MT5763: Software for Data Analysis

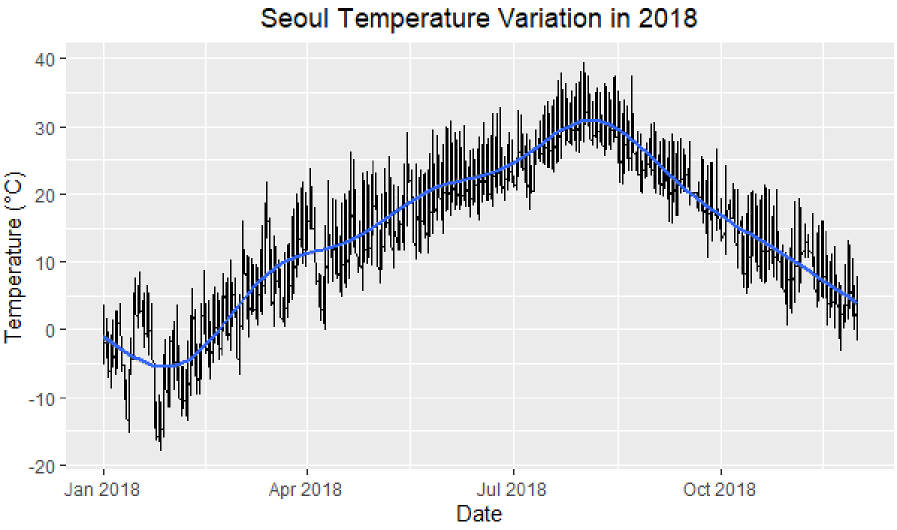
Assignment I Report

Link to Github repository: https://github.com/Magnifico1/MT5763_1_-180008620-

DATA VISUALISATION

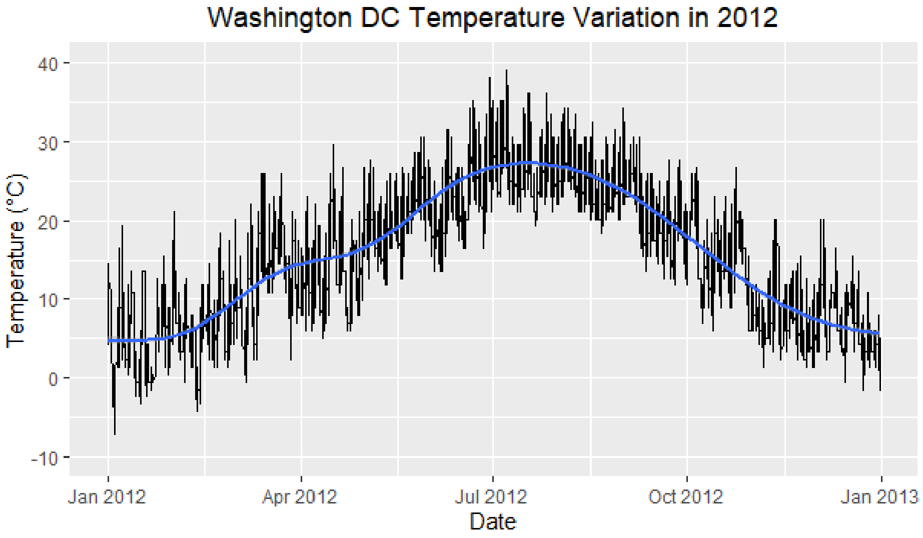
How does air temperature vary over the course of a year?

Seoul:



Above is shown a graph of temperature against the date for the chosen year 2018 in Seoul. The daily temperature is very changeable on consecutive days, resulting in the fluctuating pattern seen. The smoothed nonparametric line clearly shows a gradual increase in average temperature from around February, where average temperatures are near -5°C to around the beginning of August, where they are just above 30°C. At this point the temperature begins to decline towards the end of the year.

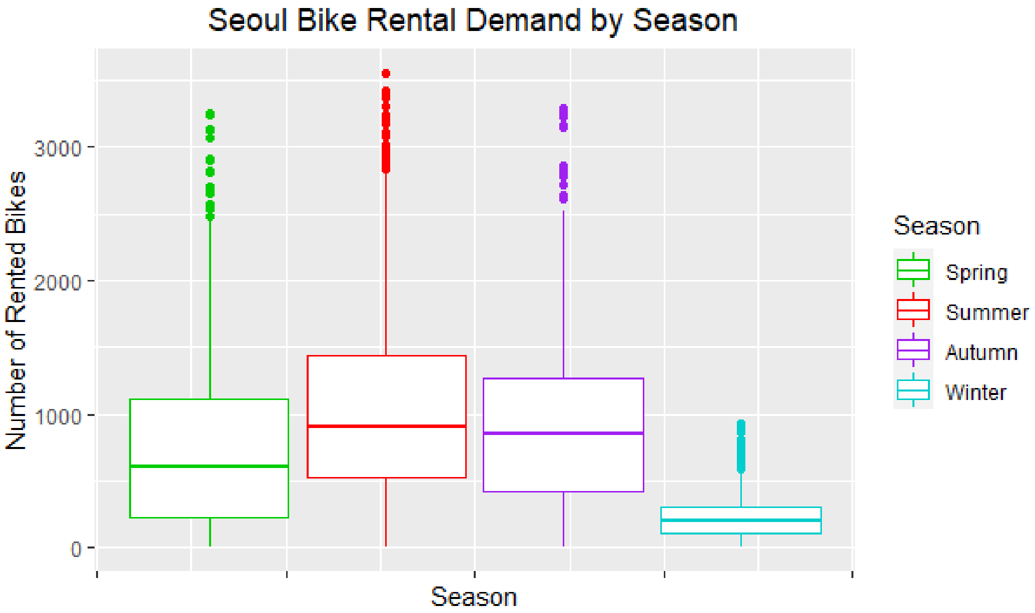
DC:



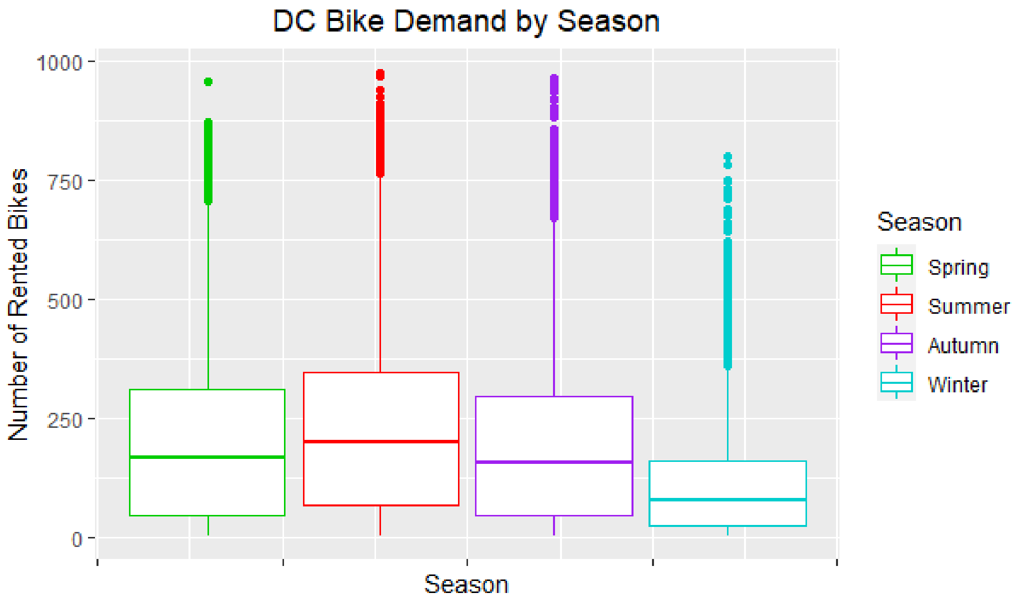
DC's temperature variation over the chosen year of 2012 appears less extreme than for Seoul. As with Seoul's case, the temperature increases from around February, with an average of 5°C, to summer. The average temperature starts declining from around August, when average temperatures are about 27°C, to the end of the year when temperatures fall to an average of 5°C again.

Both graphs are in line with expectation, as they show the hottest temperatures in summer, and the coldest in winter.

Do seasons affect the average number of rented bikes?
Seoul:



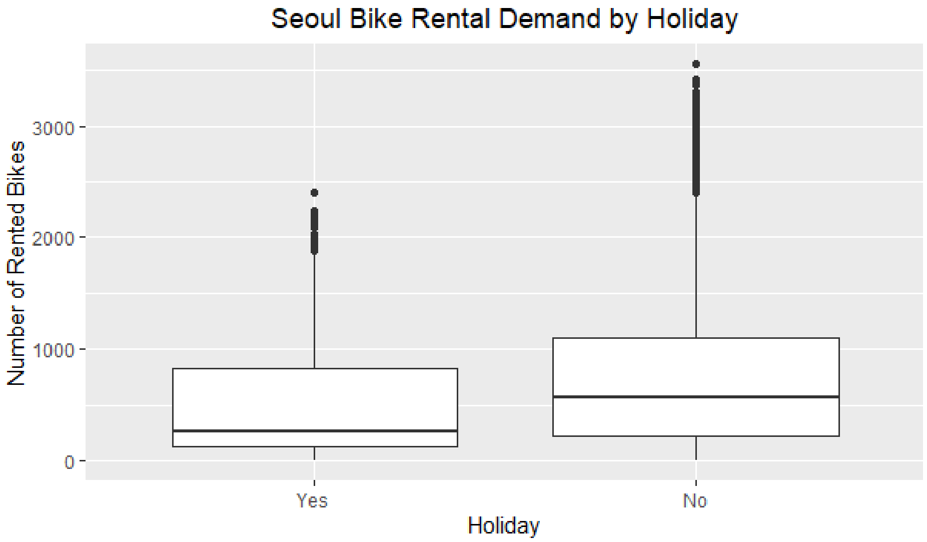
The seasons of spring and autumn have comparable means and interquartile ranges, thus not showing much difference in bike demand between these seasons. There is, however, a noticeable increase in the response variable from summer to spring, as shown by a greater median for summer. Winter is clearly very different from the other two seasons in its distribution of the response variable. The mean is much lower than the three other seasons, and both the maximum value and interquartile range are significantly smaller than the other seasons'. Thus, we see a pattern in the response variable described by increase towards summer and subsequent decline towards winter.



We see a similar pattern in bike demand for Washington DC as with Seoul, in that the average number of rented bikes increases from spring to summer and continues to decline afterwards towards winter, the season with the lowest average number of rented bikes. In contrast to Seoul, the distribution of rented bikes in winter is not so different to the other seasons. Although both the IQR and median of the response in winter are less than in the other seasons in DC, they are proportionally significantly greater than in Seoul's case, resulting in a less dramatic decline in bike demand in winter.

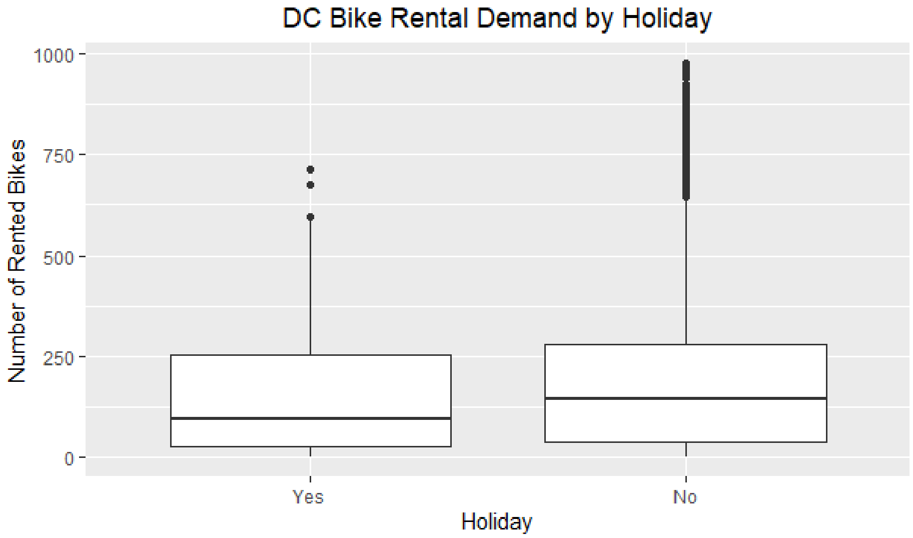
Overall, it can be seen that seasons do affect the average number of rented bikes. This number is highest in summer and lowest in winter with spring and autumn demand lying at a similar level inbetween.

Do holidays increase or decrease the demand for rented bikes?
Seoul:



Seoul shows a marked difference in bike demand between holidays and days which are not holidays. Holidays can be seen to have a reduced bike rental demand than non-holidays, as shown by the lower median average. The maximum value of the response for holidays is also significantly less than for non-holidays.

DC:

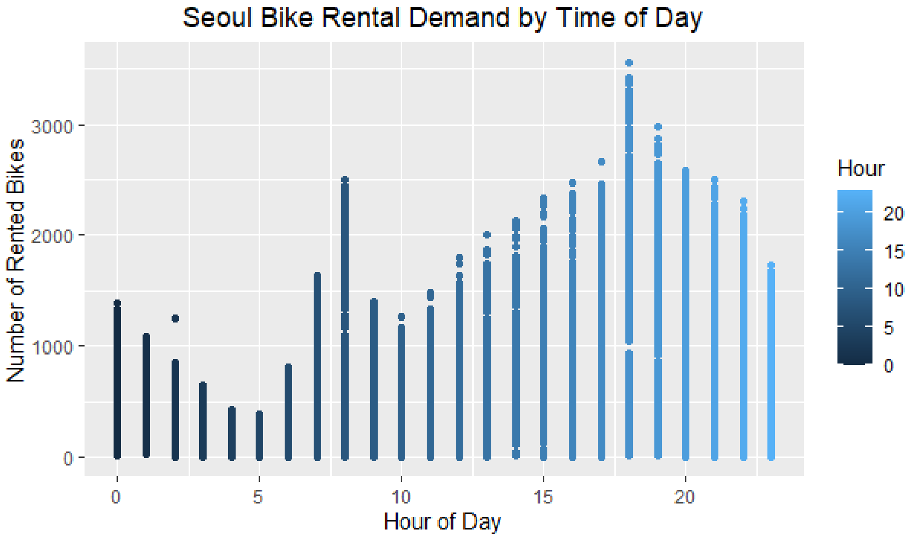


In DC the median number of bikes rented on non-holidays is also greater than on holidays, however, there is a less marked difference than in Seoul, and the IQRs are very similar. Despite this, the maximum value of the response is proportionally similarly low for holidays as compared to non-holidays.

Overall, average bike demand is slightly decreased on days which are holidays.

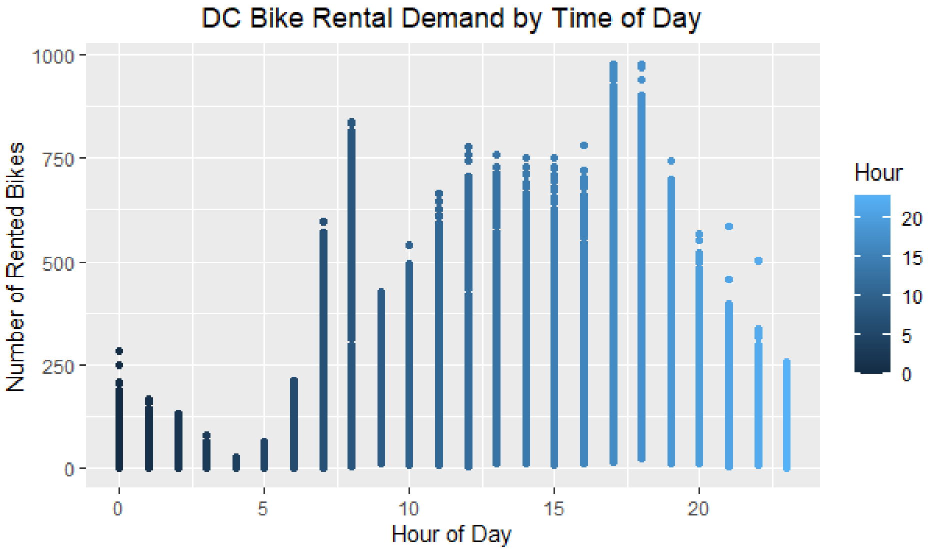
How does time of day affect demand for rented bikes?

Seoul:



Bike rental demand sees a general increase from a low at 5AM to a high at 6PM, at which time demand begins to decline again. There is a noticeable spike in demand on top of this during the hours of 7AM and 8AM, presumably caused by commuters to work.

DC:

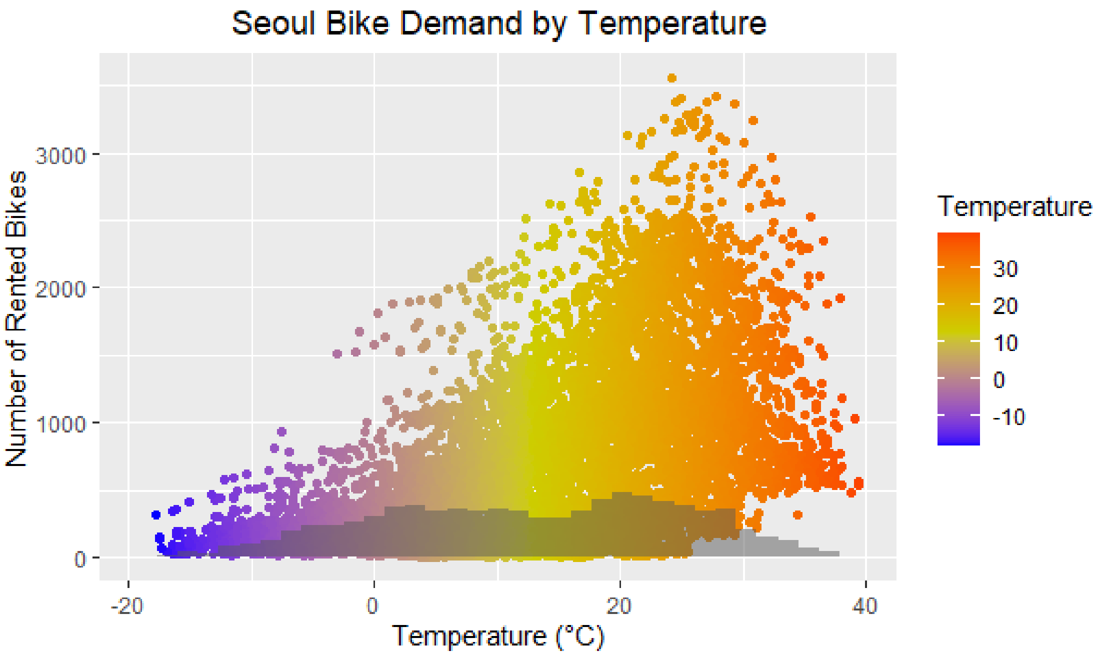


Bike demand follows an almost identical pattern to Seoul - it increases from the early morning (4AM) onwards, with a spike in demand during the hours of 7AM and 8AM. Like Seoul, the peak of demand occurs in the early evening, with the hours of 5PM and 6PM showing the greatest numbers of bike rentals. Bike demand subsequently dwindles until the early morning the next day. Unlike Seoul, however, there seems to be a plateau in demand from 12PM to 4PM.

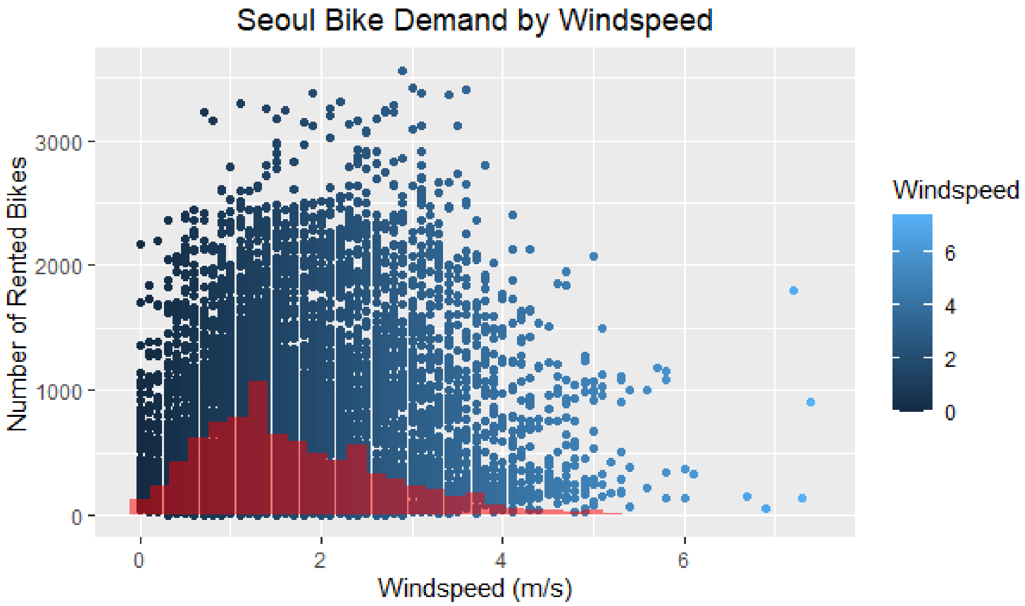
Overall it can be seen that bike demand is certainly affected by time of day, as it generally increases from the early morning to the early evening, when it starts to decline again. Peaks in demand can be seen at times when commuters are travelling to and from work.

Is there any association between bike demand and the three meteorological variables (air temperature, windspeed, and humidity)?

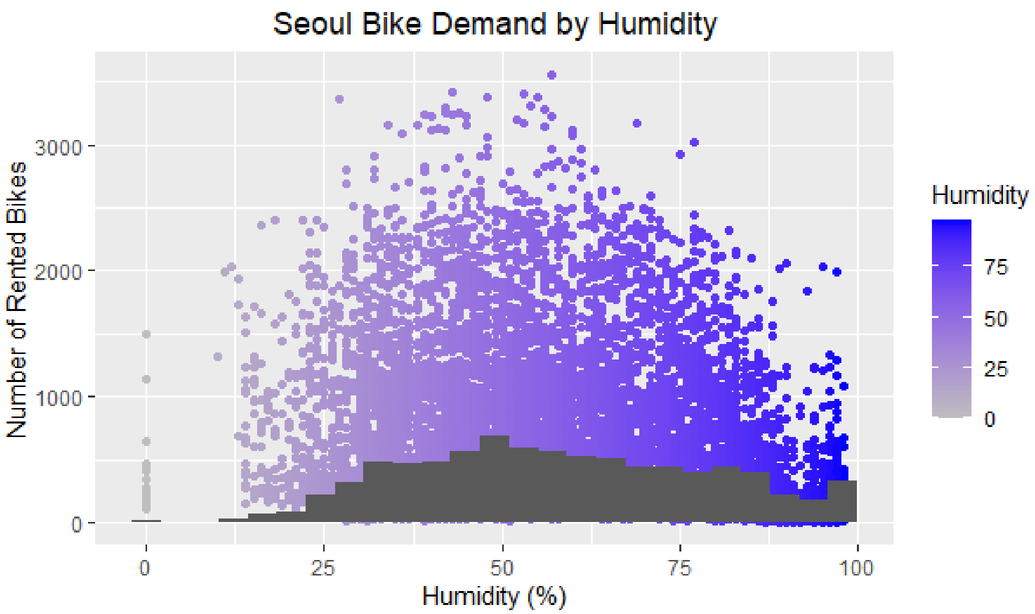
Seoul:



The above plot shows a deviation of bike demand from what we would expect if there were no association between bike demand and temperature. The distribution of temperatures is given by a histogram for comparison with the data. Since the points are heavily left-skewed, it can be seen that bike demand generally increases with increasing temperature. This demand seems to drop off at temperatures higher than about 23°C. However, at temperatures greater than this, there is also a higher minimum demand, with few instances where the number of rented bikes was very low.

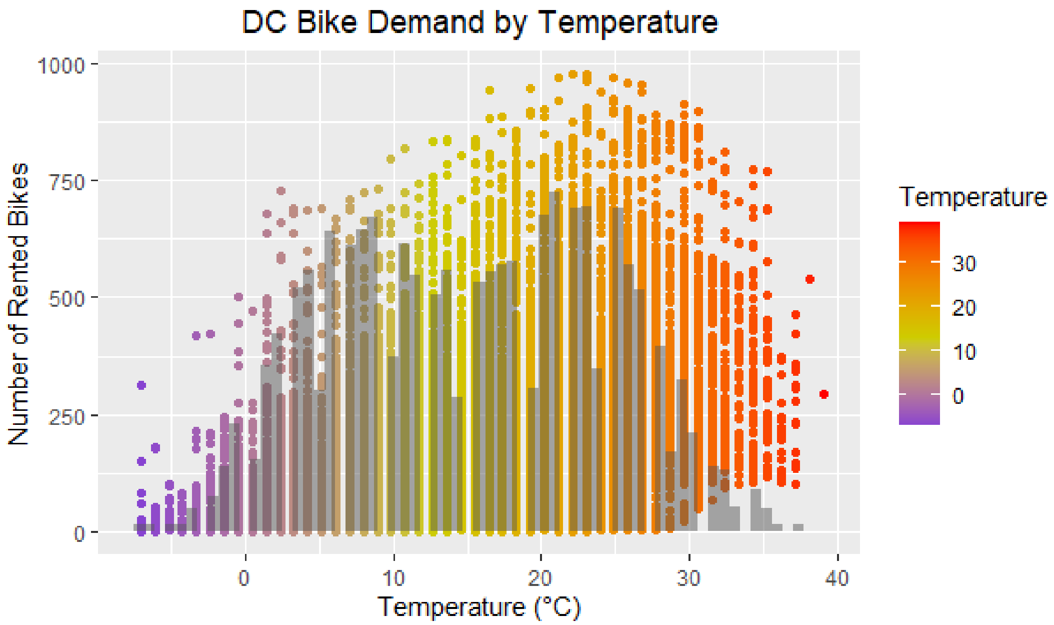


Looking at windspeed we can see that more bikes tend to be rented on days with lower windspeeds. However, the distribution of windspeeds, shown by the red histogram, is such that most of the days have fairly low windspeeds. Therefore, there doesn't seem to be much of an association in Seoul between windspeed and bike demand, as demand is fairly uniform across the range of windspeeds in which most data points lie (0-4m/s). It may be said that the days on which the most bikes are rented have a windspeed between 1m/s and 4m/s.

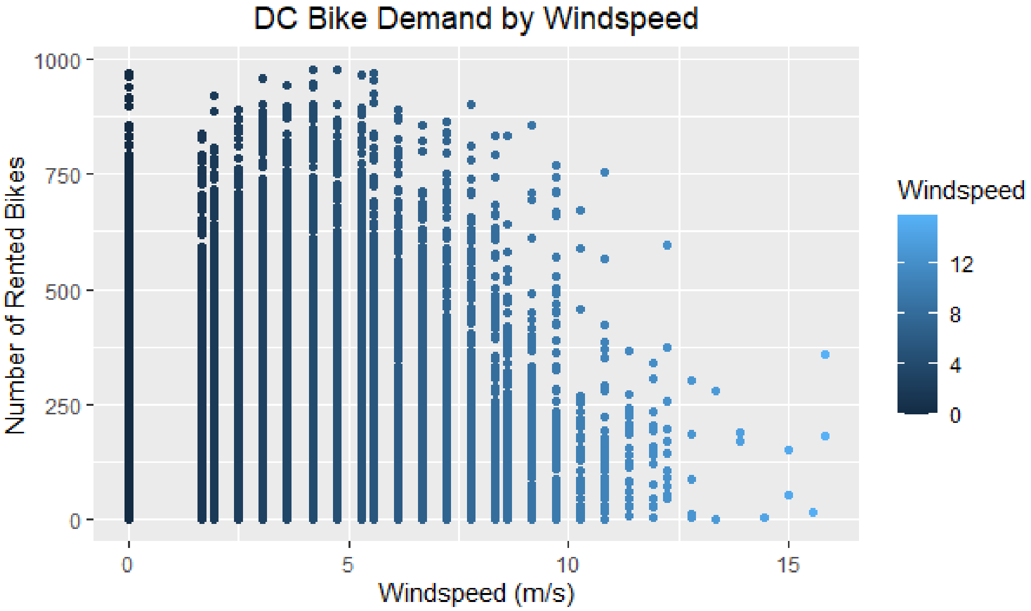


The humidity variable shows some association with bike demand, since the greatest numbers of rented bikes were on days of moderate humidity (around 50%) and the numbers fall off either side of this.

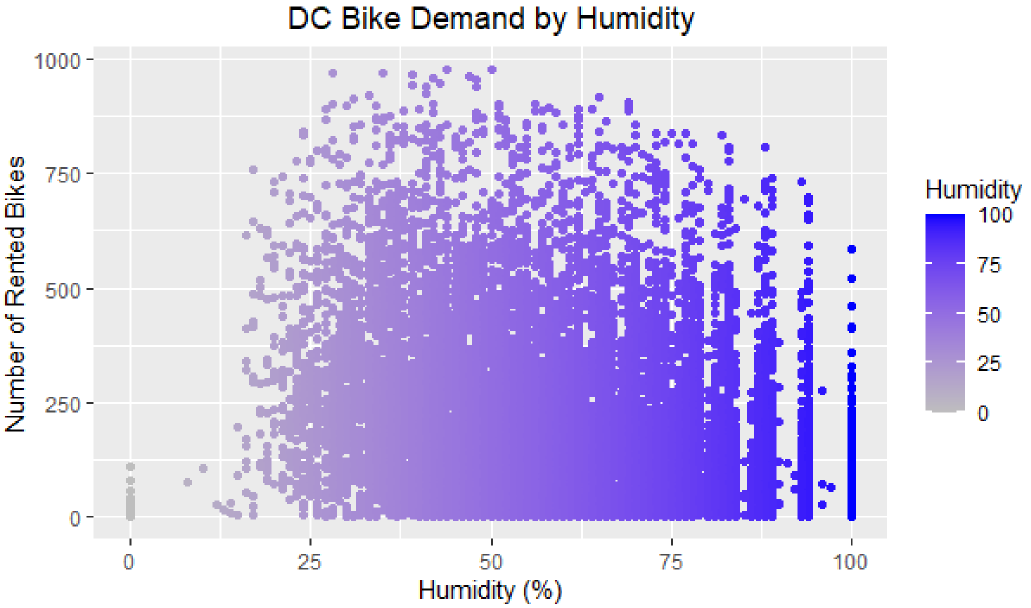
DC:



As in Seoul, bike demand in DC tends to be higher with increasing temperature, up to about 25°C, after which demand declines slightly. The left-skewness of the temperature data suggests positive association between temperature and bike demand.



As windspeed increases, the maximum number of bikes rented for a given windspeed seems to decline at speeds greater than 5m/s. Although there are relatively few data points for windspeeds around 15m/s, they all represent days of low numbers of rented bikes. DC has a greater range of windspeeds than does Seoul, and this can perhaps explain the slightly stronger effect of windspeed on daily bike demand.



The bike demand by humidity graph shows not much variation in the response over the majority of the range of humidities, i.e. from around 20% to 100%. The highest number of bikes are rented on days with moderate to low humidity. There does, however, seem to be a slight decline in the maximum number of bikes rented for given humidities as the humidity rises to 100%. As such, there appears to be weak to moderate association between the two variables.

Overall, there appears to be a distinct yet moderate effect of the three meteorological variables on bike demand across the two cities.

STATISTICAL MODELLING

Fit a linear model with log count as outcome, and season, air temperature, humidity and wind speed as predictors. Print out a summary of the fitted models, comment on the results and compare across the two cities.

Seoul:

```
> summary(Seoulfit)
Call:
lm(formula = log(Count) ~ Season + Temperature + Humidity + WindSpeed,
    data = BikeSeoul)
Residuals:
    Min      1Q  Median      3Q     Max
-5.1073 -0.4281  0.0812  0.5493  2.4352
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7336965  0.0467062 144.171  < 2e-16 ***
SeasonSummer  0.0036038  0.0327843   0.110  0.91247
SeasonAutumn  0.3733211  0.0261578  14.272  < 2e-16 ***
SeasonWinter -0.3830362  0.0349918 -10.946  < 2e-16 ***
Temperature   0.0492700  0.0015053  32.732  < 2e-16 ***
Humidity     -0.0224974  0.0004844 -46.441  < 2e-16 ***
WindSpeed     0.0253809  0.0093544   2.713  0.00668 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.8276 on 8458 degrees of freedom
Multiple R-squared:  0.4941, Adjusted R-squared:  0.4937
F-statistic:  1377 on 6 and 8458 DF,  p-value: < 2.2e-16
```

DC:

```
> summary(DCfit)
Call:
lm(formula = log(Count) ~ Season + Temperature + Humidity + WindSpeed,
    data = BikeDC)
Residuals:
    Min      1Q  Median      3Q     Max
-5.4834 -0.6069  0.2458  0.8440  3.5203
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.6264010  0.0576892  80.195  < 2e-16 ***
SeasonSummer -0.3651680  0.0300276 -12.161  < 2e-16 ***
SeasonAutumn  0.5361839  0.0289332  18.532  < 2e-16 ***
SeasonWinter  0.1046103  0.0341346   3.065  0.00218 **
Temperature   0.0797914  0.0017401  45.856  < 2e-16 ***
Humidity     -0.0233425  0.0005317 -43.901  < 2e-16 ***
WindSpeed     0.0245022  0.0044358   5.524 3.37e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.263 on 17372 degrees of freedom
Multiple R-squared:  0.278,  Adjusted R-squared:  0.2777
F-statistic:  1115 on 6 and 17372 DF,  p-value: < 2.2e-16
```

In the model for Seoul, all regression coefficients except `SeasonSummer` are significant at the 5% level. In the model for DC, all regression coefficients are significant at this level. The adjusted R-squared values are somewhat different, with Seoul’s model reporting a moderate 0.49 and DC’s reporting a low 0.28. This suggests that the Seoul model can explain more of the variability in its data than the DC model can. The p-values of the *F*-statistics given in the summaries are both highly significant, suggesting that both models provide a fit better than one with no independent variables.

Display the 97% confidence intervals for the estimated regression coefficients. Do you think these confidence intervals are reliable?

Seoul:

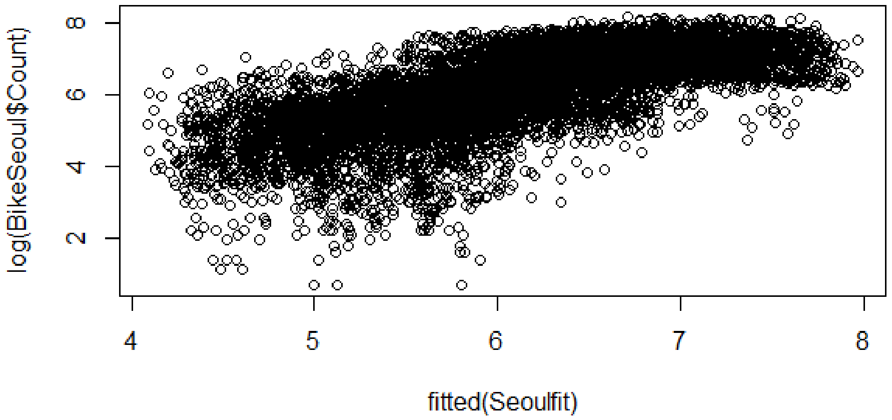
```
                    1.5 %      98.5 %
(Intercept)   6.632322686  6.83507030
SeasonSummer -0.067553139  0.07476072
SeasonAutumn  0.316546593  0.43009553
SeasonWinter -0.458984431 -0.30708797
Temperature   0.046002904  0.05253719
Humidity     -0.023548780 -0.02144592
Windspeed     0.005077663  0.04568421
```

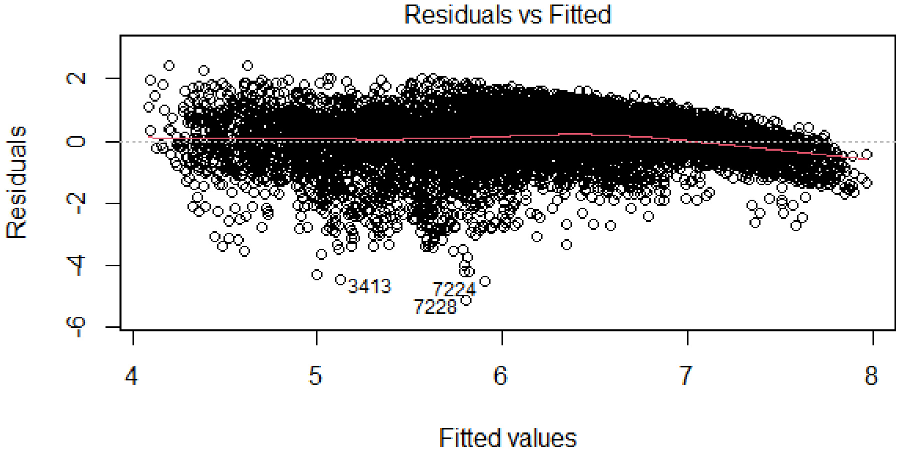
DC:

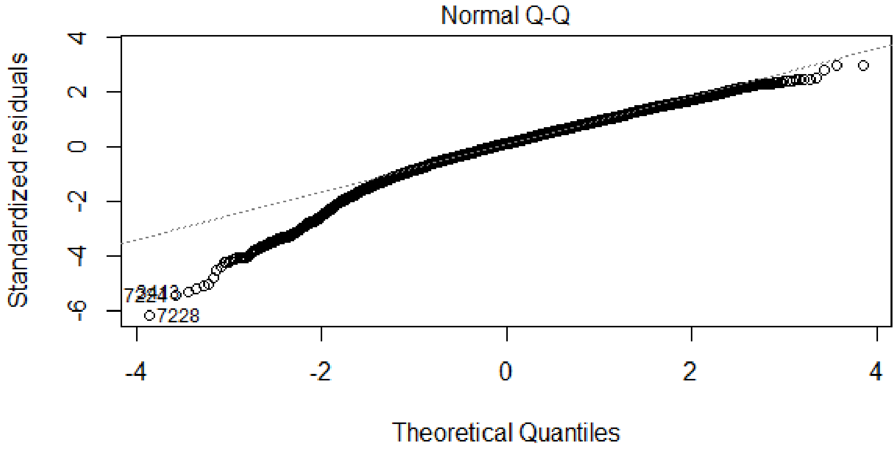
```
(Intercept)   4.50119998  4.75160198
SeasonSummer -0.43033590 -0.30000019
SeasonAutumn  0.47339115  0.59897666
SeasonWinter  0.03052896  0.17869159
Temperature   0.07601506  0.08356781
Humidity     -0.02449639 -0.02218851
Windspeed     0.01487540  0.03412904
```

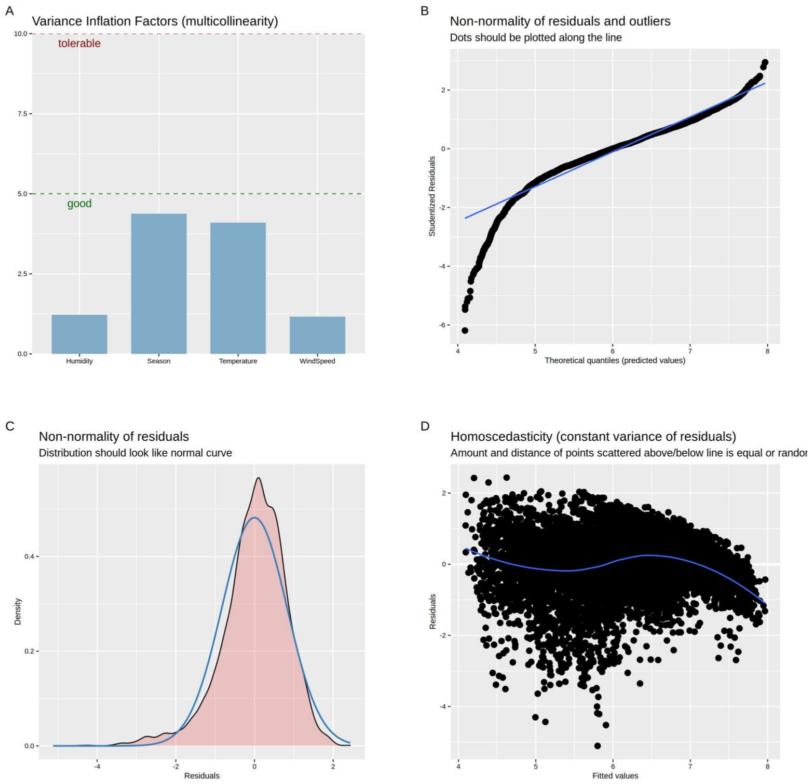
To ascertain whether these results are reliable, we must test the model assumptions.

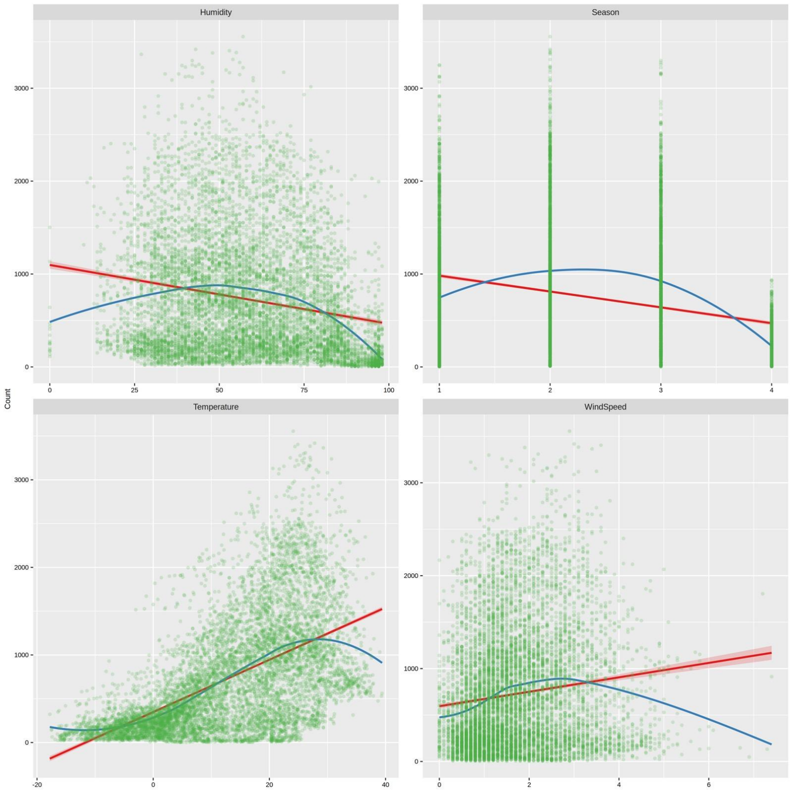
Seoul:


log(BikeSeoul$Count)
2
4
6
8
4
5
6
7
8
fitted(Seoulfit)


Residuals vs Fitted
Residuals
-6
-4
-2
0
2
3413
7224
7228
4
5
6
7
8
Fitted values

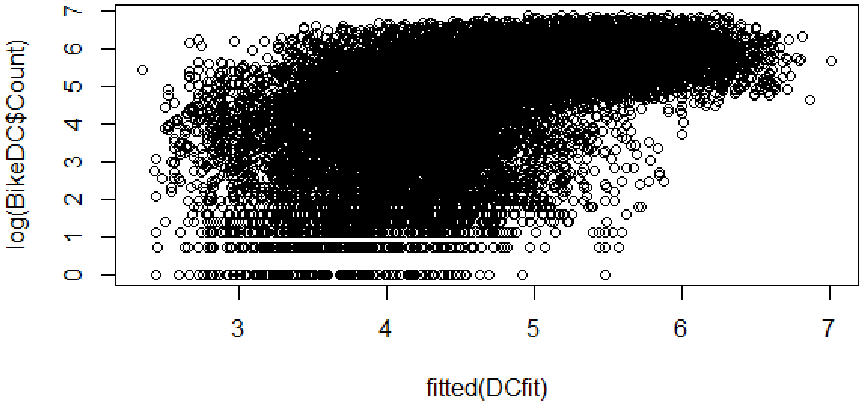

Normal Q-Q
Standardized residuals
-6
-4
-2
0
2
4
7224
3413
7228
-4
-2
0
2
4
Theoretical Quantiles

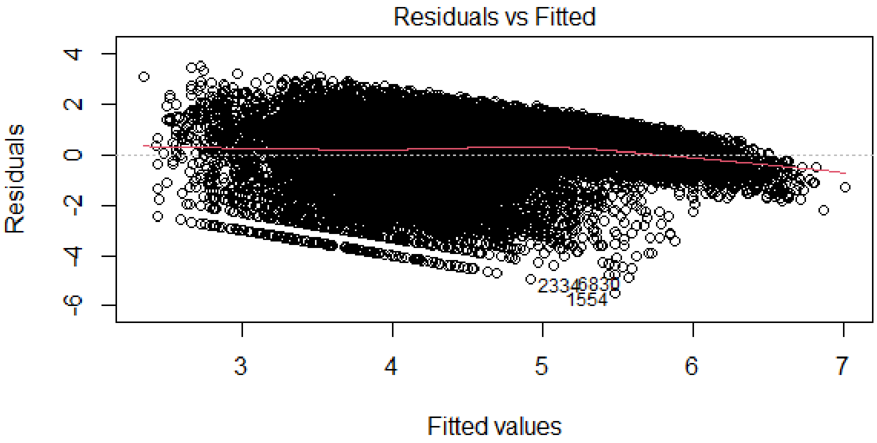

A
Variance Inflation Factors (multicollinearity)
tolerable
good
10.0
7.5
5.0
2.5
0.0
Humidity
Season
Temperature
WindSpeed
B
Non-normality of residuals and outliers
Dots should be plotted along the line
Studentized Residuals
2
0
-2
-4
-6
4
5
6
7
8
Theoretical quantiles (predicted values)
C
Non-normality of residuals
Distribution should look like normal curve
Density
0.4
0.2
0.0
-4
-2
0
2
Residuals
D
Homoscedasticity (constant variance of residuals)
Amount and distance of points scattered above/below line is equal or randor
Residuals
2
0
-2
-4
4
5
6
7
8
Fitted values

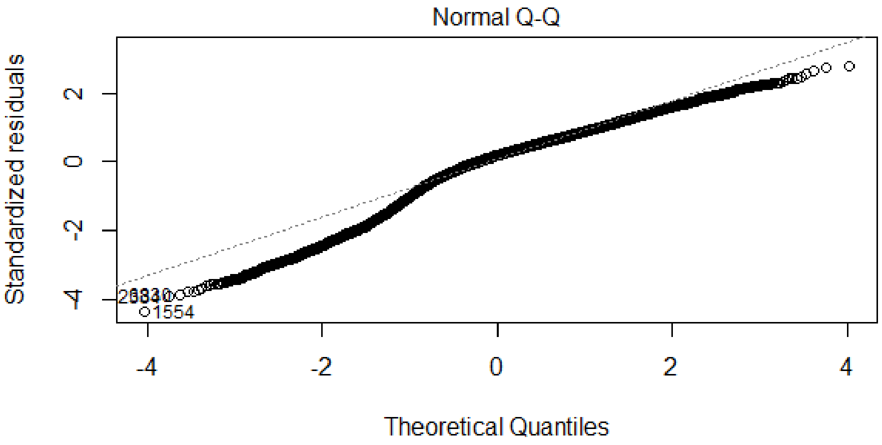

Humidity
Season
Temperature
WindSpeed
Count
3000
2000
1000
0
0
25
50
75
100
1
2
3
4
-20
0
20
40
0
2
4
6

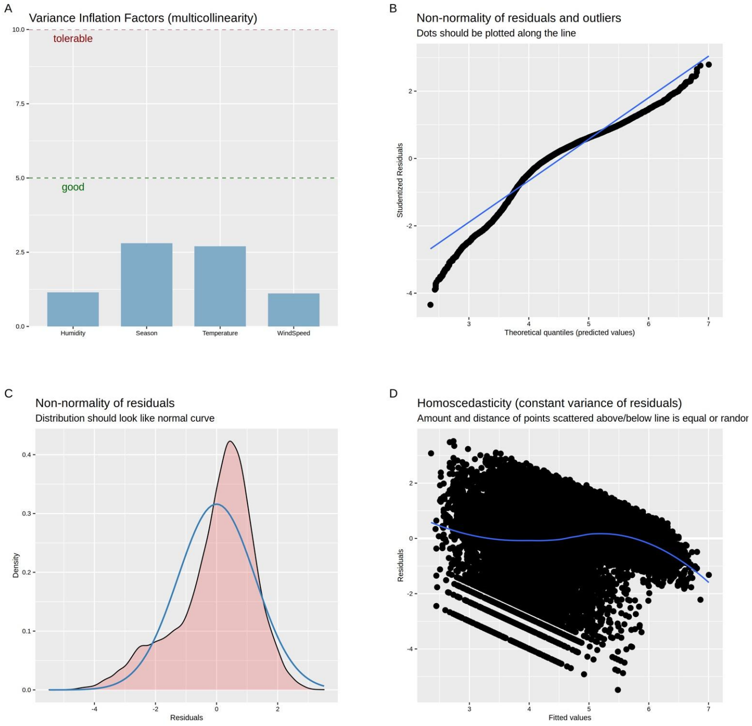
From the plot of the fitted versus observed values, it can be seen there may not be a linear relationship between the explanatory variables and the response, meaning the linearity assumption may not be met. The explanatory variables do not seem to be linearly related to the response. However, there appears to be no issue with multicollinearity and the residuals of the model for Seoul appear to be fairly homoscedastic, although the residuals do display a noticeable but weak pattern. The residuals also appear to be fairly Normally distributed, but with some deviation from Normal at lower predicted values.

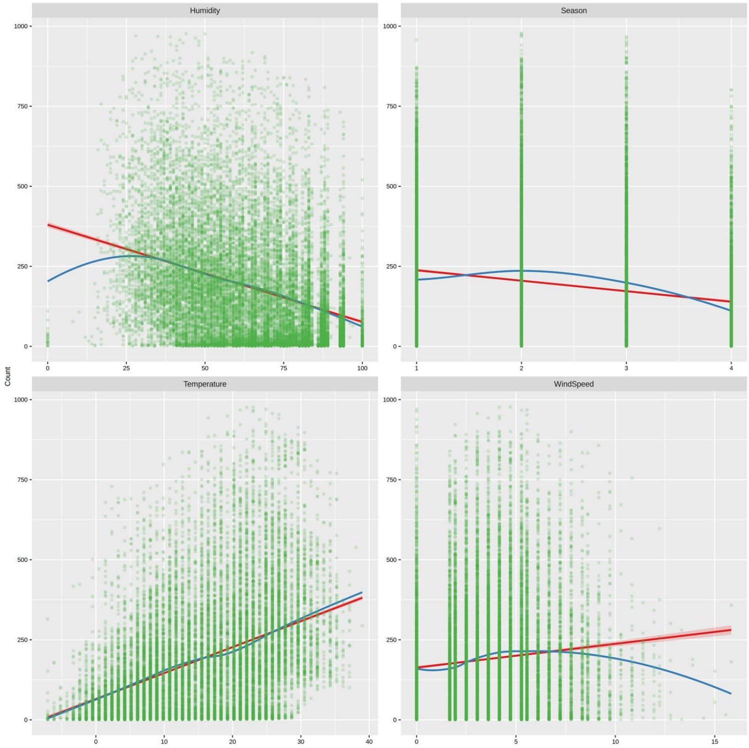
DC:






Normal Q-Q
Standardized residuals
Theoretical Quantiles
-4
-2
0
2
4
1554


A
Variance Inflation Factors (multicollinearity)
tolerable
good
10.0
7.5
5.0
2.5
0.0
Humidity
Season
Temperature
WindSpeed
B
Non-normality of residuals and outliers
Dots should be plotted along the line
Studentized Residuals
Theoretical quantiles (predicted values)
C
Non-normality of residuals
Distribution should look like normal curve
Density
Residuals
0.4
0.3
0.2
0.1
0.0
D
Homoscedasticity (constant variance of residuals)
Amount and distance of points scattered above/below line is equal or randor
Residuals
Fitted values



Based on the plots for DC, the model assumptions appear to be somewhat violated. The fitted and observed values do not show a linear relationship. The relationships between each explanatory variable and the response also do not appear linear. The residuals also appear to decrease with increasing fitted value. As with Seoul, the residuals also appear to violate the Normality assumption at low theoretical quantiles. Despite this, there appears to be no multicollinearity.

Based on all this, the calculated confidence intervals may not be reliable.

Assuming the model is trustworthy, what's the expected number of rented bikes in winter when the air temperature is freezing (0°C), in the presence of light wind (0.5 m/s) and a humidity of 20%. Provide the 90% prediction intervals and comment on the results.

Seoul:

```
fit      lwr      upr

369.9633 94.74608 1444.628
```

DC:

```
fit      lwr      upr

71.9818  9.006506 575.293
```

The model for Seoul predicts an expected number of rented bikes of 370 (to the nearest integer). The model for DC predicts an expected number of rented bikes of 72, and significantly less than the prediction for Seoul.