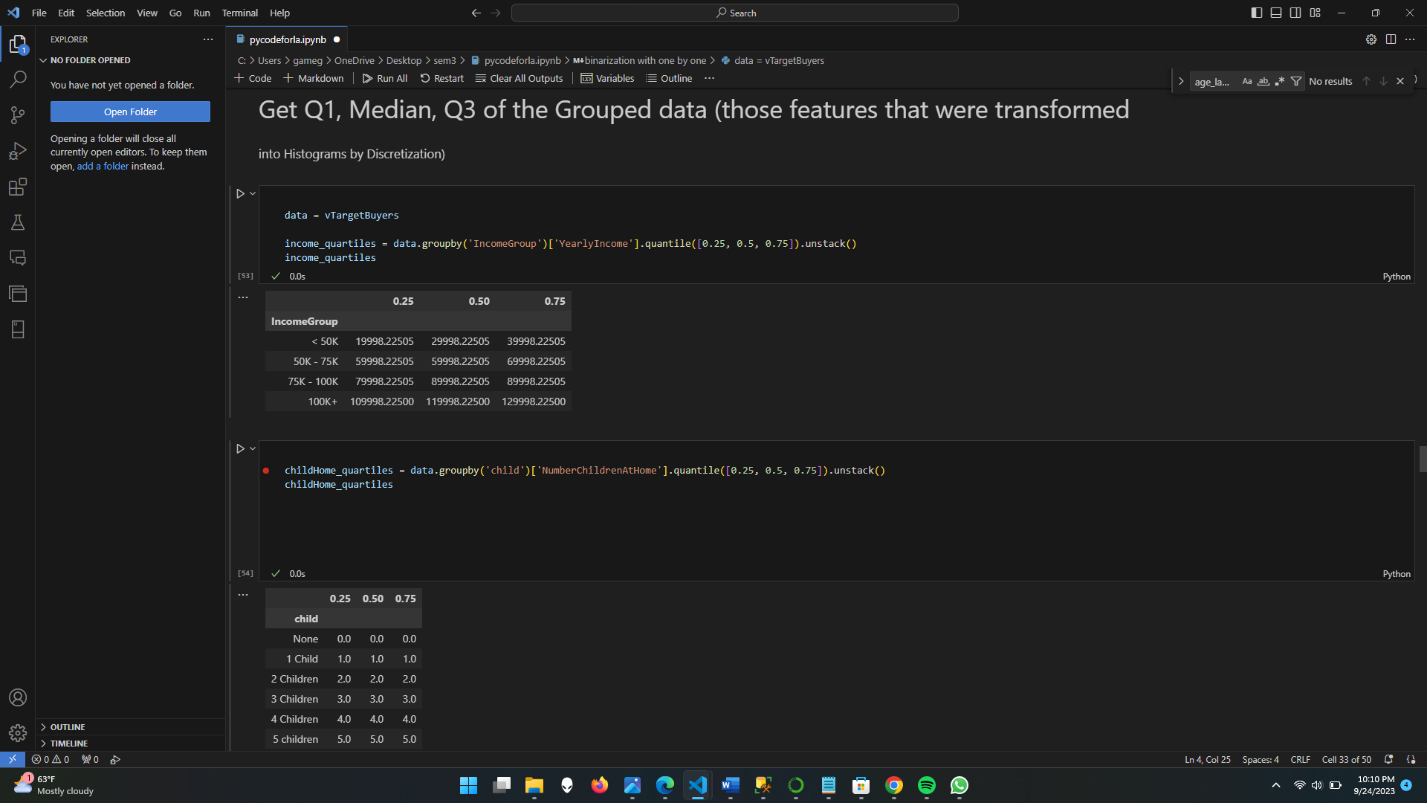




**Get Q1, Median, and Q3 of the Grouped data (those features that were transformed into Histograms by Discretization)**

In this section I did the following Q1,Q2, Q3Of grouped data, I easily get the insides how many people are under what area, also, it is important to understand the range of data to get the proper information in each area of features

Also, it helps to find outliers.

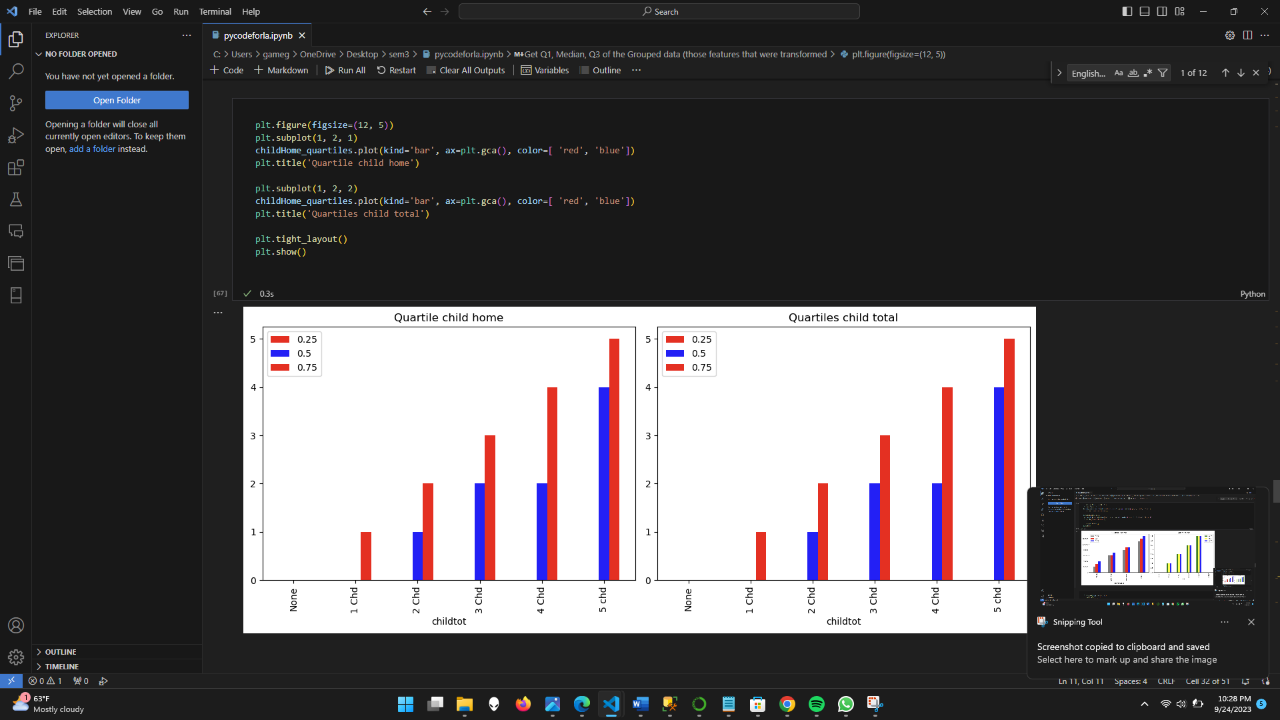


A screenshot of a computer

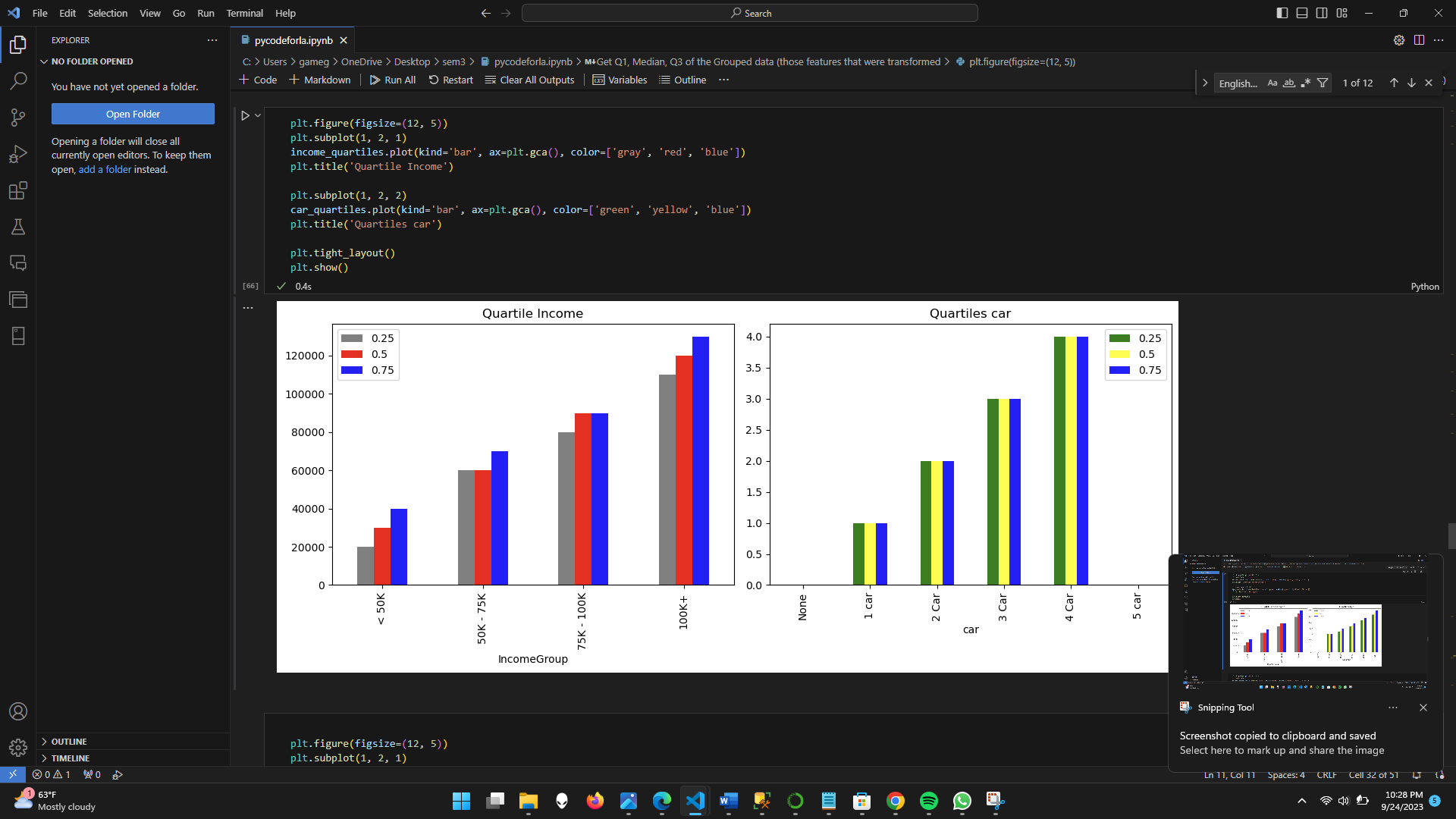
Description automatically generated

A screenshot of a computer

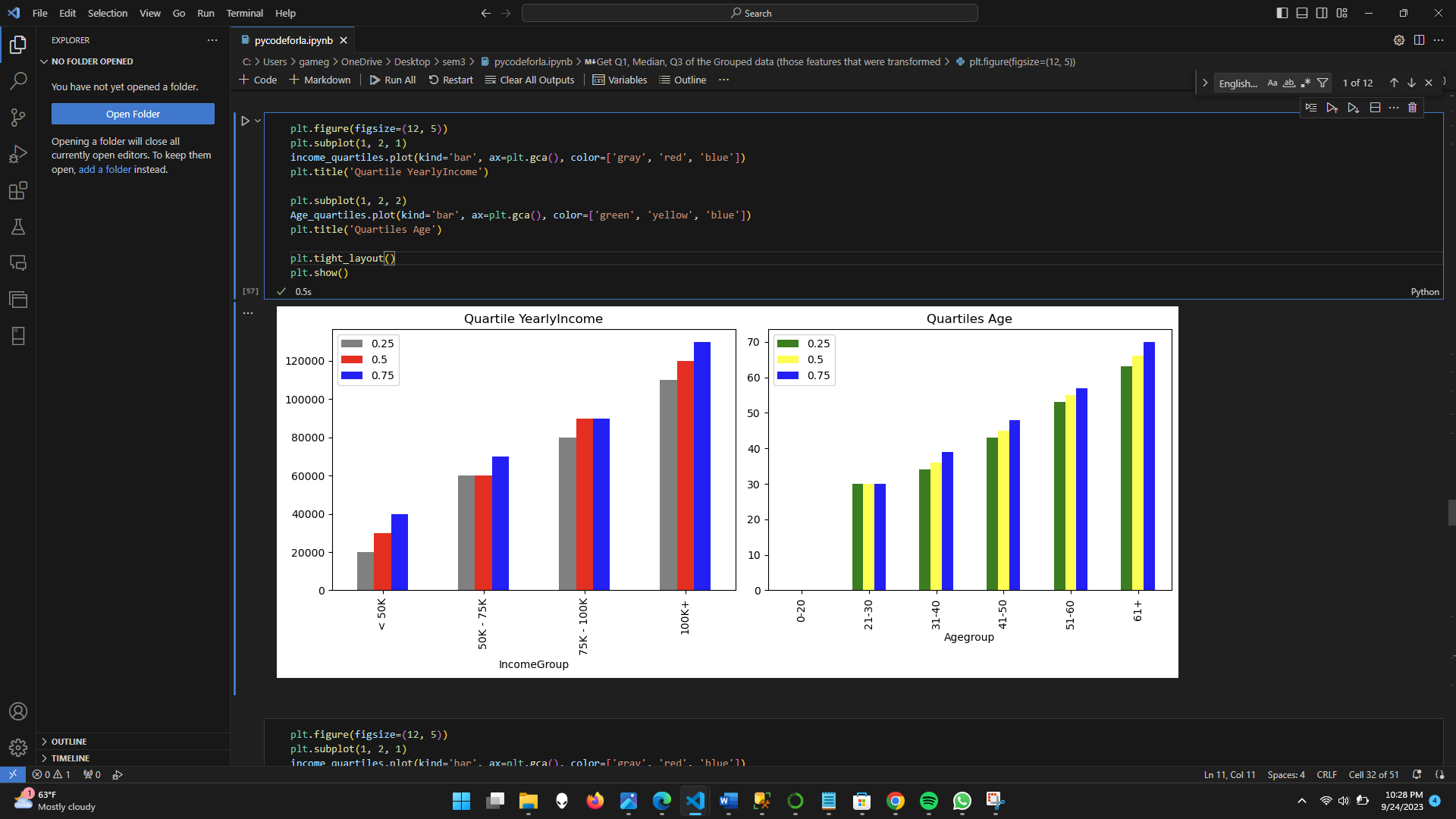
Description automatically generated



Over here I did some graphing of quartile to represent that my data cleaning methods show that high-income groups are most likely to have a high number of cars



Over here this graph shows that income group has a co-relation with age so according to that 31 to 51 age is best people to target for the bike purchase because they have income and age to ride .



* **Binarization**

this is used to give the values as 0 and 1 for the shorting also we can use this for machine learning and that helps to find the right answer with the multi features that’s important for machine learning output and that’s accurate with the 0 and 1 value and easy understand for machine learning

A screenshot of a computer

Description automatically generated

**• Normalization or Standardization**

Over here I did normalization of age and yearly income for that I created first mean and the std to find the z-score for this data so first I created the required data and then inserted that value in the equation to find the z-score

And I did that with the income group, age.

It helps to find different data set to achieve the easy comparison between similar kind of data that’s the advantage of the z score

A screenshot of a computer program

Description automatically generated