Introduction

Our group has set out to explore the question of whether the prices on AirBnB (Toronto) are affected by certain explanatory variables to help us find faired-price rental accommodation for our future Coop terms. There are a few criteria that we are particularly interested in. First, we want to see how the location (represented in latitude) of the property affects the average listing price. Second, we also want to test how much the price will change for every increase in people the property accommodates and for every increase in bedrooms. Finally, we are also interested in whether there will be a substantial difference in price among different property types, and how much the additional parking spot is accounted in the final listing price. After building a statistically significant model, we hope to compare the prices predicted by our model to the actual price of the listing to see if we are getting a good deal. We will discuss these explanatory variables in detail in the next section.

Data set

The data set we used was found on the AirBnB Canada website. The data set contains all AirBnB listing in the Toronto area. Our first step was to filter out all the data that had little relevance to our decision criteria. Given that we were only interested in places in one particular area of Toronto, we filtered the data by longitude and latitude. Our requirements were that the listings were within the longitude range of [-79.45, -79.35], and latitude range of [43.65, 43.70]. Latitude is a measurement of how North or South a point is relative to the equator. So in our case, a latitude of 43.70 is more North than 43.65. Longitude refers to how west or east a point is. The point -79.45 is more west than -79.35.

Furthermore, other variables that were included in the data set include: the number of bedrooms and bathrooms, binary variables indicating whether or not the listings had swimming pools, parking or fireplaces. In addition, the data set had a column dedicated to how many people the place could accommodate and whether or not the property was a house, condo, or apartment.

Upon further investigation, many of these explanatory variables had no correlation with price and had no influence on our housing situations, so we decided to remove these explanatory variables to simplify the model. In addition, we used various tests to test how much variability those variables were able to explain. Given that they showed little significance to the models, we decided to eliminate those variables before conducting further model testing.

As a result, the main variables that we used in a consistent basis throughout the project include: *accommodates, beds, latitude, parking and property type*.

Accommodates: Accommodates refers to the number of people the AirBnB listing can accommodate. This was of significance to us given that we were interested to see whether or not it would be cheaper to live with more friends or individually. For example, a data point with an accommodates value of 3 means that the AirBnB listing can house 3 people at a time. This explanatory variable is discrete in nature given that it only takes in numeric data and that it can only take integer values.

Beds: This refers to the number of beds in the AirBnB accommodation. This is another important factor because it tells us how many beds are in the apartment. Once again, similar to the accommodates explanatory variable, beds is a discrete variable because it only takes integer values.

Latitude: Latitude is a continuous variable given that it can take on any value in the range listed above. As mentioned previously, latitude is measures how North or South a property is in comparison to the equator. Listing with a larger latitude are more North than those with smaller latitude values.

Parking: Parking is a binary explanatory variable. It is either True or False. True means that the listing has on-site parking, and False means that there is no on-site parking. This may be of interest to us, as parking sometimes includes the value of the listing.

Property type: Property type is a categorical variable. A listing can be one of apartment, condominium or other. This may be of importance for us, as it may indicate whether or not one property type is more expensive than another.

Price: Finally, we will be using all of these explanatory variables to create a model to predict our response variable: price. Price is a continuous variable and is the cost of the listing on a per month basis.

Analysis

The following paragraphs are a summary of our analysis and thought process in the model selection process from submission 2, with a couple of additional models.

In the previous section under data set, we mentioned that we had already reduced the number of explanatory variables to five, given that many of the original explanatory variables had little additional value to the quality of our model. We also performed pairwise correlation between the explanatory variables and concluded there are no concerns for collinearity between all explanatory variable pairs, except for # of beds and accommodates with pairwise correlation of 0.7.

Using the 5 explanatory variables, accommodates, beds, parking, latitude, and property type we built various models to test the significance of each of these variables in relation to price. Our first 2 models were just simple linear regression models involving price and one explanatory variable. The explanatory variables in the first two models were property type (model 1) and parking (model 2). We chose these two explanatory variables because they are both non-numerical, so we wanted to see if there was a relationship between these variables and price. Both of theses models explained very limited variability

given that the adjusted R^2 values were both below 0.05. (See models 1 and 2 in the appendix).

```
lm(formula = listing$price ~ listing$property_type)

Residuals:
    Min      1Q  Median      3Q     Max
-105.73  -19.73   -1.63   21.50   48.37

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       101.630      0.687 147.930  < 2e-16 ***
listing$property_typeCondominium   14.743      1.322  11.152  < 2e-16 ***
listing$property_typeOther          4.104      1.276   3.215  0.00132 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.61 on 2664 degrees of freedom
Multiple R-squared:  0.04461,   Adjusted R-squared:  0.04389
F-statistic: 62.19 on 2 and 2664 DF,  p-value: < 2.2e-16
```

```
lm(formula = listing$price ~ listing$`Parking?`)

Residuals:
     Min       1Q   Median       3Q      Max
-107.308  -18.384   -3.384   21.692   46.616

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           103.3845     0.8062  128.24  < 2e-16 ***
listing$`Parking?`TRUE   3.9234     1.0633    3.69 0.000229 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.15 on 2665 degrees of freedom
Multiple R-squared:  0.005083,  Adjusted R-squared:  0.004709
F-statistic: 13.61 on 1 and 2665 DF,  p-value: 0.000229
```

From there we decided to create a few multiple variable linear regression models. Model 3, had the explanatory variables accommodates and beds. Given that the pairwise correlation was around 0.7, we wanted to see if there were any collinearity concerns between these variables. So we built model 3, and calculated the VIF values for this model. This model explained minimal amounts of variability given that the adjusted R squared value was 0.04323 and the VIF model was 2.28. Note the R squared value of model 3 was lower than model 2. Despite the high correlation, the VIF value of 2.28 suggests that their are minimal collinearity concerns.

```
lm(formula = listing$price ~ listing$accommodates + listing$beds)

Residuals:
     Min       1Q   Median       3Q      Max
-114.927  -19.893   -1.225   20.439   52.443

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          92.3239     1.3103  70.460  < 2e-16 ***
listing$accommodates  3.6681     0.5993   6.121 1.07e-09 ***
listing$beds          1.5646     1.0341   1.513     0.13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.62 on 2664 degrees of freedom
Multiple R-squared:  0.04395,   Adjusted R-squared:  0.04323
F-statistic: 61.23 on 2 and 2664 DF,  p-value: < 2.2e-16

> vif(model3)
listing$accommodates         listing$beds
            2.282807             2.282807
```

In addition, we also wanted to see if an interaction term would improve our explanation of variability. So we created Model 4 which used beds, accommodates, and an interaction term between beds and accommodates to predict price. Once again, this model explained very little variability as the adjusted R^2 value was 0.04672. Looking at the p-value for the interaction term, we see that the p-value is 0.001. This means that we reject the null-hypothesis at a 5% significance and that the interaction term may be beneficial for our model.

FInally, given that our first 4 models were not very successful in terms of the variability explained by the model, we decided to create a model that included all of our explanatory variables to see if we could find a better model. This model was called model_big. The adjusted R^2 value for this model was 0.09857 which still suggests that very little variability is being explained. However it is explaining a little more variability compared to the first 4 models. In addition, something worth noting is the p-value for the variable beds. As mentioned previously, there were concerns about collinearity between these two variables. This appears to have showed up in the model_big as the p-value for the explanatory variable beds is 0.0458. This is close to our 5% significance cut off and suggests that maybe we cannot reject the null hypothesis and that no additional variability is being explained by the explanatory variable beds.

Following this observation we made 2 new models. Model_big2 and model_big3. Model_big2 is the same as model_big without the explanatory variable beds. The adjusted R^2 value for this model was 0.9755, which is almost the same as model_big. This means that it may not be necessary to include the explanatory variable beds.

Looking at model_big3. Model_big3 is the same as model_big, but without the explanatory variable accommodates. Since beds and accommodates were potentially correlated, we wanted to determine which of the two explanatory variables would be better for the final model. The adjusted R^2 value was 0.08693, hence model_big2 was better than model_big3. Therefore if we were to keep one of beds or accommodates, we would choose accommodates because model_big2 has a larger adjusted R^2 value.

Now that we have summarized our finding from submission 2, we will apply some of the newer concepts learned to confirm our results found in submission 2.

Further Analysis Using Cp Statistics and AIC

First we will look at the Cp statistic. As seen in figure below, the Cp value for eleventh model (price = latitude parking+accommodates) is 5.811245 which is closest to the ideal value of p+1 = 4. As seen all of the other models have much higher cp statistics in reference with the p+1 value. This implies that the smaller model is still adequate and its mean square error is only a bit inflated compared to the full model. This confirms our findings from our submission 2 model testing, where we determined that the best model was the one with explanatory variables accommodates, parking, and latitude.

```
$which
   listing.latitude listing.parking_bool listing.accommodates listing.beds
1             FALSE                FALSE                  TRUE        FALSE
1             FALSE                FALSE                 FALSE         TRUE
1              TRUE                FALSE                 FALSE        FALSE
1             FALSE                 TRUE                 FALSE        FALSE
2              TRUE                FALSE                  TRUE        FALSE
2              TRUE                FALSE                 FALSE         TRUE
2             FALSE                 TRUE                  TRUE        FALSE
2             FALSE                FALSE                  TRUE         TRUE
2             FALSE                 TRUE                 FALSE         TRUE
2              TRUE                 TRUE                 FALSE        FALSE
3              TRUE                 TRUE                  TRUE        FALSE
3              TRUE                FALSE                  TRUE         TRUE
3              TRUE                 TRUE                 FALSE         TRUE
3             FALSE                 TRUE                  TRUE         TRUE
4              TRUE                 TRUE                  TRUE         TRUE

$label
[1] "(Intercept)"          "listing.latitude"     "listing.parking_bool" "listing.accommodates" "listing.beds"

$size
 [1] 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5

$Cp
 [1]  56.947675  92.832515 138.680810 165.081002  11.191743  47.359472  53.545923  56.612233  87.124176 123.003786   5.811245  10.247630
[13]  39.309751  53.322004   5.000000
```

We then performed the automatic model selection code in R with all 5 explanatory variables as potential candidates. All 3 methods, backward, forward and stepwise all arrived at the same conclusion that the model with the lowest AIC is the one that includes all explanatory variables. However, this merely serves as a reference for us as there are more factors to consider in the best model, like R squared adjusted.

```
listing$price ~ listing$property_type + listing$accommodates +
    listing$latitude + listing$`Parking?`

               Df Sum of Sq     RSS   AIC
+ listing$beds  1    2665.8 1775544 17352
<none>                      1778209 17354
```

From the figure above we can see that the model before the last step had an AIC 2 greater than the AIC for the final model when using the forward automatic model. This shows that by adding beds the model

has improved marginally, which aligns with our conclusions found earlier when comparing adjusted R^2 values for the 2 models.

```
> model_big <- lm(listing$price~listing$accommodates+listing$latitude+listing$property_type+listing$`Parking?`);AIC(model_big)
[1] 24924.54
> summary(model_big)

Call:
lm(formula = listing$price ~ listing$accommodates + listing$latitude +
    listing$property_type + listing$`Parking?`)

Residuals:
     Min       1Q   Median       3Q      Max
-117.484  -18.957   -1.419   20.824   57.216

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                   11399.4325  2018.4953   5.647 1.80e-08 ***
listing$accommodates              4.3354     0.3922  11.054  < 2e-16 ***
listing$latitude               -259.0473    46.2273  -5.604 2.31e-08 ***
listing$property_typeCondominium   13.3934     1.2982  10.317  < 2e-16 ***
listing$property_typeOther          1.8839     1.2597   1.495    0.135
listing$`Parking?`TRUE              2.5050     1.0297   2.433    0.015 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.85 on 2661 degrees of freedom
Multiple R-squared:  0.09924,   Adjusted R-squared:  0.09755
F-statistic: 58.64 on 5 and 2661 DF,  p-value: < 2.2e-16

> anova(model_big)
Analysis of Variance Table

Response: listing$price
                       Df  Sum Sq Mean Sq  F value    Pr(>F)
listing$accommodates    1   85132   85132 127.3957 < 2.2e-16 ***
listing$latitude        1   33166   33166  49.6318 2.353e-12 ***
listing$property_type   2   73666   36833  55.1187 < 2.2e-16 ***
listing$`Parking?`      1    3955    3955   5.9185   0.01505 *
Residuals            2661 1778209     668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> vif(model_big)
                         GVIF Df GVIF^(1/(2*Df))
listing$accommodates  1.036714  1        1.018192
listing$latitude      1.027531  1        1.013672
listing$property_type 1.055648  2        1.013631
listing$`Parking?`    1.034215  1        1.016963
```

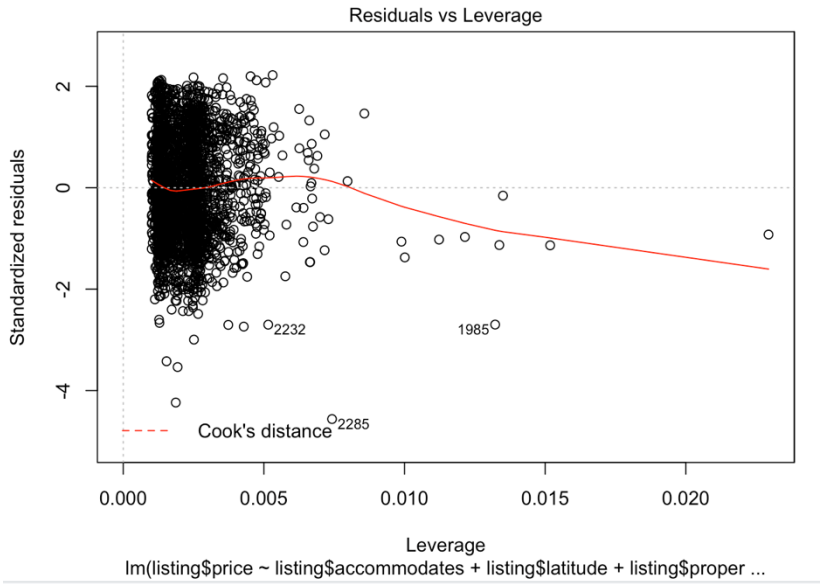
```
> model_int <- lm(listing$price~listing$accommodates+listing$beds+listing$latitude+list
ing$property_type+listing$`Parking?`+listing$accommodates*listing$beds);AIC(model_int)
[1] 24917.88
> summary(model_int)

Call:
lm(formula = listing$price ~ listing$accommodates + listing$beds +
    listing$latitude + listing$property_type + listing$`Parking?` +
    listing$accommodates * listing$beds)

Residuals:
     Min       1Q   Median       3Q      Max
-112.504  -19.385   -1.404   20.613   57.980

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     11199.8539  2018.5368   5.549 3.17e-08 ***
listing$accommodates                4.9499     0.8181   6.051 1.64e-09 ***
listing$beds                        5.2914     1.6218   3.263  0.00112 **
listing$latitude                 -254.6050    46.2259  -5.508 3.98e-08 ***
listing$property_typeCondominium    13.3406     1.2967  10.288  < 2e-16 ***
listing$property_typeOther           1.7307     1.2617   1.372  0.17027
listing$`Parking?`TRUE               2.4090     1.0285   2.342  0.01925 *
listing$accommodates:listing$beds   -0.7859     0.3049  -2.578  0.01000 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.81 on 2659 degrees of freedom
Multiple R-squared:  0.1028,    Adjusted R-squared:  0.1005
F-statistic: 43.54 on 7 and 2659 DF,  p-value: < 2.2e-16

> anova(model_int)
Analysis of Variance Table

Response: listing$price
                                 Df  Sum Sq Mean Sq  F value    Pr(>F)
listing$accommodates              1   85132   85132 127.8097 < 2.2e-16 ***
listing$beds                      1    1622    1622   2.4351   0.11877
listing$latitude                  1   33589   33589  50.4277  1.58e-12 ***
listing$property_type             2   74329   37165  55.7957 < 2.2e-16 ***
listing$`Parking?`                1    3913    3913   5.8744   0.01543 *
listing$accommodates:listing$beds 1    4426    4426   6.6444   0.01000 *
Residuals                      2659 1771118     666
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> vif(model_int)
                                     GVIF Df GVIF^(1/(2*Df))
listing$accommodates             4.524665  1        2.127126
listing$beds                     5.972011  1        2.443770
listing$latitude                 1.030807  1        1.015287
listing$property_type            1.065204  2        1.015917
listing$`Parking?`               1.035265  1        1.017480
listing$accommodates:listing$beds 11.199658  1        3.346589
```

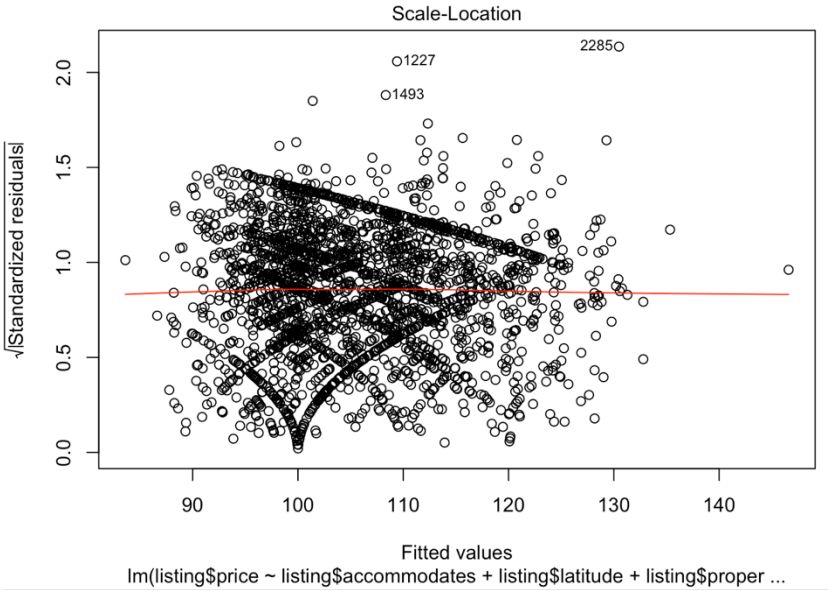
As a matter of fact, we have also considered a model with all 5 explanatory variables with the addition of an interaction term between number of accommodates and bedrooms, where we call this model_int. We compared model_int with model_big where mode_big includes all explanatory variables except # of beds for reasons we have discussed above. We see that R squared adjusted for the interaction model is 0.1005 which is slightly bigger than the corresponding R squared adjusted of 0.09755 for model_big. However, looking at the VIF for the interaction model, we see that accommodates and beds have VIF values of around 5 , which suggests multicollinearity issues, which was expected by finding pairwise correlation of around 0.7 between beds and accommodates in prior submission. This further reinforces the idea for us to use a model with all explanatory variables except # of beds as this variable conveys similar information as # of accommodates.

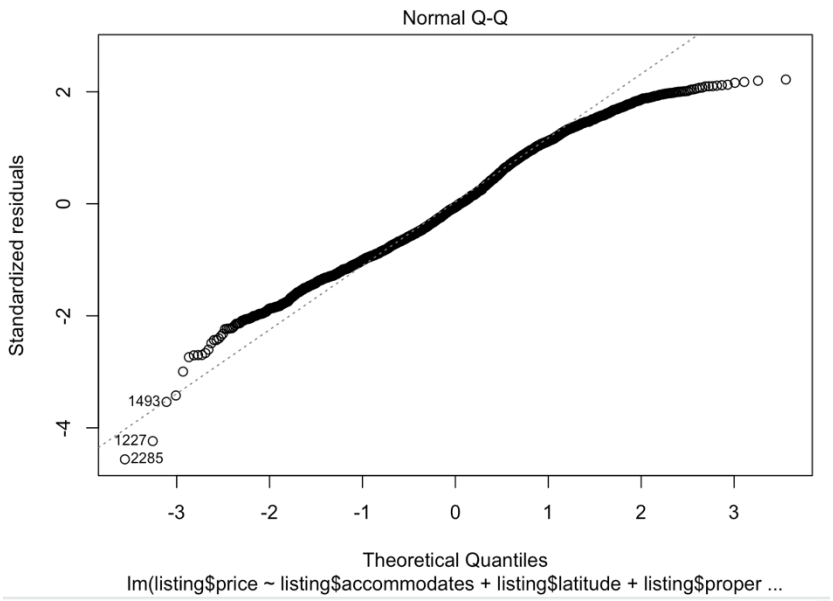
Additional Sum of Squares

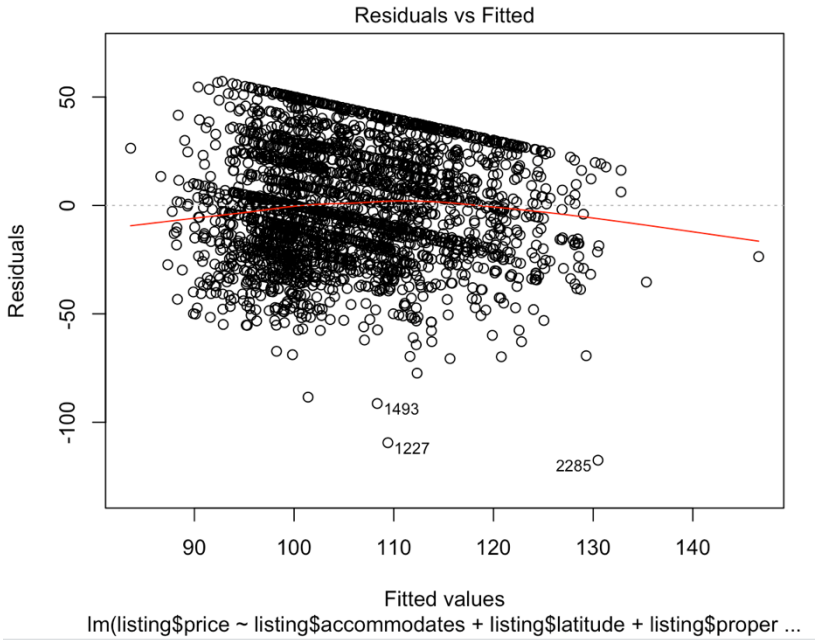
We can also quickly read off the additional Sum of Squares R has produced with our model with all 5 explanatory variables. Notice the f-value of beds given all other variables in the model is around 0.119, which using a standard 5% L.O.S suggests we do not have to reject the null hypothesis that the beta beds = 0. This shows that model_big2 is adequate in predicting price, in comparison to model_big.

Checking Model Assumptions









Normality

Above we have our 4 diagnostic plots. Checking our normality assumptions, we can conclude that the residuals are normal given that the data lies on a straight line in the QQ plot. However, on could argue that the tail found on the right most portion of the graph may signal that the normality assumption is not met. To fix this potential issue, we tried transforming by taking the log of the residuals, however this provided limited success. Next we tried transforming the residuals by taking the reciprocal, however this was also unsuccessful. WIth all that being said, we still believe that the normality assumption has been met given that the majority of the data lies on the straight line.

Mean of 0

The next assumption that must be met is that the residuals have a mean of 0. We can look fitted values vs Residuals plot to see that there is an even proportion of data points above and below the line y = 0. This means our assumption of a mean of 0 is satisfied.

Constant Variance

Next we will look at constant variance. Looking at the fitted values vs residuals plot and the scale location plots, we can conclude that the variance is constant. The red-line on the scale location plot is horizontal across the whole graph which tells us that the variance is constant. In addition, we can see a relatively constant band on the fitted values vs residual plot. (with the exception of the downwards sloping line which we will analyze in the independence section).

Independence

We can see that there is a pattern on the residuals versus fitted values plot. We can see a downwards sloping line on the plot. A simple explanation for this is that when downloading the data we filtered the data by price to eliminate some of the extra data we were seeing. Therefore, we only have this trend on the top part of the graph and not the lower part. So this is not a huge concern given that we know why the trend is occuring.

In summary, all of our model assumptions are met according to the 4 plots above.

Outlier and Leverage Analysis

With regards to leverage, we can use the 2 time h bar cut off to determine if any of the points have high leverage. Two times h bar for our model is equal to 2(4 + 1)/2667 = 0.00374. Therefore, we can see that we have around 20-30 points with high leverage. We now move on to see if these high leverage points are also outliers to conclude whether we have any influential points. In fact, looking at the scale location plot, we see that all our observations have square root standardised residual of under 3, so we don't have any outliers in the y direction.According to our Residual vs leverage plot from the section above, the line for Cook's distance is not visible in our plot. That means that all of our observations are within the boundary defined by Cook's distance and hence there are no influential points.

Results

After performing analysis on the major topics covered in the course, we have determined that the best model for our data set is model_big2. Where price is predicted using the explanatory variables accommodates, parking, and property type. In addition, the model does not need an interaction term as it marginally improves R^2 value. With that being said, despite model_big2 being the best model, all of the models created were very statistically weak. With that being said, the second best model was the model with all the explanatory variables used in our analysis. The two models explained roughly the same amount of variability so we decided to choose the smaller model. Furthermore, these decisions were confirmed by the Cp statistic and additional sum of squares concepts. Our final regression equation is $Y = 11399.4325 + 4.3354x_1 - 259.0473x_2 + 13.3934x_3 + 1.8839x_4 + 2.5050x_5$, where each of the x's corresponds to the variables found in the table on the next page. If we were to make assumptions about AirBnB listing prices based on our model, model_big2, the assumptions would be as follows:

1. An apartment that accommodates 0 people, is at 0 degrees latitude, and has no parking would cost 11399.4325 per month.

2. WIth all other variables fixed, each additional person that the listing accommodates increases the listing price by 4.3354.
3. Properties closer to the equator are cheaper. In the context of Toronto, prices that are closer to the South of Toronto (downtown) are more expensive. ($259.04 more for every unit of latitude increase, with all other variables fixed)
4. Condominiums are a little more expensive than apartments. ($13.39 per month, with all other variables fixed)
5. Other properties are generally the same price as apartments, with everything else remaining fixed. (with all other variables fixed)
6. Finally, parking costs an extra $2.51 per month with everything else being fixed.

```
Coefficients:
                                  Estimate 
(Intercept)                     11399.4325
listing$accommodates                4.3354
listing$latitude                 -259.0473
listing$property_typeCondominium   13.3934
listing$property_typeOther          1.8839
listing$`Parking?`TRUE              2.5050
---
```

With all of these conclusion, if we are looking for the cheapest option, the biggest factor we should consider is how close the listing is relative to downtown and whether the listing is an apartment, condominium, or other. For the best value, we should live in an apartment, or any other property type in the other category.

Limitations of study and conclusion

After receiving our underwhelming results and lack of success, we realized a couple of things. To predict real estate rental prices, our model requires more explanatory variables that were not included in our data set. Potential explanatory variables that may have been useful for our model include, neighbourhood score, walking scores, transit scores, builder, seasonality etc. These are all crucial explanatory variables that we believe could've helped our model. In hindsight, we realized that comparing the number of bedrooms in a listing would give us very little info given that location is one of the largest factors that impacts price. In other words, our model was missing variables that had good correlation with price. All of our explanatory variables had effects on price, but they were minimal. In addition, as we mentioned earlier, we have filtered the dataset to work with data that fits our minimum preference, for example, it is simply not feasible for us to rent a place with more than $ 1,000 per month. Hence, this placed a cap on our response variable and might have prevented us from creating the most statistically significant model for the entire dataset.

All in all, our group learned a lot of valuable things from this project. There are many factors that must be considered when building complex models to predict certain things.

Appendix

R-code used for all model testing and analysis.

```r
install.packages("car",dependencies=TRUE)
library("carData")
library("car")
library(readxl)

## 3 - introduction of dataset, variables, and summaries
#change directory
listing <- read_excel("Desktop/Toronto AB final.xlsx")
View(listing)
summary(listing)

#this allows R to recognize qualitative variables
listing$property_type = factor(listing$property_type)
listing$`Parking?`=factor(listing$`Parking?`)

#variable choice
# - lattitude
# - beds
# - accommodates
# - property type
# - Parking

#variable summary
summary(listing$price)
summary(listing$latitude)
plot(listing$price~listing$latitude)
summary(listing$beds)
boxplot(listing$price~listing$beds)
as.data.frame(table(listing$beds))
histogram(listing$beds)
summary(listing$accommodates)
as.data.frame(table(listing$accommodates))
```

```r
boxplot(listing$price~listing$accommodates)
summary(listing$property_type)
boxplot(listing$price~listing$property_type)
summary(listing$`Parking?`)
boxplot(listing$price~listing$`Parking?`)

quan <- data.frame(price = listing$price, accommodates = listing$accommodates,
                   beds = listing$beds, latitute = listing$latitude)
View(quan)
pairs(quan,main="Simple Scatterplot Matrix")
splom(quan, panel = panel.smoothScatter,xaxt="n",ann=FALSE)

model1 <- lm(listing$price~listing$property_type)
summary(model1)
anova(model1)

model2 <- lm(listing$price~listing$`Parking?`)
summary(model2)
anova(model2)

model3 <- lm(listing$price~listing$accommodates+listing$beds)
summary(model3)
vif(model3)

anova(model3)

model4 <- lm(listing$price~listing$accommodates+listing$beds+
               listing$beds*listing$accommodates)
summary(model4)
anova(model4)

model_big <- lm(listing$price~listing$accommodates+listing$beds+
```

```
                  listing$latitude+listing$property_type+listing$`Parking?`)
summary(model_big)
anova(model_big)
vif(model_big)

model_big2 <- lm(listing$price~listing$accommodates+listing$latitude+
                   listing$property_type+listing$`Parking?`)
summary(model_big2)
anova(model_big2)
vif(model_big2)

model_big4 <- lm(listing$price~listing$accommodates+listing$latitude+
                   listing$property_type+listing$`Parking?`+
                   listing$beds*listing$accommodates)
summary(model_big4)
anova(model_big4)
vif(model_big4)

model_big3 <- lm(listing$price~listing$beds+listing$latitude+
                   listing$property_type+listing$`Parking?`)
summary(model_big3)
anova(model_big3)
vif(model_big3)

plot(model_big2)
leveragePlots(model_big2)

model_int <- lm(listing$price~listing$accommodates+listing$beds+
                  listing$latitude+listing$property_type+listing$`Parking?`+
                  listing$accommodates*listing$beds)
summary(model_int)
anova(model_int)
listing$logprice <- log(listing$price)
View(listing)
model_trans1 <- lm(listing$logprice~listing$accommodates+listing$beds+
                     listing$latitude+listing$property_type+listing$`Parking?`)
summary(model_trans1)
plot(model_trans1)
listing$invprice <- 1/listing$price
model_trans2 <- lm(listing$invprice~listing$accommodates+listing$beds+
                     listing$latitude+listing$property_type+listing$`Parking?`)
summary(model_trans2)
plot(model_trans2)

model_acc <- lm(listing$accommodates~listing$beds)
summary(model_acc)
model_beds <- lm(listing$beds~listing$accommodates)
summary(model_beds)

#automatic model selection results
step(lm(listing$price~1), scope=~listing$accommotdates+listing$beds+
       listing$latitude+listing$property_type+listing$`Parking?`,direction = "forward")
step(lm(listing$price~listing$accommodates+listing$beds+listing$latitude+
          listing$property_type+listing$`Parking?`),direction="backward")
step(lm(listing$price~listing$accommodates+listing$beds+listing$latitude+
          listing$property_type+listing$`Parking?`), scope=~listing$accommodates+
       listing$beds+listing$latitude+listing$property_type+listing$`Parking?`,
     direction = "both")
#automatic model selection with interaction term
step(lm(listing$price~1), scope=~listing$accommodates+listing$beds+
       listing$latitude+listing$property_type+listing$`Parking?`+
       listing$beds*listing$accommodates,direction = "forward")
step(lm(listing$price~listing$accommodates+listing$beds+listing$latitude+
          listing$property_type+listing$`Parking?`+listing$beds*listing$accommodates),
```

```
     direction="backward")
step(lm(listing$price~listing$accommodates+listing$beds+listing$latitude+
          listing$property_type+listing$`Parking?`+listing$beds*listing$accommodates),
     scope=~listing$accommodates+listing$beds+listing$latitude+
       listing$property_type+listing$`Parking?`+listing$beds*listing$accommodates,
     direction = "both")

#analyze different test stats
#AIC, BIC, R^2 adj, MSE, Cp stat, PRESS

View(listing)
listing$parking_bool = 0
listing$parking_bool[listing$`Parking?`==TRUE] = 1
library(leaps)
x <- data.frame(listing$latitude,listing$parking_bool,listing$accommodates,
                listing$beds)
b <- leaps(x=x,y=listing$price,names=names(x))
b
```

Project done by : Brent Huang, Kyumin Shim and Ian Cheung