Homework 3

Soc 225: Data & Society

Brian Park

2023-04-18

Contents

Problem 1: Google Trends **1**

1.1 1

Problem 2: Join data frames **2**

2.1. Set up your environment: 2

2.2 Find data frames 2

2.3. Find common keys 3

2.4. Join the two data frames 3

2.5. Build on your answer 3

2.6. Use the data to answer the below questions 4

2.7. 4

Problem 3: Your research question **4**

Write all code in the chunks provided!

Remember to unzip to a real directory before running everything!

Problems should be roughly analogous to what we've done in class, with a few extensions. There are hints at the bottom of this document if you get stuck. If you still can't figure it out, go to google/stack exchange/ask a friend. Finally, email your TA or come to office hours :).

Problem 1: Google Trends

1.1

Go to Google Trends and search for "covid-19 vaccine". Look at variations by time and by region in US. What do you observe?

- This search tool is a very interesting website to find people's online interests in certain social trends based on my searched keywords. When I searched for "covid-19 vaccine", I could find that people's interest in social trends varies with respect to the specified time frame and regions. The rank of interests based on my keyword varies every time I change the time frame and regions. For example, for the past 12 months, Delaware was ranked at the very top, representing people in this state who were mostly interested in the "covid-19 vaccine" for the past 12 months. On the other hand, once I changed it to for the past 7 days, New Hampshire was ranked at the top, showing that people living in New Hampshire were mostly interested in the search keyword for the past 7 days.

Problem 2: Join data frames

In this problem we will use data in the `nycflightdata13` package to perform joining of data frames.

It includes five dataframes, some of which contain missing data (`NA`).

- `flights`: flights leaving JFK, LGA or EWR in 2013
- `airlines`: airline abbreviations
- `airports`: airport metadata
- `planes`: airplane metadata
- `weather`: hourly weather data from JFK, LGA and EWR

Note these are **separate data frames**, each needing to be loaded separately using `data()`.

2.1. Set up your environment:

a. Install and load the `nycflights13` package. Load the `tidyverse` package.
b. Load data sets `flights`, `planes`, `airlines`

```
#install.packages('nycflights13')
library(nycflights13)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
data(flights)
data(planes)
data(airlines)
```

2.2 Find data frames

We'll be looking at who manufactures the planes that flew to Seattle. Which are the two data frames we need to join?

- To determine who manufactures the planes headed to Seattle, we need to join these two following data frames, `flights` and `planes`.

2.3. Find common keys

Take a look at variables contained the two data frames. Which variable(s) should be used as the key to join?

- Among the variables in these two data frames, the variable `tailnum` should be used to join as the key variable.

2.4. Join the two data frames

```
combined_data <- left_join(flights, planes, by = "tailnum")
combined_data %>%
  head()
```

```
## # A tibble: 6 x 27
##   year.x month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1   2013     1     1      517            515         2      830            819
## 2   2013     1     1      533            529         4      850            830
## 3   2013     1     1      542            540         2      923            850
## 4   2013     1     1      544            545        -1     1004           1022
## 5   2013     1     1      554            600        -6      812            837
## 6   2013     1     1      554            558        -4      740            728
## # i 19 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>, year.y <int>, type <chr>,
## #   manufacturer <chr>, model <chr>, engines <int>, seats <int>, speed <int>,
## #   engine <chr>
```

2.5. Build on your answer

For flights with a destination of Seattle, who are the largest manufacturers? Give top five of the manufacturers. (Check hints if you have troubles)

```
combined_data %>%
  filter(dest == "SEA") %>%
  count(manufacturer) %>%
  arrange(desc(n)) %>%
  head(5)
```

```
## # A tibble: 5 x 2
##   manufacturer          n
##   <chr>             <int>
## 1 BOEING             2659
## 2 AIRBUS              475
## 3 AIRBUS INDUSTRIE    394
## 4 <NA>                391
## 5 BARKER JACK L         2
```

2.6. Use the data to answer the below questions

We'd like to know which airlines had the most flights to Seattle from NYC. Which are the two data frames we need to join, and on which key variable(s)?

- To determine which airlines had the most flights to Seattle from NYC, we need to join `flights` and `airlines` and use `carrier` as the key variable.

```
combined_data2 <- left_join(flights, airlines, by = "carrier")
combined_data2 %>%
  head()
```

```
## # A tibble: 6 x 20
##    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
## 1  2013     1     1      517            515         2      830            819
## 2  2013     1     1      533            529         4      850            830
## 3  2013     1     1      542            540         2      923            850
## 4  2013     1     1      544            545        -1     1004           1022
## 5  2013     1     1      554            600        -6      812            837
## 6  2013     1     1      554            558        -4      740            728
## # i 12 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>, name <chr>
```

2.7.

Join the two data frames in 2.6 and list the top five airlines.

```
combined_data2 %>%
  filter(dest == "SEA") %>%
  count(name) %>%
  arrange(desc(n))
```

```
## # A tibble: 5 x 2
##   name                       n
##   <chr>                  <int>
## 1 Delta Air Lines Inc.    1213
## 2 United Air Lines Inc.   1117
## 3 Alaska Airlines Inc.     714
## 4 JetBlue Airways          514
## 5 American Airlines Inc.   365
```

Problem 3: Your research question

Think about the research question you have in mind. Plot is a great way to understand patterns, key relationships and uncertainties in a data set. Here we'll ask you to plan about plotting your variables of interest for your research question. Try to think about **3 plots** below:

For each of the 3 plots, provide:

A. The purpose of the plot: what do you want people to understand when they see this?

B. The type of plot: what geom functions will you use to present the plot? Why are those the best choices?

C. Limitations/biases: What is missing from this presentation? Could someone get the wrong idea? What can you do to help limit the negative possibilities here?

```
mlb_data <- read.csv('data/mlb.csv') %>%
  select(last_name, first_name, xslg, launch_angle_avg, exit_velocity_avg)

team_one <- mlb_data %>%
  filter(launch_angle_avg >= 22 & launch_angle_avg <= 28 & exit_velocity_avg >= 90) %>%
  mutate(team = "team1", mid_slg = median(xslg), mean_slg = mean(xslg))

team_two <- mlb_data %>%
  filter(launch_angle_avg < 22 | launch_angle_avg > 28 & exit_velocity_avg < 90) %>%
  mutate(team = "team2")
```

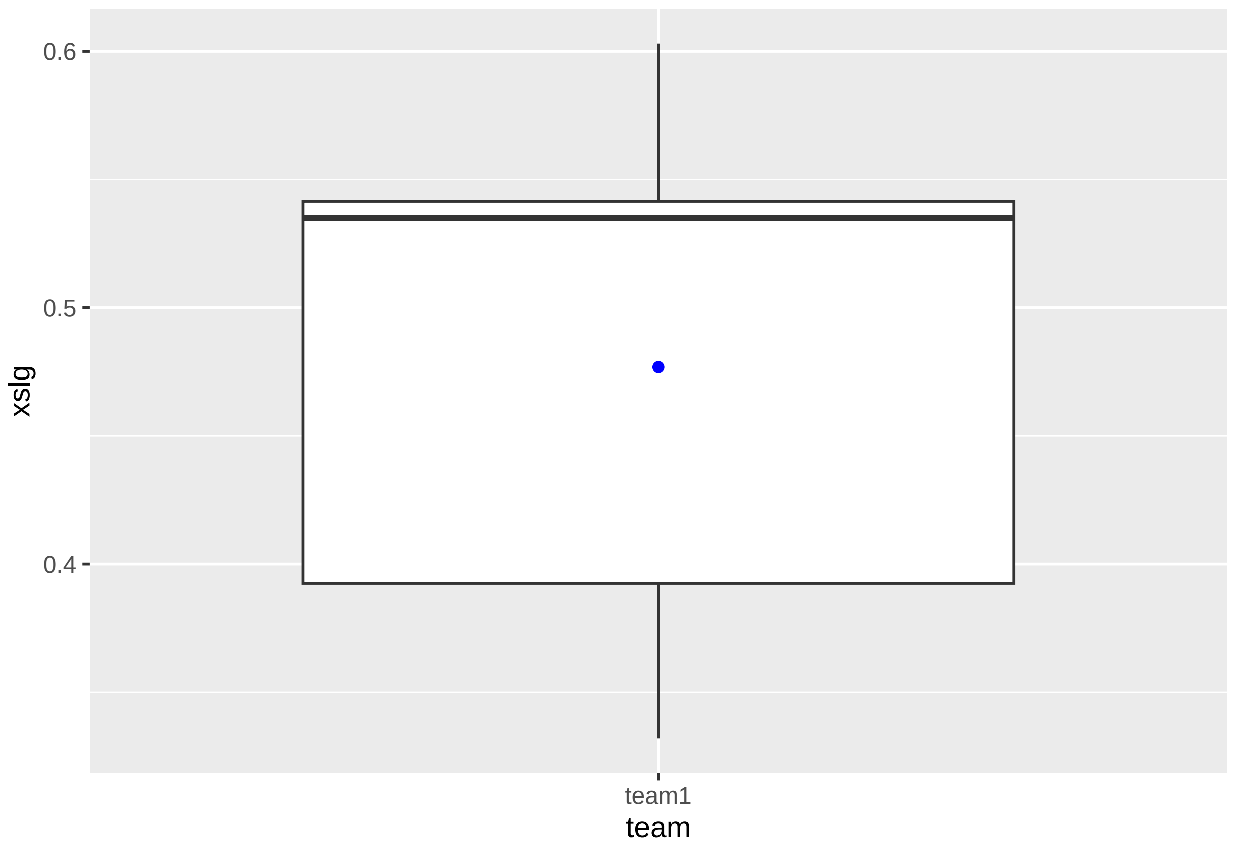
Plot idea 1

A. The plot 1 will display median and mean of `xslg` between two groups who fall into the ideal condition (22 <= `launch_angle_avg` <= 28 and `exit_velocity_avg` >= 90) and who do not meet this condition. With this plot, people will understand the difference of median and mean between two groups.

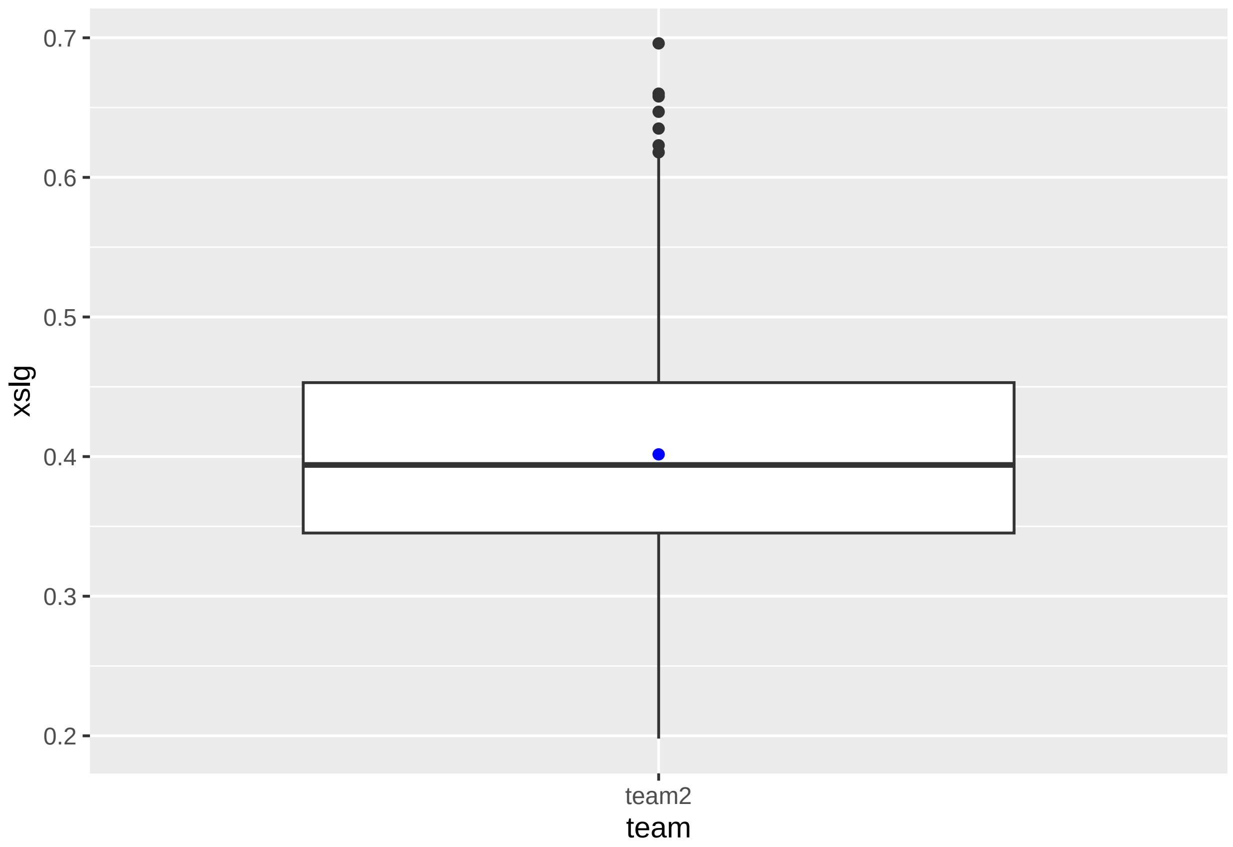
B. The box plot will be the most suitable geom function to represent the purpose of this plot because it will only need to show two values of median and mean together.

C. The mean number is to divide the sum of `xslg` by the total number of players in each group. Since there are only a few players in a group meeting the ideal condition, I can put how many players belong to each group to provide more transparent information to people seeing this data.

```
ggplot(team_one, aes(x = team, y = xslg)) +
      geom_boxplot() +
      stat_summary(fun=mean, geom="point", color="blue", fill="blue")
```



```
ggplot(team_two, aes(x = team, y = xslg)) +
      geom_boxplot() +
      stat_summary(fun=mean, geom="point", color="blue", fill="blue")
```



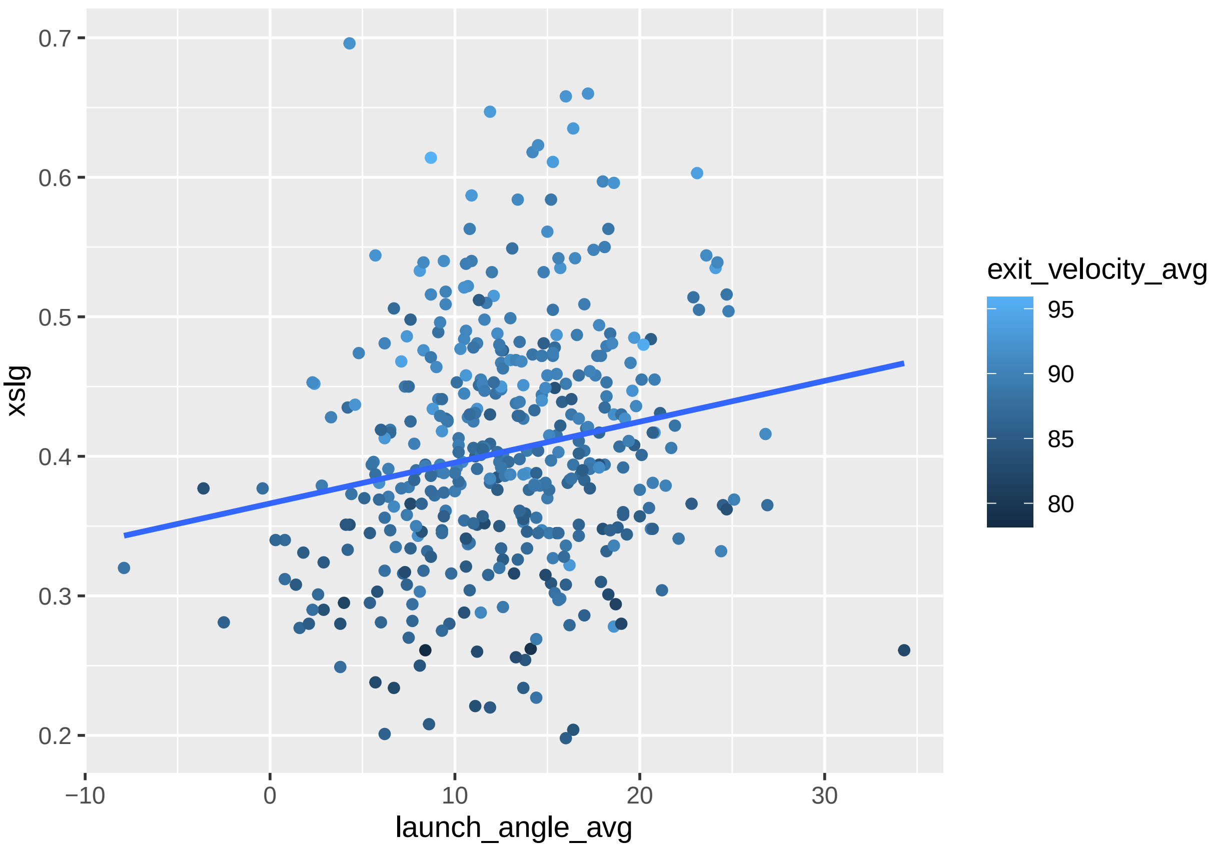
Plot idea 2

A. The plot 2 will provide the relationship between `xslg` and `exit_velocity_avg` to represent how `xslg` varies with respect to `exit_velocity_avg`.
B. The scatterplot will be the most suitable geom function to visualize the relationship between those variables whether `xslg` is proportional to `exit_velocity_avg`.
C. If this plot is expected to be too scattered to see the relationship, I would provide the best fitting line on the scatterplot to help them understand the relationship.

```
plot1 <- ggplot(mlb_data, aes(x = launch_angle_avg, y = xslg, color = exit_velocity_avg)) +
         geom_point() +
         geom_smooth(method="lm", se = FALSE)
plot1
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



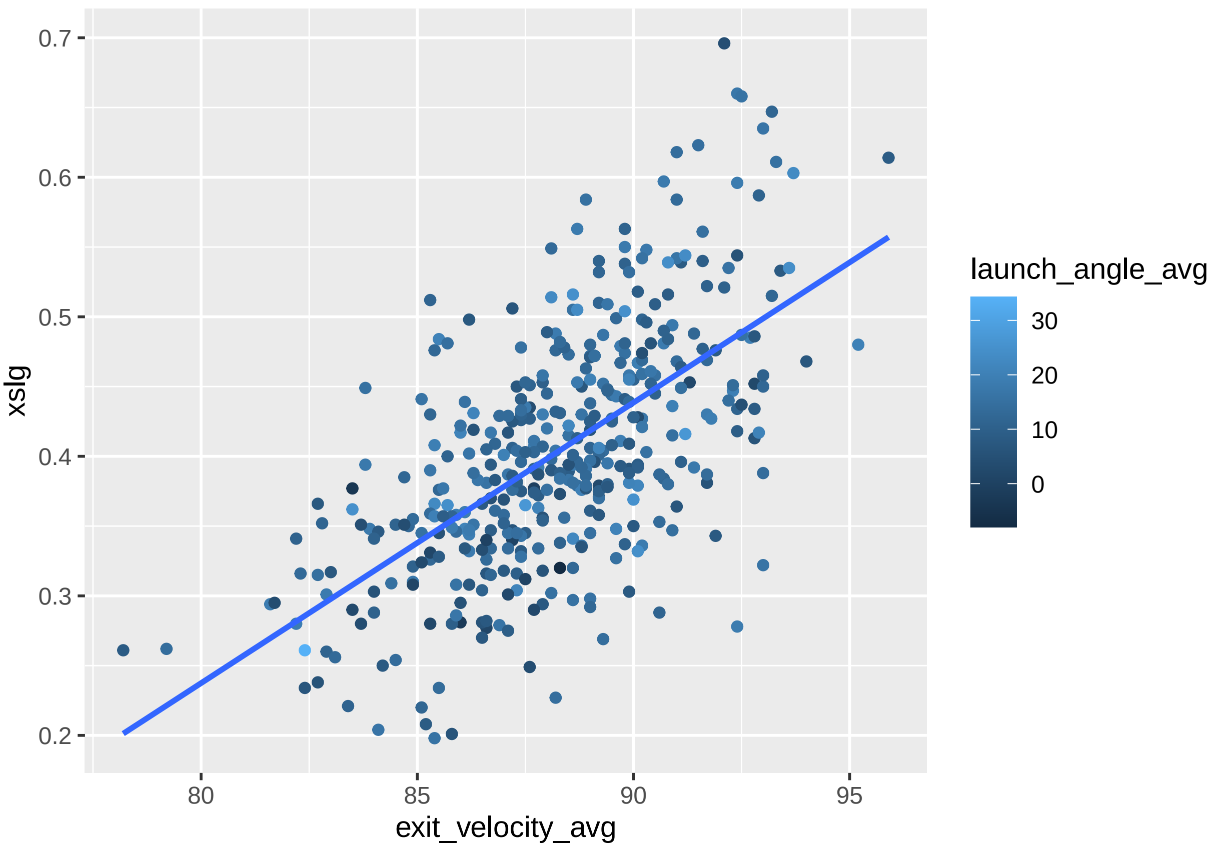
Plot idea 3

A. The plot 3 will provide the relationship between `xslg` and `launch_angle_avg` to represent how `xslg` varies with respect to `launch_angle_avg`.
B. The scatterplot will be the most suitable geom function to visualize the relationship between those variables whether `xslg` is proportional to `launch_angle_avg`.
C. If this plot is expected to be too scattered to see the relationship, I would provide the best fitting line on the scatterplot to help them understand the relationship.

```
plot2 <- ggplot(mlb_data, aes(x = exit_velocity_avg, y = xslg, color = launch_angle_avg)) +
         geom_point() +
         geom_smooth(method="lm", se = FALSE)
plot2
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



Hint

2.5

use `left_join()` by "tailnum" to join the two data frames, then count() observations by manufacturer, and then use arrange() with descending order.