LETTERKENNY INSTITUTE OF TECHNOLOGY
ASSIGNMENT COVER SHEET

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

Assessment Title: **CA3 – Data Analysis**

Work to be submitted to: _____

Date for submission of work: **2020-05-10**

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 3 - Data Analysis

Health in Ireland

Title:	Trends in Irish hospital 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

Waiting lists for procedures in Irish public hospitals are some of the longest in Europe. The National Treatment Purchase Fund (NTPF) is the organisation responsible with collecting data about patients on these lists. This project uses the NTPF data to examine the question of what trends are present within it and determine what similarities or differences exist between those trends. To answer this research question the Mann-Kendall and Sen's slope statistical tests are applied to the data to verify if trends exist and the magnitude of those trends respectively.

The results returned from applying these tests show that, with the exception of the group containing all patents waiting under a year for a procedure, increasing trends are the norm. This result confirms that serious issues with the operation of the Irish health system need to be addressed and further research could build on what is presented here to investigate the problems in finer detail.

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.

A description of the data and steps taken to clean it are the feature of the next section. This data description section also includes justification for choices of statistical methods to aid in answering the research question displayed below. The sections that follow this cover the hypothesis testing to be carried on the waiting list data and reporting and discussion of results obtained from the analysis undertaken.

Research Question

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

> What differences, if any, exist among trends found in Irish hospital group waiting list figures from recent years?

Data Description

The data for this project comes from that collected by NTPF for outpatient (OP)[1], inpatient/day case (IPDC)[2] and GI endoscopy (IPDC GI)[3] waiting list numbers for all public hospitals across Ireland. The data that is being considered is monthly totals from January 2014 to December 2019. This section describes how the data was prepared for analysis and which statistical methods were chosen to assist in answering the research question.

[1]https://data.ehealthireland.ie/dataset/op-waiting-list-by-group-hospital
[2]https://data.ehealthireland.ie/dataset/ipdc-waiting-list-by-group-hospital
[3]https://data.ehealthireland.ie/dataset/ipdc-gi-endoscopy-by-group-hospital

Table 1: Waiting list counts.

Waiting list	Patient count
OP:	481643
IPDC:	294977
IPDC GI:	36978
Overall:	813598

Data Cleaning

R code referenced in this section is found in `data_prep/data_transformation.R`.

Amalgamating source data

The first step necessary is to combine all data into a single dataset. This is accomplished by initially combining all csv files for each waiting list category with the `combine_csv_data()` function in `helpers/data_prep_helper.R` which uses the `vroom` package. Then the 3 resulting tibbles are combined with `dplyr::bindrows`. The row counts for the individual and combined datasets can be seen in Table 1 while Table 2 shows a sample of rows from the overall data. The format of archive dates varies between Y-m-d and d/m/Y so the `convert_dates()` function that makes use of `libridate::parse_date_time` to convert all to a single datetime format was added to the `data_prep_helper.R` helper file.

Table 2: Random sample of rows from the combined data.

Archive_Date	Hospital_Group	Hospital_HIPE	Hospital_Name	Speciality_HIPE	Speciality_Name	Case_Type	Adult_Child	Age_Profile	Time_Bands	Total
2015-02-26	Children's Hospital Group	0941	Our Lady's Children's Hospital Crumlin	2600	General Surgery	NA	Child	0-15	15-18 Months	36
2014-05-30	University of Limerick Hospital Group	0305	Ennis Hospital	2600	General Surgery	NA	Adult	65+	6-9 Months	43
2017-05-31	Children's Hospital Group	0941	Our Lady's Children's Hospital Crumlin	1905	Paed Endocrinology	NA	Child	0-15	6-9 Months	11
2015-03-26	Ireland East Hospital Group	0908	Mater Misericordiae University Hospital	7000	Dental Surgery	Day Case	Adult	16-64	3-6 Months	2
2016-05-31	RCSI Hospitals Group	0400	Louth County Hospital	2600	General Surgery	NA	Adult	65+	3-6 Months	7
30/01/2020	Ireland East Hospital Group	0101	St. Columcille's Hospital	2100	Psychiatry	NA	Adult	65+	6-9 Months	2
2016-08-31	Saolta University Health Care Group	0500	Letterkenny General Hospital	0100	Cardiology	Day Case	Adult	65+	0-3 Months	22
2014-10-31	University of Limerick Hospital Group	0304	Nenagh Hospital	2604	Vascular Surgery	NA	Child	0-15	3-6 Months	1
2015-07-30	Dublin Midlands Hospital Group	0102	Naas General Hospital	5000	General Medicine	Day Case	Adult	16-64	6-9 Months	50
28/02/2018	Saolta University Health Care Group	803	Roscommon University Hospital	2600	General Surgery	Day Case	Adult	65+	0-3 Months	12
20/12/2018	South/South West Hospital Group	913	Mercy University Hospital	8003	Pain Relief	NA	Adult	16-64	3-6 Months	77
2016-09-29	Ireland East Hospital Group	0601	St. Luke's General Hospital Kilkenny	1503	Gynaecology	NA	Adult	16-64	12-15 Months	57
31/05/2018	Ireland East Hospital Group	202	Midland Regional Hospital Mullingar	1700	Ophthalmology	NA	Child	0-15	9-12 Months	1
2014-09-26	Ireland East Hospital Group	0908	Mater Misericordiae University Hospital	0300	Dermatology	NA	Adult	65+	6-9 Months	100
25/04/2019	Ireland East Hospital Group	0908	Mater Misericordiae University Hospital	0700	Gastro-Enterology	NA	Adult	65+	9-12 Months	12

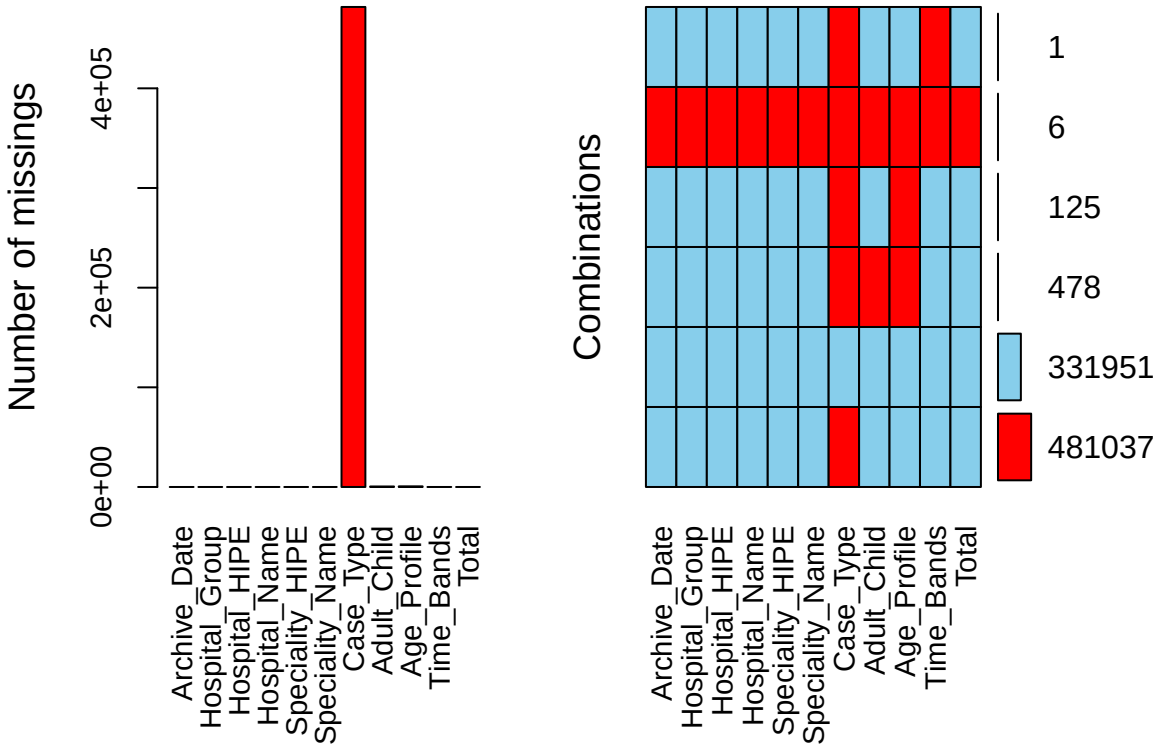

Figure 1: Missing values.

Dealing with missing values

Missing values from the combined data are displayed in Figure 1. The following steps were taken to handle these.

The missing case types are expected for all outpatient records because that column does not exist in the source csv so those are assigned a value of "Outpatient":

```
all_waiting_lists$Case_Type[is.na(all_waiting_lists$Case_Type)] <- "Outpatient"
```

There are 6 blank records introduced from the source datasets explaining the missing values for Archive_Date, Hospital_Group, Hospital_HIPE, Hospital_Name, Speciality_HIPE, Speciality_Name, Time_Bands and Total so these are dropped:

```
all_waiting_lists <- subset(all_waiting_lists,
                            !is.na(all_waiting_lists$Hospital_Name))
nrow(all_waiting_lists)
```

```
## [1] 813592
```

The 1 remaining row with a missing Time_Bands value is also removed because time band is vital for the analysis that follows:

```
nrow(all_waiting_lists)
```

```
## [1] 813592
```

```
all_waiting_lists <- all_waiting_lists[!is.na(all_waiting_lists$Time_Bands),]
nrow(all_waiting_lists)
```

```
## [1] 813591
```

All other missing values are deemed irrelevant for the current project, so no action is taken.

Time band formatting

The time bands variable originally had variations of the same values:

```
unique(all_waiting_lists[, c("Time_Bands")])
```

```
## # A tibble: 12 x 1
##    Time_Bands
##    <chr>
##  1 "0-3 Months"
##  2 "3-6 Months"
##  3 "6-9 Months"
##  4 "9-12 Months"
##  5 "15-18 Months"
##  6 "18+ Months"
##  7 "12-15 Months"
##  8 "18 Months +"
##  9 " 0-3 Months"
## 10 " 9-12 Months"
## 11 " 3-6 Months"
## 12 " 6-9 Months"
```

To rectify these differences whitespace was removed from all and the format for "18+ months" was standardised:

```
all_waiting_lists$Time_Bands <- as.character(all_waiting_lists$Time_Bands)
all_waiting_lists$Time_Bands <- trimws(all_waiting_lists$Time_Bands)
all_waiting_lists$Time_Bands[
  all_waiting_lists$Time_Bands == "18 Months +"] <- "18+ Months"
```

With the result looking like so:

```
unique(all_waiting_lists[, c("Time_Bands")])
```

```
## # A tibble: 7 x 1
##   Time_Bands
##   <chr>
```

```
## 1 0-3 Months
## 2 3-6 Months
## 3 6-9 Months
## 4 9-12 Months
## 5 15-18 Months
## 6 18+ Months
## 7 12-15 Months
```

To perform the analysis for answering the defined research question it was decided to reduce time band to 2 groupings, less than a year and greater than a year waiting:

```
unique(all_waiting_lists[, c("Time_Bands")])
```

```
## # A tibble: 2 x 1
##   Time_Bands
##   <fct>
## 1 < 1 Year
## 2 > 1 Year
```

Removal of unnecessary variables

Columns that aren't necessary for this project were removed from the dataset. The analysis will examine trends by using the 7 hospital groups and the 2 newly created waiting list time band so only the Archive_Date, Hospital_Group, Time_Bands and Total variables are kept.

Structure before removal:

```
str(all_waiting_lists, width = 70, strict.width = "cut")
```

```
## tibble [813,591 x 11] (S3: tbl_df/tbl/data.frame)
##  $ Archive_Date   : chr [1:813591] "31/10/2019" "31/10/2019" "31/10"..
##  $ Hospital_Group : chr [1:813591] "Children's Health Ireland" "Chi"..
##  $ Hospital_HIPE  : chr [1:813591] "0000" "0000" "0000" "0000" ...
##  $ Hospital_Name  : chr [1:813591] "Children's Health Ireland" "Chi"..
##  $ Speciality_HIPE: chr [1:813591] "0000" "0601" "0601" "0601" ...
##  $ Speciality_Name: chr [1:813591] "Small Volume Specialities" "Pae"..
##  $ Case_Type      : chr [1:813591] "Day Case" "Day Case" "Day Case""..
##  $ Adult_Child    : chr [1:813591] "Child" "Child" "Child" "Child" ...
##  $ Age_Profile    : chr [1:813591] "0-15" "0-15" "0-15" "0-15" ...
##  $ Time_Bands     : Factor w/ 2 levels "< 1 Year","> 1 Year": 1 1 1 ..
##  $ Total          : num [1:813591] 1 71 26 20 1 2 4 4 2 136 ...
```

After removal:

```
str(all_waiting_lists, width = 70, strict.width = "cut")
```

```
## tibble [813,591 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Archive_Date  : chr [1:813591] "31/10/2019" "31/10/2019" "31/10/"..
```

```
##  $ Hospital_Group: chr [1:813591] "Children's Health Ireland" "Chil"..
##  $ Time_Bands    : Factor w/ 2 levels "< 1 Year","> 1 Year": 1 1 1 1..
##  $ Total         : num [1:813591] 1 71 26 20 1 2 4 4 2 136 ...
```

Final clean up

The final steps in the data cleaning process involved abbreviating the hospital group names and excluding records from 2020 in the data as these only consist of January figures. The name abbreviation was done to accommodate displaying the names in charts that follow.

Note: The Children's Hospital Group was renamed to Children's Health Ireland in 2018 so these are classed as a single group.

Original hospital group names:

```
unique(all_waiting_lists[, c("Hospital_Group")])
```

```
## # A tibble: 8 x 1
##   Hospital_Group
##   <chr>
## 1 Children's Health Ireland
## 2 Dublin Midlands Hospital Group
## 3 Ireland East Hospital Group
## 4 RCSI  Hospitals Group
## 5 Saolta University Health Care Group
## 6 South/South West Hospital Group
## 7 University of Limerick Hospital Group
## 8 Children's Hospital Group
```

Abbreviated hospital group names:

```
unique(all_waiting_lists[, c("Hospital_Group")])
```

```
## [1] "CHI"              "Dublin Midlands"  "Ireland East"     "RCSI"
## [5] "Saolta"           "South/South West" "UL"
```

A csv file named `combined_waiting_lists.csv` containing the processed data is found in the `data_prep` directory.

Choice of Statistical Methods

Figure 2 shows all waiting lists plotted as a bar chart to give a general view of how the data has changed over time. The trend here appears to be generally ever growing numbers of patients waiting more than a year for procedures while those waiting less than a year saw a steady increase from 2014 to 2017 and a stabling thereafter.

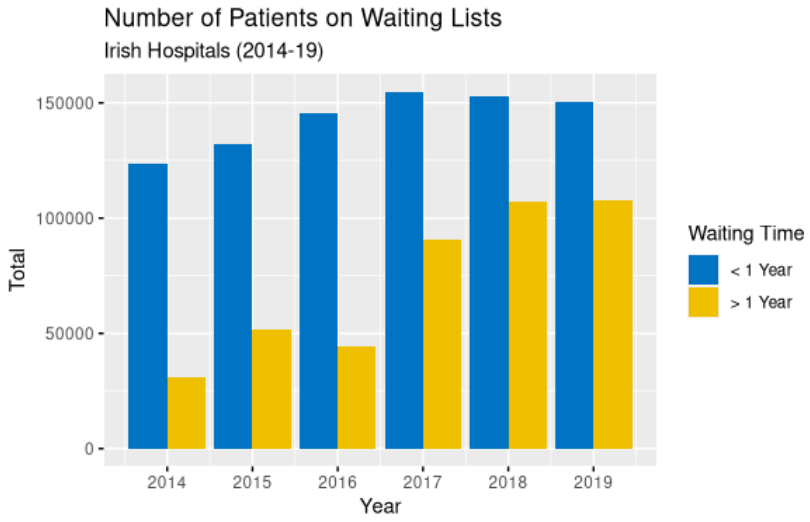

Figure 2: Totals waiting list numbers by year.

To help decide what type of statistical methods are suitable for use on this data, density plots for each hospital group were generated, see Figure 3. All groups have skewed distributions meaning some type of non-parametric methods must be used.

Based on the data being time series in nature and not normally distributed the chosen statistical tests to answer the research question are the Mann-Kendall test to determine if monotonic trends are present and the Sen's slope test to check the magnitude of trends, if they exist.

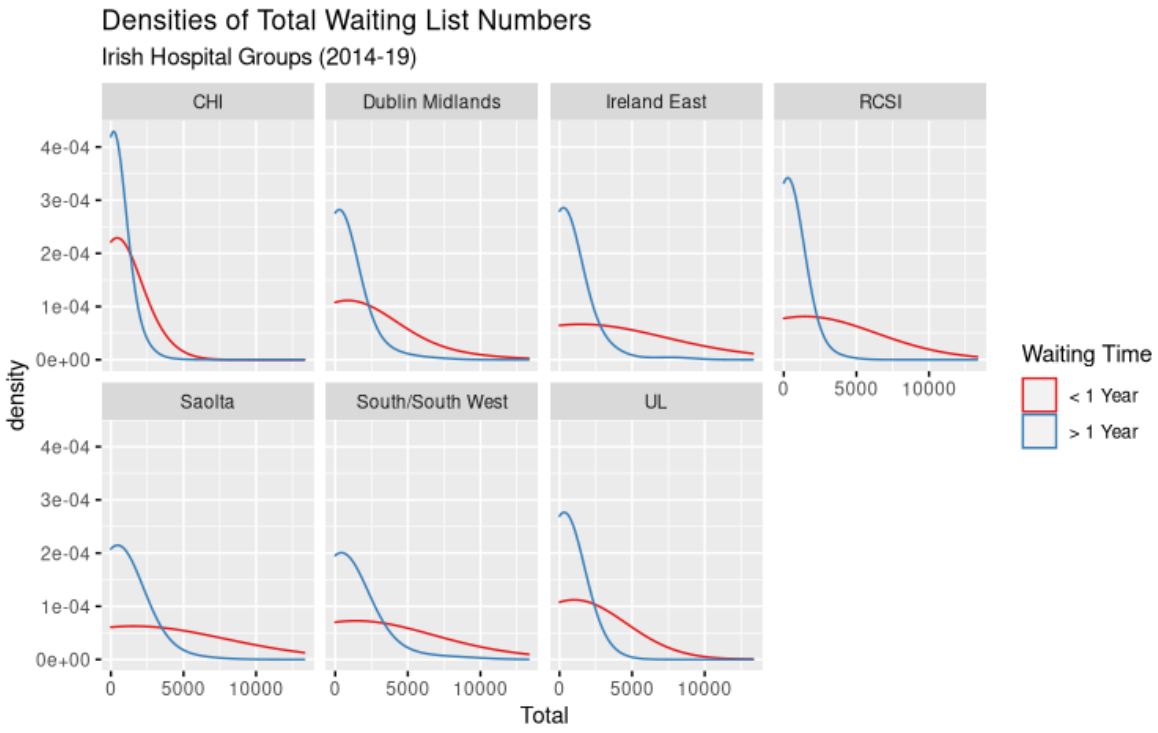

Figure 3: Density plot for each Irish hospital group.

Hypothesis Testing

The hypothesis being tested on the waiting list time series data with the Mann-Kendall test is as follows:

H_0: No monotonic trend exists

H_1: Monotonic trend exists

Code relating to the testing of this hypothesis is found in the `time_series_analysis` function in `helpers/descriptive_statistics_helper.R`.

The Mann-Kendall test is performed with `Kendall::MannKendall` which returns a score and a p-value, 2-sided in this case. For this test, a confidence interval of 95% is used meaning a p-value of less than 0.05 is necessary to reject the null hypothesis

The Sen's slope test is performed to determine the magnitude of the monotonic trend, if present. The `trend::sens.slope` function provides this test.

The overall data is split into individual datasets for each hospital group in order to facilitate testing. The next section discusses the results obtained by using this combination of tests.

Results

Results from performing the Mann-Kendall and Sen's Slope tests on the overall data and each of the individual hospital groups can be seen in Table 3. Code relevant to this section is found in `data_analysis/analyse_prepared_waiting_list_data.`. Discussion about the results is featured below and plots of the decomposition of each time series group can be found in the Appendix.

Table 3: Time series analysis results

		Mann-Kendall			**Sen's Slope**	
		Score	**p-value**	**Result**	**Slope**	**p-value**
All	< 1 Yr (Fig. 4)	232	0.26146	No Trend	58.38	0.2615
	> 1 Yr (Fig. 5)	1084	< 0.0001	Increasing	77.89	< 0.0001
CHI	< 1 Yr (Fig. 6)	1508	< 0.0001	Increasing	126.61	< 0.0001
	> 1 Yr (Fig. 7)	2238	< 0.0001	Increasing	312.91	< 0.0001
Dub. Mid.	< 1 Yr (Fig. 8)	1966	< 0.0001	Increasing	158.69	< 0.0001
	> 1 Yr (Fig. 9)	2098	< 0.0001	Increasing	415.79	< 0.0001
Ire. East	< 1 Yr (Fig. 10)	2180	< 0.0001	Increasing	158.69	< 0.0001
	> 1 Yr (Fig. 11)	2314	< 0.0001	Increasing	515.56	< 0.0001
RCSI	< 1 Yr (Fig. 12)	884	< 0.0001	Increasing	65.95	< 0.0001
	> 1 Yr (Fig. 13)	648	0.0016596	Increasing	103.62	0.00166
Saolta	< 1 Yr (Fig. 14)	2042	< 0.0001	Increasing	235.46	< 0.0001
	> 1 Yr (Fig. 15)	2078	< 0.0001	Increasing	382.67	< 0.0001
UL	< 1 Yr (Fig. 16)	2152	< 0.0001	Increasing	171.69	< 0.0001
	> 1 Yr (Fig. 17)	2404	< 0.0001	Increasing	284.23	< 0.0001
South/SW	< 1 Yr (Fig. 18)	931	< 0.0001	Increasing	134.20	< 0.0001
	> 1 Yr (Fig. 19)	2110	< 0.0001	Increasing	521.42	< 0.0001

Note: p-values have been rounded for readability when very small values were found. For example, the actual p-value for MK test on all waiting list < 1 year is 1.1921e-07, displayed here as < 0.0001. Sen's slope values have also been rounded (to 2 decimal places).

In general, increasing monotonic trends were the norm found when applying the stated statistical techniques to the groups of waiting lists. The result for the group containing all patients waiting less than a year is the only one that does not conform to this generality. Its Mann-Kendall score (232) indicates it is a mainly increasing time series, although from the corresponding graph (Figure 4) it is clear there is no consistent trend present. With the exception of the greater than a year waiting group for RCSI, all other p-values for the Mann-Kendall test are extremely low. Numbers for that RCSI group have seen a steady decline since 2018 (Figure 13) thus explaining the reduced confidence level of the test.

Using the Sen's slope test gives an indication of how great the magnitude of a given trend is which provides a useful means for comparing similar trends. The patients waiting more than a year for procedures in South/South West Hospital group (Figure 19) returned the highest slope value. This signifies that the increasing trend observed has the largest average growth rate of the groups studied.

Conclusions

The stated aim of this project was to examine trends in waiting list data for Irish public hospitals and determine if any significant differences are observable. To enable this work, data collected by the NTPF for each category of waiting was gathered and processed. The processing involved combining csv files containing monthly waiting list figures going back as far as January 2014. The source data contains records for each list speciality, hospital and trimonthly time band slots. Categorisation was performed on the overall dataset, focused on dividing it into hospital groupings and waiting time bands of less than or greater than a year to facilitate the analysis section of this research.

To answer the project's research question, the Mann-Kendall and Sen's slope statistical trend tests were applied to each of the groups created from the waiting list data. The general result found from these tests was that all of the observed waiting list groups feature an increasing monotonic trend, except the combination of all patients waiting for less than a year. The group containing patient numbers waiting for longer than a year with South/South West Hospital group hospitals returned the highest magnitude from the use of the Sen's slope test. These tests confirmed the assumption made when the total figures were examined initially that increasing trends are the norm (see Choice of Statistical Methods)

Further work on this research that was neglected from the scope of this project due to time constraints is to examine how trends differ between individual procedure waiting lists. Knowing where the main problem areas exist with evidence confirmed with statistical methods could aid the Department of Health in distributing resources to areas most in need.

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].

Appendix

Decomposed Time Series Plots

The following plots show the seasonal and trend decomposition for each of the waiting list groups using Loess (locally estimated scatterplot smoothing). This was achieved by calling the `stats::stl` function on each time series when plotting.

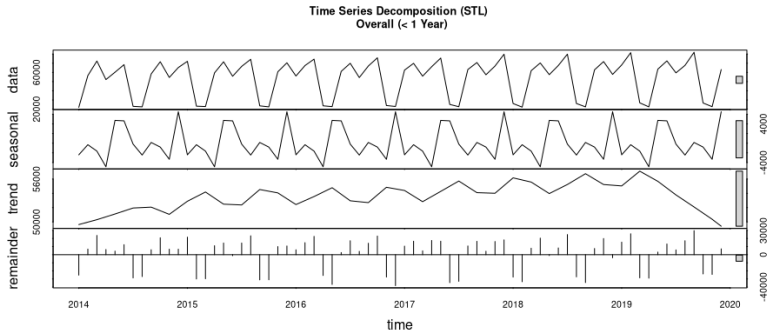

Figure 4: Time Series Decomposition - Overall patient numbers waiting less than a year.

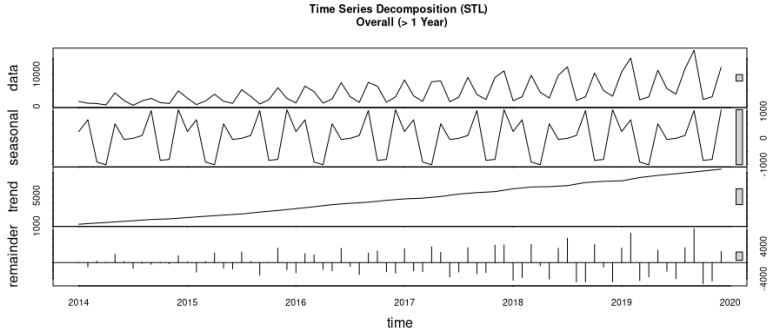

Figure 5: Time Series Decomposition - Overall patient numbers waiting more than a year.

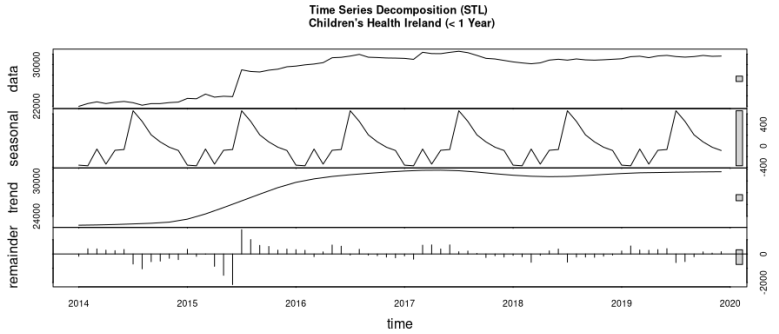

Figure 6: Time Series Decomposition - Children's Health Ireland patient numbers waiting less than a year.

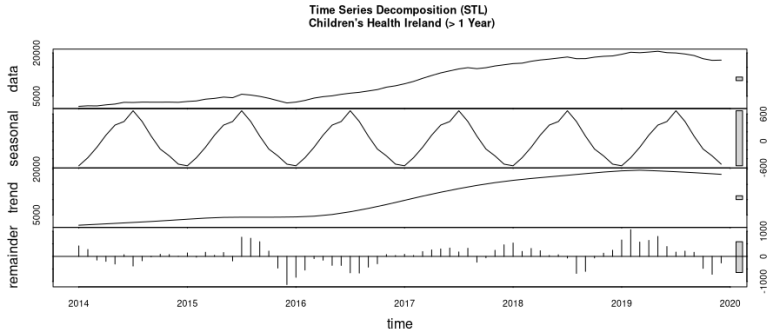

Figure 7: Time Series Decomposition - Children's Health Ireland patient numbers waiting more than a year.

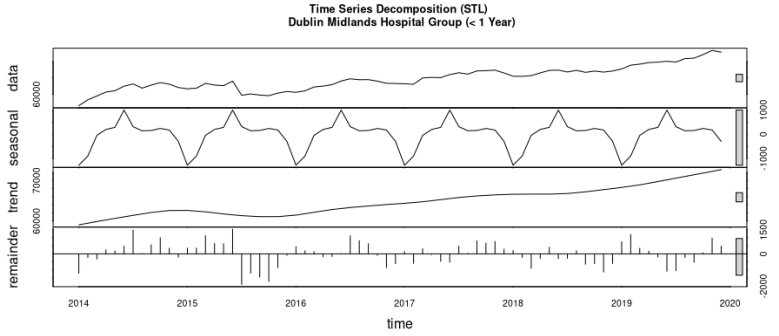

Figure 8: Time Series Decomposition - Dublin Midland Hospital Group patient numbers waiting less than a year.

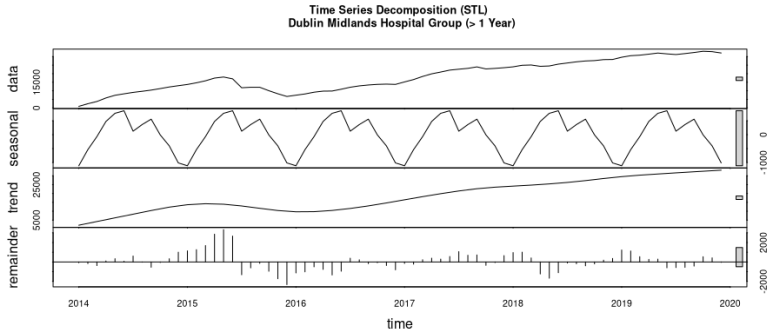

Figure 9: Time Series Decomposition - Dublin Midland Hospital Group patient numbers waiting more than a year.

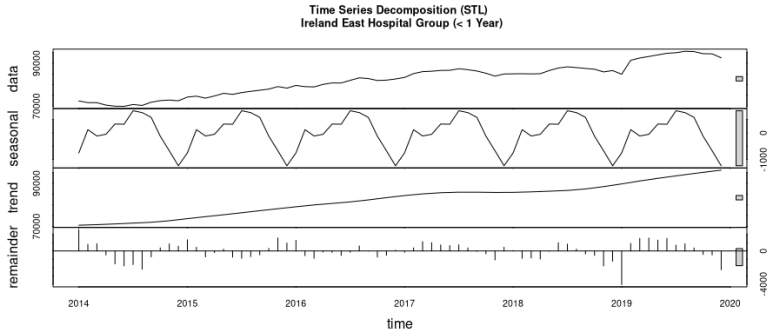

Figure 10: Time Series Decomposition - Ireland East Hospital Group patient numbers waiting less than a year.

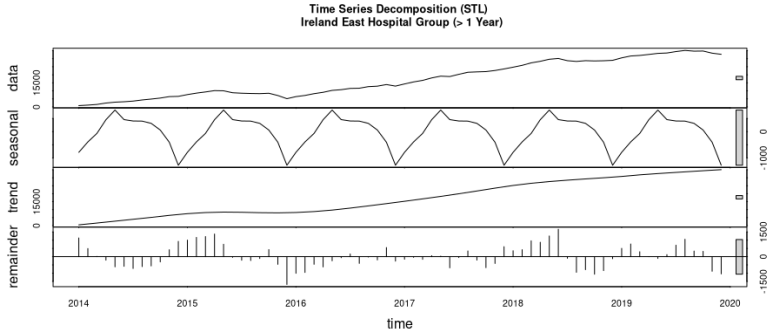

Figure 11: Time Series Decomposition - Ireland East Hospital Group patient numbers waiting more than a year.

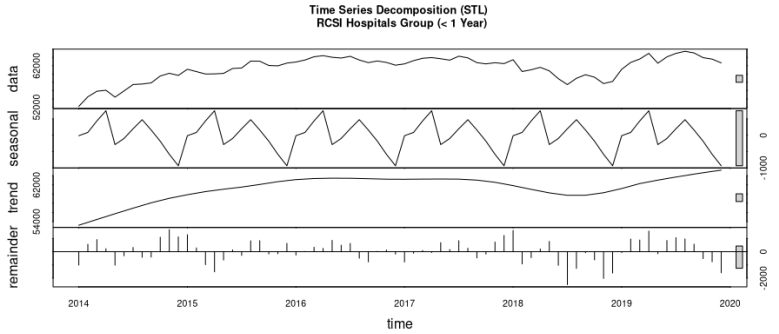

Figure 12: Time Series Decomposition - RCSI Hospitals Group patient numbers waiting less than a year.

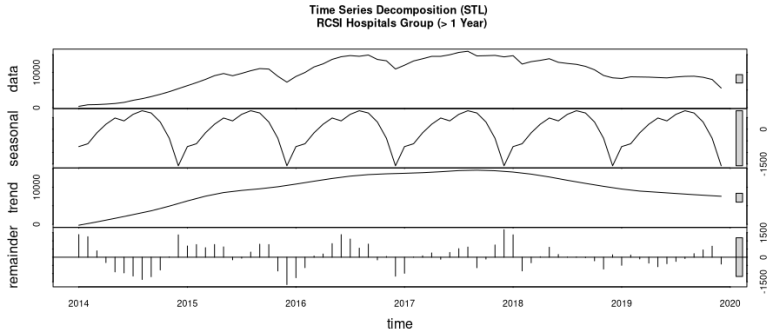

Figure 13: Time Series Decomposition - RCSI Hospitals Group patient numbers waiting more than a year.

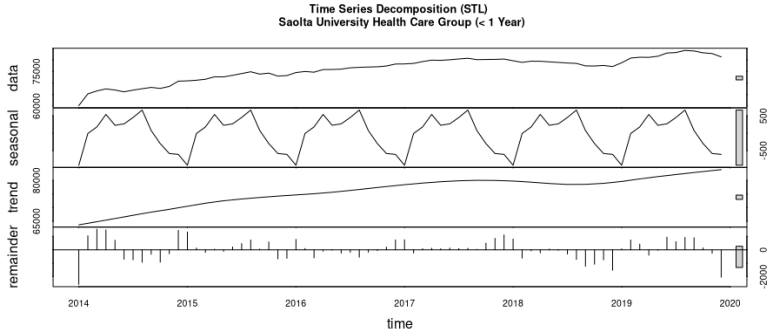

Figure 14: Time Series Decomposition - Saolta University Health Care Group patient numbers waiting less than a year.

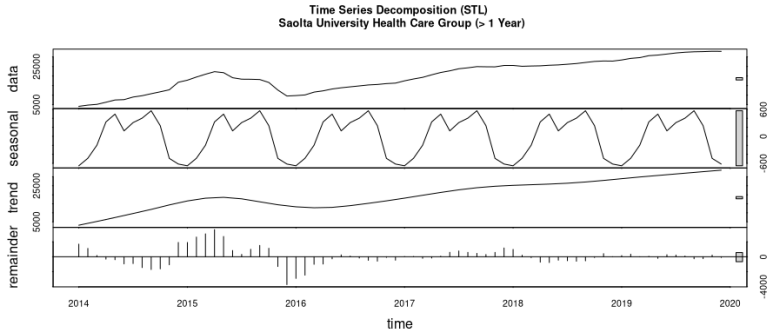

Figure 15: Time Series Decomposition - Saolta University Health Care Group patient numbers waiting more than a year.

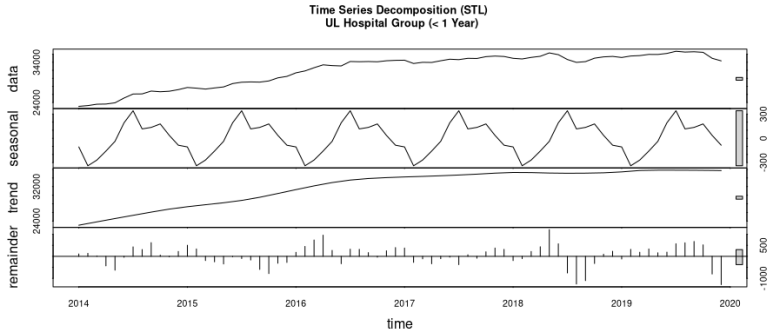

Figure 16: Time Series Decomposition - University of Limerick Hospital Group patient numbers waiting less than a year.

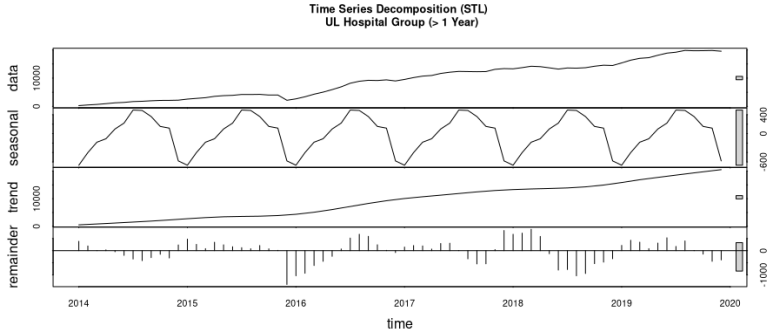

Figure 17: Time Series Decomposition - University of Limerick Hospital Group patient numbers waiting more than a year.

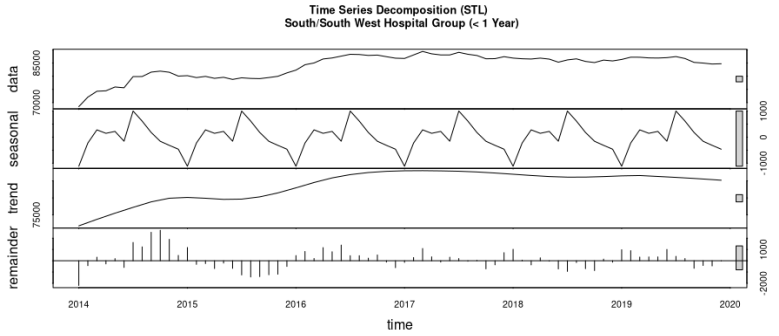

Figure 18: Time Series Decomposition - South/South West Hospital Group patient numbers waiting less than a year.

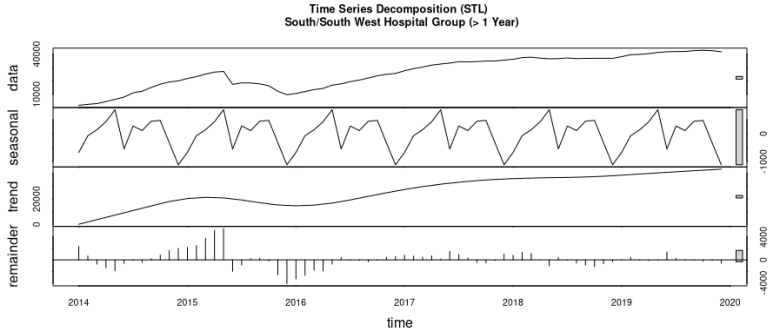

Figure 19: Time Series Decomposition - South/South West Hospital Group patient numbers waiting more than a year.