Supplemental Exercise - Data Manipuation & EDA

Jillian Morrison

October 1, 2019

Below are exercises you can use to practice data manipulation and EDA. See solutions on moodle.

Data Manipulation

`Fertility` dataset

1. Load the dplyr package. load the `Fertility` dataset from the `{AER}` package. Use `glimpse()` to see what is in the dataset.

```
> #install.packages("AER")
> library(AER)
> data("Fertility")
```

2. Save rows 35 to 50 of the age and work variables to a new dataset calles `Fert`. Hint: Use `slice()` and `%>%`
3. Count how many women proceeded to have a third child.
4. There are four possible gender combinations for the first two children. Which is the most common?
5. By racial composition what is the proportion of woman working four weeks or less in 1979?
6. Filter out a subset of woman between the age 22 and 24 and calculate the proportion who had a boy as their firstborn
7. Add a new column, age squared, to the dataset.
8. Calculate the proportion of women who have a third child by gender combination of the first two children?
9. Out of all the racial composition in the dataset which had the lowest proportion of boys for their firstborn.

Exploratory Data Analysis

`diamonds` dataset

1.Using `diamonds` dataset in `{datasets}` package, COnstruct a barplot of cut. Add colors, a legend, and titles to the plot.

2. Create boxplots of Price by cut of diamonds. Add titles and labels
3. Construct barplot of mean Price by cut and clarity
4. Construct barplot of mean carat by cut and clarity. Rearange the order of the grouping variables and choose the order that makes the most sense

5. Select the observations/diamonds that have carat less than 3. construct histograms of carat and group by cut.
6. Notice that you cannot say much about the histogram in 5 above. try using `geom_freqpoly()` instead of `geom_histogram()`. Compare the result to the result in 5.
7. Try the following code for a histogram:

What kinds of questions does this generate about diamonds?

`mpg` dataset

8. Create boxplots of highway mileage by class
9. Re-order the plot in 8 by the median for each class. Hint: for the x variable, use x = `reorder(class, hwy, FUN = median)`
10. Use layering (i.e. `+ coord_flip()`) to flip the plots 90 degrees to horizontal boxplots instead.