HUMBER INSTITUTE OF TECHNOLOGY AND ADVANCED LEARNING (HUMBER COLLEGE)

MACHINE LEARNING 1 - ASSIGNMENT 1

Assessing the Influence of Urban Rivers and Highway Proximity on Property Values: Insights from Boston Housing Dataset

Submitted by: **Grade/Comments**

Last Name	First Name	Student Number
Angelia	Inge	N01536053
Carvalho	Risa	N01578737
Edward	Rodney	N01600536
Kuriakose	Divya	N01579609
Moromisato	Renzo	N01578865

Submitted to:
Professor Salam Ismaeel

Submission Date: 1st of February 2024

Assessing the Influence of Urban Rivers and Highway Proximity on Property Values: Insights from Boston Housing Dataset

Introduction:

Presence of urban rivers in residential area and how they affect property value

The scarcity of open spaces in urban areas and the relentless pace of modern life have heightened the vulnerability of humans to stress, anxiety, depression, and aggressive behaviors. Researchers have been exploring how the absence of urban rivers exacerbates these negative health impacts. Rivers play a crucial role in our ecosystem, providing natural habitats for flora and fauna and offering significant health benefits to society, particularly in urban areas where green spaces are limited. Studies conducted by UCLA's Center for Occupational and Environmental Health indicate that urban rivers enhance physical and mental well-being, promote community health, and create opportunities for nature-based activities. This is especially evident when rivers are integrated into essential urban features such as homes, parks, workplaces, and schools (American Trails, 2014).

From the viewpoint of homeowners and investors, riverfront properties offer more than just aesthetic appeal. Numerous studies have sought to quantify the positive impact of urban rivers on property prices. This is supported by a study conducted by Nicholls (2017), which summarized the results of 25 studies showing significant positive effects on property prices, especially in urban areas.

However, despite their beauty and potential health benefits, urban rivers also present the risk of water pollution, which can have detrimental effects on health if not properly managed. A case study in North Carolina and Tennessee illustrates this concern. Pollution from the Pigeon River has led to decreased property values in both states, irrespective of whether individuals benefit economically from the Paper Mill responsible for the pollution (Cho et al., 2011).

Presence of radial highways in residential area and how they affect property value

The presence of highways within the proximity of residential areas could be a double-edged sword. While it offers convenience and increases the area’s accessibility, it could also increase the exposure to traffic intensity and noise pollution that could affect potential homeowner’s decisions when purchasing a property. A research conducted by Levkovich et

al. (2015, p.1) in the Netherlands showed that the development of highways correlated positively with the effect on the house prices on the nearby municipalities. On the other hand, a study conducted by Kim et al. in 2007 showed completely different results in Seoul, South Korea, where the 1% increase in highway traffic noise corresponded to 1.3% decline in land price.

Purpose and Approach:

This report aims to investigate the relationship between the presence of urban rivers, specifically the Charles River (CHAS), and radial highways (RAS) on housing prices in Boston. By examining the correlation between these urban features and property values, we seek to understand how proximity to rivers and highways influences housing prices in the Boston area.

To achieve our research objective, we will conduct a quantitative analysis using data from the Boston housing database that has been collected by Harrison and Rubinfeld (1978). Utilizing Python programming, we will calculate correlation coefficients to assess the strength and direction of the relationship between river and highway proximity and housing prices. By employing this approach, we aim to provide insights into the impact of urban rivers and radial highways on property values in Boston, without utilizing regression analysis.

Data Overview:

The data consists of 14 variables with 506 entries that were collected in 1978. In this data, the main objective is to see the median value of the property price based on the independent variables listed in the following table below in order:

CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	Proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	Nitric oxides concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centres
RAD	Index of accessibility to radial highways (from 1 to 24)

TAX	Full-value property-tax rate per $10,000
PTRATIO	Pupil-teacher ratio by town
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	Lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

Table 1: Independent variables

Data Cleaning:

To ensure data integrity, several necessary steps were taken to clean the data by using Python. Employing essential tools such as Pandas, Matplotlib, Seaborn, and Numpy libraries within the Jupyter Notebook environment.

Checking data type

The first step is to check the independent variable data type associated in this dataset. From this step it is evident that all variables are in the correct data type

Variable	**CRIM**	**ZN**	**INDUS**	**CHAS**	**NOX**	**RM**	**AGE**
Data Type	float64	float64	float64	int64	float64	float64	float64
Variable	**DIS**	**RAD**	**TAX**	**PTRATIO**	**LSTAT**	**MEDV**	**CAT. MEDV**
Data Type	float64	int64	int64	float64	float64	float64	int64

Table 2: Variable Data Type

Renaming column

In order to minimize error, the spaces in the CAT. MEDV column was eliminated and renamed into CAT_MEDV.

Checking null values

The next step in data cleaning is to check if there are any null values that might affect the descriptive statistics of the dataset in the future calculations. As shown in Table 2, every variable shows the same count value at 506, and by utilizing the isnull() and sum() function, the dataset shows there are no null values detected at all.

Var.	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
Length	506	506	506	506	506	506	506
Missing Value	0	0	0	0	0	0	0
Var.	**DIS**	**RAD**	**TAX**	**PTRATIO**	**LSTAT**	**MEDV**	**CAT_MEDV**
Length	506	506	506	506	506	506	506
Missing Value	0	0	0	0	0	0	0

Table 3: Count and Missing Values of Each Variables

Replacing null values with NaN

After checking for the presence of any null values in the Boston housing data, we carried out measures to replace undetected missing values with NaN, which stands for “Not a Number”. This is an intermediary step taken to provide a consistent representation for missing and undefined values across different data types. Replacing null values with NaN is essential in certain scenarios where it is necessary to clearly identify missing or undefined values. Some examples include data manipulation with Python and corresponding data libraries as well as statistical operations to compute the mean, median, or standard deviation by excluding NaN values.

Data Imputation:

Despite the absence of null values, we adhered to data imputation protocols that will replace missing values with the average of that particular value’s attribute column. This is done as a precaution in the event that the data cleaning steps performed previously fail to catch missing values. For example, if a missing value was discovered within the RAD attribute column, that missing value would be replaced with the average of that column’s existing values.

Var.	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE
Average	3.61352	11.3636	11.1367	0.06917	0.55469	6.28463	68.5749
Var.	**DIS**	**RAD**	**TAX**	**PTRATIO**	**LSTAT**	**MEDV**	**CAT_MEDV**
Average	3.79504	9.54940	408.237	18.4555	12.6530	22.5328	0.1660

Table 4: Average value of each variable of the dataset

Descriptive Statistics:

After performing the respective cleaning and imputation of variables we have the following results on how a property being near the Charles River and the index amount of radial highways impacts the median value of owner-occupied homes, as shown in figure 1.

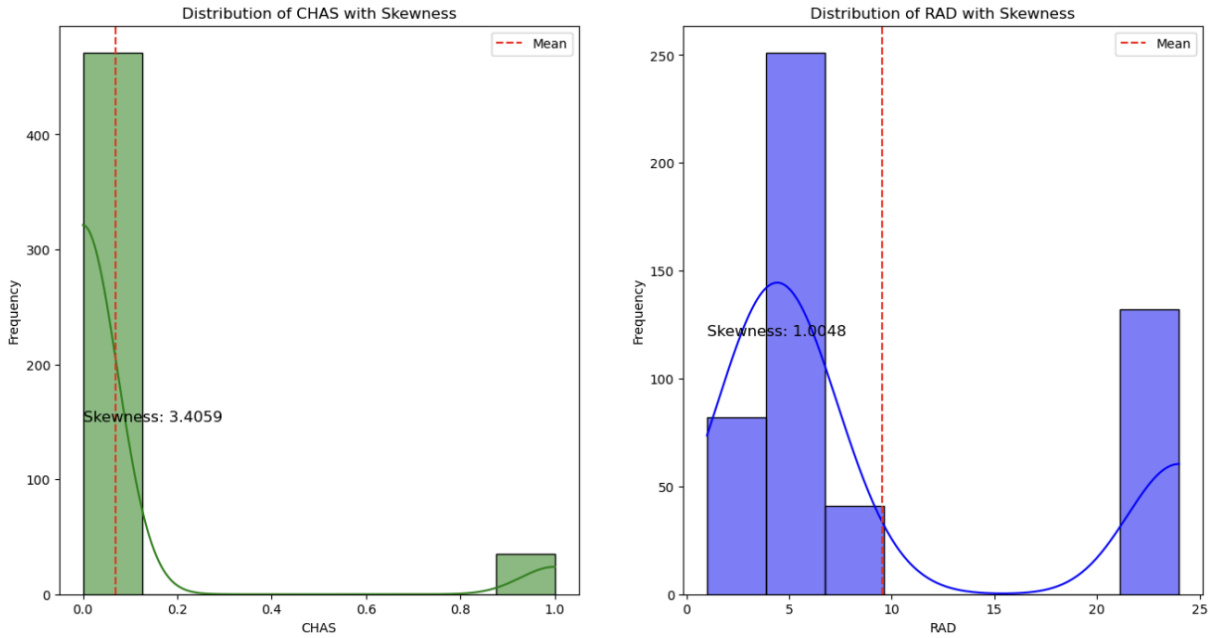

Figure 1 : Histogram and skewness of the variables "CHAS" and "RAD"

As we can observe, the histogram of the variable CHAS, shows only 2 values (0 and 1) as it is a boolean variable, however, the amount of properties in close proximity to the Charles River are significantly lower than the properties that are not close to the river. On the other hand, most properties have between 1 and 10 radial highways, with the exception of a couple of properties that have a significant amount of radial highways.

In Figure 1, we identified the skewness of the CHAS and RAD variables. Typically, as a rule of thumb, skewness value exceeding 1 indicates significant skewness, with a positive value suggesting right-skewness (Jagdeesh, n.d). The CHAS variable has a skewness of 3.405904, while the RAD variable exhibits a skewness of 1.004815. Both variables are heavily skewed to the right, meaning that outliers would tend to fall on the right side of the data.

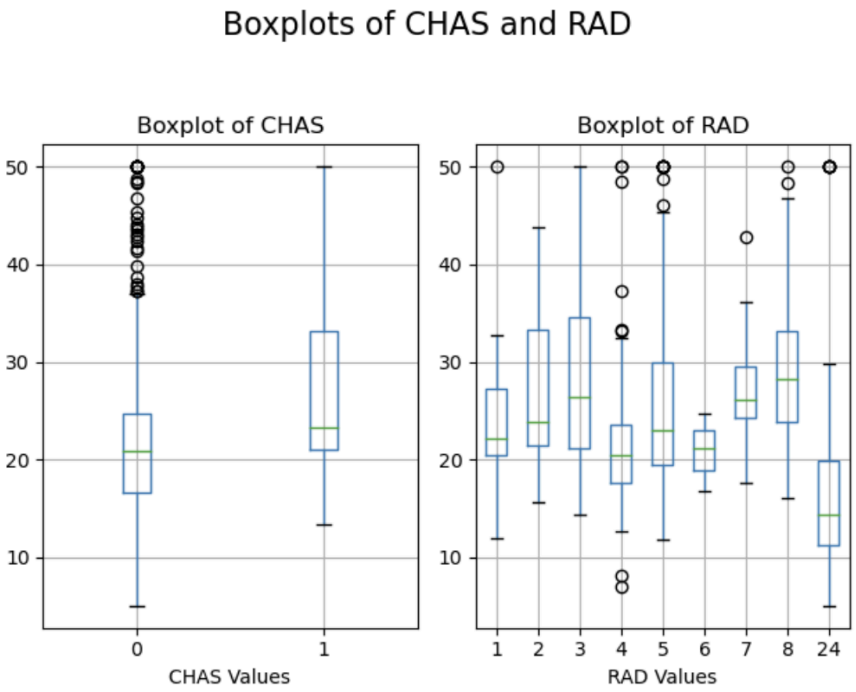

Figure 2 : Boxplots of the variables "CHAS" and "RAD"

From the boxplots shown in figure 2, we can observe that the median price in properties with no proximity to the Charles Rivers have a significant amount of outliers compared to the ones in proximity to the river. Moreover, As a property has more accessibility to radial highways, the boxplots show more outliers.

Correlation Analysis:

Following the descriptive analytics using the histogram and boxplots, we performed a correlation analysis of the Boston housing data which is captured in the heatmap below. The analysis studies how the median value of owner-occupied homes (MEDV) correlates with the presence of the Charles River (CHAS) and how it correlates with accessibility to radial highways (RAD). From viewing the heatmap, we can see that the correlation coefficient between MEDV and CHAS is 0.2. This indicates a relatively weak positive relationship between MEDV and CHAS. Therefore, if a tract is bound by the Charles River, it would have little positive influence on the median value of owner-occupied homes.

On the other hand, the correlation coefficient between MEDV and RAD is -0.4. This is indicative of a relatively moderate negative relationship between MEDV and RAD. Therefore, accessibility to radial highways would have a moderately inverse influence on the median value of owner-occupied homes.

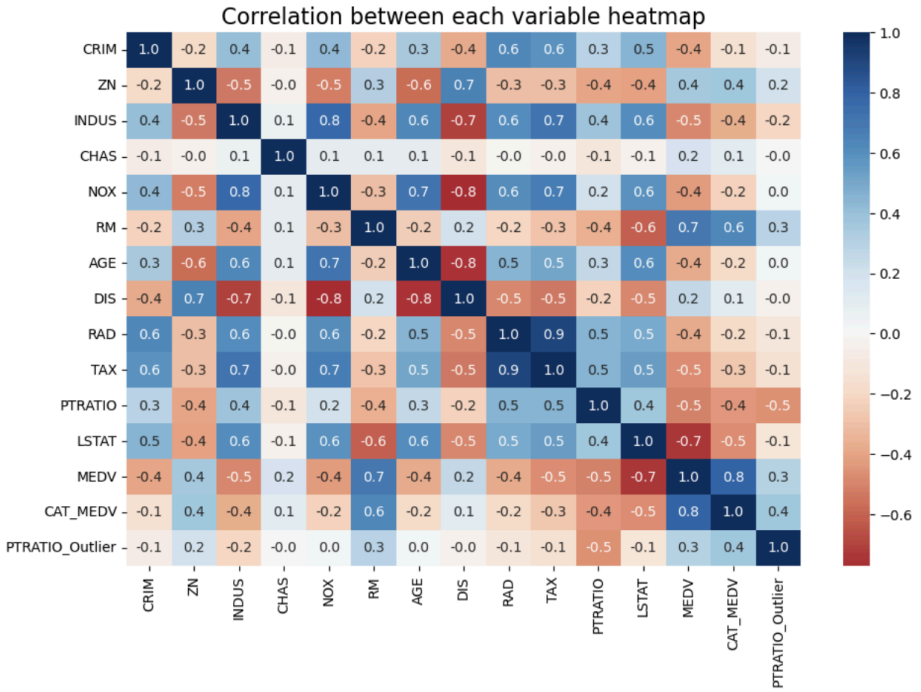

Figure 3: Correlation between each variable heatmap

Results and Discussion:

Impact of Urban Rivers (CHAS) on Property Values:

Property values in Boston are significantly correlated with being close to the Charles River (CHAS), according to the data. The median price of real estate near the Charles River is often higher (MEDV). This result is consistent with other research that indicates urban rivers raise residential areas' aesthetic appeal and desirability, which in turn raises property values. The Charles River is there, adding to the area's natural beauty while also providing recreational options and enhancing the wellbeing of the locals. However, as the Pigeon River case study makes clear, it's critical to recognize the possible risk of water pollution connected to urban waterways. Using efficient management techniques is essential to reducing the hazards of pollution and maintaining the beneficial effects of urban rivers on property values.

Effect of Radial Highways (RAD) on Property Values:

The data reveals a moderate negative correlation between property values in Boston and the proximity to radial roadways (RAD). As the RAD index increases, indicating greater accessibility to highways, there is a trend of decreasing median home prices. This suggests that while optimal highway connectivity may initially raise property values, an excessive presence of radial highways can lead to a decline in home prices due to associated issues such as noise pollution and traffic volume.

Interestingly, residences with RAD scores of 8 exhibit the highest median prices, indicating that there is an ideal level of highway connectivity that positively impacts property values. However, beyond a certain threshold, typically indicated by a RAD index above 3, the median home price tends to decline. This highlights the delicate balance between accessibility and livability in urban areas, where an excessive presence of radial highways may detract from the desirability of residential properties.

In summary, while highway connectivity plays a significant role in shaping property values, it is essential to consider the potential negative effects of excessive accessibility on homeowner decisions and property values. Urban planning strategies should aim to optimize highway infrastructure while mitigating associated drawbacks to ensure sustainable and desirable living environments in Boston.

Correlation Analysis and Data Insights:

The correlation study highlights how several factors interact to influence Boston real estate values. Robust associations among specific variables, like TAX and RAD, underscore the interdependent character of urban attributes and socio-economic determinants. Removing superfluous variables such as TAX and RAD may improve the quality of the dataset while maintaining meaningful correlations. Additionally, since they might point to underlying socioeconomic inequalities and environmental factors influencing property values, skewed distributions in variables like CRIM, ZN, and CHAS merit investigation.

Conclusions:

To sum up, this study's results provide insight into the complex interplay of Boston's radial freeways, urban rivers, and property values. The analysis revealed that proximity to the Charles River, which is typically associated with higher property values, does not always generate a consistent trend. Other features, apart from proximity to the Charles River, may have enough power to command higher prices. Likewise, the accessibility of a property to a radial highway affects its value, highlighting the impact of transportation infrastructure on residential choices and market dynamics. To optimize the benefits on real estate prices while tackling any environmental and social issues, efficient urban river management and transit planning are crucial. In order to create thriving and resilient communities in Boston and elsewhere, politicians, urban planners, and stakeholders should give priority to sustainable development initiatives that improve urban livability and encourage fair access to facilities.

References:

American Trails. (2014). Study Cites Health Benefits of Urban River Parkways. *American Trails.* Retrieved from https://www.americantrails.org/resources/study-cites-health-benefits-of-urban-river-parkways

Cho, S.H., Roberts, R.K, & Kim, S.G. (2011). Negative externalities on property values resulting from water impairment: The case of the Pigeon River Watershed. *Ecological Economics 70,* 2390 - 2399. https://doi.org/10.1016/j.ecolecon.2011.07.021

Harrison, D., & Rubinfeld, D. L. (1978). Boston Housing. *Carnegie Mellon University*. Retrieved from *http://lib.stat.cmu.edu/datasets/boston*

Jagdeesh. (n.d). Skewness and Kurtosis – Peaks and Tails, Understanding Data Through Skewness and Kurtosis. *Machine Learning* +. Retrieved from https://www.machinelearningplus.com/statistics/skewness-and-kurtosis/?utm_content=cmp-true

Kim, K.S., Park, S.J., & Kweon, Y.J. (2007). Highway traffic noise effects on land price in an urban area. *Transportation Research Part D: Transport and Environment 12,* 275 - 280. https://doi.org/10.1016/j.trd.2007.03.002

Levkovich, O., Rouwendal, J. & van Marwijk, R. (2016). The effects of highway development on housing prices. *Transportation 43*, 379–405. https://doi.org/10.1007/s11116-015-9580-7

Nicholls, Sarah & Crompton, J.. (2017). The effect of rivers, streams, and canals on property values. *River Research and Applications*. https://doi.org/10.1002/rra.3197