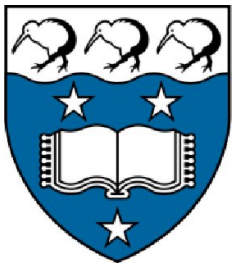
Data Storytelling Dashboard for Exploring Auckland Air Quality



ZHAOMING SU

DEPARTMENT OF STATISTICS

THE UNIVERSITY OF AUCKLAND

supervised by

DR. EARO WANG

PROF. CHRIS WILD

A dissertation submitted in partial fulfilment of the requirements for the degree of Bachelor of Science (Honours) in Statistics, The University of Auckland, 2021.

Contents

Abstract v

Acknowledgements vii

Copyright notice ix

Preface xiii

1 Introduction (Template Demo) 1
- 1.1 Rmarkdown . . . 1
- 1.2 Data . . . 2
- 1.3 Figures . . . 2
- 1.4 Results from analyses . . . 3
- 1.5 Tables . . . 3

2 Background and related works 5
- 2.1 Time series graphics toolbox . . . 5
- 2.2 HTML widgets for interactive graphics . . . 5
- 2.3 Environmental visual analysis . . . 5

3 Auckland air quality data 7
- 3.1 Introduction . . . 7
- 3.2 Data quality and cleaning . . . 8
- 3.3 Data enrichment . . . 10

4 Design layout and philosophy 11

5 Linked interactive graphics 13

6 Modelling 15

7 Conclusion and future works 17

A Additional stuff 19

Bibliography 21

Abstract

The abstract should outline the main approach and findings of the thesis and must not be more than 500 words.

Acknowledgements

I would like to thank my pet goldfish for ...

Copyright notice

(Standard thesis)

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

(Thesis including published works declaration)

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes (insert number) original papers published in peer reviewed journals and (insert number) submitted publications. The core theme of the thesis is (insert theme). The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the (insert name of academic unit) under the supervision of (insert name of supervisor).

(The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.) Remove this paragraph for theses with sole-authored work

In the case of (insert chapter numbers) my contribution to the work involved the following:

I have / have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student name: ZHAOMING SU

Student signature:

Date:

Thesis chapter	**Publication title**	**Status (published, in press, accepted or returned for revision)**	**Nature and % of student contribution**	**Co-author name(s), nature and % of co-author's contribution**	**Co-author(s), Monash student Y/N**
2	xx	xx	xx	xx	N
3	xx	xx	xx	xx	N
4	xx	xx	xx	xx	N
5	xx	xx	xx	xx	N

Preface

The material in Chapter 1 has been submitted to the journal *Journal of Impossible Results* for possible publication.

The contribution in Chapter 2 of this thesis was presented in the International Symposium on Nonsense held in Dublin, Ireland, in July 2015.

Chapter 1

Introduction (Template Demo)

This is where you introduce the main ideas of your thesis, and an overview of the context and background.

In a PhD, Chapter 2 would normally contain a literature review. Typically, Chapters 3–5 would contain your own contributions. Think of each of these as potential papers to be submitted to journals. Finally, Chapter 6 provides some concluding remarks, discussion, ideas for future research, and so on. Appendixes can contain additional material that don't fit into any chapters, but that you want to put on record. For example, additional tables, output, etc.

1.1 Rmarkdown

In this template, the rest of the chapter shows how to use Rmarkdown. The big advantage of using Rmarkdown is that it allows you to include your R code directly into your thesis, to ensure there are no errors in copying and pasting, and that everything is reproducible. It also helps you stay better organized.

For details on using *R Markdown* see `http://rmarkdown.rstudio.com`.

1.2 Data

Included in this template is a file called `sales.csv`. This contains quarterly data on Sales and Advertising budget for a small company over the period 1981–2005. It also contains the GDP (gross domestic product) over the same period. All series have been adjusted for inflation. We can load in this data set using the following command:

```
sales <- ts(read.csv("data/sales.csv")[, -1], start = 1981, frequency = 4)
```

Any data you use in your thesis can go into the data directory. The data should be in exactly the format you obtained it. Do no editing or manipulation of the data outside of R. Any data munging should be scripted in R and form part of your thesis files (possibly hidden in the output).

1.3 Figures

Figure 1.1 shows time plots of the data we just loaded. Notice how figure captions and references work. Chunk names can be used as figure labels with `fig:` prefixed. Never manually type figure numbers, as they can change when you add or delete figures. This way, the figure numbering is always correct.

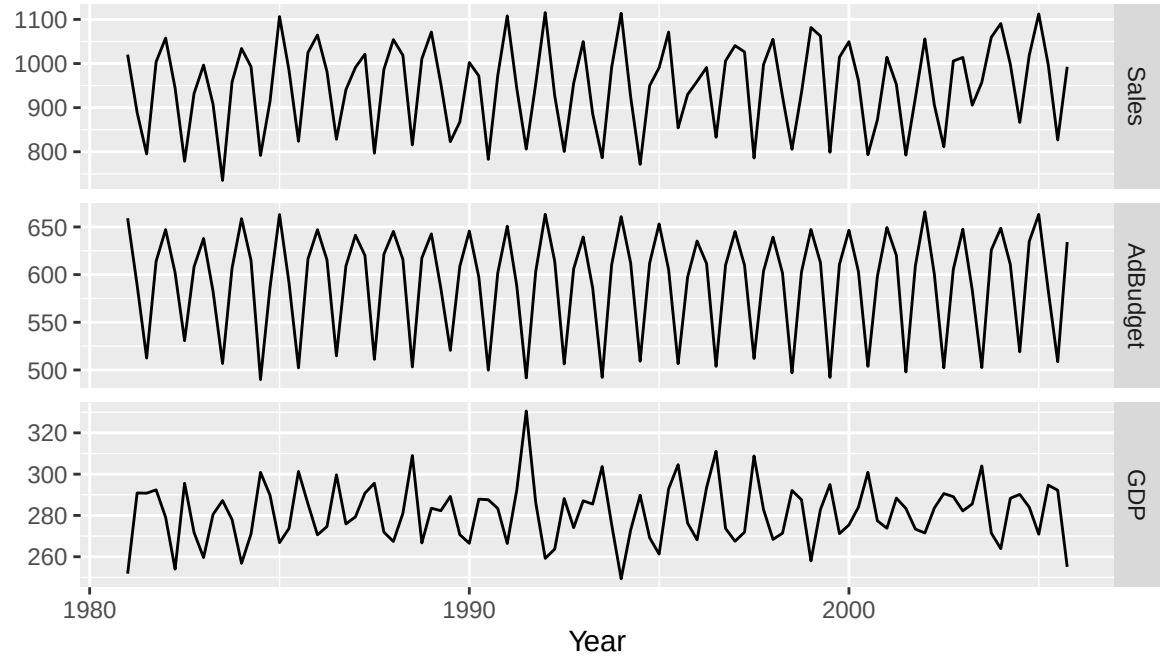


Figure 1.1: *Quarterly sales, advertising and GDP data.*

1.4 Results from analyses

We can fit a dynamic regression model to the sales data.

If y_t denotes the sales in quarter t, x_t denotes the corresponding advertising budget and z_t denotes the GDP, then the resulting model is:

$$y_t - y_{t-4} = \beta(x_t - x_{t-4}) + \gamma(z_t - z_{t-4}) + \theta_1\varepsilon_{t-1} + \Theta_1\varepsilon_{t-4} + \varepsilon_t \tag{1.1}$$

where $\beta = 2.28$, $\gamma = 0.97$, $\theta_1 = NA$, and $\Theta_1 = -0.90$.

1.5 Tables

Let's assume future advertising spend and GDP are at the current levels. Then forecasts for the next year are given in Table 1.1.

Again, notice the use of labels and references to automatically generate Table numbers. In this case, we need to generate the label ourselves.

Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
1000.2	947.7	1052.7	919.9	1080.5
1013.1	959.3	1066.8	930.9	1095.3
1076.7	1022.9	1130.6	994.4	1159.0
1003.5	949.7	1057.4	921.2	1085.8

Table 1.1: *Forecasts for the next year assuming Advertising budget and GDP are unchanged.*

The `knitLatex` package is useful for generating tables from R output. Other packages can do similar things including the `kable` function in `knitr` which is somewhat simpler but you have less control over the result. If you use `knitLatex` to generate tables, don't forget to include `results="asis"` in the chunk settings.

Chapter 2

Background and related works

2.1 Time series graphics toolbox

- Calendar plot
- Seasonal plot
- Line plot and ACF (FPP)

2.2 HTML widgets for interactive graphics

- shiny, plotly, echarts4r, leaflet and reactable

2.3 Environmental visual analysis

- Wind rose and pollution rose
- Trend estimate

Chapter 3

Auckland air quality data

3.1 Introduction

Air quality index (AQI) is a critical indicator of overall air quality by measuring key air pollutant concentrations at a given time. The constitution of AQI consists of ambient air pollutants listed in the National Environmental Standards for Ambient Air Quality which defines the threshold target for calculating AQI (Auckland Regional Council, 2020a). The national standard defines AQI as the maximum ambient air pollutant measurement ratio to the national target as a percentage (Auckland Regional Council, 2020b). Over 10 stations in Auckland monitor a subset of the standard-listed pollutants in an hourly interval.

It is of interest to explore the variation and relationships of AQI and its constituent pollutants with other environmental and meteorological parameters over time. The data, provided by Auckland Regional Council (2021), includes 14 parameters from 10 monitoring stations in Auckland from as North as Takapuna to as South as Patumahoe. Available parameters consist of air quality index (AQI), 10 pollutant levels and four other meteorological variables as per Table 3.1, with various starting dates (since as early as 2003 in Takapuna) until April 2021.

Only six of the standard-listed air pollutants are monitored and available in the data, and each station independently monitors a subset of the six pollutants. As such, the calculation of AQI, based on available data, may be simplified to

$$AQI = 100 \times \max\{\frac{PM_{2.5}}{25}, \frac{PM_{10}}{50}, \frac{NO_2}{200}, \frac{SO_2}{350}, \frac{CO}{10}, \frac{O_3}{150}\} \tag{3.1}$$

Parameter	Unit	Note
AQI		Air quality index
BC(370)	ngm^{-3}	Black carbon at 370nm wavelength
BC(880)	ngm^{-3}	Black carbon at 880nm wavelength
CO*	mgm^{-3}	Carbon monoxide concentration
NO	μgm^{-3}	Nitrogen monoxide concentration
NO_2*	μgm^{-3}	Nitrogen dioxide concentration
NOx	μgm^{-3}	Nitrogen oxides concentration
O_3*	μgm^{-3}	Ozone concentration
PM2.5*	μgm^{-3}	Particulate matter with diameter <2.5μm
PM10*	μgm^{-3}	Particulate matter with diameter <10μm
SO_2*	μgm^{-3}	Sulphur dioxide concentration
Relative Humidity	%	
Temperature	°C	
Wind Speed	ms^{-1}	
Wind Direction	°	

Table 3.1: *Parameters available in raw data.*
* *AQI-related ambient air pollutants*

It is noteworthy that the availability of air quality parameters in each monitoring station varies from year to year. Besides, the extreme values addressed in Section 3.2.1 are more frequent in earlier years. The final data set is subsetted from the year 2016.

3.2 Data quality and cleaning

The raw data consists of two separate data sets, each with a different data structure. Cleaning and manipulation are needed to ensure that the two data sets are consistent in structure and free from error. The raw data sets are individually inspected and cleaned before combination. This section outlines the issues found and methods to address them.

3.2.1 Abnormal and missing values

Abnormal or missing values arise from instrumental or input errors. Upon inspection, 104,332 records were found to have a negative value. Nevertheless, all pollutants are reported in units in the form of mass per unit volume, and other parameters, except for

temperature, are only sensible if positive as of Table 3.1. Therefore, 104,257 records of insensible negative values are removed. Besides, conspicuously anomalous records of AQI are found in data, including consecutive hours of >1,000 AQI in Takapuna and numerous AQI values being inconsistent with Formula 3.1 based on available pollutants in the same data set. The anomalous records are nonetheless kept as-is for further verification.

In addition, preliminary inspection finds that 0.81% of records are explicitly missing. Yet after filling the implicit time gaps in the data, 53.71% of records are implied to be missing.

3.2.2 Date and time

A consistent format in date and time is crucial to the accuracy of temporal data. Observations with inconsistent time format are present in the data, where some are recorded in `hh:mm:ss` whilst others in `hh:mm`. The inconsistency in the time format is correctable due to the hourly nature of the data. 0.06% of records with missing time are removed.

The time zone of New Zealand changes by +1 during daylight saving. To avoid duplicated index upon boundaries of daylight saving upon data visualisation, all time-stamps are presented in NZST (UTC+12). On the other hand, the date and time in the cleaned data file are stored as a single variable, with its format in compliance with ISO 8601 (International Organization for Standardization, 2019; Wickham, Hester, and Francois, 2018).

3.2.3 Duplicate records

Temporal data should not present duplicate records. Of the 7,292,038 valid records, 239,374 (3.28%) are duplicate with 120,207 redundant records. Further checking reveals that 230,822 of the duplicates have inconsistent values. However, as the scale of the inconsistency of most duplicate records is reasonably small, the first-appearing records of each duplicate are kept.

3.2.4 Structural difference in raw data sets

The primary data set, which records all parameters except for wind direction, is in long format, with each observation consisting of a single record of one parameter for one station at a given hour. Nevertheless, each observation of the wind direction data set consists

of wind direction records of all stations at a given hour. Each data set is pivoted to the structure such that each observation is uniquely identified by the date-time and station with records of all parameters before combination to ensure structural consistency.

3.3 Data enrichment

- Categorisation of AQI
- Categorisation of wind direction

Chapter 4

Design layout and philosophy

- Overview of AQI
  - Spatial
  - Temporal (AQI and its constituent)
    - Calendar
    - Drill-down line plot
- Data enrichment
  - Explore relationship AQI with wind speed and direction
  - Meteorological data
- Trend analysis

Chapter 5

Linked interactive graphics

- Introduction
- Implementation of interactive linking
- Modularisation of Shiny App

Chapter 6

Modelling

Chapter 7

Conclusion and future works

Appendix A

Additional stuff

You might put some computer output here, or maybe additional tables.

Note that line 5 must appear before your first appendix. But other appendices can just start like any other chapter.

Bibliography

Auckland Regional Council (2020a). *Auckland Ambient Air Quality Guidelines*. Accessed: 2021-09-22. Auckland, New Zealand. https://drive.google.com/drive/folders/1D-WsQ3ISmYWjJlSI29aA7rlgYPrzOv6F.

Auckland Regional Council (2020b). *Auckland Ambient Air Quality Targets*. Accessed: 2021-09-22. Auckland, New Zealand. https://unitaryplan.aucklandcouncil.govt.nz/Images/Auckland%20Unitary%20Plan%20Operative/Chapter%20E%20Auckland-wide/1.%20Natural%20Resources/E14%20Air%20quality.pdf.

Auckland Regional Council (2021). *The Auckland Environmental Data*. Accessed: 2021-05-25. Auckland, New Zealand. https://environmentauckland.org.nz/Data/DataSet/.

International Organization for Standardization (2019). *ISO 8601: Date and Time Format*. https://www.iso.org/iso-8601-date-and-time-format.html.

Wickham, H, J Hester, and R Francois (2018). *readr: Read Rectangular Text Data*. R package version 1.3.1. https://CRAN.R-project.org/package=readr.