Portfolio Project: Football Club Twitter Analysis

Stephen Ellingson

TABLE OF CONTENTS

INTRODUCTION

Over the course of Summer 2021, I have been taking Codecademy's Data Analyst certification course, which educates on tools like Python and SQL to obtain and clean data, run analyses, gather insights, and create visualizations from various forms of data.

This final portfolio presentation, which is a completely open-ended project, displays the most important tools learned during this course. Through the next few slides, I'll be running through the last 3,250 tweets from each "big six" football club (soccer club in the U.S.) in England's top-flight division, the Premier League.

I will explain the steps I took in gathering my final insights to provide a window into my methods, and the code is available in this project’s GitHub repository. Thank you for viewing!

04

INSPECTING AND PREPARING DATA

INSPECTING AND PREPARING DATA

To start, I knew I needed a source of data that had a large amount of information. I visited Kaggle and found an amazing set of CSV files which contained scraped Tweet data from six different clubs: Arsenal, Chelsea, Manchester City, Manchester United, Liverpool, and Leicester City. After downloading these files, I hopped onto Jupyter Notebook and loaded these CSV files into separate Pandas dataframes.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

arsenal_df = pd.read_csv('Arsenal.csv')
chelsea_df = pd.read_csv('ChelseaFC.csv')
mancity_df = pd.read_csv('ManCity.csv')
manutd_df = pd.read_csv('ManUtd.csv')
liverpool_df = pd.read_csv('LFC.csv')
leicester_df = pd.read_csv('LCFC.csv')
```

INSPECTING AND PREPARING DATA (cont.)

After creating the dataframes, I inspected the content and data types of each table.

```
print(leicester_df.head(5))
print(leicester_df.dtypes)
```

```
created_at    object
full_tweet    object
tweet_type     int64
retweets       int64
likes          int64
mentions      object
dtype: object
```

created_on	full_tweet	tweet_type	retweets	likes	mentions
2021-06-29 06:52:00	Happy birthday to former Fox Ali Mauchlen! 🎂 h...	0	7	112	NaN
2021-06-28 19:07:00	2️⃣ years of @JamesJustin98 as a 🦊\n\nIs this ...	0	20	444	jamesjustin98
2021-06-28 17:24:00	Club Historian John Hutchinson's **Links With Th**...	0	9	169	NaN
2021-06-28 15:59:31	**JJ** ✨\n\n@JamesJustin98's rise to the top 📈➡️ h...	0	5	148	jamesjustin98
2021-06-28 13:54:35	"Everyone knows he's a world-class goalkeeper....	0	19	327	kschmeichel1

INSPECTING AND PREPARING DATA (cont.)

After looking at the data itself, everything seemed to be in order except for the “created_at” column, which had a data type “object”.

To convert this column from string to datetime, I used the following code.

After this, I was ready to start my analysis!

```
from datetime import datetime
clubs_df["created_at"] =
pd.to_datetime(clubs_df["created_at"])
```

08

SUMMARY STATISTICS

VIRTUAL PRESENCE

TOTAL FOLLOWERS

Man Utd: 26.3M
Arsenal: 17.8M
Liverpool: 17.7M
Chelsea: 17.6M
Man City: 10.3M
Leicester: 2.1M

AVERAGE LIKES

Chelsea: 13,978
Man Utd: 10,730
Liverpool: 9,453
Arsenal: 8,948
Man City: 3,781
Leicester: 1,418

AVERAGE RETWEETS

Chelsea: 1,500
Man Utd: 989
Liverpool: 795
Arsenal: 753
Man City: 344
Leicester: 125

LIKE COUNT QUARTILES

	Arsenal	Chelsea	Manchester City	Manchester United	Liverpool	Leicester City
Min	3	4	0	3	6	3
25%	2,349	3,677.25	608.25	3,229	2,114	98
50% (Median)	5,180	7,146	1,187	5,644.5	4,592	205
75%	11,009	15,714	2,502.75	11,744.5	11,291.75	610.75
Max	229,024	235,164	197,154	281,574	311,247	172,622

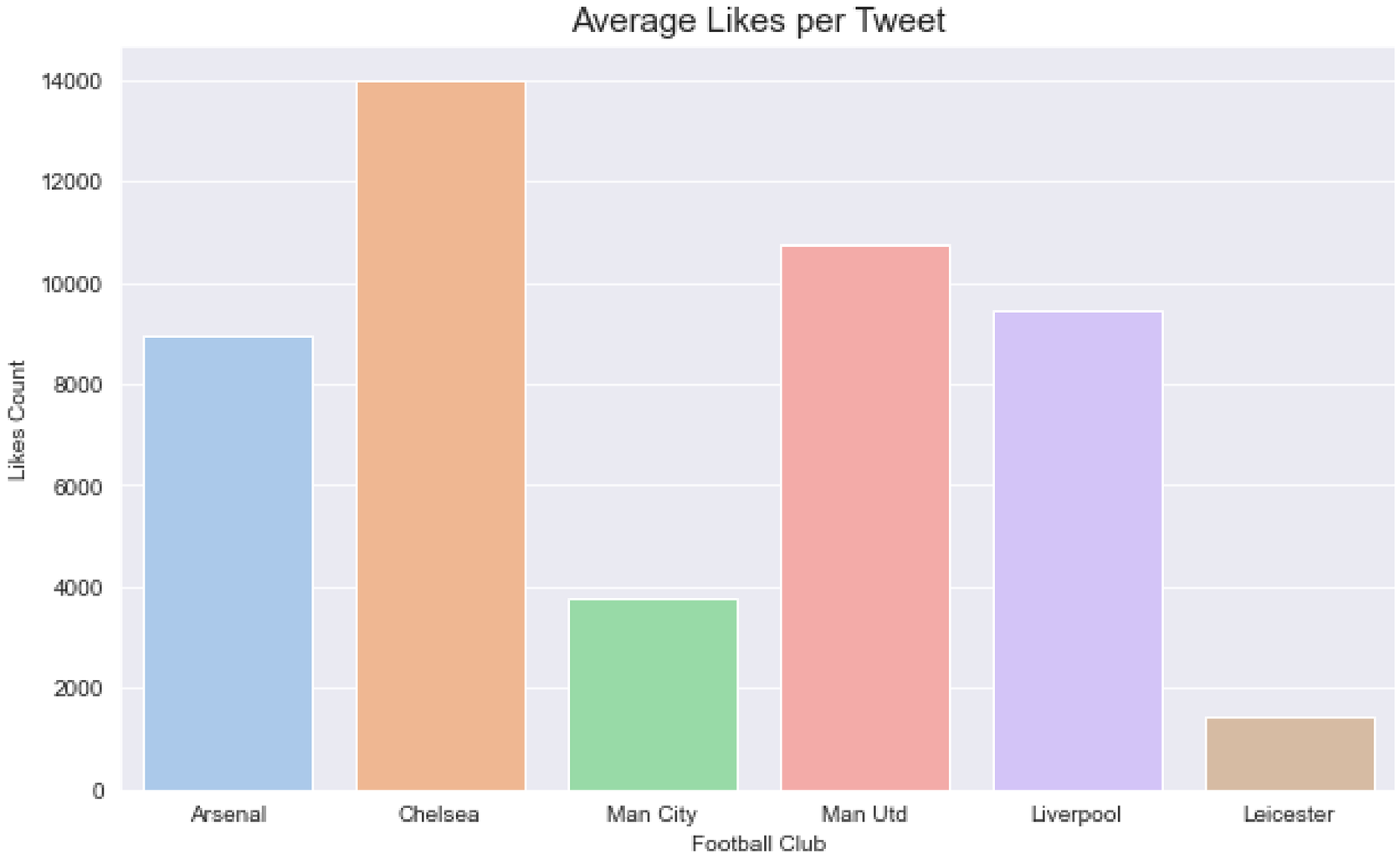
RETWEET COUNT QUARTILES

	Arsenal	Chelsea	Manchester City	Manchester United	Liverpool	Leicester City
Min	0	0	0	1	1	0
25%	149	257	39	252	158	7
50% (Median)	334	562.5	77	429	324	14
75%	782	1,346.5	183	970.75	744.75	36
Max	44,397	91,662	32,016	52,321	34,620	18,131

12

VISUALIZATIONS

VISUALIZATIONS – BARPLOT



To start off our visualizations, I thought that comparing mean likes per tweet by club would be a good indicator of social media presence.

Insights:

- Chelsea has a higher mean count than the next three clubs, which are fairly close to each other.
- Leicester City is the only club with a mean like count of less than 2,000. This can be expected as the club is a newcomer to the “big six”.

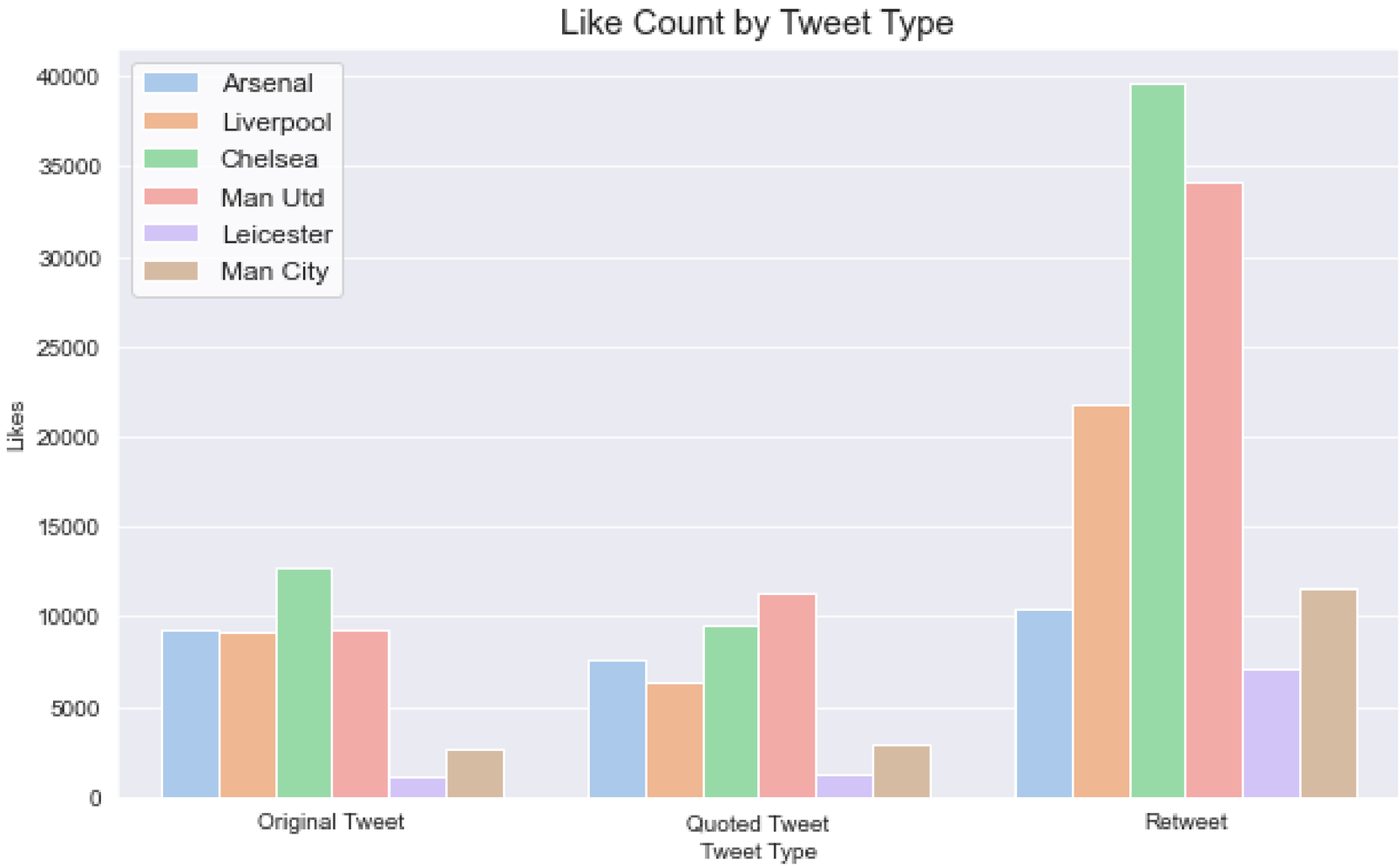
VISUALIZATIONS – BARPLOT

Moving away from average like counts, we can view total likes for each tweet type by each club. This gives us an idea of not just how many likes each club receives, but which kinds of posts those likes are directed towards.

Insights:

- Retweets are understandably much higher in counts than other tweet types.
- Interesting to note, however, that Man Utd has more total likes in quoted tweets than Chelsea.

.

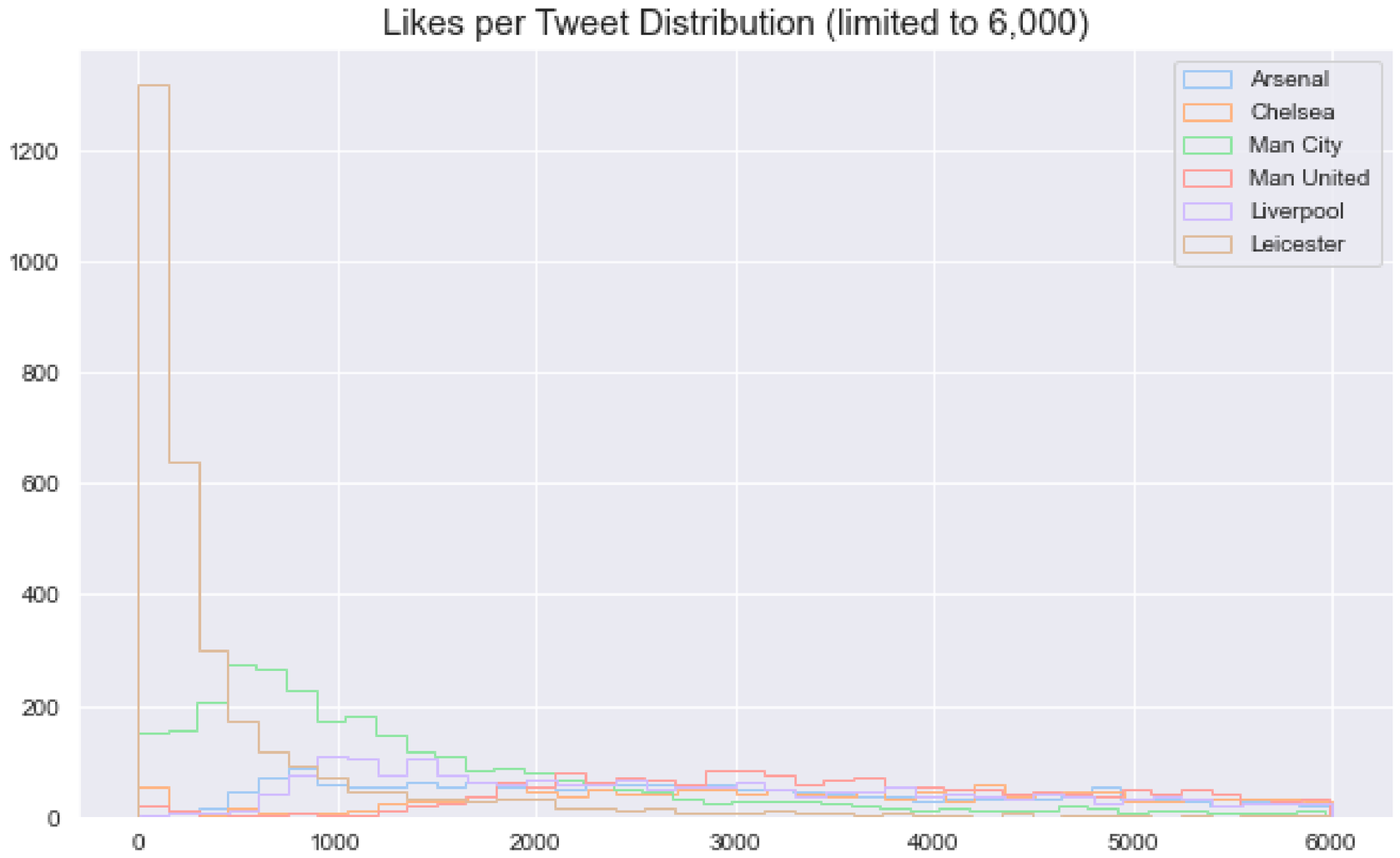


VISUALIZATIONS – HISTOGRAM

Here we can view like distributions for each club. This should give us a good idea of where each team stands compared to each other.

Insights:

- Leicester City has much more posts that received fewer likes than the other five clubs.
- A boxplot or violin plot would be better for viewing spread in a comparative manner.

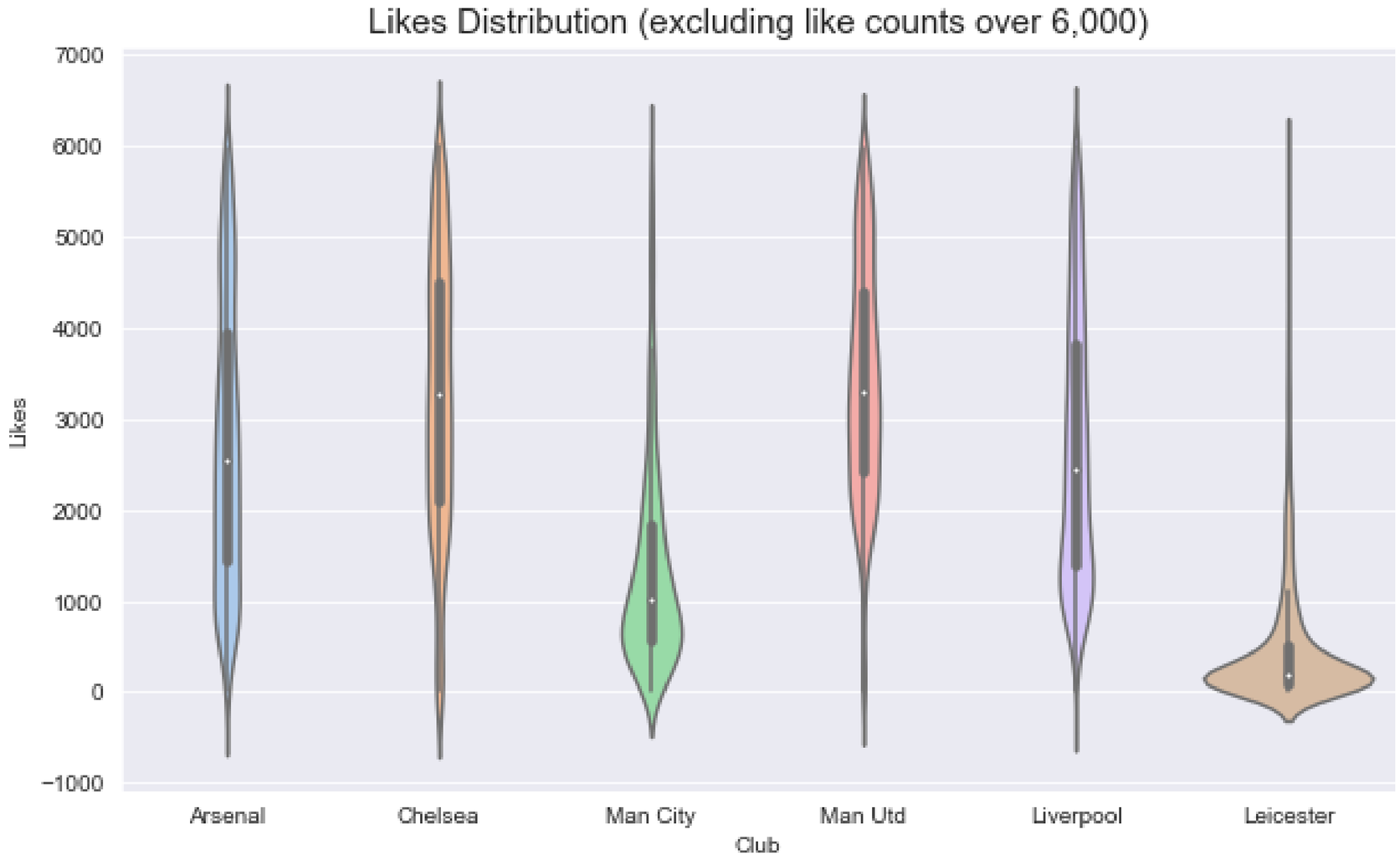


VISUALIZATIONS – VIOLIN PLOT

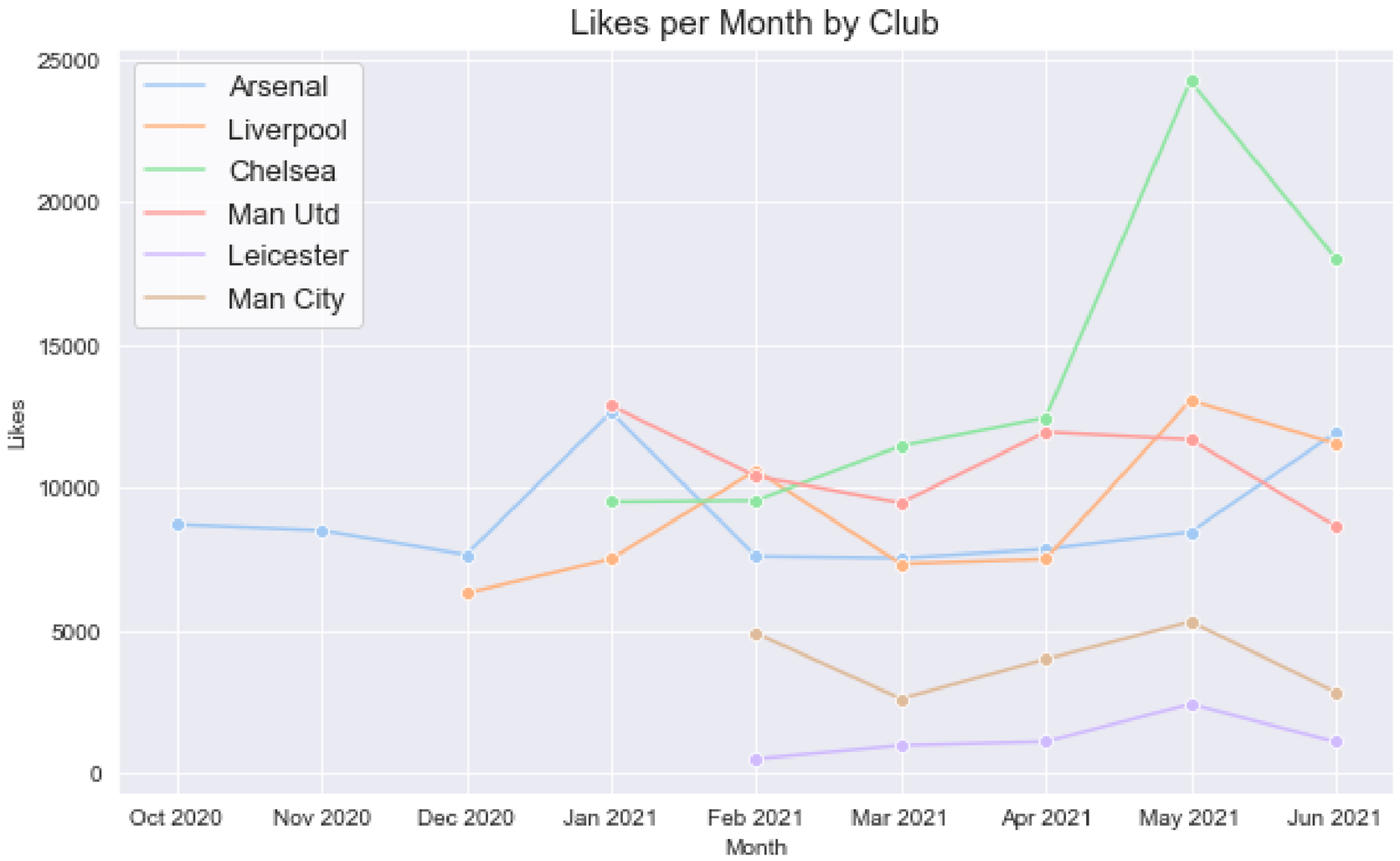
Using a violin plot has significantly helped with viewing distributions in a competitive manner. As like distributions have many outliers, counts over 6,000 have been excluded for the sake of visual clarity.

Insights:

- Leicester City and Man City have many more like counts under 1,000 than other clubs.
- Like distribution is fairly even amongst the other clubs, with no large clusters.



VISUALIZATIONS – LINE PLOT

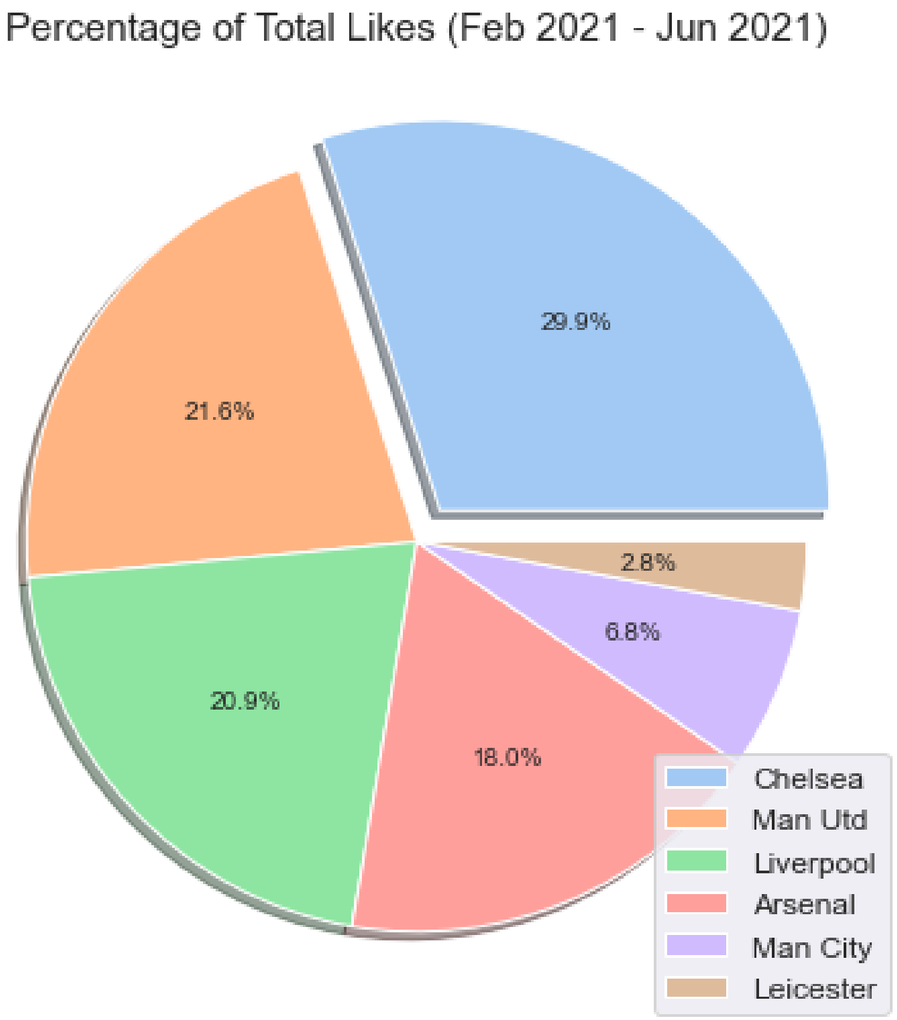


Moving on to time series analysis, we can see the like counts per month of each club. Since data was gathered based on the number of tweets, not every club has the same time frame.

Insights:

- All clubs save for Man Utd experienced a nice increase in like counts from April to May 2021. FA Cup and Champion's League finals may be responsible.
- Leicester and Man City have lower like counts but a higher tweet count per month due to the shorter time frame.

VISUALIZATIONS – PIE PLOT



After the previous slide, we can see that from February 2021 onwards, data from all six clubs is being collected. I created this pie plot to see who takes up the most likes in this time frame.

Insights:

- Chelsea's Twitter account lays claim to the highest percentage of likes with almost a third of the total. This is expected, as the team won the Champions League in May.
- Man United, Liverpool, and Arsenal are also thriving while sticking around the 20% range.
- Leicester once again comes up with considerably lower figures than the rest of the pack, despite winning the FA Cup Final during this time frame.

19

CONCLUSION

CONCLUSION

What insights did we gain from looking at the data? Here are a few:

1. Despite having less followers than Manchester United, Arsenal, and Liverpool, Chelsea's posts received more likes and retweets on average and in total than anyone else.
2. Chelsea had a spike in like counts from April 2021 to May 2021, which was due to the team winning the Champions league during that time. With this influx in mind, it is likely that had Chelsea not succeeded in this international competition, they would not be leading in average or overall activity.
3. Leicester was an obvious outlier in social activity. This is due to the team being a relative newcomer to the "big six", having won their first and only Premier League title in the 2015/16 season.
4. While Leicester and Manchester City enjoyed less activity, the time frame in which they tweeted the same number of posts as other clubs is shorter, meaning that these clubs have denser posting schedules.

I sincerely hoped you enjoyed reading through this presentation. If you have any questions, please let me know, and feel free to check out my GitHub page for more projects and reports!