140.886: Intro to R

Homework 2

Assigned Date: 1/8/2013

Due Date: 1/10/2013

Name:

Please submit:

1. Annotated R analysis script (emailed to me)
2. Answers written/typed on this sheet (turned in tomorrow morning)

For (a), the script should be source-able (ie have no errors) and divided into sections (corresponding to questions) using comments. The answer to the question should be the last line(s) of code evaluated in that section. For example:

#####

# q1

[code goes here]

[have answer print here]

Instructions:

1. Get the dataset: <http://biostat.jhsph.edu/~ajaffe/files/kaggleCarAuction.csv>
2. Read the “dictionary”: <http://biostat.jhsph.edu/~ajaffe/files/Carvana_Data_Dictionary.txt>

This is a dataset from the “Kaggle” website, which hosts competitions for prediction and machine learning. More details on this dataset are here:

<http://www.kaggle.com/c/DontGetKicked/details/Background>

Questions

1. Read in the dataset itself, naming the R object “cars”, and the data dictionary, naming it “key”, into R (as separate objects). Save them together in an “.rda” file so you can access the data offline.
2. How many cars are in the dataset? How many variables are recorded for each car?
3. What is the range of the manufacturer's years of the vehicles? How many cars were from before 2004, and what percent/proportion of the cars are these older models?
4. **Drop** any vehicles that cost less than $1000 – how many vehicles were removed, and how much were they? **The rest of the questions expect answers based on this reduced dataset.**
5. How many different vehicle a) manufacturers/makes b) models and c) sizes are there?
6. Which vehicle a) make, b) model and c) color had the highest average acquisition cost paid for the vehicle at time of purchase, and what was this cost?
7. Which vehicle a) make, b) model and c) color had the highest **variability** in acquisition cost paid for the vehicle at time of purchase?
8. Display the relationship between acquisition cost and mileage, and describe this relationship
9. Which variables of cost, odometer reading, and/or warranty (if any) visually appear to associate with a car being a “lemon”/bad purchase?
10. How many vehicles:
    1. Were red and have fewer than 50,000 miles?
    2. Are made by GM and were purchased in Florida?
    3. Are green or white?
    4. Are made by Mazda or Nissan and are black or silver?
    5. Are automatic, blue, Pontiac cars with under 60,000 miles?