Munging and EDA Exam 1

P. McDonald

October 4, 2016

This is a 60 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 60 minute period.

Navigate to the following page:

https://archive.ics.uci.edu/ml/datasets/Automobile

Read the annotation.

Importing data

1. Read the data into R as a csv file.
2. Use the following vector to name the features of the csv file:

```
features <- c("symboling", "normalized-losses","make",
              "fuel-type","aspiration", "num-of-doors","body-style",
              "drive-wheels", "engine-location", "wheel-base",
              "length", "width","height", "curb-weight", "engine-type",
              "num-of-cylinders", "engine-size", "fuel-system",
              "bore", "stroke", "compression-ratio", "horsepower",
              "peak-rpm", "city-mpg", "highway-mpg", "price")
```

Initial vetting

3. How many complete cases are there?
4. Subset the data on complete cases.
5. Working with complete cases, what is the range of values for the feature "horsepower"? What is the mean value?

Subsetting and binning

6. Subset the data to include "make"" and ten of the last eleven variables, omitting "fuel-system". Assign the outcome to a file called "car_data_reduced". Convert the last ten variables of "car_data_reduced" to numeric.
7. Bin the "horsepower" feature as 5 intervals of equal length, with the right-hand-endpoint of the last interval determined by maximal value of the feature and the left-hand-endpoint determined by the minimal value of the feature. Add this "binned information" as a feature to car_data_reduced.
8. What is the make of the car belonging to the third interval and having maximal price?

Aggregation

9. Compute the median values for all variables in car_data_reduced except "make" and the binned information, aggregating on binned horsepower and the number of cylinders.
10. What pair of the last 10 variables (omitting the binned information) is maximally correlated? Construct a scatterplot for these variables.

Data reduction

11. Consider a data frame consisting of the last 11 features of the original data. Perform a principal component analysis using the tool of your choice. How many components are required to account for 90% of the data?

12. (Extra credit)Does the answer change if the variables are standardized prior to performing PCA?