ds-midterm

January 9, 2023

1 I. Introduction

1.0.1 Domain specific area

For this project we chose to analyze data on the domain of social networks given its content related to natural language processing. We are analyzing texts coming from both instagram and twitter.

1.0.2 Dataset

For the first part of this project, the dataset we chose is a CSV file comprised of information about the top 50 most followed abbount on twitter. It contains features such as the Owner, if it is a brand or an account, the number of followers in millions, the profession of the account owner and the country/continent the account operates at. We will use this for gathering insights about the data and also statistics.

Since the dataset mentioned above has very few data, we chose a second dataset consisting of 3218 tweets from Elon Musk in order to apply a machine learning trained model to analyze the sentiment of each tweet.

Both the datasets are taken from kabble. Check the links below: 1. Top 50 list of most-followed instagram account 2. Elon Musk's Tweets

1.0.3 Objectives

The goal of this project is to collect some statistical information on top of the first dataset, such as: 1. The relationship between profession/activity and number of followers 2. Basic statistics such as the mean, standard deviation, maximum and minimum of followers.

Create some visualization for these statistics and then use a pre-trained Machine Learning model to classify the sentiments of the second dataset (a list of twitters from Elon Musk).

2 II. Implementation

2.0.1 Loading the data

[38]:
```
import pandas as pd

df = pd.read_csv('dataset.csv', encoding = "ISO-8859-1")
df
```

[38]:
```
    Rank              Username                 Owner Brand\naccount  \
0      1            @instagram             Instagram               +
1      2            @cristiano     Cristiano Ronaldo               -
2      3             @leomessi          Lionel Messi               -
3      4          @kyliejenner          Kylie Jenner               -
4      5          @selenagomez          Selena Gomez               -
5      6              @therock        Dwayne Johnson               -
6      7         @arianagrande         Ariana Grande               -
7      8         @kimkardashian        Kim Kardashian              -
8      9              @beyonce               Beyoncé               -
9     10       @khloekardashian      Khloé Kardashian               -
10    11         @justinbieber          Justin Bieber               -
11    12         @kendalljenner        Kendall Jenner               -
12    13                 @nike                  Nike               +
13    14               @natgeo    National Geographic               +
14    15          @taylorswift          Taylor Swift               -
15    16                  @jlo         Jennifer Lopez               -
16    17          @virat.kohli           Virat Kohli               -
17    18           @nickiminaj           Nicki Minaj               -
18    19      @kourtneykardash   Kourtney Kardashian               -
19    20            @neymarjr                Neymar               -
20    21           @mileycyrus           Miley Cyrus               -
21    22            @katyperry             Katy Perry               -
22    23              @zendaya               Zendaya               -
23    24        @kevinhart4real            Kevin Hart               -
24    25            @iamcardib               Cardi B               -
25    26            @ddlovato            Demi Lovato               -
26    27            @kingjames           LeBron James               -
27    28           @badgalriri               Rihanna               -
28    29           @realmadrid         Real Madrid CF               +
29    30         @theellenshow        Ellen DeGeneres               -
30    31        @champagnepapi                 Drake               -
31    32   @chrisbrownofficial           Chris Brown               -
32    33          @fcbarcelona           FC Barcelona               +
33    34         @billieeilish          Billie Eilish               -
34    35       @championsleague  UEFA Champions League               +
35    36            @gal_gadot              Gal Gadot               -
36    37             @k.mbappe          Kylian Mbappé               -
37    38             @dualipa               Dua Lipa               -
38    39                 @nasa                  NASA               +
39    40         @lalalalisa_m                  Lisa               -
40    41            @vindiesel             Vin Diesel               -
41    42        @priyankachopra        Priyanka Chopra               -
42    43             @khaby00             Khaby Lame               -
43    44            @snoopdogg             Snoop Dogg               -
44    45              @shakira               Shakira               -
45    46       @shraddhakapoor        Shraddha Kapoor               -
```

46	47	@davidbeckham	David Beckham	-
47	48	@gigihadid	Gigi Hadid	-
48	49	@victoriassecret	Victoria's Secret	+
49	50	@aliaabhatt	Alia Bhatt	-

	Followers(millions)[2]	Profession/Activity \
0	583.0	Social media platform
1	525.0	Footballer
2	411.0	Footballer
3	376.0	Television personality, model and businesswoman
4	366.0	Musician, actress, and businesswoman
5	355.0	Actor and professional wrestler
6	346.0	Musician, actress and businesswoman
7	337.0	Television personality, model and businesswoman
8	288.0	Musician, actress and businesswoman
9	285.0	Television personality and model
10	271.0	Musician
11	268.0	Model and television personality
12	259.0	Sportswear multinational
13	252.0	Magazine
14	237.0	Musician and actress
15	230.0	Musician and actress
16	229.0	Cricketer
17	207.0	Musician
18	206.0	Television personality and model
19	198.0	Footballer
20	191.0	Musician and actress
21	184.0	Musician
22	162.0	Actress and singer
23	161.0	Comedian and actor
24	146.0	Musician and actress
25	145.0	Musician and actress
26	140.0	Basketball player
27	139.0	Musician and businesswoman
28	129.0	Football club
29	129.0	Comedian and television personality
30	128.0	Musician
31	127.0	Musician
32	115.0	Football club
33	107.0	Musician
34	101.0	Club football competition
35	92.6	Actress
36	91.8	Footballer
37	87.5	Musician
38	86.0	Space agency
39	85.7	Musician
40	85.2	Actor

41	84.1	Actress and musician
42	80.2	Social media personality
43	78.5	Musician
44	77.9	Musician
45	76.9	Actress
46	76.6	Former FootballerPresident of Inter Miami
47	76.4	Model
48	73.8	Lingerie company
49	73.6	Actress and musician

	Country/Continent
0	United States
1	Portugal
2	Argentina
3	United States
4	United States
5	United States
6	United States
7	United States
8	United States
9	United States
10	Canada
11	United States
12	United States
13	United States
14	United States
15	United States
16	India
17	Trinidad and Tobago United States
18	United States
19	Brazil
20	United States
21	United States
22	United States
23	United States
24	United States
25	United States
26	United States
27	Barbados
28	Spain
29	United States
30	Canada
31	United States
32	Spain
33	United States
34	Europe
35	Israel

```
36                                    France
37                 United Kingdom Albania
38                          United States
39                               Thailand
40                          United States
41                                  India
42                          Italy Senegal
43                          United States
44                               Colombia
45                                  India
46                         United Kingdom
47                          United States
48                          United States
49                   United Kingdom India
```

2.0.2 Statistical Summary and Data Visualization

[7]:
```python
# Inspecting a little more on the dataframe: 50 rows and 7 columns
df.shape
```

[7]:
```
(50, 7)
```

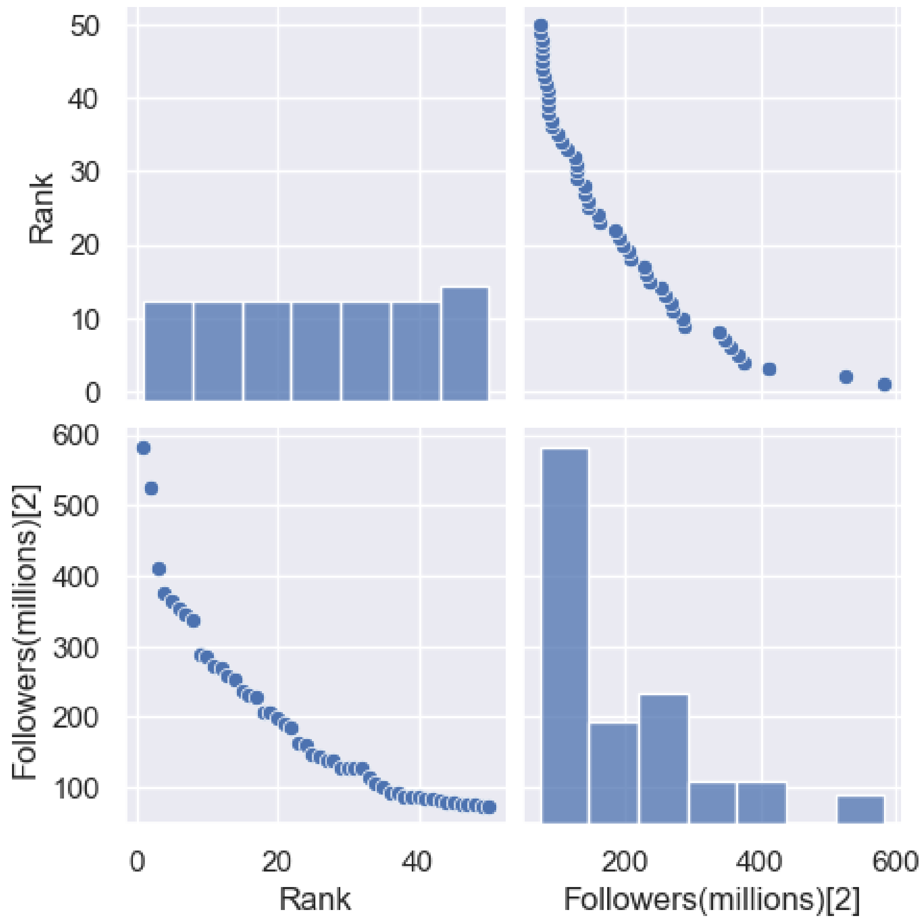
[10]:
```python
# Inspecting information about the datatypes
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 7 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Rank                    50 non-null     int64
 1   Username                50 non-null     object
 2   Owner                   50 non-null     object
 3   Brand
account             50 non-null     object
 4   Followers(millions)[2]  50 non-null     float64
 5   Profession/Activity     50 non-null     object
 6   Country/Continent       50 non-null     object
dtypes: float64(1), int64(1), object(5)
memory usage: 2.9+ KB
```

[11]:
```python
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()
```

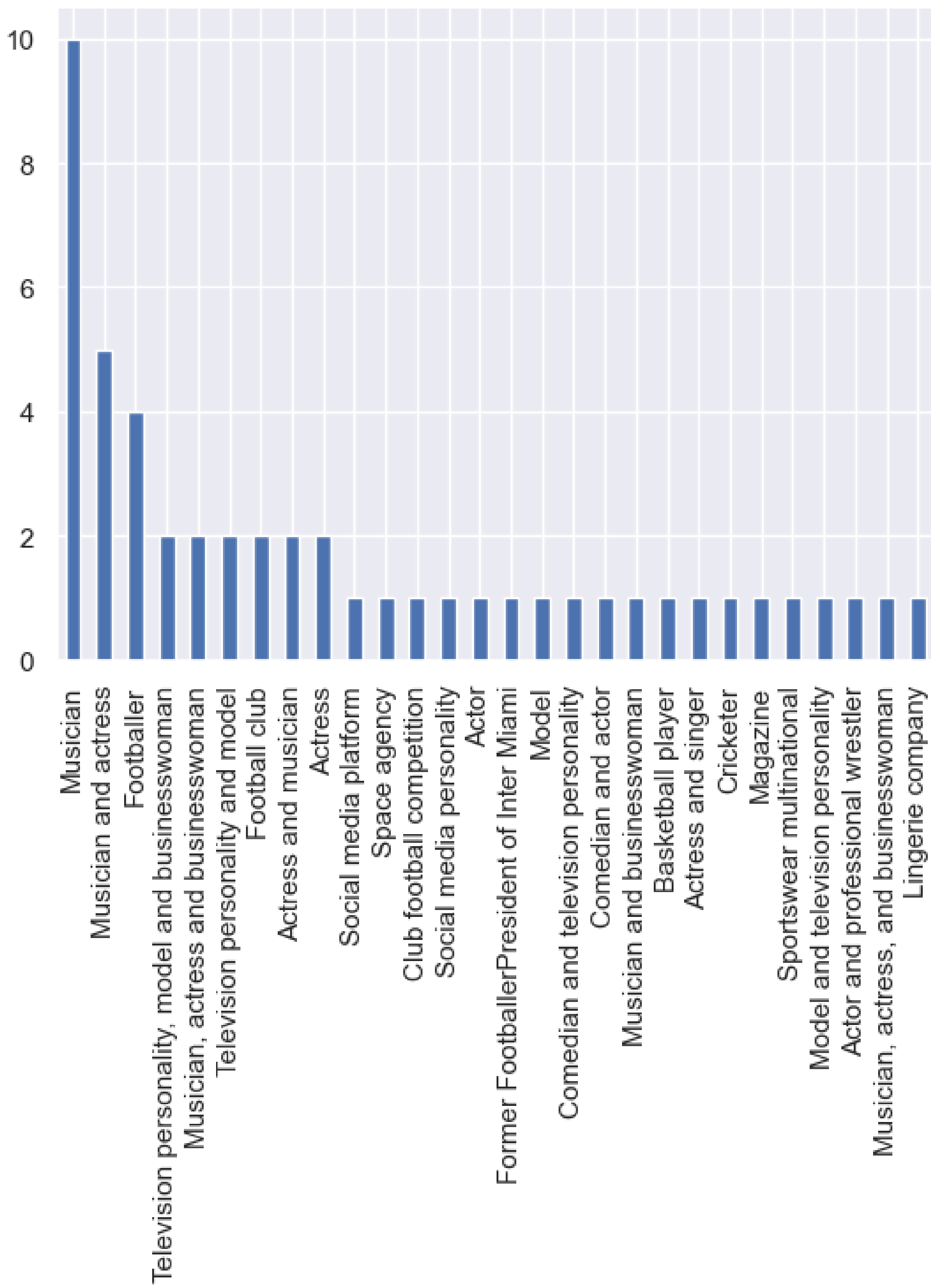
[14]:
```
# Finding nulls
sns.pairplot(df)
```

[14]: <seaborn.axisgrid.PairGrid at 0x14721a140>

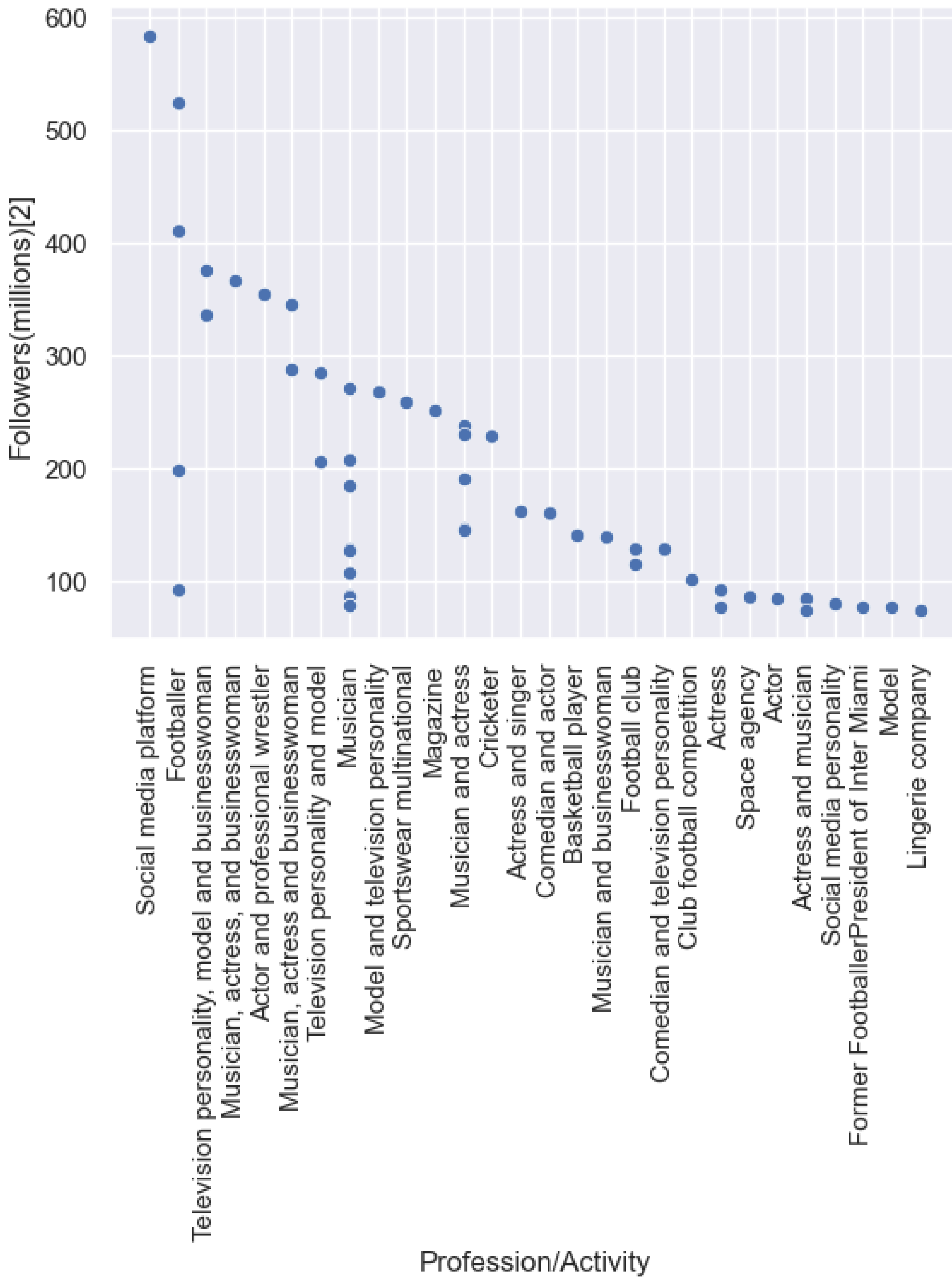


[17]:
```
# A value count by distinct professions
df['Profession/Activity'].value_counts().plot(kind='bar')
```

[17]: <AxesSubplot: >



[21]:
```
# A scatterplot of number of followers by profession activity
sns.scatterplot(x=df['Profession/Activity'],y=df['Followers(millions)[2]'])
plt.xticks(rotation=90)
plt.show()
```



[39]:
```python
# Some more statistical information about the numbers of followers
print('Mean of followers (in millions):',df['Followers(millions)[2]'].mean())
print('Standard deviation (in millions):',df['Followers(millions)[2]'].std())
```

```
print('Minimum number of followers (in millions):',df['Followers(millions)[2]'].
  ↪min())
print('Maximum number of followers (in millions):',df['Followers(millions)[2]'].
  ↪max())
print('90percentile: (in millions):',df['Followers(millions)[2]'].quantile(0.
  ↪90))
print('70percentile: (in millions):',df['Followers(millions)[2]'].quantile(0.
  ↪70))
print('25percentile: (in millions):',df['Followers(millions)[2]'].quantile(0.
  ↪25))
```

```
Mean of followers (in millions): 191.19600000000003
Standard deviation (in millions): 121.43767186066594
Minimum number of followers (in millions): 73.6
Maximum number of followers (in millions): 583.0
90percentile: (in millions): 356.1
70percentile: (in millions): 232.09999999999997
25percentile: (in millions): 88.575
```

2.1 Preprocessing

[61]:
```
## Preprocessing
# Defining the method that is going to preprocess the data so we can use it␣
  ↪later
import nltk
from nltk.corpus import stopwords

# Preprocess the sentences
def preprocess(sentences):
    # Tokenize the sentences: transforming the whole string into a list of␣
  ↪tokens.
    sentences = [nltk.word_tokenize(sentence) for sentence in sentences]
    # Removing stopwords
    stop_words = stopwords.words('english')
    sentences = [word for word in sentences if not word in stop_words]

    return sentences
```

2.1.1 Machine Learning Model

[62]:
```
# Loading the new CSV
tweets_df = pd.read_csv('data_elonmusk.csv', encoding='latin-1')
tweets_df.head(10)
```

```
[62]:   row ID                                                Tweet  \
0   Row0  @MeltingIce Assuming max acceleration of 2 to …
1   Row1  RT @SpaceX: BFR is capable of transporting sat…
2   Row2                                   @bigajm Yup :)
3   Row3                      Part 2 https://t.co/8Fvu57muhM
4   Row4  Fly to most places on Earth in under 30 mins a…
5   Row5  RT @SpaceX: Supporting the creation of a perma…
6   Row6  BFR will take you anywhere on Earth in less th…
7   Row7  Mars City\nOpposite of Earth. Dawn and dusk sk…
8   Row8             Moon Base Alpha https://t.co/voY8qEW9kl
9   Row9  Will be announcing something really special at…

                  Time Retweet from      User
0  2017-09-29 17:39:19          NaN  elonmusk
1  2017-09-29 10:44:54       SpaceX  elonmusk
2  2017-09-29 10:39:57          NaN  elonmusk
3  2017-09-29 09:56:12          NaN  elonmusk
4  2017-09-29 09:19:21          NaN  elonmusk
5  2017-09-29 08:57:29       SpaceX  elonmusk
6  2017-09-29 08:53:00          NaN  elonmusk
7  2017-09-29 06:03:32          NaN  elonmusk
8  2017-09-29 05:44:55          NaN  elonmusk
9  2017-09-29 02:36:17          NaN  elonmusk
```

[64]:
```
# Preprocessing the tweets so we can later analyze
sentences = preprocess(tweets_df.Tweet.values)
```

2.1.2 Using Vader, a pre-trained Machine Learning Model to analyze the sentiment of the tweets

[68]:
```
# Analyzing if tweet is positive or negative using VADER's Sentiment Analyzer
from nltk.sentiment import SentimentIntensityAnalyzer

nltk.download('vader_lexicon')

sia = SentimentIntensityAnalyzer()

def get_polarity_score(tweet: str) -> bool:
    return sia.polarity_scores(tweet)


def classify(sentences):
    analysis_result_json = []
    for element in sentences:

        sentiment = 'POS' if get_polarity_score(element)["compound"] > 0 else␣
 ↪'NEG'
```

```python
        compound = get_polarity_score(element)["compound"]

        sentiment_dictionary = {
            "Sentence": element,
            "Compound Score": compound,
            "Sentiment": sentiment
        }
        analysis_result_json.append(sentiment_dictionary)
    return analysis_result_json;

classify(tweets_df.Tweet.values[:10])
```

```
[nltk_data] Downloading package vader_lexicon to
[nltk_data]     /Users/lucas.viola/nltk_data…
[nltk_data]   Package vader_lexicon is already up-to-date!
```

[68]:
```
[{'Sentence': "@MeltingIce Assuming max acceleration of 2 to 3 g's, but in a
comfortable direction. Will feel like a mild to moder? https://t.co/fpjmEgrHfC",
  'Compound Score': 0.8271,
  'Sentiment': 'POS'},
 {'Sentence': 'RT @SpaceX: BFR is capable of transporting satellites to orbit,
crew and cargo to the @Space_Station and completing missions to the Moon an?',
  'Compound Score': 0.3818,
  'Sentiment': 'POS'},
 {'Sentence': '@bigajm Yup :)', 'Compound Score': 0.4588, 'Sentiment': 'POS'},
 {'Sentence': 'Part 2 https://t.co/8Fvu57muhM',
  'Compound Score': 0.0,
  'Sentiment': 'NEG'},
 {'Sentence': 'Fly to most places on Earth in under 30 mins and anywhere in
under 60. Cost per seat should be? https://t.co/dGYDdGttYd',
  'Compound Score': 0.0,
  'Sentiment': 'NEG'},
 {'Sentence': 'RT @SpaceX: Supporting the creation of a permanent, self-
sustaining human presence on Mars. https://t.co/kCtBLPbSg8
https://t.co/ra6hKsrOcG',
  'Compound Score': 0.6124,
  'Sentiment': 'POS'},
 {'Sentence': 'BFR will take you anywhere on Earth in less than 60 mins
https://t.co/HWt9BZ1FI9',
  'Compound Score': 0.0,
  'Sentiment': 'NEG'},
 {'Sentence': 'Mars City\nOpposite of Earth. Dawn and dusk sky are blue on Mars
and day sky is red. https://t.co/XHcZIdgqnb',
  'Compound Score': 0.0,
  'Sentiment': 'NEG'},
 {'Sentence': 'Moon Base Alpha https://t.co/voY8qEW9kl',
  'Compound Score': 0.0,
```

```
  'Sentiment': 'NEG'},
 {'Sentence': "Will be announcing something really special at today's talk
https://t.co/plXTBJY6ia",
  'Compound Score': 0.4576,
  'Sentiment': 'POS'}]
```

3 III. Conclusions

In terms of the first dataset (the list of the most followed accounts on instagram), we can see that it is easier to gather insights on the data than it is to do any ML evaluation in it. This is due to the properties of the data in the spreadsheet which are most objects. In our analysis we were able to capture insights such as information about the kind of professions that have the most followers, we showed this information both in a scatterplot format and in a bar plot format, and also we gathered some basic statistics on top of the datasetm such as the mean of followers and standard deviation comparing all acounts.

As for the second dataset, it was easier then to apply Machine Learning to it because we had more data in the form of sentences in the english language which we could use to do a Sentiment Analysis. Since for this dataset we did not have any previous classification in order to train a new model we decided to use a pre-trained model called Vader which is part of the NLTK library. With this model we could check information such as if the sentiment of the tweet is positive or negative and also the compound score (the polarity score) of each tweet.