The project looked at 3 different data frames of unclean origin with the aim of cleaning, tidying, merging, analysing and generating insight from these dataframes. I first brought in the data frames by the different means available to me from the document.

Upon bringing in the dataframes I noticed errors such as improper extraction that I generated a code to combat in the first dataframe, duplicate cells were also noticed in which I cleared these to increase the quality of the dataset. I then checked manually for errors such as inconsistency in the name structure which I rectified by capitalising the names, invalid input such as names containing only the letter a were then removed to increase the quality of the data frame . I then looked to see how I could have reduced the number of columns without any input in them by clearing the empty cells. The time_frame column was in a format that was not favourable for the type of analysis I would be carrying out as such I changed the data type such that only the day of the week it falls on will be displayed.

There was a problem with the tidiness of the data frame as the different dog stages could be merged into one column reducing the need for the empty cells in the dataframe.

I then generated insight using the different charts available to me to be able to make useful decisions on the day of the week with the most tweets, the name of the person with the highest tweets, the dog breeds with the highest tweet counts.

I then looked at finding the correlation between the confidence interval for p1 and the favourite count. This helped me see there was a very week correlation between both of them as well as for p2 and p3