group_proj

Rachel Montgomery

2022-11-02

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v dplyr   1.0.9
## v tibble  3.1.8     v stringr 1.4.1
## v tidyr   1.2.0     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)

library(tidytext)
library(textdata)
```

```
library(readr)
UFO_and_Weather <- read_csv("UFO_and_Weather.csv",
    col_types = cols(month = col_character(),
        hour = col_time(format = "%H")))
```

```
## New names:
## * `` -> `...1`
```

```
View(UFO_and_Weather)
```

```
#making new column with month names

UFO_and_Weather <- UFO_and_Weather %>%
  mutate(Month= case_when(
    month=="1" ~ "January",
    month=="2" ~ "February",
    month=="3" ~ "March",
    month=="4" ~ "April",
    month=="5" ~ "May",
    month=="6" ~ "June",
    month=="7" ~ "July",
    month=="8" ~ "August",
    month=="9" ~ "September",
    month=="10" ~ "October",
    month=="11" ~ "November",
    month=="12" ~ "December"))
```

Goals: Are UFO sightings more prevalent during certain months of the year? What time of day do UFO sightings occur? Are reports less reliable in these times (e.g. Friday/Saturday night, low vision) How long does the average UFO sighting last?

```
#Understanding what the data represents
View(UFO_and_Weather)

head(UFO_and_Weather)
```

```
## # A tibble: 6 x 19
##    ...1 city   state date_time           shape text  city_~1 city_~2  year month
##   <dbl> <chr>  <chr> <dttm>              <chr> <chr>   <dbl>   <dbl> <dbl> <chr>
## 1     0 Chest~ VA    2019-12-12 18:43:00 light My w~    37.3   -77.4  2019 12
## 2     1 Rocky~ CT    2019-03-22 18:30:00 circ~ I th~    41.7   -72.6  2019 3
## 3     2 Ottawa ON    2019-04-17 02:00:00 tear~ I wa~    45.4   -75.7  2019 4
## 4     3 Kirby~ TX    2019-04-02 20:25:00 disk  The ~    30.7   -94.0  2019 4
## 5     4 Tucson AZ    2019-05-01 11:00:00 unkn~ Desc~    32.3  -111.   2019 5
## 6     5 Gold ~ AZ    2019-04-10 17:00:00 circ~ Apr.~    33.4  -111.   2019 4
## # ... with 9 more variables: day <dbl>, hour <time>, temperature <dbl>,
## #   relative_humidity <dbl>, precipitation <dbl>, snow <lgl>,
## #   wind_direction <dbl>, wind_speed <dbl>, Month <chr>, and abbreviated
## #   variable names 1: city_latitude, 2: city_longitude
```

```
dim(UFO_and_Weather)
```

```
## [1] 22482    19
```

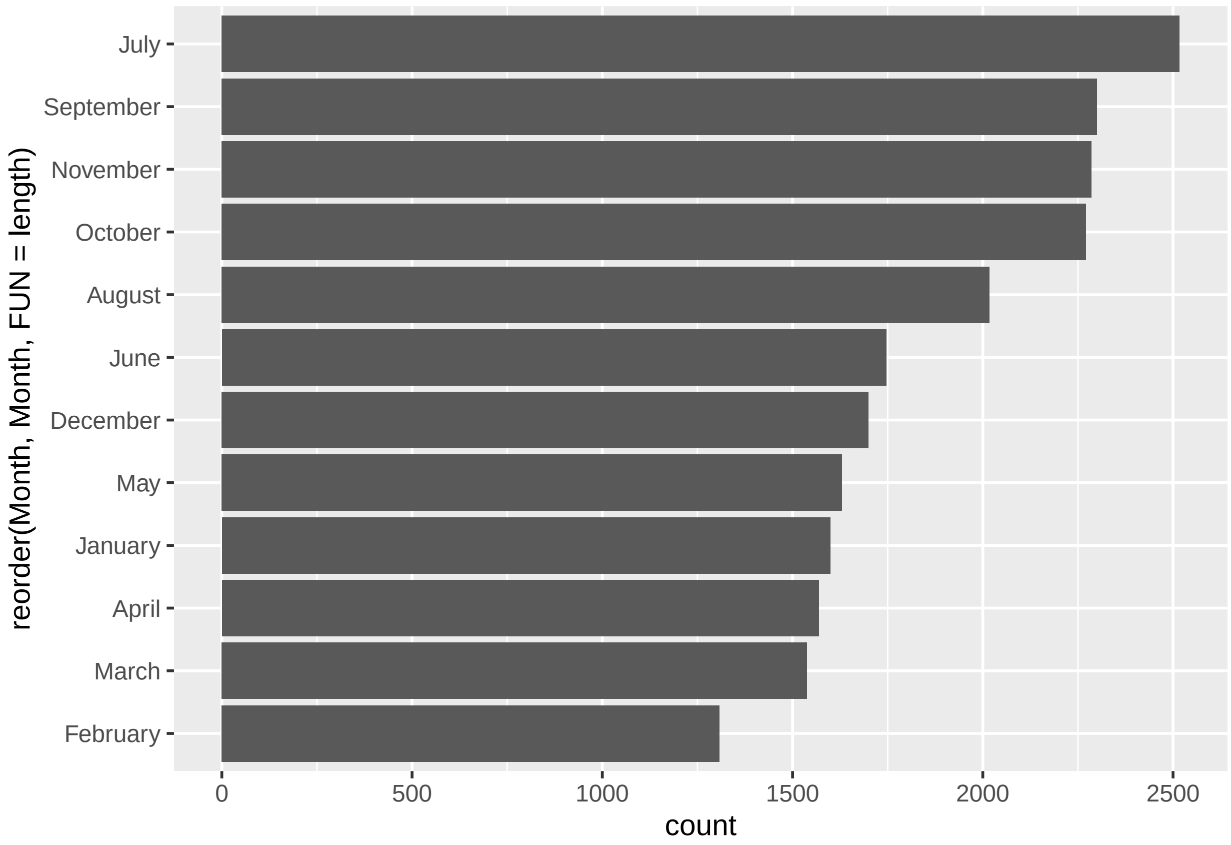
```
str(UFO_and_Weather)
```

```
## tibble [22,482 x 19] (S3: tbl_df/tbl/data.frame)
##  $ ...1             : num [1:22482] 0 1 2 3 4 5 6 7 8 9 ...
##  $ city             : chr [1:22482] "Chester" "Rocky Hill" "Ottawa" "Kirbyville" ...
##  $ state            : chr [1:22482] "VA" "CT" "ON" "TX" ...
##  $ date_time        : POSIXct[1:22482], format: "2019-12-12 18:43:00" "2019-03-22 18:30:00" ...
##  $ shape            : chr [1:22482] "light" "circle" "teardrop" "disk" ...
##  $ text             : chr [1:22482] "My wife was driving southeast on a fairly populated main side r
##  $ city_latitude    : num [1:22482] 37.3 41.7 45.4 30.7 32.3 ...
##  $ city_longitude   : num [1:22482] -77.4 -72.6 -75.7 -94 -110.9 ...
##  $ year             : num [1:22482] 2019 2019 2019 2019 2019 ...
##  $ month            : chr [1:22482] "12" "3" "4" "4" ...
##  $ day              : num [1:22482] 12 22 17 2 1 10 18 12 11 15 ...
##  $ hour             : 'hms' num [1:22482] 18:00:00 18:00:00 02:00:00 20:00:00 ...
##   ..- attr(*, "units")= chr "secs"
##  $ temperature      : num [1:22482] 3.9 5.6 NA 20 13.9 19 NA NA 26 33.9 ...
##  $ relative_humidity: num [1:22482] 46 82 NA 26 55 21 NA NA 23 38 ...
##  $ precipitation    : num [1:22482] 0 0.8 NA 0 0 NA NA NA NA 0 ...
##  $ snow              : logi [1:22482] NA NA NA NA NA NA ...
##  $ wind_direction    : num [1:22482] NA 310 NA 350 180 260 NA NA 350 150 ...
##  $ wind_speed        : num [1:22482] 9.4 9.4 NA 7.6 0 27.7 NA NA 16.6 25.9 ...
##  $ Month             : chr [1:22482] "December" "March" "April" "April" ...
```

22,482 obs and 18 variables.

#1 Are UFO sightings more prevalent during certain months of the year?

```
ggplot(UFO_and_Weather, aes(x=reorder(Month, Month, FUN=length)))+
  geom_bar()+
  coord_flip()
```

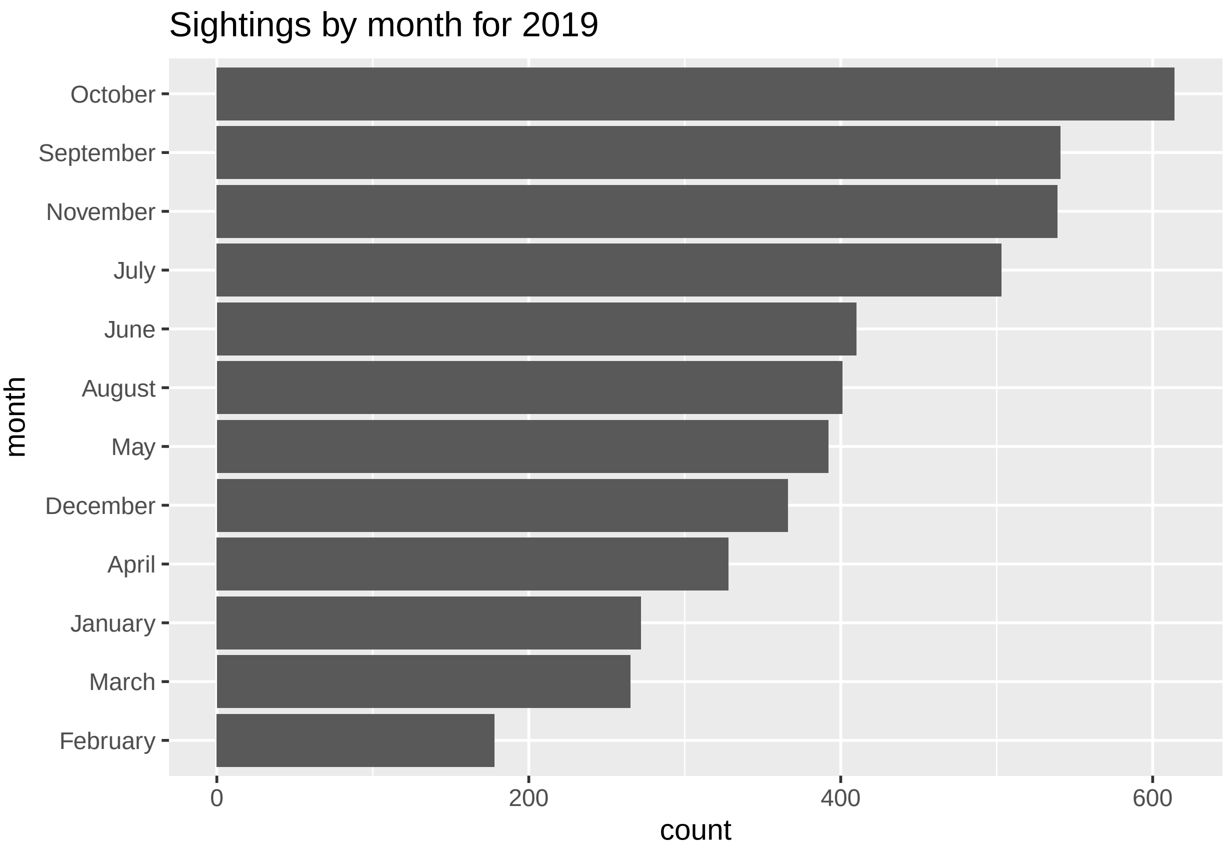


```
#July has the most
```

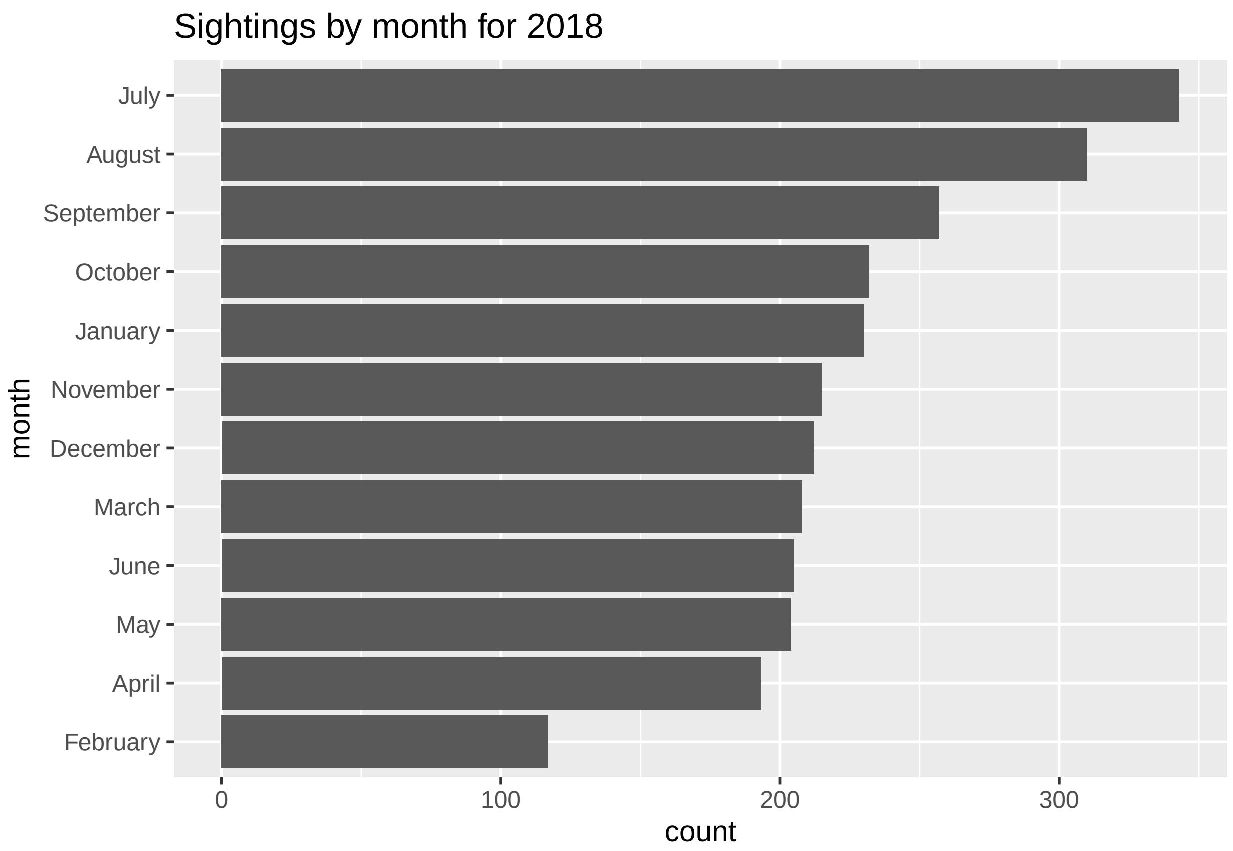
We can see that July has the most by quite a good margin. When we examine year by year, for 4/5 years July has either the most or the 2nd most sightings. This has me think that we should plot the number of sightings per each day in July. Maybe this is because of the 4th of july - a lot of fireworks and air shows happening? Because its summer more people are out and about and looking at the sky??

```
#let's see if this pattern of mostly being in the fall is a yearly pattern!


weather_2019 <- UFO_and_Weather %>%
    filter(year == 2019)
```

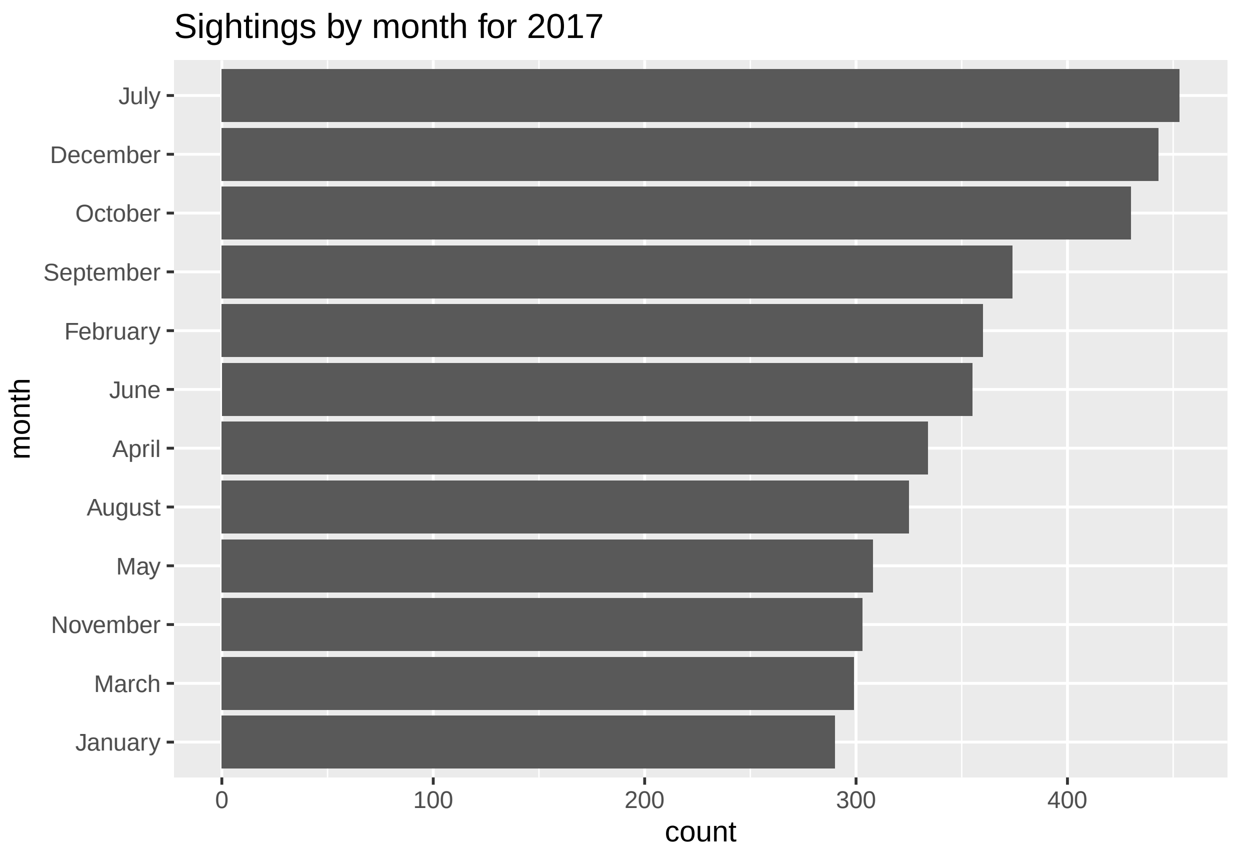
```
#plotting
ggplot(weather_2019, aes(x=reorder(Month, Month, FUN=length)))+
  geom_bar()+
  labs(title="Sightings by month for 2019", x="month", y="count")+
  coord_flip()
```



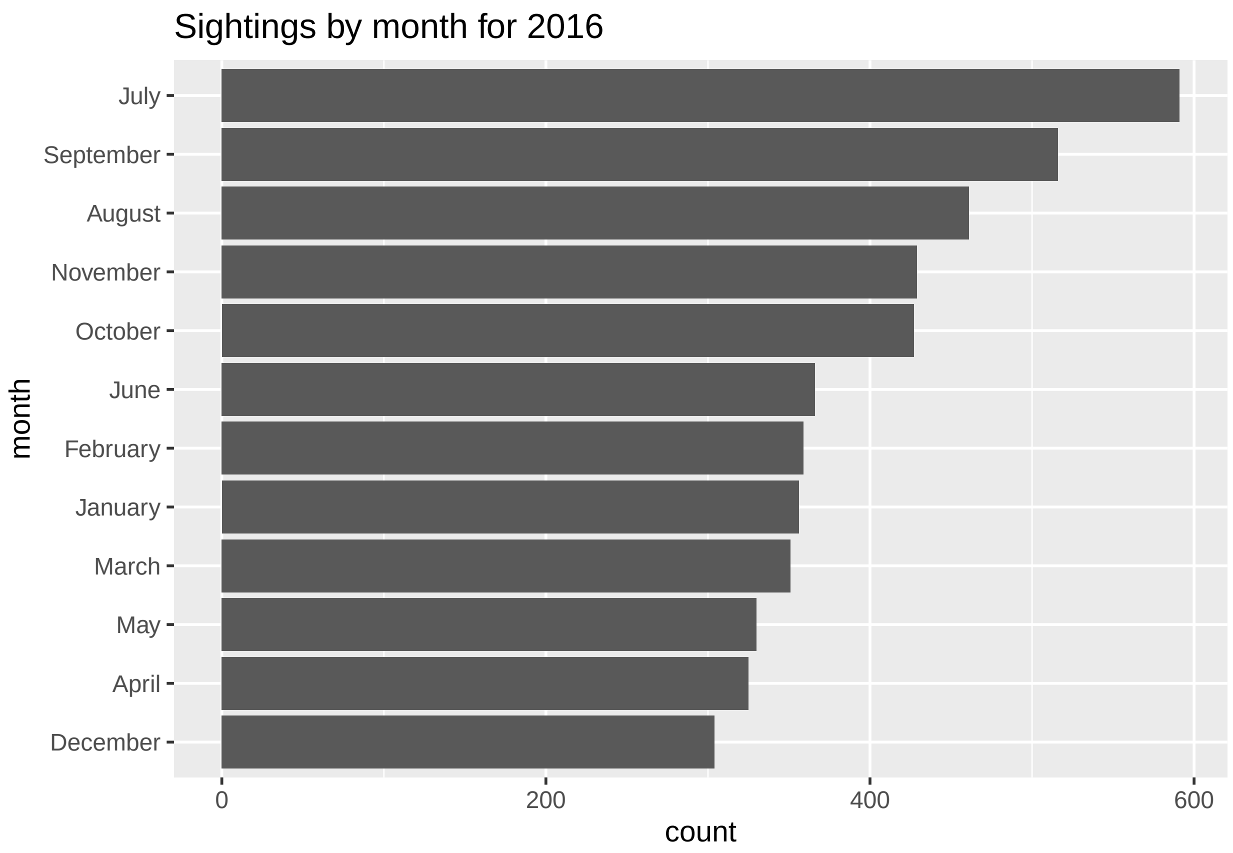
```
weather_2018 <- UFO_and_Weather %>%
    filter(year == 2018)
#plotting
ggplot(weather_2018, aes(x=reorder(Month, Month, FUN=length)))+
  geom_bar()+
  labs(title="Sightings by month for 2018", x="month", y="count")+
  coord_flip()
```



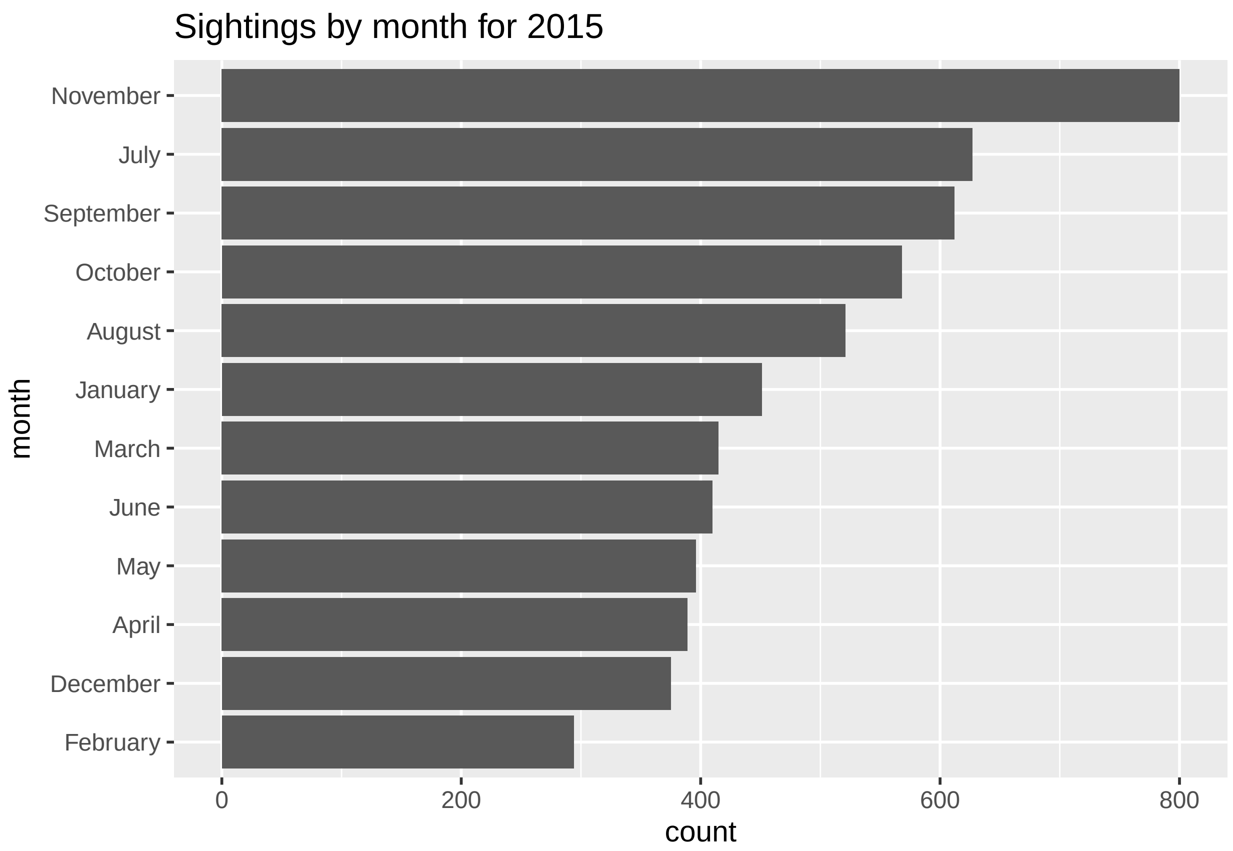
```
weather_2017 <- UFO_and_Weather %>%
    filter(year == 2017)
#plotting
ggplot(weather_2017, aes(x=reorder(Month, Month, FUN=length)))+
  geom_bar()+
  labs(title="Sightings by month for 2017", x="month", y="count")+
  coord_flip()
```



```
weather_2016 <- UFO_and_Weather %>%
    filter(year == 2016)
#plotting
ggplot(weather_2016, aes(x=reorder(Month, Month, FUN=length)))+
  geom_bar()+
  labs(title="Sightings by month for 2016", x="month", y="count")+
  coord_flip()
```

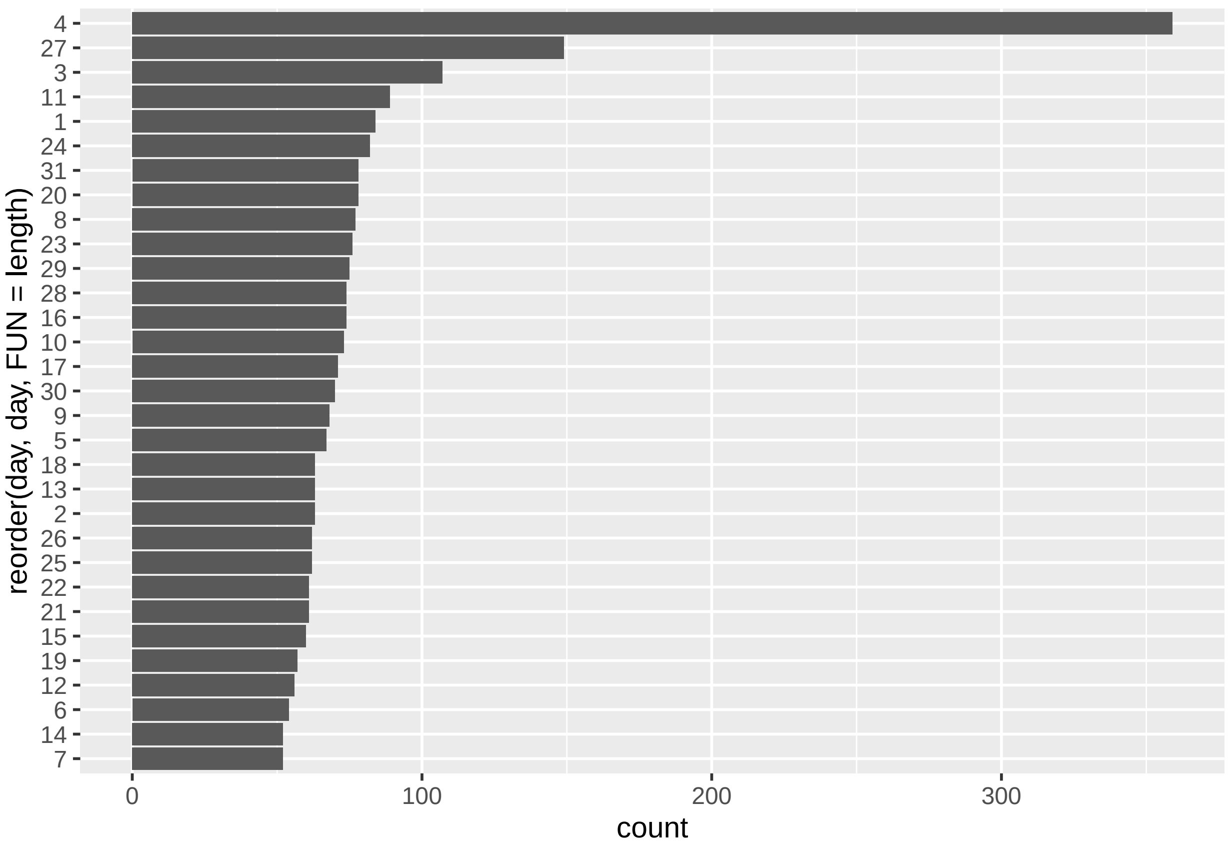


```
weather_2015 <- UFO_and_Weather %>%
    filter(year == 2015)
#plotting
ggplot(weather_2015, aes(x=reorder(Month, Month, FUN=length)))+
  geom_bar()+
  labs(title="Sightings by month for 2015", x="month", y="count")+
  coord_flip()
```



let's look at just July

```
just_july <- UFO_and_Weather %>%
  filter(month== 7)

ggplot(just_july, aes(x=reorder(day, day, FUN=length))+
  geom_bar()+
  coord_flip()
```

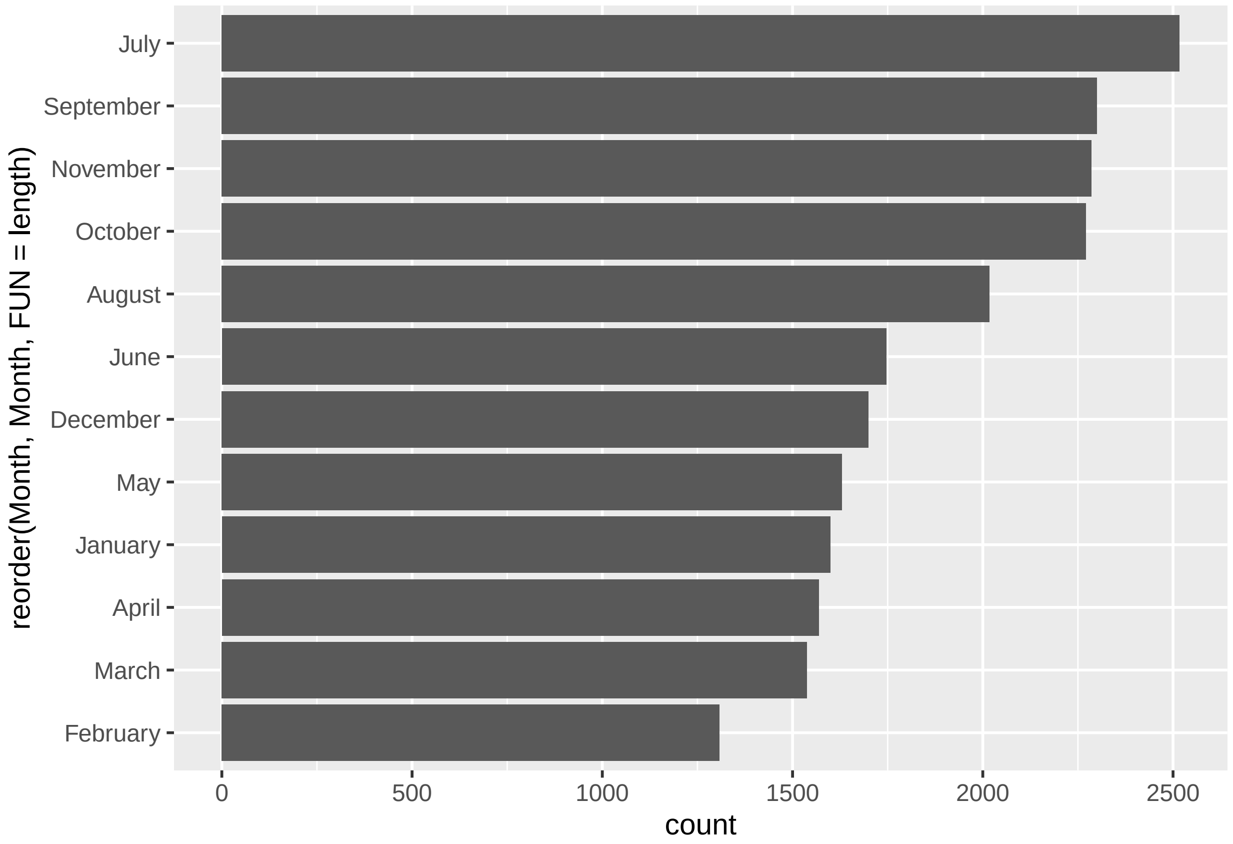


```
#July has the most
```

Wow!!! The fourth of july has about 2.5x more than all other days!! Some possible explanations for this are because there are a ton of fireworks going off, which cause light and sound to be in the sky, and could be mistaken for UFOS. Also there are air shows that occur, which could also be reported as UFOS. Because of these festivities, many people are out at night and looking at the sky, so mere exposure effect could be play. Additionally, it should be noted that this is a national holiday, and one that is often celebrated with alcohol, which conflicts the reliability of these reports.

Overall, the 4th of July is an outlier and should be treated as such in following analysis.

```
ggplot(UFO_and_Weather, aes(x=reorder(Month, Month, FUN=length)))+
  geom_bar()+
  coord_flip()
```



```
#July has the most
```

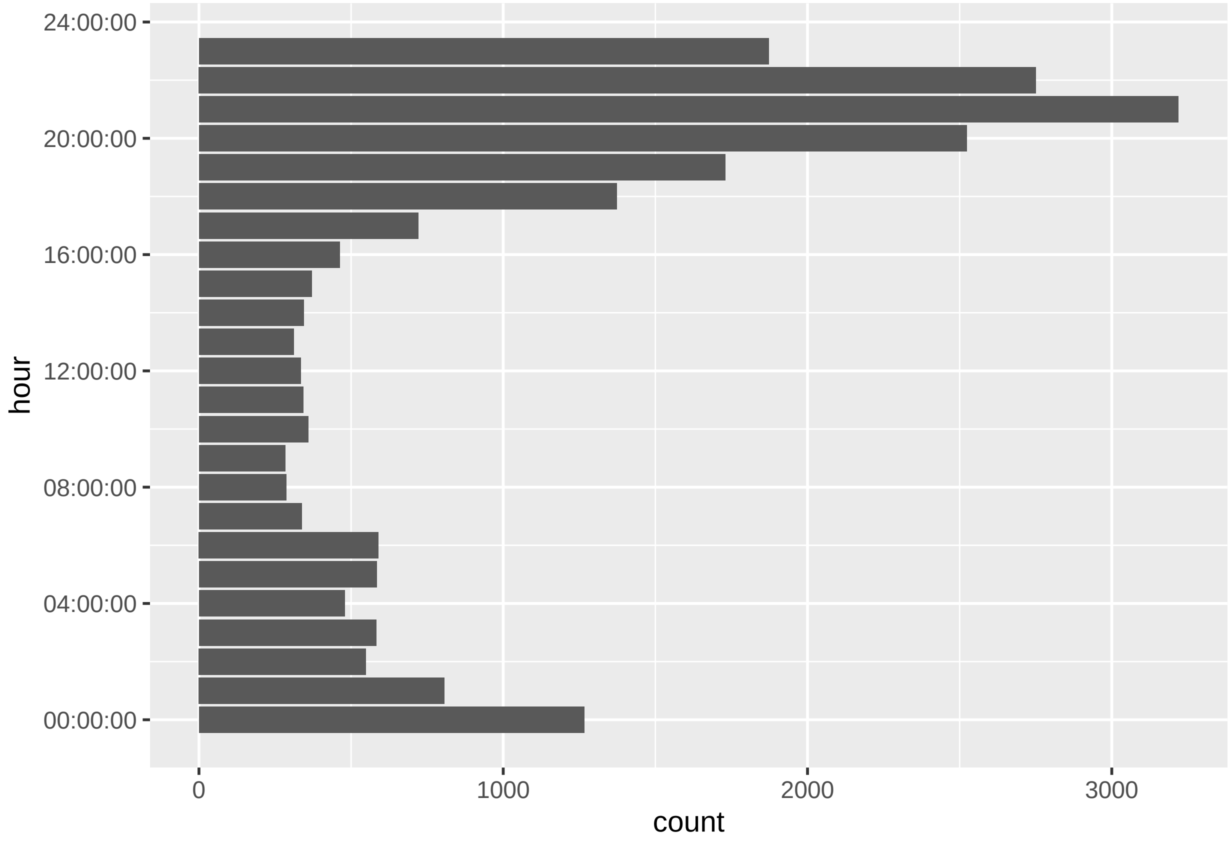
1.2 Is there a most popular season for UFO sightings?

```
#first creating a season variable
UFO_and_Weather <- UFO_and_Weather %>%
  mutate(season= case_when(
    month=="12" | month=="1" | month=="2" ~ "Winter"   ,
    month=="3" |  month=="4" | month=="5" ~ "Spring",
    month=="6" | month=="7" | month=="8" ~ "Summer",
    month=="9" | month=="10" |  month=="11" ~ "Fall"))

#spring is march - may (3,4,5)
#summer is june to august (6,7,8)
#fall is september to november (9,10,11)
#winter is december to february (12, 1, 2)
```

#2 What time of day do UFO sightings occur?

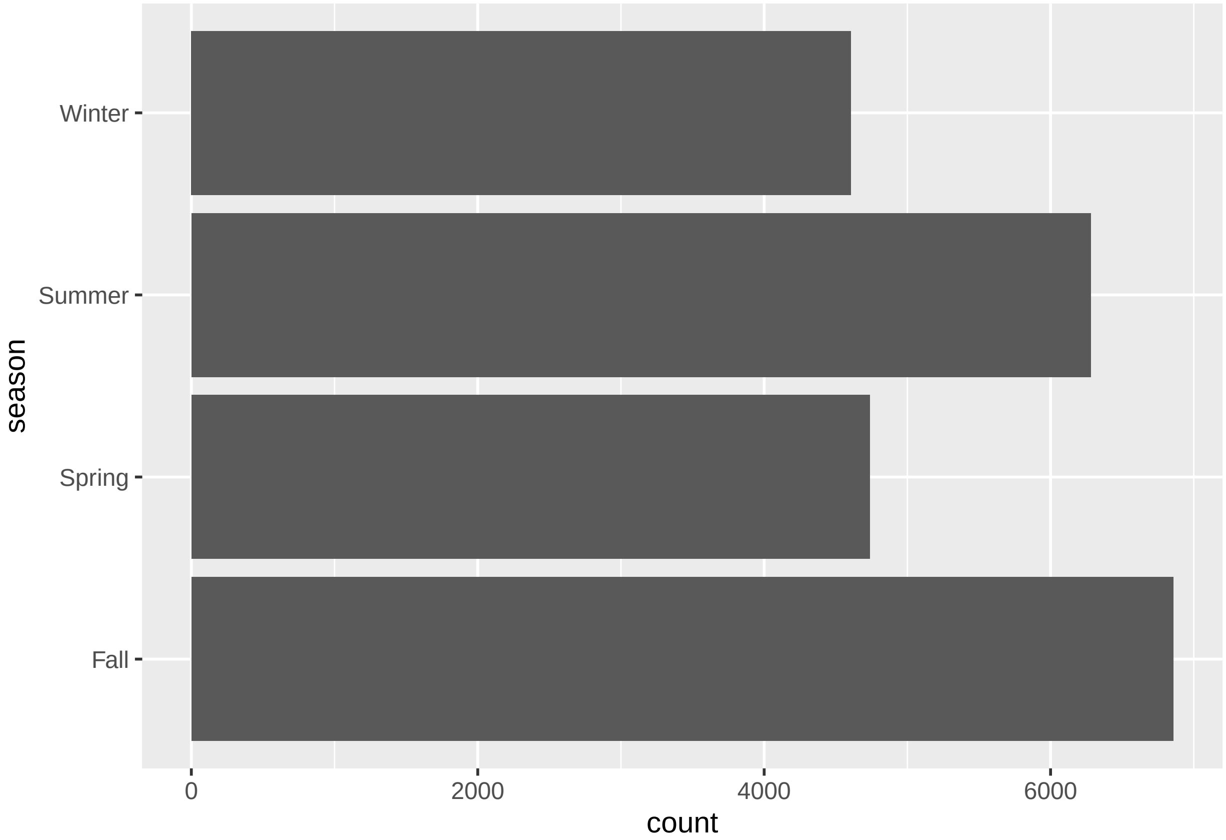
```
#first making time into am and pm hour times
UFO_and_Weather <- UFO_and_Weather %>%
  mutate(time_of_day= case_when(
    hour >=6 & hour <= 18 ~ "day",
    hour < 5  ~ "night",
    hour >18 ~ "night"))

#next, making day and night variables
#day is 6 am to 6 pm (6  to 18)
#night is 12 am to 6am, and 6 pm to 12 pm (0-6, and 18-24)
UFO_and_Weather <- UFO_and_Weather %>%
  mutate(time_of_day= case_when(
    hour >=6 & hour <= 18 ~ "day",
    hour < 5  ~ "night",
    hour >18 ~ "night"))
```

```
ggplot(UFO_and_Weather, aes(x=hour))+
  geom_bar()+
  coord_flip()
```



We can see that the most sightings are at night, specifically between the hours of 6 pm and 11 pm.

```
ggplot(UFO_and_Weather, aes(x=season))+
  geom_bar()+
  coord_flip()
```



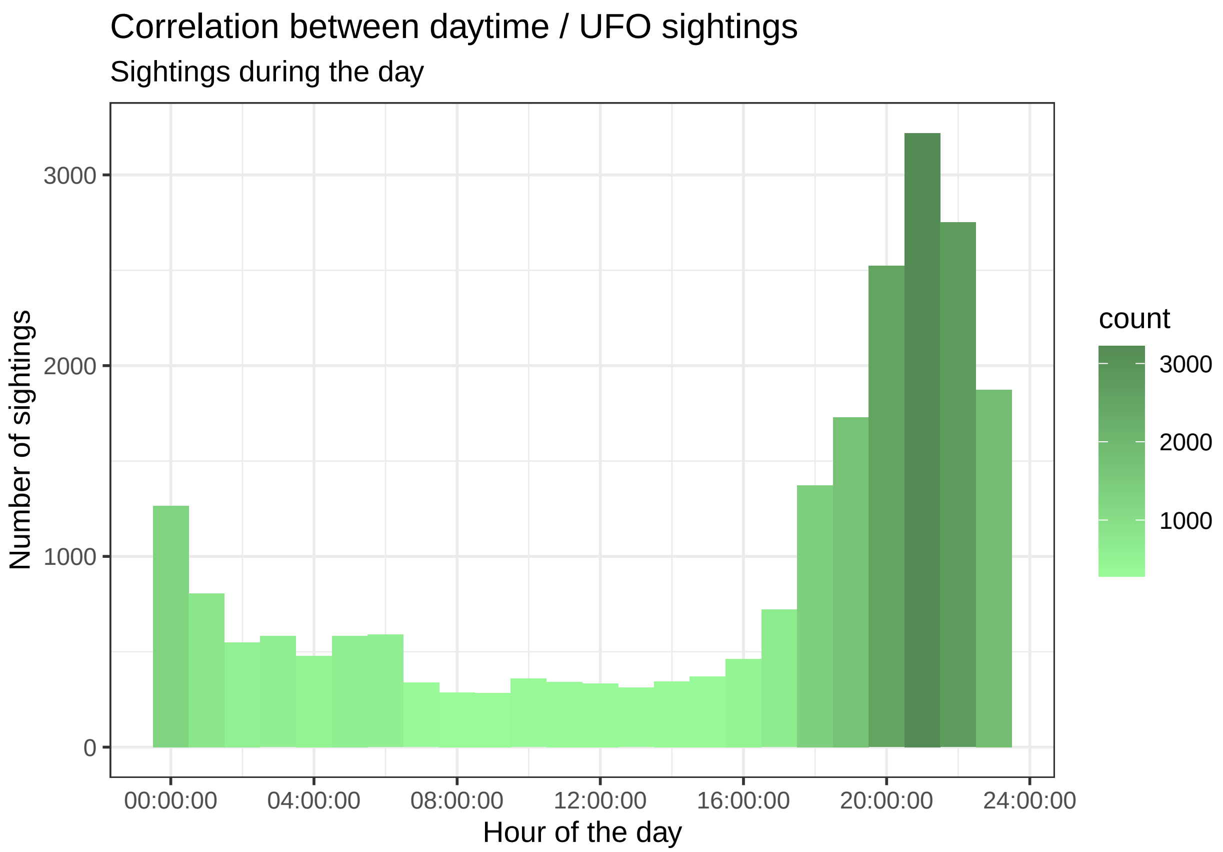
We can see that the fall months have the most sightings, with summer having the 2nd most.

avg wind speed vs count (and do that for months or season)

#3 How long does the average UFO sighting last? We can't do this because the duration variable was taken out.

#4 Is there a correlation between the UFO sightings and the time? You can see if there's any correlation between the number of sightings and the time of day. Also, you could test with the month and the year.

```
ggplot(UFO_and_Weather, aes(x=hour)) +
  geom_histogram(bins=24, aes(fill=..count..)) +
  theme_bw() +
  scale_fill_gradient(low = "palegreen", high = "palegreen4") +
  labs(x = "Hour of the day", y = "Number of sightings",
       title="Correlation between daytime / UFO sightings",
       subtitle = "Sightings during the day")
```



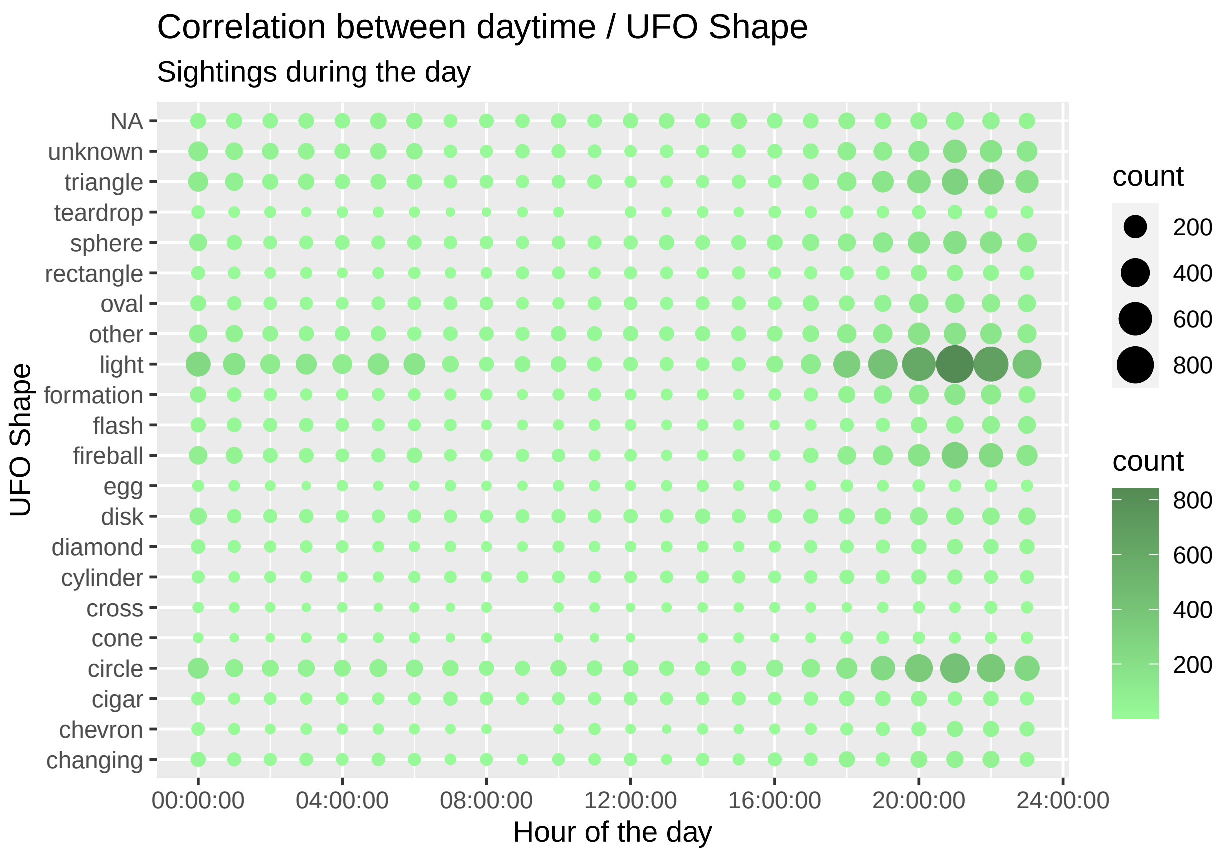
The histogram clearly shows that the majority of the sightings occur when there is no or very little light. However, it's worth noting that there have been reports of UFOs during the day as well.

#5 Is there a correlation between the UFO shapes and time? You can look at the relationship between the shapes and the time of day. This may explain why the light shape is the most frequently reported shape.

```
shapesDaytime <-
  UFO_and_Weather %>%
  group_by(hour, shape) %>%
  summarize(count=n());
```

```
## 'summarise()' has grouped output by 'hour'. You can override using the
## '.groups' argument.
```

```
ggplot(shapesDaytime, aes(x=hour, y=shape)) +
  geom_point(aes(color=count, size=count)) +
  scale_colour_gradient(low = "palegreen", high="palegreen4") +
  labs(x = "Hour of the day", y = "UFO Shape",
       title="Correlation between daytime / UFO Shape",
       subtitle = "Sightings during the day")
```



You can see from the plot that the shapes are more prevalent/persistent at night as well. We can see that light appears more frequently at night and in the evening, but less frequently during the day than other common shapes.

Now doing the “chi-square” test to see if there’s a link between the time of day and the shapes.

The Chi-Square independence test is performed, assuming that: Each sample observation is independent. Each case contributes to at least one entry.

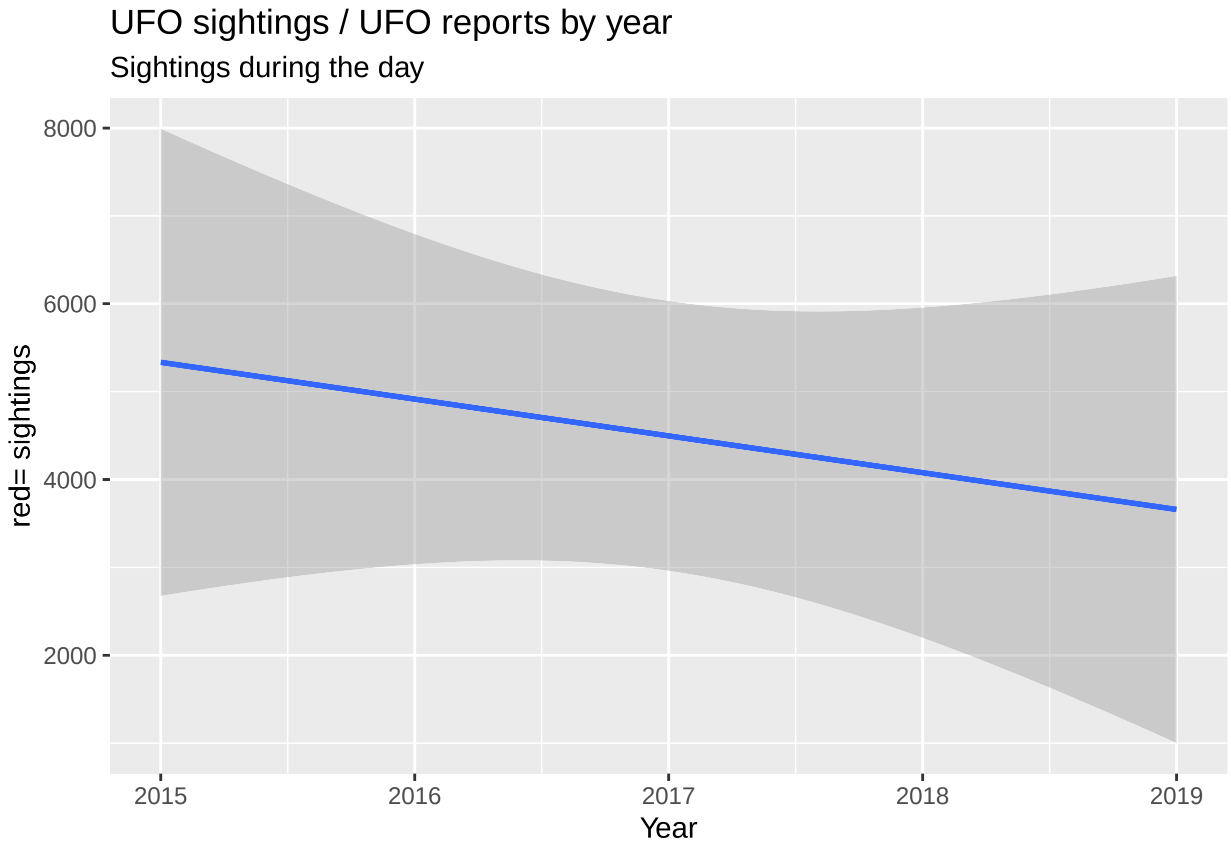
```
#chisq.test(UFO_and_Weather, hour, simulate.p.value=T);
```

this does not work :(

#6 Has the frequency of UFO sightings risen over time?

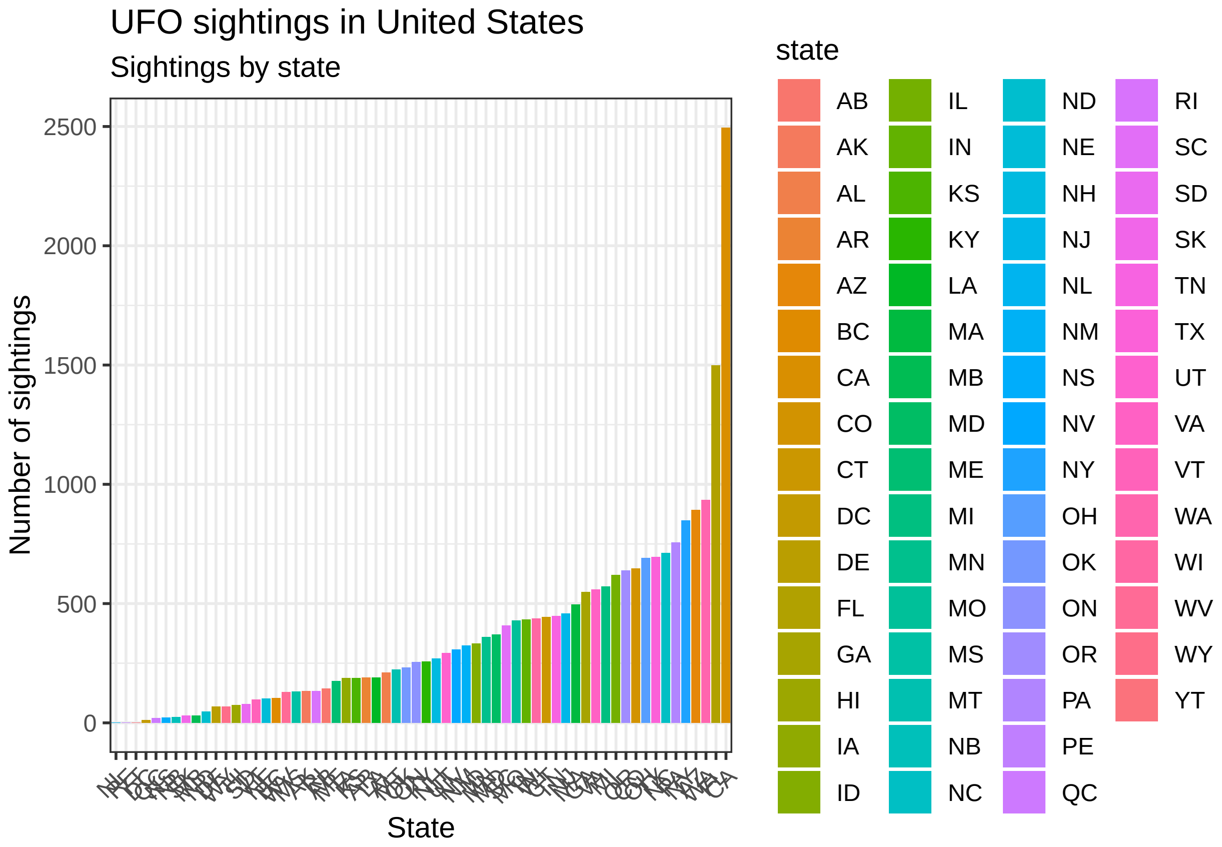
```
# SIGHTINGS BY YEAR
sightingsYear <-
  UFO_and_Weather %>% group_by(year) %>%
  summarize(count=n());

# REPORTS BY YEAR
reportsYear <-
  UFO_and_Weather %>% group_by(year) %>%
  summarize(count=n());
ggplot(sightingsYear, aes(x=year, y=count))  +
  geom_smooth(method="lm") +
  labs(x = "Year", y = "red= sightings",
       title="UFO sightings / UFO reports by year",
       subtitle = "Sightings during the day")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



#which states have the most findings?

```
# SIGHTINGS IN UNITED STATES
ggplot(UFO_and_Weather, aes(x=reorder(state, state, FUN=length), fill=state)) +
  stat_count() +
  theme_bw() +
  theme(axis.text.x = element_text(angle=45, size=9, hjust=1)) +
  labs(x = "State", y = "Number of sightings",
       title="UFO sightings in United States",
       subtitle = "Sightings by state")
```



To sum up, your findings demonstrate that the United States, namely the state of California, has the highest number of sightings. However, an important point to note is that, based on population density, the state of Washington has a higher density of sightings than California. California and Washington have both legalized marijuana for recreational use, which could explain why there have been so many sightings on the west coast.

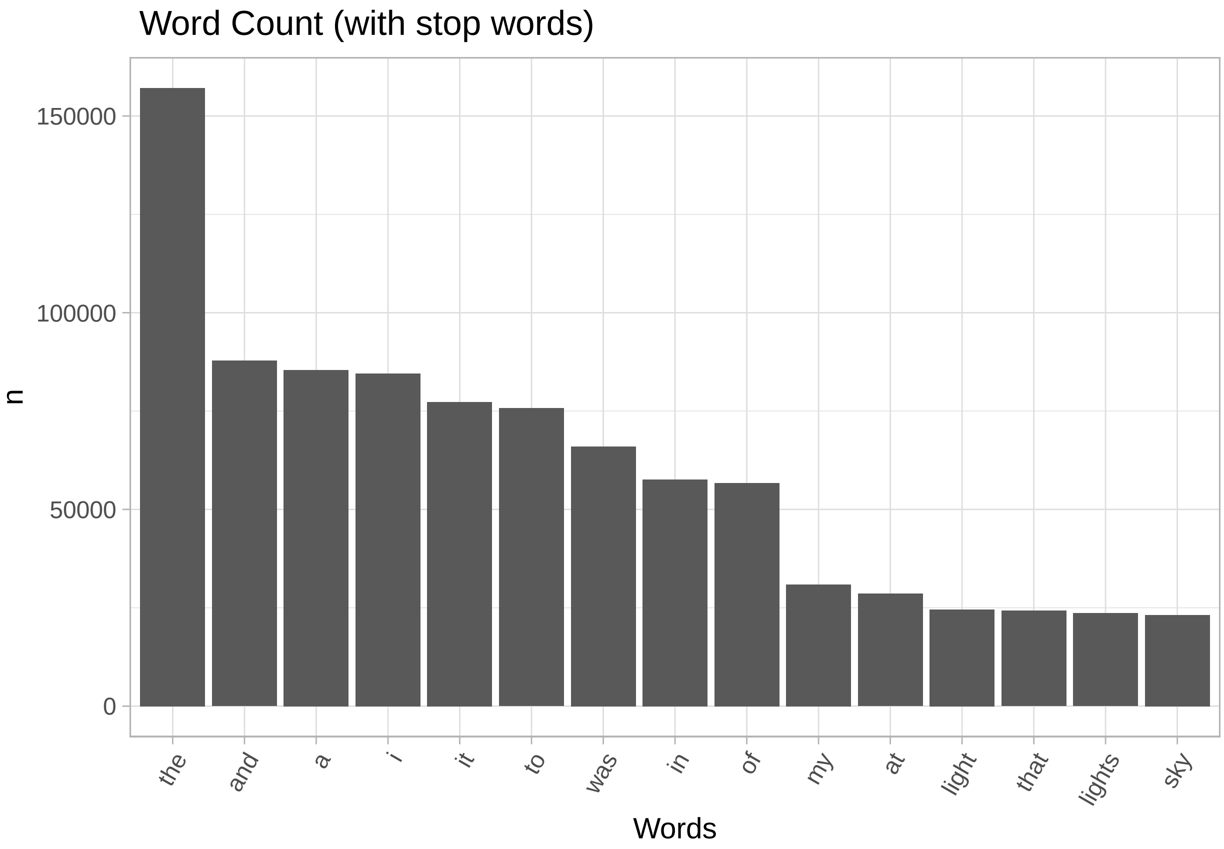
Text analysis

First, lets make a word frequency

```
#Step 1:tokenize corpus

words <- UFO_and_Weather %>%
  select(text) %>%
  unnest_tokens(word, text)

head(words)
```

```
## # A tibble: 6 x 1
##   word
##   <chr>
## 1 my
## 2 wife
## 3 was
## 4 driving
## 5 southeast
## 6 on
```

```
#Now, we'll generate a count of the words, sort by the number of times the word occurs, and then plot t

#x= season and poem=line

words %>% count(word, sort = T) %>% slice(1:15) %>%
  ggplot(aes(x = reorder(word, n, function(n) -n), y = n)) +
  geom_bar(stat = 'identity') +
  theme_light() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  xlab("Words") +
  ggtitle(" Word Count (with stop words)")
```



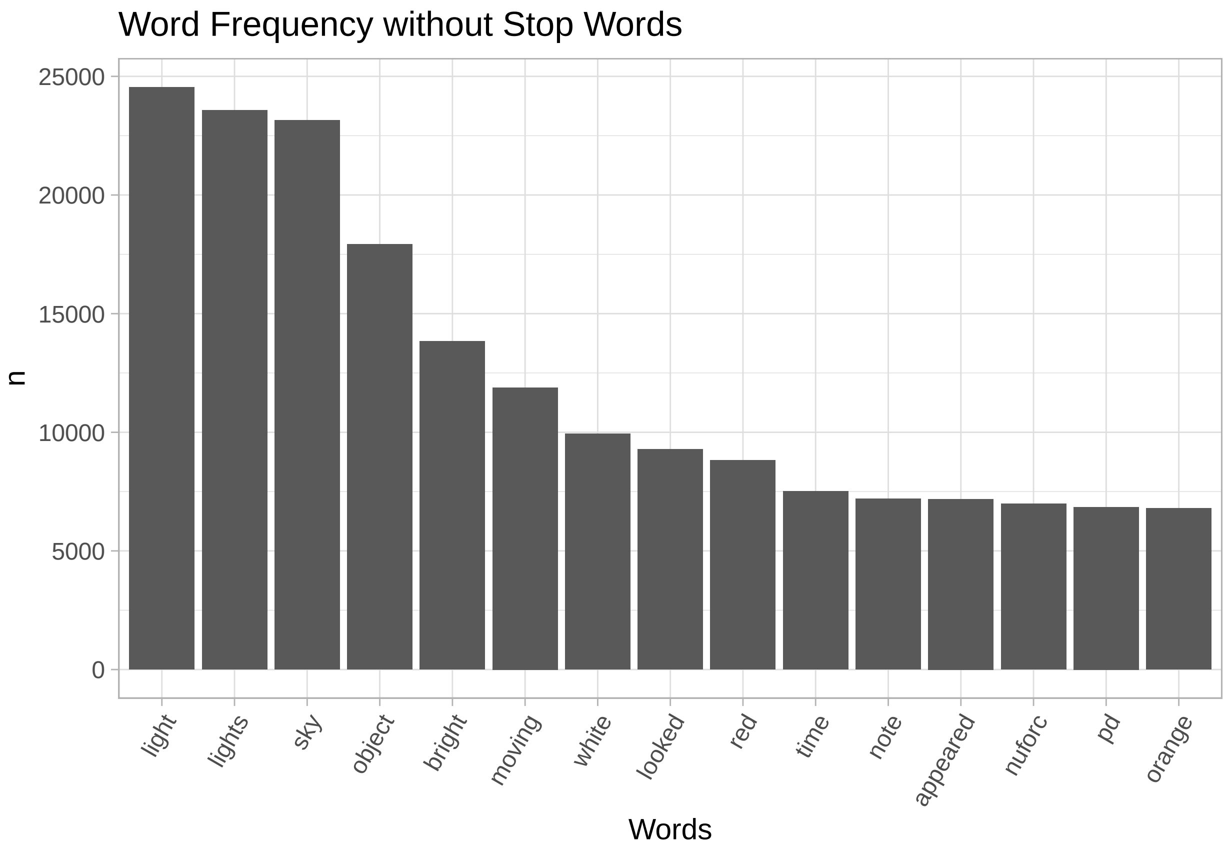
```
#as we can see, the most popular words (at the moment) are stop words. This isn't very helpful for our
```

now lets create stop words

```
#Step 2: Using the `TidyText` package, remove stop words and generate a new word count
ufo_no_stop <- words %>%
  anti_join(stop_words)
```

```
## Joining, by = "word"
```

```
ufo_no_stop %>%
  count(word, sort = T) %>%
  slice(1:15) %>%
  ggplot(aes(x = reorder(word, n, function(n) -n), y = n)) +
  geom_bar(stat = "identity") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  xlab("Words") +
  ggtitle("Word Frequency without Stop Words")
```



We can see the most common words are those typical for a UFO report. Light, sky, object, moving, and looked all make sense here.

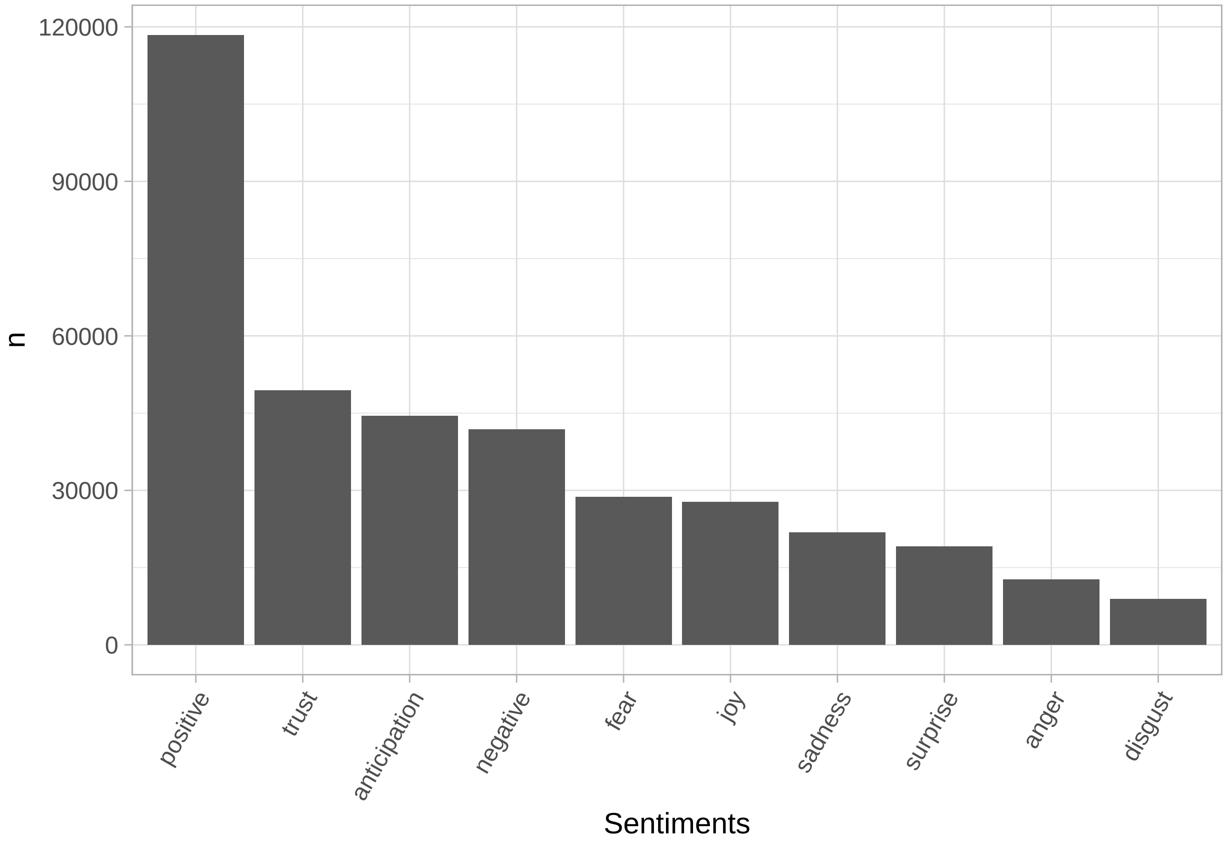
Let’s explore sentiment analysis.

```
sentiments <- get_sentiments("nrc")

df_sentiments1 <- ufo_no_stop %>% left_join(sentiments)
```

```
## Joining, by = "word"
```

```
df_sentiments_filtered1 <- df_sentiments1 %>%
  filter(!is.na(sentiment)) %>%
  group_by(sentiment) %>%
  summarize(n = n())
```

```
df_sentiments_filtered1 %>%
  ggplot(aes(x = reorder(sentiment, n, function(n) -n), y = n)) +
  geom_bar(stat = "identity") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  xlab("Sentiments")
```



Okay I think that this a lot of baloney

BUT there is a high amount for positive and trust! which is neato

ok trying