Homework Week 1 Mark Cappiello

1. Using the material from this chapter and possibly other information that you look up, write a brief definition of these terms in your own words: mean, median, mode, variance, standard deviation, histogram, normal distribution, and Poisson distribution.

Mean – The arithmetic mean. It is the most used measure of central tendency. It is calculated by adding up the numbers in a set of data, or vector, and then dividing by the count of those numbers. It is what we commonly call the average.

Median – The middle value in a set of numbers that are arranged in ascending or descending order. If there is an even number of values, then you would take the average or mean of the two middle values.

Mode – The value that appears most frequently in a set of numbers. It is possible for a set of numbers to have no mode, one mode or multiple modes.

Variance – it is a measurement of how much the values in a data set deviate from the mean. The variance is calculated by taking the difference of each value and the mean, squaring those differences and then taking the average of the squares.

Standard Deviation – is the square root of the variance. It gives us an idea of the dispersion of our data.

Histogram – It is a graphical representation of the distribution of our data set. It is made of rectangular bars which represent the number of data points in a given range of values. The higher the number of data points in a range the longer that bar will be.

Normal Distribution – when looking at a normal distribution on a histogram this will be represented by what we recognize as a typical bell curve. This means that the data will be symmetrical around the mean and have the most frequently occurring values toward the middle of the histogram and the values will gradually become lower as we move away from the middle toward each end.

Poisson Distribution – a discrete frequency distribution which gives the probability of a number of independent events occurring in a fixed time.  
3. Use the data() function to get a list of the data sets that are included with the basic installation of R: just type “data()” at the command line and press enter. Choose a data set from the list that contains at least one numeric variable—for example, the Bio‑ chemical Oxygen Demand (BOD) data set. Use the summary() command to summarize the variables in the data set you selected—for example, summary(BOD). Write a brief description of the mean and median of each numeric variable in the data set. Make sure you define what a “mean” and a “median” are, that is, the technical definition and practical meaning of each of these quantities.

data()  
summary(airquality)

## Ozone Solar.R Wind Temp   
## Min. : 1.00 Min. : 7.0 Min. : 1.700 Min. :56.00   
## 1st Qu.: 18.00 1st Qu.:115.8 1st Qu.: 7.400 1st Qu.:72.00   
## Median : 31.50 Median :205.0 Median : 9.700 Median :79.00   
## Mean : 42.13 Mean :185.9 Mean : 9.958 Mean :77.88   
## 3rd Qu.: 63.25 3rd Qu.:258.8 3rd Qu.:11.500 3rd Qu.:85.00   
## Max. :168.00 Max. :334.0 Max. :20.700 Max. :97.00   
## NA's :37 NA's :7   
## Month Day   
## Min. :5.000 Min. : 1.0   
## 1st Qu.:6.000 1st Qu.: 8.0   
## Median :7.000 Median :16.0   
## Mean :6.993 Mean :15.8   
## 3rd Qu.:8.000 3rd Qu.:23.0   
## Max. :9.000 Max. :31.0   
##

The airquality dataset shows daily air quality measurements in New York, May to September 1973.

Ozone - Mean ozone in parts per bilion from 1pm to 3pm at Roosevelt Island Mean = 42.13 Median = 31.50

The mean ozone level is higher than the median ozone level. This tells us that the distribution of ozone levels are right-skewed, with some higher values pushing the mean upwards. A good Ozone level is in the range of 0-50 parts per billion.

Solar.R - The solar radiation measurements from 8am - noon at Central Park. measured in Langleys Mean = 185.9 Median = 205.0

The mean Solar.R value is higher than the median Solar.R value. This tells us that the distribution of Solar.R levels may be right-skewed, with some higher values pulling the mean upwards. A moderate Solar.R range is 100-250 Langleys

Wind - the average wind speed per day at LaGuardia airport from 7am- 10am Mean = 9.958 Median = 9.700

The mean Wind level is higher than the median Wind level. This tells us that the distribution of Wind values are right-skewed, with some higher values pushing the mean upwards. A moderate wind speed between 3-15 mph helps disperse air pollutants.

Temp - the maximum daily temperature in degrees Farenheit at LaGuardia Airport Mean = 77.88 Median = 79.00

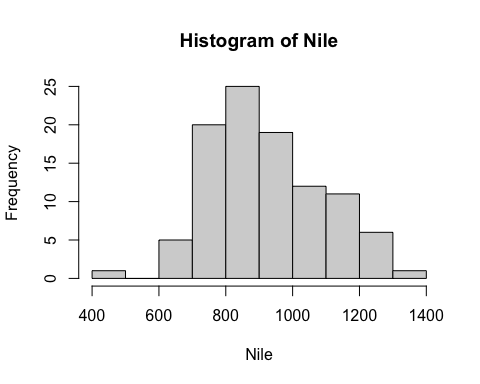
The mean Temperature value is lower than the median Temperature value. This tells us that the distribution of Temp values are left-skewed, with lower values pushing the mean lower.

Month - month the data was collected Mean = 6.993 Median = 7.000

Day - the day of the month that the data was collected Mean = 15.8 Median = 16.0

1. As in the previous exercise, use the data() function to get a list of the data sets that are included with the basic installation of R. Choose a data set that includes just one variable, for example, the LakeHuron data set (levels of Lake Huron in the years 1875 through 1972). Use the hist() command to create a histogram of the variable—for example, hist(LakeHuron). Describe the shape of the histogram in words. Which of the distribution types do you think these data fit most closely (e.g., normal, Poisson). Speculate on why your selected data may fit that distribution. (I considered the following R data sets: Nile, lynx or rivers)

hist(Nile)



The Nile data set shows the flow of the Nile river at Aswan from 1871-1970

The data is unimodal, which means that the histogram has one peak, so the data is clustered around one central point.

The histogram is right-skewed or positively skewed. This means that there are more values that are greater than the median which pushes the mean to the right.

The spread of the data shows that thee is a wide variety of values in the data set.

The fact that there is a gap in the histogram on the left and that the data is right-skewed shows that there are outliers in the dat set.