Poli281

Spring, 2020

Assignment 4

Questions 1-3 are worth 10 points each. Q4 is OPTIONAL, for up to 3 bonus points. Please submit this filled-in word document, as well as an R script file showing how all your figures were made.

Q1: I have a conjecture that Raleigh police officers pull fewer people for traffic violations in the winter than in the summer. After all, police officers don’t necessarily like getting out of their cars when the weather is crummy!

Using our traffic stops dataset, produce a figure illustrates whether the data support my conjecture. Specifically,

* Make a bar graph using geom\_col()
* Make the category variable be month in which the stop occurred (January, February, etc.)
* Make the outcome variable be the number of stops that occurred. (Note: NOT the number of searches, as we have been examining so far.)
* For tidiness, put the categories on the y-axis, similar to Figure 1 from our in-class example.
* Give the x-axis an informative label.

Here are hints and tips:

* You’ll need to have R consider the month as a factor variable. And to get things in the right order on the graph (January at the top, December at the bottom), you’ll need to specify the order manually. See the template for how to do this.
* You can use a new function—tally—to count up the number of stops. (See [here](https://dplyr.tidyverse.org/reference/tally.html) for an example.)

Place the figure your produced below:

Chart, bar chart

Description automatically generated

After you make this figure, comment on whether or not it supports my conjecture:

The figure does show that July has the highest number of stops, but January also has the second highest over other summer months like June and August. However, other winter months have fewer stops than the summer months – most notably November and February. November looks like an outlier from the rest of the data, so the fewer stops might not have to do anything with it being a winter month. Overall, there may be a slight difference between summer months (June, July, August), and winter months (December, January, February) when looking at the total number of stops during the time. Per month, however, I am unable to conclude firmly whether there is a notable difference in the number of traffic stops on a seasonal basis.

Q2: Next, turn your attention to the jobs.csv file Sakai. This is real data compiled by the Census Bureau and the Bureau of Labor Statistics. ([Source](https://github.com/rfordatascience/tidytuesday/tree/master/data/2019/2019-03-05), but use the dataset posted to Sakai.) Import this dataset as an object called jobs. This dataset has information about male and female earnings, by occupation and year. Here is a description of variables:

|  |  |
| --- | --- |
| year | year |
| occupation | Specific occupation (most specific category) |
| major\_category | Broad job sector (8 levels) |
| minor\_category | More specific job sector (23 levels) |
| workers\_male | Estimated male workers in this occupation |
| workers\_female | Estimated female workers in this occupation |
| total\_earnings | Estimated median earnings in this occupation |
| total\_earnings\_male | Estimated male median earnings |
| total\_earnings\_female | Estimated female median earnings |

Your first job is to make a figure that examines the wage gap between male and female workers, and how it varies by broad job (major\_category) sector. Please do this as follows:

1. Use “mutate” to create a new column in the jobs dataframe called f\_pct\_m. Define this new variable as total\_earnings\_female divided by total\_earnings\_male. Of course, this variable represents female earnings as a proportion of male earnings, for a particular occupation.
2. Use dplyr tools to calculate the *weighted* average of f\_pct\_m, where each occupation is weighted by total\_workers. (The idea is that sectors with more workers should count more heavily in our average, just like assignments that are more important count more heavily in your final course grades.) Here is an example of dplyr syntax that calculates the weighted average of the variable x, with w representing the weights:  
     
   summarize(results = weighted.mean(x, w, na.rm = TRUE))
3. Create a bar graph, using the same basic approach as in Figure 1 of our in-class exercise. You should flip the axes for readability. I also recommend placing the categories in descending order and adding a reference line at 1.0.

Place the figure your produced below:

Chart, bar chart

Description automatically generated

After you make this figure, comment on what it shows:

The figure shows that the greatest wage difference is in management, business, and financial, because the weighted average of the proportion of female earnings compared to male earnings is furthest from 1, which signifies equal wages. Overall, the highest weighted average wage proportion is in natural resources, construction, and maintenance, which lies around 0.87. Both the education/legal/community service/arts/media category and the computer, engineering, and sciences category have about the same weighted average. Unfortunately, the ultimate value that stands out is the 1.0, which shows that no category of work listed has equal wages for male workers and female workers – on average, male workers have higher earnings than females.

Q3: I have a second conjecture. It is that the wage gap is smaller in occupations that are a higher proportion female. (My thinking is that there will be more female managers and executives in these occupations, and they will work harder to address wage differentials.)

Make a figure that assesses my conjecture. Please do this as follows:

1. Limit your analysis to the Computer, Engineering, and Science sector. (This is the only sector I want my theory to apply to.)
2. Create a variable called prop\_male, defined as the proportion of workers in each occupation that are male. For instance, if there were 45 male workers and 27 female workers in a particular occupation, that occupation’s value for prop\_male would be .625.
3. Remove two outlier observations (see template)
4. Create a scatterplot where prop\_male is the x-axis and f\_pct\_m is the y-axis.
5. To pretty it up,
   1. Give the axes informative names
   2. Add a reference line at y=1.0
   3. Add a line of best fit, using: stat\_smooth(method="lm", se=FALSE)

Place the figure your produced below:

Chart, scatter chart

Description automatically generated

After you make this figure, comment on what it shows:

The figure shows that the general average weighted proportion wage gap gets closer to 1 as the job becomes more male-dominated. However, the increase is at a fairly slow rate, meaning that there is a general tendency throughout for the wage gap to be at least 0.7 no matter the proportion of male workers in the industry. There are outliers around weighted = 0.5 and above 1.25, the most notable being the point at average weighted wage gap = 1.50. The scatter of the graph is fairly centralized to the line of best fit otherwise.

Q4: This question is optional, and is worth up to three bonus points. (two points for a good figure, and three points for an outstanding figure.) Also, this is an **individual** question. You cannot work together on this question.

Using *any* of the datasets we’ve been focusing on—soccer, traffic stops, or jobs—create a visually appealing figure that tests an original idea that you had. The figure should involve at least two variables (i.e. it can’t just be a histogram), but other than that you have free rein.

Below, paste the figure you produced, and write one paragraph that explains what idea it tests, and what it reveals.