MEMORANDUM BohnDA-gram

To: Students in Big Data Analytics

Subject: Importing and filtering data, using EPA mileage as an example

From: Prof. Roger Bohn

Date: April 17, 2016

The EPA data on auto mileage was provided in a spreadsheet, that you can convert into a CSV file. The original file, and the converted one from this memo, are both accessible at https://https:// irgn452.wordpress.com/data-sets/ irgn452.wordpress.com/data-sets/

You know how to clean up data like this using Excel and using Rattle. This memo shows how to do it using raw R.

The cleaned data, with some variables that I added, is now in a file called *EPA mileage 1984-16 cleanRB.csv* You are welcome to use it.

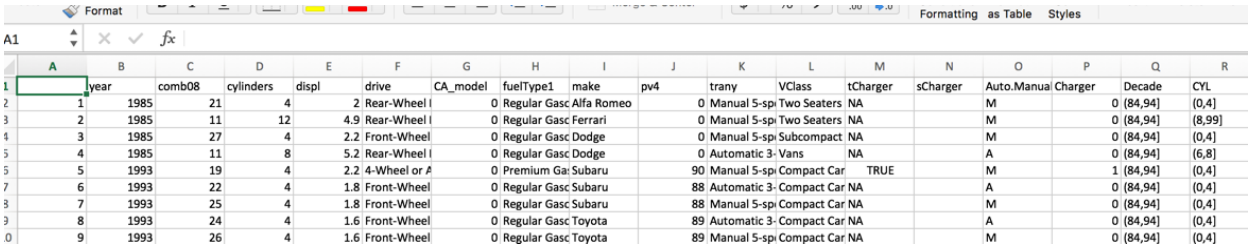
The main changes I made were:

Cutting out many columns

Adding some categorical variables, such as a year category and a cylinders category. Also a "Charger" variable which has 3 levels and includes both supercharger and turbocharger.

Removing all the electric cars, which were messing up some of the results. Only 3 fuel types remain: Diesel, Premium, and Regular (gasoline).

After saving the file as .CSV, I read it into Excel to check it. This is what it looks like.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1		year	comb08	cylinders	displ	drive	CA_model	fuelType1	make	pv4	trany	VClass	tCharger	sCharger	Auto.Manual	Charger	Decade	CYL
2	1	1985	21	4	2	Rear-Wheel	0	Regular Gasc	Alfa Romeo	0	Manual 5-sp	Two Seaters	NA		M	0	(84,94]	(0,4]
3	2	1985	11	12	4.9	Rear-Wheel	0	Regular Gasc	Ferrari	0	Manual 5-sp	Two Seaters	NA		M	0	(84,94]	(8,99]
4	3	1985	27	4	2.2	Front-Wheel	0	Regular Gasc	Dodge	0	Manual 5-sp	Subcompact	NA		M	0	(84,94]	(0,4]
5	4	1985	11	8	5.2	Rear-Wheel	0	Regular Gasc	Dodge	0	Automatic 3-	Vans	NA		A	0	(84,94]	(6,8]
6	5	1993	19	4	2.2	4-Wheel or A	0	Premium Ga	Subaru	90	Manual 5-sp	Compact Car	TRUE		M	1	(84,94]	(0,4]
7	6	1993	22	4	1.8	Front-Wheel	0	Regular Gasc	Subaru	88	Automatic 3-	Compact Car	NA		A	0	(84,94]	(0,4]
8	7	1993	25	4	1.8	Front-Wheel	0	Regular Gasc	Subaru	88	Manual 5-sp	Compact Car	NA		M	0	(84,94]	(0,4]
9	8	1993	24	4	1.6	Front-Wheel	0	Regular Gasc	Toyota	89	Automatic 3-	Compact Car	NA		A	0	(84,94]	(0,4]
10	9	1993	26	4	1.6	Front-Wheel	0	Regular Gasc	Toyota	89	Manual 5-sp	Compact Car	NA		M	0	(84,94]	(0,4]

R code used to edit the EPA file.

I have added row numbers to make it easier to read. Obviously, undo them before pasting into R. (To enable that, I will provide this file as a Microsoft Word document as well as a PDF.)

1. setwd("~/Documents/REB docs '12/Teaching/Big data IRGN 452/R data dir/EPA fuel efficiency 2015")
2. EPA.DATA.vehicles.86.to.15.shrunk <- read.csv("~/Documents/REB docs '12/Teaching/Big data IRGN 452/R data dir/EPA fuel efficiency 2015/EPA DATA vehicles 86 to 15 shrunk.csv") **#done automatically using the Tools/Import Dataset menu in Rstudio**.
3. View(EPA.DATA.vehicles.86.to.15.shrunk)
4. Df <- EPA.DATA.vehicles.86.to.15.shrunk str(Df)
5. table(Df$fuelType1) **#simple one-way table of occurrences**
6.
7.
8. OK.fuel <- Df$fuelType1 == "Diesel"
9. sum(OK.fuel) **#counts number of occurrences, to check that OK.fuel is what I want.**
10. OK.fuelP <- Df$fuelType1 == "Premium Gasoline"
11. sum(OK.fuelP)
12. OK.fuelR <- Df$fuelType1 == "Regular Gasoline"
13. sum(OK.fuelR)
14. OK.fuel <- OK.fuel | OK.fuelP |OK.fuelR **#combine all 3 fuel types, and only those fuel types**
15. sum(OK.fuel)
16. dim (Df2 <- Df[OK.fuel,]) **# This line is poor coding. It creates Df2, and reads its size, in the same line.**
17. summary(Df2$year)
18. Breaks <- c(1994, 2004, 2014, 2017)
19. Yr.cat <- cut(Df2$year, Breaks)
20. head(Yr.cat)
21. Yr.cat <- cut(Df2$year-1900, Breaks)
22. head(Yr.cat)
23. Breaks <- c(1984,Breaks)
24. Yr.cat <- cut((Df2$year-1900), Breaks-1900)
25. Df2$Decade <- Yr.cat
26. Df3 <-Df2 **# Now let's get rid of some columns**
27. Df3 <- subset(Df3, select = -id) **#This is a way to get rid of columns by name, not number. Note the minus sign!**
28. Df3 <- subset(DF3, select = c(-city08,-co2TailpipeGpm, -fuelCost08))
29. Df3 <- subset(Df3, select = c(-city08,-co2TailpipeGpm, -fuelCost08))
30. Df3 <- subset(Df3, select = c(-highway08, -id.1, -year.1, -youSaveSpend))
31.
32. Df3 <- subset(Df3, select = c(-engId, -eng_dscr, -model))
33. sum(Df3$Auto.Manual == "") # what is mysterious third transmission? Only 2 cars have it.

34. CYL <- cut(Df3$cylinders, c(0,4,6,8,99)) **#categorical variable for number of cylinders: 4, 5+6, 7+8, and 9 or higher**
35. Df3$CYL <- CYL
36. write.csv(Df3,file="EPA clean") **#writes results to a file, in your working directory.**
37. summary(Df3) **#this is always a good idea when loading any data.**
38. str(Df3) **#so is this. It's more concise, and shows actual values better.**
39. head(Df3) **# this shows the first 6 rows**