

STAT9005: Time Series Project

United States Covid Data

Brian Higgins – R00239570

Introduction

For this project, we will be using a Covid dataset from “ourworldindata.org” which shows detailed information for the Covid Pandemic with the number of cases, tests, vaccinations, deaths, etc recorded for countries worldwide. We chose the United States (Figure 1) as the United States has good recorded data and the plot shows some interesting fluctuations which should make for some interesting Time Series Analysis.

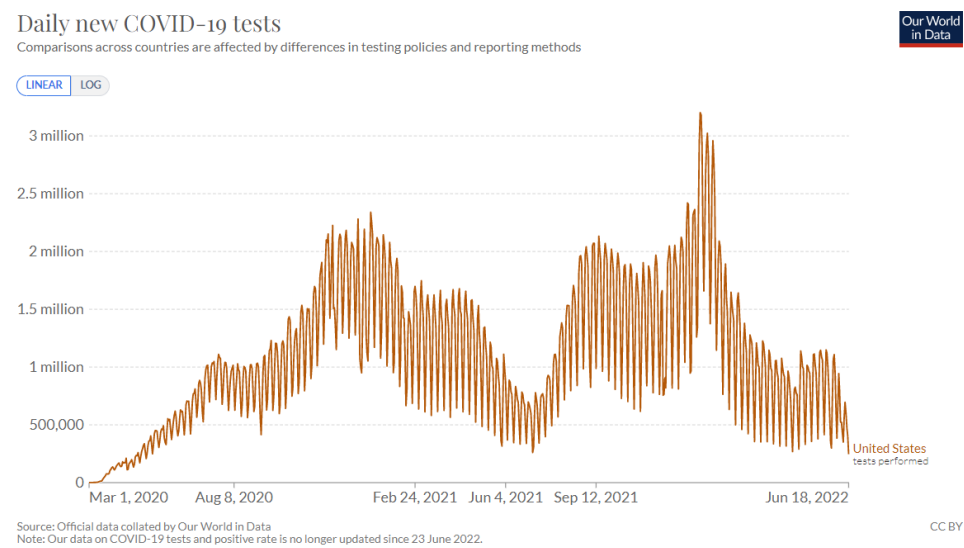


Figure 1: United States

EDA

Time Series Variable Selection

There are 67 descriptive columns covering vaccinations, cases, deaths, etc and 304,355 observations covering a period from 2020-01-03 to 2023-04-14 for 255 worldwide countries and areas. In Figure 1, we saw the “new_cases” data which is very promising. While the new tests variable looks promising to use as data for our Time Series Project on I wanted to investigate some other variables to offer a comparison for this decision.

Table 1 shows 4 variables; New Cases, New Deaths, New Tests, and New Vaccinations. With this initial table we can see that New Cases and New Deaths have nearly all the data whereas New Tests and New Vaccinations have very high rates of missing values.

	New Cases	New Deaths	New Tests	New Vaccinations
Missing Data	0	1	363	347
Percentage of Missing data	0%	0.08%	30.17%	28.84

Table 1: Variable Selection

On this basis alone I might have chosen New Cases or New Deaths but as we can see in Figure 2 I do not find New Cases or New Deaths to have as interesting Time Series fluctuations. Both have large periods at the end that would not be helpful. So while they both have more observations than the other two variables, when you remove the flat periods at the beginning and end, the difference in the number of observations is not so large as the other two variables. New Tests does still have a little more missing data but later in the project when we split the data into three waves I believe it gives more interesting waves to complete our Time Series Analysis on and so this is why it was chosen. New Vaccinations is also not as interesting because it is missing a large amount of data and it has three similar periods which would not be interesting to investigate once we broke the data up into three daily waves.

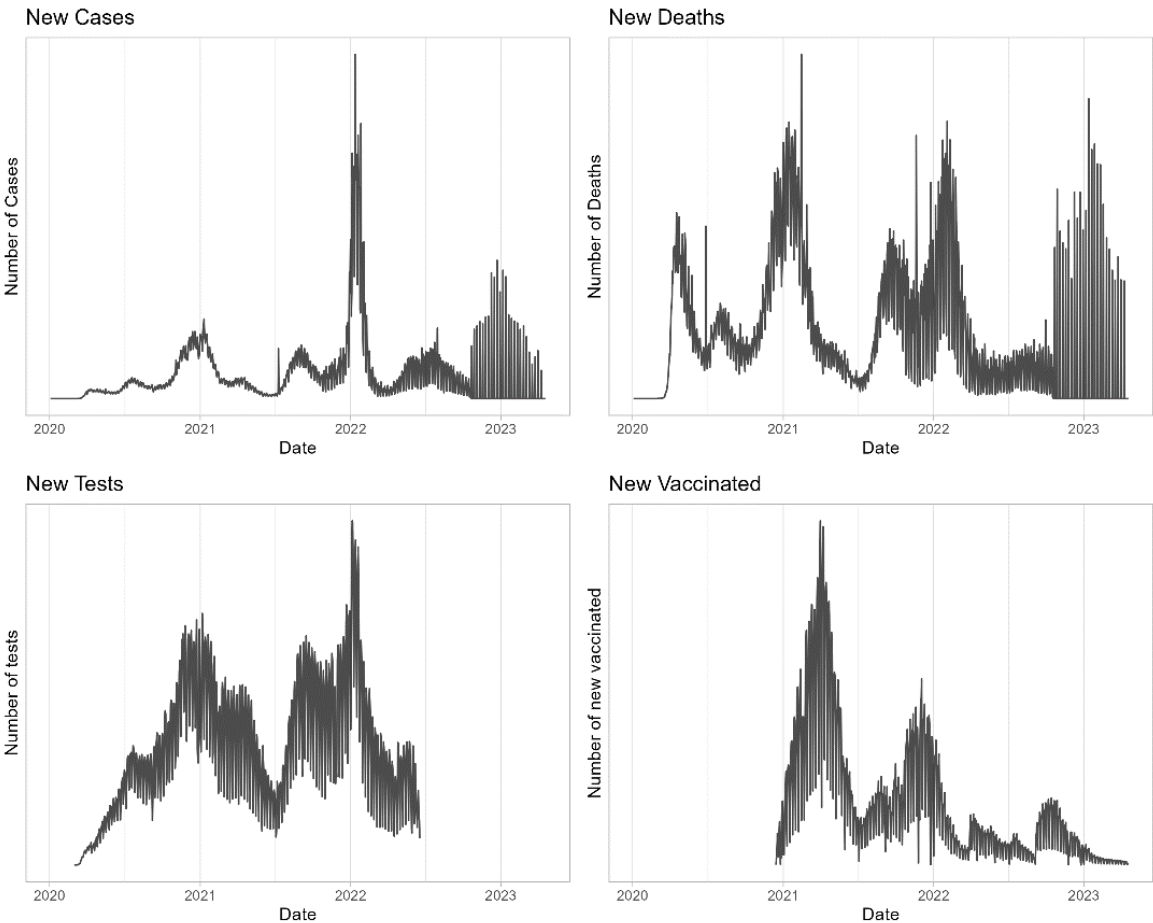


Figure 2: Compare Variables

Chosen Variable: New Test Missing Data

As noted we have chosen New Tests as our data for our Time Series Analysis but we still have to deal with the Missing Values. We choose the United States because they would likely have good records and while there are 363 missing data points, once we remove the points from the beginning (58 values) and the missing values at the end of the data (305 data points) there were no missing values inside the remaining data so we are now left with our final Time Series Data which we can see in Fig 4. As you can see it is similar to the plot we saw on the ourworldindata.org website in Figure 1.

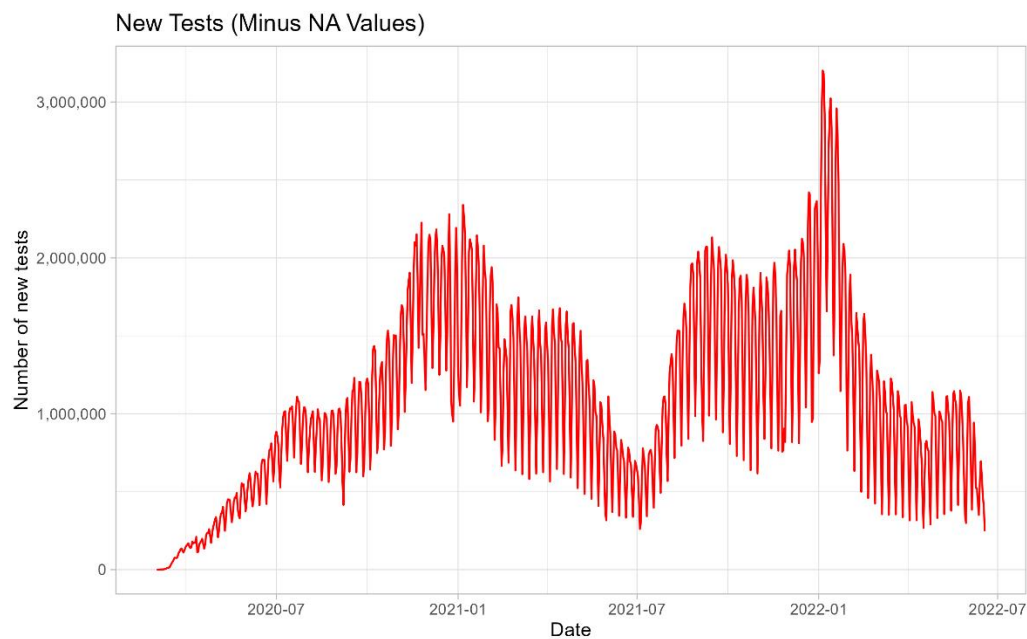


Figure 3: New Tests Time Series Plot

Preliminary Analysis

Summary Analysis and Plots to describe the three time periods

From the summary statistics in Table 2, we can see the number of tests per day, week and month varies significantly. The daily values have the smallest range and the monthly values the highest unsurprisingly. The mean also increases in the same way. The IQR also increases suggesting that the data has a wider distribution as the time scale increases.

Summary statistics for each period:

	Daily	Weekly	Monthly
Minimum	515.0	8157	1533392
1st Quartile	663230.5	5379211	22131045
Median	1011561.0	7355708	30871745
Mean	1496854.5	10282004	42873039
3rd Quartile	1087924.6	7606406	32598885
Maximum	3201706.0	18365533	66249955

Table 2: Summary Statistics

Figure 4 shows the three periods, daily in the blue line has many data points, the middle red line plot shows the weekly data and the right plot with the green line shows the monthly tests. Each line plot has less detail and also note the increase axis values as we move from daily to weekly and then monthly.

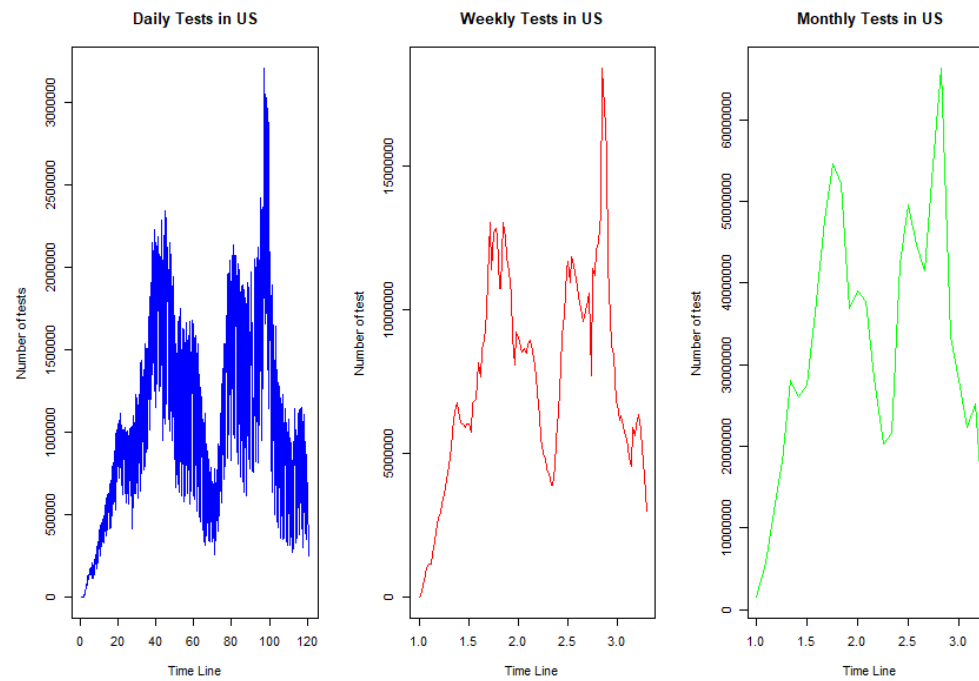


Figure 4 line plots

Figure 5 shows the histograms for the three-time ranges. The Daily data shows a right skew with a peak of 1 million per day. There are only a few low and even less very high test days. Most fall within a range of 500k to 1.5 million. The weekly histogram is also right skewed with a peak of around 7.5 million. The monthly histogram shows a different pattern with a more symmetrical distribution, (with a hole in the middle that could be showed more or hidden by adjusting the bins). Some of the differences in the monthly distribution could be because of recording issues or because of aggregating the data over time which could cause this.

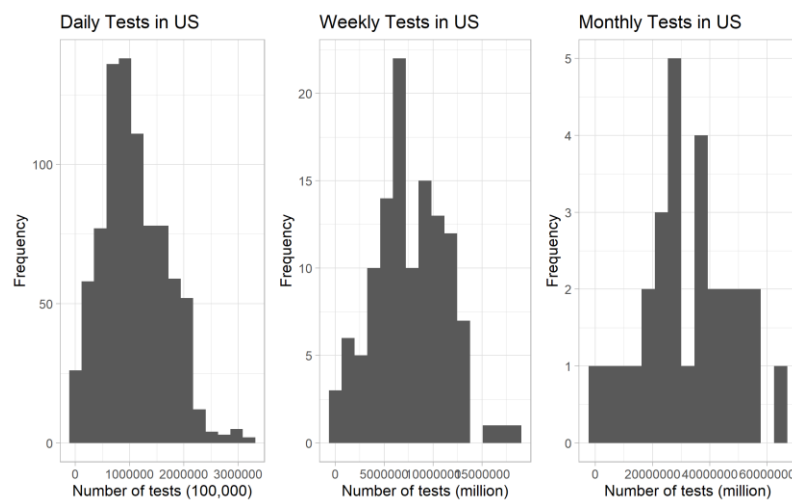


Figure 5: Summary Histograms

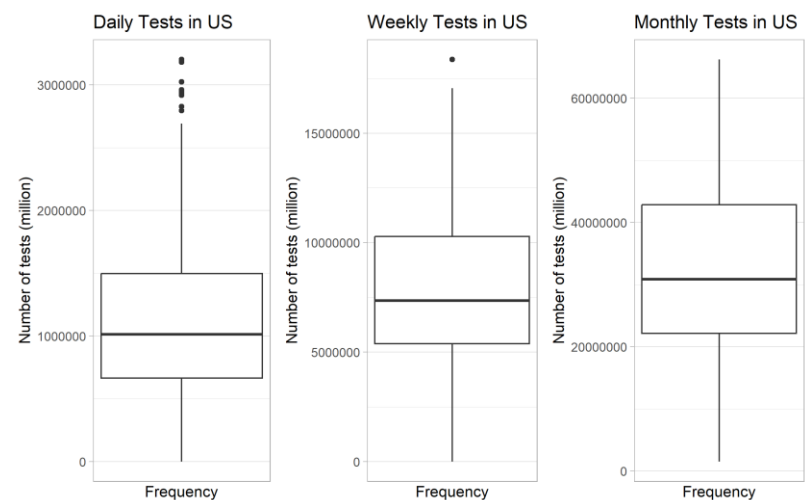


Figure 6: Summary Boxplots

Figure 6 shows boxplots for the same periods, generally, the information is the same, but you can see that the Daily Boxplot has some outliers, the weekly just one and the monthly none. The average number of tests goes up as we increase the time line but over all the range stays similar. The skew is not as clearly seen in the boxplots.

Decompose the three time periods

Daily Time Period

Below in Fig 7, we can see the Additive and Multiplicative models. As we can see the Trend looks the same in both. The seasonality is very similar, if you zoom in you can see very slight differences. The scale is larger for the Additive model but this does not necessarily mean much when comparing the two models as the scale is a natural outcome on how the models are prepared.

The randomness of the two models is the only part that gives us some indication. It looks more random for the Multiplicative model. There is not a huge difference. But in this case, I would use the Multiplicative model as the random component looks more random even if it is a more complicated model.

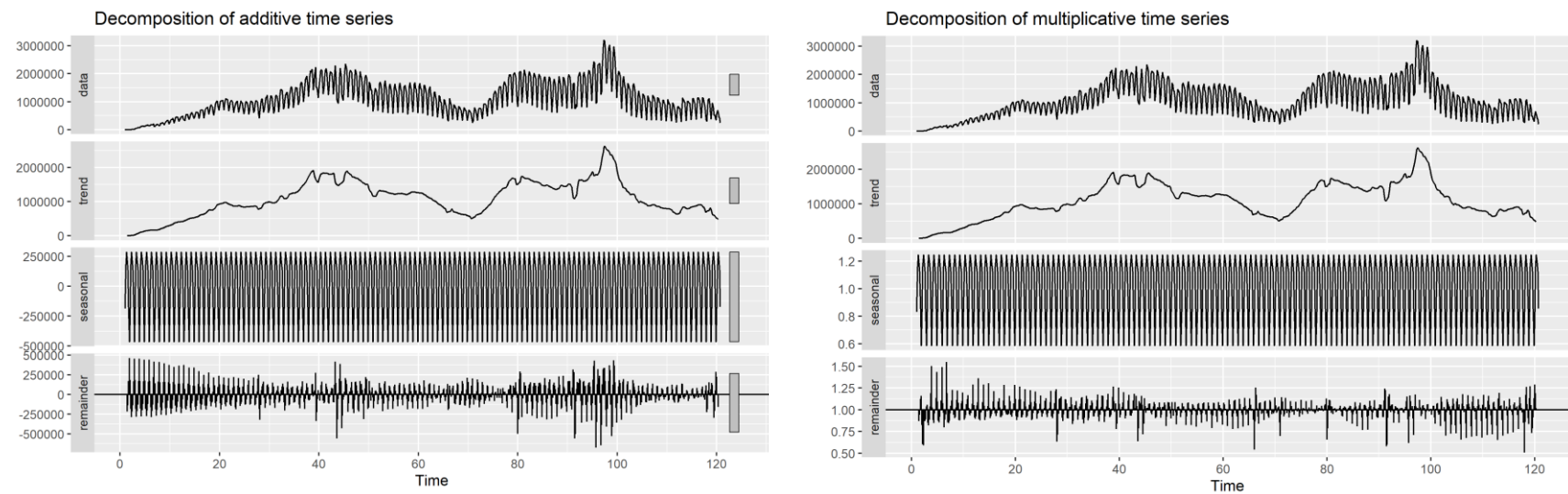


Figure 7

Address Issues:

In the below plots in Figure 8, we will use a log transformation on the models to see if it has much of an impact. Once again the Trend is the same, seasonality looks similar and the random component also looks similar. In this case I would not use logs as it does not improve our results and its better to have as simple a model as possible.

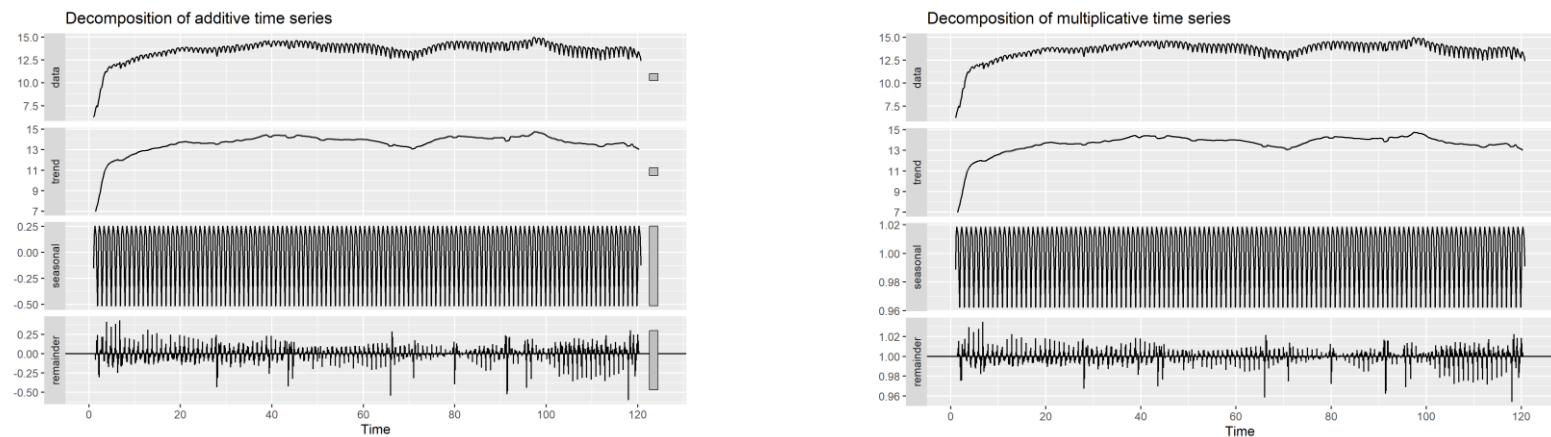


Figure 8

Periodicity: We used MSTL as there are clearly two peaks and looking at periodic you can see in Figure 9 on the left an s.window is set to 7 and on the right a value of “periodic” to capture any longer-term changes in the trends in the seasonal pattern. The left plot with the s.window of 7 is not capturing the data well, while the trend is the same, the seasonality is not regular.

The right plot with a “periodic” value for s.window looks much better as the seasonality component looks better. This is capturing a seasonal component over all the data and there looks like there are two peaks centred on the beginning of 2021 and 20222 so this makes sense. You can note the random component looks the same as the Additive model above.

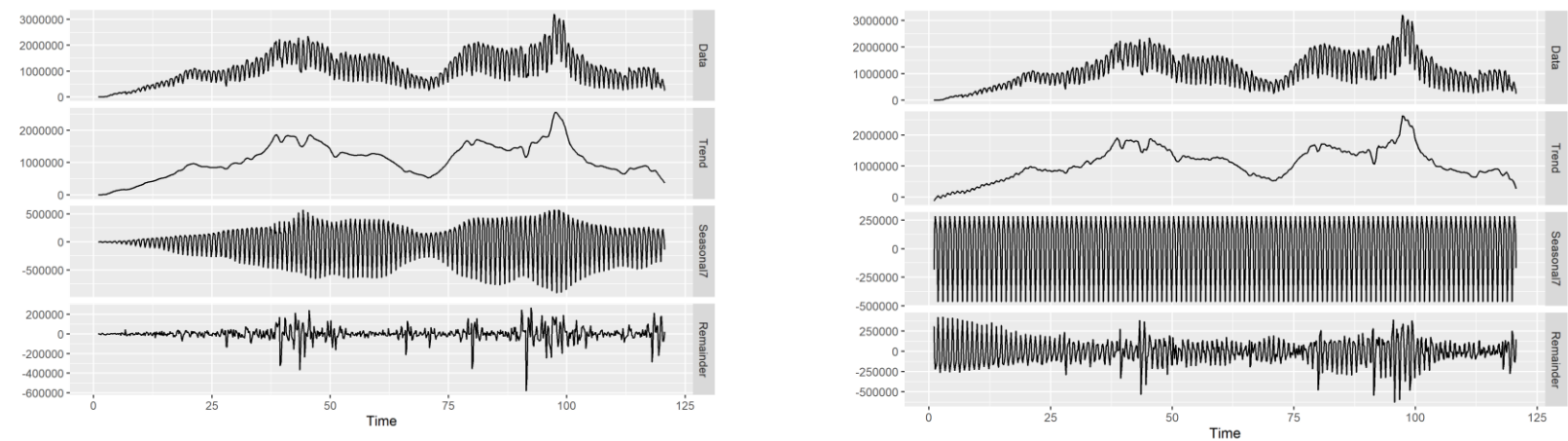
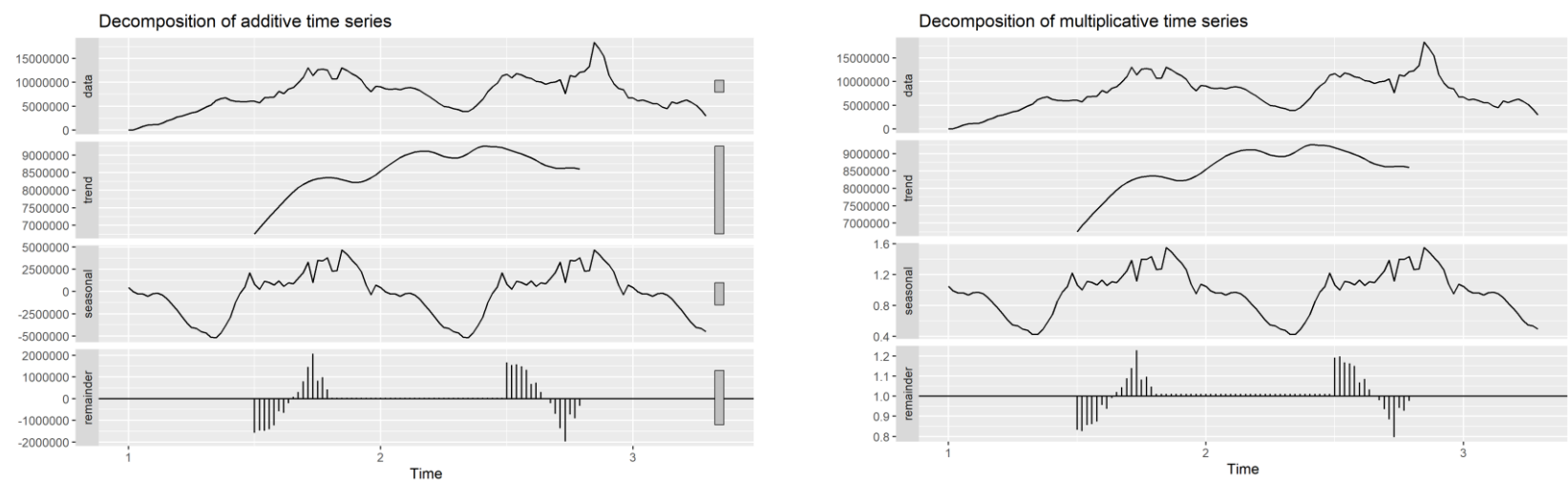


Figure 9

Adding complexity to the model with transformations and periodicity, the Multiplicative model without these seen in Figure 7 is the best model for capturing the Daily Time Series. While it is more complex than the additive model It looks better.

Weekly Time Period

Figure 10 shows the Additive and Multiplicative models for the weekly Time Series data. We can see that they are very similar, the trend is the same and seasonality looks the same also. The random component has the same shape in both and it does not look random, it looks the same with a peak at the beginning and end and a straight line between. So not random at all. The Additive model would be the better of the two here but let’s look to see if we can address the issues in the random component.



Address Issues:

Using log transformation has very little effect on the models so it is of little use.

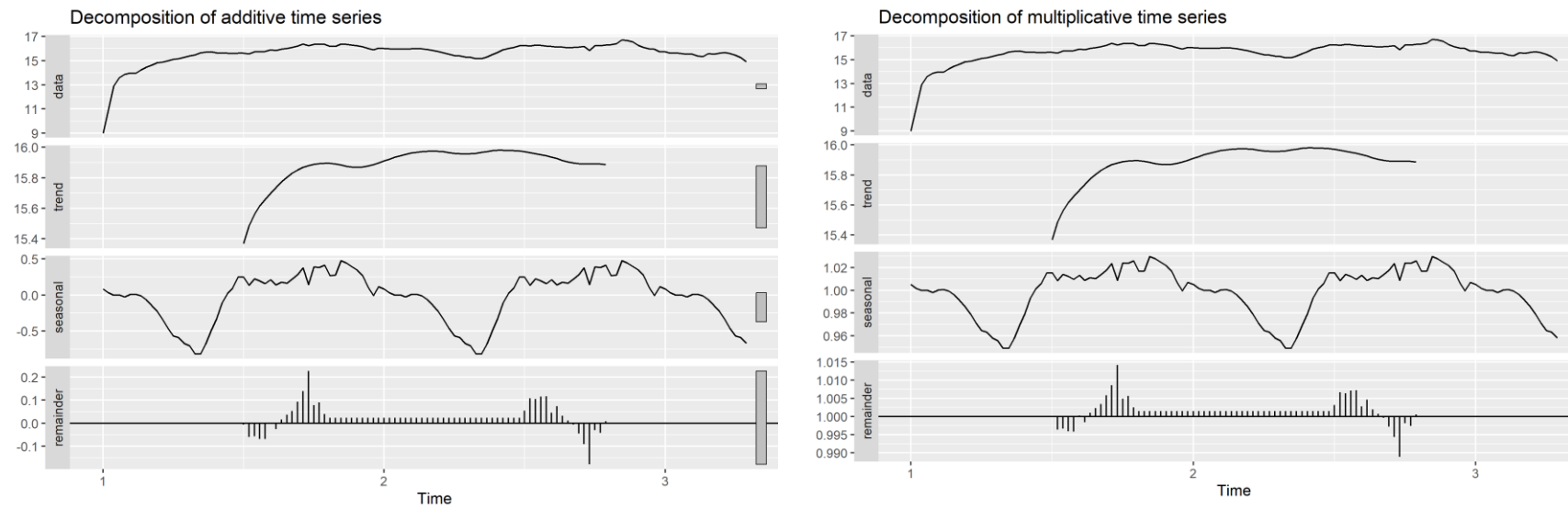


Figure 11

Periodicity:

Figure 12 shows two plots when we look at the periodicity, the left has a s.window value of 52 and the right uses a “periodicity” value. As we can see they give the same result. As they are the same, the periodicity was chosen with a period of 52 which indicates a strong yearly seasonal pattern. This does make sense as we can see two peaks in January of each year.

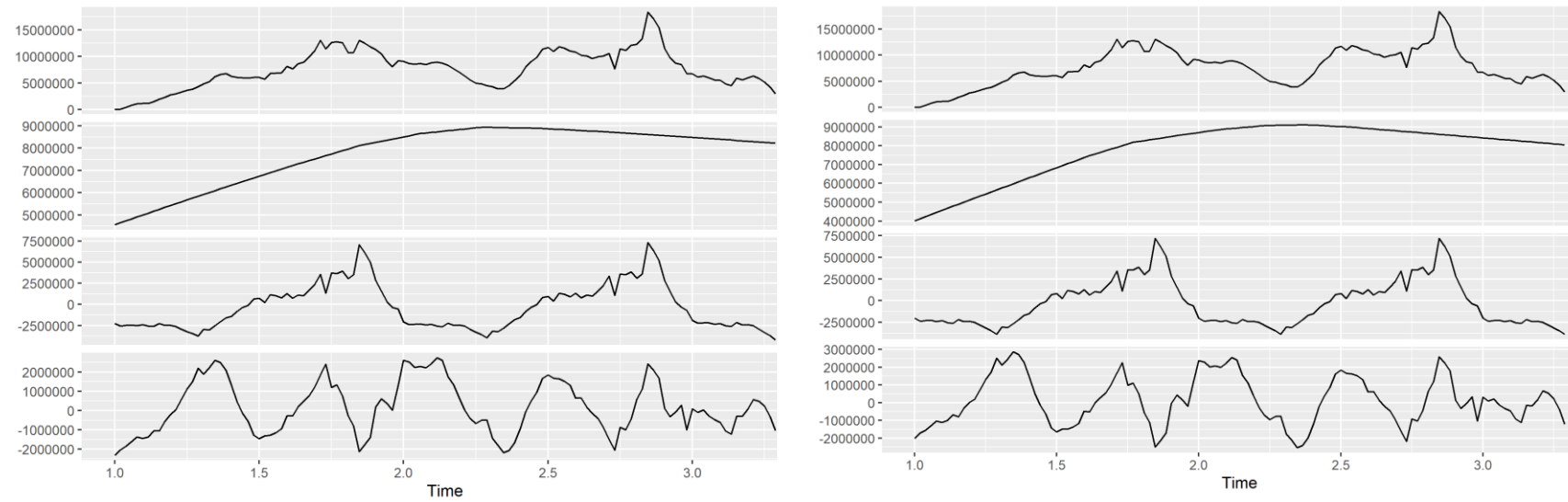


Figure 12

However, we still have a random component that is not looking very random so it still looks like the models are not capturing the underlying patterns in the data. Figure 13 shows a MSTL model with a periodicity of 52 with log transformations on the left. But the random component still does not look random, you can see the two peaks match the seasonality. The right model uses STL with a periodicity of 52 and log transformations but again we can see the same non-randomness. Neither of these options are any better.

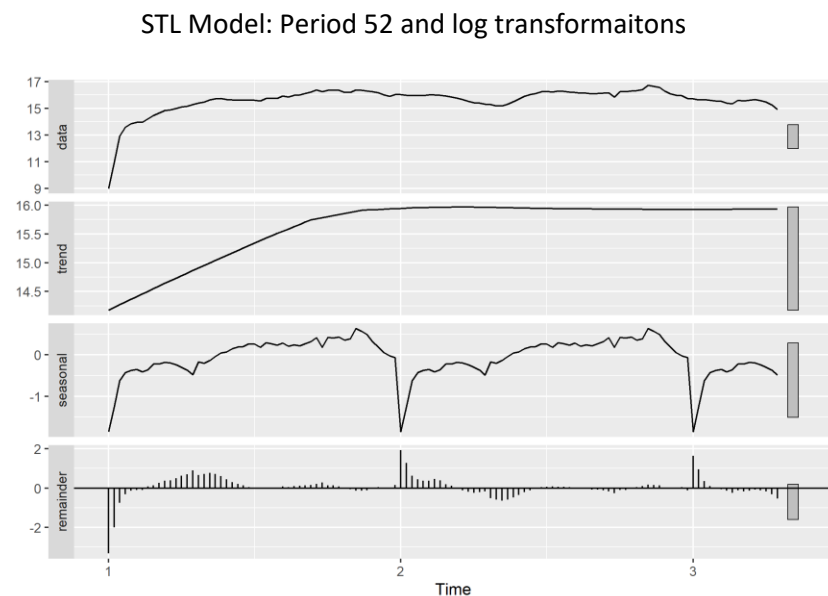
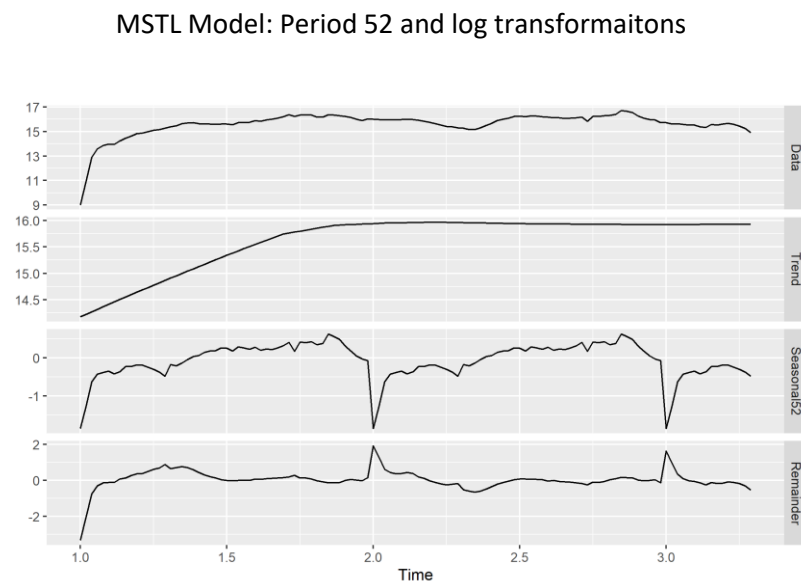


Figure 13

As the random component is still appearing to not be fully random to address the issues I would continue to look at other time series models such as Classical and ARIMA. Since we will look at these techniques later we will come back to exploring these techniques in the next sections.

The data used is complex so it may not be possible to capture all the components in the data. The variance in the weekly data and its seasonality seem to be causing the most issues that we can see so far.

Monthly Time Period.

Figure 14 shows a very similar situation to the weekly data models. Trend and Seasonality look good but the random component is still an issue. Again we have two peaks for the increase of new cases each January.

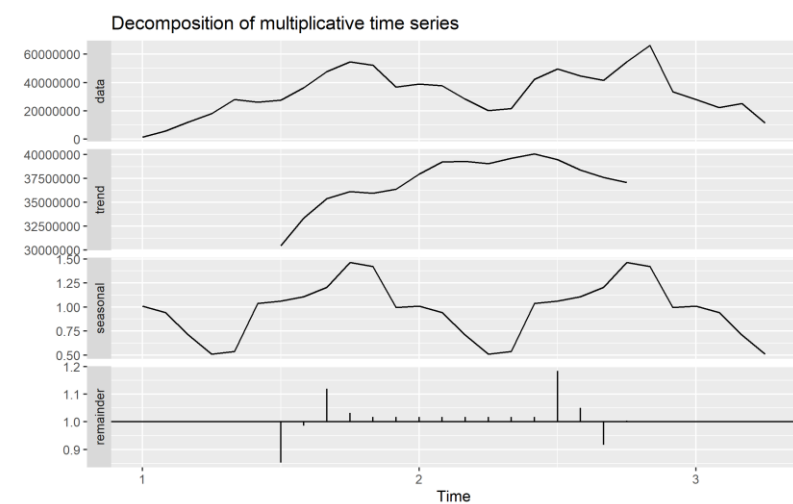
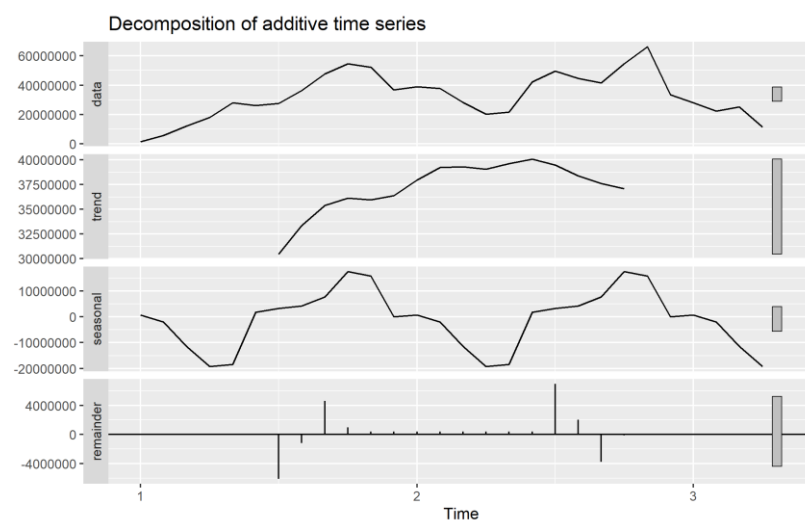


Figure 14

Periodicity:

Figure 15 shows a periodicity of 12, for 12 months and the periodicity value in the s.window picks up the same value again. We tried MSTL and STL with logs again but no improvements. Again the random component is still not looking random.

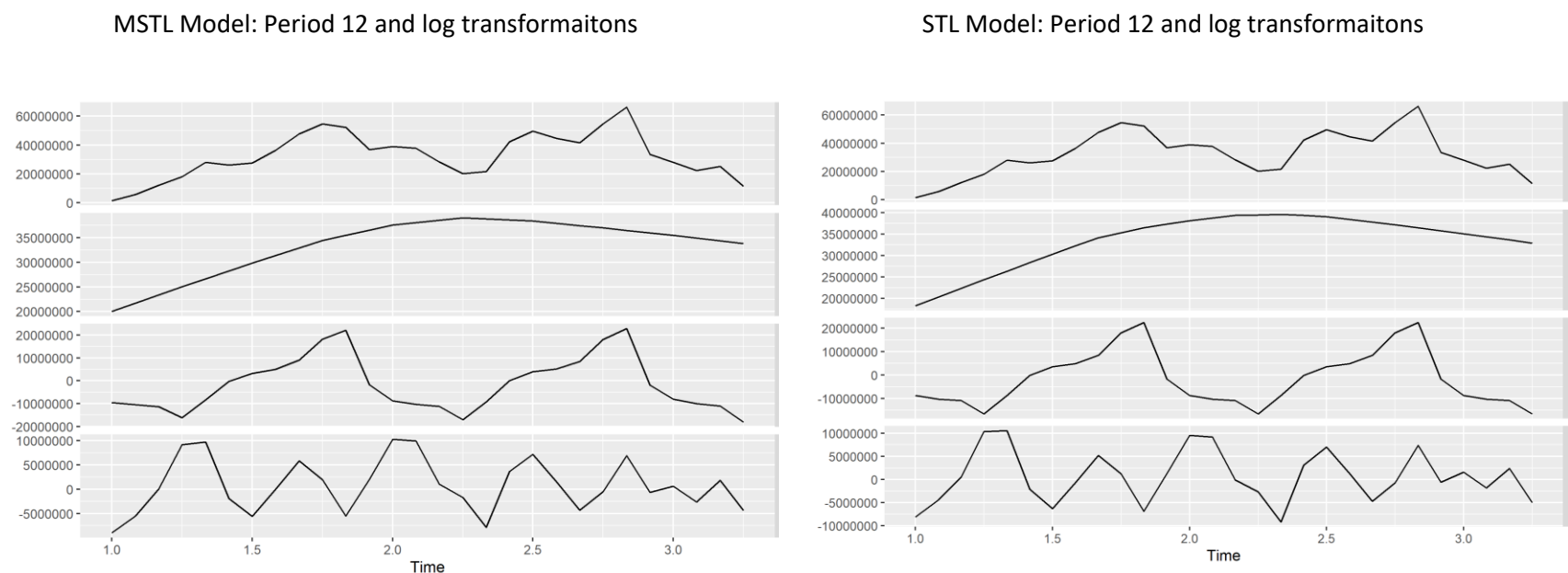


Figure 15

The Monthly Time Series is looking very similar to the Weekly Time series so once again we would look at Classical Models and ARIMA models to try and explain the Random Component as there is some underlying pattern that is not being picked up. We will investigate this further using these techniques in the next sections.

Time Series Modelling

Splitting up the daily dataset into three sections

We will now take the time series data and create three new daily periods for the New Tests data. Figure 16 shows how the original dataset has been divided into three new periods or “waves” and what those waves look like. What is interesting is we were having some issues explaining the random component in the plots above and now with the three sections it’s easier to see more variance and seasonality inside each weekly season. This is likely as the reporting of new tests was not given at the weekends and so each weekend there is a dip with a big release of test results at the beginning of the week.

The three waves are of different time lengths as I choose three periods that would be interesting to analysis with time series techniques rather than three equal time lengths. This gives unique and interesting differences in the three waves chosen.

- Wave 1: 2020-03-02 - 2020-11-29 – 9 Months
- Wave 2: 2020-11-30 - 2021-12-26 – 13 Months
- Wave 3: 2022-01-02 - 2022-06-18 – 7 Months

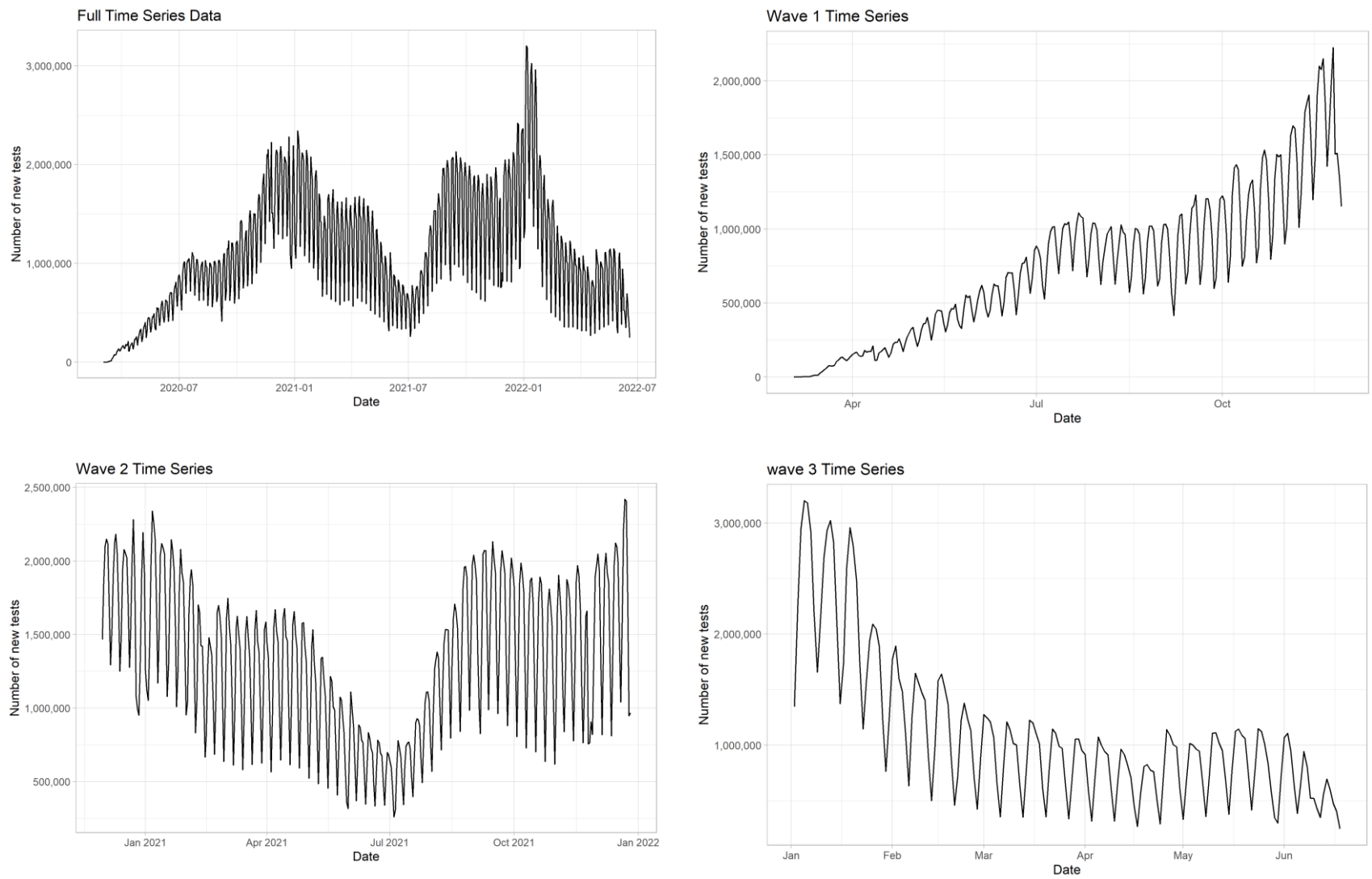
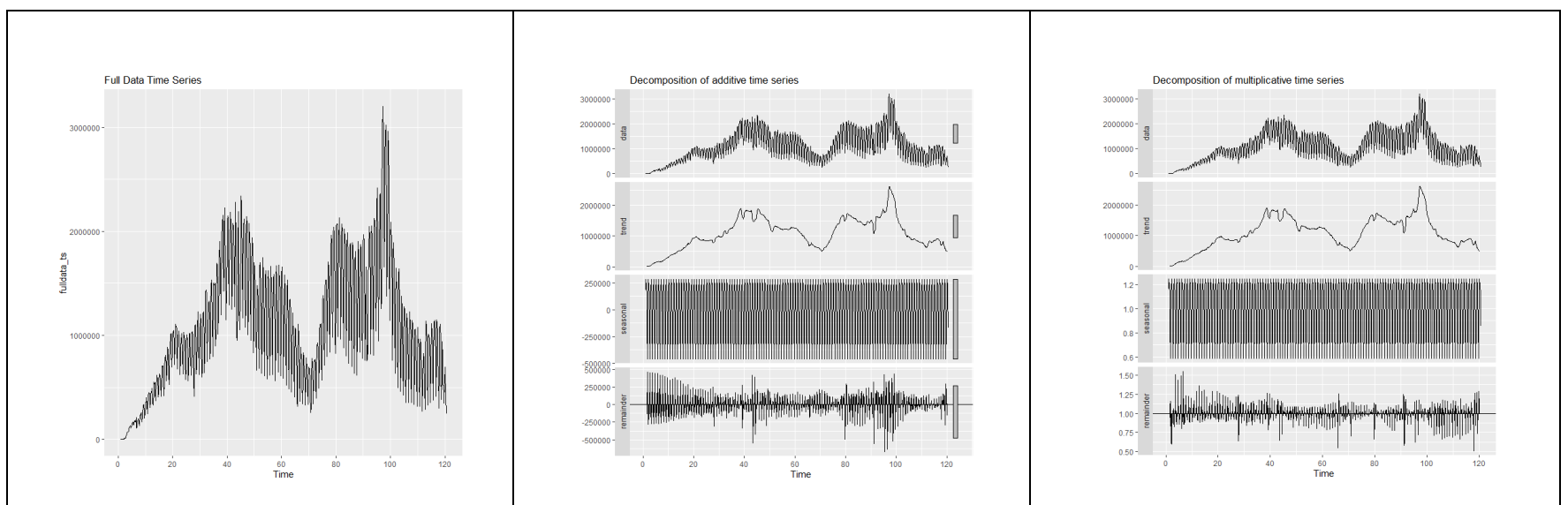
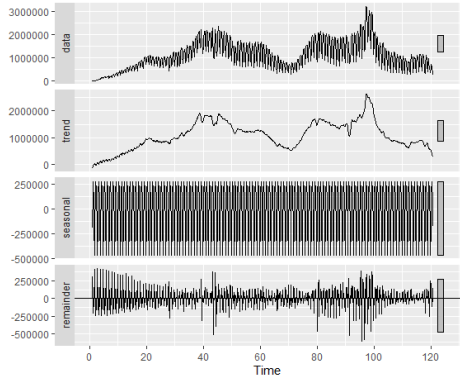
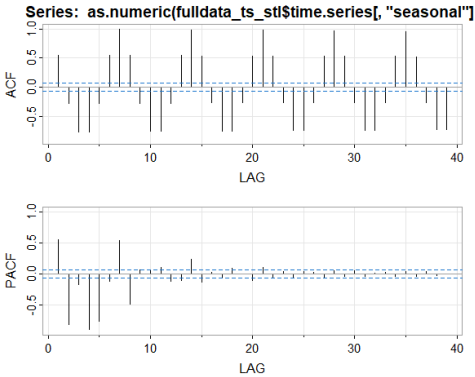
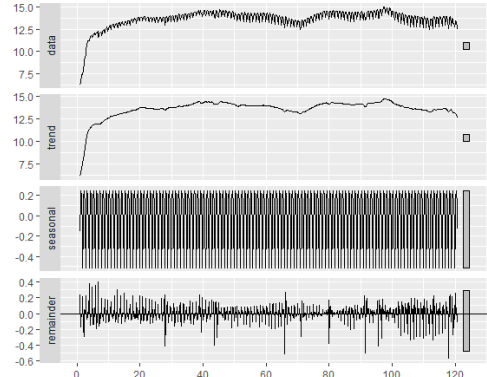
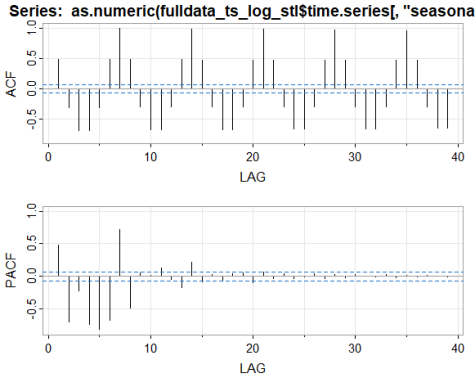
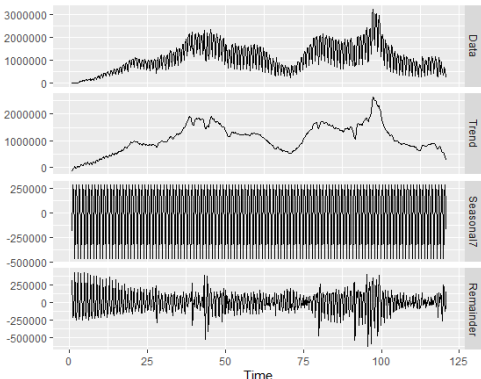


Figure 16

Daily Full Data: Classical Models

I have included the most interesting plots below, a lot more can be found in the R code as needed.

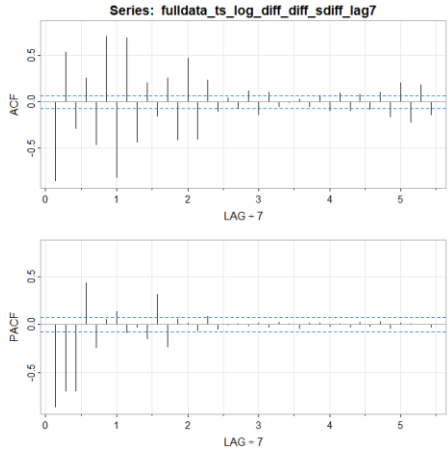


<p>1. Full data Time Series</p> <p>We have our full data time series where we can see a seasonal trend over a two-year period. Also, there is a lot of variation and seasonality within each week which is expected with the release of new_tests results. There are two peaks.</p>	<p>2. Decomposition with Additive</p> <p>The trend has the most impact and also looks like seasonal is capturing the data but the random component could perhaps be better.</p> <p>There appears to be a pattern at the beginning and end of the random component.</p>	<p>3. Decomposition with Multiplicative</p> <p>For a Multiplicative model, we see similar results for Trend and Seasonality but the random component looks better. Although a more complex model it captures the data in a better way.</p>
		
<p>4. STL Model</p> <p>The trend has a small rectangle similar to the original data and so is more important. There are two peaks in the overall Time series. Seasonal looks ok. The random component could be better, lets compare to other models.</p>	<p>STL Model</p> <p>The ACF plot for STL still has a very strong seasonally component. So we will look at other options.</p> <p>The PACF plot is looking ok.</p>	<p>5. STL with Log</p> <p>Used log of the Time Series data to see if it makes an improvement in removing the seasonality but it does not.</p> <p>The random component looks the same as above.</p>
		
<p>STL with Log</p> <p>The ACF does not show an improvement so the extra complexity of adding log does not improve the model with STL.</p>	<p>6. MSTL</p> <p>There are two peaks so MSTL is a better model to try. We can see that it picks up the Seasonality as 7 which we have in the weekly data.</p> <p>The Random component looks same as first STL model.</p>	<p>7. Holts Winter</p> <p>Tried Holts Winter as I thought it would be a good model to try as the data had Trend and Seasonality but received an error as the data was not a good fit when displaying decompose models.</p> <p>We will use it later for some forecasting as a comparison.</p>

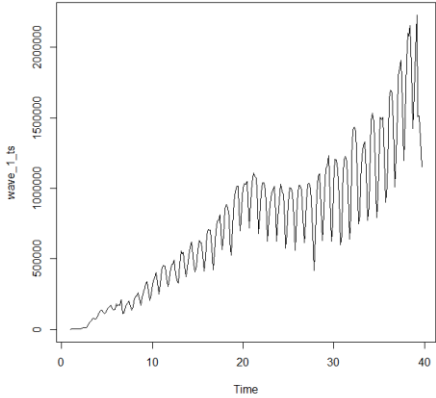
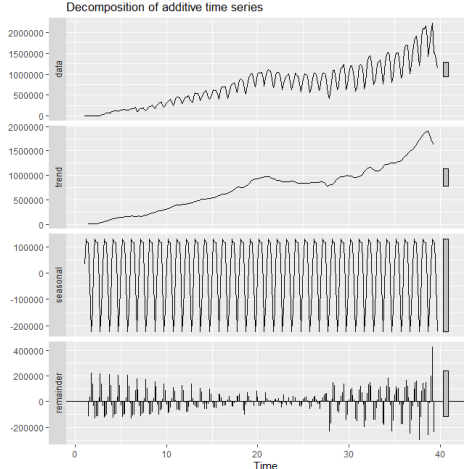
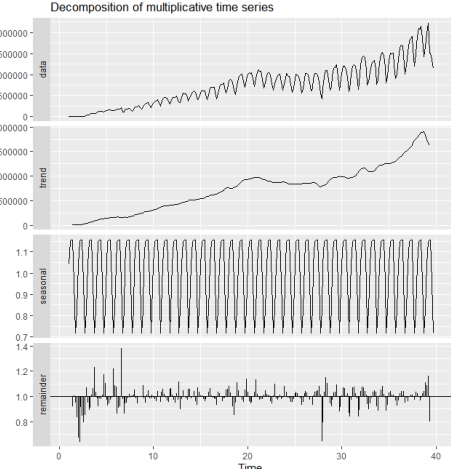
Generally it looks like the classical models are not doing well at capturing the patterns in the data. This does happen and sometimes it is not possible to get a good fit with certain models. We will continue on to the ARIMA models to look if they can do a better job at capturing the data.

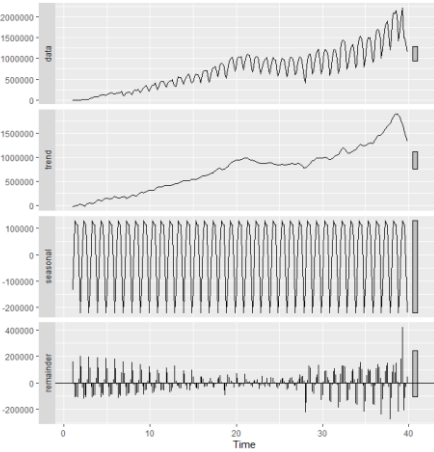
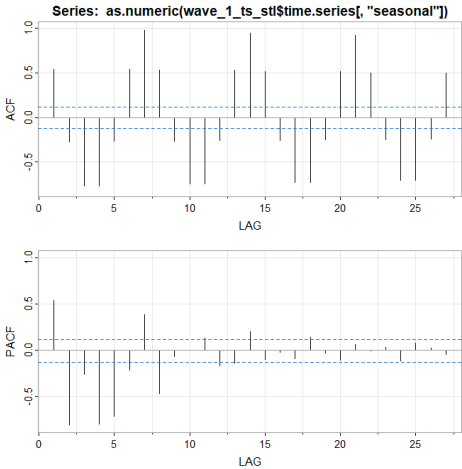
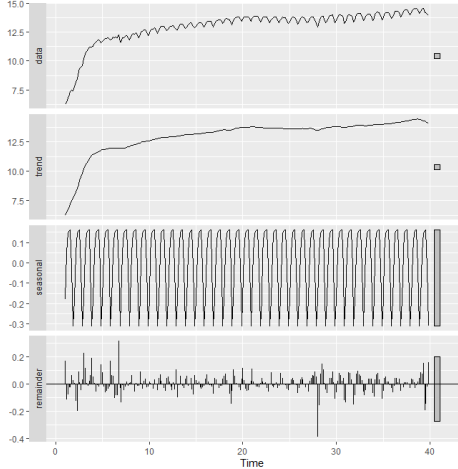
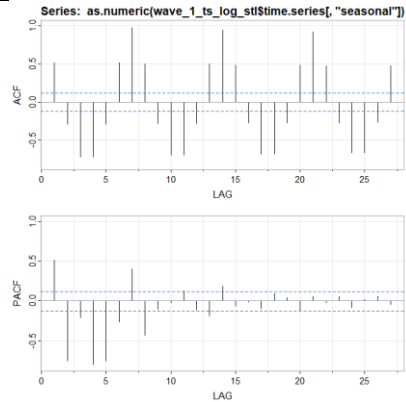
Daily Full Data: Arima Models

<div><div>Augmented Dickey-Fuller Test</div><div>data: fulldata_ts Dickey-Fuller = -1.8588, Lag order = 9, p-value = 0.6381 alternative hypothesis: stationary</div><div>KPSS Test for Level Stationarity</div><div>data: fulldata_ts KPSS Level = 2.5676, Truncation lag parameter = 6, p-value = 0.01</div></div>		
<div><div>1. Formal Tests</div><div>The ADF test has a p-value of 0.6404 so we fail to reject the HO. The KPSS test is 0.01 and so we reject the HO. So both tests suggest the time series is non-stationary and will need to be dealt with this issue.</div></div>	<div><div>2. Graphical Tools</div><div>Again, the full data has trend and seasonality and we notice in the ACF especially there is a very slow decay with a seasonality aspect as well. PACF is also showing a seasonality aspect after 7 lags.</div></div>	<div><div>3. Apply logs</div><div>Applying logs helps the seasonality in the ACF plots but there is still a slow decay. PACF shows an improvement but there is still a seasonality aspect. Not sure it has help.</div></div>
<div><div>4. Logs and Trend Differencing</div><div>Using Logs and Trend Diff improves the long decay but there is still a seasonality, perhaps this is still expected as we have seasonality in the weekly data. PACF Plot is showing a seasonality aspect still.</div></div>	<div><div>5. Logs, Trend Diff and Seasonal Diff=7</div><div>Use of a seasonal differencing of 7 has not made a huge difference. We can clearly still see a seasonal component in both.</div></div>	<div><div>6. Logs, Trend Diff and Seasonal Diff=7, lag=7</div><div>Use of a lag value of 7 has improved the model and reduced the amount of seasonality by a large amount. We can still see a seasonally aspect in both ACF and PACF at the values of 1 and 2, etc.</div></div>

<ul style="list-style-type: none">- No logs, Trend Differencing- Logs, Trend Diff, Trend Diff (2nd time)- Logs, Trend diff, Trend Diff (2nd) Seasonal diff=7- Seasonal diff =7 only (unlikely but attempted)<ul style="list-style-type: none">- Logs, Seasonal diff=7		<p>Augmented Dickey-Fuller Test</p> <p>data: fulldata_ts_log_diff_sdiff_sdiff_lag7 Dickey-Fuller = -68.972, Lag order = 9, p-value = 0.01 alternative hypothesis: stationary</p> <p>KPSS Test for Level Stationarity</p> <p>data: fulldata_ts_log_diff_sdiff_sdiff_lag7 KPSS Level = 0.0086874, Truncation lag parameter = 6, p-value = 0.1</p>
<p>7. More attempts to reduce the seasonality</p> <p>Tried different variations as seen above with the plots in the R file to try and remove the seasonality but with no big changes from these attempts.</p>	<p>8. Logs, Trend Diff, Seasonal Diff =7 Twice, Lag =7</p> <p><u>BEST MODEL</u></p> <p>This model shows the best attempt to reduce the high amount of seasonality in the ACF and PACF plots as we see above. It's a massive improvement and removes the seasonality.</p> <p>I thought it was unusual to do seasonal diff twice but its what worked.</p>	<p>Formal Tests</p> <p>The ADF test now says the data after the changes in Model 8 are stationary so this is a good result for a complex Time Series Model.</p> <p>Our model should be more reliable and accurate now. Which we will test later.</p> <p>The KPSS test is inconclusive but since our ADF is stationary and the ACF/PACF plots look good in Model 8 this is the best result we can get for now.</p>

Daily Wave 1: Classical Wave

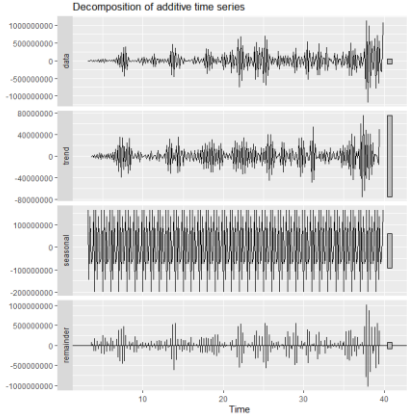
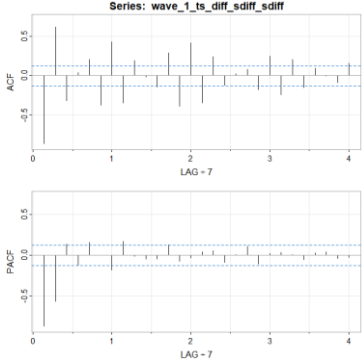
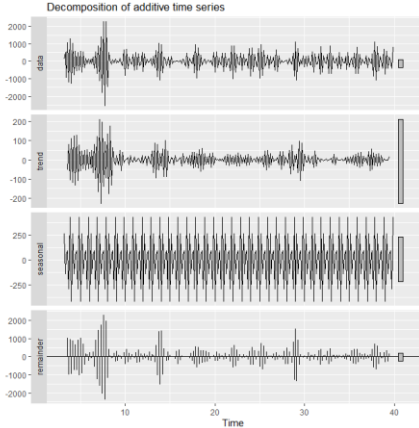
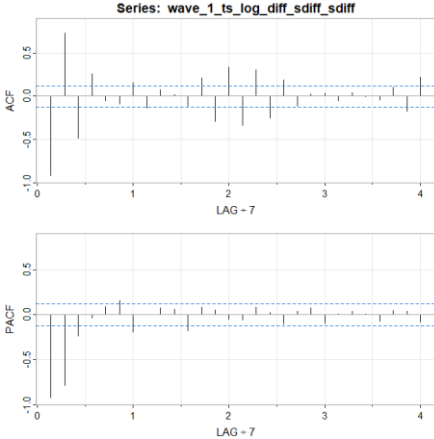
		
<p>1. Wave 1 Time Series</p> <p>The first wave shows a clear upward trend.</p> <p>We can more clearly see a seasonally variation inside each week that has been mentioned before as it is clearly increasing for the first half of the trend and after that becomes more stable.</p>	<p>2. Decomposition with Additive</p> <p>Trend has the biggest affect which makes sense given the big trend in Plot 1. Seasonality looks like it could be better. The random component has a pattern with higher edges so not capturing the data well.</p>	<p>3. Decomposition with Multiplicative</p> <p>The Multiplicative model does better as we can see the random component is much better and more random. This is due to stabilising the variance more than likely. So transformations help again.</p>

		
<p>4. STL Model</p> <p>Trend has the biggest effect which is expected. I'm surprised the 7 day seasonality does not have more. We can see the same randomness as the additive model so we will try transformations again.</p>	<p>STL Model</p> <p>Once again we can see the same seasonality in the ACF part. Similar to what we have seen before.</p> <p>PACF does appear to show a pattern at 7 and 14.</p>	<p>5. STL with Log</p> <p>Used log of the Time Series data to see if it makes an improvement in removing the seasonality. The random component looks better. We can still see the Trend has the biggest effect by the small Shaded box.</p>
		
<p>STL with Log</p> <p>Similar to before, there is a lot of seasonality in the ACF plot but the PACF plot looks good.</p>	<p>MSTL</p> <p>Did not use MSTL as there are not two peaks this time.</p>	<p>Holts Winter</p> <p>Later for forecasting we will use Holts Winter</p>

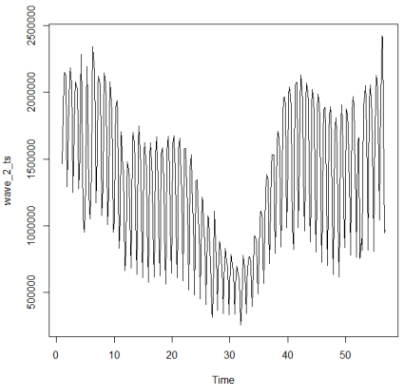
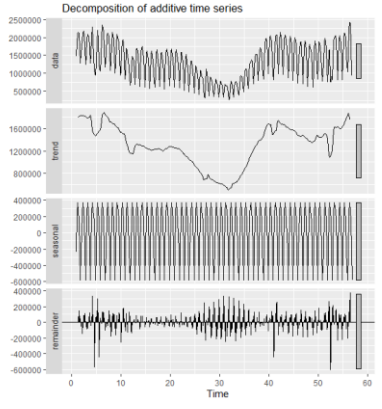
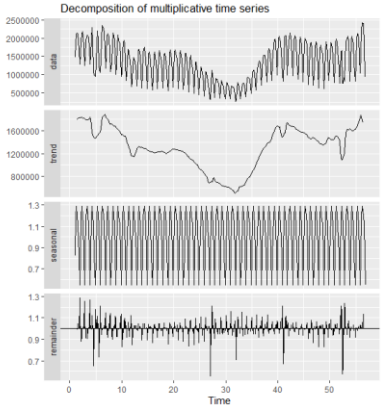
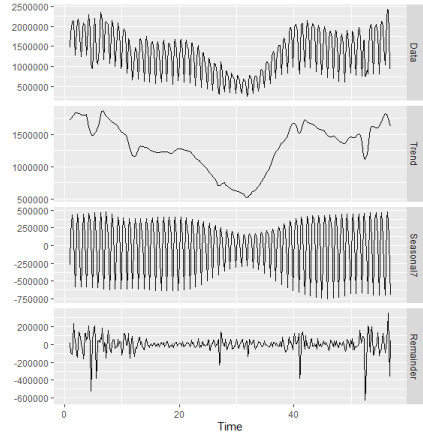
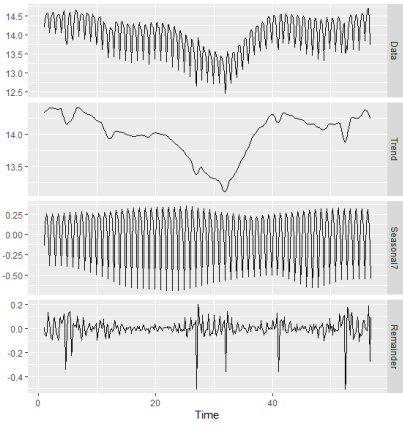
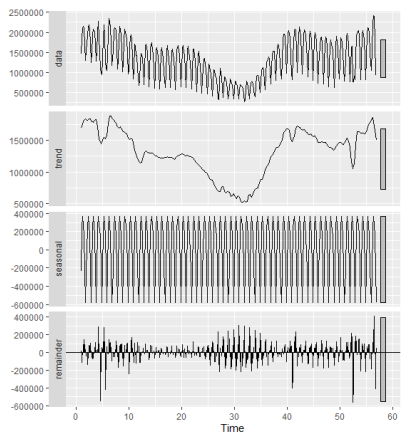
Once again, the Classical models are not capturing the data as well. I believe the ARIMA models will have much better success as there are clearly strong seasonal components in the time series data and SARIMA models might work really well.

Daily Wave 1: Arima Models

<div><p>Augmented Dickey-Fuller Test</p><pre>data: wave_1_ts Dickey-Fuller = -1.4706, Lag order = 6, p-value = 0.799 alternative hypothesis: stationary</pre><p>KPSS Test for Level Stationarity</p><pre>data: wave_1_ts KPSS Level = 4.1647, Truncation lag parameter = 5, p-value = 0.01</pre></div>		
<div><p>1. Formal Tests</p><p>The ADF test the p-value is 0.799 so fail to reject the HO. The KPSS test is 0.01 and so we reject the HO.</p><p>So both tests suggest the time series is non-stationary and will need to be deal with this issue.</p></div>	<div><p>2. Graphical Tools</p><p>Again, the full data has trend and seasonality and we notice in the ACF especially there is a very slow decay with a seasonality aspect as well. PACF is looking ok but we can perhaps improve it as well.</p></div>	<div><p>3. Trend Differencing</p><p>As we have removed Trend Differencing we see that the Trend now has the least importance but there is a smallish pattern at the end. The Random Component is the most important which is a surprise as I would have expected the seasonality still to be a big impact, it still has a reverse pattern. Let’s see next.</p></div>
<div><p>Logs and Trend Differencing</p><p>We can see in the ACF plot that there is still a lot of seasonality but the slow decay has gone. PACF looks to have no seasonality which is a good sign.</p></div>	<div><p>4. Trend Diff and Seasonal Diff=7</p><p>Use of a seasonal differencing of 7 further reduced the Trend so is a good addition. The seasonality and the random component look better.</p></div>	<div><p>Trend Diff and Seasonal Diff=7</p><p>The pattern has been reduced in the ACF plot which is good but the PACF looks ok, maybe a little worse than the last one.</p></div>

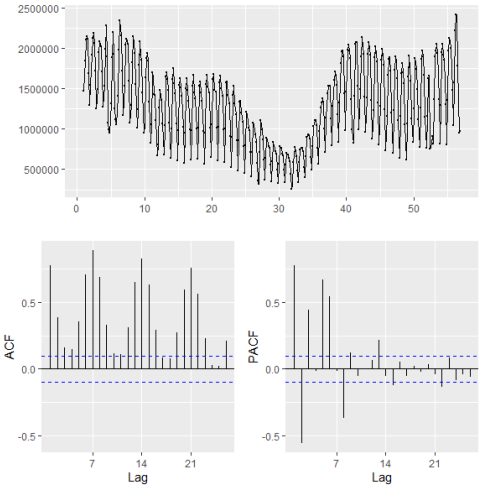
		<ul style="list-style-type: none">- Transform with Box Cox- Box Cox with Trend Diff- Box Cox with Trend Diff, Seasonally Diff =7
<p>5. Trend Diff, Seasonality Diff =7 twice</p> <p>Trend is nearly the same shape so it is not the main drive. The random component has the smallest shaded box to the data so showing the biggest affect now which is a good sign.</p>	<p>Trend Diff, Seasonality Diff =7 twice</p> <p>ACF still has seasonal pattern but the PACF looks good.</p> <p>We will try some models in the code to save space as I am also curious again that I have to use seasonal Differencing twice for the best results.</p>	<p>6. Other attempts in R Code</p> <p>Gaining experience by trying the effects of other options and how they affect the ACF plots in particular</p> <p>No improvements from these options, still seeing a seasonal pattern in the ACF plot.</p>
		<p>Augmented Dickey-Fuller Test</p> <p>data: wave_1_ts_log_diff_sdif_sdif Dickey-Fuller = -45.258, Lag order = 6, p-value = 0.01 alternative hypothesis: stationary</p> <p>KPSS Test for Level Stationarity</p> <p>data: wave_1_ts_log_diff_sdif_sdif KPSS Level = 0.017563, Truncation lag parameter = 5, p-value = 0.1</p>
<p>7. Log, Trend Diff, Seasonality Diff =7 twice, lag=7</p> <p><u>BEST MODEL</u></p> <p>A very similar model to the full data but this time we do not need the lag of 7. The Data, trend and random look similar meaning they are not contributing to the data very much.</p>	<p>Log, Trend Diff, Seasonality Diff =7 twice, lag=7</p> <p>The ACF Plot and the PACF plot look very good with the seasonally pattern gone.</p>	<p>Formal Test</p> <p>Taking the formal tests again we can show with the p-values that the data is now stationary with the ADF results.</p>

Daily Wave 2 Data: Classical Models

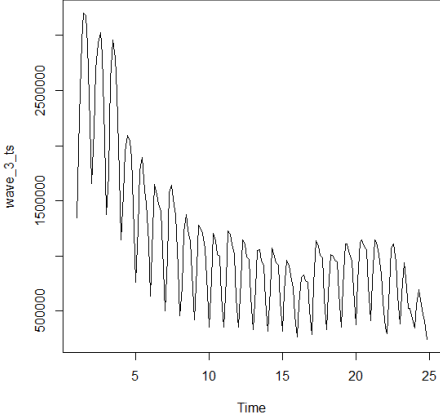
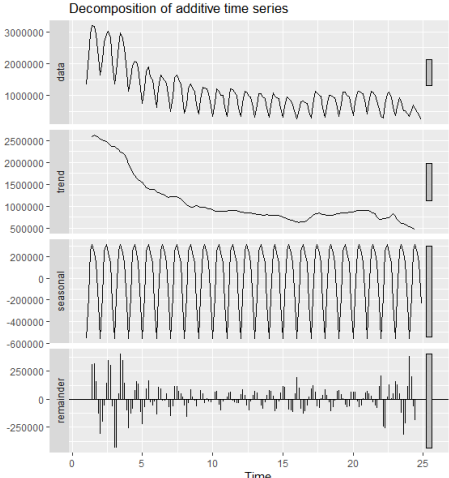
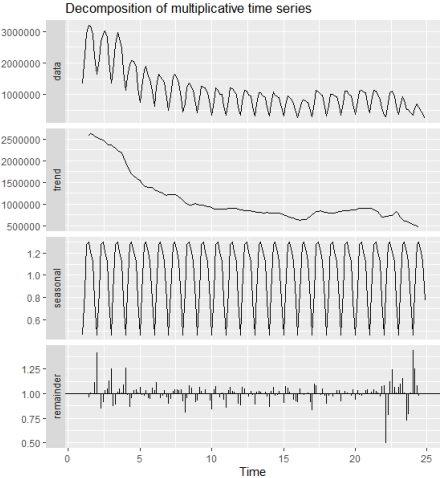
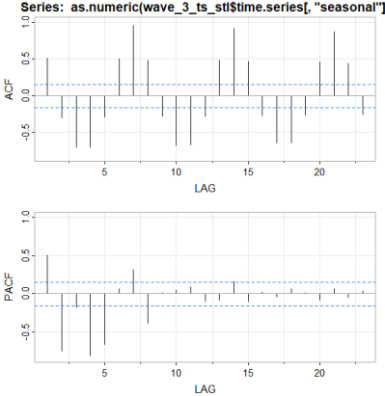
		
<p>1. Wave 2 Time Series</p> <p>The second wave is more interesting, it has less of an increasing in variation in the weekly data as the fluctuations look similar but the trend has a decrease and then an increase when covid dropped in the summer between the two years.</p>	<p>2. Decomposition with Additive</p> <p>Interestingly the shaded boxes do not show any component that shows a big affect on the data. Trend is closest which makes sense given the valley shape of the trend. The Seasonality looks good but the random component has a pattern. Perhaps the valley shape affects the model to cancel out components.</p>	<p>3. Decomposition with Multiplicative</p> <p>The Multiplicative model does not make as much of an improvement in this case so I would stick with the Additive model as we don't need the extra complexity of the Multiplicative model.</p>
		
<p>4. MSTL Model</p> <p>I thought MSTL might offer good results as there are two peaks either side but it does not do a good job as we can see in the seasonality and random component.</p>	<p>5. MSTL with logs</p> <p>MSTL with logs is no better. Which it captures the right period of 7, but the seasonality component does not look good.</p>	<p>6. STL</p> <p>STL is giving the best results at this point. I believe like for the other time series waves that ARIMA models will do a better job at capturing the data.</p>

Daily Wave 2 Data: ARIMA Models

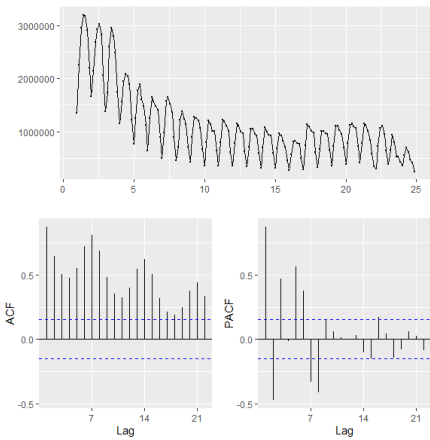
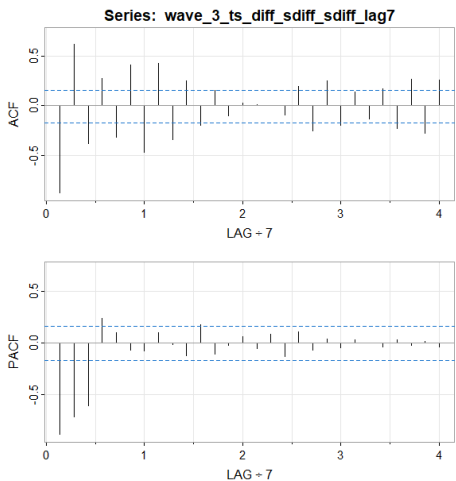
We can Conclude from the first two Daily Datasets that the Classical models are not capturing the complex Time Series patterns as well as the ARIMA models mainly due to the seasonally pattern in the week, so we will try those first with ARIMA models.

<div><div>Augmented Dickey-Fuller Test</div><div>data: wave_2_ts Dickey-Fuller = -1.7746, Lag order = 7, p-value = 0.6724 alternative hypothesis: stationary</div><div>KPSS Test for Level Stationarity</div><div>data: wave_2_ts KPSS Level = 1.1386, Truncation lag parameter = 5, p-value = 0.01</div></div>		<ul style="list-style-type: none">- Transform with logs- logs with Trend Diff- logs with Trend Diff, Seasonally Diff =7- Trend Diff, Seasonally Diff=7 Twice, Lags =7- Logs with Trend Diff, Seasonally Diff=7, lags=7- Logs with Trend Diff, Seasonally Diff=7 Twice, lags=7
<div><div>1. Formal Tests</div><div>The ADF test the p-value is 0.6724 so fail to reject the HO. The KPSS test is 0.01 and so we reject the HO. So both tests suggest the time series is non-stationary and will need to be deal with this issue.</div></div>	<div><div>2. Graphical Tools</div><div>We do not have the long decay we had before. But we can still see a large seasonality pattern in the ACF plot.</div></div>	<div><div>3. Options Tried</div><div>Difficult to capture the Time Series model for this wave as the complexity of the weekly seasonally we have seen before but also the quick trend down and then increase. We are not able to find a model that is stationary.</div><div>More time might help but I am also curious how the Auto.ARIMA will do and will perhaps be able to help.</div></div>

Daily Wave 3 Data: Classical Models

		
<p>1. Wave 2 Time Series</p> <p>The third wave has a clear downward trend with a very large clear weekly seasonality. There is also some variation in the weekly height of the seasonality but not as much as seen in wave 1, perhaps more than wave 2.</p>	<p>2. Decomposition with Additive</p> <p>Trend has the biggest affect which is expected given the clear downward trend. Seasonality looks good but the random component is still seeing the weekly seasonality inside of it.</p>	<p>3. Decomposition with Multiplicative</p> <p>The Multiplicative model does not make as much of an improvement in this case, while the random component has lower heights, there is still a pattern in it.</p>
		
<p>4. STL Model</p> <p>No Surprise, once again we are still seeing the classical models not able to capture the weekly seasonality well.</p>	<p>Holts Winter</p> <p>Later for forecasting we will use Holts Winter</p>	

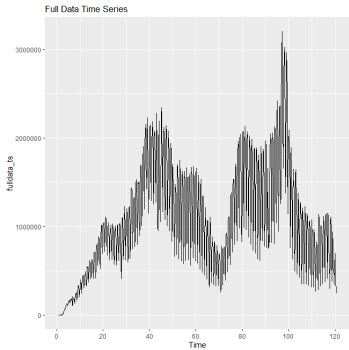
Daily Wave 3 Data: ARIMA Models

<div><div>Augmented Dickey-Fuller Test</div><div>data: wave_3_ts Dickey-Fuller = -2.4946, Lag order = 5, p-value = 0.3698 alternative hypothesis: stationary</div><div>KPSS Test for Level Stationarity</div><div>data: wave_3_ts KPSS Level = 2.2276, Truncation lag parameter = 4, p-value = 0.01</div></div>		
<div><div>1. Formal Tests</div><div>The ADF test the p-value is 0.3698 so fail to reject the HO. The KPSS test is 0.01 and so we reject the HO. So both tests suggest the time series is non-stationary and will need to be deal with this issue.</div><div>The ADF is the lowest we have seen but still high enough to reject the HO.</div></div>	<div><div>2. Graphical Tools</div><div>Once again we see a long decay with the ACF plot and a seasonality patter. So no overall changes. We also see a regular pattern in the PACF plot.</div><div>We can save time by trying similar Differencing we saw for other waves</div></div>	<div><div>3. Log, Trend Diff, Seasonality Diff =7 twice, lag=7</div><div>Once again, to find the best model we have very similar steps to remove the non-stationary and make the data stationary. The formal tests and plots back this up in the code.</div></div>

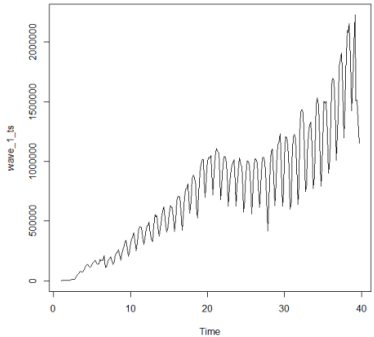
Comparison of models in each wave

Classical models did not capture the data as well as ARIMA models so we have looked at those in more detail to compare the AIC values of the better models for each Daily Dataset.

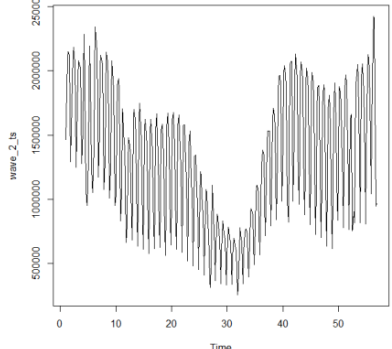
Full Data: Comparison of models

	Models:	AIC:	The best model is the last. Yes the AIC value is the largest but an AIC score balances Goodness of Fit and Level of Complexity.
	Log, Seasonally =7	5111.68	
	Log, Trend Diff, Seasonally Diff =7	6140.77	
	Log, Trend Diff, Seasonally Diff Twice, lag=7	10883.82 Best Model	The first two examples do not explain the data as well as the third model.

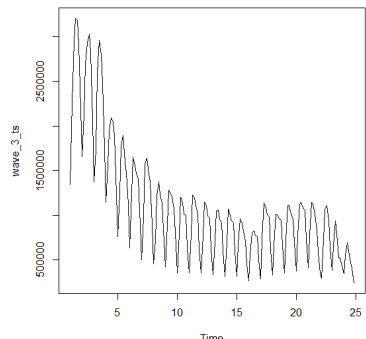
Wave 1: Comparison of models

	Models:	AIC:	<p>The best model results comes from the Third model which has the lowest AIC Score of 4017.16.</p> <p>We can notice big improvements in the 3rd model's AIC value even as the model is more complex. A great result.</p>
	Trend Diff, Seasonally Diff = 7 Twice.	10793.82	
	Trend Diff, Seasonally Diff = 7 Twice, lag=7	8594.18	
	Log, Trend Diff, Seasonally Diff Twice	4017.16 Best Model	

Wave 2: Comparison of models

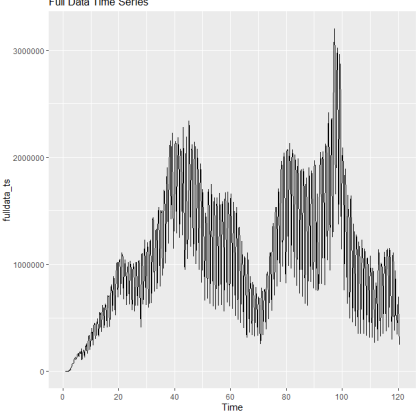
	Models:	AIC:	<p>Much harder time series to remove the Trend and Seasonality and so not as good results.</p> <p>We can see the AIC levels go up as the models get more complex. I was not able to do as well as the other models for this data. Perhaps given its quick Trend down and then up again.</p>
	Log, Trend Diff, Seasonally Diff=7	2838.61	
	Log, Trend Diff, Seasonally Diff=7 Twice	6244.45	
	Log, Trend Diff, Seasonally Diff=7 Twice, Lag=7	6730.05 Best Model	

Wave 3: Comparison of models

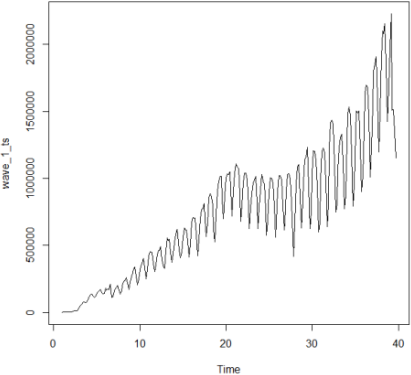
	Models:	AIC:	<p>The best model results comes from the Third model which has the lowest AIC Score of 1661.22.</p> <p>Again, big improvements in models as we removed Trend and the seasonality.</p> <p>The lowest AIC value is the lowest which is a great result given the complexity increases.</p>
	Trend Diff, Seasonally Diff = 7 Twice.	6525.2	
	Trend Diff, Seasonally Diff = 7 Twice, lag=7	6212.64	
	Log, Trend Diff, Seasonally Diff Twice	2661.22 Best Model	

1.1 Auto Arima Models and compare to models

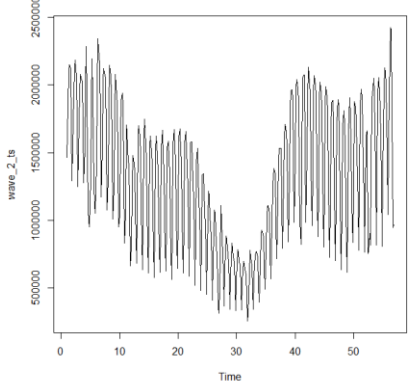
Full Data: Comparison of Auto.ARIMA and Manual modelling

	ARIMA Models	Comparison and Best Model
	My Model ARIMA(1,1,1)(3,2,1)[7] AIC=10883.82	Auto.ARIMA first picks up the Seasonality as 365 and this was something I tied as well but it did not explain the time series as well. It also caused issues with some classical models as the frequency was too high. When Seasonality was used it rightly picks up the seasonal pattern of 7. While my model was more complex with a second Seasonal Diff, Log and Lag. My AIC score was the best with a value of 10883.82. These are complex time series and the seasonality in the ‘year’ and ‘week’ elements is hard to pick up. My manual results did better than the Auto.ARIMA in this case.
	Auto. ARIMA ARIMA(3,1,3)(0,1,0)[365] AIC = 13043.6	
	Auto.ARIMA (Seasonality) SARIMA(3,0,2)(0,1,2)[7] AIC=21572.28	

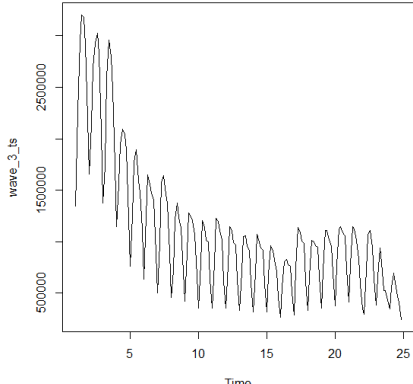
Wave 1 Data: Comparison of Auto.ARIMA and Manual modelling

	ARIMA Models	Comparison and Best Model
	My Model ARIMA(1,1,1)(3,2,1)[7] AIC=10883.82	Same results for both Auto.ARIMA models, not surprising as there is no second seasonality that confuses the model with this wave. Interesting with the Auto.ARIMA models have not included Trend Differencing which I would have expected in this wave especially. Showing that Auto.ARIMA models are not always the best. My model gets an AIC score of 4017.16 so once again is both more complex but has a much lower score so my model has done better.
	Auto. ARIMA ARIMA(1,0,1)(0,1,1)[7] AIC=6637.4	
	Auto.ARIMA (Seasonality) SARIMA(1,0,1)(0,1,1)[7] AIC=6637.4	

Wave 2 Data: Comparison of Auto.ARIMA and Manual modelling

	ARIMA Models	Comparison and Best Model
	My Model ARIMA(1,1,1)(2,2,1)[7] AIC=6730.05	No change in the seasonality so I am not sure the Auto.ARIMA models are picking up the seasonality as well as my manually modelling due to the high AIC score and model format. My AIC of 6730.05 is lower but again is a more complex model. I would choose mine at this point.
	Auto. ARIMA ARIMA(3,0,2)(1,1,2)[7] AIC=10093.69	
	Auto.ARIMA (Seasonality) SARIMA(3,0,2)(1,1,2)[7] AIC=10093.69	

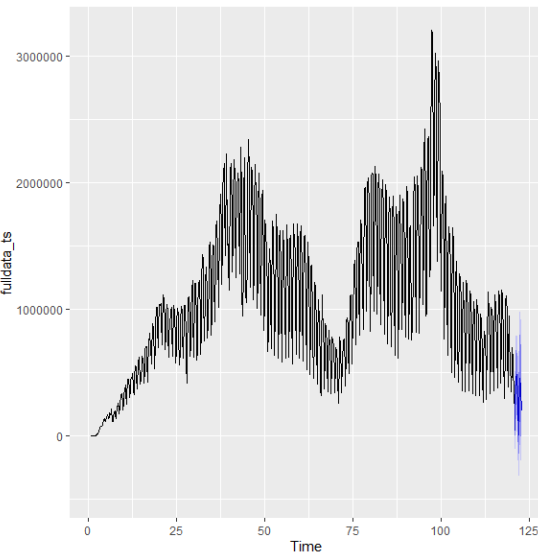
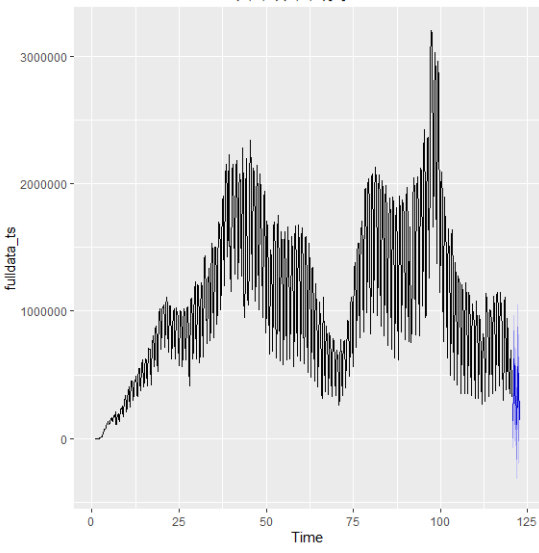
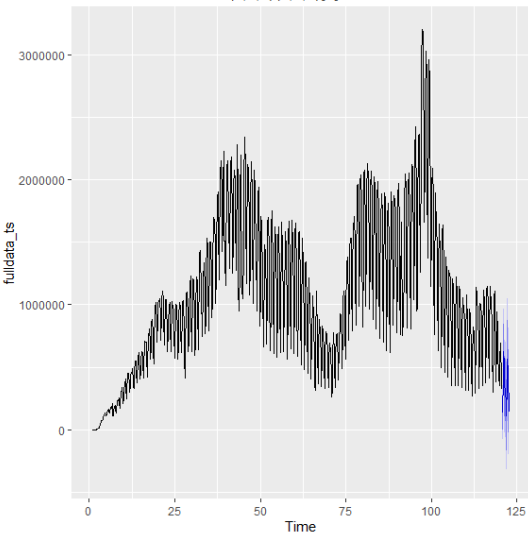
Wave 3 Data: Comparison of Auto.ARIMA and Manual modelling

	ARIMA Models	Comparison and Best Model
	My Model ARIMA(1,1,1)(1,2,2)[7] AIC=2661.22	This time it Auto. ARIMA has used Trend Differencing which surprises me when it didn't for Wave 1. My score is 2661.22 which once again is an improvement even given the greater complexity so I would choose my model again.
	Auto. ARIMA ARIMA(1,1,3)(0,1,1)[7] AIC=4141.33	
	Auto.ARIMA (Seasonality) SARIMA(1,1,3)(0,1,1)[7] AIC=4141.33	

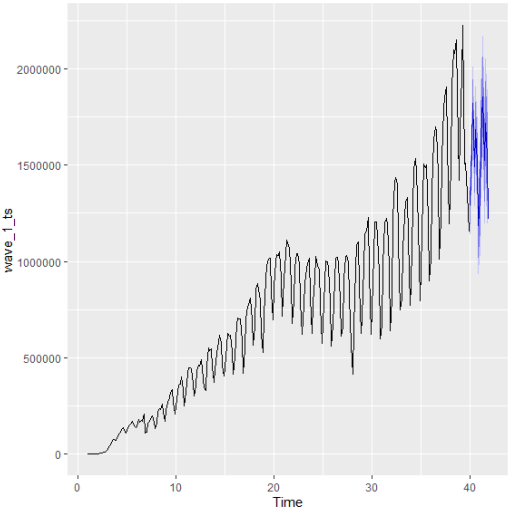
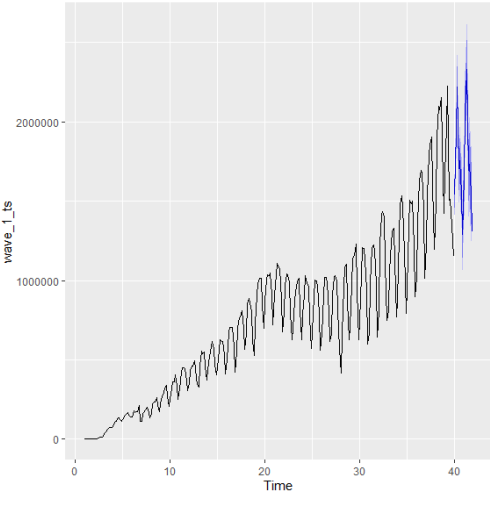
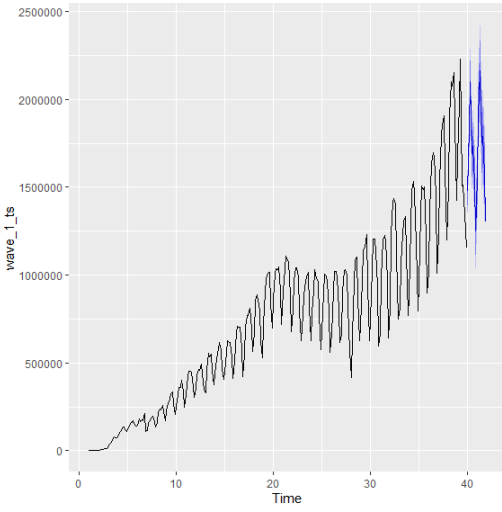
Forecast

Compare Classical, Manual ARIMA and Auto.ARIMA

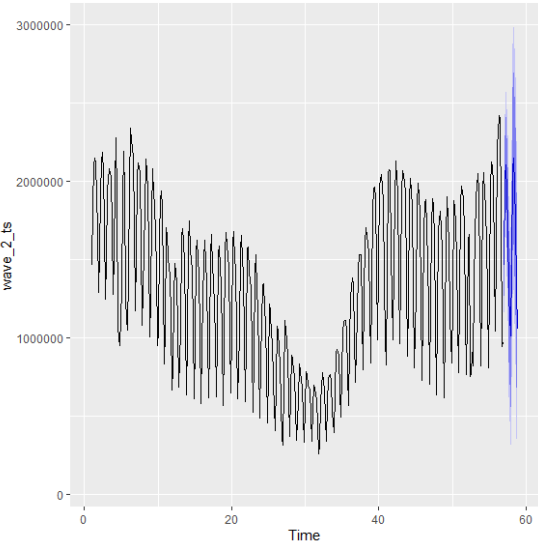
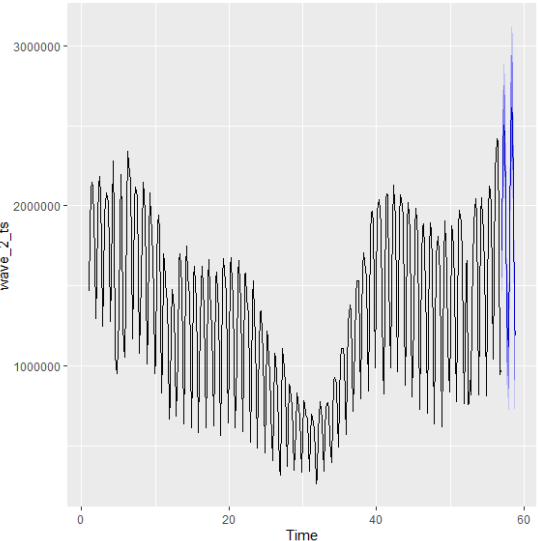
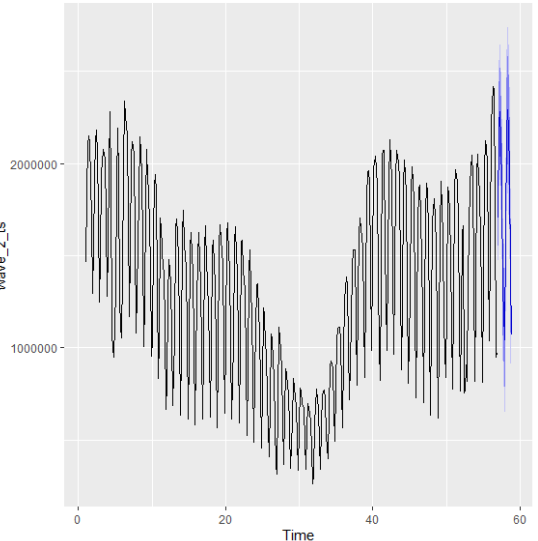
Full Data: Comparison of Forecasts

Classical Model: Holts Winter	Manual ARIMA Model	Auto.ARIMA Models
<div><p>Forecasts from Holt-Winters' additive method</p></div> <p>The classical model does the least well of the three. We have seen this while looking to explain the different parts of the data so this is not surprising.</p> <p>The Confidence Intervals are a lot bigger in the Holts Winter model so it is less sure of its forecast.</p>	<div><p>Forecasts from ARIMA(3,1,1)(1,2,1)[7]</p></div> <p>My manual model does the best when you zoom in. It is not very clear but the area where it bets the Auto.ARIMA model is the shaded areas. The Confidence Areas are better in my model.</p> <p>The model itself is more complex but the AIC value is 10883.82 which is less than half the Auto.ARIMA value and so this complexity is worth it for a better model.</p>	<div><p>Forecasts from ARIMA(3,1,1)(1,2,1)[7]</p></div> <p>The Auto.ARIMA model also does very well and only loses out to the manual model as the AIC value is much lower.</p>

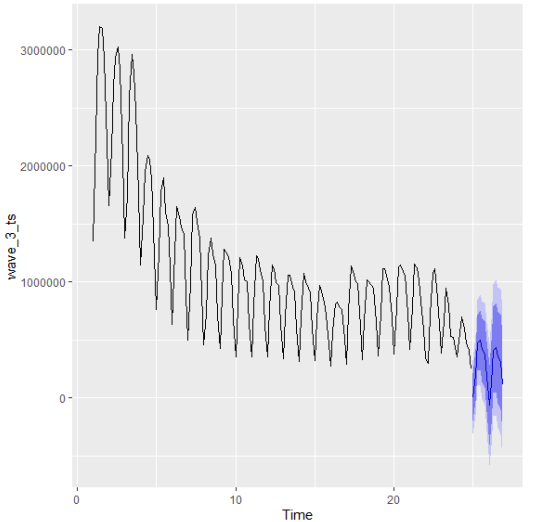
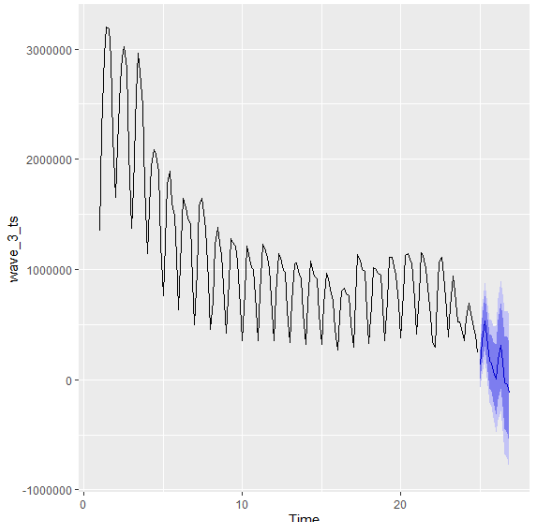
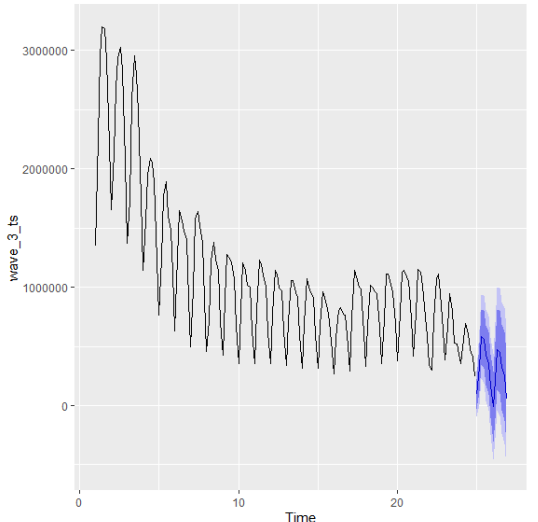
Wave 1 Data: Comparison of Forecasts

Classical Model: Holts Winter	Manual ARIMA Model	Auto.ARIMA Models
<div><p>Forecasts from Holt-Winters' additive method</p></div>	<div><p>Forecasts from ARIMA(2,1,1)(1,2,1)[7]</p></div>	<div><p>Forecasts from ARIMA(1,0,1)(0,1,1)[7] with drift</p></div>
<p>We can see a clear dip in the forecasted line and the confidence intervals are much bigger.</p> <p>This is not surprisingly as from earlier in the report the classical models were not doing as well capturing the patterns in the data.</p>	<p>The Manual model has done very well. It looks like a very realistic forecast for this model and the confidence intervals are very small so the model is confident of the range outside of this.</p> <p>Once again, the AIC is much better than the Auto.ARIMA model so I would choose this model. However, the Auto.ARIMA model is perhaps slightly better.</p>	<p>The Auto.ARIMA model has also done very well, perhaps just a little better but Its AIC value is much higher.</p> <p>Hard one to say at this point. My model is more complex but has a lower AIC value. The AA model looks a little better. I would go with my model due to the AIC values in this case.</p>

Wave 2 Data: Comparison of Forecasts

Classical Model: Holts Winter	Manual ARIMA Model	Auto.ARIMA Models
<div>Forecasts from Holt-Winters' additive method</div> 	<div>Forecasts from ARIMA(2,1,1)(1,2,1)[7]</div> 	<div>Forecasts from ARIMA(3,0,2)(1,1,2)[7]</div> 
<p>The holts winter in this case has perhaps done better than my model. It has large confidence intervals but they are not as high in range which I think is better.</p>	<p>I knew my model had not captured all the data so in this case the Auto.ARIMA is the better result. I found it hard to capture all the patterns in the time series. Its not bad, just is a little too high range with the confidence intervals.</p>	<p>This time the Auto.ARIMA model has done the best but I expected this as I was finding it hard to completely remove the seasonally patterns from my manual ARIMA models.</p> <p>My AIC values are better but In this case I would investigate more to see if I could improve the results of the forecasting.</p>

Wave 3 Data: Comparison of Forecasts

Classical Model: Holts Winter	Manual ARIMA Model	Auto.ARIMA Models
<div>Forecasts from Holt-Winters' additive method</div> 	<div>Forecasts from ARIMA(2,1,1)(1,2,1)[7]</div> 	<div>Forecasts from ARIMA(1,1,3)(0,1,1)[7]</div> 

The Holts Winter models has once again the highest Confidence intervals so the other two ARIMA models look better. Classical Models do now capture the data as well.	Very similar results. You can see that the peaks in the forecasting are a little sharper and with the lower AIC values I would chose my model.	The Auto.ARIMA model also does a good job in this situation. Its AIC values is higher than my model and my confidence intervals look better I think.
---	--	--

Power of Prediction

The question did not specify train and test so I took all the data for some options, then took the RMSE value and then compare it to my ARIMA model and the Auto.ARIMA Models to see how they compared.

	ARIMA model RMSE	Auto.ARIMA Model RMSE	Full data RMSE	Last 10% of values RMSE
Fulldata_ts	108047.1	101529.7	111041.7	87501.62
Wave_1_ts	61163.89	61163.89	62089.5	105486.5
Wave_2_ts	114641.3	114641.3	126195.6	221604.5

We see some patterns when we compare the results. For the Full data, my model and the Auto.ARIMA models had different results. But for the two waves they were the same. Perhaps because the size of the time series was smaller or both had similar results, which they did.

We also notice a pattern where when the Full data was taken for the RMSE value it was much more accurate than taking the last 10% of the Time Series data to get an RMSE value. The questions said to use the last points so I choose 10% as a value to take to compare. This is interesting and shows that taking the last value and using that for forecasting is not as beneficial as using as a larger time series section of data. At least in this situations.

At times, especially for the Second Wave the RMSE value for the last 10% was much worse. Double in this case.