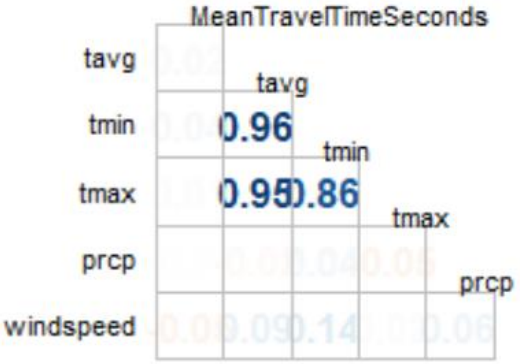
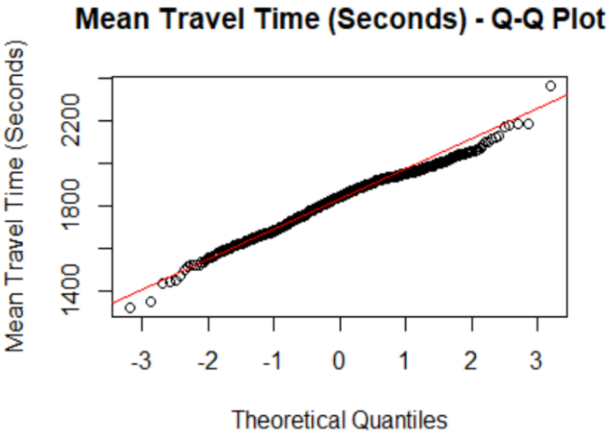
Upon splitting the uberLondon dataset, there is 699 observations remain in the training set. Using the dfSumary(),the following is found.

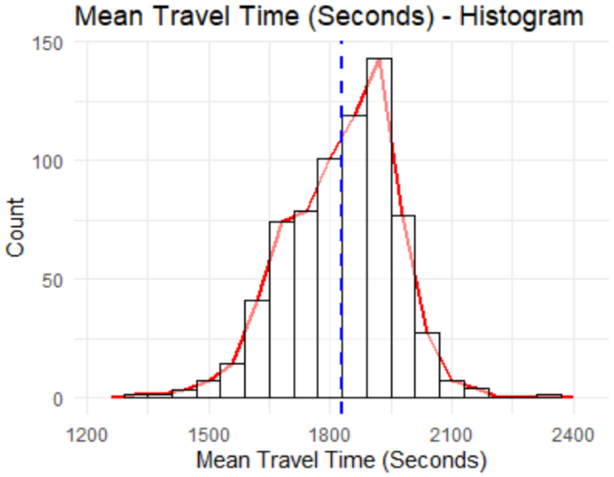
	MTTS	tavg	tmin	tmax	prcp	windspeed
Mean	1829	12.06	8.423	16.01	1.563	0.927
IQR	191.4	8.6	8.2	9.5	1.3	0
SD	132.8	5.5	5.3	6.3	3.7	0.6
No. of NA	0	0	5	6	1	0

Below, a correlation plot is shown. Apart from the temperature factors, all other variables have truly little correlation. The correlations of weather factors and mean travel time are remarkably close to 0.00. This means weather is not an influential factor in affecting the length and duration of travel using uber.



Below is a Q-Q plot and a histogram of the Mean Travel Time (Seconds). The concerned variable is generally normally distributed with a few of outliers deviating away from the Q-Q line. From the histogram with a bin width of 60, given the blue dotted line is the mean, the mean travel time appears to be slightly negatively skewed.

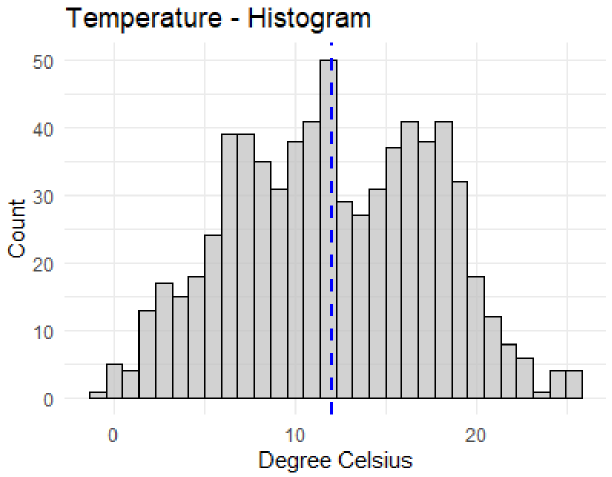


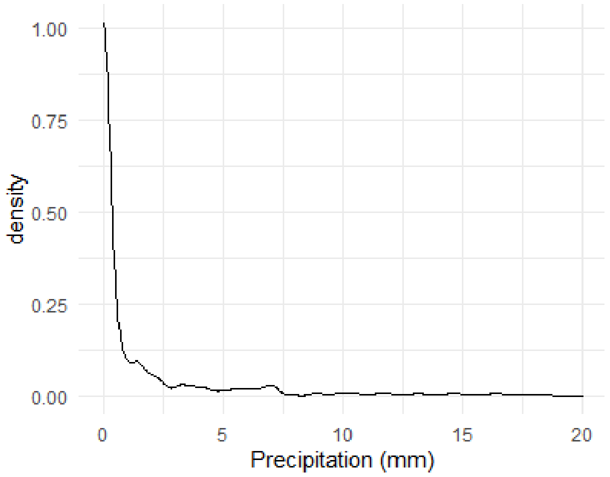


Using the boxplot() function, it is discovered all 5 outliers illustrated below, as defined by the 1.5 IQR rule, were bizarrely interestingly in January. While, using the Rosner's test, only 09-Jan-17's value is a true outlier. Despite the other 4 are not true outliers as defined by Rosner's test, it is still worth investigating why all these 5 values saturated in January.

Date	02-Jan-16	04-Jan-16	05-Jan-16	03-Jan-17	09-Jan-17
Value	1441.684	1326.496	1355.872	1446.048	2363.019

Below is a histogram for the average temperature and a density plot of the precipitation. From the histogram which has 30 bins, the average temperature was deemded to follow a normal distribution. Whilst, in the Density Plot, we can find that most of the values are near to zero, meaning that there is little or no rain fall during the observed period.





There are 5 missing values for the tmin and 6 missing values for tmax. Whilst the value of the tavg of these 6 days is available. This is deemed to be abnormal and counter-intuitive since it is hard to produce the average figure for the day without the recording all values of the day. Besides, as at 2017-01-18, tmin is available but there is no tmax value. All these imply that there is serious problem in the collection of data which could be a lack of quality assurance in the data collection process or there are some IT restrictions. Thorough investigation is advised.

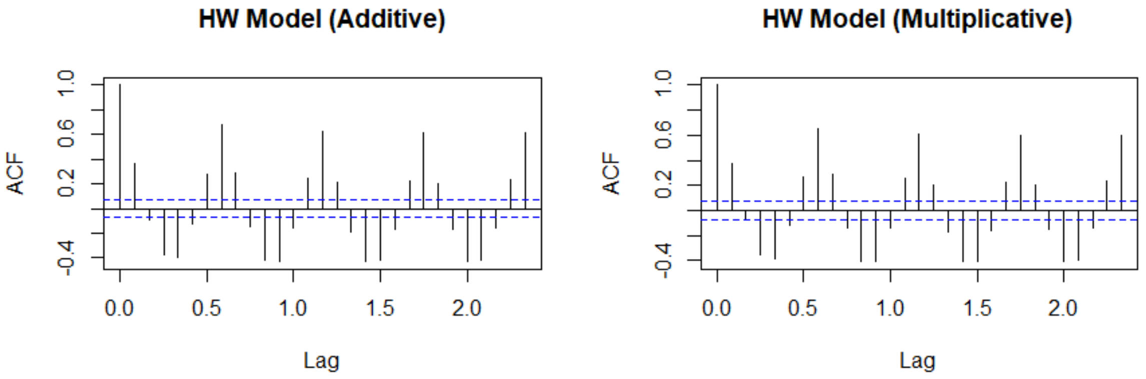
Owing to the extremely little correlation between the mean travel time with other weather-related variables, vector autoregression, which considers all variables affecting with one another, is regarded as not suitable. Whilst, the variance appears to be constant across the time, Box-Cox transformation is deemed to be ineffective and hence has not been conducted.

To compare different models with diverse features, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) are decided to be the benchmark of comparison among the model attempted.

Generated by using auto.arima() function, an ARIMA (1,1,5) model, hereby naming as *autoarima1*, with an AIC of 8470.42 recorded. While, setting approximation and stepwise to False in the auto.arima() function, hereby naming it as *autoarima2*, a same model of ARIMA(1,1,5), with an slight lower AIC of 8353.13 is obtained. Despite the forecasts using *autoarima2* provides a higher value of RMSE (323.09) than the *autoarima1* model (250.64), the forecasts of *autoarima2* provides a lower MAPE value (0.08222229) and a lower MAE value (147.5077) than its counterpart (MAPE = 0.08687857, MAE = 158.0475). However, the AIC of these 2 models is so high that indicates there may be other more suitable models.

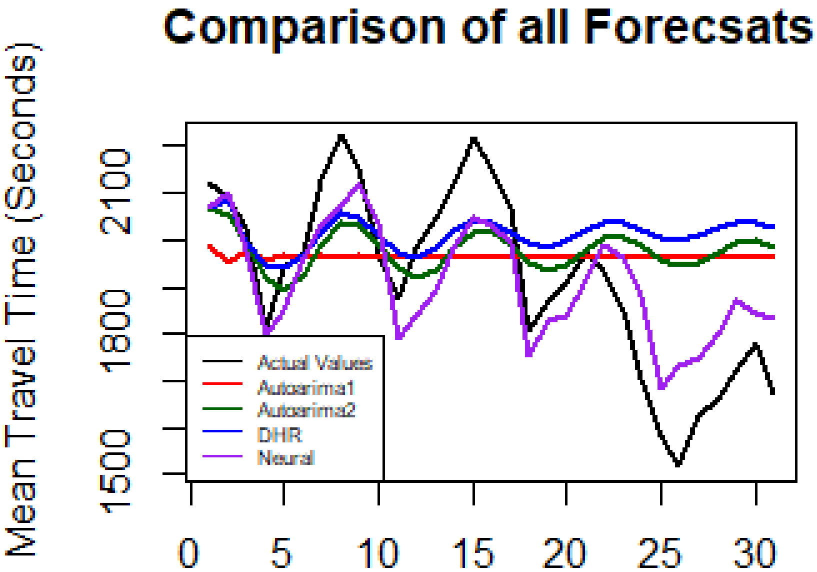
Dynamic Harmonic Regression (DHR) is suitable for data with any length of seasonality and short-term dynamic. Hence DHR is attempted. Ranging from 1 to 5, number of Fourier sin and cos pairs, denoted as K, were fit to find out the best smoothness of the seasonal pattern. The model with K = 2, named as *dhr2*, provided the lowest AIC (-2117.48) among all DHR. The *dhr2* provides a MAE of 157.4607, MAPE of 0.08898 and RMSE of 516.3021.

Holt Winter Method, both additive and multiplicative, have been attempted. However, the ACF plot of the residuals does not look like white noise. This model will not be chosen.



Artificial Neural Network (ANN) provided the best forecast among all models illustrated in this paper. With a RMSE of 43.7350, MAPE of 0.0512, MAE of 95.6152, the nnetar() function offers a model of NNAR(28,14). There are in total 28 lagged inputs and 14 nodes in the hidden layer. However, the major shortcoming of this method is that the forecast value varies when setting different seed, which the result deviates from each other a lot. Hence a solution of this is to run the model with multiple seeds and choose the one with the best forecasts. Other shortcomings of using ANN include high dependencies on the hardware computational power, and time-consuming during computation.

On the right, a comparison of all models is illustrated. For simplicity, only the concerned 31 days in December 2017 are shown. Values in the testing set is shown as black. It is visible that the neural network method, which is in purple, provided the best forecasts in graphic terms. It may be because of the learning algorithm in the neuron network help reduce the forecast error.


Comparison of all Forecsats
Mean Travel Time (Seconds)
Actual Values
Autoarima1
Autoarima2
DHR
Neural