Munging and EDA Exam 2

P. McDonald

December 6, 2016

This is a 90 minute exam. You are to write an R Markdown document that provides scripts for carrying out the following tasks. It is the work, not the answers, which will be evaluated.

You may use the electronic resources at your disposal, but please do your own work.

Do as much as you can. Mail your pdf to mcdonald@ncf.edu at the end of the 90 minute period.

Navigate to the following page:

https://www.kaggle.com/lislejoem/us_energy_census_gdp_10-14

Read the annotation. Download the dataset and read it into memory (this may require you to register at kaggle - it's easy). Alternatively, grab the data from the canvas page under "files," then "Data."

Initial exploration

1. Vet the data for completeness. Briefly discuss your conclusions.
2. Filter out any observations which do not correpsond to US states.
3. Subset the energy variables to produce a data frame which contains information concerning coal and electricity data (as well as all variables not contained in the energy variables).
4. Using ggplot2, create histograms and boxplots for total coal consumption and total coal production for the year 2010. Identify any production outliers above median.
5. Using ggplot2, create a scatterplots for total coal consumption/production and total electricity consumption/production for the year 2010. Discuss correlation.

Aggregation

6. Aggregate total coal consumption and total coal production by region. Report the corresponding IQRs.
7. Aggregate total coal consumption and total coal production by division. Report the corresponding medians.

Mutation and filtering

8. For the year 2010, normalize total coal consumption and total coal production by population and add this as another variable. Report the range of the new variable
9. Filter the new variable to retain the observations corresponding to the top ten values. Report the state abbreviations of the 2010 populations of the corresponding states.

Plotting multiple variables

10. Using ggplot2, facet on year and generate boxplots for total coal consumption (you might consider reshaping the data).
11. Generate a scatterplot of coal production and population for the year 2010 and color by region. Generate a second plot colored by division.
12. Generate a scatterplot of coal production and population, colored by region, faceted on year.