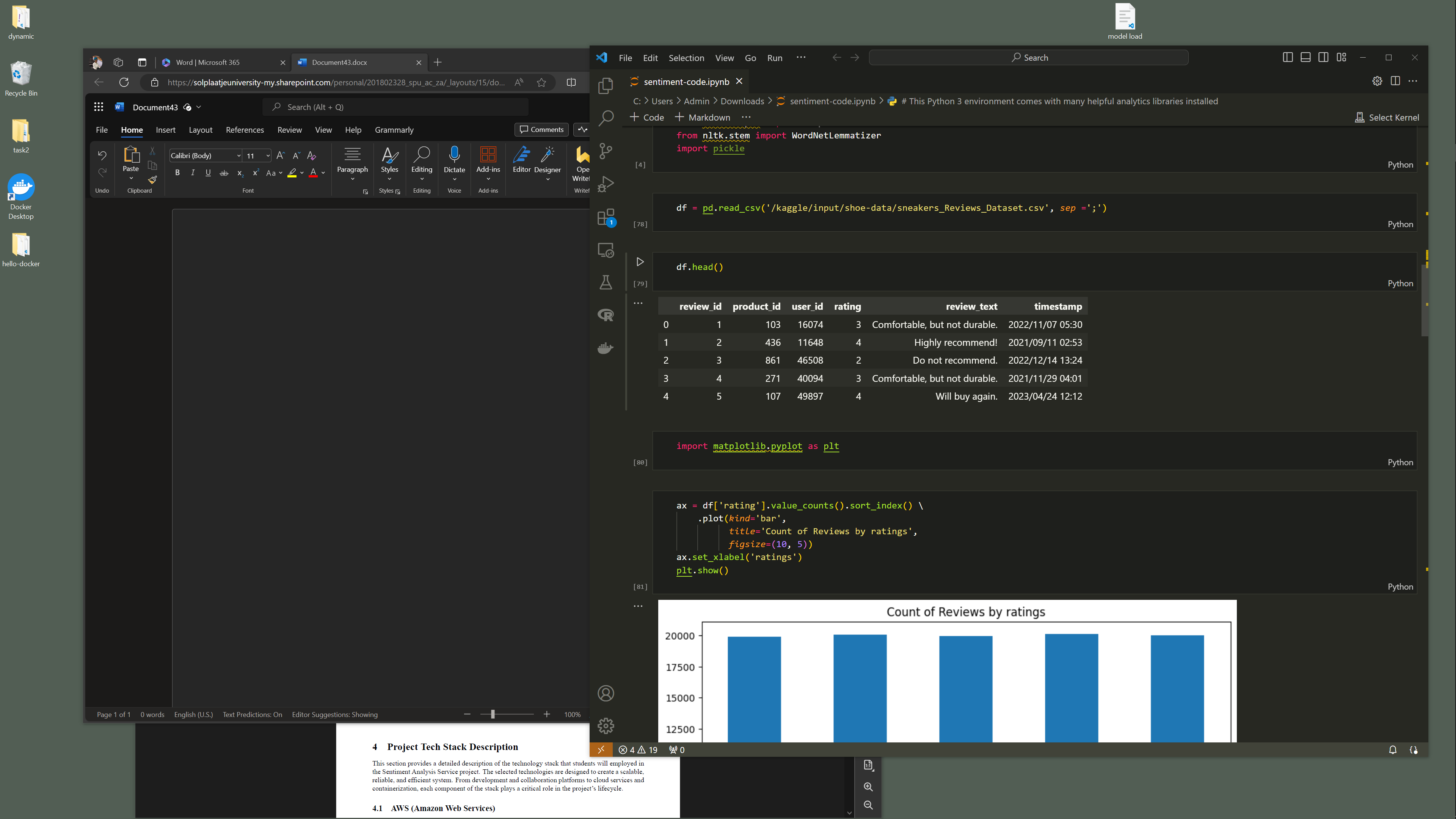
Hanna sebolai

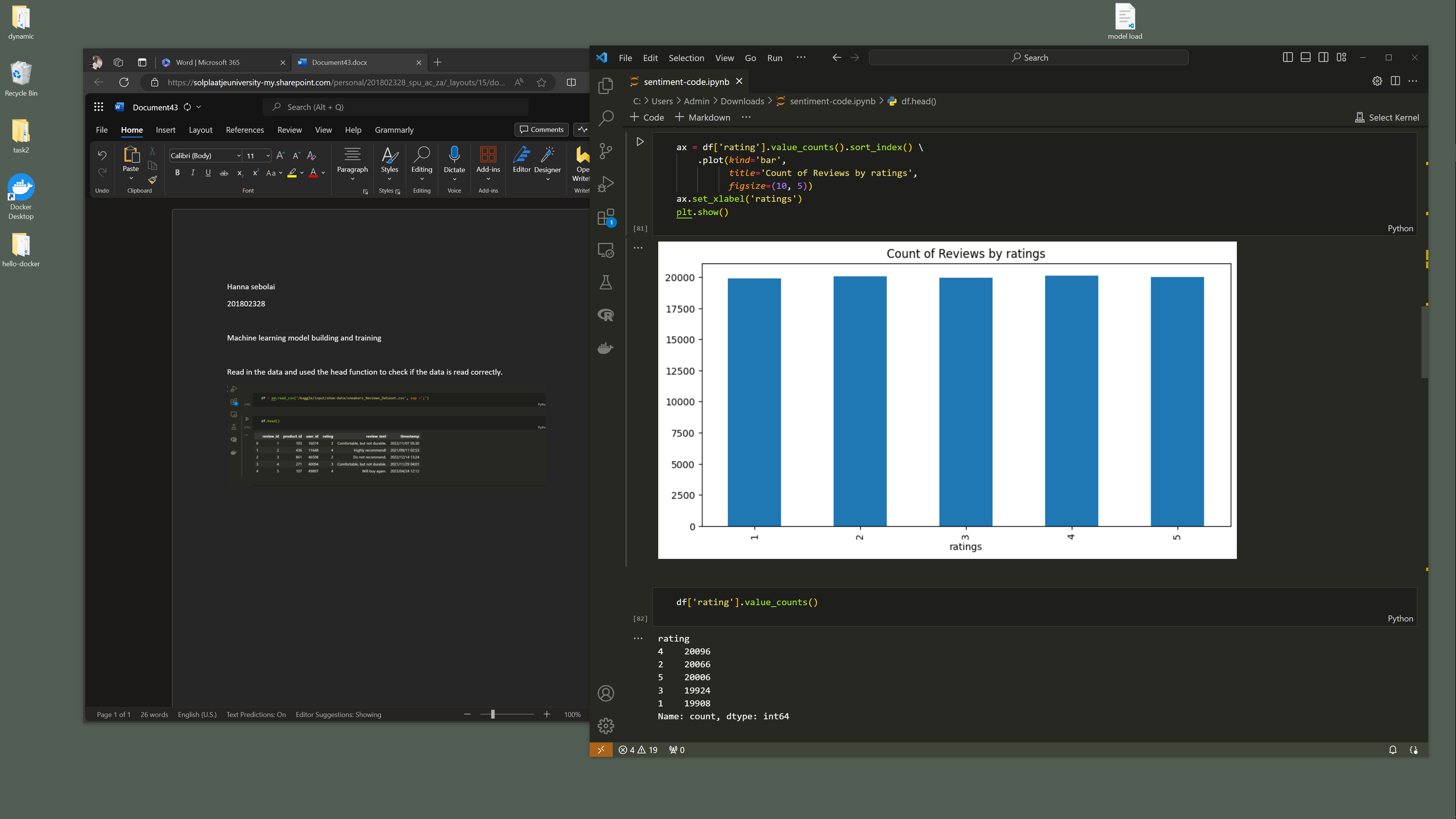
201802328

Machine learning model building and training

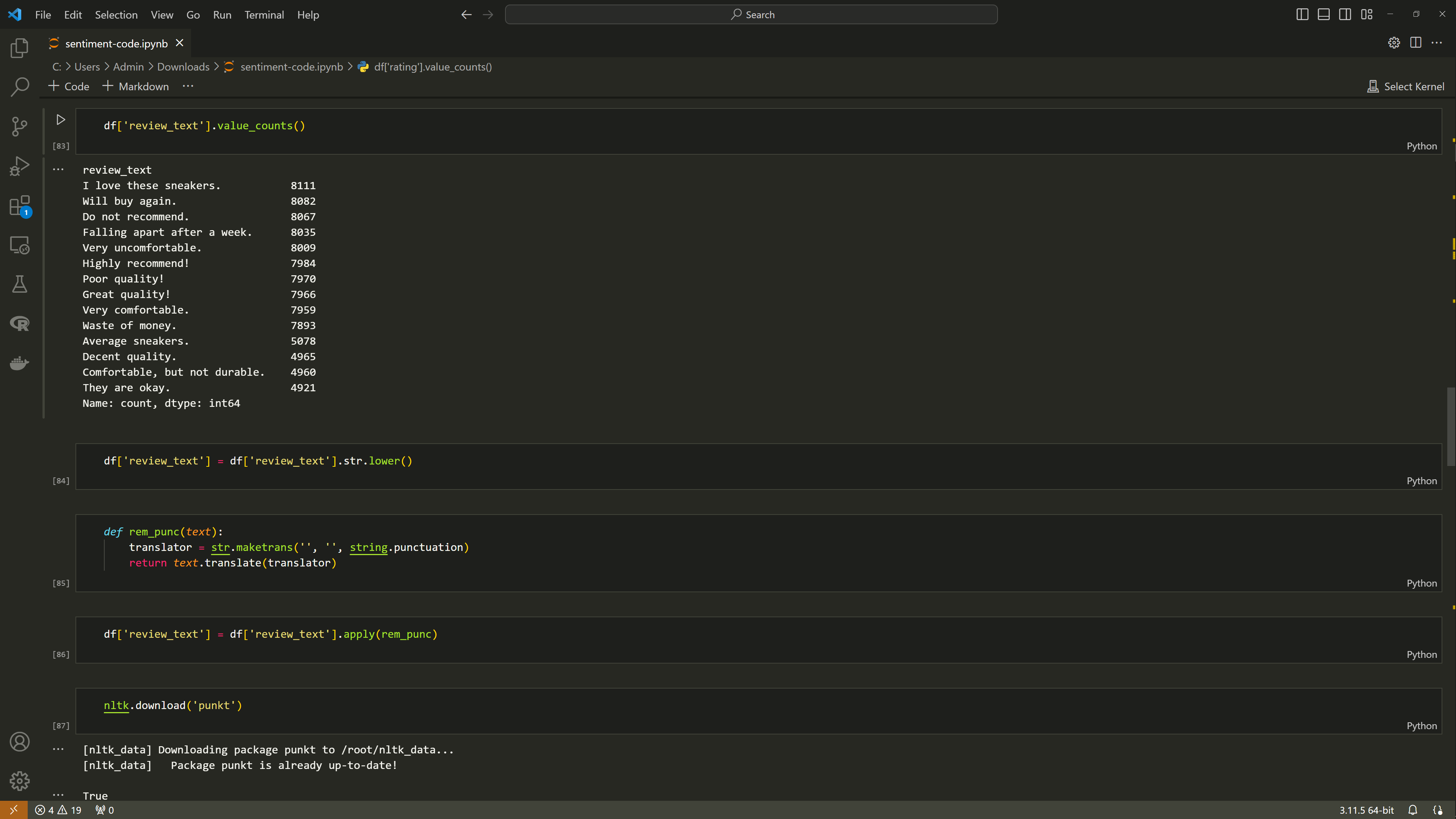
Read in the data and used the head function to check if the data is read correctly.



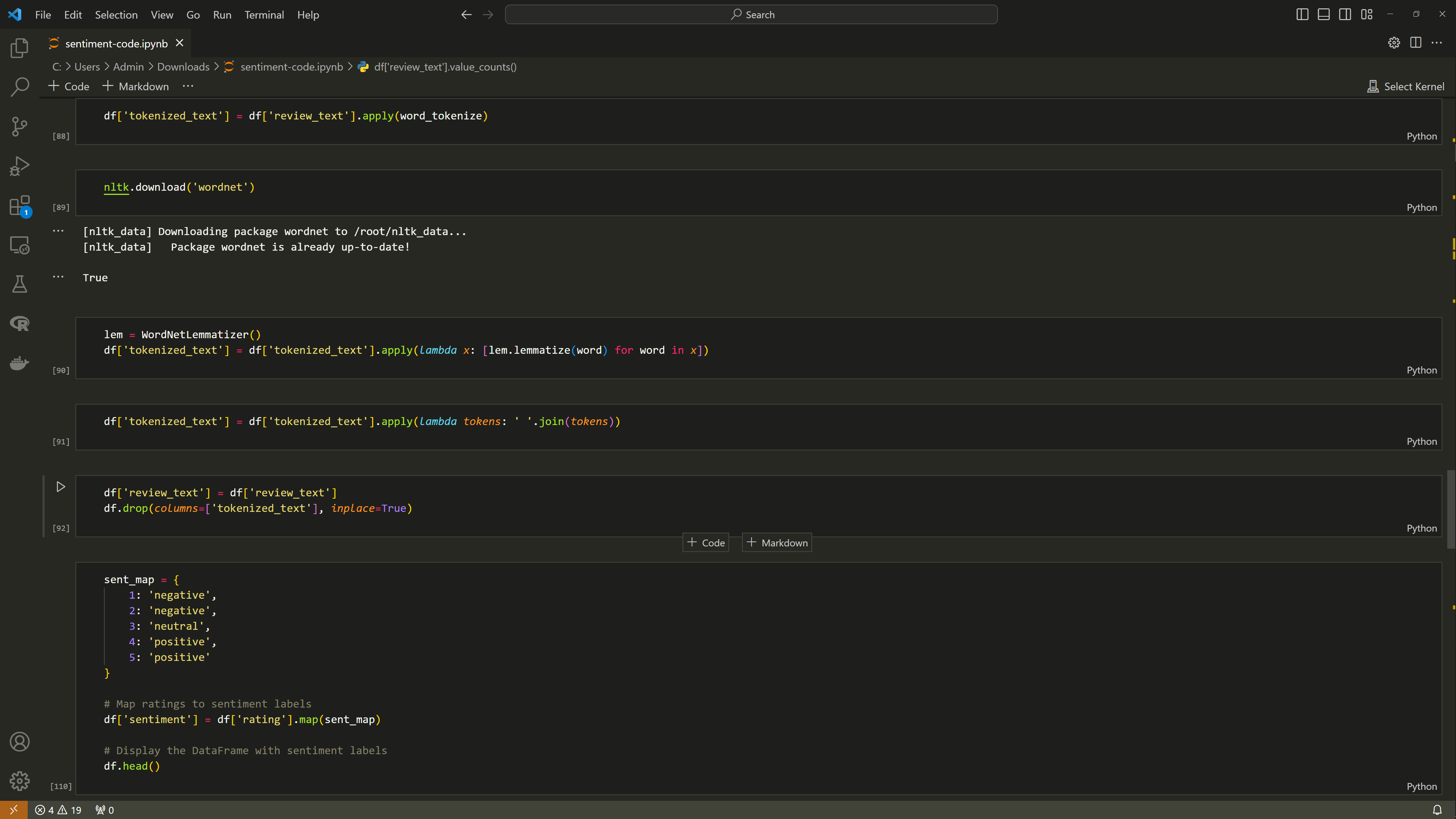
The code below is an output of the counts for all the ratings, we want to check if there is an even amount of ratings to avoid unbias in the data.



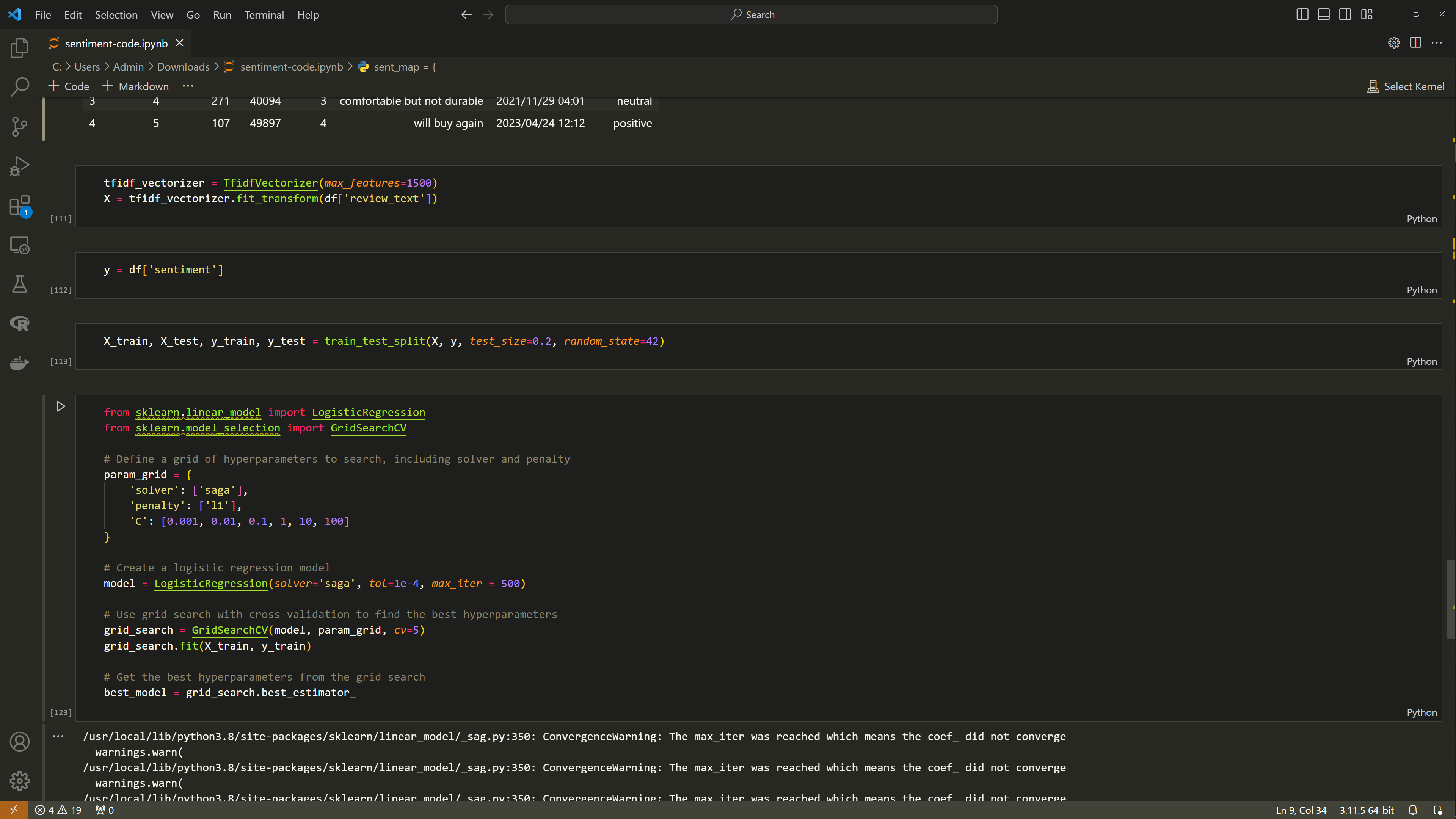
The code below shows the preprocessing of the column containing the comments of the customers. The data is preprocessed where it: will be converted to lower case, remove white spaces.



We will use a tokenizer from the nltk library to tokenize our text that was preprocessed and that will become the new text that will be used for modeling. A dictionary is created that assigns each rating a sentiment and then it is mapped to the ratings in the dataframe to create a new column sentiment. s



We move on the training where the x and y variables are defined. We will then split our data into training and testing using a 80/20 split. The model selected is logistic regression and we used GridsearchCv to try and improve the model by obtaining better results . The model is then trained and given captured under best\_model.



Even though we tried to obtain optimal parameters the model still overfitted because based on the metrics used for evaluation. Even when reducing the data by 50% and using the regularizer the model was still overfitting.

* F1\_score = 1
* Precision =1
* Recall = 1
* Accuracy = 1

