CSI 2300: Intro to Data Science

In-Class Exercise 11: Data wrangling and cleaning

1. Take a look at the script
   `readAudiLecture11.R` that creates the final **AudiA4** data frame.

In the first line to read in the raw data,

```
AudiA4Raw <- scan("dat/rawCars.txt", what = "a", sep= "\n")
```

what is does `sep="\n"` do? What would happen if this was omitted?

sep="/n" is the separator for the scan function. It is used to separate the data in the file. If it is omitted, the data will be read as a single string.

2. See lines 56 to 62 in `readAudiLecture11.R` to explain how the variable "distance to dealership" is being extracted from the raw data. One hint on understanding what is happening in a loop is to set `k <- 1`, and then run the lines inside of the loop.

```
ind <- grep('80305', work) #get the index of the zipcode
if(length(ind) != 0){ #make sure the car has a zipcode
  temp <- scan(text = work[ind], what= "a", quiet = TRUE) #takes the line with the z
  ind2 <- grep("mi.", temp, fixed = TRUE) #get the index of the miles
  distance[k] <- temp[ind2-1] #stores word before mi. in our distance vector for the
```

3. **grep** refers to a UNIX function designed to do all kinds of matching of strings and uses some characters that have special meaning (known as a grep *meta-character*). In particular the `$` by default in `grep` is used to match the end of a line. Find the place in the `readAudiLecture11.R` script where the asking price is extracted, and explain how the dollar amount is matched as a dollar sign and not as a special `grep` meta-charcter.

It had a gsub with square brackets to specify the dollar sign is not a meta syntax.

4. Load the `AudiA4.rda` data, and plot the asking price (y axis) against mileage (x axis), and color the points by the model year. Is there a straight line relationship between price and mileage? Do you think adjusting for both year and mileage will make much difference in predicting the asking price?

To create a set of colors based on a variable here is one way

```
library(fields)
load('dat/AudiA4.rda')
# this will map the years to 17 colors on the rainbow color range.
yearColors <- color.scale(AudiA4$year, col = rainbow(17))

plot(AudiA4$mileage, AudiA4$price, col = yearColors, type='p', pch = 19,
     xlab= "Mileage", ylab = "Price (USD)")
```

Now in your plot commands use `col = yearColors` in the plot function to get symbols coded by these years. Seventeen colors may be too many, so you can experiment with the number of colors and the colors.

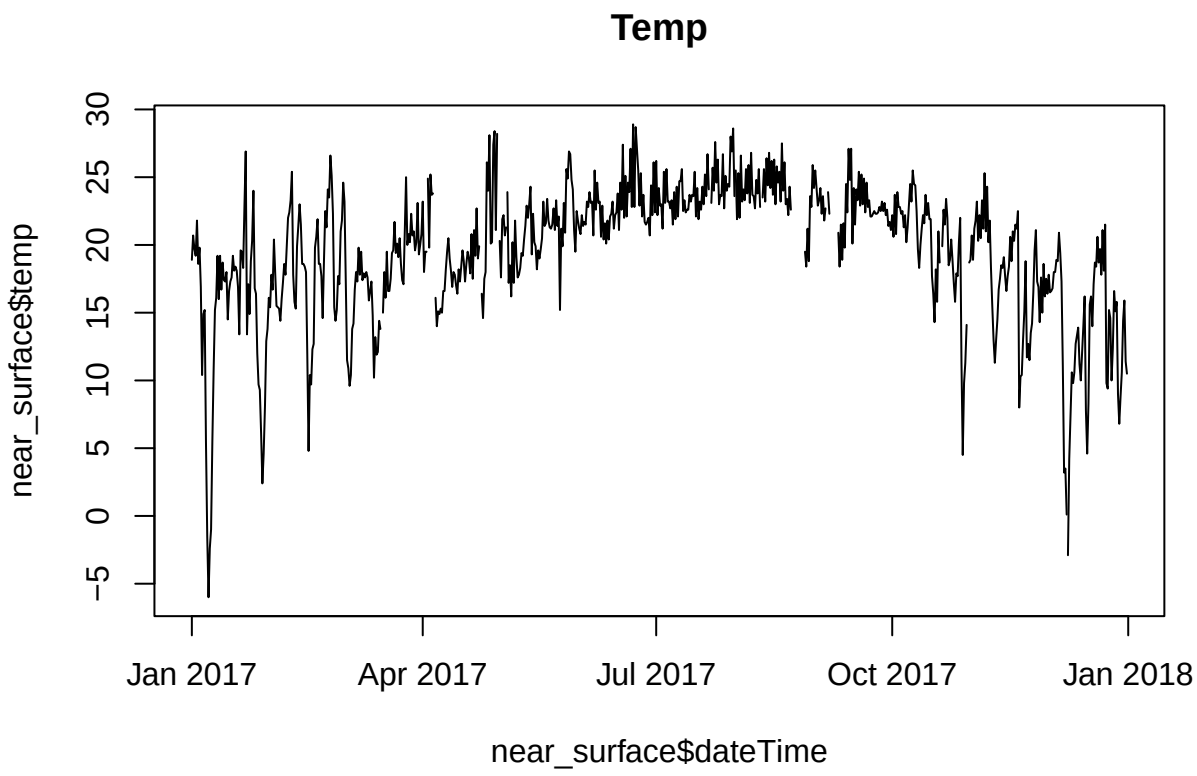
5. For your plot in 4. are there any cars you would consider to be outliers? Any potential bargains?

6. Load the Corpus Christi radiosonde record, `CorpusCristi.rda`. Plot the **temp**, **windSpeed** and **windDir** variables at the *near surface* pressure level (the one equal to 925 mb) over time. To obtain the observations at 925 mb, one way is to use the logical

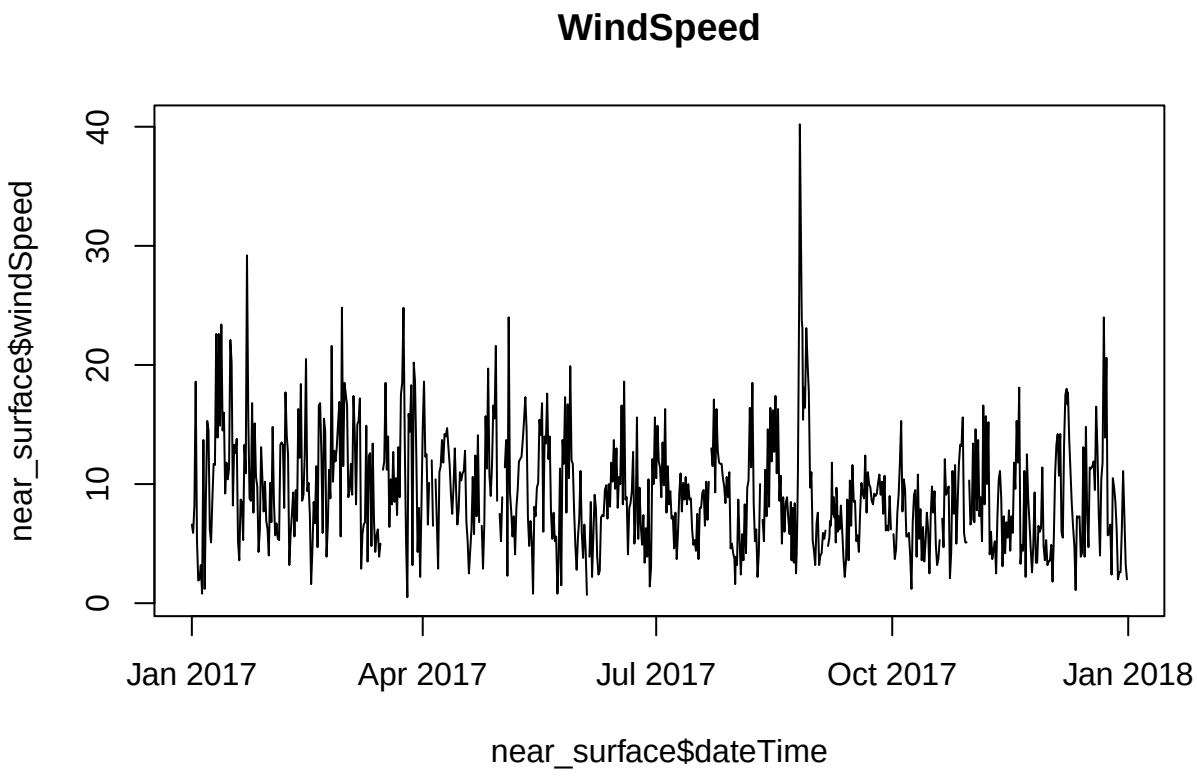
`ind <- CorpusCristi$pressure == 925.`

Comment on any unusual values or patterns.

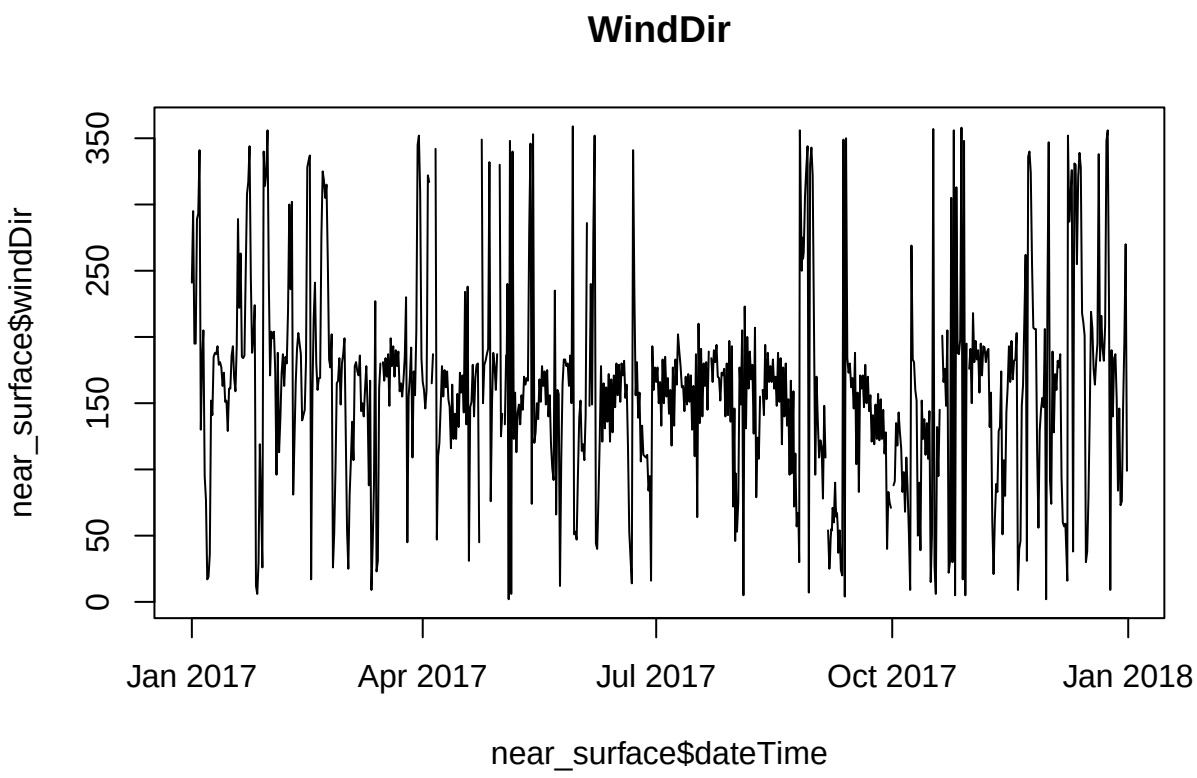
```
load('dat/CorpusCristi.rda')

near_surface <- CorpusCristi[CorpusCristi$pressure == 925,]
plot(near_surface$dateTime, near_surface$temp, type = 'l', main='Temp')
```




```
plot(near_surface$dateTime, near_surface$windSpeed, type = 'l', main='WindSpeed')
```




```
plot(near_surface$dateTime, near_surface$windDir, type = 'l', main='WindDir')
```



In windspeed plot, in between Jul 2017 and Oct 2017, there is a huge spike. In the temp plot, there is a huge drop in temp to -5 degrees that isn't seen across the other years. In the wind direction plot, there is no huge outliers that is visible.

7. Do the missing values in temperature appear to be random across pressure levels, or do they follow a pattern?

```
table(CorpusCristi$pressure)
#
#    7   10   20   30   50   70  100  150  200  250  300  400  500  700  850  925
#  748  748  748  748  748  748  748  748  748  748  748  748  748  748  748  748
# 1000
#  748
table(CorpusCristi$pressure, is.na(CorpusCristi$temp))
#
#       FALSE TRUE
#   7     108  640
#   10    604  144
#   20    706   42
#   30    715   33
#   50    718   30
```

```
#    70      718    30
#    100     719    29
#    150     723    25
#    200     722    26
#    250     722    26
#    300     724    24
#    400     724    24
#    500     723    25
#    700     724    24
#    850     726    22
#    925     718    30
#    1000    712    36
```

The pressure of 7 is showing up as true 640 times which is the opposite of a pressure of 95 which is the normal surface temperature. Meaning the data is not random and is following a pattern. Also the data may be inaccurate in terms of having a temp value of 7 as the placeholder for missing data.