LETTERKENNY INSTITUTE OF TECHNOLOGY
ASSIGNMENT COVER SHEET

Lecturer's Name: **Dr James Connolly**

Assessment Title: **CA 4 – Predictive Modelling**

Work to be submitted to: _____

Date for submission of work: **2020-05-24**

Place and time for submitting work: _____

To be completed by the Student

Student's Name: **Danny Regan**

Class: **MSc in Big Data Analytics**

Subject/Module: **Data Science**

Word Count (where applicable): _____

I confirm that the work submitted has been produced solely through my own efforts.

Student's signature: _____Date: _____

Notes

Penalties: The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero.

Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment shall normally be carried forward from the original examination to the repeat examination.

Declaration:

I declare that this work is entirely my own and does not contain the words or ideas of someone else, whether published or not, without specific acknowledgement by relevant referencing. I have read and understood the LYIT Plagiarism Policy on the "Student & Academic Policies" section of the LYIT Website and understand plagiarism to include:

- Direct copying of text, images and other materials (electronic or otherwise) from a book, article, fellow student's essay, handout, web page or other source without proper acknowledgement.
- Claiming individual ideas derived from a book, article etc. as one's own and incorporating them into one's work without acknowledging the source of these ideas.
- Overly depending on the work of one or more other sources without proper acknowledgement of the source, by constructing an essay, project etc., extracting large sections of text from another source and merely linking these together with a few of one's own sentences.

I understand that it is my responsibility to familiarise myself with and to follow the Institute's Assessment Regulations. I acknowledge that Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations and that penalties will be applied if I breach this policy.

Signed:_____ Date:_____

CA 4 - Predictive Modelling

Health in Ireland

Title:	Forecasting Saolta University Hospital Group Waiting List Figures
Author:	Danny Regan
Supervisor:	Dr James Connolly
Degree:	MSc in Big Data Analytics
Module:	Data Science
Github:	https://github.com/ancodia/hospital_waiting_lists

Abstract

This project aims to determine how accurately waiting list figures for the Saolta University Hospital Group can be forecasted through predictive modelling. Time series forecasting with ARIMA modelling was chosen as the method to use for predictions due to the time-based nature of the source data. ARIMA(1,1,0) with drift included was found to offer the greatest accuracy levels after comparison of all model configurations implemented.

A correlation accuracy of 83% was found between the actual values and those predicted with the ARIMA(1,1,0) with drift model which implies a high level of confidence in the accuracy of predictions. This ARIMA configuration was then used to forecast monthly waiting list totals for 2020. A mean percentage increase of 3.2% was found between the actual totals from 2019 and those forecasted for 2020, indicating that the increasing trend in the data is likely to continue upwards.

Introduction

The volume of patients waiting for hospital procedures and the length of these waits constitute a major shortfall in the Irish public healthcare system. According to the most recent Euro Health Consumer Index (Health Consumer Powerhouse 2018, p.15), Ireland has the longest waiting times in Europe despite having one of the greatest levels of expenditure on health (OECD and European Union 2018, p.133).

The National Treatment Purchase Fund (NTPF) is the organisation assigned the task by the Irish government of collecting, collating and validating data about individuals who are waiting for treatment in public hospitals. This is the source of data for the current project. The objective of this project is to apply predictive modelling to the NTPF waiting list data, evaluate its accuracy and forecast future numbers of patients waiting for hospital procedures. The Saolta University Hospital Group which hospitals in the west and north-west of Ireland will be the subject of this experimentation.

The next section (Predictive Model Selection) of this document covers the model selection process including justification through analysis and visualisations. Following that in the Build and Evaluate Predictive Model section is a discussion around the construction and evaluation of the selected predictive model configurations. The accuracy of the forecast returned by the models is documented and the best performing one is identified in the section called Model Validation. Finally in the Model Forecasting and Appraisal section, the chosen model is used to predict future waiting list numbers for the Saolta group and these are compared to the latest year from the source data. All R code referenced in this document is found in `predictive_modelling/time_series_predictions.R` which can be accessed in the Github repository listed on the cover page.

Research Question

The NTPF waiting list data will be used in this project to answer the following research question:

> How accurately can waiting list figures for the Saolta University Hospital Group be forecasted using predictive modelling?

Predictive Model Selection

Time series forecasting is the most suitable method for predictive modelling on the waiting list due to its time-based nature. The monthly totals of patients waiting for hospital procedures are the main point of interest and autoregressive integrated moving average (ARIMA) modelling is fit for this task. This section discusses the preparation of waiting list data for the Saolta University Hospital Group for predictive

modelling and includes analysis of the constitution of the data to determine what type of ARIMA model should be implemented.

The combined waiting list data that was created during CA3 is loaded and records for the Saolta University Hospital group extracted from it. The day part of the archive date variable is excluded so that it can be guaranteed that only one record per month is in the resulting dataset. Monthly waiting list totals are then aggregated and the number of rows present is now 72 as expected - 12 months x 6 years (2014-19).

```
waiting_lists <- read_csv("data_prep/combined_waiting_lists.csv")

saolta <- subset(waiting_lists, Hospital_Group == "Saolta")

saolta$Archive_Date <- format(as.Date(saolta$Archive_Date,
                                      format="%Y-%m-%d"),
                              "%Y-%m")

saolta <- aggregate(cbind(Total) ~
                      Archive_Date,
                    data = saolta, sum)
nrow(saolta)
```

```
## [1] 72
```

To facilitate ARIMA forecasting, the waiting list data is converted to a time series object. A frequency parameter of 12 is used because the time points found in the source data are monthly and the series is set to begin from January 2014. The content of the time series can be seen in Table 1 below.

```
saolta_ts <- ts(saolta$Total, frequency = 12, start = c(2014, 1))
```

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2014	64351	70105	71743	73760	74442	73853	75933	77240	78870	79422	81328	87585
2015	88784	90709	92611	94974	94446	92477	92487	93211	92076	90953	85676	82791
2016	84328	85118	86299	88091	89099	89885	90962	91652	92290	92636	93437	94542
2017	95784	96980	98626	100604	101702	102839	104249	105063	105096	105082	105101	105939
2018	105221	104076	104741	104804	104832	104788	104863	105064	104547	105052	105510	104987
2019	107203	110075	110833	111803	112612	114542	115328	116622	116588	115910	115755	114253

Table 1: Saolta waiting lists time series.

There are values for each month so no additional effort is required for cleaning the time series. Figure 1 features a plot of the time series. The time series is additive due to the consistent growth with no dramatic spikes in the peaks and troughs. The plot shows signs of an upward trend which needs to be investigated further and removed before deciding on which ARIMA model parameters to use. This is handled in the next section. The `graphics::abline()` function provides the straight line in the chart

and indicates that the data has quite a strong linear relationship between numbers of patients waiting and time.

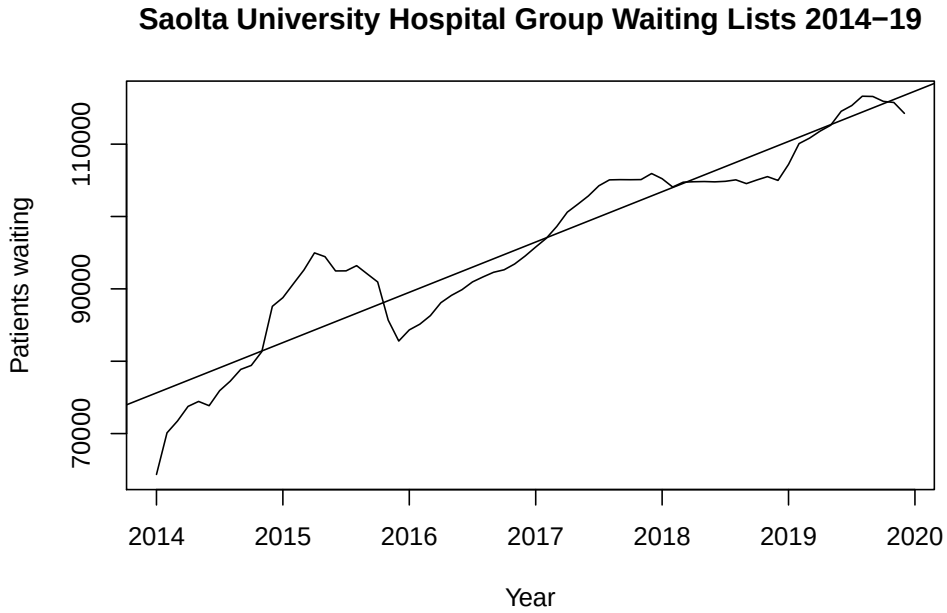

Figure 1: Plot of waiting list time series.

A visual check for seasonality can be achieved with a box plot of the cycles contained in the time series (Figure 2). The median monthly value remains reasonably consistent throughout meaning that no strong link between total number of patients on waiting lists and the time of year is present. Statistical validation of this is featured in the next section of this document.

Build and Evaluate Predictive Model

This section discusses the steps taken in determining the appropriate parameters to use for ARIMA modelling of the time series data and the construction and evaluation of the resulting models. To build a non-seasonal ARIMA model the following values are required:

- `p`: the number of autoregressive terms from the autocorrelation function (ACF) - AR order.
- `d`: the number of non-seasonal differences needed to make the time series stationary.
- `q`: the number of lagged forecast errors from the partial autocorrelation function (PACF) - MA order.

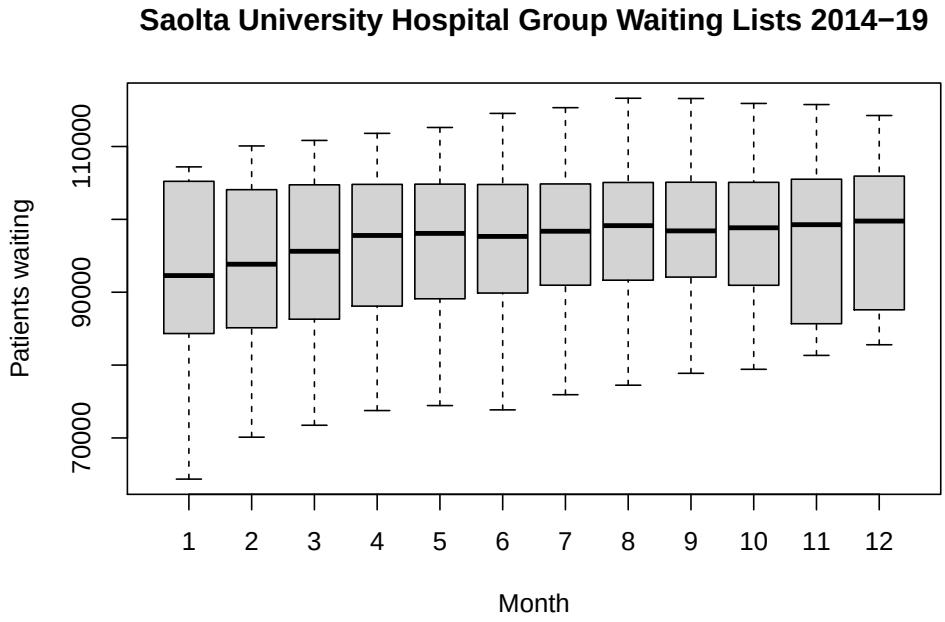

Figure 2: Box plot of waiting list time series.

Initially, the value for `d` will be found by verifying that the assumption of non-seasonality is true and then applying differencing to the time series to introduce stationarity. The `p` and `q` values are found by plotting the ACF and PACF of the stationary time series respectively.

Stationarity and Seasonality

Although the decomposed time series plot (Figure 3) appears to reveal seasonality in the data, the box plot of the series (Figure 2) tells otherwise.

The proportion of variance that each element in the time series makes up can help to determine if there is a higher level of seasonality than expected:

```
apply(saolta_ts_decomposed$time.series, 2, var) / var(saolta_ts)
```

```
##   seasonal      trend  remainder
## 0.00458736 0.97523453 0.01495276
```

Seasonality explains only 0.005% of variance in the time series, confirming the assumption of the lack thereof. The `seastests::isSeasonal()` function also offers a method of checking for seasonality and returns false as anticipated:

```
seastests::isSeasonal(saolta_ts)
```

```
## [1] FALSE
```

Trend accounts for almost all variance in the time series so it needs to be removed to make the data stationary and enable ARIMA modelling. Non-stationarity can be

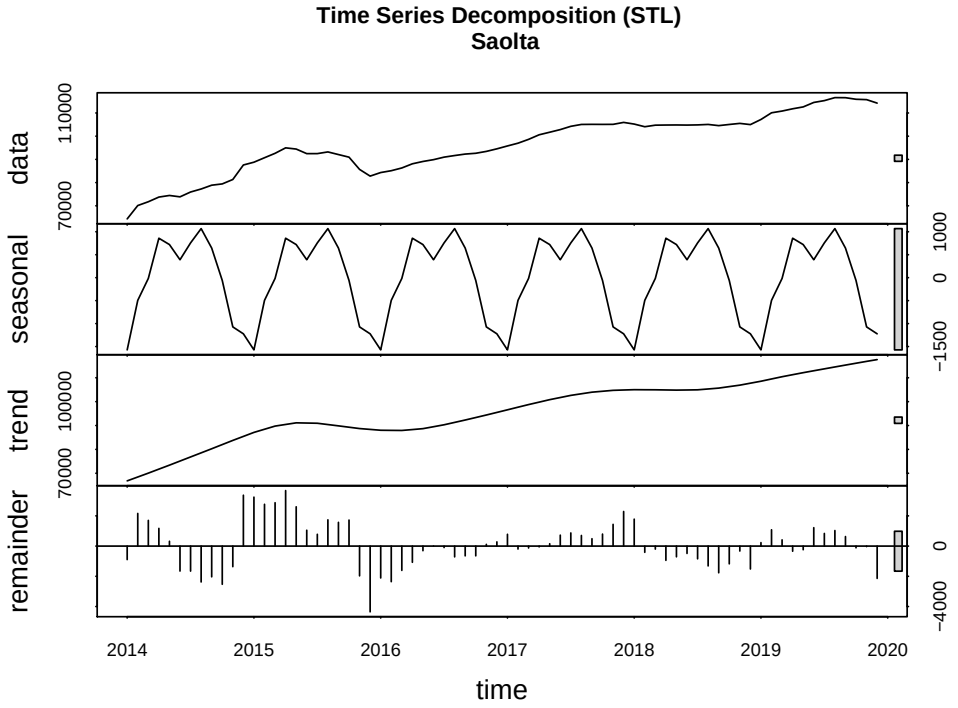

Figure 3: Decomposed time series.

visualised by plotting the autocorrelation function (ACF) applied to the data, see Figure 4. The slow drop off towards 0 seen in the ACF plot demonstrates that the data is not stationary. The ACF of a stationary time series will drop quickly. Differencing must be applied to the data to get it into a stationary form.

Find the ARIMA d value

Differencing must be applied to the time series to make it stationary, the ARIMA `d` value is the number of differences required. The `forecast::ndiffs()` function returns an estimation of how many times differencing should be applied for stationarity to be introduced. In this case, one difference is the value returned:

```
ndiffs(saolta_ts)
```

```
## [1] 1
```

To apply differencing to the time series, the `diff()` function is applied:

```
saolta_ts_diff <- diff(saolta_ts, differences = 1)
```

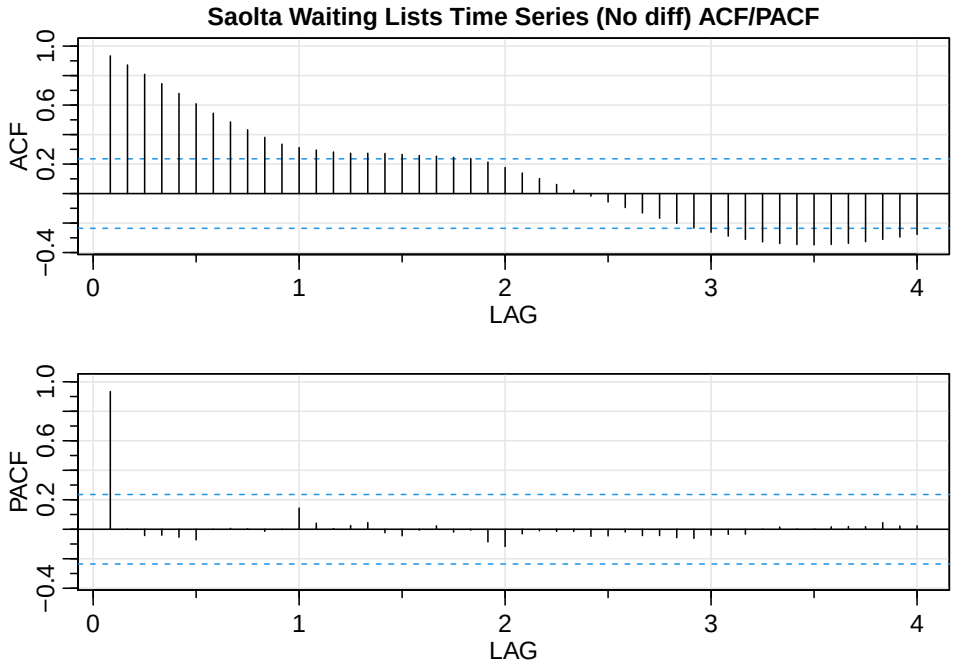

Figure 4: ACF/PACF plots of the time series before differencing.

The plot of the differenced time series is shown in Figure 5 and appears to be stationary. Running `ndiffs()` on the differenced time series now shows that no more differences are necessary for stationarity:

```
ndiffs(saolta_ts_diff)
```

```
## [1] 0
```

To verify that the time series is now stationary, the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests are applied. The H_0 for ADF is that the time series is not stationary, while H_0 for KPSS is that it is stationary.

ADF (note: using the ADF function from the `urca` package rather than the `tseries` version as it can be run with including drift/trend (already removed with differencing):

```
adf_pvalue <- urca::ur.df(saolta_ts_diff)@testreg[["coefficients"]][1,4]
adf_pvalue
```

```
## [1] 0.0002702675
```

p-value is 0.00027 so the null hypothesis of non-stationarity can be rejected.

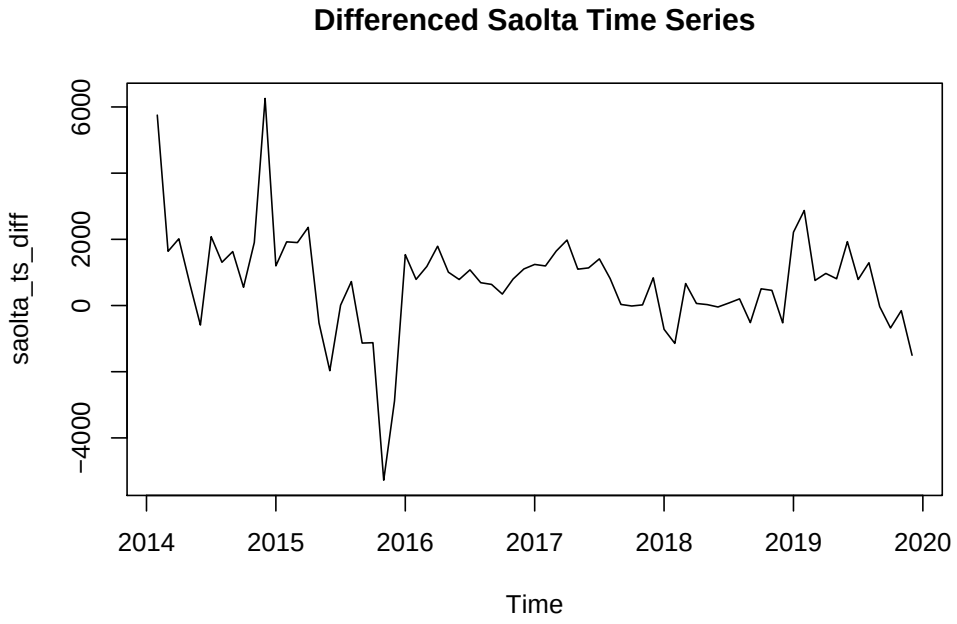

Figure 5: Differenced time series.

KPSS:

```
kpss <- tseries::kpss.test(saolta_ts_diff)
```

```
## Warning in tseries::kpss.test(saolta_ts_diff): p-value greater than printed p-
## value
```

```
kpss$p.value
```

```
## [1] 0.1
```

A p-value of greater than 0.1 means that the null hypothesis cannot be rejected thus confirming that the time series is in a stationary state.

This one required difference gives an optimum `d` value of 1 for the ARIMA model.

ARIMA p and q values from ACF/PACF

When the ACF and PACF are plotted for the differenced time series (Figure 6), they both quickly drops below the dotted line indicating that the majority of values are not significantly different from 0. The `p` and `q` values can now be taken from this plot, both cut off after 1 lag so this is the value assigned to both. Along with the `d` value of 1, this provides three possible ARIMA(p,d,q) configurations: ARIMA(1,1,0), ARIMA(0,1,1) and ARIMA(1,1,1).

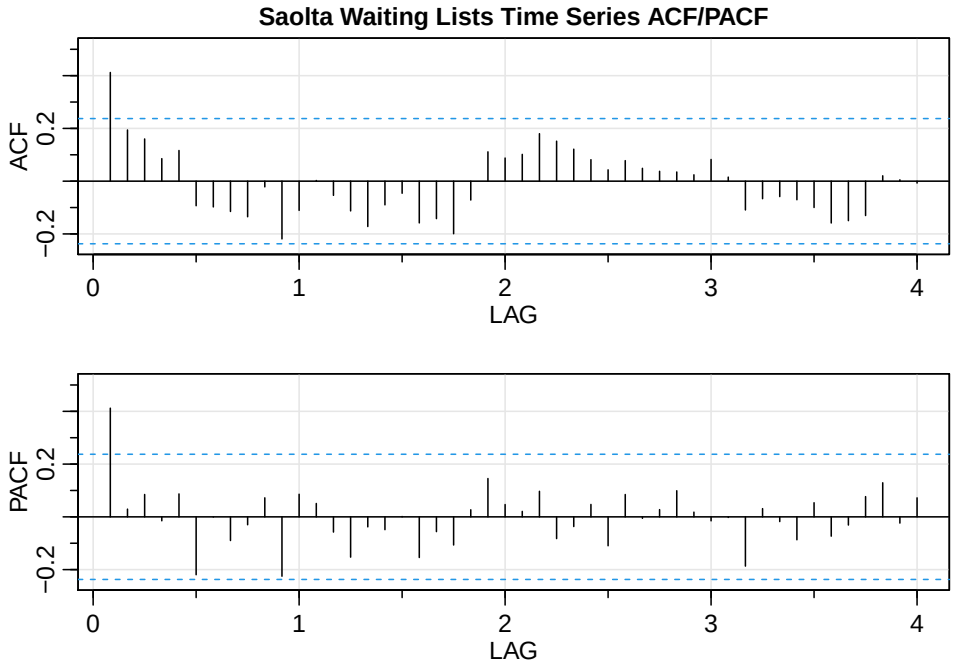

Figure 6: ACF/PACF plots of the time series after differencing.

Build Models

Using the parameters found above, three ARIMA models will be implemented and `forecast::auto.arima()` is then used to evaluate if the parameters provided are correct. The original time series must first be split into train and test sets. The training time series includes all cycles from 2014 to 2018, while the test set is made up of all 2019 data.

```
train <- window(x = saolta_ts, start = c(2014, 1), end = c(2018, 12))
test <- window(x = saolta_ts, start = c(2019, 1), end = c(2019, 12))
```

The training data is used to fit the ARIMA models:

```
arima_model1 = forecast::Arima(train, order = c(1, 1, 0))
arima_model2 = forecast::Arima(train, order = c(0, 1, 1))
arima_model3 = forecast::Arima(train, order = c(1, 1, 1))
auto_arima_model <- auto.arima(train)
```

The model calculated by `auto.arima()` is ARIMA(1,1,0) with drift. This drift is the amount of change over time and uses the average change seen in historical data. Drift was not considered during the manual model specification but makes sense with the nature of the time series under investigation. The ARIMA parameters (1,1,0) match that of the first manually specified one, providing additional confidence that the earlier process in determining these parameters was completed correctly. Comparison of these models is the focus of the next section.

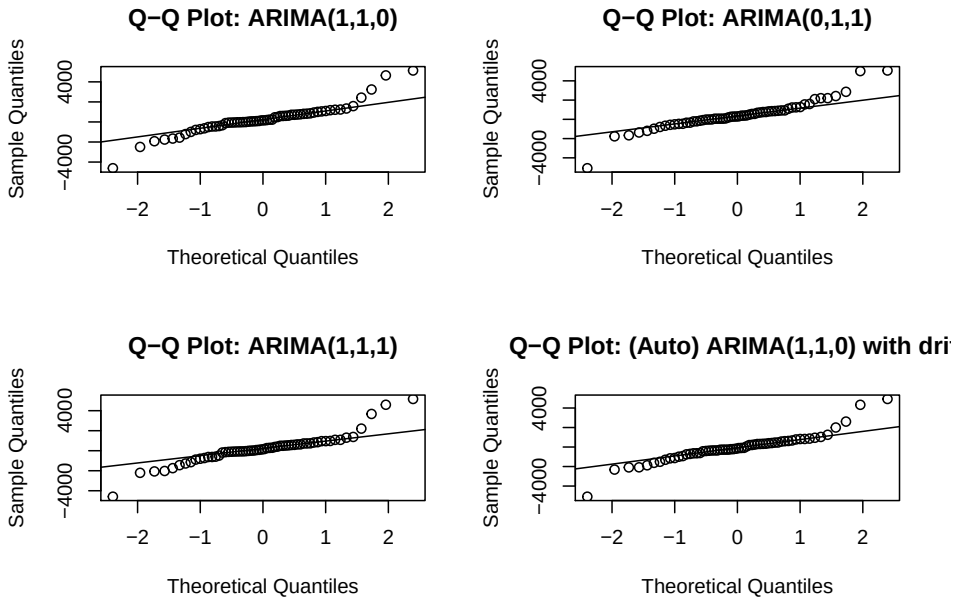

Figure 7: Quantile-Quantile plots for each ARIMA model.

Model Validation

The following outlines the methods used to evaluate the accuracy of the models proposed in the previous section and determine the best fit for forecasting waiting list totals.

The residuals of each ARIMA model are used to check the models for the presence of normal distributions. Quantile-Quantile plots for each model are featured in Figure 7 while in Figure 8 histograms are the other form of visualisation utilised. From visually inspecting these plots, all implemented ARIMA models appear to be normally distributed. The Ljung-box test which checks for randomness in ARIMA residuals was performed on each set of residuals with p-values recorded in Table 2. These are all greater than 0.05 meaning that the residuals are independent of each other which is the anticipated result.

Prediction Accuracy

Akaike information criterion (AIC) and the mean absolute percentage error (MAPE) for each of the four models are found in Table 2. The `auto.arima()` generated model (ARIMA(1,1,0) w/ drift) performs best based on both metrics i.e. it has the lowest score for each. This means this model is expected to have the most accurate predictions.

Predictions for 2019 using each model were calculated with the `forecast::forecast()` method. Plots for each of these forecasts are displayed in Figure 9. The ARIMA(1,1,0) w/ drift model's plot looks to be the most accurate due to the inclusion of drift,

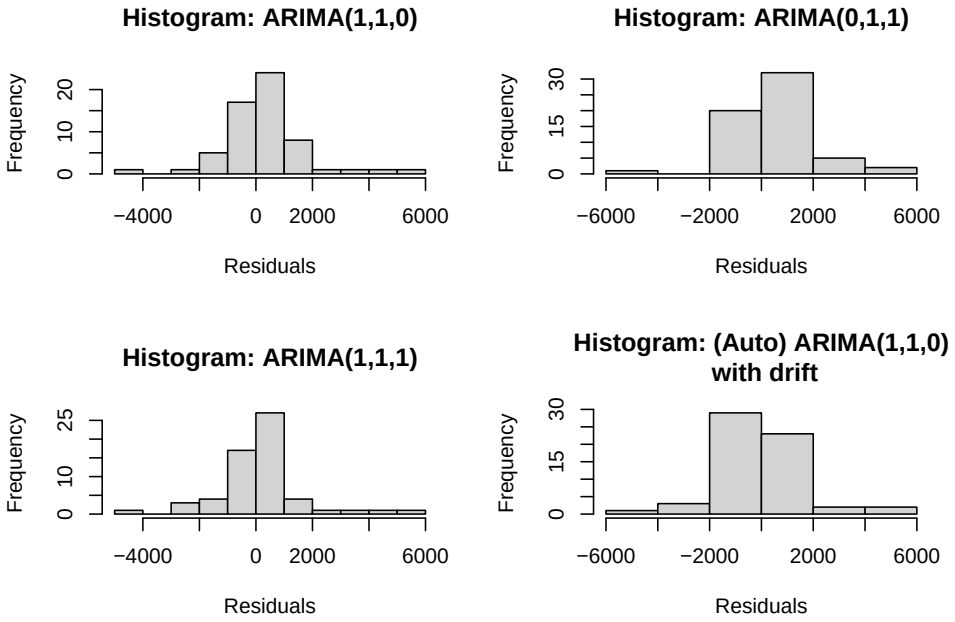

Figure 8: Histograms for each ARIMA model.

capturing the general upward growth found in the time series. Whereas each of the other model's mean forecasted values are almost horizontal. The mean forecasted results, as well as the actual figures for each month can be inspected in Table 3.

Correlation accuracy measures for the predictions are listed in Table 4. All but the model with drift applied were found to be negatively correlated which contradicts the actual waiting list figures for 2019. With these findings, the ARIMA(1,1,0) w/ drift model is selected for forecasting future values in the next section.

Table 2: Evaluation metrics for each implemented ARIMA model.

Model	AIC	MAPE	Ljung-box
ARIMA(1,1,0)	1032.688	1.093352	0.18205
ARIMA(0,1,1)	1036.507	1.135171	0.34483
ARIMA(1,1,1)	1034.202	1.071500	0.49092
(Auto) ARIMA(1,1,0) w/ drift	1031.205	1.070920	0.52420

Note: Values listed for Ljung-box are p-values.

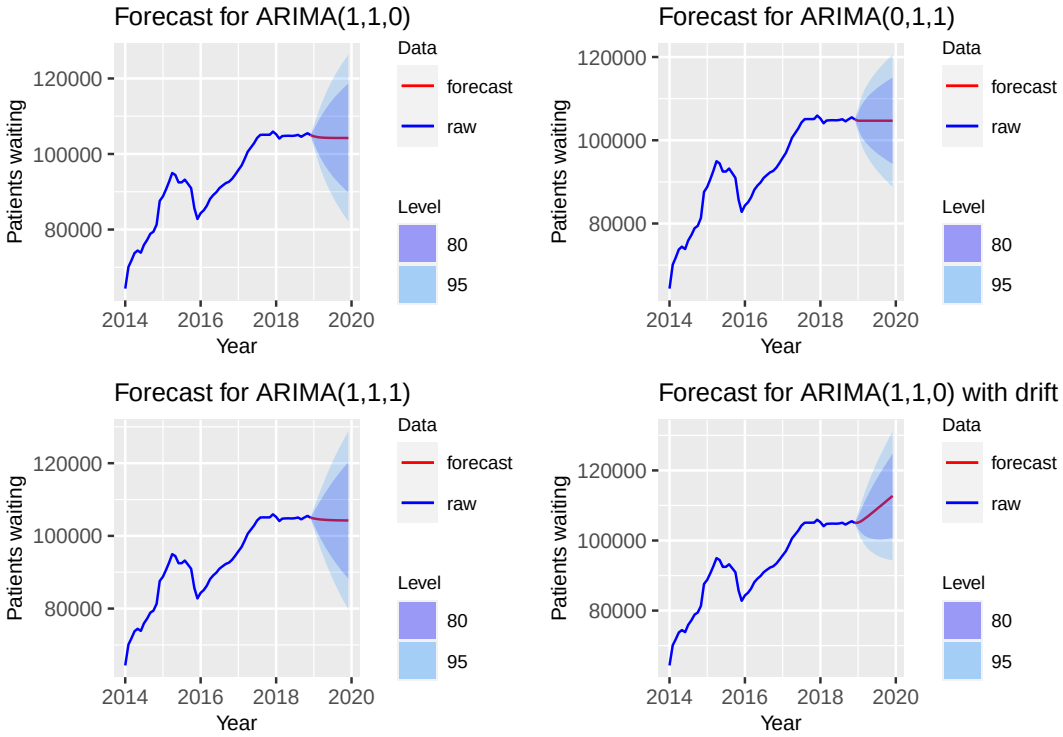

Figure 9: Forecast plots for each ARIMA configuration (2019).

Table 3: Actual vs Predicted Totals for 2019.

2019	Actual	ARIMA(1,1,0)	ARIMA(0,1,1)	ARIMA(0,1,1)	(Auto) ARIMA(1,1,0) w/ drift
1	107203	104678.4	104694.7	104791.9	105100.0
2	110075	104496.2	104694.7	104646.4	105531.8
3	110833	104388.8	104694.7	104537.8	106123.3
4	111803	104325.3	104694.7	104456.8	106795.0
5	112612	104287.9	104694.7	104396.5	107506.9
6	114542	104265.8	104694.7	104351.4	108238.9
7	115328	104252.8	104694.7	104317.8	108981.0
8	116622	104245.1	104694.7	104292.7	109728.1
9	116588	104240.6	104694.7	104274.0	110477.8
10	115910	104237.9	104694.7	104260.1	111228.7
11	115755	104236.3	104694.7	104249.7	111980.3
12	114253	104235.4	104694.7	104241.9	112732.2

Table 4: Actual vs Predicted correlation accuracy for 2019.

	Actual	ARIMA(1,1,0)	ARIMA(0,1,1)	ARIMA(0,1,1).1	(Auto) ARIMA(1,1,0) w/ drift
Actual	1.0000000	-0.9241134	NA	-0.9545271	0.8320844
ARIMA(1,1,0)	-0.9241134	1.0000000	NA	0.9736906	-0.7855254
ARIMA(0,1,1)	NA	NA	1	NA	NA
ARIMA(0,1,1).1	-0.9545271	0.9736906	NA	1.0000000	-0.8993836
(Auto) ARIMA(1,1,0) w/ drift	0.8320844	-0.7855254	NA	-0.8993836	1.0000000

Model Forecasting and Appraisal

The final stage in this project involves using the most suitable ARIMA model found from the previously performed validation to forecast future waiting list totals for the Saolta University Hospital Group. This forecasting will use the ARIMA(1,1,0) w/ drift model fitted with the entire original time series (2014-19) to predict monthly values for 2020.

The generated forecast is plotted in Figure 10 with the corresponding 2020 forecasted totals compared to 2019 actuals detailed in Table 5. The mean percentage difference when comparing forecasted to actual is 3.202441%, meaning the model is predicting the upward trend in the number of patients waiting to continue growing. The cycles of the 2019 time series are plotted those from the forecasts for 2020 in Figure 11, visually representing the predicted difference.

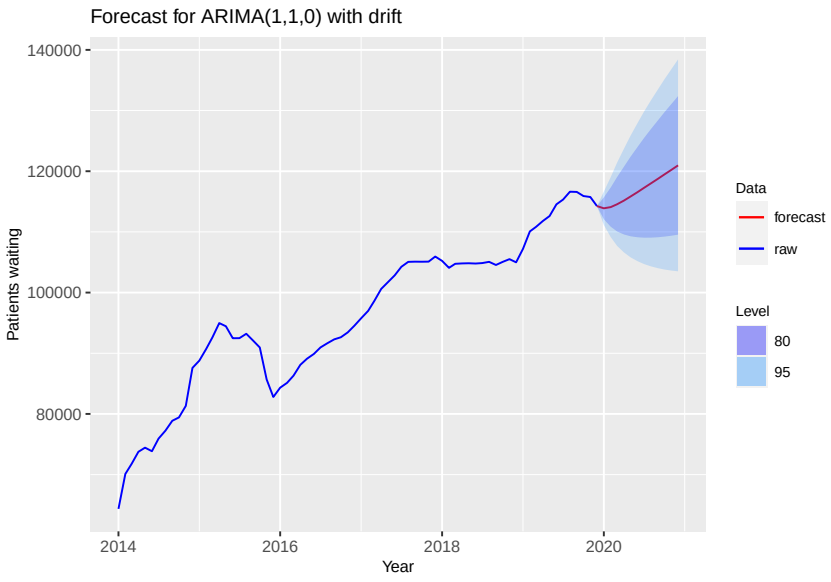

Figure 10: Forecast plot for 2020 patient numbers.

Table 5: Total Patients waiting monthly - 2019 vs 2020

2019 (Actual)	2020 (Forecasted)	Percentage Change
107203	113889.1	6.236863
110075	114085.8	3.643680
110833	114558.6	3.361422
111803	115167.3	3.009173
112612	115843.1	2.869243
114542	116551.9	1.754703
115328	117276.9	1.689860
116622	118009.9	1.190082
116588	118746.9	1.851695
115910	119485.8	3.084939
115755	120225.6	3.862130
114253	120965.9	5.875500

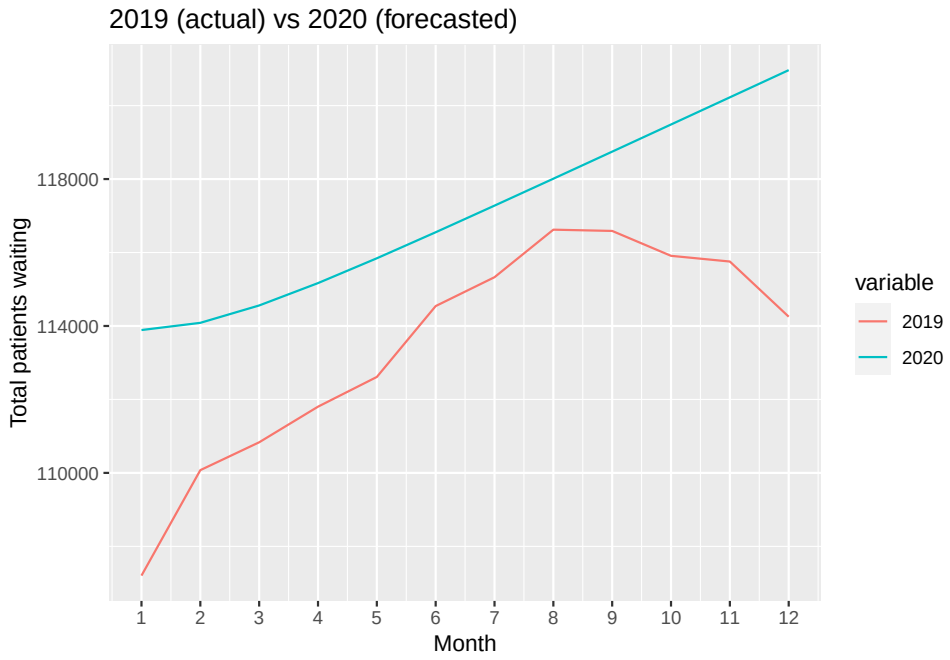

Figure 11: 2019 (actual) vs 2020 (forecasted).

Conclusions

The purpose of this project was to determine how accurately waiting list figures for the Saolta University Hospital Group can be forecasted using predictive modelling. ARIMA forecasting was determined to be the best fit for generating predictions from the time-based source data. Inspection of the time series created from the NTPF waiting list data revealed an upward trend but no strong seasonality in the data.

To find the parameters required for proposing ARIMA models, first non-seasonal differencing was applied to the time series to obtain `d` and `p` and `q` values were found by plotting the ACF and PACF applied to the differenced time series. Three models were proposed from these parameters; ARIMA(1,1,0), ARIMA(1,1,1) and ARIMA(0,1,1), while the R `auto.arima()` function generated a fourth model, ARIMA(1,1,0) w/ drift. Evaluation of the group of models revealed that the automatically generated model provided the lowest level of prediction error and so was determined as the best choice for forecasting future values. The correlation accuracy can be used to answer the stated research question, with the ARIMA(1,1,0) w/ drift model giving a value of 83% which represents a strong level of prediction accuracy.

Finally, the chosen ARIMA configuration was used to forecast monthly waiting list totals for 2020. The mean percentage increase from 2019 to 2020 was calculated as 3.2% which if accurate would imply that the upward trend present in the existing data is set to continue to grow.

References

Health Consumer Powerhouse (2018) *Euro Health Consumer Index 2018* [online], Health Consumer Powerhouse, available: https://healthpowerhouse.com/media/EHCI-2018/EHCI-2018-report.pdf [accessed 17 Mar 2020].

OECD, European Union (2018) *Health at a Glance: Europe 2018: State of Health in the EU Cycle* [online], Health at a glance: Europe, OECD, available: https://www.oecd-ilibrary.org/social-issues-migration-health/health-at-a-glance-europe-2018_health_glance_eur-2018-en [accessed 18 Mar 2020].