Impact of COVID-19 on international tourism expenditure

Byron Smith, 300504707

Group Members: Lauren Halka, Iona Sammons

Date of submission: 16-Oct-20

Executive Summary

The goal of this project is to answer the question, 'How much international tourism expenditure was lost due to non-New Zealand passport holders being prevented from entering New Zealand due to COVID-19?', using ARIMA models.

I first extracted and prepared the relevant data concerning daily arrivals of non-New Zealand passport holders. This data subset has two columns, 'Date' of type date and 'value', a discrete numerical variable. The dates range from the first day of 2016 to 14 October 2020.

I determined the date that COVID-19 started visually using a chart showing each year's daily values overlayed on each other. This date was 7 Feb 2020. I split the data into a before and after set and converted the before data to time series format.

I then used 'auto.arima' to find the best ARIMA model for this data. Using this model, I forecast the values from 7 Feb to 14 Oct 2020. The total of all the forecasts is 2,451,207, which interpretable as the estimated number of non-New Zealand passport holders that would have entered New Zealand had COVID-19 had no effect. The 95% confidence interval for this is (1,841,415, 3,060,998), quite high due to adding the confidence intervals for each daily prediction.

To tighten the 95% confidence interval I repeated the above steps of fitting an ARIMA model but with monthly values. This produced a total of all monthly predictions for the period of 2,593,293 with upper and lower 95% confidence intervals of 2,354,225 and 2,832,360.

Using this prediction I subtracted the actual number of international arrivals from the predicted number of arrivals. The difference is 2,008,450 with the 95% confidence interval (1,769,382, 2,247,517).

I used date from the tourism satellite account 2019 (TSA), which international tourism expenditure per year. Using this and arrival data I could determine the international tourism expenditure per arrival for the years 2016 – 2019. As I had limited historical data I could not accurately the international tourism expenditure per arrival for 2020 so I made the assumption that the increase in this value would be the same from 2019 to 2020 as the increase from 2018 to 2019 which was NZ$181. This gave me an estimated value for 2020 of NZ$4,359.

The international tourism expenditure per arrival multiplied by the difference between actual and predicted non-New Zealand passport holder arrivals produced the final answer of NZ$8,754,831,515 in international tourism expenditure lost due to non-New Zealand passport holders being prevented from entering New Zealand due to COVID-19, with the 95% confidence interval (NZ$7,712,735,084, NZ$9,796,927,946).

Background

The aim of the initial project was broad with the only concrete guidelines being the use of time series analysis and the COVID-19 portal data set. As a group we performed EDA on the entire COVID-19 portal data set and found that arrival and departure data had frequent observations (daily) which covered a large time period (5+ years). This amount of data points is suitable for constructing ARIMA models. Using this data, we could look at the difference between actual arrivals / departures compared to predicted arrivals / departures given COVID-19 had no effect.

The arrival / departure data also contained sub-categories, such as city and passport type. I started to look for specific questions that could be answered using these sub-categories. While looking for additional data relating to arrivals / departure I found information about international tourism expenditure which is closely related to a sub-category of arrivals, non-New Zealand passport holder arrivals. This data could be used with the arrivals data from the COVID-19 portal dataset to answer the question, ‘How much international tourism expenditure was lost due to non-New Zealand passport holders being prevented from entering New Zealand due to COVID-19?’.

To answer this question I need carry out a series of steps:

- Predict the number of arrivals of non-New Zealand passport holders had COVID-19 had no effect.
- Find the difference between this predicted number of arrivals and the actual number of arrivals.
- Determine the international tourism expenditure per arrival.
- Determine the loss of international tourism expenditure.

Data Description

The data used for this project is a subset of the COVID-19 data portal dataset, and data from the tourism satellite account 2019 (TSA), both sourced from StatsNZ.

The COVID-19 portal dataset contains time-series data about numerous sectors and industries that are in or have an effect on New Zealand.
The dataset is structured in a tree like pattern, where there are four main classes, which each have a number of categories with a number of indicators. Some indicators also contain subcategories which splits the data into categories such as City or Type of Passport. This is represented in the dataset with columns called ‘class’, ‘category’, ‘indicator_name’ and ‘sub_series’. Some indicators also have multiple series which cover the same time period.

For example the indicator `Card transaction total spend` has, as sub categories, the actual amount spent, the seasonally adjusted amount spent and the number of transactions. There are three columns which contain the actual time series information. These columns have the date, value and unit of measure. The values of different indicators have a number of units of measure including dollars, tonnes, percentage, percentage per annum, index and number. The last column is the date that the data for the indicator was last updated. This appears to apply to all the indicator values for all dates and not just the most recent value added to the indicator.

The dataset has 76 rows that have missing values in the `value` column. These missing values are mostly limited to three indicators and cover a certain time period for each indicator: Electricity Grid Demand (Oct 2003 - Mar 2005), New Jobs Posted Online (Oct 2003 - Dec 2004), Christchurch Heavy Vehicles (03 Dec 2018 - 05 Mar 2019). There is also one row with a missing value for the indicator, Fuel Supply.

The data of interest for this project is the indicator 'Daily border crossings – arrivals' and 'Daily border crossings – departures'. These indicators are contain a number of sub-series name, specified by the 'series_name' column breaking both indicators down into passport type, city and total. Passport type has two levels (NZ passort and non-NZ passport) and city has five levels (Auckland, Christchurch, Queenstown, Wellington and Other). After extracting this subset of the data there are a number of columns which are not useful as they are either used for higher level filtering ('class', 'category') or have the same values for all rows in this subset ('sub_series_name', 'units', 'date_last_updated').
We are left with four columns:

- 'indicator_name', which specifies whether the row concerns departure or arrival data.
- 'series_name', which specifies the sub category. eg. Total, City, Passport Type.
- 'parameter', the date.
- 'value', number of arrivals / departures.

These four columns are of type character as they are but it is clear that 'parameter' represents a date and 'value' represents a numeric value. This subset does not contain any missing values.

The tourism satellite account 2019 (TSA) is represented as an excel spreadsheet with multiple sheets. Each sheet has yearly data about a different aspect tourism in New Zealand. Some examples of the data the sheets contain are, tourism expenditure by type of tourist, tourism expenditure by product, cruise ship expenditure and arrivals into New Zealand by country of origin.
The sheet that is of interest is 'Table 2' which contains tourism expenditure in New Zealand from 1999 through to 2019, broken down into tourist type (international or domestic). This sheet also contains the percentage change of expenditure by type of tourist and international tourism as a percentage of total exports. There are not missing values in this data.

Ethics, Privacy and Security

Ethical Considerations

Ethics in the data science process, at its core, revolves around the idea of ensuring that any data associated with a given project is ethically sourced and used. In the context of the COVID-19 portal dataset, this in turn, firstly, involves considering whether informed consent was collected prior to the gathering/integrating of new and existing data, to ensure that all of those individuals represented in the figures and statistics of the portal, are aware that data linking back to them is being used and for what reason.

As the COVID-19 portal dataset brings together data from multiple external sources (as well as utilising its own), ensuring informed consent was received in this particular setting is clearly questionable. The process of tracking down and informing each individual (who contributed their data to the portal) of the prospective use and distribution of their data would be an extremely complex and timely exercise. Because the data in the portal itself takes an aggregate format, and thus does not present data of individuals, it could also be questioned whether the above detail in terms of informed consent is required. As the portal utilises data from other government organisations, it could be assumed that appropriate agreements have been made between each source and those individuals whose data were collected, to distribute it for research purposes.

Privacy Considerations

To ensure that we have adequately taken all privacy considerations into account we should refer to the law governing data privacy in New Zealand, Privacy Act 1993. This act has 12 principles that need to be considered when handling data. As we are not collecting the data ourselves and the data does not identify any single party, as stated by StatsNZ on their Privacy, Security and Confidentiality page, we can be confident using this data will be in line with all regulations.

Security Considerations

In order to ensure the security of our data and research, we should ensure the Confidentiality, Integrity and Availability of all our work. The research will be carried out on a personal machine which is only accessible to myself, with all files and data kept locally, however at times the project will need to be given to others.

Confidentiality

To preserve confidentiality of the files on my personal machine I use a strong password which is resistant to brute force attacks. This password consists of a mix of symbols, numbers and letters and should theoretically take 268 million years to guess using brute force methods. The other attack which could potentially allow un-authorized access to my files is a phishing attack where I am tricked into giving out my password. To prevent this I

need to be aware of these methods and ensure I do not give out my password under any circumstances.

When sending files to others I will encrypt the files and supply a password to the recipient in person to ensure that access to the files is only possible for those it is intended for.

Integrity

I can be sure that the integrity of the files on my local machine is maintained by following the steps concerning confidentiality.

Integrity of the files will be most at risk when sending the files. A man in the middle attack could be carried out as the files are en-route to the recipient. To prevent this, I will first encrypt the files with the recipients' public key and then encrypt this with my private key. This will ensure the integrity and confidentiality of the contents. If the recipient does not have a public key, I will deliver the files in person.

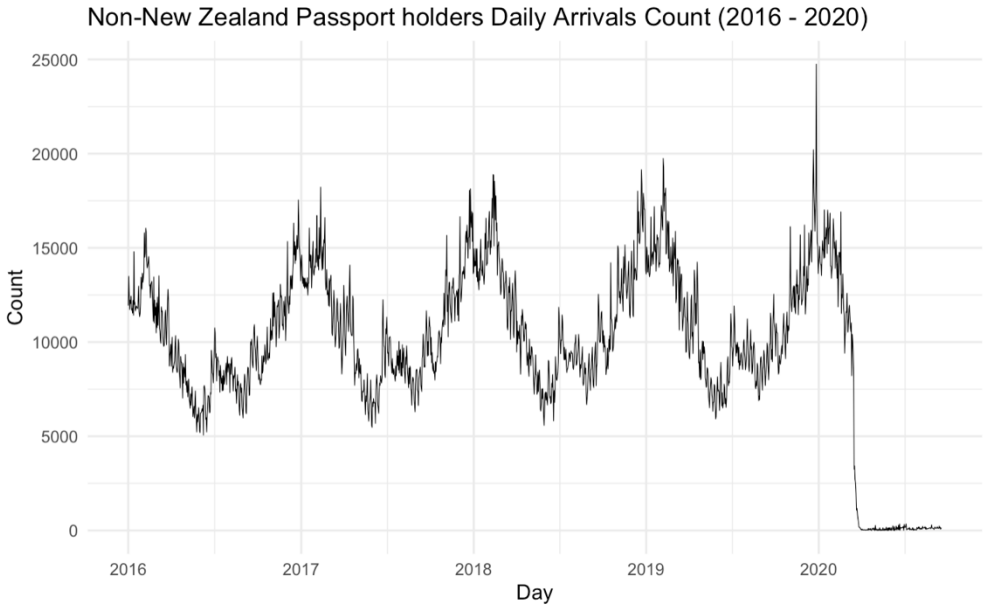
Availability

As the research files and data are kept on my personal machine there will be no availability issues as long as this continues functioning correctly.
The COVID-19 portal dataset is updated daily, and I will be downloading the most recent data periodically. There are no apparent issues with the availability of being able to download the data and potential problems cannot will need to be addressed by StatsNZ.

EDA

This exploratory data analysis is focused on the subset of the COVID-19 portal dataset concerning arrivals and departures from New Zealand, and the data contained in the sheet named 'Table 2' in tourism satellite account 2019.

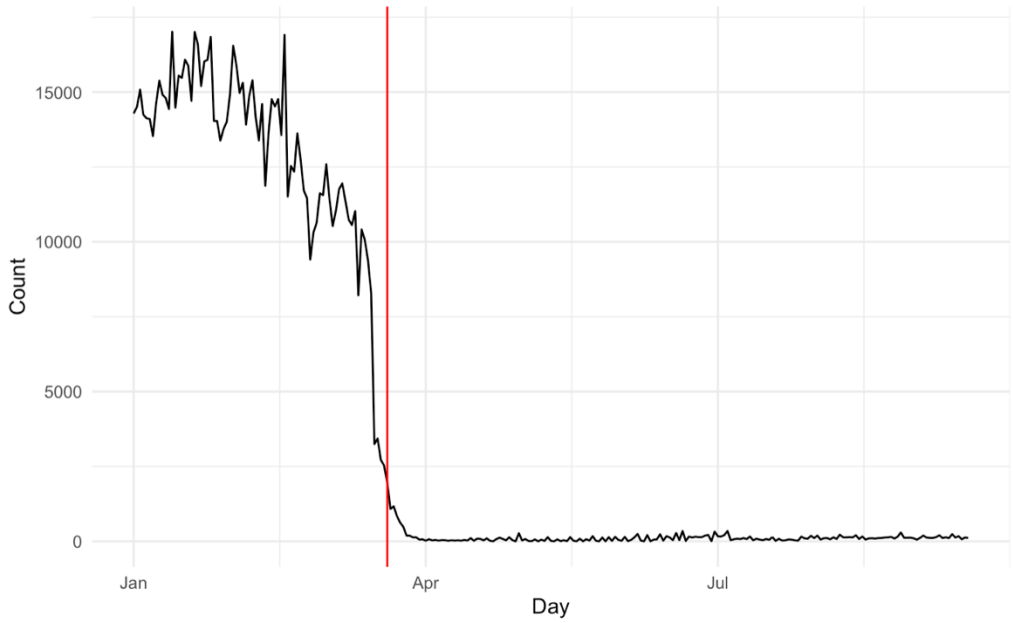
The question, how much international tourism expenditure was lost due to non-New Zealand passport holders being prevented from entering New Zealand due to COVID-19?, can be broken down into two parts. Firstly we will need to forecast the expected international arrivals, had COVID-19 had no effect, and secondly use the predicted arrival count versus actual arrival count to determine the international tourism expenditure loss. We will start the EDA by looking at the data concerning the first part of the question, relating to non-New Zealand passport holders arrivals.



We can see some clear seasonal trends in the number of arrivals of non-New Zealand passport holders. The number of arrivals peaks twice each year, at the end of the year (December) and again at the start of the year (January/ February). At the peaks the arrivals count is around 12,500 more than the lowest count for that year. The lowest point for each year is consistently just before the half year (June). We can also see that arrivals counts increase steeply between June and July and then begin to decrease steadily until September.

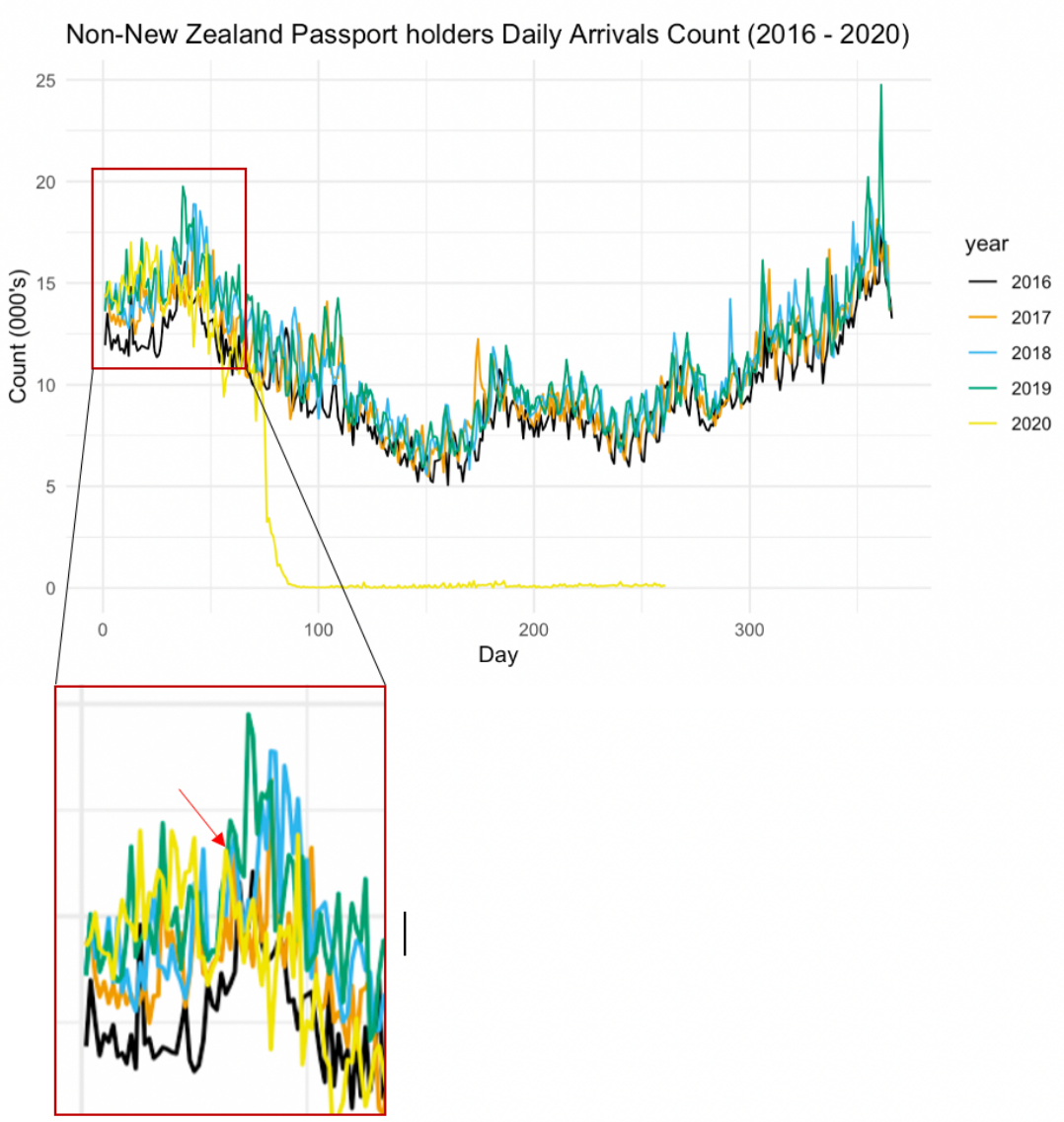
In December 2020 we can see that the arrivals count was significantly higher than all other counts. Looking at the arrivals data by city for this day the count for 'Other' is much higher than usual, with the previous maximum being around 4200. This could indicate that this peak is caused by cruise ships docking at city's' ports that are part of the 'Other' category. It is hard to determine whether this is a mistake or caused by a number of factors that we do not know of.

We can also see a sharp decrease in arrivals in March 2020. The chart below shows arrivals for 2020 only and the red line indicates the day that border restrictions began (20 March).



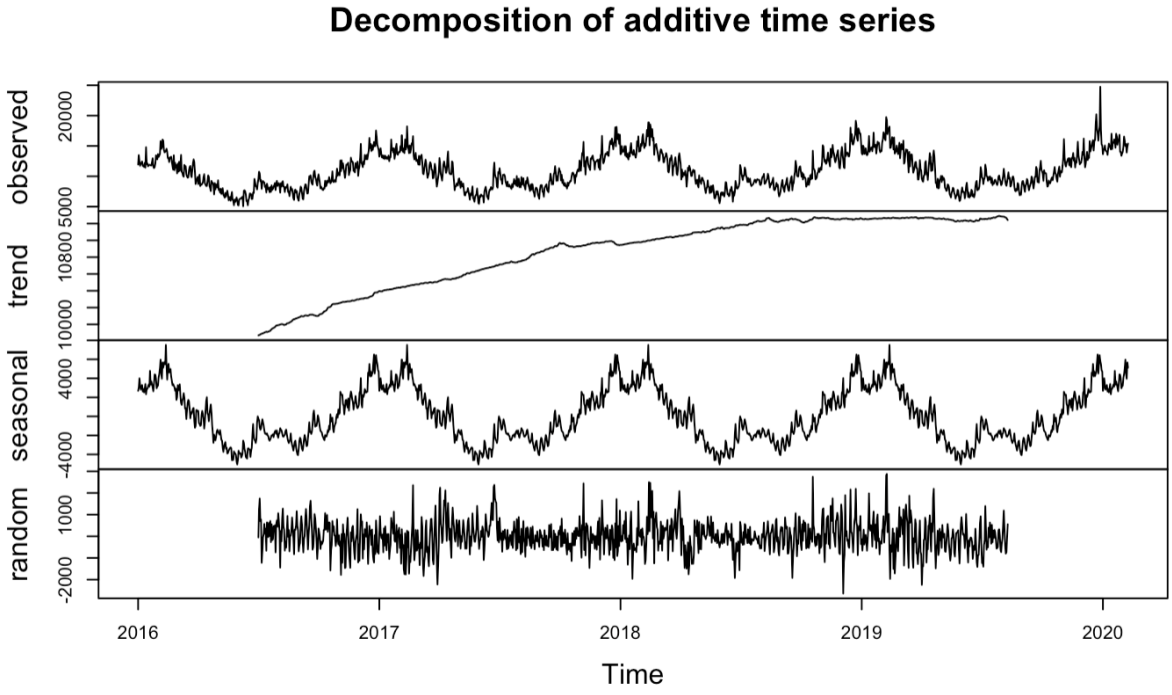
This clearly shows that the introduction of border restrictions was not the only factor influencing the decrease in international arrivals.

It is important for us to know when COVID-19 started to have an effect on the number of non-New Zealand passport holder arrivals. As we saw above it was not only affected by border restrictions. There may be a certain announcement or event that triggered the decline but it is not important for us to know in order to be able to answer the question. We can just look at the daily arrivals for each year and visually determine the date that arrivals began to deviate from the trend.



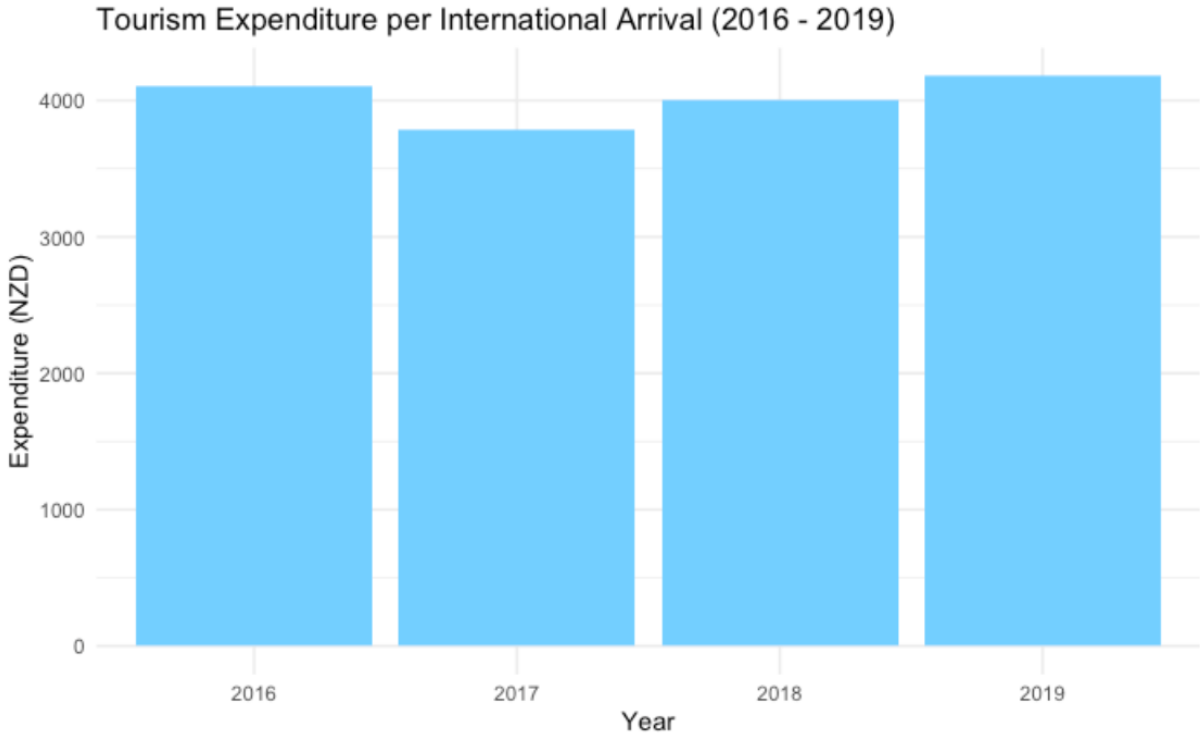
We can see that the arrivals count in 2020 is relatively similar to previous years until the 7th of February, shown by the red arrow. In the detailed analysis we can use this date as a cutoff between the values and dates we use for building a model and the dates for which the values we are predicting.

Now that we know the data that will be used to build the model, we can analyze this on its own.
Time series data can be broken down into three components, trend, seasonal and random. Let's look at these three components for the data for before the cutoff date we identified above.



The trend component shows that the number of international arrivals is increasing steadily from 2016 to late 2018 and then plateaus. The seasonal component shows a clear seasonal trend which is not surprising as we identified a number of seasonal trends in the first part of the EDA. The random component shows that the number of international arrivals is influenced by other factors unrelated to the general trend and the seasonal trend.

The second part of the question concerns international tourism expenditure. To answer this we will need to know the average amount spent for each non-New Zealand passport holder entering New Zealand.



This chart is showing the average tourism expenditure for each non-New Zealand passport holder that arrives in New Zealand. We can see that this average varies slightly year by year. We are limited by the amount of historical data that we have to accurately determine the trend and find an expected value for the average expenditure by international arrival in 2020. This will lower the confidence we have that our final answer is correct.

Detailed Analysis

In the EDA section, we extracted the data for the number of arrivals of non-New Zealand passport holders from the COVID-19 portal dataset. To prepare this for use we first need to filter out the values corresponding to the dates we wish to predict (all after 07/02/20). We then need to convert this data into a time series for compatibility with the R functions we are going to use. To make our data compatible we need to create a time series which can be done using the 'ts' function, where we give the parameters, values for each date, the start date and the frequency (observations per unit of time).

To forecast the values from the date 07/02/2020 we will use an ARIMA (autoregressive integrated moving average) model. An ARIMA model has three parameters *p*, *d* and *q* where:

- p is the number of autoregressive terms;
- d is the number of differences; and
- q is the number of moving averages.

We can use the 'auto.arima' function from the 'forecast' package in R, on our data, which performs a stepwise selection of the p, d and q parameters, for both the seasonal and non-seasonal parts of the model to find the best ARIMA model specification for the data.

The model we get from using 'auto.arima' is the following:

```
Series: arr_full
ARIMA(5,1,5)(0,1,0)[365]

Coefficients:
          ar1      ar2      ar3      ar4     ar5     ma1      ma2     ma3      ma4      ma5
      -0.7174  -0.1225  -0.5923  -0.6065  0.2118  0.0414  -0.4234  0.2478  -0.0615  -0.7465
s.e.   0.0596   0.0721   0.0778   0.0402  0.0516  0.0513   0.0666  0.0685   0.0543   0.0597
```

This model specification shows that the best model found using a stepwise search for the optimal parameters has five autoregressive terms and five moving average terms, for which we can see the coefficients here. The model specification also tells us that to make the data stationary we need to apply one seasonal difference with a period of 365 time periods (days) and one non-seasonal difference.

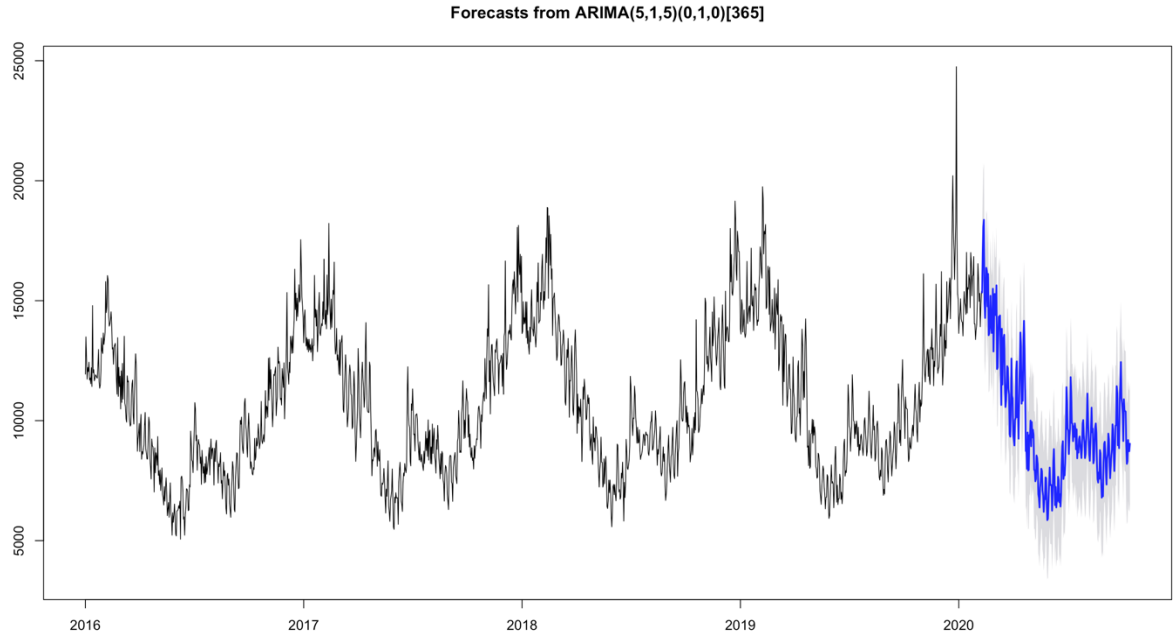
Training set evaluation metrics:

```
                      ME     RMSE      MAE
Training set -43.37887 957.5852 610.8169
```

The Mean Error of the model is relatively low at -42.38. This metric is appropriate here as we are not worried about positive and negative errors cancelling each other out but welcoming it, as we will be using the sum of the predictions to answer our question. However we cannot count on this metric to hold when performing predictions. The Mean Absolute Error is much higher (610.82) and is more informative about the models

performance. Our mean value is 10865.74, meaning the MAE is around 5.62% of the true value.

To see what the predictions look like we'll plot predictions from 08/02/2020 to today's date (14/10/20), 250 predictions, and include the 95% confidence interval of the predictions.



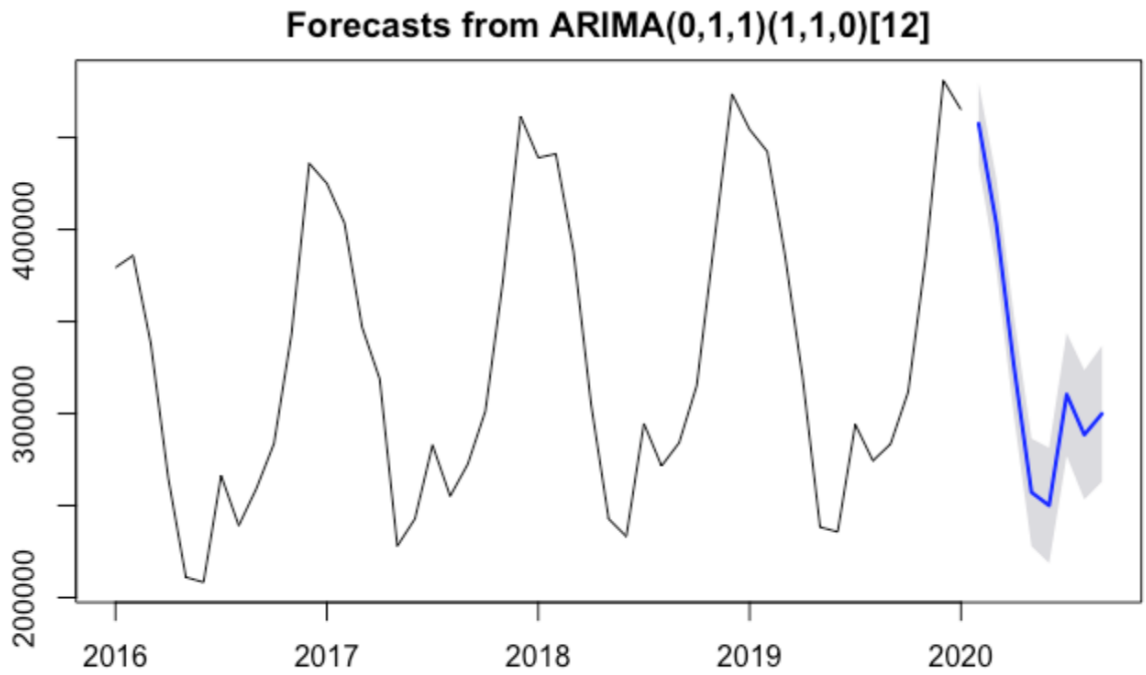
We can see that the predictions look reasonable and the 95% confidence prediction intervals are relatively tight around the predictions. The total value of the predictions is 2,451,207. The lower and upper 95% confidence interval totals are 1,841,415 and 3,060,998. We can see that the prediction intervals add up quickly and leave us with a difference between the lower and upper CI of over 1 million.

As it is only necessary to know the total, sum of all the predictions, predicting the daily value will decrease the confidence of our prediction total, as we need to take into account all the prediction intervals. We can instead use monthly values which will decrease of the number of predictions we need to make from 250 to 8, and also the corresponding prediction intervals.

Using monthly values, 'auto.arima' finds the best model to be:

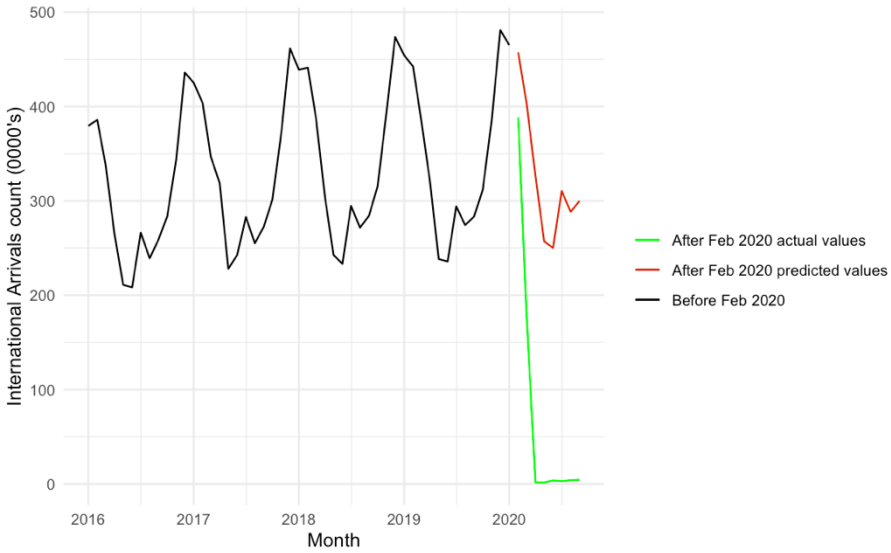
```
Series: arr_ts_month
ARIMA(0,1,1)(1,1,0)[12]

Coefficients:
          ma1     sar1
      -0.4843  -0.6576
s.e.   0.1975   0.1359
```

The interpretation of this model is that the model contains one moving average term, one seasonal autoregressive term and performs one seasonal and one non-seasonal difference. Plot showing monthly predictions for February through to September with 95% confidence interval:



We can see again that the predictions look reasonable and the confidence intervals are relatively tight. The total arrivals for the predicted period of 8 months is 2,593,293 with upper and lower confidence intervals of 2,354,225 and 2,832,360. The difference between the upper and lower confidence interval is 478,135. Using monthly values has given us a much smaller confidence interval than using daily values.

Plot showing actual vs predicted values:



Looking at the plot above, we can see there is a large difference between the predicted and actual values. We need to determine the difference between the predicted values and the actual values which will give us the total number of non-New Zealand passport holders prevented from entering New Zealand due to COVID-19. The difference is 2,008,450 with the 95% confidence interval (1,769,382, 2,247,517).

We now have our predictions and corresponding confidence intervals so we will now continue to the second part of the analysis in order to answer our question, 'How much international tourism expenditure was lost due to non-New Zealand passport holders being prevented from entering New Zealand due to COVID-19?'.

In the EDA section we looked at international tourism expenditure per international arrival and saw that that due to a limited amount of historical data it was hard to identify a trend. In order to complete our analysis we'll assume that expenditure per international arrival would increase from 2019 to 2020 by the same amount as the increase between 2018 and 2019 which was NZ$181. This gives us an estimated value for international tourism expenditure per international arrival in 2020 of NZ$4,359.

Using our international tourism expenditure per international arrival estimate and the estimated difference of predicted and actual number of arrivals from February to September 2020 we can calculate the difference in international tourism expenditure for this period. It should be noted that this difference in expenditure will also assume that the international arrivals over the period that we predicted also maintained the international tourism expenditure of NZ$4,459. The calculated difference is NZ$8,754,831,515 with the 95% confidence interval (NZ$7,712,735,084, NZ$9,796,927,946).

Conclusions and Recommendations

Conclusion

We now have the final answer to our question, 'How much international tourism expenditure was lost due to non-New Zealand passport holders being prevented from entering New Zealand due to COVID-19?'. NZ$8,754,831,515 worth of international tourism expenditure in New Zealand was lost due to non-New Zealand passport holders being prevented from entering New Zealand, during the months February to September, due to COVID-19. This assumes that increase in international tourism expenditure per international arrival was the same between 2019 and 2020 as the increase between 2018 and 2019. We also assume that non-New Zealand passport holders that arrived in New Zealand between February and September 2020 contributed on average NZ$4,459 to international tourism expenditure.

Recommendations and future uses

These findings would be most useful for the Government to make high level decisions about the state of the tourism industry and assess the level of financial assistance that will be needed. These findings could be used with findings about other industries to determine which industry is most affected and most in need. Another use could be to use these findings to determine how dependent the New Zealand economy is on international tourism expenditure and whether or not there is any reason we should consider moving away from a strict health first based approached. These findings could also be used in another study about a different sector to determine how the loss of international tourism expenditure has had an indirect effect on that sector.

Limitations

During this study we were very limited by the amount of historical data that we had on international tourism expenditure per arrival. We had international tourism expenditure data spanning from 1999 to present but only had 4 years of arrivals data. As we needed to determine the expenditure per arrival we were limited to only 4 years of historical data. I decided to make an assumption which meant the final answer was also based on this assumption. To improve the final answer I could perform a more in depth analysis on the expenditure per arrival, which would produce a more reliable result. Another assumption that was made is that the non-New Zealand passport holder that entered the country between February and October 2020 were also maintaining the international tourism expenditure average but this may not be the case as the purpose of their visit is unlikely to be for tourism.