<Title Page>

*We give consent for this to be used as a teaching resource. / We DO NOT give consent for this to be used as a teaching resource.*

# Executive Summary

# Table of Contents

[Executive Summay 2](#_Toc52110088)

[Table of Contents 3](#_Toc52110089)

[2 Defining the Problem 4](#_Toc52110090)

[2.1 4](#_Toc52110091)

[2.2 S 4](#_Toc52110092)

[3 Sourcing Methodology 5](#_Toc52110093)

[3.1 5](#_Toc52110094)

[4 Modelling Methodology 6](#_Toc52110095)

[5 Outcomes and Insights 7](#_Toc52110096)

[6 Summary 8](#_Toc52110097)

# Defining the Problem

## What is happening

According to the 2019 Australian Election Study, public trust in democratic institutions in Australia is at the lowest it has been since the 1970s. Just 59% of Australian’s surveyed said they were satisfied with how democracy is working, down from a record high of 87% in 2007, and only 12% said they believed the government was run for ‘all the people’1. This displeasure with government institutions is described further in Figure 1.

Figure 1: Trends in Australian Trust in Politicians and Government Institutions

Stories of government corruption have filled news articles in recent years such as the now infamous Sports Rorts scandal, Barnaby Joyce’s $80M water buybacks and the Department of Infrastructure and Transports recent $30M purchase of land for the Western Sydney Airport which was valued at just $3M.

The trends above shows that public is becoming more sceptical of government, justified by these frequent corruption scandals. Corruption is not new but, with the availability of data in the modern era, the solutions to it can be. Rapidly exposing this corruption and informing voters is the most powerful way to discourage this behaviour.

## How will the problem be addressed

There are many forms of corruption in government; the type of corruption that this project aims to expose is pork barrelling – the use of public funds in a manner which is designed to win elections, regardless of the needs of the people. This first phase of the project will serve as a ‘proof of concept’, focussing on the use of Queensland state funds and state elections. The data collected will include electoral maps, electoral results, the Queensland State Infrastructure Plan and census data. With this information, seats will be categorised broadly into their parties and safe/marginal. From there, relationships between spending and other parameters will be explored using basic exploratory data analysis. Any potential correlations found in the data will have a model developed to determine if there is a statistically significant relationship. Finally, a web tool will be used to convey to voters what behaviour they can expect from their elected officials.

# Sourcing Methodology

## Getting the data

In order to answer to our question and find out whether spending project following a fair pattern, we gathered different type of dataset form various sources listed as below:

1. Queensland state election results and statistics:

This source provides data regarding State election results from 2009 to 2017 by district List. It also includes information about declared seats summary, preferences total candidates nominated for election by each party and two candidate results after distribution of preferences.

<https://www.ecq.qld.gov.au/elections/election-results>

1. Queensland Spatial Catalogue – Q-Spatial:

This service is a Queensland Government initiative to provide improved public access to a variety of spatial and associated data.

This dataset provides spatial data regarding 93 electoral boundaries in 2017

[http://qldspatial.information.qld.gov.au/catalogue/custom/detail.page?fid={079E7EF8-30C5-4C1D-9ABF-3D196713694F}](http://qldspatial.information.qld.gov.au/catalogue/custom/detail.page?fid=%7b079E7EF8-30C5-4C1D-9ABF-3D196713694F%7d)

1. The Queensland government open data portal provides public with the data on current and future Queensland government capital works building projects undertaken across Queensland. This dataset is a CSV file with more than 20 attributes such as the title department and the total estimated cost

<https://www.data.qld.gov.au/dataset/b518dada-3d2a-4d85-bd2c-febe197863c7/resource/410fb21f-8c5a-43a1-8b57-a74a3329d1d0>

1. Census Data based on LGA (local Government areas) owned by Australian Bureau of Statistics. 2016 Census data consists of more than 15,000 features for 89 State Electoral Divisions. Initial EDA and feature selection were carried out based on its metadata. <https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/2016%20search%20by%20geography>

<https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016>

<https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1410.02014-19?OpenDocument>

## Data Ingestion

The first step to ingest the collated data is to find out whether our data is consistence or not and how all data from different sources could be linked together.

From Queensland Spatial Catalogue, a set of vectors for 93 electoral boundaries describing polygons with a latitude and longitude extracted.

One of the attributes in infrastructure dataset is the physical address of each project. By deploying ggmap() library in R, the longitude and latitude of every project from its physical address have been extracted and consequently their corresponding electoral district have been found out.

The census data was based on LGA (local Government areas), there are three levels of local government areas and the third level has same geographical boundaries as electoral district. Filtering helped us to narrow down the data and extract the census data we need for each electoral boundary.

There are 93 entities with more than 1000 major attributes in the combined files. In order to prepare our data for EDA, we merged attributes together and delete unnecessary ones.

For example, gender, background and ethnicity group did not play a role in our hypothesis. Therefore, such this information has been deleted and we manage to reduce the number of attributes.

One thing which we might have done differently if we had time is to categorise attributes before deleting them. In this case we were able to find correlations between them that might result in finding some interesting outcomes.

In this stage, avoid redundancy was one of challenges we encounter in this project. Using Github and creating a database help us to nail in this purpose to a great extent.

## Data Exploration

The integrated dataset was accurate and can be used for our intended purpose. In terms of data quality, the consistency was fashioned. However, for some electoral district (records) some attributes were missed mainly because the measure wasn’t available for that records.

The lost data were not re creatable and therefore imputation methods were not applied to enrich dataset.

# Modelling Methodology

The first step in order to test our hypothesis was conducting an Explanatory Data Analysis approach to find out significant correlation between variables. We perform EDA by developing a correlation matrix showing correlation coefficients between each two variables.

Creating a correlation matrix helped us to summarise data, identify which variables worth taking into a more advanced analysis.

It is worth noting that in the correlation matrix, correlations with p-value > 0.01 are considered as insignificant.

We were surprised at how initial hypothesis of a negative correlation between spending in the same margin was unfounded.

we did find a week linear relationship between spending and the margin for the winning party

prompted us to look at whether the spending was focus particularly in states which belong to the winning party.

approach this with a logistic model and found a relationship that had statistical significance. The scattergun approach to find correlations and nearly 100 variables meant we needed to determine if statistical significance to spending. Truly genuine for a handful of variables. The lowest p value in the data set is for the proportion of the population who walks to work.

the model assumption on residuals were validated and proved to be reasonably acceptable. QQ plot shows residuals are normally distributed. Residuals also have a fairly constant variants around zero mean.

# Outcomes and Insights

## Project outcome

Based on our objective in this project, a web tool is developed to display the data in two different maps in the first map based on the last political districts.

The website also depicts the distribution of projects and their locations on each electoral district. Expenditure and the total estimated cost in that Each project related to one of 13 different departments and the relationship between projects with population of each region.

The website also provides electoral results from 2004 to 2017.

by following data science process from the finding a human centre problem, getting a fit data for use and making the data confess, found some evidence for the hypothesis of political motivation on how to government spend, correlation to the working population.

However, no correlation to the population size in a median household income.

Improved our understanding of government spending, we suggest following more sophisticated statistical techniques widen the scope of the data.

# Summary

## Response to Feedback

# Appendix A – References

## Literature

[1] <https://australianelectionstudy.org/> - REFERENCE TO BE CORRECTLY DETAILED

## Datasets

## Code Libraries and Tools