https://github.com/NavarroJulia/MT5763_1_220030294.git

Introduction

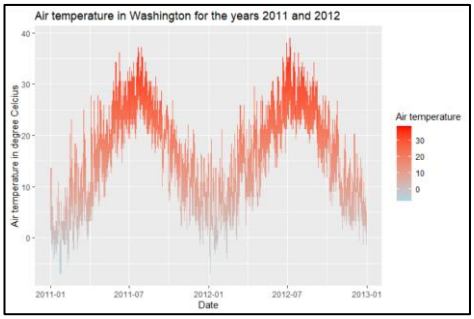
This document contains a brief report and analysis performed in R on two data sets. These data sets contain information that could prove useful in helping the cities of Washington DC and Seoul in timing their bike maintenance more conveniently for the public. Both data frames were cleaned and we ensured that variables names in both sets matched for comparison for further analysis.

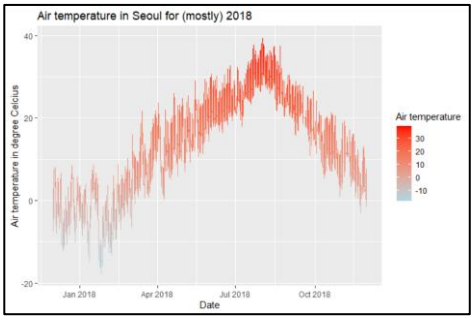
We will present the data both visually and by using a linear model. Of interest is the potential link between the weather and the number of bicycles rented throughout the day.

Main Analysis: Data visualization and statistical modelling

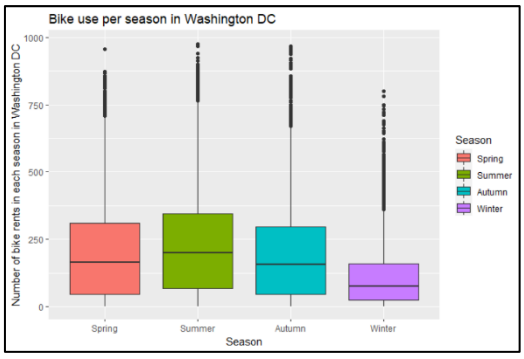
Data visualization

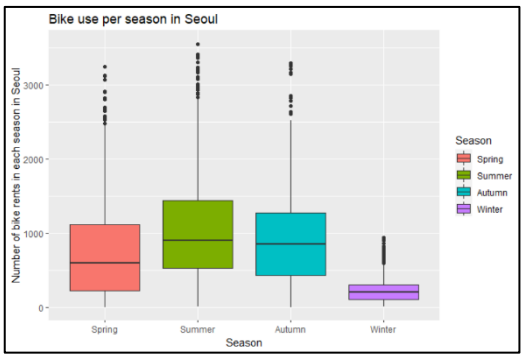
We firstly investigate the fluctuation of air temperature over the year in the two cities and observe how it varies; colder in the winter months and warmer in the summer months. The temperature in Washington DC ranges from (-7.06°C, 39°C) and in Seoul varies from (-17.8°C, 39.4°C). Thus, we naturally conclude that the temperature does differ throughout the seasons and hence over the course of the year in both cities. Note that the plot for Washington DC shows the distribution over two years instead of one and that the temperature differences in winter are more pronounced between the two cities: it generally does not decrease beyond -10°C in Washington DC compared to Seoul.





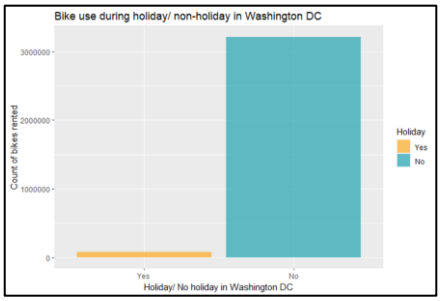
Next we investigate the visual relationship of bike renting in one of the four seasons. In the boxplots below we see that the number of bike rents decreases in winter. However, in the figure associated to Seoul, this decrease in winter is very significant as even the outliers for that season are around the same as the medians in the other three seasons. We could perhaps argue that this is due to the temperatures for that season being much lower than in Washington DC for the same time frame.

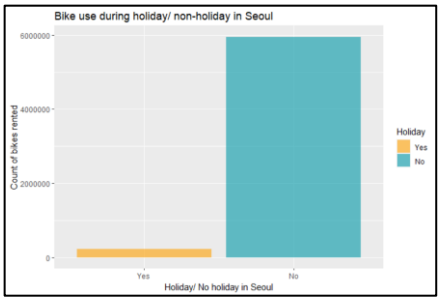




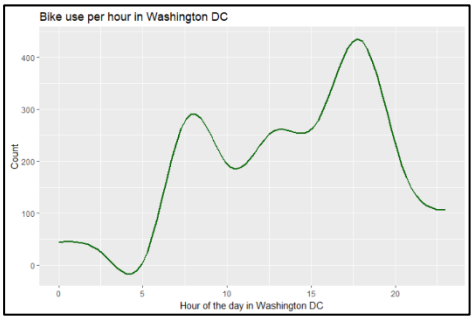
The next question we would like to ask is whether holidays affects the likelihood of a person renting a

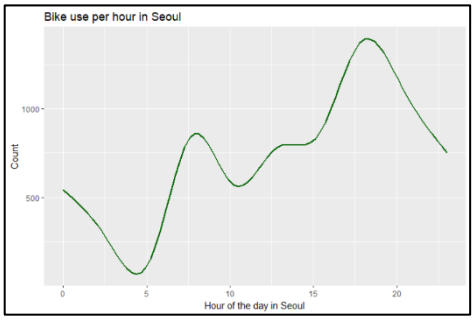
bike. From the two figures below, we can see that the demands are not equal. However, we would need to consider the number of days which would count as a holiday as they make up only a small part of the overall data. They have the potential for higher bike use (by proportion).



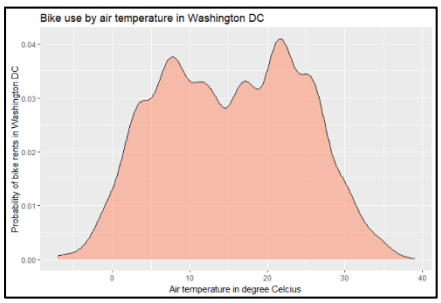


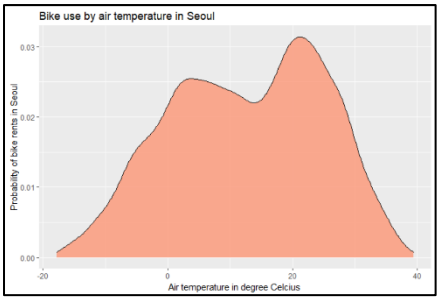
Another factor that could influence the number of bikes rented is the time of the day. One would hypothesise that during rush hour (in the mornings and late afternoons) rents would show a spike in popularity. We end up seeing that in both cities there is an apparent sudden increase prior to 6 am and another rise at around 3 pm onwards until 5 pm. Thus, we conclude that bike rents are affected by the hour of the day.





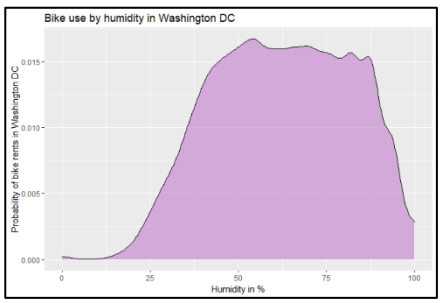
Next we examine how meteorological conditions influence the count of bike rents. Below we can see plots of air temperature, humidity and windspeed against the probability of renting (Washington DC vs Seoul).

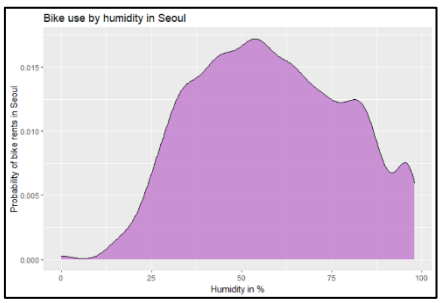




Whilst bike renting at the extreme ends of the temperature distribution appears low, there seems to be a slight preference for renting just above 25°C, but overall it does not appear to be a strong indicator. Overall, the density plot above is fairly spread out, noting that at its peak the probability is just 0.04.

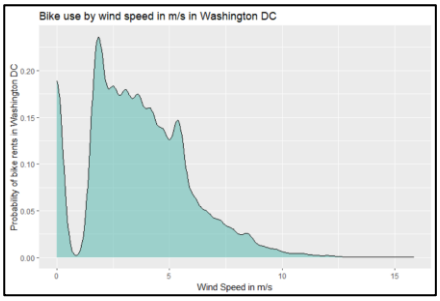
It does appear as though a humidity percentage of above 25% increases the likelihood of one renting, but one must note that both cities

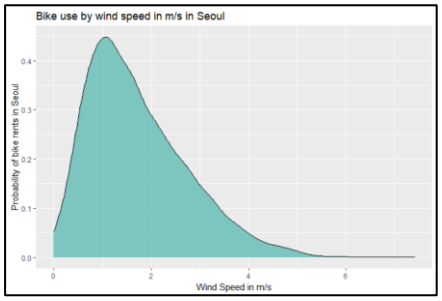




are generally humid throughout the year (see meteorological data online [1]). And similarly as with the density plot for the temperature, we do observe that the spread is quite high (peak probability is just slightly over 0.015).

Lastly, there is most certainly a inclination to rent a bike when the wind speed is less than 10 m/s and 6 m/s in Washington DC and Seoul, respectively. It appears that wind speed has the largest influence on bike count compared to the other predictors we have investigated.





Statistical Modelling

We have fitted a linear model to the two data sets of the following form:

$$BikeCount_{City} = \beta_0 + \beta_1 Season + \beta_2 Temperature + \beta_3 Humidity + \beta_4 WindSpeed$$

```
Call:
lm(formula = log(Count) ~ Season + Temperature + Humidity + WindSpeed,
    data = BikeW)

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 0.0000000337 ***
---
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

> summary(LMSeoul)

Call:
lm(formula = log(Count) ~ Season + Temperature + Humidity + WindSpeed,
    data = BikeS)

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)   6.7338965  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
```

For the linear model associated with the Washington data set (*BikeW*), it appears that all explanatory variables are good predictors of bike count (p-values are noted with an *). However, the fit of the model does not appear to be good as we find a low multiple R squared value of 0.278 associated to it.

Comparing this output to the Seoul data set (*BikeS*) below the only variable which appears to be insignificant is the factor *summer* in seasons (with a p = 0.912). We, moreover, find that the fit of the model appears to be improved compared to the first model output (we have a multiple R squared = 0.494) with an overall p-value of < 2.2e-16.

Next, we investigate the 97% confidence intervals for the estimated regression coefficients. We note that in both cases all but the intercept coefficients lie near zero. In the output for *LMSeoul* the confidence interval for the coefficient associated with the factor level *summer* contains 0, which would imply that the value of 0 is a possibility, i.e., that the factor *summer* is possibly insignificant and does not contribute to the model. However, many of the confidence intervals lie very close to zero, and thus there could be inaccuracies when attempting to predict using these models.

```
> confint(LMWashingDC, level = 0.97)
                                1.5 %      98.5 %
(Intercept)                4.50119998  4.75160198
as.factor(Season)Summer -0.43033590 -0.30000019
as.factor(Season)Autumn  0.47339115  0.59897666
as.factor(Season)Winter  0.03052896  0.17869159
Temperature              0.07601506  0.08356781
Humidity                -0.02449639 -0.02218851
WindSpeed                0.01487540  0.03412903
> confint(LMSeoul, level = 0.97)
                                1.5 %      98.5 %
(Intercept)              6.632322686  6.83507030
as.factor(Season)Summer -0.067553139  0.07476072
as.factor(Season)Autumn  0.316546593  0.43009553
as.factor(Season)Winter -0.458984431 -0.30708797
Temperature              0.046002904  0.05253719
Humidity                -0.023548780 -0.02144592
WindSpeed                0.005077663  0.04568421
```

We can, assuming the model is reliable, predict the number of bikes rented. Say we are interested in knowing the bike count when the season is winter, the air temperature is 0°C, the humidity is 20% and when the wind speed is 0.5 m/s. The 90% prediction intervals are:

```
> predict(LMWashingDC, newdata = NewData, interval = "confidence", level = 0.90)
       fit      lwr      upr
1 4.276413 4.215417 4.33741
>
> predict(LMSeoul, newdata = NewData, interval = "confidence", level = 0.90)
       fit      lwr      upr
1 5.913404 5.865934 5.960874
```

For Washington DC the expected number of bike rents is approximately 4 and for Seoul it is around 6.

Summary

Overall, we cleaned the two data sets so that statistical analysis could be performed on them. We began by visualizing the relationship between the temperature and time of the year, as well as the relationships between the use of the bike renting system at different seasons. Although it appeared that bike usage was the highest on non-holidays, we need to account for the fact that holidays only make up a small percentage of the year. This could lead to, that by proportion, bike usage is the same or higher on holidays, but further investigation is required to draw this conclusion. Next, we observed, in both data sets, an increase in the bike renting service around the morning and early evening rush hours. Lastly, the relationship between meteorological conditions and bike renting was analysed. We could infer that the use of the renting system appeared to depend mostly on whether the temperature was below freezing and if the wind conditions were such that the wind speed exceeded 10 m/s and 6 m/s in Washington DC and Seoul, respectively.

The statistical modelling part of this report focused on fitting a linear model to our two cleaned data sets. The goal was to try and use the model to predict the rented bike count based on the following explanatory variables: season, temperature, humidity and wind speed. Although most of the explanatory variables read as being significant, the fit of the models appeared to be low. The confidence intervals associated with the coefficients also appeared to lie near zero which could lead us potentially to making an incorrect prediction. However, if we were to assume that this is not the case and the two models are reliable, we make two predictions and find that the count was on the lower end (predicted ~ 4 rents in Washington DC and ~ 6 rents in Seoul given the conditions).

References

[1]
Weather and climate (2010). *Weather and Climate information for every country in the world.* [online] Weather-and-climate.com. Available at: https://weather-and-climate.com/.