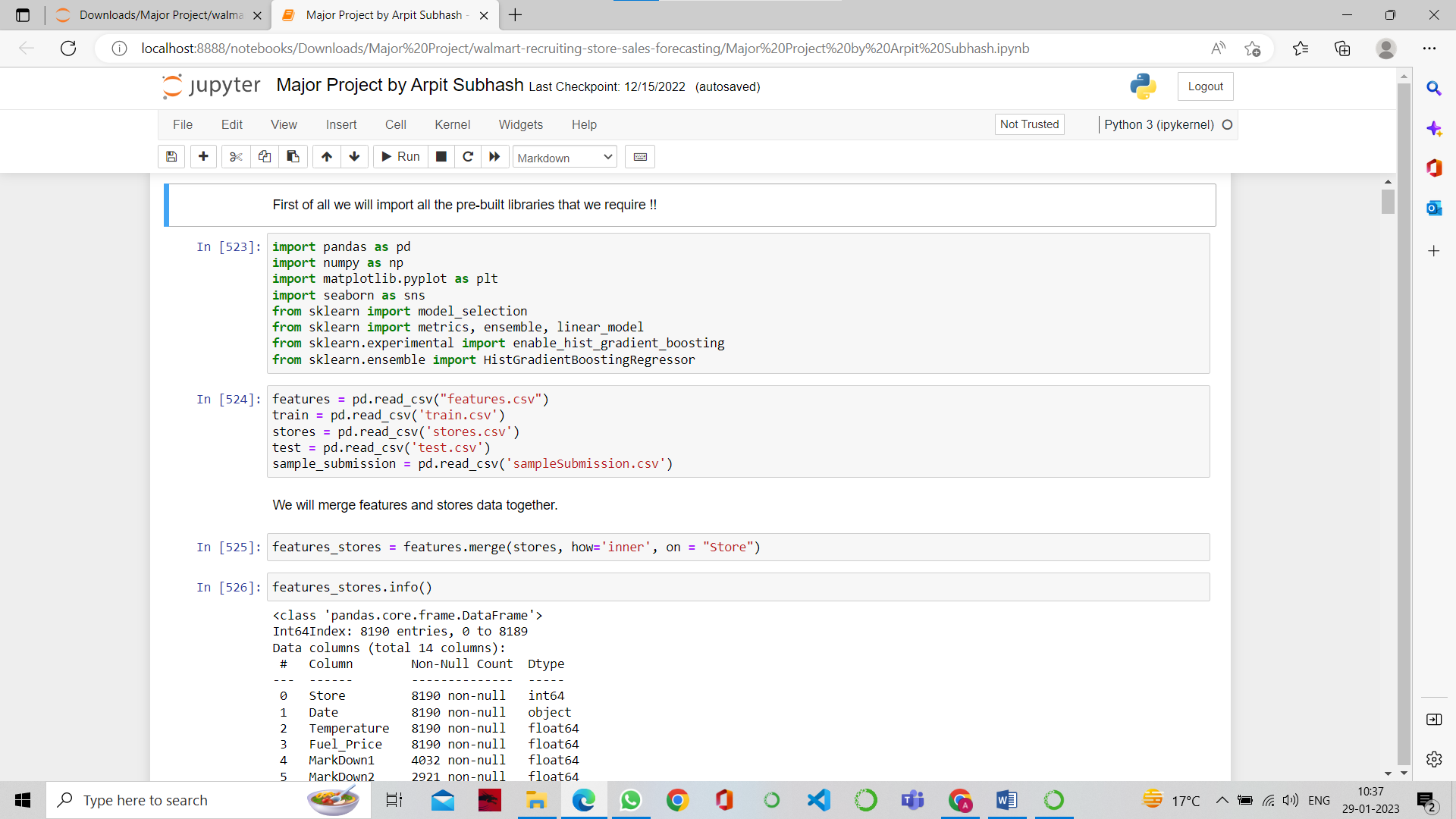
**Report on Data Science project**

Name: Arpit Subhash

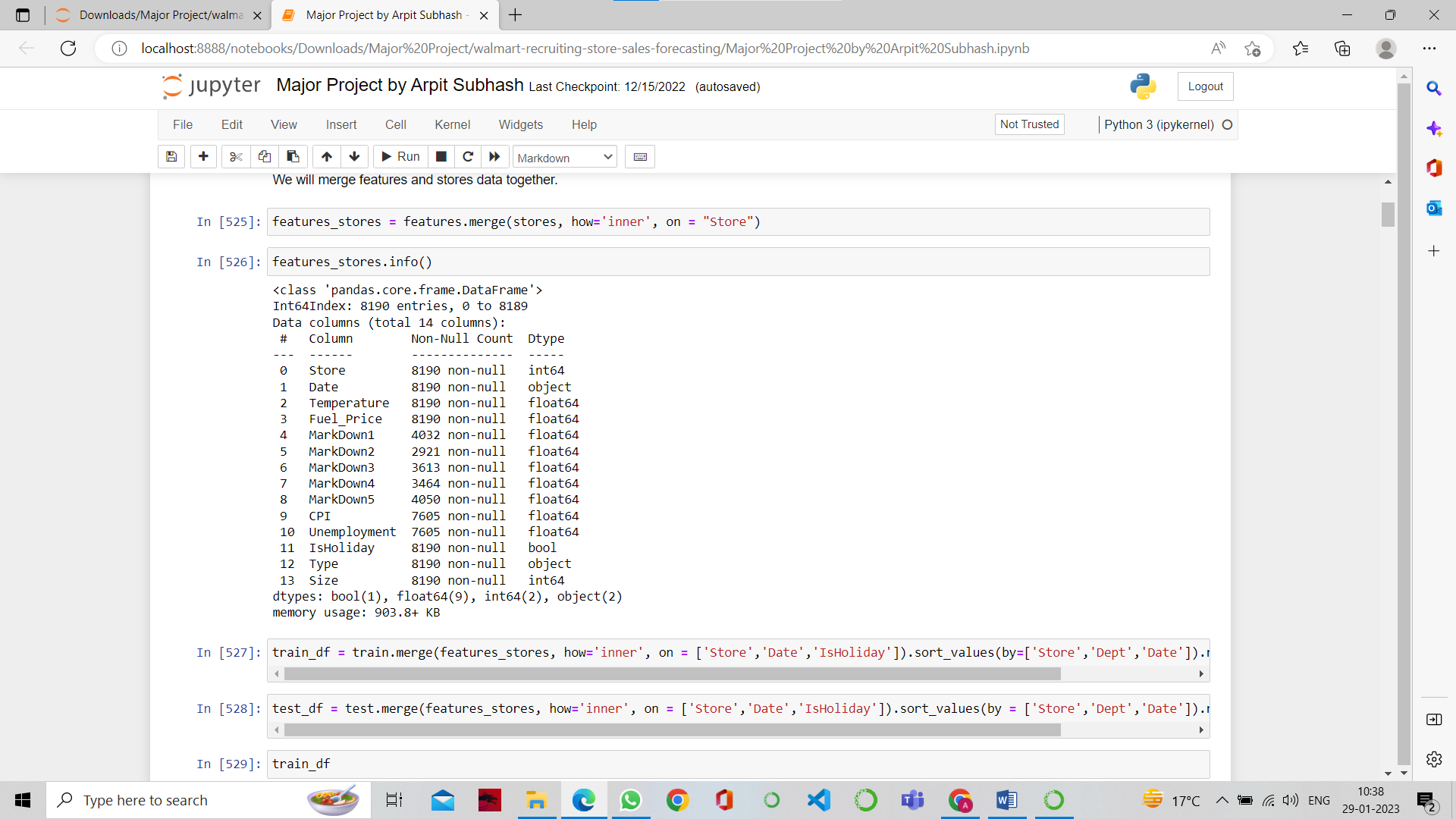
Topic: Prediction on Walmart’s departments of different stores through all of the world

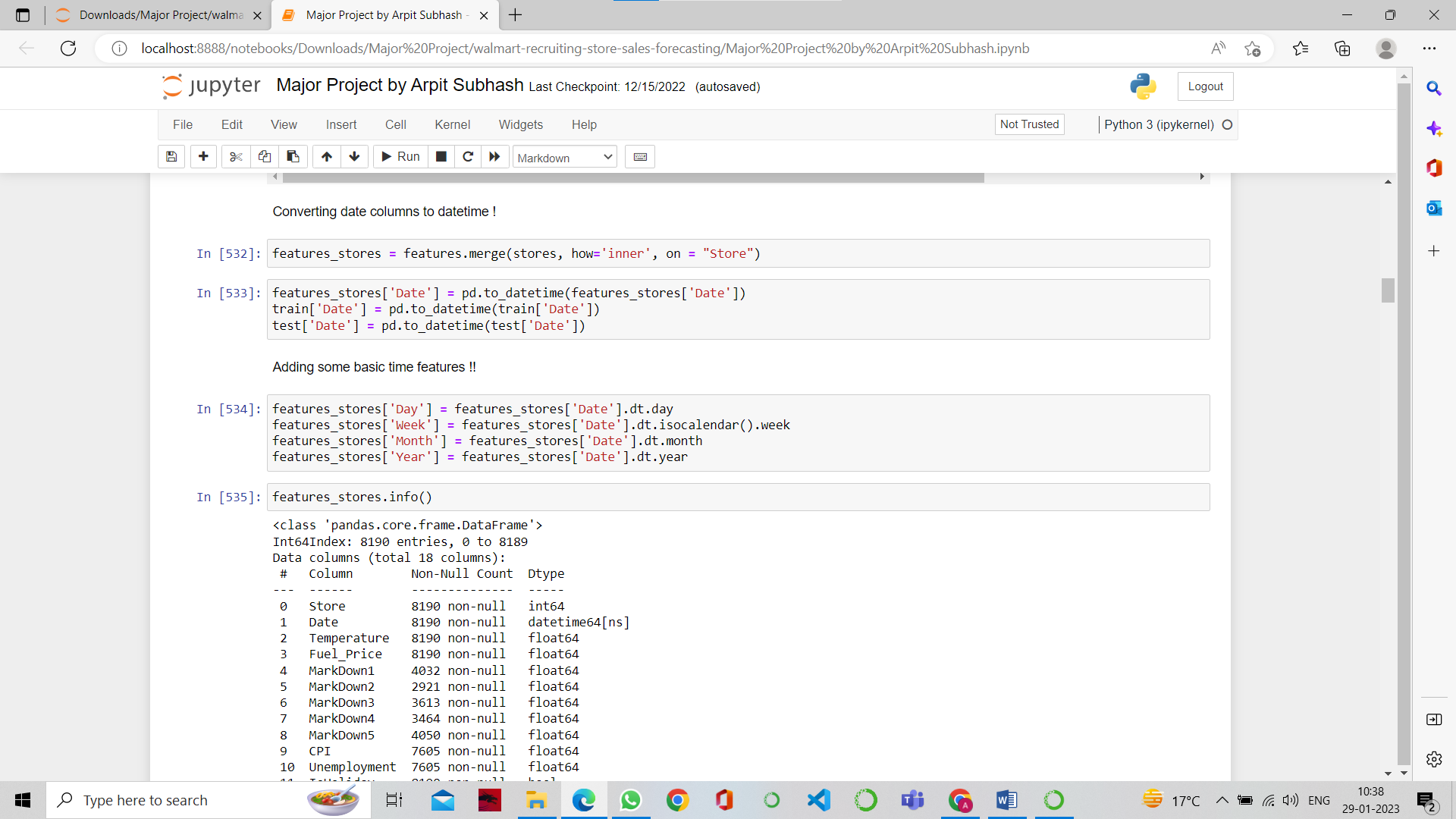
Table of contents:

1. Libraries – I have used libraries like pandas, numpy, matplotlib, seaborn, sklearn kit (model\_selection, ensemble, linear\_model, eli5).
2. Data loading – features, stores, train, test (csv files).

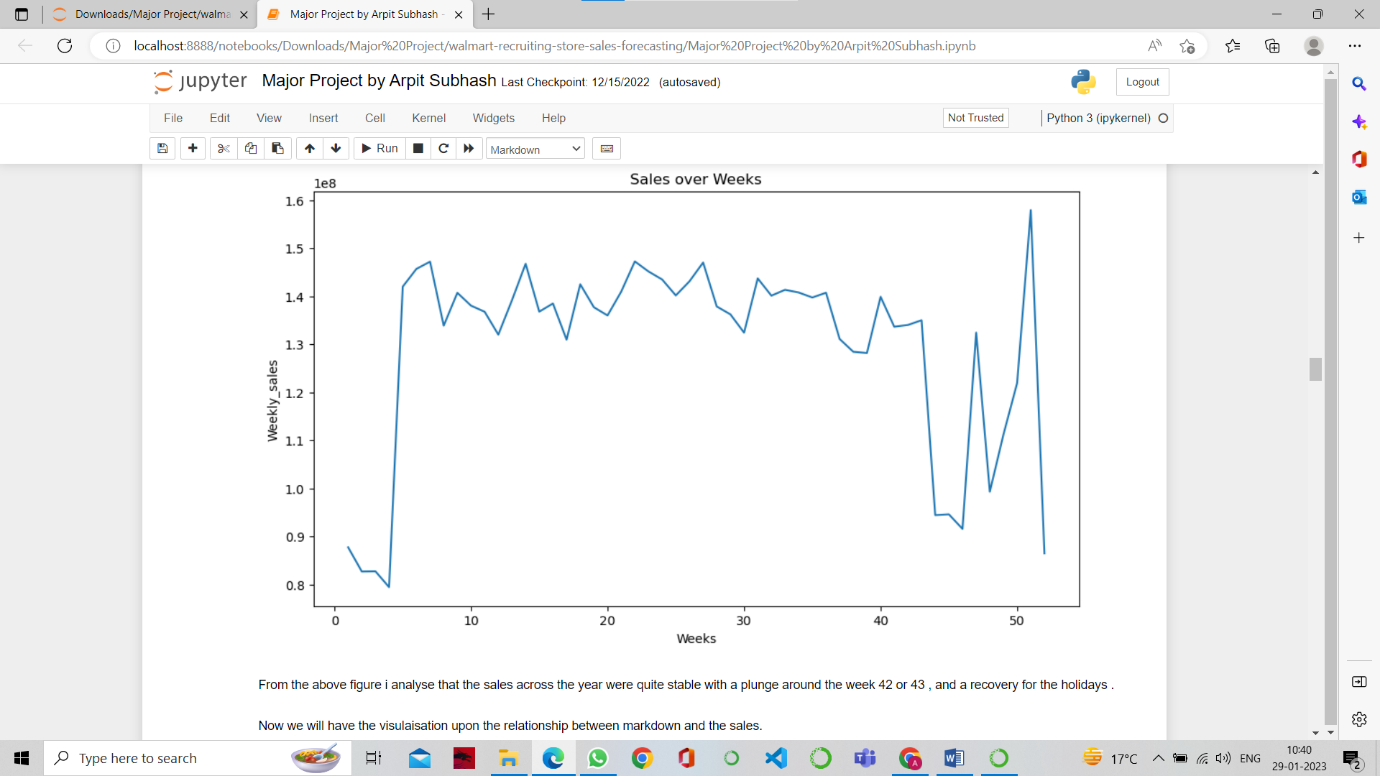


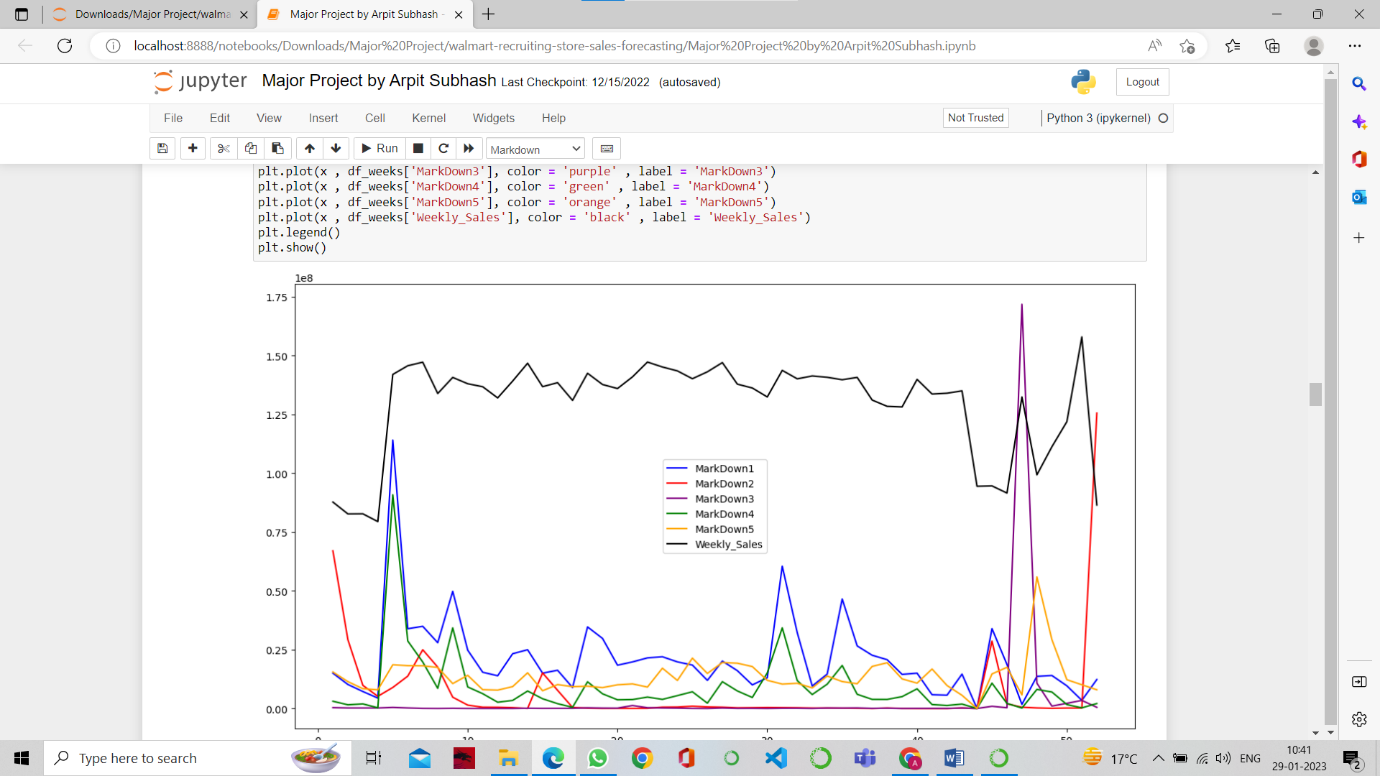
1. EDA – First of all I converted the date columns to date time, then I added some basic features related to all types of days that is, data, time, week, month, year and I merged four data sets to form two new data sets that is, train with features and stores and test with features and stores.

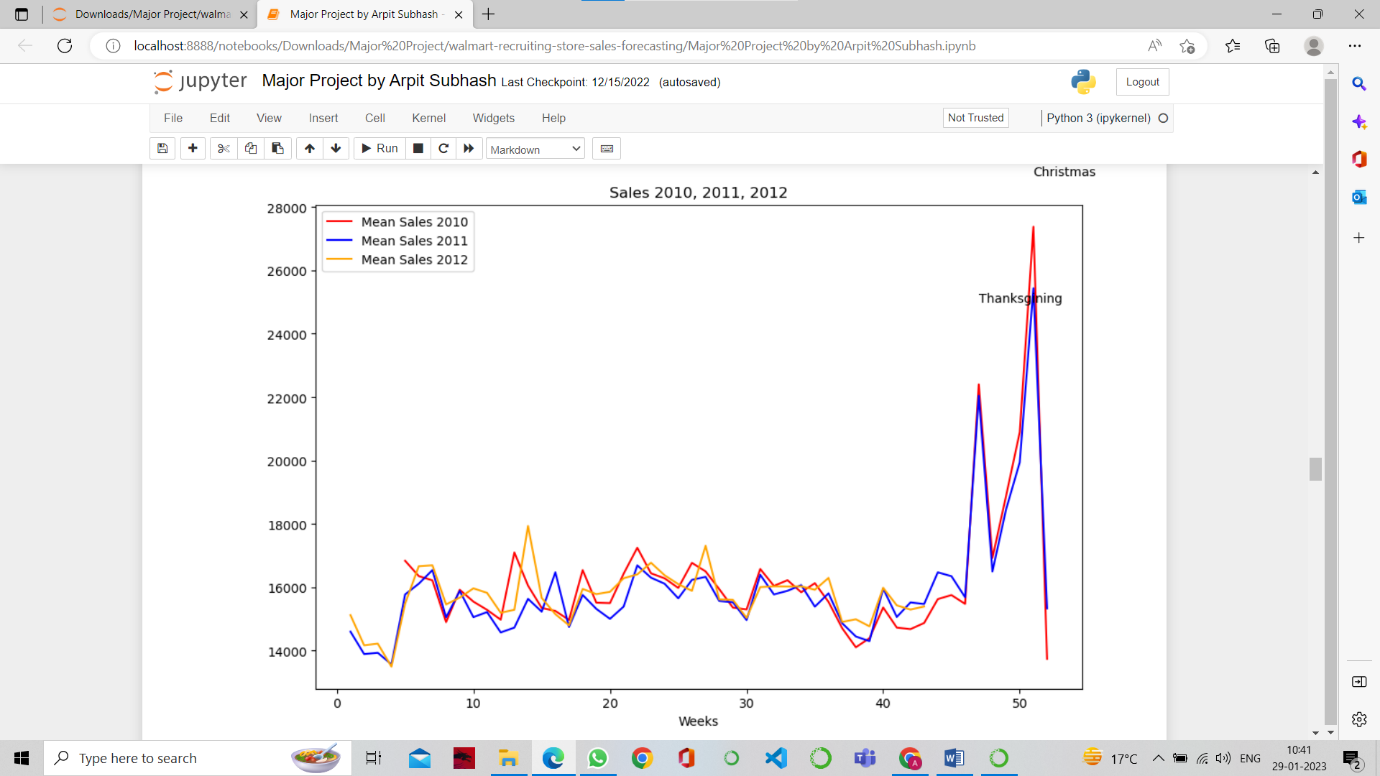


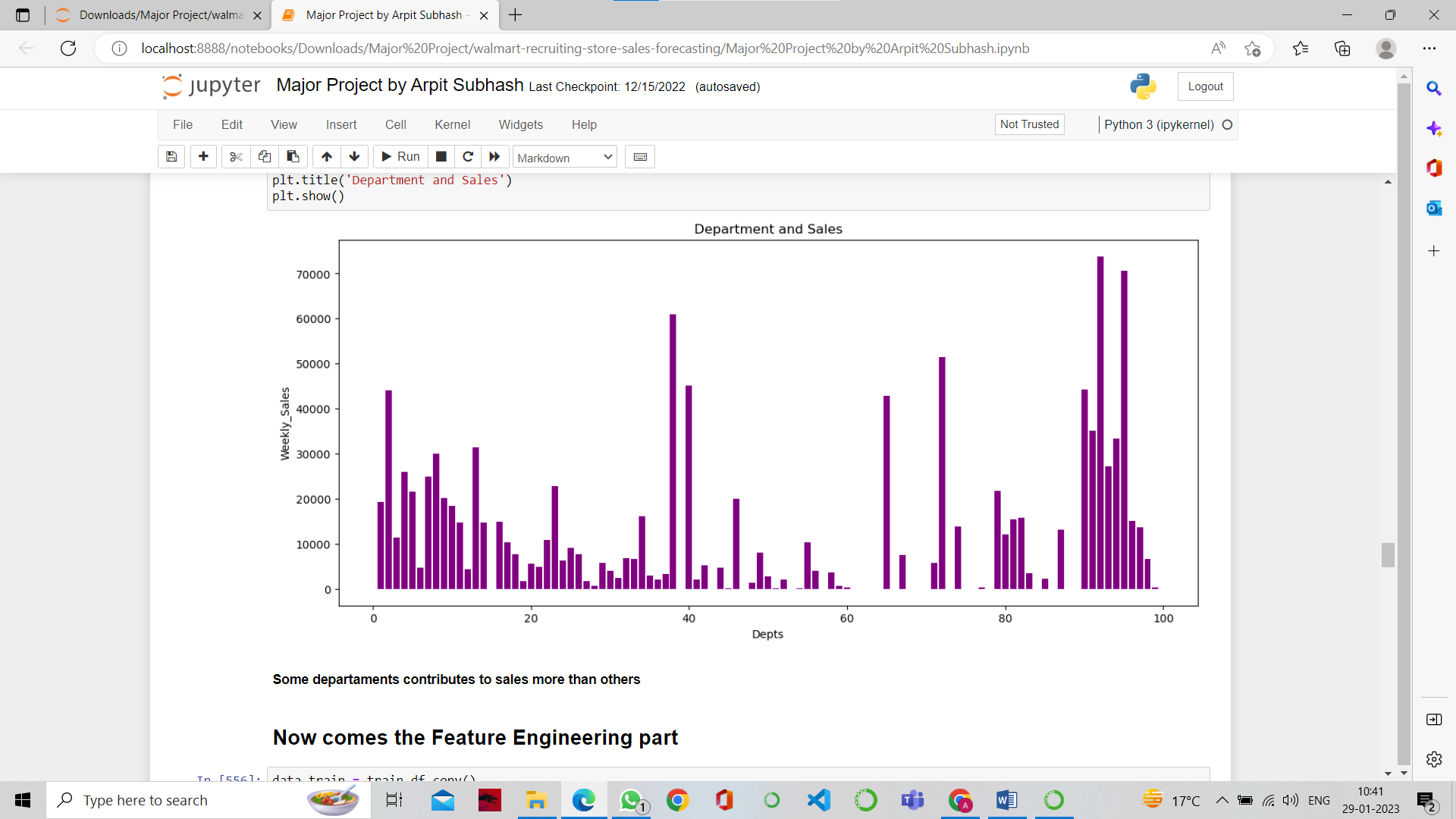


1. Data Visualisation – sales over weeks, sales vs all the markdowns, sales over weeks of year 2010, 2011, 2012 and relationship between store size and sales and relationship between department and sales.

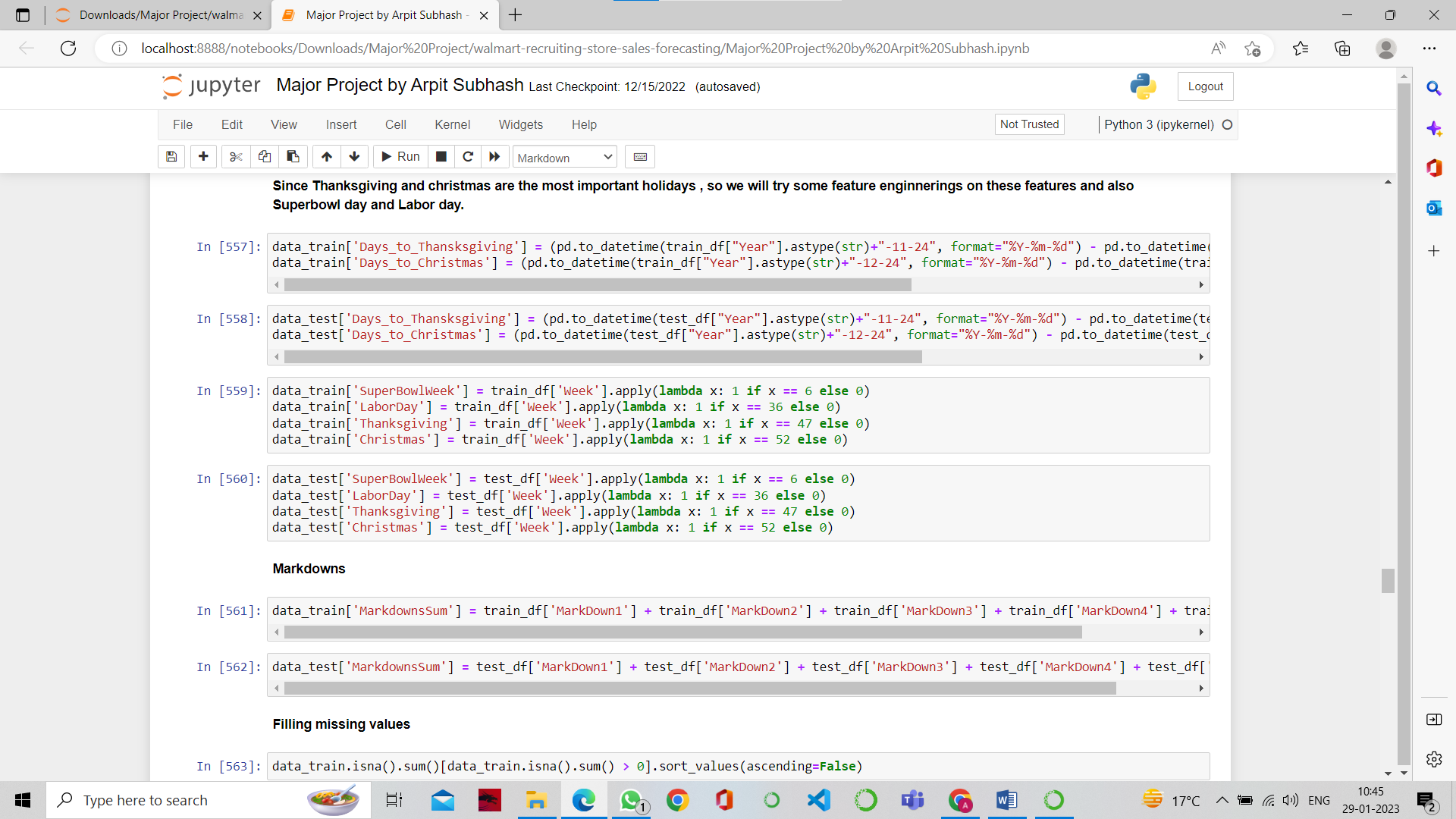




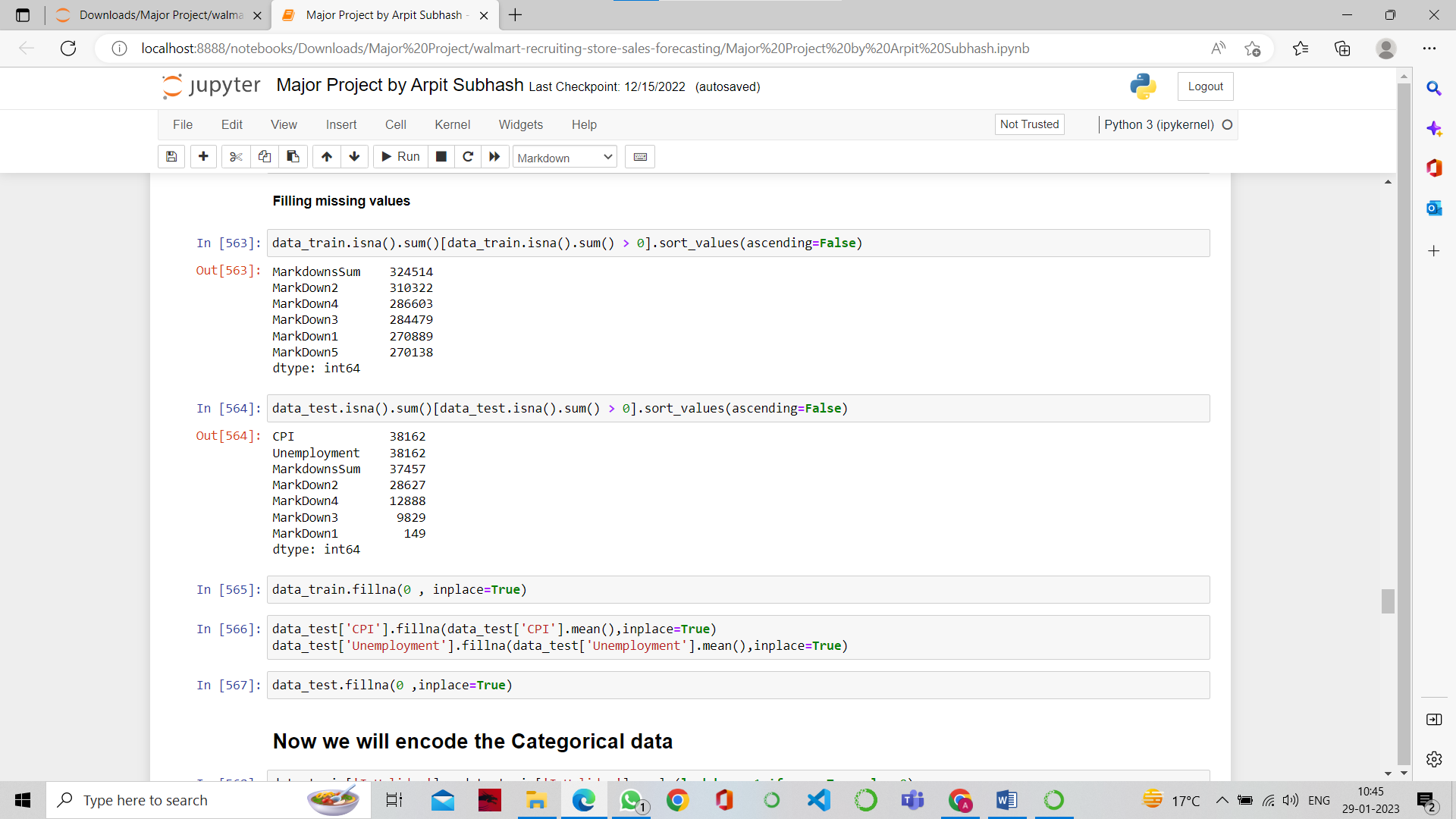




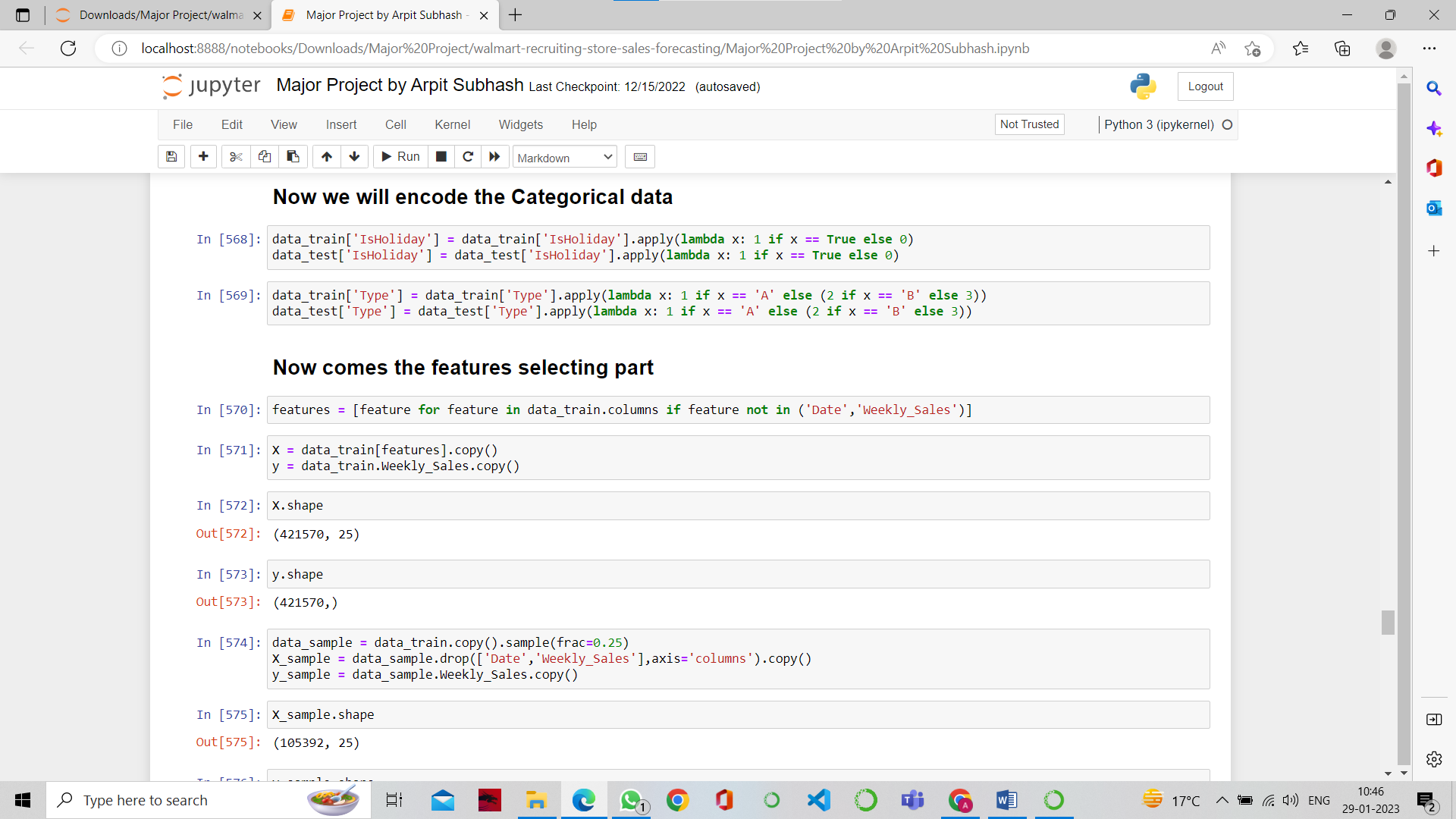
1. Feature Engineering – Since Thanksgiving and Christmas were the most important holidays, I tried some basic feature engineering on the days like Super Bowl day and Labor day. Also I added all the markdown to form two parts data train and data test.



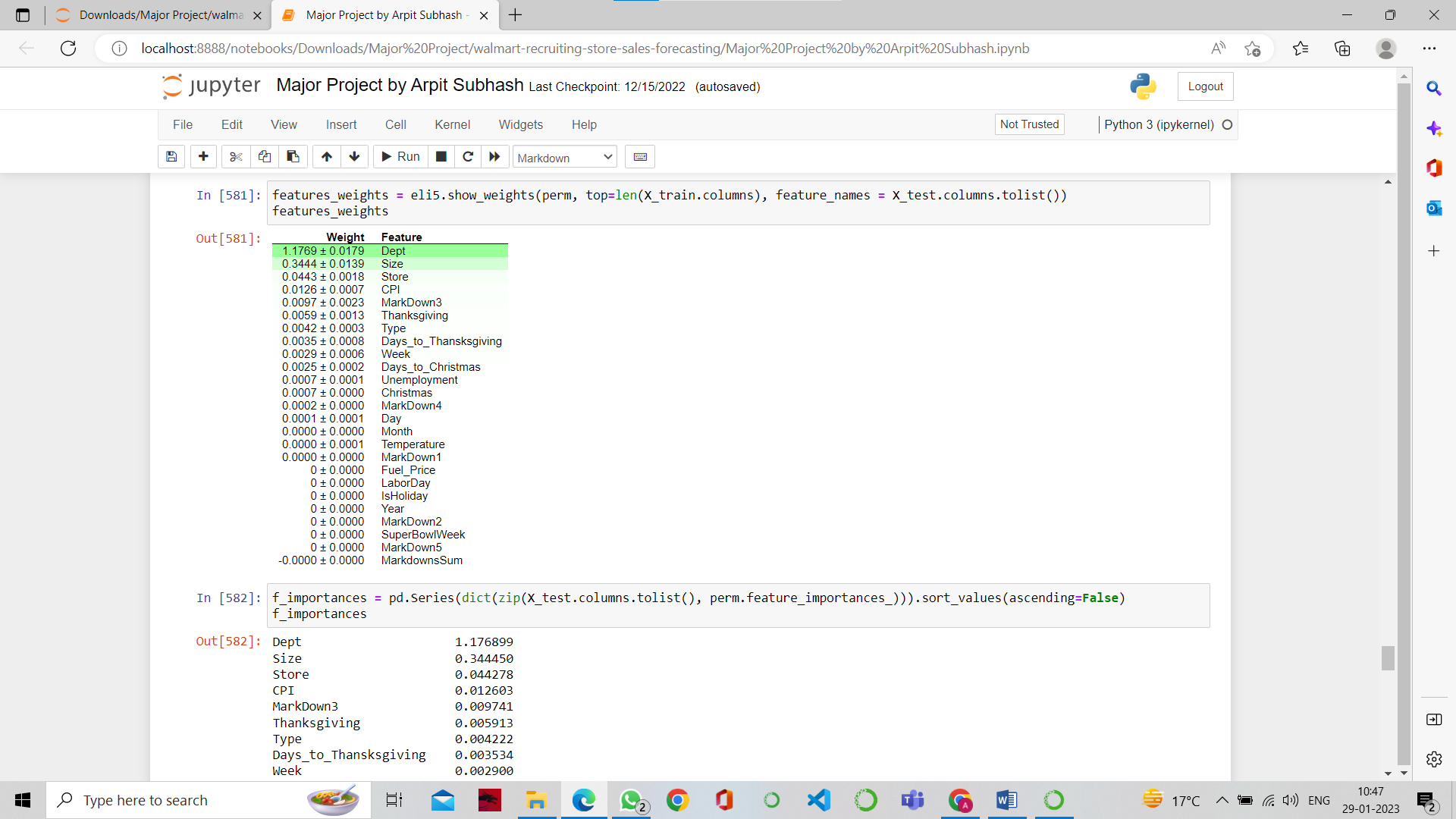
1. Preprocessing – Since all the Markdowns were having null values so I filled all the null values.



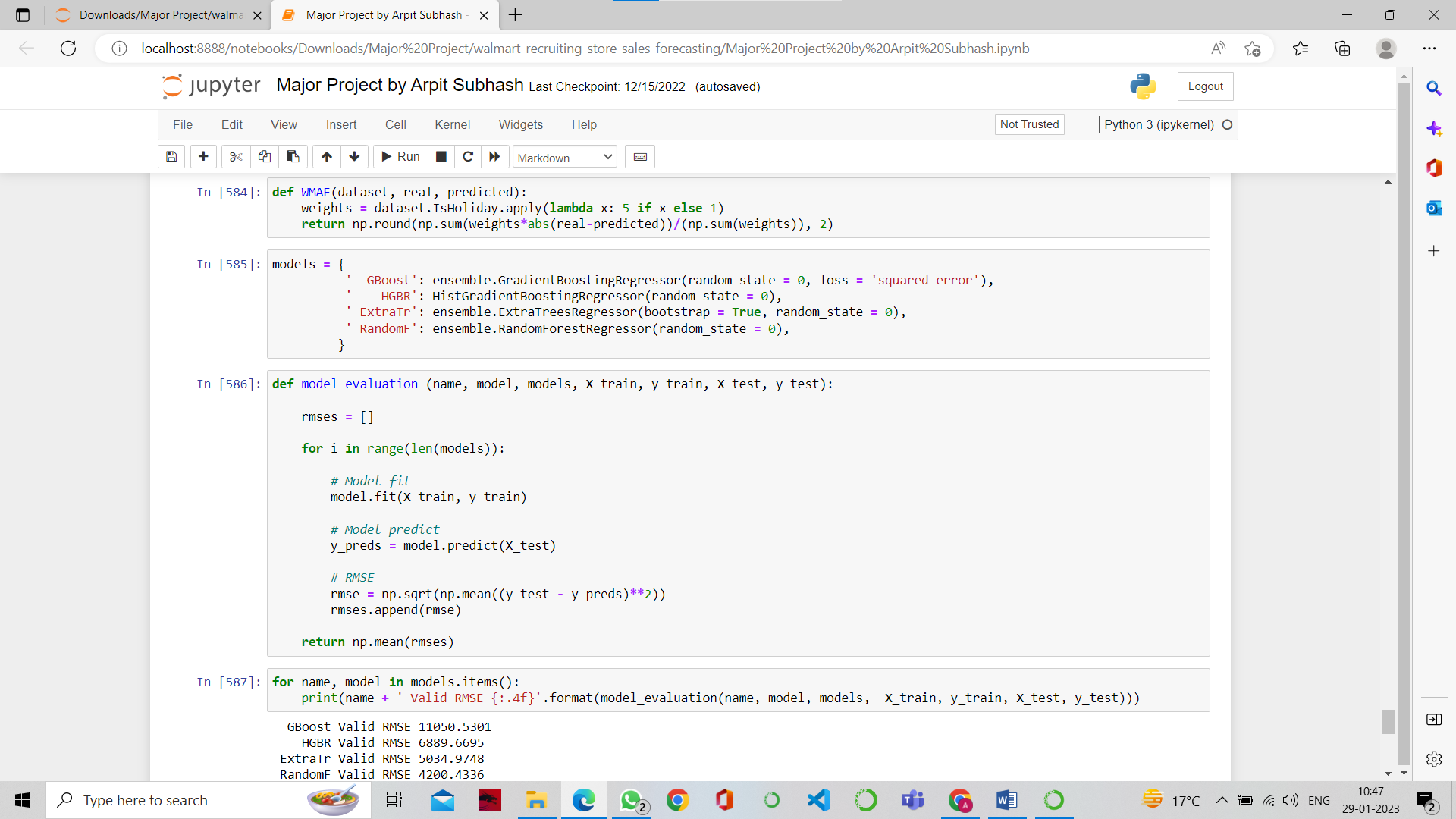
1. Encoding the categorical data – Like if there is holiday then it is 1 otherwise 0 and some more.



1. Feature Selection - First I selected the top 5 features that were Dept, Size, Store, CPI, Markdown3 by using one library eli5.



1. Models – Then I used the models like GBoost, HGBR, ExtraTrees, Random Forest and it comes out that the Random Forest was giving most accurate value or the least error.



1. Baseline model – At last I created a baseline model with Random Forest and Extra Trees Regressor.

