**Question #1**

We have two datasets for Waterloo and New York. To compare mean values of temperatures for pre-, during-, and post-quarantine periods, we divided data into three subsets by Year: Pre-Quarantine Period is year 2019, During-Quarantine Period is year 2020, and Post-Quarantine Period is year 2021.

waterloo\_clean <- read.csv('./datasets/Waterloo\_Municipal\_IowaClean.csv')

nyc\_clean <- read.csv('./datasets/JFK\_International\_NewYorkClean.csv')

##Data splitting by year for Waterloo

data\_w<-split(waterloo\_clean, waterloo\_clean$Year)

waterloo\_pre<-data\_w$`2019`

waterloo\_during<-data\_w$`2020`

waterloo\_post<-data\_w$`2021`

w\_pre\_monthly<-waterloo\_pre[!is.na(waterloo\_pre$MonthlyMeanTemperature),]

w\_d\_monthly<-waterloo\_during[!is.na(waterloo\_during$MonthlyMeanTemperature),]

w\_post\_monthly<-waterloo\_post[!is.na(waterloo\_post$MonthlyMeanTemperature),]

##Data splitting by year for New York:

data\_nyc<-split(nyc\_clean, nyc\_clean$Year)

nyc\_pre<-data\_nyc$`2019`

nyc\_during<-data\_nyc$`2020`

nyc\_post<-data\_nyc$`2021`

nyc\_pre\_monthly<- nyc\_pre[!is.na(nyc\_pre$MonthlyMeanTemperature),]

nyc\_d\_monthly<- nyc\_during[!is.na(nyc\_during$MonthlyMeanTemperature),]

nyc\_post\_monthly<- nyc\_post[!is.na(nyc\_post$MonthlyMeanTemperature),]

We can see the summary information about average temperature in Waterloo by month for three years:

* Summary information for Waterloo:

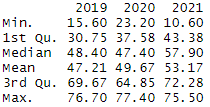
w\_summary<-round(cbind(summary(w\_pre\_monthly$MonthlyMeanTemperature), summary(w\_d\_monthly$MonthlyMeanTemperature), summary(w\_post\_monthly$MonthlyMeanTemperature)), 2)

colnames(w\_summary)=c("2019", "2020", "2021")

print("Monthly Average Temperatyre in Waterloo for pre-, during-, and post- COVID period:")

w\_summary

The Output:



* Summary information for New York:

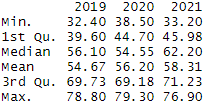
nyc\_summary<-round(cbind(summary(nyc\_pre\_monthly$MonthlyMeanTemperature), summary(nyc\_d\_monthly$MonthlyMeanTemperature), summary(nyc\_post\_monthly$MonthlyMeanTemperature)), 2)

colnames(nyc\_summary)=c("2019", "2020", "2021")

print("Monthly Average Temperatyre in New York for pre-, during-, and post- COVID period:")

nyc\_summary

The Output:



To compare Monthly Average Temperature for these three periods, we will plot the graph to see if there are any significant differences in the graphs:

* Waterloo:

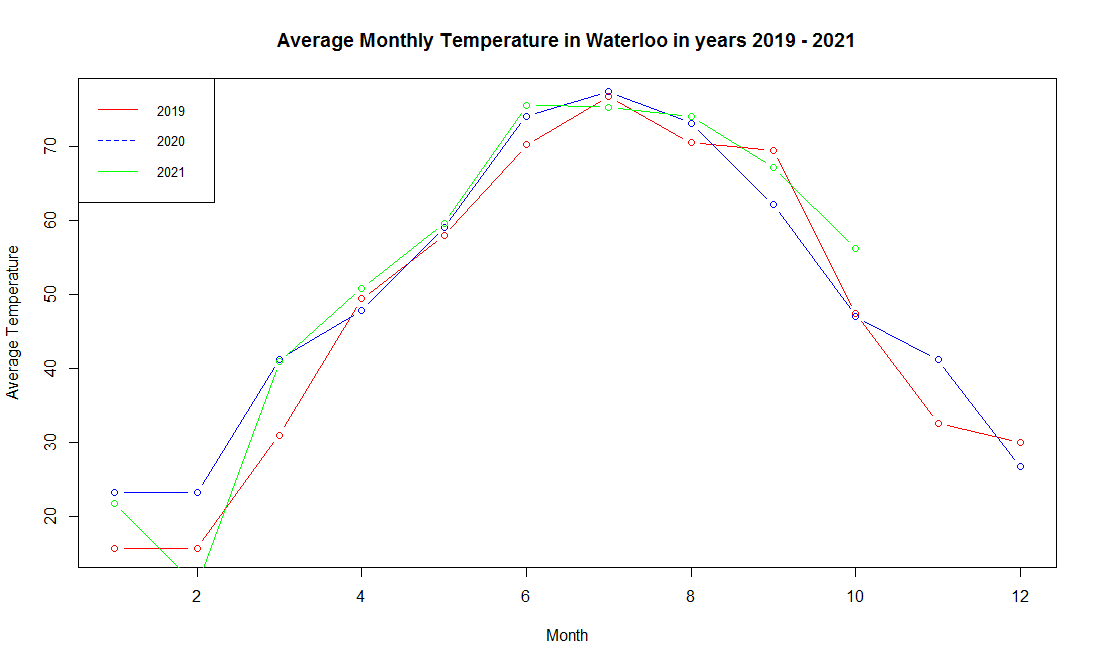
plot(w\_pre\_monthly$Month, w\_pre\_monthly$MonthlyMeanTemperature, type = "b", col= "red", xlab = "Month", ylab = "Average Temperature", main = "Average Monthly Temperature in Waterloo in years 2019 - 2021")

lines(w\_d\_monthly$Month, w\_d\_monthly$MonthlyMeanTemperature, type = "b", col = "blue")

lines(w\_post\_monthly$Month, w\_post\_monthly$MonthlyMeanTemperature, type = "b", col = "green")

legend("topleft", legend = c(2019, 2020, 2021), col = c("red", "blue", "green"), lty = 1:2, cex = 0.8)

The Output:



* New York:

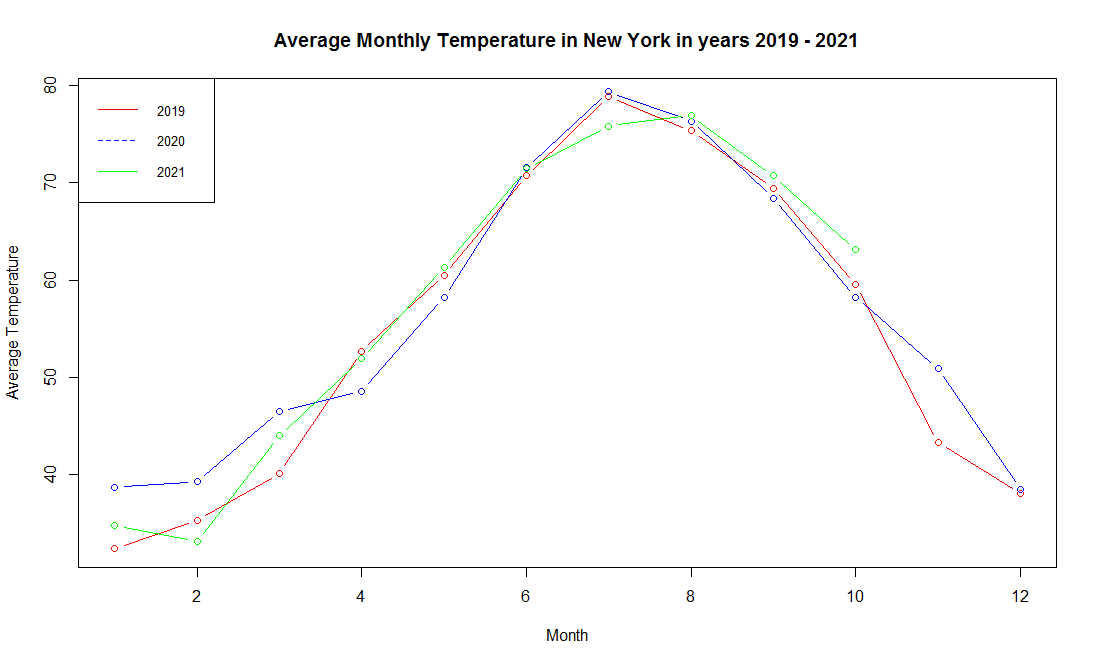
plot(nyc\_pre\_monthly$Month, nyc\_pre\_monthly$MonthlyMeanTemperature, type = "b", col= "red", xlab = "Month", ylab = "Average Temperature", main = "Average Monthly Temperature in New York in years 2019 - 2021")

lines(nyc\_d\_monthly$Month, nyc\_d\_monthly$MonthlyMeanTemperature, type = "b", col = "blue")

lines(nyc\_post\_monthly$Month, nyc\_post\_monthly$MonthlyMeanTemperature, type = "b", col = "green")

legend("topleft", legend = c(2019, 2020, 2021), col = c("red", "blue", "green"), lty = 1:2, cex = 0.8)

The Output:



As we can see, we cannot say if there is a difference between the average monthly temperature in pre-, during-, and post-covid periods in Waterloo or New York. Thus, we heed to test hypothesis about the equivalence of mean values for these periods. We will use t-test:

1. Waterloo:

t\_w1<-t.test(w\_pre\_monthly$MonthlyMeanTemperature,w\_d\_monthly$MonthlyMeanTemperature)

t\_w2<-t.test(w\_pre\_monthly$MonthlyMeanTemperature,w\_post\_monthly$MonthlyMeanTemperature)

t\_w3<-t.test(w\_d\_monthly$MonthlyMeanTemperature,w\_post\_monthly$MonthlyMeanTemperature)

##Results:

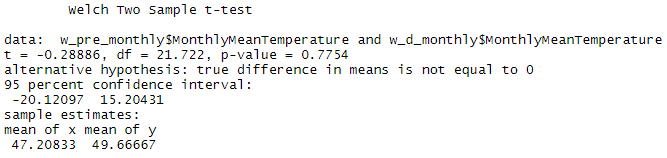
t\_w1

t\_w2

t\_w3

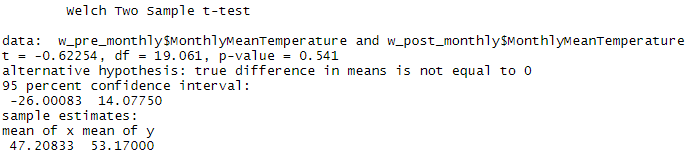
The results for t-tests:

* Comparing mean values of the average monthly temperature in Waterloo for pre-COVID and during-COVID periods:



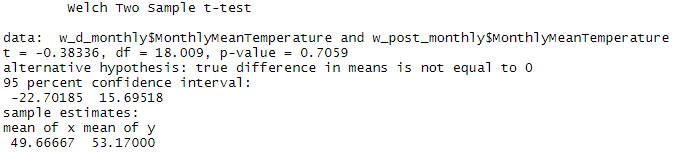
As we can see, the p-value is 0.7754, which is greater than . It means that we cannot reject the null-hypothesis about equality of means. Thus, the true difference in means for pre-Covid and during-Covid period in Waterloo is equal to zero.

* Comparing mean values of the average monthly temperature in Waterloo for pre-COVID and post-COVID periods:



As we can see, the p-value is 0.541, which is greater than . It means that we cannot reject the null-hypothesis about equality of means. Thus, the true difference in means for pre-Covid and post-Covid period in Waterloo is equal to zero.

* Comparing mean values of the average monthly temperature in Waterloo for during-COVID and post-COVID periods:



As we can see, the p-value is 0.7059, which is greater than 0.05. It means that we cannot reject the null-hypothesis about equality of means. Thus, the true difference in means for during-Covid and post-Covid period in Waterloo is equal to zero.

1. New York:

t\_ny1<-t.test(nyc\_pre\_monthly$MonthlyMeanTemperature,nyc\_d\_monthly$MonthlyMeanTemperature)

t\_ny2<-t.test(nyc\_pre\_monthly$MonthlyMeanTemperature,nyc\_post\_monthly$MonthlyMeanTemperature)

t\_ny3<-t.test(nyc\_d\_monthly$MonthlyMeanTemperature,nyc\_post\_monthly$MonthlyMeanTemperature)

##Results:

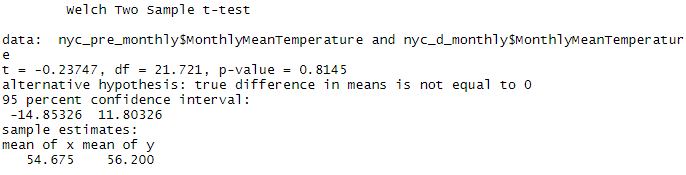
t\_ny1

t\_ny2

t\_ny3

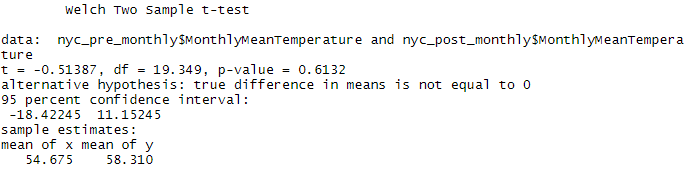
The results for t-tests:

* Comparing mean values of the average monthly temperature in New York for pre-COVID and during-COVID periods:



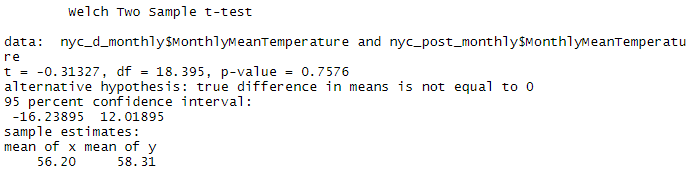
As we can see, the p-value is 0.8145, which is greater than . It means that we cannot reject the null-hypothesis about equality of means. Thus, the true difference in means for pre-Covid and during-Covid period in New York is equal to zero.

* Comparing mean values of the average monthly temperature in New York for pre-COVID and post-COVID periods:



As we can see, the p-value is 0.6132, which is greater than . It means that we cannot reject the null-hypothesis about equality of means. Thus, the true difference in means for pre-Covid and post-Covid period in New York is equal to zero.

* Comparing mean values of the average monthly temperature in New York for during-COVID and post-COVID periods:



As we can see, the p-value is 0.7576, which is greater than 0.05. It means that we cannot reject the null-hypothesis about equality of means. Thus, the true difference in means for during-Covid and post-Covid period in New York is equal to zero.

Thus, as we can see, the average monthly temperature in pre-, during-, and post-COVID periods does not have statistically significant difference, i.e. we may conclude that the average monthly temperature did not change significantly in COVID period both in Waterloo and New York.

**Question #2**

At first, we will use 'Year' variable as factor variable:

nyc\_clean$Year<-as.factor(nyc\_clean$Year)

By using this data and 'Year' as factor, we can build a boxplot to see the average daily temperature in New York in years 2019 (pre-COVID time), 2020 (during-COVID time), and 2021 (post-COVID time):

library(ggplot2)

library(RColorBrewer)

ggplot(data = nyc\_clean, aes(y = DailyAverageDryBulbTemperature, x = Year), fill = "class") +

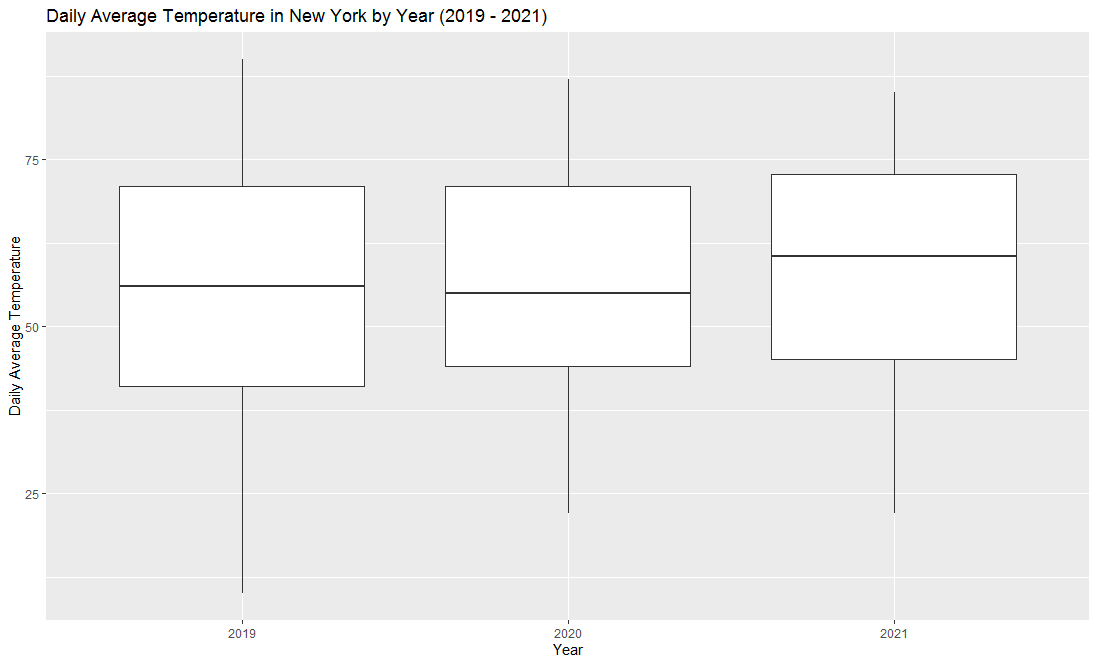
geom\_boxplot()+

xlab("Year")+

ylab("Daily Average Temperature")+

ggtitle("Daily Average Temperature in New York by Year (2019 - 2021)")

The output:

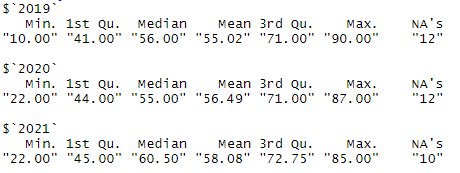


As we can see from the boxplot, we can assume that the average daily temperature in New York during COVID-19 has changed: it is slightly greater for year 2020 compared to 2019; it's greater for year 2021 compared to 2019 and 2020.

The summary statistics:

tapply(nyc\_clean$DailyAverageDryBulbTemperature, nyc\_clean$Year, function(x) format(summary(x)))

The output:



As we can see from the summary statistic, our assumption should be right. Let check it with t-test:

tt\_ny1<-t.test(nyc\_pre$DailyAverageDryBulbTemperature, nyc\_during$DailyAverageDryBulbTemperature)

tt\_ny2<-t.test(nyc\_pre$DailyAverageDryBulbTemperature, nyc\_post$DailyAverageDryBulbTemperature)

tt\_ny3<-t.test(nyc\_during$DailyAverageDryBulbTemperature, nyc\_post$DailyAverageDryBulbTemperature)

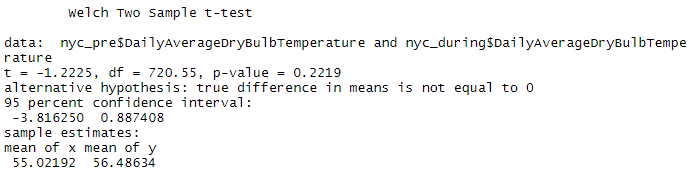
tt\_ny1

tt\_ny2

tt\_ny3

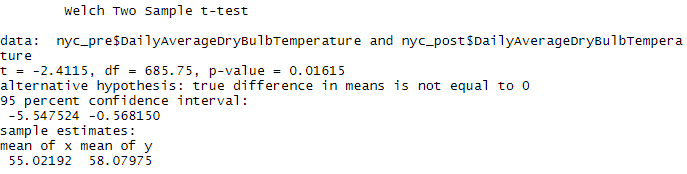
The results:

* Comparing the means for Daily Average Temperature in New York in pre-COVID and during-COVID periods:



As we can see, the p-value of the test is 0.2219, which is greater than . It means that at 95% confidence level, we cannot reject the null-hypothesis: the true difference in means is equal to zero. Thus, there is no difference in daily average temperature in New York in pre- and during-COVID periods.

* Comparing the means for Daily Average Temperature in New York in pre-COVID and post-COVID periods:

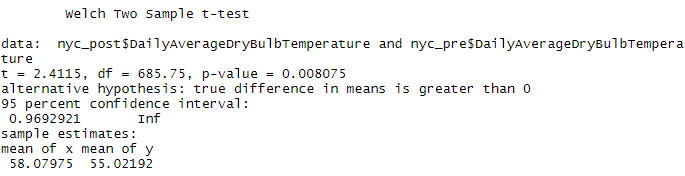


As we can see, the p-vlue is 0.01615, which is lower than . Thus, we should reject the null-hypothesis about equality of means. Thus, there is a statistically significant difference in pre- and post- daily average temperature in New York. We can check if our assumption is true and the post-COVID average daily temperature is higher than in pre-COVID period in NY:

tt\_ny2\_new<-t.test(nyc\_post$DailyAverageDryBulbTemperature, nyc\_pre$DailyAverageDryBulbTemperature, alternative = "greater")

tt\_ny2\_new

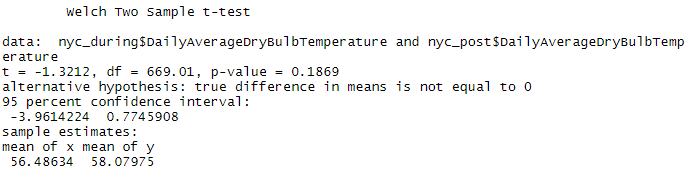
The results:



As we can see, the p-value is 0.008075, which is lower than : we should reject the null-hypothesis. It means that the true difference in means between post-COVID daily average temperature and pre-COVID daily average temperature is greater than zero. Thus, the post-COVID daily average temperature is higher than the pre-COVID daily average temperature in New York.

Let see what we have for during- and post-COVID daily average temperature difference:

* Comparing the means for Daily Average Temperature in New York in during-COVID and post-COVID periods:



As we can see, the p-value is 0.1869, which is greater than , which means that we cannot reject the null-hypothesis. Thus, the difference between the mean values for daily average temperature in during- and post-COVID periods are equal to zero. Thus, there is no statistically significant difference in daily average temperatures in during- and post-COVID periods.

Thus, we can conclude that the post-COVID daily average temperature in New York is higher than in pre-COVID period, but it's not different from during-COVID period. The daily average temperature in pre-COVID and during-COVID period are not significantly different.

**Question #3**

At first, we will use 'Year' variable as factor variable:

waterloo\_clean$Year<-as.factor(waterloo\_clean$Year)

By using this data and 'Year' as factor, we can build a boxplot to see the average daily temperature in Waterloo in years 2019 (pre-COVID time), 2020 (during-COVID time), and 2021 (post-COVID time):

ggplot(data = waterloo\_clean, aes(y = DailyAverageDryBulbTemperature, x = Year), fill = "class") +

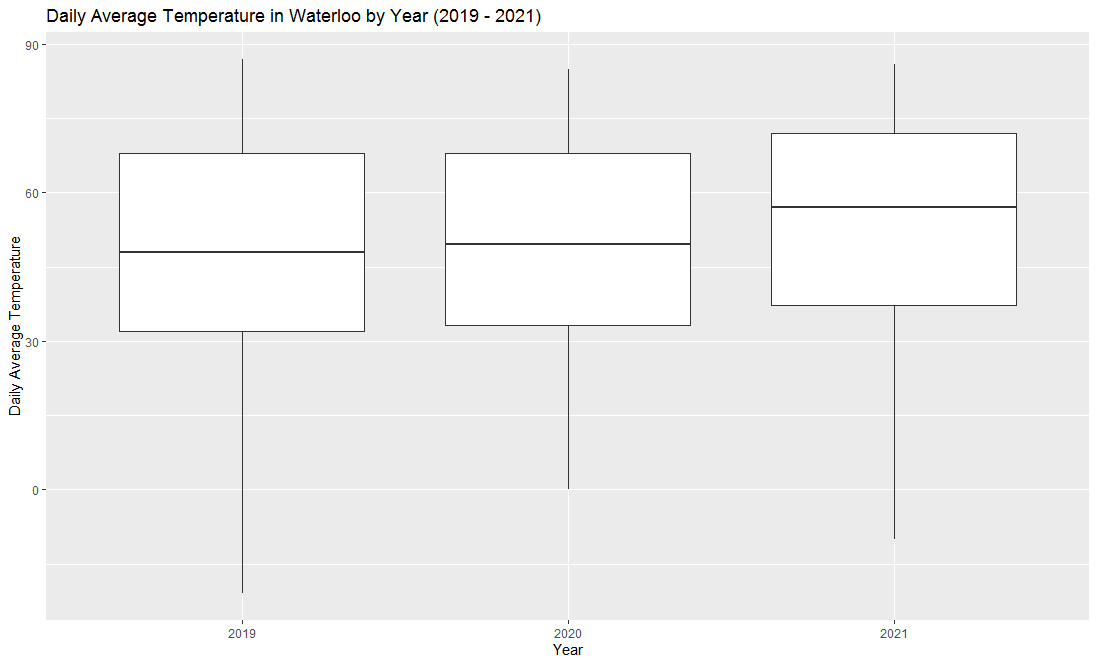
geom\_boxplot()+

xlab("Year")+

ylab("Daily Average Temperature")+

ggtitle("Daily Average Temperature in Waterloo by Year (2019 - 2021)")

The output:

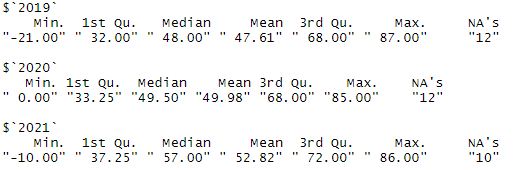


As we can see from the boxplot, we can assume that the average daily temperature in this city during COVID-19 has changed: it is slightly greater for year 2020 compared to 2019; it's greater for year 2021 compared to 2019 and slightly greater compared to year 2020.

The summary statistics:

tapply(waterloo\_clean$DailyAverageDryBulbTemperature, waterloo\_clean$Year, function(x) format(summary(x)))

The output:



As we can see from the summary statistic, our assumption should be right. Let check it with t-test:

tt\_w1<-t.test(waterloo\_pre$DailyAverageDryBulbTemperature, waterloo\_during$DailyAverageDryBulbTemperature)

tt\_w2<-t.test(waterloo\_pre$DailyAverageDryBulbTemperature, waterloo\_post$DailyAverageDryBulbTemperature)

tt\_w3<-t.test(waterloo\_during$DailyAverageDryBulbTemperature, waterloo\_post$DailyAverageDryBulbTemperature)

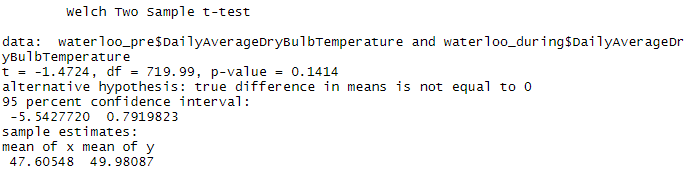
tt\_w1

tt\_w2

tt\_w3

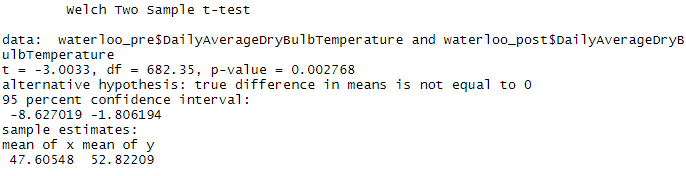
The results are:

* Comparing the means for Daily Average Temperature in Waterloo in pre-COVID and during-COVID periods:



As we can see, the p-value of the test is 0.1414, which is greater than . It means that at 95% confidence level, we cannot reject the null-hypothesis: the true difference in means is equal to zero. Thus, there is no difference in daily average temperature in Waterloo in pre- and during-COVID periods.

* Comparing the means for Daily Average Temperature in Waterloo in pre-COVID and post-COVID periods:

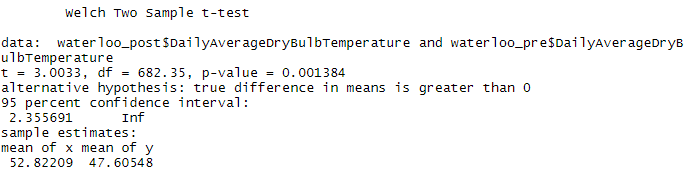


As we can see, the p-value is 0.002768, which is lower than . Thus, we should reject the null-hypothesis about equality of means. Thus, there is a statistically significant difference in pre- and post- daily average temperature in Waterloo. We can check if our assumption is true and the post-COVID average daily temperature is higher than in pre-COVID period in Waterloo:

tt\_w2\_new<-t.test(waterloo\_post$DailyAverageDryBulbTemperature, waterloo\_pre$DailyAverageDryBulbTemperature, alternative = "greater")

tt\_w2\_new

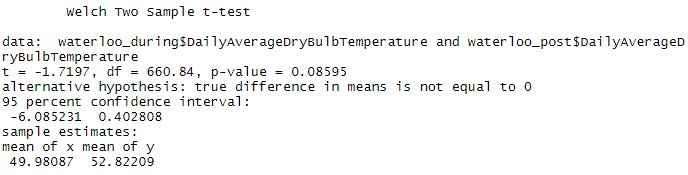
The result:



As we can see, the p-value is 0.001384, which is lower than : we should reject the null-hypothesis. It means that the true difference in means between post-COVID daily average temperature and pre-COVID daily average temperature is greater than zero. Thus, the post-COVID daily average temperature is higher than the pre-COVID daily average temperature in Waterloo.

Let see what we have for during- and post-COVID daily average temperature difference:

* Comparing the means for Daily Average Temperature in Waterloo in during-COVID and post-COVID periods:

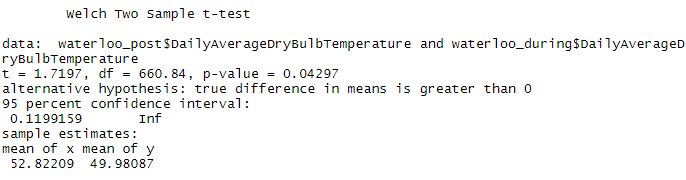


As we can see, the p-value is 0.08595, which is greater than . But it is lower than , which means that at 90% of confidence we can say that the mean values of daily average temperature for during- and post-COVID periods are different. We can check this:

tt\_w3\_new<-t.test(waterloo\_post$DailyAverageDryBulbTemperature, waterloo\_during$DailyAverageDryBulbTemperature, alternative = "greater")

tt\_w3\_new

The result:



As we can see, the p-value is 0.04297, which is lower than : we should reject the null-hypothesis. It means that the true difference in means between mean values of daily average temperature in post- and during-COVID period is greater than zero. Thus, the mean value of post-COVID daily average temperature is higher than the mean value of during-COVID daily average temperature in Waterloo.

Thus, we can conclude that the post-COVID daily average temperature in Waterloo is higher than in pre-COVID and during-COVID periods. The daily average temperature in pre-COVID and during-COVID period is not significantly different.

**Question #4**

Based on the previous analysis, we can say that the quarantine has not been effective in reducing daily mean temperature or average temperature in general. We can see, that post-COVID average daily temperature is higher than pre-COVID average daily temperature both for New York and waterloo. The reason that people during the COVID and in post-COVID period stay in their homes and use more electricity, cars, buy more food, etc.