**Target Protein Prediction Model**

Target Protein Prediction model can be used to predict the bioactivity of the protein which may be an enzyme or an inhibitor or promotor or enhancer depending on the location were it is present. Prediction Models built with the help of various supervised and unsupervised regression models have a huge application in drug discovery and protein bioactivity prediction. Here in this project, I have predicted the bioactivity of HER2 gene in breast cancer.

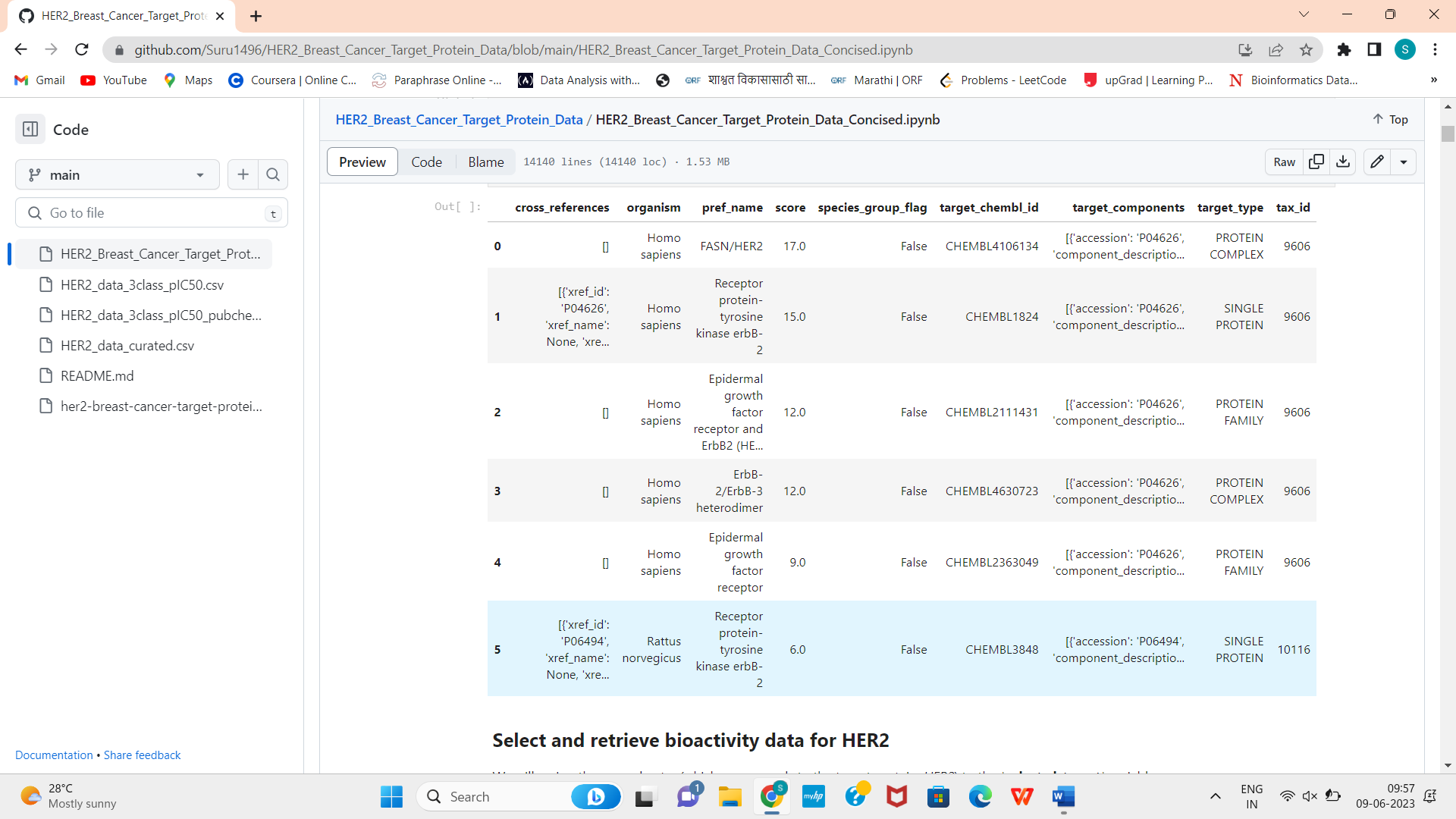
**Introduction: -**

HER2 stands for human epidermal growth factor receptor 2. It is a gene that makes a protein found on the surface of all the breast cells. It is involved in the normal cell growth and also promotes growth of cancer cells. HER2 is a membrane tyrosine kinase and the oncogene that is overexpressed and gene amplified in about 20% of breast cancers. When activated it provides the cell with potent proliferative and anti-apoptosis signals and is major driver of tumor development and progression for this subset of breast cancer.

In order to create prediction model, we need to follow several steps :-

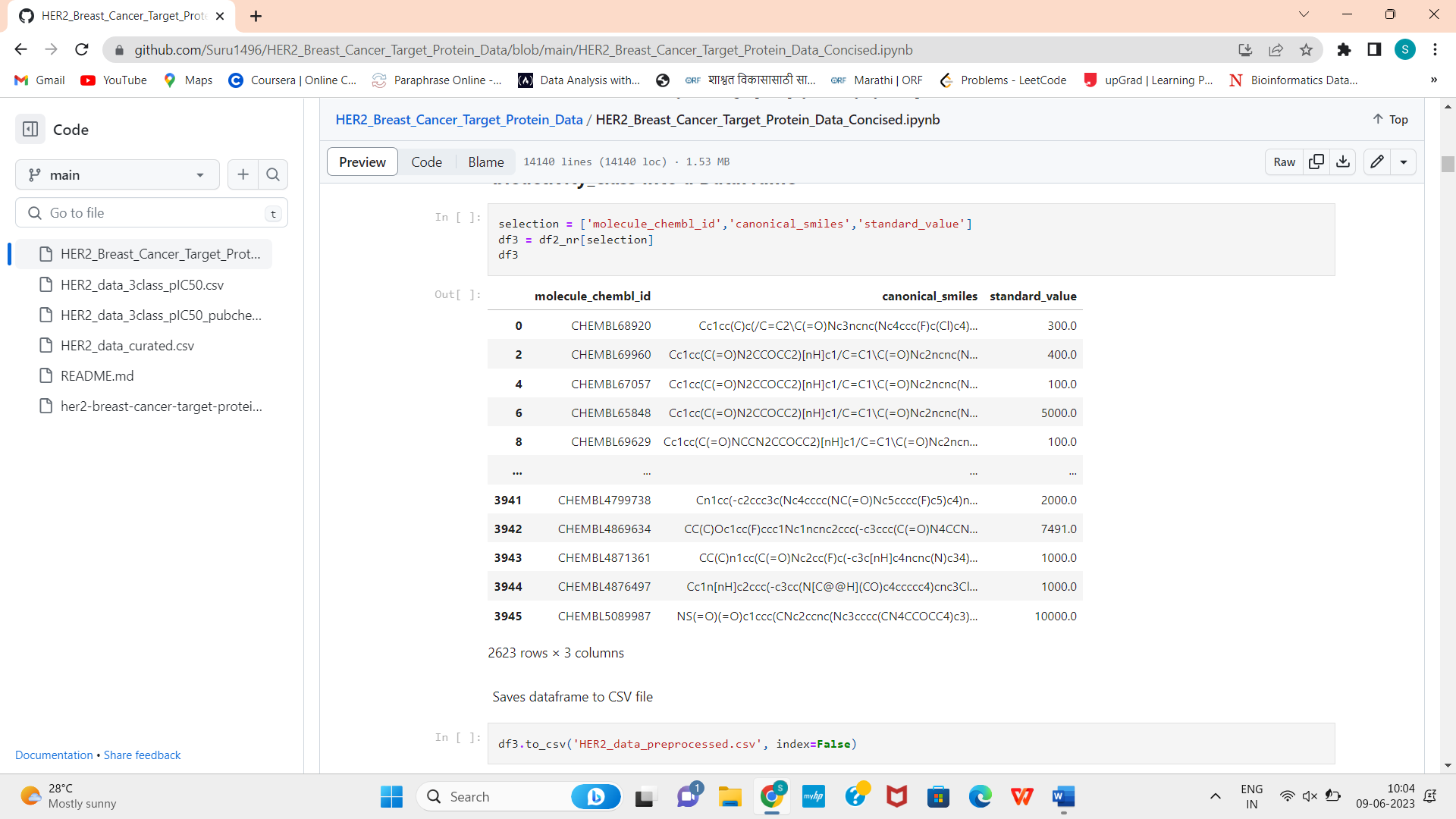
1. Data collection
2. Data pre-processing
3. Descriptor Calculation
4. Exploratory data analysis
5. Model Building
6. Model Comparison
7. Data visualization
8. Data Collection:

In order the curated data for the target protein or a gene we need to install ChEMBL web service package. ChEMBL database is a database that contains curated bioactivity data for more than 2 million compounds. After importing the necessary libraries like pandas ,ChEMBLwe will be able to import the data from ChEMBL database.

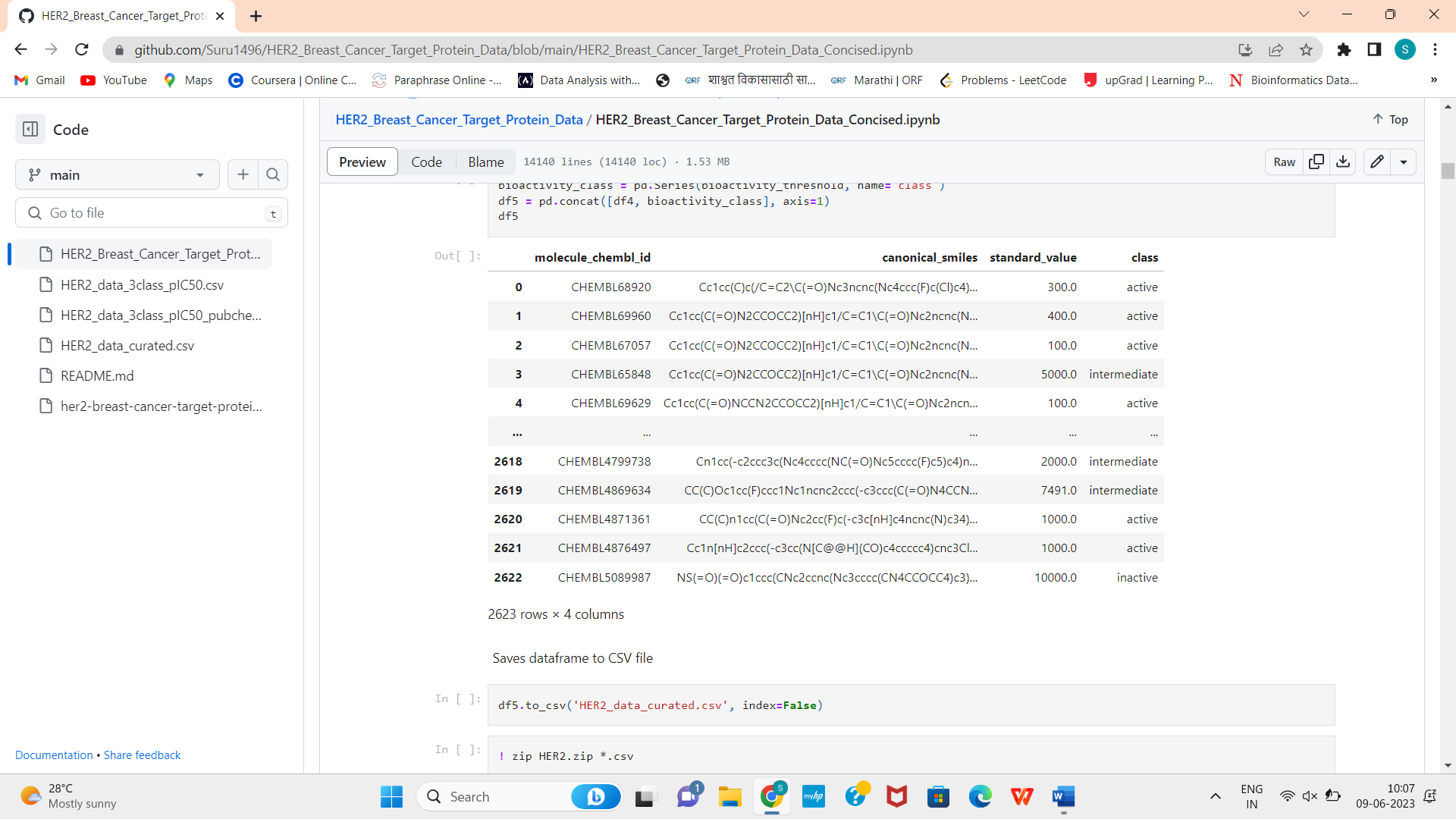


1. Data Pre-Processing:

Now select the target variable from the imported data. We will retrieve only bioactivity data for HER2 (CHEMBL1824) that are reported as pChEMBL values. If any compounds have missing value for the **standard\_value** and **canonical\_smiles** column then drop it.



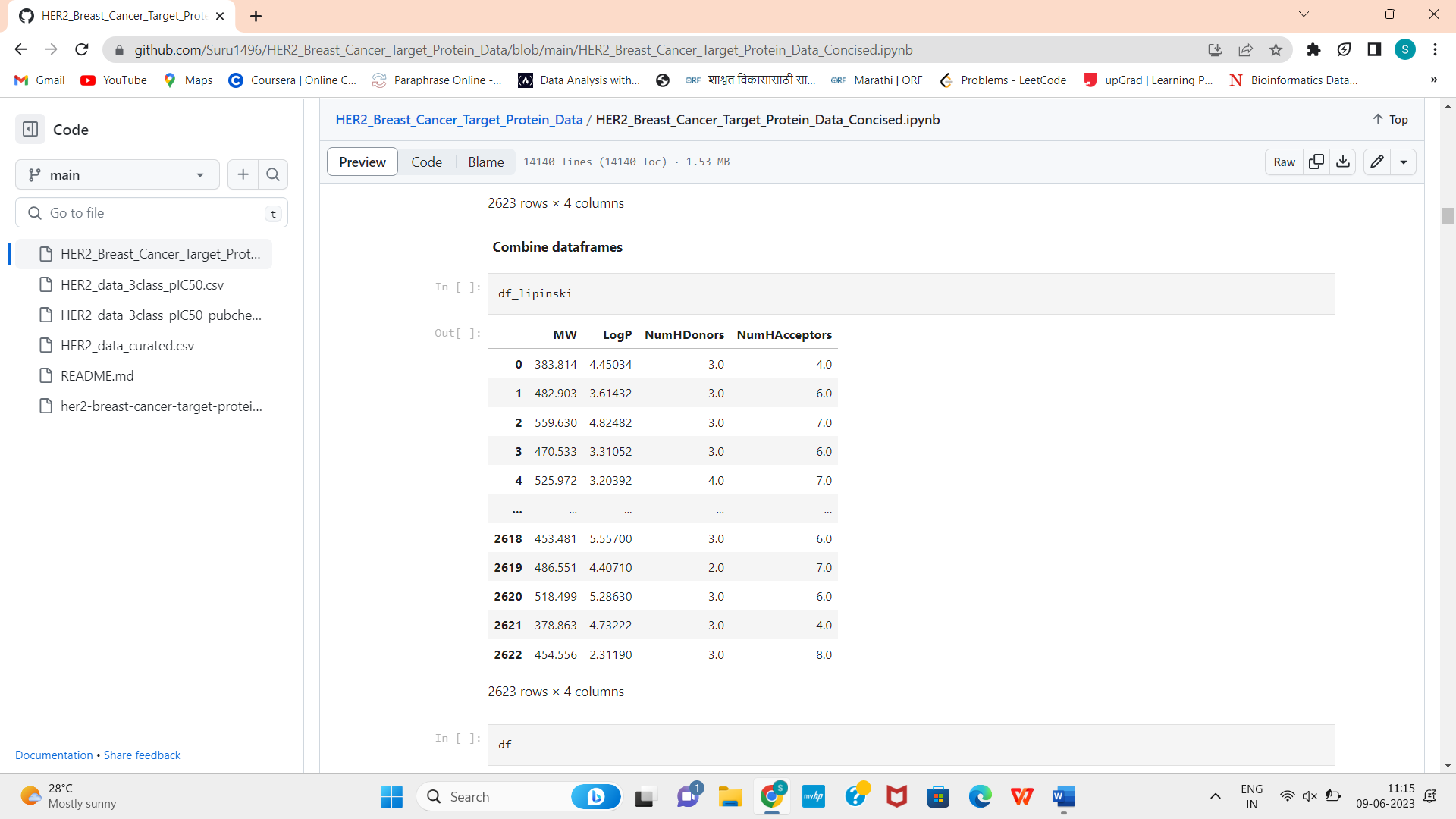
The bioactivity data is in the IC50 unit. Compounds having values of less than 1000 nM will be considered to be **active** while those greater than 10,000 nM will be considered to be **inactive**. As for those values in between 1,000 and 10,000 nM will be referred to as **intermediate**.



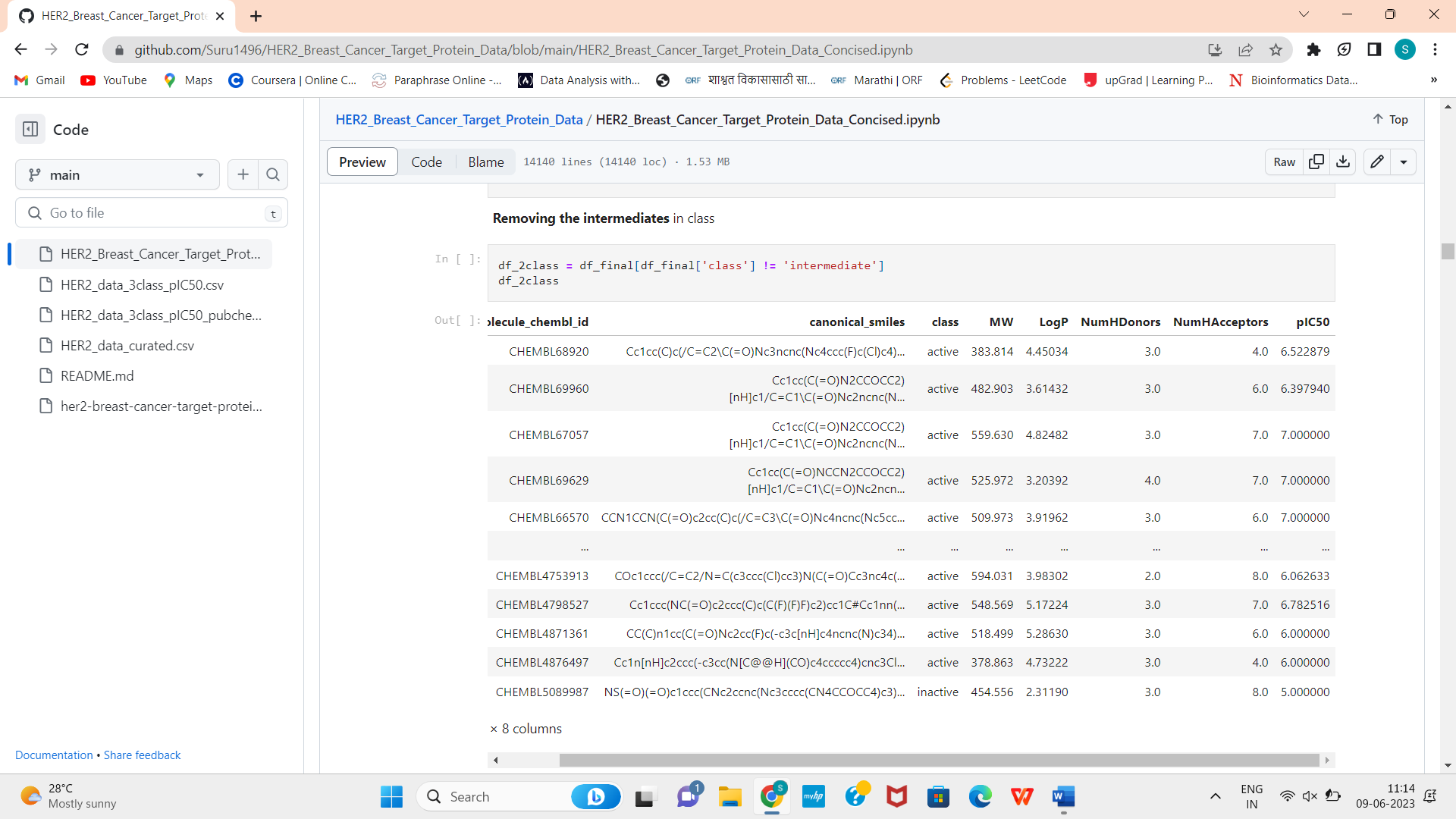
1. Descriptor Calculation:

To perform descriptor Calculations for predictions, we use here lipinski’s rule. Lipinski states that the molecular weight of the drug should be less than 500 Daltons and octanol-water partition coefficient (LogP) should be less than 5.3 Hydrogen bond donors which should be less than 5.4 Hydrogen bond acceptors which is again less than 10.

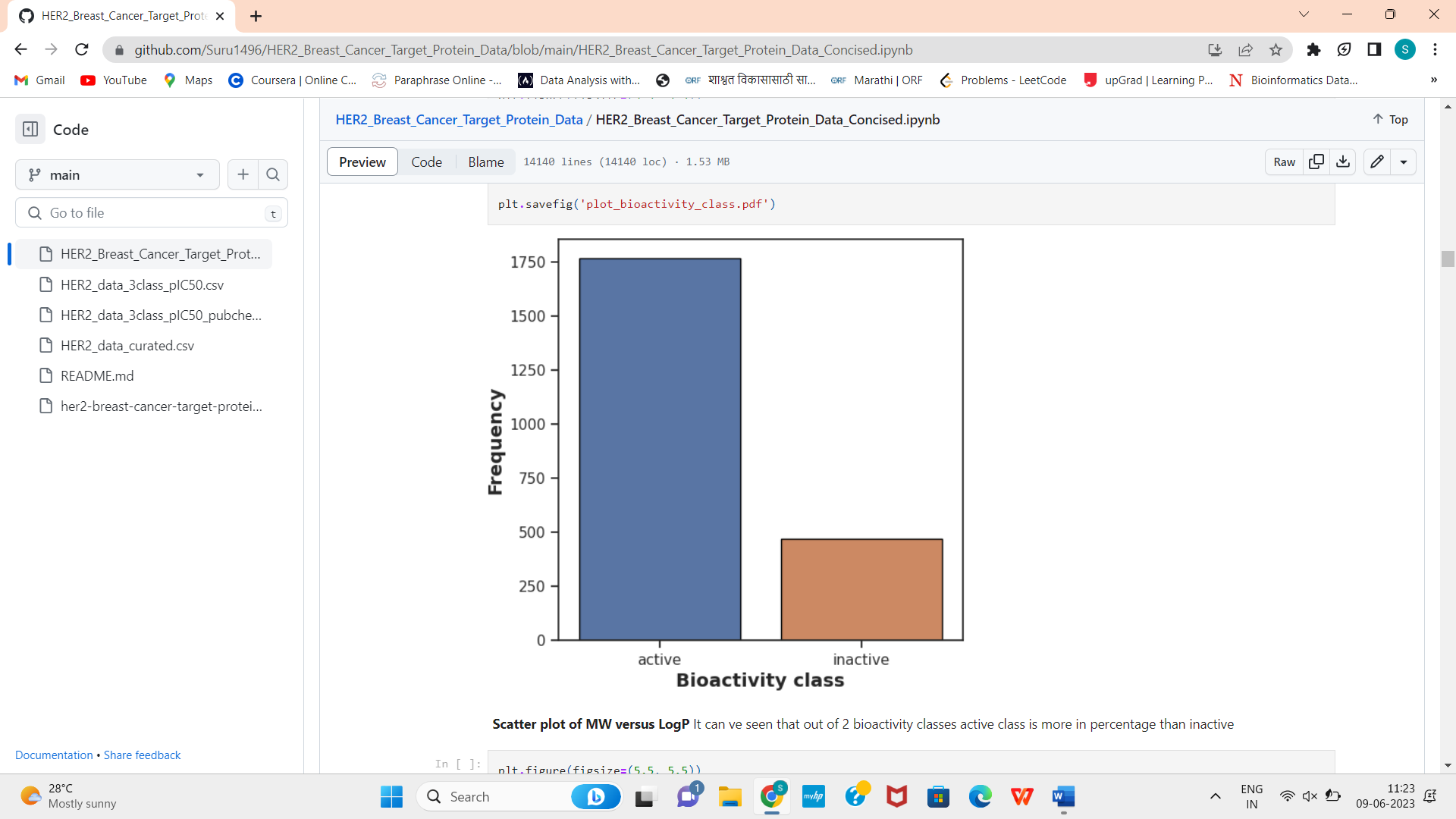
In order to perform Lipinski descriptor, we need to import rdkit package, and then combine both the data frames the one formed recently using descriptor and the original formed.

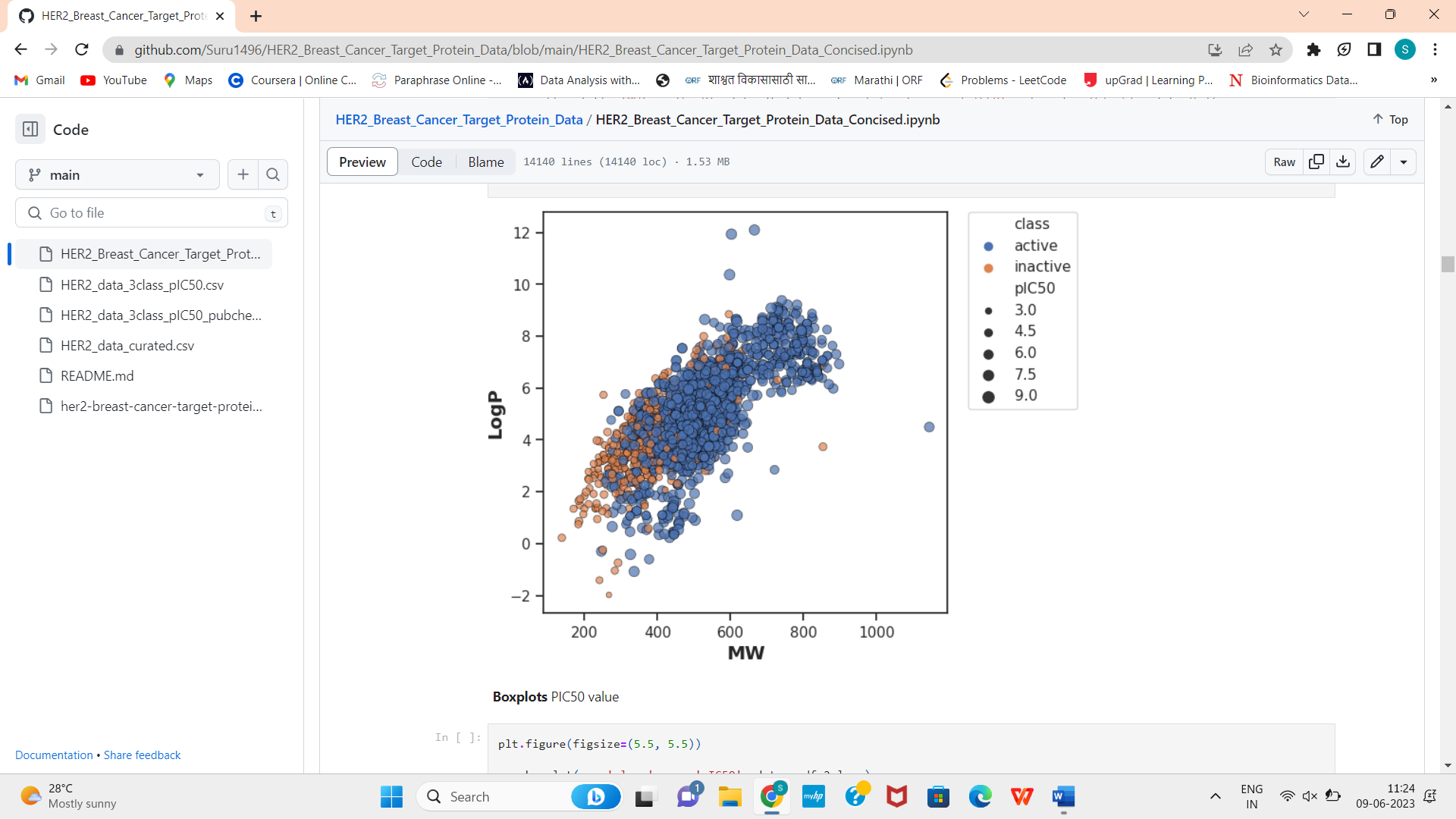


Now we need to convert the IC50 into pIC50 to form a more uniformly distributed data. You need to follow the given procedure in the. ipynb file. At the end of the procedure of descriptor calculation we get a new column of pIC50 added to the combined dataset.

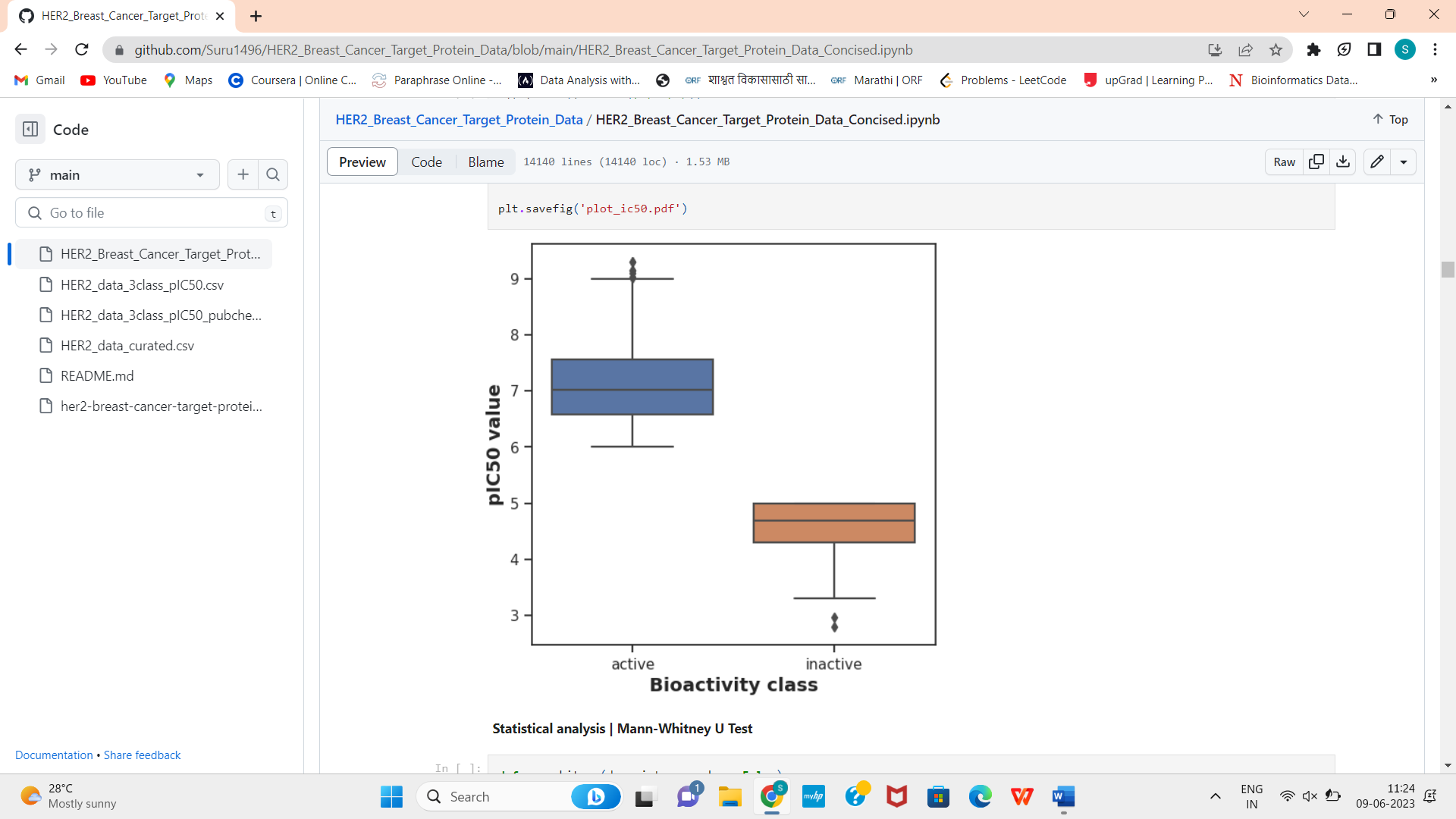


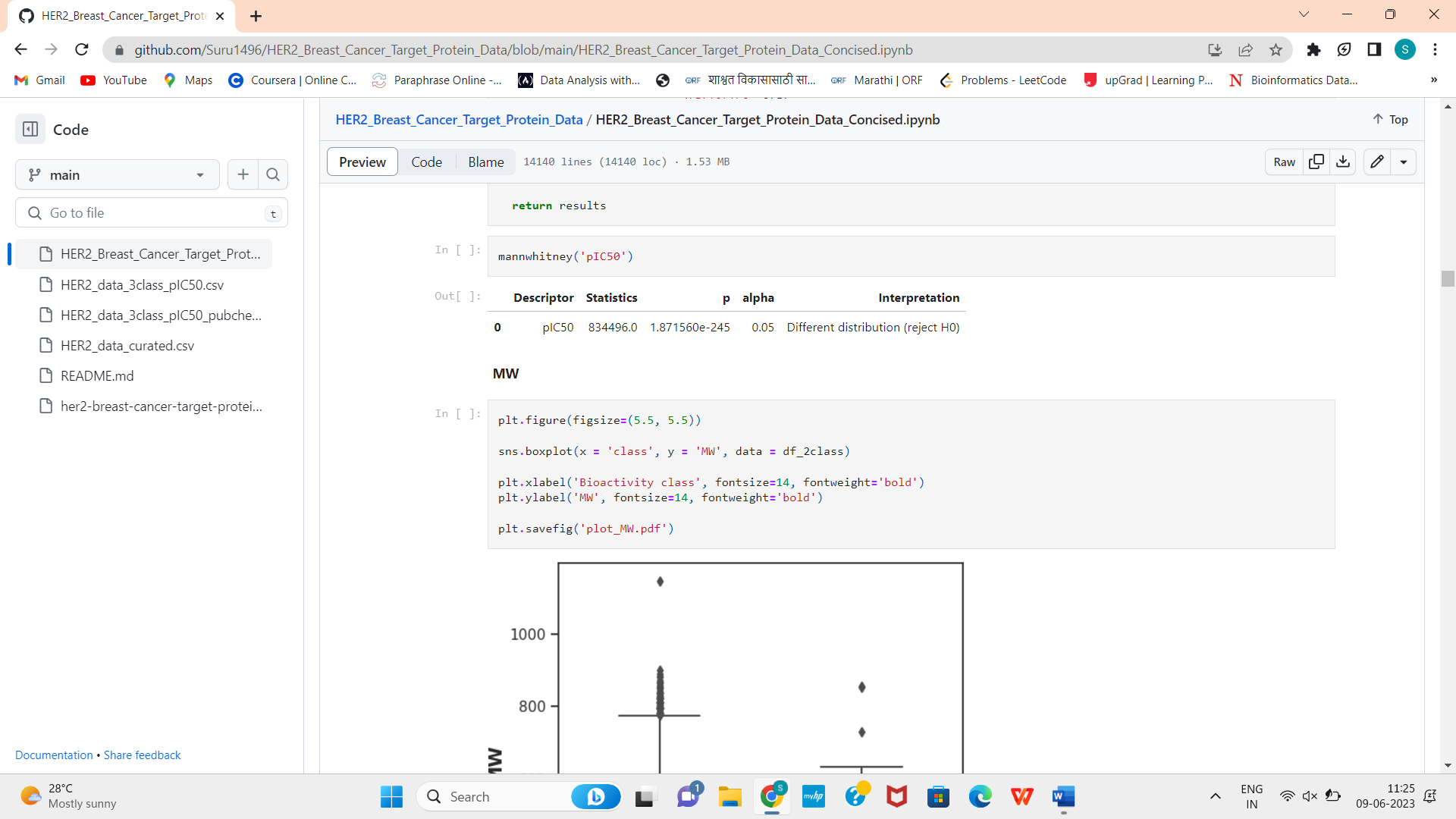
1. Exploratory data analysis:

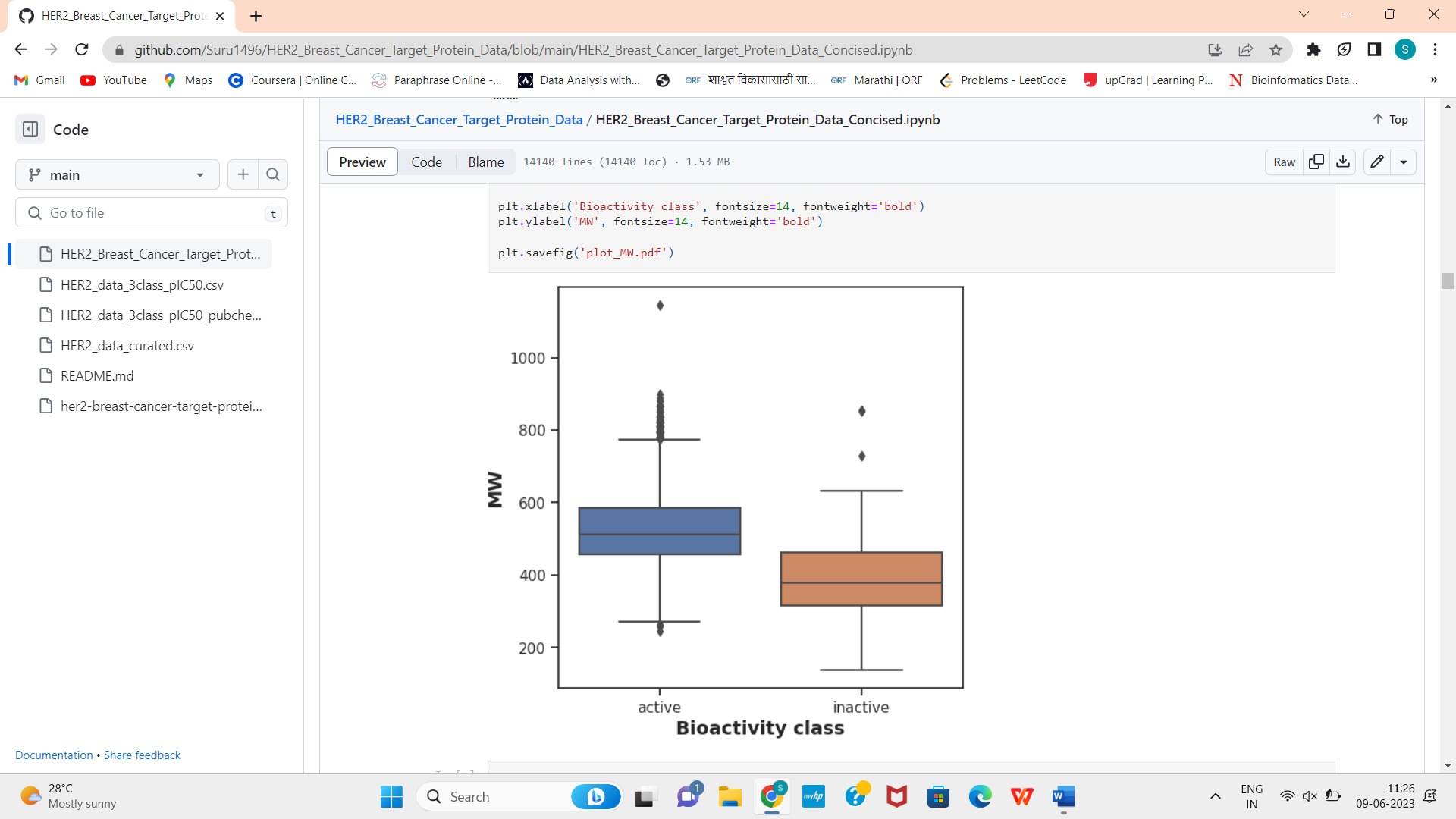


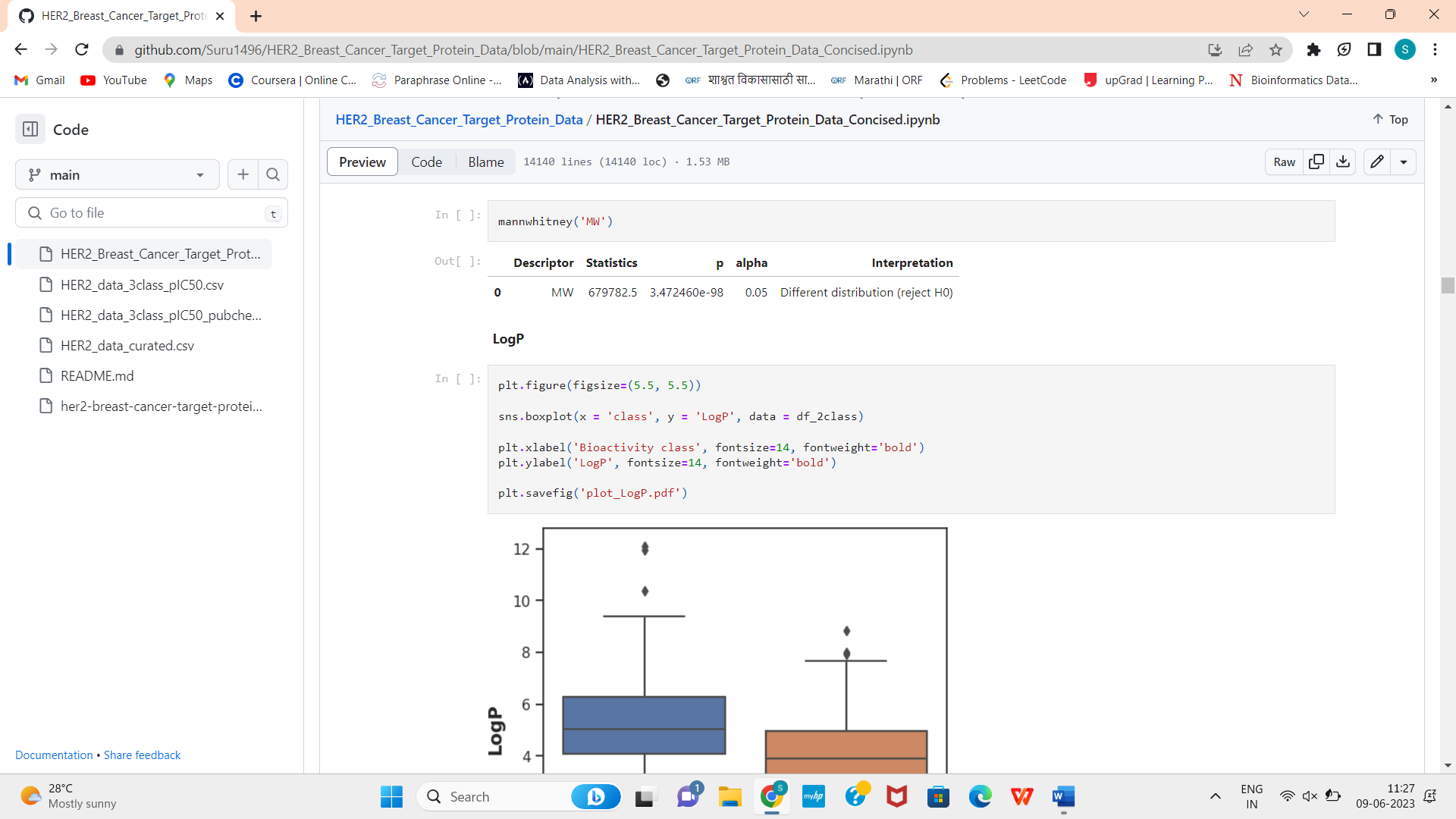


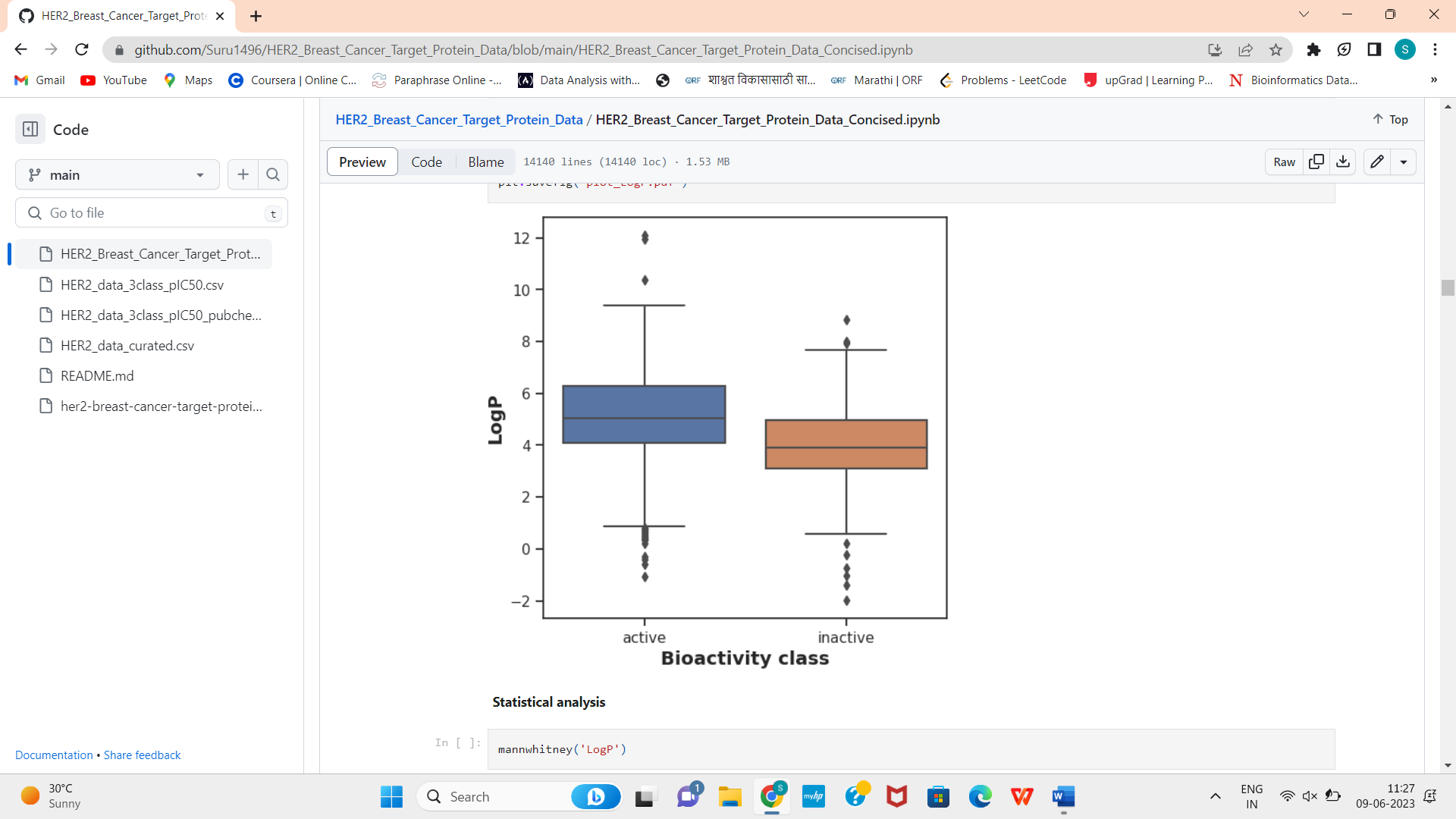
Statistical analysis using Mann-Whitney U test: -

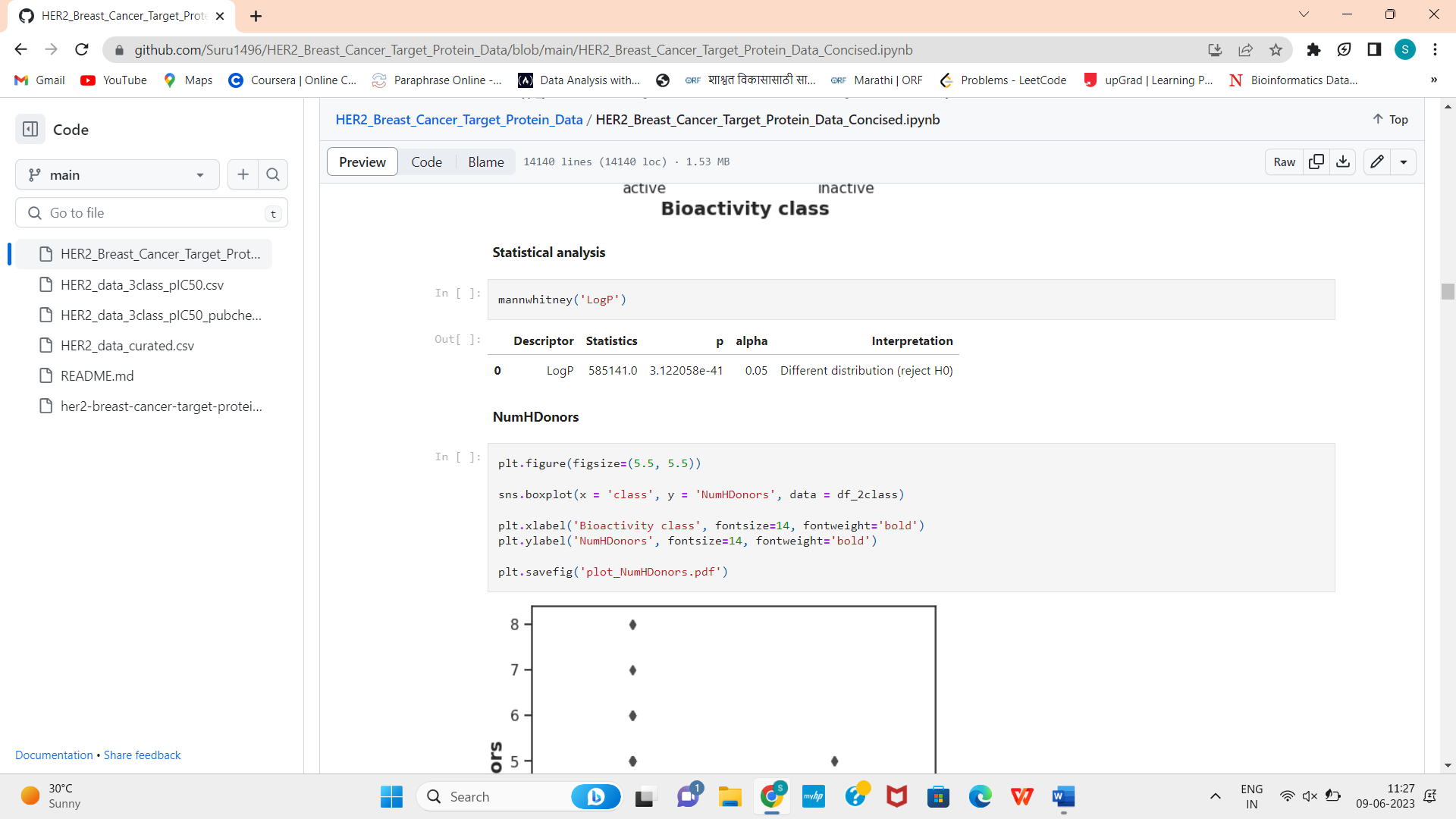


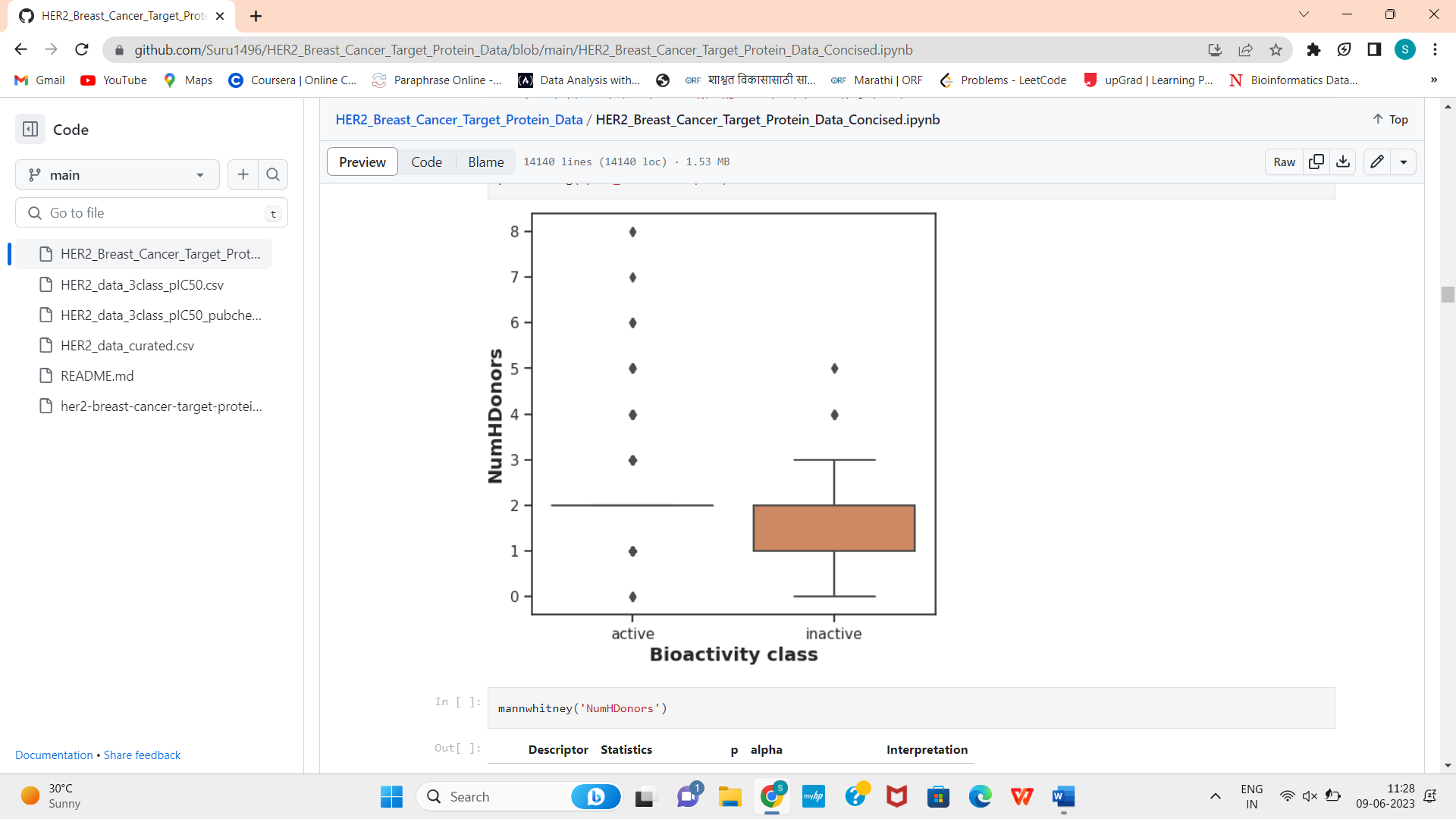


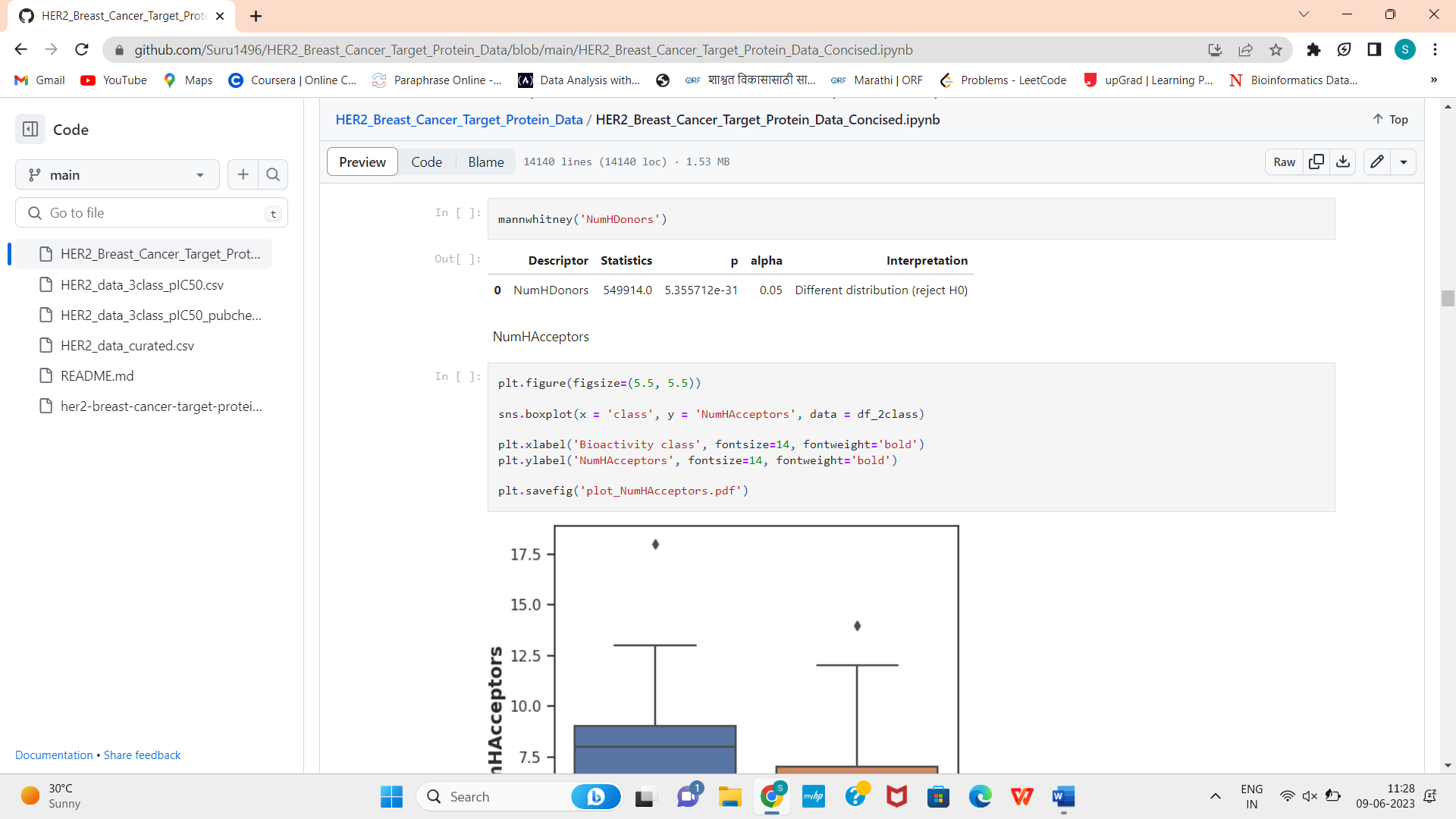


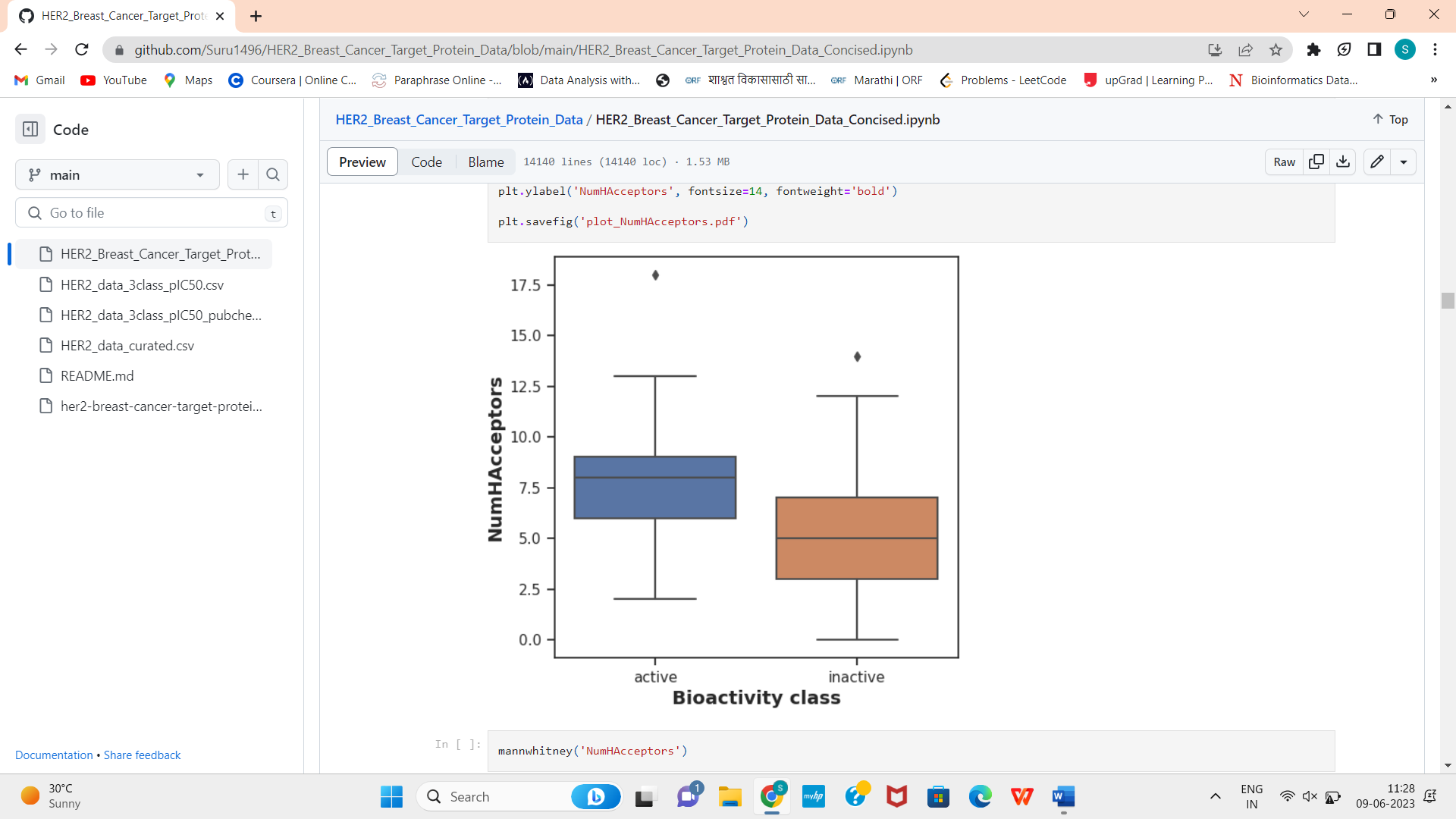


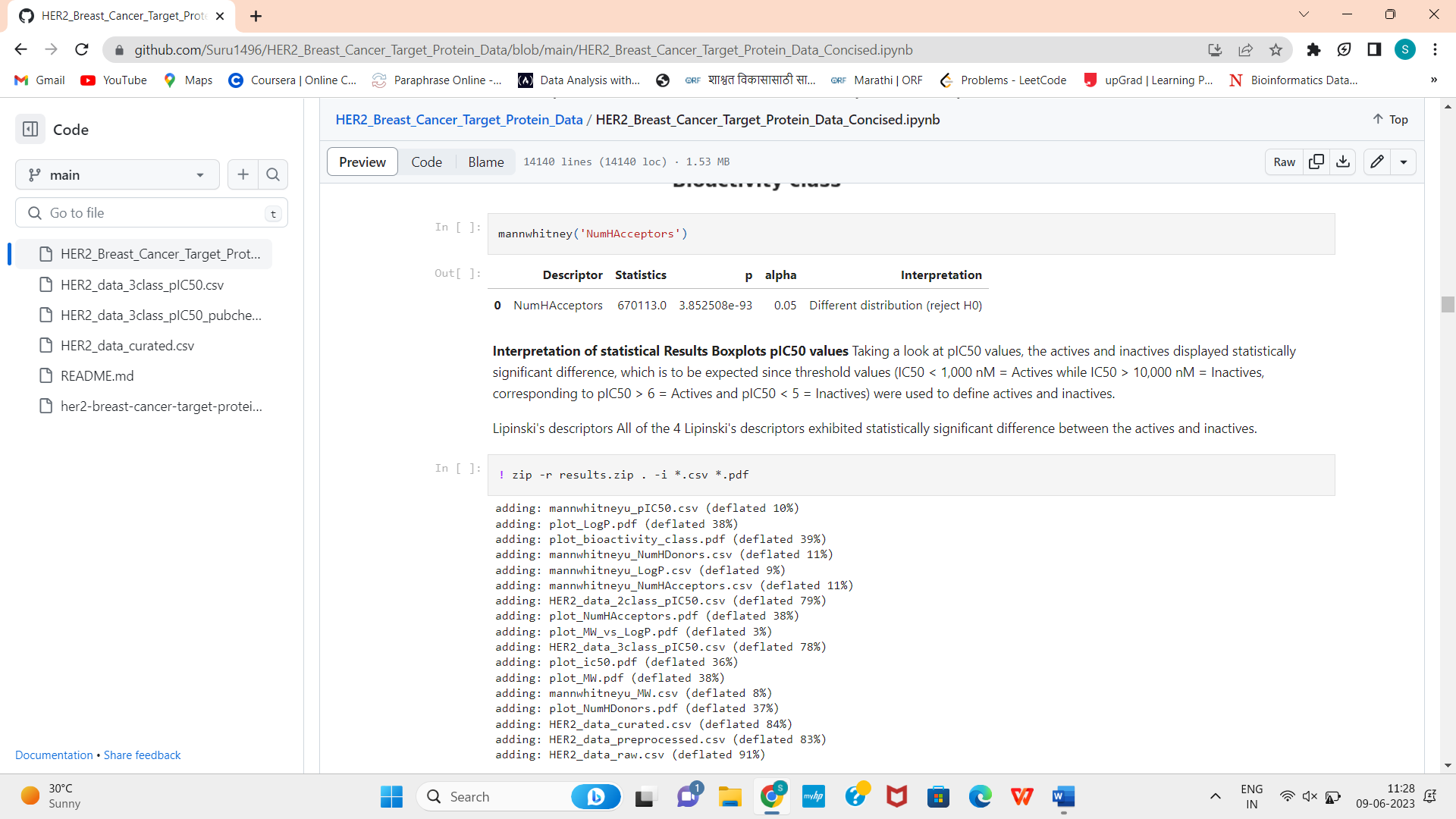










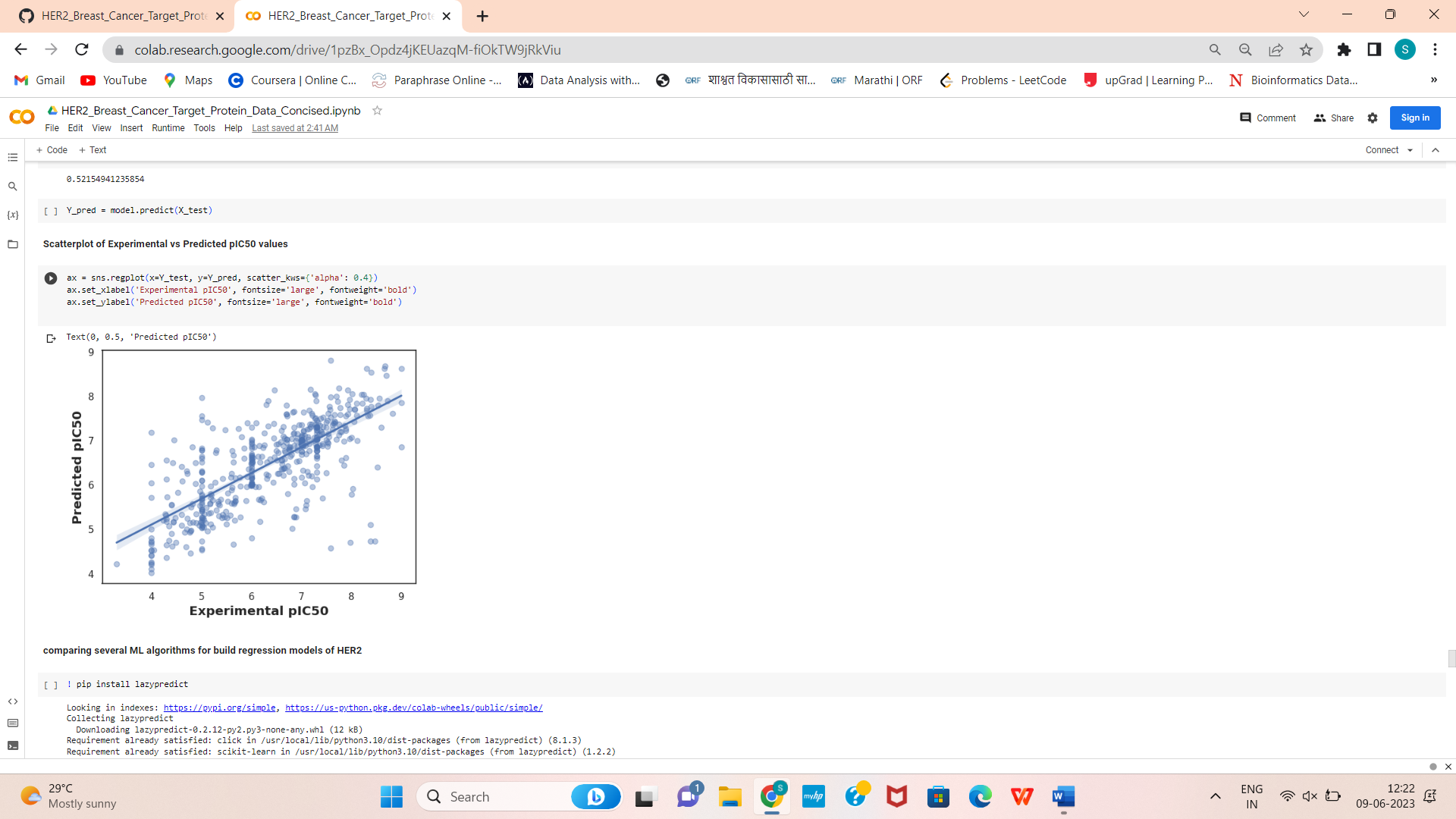


1. Model Building: -

Model building involves several steps which are as follows:

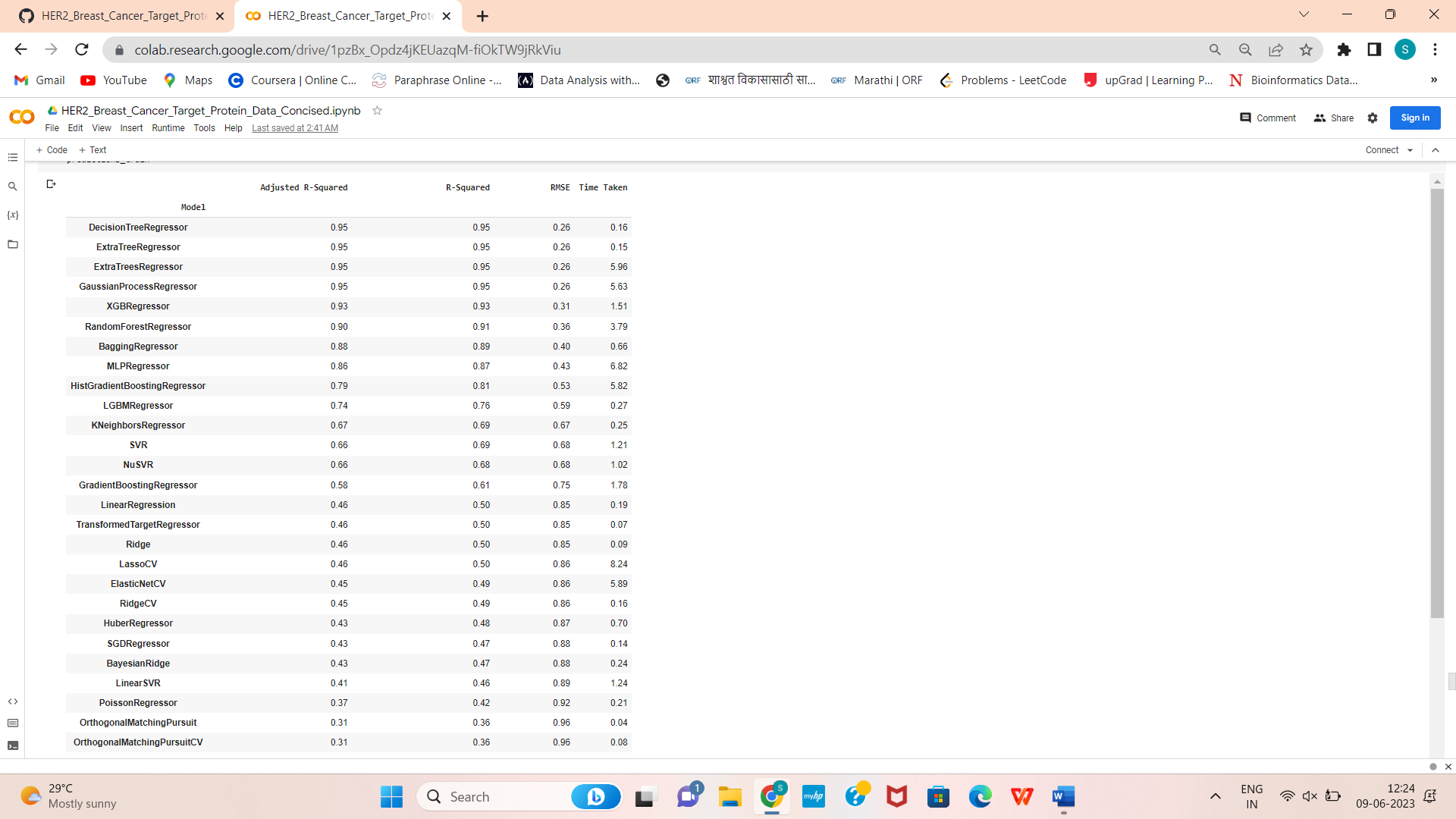
1. Splitting the dataset
2. Choosing the regression model
3. Fitting the model
4. Predicting
5. Find the accuracy score and scatterplot

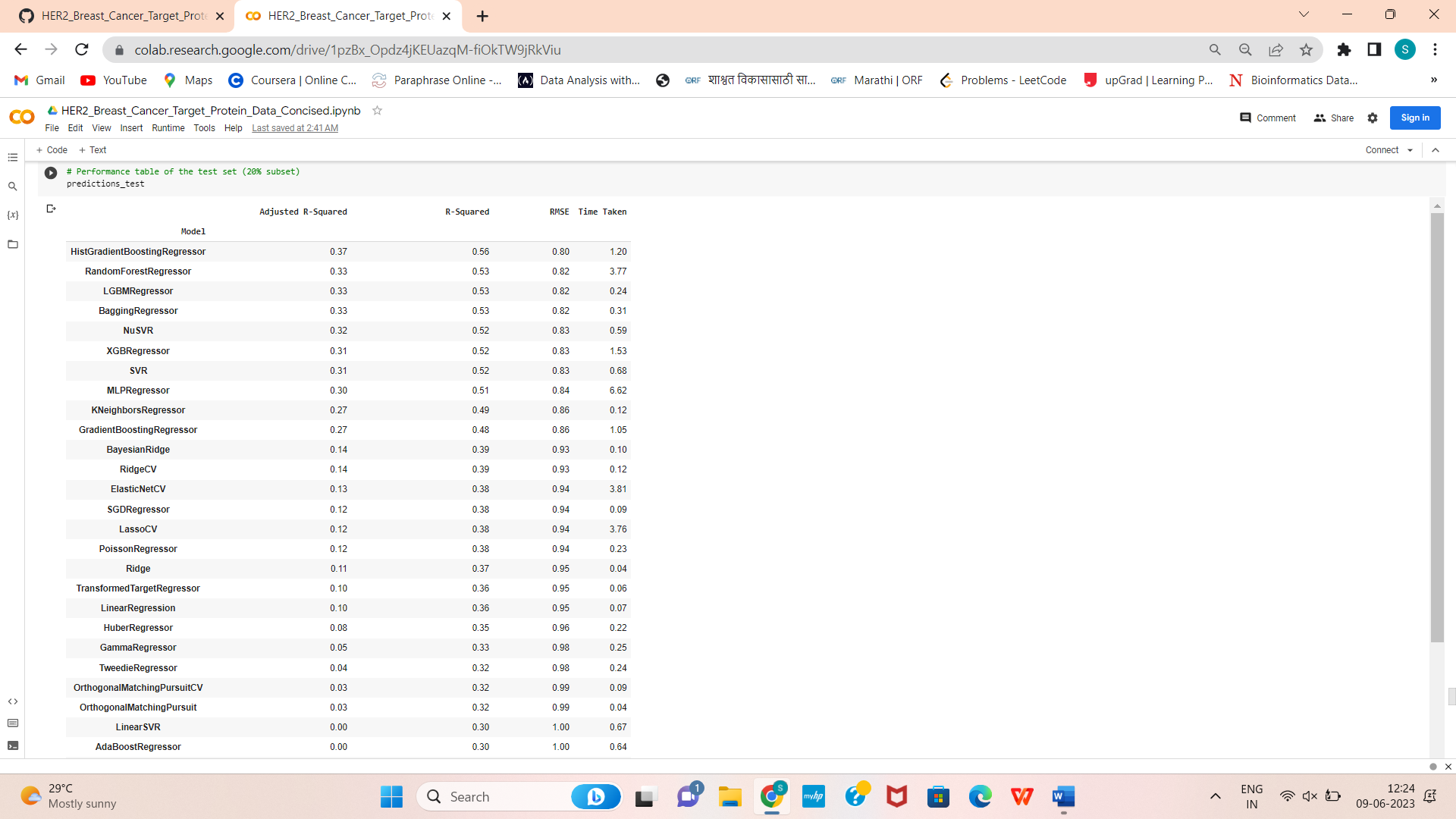
We draw scatterplot in order to analyze the relation between the experimental and predicted data.



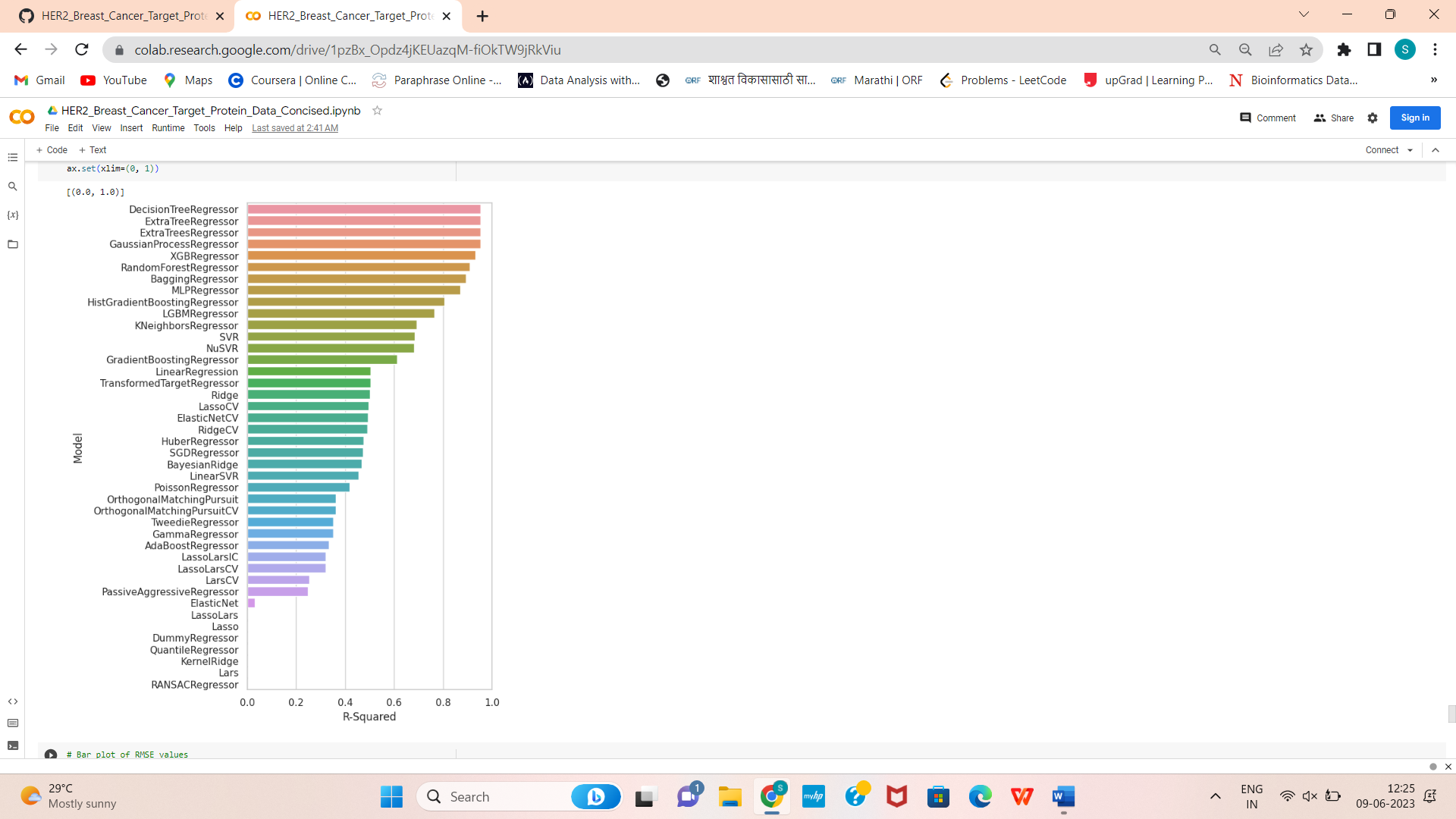
1. Model Comparison: -

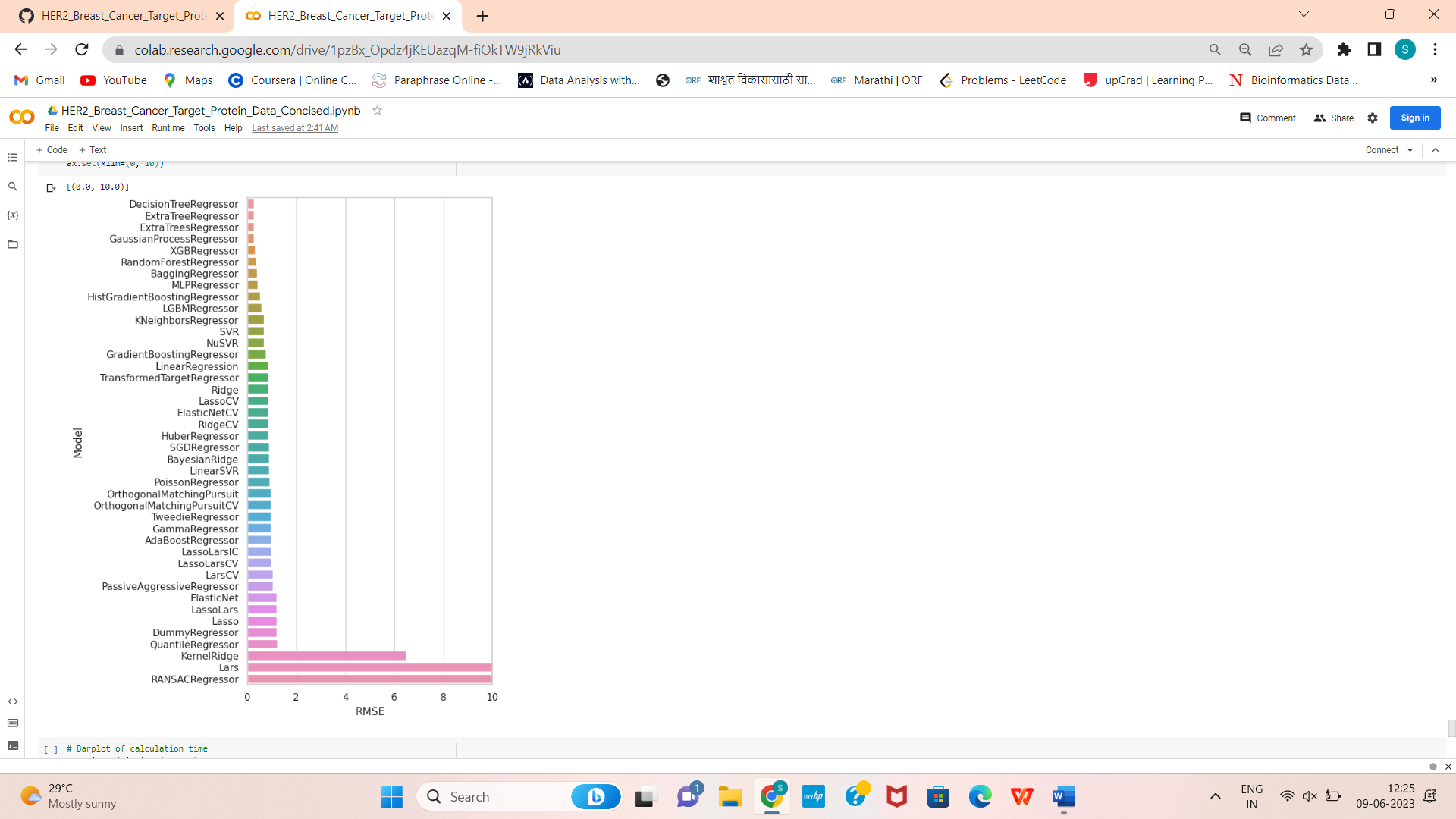
In this step we compare several regression models in order to choose the most appropriate model by comparing their R-squared, RMSE and Time taken values. And then we compare both the train as well test values.

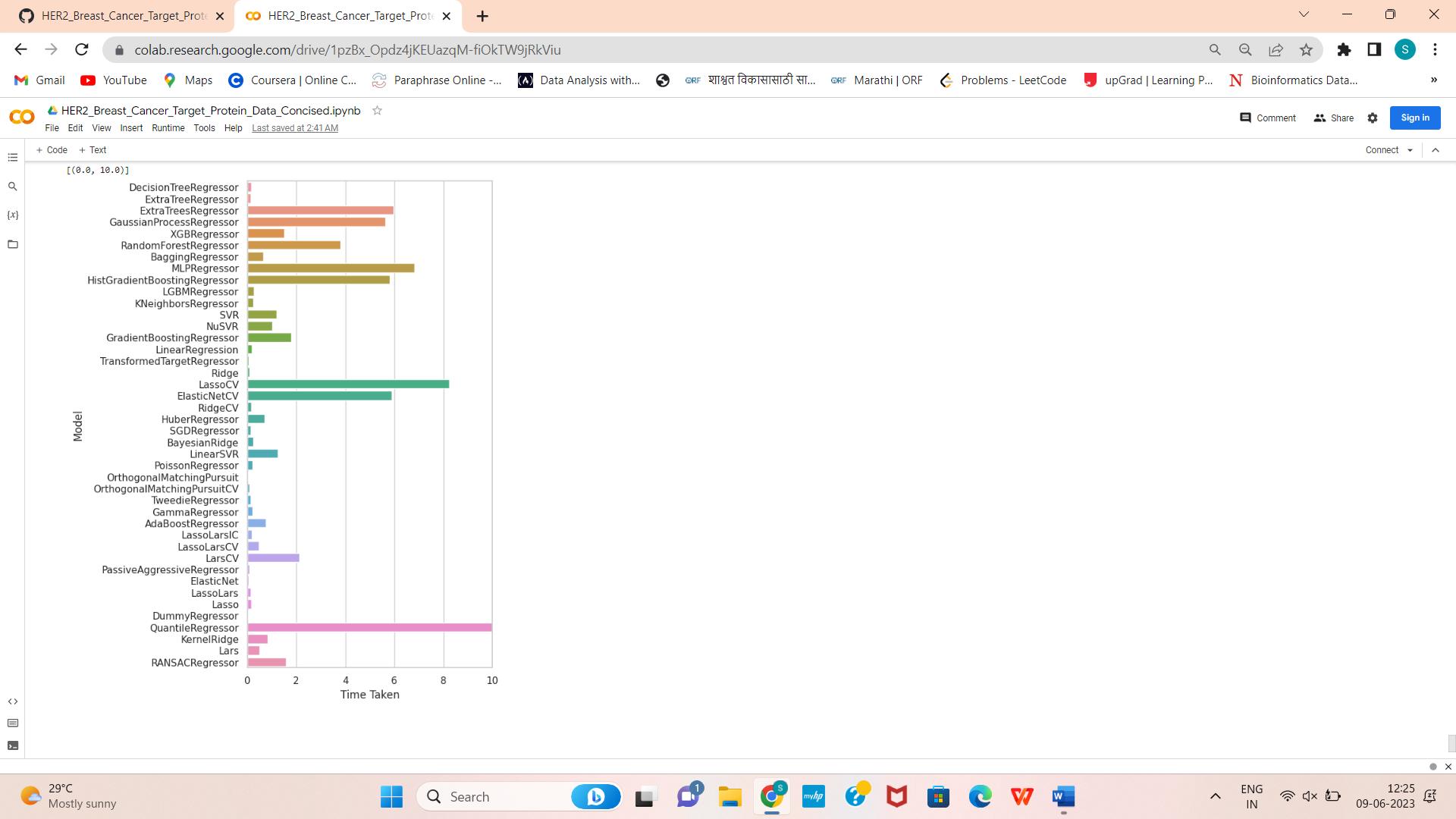




1. Data visualization: -







Through this we can conclude that decision tree model should be more appropriate and efficient to build the predictions of the Target protein. But it will also be time consuming. So to replace it, random forest algorithm can be used.