MAST30034
Applied Data Science

Real Estate Industry Project 2022

Group 30:
Arshia Azarhoush
Sophie Sarwesvaran
Ayesha Tabassum
Sureen Tiwana
Andrew Dharmaputra

THE UNIVERSITY OF MELBOURNE

Objective

What did we want to answer?

Which 10 suburbs have the highest predicted growth rate?

What are the most liveable and affordable suburbs?

What are the most important internal and external features in predicting rental prices?

Project Logistics Timeline

Time frame and communication

- 6 sprints over a span of 6 weeks.
- Utilized Trello for project-mangement and distributing tasks to each member.
- Communication through Facebook Messenger and Zoom.

Sprint 1: Project planning. Start web scraping for property data, create Python scripts to download other external datasets. This took roughly 10 hours of work in total.

Sprint 2: Further web scraping, start data preprocessing and visualisations for external datasets (including the use of openrouteservices API). This took approximately 6 hours.

Sprint 3: Finish web scraping for property data, and further data preprocessing and visualisations, including property data. This took 6 hours in total.

Sprint 4: Finish preprocessing and visualisations, then join those datasets to be used for modelling. 7 hours was allocated to this.

Sprint 5: Modelling and feature analysis. Exploring different models and tuning parameters took roughly 8 hours.

Sprint 6: Finish modelling, and begin presentation preparations. On top of this, quality testing, general improvements and bug fixing took a total of 7 hours.

Literature Review

Overview of Two Similar Research Projects/Studies

California Rental Price Prediction Using Machine Learning Algorithms (Fei Yue, 2020)

Aim: predicting possible rental prices in California
Scope: Used dataset containing all properties in California listed on Airbnb
After feature analysis, the number of bedrooms and the property type were found to be the features most highly correlated with the rental price
The model XGBoost produced the best prediction result.

Benefit and cost: Huge sample of data. Great for the model accuracy but computationally expensive.

Predicting Rental Prices in Amsterdam

The aim was to make predictions for the real estate market in Amsterdam. Data gathered from Amsterdam rental website Pararius.
The two most important features were the postcodes - namely, proximity to city - and the square meterage of the property.
Obtained a 94.75% prediction accuracy with Random Forest
Conclusion: Realised the true complexity underlying the real estate market from the number of variables involves - e.g. more than just postcode and square meterage, but also proximity to good restaurants, etc.

Liveability of Suburbs

What makes a suburb "liveable"?

Liveability is measured by factors that promote quality of life, such as:

- Low crime rates
- Easeful access to public transport
- Higher median family and personal income

External Data

What data did we use and where did it come from?

- Census data and suburb geospatial information from ABS.
- Public transport information from PTV.
- Crime data from crimestatistics.vic.gov.au
- Openrouteservice was used to create isochrone maps.

Openrouteservice API

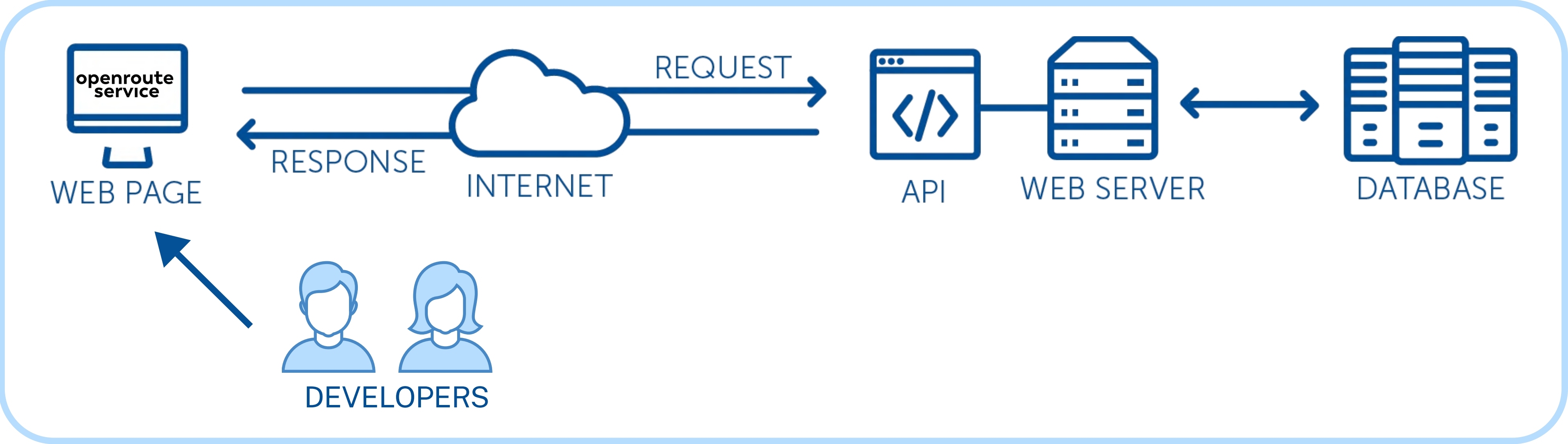
What is an API and how did we use it?

What is an API?

Mechanisms that enable two software components to communicate with each other.

Why did we use the Openrouteservice API?

Created isochrone maps around train stations.



Security and Ethics

How have we ensured a secure, ethical analysis?

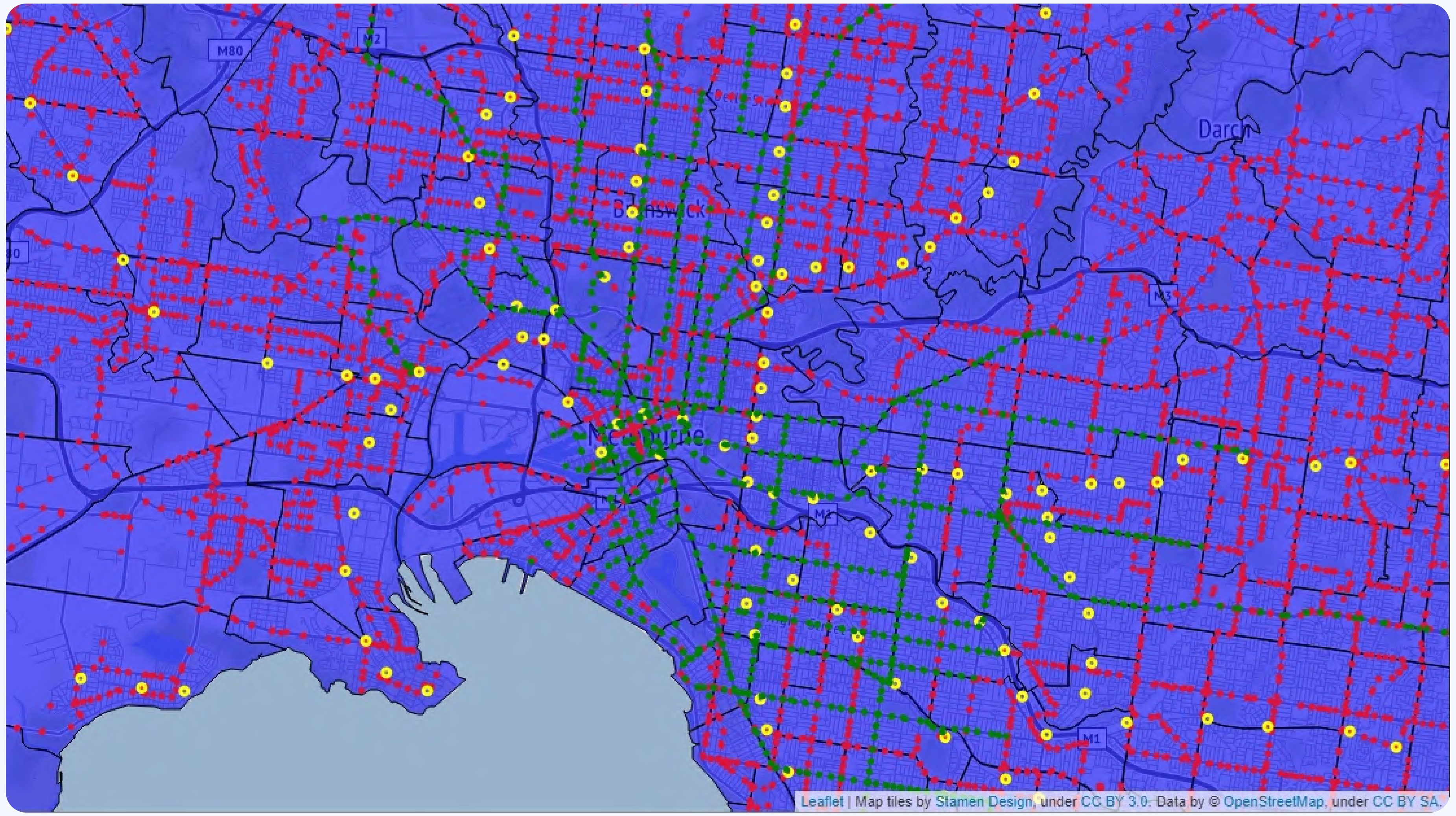
- Complied with the API terms of service and usage limits.
- Stored API keys offline, in a local file as an environment variable.
- Data kept within repository, with no datasets being uploaded and accessed via external means. Datasets downloaded via data download scripts.
- Data stripped of any personal identifiers.

PTV & Suburb Data

Visualisation of PTV data.

- PTV data has over 20,000 metro and regional bus stops, 330 metro and regional train stations and 1665 tram stops.
- PTV locations overlaid on SA2 map.
- We used these datasets for our geospatial visualisation and analysis

Location of PTV Stops & Stations

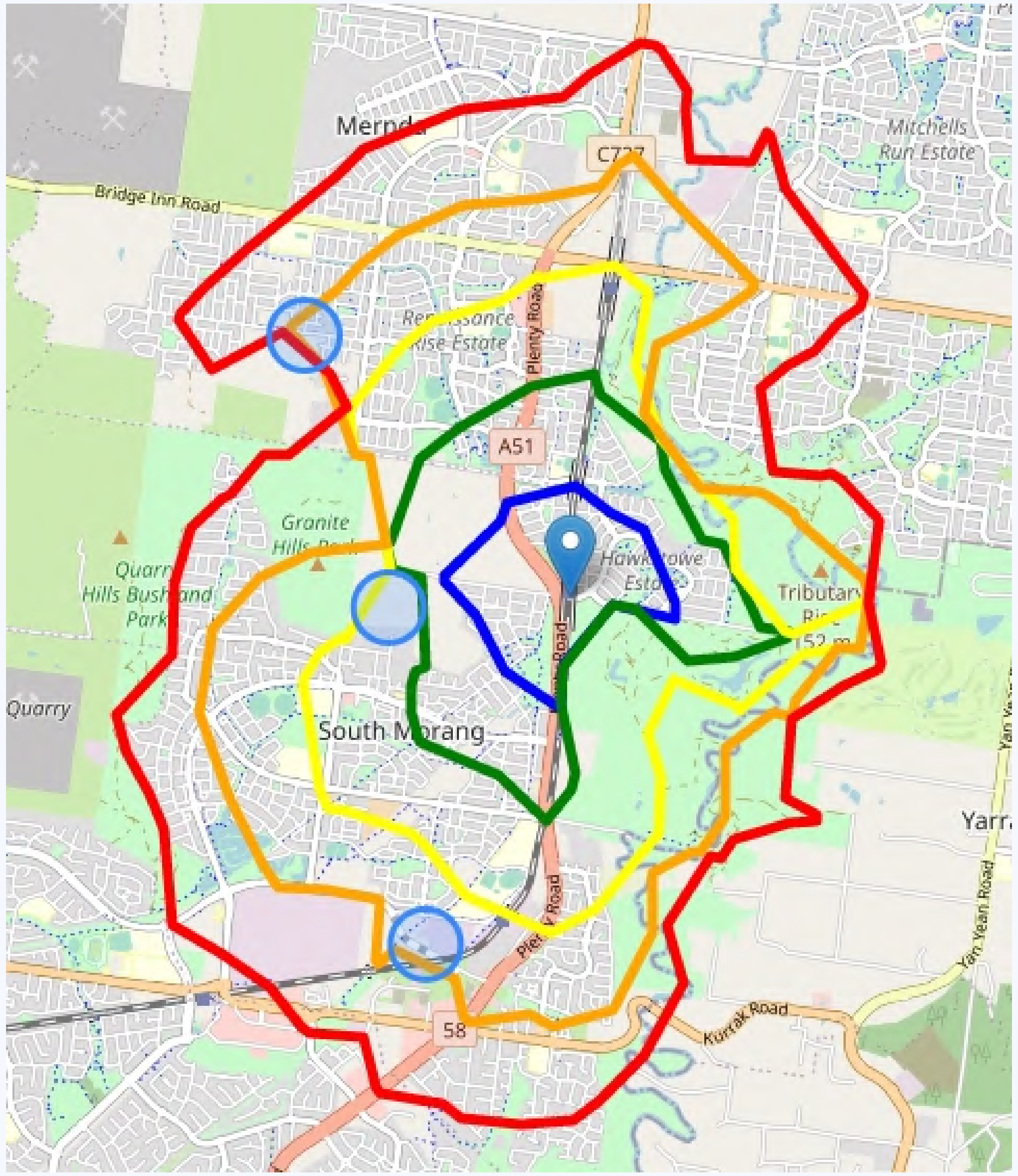

M80
M2
Darc
Brunswick
Melbourne
M1
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under CC BY SA.

Isochrone Map

Visualisation of train station isochrone maps.

- Isochrone map for 330 train stations.
- Area within each polygon represents 10, 20, 30, 40 and 50 minute walking distances respectively to each train station.
- Blue circles represent suburb centroids within 40 min walking distance.
- Created new information for each property.

Isochrone Map of Train Station "52160"



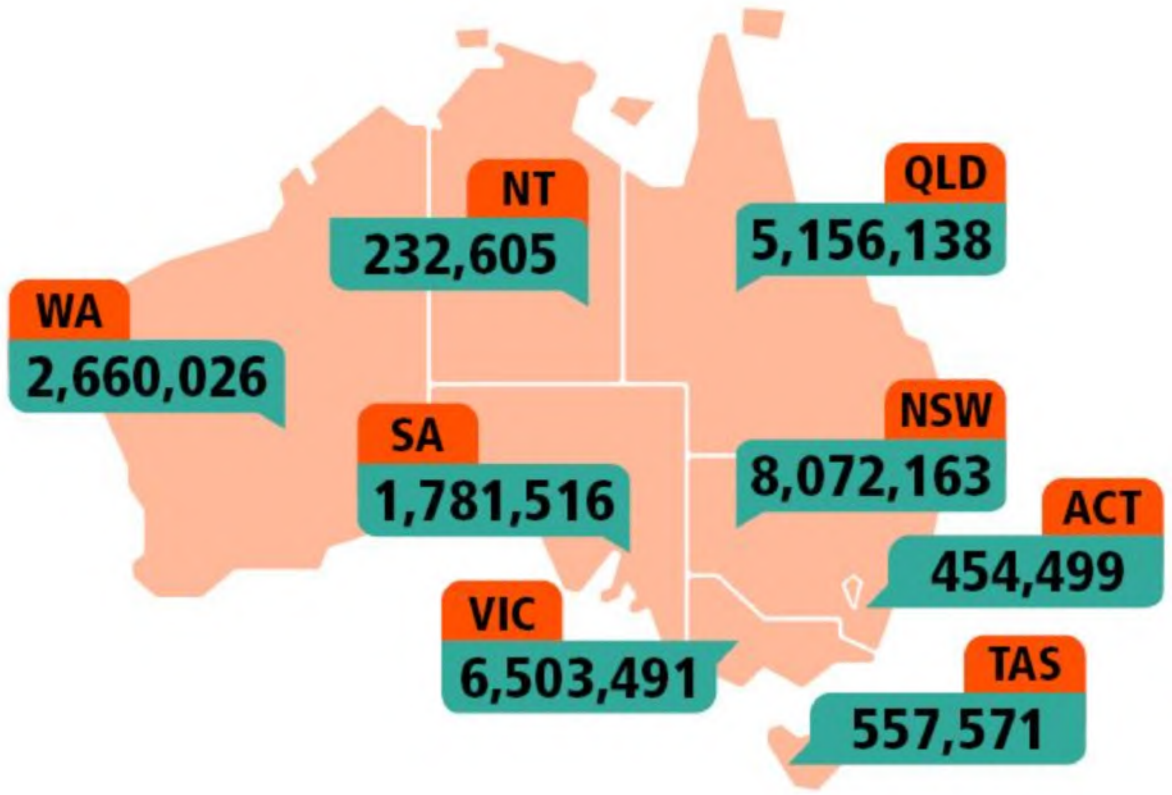
Property Data

Data for rental properties across Victoria were scraped from domain.com.au. Data was retrieved by suburb, and features such as:

- Number of Beds
- Number of Bathrooms
- Number of Parking Spaces
- Longitude and Latitude
- Property Type
- Description Header

Were gathered using regex. As Domain typically imposes a 50 page result limit on queries, we needed to divide the data into smaller portions to download piece by piece. To accurately obtain URLs inline with the naming scheme for suburbs used by Domain, a script was made to make use of the 'autocomplete' feature present within the website- this allowed for a significant increase in the volume of data able to be retrieved.

Census Data



Census population count by state and territory 2021,
Australian Bureau of Statistics

- Census data collected from ABS (Australian Bureau of Statistics) for 2011, 2016 and 2021
- Total **population** for each SA2 code calculated using the number of people in each income group including all genders
- Median **weekly income** data for each household, family and person

Pre-processing External Datasets

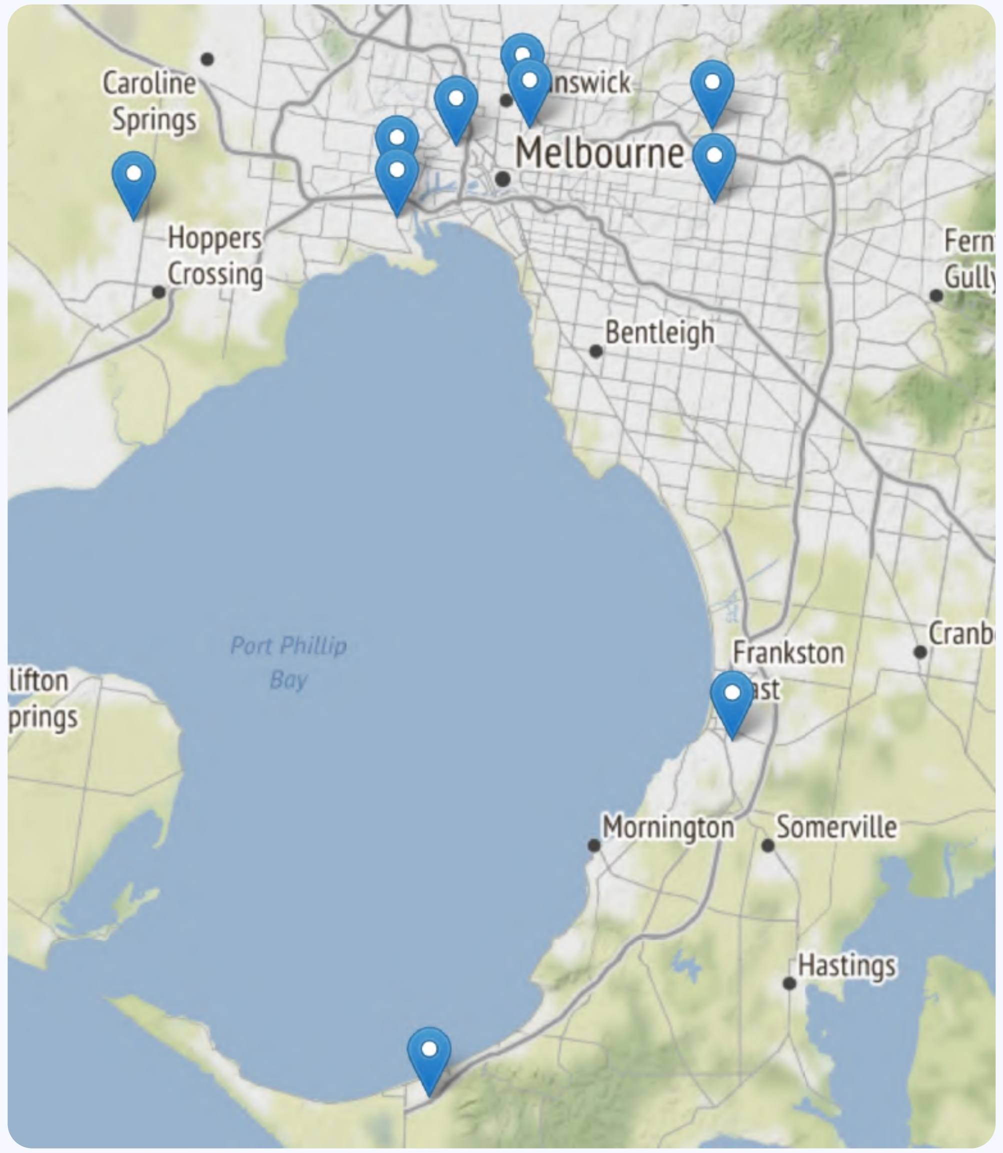
- Checked the distribution of the variables
- Removed redundant variables
- Created new variables where necessary
- Removed outliers based on inter-quartile range (IQR)
- Removed negative and zero values (i.e. areas with zero population- Royal Botanic Gardens, Moorabbin Airport, Essendon Airport)
- Mapped with the suburb geospatial data to create visualisations and further analysis

Population

Areas with the highest population:

- Doncaster
- Tarneit - Central
- Rosebud - McCrae
- Frankston
- Box Hill

Locations with highest population

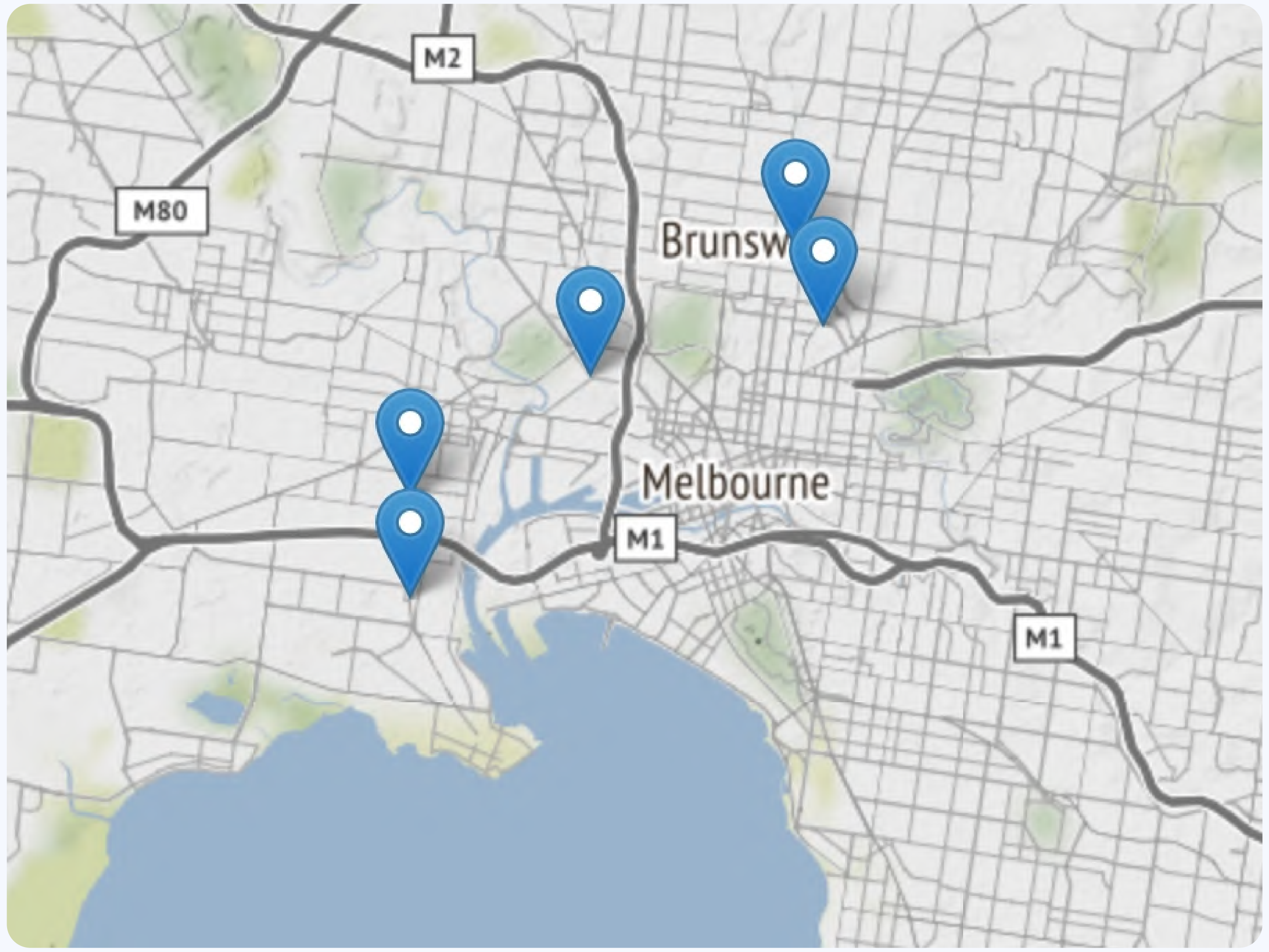


Income

Suburbs with highest median personal income per week:

- Yarraville
- Kensington
- Brunswich East
- Fitzroy North
- Newport

Locations with maximum median personal income

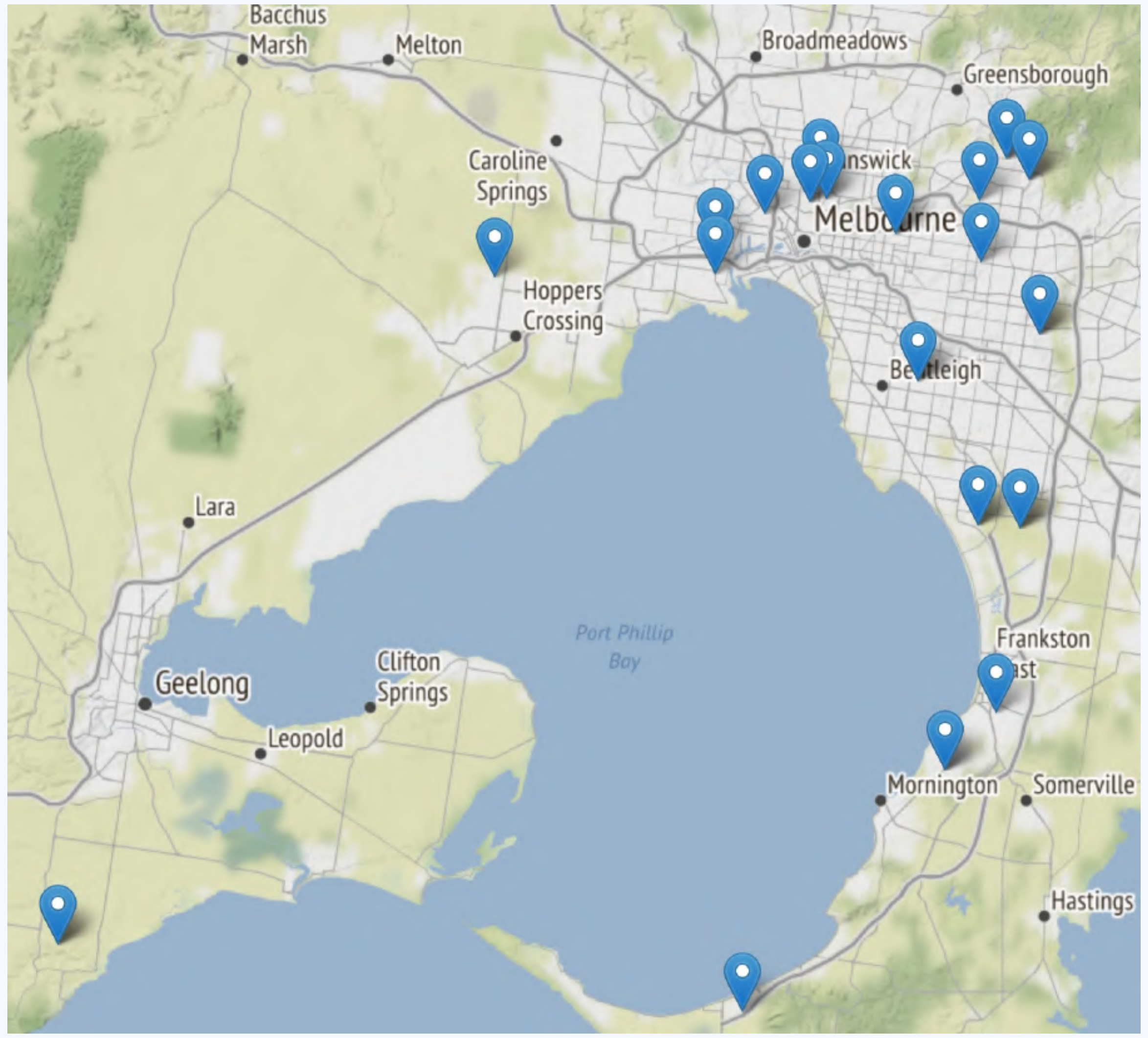


Median Weekly Rent

Suburbs with highest median rent per week:

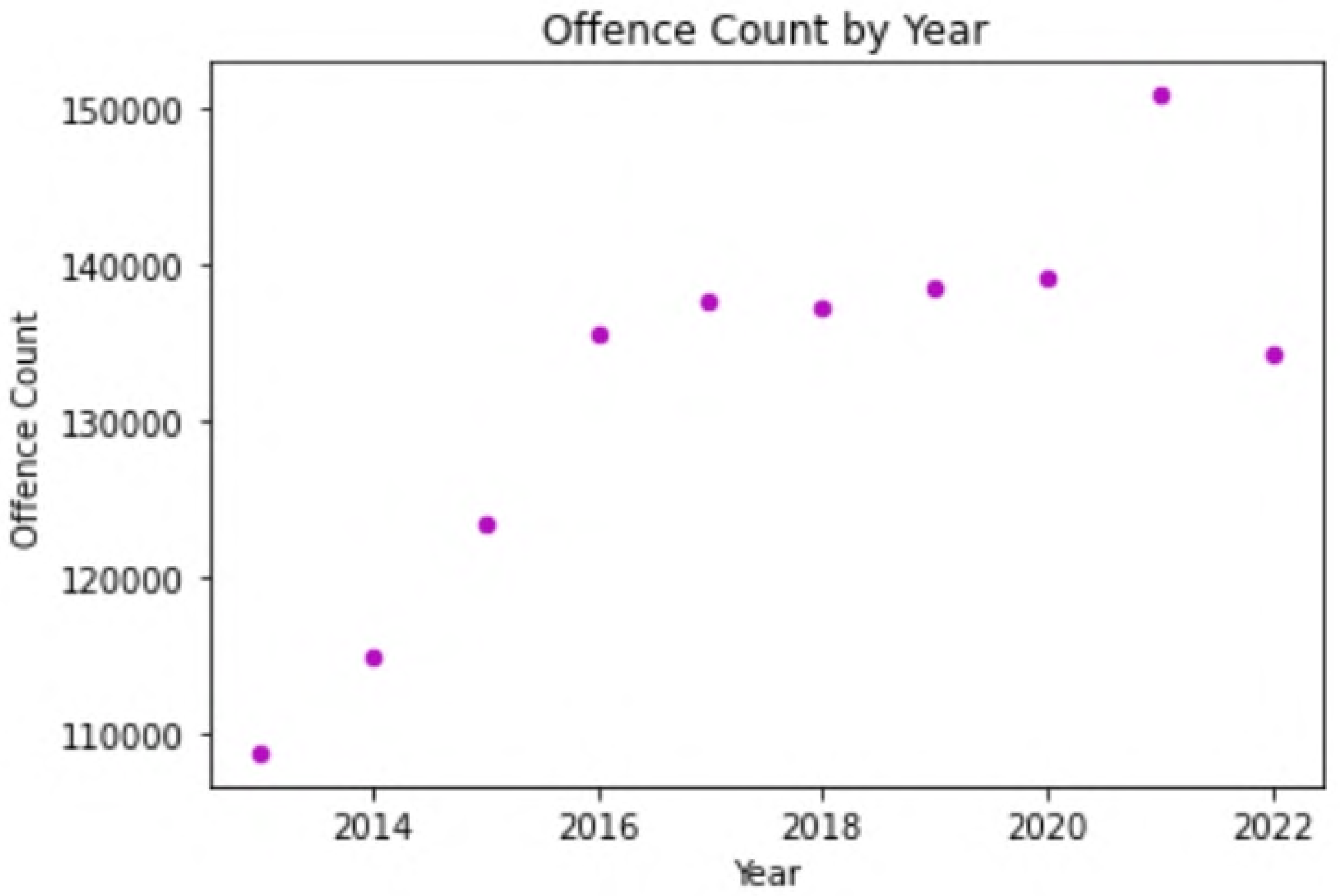
- Carlton North - Princess Hill
- Kensington - South
- Templestowe
- Mount Eliza
- Aspendale Gardens

Suburbs with the highest median rent per week (AUD)



Crime Data

- Data from 2013-2022 which ties suburbs to total offence counts.
- We observe an increase in crime by year, with a peak in 2021. However, keep in mind 2022 is not yet over, and we may see a peak in the holiday period.
- **We hypothesised that higher crime rates would be associated with a lower rental cost.** Crime is generally known to have a negative impact on the rental market.
- We see the growth in crime is not perfectly linear.



Crime Observations

1. Top 20 suburbs with the highest crime rates across 2013-2022, with the top 5 being:

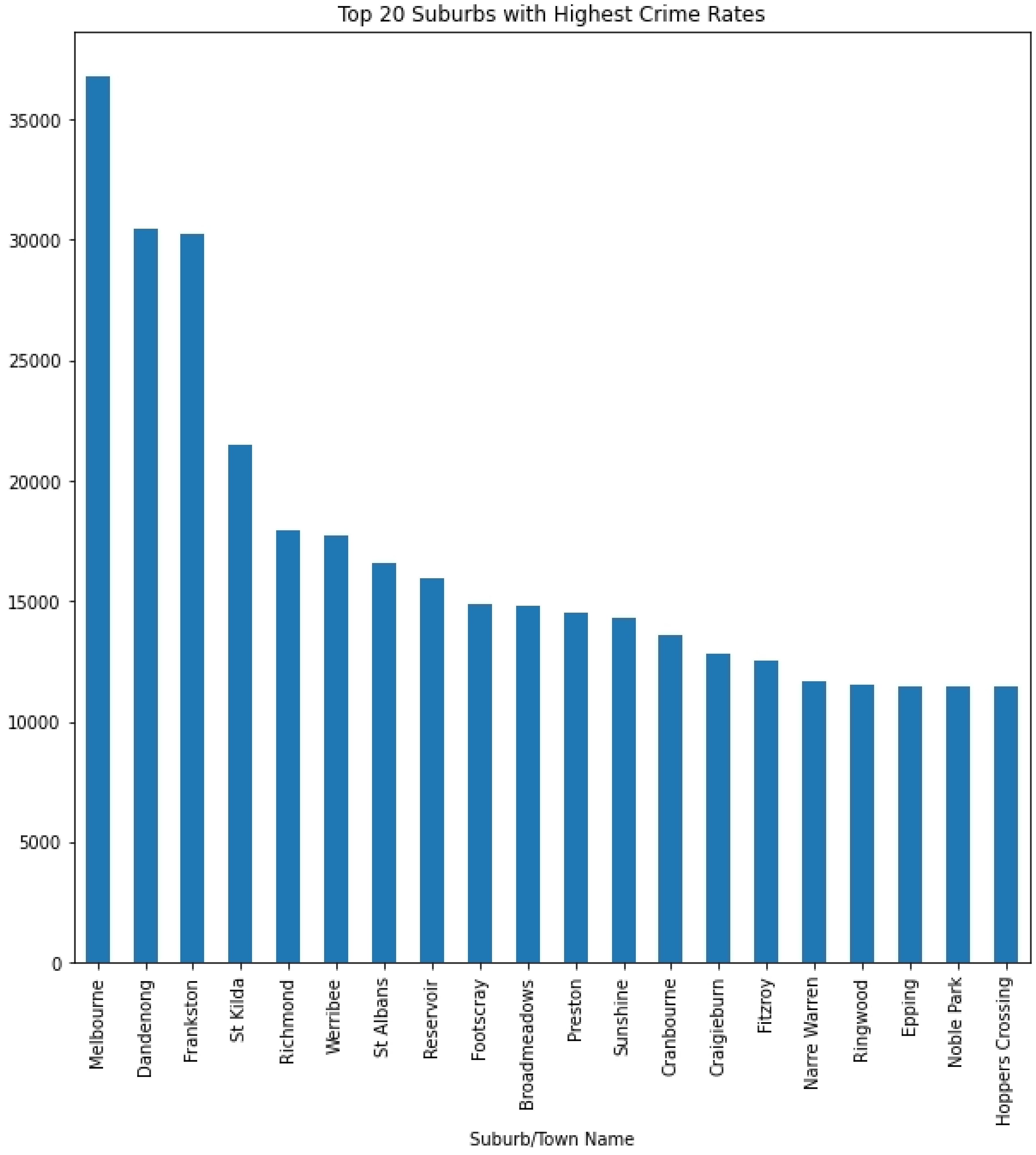
1) Melbourne

2) Dandenong

3) Frankston

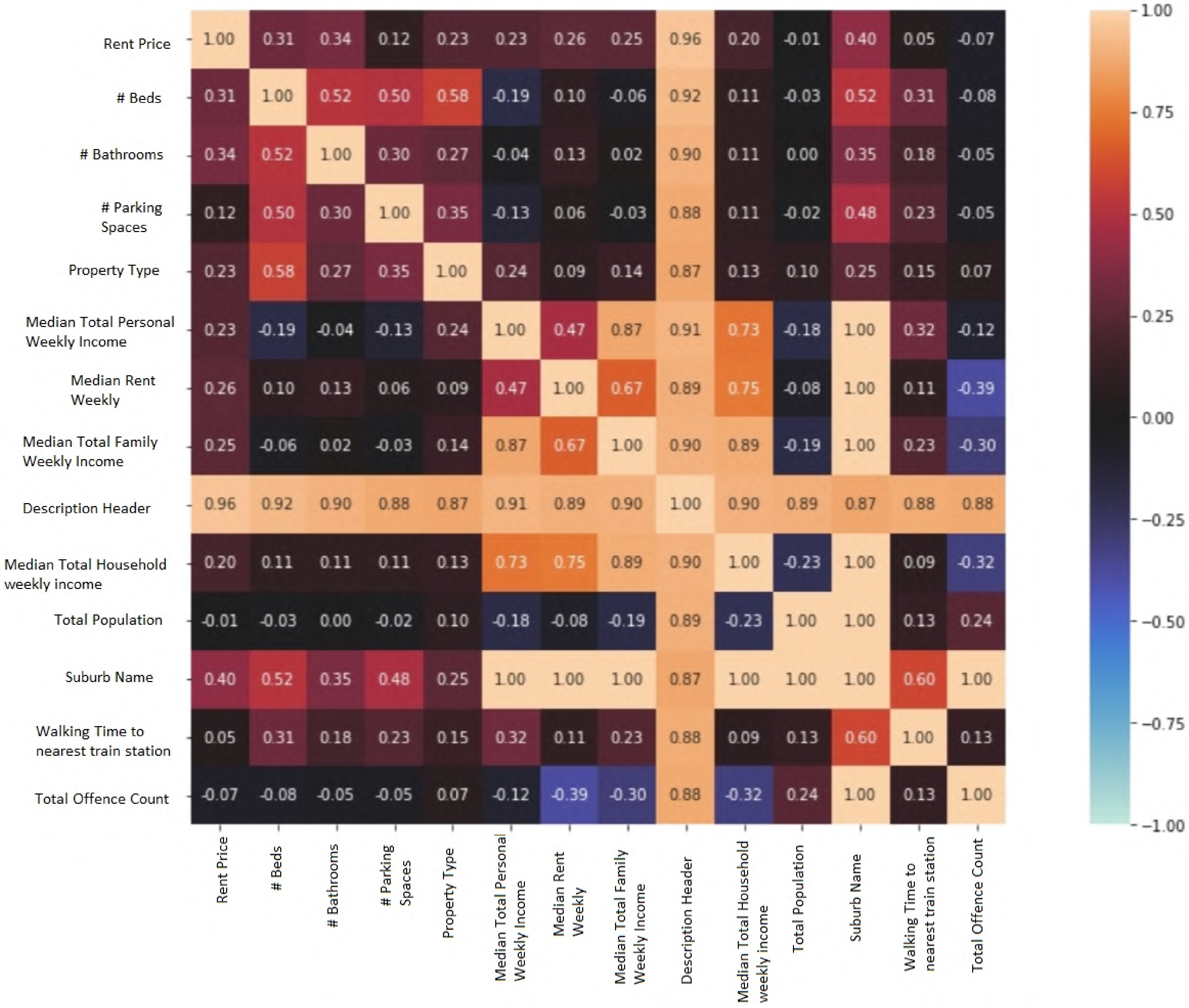
4) St Kilda

5) Richmond



Correlation Analysis

- We see the description header is highly correlated with every variable due to its uniqueness in the dataset. However, we chose to keep it to demonstrate its utility for future feature improvements - e.g. implementation of natural language processing algorithm
- **Interestingly**, we do not see much correlation between the target variable rent price and the predictor variables, which may lead to less than desired model performance
- **Contrary to** our hypothesis, we also didn't notice a significant correlation between crime (offence count) and rent price.
- However, we do notice **slight correlations** between the predictor cost variable and the number of beds and baths (0.31 and 0.34 respectively)



	Rent Price	# Beds	# Bathrooms	# Parking Spaces	Property Type	Median Total Personal Weekly Income	Median Rent Weekly	Median Total Family Weekly Income	Description Header	Median Total Household weekly income	Total Population	Suburb Name	Walking Time to nearest train station	Total Offence Count
Rent Price	1.00	0.31	0.34	0.12	0.23	0.23	0.26	0.25	0.96	0.20	-0.01	0.40	0.05	-0.07
# Beds	0.31	1.00	0.52	0.50	0.58	-0.19	0.10	-0.06	0.92	0.11	-0.03	0.52	0.31	-0.08
# Bathrooms	0.34	0.52	1.00	0.30	0.27	-0.04	0.13	0.02	0.90	0.11	0.00	0.35	0.18	-0.05
# Parking Spaces	0.12	0.50	0.30	1.00	0.35	-0.13	0.06	-0.03	0.88	0.11	-0.02	0.48	0.23	-0.05
Property Type	0.23	0.58	0.27	0.35	1.00	0.24	0.09	0.14	0.87	0.13	0.10	0.25	0.15	0.07
Median Total Personal Weekly Income	0.23	-0.19	-0.04	-0.13	0.24	1.00	0.47	0.87	0.91	0.73	-0.18	1.00	0.32	-0.12
Median Rent Weekly	0.26	0.10	0.13	0.06	0.09	0.47	1.00	0.67	0.89	0.75	-0.08	1.00	0.11	-0.39
Median Total Family Weekly Income	0.25	-0.06	0.02	-0.03	0.14	0.87	0.67	1.00	0.90	0.89	-0.19	1.00	0.23	-0.30
Description Header	0.96	0.92	0.90	0.88	0.87	0.91	0.89	0.90	1.00	0.90	0.89	0.87	0.88	0.88
Median Total Household weekly income	0.20	0.11	0.11	0.11	0.13	0.73	0.75	0.89	0.90	1.00	-0.23	1.00	0.09	-0.32
Total Population	-0.01	-0.03	0.00	-0.02	0.10	-0.18	-0.08	-0.19	0.89	-0.23	1.00	1.00	0.13	0.24
Suburb Name	0.40	0.52	0.35	0.48	0.25	1.00	1.00	1.00	0.87	1.00	1.00	1.00	0.60	1.00
Walking Time to nearest train station	0.05	0.31	0.18	0.23	0.15	0.32	0.11	0.23	0.88	0.09	0.13	0.60	1.00	0.13
Total Offence Count	-0.07	-0.08	-0.05	-0.05	0.07	-0.12	-0.39	-0.30	0.88	-0.32	0.24	1.00	0.13	1.00

Most Liveable & Affordable Suburbs

- Plan Melbourne 2017-2050 introduces the idea of "20-minute neighbourhoods".
- Suburbs which are a 20-minute walk to the nearest train station.
- Low crime rate suburbs are in the 25th percentile for crime rates.
- High median weekly family income is considered any income above the mean.

Rank	SA2 Names
1st	Campbellfield - Coolaroo
2nd	Upwey - Tecoma
3rd	Laverton
4th	South Morang
5th	Fawkner
6th	Alphington - Fairfield
7th	Clayton South
8th	Essendon (West) - Aberfeldie
9th	West Melbourne - Residential
10th	Altona

Top 10 Most "Liveable" Suburbs

Assumptions

- Due to the lack of historical property data, we assumed that with time, the rental prices of suburbs tend to shift towards that of expensive areas such as the CBD.
- In order to perform a prediction, we also operated under the assumption that the growth in external factors such as income, population and crime, are intrinsically linked to the growth in rent prices.
- In addition, we also assumed that the growth of external features (crime, population, and income) is linear.

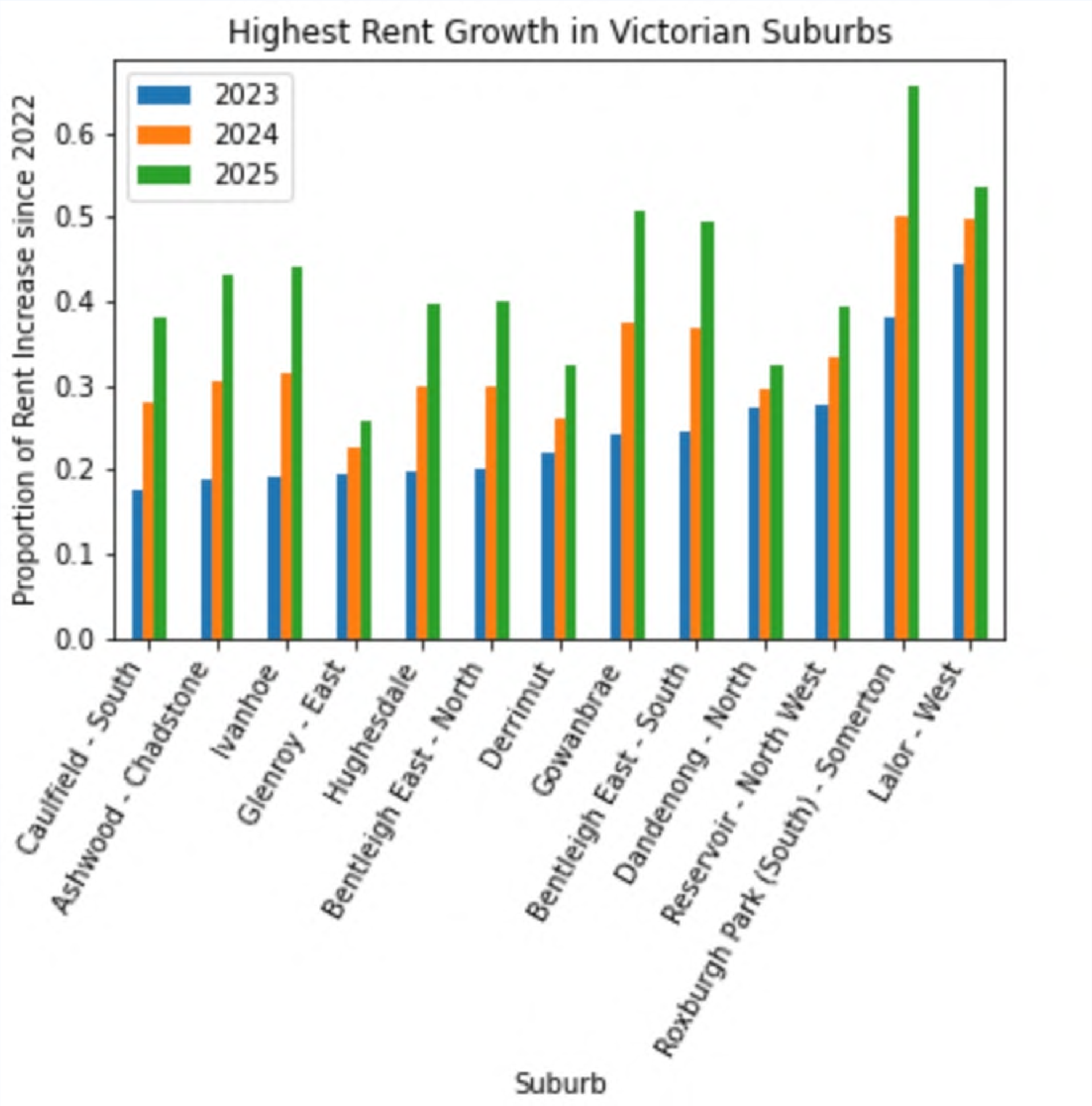
Which 10 suburbs have the highest predicted growth rate?

- In order to determine which suburbs have the highest predicted rent growth, we first determined the average property for each Victorian suburb. We then feed all of these into our model, and output a prediction of the house's price in 2022.

- We then take the same average house, but with the predicted values for the years of 2023-5 for our external datasets (crime, population, and income), feeding these into the model.

- The proportional growth of these values gives us an estimate of which Victorian suburbs will have the highest rent growth.

- The figure displays our results, as predicted by the Neural Network.



Limitations and Difficulties

Limitation	Explanation	Possible Improvements
Lack of Historical Housing Data	As only currently available properties are listed on Domain, we could not gather rental data from previous months/years.	Using a site that has historical rent data which is easily accessible, or scraping data across a few different months may lead to better growth models.
Low property numbers in more rural suburbs	For regional suburbs in particular, there were very few data points to construct a model from. This made predictions in those suburbs have high variability.	Relying on past data for these suburbs, or using bootstrapping to increase the counts for these rare instances may greatly improve the predictions.
Highly extrapolated predictions	The above issues of low data for some suburbs as well as a lack of historical housing data means that there is a significant amount of extrapolation	Using a greater range of data for both the internal and external datasets may work to reduce some variability in the predictions.
Additional features that were not considered	There were features such as the description header which were scraped, but not used as a feature in the model	Applying a scoring system to perform a numeric sentiment analysis may provide better modelling results, as certain words such as 'luxurious' will be correlated with higher rent prices
Lack of API requests	We could not create isochrone maps for every form of public transport as we were restricted to 500 isochrone requests a day.	Split the 27,000+ public transport stops and stations into buckets of size 500 and have all members run the API once a day.

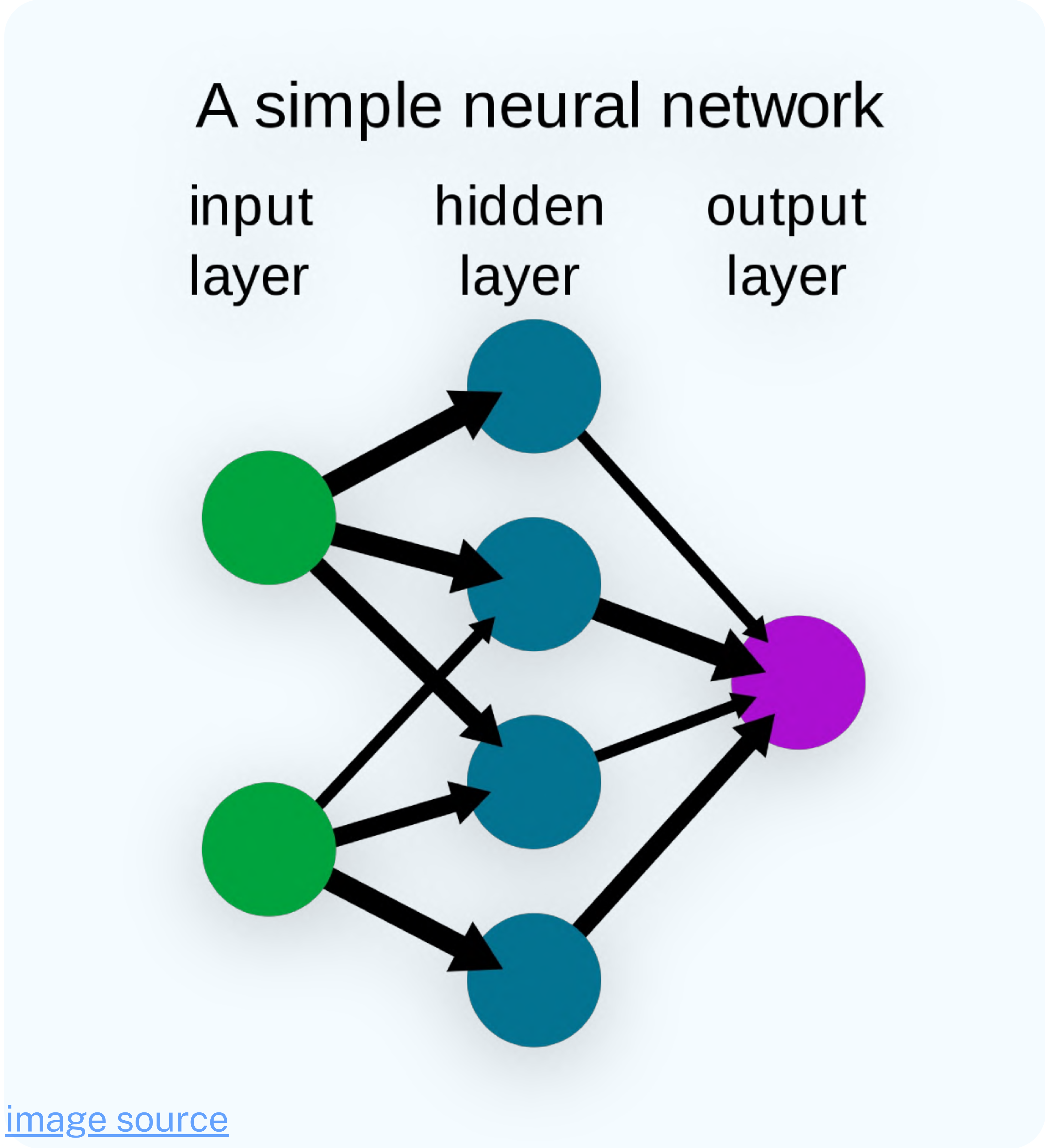


image source

Modelling

Understanding our data

- Basic Linear Models for external datasets to predict growth by suburb
- Neural Network over entire property dataset
- XGBoost Tree as an alternative regressor over dataset

Neural Network

A Multi-Layer Perceptron Model

- A Multi-Layer Perceptron Regressor was applied to the full property dataset
- The structure of the resulting model is depicted in the diagram.
- When the full dataset was put into the regressor, the model only had a coefficient of determination of 0.31.
- Likely due to the large amount of variation and noise in the data, this first model was not able to accurately determine the rent price.

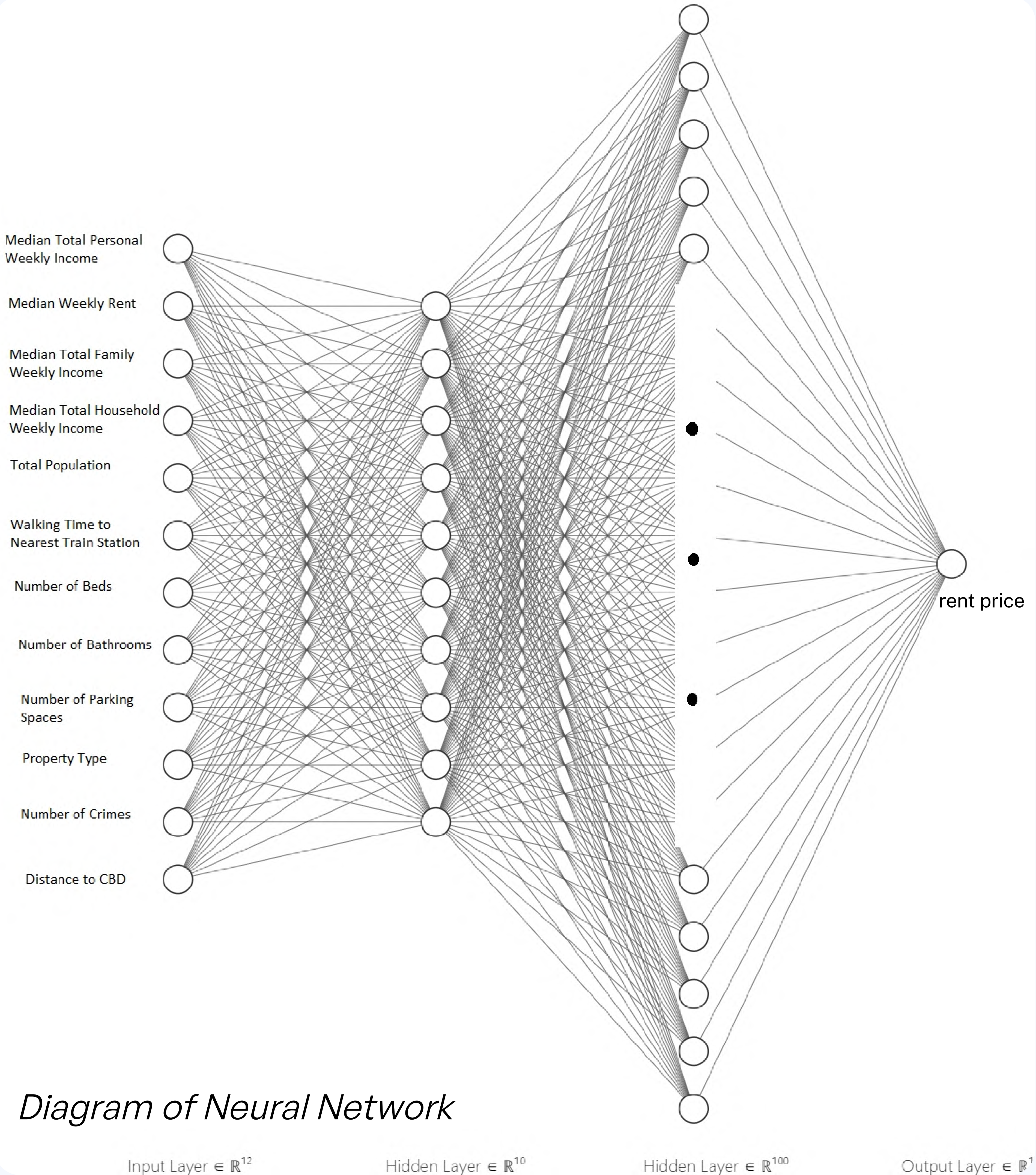

Median Total Personal Weekly Income
Median Weekly Rent
Median Total Family Weekly Income
Median Total Household Weekly Income
Total Population
Walking Time to Nearest Train Station
Number of Beds
Number of Bathrooms
Number of Parking Spaces
Property Type
Number of Crimes
Distance to CBD
rent price
Input Layer $\in \mathbb{R}^{12}$
Hidden Layer $\in \mathbb{R}^{10}$
Hidden Layer $\in \mathbb{R}^{100}$
Output Layer $\in \mathbb{R}^{1}$

Diagram of Neural Network

Neural Network

A Multi-Layer Perceptron Model

- In order to reduce the noisiness of the data, the data was instead split into 4 smaller datasets based on distance from the CBD. A model was then trained on each of these datasets.

- Though the coefficient of determination still remained low for the areas closest to the CBD (R^2 = 0.22), the second 'ring' had a R^2 of 0.4, and the third and fourth rings had a R^2 of 0.5.

- It is likely that the reason our first ring has such a low R^2 is due to our assumption in modelling- as we are assuming that areas slowly shift closer to the rent prices of high rent suburbs like the CBD, our model does not have a good understanding of how the CBD itself should grow with time.

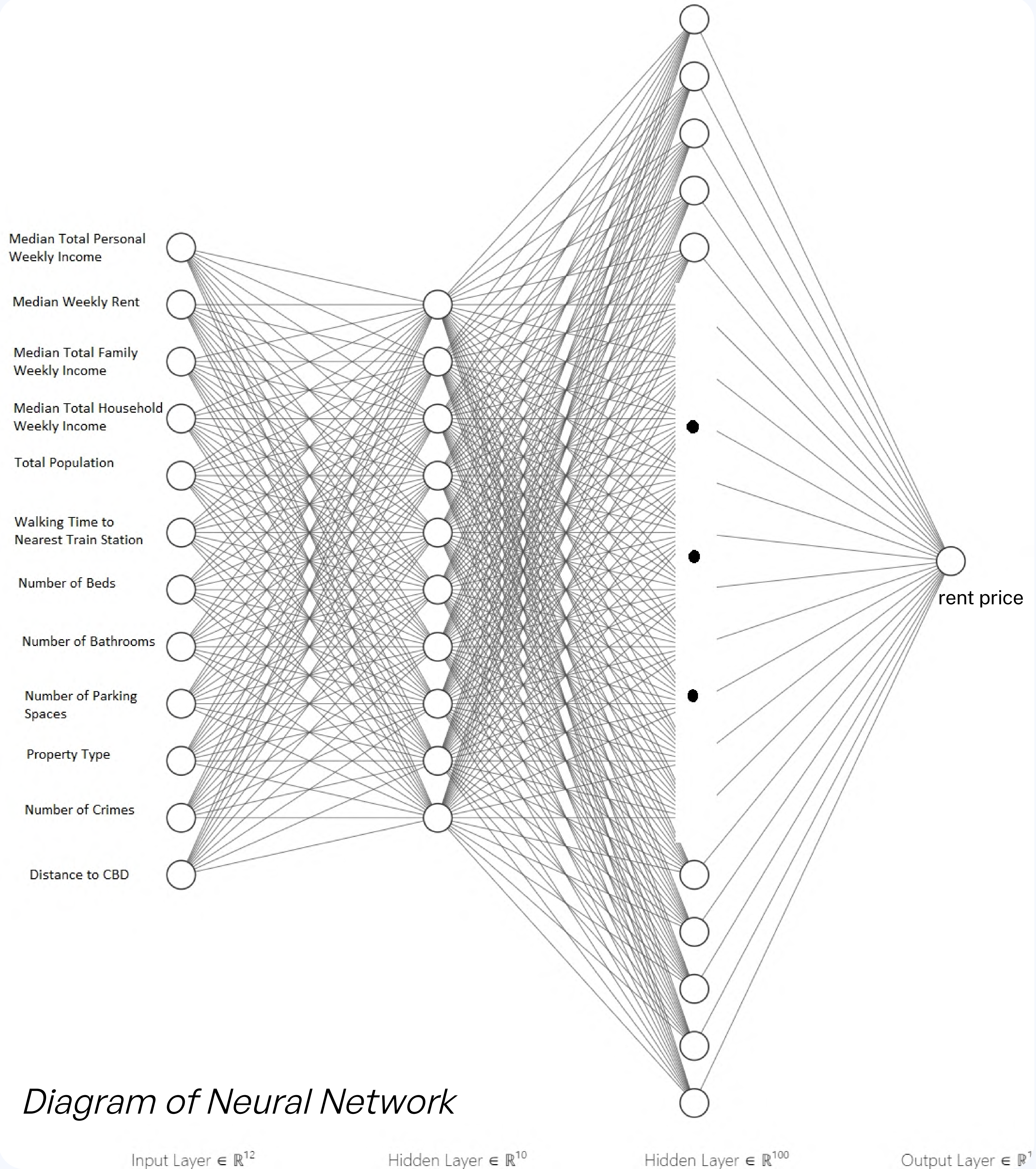


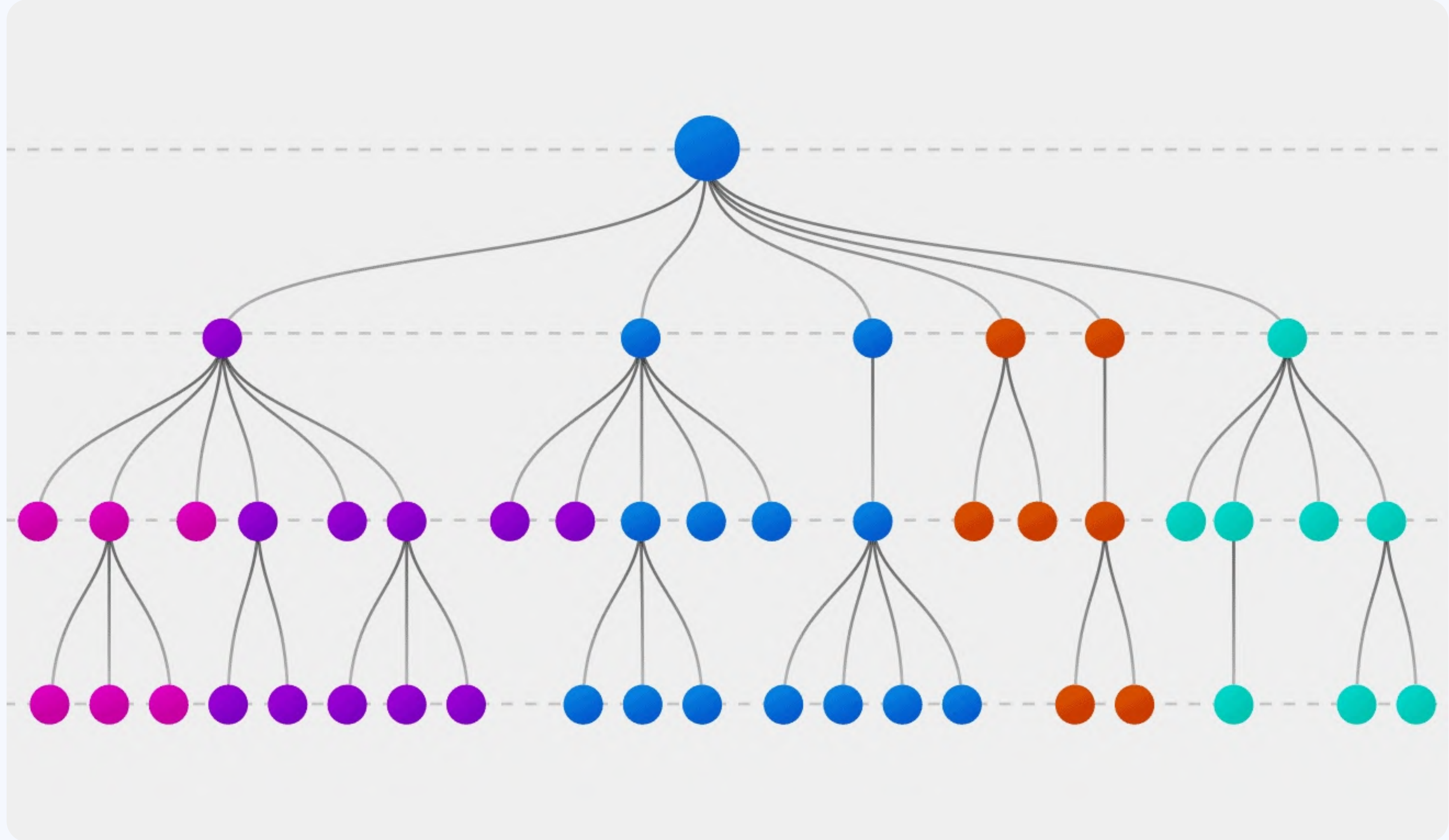
Diagram of Neural Network

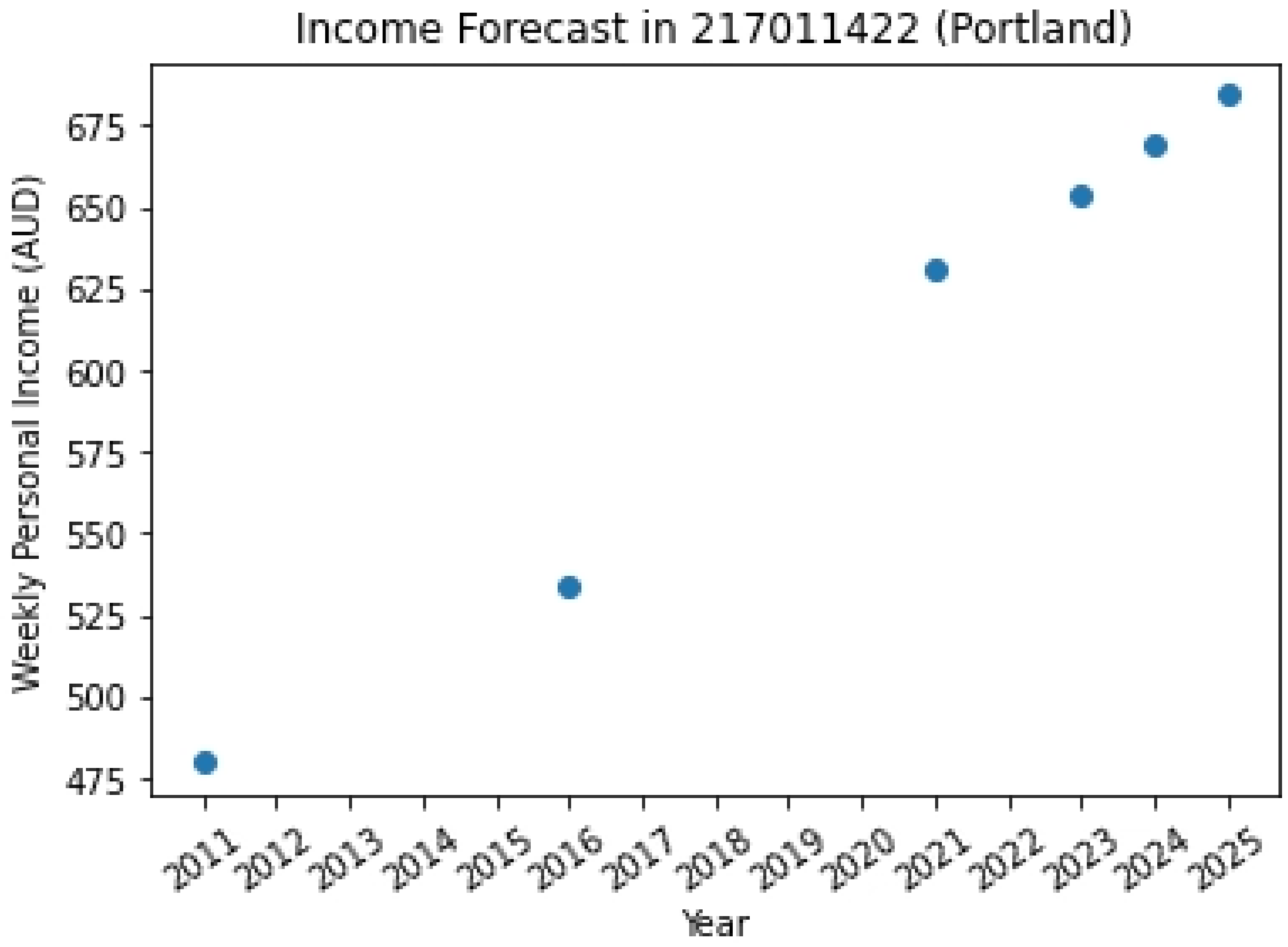
XGBoost

A gradient boosted trees model.

- Implementation of gradient-boosted decision trees designed for speed and performance.
- Decision trees perform well with highly correlated data.
- Outperformed by neural network model.

Decision Trees Clasiffier





Linear Model

Simple Linear Regression

For each suburb (SA2), a simple linear regression was implemented to forecast the values in 2023-5, of:

- Crime
- Income
- Population

based on the available data from current and past years we acquired. We found that according to our linear model, all features increase in the future, unsurprisingly.

Recommendations

What should we do in the future?

Invest in high rent growth suburbs

Investment in high rent growth suburbs may lead to a significant return on investment.

- Maximising return on investment by choosing to invest in a suburb that will experience high growth will be quite profitable to investors

Plan home purchases based on predicted growth

First home-buyers/ renters may be shocked by the sudden increase in price in certain suburbs if rent growth is not considered.

- Taking rent growth into account when purchasing can help younger individuals have a better plan for moving into a new home
- Even for older people, being able to have a solid foundation for estimating the expenses involved in moving to a new area is important.

References

- https://escholarship.org/uc/item/0h04h8ms (California Rental Price Prediction Using Machine Learning Algorithms - Literature Review)
- https://towardsdatascience.com/ai-and-real-state-renting-in-amsterdam-part-1-5fce18238dbc (Machine Learning and Real State: Predicting Rental Prices in Amsterdam)
- https://medium.com/@knoldus/how-to-find-correlation-value-of-categorical-variables-23de7e7a9e26 (Guide on using Dython to find correlations between categorical and continuous variables)
- https://www.crimestatistics.vic.gov.au/crime-statistics/latest-victorian-crime-data/download-data (crime data)
- https://www.domain.com.au/ (Real Estate Data)
- https://openrouteservice.org/ (Open Route Service)
- https://www.abs.gov.au/statistics (ABS Statistics for Census Data)
- https://www.ptv.vic.gov.au/footer/data-and-reporting/ (PTV Data)