Attribution Analysis of the House Sale Prices Based on the Multiple Linear Regression

Li Si Xuan

21/12/2020

Link to a Github repository

Code and data supporting this analysis is available at: https://github.com/LLINDAx/STA304-Final-Project

Abstract

The visual analysis of the cleaned data and further imputation analysis of the house sale prices were performed. Since the relationship between other variables and the house sale prices was of concern, a multiple linear regression model was used and the most significant house's characteristics affecting the house sale prices were found to be the number of bedrooms, the square footage of the land space and the number of floors.

Keywords

Multiple Linear Regression, Real Estate, Prediction, House Sale Prices

Introduction

Statistics play an important role in many fields, such as real estate. For most people, buying or selling a house is a relatively large financial transaction in their lifetime. First-time buyers might meet overpaying. They may have met a seller with a high bid, but the house is not worth that much money. On the other hand, current homeowners might meet problems with losing money in a sale because they don't know how much their house should sell. Hence, it can be seen that both buyers and sellers may encounter complex problems in real estate transactions.(S., 2018)

Furthermore, real estate investment has always been a hot topic. Many people choose to invest in real estate. Because compared with other investment projects, such as stocks and bonds, there tends to be less volatility in real estate. Also, real estate investment is relatively safe because it has a high tangible asset value and a large appreciation space(R., 2020).

Whether you want to buy a house for your own living or for investment, you need to be able to understand the housing prices in recent years. And what will affect the price of a house, such as the location, size, etc. Therefore, we analyzed the regression of these houses' characteristics on the house sale prices by building a multiple linear regression model. Further, the evaluation of the multiple linear regression model led to some conclusions on the relationship between the house's characteristics and the house sale prices, and to achieve our goal of predicting housing prices. Finally, in the discussion of the experimental procedure, it is suggested that the model may have some stability problems caused by endogeneity and some directions for improvement are proposed.

Methodology

Data

The dataset includes houses sold between May 2014 and May 2015 for King Country, USA. And this dataset comes from Kaggle, which is a website that allowed users to find freely available dataset (Harlfoxem, 2016). There are 21613 observations and 21 variables in the dataset.

The variables in the dataset are:
id: Unique ID for each home sold
date: Date of the home sale
price: Price of each home sold
bedrooms: Number of bedrooms
bathrooms: Number of bathrooms, where .5 accounts for a room with a toilet but no shower
sqft_living: Square footage of the apartments interior living space
sqft_lot: Square footage of the land space
floors: Number of floors
waterfront: - A dummy variable for whether the apartment was overlooking the waterfront or not
view: An index from 0 to 4 of how good the view of the property was
condition: - An index from 1 to 5 on the condition of the apartment,
grade: An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.
sqft_above: The square footage of the interior housing space that is above ground level
sqft_basement: The square footage of the interior housing space that is below ground level
yr_built: The year the house was initially built
yr_renovated: The year of the house's last renovation
zipcode: What zipcode area the house is in
lat: Lattitude
long: Longitude
sqft_living15: The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15: The square footage of the land lots of the nearest 15 neighbors
(Mantero, 2020)

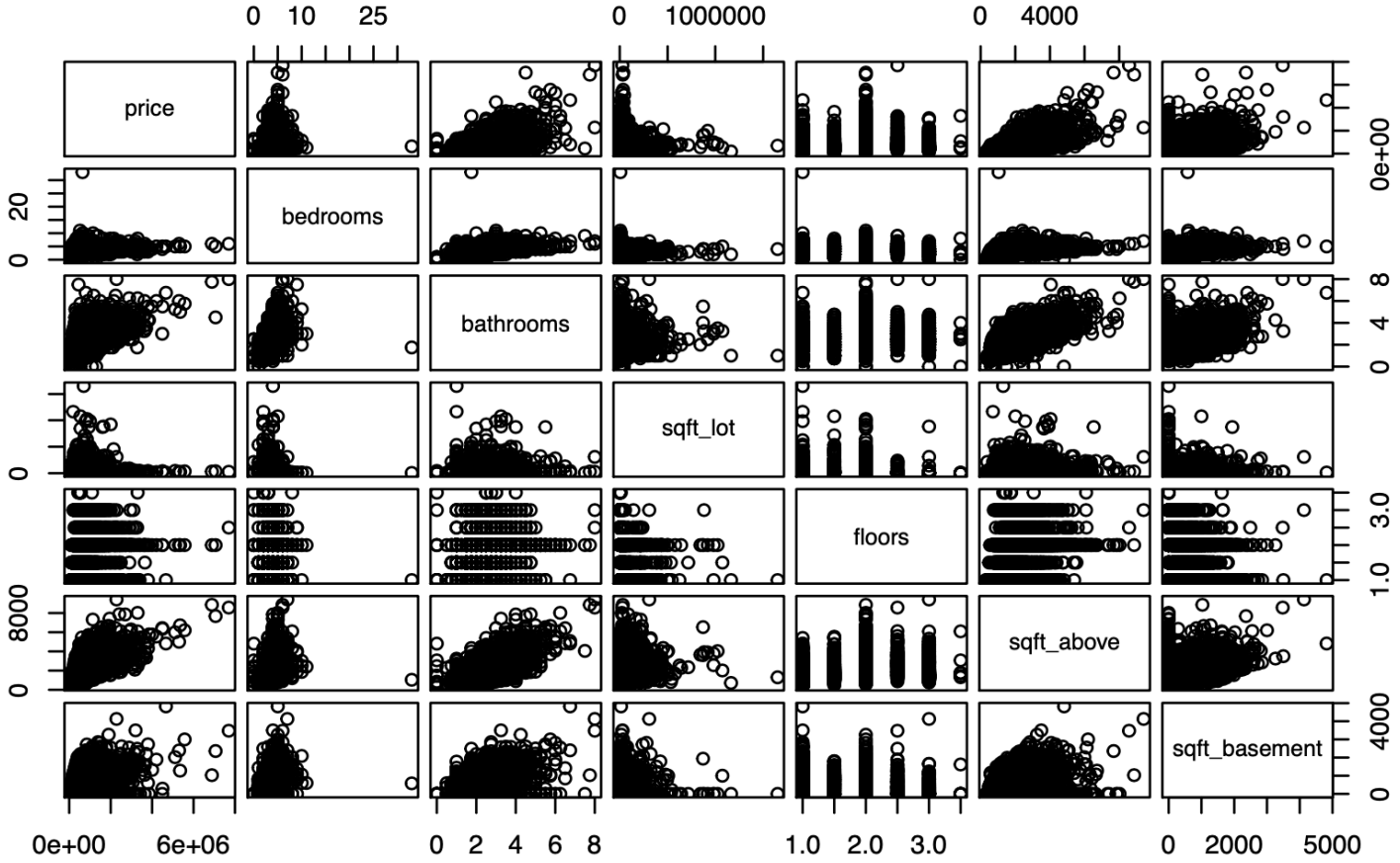
Since our goal is to predict housing prices, we need to remove some unrelated variables and some unuseful variables, such as **id**, **lat**, **date**, etc. Also, we need to remove overly complex variables and some variables are almost the same, such as **zipcode**, **sqft_lot15**, **sqft_living**, etc. Then, we use the remaining variable to form a new dataset. There are 7 variables in it. The data is cleaned and analyzed by RStudio. The following table is showing summary statistics of this dataset:

Table 1: Summary Statistics of Data

	Overall
n	21613
price (mean (SD))	540088.14 (367127.20)
bedrooms (mean (SD))	3.37 (0.93)
bathrooms (mean (SD))	2.11 (0.77)
sqft_lot (mean (SD))	15106.97 (41420.51)
floors (mean (SD))	1.49 (0.54)
sqft_above (mean (SD))	1788.39 (828.09)
sqft_basement (mean (SD))	291.51 (442.58)

From the table 1, we can see the means and the SDs of each variables in the dataset.

Figure 1: Scatterplot Matrix



The above figure is the scatterplot matrix for all pairs of variables in the data. From the plot, we can see the rough relationships between any of two variables. For instance, the scatterplot of price and the square footage of the interior housing space that is above ground level shows that there is a positive relationship between these two variables. Hence, we can know that as the square footage of the interior housing space that is above ground level increases, the price will increases.

Model

The chosen model is the multiple linear regression model. A multiple linear regression model "is a statistical technique that uses several explanatory variables to predict the outcome of a response variable"(Kenton,2020). Since we have only one variable of primary interest, the response variable, and our analytical interest lies in the effect of other variables on him, we use a multiple linear regression model here. We want to see how **bedrooms**, **bathrooms**, **sqft_lot**, **floors sqft_above** and **sqft_basement** can affect **price**. I fitted an additive linear regression model with these predictors variable for sale price. Then, I use stepwide regression with AIC to try to find a parsimonious model. Backward elimination starts with all the potential predictors in the model, then removes the predictors with the largest p-value each time to give a smaller AIC. Thus, after using this method, the estimated multiple linear regression model between response variable and explanatory variables is:

$\hat{price} = \hat{\beta}_0 + \hat{\beta}_1 bedrooms + \hat{\beta}_2 sqftlot + \hat{\beta}_3 floors + \hat{\beta}_4 sqftabove + \hat{\beta}_5 sqftbasement.$

Results

Table 2: Summmary of Fitted Model

Coefficients	Estimate	Std. error	t value	p-value
(intercept)	75071.59272	7736.00748	9.704	< 0.0001
bedrooms	-59530.13626	2315.45880	-25.710	< 0.0001
sqft_lot	-0.34983	0.04342	-8.057	< 0.0001
floors	13194.49708	3963.17361	3.329	0.000872
sqft_above	308.57544	2.88138	107.093	< 0.0001
sqft_basement	340.97734	4.39399	77.601	< 0.0001

Residual standard error	Multiple R-squared	Adjusted R-squared	Pr(>F)
257100	0.5095	0.5094	< 0.0001

The fitted model is:

$\hat{price} = 75071.59272 - 59530.13626 * bedrooms - 0.34983 * sqftlot + 13194.49708 floors + 308.57544 * sqftabove + 340.97734 * sqftbasement$

The **standard error** of $\hat{\beta}_0$ is **7736.00748**.

The **standard error** of $\hat{\beta}_1$ is **2315.45880**.

The **standard error** of $\hat{\beta}_2$ is **0.04342**.

The **standard error** of $\hat{\beta}_3$ is **3963.17361**.

The **standard error** of $\hat{\beta}_4$ is **2.88138**.

The **standard error** of $\hat{\beta}_5$ is **4.39399**.

For β_0:

Assume $H_0 : \beta_0 = 0$, $H_a : \beta_0 \neq 0$. The **p-value** < 0.0001, which is **smaller** than $\alpha = 0.05$. So we have very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_0 \neq 0$.

For β_1:

Assume $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$. The **p-value** < 0.0001, which is **smaller** than $\alpha = 0.05$. So we have very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_1 \neq 0$, which means there is a relation between the number of bedrooms and the house prices.

For β_2:

Assume $H_0 : \beta_2 = 0$, $H_a : \beta_2 \neq 0$. The **p-value** < 0.0001, which is **smaller** than $\alpha = 0.05$. So we have very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_2 \neq 0$, which means there is a relation between the square footage of the land space and the house prices.

For β_3:

Assume $H_0 : \beta_3 = 0$, $H_a : \beta_3 \neq 0$. The **p-value** $= 0.000872$, which is **smaller** than $\alpha = 0.05$. So we have very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_3 \neq 0$, which means there is a relation between the number of floors and the house prices.

For β_4:

Assume $H_0 : \beta_4 = 0$, $H_a : \beta_4 \neq 0$. The **p-value** < 0.0001, which is **smaller** than $\alpha = 0.05$. So we have very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_4 \neq 0$, which means there is a relation between the square footage of the interior housing space that is above ground level and the house prices.

For β_5:

Assume $H_0 : \beta_4 = 0$, $H_a : \beta_4 \neq 0$. The **p-value** < 0.0001, which is **smaller** than $\alpha = 0.05$. So we have very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_5 \neq 0$, which means there is a relation between the square footage of the interior housing space that is below ground levels and the house prices.

The interpretation of the estimates of the regression coefficients will be explained in discussion section.

The $\sqrt{MSE}$ of this estimated model is **257100**. And the R^2 of this estimated model is **0.5095**, which means 50.95% of the total variation in y is explained by the regression line. The adjusted R^2 is **0.5094**, which is adjusted for the number of predictors in the model. It is better to use instead of R^2. So, we can see that the adjusted R^2 is not high, which the total variation in y is not explained well by the regression line. The **p-value** for global F test is < 0.0001, which implies that the model contains at least one significant predictor among the set of p predictors.

Table 3: Variance Inflation Factor

bedrooms	sqft_lot	floors	sqft_above	sqft_basement
1.515833	1.057233	1.496958	1.860852	1.236078

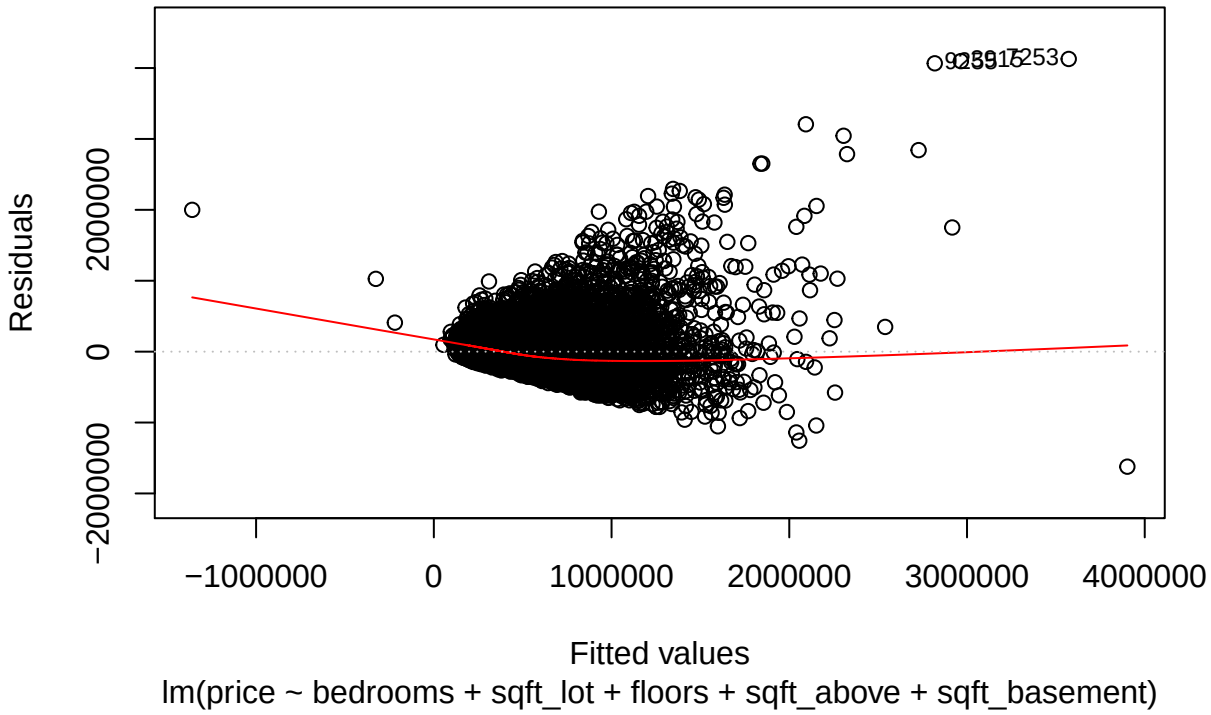
"A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model."(Staff, 2020) Multicollinearity occurs when predictors are highly correlated, which might cause that the fitted equation is unstable, and the estimated regression coefficients have opposite sign. We can see that VIFs are not high, which means predictors are not highly correlated.

Table 4: Correlation

1	2	3	4	5
sqft_above	sqft_basement	bedrooms	floors	sqft_lot
0.60556730	0.32381602	0.30834960	0.256793888	0.089660861

The above table is the ranking of the predictors for sale price, in terms of their correlation coefficient, from highest to lowest. We can see that the square footage of the interior housing space that is above ground level has highest correlation with the house sale prices.

Figure 2: Residuals vs Fitted

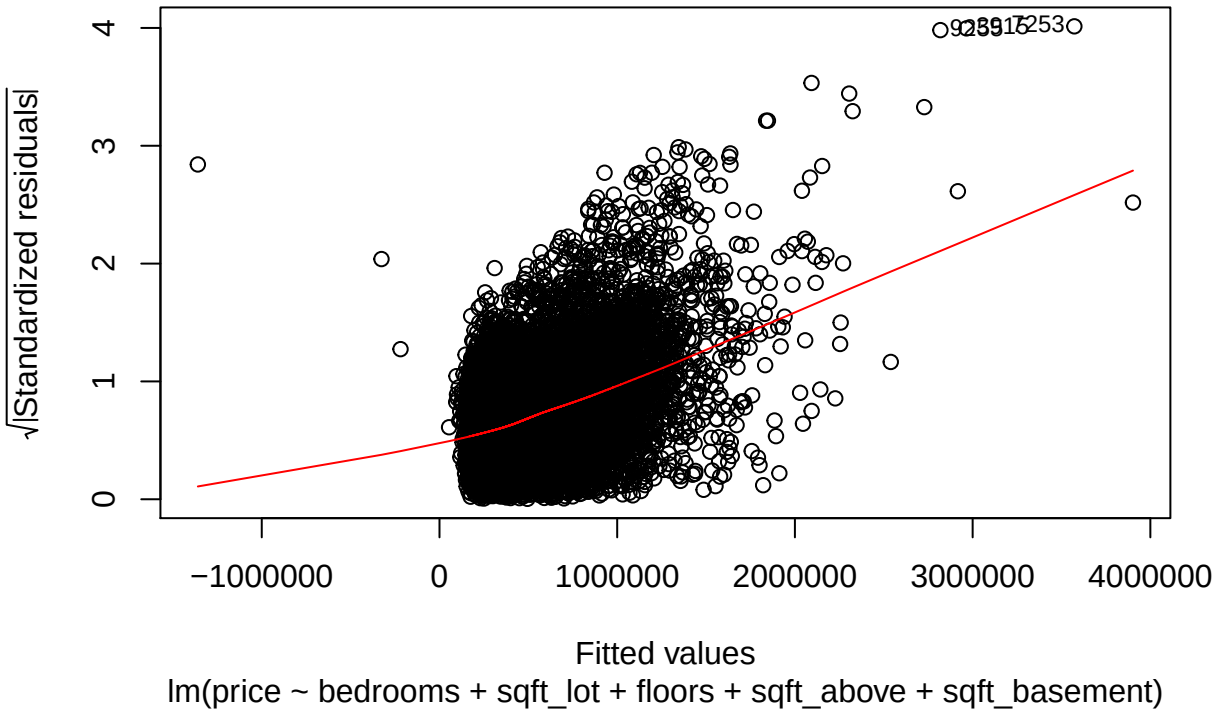


The **Residuals vs Fitted plot** shows if residuals have non-linear patterns. From the plot, we can see residuals are not equally spread around the horizontal line, and there is a increasing trend. Hence, the fitted model does not meet the assumption of linearity. In addition, we can see that there are some points in the upper right corner with large positive residuals.

Figure 3: Normal Q−Q

The **Normal Q-Q plot** shows if residuals are normally distributed. From the plot, we can see that residuals do not follow a straight dashed line well, especially for the tail part, which is called heavy-tailed. Hence, the fitted model does not meets the assumption of normality. We can transform Y or fit a time series model as remedies. In addition, we can see that there are some high standardized residuals in the upper right corner, so these points are outliers. We need to do further observations on these point and check if the fit will be better if these point are removed.

Figure 4: Scale−Location



The **Scale-Location plot** shows if residuals have constant variance. From the plot, we can see that the variance does not seem to be constant because we cannot see a horizontal line with equally spread points. And the variance has an increasing trend. Hence, the fitted model has a violation with the assumption of homoscedasticity. We need to transform x and/or Y, or we can do weighted least squares as remedies.

Discussion

Summary

We got the dataset comes from Kaggle and performed a preliminary cleaning of the data. Then, we geneate a new dataset with the remaining variables which are suitable to predict house price. There are 7 variables and 21613 observations in the new dataset. Then, we analyze the data by building a multiple linear regression model. The explanatory variable are **bedrooms**, **sqft_lot**, **floors sqft_above** and **sqft_basement**, and the response variable is **price**. The purpose is to see how these predictors affect housing prices and use these predictors to predict housing prices.

Conclusions

The fitted multiple linear regression model is:

$\hat{price} = \hat{\beta}_0 + \hat{\beta}_1 bedrooms + \hat{\beta}_2 sqftlot + \hat{\beta}_3 floors + \hat{\beta}_4 sqftabove + \hat{\beta}_5 sqftbasement = 75071.59272 - 59530.13626 * bedrooms - 0.34983 * sqftlot + 13194.49708 floors + 308.57544 * sqftabove + 340.97734 * sqftbasement$

From the summary of the fitted model, we can see that the t-test results of all the estimated model coefficients were significant. When the number of bedrooms increases one unit, on average sale price decreases 59530.13626 dollars, holding all other explanatory variables in the model fixed. When the square footage of the land space increases one unit, on average sale price decreases 0.34983 dollars, holding all other explanatory variables in the model fixed. When the number of floors increases one unit, on average sale price increases 13194.49708 dollars, holding all other explanatory variables in the model fixed. When the square footage of the interior housing space that is above ground level increases one unit, on average sale price increases 308.57544 dollars, holding all other explanatory variables in the model fixed. When the square footage of the interior housing space that is below ground level one unit, on average sale price increases 340.977346 dollars, holding all other explanatory variables in the model fixed.

Therefore, we believe that the house's characteristics all have a significant impact on the house sale price. The result can be used to predict the house price in the future. We just need to know these house's characteristics, then we can get the estimated house sale price. This allows first-time buyers, current homeowners, and people who want to invest in the real estate to predict the house prices before deciding to buy or sell.

Weakness & Next Steps

From the result section, three diagnostic plots were drawn. We can see that the normal error multiple linear regression assumptions are not satisfied, which means that "the forecasts, confidence intervals, and scientific insights yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading."(F., n.d.) Hence, the next step is that we need to use some remedies to deal with model assumption violations.

For the Linearlity, we can add additional predictors, or transform x and/ or Y as remedies. For the Normality, we can transform Y, or fit a generalized linear model, or fit a time series model as remedies. In additional, we can do the Box-Cox Transformations,which makes the transformed variable close to normally distributed. For the Homoscedasticity, we can transform x and/ or Y, or do weighted least squares, or fit a generalized linear model(models variance as a function of the mean) as remedies.

Furthermore, it is also important to do some further observations on noteworthy points, points with high leverage, outliers, and influential points. We can check the Cook's distance, the leverage, and absolute standardized residuals in more detail by code. Then we need to check if the fit will be better if these points are removed.

In order to make this model better and predict housing prices more accurately, we need to implement the steps mentioned above.

References

Harlfoxem. (2016, August 25). House Sales in King County, USA. Retrieved December 20, 2020, from https://www.kaggle.com/harlfoxem/housesalesprediction

Mantero, T. (2020, November 23). Predicting House Prices (Keras - ANN). Retrieved December 20, 2020, from https://www.kaggle.com/tomasmantero/predicting-house-prices-keras-ann

Kenton, W. (2020, September 21). How Multiple Linear Regression Works. Retrieved December 20, 2020, from https://www.investopedia.com/terms/m/mlr.asp

Staff, I. (2020, October 23). Variance Inflation Factor (VIF). Retrieved December 22, 2020, from https://www.investopedia.com/terms/v/variance-inflation-factor.asp

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

F. (n.d.). Regression diagnostics: testing the assumptions of linear regression. Retrieved December 22, 2020, from http://people.duke.edu/~rnau/testing.htm

R. (2020, August 21). Is real estate still a good Investment? Yes, and here's why. Retrieved December 23, 2020, from https://blog.remax.ca/is-real-estate-a-good-investment/

S. (2018, October 05). Why Real Estate Statistics Matter. Retrieved December 23, 2020, from https://www.showingtime.com/blog/why-real-estate-statistics-matter/

Kazuki Yoshida and Alexander Bartel (2020). tableone: Create 'Table 1' to Describe Baseline Characteristics with or without Propensity Score Weights. R package version 0.12.0. https://CRAN.R-project.org/package=tableone

John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/