STAT 425 Fall 2021 Final Project Report

Paul Holaway: Wrote R script as the primary coder in the group, EDA, & Appendix

Jefferson Mathews: Conclusion, Assisted in writing R script & Appendix

Rebecca Oberhelman: Introduction, Methods, & Assisted in writing R script

University of Illinois Urbana-Champaign

Introduction

For the final project, we looked at a publication by I-Cheng Yeh and Tzu-Kuang Hsu about proposing an innovative real estate valuation approach. The purpose of the publication is to solve the shortcomings of correction coefficients for traditional comparison methods. The original publication looks at four existing real estate appraisals, income approach, cost approach, comparative approach, and hedonic price approach. While each approach has its advantages, they also lack in certain areas. Fixing these problems with these four individual approaches is where the new valuation approach is proposed. The Quantitative Comparative Approach finds the price per unit area of real estate by multiplying the average price per unit area[1] and the product of adjustment coefficients of factors. The data for the original publication originates from Taiwan's real estate market. Information on each variable, or factor, is either scraped from public databases from the Ministry of the Interior from June 2012 to May 2013 from two districts in Taipei City, and two districts in New Taipei City, or calculated on their own from separate sources, such as Google Maps.

The goal of the final project is to create another model using the categorical and quantitative variables described in the original dataset to predict the house price of the unit area. The data and information for our final project are sourced from the original publication described below. To create a better-proposed model, we made a multiple linear regression and an ANCOVA model, each with and without weights. With each model we created, we performed transformations on certain variables to create optimal models. Each variable is described below in Table 1.

[1] 10,000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 squared meters.

Table 1: Variable Descriptions

Variable	Description
House_Price_Unit_Area	**(RESPONSE VARIABLE)** Predicted price of the property per unit area in meters.
Trans_Date	When the property was purchased, split into Trans_Month (Month purchased) and Trans_Year (Year purchased).
House_Age	How old the property is in years.
MRT_Dist	Distance between property and nearest MRT station in meters.
Conv_Store_No	Number of convenience stores in the living circle, 500 m, on foot.
Latitude	At what latitude the property is located.
Longitude	At what longitude the property is located.

Exploratory Data Analysis

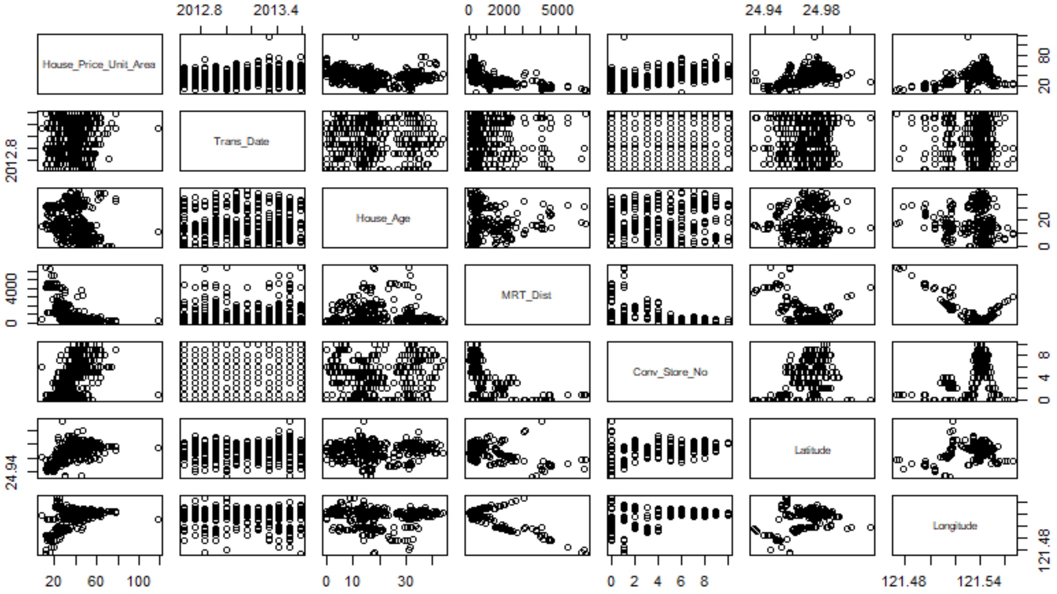
When beginning the EDA process, we began by making a scatterplot matrix of the data (See Figure 1 below). We did not include the predictors Trans_Month and Trans_Year as the former was categorical and the latter only had two possible values. Latitude and Longitude look to have a close linear relationship. However, the data points fan out for larger values warranting further testing into potential transformations. House_Age looks like a null plot at first glance so it may not be significant in the model. MRT_Dist has a shape that would signify either a log or inverse transformation. Further testing will be done for MRT_Dist as well. There does not appear to be anything significant about Trans_Date or Conv_Store_No other than a small positive relationship between Conv_Store_No and House_Price_Unit_Area. The following figures are the type (categorical or quantitative), numeric description (Table 2), and graphical description (Figure 1) of the variables.

CATEGORICAL VARIABLES	`Conv_Store_No, Trans_Month, Trans_Year`
QUANTITATIVE VARIABLES	`Trans_Date, House_Age, MRT_Dist, Latitude, Longitude,` **`House_Price_Unit_Area (Response)`**

Table 2: Descriptive Statistics

Variable	Sample Size[2]	Mean	SD	Minimum	Maximum	Skewness
House_Price _Unit_Area	414	37.98	13.61	7.60	117.50	0.60
Trans_Date	414	2013	0.28	2013	2014	-0.15
House_Age	414	17.713	11.39	0.000	43.80	0.38
MRT_Dist	414	1,083.89	1,262.11	23.38	6,488.02	1.88
Conv_Store_No	414	4.094	2.95	0.000	10	0.15
Latitude	414	24.97	0.01	24.93	25.01	-0.44
Longitude	414	121.5	0.02	121.47	121.57	-1.21
Trans_Month	414	–	–	–	–	–
Trans_Year	414	2013	0.46	2012	2013	-0.85

Figure 1: Scatter Plot Matrix of Real Estate Data



[2] Three observations deemed to be outliers were removed in the final regression model.

When looking at the histograms of the data (See Appendix: Figure 2), we can see that House_Price_Unit_Area, Latitude, and Longitude appear to have approximately normal distributions. However, only Latitude does not appear to have some kind of skew. Trans_Date and House_Age are multimodal while the histogram of MRT_Dist shows signs of an inverse of log-transformation needing to be done. Further evidence of this is looking at the skewness of the variables. Latitude and MRT_Dist have the worst skews (both having $|\text{skew}| > 1.00$) and House_Price_Unit_Area is the next closest. When looking at the frequency histograms of the data (See Appendix: Figures 4 & 5), we can observe some interesting aspects. There are large spikes in properties being bought in February and June/July. A possible reason for the February spike may be due to bonuses being given out during Chinese New Year (Mid January - Mid February). For the June/July spike, many universities in Taiwan get out in June so it would make sense as recent graduates need to find a new place to live. All these things will be kept in mind when creating the linear models.

Methods

MLR Methodology:

First we attempted a multiple linear regression model using quantitative variables. The model was sufficient, however, the R^2 was not large, at 0.5824, and there was a skew on the House_Price_Unit_Area variable. We figured that there would have to be some kind of transformation done to the variable. In multiple linear regression, R^2indicates the percentage of variation accounted for by the predictors. A higher R^2 implies a stronger relationship between the model and what we are trying to predict. We started with a transformation of the response variable because it is a skewed distribution (See Appendix: Figure 2 & 3). To alter our data so it more closely resembles a normal distribution, we concluded that the optimal transformation was a logarithmic one because, although it was not in the confidence interval for the test, it was the closest practical transformation. After fitting the new log-transformed House_Price_Unit_Area variable, the R^2 went up to 0.6858, and the original standard error dropped significantly. After that first transformation, we attempted additional log transformations on both the MRT_Dist and Longitude since those variables had large skews (See Appendix: Figure 2 & 3). After going through the transformations, we acquired a good model with all our predictors being significant, including the intercept. However, we wanted to double-check if there could be a better model. Using variable selection techniques, including backward elimination method, AIC, BIC, Adjusted R^2, and Mallow's CP, we discovered that we did not want to remove any continuous predictors. Therefore, the model with all continuous predictors, including the aforementioned transformations, was the best.

We decided to run through diagnostics, which checked for high leverage points, Cook's Distance for highly influential points, and studentized residuals for outliers. We found three observations as outliers. However, we did not have any influential points or high leverage points. It was evident that we should take out those outliers and refit the model, resulting in all variables and the intercept being highly significant, the residual standard error decreasing, and an increase in R^2. After running the same diagnostics again, there were no outliers, highly influential points, or collinearity. However, we found problems with non-constant variance, the residuals being abnormal. Due to these issues, we theorized that we would have to try weighted least squares, with weight being $\frac{1}{(Residuals)^2}$. A WLS model means that observations with higher error have less weight, and observations with lower error have more weight. Due to this, we applied the same transformation test as earlier, with the optimal transformation being a log transformation.

With this new model, we ran the same model diagnostics as before, finding no outliers, highly influential points, or collinearity. We could not solve the non-constant variance issue, but this was not a possible fix, given our time constraint. To compensate for the non-constant variance, we calculated robust standard errors. In the end, our final MLR model was a weighted least squares using all quantitative variables with their respective transformations (See Appendix: Table 5 for model coefficients and SE).

MLR Predictions:

With our two final MLR models, we used a ten k-fold cross-validation to find the prediction errors. The prediction error in the non-weighted least squares model was 0.1780 and the prediction error for the weighted least squares model was 0.1783. The prediction error between the two models is very small, which indicates that both models will perform about the same in terms of prediction. In addition, the prediction error itself is small, meaning more accurate predictions. Five predictions are listed below for both MLR models.

Table 3: Model Predictions[3]

House_Price_Unit_Area	M4_2 (MLR Model)	M4_2W0 (WLS MLR Model)
37.9	42.2	54.8
32.1	40.3	18.8
41.4	58.1	23.8
50.5	70.1	42.3
29.3	51.6	47.9

[3] In all final models a log-transformation of `House_Price_Unit_Area` was used. The results in the table are the un-transformed equivalents of the predictions. The results above are for the first five observations in the data set.

ANCOVA Methodology:

For the ANCOVA model, we started with a model using all predictors with `Conv_Store_No` and `Trans_Month` as a factor but removed `Trans_Date` because it was perfectly collinear. After looking at that model, we removed `Trans_Month` since none of the month coefficients were significant. When we used an ANOVA comparison of models test we decided that the reduced model, with `Trans_Month` removed, was adequate. We also decided to determine if a log transformation on `House_Price_Unit_Area` would be beneficial for the model and found that it was. Using the transformed model, the R^2 was higher and the `Conv_Store_No` coefficients were more significant than the non-transformed model. The problem was that we lost the significance of `Longitude`, so we tried a log-transformation of the variable. However, that did not work so we ran another ANOVA test and found that the reduced model without `Longitude` was adequate so we dropped that variable. Following the ANOVA test, we added an interaction term between the `Trans_Month` and `Trans_Year` out of curiosity and discovered it was insignificant. Only one interaction term was significant with the rest having high P-values, so we dropped the interaction. We decided to log-transform the `MRT_Dist`, as we did in the first model. For the ANCOVA model, we found that it made the variable much more significant and normally distributed (See Appendix: Figure 2). We did lose significance in the `Conv_Store_No` coefficients. After doing the log-transformation on `MRT_Dist`, we tried models with both `Longitude` and log-transformed `Longitude` and found that `Longitude` became significant after the transformation of `MRT_Dist`, so we included both log-transformed `MRT_Dist` and `Longitude` in the model. We then performed our same variable selection techniques as above, resulting in support of the model re-including the log-transformation of `Longitude`. The diagnostics gave very similar results in both models and to the MLR model as well, with the same points as before being outliers, having no highly influential points, having non-normality in our residuals, and having no collinearity. We found similar issues after running diagnostics as we had in the MLR model. We could not fix the non-constant variance but again, we did not have enough time to find a solution. We made a weighted least squares ANCOVA model, as we did for the MLR, and the results were the same. Our final decision between models would have to be decided based on prediction error (See Appendix: Table 5 for model coefficients and SE).

ANCOVA Predictions:

With our two final ANCOVA models, we used a ten k-fold cross-validation to find the prediction errors. Similar to the results from above, the prediction error between the weighted and non-weighted models, and the prediction errors themselves, was small. The prediction error in the non-weighted least squares model was 0.1753 and the prediction error for the weighted least squares model was 0.1759. In both cases, the non-weighted least squares performed slightly better; however, since it was not a large difference, we chose to include all models. It is important to note that the ANCOVA models performed better than the MLR models in terms of prediction. Five predictions are listed below for both ANCOVA models.

Table 4: Model Predictions[4]

House_Price_Unit_Area	PM14_2 (ANCOVA Model)	PM14_2W0 (WLS ANCOVA Model)
37.9	47.3	43.1
32.1	46.7	22.1
41.4	39.3	34.3
50.5	37.4	47.7
29.3	24.6	38.8

[4] In all final models a log-transformation of House_Price_Unit_Area was used. The results in the table are the un-transformed equivalents of the predictions. The results above are for the first five observations in the data set.

Conclusion

No one model performed significantly better than any other in a way that would suggest it be used as the predictive model. This is to be expected as it can be very difficult to predict something as dynamic as housing cost per unit. The final four models that provided predictions with the least amount of error were a Multiple Linear Regression model, an ANCOVA model, and a Weighted Least Squares version of each model. When comparing the MLR and ANCOVA models, there was no lack of fit for the MLR model. These models were created from a dataset that removed three outliers to improve accuracy. These models each had no autocorrelation, minimized high leverage points and used predictors with very little collinearity. This allows us to accept key assumptions required for building a successful model. Their main strength was the very small RSME, giving a reliable amount of consistency to the predictions of the models. A disadvantage to each model is that the errors are not normally distributed, according to a test that was run, and this could make our model less accurate. The main conclusions of this paper are as follows.

- All 4 models had a low predictive error with the unweighted ANCOVA model having the lowest prediction error.
- All 4 models had an adjusted R^2 value of 0.75 or greater with the weighted MLR model having the highest value
- All 4 models had heteroskedasticity and as a result of this, robust standard errors were calculated. This most likely is due to the unique nature of real estate pricing.

Using a linear regression approach, these models are among the best for minimizing prediction error. However, other approaches such as a Time-Series model could account for the shortcomings of these models and potentially predict the response variable more accurately.

Appendix

Figure 2: Histograms of Quantitative Variables

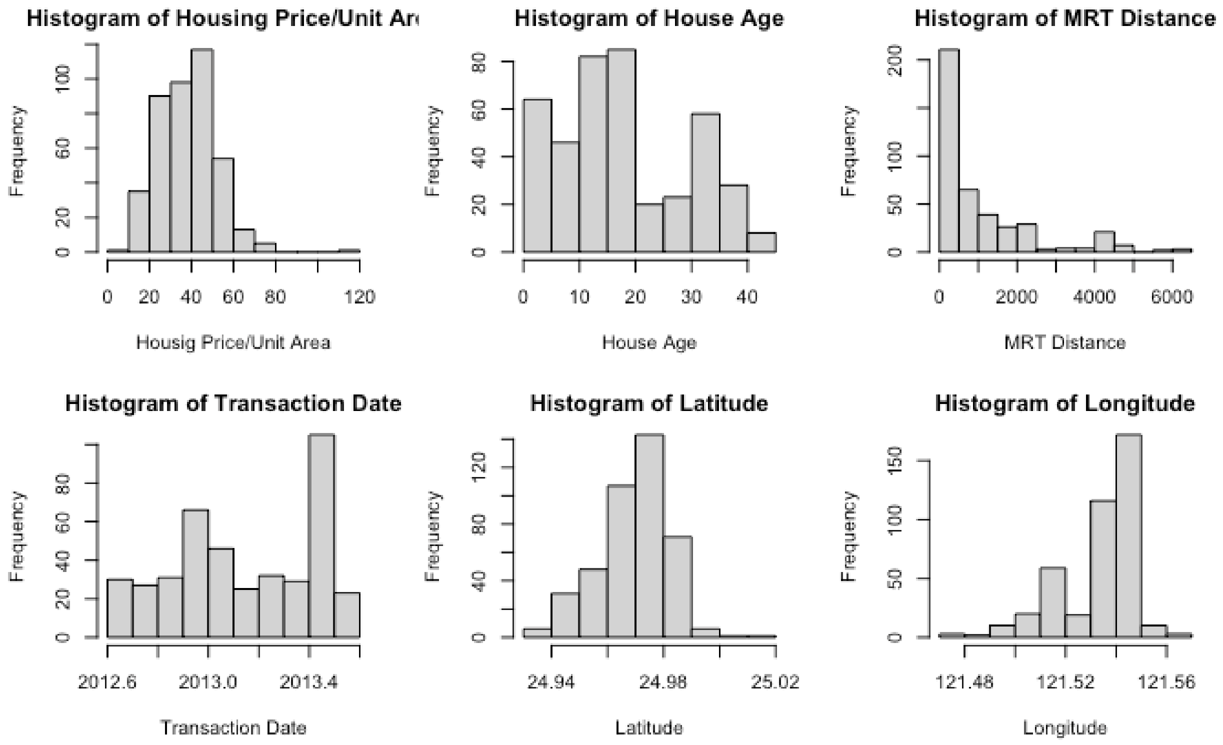


Figure 3: Histograms of log-transformed Quantitative Variables

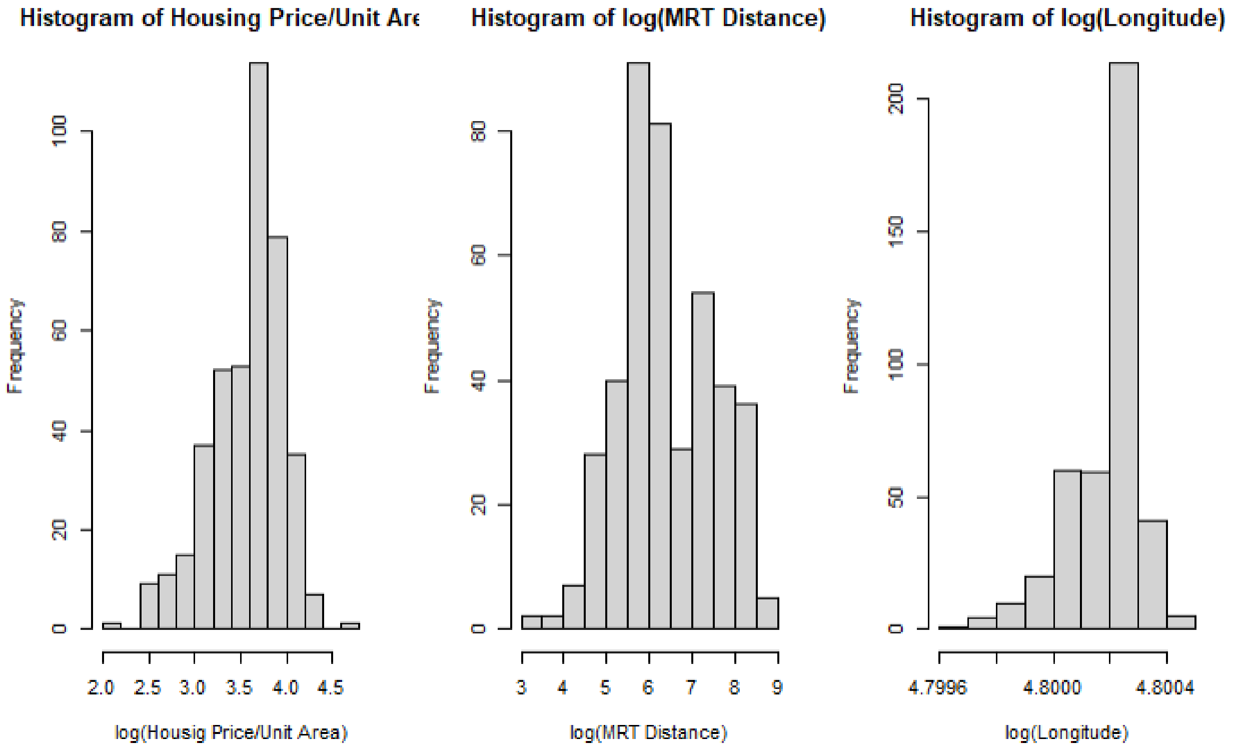


Figure 4: Frequency Histogram for `Trans_Month`

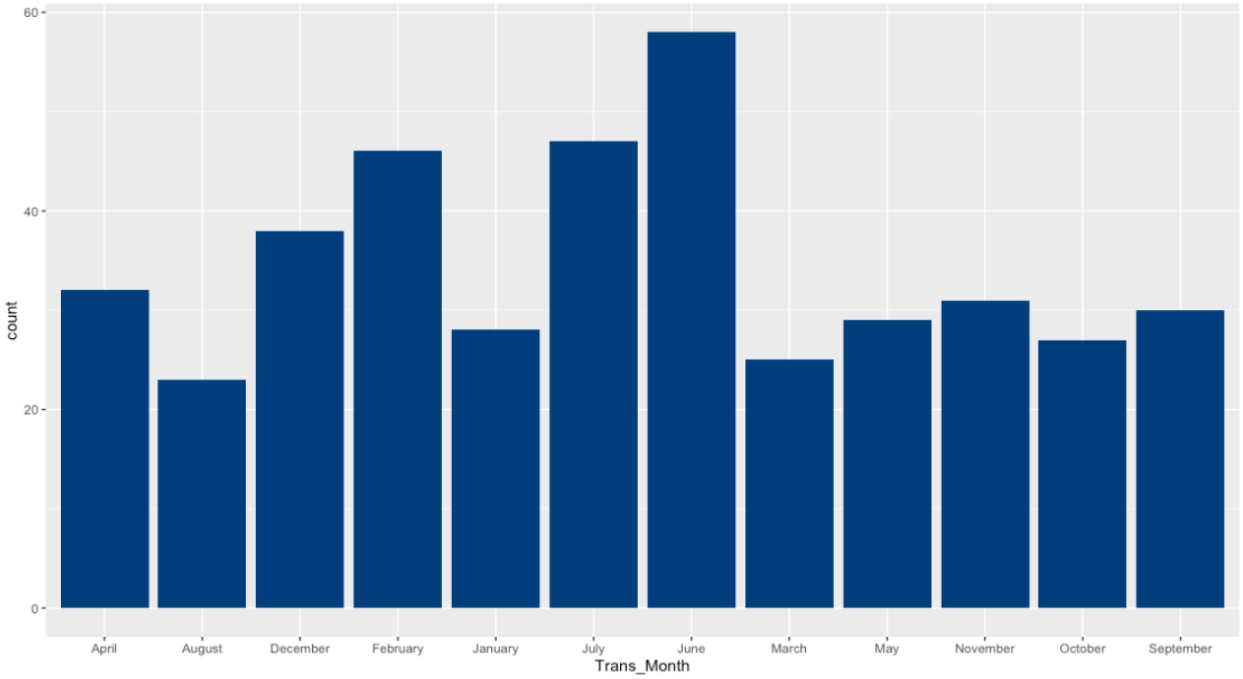


Figure 5: Frequency Histogram for `Conv_Store_No`

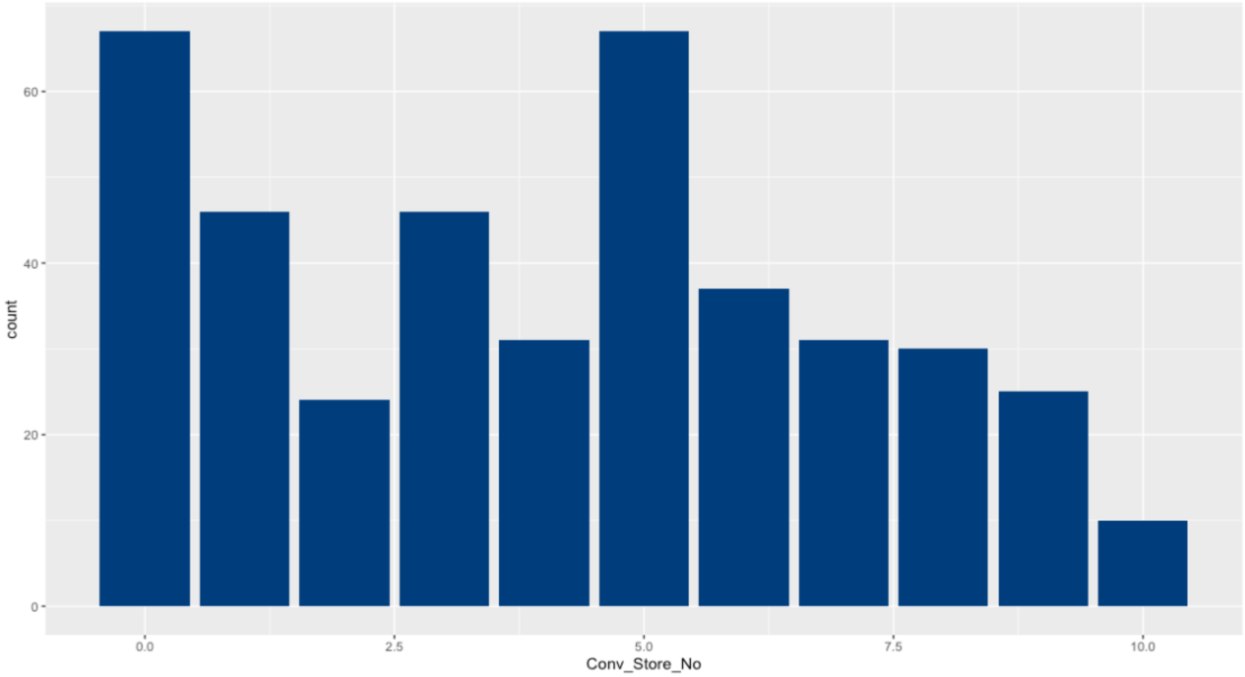


Table 5: Model Summaries

Predictor	M4_2 (MLR)	M4_2W0 (WLS MLR)	PM14_2 (ANCOVA)	PM14_2W0 (WLS ANCOVA)
(Intercept)	-2,267.0 (449.9)[5] (420.6)[6]	-2530.2 (480.7) (440.5)	-1689.3 (461.4) (452.92)	-1,948.1 (519.0) (467.97)
Trans_Date	0.1564 (0.0319) (0.0314)	0.1551 (0.0315) (0.0307)	-	-
House_Age	-0.0060 (0.0008) (0.0009)	-0.00604 (0.0007) (0.0009)	-0.00592 (0.0008) (0.0009)	-0.0059 (0.0008) (0.0009)
log(MRT_Dist)	-0.0166 (0.0314) (0.0134)	-0.1609 (0.0124) (0.0137)	-0.1683 (0.01376) (0.01458)	-0.1589 () (0.0144)
Conv_Store_No	0.01458 (0.0043) (0.0047)	0.01336 (0.0041) (0.0046)	-	-
1[7]	-	-	0.0287 (0.0354) (0.0405)	0.0122 (0.0389) (0.0417)
2	-	-	0.0514 (0.0445) (0.0418)	0.0461 (0.0472) (0.0446)
3	-	-	0.0794 (0.0358) (0.0416)	0.0731 (0.0391) (0.0435)
4	-	-	0.0993 (0.0424) (0.0476)	0.0811 (0.0428) (0.0488)

[5] The first number in parentheses under the coefficient is the standard error of the coefficient estimate. The number underneath in parentheses is the robust standard error.

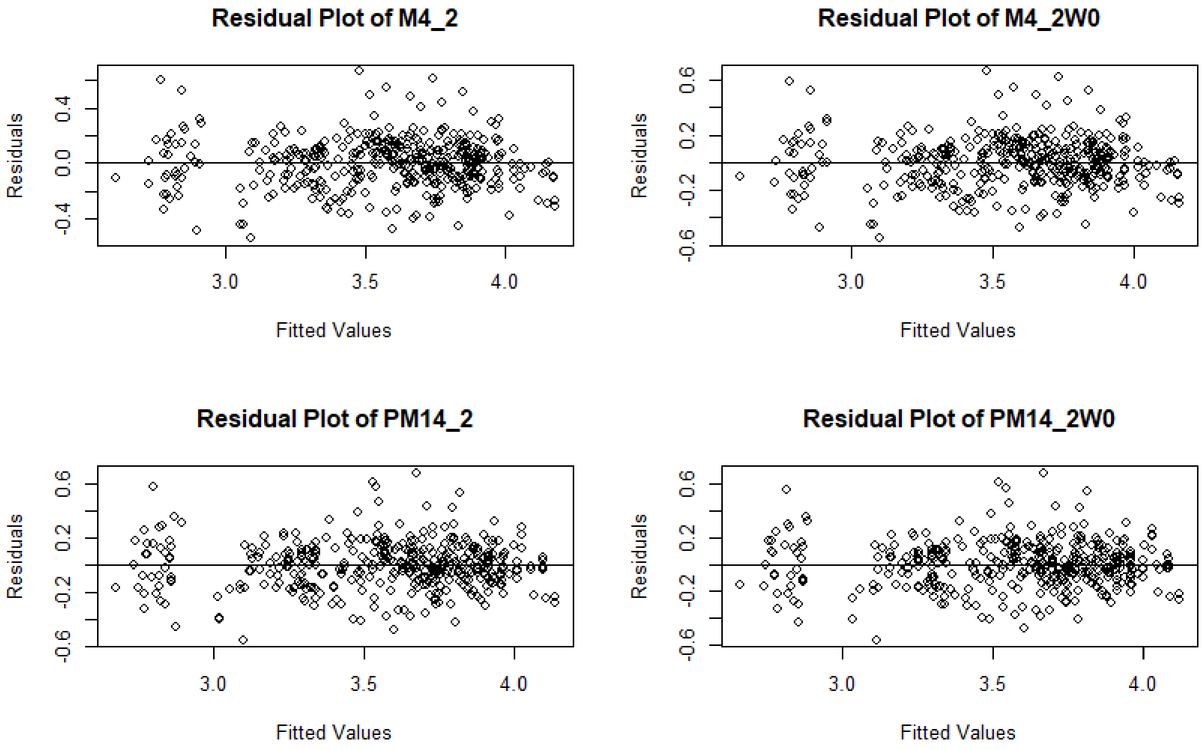
[6] Robust Standard error calculations used due to heteroskedasticity in the models. Attempts were made to remove it but were unsuccessful.

[7] Conv_Store_No ran as a factor for ANCOVA models.

5	-	-	0.1272 (0.0374) (0.0413)	0.1199 (0.0372) (0.0412)
6	-	-	0.1867 (0.0432) (0.0447)	0.1859 (0.0412) (0.0451)
7	-	-	0.0882 (0.0466) (0.0475)	0.0779 (0.0453) (0.0467)
8	-	-	0.1601 (0.0468) (0.0499)	0.1527 (0.0453) (0.0489)
9	-	-	0.1379 (0.0507) (0.0616)	0.1244 (0.0452) (0.0579)
10	-	-	0.0081 (0.0679) (0.0678)	0.0670 (0.613) (0.0679)
Latitude	10.1340 (0.8427) (1.0565)	9.9829 (0.8486) (0.9974)	10.1830 (0.9101) (1.1936)	10.1000 (0.9004) (1.091)
log(Longitude)	354.91 (93.71) (88.45)	411.11 (100.4) (91.981)	299.91 (96.87) (95.502)	354.30 (108.9) (98.647)
Trans_Year(2013)	-	-	0.0804 (0.0194) (0.0188)	0.0859 (0.0189) (0.0182)
Sample Size	411[8]	411	411	411
Adjusted R^2	0.7783	0.7795	0.7540	0.7494

[8] Three observations were removed when making the models.

Figure 6: Residual Plots of Final Models


Residual Plot of M4_2
Residuals
Fitted Values
Residual Plot of M4_2W0
Residuals
Fitted Values
Residual Plot of PM14_2
Residuals
Fitted Values
Residual Plot of PM14_2W0
Residuals
Fitted Values