Table of Contents

Exploring Urban Scaling Laws in the Unique Context of Western Australia: A Comprehensive Analysis

Interpretation and Discussion

Import Necessary libraries and Load dataset

first of all, I set up and loaded several R packages that are involved in data manipulation, analysis, visualization and Machine learning. Next, I imported a dataset named "wa_dataset.xlsx" using the rio:-Import function. This dataset mostly consisted of data about Western Australia. Installing and loading the packages in R makes ready for carry on data analysis, exploration, and modeling using R.

Data Preprocessing and Exploratory data analysis

In thisSection, I performed data processing tasks on the main dataframe named dataset.

First, I took only specific columns from the dataset and saved them in a new dataframe which was given the name of regions. These variables are the 1st, 2nd, and 44th elements of the given dataset. After that, I saved this subset of data into a file named "regions.csv", which is CSV file.

For the next step I eliminated the 1st, 2nd, and 44th columns in the main dataset using negative indexing and stored the remaining columns in a new data frame named data. Such columns can be understood as the measures that one wants to examine. I proceeded to export this modified dataset into a CSV file called "data.csv".

Additionally, I extract the names of all indicators out of the data frame and saved them to a variable named indicator_names. The purpose of this variable is to designate the names of the columns that represent the various indicators of the dataset.

On the whole, this enabled me to single out distinct groups of data within the big dataset and also to eliminate columns that were not necessary for further examination of the data.

Scatter Plots for Exploratory Data Analysis

I plotted scatter plots for each indicator variable which were present in the dataset.

Initially, I verified the existence of the directory "NEW". Otherwise, I developed one to save the produced plots.

After that, I calculated the log-transformed densities of the population and the corresponding indicator variable for each of the indicator variables. The logarithms of both population and indicator densities, which were adjusted by the land area, were utilized to determine these densities.

Then, I prepared the data for plotting, excluding rows with infinite values.

I employed the ggplot2 function to generate scatter plots in which log-transformed population density is on the x-axis and log-transformed indicator density on the y-axis. A plot of each was with blue dots for the data and the red line fitted with the linear regression method.

The title of every plot given the name of the indicators and axes were appropriately labeled.

Lastly, I saved each plot as a .jpg file in the "NEW" directory and named them according to the indicator title.

In general, scatter plots shows the relationship between population density and various parameters, which was the basis of the exploratory data analysis.

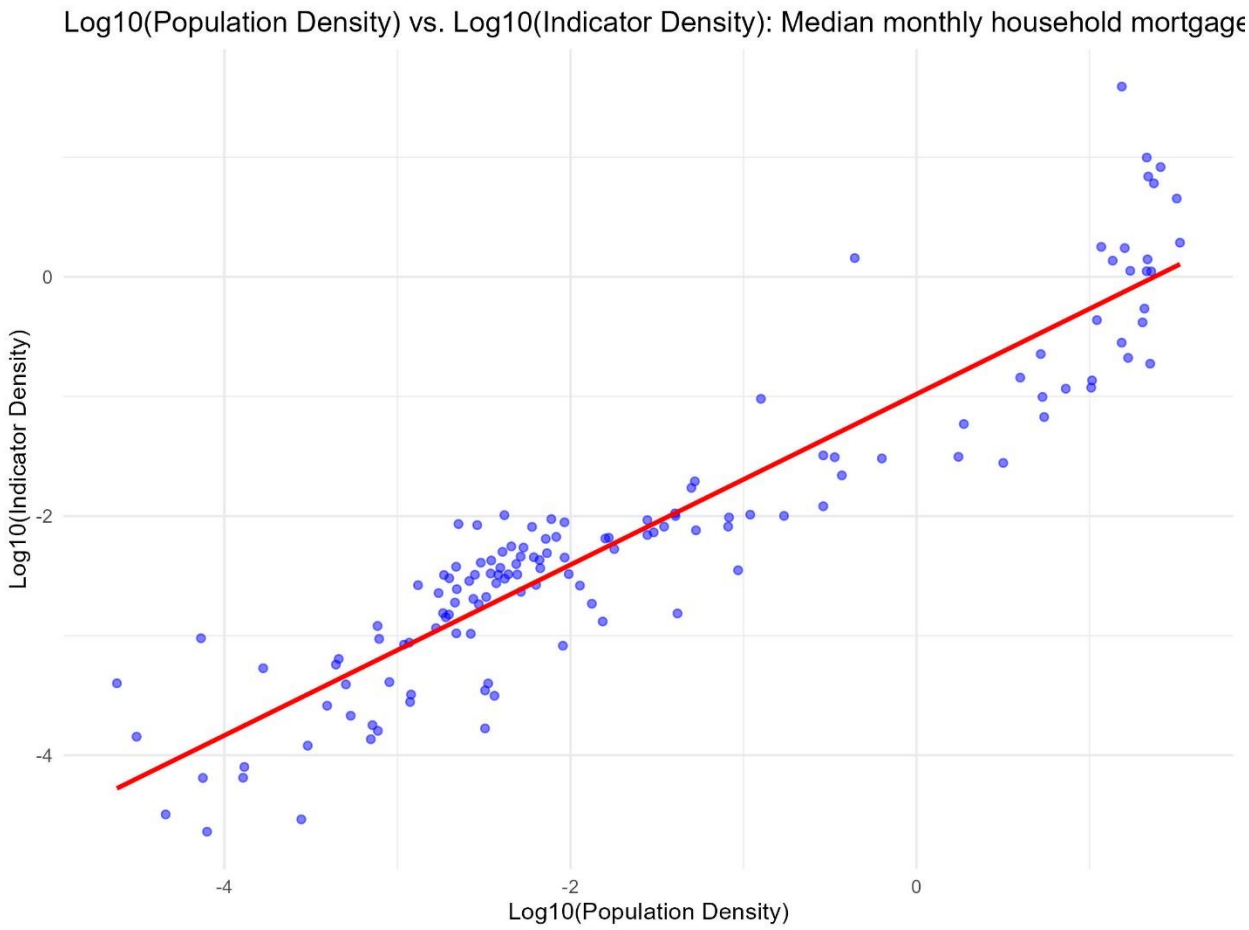
Insights from Scatter Plots

Positive Slope: In the case of a red line having a positive slope, it means that there is a positive correlation between the indicator and the population density. With growing population, the indicator density too increases accordingly. In the image below its clear that log population density vs. log median employee income density are correlated.

Negative Slope: A negative slope stands for a negative correlation. The higher the population density, the lower the indicator density may be.

Flat Slope: A flat line tells of a medium relationship between population density and the indicator.
Strength of Correlation: A steeper line indicates a stronger association.

Data Spread: The scatter of the points around the trend line indicates the variability in the relationship



Processing residuals for indicator variables

Calculation of residuals for all explanatory variables over the dataset was made after building up a linear regression model. I first created a result matrix of the dimensions that corresponded to the numbers of rows in the dataset and the indicator variables. Neatly, for each indicator I computed the log10 of population density (x) and log10 of indicator density (y). Then, instructing 'lm()' on the behavior for the infinity values in 'y' and replacing them with 'NA' we fitted a linear model. The output of the model to the matrix named 'resdat' is the residuals. Then, I saved the 'resdat' matrix into a .csv file which I named 'resdat.csv' storing the row names. Through this, I can calculate the difference between the actual and the expected results for every independent variable.

Insights from Histograms:

Normality: Ideally, the histograms have to be like a bell-shaped curve ("normal distribution"), so that the residuals would be scattered at random around zero. It is implied that the trends are the main focus of the linear models and capture the visualized patterns.

Non-normality: Skewed or odd histograms, especially, might suggest violations of the assumptions of linear regression.

Outliers: Off-the-charts residuals may flag outliers or issues with input data. The subsequent study may be done.

Essentially, assessing the distribution of residuals informs the model suitability for depicting the relationship between the population density and urban characteristics variables in your Western Australia dataset.

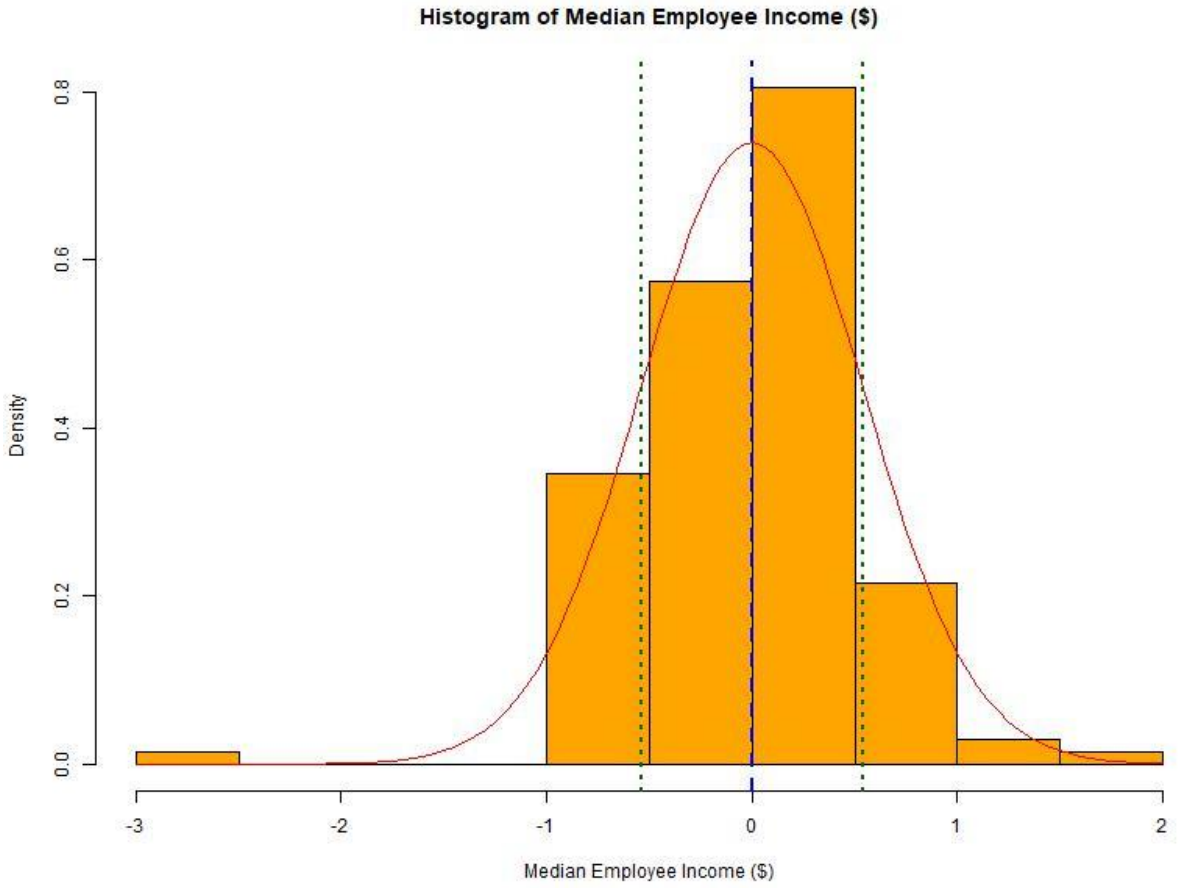
shape of histogram

The perfect shape for a histogram of residuals in linear regression is an ordinary or bell-shaped curve which is also known as normal distribution.

X-axis: Therefore, this means that the number of residuals is the value. In simple words, residuals are the variations between the actual data points and the expected values predicted using the linear fit.

Y-axis: This represents the frequency or count of residuals that fall within a specific range of values on the x-axis.

Overall, the shape of the histogram of residuals gives insights into the distribution of errors in linear models. Ideally, a symmetrical bell shape is preferred for valid regression analysis.



Assessing Homoscedasticity

In this section I used the homoscedasticity assumption to examine the data for each indicator variable. First, I have the directory directory named "VPlot" to store the plots. I subsequently deal with each indicator variable, calculating log-transferred population density and log-transformed indicator density. Then, it was the turn of constructing a plot per each indicator to graph the interrelationship between the fitted values and residuals of a linear regression model.

In order to know if heteroscedasticity has occurred, I used linear regression model and then plotted the residuals against the fitted values. Each plot was named after those indicators and saved as .jpgs in the "VPlot" directory.

Insights From Plot

X-axis: Scaled Values These consist of predicted model values for those indicators I already used (log of population density versus indicator density).

Y-axis: Residuals, Such quantities represent the differences between the measured data and the fitted values.
Scatter Points: This is a collection of individual data points for regions, after excluding missing values.

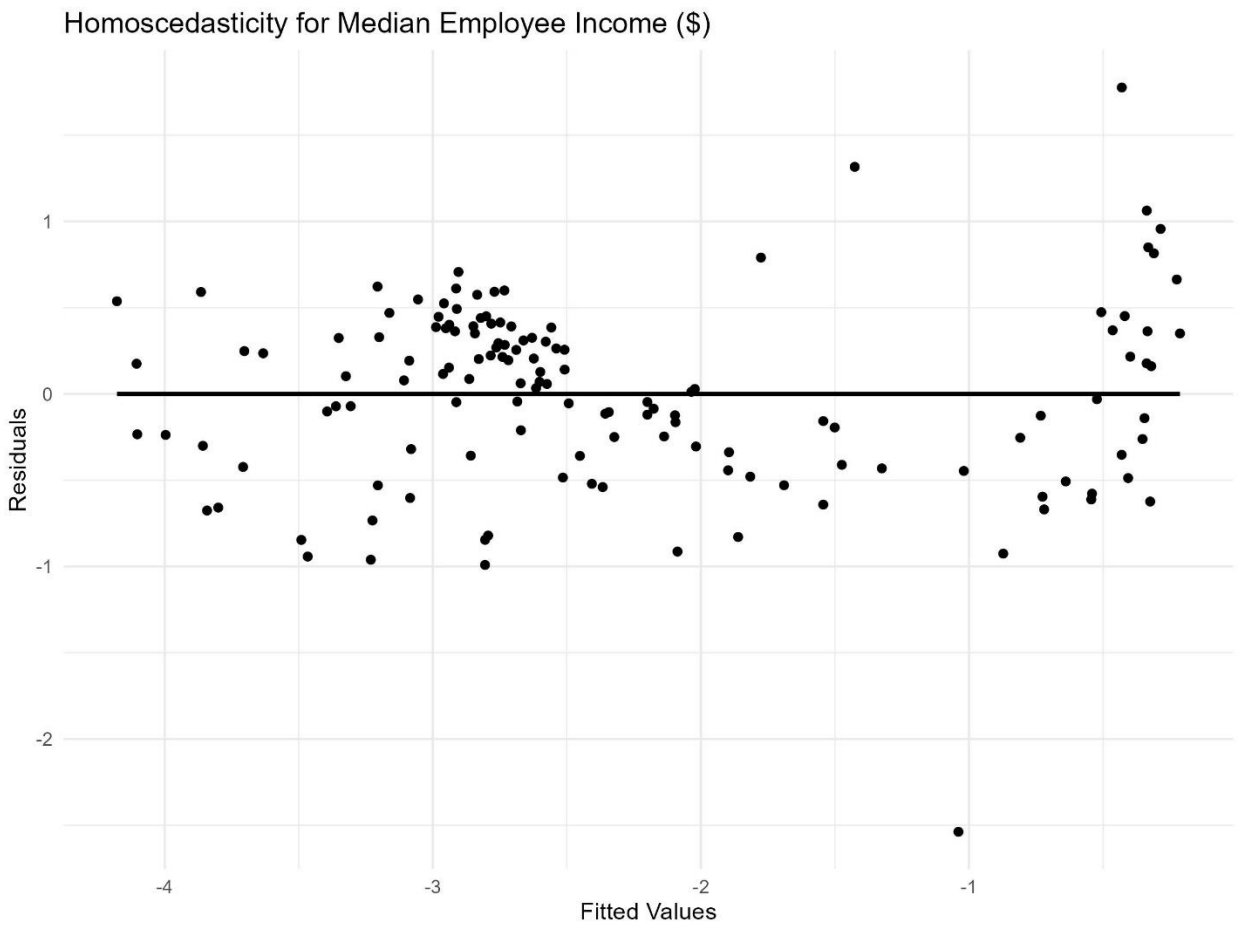
Smooth Line: This would most likely be a linear regression line fitted to the residuals of latex x vs. fitted values.

Shape and Interpretation:

The perfect hope for homoscedasticity (constant variance) is when spread points randomly appear around a horizontal line passing through zero on the y-axis (residuals). This implies that the pattern of the residual's variance is uniform with the change of the fitted value in the data.

Non-random pattern: You can suspect heteroscedasticity (unequal variances) by looking at the plot of residuals and notice a clear pattern, such as widening funnel shape (fitted values increasing) or a U-shape. You may be violating the fundamental regulations of the linear regression and this could negatively influence the reliability of your analysis

Outliers: If there are points significantly different from the main cluster on the residual plot, then they probably represent highly influential points, which can give bias in the outcomes.



Residual Heatmap

I performed the residuals' analysis using the one gained in the previous linear regression modelling.

Firstly, I established a general color range using colorRampPalette function, which provided 100 gradients in the range from very dark blue to reddish-brown with bare white color in the middle.

As a result, I performed the correlation of the residuals (resdat) using Pearson's correlation coefficient. Under the SMAC technique, the residuals could be interpreted to reveal any hidden and underlying patterns or connections among them.

Next, I rotated the residual matrix so that it gave us another matrix of correlations. This allowed for a closer look of the different variables' relationships.

I created a heatmap visualization by using the pheatmap function of ggplotR. This heatmap mapped the relevancy of some indicators based on their residuals.

Then, I saved the heatmap plot in a PNG file that will be called "heatmap.png" helping me with the analysis of the residual patterns and also with the identification of the relationships among the indicators.

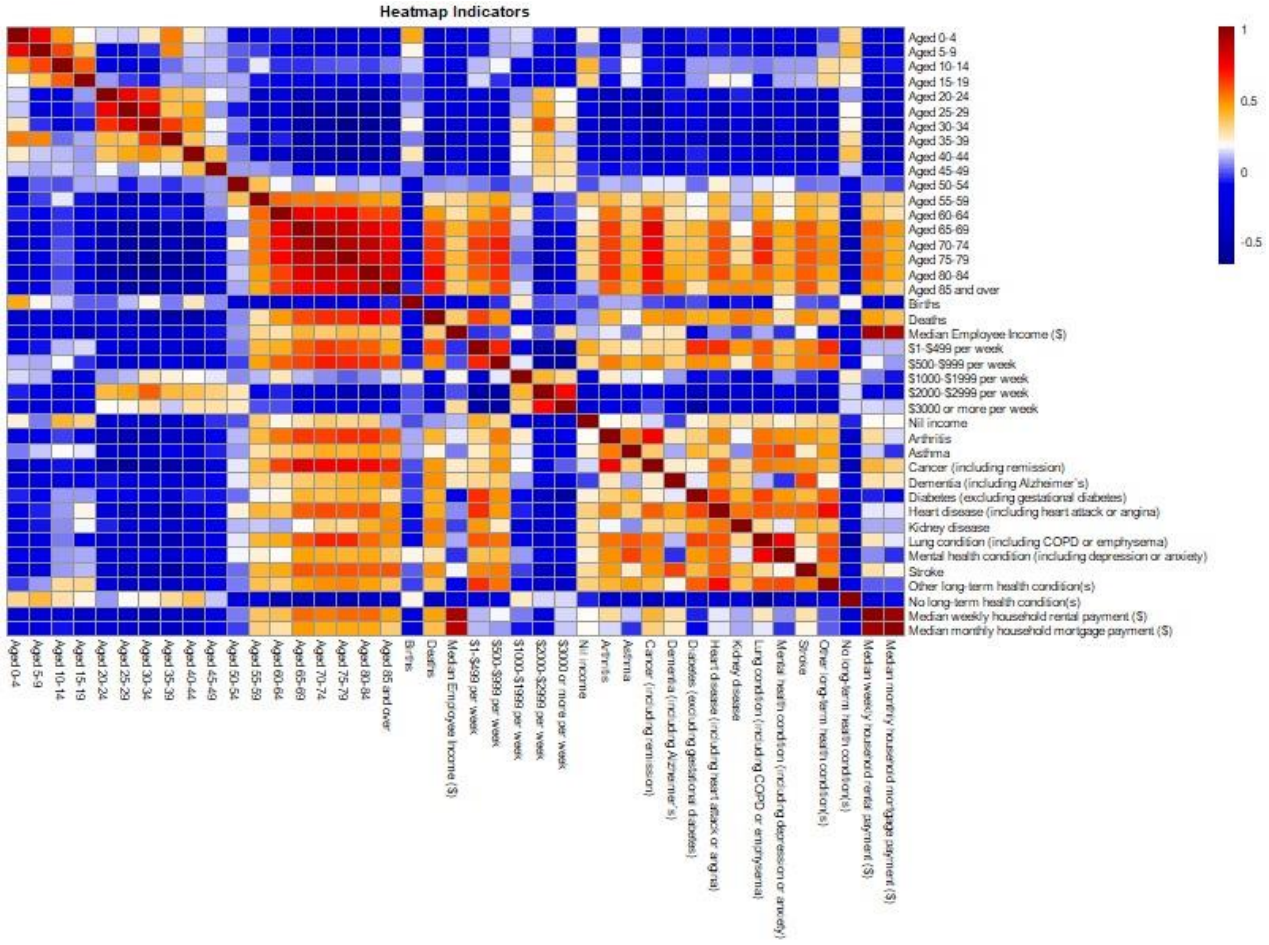
Insights from Heatmap

Grid: Heat map is represented as a grid of squares where each square comprises the correlation coefficient between two residuals of the same indicators.

Color Coding: The color intensity (blue to red) probably denotes the amount and direction of such a connection. "Blue" shows the negative trend (opposite one), "white" means that it is not correlated, and "red" refers to the positive trend.

Correlations Between Residuals: The heatmap helps in locating the degree to which the linear models with respect to different indicators are related from the residuals (errors). High positive or negative which may imply the existence of another factor which probably will be affecting the indicators in the same period.

Grouping Indicators: Pursuing for centered groups of the same color and I already have indicators that shares similar patterns in their residuals. Such an approach could be an effective to determine the pathways by which all the identified indicators are affected.



Scaling Exponent Analysis

In this I focused, in particular, on the finding of the scaling exponent of each indicator's plot, a result of the relationship between population density and any indicator. To start with, a dataframe that is an empty space for keeping coefficient as well as their confidence intervals for every indicator was created. The result of the process of interaction modeling is the log-transformation of particular variables and performing the procedure of linear regression that gives the coefficients and their confidence intervals. The computation shows the intervals of the coefficient for Beta1 under different indicators and hence offered crucial information on the relations of population density to such indicators. This complete method gave a coincidence for a scaling indicator and completing a dataset for the later modeling tasks.

Why Imputing Null Values is necessary for ML models?

imputation is necessary for machine learning models, as it results a model which is either biased or ineffective. Through imputing the missing values, the dataset becomes complete one which the model can use as an instructor to learn from all existing data information. Also, this makes it possible to avoid loss of information and as a result, the make of the model's prediction is more precise and dependable.

Naïve Bayes' Model

This machine learning Naive Bayes classification algorithm to classify the data based on the employee's median income and other variables. T specified this as the target variable "Median Employee Income ($)" and separated it from other predictors. Next, I transformed the predictors and target variable upturn. To meet compatibility in the operation of a Naive Bayes algorithm, I changed the target variable into factor if it still was not a factor. Lastly, I trained a Naive Bayes model using predictors together with a target variable storing the model in the variable naive_bayes_model. These now awaiters models can now be used for prediction of new data points.

Data Stream Mining Algorithms (DSMAs)

I use Data Stream Mining Algorithms (DSMAs), and clustering techniques.

DenStream (DBSCAN-based):

It sets the starting parameters of DBSCAN clustering process, including epsilon (eps) and minPts. Thereafter, it performs DBSCAN clustering to all the numeric columns of the filled-in dataset, identifying clusters as well as noise points. Lastly, the clustering result is printed in a visual presentation, a scatter plot, the image of which is saved in a file.

K-Means:

The number of clusters (K) is specified. K-Means clustering is performed. The data will be divided into K groups by the algorithm, basing on the central points of the clusters and the Euclidean distance between the hypotheses. The printing outcome is there, with cluster centers and assignments for each data point on the given set.

K-Medoids:

This covers the number of clusters (k) and performs the utilization of the Partitioning Around Medoids algorithm (PAM) for the K-Medoids clustering. Unlike K Means, K-Medoids uses actual data points as cluster representatives since these become medoids and therefore the method is more robust to outliers. The result of clustering is printed, that gives the cluster medoids and the cluster handout to each data point.

In general, these properly structured algorithms lead to a discovery of the internal structure of the dataset, which in turn give rise to data clustering and patterns for further exploration and decision-making.

Self-Organizing Maps (SOMs)

I focused on applying Self-Organizing Maps (SOMs) to two different scenarios: social feature study and area-based health and socioeconomic classifications.

Demographic Feature Learning:

First of all, the feature demographic is pulled out by the dataset, namely age distribution. Further, these features are normalized so as to regularize the scaling process. A SOM (Self-Organizing Maps) model is trained on the normalized demographic data by using a defined grid size as an input.
The U-matrix which encodes unit-to-unit topological relations in the SOM grid, is subsequently extracted and reshaped. The unit labels are calculated for the u-matrix by the amount of data points put into each unit. Topographic error and quantization error computations are done to access the model's accuracy with respect to representing data and preserving the topological structure.

Area wise and socioeconomic factors

Features related to health and socioeconomic factors which are considered important for a region are selected. Feature engineering is essential to avoid the 'Area' feature. The rest is normalized for SOM.A SOM model is fitted on the scaled features through a predetermined grid setting. The output error that is printed reviews the model accuracy in representing the input data.

Moreover, the manual calculation of the topological error is employed to control the validity of spatial relationships between gridcells of the SOM grid.

Specific information is delivered, for example, the number of iterations and errors printing to provide knowledge about the performance of SOM profiling health and socioeconomic features across the areas.

Conclusion/Results

Through my extensive research of urban scaling laws in Western Australia, I witnessed a journey of diverse field that covers all the elements which tells the secret about the development of urban towns. At first, we dealt with the raw dataset starting with missing values and removing redundant features, that is, the socio-economic and demographic ones. Machine learning algorithms including Naive Bayes, DBSCAN and K-Means, and (SOMs) Self-Organizing Maps were used for us to go deeper into the hidden patterns and relationships. Visualizations such as scatter plots, heatmaps, and coefficient interval plots made it possible for me to see the associations among indicators variables, demographic features, and population density at a glance. My findings give insight on the complicated factor interactions between the economic situation, place of residence, and health in a western Australian state. Using advanced research, I not only found out regulation laws in metropolitan area but also paved the way for the future research relating to sustainable development of the urban and faced the societal problems of the region.