Population, Shapes, Phenomena

Edmund Hui, Rio Jia, Rachel Montgomery, Yuning Wu

2022-11-02

Observations

- Why are there duplicate observations in the data?

Dependencies & Read in Data

```
#install.packages("tidyverse")
#install.packages("stringr")
#install.packages("usmap")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(usmap)
```

```
df <- read_csv("data/UFO_and_Weather.csv")
```

```
## New names:
## Rows: 22482 Columns: 18
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (4): city, state, shape, text dbl (12): ...1, city_latitude, city_longitude,
## year, month, day, hour, temp... lgl (1): snow dttm (1): date_time
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
df
```

```
## # A tibble: 22,482 x 18
##      ...1 city  state date_time           shape text  city_~1 city_~2  year month
##     <dbl> <chr> <chr> <dttm>              <chr> <chr>   <dbl>   <dbl> <dbl> <dbl>
##  1      0 Ches~ VA    2019-12-12 18:43:00 light My w~    37.3   -77.4  2019    12
##  2      1 Rock~ CT    2019-03-22 18:30:00 circ~ I th~    41.7   -72.6  2019     3
##  3      2 Otta~ ON    2019-04-17 02:00:00 tear~ I wa~    45.4   -75.7  2019     4
##  4     3 Kirb~ TX    2019-04-02 20:25:00 disk  The ~    30.7   -94.0  2019     4
##  5     4 Tucs~ 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
##  7     6 Broo~ IN    2019-06-18 21:00:00 sphe~ Meta~    39.4   -85.0  2019     6
##  8     7 Melb~ FL    2019-06-12 22:00:00 unkn~ We t~    28.0   -80.5  2019     6
##  9     8 Carr~ NM    2019-06-11 22:00:00 chan~ I wa~    33.8  -106.   2019     6
## 10     9 Waco  TX    2018-06-15 01:00:00 circ~ I wa~    31.6   -97.1  2018     6
## # ... with 22,472 more rows, 8 more variables: day <dbl>, hour <dbl>,
## #   temperature <dbl>, relative_humidity <dbl>, precipitation <dbl>,
## #   snow <lgl>, wind_direction <dbl>, wind_speed <dbl>, and abbreviated
## #   variable names 1: city_latitude, 2: city_longitude
```

Question 1: Do UFO sightings happen in more densely populated areas?

We would have to add in some sort of population/census data, but could be interesting to look into

```
citypop <- read_csv("data/populations_by_city.csv")
```

```
## Rows: 81372 Columns: 4
## -- Column specification ---------------------------------------------------------
## Delimiter: ","
## chr (3): TYPE, SHORTNAME, STSHORT
## dbl (1): POPESTIMATE2021
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Filter only the municipalities
citypop <- citypop %>% filter(TYPE %in% c("city", "town", "village"))
citypop
```

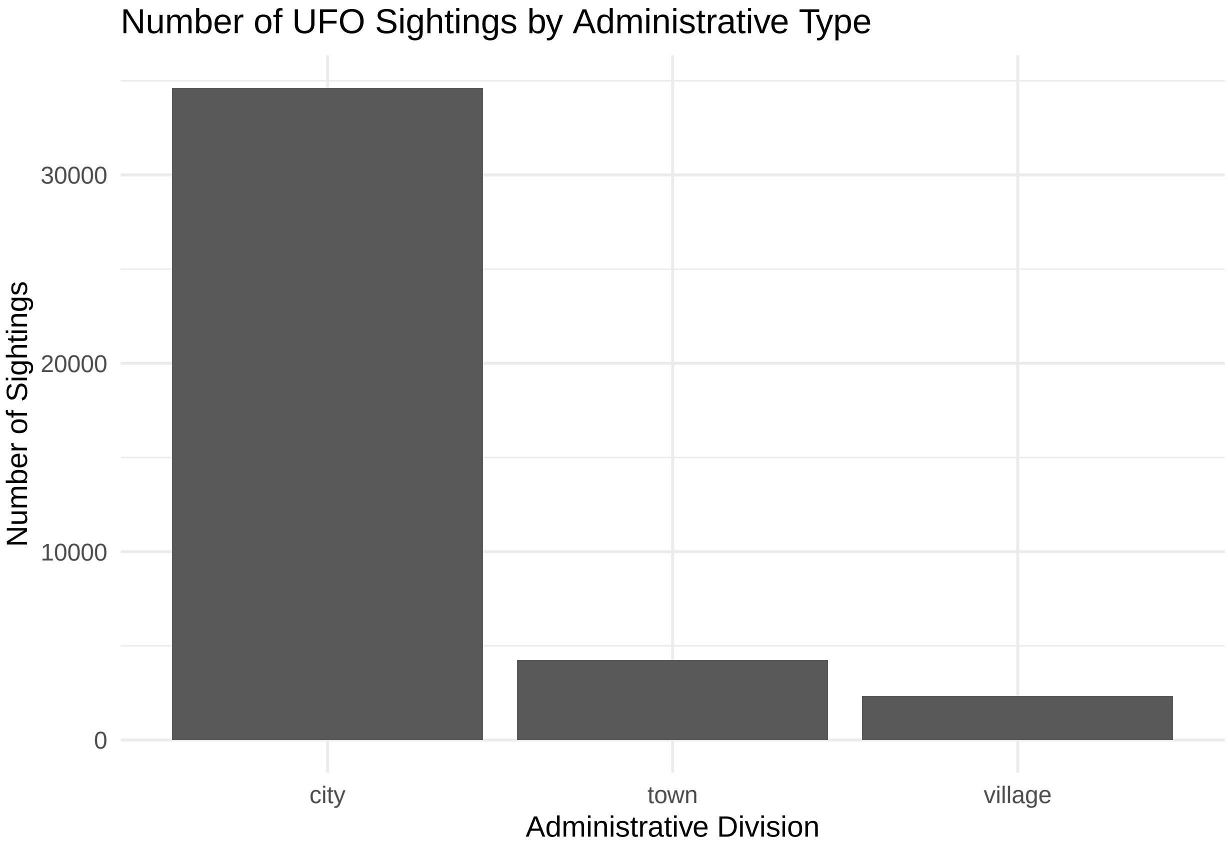
```
## # A tibble: 48,555 x 4
##    POPESTIMATE2021 TYPE  SHORTNAME      STSHORT
##              <dbl> <chr> <chr>          <chr>
##  1            2379 city  Abbeville      AL
##  2            4294 city  Adamsville     AL
##  3             668 town  Addison        AL
##  4             226 town  Akron          AL
##  5           33676 city  Alabaster      AL
##  6           22522 city  Albertville    AL
##  7           14618 city  Alexander City AL
##  8            2123 city  Aliceville     AL
##  9             545 town  Allgood        AL
## 10             951 town  Altoona        AL
## # ... with 48,545 more rows
```

```
df <- df %>%
  left_join(citypop, by = c("city"="SHORTNAME", "state"="STSHORT")) %>%
  rename(population_2021 = POPESTIMATE2021, geo_class=TYPE)

df
```

```
## # A tibble: 45,206 x 20
##     ...1 city  state date_time           shape text  city_~1 city_~2  year month
##    <dbl> <chr> <chr> <dttm>              <chr> <chr>   <dbl>   <dbl> <dbl> <dbl>
##  1     0 Ches~ VA    2019-12-12 18:43:00 light My w~    37.3   -77.4  2019    12
##  2     1 Rock~ CT    2019-03-22 18:30:00 circ~ I th~    41.7   -72.6  2019     3
##  3     2 Otta~ ON    2019-04-17 02:00:00 tear~ I wa~    45.4   -75.7  2019     4
##  4     3 Kirb~ TX    2019-04-02 20:25:00 disk  The ~    30.7   -94.0  2019     4
##  5     3 Kirb~ TX    2019-04-02 20:25:00 disk  The ~    30.7   -94.0  2019     4
##  6     4 Tucs~ AZ    2019-05-01 11:00:00 unkn~ Desc~    32.3  -111.   2019     5
##  7     4 Tucs~ AZ    2019-05-01 11:00:00 unkn~ Desc~    32.3  -111.   2019     5
##  8     5 Gold~ AZ    2019-04-10 17:00:00 circ~ Apr.~    33.4  -111.   2019     4
##  9     6 Broo~ IN    2019-06-18 21:00:00 sphe~ Meta~    39.4   -85.0  2019     6
## 10     6 Broo~ IN    2019-06-18 21:00:00 sphe~ Meta~    39.4   -85.0  2019     6
## # ... with 45,196 more rows, 10 more variables: day <dbl>, hour <dbl>,
## #   temperature <dbl>, relative_humidity <dbl>, precipitation <dbl>,
## #   snow <lgl>, wind_direction <dbl>, wind_speed <dbl>, population_2021 <dbl>,
## #   geo_class <chr>, and abbreviated variable names 1: city_latitude,
## #   2: city_longitude
```

Group by municipalities (administrative division)

```
muni <- df %>% drop_na(geo_class) %>% group_by(geo_class) %>% summarise(count=n())
muni %>% ggplot(aes(x=geo_class, y=count)) + geom_col() +
  theme_minimal() +
  labs(title="Number of UFO Sightings by Administrative Type") +
  xlab("Administrative Division") +
  ylab("Number of Sightings")
```



## Group By Population

Use case_when to split population into even & logical levels

```
#mean(citypop$POPESTIMATE2021)
#mean(df$population_2021, na.rm=TRUE)
```

UFO Reports Per Capita / Per State

```
states <- read_csv("data/statepop.csv", col_names=FALSE)
```

```
## Rows: 51 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): X1
## num (3): X2, X3, X4
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

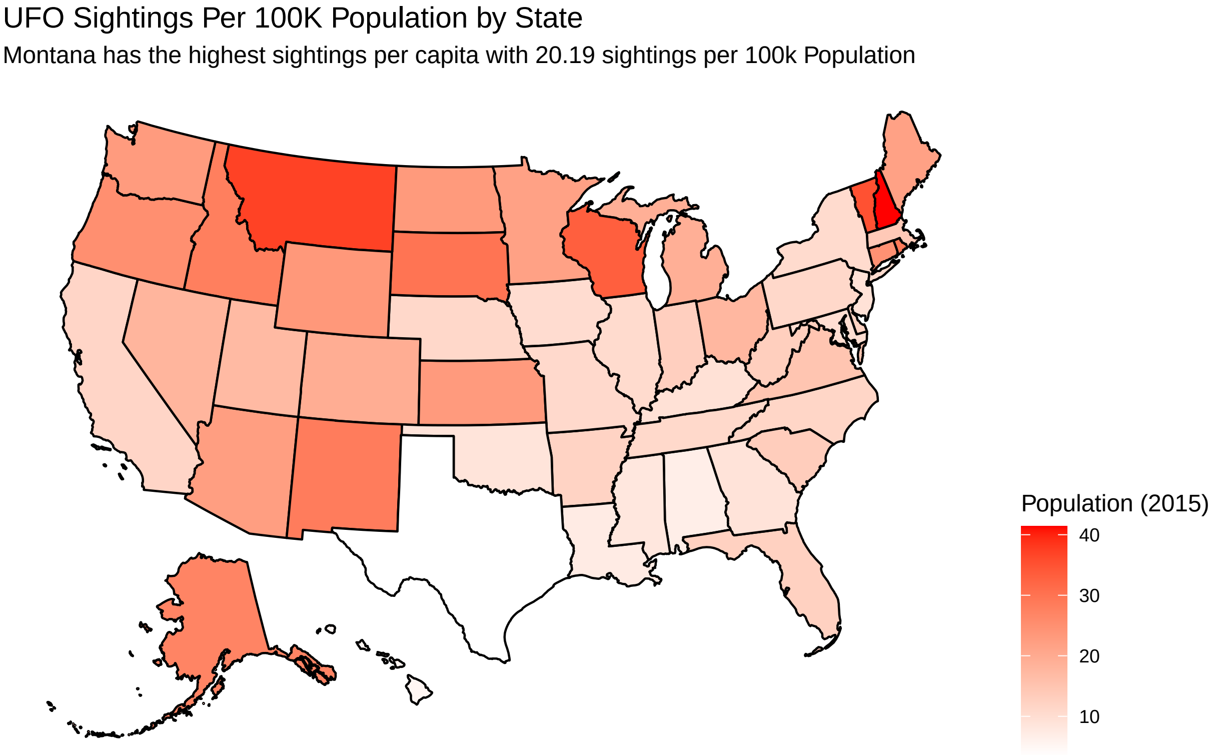
```
states <- states %>% select(X1, X4)
names(states) <- c("state", "population_2021")
states$state <- substr(states$state, 2, 100)
states$state <- state.abb[match(states$state,state.name)]
states <- states %>% drop_na()
```

```
sight_counts <- df %>% group_by(state) %>% summarise(count=n())
states <- states %>% left_join(sight_counts, by="state") %>% mutate(obs_100k = (count/population_2021)*
```

```
states %>% arrange(obs_100k)
```

```
## # A tibble: 50 x 4
##    state population_2021 count obs_100k
##    <chr>           <dbl> <int>    <dbl>
##  1 TX           29527941  1014     3.43
##  2 HI            1441553    74     5.13
##  3 AL            5039877   342     6.79
##  4 LA            4624047   339     7.33
##  5 MS            2949965   245     8.31
##  6 OK            3986639   351     8.80
##  7 GA           10799566   972     9.00
##  8 NJ            9267130   858     9.26
##  9 KY            4509394   424     9.40
## 10 MD            6165129   616     9.99
## # ... with 40 more rows
```

```
plot_usmap(data = states, values = "obs_100k", color = "black") +
  scale_fill_continuous( low = "white", high = "red", name = "Population (2015)", label = scales::comma
  theme(legend.position = "right") +
  labs(title="UFO Sightings Per 100K Population by State", subtitle="Montana has the highest sightings p
```



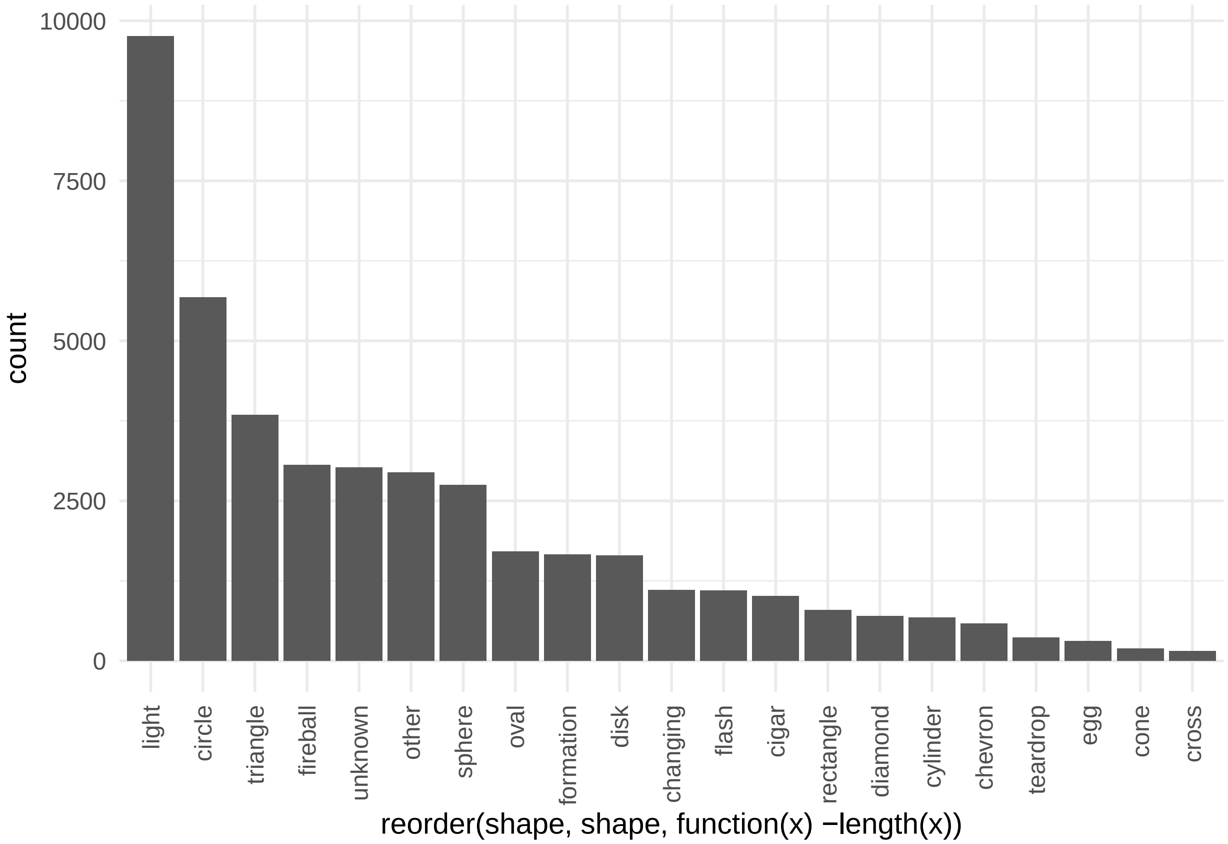
UFO Sightings Per 100K Population by State

Montana has the highest sightings per capita with 20.19 sightings per 100k Population

Question 2: What are the most common UFO descriptions?

General Shape Analysis

```
df %>% drop_na(shape) %>%
  ggplot() +
  geom_bar(aes(x = reorder(shape, shape, function(x)-length(x)))) +
  scale_x_discrete(guide = guide_axis(angle = 90)) +
  theme_minimal()
```



Which Shapes are Most Common in Each State?

```
df %>%
  group_by(state, shape) %>%
  summarise(count=n()) %>%
  group_by(state) %>%
  top_n(1, count)
```

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

```
## # A tibble: 65 x 3
## # Groups:   state [62]
##    state shape count
##    <chr> <chr> <int>
##  1 AB    light    37
##  2 AK    light    39
##  3 AL    light    95
##  4 AR    light    83
##  5 AZ    light   425
##  6 BC    light    21
##  7 CA    light  1018
##  8 CO    light   218
##  9 CT    light   168
## 10 DC    other     4
## # ... with 55 more rows
```

Question 3: Do certain cultural phenomena influence UFO sightings?

— We can add in cultural data like # of sci fi movies released in a year and see if there is a correlation, if a war is happening, etc

3.1 Investigate the relationships between google trends data and UFO sightings

```
# Read in google trend data
trends <- read_csv("data/multiTimeline.csv", skip=1)
```

```
## Rows: 72 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Month
## dbl (2): ufo: (United States), alien: (United States)
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

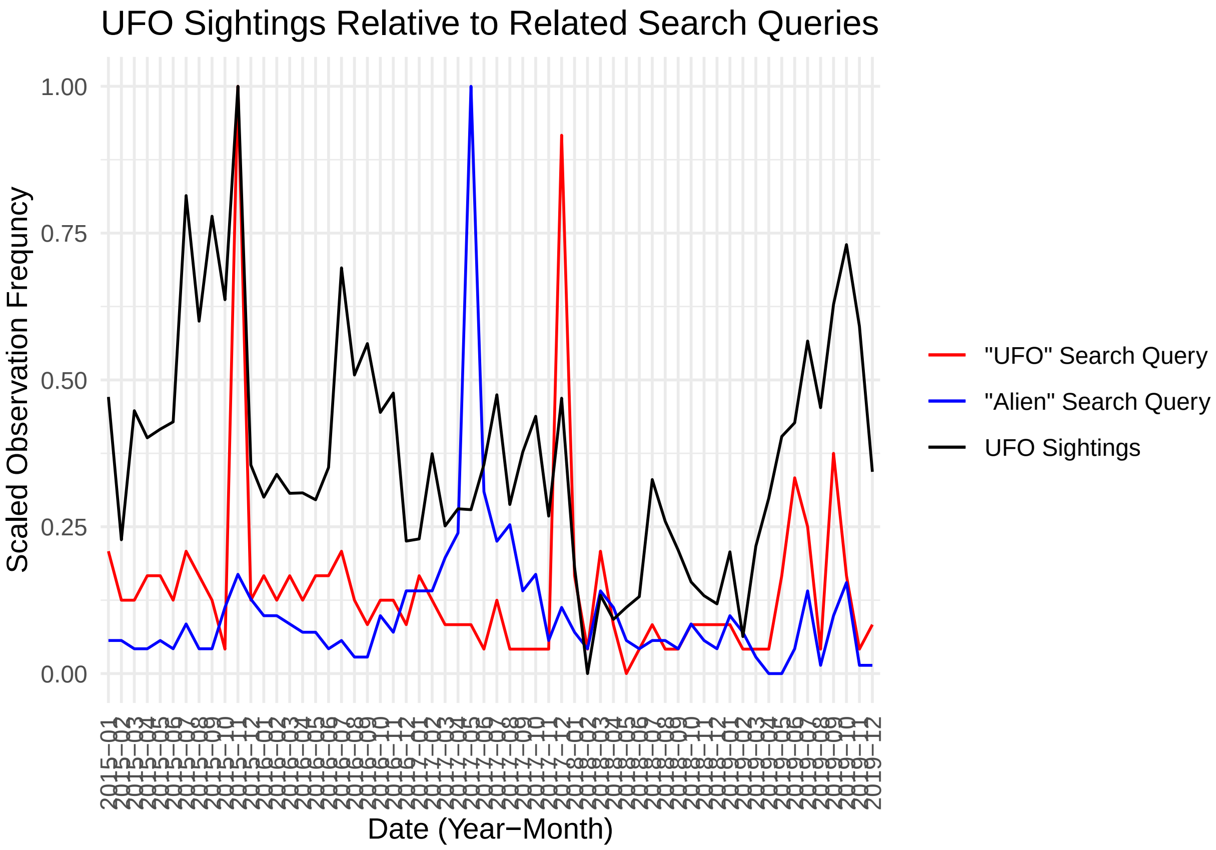
```
# Rename columns
trends <- trends %>% rename(month=Month,ufo=`ufo: (United States)`, alien=`alien: (United States)`)

# Take only the years we need
trends <- trends %>% filter(substr(month, 1,4) %in% c("2015","2016","2017","2018","2019"))

# Count occurances each month
df_counts <- df %>% group_by(year, month) %>% summarise(count=n())
```

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

```
# Add counts to the trends dataframe
trends$count <- df_counts$count
```

```
# Min-Max scale as we are only interested in relative movements
trends$ufo_scaled <- (trends$ufo-min(trends$ufo))/(max(trends$ufo)-min(trends$ufo))
trends$alien_scaled <- (trends$alien-min(trends$alien))/(max(trends$alien)-min(trends$alien))
trends$count_scaled <- (trends$count-min(trends$count))/(max(trends$count)-min(trends$count))
```

```
trends %>% ggplot(aes(x=month, y=ufo_scaled, group=1)) +
  geom_line(aes(colour="\"UFO\" Search Query")) +
  geom_line(aes(y=alien_scaled, colour="\"Alien\" Search Query")) +
  geom_line(aes(y=count_scaled, colour="UFO Sightings")) +
  labs(x = "Date (Year-Month)",
       y = "Scaled Observation Frequncy",
       color = "Legend") +
  scale_colour_manual("",
                      breaks = c("\"UFO\" Search Query", "\"Alien\" Search Query", "UFO Sightings"),
                      values = c("red", "blue", "black")) +
  labs(title="UFO Sightings Relative to Related Search Queries") +
  scale_x_discrete(guide = guide_axis(angle = 90)) + theme_minimal()
```



```
#theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```