**Phase 2 Report**

**Vamsi Draksharam - 16291789**

**Divya Reddy - 16281700**

**Sai Charan Kottapalli -16247878**

**Abstract:**

1. Performing analysis on the Twitter data being collected by the means of various queries.
2. Visualization of data.

**Used Technologies:**

1. Python
2. Spark

**Tools:**

Spyder(Python 3.7), Tableau

**Queries:**

**Query 1:** List of Trending Players of National Basketball Association(NBA) and the visualization through bar graph.

df = spark.read.json("C:/Users/Vamsi Draksharam/PycharmProjects/PB-Vamsi/phase2/data2.json")

df.createOrReplaceTempView("NBA")

sqlhash = spark.sql("SELECT 'Quinn' player,count(text) as count \

FROM NBA\

WHERE 1=1\

AND (upper(text) LIKE '%COOK%' or upper(text) LIKE '%QUINN%' or upper(text) LIKE '%QUI%')\

GROUP BY player\

UNION\

SELECT 'Klay' player,count(text) as count \

FROM NBA\

WHERE 1=1\

AND (upper(text) LIKE '%KLAY%' or upper(text) LIKE '%THOMPSON%')\

GROUP BY player\

UNION\

SELECT 'Stephen' player,count(text) as count \

FROM NBA\

WHERE 1=1\

AND (upper(text) LIKE '%STEPHEN%' or text LIKE '%stephen%')\

GROUP BY player\

UNION\

SELECT 'Draymond' player,count(text) as count\

FROM NBA\

WHERE 1=1\

AND (upper(text) LIKE '%DRAYMOND%' or upper(text) LIKE '%GREEN%')\

GROUP BY player\

UNION\

SELECT 'DAMIAN' player,count(text) as count \

FROM NBA\

WHERE 1=1\

AND (upper(text) LIKE '%DAMIAN%' or text LIKE '%damian%')\

GROUP BY player\

UNION\

SELECT 'Jordan' player,count(text) as count \

FROM NBA\

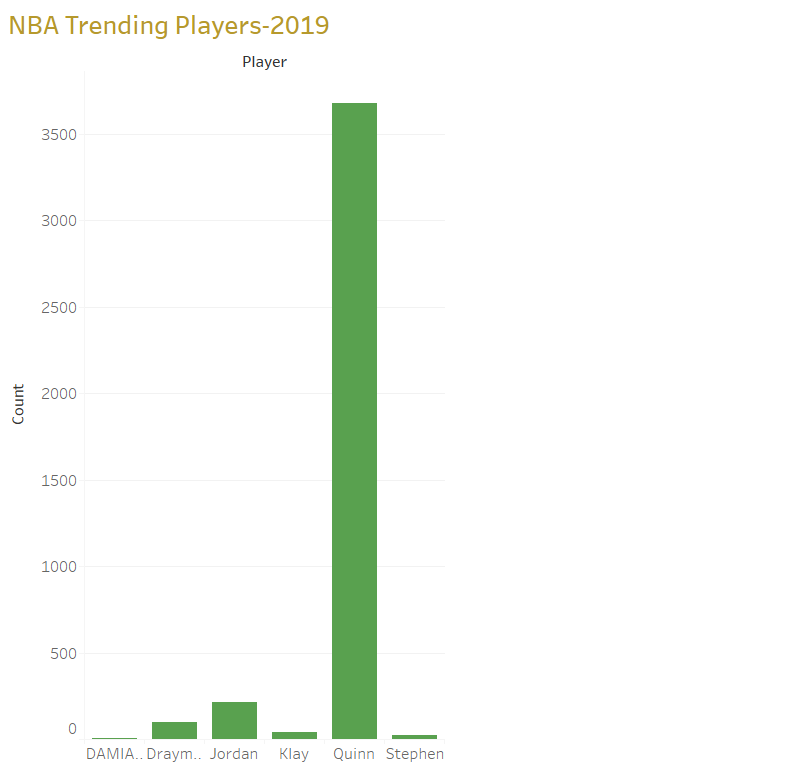
WHERE 1=1\

AND (upper(text) LIKE '%JORDAN BELL%' or upper(text) LIKE '%JORDAN%' or upper(text) LIKE '%BELL%')\

GROUP BY player")

sqlhash.show()

sqlhash.toPandas().to\_csv('1.csv')



**Query2:** NBA 2019 - Number of matches being held in various cities and the visualization using pie-chart.

df = spark.read.json("C:/Users/Vamsi Draksharam/PycharmProjects/PB-Vamsi/phase2/data2.json")

df.createOrReplaceTempView("nba")

sqldf= spark.sql("SELECT 'Staples Center' Arena,'Los Angeles' City,count(\*) FROM nba WHERE upper(text) LIKE '%LOS ANGELES%' or text like '%los angeles%' \

UNION \

SELECT 'Amway Center' Arena,'Orlando' City,count(\*) FROM nba WHERE upper(text) LIKE '%ORLANDO%' or text like '%orlando%' \

UNION \

SELECT 'TD Garden' Arena,'Boston' City,count(\*) FROM nba WHERE upper(text) LIKE '%BOSTON' or text like '%boston%' \

UNION \

SELECT 'American Airlines Center' Arena,'Dallas' City,count(\*) FROM nba WHERE upper(text) LIKE '%DALLAS%' or text like '%dallas%' \

UNION \

SELECT 'Madison Square Garden' Arena,'New York' City,count(\*) FROM nba WHERE upper(text) LIKE '%NEW YORK%' or text like '%new york%' \

UNION \

SELECT 'Veterans Memorial Coliseum' Arena,'Portland' City,count(\*) FROM nba WHERE upper(text) LIKE '%PORTLAND%' or text like '%portland%' \

UNION \

SELECT 'Wells Fargo Center' Arena,'Philadelphia' City,count(\*) FROM nba WHERE upper(text) LIKE '%PHILADELPHIA%' or text like '%philadelphia%' \

UNION \

SELECT 'Golden 1 Center' Arena,'Sacramento' City,count(\*) FROM nba WHERE upper(text) LIKE '%SACRAMENTO%' or text like '%sacramento%' \

UNION \

SELECT 'Barclays Center' Arena,'Brooklyn' City,count(\*) FROM nba WHERE upper(text) LIKE '%BROOKLYN%' or text like '%brooklyn%' \

UNION \

SELECT 'AT&T Center' Arena,'San Antonio' City,count(\*) FROM nba WHERE upper(text) LIKE '%SAN ANTONIO%' or text like '%san antonio%' \

UNION \

SELECT 'Little Caesars Arena' Arena,'Detroit' City,count(\*) FROM nba WHERE upper(text) LIKE '%DETROIT%' or text like '%detroit%' \

UNION \

SELECT 'Chase Center' Arena,'San Francisco' City,count(\*) FROM nba WHERE upper(text) LIKE '%SAN FRANCISCO%' or text like '%san francisco%' \

UNION \

SELECT 'Talking Stick Resort Arena' Arena,'Phoenix' City,count(\*) FROM nba WHERE upper(text) LIKE '%PHOENIX%' or text like '%phoenix%' \

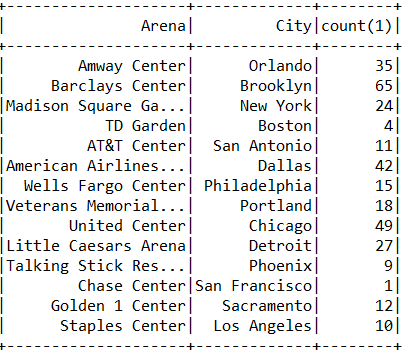
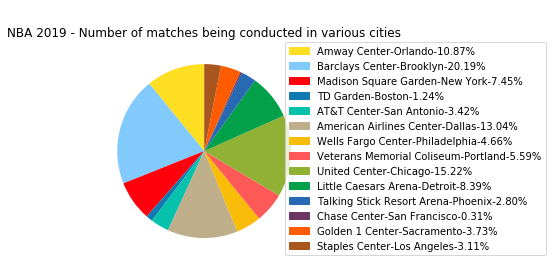
UNION \

SELECT 'United Center' Arena,'Chicago' City,count(\*) FROM nba WHERE upper(text) LIKE '%CHICAGO%' or text like '%chicago%'")

sqldf.show(150)

sqldf.toPandas().to\_csv('2.csv')

**Output:**

****

**Tableau Visualization:**

****

**Query3:** Display of Tweets from top 20 languages and the visualization using bar-graph

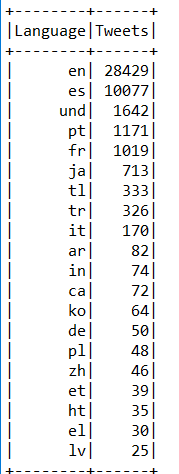
df = spark.read.json("C:/Users/Vamsi Draksharam/PycharmProjects/PB-Vamsi/phase2/data2.json")

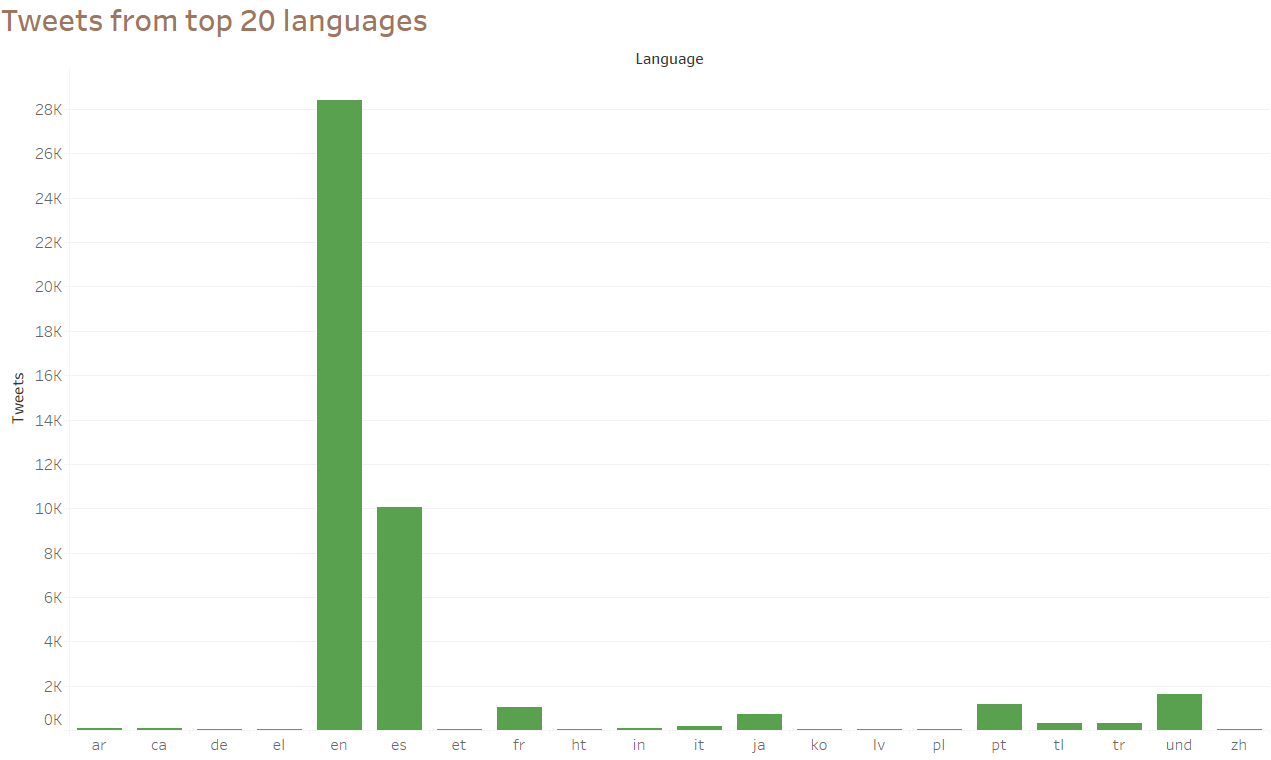
df.createOrReplaceTempView("nba")

sqldf= spark.sql("SELECT nba.lang Language,count(\*) Tweets FROM nba WHERE nba.lang is NOT NULL GROUP BY nba.lang ORDER BY 2 DESC limit 20")

sqldf.show(150)

**Output:**





**Query 4:**  Displaying the supporters and hatters of LeBron James in NBA-2019 and the visualization using Donut Pie-Chart.

df = spark.read.json("C:/Users/Vamsi Draksharam/PycharmProjects/PB-Vamsi/phase2/data2.json")

date= df.select("created\_at")

def dateMTest(dateval):

dt=datetime.datetime.strptime(dateval, '%a %b %d %H:%M:%S +0000 %Y')

return dt

d = udf(dateMTest , DateType())

df=df.withColumn("created\_date",d(date.created\_at))

df.createOrReplaceTempView("nba")

sqldf= spark.sql("SELECT id,text,created\_date FROM nba WHERE 1=1 AND (upper(text) LIKE '%LEBRON%'AND text LIKE '%nba%')")

i=0

positive=0

neutral=0

negative=0

for t in sqldf.select("text").collect():

i=i+1

analysis = TextBlob(str((t.text).encode('ascii', 'ignore')))

print(analysis.sentiment.polarity)

if (analysis.sentiment.polarity<0):

negative=negative+1

print(i," in negative")

elif(analysis.sentiment.polarity==0.0):

neutral=neutral+1

print(i," in neutral")

elif(analysis.sentiment.polarity>0):

positive=positive+1

print(i," in positive")

print("The total negative percentage is",((negative)\*100)/i)

print("The total neutral percentage is",((neutral)\*100)/i)

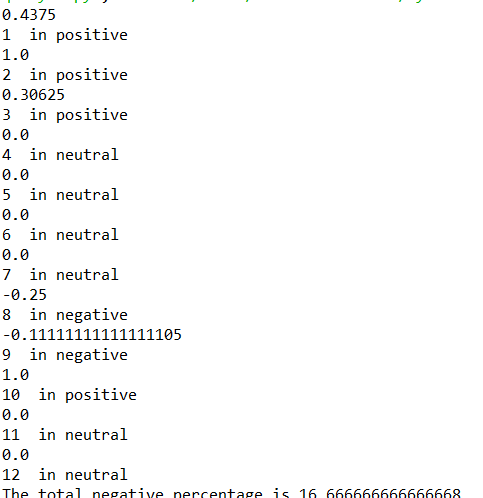
print("The total positive percentage is",((positive)\*100)/i)

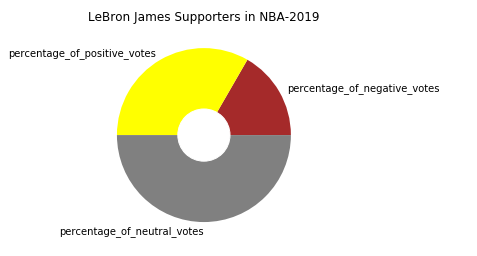
percentage\_of\_negative\_votes=((negative)\*100)/i

percentage\_of\_positive\_votes=((positive)\*100)/i

percentage\_of\_neutral\_votes=((neutral)\*100)/i

**Output:**

****

****

**Query5:** List of top NBA players and their occurrences.

df = spark.read.json("C:/Users/Vamsi Draksharam/PycharmProjects/PB-Vamsi/phase2/data2.json")

df.createOrReplaceTempView("NBA")

sqlDF = spark.sql("SELECT 'Chris Paul' as Player, count(\*) as Occurrences from nba where text like '%chris paul%' or text like '%nba%' or upper(text) like '%CHRIS PAUL%' or upper(text) like '%NBA%'\

UNION\

SELECT 'Stephen Curry' as Player, count(\*) as Occurrences from nba where text like '%curry%' or upper(text) like '%CURRY%'\

UNION\

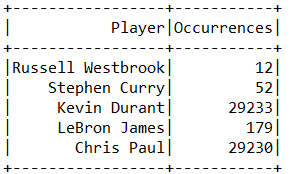
SELECT 'Kevin Durant' as Player, count(\*) as Occurrences from nba where text like '%kevin durant%' or upper(text) like '%KEVIN DURANT%' or text like '%nba%' or upper(text) like '%NBA%' UNION\

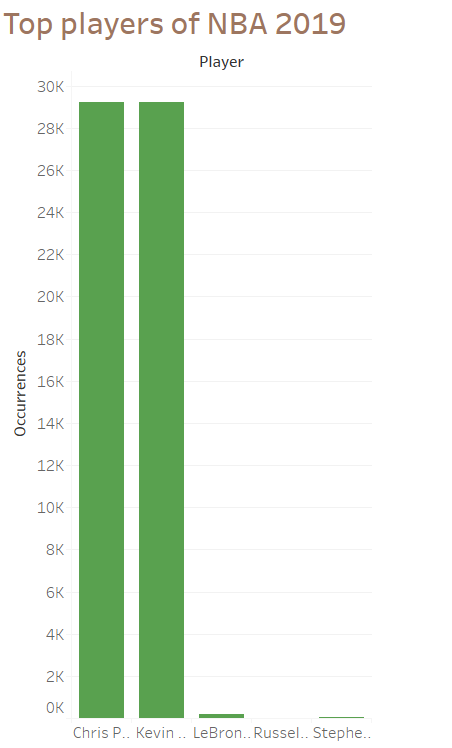
SELECT 'LeBron James' as Player, count(\*) as Occurrences from nba where text like '%lebron%' or upper(text) like '%LEBRON%' or text like '%LeBron James' UNION\

SELECT 'Russell Westbrook' as Player, count(\*) as Occurrences from nba where text like '%westbrook%' or upper(text) like '%WESTBROOK%'")

pd = sqlDF.toPandas()

**Output:**

****

****

**Query6:** Tweets from top Users and the visualization using line-graph.

df = spark.read.json("C:/Users/Vamsi Draksharam/PycharmProjects/PB-Vamsi/phase2/data2.json")

df.createOrReplaceTempView("Users")

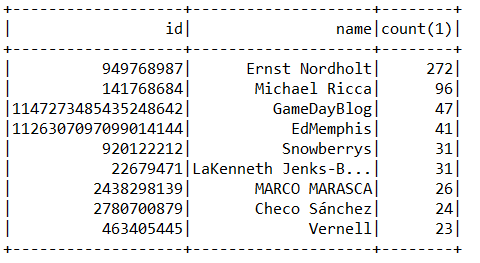
sqldf = spark.sql(

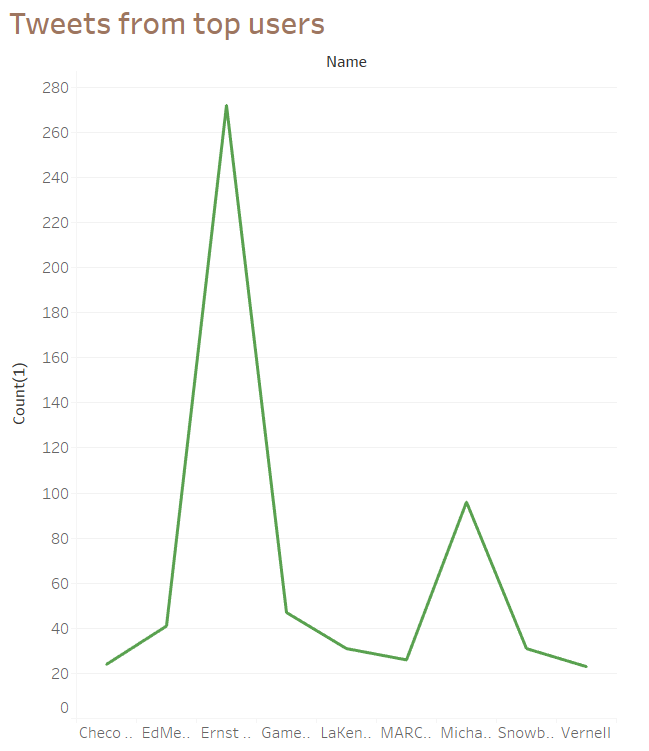
"SELECT user.id,user.name,count(\*) FROM Users"

" WHERE (user.id is not null and user.name is not null) group by user.id,user.name order by 3 desc limit 9")

sqldf.show(150)

**Output:**

****

****

**Query7:** List of Popular Awards of NBA.

df = spark.read.json("C:/Users/Vamsi Draksharam/PycharmProjects/PB-Vamsi/phase2/data2.json")

df.createOrReplaceTempView("nba")

sqldf = spark.sql("SELECT 'All-Star Game MVP' award,count(text) as count \

FROM nba\

WHERE 1=1\

AND (upper(text) LIKE '%MVP%' or upper(text) LIKE '%LEBRON JAMES%' or text like '%LeBron James%' or upper(text) LIKE '%LEBRON%' or text like '%LeBron%')\

GROUP BY award\

UNION\

SELECT 'Rookie of the Year' award,count(text) as count \

FROM nba\

WHERE 1=1\

AND (upper(text) LIKE '%ROOKIE%' or (upper(text) LIKE '%STEPHEN%' or upper(text) LIKE '%CURRY%' or text like '%stephen%')\

GROUP BY award UNION\

SELECT 'Most Valuable Player' award,count(text) as count \

FROM nba\

WHERE 1=1\

AND (upper(text) LIKE '%MOST VALUABLE PLAYER%' or upper(text) LIKE '%KEVIN DURANT%' or text like '%kevin durant%' or text like '%Durant%')\

GROUP BY award UNION\

SELECT 'Coach of the Year' award,count(text) as count \

FROM nba\

WHERE 1=1\

AND (upper(text) LIKE '%COACH OF THE YEAR%' or upper(text) LIKE '%ANTHONY DAVIS%' or text like '%anthony davis%')\

GROUP BY award UNION\

SELECT 'NBA Finals Most Valuable Player' award,count(text) as count \

FROM nba\

WHERE 1=1\

AND (upper(text) LIKE '%NBA FINALS MOST VALUABLE%' or upper(text) LIKE '%JAMES HARDEN%' or text like '%james harden%')\

GROUP BY award UNION\

SELECT 'Executive of the Year' award,count(text) as count \

FROM nba\

WHERE 1=1\

AND (upper(text) LIKE '%EXECUTIVE OF THE YEAR%' or upper(text) LIKE '%ANTETOKOUNMPO%' or text like '%antetokounmpo%')\

GROUP BY award UNION\

SELECT 'Citizenship Award' award,count(text) as count \

FROM nba\

WHERE 1=1\

AND (upper(text) LIKE '%CITIZENSHIP AWARD%' or upper(text) LIKE '%EMBIID%' or text like '%Embiid%')\

GROUP BY award UNION\

SELECT 'Defensive Player of the Year' award,count(text) as count \

FROM nba\

WHERE 1=1\

AND (upper(text) LIKE '%DEFENSIVE PLAYER%' or upper(text) LIKE '%RUSSELL WESTBROOK%' or text like '%Westbrook%')\

GROUP BY award UNION\

SELECT 'Sixth Man of the Year' award,count(text) as count \

FROM nba\

WHERE 1=1\

AND (upper(text) LIKE '%SIXTH MAN OF THE YEAR%' or upper(text) LIKE '%PAUL GEORGE%' or text like '%Paul George%')\

GROUP BY award UNION\

SELECT 'Most Improved Player' award,count(text) as count \

FROM nba\

WHERE 1=1\

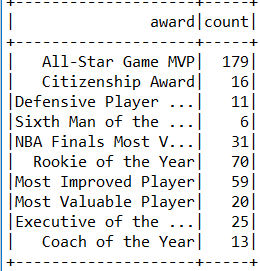
AND (upper(text) LIKE '%MOST IMPROVED PLAYER%' or upper(text) LIKE '%KAWHI LEONARD%' or text like '%Kawhi Leonard%' or text like '%Kawhi%')\

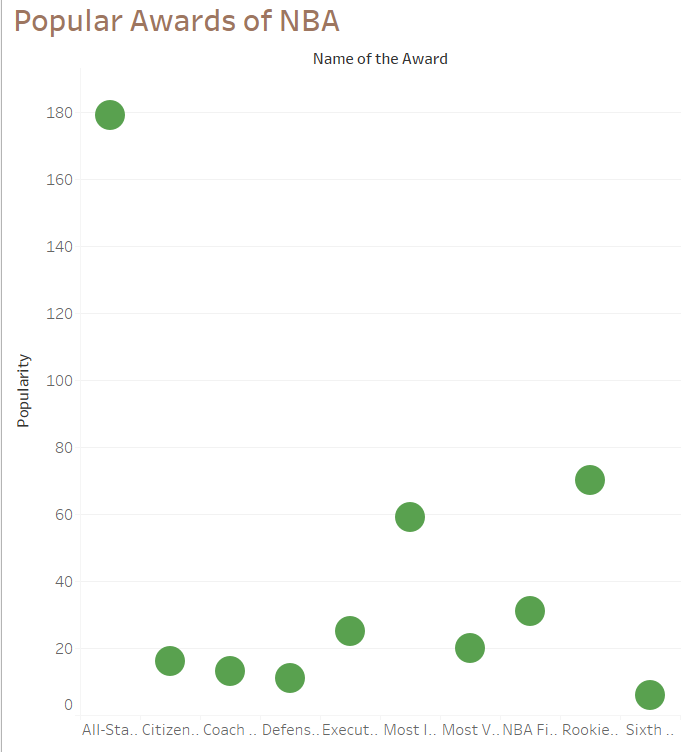
GROUP BY award")

sqldf.show(150)

sqldf.toPandas().to\_csv('7.csv')

**Output:**

****

****

**Query8:** To display the number of retweets from top pages and visualization using pie-chart.

df = spark.read.json("C:/Users/Vamsi Draksharam/PycharmProjects/PB-Vamsi/phase2/data2.json")

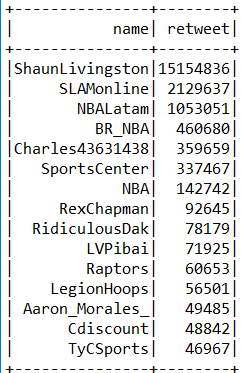
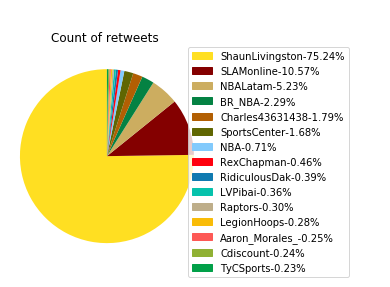
df.createOrReplaceTempView("nba")

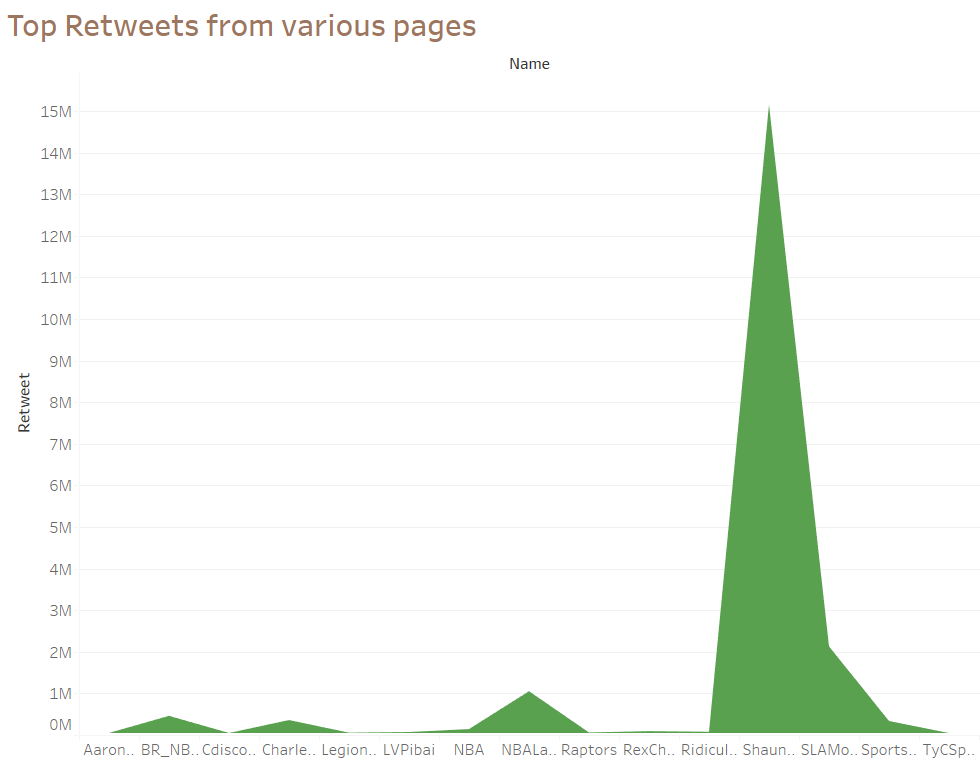
sqldf = spark.sql(

"SELECT name,SUM(cnt) as retweet FROM (SELECT quoted\_status.user.screen\_name AS name,quoted\_status.retweet\_count AS cnt FROM nba WHERE quoted\_status.retweet\_count>0)GROUP BY name ORDER BY retweet DESC LIMIT 15")

sqldf.show(150)

**Output:**

** **

****

**Query9:**  The list of Top 5 Hashags of NBA 2019 and visualization using bar-graph.

df = spark.read.json("C:/Users/Vamsi Draksharam/PycharmProjects/PB-Vamsi/phase2/data2.json")

words = df.select(

explode(

split(df.text, " ")

).alias("word")

)

def extract\_tags(word):

if word.lower().startswith("#"):

return word

else:

return "nonTag"

extract\_tags\_udf = udf(extract\_tags, StringType())

resultDF = words.withColumn("tags", extract\_tags\_udf(words.word))

resultDF.createOrReplaceTempView("hashtag\_count")

sqlhash = spark.sql("SELECT Hashtag,\

Occurrences\

FROM (SELECT upper(tags) Hashtag,\

count(\*) Occurrences\

FROM hashtag\_count\

WHERE 1=1\

AND tags!='nonTag'\

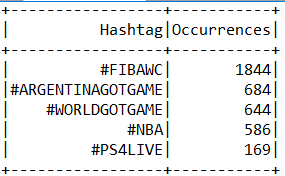
GROUP BY upper(tags)\

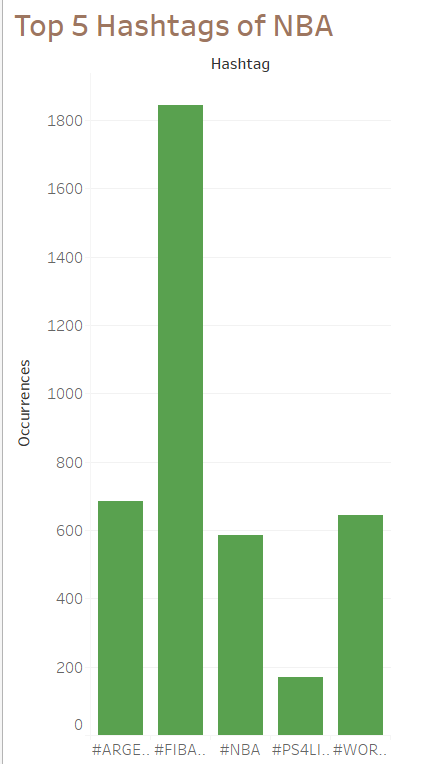
ORDER BY Occurrences desc, Hashtag asc) limit 5")

sqlhash.show(70)

sqlhash.toPandas().to\_csv('9.csv')

**Output:**





**Query10:** Player recognition in NBA and the visualization using bar-graph.

df = spark.read.json("C:/Users/Vamsi Draksharam/PycharmProjects/PB-Vamsi/phase2/data2.json")

df.createOrReplaceTempView("nba")

sqlDF = spark.sql("SELECT 'Jordan Clarkson' as Player, count(\*) as Count from nba where text like '%jordan%' and text like '%nba%'\

UNION\

SELECT 'Stephen Curry' as Player, count(\*) as Count from nba where text like '%curry%' and text like '%nba%'\

UNION\

SELECT 'LeBron James' as Player, count(\*) as Count from nba where text like '%lebron%' and text like '%nba%' UNION\

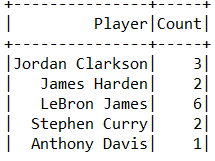
SELECT 'James Harden' as Player, count(\*) as Count from nba where text like '%harden%' and text like '%nba%' UNION\

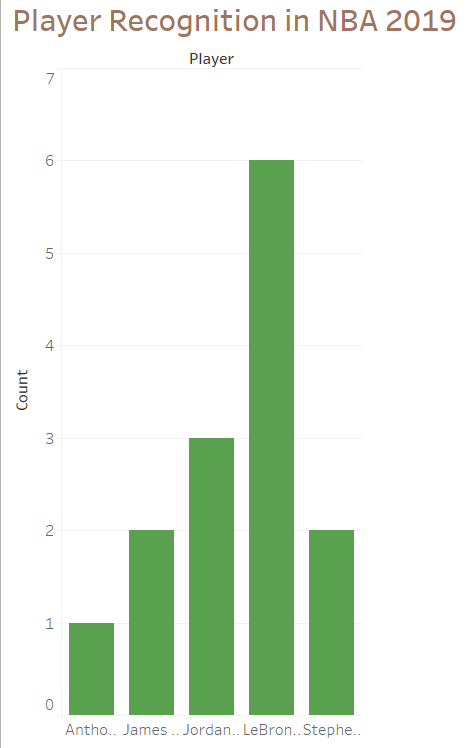
SELECT 'Anthony Davis' as Player, count(\*) as Count from nba where text like '%anthony%' and text like '%nba%'")

pd = sqlDF.toPandas()

pd.to\_csv('10.csv', index=False)

**Output:**

****

****