University of Amsterdam
Economics and Business Economics
Bachelor's Thesis

The Impact of Train Station Presence and Train Traffic on House Prices in Dutch Municipalities

Author:

Levente Szabó – 13614878

Supervisor:

Lisa Marie Timm

June 2024

Statement of Originality

This document is written by Levente Szabó who declares to take full responsibility for the contents of this document. I declare that the text and the work presented in this document are original and that no sources other than those mentioned in the text and its references have been used in creating it. I have not used generative AI (such as ChatGPT) to generate or rewrite text. UvA Economics and Business is responsible solely for the supervision of completion of the work and submission, not for the contents.

Table of Contents

Introduction

Passenger rail transport holds an important place in the Dutch mobility sphere. According to the NS Annual Report, the share of train travel in total mobility in the Netherlands exceeds 12%, and total traffic has been sharply increasing since the COVID-19 crisis (Nederlandse Spoorwegen, 2024). Rail infrastructure is also extensive, featuring 401 train stations - more than the number of municipalities. Concurrently, housing prices have exhibited an upward trend, with a 30% increase in the average sales price of owner occupied homes from 2020 to 2024 (Centraal Bureau voor de Statistiek, 2024). Understanding these phenomena and examining whether a connection between rail infrastructure, rail travel and housing prices exist in The Netherlands is crucial for policy considerations.

The link between accessibility and house prices is a central topic of spatial and transport economics and thus have been widely documented in the past. The idea that basic economic models are built upon is that accessibility can improve connectivity by reducing transportation costs to and from an area, and by doing so, increasing local property prices (Alonso, 1964; von Thünen, 1826). Higher connectivity can also stimulate local economies leading to economic growth (Jiao et al., 2020). In the case of rail traffic and rail infrastructure, the added value of accessibility it brings to an area can not be ignored, making the effects of these variables a widely researched topic in the field of transport economics. And while the general consensus seems to suggest that a positive relationship exists, results vary across contexts, methodology, and datasets used leading to a discussion on the added value of accessibility versus the negative externalities higher train traffic might induce. Despite extensive research, findings are mixed even in a small country like The Netherlands.

Most existing Dutch studies deal with the direct proximity of a train station and investigate the effect on prices of property transactions nearby, consequently zooming in around specific stations and handling local effects. Furthermore, there is also a lack of papers examining the potential effects of train frequency. On the contrary, this paper is to be interpreted as a descriptive analysis that takes a bird's-eye view of the entire country which compares municipalities by train station presence, train frequency and average house prices. Thus, the central question of the paper is not necessarily about the local effects of the proximity of a train station - as previous research has adressed this various contexts - but rather, whether the price effects around a train station in turn also impact the average level in the municipality. Does train station presence effect house prices so significantly that it has to be taken into account in infrastructure related policies on the municipal level? What is the impact of train frequency? How are frequency effects different in municipalities with lower of higher population density, or distance from urban centers?

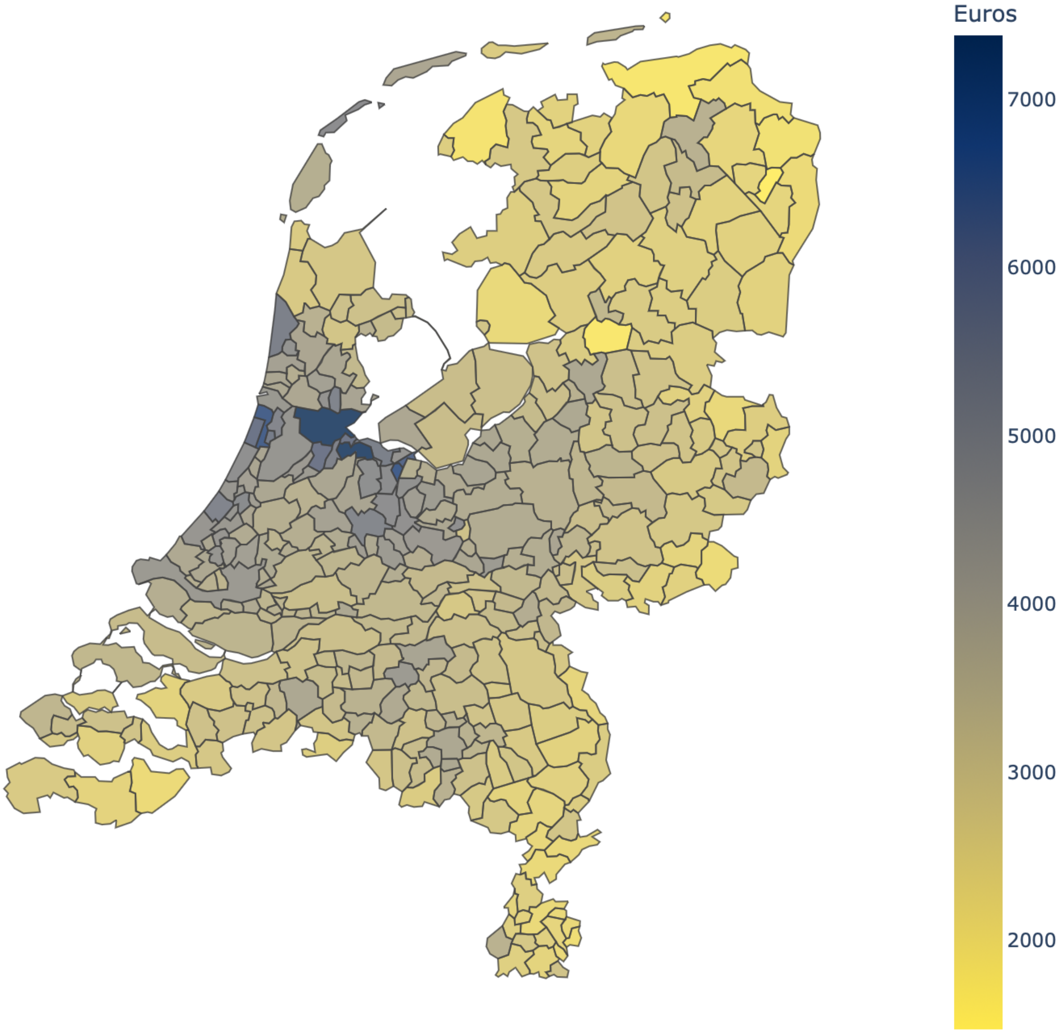


Figure 1: Square Meter House Prices by Municipality in The Netherlands (2023)

Consequently, model selection will be reconsidered. Studies on the effect of the proximity of train station on housing prices generally apply a model called hedonic pricing model, which focuses on the internal characteristics of an individual property, and the external characteristics of it's location (Rosen, 1974). This paper uses a model that is similar to the hedonic pricing model, however instead of focusing on individual properties, it focuses on real estate markets of municipalities. It tries to capture the internal structure of the housing market of a municipality through incorporating variables such as ratio of multy-family homes, or homes per capita while controlling for external factors affecting housing prices, such as average income level. The purpose of this approach is to potentially correct for station specific variability and to take a broader perspective.

Overall, this paper seeks to provide insight for policymakers by mapping the effects of municipal train station presence and volume of train traffic on housing prices and in doing so, trying to reduce station-specific variability. By deviating from past research and introducing a nation wide model with a broader focus, it seeks to uncover municipal level effects that might be obscured by in a more granular analysis. Furthermore, it tries to bridge the gap in literature by conducting a thorough analysis on the effects of train frequency on house prices in The Netherlands.

The paper is structured as follows: in the next section, relevant literature will be dealt with, mainly focusing on the effect of train station presence and train frequency on housing prices. Findings, and reasons for their variability will be examined in general and in a national context. The subsequent sections introduce the dataset and methodology used. The results of the analysis are then presented and the paper concludes with a discussion of the findings, limitations, and implications for future research.

Literature Review

Effect of Accessibility on Property Prices

The effect of accessibility on property prices is a popular topic in economic research, with significant implications for regional development, urban planning and transport economics. The first link between the two was documented by von Thünen (1826). He tried to explain variations in the rent price of agricultural land by examining their distance from the market. According to von Thünen, land rent prices can be explained by the following formula:

$$R = Y(p - c) - YFm$$

This formula is specified as follows:

R - Rent price of land,
Y - Yield on land,
p - Market price,
c - Production costs,
F - Transportation costs,
m - Distance to market

Based on this model, it is evident that for agricultural products with higher transport costs, such as dairy which has to be transported quickly, or wood which is difficult to transport, the distance to market from the land should be low in order to maximize profits.

The theories of von Thünen later contributed to the work of William Alonso, who developed the bid rent model. According to the bid rent theory, economic agents all compete for land closest to the central business district (CBD), and are willing to pay a certain amount, called bid rent (Alonso, 1964). According to Alonso, residential, commercial and manufacturing agents are endowed with different bid rent curves, with residential agents willing to pay the least, and commercial agents willing to pay the most for land close to the CBD, as here, the concentration of customers is the highest - meaning transportation costs are the lowest. This leads to a gradually decreasing bid rent curve as one moves away from ther CBD, and explains why many of the commercial and office buildings are usually located in the city center, and residential buildings on the outskirts.

Overall, it is clear that basic economic models build upon the idea that land value is higher closer to a central point where demand concentration is high. For economic agents to maximize profits, they need to reduce transportation costs to the CBD. Improved accessibility is a key means of achieving this.

Effect of Train Station Presence and Train Traffic on House Prices

There is a general consensus in literature over the accessibility benefits of train stations raising house prices (Paliska & Drobne, 2020), however past papers on the impact of a train station on property values present the reader with mixed results. However, these studies vary across economic context, countries, type of railway (light rail, metro, commuter rail, or heavy rail), and several other factors. This section of the literature review is dedicated to present some of these findings, and to discuss potential reasons for their variability.

There is a substantial body of literature dealing with positive effects of train station proximity and train frequency on house prices. Dubé et al. (2013) investigated the impact of implementing a commuter train system between Montreal, Canada and it's southern periphery on house values. They utilize a difference-in-differences hedonic pricing model for single-family house sales between between 1992 and 2009. Their findings show that implementing the rail system results in a 2.6% overall mean house price for the entire southern area, which translates to a total increase of more than a billion dollars for the local housing market. Similarly, Syabri (2011) conducted an analysis utilizing a spatial hedonic pricing model for properties around the Serpong station in in Jakarta, Indonesia. He examines the effect of the proximity of the train station on rent prices in the neighbourhood. His findings reveal a positive relationship, that is, a negative rent price gradient as one moves away from the station. Bohman and Nilsson (2016) examine the effect of commuter rail station proximity on property prices across various market segments in the Scania region of Sweden. They find that there is a negative relationship between station distance and property prices for all price segments, and this effect is stronger for lower segments. This also suggests a negative price gradient when the distance to the station increases. Additionally, they include departure frequency in their model, and find that higher train frequency leads to higher property prices for all segments. Lieske et al. (2021) also find evidence for this: their results show that increasing relative frequency of peak hour trains leads to higher accessibility and thus higher property prices.

On the contrary, there exist studies that either do not find any significant effects, or the effects are negative. Camins-Esakov and Vandegrift (2018) study the impact of an extension to a light-rail line in New Jersey, examining the differences in prices of homes in the area pre and post station announcement. They find that no significant effects can be derived. Lieske et al. (2021) find disamenity effects in the 400m radius of train stations in the Western Sydney area. They find negative effects specifically for stations that include a parking lot. Additionally, Paliska and Drobne (2020) also conducted a study in rural Slovenia, and show that there is no significant price effect resulting from proximity to station. This indicates that price effects can differ in rural and urban areas.

There are several studies that seek to explain these mixed findings. There exists an argument that exposure to environmental stressors in close proximity to railway lines may offset the benefits of improved accessibility. For example, Maclachlan et al. (2018) conducted a study in Sweden on the relationship between distance from the railway and vibration-related annoyance. Their results indicate that, depending on the train type, annoyance can be moderate to high up to 400 meters. They argue that this has significant implications for property planning. Furthermore, Bowes and Ihlanfeldt (2001) also identify two factors that counteract the positive price effects of connectivity and accessibility. Firstly, there are emission-related externalities, such as noise, pollution, or unsightliness of the station, especially if it includes a parking lot. Secondly, they argue that a train station provides easier access to the area for outsiders, which might translate to higher crime rates

Due to the mixed results in present in literature, Debrezion et al. (2007) conducted a meta-analysis featuring 57 studies in order to seek out a systematic conclusion on the impact of railway station proximity on property values. They use seven categorical variables in order to isolate variation in results, namely: type of property under consideration, type of railway station, type of model used to derive the valuation, the presence of specific variables related to accessibility, demographic features and the time of the data. Their findings show that commuter railway station affect house prices significantly more than other types of stations. Furthermore, they find that the price effect on commercial properties differs from residential properties. According to their research, locally, within a quarter mile range from the station, commercial properties sell or rent 12% higher than residential properties. However, on a global scale, for every 250m coming closer to the station the railway station effect is 2.3% higher for residential properties than commercial properties. Lastly, they investigate the effects of including other types of accessibility variables. They find that when other accessibility variables are part of the model besides station proximity, station effects are generally smaller. They specifically find omitted variable bias, when highway accessibility is not explicitly considered in a model.

In conclusion, it is evident when looking at the general literature, that results vary due to several factors. In many cases results are mitigated by negative externalities such as noise pollution or increased crime rate and can be overestimated due to omitted variable bias. Findings can further vary based on property and railway categories and whether the station is located in an urban or rural area. However, it is also clear that accessibility benefits of train stations and train frequency is well documented and cannot be ignored.

Effect of Traffic and Proximity in The Netherlands

In order to further explore the research question at hand, this section of the literature review deals with research conducted in the national context. Studies done in The Netherlands show similar results and goals as those of international context and the results also vary from study to study. The main focus of literature remains to be train station proximity in general, however, some papers do highlight the effect of traffic volume.

Debrezion et al. (2005) conducted a study on the impact of railway stations on the Dutch housing market. This provides context for understanding local effects in The Netherlands. Their dataset includes the entirety of The Netherlands, with data ranging from 1985 to 2001. Their research, utilizing a cross-sectional hedonic price model, reveals a significant positive relationship between train frequency and house prices, with an elasticity close to 0.03 for properties within 2 kilometers of a station. This implies that a doubling of train frequency can increase house prices by 3%. Additionally, they highlight the effects of proximity to railway lines. According to their research, while immediate station proximity generally increases property values, extreme closeness can introduce negative externalities such as noise, which could negate these benefits. The study also considers highway accessibility, finding optimal benefits at 4-5 kilometers from highway entry points. These findings strengthen the results of research in international context, specifically that there is a balance between accessibility benefits, and negative externalities of a train station leading to intricacies in the results.

Koster et al. (2012) investigate the effects of new railway station openings on house prices in Dutch cities between 1995 and 2007 using an extensive repeated sales dataset. Contrarily, their findings do not reveal any statistically significant impact of station openings on house prices. They attribute this lack of effect to several factors: the small size and suburban location of the new stations, which might offer limited travel time savings, potential negative externalities like noise and crime, and the relatively low share of train trips in overall travel, suggesting that railway proximity is not a major determinant factor of house prices for a large portion of the Dutch population.

Lastly, Debrezion et al. (2011) conducted research on the impact of railway stations on residential property values in the Netherlands. They utilized a hedonic pricing model based on sales data from three metropolitan areas: Amsterdam, Rotterdam, and Enschede. The study measured railway accessibility through both the distance to the nearest railway station and the quality of services provided at the station, represented by the Railway Service Quality Index (RSQI). They find that the model considering the most frequently chosen station by residents outperforms the one based on the nearest station in estimating the effect of railway accessibility on property prices. Their results also show significant differences in the impacts of railway accessibility between more and less urbanized areas, with urbanized areas experiencing a stronger effect. The results of the study imply that next to railway proximity, the quality of railway services may also influence local real estate values.

Overall, these studies place the findings of international research in the national context of The Netherlands. Nevertheless, the conclusions prove to be the same. While the proximity of railway stations and frequency of traffic seem to increase house prices in general, no definitive conclusions can be made due to negative externalities and variations in context and model choice.

Data

This section provides an overview of the data used in the analysis. The dataset comprises information on all 343 municipalities of The Netherlands, focusing on housing prices, demographic characteristics, and rail infrastructure. The data consists of observations only from the year 2023. The reason for this is to reduce complexity, specifically that municipality borders frequently change in The Netherlands, making it difficult to create a general model based on multi-year panel data that includes all of the municipalities. The data was obtained from multiple sources, including CBS Statline and Rijden De Treinen [1]. The dataset contains the following variables:

- **municipality**: The name of the municipality.
- **m2_price**: Average price per square meter of housing, measured in euros.
- **pop_density**: Population density, measured as the number of people per square kilometer.

[1] The full list of data sources can be found in Appendix 1.

- **avg_income**: Average annual income per capita in thousands of euros.

- **homes_per_capita**: Total number of owner-occupied homes per capita in the municipality.

- **unemp_rate**: Unemployment rate as a percentage of the labor force.

- **net_labor_participation**: Net labor participation rate, measured as the percentage of population (labour force and not labour force).

- **multy_family**: The share of multi-family housing units in total number of owner-occupied homes in the municipality.

- **distance_to_urban_center**: Distance from the central point of the municipality to the nearest urban center in kilometers.

- **station_count**: Number of railway stations within the municipality.

- **traffic**: Average daily traffic at railway stations, measured by the number of trains scheduled to stop at any of the municipality's stations.

- **has_station**: Indicator variable for the presence of at least one railway station in the municipality (1 if there is at least one station, 0 otherwise).

The summary statistics for all numeric variables are presented in Table 1. The table provides the mean, standard deviation, minimum, and maximum values, as well as the 25th, 50th, and 75th percentiles for each variable.

To elaborate on the most important variables, firstly it can be seen that the average price per square meter (*m2_price*) varies significantly, with a mean of 3196.45 euros and a standard deviation of 878.96 euros. The highest observed price per square meter is 7378.71 euros, while the lowest is 1467.73 euros.

Secondly, the variables *homes_per_capita* and *multy_family* were included in order to control for the internal structure of the housing market in municipalities. The table shows that *homes_per_capita* has a mean of 0.45 and ranges from 0.32 to 0.62 with a standard deviation of 0.04, while *multy_family* shows a mean of 23.68, a standard deviation of 13.23 and ranges from 1.98 to 87.36.

Thirdly, railway infrastructure and frequency related variables are included. The variable *station_count* indicates the number of train stations within each municipality, with an average of 1.16 stations. However, this number ranges from 0 to 11, demonstrating that some municipalities do not have any train stations, while others have multiple. Based

Variable	mean	std	min	25%	50%	75%	max
m2_price	3196.45	878.96	1467.73	2585.31	3089.78	3685.78	7378.71
pop_density	910.96	1077.66	23.00	248.25	478.00	1184.00	6827.00
avg_income	36.92	5.26	27.40	34.50	36.20	38.50	81.40
homes_per_capita	0.45	0.04	0.32	0.42	0.44	0.46	0.62
unemp_rate	3.11	0.46	2.40	2.80	3.00	3.30	5.30
net_labor_participation	73.41	2.71	59.50	71.90	73.80	75.20	81.30
multy_family	23.68	13.23	1.98	14.31	19.98	29.43	87.36
distance_to_urban_center	37.32	25.79	0.26	17.08	30.38	51.32	112.97
station_count	1.16	1.55	0.00	0.00	1.00	2.00	11.00
traffic	153.64	314.50	0.00	0.00	71.00	177.75	3777.00
has_station	0.56	0.50	0.00	0.00	1.00	1.00	1.00

[1] Number of observations: 343 (for all columns).

Table 1: Summary statistics for all numeric variables

on this variable, the dummy indicator *has_station* is created and with a mean of 0.56, it shows that a train station is present in 56% of Dutch municipalities. The variable *traffic* measures the volume of daily train traffic and ranges from 0 to 3777 trains per day, with a mean of 153.64. This wide range suggests significant variation in train traffic volume across municipalities.

Lastly, several demographic, geographic and economic control variables have been included. Population density (*pop_density*), with a mean of 910.96 shows significant variation with values ranging from as low as 23.00 to 6827.00. The average income (*avg_income*) per capita is also diverse, with a mean of 36.92 thousand euros and a range from 27.40 to 81.40 thousand euros. Furthermore, *distance_from_urban_center*, with a mean of 37.32 and a standard deviation of 25.79 has been included. Seven urban centers were considered, specifically: Amsterdam, Rotterdam, Den Haag, Utrecht, Eindhoven, Maastricht, and Groningen. Lastly, unemployment rate (*unemp_rate*) and net labor participation (*net_labor_participation*) were considered to further control for economic conditions in the municipalities.

In conclusion, the dataset proves to provide a comprehensive overview of housing prices, demographic characteristics, and rail infrastructure in The Netherlands for 2023. There are 11 variables measured for each municipality, totaling to 343 observations. The range and diversity of variables allow for a comprehensive analysis across the country, and the size of the dataset provides a valid basis for statistical estimation. That being said, in the following section, the methodology used in the analysis will be dealt with.

Methodology

This section outlines the methodological approach used to analyze the relationship between housing prices and rail infrastructure in the municipalities of The Netherlands. Using a set of controls, the analysis will be conducted in two phases, the first focusing on train station presence in all municipalities, and the second focusing on train traffic volume in municipalities with a train station. The section is laid out as follows. Firstly, a subsection is granted to justify data exclusion and variable transformations. Then, the traditional hedonic pricing model will be explained and will be compared to the model utilized as part of this study. Lastly, the final model specifications and diagnostics will be discussed and a step-by-step overview of the analysis is presented. A subsection will be granted to elaborate on tackling multicollinearity as well.

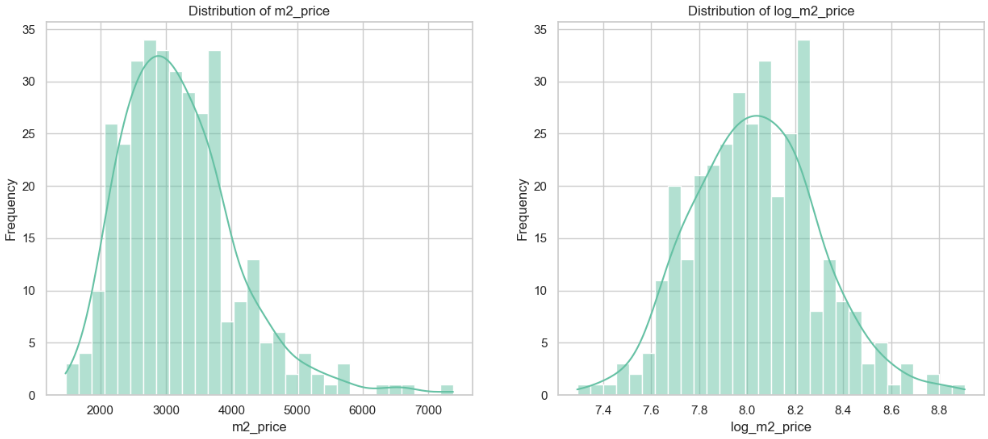


Figure 2: Distribution of Square Meter Prices Before and After Log Transformation

Data Preprocessing

Given the significant variation in the scale of different variables, log transformations were applied to certain variables to improve the fit and adherence to ordinary least squares assumptions. Specifically, the dependent variable *m2_price*, and the predictor variables *avg_income*, *pop_density*, *distance_to_urban_center*, and *traffic* were log-transformed. Figure 2 shows the the distribution of the dependent variable before and after transformation, Figure 5 depicts the distributions of the non-transformed predictor variables, and their scatter plots against *m2_price*, and Figure 6 depicts the scatter plots and distributions of variables after final log transformations. The distribution of daily traffic (*traffic)* is not included in these as sample sizes differ after log transformation [2]. For this, see Figure 3. From the figures, it is evident that log-transforming these variables brings them to a similar scale, reducing the effect of outliers and numerical issues, and improves linearity.

Next, a correlation matrix of all independent variables was estimated to examine potential multicollinearity problems among predictors. High correlation, (values close to 1 or -1) between variables leads to multicollinearity, which can distort the results of regression analysis by inflating the p-values and confidence intervals (Paul, 2006). Based on the correlation matrix (see Figure 4), variables with high correlation values were identified and considered for exclusion. Specifically, *pop_density*, *unemp_rate*, *station_count* and *net_labor_participation* were excluded from the final model due to high correlations with other variables.

[2] Note that *has_station* was not plotted due to it's binary nature.

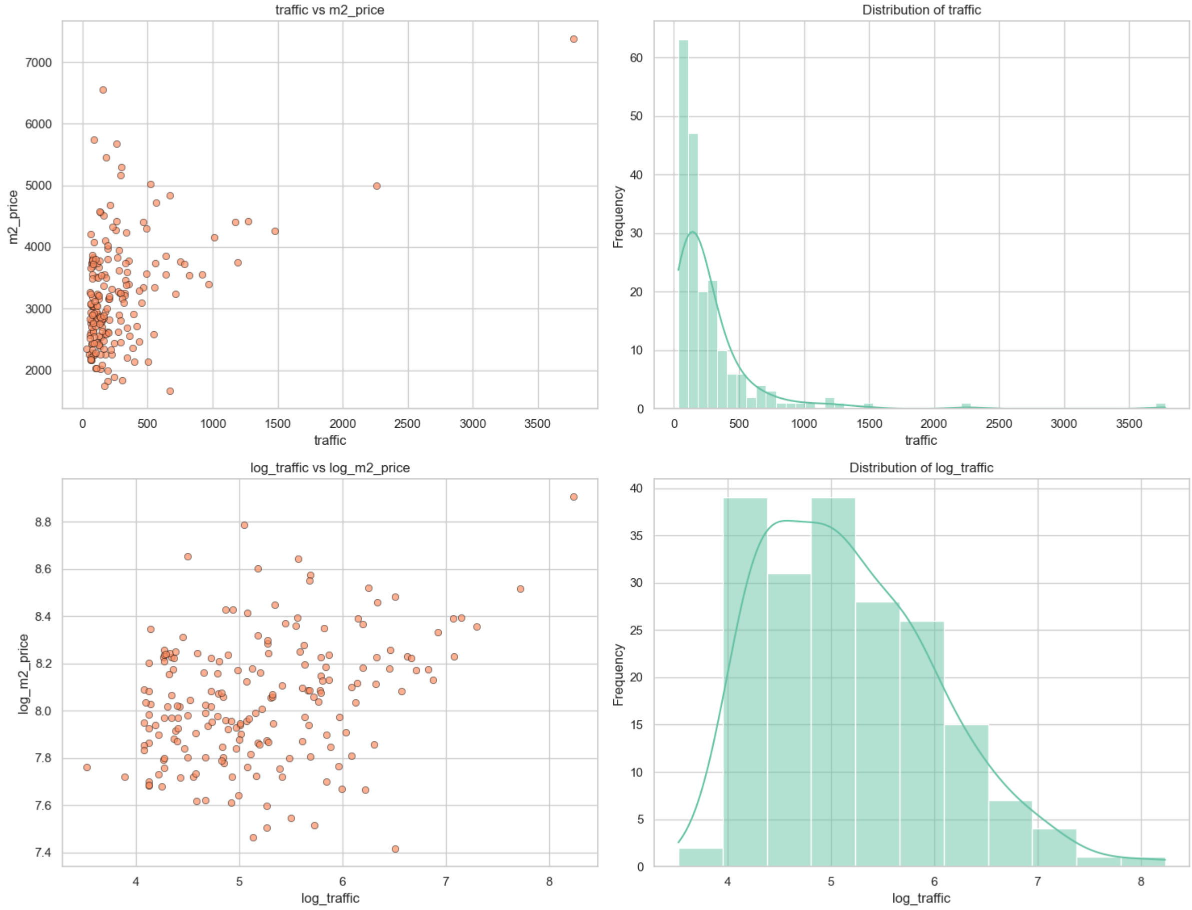
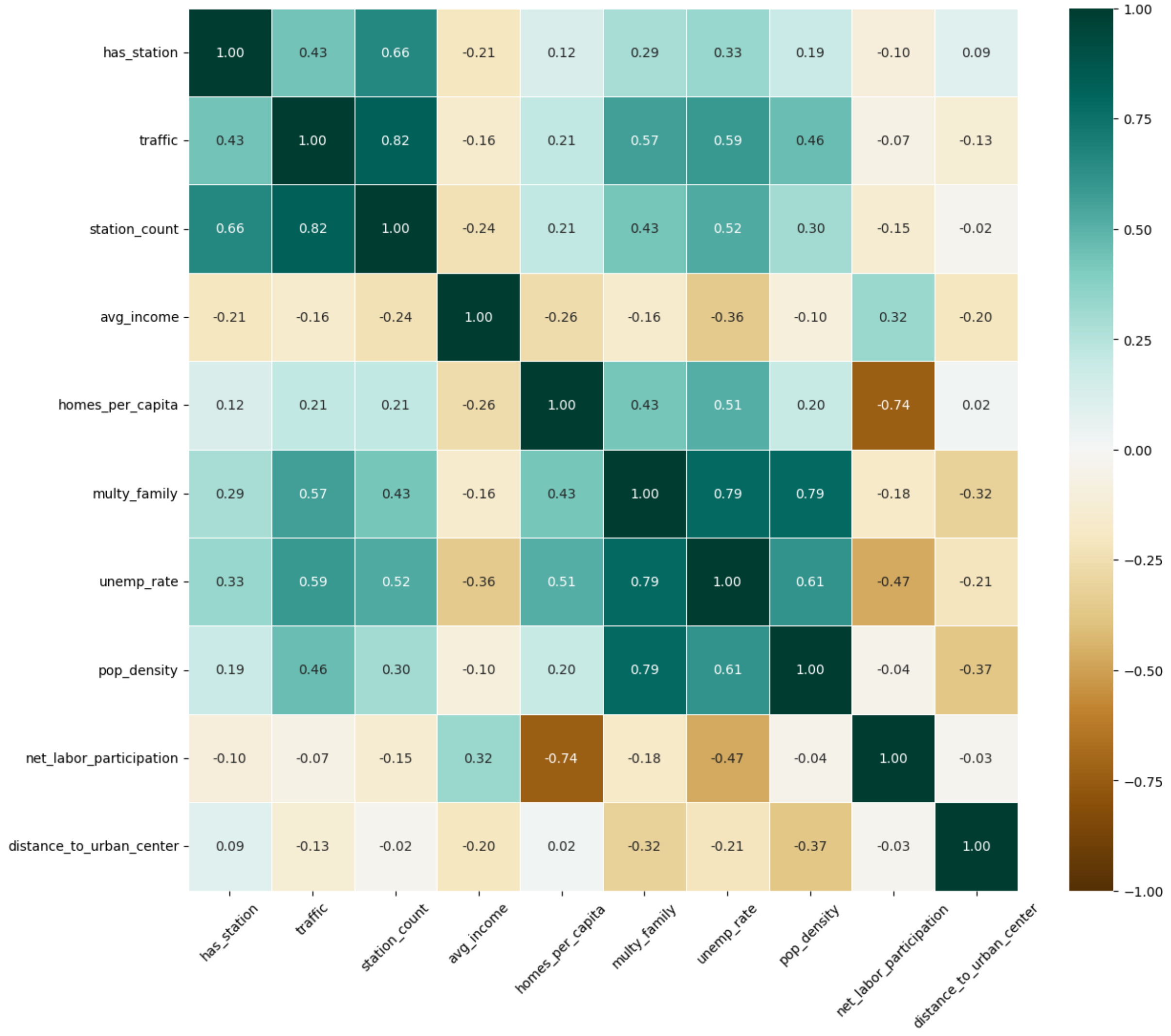


Figure 3: Scatterplots and Distributions of Traffic (N=194)

Deviations from the Hedonic Pricing Model

The hedonic pricing model is a widely used methodological approach in real estate economics that determines the price of a property based on its internal characteristics and the characteristics of its location. According to this model, the price of a property is a function of various attributes, which can be classified into internal characteristics, such as the size, age, and condition of the property and external characteristics, such as the neighborhood quality, accessibility, and environmental factors (Rosen, 1974). In the context of analyzing the impact of rail infrastructure on housing prices, previous applications of the hedonic pricing model have typically focused on micro-level data, examining how proximity to a train station influences the prices of individual properties, utilizing data from local real estate markets. This approach is widely reflected in the papers referenced in this study.

This paper, however, adopts a broader perspective by applying a model that is similar to the hedonic pricing model, however operates on the municipal level. Instead of focusing on individual properties, this model aggregates data to the level of municipalities, seeking to capture the overall real estate market dynamics within each municipality. This approach involves including variables that reflect both the internal structure of the housing market, such as the ratio of multi-family homes and homes per capita, and external factors affecting housing prices, such as average income level and distance to urban centers. The reason for this is that while conducting analysis on property level data across the entire country has been done before (Debrezion et al., 2005), recreating a similar dataset would be out of scope for this research due to time constraints. Consequently, the methodology



	has_station	traffic	station_count	avg_income	homes_per_capita	multy_family	unemp_rate	pop_density	net_labor_participation	distance_to_urban_center
has_station	1.00	0.43	0.66	-0.21	0.12	0.29	0.33	0.19	-0.10	0.09
traffic	0.43	1.00	0.82	-0.16	0.21	0.57	0.59	0.46	-0.07	-0.13
station_count	0.66	0.82	1.00	-0.24	0.21	0.43	0.52	0.30	-0.15	-0.02
avg_income	-0.21	-0.16	-0.24	1.00	-0.26	-0.16	-0.36	-0.10	0.32	-0.20
homes_per_capita	0.12	0.21	0.21	-0.26	1.00	0.43	0.51	0.20	-0.74	0.02
multy_family	0.29	0.57	0.43	-0.16	0.43	1.00	0.79	0.79	-0.18	-0.32
unemp_rate	0.33	0.59	0.52	-0.36	0.51	0.79	1.00	0.61	-0.47	-0.21
pop_density	0.19	0.46	0.30	-0.10	0.20	0.79	0.61	1.00	-0.04	-0.37
net_labor_participation	-0.10	-0.07	-0.15	0.32	-0.74	-0.18	-0.47	-0.04	1.00	-0.03
distance_to_urban_center	0.09	-0.13	-0.02	-0.20	0.02	-0.32	-0.21	-0.37	-0.03	1.00

Figure 4: Correlation Heatmap of Predictors

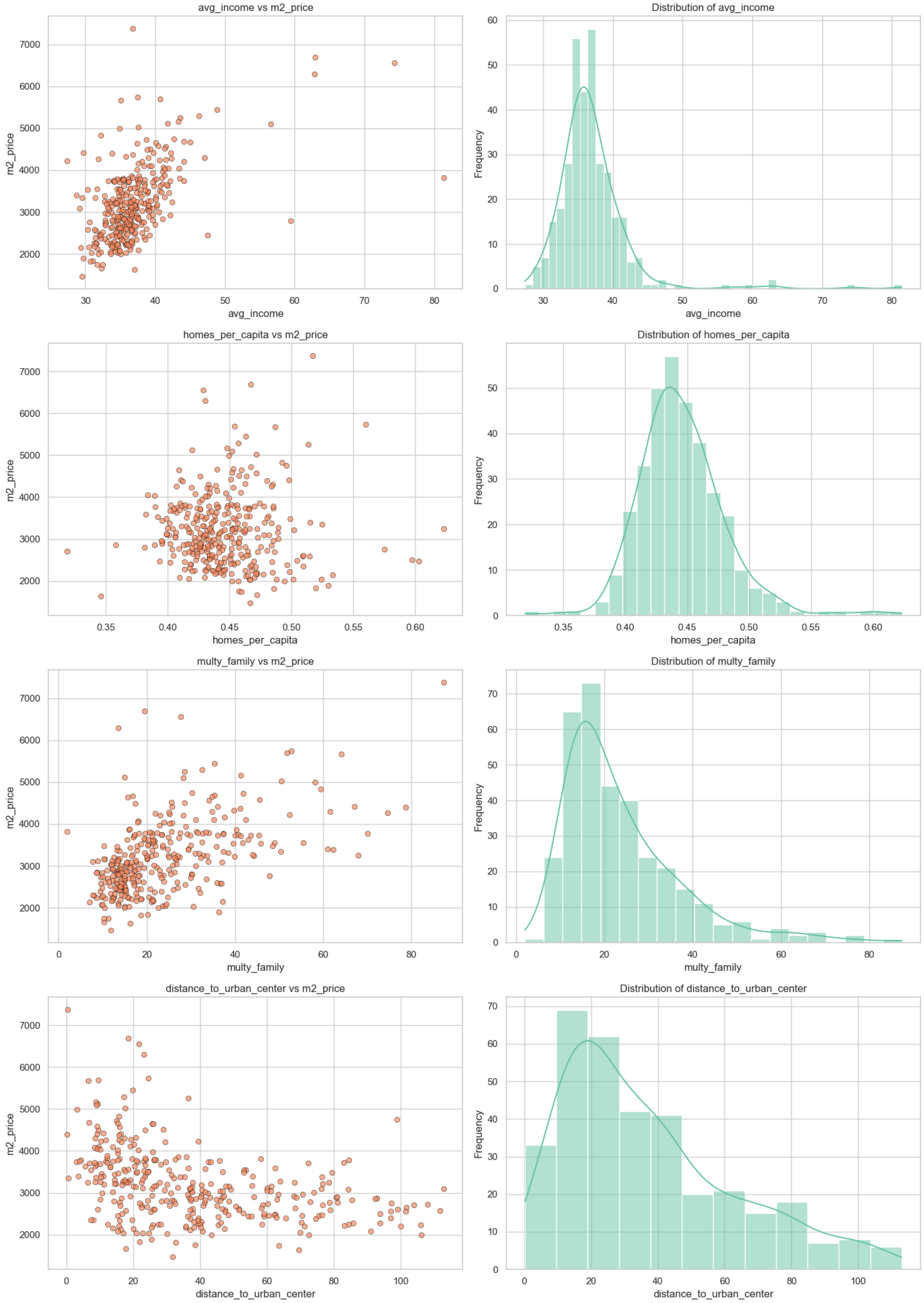


Figure 5: Scatter Plots and Distributions of Predictors (N=343)

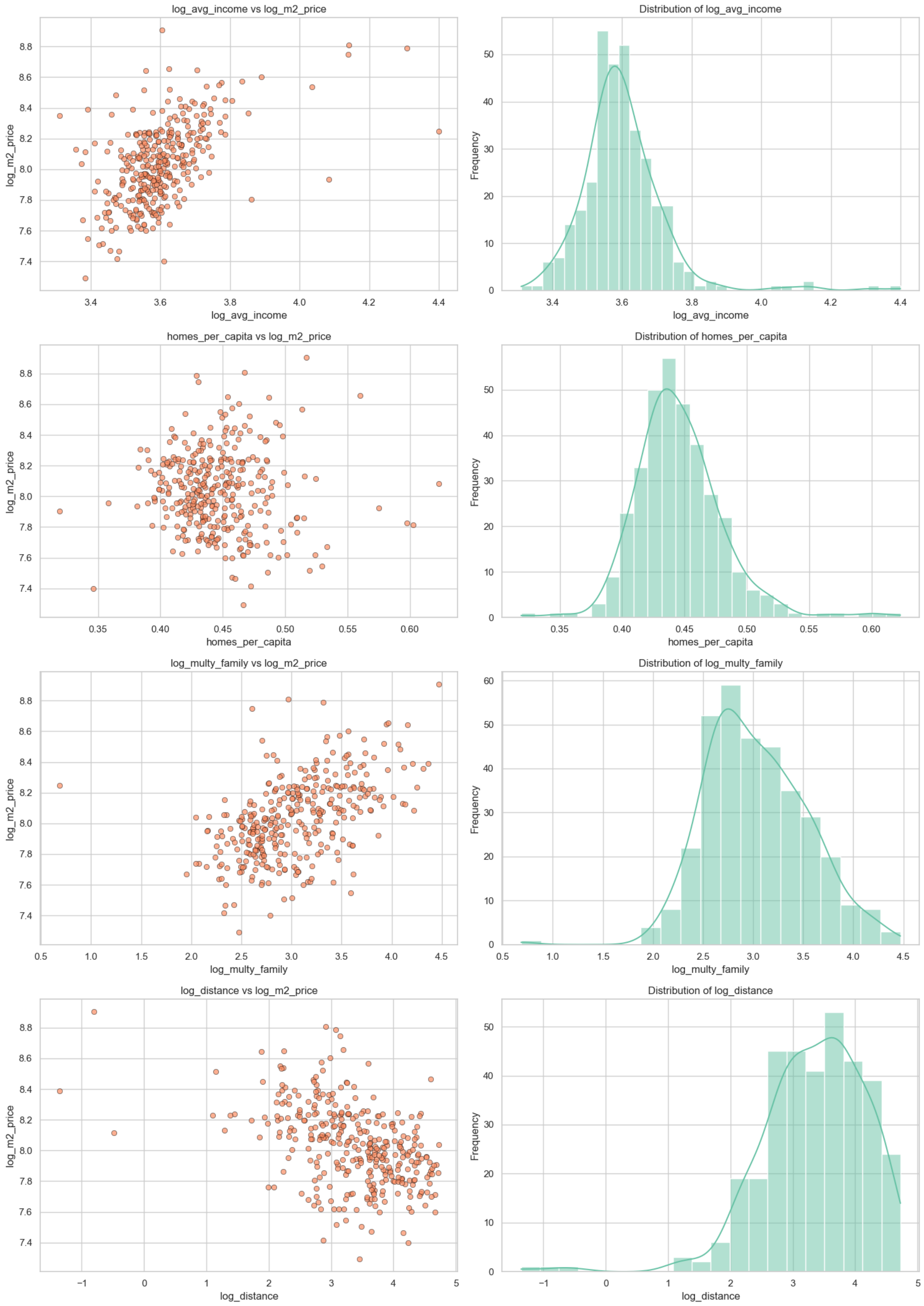


Figure 6: Scatter Plots and Distributions of Predictors (Log Transformed, N=343)

used in this paper aims to provide insights at the macroeconomic level, and take a broader, birds-eye view analyzing effects on the municipal level. This approach is also helpful in mitigating station specific variability that may be present in studies that focus in around a particular group of stations.

Model Specification

That being said, the specific regression models utilized will now be discussed. The analysis was conducted in two phases, each phase focusing on a different aspect of rail infrastructure's impact on housing prices, using an Ordinary Least Squares (OLS) based approach. The first phase includes a equation with *has_station* as a key independent variable, indicating the presence of at least one railway station in the municipality. The second phase focuses on municipalities with a train station present and includes an equation with *traffic*, representing the average daily traffic at railway stations. In the first phase (N = 343), the following two linear regression equations will be utilized:

$$\begin{aligned}\log(m2\_price) = {} & \beta_0 + \beta_1 has\_station + \beta_2 \log(avg\_income) \\ & + \beta_3 \log(multy\_family) + \beta_4 \log(distance\_from\_urban\_center) \\ & + \beta_5 homes\_per\_capita + \epsilon \end{aligned} \tag{1}$$

$$\begin{aligned}\log(m2\_price) = {} & \beta_0 + \beta_1 has\_station + \beta_2 \log(avg\_income) \\ & + \beta_3 \log(multy\_family) + \beta_4 \log(distance\_from\_urban\_center) \\ & + \beta_5 homes\_per\_capita + \beta_6 (has\_station \times \log(multy\_family)) \\ & + \beta_7 (has\_station \times \log(distance\_from\_urban\_center)) + \epsilon \end{aligned} \tag{2}$$

As for the second phase, the dataset will be restricted to municipalities with a train station (N=194) and the predictor *has_station* will be replaced by *traffic*. The corresponding regression equations are as follows:

$$\begin{aligned}\log(m2\_price) = {} & \beta_0 + \beta_1 \log(traffic) + \beta_2 \log(avg\_income) \\ & + \beta_3 \log(multy\_family) + \beta_4 \log(distance\_from\_urban\_center) \\ & + \beta_5 homes\_per\_capita + \epsilon \end{aligned} \tag{3}$$

$$\begin{aligned}\log(m2\_price) = {} & \beta_0 + \beta_1 \log(traffic) + \beta_2 \log(avg\_income) \\ & + \beta_3 \log(multy\_family) + \beta_4 \log(distance\_from\_urban\_center) \\ & + \beta_5 homes\_per\_capita + \beta_6(\log(traffic) \times \log(multy\_family)) \\ & + \beta_7(\log(traffic) \times \log(distance\_from\_urban\_center)) + \epsilon \end{aligned} \tag{4}$$

Measuring and Mitigating Multicollinearity

This study employs several metrics to measure the level of multicollinearity present in the regression models. In addition to the correlation matrix discussed above, VIF scores were calculated for each model to quantify the severity of multicollinearity. VIF is a measure that indicates how much the variance of a regression coefficient is inflated due to dependence other predictors in the model (Paul, 2006). In general, VIF scores greater than 5 to 10 suggest significant multicollinearity.

Furthermore, condition indices were computed as part of the multicollinearity diagnostics. The condition index is a metric derived from the eigenvalues of the standardized independent variable matrix (Kim, 2019). High condition index values (a rule of thumb is generally 10 to 30) indicate potential multicollinearity problems.

To minimize the level of multicollinearity, besides excluding predictors showing high correlation with each other, the models were also computed using standardized regression coefficients. Standardizing helps in mitigating multicollinearity and presenting a more trustworthy model by bringing independent variables to the same scale (Azubuike & Tobe, 2019). When variables are standardized, the coefficients represent the change in the dependent variable for a one standard deviation change in the independent variable, rather than a one-unit change. This step reduces the economic interpretability of the models, however as coefficients are scaled to a mean 0 and standard deviation of 1, these regressions provide insights on the relative importance of independent variables. Standardized models will not be given much emphasis in order to stay within the scope of this thesis, they are merely presented as a means to tackle multicollinearity.

OLS Assumptions and Model Diagnostics

For the OLS estimates to be valid and reliable, certain assumptions must be met. This subsection outlines these assumptions according to Fox (2015) and the corresponding diagnostic tests and visual checks conducted to verify adherence in the estimated models.

Linearity

The assumption of linearity states that the relationship between the independent variables and the dependent variable is linear. To ensure linearity, some predictors were log-transformed and scatter plots of the observed values versus against the dependent variable were examined (see Figure 5).

Independence

The independence assumption requires that the residuals are independent of each other. To account for this, residuals were plotted against corresponding fitted values in determine any possible pattern. Furthermore, the Durbin-Watson test statistic was used in order to detect the presence of autocorrelation in residuals. A Durbin-Watson value close to 2 indicates no autocorrelation, while values significantly less than 2 suggest positive autocorrelation, and values significantly greater than 2 suggest negative autocorrelation (Montgomery et al., 2021).

Homoskedasticity

Homoskedasticity implies that the variance of the residuals is constant across all levels of the independent variables. This effects the efficiency of the least squares estimator (Fox, 2015). To test for homoskedasticity, residual plots were analyzed for any visible funnel-shaped patterns, which could indicate different levels of variance for different values of the dependent variable.

Normality of Residuals

The normality assumption requires that the residuals of the model are normally distributed, with mean zero. Although in large samples the impact of non-normal residuals can be negligible, the least-squares estimator is only the most efficient unbiased estimator when this assumption is met (Fox, 2015). This assumption was assessed using two methods. Firstly, histograms of the residuals were plotted to visually inspect their distribution. Secondly, the Omnibus and Jarque-Bera tests were used to statistically test for normality. Non-significant results from these tests indicate that the residuals are normally distributed.

Comparison and Improvement of Model Fit

This study utilizes two metrics to evaluate model fit, namely the Bayesian Information Criterion (BIC) and the coefficient of determination (R^2). The BIC balances model fit with model complexity by including a penalty for the number of parameters in the model.

A lower BIC value indicates a better model, suggesting a good fit with fewer predictors (Neath & Cavanaugh, 2012). The R^2 value indicates the proportion of variance in the dependent variable explained by the model. Higher (R^2) values indicate a better fit, meaning that the model accounts for a larger portion of the variability in housing prices (Montgomery et al., 2021).

To improve model fit, exclusion of outliers in the data was considered. Outliers can disproportionately influence the results and residuals of a regression model (Fox, 2015). To identify and mitigate their impact, firstly the Cook's distance was calculated for each observation, which indicates the residuals that are the most influential in the regression results (Fox, 2015). Then, the models were re-estimated after removing the 10 observations with the highest Cook's distance values and changes in the results were assessed.

In conclusion, the methodological approach used to analyse the research question at hand has been adequately assessed. Justifications for variable exclusion have been presented due to multicollinearity problems. Furthermore, the section introduced the relevant regression equations, and the steps taken to reduce multicollinearity and insure adherence to Ordinary Least Squares assumptions. In the following section the concrete outline of the analysis will be presented and the results will be discussed and compared across models.

Analysis

This section of the paper will sequentially deal with both phases of the analysis. In general, the phases will adhere to the following concrete steps:

1. Estimating the relevant regression model.

2. Introduce interaction terms between the key independent variable and *multy_family* and *distance_from_urban_center* in order to check for whether effects vary in more rural and more densely populated municipalities.

3. Standardize the coefficients in order to further mitigate multicollinearity.

4. Compare the estimated models based on BIC, VIF and condition index values.

5. Conduct residual analysis for the non-standardized model with interaction terms. The residual distribution, the scatter plot of residuals against fitted values, and a histogram of Cook's distances are plotted in order to validate OLS assumptions.

6. Remove the largest 10 Cook's distance values, then re-estimate the models without these outliers (N=333 (first phase) and N=184 (second phase)).

7. Compare changes in estimations, model fit and multicollinearity after dropping the influential outliers.

8. Compare Omnibus, Jarque-Bera, Durbin-Watson and R^2 scores for all models and assess further adherence to OLS assumptions.

Effect of Station Presence on House Prices

To commence with the first phase of the analysis, a boxplot of house prices in municipalities with a train station is paired with one in municipalities without a train station (see Figure 7). Next, Table 2 shows statistical tests conducted in order to compare the two distributions.

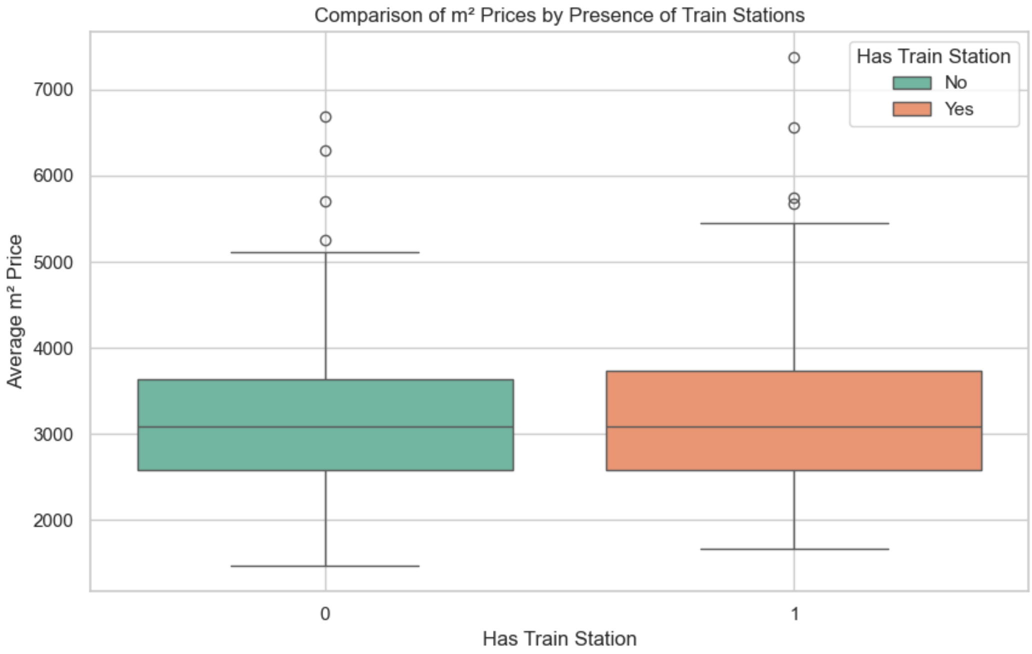


Figure 7: Boxplots of Prices by Station Presence

Test Name	Test Statistic	P-value
Shapiro-Wilk Test (With Station)	0.9297	0.0000
Shapiro-Wilk Test (Without Station)	0.9402	0.0000
Levene Test for Equal Variances	0.1279	0.7208
Mann-Whitney U test	14816.0000	0.6298

Table 2: Test Results of Comparing Distributions of Prices by Station Presence

Firstly, a Shapiro-Wilk test is conducted for both distributions to check for normality. For both municipalities with and without a train station, the p-value is below 0.001 suggesting that house prices are not normally distributed. Next, a Levene Test is conducted (p=0.1279), showing that the variances are approximately equal for both groups. Since the normal distribution assumption of the two-sample t-test is violated, a Mann-Whitney U test is used to compare the distributions (p = 0.6298). It is evident from the boxplots shown in Figure 7 and the test outcome that the two distributions are not significantly different.

Next, the linear regressions models specified in Equation 1 (Simple) and 2 (Interaction) are presented along with a standardized version of the interaction model. The results of the regression models are presented in Table 3 (N = 343). The simple regression model shows that the presence of a train station (has_station) has a negligible and statistically insignificant effect on house prices (β = -0.0087, p > 0.05). Average income (*log_avg_income*) and the proportion of multi-family homes (*log_multy_family*) have significant positive effects on house prices (β = 1.2649, p < 0.001 and β = 0.3359, p < 0.001, respectively). The number of homes per capita (*homes_per_capita*) and the distance to the nearest major city (*log_distance*) show significant negative effects on house prices (β = -1.0032, p < 0.001 and β = -0.0274, p < 0.05, respectively).

The interaction terms (*station_x_multy_fam* and *station_x_distance*) are not statistically significant, indicating no substantial interaction effects. The standardized interaction model confirms these findings, showing consistent effects with the interaction model. Overall, it is evident that income and the number of homes per capita are the strongest predictors of house prices, while the presence of a station does not significantly influence them.

	Simple	**Interaction**	**Standardized**
const	3.0051 (0.3328)***	3.1678 (0.3547)***	7.9402 (0.0711)***
has_station	-0.0087 (0.0175)	-0.2180 (0.1620)	-0.2180 (0.1620)
log_avg_income	1.2649 (0.0740)***	1.2542 (0.0745)***	0.1509 (0.0090)***
log_multy_family	0.3359 (0.0205)***	0.3071 (0.0296)***	0.1551 (0.0149)***
homes_per_capita	-1.0032 (0.2578)***	-1.0205 (0.2636)***	-0.0361 (0.0093)***
log_distance	-0.0274 (0.0111)*	-0.0376 (0.0197)	-0.0317 (0.0166)
station_x_multy_fam	-	0.0509 (0.0378)	0.0823 (0.0611)
station_x_distance	-	0.0169 (0.0234)	0.0306 (0.0423)

[1] Standard errors in parentheses.

[2] * p < 0.05, ** p < 0.01, *** p < 0.001

[3] Note that the second variable is log transformed in the interaction terms.

Table 3: Regression Model Results (Eq. 1 and 2)

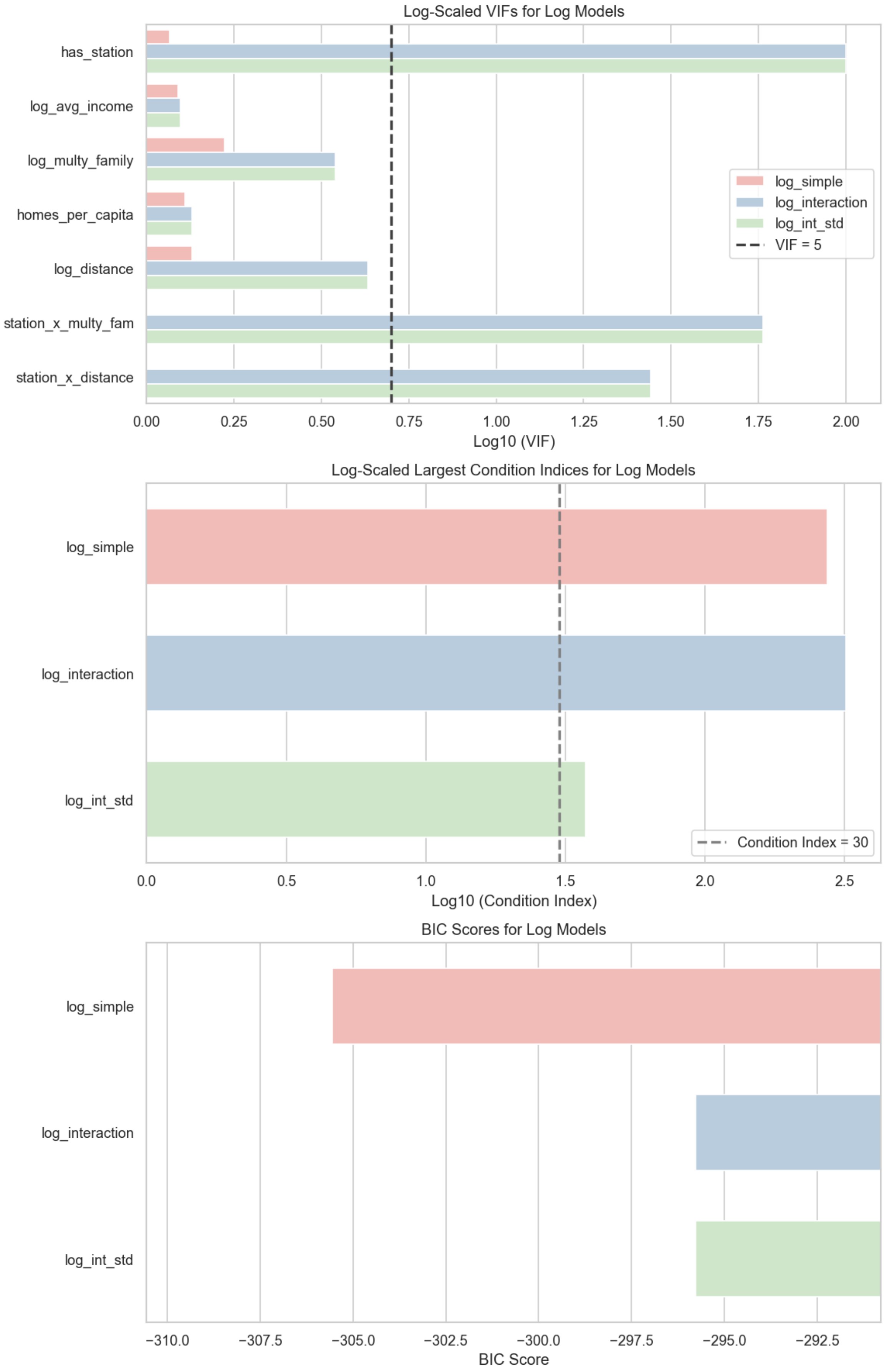


Figure 8: Multicollinearity Indicators and BIC Scores (Table 3)

Figure 8 depicts multicollinearity metrics and BIC scores relevant to the models presented in Table 3. The top plot, which presents log-scaled VIFs, indicates that while VIFs scores are well below the treshold for the simple model, adding interaction terms introduces multicollinearity to the regressions. The middle plot shows the largest condition index for the models. It is evident that while the simple model does not suffer from multicollinearity according to the VIF value, the largest condition index is still above 30,

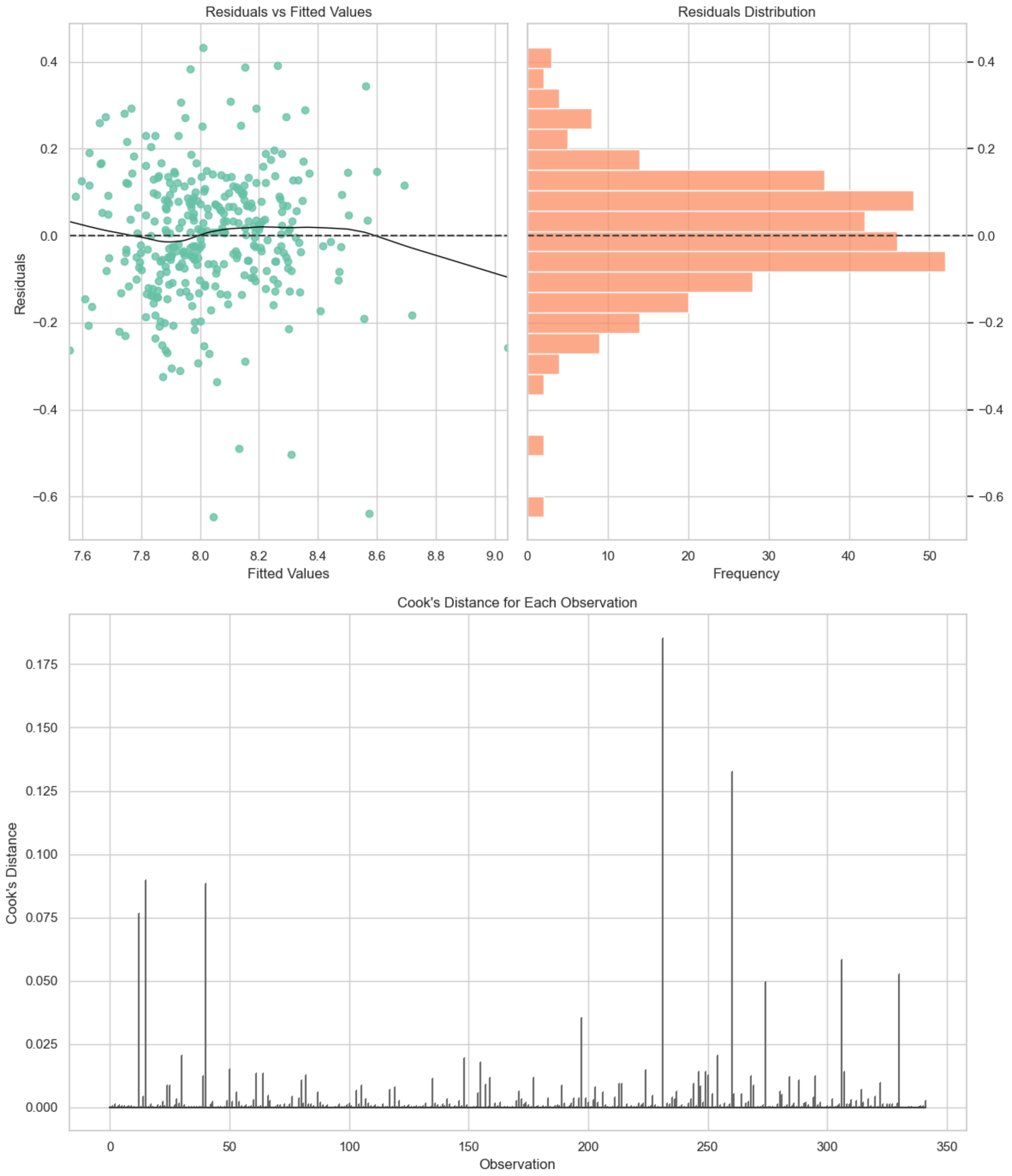


Figure 9: Residual Analysis of Interaction Model (Table 3)

which could indicate potential numerical issues. Standardizing independent variables decreases the largest condition index, however it is still above the level of 30. The chart on the bottom shows BIC values obtained for all models. The lowest BIC value (a little under -305) belongs to the model without interaction terms. This indicates that introducing interaction terms did not improve model, instead, it led to potential overfitting and multicollinearity issues due to the additional predictors. Furthermore, It is evident from the figures that standardizing predictors did not improve VIF or BIC scores for the models, although it did reduce the value of the largest condition index.

Next, on Figure 9 the residual analysis of the interaction model is presented. The top left plot displays the residuals versus fitted values, showing no apparent pattern, which indicates that the assumption of independence and homoskedasticity is met. On the top right plot the distribution of residuals is shown. While the residuals do seem to be distributied around mean zero, it is evident from the figure that some outliers skew the distribution right leading to a potentially non-normal outlier distribution. Furthermore, Cook's indices for all observations are presented on the bottom plot, showing that several outliers substantially spike out.

After dropping the outliers with the 10 largest Cook's distances [2], the regression models were re-estimated, and the results are summarized in Table 4 (N = 333). The effect of the presence of a train station remains statistically insignificant, similar to the initial analysis. Average income (*log_avg_income*) and the proportion of multi-family homes (*log_multy_family*) continue to show significant positive effects on house prices. The number of homes per capita (*homes_per_capita*) also retains its significant negative effect. The interaction terms still remain statistically insignificant.

	Simple	**Interaction**	**Standardized**
const	2.5442 (0.3231)***	2.4455 (0.3335)***	8.0174 (0.0612)***
has_station	-0.0158 (0.0153)	-0.0311 (0.1422)	-0.0311 (0.1422)
log_avg_income	1.4246 (0.0719)***	1.4417 (0.0710)***	0.1590 (0.0078)***
log_multy_family	0.3570 (0.0182)***	0.3474 (0.0264)***	0.1732 (0.0132)***
homes_per_capita	-1.5400 (0.2441)***	-1.5416 (0.2446)***	-0.0512 (0.0081)***
log_distance	-0.0069 (0.0099)	0.0124 (0.0178)	0.0101 (0.0145)
station_x_multy_fam	-	0.0276 (0.0328)	0.0443 (0.0527)
station_x_distance	-	-0.0204 (0.0208)	-0.0368 (0.0376)

[1] Standard errors in parentheses.
[2] * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
[3] Note that the second variable is log transformed in the interaction terms.

Table 4: Regression Model Results (Dropped Outliers, Eq. 1 and 2)

[2] The names of the excluded municipalities are listed in Appendix 2.

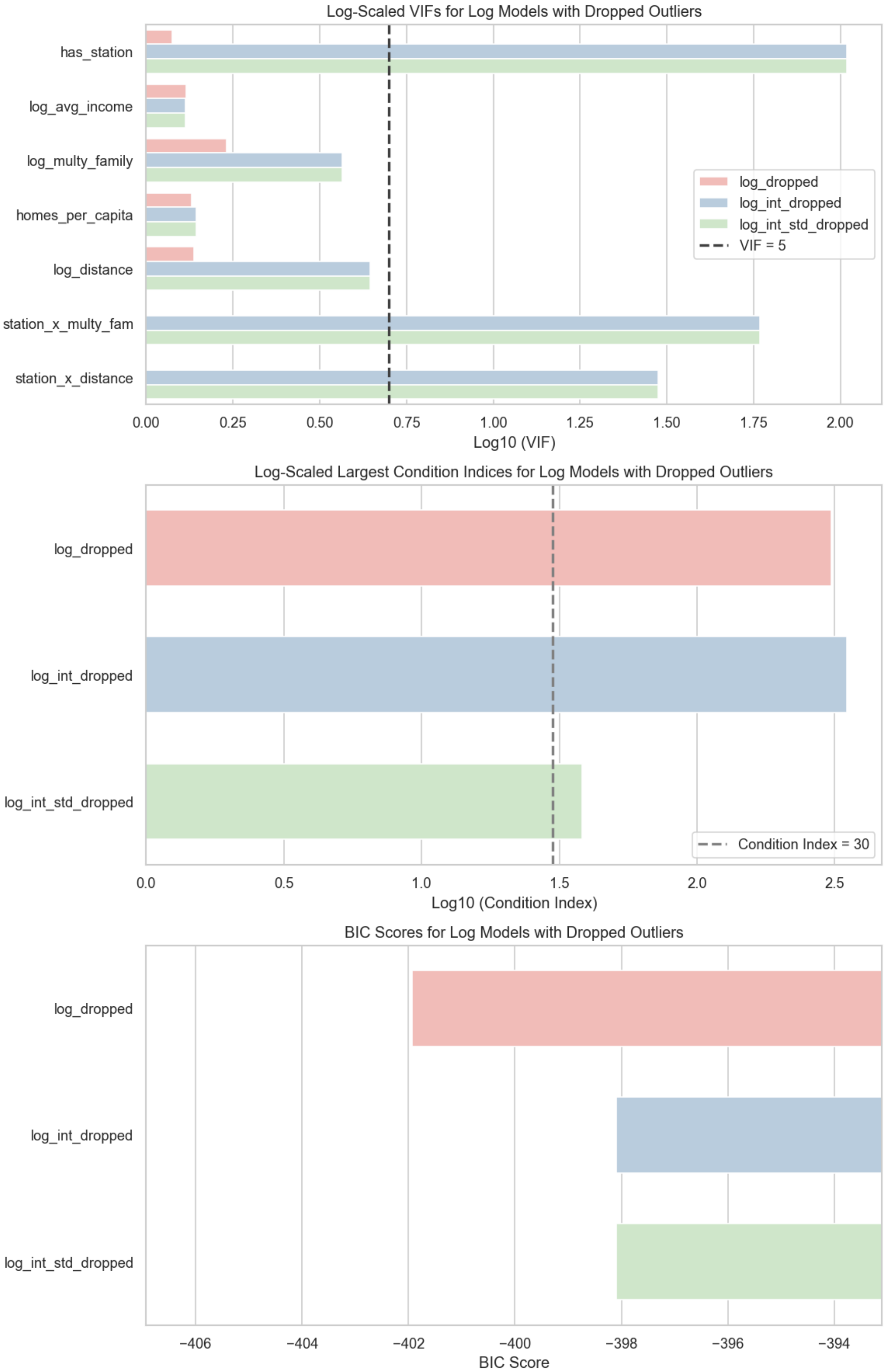


Figure 10: Multicollinearity Indicators and BIC Scores (Dropped Outliers, Table 4)

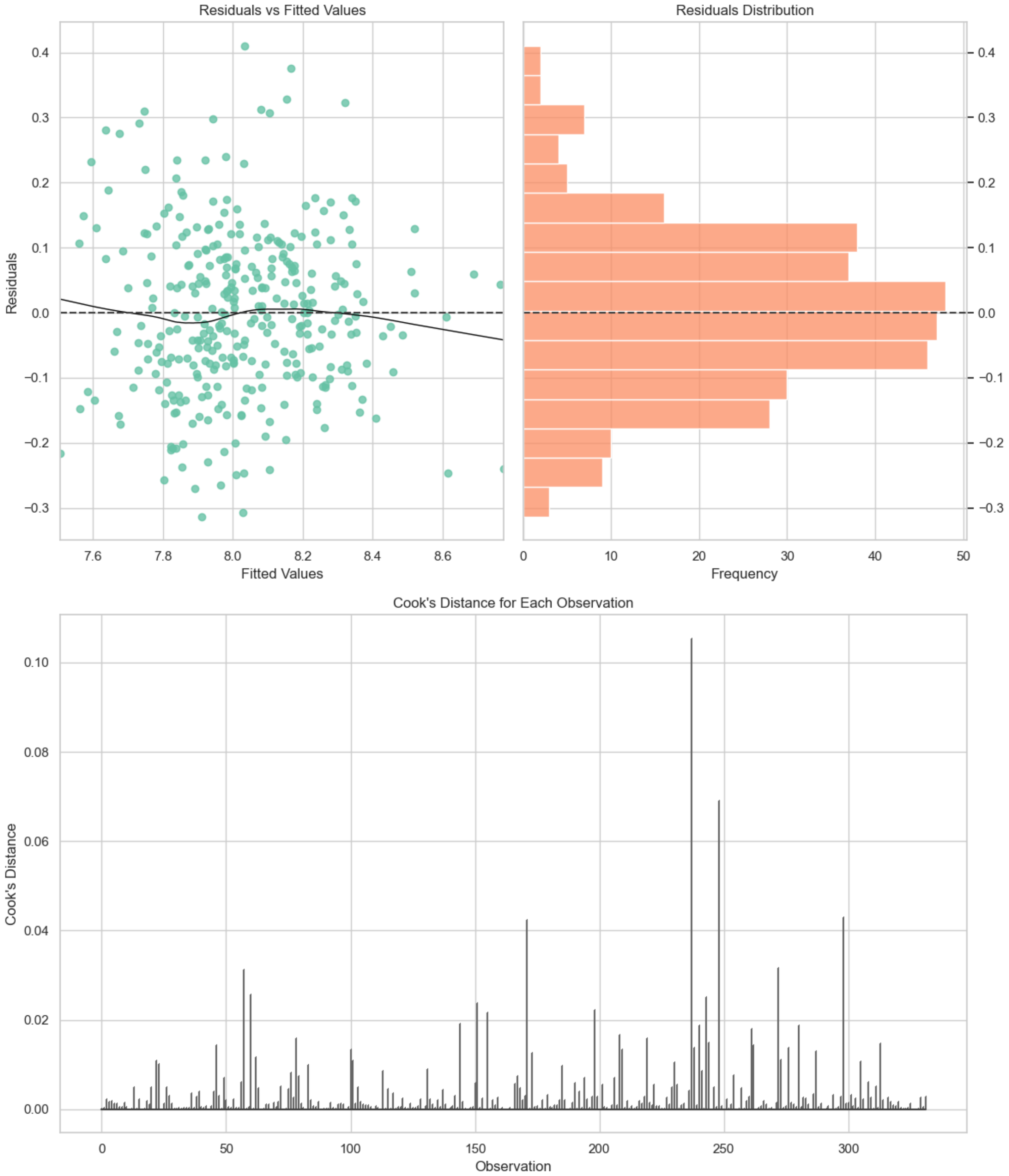


Figure 11: Residual Analysis of Regression Models (Dropped Outliers, Table 4)

Multicollinearity metrics are similar to before, however, BIC scores (Figure 10) indicate some improvements after dropping outliers. The BIC values slightly decrease (by about 100), indicating a better model fit after removing the influential observations. Furthermore, it is visible on Figure 11 that dropping influential values improved the skewness of the residual distribution - exhibiting better fit to normality - and the bottom plot shows a more balanced distribution of Cook's distances as well.

Next, Table 5 shows normality and autocorrelation tests, along with R^2 and adjusted R^2 scores. The Omnibus and Jarque-Bera tests indicate significant deviations from normality in the residuals for the simple and interaction models ($p < 0.001$ for both). However, after removing outliers, the test results improve substantially, with p-values above 0.1, suggesting a better fit to normality. Furthermore, the Durbin-Watson statistic shows values close to 2 for all models, indicating no significant autocorrelation issues in the residuals. Lastly, the R^2 and adjusted R^2 values, increase from around 0.68 to 0.75 after dropping outliers, indicating an improved model fit. Overall, while removing outliers proves to slightly improve model fit, the conclusions remain the same: the presence of a train station in a municipality does not significantly influence house prices, whereas income, total homes per capita and ratio of multy-family housing, are strong predictors.

Metric	**Simple**	**Interaction**	**Dropped**	**Dropped (Int.)**
Omnibus	33.707	31.068	4.611	1.816
Omnibus p-value	0.000	0.000	0.100	0.403
Jarque-Bera	80.077	73.307	4.323	1.657
Jarque-Bera p-value	0.000	0.000	0.115	0.437
Durbin Watson	2.211	2.223	2.255	2.245
$\mathbf{R^2}$	0.681	0.683	0.745	0.751
Adjusted $\mathbf{R^2}$	0.676	0.676	0.741	0.745

[1] Simple and Interaction refer to Equation 1 and 2 respectively. Dropped and Dropped (Int.) are their counterparts with outliers removed.

Table 5: Summary of Test Scores and R^2 for Regression Models (Table 3 and 4)

Effect of Traffic in Municipalities with a Train Station

In the second phase of the analysis, the focus shifts to examining the effect of traffic levels on house prices within municipalities that have a train station. The results of the regression models are summarized in Table 6 ($N = 194$). The simple regression model shows that traffic (*log_traffic*) has a small but statistically significant positive effect on house prices ($\beta = 0.0259$, $p < 0.05$), so according to the model a 1% increase in train traffic in a municipality leads to a 0.0259% increase in square meter house prices. Consistent with the findings in the previous models, average income (*log_avg_income*) and the proportion of multi-family homes (*log_multy_family*) exhibit strong positive effects on house prices ($\beta = 1.2601$, $p < 0.001$ and $\beta = 0.3771$, $p < 0.001$, respectively). Furthermore, the number of homes per capita (*homes_per_capita*) shows a significant negative effect on house prices ($\beta = -2.2141$, $p < 0.001$), this is also consistent with the findings before.

	Simple	**Interaction**	**Standardized**
const	3.2606 (0.4804)***	3.9129 (0.6791)***	8.0417 (0.0087)***
log_traffic	0.0259 (0.0140)*	-0.1698 (0.1039)	-0.1417 (0.0867)
log_avg_income	1.2601 (0.1022)***	1.3197 (0.0991)***	0.1382 (0.0104)***
log_multy_family	0.3771 (0.0280)***	0.0369 (0.1320)	0.0183 (0.0655)
homes_per_capita	-2.2141 (0.3919)***	-2.1312 (0.3774)***	-0.0663 (0.0117)***
log_distance	-0.0177 (0.0115)	0.0564 (0.0556)	0.0532 (0.0524)
traffic_x_multy_fam	-	0.0671 (0.0249)***	0.3274 (0.1214)***
traffic_x_distance	-	-0.0103 (0.0097)	-0.0546 (0.0512)

[1] Standard errors in parentheses.

[2] * p < 0.05, ** p < 0.01, *** p < 0.001

[3] Note that both variables are log transformed in the interaction terms.

Table 6: Regression Model Results (Eq. 3 and 4)

Interestingly, the interaction model shows that the interaction term between traffic and multi-family homes is significant (β = 0.0671, p < 0.01), indicating that the effect of traffic on house prices is stronger in areas with a higher proportion of multi-family homes. The interaction term between traffic and distance from the nearest major city is not significant, suggesting no substantial variation in the effect of traffic across different distances from urban centers. It is also worth to note that although it is not statistically significant, the effect of traffic is negative (β = -0.1698, p > 0.1).

The VIF scores and condition index values are similar to the models presented before (see Figure 12). The plots show that although the VIF scores of the simple model are well below the treshold of 5, the largest condition index persists to be an issue. Similarly, introducing interaction terms increased VIFs and condition indices, thus presenting further multicollinearity concerns. Standardizing the coefficients improved the value of the largest condition index, but the VIFs remain unchanged. However, contrarily to the previous models, BIC scores of the interaction models are lower than those of the model without interaction terms, indicating that the improved model fit outweighs the overfitting concerns present due to the introduction of new independent variables.

Figure 13 summarizes the results of the residual analysis for the interaction model presented in Table 6. Similarly to before, the scatter plot of residuals versus fitted values exhibit no funnel-shaped or general pattern indicating independence and no autocorrelation. However, the presence of outliers influences the skewness of the residual distribution, and it is visible on the bottom plot presenting Cook's distances, that three specific observations (around index 7, 23 and 185 respectively) exhibit very large influence on the coefficients of the model compared to the rest of the residuals. Again, exclusion of the observations with the largest 10 Cook's indexes is considered.

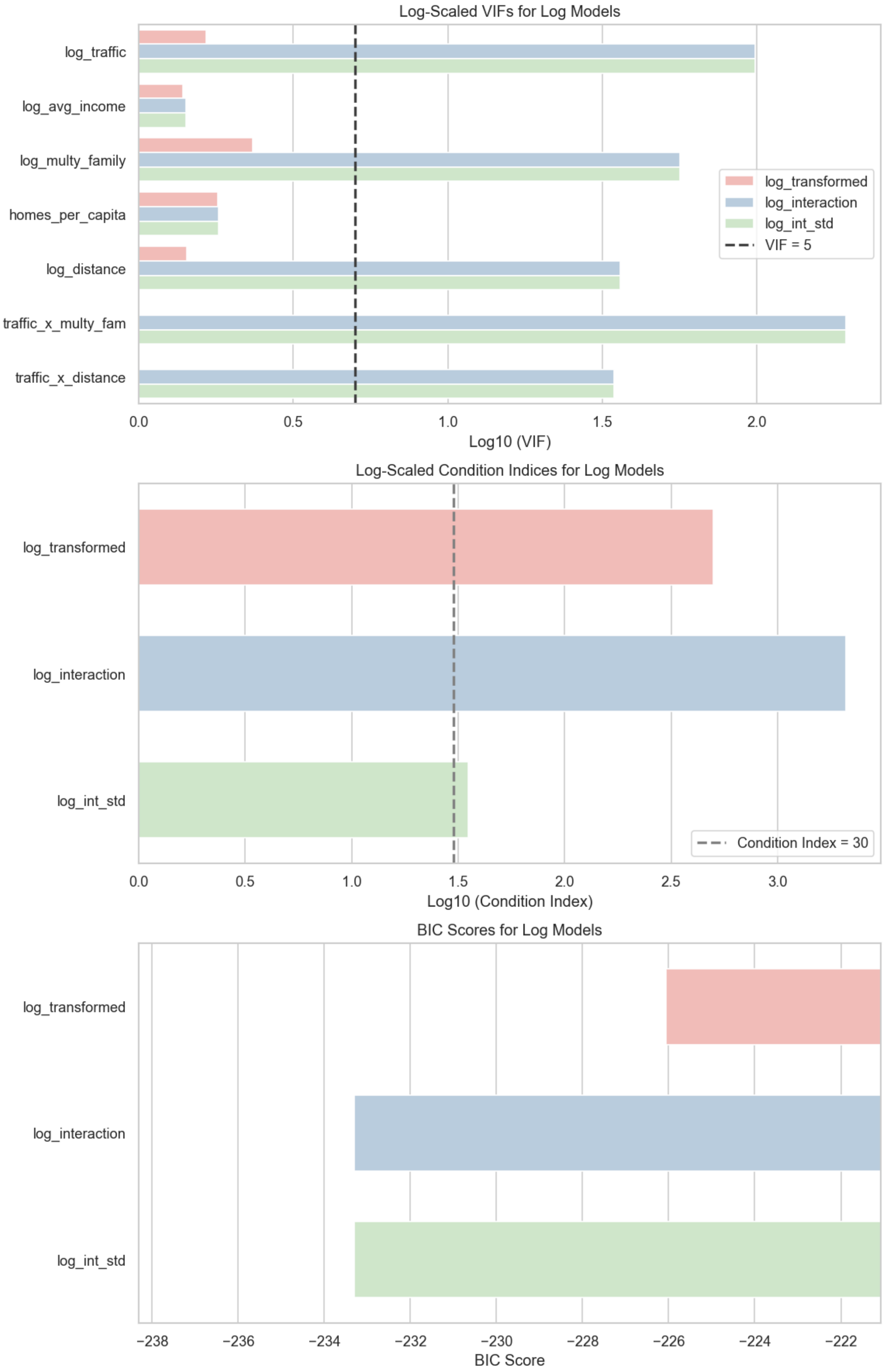


Figure 12: Multicollinearity Indicators and BIC Scores (Table 6)

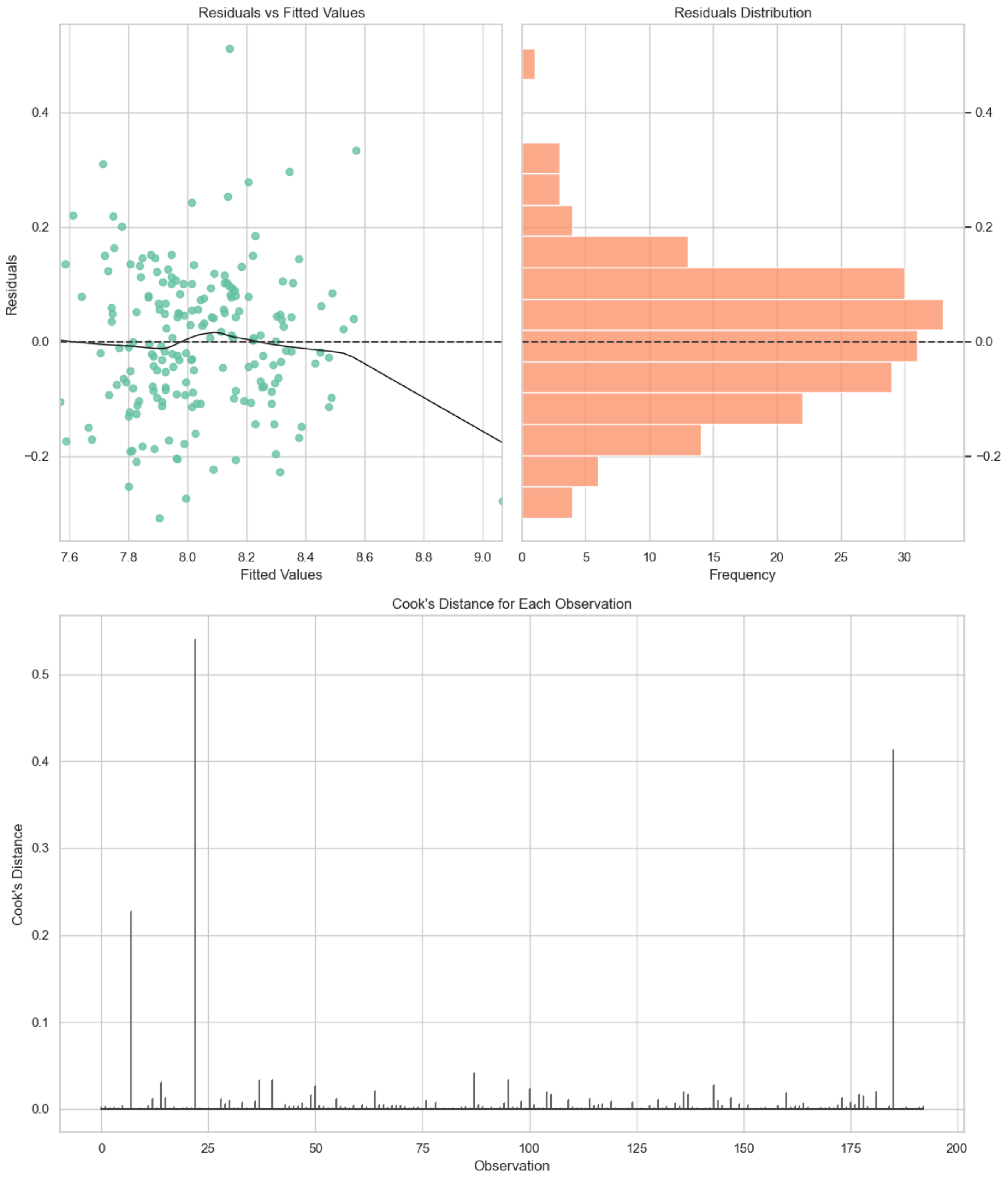


Figure 13: Residual Analysis of Regression Models (Table 6)

The results of the regressions after dropping influential outliers are presented in Table 7 (N = 184). The simple regression model shows that the effect of traffic (*log_traffic*) on house prices remains positive and statistically significant ($\beta = 0.0279$, p < 0.01). This result suggests that an increase in train traffic continues to positively impact house prices, and the effect is slightly more pronounced compared to the initial model.

	Simple	**Interaction**	**Standardized**
const	3.1177 (0.4999)***	4.0180 (0.6849)***	8.0229 (0.0076)***
log_traffic	0.0279 (0.0131)**	-0.2587 (0.0985)***	-0.2027 (0.0772)***
log_avg_income	1.3358 (0.1085)***	1.4793 (0.1064)***	0.1333 (0.0096)***
log_multy_family	0.3552 (0.0258)***	-0.0702 (0.1194)	-0.0327 (0.0557)
homes_per_capita	-2.3956 (0.3672)***	-2.4552 (0.3446)***	-0.0718 (0.0101)***
log_distance	-0.0141 (0.0106)	-0.0126 (0.0652)	-0.0097 (0.0503)
traffic_x_multy_fam	-	0.0831 (0.0225)***	0.3688 (0.1001)***
traffic_x_distance	-	0.0031 (0.0120)	0.0134 (0.0520)

[1] Standard errors in parentheses.

[2] * p < 0.05, ** p < 0.01, *** p < 0.001

[3] Note that both variables are log transformed in the interaction terms.

Table 7: Regression Model Results (Dropped Outliers, Eq. 3 and 4)

The interaction model indicates that the interaction term between traffic and multi-family homes remains statistically significant (β = 0.0831, p < 0.001), reinforcing the conclusion that the effect of traffic on house prices is stronger in areas with a higher proportion of multi-family housing. The interaction term between traffic and distance from the nearest urban center remains insignificant. Contrarily to the previous model, after dropping influential observations, the interaction model shows a negative, highly significant effect of traffic (β = -0.2587, p < 0.001). This means that on average, 1% in daily train traffic in a municipality leads to a -0.2587% decrease in house prices. However, this value also depends on the share of multy-family housing in the municipality.

Furthermore, average income (*log_avg_income*) continues to show a strong positive effect on house prices (β = 1.4793, p < 0.001). The proportion of multi-family homes (*log_multy_family*) and the number of homes per capita (*homes_per_capita*) maintain their significant roles in the regression as well. Overall, it is clear that the exclusion of outliers led to more robust results, reinforcing the previous conclusions. Increased traffic has a negative effect on house prices, and this negative effect is reduced and may be positive in municipalities with a higher share of multy-family housing.

Figure 14 depicts the BIC scores and multicollinearity metrics for the models. In general, conclusions remain the same as before. Dropping the outliers did not improve multicollinearity metrics significantly. Standardizing the coefficients seems to improve condition indices slightly, however VIF values are still a problem for regressions with interaction terms included. Furthermore, the bottom plot shows that BIC scores decrease by about 50, indicating that the exclusion of influential outliers led to better model fit. The plot further indicates that the interaction models fit the data better than the simple model, similarly to the model with the outliers included.

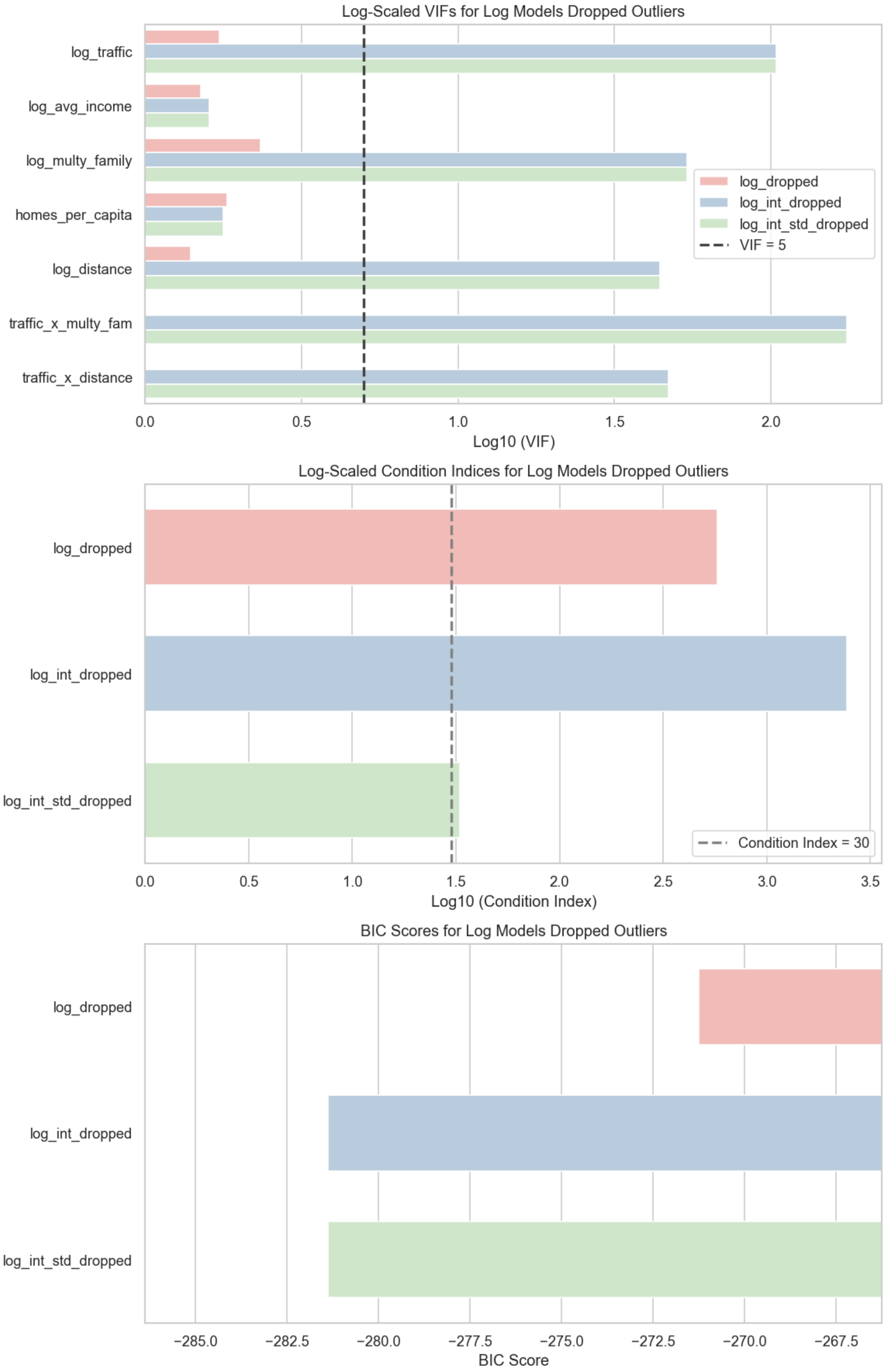


Figure 14: Multicollinearity Indicators and BIC Scores (Dropped Outliers, Table 7)

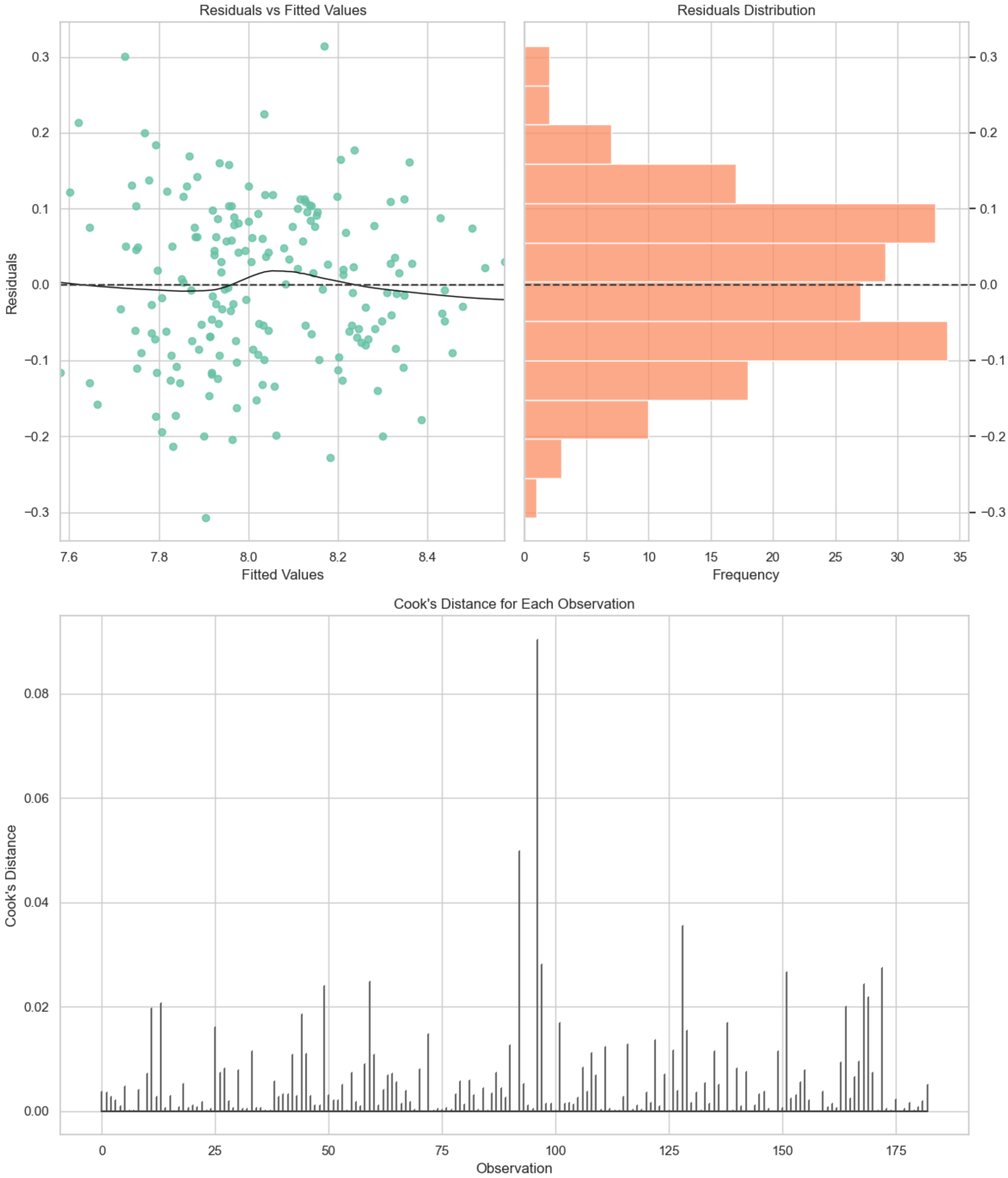


Figure 15: Residual Analysis of Regression Models (Dropped Outliers, Table 7)

Residual plots of the interaction model are presented on Figure 15. The plots show significant improvement after excluding the influential outliers. The distribution and the scatterplot indicate no autocorrelation, no dependency and no significant deviancy from the normal distribution. The visualization of Cook's distances indicates that the removal of outliers largely improved the equivalence of residual influences on fitted values.

Lastly, normality and autocorrelation tests along with R^2 values are presented in Table 8. The table indicates that in the case of the initial models, normality of residuals is rejected ($p < 0.01$). However, dropping influential outliers improved test statistics and p-values across all tests, with both the Simple and the Interaction model showing very high p-values ($p = 0.991$ and 0.827 for the Omnibus, and $p = 0.958$ and 0.775 for the Jarque-Bera, respectively). Furthermore, Durbin-Watson statistics also show slight improvement after the removal of outliers, however, values are close to 2 across all models indicating no significant risk of autocorrelation. The table also shows that R^2 values have improved by dropping influential outliers, with the interaction model explaining higher proportion of the variance.

Metric	**Simple**	**Interaction**	**Dropped**	**Dropped (Int.)**
Omnibus	9.416	14.185	0.018	0.381
Omnibus p-value	0.009	0.001	0.991	0.827
Jarque-Bera	11.819	25.722	0.085	0.510
Jarque-Bera p-value	0.003	0.000	0.958	0.775
Durbin Watson	1.747	1.737	1.869	1.777
R^2	0.772	0.792	0.796	0.822
Adjusted R^2	0.766	0.784	0.790	0.815

[1] Simple and Interaction refer to Equation 3 and 4 respectively. Dropped and Dropped (Int.) are their counterparts with outliers removed.

Table 8: Summary of Test Scores and R^2 for Regression Models (Table 6 and 7)

Discussion

In conclusion, the effect of rail infrastructure on house prices in The Netherlands has been assessed and this section seeks to discuss them in light of the literature. The analysis consisted of two phases. The first phase focused on investigating the effect of having a train station in a municipality. The results showed that the presence of a train station has no significant impact on house prices. Key predictors of house prices were average income and the proportion of multi-family homes, both showing positive and significant effects, while the number of homes per capita had a significant negative effect. The insignificant effect of train station presence is generally in line with the literature for various reasons. It has been established that both in international and Netherlands-specific context, the evidence on the impact of rail infrastructure on house prices is mixed. Externalities such as emissions, noise and crime (Bowes & Ihlanfeldt, 2001) can balance out the positive accessibility benefits of a train station which may lead to statistically insignificant result. Furthermore, most studies presented look at effects at a local scale, and conclude that if benefits are found, they effect properties within a short distance from the station (Debrezion et al., 2005). Municipalities are relatively large however, and such local effects may not have an impact on average housing prices.

In the second phase, the analysis focused on municipalities with train stations to explore the effect of daily traffic levels on house prices. The simple regression model indicated a small but significant positive effect of traffic on house prices. The interaction model revealed that the effect of traffic is not significant, however the interaction term with the proportion of multi-family homes is positive and statistically significant. After excluding influential outliers, the interaction model showed that increased traffic generally has a negative effect on house prices, which is mitigated and can turn positive in municipalities with a higher share of multi-family housing. As the share of multy-family housing is highly correlated with population density (see Figure 4), the results of the analysis indicate that in rural municipalities with low population density, the negative price effects related to externalities balance out or even outweigh the positive impact of connectivity, while in urban areas with large population density the positive accessibility benefits are dominant. This result highlights the context-specific nature of the research. The results are in line with several studies that find positive railway accessibility effects in urban areas, (Dubé et al., 2013; Lieske et al., 2021; Syabri, 2011). Additionally, they confirm findings of studies that find weaker effects in rural or less urban regions (Debrezion et al., 2011; Paliska & Drobne, 2020).

Conclusion

The paper focused on the impact of train station presence and daily train traffic level on house prices in Dutch municipalities. Firstly, a literature review was conducted, establishing the connection between accessibility and property prices, rail infrastructure and property prices, and lastly rail infrastructure and property prices in The Netherlands. The referred studies provide mixed findings, which depend on several factors such as urban context, model choice, or type of railway. Furthermore, the effects of externalities such as noise, vibration, pollution or crime, has been assessed in both national and international context.

The study employed a dataset consisting of square meter house prices, railway infrastructure data and several economic, geographic and demographic control variables across all municipalities of The Netherlands for the year 2023. The study utilized Ordinary Least Squares as a methodology, using multiple regressions with interaction effects. The analysis was conducted in two phases: first, examining the effect of having a train station in a municipality, and second, analyzing the impact of daily train traffic levels in municipalities with train stations.

The results of the first phase indicate that the presence of a train station does not significantly affect house prices at the municipal level. This finding aligns with existing literature, which suggests that accesibility benefits of a train station are often balanced out by negative externalities. The second phase presents a granular impact of train traffic on house prices. While the initial analysis suggests a positive relationship, further examination with interaction terms shows that this effect is context-dependent. Specifically, the analysis shows that in itself, 1% increase in daily train traffic leads to a 0.26% decrease in house prices, however this value depends on the share of multy-family housing in the municipality. In rural municipalities with lower share of multy-family housing, increased train traffic tends to have a negative impact on house prices, likely due to negative externalities. In contrast, in more urbanized areas with a higher share of multi-family housing, the positive accessibility benefits of increased train traffic appear to outweigh the negative impacts.

However, this paper is subject to several limitations. Firstly, due to time constraints, highway data was not incuded in the dataset. Debrezion et al. (2007) find significant omitted variable bias in models on the impact of station proximity on property values that do not include highway accessibility related controls. Bohman and Nilsson (2016) also find that highway accessibility significantly impacts house prices. This suggests that the study may be subject to omitted variable bias and effects might be smaller than presented. Furthermore, due to the changing borders of municipalities, building a multy-year panel dataset of the entire country would have been challenging considering the scope of this thesis. However, this suggests that the ability of this study to predict long-term trends is limited. Furthermore, indicators such as mortgage rates which might affect the year-to-year development of housing prices are also missing due to the nature of the data. Future research could benefit from collecting data for highway accessibility and employing a longitudinal dataset over several years. This, however, presents significant challenges due to the frequent changes of municipality borders.

The paper contributes to literature by presenting a detailed analysis on the municipal house price effects of rail frequency. Furthermore, it confirms the mixed findings of previous literature on the impact of station presence. The findings of this research offer important insights to municipal policymakers and important implications for rail infrastructure planning and real estate projects, as the detailed effects of daily train frequency in rural and urban areas provide a more granular understanding of its impact on house prices.

References

Alonso, W. (1964). *Location and land use: Toward a general theory of land rent*. Harvard University Press.

Azubuike, I. M., & Tobe, N. M. (2019). A review of standardization and centering techniques in a multicollinear regression model. *Scientia Africana*, *18*(3), 97–110.

Bohman, H., & Nilsson, D. (2016). The impact of regional commuter trains on property values: Price segments and income. *Journal of Transport Geography*, *56*, 102–109.

Bowes, D. R., & Ihlanfeldt, K. R. (2001). Identifying the impacts of rail transit stations on residential property values. *Journal of urban Economics*, *50*(1), 1–25.

Camins-Esakov, J., & Vandegrift, D. (2018). Impact of a light rail extension on residential property values. *Research in Transportation Economics*, *67*, 11–18.

Centraal Bureau voor de Statistiek. (2024). Bestaande koopwoningen; verkoopprijzen prijsindex 2020=100 [Accessed: 2024-06-16]. https://opendata.cbs.nl/statline/#/CBS/nl/dataset/85773NED/table?ts=1718560203565

Debrezion, G., Pels, E., & Rietveld, P. (2005). Impact of railway station on dutch residential housing market.

Debrezion, G., Pels, E., & Rietveld, P. (2007). The impact of railway stations on residential and commercial property value: A meta-analysis. *The journal of real estate finance and economics*, *35*, 161–180.

Debrezion, G., Pels, E., & Rietveld, P. (2011). The impact of rail transport on real estate prices: An empirical analysis of the dutch housing market. *Urban studies*, *48*(5), 997–1015.

Dubé, J., Thériault, M., & Des Rosiers, F. (2013). Commuter rail accessibility and house values: The case of the Montreal South Shore, Canada, 1992–2009. *Transportation Research Part A: Policy and Practice*, *54*, 49–66.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage publications.

Jiao, J., Wang, J., Zhang, F., Jin, F., & Liu, W. (2020). Roles of accessibility, connectivity and spatial interdependence in realizing the economic impact of high-speed rail: Evidence from China. *Transport Policy*, *91*, 1–15.

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean journal of anesthesiology*, *72*(6), 558.

Koster, H., van Ommeren, J., & Rietveld, P. (2012). *The gains of trains: The effect of station openings on house prices* (tech. rep.). Tinbergen Institute.

Lieske, S. N., van den Nouwelant, R., Han, J. H., & Pettit, C. (2021). A novel hedonic price modelling approach for estimating the impact of transportation infrastructure on property prices. *Urban studies*, *58*(1), 182–202.

Maclachlan, L., Ögren, M., Van Kempen, E., Hussain-Alkhateeb, L., & Persson Waye, K. (2018). Annoyance in response to vibrations from railways. *International journal of environmental research and public health*, *15*(9), 1887.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Neath, A. A., & Cavanaugh, J. E. (2012). The bayesian information criterion: Background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 199–203.

Nederlandse Spoorwegen. (2024). Ns annual report 2023 [Accessed: 2024-06-16]. https://www.nsannualreport.nl/external/asset/download/project/f9822a0a-03ec-0000-4a30-4363ddceac2a/name/NS_annualreport_2023.pdf

Paliska, D., & Drobne, S. (2020). Impact of new motorway on housing prices in rural North-East Slovenia. *Journal of Transport Geography*, *88*, 102831.

Paul, R. K. (2006). Multicollinearity: Causes, effects and remedies. *IASRI, New Delhi*, *1*(1), 58–65.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of political economy*, *82*(1), 34–55.

Syabri, I. (2011). The Influence of Railway Station on Residential Property Values-Spatial Hedonic Approach the Case of Serpong's Railway Station. *Jurnal Teknik Sipil ITB*, *18*(3), 291–300.

von Thünen, J. (1826). *Der isolirte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Wirtschaft & Finan.

Appendix

Appendix 1

- **CBS Statline**: This source provided data on population density (*pop_density*), average annual income (*avg_income*), unemployment rate (*unemp_rate*), net labor participation (*net_labor_participation*), housing prices (*m2_price*), homes per capita (*homes_per_capita*), and share of multi-family housing units (*multy_family*).
- **Rijden De Treinen**: This source provided data on number of railway stations (*station_count*), average daily traffic at railway stations (*traffic*), and the presence of at least one railway station (*has_station*).
- **Publieke Dienstverlening Op de Kaart (PDOK)**: This source provided a GeoJSON outline of Dutch municipalities as of 2023. This file was used to calculate the *distance_from_urban_center* variable, and to create the map shown on Figure 1.

Appendix 2

In both phases of the analysis, 10 municipalities with the highest Cook's index values are dropped.

Phase 1	**Phase 2**
Landsmeer	Maastricht
Noord-Beveland	Beek
Zandvoort	Leiden
Bloemendaal	Delft
Ameland	Diemen
Tubbergen	Eijsden-Margraten
Wageningen	Amsterdam
Amsterdam	's-Gravenhage
Staphorst	Zandvoort
Renswoude	Bloemendaal

Table 9: List of Dropped Municipalities