---

### 0.0.1 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.

A marketing research firm might be interested in analyzing tweets for a better understand how customer's hobby trend. They can know what are people's interest and made decision on their action, by collecting data. They also can sent specific advertisments to people who is intrested in it, so it is a very useful tool.

---

#### 0.0.2 Question 2e

What might we want to investigate further based on the plot in 2d above? Write a few sentences below.

We could check when they post the messages. We got "created_at" column for times, so we can do some analyse there. We may find when they usually post. I noticed "favorite_count" column. We could check which kinda of message people like. "possibly_sensitive" also sounds interesting. maybe we could know how is always post sensitive message.

---

#### 0.0.3 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?

Proportions can show a better view of how ecah one used resource. It is a better way to do comparison. If two people got same number on "Twitter for iPhone", but for one of them, he just used it as 1% of his total quantity. Mabe we can see different things.

---

#### 0.0.4 Question 3b

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?

From 0-6 AOC and Elon Musk usually post a lot, but Cristiano usually do not post at this time. From 6-12, AOC don't post at this time, all other two post. But during this time period, Elon Musk seems to lack of vitality comparing other times. From 12 to 24 all of them are active during this time period. I can tell they have different schedual. They may just sleep in different time during a day or they are in different time zone. For Musk, he post 24 hour. Maybe he sleep randomly. Rich live maybe buzy or maybe just happy all the time, hard to tell. I think data plotted above seem reasonable.

---

### 0.0.5 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?

police 1 order 0 Democrat 0 Republican 0 gun -1 dog 1 technology 0 TikTok 0 security 0 face-mask -1 science 0 climate change 0 vaccine 0

These are all noun, it is hard for me to tell if they are positive or negitive.

---

#### 0.0.6 Question 4g

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

I think median function maybe better. Because mean function will be affected by extreamly large or extreamly small number. Maybe we could see a better picture by using median function.

---

### 0.0.7 Question 5a

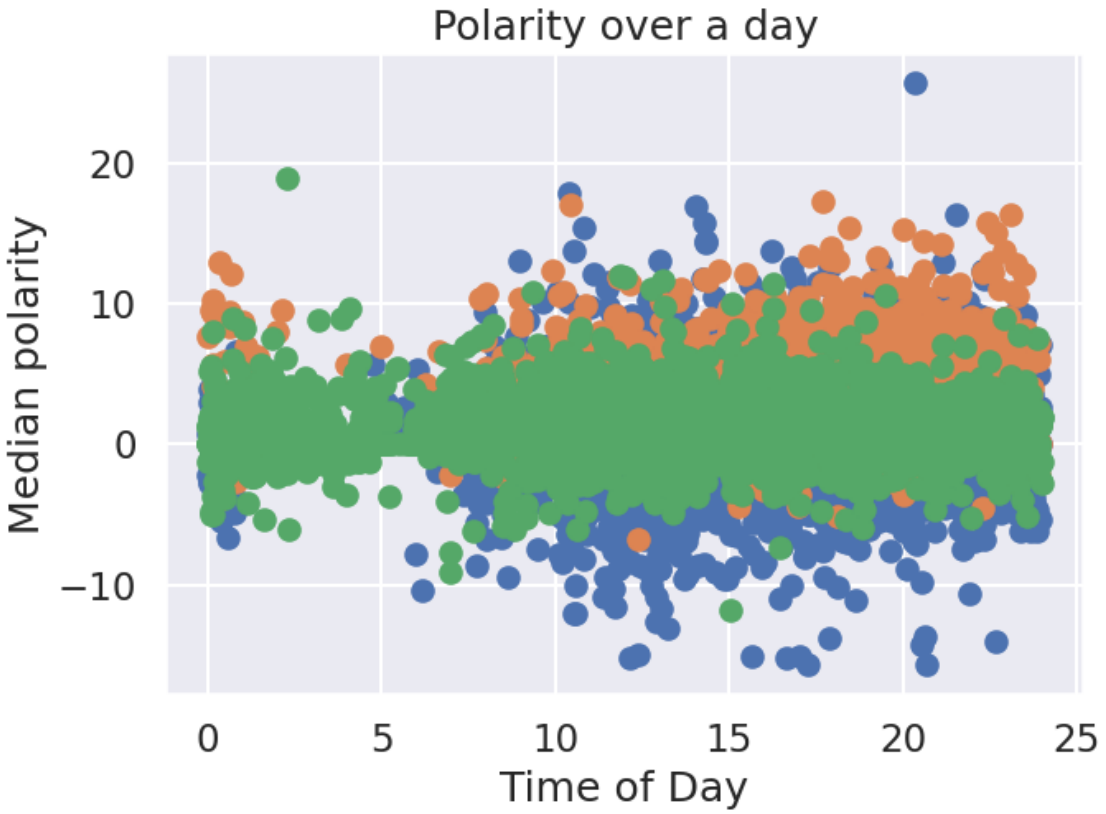
Use this space to put your EDA code.

```
In [113]: def mention_polarity_median_plot(df, mention_df color):
              #return scatter plot hour of a dat vs polarity
              df_q5 = df.join(mention_df)[["converted_hour", "polarity", "mentions"]]
              group_q5 = df_q5.groupby("id").median()

              plt.title("Polarity over a day")
              plt.xlabel("Time of Day")
              plt.ylabel("Median polarity")

              x = group_q5['converted_hour']
              y = group_q5['polarity']
              plt.scatter(x, y, c = color)

          aoc_mention_polarity = mention_polarity_median_plot(tweets["AOC"], mentions["AOC"])
          Cristiano_mention_polarity = mention_polarity_median_plot(tweets["Cristiano"], mentions["Cris
          elonmusk_mention_polarity = mention_polarity_median_plot(tweets["elonmusk"], mentions["elonmu
```


Polarity over a day
Median polarity
20
10
0
−10
0
5
10
15
20
25
Time of Day

---

#### 0.0.8 Question 5b

Use this space to put your EDA description.

I am trying to look at who is more positive and who is more negative. I also want to check if the times will affect polarity.

So I first join two table inorder to get “converted_hour”, “polarity”, “mentions” columns and then I grouped them by ‘id’ index to get table that shows every hour and polarity. Then I draw the scatter plot and see what’s the different. Blue dots are for AOC, Orange dots are for Cristiano, Green dots are for elonmusk

It seems everybody tends to be positive except AOC. She is neutral. Her -10 to -15 range spreads more than 10 to 15 range. It is hard to tell if the times affect polarity. Druring 1 - 6 am of the day, they are in neutral state. After that they tend to be “emotional”. This may cause by they sleep during 1-6 am.