Math 321 – Spring 2019 – R Activity 1

We are going to look at oil production data from the state of Montana from 1981 to 2018.

Data source: https://www.eia.gov/opendata/qb.php?sdid=PET.MCRFPMT1.M

The data is saved in a text file named: `MT_oil_production.txt`. The columns of data are volume of oil produced in thousands of barrels, numerical month (1=Jan,2=Feb,...,12=Dec), and numerical year. Create a folder somewhere on the computer where you will consistently work from for this course. Download the MT oil production data file and save it there.

Throughout this exercise and future exercises, R commands will usually be given in `blue typewriter style text` and after a "`>`" symbol to represent the command prompt in R Studio. Generally, you will always type in a command and hit the `Enter` or `Return` key.

Getting started. The first step is to load the data into R. Make sure you are working on a computer with both R and RStudio installed. Open Rstudio.

Setting your working directory. We need to set our working directory so that R will know where to open files from and save output to. In RStudio you can do this from the main menu with:

Session → Set Working Directory → Choose Directory...

Otherwise we'll need to set it manually.

Let's say your directory is `C:\Users\myname\files\statsRwork`, then you will need to type:

```
> setwd("C:/Users/myname/files/statsRwork")
```

Notice that the slashes have reversed direction!!!

Check that your working directory is where you want: `> getwd()`
View all files that are in that directory: `> list.files()`

Load the data. Now that you have your working directory set in RStudio, and the data text file saved to that location.

```
> mtop = read.table("MT_oil_production.txt", header=T)
```

A *table* is a specific kind of data structure in R mad up of rows and columns. Note that the filename is in double quotes. The command "`header=T`" tells R to read the first line as the names of the data stored in the columns. If the text file did not have column names, we would eliminate this command. Now the data table is stored under the name `mtop` (short for "Montana Oil Production").

There are three columns of information: Oil Production, Month, and Year. The data is already sorted in order by date as well.

Accessing a specific part of a data object in R. We have a data table with header names `oilprod`, `month`, and `year`. We want to be able to extract these separately. This is done with the dollar sign "$".

Let's save the actual oil production data as `x`:

```
> x = mtop$oilprod
```

Now type `> x` and hit return, and you will see the list of numerical data. If the list is very large, say with thousands of tens of thousands of data points, this will only show a part of the data.

Now we are ready to analyze the data!!!

Complete the following to analyze this dataset.

1. Plot oil production over time. The data is already sorted temporally. We'll plot it a few different ways.
   `> plot(x)` (This just plots the data we have stored under the name `x`.)

2. Plot oil production by year only. This will have all 12 months of each year clustered in a vertical line.
   `> plot(mtop$year,mtop$oilprod)`

3. Calculate mean, median, minimum, and maximum oil production over the entire period:
   `> mean(x)`
   `> median(x)`
   `> min(x)`
   `> max(x)`

4. Calculate the summary statistics. This will give us all of the above plus the quartiles:
   `> summary(x)`

5. Calculate the sample size:
   `> length(x)`

6. Sort the data in numerical increasing order. We'll use this to manually calculate the quartiles and median in order to make sure we believe the output hat the above commands gave us.
   `> y = sort(x)`

7. You should have gotten $n = 454$, thus we have an even dataset. So the median is the average of the two middle values. $454/2 = 227$ thus we will average the 227^{th} and 228^{th} data values in order to ge the median. In R, square brackets are used to access specifi elements of a list, `y[i]` is the i^{th} element in list `y`.
   `> xtilde = (y[227]+y[228])/2` (sample median is denoted by $\tilde{x}$ "tilde over an x")
   Does this match the output from `> median(x)` ?

8. Let's manually calculate the quartiles. Each half of the dataset has 227 points. This is an odd number, and $227/2 = 113.5$ so we will get the 114^{th} data point for the 1^{st} quartile and the $227 + 114 = 341^{th}$ data point for the 3^{rd} quartile.
   `> y[114]`
   `> y[341]`
   Notice that `y[114]` does not match up with the 1^{st} quartile given to you by the `summary()` command! That is normal. There are at least 9 different ways to calculate quartiles. My preferred method is Method 1 here: https://en.wikipedia.org/wiki/Quartile. Don't worry about learning the different methods, as long as you can do one of them.

9. Calculate the *interquartile range*:
   `> IQR(x)`

10. Calculate the *standard deviation* and *variance*:
    ```
    > sd(x)
    > var(x)
    ```

11. Plot a boxplot:
    ```
    > boxplot(x)
    > boxplot(x,horizontal=TRUE)
    ```
    (This makes it a horizontal boxplot instead of vertical.)

12. Plot some different histograms:
    ```
    > hist(x)
    ```
    (This plots a *frequency* histogram.)
    ```
    > hist(x,freq=FALSE)
    ```
    (This plots a *density* histogram.)
    ```
    > hist(x,breaks=25)
    ```
    (This histogram will have roughly 25 bins.)
    ```
    > hist(x,breaks=c(1000,1500,2000,2500,3000,3500))
    ```
    (This histogram has precise bins [1000, 1500], (1500, 2000], (2000, 2200], (2500, 3000], (3000, 3500].)

13. Try this really nice-looking histogram with 10 bins from 1,000 to 3,200:
    ```
    > hist(x,breaks=seq(1000,3200,l=11),
          freq=TRUE,col="light blue",
          main="Monthly MT Oil Production 1918-2018\n (thousands of barrels)",
          xlab="oil production",ylab="frequency",yaxs="i",xaxs="i")
    ```

Assignment for submission:

SUBMISSION INSTRUCTIONS:
- Submit this assignment via email. DUE DATE: Wednesday, 1/30/2019
- Use your preferred wordprocessing software (e.g. MS Word).
- Be sure to insert any graphics requested. Make sure they are legible!
- Answer each question below.
- Be sure to indicate what your answers refer to, e.g. if I am asking for a mean, don't just give the numerical answer, state that it is the mean and what it is the mean value of.
- Include all R commands and their output entered, e.g. if you used mean(x), copy and paste it into the document that you turn in. For example:

```
> mean(x)
[1] 1980.601
```

Turn in the answers to these questions:

1. Include the descriptive statistics: minimum, maximum, mean, median, 1^{st} and 3^{rd} quartiles, interquartile range, standard deviation, and variance for monthly oil production.

2. Use the last fancy histogram command above to create a histogram for oil production with a color chosen from Rcolor.pdf posted on my website (choose a unique color; be creative!). Also choose a new list of at least 20 bins. You can do this using `seq(min,max,n)` (`n` is the number of bins + 1) or `c(b0,b1,b2,...,bn)` (to get bins $[b_0, b_1], (b_1, b_2], ..., [b_{n-1}, b_n]$. Use the example histogram commands above as a guide. To save your histogram as an image, look for the Export button at the top of the plot window. Note that you can adjust the dimensions of your output image file before saving it.