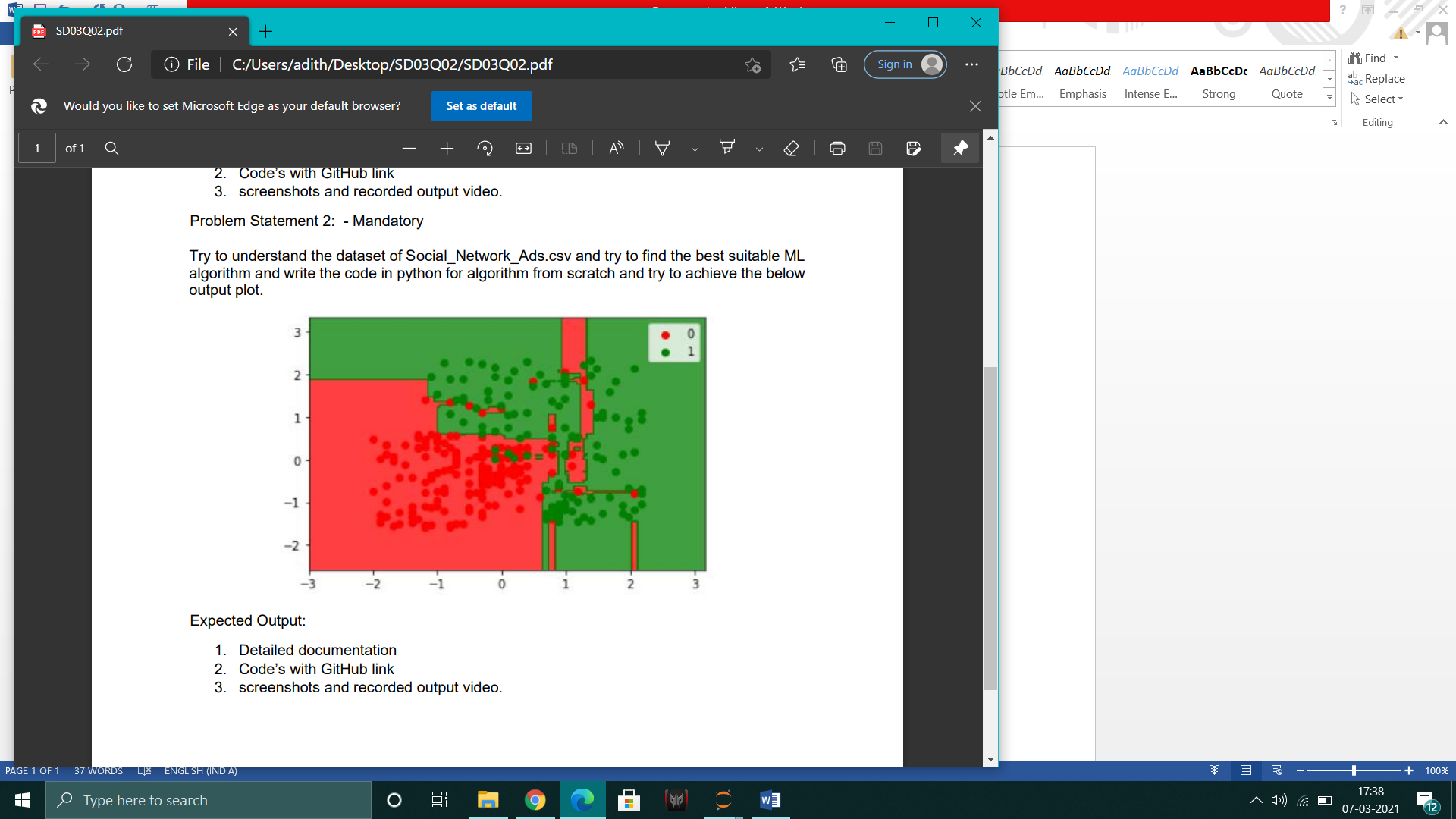
Adithya B

1832002

**DOCUMENTATION FOR SOCIAL NETWORK ADS PROJECT**

**Problem statement 2 :**

Try to understand the dataset of Social\_Network\_Ads.csv and try to find the best suitable ML algorithm and write the code in python for algorithm from scratch and try to achieve the below output plot.



**Solution :**

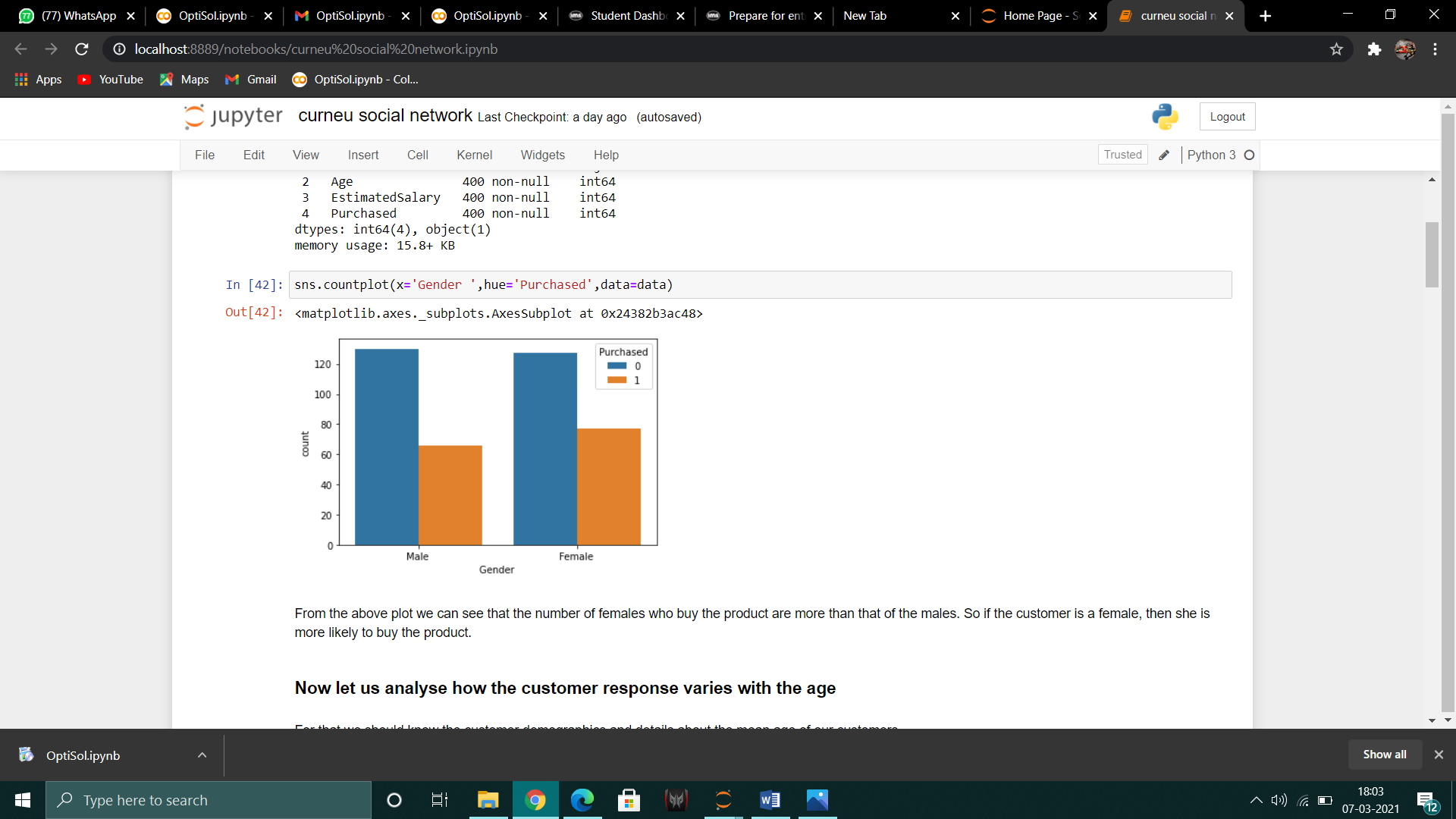
The data set consists of 4 independent variables namely user id, gender, estimated salary and age and the target variable is ‘purchased’ . The task is to predict whether a customer will purchase the product by clicking the ads.

The dataset had 400 entries. As a part of pre-processing , I checked for missing values in the dataset. But there were no missing values. Also the dataset contained categorical variable Gender. The gender was is in the form string. Since strings cannot be processed by the model , they need to be converted into numerical values. So the males were given 1 and females were given 0 .

Basic exploratory analysis was done, to find the significance of each attribute.

**CUSTOMER RESPONSE VARIATION WITH RESPECT TO GENDER :**

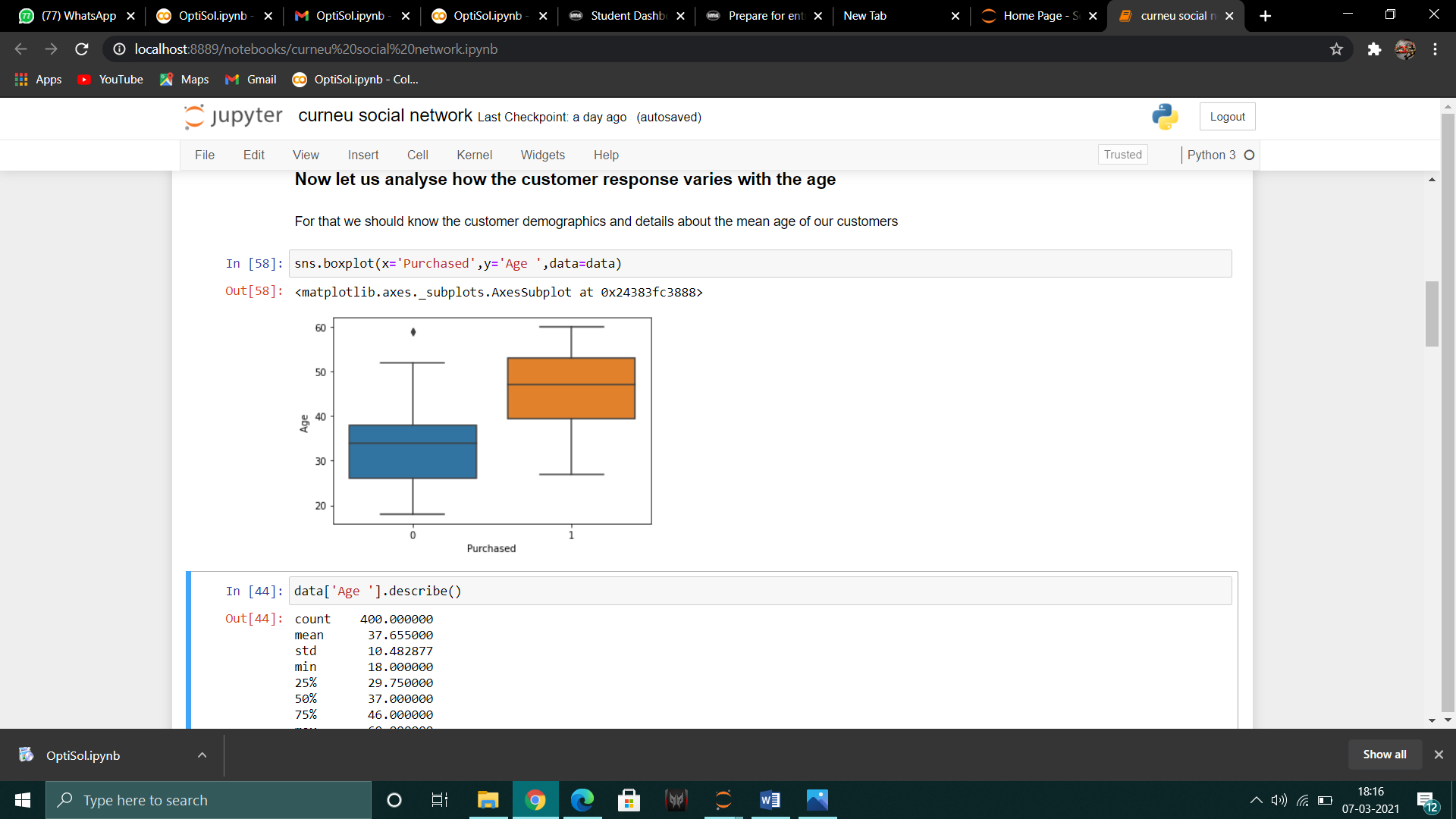
The number of male and female who purchased the product was plotted using the countplot. It also showed the ratio of people who purchased to the people who did not purchased for both the genders.

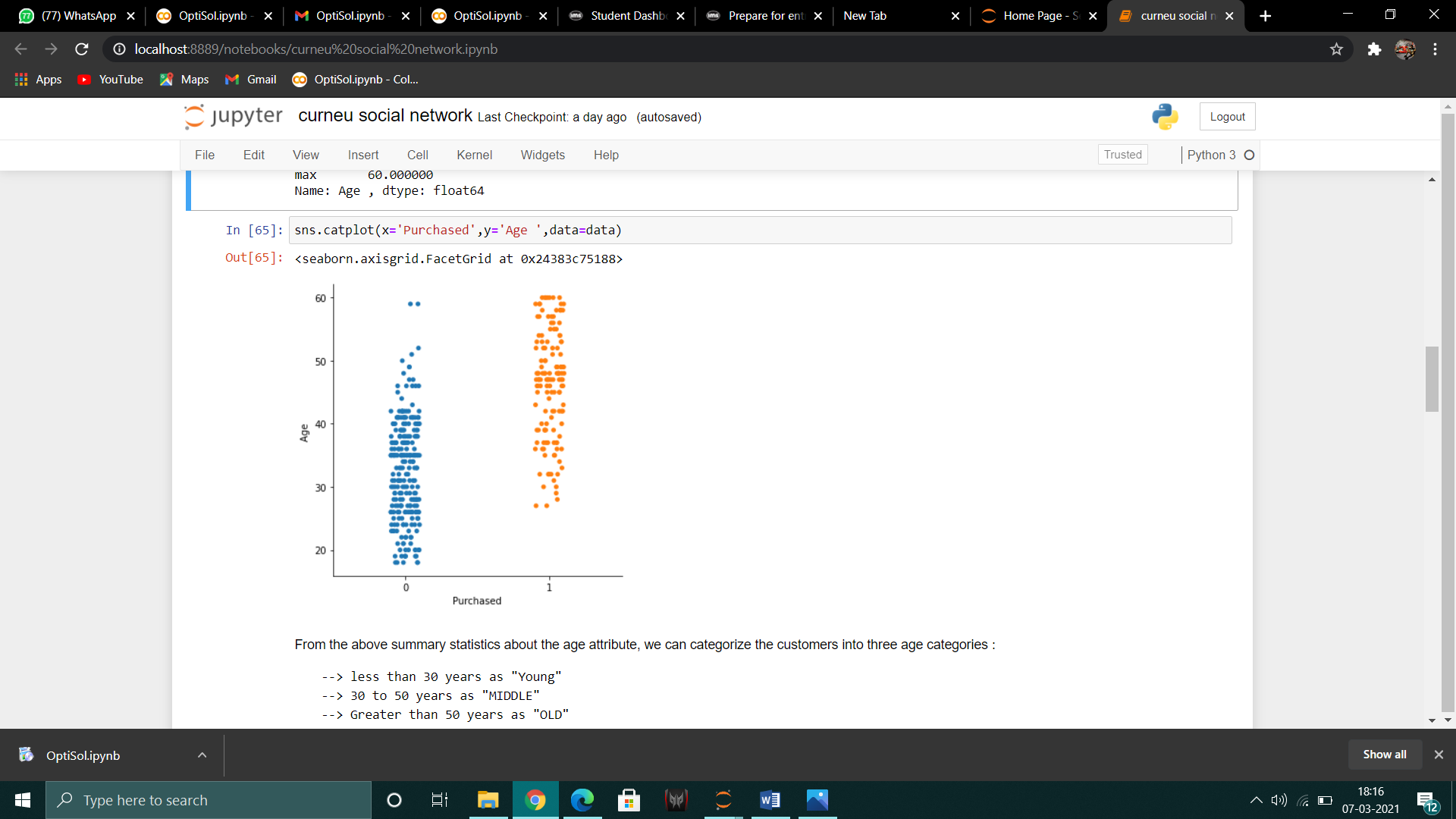


From the above plot we can see that the number of females who buy the product are more than that of the males. So if the customer is a female, then she is more likely to buy the product.

**CUSTOMER RESPONSE VARIATION WITH RESPECT TO AGE :**

For analysing the how the response of a customer is dependent on age, we should first know the demographics of the customer group. For that a simple summary statistic was calculated to find the interquartile range which shows the age range where maximum customers come under. Also the age groups were divided into 3 categories and plotted to get a clear understanding of the customer behavior. A box plot and cat plot were made to easily interpret the dataset. The box plot is useful in removing the outliers .





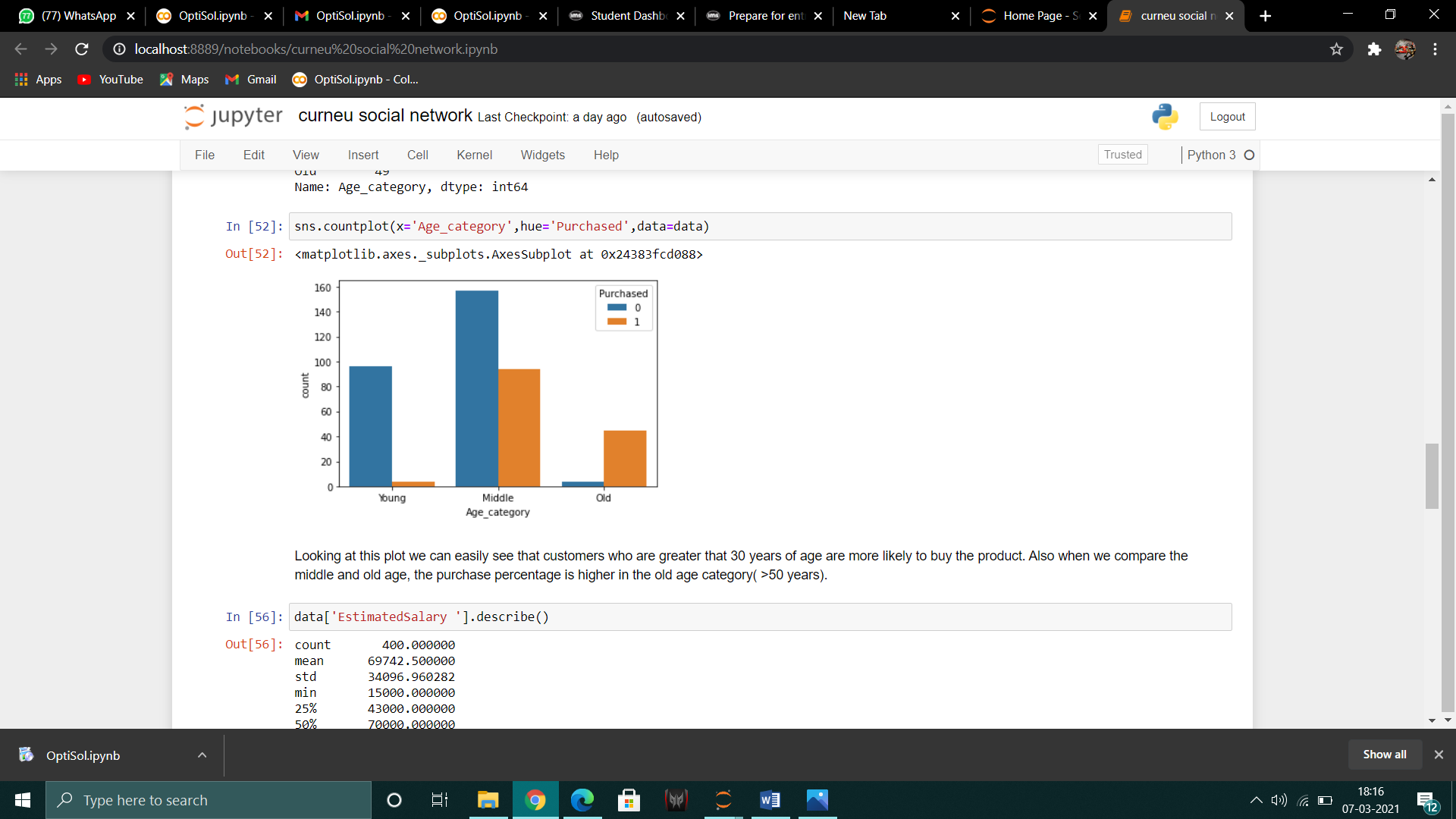
From the above summary statistics about the age attribute, we can categorize the customers into three age categories :

--> less than 30 years as "Young"

--> 30 to 50 years as "MIDDLE"

--> Greater than 50 years as "OLD"

We can now plot the count of customers in each category who have purchased our product.



Number of customers in each category :

Middle 251

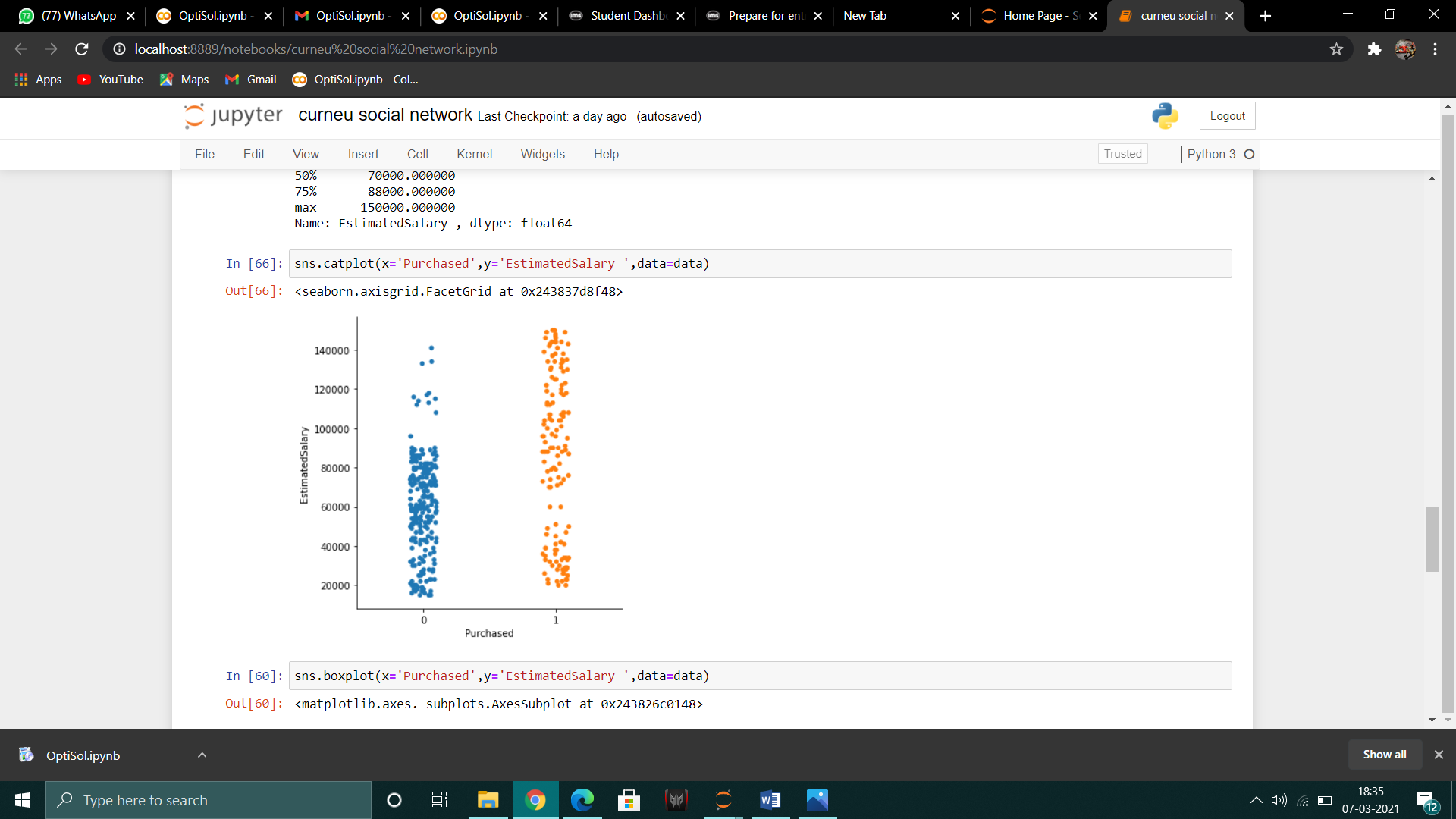
Young 100

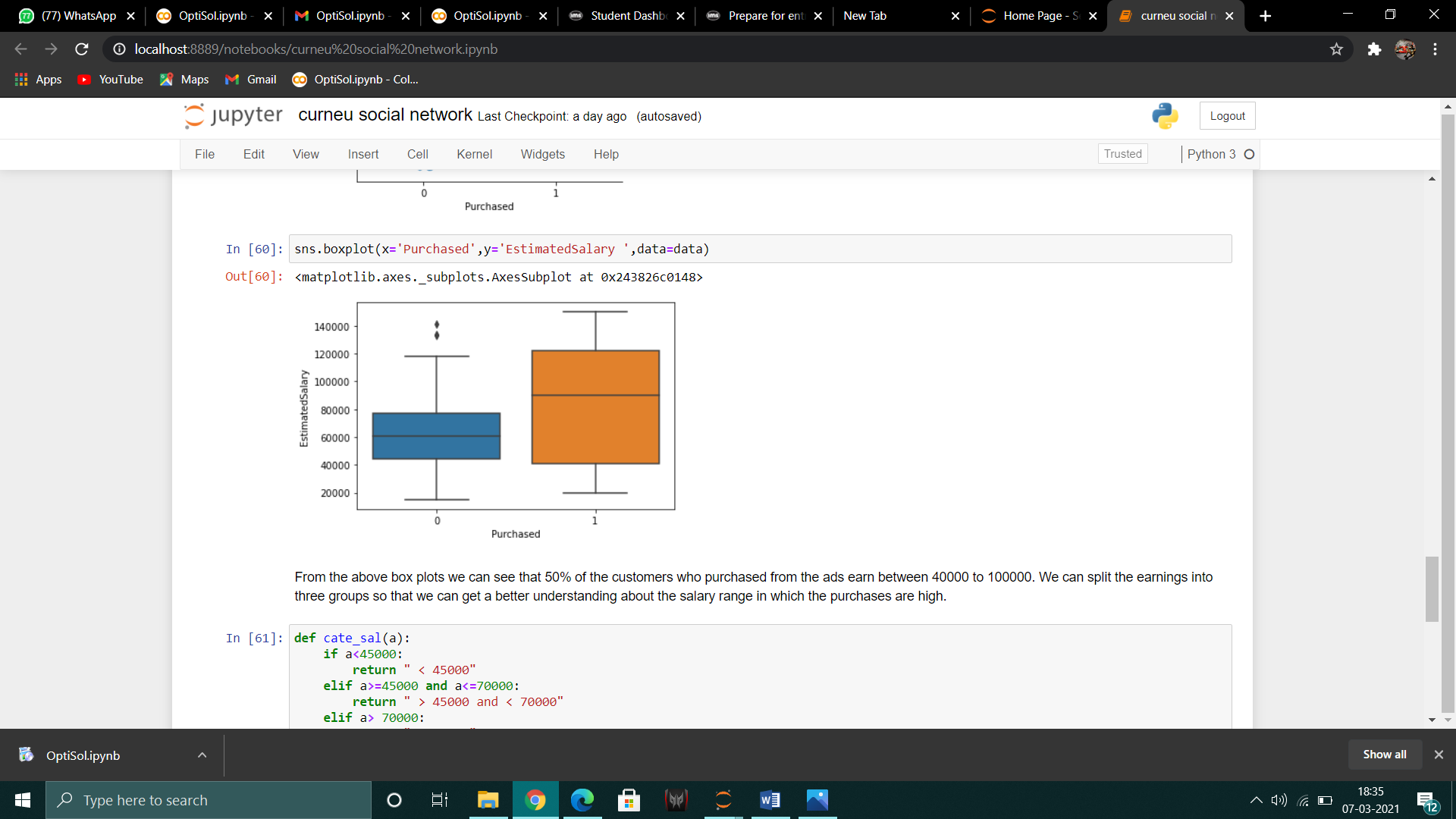
Old 49

Looking at this plot we can easily see that customers who are greater that 30 years of age are more likely to buy the product. Also when we compare the middle and old age, the purchase percentage is higher in the old age category( >50 years).

**CUTOMER VARIATION RESPONSE WITH RESPECT TO SALARY :**

As we did for finding the customer demographics, we should find the average salary of maximum number of people. Also categorize them into 3 salary ranges so that we can get a better intuition of relationship between the salary and the probability of the customer to buy to the product .





From the above box plots we can see that 50% of the customers who purchased from the ads earn between 40000 to 100000. We can split the earnings into three groups so that we can get a better understanding about the salary range in which the purchases are high.

The customers can be categorized into three groups as follows :

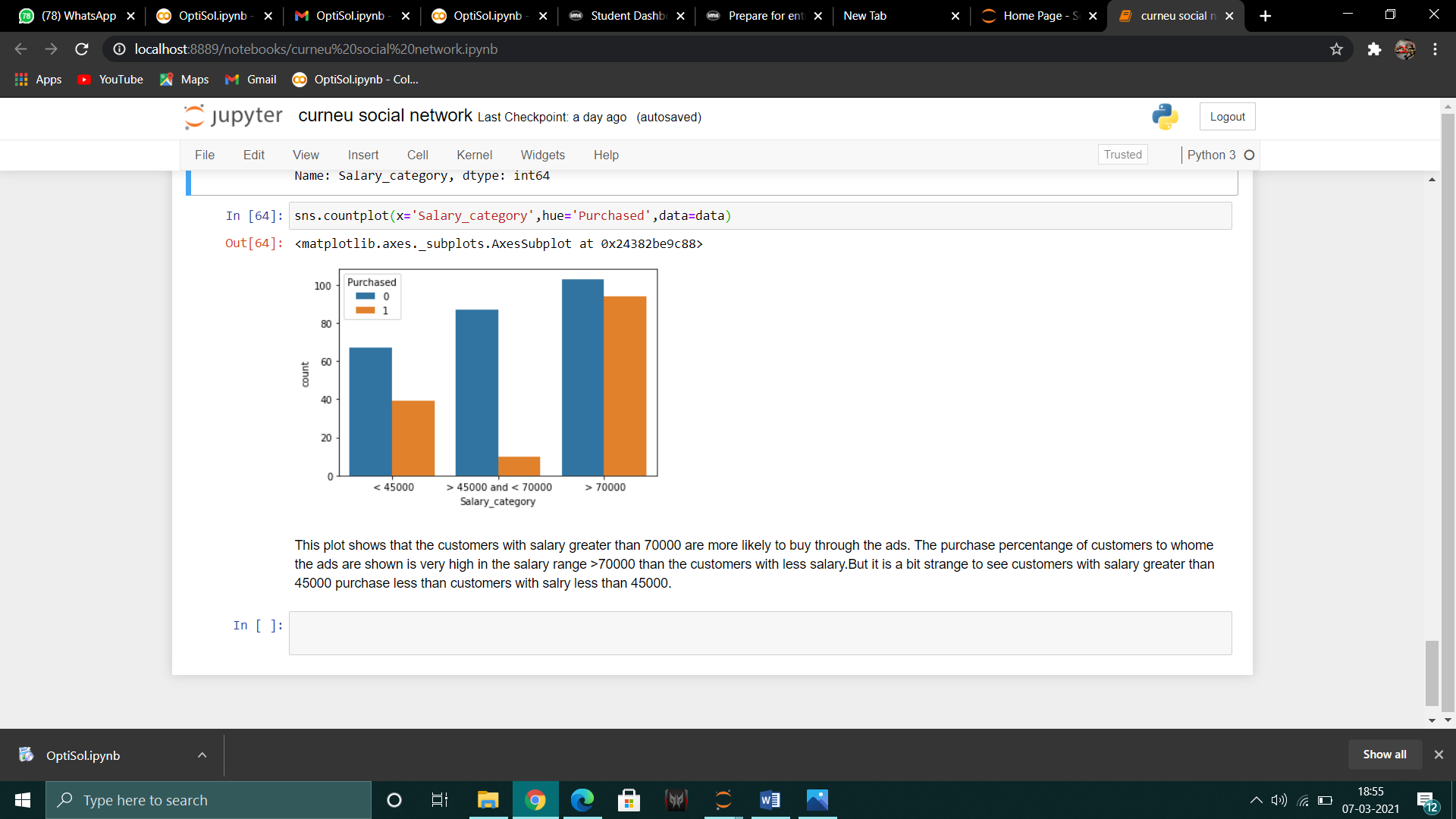
* More than 70000
* 45000 to 70000
* Less than 45000

Number of customers in each category :

> 70000 197

< 45000 106

> 45000 and < 70000 97



This plot shows that the customers with salary greater than 70000 are more likely to buy through the ads. The purchase percentange of customers to whome the ads are shown is very high in the salary range >70000 than the customers with less salary.But it is a bit strange to see customers with salary greater than 45000 purchase less than customers with salry less than 45000.

The exploratory analysis part is done. Before training the model we can normalize the dataset for better results.

**MODEL FITTING AND ACCURACY :**

Two models were fit. First a logistic regression model was fit to the data and it was trained. The test set was then used to test the model and the accuracy was found to be 63% .

Next K-Nearest neighbours model was fit and the accuracy was calculated to be 92%