Effect of Location and Rooms on Toronto Apartment Rental Price

STA 130 Final Project

Mimis Chlympatsos, Daniel Du

What is the goal of our project?

Our project aims to analyze the relationship between Toronto apartment rental prices and key factors such as the number of rooms and distance from a central location. We wanted to develop a predictive model that can accurately forecast rental prices based on the variables in our dataset. To achieve this, we constructed a multivariate linear regression model using our factors.

To gain further insight into the significance of these factors, we generated graphs which illustrate both the relative importance of the different predictors in predicting apartment rental prices and the extent of their impact.

By providing insights and a predictive model, our project seeks to inform and aid decision-making regarding the Toronto rental market, especially for renters.

What are our project questions?

Question 1: How can information on the number of bedrooms/dens, bathrooms, latitude, and longitude be used to predict the rental price of an apartment?

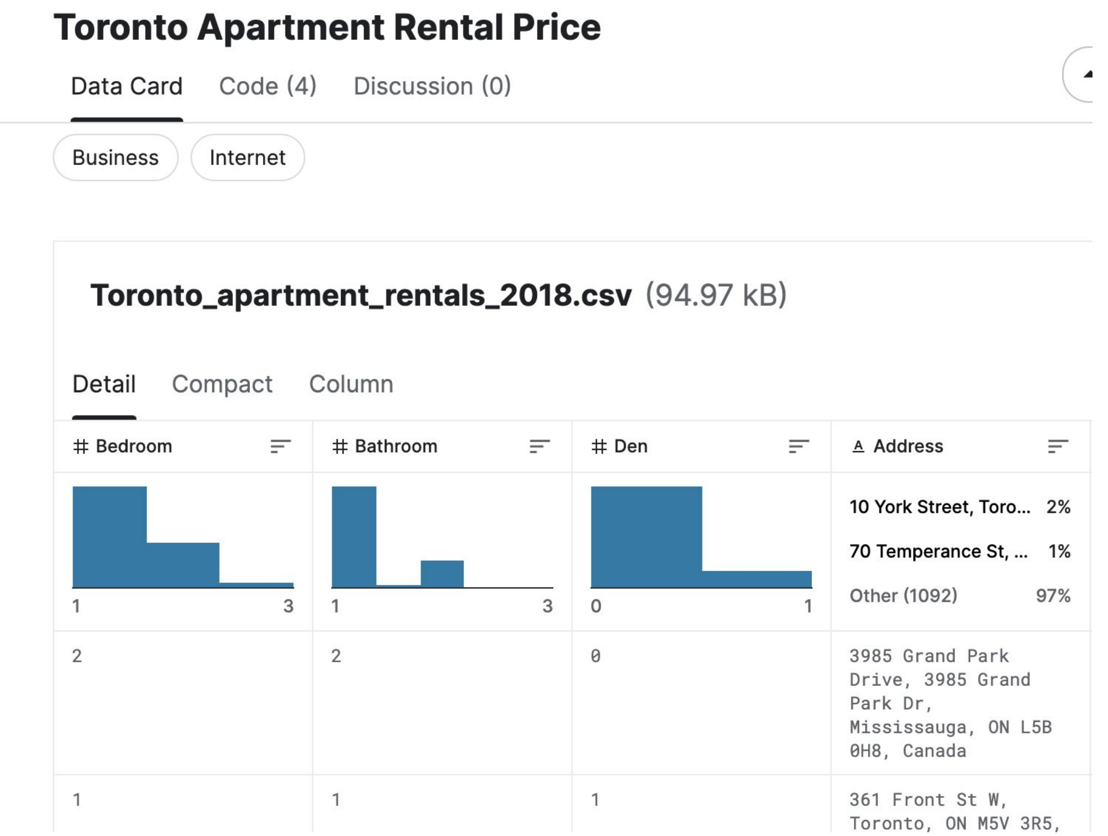
We will perform multivariable linear regression to create a model that seeks to answer this question.

Question 2: How strong is the effect of each explanatory variable (overall) for the rental price prediction, and how do these effects compare relative to each other?

To answer this question, we will determine, appropriately visualize, and analyze the effect sizes and distributions of our explanatory variables.

What data did we use?

We used a collection of Toronto apartment rental prices gathered from Kijiji in 2018. In total, there was data for 1124 apartments. The data had 5 variables which were utilized in our model - the number of bedrooms, bathrooms, dens, and coordinates for latitude and longitude - and the price of the apartment, which is what we intended to predict.



Toronto Apartment Rental Price

Data Card Code (4) Discussion (0)

Business Internet

Toronto_apartment_rentals_2018.csv (94.97 kB)

Detail Compact Column

# Bedroom	# Bathroom	# Den	Address
1 3	1 3	0 1	10 York Street, Toro... 2% 70 Temperance St, ... 1% Other (1092) 97%
2	2	0	3985 Grand Park Drive, 3985 Grand Park Dr, Mississauga, ON L5B 0H8, Canada
1	1	1	361 Front St W, Toronto, ON M5V 3R5,

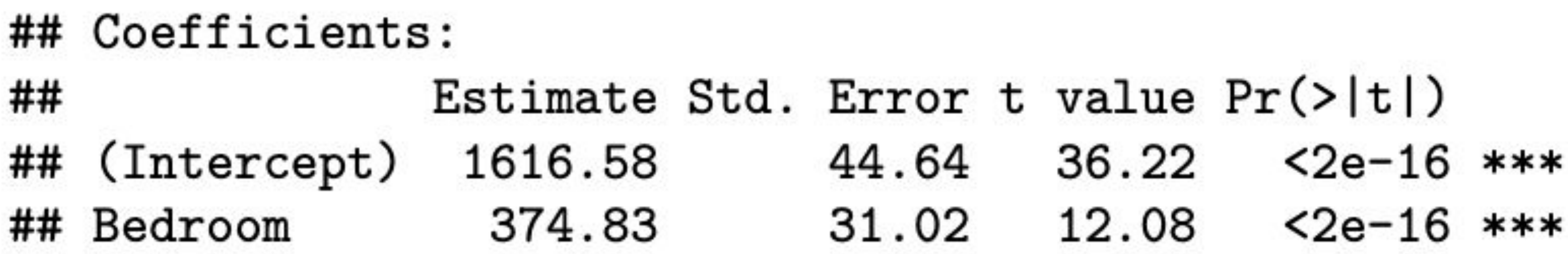
Cleaning: Prices for some observations were below $80/month. These were removed as they are extreme outliers, and most likely were faulty observations. We also decided to remove the observations with more than 2 bathrooms, as there were only 8 observations in total.

Data - Creating Variables for Model

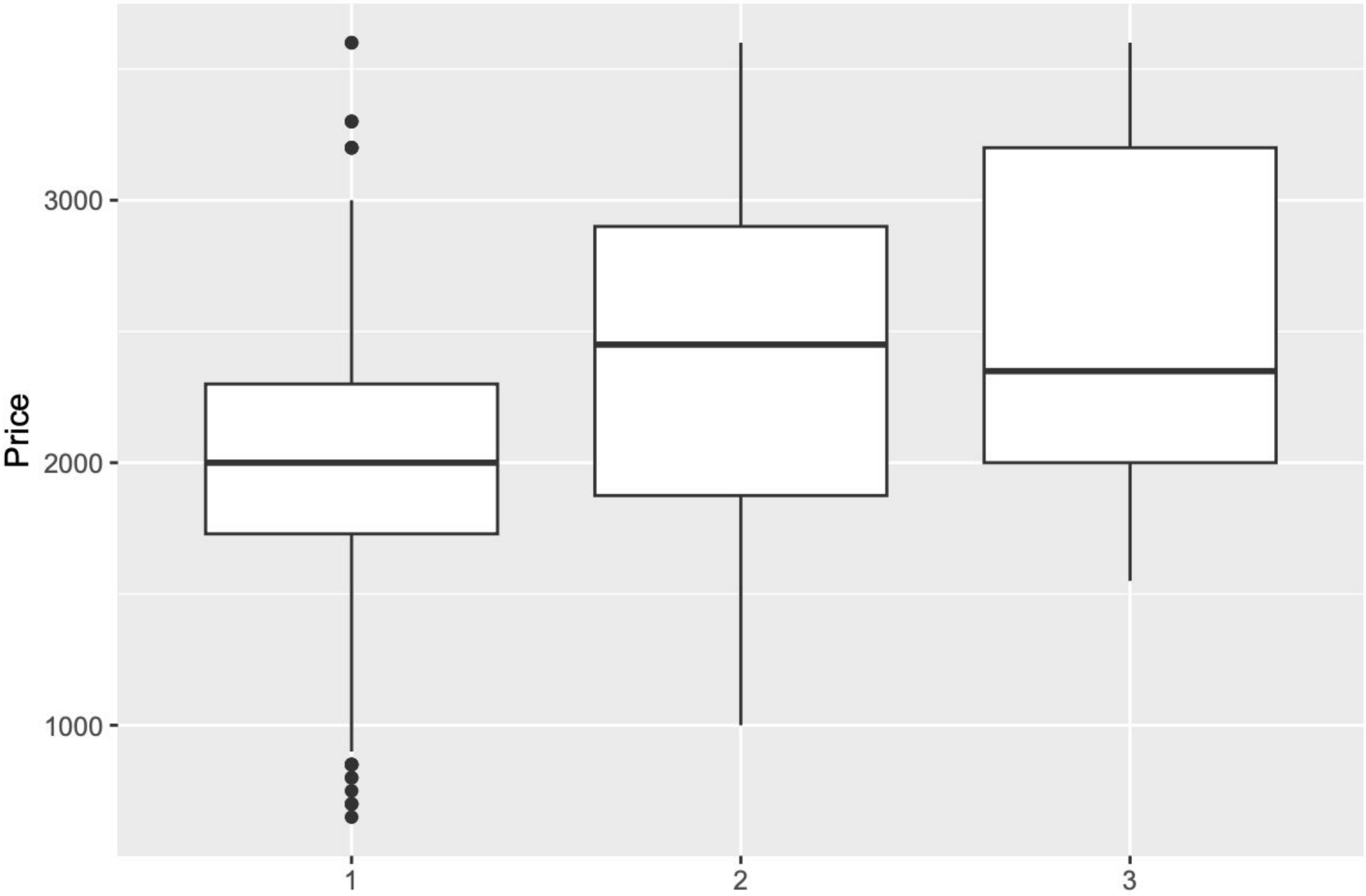
Number of Each Type of Room

Before including the number of bedrooms, bathrooms, and dens as explanatory variables for our multivariate linear regression model, we wanted to determine if there was a relationship between those variables and the price. Firstly, we did this by creating boxplots.

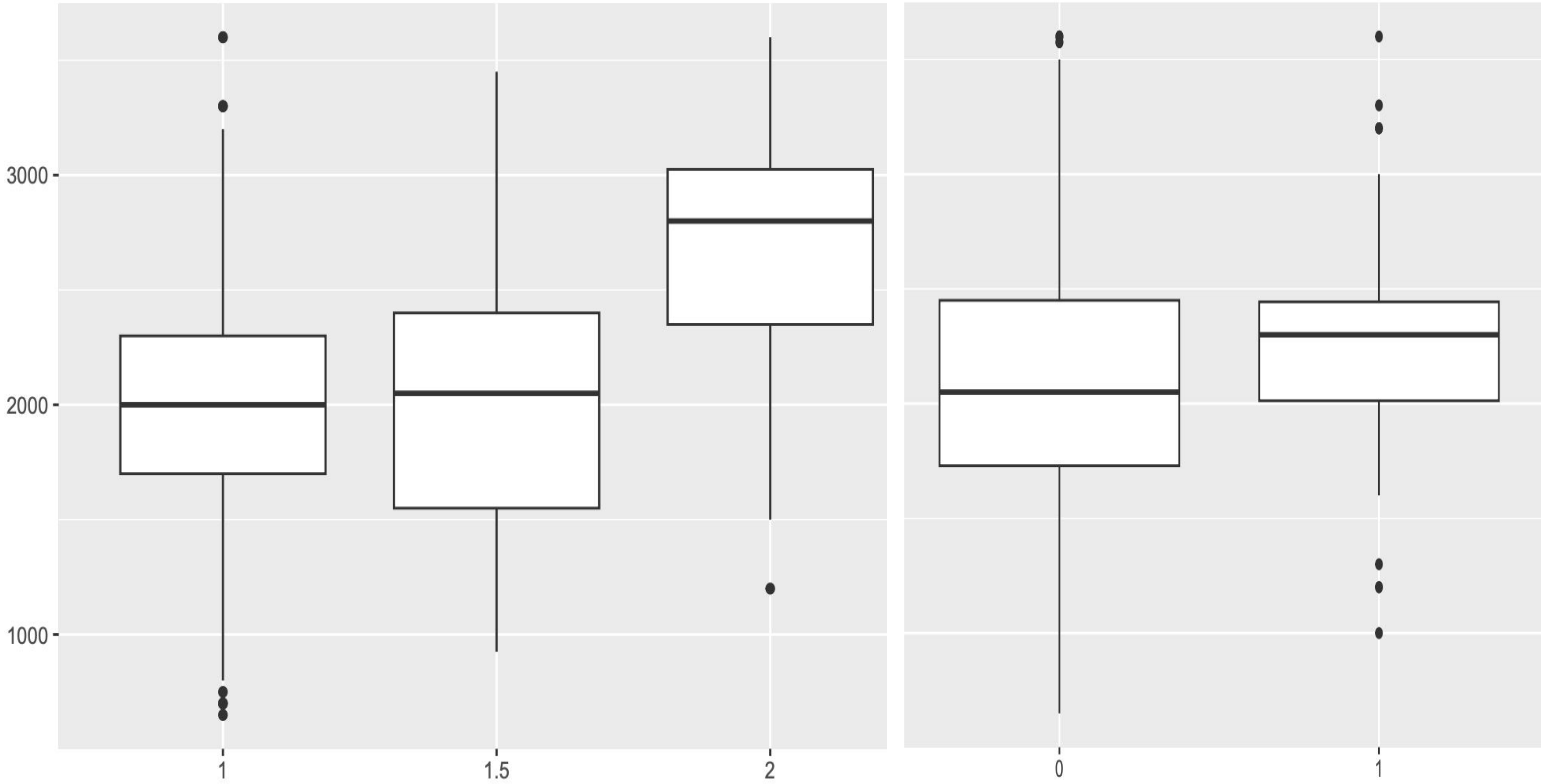
We also computed a t-test (one example below) for the slope for each of our variables, and found that the p-values were all very small (suggesting our variables are indeed related to price).



```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1616.58      44.64   36.22   <2e-16 ***
## Bedroom       374.83      31.02   12.08   <2e-16 ***
```



Number of Bedrooms



1 Bathroom 1.5 Bathrooms 2 Bathrooms No Den Den

y - axis represents the apartment rental price in dollars.

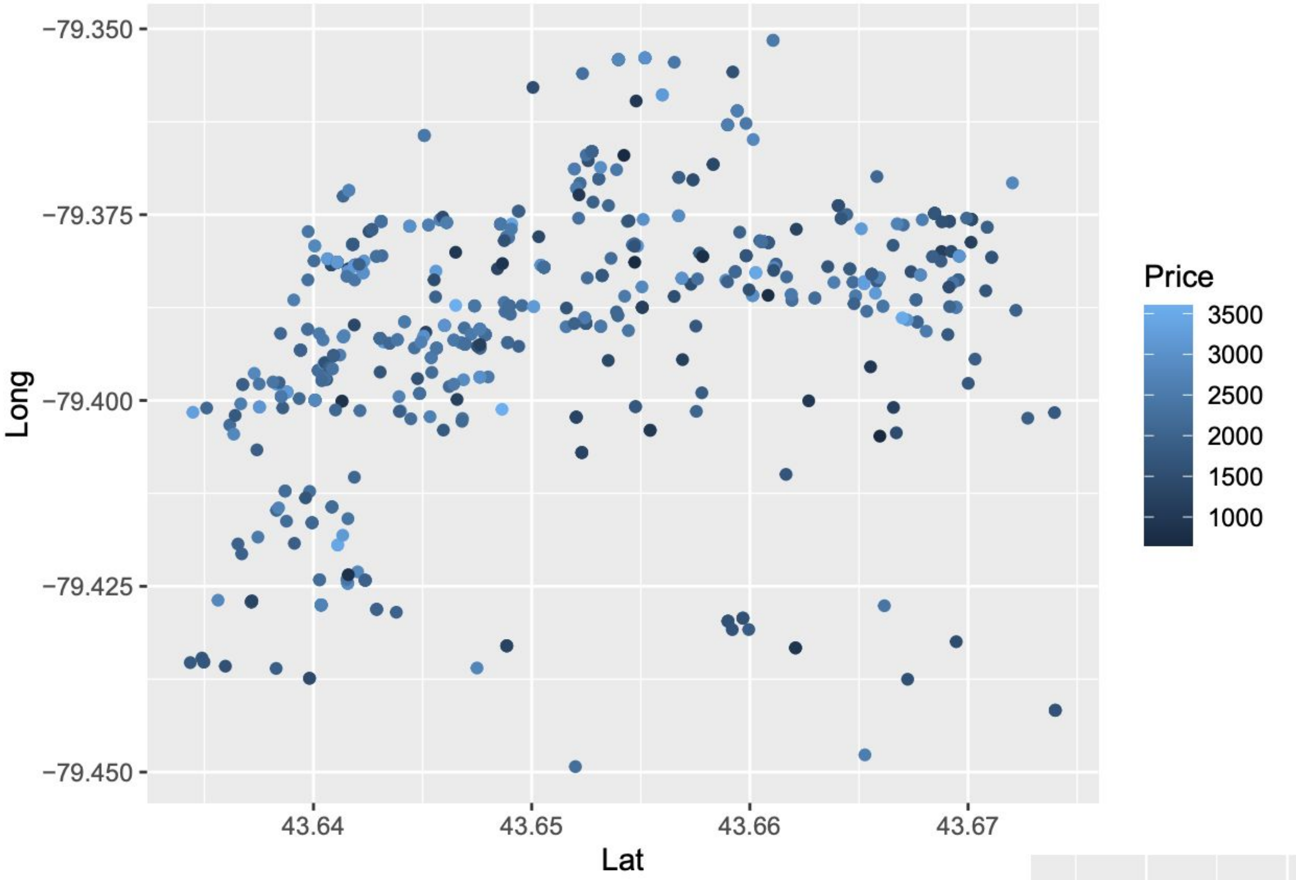
Data - Creating Variables for Model

Location

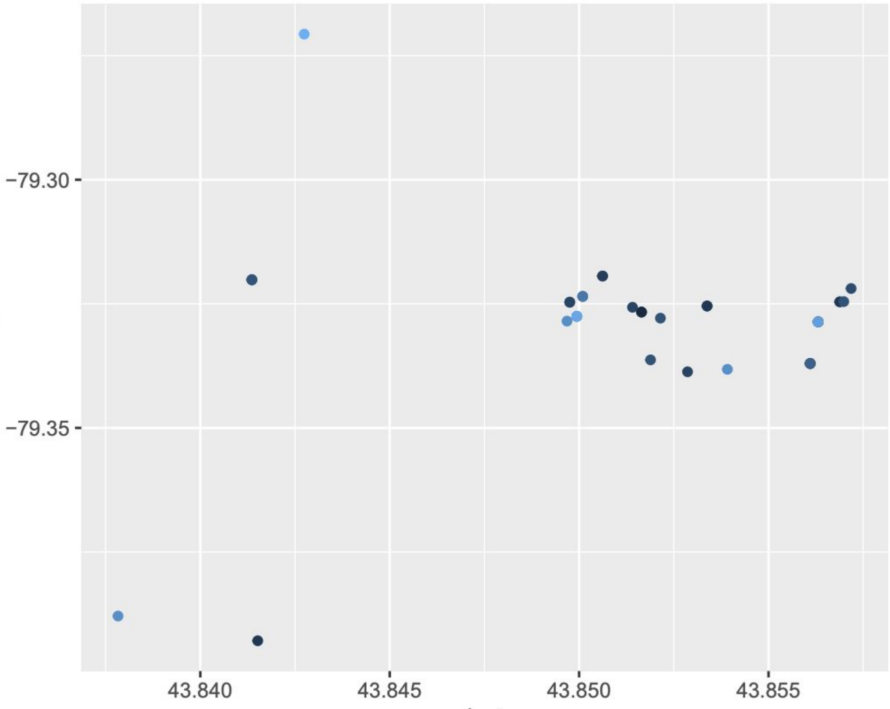
We created two variables to account for the location of the observations.

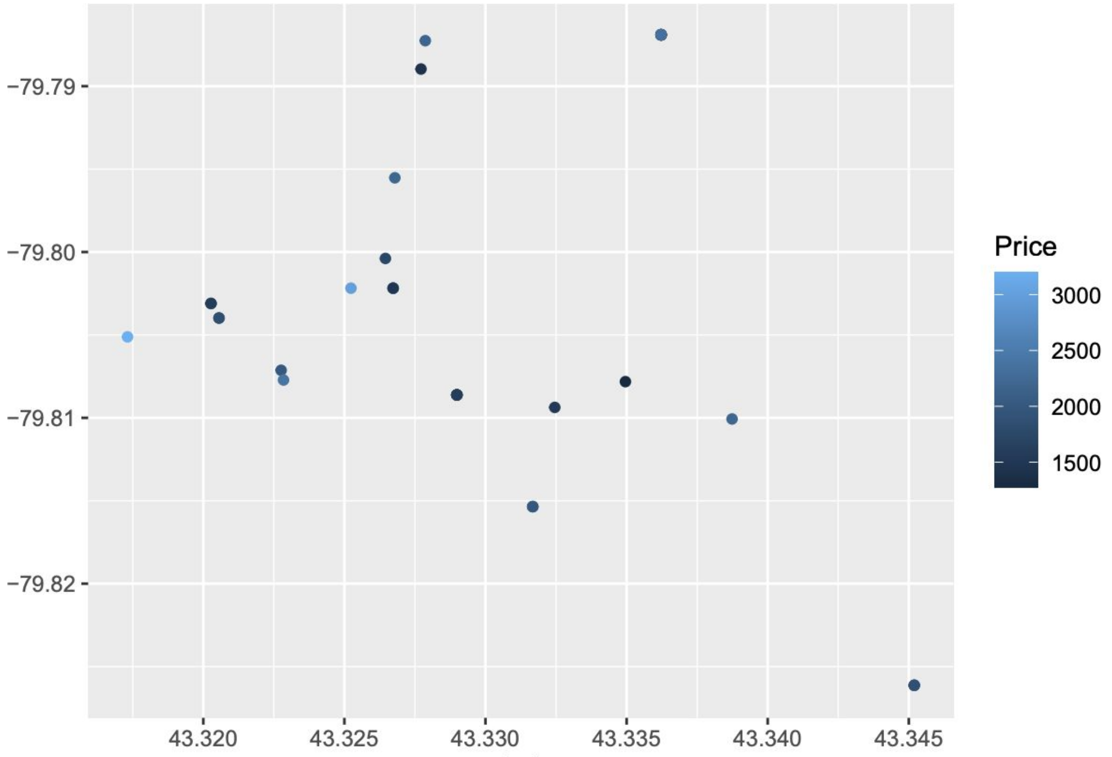
The first variable was the apartment’s proximity to the center of Toronto; to represent this, we used the location of our STA130 lectures. After this computation, we removed a few observations that were so far away from central Toronto that they were not relevant to our project.

The second variable was each apartment’s distance to the closest of the most expensive apartments. These valuable properties act as a proxy for the wealthiness of the neighborhood. First, we split the properties into groups depending on the number of bedrooms. We created a scatterplot to map out the prices of the different observations. We then zoomed into three “high concentration” areas of the plot.



Scatterplots show the three high concentration areas we zoomed into. Note that the latitude-longitude coordinates are different for each, as well as the price ranges.





Inclusion of Variables

- Based on the boxplots and the t-tests, all of our explanatory variables (beds, baths, den, DTGA, and DIS) appear to be correlated with the Price variable.
- This justifies their inclusion in our multivariate linear regression model.

Inclusion terms:

- An interaction term between two or more of the features should be included when the effect of one explanatory variable on the response variable depends on the value of another explanatory variable.
- This appears to be true in our case between the number of bathrooms and the number of bedrooms.
- Thus, we decided to integrate into our multivariate regression model this interaction between BED and BATH.

Final Model

$$\widehat{PRICE} = \hat{\beta}_0 + \hat{\beta}_1 * BED + \hat{\beta}_2 * BATH + \hat{\beta}_3 * DIS + \hat{\beta}_4 * DTGA + \hat{\beta}_5 DEN + \hat{\beta}_6 * BB$$

After utilizing R to compute the coefficients that minimize the Mean Squared Error, we yielded the following complete model.

(coefficients here have been rounded for visual purposes)

$$\widehat{PRICE} = 1893 + 220 * BED + 276 * BATH - 1710 * DIS - 168 * DTGA + 332 * DEN + 97 * BB$$

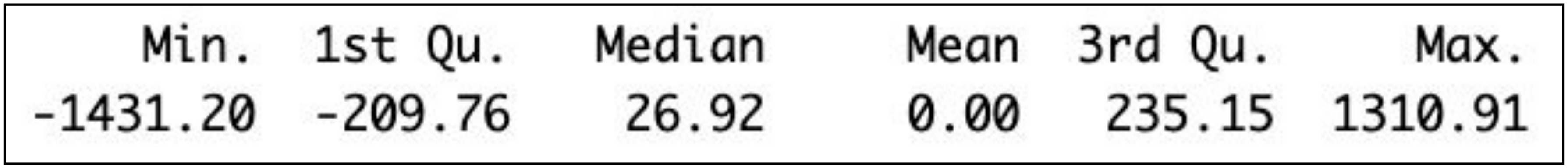
Normality of Residuals I

We wanted to determine whether the “Normality” condition is met (whether the residuals produced by our model follow a normal distribution).

To assess the “normality” of the residuals, we decided to produce a histogram of the residual distribution, seen in the next slide.

Although the distribution of the residuals is not a perfect normal bell-curve distribution (this is rarely the case in the real world), it appears to be roughly normal and symmetric, and no significant skew is apparent.

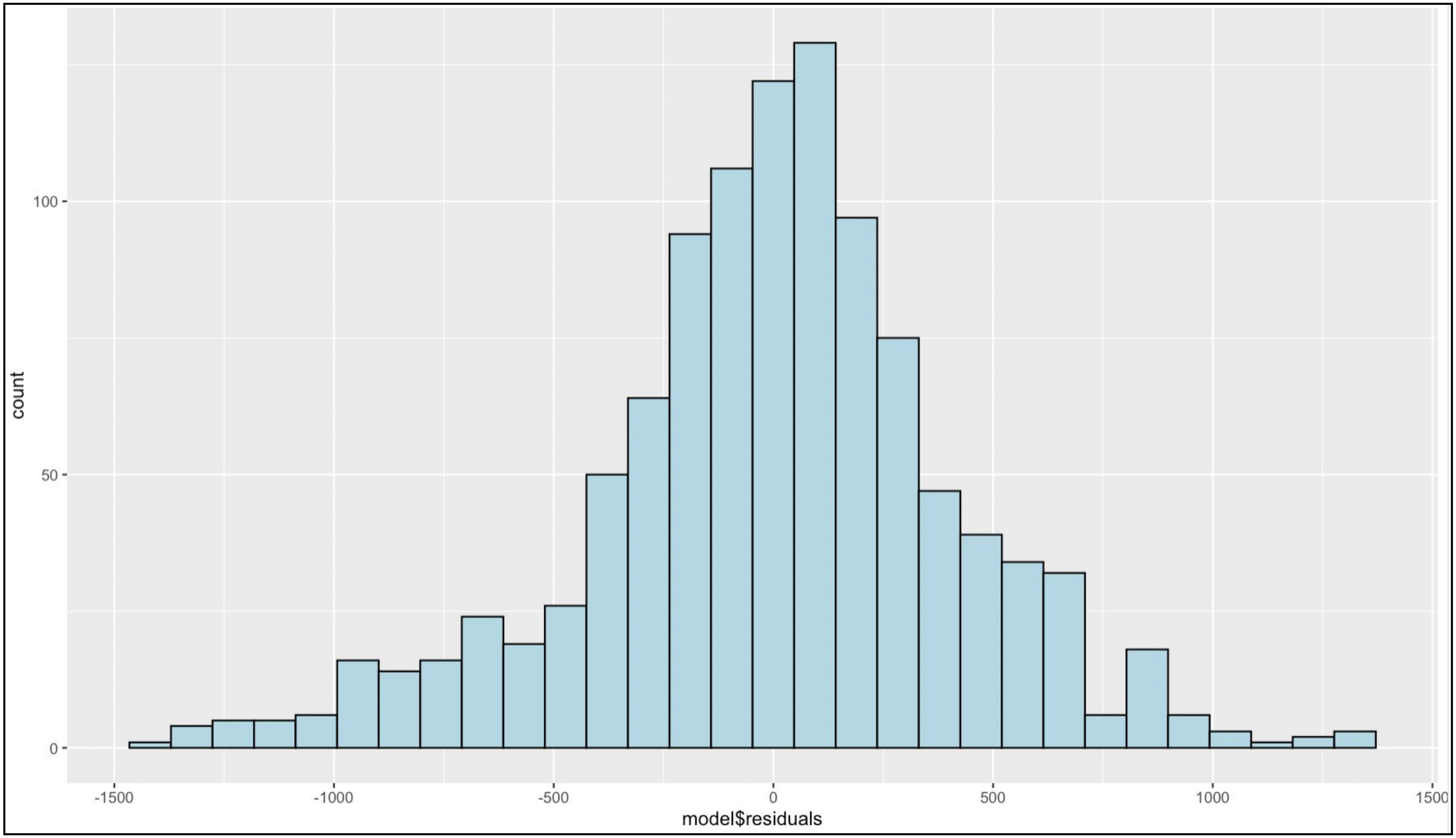
Thus, we can confidently say that the *Normality* assumption of the linear regression is indeed satisfied by our model and data.



```
    Min.  1st Qu.    Median      Mean  3rd Qu.      Max.
-1431.20  -209.76     26.92      0.00   235.15   1310.91
```

Normality of Residuals II

Histogram of residuals caused by our final model



Effects I

In the context of a linear regression model, the effect size of an explanatory variable X for a particular observation W is the coefficient of that variable (in the model) multiplied by the value of the variable for the particular observation W.

In the context of a linear regression model, the relative effect of an explanatory variable X for a particular observation W is the proportion of the total *effect sizes* of all variables (for the observation W) that is attributed to the contribution of the explanatory variable X. For our model, the relative effect of an explanatory variable X for a particular apartment observation W is the proportion of the estimated price (minus the model's intercept) that is attributed to X.

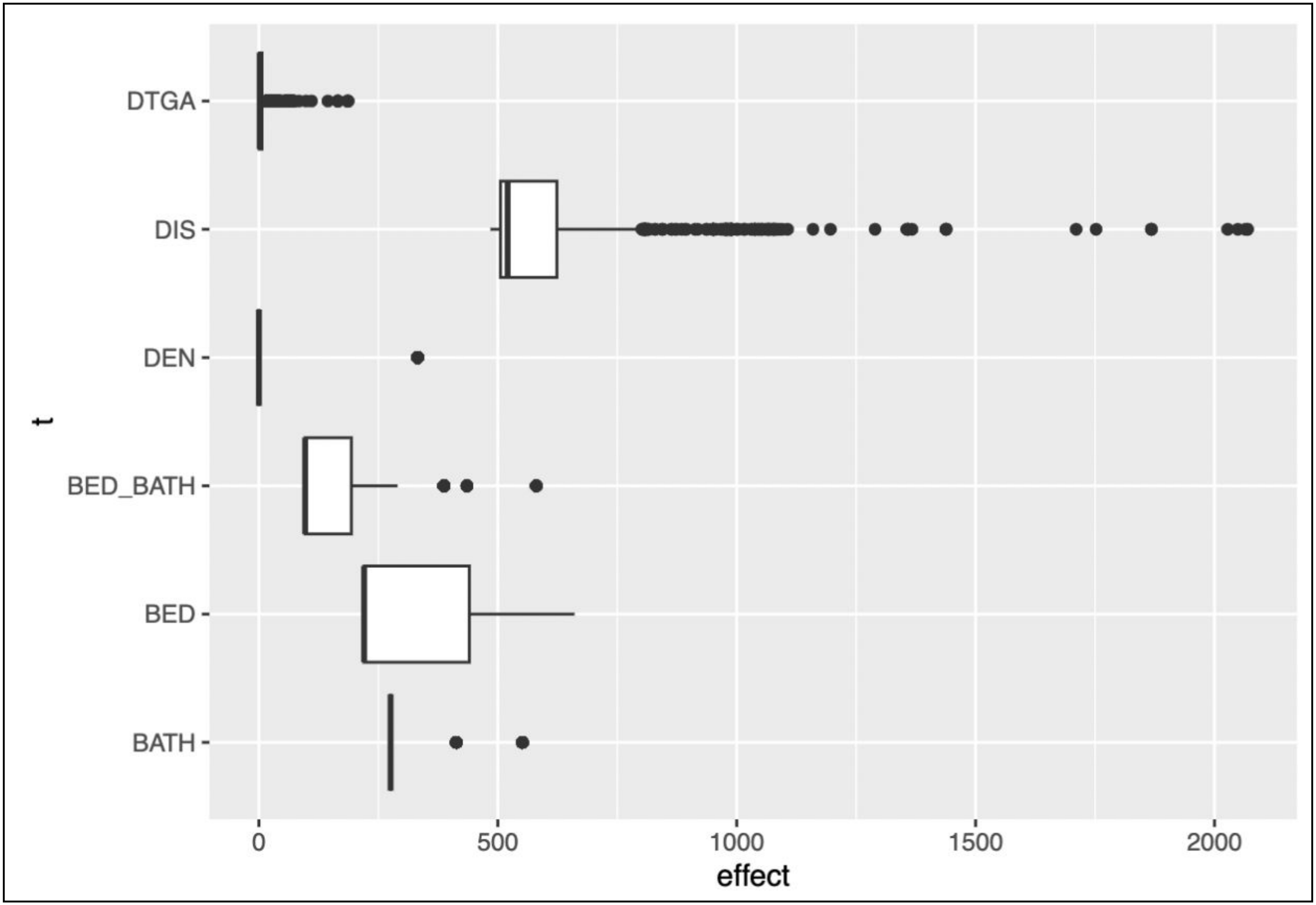
Effect Distributions I

Our first objective is to extract the distributions of the effects for each of the explanatory variables in our model (six variables in total), and create a plot with "stacked" boxplots (one per explanatory variable) that shows the distribution of the effects of each variable (across all observations).

For explanatory variables such as DIS and BATH, we consider the magnitude (absolute value) of the effects in order to be more easily compare them visually with the effect distributions of the remaining explanatory variables.

Once we have a list of *effects* of a variable across all observations, for all explanatory variables, we plot the distribution of the effects of each variable using boxplots.

Effect Distributions of the Explanatory Variables



Relative Effects

We now want to answer the following subquestion:

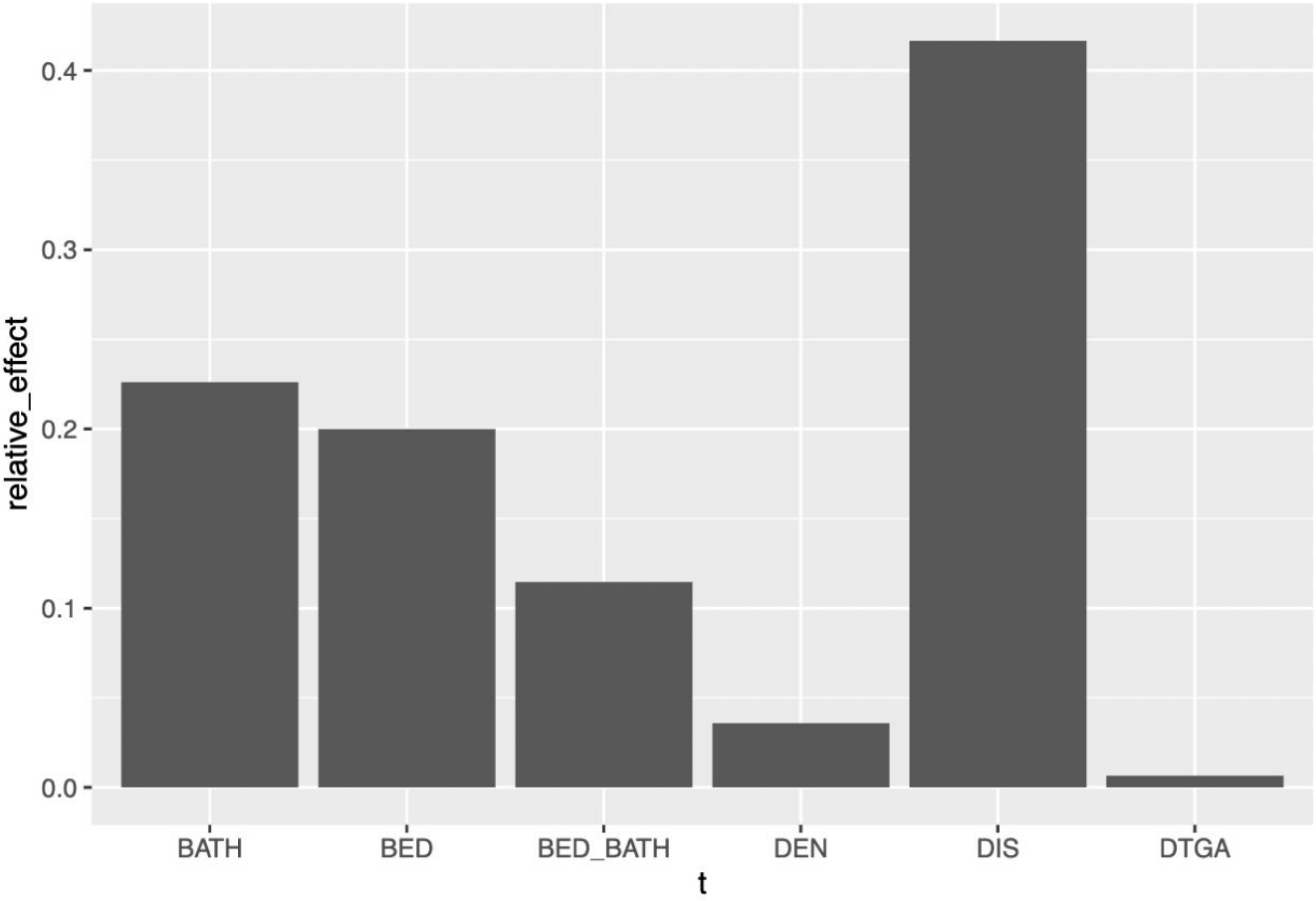
“Across all observations, what proportion of the total effects is attributed to each explanatory variable?”

To compute this "proportion" we do the following:

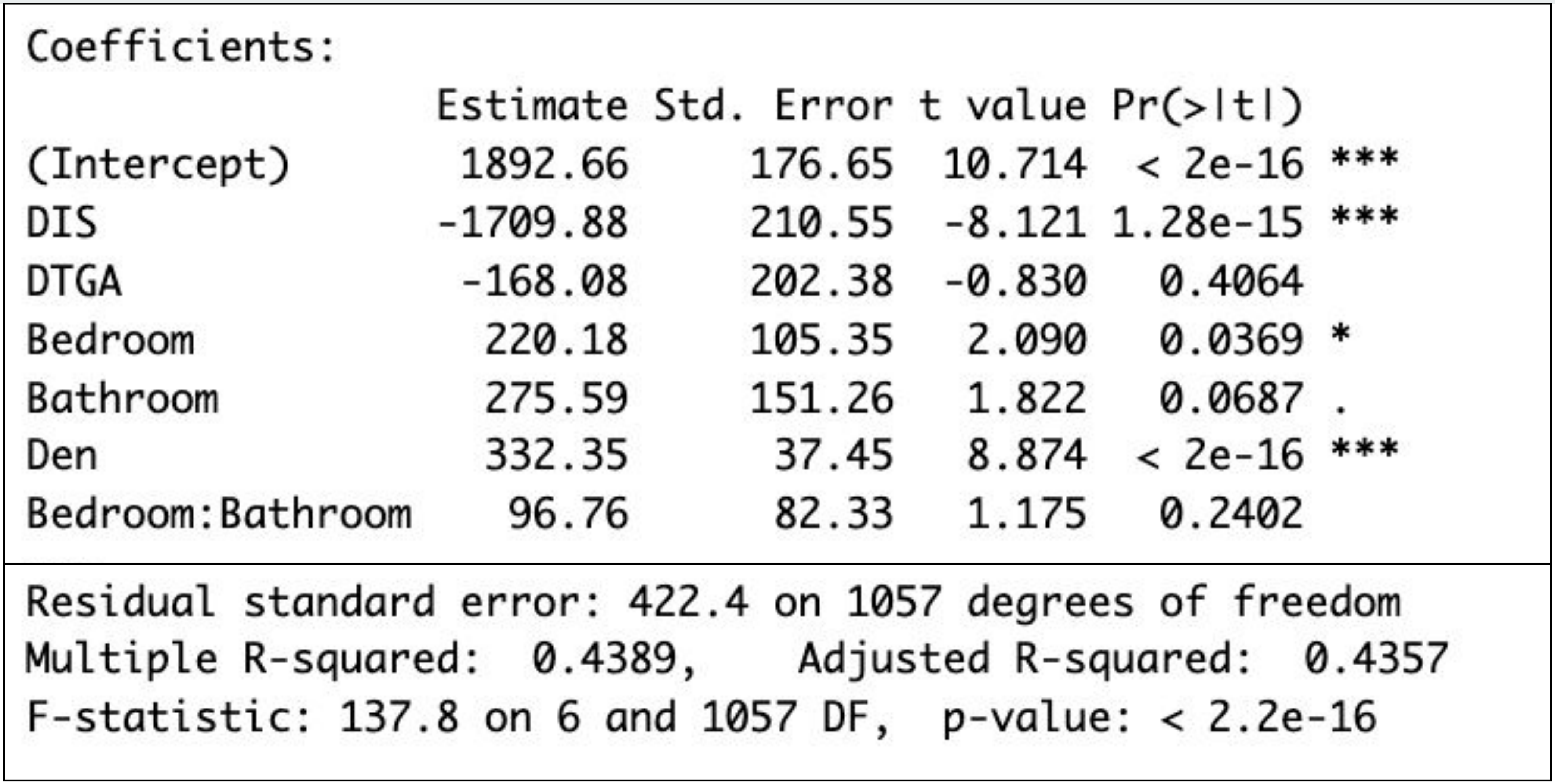
1. Compute the total effect of a particular feature across all observations. That is, calculate what the effect of a particular explanatory variable is for each of the observations, and sum up these effects (for all observations).
2. Divide the quantity from (1) by the sum of the effects of all features across all observations.

Thus, we define the "relative effect" of a feature/explanatory variable to be the proportion of the *total effects* of all features (across all records) that this feature is responsible for.

Histogram of Relative Effects



Model Effectiveness



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        1892.66     176.65  10.714  < 2e-16 ***
DIS               -1709.88     210.55  -8.121 1.28e-15 ***
DTGA               -168.08     202.38  -0.830   0.4064
Bedroom             220.18     105.35   2.090   0.0369 *
Bathroom            275.59     151.26   1.822   0.0687 .
Den                 332.35      37.45   8.874  < 2e-16 ***
Bedroom:Bathroom     96.76      82.33   1.175   0.2402
```

```
Residual standard error: 422.4 on 1057 degrees of freedom
Multiple R-squared:  0.4389,    Adjusted R-squared:  0.4357
F-statistic: 137.8 on 6 and 1057 DF,  p-value: < 2.2e-16
```

Our model as a whole explains 43.89% of the variation of the price in our dataset.

The explainability of our model is satisfactory given the limited number of variables our dataset has. It could have been higher if we had more information for each apartment record, such as the size of the unit (square feet), utilities, building age, etc.

Effect Distributions - Results

Interestingly, the (unsigned) effect distributions for all explanatory variables appear to be **skewed right**. This right skewness can be interpreted as follows: for each of the variables, there are many observations for which the net effect of the variable is much greater than what the general effect of that variable is.

This right skewness is most evident through the DIS explanatory variable, which also appears to be the variable with the largest effect to the prediction of a price.

One possibility for this skewness: there is some other variable that is missing from our model, and whose explainability is made up for by (in absolute terms) higher coefficients for the variables that are indeed included, leading to some effects that are high enough to be outliers.

In particular, for this particular analysis model, the effect boxplots for the BATH, BED, and DEN variables are of limited value due to the limited amount of values that effects of these variables can take.

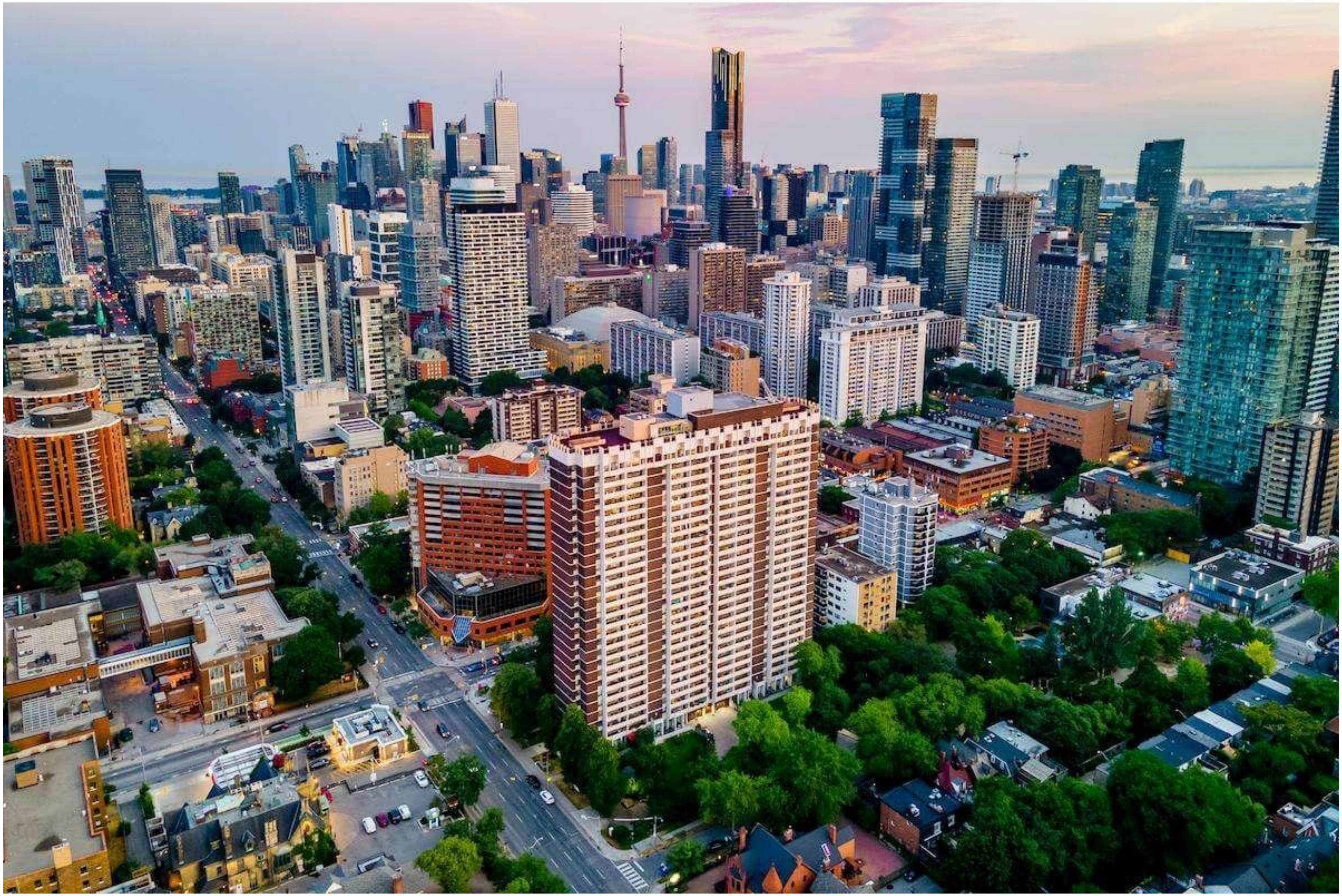
Relative Effects - Results

The DIS variable, which denotes the distance of an apartment from the center of Toronto, has the largest overall effect on the Price of an apartment.

The den and DTGA variables appear to be rather insignificant:

DEN: Although whether an apartment has a den is correlated to its price, it does not make a massive difference to the apartment price, hence the low overall effect.

DTGA: Although one would expect that the distance of an apartment to a “good” area to have a significant effect on the price of an apartment, our model disagrees. One possible explanation for this contradiction is that some variables “explain the same thing”; that is, their individual relationship with Price is based on the same underlying factors. This means that a variable could explain the same variation in the price data that another variable (in our case, DIS) does, making one redundant. This also makes sense as we know from our t-tests that DTGA has a correlation with price when the two variables are individually compared.



STOP
ALL WAY
40