How can we combat income disparity in Victoria?
Group W12G02

Jessica Le	Hanyang (Martin) Ma	Wenxuan Qiu	Inesh Pinjani
1353984	1403199	1313864	1424652
jele3@student.unimelb.edu.au	hanyangm@student.unimelb.edu.au	wenxuanq@student.unimelb.edu.au	

Executive Summary

This report analyses the key socioeconomic factors influencing low personal income in communities across Victoria. The primary objective is to identify and examine these factors to provide actionable insights that can inform policy decisions. Correlation and regression analysis explored the relationships between various features such as unemployment, English proficiency, access to education, and gambling losses, with the proportion of people earning less than $400 per week. Preprocessing techniques were used to clean and transform the data before employing supervised data modelling to predict future changes.

Overall, the findings suggests that educational attainment was the most significant factor. Notably, the regression model predicts a 3% decrease in the proportion of low-income individuals when the percentage of degree holders increases by 20%, and individuals not completing Year 12 decreases by 20%. These results stipulate that targeted interventions aimed at improving overall education levels can effectively alleviate income disparity across Victoria. Initiatives that enhance access to higher education and vocational training are recommended to combat high proportions of low-income earners. Ultimately, improving educational outcomes benefits not only the individual but also economic health and stability across Australia. It is imperative that policies prioritizing a general literacy increase be implemented.

Introduction

The basis of this report is formed around the prevalence of low-income earning individuals in Victorian communities. According to a study conducted by The University of Melbourne in 2012, a working individual earning less than $470/week or an unemployed individual receiving less than $380/week would be considered living under the poverty line. A quick exploration into data collected from Victorian communities shows that the mean proportion of individuals earning less than $400/week was 39.56%. This initial figure was alarming and therefore set the basis of the report's investigation onto how to effectively mitigate this issue. Moreover, a study conducted in 2022 suggests that at least 1 out of 8 Australians live below the poverty line (Davidson, Bradbury, & Wong, 2022), indicating that the issue of low income has persisted throughout the last decade. Such statistics underline the urgency and need for addressing income disparity across Victorian communities.

This research serves as a guide for targeted policy reforms and promotes greater societal equity. By employing a combination of correlation analysis and various regression models, the relationships between these socioeconomic features and low personal income levels are made clearer. Understanding these factors in depth is crucial for identifying high-impact interventions that can address income inequality.

Methodology

To explore the factors that potentially contribute to the prevalence of low-income earners, several methods, techniques, and tools were used to prepare, analyse and interpret the data. First, factors which were suggested by previous studies to have an impact on income were modelled on a scatterplot to better visualise the correlation in Victorian communities. After, the datasets were cleaned by addressing missing or null values in numeric columns and replaced by their respective median values to avoid skewed data due to outliers. A correlation heatmap was then generated to identify potential relationships and their strengths. The highest correlated factors were selected for further analysis.

To model predictions on the contributors of income disparity, multiple regression techniques were employed, including linear regression, polynomial regression, and gradient boosting. Each model's performance was evaluated using Mean Squared Error (MSE) and R-squared scores (R^2) alongside cross validation as the primary metrics to assess accuracy and precision. Among the three models created and tested, linear regression had the lowest MSE suggesting that it performed the best. Finally, this model was used to predict potential changes in the proportion of low-income earners when different variables were increased or decreased.

Data Exploration and Preprocessing

To ensure the dataset was suitable for analysis, several preprocessing techniques were carried out to clean and transform the data to prepare for modelling. These steps were used to improve the quality and accuracy of the results obtained from the dataset. Based on the concept discussed in the National Center for Education Statistics' paper, "Annual earnings by educational attainment," a scatterplot was created to visualize the relationship between education and income levels, specifically the proportion of people who have not completed year 12 and the proportion of people who earn less than $400/week. The columns of interest were converted to numeric type (integers or floats) as they existed as strings in their original csv file. The function pd.to_numeric() will always replace columns containing non-numeric characters as null, NaN. This is crucial as the following steps involve numerical operations. For this graph, potential outliers were kept getting a better visualization of the distribution.

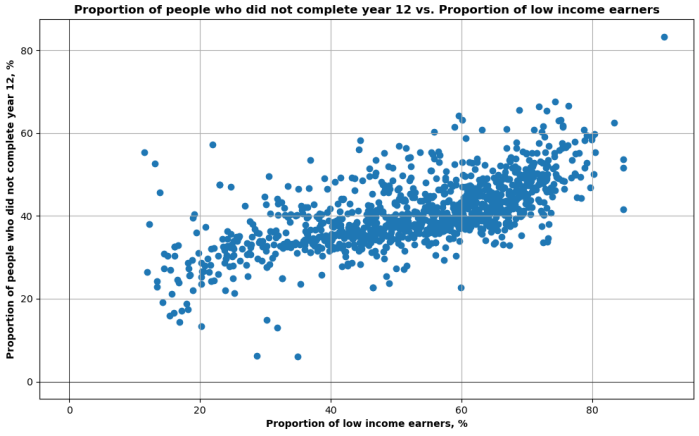

Figure 1: Scatterplot depicting the relationship between the proportion of people who did not complete year 12 versus the proportion of people who earn less than $400/week, categorised in this study as low-income earners.

A heatmap was then created to identify relationships between a broader range of variables such as crime rates, losses from gambling, location, and English proficiency. The heatmap provides a visual representation of strength and direction of relationships and a correlation value. To generate this heatmap, a combination of preprocessing techniques was utilised, including:

- removing outliers in each relevant column as extreme values could skew and distort the results of future analysis. They were identified using the Interquartile Range method.
- handling any missing or null values to lead to a more reliable imputation. Mean imputation was conducted since the data is continuous numerical and free from outliers, meaning mode and median imputation were would not as accurately maintain the trends in the data.
- converting non-numeric data (mostly column headings) to lowercase. Since the heatmap was generated from several different datasets, formatting errors were imminent. For example, the "City Of Melbourne" exists in one document whereas another document referred to this as "City of Melbourne." To ensure that equivalent data was being recognized correctly, everything was converted to lowercase to avoid mismatches due to case differences.
- Aggregating like data. There was two ways in which aggregation was used during the preprocessing stage. First, to create a new column in the relevant data frame which combines the number of all school types (e.g. primary, secondary TAFE) into a total value. This helps simplify the analysis. Another example of aggregation was to group the total number of offences by their local government area instead of the offence type.
- manually cleaning and replacing names which are equivalent. For example, a local government area is commonly referred to be both "Merribek" and "Moreland;" "Merribek" now exists as "Moreland."
- using regular expression as a tool to find a desired phrase. The document containing information of gambling also focuses on local government areas, however, have named their rows slightly differently. To combat this, regex was used to drop prefixes and suffixes such as "City of" and "Shire of" and therefore be matched with the LGA's in the other documents. Regex was also used to extract only the numeric data from "location" which contained data such as "3km from Melbourne".

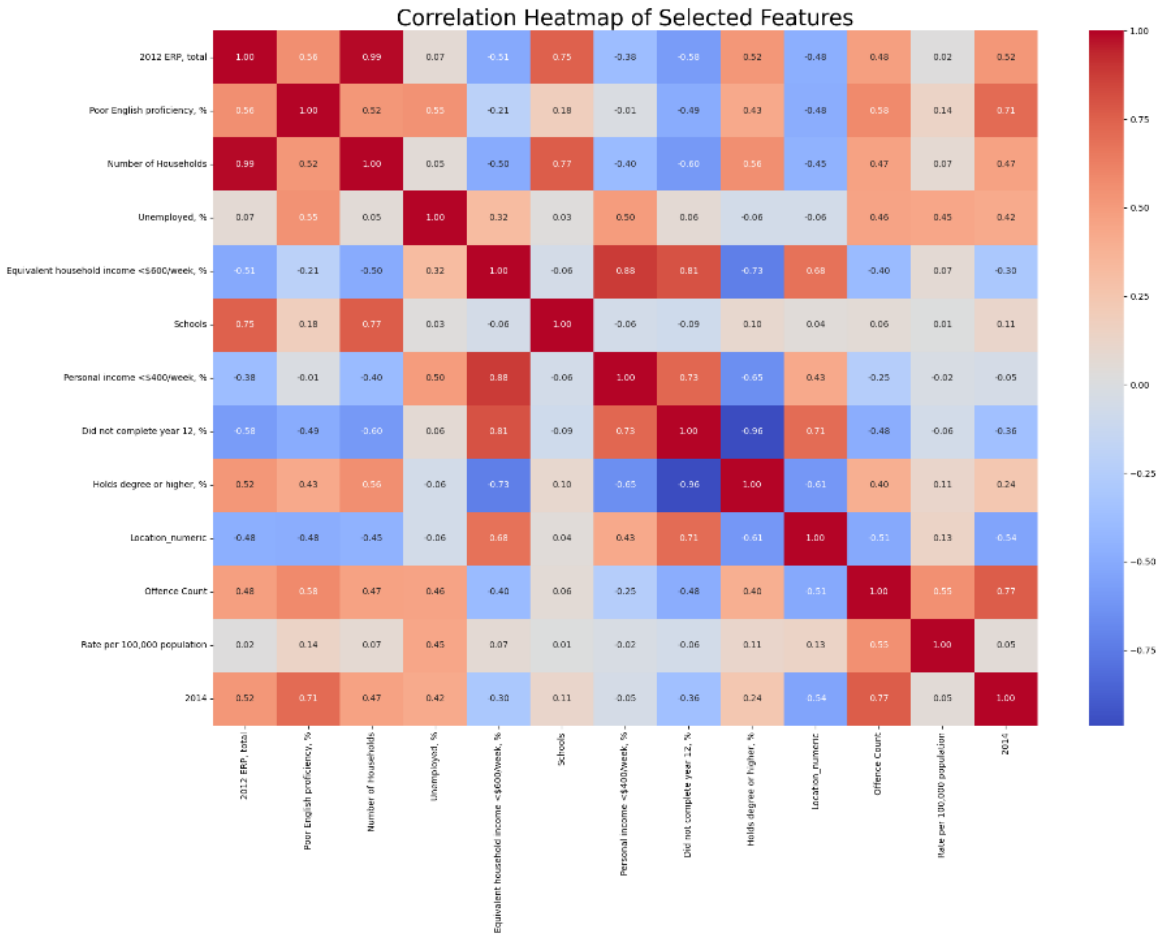

Figure 2: A heat map depicting the correlations between a variety of factors.

The heatmap revealed several key correlations between some factors but also debunking some relationships that were thought to may have been correlated. Education emerged as a critical factor and showed a positive correlation with income. The comparison was made between personal income being less than $400 and the proportion of people who did not complete year 12 (0.73), as well as whether someone holds a degree or higher (-0.65). Unemployment was also positively linked to low income (0.50). Gambling losses (represented by “2014” in the heatmap as it shows the total number of gambling losses in 2014), was moderately correlated, indicating a potential relationship between economic hardship and gambling. On the other hand, crime showed little to no correlation with income. A heatmap containing the selected features was created to better visualise the factors of interest (see below).

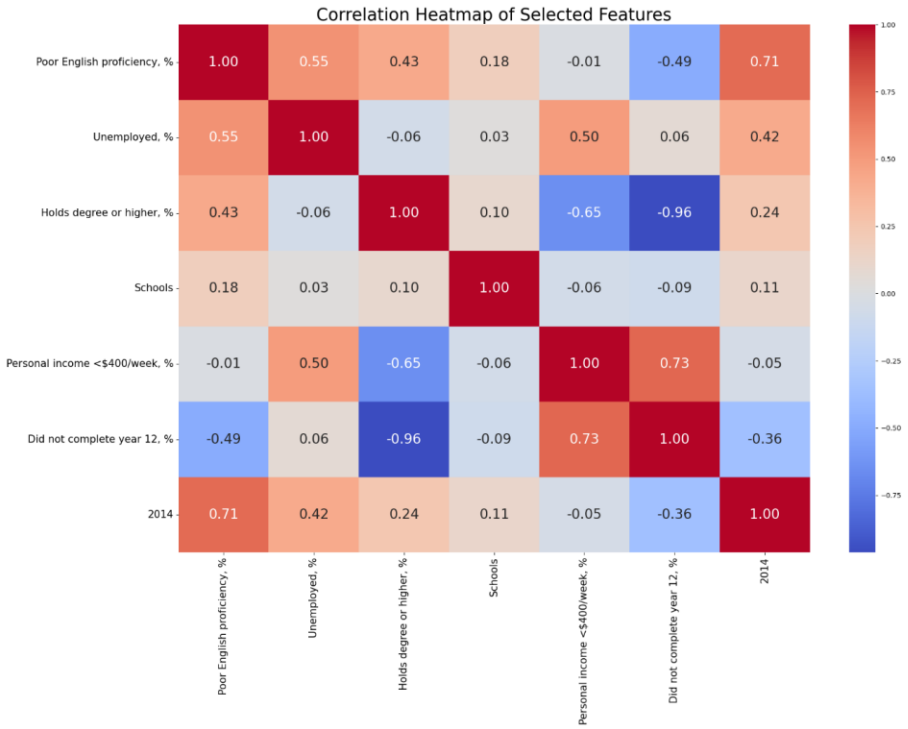

Figure 3: Heatmap based on selected features, reduced some features for formatting, all correlations are calculated with Pearson's coefficient (Ma, 2024).

Data Modelling

For this analysis, supervised learning models were chosen, specifically regression analysis, as they are well-suited for examining continuous data and quantifying relationships between variables. Regression enables us to tangibly assess how changes to independent variables (e.g., educational attainment, unemployment rates) can predict changes in the dependent variable, providing clear insights into the impact of each attribute.

A total of three regression methods were implemented. Linear regression, quadratic regression, and gradient boosting. In the following sections, a brief introduction is included for each model, discussing their strengths and weaknesses.

Linear Regression (Multi-variable)

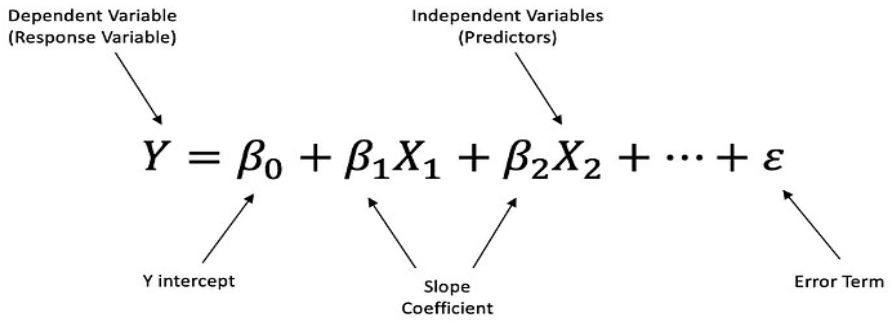

Figure 4: Multi-variable regression model illustrating the prediction model given independent variables, intercept, slope coefficients, and error term (Dawson, 2021).

As illustrated in Figure 4, this multi-variable regression model served as the baseline, allowing a direct test for a linear relationship between independent features and the proportion of low-income earners. Some key assumptions for this model include the belief that the relationship as strictly linear, and that the residuals/errors are independent from each other (the distance between predicted value and actual value).

The main advantages of linear regression include its simplicity, and low computational cost. Since we are computing the line of best fit, linear regression is a great starting point for further analysis. A limitation the linear regression model suffers from is its sensitivity to noise and outlier in the input data. To mitigate this issue, some pre-processing steps included the removal of extreme outliers, which has proven to be an effective way to address this problem. Additionally, the model is unable to accurately predict features when the relationship between variables is more complex. Given this disadvantage, polynomial regression was used to capture potential non-linear relationships.

Polynomial Regression (Quadratic)

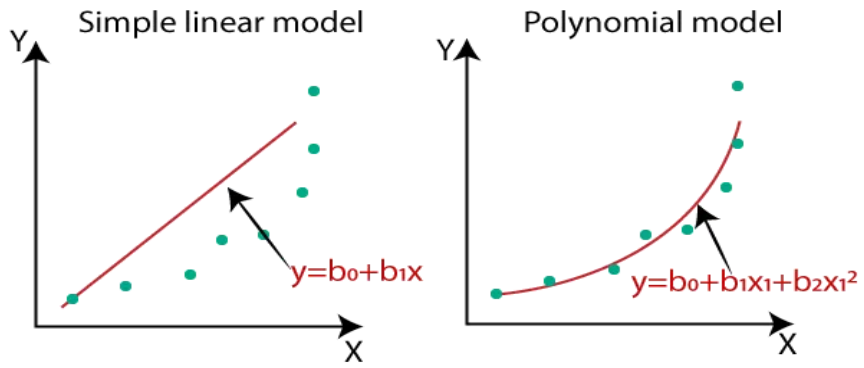

Figure 5: Comparison between linear and polynomial models of regression, highlights the polynomial model capturing a quadratic relationship (Abhigyan, 2020).

Like Figure 5, the polynomial model was used as a secondary model to compare the performance of the linear regression model. This was done to test if the linear model had inadvertently missed any non-linear correlations between variables.

While the relationship can be non-linear, this model still presumes that the variable can be expressed in terms of a polynomial function. This means that the relationship between the transformed independent variables and the dependent variable is still linear. For instance, predictions must be expressed in the form of the equation for the polynomial model shown in Figure 5.

The main advantage of using this model includes an added degree of flexibility, as not only can the model capture more complex relationships between variables, but it can also adjust the degree of the exponent to model any type of polynomials. A disadvantage stems from overfitting, in which higher-degree data tends to fit the training data very closely but perform poorly on unseen data. To combat the risk of overfitting with the model, gradient boosting was implemented. This technique utilizes a combination of decision trees to assist with predictive accuracy, while incorporating techniques that avoid overfitting.

Gradient Boosting

Gradient boosting is the final model implemented and is again used to validate the results of our previous models. It relies on the intuition that the next best model, and the previous models can work together to reduce prediction error (Hoare, 2017). The obvious strengths to this model include its ability to achieve a highly accurate prediction, regardless of the relationship between variables, and general avoidance of overfitting. However, this model is not only computationally demanding, but also considerably more volatile than both previous models, especially in terms of its performance metrics such as Mean Squared Error (MSE). As a result, the predictive accuracy may fluctuate across the specific data used, making it difficult to standardize.

Implementation and Evaluation Metrics

All three models were implemented using the Scikit-learn (sklearn) library in Python, which provides powerful tools for machine learning. A selected a set of key features was used to train the models, focusing on the proportions of individuals with poor English proficiency, unemployment rates, gambling losses in 2014, the percentage of the population holding a bachelor's degree or higher, and those who did not complete year 12. This selection was made to capture critical socio-economic factors that are likely to influence income disparity in Victoria. A specific pre-processing measure was done for the Polynomial model, where the data set values were squared before training.

The main evaluation metrics were Mean Squared Errors (MSE) and R-Squared Scores (R^2), used in combination with cross-validation. Evaluating these metrics individually may not provide a comprehensive understanding of the models' strengths, however, when used together offer a robust measure of performance.

MSE focuses on the accuracy of predictions by quantifying the average squared difference between predicted and actual values, while R^2 assesses how well the independent variables explain the variance in the dependent variable. This complementary approach ensures that we not only address predictive accuracy but also understand the explanatory power of our models.

Furthermore, by performing cross-validation, which partitions our dataset into multiple subsets, the models were trained some subsets and tested them on others across each fold. This process provided a reliable estimate of the models' accuracy across all three models.

Table 1: Performance metrics of various regression models evaluated in this analysis, including Mean Squared Error (MSE) and R-Squared (R^2) scores (Ma, 2024).

Model	Mean MSE	Mean R2 Score
Linear Regression	9.64	0.57
Polynomial Regression	18.41	0.18
Gradient Boosting	15.00 - 16.50	0.40-0.44

According to Table 1, linear regression was highlighted as the best performing, not only with the lowest mean MSE of 9.64, but also with the highest R^2 score of 0.57, indicating that this model can explain a substantial portion of the variance in low-income proportions. Building on the discussion with polynomial regression, the tendency to overfit is exposed through cross validation, as the low R^2 score of 0.18 is in line with the limitation that the model tends to perform poorly on unseen data.

Interestingly, gradient boosting demonstrated a range of MSE values between 15.00 and 16.50, and an R^2 score between 0.40 and 0.44, representing that although it offers some level of predictive capability, its performance was not as effective nor consistent than linear regression.

In relation to addressing income disparity, the analysis across the three models concludes that linear regression performs the best, indicating that the relationship between our independent variables and income is still predominantly linear. The two strongest models, linear regression, and gradient boosting have been selected for controlled experiments by adjusting specific independent variables, such as education levels, to assess their impact on income disparities. This foundation will support our subsequent discussion on potential policy inferences.

Discussion and Interpretation

Our overall analysis has yielded results that mostly aligned with our expectations, such as the clear relationship between education and income. Despite specific findings that seemed unexpected, such as the lack of correlation between low income and crime rates, the insights gathered from this study remains informative and provides grounds for comprehensive solutions in addressing income disparity in Victoria. As we explore these findings, we also wish to suggest potential policies that promote income equity across not only Victoria, but Australia in general.

Education vs. Income

Our analysis indicates that education is the most influential predictor in income, in line with past empirical studies.

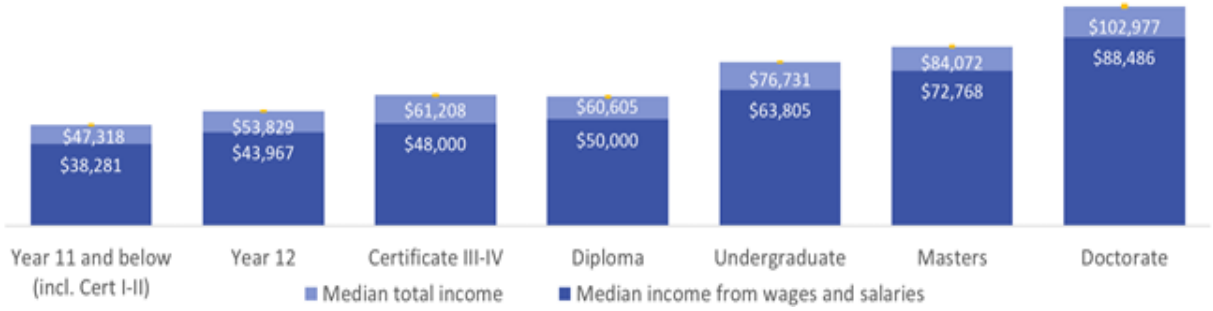

Figure 6: Bar chart highlighting that increased education level correlate with a higher median salary (Clarke, 2022)

As Figure 6 indicates, individuals who complete Year 12 or attain higher education levels are significantly more likely to secure higher-paying jobs and escape low-income brackets. This is further evident through the correlations test in our data set. Even though the study primarily captures linear relationships, our heatmap (Figure 3) reveals a Pearson's coefficient of –0.65 between the percentage of individuals holding a bachelor's degree or higher, and those earning less than $400 a week. This strong negative correlation indicates that if more people were to graduate high school or complete a tertiary degree, the likelihood of earning low-income decreases.

To test this, regression models were used to predict the impact of increasing the percentage of individuals holding a degree or higher by 20%. To ensure the experiment was controlled, the proportion of those who did not complete Year 12 was reduced by 20%, as these two variables exhibit an almost perfect negative

correlation. Across the two most effective models—linear regression and gradient boosting—an average 3% decrease in the proportion of low-income earners was found.

If it were that simple, why does Victoria struggle to keep their kids in school? The main culprit is likely to be the imbalance of funding. Public schools in Victoria, despite holding most students, are being underfunded by 1.8 billion dollars each year (AEU, 2024). On the other hand, privately owned institutions are being significantly over-funded by the Commonwealth Government (SOS, 2024). There is a hint of classism in this decision, as public schools tend to serve disadvantaged communities while private schools cater to more privileged students who already have access to additional resources and opportunities. This disproportion continues to perpetuate income disparity.

To address this imbalance, a more equitable distribution of funding is recommended, especially to those who need it. By reallocating funds based on student needs rather than the type of school, more students can be given the opportunity to complete their education, thereby increasing their chances of escaping low-income brackets.

While education plays a vital role in determining income, it is also closely tied to employment opportunities. Although higher education acts as a gateway to better employment prospects, disparities in job availability and the stability of industries across Victoria contribute significantly to income inequality. In some regions, even individuals with higher education may face limited employment options, which exacerbates the income gap, particularly in areas reliant on declining industries.

Employment as a Pathway Out of Poverty

Unemployment was another critical factor that emerged from the analysis, which aligns with the understanding that unemployment is often strongly associated with low-income levels. The results highlight the importance of *stable employment* as a key driver for individuals to escape poverty and enter higher income brackets. Having employment is not enough—long term job stability is essential for financial sustainability.

In the context of Victoria, this suggests the need to prioritize vocational training and re-skilling programs to help individuals adapt to the fast-evolving job market. As industries shift and technology advances, workers must be equipped with the skills necessary to remain competitive. Global research emphasizes that tackling unemployment is essential not only for reducing poverty but also for ensuring long-term economic growth (Kreishan, 2011). Our findings reinforce this perspective, showing that addressing unemployment will profoundly reduce income inequality in Victoria.

By investing in programs such as *Transition to Work* (Employment Services, 2023), alongside education-focused initiatives such as increasing public school funding, which equips individuals—especially youth—with the skills needed for stable employment, we can more effectively address income disparity across Australia.

Income and Crime: A Surprising Lack of Correlation

During our project, one of the most surprising findings was the lack of a strong correlation between income levels and crime prevalence. Prior to running the correlation tests, the consensus was that communities with a higher proportion of low-income individuals would exhibit higher crime rates due to economic hardship, as supported by previous studies (Sariaslan et al., 2014). Despite research suggesting that low socioeconomic status is a clear predictor of criminal behaviour, this was not reflected in the analysis.

Initially, it was hypothesized that reducing the number of low-income earners in a community would correspond with a decrease in crime. The approach involved tallying the number of crimes in each community but discovered a weak negative correlation of –0.25 between income levels and crime rates, which seemed counterintuitive. In response, our analysis was refined and focused solely on petty crimes, such as burglary and theft. This yielded a faint positive correlation of 0.15, slightly more aligned with the hypothesis, but still insufficient to infer a strong relationship between income levels and crime prevalence.

While low socioeconomic status does predict criminal behaviour, this trend is most prominent among children growing up in disadvantaged conditions. Early exposure to poverty may contribute to higher crime rates in adulthood, as these individuals face developmental challenges and limited access to education and resources (Sariaslan et al., 2014). This insight reiterates the imperativeness of implementing initiatives such as fund reallocation to public schools, to ensure children from disadvantaged backgrounds have equal opportunities for success, as well as mitigate future criminal behaviour.

Smaller indicators of low-income

Although the inclusion of English proficiency, number of Schools, and gambling losses during the training stage boosted our model's performance, our experiments revealed that changes to these factors individually did not significantly impact the model's predictions for low-income proportions. This suggests that while these features play a role in shaping income disparity, their influence comes from a convolution with other factors, such as unemployment and education. Their limited standalone effect also highlights that simply addressing these issues alone may not lead to significant reduction in income inequality, but they are still important components of a multifaceted approach to improving socioeconomic outcomes.

Limitations and improvement opportunities

There are limitations that must be considered when interpreting the results of this report.

- Data used to conduct the analysis was harvested from different datasets and spanned across separate times. Majority of the factors were taken from 2012 (the most recent year found in "communities.csv"), however, data in relation to offence count, crimes and gambling losses were from 2014. The discrepancies between 2012 and 2014 populations may lead to inaccurate calculations and therefore inaccurate results. For example, taking the crime rate by dividing the total offence count of 2014 by the total population of 2012 is likely to yield a higher crime rate compared to the actual crime rate. Additionally, socioeconomic factors such as income, education and crime rates fluctuate overtime therefore trends from 2012/2014 cannot be generalised over different periods of time. In future, data can be consistently harvested from a longer period (for example, over 4 years) to not only analyse trends but also how trends may change and what factors affect them.
- The decision to remove outliers worked to create a more general trend within the data however may have led to the exclusion of extreme but potentially meaningful data points. Outliers might represent important socioeconomic disparities that are relevant the proportion of low-income earners in Victoria. In future, having a closer look at the community understanding their unique challenges beyond the scope of generalised labels such as unemployment and explore the reasons for which have led the community to such a condition. Aboriginal Australians were recorded to have the highest rate of unemployment in 2011, suffering from high rates of alcoholism and incarceration, contributing to a cycle of poverty (Australian Bureau of Statistics, 2011).

- The regression model was trained and tested using information based off LGAs. This was beneficial to uncover underlying trends, e.g. to test whether the number of schools affect income, one suburb might not have even one school but one LGA may have 3 schools. However, a downside of using LGAs as labels is the reduction in sample size to 61 LGAs.
- As the research question was clearly defined, what factors contributed to low income, various supervised models were used in favour of a combination of supervised and unsupervised models. This is because unsupervised model is more beneficial to explore and understand the data, especially one with no labels, whereas linear regression is better suited to build prediction models. In future, it could be interesting to explore a broader research question such as "How does earning a low income affect a community?" Here, PCA could have first helped reduce the dimensionality of the dataset and K-means clustering could be used to group communities with equivalent properties, e.g. similar proportions of education-related factors and income. Using the clusters identified through K-means, linear regression could be improved by conducting an analysis of each cluster separately.

Conclusion

This report has highlighted the key socioeconomic factors contributing to income disparity across Victorian communities. Through comprehensive analysis, it is evident that education remains the most significant predictor of income levels, with higher educational attainment intricately linked to poverty reduction. As such, the report strongly advocates for reallocating resources towards public education to provide equitable opportunities for all.

Additionally, unemployment was identified as another crucial factor influencing income disparity. Stable, long-term employment is essential for individuals to rise above the poverty line. Vocational training initiatives, such as *Transition to Work*, were recommended as key solutions to address shifting market demands, equipping the workforce with the necessary skills to remain competitive.

Interestingly, the analysis found a weak correlation between income levels and crime in adult populations. However, early exposure to low socioeconomic conditions appears to contribute to criminal behaviour in adulthood. This emphasizes the need for increased educational resources in disadvantaged communities as a preventive measure against future crime.

Finally, smaller indicators such as English proficiency, school numbers, and gambling losses also play a role in shaping income disparity. Tackling these issues alongside larger factors can provide a more comprehensive solution. Our analysis provides a solid foundation for policy refinements aimed at reducing income inequality in Victoria. By addressing these core issues, these findings could potentially be applied across Australia, offering a blueprint for tackling income disparity nationwide.

References

[1] Davidson, P., Bradbury, B., & Wong, M. (2022). *POVERTY IN AUSTRALIA 2022 A SNAPSHOT*. https://povertyandinequality.acoss.org.au/wp-content/uploads/2022/10/Poverty-in-Australia-2020_A-snapshot_print.pdf

[2] Ma, H. (2024). Correlation heatmap of selected features [Unpublished manuscript]. University of Melbourne.
[3] Ma, H. (2024). Machine Learning Model Evaluation [Unpublished manuscript]. University of Melbourne.
[4] Dawson, C. (2021). *Understanding Multiple Linear Regression*. The Startup. https://medium.com/swlh/understanding-multiple-linear-regression-e0a93327e960
[5] Abhigyan. (2020, August 2). *Understanding Polynomial Regression!!!* Medium. https://medium.com/analytics-vidhya/understanding-polynomial-regression-5ac25b970e18
[6] Hoare, J. (2017). *Gradient Boosting Explained - The Coolest Kid on The Machine Learning Block*. Displayr. https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/
[7] Clarke, M. (2022). *Income - Department of Education, Australian Government*. Department of Education. https://www.education.gov.au/integrated-data-research/benefits-educational-attainment/income
[8] New research shows urgent need for full funding for Victoria's schools | Australian Education Union (AEU) Victorian Branch. (2024). Aeuvic.asn.au. https://www.aeuvic.asn.au/new-research-shows-urgent-need-full-funding-victorias-schools
[9] Private Schools Serving Richest Victorian Families Over-Funded by Millions – SOS Australia. (2024). Saveourschools.com.au. https://saveourschools.com.au/funding/private-schools-serving-richest-victorian-families-over-funded-by-millions/
[10] Sariaslan, A., Larsson, H., D'Onofrio, B., Långström, N., & Lichtenstein, P. (2014). Childhood family income, adolescent violent criminality, and substance misuse: quasi-experimental total population study. *British Journal of Psychiatry*, *205*(4), 286–290. https://doi.org/10.1192/bjp.bp.113.136200
[11] Kreishan, F. M. (2011). Economic growth and unemployment: An empirical analysis. Journal of social sciences, 7(2), 228-231.
[12] Employment services. (2023). Australian Institute of Health and Welfare. https://www.aihw.gov.au/reports/australias-welfare/employment-services
[13] Australian Bureau of Statistic (2012). 2076.0 - Census of Population and Housing: Characteristics of Aboriginal and Torres Strait Islander Australians, 2011 https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/2076.0main+features802011
[14] National Center for Education Statistics. (2024). Annual Earnings by Educational Attainment. *Condition of Education*. U.S. Department of Education, Institute of Education Sciences. From https://nces.ed.gov/programs/coe/indicator/cba.
[15] The University of Melbourne. (2012). *Poverty Lines: Australia – March Quarter 2012. https://melbourneinstitute.unimelb.edu.au/assets/documents/poverty-lines/2012/Poverty-Lines-Australia-March-Quarter-2012.pdf*