STA302 Project part 1

Hilary Zou, Luka Tosic, Henry Zhang

2024-12-06

Introduction

Understanding the variables influencing rental pricing in Toronto is essential since the city's competitive real estate market and affordability concerns are escalating. This study investigates the effects of time on the market and property attributes (such as the number of bedrooms, number of bathrooms, building type, and pet friendliness) on rental pricing. This study attempts to answer "How do property features and time on the market influence rental prices in Toronto?". The findings will help regulators, landlords, and tenants understand pricing trends and help them benefit.

The relevancy of the study is supported by the three journal articles we found. Muraleedharan (2019) demonstrates how pet-friendly legislation can increase a property's attractiveness and value in his article. Although they only looked at property sales, Krashinsky and Milne (1987) emphasized the effect of time on real estate market demand. Pi (2017) showed how facilities like parking can significantly impact rental costs. Combining these results, this study uses the Toronto rental data set to examine the various effects on rental pricing.

Linear regression works best for our research because it can quantify the relationship between rental prices and a number of predictors. By employing this methodology, the study fills the gap in the current research by examining a combination of factors—property features and time on the market—that have not been fully explored together in the context of Toronto's rental market and provides a way to better understand Toronto's complex rental market.

The following sections detail the process of arriving at a final descriptive model.

Methods

Model Creation

First, we start with two candidate models model1 and model2 for the response variable price. The rationale behind starting with two candidate models initially is that we are trying to create detailed desciptive models with predictors that are contextually meaningful. To that end, we had a broad idea of what predictors we were interested in already, and we wanted to avoid going through the process of using the All Possible Subsets method or Automated Selection Methods as there were some potential predictors that were not interested in either because of missing data or due to lack of relevance. Moreover, of the predictors we are interested in, there were several models that the ASP or automated selection methods would have fit that would have been missing key predictors. In the end, we decided that only the number of bathrooms and rental type were predictors we felt could be left out of a model, so we fit one with one of each. A final note to be expanded on in the limitations section is that there are other potentially strong predictors in the dataset, but we were looking to avoid overfitting individual models, and did have the capacity to compare more models at this stage.

Model Diagnostics

Residuals

Before these models are compared, several diagnostics are run on them individually to determine their overall quality and whether they should be refit or transformed in any way. The first is to create residual plots for both models, those being residuals vs fitted, residuals, vs each predictor, and a quantile-quantile plot. The residuals are the sample errors given by $\hat{e}_i = y_i - \hat{\mathbb{E}}[y_i|x_i]$. The residual plots are created to assess the assumptions on linear regression that are made implicitly any time a model is fit. If we are to have meaningful results from our model, we want to make sure that our model follows these assumptions as well as possible.

We plot the residual plots described above and analyze them for the following patterns: any systematic patterns can point to violations of all but normality, any functional relationships of predictors pointing to violations of linearity, clustering patterns pointing to violations of uncorrelated errors, fanning patterns pointing to constant variance violations, and significant deviations from the linear trend in the qq plot pointing to normality violations.

Additionally, we analyzed the residual plots and predictor vs predictor plots for Conditional Mean Response and Predictor violations.

Multicollinearity and Significance

We also analyzed the VIF of both models for their numerical predictors to determine the multicollinearity of the predictors. No refitting was done based on this.

As well, we performed an ANOVA test for both models to test their overall significance, and since we would not be removing any predictors, we did not create hypothesis tests on the significance of individual predictors, or a Partial - F test.

Mitigating Violations

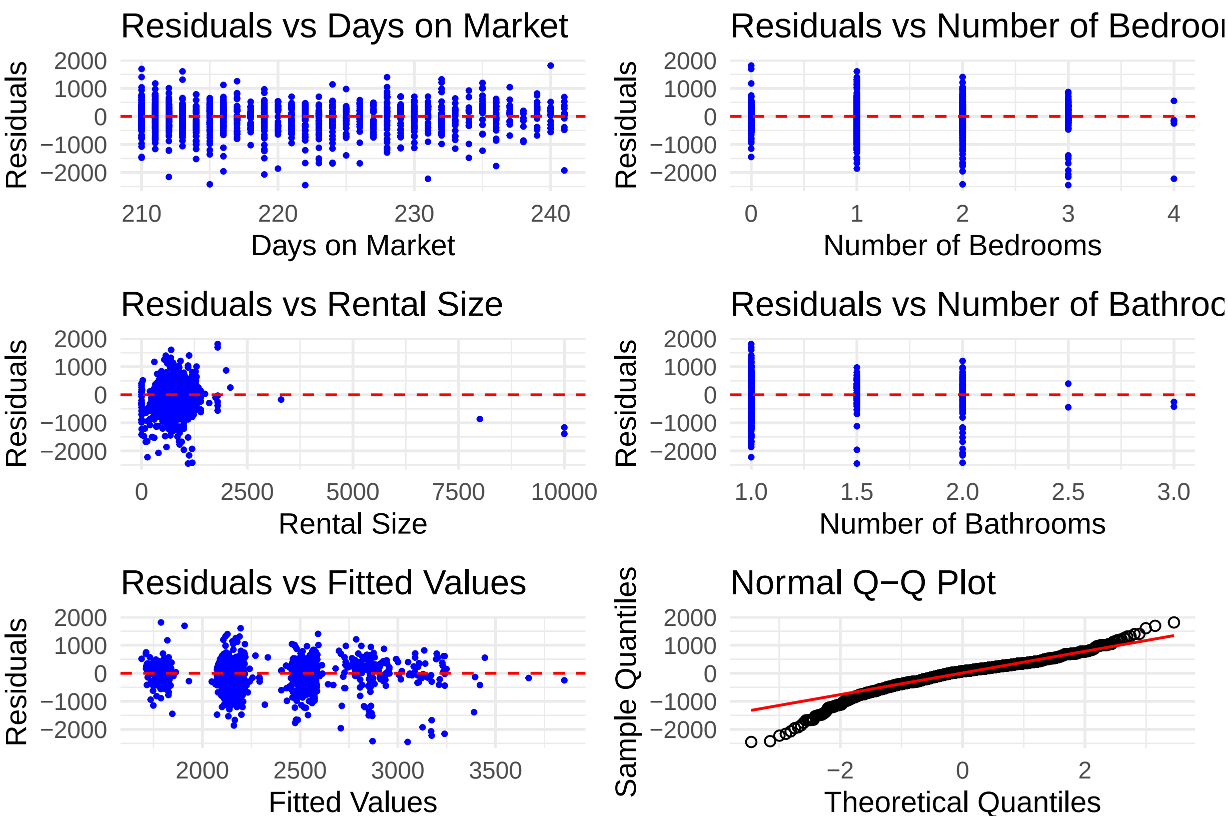
As will be discussed in the results section below, both models saw severe violations of uncorrelated errors, for which there is no applicable method to mitigate the impact of. Additionally, some moderate violation of normality was observed for both, so a box-cox transformation would be applied to both models individually.

Model Comparison

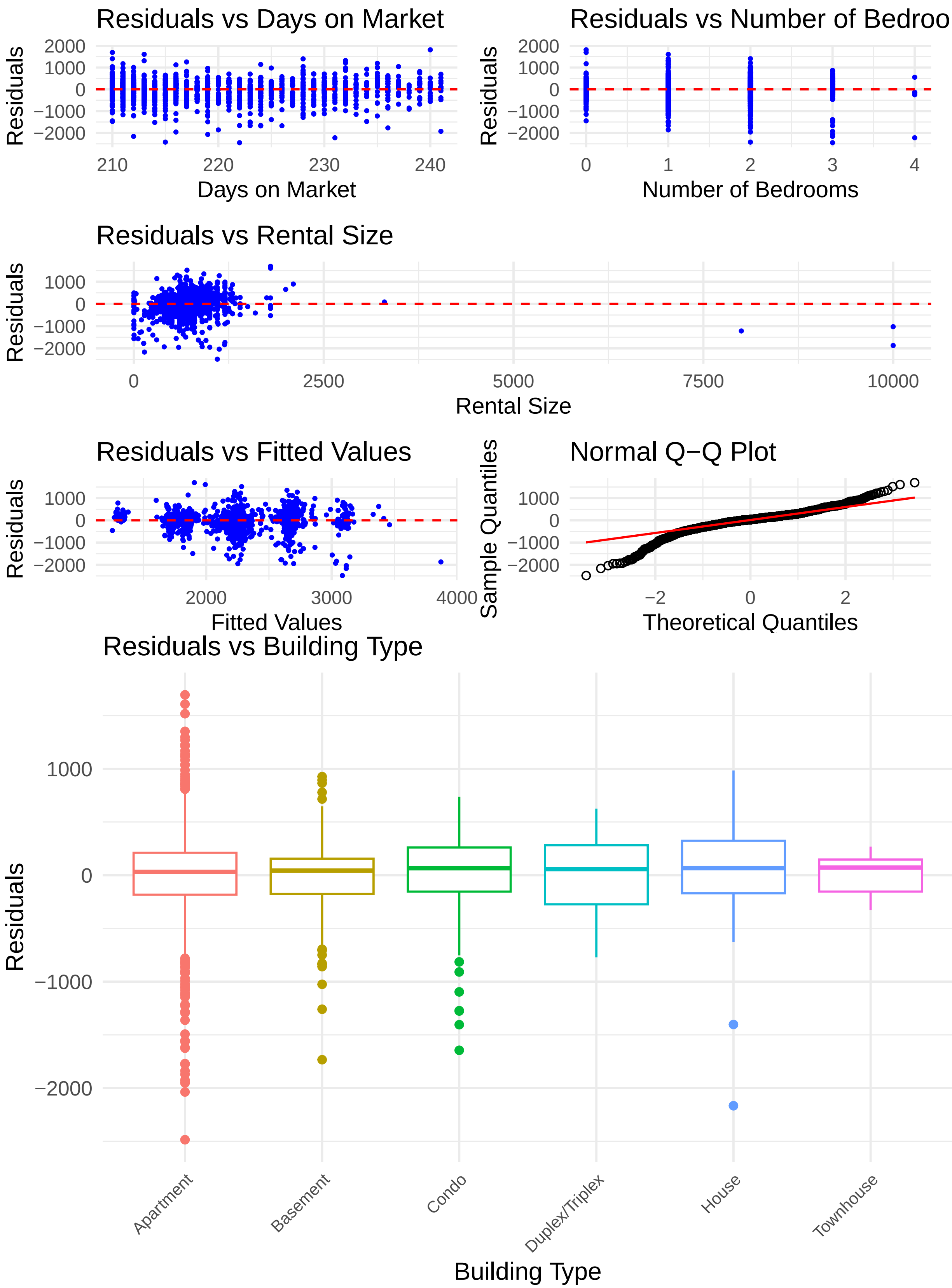
We are looking to choose only one of the two models we initially fit, and after running diagnostics to check for major issues with either model, we compare them directly via numerical measures of goodness. The ones we chose are the adjusted R squared R^2_{adj}, Akaike's Information Criteria AIC, and Bayesian Information Criteria BIC. We felt that these would be sufficient for us to be able to decide on one model.

Results

Residuals of Model1

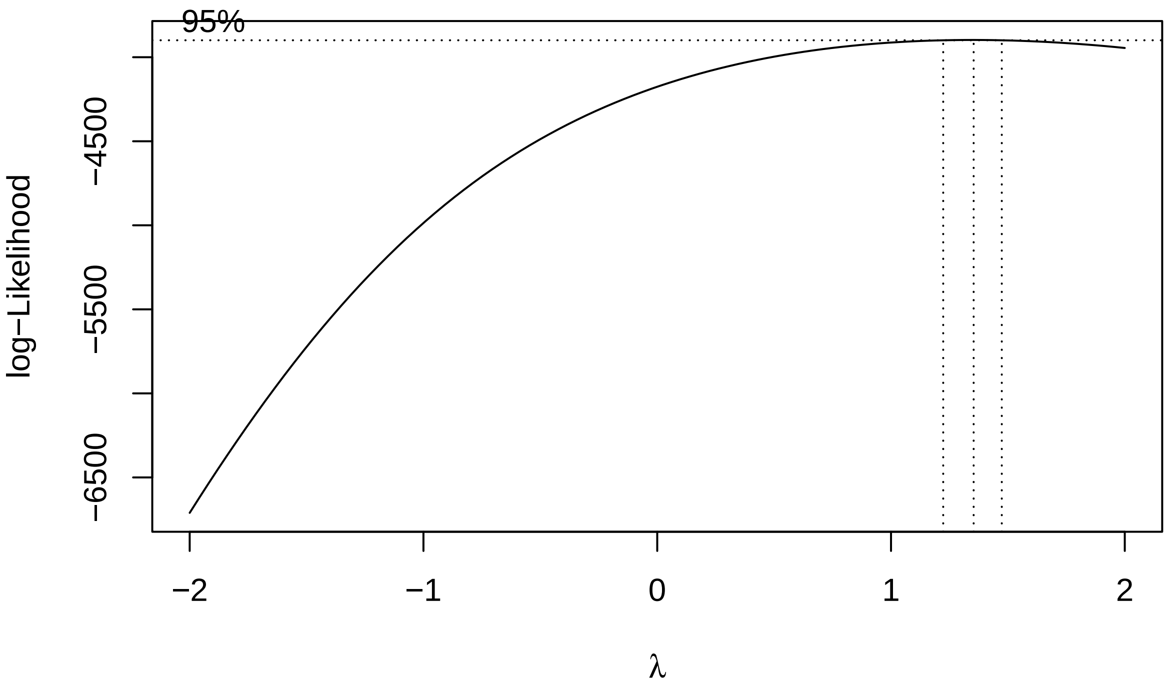


Residuals of Model2



1. Residuals vs. Fitted Values In the plot, residuals are plotted against the fitted values . The residuals do not show a clear pattern but cluster slightly, suggesting the linearity assumption is fitted.
2. Histogram of Residuals The histogram shows the distribution of residuals. While the residuals appear close to normal, small deviations affect confidence intervals and p-values. If deviations from normality are significant, we can apply a Box-Cox to help with analysis.
3. Normal Q-Q Plot The residuals mostly follow the reference line, but there are slight deviations at the tails. This suggests that the residuals are approximately normal, but the deviations at the tails might indicate some outliers.

Violation Correcting



##

Model Selection Criteria

```
##
## Call:
## lm(formula = Price ~ Days_on_Market + Number_of_Bedrooms + Number_of_Bathrooms +
##     Rental_Size, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2450.13  -249.25    70.87   271.00  1811.84
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2356.6545   271.2164   8.689  < 2e-16 ***
## Days_on_Market        -4.0176     1.2386  -3.244   0.0012 **
## Number_of_Bedrooms   360.5152    15.6104  23.095  < 2e-16 ***
## Number_of_Bathrooms  294.7471    42.1571   6.992  3.8e-12 ***
## Rental_Size            0.0561     0.0259   2.166   0.0305 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 449.1 on 1815 degrees of freedom
```

```
## Multiple R-squared:  0.3504, Adjusted R-squared:  0.349
## F-statistic: 244.8 on 4 and 1815 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = Price ~ Days_on_Market + Number_of_Bedrooms + Building_Type +
##     Rental_Size, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2485.26  -181.61    33.86   212.10  1694.57
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2442.99103  250.15709   9.766  < 2e-16 ***
## Days_on_Market               -2.90905    1.14564  -2.539 0.011193 *
## Number_of_Bedrooms          396.64347   13.58128  29.205  < 2e-16 ***
## Building_TypeBasement      -534.19880   27.65160 -19.319  < 2e-16 ***
## Building_TypeCondo           84.76074   40.92452   2.071 0.038486 *
## Building_TypeDuplex/Triplex -127.05225  125.23295  -1.015 0.310467
## Building_TypeHouse         -253.98219   66.79910  -3.802 0.000148 ***
## Building_TypeTownhouse     -184.32283  156.64821  -1.177 0.239483
## Rental_Size                   0.08923    0.02389   3.736 0.000193 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 412.5 on 1811 degrees of freedom
## Multiple R-squared:  0.4533, Adjusted R-squared:  0.4509
## F-statistic: 187.7 on 8 and 1811 DF,  p-value: < 2.2e-16

## [1] "ANOVA Test Results:"

## Analysis of Variance Table
##
## Model 1: Price ~ Days_on_Market + Number_of_Bedrooms + Number_of_Bathrooms +
##     Rental_Size
## Model 2: Price ~ Days_on_Market + Number_of_Bedrooms + Building_Type +
##     Rental_Size
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1   1815 366123959
## 2   1811 308119829  4  58004130 85.231 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] "VIF for Model 1:"

##      Days_on_Market  Number_of_Bedrooms Number_of_Bathrooms         Rental_Size
##            1.012598            1.348865            1.171498            1.176335

## [1] "VIF for Model 2:"

##                        GVIF Df GVIF^(1/(2*Df))
## Days_on_Market     1.027165  1        1.013491
## Number_of_Bedrooms 1.210527  1        1.100240
## Building_Type      1.054538  5        1.005324
## Rental_Size        1.185961  1        1.089018
```

Ethics Discussion

Conclusion and Limitations