Homework 2

Soc 225: Data & Society

Brian Park

2023-04-18

Contents

Write all code in the chunks provided!

Remember to unzip to a real directory before running everything!

Problem 1 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: Piping Hot Variables

This problem uses `dplyr` verbs to answer questions about an Airbnb data set.

1.1: Get the data

Go to Inside Airbnb and download the "Detailed Listings" data for Seattle, `listings.csv.gz`. This file has many more variables than the "Summary" file we've been using in class. Put it in a `data/` subfolder in your `hw-02` project folder.

[This is a compressed (gzipped) file, but R should be able to handle it as-is. If you run into trouble, try unzipping the file before reading it into R.]

1.2: Set up your R environment

a. Load the tidyverse
b. Read the detailed Airbnb data into R

```
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
```

```
listings <- read.csv(gzfile("data/listings.csv.gz"))
```

1.3: Use the data to answer a question

For how many units does the host live in a different neighborhood from the listing? For how many units does the host live in the same neighborhood as the listing?

Try to figure out which variables to use from their names, and think about which verbs you've learned about might work to answer this question. See the hints at the end if you need help.

```
listings %>%
  select(neighbourhood_cleansed, host_neighbourhood) %>%
  filter(!(neighbourhood_cleansed %in% host_neighbourhood)) %>%
  glimpse()
```

```
## Rows: 14,246
## Columns: 2
## $ neighbourhood_cleansed <chr> "Oostelijk Havengebied - Indische Buurt", "Cent~
## $ host_neighbourhood     <chr> "Indische Buurt", "Grachtengordel", "Grachtengo~
```

```
listings %>%
  select(neighbourhood_cleansed, host_neighbourhood) %>%
  filter(neighbourhood_cleansed %in% host_neighbourhood) %>%
  glimpse()
```

```
## Rows: 1,870
## Columns: 2
## $ neighbourhood_cleansed <chr> "Bos en Lommer", "Bos en Lommer", "Slotervaart"~
## $ host_neighbourhood     <chr> "Oud-West", "Bos en Lommer", "Slotervaart", "Bo~
```

There are 10,405 units where the host lives in a different neighborhood from the listing. There are 5,711 units where the host lives in the same neighborhood as the listing.

1.4: Build on your answer

Building on that work, what is the average number of listings for hosts that live in the same neighborhood as their listing? What's the average for hosts who live in different neighborhoods from their listing?

The `mean` function will take the average of a variable, but you might need to look up how to use it. See the hints for more suggestions if you get stuck.

```
listings %>%
  filter(neighbourhood_cleansed %in% host_neighbourhood) %>%
  group_by(neighbourhood_cleansed) %>%
  summarise(ave = mean(host_listings_count))
```

```
## # A tibble: 4 x 2
##   neighbourhood_cleansed   ave
##   <chr>                  <dbl>
## 1 Bos en Lommer          NA
## 2 Osdorp                  1.11
## 3 Slotervaart             1.43
## 4 Watergraafsmeer         1.31
```

```
listings %>%
  filter(!(neighbourhood_cleansed %in% host_neighbourhood)) %>%
  group_by(neighbourhood_cleansed) %>%
  summarise(ave = mean(host_listings_count))
```

```
## # A tibble: 18 x 2
##    neighbourhood_cleansed                   ave
##    <chr>                                  <dbl>
##  1 Bijlmer-Centrum                         1.27
##  2 Bijlmer-Oost                            1.25
##  3 Buitenveldert - Zuidas                  1.59
##  4 Centrum-Oost                           NA
##  5 Centrum-West                            3.00
##  6 De Aker - Nieuw Sloten                  1.43
##  7 De Baarsjes - Oud-West                 NA
##  8 De Pijp - Rivierenbuurt                 1.68
##  9 Gaasperdam - Driemond                   1.57
```

```
## 10 Geuzenveld - Slotermeer                  1.74
## 11 IJburg - Zeeburgereiland                 1.85
## 12 Noord-Oost                              NA
## 13 Noord-West                               1.27
## 14 Oostelijk Havengebied - Indische Buurt  1.31
## 15 Oud-Noord                                2.09
## 16 Oud-Oost                                 1.27
## 17 Westerpark                               1.53
## 18 Zuid                                     1.65
```

1.5: Reflect and interpret

Reflect on your answer to 1.4. What might cause the results you got? How does that connect to the idea that Airbnb might be changing neighborhoods?

The average number of listings for hosts living in the same neighborhood is smaller than those living in the different neighborhood. This can be interpreted that the majority of hosts in this neighborhood do not tend to host their houses nearby. Therefore, this result is the reason why Airbnb might be changing neighborhoods to other places which have more hosts living in the same neighborhoods.

2. Prepare and Visualize data

2.1. Set up your environment

Set up your environment by:

Reading the Airbnb data: There's another new data set in the `data/` folder. This one has almost 10,000 cases and the census data by zipcode. These data are from New York City, not Seattle!

```r
new_data <- read.csv('data/census.csv')
summary(new_data)
```

```
##     zipcode          white           black           asian      
##  Min.   :10001   Min.   :  160   Min.   :   17   Min.   :   67  
##  1st Qu.:10250   1st Qu.: 3477   1st Qu.: 1064   1st Qu.: 1044  
##  Median :11104   Median :10119   Median : 3243   Median : 3220  
##  Mean   :10790   Mean   :15586   Mean   :10763   Mean   : 5898  
##  3rd Qu.:11357   3rd Qu.:21800   3rd Qu.:14634   3rd Qu.: 6742  
##  Max.   :11693   Max.   :58646   Max.   :77175   Max.   :60670  
##                  NA's   :1                                      
##      latinx         full_pop      white_proportion  black_proportion 
##  Min.   :   91   Min.   :  1685   Min.   :0.00539   Min.   :0.003838  
##  1st Qu.: 4175   1st Qu.: 26970   1st Qu.:0.10244   1st Qu.:0.030672  
##  Median : 7773   Median : 41847   Median :0.37689   Median :0.082469  
##  Mean   :13530   Mean   : 47026   Mean   :0.37072   Mean   :0.212112  
##  3rd Qu.:16551   3rd Qu.: 66426   3rd Qu.:0.63733   3rd Qu.:0.300999  
##  Max.   :81093   Max.   :109931   Max.   :0.87819   Max.   :0.903535  
##                                   NA's   :1                          
##   modal_race       
##  Length:172        
##  Class :character  
##  Mode  :character  
```

```
##
##
##
##
```

We've given you absolute populations and proportions for the racial composition of the zipcode for each listing. We've also made a variable called 'modal_race' which is the race with the largest proportion in that neighborhood.

These variables are all in the last columns of the data set—you can try selecting them and using `summary()` to get a sense for what they contain.

2.2: Turn `price` into a number

`price` includes dollar signs, which means that R interprets it as a character. We want it to be a numeric variable instead. Turn `price` into a numeric variable in the chunk below.

There are a few ways to do this using `tidyverse` functions. See the hints below for some suggestions.

2.3: Make a scatterplot

Use a scatter plot to compare how unit prices change with the proportion of a particular race.

Bonus: try grouping by zipcode (in any fashion) for this plot

2.4: Make a boxplot

Use the `modal_race` variable to plot a boxplot comparing race and price. You may have to look up how to make a boxplot in `ggplot2`—what geom do you need?

Bonus: try showing how this comparison differs by neighborhood group.

2.5: Interpret your answer

Interpret your answer to 2.4. Check the hints if you need help.

Your answer should be at least a few sentences here

Bonus: how did we make the data?

There's another file in the data folder, census.csv. Read it into R and have a look at it.

Download the full listings for New York City from Inside Airbnb, and see if you can join the Census data to it by zipcode using `left_join`. You'll have to filter out some weird values for zipcode before you can merge.

3. Your own data

3.1. Looking at the dataset you chose for homework 1, think about a research question you'd like to investigate (try search about existing studies around your question). What variables do you plan to use to answer your question?

I will use three variables, `xslg`(Expected Slugging Percentage), `launch_angle_avg`, and `exit_velocity_avg` to examine the correlation of how the launch angle of a ball and its velocity affects the xslg. The ideal angle and velocity are from 22 to 28 degrees and faster than 90 mph, respectively. I expect that the players more close to the ideal condition would likely have higher xslg.

3.2. What is one way that you have to modify or examine your data to begin to answer your question?

I will have to filter the players having launch angle from 22 to 28 degrees and exit velocity faster than 90 mph. After that, I will examine its mean and median values to answer my research question.

3.3. Using the functions we've worked with in class (select, filter, arrange, mutate), plus any others you'd like to use, clean and transform your data set to make it ready for further exploration.

You must:

a. Create a new dataset that only includes the variables you're interested in

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

head(new_dataset)
```

```
##   last_name first_name  xslg launch_angle_avg exit_velocity_avg
## 1    Pujols     Albert 0.394             16.4              88.6
## 2   Cabrera     Miguel 0.515             12.1              93.2
## 3    Mathis       Jeff 0.348             20.6              89.6
## 4      Choo   Shin-Soo 0.455             11.4              90.0
## 5    Molina     Yadier 0.385             12.3              84.7
## 6      Cano   Robinson 0.481              6.2              90.4
```

b. Output a version of that dataset that only includes certain values of observations, hopefully ones you're interested in.

```
team_one <- new_dataset %>%
  filter(launch_angle_avg >= 22 & launch_angle_avg <= 28 & exit_velocity_avg >= 90)

summary(team_one)
```

```
##   last_name          first_name             xslg        launch_angle_avg
##  Length:7           Length:7           Min.   :0.3320   Min.   :23.10
##  Class :character   Class :character   1st Qu.:0.3925   1st Qu.:23.85
##  Mode  :character   Mode  :character   Median :0.5350   Median :24.20
##                                        Mean   :0.4769   Mean   :24.47
##                                        3rd Qu.:0.5415   3rd Qu.:24.75
##                                        Max.   :0.6030   Max.   :26.80
##  exit_velocity_avg
##  Min.   :90.00
##  1st Qu.:90.45
##  Median :91.20
##  Mean   :91.51
##  3rd Qu.:92.40
##  Max.   :93.70
```

```r
team_two <- new_dataset %>%
  filter(launch_angle_avg < 22 | launch_angle_avg > 28 & exit_velocity_avg < 90)

summary(team_two)
```

```
##   last_name          first_name             xslg        launch_angle_avg
##  Length:398         Length:398         Min.   :0.1980   Min.   :-7.900
##  Class :character   Class :character   1st Qu.:0.3453   1st Qu.: 8.925
##  Mode  :character   Mode  :character   Median :0.3940   Median :12.400
##                                        Mean   :0.4016   Mean   :12.227
##                                        3rd Qu.:0.4530   3rd Qu.:15.775
##                                        Max.   :0.6960   Max.   :34.300
##  exit_velocity_avg
##  Min.   :78.20
##  1st Qu.:86.60
##  Median :88.30
##  Mean   :88.22
##  3rd Qu.:89.90
##  Max.   :95.90
```

c. Order your data by the values of one variable you're interested in.

```r
team_one %>% arrange(desc(xslg))
```

```
##   last_name first_name  xslg launch_angle_avg exit_velocity_avg
## 1     Trout       Mike 0.603             23.1              93.7
## 2    Buxton      Byron 0.544             23.6              91.2
## 3     Smith       Will 0.539             24.2              90.8
## 4   Chapman       Matt 0.535             24.1              93.6
## 5     Gallo       Joey 0.416             26.8              91.2
## 6   Greiner    Grayson 0.369             25.1              90.0
## 7   Meadows     Austin 0.332             24.4              90.1
```

```r
team_two %>% arrange(desc(xslg))
```

```
##          last_name  first_name  xslg launch_angle_avg exit_velocity_avg
## 1             Soto        Juan 0.696              4.3              92.1
## 2          Freeman     Freddie 0.660             17.2              92.4
## 3           Harper       Bryce 0.658             16.0              92.5
## 4           Seager       Corey 0.647             11.9              93.2
## 5            Ozuna     Marcell 0.635             16.4              93.0
## 6             Rios       Edwin 0.623             14.5              91.5
## 7            Perez    Salvador 0.618             14.2              91.0
## 8        Tatis Jr.    Fernando 0.614              8.7              95.9
## 9        Hernandez     Teoscar 0.611             15.3              93.3
## 10            Belt     Brandon 0.597             18.0              90.7
## 11       Acuna Jr.      Ronald 0.596             18.6              92.4
## 12           Abreu        Jose 0.587             10.9              92.9
## 13           Myers         Wil 0.584             13.4              91.0
## 14        Voit III        Luke 0.584             15.2              88.9
## 15    Springer III      George 0.563             18.3              88.7
```

```
## 16        Smith     Dominic 0.563              10.8             89.8
## 17       Stassi         Max 0.561              15.0             91.6
## 18         Lowe     Brandon 0.550              18.1             89.8
## 19        Walsh       Jared 0.549              13.1             88.1
## 20       Turner      Justin 0.548              17.5             90.3
## 21      Jimenez        Eloy 0.544               5.7             92.4
## 22  Castellanos        Nick 0.542              16.5             91.0
## 23      Machado       Manny 0.542              15.6             90.2
## 24     Cruz Jr.      Nelson 0.540               9.4             91.6
## 25       Slater      Austin 0.540              10.9             89.2
## 26      Stanton   Giancarlo 0.539               8.3             91.1
## 27  Cronenworth        Jake 0.538              10.6             89.8
## 28        Judge       Aaron 0.535              15.7             92.2
## 29     d'Arnaud      Travis 0.533               8.1             93.4
## 30     Bichette          Bo 0.532              12.0             89.2
## 31       Dalbec       Bobby 0.532              14.8             89.9
## 32     Longoria        Evan 0.522              10.7             91.7
## 33       Winker       Jesse 0.521              10.5             92.1
## 34       Cooper     Garrett 0.518               9.5             90.1
## 35       Hosmer        Eric 0.516               8.7             90.8
## 36      Cabrera      Miguel 0.515              12.1             93.2
## 37       Castro       Willi 0.512              11.3             85.3
## 38  Goldschmidt        Paul 0.510              11.7             89.2
## 39      Calhoun        Kole 0.509              17.0             89.4
## 40       Turner        Trea 0.509               9.5             90.5
## 41     Anderson         Tim 0.506               6.7             87.2
## 42       Gyorko        Jedd 0.505              15.3             88.6
## 43      Pollock          AJ 0.499              13.0             89.6
## 44     Iglesias        Jose 0.498               7.6             86.2
## 45     Realmuto        J.T. 0.498              11.6             90.2
## 46    Arozarena       Randy 0.496               9.2             90.3
## 47    Dickerson        Alex 0.494              17.8             90.9
## 48       Tellez       Rowdy 0.490              10.6             90.7
## 49       Taylor       Chris 0.489               9.1             88.0
## 50      Urshela         Gio 0.488              12.3             91.4
## 51  Yastrzemski        Mike 0.488              18.4             88.2
## 52        Soler       Jorge 0.487              15.5             92.5
## 53    Bellinger        Cody 0.487              16.6             89.3
## 54        Hayes     Ke'Bryan 0.486              7.4             92.8
## 55       Castro       Jason 0.485              19.7             92.7
## 56         Cron        C.J. 0.484              20.6             85.5
## 57  Gurriel Jr.     Lourdes 0.484              10.5             90.8
## 58      Grisham       Trent 0.482              13.5             88.3
## 59         Cano    Robinson 0.481               6.2             90.4
## 60       Miller        Brad 0.481              11.2             89.8
## 61        Betts      Mookie 0.481              18.5             90.7
## 62       France          Ty 0.481              14.8             85.7
## 63       Taylor  Michael A. 0.480              12.4             89.0
## 64         Sano      Miguel 0.480              20.2             95.2
## 65    McCutchen      Andrew 0.479              18.2             89.7
## 66        Votto        Joey 0.478              15.4             87.4
## 67     Conforto     Michael 0.478              11.0             88.4
## 68      Jeffers        Ryan 0.477              10.3             91.6
## 69    Dietrich        Derek 0.476              12.6             85.4
```

```
## 70        Moreland       Mitch 0.476                 12.5                 88.2
## 71           Moran       Colin 0.476                  8.3                 91.9
## 72           Braun        Ryan 0.474                 15.3                 89.8
## 73            Bohm        Alec 0.474                  4.8                 90.2
## 74           Muncy         Max 0.473                 14.2                 88.5
## 75           Bruce         Jay 0.472                 15.3                 89.0
## 76          Suarez     Eugenio 0.472                 17.9                 89.1
## 77          Seager        Kyle 0.472                 17.7                 89.1
## 78           Osuna        Jose 0.472                 14.7                 89.1
## 79         Swanson      Dansby 0.472                 14.7                 89.1
## 80        Bogaerts      Xander 0.471                  8.7                 89.0
## 81          Naquin       Tyler 0.469                 13.0                 91.7
## 82      Candelario      Jeimer 0.469                 13.3                 90.2
## 83          Yelich   Christian 0.468                  7.1                 94.0
## 84           Riley      Austin 0.468                 13.6                 91.0
## 85            Nola      Austin 0.467                 12.5                 89.7
## 86          Rendon     Anthony 0.467                 19.5                 90.1
## 87            Happ         Ian 0.464                  9.0                 91.1
## 88         Grichuk      Randal 0.463                 12.6                 88.9
## 89           Moore       Dylan 0.461                 17.3                 90.4
## 90          Alonso        Pete 0.459                 15.5                 90.2
## 91           Gomes         Yan 0.458                 17.6                 89.9
## 92          McCann       James 0.458                 15.0                 90.5
## 93          Devers      Rafael 0.458                 10.6                 93.0
## 94          Robert        Luis 0.458                 16.7                 87.9
## 95            Choo    Shin-Soo 0.455                 11.4                 90.0
## 96           Story      Trevor 0.455                 20.8                 89.9
## 97         McBroom        Ryan 0.455                 20.1                 89.0
## 98        LeMahieu          DJ 0.453                  2.3                 91.3
## 99         Trevino        Jose 0.453                 10.1                 87.9
## 100        Valaika         Pat 0.453                 18.2                 88.7
## 101         Arraez        Luis 0.453                 12.1                 87.5
## 102           Pham       Tommy 0.452                  2.4                 92.8
## 103        Aguilar       Jesus 0.452                 16.0                 89.3
## 104         Walker   Christian 0.452                 11.5                 90.4
## 105        Heyward       Jason 0.451                 11.3                 87.6
## 106         Pinder        Chad 0.451                 13.7                 92.3
## 107   Kendrick III       Howie 0.450                  7.3                 88.8
## 108         Garcia       Leury 0.450                  7.5                 87.3
## 109       Pederson         Joc 0.450                 12.5                 93.0
## 110       Harrison        Josh 0.449                 15.4                 83.8
## 111         Tucker        Kyle 0.449                 14.9                 91.1
## 112            Ruf       Darin 0.448                 12.5                 89.4
## 113          Olson        Matt 0.447                 19.6                 92.3
## 114        Frazier       Clint 0.447                 11.6                 89.4
## 115        Santana      Carlos 0.445                 12.2                 88.0
## 116     Marmolejos        Jose 0.445                 10.5                 90.5
## 117       Martinez        J.D. 0.444                 14.7                 89.5
## 118         Farmer        Kyle 0.443                 18.2                 89.6
## 119      Contreras     Willson 0.441                  9.1                 89.8
## 120         Jansen       Danny 0.441                 16.3                 85.1
## 121        Carlson       Dylan 0.441                  9.3                 87.4
## 122         Murphy        Sean 0.440                 14.7                 92.2
## 123     Merrifield        Whit 0.439                 15.8                 86.1
```

```
## 124          Lindor   Francisco 0.439             13.5              89.9
## 125       Cervelli   Francisco 0.438             13.3              89.0
## 126   Guerrero Jr.    Vladimir 0.437              4.6              92.5
## 127            Shaw      Travis 0.436             19.8              90.9
## 128           Pence      Hunter 0.435              4.2              87.6
## 129         Rosario       Eddie 0.435             18.1              87.5
## 130           Reyes     Franmil 0.434             11.2              92.4
## 131       Schwarber        Kyle 0.434              8.8              92.8
## 132           Hiura      Keston 0.433             14.3              87.4
## 133           Hicks       Aaron 0.432             11.1              88.2
## 134           Nunez      Renato 0.431             21.1              86.3
## 135           Lewis        Kyle 0.431             11.1              88.3
## 136           Upton      Justin 0.430             18.6              91.7
## 137            Kemp        Matt 0.430             11.9              85.3
## 138       Moustakas        Mike 0.430             16.3              88.8
## 139          Flores      Wilmer 0.430             19.0              87.9
## 140     Mountcastle        Ryan 0.430             10.8              87.4
## 141        Blackmon     Charlie 0.429             13.5              86.9
## 142          Pillar       Kevin 0.429             13.4              87.1
## 143          Ohtani      Shohei 0.429              9.2              89.1
## 144           Davis        J.D. 0.428              3.3              90.1
## 145           Reyes      Victor 0.428             10.7              90.0
## 146         Cabrera    Asdrubal 0.427             13.7              89.5
## 147            Lamb        Jake 0.427             16.7              90.2
## 148         Sanchez        Gary 0.427             19.2              91.8
## 149        Severino       Pedro 0.427              9.5              87.6
## 150        Anderson       Brian 0.426              9.6              87.4
## 151           Jones      JaCoby 0.425             11.0              89.5
## 152           Arcia     Orlando 0.425              9.6              89.0
## 153           Nimmo     Brandon 0.425              7.6              87.2
## 154          Kepler         Max 0.422             21.9              88.5
## 155           Bader    Harrison 0.422             15.7              86.0
## 156        Tsutsugo       Yoshi 0.421             17.2              90.2
## 157       La Stella       Tommy 0.420             17.1              88.0
## 158           Ramos      Wilson 0.419              6.5              89.0
## 159          Guzman      Ronald 0.419              6.0              86.3
## 160            Bote       David 0.418              9.3              92.4
## 161           Marte    Starling 0.417              6.5              87.1
## 162         Polanco     Gregory 0.417             20.8              92.9
## 163            Odor     Rougned 0.417             20.7              86.0
## 164          Albies       Ozzie 0.417             17.8              86.7
## 165          Solano     Donovan 0.415             15.5              88.5
## 166       Brosseau        Mike 0.415             15.1              90.9
## 167   Brantley Jr.     Michael 0.413             10.2              88.7
## 168       Donaldson        Josh 0.413              6.2              92.8
## 169           Rizzo     Anthony 0.411             16.7              87.7
## 170           Canha        Mark 0.411             19.4              89.7
## 171           Knapp      Andrew 0.409             11.9              86.8
## 172          Arroyo   Christian 0.409              7.8              89.9
## 173        Barnhart      Tucker 0.408             19.7              85.4
## 174       Vogelbach      Daniel 0.408             10.2              89.5
## 175           Davis       Khris 0.407             18.9              87.7
## 176           Gamel         Ben 0.407             11.5              87.9
## 177          Profar   Jurickson 0.406             11.0              87.2
```

```
## 178         Castro      Harold 0.406             11.6              89.0
## 179         DeJong        Paul 0.406             21.7              89.2
## 180         McNeil        Jeff 0.405             11.5              86.6
## 181        Gurriel        Yuli 0.404             13.9              89.3
## 182      Carpenter        Matt 0.404             17.0              88.2
## 183          Perez     Michael 0.404             14.5              87.3
## 184        Grandal     Yasmani 0.403             15.6              90.3
## 185       Laureano       Ramon 0.403             12.3              87.7
## 186       Reynolds       Bryan 0.403             10.2              87.5
## 187          Dubon    Mauricio 0.402             16.7              86.2
## 188       Gonzalez      Marwin 0.401             12.6              89.2
## 189         Correa      Carlos 0.401             11.4              88.6
## 190     Nottingham       Jacob 0.401             20.1              87.0
## 191      Dickerson       Corey 0.400             11.1              85.7
## 192       Piscotty     Stephen 0.398             13.5              88.1
## 193       Grossman      Robbie 0.397             15.2              89.0
## 194      Hernandez       Cesar 0.396              5.6              89.1
## 195       Crawford     Brandon 0.396             12.4              88.7
## 196           Diaz       Elias 0.396             12.9              87.4
## 197           Ward      Taylor 0.396             10.4              91.1
## 198        Renfroe      Hunter 0.395             17.3              89.4
## 199         Pujols      Albert 0.394             16.4              88.6
## 200        Escobar     Eduardo 0.394             18.1              88.6
## 201      Gregorius        Didi 0.394             17.8              83.8
## 202       Gonzalez        Erik 0.394              5.5              88.5
## 203         Franco      Maikel 0.394              8.4              86.7
## 204        McMahon        Ryan 0.394              9.2              90.1
## 205           Ford        Mike 0.393              8.4              89.7
## 206        Arenado       Nolan 0.392             19.1              87.8
## 207        Stewart          DJ 0.392             17.8              91.4
## 208         Adames       Willy 0.392             12.5              88.8
## 209          Tromp    Chadwick 0.392             10.1              90.1
## 210         Segura        Jean 0.391             11.2              87.7
## 211         Machin      Vimael 0.391              6.4              89.9
## 212        Bregman        Alex 0.391             17.3              88.9
## 213        Brinson       Lewis 0.390              7.9              88.1
## 214        Kingery       Scott 0.390             16.9              85.3
## 215         Andrus       Elvis 0.389              9.2              88.5
## 216        Torrens        Luis 0.388             13.9              93.0
## 217        Hampson     Garrett 0.388             14.4              86.3
## 218         Chavis     Michael 0.388             10.0              88.3
## 219          Solak        Nick 0.388              9.4              89.9
## 220         Castro     Starlin 0.387             16.8              87.1
## 221          Eaton        Adam 0.387              5.7              87.8
## 222          White        Evan 0.387             13.7              91.7
## 223       Haggerty         Sam 0.387             13.0              90.6
## 224         Schoop    Jonathan 0.386              8.7              87.2
## 225       Plawecki       Kevin 0.386             12.7              88.9
## 226         Molina      Yadier 0.385             12.3              84.7
## 227        O'Hearn        Ryan 0.384             11.9              90.7
## 228         Senzel        Nick 0.384             16.3              88.3
## 229      Hernandez     Enrique 0.383             16.2              88.5
## 230         Dozier      Hunter 0.383             17.0              86.4
## 231         Naylor        Josh 0.383              7.8              86.8
```

```
## 232          Diaz     Aledmys 0.382              10.2              87.3
## 233         Rojas      Miguel 0.381              11.9              87.3
## 234       Goodwin       Brian 0.381              20.7              89.9
## 235       Polanco       Jorge 0.381              16.1              86.6
## 236          Bell        Josh 0.381               5.9              91.7
## 237        Torres     Gleyber 0.381              14.9              88.6
## 238          Baez      Javier 0.380              10.3              89.4
## 239        Tejeda    Anderson 0.380              14.3              90.8
## 240        Thames        Eric 0.379              14.6              88.7
## 241         Adams        Matt 0.379              21.4              90.1
## 242        Alfaro       Jorge 0.379               2.8              89.2
## 243       Taveras       Leody 0.379              14.3              88.9
## 244        Margot      Manuel 0.378               7.5              89.4
## 245     Kiermaier       Kevin 0.377              -0.4              87.7
## 246      Locastro         Tim 0.377              17.3              85.6
## 247          Lowe   Nathaniel 0.377               7.1              88.9
## 248        Garcia        Luis 0.377              -3.6              83.5
## 249       Frazier        Adam 0.376              12.3              85.5
## 250         Engel        Adam 0.376              14.0              87.2
## 251       O'Neill       Tyler 0.376              15.1              88.0
## 252         Sisco      Chance 0.376              20.0              88.8
## 253        Garcia     Avisail 0.375               8.7              87.4
## 254         Marte       Ketel 0.375              10.0              89.2
## 255         Ahmed        Nick 0.374               9.4              87.7
## 256   Bradley Jr.      Jackie 0.373               4.4              88.3
## 257        Maybin     Cameron 0.372               8.9              87.8
## 258       Peralta       David 0.371               6.4              89.2
## 259       Gardner       Brett 0.370              15.0              89.2
## 260        Wendle        Joey 0.370               5.1              86.7
## 261       Verdugo        Alex 0.369               5.9              87.0
## 262      Olivares      Edward 0.366               7.6              82.7
## 263         Edman       Tommy 0.366               8.2              86.5
## 264        Mazara       Nomar 0.364               6.7              91.0
## 265       Frazier        Todd 0.363              20.5              87.8
## 266      Markakis        Nick 0.361               9.5              89.0
## 267       Gimenez      Andres 0.361              13.5              86.8
## 268     Maldonado      Martin 0.360              19.1              86.1
## 269      Gosselin        Phil 0.359              13.8              85.3
## 270       Reddick        Josh 0.358              19.1              85.9
## 271        Galvis      Freddy 0.358              13.6              87.0
## 272      Peterson        Jace 0.358               7.4              89.2
## 273        Peraza        Jose 0.357              20.0              85.4
## 274      Crawford        J.P. 0.357              11.5              85.8
## 275         Arauz    Jonathan 0.357               9.4              85.6
## 276      Martinez        Jose 0.356               6.2              87.9
## 277       Vazquez   Christian 0.356              14.4              88.4
## 278      Phillips       Brett 0.355              13.7              84.9
## 279          Ruiz         Rio 0.354              10.5              87.9
## 280       Mondesi  Adalberto 0.353              13.7              90.6
## 281        Gordon        Alex 0.352              11.6              82.8
## 282          Hays      Austin 0.352              11.0              87.0
## 283        Fowler      Dexter 0.351              11.2              84.5
## 284         Kelly      Carson 0.351              16.7              86.3
## 285      Madrigal        Nick 0.351               4.3              83.7
```

```
## 286       Fletcher       David 0.351              4.1              84.7
## 287            Jay         Jon 0.350             12.4              84.8
## 288          Beaty        Matt 0.350              7.9              90.0
## 289           Kemp        Tony 0.349             18.8              85.8
## 290         Mathis        Jeff 0.348             20.6              89.6
## 291         Suzuki        Kurt 0.348             18.0              83.9
## 292         Bryant        Kris 0.348             20.7              86.1
## 293         Altuve        Jose 0.347              9.3              86.7
## 294        Santana       Danny 0.347             14.7              90.9
## 295          Lopes         Tim 0.347              6.5              87.2
## 296         Varsho     Daulton 0.347             18.4              86.2
## 297        Wolters        Tony 0.346              8.2              84.1
## 298        Andujar      Miguel 0.346             13.9              85.9
## 299         Murphy      Daniel 0.345             15.5              85.1
## 300        Santana     Domingo 0.345              5.4              85.5
## 301          Panik         Joe 0.345              9.3              87.5
## 302       VanMeter        Josh 0.345             15.1              89.0
## 303   Chisholm Jr.        Jazz 0.345             15.6              87.1
## 304        Espinal    Santiago 0.345             14.5              87.3
## 305         Semien      Marcus 0.344             19.3              86.2
## 306       Sandoval       Pablo 0.343              8.0              91.9
## 307         Biggio       Cavan 0.343             16.7              87.4
## 308         Aquino   Aristides 0.341             10.6              82.2
## 309        Fuentes      Joshua 0.341             10.6              84.0
## 310   Kiner-Falefa       Isiah 0.340              0.8              87.2
## 311        Haseley        Adam 0.340              0.3              86.6
## 312       Hilliard         Sam 0.338             10.8              88.3
## 313      Guillorme        Luis 0.337             10.7              89.8
## 314        Goodrum        Niko 0.336             16.0              88.8
## 315         Hedges      Austin 0.336             18.6              90.2
## 316         Riddle          JT 0.335              6.8              88.8
## 317          Joyce        Matt 0.334             12.5              86.7
## 318           Toro     Abraham 0.334              7.6              86.1
## 319        Moncada        Yoan 0.334             13.9              87.8
## 320            Lux       Gavin 0.334             13.9              87.1
## 321        Rosario        Amed 0.333              4.2              86.5
## 322         Kipnis       Jason 0.332             18.2              86.2
## 323           Cave        Jake 0.332              8.5              87.4
## 324          Tapia      Raimel 0.331              1.8              85.3
## 325         Newman       Kevin 0.328              8.7              85.5
## 326          Straw       Myles 0.328             15.9              87.4
## 327          Smoak      Justin 0.327             15.3              89.6
## 328          Avila        Alex 0.326             12.6              85.3
## 329           Wade       Tyler 0.326             13.4              86.6
## 330        Akiyama       Shogo 0.324              2.9              85.1
## 331        Flowers       Tyler 0.322             16.2              93.0
## 332       Tauchman        Mike 0.321             10.6              84.9
## 333      Stallings       Jacob 0.320             12.4              88.6
## 334           Diaz       Yandy 0.320             -7.9              88.3
## 335       Heineman       Scott 0.318              8.3              87.0
## 336       Caratini      Victor 0.318              6.2              87.9
## 337         Murphy   John Ryan 0.317              7.3              83.0
## 338          Berti         Jon 0.316              7.2              86.6
## 339        Alberto      Hanser 0.316             13.2              82.3
```

```
## 340         Camargo      Johan 0.316             9.8              87.3
## 341       Gallagher        Cam 0.315            14.9              82.7
## 342         Bemboom    Anthony 0.315            11.8              86.7
## 343         Hoerner       Nico 0.312             0.8              87.5
## 344            Leon      Sandy 0.310            17.9              84.9
## 345          Sogard       Eric 0.309            15.2              84.4
## 346        Adrianza      Ehire 0.308            16.0              85.9
## 347         Mendick      Danny 0.308             7.4              86.2
## 348           Lopez      Nicky 0.308             1.4              84.9
## 349            Vogt    Stephen 0.304            21.2              87.3
## 350            Wong     Kolten 0.304            10.8              86.5
## 351       DeShields     Delino 0.303             5.8              84.0
## 352       Smith Jr.     Dwight 0.303             8.1              89.9
## 353           White        Eli 0.302            15.4              88.1
## 354       Bonifacio      Jorge 0.301            18.3              82.9
## 355        Long Jr.       Shed 0.301             2.6              87.1
## 356            Choi     Ji-Man 0.298            15.7              89.0
## 357      Mullins II     Cedric 0.297            15.6              88.6
## 358        Harrison      Monte 0.295             4.0              81.7
## 359           Rojas       Josh 0.295             5.4              86.0
## 360          Romine     Austin 0.294             7.7              87.9
## 361         Narvaez       Omar 0.294            18.7              81.6
## 362            Bart       Joey 0.292            12.6              89.0
## 363         Estrada     Thairo 0.290             2.9              83.5
## 364           Urias       Luis 0.290             2.3              87.7
## 365            Holt      Brock 0.288            10.5              84.0
## 366           Adell         Jo 0.288            11.4              90.6
## 367            Dahl      David 0.286            17.0              85.9
## 368         Barrero       Jose 0.282             7.7              86.6
## 369           Perez    Roberto 0.281            -2.5              86.0
## 370         Simmons  Andrelton 0.281             6.0              86.5
## 371          Vargas   Ildemaro 0.280             2.1              85.3
## 372           Quinn      Roman 0.280             9.7              85.8
## 373          Sierra  Magneuris 0.280             3.8              83.7
## 374          Robles     Victor 0.280            19.0              82.2
## 375          Barnes     Austin 0.279            16.2              86.9
## 376          Garver      Mitch 0.278            18.6              92.4
## 377          Villar   Jonathan 0.277             1.6              86.6
## 378         Cameron        Daz 0.275             9.3              87.1
## 379         Paredes      Isaac 0.270             7.5              86.5
## 380         Calhoun     Willie 0.269            14.4              89.3
## 381 Strange-Gordon        Dee 0.262            14.1              79.2
## 382        Inciarte      Ender 0.261             8.4              78.2
## 383        Heineman      Tyler 0.261            34.3              82.4
## 384     Hechavarria     Adeiny 0.260            11.2              82.9
## 385          Tucker       Cole 0.256            13.3              83.1
## 386        Chirinos   Robinson 0.254            13.8              84.5
## 387        Starling      Bubba 0.250             8.1              84.2
## 388         Rengifo       Luis 0.249             3.8              87.6
## 389           Dyson     Jarrod 0.238             5.7              82.7
## 390       Velazquez     Andrew 0.234             6.7              82.4
## 391             Lin    Tzu-Wei 0.234            13.7              85.5
## 392         Mercado      Oscar 0.227            14.4              88.2
## 393          Garcia       Greg 0.221            11.1              83.4
```

```
## 394        Kieboom      Carter 0.220             11.9                85.1
## 395     Benintendi      Andrew 0.208              8.6                85.2
## 396         Zimmer     Bradley 0.204             16.4                84.1
## 397          Davis       Chris 0.201              6.2                85.8
## 398          Ervin     Phillip 0.198             16.0                85.4
```

d. Create a modified version of one of your variables (many of you will *need* to do this, but even if you don't, I want to see that you can)

```
new_dataset2 <- new_dataset %>% mutate(team_1_ave_xslg = mean(team_one$xslg))

glimpse(new_dataset2)
```

```
## Rows: 414
## Columns: 6
## $ last_name         <chr> "Pujols", "Cabrera", "Mathis", "Choo", "Molina", "Ca~
## $ first_name        <chr> " Albert", " Miguel", " Jeff", " Shin-Soo", " Yadier~
## $ xslg              <dbl> 0.394, 0.515, 0.348, 0.455, 0.385, 0.481, 0.366, 0.4~
## $ launch_angle_avg  <dbl> 16.4, 12.1, 20.6, 11.4, 12.3, 6.2, 22.8, 7.3, 18.0, ~
## $ exit_velocity_avg <dbl> 88.6, 93.2, 89.6, 90.0, 84.7, 90.4, 85.4, 88.8, 83.9~
## $ team_1_ave_xslg   <dbl> 0.4768571, 0.4768571, 0.4768571, 0.4768571, 0.476857~
```

e. Look up and try out one new verb for data transformation. The RStudio data transformation cheat sheet is a fantastic place to start: https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf

For e., we'd recommend using `group_by` + `summarize`. You can group your data by one variable, and then see the mean (or similar) of another variable within each of those groups.

Use as many code blocks as you need for a-e

```
team_one %>%
  summarise(mid_slg = median(xslg), mean_slg = mean(xslg))
```

```
##   mid_slg  mean_slg
## 1   0.535 0.4768571
```

```
team_two %>%
  summarise(mid_slg = median(xslg), mean_slg = mean(xslg))
```

```
##   mid_slg  mean_slg
## 1   0.394 0.4015955
```

In conclusion, the players having launch angle from 22 to 28 degrees and exit velocity faster than 90 mph have higher number of mean and median values of xslg.

Hints

1.3 Try using these steps:

- Step 1: identify the variables you need
  - Listing neighborhood: `neighbourhood`
  - Host's neighborhood: `host_neighbourhood`
- Step 2: Filter the data to only include the rows where those variables are not equal. Look back to Module 2 (or look online) if you need a reminder about how to write "equal", "not equal", and so on in R.
- Step 3: How many rows are left in the filtered data?

Extra food for thought: how do "NA" (missing) values get handled here? Do you think that makes sense? Should you do something else with them, maybe using `is.na`?

1.4 The variable for number of listings is `host_listings_count`. You might want to make a new variable indicating if a host is a local host (your answer to 1.3 will help here!). There are many ways to use `mean` on a subset of data, but the best approach is one we introduce in Module 5: `group_by` + `summarize`. Try it out now if you can! For this problem, don't worry about NAs.

2.2

Use `mutate` for this. You can replace the original `price` variable, or name it something else. There are a couple things you can use on `price` inside the mutate:

- `parse_number`, a function in the `readr` package, does a good job of converting currency to numbers on its own.
- `str_extract` with `pattern = "\\d+"`, then `as.numeric`, will extract numbers from a string, then convert the new (sub)string to a number.
- `str_remove_all`, with `pattern = "[\\$|,]"`, then `as.numeric`, will remove all dollar signs and commas.

2.5

Check out these resources if you're not sure about interpreting box plots:

https://magoosh.com/statistics/reading-interpreting-box-plots/

https://www.youtube.com/watch?v=oBREri10ZHk

3.3

a. use select()
b. use filter()
c. use arrange()
d. use mutate()
e. use group_by(var1) %>% summarise(mean = mean(var2))