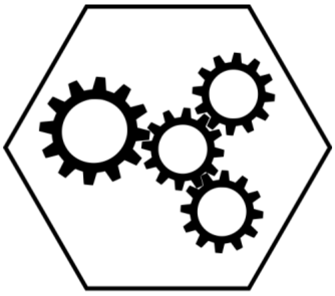
Dissertation for MSc Programmes (ODL)

Project Proposal



Please only complete this form if you intend to take STATS5084P.

About you

Name	Carlos YANEZ SANTIBANEZ
Student number	2603873Y
Version / Date form submitted	27 August 2022
Programme	MSc in Data Analytics (ODL)
Route	Regular route (3 years)
Courses you have already taken (not required if you are on the standard regular or faster route)	Click or tap here to enter text.

Submission

Desired submission deadline	April submission

Students who select a May submission will be able to graduate in June (provided they have no pending reassessments). Students who submit in June will be able to graduate in November/December (provided they have no outstanding reassessments).

Summary of proposed project

What is the nature of the project? *Tick all that apply.*	X Data analysis ☐ Software (and/or methodological) development

Overview

Summarise the project in around 50 words.

This project will aim to analyse what are they factors driving primary voting in Australia's lower house electorates. Using data from 2007 onwards, I will attempt to understand if the results in the 2019 and 2022 elections in each electorate can be explained by local demographic changes, state and make-up of local economy and effects of climate change.

Detailed Description

Give a detailed description of the project in around 200 words.

Like in many countries over the last decade, the Australian political system has experienced some radical changes. Demographic changes (age, migration, city/rural divide), climate change and the economics have been highlighted by electoral experts as drivers for change – however sometimes how this drive the voting patterns is unclear (e.g. traditionally conservative electorates showing a Green surge, despite their "economically left" policies). Being able to identify and profile voting patterns in different electorates is not only interesting to political parties and analysts but to advocacy organisations as well. Elections to Australia's Federal House of Representatives are interesting as an analysis subject, due to compulsory voting and a preferential voting system, where minor parties can influence election outcomes. This is complemented by good data from government sources like the Australian Census, which is held every 5 years.

For this project, I propose to conduct an analysis of election results from 2007 onwards (2010, 2013, 2016), which combined with a number of data sources may help to address the following questions:

- Is it possible to quantify the main drivers of *primary voting* (candidate voted as number one in a preferential ballot, see *Additional Information)?*
- Has climate change been one of the drivers? Is it possible to quantify it in any way (using for instance meteorological data)
- Would have been possible to predict the outcomes of the 2019 election (lost by the party predicted to win by all analysts), using the previous data and regression or classification methods?
- Would have been possible to predict the outcome of the past 2022 election using the same model?

Project details (Data analysis)

Only complete this section if your project has a data analysis component.

Overview

Objective

What are the questions of interest you want to answer (using data)?

I want to analyse demographics, economics and climate events have influenced Australian federal election results over the last decade. This can be divided in the below questions:

- How do demographics and the economy affect the *primary vote* (number 1 choice in ballot) in each electorate.
- If the impact of climate events/climate change can be quantified in terms of how it has altered voting patterns.
- How does the above affect "traditional voting patterns" (e.g. rise of green voting in traditionally conservative electorates, city/rural divide).
- It is possible to quantify the trends in voting patterns to see which electorates are changing and in which direction (i.e. the effect of migration and ageing in electorates' primary vote)?

Methodology

Select all that apply:

Regression	X	Classification	X	Clustering	X
Time series	□	NLP	□	Deep Learning	□

Other specialised analysis (Graph/Network analysis etc) please give more details in following box. □

Click or tap here to enter text.

Please briefly describe your proposed methodology in 50 words

I propose the following methodology of analysis:

1. Consolidate all the data by electorate, to profile each of them.
2. Select the most relevant variables and cluster electorates in groups.
3. Use regression techniques to find what drives the primary vote for each block.
4. Fit a model using the results from elections from 2010 to 2019.
5. Use the model to estimate primary vote results for April 2022 election.

Prior work

If the answer to the next question is "Yes", please complete the prior work section in the next section.

Has this problem / data set already be analysed (either by you or someone else)?	Yes
Is your approach different to previous approaches?	I know that this electoral profiling and forecasting work has been attempted by political / advocacy organisations in many countries. However, these work/ these models are not public, so I am not able to specify how this work is different.

Reasonable pipeline, but need a little more details on the type of regression methods that you are planning on using here

How do you plan on detailing with Ranking?
(perhaps want to use a model which deals with it explictly?)

Detailed Description

What techniques and methods will you use to analyse the data? How do you think will they be applied in the context of your project? (In a regression context, explain what variables will be used as response and what variables will be used as predictors, as well as what types or regression models you could employ.)

I foresee this work will require the following methods and techniques:

- Regression and/or classification techniques for variable selection, to narrow down the most important drivers.
- Potentially, use of principal component analysis or factor analysis techniques to reduce/transform the main drivers.
- Clustering techniques to find out similar electorates and population groups.
- Regression techniques (including methods covered in Advanced Predictive Models) to attempt to predict primary vote outcomes.

Depending on time, if social media data can be included, sentiment analysis will be also included.

Will analysing the data require methodological development or substantial software development?	Not foreseen

If so, please also complete the section *Project details (Software / methodological development).*

Prior work

Is similar analysis to the proposed project published in the scientific	NA

literature or online (such as on Kaggle)?	

If so, please provide citations and/or web links.

Click or tap here to enter text.

Please explain in detail, how your work will be different to prior work.

It is well-known that data analytics has been used to understand electorate demographics and voting intention. Not only political parties but advocacy organisations want to understand the electoral mood. However, while searching for antecedents on the Internet and in the Library search, it was not possible to find any similar published work.

For the project, I intend to focus the analysis on long-term trends, to try to understand the effect of demographic/values changes instead of opinion polls. I also will incorporate data beyond basic demographics – such as climate data and economic statistics.

Data (Data analysis)

What data do you want to use to answer those questions? Describe both the nature and the quantity of data you want to use.

Overview

Give a brief description of your dataset:

For this project, I plan to use several datasets. The two main series of data sets are:

- Election results for Australian Federal elections in the last 10 years
- Regional Statistics by Local Government Area, for the last 10 years.

This data will be complemented with the following

- Taxation statistics
- Meteorological

If time allows it, historical social media data will be added, which can be source from Twitter.

Where is this dataset from:

Most of the data will be sourced from government sources, namely:

- Australian Electoral Commission
- Australian Bureau of Statistics
- Australian Taxation Office
- (Australian) Bureau of Meteorology

Social media data will be sourced from Twitter via their API.

Is the data publicly available?	Yes
If not, do you have permission to obtain and process/analyse the data?	NA

If the data is not in the public domain, we strongly recommend that you obtain such permission in writing and/or that a three-way non-disclosure agreement is established.

Have you already obtained the data (rather than just have access to it)?	No

If not, what do you need to do to obtain the data?

Data is publicly available on government websites, can be downloaded at any time.

Detailed Description

Please provide as where the data is obtained from, how the data was/is being collected, what observations represent and what information (variable) is available in the data.

Data will be downloaded from the website of each institution or using R packages if available, i.e.

- https://results.aec.gov.au/
- https://www.abs.gov.au/census/find-census-data/historical
- https://www.abs.gov.au/statistics/labour/employment-and-unemployment/labour-force-australia-detailed/latest-release#labour-market-regions-sa4
- https://explore.data.abs.gov.au/?fs[0]=Data%20by%20region%2C0%7CLocal%20Government%20Areas%23LGA%23&pg=0&fc=Region
- https://www.ato.gov.au/about-ato/research-and-statistics/tax-and-superannuation-statistics/
- https://ropensci.github.io/bomrang/

All the datasets from the above list have been collected by government institutions as part of the core work - election day, Census, surveys, tax records, and meteorological measurements.

In general terms, the datasets contain a wide range of variables characterising the population of Australia and the social, economic and physical environment where they live.

For past tweets, they can be obtained from the Twitter archive: https://developer.twitter.com/en/docs/tutorials/getting-historical-tweets-using-the-full-archive-search-endpoint

Will you be able to provide the data (or a representative subset of the data) so that your analyses can be reproduced by the markers of the project? *You might require to obtain permission for this from the owner of the data.* *(see below for confidentiality agreements)*	Yes

If not, please explain why this is not possible.

Click or tap here to enter text.

What variables in the data (or what part of the data in more general terms) do you think will help you answer the questions of interest?

From common knowledge, variables related age and income have been traditionally associated with voting patterns. For this project, I would like to examine which other variables may be associated with voting trends, for instance how employment in particular sectors or migration rates affect voting patterns, and I would also like to qualify the impact of climate events – which has been highlighted as a key election issue in the last decade.

What are the data types of these variables? Please also discuss potential issues with data quality and missing values.

Most variables are numerical in nature.
Since the data comes from trusted data sources, I don't foresee issues with data quality or missing values.
However, a challenge may arise when mapping data to the different electorates since not all data will be organised by electorates.

Project details (Software / methodological development)

Only complete this section if your project has a methodological or substantial software development component.

Overview

Objective
What will the method and/or software you want to develop allow the user to do? How and in what context can it be used?

Click or tap here to enter text.

Methods

Broad Category

Select all that apply:
Regression □ Classification □ Clustering □
Time series □ NLP □ Deep Learning □
Other specialised analysis (Graph/Network analysis etc) please give more details in following box. □

Click or tap here to enter text.

Please briefly describe your proposed methodology in 50 words

Click or tap here to enter text.

Detailed Description

Describe in detail what data analytical models and/or methods you will implement in your project.

Click or tap here to enter text.

Will the project require you to develop novel methodology or adapt existing methods? (This is not required)	Choose an item.

If so, describe what methods need to be developed or how existing ones need to be adapted.

Click or tap here to enter text.

Prior work

Does software with similar functionality already exist?	Choose an item.

If the answer to this question is “Yes”, please complete the remainder of this section. Otherwise proceed to the next section.

Are such implementations publicly available?	Choose an item.

If so, please provide citations and/or web links.

Click or tap here to enter text.

Please explain how your work will be different to prior work.

Click or tap here to enter text.

Connection to programme

What methods and topics covered in the MSc programmes will your project relate to?

Click or tap here to enter text.

What data analytics methods and topics <u>not</u> covered in the MSc programmes will your project relate to?

Click or tap here to enter text.

Software

Which software are you intending to use for the project?
In case of a software development project, list the software your project will depend on.

Overview

What software do you intend to use for the analysis? Tick all that apply.	X R ☐ Python ☐ SAS ☐ Other (please specify below)
If you project has additional requirements in terms of computational resources (CPU/GPU, memory and/or storage), explain what resources you will use. (You should complete this field if you plan to make use of software such as use Tensorflow or Spark).	NA

Detailed Description

What R packages / Python libraries are you likely to use? (other than standard packages such as ggplot2, tidyverse, pandas or numpy) Also use this field to list other software you intend to use.	Apart from standard regression/classification libraries, I also expect to use any existing package to access the data, such as {bomrang} https://ropensci.github.io/bomrang/ for data from the Bureau of Meteorology.
Is there substantial amount of software development required for this project (i.e. the re-implementation of methods, above and beyond using data science libraries)	No – unless it makes sense to package any code use to download and pre-process the data functions to access data where it has not been developed yet (i.e. R interface for Australian Census data)
Is there any other special software requirements of this project	No

Ethical approval

Overview

Will your project involve working with human/animal subjects, human/animal material or data about individuals that is not anonymised and not already in the public domain?	No

If the answer to this question is "Yes", please complete the remainder of this section. Otherwise proceed to the next section.

Can the research and data be considered clinical?	NA
Have you already obtained ethical approval from the University or another institution?	NA
Have you already applied for ethical approval from the College Ethics Committee (or in case of clinical data the NHS)?	NA

Comments

Click or tap here to enter text.

Data Protection (GDPR)

Overview

Will the project involve working with non-anonymised personal data that could be subject to GDPR?	No

If the answer to this question is “Yes”, please complete the remainder of this section. Otherwise proceed to the next section.

Will the data provided to you be provided by an organisation who will benefit from the deliverables of the projects or can you be seen as an agent acting on behalf of that organisation?	NA

If the answer to this question is “Yes”, please complete the remainder of this section. Otherwise proceed to the next section.

Do you have confirmation (ideally in writing) from the organisation confirming that they will be the data controller under GDPR?	NA

Comments

No personal data will be used.

Confidentiality

Overview

Is any data you will work with and/or information you will mention in the portfolio confidential?	No

If the answer to this question is “Yes”, please complete the remainder of this section. Otherwise proceed to the next section.

Do you require the University to enter a confidentiality agreement with you and/or the owner of the information and intellectual property in question?	No

If so, please provide who would be parties to such a confidentiality agreement. We strongly recommend using a non-disclosure agreement to which you (and not just your employer) are a party as well.

Click or tap here to enter text.

Comments

Click or tap here to enter text.

Risks

Describe the main risks that could prevent you from successfully completing the project. These risks should be specific to the project (such as data not becoming available or being of a different nature or the complexity of the project getting out of control). How likely are these risks? What would be their impact? What can you do to prevent these risks ore mitigate the effects?

Description of the risk	Likelihood / Impact	Prevention / mitigation measures
Too many variables from source data.	High/ Medium	Use regression or another method for variable selection, combined with hypotheses from common knowledge of Australian politics. Focus on selection on the first quarter of the work.
Variables don't have enough predictive power.	Medium/ High	This is a conclusion on its own and should be treated as such. However, there are other sources of data that can be obtained for academic purposes (like attitude surveys), for which I will request access.
Project takes too long	Medium/Medium.	Structure project work in stages, as defined in the methodology. Establish intermediate objectives and expected timelines.
Click or tap here to enter text.	Choose an item. Choose an item.	Click or tap here to enter text.
Click or tap here to enter text.	Choose an item. Choose an item.	Click or tap here to enter text.
Click or tap here to enter text.	Choose an item. Choose an item.	Click or tap here to enter text.
Click or tap here to enter text.	Choose an item. Choose an item.	Click or tap here to enter text.

Comments

Click or tap here to enter text.

Additional information

If there is any additional information about the proposed dissertation topic you would like to provide to the project team, please use the space below.

Background information on the Australian Electoral system:

As a constitutional parliamentary democracy, Australians elect their government by choosing MPs to represent local electorates. Electorates are defined by the Australian Electoral Commission to contain approximately the same number of voters.

Candidates are selected by preferential voting (https://en.wikipedia.org/wiki/Instant-runoff_voting), where voters are asked to rank the candidates by order of preference. Votes are first assigned to the candidates marked as first preference. If no candidate achieves an absolute majority, the candidate with the lowest number of votes is removed and their votes are assigned based on second preferences. The process is repeated until one candidate obtains 50%+1 of all the ballots.

Voting is compulsory for all eligible voters (Australian citizens and British subjects who were enrolled prior to January 1984), and there is automatic enrolment. Voting participation usually is over 95%.