COMP20008 Assignment 2: Data Science Project
Group Report
October 15, 2021

The University of Melbourne
Members: Sihan Chen; Yitong Kong; Zheqi Shen; Junfei Zhang
Git: https://github.com/COMP20008/assignment-2-comp20008-assignment-2-group-65.git

Research Question

Is the inequality in living status faced by Aboriginals correlated with the socio-economic levels in different regions of Victoria? [Liveability / Inclusiveness]

Abstract

This project focuses on the inequality in living status faced by Aboriginal communities in Victoria, investigating whether the situation is related to socio-economic level; if so, which aspects of their living are associated more severely.

The inequality is measured based on a population census from PHIDU, which includes the data of Indigenous and Non-Indigenous communities[1], including low-income families, unemployment rate, primary school dropout rate[2], middle school participation rate[3], internet access, and social engagement rate[4]. Socio-economic level is referring to the Socio-Economic Indexes for Areas (SEIFA), which includes relative socio-economic disadvantage, relative socio-economic advantage and disadvantage, economic resources, and education and occupation[5].

This project is related to the theme of liveability and inclusiveness as we aim to compare the current living standards of the Indigenous people with the rest of society to get an understanding of the inequality. The outcome would imply whether the current policies and welfare given to Aboriginals are sufficient and efficient.

Data

Two datasets are used in this project:

1. *PHIDU - INDIGENOUS STATUS COMPARISON: SOCIAL HEALTH ATLAS OF AUSTRALIA*
   - Format: XLS/XLSM/XLSX
   - Size: each sheet has 457 rows, 47 datasheets in total
   - Source: phidu.torrens.edu.au

[1] Detailed information please refer to https://phidu.torrens.edu.au/social-health-atlases/data
[2] People who left school under age 10
[3] Full time schooled students at age 16
[4] Population either learning or earning
[5] Detailed information please refer to https://www.abs.gov.au/websitedbs/censushome.nsf/home/seifa

- Granularity: Indigenous Areas (IARE)[6]
- Information: the statistics about the current living situation, such as medical care and education level, of Aboriginals and Non-Indigenous people according to regions
- Legality: open source; legally obtained from PHIDU website

2. *ABS - Socio-Economic Indexes for Areas (SEIFA), Australia, 2016*
- Format: XLS/XLSM/XLSX
- Size: each sheet is around 55,000 rows, 6 datasheets in total
- Source: portal.aurin.org.au
- Granularity: Statistical Area Level 1 (SA1)[7]
- Information: indices evaluating the socio-economic level of a certain region
- Legality: open source; legally obtained from AURIN repository

The PHIDU dataset is used to compute the level of inequality in living conditions of the two communities; combining with the ABS dataset, the correlation between the inequality and the socio-economic status among regions in Victoria can be calculated.

Methodology

Index Linkage

code: index_linkage.py
output: data/index_linkage.csv

The two datasets use different geography standards (IARE and SA1). For maintaining higher accuracy, the final result will be presented in terms of IARE, which has a larger granularity. To link the two datasets together, an index linkage is created. By using the `openpyxl`, `pandas`, and `re` python libraries, the IARE codes starting with 2[8], IARE names, and tokenized IARE names are read to a dataframe. By using the 'Victoria' section of *Table 7*[9] from the ABS dataset, three attributes—7-digit SA1 code, corresponding LGA name, and corresponding SA2 name—are extracted. While tokenizing the LGA names, a similarity value of the LGA names and IARE names are obtained using the Sørensen-dice similarity formula:$Sim_{dice}(S_1, S_2) = \frac{2\times|S_1 \cap S_2|}{|S_1| + |S_2|}$[10]. The LGA names with the largest similarity value that is greater than 0.6[11] are considered to be a match and written to the dataframe. LGA names are primarily compared; SA2 names will be compared if there is no mapping by using the LGA names. The data frame is then exported as a CSV file, serving as a linkage mapping multiple entries of the ABS data with a single PHIDU entry. This linkage is name-based and is conceptual, as we aim to investigate the inequality based on community instead of geography[12].

---

[6] IARE is an Australian Statistical Geography Standard (ASGS) medium sized structure with approximately 540 regions in Australia.
[7] SA1 is an ASGS main structure, with approximately 57523 regions in Australia.
[8] IARE codes starting with 2 represent regions in Victoria.
[9] Contains the names and codes of various geographic structures.
[10] S1 and S2 are the LGA tokens and IARE tokens respectively.
[11] The names are considered possibly similar if they are at least 60% the same.
[12] Explained in the limitations section.

Data Cleansing

code: indigenous_cleanse.py, abs_data_integrate.py
output: data/indigenous_cleanse.csv, data/integrate_abs_data_raw.csv

Data from the two datasets are cleansed to two CSV files. Taking the "iare_code" column from the index linkage as keys, selected attributes from each worksheet of the PHIDU dataset, including low-income families, unemployment rate, primary school dropout rate, middle school participation rate, household internet access, and social engagement rate, are extracted into dataframes. The ABS file is also processed according to the index linkage: for each IARE code region, a weighted average of socio-economic index of the corresponding SA1 codes is calculated. For each set of SA1 codes, the socio-economic index categories, including the index of relative socio-economic disadvantage, index of relative socio-economic advantage and disadvantage, index of resources score, and index of education and occupation score, are summed and weighted according to the population ratio of each region.

Missing Value Imputation

code: missing_value.py
output: data/integrate_abs_data.csv

There are several missing values in the integrated file from the previous step; To deal with this, An imputer is used to impute the missing value with the mean value of the remaining data in that column.

Calculating Equality/Inequality Index

code: inequality_index.py
output: data/inequality_index_unormalised.csv

For the attributes that are directly proportional to level of equality, such as education and internet access, the formula $\frac{percentage\ of\ Indigenous\ people\ involved}{percentage\ of\ Non-Indigenous\ people\ involved}$[13] is used to calculate the equality index. For attributes that are inversely proportional to equality level, including unemployment and low-income families, an inequality index is calculated based on the same formula. Two different indexes are calculated instead of a single inequality index since there is no evidence to show that those two qualities are binary and opposite[14].

Normalisation of Equality Index

code: inequality_index.py
output: data/inequality_index.csv

The equality/inequality indices of different aspects vary in scope. For instance, "Unemployment inequality index" varies between 1 and 4 while "Internet access equality index" varies between 0 and 1. This may cause a loss of precision when visualising the result,

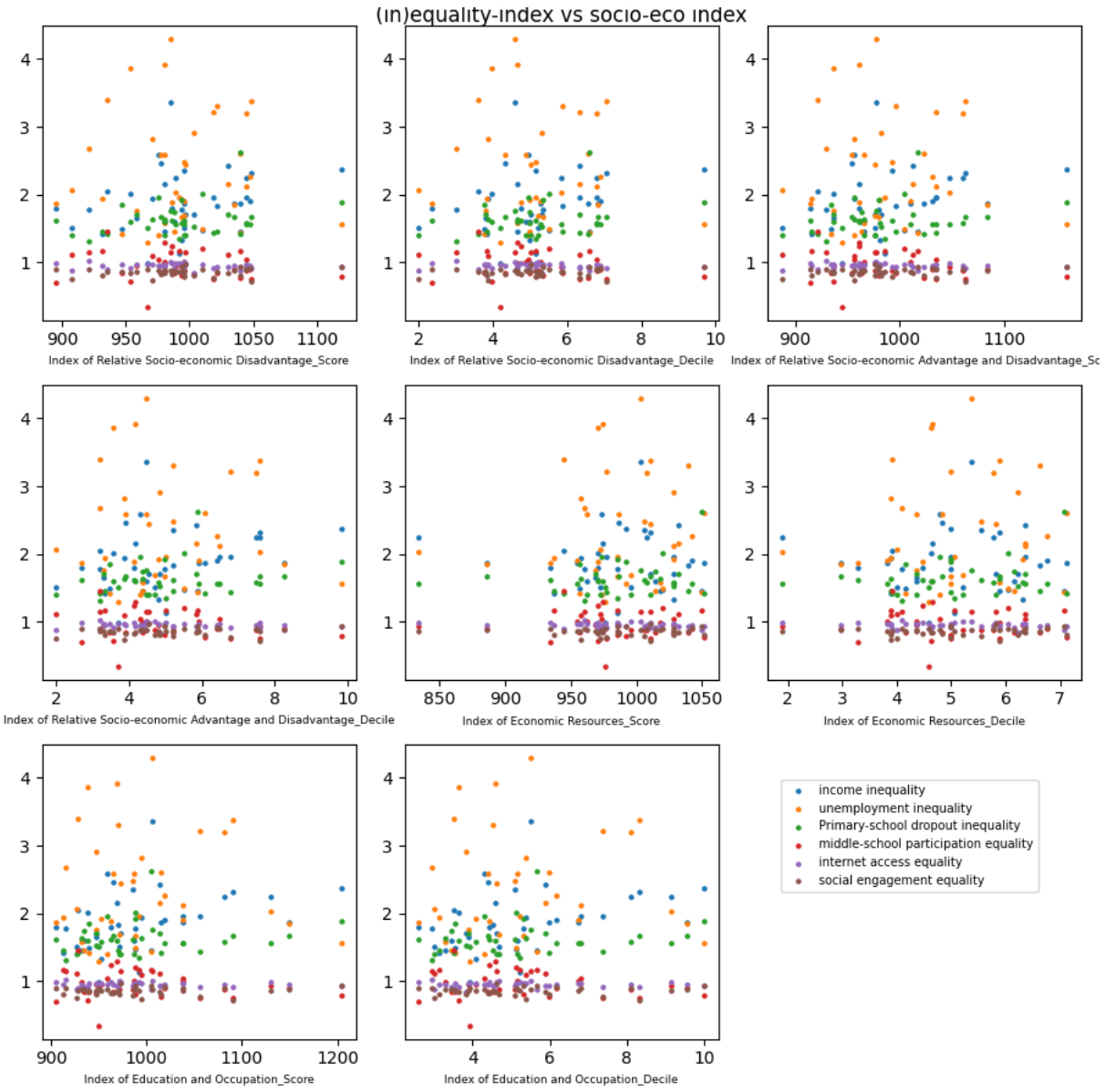
[13] As we aim to compare the inequality status rather than direct number.
[14] Explained in limitations section.

since the one with smaller scope will be squeezed into a line. Hence, each index is normalised using `numpy` to ensure they have the same scope.

Results & Analysis

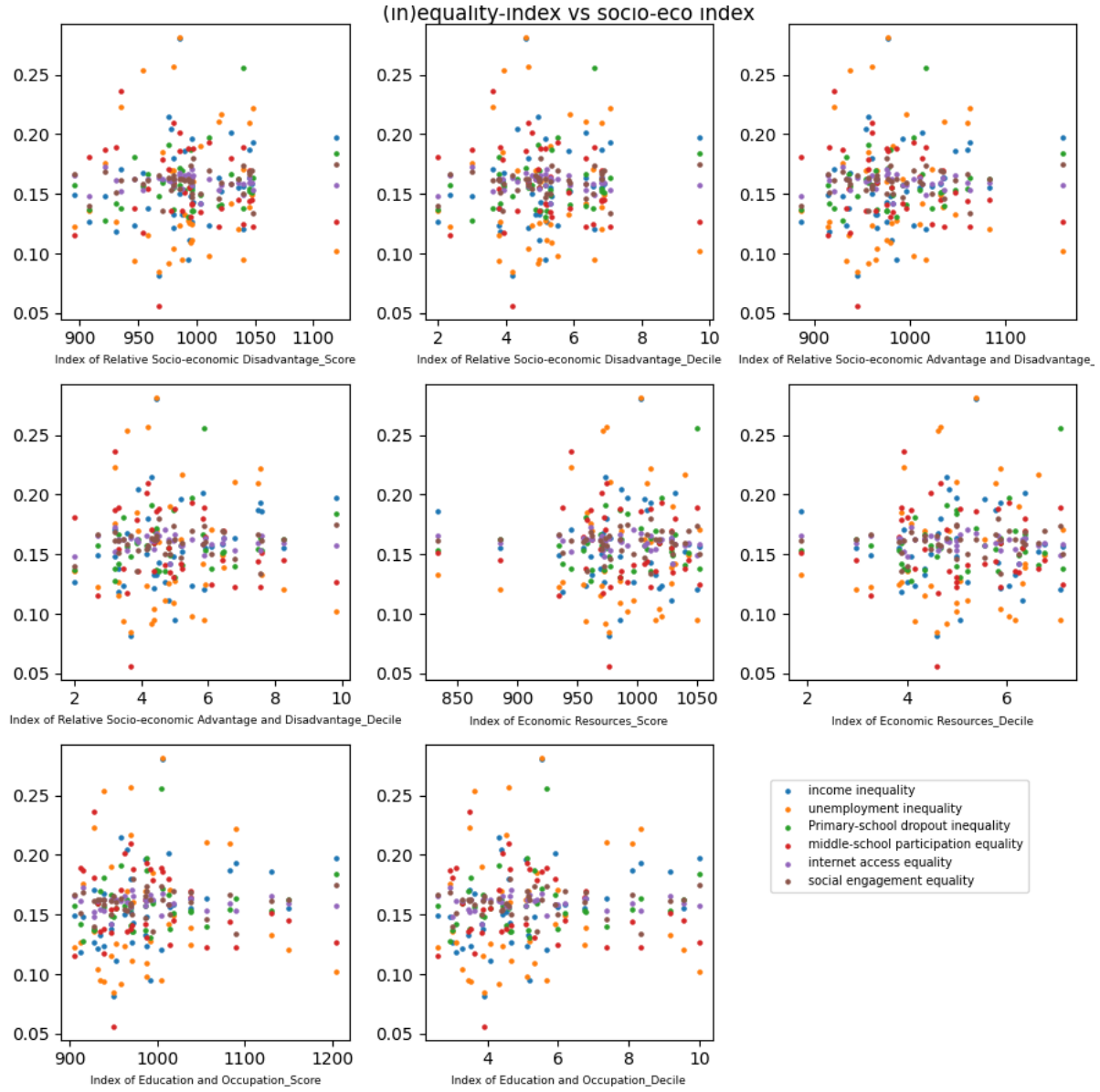
Visualisation: Scatter
code: scatter_plot.py
output: data/scatter(unormalised).png, data/scatter.png

Since this research investigates correlation, a scatter plot is created by using `pandas` and `matplotlib` libraries, with x-axis being the socio-economic index and their weighted index, and y-axis being the six aspects of equality/inequality indexes.



As this plot of unnormalised data shows, inequality is most severe in unemployment rate. For areas that have a moderate socio-economic status, the unemployment rate of indigenous

people is approximately 2 to 3 times higher than that of non-indigenous people. For some areas, this ratio even reached 4. The income inequality index also follows a similar pattern. Primary-school dropout inequality index fluctuates between 1 and 2, meaning that indigenous community has a higher percentage of people who did not finish primary school. The three equality indexes—middle-school participation, internet access, and social engagement—remain steady around 1. This means that the two communities are respectively equal in those three aspects.



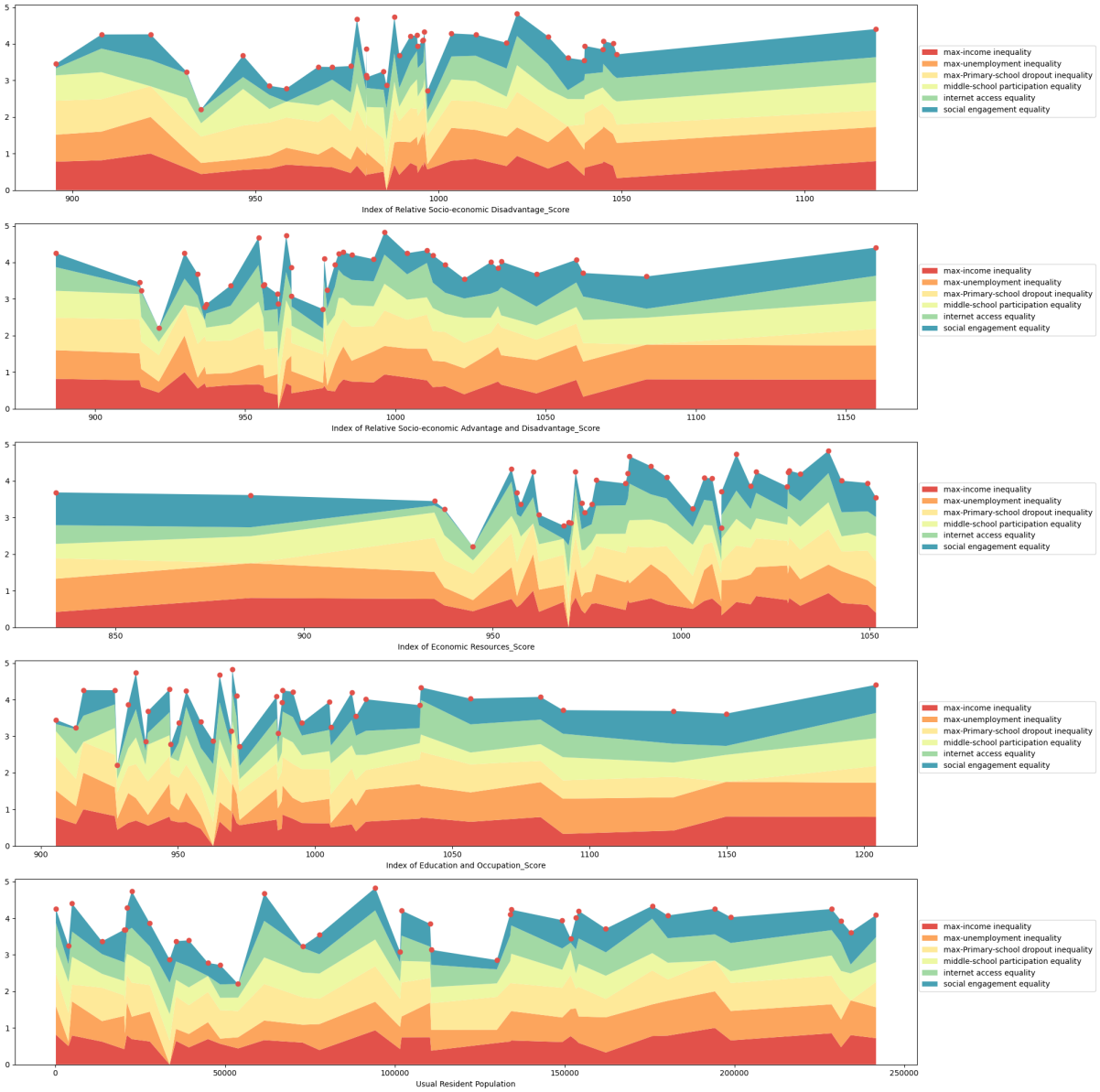
As shown by this scatter plot of normalised data, no obvious trends are observed. The data of internet access and social engagement equality indices are relatively vertically condensed whereas the rest are relatively vertically spread out. Horizontal outliers are not removed, since the socio-economic indexes are already weighted according to population. However, the vertical outliers must be analysed more carefully. There are two obvious extraordinarily high data points on each subplot—income and unemployment inequality indices of "Warrnambool". This means that the percentage of indigenous people that have low income

or are unemployed in Warrnambool is relatively high. Another possible outlier is the middle-school participation equality index of Brimbank. These outliers may be because those two areas are both developing areas with a large branch of indigenous community, and policies regarding welfare are still being implemented.

Visualisation: Stacked-Area Plot

code: stacked_area_plot.py

output: data/stacked_area_eq.png, data/stacked_area_sc.png



After uniformalised the data, two stacked plots are created to show the integrated trend. The axes are set such that higher x is referring to higher equality[15]. As can be seen from the plot

[15] This inconsistency is explained in limitations section.

concerning the index of education and occupation, after a critical value of around 1030, the equality indices remain relatively stable. Compared to other socio-economic indices, the index of education and occupation has the most data points which are less fluctuated. This is possibly implying that the index of education and occupation score has a high potential of having relation to equality indices.

As illustrated by the plot concerning primary-school dropout, after a critical value of around 0.06, socio-economic indices perform an overall decreasing trend. Moreover, the plot concerning social engagement shows that before a value at around 0.147, social-economic indices show an increasing trend. Thus, it is more likely that there are certain relations between socio-economic indices with the equality in social engagement and middle-school participation.

Hypothesis Test

A two-tailed hypothesis test is conducted:

Null hypothesis H_0: $\rho = 0$

- There is no association between inequality/equality index and socio-economic index.

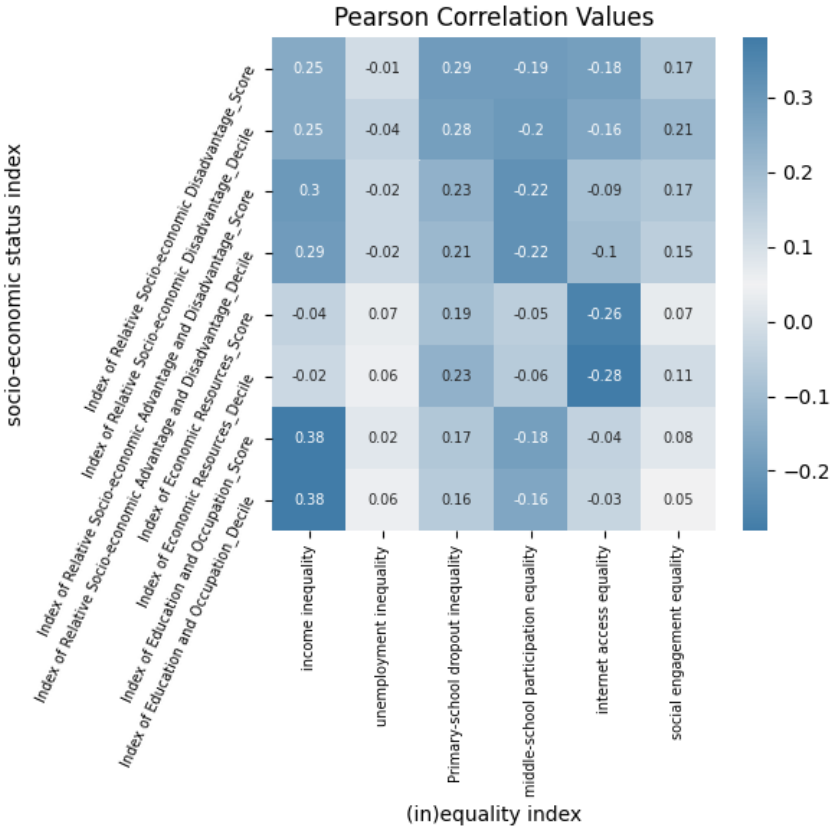
Alternative hypothesis H_A: $\rho \neq 0$

- There is an association between inequality/equality index and socio-economic index.

A significant level of $\alpha = 0.05$ is chosen. The degree of freedom for both pearson and spearman's correlation is $df = n - 2 = 39 - 2 = 37$. According to the table of critical values for pearson correlation coefficient[16], we can reject the null hypothesis if $|\rho| \geq 0.28$. Regarding spearman's correlation, we can reject the null hypothesis if $|\rho| \geq 0.33$[17].

Pearson Correlation

code: pearson_correlation.py
output: data/pearson_heatmap.png



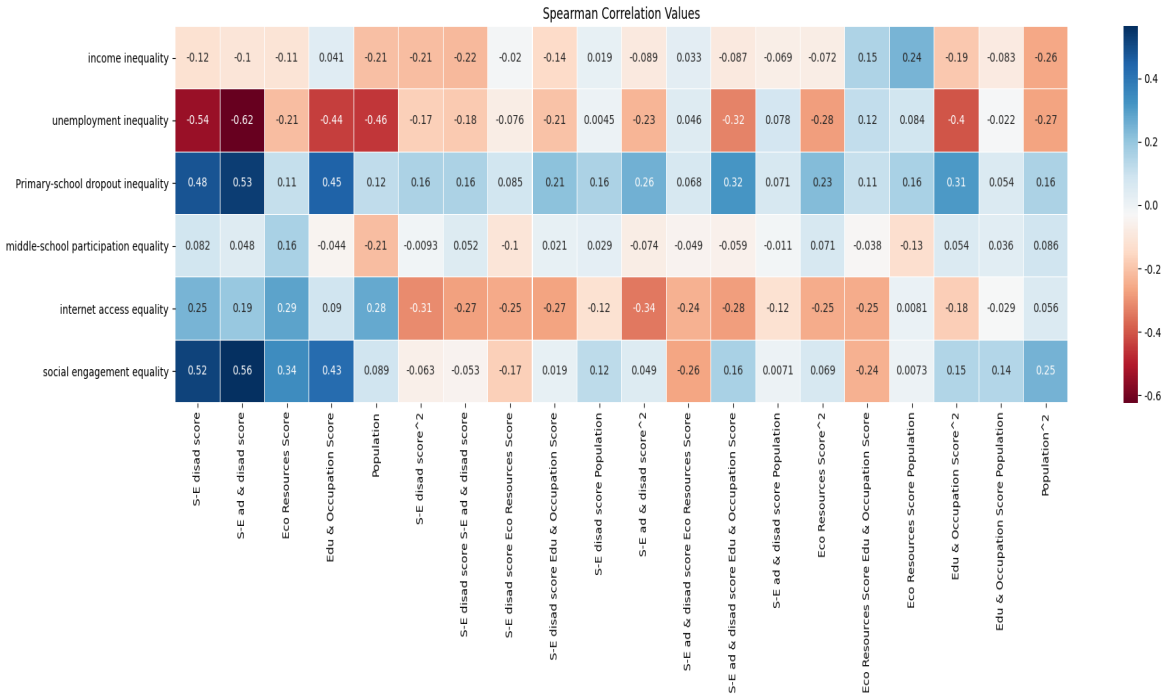
[16] Here we use the critical value approach. Table refer to https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118342978.app2.
[17] Table refer to http://webspace.ship.edu/pgmarr/geo441/tables/spearman%20ranked%20correlation%20table.pdf.

Pearson correlation is calculated by using the `scipy.stats` library[18]. The null hypothesis is rejected between these 4 pairs of indices which have an absolute correlation value exceed the critical value: income inequality index and index of relative socio-economic advantage and disadvantage score (0.29) and index of education and occupation score (0.38), primary-school dropout inequality index and index of relative socio-economic disadvantage score (0.28), and internet access equality index and index of economic resources score (-0.28). There is evidence to believe that those indices are likely to be associated.

Spearman's Correlation

code: spearman_correlation.py

output: data/spearman_heatmap.png



Spearman's correlation value is also calculated.

The unemployment inequality has 5 values exceeding the critical value. Thus, the null hypothesis is rejected for socio-economic disadvantage score (-0.54), socio-economic advantage and disadvantage score (-0.62), education and occupation score (-0.44), population (-0.46), and the squared education and occupation score (-0.4).

Social engagement equality index also illustrates a relatively high absolute value. The null hypothesis is rejected for socio-economic disadvantage score (0.52), socio-economic

[18] The indices with 'Decile' represent weighted indices.

advantage and disadvantage score (0.56), economic and resources score (0.34), and education and occupation score (0.43). This conclusion is consistent with the analysis of the stacked plot.

Regarding primary-school dropout inequality index, the null hypothesis is rejected for socio-economic disadvantage score (0.48), socio-economic advantage and disadvantage score (0.53), and education and occupation score (0.45).

In terms of internet access, the null hypothesis is rejected for the squared value of socio-economic advantage and disadvantage score ($r = -0.34$). This is also consistent with the implication of the stacked plot.

Regarding all of the socio-economic indices, the socio-economic disadvantage score, socio-economic advantage score and disadvantage score, and education and occupation score have relatively higher correlation values with inequality/equality indices.

Conclusion

Regarding the current status among the two communities in Victoria, inequality is most severe in the income and unemployment status, with indigenous communities having a higher unemployment rate and lower-income than the non-indigenous community in most areas. Indigenous communities are also at a disadvantage in the primary school dropout rate. Equality is relatively achieved in aspects including middle-school participants, internet access, and social engagement. This shows that the current policies are not sufficient for maintaining equality in employment related aspects and primary education.

Secondly, regarding the correlation, there is evidence to believe that the inequality in the percentage of low-income families, primary-school dropouts, and internet access is likely to be positively associated with the level of socio-economic disadvantage. This is implying that for areas that perform well at socio-economic level, indigenous communities are likely to face a larger inequality in income, primary and secondary schooling, and internet access. Hence, when an area is under rapid development and is achieving higher socio-economic indices, policies regarding those aspects should be considered to incline more towards indigenous community. On the other hand, there is evidence to believe that the inequality in unemployment rate is negatively associated with socio-economic level, and the equality in social engagement is positively associated with socio-economic level. This means that policies are relatively efficient in those two aspects.

Combining all the conclusions, current policy which aims to enhance equality between the two communities is suggested to focus on employment related aspects and primary education in the earlier stage, while continuing to pay attention to inequality in income, primary and secondary schooling, and internet access.

Significance

The key stakeholders of the results are the Victorian Government and the Aboriginals in Victoria. As no direct data is showing the correlation between the inequality in various aspects of living status faced by Aboriginals and the socio-economic level, our data will provide innovative information that will help the government decide which aspects should the policies be made more inclined towards Aboriginals in the long term.

Limitations & Improvements

This project did poorly in converting the geographical granularity of the two datasets. The index linkage is based on the assumption that the region is consistent with their names. However, this may decrease the accuracy since the conceptual naming may differ between different systems. Hence, it is better to choose datasets that use the same system or to introduce a better method of linkage.

Problems could also arise due to treating inequality and equality indices separately; this may cause inconsistency in analysation. A better method would be to convert them to a single index under the same standard.

Acknowledgment

"Australian Statistical Geography Standard (ASGS): Volume 3 - Non ABS Structures, July 2016 ." *Australian Bureau of Statistics*, https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by Subject/1270.0.55.003~July 2016~Main Features~State Suburbs (SSC)~9.

"Indigenous Structure." *Australian Bureau of Statistics*, https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/indigenous-structure.

"Socio-Economic Indexes for Areas." *Australian Bureau of Statistics*, https://www.abs.gov.au/websitedbs/censushome.nsf/home/seifa.

"Appendix B Statistical Tables." *Onlinelibrary.Wiley*, https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118342978.app2.

"Critical Values of the Spearman's Ranked Correlation Coefficient." *Webspace.ship.edu*, http://webspace.ship.edu/pgmarr/geo441/tables/spearman%20ranked%20correlation%20table.pdf.