EODP Assignment 2 report

Group Name: H1 Fighter
Sheldon Liu (1455927), Zihan Dai (1488802), Jiahao Ni (1455828),
Phone Moe Thway (1342981)

October 7, 2024

1 Executive Summary

In the past decade, social-economic factors have played an increasingly crucial role in influencing crime rates within urban communities. In Victoria, Australia, understanding the interplay between different factors and crime is essential for effective policy-making and community development.

In this project, we aim to analyze what factors about communities in Victoria have the greatest impact on the average crime rate of all Victoria Suburbs. To achieve this goal, we first performed data preprocessing and then we used Exploratory Data Analysis to extract meaningful insights. Next, we moved on to correlation analysis.mBased on our correlation analysis, we also found some Economic, Regional, Social, and Educational indicators that correlate with crime rate the most.

Eventually, we created supervised models that measure what will average crime rate in Victoria Suburbs will be (in percentage), depending on how government manipulates other socioeconomic factors. We first performed feature engineering, then we selected the most relevant ones out of the features from the combined dataset in order to elevate their predicting power on crime rate. The government can use this model to verify their estimates on to extent a factor influences crime rate(thus leading to better planning).

Throughout the investigation, we figured out some key findings that crime rate is closely related to social factors, economic factors, and regional factors. These factors will be explored in more detail in later sections. We elaborated on a few key recommendations for the government to control the crime rate which will be further discussed. The government needs to boost the economy, focusing on better urban planing and controlling suburb population diversity.

2 Introduction

Our research question is to examine which social-economic factors have the strongest impact on crime rate in communities across Victoria. As such Victoria government can work on improving these aspects in order to avoid community's exposure to crime.

From communities csv File, We selected 25 factors that are intuitively relevant to crime rate. We also use house price dataset, this dataset contains 774 distinct suburbs and house price from 2013 to 2023. Additionally, EGM loss dataset is used. In this dataset, there are 57 distinct LGAs recorded. And the total amount of money losses from 2011 to 2020 are recorded. We also used Table 01 and Table 03 from LGA_Offences. Table 01 kept track of total offence count and crime rate broken down by 79 unique LGAs. Table 03 kept records of Offence count broken down by offence divisions. Furthermore, the specific suburb where the offence occurred is reported.

3 Method

3.1 Pre-processing

We begin by cleaning and preparing each dataset to ensure consistency and reliability for our analysis.

For the **Houses by Suburb** dataset, we first remove any rows containing missing or invalid data, specifically those with entries marked as '-'. There are 792 suburbs in total and only 12 empty entries are being removed, therefore we assume the impact of removal is minimal. We standardize the suburb names in the *Locality* column by converting them to title case to maintain consistency across datasets.

In the **Victorian Communities** dataset, we filter the data to include only suburbs—entries where the *Community Name* ends with "(Suburb)"—and remove the "(Suburb)" suffix for uniformity. We select pertinent features such as population density, area, land use percentages, socioeconomic indicators, and the availability of health and educational facilities. Missing values are imputed with the mean of their respective columns to handle incomplete data. The population of each suburb is calculated by multiplying the population density by the area. We also standardize the *LGA* names by removing designations like "(C)" and "(S)" to match naming conventions across datasets.

For the crime data from the **LGA Offences** dataset, we read and clean the data by removing unnecessary columns such as *Year ending* and *Police Region*, and exclude rows that do not correspond to actual LGAs (e.g., total counts or unincorporated areas). We rename columns for clarity and group the data by *LGA* to compute the mean offence count and crime rate per 100,000 population over the years. Consistency in *LGA* naming conventions is ensured by trimming whitespace and standardizing names.

In the **Electronic Gaming Machine (EGM)** dataset, we clean the *LGA Name* entries by removing prefixes such as "City of," "Shire of," and "Rural City of," and convert them to title case for consistency.

After preprocessing the individual datasets, we merge them to form a consolidated dataset for analysis. The *Houses by Suburb* dataset is merged with the *Victorian Communities* dataset on the *Locality* field. This combined dataset is then merged with the crime data on the *LGA* field, and subsequently with the EGM dataset also on the *LGA* field.

We calculate the average offence count per suburb using the **Suburb Crime Rate** data by grouping offences by suburb and averaging over the years. The *Suburb Crime Rate Percentage* is computed by dividing the average offence count by the suburb's population and multiplying by 100 to express it as a percentage.

To account for the proportion of the LGA's population residing in each suburb, we calculate the total population per LGA by summing the populations of all suburbs within each LGA. The *Population Percentage* for each suburb is then determined by dividing the suburb's population by the total LGA population. Using this proportion, we estimate the *Suburb EGM Losses* by weighting the LGA's average EGM losses by the suburb's population percentage.

Finally, we clean the consolidated dataset by removing unnecessary columns such as individual year data and redundant identifiers. This results in a prepared dataset that includes key variables like average house prices, socioeconomic indicators, crime rates, and EGM losses at the suburb level, ready for further analysis.

3.2 Feature engineering

We calculate the average house price for each suburb by averaging the prices from 2013 to 2023. We also calculate the average EGM losses for each LGA over the years 2011 to 2020 by averaging the losses across these years. By using the averages, we smooth out short-term fluctuations and capture the long-term trend of housing prices in a suburb. Moreover, it would decrease the number of features in the model to avoid curse of dimensionality, reducing potential multicollinearity and overfitting within our models in later steps.

Additionally, we create new features, *Number of Hospitals* and *Number of Aged Care and Schools*, by summing the counts of related facilities in the Communities dataset. These two features act as important indicators of health and education services quality of a suburb which can be a decisive influencer for crime rate.

3.3 Feature selection

To reduce the risk of overfitting and enhance the model's predictive performance, we perform feature selection based on correlation analysis and multicollinearity checks. Features exhibiting high correlation with the target variable but low inter-correlation with other features are preferred.

We generate a correlation heat map to visualize the relationships among features and identify any highly correlated pairs that may introduce redundancy. If two features are highly correlated, one of them may be removed to avoid multicollinearity.

Born overseas and Poor English proficiency have strong correlation between each other, thus one of them has to be removed to avoid multicollinearity.

Also, we remove any variables that are weakly correlated (absolute value of correlation factor ¡ 0.1) with target variable.

Based on the correlation analysis and domain knowledge, we select the following key features for our predictive models. They are Average Price, Change Percentage 13-23, Number of Hospitals, Unemployed (%), Commercial (%), Residential (%), Born Overseas (%), % dwellings which are public housing, Personal income < $400/week (%), Equivalent household income ¡$600/week (%).

3.4 Regression Model

We employed two different regression models: linear regression and decision tree regression.

Reasons for model choice:

- Regression models are good at predicting continuous variables, suburb crime rate(variable of interest) is continuous.
- **To better help government plan changes for a better community, we desire a model that tells the government, if they reduce poverty factors (low income, unemployment etc.) to a certain percentage, or raise values of other factors(amount of EGM loss, rise percentage of house price etc.), what will crime rate in Victoria suburbs be as a result of governments control over other variables. Regression model monitors how independent variables influence dependent variables in a clear way, so we believe this is the best model to show how factors affect crime rate.**
- Linear regression is effective for modeling linear relationships. However, non-linear relationships may also exist between relevant factors, we also used a decision tree regressor to better capture this.

We observe that the target variable, *Suburb Crime Rate %*, exhibits a right-skewed distribution. To normalize it, we apply a logarithmic transformation using $y_{\log} = \log(1 + y)$, where y represents the original crime rate percentage. This transformation helps in stabilizing the variance and making the data more suitable for regression modeling.

To ensure that all features contribute equally to the model and to improve convergence during training, we standardize the features using the StandardScaler. This scales the features to have a mean of zero and a standard deviation of one. With the selected features, we proceed to model the relationship between the predictors and the target variable using regression techniques.

In order to prevent model from overfitting, we utilized 80% dataset as training data and 20% dataset of testing data. And we used cross-validation with five folds.

3.4.1 Linear Regression

As a baseline, we implement a linear regression model to establish a point of reference for model performance. The linear regression model is trained on the standardized features and the log-transformed target variable. We evaluate the model using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2 score).

3.4.2 Decision Tree Regression

To capture nonlinear relationships and interactions between features, we employ a Decision Tree Regression model. Decision trees partition the feature space into regions with similar target values, making them effective for modeling complex datasets without requiring extensive data preprocessing.

The Decision Tree Regression model is trained on the same standardized features and evaluated using the same set of performance metrics as the linear regression model. We also perform hyperparameter tuning, such as adjusting the maximum depth of the tree, to prevent overfitting and improve the model's generalization capability.

By comparing the performance of the linear regression and decision tree models, we aim to determine whether a nonlinear model provides a significant improvement in predicting suburb crime rates based on the selected socioeconomic features.

3.5 Model Evaluation Metrics

3.5.1 MSE

Mean Squared Error (MSE) measures the average squared difference between the actual values (observed) and the predicted values (estimated) by the model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3.5.2 MAE

Mean Absolute Error (MAE) measures the average of the absolute differences between the actual values (observed) and the predicted values (estimated) by the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

3.5.3 R Square

R-squared (coefficient of determination) explains how well the independent variables in a regression model predict the dependent variable. R^2 ranges from 0 to 1, with 1 indicating that the model perfectly predicts all the data points.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

4 Data Analysis

4.1 Exploratory Data Analysis and Trend Analysis

We believe an efficient way to investigate crime rate is by first analyzing a community that is most representative.

Melbourne is the most populated LGA in Victoria and might be the most representative(larger sample). It has the highest number of offence counts in all types of crimes, meaning studying this LGA can give us a more comprehensive reflection on crime rates.

According to The Sydney Morning Herald, Melbourne was no longer the NO.1 most livable place in 2017 due to many consequences of urban development, including burglary and theft "crime rising rapidly" and "Housing affordability plummeted". We seek to explore how these 2 problems might interweave with each other. This analysis might better help Victoria government to tackle property safety crime in modern society that is becoming more urbanized.

The number of property and deception offences (including theft and burglary) in Melbourne is significantly higher than all other LGA. We believe we can better examine how EGM loss and rising house prices relate to property crime by looking at a region with the most concentrated crime occurrence.

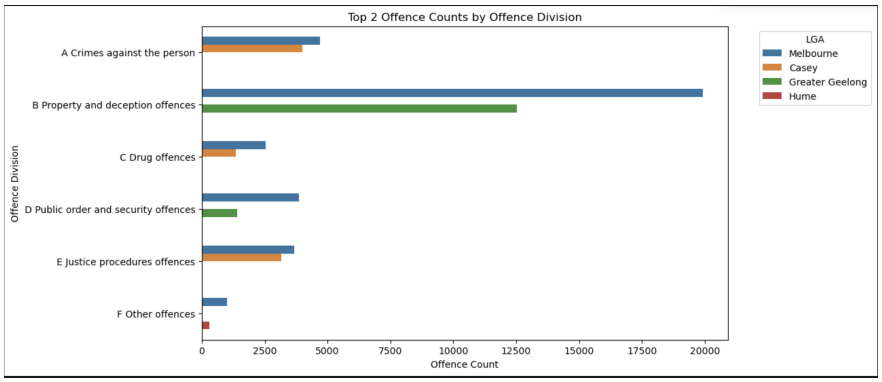

Figure 1: This figure shows the Top 2 offence count by division.

The gap in this figure was caused by minor formatting issues, the message is still accurate Therefore, Melbourne may be the best starting point for our data analysis.

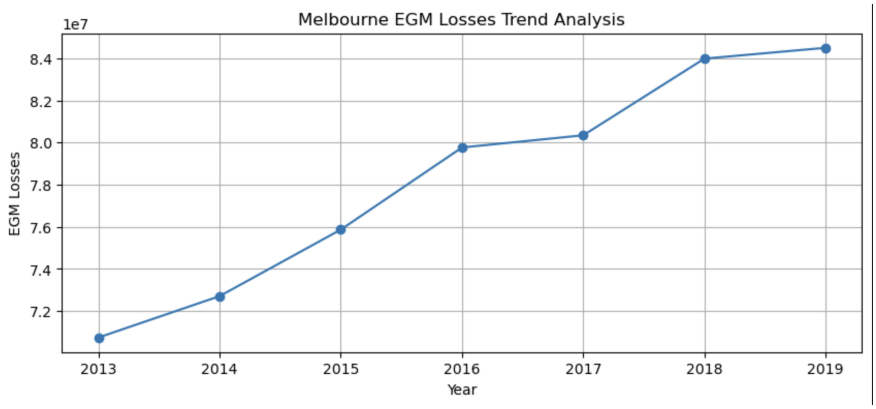

Figure 2: This figure shows the trend of Melbourne EGM losses .

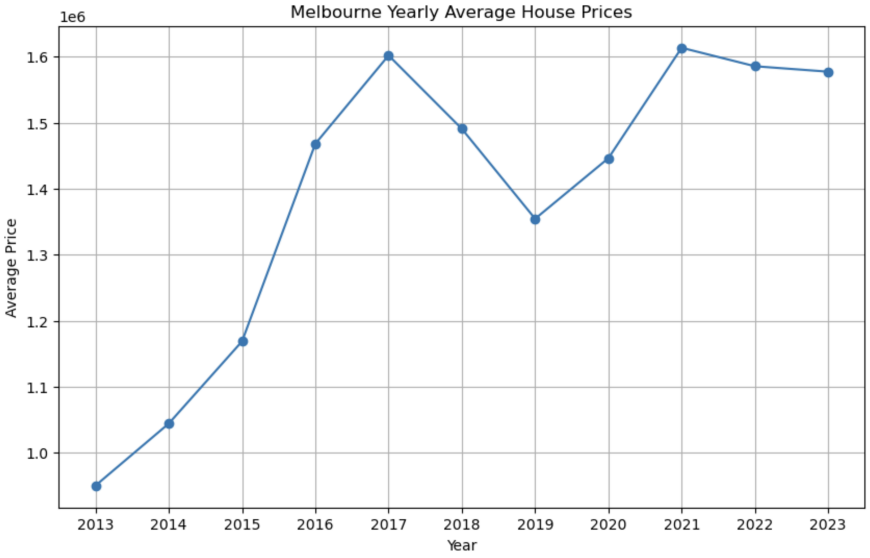

Figure 3: This figure shows the trend of Melbourne average house price.

Both EGM loss and house price experienced the greatest rate of growth from 2014 - 2016, and their respective value dropped or grew more slowly during and after 2017. Coincidentally, offence count of money-related crime also spiked high from 2014 - 2016, and decreased rather rapidly during and after 2017(we count money related crime using the total count of B40 Theft, B30 Burglary/Break and enter, A50 Robbery from offence division data set).

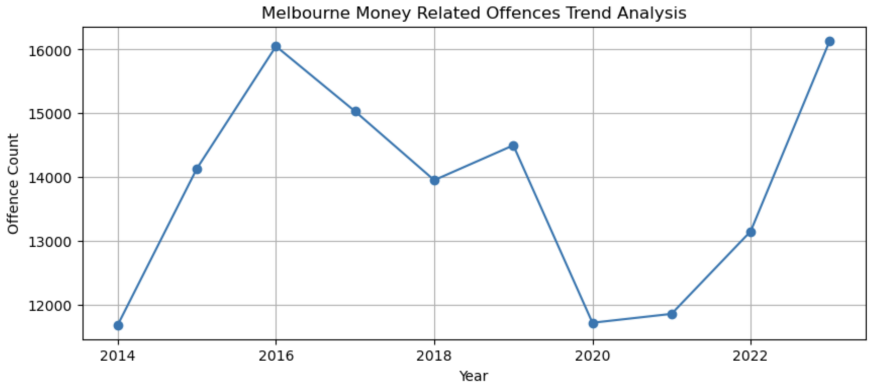

Figure 4: This figure shows the trend of Melbourne money-related offence count.

We looked at real-world polices that might caused the similarity between the trends of EGM loss and property crime rate.

(DATA VIC gambling statics): From 2014, the Victoria government has gained 8 % budget from Melbourne EGM tax revenue. In 2016, this proportion rose to around 10%. In 2017, the Australian institution of family estimated $100 increase in EGM expenditure, and there was a 0.65% increase in property crime. Crime types include digital theft with a maximum amount up to millions and physical burglary/robbery.

Major EGM reforms began in 2017 include:

- Establishment of the Victorian Gambling and Casino Control Commission which is a newly formed regulator (VGCCC).
- EGM machines have mandatory pre-commitment limits to reduce loss.
- Mandatory closure period that reduces EGM access in non-casino area followed by 7.4% in total property criminal offenses from 2017 to 2018

Similarly, many policies may explain the similarity between the trends of house price and property crime rate.

Since 2014, Victoria government introduced Skilled and Business Migration Programs which resulted in a rising population that naturally drove prices of limited housing supply. According to Crime Statistics Agency Victoria, the average house price in Melbourne hovered around AUD 1,614,070 from 2013 to 2014, some experts speculate 15% more homes are perceived as high-value targets can attract more property crime. 37% of individuals who committed property crimes do financially struggle with aspects including commendation. Although some researches seemingly align with the assumption that EGM loss and house price do affect property crime, we also encounter reports and opinions that suggest otherwise.

Hence, more analysis needs to be done by the government for a more definitive answer of whether a causal relationship truly exists. Our analysis serves as an entry point for government to formulate further trend studies.

In the previous trend analysis, we did not make interpretation on property crime for years after 2019. ABC suggests covid-19 in 2020 leads to major economic recession, which can a significant confounding variable for analysis on money related crime. For this research question, We want to conduct study on data set not affected by covid - 19, but if government can provide more

information about the epidemic, analysis about covid-19 and crime rate can be explored in other research questions.

Australian Institute of Criminology suggests although improving social-economic factors like employment and living quality may discourage criminal behavior, imperative intervention like legal punishment and police intervention is also crucial as it deters crime. We believe the feature most relevant to this notion is investigation status of crime in LGA we are provided with.

Using Melbourne as an example, 4 types of investigation status shows:

- 50% Arrest/Summons - community has a decent ability to deliver punishment/sentences for criminal activity
- 34.5 % Unsolved - the ability to investigate criminal activity can be improved, so offenders are not indulged by escaping punishment.
- 3.6% Not authorised - good management over investigation process.
- 12.7% of other status - government may need to further examine what these statuses actually are and what positive or negative impacts they have.

Example of Not authorised investigations may be police search ones' private property without the warrant required, or using unauthorised violence against suspect. This aspect may reflect problems of methods used to enforce law and poor management within justice departments, which is crucial to delivering systemic justice to Australian legal system.

Other LGA can also use similar pie charts to assess strengths and weaknesses of their crime investigation, and then improve upon it to better handle crime occurrence.

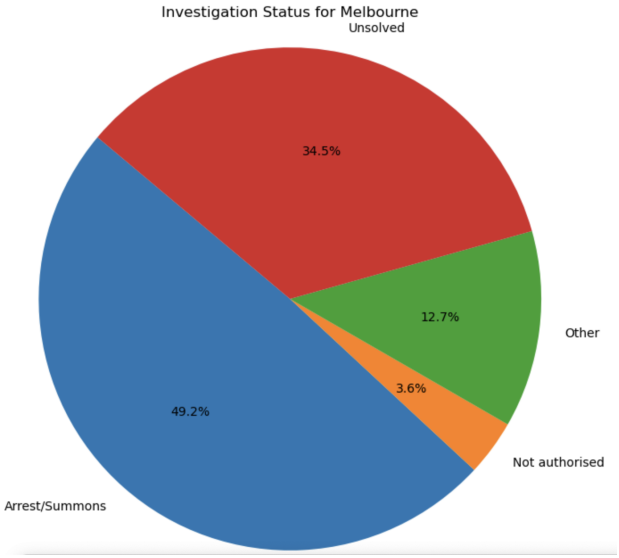

Figure 5: This figure shows the investigation status piechart.

4.2 Correlation Analysis

After exploring the crime rate on the smaller scale of Melbourne, we also believe the government can gain an overview of more causes of crime on the grand scale of the entire Victoria (all the suburbs), so they can take more factors into consideration.

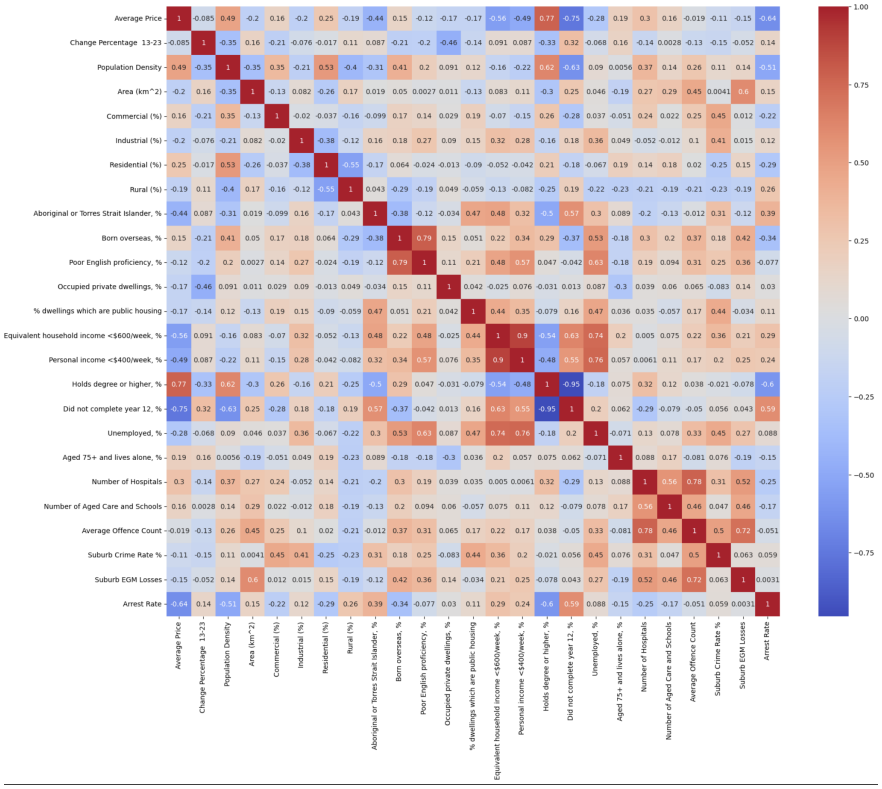

Figure 6: This figure shows the heatmap.

To achieve this, we use a correlation heatmap to find correlations between social factors (features) to a new target variable, Suburb Crime Rate (%).

Here are the findings that are intuitive:

Unemployment (correlation factor : 0.45), weekly income less than $400 (0.2), weekly income less than $600 (0.36), are positively correlated to crime rate. Affirming common notion that improving economical condition of Victoria will likely reduce crime rate. % of dwellings which are public housing (0.44) also shows rather strong positive correlation with crime occurrences. Department Of Treasury And Finance in Victoria found economically disadvantaged individuals are more likely to live in public housing. So the correlation might reflect that crime occurrence affect financially stressed individuals (live in public housing) more. This is an issue deserve government attention in time of growing economical disparity in modern society.

Results from the heatmap is different to analysis of Melbourne. It suggests EGM loss and house price is less related to crime rate: Crime Rate of all suburbs in Victoria(%) has a negative correlation with average Price of houses(-0.11), house price Change Percentage 13-23 (-0.15). It also little correlation with Suburb Egm loss (0.063).

This is does not necessarily imply mistake in our approach. The states of many social-economic factors are different in smaller regions compared to the whole of Victoria overall. (E.g. Melbourne may be a region affected more by EGM loss and subsequent crime, whereas gambling is less popular and less of a lead to crime in other parts of Victoria). As a result, different observations come naturally when making assessments from scopes of different area scale.

Model	MSE	RMSE	MAE	R^2 Score
Linear Regression	0.0907	0.3	0.2291	0.6414
Decision Tree	0.1608	0.4006	0.3222	0.3595

Table 1: Performance Comparison of Regression Models

5 Model Evaluation

Based on the above model performance, we can observe that linear regression is a superior model that has less MSE,RMSE,MAE than decision tree regressor and an R square score 0.3 higher.

There are several potential reasons why this is the case: 1. Features and target variable themselves are indeed more linear related.

2. For decision tree regressor to have better performance, more effort of hyper parameter tuning is required compared to linear regression model.

e.g. we need to choose which combination of max_depth, max_features and max_leaf_nodes yields the best result (out of many possible choice of combinations).

Initial R square value is close to 0, we tuned the model to raise R square value to 0.3595. Due to limited time given, this is the best improvement can be made. We understand the difficulty of hyper parameter tuning is the limitation of this model.

6 Discussion

Trend plots and research of polices in Melbourne suggests EGM loss and House price is strongly related to property crime. However, correlation analysis suggests there is minimal correlation between these factors. As mentioned before, scopes of large scale (whole Victoria) yield different results compared to a specific LGA. We recommend government to do further study on whether EGM and house price driving property crime is a particularly potent issue in Melbourne. Following this logic, we also advise government to use a similar approach, and examine why some factors raise crime rate more in specific LGA, even though from the perspective of the whole Victoria, the factors seem to relate less to crime. (Better planning of more specific regions)

There are many unexpected or meaningful insights from our correlation heatmap:

- Low correlation level between "not completing year 12" and "suburb crime rate". TAFE Victoria argues that Victoria government provides good career/living quality without the need to complete Year 12. This may suggest if Victoria government continues to provide more life/career options without high demand for academic achievements can avoid potential crime related to low education level.

- Research from the Australian Institute of Criminology suggests legal punishment is needed to deter crime. On the heatmap, there is a low correlation between "arrest rate" and "Suburb Crime Rate" (correlation factor = 0.059). Based on this, we advise Victoria government to study whether crime rate is better reduced through legal punishment or working on other social benefits (boosting economy or better prisoner re-education).

- % of Commercial (0.45) industrial (0.41) are highly correlated to average suburb crime Rate % residential (-0.25) rural(-0.23) are negatively correlated to the target variable. We believe this result is reasonable since Commercial and industrial regions involve more human and business activity, which naturally creates the potential for more crime. The opposite logic applies to

residential and rural areas. When developing the community, government might want to balance the proportion of types of regions so that they keep crime rate at a manageable level.

- Percentage of Aborigines dwellers has correlation coefficient of 0.31(one of the highest value). Minister of indigenous Australian suggests effects of colonisation and institutionalised racism leads to more indigenous people being imprisoned or deteriorating aspects of their living quality, which might drive them to crime. Based on these observations, we hope government can provide more data relating to how they treat indigenous people, so that we can analyse which aspect of indigenous care can be done better to lower indigenous crime rate.

- % population "born overseas" has a correlation factor of (0.18), which is not too minimal to be ignored. Considering Australia has open immigration policy and illegal immigrant/refugee issues (ABC news), government might need to research more on how foreigners influence crime rate and why.

We have analysed investigation status for crime from 2014 - 2023 in Melbourne, identifying that the ability to investigate unsolved cases can be improved. Punctual crime intervention (manifested through 50% arrest and summons) and legitimate investigation process (manifested through a low rate of unauthorized type of investigations) are positive performances that need to be maintained in order to better thwart crime. We believe government can use similar methods to identify strengths and weaknesses of other regions to improve crime intervention.

Summarised answer to research question:

- Economical factors that influence average suburb Crime rates include : unemployment, % Personal income less than $400, % Household income less than $600

- Social factors include Percentage of Aboriginal or Torres Strait Islander and Percentage of population Born Overseas

- Region factors are % of a region type: (Commercial, industrial, residential, or rural), % dwellings which are public housing

7 Limitation and Improvement Opportunities

- Limited definition of investigation status limits us from making better interpretations. For instance: What does "other status" represent? Does "summon" status include summoning witnesses to court(thwart crime with public effort) instead of only summoning the suspect to court for a sentence?

- The major limitation in terms of data set is that features in Communities CSV file have only values for year 2014. Since we use these features in our correlation heap and regression models, having data on features for different years might yield better results. We can also perform trend analysis on some factors covered in this file if we know how these factors change over the years.

- Features like EGM loss, suburb crime rate (derived from LGA crime rate), and increase in house price, which does have values over the years. We found the average value over the years and used them for our models, which have positive and negative impacts. Positively, it smooths out short-term fluctuations and assesses overall trends throughout the years. On the other hand, very extreme outliers may exist and produce exaggerated average values. We

may want to improve our code so that it defines what values are considered extreme outliers in different cases and ignores them for the calculation.

- The weakness of only using regression models trained on solely single average values over the years, is that it ignores the effect of time. We may want to learn about time series analysis model (and other models beyond the scope of this subject) for analysis in the future.

- We are aware that our target variables can be more detailed, including rates of other types of crime instead of just property crime. The approach of our data analysis can be replicated on other crime type. According to ALFS, Brimbank, a suburb of Melbourne with a high density of EGM venues, local reports show concerns over drug use, violence, and other crimes near gambling venues. Our Trend plot can also evaluate how EGM loss influences Drug Crimes/ Violent crime by simply switching the target variable.

8 Conclusion

Throughout this project, we find some Economical, Social, Education and Region Factors that have the strongest effect on average crime rate all suburbs across the whole Victoria. These results can help government manage crime rate through these 4 perspectives(Economical, Social, Education and Region). Our results also reflect that looking at more specific region, such as LGA Melbourne, may provide different insights compare to the scope of whole Victoria. Therefore we advise government to do more focused study at smaller scopes if they intend to lower crime rate of a particular region. Major limitations of the data set include: somehow unclear definition of key categories (investigation status) and lack of data/trend over years for many features. If government solve these data quality issues, we can make more trend analysis and have clearer direction when finding real implication of our results. Our group created decent regression models that capture what(%) crime rate will become as a response of government's control over other social economic factors as we intended. We believe it a very intuitive model that can help with government's planning of a community with less crime.However, we need to study more advanced models that require less hypertuning effort and can better examine effect of time.

References

source: https://www.smh.com.au/opinion/five-reasons-why-melbourne-is-no-longer-the-worlds-most-liveable-city-20170215-gudbv7.html

ALFS: https://aifs.gov.au/sites/default/files/publication-documents/1902_gambling_in_suburban_australia_0.p

Crime Statistics Agency Victoria: https://www.crimestatistics.vic.gov.au/crime-statistics/latest-crime-data-by-area

Australian Institute of Criminology. https://www.aic.gov.au/publications/crm/crm27

Indigenous Crime rate: https://www.youtube.com/watch?v=r5vXtYrFO-U