Introduction

This project focuses on estimating demand at each of the 13 bike-sharing stations in the city of Nyon, Switzerland. The stations are Changins, Débarcadère, Gare Sud, Piscine du Cossy, Hôpital, Petit Perdtemps, Hostel, Place de Savoie, La Plage, Stade de Colovray, Triangle de l'Etraz, Gare Nord, and Château. **Fig 1** shows how the stations are spread out across Nyon. Apart from the two bike stations around Nyon Gare, most stations are spread out. There are four stations clustered together near the lake front as well. However, this is a population dense part of town, therefore all these bike stations are likely to come in use separately from each other. **Note**: Most of the analysis was done on data from 03rd March to 15th May. The final forecast makes use of data until 22nd May, however.

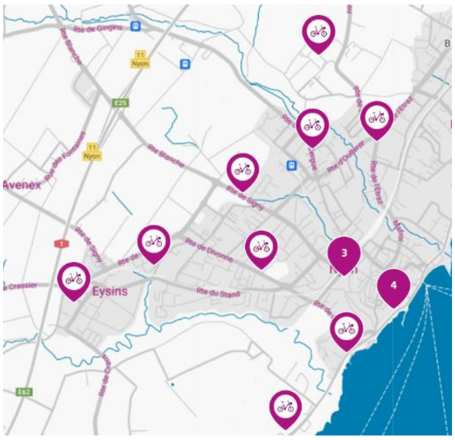


Fig 1.1 – Distribution of Stations

Exploratory Analysis and Data Wrangling

Before data exploration can begin, there are a few limitations that need to be dealt with. Firstly, the data had several missing values. While most of the data was structured around 10-minute intervals, there were several dates where it had 30-minute intervals. On investigation, it was discovered that bike demand does not vary dramatically between 30-minute intervals. Therefore, the gaps were filled with the last available value using the *zoo* package. *Dates* were converted into R's *POSIXct* format so that they could be easily manipulated into tsibbles. *TotalBikes* was created as a sum of *Bikes* and *E-Bikes*. **Fig 2** shows how the demand of total bikes across Nyon varies. There's a lot of variability within the days and while there are peaks and troughs, there does not appear to be a long-term trend or longer than weekly seasonal pattern.

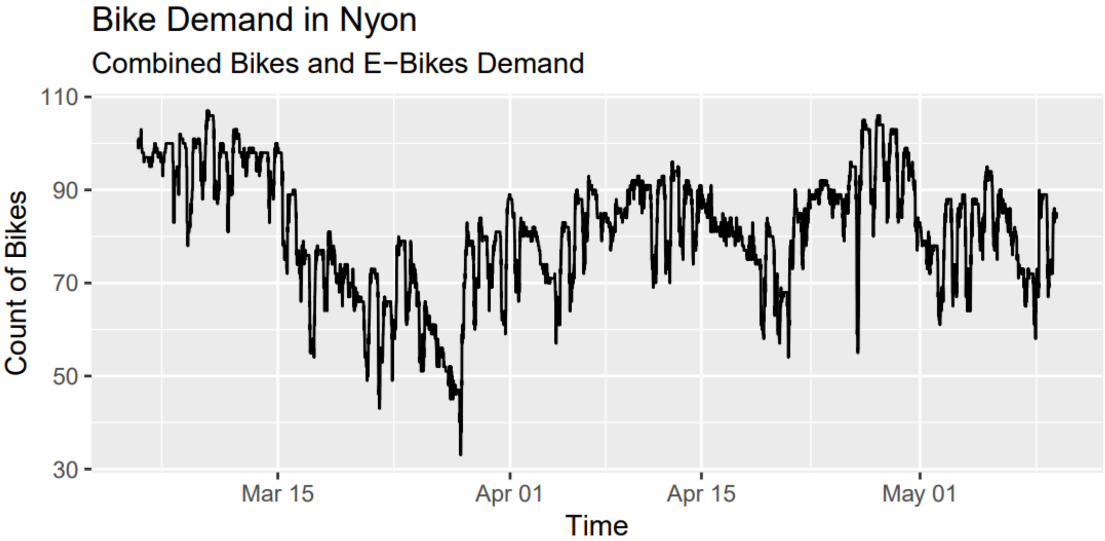


Fig 2 – Total Bike demand in Nyon over Time

In order to investigate whether any stations had similar demand patterns and could be joined together because of their proximity, *TotalBikes* were plotted against time for every region as shown in **Fig 3.** This reveals a variety of daily patterns and differing demands even among the two stations around *Gare*. Therefore, the prudent approach was to keep all stations demand separate rather than creating new stations.

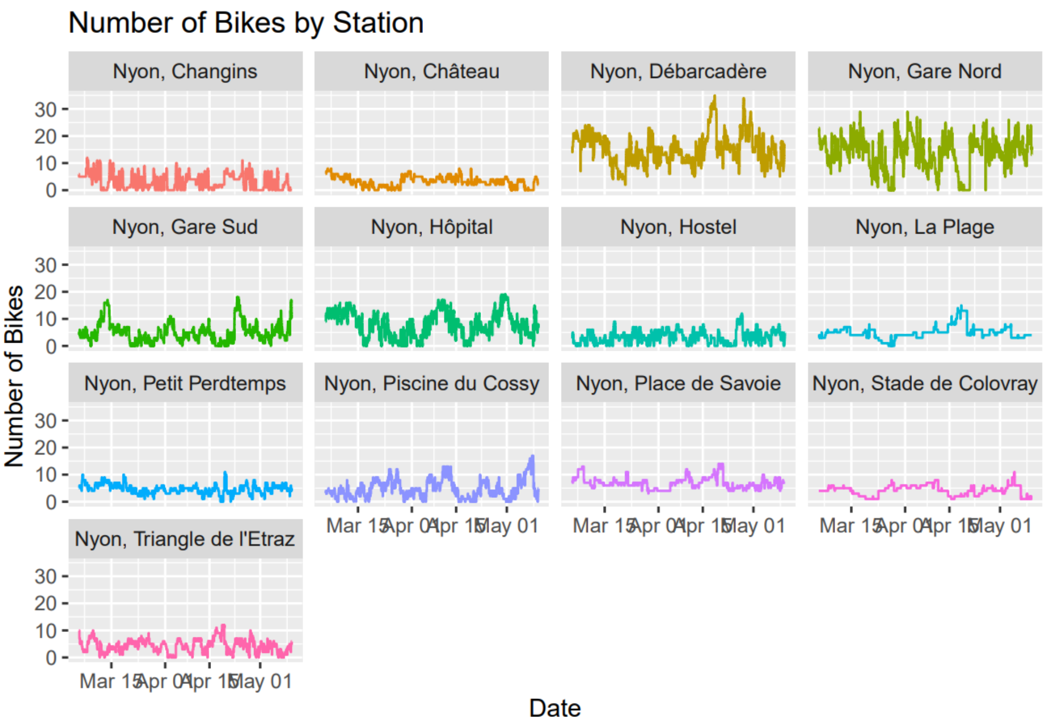


Fig 3 – Demand Across All Stations

Demand for *Bikes* and *E-Bikes* was also investigated to determine if their time-series differed enough to justify separate forecasting. This correlation can be found in **Appendix 1**. There was not enough of a difference between the two time-series to justify separate forecasting thus *TotalBikes* would be forecasted eventually. Next, three variables were created *timeofday*, *weekday* and *holiday* that were 0 or 1 depending on whether the day in question was a weekday or weekend or whether it was a holiday. *Timeofday* was later removed as this variable was found to be covered within the seasonality of the timeseries itself and including it resulted in multi-collinearity between the variables. Investigation on these variables impact can be found in **Appendix 2**. Lastly, weather data was added from the *WorldWeatherOnline's* API which included many variables including temperature, snow and rain. A basic plot of *TotalBikes* demand changing with weather is depicted in **Fig 4** while **Fig 5** shows a PCA against temperature detecting the outliers as days where temperature was considerably lower. More outlier detection and analysis can be found in **Appendix 3**.

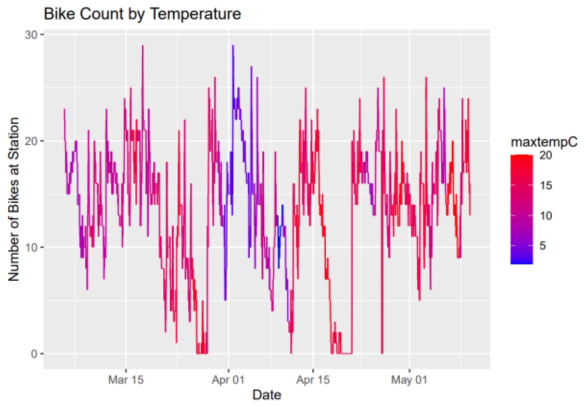


Fig 4 – Bike Demand by Temperature

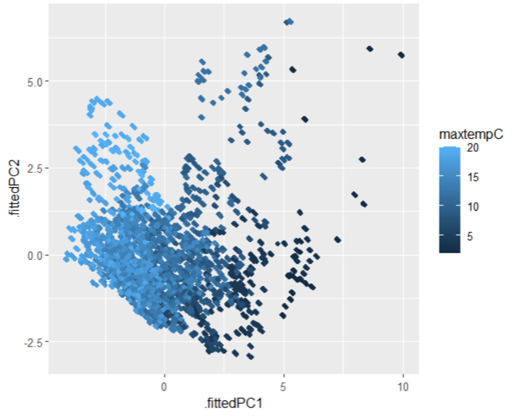


Fig 5 – PCA for Outliers

Modelling

Before the modelling can begin in earnest, the first step is to see if there is some sort of seasonality that has not been captured in the earlier EDA. The *STL* function from the *fpp3* package was utilized to create decompositions of all areas. For reference, '*Nyon, La Plage*' is shown here as an example. The *STL* function is adept at detecting multiple seasonalities and allowing modelling based on those. **Fig 6** shows this with the *STL* function decomposing the time-series into daily and weekly seasonalities which can be used.

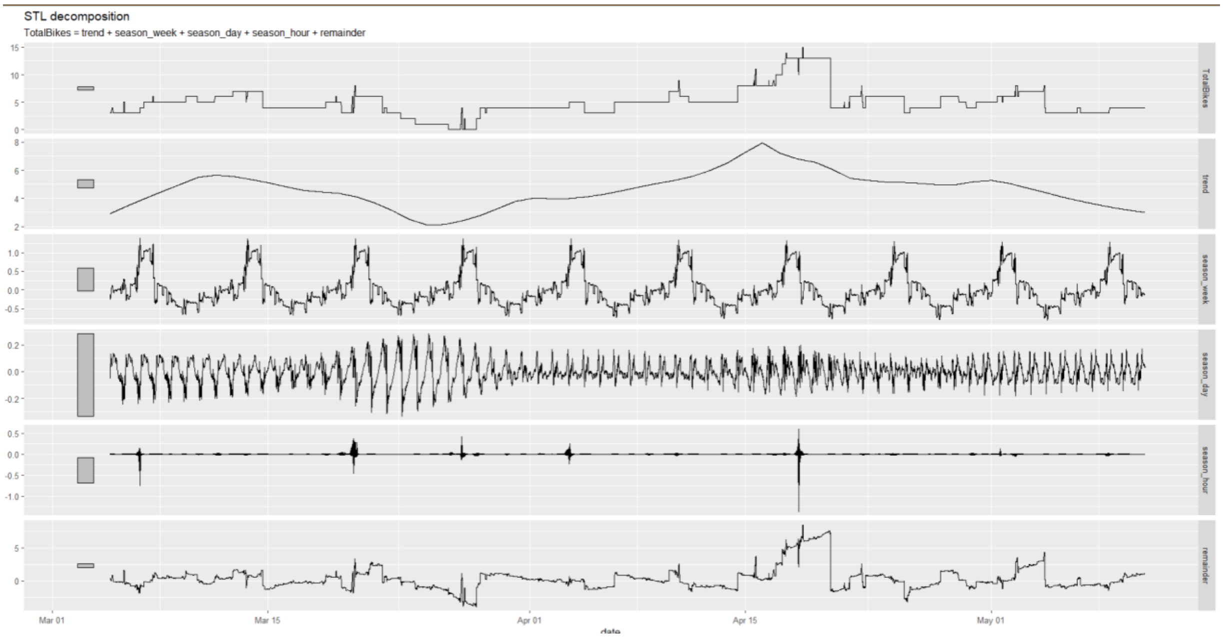


Fig 6 – STL Decomposition of Nyon, La Plage

From the figure, it can be seen that there is a clear pattern in terms of weekly seasonality (3rd box). Number of bikes at the station are low during the weekdays and jump dramatically during the weekend. Similarly, a daily seasonality can be seen (4th box). Demand tends to be lower during the night and higher during the day. There are exceptions. April 18th (Easter Monday) where the residuals were high. Moreover, daily demand seasonality was different in the last week of March (snow in Western Switzerland). Therefore, the STL decomposition is powerful enough that additional variables are not required. We will build an ETS model based on this decomposition as our first model using the seasonally adjusted data. The fitted values for *Gare Nord* are shown in **Fig 7**.

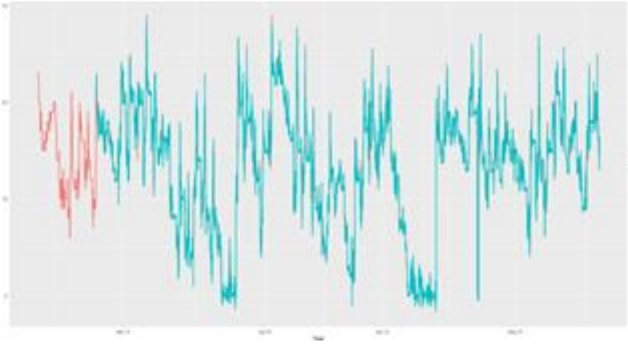


Fig 7 – Fitted Values STL Model

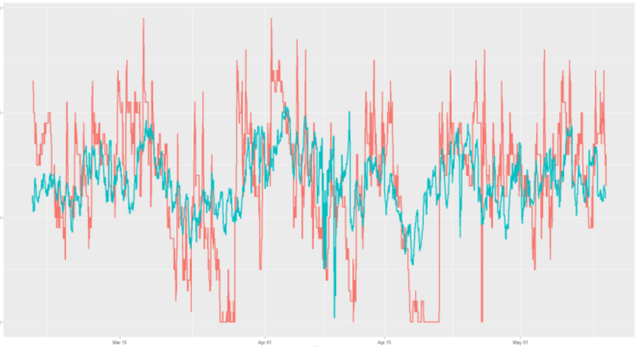


Fig 8 – Fitted Values TSLM Model

The relationship between weather patterns and bike demand has already been established. Therefore, the TSLM model which allows exogenous variables to be taken into account can also be considered. The TSLM modelling was done using the following variables: weekday, snow in cm, uvIndex, Heat, wind speed and chill, visibility, cloudcover, humidity, holiday and obviously, temperature. The model was run and the fitted values are shown in **Fig 8**. Data on residual analysis of the TSLM can be found in **Appendix 4**. Two other models were run as baseline comparisons for these two models – basic ETS and ARIMA. Details regarding why these models and their results can be found in **Appendix 5.**

From **Fig 7** and **Fig 8**, we can see that the STL model tends to overfit. The TSLM model does seem to be picking up the ebbs and flows of the overall time-series just as well. Demand drops during the holiday period as expected and also is lower during the last week of March when the temperature was low and the weather was inclement. The fear with the STL model here is that because of the overfitting, the forecast developed using it will have a large confidence interval and the forecast will be mostly useless. In order to pick between all the models developed, data was split into training and test sets. Data prior to 11th May was treated as the training set whereas data from May 11th to May 15th would be used to test the accuracy of the forecast models. The forecasts for *Gare Nord* of all the models developed can be seen in **Fig 9**.

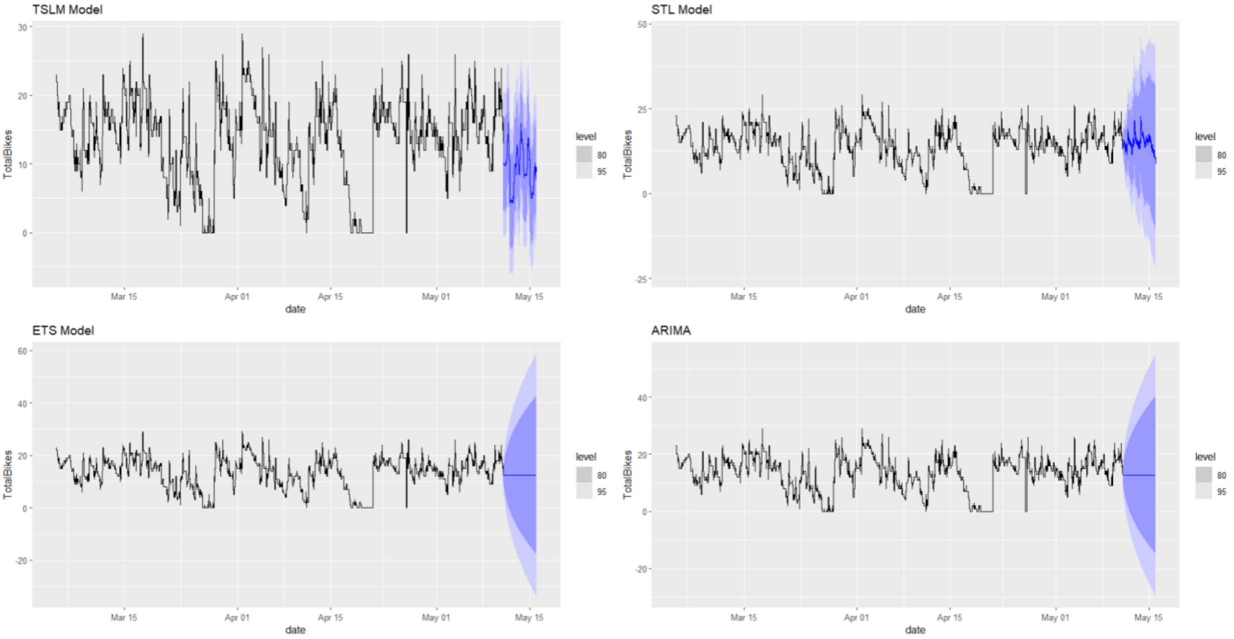


Fig 9 – Five-Day Forecast for All Models

There are some interesting observations to be made here – the ETS and ARIMA model overfit in general and do not provide good accuracy at all. Their confidence intervals are higher than the entire series values. The TSLM model tends to provide the smallest confidence intervals. However, it must be noted that the TSLM overstates the effect of the trend. In this case, and in cases of other stations, the forecast would continue on declining below 0. The STL model, as feared, has a high confidence interval. However, the STL model also successfully manages to predict the trend of the series correctly. The four models were judged on ME, MAE and RMSE. The full accuracy table can be found in **Appendix 6.** Similar analysis was done on a 24-hour forecast but the results did not differ. In the end, the STL decomposition model performed best in most locations in all three accuracy measures. Therefore, it was the model selected going forward.

The last thing to be considered was weather a bottom-up hierarchical approach would work best or whether each time-series should be considered individually and forecasted accordingly. In order to test this, the selected STL decomposition model was used to create a hierarchical cluster of all locations and then a bottom-up approach was used to reconcile and forecast the next five days. A summary of the results can be found in **Table 1**. It is clear from the results that the bottom approach proves superior. Therefore, it will be used for the final forecast.

Location	**ME**		**RMSE**		**MAE**	
	Bottom-Up	Regular	Bottom-Up	Regular	Bottom-Up	Regular
Changins	-0.4	-0.5	1.0	1.1	0.7	0.7
Chateau	-0.6	-0.5	1.3	1.3	1.0	1.1
Debarcadere	-0.4	5.7	1.9	7.6	1.6	6.1
Gare Nord	-0.5	2.0	1.7	3.7	1.5	2.8

Gare Sud	-3.4	-6.5	5.2	7.0	4.0	6.6
Hopital	-0.1	-1.5	2.1	3.1	1.6	2.3
Hostel	-1.5	-2.9	2.0	3.4	1.5	3.0
La Plage	1.7	1.5	2.9	2.6	1.8	1.8
Petit Perdtemps	-0.4	-1.0	0.9	1.4	0.6	1.1
Piscine du Cossy	-1.0	-0.2	1.8	2.0	1.3	1.6
Place de Savoie	0.1	-1.7	0.9	2.6	0.6	2.3
Stade de Colovray	0.5	1.8	0.8	2.2	0.5	1.8
Triangle de l'Etraz	0.2	-1.9	1.9	2.8	1.7	2.4
Aggregated	-5.8	NA	7.8	NA	5.9	NA

Table 1 – Analysis of Accuracy Measures

Forecast

Fig 10 shows the final forecast developed using the bottom-up STL model:

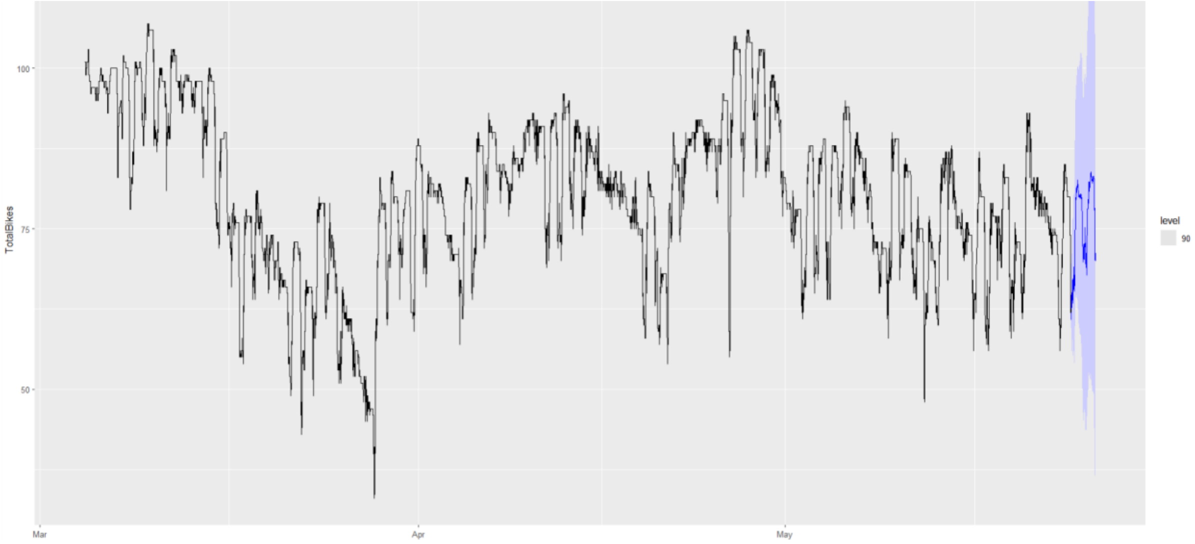


Fig 10 – Final Forecast of Total Bikes

Limitations

The bottom-up STL decomposition model proved to be the most effective one when judged on RMSE/MAE/ME. However, there are some drawbacks of this model: (1) The confidence intervals are very large and become larger as the forecast horizon increases (2) the model does not take into account exogenous variables such as weather (3) Only the pre-evaluated seasonality was used. (4) An ETS model was used, further research could be conducted on ARIMA models using Fourier Transformations in time-series with complex and multiple seasonality. (5) The demand for bikes and e-bikes was combined. Future research could evaluate whether separate forecasting yields better results. (6) Population was not considered as an exogenous variable due to computational complexity which probably impacts results to a large degree.

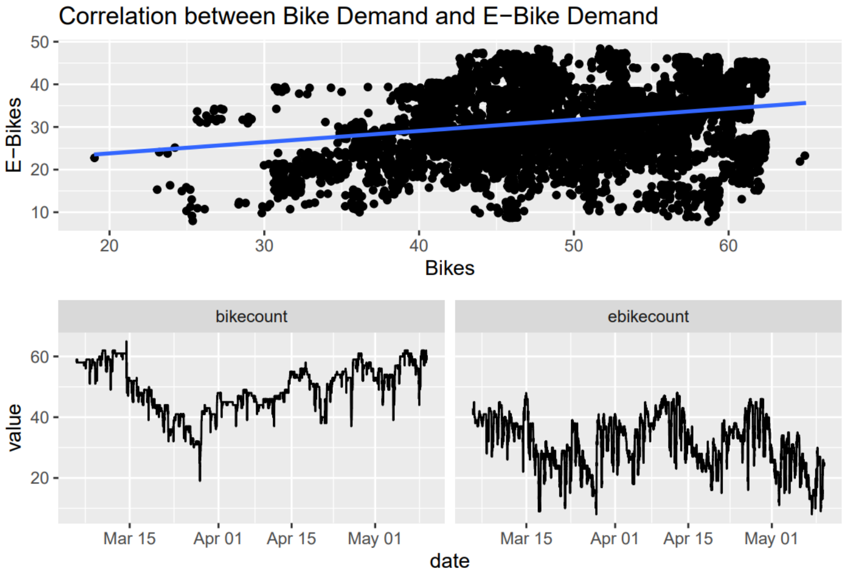
Evaluation and Conclusion

This report has endeavored to evaluate different models in estimating bike-sharing demand. The STL decomposition model was selected in the end and despite its aforementioned drawbacks. In terms of the forecast results, unlike the TSLM model, they are reasonably spaced and do not display a perpetually increasing or decreasing trend. However, because of the large confidence intervals, the prediction is not as cogent as the group would like. Nonetheless, the forecast captures the daily and weekly seasonality accurately enough and presents a forecast that can be credible. In short, this report has shown that even erratic time periods with complex seasonality can be modelled accurately.

Appendix

Appendix 1

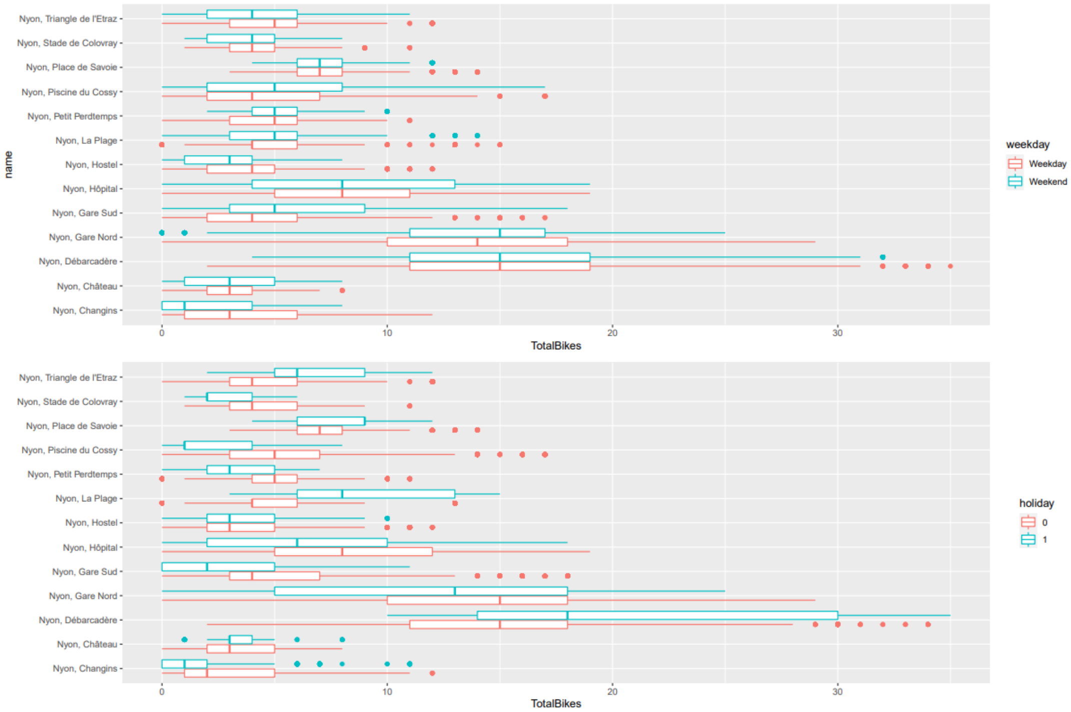
This appendix shows the correlation between bike demand and e-bike demand. There is some minor correlation between the two but not enough to suggest that increases in one lead to a directly proportional increase in another. However, the two time series are not completely distinct so the work clumped them together.



Correlation between demand of bikes and e-bikes

Appendix 2

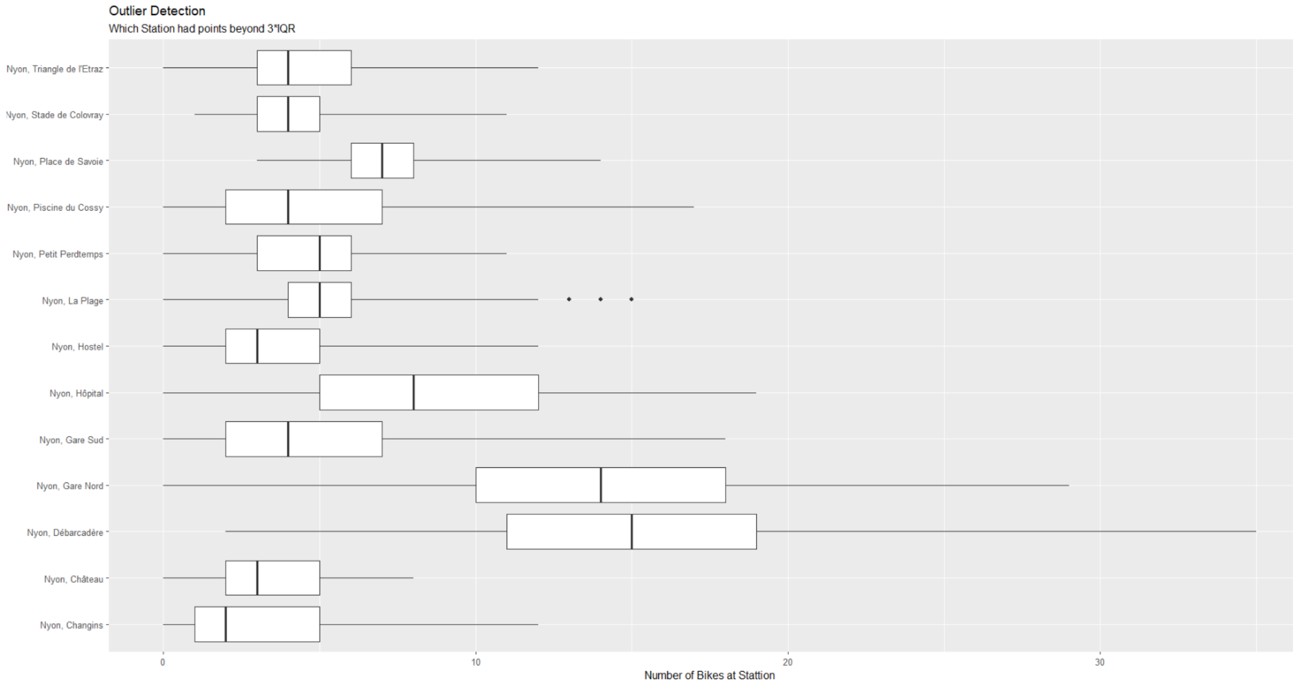
This shows the difference between bikes at a station for weekday/weekends and holidays. While there is not a lot of difference between weekends and weekdays, holidays tend to have different effects on different stations. The holiday mean is lower at Nyon, Hopital but higher at Nyon, Triangle.



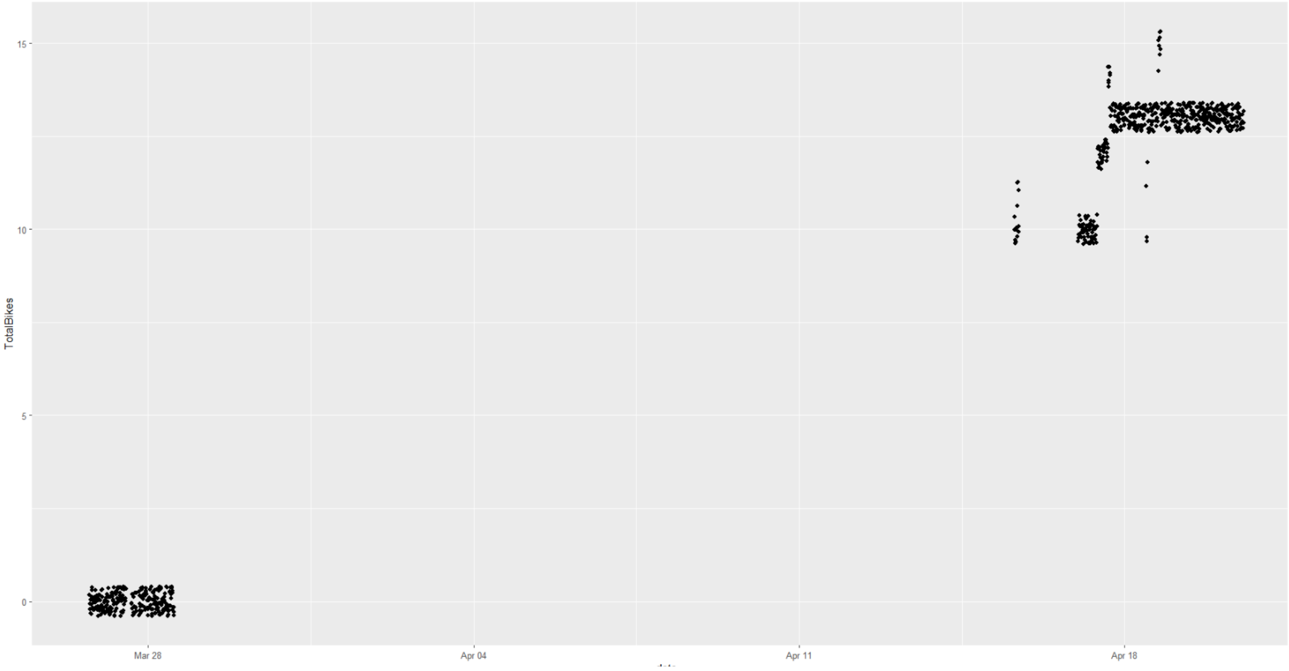
Difference in means

Appendix 3

Outliers were classified as points beyond the 3*IQR Range. There were no outliers except for La Plage which had outliers on 28th March (Snow) and 18th April (Easter Monday). These outliers were kept in the data as the TSLM model took into account both of these variables and the STL decomposition model did not show dramatic differences in seasonality by keeping these in



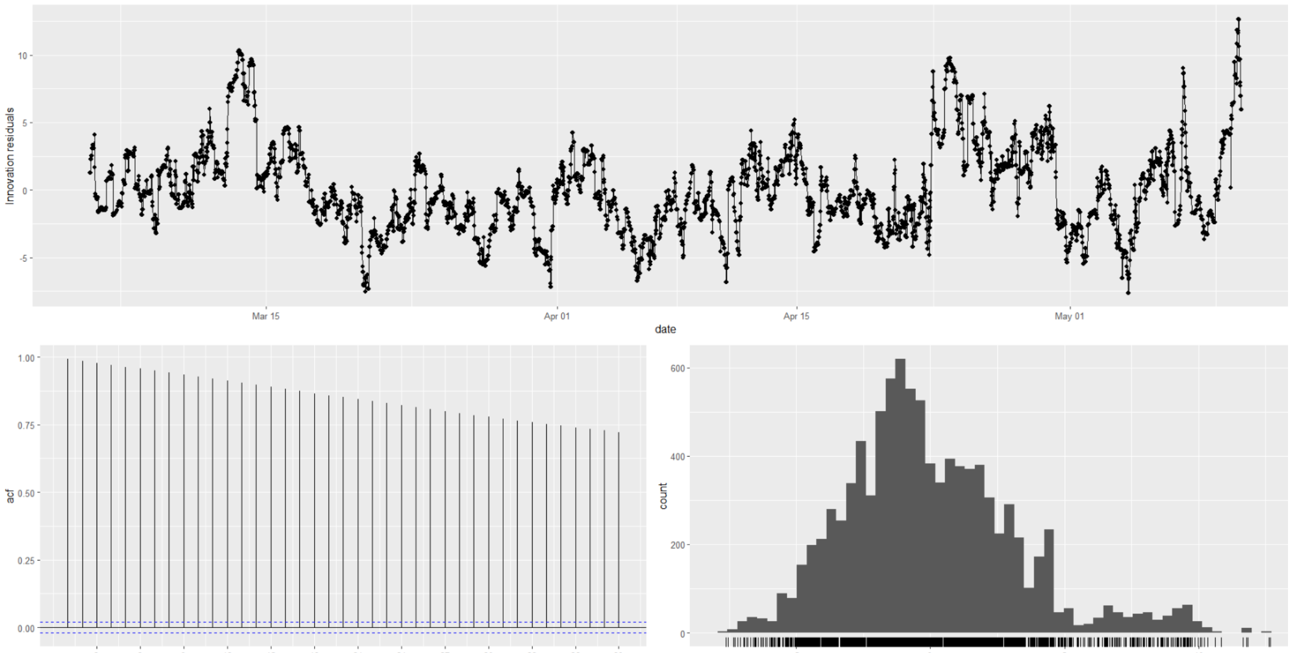
Outliers 1



Outliers 2

Appendix 4

This appendix shows the residual analysis for the TSLM model. The residual analysis was done to make sure that the residuals were normally distributed and were not correlated. As can be seen, there is no pattern to the residuals and they are normally distributed. The auto-correlation is to be expected in a seasonal time series.



Residual Analysis of TSLM

Appendix 5

The ARIMA model was built using the *ARIMA* function without any *pdq* values specified. *ETS* was built on Additive *Error* and *Seasonal* components. Trend was reduced to *N* as the bike demand is seen to be constant over time. The results of the ME can be found below (only ME is shown for brevity, however they performed similar across all measures):

		ME		
Station	**STL**	**TSLM**	**ETS**	**ARIMA**
Nyon, Changins	-0.5	0.0	1.1	1.1
Nyon, Chateau	-0.5	0.9	-0.2	-0.2
Nyon, Debarcadere	5.7	1.7	0.6	0.6
Nyon, Gare Nord	2.0	7.5	4.5	4.2
Nyon, Gare Sud	-6.5	0.7	-5.7	-5.7
Nyon, Hopital	-1.5	-2.2	-2.0	-1.9
Nyon, Hostel	-2.9	-2.1	-3.1	-3.1
Nyon, La Plage	1.5	0.5	1.5	1.5
Nyon, Petit Perdtemps	-1.0	-0.4	-1.1	-1.1
Nyon, Piscine du Cossy	-0.2	-4.2	-1.3	-1.2
Nyon, Place de Savoie	-1.7	-1.5	-1.4	-1.4
Nyon, Stade de Colovray	1.8	1.5	1.3	1.3
Nyon, Triangle de l'Etraz	-1.9	-1.6	-2.8	-2.8