Homework #3: Factors, aggregation with SQL, and ggplot2

The purpose of this homework is to give you more experience with data aggregation in R, use of factors, SQL access, and visualization using ggplot() on a real-world dataset.

There are two parts.  In the first part, you'll load data into a SQLite database using Python.  In the second part, you'll retrieve the data from your R code, aggregate to compute various summaries and visualize your statistics using ggplot().

The dataset contains year-by-year fuel economy data for many different manufacturers and models of vehicles. The data used was originally downloaded from [http://www.fueleconomy.gov/feg/ws/index.shtml  (Links to an external site.)](http://www.fueleconomy.gov/feg/ws/index.shtml) but you don't need to download the data again, as it is provided in the attached zip file **si618hw3.zip**

Your goal is to generate a report that looks like 'si618hw3\_sample\_solution.html' in the attached zip file. Your report doesn't need be pixel-for-pixel identical, but should capture the same set of data summaries with the same general graphic formatting as in the given plot forms.

Part 1. Data Preparation (30 points)

Write Python code to load data in the vehicles.csv file into a SQLite3 database named vehicles.db. You don't need to include all the columns, just these:

['year', 'make', 'model', 'VClass', 'cylinders', 'displ', 'trany',  'city08', 'highway08', 'comb08']

When this step is done, you should have a pretty big vehicles.db file on your hard drive.

Note that there are some rows in the vehicles.csv file with missing 'cylinders' data, and in some rows, displ is 0, which doesn't make sense for V8 engines. You should remove these rows from consideration and not load them into the database.

Part 2. Data Aggregation and Visualization in R (70 points)

In this second part, you'll write a RMarkdown document that will generate a report looking like the provided si618hw3\_sample\_solution.html. The steps involved and grading rubric are described there.  In the provided RMarkdown template si618hw3\_sample\_template.rmd, you'll need to fill in the R code.

You should use the ggplot() function in the ggplot2 package and functions in the plyr package to make your plots. You'll generate quite a few plots here, and you don't want to generate them all manually by cutting and pasting code. One way could be to write a "for" loop in R to generate the two plots for each VClass in one iteration of the loop.  (For an extra challenge, try to see if it's possible to do without the 'for' loop.)

Note that in your code, using the print statement with the ggplot object, like this:

ggobj = ggplot( \*your settings here\*)  \*and possibly here\*

print(ggobj)

should cause the plot to appear in the HTML file generated by RMarkdown.

What to submit:

A zip file named 'si618hw3\_youruniquename.zip' containing:

* Your Python code for Part 1 named 'si618hw3part1\_youruniquename.py'
* The R Markdown file you wrote for Part 2 named 'si618hw3report\_youruniquename.Rmd'
* The HTML report generated by your R Markdown file.