Week 1

To start-off the semester, we were introduced to the overview of the ADS2002 unit, along with tasks we are expected to complete over the following months. Overall, I feel quite excited for the semester ahead and continue building on the data science concepts I undertook in ADS2001. The highlight of this week's studio was exploring the potential options for the group project. After watching the project videos and viewing the respective datasets, I selected the World Cup project as my first choice. For me, this was a relatively straightforward choice due to my strong interest in football. Since this is a project lasting over the whole semester, it will be highly beneficial if I'm analyzing a subject that's also one of my personal interests. In addition, it focuses on a different field of study compared to my previous projects, and utilizes data science techniques such as text-analysis which will be new to me. As a result, I'm looking forward to hopefully working on this project.

Week 2

This week, we were allocated into our project groups, where I will be undertaking the World Cup social media project for the semester ahead. Most of this week's studio was spent meeting our project mentor and brainstorming ideas during our group discussion, such as the objective of our research. Overall, this activity was highly beneficial since it increased my understanding of the dataset's characteristics and the aims of this project. Receiving this guidance was also important since this project is heavily reliant on text-analysis, which is an area of data science that I have extremely limited experience prior to this unit. During the group discussion, some of the ideas I suggested included analyzing the evolution of main trends throughout the tournament's progression, or creating a new variable classifying the news in various categories, such as political, player, or match-related articles. Since the nature of this project is different compared to the previous projects I worked on (which focused more on predictive modeling), I will need to spend a significant amount of time outside of class reviewing the new concepts. For next week, our group's goal is to begin the data wrangling stage where we aim to remove irrelevant symbols and characters from the raw data. This will result in simplified article titles which only contain the relevant words that describe its subject. Outside of class, I spent some time working on the twitter dataset and performing some exploratory data analysis. In this project, it will be crucial to utilize the NLTK toolkit to perform the necessary text processing to simplify the text data.

Code/Figures

```
twitter = pd.read_excel("twitter.xlsx")
twitter.head()
```

	title	content	from	locatiion	date	images	url	by	likes	shares
0	NaN	Kalvin Phillips hopeful of World Cup inclusion...	myKhel.com	Bengaluru	2022-11-08 10:40:56	[]	https://twitter.com/mykhelcom/status/158981016...	myKhel.com	0	0
1	NaN	Qatar promises a 'carbon-neutral' FIFA World C...	Republic	Mumbai, India	2022-11-08 10:58:55	[]	https://twitter.com/republic/status/1589814697...	Republic	10	1
2	NaN	Qatar's promise of 'carbon-neutral' World Cup ...	Carbon Credit Research	Worldwide	2022-11-08 11:08:09	[]	https://twitter.com/CarbonCreditRes/status/158...	Carbon Credit Research	0	0
3	NaN	Around 6,000 Argentine Fans Banned from Stadiu...	Viral Cyprus	Worldwide	2022-11-08 12:00:50	[]	https://twitter.com/viralposthq/status/1589830...	Viral Cyprus	0	0
4	NaN	Happy World Cup Final Day!. Argentina to win 2...	MrX_NFT	NaN	2022-12-18 16:21:57	[]	https://twitter.com/MrX__NFT/status/1604391505...	MrX_NFT	0	0

- Loading the dataset

	Title	Publisher	Location	Date	Likes	Shares
0	Kalvin Phillips hopeful of World Cup inclusion...	myKhel.com	Bengaluru	2022-11-08 10:40:56	0	0
1	Qatar promises a 'carbon-neutral' FIFA World C...	Republic	Mumbai, India	2022-11-08 10:58:55	10	1
2	Qatar's promise of 'carbon-neutral' World Cup ...	Carbon Credit Research	Worldwide	2022-11-08 11:08:09	0	0
3	Around 6,000 Argentine Fans Banned from Stadiu...	Viral Cyprus	Worldwide	2022-11-08 12:00:50	0	0
4	Happy World Cup Final Day!. Argentina to win 2...	MrX_NFT	NaN	2022-12-18 16:21:57	0	0
...	...	...	...	...	...	...
27503	FIFA World Cup 2022: Sometimes things don't go...	Daily News hunt 24	Mumbai	2022-11-28 21:24:41	0	0
27504	⚽ FIFA World Cup Qatar 2022 Match schedule:⚽...	K8 Official	NaN	2022-11-29 03:13:17	5	1
27505	Please am I only one that has not yet watched ...	Anas zurmi	Nigeria.	2022-11-29 03:13:26	5	0
27506	This is what the #FIFAWorldCup is all about;\n...	FIFA World Cup Stats	Global 🌍	2022-11-28 21:47:41	124	28
27507	Only if i knew that tuning into serbia vs came...	TradeStar	Mumbai, India	2022-11-28 21:47:47	0	0

27508 rows × 6 columns

```
twitter.dtypes
```

```
Title                object
Publisher            object
Location             object
Date         datetime64[ns]
Likes                 int64
Shares                int64
dtype: object
```

```
twitter.shape
```

```
(27508, 6)
```

```
twitter.describe()
```

	Likes	Shares
count	27508.000000	27508.000000
mean	186.251854	45.223971
std	4939.797974	1552.419537
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	3.000000	1.000000
max	629808.000000	236759.000000

- Revised dataset selecting relevant variables + Data exploration

Week 3

Prior to this week's studio, I dedicated time outside of class on this project to enhance my overall understanding. For instance, our group scheduled an on-campus meet-up to share our progress, solve challenges, and further discuss our approach to the project. This was important since we are analyzing three datasets, where each member is allocated different tasks. A key takeaway from this meeting was having a clearer idea of our research focus and method of analysis. One of my suggestions in the discussion which my group adopted was splitting our investigation into three main periods; pre-tournament, during the tournament focusing on each round, and post-tournament, where we aim to gather insight on the trending events throughout the tournament's progression based on online reactions. In terms of the dataset progress, I continued cleaning the data, primarily with the objective of converting the titles into a format suitable for text analysis. This includes ensuring that the relevant N/A values were dropping, and utilizing the NLTK package to simplify the titles. So far, the steps I have taken are eliminating the stopwords to prevent redundancy, and converting the title phrases into a tokenized version. This was achieved through utilizing NLTK pre-built functions which downloaded the complete list stopwords, and implementing it to our dataset. For next week, I aim to complete the data wrangling stage, which is a goal shared by our group for each of the datasets we are observing.

Code/Figures

```
twitter.isna().sum()
```

```
Title            0
Publisher        2
Location      8365
Date             0
Likes            0
Shares           0
dtype: int64
```

```
date_counts = twitter['Date'].value_counts()

date_counts
```

```
2022-11-21    2802
2022-11-20    1909
2022-11-19    1329
2022-11-18    1089
2022-11-22     998
              ...
2023-01-04       4
2023-01-08       3
2023-01-09       3
2023-01-07       2
2023-01-06       1
Name: Date, Length: 92, dtype: int64
```

- Ensuring NA values were dropped (for relevant variables) + Further data exploration

```
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')

stop_words = set(stopwords.words('english'))
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/farrelw/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /Users/farrelw/nltk_
data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /Users/farrelw/nlt
k_data...
[nltk_data]   Package wordnet is already up-to-date!
```

```
twitter['Title'] = twitter['Title'].apply

twitter
```

	Title
0	Kalvin Phillips hopeful World Cup inclusion En...
1	Qatar promises 'carbon-neutral ' FIFA World Cu...
2	Qatar 's promise 'carbon-neutral ' World Cup r...
3	Around 6,000 Argentine Fans Banned Stadiums Qa...
4	Happy World Cup Final Day ! . Argentina win 2-...
...	...
27503	FIFA World Cup 2022 : Sometimes things ' go wa...
27504	⚽ FIFA World Cup Qatar 2022 Match schedule : ...
27505	Please one yet watched even single match Qatar...
27506	# FIFAWorldCup ; 🏆 playing street football big...
27507	knew tuning serbia vs cameroon match accidenta...

- Downloading NLTK packages + 1st step of removing stopwords

Week 4

This week, my primary focus was continuing the data wrangling process of the Twitter dataset. I resumed my work on the Natural Language Toolkit platform in order to convert the text data into a suitable form for text analysis. After successfully removing the stopwords from the entries last week, I continued by lemmatizing the text as the next step. The goal of lemmatization is to reduce a word to its root form. For instance, the word “scoring” would be identified as “score”, and words, verbs, and adjectives with multiple extensions would all be converted into its base form. Lemmatization is a crucial implementation when performing text-analysis since it eliminates the redundancies caused by similar-definition words, normalizing the text which leads to reducing word complexity. Overall, this will create a more accurate form of analysis and better insights. Afterwards, I encountered some challenges in constructing the code for removing the emojis and punctuation from the text, with my current code resulting in several errors. As a result, I tried searching online tutorials and reaching out to my colleagues for help, but I’m still in the process of trying to fully solve this problem. Beyond data wrangling, I also analyzed the dataset’s characteristics further, such as examining the distribution of likes and shares among the entries. A key observation I noticed was the positively-skewed distribution of these values. For instance, over half of the dataset contained zero likes, while in contrast, the top 1 percentile featured entries ranging from tens to hundreds of thousands of likes and interactions, which might be an important consideration for our subsequent analyses. Before next week’s studio, I aim to spend some time outside of class implementing the NLTK package on the data.

Code/Figures

```
translator = str.maketrans("", "", string.punctuation)
twitter['Title'] = twitter['Title'].apply(lambda tokens: [word for word in tokens if word

def lemmatize_text(text):
    lemmatizer = WordNetLemmatizer()
    words = nltk.word_tokenize(text)
    lemmatized_words = [lemmatizer.lemmatize(word, pos='v') for word in words]
    return ' '.join(lemmatized_words)

twitter['Title'] = twitter['Title'].apply(lemmatize_text)

twitter.head()
```

	Title	Publisher	Location	Date	Likes	Shares
0	[Kalvin, Phillips, hopeful, World, Cup, inclus...	myKhel.com	Bengaluru	2022-11-08	0	0
1	[Qatar, promise, 'carbon-neutral, FIFA, World,...	Republic	Mumbai, India	2022-11-08	10	1
2	[Qatar, promise, 'carbon-neutral, World, Cup, ...	Carbon Credit Research	Worldwide	2022-11-08	0	0
3	[Around, 6,000, Argentine, Fans, Banned, Stadi...	Viral Cyprus	Worldwide	2022-11-08	0	0
4	[Happy, World, Cup, Final, Day, Argentina, win...	MrX_NFT	NaN	2022-12-18	0	0

- Removing punctuation + Implementing text lemmatization

```
count     27508.000000
mean        186.251854
std        4939.797974
min           0.000000
50%           0.000000
60%           1.000000
70%           2.000000
75%           3.000000
80%           5.000000
85%          10.000000
90%          23.000000
95%          75.000000
96%         109.720000
97%         174.790000
98%         347.000000
99%        1307.670000
99.5%      4133.905000
99.9%     43192.497000
max      629808.000000
Name: Likes, dtype: float64
```

```
count      27508.000000
mean          45.223971
std         1552.419537
min            0.000000
50%            0.000000
75%            1.000000
85%            3.000000
90%            7.000000
95%           23.000000
96%           34.000000
97%           52.000000
98%          102.000000
99%          482.930000
99.5%       1381.000000
99.9%       7717.000000
max       236759.000000
Name: Shares, dtype: float64
```

- Viewing distribution of likes and shares

Week 5 - Outside of class, I concentrated on finalizing the data wrangling stage, particularly removing the emojis, unnecessary prefixes, and miscellaneous characters such as ””. There were some challenges encountered, such as ensuring I had the complete list of emojis to be removed. This took several runs of trial and error by running my code to ensure the dataset text was in a cleaned format. Relative to our group’s expectations formed at the start of the project, the data wrangling stage has taken longer than we initially anticipated. However, at the same time, we likely underestimated the complexity of this task. From a personal perspective, this was also quite a challenging process due to my limited previous exposure in this area of data science prior to this project, specifically the required data cleaning in order to carry out text-analysis. Additionally, despite our progress, our dataset is not guaranteed to be in a fully cleaned condition since it's quite difficult to judge due to the large number of observations. However, we have proceeded with our exploration, where our group can address any potential dataset issues later on as we gain more insight. To continue our analysis, I split the ‘cleaned’ dataset into three sub-datasets, separating the entries into tweets before, during, and after the tournament. To facilitate this analysis, the entries were categorized based on date, where I was able to view the top ‘x’ daily entries according to likes and shares. Although I have not reached any meaningful insights yet, this categorization is a good place to start with. This ties with aiming to understand one of the research questions I brainstormed in a recent group discussion, aiming to examine the evolution of trending topics throughout the tournament’s progression, and also whether there are any specific periods where certain controversies surrounding the tournament are most highlighted. For next week, our group will commence with the player and team analysis.

Code/Figures

```
def remove_emojis(text):
    emoji_pattern = re.compile("["
        u"\U0001F600-\U0001F64F"
        u"\U0001F300-\U0001F5FF"
        u"\U0001F680-\U0001F6FF"
        u"\U0001F1E0-\U0001F1FF"
        u"\U00002500-\U00002BEF"
        u"\U00002702-\U000027B0"
        u"\U00002702-\U000027B0"
        u"\U000024C2-\U0001F251"
        u"\U0001f926-\U0001f937"
        u"\U00010000-\U0010ffff"
        u"\u2640-\u2642"
        u"\u2600-\u2B55"
        u"\u200d"
        u"\u23cf"
        u"\u23e9"
        u"\u231a"
        u"\ufe0f"
        u"\u3030"
                      "]+", re.UNICODE)
    return emoji_pattern.sub(r'', text)
```

	Title	Publisher	Location	Date	Likes	Shares
0	[Kalvin, Phillips, hopeful, World, Cup, inclus...	myKhel.com	Bengaluru	2022-11-08	0	0
1	[Qatar, promise, 'carbon-neutral, FIFA, World,...	Republic	Mumbai, India	2022-11-08	10	1
2	[Qatar, promise, 'carbon-neutral, World, Cup, ...	Carbon Credit Research	Worldwide	2022-11-08	0	0
3	[Around, 6,000, Argentine, Fans, Banned, Stadi...	Viral Cyprus	Worldwide	2022-11-08	0	0
4	[Happy, World, Cup, Final, Day, Argentina, win...	MrX_NFT	NaN	2022-12-18	0	0

- Code to remove emojis + Final state of ‘cleaned’ dataset

```
pre_tournament['Date'] = pd.to_datetime(pre_tournament['Date'])

pre_tournament_sorted = pre_tournament.sort_values(by=['Date', 'Likes'], ascending=[True, False])

pre_tournament_likes = pre_tournament_sorted.groupby('Date').head(3)

pre_tournament_likes
```

	Title	Publisher	Location	Date	Likes	Shares
12780	[Paul, Pogba, miss, Qatar, 2022, World, Cup, d...	Fabrizio Romano	Milano, Italia	2022-11-01	56098	4540
2763	[Official, Arsenal, defender, Takehiro, Tomiya...	afcstuff	NaN	2022-11-01	13012	670
13391	[Welcome, World, Cup, month, Qatar2022, WorldC...	Road to 2022	NaN	2022-11-01	12342	1515
12759	[Qatar, pay, travel, flight, hotel, 50, Dutch,...	Transfer News Live	twittersupplement@gmail.com	2022-11-02	2374	184
2978	[bid, host, World, Cup, 're, ask, whole, world...	Anthony DiCicco	NaN	2022-11-02	1289	149
2952	[Contrary, report, tell, PSG, prepare, offer, ...	Ben Jacobs	London & New York	2022-11-02	611	38

- Example: Examining the top liked ‘x’ entries based pre-tournament data

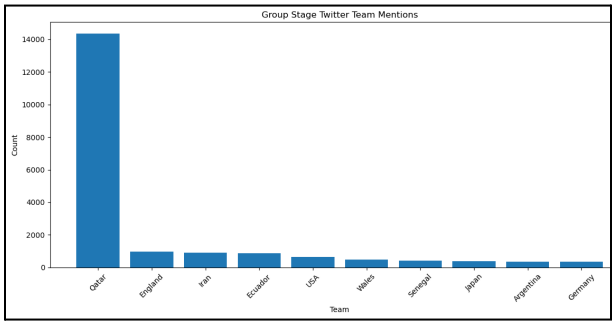
Week 6

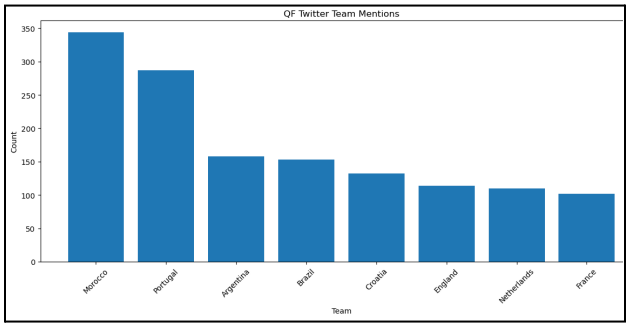
This week, our group began our player and team analysis, which followed our plans from last week during our group discussions. To facilitate our analysis, we utilized a public dataset posted in Kaggle regarding statistical player performance during the 2022 Fifa World Cup. Although our research isn't concerned about this subject, this was helpful since we were able to utilize this data to extract all player and team names participating in the World Cup, which we will apply when performing our player and team analysis. Furthermore, we brainstormed the various approaches our group could take in transforming the data into meaningful analysis. Some of the main ideas we considered included analyzing the types of content generating the most interest in terms of interactions, along with sentimental analysis such as examining the content associated with positive or negative reactions, and evaluating the level of engagements it generates. In terms of personal progress, I performed some team analysis, exploring the trending teams during the three main periods surrounding the world cup - before, during, and after. In terms of the tournaments, I also split the data into the group stage, and categorized each knockout round to only include the teams still participating in the tournament. As a result, I visualized this information by constructing bar graphs displaying the amount of tweets containing a given team's name. However, examining the quantity of mentions is simply just the beginning of the investigation. In order to convert this into meaningful analysis, it would be beneficial to take a deeper perspective such as analyzing the level of engagement attributed to each team in different stages, as well as possibly performing sentiment analysis. With this in mind, my action plan ahead of next week's meeting will be taking a step towards accomplishing these outcomes.

Code/Figures

```
unique_teams = team['team'].unique()
unique_teams = [team.replace('IR Iran', 'Iran').replace('Korea Republic', 'Korea')
unique_teams
```

- Extracting team names





- Graph analyzing amount of team mentions
    - Figure 1: Group Stage total
    - Figure 2: Quarter Finals total - only including non-eliminated teams

Reflection #1

Overall, this has been an interesting project to work on. At the start of this semester, I was quite satisfied with receiving this project topic since I have a strong interest in soccer, especially the World Cup which took place last year. As a result, this helped with my engagement when undertaking my tasks in this project. Additionally, my background knowledge can also be helpful in understanding the context behind the data, where I hypothesize that this will come into play more as our group reaches the later stages of our project with our generated findings and key ideas.

However, my experience working in this project has definitely not been simple, as my group and I have encountered several challenges throughout the last few weeks. As mentioned before, this nature of text-analysis and clustering performed in this project is relatively new to me since my previous projects primarily involved regression models utilized to construct predictive models as a forecasting tool. As a result, it was important for me to do some background research outside of class to better understand the data science concepts involved in a project of this nature. Additionally, cleaning the dataset was particularly time-consuming as there were a significant amount of steps required. For instance, we had to carefully plan the various elements of the text to eliminate (symbols, emojis), along with the text attributes (tokenization, lemmatization) in order to convert the text to a suitable form for text-analysis, which was definitely easier said than done for me. Furthermore, as a group, we aren't fully sure on how to categorize the dataset via code since the entries have a variety of subjects such as player news, team news, match interactions, and general reactions to the tournament.

In terms of my personal contribution, I am responsible for analyzing the twitter dataset. So far, my progress includes performing basic dataset exploration variable analysis, data cleaning, and began analyzing team data by utilizing the customized dataset I have created based on the tournament stages. In terms of our group cooperation, I believe that there are no significant issues with our general collaboration or with any specific members of the group. However, one area that we could possibly work on is our communication outside of class. For instance, at the start of the semester, we planned a group meeting every Tuesday to check-in with each other and update our progress, but over the past few weeks our meetings have been quite inconsistent in terms of participation. At the same, I believe that this will improve in the coming weeks as our workload becomes more intensive during the later stages of the project. Overall, I look forward to continuing to work on this project over the next half of the semester, where I'm eager to progress through the project and hopefully generate some interesting findings.

Week 7

In this week's studio, our group continued our progress in the social media project by performing further analysis on our respective datasets. In terms of my exploration in the Twitter dataset, I constructed a graph displaying the top players mentioned in the dataset, highlighting the most popular players during the World Cup. Overall, this discovery connects to the idea of analyzing the trending topics during the tournament. To achieve this, I collaborated with my teammates where we cross-checked on our progress so far. For instance, I shared how I performed my team analysis the previous week and the code I implemented to split the tournament into different stages, whereas I received inspiration from my teammates on how to gauge the player mentions. This required several steps, such as ensuring that the accents were removed from the strings which represented a given player's surname to include all authentic references. This was important when dealing with names containing an accent, as social media users in the dataset could refer to their surnames both with or without the accent, with Mbappe's name being a prominent example of this. In addition, I also constructed a word cloud visualizing the most frequently occuring words. However, further analysis is required to truly understand the highlights of the tournament, since the findings obtained so far are relatively surface-level. With the mock group presentations coming up next week, hopefully our group will have a better vision of how we are going to approach the final month of our research in this project.

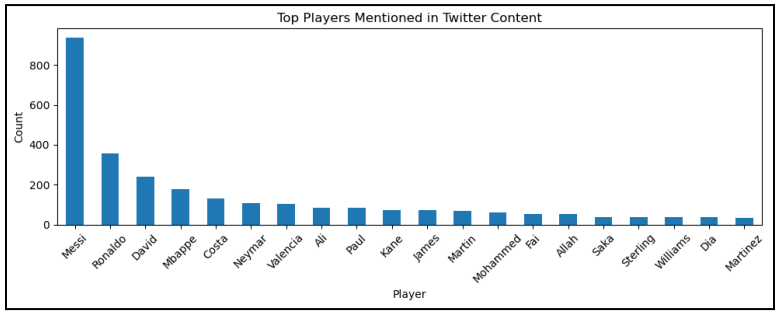
Code/Figures

```
def remove_accents(input_str):
    nfkd_form = unicodedata.normalize('NFKD', input_str)
    return ''.join([c for c in nfkd_form if not unicodedata.combining(c)])

players['player'] = players['player'].apply(remove_accents)
```

- Code to remove accents from player surname

```
def extract_surname(full_name):
    names = full_name.split()
    return names[-1]

players['Name'] = players['player'].apply(extract_surname)

player_counts = {}
for player in players['Name']:
    count = twitter['content'].str.count(player).sum()
    player_counts[player] = count
```

- Code to count player surname mentions



- Most mentioned players + Twitter dataset word cloud

Week 8

This week, I made some changes on the finalized Twitter dataset, making some updates to the data wrangling. In our recent group meeting, a prevalent topic of discussion was potentially exploring the role of hashtags and their impact on factors such as engagement rate. This particularly applies to the Twitter dataset I'm working on, since hashtags are more commonly used in this platform compared to other various social media apps such as Facebook. As a result, I revised the construction of the main dataset, where I extracted the hashtags from entries which contained them, putting this in a new, separate column, displaying all hashtags utilized in a given tweet. During this week's studio, we conducted our mock presentations where our group presented our progress so far, along with the preliminary findings of our project. This included introducing the context and background to our project, explaining how we implemented data cleaning and preprocessing, and covering our exploratory data analysis such as the word clouds, top performing users, and the most mentioned players and teams. Overall, this activity was highly useful as our group received valuable feedback, which gives us a better understanding of our priorities we need to fulfill in the coming weeks. This includes narrowing our set of research questions and improving the cohesion of our analysis. For example, we'll need to better explain how the findings in each social media platform are linked to each other and their connections to the research question. Afterwards, I spent some time outside of class to perform sentiment analysis, where sentiment scores were obtained for each tweet to quantify their emotional tone on a scale of -1 (negative) to 1 (positive).

Code/Figures

```
def extract_hashtags(text):
    hashtags = [tag for tag in text.split() if tag.startswith('#')]
    return ' '.join([word for word in text.split() if not word.startswith('#')]), ' '.join(hashtags)

def remove_links(text):
    # Use regular expression to remove URLs
    return re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILINE)

twitter['content'] = twitter['content'].apply(remove_characters)
twitter['content'] = twitter['content'].apply(remove_words)
twitter['content'] = twitter['content'].apply(remove_standalone_s)
twitter['content'] = twitter['content'].apply(merge_hashtags)
twitter['content'] = twitter['content'].apply(remove_links)
```

```
def get_sentiment(text):
    analysis = TextBlob(text)
    return analysis.sentiment.polarity

twitter['Sentiment_Score'] = twitter['content'].apply(get_sentiment)
```

- Code to extract hashtags and performing sentiment analysis

	content	Publisher	Date	Likes	Shares	Hashtag	Sentiment_Score
18924	Paul Pogba miss Qatar due new injury confirm a...	World Cup 2022	2022-11-01	12	1	#Pogba	-0.092013
2774	Qatar face renew pressure migrant worker LGBT ...	Colossus Diplomacy	2022-11-01	0	0		0.095238
2775	PUTERA_Miguel Daniel need help work ambitious ...	Blackburn Roverseasson	2022-11-01	0	0		0.233333
2776	England boss Gareth Southgate say Qatar worker...	SPORTS CIRCUS INT.	2022-11-01	0	0		0.000000
2777	Mikel Arteta Thomas Partey wait happen happy h...	King David	2022-11-01	0	0		0.400000

- Revised dataset

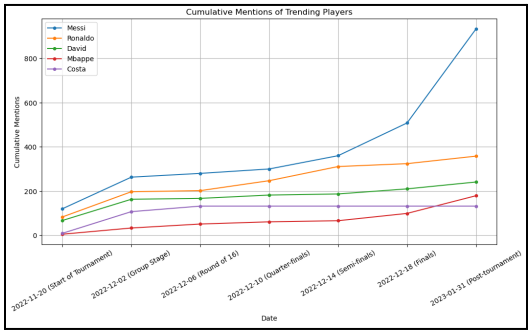
Week 9

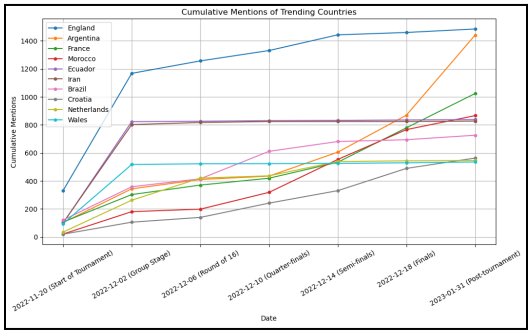
This week, our group held a lot of active discussions during the studio, where it was important to reach a final decision on our project's approach. Acting on the feedback received from last week's presentation, our group narrowed our potential research questions and selected the following subjects that we felt were most important:

1. "What type of content specifically generates the most interest in terms of engagement? Is engagement dependent on the type of user?"
2. "Which hashtags are the most used in posts? Do hashtags result in higher engagement?"
3. "Does content with positive sentiment generate more engagement than those with negative?"
4. "How do the trending topics evolve throughout the tournament's progression?

With our finalized research questions constructed, it is evident that a lot of progress is required in the coming weeks since our group is currently a bit behind schedule. Since I created some player and team analysis visualizations a few weeks ago, I decided to focus on this research question first, expanding on my preliminary findings. Instead of having separate graphs for each stage, I decided to combine them into a time-series plot. This required a lot of time and coding effort, but it was effective as the trends are more clearly visualized, where we are able to see when trending players and teams generate the most interest. A key finding from these following graphs is the strong correlation between a team's performance and their chances of trending, with the majority of trending teams being nations that progressed deep into the tournament. This trend was also present among the players, with interest surrounding Messi significantly rising during the final stages of the World Cup. I also made some progress on RQ #3 by categorizing the average sentiment score, grouping them by each team.

Code/Figures





- Time series graph visualizing cumulative player and team mentions throughout the World Cup

```
average_sentiment_scores = {'Country': [], 'Average_Sentiment_Score': []}

for country in unique_teams:
    country_data = twitter[twitter['content'].str.contains(country, case=False)]

    average_sentiment_score = country_data['Sentiment_Score'].mean()

    average_sentiment_scores['Country'].append(country)
    average_sentiment_scores['Average_Sentiment_Score'].append(average_sentiment_score)

avg_sentiment = pd.DataFrame(average_sentiment_scores)
```

Country	Average_Sentiment_Score
Serbia	0.258443
Argentina	0.222221
Brazil	0.211419
Morocco	0.199563
Japan	0.179274

- Grouping average sentiment score by country

Week 10

This week, I continued to spend time tackling the research questions. Recently, our group has attempted to speed up our progress by communicating outside of class more often and meeting up more frequently, such as a group meeting we held this Tuesday. Overall, I mainly spent this week on hashtag analysis, where I created a word cloud to visualize the most utilized hashtags. I also compared this to the word cloud created by my teammates, which led to some interesting findings. For instance, the Instagram word cloud primarily focused on the popular players and teams. Although this is also present in the Twitter dataset, it's interesting to note the emergence of some political-related hashtags, which is more prevalent in Twitter compared to Facebook. This might relate to the differences between both social media platforms and their intended purposes for users, which could be a possible talking point. To quantify the effect of hashtags on engagement, I filtered the dataset to include users with both hashtag and non-hashtag entries, and calculated the mean likes for both categories. It revealed that tweets containing hashtags did average higher levels of engagement in the platform. Meanwhile, I approached the first research question by plotting the users with the highest cumulative number of likes. However, I'm still unsure on how to proceed, specifically on how to categorize the content from the dataset.

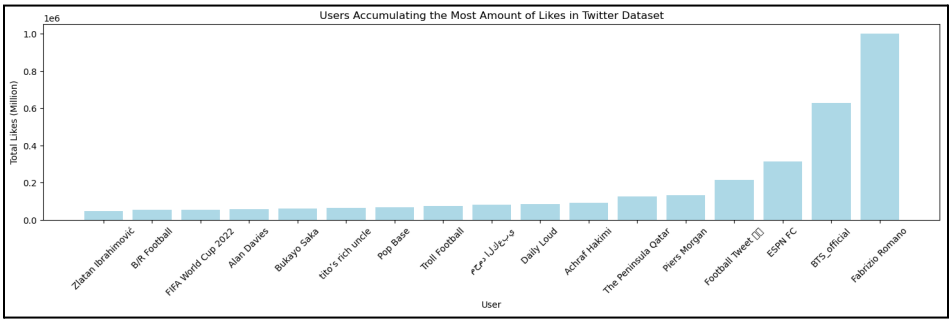
Code/Figures


```
wordcloud = WordCloud(
    width=600,
    height=400,
    random_state=1,
)
data = twitter['Hashtag']
text_data = ' '.join(data)
wordcloud.generate(text_data)
wordcloud.to_image()

all_words_lem = ' '.join([word for word in twitter['Hashtag']])

wordcloud = WordCloud(width=600,
                      height=400,
                      random_state=1,
                      max_font_size=100).generate(all_words_lem)
plt.figure(figsize=(6, 4))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
```

- Code + Visualization of hashtag word cloud

```
hashtag_likes = twitter[twitter['Hashtag'].str.contains('#', regex=False)].groupby('Publisher')['Likes'].mean()
no_hashtag_likes = twitter[twitter['Hashtag'].eq('')].groupby('Publisher')['Likes'].mean()

combined_df = pd.concat([hashtag_likes, no_hashtag_likes], axis=1)
combined_df.columns = ['hashtag_likes', 'no_hashtag_likes']
```

- Hashtag analysis code



- Graph plotting the cumulative likes of top users in Twitter dataset

Week 11

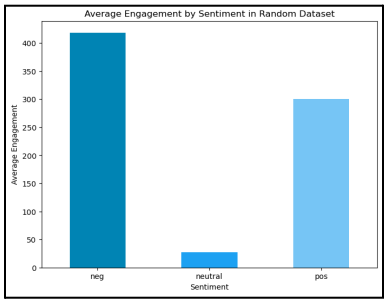
With research questions #2 and #4 answered, this week's focus was to finish exploring the first and third research question. After struggling to come up with an approach for the RQ #1, I came up with an idea which I shared with my group during our meetings. Instead of adopting a computational approach, I decided to simply analyze the top user: Fabrizio Romano, a sports journalist who frequently tweeted during the World Cup to provide news and information. Due to his genre of tweets covering a variety of topics throughout the World Cup, I analyzed his most-engaged content. Interestingly, his most-liked tweets reveal a pattern, where it centers on news regarding the most famous players such as Messi and Ronaldo. Furthermore, these posts were uploaded after the ending of significant games during the tournament, such as the World Cup final, and were very emotional in their content, thus generating huge online interest. As a result, I arrived at the conclusion that all these factors combined with Romano's influence on Twitter contribute to this type of content generating the most engagement. To answer RQ#3, I reached out to my teammates for assistance, where the methodology we came up with was to utilize an SVM model to classify the content based on sentiment, where this model was run on a randomly generated dataset containing 100 entries to obtain a predicted average amount of engagement for each sentiment category. With all the research questions answered, we aim to spend the weekend finalizing the slides and rehearsing for next week's final presentation.

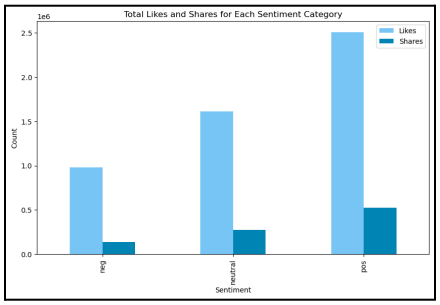
Code/Figures

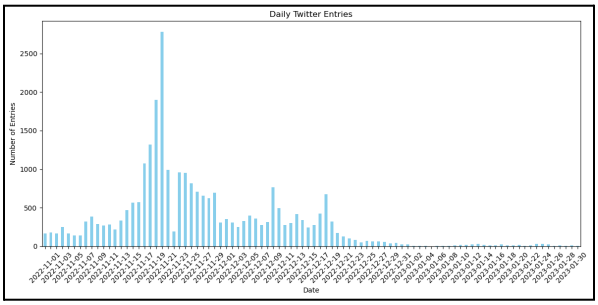
```
X = twitter['content']
y = twitter['Sentiment']

vectorizer = TfidfVectorizer()
X_vectorized = vectorizer.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X_vectorized, y, test_size=0.2, random_state=42)

svm_model = SVC(kernel='linear')
svm_model.fit(X_train, y_train)

y_pred = svm_model.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
```

```
np.random.seed(0)
random_data = twitter.groupby('Sentiment', group_keys=False).apply(lambda x: x.sample(min(len(x), 300)))

X_random = random_data['content']
y_random = random_data['Sentiment']

X_vectorized_random = vectorizer.transform(X_random)
y_pred_random = svm_model.predict(X_vectorized_random)

data_random = pd.DataFrame({'Sentiment': y_pred_random, 'Engagement': random_data['Likes']})

average_engagement_random = data_random.groupby('Sentiment')['Engagement'].mean()

plt.figure(figsize=(8, 6))
average_engagement_random.plot(kind='bar', color=['#0084b4', '#1da1f2', '#76c5f5'])
plt.title("Average Engagement by Sentiment in Random Dataset")
plt.xlabel("Sentiment")
plt.ylabel("Average Engagement")
plt.xticks(range(len(average_engagement_random.index)), average_engagement_random.index, rotation=0)
plt.show()
```

- Code to create SVM model & predicted engagement plot







- Relevant plots on sentiment analysis, focusing on engagement + improving graph's visual appeal

Week 12 - Reflection #2

This week, we presented our final presentations in the last studio. Overall, the past few days have been very tiring but simultaneously rewarding since it was the end of a project that I worked hard on, especially over the past few weeks. As a result, I was quite satisfied with our group's final product.

Overall, working on this project was definitely a valuable experience. Compared to the projects I undertook throughout the previous semesters, I consider this to be the most challenging since it involved data science concepts that I had limited knowledge and experience with at the start of the semester, primarily dealing with text and sentiment analysis. I tried to overcome this challenge by researching these data science topics outside of class to improve my understanding. In addition, the potential research approaches were less defined and a lot more open-ended compared to first-year projects, thus requiring a higher level of critical thinking. There were also multiple datasets involved in the project, compared to my past experiences of mostly exploring just a singular dataset. As a result, it became more important for me to collaborate well with my group to ensure that our findings for each dataset synced and flowed well during the presentation, while also justifying the reasoning behind these connections. To summarize my contributions, I was singlehandedly in charge of the Twitter dataset, where I performed analysis on this dataset from start to finish. This included performing the necessary data cleaning and text-preprocessing, and answering all the research questions for the Twitter section. I also contributed some significant ideas to the overall project, such as the suggestion of RQ#4 and the idea of constructing the time-series graph to capture the evolution of player and team trends.

In this project, there were several challenges that arose. I believe that the main obstacle our group faced was deciding on our approach to this investigation, especially with the World Cup being a topic with so many possible routes for analysis, along with the multiple datasets measuring so many different variables. As a result, dealing with this was extremely time-consuming which slowed down our overall progress, especially during the first half of the semester where we were unsure on our research questions until after the mock presentation conducted in week 8. Another challenge we faced was the implementation of a predictive model on a project dominated by text-analysis, which was the focus of RQ#3. Eventually, we tried our best to tackle this issue through the packages, but it was perhaps an area of the presentation that could have still been improved. Personally, one area I could have improved on would be cross-checking with my teammates more often at the start of the semester, which was really important since they were all working on a different dataset. This made our progress less effective as we didn't really examine the differences between each platform until the final few weeks of the semester. As a result, we really had to increase our effort and time dedicated to these projects over the past month as we had a lot of catching-up to do.