ETL Project Report

Anne Wieggers, Catherine Sloan, Danielle Cahill, Evrard Harris

Project Overview

We will extract, transform and load data related to the 2016 Australian Federal Elections and the 2017 Australian Marriage Law Postal Survey. It is expected that this project will generate a database that can be used for trend analysis on election and survey results in relation to the Australian electoral divisions. In particular, the data will allow for an analysis of the relationship between the results of the 2016 Federal Election and the outcomes of the Marriage Law Postal Survey. The database will also enable further exploration of the Australian electoral divisions and whether there is a relationship between the socio-economic factors of each division and the results in either the election or the survey. This analysis is possible with the inclusion of Commonwealth Electorate Data from the Australian Bureau of Statistics which considers various social and economic determinants for each electorate. The data in question will be valuable in exploring if there is a relationship between socio-economic factors and the way Australian's vote, both in terms of their voting decisions and their voting methods (such as postal voting, in person voting). It will also be valuable in considering a relationship between each electorates political party and the way they responded to the Marriage Law Postal Survey. Any patterns identified through an analysis of the database could be useful for political parties when running their campaigns for future elections or for the campaign organisers of any possible future postal surveys.

Data Sources

Data source	**Description**	**Retrieval**
1. 2016 Federal Election Vote Types By Division - Australian Electoral Commission Date accessed: 23/02/21	The data source contains the number of vote types per division for the 2016 Federal Election. This includes the number of ordinary votes, absent votes, provisional votes, pre-poll votes and postal votes. The total votes, total percentage of votes and enrolment numbers per division are also included.	Downloaded the 'Votes by division' file as CSV. The file was then read using pd.read_csv.

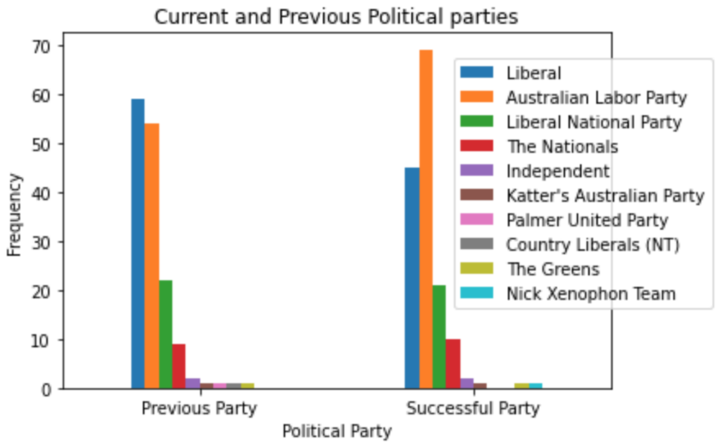
2. 2016 Federal Election results - Australian Electoral Commission Date accessed: 24/02/21	The data source contains divisional classifications of all states and territories. This includes the parties elected following the 2013 election and the 2016 election, along with the seat status at that time. The enrolment number for each division is also included, as is the demographic of the division.	The data was unavailable for download so it was accessed using web scraping. Specifically, Pandas read_html() was used to scrape the data from the HTML table on the AEC website.
3. Australian Marriage Law Postal Survey: Electorate Results - Kaggle Date accessed 23/02/21	The data source contains the results by electorate for the voluntary Australian Marriage Law Postal Survey conducted in 2017. The survey asked “should the law be changed to allow same-sex couples to marry?”, with respondents required to mark either the ‘Yes’ box or ‘No’ box.	Downloaded the electorate-results file as CSV. The file was then read using pd.read_csv.
4. Australian Marriage Law Postal Survey: Participant Information - Kaggle Date accessed 23/02/21	The data source contains participant information for the Australian Marriage Law Postal Survey. It outlines the number of participating and eligible voters in terms of their electoral division, state, gender and age range.	Downloaded the participant information file as CSV. The file was then read using pd.read_csv.
5. Commonwealth Electorate Data - Australian Bureau of Statistics Date accessed: 25/02/21	The source contains eight tables of Commonwealth Electoral Division data. The tables used include: 1. Population, 2. Age Groups, 5. Cultural diversity and 8. Education. Tables 1-2 are based on Estimated Resident Population data as of 30 June 2017. Tables 5 and 8 are based on the 2016 Census of Population and Housing.	The data was downloaded as an excel file. It was accessed using pd.read_excel with the specific sheet name referenced in some instances. It was also converted to csv files and read using pd.read_csv.

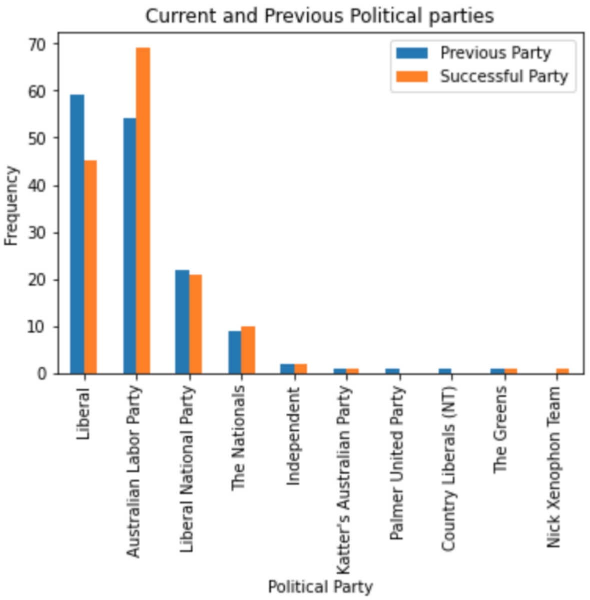
Exploratory Data Analysis

Exploratory data analysis was conducted in order to investigate the datasets and better understand the resources we were working with. Summary statistics were implemented to view relationships between the columns and the data set as a whole. They also allowed us to see that the data would be useful and appropriate for this type of analysis. Visualisation methods were used as a way to spot anomalies, understand features of the data and identify the number of observations. Plotting the data using bar charts helped visualise the values and determine if there were any inconsistencies with the datasets.

02- election_results

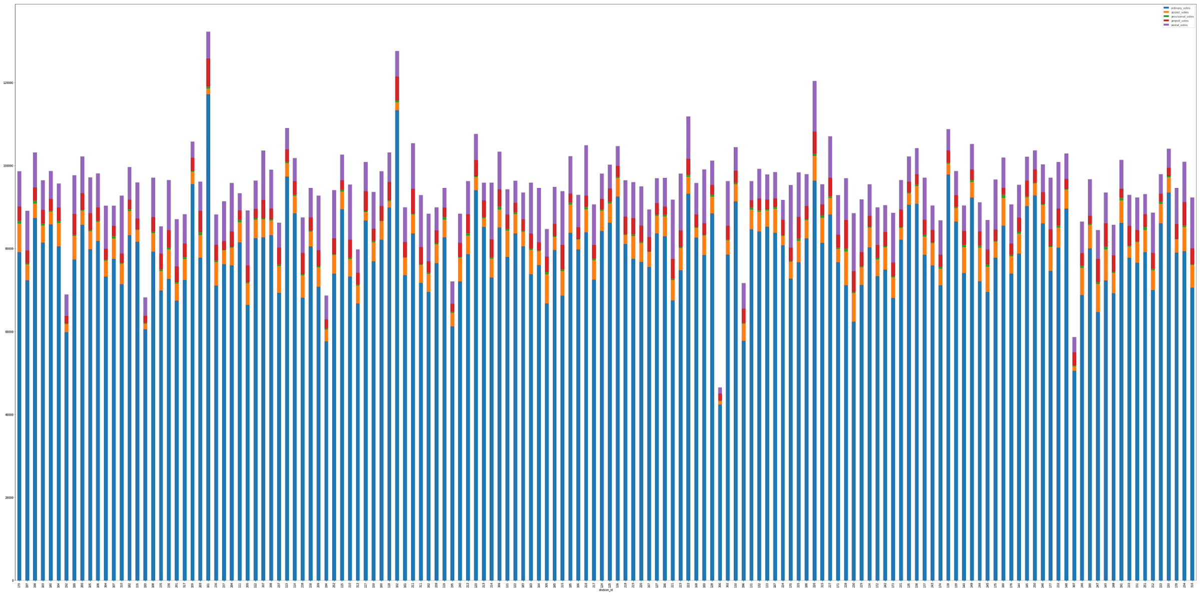
Bar charts were used to visualise the particular values that were present in the dataset. A comparison was made between the frequency of each political party following the 2016 election, and the frequency of each political party following the 2013 election. This allowed it to be noted that there was a differing number of parties represented in each year, and that some parties were not represented across both years. This led to a greater understanding of the data being considered.





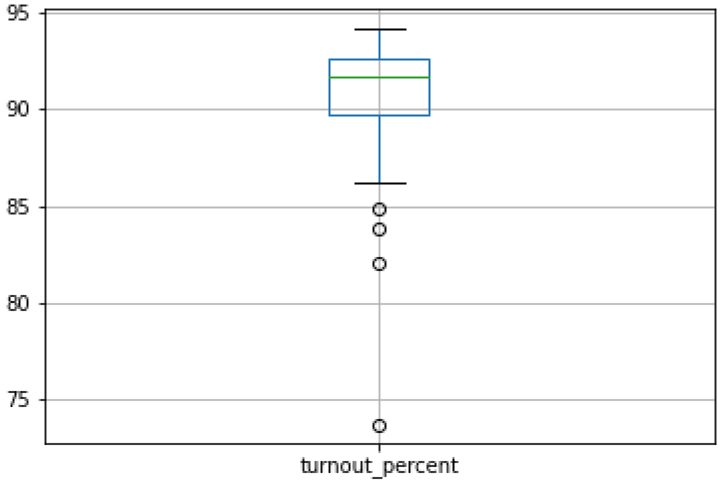
03-election_vote_types

A stacked bar chart was created to visualise how many different votes there were for each type in each electorate. All five vote types were visible for all 150 electoral divisions.



04 - election_turnout

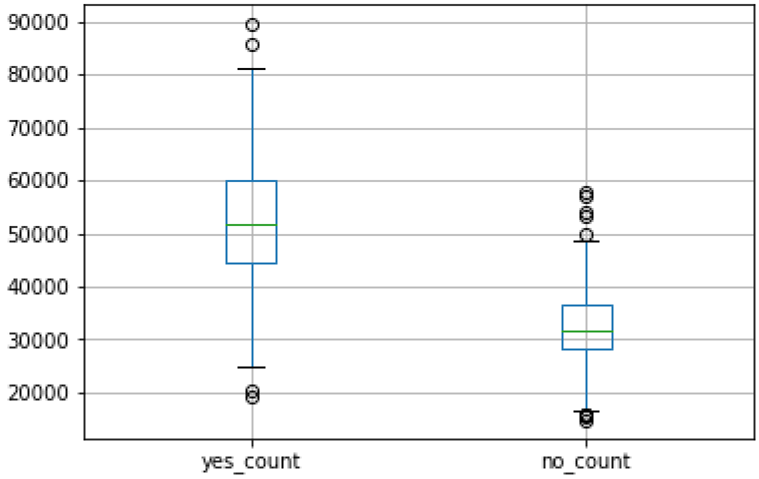
A box and whisker plot was created to investigate the turnout percentage of the federal election. This was done to see how the data is spread out across the 150 electoral divisions. As you can see from the image below there are potential outliers in this dataset.



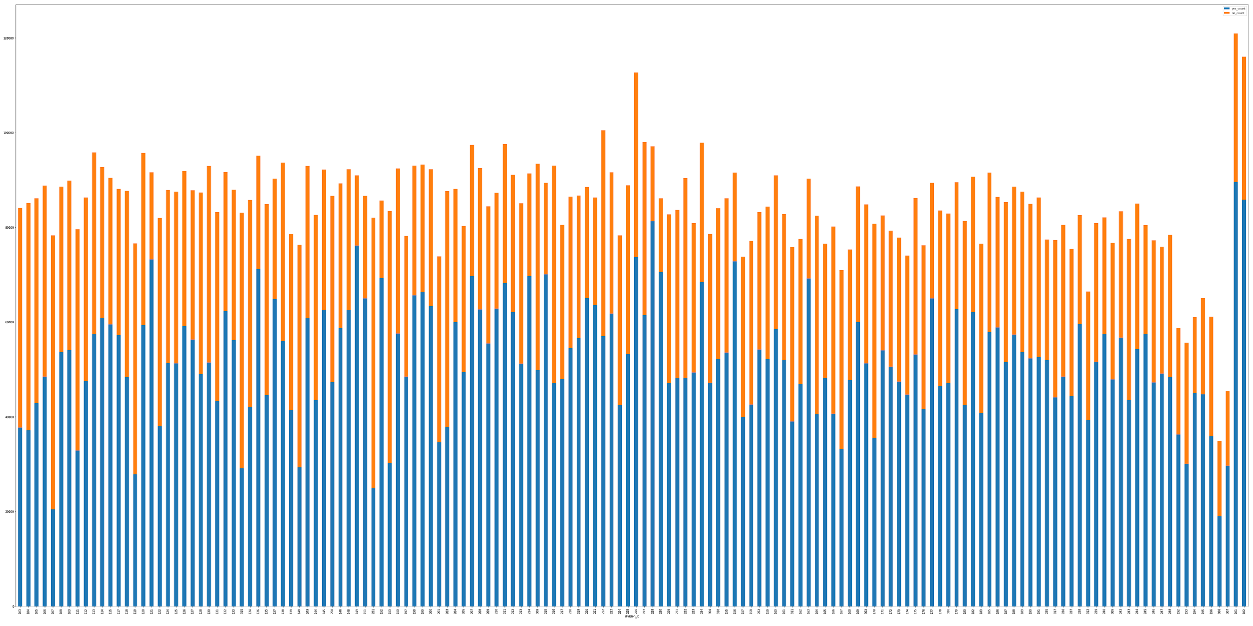
05 - marriage_postal_results

Created two different visualisations to investigate the yes and no count of the marriage postal survey.

Box and whisker plot below shows that there are potential outliers for both vote counts.

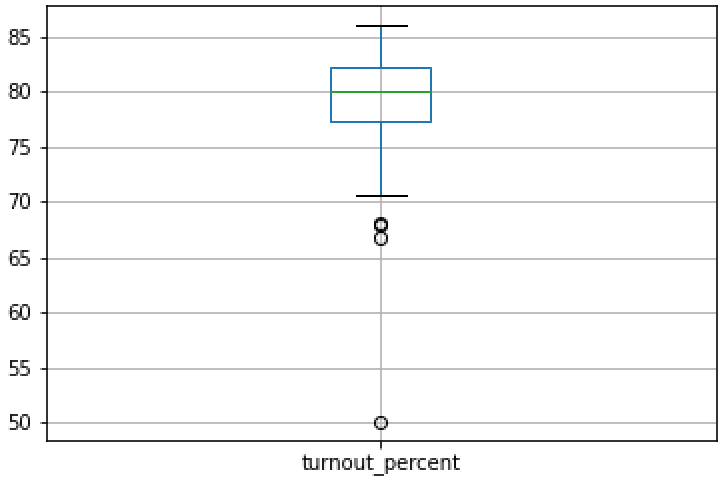


The stacked bar chart shows a visual comparison of yes and no votes and that both types were recorded for all 150 electoral divisions.



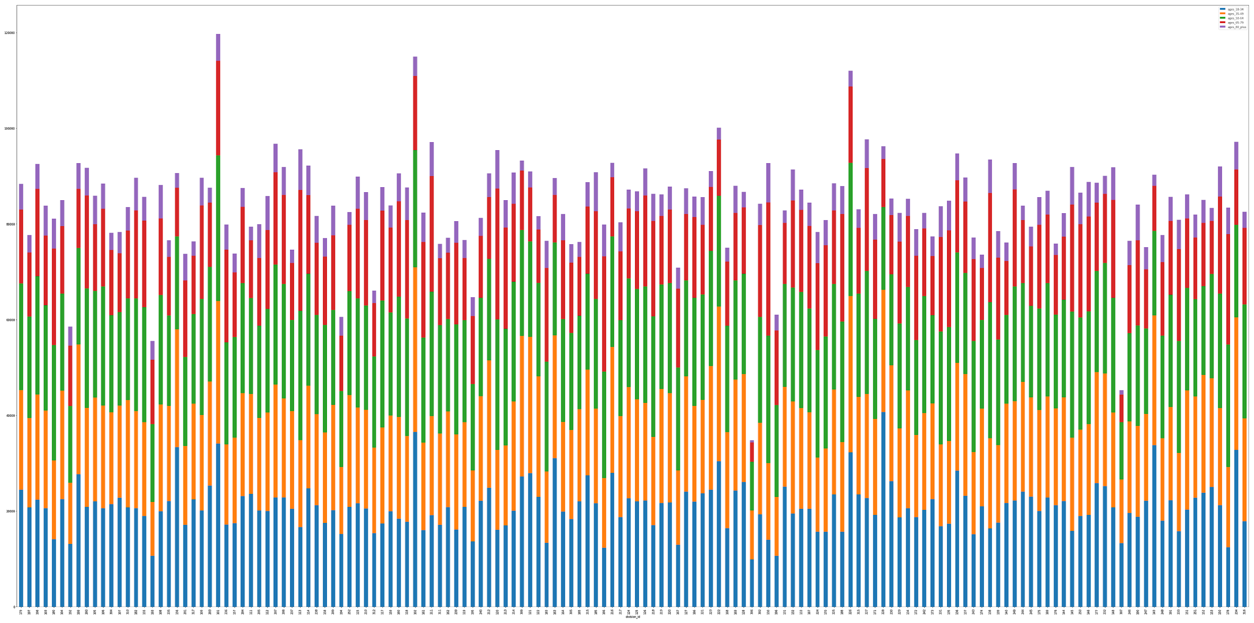
06 - marriage_postal_turnout

A box and whisker plot was created to investigate the turnout percentage of the marriage postal survey which shows that there are potential outliers in this data set.



07 - marriage_postal_participants_by_age

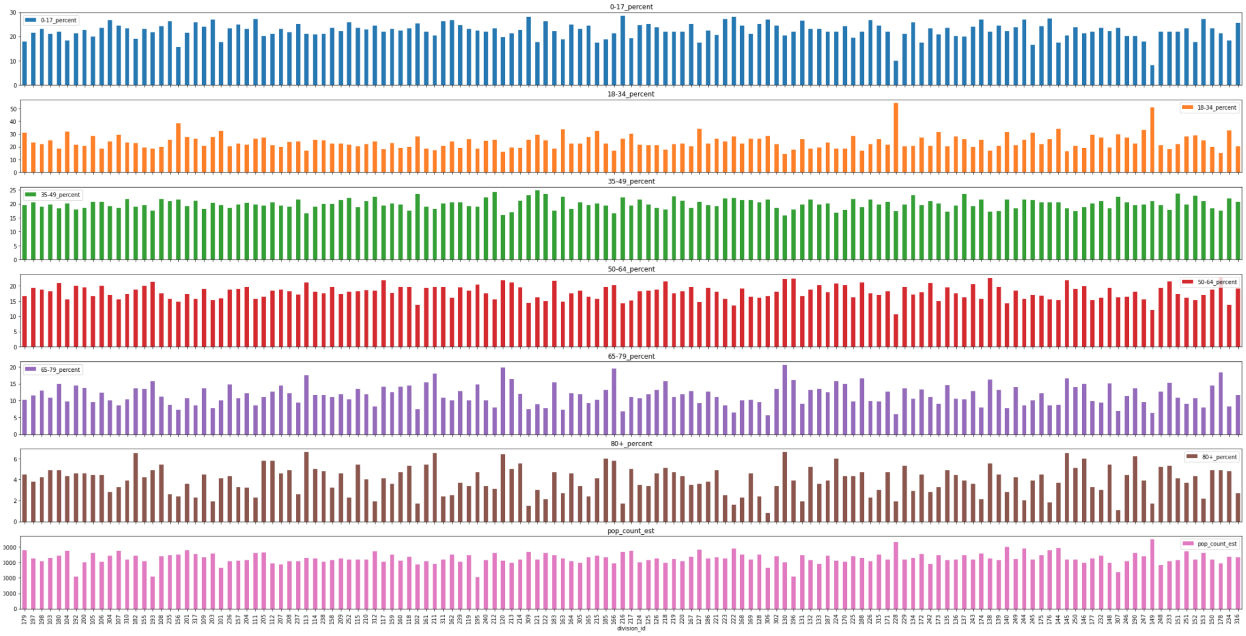
A stacked bar chart was created to show a visual comparison of the number of participants in each age group for all 150 electoral divisions. All age groups are represented across all 150 electoral divisions.



08 - 2017_population_agedemo

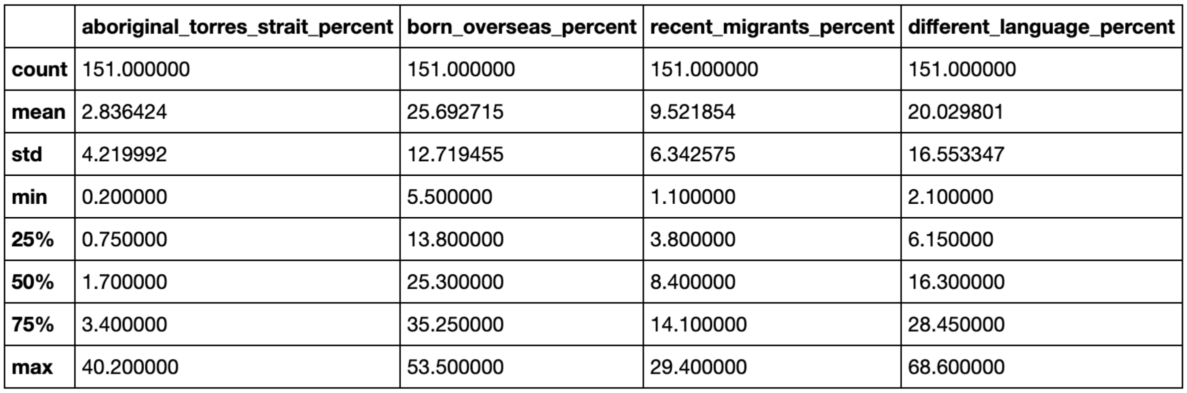
	2017 Estimated Resident Population counts
count	151.000000
mean	162897.536424
std	18777.022083
min	102079.000000
25%	153909.000000
50%	161803.000000
75%	173616.500000
max	223820.000000

A summary statistics table for the age demographic data set was created using the '.describe' function to better understand the data and obtain an insight into some possible trends. Created a subplot for each of the age groups which showed the percentage of people in that respective age group for each division. The visualisation helped determine which age groups contained the most variation and the value of comparing the age groups and divisions.

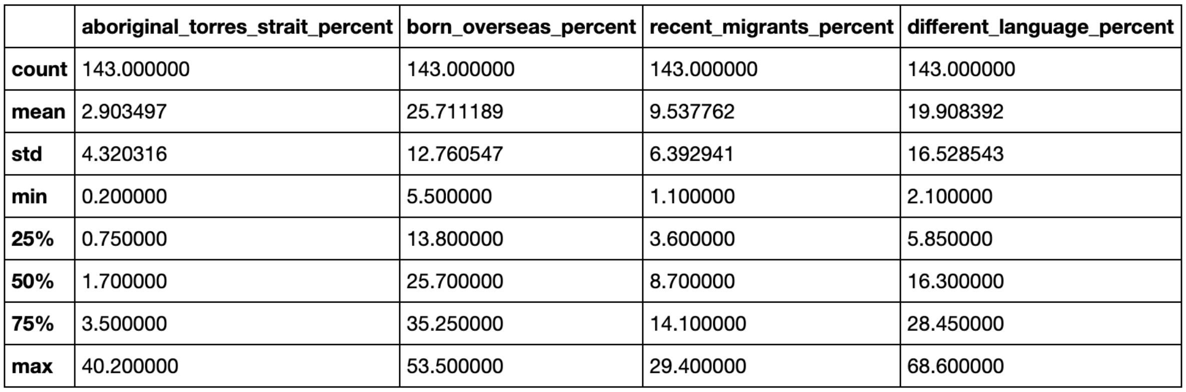


09 - cultural_diversity

Two summary statistics tables were created through the ‘.describe’ function. One was calculated before the cultural diversity data set was merged with the division id data set and one was calculated following the merge. This showed the extent of the statistical differences the inner merge caused as some electoral divisions were not present in the final data set.

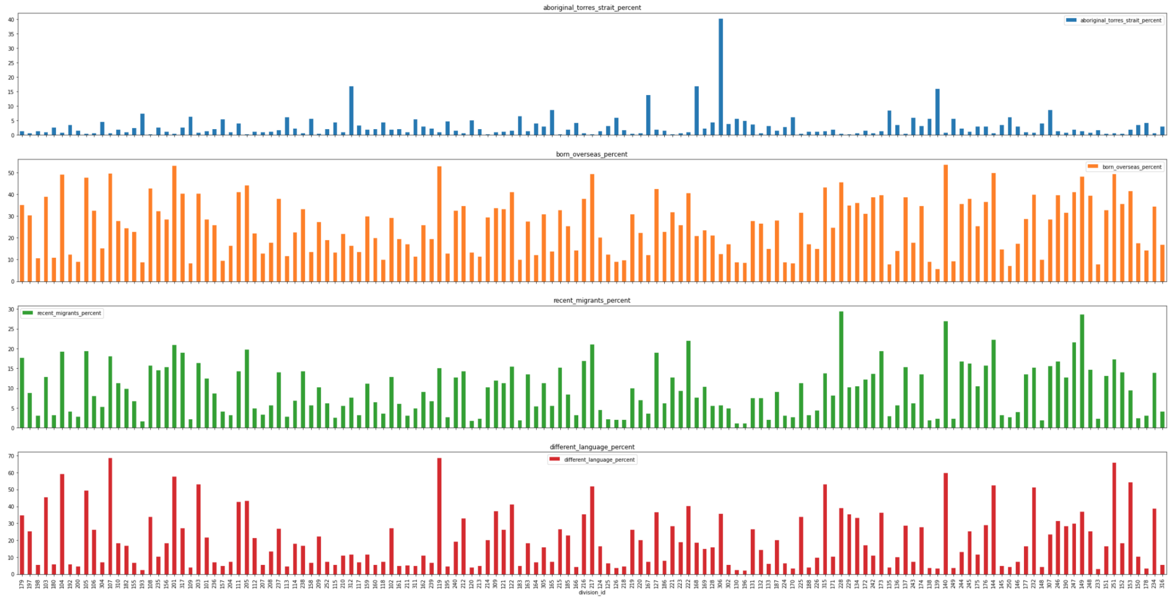


	aboriginal_torres_strait_percent	born_overseas_percent	recent_migrants_percent	different_language_percent
count	151.000000	151.000000	151.000000	151.000000
mean	2.836424	25.692715	9.521854	20.029801
std	4.219992	12.719455	6.342575	16.553347
min	0.200000	5.500000	1.100000	2.100000
25%	0.750000	13.800000	3.800000	6.150000
50%	1.700000	25.300000	8.400000	16.300000
75%	3.400000	35.250000	14.100000	28.450000
max	40.200000	53.500000	29.400000	68.600000



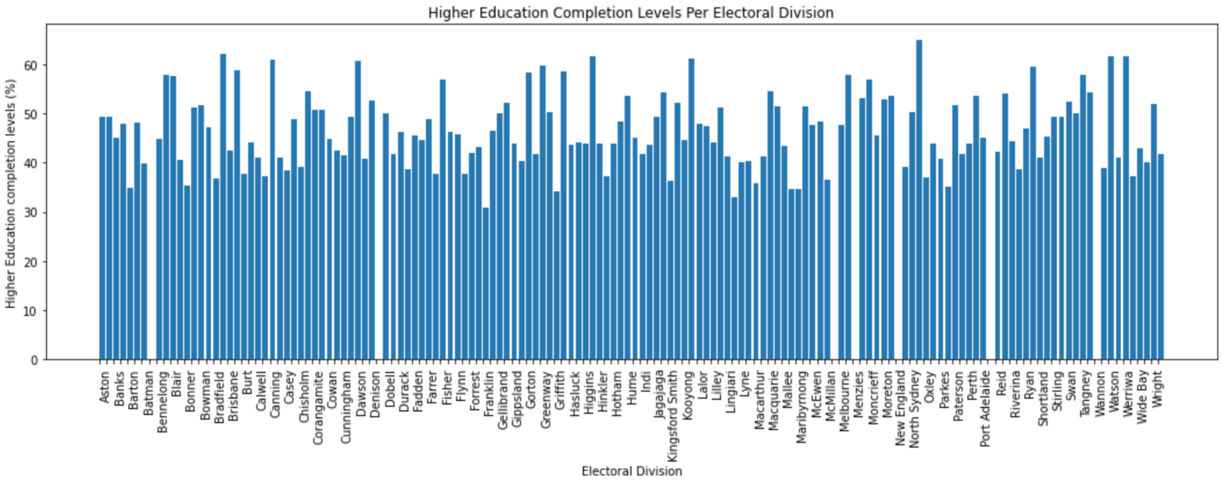
	aboriginal_torres_strait_percent	born_overseas_percent	recent_migrants_percent	different_language_percent
count	143.000000	143.000000	143.000000	143.000000
mean	2.903497	25.711189	9.537762	19.908392
std	4.320316	12.760547	6.392941	16.528543
min	0.200000	5.500000	1.100000	2.100000
25%	0.750000	13.800000	3.600000	5.850000
50%	1.700000	25.700000	8.700000	16.300000
75%	3.500000	35.250000	14.100000	28.450000
max	40.200000	53.500000	29.400000	68.600000

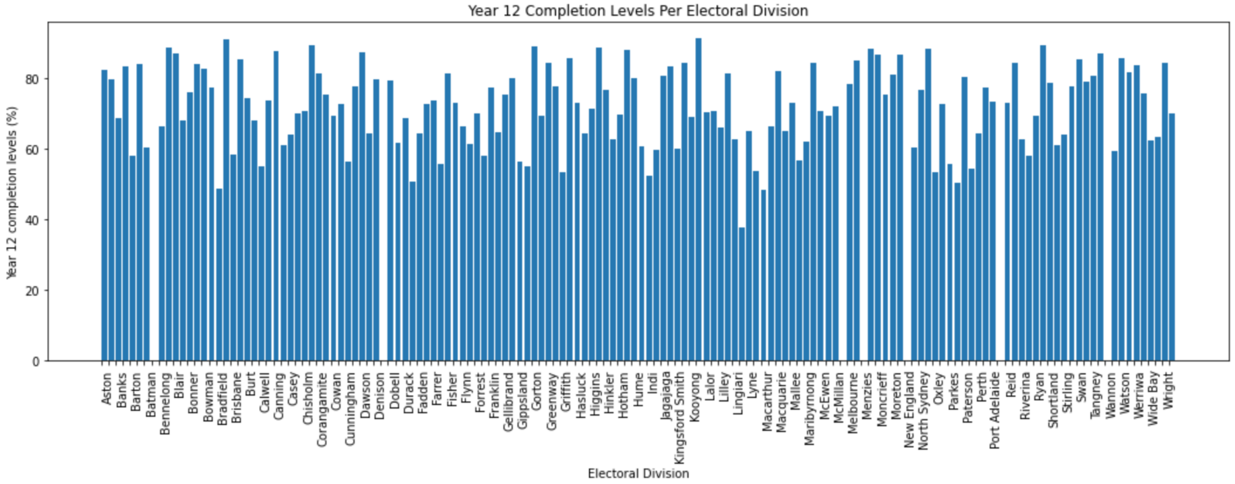
A subplot was created for each of the cultural diversity classifications which showed the percentage of people in that classification for each division. The visualisation helped determine which classifications contained the most variation and the value of comparing the data.



10-education

The creation of bar charts helped to identify NaN values within the dataset. This led to further investigation and exploratory analysis to discover how this had occurred. It was concluded that the merging of two data sets had resulted in the NaN values. It had been assumed that the two columns on which the data frames were being merged were of the same length and contained the same values. However, this assumption was found to be incorrect. The differences between these two datasets and the effect this would have on the analysis then had to be considered. Both data sets contained electoral division information on which the data frames were being merged. However, upon further investigation it was found that data source 5 had its Commonwealth Electoral Division boundaries based on the 2018 redistributions made by the Australian Electoral Commission. All other data sets contained Commonwealth Electoral Division boundaries before these changes had been made. In turn, the data sources contained a differing number of electoral divisions and multiple electoral divisions were specific to individual data sets. Therefore, it was decided that when merging tables from data source 5 and the 01 - electoral_division table an inner merge should be used. This created a merged data frame containing only the common electoral divisions and would not produce NaN values.





Transformations

01 - electoral_division

Aim was to create a table with the 3 columns; division ID, electoral division and state for each of the 150 divisions in Australia. All other tables in the database will also have a primary key of division ID. The purpose is so table 01 can be joined with all other tables in the database to provide the division name and what state they belong to.

- Data source 1, as it included the division ID for each electoral division
- Removed any NaN values, checked length to ensure all 150 divisions still present
- Checked each division ID was unique
- Dropped columns other than the 3 interested in and renamed them
- Set the division ID as the index

02 - election_results

Aim was to create a table that outlined the successful parties for the 2013 and 2016 Federal elections, along with their respective seat status for each electoral division. The table will also include the number of enrolments for each electorate and the demographic.

- Data Source 2 was used for this table
- Dropped 'State' column as this was no longer needed. The state for each Electoral Division is included in the 01-electoral_division dataframe. It is therefore not required in the other tables. If it was required for analysis the tables would only need to be merged with 01-electoral_division for this to be possible.
- Renamed the remaining columns so they are appropriately named for the database.
- Identified and removed NaN values.
- Checked the data types, confirming they were suitable.
- Merged with a secondary dataframe, 01-electoral_division to add an index. The DataFrames were merged using the electoral division name columns.
- Dropped electoral division column following the merge. This column was no longer required in the output as the electoral ID was now being used to identify each row. The names of each electorate would also be included in 01-electoral_division and could therefore be referenced in that table.

03- election_vote_types

To create a table that holds the data for the different vote types for each electoral division from the 2016 Federal Election.

- Data source 1
- Checked the data types
- Removed any NaN values, checked length to ensure all 150 divisions still present
- Checked each division ID was unique
- Reduced the columns to ones of interest and renamed
- Set the division ID as the index

04 - election_turnout

To create a table that holds the total number of enrollments, total number of votes and turnout percentage for each electoral division in the 2016 Federal Election.

- Data source 1
- Checked the data types
- Removed any NaN values, checked length to ensure all 150 divisions still present
- Checked each division ID was unique
- Turnout percentage was already calculated in column ‘TotalPercentage’
- Reduced the columns to ones of interest and renamed
- Rounded turnout percentage to 1 decimal place
- Set the division ID as the index

05 - marriage_postal_results

To create a table that contains marriage postal survey results

- Data source 3
- Merged data frame with table 1, joined on electoral division
- Dropped state and electoral division columns
- Renamed column headers
- Checked division_id was unique and set as index

06 - marriage_postal_turnout

Aim is to create a table that groups the participant information from marriage postal survey by electoral division. Perform aggregations on total number of eligible and participants. Calculate turnout percent and add to table.

- Data source 4
- Converted 2 columns from object to string then integers
- Groupby on electoral division and performed 2 aggregates - sum of eligible and sum of participants.
- Calculated turnout percentage using aggregates in above step, created a new dataframe with the 3 new columns
- Rounded turnout percentage to 1 decimal place
- Merged new dataframe with table 1, joined on electoral division
- Dropped columns didn’t need
- Checked division_id was unique and set as index

07 - marriage_postal_participants_agedemo

To create a table that held the marriage postal participant data stored in the same age groups as the commonwealth electoral data (data source 5 - table 2). Aim was to create a table that showed how many participants there were for each age group in the marriage postal survey

- Data source 4
- Participant data stored as object, converted to a string and then to a integer
- Checked length and dropped any na values

- Created new bins to match the age ranges of data source 5, excluding 0-17 year olds, placed new ‘age groups’ in a new column inside of the dataframe
- Group by of both electoral division and new age groups
- Applied sum aggregate to number of participants for each age group in each division
- Unstacked the age groups and electoral divisions so that age groups became column headers and only one row of data for each electoral division
- Renamed columns
- Merged new dataframe with table 1, joined on electoral division
- Dropped columns didn’t need
- Checked division_id was unique and set as index

08 - 2017_population_agedemo

Aim was to create a table that identified the percentage age distributions for each electoral division. This would be accompanied by a population estimate for each division. Table 2 from data source 5 is used for the percentage age distributions and table 1 from data source 5 is used for the population estimate.

- Read in the percentage age csv
- Checked the columns headers and data types, confirming they were as expected
- Counted the rows (151) and dropped NaN values
- Checked for duplicates
- In order to create a unique identifying division ID for each of the rows, this database was merged with the 01-electoral_division table through an inner join on the electoral division name column
- For completeness, a check was performed for duplicates of divisions and datatypes. No duplicates were present, and as such, no transformation was required.
- Read in the population estimate csv
- Checked the columns headers and data types, confirming they were as expected
- Counted the rows (151) and dropped NaN values
- Checked for duplicates
- Merged the previously merged dataframe (above) and the population dataframe through an inner join on electoral division name.
- Dropped the state and electoral_division columns for purposes of normalisation.
- Renamed columns and changed formatting related to age for purposes of clarity
- Set the index to division_id
- Counted the rows (as discussed above, the number of rows have decreased through the inner join)

09 - cultural_diversity

Aim was to create a table that displayed the percentage of each electoral division that identified as Aboriginal and/or Torres Strait Islander peoples, were born overseas, were a recent migrant (arrived 2006-2016), and spoke a language other than english at home.

- Data source 5, Table 5
- Removed headers / non relevant rows from the file and read in the csv;
- Checked all columns and datatypes. As expected, these were floats (as they reflect percentages) and one object (division name / text value);
- Checked the electoral_division column is unique to ensure there are no duplicates;
- Counted the number of rows (151), then dropped any NaN values. Counted the number of rows again (151)
- Renamed columns for purposes of clarity;
- Multiplied the floating numbers * 100 to show percentages;
- Created a summary statistics table through the “.describe” function.
- In order to create a unique identifying division ID for each of the rows, this database was merged with the 01-electoral_division table. For completeness, a check was performed for duplicates of divisions and datatypes. No duplicates were present, and as such, no transformation was required.
- Merged the cultural diversity database with division_id through an inner join on “electoral division” name.
- For purposes of normalisation, the “state” column and “electoral_division” column were dropped.
- The index was set to “division_id”.

10 - education

Aim was to create a table that displayed the percentage of each electoral division that had completed year 12 and the percentage of each electoral division that had completed higher education.

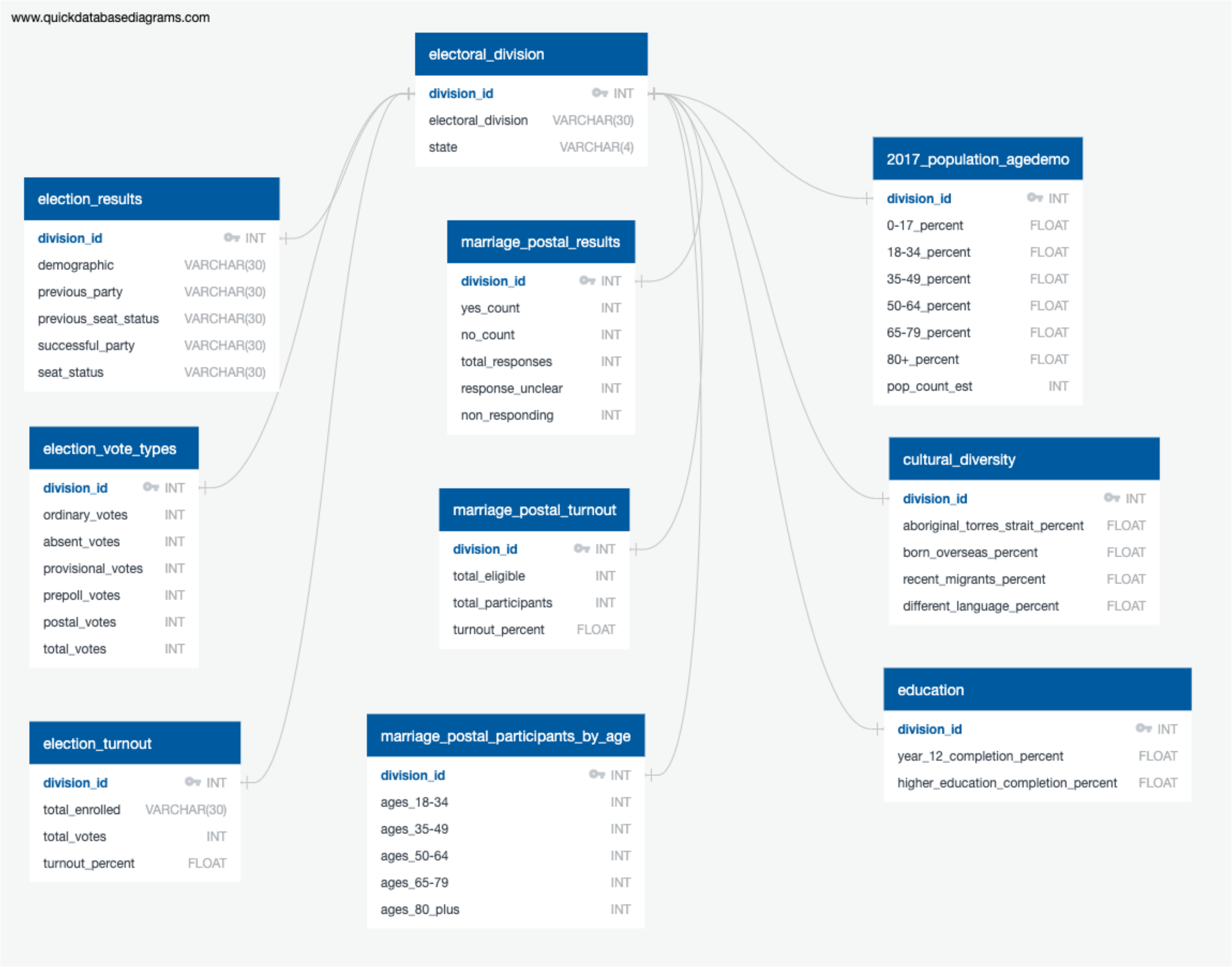
- Table 8 from data source 5 was used for this section
- Removed unwanted blank columns that were created when converting the excel spreadsheet to a CSV file
- Identified and removed NaN values.
- Renamed columns so they were suitable for the database.
- Checked the data types and found that the numerical columns were objects when they needed to be floats. First converted the data from objects to strings. Then converted from strings to floats, removing percentage signs in the process.
- Merged with a secondary dataframe, 01-electoral_division to add an index to the data.
- The two data frames were merged using an inner merge on the electoral division column. As discussed in the transformation section an inner merge was necessary due to the differences in the electoral division column between the data sets.
- Dropped Electoral division column following merge. This column was used for the join but was no longer required in the output as the electoral ID was being used to identify each row. The electoral division names could then be found in the 01-electoral_division table if needed.

Load

A relational database, Postgresql, has been chosen to store the data which has been produced. The structured nature of the tables created lends well to the tabular nature of a relational database. The tables consist of columns and rows where each row represents a single data entry and each column contains specific information. The columns also contain entries of the same data type, which can be specified in the database schema. Each table was developed with an index, meaning each row was uniquely identifiable by this value. This index could then be set as a primary key in the relational database. The primary key of electoral_id allows each row to be identified and also creates a way for the tables to link with each other.

Data normalisation led to the creation of the 01-electoral_division table which contains each electoral ID, the electoral division name for the respective ID and the state for that electorate. This meant that the other nine tables did not require columns for the electoral division name or the state. Therefore, the only column required alongside the information in each table was an electoral ID column. If further information about the electorate name or state was required, the electoral_division table could be referenced.

The following Entity Relationship Diagram (ERD) demonstrates the relationships between these tables:



electoral_division	
division_id	🔑 INT
electoral_division	VARCHAR(30)
state	VARCHAR(4)

election_results	
division_id	🔑 INT
demographic	VARCHAR(30)
previous_party	VARCHAR(30)
previous_seat_status	VARCHAR(30)
successful_party	VARCHAR(30)
seat_status	VARCHAR(30)

election_vote_types	
division_id	🔑 INT
ordinary_votes	INT
absent_votes	INT
provisional_votes	INT
prepoll_votes	INT
postal_votes	INT
total_votes	INT

election_turnout	
division_id	🔑 INT
total_enrolled	VARCHAR(30)
total_votes	INT
turnout_percent	FLOAT

marriage_postal_results	
division_id	🔑 INT
yes_count	INT
no_count	INT
total_responses	INT
response_unclear	INT
non_responding	INT

marriage_postal_turnout	
division_id	🔑 INT
total_eligible	INT
total_participants	INT
turnout_percent	FLOAT

marriage_postal_participants_by_age	
division_id	🔑 INT
ages_18-34	INT
ages_35-49	INT
ages_50-64	INT
ages_65-79	INT
ages_80_plus	INT

2017_population_agedemo	
division_id	🔑 INT
0-17_percent	FLOAT
18-34_percent	FLOAT
35-49_percent	FLOAT
50-64_percent	FLOAT
65-79_percent	FLOAT
80+_percent	FLOAT
pop_count_est	INT

cultural_diversity	
division_id	🔑 INT
aboriginal_torres_strait_percent	FLOAT
born_overseas_percent	FLOAT
recent_migrants_percent	FLOAT
different_language_percent	FLOAT

education	
division_id	🔑 INT
year_12_completion_percent	FLOAT
higher_education_completion_percent	FLOAT

Examples of Analysis

Possible avenues to consider for analysis include:

- Considering if there is a relationship between the elected party in each federal electorate and the outcome of each electorate's vote in the Marriage Law Postal Survey.
- Extracting the number of postal votes for each electoral division in the Federal Election and joining with the Marriage Law Postal Survey. This may identify certain electorates that have a higher preference for postal votes and those that do not.
- Analysing whether electorates with similar percentage values for the cultural diversity columns voted in a similar way in both the Federal Election and the Marriage Law Postal Survey.
- Considering if there is a link between higher education completion in electorates and the outcome of the Federal Election and Marriage Law Postal Survey results.
- Considering whether there is a relationship between age demographics and the Federal Election and Marriage Law Postal Survey results.
- Calculating how many people participated in the Marriage Law Postal Survey for each age group as a comparison to how many people make up that age group in their electorate.
- Determining the number of people who were eligible for the Marriage Law Postal Survey and those who participated. Were people of certain electorates or characteristics more likely to participate?