Statistical Analysis of Aggregated Hedonic Dataset regarding Property Prices in Busan, South Korea

Project Report

Souparna Pal
Priyankar Biswas
Aayusman Mallick
Sayan Dewan

Introduction to Statistics and
Computation with Data

April 17, 2025

Abstract

This project attempts to analyse the data given in 'Aggregated hedonic dataset with a green index: Busan, South Korea' by Sihyun An, Seongeun Bae, Yena Song, Kwangwon Ahn.

The goal of this study is to investigate the association between hedonic variables and property prices in the Busan Metropolitan City of South Korea, focusing on green index, which is a measure of the degree of urban street greenness exposed to residents and pedestrians, introduced in the paper. This report provides a summary of the methods and techniques we used to complete this project, and an analysis conducted using these techniques on the data.

Contents

Introduction

In recent years, rapid urban growth has often come at the cost of natural environments, replaced by built-up infrastructure in cities. However, there has been a growing focus on developing green cities that emphasize environmental well-being. Numerous studies have highlighted the economic benefits of urban greenery, linking it to increased housing prices due to factors like enhanced physical activity and reduced pollution.

The paper[5] provides an aggregated dataset for investigating the association between hedonic variables, and property prices in the Busan Metropolitan City of South Korea. This dataset offers additional value in exploring the relationship between green amenities and housing prices, alongside variables commonly used in housing price assessment models. Hedonic variables include various factors that include property prices such as property characteristics, environmental amenities, local built environments, local demographic characteristics, and seasonal controls.

The paper also introduces the green index, which quantifies the degree of urban street greenness exposed to residents and pedestrians.

We have analysed the relationship between property prices and the hedonic variables using various techniques. We have employed Factor Analysis to identify the underlying factor structure of the variables affecting property prices. We have also conducted regression analysis using the factors obtained previously and using all the variables in an attempt to give a model to predict property prices using the hedonic variables. Finally we have done hypothesis testing using the coefficients obtained from regression to find the relationship between green index and property prices.

Chapter 1

Dataset and Descriptive Statistics

1.1 Dataset Structure

The dataset includes 52,644 observations and 28 variables, making it a valuable resource for benchmarking or exploring urban and real estate studies. It consists of property prices and various hedonic variables, which are categorised into five groups, namely, property characteristics, environmental amenities, local built environment, local demographics, and sales period controls. The green index, introduced in this dataset as a measure of urban greenness, is classified under environmental amenities. The dataset uses demographic variables measured at the 'Dong' level, which is equivalent to an 'Administrative District' in India.
Table 1.1 summarises the hedonic variables, including their names, scales, and details. The variable column presents the column names of the variables in the hedonic dataset and the denominated names in the descriptive statistics. Scale stands for the measurement scale, and Detail describes each variable. Table 1.2 presents the descriptive statistics for the green index and other hedonic variables.

1.2 Acquisition of Source Dataset

Busan was taken as the study area for the paper because of its diverse characteristics, including a large population and various environmental amenities such as waterfronts, seashores, and natural parks. The Ministry of Land, Infrastructure and Transport (MLIT) provided apartment transaction records that include transaction prices, property addresses, and related characteristics. The condominium was focused on as the representative housing type because the MLIT supplies geographic coordinates for transaction points, facilitating spatial analyses, and apartments serve as

Variable	Scale	Detail
Property prices	Ratio	Log-transformed Korean won per square meter (won/m^2)
Longitude	Ratio	Longitude in the Cartesian coordinate system
Latitude	Ratio	Latitude in the Cartesian coordinate system
Property Characteristics		
Size	Ratio	Unit size aggregated in square meters (m^2)
Floor	Interval	A floor of transacted property
Highest Floor	Ratio	Highest floor in an apartment complex
Units	Ratio	Number of households in an apartment complex
Parking	Ratio	Number of parking spaces divided by the number of households
Heating	Nominal	A heating type of each housing: city gas = 0; others = 1
Year	Date	Year of construction of each apartment complex
Environmental Amenities		
Dist. Green	Ratio	Log-transformed network distance to the nearest park, hill, or mountain in meters
Dist. Water	Ratio	Log-transformed network distance to the nearest river, stream, pond, or seashore in meters
Green Index	Ratio	Degree of street greenness exposed to pedestrians
Local Built Environment		
Dist. Subway	Ratio	Log-transformed network distance to the nearest subway station in meters
Bus Stop	Ratio	Number of bus stops within a 400-meter radius of a property
Dist. CBD	Ratio	Network distance to the city hall in meters
Top Univ.	Ratio	Number of Seoul National University entrants from high schools within a 5-km radius of properties
High School	Ratio	Number of high schools within a 5-km radius of a property
Local Demographics		
Sex Ratio	Ratio	Percentage of the number of men divided by the number of women
Population	Ratio	Number of people in a neighborhood
Pop. Density	Ratio	Number of people per square kilometer (km^2)
Higher Degree	Ratio	Percentage of people with higher degrees divided by people aged 15+ years
Young Population	Ratio	Percentage of people aged less than 15 years divided by total population
Median Age	Ratio	Percentage of people aged 15 to 65 years divided by total population
Old Population	Ratio	Percentage of people aged 65+ years divided by total population
Seasonality Control		
Spring	Nominal	Seasonal dummy indicating transaction occurred in March, April, or May
Fall	Nominal	Seasonal dummy indicating transaction occurred in September, October, or November
Winter	Nominal	Seasonal dummy indicating transaction occurred in December, January, or February

Table 1.1: Delineation of hedonic variables

the predominant housing type in South Korea[7, 2].
Raw data for the dataset has been collected from all apartment transaction records in Busan for the years 2018 and 2019. After excluding missing values and outliers through exploratory data analysis (EDA) and descrip-

	Mean	Std.	Min.	Max.
Property prices	10.187	0.568	6.908	12.934
Property Characteristics: Unit-related				
Size	77.820	28.916	12.488	269.680
Floor	11.819	8.201	-1.000	77.000
Property Characteristics: Complex-related				
Units	937.087	885.506	4.000	5239.000
Buildings	9.227	8.910	1.000	77.000
Year	2003.496	10.153	1969.000	2019.000
Heating	0.093	0.291	0.000	1.000
Parking	1.101	0.619	0.000	77.000
Highest Floor	23.249	10.347	2.000	84.000
Environmental Amenities				
Dist. Green	7.277	2.222	0.808	10.714
Dist. Water	6.268	1.181	-0.170	8.601
Green Index	10.733	2.098	4.163	18.927
Local Built Environment				
Dist. Subway	6.892	1.016	3.366	9.978
Bus Stop	18.105	10.840	0.000	63.000
Dist. CBD	237737.312	192493.328	243.860	398856.132
Top Univ.	11.179	6.442	0.000	27.000
High school	14.274	7.295	0.000	30.000
Local Demographics				
Population	25888.373	14125.007	1208.000	83116.000
Pop. Density	13220.993	10687.626	1.003	118181.818
Higher Degree	30.303	9.847	10.356	61.289
Young Population	12.120	4.121	3.151	26.285
Old Population	16.325	4.455	5.712	33.292
Medium Age	42.603	3.637	32.700	55.400
Sex Ratio	95.750	4.656	81.024	124.508
Sales Period Control				
Spring	0.215	0.411	0.000	1.000
Fall	0.343	0.475	0.000	1.000
Winter	0.243	0.429	0.000	1.000

Table 1.2: Descriptive statistics of variables

tive statistics analysis, the aggregated hedonic dataset consists of a total of 52,644 observations, including property prices and other hedonic variables. For the green index, Google Street View (GSV) images were used, which have proven to be a reliable and promising source to examine urban areas, especially when compared to less developed or rural regions.
GSV images are particularly valuable owing to their comprehensive panoramic landscape information. For the dataset in the paper, 409,390 GSV images from within Busan's administrative boundary were retreived, all generated in 2017 and 2018. During the image collection, location tokens with geographical information, such as latitude and longitude were included, to facilitate the spatial interpolation based on geographic coordinates.

Subsequently, images with artificial greenness, such as those depicting playgrounds or tunnels painted green, were been filtered out, resulting in 306,425 cleaned GSV images. This cleaning procedure was essential to accurately assess the impact of natural greenness.
Image processing was performed by converting the images to the hue, saturation, and value (HSV) colour space and setting upper and lower boundaries to identify natural greenness. Images were gray scaled to determine whether each pixel fell within these boundaries. Then Green Index was individually calculated for each image. Finally, spatial interpolation was conducted to address missing values and calculated green indices were integrated into the aggregated hedonic dataset. In summary, a four-step approach (Figure 1.1) was used in the paper to construct the green index data: GSV image collection, colour space conversion and masking, green index calculation, and spatial interpolation.

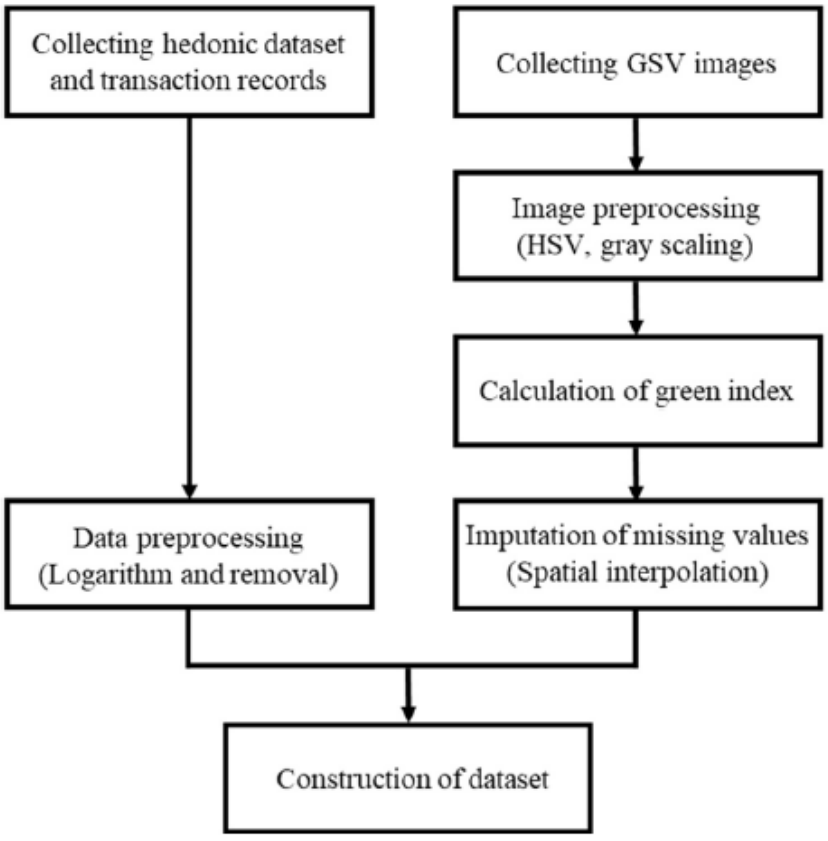

Figure 1.1: Flowchart illustrating the procedure for constructing the hedonic dataset.

In addition to GSV images, property data for the years 2018 and 2019 were also aggregated and cleaned. The resulting dataset encompasses transaction records and other hedonic variables sourced from public data repositories, including the Korea Transport Database, Statistics Korea, MLIT,

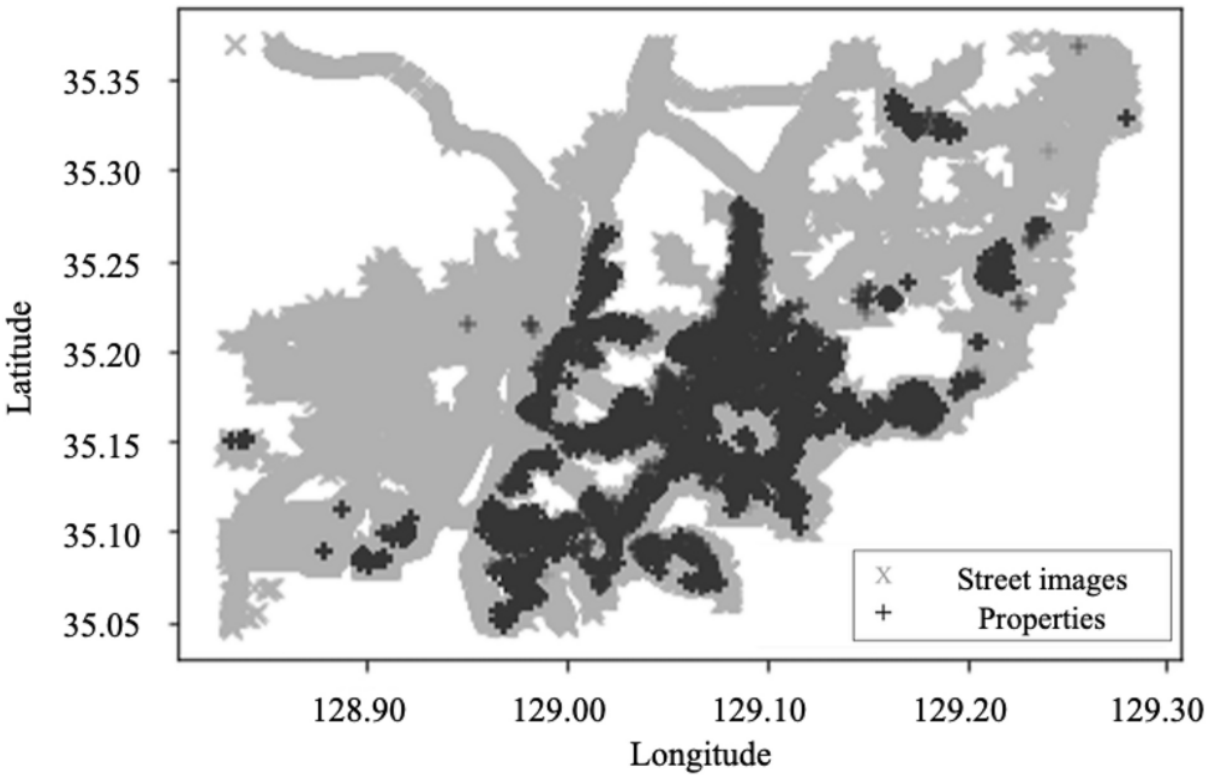

Figure 1.2: Spatial distribution of street images and property units

and the Spatial Geographic Information Service. Supplementary property information was also obtained from the websites of three private real estate companies: `kbland.kr`, `land.naver.com`, and `realty.daum.net`. During the data collection process, each transaction record was treated as an individual observation. The compiled transaction records included geographic coordinates and property-specific variables such as the property price, address, year built, floor, floor area, and transaction date. Next, data on environmental amenities, local built environment, and local demographic variables for each property was collected. Specifically, distance variables were calculated based on road network distances from each housing unit to the nearest environmental amenity or subway station. The highly skewed variables were transformed into a logarithmic scale to be close to a normal distribution. Additionally, the variable for bus stops was calculated to reflect the number of bus stops within a 400-m radius buffer around each housing unit using the ArcMap.

1.3 Calculation of the green index

The green index, as quantified through the previous steps, uses GSV images captured from a three-dimensional perspective. This approach allows

us to assess the degree of natural greenness visible to pedestrians.

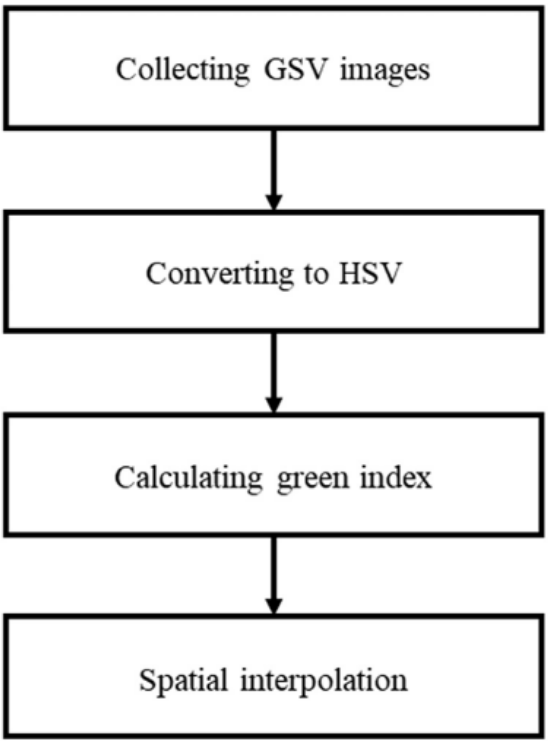

Figure 1.3: Four-step process for quantifying the green index.

After collecting street view images, each of the GSV images were transformed from a red, green, and blue colour space to the HSV colour space to improve image clarity[24]. Next, the upper and lower boundaries were set to capture natural greenness based on the HSV colour space, based on which all images were scrutinised according to whether or not each of the pixels falls between boundaries, as described before. If a pixel value fell out of boundaries, then that pixel was masked by assigning a value of zero.

Figure 1.4: Original (left) and converted (right) images with masked pixels

Therefore, the green index of each GSV image was calculated as follows:

$$\text{Green index}_i = \frac{\text{pixel}_{\text{non-zero}}}{\text{pixel}_{\text{total}}} \times 100 \tag{1.1}$$

where $\text{pixel}_{\text{non-zero}}$ denotes the number of non-zero pixels and $\text{pixel}_{\text{total}}$ represents the total number of pixels in an image.

1.4 Spatial interpolation

Notably, less GSV images are taken of city outskirts, because natural barriers, such as mountains, rivers, and forests, frequently determine administrative boundaries, such that city outskirts tend to have a smaller number of street images. This scenario indicates that the use of GSV images can potentially suffer from uneven spatial distribution, which leads to a biased quantification of the green index.
To address this issue, spatial interpolation was used in the paper to transform point data into areal information[9]. This relies solely on longitude and latitude information, making it accessible for spatial analysis and urban studies. The spatial interpolation assumes that adjacent properties are likely to exhibit similar levels of street greenness.

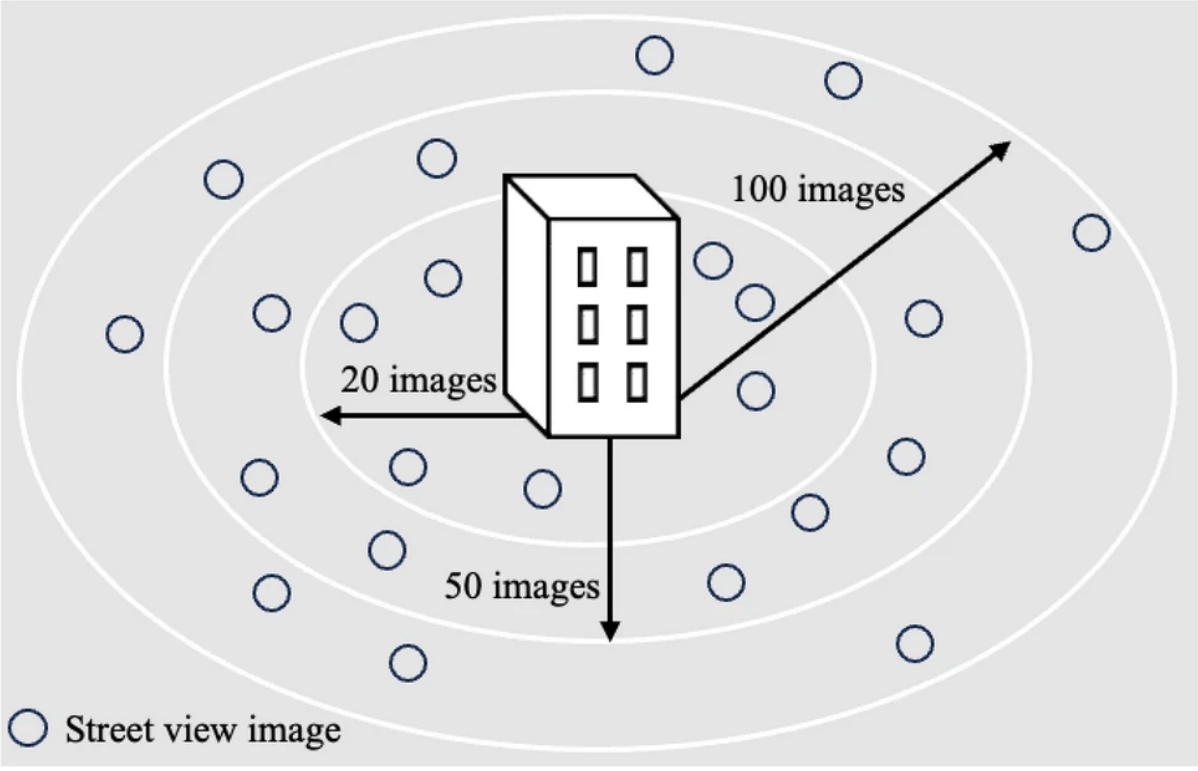

Figure 1.5: Graphical description of spatial interpolation

The implementation of spatial interpolation involves two major steps: calculating the haversine distance and averaging the green index. First, longitude and latitude information was been used to calculate the distances between target properties and all green indices using the haversine formula[1]. This formula is commonly used to calculate the distance between two points based on their longitude and latitude coordinates. The mathematical expression for the haversine formula is as follows:

$$d_{\text{haversine}} = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\text{lat}}{2}\right) + \cos(\text{lat}_p)\cos(\text{lat}_g)\sin^2\left(\frac{\Delta\text{lng}}{2}\right)}\right) \tag{1.2}$$

where R is the Earth's radius, set at 6,371 km; Δlat is the difference between the latitude of the target property (lat_p) and the latitude of the green index (lat_g); and Δlng is the difference in longitudes between the target property and the green index. Next, the calculated haversine distances were arranged in ascending order and the mean values of the green indices were assigned to each property unit by aggregating the values of a specific number of the nearest green indices. At the averaging stage, the number of nearest green indices can be adjusted as a parameter. After being experimented with different numbers of nearest images, specifically 50, 100, and 150, given data paper focused on using 50 nearest images to explore the relationship between urban greenness and property prices.

1.5 Descriptive Statistics

In this section some preliminary analysis of the given data has been done to visualize the data roughly. However, in the next chapters analysis of the data has been done at length using factor analysis and regression. Here analysis of some basic expected and observed nature of the data has been shown.

1.5.1 Normality of Property Price and Green Index

In the given paper the two most important variables are "Property Prices" and "Green Index". It was necessary to check the normality of these variables. The histograms were approximately normal and qq plots were more or less over the fixed lines except for some outliers (Figure 1.6).
Hence it can be concluded that these variables are approximately normal.

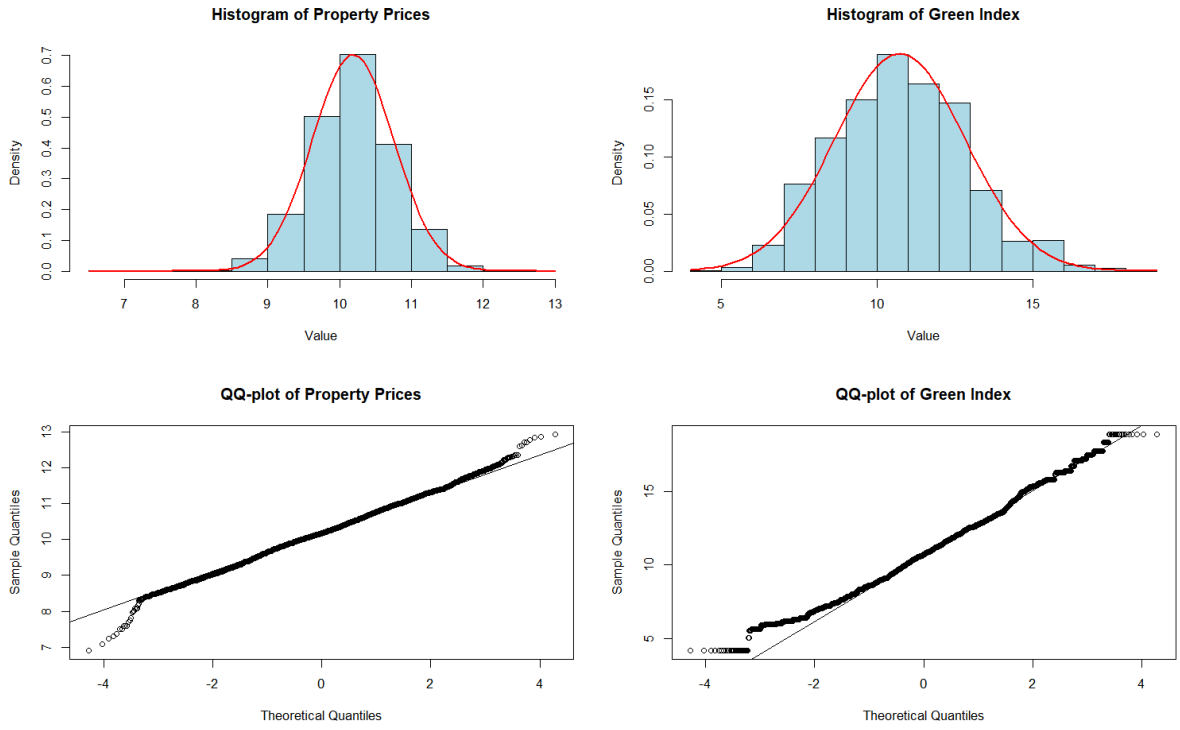

Figure 1.6: Histogram and QQ plot of property prices and green index

1.5.2 Relation between Property Prices and Green Index

The paper claimed that Green Index is a robust measure of urban greenness. So, it was expected that "Property Prices" will be significantly affected by the Green Index. It was also expected that properties having higher green index would be more in demand and thus would have higher price compared to low green index properties. This might not be true in every instance, but as we have a very large sample, this trend was expected. But the scatterplot (Figure 1.7) between Green Index and Property Prices didn't show any particular pattern. Also, the correlation between green index and property price was observed to be -0.13, which was very weak and also negative. It was also observed how property price varies across different green index levels, using boxplots (Figure 1.8). This too didn't reveal any pattern, although some observations were made. There were a lot of outliers for medium and high green index levels, the price varied more when the green index was very high and the median property price was almost same over all levels.
These observations seem to contradict the hypothesis that property price is heavily affected by green index and property price increases with increase in green index.

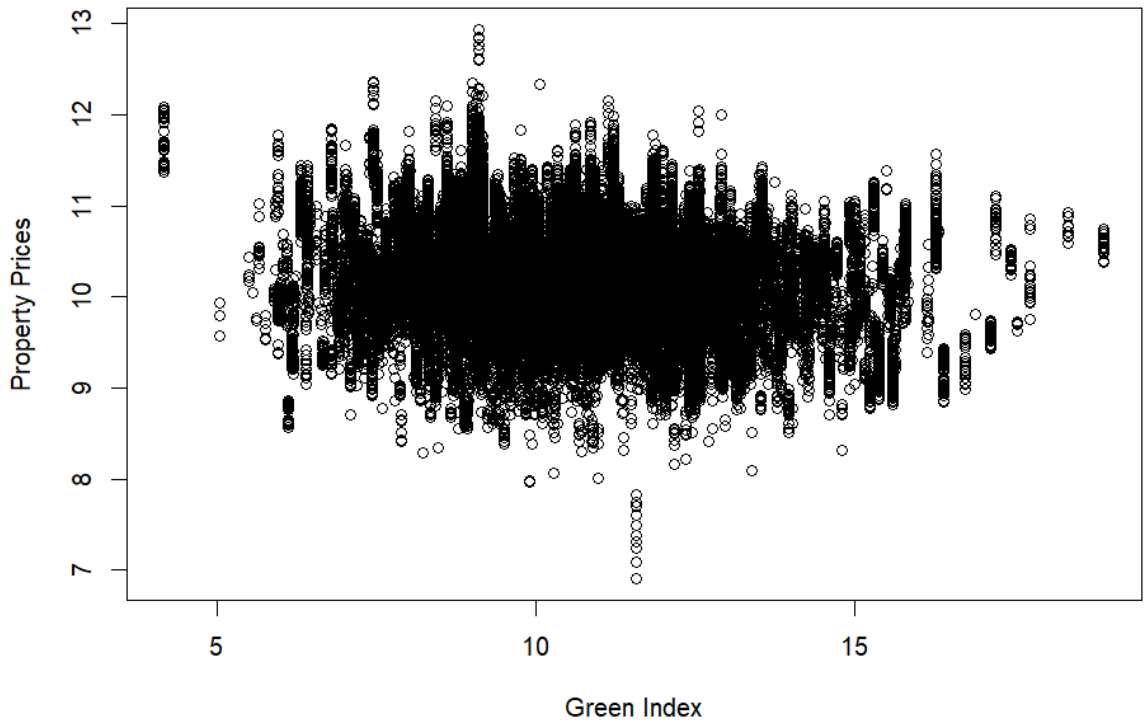

Figure 1.7: Scatterplot Between green index and property prices

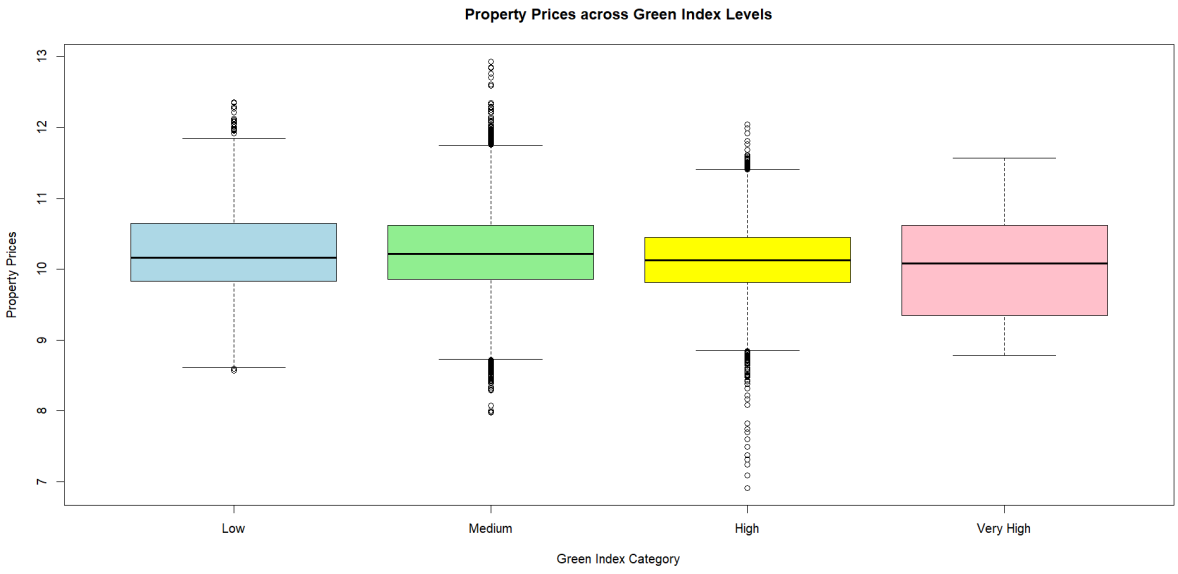

Figure 1.8: Box plot of property prices across levels of green index

1.5.3 Spatial variability in Green Index and Property Prices

The property prices and green index were also expected to vary over latitude and longitude. Normally, the price of property would be higher nearer to the centre of the city and lower further out. Similarly, more greenery is expected in the suburbs than in the main city. Thus, heatmaps were made of both, but we couldn't find any legible patterns.

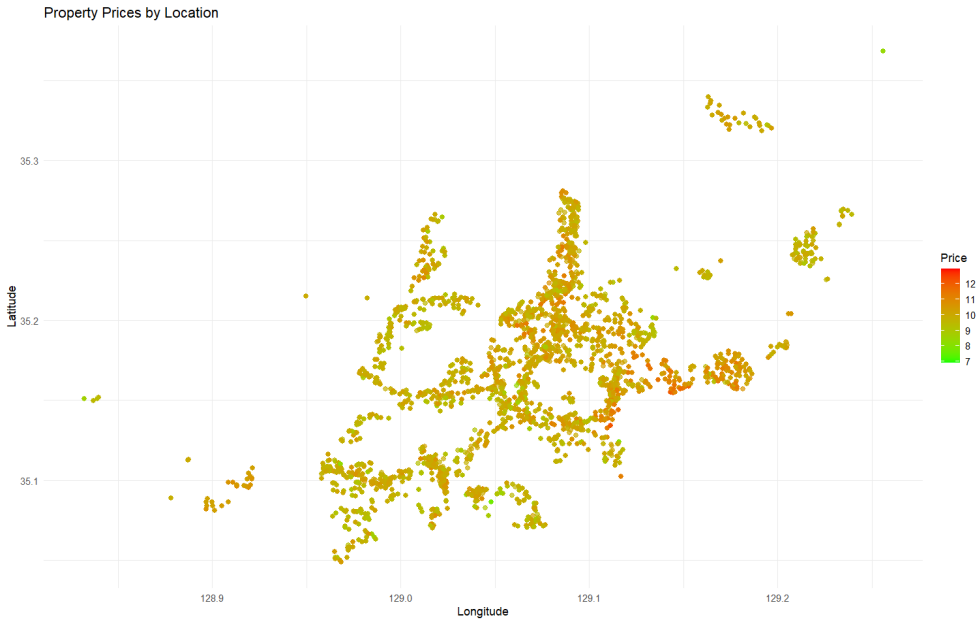

Figure 1.9: Property Prices vs Location

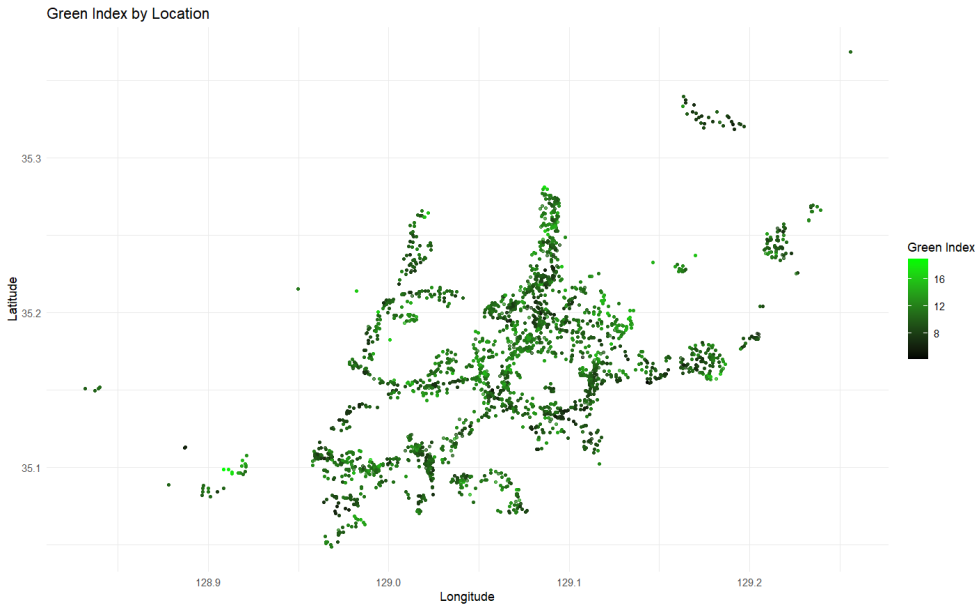

Figure 1.10: Green Index vs Location

This could be because Busan has evenly distributed greenery, even in the city centre and the property prices are also evenly distributed.

1.5.4 Relation Between Property Price and Year

Normally, the older the property, lower is its expected price. So, we expected an increasing relation in the scatterplot between Year and Property Price.

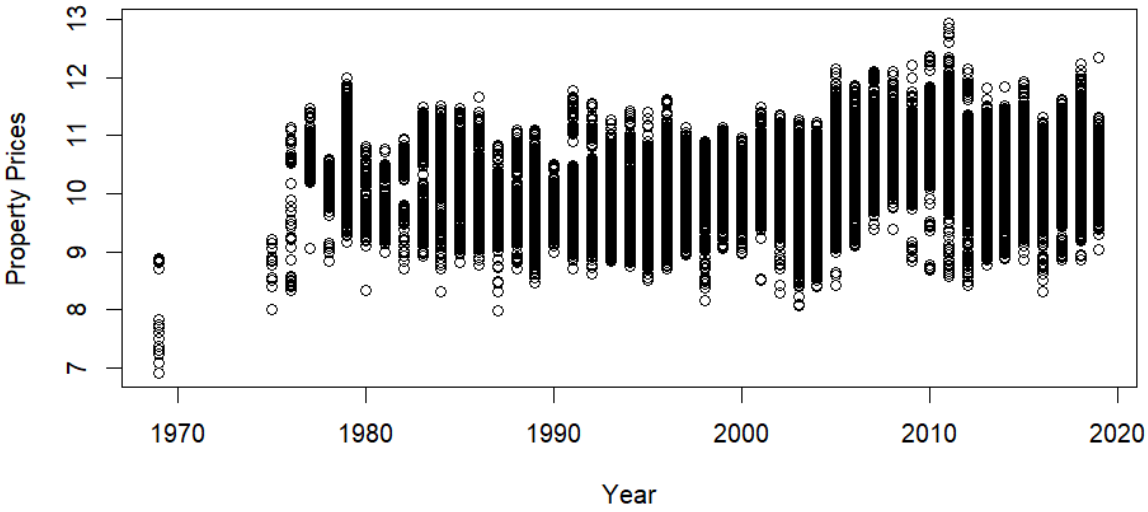

Figure 1.11: Scatterplot between property prices and year

It was observed that the variation in property price is higher in recent buildings. The correlation between the two was 0.27. Although, it was not significant, it hints at an increasing relationship like we claimed.

We were unable to find any clear relationships between the variables using graphical methods. So, we proceeded with factor analysis to explain the correlations between the variables.

Chapter 2

Factor Analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors[15]. It aims to interpret the factors in subject-matter terms and provide estimates of the individual values of the factors to be used in further analysis.
For example, it is possible that variations in six observed variables mainly reflect the variations in two unobserved variables. The aim of factor analysis is to find these factors and the extent to which the variables are related to a given factor to ultimately reduce the number of variables in the dataset.
We've used a technique within factor analysis called Exploratory Factor Analysis (EFA). This is used when there is no a priori hypothesis about factors or patterns of measured variables[20]. EFA is based on the common factor model. In this model, observed variables are expressed as a function of common factors, unique factors, and errors of measurement. Each unique factor influences only one observed variable, and does not explain correlations between observed variables. Common factors influence more than one observed variable and "factor loadings" are measures of the influence of a common factor on a observed variable.

2.1 Model[12]

Let there be p observed variables $X_1, \cdots, X_p$ and m underlying factors $F_1, \cdots, F_m$. We assume that the variables have zero mean and one vari-

ance, beforehand. The Factor Analysis model is defined as -

$$\begin{aligned} X_1 &= \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1m}F_m + \epsilon_1 \\ X_2 &= \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2m}F_m + \epsilon_2 \\ &\vdots \\ X_p &= \lambda_{p1}F_1 + \lambda_{p2}F_2 + \cdots + \lambda_{pm}F_m + \epsilon_p \end{aligned} \tag{2.1}$$

Here λ_{ij} are called factor loadings and ϵ_i are the unique error terms. This can be written compactly as -

$$X = \Lambda F + \varepsilon \tag{2.2}$$

Here X is the $p \times 1$ vector of observed variables, $\Lambda = [\lambda_{ij}]$ is the $p \times m$ matrix of factor loadings, F is the $m \times 1$ vector of factors and ε is the $p \times 1$ vector of error terms.
The assumptions in this model are -

1. $E(\epsilon_i) = 0$, $E(F_j) = 0$ and $Var(F_j) = 1$.
2. ε and F are uncorrelated among themselves and with each other i.e. $Cov(F_i, F_j) = Cov(F_i, \epsilon_j) = Cov(\epsilon_i, \epsilon_j) = 0$.
3. The covariance matrix Ψ of the residuals is diagonal, with diagonal elements $\psi_i = Var(\epsilon_i)$.

From the assumptions, it is clear that

$$\begin{aligned} Var(X_i) &= \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \psi_i \\ &= h_i^2 + \psi_i \end{aligned} \tag{2.3}$$

where $i = 1, \ldots, p$ and h_i^2 is called communality. It denotes the amount of variance in the variable, that is explained by the factors. ψ_i is called the uniqueness variance and represents the variance unexplained by the factors.
We also can calculate from the assumptions that

$$Cov(X_i, X_k) = \lambda_{i1}\lambda_{k1} + \lambda_{i2}\lambda_{k2} + \cdots + \lambda_{im}\lambda_{km} \tag{2.4}$$

Using equations (2.3) and (2.4), we can write

$$\Sigma = \Lambda\Lambda^T + \Psi \tag{2.5}$$

where Σ is the covariance matrix of the variables.

Hence, the $\frac{p(p+1)}{2}$ non-redundant elements of Σ can be reproduced exactly by the pm factor loadings and the p unique variances. This is very useful when p is comparatively much larger than m, as $\frac{p(p+1)}{2}$ is much greater than $pm + p$ in those cases.

2.2 Preparation of Data

It was assumed in the model that observed variables have zero mean and one variance. So, z-score standardization should be performed to transform the variables to zero mean and one variance.

$$Z_j = \frac{X_j - \overline{X_j}}{s_j} \tag{2.6}$$

It is also needed to be ensured that no variables are highly correlated ($|r| > 0.9$). This is because highly correlated variables can lead to unstable and difficult to interpret factor loadings.
It is undesirable to have high multicollinearity in the data. To check for this we have a measure called Variance Inflation Factor (VIF).

2.2.1 Variance Inflation Factor (VIF)

It is the ratio of the variance of a parameter estimate when fitting a full model that includes other parameters to the variance of the parameter estimate if the model is fit with only the parameter on its own[18]. The VIF provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity.

$$VIF_j = \frac{1}{1 - R_j^2} \tag{2.7}$$

where R_j is the coefficient of determination from regressing X_j against all other observed variables. VIF value greater than 5 indicates multicollinearity.

2.3 Assessing Suitability of Data

2.3.1 Kaiser–Meyer–Olkin Test

It is a statistical measure to determine how suited data is for factor analysis. The test measures sampling adequacy for each variable in the model

and the complete model. The statistic is a measure of the proportion of variance among variables that might be common variance[22]. The higher the proportion, the higher the KMO value, the more suited the data is to factor analysis.
In this test, at first the Measure of Sampling Adequacy (MSA) is calculated for each variable-

$$MSA_j = \frac{\sum\limits_{k \neq j} r_{jk}^2}{\sum\limits_{k \neq j} r_{jk}^2 + \sum\limits_{k \neq j} p_{jk}^2} \tag{2.8}$$

Here r_{jk} is the correlation between the variable in question and another, and p_{jk} is the partial correlation.
The Kaiser–Meyer–Olkin criterion is given by-

$$KMO = \frac{\sum\limits_{j \neq k} \sum r_{jk}^2}{\sum\limits_{j \neq k} \sum r_{jk}^2 + \sum\limits_{j \neq k} \sum p_{jk}^2} \tag{2.9}$$

Both MSA_j and KMO returns values between 0 and 1. $KMO > 0.6$ indicates that the data is suitable factor analysis, while $MSA_j > 0.6$ indicates that the variable is not problematic.

2.3.2 Bartlett's Test of Sphericity

This tests whether a matrix is significantly different from an identity matrix[3]. If the correlation matrix is equal to an identity matrix, we cannot proceed with EFA, since there is no correlation between variables.

H_o : Variables not correlated i.e. the correlation matrix is identity.

H_a : Variables are significantly correlated i.e. the correlation matrix is significantly different from identity.

The test statistic is-

$$\mathcal{T} = \log\left[\det R\left(n - 1 - \frac{2p+5}{6}\right)\right] \sim \chi^2_{\frac{p(p-1)}{2}} \tag{2.10}$$

Here R is the correlation matrix of the dataset, n is the number of sample elements and p is the number of variables.

2.4 Determining number of Factors

The choice of number of underlying factors to be used in exploratory factor analysis depends on various considerations. Choosing too few factors can lead to underfitting, where important structures in the data are missed, while selecting too many can result in overfitting, capturing noise instead of meaningful patterns. There are three ways to determine the appropriate number of factors.

2.4.1 Kaiser's Criterion

The eigenvalues of the correlation matrix is calculated in this method and the number of eigenvalues greater than 1 gives the number of factors to include in the model.

2.4.2 Scree Plot

The eigenvalues of the correlation matrix are plotted from largest to smallest. The graph is examined to determine the last substantial drop in the magnitude of eigenvalues. The number of plotted points before the last drop is the number of factors to include in the model.

2.4.3 Parallel Analysis

Parallel analysis[21] is regarded as one of the more accurate methods for determining the number of factors or components to retain. The eigenvalues are plotted from largest to smallest along with a set of random eigenvalues. The number of eigenvalues before the intersection point indicates how many factors to include in the model.

None of these methods are entirely foolproof. The Kaiser criterion, for example, tends to overestimate the number of factors, especially in large datasets. Scree plots rely on visual interpretation, which can be subjective. Parallel analysis is more robust but still depends on simulation assumptions that may not perfectly align with the data structure. Because these methods can yield different results, and none are universally reliable, it's essential to supplement them with theoretical reasoning and domain knowledge. Ultimately, determining the number of factors is both a statistical and conceptual decision, requiring a balance between empirical indicators and interpretability of the factor model.

2.5 Factor Extraction

In this step the matrix Λ of factor loadings is estimated using various methods. The result, called the initial factor solution, is often not easily interpretable. There are several methods to extract factors like Principal Component Method (PC), Principal Axis Factoring (PAF), Maximum Likelihood Estimate (MLE) and Weighted Least Squares (WLS). The estimate of the factor loading matrix Λ thus obtained is then used to interpret the factors.
When factor analysis is performed, it is expected that each variable will depend highly on exactly one factor and each factor will have atleast one variable depending highly on it. Such a solution is easily interpretable because all the variables can be grouped under the factors without any overlaps or omissions.
Ease of interpretation is obtained in a solution that exhibits a "simple structure". A "simple structure" is characteristic of the following pattern of factor loadings-

1. Each observed variable loads highly on a single factor and has small loadings on the remaining.

2. For each factor there is atleast one variable saliently loading on it.

As a rule of thumb, we can consider a loading to be prominent if it is larger than 0.3. This is because $\lambda_{ij} > 0.3 \implies \lambda_{ij}^2 > 0.09$ and recall that

$$Var(X_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \psi_i = 1$$

by assumption. Hence, the factor F_j explains atleast close to 10% of the variance of X_i. If a factor loads prominently on multiple factors, it is known as a cross-loading.

2.6 Factor Rotation

The initial factor solution obtained from factor extraction step is not unique. Let T be an invertible $m \times m$ matrix. Then, we can write the model as-

$$X = \Lambda F + \varepsilon = \Lambda T^{-1} T F + \varepsilon = \Lambda^* F^* + \varepsilon \tag{2.11}$$

where Λ^* is another possible loading matrix. In factor rotation, we choose this matrix T in such way that the factor loadings are more interpretable.

The resulting Λ^* is called the rotated factor loading matrix. Notice that ε is unchanged in this transformation and thus rotation does not change the total variance explained by the model, it just redistributes the variance across factors.
Recall that the factors obtained, in previous step, are uncorrelated, by construction. There are two types of rotation based on the correlation between resulting factors.

1. **Orthogonal Rotation**: The resulting factors are still uncorrelated.
2. **Oblique Rotation**: The resulting factors are allowed to be correlated.

A popular type of orthogonal rotation is Varimax. Based on Varimax, we can do an oblique rotation called Promax[6].

2.6.1 Kaiser's Varimax Criterion

This criterion[13] is based on a measure, denoted V, of closeness to simple structure, which needs to be maximized across all possible orthogonal rotations.

$$\begin{aligned} V &= \text{Sum of variances of squared scaled loadings on factors} \\ &= \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\left(\frac{\lambda_{ij}}{h_i}\right)^4 - \frac{1}{p}\left[\sum_{i=1}^{p}\left(\frac{\lambda_{ij}}{h_i}\right)^2\right]^2\right] \end{aligned} \tag{2.12}$$

This maximization has the tendency to polarize the factor loadings so that they are either high or low, thereupon making the loading matrix more interpretable. The result of this rotation is the creation of groups of large and of negligible loadings in any column of the factor loading matrix Λ.

2.7 Estimating Factor Scores

Once we carry out factor extraction, we actually furnish estimates of the factor loading matrix Λ and the error covariance matrix Ψ. In order to obtain factor scores, we may consider these matrices as known. With this in mind

$$X = \Lambda F + \varepsilon$$

is just the general linear model. Then, we can use ordinary least squares regression to obtain an estimate of F.

2.8 Goodness of Fit Indices

2.8.1 Chi-Square Test

The model implied covariance matrix is given by

$$\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}^T + \hat{\Psi} \tag{2.13}$$

where $\hat{\Lambda}$ is the estimated factor loading matrix and $\hat{\Psi}$ is the estimated unique variance matrix. Let S be the observed covariance matrix, then the test statistic[8] is

$$\chi^2 = n\left[\log|\hat{\Sigma}| - \log|S| + tr(S\hat{\Sigma}^{-1}) - p\right] \tag{2.14}$$

If n is large, χ^2 follows a chi-square distribution with

$$df = \frac{p(p-1)}{2} - pm + \frac{m(m-1)}{2} \tag{2.15}$$

The hypotheses of this test are-
$H_o : \hat{\Sigma} = S$ i.e. the observed covariance matrix is not significantly different from the model-implied covariance matrix and thus the model fits the data well.
$H_a : \hat{\Sigma} \neq S$ i.e. the observed covariance matrix is significantly different from the model-implied covariance matrix and the model isn't a good fit.

The model fits well if p-value$>$ 0.05, as we don't reject H_o in this case. But, this test is highly sensitive to sample size and often fails for large samples.

2.8.2 Root Mean Squared Error of Approximation (RMSEA)

This avoids issues of sample size by analysing the discrepancy between the hypothesized model, with optimally chosen parameter estimates, and the population covariance matrix.

$$RMSEA = \sqrt{\max\left(\frac{\chi^2/df - 1}{n-1}, 0\right)} \tag{2.16}$$

The RMSEA[23] ranges from 0 to 1, with smaller values indicating better model fit. A value of 0.05 or less is indicative of good model fit and a value less than 0.08 is considered acceptable.

2.8.3 Standardized Root Mean Squared Residual (SRMR)

This is the square root of the discrepancy between the sample covariance matrix and the model implied covariance matrix. The SRMR[23] also ranges from 0 to 1. A value of 0.05 or less is indicative of good model fit.

$$SRMR = \sqrt{\frac{2\sum_{i=1}^{p}\sum_{j=1}^{i}\left(\frac{S_{ij}-\hat{\Sigma}_{ij}}{S_{ii}S_{jj}}\right)^2}{p(p+1)}} \tag{2.17}$$

where S_{ij} are the elements of observed covariance matrix and $\hat{\Sigma}_{ij}$ are the elements of the model implied covariance matrix.

2.8.4 Tucker-Lewis Index (TLI)

It is a relative measure of the difference between the chi-squared value of the hypothesized model and the chi-squared value of the null model, which assumes that all observed variables are uncorrelated with each other[23].

$$TLI = \frac{\left(\frac{\chi^2_{\text{null}}}{df_{\text{null}}}\right) - \left(\frac{\chi^2_{\text{model}}}{df_{\text{model}}}\right)}{\left(\frac{\chi^2_{\text{null}}}{df_{\text{null}}}\right) - 1} \tag{2.18}$$

where the χ^2_{null} and df_{null} are the test statistic and degrees of freedom for the null model.

2.8.5 Bayesian Information Criterion (BIC)

This is a model selection criterion that balances model fit and model complexity. It penalizes overly complex models to avoid overfitting.

$$BIC = \chi^2 - df \log n \tag{2.19}$$

Lower BIC implies better fit.

2.9 Implementation on our dataset

Our dataset contains 27 variables that influence property prices. To uncover the latent structure among these variables, we employed factor analysis. This allowed us to reduce dimensionality and identify underlying factors. We then used regression analysis to examine how these extracted factors impact property prices.

2.9.1 Preparation of Data

At first, all the variables were standardized to their z-scores. Then variables having $|r| > 0.9$ with some other variable and having VIF greater than 5 were removed, as these indicated high collinearity.
In our data, we found that the variables "Dist. Green" and "Dist. CBD" have correlation -0.9433992 and both have VIF greater than 5. We calculated VIF using `vif()` from `car` library in R (Table 2.1). We decided to remove the variable "Dist. CBD" as the variable "Dist. Green" had lower VIF and seemed more important as we focus more on effect of environmental amenities on property price.

Variable	**VIF**	**Variable**	**VIF**
Longitude	1.6716	Latitude	1.5106
Size	1.3233	Floor	1.4452
Highest floor	2.1226	Units	1.3122
Parking	1.2611	Heating	1.2632
Year	1.3533	Dist. Green	9.7760
Dist. Water	1.1866	Green Index	1.0849
Dist. Subway	1.3071	Bus Stop	1.3235
Dist. CBD	11.2446	Top Univ.	3.1226
High School	3.1882	Sex Ratio	1.4670
Population	1.9396	Pop. Density	1.1886
Higher Degree	1.6227	Young Population	3.1493
Median Age	2.1227	Old Population	3.1433
Spring	1.6482	Fall	1.8308
Winter	1.6874		

Table 2.1: VIF values of all variables

We generated the heatmap of correlation matrix shown in Figure 2.1 using `ggcorrplot()` from `ggcorrplot` library of R.

2.9.2 Assessing Suitability of Data

Kaiser–Meyer–Olkin Test

$KMO > 0.6$ indicates that the data is suitable factor analysis. We calculated KMO using `KMO()` from `psych` library of R. The overall KMO value of our dataset came out to be 0.61, which was adequate for factor analysis. But when MSA was calculated for each of the variables (Table 2.2), it was discovered that MSA was below 0.5 for the variables "Year", "Fall", "Spring", "Winter" and "Latitude".

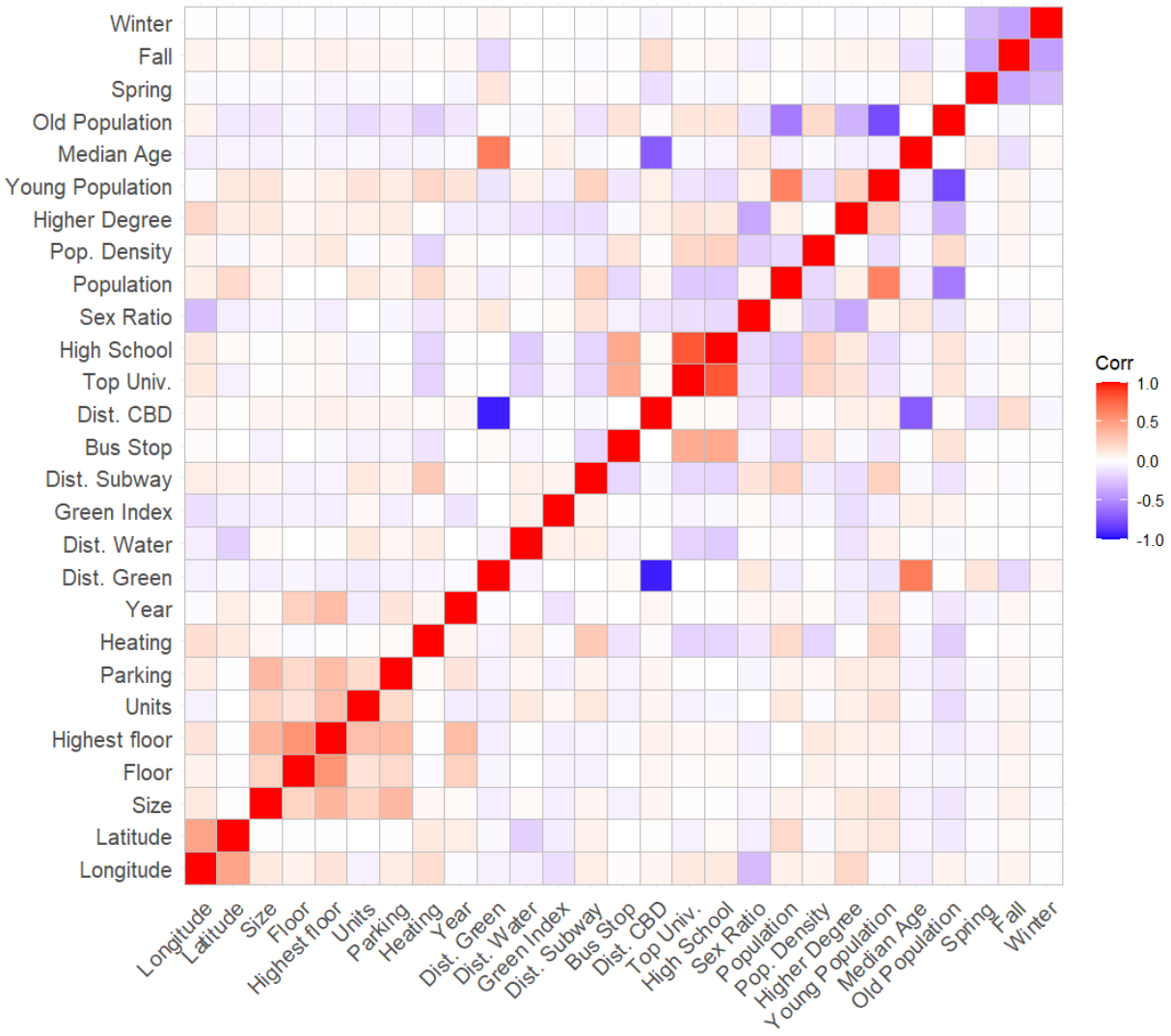

Figure 2.1: Heatmap of correlation matrix of all variables

We tried removing various combinations of variables to improve MSA of each variable. The best results were obtained on removing "Fall", "Spring", "Winter", "Latitude", "Longitude" and "Sex Ratio". In this case, the KMO came out to be 0.68 and MSA improved to greater than 0.5 for each of the remaining variables (Table 2.3).

Bartlett's Test of Sphericity

This test was performed on the remaining 20 variables. This was done in R using cortest.bartlett() from psych package. The output is shown in Figure 2.2. The p-value came out to be very close to zero, so the hypothesis was rejected at 5% level. Hence, it was deduced that the variables were sufficiently correlated for factor analysis.

Variable	MSA	Variable	MSA
Longitude	0.52	Latitude	0.49
Size	0.76	Floor	0.72
Highest floor	0.61	Units	0.56
Parking	0.80	Heating	0.70
Year	0.49	Dist. Green	0.54
Dist. Water	0.62	Green Index	0.59
Dist. Subway	0.65	Bus Stop	0.86
Top Univ.	0.64	High School	0.65
Sex Ratio	0.56	Population	0.82
Pop. Density	0.75	Higher Degree	0.57
Young Population	0.71	Median Age	0.54
Old Population	0.67	Spring	0.31
Fall	0.37	Winter	0.29

Table 2.2: MSA of all variables

Variable	MSA	Variable	MSA
Size	0.74	Floor	0.70
Highest floor	0.63	Units	0.63
Parking	0.80	Heating	0.79
Year	0.54	Dist. Green	0.50
Dist. Water	0.68	Green Index	0.59
Dist. Subway	0.66	Bus Stop	0.85
Top Univ.	0.65	High School	0.65
Population	0.85	Pop. Density	0.80
Higher Degree	0.65	Young Population	0.70
Median Age	0.51	Old Population	0.69

Table 2.3: MSA of remaining variables

```
> cortest.bartlett(cor(new_data), n = nrow(new_data))
$chisq
[1] 283926.6

$p.value
[1] 0

$df
[1] 190
```

Figure 2.2: Result of Bartlett's test

2.9.3 Determining number of Factors

parallel() from `nFactor` library of R was used to perform parallel analysis. `nScree` was used to check the number of factors found via other two methods and decide the optimal number. `plotnScree` was used to create the scree and parallel analysis plot (Figure 2.3).
The optimal number of factors is given to be 4, but we've used 7 factors which is recommended by parallel analysis, because 4 factors caused underfitting and gave uninterpretable solutions.

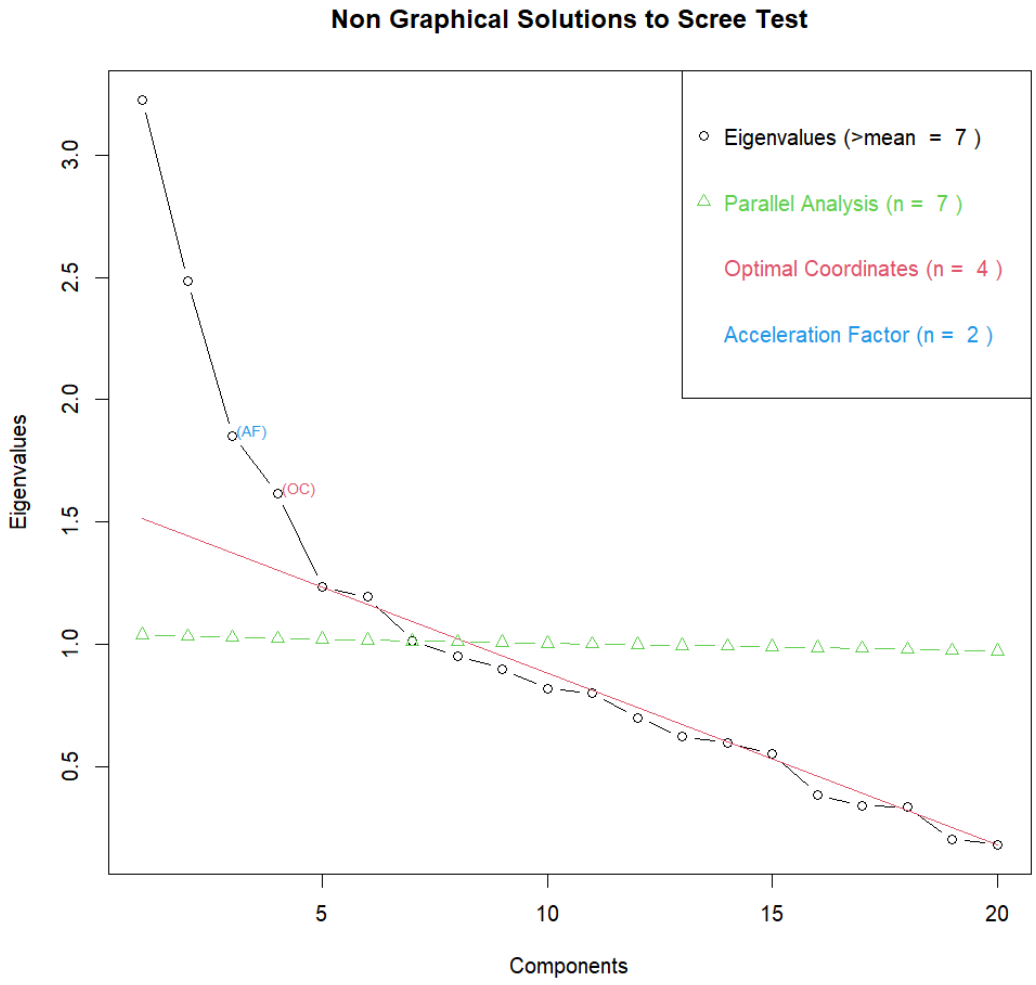

Figure 2.3: Scree plot

2.9.4 Factor Extraction and Rotation

In R, the factor extraction and rotation can be done using `fa()` from `psych` library. WLS method was used for factor extraction. The library `GPArotation` was needed for calculating the rotated loading matrix. In

this case, it was expected that the factors would be correlated. This was because the variables were expected to be roughly grouped under factors like "Property Characteristics", "Demographics" and "Environmental Amenities", which are logically related. So, Promax oblique rotation was used. Hence, factor analysis was performed on the dataset of 20 variables with 7 factors, using WLS to extract factors and Promax to rotate the loading matrix.
The R code used to perform factor analysis and view estimated loading matrix was-

```
efa_results <- fa(new_data, nfactors = 7, rotate =
   "promax", fm = "wls", scores = "regression")
print(efa_results$loadings, cutoff = 0.3)
```

We have printed the estimated factor loading matrix (Figure 2.4) with blanks at places where loading is less than 0.3. Note that the factors are written as WLS1,..., WLS7 in the code output.

```
Loadings:
                WLS1   WLS3   WLS2   WLS4   WLS5   WLS6   WLS7
Size                          0.435
Floor                         0.581
Highest floor                 1.008
Units                         0.333
Parking                       0.419
Heating
Year                          0.305
Dist. Green                         0.911
Dist. Water                                                1.029
Green Index                                  1.008
Dist. Subway
Bus Stop                0.539
Top Univ.               0.954
High School             0.951
Population        0.681
Pop. Density                                         1.004
Higher Degree
Young Population  0.967
Median Age                          0.747
Old Population   -0.893

                 WLS1  WLS3  WLS2  WLS4  WLS5  WLS6  WLS7
SS loadings    2.439 2.252 1.934 1.411 1.079 1.033 1.141
Proportion Var 0.122 0.113 0.097 0.071 0.054 0.052 0.057
Cumulative Var 0.122 0.235 0.331 0.402 0.456 0.507 0.564
```

Figure 2.4: Factor loadings using 7 factors

2.9.5 Factor Interpretation

The key observations from the result were-

1. "Population", "Young Population" and "Old Population" had high loadings on factor 1.
2. "Size", "Floor", "Highest Floor", "Units", "Parking" and "Year" had high loadings on factor 2.
3. "Bus Stop", "Top Univ." and "High School" had high loadings on factor 3.
4. "Dist. Green" and "Median Age" had high loadings on factor 4.
5. "Green Index", "Pop. Density" and "Dist. Water" had high loadings on factor 5, 6 and 7, respectively.
6. This model explained 56.4% of the total variance of the variables.

The problems encountered were-

1. "Heating", "Dist. Subway" and "Higher Degree" had negligible loadings on all factors.
2. Factor 4 couldn't be interpreted as any real world construct.
3. 56.4% explained variance could not be considered as very significant.

A possible solution in this case was to increase number of factors as some variables loaded highly on none of the existing ones and two unrelated variables loaded highly on the same factor. On increasing the number of factors to 8, we got 60.9% explained variance, although the other two problems persisted. The results worsened on increasing the number of factors beyond 8 or decreasing them below 7. Thus, 8 was the optimal number of factors (Figure 2.5) in this model. We also tried rerunning the EFA by removing the variables, which were creating problems, but in each case we encountered further issues like other variables loading negligibly, cross-loadings for some variables and explained variance reducing further.

The observations from the improved result were-

1. "Population", "Young Population" and "Old Population" had high loadings on factor 1.

```
Loadings:
                 WLS1   WLS3   WLS2   WLS4   WLS6   WLS5   WLS8   WLS7
Size                           0.517
Floor                          0.534
Highest floor                  0.876
Units                          0.439
Parking                        0.452
Heating
Year                                                                0.830
Dist. Green                           0.823
Dist. Water                                  1.052
Green Index                                         1.031
Dist. Subway
Bus Stop                0.548
Top Univ.               0.958
High School             0.953
Population        0.694
Pop. Density                                                  0.992
Higher Degree
Young Population  1.002
Median Age                            0.820
Old Population   -0.894

                 WLS1  WLS3  WLS2  WLS4  WLS6  WLS5  WLS8  WLS7
SS loadings     2.514 2.281 1.768 1.367 1.171 1.121 1.011 0.938
Proportion Var  0.126 0.114 0.088 0.068 0.059 0.056 0.051 0.047
Cumulative Var  0.126 0.240 0.328 0.397 0.455 0.511 0.562 0.609
```

Figure 2.5: Factor loadings using 8 factors

2. "Size", "Floor", "Highest Floor", "Units" and "Parking" had high loadings on factor 2.

3. "Bus Stop", "Top Univ." and "High School" had high loadings on factor 3.

4. "Dist. Green" and "Median Age" had high loadings on factor 4.

5. "Green Index", "Dist. Water", "Year" and "Pop. Density" had high loadings on factor 5, 6, 7 and 8, respectively.

6. This model explained 60.9% of the total variance of the variables.

It was not possible to improve this result further by tweaking the dataset or changing the number of factors. Hence, we proceeded with goodness of fit tests for our model.

2.9.6 Goodness of Fit Tests

The fit indices are calculated automatically when we do `fa()` and can be seen by `summary(efa_results)`. In chi-square test, the test statistic was

$\chi^2 = 24187.17$ for the 7 factor model and $\chi^2 = 16833.42$ for the 8 factor model. The p-value is zero in both cases, but this could be because our sample size is very high. The value of the other fit indices for the two models are given below (Table 2.4). It was observed that our model was

Fit Index	**EFA(7 factors)**	**EFA(8 factors)**	**Accepted Region**
RMSEA	0.0803	0.0741	< 0.08 (Acceptable)
SRMR	0.0351	0.0265	< 0.05 (Good Fit)
TLI	0.773	0.806	> 0.90 (Good Fit)
BIC	23415.3	16202.88	Lower is better
Variance Explained	56.4%	60.9%	$> 60\%$

Table 2.4: Fit indices for 7 factor and 8 factor models

an acceptable fit in 2 out of the 3 fit indices and the 8 factor model was better according to BIC and also by total variance explained. Hence, it was concluded that the 8 factor model was an acceptable fit for the data.

2.9.7 Estimating Factor Scores

In R, factor scores are also estimated by `fa()`, using the parameter `scores = "regression"`. The factor matrix can be extracted by `efa_results$scores`. The estimated factor scores were used for regressing the variable "Property Prices" against the latent factors. This has been discussed in detail in Chapter 3.

2.9.8 Results

Although the final model had two issues, it explained more than 60% variance and the goodness of fit indices suggested that it was an acceptable fit. Hence, we conclude that the factor structure found by us (Table 2.5) explains the data well.
Figure 2.6 depicts the factor loadings of the variables diagrammatically. The circles are the factors while the squares are the variables. The thickness of the lines indicate the strength of loading. Red indicates negative loading while green indicates positive loading.

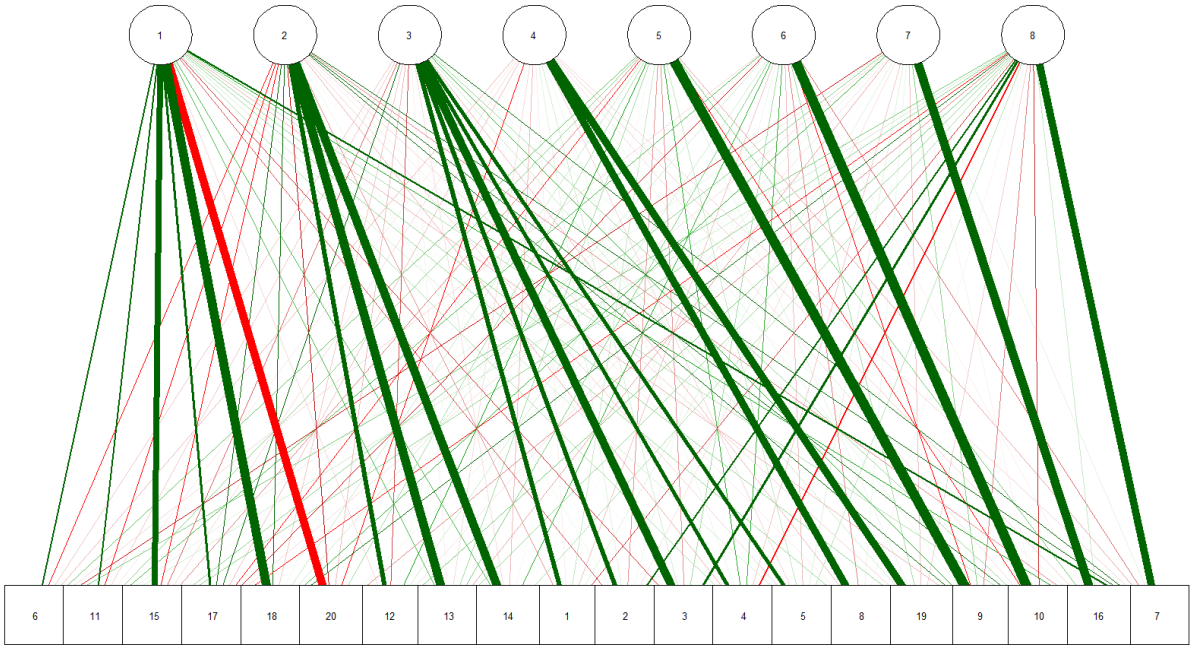

Figure 2.6: Factor loadings diagram

Factor	Variable	Factor Loading	Variance Explained
Demographics (WLS1)	Population	0.694	12.6%
	Young Population	1.002	
	Old Population	-0.894	
Property Characteristics (WLS2)	Size	0.517	8.8%
	Floor	0.534	
	Highest Floor	0.876	
	Units	0.439	
	Parking	0.452	
Access to Education (WLS3)	Bus Stop	0.548	11.4%
	Top Univ.	0.958	
	High School	0.953	
WLS4	Median Age	0.820	6.8%
	Dist. Green	0.823	
Green Index (WLS5)	Green Index	1.031	5.6%
Dist. Water (WLS6)	Dist. Water	1.052	5.9%
Building Age (WLS7)	Year	0.830	4.7%
Population Density (WLS8)	Pop. Density	0.992	5.1%

Table 2.5: Factor structure of data

Chapter 3

Multiple Linear Regression

In statistics, regression analysis is a set of statistical processes for estimating the relationship between a dependent variable and independent variables[4]. We consider the problem of regression when the study variable depends on more than one explanatory or independent variables. It is called a multiple linear regression[14] model, if the relation between the variables is assumed to be linear.

3.1 Model[10]

Let y be a variable depending on n independent variables, $x_1, x_2, \ldots, x_n$. y can be written as,

$$y = b_0 + x_1 b_1 + x_2 b_2 + \cdots + x_n b_n + \epsilon \tag{3.1}$$

The parameters $b_0, b_1, \cdots, b_n$ are the regression coefficients associated with $x_1, x_2, \cdots, x_n$ respectively and ϵ is the random error component reflecting the difference between the observed and fitted linear relationship.

Let there be m set of values of y with respect to n observed variables. The m-tuples of observations are also assumed to follow the same model. Thus they satisfy,

$$\begin{aligned} y_1 &= b_0 + x_{11} b_1 + x_{12} b_2 + \cdots + x_{1n} b_n + \epsilon_1 \\ y_2 &= b_0 + x_{12} b_2 + x_{22} b_2 + \cdots + x_{2n} b_n + \epsilon_2 \\ &\vdots \\ y_m &= b_0 + x_{m1} b_1 + x_{m2} b_2 + \cdots + x_{mn} b_m + \epsilon_n \end{aligned} \tag{3.2}$$

This can also be written in matrix form as,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix} \tag{3.3}$$

in short $Y = X\beta + \epsilon$ where,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \beta = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix} \tag{3.4}$$

In this model, assumptions are,

1. $E(\epsilon) = 0$
2. $E(\epsilon\epsilon^T) = \sigma^2 I$, where I is identity matrix of order m.
3. $Rank(X) = n + 1$
4. $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

3.1.1 Estimation of $\boldsymbol{\beta}$

We will estimate β by using ordinary least squares (OLS). We have to find $\hat{\beta}$ from the set of all β such that the sum of squared residuals

$$S(\beta) = \|Y - X\beta\|^2 = (Y - X\beta)^T(Y - X\beta) = \epsilon^T\epsilon = \sum_{i=1}^{m} \epsilon_i^2 \tag{3.5}$$

is minimized. To minimize $S(\beta)$, we take the derivative with respect to β and set it to zero:

$$\begin{aligned} \frac{\partial S}{\partial \beta} &= -2X^T(Y - X\beta) = 0 \\ \Rightarrow X^TY &= X^TX\hat{\beta} \\ \Rightarrow \hat{\beta} &= (X^TX)^{-1}X^TY \end{aligned}$$

Thus, the estimated coefficients $\hat{\beta}$ in multiple linear regression are obtained using the formula:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{3.6}$$

We define the estimation of Y,

$$\hat{Y} = X\hat{\beta} \tag{3.7}$$

and the residual

$$\hat{\epsilon} = Y - \hat{Y} \tag{3.8}$$

3.1.2 Distribution of $\hat{\beta}$

From the previous calculations,

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ \hat{\beta} &= (X^T X)^{-1} X^T (\beta X + \epsilon) = \beta + (X^T X)^{-1} X^T \epsilon \\ \hat{\beta} - \beta &= (X^T X)^{-1} X^T \epsilon \Rightarrow E(\hat{\beta}) = E(\beta) \end{aligned} \tag{3.9}$$

$$\begin{aligned} Var(\hat{\beta}) &= E((\hat{\beta} - \beta)(\hat{\beta} - \beta)^T) \\ &= E(((X^T X)^{-1} X^T (\epsilon))((X^T X)^{-1} X^T (\epsilon))) \\ &= \sigma^2 (X^T X)^{-1} \end{aligned} \tag{3.10}$$

Hence, $E(\hat{\beta}_j) = E(\beta_j)$ and $Var(\hat{\beta}_j) = \sigma^2 (X^T X)^{-1}_{jj}$.

As the $\hat{\beta}$ follows multivariate normal, the $\hat{\beta}_{ii}$ follows normal. And we have previously derived mean and variance of $\hat{\beta}_{ii}$. Hence,

$$\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2 (X^T X)^{-1}_{ii}) \tag{3.11}$$

3.1.3 Distribution of Sum of Squared Residuals (SSR)

The SSR is defined as:

$$\text{SSR} = \hat{\epsilon}^T \hat{\epsilon} = \|\hat{\epsilon}\|^2 = \sum_{i=1}^{m} \hat{\epsilon}_i^2 \tag{3.12}$$

Here,

$$\begin{aligned}\hat{\epsilon} &= Y - X\hat{\beta} \\ &= Y - X(X^TX)^{-1}X^TY \\ &= (\mathbf{I_m} - X(X^TX)^{-1}X^T)Y \end{aligned} \tag{3.13}$$

Let

$$P = X(X^TX)^{-1}X^T \tag{3.14}$$

So,

$$\hat{\epsilon}^T\hat{\epsilon} = Y^T(\mathbf{I_m} - P)^2Y = Y^T(\mathbf{I_m} - P)Y \tag{3.15}$$

As $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_m)$ and $Y = X\beta + \epsilon$ we get that,

$$Y \sim \mathcal{N}(X\beta, \sigma^2\mathbf{I}_m) \tag{3.16}$$

Note that $P = P^T$ and $P = P^2$. Hence, P is an orthogonal projector[11] into column space of P along column space of $\mathbf{I}_m - P$ which is the same as the null space of P. So, $\mathbb{R}^m = \mathrm{C}(P) \oplus \mathrm{C}(\mathbf{I}_m - P)$, where $\mathrm{C}(P)$ denotes the column space of P. Thus, rank of $\mathbf{I}_m - P$ is $m-(n+1)$, as $rank(P) = rank(X) = n+1$. Also $\mathbf{I}_m - P$ is symmetric and idempotent, it follows that,

$$\frac{\mathrm{SSR}}{\sigma^2} = \frac{Y^T(\mathbf{I}_m - P)Y}{\sigma^2} \sim \chi^2_{m-(n+1)}$$

This result follows from the theory of quadratic forms of multivariate normal distributions, If $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathbf{A}$ is symmetric, idempotent, and of rank r, then,

$$\frac{\mathbf{z}^T\mathbf{A}\mathbf{z}}{\sigma^2} \sim \chi^2_r \tag{3.17}$$

Therefore, the sum of squared residuals divided by the variance follows a chi-squared distribution,

$$\frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2_{m-n-1} \tag{3.18}$$

3.1.4 Confidence Interval of $\boldsymbol{\beta_i}$

Previously, we have discussed that,

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 (X^T X)^{-1}_{ii})$$

and

$$\frac{\text{SSR}}{\sigma^2} = \frac{(m-(n+1))s^2}{\sigma^2} \sim \chi^2_{m-(n+1)}$$

where

$$s^2 = \sum_{i=1}^{m} \frac{\hat{\epsilon}_i^2}{m-(n+1)} = \sum_{i=1}^{m} \frac{(y_i - \hat{y}_i)^2}{m-(n+1)} \tag{3.19}$$

Note that s^2 is an unbiased estimator of σ^2. Consequently,

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{s^2 (X^T X)^{-1}_{ii}}} \sim t_{m-(n+1)} \tag{3.20}$$

Based on this result, $1-\alpha$ confidence interval for β_i is,

$$\left(\hat{\beta}_i - s t_{\alpha/2, m-(n+1)} \sqrt{(X^T X)^{-1}_{ii}}, \hat{\beta}_i + s t_{\alpha/2, m-(n+1)} \sqrt{(X^T X)^{-1}_{ii}} \right) \tag{3.21}$$

Here $t_{\alpha/2, m-(n+1)}$ refers to the $\alpha/2$ quantile of a t-distribution with $m-(n+1)$ degrees of freedom.

The summarised results are-

1. OLS estimator of β is $\hat{\beta} = (X^T X)^{-1} X^T Y$.
2. $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2 (X^T X)^{-1}_{ii})$
3. Confidence interval of β_i is

$$\left(\hat{\beta}_i - s t_{\alpha/2, m-(n+1)} \sqrt{(X^T X)^{-1}_{ii}}, \hat{\beta}_i + s t_{\alpha/2, m-(n+1)} \sqrt{(X^T X)^{-1}_{ii}} \right)$$

3.2 Diagnostics

3.2.1 Multiple R-squared

If y is the dependent variable estimated by $\hat{y}$, then the multiple coefficients of determination[16] R^2 is given by,

$$R^2 = 1 - \frac{SSR}{SS_{yy}} = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_i)^2} \tag{3.22}$$

This represents the proportion of the total sample variation in y that can be explained by the multiple regression model.

3.2.2 Adjusted R-squared

The use of an adjusted R^2 is an attempt to account for the phenomenon of the R^2 automatically increasing when extra explanatory variables are added to the model. There are many different ways of adjusting. This is by far the most used one. We will denote it by $\bar{R}^2$.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{m-1}{m-n-1} \tag{3.23}$$

3.2.3 Leverage

The leverage[19] score for the ith independent observation x_i is given as:

$$h_{ii} = X_i^T (X^T X)^{-1} X_i = \frac{\partial \hat{y}_i}{\partial y_i} \tag{3.24}$$

Here, X_i is the ith column of X. It can be interpreted as the degree by which the ith dependent value influences the ith fitted (predicted) value. High leverage points have a stronger capacity to shift the regression line and can be influential, causing the outcome and accuracy to be distorted. A common rule to identify high leverage is to check whether it is greater than $\frac{2}{n}\sum_{i=1}^{n} h_{ii}$.

3.2.4 Cook's Distance

Data points with large residuals and/or high leverage may distort the outcome and accuracy of a regression. Cook's distance[17] measures the effect of deleting a given observation. Cook's distance of ith observation

is defined as,

$$D_i = \frac{1}{(n+1)s^2}\sum_{j=1}^{m}(\hat{y}_j - \hat{y}_{j(i)})^2 = \frac{e_i^2}{(n+1)s^2}\left[\frac{h_{ii}}{(1-h_{ii}^2)}\right] \tag{3.25}$$

where $\hat{y}_{j(i)}$ is the estimated value of y_j deleting the ith observation from the initial data. If Cook's distance of large number of variable X_i's are high (generally greater than 1), then this situation is problematic.

3.2.5 Diagnostic plots

1. **Residual vs Fitted plot**
   It is a scatter plot of residuals on the y axis and fitted values (estimated responses) on the x axis. The plot is used to detect non-linearity, unequal error variances, and outliers. In an ideal residual vs. fitted plot, points should be randomly scattered around zero, without any discernible patterns or trends, suggesting a good fit of the model and adherence to the assumptions of linearity and homoscedasticity.

2. **Normal QQ-plot of Residuals**
   This is a scatter plot with theoretical normal quantiles on the x-axis and quantiles of residues in y-axis. This plot diagnoses if the residuals are normally distributed. Ideally, the points in the plot should fall along a specific straight line, which indicates that the residuals follow a normal distribution.

3. **Scale location plot**
   The standardised residual is given by $\frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$. The intensity of the discrepancy between actual and predicted values is measured by the standardised residual. A scale-location plot is a scatter plot that shows the square root of the absolute value of standardised residuals against fitted values. If residuals are spread uniformly across predictor ranges, then our assumption of homoscedasticity is valid. This is how we may test the equal variance assumption graphically.

4. **Residuals vs Leverage Plot**
   This is a scatterplot of standardized residulas against leverage. Influential data points that might distort the outcome and accuracy of a regression can be detected using this plot. The values on the top right or bottom right corners outside the Cook's distance curve are considered problematic.

3.3 Regresssion using Factor Scores

The factor scores obtained from factor analysis were used to do regression against "Property Prices". We previously obtained the $m \times n$ matrix F of factor scores, where m was the number of factors and n was the number of data points in sample. So, our model for linear regression is

$$Y = F'\beta + \epsilon \tag{3.26}$$

where Y is the $n \times 1$ vector of property prices and F' is the $n \times (m+1)$ matrix defined by $F' = [1|F^T]$.

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.4790 -0.4419  0.0240  0.4760  3.1577

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.516e-15  3.247e-03   0.000 1.000000
WLS1         7.032e-02  4.190e-03  16.782  < 2e-16 ***
WLS3         8.481e-02  4.416e-03  19.203  < 2e-16 ***
WLS2         6.713e-01  3.860e-03 173.929  < 2e-16 ***
WLS4        -3.652e-02  3.751e-03  -9.737  < 2e-16 ***
WLS6        -1.537e-02  4.089e-03  -3.758 0.000172 ***
WLS5        -4.658e-02  3.818e-03 -12.200  < 2e-16 ***
WLS8         4.816e-02  4.038e-03  11.926  < 2e-16 ***
WLS7         1.014e-01  4.122e-03  24.601  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7449 on 52635 degrees of freedom
Multiple R-squared:  0.4452,    Adjusted R-squared:  0.4451
F-statistic:  5279 on 8 and 52635 DF,  p-value: < 2.2e-16
```

Figure 3.1: Result of regression using factor scores

Regression was performed using `lm()` from `stats` library in R. The output is shown in Figure 3.1. It was observed that the variance explained (Multiple R-squared) was 44.5% which indicated a poor fit of the linear model.

3.3.1 Diagnostic Plots

The diagnostic plots of the regression model are given in Figure 3.2. The key observations are-

1. In the residuals vs fitted plot, there is some linearity towards the end.
2. The QQ-plot indicates that the residuals are approximately normal.

3. In the scale location, the points are not spread randomly. We can see a wave like pattern.

4. None of the points cross the Cook's distance line.

5. Point 6114 is an outlier, as it significantly disrupts the pattern in each graph.

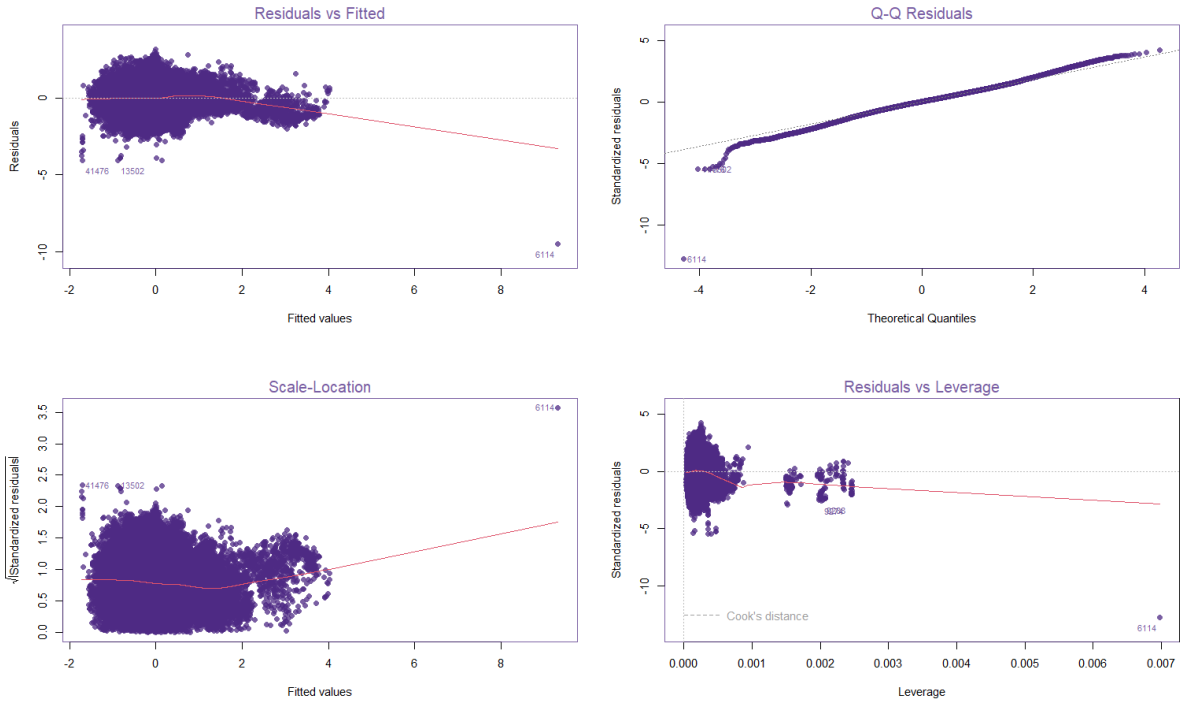

Figure 3.2: Diagnostic plots of regression using factor scores

It is apparent that the model isn't a good fit to the data as there exists significant linearity and heteroscedasticity, along with high unexplained variance. Thus, the model doesn't explain the data well.

3.4 Regression using All Variables

We failed to find an acceptable linear model to predict "Property Prices" through EFA. Hence, we attempted to do regression on all the variables directly. This improved the explained variance to 73.5%, which indicates a satisfactory fit. The results are given in Figure 3.3.

```
Residuals:
     Min       1Q   Median       3Q      Max
-11.0248  -0.3032   0.0093   0.3111   2.2201

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.426e-14  2.246e-03   0.000 1.000000
Longitude         1.982e-01  2.903e-03  68.269  < 2e-16 ***
Latitude         -2.676e-02  2.760e-03  -9.694  < 2e-16 ***
Size              5.842e-01  2.583e-03 226.150  < 2e-16 ***
Floor             3.160e-02  2.700e-03  11.705  < 2e-16 ***
`Highest floor`   8.209e-02  3.272e-03  25.090  < 2e-16 ***
Units             1.277e-01  2.573e-03  49.626  < 2e-16 ***
Parking           8.935e-02  2.522e-03  35.428  < 2e-16 ***
Heating           4.268e-02  2.524e-03  16.911  < 2e-16 ***
Year              2.070e-01  2.612e-03  79.223  < 2e-16 ***
`Dist. Green`    -5.306e-02  7.022e-03  -7.557 4.20e-14 ***
`Dist. Water`    -2.632e-02  2.446e-03 -10.759  < 2e-16 ***
`Green Index`    -1.284e-02  2.339e-03  -5.490 4.03e-08 ***
`Dist. Subway`   -1.565e-01  2.567e-03 -60.963  < 2e-16 ***
`Bus Stop`       -1.683e-02  2.584e-03  -6.514 7.38e-11 ***
`Dist. CBD`      -2.633e-02  7.531e-03  -3.496 0.000472 ***
`Top Univ.`      -1.626e-03  3.968e-03  -0.410 0.681991
`High School`     5.179e-02  4.010e-03  12.916  < 2e-16 ***
`Sex Ratio`      -3.432e-02  2.720e-03 -12.619  < 2e-16 ***
Population       -1.030e-01  3.128e-03 -32.939  < 2e-16 ***
`Pop. Density`    3.864e-02  2.448e-03  15.781  < 2e-16 ***
`Higher Degree`   3.896e-02  2.861e-03  13.617  < 2e-16 ***
`Young Population` 1.429e-01 3.985e-03  35.847  < 2e-16 ***
`Median Age`      1.174e-02  3.272e-03   3.588 0.000333 ***
`Old Population`  1.083e-03  3.981e-03   0.272 0.785710
Spring            1.262e-02  2.883e-03   4.377 1.20e-05 ***
Fall              4.997e-02  3.039e-03  16.445  < 2e-16 ***
Winter            3.500e-02  2.917e-03  11.999  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5153 on 52616 degrees of freedom
Multiple R-squared:  0.7346,    Adjusted R-squared:  0.7345
F-statistic:  5395 on 27 and 52616 DF,  p-value: < 2.2e-16
```

Figure 3.3: Result of regression using all variables

3.4.1 Diagnostic Plots

The diagnostic plots of the new regression model are given in Figure 3.4. The key observations are-

1. The residuals vs fitted plot has improved, but still shows significant linearity.

2. More points are on the line. Thus, the residuals are closer to normal.

3. The scale-location plot too has improved but heteroscedasticity is still present.

4. The point 6114 is beyond the Cook's distance line and thus influences the regression severely.

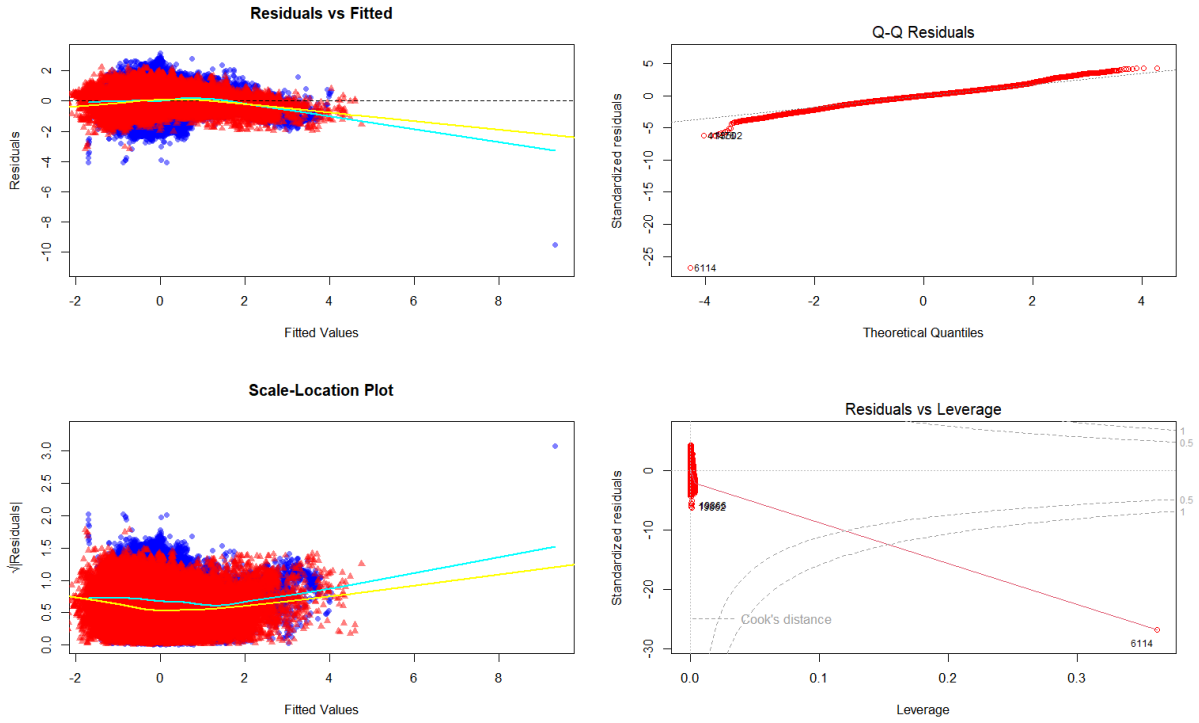

Figure 3.4: Diagnostic plots of regression using all variables

Hence, direct regression gives a slightly better fit than regression after EFA in our case. But, there still are significant issues in the diagnostics that indicate a poor fit of the linear model.

3.5 Hypothesis Test

A statistical hypothesis test is a method of statistical inference used to decide whether the data provide sufficient evidence to reject a particular hypothesis. A statistical hypothesis test typically involves a calculation of a test statistic. Then a decision is made, either by comparing the test statistic to a critical value or equivalently by evaluating a p-value computed from the test statistic.

We have tested the sign of the coefficient of Green Index in our regression model using all variables. The claim was that the coefficient would be positive.

Let the coefficient in question be β_g.

$H_o : \beta_g \geq 0$
$H_a : \beta_g < 0$

The test statistic is

$$T = \frac{\hat{\beta}_g - \beta_g}{\sqrt{s^2 (X^T X)^{-1}_{gg}}} \sim t_{n-(m+1)}.$$

Using R the p-value is obtained as $2.01 \times 10^{-8} < 0.05$. Hence, we reject the hypothesis at 5% level.

Hence, property price decreases linearly with green index if all other variables are fixed.

Chapter 4

Conclusions

The aim of this project was to analyse a hedonic dataset presented in the given data paper[5], which investigated the relationship between property prices and various factors, with a particular focus on urban greenness. The dataset comprised property prices and 27 associated variables, including a newly proposed green index intended to quantify urban vegetation and green amenities.

Our analysis began with an exploratory factor analysis (EFA) to uncover the latent structure underlying the 27 variables. A well-fitting factor structure was identified suggesting the presence of interpretable underlying constructs among the hedonic variables.

Subsequently, we performed a multiple regression analysis using the factor scores obtained from the EFA. However, this model did not exhibit a strong fit, indicating that the factor scores may not sufficiently capture the variability in property prices. In contrast, a full regression model using all 27 variables as predictors resulted in a better fit, but diagnostic plots revealed violations of key assumptions. These issues call for cautious interpretation and perhaps the use of more robust or non-linear modeling techniques in future work.

A key component of our investigation involved testing the hypothesis — that an increase in the green index leads to higher property prices. Logically we'd be inclined to believe that this was true, as people prefer to live in green areas of the city. Contrary to this, our hypothesis test for the regression coefficient of the green index revealed that its effect on property price is negative and statistically significant. Furthermore, the correlation between the green index and property prices was found to be negative, providing additional evidence against the original claim.

These findings highlight the importance of validating empirical claims using robust statistical techniques. While urban greenness is often associ-

ated with positive environmental and health outcomes, its economic valuation—particularly in real estate contexts—can be complex and context-dependent. Future research could benefit from incorporating non-linear models or spatial effects to better understand these relationships.
In summary, our project was able to find an underlying factor structure of the hedonic variables affecting property prices in Busan, South Korea and also found evidence to challenge the conclusion that urban greenness increases property prices in the area. Through the use of factor analysis, regression modeling, and hypothesis testing, we have demonstrated how rigorous statistical methodology can offer deeper insights into real-world data and question surface-level interpretations.

4.1 Limitations

While the analysis in this project offers several insights, it is not without its limitations:

- **Model Assumptions:** The multiple linear regression model using all 27 variables violated key assumptions. These violations suggest that the linear model may not be the most appropriate for capturing the complex relationships present in the data.
- **Factor Interpretability:** Though exploratory factor analysis helped reduce dimensionality, the interpretability of some factors was limited. In certain cases, the loadings did not suggest a clear thematic connection, making practical interpretation difficult.
- **Outliers and Influential Observations:** The dataset contained several outliers and influential points, which may have skewed the regression results. While some diagnostics were performed, a more thorough treatment (e.g., robust regression) was not implemented.
- **Green Index Validity:** The green index used in the dataset, although novel, may not be a comprehensive representation of urban greenness. It is constructed based on google street view and may not fully capture accessibility, usability, or perceived value of green spaces by residents.
- **Geographic and Temporal Scope:** The dataset is limited to a single city (Busan, South Korea) and a single time frame. Thus, the findings may not generalize to other urban contexts or time periods.

4.2 Future Scope

This study opens several avenues for further research and methodological improvement:

- **Robust and Nonlinear Modeling:** Future analyses could explore nonlinear models (e.g., generalized additive models) or robust regression methods that are less sensitive to outliers and assumption violations.
- **Confirmatory Factor Analysis (CFA):** While EFA is exploratory, follow-up work could involve CFA to test specific hypothesized factor structures and validate measurement models.
- **Spatial Analysis:** Given the geographic nature of the data, incorporating spatial regression models or geographically weighted regression could yield more location-aware insights.
- **Broader and Longitudinal Data:** Expanding the dataset to include other cities and multiple years could help test the robustness and generalizability of the findings across different urban environments and over time.
- **Alternative Measures of Greenness:** Developing or incorporating more nuanced indicators of green space—such as accessibility, maintenance, or user perception—could lead to more accurate valuation of environmental amenities.

Bibliography

[1] R.A. Azdy and F. Darnis. Use of haversine formula in finding distance between temporary shelter and waste end processing sites. *Journal of Physics: Conference Series*, 1500(1):012104, 2020.

[2] Chul Baek, Sungho Tae, Ran Kim, and Seungjun Shin. Life cycle co2 assessment by block type changes of apartment housing. *Sustainability*, 8(8):752, 2016.

[3] M. S. Bartlett. A further note on tests of significance in factor analysis. *British Journal of Statistical Psychology*, (1):1–2, Mar. 1951.

[4] Contributors to Wikimedia projects. Regression analysis - wikipedia, July 2004.

[5] Sihyun An et al. Aggregated hedonic dataset with a green index: Busan, south korea. *Data in Brief*, page 111009, Dec. 2024.

[6] Alan E. Hendrickson and Paul Owen White. Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, (1):65–70, May 1964.

[7] Hyeongmin Jang, Kwangwon Ahn, Dongil Kim, and Yena Song. Detection and prediction of house price bubbles: Evidence from a new city. In *Lecture Notes in Computer Science*, volume 10862, pages 782–795. Springer, 2018.

[8] K. G. Jöreskog. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, (2):183–202, June 1969.

[9] Jinfeng Li and Andrew D Heap. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics*, 6(3-4):228–241, 2011.

[10] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, 3rd edition, 2006.

[11] A. Ramachandra Rao and P. Bhimasankaram. *Linear Algebra*. Springer, 2000.

[12] Tenko Raykov and George A. Marcoulides. *An Introduction to Applied Multivariate Analysis*. Routledge, 2008.

[13] Muni S. Srivastava. *Methods of Multivariate Statistics*. John Wiley & Sons, 2002.

[14] Contributors to Wikimedia projects. Linear regression - wikipedia, May 2001.

[15] Contributors to Wikimedia projects. Factor analysis - wikipedia, June 2003.

[16] Contributors to Wikimedia projects. Coefficient of determination - wikipedia, Feb. 2005.

[17] Contributors to Wikimedia projects. Cook's distance - wikipedia, Mar. 2006.

[18] Contributors to Wikimedia projects. Variance inflation factor - wikipedia, Oct. 2007.

[19] Contributors to Wikimedia projects. Leverage (statistics) - wikipedia, Dec. 2008.

[20] Contributors to Wikimedia projects. Exploratory factor analysis - wikipedia, Oct. 2009.

[21] Contributors to Wikimedia projects. Parallel analysis - wikipedia, Jan. 2019.

[22] Contributors to Wikimedia projects. Kaiser–meyer–olkin test - wikipedia, May 2021.

[23] Li tze Hu and Peter M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, (1):1–55, Jan. 1999.

[24] Y Wan and Q Chen. Joint image dehazing and contrast enhancement using the hsv color space. In *Visual Communications and Image Processing (VCIP)*, 2015.