Homework 3

rgk359 Rohit Kamat

This homework is due on Feb. 7, 2017 at 7:00pm. Please submit as a PDF file on Canvas.

In this homework, you are asked to evaluate two data sets and determine if they are tidy data sets. *We are referring to a very specific definition of "tidy", so if this term is unfamiliar to you, please review the lecture materials*.

Problem 1: (2 pts) The dataset `mdeaths` built into R lists monthly deaths from lung diseases in the UK in 1974-1979. You can run `?mdeaths` to learn more about this data set.

```
mdeaths
```

```
##       Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
## 1974 2134 1863 1877 1877 1492 1249 1280 1131 1209 1492 1621 1846
## 1975 2103 2137 2153 1833 1403 1288 1186 1133 1053 1347 1545 2066
## 1976 2020 2750 2283 1479 1189 1160 1113  970  999 1208 1467 2059
## 1977 2240 1634 1722 1801 1246 1162 1087 1013  959 1179 1229 1655
## 1978 2019 2284 1942 1423 1340 1187 1098 1004  970 1140 1110 1812
## 1979 2263 1820 1846 1531 1215 1075 1056  975  940 1081 1294 1341
```

Using the formal definition of tidy data that we learned in lecture, is this data set tidy? Explain why or why not.

This data set is not considered tidy data. The variable year is not on the column side. There should be three columns of year, month, and count.

The data set `CO2` built into R contains CO2 uptake of plants from Quebec and Mississipi measured at different levels of ambient CO2 concentrations. You can run `?CO2` to learn more about this data set.

```
head(CO2)
```

```
##   Plant   Type  Treatment conc uptake
## 1   Qn1 Quebec nonchilled   95   16.0
## 2   Qn1 Quebec nonchilled  175   30.4
## 3   Qn1 Quebec nonchilled  250   34.8
## 4   Qn1 Quebec nonchilled  350   37.2
## 5   Qn1 Quebec nonchilled  500   35.3
## 6   Qn1 Quebec nonchilled  675   39.2
```

Using the formal definition of tidy data that we learned in lecture, is this data set tidy? Explain why or why not.

Yes this data set is considered tidy data. The variable of the data set form the columns and each observation forms the row.

Problem 2: (2 pts) The `Cars93` dataset from the "MASS" package contains information about passenger car models from 1993. You should be familiar with this data set from Homework 2.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
head(Cars93)
```

```
##   Manufacturer   Model    Type Min.Price Price Max.Price MPG.city
## 1        Acura Integra   Small      12.9  15.9      18.8       25
## 2        Acura  Legend Midsize      29.2  33.9      38.7       18
## 3         Audi      90 Compact      25.9  29.1      32.3       20
## 4         Audi     100 Midsize      30.8  37.7      44.6       19
## 5          BMW    535i Midsize      23.7  30.0      36.2       22
## 6        Buick Century Midsize      14.2  15.7      17.3       22
##   MPG.highway            AirBags DriveTrain Cylinders EngineSize
## 1          31               None      Front         4        1.8
## 2          25 Driver & Passenger      Front         6        3.2
## 3          26        Driver only      Front         6        2.8
## 4          26 Driver & Passenger      Front         6        2.8
## 5          30        Driver only       Rear         4        3.5
## 6          31        Driver only      Front         4        2.2
##   Horsepower  RPM Rev.per.mile Man.trans.avail Fuel.tank.capacity
## 1        140 6300         2890             Yes               13.2
## 2        200 5500         2335             Yes               18.0
## 3        172 5500         2280             Yes               16.9
## 4        172 5500         2535             Yes               21.1
## 5        208 5700         2545             Yes               21.1
## 6        110 5200         2565              No               16.4
##   Passengers Length Wheelbase Width Turn.circle Rear.seat.room
## 1          5    177       102    68          37           26.5
## 2          5    195       115    71          38           30.0
## 3          5    180       102    67          37           28.0
## 4          6    193       106    70          37           31.0
## 5          4    186       109    69          39           27.0
## 6          6    189       105    69          41           28.0
##   Luggage.room Weight  Origin          Make
## 1           11   2705 non-USA Acura Integra
## 2           15   3560 non-USA  Acura Legend
## 3           14   3375 non-USA       Audi 90
## 4           17   3405 non-USA      Audi 100
## 5           13   3640 non-USA      BMW 535i
## 6           16   2880     USA Buick Century
```

Pick a car type and a car origin of your choosing. What is the **median** price of the cars with the type and origin that you chose? State your answer in a sentence.

I am choosing a midsize car that is a non-USA origin.

```
Cars93 %>% filter(Type=="Midsize", Origin=="non-USA") %>% summarize(median.price=
median(Price))
```

```
##   median.price
## 1           29
```

The median price of Midsize cars that are non-USA origin is 29.

Problem 3: (4 pts) Which manufacturer has the largest difference between city MPG and highway MPG for large cars? List at least the top 5 and state your answer in a sentence. **HINT:** Use the functions `max()` and `min()` to determine the difference in MPG.

```
car.group <- Cars93 %>% group_by(Manufacturer) %>% filter(Type=="Large") %>% summarise(d
iff.Cars= max(MPG.highway)-min(MPG.city)) %>% arrange(desc(diff.Cars))
head(car.group)
```

```
## # A tibble: 6 × 2
##   Manufacturer diff.Cars
##         <fctr>     <int>
## 1        Buick        12
## 2     Cadillac         9
## 3    Chevrolet         9
## 4   Oldsmobile         9
## 5      Pontiac         9
## 6     Chrylser         8
```

The top 5 manufacturers that have for the largest difference between city MPG and highway MPG for large cars are 1.Buick 2.Cadillac 3.Chevrolet 4.Oldsmobile and 5.Pontiac.

Problem 4: (2 pts) Ask a question about the `Cars93` data set. Describe in 1-2 sentences how you would answer this question with an analysis or a graph.

Is there a broad trend for fuel tank capacity based on size of engine? I would do a ggplot with my x-axis as the size of engine and the y-axis as the fuel tank capacity. Using a scatter plot using geom_point() and using a curve, geom_smooth(), I would be able to see if there is a trend between the two variables.