Assignment 1

MT5763: Software for Data Analysis

Find the R code at https://github.com/erna1997/MT5763_1_22001330.git

Erna Kuginyte
Student number: 220013309

1. Data Wrangling

The Data Wrangling section has been fully completed in the MT5763_220013309.r script, which is uploaded on the private GitHub repository (link can be found above). This part has been fully executed and now the Seoul and Washington data frames have the same set of consistently named and comparable columns.

2. Data Visualisation

2.1 Seoul and Washington DC Temperature
2.2 Seoul and Washington DC Bikes Rented by Season
2.3 Bikes Rented on Weekdays VS Holidays
2.4 Bikes Rented by Time of Day
2.5a Bike Demand and Meteorological Variables. Seoul
2.5b Bike Demand and Meteorological Variables. Washington DC

3. Statistical Modelling

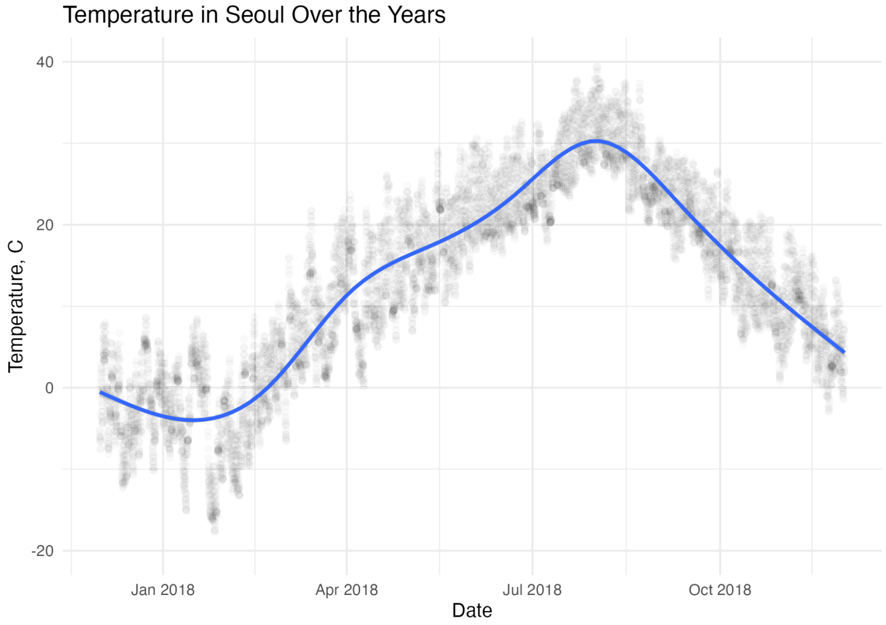
3.1a Linear Model Summary. Seoul
3.1b Linear Model Summary. Washington DC
3.2a Confidence Intervals for Estimated Regression Coefficients. Seoul
3.2b Confidence Intervals for Estimated Regression Coefficients. Washington DC
3.3a Expected Number of Rented Bikes. Seoul
3.3b Expected Number of Rented Bikes. Washington DC

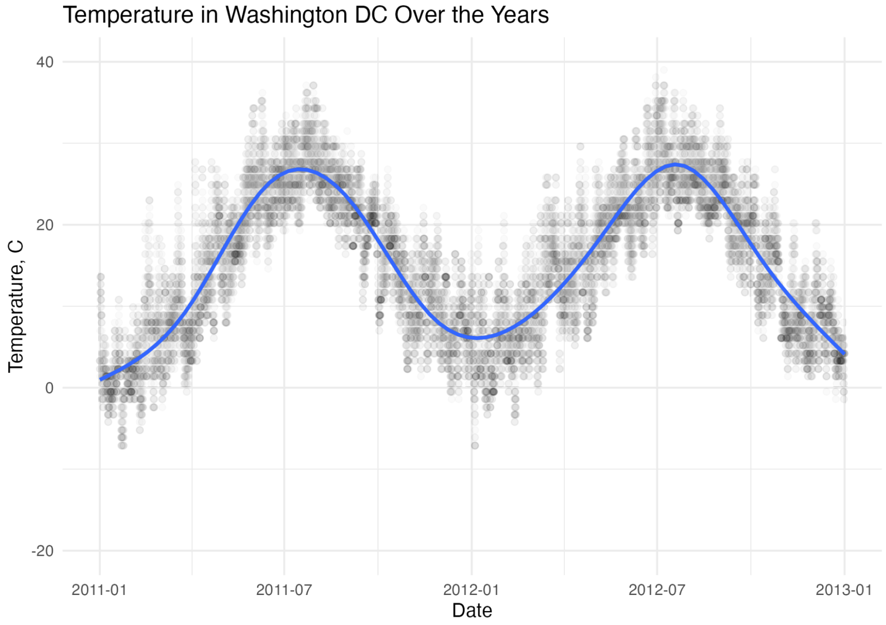
4. References

University of St Andrews

2. Data Visualisation

2.1 Seoul and Washington DC Temperature

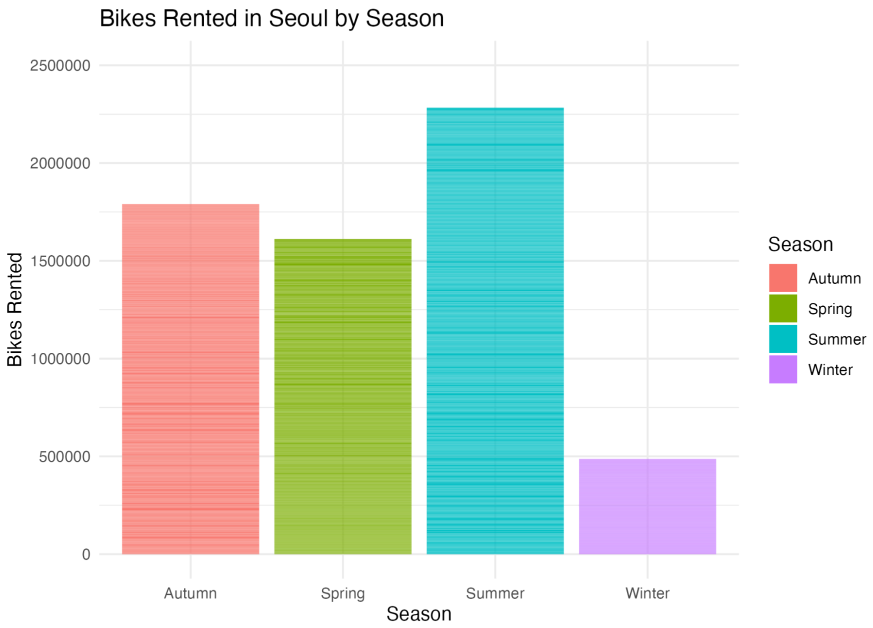


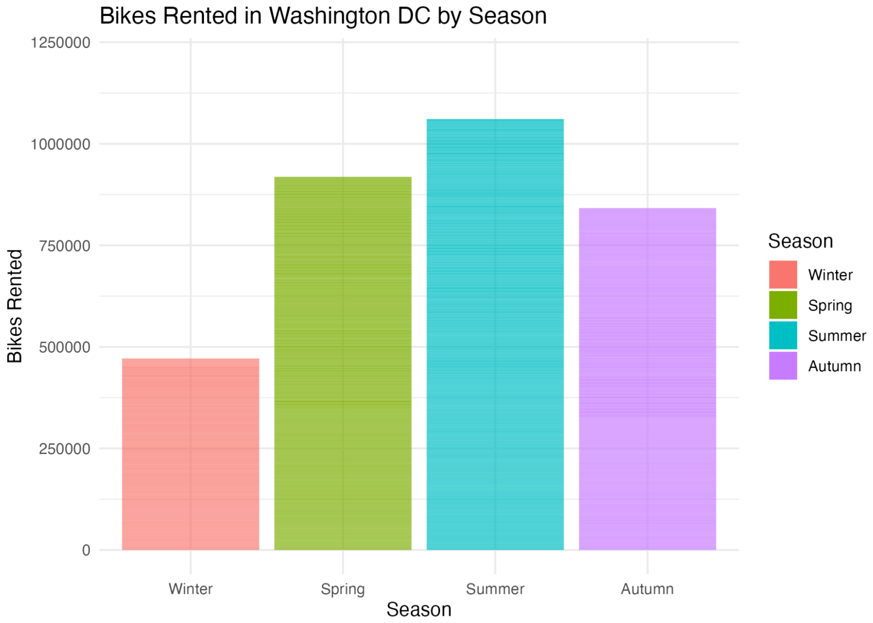


We can see that the amplitudes of the temperatures in Seoul and Washington DC are slightly different. Looking at the first graph for temperature in Seoul Over 2017 until the end of 2018, we can see that the temperature drops to around -4 degrees Celsius, and reaches as high as 31 degrees Celsius. The second graph - Temperature in Washington DC Over shows that the temperature amplitude from 2011 until 2013 was from around 1 to 27 degrees Celsius.

Noticeably, the temperature in April 2018 in Seoul had a little spike, whereas the temperature in Washington DC would increase more steadily from Winter until Summer time.

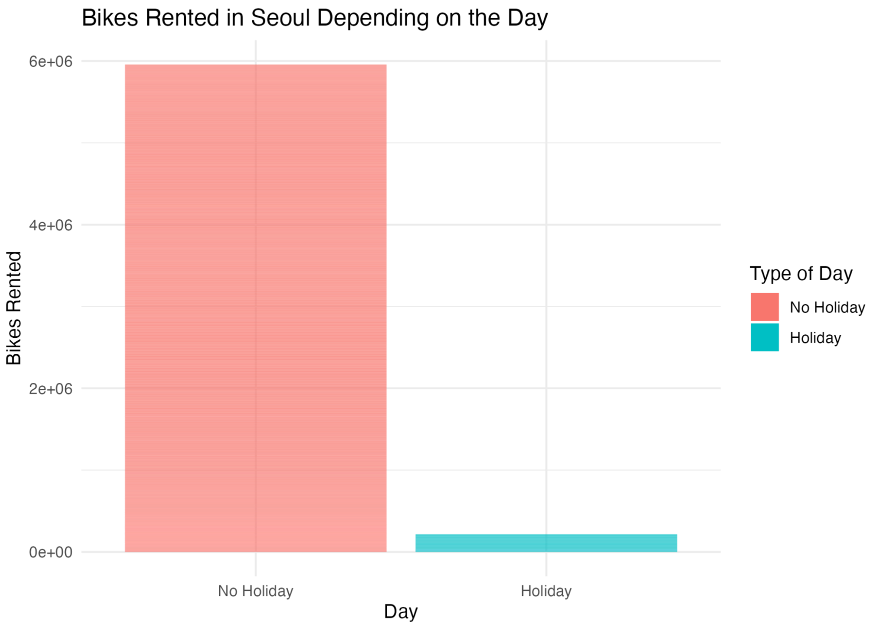
2.2 Seoul and Washington DC Bikes Rented by Season

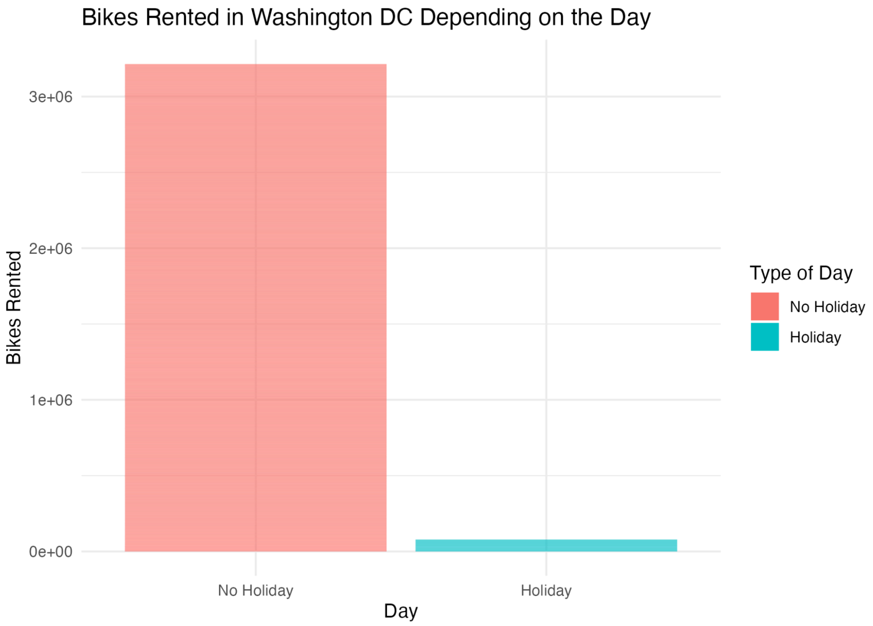




Seasons really affect the number of rented bikes in both cities. Summer time seems to be the peak time for bike renting across both cities, and the bike rental during Winter really dips.
Seoul rents out less than a quarter of bikes during Winter time compared to Summer, whil Washington DC rents out half the number of bikes during Winter compared to Summer.

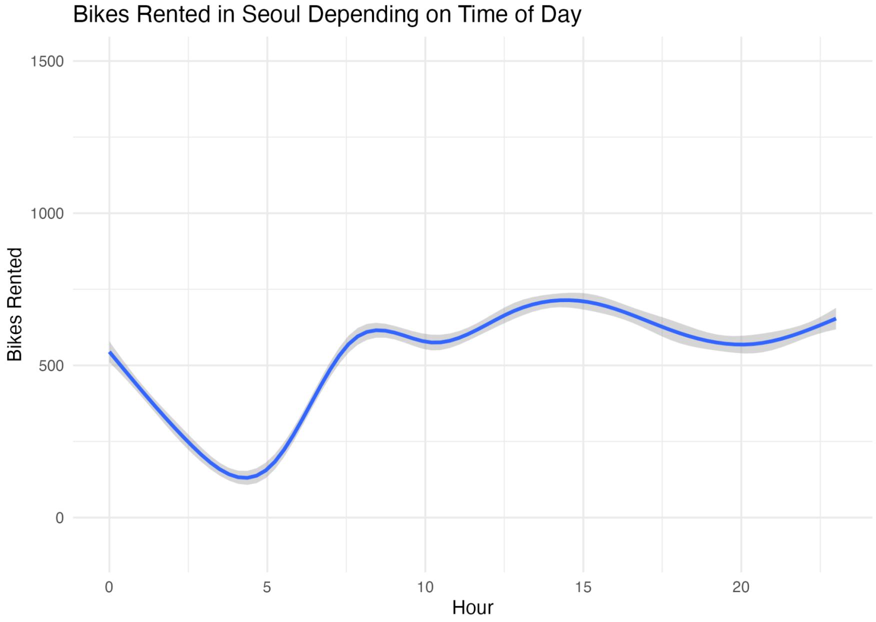
2.3 Bikes Rented on Weekdays VS Holidays

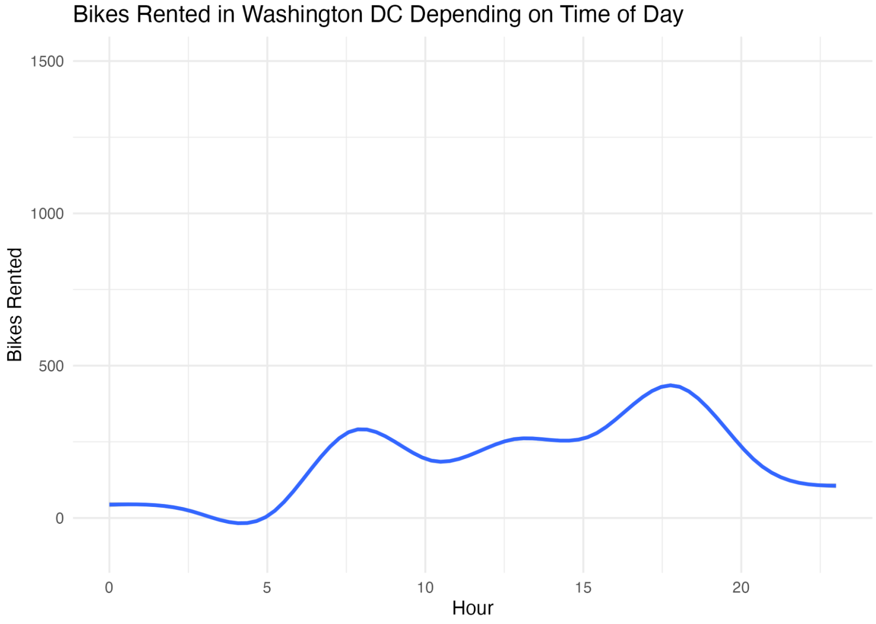

Bikes Rented in Seoul Depending on the Day


Bikes Rented in Washington DC Depending on the Day

There seems to be a massive association between the number of bikes rented across both cities depending whether it is a weekday or a holiday. People don't seem to be keen to rent bikes during holidays. The ratio of rented bikes during a weekday vs a holiday looks like around 40:1 in Seoul and 32:1 in Washington DC.

2.4 Bikes Rented by Time of Day



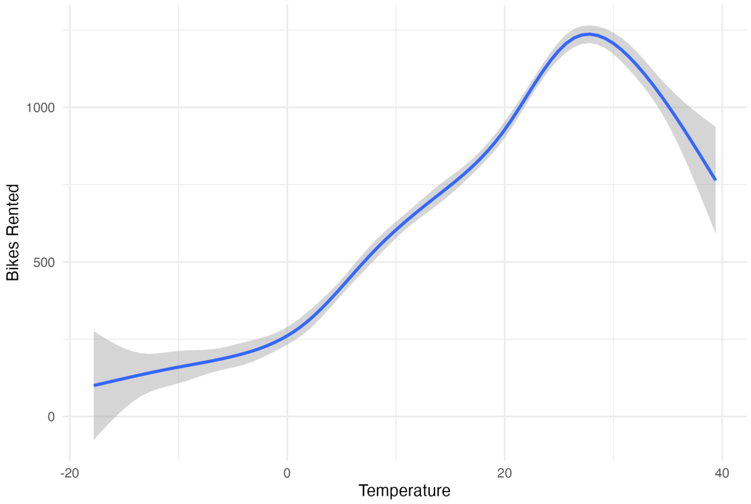


We can see from the Seoul graph that the most popular time to rent a bike is around 1-2pm. There are three spikes overall: around 9am, 1-2pm and 11pm-12am. There are very few bikes rented around 3-5am.

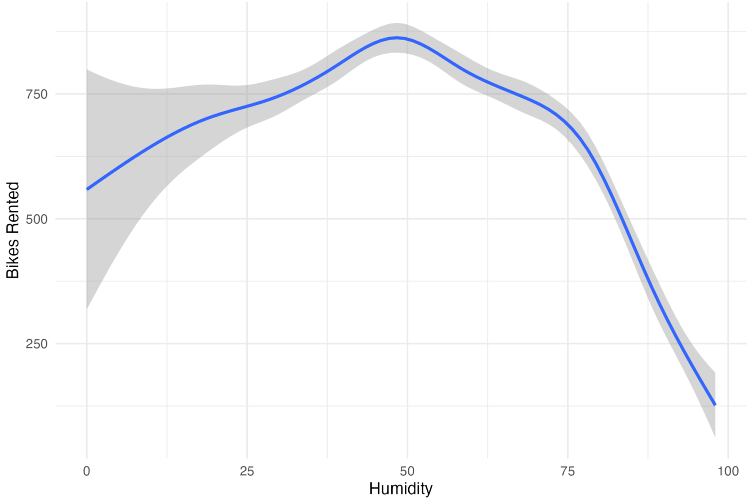
The case of Washington DC also has three spikes overall: around 8am, 12am-1pm, 5-6pm. The most popular time is 5-6pm. There are very few rented bikes around 9pm-5am.

2.5a Bike Demand and Meteorological Variables. Seoul

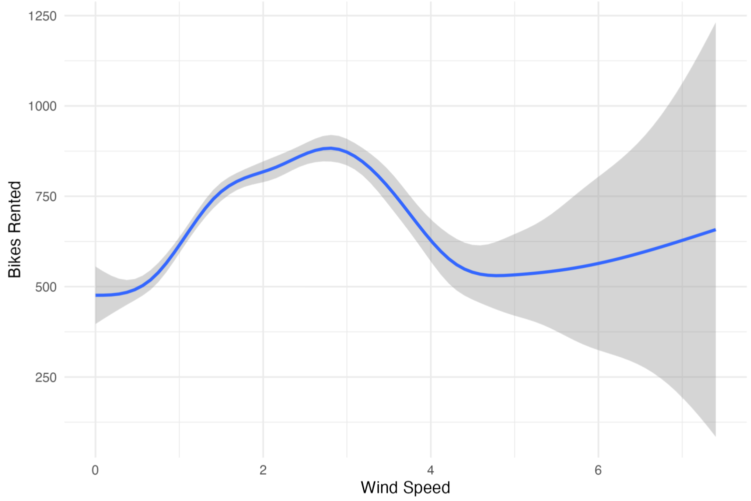
Bikes Rented in Seoul Depending on Temperature



Bikes Rented in Seoul Depending on Humidity



Bikes Rented in Seoul Depending on Wind Speed

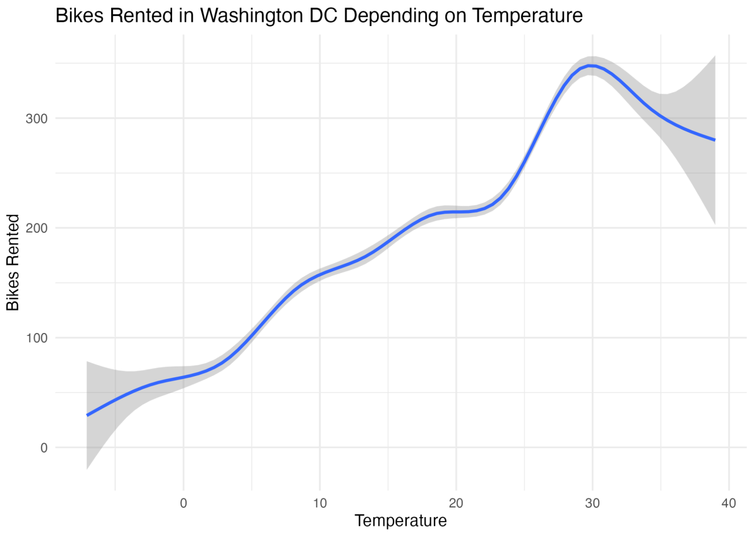


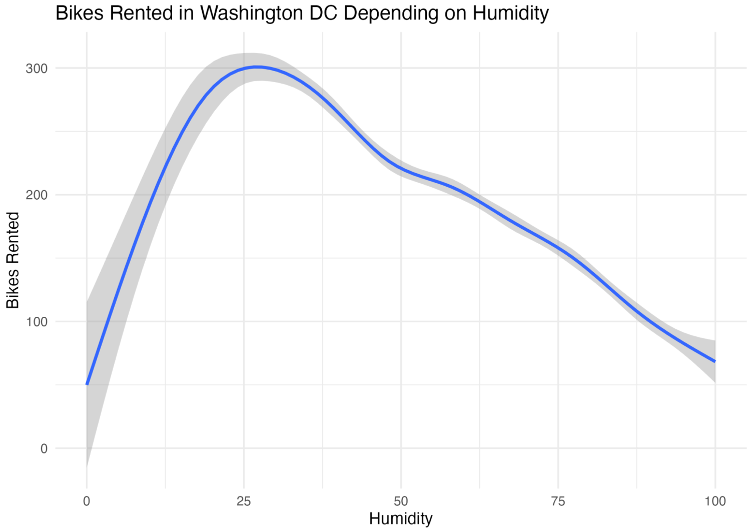
There seems to be a high dependency between the number of bikes rented and the air temperature in Seoul. The peak temperature is around 27 degrees Celsius.

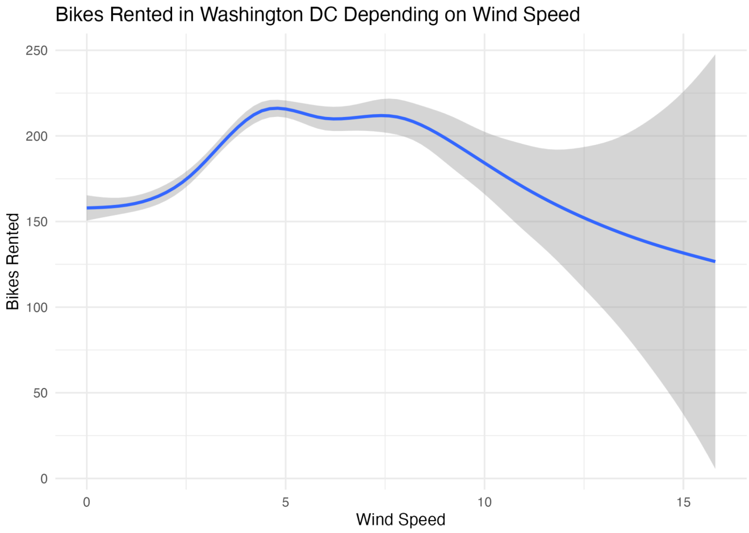
Humidity also seems to have a correlation with the number of bikes rented. The peak is at around 50% humidity. The count really drops at 100% humidity (most likely when it’s raining/snowing).

The wind speed seems to have an association with the bike count: the peak is at around 3m/s, the drops are at 0-1m/s, and 4-5m/s. The data points are a lot more spread out at the higher wind speed (from 5m/s onwards), which might mean that there is another factor that plays out in the count, it could be temperature or humidity for example.

2.5b Bike Demand and Meteorological Variables. Washington DC







There seems to be a high dependency between the number of bikes rented and the air temperature in Washington DC as well. The peak temperature is around 29-30 degrees Celsius.

Humidity seems to have a correlation with the number of bikes rented, but it differs from Seoul: the peak is at around 25-30% humidity. The count really drops at 0% and 100% humidity (most likely when it’s raining/snowing).

Wind speed seems to also have an association with the bike number: the peaks are at 4-5m/s and 7-8m/s. The drop of the count starts at 8-9m/s.

It might be interesting to explore the effects of humidity, wind speed and temperature all together associated with the count of bikes rented by using correlation coefficients.

3. Data Visualisation

3.1a Linear Model Summary. Seoul

Call

lm(formula = count ~ season + temperature + humidity + wind_speed, data = bike_seoul)

Residuals

Min	***1Q***	***Median***	***3Q***	***Max***
-1415.42	*-305.22*	*-54.27*	*197.71*	*2465.39*

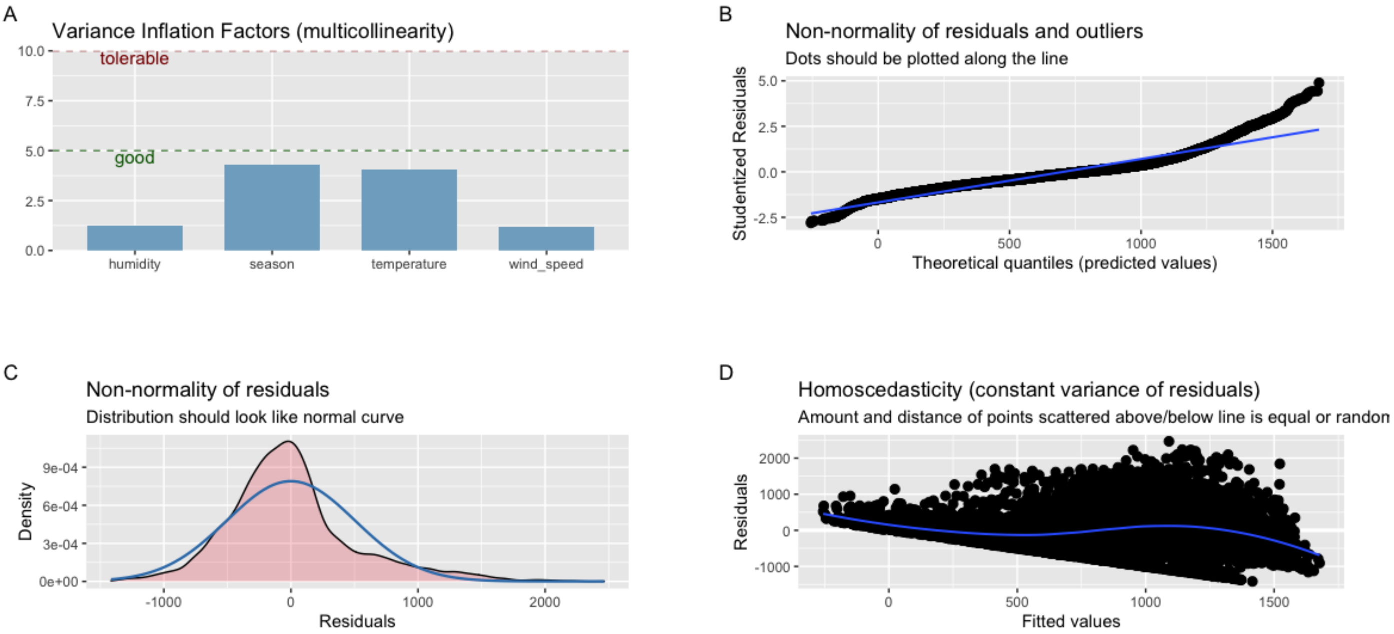
Coefficients

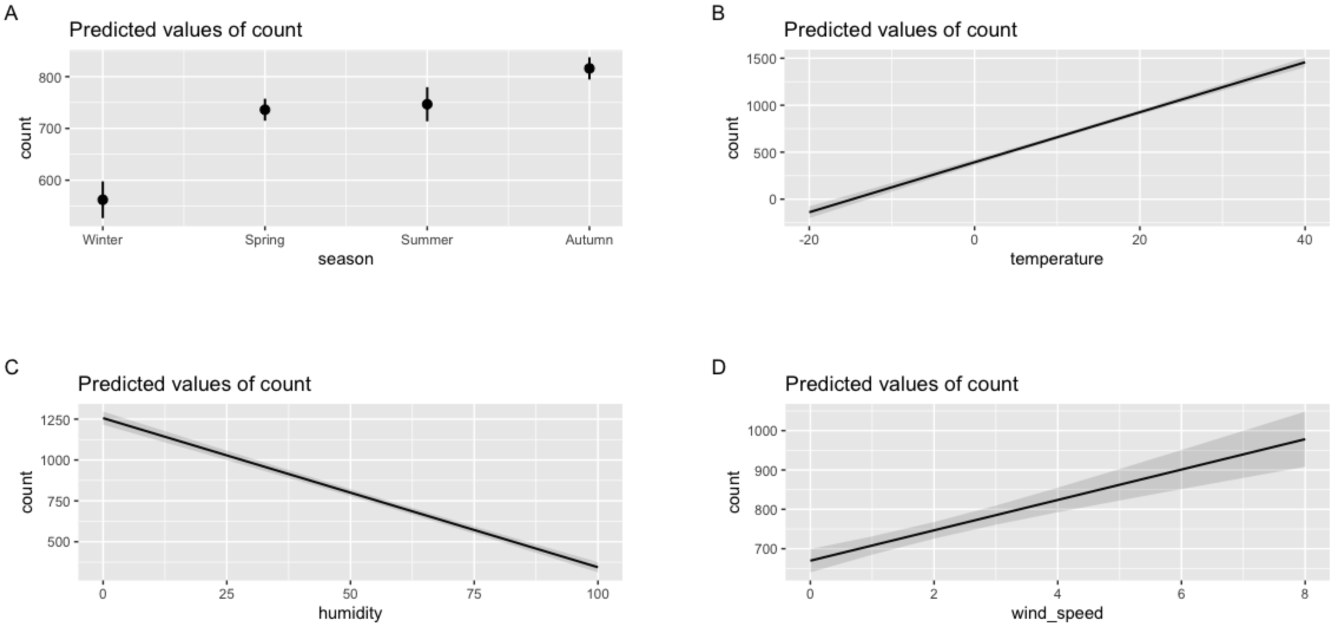
	Estimate	***Std. Error***	***t value***	***Pr(>|t|)***
(Intercept)	*926.4531*	*27.7956*	*33.331*	*< 2e-16 ****
seasonSpring	*-79.7534*	*15.4250*	*-5.170*	*2.39e-07 ****
seasonSummer	*-69.1772*	*19.1634*	*-3.610*	*0.000308 ****
seasonAutumn	*-253.6003*	*22.1001*	*-11.475*	*< 2e-16 ****
temperature	*26.6184*	*0.9081*	*29.314*	*< 2e-16 ****
humidity	*-9.1240*	*0.2923*	*-31.216*	*< 2e-16 ****
wind_speed	*38.5847*	*5.6030*	*6.886*	*6.11e-12 ****

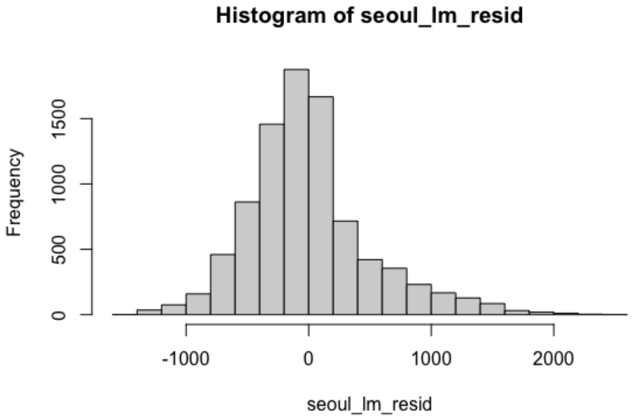
Signif. codes: 0 ‘*’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

Residual standard error:	*505 on 8753 degrees of freedom*
Multiple R-squared:	*0.3875*
Adjusted R-squared:	*0.3871*
F-statistic:	*923 on 6 and 8753 DF*
p-value	*< 2.2e-16*

Above we have the summary of the linear model with log count as outcome, and season, air temperature, humidity and wind speed as predictors.

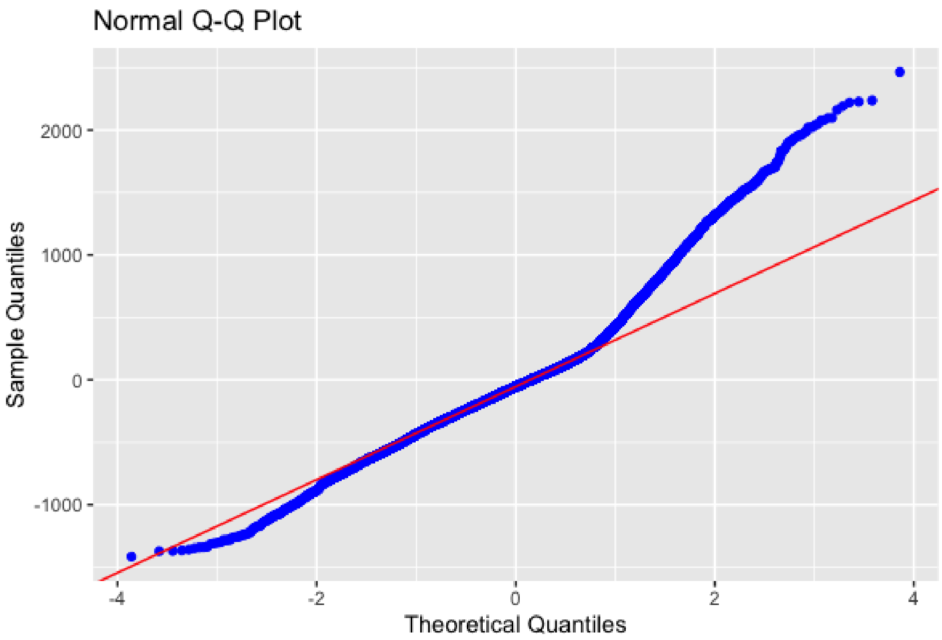

A
Variance Inflation Factors (multicollinearity)
10.0
7.5
5.0
2.5
0.0
tolerable
good
humidity
season
temperature
wind_speed
B
Non-normality of residuals and outliers
Dots should be plotted along the line
Studentized Residuals
5.0
2.5
0.0
-2.5
0
500
1000
1500
Theoretical quantiles (predicted values)
C
Non-normality of residuals
Distribution should look like normal curve
Density
9e-04
6e-04
3e-04
0e+00
-1000
0
1000
2000
Residuals
D
Homoscedasticity (constant variance of residuals)
Amount and distance of points scattered above/below line is equal or randon
Residuals
2000
1000
0
-1000
0
500
1000
1500
Fitted values


A
Predicted values of count
count
800
700
600
Winter
Spring
Summer
Autumn
season
B
Predicted values of count
count
1500
1000
500
0
-20
0
20
40
temperature
C
Predicted values of count
count
1250
1000
750
500
0
25
50
75
100
humidity
D
Predicted values of count
count
1000
900
800
700
0
2
4
6
8
wind_speed


Histogram of seoul_lm_resid
Frequency
1500
1000
500
0
-1000
0
1000
2000
seoul_lm_resid

Above we can see the graphs of noise and signal model diagnostics, the histogram of residuals (that looks relatively normal), it helps to visualise the outliers and their distance from the normal distribution.

The residuals from median to 1Q and median to 3Q are very symmetrical. However the Min and Max values from the median are not symmetrical, we can see from the plot below that there are values that don't follow the normality line.[1]



The **intercept estimate** at 926.45 could potentially be not very meaningful since all the factors would be at 0 (temperature, wind speed and humidity).[2]

The **Pr(>t)** values are very small and indicate a strong relationship between the values and the Bike Count outcome.

Residual standard error shows the quality of a linear regression fit and in this case it's rather large (505 / 926,4531 = 54.5%) and indicates that the model is not very well fitted.[3]

Multiple R-squared 0.3875 and **adjusted R-squared** 0.3871, so roughly 38.7% of the variance found in the response variable Count can be explained by the predictor variables. There might be other factors that have stronger association with bike Count that have not been taken into consideration when modelling lm.[4]

[1] (Quick Guide: Interpreting Simple Linear Model Output in R, 2022)
[2] (Zach, 2022)
[3] (Zach, 2022)
[4] (Quick Guide: Interpreting Simple Linear Model Output in R, 2022)

F-statistic 923 on 6 and 8753 DF - F-statistic is very large, so there should be strong association between predictors and the response variable.

p-value < 2.2e-16 - is very small, meaning there is a great statistical significance between the predictors and the response variable.

3.1b Linear Model Summary. Washington DC

Call

lm(formula = count ~ season + temperature + humidity + wind_speed, data = bike_washington_DC)

Residuals

Min	1Q	Median	3Q	Max
-372.07	-102.58	-30.38	64.24	731.20

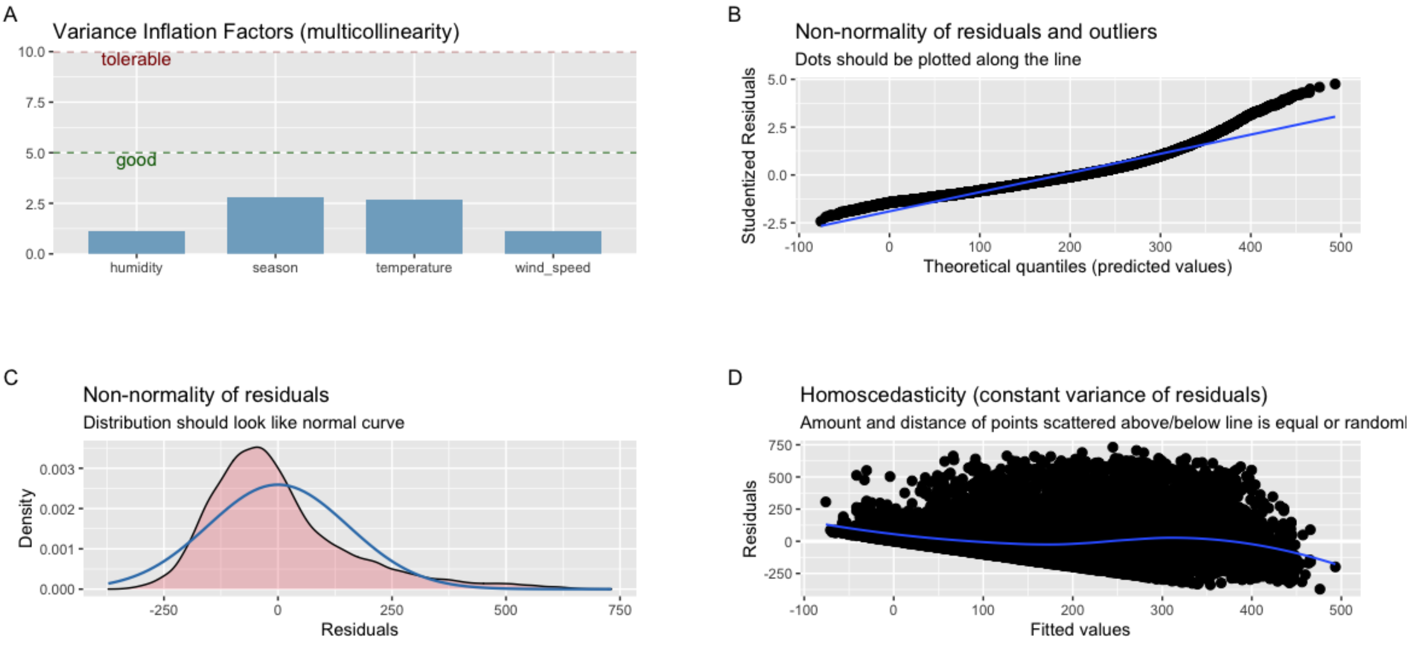
Coefficients

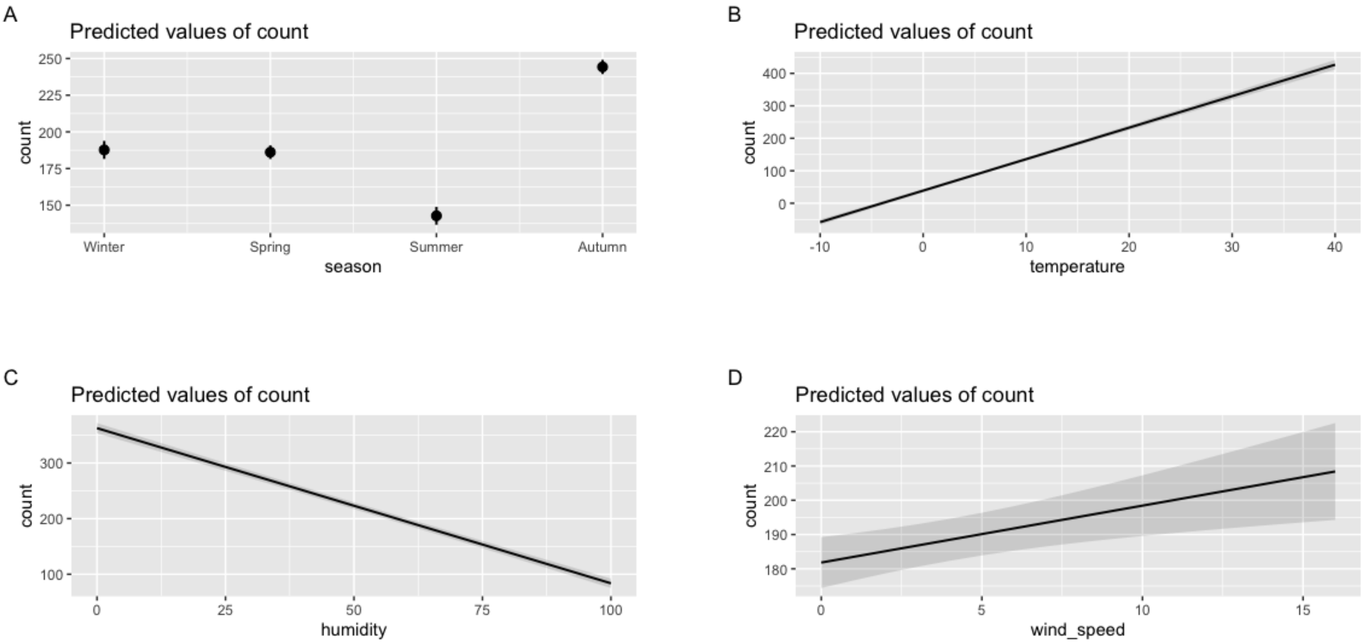
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	207.92546	5.65747	36.752	< 2e-16 ***
seasonSpring	-1.56229	4.15791	-0.376	0.70712
seasonSummer	-44.91728	5.30285	-8.470	< 2e-16 ***
seasonAutumn	56.54880	3.65380	15.477	< 2e-16 ***
temperature	9.69754	0.21197	45.750	< 2e-16 ***
humidity	-2.79008	0.06477	-43.077	< 2e-16 ***
wind_speed	1.66116	0.53977	3.078	0.00209 **

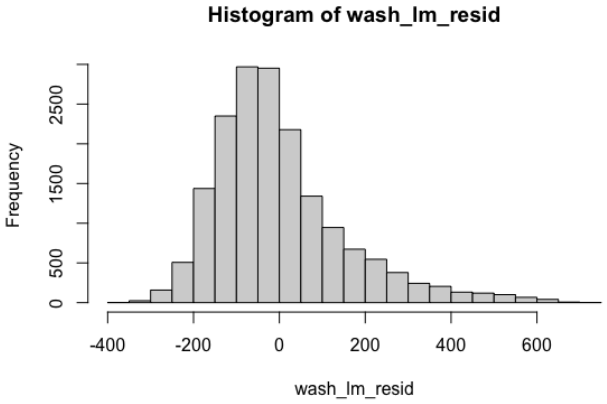
*Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error:	153.8 on 17372 degrees of freedom
Multiple R-squared:	0.2809
Adjusted R-squared:	0.2806
F-statistic:	1131 on 6 and 17372 DF
p-value	< 2.2e-16

Above we have the summary of the linear model with log count as outcome, and season, air temperature, humidity and wind speed as predictors.

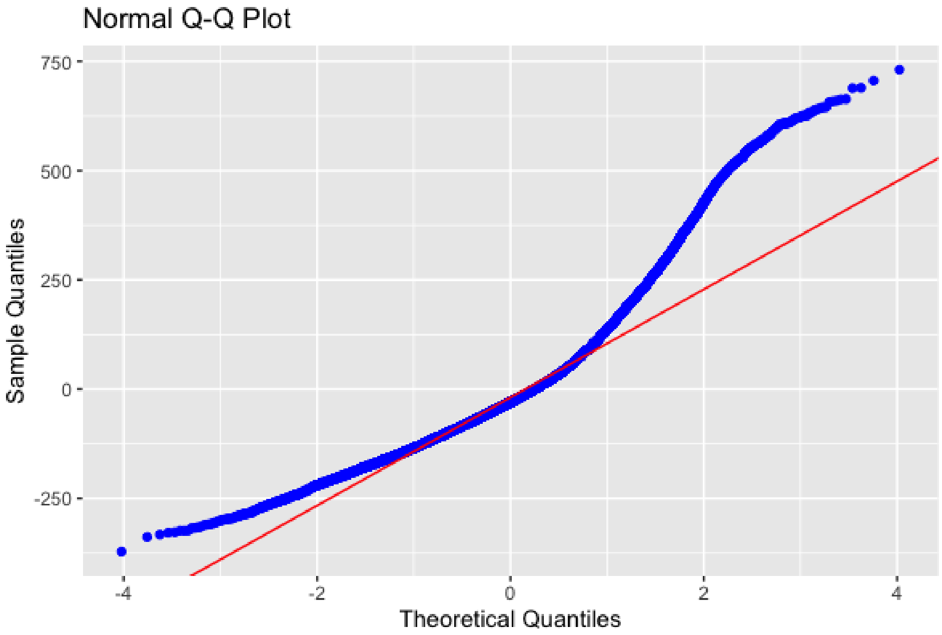






Above we can see the graphs of noise and signal model diagnostics, the histogram of residuals (that looks ever so slightly right-skewed), it helps to visualise the outliers and their distance from the normal distribution.

The residuals from median to 1Q and median to 3Q are unsymmetrical and the Min and Max values from the median are very asymmetrical (different case from the Seoul data), we can see from the plot below that there are a lot of residuals that don't follow the normality line, meaning there are a lot of values that are more extreme than expected. [5]



The **intercept estimate** at 207.92546 could potentially be not very meaningful since all the factors would be at 0 (temperature, wind speed and humidity).[6]

Not all of the **Pr(>t)** values are very small (differs from Seoul data), but most of them still indicate a strong relationship between the values and the Bike Count outcome. We could possibly state that the season_spring predictor does not have an effect on the Bike Count outcome. The wind_speed predictor is at 2%, thus we could possibly state that it does not have an effect on the Count log.

Residual standard error shows the quality of a linear regression fit and in this case it's very large (153.8 / 207.925 = 73.97%) and indicates that the model is not very well fitted.[7]

Multiple R-squared 0.2809 and **adjusted R-squared** 0.2806, so roughly 28% of the variance found in the response variable Count can be explained by the predictor variables. There might be

[5] (Quick Guide: Interpreting Simple Linear Model Output in R, 2022)
[6] (Zach, 2022)
[7] (Zach, 2022)

other factors that have stronger association with bike Count that have not been taken into consideration when modelling lm (this case is very similar to Seoul).[8]

F-statistic 1131 on 6 and 17372 DF - F-statistic is very large, so there should be strong association between predictors and the response variable.

p-value < 2.2e-16 - is very small, meaning there is a great statistical significance between the predictors and the response variable.

Comparing across two cities, both models show some drawbacks, but there are strong indications that there are associations between the predictors and the variables.

3.2a 97% Confidence Intervals for Estimated Regression Coefficients. Seoul

	1.5%	***98.5%***
(Intercept)	*866.124343*	*986.781843*
seasonSpring	*-113.232525*	*-46.274327*
seasonSummer	*-110.770376*	*-27.584011*
seasonWinter	*-301.567313*	*-205.633189*
temperature	*24.647535*	*28.589313*
humidity	*-9.758449*	*-8.489643*
wind_speed	*26.423761*	*50.745731*

The width of the confidence interval around the intercept is not very narrow, the margin of error is relatively big, and there is a great range of plausible values.[9] The predictors temperature and humidity have really narrow confidence intervals. However, the CIs for the season predictors are huge, indicating there probably is a big variability of the number of bikes rented. Similar is the case with wind_speed.

The 97% confidence intervals provide greater intervals than let's say a 95% CI would, therefore, in this case we would have more outliers covered in the predictions.

These confidence intervals look reliable.

[8] (Quick Guide: Interpreting Simple Linear Model Output in R, 2022)

[9] (Frost, 2022)

3.2b 97% Confidence Intervals for Estimated Regression Coefficients. Washington DC

	1.5%	***98.5%***
(Intercept)	*195.6472182*	*220.203698*
seasonSpring	*-10.5860809*	*7.461498*
seasonSummer	*-56.4258979*	*-33.408666*
seasonWinter	*48.6190773*	*64.478525*
temperature	*9.2375165*	*10.157569*
humidity	*-2.9306494*	*-2.649513*
wind_speed	*0.4897278*	*2.832602*

The Washington DC case is different to the Seoul one: the width of the confidence interval around the intercept is narrow, thus the margin of error is small, and there is not a big range of plausible values.[10] The narrow values are partly due to the big sample size.

All of the predictors, except from the season_summer, have really narrow confidence intervals.

The narrowness of the intervals could be changed if we wanted to trust that our prediction model would cover the outliers, to, for example, CI of 99%.

The confidence intervals for the Washington DC case do not look very reliable since they are really narrow.

3.3a Expected Number of Rented Bikes. Seoul

fit	***lwr***	***ups***
509.6643	*-321.5013*	*1340.83*

Assuming the model is trustworthy, the expected number of rented bikes in winter when the air temperature is freezing (0 C), in the presence of light wind (0.5m/s) and a humidity of 20% is 509.6643. The 90% prediction intervals [-321.5013, 1340.83]. The result seems slightly high for the 0 C temperature, and the other predictors have most likely increased the number by quite a bit.

[10] (Frost, 2022)

3.3b Expected Number of Rented Bikes. Washington DC

fit	lwr	ups
152.9544	-100.2205	406.1294

Assuming the model is trustworthy, the expected number of rented bikes in winter when the air temperature is freezing (0 C), in the presence of light wind (0.5m/s) and a humidity of 20% is 152.9544. The 90% prediction intervals [-100.2205, 406.1294]. The result seems reasonable align with the collected data.

4. References

1. Feliperego.github.io. 2022. *Quick Guide: Interpreting Simple Linear Model Output in R*. [online] Available at: <https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R> [Accessed 10 October 2022].

2. Zach, V., 2022. *How to Interpret the Intercept in a Regression Model (With Examples)*. [online] Statology. Available at: <https://www.statology.org/intercept-in-regression/> [Accessed 10 October 2022].

3. Frost, J., 2022. *Confidence Intervals: Interpreting, Finding & Formulas*. [online] Statistics By Jim. Available at: <https://statisticsbyjim.com/hypothesis-testing/confidence-interval/> [Accessed 10 October 2022].