

What is Affecting Toronto Condo Asking Price?*

Sijian Wu

27 April 2022

Abstract

Toronto Housing market and real estate has boomed over the last decade. Investors would often face questions such “what is affecting Toronto real estate prices?” This analysis consists of logistic Multiple Linear Regression performed with the statistical programming language R. The results estimate the statistical linear relationship of Toronto condo asking price with Distance to Downtown, Building Age Maximum, and Square Foot Minimum. We aimed to provide reasonable information for audience like buyers, sellers, and agencies who are interested in condo trading in Toronto with price lower than \$1.5 million.

Keywords: Toronto, Real Estate, Condo, MLR, Asking Price, Regression, Distance to Downtown

Contents

1	Introduction	2
2	Data	3
2.1	Data Source	3
2.2	Methodology and Data Collection	3
3	Model	6
3.1	Statistical Model for Multiple Linear Regression	6
3.2	Validation and Assumptions	6
4	Results	8
5	Discussion	17
5.1	Implications and Ethical Considerations	17
5.2	Caveats and Limitations	17
5.3	Recommendations and Next Steps	17
6	Appendix	19
6.1	Appendix1	19
6.2	Appendix2	21
6.3	Appendix3	22
6.4	Appendix4	23
References		24

*Code and data are available at: <https://github.com/NnnnninaW/Totonto-Condo-Asking-Price-Analysis.git>. Complementary Survey available at:https://docs.google.com/forms/d/e/1FAIpQLScxK6foA19kYJqFzpFWQokcmxL_R3VimvgbpNlmX-YU3opg/viewform.

1 Introduction

The Toronto Housing market has boomed over the last decade. Investors would often face questions like “what is affecting Toronto real estate prices?” First, we need to understand if the housing market is mainly fueled by supply and local demand, or if there are external factors. Also, condos are an entirely different segment than detached houses, in the same way, Hamilton is an entirely different market than Toronto (Scrinko, 2021). Real estate investment has always been an essential part of portfolio management for individuals since it can be extremely durable. Investing in real estate isn’t just for the rich and famous. Anyone can do it, even if you only have a small down payment. With the right preparation and research, anyone can enjoy the profits real estate offers (“8 Major Reasons Why You Should Invest in Real Estate Starting Today” 2021).

Moreover, the real estate market is usually localized due to geographic and economic differences. All real estate is local, with every property being unique in terms of location, physical structure, and financing (Vazquez 2015). So, our analysis would focus only on the Toronto Condo Apartment market. This category includes all apartments and condominiums. These are complexes of residential units with common areas such as hallways, parking lots, stairwells, etc. (WOWA, 2021). Since the average price of condos was \$715,104 in November 2021, we would narrow our research focusing on Condos with a price lower than \$1.5 million despite those with extremely high prices.

We retrieved data from the website Condos.ca(<https://www.condos.ca/>) on 2021, December 5 for all kinds of information and asking prices that are public to potential buyers and sellers. Our research would be conducted through the Multiple Linear Regression (MLR) model to estimate the statistical linear relationship between one outcome/response variable with two or more explanatory/predictor variables (Sheather 2009). Our final model has response variables as Toronto condo asking price, and predictor variables such as Distance to Downtown, Building Age Maximum, and Square Foot Minimum. We aimed to provide reasonable information for audiences like buyers, sellers, and agencies who are interested in condos in Toronto with a price lower than \$1.5 million.

We should also keep in mind that the real estate market works very differently. When an investor buys a property, there is a risk that the seller is withholding information, or that the seller is unaware of problems. Therefore, research and inspections are important when buying real estate. And if you buy a property sight-unseen, such as at auction, be sure to consider the higher risk when making an offer (Vazquez 2015). Real estate investment is also considered very illiquid, so it is necessary for investors to consider cautiously before investing.

We would introduce our methods, results, interpretations, and limitations in the following sections. Section 2 would illustrate the source of our data and collection methods. Section 3 would introduce the MLR model including assumptions and measurements. Section 4 would demonstrate our analysis process and present our results for reference. Section 5 would evaluate our findings with ethical considerations, limitations, and recommendations for the next step. Section 6 includes complimentary information mentioned in our analysis. The last References Section is for citations for our research.

2 Data

We started our analysis by using R (R Core Team 2020), tidyverse(Wickham et al. 2019), here(Müller 2020), car(Fox and Weisberg 2019), GGally(Schloerke et al. 2021), leaps(Fortran code by Alan Miller 2020), MASS(Venables and Ripley 2002), knitr(Xie 2021), and ggplot2(Wickham 2016).

2.1 Data Source

We retrieved data from the website Condos.ca(<https://www.condos.ca/>) on 2021, December 5 for all kinds of information and asking prices that are public to potential buyers and sellers(“Condos.ca” 2021). Indeed, there are various real estate trade and listing websites available. However, we chose to obtain data from Condos.ca because it is the website that focuses on the condo trading market in Toronto which perfectly match our topic. It also contains substantial condo transactions with available information on all kinds of features of the condos. The framework of basic information is well structured on the website. Condo listings would provide needed information accordingly. This is an essential reason to choose this data since it indicates that people would be exposed to similar information so that it could limit impacts from external unknown factors. Even though the data may not be able to capture every transaction happened in Toronto, the well formatted information framework would provide a solid foundation for us to formalize our model. Also, the differences in prices would not be significant if the units have similar condition, especially same locations and areas.

There are 2511 entries in the original data set. As we would like to focus on condo prices, Figure 1 shows that there are sufficient observations for Condo so we could continue our analysis. The data provides information about identification, physical measurement, community, and current conditions of the condo. Most of the variables are numerical while most non-numeric variables are identification records of the subject which would not be analyzed. Usually, sellers would have a listing price that incorporates considerations of all kinds of information about the selling unit. While most of them would base their asking price logically, some of them may also misprice due to their financial needs like loans or mortgages.

2.2 Methodology and Data Collection

Before we start, we cleaned the dataset to focus on only condos in Toronto and only the asking price accordingly. Asking price, also known as the listing price, refers to how much a seller has listed a property for, and sale price refers to the amount it actually sells for. It usually reflects the market condition as sellers would provide the asking price according to all available information about the unit selling.

There is also much excess information like other property types and many inapplicable variables that would be problematic for us to build the model. We created a new predictor variable called “Distance to Downton” which combines all location information by using the CN Tower as a benchmark. To avoid extreme cases with abnormally high trading prices, we filtered condos with an asking price not higher than \$1.5 million. Some variables like bedrooms, bathrooms, and parking spots were inconsistent for comparison, so we transformed them as needed to keep the same unit measure. We also reverse selected variables that are inapplicable or unrelated to our topic. Therefore, there are 2002 observations and 22 variables after cleaning the data for further analysis.

Locations have been a popular topic whenever people think of real estate investment. Generally, the real estate market could be classified geographically since it also incorporates features like weather, environment, income level, economy, politics, security, financial policies, and transportation for investors to consider. To have a glimpse into our data, Figure 2 shows how the asking price is distributed according to the distance to downtown among different buildings. Referring to Figure 2, we could see that there are more relatively new buildings near downtown. Also, we could see large clusters of points below 1,000,000 and above 500,000 which implies that most of the condos have an asking price between \$0.5 and \$1 million dollars.

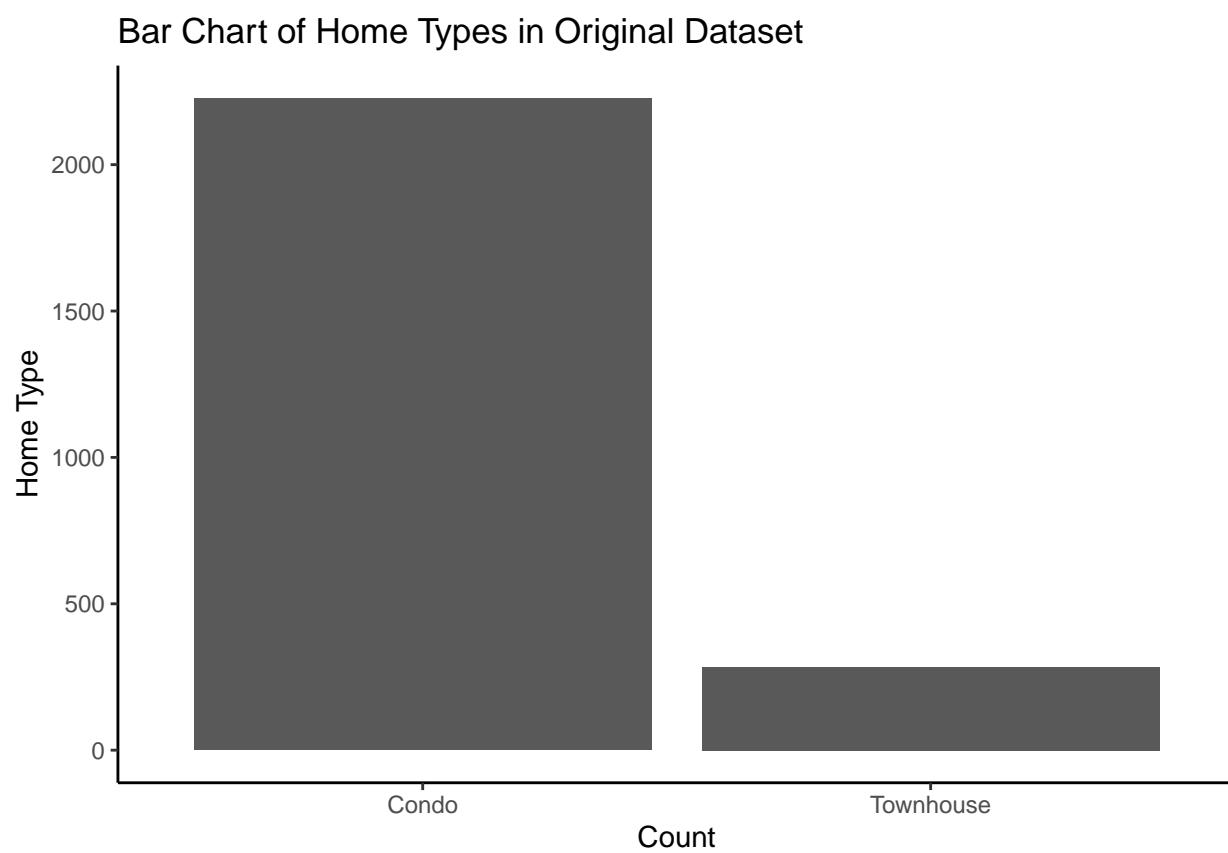


Figure 1: Bar Chart of Home Types in Original Dataset

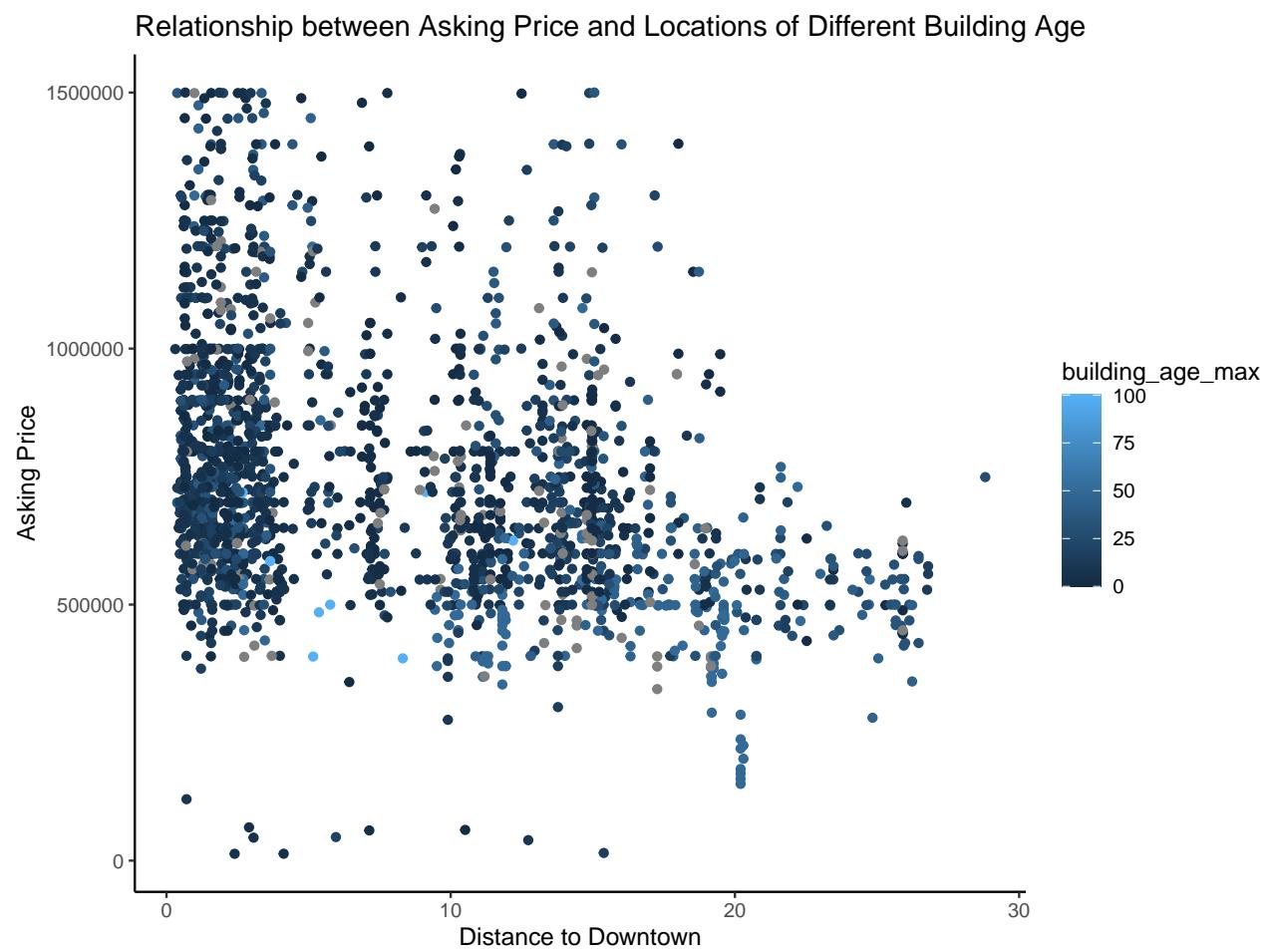


Figure 2: Relationship between Asking Price and Locations of Different Building Age Max

3 Model

To provide useful information for people interested in real estate investment, we need to estimate a relationship that describes how the change in our predictors corresponds to values in our response. As we are trying to express the statistical relationship, we are not restricted to only looking at the relationship between one predictor and the response. Moreover, when assessing whether the collected data fits our model, it is always important to assess whether any of the model's assumptions are violated. We also provided a Model Card as shown in the Appendix 6.1 (Stanton 2001; Sinharay 2010; Alam 2020; Brownlee 2021).

3.1 Statistical Model for Multiple Linear Regression

It is common to have more than one factor to influence an outcome. By applying Multiple Linear Regression Model (MLR), we expect to find numerous factors that are related to our response variable. Unlike in Simple Linear Regression, where there is only one predictor for the response variable, in MLR, the number of predictors is $p \in \mathbb{N}$ predictors.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

- Y_i is the response variable, a.k.a. dependent variable.
- X_{ij} is the explanatory variable, a.k.a. independent variable.
- β_0 is the intercept parameter. Statistically, it is the expected value of Y_i when $X_{ij} = 0$.
- β_i is the slope parameter.
- ε_i is the random error, caused by variation of measure that we cannot count for.

Since we are working with observations from a distribution, we need to incorporate error to represent that the data comes from a distribution that itself has variation. Just like Simple Linear Regression, the Multiple Linear Regression model also assumes that $\varepsilon_i \stackrel{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. Therefore, the Simple Linear Regression is a special case of Multiple Linear Regression Model, where $p = 1$. With this assumption, we know that the model has $p + 2$ parameters in total, $(p + 1) \beta_i$'s, and the variance of the error term, σ^2 . Since there are $(p + 2)$ parameters, we will need at least $p + 2$ observations in order to fit the model and get an estimate for the parameter.

3.2 Validation and Assumptions

First of all, it is crucial that we could validate our final model, so we use the Validation Methods to check if our final model is good enough not only for our collected sample but also for others from the same population. We randomly created a “train dataset” that would be used to implement all of our analysis and a “test dataset” for evaluations.

Secondly, we would perform an Exploratory Data Analysis (EDA) to identify potential model violations. Recall that Multiple Linear Regression model also assumes that $\varepsilon_i \stackrel{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. Therefore, we would also assess the assumptions mentioned below formally with Residual Scatter Plots. Residuals are the distance between the true values and our fitted value so that any violations in assumptions should also appear in the residuals. The estimated residual is defined by

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$$

We would like to visually assess whether an appropriate regression line has been fitted to the data. If no pattern is found in the residual scatter plots, then it indicates that the model provides an adequate summary of the data. If a pattern is found, then the shape of the pattern would provide information on the function of X that is missing from the model.

We also need to check two conditions first to examine if a linear relation model were inappropriate and if there are highly non-linear relations between our predictors. Simultaneously, we do not want any strong correlations between any of the predictor variables which is Multicollinearity. Multicollinearity could lead to misunderstanding and much larger standard errors. So, we would use the Variance Inflation Factor (VIF)

to measure how inflated the errors will be due to the relationship of one predictor to the others. If any significant violations are detected, we would see whether a transformation was needed.

Overall, we need to make sure it is reasonable for us to apply the MLR Model by checking these assumptions:

- Response variable is related to predictor variables by the linear regression model that only the predictors we are including as predictors are related to the response.
 - Do not want to see any systematic patterns in residual plots, such as a curve.
- The errors are uncorrelated.
 - Do not want to see large clusters of residuals that have obvious separation from the rest.
- The errors have a common variance.
 - Do not want to see any pattern especially a fanning pattern where residuals gradually become more spread out.
- The errors are Normally distributed.
 - Do not want to see severe deviations in residuals.

Then, we need to analyze whether there are problematic observations such as Leverage points, Outliers, and Influential points according to their measurements and cutoff values in Appendix 6.2. An **outlier** is a point located off the trend of the other observations. An outlier with respect to the response variable will lead to large residuals and can potentially affect the estimated regression line. A **leverage point** is characterized as being a single observation whose predictor values are overall quite distant from the center of mass of all predictors. They can have a disproportionate impact on the placement of the regression line (or equivalently on their own fitted value). **Influential points** generally do impact the estimated regression line but may not necessarily be a leverage point or outlier or both. It is important to check all three of these types of problematic observations for us to fully understand why the estimated regression line is being placed where it is.

Finally, we would go over the automated selection processes to learn about what are the suggested models. Since we are expected to have a large number of predictors initially, automated ways could be especially helpful to obtain an overall sense of preferred models. Throughout our analysis, we would also refer to measurements like Adjusted R^2 as complementary benchmarks. It is named adjusted R^2 because it is adjusted for the number of predictors in the model. It is better to use adjusted R^2 instead of R^2 since R^2 would no longer be helpful to tell whether additional predictors are useful for explaining the response variable as the number of variables (p) increases. Combine with contextual reasons, we would generate three preferred models for validation. We would fit them in our test dataset to compare their properties, understand potential limitations, and choose the final model. In the end, instead of modifying our previous analysis, we would conclusively discuss our results with observed limitations.

4 Results

Firstly, we look at the distribution of our response variable which is applicable for an MLR. From Figure 3, we could see the distribution of Asking Price is mostly normally distributed but slightly right-skewed. But it is Unimodal and good enough to proceed with the Multiple Linear Regression model.

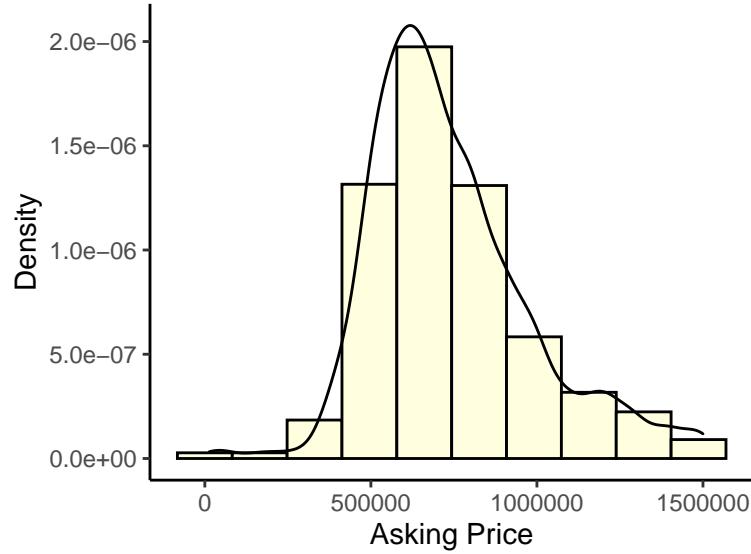


Figure 3: Asking Price Distribution Histogram

We also performed some Exploratory Data Analysis (EDA) to identify potential model violations. Our EDA results suggested there are potential linear relations between our response and predictor variables with no significant assumption violations. Figure 4 and Figure 5 shows the potential relation between numerical and categorical predictor variables and our response variable – Asking Price. However, we detected some variables that only have one level as shown in Appendix 6.3. Since they would be inapplicable and not meaningful for us to apply the MLR model, we removed them before we fit our initial full model.

Scatterplots of Numerical Predictor Variables

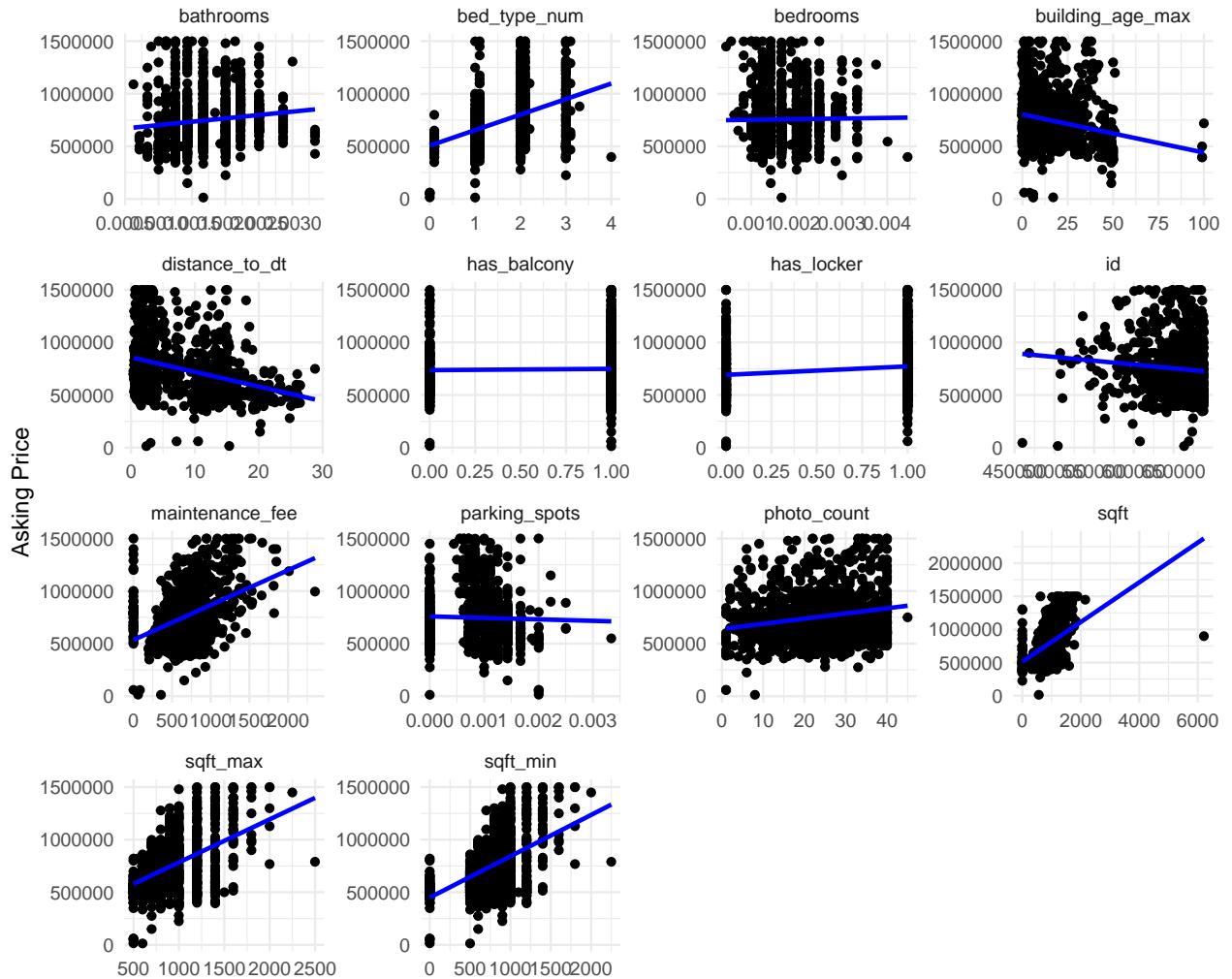


Figure 4: Scatterplots of Numerical Predictor Variables

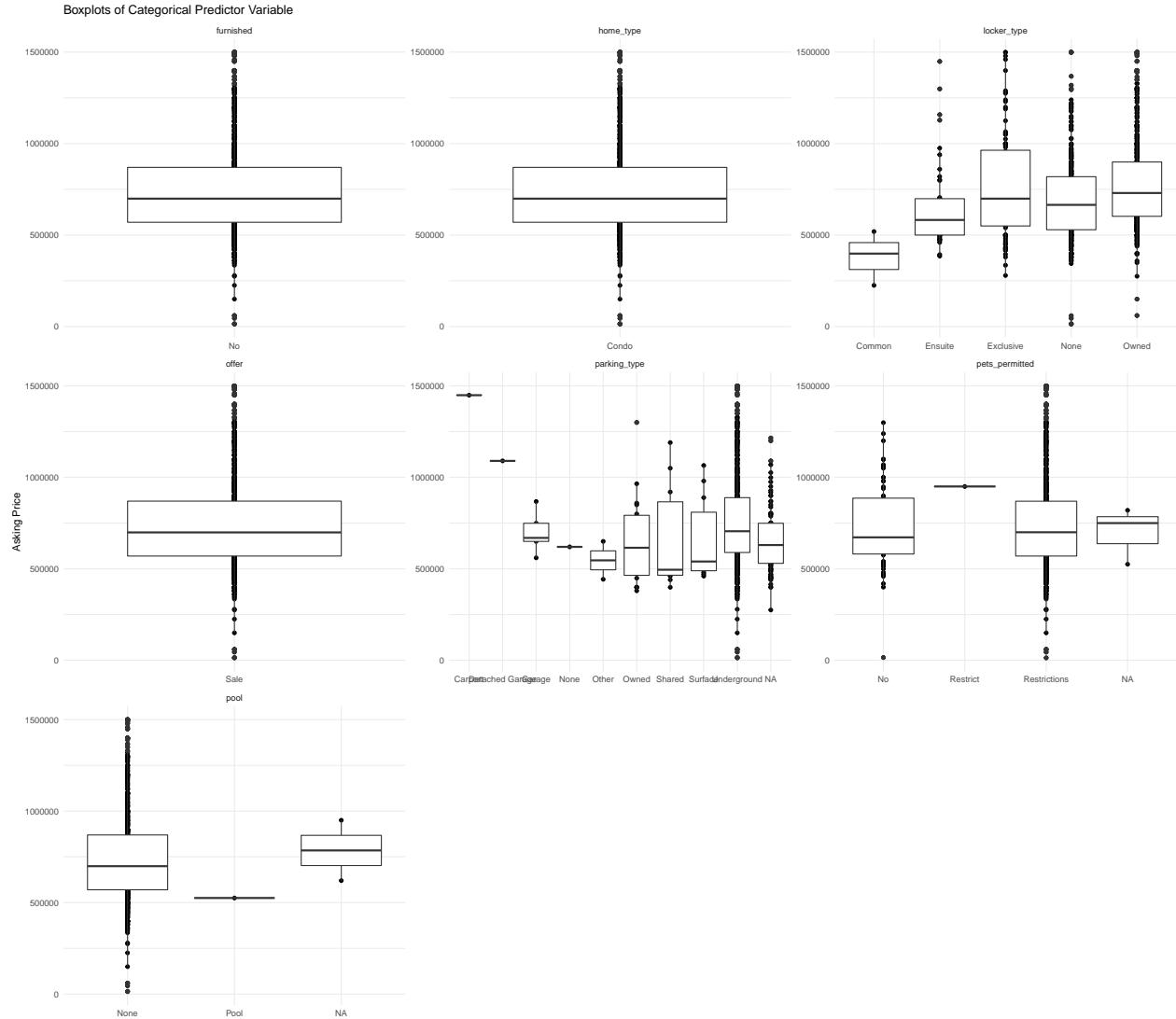


Figure 5: Boxplots of Categorical Predictor Variable

Problems of interpretation would arise when we have correlated predictors. The issue is called Multicollinearity. The initial full model we fitted has large risk of multicollinearity. So we further eliminated the variable "Square Feet Maximum" according to the VIF results as shown in Appendix 6.4.

We also rechecked our model to make sure the two Priority Condition are satisfied. Satisfying these two conditions would help us interpret our residual plots. We can use a plot of the response against the fitted values to determine whether Condition1 holds. As shown in Figure 6, the points are randomly scattered around with no significant deviations among the functions. We also rechecked Condition 2 after we made modifications to our full model. As shown in Figure 7, there are no predictor variables with highly non-linear relations. Since both conditions are satisfied, we could analyze the residual plots and identify potential limitations.

Overall, the residual plots in Figure 8 are randomly distributed with no distinguishable patterns. Admittedly, there is potential for Distance to Downtown, Bathrooms, and Photo Count to violate the constant error variance assumption as there are mild fanning patterns in their residual plots. However, our Likelihood ratio test results suggested that no transformation is needed. Moreover, our Normal QQ Plot shown in Figure 9 has good relationship without severe deviations, so our fitted model is satisfactory to proceed.

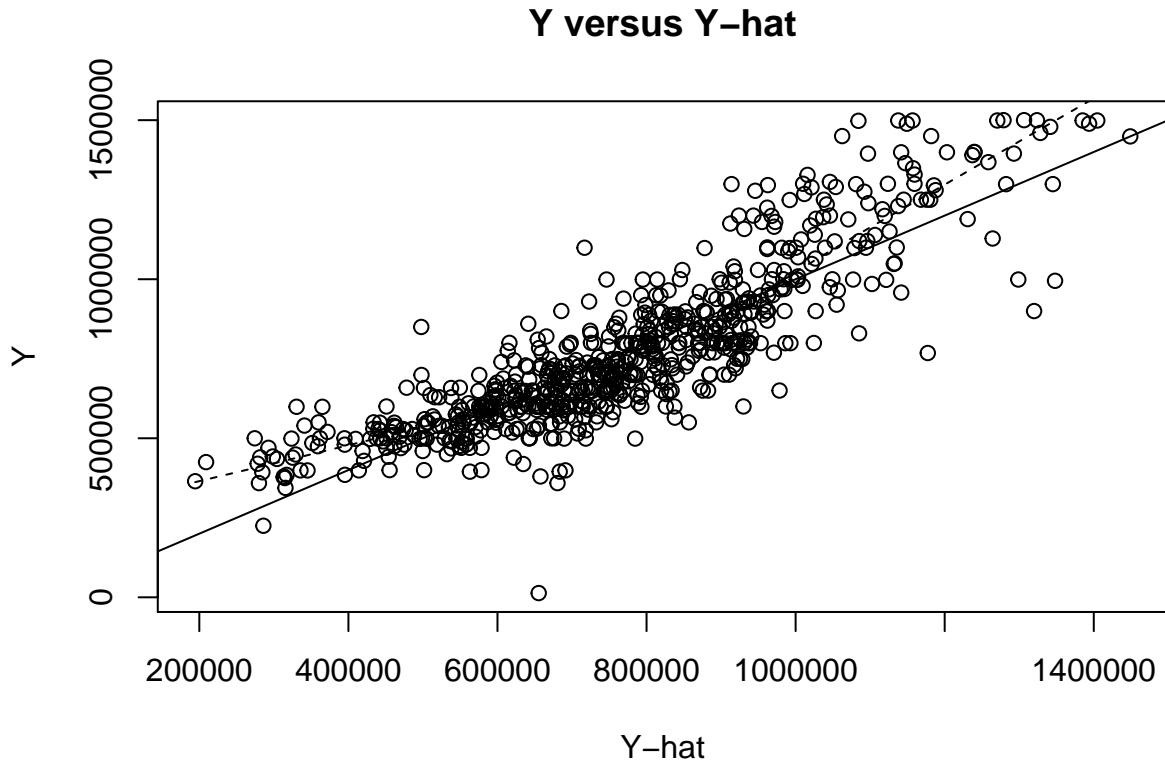


Figure 6: Condition1 - Response & Fitted Value

The final check is to visualize the effects of Leverage points, Outliers, and Influential observations. By using the measurements and cutoffs accordingly as shown in Appendix 6.2, we do have some Leverage points shown in Figure 10. However, due to the large sample size, this is not very concerning. Also, we only detected 4 Outliers as shown in Figure 11 which means there is a small potential for our trend to be largely impacted. We did have more Influential observations shown in Figure 12, but many are repeated captures with Leverage points which is a good sign as we want to have a relatively fixed group of problematic observations. However, since we do not have any contextual reason to delete these problematic observations, we would simply acknowledge their existence when discussing the limitations of our model.

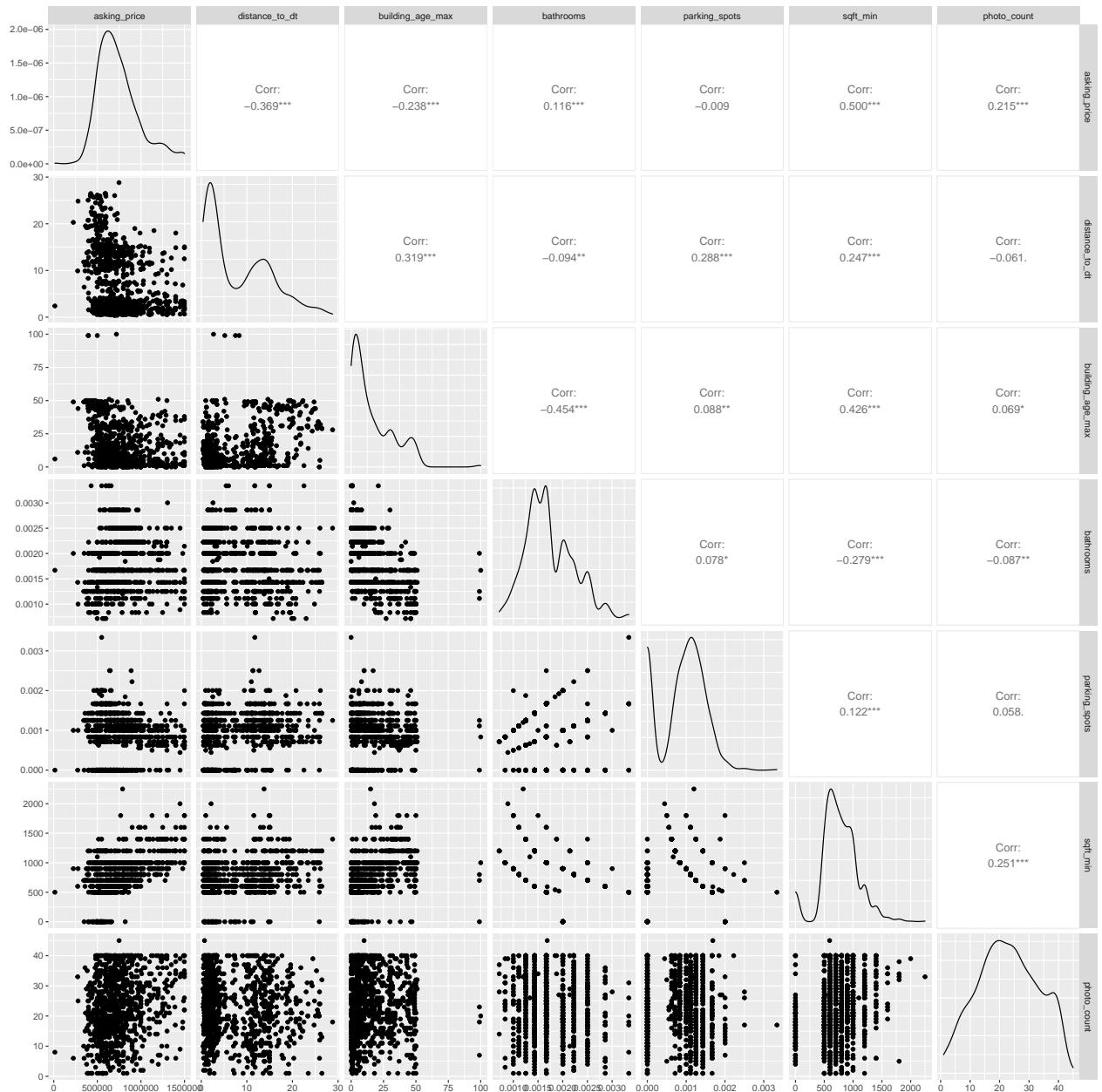


Figure 7: Condition 2 Rechecked - Relationships Among Variables

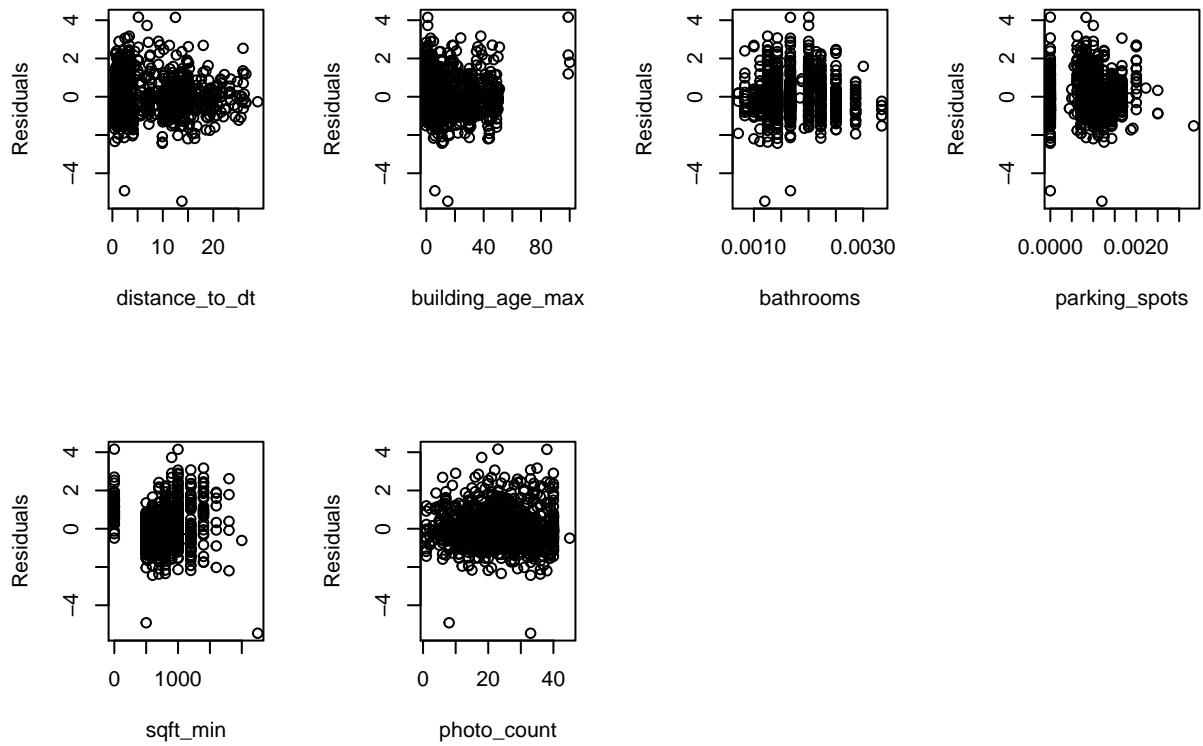


Figure 8: Residuals versus Predictor Plots

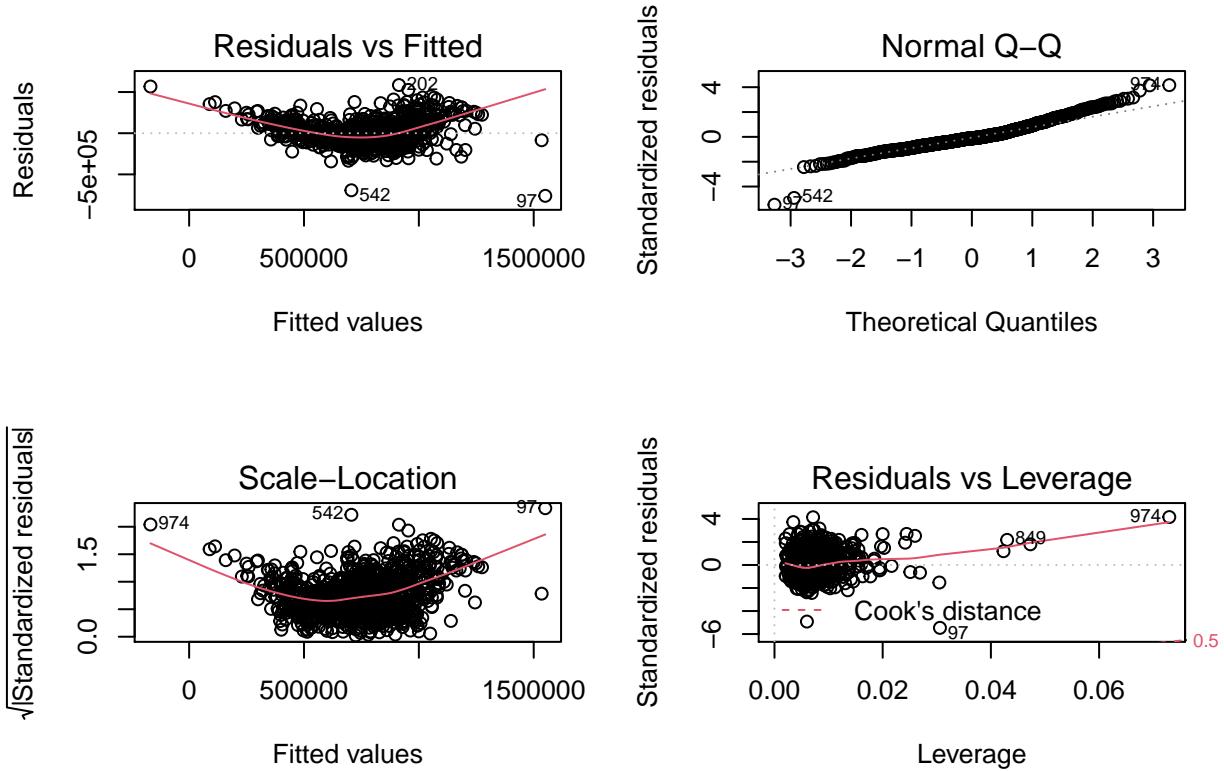


Figure 9: Normal Quantile-Quantile(QQ) Plots and Residuals versus Fitted Values Plots

The regression equation will not be useful if it is very complicated or a function of a large number of explanatory variables. So, we generally want the model to be as simple as possible to adequately describe the response variable. However, without the knowledge of the exact logical connection between the response and explanatory variables, it is unwise to think there exists an exact, discoverable equation. Many possible models will exist. To develop our preferred models, we went through the All Possible Subsets selection process and Stepwise selection to overcome the issue of changing statistical significance and account for the conditional nature of the model. We generated our first preferred model from the results of Stepwise selection using AIC as criteria. There is a consensus on the resulted model from all possible processes and stepwise selection, so we chose it as the second model. However, their strict adherence to guidelines could result in overfitting regardless of contextual information. We generated our third model which eliminated the predictor variable "Bathrooms" since we believe the area of bathrooms in a condo also depends on the square foot minimum.

As the automated selection methods would proceed without warnings of model violations or multicollinearity, we checked assumptions for each preferred model. After comparisons from the summary between the preferred model and the corresponding test model, we reached a result of choosing Preferred Model 3 as the final model focusing on the perspectives of coefficients as parameters and adjusted R squared. Our final model suggested that Toronto condo Asking Price ~ Distance to Downtown + Building Age Maximum + Square Foot Minimum with the Adjusted R^2 equal to 0.6675. It implies that our model explains approximately 66.75% of the variations in the response. In other words, distance to Downtown, maximum building age, and minimum square foot could explain 66.75% of the variations in Toronto condos' asking price.

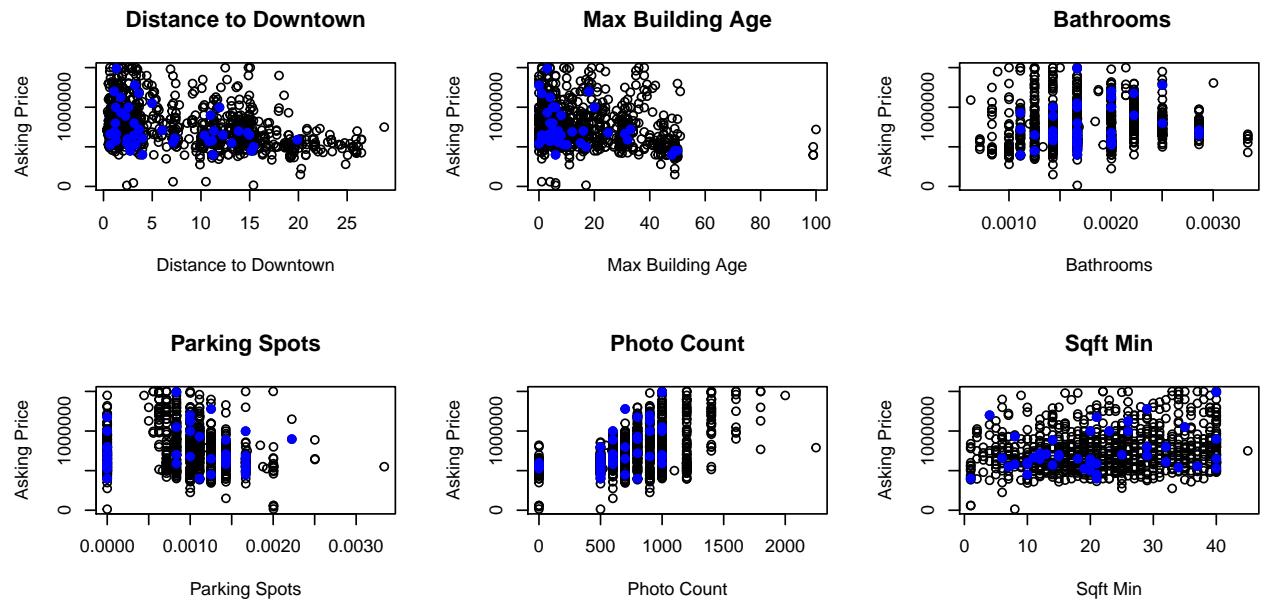


Figure 10: Leverage Points

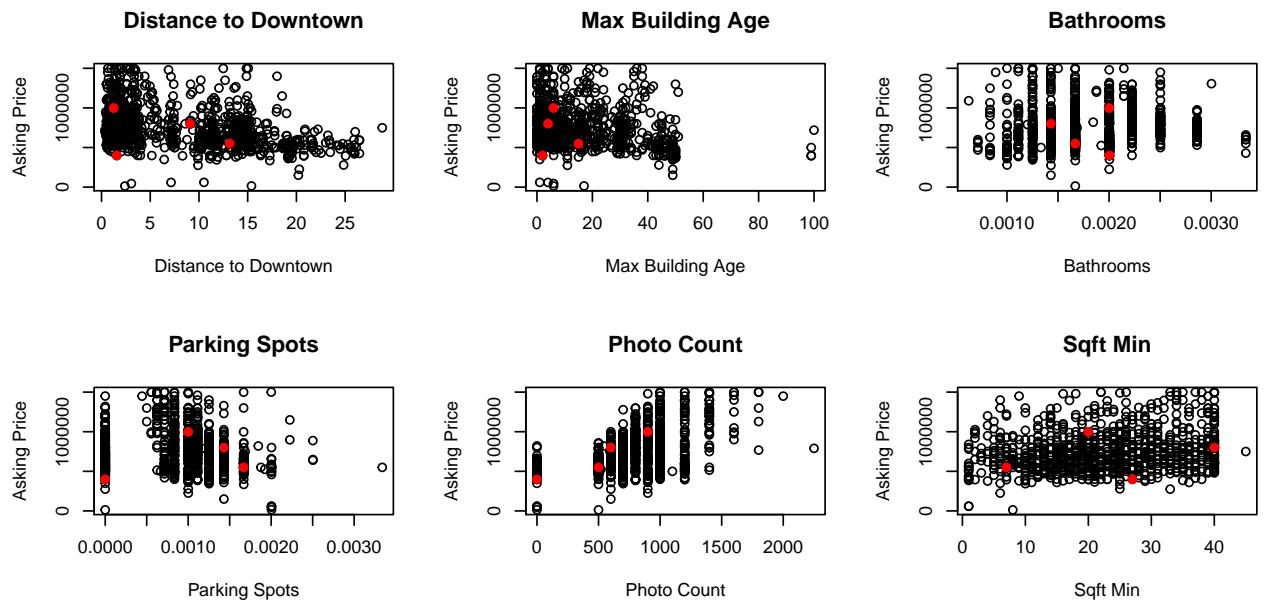


Figure 11: Outliers

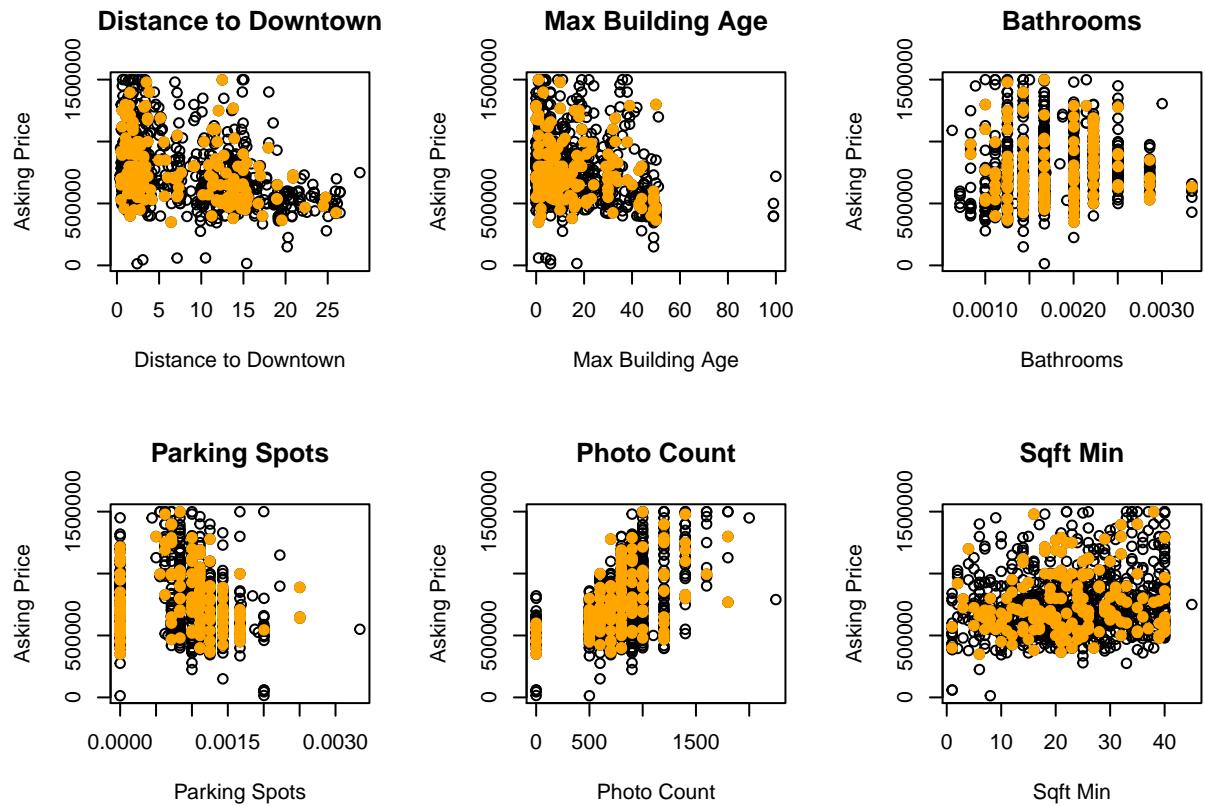


Figure 12: Influential Points

5 Discussion

5.1 Implications and Ethical Considerations

Our final model presents a reasonable estimation of the relations between Toronto condo asking price with the condo's location, building age, and area. "Distance to Downtown" has the largest absolute value of its coefficient as parameters which implied having the largest impact on the asking price. As the relationship between distance to downtown and asking price is negative, it indeed implies the significance of locations. However, there is an issue with confidentiality since the access to residence data may require voluntary participation. As we only analyzed one city, it may not be sufficient for the audience to obtain conclusions that could be applied to a broader range. It is not supportive enough to apply the discovered relationship globally.

Surprisingly, the physical area of the unit is not the most influential predictor. Building Age Maximum also has significant impact on the asking price. It is intriguing for us to find out what aspects related to building age is creating this strong momentum driving the differentiation in condo prices. It may indicate that the investors and sellers are assuming the old condo buildings are relatively less attractive than the new buildings. While we are not sure if this is because of safety concerns, it would be helpful to verify what features of old buildings are making it worth less than new buildings. If the biased perspective indeed exists, it would be crucial for both sellers and buyers to rationally modify price expectations to capture the meaningful value for themselves.

5.2 Caveats and Limitations

We built our final model with caution to avoid overfitting and realistic contextual reasons. However, our response variable has no perfect Normal distribution which may imply the inappropriateness of implementing MLR. Also, we do have some problematic observations. We have many influential points that are likely to impact our estimated regression line. However, they seem to be randomly distributed among all predictor variables with complementary directions which may neutralize their impacts. Since there are many repeated capture of Leverage points, it seems like our regression line is affected by some newly built condos that are in the center of downtown with an extremely high asking price. Importantly, even if we have these problematic observations, we should never delete them to make our model look better.

As we mentioned, the asking price, also known as the listing price, refers to how much a seller has listed a property for, and the sale price refers to the amount it actually sells for. So, the response variable we used for analysis does not reflect the final trading value of each unit. If a seller is looking for a quick sale that will generate a lot of interest, they will typically list below market value. If the seller's property is in high demand but they are still looking for a quick sale, then they may list at market value. If sellers have a firm idea of what their home is worth and are not in an immediate rush to sell, they may test out the market and list above market value. Therefore, there is a limited range of audiences to utilize our results. When considering market value, which is the current value of a home-based on local sales prices that are comparable to the subject property, people should keep in mind that many factors can affect fair market value from unit to units, such as supply, demand, renovations, bedrooms, bathrooms, lot size, house size, suites, age, and more. When determining market value, it is important that you take comparable sales in the same or nearby neighborhoods and weigh the pros and cons of each sale based on the subject property to determine an estimate of what people will pay for the property(Baerg, n.d.).

5.3 Recommendations and Next Steps

Our analysis also showed that our model explains approximately 67% of the variation in the response which addressed the first question for investors that there may be 33% of external factors like economy conditions are affecting Toronto condo prices. It could be testified if we could build a model to analyze the relation between condo price and some external factors as next step. Our results should not be used conclusively when investors are making decisions. Instead, results generated by regression model is helpful for predictions and forecasting. Our research provides a multipurpose implication on the real estate market in Toronto that could be referred to by any individual interested in this market. Moreover, real estate investment is usually

an important section of individual's portfolio management. Therefore, our analysis could also help investors arrange their portfolios. They could utilize our results to decide whether they should and how to participate in real estate investment business.

In addition, we only obtained data about the building, but there is not much information about investors and buyer's personal information like income, education, marriage status, and financial conditions which could also have essential impacts on asking prices. To fully capture the movements in the real estate market, it is recommended to investigate from the perspective of investors. We also included a complementary questionnaire to initialize the next step. Data generated through this survey could be used to analyze opinions and behavioral decisions of investors. As personal preferences play an important role when making private real estate investments, it is crucial for the researchers to develop a well-structured classification of individuals for future investigations.

6 Appendix

6.1 Appendix1

6.1.1 Model Card

Model Card - Multiple Linear Regression

Model Details

- Developed by Karl Pearson as a rigorous treatment of the mathematics of the Pearson Product Moment Correlation (PPMC).
- Galton laid the imaginative groundwork that Pearson later developed into a rigorous mathematical treatment.
- Pearson's subsequent work included further development of multiple regression as well as innovative progress on other statistics such as chi-square (Pearson 1938).
- Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors (<https://doi.org/10.1080/10691898.2001.11910537>).

Intended Use

- Quantifying relationships: a statistical procedure to predict the values of a response (dependent) variable from a collection of predictor (independent) variable values.
- Prediction.
- Forecasting: Vector autoregression and panel data regression are two powerful forecasting techniques that use the principles of multiple regression.
- Have been extensively used in education.

Factors

- In an application of linear regression, the observations are assumed to be independent.
- Linear regression models apply when the response variable can be assumed to be a continuous variable or to be normally distributed.

Metrics

- Metrics for regression involve calculating an error score to summarize the predictive skill of a model and on average how close predictions were to their expected values.
- Mean Squared Error (MSE) is a popular error metric for regression problems which is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.
- We use the coefficient of determination $R^2 = \frac{SSReg}{SST} = 1 - \frac{RSS}{SST} = \frac{\bar{y}^T(H - \frac{1}{n}J)\bar{y}}{\bar{y}^T(J_{n \times n} - \frac{1}{n}J)\bar{y}}$ to describe the percentage of variance of the response variable explained by the model. In Multiple Linear Regression Model, we use the adjusted R^2 because it is adjusted for the number of predictors in the model.

Training Data

- Training Data Set to perform all model building and diagnostics until we have decided on a final (or a few final) models.

Evaluation Data

- Test Data Set will only be used to evaluate the performance of the models.

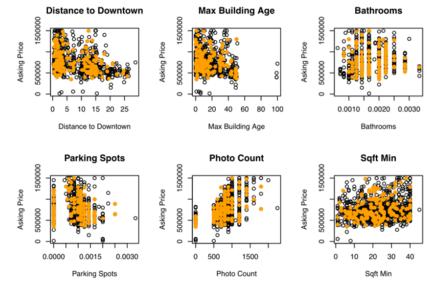
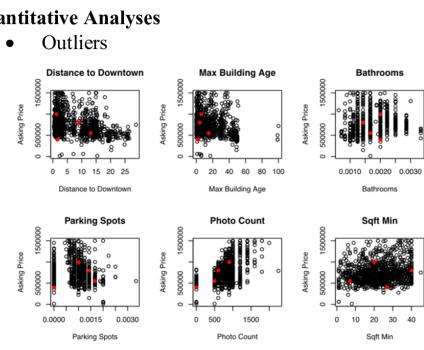
Ethical Considerations

- Real estate trading prices largely depend on geographic and economic conditions.
- Confidentiality of buyers and sellers.
- Access to listing information may require voluntary participation.

Caveats and Recommendations

- Strict assumptions required for the data.
- Able to determine the relative influence of one or more predictor variables.
- Able to identify violations and problematic observations.
- Recommended for prediction and forecasting.

Influential Points



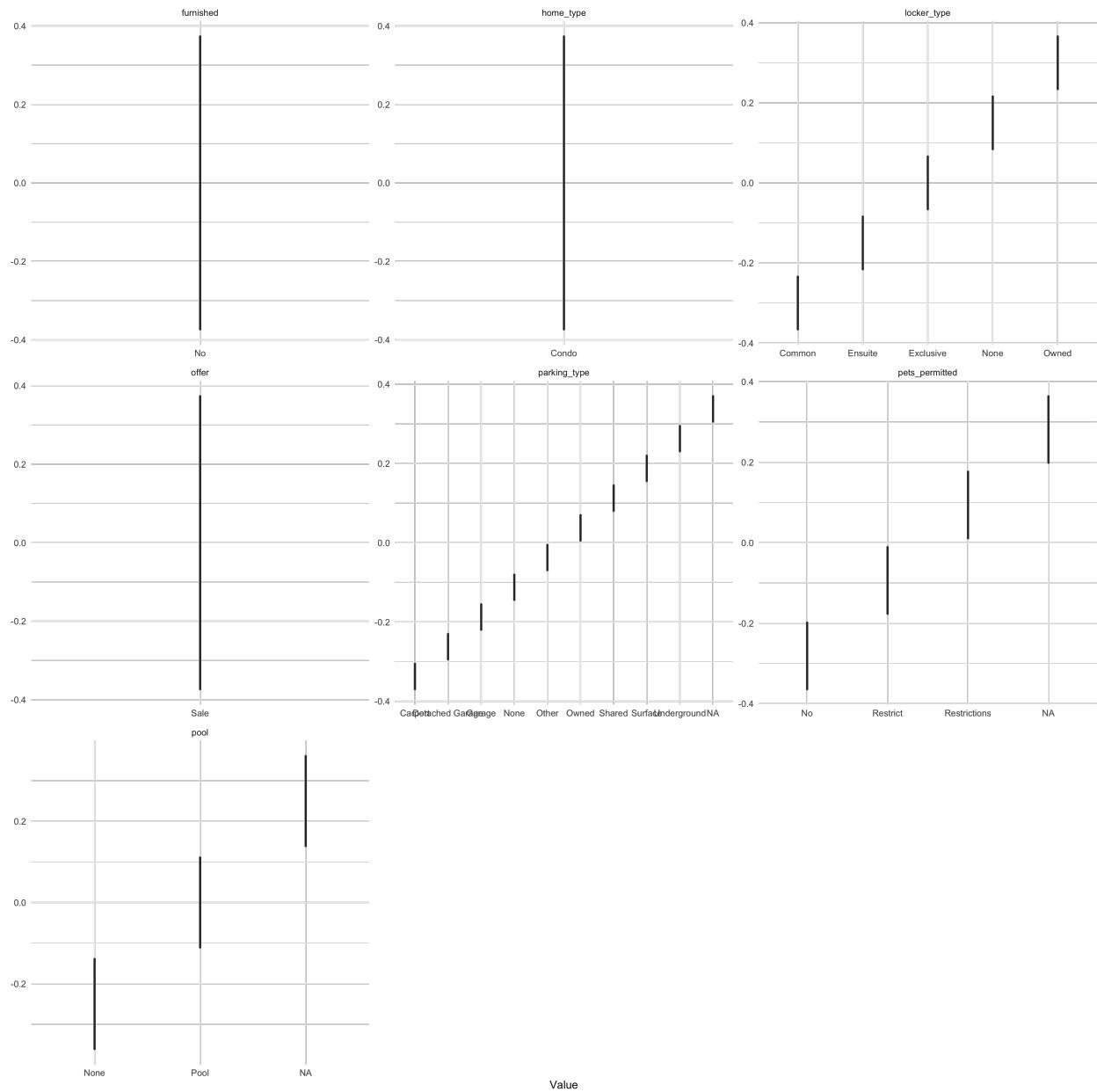
6.2 Appendix2

6.2.1 Cutoffs for Problematic Observations

Type of Point	Measure	Cutoff
Leverage	h_{ii}	$h_{ii} > 2\frac{p+1}{n}$
Outlier	r_i	$r_i \notin \begin{cases} [-2, 2] & \text{if "small" dataset} \\ [-4, 4] & \text{if "large" dataset} \end{cases}$
Influential	D_i $DFFITS_i$ $DFBETAS_{j(i)}$	$D_i > 50\text{th percentile of } F(p+1, n-p-1)$ $ DFFITS_i > 2\sqrt{\frac{p+1}{n}}$ $ DFBETAS_{j(i)} > \frac{2}{\sqrt{n}}$

6.3 Appendix3

6.3.1 Levels of Certain Predictor Variables



6.4 Appendix4

6.4.1 VIF After Modification

Comparison between Initial Fitted Full Model and Modified Fitted Full Model

Response Variable	Initial Fitted Full Model	Modified Fitted Full Model (Reducing Multicollinearity)	
	<i>Asking Price ~</i>	<i>Asking Price ~</i>	VIF (After Modification)
<i>Predictor Variables</i>	<i>Distance to Downtown</i>	<i>Distance to Downtown</i>	1.243751
	<i>Building Age Maximum</i>	<i>Building Age Maximum</i>	1.535524
	<i>Bedrooms</i>	<i>Bathrooms</i>	1.305355
	<i>Bed Type Number</i>		
	<i>Bathrooms</i>		
	<i>Parking Spots</i>	<i>Parking Spots</i>	1.118574
	<i>Parking Type</i>		
	<i>Locker Type</i>		
	<i>Square Feet Minimum</i>	<i>Square Feet Minimum</i>	1.351787
	<i>Square Feet Maximum</i>		
	<i>Square Feet</i>		
	<i>Maintenance Fee</i>		
	<i>Photo Count</i>	<i>Photo Count</i>	1.092865
	<i>Pets Permitted</i>		
	<i>Has Balcony</i>		

References

- “8 Major Reasons Why You Should Invest in Real Estate Starting Today.” 2021.%20<https://finance.yahoo.com/news/8-major-reasons-why-invest-175703808.html>.
- Alam, Mahbubul. 2020. “Multiple Regression as a Machine Learning Algorithm.” <https://towardsdatascience.com/multiple-regression-as-a-machine-learning-algorithm-a98a6b9f307b#:~:text=Multiple%20regression%20is%20a%20machine,predictions%20and%20time%20series%20forecasting>.
- Baerg, Mariko. n.d. “Asking Price Vs Selling Price: What Do They Mean and How Are They Different?” <https://bridgewellgroup.ca/asking-price-vs-selling-price/>.
- Brownlee, Jason. 2021. “Regression Metrics for Machine Learning.” <https://machinelearningmastery.com/regression-metrics-for-machine-learning/>.
- “Condos.ca.” 2021. Property.ca Inc. Brokerage.%20<https://www.condos.ca/>.
- Fortran code by Alan Miller, Thomas Lumley based on. 2020. *Leaps: Regression Subset Selection*. <https://CRAN.R-project.org/package=leaps>.
- Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.
- Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. <https://CRAN.R-project.org/package=here>.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Schloerke, Barret, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. 2021. *GGally: Extension to 'Ggplot2'*. <https://CRAN.R-project.org/package=GGally>.
- Sheather, Simon. 2009. *A Modern Approach to Regression with r*. Springer Science & Business Media.
- Sinharay, S. 2010. “An Overview of Statistics in Education.” In *International Encyclopedia of Education (Third Edition)*, edited by Penelope Peterson, Eva Baker, and Barry McGaw, Third Edition, 1–11. Oxford: Elsevier. <https://doi.org/https://doi.org/10.1016/B978-0-08-044894-7.01719-X>.
- Stanton, Jeffrey M. 2001. “Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors.” *Journal of Statistics Education* 9 (3): null. <https://doi.org/10.1080/10691898.2001.11910537>.
- Vazquez, Jorge. 2015. “6 Unique Characteristics of Real Estate Investments.” <https://realestate4investing.com/articles/real-estate-investments/6-unique-characteristics-real-estate-investments>.
- Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. <https://www.stats.ox.ac.uk/pub/MASS4/>.
- Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Xie, Yihui. 2021. “Knitr: A General-Purpose Package for Dynamic Report Generation in R.” <https://yihui.org/knitr/>.