Siamak Adeli Koodehi 46642800

Cristhyan Cardona Garcia 4666772

Matthew Colwell 4292619

Christopher Symons 46587947

Hai Hung Vu 46634010

Abstract

Does government allocate capital spending fairly according to societal need, or do electoral pressures mean that some places get more than their fair share? How can a taxpayer make sense of the spending data to decide for themselves?

Group 8

DATA7001 Introduction to Data Science

Group Project

6 November 2020

*We give consent for this to be used as a teaching resource.*

# Executive Summary

The project aims to reveal relationships between the spatial distribution of government capital project expenditure across Queensland and the electoral politics of the electorates in which projects are located.

The project will deliver a web-based analytics tool to empower a voter to visualise capital projects, associated expenditure, and electoral political status of electorates in relation to one another and in the context of relevant population statistics.

The project will also apply exploratory data analysis and statistical analysis to the same data to attempt to investigate the possibility that electoral pressure causes expenditure decisions to deviate from what would best serve societal need.

Key take-away messages: (1) Some evidence of political motivation clouds government spending. Correlation to the proportion of the working population but not to the total population or income. (2) What we could do next: more advanced statistical techniques, temporal spending data, more reliable method for classifying seat margin under the preferential voting system, more regions.

# Table of Contents

[Executive Summary 2](#_Toc55483353)

[Table of Contents 3](#_Toc55483354)

[1 Defining the Problem 1](#_Toc55483355)

[1.1 What is happening 1](#_Toc55483356)

[1.2 How will the problem be addressed? 2](#_Toc55483357)

[1.1 What related works exist? 2](#_Toc55483358)

[2 Sourcing Methodology 3](#_Toc55483359)

[2.1 Getting the data 3](#_Toc55483360)

[2.2 Data Ingestion 3](#_Toc55483361)

[2.3 Data ingestion of electoral boundaries 3](#_Toc55483362)

[2.4 (Atttempt at) Data ingestion of PDF spending data 4](#_Toc55483363)

[2.5 Making the Data Fit for Use 5](#_Toc55483364)

[3 Modelling Methodology 6](#_Toc55483365)

[4 Outcomes and Insights 9](#_Toc55483366)

[4.1 Project outcome 9](#_Toc55483367)

[5 Incorporation of feedback and recommendations 14](#_Toc55483368)

[5.1 Presentation day 14](#_Toc55483371)

[5.2 Peer review 14](#_Toc55483372)

[Appendix A – References 15](#_Toc55483373)

[Literature 15](#_Toc55483374)

[Datasets 15](#_Toc55483375)

[Code Libraries and Tools 15](#_Toc55483376)

**Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.Error! Hyperlink reference not valid.**

# 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 Australians 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 demonstrated in Figure 1.

Figure : 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 show that public is increasingly more sceptical of government’s agenda due to these frequent scandals. The use of big data in public domains may help to increase political awareness of concerned tax payers. Rapidly exposing and informing voters of inappropriate agenda is the most powerful tool to discourage this behaviour.

## How will the problem be addressed?

There are many forms of inappropriate behaviours of the government; of which one is “pork-barrelling” which this project aims to expose. “Pork-barrelling” is a form of misusing public funds which is designed to promote one’s political influence 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.

## What related works exist?

Many groups have analysed and visualised government expenditure, and several groups in Australia and elsewhere have published analytic tools aimed at allowing a layperson to visualise government expenditure. But we have found that few of these offerings are focused on the spatial dimension of the data, and, surprisingly, that none have attempted to link expenditure data with electoral political data by using electoral boundaries to attribute an electorate to each expenditure line item.

The Truii / Advance Queenland project qld.govspend.info provides an interactive spatial visualisation of Queensland government expenditure. It includes informative storytelling elements. But it does not attempt to link to electoral political data.

# 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, first preference totals for each candidate 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. Based on its metadata, data related to population, income, education level, employment rate and methods of travel to works were selected for modelling purposes. <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 was to find out whether the data was consistent or not and how all data from different sources could be linked together.

## Data ingestion of electoral boundaries

A set of spatial descriptions of all 93 Queensland electorates was identified at the publicly available Queensland Spatial Catalogue (QSpatial). Each electorate in this data set is represented by a polygon with latitude and longitude coordinates. Upon inspecting the published metadata, we identified that the data was available in several geospatial vector data formats including the commonly used ESRI shapefile format (.shp), for which we discovered a handling method in R. This method uses libraries sp, rgdal, rgeos, and ggmap, and is demonstrated in accompanying Jupyter notebook find\_electorate\_by\_latlong.ipynb. We encountered some complexities with the gContains()function reporting warnings about its arguments (spatial point and polygon) having inconsistent projection/coordinate system. We determined that efficiently resolving this inconsistency would likely require geospatial expertise so we chose to use suppressWarnings()to simply suppress these warnings. We checked and confirmed by testing that, notwithstanding the warning, the function call correctly and reliably returns the correct true/false value.

One of the attributes in the infrastructure dataset is the physical address of each project. By deploying the ggmap library in R, the longitude and latitude of every project could be retrieved from its physical address and consequently their corresponding electoral district could be determined.

## (Atttempt at) Data ingestion of PDF spending data

We located recent spending data in the Queensland Government Capital Program 2020 Update at <https://www.dsdmip.qld.gov.au/resources/plan/capital-program-sept2020.pdf> and attempted to ingest the tables in this document into R. We had some success doing so with the R libraries pdftools and tm but our solution relied on reading each page of the document into an array of each piece of text with its horizontal and vertical position, and then attempting to stitch these pieces together into a table. We had some success but the resulting dataframe would have required substantial manual cleaning to be fit for use because of spacing inconsistencies in the document tables. We abandoned this effort when we located an alternative dataset available in comma separated value (.csv) format.

As part of this effort we successfully developed a Python script to scrape the suburb index from the Australia Post website for use in searching the pdf data for textual locations. This is demonstrated in accompanying Jupyter notebook scrape\_AP\_suburb\_list.ipynb.

[next section]

The census data was grouped by LGAs (local Government areas). There are three levels of local government areas and the third level has same geographical boundaries as electoral districts. Filtering helped narrow down the data and extract the census data needed for each electoral boundary.

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

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.

Avoiding data redundancy was one of the major challenges encountered when ingesting the data. Using GitHub and creating an SQL database helped to mitigate this issue.

## Making the Data Fit for Use

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.

ALL, I wrote a paragraph for this section, see below. Happy to chop and change with above.

To develop a modelling methodology, field linking needed to be applied so that questions such as “does margin of victory at the previous election effect government spending?” could be answered. A question such as this requires relating the spending data to the election results. A question such as “is electorate population strongly correlated to government spending?” required linking the census data with the spending data.

Unfortunately, only 77 out of 93 total electorates have spending allocated to them. This meant that 16 electorates had a 0 or null value which needed to be handled (although many projects involve multiple electorates which were removed e.g. highway upgrades). Since it was not necessarily true that those electorates had zero spending, and there is no valid way to impute their values from other electorates, they were removed from the dataset for modelling.

A similar issue occurred with the census data. Queensland electorates were redrawn prior to 2017 election, whereas the census was conducted in 2016 with the old electorates. Using a document “26.5.17\_Extraordinary-Gazette\_QRC-Final-Determination”, relationships between 2016 and 2017 electorates were constructed and summarised in Table 1. The new 2017 electorates’ Census data were estimated as average of all 2016 electorates, which they were drawn from. For example, Bancroft would have estimated Census data that were calculated from Murrumba, Kurwongbah, Morayfield, Pumicestone. There were some name changes, where Census data remained the same. This transformation helped reducing data loss.

|  |  |
| --- | --- |
| 2017 State Electoral Divisions (SEDs) | Drawn from equivalent 2016 SEDs |
| Bancroft | Murrumba, Kurwongbah, Morayfield, Pumicestone |
| Bonney | Southport, Broadwater |
| Cooper | Ashgrove |
| Hill | Dalrymple |
| Jordan | Bundamba, Inala, Algester, Lockyer, Logan |
| Kurwongbah | Kallangur |
| Macalister | Waterford, Coomera, Albert, Redlands |
| Maiwar | Mount Coot-tha |
| Miller | Indooroopilly, Yeerongpilly |
| McConnel | Brisbane Central |
| Ninderry | Noosa, Nicklin, Buderim, Maroochydore |
| Oodgeroo | Cleveland |
| Scenic Rim | Beaudesert |
| Theodore | Albert |
| Toohey | Sunnybank |
| Traeger | Mount Isa |

Table 1: 2016 and 2017 SED Comparison

Census population data was combined into three groups by age, which are 0-20 year-old (Under\_Age), 21-65 year-old (Working\_Age) and over 65 year-old (Retired) groups. This transformation was based on our understanding that each age group might have different social needs, which might impact the government’s spending. For example, public health, transport infrastructure, education needs were believed to differ amongst the groups. The count data of each group was calculated as a proportion of total population for each electorate. Therefore, during modelling stage, one of the groups (Under\_Age) was excluded to avoid collinearity with the other two (i.e. Under\_Age + Working\_Age + Retired = 100%).

Finally, for data preparation, all spending and income data needed to be transformed to its natural logarithm to linearise the data.

# Modelling Methodology

The first step towards making the data confess was to test the project’s hypothesis by conducting an Explanatory Data Analysis. Figure 2 below shows the outcome of this exercise for some key variables; for the purpose of this analysis, the most important row in the “Log Spending” row.

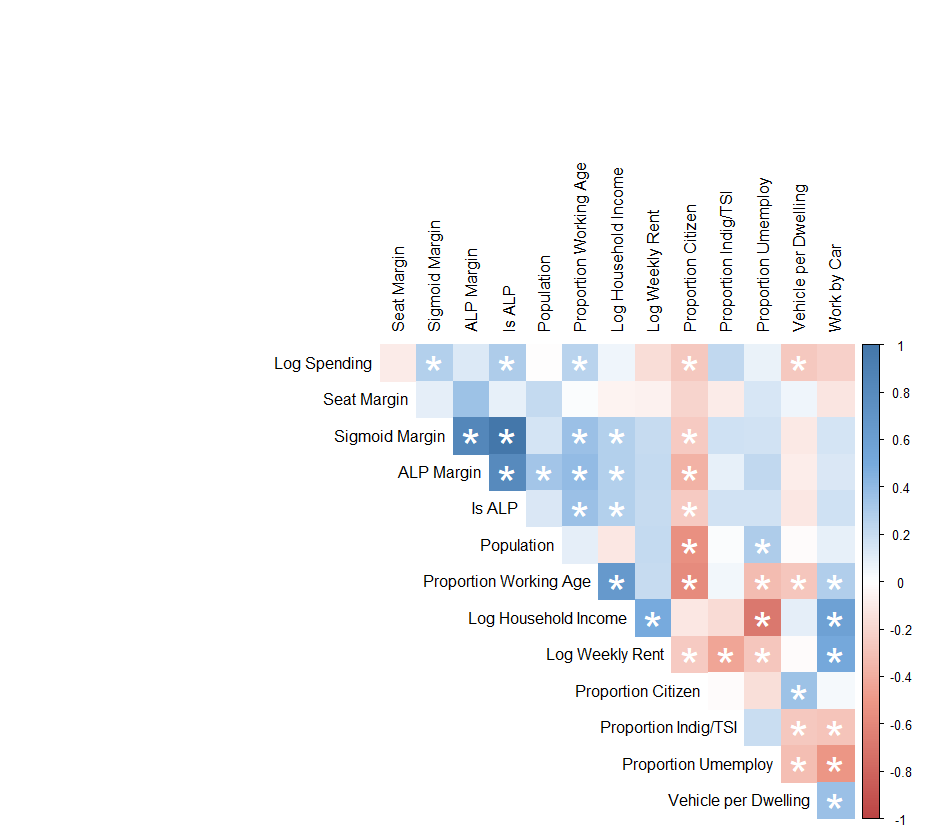


Figure : Preliminary Correlation Matrix. Significant correlations (p > 0.05) denoted with \*.

From the above it can be seen that a weak negative relationship exists between spending and the seat margin, suggesting weak evidence for the original hypothesis. However, Sigmoid Margin, which is a transformation of the ALP Margin variable, has statistically significance. Since the ALP is the party in power, it is somewhat concerning that they could potentially be spending more heavily in electorates where they won. To illustrate this, a logistic model was developed and is shown below in Figure 3.

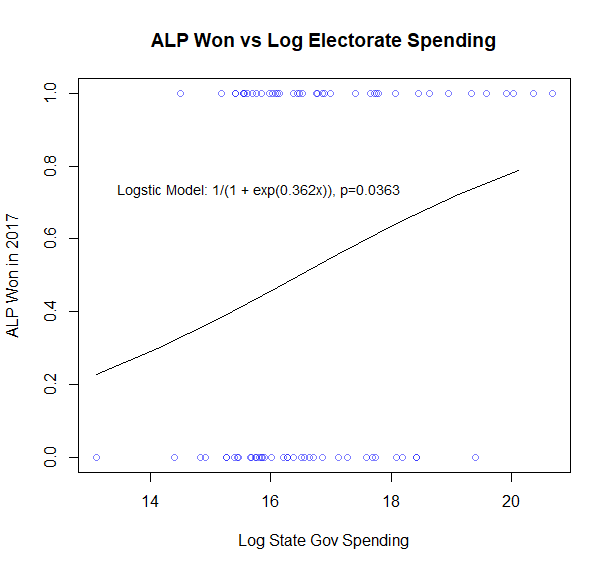


Figure : Logistic Model for Spending as a Predictor for ALP Winning Electorate in the 2017 Election

Because of the number of feature variables compared, several spurious relationships were identified. The strongest correlation by p-value was between spending and the proportion of the population who walks to work. However, these relationships did not satisfy the assumptions of the linear model, as seen in below.

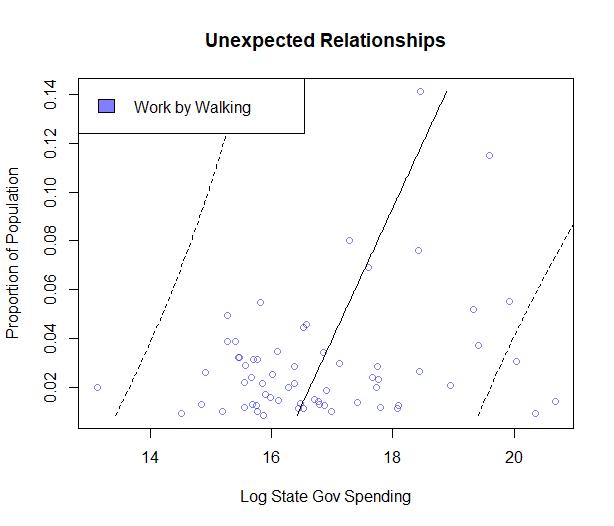


Figure : Linear model for spending as a predictor of the proportion of the population who walks to work

A multiple linear regression model was also constructed using seven predictors, namely Working\_Age, Retired, logarithmic of Median weekly family income, number of full-time employment, isALP and ALP\_Safety\_Ranking\_Margin. The response variable is logCost, which is a logarithmic of TotalEstimatedCost variable. The transformation turned the variable to a normal distribution as shown in Figure 5.

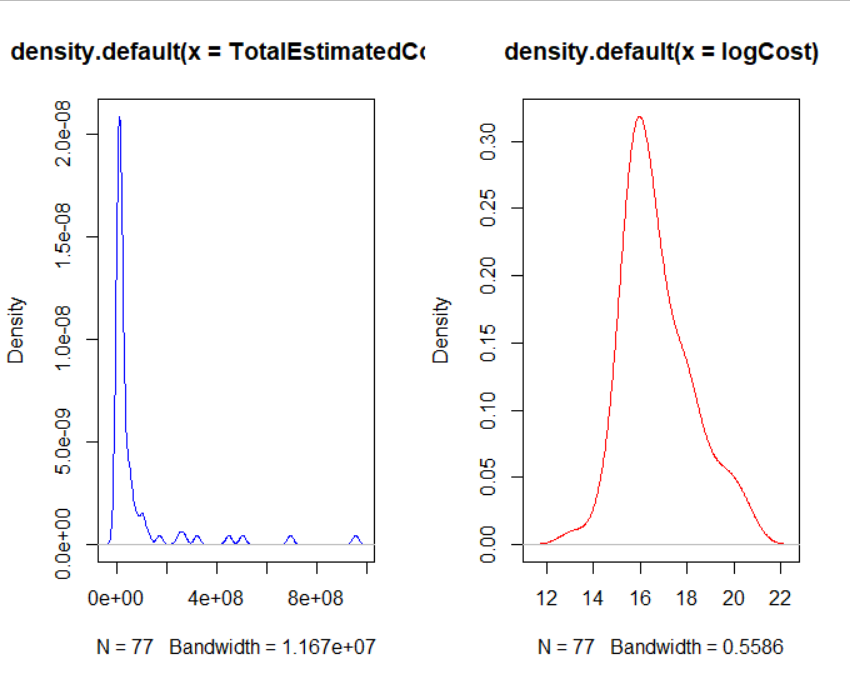


Figure 5: TotalEstimatedCost and logCost distribution

ALP\_Safety\_Ranking\_Margin was formulated based on the two-party preferred election results. As the voting preference ranking was unclear, assumptions were made such that following logics were applied:

* Safety is ranked from 0 to 5, where 0 means ALP dominating an electorate and 5 means ALP being a minority
* If ALP was voted the first preferred party and higher than the second preferred party by a margin that was greater than 10%, safety ranking is 0
* If ALP was the first preferred party and a margin was between 6-10%, safety ranking is 1
* If ALP was the first preferred party and a margin was lower than 6%, safety ranking is 2
* If ALP was not the first preferred party and the first preferred party had a margin less than 6%, safety ranking is 3
* If ALP was not the first preferred party and the first preferred party had a margin greater than 6% and less than 10%, safety ranking is 4
* If ALP was not the first preferred party and the first preferred party had a margin greater than 10%, safety ranking is 5

These assumptions on safety ranking need to be validated by a political science expert. In this project, the ranking was only used to explore if there was any relationship with the government’s spending. The model’s assumptions were validated by checking various plots (Figure 6-9). The assumptions are considered reasonably held.

Chart, scatter chart

Description automatically generated

Figure 6: Residual vs Fitted Values

Chart, scatter chart

Description automatically generated

Figure 7: QQ Plot

Chart, scatter chart

Description automatically generated

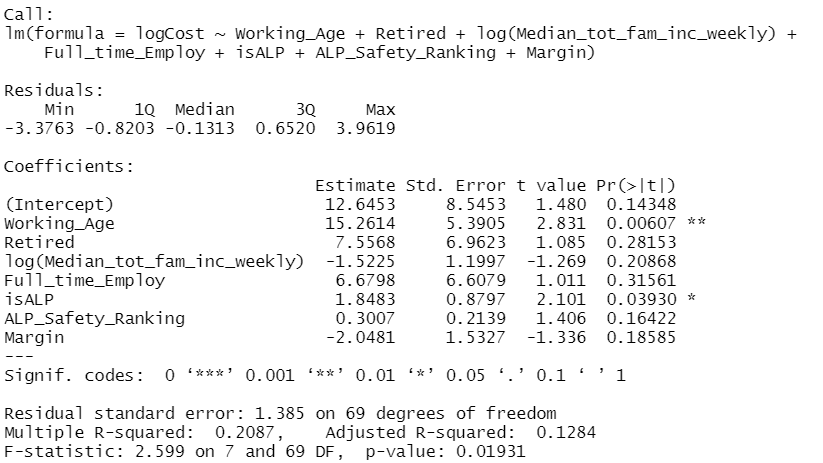
Figure 8: Influential Data - Cook's Distance

Graphical user interface, diagram

Description automatically generated

Figure 9: Checking Linearity with Predictors

The multiple linear regression model’s results are shown below.



Based on F-statistic (p-value of 0.019), there is a moderate evidence against the null hypothesis, which suggests there is a relationship between logCost and the predictors, mainly Working\_Age and isALP. Figure 8 shows a reasonable linearity between logCost and Working\_Age.

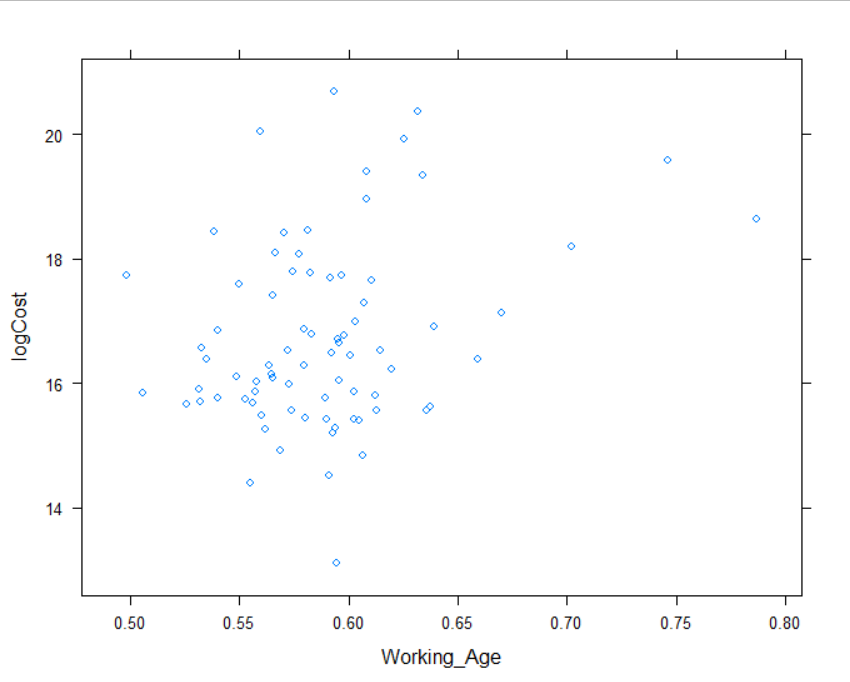
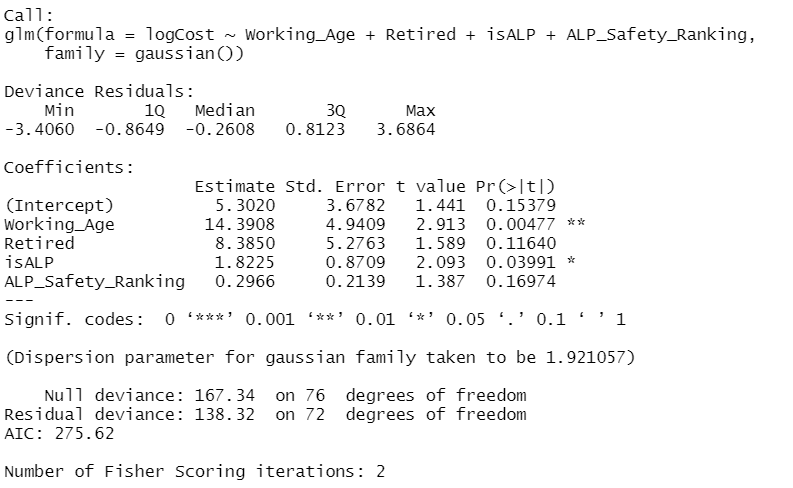


Figure 10: Linearity betwen logCost and Working\_Age

As the response variable is positive real number, a generalised linear regression model was also fitted for four predictors (Working\_Age, Retired, isALP, ALP\_Safety\_Ranking) using identity link function. Its results are shown below. This model has similar AIC as the multiple linear regression model (276 vs 278).



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 > 0.05 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 [[https://data7001group8.com]](https://data7001group8.com/) 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.

Map

Description automatically generated

The website also provides electoral results from 2004 to 2017.

Chart, bubble chart

Description automatically generatedChart, bubble chart

Description automatically generated

Chart, bubble chart

Description automatically generatedChart, bubble chart

Description automatically generated

Chart, bubble chart

Description automatically generatedChart, bubble chart

Description automatically generated

The website also provides electoral results divided by each district.

Map

Description automatically generatedMap

Description automatically generated [[https://data7001group8.com]](https://data7001group8.com/) also has the feature to illustrate the location of each project by drilling down on the map

Map

Description automatically generatedGraphical user interface, application, table, Excel

Description automatically generated

Based on findings of the regression models, there was no clear evidence of pork-barrelling theory which suggested the government would allocate more budget to swing electorates. However, there is a moderate evidence for political motivation where the winning party allocated more budget to their won electorates. There is also an evidence that more spending was allocated to electorates that have higher proportion of working age population. However, the total population size and median household income have no correlation to the spending.

The project team was aware that our data was only a snapshot of the government’s spending, election results and Census data. Due to time limitation and significant amount of data, the team decided it would be more efficient to explore the snapshot data to find features that are the most relevant in understanding the government’s spending. Based on these findings, the project’s scope and data would then be extended if time allowed.

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 age group (21-65 years old).

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.

# Incorporation of feedback and recommendations

## Presentation day

[yet to be done by CJRS]

## Peer review

We are grateful to have received peer review comments on our 28/10/2020 presentation from nineteen (19) of our classmates. We have distilled the comments into a list, giving weight to repeated themes and to sentiments expressed as actionable, specific, and objective recommendations. The list is presented below, interspersed with our responses in italics:

* Don’t show code on your PowerPoint slides. It’s not possible to make sense of it in the time available and perhaps not even legible due to text size and resolution available.
  + *Agreed. We appreciate that 30 or 60 seconds is not sufficient to make sense of half a screen of code. This comment relates only to the presentation and does not have any bearing on the report. No action required to resolve.*
* Support your conclusions more clearly with inferences drawn from data. (Or, at least, put conclusions in the context of the data and findings.)
  + *Agreed. In our enthusiasm to maximise inclusion of interesting content, we did not retain enough detail in the conclusion section linking back to what we found in the data. Action complete: We have included more complete linking of conclusions back to data in [section 4] of the present report.*
* Link your conclusions back to aim of project / stated problem.
  + *Agreed. Similar to above item. Action complete: We have included more complete linking of conclusions back to the aim of the project in [section 4] of the present report.*
* Try to use time series / historical spending data. (Including because election year yes / no might be an interesting dependent variable.) This may allow more and more convincing inferences from data.
  + *We agree that this would add much richness to the data and open up new possibilities to draw more and more convincing inferences. Despite an extensive search for time series / historical data, we were limited by availability of open data here. No action.*
* State the topic of your project clearly and at the start.
  + *Agreed. Action complete: We have stated the topic on the cover page, in the executive summary, and in [Section 1] of the present report.*
* Identify your Storytelling with Data section more clearly. It seemed to be missing or perhaps blended with other sections.
  + *Understood. In our enthusiasm to use the inverted pyramid storytelling structure, we perhaps deemphasised our Storytelling with Data component by splitting it between the web tool demonstration and an integrated part of the Making the Data Confess section. Action complete: We have included a distinct section for Storytelling with Data [Section 5] in the present report.*

# Appendix A – References

## Literature

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

Queensland Redistribution Commission 2017, Queensland Government Gazette Extraordinary, Queensland Redistribution Commision, <https://www.ecq.qld.gov.au/\_\_data/assets/pdf\_file/0021/4944/26.5.17\_Extraordinary-Gazette\_QRC-Final-Determination.pdf>

## Datasets

|  |  |  |
| --- | --- | --- |
| Data Source | File | Description |
| Australian Bureau of Statistics | 2016\_GCP\_SED\_for\_Qld\_short-header.zip | 110 CSV data files, metadata and readme folders.  Based on the metadata file (Metadata\_2016\_GCP\_DataPack.xlsx), following files were selected for modelling:   * 2016Census\_G01\_QLD\_SED.csv * 2016Census\_G02\_QLD\_SED.csv * 2016Census\_G18\_QLD\_SED.csv * 2016Census\_G30\_QLD\_SED.csv * 2016Census\_G40\_QLD\_SED.csv * 2016Census\_G59\_QLD\_SED.csv |
|  | SED CODES to NAMES.xlsx | Created based on 2016Census\_geog\_desc\_1st\_and\_2nd\_release.xlsx |
| Electoral Commission Queensland | 26.5.17\_Extraordinary-Gazette\_QRC-Final-Determination | Details how 2017 electorate boundaries were redistributed |
|  |  |  |
|  |  |  |
|  |  |  |

## Code Libraries and Tools

find\_electorate\_by\_latlong.ipynb (CJRS note: on GitHub as at 5/11/2020)

scrape\_AP\_suburb\_list.ipynb (CJRS note: on GitHub as at 5/11/2020)

|  |  |
| --- | --- |
| File | Description |
| *Census data with 2017 State Electoral Division Names.R* | Data preparation:   * Importing relevant census data with some transformation * Converted SED codes to names * Rectified discrepancy of SED due to 2017 boundary redistribution   Output: Census\_3\_Corrected SED.csv |
| *MLR model.R* | Input data:   * Census\_3\_Corrected SED.csv * projectscsv.csv * election\_results\_two\_preferred.csv   Output:   * EDA with boxplot, histogram, density plots * Summarise cost data * Calculate APF\_Safety\_Ranking\_Margin * Multiple linear regression with all predictors * Multiple linear regression with 7 predictors. Checked assumptions and reporting model’s output * Generalised linear regression using identity link function. |
|  |  |
|  |  |
|  |  |
|  |  |