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:

```
           2019  2020  2021
Min.      15.60 23.20 10.60
1st Qu.   30.75 37.58 43.38
Median    48.40 47.40 57.90
Mean      47.21 49.67 53.17
3rd Qu.   69.67 64.85 72.28
Max.      76.70 77.40 75.50
```

- 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:

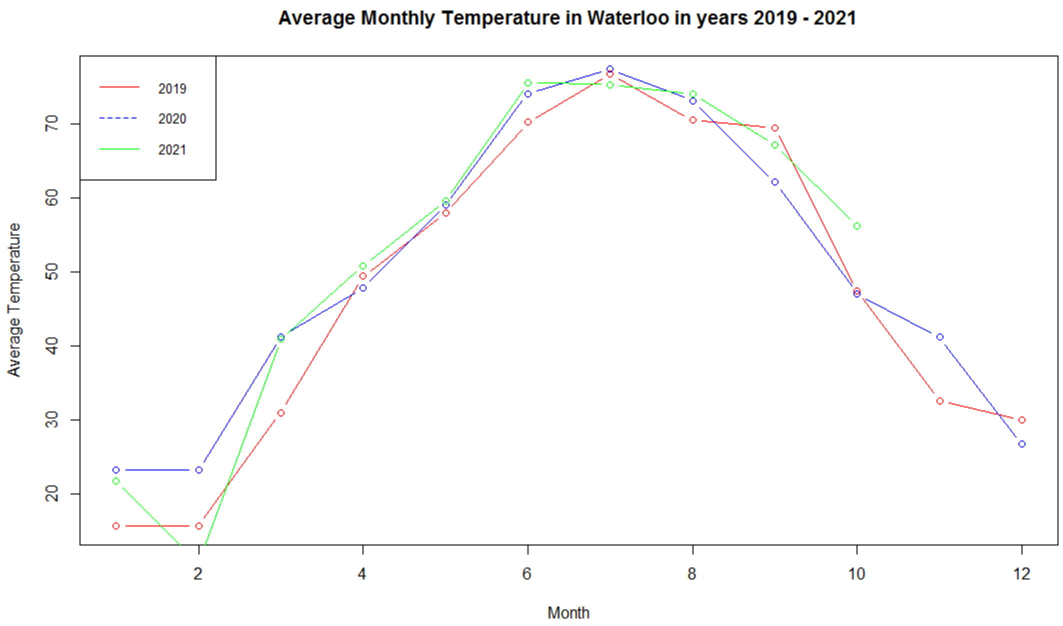
```
           2019  2020  2021
Min.      32.40 38.50 33.20
1st Qu.   39.60 44.70 45.98
Median    56.10 54.55 62.20
Mean      54.67 56.20 58.31
3rd Qu.   69.73 69.18 71.23
Max.      78.80 79.30 76.90
```

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)
```

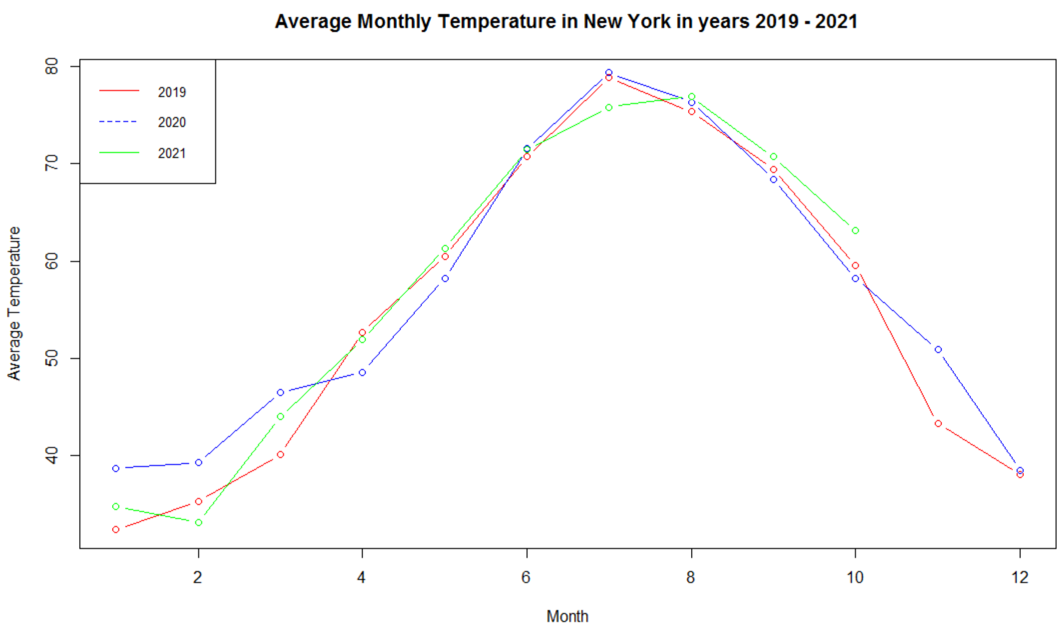
The Output:



- 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:

a) Waterloo:

```
t_w1<-
t.test(w_pre_monthly$MonthlyMeanTemperature,w_d_monthly$MonthlyMeanT
emperature)
t_w2<-
t.test(w_pre_monthly$MonthlyMeanTemperature,w_post_monthly$MonthlyMe
anTemperature)
t_w3<-
t.test(w_d_monthly$MonthlyMeanTemperature,w_post_monthly$MonthlyMean
Temperature)
##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:

```
        Welch Two Sample t-test

data:  w_pre_monthly$MonthlyMeanTemperature and w_d_monthly$MonthlyMeanTemperature
t = -0.28886, df = 21.722, p-value = 0.7754
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -20.12097  15.20431
sample estimates:
mean of x mean of y
 47.20833  49.66667
```

As we can see, the p-value is 0.7754, which is greater than $\alpha = 0.05$. 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:

```
        Welch Two Sample t-test

data:  w_pre_monthly$MonthlyMeanTemperature and w_post_monthly$MonthlyMeanTemperature
t = -0.62254, df = 19.061, p-value = 0.541
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26.00083  14.07750
sample estimates:
mean of x mean of y 
 47.20833  53.17000
```

As we can see, the p-value is 0.541, which is greater than $\alpha = 0.05$. 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:

```
        Welch Two Sample t-test

data:  w_d_monthly$MonthlyMeanTemperature and w_post_monthly$MonthlyMeanTemperature
t = -0.38336, df = 18.009, p-value = 0.7059
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.70185  15.69518
sample estimates:
mean of x mean of y 
 49.66667  53.17000
```

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.

b) New York:

```
t_ny1<-
t.test(nyc_pre_monthly$MonthlyMeanTemperature,nyc_d_monthly$MonthlyM
eanTemperature)
t_ny2<-
t.test(nyc_pre_monthly$MonthlyMeanTemperature,nyc_post_monthly$Month
lyMeanTemperature)
t_ny3<-
t.test(nyc_d_monthly$MonthlyMeanTemperature,nyc_post_monthly$Monthly
MeanTemperature)
##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:

```
        Welch Two Sample t-test

data:  nyc_pre_monthly$MonthlyMeanTemperature and nyc_d_monthly$MonthlyMeanTemperatur
e
t = -0.23747, df = 21.721, p-value = 0.8145
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.85326  11.80326
sample estimates:
mean of x mean of y
   54.675    56.200
```

As we can see, the p-value is 0.8145, which is greater than $\alpha = 0.05$. 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:

```
        Welch Two Sample t-test

data:  nyc_pre_monthly$MonthlyMeanTemperature and nyc_post_monthly$MonthlyMeanTempera
ture
t = -0.51387, df = 19.349, p-value = 0.6132
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.42245  11.15245
sample estimates:
mean of x mean of y
   54.675    58.310
```

As we can see, the p-value is 0.6132, which is greater than $\alpha = 0.05$. 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:

```
        Welch Two Sample t-test

data:  nyc_d_monthly$MonthlyMeanTemperature and nyc_post_monthly$MonthlyMeanTemperatu
re
t = -0.31327, df = 18.395, p-value = 0.7576
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -16.23895  12.01895
sample estimates:
mean of x mean of y
    56.20     58.31
```

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

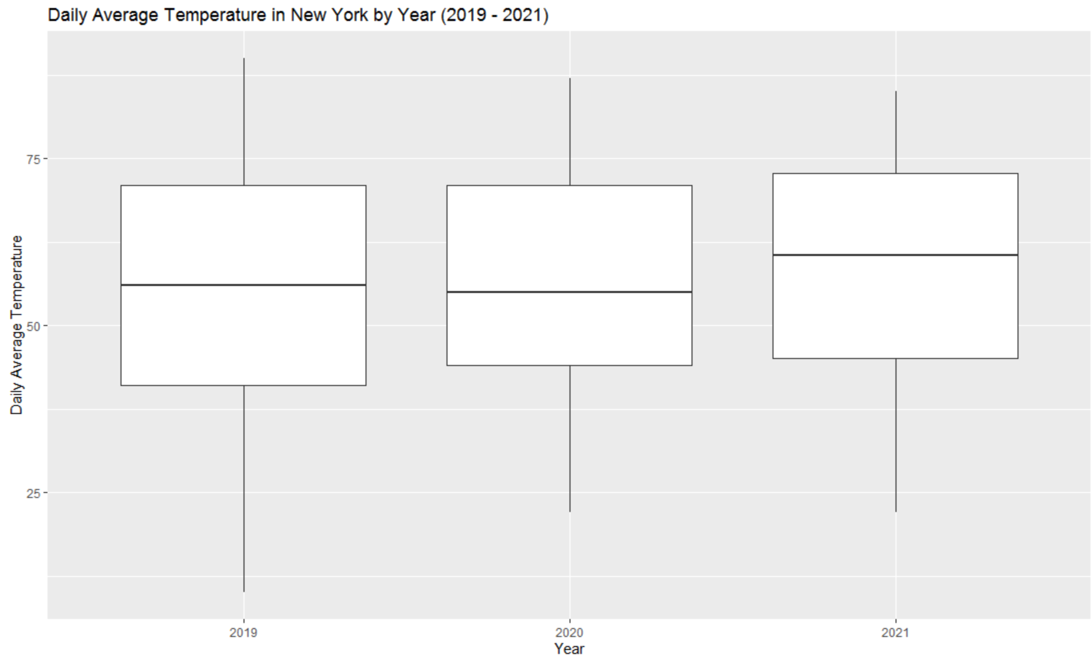
At first, we will use 'Year' variable as factor variable:

```
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:

```
$`2019`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
"10.00" "41.00" "56.00" "55.02" "71.00" "90.00"    "12" 

$`2020`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
"22.00" "44.00" "55.00" "56.49" "71.00" "87.00"    "12" 

$`2021`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
"22.00" "45.00" "60.50" "58.08" "72.75" "85.00"    "10" 
```

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:

```
        Welch Two Sample t-test

data:  nyc_pre$DailyAverageDryBulbTemperature and nyc_during$DailyAverageDryBulbTempe
rature
t = -1.2225, df = 720.55, p-value = 0.2219
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.816250  0.887408
sample estimates:
mean of x mean of y
 55.02192  56.48634
```

As we can see, the p-value of the test is 0.2219, which is greater than $\alpha = 0.05$. 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:

```
        Welch Two Sample t-test

data:  nyc_pre$DailyAverageDryBulbTemperature and nyc_post$DailyAverageDryBulbTempera
ture
t = -2.4115, df = 685.75, p-value = 0.01615
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.547524 -0.568150
sample estimates:
mean of x mean of y
 55.02192  58.07975
```

As we can see, the p-vlue is 0.01615, which is lower than $\alpha = 0.05$. 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:

```
        Welch Two Sample t-test

data:  nyc_post$DailyAverageDryBulbTemperature and nyc_pre$DailyAverageDryBulbTempera
ture
t = 2.4115, df = 685.75, p-value = 0.008075
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.9692921       Inf
sample estimates:
mean of x mean of y 
 58.07975  55.02192 
```

As we can see, the p-value is 0.008075, which is lower than $\alpha = 0.05$: 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:

```
        Welch Two Sample t-test

data:  nyc_during$DailyAverageDryBulbTemperature and nyc_post$DailyAverageDryBulbTemp
erature
t = -1.3212, df = 669.01, p-value = 0.1869
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.9614224  0.7745908
sample estimates:
mean of x mean of y 
 56.48634  58.07975 
```

As we can see, the p-value is 0.1869, which is greater than $\alpha = 0.05$, 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

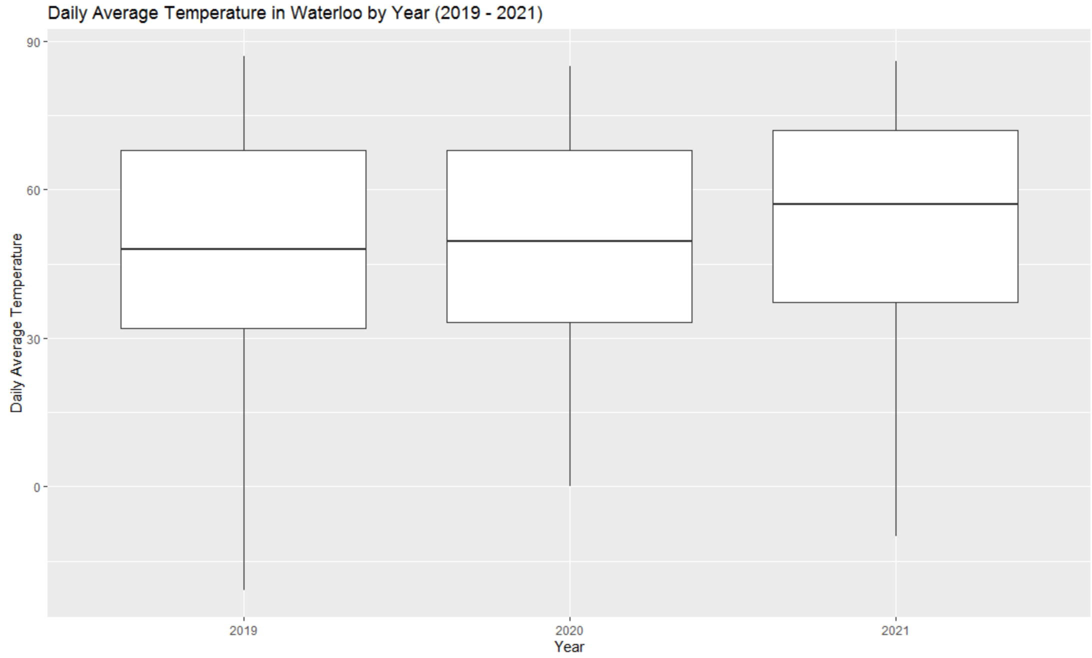
At first, we will use 'Year' variable as factor variable:

```
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:

```
$`2019`
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
"-21.00" " 32.00" " 48.00" " 47.61" " 68.00" " 87.00"      "12"

$`2020`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
" 0.00" "33.25" "49.50" "49.98" "68.00" "85.00"    "12"

$`2021`
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
"-10.00" " 37.25" " 57.00" " 52.82" " 72.00" " 86.00"      "10"
```

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:

```
        Welch Two Sample t-test

data:  waterloo_pre$DailyAverageDryBulbTemperature and waterloo_during$DailyAverageDr
yBulbTemperature
t = -1.4724, df = 719.99, p-value = 0.1414
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.5427720  0.7919823
sample estimates:
mean of x mean of y
 47.60548  49.98087
```

As we can see, the p-value of the test is 0.1414, which is greater than $\alpha = 0.05$. 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:

```
        Welch Two Sample t-test

data:  waterloo_pre$DailyAverageDryBulbTemperature and waterloo_post$DailyAverageDryB
ulbTemperature
t = -3.0033, df = 682.35, p-value = 0.002768
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.627019 -1.806194
sample estimates:
mean of x mean of y
 47.60548  52.82209
```

As we can see, the p-value is 0.002768, which is lower than $\alpha = 0.05$. 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:

```
        Welch Two Sample t-test

data:  waterloo_post$DailyAverageDryBulbTemperature and waterloo_pre$DailyAverageDryB
ulbTemperature
t = 3.0033, df = 682.35, p-value = 0.001384
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 2.355691      Inf
sample estimates:
mean of x mean of y 
 52.82209  47.60548 
```

As we can see, the p-value is 0.001384, which is lower than $\alpha = 0.05$: 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:

```
        Welch Two Sample t-test

data:  waterloo_during$DailyAverageDryBulbTemperature and waterloo_post$DailyAverageD
ryBulbTemperature
t = -1.7197, df = 660.84, p-value = 0.08595
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.085231  0.402808
sample estimates:
mean of x mean of y 
 49.98087  52.82209 
```

As we can see, the p-value is 0.08595, which is greater than $\alpha = 0.05$. But it is lower than $\alpha = 0.10$, 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:

```
        Welch Two Sample t-test

data:  waterloo_post$DailyAverageDryBulbTemperature and waterloo_during$DailyAverageD
ryBulbTemperature
t = 1.7197, df = 660.84, p-value = 0.04297
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.1199159       Inf
sample estimates:
mean of x mean of y 
 52.82209  49.98087 
```

As we can see, the p-value is 0.04297, which is lower than $\alpha = 0.05$: 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.