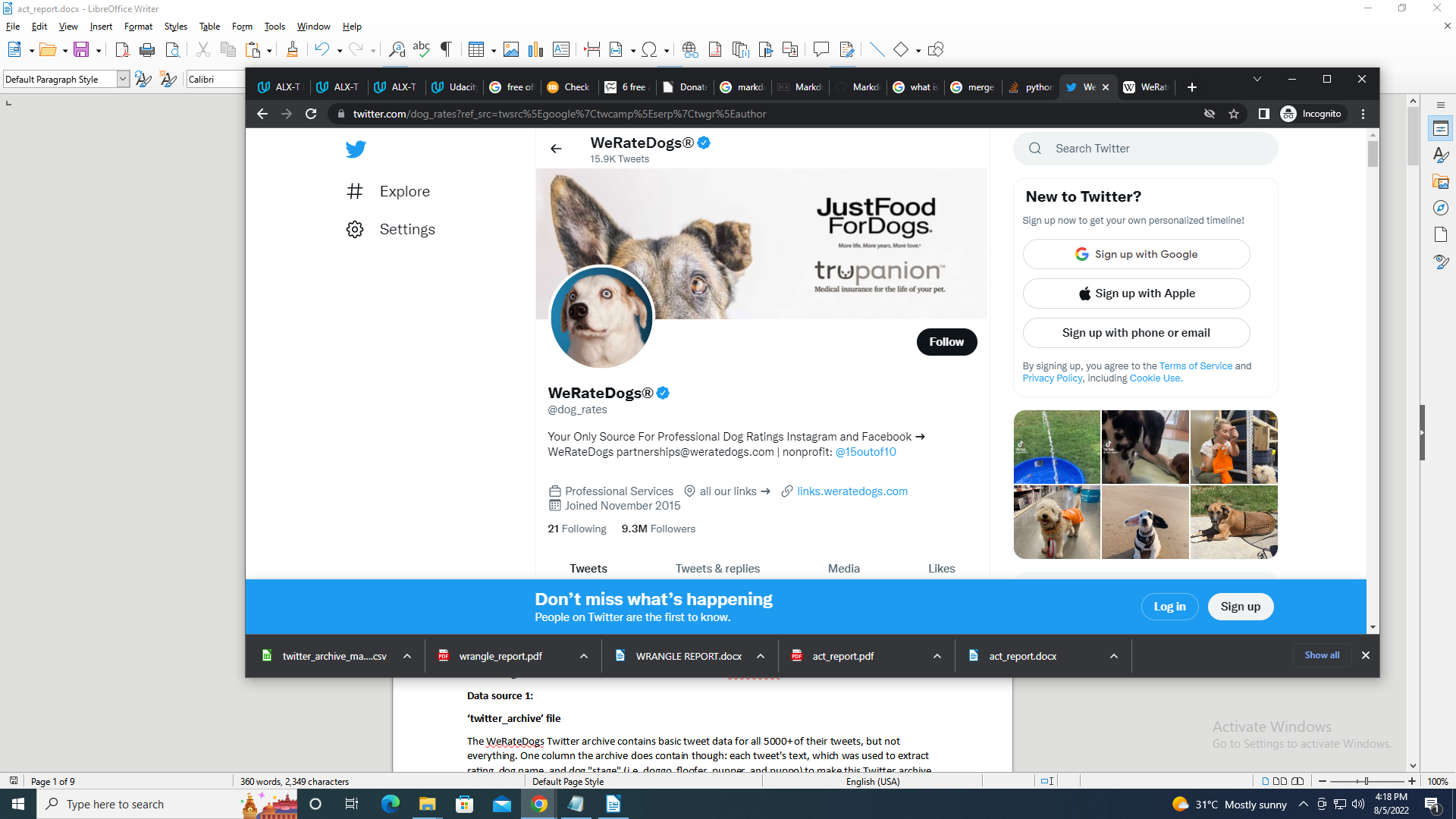
**‘WeRateDogs’**



**INTRODUCTION**

‘WeRateDogs’ is a Twitter account which became popular for its smile-evoking comments and ratings of dog pictures

These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "[they're good dogs Brent](http://knowyourmeme.com/memes/theyre-good-dogs-brent)." WeRateDogs has over 9 million followers and has received international media coverage.

This project is focused on data wrangling and analysis. However, visualisations were made and insights were gathered

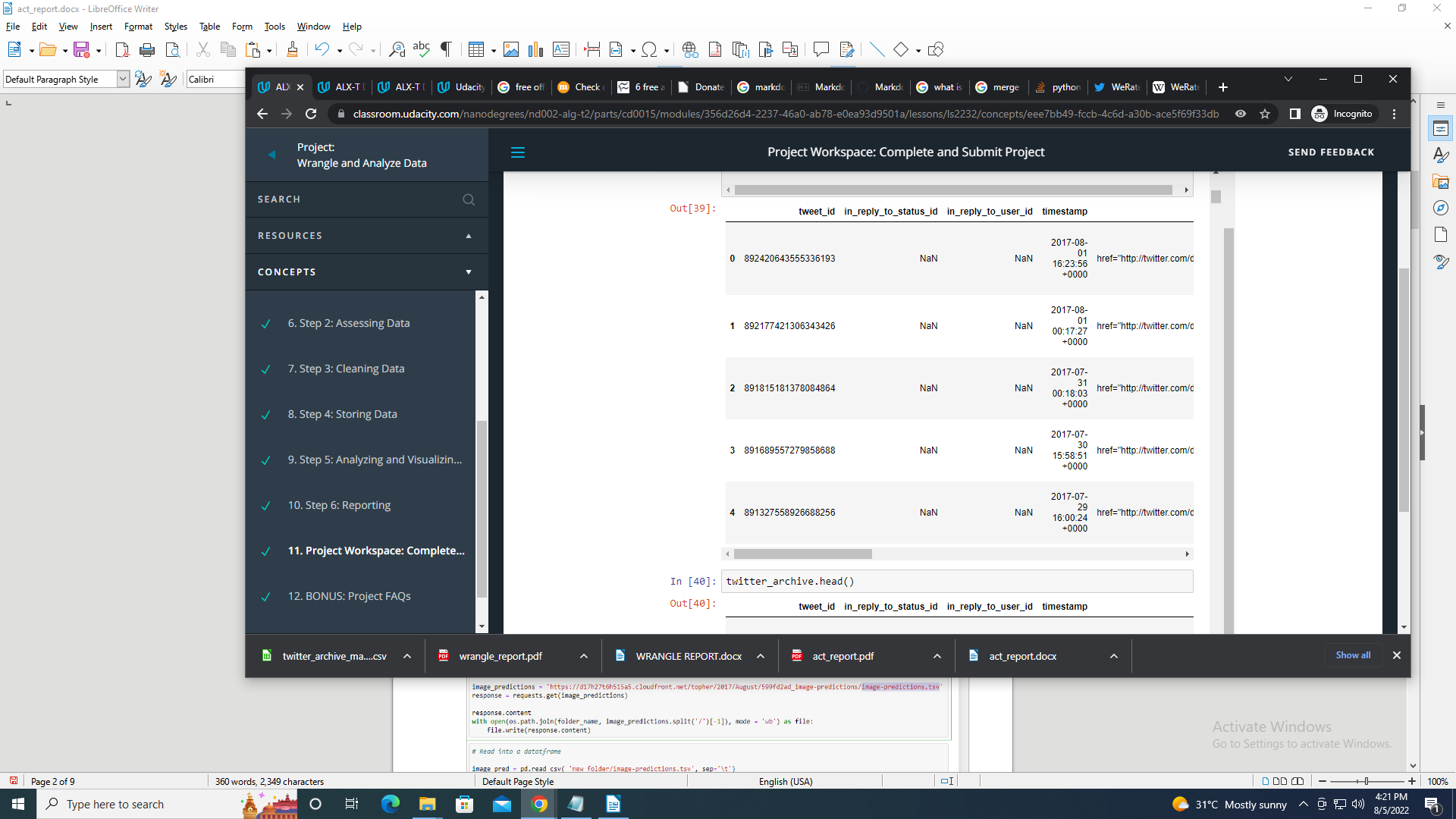
**DATA GATHERING**

Data was gathered from three data sources and read into dataframes

**Data source 1:**

**‘twitter\_archive’ file**

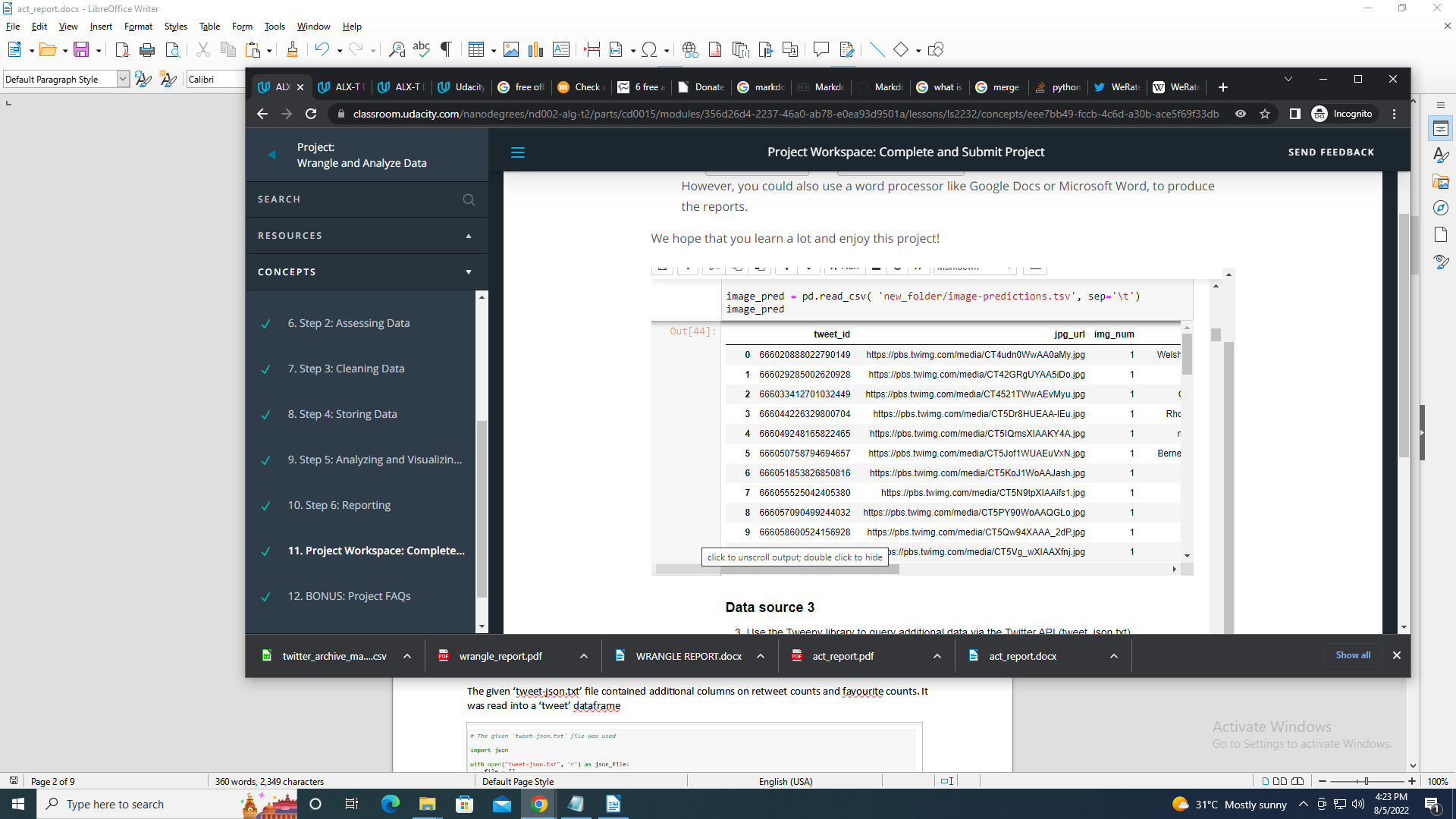
The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which was used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, they have been filtered for those tweets with ratings only (there are 2356). This is the ‘twitter-archive-enhanced’ file which is then read into the twitter\_archive dataframe



**Data source 2:**

**‘image-predictions.tsv’**

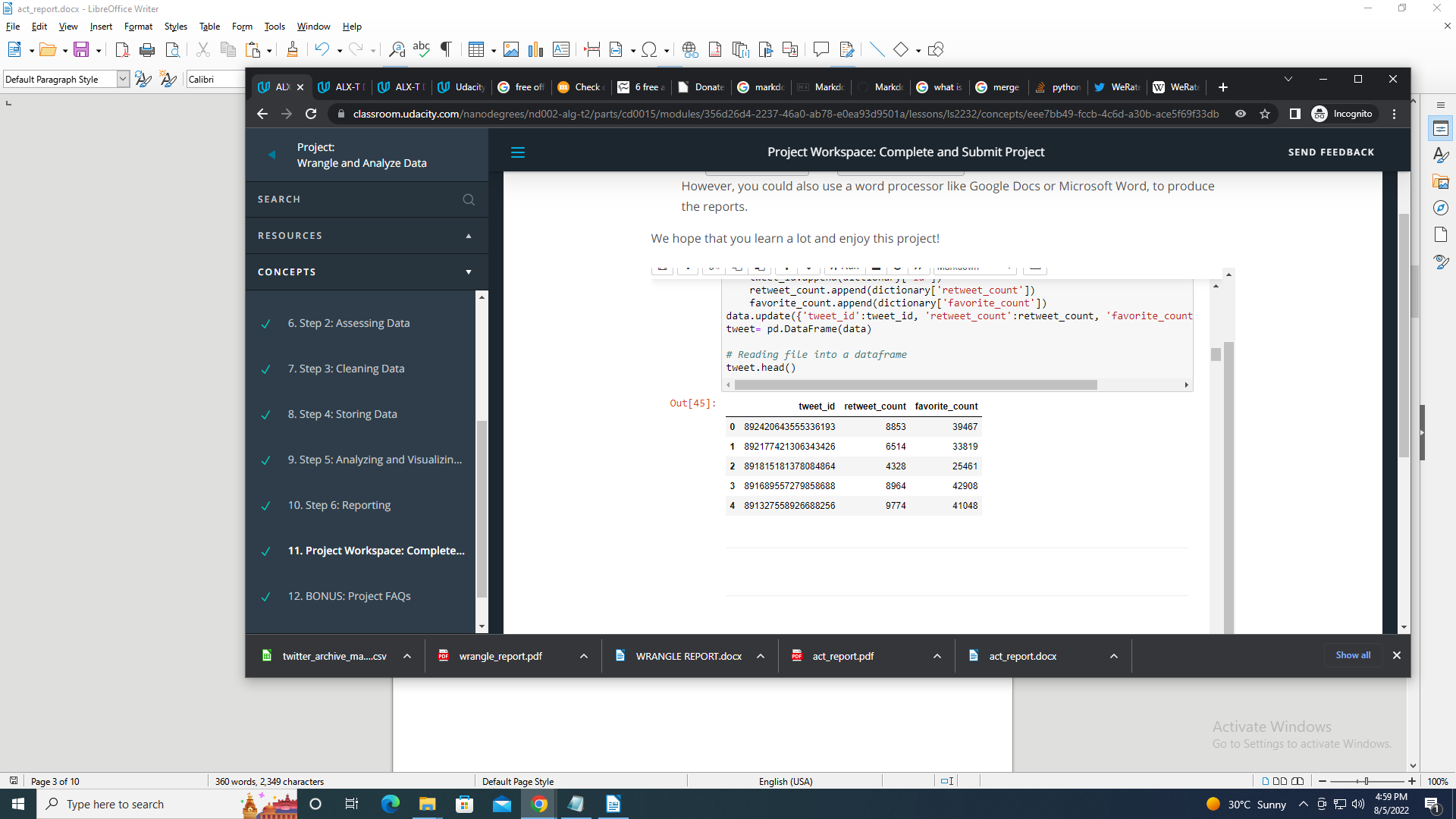
The ‘image\_predictions.tsv’ file contained data on image predictions of the dogs from the tweets. It was extracted from the given URL into a folder and then read into ‘image\_pred’ dataframe



**Data source 3:**

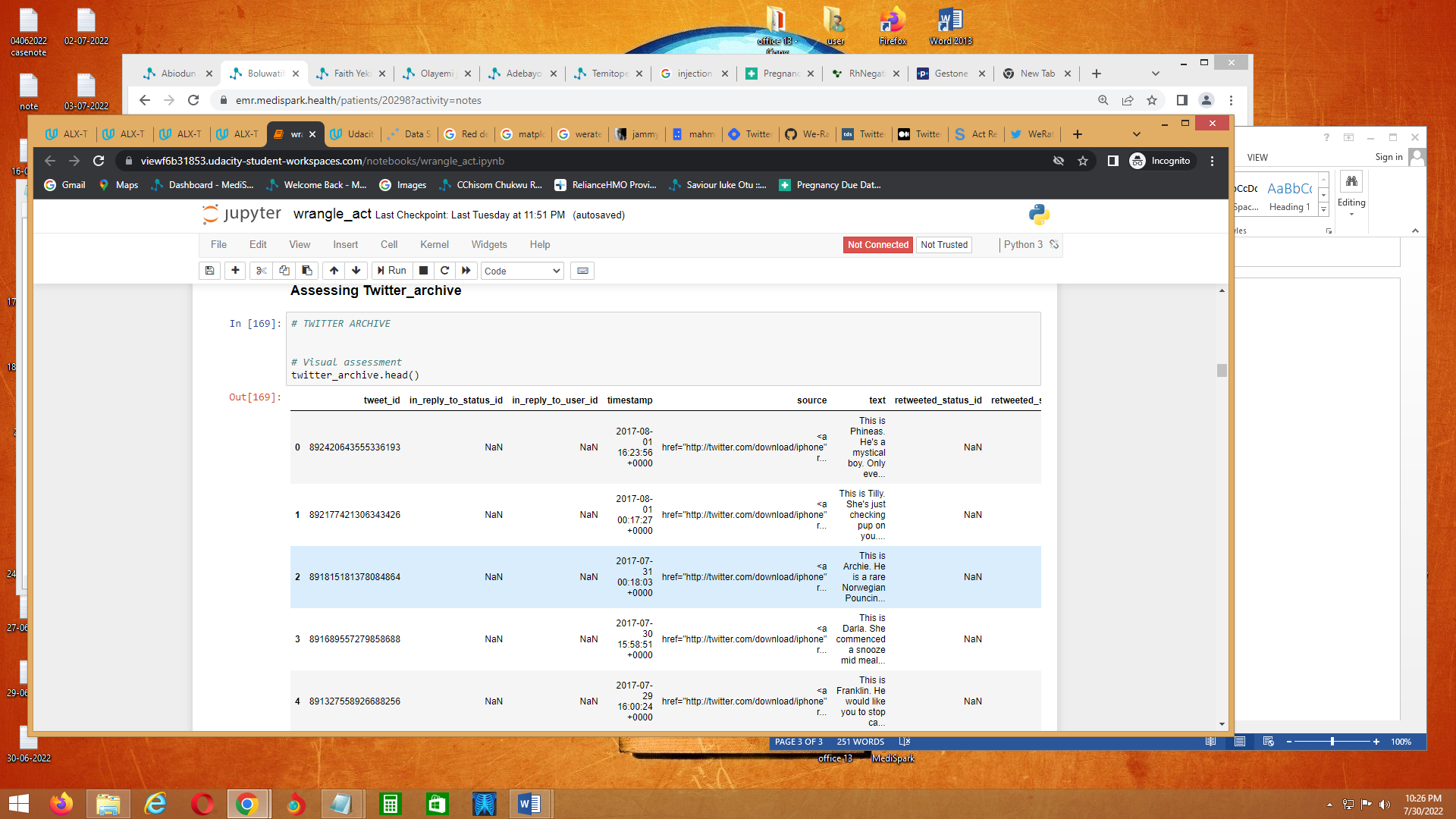
**‘tweet-json.txt’**

The given ‘tweet-json.txt’ file contained additional columns on retweet counts and favourite counts. It was read into a ‘tweet’ dataframe

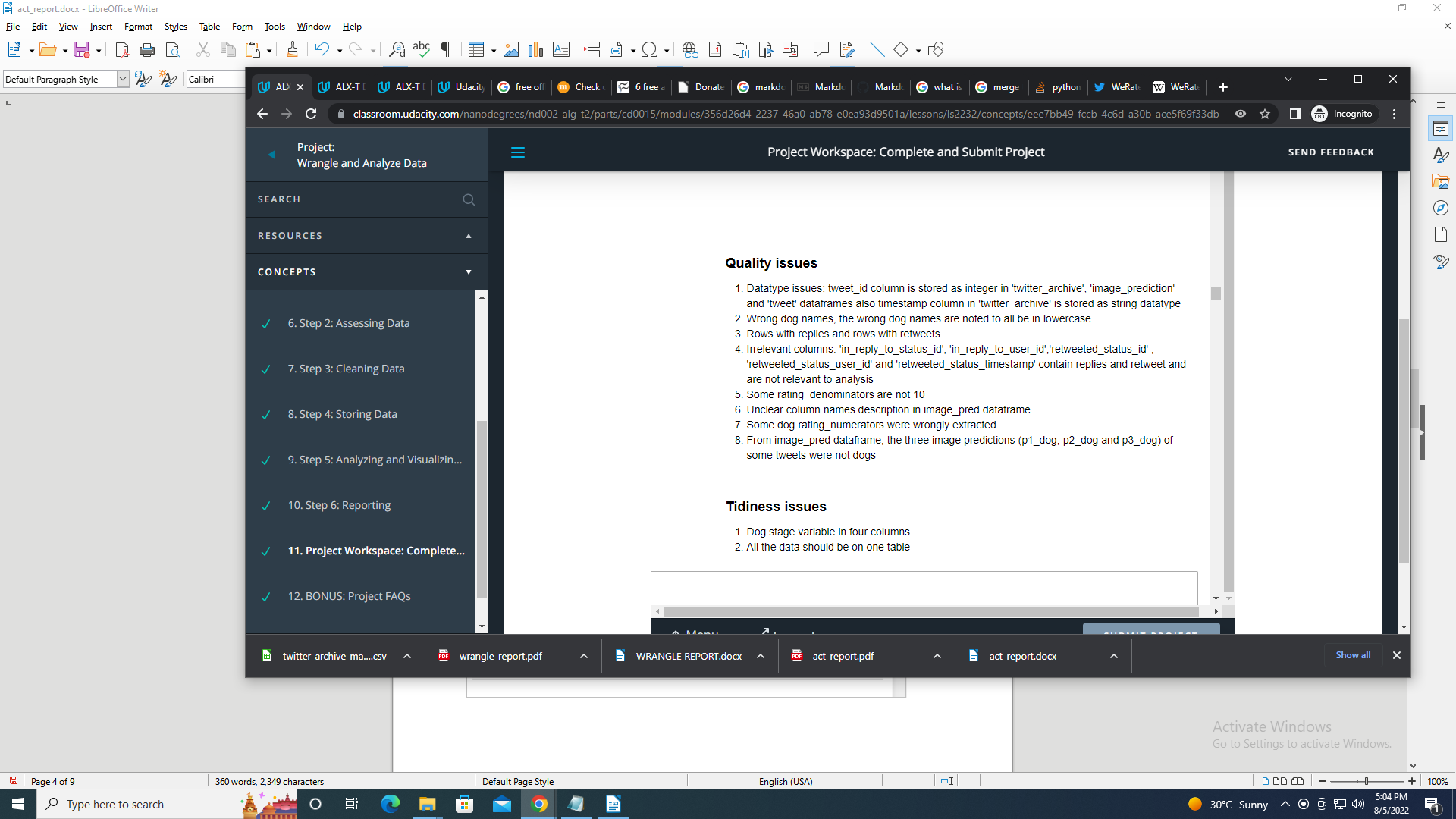


**ASSESSING DATA**

Visual and programmatic assessments were used to identify quality and tidiness issues



**Identified issues:**



**CLEANING DATA**

The identified issues were addressed during the cleaning process

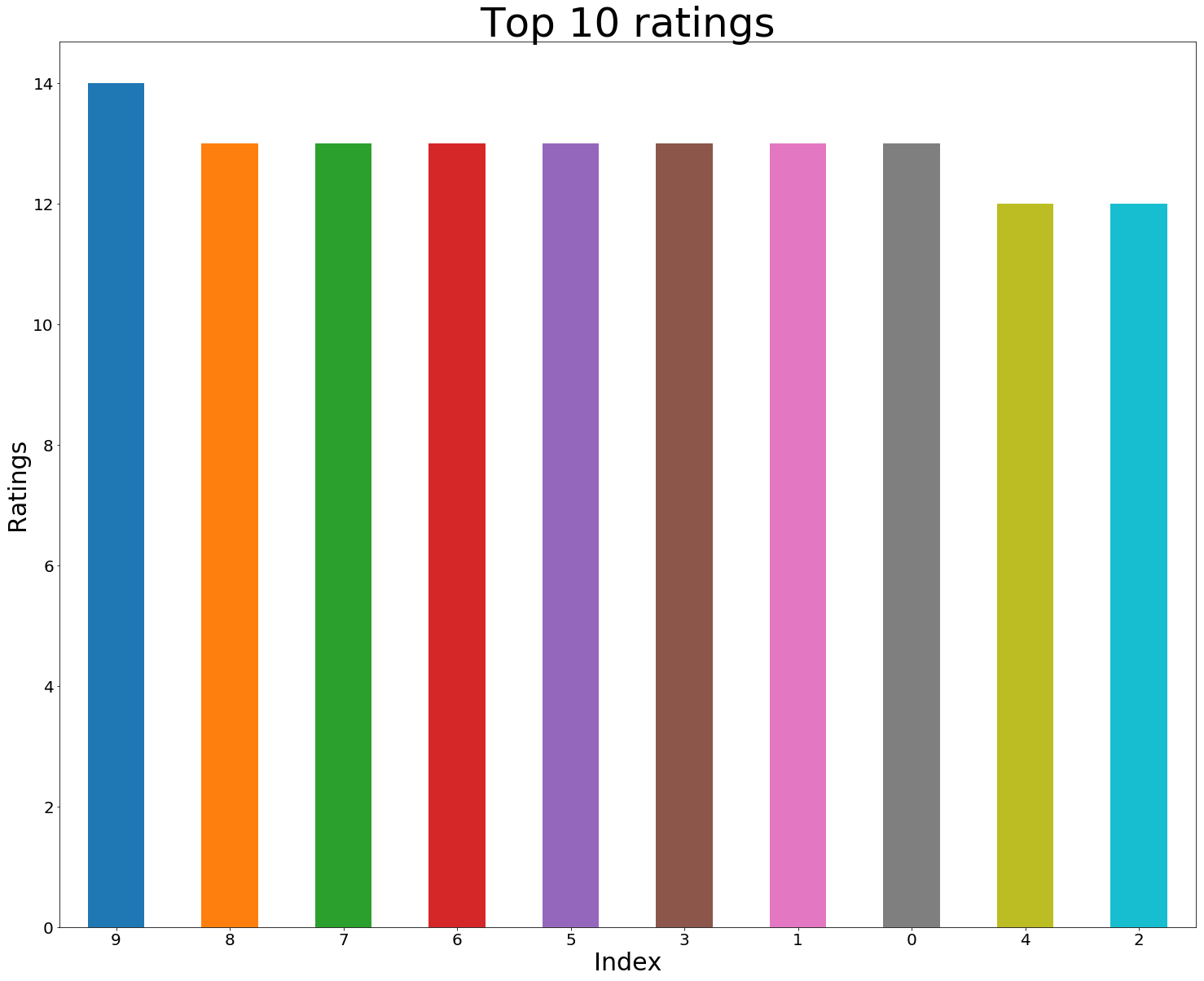
The Define-Code-Test cycle was used to clean each issue identified

**SAVING DATA**

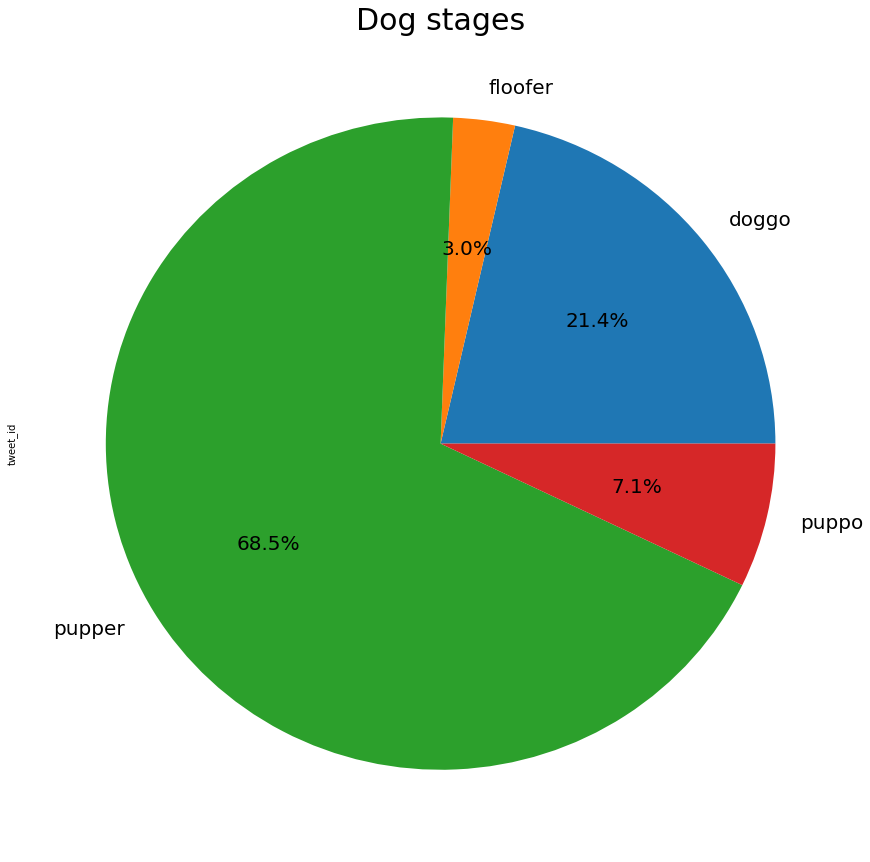
The master dataset was saved into a csv file named ‘twitter\_archive\_master.csv’

**ANALYSIS AND VISUALISATIONS**

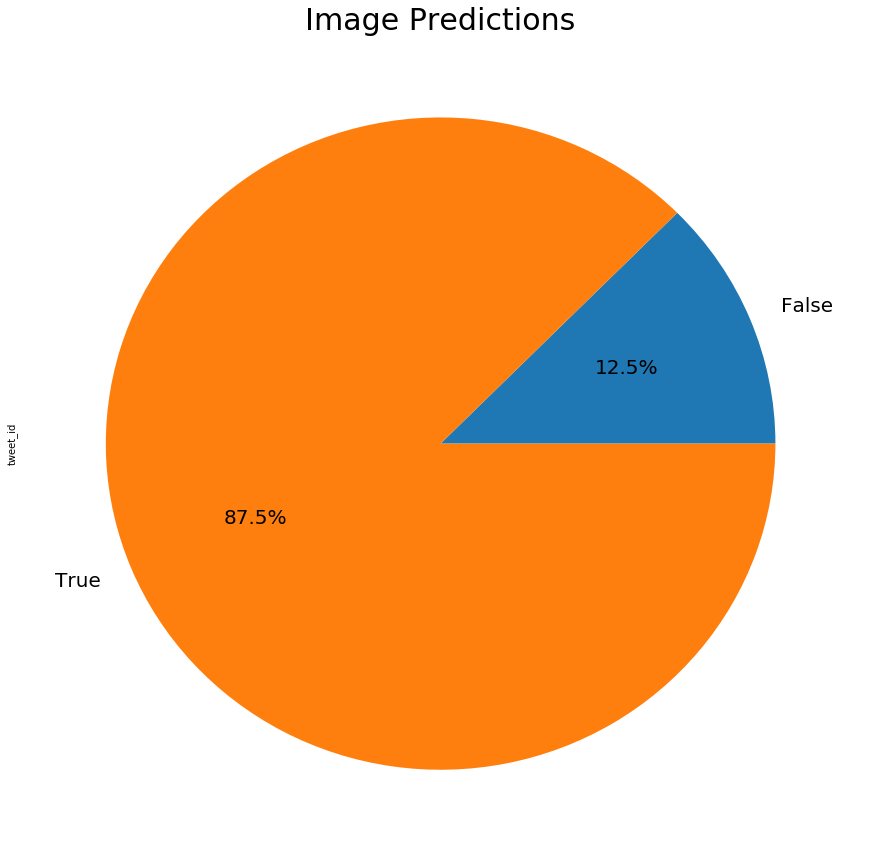
The analysis and visualisations made are as below:



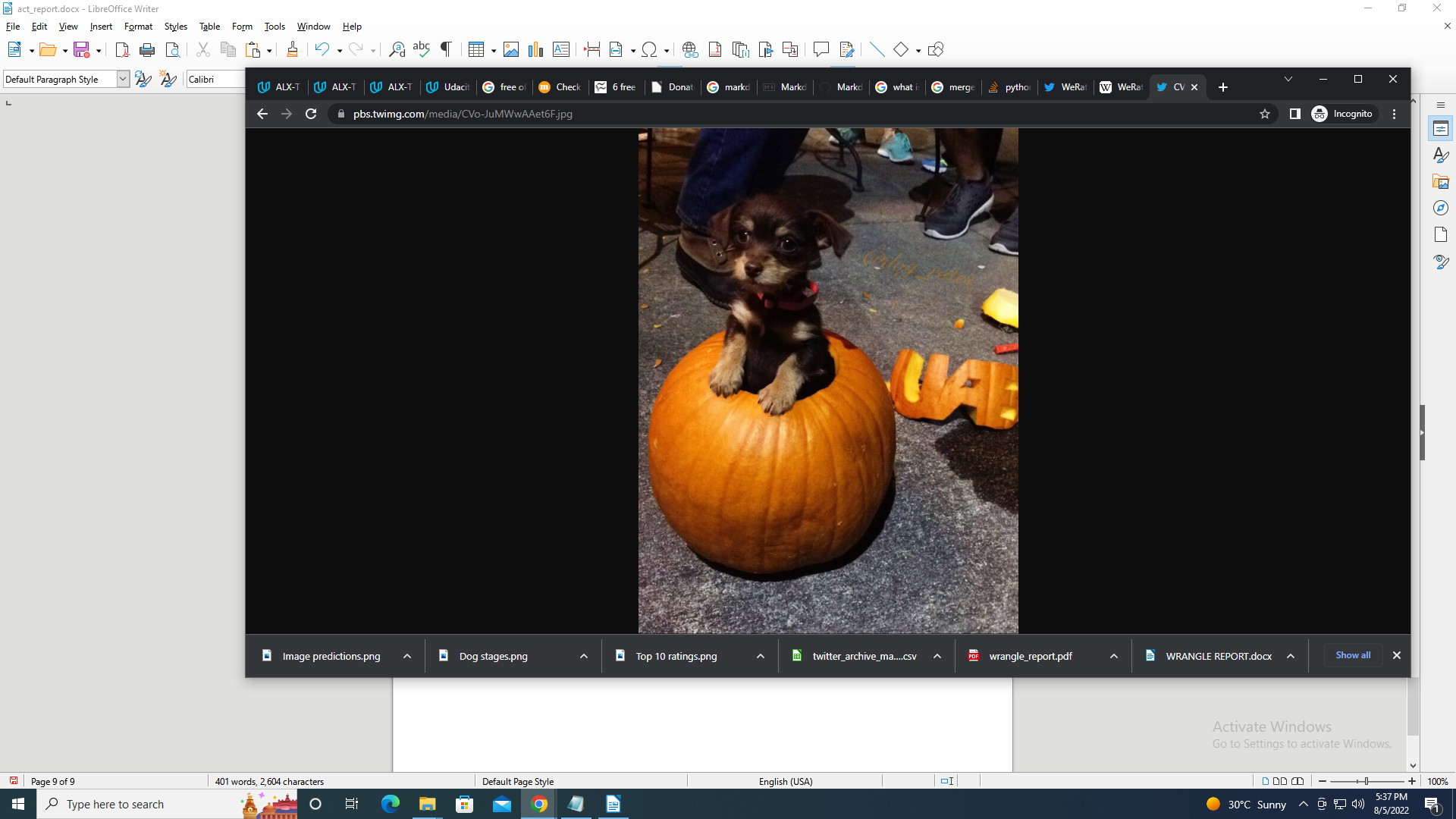
* This showed that the highest rating is 14/10



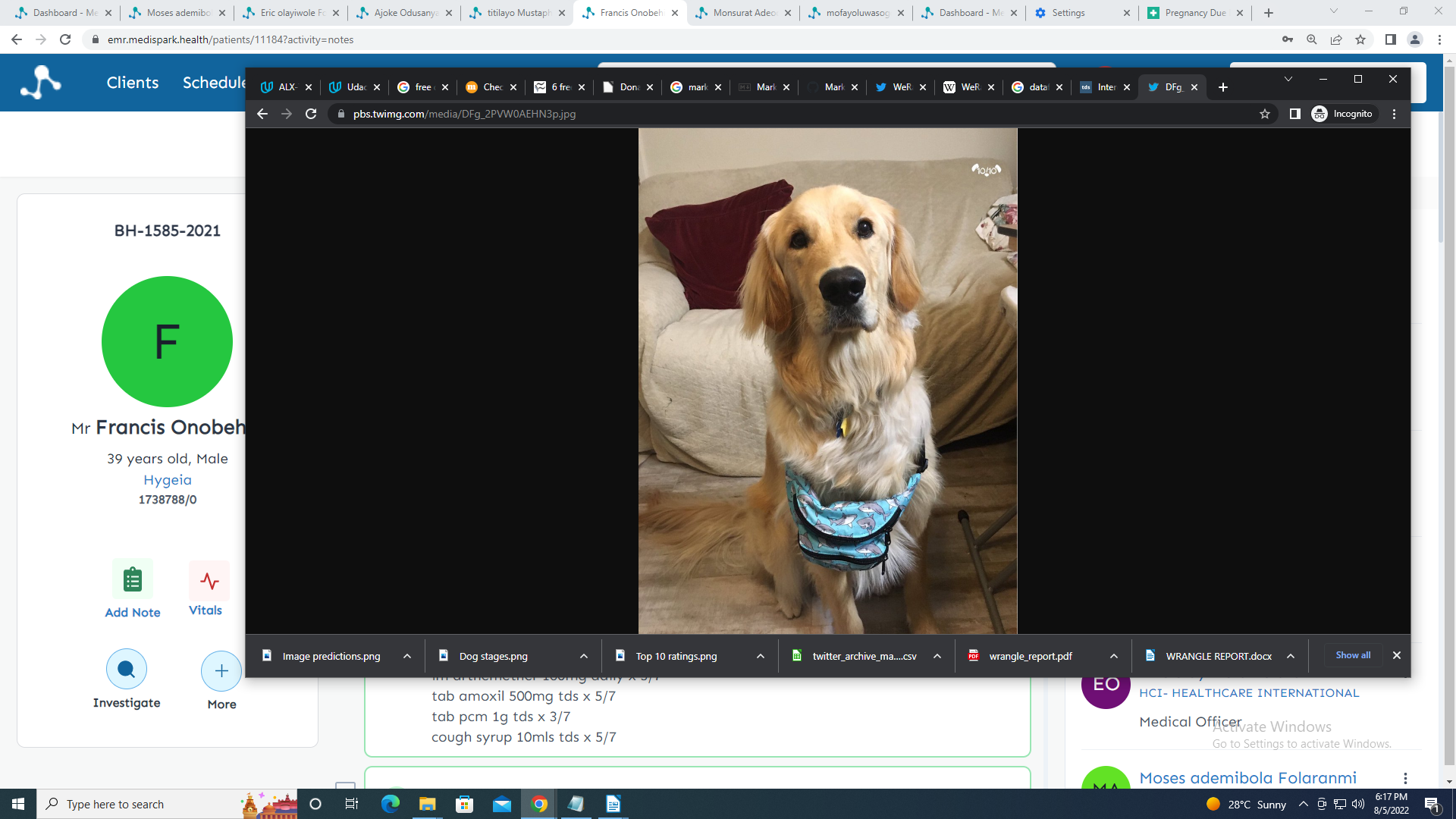
* Of the tweets with dog stages, the pupper stage constituted about two-thirds



* Over three-fourths of the first image predictions by the algorithm turned out to be dogs

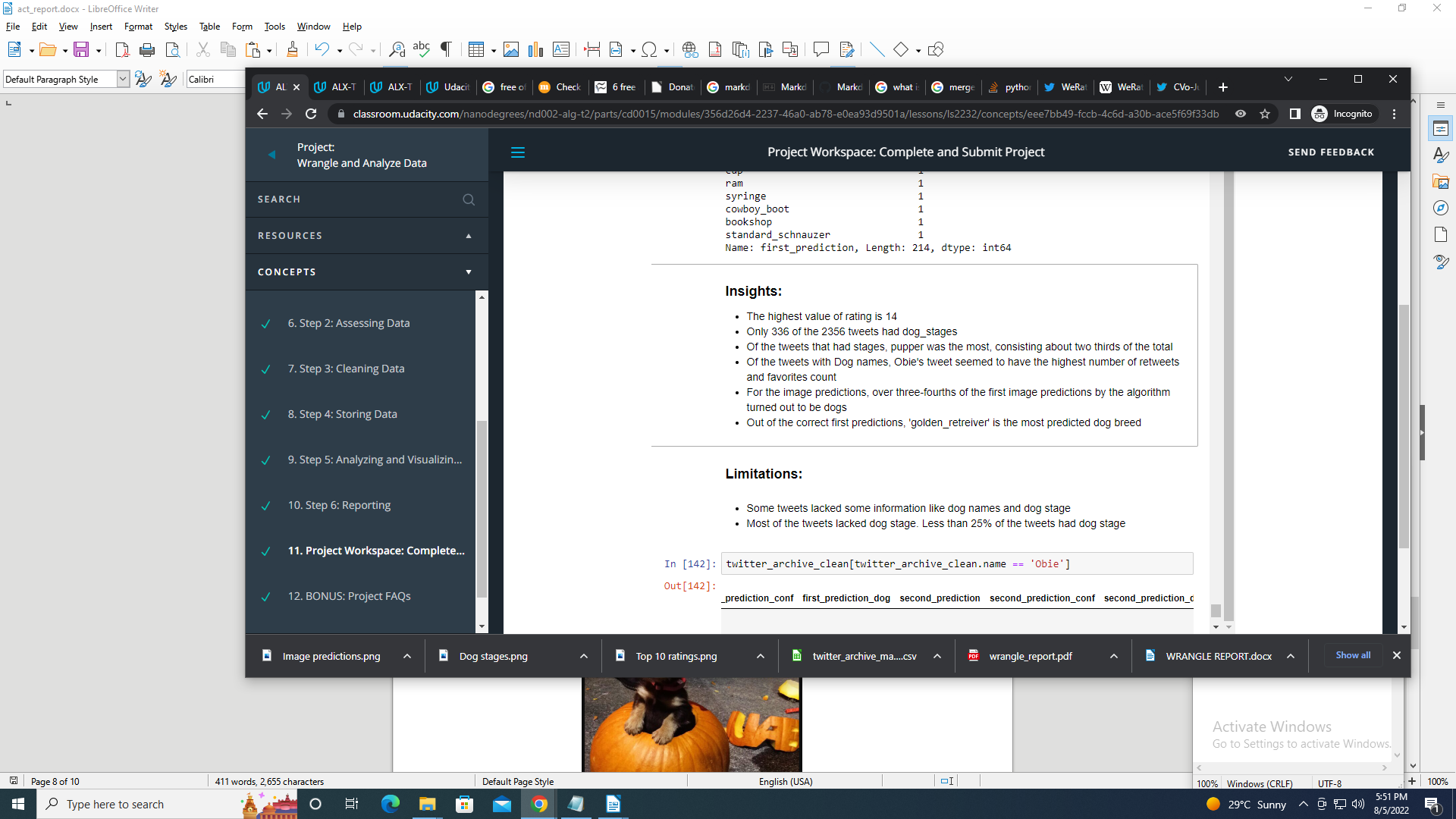


My name is Obie. I have the highest favorite count



Hello, my name is Stuart. I’m a golden retriever

**INSIGHTS**

****

**CONCLUSION**

The ‘WeRateDogs’ tweets served as a good data source to practice data wrangling and analysis

The most tweeted dog stage are those in the ‘pupper’ stage

The algorithm used to predict the dog types was had about 87.5% correct dog prediction rate for the first prediction

In-depth assessment of the dataframes is needed to identify all the possible issues and properly wrangle and analyse the data for better insight.