**Dataset Details**

Dataset used: <https://www.kaggle.com/gpreda/covid19-tweets>

Dataset size: 69mb

Number of tweets: 178638 tweets

Sentiment assigned: False

Timeframe of tweets: July 2020- August 2020

**Platform and System Details**

SYSTEM DETAILS:

CPU: AMD Ryzen 3 1200 @ 3.10 GHz

RAM: 16.0 GB

OS: Windows 10 Home 64-bit

GPU: Nvidia Geforce 1060 6 Gb

ENVIRONMENT DETAILS:

Running jupyter notebook on visual studio code using python v3.10.0

Version Control via github

CODING and its process

**Initial stuff**

Imported base libraries namely sklearn, pandas, matplotlib and numpy.

Read the dataset as twt using pandas

Displayed the head of the dataset using twt.head()

**Dataset exploration and preprocessing:**

**Exploration**

1. Show missing values Done
2. Print twt.info() Done
3. Print twt.describe() Done
4. Barplot of tweet locations Done
5. WordCloud before stopword removal Done

**PreProcessing**

1. Lowercasing: Done
2. Remove URLs: Done
3. Remove punctuations: Done
4. Remove words with numbers: Done
5. Remove emojis: Done
6. Stemming: Done
7. Stop word removal: Done
8. Assign sentiment to tweets using Sentiment Intensity Analyzer model from Vader: Done
9. If score < -0.25 its negative -0.25 < score < 0.25 its neutral 0.25 < score its positive: Done
10. Display sentiment of tweets vs days of the month: Done
11. Word cloud of different sentiment tweets: Done

**Feature Extraction**

1. Train/Test split: Done
2. TFIDF: Done