**PRINCIPLES OF BIG DATA MANAGEMENT**

**PHASE 2**

**Twitter Analysis on Football Data Report**

**Team Member:**

**V H N V R S Sai Pavan (16296461)**

**Pallavi Arikatla (16301244)**

**Sai Krishna Reddy Katta (16297003)**

**Link**

Tweets data:

<https://drive.google.com/file/d/1PNxu_YzMjVvnzpCAodXf65msFHZklCIi/view?usp=sharing>

Github Code Link:

<https://github.com/vattemsaipavan/Principles_of_bigdata>

**OBJECTIVE**

* Stream tweets from Twitter (we collected more than 500,000 tweets).
* Write at least 11 analytic queries to explore and understand the collected tweets.
* Create Visualizations on the results.
* Create a web page to view the visualizations.

**SOFTWARES PLANNING TO USE**

* Apache Spark
  + Spark Streaming to collect tweets.
  + Spark SQL to store and execute queries.
* Matplot, Seaborn and word cloud for visualization
* HTML, CSS, Bootstrap and JavaScript for web application

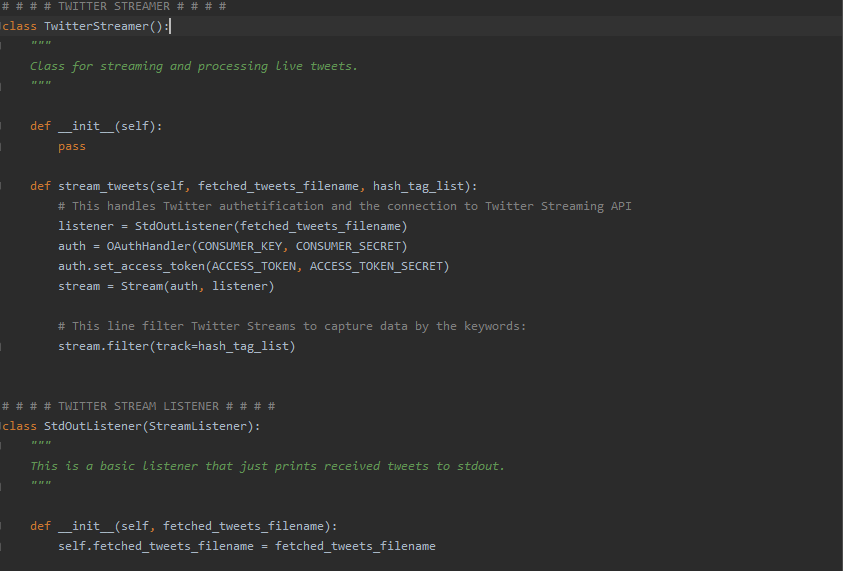
**IMPLEMENTATION**

* Initially, we wrote a python program to stream the tweets and save them into JSON format which will have details of the tweets.
* The extracted JSON tweets are persisted into the Apache SparkSQL in the form of Views.
* We used PyCharm IDE to write queries and visualize the outputs using tables and charts.
* The web application is developed using HTML, CSS and Java script.

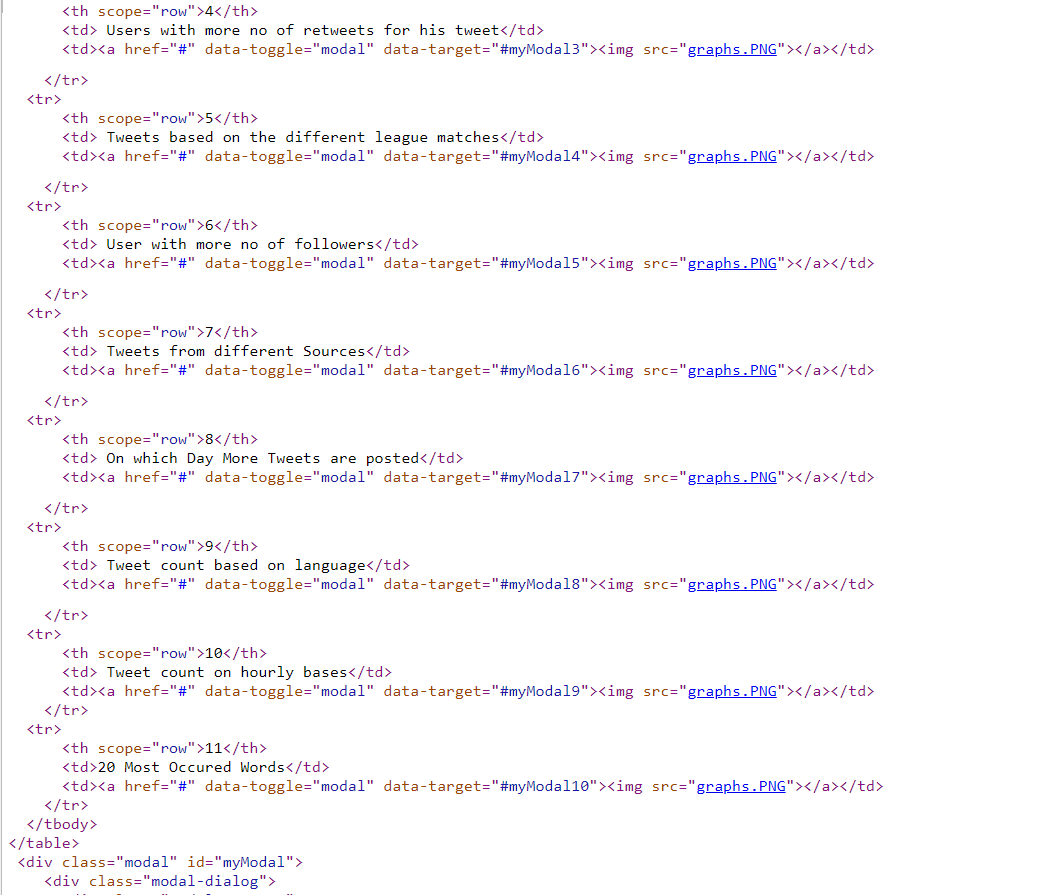
Libraries used:

* Matplot for visualizations
* Text Blob to get sentimental score
* Flask for web application

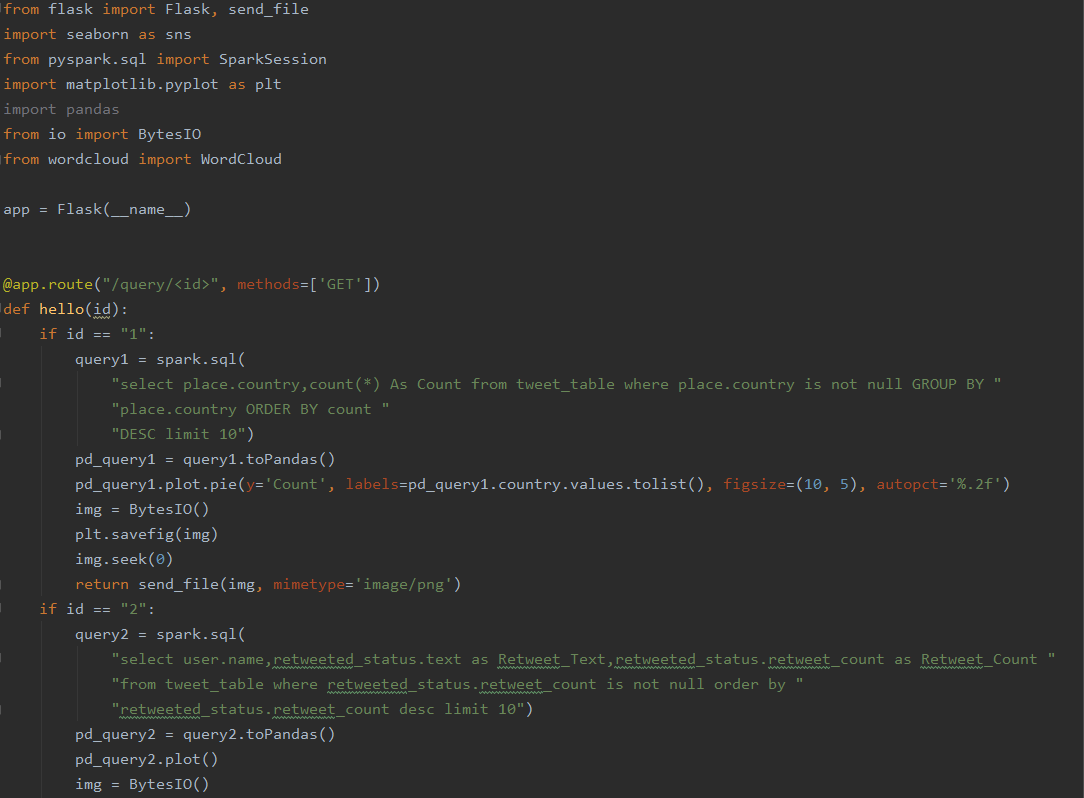
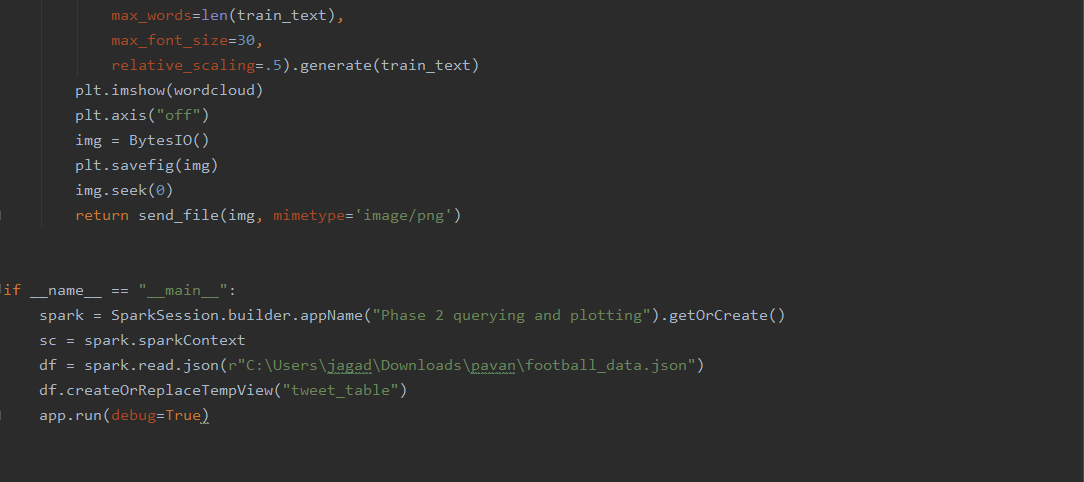
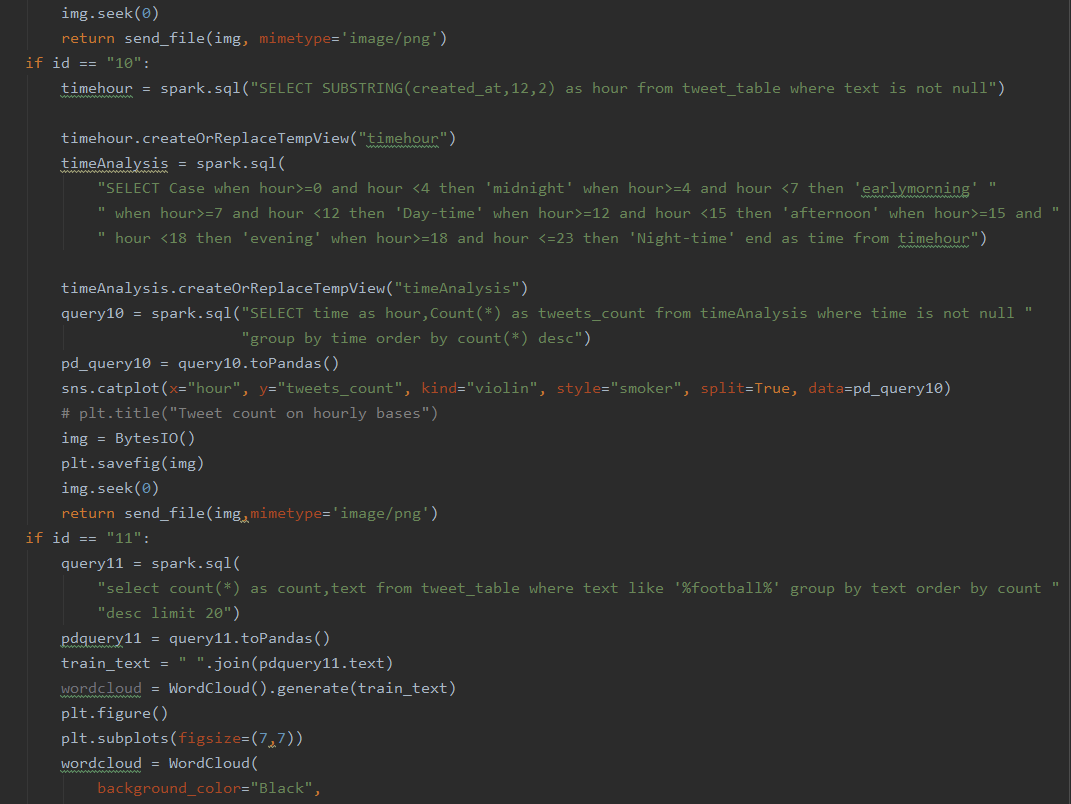
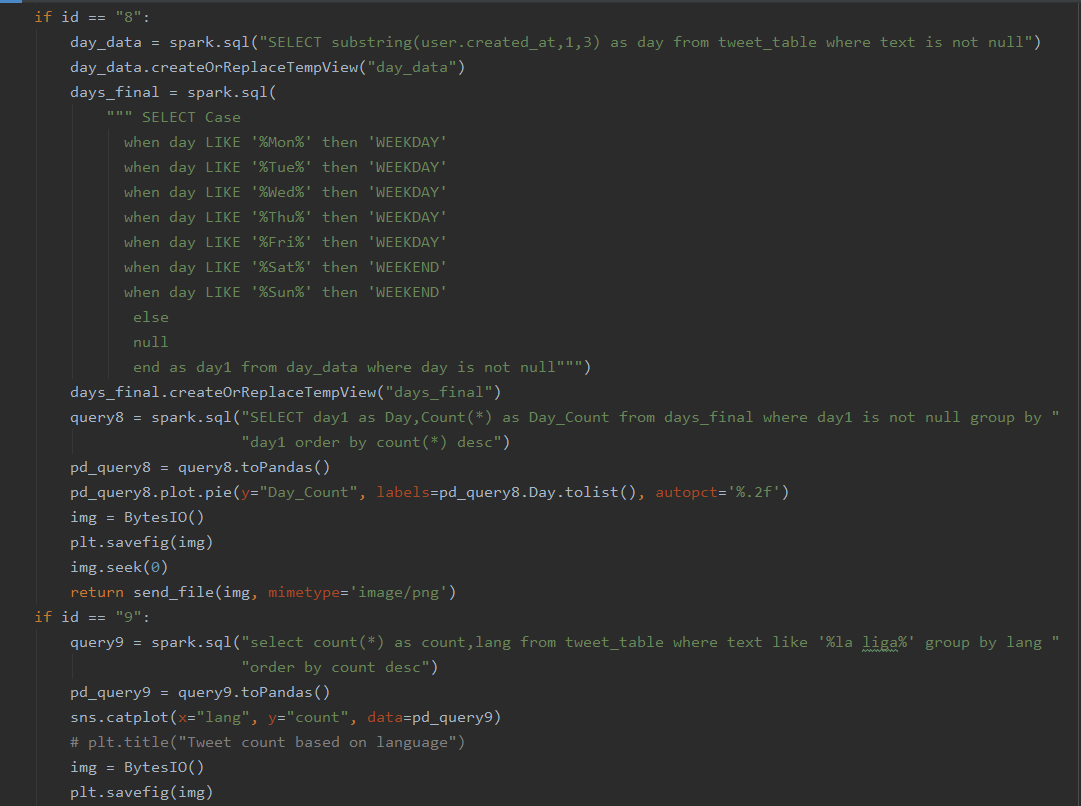
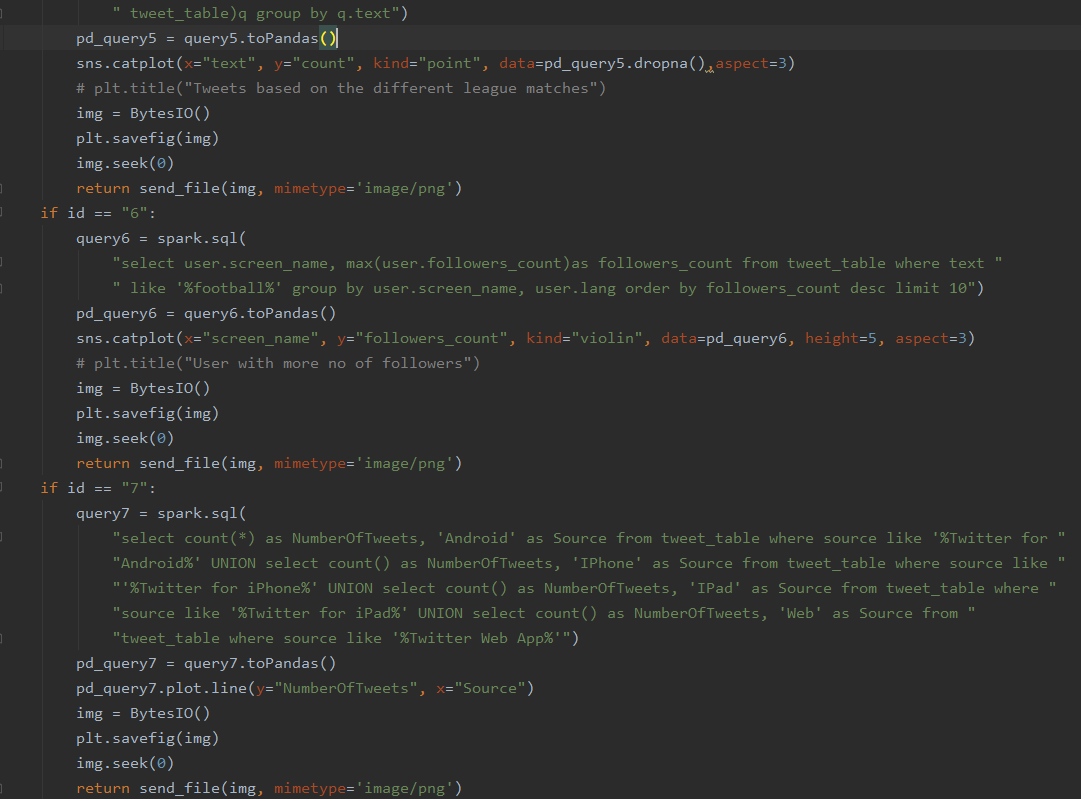
Code for Collecting the Tweets:





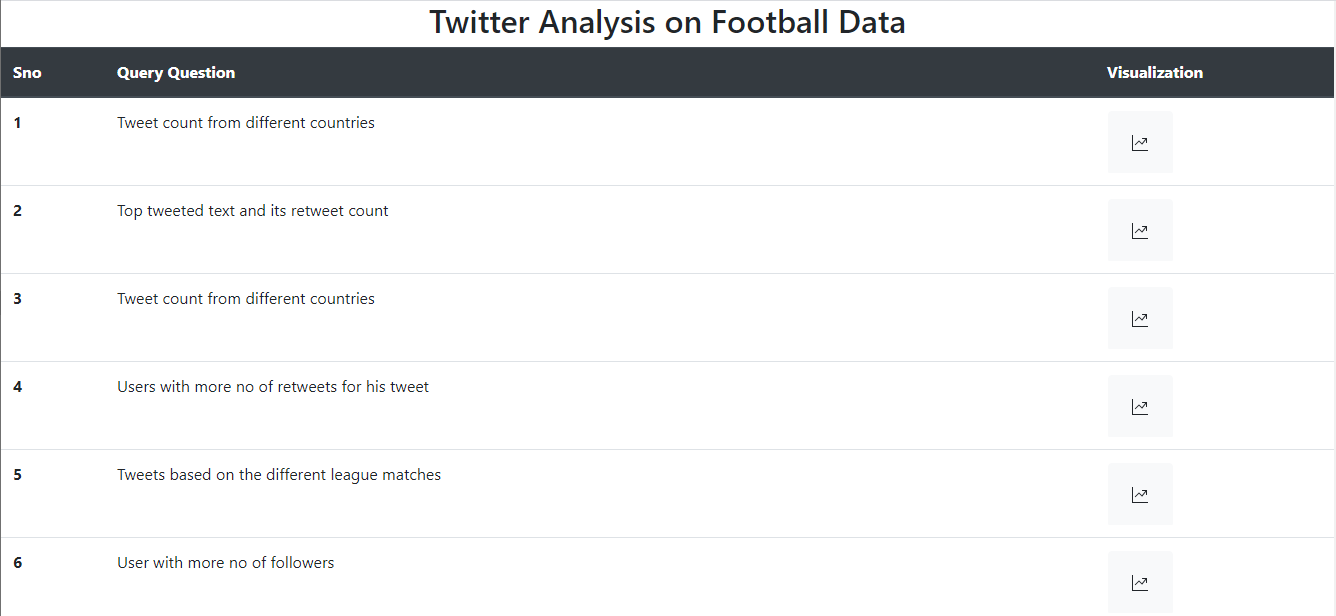
HTML and JavaScript code for UI: 

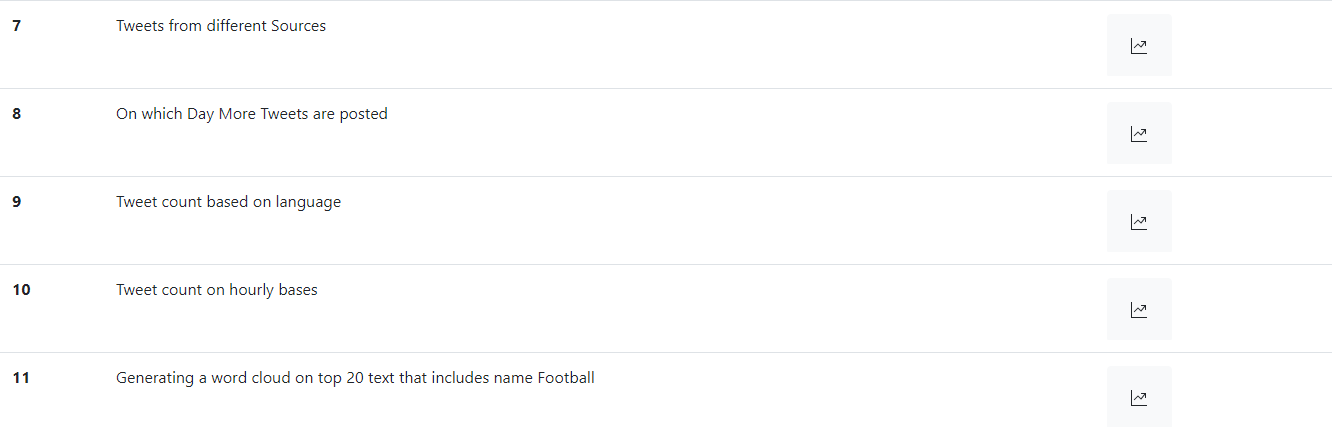


Code For 10 Queries: 

**Query Execution and Results:**

**Home Page**



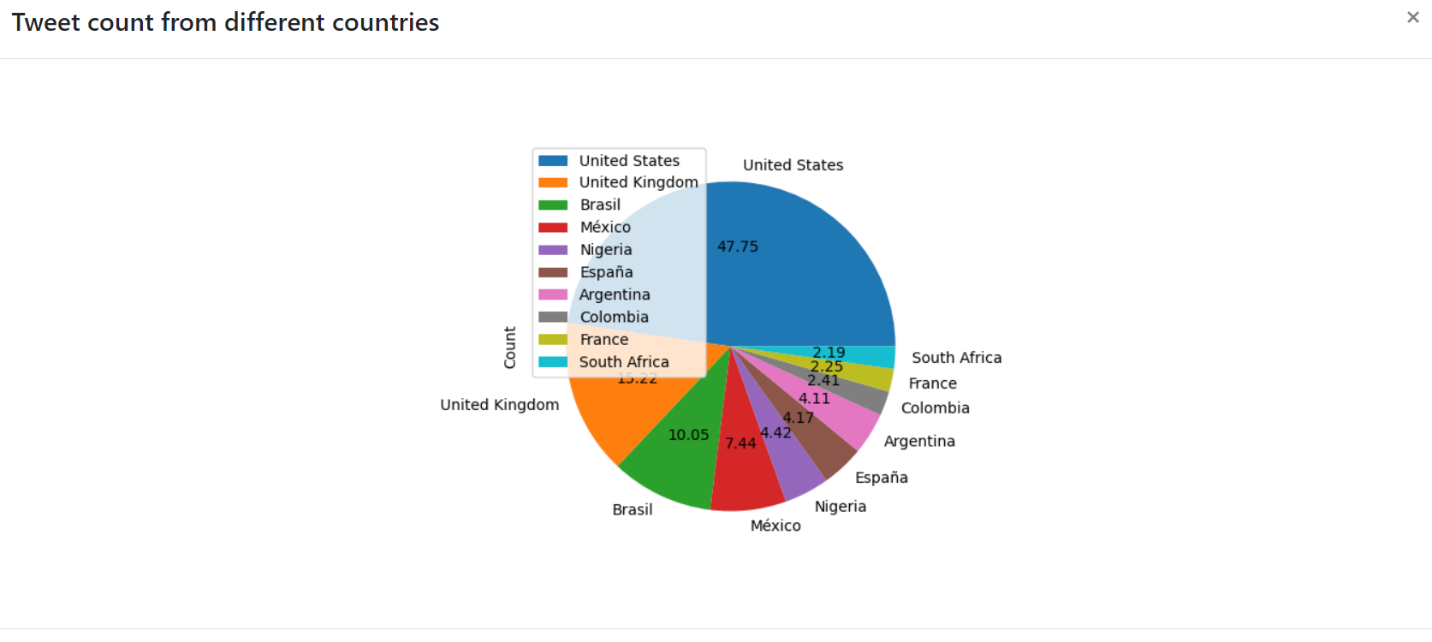


We have executed and visualized totally 10 queries.

1). Tweet count from different countries using Pie Chart.

Query result in table form :

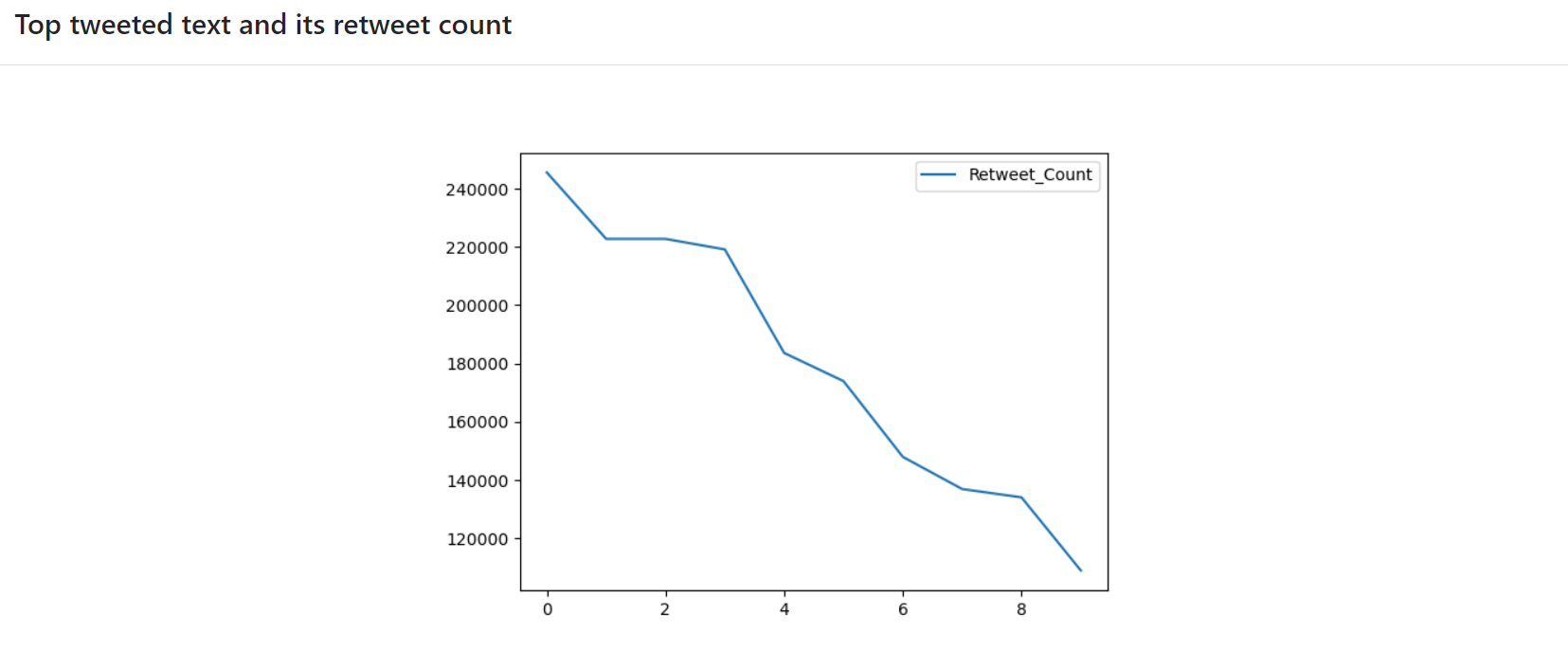




2) Query to display the Top tweeted text and its retweet count using “line plot”.

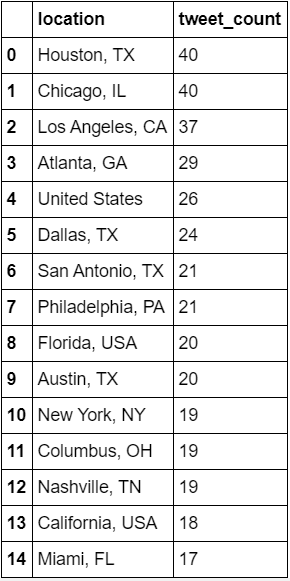
Result:

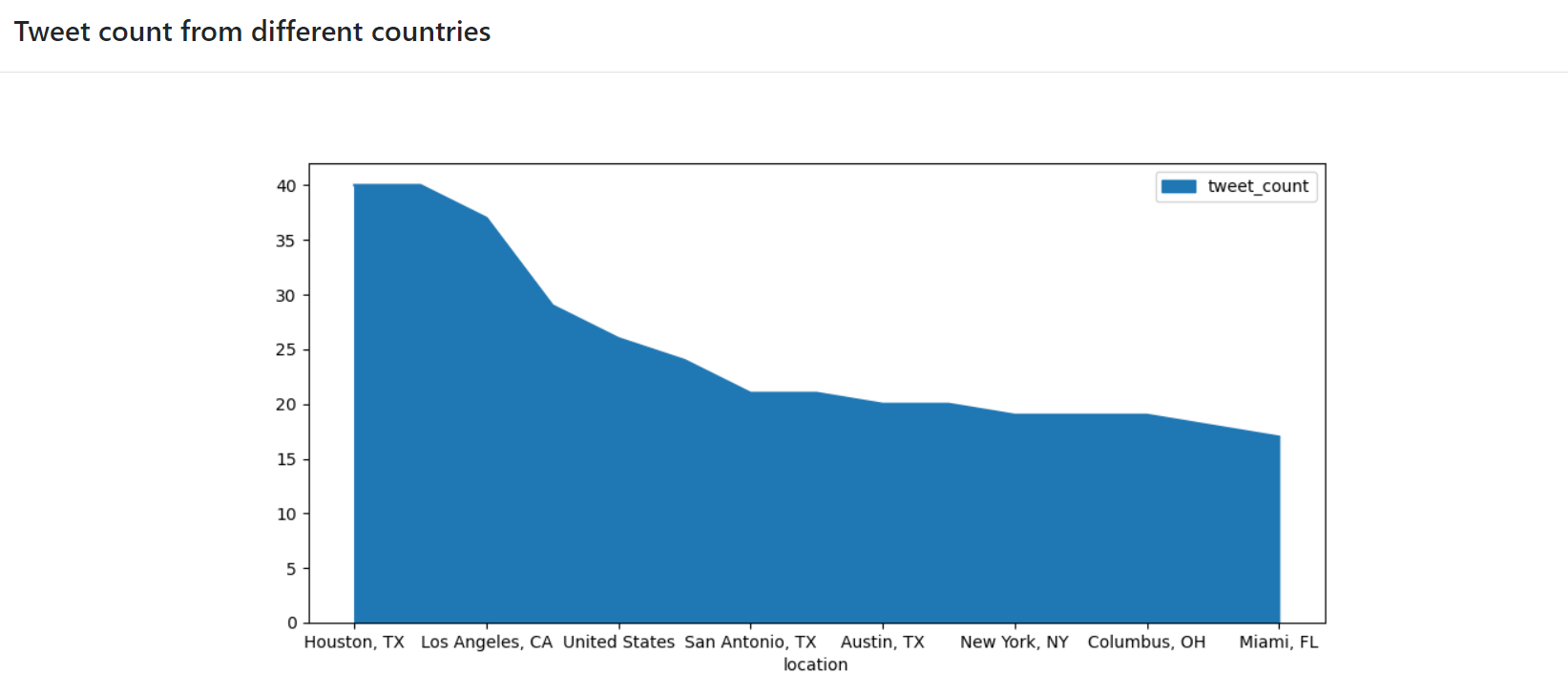




3). Tweet count from different countries using “Area Chart”

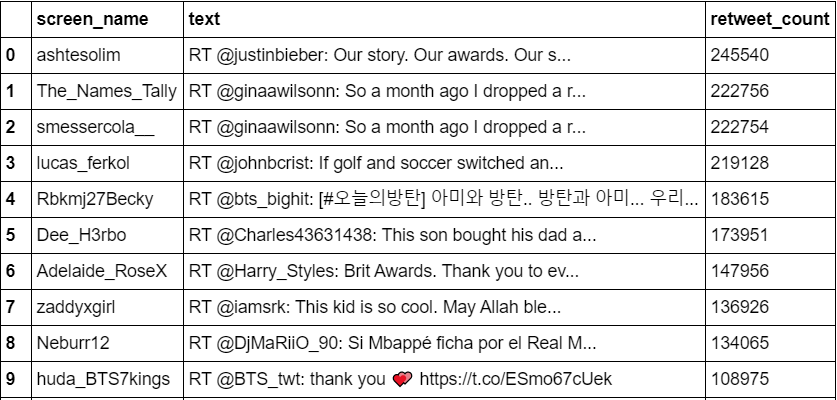
Result:

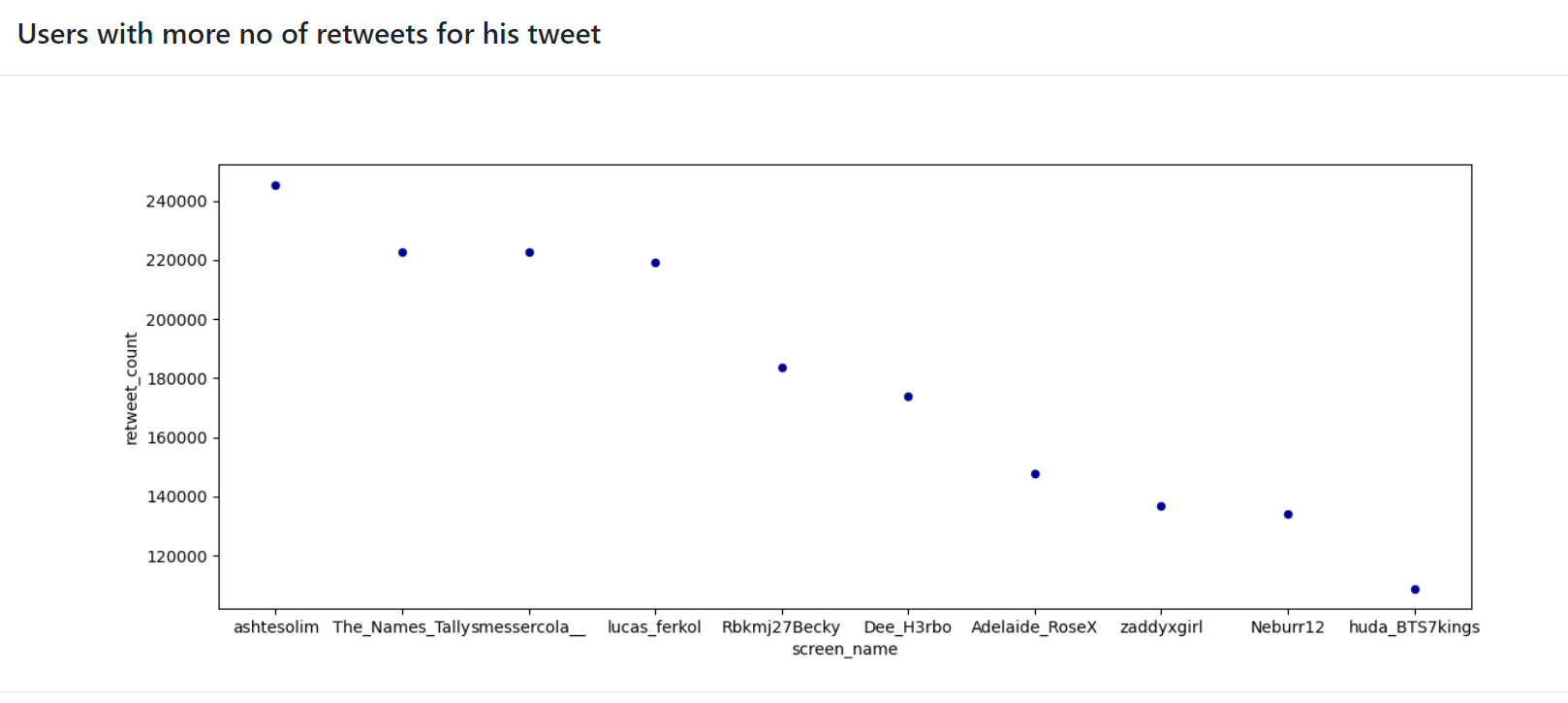




4). Users with more no of retweets for his tweet using “Scatter plot”

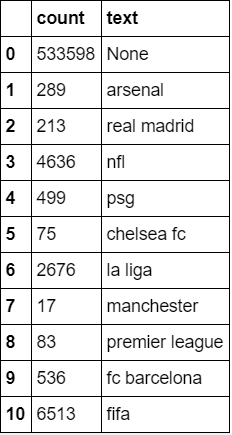
Result:

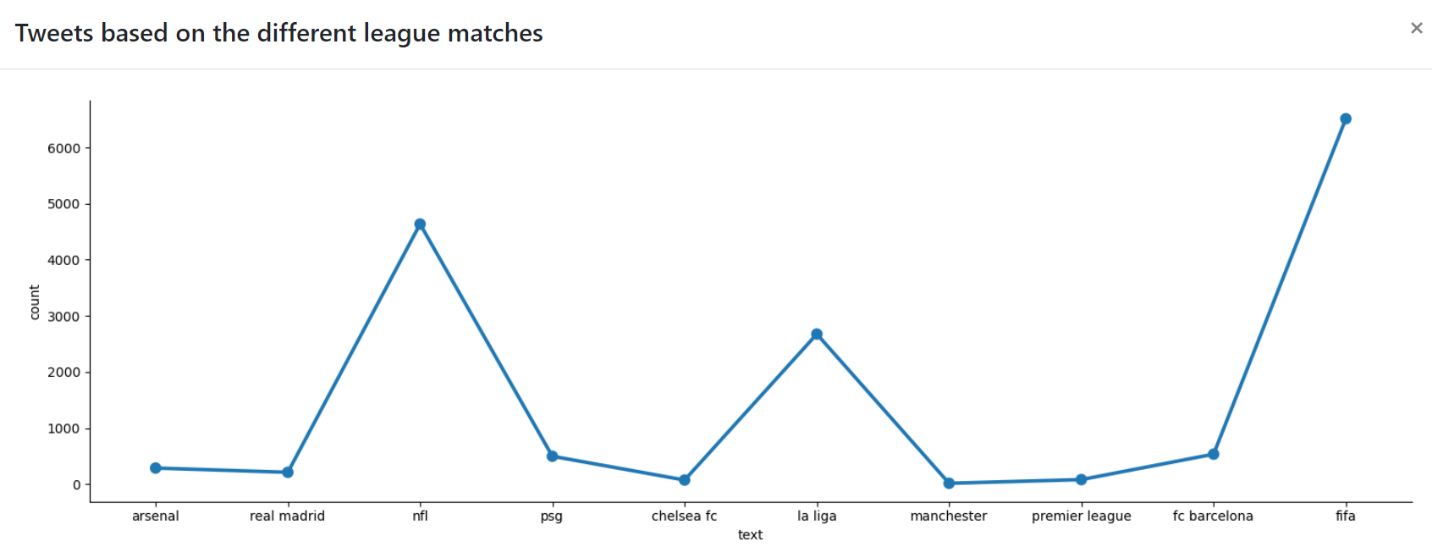




5). Tweets based on the different league matches using “Seaborn Point Code”

Result:

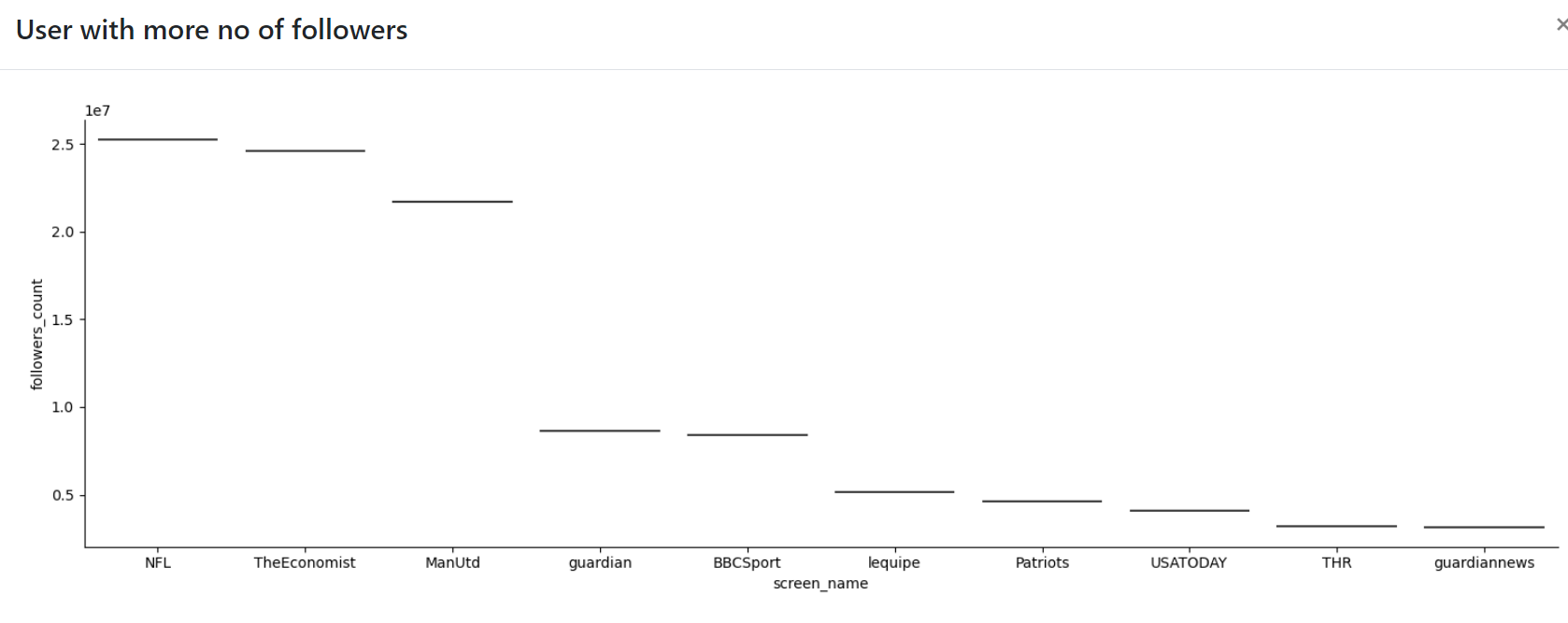




6). User with more no of followers using “Seaborn violin Graph”

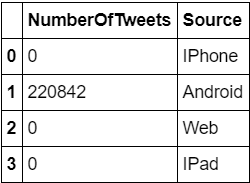
Result:

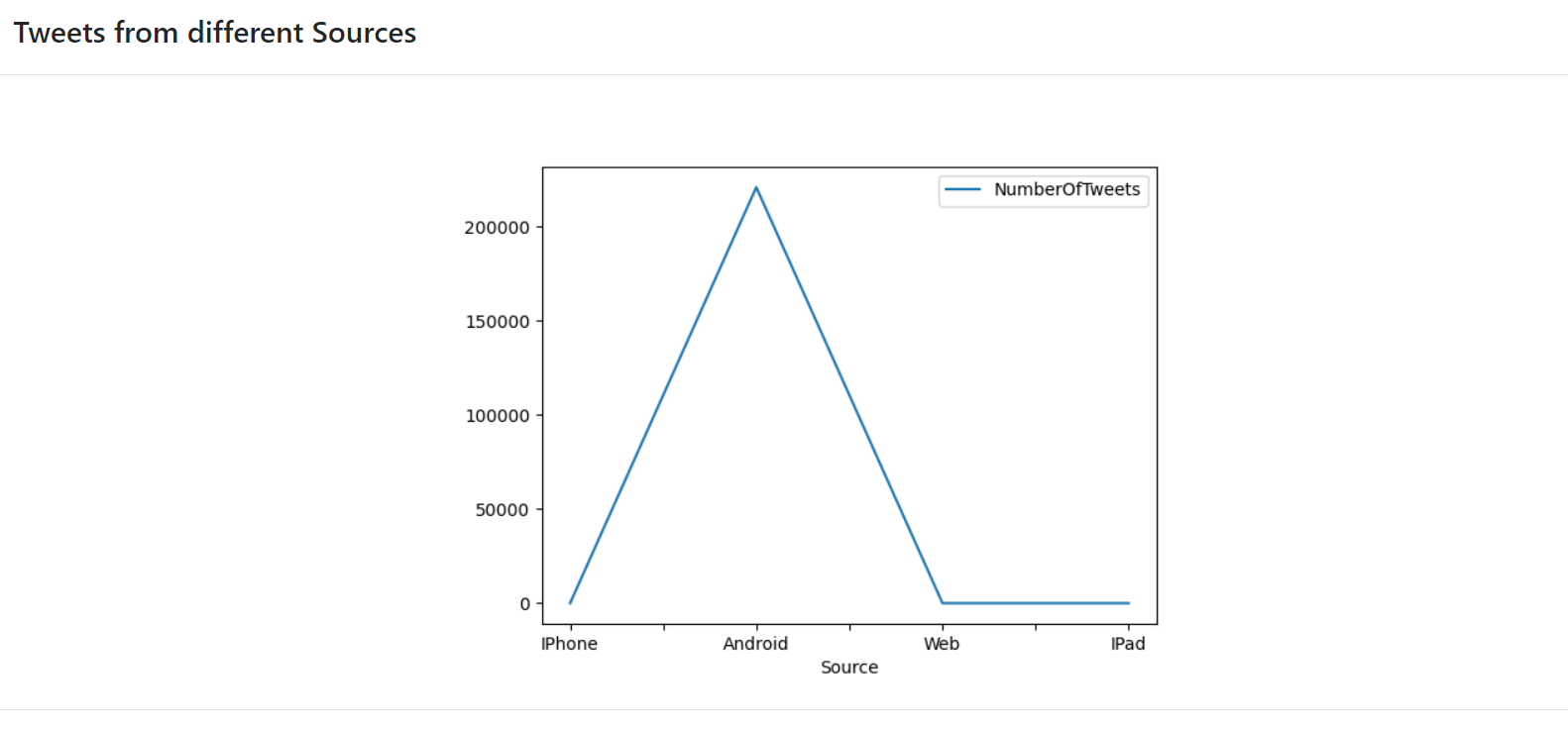




7). Tweets from different Sources using (Line Graph)

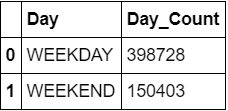
Result:

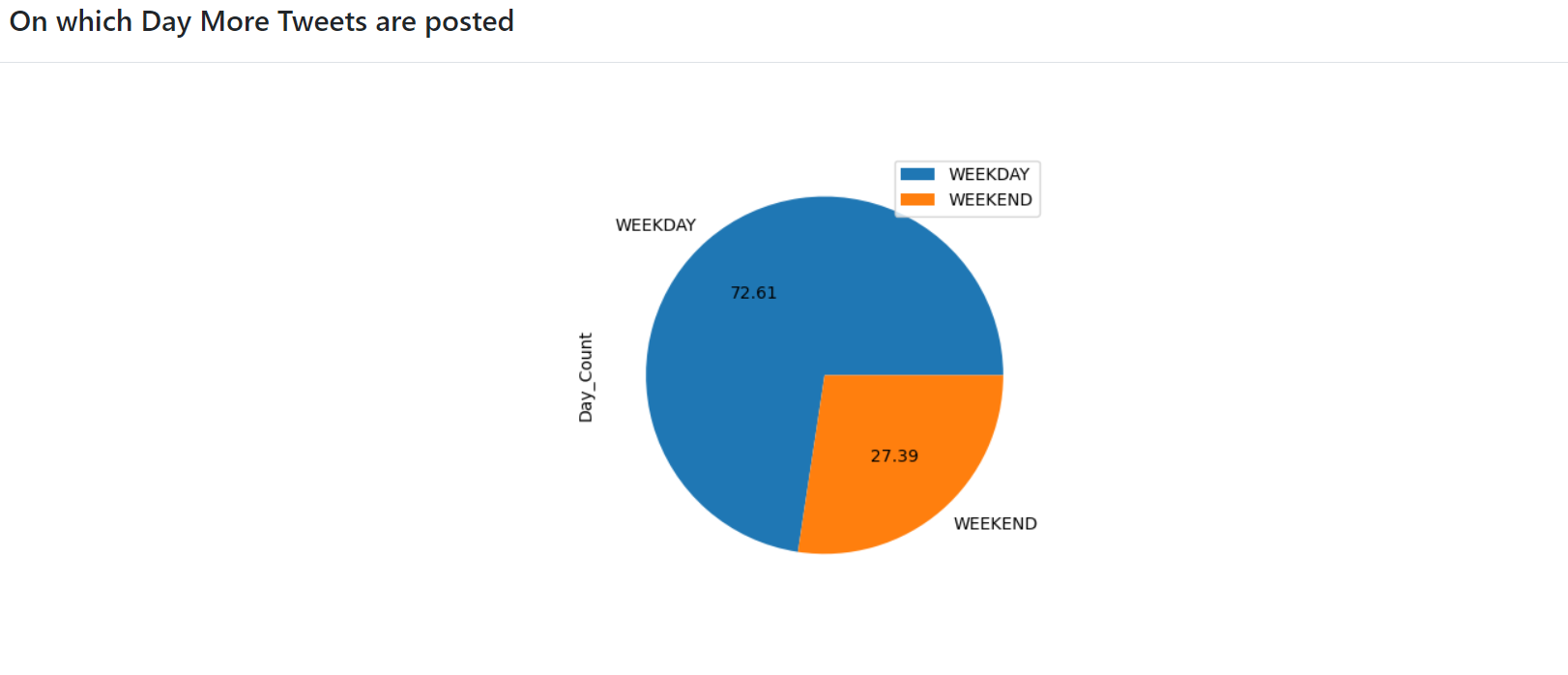




8). On which Day More Tweets are posted using “Pie Graph”

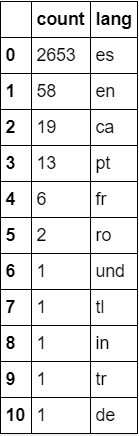
Result:

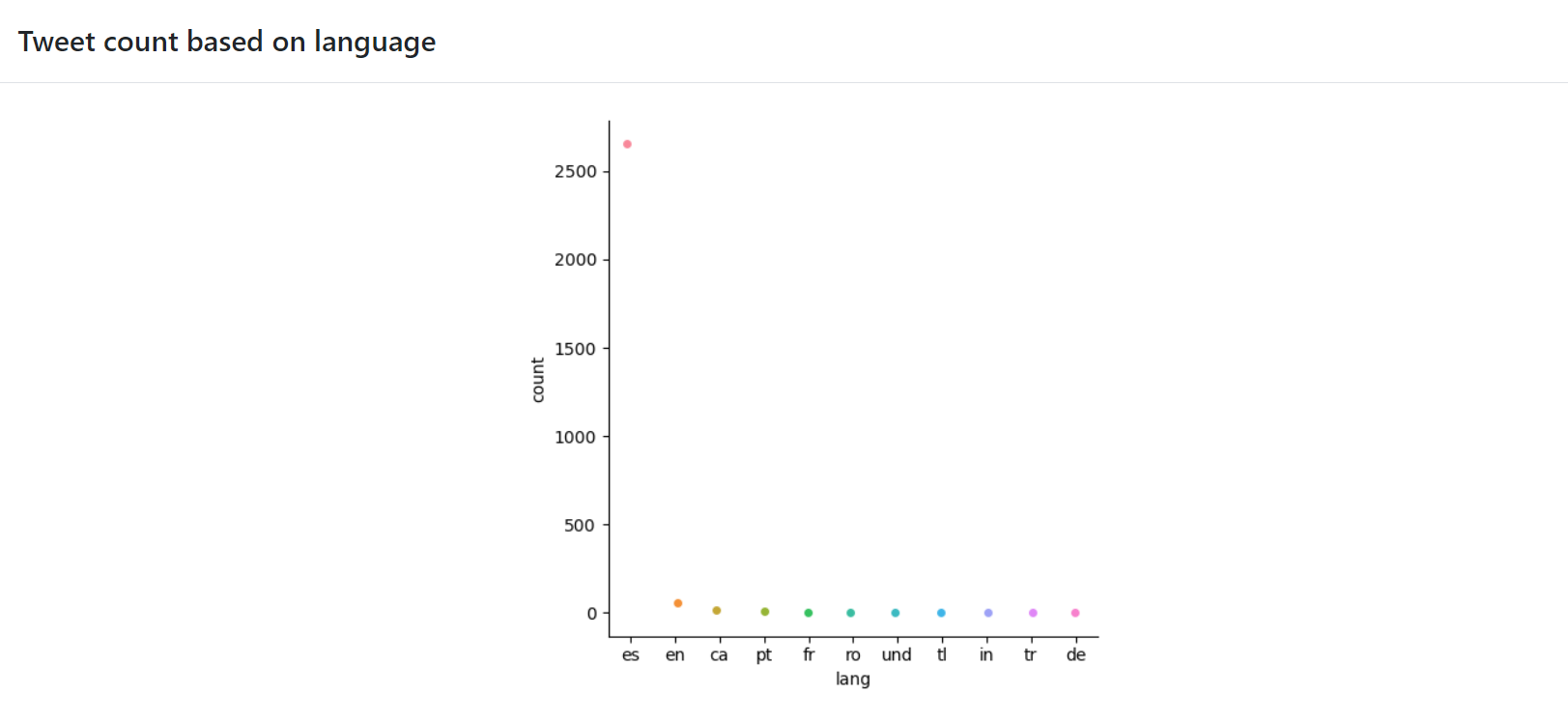




9). Tweet count based on language using “Seaborn Category Plot)

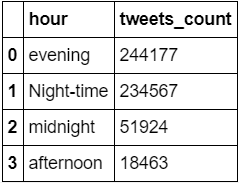
Result:

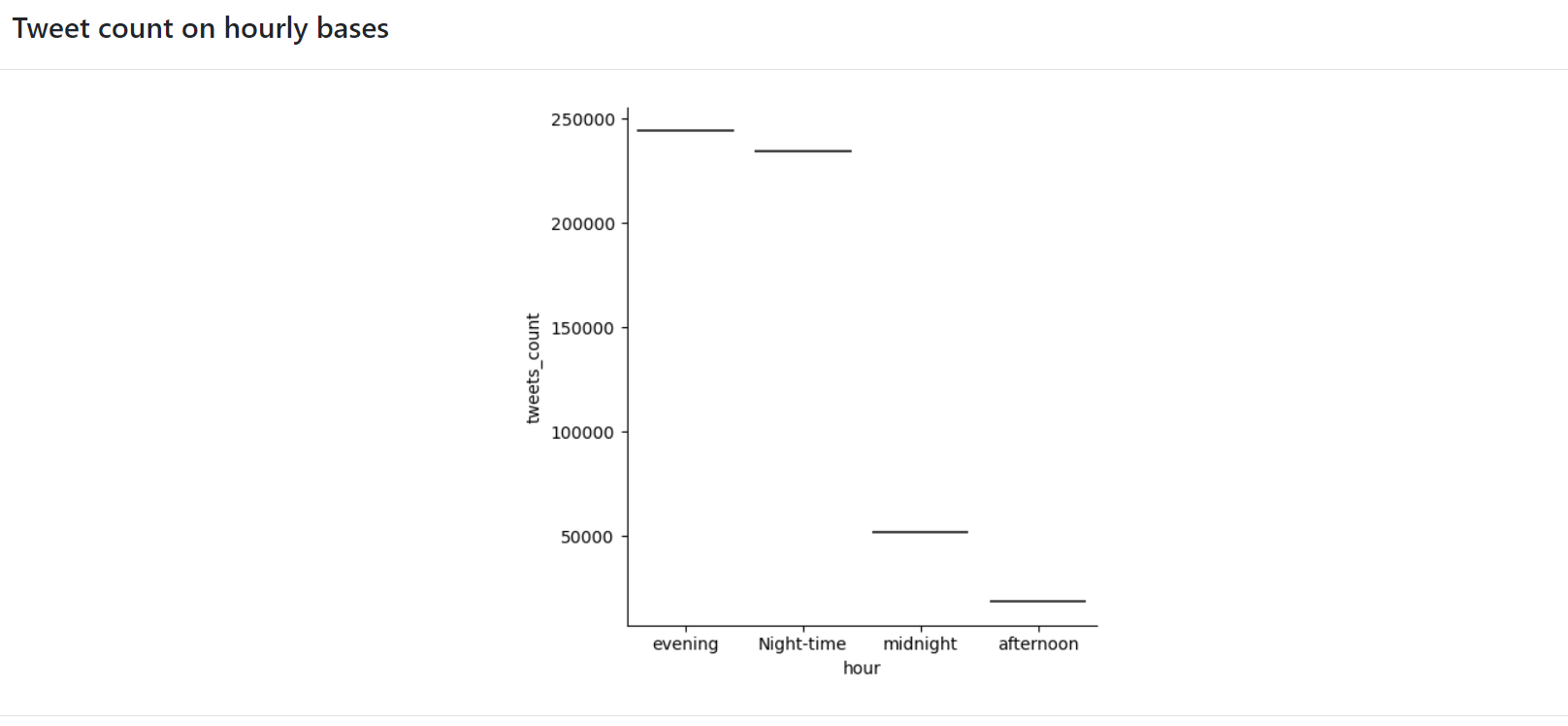




10). Tweet count on hourly bases using “Seaborn Category plot”

Result:





11). 20 Most Occurred Words using “Seaborn Category plot”

Result:

