Question	Points
2c	2
2d	2
2e	2
2f	1
3a	1
3b	1
3c	1
4a	1
4b	1
4ci	1
4cii	1
4d	1
4e	2
4f	2
4g	2
5a	2
5b	2
Total	**35**

Question 1: Importing the Data

The data for this assignment was obtained using the Twitter APIs. To ensure that everyone has the same data and to eliminate the need for every student to apply for a Twitter developer account, we have collected a sample of tweets from several high-profile public figures. The data is stored in the folder `data` . Run the following cell to list the contents of the directory:

In [147…

```
from os import listdir
for f in listdir("data"):
    print(f)
```

```
AOC_recent_tweets.txt
BernieSanders_recent_tweets.txt
BillGates_recent_tweets.txt
Cristiano_recent_tweets.txt
EmmanuelMacron_recent_tweets.txt
elonmusk_recent_tweets.txt
```

Question 1a

Let's examine the contents of one of these files. Using the `open` function and `read` operation on a python file object, read the first 1000 characters in `data/BernieSanders_recent_tweets.txt` and store your result in the variable `q1a`. Then display the result so you can read it.

Caution: Viewing the contents of large files in a Jupyter notebook could crash your browser. Be careful not to print the entire contents of the file.

Hint: You might want to try to use `with`:

```
with open("filename", "r") as f:
    f.read(2)
```

In [148…

```
with open("data/BernieSanders_recent_tweets.txt","r") as f:
    q1a = f.read(1000)
q1a
```

Out[148…

```
'[{"created_at": "Sat Feb 06 22:43:03 +0000 2021", "id": 1358184460794163202, "id_str": "1
358184460794163202", "full_text": "Why would we want to impeach and convict Donald Trump
\\u2013 a president who is now out of office? Because it must be made clear that no presid
ent, now or in the future, can lead an insurrection against the government he or she is sw
orn to protect.", "truncated": false, "display_text_range": [0, 243], "entities": {"hashta
gs": [], "symbols": [], "user_mentions": [], "urls": []}, "source": "<a href=\\"http://twi
tter.com/download/iphone\\" rel=\\"nofollow\\">Twitter for iPhone</a>", "in_reply_to_statu
s_id": null, "in_reply_to_status_id_str": null, "in_reply_to_user_id": null, "in_reply_to_
user_id_str": null, "in_reply_to_screen_name": null, "user": {"id": 216776631, "id_str":
"216776631", "name": "Bernie Sanders", "screen_name": "BernieSanders", "location": "Vermon
t", "description": "U.S. Senator for Vermont. Not me, us.", "url": "https://t.co/jpg8Sp1Gh
R", "entities": {"'
```

In [149…

```
grader.check("q1a")
```

Out[149…

q1a passed!

Question 1b

What format is the data in? Answer this question by entering the letter corresponding to the right format in the variable `q1b` below.

a. CSV
b. HTML
c. JavaScript Object Notation (JSON)
d. Excel XML

In [150…

```
q1b = "c"
```

In [151…

```
grader.check("q1b")
```

Out[151…

q1b passed!

Question 1c

Pandas has built-in readers for many different file formats including the file format used here to store tweets. To learn more about these, check out the documentation for `pd.read_csv` , `pd.read_html` , `pd.read_json` , and `pd.read_excel` .

1. Use one of these functions to populate the `tweets` dictionary with the tweets for: `AOC` , `Cristiano` , and `elonmusk` . The keys of `tweets` should be the handles of the users, which we have provided in the cell below, and the values should be the dataframes.
2. Set the index of each dataframe to correspond to the `id` of each tweet.

Hint: You might want to first try loading one of the DataFrames before trying to complete the entire question.

In [152…

```
tweets = {
    "AOC": pd.read_json('data/AOC_recent_tweets.txt').set_index('id'),
    "Cristiano": pd.read_json('data/Cristiano_recent_tweets.txt').set_index('id'),
    "elonmusk": pd.read_json('data/elonmusk_recent_tweets.txt').set_index('id')
}
```

In [153…

```
grader.check("q1c")
```

Out[153…

`q1c` passed!

If you did everything correctly, the following cells will show you the first 5 tweets for Elon Musk (and a lot of information about those tweets).

In [154…

```
tweets["elonmusk"].head()
```

Out[154…

	created_at	id_str	full_text	truncated	display_text_r
id					
1357991946082418690	2021-02-06 09:58:04+00:00	1357991946082418688	The Second Last Kingdom https://t.co/Je4EI88HmV	False	[
1357973565413367808	2021-02-06 08:45:02+00:00	1357973565413367808	@DumDin7 @Grimezsz Haven’t heard that name in years …	False	[1
1357972904663687173	2021-02-06 08:42:25+00:00	1357972904663687168	@Grimezsz Dogecake	False	[1
1357970517165182979	2021-02-06 08:32:55+00:00	1357970517165182976	YOLT\n\nhttps://t.co/cnOf9yjpF1	False	[
1357964347813687296	2021-02-06 08:08:24+00:00	1357964347813687296	@Kristennetten That’s Damian	False	[1

5 rows × 30 columns

Question 1d

There are many ways we could choose to read tweets. Why might someone be interested in doing data analysis on tweets? Name a kind of person or institution which might be interested in this kind of analysis. Then, give two reasons why a data analysis of tweets might be interesting or useful for them. Answer in 2-3 sentences.

Apple might be interested in data analysis on tweets. First, since they released iphone 13 recently, they are able to obtain sense of popularity or satisfaction score about their New iphone through data analysis of tweets. Second, they are able to obtain sentiments of customers to improve their products for the future.

Question 2: Source Analysis

In some cases, the Twitter feed of a public figure may be partially managed by a public relations firm. In these cases, the device used to post the tweet may help reveal whether it was the individual (e.g., from an iPhone) or a public relations firm (e.g., TweetDeck). The tweets we have collected contain the source information but it is formatted strangely :(

In [155…

```
tweets["Cristiano"][["source"]]
```

Out[155…

id	source
1358137564587319299	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
1357379984399212545	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
1356733030962987008	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
1355924395064233986	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
1355599316300292097	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
...	...
32514882561638401	<a href="http://www.whosay.com" rel="nofollow">WhoSay</a>
32513604662071296	<a href="http://www.whosay.com" rel="nofollow">WhoSay</a>
32511823722840064	<a href="http://www.whosay.com" rel="nofollow">WhoSay</a>
32510294081146881	<a href="http://www.whosay.com" rel="nofollow">WhoSay</a>
32508748819857410	<a href="http://www.whosay.com" rel="nofollow">WhoSay</a>

3198 rows × 1 columns

In this question we will use a regular expression to convert this messy HTML snippet into something more readable. For example: `<a href="http://twitter.com/download/iphone"`

rel="nofollow">Twitter for iPhone</a> should be "Twitter for iPhone".

Question 2a

We will first use the Python `re` library to cleanup the above test string. In the cell below, write a regular expression that will match the **HTML tag** and assign it to the variable `q2a_pattern`. We then use the `re.sub` function to substitute anything that matches the pattern with an empty string `""`.

An HTML tag is defined as a `<` character followed by zero or more non-`>` characters, followed by a `>` character. That is `<a>` and `</a>` are both considered *separate* HTML tags.

In [156…
```
q2a_pattern = r"<.*?>"
test_str = '<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone
re.sub(q2a_pattern, "", test_str)
```

Out[156…
```
'Twitter for iPhone'
```

In [157…
```
grader.check("q2a")
```

Out[157…

q2a passed!

Question 2b

Rather than writing a regular expression to detect and remove the HTML tags we could instead write a regular expression to **capture** the device name between the angle brackets. Here we will use **capturing groups** by placing parenthesis around the part of the regular expression we want to return. For example, to capture the `21` in the string `08/21/83` we could use the pattern `r"08/(..)/83"`.

Hint: The output of the following cell should be `['Twitter for iPhone']`.

In [158…
```
q2b_pattern = r">(.*)<"
test_str = '<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone
re.findall(q2b_pattern, test_str)
```

Out[158…
```
['Twitter for iPhone']
```

In [159…
```
grader.check("q2b")
```

Out[159…

q2b passed!

Question 2c

Using either of the two regular expressions you just created and `Series.str.replace` or `Series.str.extract`, add a new column called `"device"` to **all** of the dataframes in `tweets` containing just the text describing the device (without the HTML tags).

In [160…
```
tweets["AOC"]["device"] = tweets["AOC"]["source"].str.extract(">(.*)<")
tweets["Cristiano"]["device"] = tweets["Cristiano"]["source"].str.extract(">(.*)<")
tweets["elonmusk"]["device"] = tweets["elonmusk"]["source"].str.extract(">(.*)<")
tweets["AOC"].head()
```

Out[160…

	created_at	id_str	full_text	truncated	display_text_range
id					
1358149122264563712	2021-02-06 20:22:38+00:00	1358149122264563712	RT @RepEscobar: Our country has the moral obligation and responsibility to reunite every single family separated at the southern border.\n\nT…	False	[0, 140]
1358147616400408576	2021-02-06 20:16:39+00:00	1358147616400408576	RT @RoKhanna: What happens when we guarantee $15/hour?\n\n💰 31% of Black workers and 26% of Latinx workers get raises.\n😷 A majority of essent…	False	[0, 140]
1358145332316667909	2021-02-06 20:07:35+00:00	1358145332316667904	(Source: https://t.co/3o5JEr6zpd)	False	[0, 33]
1358145218407759875	2021-02-06 20:07:07+00:00	1358145218407759872	Joe Cunningham pledged to never take corporate PAC money, and he never did. Mace said she'll cash every check she gets. Yet another way this is a downgrade. https://t.co/DytsQXKXgU	False	[0, 156]
1358144207333036040	2021-02-06 20:03:06+00:00	1358144207333036032	What's even more gross is that Mace takes corporate PAC money.\n\nShe's already funded by corporations. Now she's choosing to swindle working people on top of it.\n\nPeak scam artistry. Caps for cash 💰 https://t.co/CcVxgDF6id	False	[0, 197]

5 rows × 31 columns

In [161…

```
grader.check("q2c")
```

Out[161…

q2c passed!

In [162…

```
tweets["Cristiano"]['device'].value_counts().head(5)
```

Out[162…

```
Twitter for iPhone     1183
Twitter Web Client      959
WhoSay                  453
MobioINsider.com        144
Twitter for Android     108
Name: device, dtype: int64
```

To examine the most frequently used devices by each individual, implement the `most_freq` function that takes in a `Series` and returns a new `Series` containing the `k` most commonly occuring entries in the first series, where the values are the counts of the entries and the indices are the entries themselves.

For example:

```
most_freq(pd.Series(["A", "B", "A", "C", "B", "A"]), k=2)
```

would return:

```
A    3
B    2
dtype: int64
```

Hint Consider using `value_counts` , `sort_values` , `head` , and/or `nlargest` . Think of what might be the most efficient implementation.

In [163…
```
def most_freq(series, k = 5):
    return series.value_counts().head(k)

most_freq(tweets["Cristiano"]['device'])
```

Out[163…
```
Twitter for iPhone     1183
Twitter Web Client      959
WhoSay                  453
MobioINsider.com        144
Twitter for Android     108
Name: device, dtype: int64
```

In [164…
```
grader.check("q2d")
```

Out[164…

q2d passed!

Question 2e

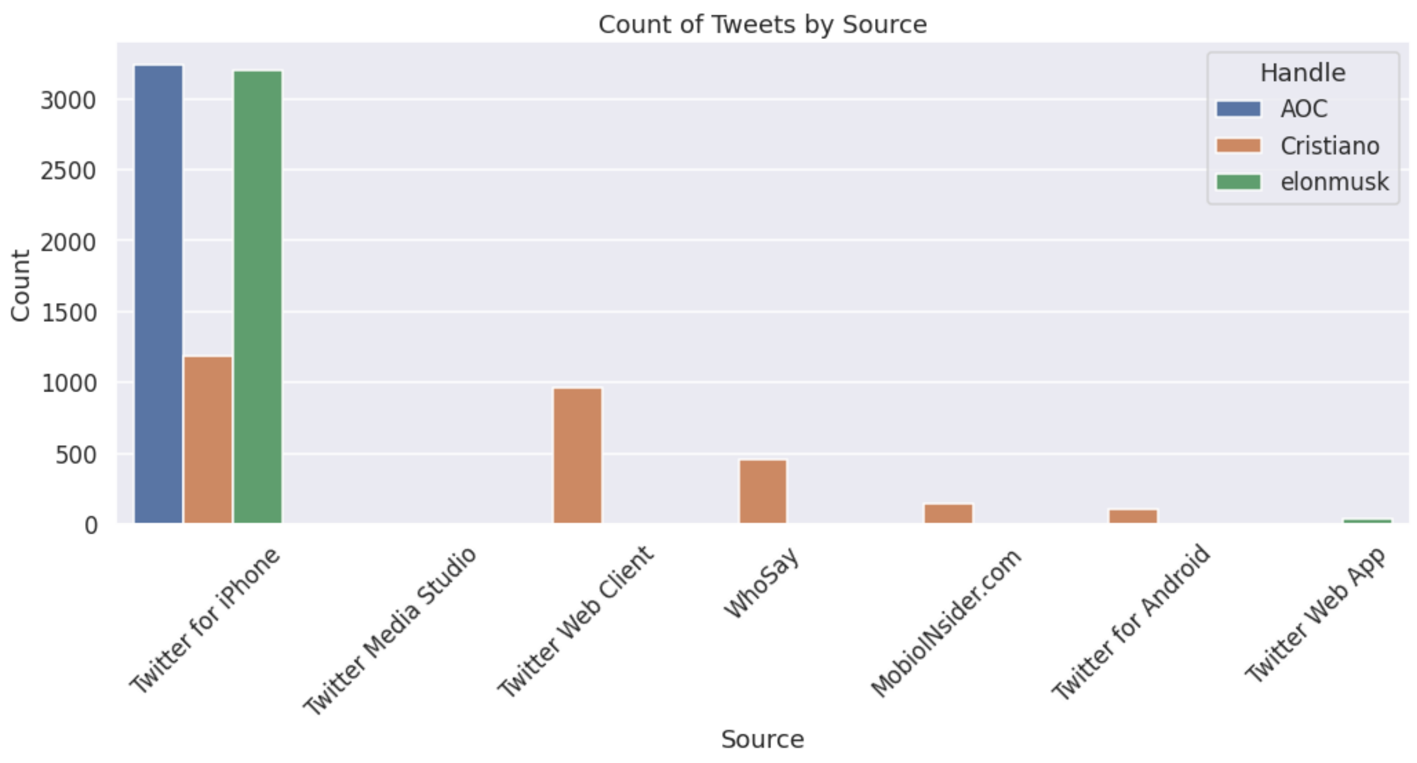
Run the following two cells to compute a table and plot describing the top 5 most commonly used devices for each user.

In [165…
```
device_counts = pd.DataFrame(
    [most_freq(tweets[name]['device']).rename(name)
     for name in tweets]
).fillna(0)
device_counts
```

Out[165…

	Twitter for iPhone	Twitter Media Studio	Twitter Web Client	WhoSay	MobioINsider.com	Twitter for Android	Twitter Web App
AOC	3245.0	2.0	0.0	0.0	0.0	0.0	0.0
Cristiano	1183.0	0.0	959.0	453.0	144.0	108.0	0.0
elonmusk	3202.0	0.0	0.0	0.0	0.0	0.0	37.0

In [166…
```
plt.figure(figsize=[15,6])
sns.barplot(x="index", y="value", hue="variable", data=device_counts.T.reset_index().melt(
plt.title("Count of Tweets by Source")
plt.ylabel("Count")
plt.xlabel("Source")
plt.legend(title="Handle");
```



What might we want to investigate further? Write a few sentences below.

We can investigate why AOC and elonmusk have higher number of tweets for iphone than Cristiano. We can figure this out by obtaining the proportion of devices that are used by each users.

Question 2f

We just looked at the top 5 most commonly used devices for each user. However, we used the number of tweets as a measure, when it might be better to compare these distributions by comparing *proportions* of tweets. Why might proportions of tweets be better measures than numbers of tweets?

It is better to compare these distributions by comparing proportions of tweet since each user have different frequency of using twitter. For example, one user tweets a lot that person will definitely have high number of tweets.

Question 3: When?

Now that we've explored the sources of each of the tweets, we will perform some time series analysis. A look into the temporal aspect of the data could reveal insights about how a user spends their day, when they eat and sleep, etc. In this question, we will focus on the time at which each tweet was posted.

In [167…

```
tweets["AOC"].head()
```

Out[167…

	created_at	id_str	full_text	truncated	display_text_range
id					
1358149122264563712	2021-02-06 20:22:38+00:00	1358149122264563712	RT @RepEscobar: Our country has the moral obligation and responsibility to reunite every single family separated at the southern border.\n\nT…	False	[0, 140]
1358147616400408576	2021-02-06 20:16:39+00:00	1358147616400408576	RT @RoKhanna: What happens when we guarantee $15/hour?\n\n💰 31% of Black workers and 26% of Latinx workers get raises.\n😷 A majority of essent…	False	[0, 140]
1358145332316667909	2021-02-06 20:07:35+00:00	1358145332316667904	(Source: https://t.co/3o5JEr6zpd)	False	[0, 33]
1358145218407759875	2021-02-06 20:07:07+00:00	1358145218407759872	Joe Cunningham pledged to never take corporate PAC money, and he never did. Mace said she'll cash every check she gets. Yet another way this is a downgrade. https://t.co/DytsQXKXgU	False	[0, 156]
1358144207333036040	2021-02-06 20:03:06+00:00	1358144207333036032	What's even more gross is that Mace takes corporate PAC money.\n\nShe's already funded by corporations. Now she's choosing to swindle working people on top of it.\n\nPeak scam artistry. Caps for cash 💰 https://t.co/CcVxgDF6id	False	[0, 197]

5 rows × 31 columns

Question 3a

Complete the following function `add_hour` that adds a new column `hour` to a tweets `DataFrame` based on the column `time_col` containing the timestamps. The `hour` column should contain the hour of the day as floating point number computed by:

$$\text{hour} + \frac{\text{minute}}{60} + \frac{\text{second}}{60^2}$$

Hint: See the following link for an example of working with timestamps using the `dt` accessors. You should use the `created_at` column to calculate the hour.

In [168…
```
def add_hour(df, time_col="created_at", result_col="hour"):
    hour = df[time_col].dt.hour
    minute = df[time_col].dt.minute
    second = df[time_col].dt.second
    df[result_col] = hour + minute/60 + second/3600
    return df

tweets = {handle: add_hour(df) for handle, df in tweets.items()}
tweets["AOC"]["hour"].head()
```

Out[168…
```
id
1358149122264563712    20.377222
1358147616400408576    20.277500
1358145332316667909    20.126389
1358145218407759875    20.118611
1358144207333036040    20.051667
Name: hour, dtype: float64
```

In [169…
```
grader.check("q3a")
```

Out[169…

q3a passed!

Question 3b

With our new `hour` column, let's take a look at the distribution of tweets for each user by time of day. The following cell helps create a density plot on the number of tweets based on the hour they are posted.

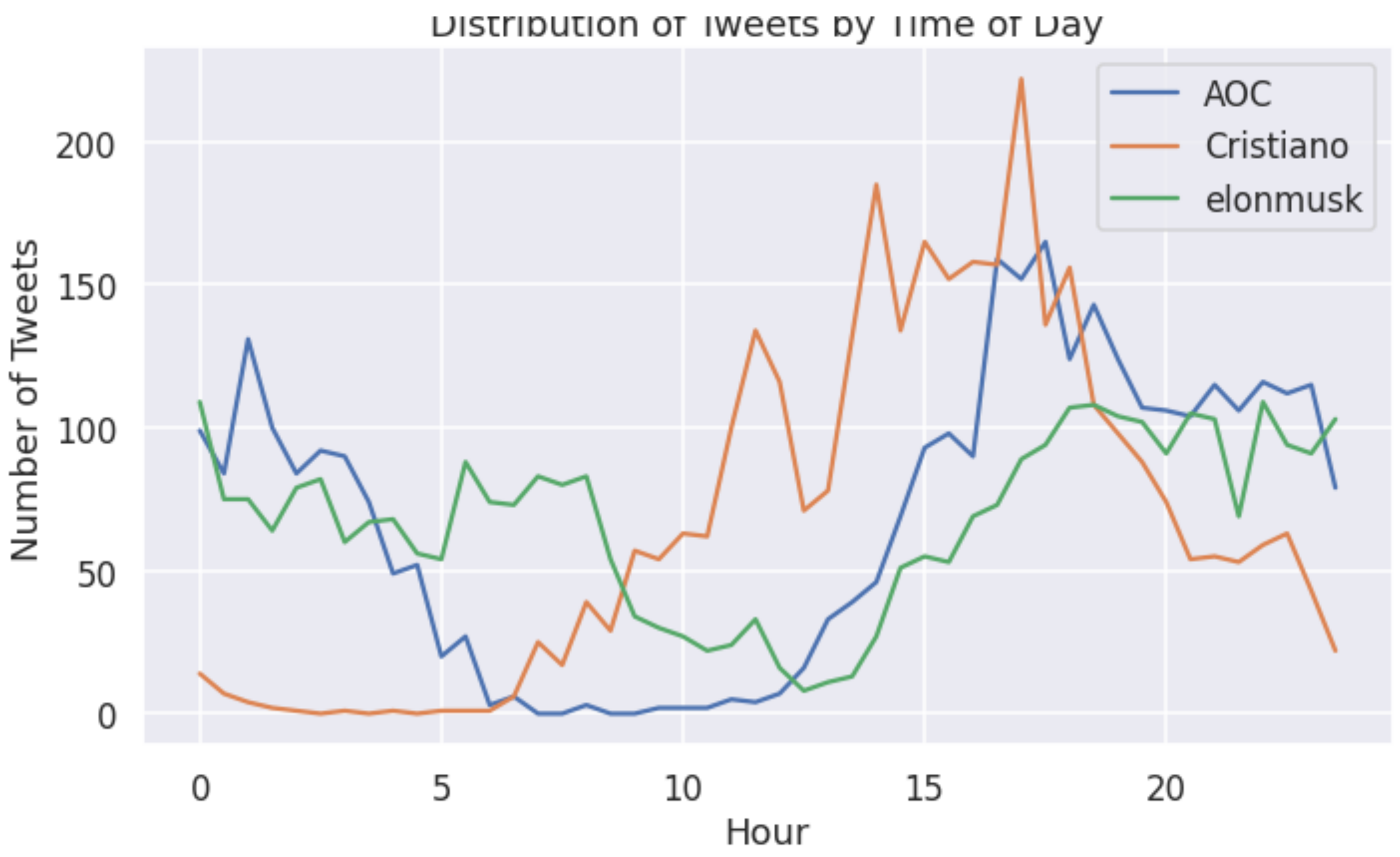
The function `bin_df` takes in a dataframe, an array of bins, and a column name and bins the the values in the specified column, returning a dataframe with the bin lower bound and the number of elements in the bin. This function uses `pd.cut`, a pandas utility for binning numerical values that you may find helpful in the future.

Run the cell and answer the following question about the plot.

In [170…
```
def bin_df(df, bins, colname):
    binned = pd.cut(df[colname], bins).value_counts().sort_index()
    return pd.DataFrame({"counts": binned, "bin": bins[:-1]})

hour_bins = np.arange(0, 24.5, .5)
binned_hours = {handle: bin_df(df, hour_bins, "hour") for handle, df in tweets.items()}

make_line_plot(binned_hours, "bin", "counts", title="Distribution of Tweets by Time of Da
               xlabel="Hour", ylabel="Number of Tweets")
```



Compare Cristiano's distribution with those of AOC and Elon Musk. In particular, compare the distributions before and after hour 6. What differences did you notice? What might be a possible cause of that? Do the data plotted above seem reasonable?

It seems like Cristiano tend to tweet less frequently than AOC and elonmusk from hour 0 to hour 6 while AOC and elonmusk tweet less than Cristiano after hour 6. Possible cause of that might be because they are living in a different timezone. The data plotted above will be reasonable when we convert the timezone.

Question 3c

To account for different locations of each user in our analysis, we will next adjust the `created_at` timestamp for each tweet to the respective timezone of each user. Complete the following function `convert_timezone` that takes in a tweets `DataFrame` and a timezone `new_tz` and add a new column `converted_time` that has the adjusted `created_at` timestamp for each tweet. The timezone for each user is provided in `timezones` .

Hint: Again, please see the following link for an example of working with `dt` accessors.

In [171…

```
def convert_timezone(df, new_tz):
    df['converted_time'] = df['created_at'].dt.tz_convert(new_tz)
    return df

timezones = {"AOC": "EST", "Cristiano": "Europe/Lisbon", "elonmusk": "America/Los_Angeles'

tweets = {handle: convert_timezone(df, tz) for (handle, df), tz in zip(tweets.items(), tin
```

In [172…

```
grader.check("q3c")
```

Out[172…

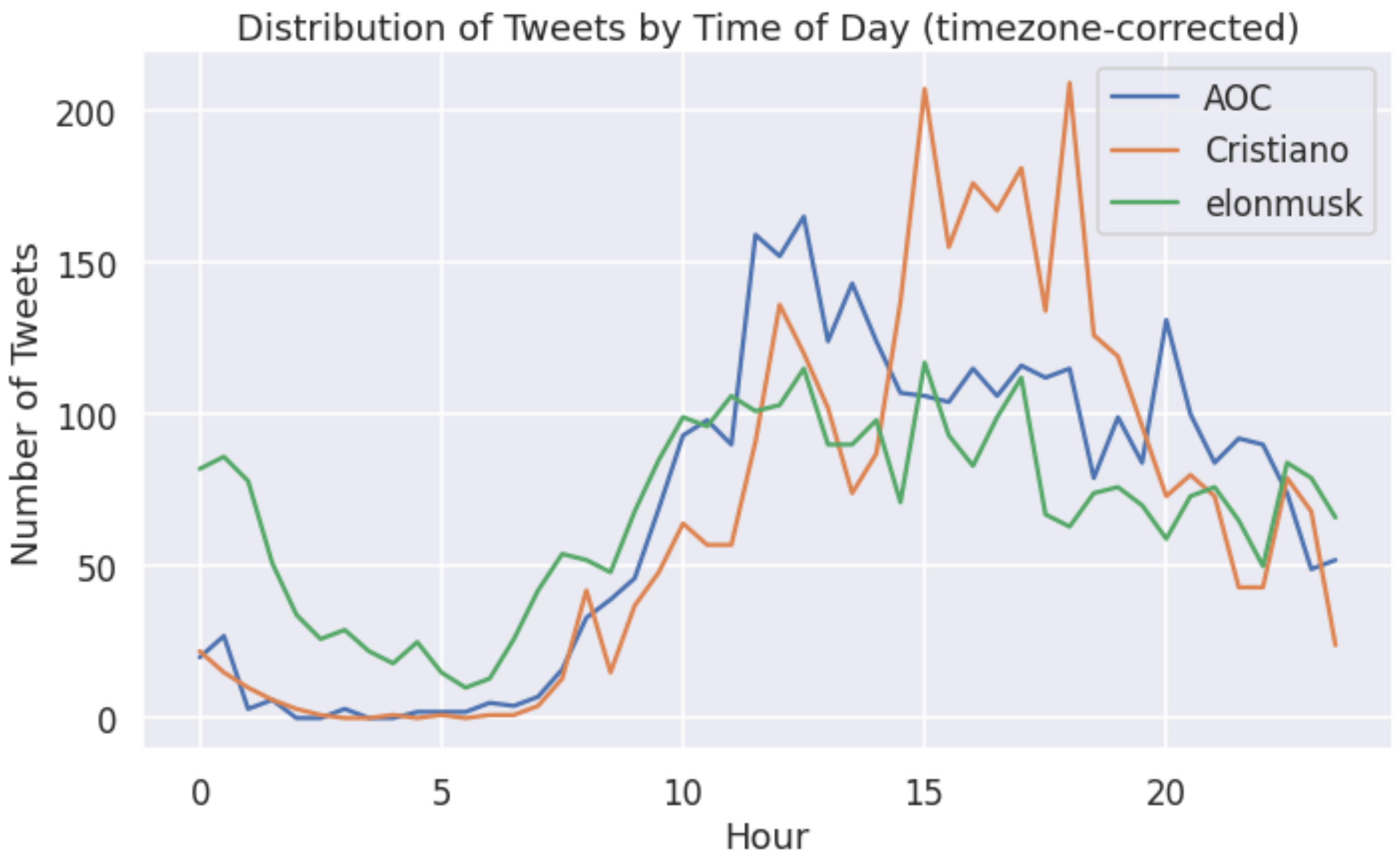
q3c passed!

With our adjusted timestamps for each user based on their timezone, let's take a look again at the distribution of tweets by time of day.

In [173…

```
tweets = {handle: add_hour(df, "converted_time", "converted_hour") for handle, df in tweet
binned_hours = {handle: bin_df(df, hour_bins, "converted_hour") for handle, df in tweets.i

make_line_plot(binned_hours, "bin", "counts", title="Distribution of Tweets by Time of Day
               xlabel="Hour", ylabel="Number of Tweets")
```



Question 4: Sentiment

In the past few questions, we have explored the sources of the tweets and when they are posted. Although on their own, they might not seem particularly intricate, combined with the power of regular expressions, they could actually help us infer a lot about the users. In this section, we will continue building on our past analysis and specifically look at the sentiment of each tweet -- this would lead us to a much more direct and detailed understanding of how the users view certain subjects and people.

How do we actually measure the sentiment of each tweet? In our case, we can use the words in the text of a tweet for our calculation! For example, the word "love" within the sentence "I love America!" has a positive sentiment, whereas the word "hate" within the sentence "I hate taxes!" has a negative sentiment. In addition, some words have stronger positive / negative sentiment than others: "I love America." is more positive than "I like America."

We will use the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon to analyze the sentiment of AOC's tweets. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media which is great for our usage.

The VADER lexicon gives the sentiment of individual words. Run the following cell to show the first few rows of the lexicon:

In [174…

```python
print(''.join(open("vader_lexicon.txt").readlines()[:10]))
```

```
$:      -1.5    0.80623 [-1, -1, -1, -1, -3, -1, -3, -1, -2, -1]
%)      -0.4    1.0198  [-1, 0, -1, 0, 0, -2, -1, 2, -1, 0]
%-)     -1.5    1.43178 [-2, 0, -2, -2, -1, 2, -2, -3, -2, -3]
&-:     -0.4    1.42829 [-3, -1, 0, 0, -1, -1, -1, 2, -1, 2]
&:      -0.7    0.64031 [0, -1, -1, -1, 1, -1, -1, -1, -1, -1]
( '}{' )        1.6     0.66332 [1, 2, 2, 1, 1, 2, 2, 1, 3, 1]
(%      -0.9    0.9434  [0, 0, 1, -1, -1, -1, -2, -2, -1, -2]
('-:    2.2     1.16619 [4, 1, 4, 3, 1, 2, 3, 1, 2, 1]
(':     2.3     0.9     [1, 3, 3, 2, 2, 4, 2, 3, 1, 2]
((-:    2.1     0.53852 [2, 2, 2, 1, 2, 3, 2, 2, 3, 2]
```

As you can see, the lexicon contains emojis too! Each row contains a word and the *polarity* of that word, measuring how positive or negative the word is.

VADER Sentiment Analysis

The creators of VADER describe the tool's assessment of polarity, or "compound score," in the following way:

"The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). This is the most useful metric if you want a single unidimensional measure of sentiment for a given sentence. Calling it a 'normalized, weighted composite score' is accurate."

As you can see, VADER doesn't "read" sentences, but works by parsing sentences into words, assigning a preset generalized score from their testing sets to each word separately.

VADER relies on humans to stabilize its scoring. The creators use Amazon Mechanical Turk, a crowdsourcing survey platform, to train its model. Its training data consists of a small corpus of tweets, New York Times editorials and news articles, Rotten Tomatoes reviews, and Amazon product reviews, tokenized using the natural language toolkit (NLTK). Each word in each dataset was reviewed and rated by at least 20 trained individuals who had signed up to work on these tasks through Mechanical Turk.

Question 4a

Please score the sentiment of one of the following words, using your own personal interpretation. No code is required for this question!

- police
- order
- Democrat
- Republican

- gun
- dog
- technology
- TikTok
- security
- face-mask
- science
- climate change
- vaccine

What score did you give it and why? Can you think of a situation in which this word would carry the opposite sentiment to the one you've just assigned?

I would give face-mask -0.4 because we are with our face-mask everyday which is really disturbing since COVID-19 happened. It will give a positive sentiment when we are finally able to get our masks off when the pandemic is over.

Optional (ungraded): Are there circumstances (e.g. certain kinds of language or data) when you might not want to use VADER? What features of human speech might VADER misrepresent or fail to capture?

Question 4b

Let's first load in the data containing all the sentiments. Read `vader_lexicon.txt` into a DataFrame called `sent`. The index of the DataFrame should be the words in the lexicon and should be named `token`. `sent` should have one column named `polarity`, storing the polarity of each word.

Hint: The `pd.read_csv` function may help here. Since the file is tab-separated, be sure to set `sep='\t'` in your call to `pd.read_csv`.

In [175…
```
sent = pd.read_csv('vader_lexicon.txt',sep='\t',names = ['token','polarity','a','b']).set_
sent = sent[['polarity']]
sent.head()
```

Out[175…

token	polarity
$:	-1.5
%)	-0.4
%-)	-1.5
&-:	-0.4
&:	-0.7

In [176…
```
grader.check("q4b")
```

Out[176…

q4b passed!

Question 4c

Before further analysis, we will need some more tools that can help us extract the necessary information and clean our data.

Complete the following regular expressions that will help us match part of a tweet that we are either interested in learning more about or want to remove.

Part 1

Assign a regular expression to a new variable `punct_re` that captures all of the punctuations within a tweet. We consider punctuation to be any non-word, non-whitespace character.

Note: A word character is any character that is alphanumeric or an underscore. A whitespace character is any character that is a space, a tab, a new line, or a carriage return.

In [177…

```
punct_re = r'[^a-zA-z0-9\s]'

re.sub(punct_re, " ", tweets["AOC"].iloc[0]["full_text"])
```

Out[177…

```
'RT  RepEscobar  Our country has the moral obligation and responsibility to reunite every
single family separated at the southern border \n\nT '
```

In [178…

```
grader.check("q4ci")
```

Out[178…

`q4ci` passed!

Part 2

Assign a regular expression to a new variable `mentions_re` that matches any mention in a tweet. Your regular expression should use a capturing group to extract the user's username in a mention.

Hint: a user mention within a tweet always starts with the `@` symbol and is followed by a series of word characters (with no space in between). For more explanations on what a word character is, check out the **Note** section in Part 1.

In [179…

```
mentions_re = r'@([a-zA-z]*)'

re.findall(mentions_re, tweets["AOC"].iloc[0]["full_text"])
```

Out[179…

```
['RepEscobar']
```

In [180…

```
grader.check("q4cii")
```

Out[180…

`q4cii` passed!

Tweet Sentiments and User Mentions

As you have seen in the previous part of this question, there are actually a lot of interesting components that we can extract out of a tweet for further analysis! For the rest of this question though, we will focus on one particular case: the sentiment of each tweet in relation to the users mentioned within it.

To calculate the sentiments for a sentence, we will follow this procedure:

1. Remove the punctuation from each tweet so we can analyze the words.
2. For each tweet, find the sentiment of each word.
3. Calculate the sentiment of each tweet by taking the sum of the sentiments of its words.

Question 4d

Let's use our `punct_re` regular expression from the previous part to clean up the text a bit more! The goal here is to remove all of the punctuations to ensure words can be properly matched with those from VADER to actually calculate the full sentiment score.

Complete the following function `sanitize_texts` that takes in a table `df` and adds a new column `clean_text` by converting all characters in its original `full_text` column to lower case and replace all instances of punctuations with a space character.

In [181…
```
def sanitize_texts(df):
    df["clean_text"] = df['full_text'].str.replace(r'[^a-zA-z0-9\s]',' ',regex = True).str
    return df

tweets = {handle: sanitize_texts(df) for handle, df in tweets.items()}
tweets["AOC"]["clean_text"].head()
```

Out[181…
```
id
1358149122264563712
                rt  repescobar  our country has the moral obligation and responsibility to
reunite every single family separated at the southern border \n\nt
1358147616400408576
               rt  rokhanna  what happens when we guarantee  15 hour \n\n  31  of black wo
rkers and 26  of latinx workers get raises \n  a majority of essent
1358145332316667909

                                    source  https   t co 3o5jer6zpd
1358145218407759875                                                        joe cunningham pledged
to never take corporate pac money  and he never did  mace said she ll cash every check she
gets  yet another way this is a downgrade  https   t co dytsqxkxgu
1358144207333036040    what s even more gross is that mace takes corporate pac money \n\ns
he s already funded by corporations  now she s choosing to swindle working people on top o
f it \n\npeak scam artistry  caps for cash   https   t co ccvxgdf6id
Name: clean_text, dtype: object
```

In [182…
```
grader.check("q4d")
```

Out[182…

q4d passed!

Question 4e

With the texts sanitized, we can now extract all the user mentions from tweets.

Complete the following function `extract_mentions` that takes in the **`full_text` column** from a tweets `DataFrame` and uses `mentions_re` to extract all the mentions in a DataFrame. The returned dataframe, renamed to `mentions`, is single-indexed and has all lower-cased characters with a separate row for each mention.

In [185…
```
def extract_mentions(full_texts):
    mentions = full_texts.str.lower().str.extract(r'@([a-zA-z]*)')
    mentions['mentions'] = mentions
    return mentions[["mentions"]].fillna(0)

mentions = {handle: extract_mentions(df["full_text"]) for handle, df in tweets.items()}
horiz_concat_df(mentions).head()
```

Out[185

	AOC mentions	Cristiano mentions	elonmusk mentions
0	repescobar	0	0
1	rokhanna	0	dumdin
2	0	0	grimezsz
3	0	0	0
4	0	0	kristennetten

In [186…

```
grader.check("q4e")
```

Out[186…

q4e passed!

Tidying Up the Data

Now, let's convert the tweets into what's called a *tidy format* to make the sentiments easier to calculate. We will use the `clean_text` column of each dataframe to create a tidy table, which will be returned by `to_tidy_format`. The index of the table will be the IDs of the tweets, repeated once for every word in the tweet. It has two columns:

1. `word` : The individual words of each tweet.

Run the following cell to convert the table into the tidy format. Take a look at the first 5 rows from the "tidied" tweets dataframe for AOC and see if you can find out how the structure has changed.

Note: Although there is no work needed on your part, we have referenced a few more advanced pandas methods you might have not seen before -- you should definitely look them up in the documentation when you have a chance, as they are quite powerful in restructuring a dataframe and converting one into a useful intermediate state!

In [187…

```
def to_tidy_format(df):
    tidy = (
        df["clean_text"]
        .str.split()
        .explode()
        .to_frame()
        .rename(columns={"clean_text": "word"})
    )
    return tidy

tidy_tweets = {handle: to_tidy_format(df) for handle, df in tweets.items()}
tidy_tweets["AOC"].head()
```

Out[187…

	word
id	
1358149122264563712	rt
1358149122264563712	repescobar
1358149122264563712	our
1358149122264563712	country
1358149122264563712	has

Adding in the Polarity Score

Now that we have this table in the tidy format, it becomes much easier to find the sentiment of each tweet: we can join the table with the lexicon table.

The following `add_polarity` function adds a new `polarity` column to the `df` table. The `polarity` column contains the sum of the sentiment polarity of each word in the text of the tweet.

Note: Again, though there is no work needed on your part, it is important for you to go through how we set up this method and actually understand what each method is doing.

In [188…

```
def add_polarity(df, tidy_df):
    df["polarity"] = (
        tidy_df
        .merge(sent, how='left', left_on='word', right_index=True)
        .reset_index()
        .loc[:, ['id', 'polarity']]
        .groupby('id')
        .sum()
        .fillna(0)
    )
    return df

tweets = {handle: add_polarity(df, tidy_df) for (handle, df), tidy_df in \
          zip(tweets.items(), tidy_tweets.values())}
tweets["AOC"][["clean_text", "polarity"]].head()
```

Out[188…

	clean_text	polarity
id		
1358149122264563712	rt repescobar our country has the moral obligation and responsibility to reunite every single family separated at the southern border \n\nt	0.0
1358147616400408576	rt rokhanna what happens when we guarantee 15 hour \n\n 31 of black workers and 26 of latinx workers get raises \n a majority of essent	1.0
1358145332316667909	source https t co 3o5jer6zpd	0.0
1358145218407759875	joe cunningham pledged to never take corporate pac money and he never did mace said she ll cash every check she gets yet another way this is a downgrade https t co dytsqxkxgu	0.0
1358144207333036040	what s even more gross is that mace takes corporate pac money \n\nshe s already funded by corporations now she s choosing to swindle working people on top of it \n\npeak scam artistry caps for cash https t co ccvxgdf6id	-6.4

In [189…

```
mentions
```

Out[189…

```
{'AOC':                        mentions
 id
 1358149122264563712  repescobar
 1358147616400408576    rokhanna
 1358145332316667909           0
 1358145218407759875           0
 1358144207333036040           0
 ...                         ...
 1181935928249606146           0
 1181932554552827905    heidinbc
 1181932460516478976           0
 1181927615340453899           0
 1181804625588051968   leaninorg
 [3247 rows x 1 columns],
 'Cristiano':                          mentions
 id
 1358137564587319299             0
 1357379984399212545             0
 1356733030962987008             0
 1355924395064233986             0
 1355599316300292097             0
 ...                           ...
 32514882561638401     thejocksays
 32513604662071296            elgc
 32511823722840064        dytrogen
 32510294081146881    realchenchen
 32508748819857410     yellow_pine

 [3198 rows x 1 columns],
 'elonmusk':                           mentions
 id
 1357991946082418690               0
 1357973565413367808          dumdin
 1357972904663687173        grimezsz
 1357970517165182979               0
 1357964347813687296  kristennetten
 ...                             ...
 1242900612897005571         flcnhvy
 1242899515268648962         enscand
 1242893395338674176        ppathole
 1242881868426612736        ppathole
 1242881125049085956         flcnhvy

 [3239 rows x 1 columns]}
```

In [190…

```
tweets["AOC"].head()
mentions["AOC"].head()
new = mentions["AOC"].merge(tweets["AOC"], how = 'left', on = 'id').reset_index().loc[:,['
new.head()
```

Out[190…

```
mentions
0                 0.402836
                  1.840000
_pamcampos        3.866667
_vulvarine        2.000000
_waleedshahid    -0.650000
Name: polarity, dtype: float64
```

Question 4f

Finally, with our polarity column in place, we can finally explore how the sentiment of each tweet relates to the user(s) mentioned in it.

Complete the following function `mention_polarity` that takes in a mentions dataframe `mentions` and the original tweets dataframe `df` and returns a series where the mentioned users are the index and the corresponding mean sentiment scores of the tweets mentioning them are the values.

Hint: You should consider joining tables together in this question.

In [193…

```
def mention_polarity(df, mention_df):
    joined_table = mention_df.merge(df,how = 'left', on = 'id')
    return joined_table.reset_index().loc[:,['mentions','polarity']].groupby('mentions').m
aoc_mention_polarity = mention_polarity(tweets["AOC"],mentions["AOC"]).sort_values(ascendi
aoc_mention_polarity
```

Out[193…
```
mentions
booker              15.4
davidscottjaffe     12.6
johnkerry           11.4
mjacobs             11.3
penasays            10.8
                    ...
amber_jane_         -7.6
meggiebaer          -8.6
hawleymo            -8.9
scotthech          -10.8
repmarktakano      -10.8
Name: polarity, Length: 984, dtype: float64
```

In [194…
```
grader.check("q4f")
```

Out[194…

q4f passed!

Question 4g

When grouping by mentions and aggregating the polarity of the tweets, what aggregation function should we use? What might be some drawbacks of using the mean?

When grouping by mentions and aggregating the polarity of the tweets, we should use mean function. Some drawbacks of using the mean will be because of outliers than can significantly change the mean value.

Question 5: You Do EDA

Congratulations! You have finished all of the preliminary analysis on AOC, Cristiano, and Elon Musk's recent tweets.

As you might have recognized, there is still far more to explore within the data and build upon what we have uncovered so far. In this open-ended question, we want you to come up with a new perspective that can expand upon our analysis of the sentiment of each tweet.

For this question, you will perform some text analysis on our `tweets` dataset. Your analaysis should have two parts:

1. a piece of code that manipulates `tweets` in some way and produces informative output (e.g. a dataframe, series, or plot)
2. a short (4-5 sentence) description of the findings of your analysis: what were you looking for? What did you find? How did you go about answering your question?

Your work should involve text analysis in some way, whether that's using regular expressions or some other form.

To aid you in creating plots, we provide the plotting helper functions in the table below. These are same helpers we have used throughout this notebook, and all accept dictionaries with a similar structure to `tweets`. That being said, if you know how to make plots, please do so! You'll be learning how to use the libraries that we're using in the helpers starting next week.

Helper	Description
`make_bar_plot`	Plot side-by-side bar plots of data like `plt.bar`
`make_histogram`	Plot overlaid histograms of data like `plt.hist`
`make_line_plot`	Plot overlaid line plots of data like `plt.plot`
`make_scatter_plot`	Plot overlaid scatter plots of data like `plt.scatter`

Each of the provided helpers is in `ds100_utils.py` and has a comprehensive docstring. You can read the docstring by calling `help` on the plotting function:

In [195…

```
help(make_line_plot)
```

```
Help on function make_line_plot in module ds100_utils:

make_line_plot(df_dict, x_col, y_col, include=None, title=None, xlabel=None, ylabel=None,
legend=True)
    Plot a line plot of two columns for each dataframe in `df_dict`.

    Uses `sns.lineplot` to plot a line plot of two columns for each
    dataframe in `df_dict`. The keys of `df_dict` are used as entries in
    the legend when `legend` is `True`.

    Parameters
    ----------
        df_dict: dict[str: pd.DataFrame]
            a dictionary mapping handles to dataframes with the data to plot
        x_col: str
            the name of a column in each dataframe in `df_dict` to plot on
            the x-axis
        y_col: str
            the name of a column in each dataframe in `df_dict` to plot on
            the y-axis
        include: list[str], optional
            a list of handles to include in the plot; all keys in `df_dict` not
            present in `include`, if specified, will *not* be included in the plot
        title: str, optional
            a title for the plot
        xlabel: str, optional
            a label for the x-axis; if unspecified, `x_col` is used
        ylabel: str, optional
            a label for the y-axis; if unspecified, `y_col` is used
        legend: bool, optional
            whether to include a legend with each key in `df_dict`
```

To assist you in getting started, here are a few ideas for this you can analyze for this question:

- dig deeper into when devices were used
- how sentiment varies with time of tweet
- expand on regexes from 4b to perform additional analysis (e.g. hashtags)
- examine sentiment of tweets over time

In general, try to combine the analyses from earlier questions or create new analysis based on the scaffolding we have provided.

This question is worth 4 points and will be graded based on this rubric:

	2 points	1 point	0 points
Code	Produces a mostly informative plot or pandas output that addresses the question posed in the student's description and uses at least one of the following pandas dataframe/series methods: `groupby` , `agg` , `merge` , `pivot_table` , `str` , `apply`	Attempts to produce a plot or manipulate data but the output is unrelated to the proposed question, or doesn't utilize at least one of the listed methods	No attempt at writing code
Description	Describes the analysis question and procedure comprehensively and summarizes results correctly	Attempts to describe analysis and results but description of results is incorrect or analysis of results is disconnected from the student's original question	No attempt at writing a description

Question 5a

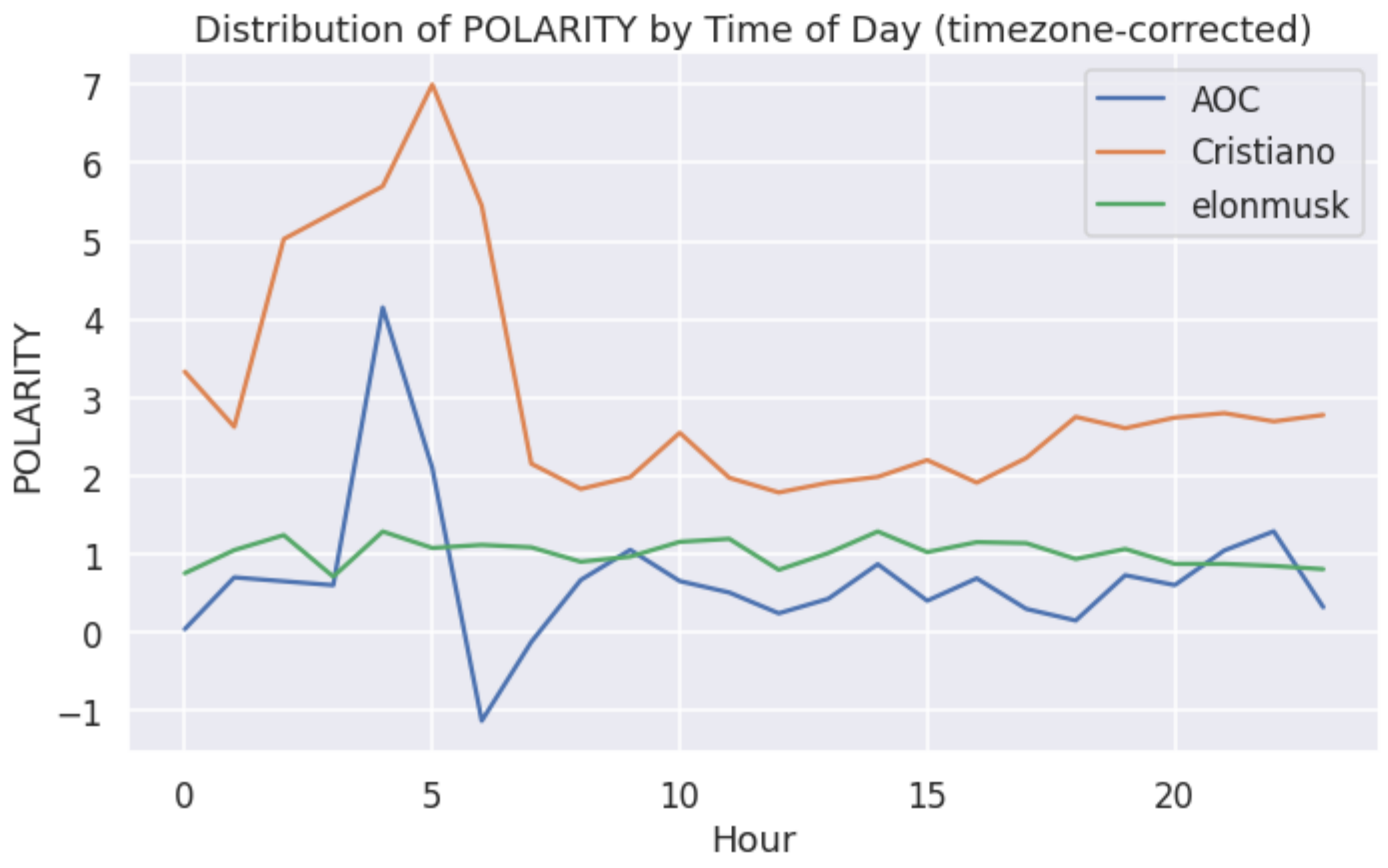
Use this space to put your EDA code.

In [196…

```
def hour_polarity(df):
    df = df.loc[:,['converted_hour','polarity']]
    df['converted_hour'] = df['converted_hour'].astype(int)
    return df.sort_values(by = ['converted_hour']).groupby('converted_hour').mean()

pol_hours = {handle: hour_polarity(df) for handle, df in tweets.items()}

make_line_plot(pol_hours, 'converted_hour', 'polarity', title="Distribution of POLARITY by
                xlabel="Hour", ylabel="POLARITY")
```



Question 5b