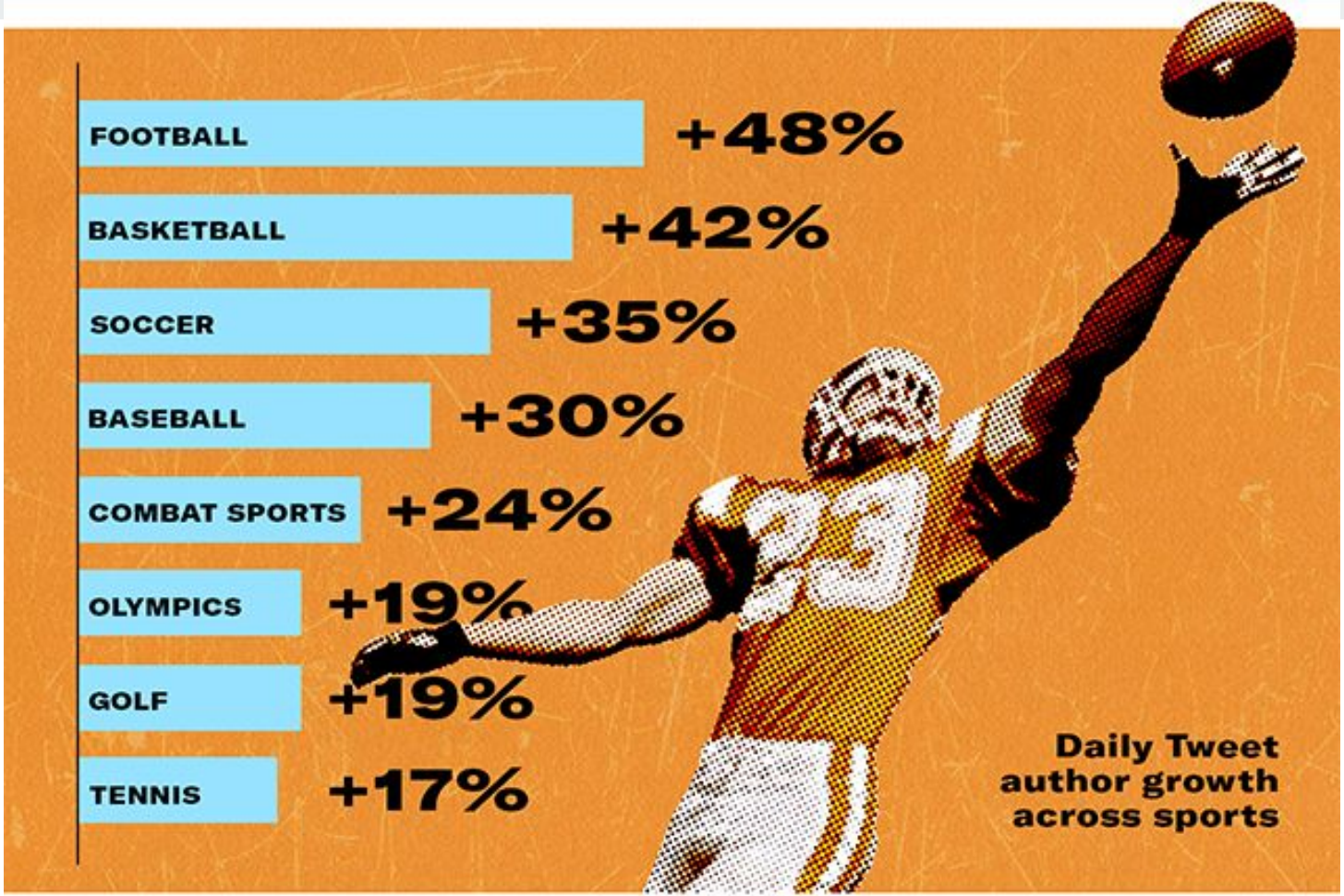
Soccer Big Data Analysis

Dhrumil Patel
Saamarth Rastogi
Soham Mane
Vishwas Chandran
Nandini Manyam
Kavish Shah

Why we chose this problem ?

Today we have this massive amount of data generated or gathered from various social media platform on daily basis of different Industries, and technical analysis is done to gain insight and to make business decisions. Sports Industry is no different, here we have chosen a few football clubs, whose data we have gathered from twitter, through which we can get the sports news/tweets using which we can analyze the stats of those clubs for getting insights. Usually these insights come handy for business investors who are investing their money into such clubs. Also it can indirectly or directly help the club teams improve their game.

Soccer has one of the highest daily average tweets.



Files:

1) Clubs.csv (Structured Data) :

   Contains data about clubs and its trophies collection.

2) User_dfs.csv (Structured Data) :

   Contains club twitter account details like followers,creation date

3) Club_tweets.json (Semi-Structured Data) :

   Contains tweets done by official club twitter handle

Goal of the project

- Work and transfer the data into such a format which is easy to visualize and gain insights from.
- Find insights and sentiments from tweets and help betting company to predict initial bet price for next season.
- Find insights and complete analysis / visualizations for club performance till last year and help news channels for pre season talks and sessions.
- Insights can help the team to improve the upcoming games.

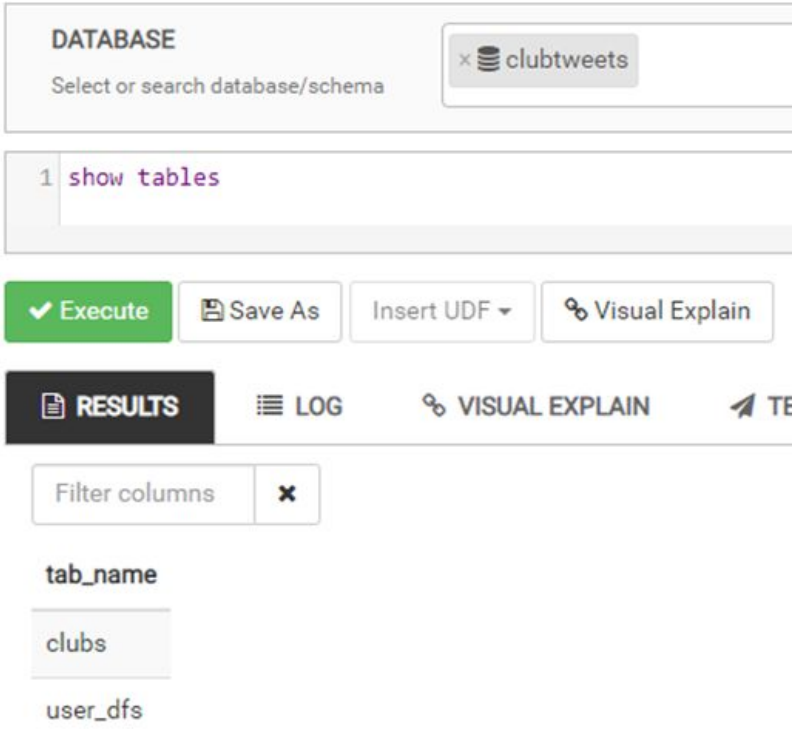
Hadoop - HIVE

- Hive is a Hadoop-based data warehouse architecture utility for processing structured data.
- Hadoop delivers tremendous scale out and fault tolerance capabilities for data storage and processing on commodity hardware.
- Hive is intended to allow for simple data summarization, ad-hoc querying, and analysis of Big Data.

Hadoop -HIVE

- Created database as **clubtweets**

- Created tables as **clubs** and **user_dfs** for uploading structured data on HIVE



Hadoop - HIVE

- Clubs table content
- Columns:

  Club, Country, Ucl, Uel, Cwc, Usc, Uic, Ic, total

DATABASE

Select or search database/schema

clubtweets

```
1 select * from clubs
```

Execute | Save As | Insert UDF | Visual Explain

RESULTS | LOG | VISUAL EXPLAIN | TEZ UI

Filter columns

clubs.club	clubs.country	clubs.ucl	clubs.uel	clubs.cwc	clubs.usc	clubs.uic	clubs.ic	clubs.total
Real Madrid	Spain	13	2	0	4	0	3	22
Milan	Italy	7	0	2	5	0	3	17
Barcelona	Spain	5	0	4	5	0	0	14
Liverpool	England	6	3	0	3	0	0	12
Juventus	Italy	2	3	1	2	1	2	11
Bayern Munich	Germany	5	1	1	1	0	2	10
Ajax	Netherlands	4	1	1	2	0	2	10
Internazionale	Italy	3	3	0	0	0	2	8

Hadoop - HIVE

- User_dfs table

```
1 select * from user_dfs
```

Execute | Save As | Insert UDF | Visual Explain

RESULTS | LOG | VISUAL EXPLAIN | TEZ UI

Filter columns

user_dfs.club name	user_dfs.contributors_enabled	user_dfs.created_at	user_dfs.default_profile	user_dfs.default_profile_imag
Real Madrid	false	2008-05-22 19:25:51+00:00	false	false

Hadoop - HIVE

- Query to find out total number of clubs in each country in descending order from clubs table.

```
1 select country, count(distinct club) numberofclub from clubs group by country order by numberofclub desc
```

Execute | Save As | Insert UDF | Visual Explain

RESULTS | LOG | VISUAL EXPLAIN | TEZ UI

Filter columns

country	numberofclub
England	13
France	12
Germany	11
Italy	11
Spain	9

Mongo DB

- **Mongo DB** is an open source **NoSQL** Database Management Program
- NoSQL is used as an alternative to traditional relational databases
- NoSQL databases are quite useful for working with large sets of distributed data.
- Mongo DB is a tool that can **manage document-oriented information, store or retrieve information**.

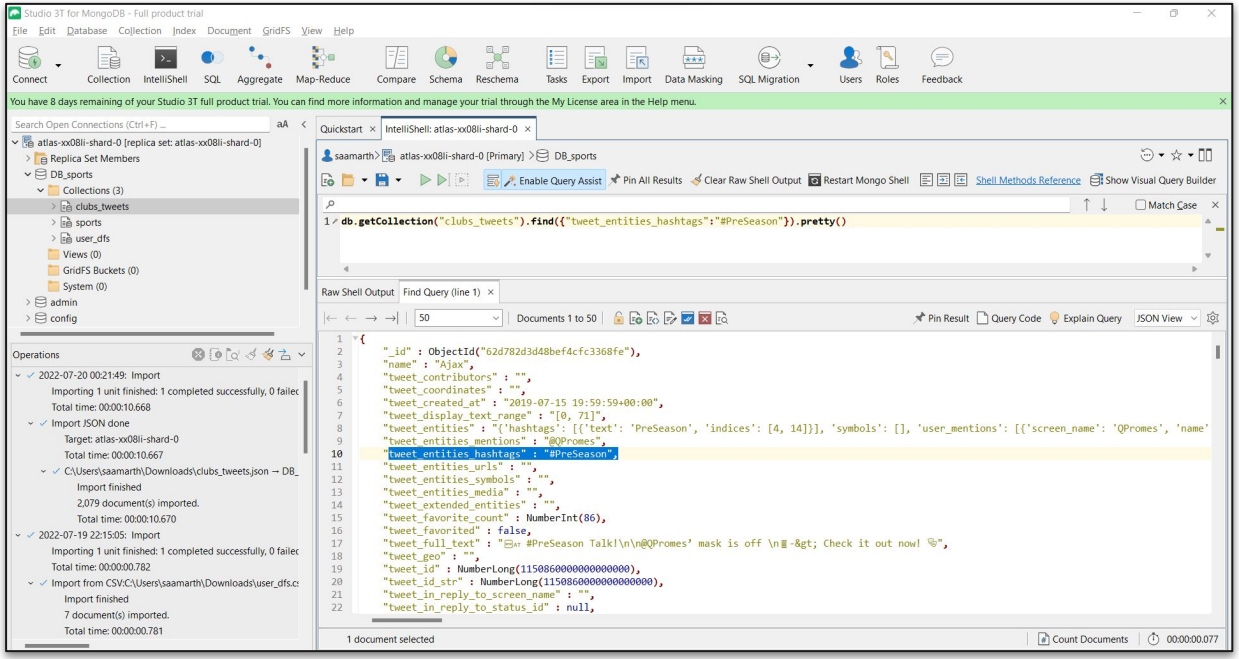
Mongo DB

- We are using The **RoboMongo(Studio 3T)**, platform as the GUI tool to **access the MongoDB server**

- **RoboMongo** allows users to easily view, access, edit MongoDB Databases from the graphical interface. It simplifies the workflow and saves time.

Mongo DB

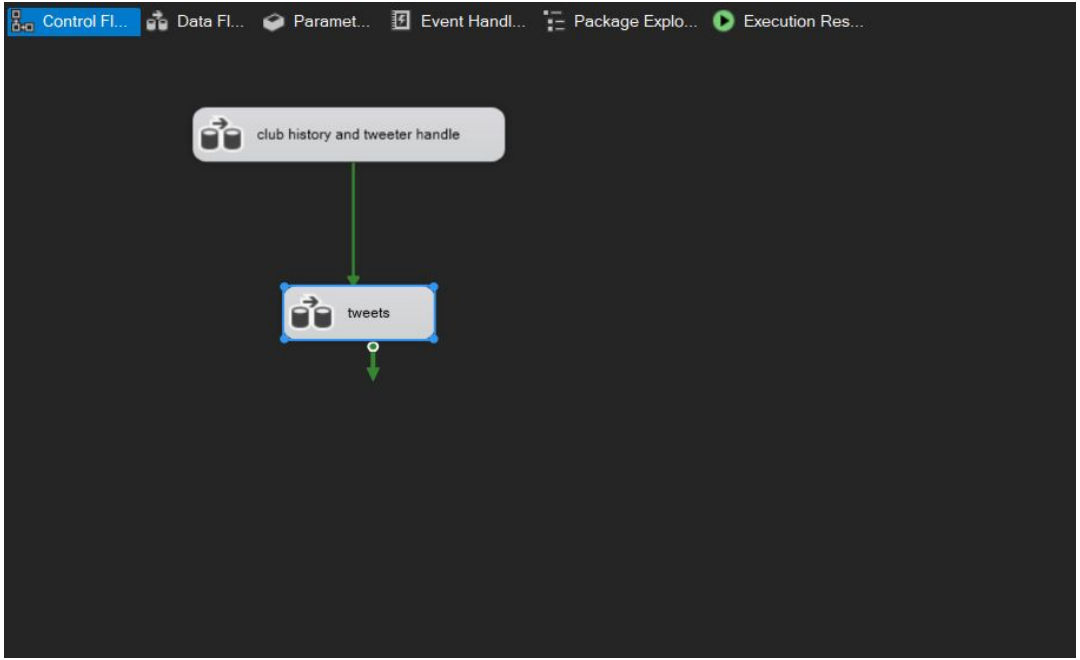
- We retrieve and store the semi-structured data related to **Club_tweets** that we have acquired from twitter onto Mongo DB.
- We then integrate MongoDB with the SSIS platform, so that the semi-structured data from Mongo DB can be merged with Relational Database from Hive, for further analysis.



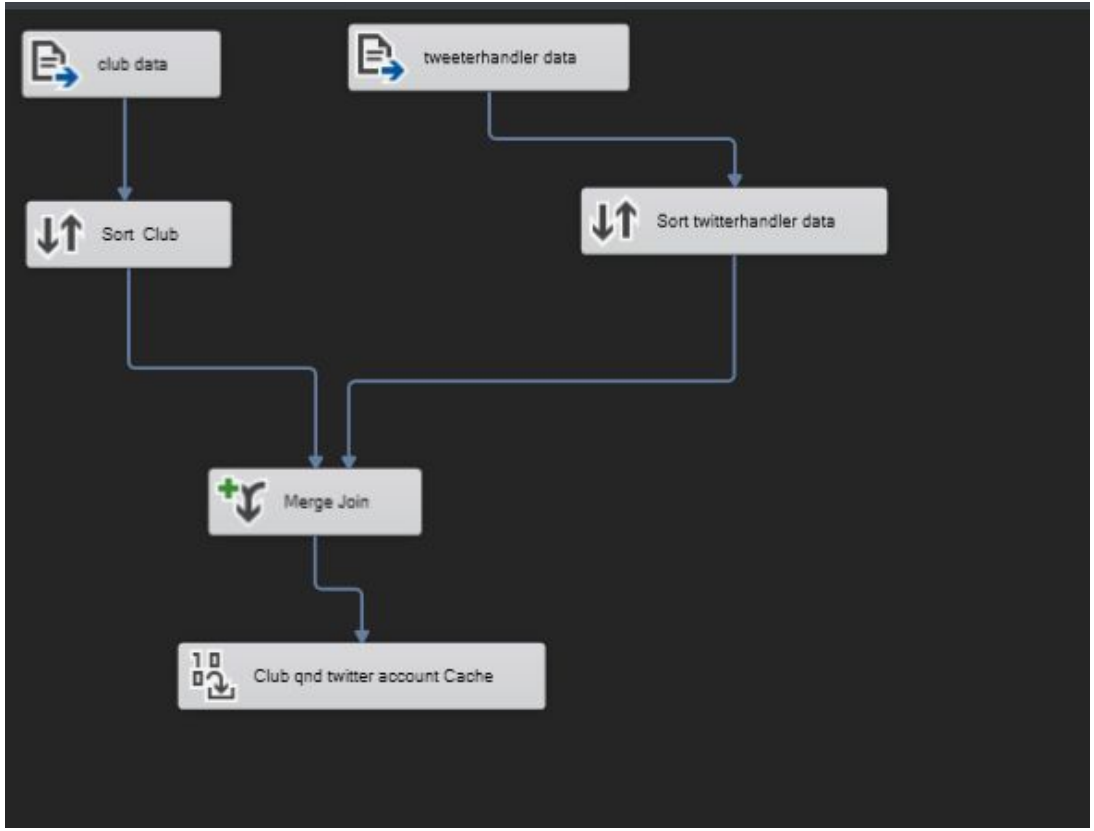
SSIS

- SSIS is etl tool and component of microsoft sql server database software.
- It is used to integrate data from various sources and various tools.
- It is easy to create pipeline, control flow and data flow using ssis compared to other softwares .

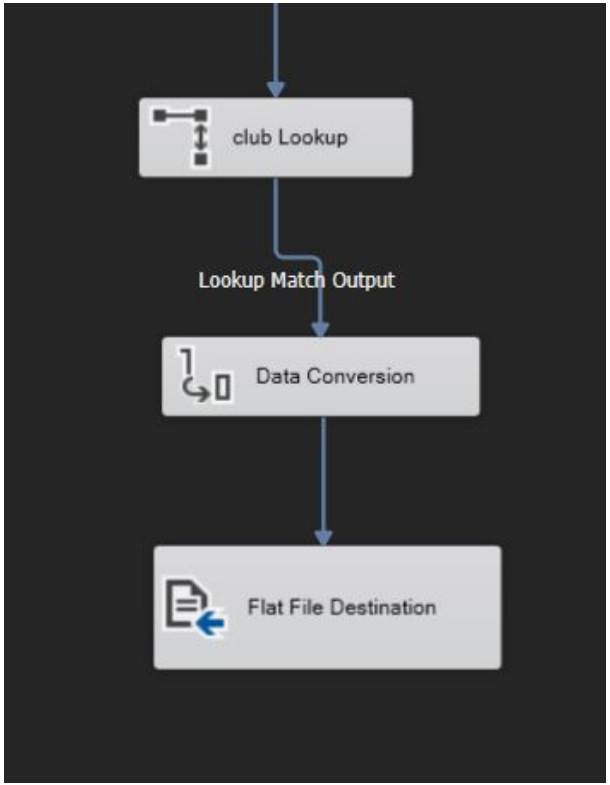
ControlFlow Diagram Of Project



DataFlow For structured data



Integration of semistructured data with structured data using lookup



Spark

- Apache Spark is an open-source, distributed processing system used for big data workloads.
- Can process multiple files at once
- Analysing data using Scala

Sorting clubs based on number of matches won

```
scala> val y = x.filter($"_c8" > 10)
y: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [_c0: string, _c1: string ... 7 more fields]

scala> y.show()
+-----------+-------+---+---+---+---+---+---+---+
|        _c0|    _c1|_c2|_c3|_c4|_c5|_c6|_c7|_c8|
+-----------+-------+---+---+---+---+---+---+---+
|Real Madrid|  Spain| 13|  2|  0|  4|  0|  3| 22|
|      Milan|  Italy|  7|  0|  2|  5|  0|  3| 17|
|  Barcelona|  Spain|  5|  0|  4|  5|  0|  0| 14|
|  Liverpool|England|  6|  3|  0|  3|  0|  0| 12|
|   Juventus|  Italy|  2|  3|  1|  2|  1|  2| 11|
+-----------+-------+---+---+---+---+---+---+---+
```

Clubs with most followers count

```
scala> club_stat.orderBy($"_c9".desc).show()
+--------------------+---------------+
|                 _c0|            _c9|
+--------------------+---------------+
|           Club name|followers_count|
| Frauen: @FCBfrauen"|      773069256|
|               Milan|        6983977|
|            Juventus|        6961616|
|         Real Madrid|       32500951|
|           Barcelona|       30165840|
|           Liverpool|       12094629|
|                Ajax|        1167488|
|       Bayern Munich|           null|
|@FCBayernEN ?? ...|           null|
|Jugend: @FCBjunio...|           null|
+--------------------+---------------+
```

Analysing medium of access

```
scala> distint_devices_sorted.orderBy($"count".desc).show(10)
+--------------------+-----+
|                 _c10|count|
+--------------------+-----+
|                null| 4103|
|  Twitter Web Client|  181|
|      Hootsuite Inc.|  125|
|Twitter Media Studio|  100|
|           TweetDeck|   77|
|  Twitter for iPhone|   37|
| Twitter for Android|   29|
|              Grabyo|   20|
|     Twitter Web App|    9|
|Twitter Ads Composer|    9|
+--------------------+-----+
only showing top 10 rows
```

Analysing top mentions in the tweet

```
scala> val top_mentions_count = a.select($"_c4").filter($"_c6" >35000).count
top_mentions_count: Long = 18

scala> val top_mentions = a.select($"_c4").filter($"_c6" > 35000)
top_mentions: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [_c4: string]

scala> top_mentions.show()
+--------------------+
|                 _c4|
+--------------------+
|                null|
|      @AntoGriezmann|
|                null|
|                null|
|      @AntoGriezmann|
|                null|
|      @AntoGriezmann|
|                null|
|            @Dembouz|
|                null|
|    @DeJongFrenkie21|
|                null|
|    @DeJongFrenkie21|
|    @DeJongFrenkie21|
|@arthurhromelo, @...|
|                null|
|         @realmadrid|
|                null|
+--------------------+
```

Visualization Using Tableau

Tableau is an excellent data visualization and business intelligence tool which can handle large volume of data and depict graphs with ease. Similarly here we have created 3 dashboard based on the 3 different datasets.

- Club Dashboard
- User_dfs Dashboard
- Club Tweets Dashboard.

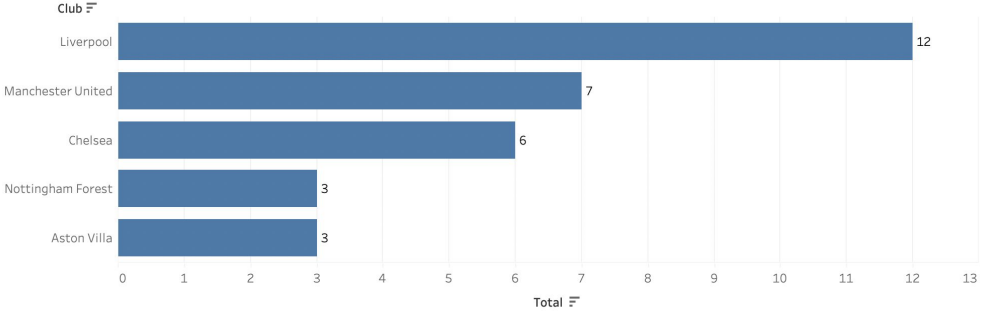
Club Dashboard

Observation

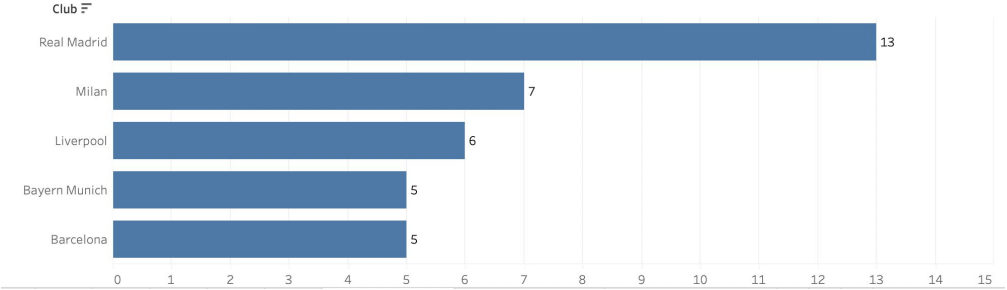
- The top 5 clubs when it comes to the region of England, are
  - Liverpool
  - Manchester United
  - Chelsea
  - Nottingham Forest
  - Aston Villa
- The Considering the UCL Championship the top 5 clubs are
  - Real Madrid
  - Milan
  - Liverpool
  - Bayern Munich
  - Barcelona

Clubs_Dashboard

Top 5 Clubs of England



Top 5 UCL Clubs



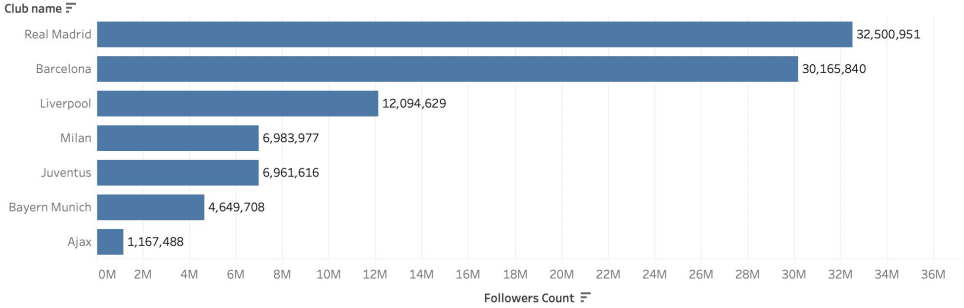
User_dfs Dashboard

Observation

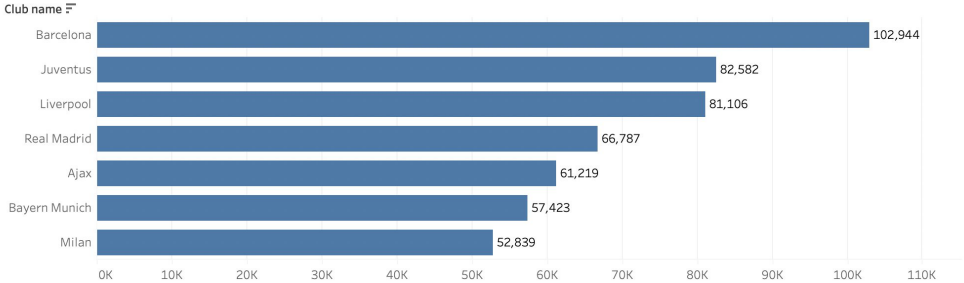
- These are ranking of the clubs based on the followers count
  - Real Madrid
  - Barcelona
  - Milan
  - Juventus
  - Bayern Munich
  - Ajax
- Ranking Based on status count
  - Barcelona
  - Juventus
  - Liverpool
  - Real madrid
  - Ajax
  - Bayern Munich
  - Milan

User_dfs_Dashboard

Ranking of Clubs Based on Followers



Ranking of Clubs Based on Statuses



Club Tweets Dashboard

Observation

- The most used hashtag is"RMliga #HalaMadrid"
- English is the most used language for tweeting.
- Ranking of clubs when it comes to retweets
    - Real Madrid
    - Barcelona
    - Liverpool
    - Juventus
    - Milan
    - Bayern Munich
    - Ajax

Club_Tweets_Dashboard

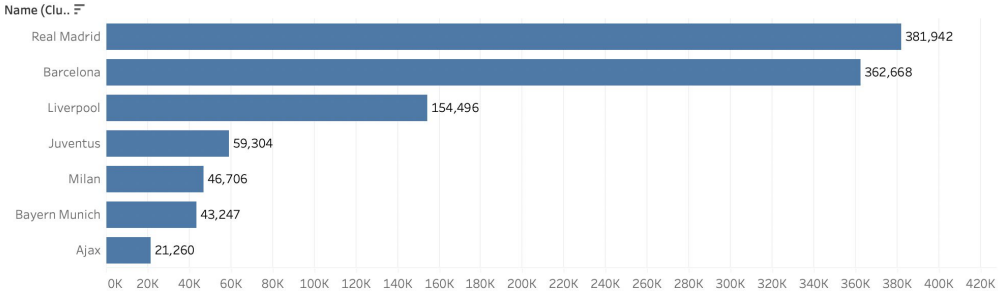
Top 10 Tweet's Hashtag.

Tweet Entities H..	
#RMLiga, #HalaMadrid	97
#RMLiga	75
#PreSeason	67
#HalaMadrid	53
#SassuoloJuve	47
#RMCity	44
#JuveFrosinone	42
#FCBBVB	33
#CARLIV	31
#F95FCB	26

Top 10 Tweet Language

Tweet..	
en	676
es	601
de	308
und	154
it	143
nl	84
pt	23
fr	18
ca	11
ht	10

Ranking of Clubs w.r.t. Retweets



Thank You