Follow the link for the project

<https://github.com/asiwach08/ML-Projects/blob/main/Project_7-Customer_Churn.ipynb>

**Application of Machine Learning**

Machine learning is the study of various algorithms that improves through experience and by data use. Machine learning algorithms build a model based on sample data, known as "[training data](https://en.wikipedia.org/wiki/Training_data)", in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as [email filtering](https://en.wikipedia.org/wiki/Email_filtering) and [computer vision](https://en.wikipedia.org/wiki/Computer_vision), where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

Because of new computing technologies, machine learning today is not like machine learning of the past. It was born from pattern recognition and the theory that computers can learn without being programmed to perform specific tasks; researchers interested in artificial intelligence wanted to see if computers could learn from data. The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to [big data](https://www.sas.com/en_us/insights/big-data/what-is-big-data.html) – over and over, faster and faster – is a recent development.

Here we are using machine learning for the analysis and prediction of customer churn.

**Introduction:-**

Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals.

Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to defect means that a company can prioritise focused marketing efforts on that subset of their customer base

**Exploratory Data Analysis:-**

Exploratory data analysis(EDA) is all about getting and overall understanding of data. It is mainly done to find it’s properties, patterns and visualizations. It helps us to assure that our data is correct and ready to use for machine learning algorithms.In this blog we are using python(Jupyter Notebook) as our programming language for the analysis purpose. Python has a wide variety of libraries like pandas , seaborn , numpy , matplotlib which can be used for this purpose. We are using the UCI Machine Learning Dataset where it shows the various diagnostic factors influencing the Heart Disease.

Exploratory Data Analysis (EDA) is a pre-processing step to explore the data. There are numerous methods and steps in performing EDA including visualization, wrangling, cleaning,etc, however, most of them are specific, focusing on either visualization or distribution , and are incomplete . Therefore, here, I will take you through step-by-step to understand, explore, and extract the information from the data to answer the questions or assumptions. There are no structured steps or methods to follow, however, this project will provide insights on EDA.

**Problem Statement**

We have to train and test mode based on customer churn dataset and find a model which can predict the churn or attrition of customers.

**Brief introduction to the used libraries:**

* *Numpy -* NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities.
* *Pandas -* Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for groping, combining and filtering data.
* *Matplotlib -* Matpoltlib is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, viz., histogram, error charts, bar chats, etc.
* *Seaborn -* Seaborn is a Python data visualization library based on [matplotlib](https://matplotlib.org/). It provides a high-level interface for drawing attractive and informative statistical graphics.

**Attributes information:-**

customerID object

gender object

SeniorCitizen int64

Partner object

Dependents object

tenure int64

PhoneService object

MultipleLines object

InternetService object

OnlineSecurity object

OnlineBackup object

DeviceProtection object

TechSupport object

StreamingTV object

StreamingMovies object

Contract object

PaperlessBilling object

PaymentMethod object

MonthlyCharges float64

TotalCharges object

Churn object

**The Outline of EDA is as follows:-**

**1.Import and analyze to know the data**

a. Importing required libraries

b. Importing Dataset

c. Knowledge of data

**2. Data Analysis**

a. Check the data type

b. Check for the data characters mistakes

c. Check for missing values and replace them

d. Check for duplicate rows

e. Statistics summary

f. Cleaning the rows and columns data

**3. EDA Concluding Remarks**

a. graphs between target variable and features

b. boxplot,and others

c. line plot

d. heatmap, correlation map

**4. Pre-processing Pipeline**

a. Outliers and how to remove them

**5. Building Machine Learning Models**

a. splitting independent and dependent features

b. splitting train and test data

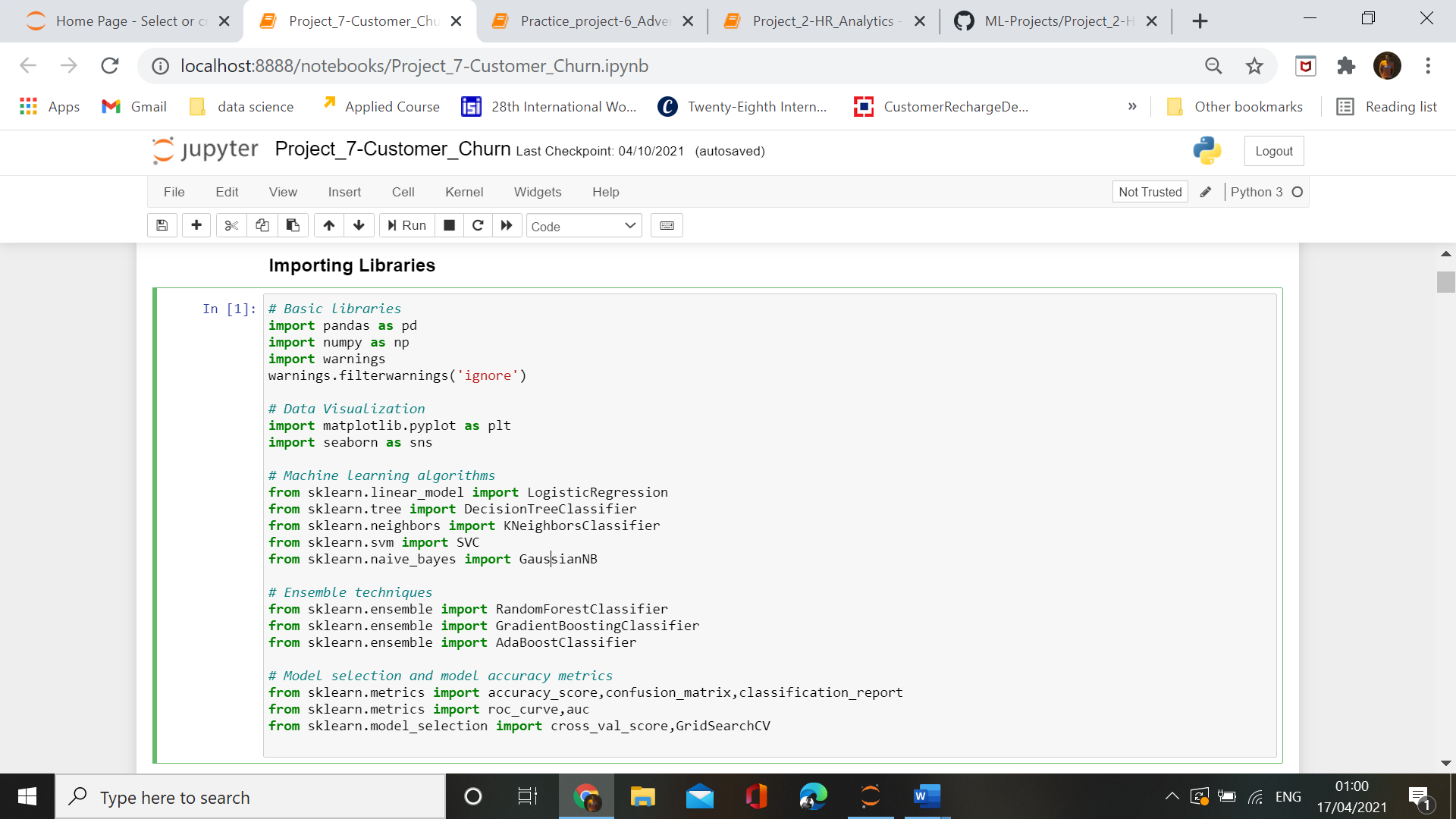
c. searching for best model

**6. Concluding Remarks**

Let’s get started…!!!

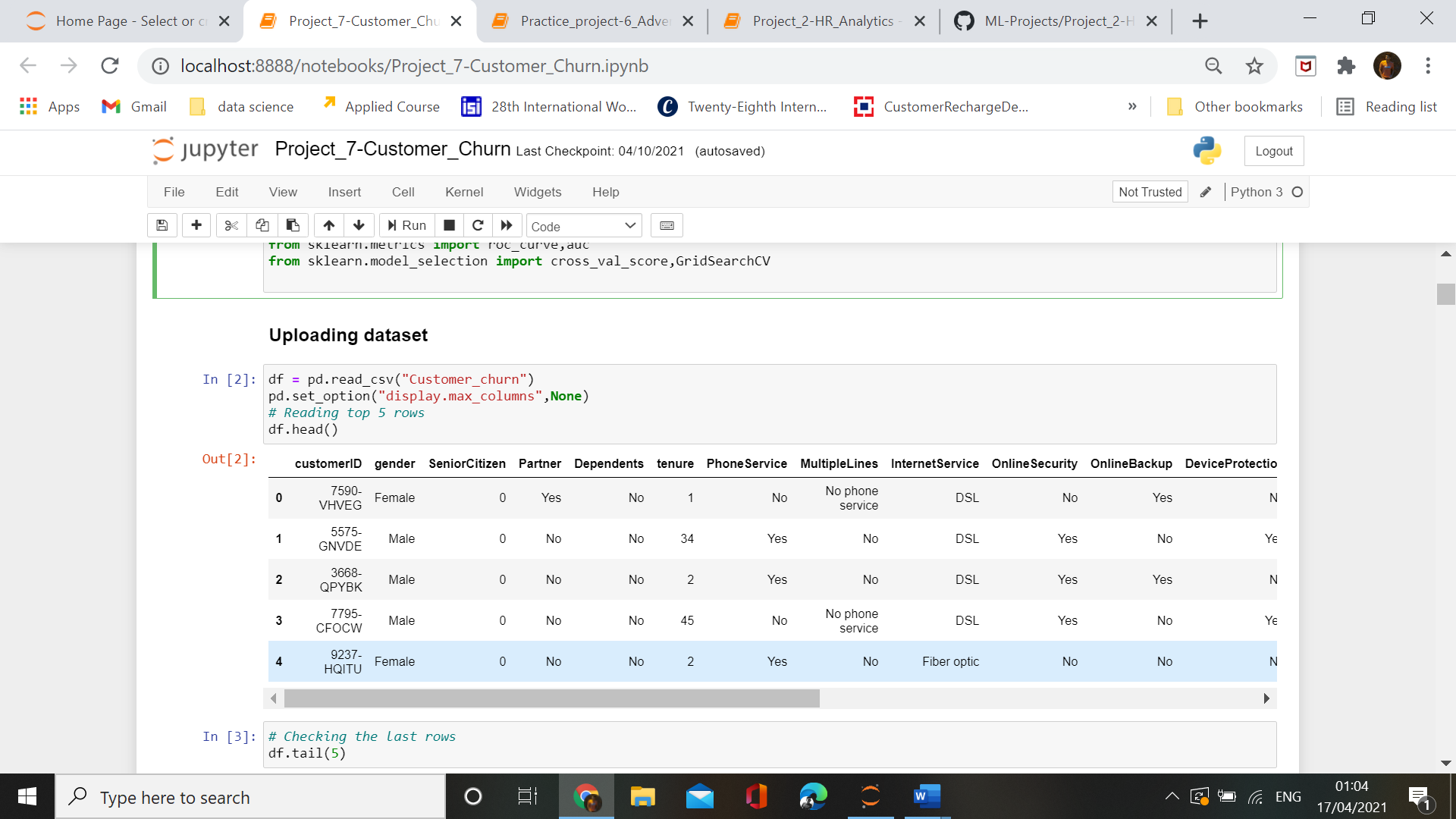
**Import and analyze to know the data**

**a.Importing required libraries**



We have imported required necessary libraries.

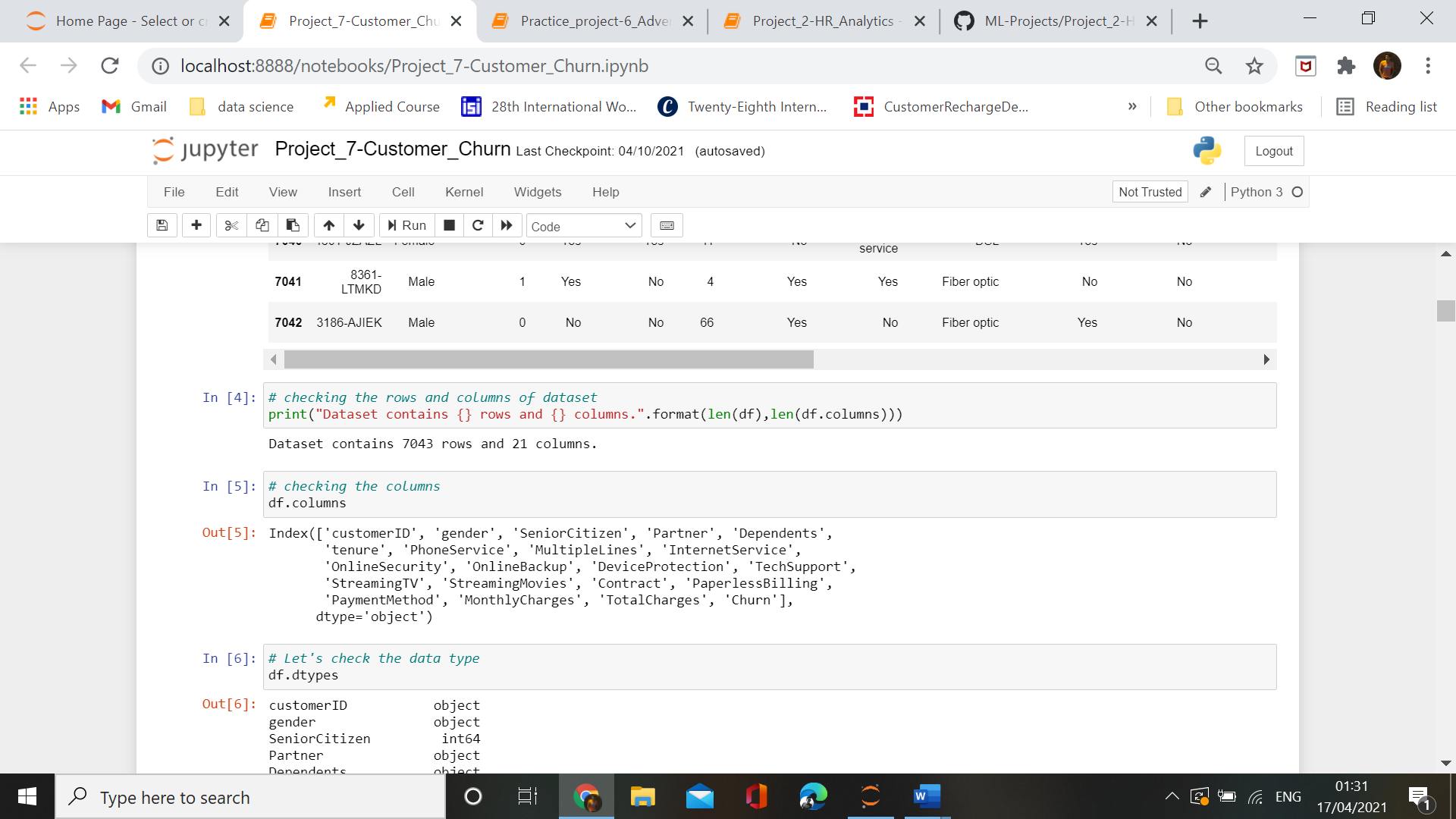
**b.Loading dataset**



We have loaded the dataset and are seeing the first 5 head rows.

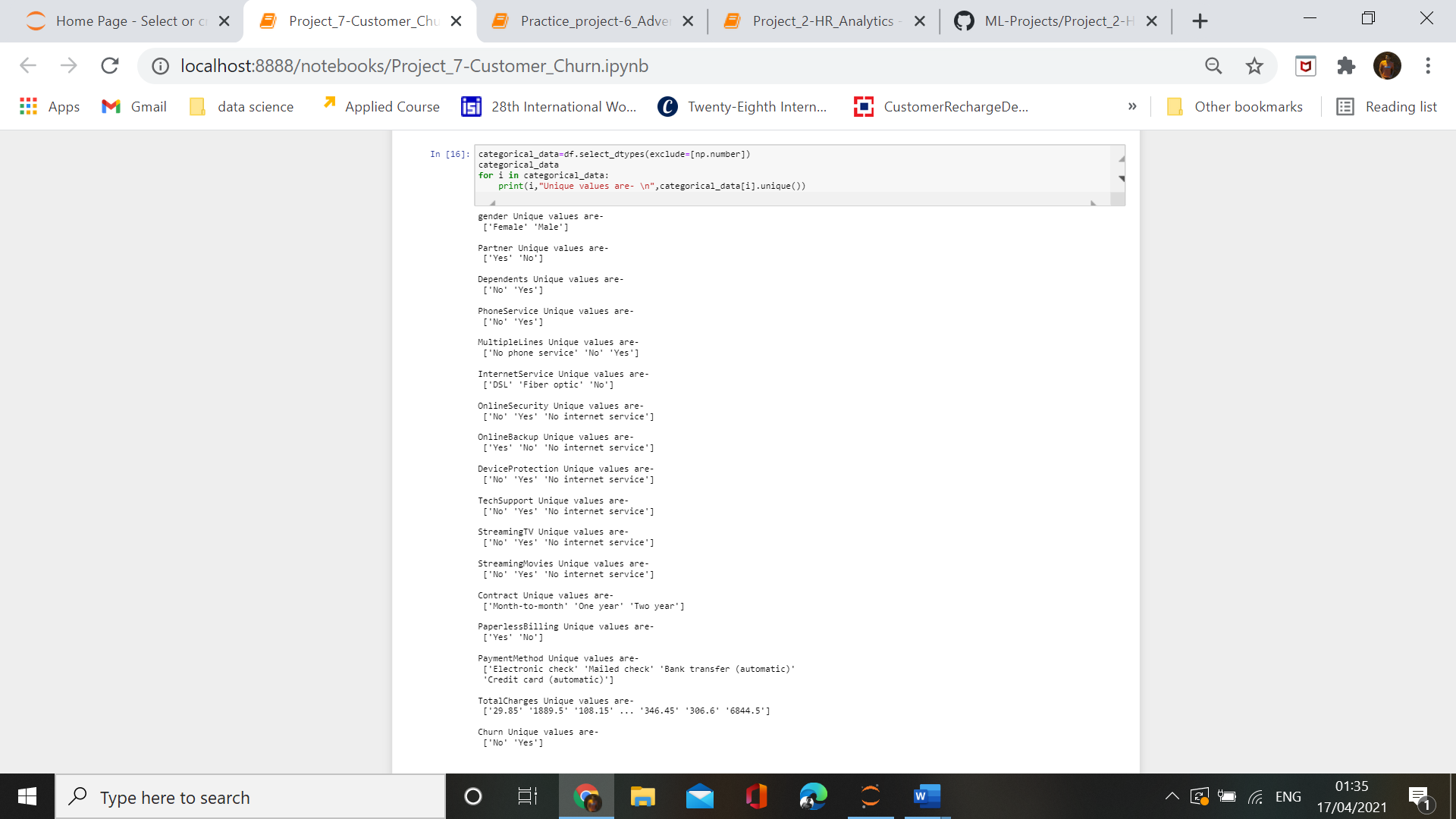
**c.Knowledge of data**

In this section , we will explore the data and try to find out it’s characteristics Collecting Basic Information about the Data.It shows the number of rows, number of columns, data types information, Memory usage, number of null values in each column.



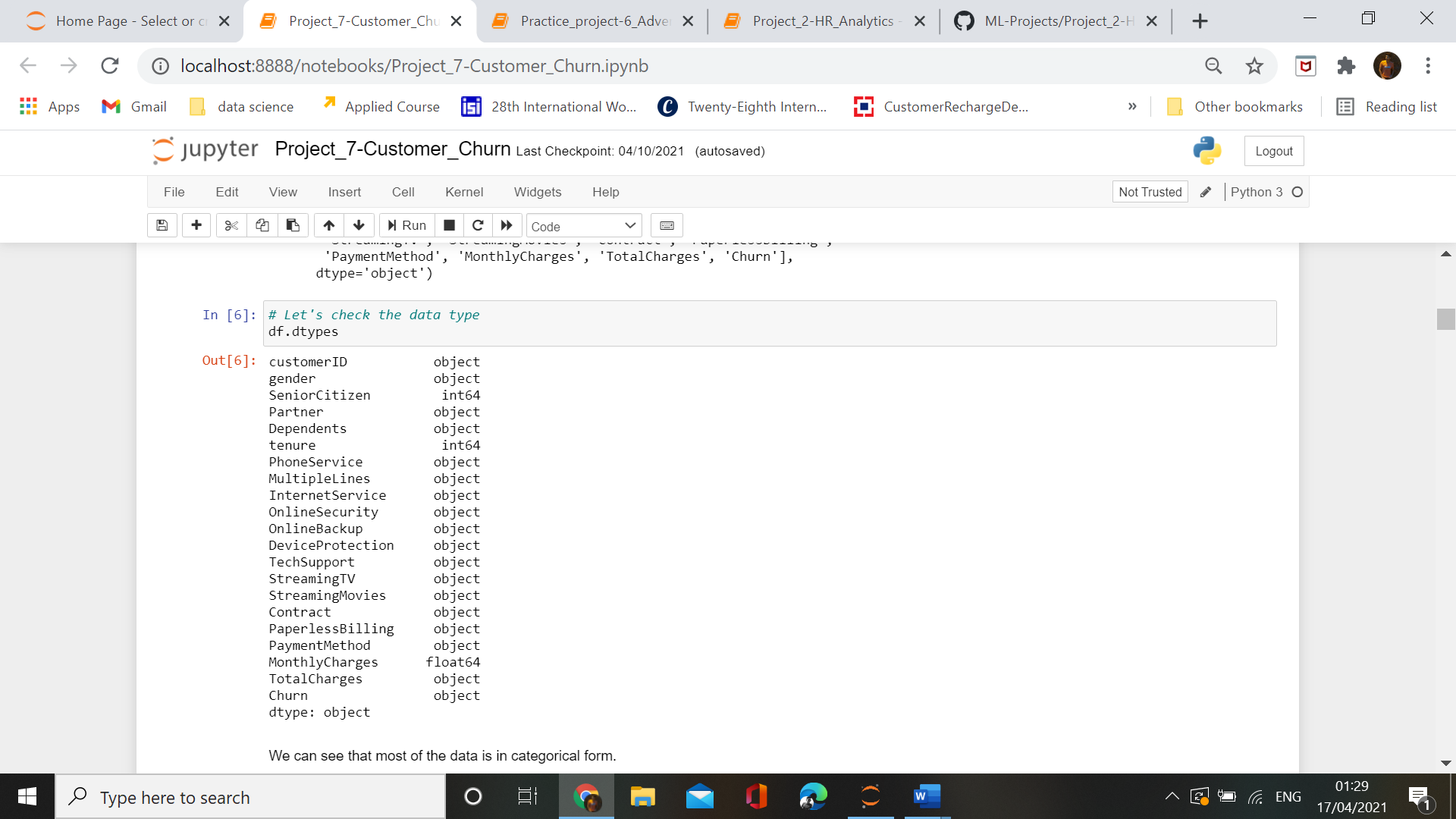
Dataset contains 7043 rows and 21 columns.

**Checking the values in different features.**



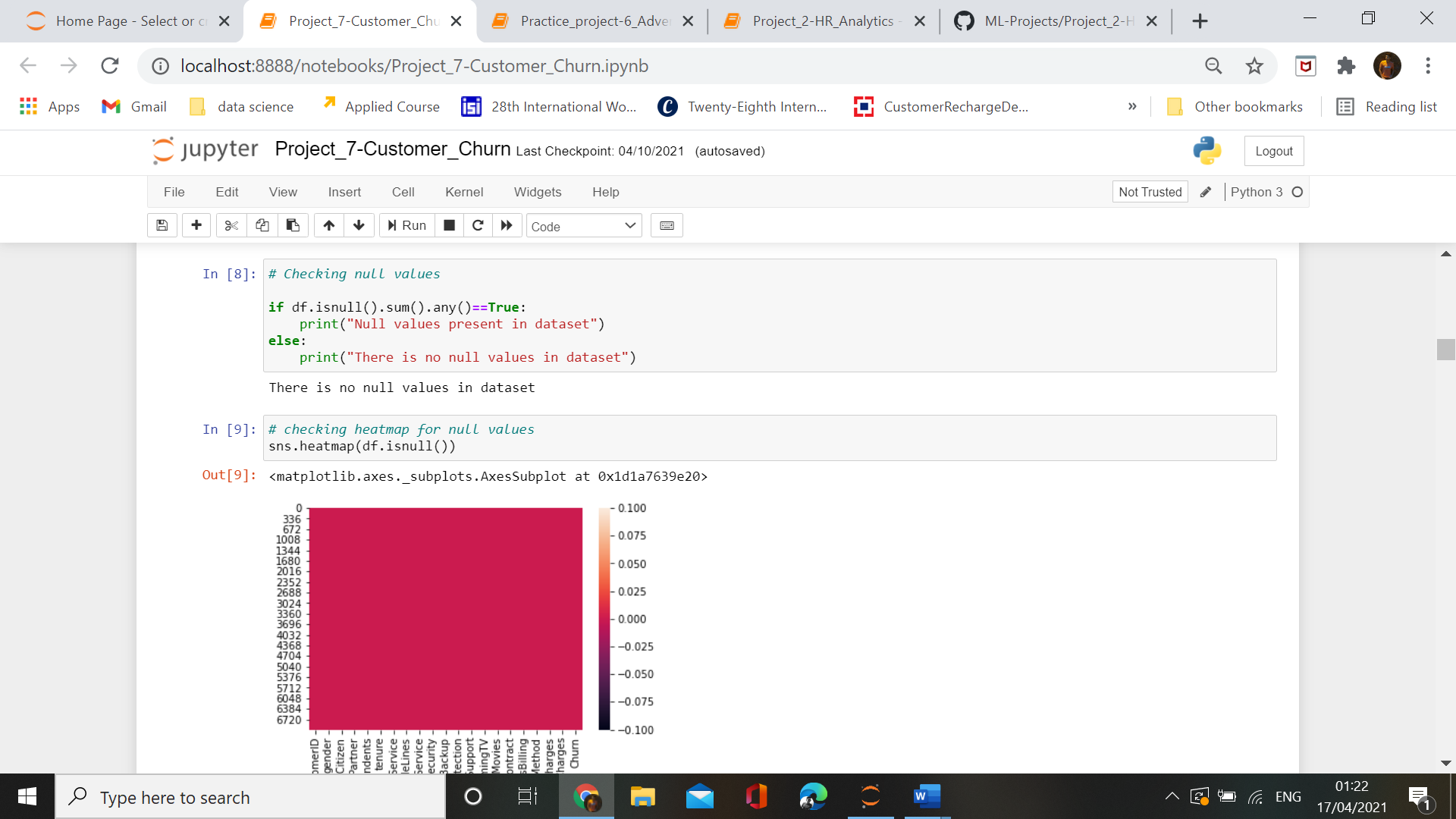
Dataset contains some features whose value is not correctly distinguished. So we will change their values in correct format.

**2. Data Analysis**

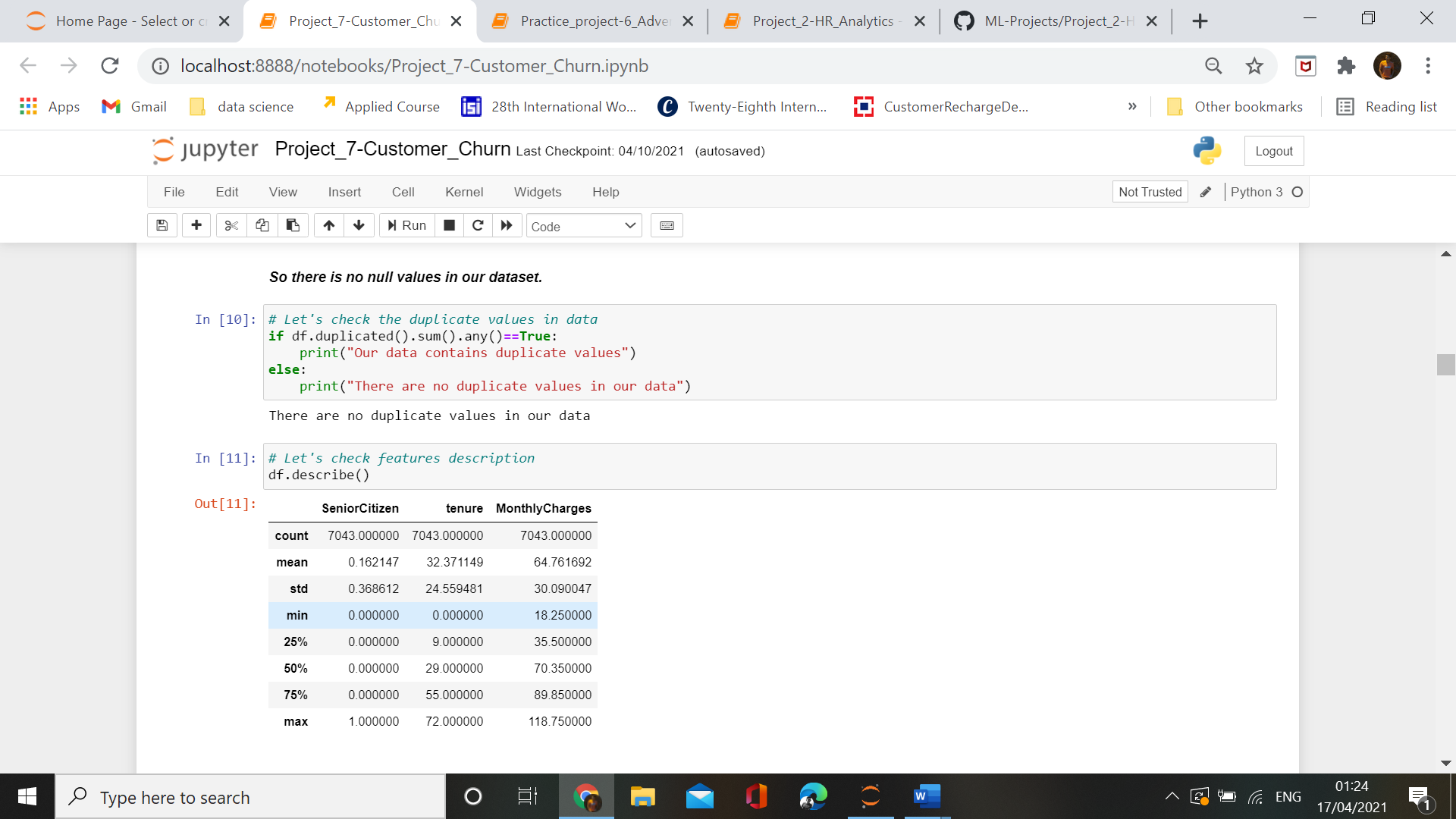


**In our dataset we have object,float and integer type of data.**

We need to check the presence of the missing values and need to replace them with mean, median and mode of data accordingly. Sometimes we have null values in the form of 0 , so we need to convert them to NaN and then replace them accordingly. The missing values can be removed also but it should be less than 5 percent of the whole dataset.and visualize the missing values using Heatmap. The missing values are represented by the horizontal lines. This map provides an informative way of visualizing the missing values located in each column, and to see whether there is any correlation between missing values of different columns.



**There are no null values in our dataset.**



**There are also no duplicates values in our dataset.**

**Checking summary statistics**



**Key Observations** :-

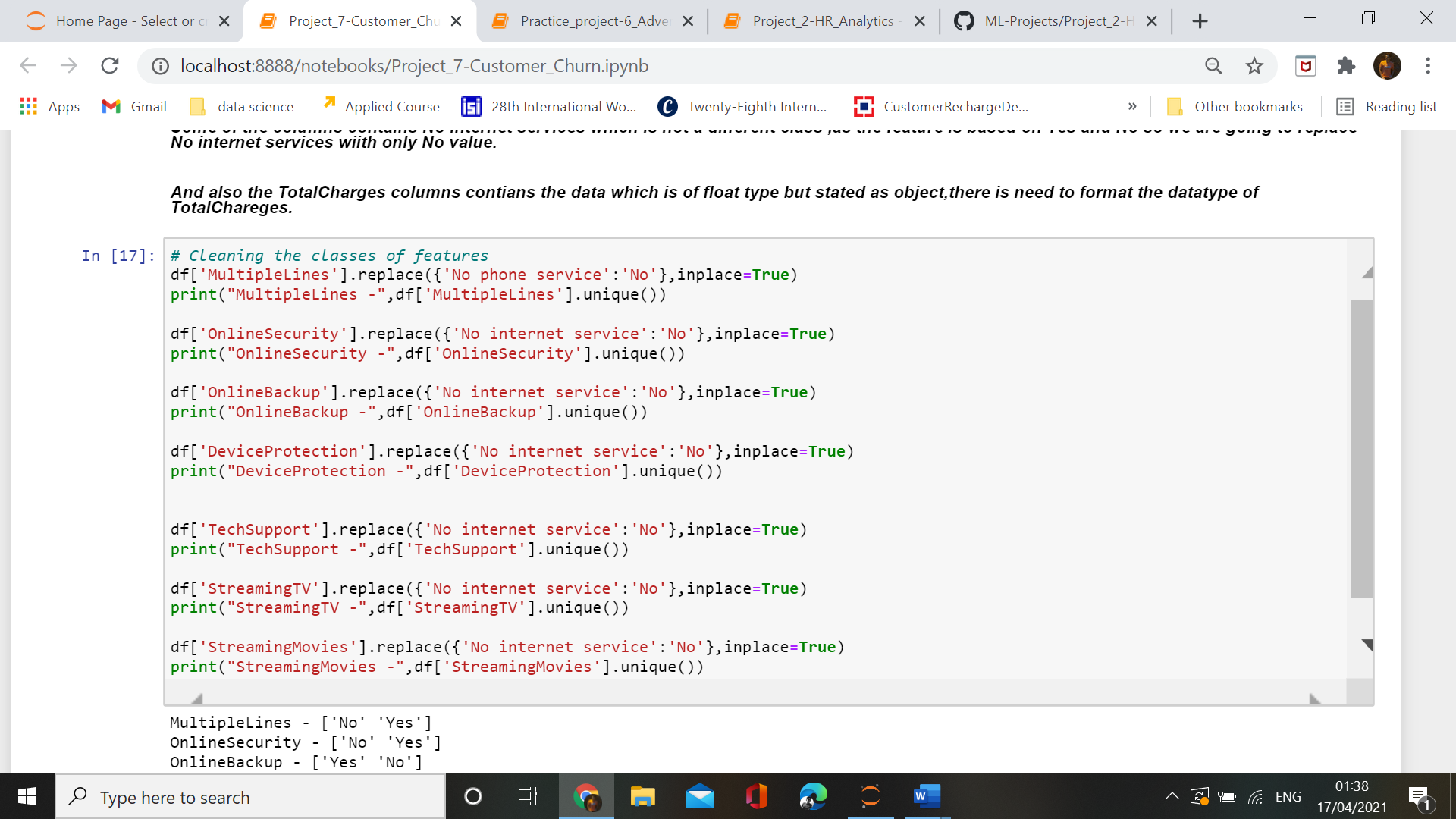
So we have only `3` numerical features which are giving us some information -

\* Seniorcitizen is a integer type of data,it suggest only two values 0 and 1.So it is not giving any usefull information although it can specify that there is imbalance in the data of SeniorCitizen.

\* Tenure is also a integer type of data ,it suggest the tenure(years) of a customer in a company, there is also a difference in mean and median of tenure data.

\* Monthlychareges is the only float datatype in our dataset,difference in mean and median is suggesting the presence of skewness in dataset,while difference in 75th and 100th percentile is suggesting the persencee of outliers.

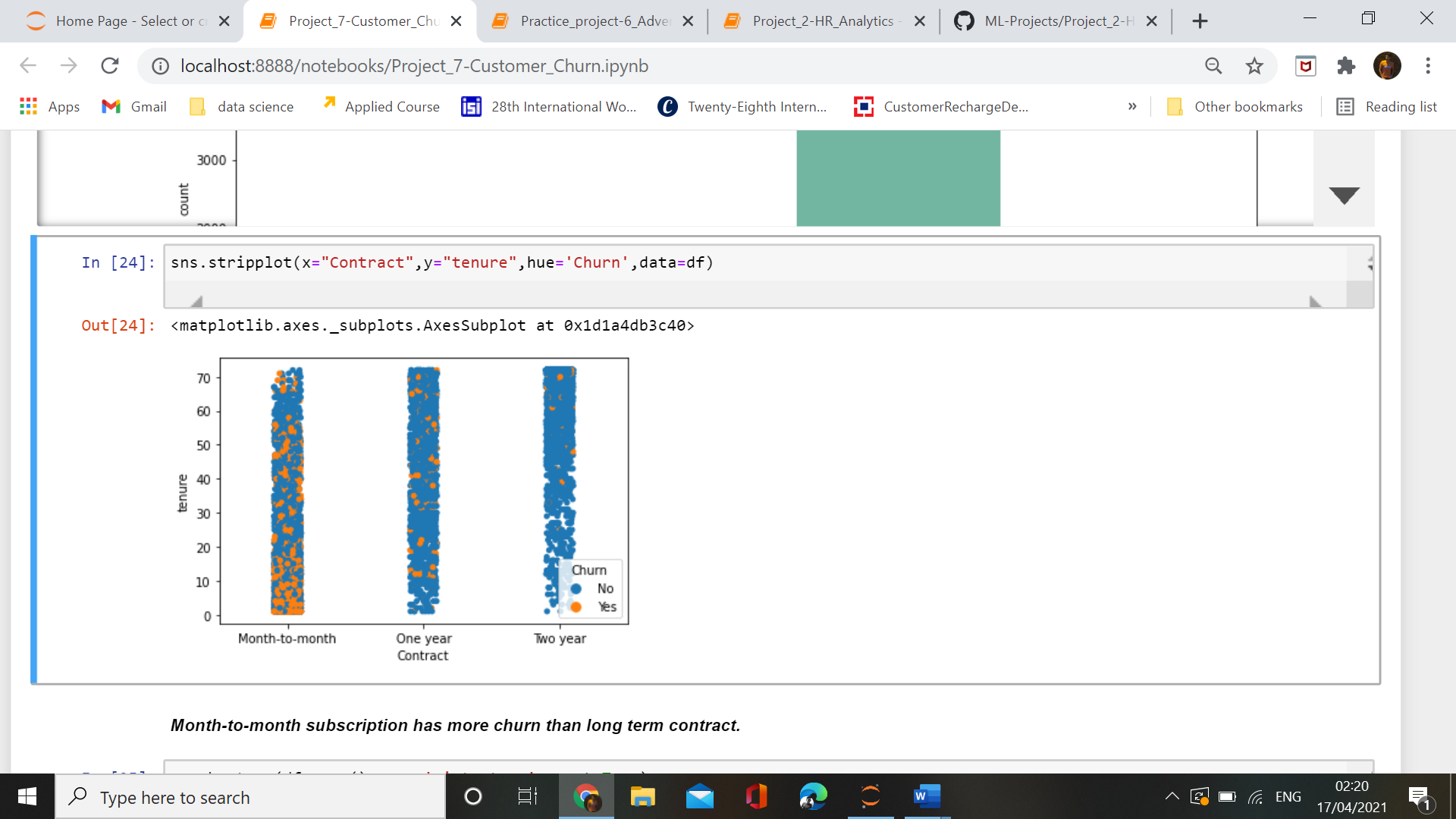
**Cleaning the features data**



We have entered the correct format of all values by changing the No internet service into No.

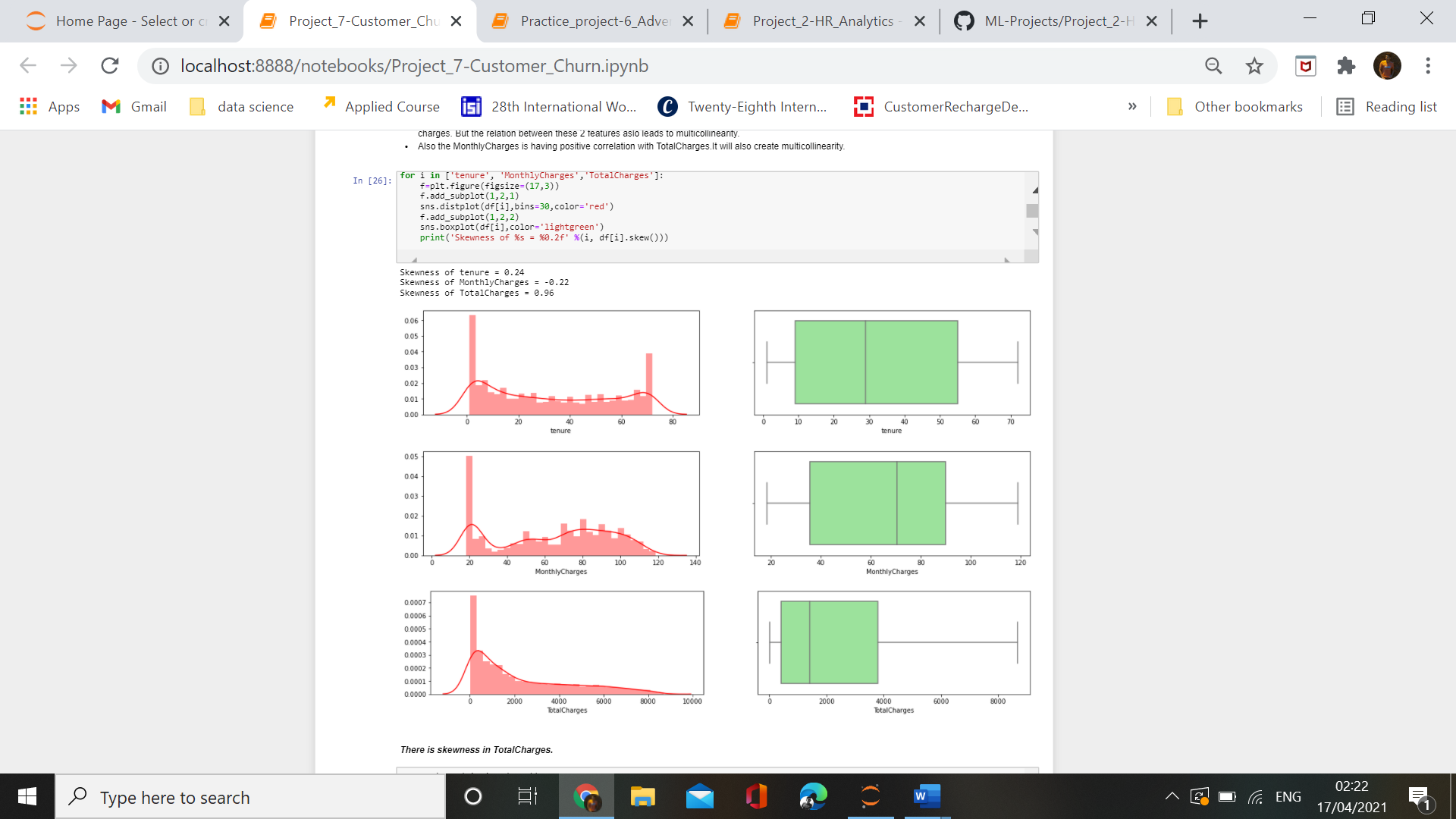
**3.EDA concluding remarks**

**a. graph between target variable and features**



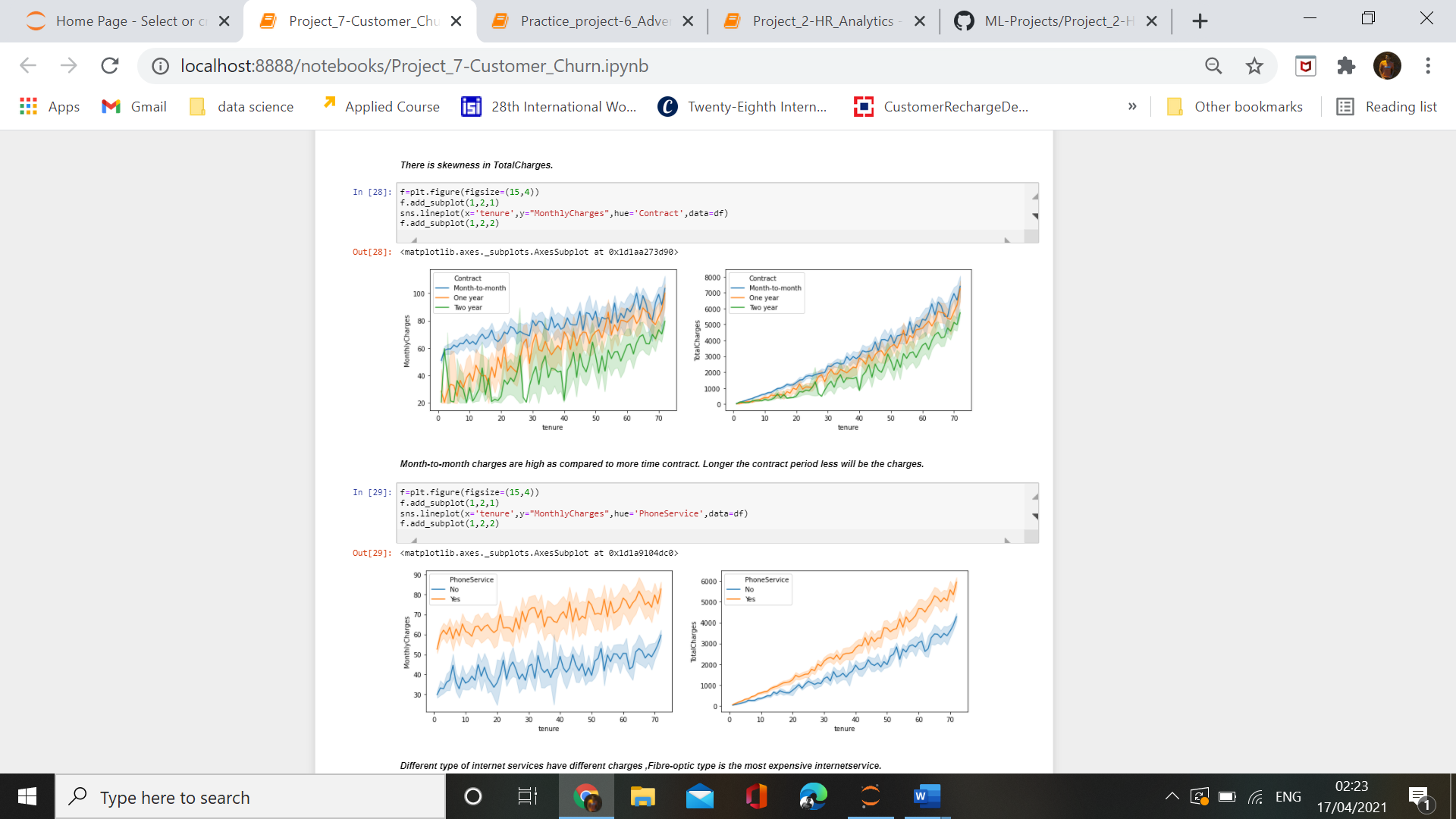
Strip plot graph between contract, tenure and target variable shows that the month-to-month contract is having high churn then two-year contract.

**b. boxplot,and others**

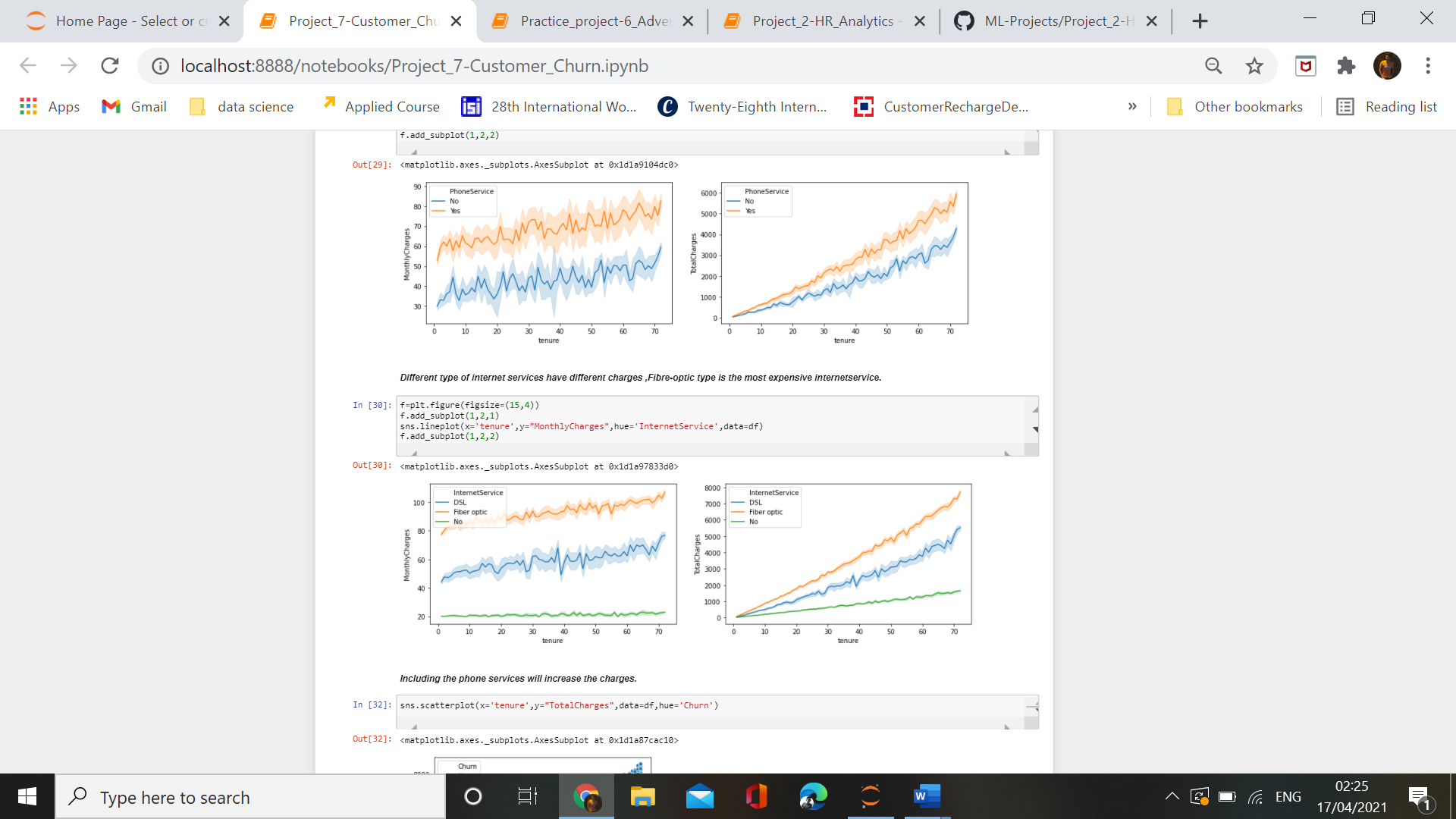


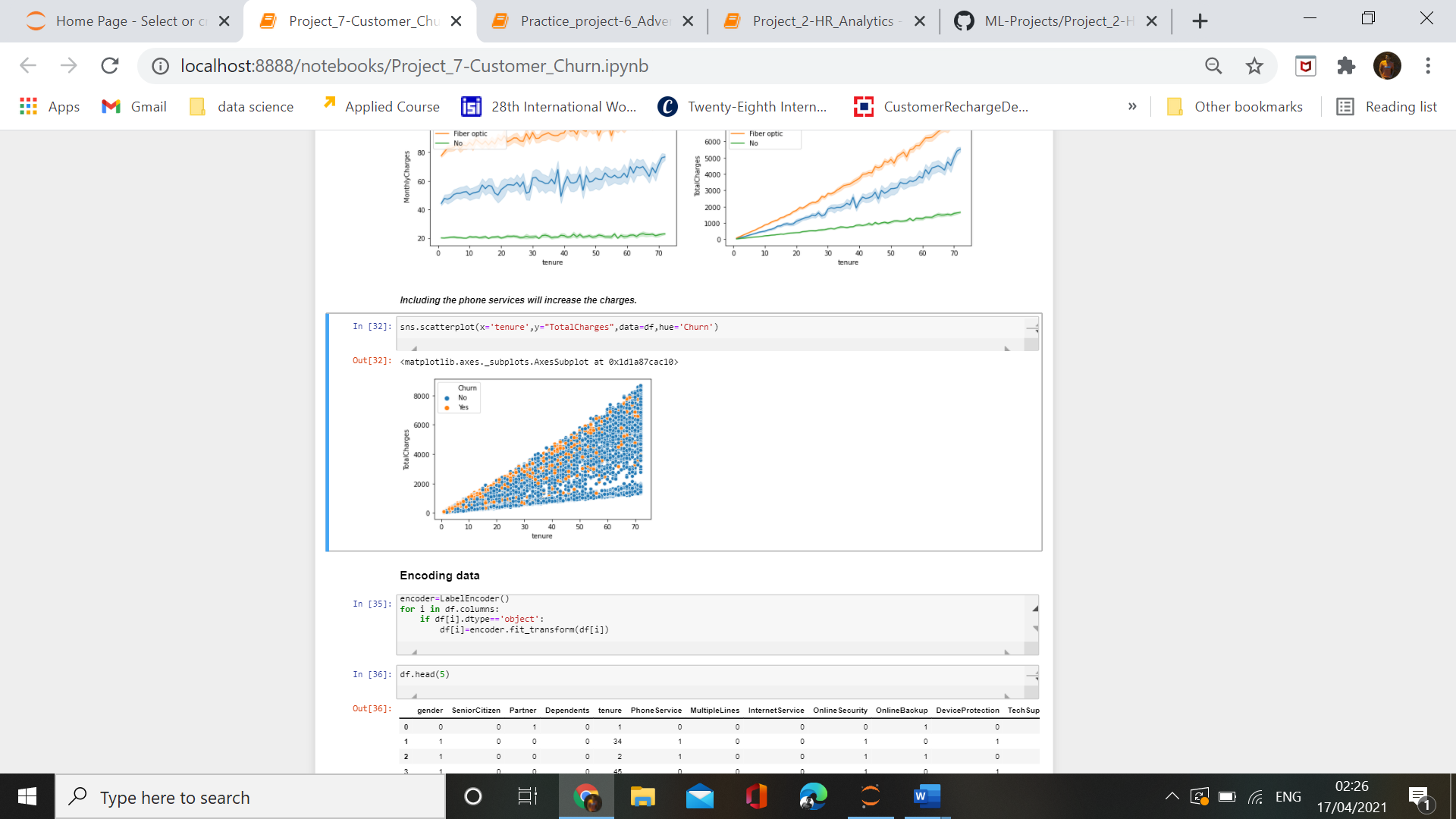
In our dataset we have only 3 numerical features ,so we have only 3 distplot which tells the distribution of data in feature. Here we can see that TotalCharges feature is having high skewness value which means that the data is not normally distributed in this feature. While by plotting boxplot, we find that there is no outliers in any of the numerical features.

**c.lineplot**



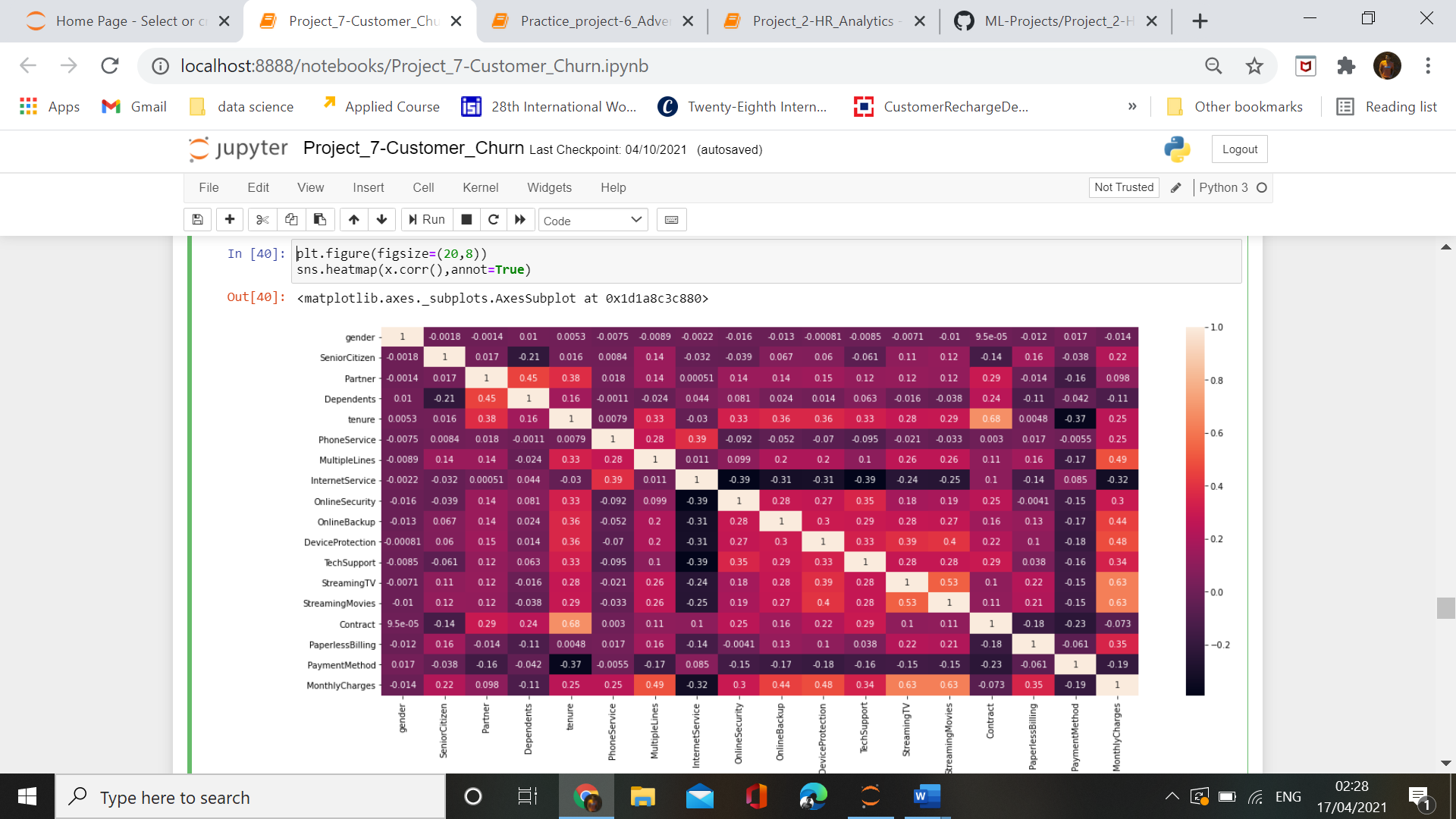
Using lineplot between tenure ,monthlyCharges and contract we find out that as the tenure increases monthlyCharges also increases but based on contract type month-to-month contract are high priced than two-year contract. So higher the contract period less will be the charges. It also shows that rate of customer churn is high where contract period is month-to-month.





Scatterplot between features and target variable is showing that lhigher the tenure higher be the totalCharges and higher be the churn. To deal with the condition of customer churn ,company should focus on the charges . They should lower their charges based on the tenure so that they can deal with customer churn problem.

**d.heatmap and correlation graph**

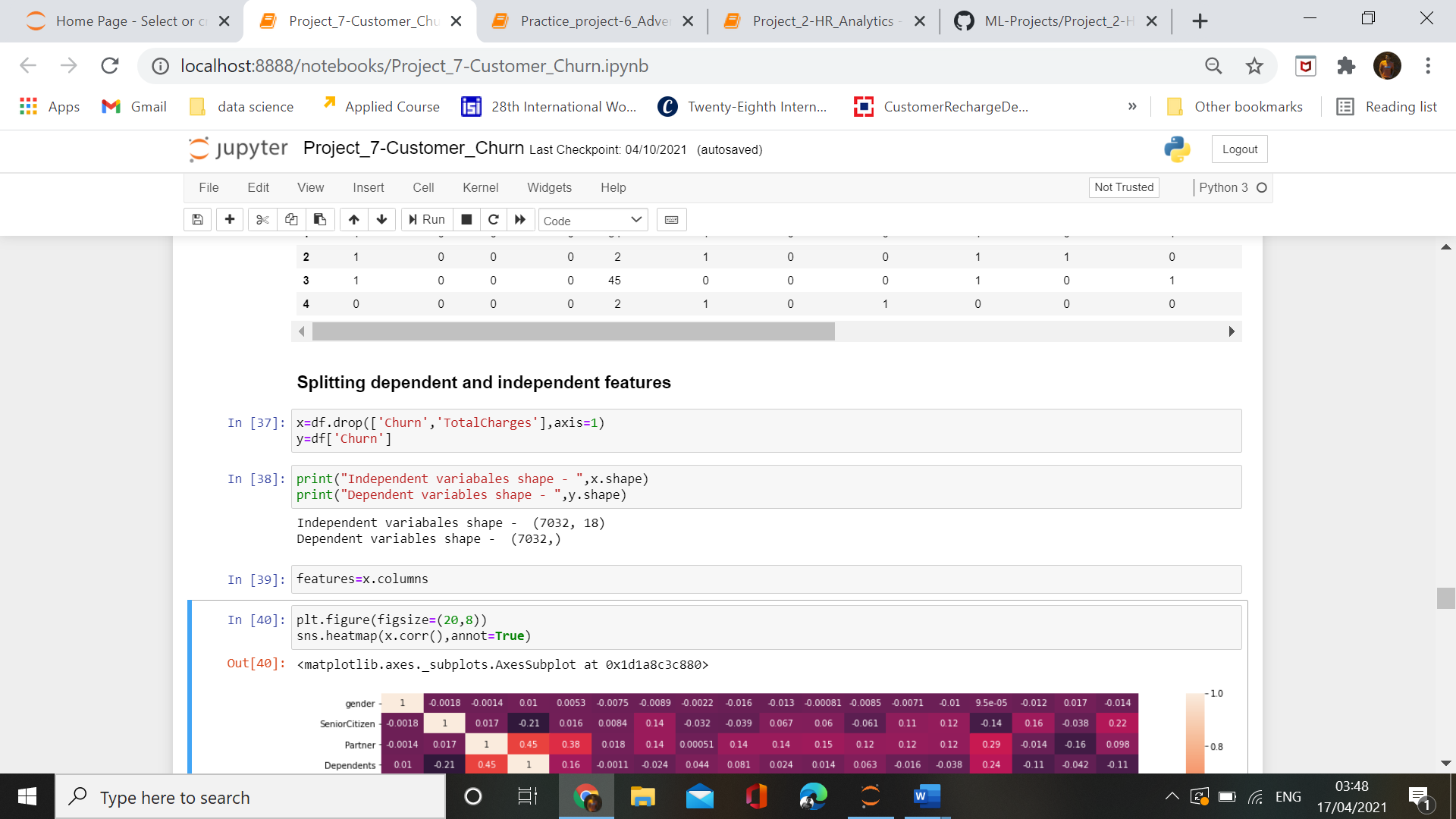


As the target variable is categorical in nature so we cann’t find the correlation with other features, As from the correlation graph we can also see that there is no strong correlation between features ,which means that there is no problem of multicollinearity in our dataset.

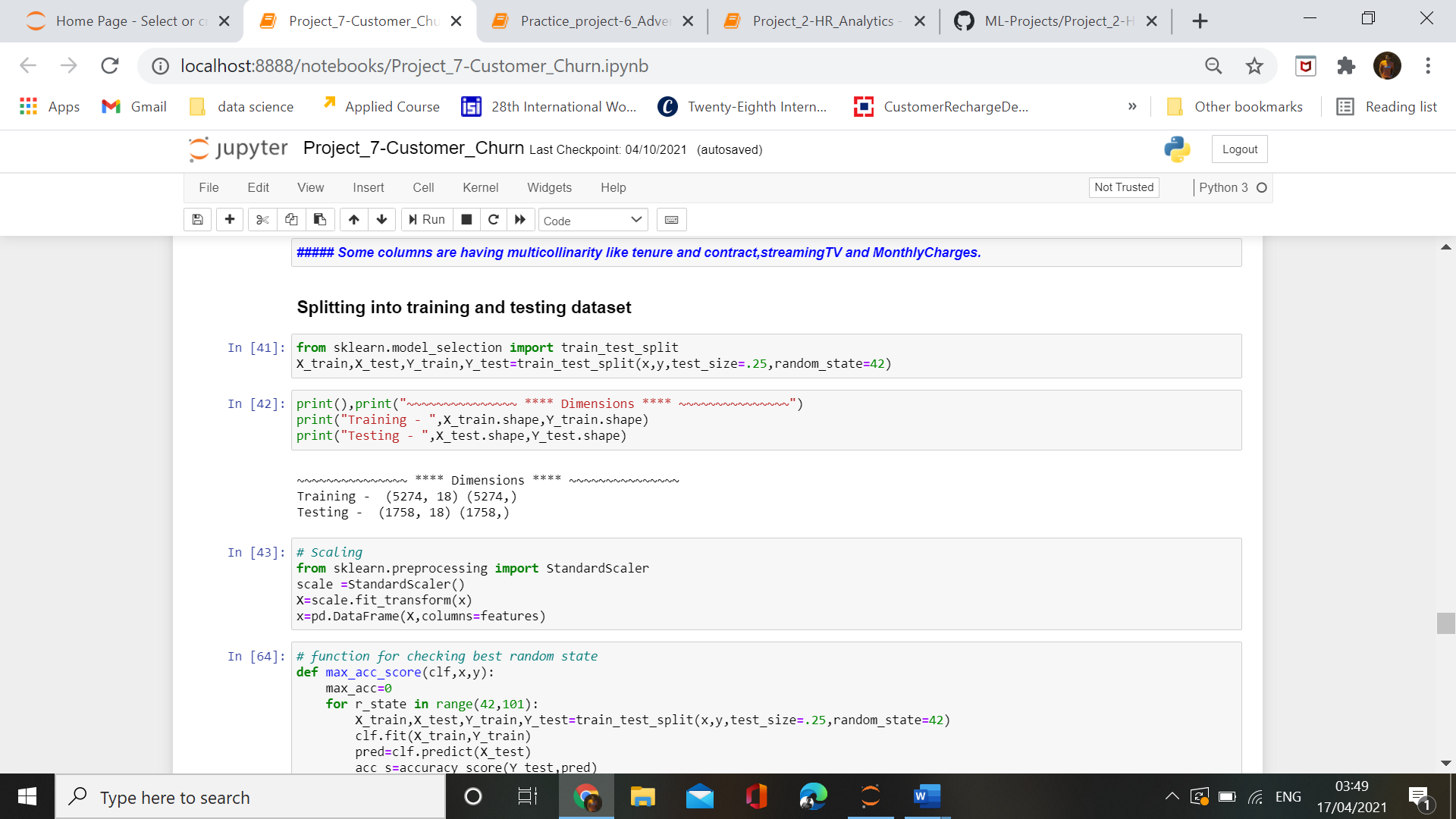
**4. Dealing with outliers and skewness**

As our dataset contains only 3 numerical features and by using boxplot we find that there is no outliers in any of the numerical features ,while we have already cleaned categorical features by converting their data into proper format. So now we have no outliers in our dataset .While talking about skewness ,we find out tha skewness was present in totalCharges and there is one another column MonthlyCharges which was highly correlated with totalCharges .So we just dropped the feature totalCharges and in this way we indirectly removed the skewness from our data.

**5. Building Machine Learning Models**



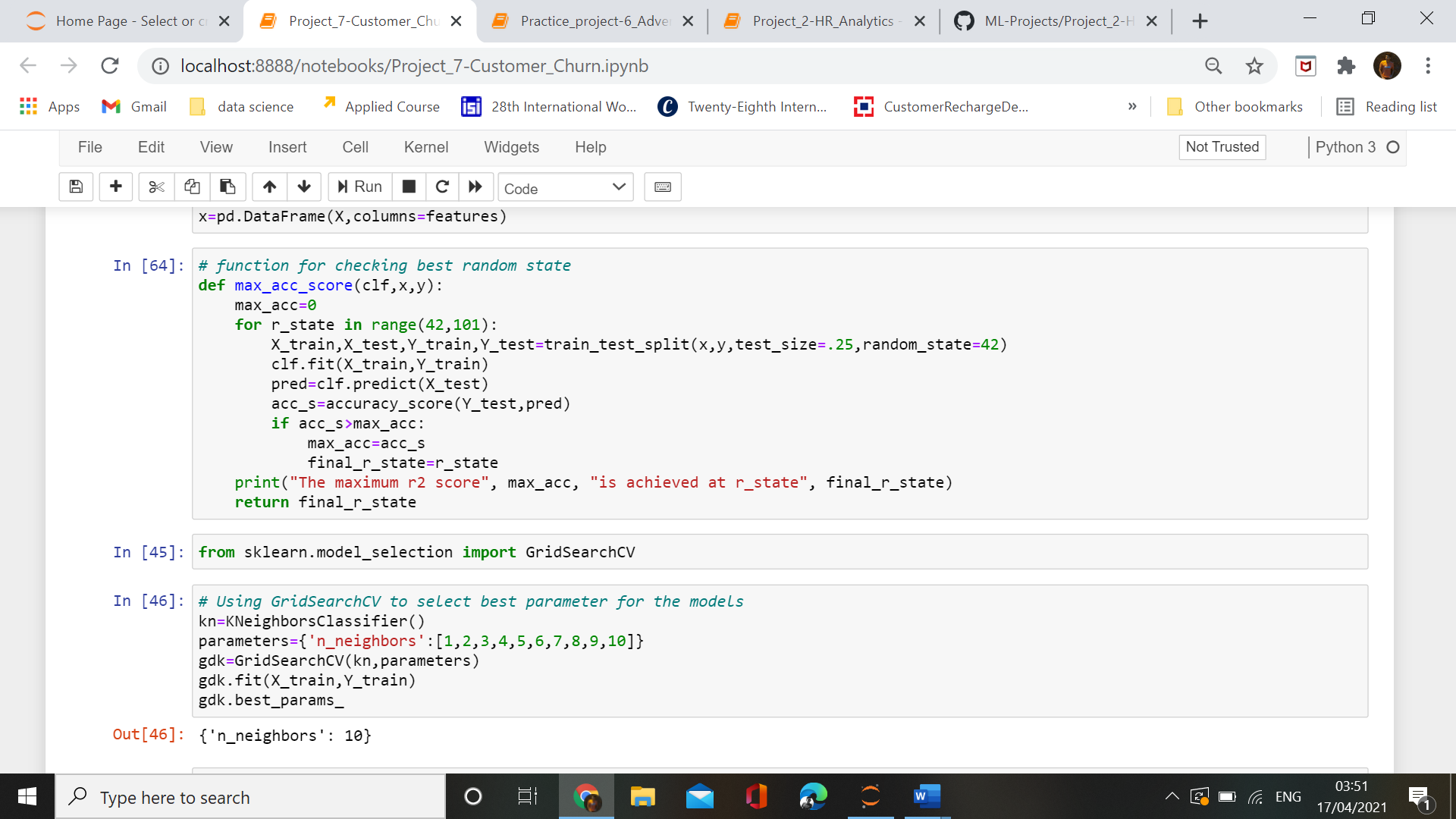
We have splitted the dependent and independent features. Y is pur dependent feature while x is our independent features.Checking the dimensions of our features ,we have 7032 rows with 18 features in x variable.



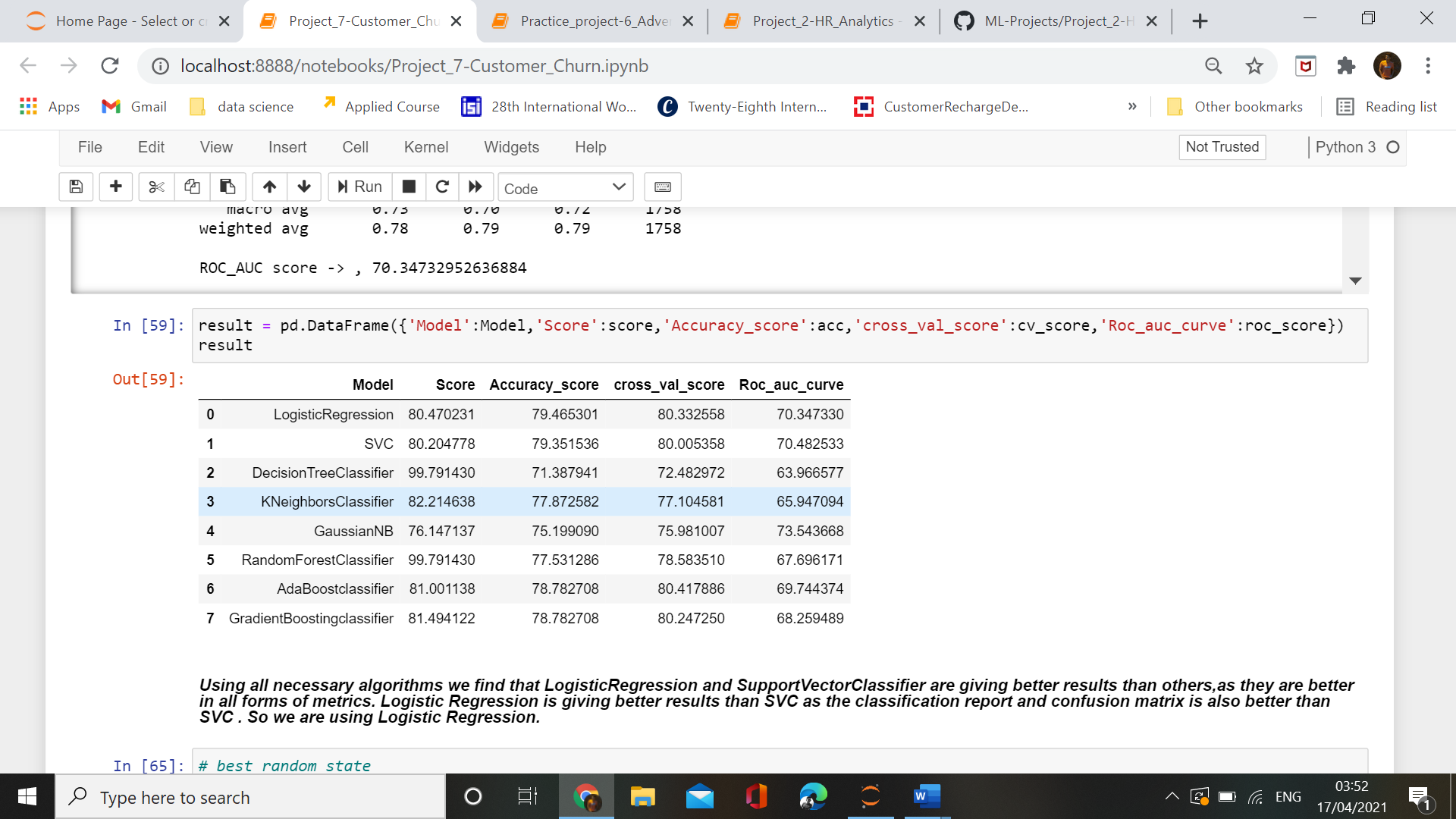
Spllitting train and test data, and by using train\_test\_split method for the splitting of dataset into testing and training data, we have 5274 rows for the mode training and 1758 rows for model testing. Using this training and testing data we are going to find the best model which will give best accuracy for our prediction.

In our dataset we are using StandardScaler for scaling our dataset so that model will not be biased. As if some features contains higher values than other features ,in this condition our model will start giving more preference to those columns which have higher values i.e why we are using standardsclaer method for scaling our data.

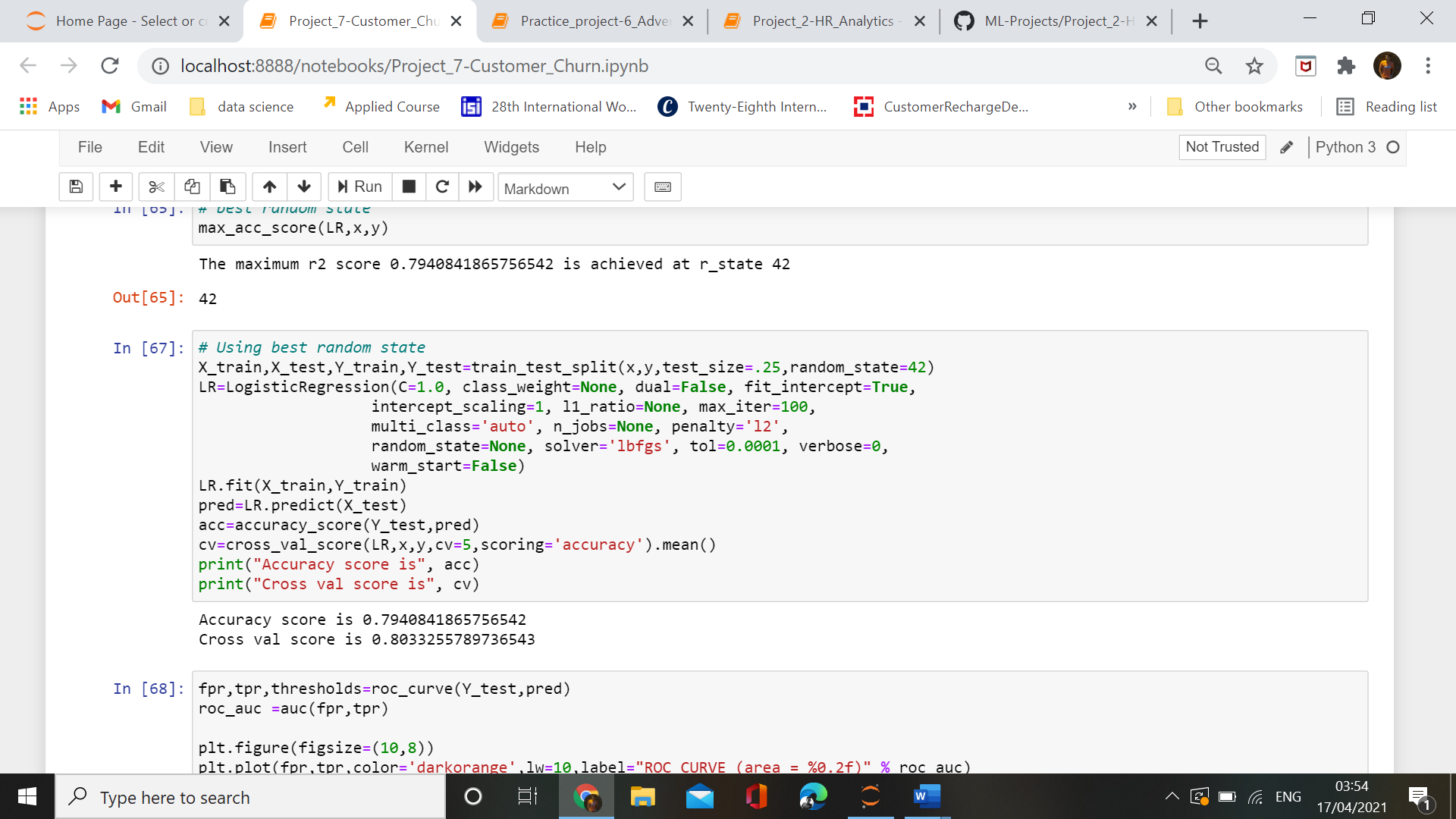
**Setting up method and GridSearchCV for model perfection.**



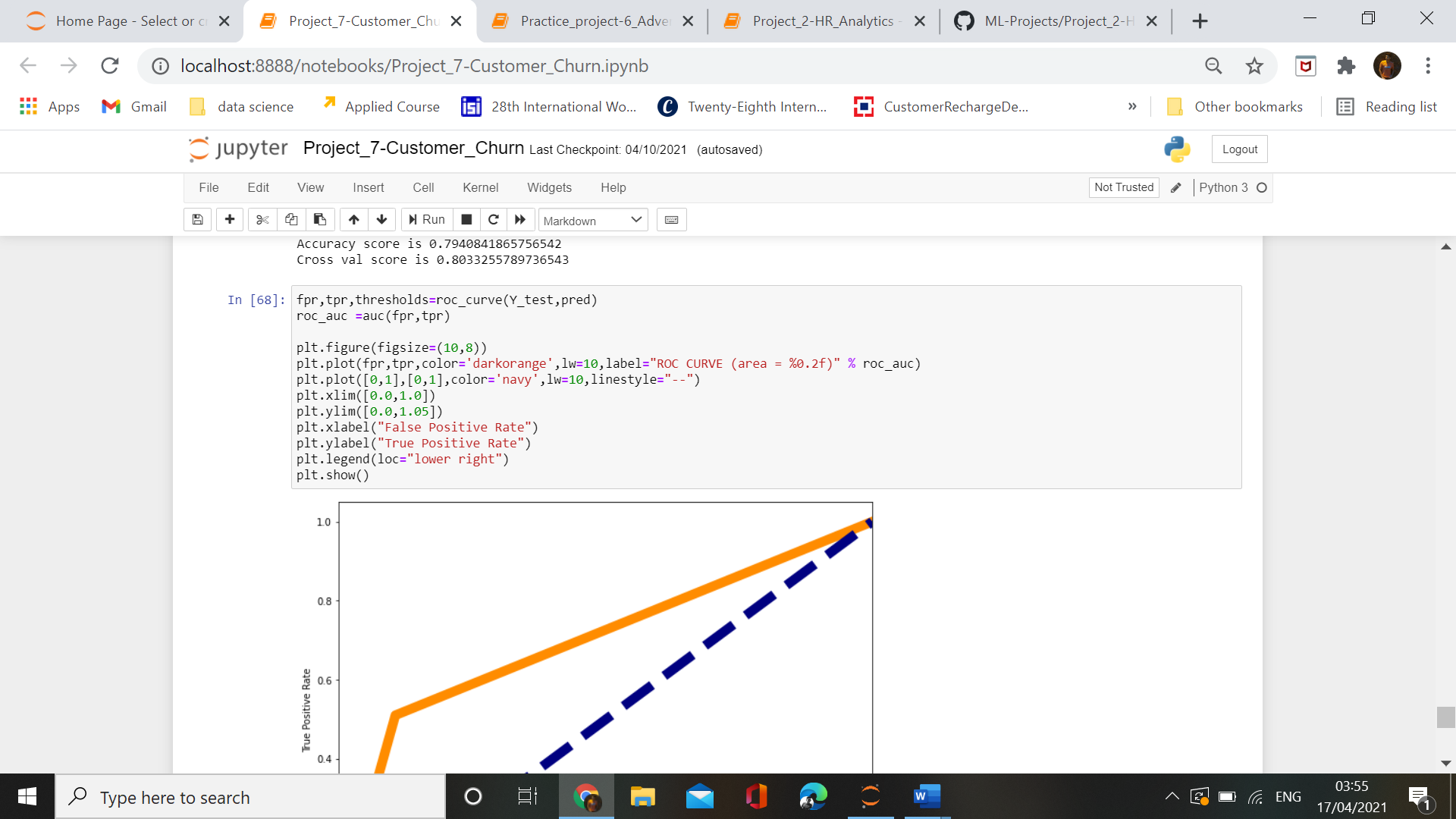
We are using GridSearchCV for the best accuracy of models.

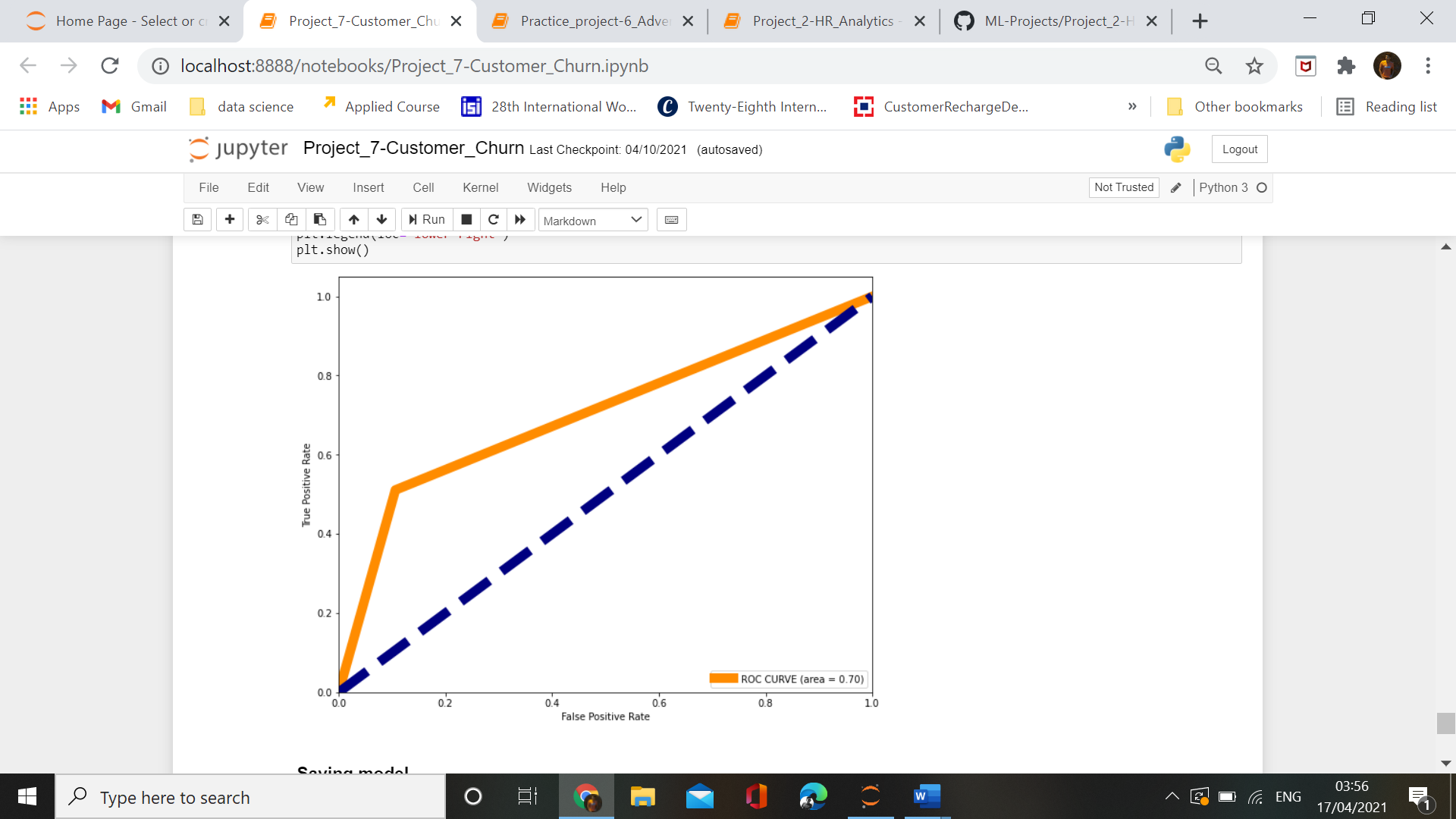


Using all necessary algorithms we find that LogisticRegression and SupportVectorClassifier are giving better results than others,as they are better in all forms of metrics. Logistic Regression is giving better results than SVC as the classification report and confusion matrix is also better than SVC . So we are using Logistic Regression.



By finalising the LogisticRegression ,we find that the accuracy\_score of our model is 79% while cross\_validation\_score is 80%.This model is giving higher accuracy and all other type of metrics necessary for the accurate perfection of new data.





**Conclusion:-**

From the above process we find out that:-

1. EDA helps in reading and understanding data more clearliy than the tabular data.

2. Using EDA we can easily find out missing values and outliers as well as the behaviour of our data.

3. using machine learning we find out the solution of customer churn that the company shoulf focus on its policies towards the charges based on the size of contract .