**R PRACTICAL WEEK 2 – 1712 – REPORT**

Task 1:

The data we have been asked to analyse contains the measurements of annual snow fall in Buffalo, New York, over the period of 109 years, specifically between the years 1910 to 2018. We have a simple dataset consisting of only two columns, one titled ‘Years’ where the years between 1910 to 2018 are listed and another titled ‘Snowfall’ which has the corresponding Snow Fall for each year.

To get a general overview and a better understanding of our data, we inputted the following command into R Studio: “summary(buffalo$snowfall)”. This gave the 5-number summary for snowfall (inches) as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 25 | 67.5 | 83.60 | 86.69 | 104.5 | 199.4 |

At first glance of this data we can see that our dataset is wide spread, having a range of 174.4 inches. This statement can also be supported by the Interquartile Range being 37 inches. Before we begin analysing our data in depth we can conclude that this dataset has a large spread of data.

Task 2:

For Task 2 we were required to find specific summary statistics, as seen below.

Firstly, for (a) we need to find the year when snowfall in Buffalo was at its lowest. Using our 5-figure summary above, we know that the minimum value of snowfall was 25 inches. R-Studio is able to locate the specific row when the snowfall is at 25. In order to find the corresponding year, we input the following R command: “which(buffalo$snowfall == "25")” which tells us that this statistics appears on row 10. To finish, we simply input “buffalo[10,1] to get the information in the 1st column (labelled ‘years’) on the 10th row. Doing so, we get that the year when snowfall was at its lowest was 1919. Suggesting that the amount of snowfall has risen over the last few decades.

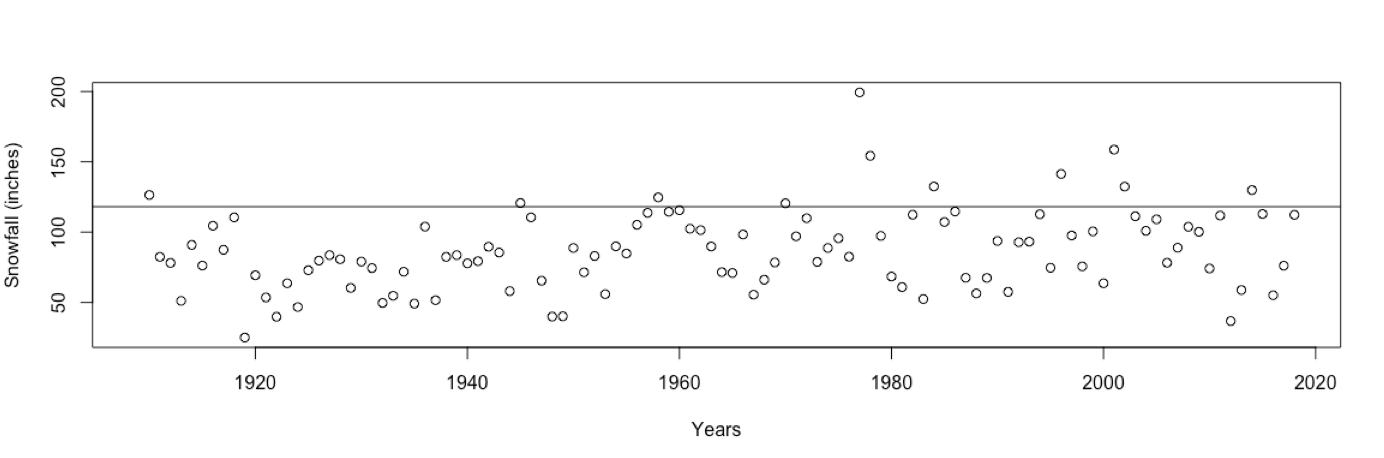
For (b) we need to find the year when snowfall in Bufallo was at its highest. We follow a similar approach as above, using our 5-figure summary to conclude that the highest recorded snowfall was 199.4 inches. To get the specific row this information is located on, we input “which(buffalo$snowfall == "199.4")” which gave us row 68. Then, we input “buffalo[68,1]” and R-Studio tells us that the year when snowfall was at its highest was 1977. This specific statistic is supported by, what American’s call it, ‘The Blizzard of 1977’ where record figures of snowfall were recorded during February 1977.

For (c), the average amount of snowfall per year in inches, we can obtain using our 5-figure summary where the ‘mean’ is calculated at 89.69 inches. We can check this by inputting “mean(buffalo$snowfall)” which indeed gives 86.69174. So the average amount of snowfall per year is 86.691 inches (3.d.p).

We see a change of measurement from (d) to (e) where we change from inches to centimetres. For (d), the standard deviation of the amount of snow in inches, we simply input “sd(buffalo$snowfall)”, giving the value 28.233 (3.d.p). Using the conversion rate of 1 inch = 2.54 centimetres, we input the following “sd((buffalo$snowfall)\*2.54)” giving the value 71.71188. We can therefore conclude that the standard deviation of the amount of snowfall was 71.7cm (3.d.p).

Finally, for (e) we need to calculate the number of years where more than 2 metres of snow fell. 2 metres is equivalent to 78.7402 inches so we will use this measurement when carrying out our calculations. We can follow a similar approach to (a) where we calculate the rows where the snow fall was greater than 78.7402 inches. In order to do this, we input “which(buffalo$snowfall > "78.7402")” which gives us many different rows. Instead of adding these up manually, we can simply use the ‘length’ command. More specifically, we input “length(which(buffalo$snowfall > 78.7402))” which gives us 64. Therefore, we can conclude that on 64 separate years, more than 2 metres of snow fell. This figure tells us that 59% of the years between 1910 to 2018 had over 2 metres of snowfall, implying that perhaps snowfall is on the rise. However, when we look closer into the data we see that every 10 years since 1910, roughly 5 to 6 years had snowfall greater than 2 metres. This suggests that the amount of snowfall above 2 metres has been at a constant rate since 1910.

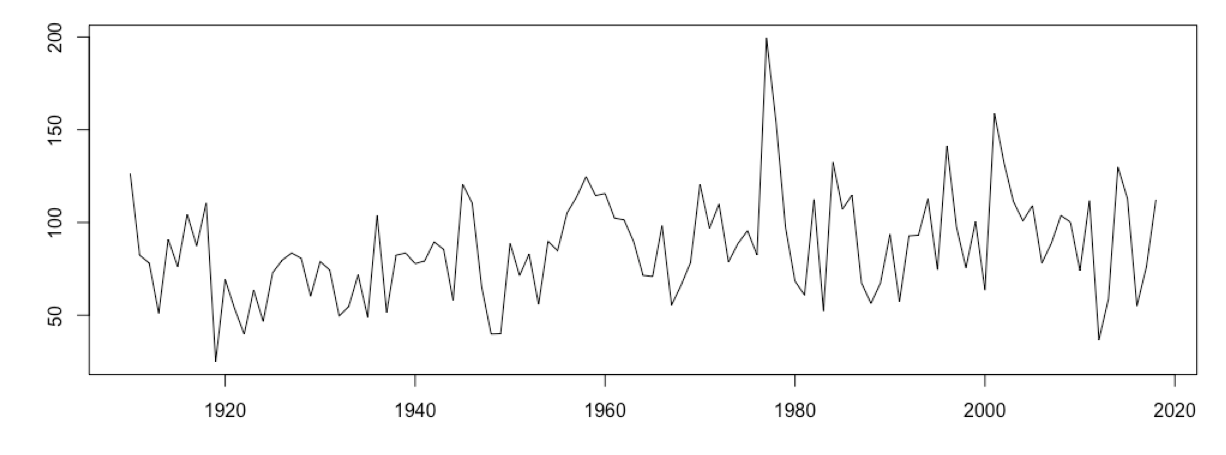
Task 3:

When plotting the amount of snowfall as a function of time, we must label the years on the x-axis and snowfall on the y-axis. When creating a plot, we input the following: “plot(buffalo, xlim=c(1909, 2018), ylab="Snowfall (inches)", xlab="Years")”. In addition, our task was to mark any years where the snowfall was greater than 3 metres. 3 metres is equivalent to 118.11 inches so we plotted the following line: “abline(h=118.11)”. This gave us the following graph:

From this particular graph, we are able to determine which years had a snowfall above 3 metres by simply reading any point that appears above our line.

We can also follow a similar process to Task 2 (e) and input “length(which(buffalo$snowfall > 118.11))” which gives that 11 years had snowfall above 3 metres, more precisely, the years 1910, 1945, 1958, 1970, 1977, 1978, 1984, 1996, 2001, 2002 and 2014.

We are also able to draw a time plot which enables us to gain a better understanding for any sharp changes in the weather pattern; perhaps where the snowfall in Buffalo dramatically increased or decreased over the period of 109 years.

To plot the graph, we input “plot(buffalo$year, buffalo$snowfall, type="l", xlab="Year", ylab="Snowfall"). This input gives the following graph:

Years

Snowfall

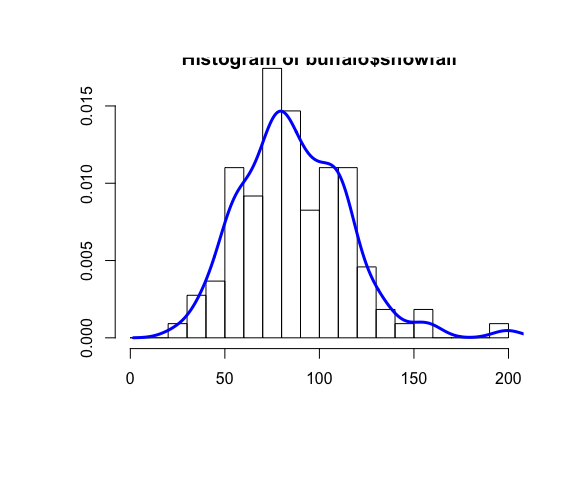
When commenting on our results from the two graphs above, we can say with certainty that the distribution of snowfall in Buffalo does not form any consistent pattern. The irregularity of the lines in the second graph imply that the snowfall is completely random. Despite there being a steady rate of snowfall above 2 metres every decade, there is no consistent pattern of increase or decrease. We may comment on the sharp increase and decrease in snowfall between the years 1976 and 1982 where between 1976 and 1977, snowfall in Buffalo increased by over 100 inches.

Scientists may argue that the fluctuation in snowfall may be as a result of unstable climate change. However, considering the area of Buffalo, New York, is only 52.5 square miles, this small area is not comparable to climate change as a whole across America or even the world.

One factor that American’s feel greatly influences the amount of snow they get is the ‘Lake Effect’. This describes the transfer of heat and moisture from the warm waters of the Great Lakes, next to Buffalo, to the cold air which then forms snow. It may be this environmental factor that contributes heavily towards the unstable formation of snow across Buffalo and one that may explain the random behaviour of the snowfall over the past 109 years.

Task 4:

Alternating the size of histogram buckets using R allows us to observe the data, noticing how changing the width of the bar can give the data one, two or three modes. Initially we used larger width bars and noticed that the wider the histogram bars, the increased likelihood that the histogram would have 1 mode. Using breaks at “10”, we obtain the following graph:



R-Command:

“>hist(buffalo$snowfall, prob=TRUE, breaks=seq(0, 200, by=10))

> lines(density(buffalo$snowfall, adjust=0.8)”

Frequency Density

Snowfall

Lowering the breaks within the histogram, in particular, to “5” then “0.45” gives us a histogram with two and three modes, as seen below:

R-Command:

Two Mode Histogram:

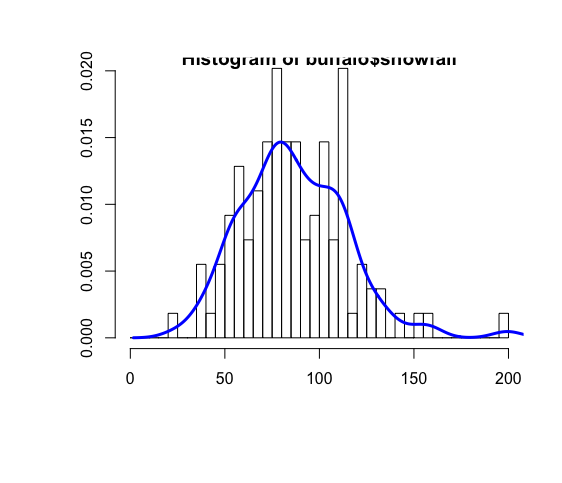
“>hist(buffalo$snowfall, prob=TRUE, breaks=seq(0, 200, by=10))

> lines(density(buffalo$snowfall, adjust=0.8)”

Three Mode Histogram:

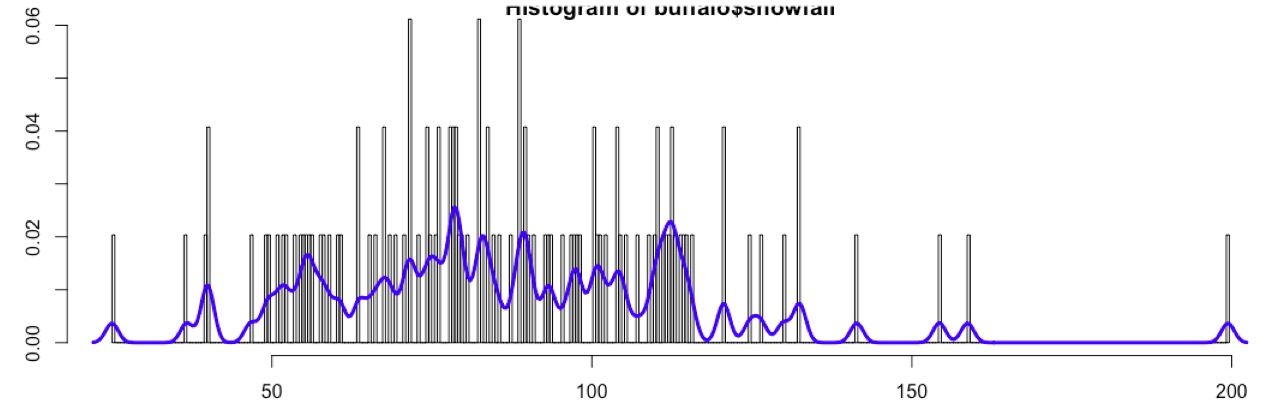
>hist(buffalo$snowfall, prob=TRUE, breaks=seq(25, 200, by=0.45))

> lines(density(buffalo$snowfall, adjust=0.1), col="blue", lwd=3)



Snowfall

Frequency Density



Snowfall

Frequency Density

Adding a density line to the data allows us to observe the fluctuation in frequency density, the larger the number of bars, the more fluctuating the density graph. We see that the density graph on the two-mode histogram is a more consistent and better fit than that of the three-mode histogram. In addition, although the one-mode graph would enable for a more stable density graph, it doesn’t represent the data as well as a two-mode histogram. We see that the best way to achieve a one-mode histogram is to decrease the number of bars and increase the “breaks” in R. This leads us to the conclusion that the ‘correct’ number of modes on a histogram is two as it gives a better, more accurate representation of the data.