Michaela Bayerlova
michaela.bayerlova@wsu.edu
011594781
STATS 419
Instructor: Monte J. Shaffer
9/4/2020

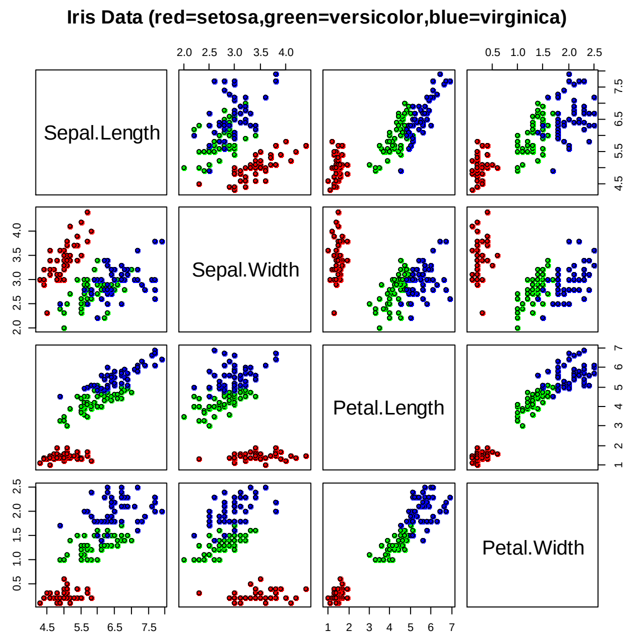
Assignment "Datasets"

1. Create the "rotate matrix" functions described in lectures from the sample matrix.

```
myMatrix = matrix ( c (
                        1, 0, 2,
                        0, 3, 0,
                        4, 0, 5
                               ), nrow=3, byrow=T);

transposeMatrix = function(mat)
{
  t(mat);
          }

             #rotateMatrix90(mat)
             #rotateMatrix180(mat)
             #rotateMatrix270(mat)
             # 3x3 matrix ... ## matrix multiplication

## ANSWER
# clockwise
rotateMatrix90 = function(mat)
   {
   t(mat[nrow(mat):1,,drop=FALSE]);
   }

rotateMatrix180 = function(mat)
   {
   rotateMatrix90(rotateMatrix90(mat));
   }

rotateMatrix270 = function(mat)
   {
   rotateMatrix90(rotateMatrix90(rotateMatrix90(mat)));
   }

rotateMatrix90(myMatrix);
rotateMatrix180(myMatrix);
rotateMatrix270(myMatrix);


# counter clockwise
rotateMatrix90_cc = function(mat)
   {
   apply(t(mat), 2, rev);
   }

rotateMatrix180_cc = function(mat)
   {
   rotateMatrix90_cc(rotateMatrix90_cc(mat));
   }

rotateMatrix270_cc = function(mat)
   {
   rotateMatrix90_cc(rotateMatrix90_cc(rotateMatrix90_cc(mat)));
   }

rotateMatrix90_cc(myMatrix);
rotateMatrix180_cc(myMatrix);
rotateMatrix270_cc(myMatrix);
```

2. Recreate the graphic for the IRIS Data Set using R. Same titles, same scales, same colors. See: https://en.wikipedia.org/wiki/Iris_flower_data_set#/media/File:Iris_dataset_scatterplot.svg

```
IRIS_Data = iris
class(iris)
# Scatterplot
pairs(IRIS_Data[1:4], main = "Iris Data
(red=setosa,green=versicolor,blue=virginica)",
    pch = 21, bg = c("red", "green", "blue")[unclass(IRIS_Data$Species)])
```



Michaela Bayerlova
michaela.bayerlova@wsu.edu
011594781
STATS 419
Instructor: Monte J. Shaffer
9/4/2020

3. Right 2-3 sentences concisely defining the IRIS Data Set. Maybe search KAGGLE for a nice template. Be certain the final writeup are your own sentences (make certain you modify what you find, make it your own, but also cite where you got your ideas from). NOTE: Watch the video, Figure 8 has a +5 EASTER EGG.

The Iris flower data set is a multivariate data set.
The data set contains four measurements (sepals length and width, petals length and width) for 150 records of flowers. Each is represented in the three species of iris: Iris setosa, Iris versicolor and Iris virginica. The Iris setosa is from a wide range across the Arctic sea. The Iris versicolor is found in North America, like Eastern United States and Eastern Canada. The Iris virginica is native to eastern North America.

4. Import "personality-raw.txt" into R. Remove the V00 column. Create two new columns from the current column "date_test": year and week. Stack Overflow may help: https://stackoverflow.com/questions/22439540/how-to-get-week-numbers-from-dates ... Sort the new data frame by YEAR, WEEK so the newest tests are first ... The newest tests (e.g., 2020 or 2019) are at the top of the data frame. Then remove duplicates using the unique function based on the column "md5_email". Save the data frame in the same "pipe-delimited format" (| is a pipe) with the headers. You will keep the new data frame as "personality-clean.txt" for future work (you will not upload it at this time). In the homework, for this tasks, report how many records your raw dataset had and how many records your clean dataset has.

The raw data set has dimensions of 838 63, which means there are 838 records.
Whereas the clean data set has dimensions of 678 63, and therefore only 678 records.

5. Write functions for doSummary and sampleVariance and doMode ... test these functions in your homework on the "monte.shaffer@gmail.com" record from the clean dataset. Report your findings. For this "monte.shaffer@gmail.com" record, also create z-scores. Plot(x,y) where x is the raw scores for "monte.shaffer@gmail.com" and y is the z-scores from those raw scores. Include the plot in your assignment, and write 2 sentences describing what pattern you are seeing and why this pattern is present.

```
doSummary = function(x)
    {
      # length
      l = length(x);
      # number of NAs
      n_NAs = colSums(is.na(x));
      # mean
      m = mean(x, na.rm = TRUE);
      # median
      med = median(x, na.rim = TRUE);
      # mode #custom function
      mode_c = doMode(x);
```

```
       # variance #custom function
       variance = doSampleVariance(x, "naive");
       # sd... built in fct but compare it to custom fct...
       sd_a = sd(x);
       sd_b = sd(x)*(sqrt((length(x)-1)/length(x)));

       }


doSampleVariance = function(x, method)
   {
   if( method=="naive")
      {
        sum((x-mean(x))^2)/(length(x)-1);
        }
        else
           {
              # two-pass algorithm
              n = sum1 = sum2 = 0;
              for (i in 1:x)
                 {
                 n += 1;
                   sum1 += x;
                   }
              m = sum1/n

              for (i in 1:x)
                 {
                 sum2 += (x-m)*(x-m);
                   }
              variance = sum2/(n-1);
              return variance;
              }

   }



doMode = function(x)
   {
   result = c();
      # freq... #high frequencies
        # ties... store all of ties
       score = 0;
       # find highest score
       for (i in 1:x)
          {
          if (score < x[i])
             {
               score = x[i];
               }
          }
       # go through x again to get all ties and store them
       for (i in 1:x)
          {
          if (score == x[i])
             {
               result = c(result,x[i]);
               }
          }

   result;
   }
```

6. Compare Will Smith and Denzel Washington. [See 03_n greater 1-v2.txt for the necessary functions and will-vs-denzel.txt for some sample code and in DROPBOX: __student_access__\unit_01_exploratory_data_analysis\week_02\imdb-example] You will have to create a new variable $millions.2000 that converts each movie's $millions based on the $year of the movie, so all dollars are in the same time frame. You will need inflation data from about 1980-2020 to make this work.

7. Build side-by-side box plots on several of the variables (including #6) to compare the two movie stars. After each box plot, write 2+ sentence describing what you are seeing, and what conclusions you can logically make. You will need to review what the box plot is showing with the box portion, the divider in the box, and the whiskers.

The box plots are showing me, that Will Smith has a higher variety, like a bigger spread of earnings for the movies he shot than Denzel Washington. The mean is a little higher than 50 Millions, whereas the mean of Denzel Washington is on about 50 Million. Will Smith also played in more movies than Denzel Washington. Based on that data you could say that Will Smith is a better actor but in reality this is a very subjective question. Every person can answer it differently.