By Iona Sammons - 300281718
Written with Lauren Halka and Byron Smith

New Zealand’s lost passengers

A look into the impacts Covid-19 had on international arrivals

Photo credit: Auckland Airport - https://www.linkedin.com/company/auckland-international-airport-limited

Executive Summary

This report looks at the impact Covid-19 has had on New Zealand, in particular the reduction in international arrivals. It uses time series data to forecast how many international arrivals New Zealand would have had this year if Covid-19 did not happen.

It found there to be an estimated loss of 1.5-2 million foreign arrivals into New Zealand due to the disease.

Background

The goal of this project is to take a set of continuous time series data and model the relationships in them in order to develop a prediction model. The dataset used was the Covid-19 data portal dataset that has been developed and maintained by Stats NZ (Stats NZ, 2020). The source of the data is from numerous sources including New Zealand government agencies such as the Reserve Bank of New Zealand, The Treasury and the Ministry of Social Development, banks and other international sources.

The question that was looked at was “how many international arrivals did New Zealand lose due to coronavirus disease 2019 (Covid-19), over the recent months?”. It used the data on the number of arrivals across New Zealand's borders and focused on anyone who entered on a non-New Zealand passport. The prediction used border crossing data from 1st January 2016 till 30th September 2020 to estimate the number of foreign arrivals there would have been from December 2019 to September 2020 if Covid-19 had not occured.

This is of interest as it would contribute to our understanding of the impact Covid-19 has had on our country. The findings could contribute to answering questions such as “how much money has the New Zealand economy lost from a lack of international tourism?”

Data Description

The dataset as a whole

Types and structure

The dataset is a time series with many variables that mostly relate to the impact Covid-19 has had on New Zealand, with a few variables relating to Covid-19 and the rest of the world.

Most of the columns in the data set are dedicated to classifying the data. It is structured in a tree-like pattern and there are four main classes. Each of these classes have a subset of categories, and in turn those categories have a subset of indicators. Some indicators also contain their own series which classifies the data further.

An example of the classification of two lines of data:

Class	Category	Indicator	Series
Economic	Employment	Monthly earnings	All industries (actual)
	Travel	Daily border crossings - arrivals	New Zealand passport

Some indicators also have multiple series which cover the same time period. For example the indicator ‘Card transaction total spend’ has three data series for each time period; 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.

All the columns in the dataset were character types. The columns ‘parameter’ and ‘value’ needed to be changed to other types before these columns could be used. However, as some values were integers while others were floats they were only changed once the data was extracted into a subset. Similarly with the dates, the formats are generally the same but differ for certain indicators.

Completeness of the data

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.

As the dataset is so large it was initially hard to determine if there were obvious errors in the data. It was not until subsets of the data were looked at that further errors were picked up. However, the data can be visualised on the Stats NZ Covid-19 data portal tool. On there, no obvious errors were seen and the data would have also been looked at by Stats NZ employees too, so we can be relatively confident the majority of the data is error free.

The analysed data

Types and structure

The data for this analysis used entries that had all of these classifications:

Class	Category	Indicator	Series
Economic	Travel	Daily border crossings - arrivals	Other passports

The daily border crossings are for both air and sea ports and 'other passports' refers to people entering the country using anything other than a New Zealand passport. The types of people that this will capture include New Zealand residents, people entering New Zealand for work and tourists.

The time series information that is used in this are dates and total counts of arrivals. These numbers have been recorded daily, and go back to 1st January 2016.

Completeness of the data

There are no missing values, however, there is one point of data that could possibly be incorrect. On 27th December 2019 the count for 'other passport' arrivals was about 25 000. This is an outlier, because even in peak seasons the highest daily count was about 20 000. A one off case of a 25% increase of the previous highest count is unlikely. This value might skew any prediction models.

Ethics, Privacy and Security

Ethics

Ethics in the data science revolves around the idea of ensuring that any data associated with a given project is both sourced and used ethically. In the context of the COVID-19 Data Portal, this involves considering whether informed consent was collected prior to the gathering/integrating of new and existing data. This would be to ensure that all of those individuals represented in the figures and statistics of the portal are aware of what their data is being used for and for what reason.

The portal brings together data from multiple external sources, as well as utilising its own. This means it is difficult to ensure informed consent was granted during the data collection. On the other hand, the data in the portal itself takes an aggregate format, and thus cannot be used to identify individuals. Therefore informed consent is not necessarily required.

Furthermore, 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.

Another angle from which ethics should be considered in this case, relates to the application of Māori Data Sovereignty principles. The COVID-19 Portal mainly concerns data relating to the state of economic, health and social factors in New Zealand. Some analysis (Martin, 2020), (Edmunds, 2020) say Māori have been shown to be more likely to die from Covid-19, and that Māori women were disproportionately affected by Covid-19 job losses. Therefore, the consideration of rights and usage of Māori data is important, but this data does not appear to address that. The reason for this is possibly related to the fact that the data has been aggregated, and thus combined with the data of people from many other cultures too, so the principles may not have been considered important. However, this increases the likelihood of other areas where Māori are disproportionately affected being missed.

Privacy

Stats NZ ensures that data published does not identify individuals, households or businesses as stated on their Privacy, Security and Confidentiality page (Stats NZ, n.d.).

Theoretically, there might be other data sets that the Covid-19 dataset could be combined with that would allow individuals to be identified. However, the data has been aggregated in such a way that would make that very hard to do. It is highly unlikely that this data could be used to identify individuals.

Security

Because we are utilising a public data source, keeping project data and results secure is more about carrying out measures to prevent work from being lost, as opposed to doing so alongside protecting sensitive data from being leaked. Consequently, to do this, we will ensure we back up all progress when working with R code through committing it to a Github repository.

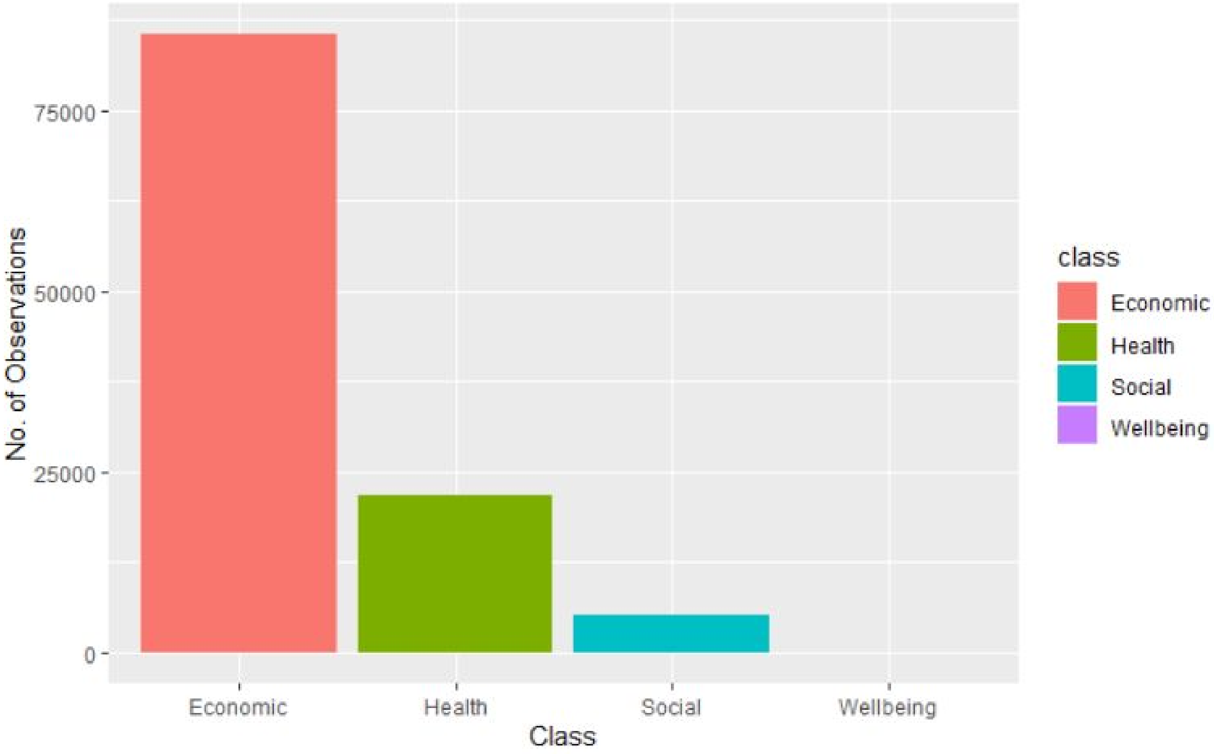
However, if we were working with sensitive data, to keep these results secure, we would take a further step and set our repository to private, ensuring that it can only be accessed by members of our team. In this scenario we would also have to be mindful of how we locally store the data we work on and communicate with our group members. It would have to be kept on non-shared devices with malware protection in place. Any communication that refers to the data or findings would be best done on a platform with end-to-end encryption.

Exploratory Data Analysis

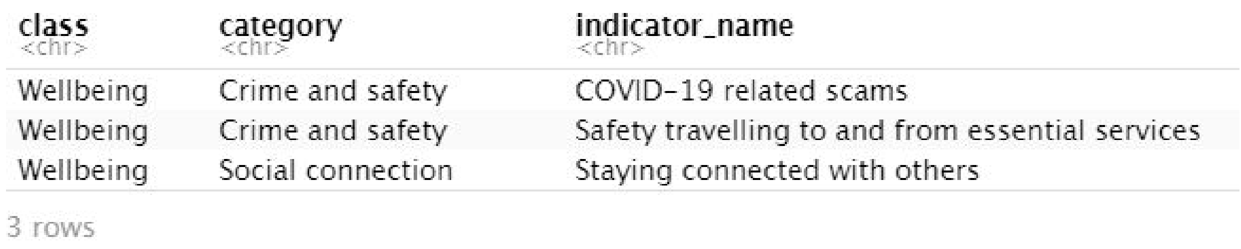
The dataset as a whole

Types of categories

As mentioned, this COVID-19 portal has its data structured like a decision tree, containing categorical variables to help group the key attribute; Parameter. The first defining attribute is that of Class, determining whether data relates to Economic, Health, Social or Wellbeing factors. Of the 112,382 observations in the cleaned version of the COVID-19 Data Portal (whereby rows containing NA values were omitted), the following visualisation shows how the classifications are distributed amongst all current observations.



The above goes to show that the clear majority of observations kept within this portal are associated with 'Economic' factors. 'Health' too contains a significant number of observations, which would be assumed as being associated with COVID-19 related statistics as that is what the portal was created for. Something interesting to note is the 24 "Wellbeing" related observations, which are so insignificant and hence do not appear on the graph. In Stats NZ's description, this classification is not one of the three mentioned, for which reason it should be further investigated, and potentially moved under the 'Health' category, if appropriate, to ensure it is accounted for in analysis.



class <chr>	category <chr>	indicator_name <chr>
Wellbeing	Crime and safety	COVID–19 related scams
Wellbeing	Crime and safety	Safety travelling to and from essential services
Wellbeing	Social connection	Staying connected with others

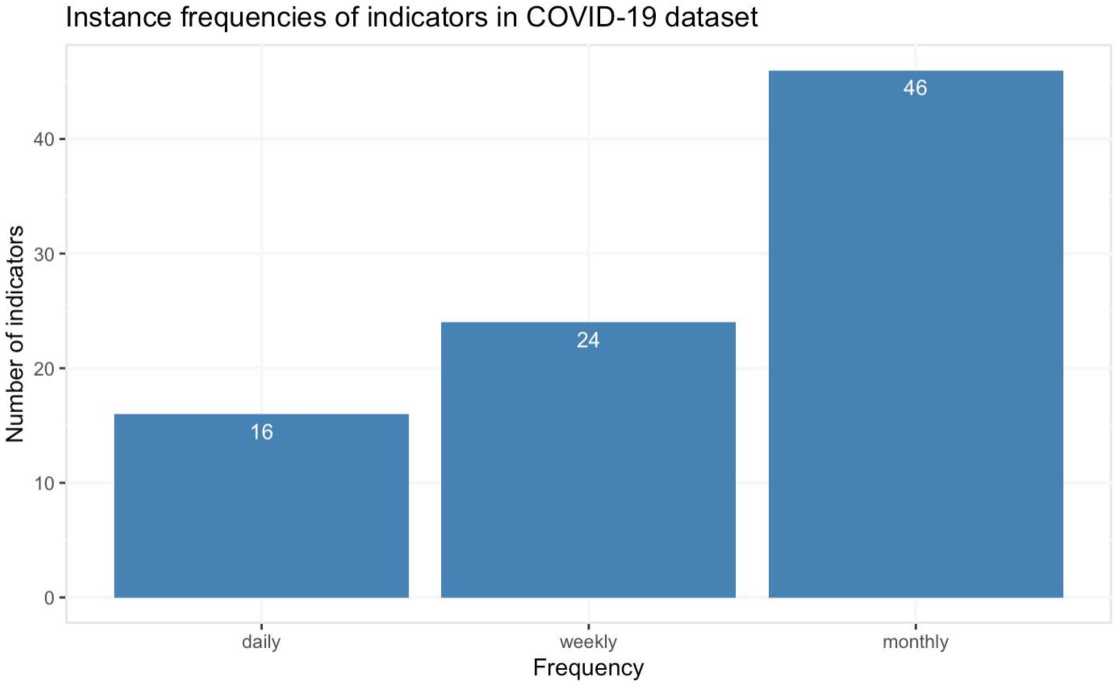
3 rows

Upon further investigation - results as shown in the above table - the 24 observations that were classified as 'Wellbeing', came under 3 distinct combinations of the attributes 'class', 'category', and 'indicator_name'. When considering they fall under categories associated with crime and social connection, it would make sense to instead move all of these observations under the 'Social' class.

Frequency of indicators

Further along those "tree branches" is the Indicators attribute. The COVID-19 data portal data set is broken down into over 90 different indicators that contain time series. First we determined the frequency of the indicators, finding out whether the indicator has daily, weekly, or monthly data. As some indicators have multiple series, we only looked at the first series for each indicator and assumed that each series is the same in terms of frequency. This meant that when using a specific indicator at a later date we needed to double check that the information we have on the indicator is correct.

We can generate a graph showing the frequencies of the indicators:



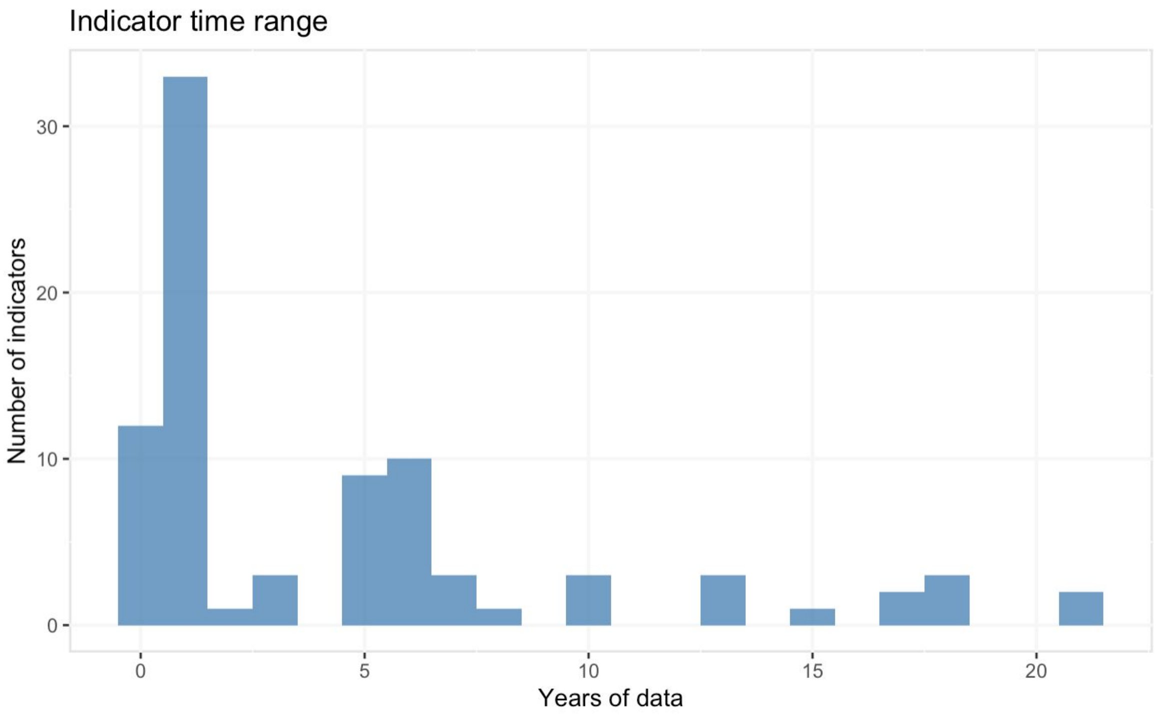
This shows that the majority of the indicators have a monthly frequency, with 46 monthly indicators, 24 weekly indicators and 16 daily indicators. It should be noted that indicators whose parameter (date) was not in the standard format were removed, removing 9 indicators. On further inspection it was because these indicators were not time series data. This included indicators such as “source of cases” where the parameters were sregions in New Zealand.

It was decided at this point that we wanted to minimise using indicators of different frequencies, especially daily and monthly, so that we can preserve as much of the information as possible. For example, if we were to merge daily with monthly, we would first have to get the total of the daily values for the month, losing any daily variation information.

Time range of indicators

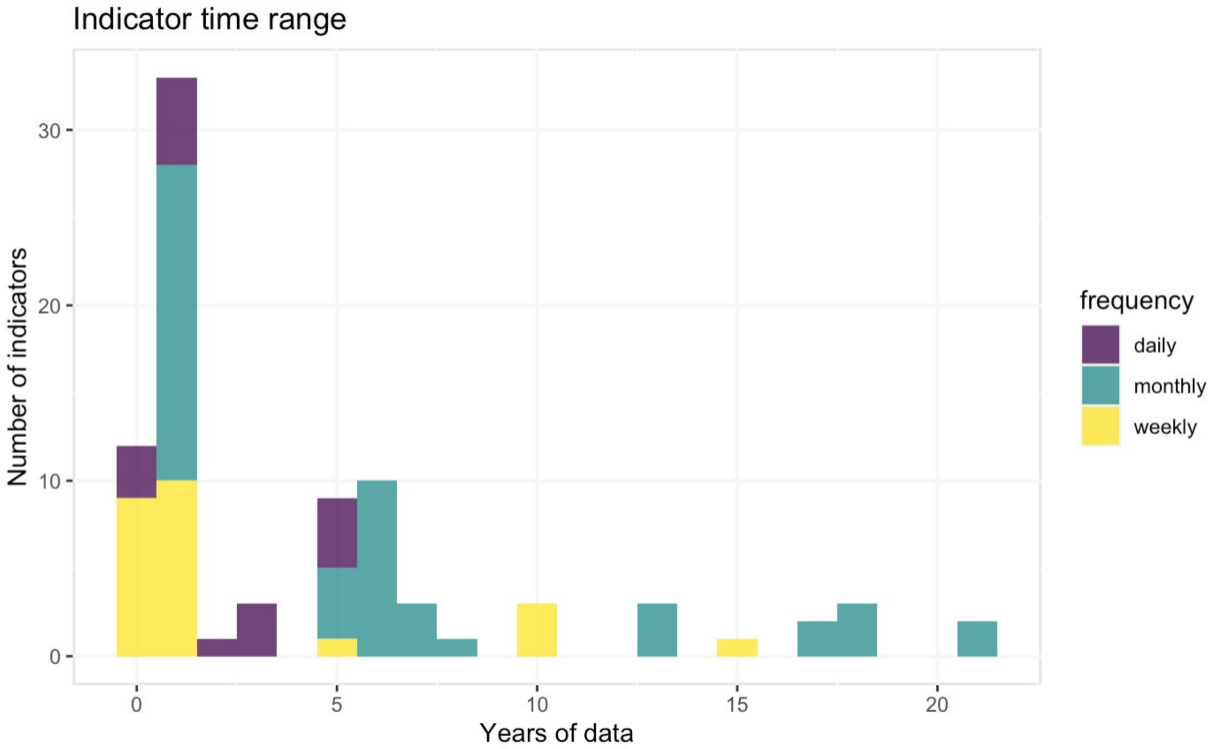
The next important detail about the indicators is the time frame that they cover. The COVID-19 data portal dataset is continuously updated and all the indicators have information up to today's date, so we will need to find the start date to determine the time frame. The length of time each indicator covered was added to the previous table to create more informative graphics.

Graph showing indicator time frame:



This shows how many years of data there is for each indicator. It appears that most of the indicators have less than two years of data. As we were planning on using ARIMA models to answer a question we needed to ensure that there was enough data for the indicators that we use the model on. As this plot only shows the time frame and not the frequency of the indicators it is hard to get an idea of the number of data points that exist for each indicator. So, we should include frequency in the plot to get a better idea of the number of data points.

Plot showing the time range and frequency of indicators:



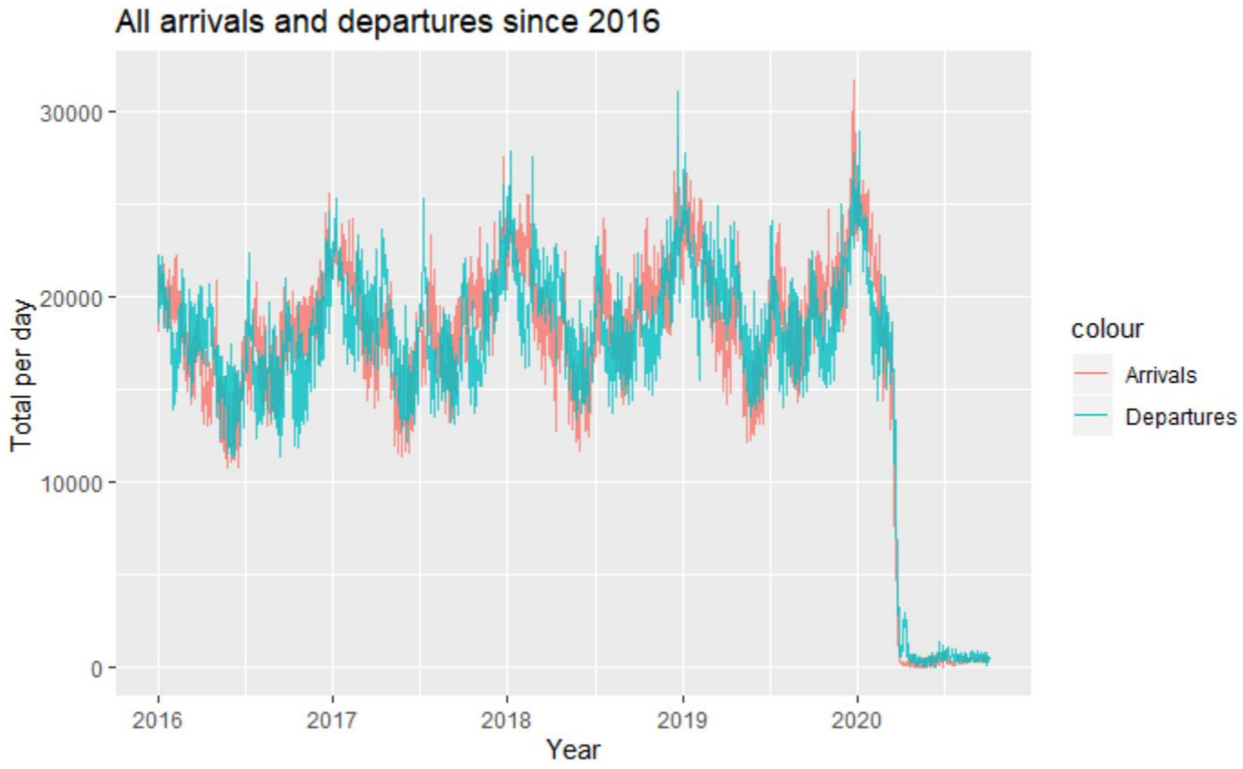
Some of the shorter time series indicators include information on the Covid-19 Income Relief Payment, which has only existed this year so we would not be able to do forecasting on that. The indicators at the other end of the scale are mostly to do with employment and spending. This includes total spends on card transactions and the jobs filled each month.

This gives us a better idea of the amount of data points for each indicator. When choosing indicators to work with for ARIMA models, it would be preferable to avoid indicators that have monthly frequencies for a short time period because that will not be enough information to do forecasting on. The indicators that we may want to use can be seen in the five year bar with daily frequency. One of these indicators concern arrivals and departures from New Zealand.

Border crossing data

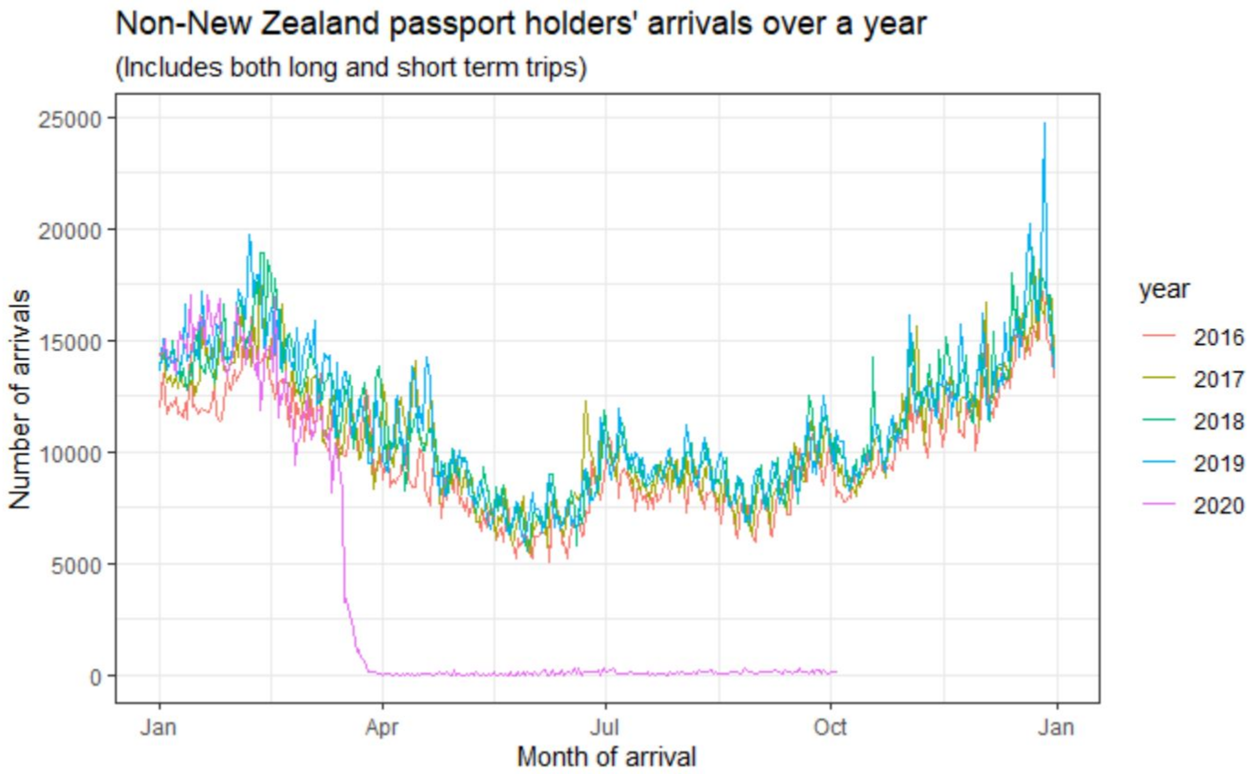
Due to the observations above, finding significant information will require some work. There is one variable however that has a clear change; the daily arrivals and departures in

New Zealand. This data also goes all the way back to 2016, so there is enough data to make some forecasts.

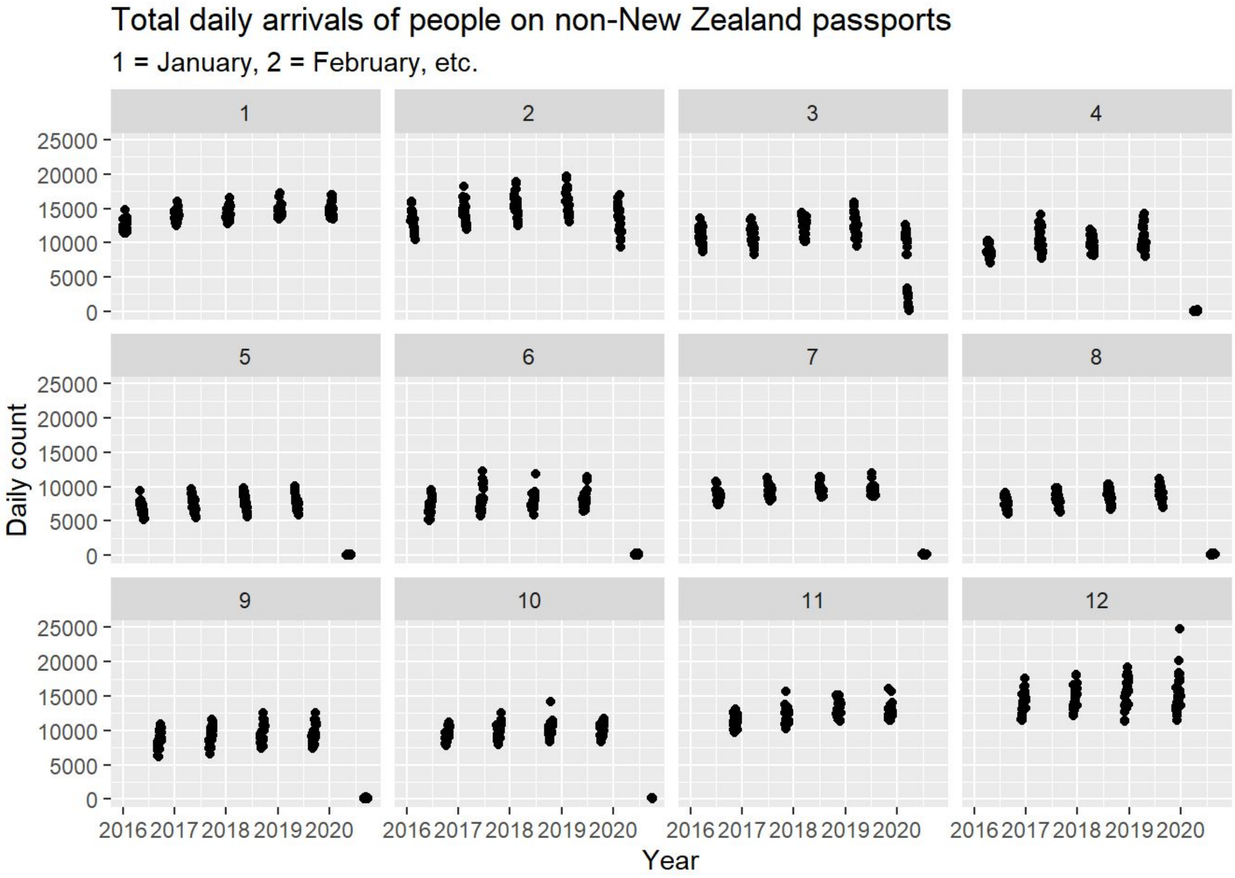


The risk of catching Covid-19 and the closing of the borders meant that people could no longer travel as freely as they could. Border crossings saw a drastic reduction in traffic.

In the graph below, we can see that every year there are seasonal trends where arrivals fluctuate. However, the 2020 arrivals decline significantly from late March onwards, to almost zero.

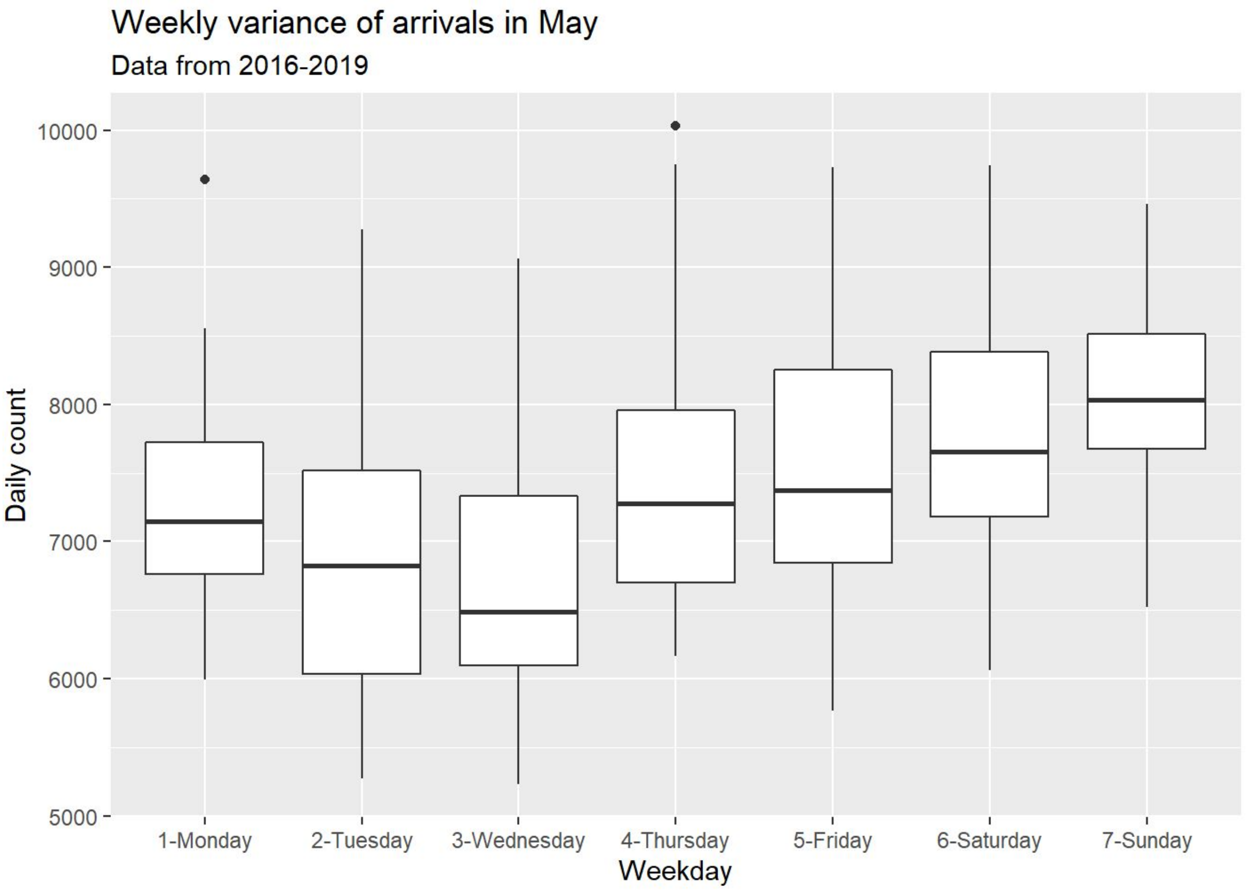


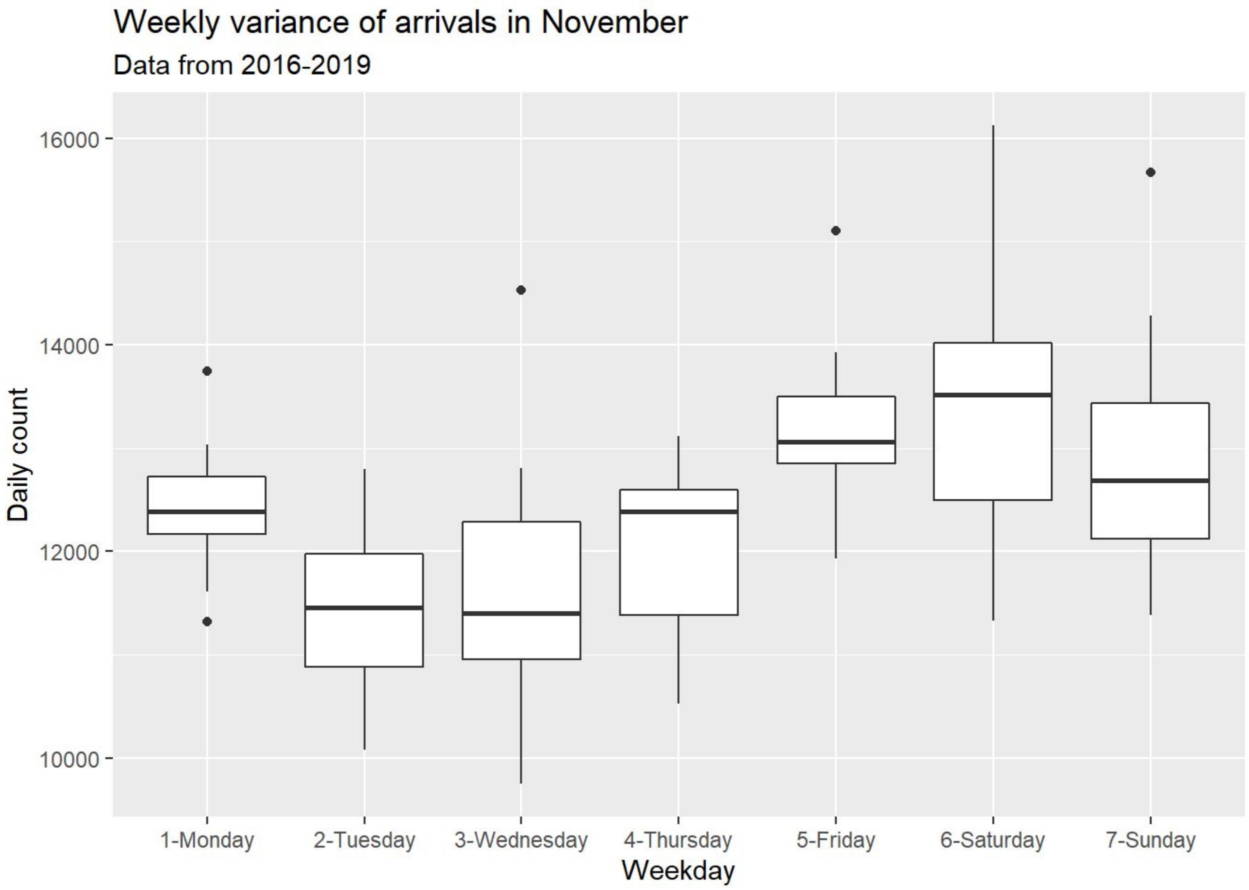
These seasonal trends are useful for answering this report's question, so we needed to figure out what they are. First the data was grouped by month:



Except for in 2020, there is a pattern of there being low arrivals over winter ranging from about 5 000 - 10 000. Then there are high numbers of arrivals over summer ranging from about 10 000 - 20 000. In the bottom right box we can also see the abnormal arrival count of ~25 000 that was mentioned in the Data Description.

Then it was grouped by weekday, however, first the 2020 arrivals had to be excluded as they would skew the results. Then the annual seasonality had to be removed somehow as the variance in those numbers would make any weekly pattern harder to see. The easiest way to do this was to look at it monthly. For simplicity, only two months were looked at; May and November. They were chosen as they also had a small range in arrival numbers of approximately 5 000, compared to December where the count ranges over approximately 10 000. This small range also helped any patterns in the weekdays stand out.


Weekly variance of arrivals in May
Data from 2016-2019
Daily count
10000
9000
8000
7000
6000
5000
1-Monday
2-Tuesday
3-Wednesday
4-Thursday
5-Friday
6-Saturday
7-Sunday
Weekday



These two graphs indicate that there is likely to be a weekly pattern throughout the year, with more arrivals during the weekend, and less during midweek. This means that to get the most accurate forecasts, each year would have to be aligned with the day of the week, e.g. the start of each year being the first Monday of the year, not 1st January being the start.

Detailed Analysis Results

Changing the data

First the data was extracted from the raw data set using the aforementioned classifications:

Class	Category	Indicator	Series
Economic	Travel	Daily border crossings - arrivals	Other passports

The “parameter” variable was converted into date format and the “value” variable was converted into numeric format.

The value was converted into a time series using the ts() function with start=2016 and frequency=365. This frequency does mean leap years were missed. However, this is not likely to impact the results as the data only covers just under five years, and there were only two leap years in that time. One of those leap years was 2020, which was not used in the prediction.

Then the tsclean() function was used. The most notable thing this did was estimate a new value for the 27/12/19 outlier that was mentioned. The point was kept but the value was lowered to be more inline with its neighbouring points.

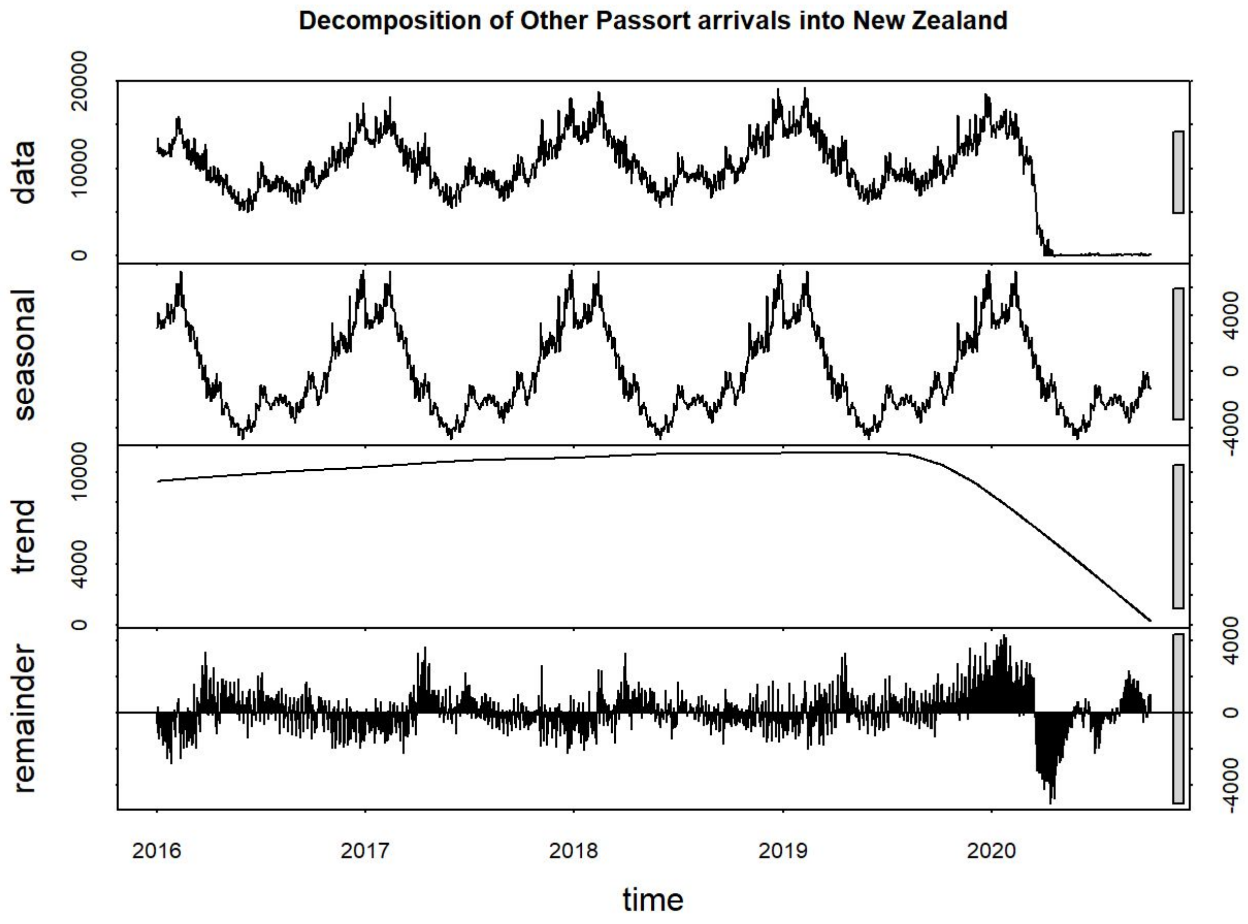
There were no NAs or missing data in this subset of data.

Techniques used

Examining seasonality

During the exploratory analysis, it was found that there were both yearly and weekly seasonal patterns. To confirm this, a decomposition of the time series was done on the cleaned data.

The following visualisation shows the data, the repeated seasonal pattern, the trend, and any irregularities that didn’t fit the seasonal and trend. The numbers on the left are the arrival count, and the number on the right is the range of variance from the average.



The annual seasonal pattern becomes even more apparent in that visualisation and the trend indicates that the number of arrivals is increasing each year. The trend line here also indicates that that increase stopped around mid to late 2019, but that is likely to be a result of smoothing data where the average arrivals was around 10 000 to suddenly being 0. The suddenness of this change means the change in the trend appears to occur earlier than it actually happened. This would also explain why there is a big remainder at the end of 2019, before Covid-19 was considered a concern by most of the world.

However, to be sure, initially only the data until 30/11/19 was used with the forecast function. This is a long length of time to be estimating, so accuracy was not good and the confidence intervals widened significantly towards the end.

Forecasting

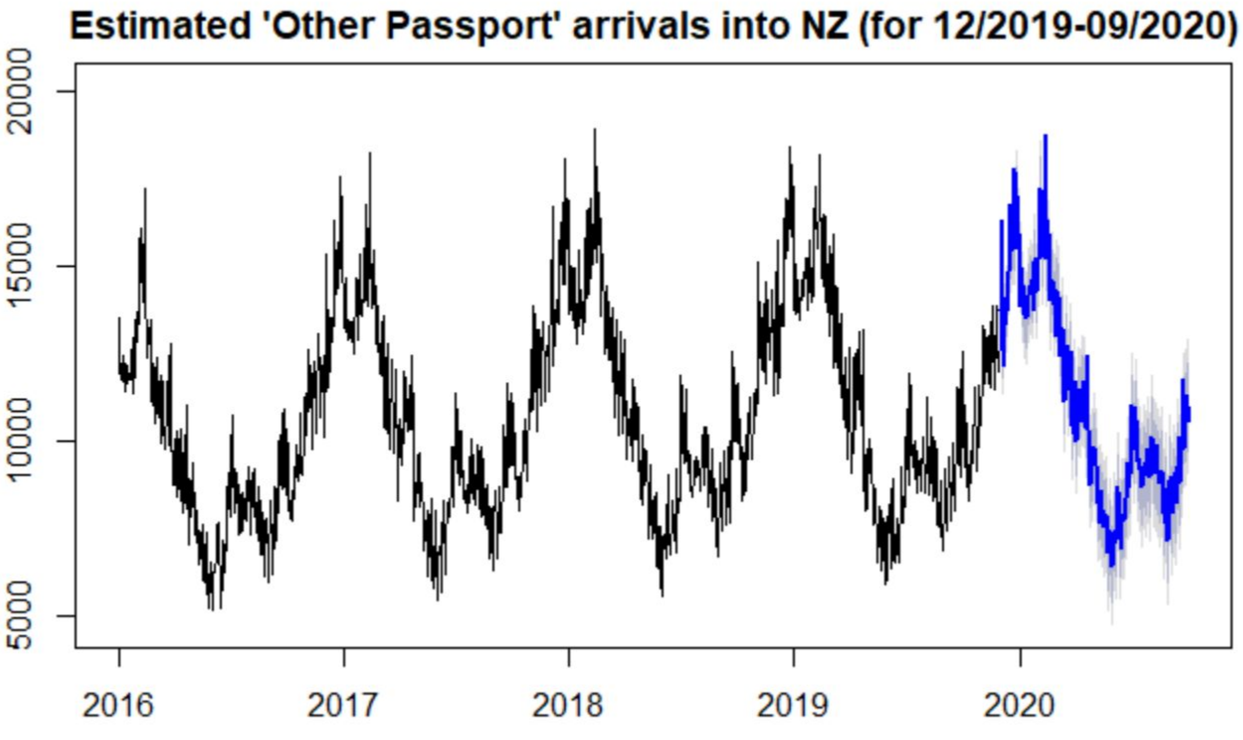
To estimate the number of arrivals of people into New Zealand that use non-New Zealand passports, the forecast() function was used. This is because some arima functions (a more technical forecasting option) have upper limits on the amount of data points that can be used. This issue could have been circumnavigated by using the weekly moving average, but only an estimated count and confidence intervals is required to answer this question, which forecast() provides. Therefore this process was deemed unnecessary.

The prediction initially uses the data from 01/01/16 - 30/11/19, and then estimates the next 305 days of arrivals. The process was then done again but with data from 01/01/16 - 29/02/20, and estimating the next 214 days of arrivals.

Predictions

From December

The forecast function produced this graph:

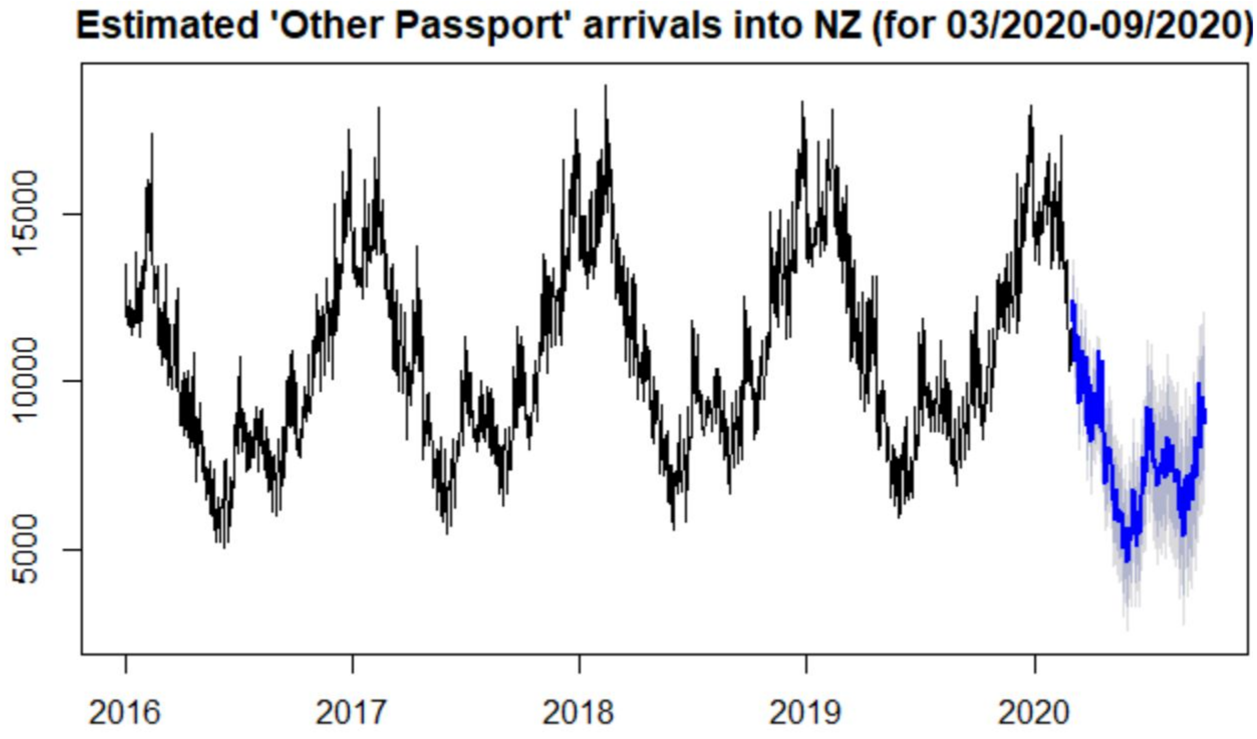


The blue line represents the estimated arrivals, and the grey parts next to it represent the 80% and 95% confidence intervals. The latter are not clearly visible, so below is a table with a summary of the total estimates. Of course this period included actual arrivals, so to work out the estimated loss of international arrivals, that needs to be subtracted.

Sum of estimated arrivals (Dec-Sep)				
Lower 95% CI	Lower 80% CI	Forecast	Upper 80% CI	Upper 95% CI
2 898 199	3 068 143	3 389 176	3 710 209	3 810 153
Less the 1 517 982 actual arrivals:				
1 380 217	1 550 161	1 871 194	2 192 227	2 362 171

95% of the time, it is expected that the number of arrivals into New Zealand between December 2019 and September 2020 would have been between 2 898 199 and 3 810 153. This means the country lost between 1 380 217 and 2 362 171 arrivals. This range of one million is very large, so the same process was repeated to reduce the time frame being estimated.

From March



The first thing to note is that the previous estimates included higher arrival counts from 01/12/19-29/02/20 than the cleaned data says there was (note the change in the y-axis). Clearly the forecast function was taking into account the upwards trend that was occurring.

Sum of estimated arrivals (Mar-Sep)				
Lower 95% CI	Lower 80% CI	Forecast	Upper 80% CI	Upper 95% CI
1 190 029	1 352 211	1 658 580	1 964 949	2 127 131
Less the 184 306 actual arrivals:				
1 005 723	1 167 905	1 474 274	1 780 643	1 942 825

The range in the 95% confidence interval is still just under one million, and the March data can not be used for a forecast as that was the month of the first reported Covid-19 case in New Zealand.

However, this second estimate has lowered the losses by approximately 400 000 arrivals for each confidence interval.

Analysis

This difference of 400 000 impacts the answer to this report's question significantly, so one must be picked.

The Dec-Sep estimate covers a large range of time, which is an unwise thing to do because it reduces the accuracy, especially as the estimate is based on only 4 years of data. However the range in the confidence intervals is only slightly larger. Furthermore, traveler confidence was decreasing by early 2020 which would have skewed the Mar-Sep estimates. This means that the Dec-Sep estimate is likely to be more accurate. However further analysis may be required to confirm this.

Going with the assumption that the Dec-Sep forecast is more accurate, that means that if the pattern of the previous four years had continued, then New Zealand has missed out on approximately 1 871 194 international arrivals in 2020. The 80% confidence interval is 1 550 161 - 2 192 227 arrivals lost, and the 95% confidence interval is 1 380 217 - 2 362 171 arrivals lost.

That is a lot, but this loss is over winter. We can see from previous years that summer is when most international arrivals happen, so the losses will be greater then.

Bias in the results

There is limited bias in the data itself, as the number of arrivals is not subject to opinion. There is bias in the results, as was briefly discussed in the analysis. The accuracy of the results are subject to my skill in using R to forecast and my assumptions of what dates should be picked (i.e. using the December forecast over the March forecast due to the assumption of decreased traveller confidence in early 2020.)

Conclusions and Recommendations

Conclusions

To answer the original question, New Zealand has lost an estimated 1.5-2 million international arrivals in the recent months due to the impact of Covid-19.

Recommendations

The only recommendation is to do further study into this. This is discussed further in Future Work.

Limitations

As the data is aggregated, we cannot yet tell what this loss of arrivals means. The lost arrivals could be tourists, skilled migrants or people on working holiday visas.

Future work

1) Repeat this study but more thoroughly. The forecast did not end up taking into account the day of the week and as there is weekly seasonality, that will skew the results.

   Also, after checking the decomposition I attempted to go through the process of an Augmented Dicky-Fuller Test, check the auto correlation and use some of the Arima functions (available in the github link below). Checking other forecasting methods would build a more accurate picture of the numbers of lost arrivals.

2) Combining the arrival data with information on entry reasons. Doing this would give a better idea on whether New Zealand has lost more tourists or more workers, and that would shape the future questions that could be answered by this.

References

Code: https://github.com/IonaCS/data301

Edmunds, S. (2020). *Māori women took hit in first wave of Covid-19 job losses, Stats NZ data shows*. Retrieved from https://www.stuff.co.nz/business/300088321/mori-women-took-hit-in-first-wave-of-covid19-job-losses-stats-nz-data-shows

Martin, H. (2020). *Coronavirus: Māori more likely to die from Covid-19, study finds.* Retrieved from https://www.stuff.co.nz/national/health/coronavirus/300096832/coronavirus-mori-more-likely-to-die-from-covid19-study-finds

Stats NZ. (2020). *COVID-19 data portal*. Retreived from https://www.stats.govt.nz/experimental/covid-19-data-portal

Stats NZ. (n.d.). *Privacy, security, and confidentiality of information supplied to Statistics NZ*. Retreived from http://archive.stats.govt.nz/about_us/legisln-policies-protocols/confidentiality-of-info-supplied-to-snz.aspx#gsc.tab=0