# Week 8 - Class Worksheet

## Scan: Outliers

Dr. Anil Dolgun

Last updated: 18 June, 2020

# Required Packages

The following packages and the function will be required or may come in handy.

```
library(readr)
library(dplyr)
library(outliers)
library(MVN)

cap <- function(x){
    quantiles <- quantile( x, c(.05, 0.25, 0.75, .95 ) , na.rm = T)
    x[ x < quantiles[2] - 1.5*IQR(x, na.rm = T) ] <- quantiles[1]
    x[ x > quantiles[3] + 1.5*IQR(x, na.rm = T) ] <- quantiles[4]
    x
}
```

# Exercises

## Wilt Data

The following exercises 1-4 will be based on wilt data set which is taken from http://archive.ics.uci.edu/ml/datasets/wilt (http://archive.ics.uci.edu/ml/datasets/wilt) containing 4839 observations and 6 variables. The data set was split by training.csv (../data/training.csv) and testing.csv (../data/testing.csv) data sets, for the purpose of this exercise training and testing sets will be joined together. It is expected to do checks on the type of the data and using the suitable transformations if necessary.

**class** : Diseased trees or all other land cover

**Mean_Green**: Mean green (G) value

**Mean_Red**: Mean red (R) value

**Mean_NIR**: Mean near infrared (NIR) value

**GLCM_pan**: Mean gray level co-occurrence matrix (GLCM) texture index

**SD_pan**: Standard deviation

Here is a quick look of the wilt data:

| class | GLCM_pan | Mean_Green | Mean_Red | Mean_NIR | SD_pan |
|---|---|---|---|---|---|
| w | 120.3628 | 205.5000 | 119.39535 | 416.5814 | 20.67632 |

| class | GLCM_pan | Mean_Green | Mean_Red | Mean_NIR | SD_pan |
|---|---|---|---|---|---|
| w | 124.7396 | 202.8000 | 115.33333 | 354.33331 | 6.70715 |
| w | 134.6920 | 199.2857 | 116.85714 | 477.85712 | 2.49671 |
| w | 127.9463 | 178.3684 | 92.36842 | 278.47371 | 4.97745 |
| w | 135.4315 | 197.0000 | 112.69048 | 532.95241 | 7.60419 |
| w | 118.3480 | 226.1500 | 138.85000 | 608.90000 | 9.07280 |

**1** Join the training.csv (../data/training.csv) and testing.csv (../data/testing.csv) data sets, and rename the combined data frame as `wilt`.

**2** Identify the univariate outliers of `Mean_Green`, `Mean_Red`, `Mean_NIR` and `GLCM_pan` variables from wilt data set using Tukey's method of outlier detection.

**3** Use z-score approach via `scores()` function to extract outliers of `Mean_Green`, `Mean_Red`, `Mean_NIR` and `GLCM_pan` variables. Find the location of the outliers. How many outliers are there per variable? Use `summary()` function to find out about the variables.

**4** Replace the outliers of `Mean_Green`, `Mean_Red`, `Mean_NIR` and `GLCM_pan` variables using `capping method`. You can use `sapply()` function to apply capping across the variables or you can do it individually. Use `summary()` function to see min and max values of the variables.

# Ozone Data

The following exercises 5-8 will be based on ozone.csv (../data/ozone.csv) data set which is taken from http://rstatistics.net/wp-content/uploads/2015/09/ozone.csv (http://rstatistics.net/wp-content/uploads/2015/09/ozone.csv) containing 366 observations and 13 variables. Variables are self explanatory however it is expected to do checks on the type of the data and using the suitable transformations if necessary.

Here is a quick look of the ozone data:

| Month | Day_of_month | Day_of_week | ozone_reading | pressure_height | Wind_speed | Humidity | Temperature_Sandburg |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 4 | 3.01 | 5480 | 8 | 20 | NA |
| 1 | 2 | 5 | 3.20 | 5660 | 6 | NA | 38 |
| 1 | 3 | 6 | 2.70 | 5710 | 4 | 28 | 40 |
| 1 | 4 | 7 | 5.18 | 5700 | 3 | 37 | 45 |
| 1 | 5 | 1 | 5.34 | 5760 | 3 | 51 | 54 |
| 1 | 6 | 2 | 5.77 | 5720 | 4 | 69 | 35 |

| Temperature_ElMonte | Inversion_base_height | Pressure_gradient | Inversion_temperature | Visibility |
|---|---|---|---|---|
| NA | 5000 | -15 | 30.56 | 200 |
| NA | NA | -14 | NA | 300 |
| NA | 2693 | -25 | 47.66 | 250 |
| NA | 590 | -24 | 55.04 | 100 |
| 45.32 | 1450 | 25 | 57.02 | 60 |
| 49.64 | 1568 | 15 | 53.78 | 60 |

**5** Investigate `ozone_reading` variable across `Month` and `Wind_speed` using univariate and bivariate box plots and scatter plots. Before taking the next step, subset the ozone data set with these variables and remove `NA` values, make appropriate adjustments.

**6** Use `mvn()` function to remove the outliers, use 2 different ways while doing this. First way will be manually removing the outliers when you find them. Second way will be simply using an argument inside the `mvn()` function.

**7** **Data Challenge**: Create a subset of ozone with ozone_reading and Temperature_Sandburg variable. Use one of the `cut()`, `case_when()` or `ifelse()` functions in `mutate()` to create a new temperature variable. You can get creative and do it in a different way. The new temperature variable is going to be categorical and grouped with 10 degrees difference. Investigate the outliers using Tukey's method of outlier detection. The subset should look like this:

| ozone_reading | Temperature_Sandburg | temp |
|---|---|---|
| 3.01 | NA | NA |
| 3.20 | 38 | (30,40] |
| 2.70 | 40 | (30,40] |
| 5.18 | 45 | (40,50] |
| 5.34 | 54 | (50,60] |
| 5.77 | 35 | (30,40] |

**8** **Bonus Exercise**: Use capping method to replace outliers in the ozone data set that you subsetted in question 5. Compare the methods you used in question 6. Which one would you pick and why? Share your own approach with your code on the discussion board. Best solution(s) will be immortalised as example solutions in this worksheet.

# Finished?

If you have finished the above tasks, work through the weekly list of tasks posted on the Canvas announcement page.

**Return to Course Website (../index.html)**