HW 4, STAT 450

Due: Friday, November 1

Directions: This assignment should be completed using Quarto and submitted to Canvas as a self-contained HTML or PDF file.

Reading: Chapters 5, 7, and 19 from R for Data Science (2e)

```r
# load packages
library(tidyverse)
library(nycflights13)
```

Exercise 1

Use `read_csv()` to read the data set `hate_crimes.csv` into R (Lecture 12). This data set was used for the FiveThirtyEight article Higher Rates Of Hate Crimes Are Tied To Income Inequality. A description of the variables can be found at this link:

https://github.com/fivethirtyeight/data/tree/master/hate-crimes

(a) The Gini Index is a measure of income inequality.[1] The Gini Index is between 0 and 1, where values closer to 1 indicate greater income inequality. Which states have the highest Gini Index? Which states have the lowest Gini Index? [Hint: use `arrange()`]

(b) Use `ggplot()` to make a scatter plot with `gini_index` on the x-axis and `avg_hatecrimes_per_100k_fbi` on the y-axis. Use `geom_smooth()` to add a smooth trend line to the scatter plot. Label the x-axis "Gini Index" and the y-axis "Average annual hate crimes per 100,000 residents". Describe the association between the two variables in the scatter plot, and identify any potential outliers.

Exercise 2

(a) What function would you use to read a file where values are separated with a semicolon `";"`?

(b) What function would you use to read a file where values are separated with a vertical bar `"|"`?

[1] https://en.wikipedia.org/wiki/Gini_coefficient

Exercise 3

Consider the following data from a Pew religion and income survey.

```
head(relig_income)
```

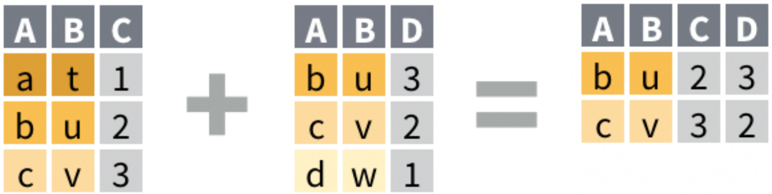
```
# A tibble: 6 x 11
  religion  `<$10k` `$10-20k` `$20-30k` `$30-40k` `$40-50k` `$50-75k` `$75-100k`
  <chr>       <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>      <dbl>
1 Agnostic       27        34        60        81        76       137        122
2 Atheist        12        27        37        52        35        70         73
3 Buddhist       27        21        30        34        33        58         62
4 Catholic      418       617       732       670       638      1116        949
5 Don’t kn~      15        14        15        11        10        35         21
6 Evangeli~     575       869      1064       982       881      1486        949
# i 3 more variables: `$100-150k` <dbl>, `>150k` <dbl>,
#   `Don't know/refused` <dbl>
```

Use the `pivot_longer()` function to reshape `relig_income` into a tidy data set, with the variables along the columns and observations along the rows. Your code should produce the following output:

```
# A tibble: 180 x 3
   religion income             count
   <chr>    <chr>              <dbl>
 1 Agnostic <$10k                 27
 2 Agnostic $10-20k               34
 3 Agnostic $20-30k               60
 4 Agnostic $30-40k               81
 5 Agnostic $40-50k               76
 6 Agnostic $50-75k              137
 7 Agnostic $75-100k             122
 8 Agnostic $100-150k            109
 9 Agnostic >150k                 84
10 Agnostic Don't know/refused    96
# i 170 more rows
```

Exercise 4

(a) What type of join operation is depicted below?



A	B	C
a	t	1
b	u	2
c	v	3

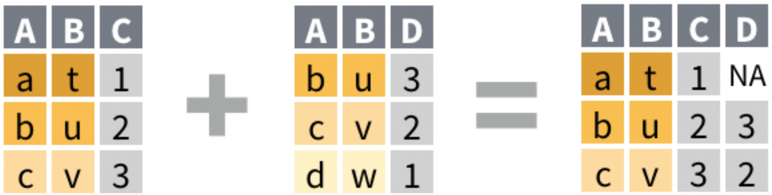
+

A	B	D
b	u	3
c	v	2
d	w	1

=

A	B	C	D
b	u	2	3
c	v	3	2

(b) What type of join operation is depicted below?



A	B	C
a	t	1
b	u	2
c	v	3

+

A	B	D
b	u	3
c	v	2
d	w	1

=

A	B	C	D
a	t	1	NA
b	u	2	3
c	v	3	2

Exercise 5

Verify that the column `tailnum` uniquely identifies each row in the `planes` table.

Exercise 6

Use `group_by()` and `summarize()` to compute the mean arrival delay for each flight destination. Then join that data frame of grouped summaries with the `airports` data frame, which contains information about each airport. This is what the resulting data frame should look like after the join:

```
# A tibble: 105 x 10
   dest  count arr_delay_mean name              lat    lon   alt    tz dst   tzone
   <chr> <int>          <dbl> <chr>           <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
 1 ABQ     254           4.38 Albuquerque ~   35.0 -107.   5355    -7 A     Amer~
 2 ACK     265           4.85 Nantucket Mem   41.3  -70.1    48    -5 A     Amer~
 3 ALB     439          14.4  Albany Intl     42.7  -73.8   285    -5 A     Amer~
 4 ANC       8          -2.5  Ted Stevens ~   61.2 -150.    152    -9 A     Amer~
 5 ATL   17215          11.3  Hartsfield J~  33.6  -84.4  1026    -5 A     Amer~
 6 AUS    2439           6.02 Austin Bergs~  30.2  -97.7   542    -6 A     Amer~
 7 AVL     275           8.00 Asheville Re~  35.4  -82.5  2165    -5 A     Amer~
 8 BDL     443           7.05 Bradley Intl   41.9  -72.7   173    -5 A     Amer~
 9 BGR     375           8.03 Bangor Intl    44.8  -68.8   192    -5 A     Amer~
10 BHM     297          16.9  Birmingham I~  33.6  -86.8   644    -6 A     Amer~
# i 95 more rows
```

Bonus

Use the data frame from Exercise 6 to visualize the spatial distribution of arrival delays. Here's some code to create a map of the United States:

```
library(maps)
library(mapproj)
states <- map_data("state")
ggplot() +
  geom_polygon(data = states, aes(x = long, y = lat, group = group),
               fill = "white", color = "black") +
  coord_map()
```

On this map, plot the coordinates (longitude, latitude) of each destination airport. Then use the `color` of the points to display the average delay time for each airport.[2] You might also what to use `filter()` to remove the airports located in Alaska and Hawaii.

[2] I recommend using the `viridis` color scale: https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html