New Zealand Tourism Data
Time Series Analysis

Timothy Lu
Springboard
Capstone 2022

Problem Statement

Forecasting appropriate travel numbers for countries has always been important. It allows them to predict and keep track of visas being distributed, the amount of security needed to protect their borders, and the expected economic benefits of having tourism to their country. When the COVID-19 pandemic hit, many countries closed their borders. New Zealand was one such country and now that we are slowly returning to normalcy, it is more important than ever to make sure we have spot-on forecasting for international visitors. Not only is it important for security purposes, but it is also now even more vital for keeping track of the health of your populace. On top of the border concerns, we must be aware of how travelers will occupy the accommodations within a country.

I have used time series analysis to answer both questions. While the complexities of forecasting data after a pandemic were challenging, I gleaned some promising results. As more data is collected, these models will only improve. By using these forecasts, the country of New Zealand can prepare itself for the future as more visitors come to the shores of Aotearoa.

Data Wrangling

The data was collected directly from the New Zealand government using the Infoshare platform (https://infoshare.stats.govt.nz/Default.aspx). I was able to collect monthly data on tourism visits from the April 1978 to May 2022. For accommodation data, there was monthly information available from January 2001 to September 2019. For the most part, the data was quite organized. There were some parts of the data such as descriptions and labeling which would not work in a machine learning model which I cleaned using python packages such as pandas.

I replaced any missing data with 0's to denote that there was no data collected during those time periods. Additionally, I made sure each feature column was clearly labeled with the type of data collected and the region. Thankfully, the government of New Zealand collected data quite cleanly with no missing dates to impute, and other than regions which started collection later, or in times of no data, there was no missing survey data. Utilizing continuous data is key for time series analysis.

Exploratory Data Analysis

As part of the exploration, I plotted out each individual time series to observe data continuity. For the most part, there was good continuity with the data and most of the time series seemed viable for analysis. However, within the Accommodations dataset, I noted that there were some series that were very disjointed. This could potentially be due to the fact that when there are 0 visitors to those regions those numbers were reported as NaN and thus did not show up in the time series at all. For our backpacker data, there were huge swathes of empty data. That makes sense as there are seasons in which there will be no backpackers. Surprisingly, we had some regions with very spotty hotel data in our smaller regions (**Figure 1)**. This could have been caused by economic factors which did not favor hotels and instead favored

smaller motels. Regardless, I saw that these datasets would not be good candidates for my time series analysis. The dataset which tracker arrival data had no such issues with continuance. There were definitely regions which had 0's due to the fact that they started collecting tourism data when those ports opened. Additionally, we see that the COVID-19 pandemic greatly impacted these data sets with border closings causing many ports to have 0 visitors for quite some time.

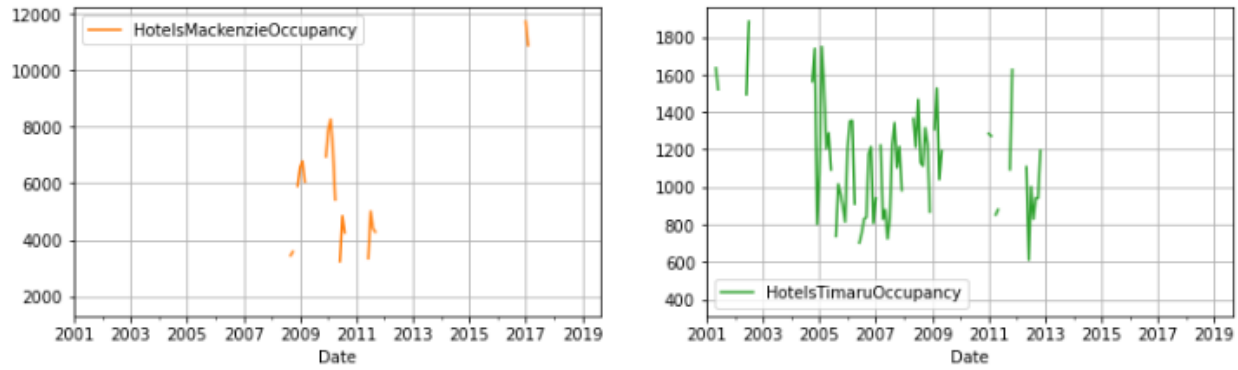

Figure 1 Example of Discontinuous Hotel Data

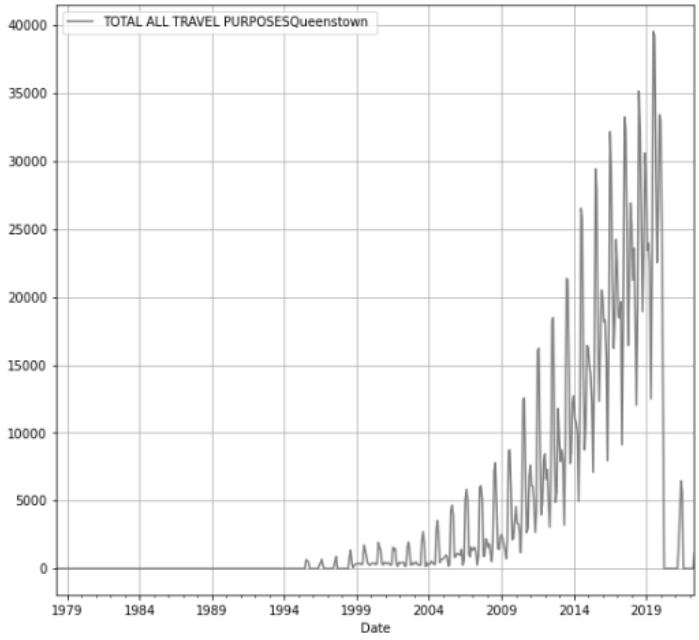

Figure 2 Example of Visitor Data with Later Collection and COVID-19 Pandemic Data

Seeing all of the time series plotted out like this was a great way to understand what approaches would need to happen. Some of the dataset really showcased strong seasonality and a clearly upward trend that would need to be accounted for in our models. Making sure I understood the factors at play with this data before modeling was very important and allowed me a greater ability to plan out my modeling approach.

In-Depth Processing (Pycaret)

Following my exploratory analysis, I needed to find a way to quickly model dozens of time series datasets and find appropriate models for each. Due to my initial lack of knowledge of all the time series models which existed, I opted to use Pycaret to help with my modeling. Pycaret was developed as a means of testing multiple models for multiple time series at once in order to generate what it considered the "best option" for modeling the data. Through this method, I would be able to learn about more time series models which existed beyond the traditional ARIMA models such as using machine learning techniques in order to create predictions.

I separately ran pycaret analysis on the accommodations and arrivals data as they were in two different CSV files. Due to the relatively small size, I ran the pycaret pipeline on the entirety of the arrivals data; however, due to the increased complexity of the accommodations data I had to make some decisions on which timeseries to model. I wanted to run pycaret on the entirety of the arrivals dataset so I can see which models it would choose for the datasets. For the final modeling, I chose those models which had a parallel in the accommodations data in order to cut down on final modeling volume and to focus on those regions which had the most continuous data.

For the accommodations data, I discarded any time series which were discontinuous. For those regions with a lot of discontinuous data I decided to not model those regions as a whole. For example, the regions of Rotorua, Hamilton, Dunedin, and Palmerston North were not included due to the lack of continuous data. Next, I chose accommodation regions which I felt were parallel to the ports of arrival. In this way, I can focus on the regions nearest to those ports and most likely to have a higher influx of visitors. Finally, I decided on the following 5 regions: Auckland, Canterbury (Christchurch), Queenstown, Wellington, and New Zealand. This allowed me to cut down on the time I would need to run the complex pycaret modeling.

Pycaret generates a list of models which had the best performance metrics for each timeseries. This gave me a starting point with which to model my datasets after and begin the formulation of models (**Fig. 3**). With one of the hardest parts out of the way, I was ready to begin modeling. I saved the list of models that pycaret chose into CSV files so that I could access them in a separate processing and modeling notebook.

	Model	MAE	RMSE	MAPE	SMAPE	MASE	RMSSE	R2	TT (Sec)	time_series
rf_cds_dt	Random Forest w/ Cond. Deseasonalize & Detrending	8349.3101	10150.3856	0.0232	0.0234	0.5438	0.5425	0.8999	0.1200	TotalAucklandOccupancy
ets	ETS	7285.5658	9498.0748	0.0454	0.0447	0.5716	0.4748	0.8676	0.0267	TotalCanterburyOccupancy
prophet	Prophet	5984.4584	7441.5653	0.0436	0.0434	0.8371	0.8323	0.6697	1.0500	TotalWellingtonOccupancy
ada_cds_dt	AdaBoost w/ Cond. Deseasonalize & Detrending	8587.296	10005.9799	0.0586	0.0604	1.203	1.1242	0.8265	0.0367	TotalQueenstownOccupancy
knn_cds_dt	K Neighbors w/ Cond. Deseasonalize & Detrending	57269.6411	68051.522	0.0332	0.0332	1.097	1.0582	0.9552	0.0333	TotalTotal New ZealandOccupancy
ets	ETS	2397.3683	2945.5814	0.0347	0.0347	0.7781	0.754	0.8022	0.0400	MotelsAucklandOccupancy
ada_cds_dt	AdaBoost w/ Cond. Deseasonalize & Detrending	3258.8863	4005.5378	0.0617	0.0593	0.9949	0.7904	0.6239	0.0367	MotelsCanterburyOccupancy
dt_cds_dt	Decision Tree w/ Cond. Deseasonalize & Detrending	1573.4806	2009.0225	0.0679	0.0643	0.7841	0.7951	0.386	0.0100	MotelsWellingtonOccupancy
bats	BATS	1769.7556	2227.6344	0.0912	0.0977	1.2692	1.2432	0.7731	6.1800	MotelsQueenstownOccupancy
lightgbm_cds_dt	Light Gradient Boosting w/ Cond. Deseasonalize...	13202.0324	16474.3382	0.027	0.0272	0.8571	0.8579	0.94	0.0133	MotelsTotal New ZealandOccupancy

Figure 3 An Example of the pycaret Modeling Pipeline Output

Modeling

This report can go on a long time about the modeling phase! With 30 models, it took a while to make sure every model ran smoothly. I will give a high-level overview of the models which had the most problems and the resolutions to those issues. Here is the variety of models which were ran on the dataset (some models used the same modeling algorithms): exponential smoothing, ARIMA, SARIMAX, ETS, BATS, Prophet, random forest, Adaboost, LightGBM, and K-Nearest Neighbors. The general process went as such: (1) break down each time series into its components of seasonality and trend (**Fig. 4)**, (2) take the pycaret model and fit the data to the model, (3) tune the model parameters as needed based on the components, and (4) forecast the data out for approximately 24-36 months to observe its performance.

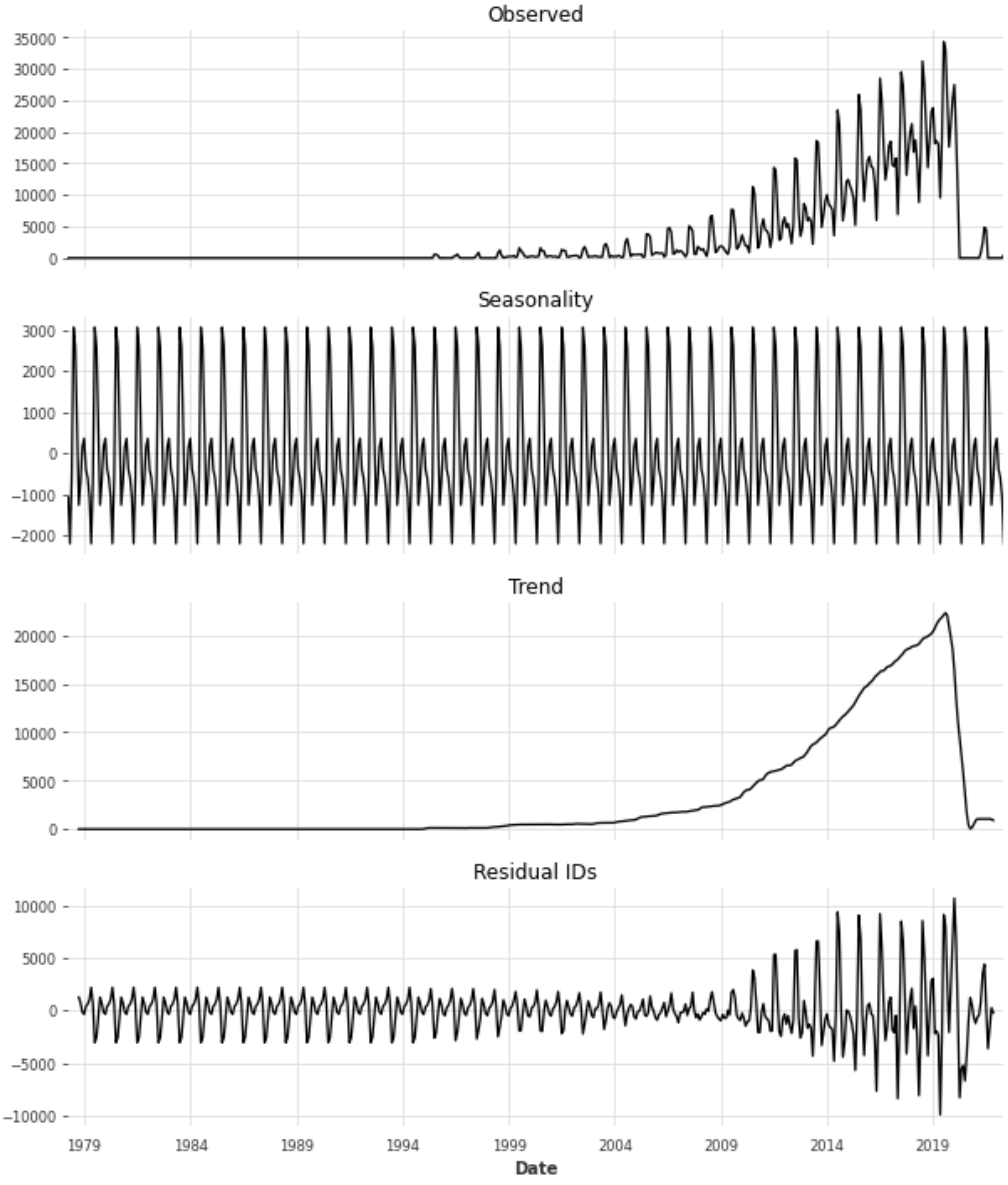

Figure 4 Seasonality and Trend Components for Queenstown Timeseries Data

I would like to first discuss modeling the tourism data. Due to the extreme situation of the COVID-19 pandemic causing the most recent data to be reported as 0's due to border closing, many of the traditional models struggled. For our unconstrained models such as SARIMAX, I noticed a severe number of negative values being predicted. The extreme dip caused chaos for these models and they struggled to adapt even when newer data had some upward trend. This is due to the fact these models make no assumptions regarding the data and only forecast based on the statistical nature of the data. If there is a dramatic downward slope, then the data will follow that trend to its conclusion and forecast a negative number which makes no sense for our data as we cannot have "negative" tourists.

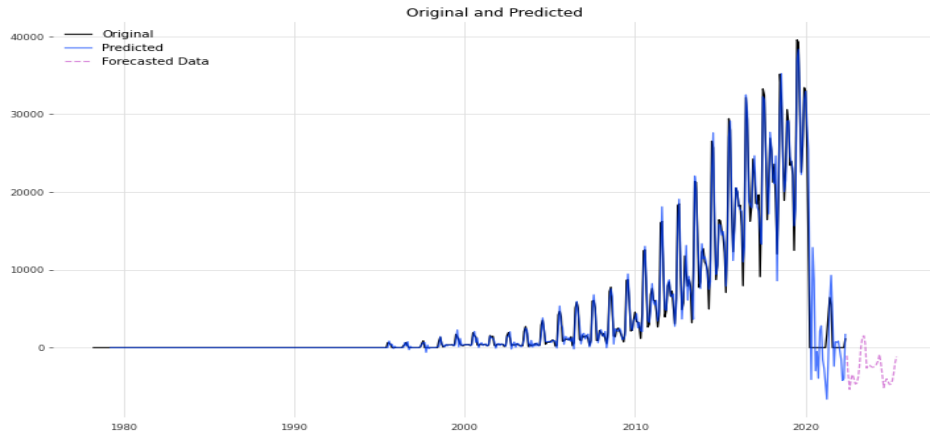

Figure 5 Auto_ARIMA Model for Queenstown Total Tourism

One of the models which struggled the most was the Auto_ARIMA modeling for Queenstown Total tourism data. As soon as the pandemic data hit, we see that the model predicted negative tourist over the historical data which propagated into the forecast (**Fig. 5**). On the other hand, a model which used BATS seemed to handle forecasting quite well and did not propagate negative trends too heavily moving forward (**Fig. 6**). This could be due to the implementation of exponential smoothing which dampens the impact of dramatic outliers such as the COVID-19 Pandemic. There is also the usage of a Box-Cox transformation and other calculations which are tuned by the DARTS package implementation of BATS.

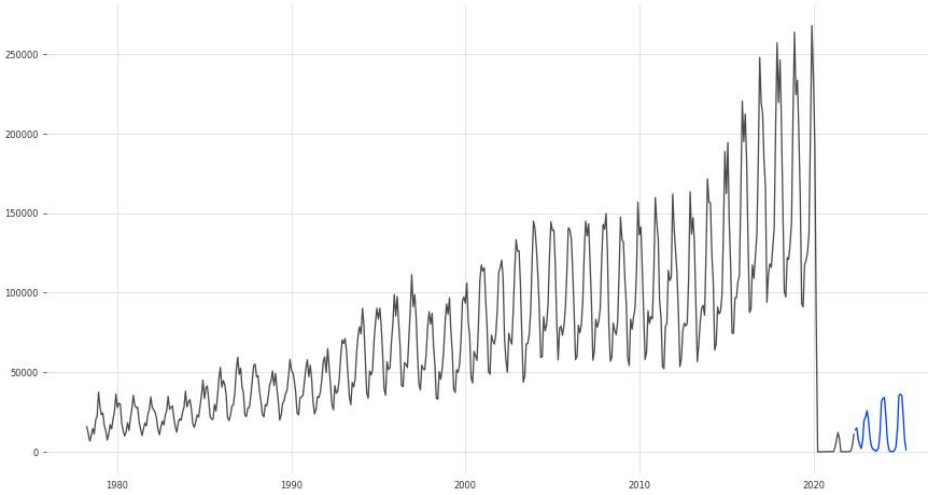

Figure 6 BATS Model of All Tourism to New Zealand

Because of this, I wanted to go beyond the pycaret recommendation of using Auto_ARIMA for the Queenstown Total Tourism data and use alternative methods in order to improve the modeling. I made two alternative models in an attempt to better the Auto_ARIMA model. First, I attempted using BATS on the Queenstown data. It modeled much more accurately than the Auto_ARIMA dataset. There were no negative values and it created a fairly accurate model on the historic data. With forecasting future data, it kept positive values with a decent representation of the seasonality (**Fig. 7**). The second approach was including exogenous variables. Exogenous variables are variables which can have an impact on the data being modeled but is not impacted by the data in the model. The exogenous variable of choice here was GDP. The hope was that due to GDP never falling to zero and going negative, we could apply that to our existing trend. Using an exogenous variable did not help with the negative forecasting due to the fact that the New Zealand GDP also collapsed greatly during the same time due to the pandemic. However, it did allow for the ability to forecast an upward trend becoming less negative over time. Given more data, I believe that the exogenous variable could have had a positive effect on the data (**Fig. 8**).

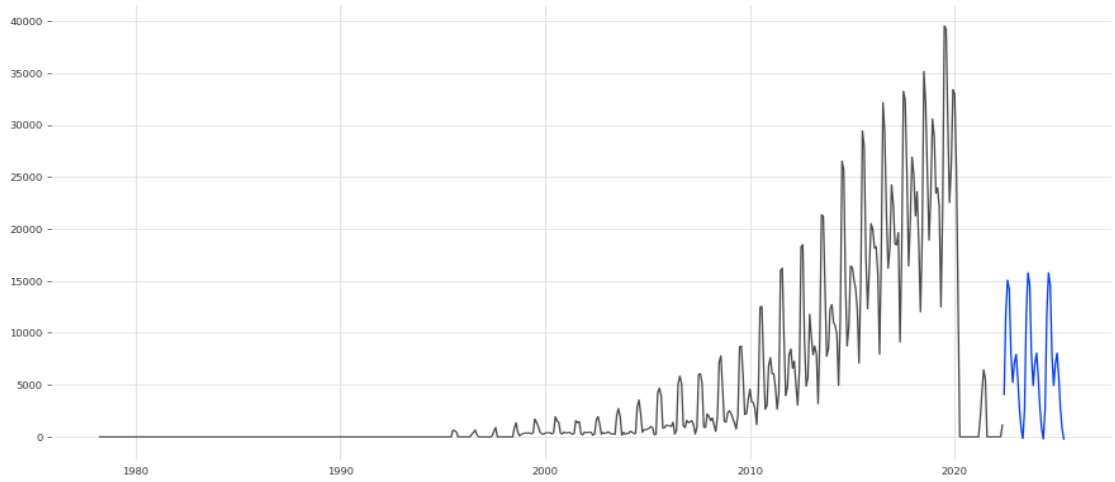

Figure 7 BATS Model on the Queenstown Total Dataset

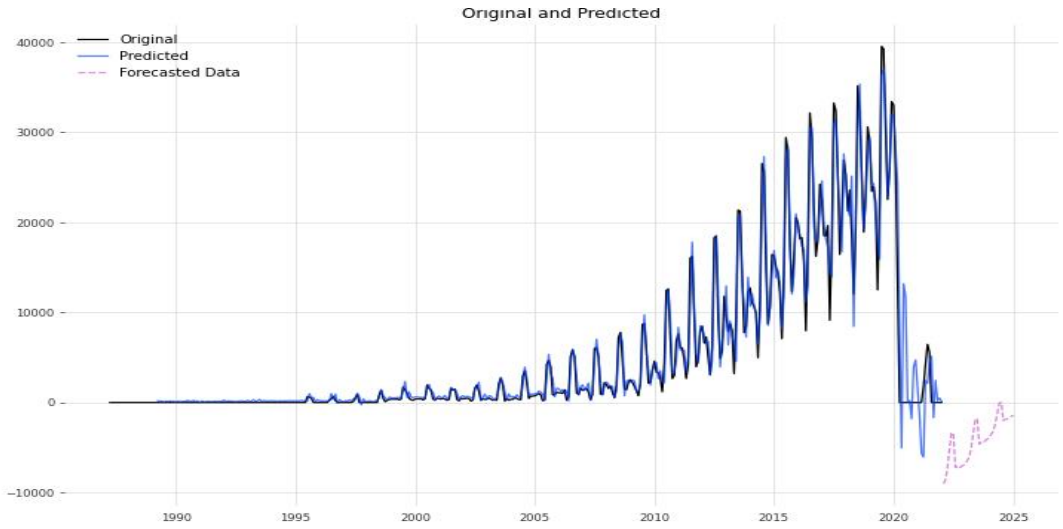

Figure 8 Attempt at Utilizing Exogenous Variables for Queenstown Total Data

After this, I modeled the accommodation dataset. Overall, the modeling went smoothly due to the lack of extremity in the data. For some of the machine learning models utilizing random forest, AdaBoost, and K-Nearest Neighbors, I utilized custom functions wrapped around the sklearn package. Part of that code was written by Michael Brownlee and posted on the blog machinelearningmastery (https://machinelearningmastery.com/random-forest-for-time-series-forecasting/). Due to the lack of pre-packaged models this was the best approach. With these machine learning models, especially AdaBoost, I found that it would converge to a mean quite quickly which is typical for longer forecast but not something I saw in other models (**Fig. 9**). This could be due to how the AdaBoost algorithm treats data and thus will converge to a mean quite quickly especially for data of high variance. For random forest and LightGBM, I discovered DARTS has these as part of its forecasting package and I utilized DARTS for these forecasts. The modeling process followed the same process as the tourism arrivals dataset. Overall, the models did really well and matched the pycaret expectations of performance. With Prophet, there was additional manipulation needed in order to get it to forecast monthly as it is generally used as a daily forecaster.

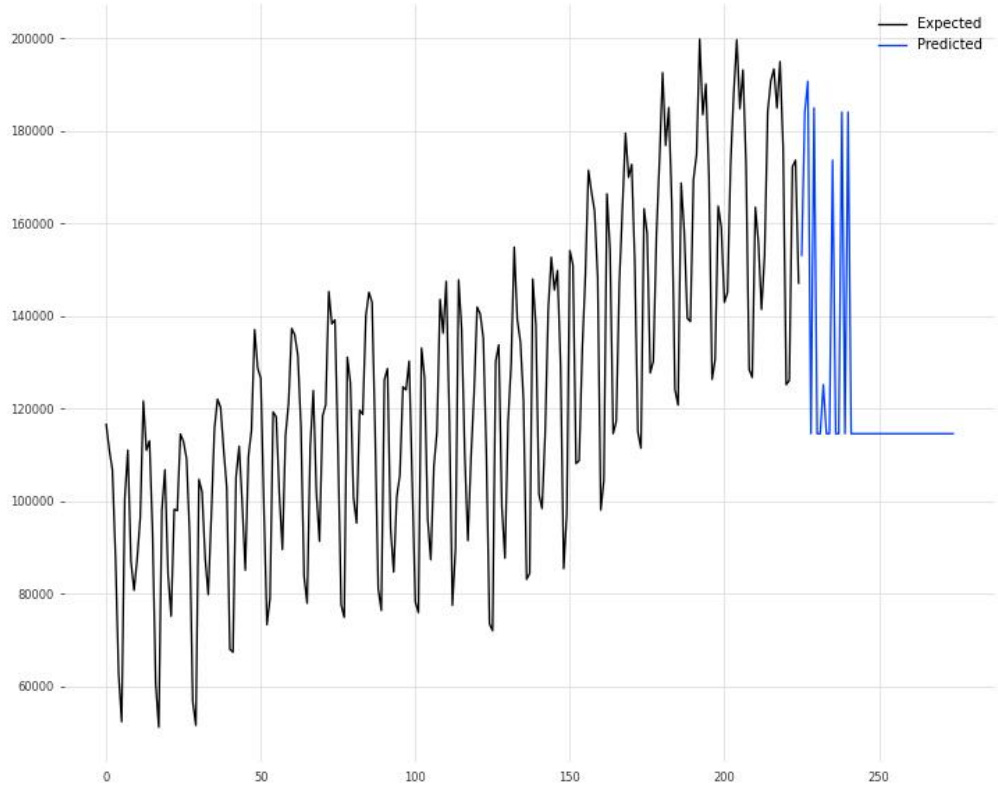

Figure 9 AdaBoost Model for Total Queenstown Accomodations Converging to a Flat Line

The most interesting model was the hotel data for Canterbury. This is because Christchurch is within the Canterbury region and in February of 2011 there was a major earthquake. Because of that, we see a huge drop in the Canterbury hotel data. While it is not quite on the same magnitude as the COVID-19 data as it does not go quite to 0, we see that the sudden negative trend had an impact on the forecast. There was a bit of a delay to the forecast until it caught up. However, as the trends returned to normalcy our forecasts were able to adapt

back to normal. This gives me hope that within a few cycles of data our forecasting ability will return especially if we continue to use models like BATS. As we can see in the historical data (**Fig. 10**) and the forecasted predictions (**Fig. 11**) there is the potential for time series models to regain their ability to accurately forecast trends once we are able to get more data back into them.

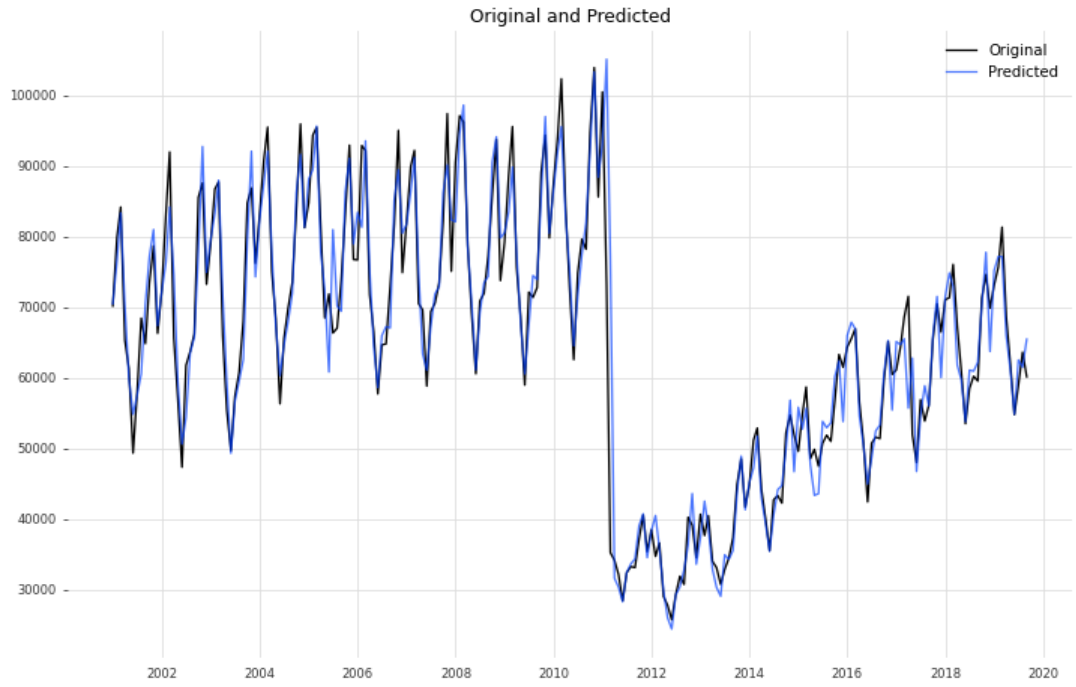

Figure 10 Comparison of Exponential Smoothing Forecast to Actual Canterbury Hotel Data

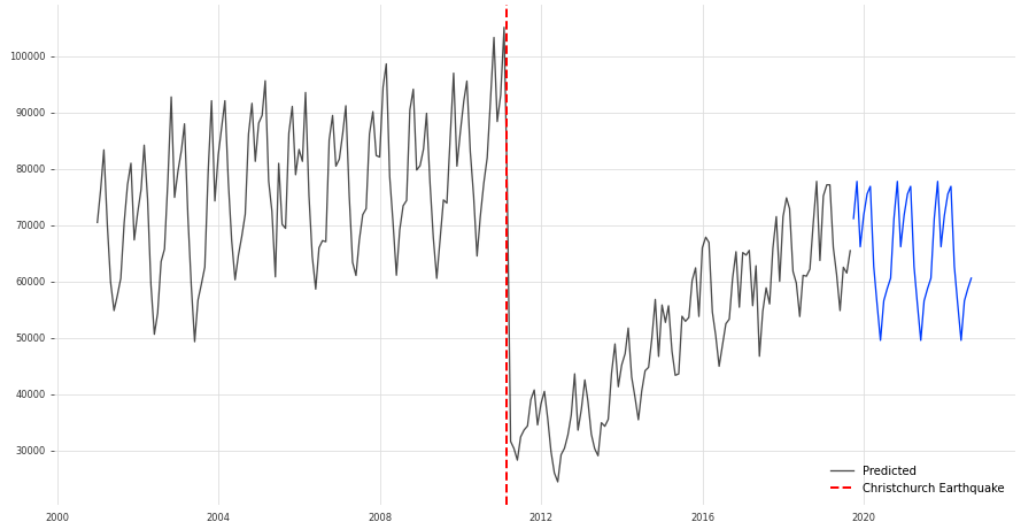

Figure 11 Forecasting for Canterbury Hotel Region Showing Only Model Predictions

Results and Discussion

The COVID-19 pandemic was truly unprecedented in many ways including for the world of time series analysis. It has created new problems for these algorithms that have not been seen before. Having the ability to work with such a variety of time series models has been a great exercise in understanding how these models react to the extremities of a pandemic. I learned a lot from my dataset and can give some recommendations based on my modeling. Firstly, I would recommend that the New Zealand government utilize BATS modeling for any datasets greatly impacted by the COVID-19 pandemic leading to a multitude of 0's. Secondly, I would recommend preparing for a ramp-up in visitors and influx of tourism. Lastly, I would make sure to input data into these datasets as quickly as possible so that the time series models are able to grab onto that data and create more accurate forecasts. Seeing how quickly the time series forecast was able to being minimizing erroneous forecasts given historical data truly demonstrates the power of these algorithms to understand trends.

That goes into my recommendation for the future. While more frequent updates take more time and resources, making sure that time-sensitive forecasting has the most up-to-date information is vital to its accurate forecasting. Some of the data used is approximately three years old. The tourism dataset is only updated annually even though there are monthly frequencies available. Creating a pipeline which allows for more frequent updates will allow for more accurate and powerful forecasting especially as we exit this pandemic.

Using pycaret was a great way to being initial modeling and I may also implement using DARTS for more machine learning and neural network modeling. These packages allow the usage of multiple models at once with minimal input from the user. They even go so far as to perform model optimization and with pycaret you can export the final optimized model for usage on your data immediately. Still, there can always be model improvements. I believe finding the time to collect exogenous variables and tailoring to the data can be a step in making models more accurate. Additionally, fine-tuning hyperparameters a little more when using the final pycaret model may improve how we can generalize using our forecasting models. Time series models are a powerful way to see the future of your economy but are only as useful as the user's ability to ensure proper modeling.