Midterm Project Report

TIDY DATA IN AUTOMOBILS FUEL ECONOMY

Haiwei Zhou | MA615 | 10/24/2016

Abstract

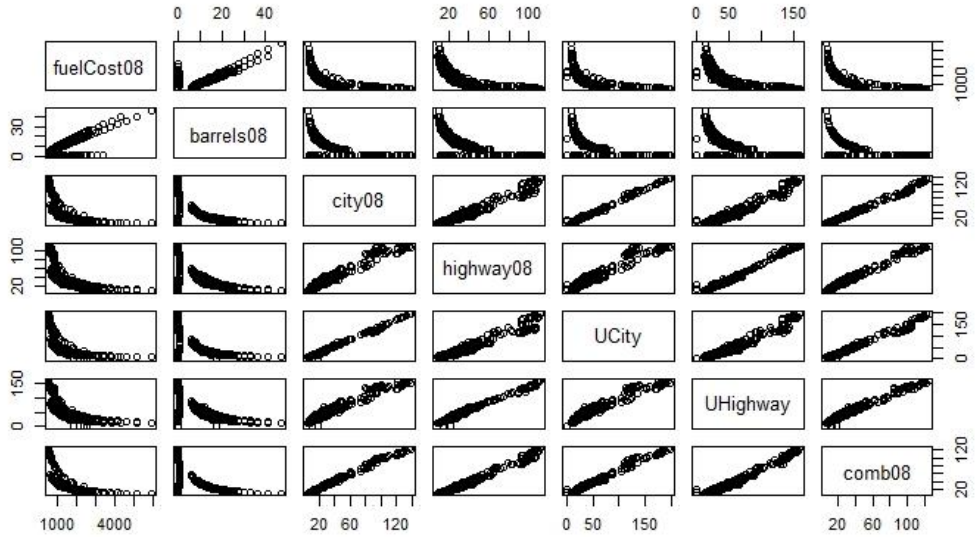
This project focuses on the data set of automobiles fuel economy. I use the tidy data tools and some package like ggplot2 to do some analysis of the clean data set. The data set "vehicles" records the data of automobiles fuel economy information from 1984 to 2017 including the data such as the type of the cars, the annual fuel cost of each type and some basic information about a vehicle especially on engines. Generally, there are two fuel types of vehicles —single fuel vehicles and dual fuel vehicles. Since the number of dual fuel vehicles is really small. In this project I only focus on single fuel vehicles. In the documents, there is an R file to demonstrate the tidy process, a raw data set and some new clean data sets. This report shows what I did to clean and organize the data.

Tidy Data Steps

1. Use read.csv to read the raw data set "vehicles" getting from http://www.fueleconomy.gov/feg/ws/index.shtml#vehicle
2. Since there are only 1422 duel fuel type vehicles from 38017 vehicles, I remove these data.
3. There are 83 columns and most of them are not needed in analysis. This step is to remove unneeded and inconsistent variables.
4. The existing variables are what we need. The missing data is given NA and some long strings are turned into letters.
5. Sort the data by years and vehicles' names and save as "vehicles1.csv".
6. Separate the data set "vehicles1" into several small data set and each data set deals with some variables such as fuel cost and mpg scores.
7. Use tidy-data tools to deal with every small data set.
8. Make some analysis to find the relationship between the data and some change in data by years.

Analysis of data

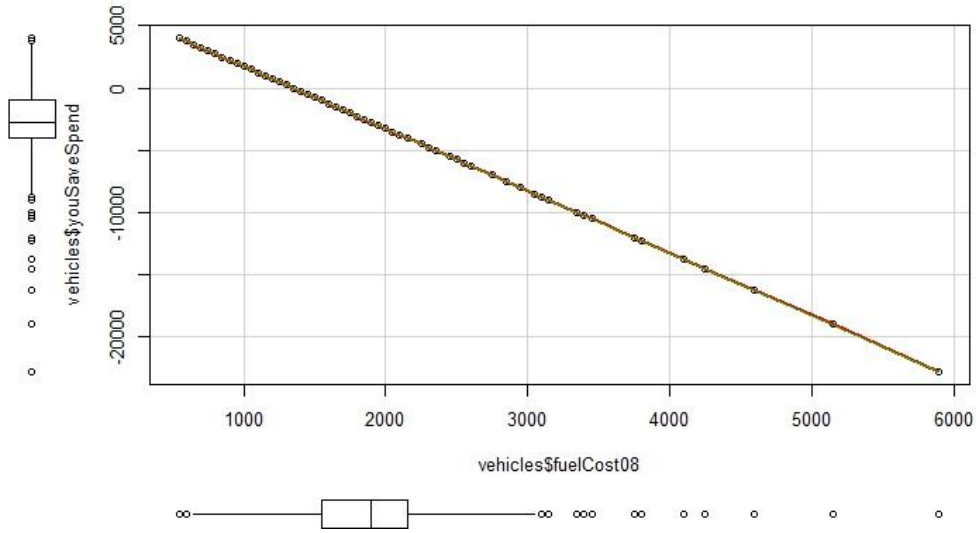
I. A group of variables represents the petroleum consumption and MPG data. It can be easily considered that there are some linear relationships between these data. Thus I make a scatterplot matrix to determine if the guess is right. The result is in Graph 1.



Graph 1: Scatter plot matrix

We can conclude that the linear relationship between annual petroleum consumption in barrels and the annual fuel cost is very strong. Similarly, the MPG data are all in strong linear relationship.

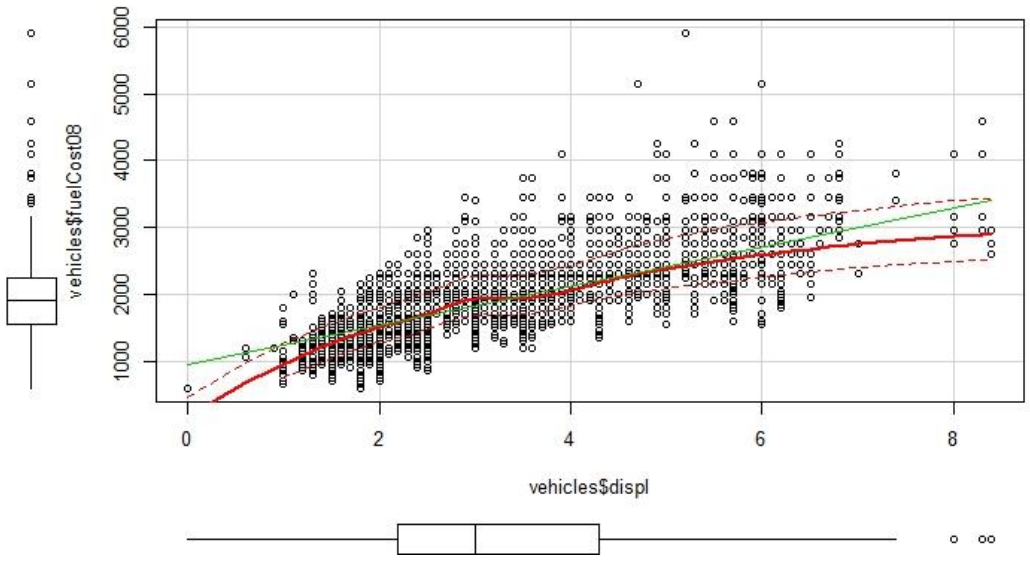
II. The most important data for customers are how much they will spend on a vehicle per year. Thus the relationship between fuel cost and annual spend deserves to be exploited.



Graph 2: scatterplot and linear regression

According to this result. We can find that the cost is a linear function of fuel cost.

III. To better understand what determines the fuel cost, studying the relationship between engine and fuel cost is needed.



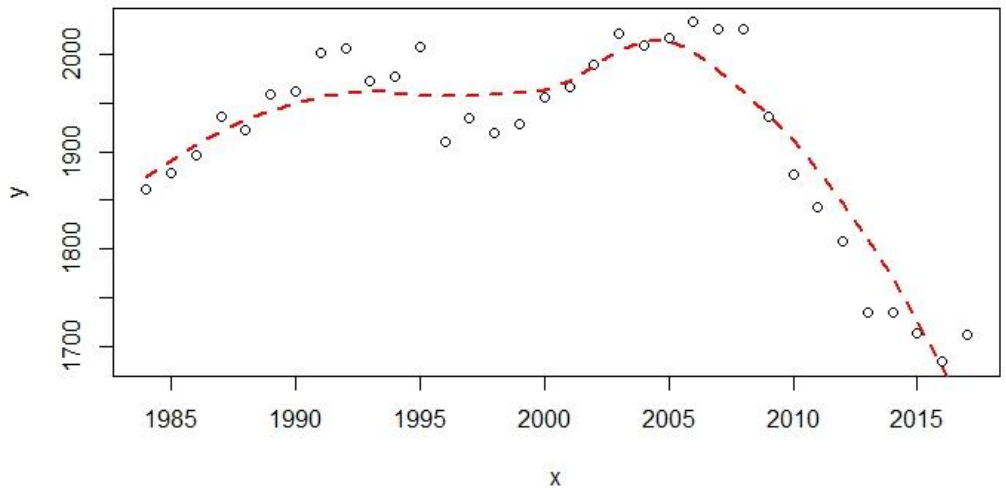
Graph 3: scatter plot and linear regression

We can find that the fuel cost is approximately linear to the engine displacement in liters.

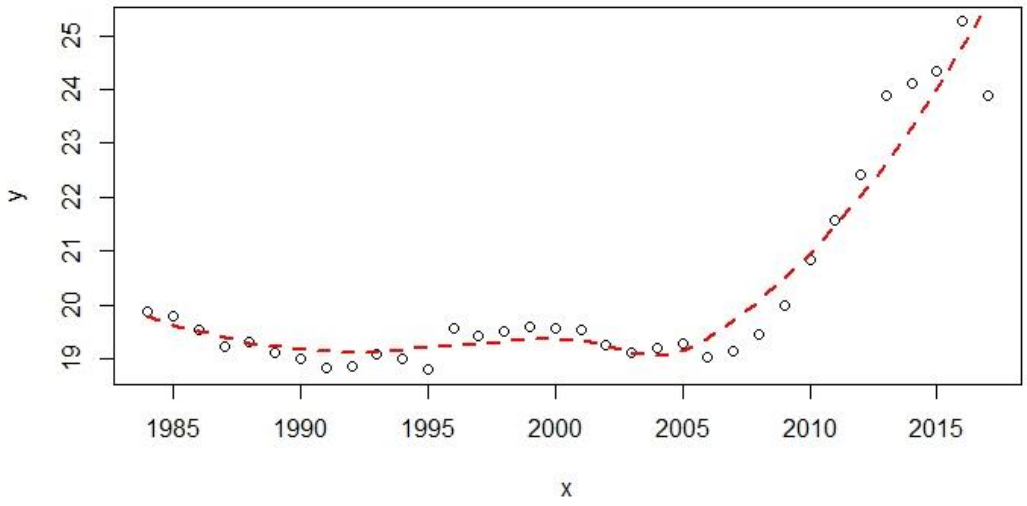
IV. The data change in years

Since the data is from 1984 to 2017. It's interesting to find the data change in years.

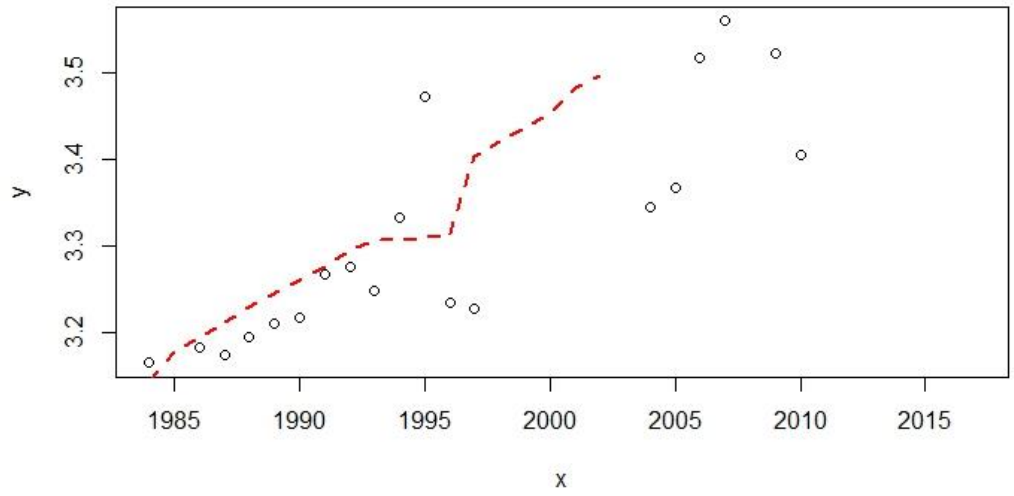
We use the mean of whole year's data to represents the data in that year. The following four graphs are the changing data of fuel cost, combined MPG, engine displacement and tailpipe CO_2 in mile.



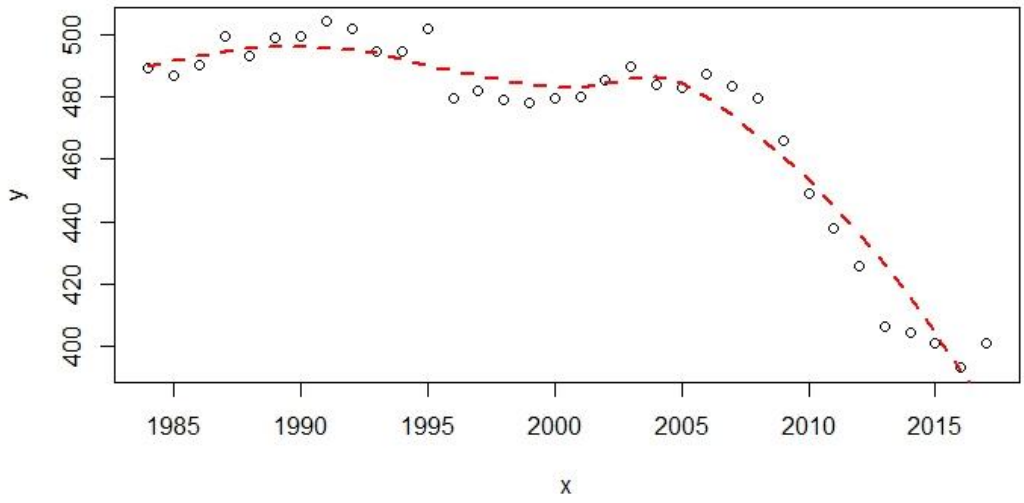
Graph 4: change of fuel cost



Graph 5: change of combined MPG



Graph 6: change of engine displacement



Graph 7: change of tailpipe CO_2 per mile

Appendix

Data Description

vehicle

atvtype - type of alternative fuel or advanced technology vehicle

barrels08 - annual petroleum consumption in barrels for fuelType1

barrelsA08 - annual petroleum consumption in barrels for fuelType2

charge120 - time to charge an electric vehicle in hours at 120 V

charge240 - time to charge an electric vehicle in hours at 240 V

city08 - city MPG for fuelType1 (2), (11)

city08U - unrounded city MPG for fuelType1 (2), (3)

cityA08 - city MPG for fuelType2 (2)

cityA08U - unrounded city MPG for fuelType2 (2), (3)

cityCD - city gasoline consumption (gallons/100 miles) in charge depleting mode (4)

cityE - city electricity consumption in kw-hrs/100 miles

cityUF - EPA city utility factor (share of electricity) for PHEV

co2 - tailpipe CO_2 in grams/mile for fuelType1 (5)

co2A - tailpipe CO_2 in grams/mile for fuelType2 (5)

co2TailpipeAGpm - tailpipe CO_2 in grams/mile for fuelType2 (5)

co2TailpipeGpm- tailpipe CO_2 in grams/mile for fuelType1 (5)

comb08 - combined MPG for fuelType1 (2), (11)

comb08U - unrounded combined MPG for fuelType1 (2), (3)

combA08 - combined MPG for fuelType2 (2)

combA08U - unrounded combined MPG for fuelType2 (2), (3)

combE - combined electricity consumption in kw-hrs/100 miles

combinedCD - combined gasoline consumption (gallons/100 miles) in charge depleting mode (4)

combinedUF - EPA combined utility factor (share of electricity) for PHEV

cylinders - engine cylinders

displ - engine displacement in liters

drive - drive axle type

emissionsList

engId - EPA model type index

eng_dscr - engine descriptor; see
http://www.fueleconomy.gov/feg/findacarhelp.shtml#engine

evMotor - electric motor (kw-hrs)

feScore - EPA Fuel Economy Score (-1 = Not available)

fuelCost08 - annual fuel cost for fuelType1 ($) (7)

fuelCostA08 - annual fuel cost for fuelType2 ($) (7)

fuelType - fuel type with fuelType1 and fuelType2 (if applicable)

fuelType1 - fuel type 1. For single fuel vehicles, this will be the only fuel. For dual fuel vehicles, this will be the conventional fuel.

fuelType2 - fuel type 2. For dual fuel vehicles, this will be the alternative fuel (e.g. E85, Electricity, CNG, LPG). For single fuel vehicles, this field is not used

ghgScore - EPA GHG score (-1 = Not available)

ghgScoreA - EPA GHG score for dual fuel vehicle running on the alternative fuel (-1 = Not available)

guzzler- if G or T, this vehicle is subject to the gas guzzler tax

highway08 - highway MPG for fuelType1 (2), (11)

highway08U - unrounded highway MPG for fuelType1 (2), (3)

highwayA08 - highway MPG for fuelType2 (2)

highwayA08U - unrounded highway MPG for fuelType2 (2),(3)

highwayCD - highway gasoline consumption (gallons/100miles) in charge depleting mode (4)

highwayE - highway electricity consumption in kw-hrs/100 miles

highwayUF - EPA highway utility factor (share of electricity) for PHEV

hlv - hatchback luggage volume (cubic feet) (8)

hpv - hatchback passenger volume (cubic feet) (8)

id - vehicle record id

lv2 - 2 door luggage volume (cubic feet) (8)

lv4 - 4 door luggage volume (cubic feet) (8)

make - manufacturer (division)

mfrCode - 3-character manufacturer code

model - model name (carline)

mpgData - has My MPG data; see yourMpgVehicle and yourMpgDriverVehicle

phevBlended - if true, this vehicle operates on a blend of gasoline and electricity in charge depleting mode

pv2 - 2-door passenger volume (cubic feet) (8)

pv4 - 4-door passenger volume (cubic feet) (8)

rangeA - EPA range for fuelType2

rangeCityA - EPA city range for fuelType2

rangeHwyA - EPA highway range for fuelType2

trans_dscr - transmission descriptor; see http://www.fueleconomy.gov/feg/findacarhelp.shtml#trany

trany - transmission

UCity - unadjusted city MPG for fuelType1; see the description of the EPA test procedures

UCityA - unadjusted city MPG for fuelType2; see the description of the EPA test procedures

UHighway - unadjusted highway MPG for fuelType1; see the description of the EPA test procedures

UHighwayA - unadjusted highway MPG for fuelType2; see the description of the EPA test procedures

VClass - EPA vehicle size class

year - model year

youSaveSpend - you save/spend over 5 years compared to an average car ($). Savings are positive; a greater amount spent yields a negative number. For dual fuel vehicles, this is the cost savings for gasoline

sCharger - if S, this vehicle is supercharged

tCharger - if T, this vehicle is turbocharged

c240Dscr - electric vehicle charger description

charge240b - time to charge an electric vehicle in hours at 240 V using the alternate charger

c240bDscr - electric vehicle alternate charger description

createdOn - date the vehicle record was created (ISO 8601 format)

modifiedOn - date the vehicle record was last modified (ISO 8601 format)

startStop - vehicle has start-stop technology (Y, N, or blank for older vehicles)

phevCity - EPA composite gasoline-electricity city MPGe for plug-in hybrid vehicles

phevHwy - EPA composite gasoline-electricity highway MPGe for plug-in hybrid vehicles

phevComb - EPA composite gasoline-electricity combined city-highway MPGe for plug-in hybrid vehicles

emissions

emissionsList

emissionsInfo

efid - engine family ID

id - vehicle record ID (links emission data to the vehicle record)

salesArea - EPA sales area code

score - EPA 1-10 smog rating for fuelType1

scoreAlt - EPA 1-10 smog rating for fuelType2

smartwayScore - SmartWay Code

standard - Vehicle Emission Standard Code

stdText - Vehicle Emission Standard

fuel prices

fuelPrices

midgrade - $ per gallon of midgrade gasoline(9)

premium - $ per gallon of premium gasoline(9)

regular - $ per gallon of regular gasoline(9)

cng - $ per gallon of gasoline equivalent (GGE) of compressed natural gas(10)

diesel - $ per gallon of diesel(9)

e85 - $ per gallon of E85(10)

electric - $ per kw-hr of electricity(10)

lpg - $ per gallon of propane(10)

yourMpgVehicle - summary of all My MPG data for this vehicle

avgMpg - harmonic mean of average MPG shared by fueleconomy.gov users

cityPercent - average % city miles

highwayPercent - average % highway miles

maxMpg - maximum user average MPG

minMpg - minimum user average MPG

recordCount - number of records for this vehicle

vehicleId - vehicle record id (links My MPG data to the vehicle record)

yourMpgDriverVehicle - summary of driver data reported for this vehicle

cityPercent - user average % city miles

highwayPercent - user average % highway miles

lastDate - date records were last updated (yyyy-mm-dd)

mpg - average MPG

state - state of residence

vehicleId - vehicle record ID (links My MPG data to the vehicle record)