A Linear Regression Analysis of the Effect of Location and Number of Rooms on Toronto Apartment Rental Price

Daniel Du, Mimis Chlympatsos

April 9, 2023

STA130 Final Project Report

Abstract

The aim of our project was to analyze the relationship between Toronto apartment rental prices and factors such as the number of rooms and distance from a central location. The motivation for this project was to identify the contributors to variation in Toronto apartment rental prices and to determine how we can optimally make predictions on rental prices given only the variables in our dataset. We constructed a multivariate linear regression model to predict the rental price of an apartment based on the number of bedrooms, bathrooms, and dens, as well as its proximity to a central location and to a good area. Furthermore, we generated two graphs to explore the relative significance of the aforementioned factors in predicting apartment rental prices, as well as the extent of their impact.

Conducting this analysis yielded a predictive model for the rental price of a Toronto apartment, using only variables directly derived from the apartment's location and number of bedrooms, bathrooms, and dens. According to the model, which explains about 44% of the overall variability in rental prices, the primary determinant of an apartment's rental price is its proximity to the University of Toronto; accounting for approximately 40% of the total *effect* of all variables on the rental price. The number of bedrooms and bathrooms are the next most important components, as both variables are responsible for approximately 20% of the overall effect on price. Future work can extend this analysis to include additional predictors, such as square area, apartment age and size, and neighbourhood characteristics, to further explore the factors that contribute to the variation in rental prices, and to come up with a model that explains an even greater part of this variation in prices.

Introduction

This project seeks to answer two 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?

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?

We conducted this analysis and attempted to give answers to these questions in order to help us understand the factors that contribute to the variability of apartment rental prices in Toronto, and create a predictive model that would help us estimate the rental price of any given apartment. This model can be useful for both landlords and tenants seeking to determine the appropriate rental price for an apartment and proceed to make choices. We carefully constructed a linear regression model, using the given variables as well as newly derived variables, to answer Question 1, yielding satisfactory results. Our analysis also involved determining, appropriately visualizing and analyzing the effect sizes and distributions of our explanatory variables -the average contribution of each explanatory variable to the predictions of our model. This allowed us to figure out which explanatory variable has the greatest contribution to the predictions of our linear regression model. The main statistical tools employed throughout this analysis are (multivariate) linear regression, hypothesis testing (evaluation), application (and adaptation) of formulae such as the one for the *effect* of a variable, and visual analysis.

Data - Collecting and Cleaning

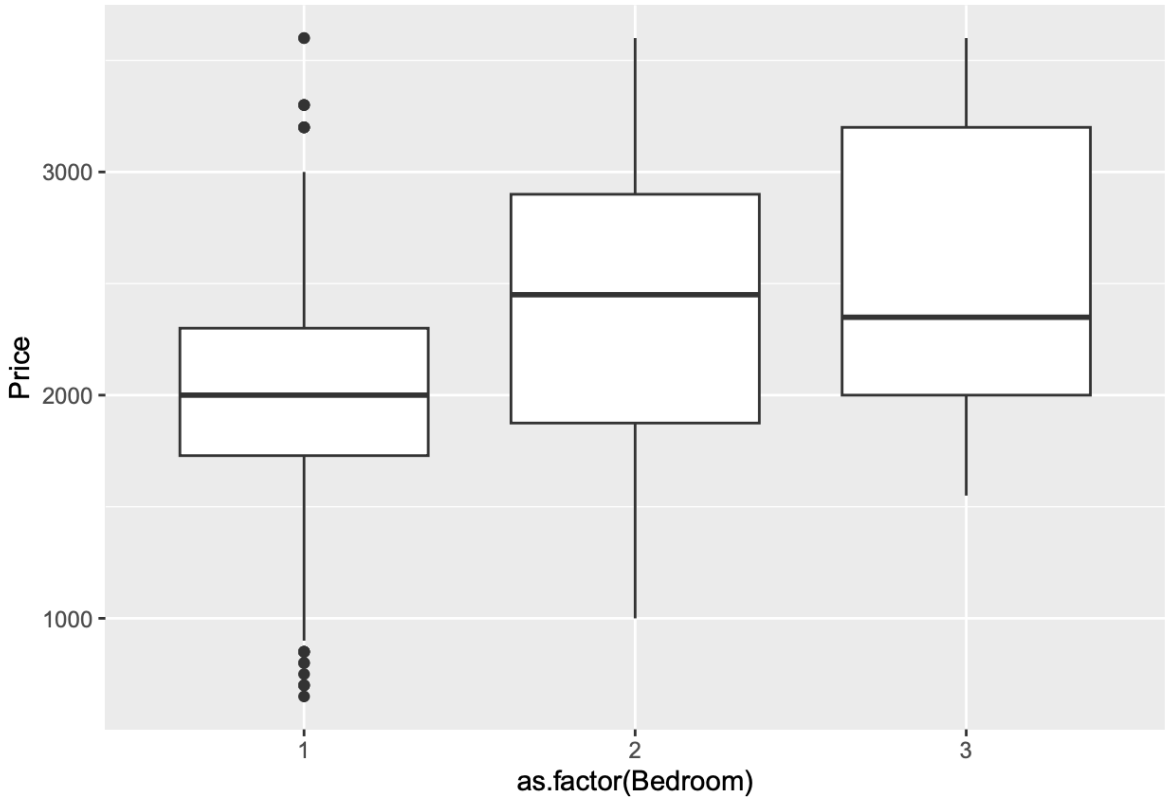
The data used is a collection of Toronto apartment rental prices gathered from Kijiji in 2018. In total, there were 1124 observations of the data. The data had 5 variables which were utilized in our model - the number of bedrooms, bathrooms, dens, and coordinates for latitude and longitude - and the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

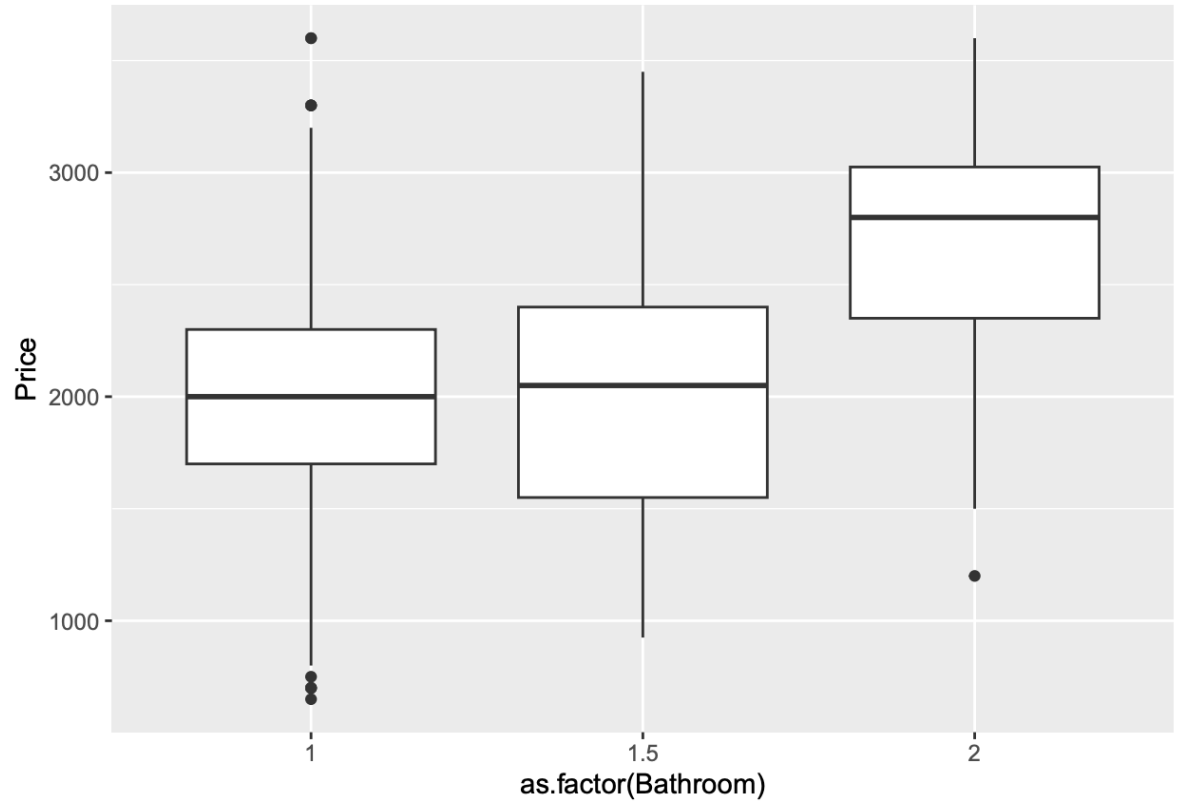
The prices for some observations were below $80/month. These were extreme outliers, and most likely were faulty observations. To avoid negatively affecting the predictability of our regression model, we removed these observations from our dataset. We also decided to remove the observations with more than 2 bathrooms, as there were only 8 observations in total, and we did not want these to have a great effect on the bathroom predictor variable in our model. There were no NA values that we had to remove.

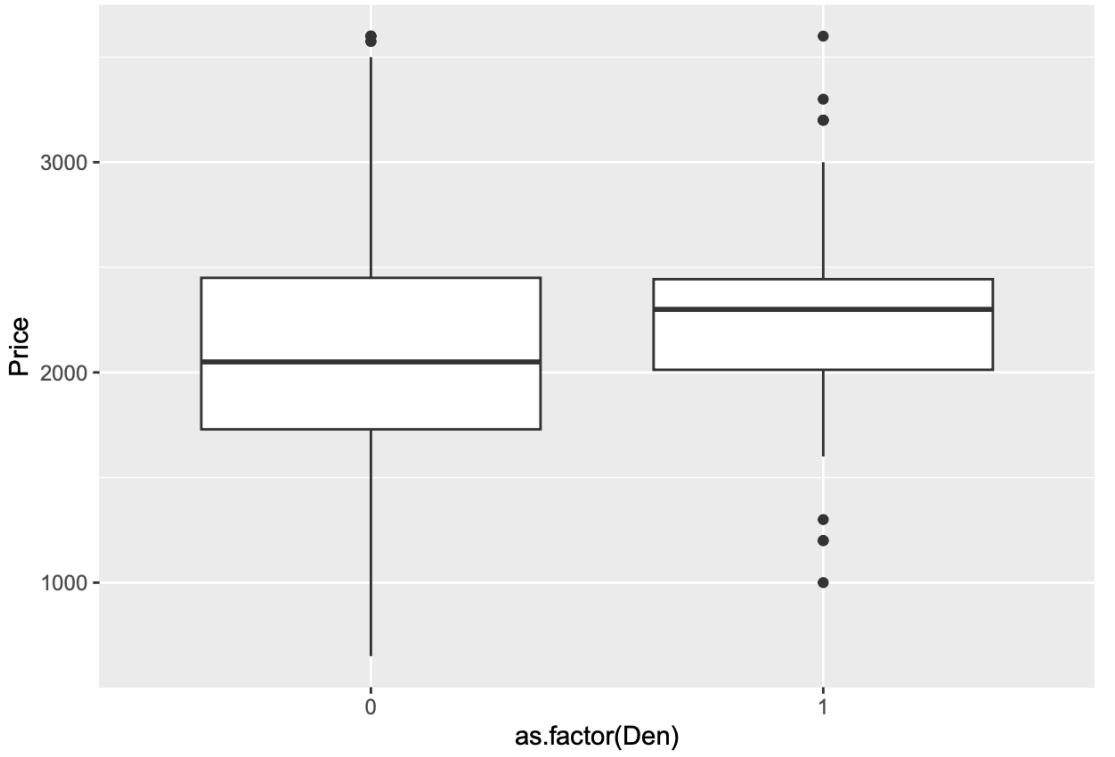
Data - Creating Model Variables

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 three sets of boxplots. We considered the number of bedrooms, bathrooms, and dens as ordinal categorical variables. We created a plot with a different boxplot for each level of the categorical variable, which in this case was the number of the room. Thus, we were able to compare the apartment price with the number of each room.


Price
3000
2000
1000
1
2
3
as.factor(Bedroom)

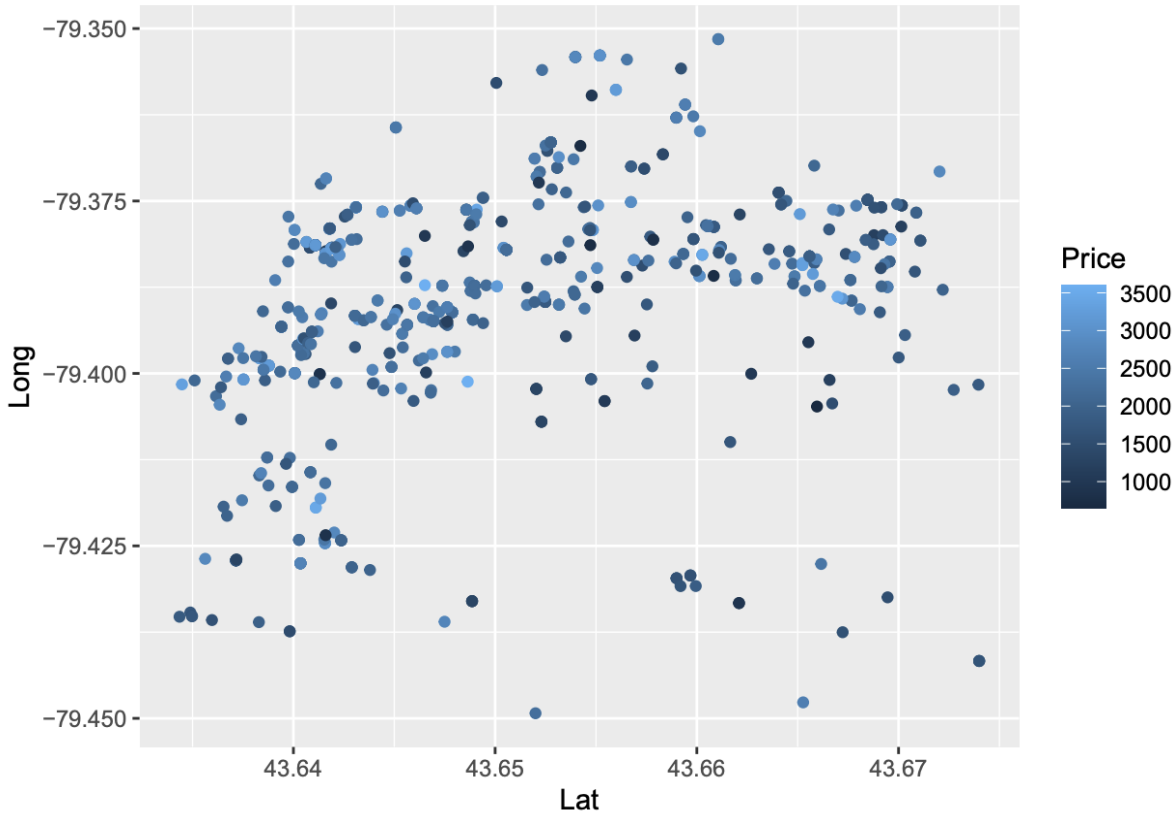

Price
3000
2000
1000
1
1.5
2
as.factor(Bathroom)

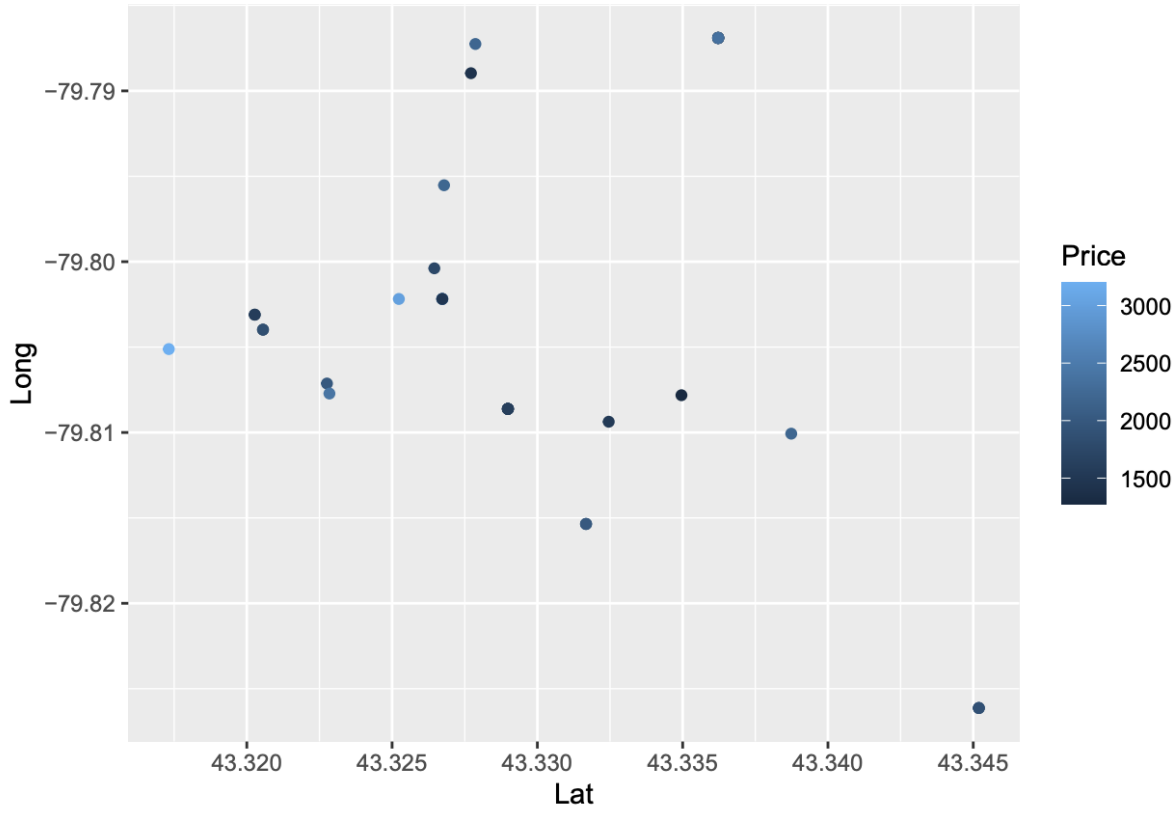


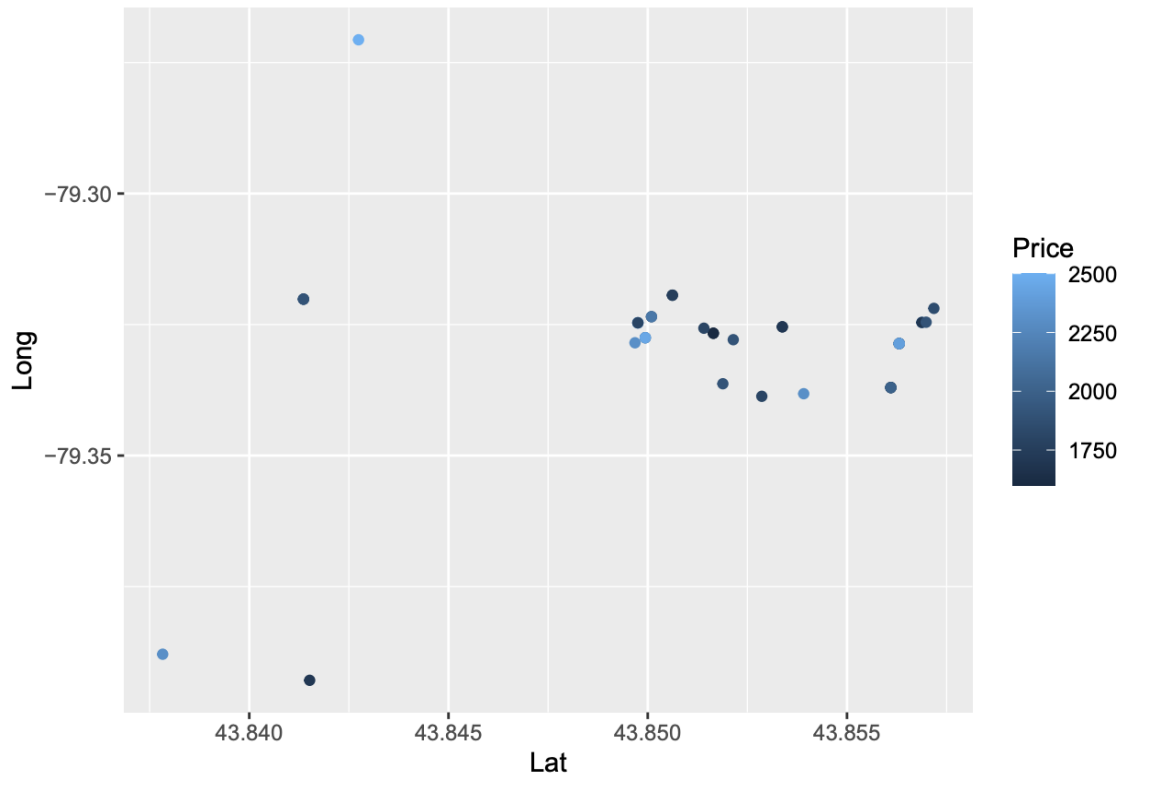
Location

We created two variables to account for the location of the observations. The first was its proximity to the center of Toronto; to represent this, we used the coordinates of the Mechanical Engineering building where our STA130 lectures are held. Due to the negligible curvature observed in a geographic area as small as the one we are interested in, we computed the distance between each apartment and the building coordinates using the Pythagorean Theorem. After this computation, we removed a few observations that were so far away from central Toronto that they were not relevant to our project.

To account for the general neighbourhood, we created a variable to ascertain whether or not an apartment was located in a wealthy area. We decided to represent this by calculating each apartment's distance to the closest of the most expensive apartments. These valuable properties act as a proxy for the wealth of the neighbourhood. First, we split the properties into groups depending on the number of bedrooms. Our earlier box plots showed that rental prices were greater as the number of bedrooms increased; when determining which properties were most valuable, we decided to control for the number of bedrooms. To achieve this, we created a scatterplot with the latitude on the x-axis and longitude on the y-axis to map out the prices of the different observations. We then zoomed into three "high concentration" areas of the plot. These scatterplots are pictured below. Note that the latitude-longitude coordinates are different for each of the plots, as we wanted to account for:


−79.350
−79.375
−79.400
−79.425
−79.450
Long
43.64
43.65
43.66
43.67
Lat
Price
3500
3000
2500
2000
1500
1000


−79.79
−79.80
−79.81
−79.82
Long
43.320
43.325
43.330
43.335
43.340
43.345
Lat
Price
3000
2500
2000
1500



For each of these three areas, we arranged by descending price and sorted the data by the number of bedrooms, resulting in three distinct datasets. We then kept the 3 highest-price observations from each dataset. Finally, we combined our three data sets of 1, 2, and 3 bedroom high-value properties, and selected spread-out data points from this new group to account for as much of the geographical area as possible. We did this for each of the aforementioned concentration areas. Due to multiple observations of apartment prices from the same apartment, there were duplicate locations; we removed these, ending up with 7 observations. We then used the latitude-longitude coordinates of these observations to represent a good location.

Variable Testing

Although our above graphs were very insightful, and suggested that a relationship did indeed exist, they are only visualizations and do not hold all the information in our data. To more concretely decide whether relationships existed between each of our explanatory variables and the response variable), we decided to compute a t-test for the slope for each of our variables. Indeed, for all aforementioned variables, the p-values were so small that we can confidently reject the null hypothesis and conclude that the variables are related to price. We have included one of these tests below; the rest have been omitted but can be found in our project code.

```
## 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 ***
```

Methods/Analysis

Explanatory Variables

As described in the previous section, all of our explanatory variables (beds, baths, den, DTGA, and DIS) appear to be correlated with the Price variable. This is suggested by the produced boxplots, as well as by negligible p-value resulting from the t-tests (which we performed for each of the individual explanatory variables. This justifies their inclusion in our multivariate linear regression model.

Generally, as taught in STA130, an interaction term between two or more of the explanatory variables/features should be included when the following is true: the (size of the) effect of one explanatory variable on the response variable depends on the value of another explanatory variable. In the context of our own housing dataset, we believe that this is in fact true between the number of bathrooms and the number of bedrooms. For instance, we would expect a second bathroom to add more value to an apartment with three bedrooms than to an apartment with just one bedroom (since the latter does not really have a use for a second bathroom anyway). Therefore, we decided to integrate into our multivariate regression model (as an explanatory variable) the interaction between the number of bedrooms and the number of bathrooms.

Final Model

Thus, our final multivariate regression model is the following:

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

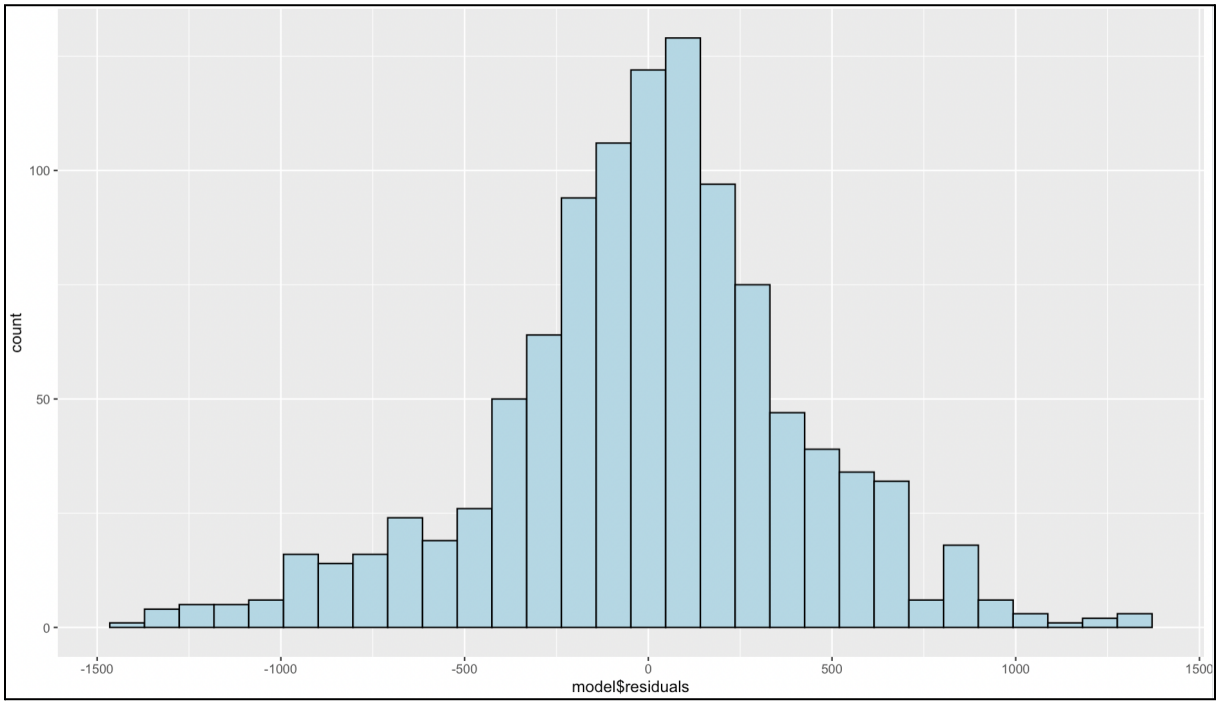
After utilizing R to compute the coefficients that minimize the Mean Squared Error, we yielded the following complete model. To display the model below, I rounded the coefficients to the nearest integer so the equation can fit and be easily readable (this rounding is insignificant, considering the size of the coefficients and of the explanatory variables).

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

Distribution of Residuals

As taught in STA130, the implementation of a linear regression model comes with certain assumptions/preconditions that must be met: Normality (of residuals), Homoscedasticity, Independence (between the explanatory variables), and Linearity (of the relationship between the features and the response variable). Although generally, it is infeasible (and often impossible) to check that all four of these are met, now that we have our model we can easily determine whether the “Normality” condition is met; that is, whether the residuals produced by our model follow a normal distribution.

First, to assess the "normality" of the residuals, we decided to produce a histogram of the residual distribution, which I have placed below. 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.



In addition, we decided to produce a summary table of this distribution of residual errors caused by our model, which can be seen below.

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

The mean of the residuals is 0.0, which is a direct consequence of the mathematical computations performed to produce the linear regression model, particularly that our loss function is the Mean Squared Error (MSE).

But the median is also insignificant, at just about 24.2 dollars. Although the "Max" and the "Min" residuals are quite large (in absolute value terms), this table shows that we are able to predict 50% of the observations' price within -approximately- a 230 dollar error, since the third quartile is 235.15 and the first quartile is -209.76. This is good news for the predictability and validity of our model!

Effects

We begin with definitions:

Definition: 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.

Definition: 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.

In the context of 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.

Our second question tries to determine which explanatory variables (from our model) are most important for predicting apartment rental price, and by how much.

We break this second question into two parts, each utilizing a different procedure:

I. Effect Distributions

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 each of the explanatory variables included in our linear regression model, we calculate its effect for all observations; that is, for every observation in our dataset, we will calculate its effect/contribution on the prediction of the price.

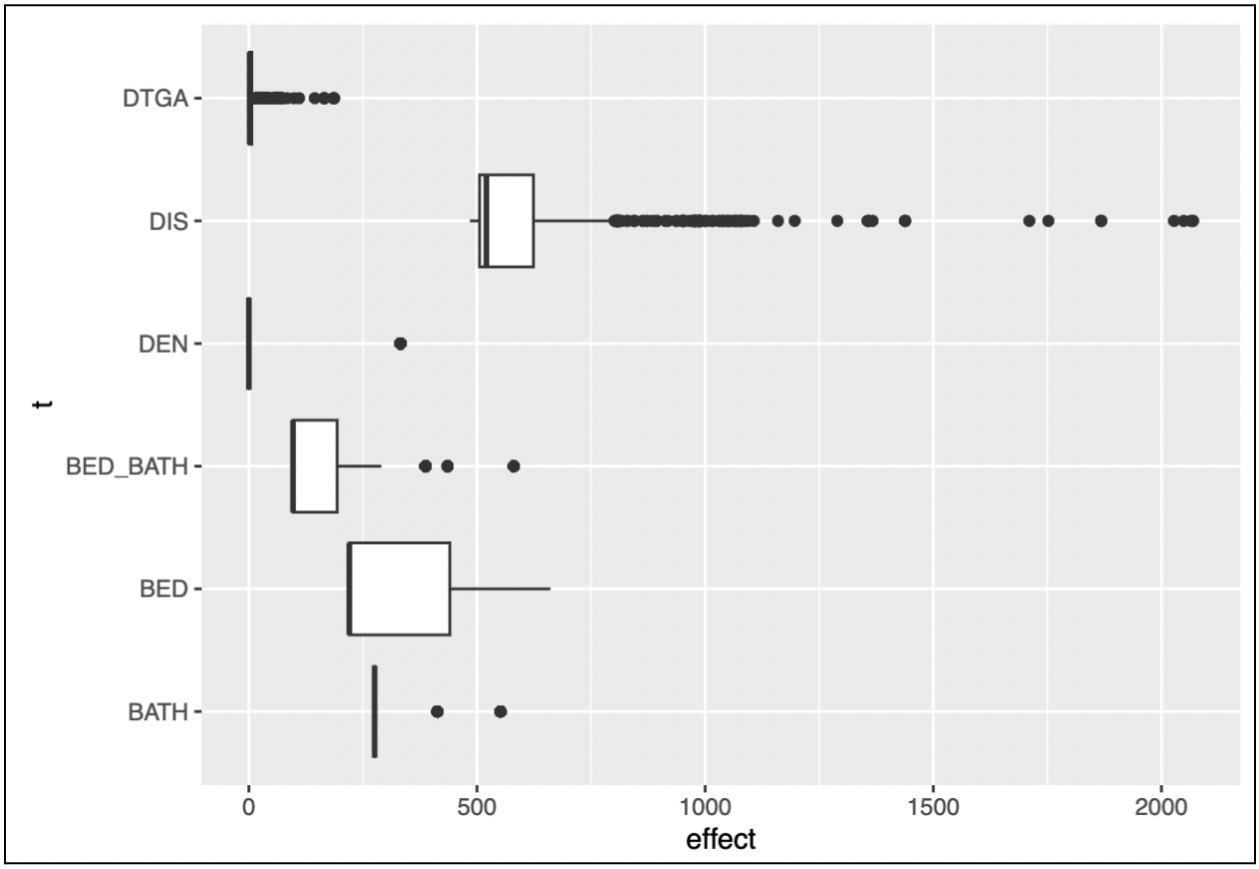
For explanatory variables such as DIS and BATH (which represent the distance to the center of Toronto and the number of bathrooms, respectively), the coefficients produced by our model are negative. This is reasonable, because, for instance, the further away an apartment is from the center of the city, the cheaper we may expect this apartment to be. Thus, in order to more easily compare them and visualize them with the effect distributions of the remaining explanatory variables, we decided to only take into account the magnitude of the effects; that is, we took their absolute value.

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. We sought to produce a plot that includes all the boxplots next to each other so that the distributions of the *effects* of the different explanatory variables can easily be compared to each other (however, the

real potential to compare the effects of the different variables will come with the second part of this question, with "relative effects").

An important design note: In order to create a plot with many boxplots next to each other (each boxplot denoting the *effects* distribution of a particular independent variable) with `ggplot` and `geom_boxplot`, we had to initialize a new data frame. The role of this new data frame will be to hold all the effects in a particular way:

Each observation of this new data frame has two variables/fields; the effect (of numeric type) and a type, which is the "label" of the explanatory variable that this effect corresponds to (e.g. "BED"). So we have two columns; "t" for type, and "effect", for the actual effect. Each observation thus represents a particular effect of an explanatory variable for a particular apartment record from our main dataset. This configuration of the data frame makes it easier to work with `ggplot` (in particular to create a plot with many boxplots stacked on top of each other, or next to each other, as per our preference).



II. **Relative Effects**

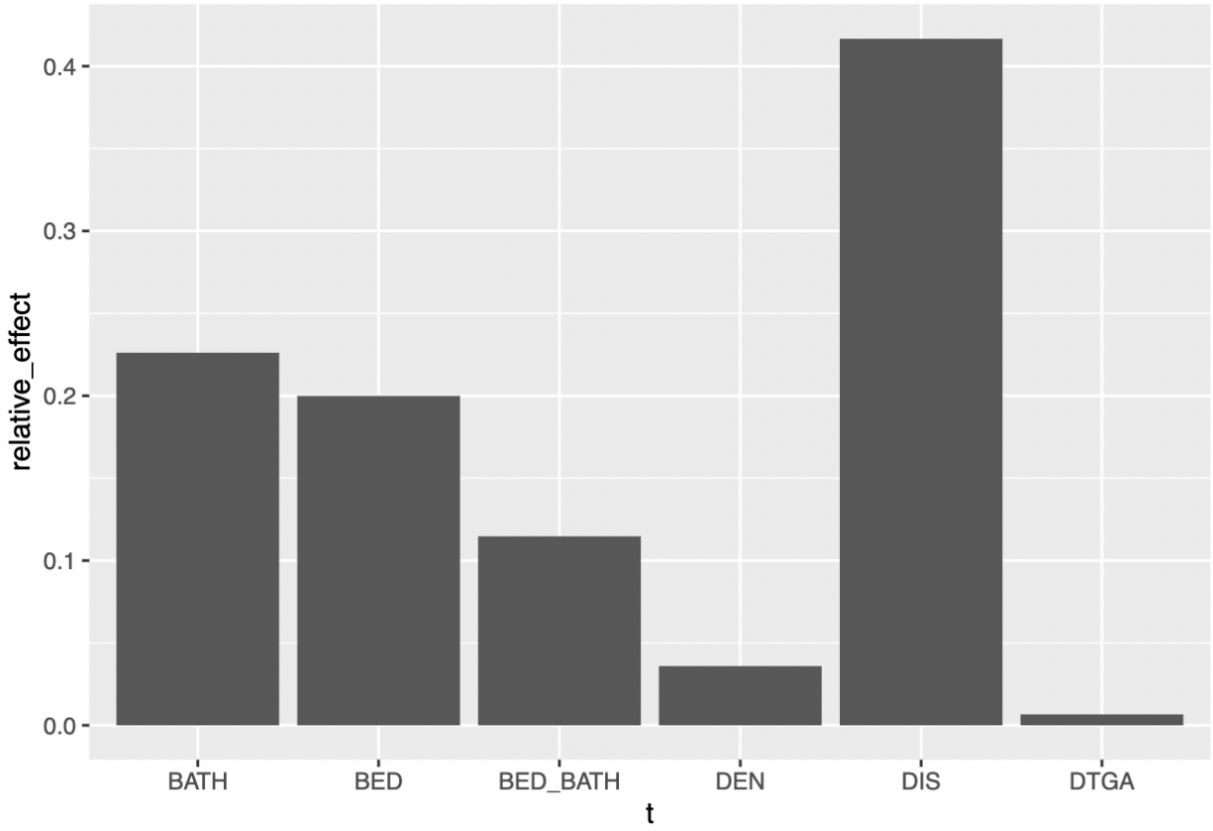
Although part I of this second question allowed us to compare the distributions of the effects of the different explanatory variables, we now want to answer the following, more specific, subquestion: across all observations, what proportion of the total effects is attributed to each explanatory variable?

To compute this "proportion" we do the following:

Firstly, we will compute the total effect of a particular feature across all observations. In other words, we calculate what the effect of a particular explanatory variable is for each of the observations, and sum up these effects (for all observations).

Then, we proceed to divide this quantity by the sum of the effects of all features across all observations. Note that for any given observation, the sum of the effects of the features is equal to the price minus the intercept.

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.



The results of the methods and procedures applied to our two questions are analyzed in the section that follows.

Results & Discussion

Q1: Multivariate Linear Regression Model

Firstly, we want to determine the effectiveness of our multivariate regression model (the output of our first question) that predicts the rental price of an apartment in Toronto. Partial analysis performed in the previous section on the distribution of the residuals showed that 50% of the dataset observation prices are predicted within approximately a 230-dollar error, which is a positive indication. However, the effectiveness of our model can more carefully be analyzed through the following part of the output of `summary(model)`:

```
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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

Many interesting things emerge from this table of statistics. The first is this very interesting contradiction:

One one hand, we have verified (as discussed in the data section) that for each of the six explanatory variables, if this variable is included alone in a model (where the response variable is, as always, the price), then the p-value of the *t-test* for the coefficient is negligible (statistically significant); suggesting indeed that the coefficient (and thus the relationship between that explanatory variable and the response variable, price) is non-zero.

On the other hand, in this summary table of our complete model, it appears that the p-value of the t-test for some coefficients is rather high, suggesting statistical insignificance. For example, the p-value of the t-test for the coefficient of the variable DTGA (which represents the distance of an apartment to a ‘good’ area) is 40.64%. Formally, this means that the probability of getting these results (from our housing dataset), assuming that the coefficient of DTGA is 0, is

40.64%. With such a high p-value, we cannot reject the null hypothesis, that is that the coefficient of DTGA is 0; which means that DTGA might have no relationship with the price. The same discussion holds for the interaction term BED*BATH, for which the corresponding t-test has a p-value of 24.02%, and -to a lesser extent- for the BATH variable, for which the corresponding t-test has a p-value of 6.87%.

It appears that all explanatory variables have a verified relationship with the price of an apartment when these variables are considered *individually*, but when these variables are combined in a model, this relationship fades away, as the p-values of the t-tests are not significant enough to prove that the coefficients are non-zero.

One possible explanation for this interesting contradiction could be 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 (included in the model) does, so including this second independent variable in the model is redundant. For instance, it is possible that the DIS variable explains all the variation explained by the DTGA variable (and potentially even more), so our inclusion of the DTGA into our model is redundant; hence the high p-value of the t-test corresponding to the coefficient of DTGA. The same possible explanation could be given for the relationship between the BED and the BATH variables.

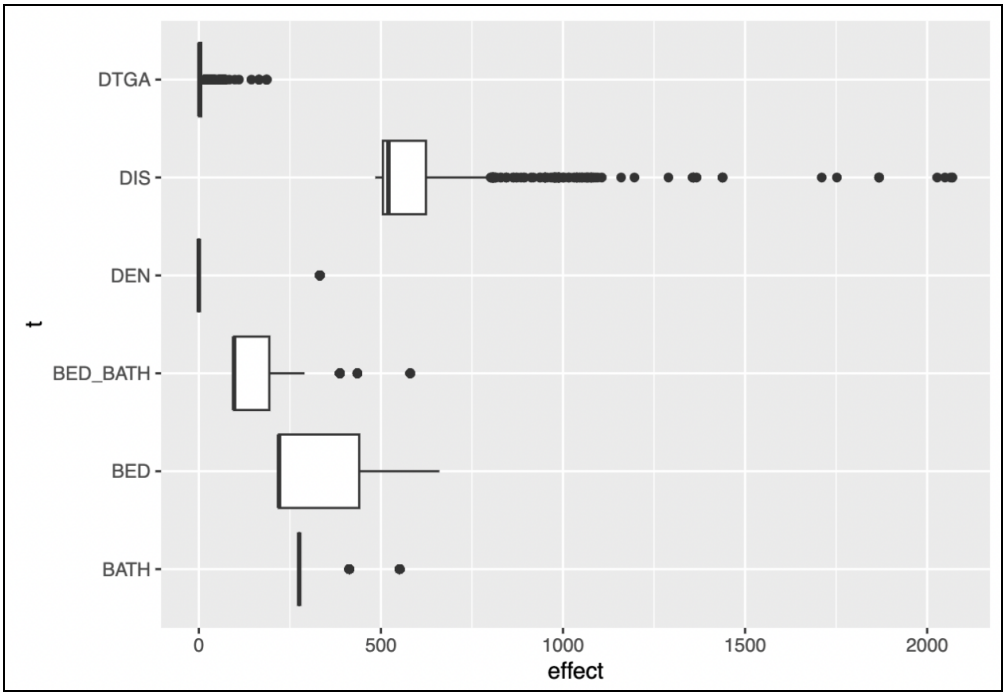
Nevertheless, based on the above summary statistics, our multivariate linear regression model as a whole explains 43.89% of the variation of the price variable in our dataset (this is the multiple R-squared, with the adjusted R-squared being only slightly smaller).

The explainability of our model is satisfactory given the limited number of variables our dataset has. This explainability could have been much if we had more information for each apartment record, such as the size of the unit (e.g. 300 square feet), the utilities included, and how old the building is.

Q2: Effects

Effect Distributions

First, will analyze the plot produced in the first part of question 2, regarding the distributions of the effects of the different explanatory variables. The plot has also been attached here for reference.



Before anything else, we should point out that the figure depicts the distributions of the "unsigned" effects of the different variables because, for the features with negative effects, we chose to take the absolute value of that effect (this process is discussed thoroughly in the *Methods* section).

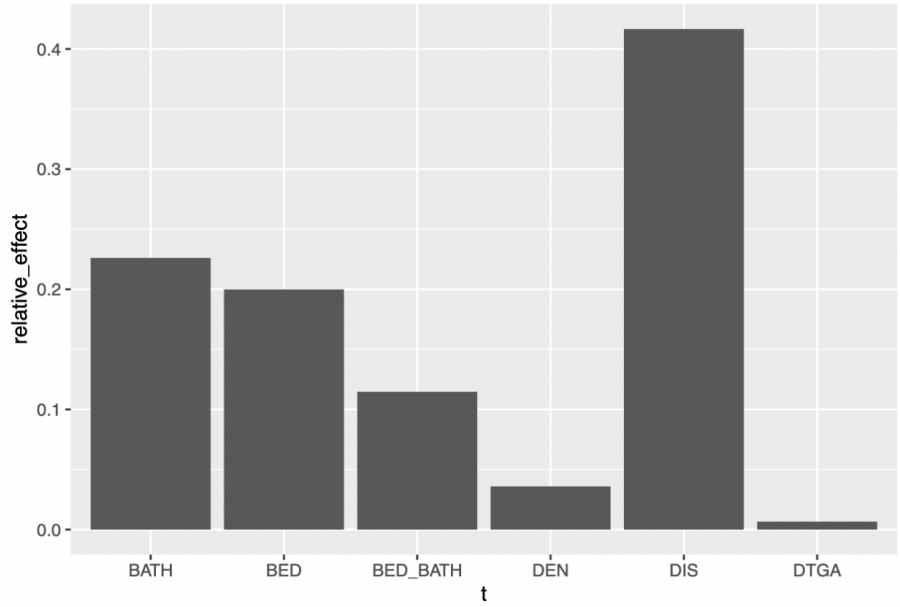
Interestingly, the (unsigned) effect distributions for all explanatory variables appear to be skewed right. This right skewness can be interpreted as follows: for each of the variables (since all distributions appear to be right-skewed), there are many observations for which the net effect of the variable 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 variable with the largest effect (out of all variables) to the prediction of a price (this assertion will be verified with the analysis of the second part, with the relative effects).

The reason for this right skewness of the effect is not clear, and any explanation we give is speculation. But one possibility that seems quite reasonable is the following: 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.

It should also be pointed out that the boxplots for the BATH, BED, and DEN variables are of limited value for the following reason: since the BATH variable has only three possible states (1 bath, 1.5 baths, or 2 baths), and the coefficient is -clearly- a constant value, there are only 3 possible effects. Same for the BED and the DEN variables, which also have a limited number of possible values.

Relative Effects

Secondly, we analyze the bar plot produced in the second part of question 2, regarding the *relative* effects of the different explanatory variables. The plot has also been attached here for reference.



As suggested by the boxplots discussed above, 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. Specifically, on average (across all observations), the DIS variable is responsible for approximately 42% of the total effect of all variables. It is followed by the BATH variables, responsible for 22% of the total effect, the BED variable, responsible for 20% of the total effect, the BED_BATH interaction term, responsible for 12% of the total effect, and then (with much lower relative effects) the DEN variable, responsible for 4% of the total effect, and finally the DTGA variable, responsible for less than 1% of the total effect.

Clearly, the DEN and the DTGA variables appear to be rather insignificant, in terms of how large their average effect is relative to the other variables. For DEN, the explanation for this insignificance could be quite straightforward. Although whether an apartment has a den is correlated to its price, it does not make a massive difference to the price of an apartment, hence the low overall effect of the DEN variable. Regarding DTGA, although one would expect the distance of an apartment to a "good" area to have a significant effect on the price of an apartment, our model disagrees. One explanation for this was that the limited information for each observation made it hard to determine which areas are "good". The low overall effect of the DTGA variable agrees with the high p-value of the t-test for the DTGA coefficient, which suggests that the DTGA variable may not be correlated with Price, or that their relation is already accounted for by other variables. Regardless, either because of the limitations of our dataset, or our own poor choice of "good areas", the average effect of the DTGA variable is negligible.

Conclusion

This project aimed to identify the factors contributing to variation in Toronto apartment rental prices and to create a model for predicting rental prices based on key variables. Taking a dataset of apartment rental prices, we recorded key variables and derived new ones to create a model to estimate prices. Through the use of multivariate linear regression and analysis of the data, we created this model and found that of our chosen factors, the primary determinant of an apartment's rental price is its proximity to the University of Toronto, followed by the number of bedrooms and bathrooms. Moving forward, future work can explore additional predictors such as square area, apartment age and size, and neighbourhood characteristics to create a more refined model and provide a more accurate prediction of apartment rental prices.

Citations

Raja CSP. (2018, November 7). *Toronto Apartment Rental Price*. Kaggle. Retrieved April 9, 2023, from https://www.kaggle.com/datasets/rajacsp/toronto-apartment-price

Acknowledgements

We would like to acknowledge Professor Speagle for aiding us by providing us with ideas and the methods necessary to create our model and calculate effect sizes.

We would also like to acknowledge our TA Adeline for her valuable support throughout the entire year, particularly in restructuring our project scope.

Appendix

As agreed to in the proposal, our group answered two research questions and had more than three visualizations. In our Progress Report, for the analysis of our first question, we mentioned the inclusion of "visualizations that plot the distance of an apartment to the nearest "central point" against price, with different graphs for each number of bedrooms, dens, bathrooms." These would be based on our linear regression model. However, because of the number of variables in our model, it was not possible for us to create visualizations that plot distance against price while still accounting for the number of rooms. To visualize this 2-dimensionally, it was necessary to create a separate visualization for each explanatory variable. However, we concluded that this would enhance our understanding of the analysis and therefore was not useful.

Individual Contribution Statements

Mimis Chlympatsos:

Although we worked collectively with my partner, my main role in the project was to produce the new variables used in our model (which were derived from the latitude and longitude data), calculate the effect sizes for the explanatory variables across all observations, and analyze the summary statistics. Regarding the final report, I mainly worked on the Methods, Results, and Discussion sections.

Throughout this project, I learned a lot of valuable lessons. First of all, the underlying relationships between different variables can be more complicated than they appear. Although we believed that our dataset consisted of enough fields (and a sufficient amount of observations) to capture and explain most of the variability in the rental prices of apartments, our final model was only capable of explaining about 44% of the variability. Thus, it appears that our dataset is missing one or more variables that could explain part of the remaining unexplained variability in the rental prices of our model. Unfortunately, a model can only be as good as the dataset on which it is based, a lesson I learned very well throughout this exploration. Moreover, I am stricken by the fact the DTGA variable seems to offer no improvement to our multivariate linear regression model since I expected that the proximity of an apartment to a "good" area would have a significant effect on the apartment's price. Perhaps a regression decision tree could have done a better job in determining how to "divide" the latitude-longitude plane representing the city of Toronto.

Daniel Du:

My particular role in the project was to help with data gathering/cleaning, creating plots to use in our analysis, and refining our model. As we were able to find time to work together quite frequently, we did not split the project into tasks for each individual to do, but rather we worked together on both questions. As a result, we both collectively contributed to the final report, both in terms of ideas and visualizations. As my partner is more knowledgeable and skilled with coding in R, most of the models/visualizations were coded by them, but aside from this, we contributed equally to the other aspects of the project.

Throughout the project, I learned that it was harder than expected to create a model. This is mainly because of two factors. Firstly, there are many confounding variables that can explain a correlation or lack thereof; frequently, a statistic may seem counterintuitive but is indeed true because of these confounding variables that exist. Secondly, it is challenging to find full, comprehensive data sets that cover all of these variables and provide information that may be valuable. Many times throughout the project, we wished that there were other variables we had data for, which would help to explain a correlation or a coefficient in our model. I believe our project and model would be much more extensive and thorough if we had a data set with more observations and more variables. We also faced challenges as our two other initial group members left the group halfway through the initial project, forcing us to begin anew. Despite this challenge, I believe our project was completed effectively.