Homework 5: Heterogeneity and Data Visualization

Econ 145

Overview

In this homework assignment, you will be using a sub-sample of arrest data from Los Angeles, and a sub-sample of arrest data from New York City.

Below is a short description of the columns that are most important in the Los Angeles arrest data set:

- `arrest_date` - The date of the arrest.
- `hour` - The hour of the time of arrest in military time (e.g. 00-23).
- `area_name` - The 21 geographic areas or patrol divisions within the city.
- `age` - The age of the arrested individual.
- `sex_code` - The sex of the arrested individual.
- `charge_group_description` - The category of the arrest charge.

Likewise, here is a short description of the columns that are most important in the New York City arrest data:

- `arrest_date` - The date of arrest.
- `ofns_desc` - A description of the arrest charge.
- `age_group` - The age group of the arrested individual.
- `perp_sex` - The sex of the arrested individual.
- `perp_race` - The race of the arrested individual.
- `longitude` - The longitude coordinate of the crime.
- `latitude` - The latitude coordinate of the crime.

To Receive Credit

- Save the scripting file as assignment_5.R. Make sure your capitalization is correct as the autograder is case-sensitive.
- Make sure all changes to the original dataset are done within the R script.
- Your one page write-up must be submitted in a .pdf to receive credit.

Grading on Coding Questions

Grading on the coding portion of the homework will come in two types of questions: *Public Questions* and *Private Questions*. Public Questions can be submitted as many times as you like to the autograder, and the autograder will give detailed feedback. Additionally, the TAs will help you on Nectir and in office hours on the Public Questions. On the other hand, Private Questions can be thought of as a mini quiz within the homework. While you still have as many times to upload your answer as you want, the autograder will not provide any feedback, and the TAs and Professor Startz will not provide any guidance or assistance (but getting advice from classmates on Nectir or elsewhere is completely okay). Private Questions will be marked on the homework assignment.

Part 1: Coding Assignment

For each graph that you are producing, please make sure your graph has the following features before submitting:

- Meaningful, readable labels for axes and categories.
- A legend is included when necessary.
- The graph is self-explanatory, i.e. someone who is not familiar with the context and data could understand its main message.

Important: For this homework, comment out any of your graphs that you created when submitting your code to Gradescope. You will submit your graphs separately on your write-up. Gradescope will fail to autograde your script if you include any graphs.

1. Read in the Los Angeles arrest data and assign it the name la_arrests. We will be working with this data first.
2. *What is the age distribution of individuals that are arrested in Los Angeles?*
    a) As a first plot, let's find the distribution of arrests by age using a density plot (e.g. `geom_density`). Change the `fill` argument to "blue" and the `alpha` argument to 0.5. Additionally be sure to label your axis and give a title (you can do this with the `labs` layer).
    b) Modifying your code from 2a, use the `fill` argument in the `aes` function to show the age distribution by `sex_code`. Does there appear to be any difference in the distribution between the sexes? Your graph should be an overlaid density plot (you can Google this if you need reference on what the graph should look like).
3. *What are the most frequent crimes individuals are arrested for in Los Angeles?*
    a) The purpose of this question will be to make a horizontally stacked barplot (e.g. the crimes on the y-axis, and the total number of crimes on the x axis, the crimes in order from most frequent to least frequent) of the most frequent crimes that individuals are arrested for in Los Angeles. To do this, we will need to make a few adjustments to the data. First, `count` the number of times each `charge_group_description` occurs setting the `sort` argument to `TRUE`. FIlter out any NA values. Save this tibble and assign it the name frequent_crimes.
    b) Using the `mutate` function, and the `fct_reorder` function (look up the documentation!), `fct_reorder` the `charge_group_description` column by `n` (e.g. the number of times the charge group occurred). Without this step, your stacked barplot will not be in descending order. Update the frequent_crimes tibble to reflect these changes (e.g. your tibble should have 2 columns and 15 rows).
    c) Now, using `ggplot2` and the `geom_col` function, create the stacked barplot. Be sure to label your axes and provide a title.
    d) (*Private Question*) Now, modify your plot from part c to show the difference in these crimes between men and women. You can do this in anyway you like, however, some potential options could be: using the `fill` argument in your previous plot, using the `facet_wrap` function in `ggplot2`. Do this in a way that is visually appealing and easy to understand to the reader.
4. Read in the New York City arrest data and assign it the name nyc_arrests.
5. *Arrest trends in New York City.*
    a) The purpose of this question will be to find if there are time trends of arrests in New York City. More specifically, were there any policies that contributed to more or less arrests that we can see in a visualization? We are going to need to do this in a few steps. First, `group_by` the `year` and `month`, and then summarize the total amount of crime that occurs in each year/month. Save this tibble as total_crime. It should have three columns: `year`, `month`, and a `total_crime` column which is the sum of all crime within a year/month.

b) Next, we want to make a new column that is of the `date` data type. The reason we want to do this, is because `ggplot2` handles date types very well (e.g. it automatically puts all of the dates in order, and has some nice customizable features). To do this, create a new column called `date` in the total_crime tibble. However, creating a date takes a little work. To do this, we will use both the `lubridate::ymd` function, and the `paste0` function. Notice that the `lubridate::ymd` function converts strings that are of the format "yyyy-mm-dd". Hence, we need to create a string variable that has this format, then pass it through to the `lubridate::ymd`. We can do this using the `paste0` function. Notice that `paste0(year, "-", month, "-1")` will create the format we want, although it will set the day of the month to the first each time. However, this is ok for our purposes. You can follow the following code snippet for creating the `date` column:

```
total_crime <- total_crime %>%
  mutate(date = lubridate::ymd(paste0(year, "-", month, "-", "1")))
```

c) Now, create the time trend plot using `geom_line` or `geom_path`. Remember to set the `group` argument to the number `1`. Additionally, be sure to label your axis, and provide a title.

d) (*Private Question*) Modifying your code from the previous parts, create a time trend for each `age_group` (e.g. same graph as part c but with one line for each age group). **Hint: you will want to `group_by` an additional column and utilize the `color` argument**.

6. *Where do robberies occur in New York City?*

a) In this problem, we will be visualizing the location of arrests due to robberies in NYC. First, load in the `ggmap` package.

b) The `ggmap` package is a great tool for making visualizations of different parts of the world. However, since we are only focused on NYC, we want to limit our map to only a view of NYC. This can be done by setting the longitude and latitude boundaries (see this tool if interested!). Hence, we will set the boundaries of the map, and tell `ggmap` the styling of map we want. Run the following two lines of code:

```
bbox<- c(left = -74.35, bottom = 40.498, right = -73.687, top = 40.90)
nyc <-get_stamenmap(bbox, zoom = 11, maptype = "terrain")
```

c) Now complete the following code to create the visualization. Remember, you can pipe within the `data` argument of `geom_point` to filter the data to only robberies. Moreover, you will need to put in the `latitude` and `longitude` columns into the `aes` function.

```
ggmap(usa) +
   geom_point(data = , aes()), alpha = 0.7) +
   ggthemes::theme_map()
```

d) (*Private Question*) Much of data wrangling is finding other people codes, and modifying it to fit your situation. In this question, we will modify the map created in part c to show the locations of robberies by-race. Using your knowledge of `ggplot2`, modify the code to replicate Figure 1.

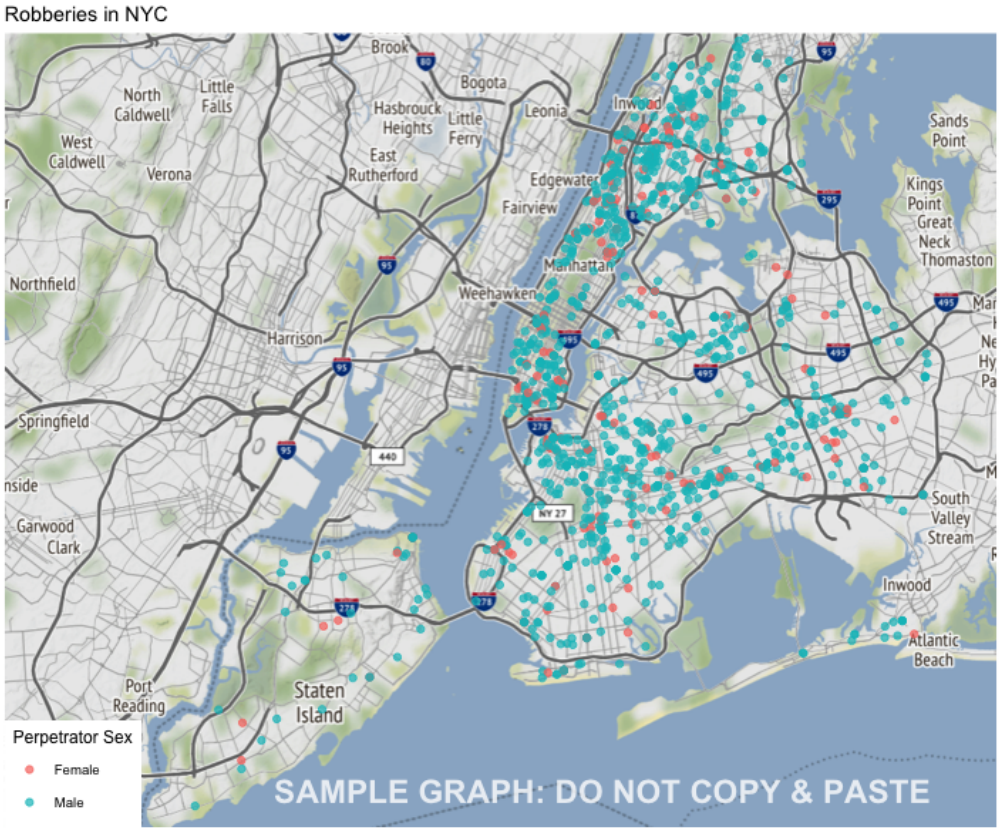

Figure 1: Visualization to be replicated in Exercise 6d

Part 2: Write-up

The purpose of this write-up is to help convey your visualizations through words in brevity and clarity. This will be an important skill for any job, as your managers will want clear and concise explanations for all of your analysis work. Clearly make sections defining which part of the write-up you are completing. This should look like a formal document you would want to submit to your supervisor. Note that this document can be more than 1 page since we anticipate the figures taking a lot of space.

- Include your graph from 2b. Your supervisor wants to know the following: Do you see any difference in the age distribution of arrests in LA between sexes? If there are differences, what might be causing these differences?
- Include your graph from 3d. Your supervisor wants to know the following: Are there certain crimes that females are arrested for more frequently than men or vice-versa?
- Include your graph from 5d. Your supervisor wants to know the following: What do you notice about the trend of arrests in NYC? What about by age group? Is there a reason for this? **Hint: you may want to look into "Stop and Frisk"**.
- Include your graph from 6d. Be sure to change the title of the graph so that it is: "FIRST NAME LAST NAME: Robberies in NYC by Gender". No need to write about this graph.