**Influencing factors for delayed and cancelled flights on Australian domestic routes**

CSC8002

Assignment 2

October 2017

Prepared by Thomas W. Body

u1101544

Table of Contents

[Abstract i](#_Toc493709174)

[1. Introduction 1](#_Toc493709175)

[1.1 Purpose of Report 1](#_Toc493709176)

[1.2 Research Questions 1](#_Toc493709177)

[1.3 Background 1](#_Toc493709178)

[2. The Data 2](#_Toc493709179)

[2.1 Original Dataset URLs 2](#_Toc493709180)

[2.2 Choice of Database 2](#_Toc493709181)

[2.3 Creating the Database and the Tables 2](#_Toc493709182)

[2.4 Cleaning the Data – The DOP and MAP Tables 2](#_Toc493709183)

[2.5 Using Excel and PostgreSQL to Clean the Data 3](#_Toc493709184)

[3. Changes made from the First Report 5](#_Toc493709185)

[3.1 The Research Questions 5](#_Toc493709186)

[3.2 Choice of Datasets 5](#_Toc493709187)

[4. Schedule 6](#_Toc493709188)

[4.1 Schedule Table 6](#_Toc493709189)

[4.2 Keeping to Schedule 7](#_Toc493709190)

[5. Progress 8](#_Toc493709191)

[5.1 Current completed tasks 8](#_Toc493709192)

[5.2 Tasks to complete 8](#_Toc493709193)

[6. Ethics and Privacy 10](#_Toc493709194)

[6.1 The Privacy Act 1988 10](#_Toc493709195)

[6.2 What is covered under Australian Privacy and Ethics Law? 10](#_Toc493709196)

[6.3 Communication of Privacy and Ethics 11](#_Toc493709197)

[6.4 Identifying Individuals 11](#_Toc493709198)

[7. Conclusion 12](#_Toc493709199)

[References 13](#_Toc493709200)

**List of Tables**

**Table 1** Weekly Schedule

**List of Abbreviations**

**BITRE** Bureau of Infrastructure, Transport and Regional Economics

**DBMS**  Database Management System

**GUI** Graphical User Interface

**CSV** Comma Separated Values

**DOP** Domestic On-time Performance

**MAP** Monthly Airline Performance

**APP** Australian Privacy Principles

# Abstract

This report outlines the changes made from the first report done for the class CSC8002, Big Data Management. The first two reports are preliminary reports for the final project report. The research questions and datasets chosen in the first report are discussed, with reasoning given as to why the research questions have not changed and no extra datasets have been added. The current progress towards the final report is discussed, and a scheduled set of tasks is outlined in a weekly format leading up to the due date of the final report. Finally, the ethics and privacy surrounding how individuals should be protected in datasets is discussed within the context of Australian law.

# Introduction

## Purpose of Report

This report was prepared to outline and discuss changes made since the initial report for the data project required for the class CSC8002, the progress made towards the final report, and the schedule on the task completion for the final report.

## Research Questions

What factors are related to cancelled and delayed flights on Australian domestic routes? How have these factors changed over time?

## Background

Flight delays and cancellations rank as the one of the top reasons for complaints to the Airline Customer Advocate in Australia (Lines 2016, p. 3). In order to help consumers make informed decisions about which airline they choose to do business with, the BITRE compile a monthly report which details the number of flights that have been cancelled or delayed across the major Australian Domestic routes (Bureau of Infrastructure 2017a). To make this report, the BITRE collected a number of variables they deemed relevant to assess the overall industry performance as well as the performance of individual airlines (Bureau of Infrastructure 2017a, p. iii).

The BITRE has also released other datasets collating relevant information on airlines and airports. One of these is the monthly figures on airport traffic data, which collates the total number of domestic and international passengers at the 20 largest airports in Australia (Bureau of Infrastructure 2017b).

# The Data

## Original Dataset URLs

The two datasets can be found at the links below. Direct download links to the datasets can be found under references.

* https://data.gov.au/dataset/airport-traffic-data (Bureau of Infrastructure 2017b).
* https://data.gov.au/dataset/domestic-airline-on-time-performance (Bureau of Infrastructure 2017c).

## Choice of Database

The database management system (DBMS) chosen for this report was PostgreSQL. There were two main reasons for this choice. The first was to utilize the graphical user interface (GUI) PgAdmin4. As the author was unfamiliar with any particular DBMS, having a simple and easy to use GUI was deemed necessary to avoid having to rely solely on command line prompts.

The second was the abundant material available which details connecting a PostgreSQL database with the programming language R.

## Creating the Database and the Tables

The database, named *domestic\_flights*, was first setup using PostgreSQL through PgAdmin4. Under this database both the airport traffic dataset and the domestic airline on time performance dataset were imported into two separate tables, called Domestic On-time Performance (DOP) and Monthly Airline Performance (MAP) respectively.

## Cleaning the Data – The DOP and MAP Tables

The domestic airline on time performance dataset overall required very little cleaning before it could be processed. There was however a number of steps that needed to be taken before the data could be imported into a SQL table.

The first of these issues was incorrect value types. For the DOP table, every column can be either one of two types, characters or integers. There are no other numeric types since the nature of the data is discrete and non-zero. For instance, the *sectors\_scheduled* column is simply a count of all scheduled flights in a month, thus this can only be a discrete, non-zero value. During the data import stage it was discovered that there were a small number of floating point values contained within the following columns: *sectors\_scheduled*, cancellations, *departures\_on\_time*, *arrivals\_on\_time*, *departures\_delayed* and *arrivals\_delayed*.

Upon closer inspection of each of these points in the *cancellations­* column, of which there were three affected rows, it appeared that there had been an error when the data was originally input. The *sectors\_scheduled* and the cancellations columns both contained values with a decimal, each occurring on the same row. To fix this, each of these values was rounded up by one. As it was such a small number of rows affected by this error, it should have a negligible impact on the overall statistics.

There was however a much larger number of floating point values in the *departures\_on\_time*, *arrivals\_on\_time*, *departures\_delayed* and *arrivals\_delayed* columns. Roughly 100 rows were affected, which is still a small number relative to the entire population. The cause of these issues may have stemmed from the how these columns were calculated from their primary resources, as many columns tended to have one value that was out in either the departures column that was carried on through to the arrivals columns. The same fix was applied as was done with the other floating point values, with each value being round up or down to its nearest whole value. Again, as it was a relatively low number of rows affected in comparison to the entire population, the effect of this should be negligible.

The second issue was the presence of null values. These occurred solely in the *cancellation* and *departures\_delayed* column. This was a trivial fix however, as the *cancellation* column can be derived by subtracting the *sectors\_flown* column from the *sectors\_*scheduled column, and the *departures\_delayed* column can be derived by subtracting *departures\_on\_time* column from the *sectors\_*scheduled column. In each null case, there were found to be no cancelled or delayed flights. Thus, each null value could be changed to a zero value.

The data in the MAP CSV file had no null values or unexpected value types, thus it did not need to be cleaned prior to importing into an SQL table.

## Using Excel and PostgreSQL to Clean the Data

To make the changes on the original datasets, the CSV files were opened in excel and the required values were changed. Although Microsoft Excel is not the most efficient way to undertake data cleaning on large datasets, there were two reasons why it was used for this report.

The first is that there were a very low number of affected rows that needed to be cleaned, and it was faster and easier to just manually update the values upon inspection rather than to try and automate this process. Also, since there were such a small number of errors, it gave the author the opportunity to understand why the errors occurred and determine the impact that these changes would make in terms of the overall dataset.

The second came from the use of the import function for CSV files in PostgreSQL. PgAdmin4 has an option where data from a CSV file can be automatically imported into a table. If there is an error in the data, the import will fail and show an error message detailing what went wrong with the import.

To utilize this function, these errors were investigated in the original CSV data files and then changed where required using Excel.

# Changes made from the First Report

## The Research Questions

There has been no change to the research questions since the first report. This is primarily due to the criteria of choosing the research questions set out by the author prior to choosing the questions to begin with. The critical points of the criteria were:

* Will there be sufficient, relevant and easily accessible data for the chosen research questions?
* Are the research questions limited enough in scope in order to be reasonably answered within the timeframe of this course?
* Are the research questions broad enough to allow a wide range of possible results to be detailed?

All of the original decisions behind these criteria were still valid, thus there was no need to update or change the research questions.

## Choice of Datasets

No datasets have been added or removed since the first report. As with the research questions the datasets were carefully selected with the aforementioned criteria in mind. The main issue in choosing whether to add more data was how it would affect the feasibility of completing the report by the end of the course. The chosen datasets needed to be large and broad enough to answer the research questions, but small enough to ensure that the report could be completed in time. After cleaning and inspecting the data for this report, it was deemed that the datasets did contain enough valid and relevant data to satisfactorily answer the research questions. Thus, to keep within the scope of this project report, no datasets were added.

# Schedule

## Schedule Table

The following is a table that outlines the weekly schedule for completing the final report starting from the week starting Monday September 18th 2017 to the week ending Sunday 22nd October 2017.

|  |  |  |
| --- | --- | --- |
| **Week Starting on** | **Week Ending on** | **Tasks to Complete** |
| 11-Sept-17 | 17-Sept-17 | * Import all data to tables in a PostgresSQL database. * Start querying the data using SQL. * Research how to make a connection from R to a PostgresSQL database. |
| 18-Sept-17 | 24-Sept-17 | * Deem which columns in both datasets are relevant. * Decide how the data will be combined. * Import the data from the PostgresSQL database to R. * Continue learning R. * Begin preliminary analysis of data. |
| 25-Sept-17 | 01-Oct-17 | * Continue learning R. * Start finalizing the data analysis section. * Proof read and complete report two. * Start writing out report three and complete a draft of the relevant project report sections and headers. |
| 02-Oct-17 | 08-Oct-17 | * Complete data analysis and begin the process of deciding on the visualization and presentation of the data. * Ensure that the data analysis has done a sufficient job of answering the research questions. * Complete the methodology, dataset, database and other methods section of the project report. |
| 09-Oct-17 | 15-Oct-17 | * Complete the conclusion and discussion section of the project report. * Complete the data visualization process and ensure all necessary sections of the project are complete. |
| 16-Oct-17 | 22-Oct-17 | * Proofread and complete the project report. |

**Table 1: Weekly Schedule**

## Keeping to Schedule

By spreading the remaining tasks over a six week period, there should be sufficient time to continue learning the basics of data manipulation in R and SQL while still having time to complete the project report.

The most ambitious aspect of this report was the decision to use R as both the tool to analyze and visualize the data. In this respect, the author is much more confident and proficient in the use of Python for data manipulation and visualization, yet as Python is not covered as part of the course material for CSC8002, R was chosen to demonstrate how the learning material covered in the course was used to complete the project. The author is however confident that the concepts from Python packages, such as *Numpy*, *Pandas* and *Matplotlib*, are just as easily covered in R packages such as *ggplot2*.

# Progress

## Current completed tasks

The following is a list of outlined tasks that have currently been completed.

* Create a database in PostgresSQL.
* Clean the data and make necessary changes to incorrect and null values.
* Import the CSV data into SQL tables.
* Do some preliminary inquiries into the data using R.

There are also a number of tasks that are scheduled to be completed prior to the due date of this report which at the current time of writing are not complete.

## Tasks to complete

The following is a succinct list of tasks that are to be completed along with a brief explanation of how they will be addressed.

* Make a connection to the database in R.
* Import the data from the PostgresSQL database to R.
* Begin preliminary analysis of data.
* Continue learning R.

All of these points will involve going over the relevant course material that covers analyzing data in R. Additional material outside of the course syllabus will also be needed to gain enough proficiency in R to complete the final report.

* Deem which columns in both datasets are relevant.
* Decide how the data will be combined.

These two tasks will be done in conjunction with the preliminary data analysis task. Deciding which columns are relevant to the research question is the last part of the data cleaning process, although in this case it is deciding which data is not useful or redundant, rather than looking for incorrect or missing values.

* Start finalizing the data analysis section.
* Proof read and complete report two.

By the time that this report is due, the data analysis section should be getting close to completion. This will give sufficient time to evaluate how well the analysis has done in answering the research questions. If extra analysis or data is required, this will be the last moment to make that decision.

* Start writing out report three and complete a draft of the relevant project report sections and headers.
* Complete data analysis and begin the process of deciding on the visualization and presentation of the data.
* Ensure that the data analysis has done a sufficient job of answering the research questions.
* Complete the methodology, dataset, database and other methods section of the project report.
* Complete the conclusion and discussion section of the project report.
* Complete the data visualization process and ensure all necessary sections of the project are complete.
* Proofread and complete the project report.

Although there are a significant number of tasks in this final group, they are all interrelated, and will most likely require going back and forth between sections as certain aspects are completed. Most of the actual project report writing will be done in this section, but all of the analysis and decisions on how the report will be structured and what information it will contain should be done by this stage. The visualization of the data should tie back to how and why it was analyzed, and should clearly answer the research questions.

# Ethics and Privacy

## The Privacy Act 1988

The *Privacy Act 1988* is Federal Australian legislation that regulates how personal information is handled (Office of the Australian Information Commissioner 2017a). The scope of this legislation applies to Australian Government agencies (including Australian Federal territories), private organizations with more than $3 million turnover, as well as private health providers and some small businesses (Office of the Australian Information Commissioner 2017a).

## What is covered under Australian Privacy and Ethics Law?

Schedule 1 of the *Privacy Act 1988* covers the Australian Privacy Principles (APP) that govern the management, transfer, use, collection, maintenance and storage of personal data (Office of the Australian Information Commissioner 2017b). These are broken down into 13 specific principles (Office of the Australian Information Commissioner 2017c).

1. Open and transparent management of personal information.
2. Anonymity and pseudonymity.
3. Collection of solicited personal information.
4. Dealing with unsolicited personal information.
5. Notification of the collection of personal information.
6. Use or disclosure of personal information.
7. Direct marketing.
8. Cross-border disclosure of personal information.
9. Adoption, use or disclosure of government related identifiers.
10. Quality of personal information.
11. Security of personal information.
12. Access to personal information.
13. Correction of personal information.

Many of these principles directly impact how large datasets can be collected, used, stored. There are a number of obligations an entity that is bound by the *Privacy Act 1988* must undertake to protect any personal information gathered in datasets. These include:

* Entities must be transparent regarding their management of personally identifiable information.
* Notice must be given upon collection of personal information.
* Personal information must be kept secure.
* Individuals must be able to correct any personal information.

## Communication of Privacy and Ethics

Both of the datasets collected for this report were from the Australian Government open data website. There, they have a blanket statement in their ‘About’ section that details their obligations in publishing and storing personal data (Australian Government 2017). They outline the provisions that they are obliged to comply with, which includes the *Privacy Act 1988* and the APP. There is also a section on how to handle complaints in terms of how the information is used, stored, or presented, and who to contact if a complaint is to be made.

## Identifying Individuals

The *Privacy Act 1988* would apply to every airline that this data was collected by, as they are all well over the minimum annual turnover of $3 million annually. Thus, all the airlines are legally responsible to ensure that these datasets meet all the provisions and principles of the act. Given the way this data has been presented, it is extremely unlikely that any individual could be identified from the data. No personal data is gathered at any point during the data collection phase. To add to this, as the data specifically only deals with airports where there is large traffic (Bureau of Infrastructure 2017b) and routes where there are a minimum amount of passengers (Bureau of Infrastructure 2017a), it would be incredibly difficult to use this data with secondary sources to identify individuals in any way.

# Conclusion

In order to complete the final project, a number of steps needed to be undertaken to clean the data to a point where the analysis stage could take place. These steps included dealing with incorrect variable types and null data. After critical examination of the datasets and the original research questions, it was determined that no changes or additions were required. The research questions did not require any additional data to answer, and the datasets had sufficient data to provide these answers. Examining the data and its collection methods also showed that they were very unlikely to identify any individuals, and that the airlines and agencies that collected and stored the data are doing so in accordance with the *Privacy Act 1988* and the APP.

# References

Australian Government 2017, *About data.gov.au*, viewed 2-September-2017, <<https://data.gov.au/about>>.

Bureau of Infrastructure, TaRE, *Domestic Airlines - On Time Performance,* 2017a, Department of Infrastructure and Regional Development.

Bureau of Infrastructure, TaRE 2017b, *Airport traffic data*, July 2017, <<http://data.gov.au/dataset/cc5d888f-5850-47f3-815d-08289b22f5a8/resource/38bdc971-cb22-4894-b19a-814afc4e8164/download/monpaxweb.csv>>.

Bureau of Infrastructure, TaRE 2017c, *Domestic Airlines - On Time Performance data*, July 2017, <<http://data.gov.au/dataset/29128ebd-dbaa-4ff5-8b86-d9f30de56452/resource/cf663ed1-0c5e-497f-aea9-e74bfda9cf44/download/otptimeseriesweb.csv>>.

Lines, J 2016, 'Response to Issues Paper and Submission', *Australian Consumer Law Review*.

Office of the Australian Information Commissioner 2017a, *Privacy Act*, viewed 15-September-2017, <<https://www.oaic.gov.au/privacy-law/privacy-act/>>.

Office of the Australian Information Commissioner 2017b, *Australian Privacy Principles*, viewed 15-September, <<https://www.oaic.gov.au/privacy-law/privacy-act/australian-privacy-principles>>.

Office of the Australian Information Commissioner 2017c, *APP Quick Reference Tool*, viewed 15-September, <<https://www.oaic.gov.au/agencies-and-organisations/guides/app-quick-reference-tool>>.