Is My Off-Campus Apartment a Good Deal?

Statistical Analysis of Factors Affecting U.S. and Boston Housing Rental Prices

Chenyu Cui, Phil Lee, Ning Li, Lu Qian, Wanning Zhou, Bowen Ma

Northeastern University Spring 2022 MATH 7343 Group C

Abstract:

Real estate is a great way to accumulate and grow personal wealth, but it can also lead to financial trouble if the decision is not coupled with research. Fortunately, we as mathematicians have the ability to analyze numerical data to increase our chances of investing in profitable properties with the highest probability of increasing our personal wealth.

This project aims to analyze and rank factors affecting rental prices of real-estate property across the continental U.S., based on a collection of listings from 2020, with special attention given to the greater Boston area.

Objective:

This study aims to employ statistical methods to rank factors affecting rental prices across U.S. cities with an additional analysis case dedicated to the greater Boston area, and to establish a reliable regression model to predict fair rent prices given a property's features.

Price of rent in dollars/month serves as the main dependent variable, and will be analyzed for correlation with various factors including property type, number of beds, number of baths, square feet footprint, dog permittance, cat permittance, wheelchair access, electric vehicle plug availability, and furnishings.

The resulting degrees of correlation will be ranked to draw a statistical inference on features most likely to yield high housing rent prices across the U.S., and finally a regression model to predict fair rental prices to assist students in their apartment searches next semester.

Study Design:

The study aims to ultimately estimate and rank correlation coefficients between rent prices vs. various factors.

Two study cases were conducted; one for the entire continental U.S., and a subset dedicated to the greater Boston area.

Descriptive Statistics was initially employed to perform preliminary analysis. Unfiltered data was sorted into categories by property types at first, with means, standard deviations, and 95% confidence intervals.

Based on the results of the descriptive statistics, property types deemed similar were tested for similarity via t-tests, followed by ANOVA, and any post-hoc tests if the means were deemed sufficiently different via ANOVA.

We then plotted the results to establish an overall pattern, and to identify any obviously anomalous data, which were then removed.

To determine which variables were most significantly correlated to the price of rent, Spearman's correlation coefficient method was chosen, primarily due to its resistance to outliers when compared to Pearson's method. The resultant correlation coefficients were then ranked in the order of significance on its impact on rent price.

Finally, using the resulting data, we constructed a linear regression model to predict fair rental prices based on a hypothetical property's specifications. Our explanatory variables consisted of both continuous (i.e. number of bedrooms) and categorical (pets allowed/disallowed) types - therefore, a multiple regression model was employed.

Assumptions:

Given our sample size of $n_{total} = 367{,}247$, this sample is considered normally distributed, and therefore nonparametric analysis (i.e. Mann-Whitney, Wilcoxon…etc) was *not* employed.

Each sample is taken from individual listings, therefore our sample is independent and identically distributed (i.i.d.).

Given the source of our data, each sample is independent and identically distributed (i.i.d.) with columns of variables populated for each; therefore, our study is considered a completely randomized design (as opposed to random block design). Although we could employ random block design, we did not see a need to do so as using completely randomized design would result in a study with higher resolution of error definition.

A significance level of $\alpha = 0.05$ was assumed.

Data Analysis

Initial Data Reduction – U.S. Continental

First, data was sorted by housing types, and simple descriptive statistics were computed to establish a pattern of our data.

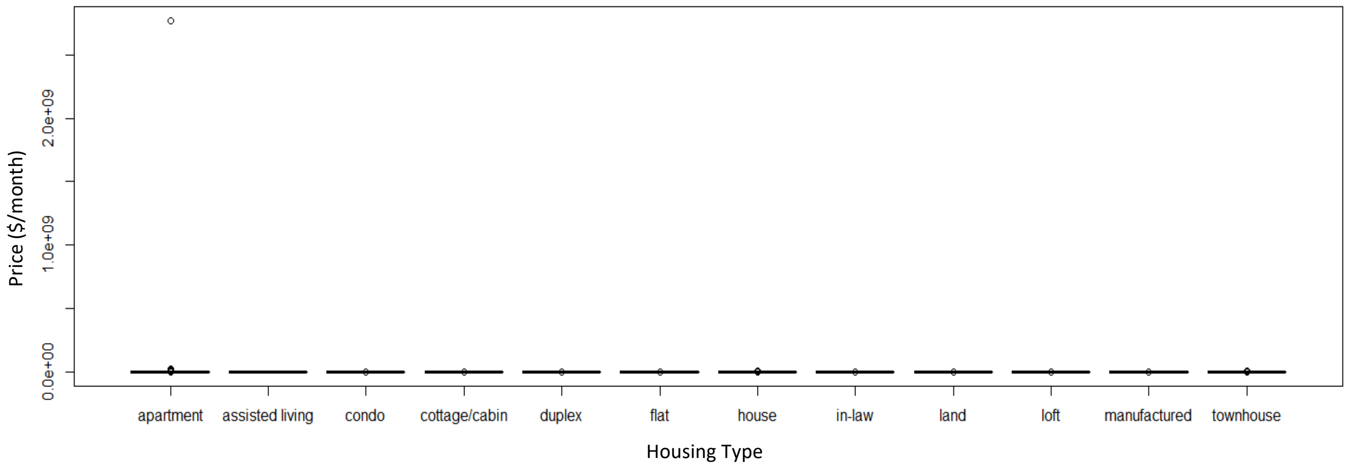


Figure 1: boxplot of housing price by type, U.S. Overall, unfiltered data

As witnessed by the above boxplot, some of our data was unreliable, such as an apartment costing 2+ million dollars, or houses with negative numbers of bedrooms. We continued to plot similar overall datasets to reduce our data into usable groups, and concluded on the following:

1. All properties with variable "price" lower than 100, or greater than 10,000 were removed.
2. All properties with variable "square feet" less than 150, or greater than 5000 were removed.
3. All properties with the variables "number of beds" and "number of baths" less than or equal to 0, and greater than or equal to 1000 were removed.

The resultant data's descriptive statistics - mean, standard deviation, and 95% confidence intervals – is plotted as follows:

Housing Type	**Mean**	**Standard Deviation**	**95% Confidence Interval**		
			Lower Bound	**Upper Bound**	**Range**
Overall	1192.198	878.4851	2366.87	2456.425	89.555
Apartment	1160.792	536.9006	1158.877	1162.707	3.83
Assisted Living	1787.5	2280.419	-18701.26	22276.26	40977.52
Condominium	1595.137	854.7913	1573.513	1616.762	43.249
Cottage	1279.6	707.5114	1225.441	1333.759	108.318
Duplex	1230.995	603.004	1214.172	1247.818	33.646
Flat	1597.846	860.9716	1522.648	1673.043	150.395
House	1380.349	887.0839	1370.713	1389.98	19.267
In-Law	1330.08	497.3621	1246.049	1414.112	168.063
Land	530	144.0486	351.14	708.86	357.72
Loft	1376.787	700.3527	1321.506	1432.067	110.561
Manufactured	917.2199	345.4665	906.6564	927.7834	21.127
Townhouse	1286.374	602.8269	1276.94	1295.808	18.868

Figure 2: Descriptive Statistics, U.S. Overall: mean, standard. Deviation, and 95% confidence interval.

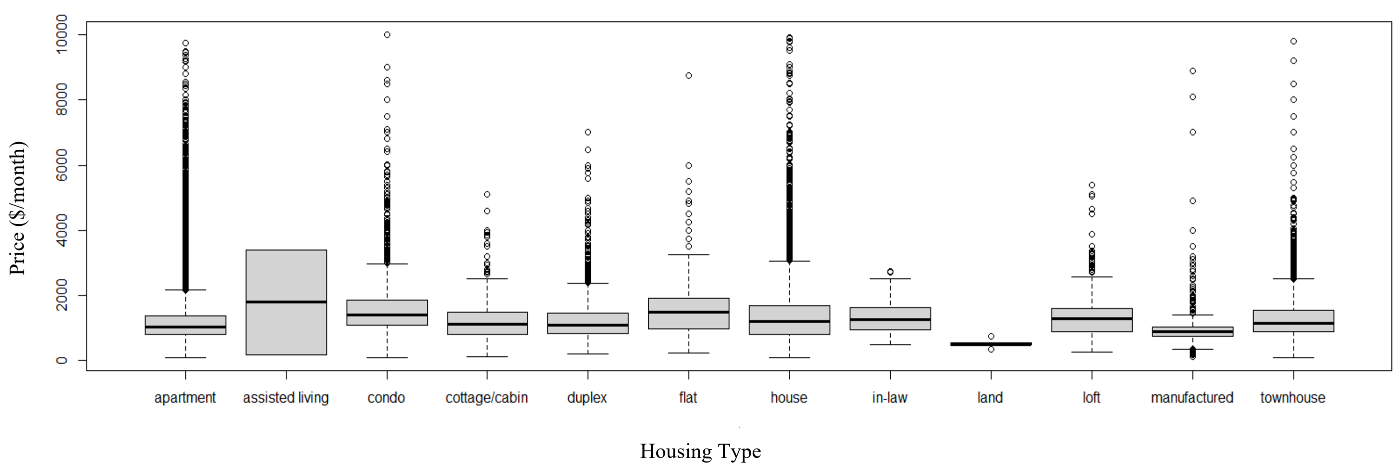


Figure 3: Boxplots, U.S. Overall, unfiltered data: House Price vs. Type

The 95% confidence intervals tell us that we are 95% confident these limits cover the true population mean for each variable. The interval that has the smallest range is Apartment at 3.83. That indicates the standard deviation of the Apartment should also be the smallest, but this does not match what we have for the sample means and standard deviations. Assisted living has the largest range of confidence interval, the difference is 40977.52 which indicates that the standard deviation of Assisted living should be the largest.

T-test Between Similar Point Estimates

Our analysis suggested that three pairs of housing types yielded similar means, to which we deemed a t-test would be appropriate:

1. Condo vs. flat (1595.137 and 1597.846, respectively)
2. Cottage vs. townhouse (1279.6 and 1286.374, respectively)
3. House vs. loft (1380.349 and 1376.787, respectively)

The results of the t-test are as follows:

1. Condo vs. Flat

   Null hypothesis H_0: $\mu_{condo} = \mu_{flat}$, vs.

   Alternative hypothesis H_a: $\mu_{condo} \neq \mu_{flat}$

   The p-value is 0.9458 which is larger than 0.05, so we failed to reject the null hypothesis at 0.05 level of significance.

2. Cottage vs. Townhouse

   Null hypothesis H_0: $\mu_{cottage} = \mu_{town\ house}$, vs.

   Alternative hypothesis H_a: $\mu_{cottage} \neq \mu_{town\ house}$

   The p-value is 0.8089 which is larger than 0.05, so we failed to reject the null hypothesis at 0.05 level of significance.

3. House vs. Loft

   Null hypothesis H_0: $\mu_{house} = \mu_{loft}$,, vs.

   Alternative hypothesis H_a: $\mu_{house} \neq \mu_{loft}$

   The p-value is 0.9008 which is larger than 0.05, so we failed to reject the null hypothesis at 0.05 level of significance.

Therefore, the means of the housing types of each test were equal within $\alpha = 0.05$ level of significance.

Analysis of Variance (ANOVA) Test

Similarly, mean price for two sets of three property types were similar, which we deemed fit for ANalysis Of VAriance (ANOVA) test.

1. Cottage, duplex, and townhouse (1279.6, 1230.995, 1286.374, respectively)
2. House, in-Law, and loft (1380.349, 1330.08, 1376.787, respectively)

The result of ANOVA are as follows:

1. Null hypothesis H_0: $\mu_{cottage} = \mu_{duplex} = \mu_{town\ house}$, vs.

   Alternative hypothesis H_a: at least one of the means is different

   Since the p-value is 1.54e-07 which is less than 0.05, we reject the null hypothesis at 0.05 level of significance.

   Post-Hoc Test: Tukey's

   Since we rejected the null hypothesis, we employed Tukey's Honestly Significant Different method to determine which variable was significantly different, which yielded the following result:

   1. p-value for duplex vs. cottage/cabin: 0.1298737
   2. p-value for townhouse vs. cottage/cabin: 0.9574927
   3. p-value for townhouse vs. duplex: 0.000

   => the difference between townhouse and duplex is significant

2. Null hypothesis H_0: $\mu_{inLaw} = \mu_{loft} = \mu_{house}$, vs.
   Alternative hypothesis H_a: at least one of the means is different

   Since the p-value is 0.798 which is greater than 0.05, we failed to reject the null hypothesis at 0.05 level of significance.

Therefore, we conclude that within α = 0.05 level of significance, there is significant variance among cottage, duplex, and townhouse, with the specific significant difference between townhouse and duplex via Tukey's HSD; however, there is insignificant variance between means of House, in-Law, and loft.

Correlation, Rental Price vs. Independent Variables

To determine the most important factors that have an impact on rental prices, we applied correlation analysis. We ran the spearman correlation test with each set of continuous variables and rental price. The categorical variables, such as cats allowed, smoking allowed, were transferred to numerical variables with 0 means not allowed and 1 means allowed. The null hypothesis for each test is that there is no linear relationship between x and y. Null hypothesis: H_0: $\rho = 0$.

The test results are as shown below:

x_i, y_i	ρ	p-value
Square Feet and Price	-0.07924597	2.2e-16
# of Bedroom and Price	0.1978706	2.2e-16
# of Bathroom and Price	0.3012577	2.2e-16
Cats allowed and Price	-0.007657359	476e-06
Dogs allowed and Price	-0.004346559	0.008437
Smoking allowed and Price	-0.1606882	2.2e-16
Wheelchair access and Price	0.06686702	2.2e-16
Electric vehicle and Price	0.09289934	2.2e-16
Comes furnished and Price	-0.006815796	3.62e-05

Figure 4: Correlation Summary, U.S. Overall: Price vs. Independent Variables

In order to get the factors that have strongest linear relationship with rental price, we ranked absolute values of correlation coefficient as followed:

x_i, y_i	ρ	Rank
# of Bathroom and Price	0.3012577	1
# of Bedroom and Price	0.1978706	2
Smoking allowed and Price	0.1606882	3
Electric vehicle and Price	0.09289934	4
Square Feet and Price	0.07924597	5
Wheelchair access and Price	0.06686702	6
Cats allowed and Price	0.007657359	7
Comes furnished and Price	0.006815796	8
Dogs allowed and Price	0.004346559	9

Figure 5: Correlation Summary, U.S. Overall: Price vs. Independent Variables, Ranked

As a result, the number of bathrooms has the most impact on rental prices for listings over the US.

Multiple Regression Model

For the multiple linear regression model equation of the continental U.S., we first plotted all variables to determine which ones could be eliminated, to simply our model. Here is the results of our initial assessment:

```
> summary(full_weight_model)

Call:
lm(formula = price ~ sqfeet + beds + baths + cats_allowed + dogs_allowed +
    smoking_allowed + wheelchair_access + electric_vehicle_charge +
    comes_furnished, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2403.0  -324.1  -105.0   200.6  8518.6

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              8.895e+02  4.275e+00 208.073  < 2e-16 ***
sqfeet                   1.858e-02  2.451e-03   7.578 3.53e-14 ***
beds                     2.437e+01  1.478e+00  16.490  < 2e-16 ***
baths                    2.673e+02  2.115e+00 126.387  < 2e-16 ***
cats_allowed            -1.085e+00  4.511e+00  -0.241     0.81
dogs_allowed            -5.480e+01  4.417e+00 -12.407  < 2e-16 ***
smoking_allowed         -1.881e+02  2.146e+00 -87.637  < 2e-16 ***
wheelchair_access        6.820e+01  3.592e+00  18.983  < 2e-16 ***
electric_vehicle_charge  5.808e+02  8.372e+00  69.374  < 2e-16 ***
comes_furnished         -1.817e+01  4.563e+00  -3.982 6.83e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 555.6 on 367237 degrees of freedom
Multiple R-squared:  0.122,	Adjusted R-squared:  0.122
F-statistic:  5670 on 9 and 367237 DF,  p-value: < 2.2e-16
```

Figure 6: Multiple Regression Model Output, U.S. Overall, All Variables

We removed the cat_allowed variable since the p-value of this parameter 0.81, indicating the variable’s low correlation to the response variable “price” – the resulting regression model’s R^2 was unchanged after its removal, further supporting our assumption.

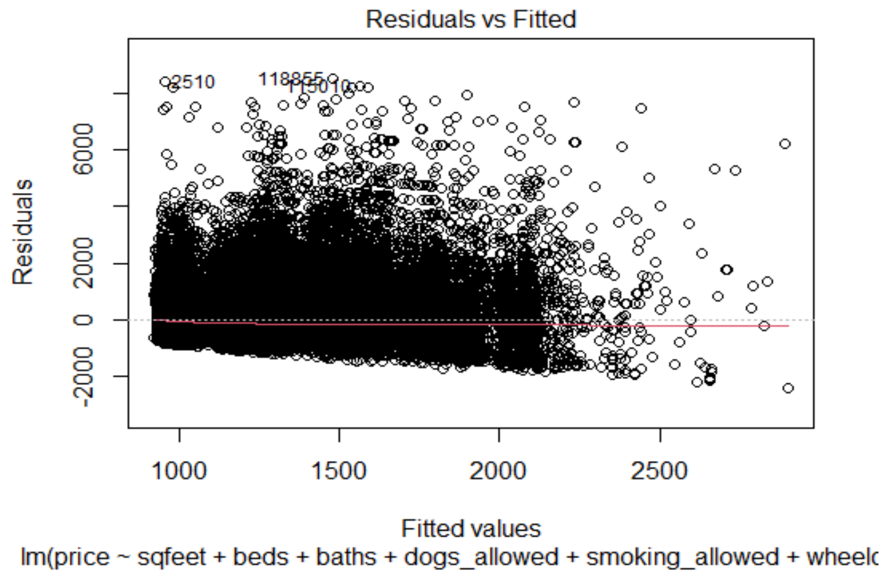
```
> summary(step.model1)

Call:
lm(formula = price ~ sqfeet + beds + baths + dogs_allowed + smoking_allowed +

    wheelchair_access + electric_vehicle_charge + comes_furnished,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2403.1  -324.0  -105.1   200.6  8518.7

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              8.893e+02  4.214e+00 211.042  < 2e-16 ***
sqfeet                   1.858e-02  2.451e-03   7.578 3.53e-14 ***
beds                     2.439e+01  1.477e+00  16.517  < 2e-16 ***
baths                    2.674e+02  2.114e+00 126.455  < 2e-16 ***
dogs_allowed            -5.574e+01  2.045e+00 -27.258  < 2e-16 ***
smoking_allowed         -1.881e+02  2.146e+00 -87.640  < 2e-16 ***
wheelchair_access        6.817e+01  3.591e+00  18.983  < 2e-16 ***
electric_vehicle_charge  5.808e+02  8.372e+00  69.376  < 2e-16 ***
comes_furnished         -1.811e+01  4.556e+00  -3.975 7.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 555.6 on 367238 degrees of freedom
Multiple R-squared:  0.122,	Adjusted R-squared:  0.122
F-statistic:  6378 on 8 and 367238 DF,  p-value: < 2.2e-16
```

Figure 7: Multiple Regression Model Output, U.S. Overall, Reduced Variables

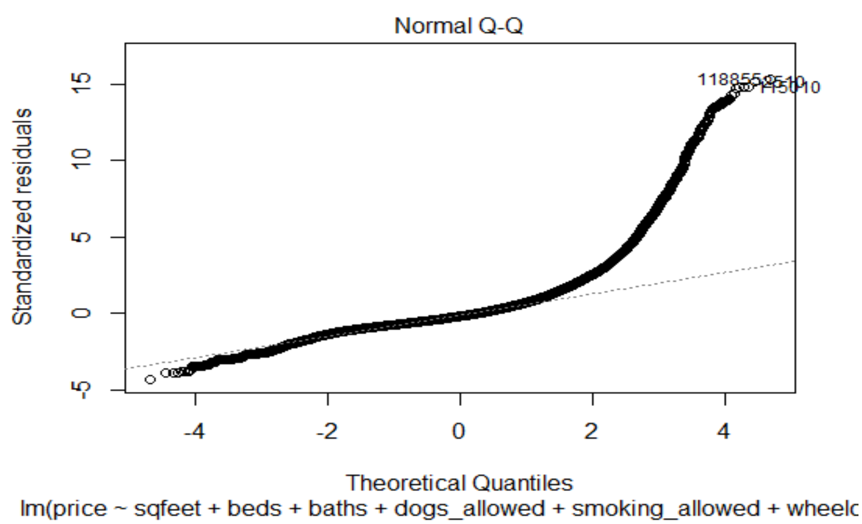
The resulting linear regression model is as follows:

$$\begin{aligned}\boldsymbol{priceUS} &= \mathbf{889.3 + 0.01858} * \boldsymbol{sqfeet} + \mathbf{24.39} * \boldsymbol{beds} + \mathbf{267.4} * \boldsymbol{baths} - \mathbf{55.74} \\ &* \boldsymbol{dogs\ allowed} - \mathbf{188.1} * \boldsymbol{somking\ allowed} + \mathbf{68.17} \\ &* \boldsymbol{wheelchair\ access} + \mathbf{580.8} * \boldsymbol{electric\ vehicle\ charge} - \mathbf{18.11} \\ &* \boldsymbol{comes\ furnished}\end{aligned}$$

To Assess the reliability of our model, we employed two residual plots: the Tukey-Anscombe (Residual vs. Fitted), and the Normal Q-Q (Quantile-Quantile) plots.



Figures 8: U.S. rental price Tukey-Anscombe



Figures 9: U.S. rental price Normal Q-Q plot

The first graph is the “Tukey-Anscombe” plot for the continental U.S. housing price. we can see that the data does not have a good fit. There are lots of values on both sides of the zero line.

The second graph is the "Normal Q-Q" plot. From the graph above, we can see that the points match up a straight line from values -3 to 1.7 which means the quantiles match. From values 2 to 4 the points do not align along a line since the data sets come from different distributions.

City of Boston Real Estate Data Analysis:

Rest assured, for our efforts were not in vain; we will conduct the above analysis on the rental data of that of Boston’s, with the aim to help students mathematically predict fair rental prices in their next semester’s apartment searches.

Initial Data Reduction – Boston

As with U.S. Continental data, data was sorted by housing types, and simple descriptive statistics were computed to establish a pattern of our data.

We performed our initial data reduction by the following:

1. All properties with variable “price” lower than 100, or greater than 10,000 were removed.
2. All properties with variable “square feet” less than 150, or greater than 5000 were removed.
3. All properties with the variables “number of beds” and “number of baths” less than or equal to 0, and greater than or equal to 1000 were removed.
4. Certain housing types were not available in Boston, or had very little data (for example, n-loft = 2) – these were excluded from our studies. These variables include: Assisted Living, Cottage, In-Law, Land, Loft, and Manufactured.

The resultant data's descriptive statistics - mean, standard deviation, and 95% confidence intervals – is plotted as follows:

Housing Type	**Mean**	**Standard Deviation**	**95% Confidence Interval**		
			Lower Bound	**Upper Bound**	**Range**
Overall	2411.648	878.4851	2366.87	2456.425	89.555
Apartment	2369.043	837.2677	2323.221	2414.864	91.643
Condominium	2556.846	835.7147	2368.442	2745.271	376.829
Duplex	3037.5	1807.449	1524.762	4550.238	3025.476
Flat	2376.429	364.4026	2039.412	2713.445	674.033
House	2843.434	1258.461	2568.641	3118.226	549.585
Townhouse	2500.353	816.4448	2080.573	2920.13	839.557

Figure 10: Descriptive Statistics, U.S. Overall: mean, standard. Deviation, and 95% confidence interval

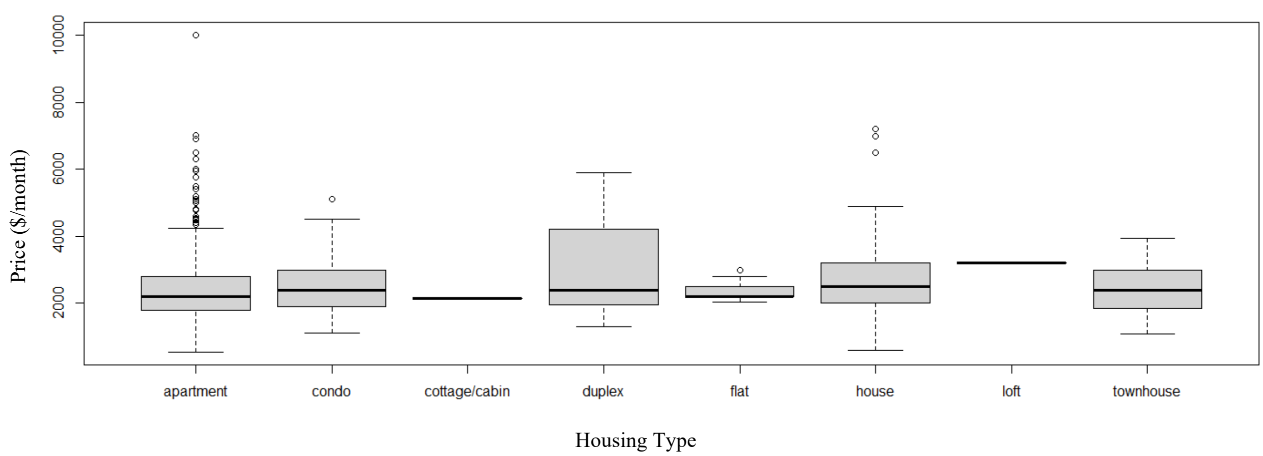


Figure 11: Boxplot of Price vs. Type, Boston

The 95% confidence intervals tell us that we are 95% confidence these limits cover the true population mean for each variable. We can see that the variable which has the smallest difference of the confidence interval is Apartment, the difference is 91.643. The variable that has the biggest difference of the confidence interval is Duplex, the difference is 3025.476 which indicates the standard deviation is the largest one also. This matches what we got for the sample means and standard deviations.

Correlation - Boston

As with U.S. Overall, rent price served as a response variable and was plotted against independent variables to determine their correlation coefficients:

x_i, y_i	ρ	p-value
Square Feet and Price	0.3932637	2.2e-16
# of Bedroom and Price	0.3952419	2.2e-16
# of Bathroom and Price	0.3708561	2.2e-16
Cats allowed and Price	0.1269025	9.614e-07
Dogs allowed and Price	0.1623481	3.295e-10
Smoking allowed and Price	0.05041759	0.0524
Wheelchair access and Price	0.07956569	0.002182
Electric vehicle charge and Price	-0.02355381	0.365
Comes furnished and Price	0.1584793	8.65e-10

Figure 12: Correlation Summary, Boston: Price vs. Independent Variables

From the p-value shown in table above, only smoking allowed and electric vehicle charge have p-values larger than 0.05, which means we fail to reject null hypotheses for these 2 factors. For the rest of the factors, the p-values are all less than 0.05. It may be concluded that there is a relationship between the other factors and rental prices. In order to get the factor that have strongest linear relationship with rental price, we ranked absolute values of correlation coefficient as followed:

x_i, y_i	ρ	Rank
# of Bedroom and Price	0.3952419	1
Square Feet and Price	0.3932637	2
# of Bathroom and Price	0.3708561	3
Dogs allowed and Price	0.1623481	4
Comes furnished and Price	0.1584793	5
Cats allowed and Price	0.1269025	6
Wheelchair access and Price	0.07956569	7

Figure 13: Correlation Summary, Boston: Price vs. Independent Variables, Ranked

Multiple Regression Model - Boston

As with the U.S. Overall regression model, we initially included all variables, then eliminated the ones deemed insignificant. Our Process is as follows:

```
> summary(full_weight_model)

Call:
lm(formula = price ~ sqfeet + beds + baths + cats_allowed + dogs_allow
ed +
    smoking_allowed + wheelchair_access + electric_vehicle_charge +
    comes_furnished, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2319.5  -451.7  -126.8   368.3  4721.1

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             852.64727   64.34120  13.252  < 2e-16 ***
sqfeet                    0.35681    0.06499   5.490 4.73e-08 ***
beds                    275.32895   27.70800   9.937  < 2e-16 ***
baths                   326.57112   44.53482   7.333 3.70e-13 ***
cats_allowed             83.93559   49.93583   1.681  0.09300 .
dogs_allowed             68.11409   52.15113   1.306  0.19173
smoking_allowed         115.12655   40.73965   2.826  0.00478 **
wheelchair_access       342.87583   87.13716   3.935 8.71e-05 ***
electric_vehicle_charge -90.53081  181.79293  -0.498  0.61857
comes_furnished         456.92335   64.67923   7.064 2.48e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 710.4 on 1471 degrees of freedom
Multiple R-squared:  0.3501,    Adjusted R-squared:  0.3461
F-statistic: 88.06 on 9 and 1471 DF,  p-value: < 2.2e-16
```

Figure 14: Multiple Regression Model Output, Boston, All Variables

By comparing the whole model and the last one, we removed the electric_vehicle_charge variable since the p-value of this parameter 0.61857, which is the biggest. The initial R^2 value is 0.3561, and the adjusted R^2 value decreased to 0.3465.

```
> summary(step.model1)

Call:
lm(formula = price ~ sqfeet + beds + baths + cats_allowed + dogs_allow
ed +
    smoking_allowed + wheelchair_access + comes_furnished, data = dat
a)

Residuals:
    Min      1Q  Median      3Q     Max
-2320.8  -451.6  -126.9   368.8  4722.8

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       850.46515   64.17542  13.252  < 2e-16 ***
sqfeet              0.35601    0.06496   5.481 4.98e-08 ***
beds              276.19933   27.64576   9.991  < 2e-16 ***
baths             327.36853   44.49466   7.357 3.10e-13 ***
cats_allowed       83.44800   49.91347   1.672  0.09477 .
dogs_allowed       68.56236   52.13004   1.315  0.18864
smoking_allowed   114.98956   40.72831   2.823  0.00482 **
wheelchair_access 339.74788   86.88828   3.910 9.64e-05 ***
comes_furnished   452.37029   64.01340   7.067 2.44e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 710.2 on 1472 degrees of freedom
Multiple R-squared:   0.35,	Adjusted R-squared:  0.3465
F-statistic: 99.08 on 8 and 1472 DF,  p-value: < 2.2e-16
```

Figure 15: Multiple Regression Model Output, Boston, Variable Reduced, Initial

Then we removed the dog_allowed variable since the p-value of this parameter 0.18864 is the second largest, the adjusted R^2 value decreased from 0.35 to 0.3462 then.

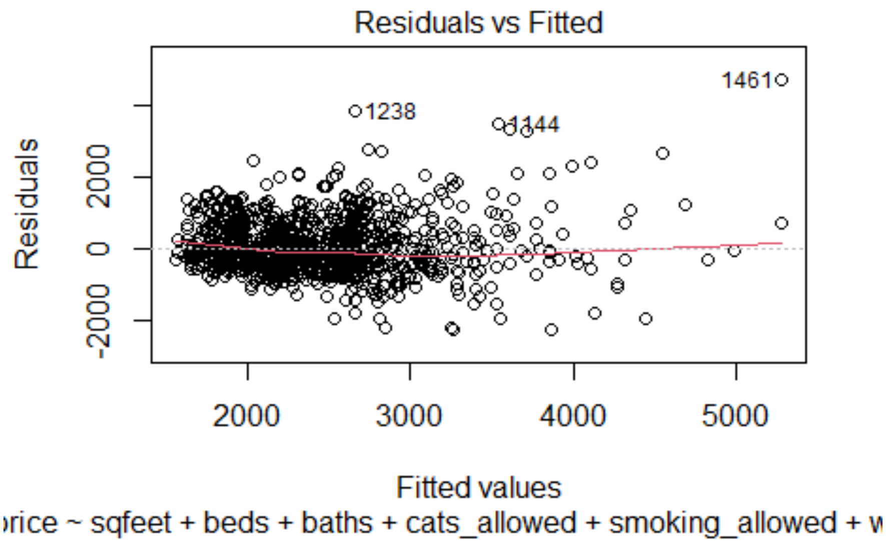
```
> summary(step.model2)

Call:
lm(formula = price ~ sqfeet + beds + baths + cats_allowed + smoking_al
lowed +
    wheelchair_access + comes_furnished, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2288.2  -458.4  -125.4   368.0  4722.8

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       844.64165   64.03835  13.190  < 2e-16 ***
sqfeet              0.35938    0.06492   5.535 3.67e-08 ***
beds              273.40820   27.57101   9.917  < 2e-16 ***
baths             333.85874   44.23112   7.548 7.70e-14 ***
cats_allowed      125.95422   38.04693   3.310 0.000954 ***
smoking_allowed   122.92883   40.28846   3.051 0.002320 **
wheelchair_access 346.38051   86.76330   3.992 6.87e-05 ***
comes_furnished   463.48403   63.46894   7.303 4.60e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 710.3 on 1473 degrees of freedom
Multiple R-squared:  0.3492,    Adjusted R-squared:  0.3462
F-statistic: 112.9 on 7 and 1473 DF,  p-value: < 2.2e-16
```

Figure 16: Multiple Regression Model Output, Boston, Variables Reduced, Final

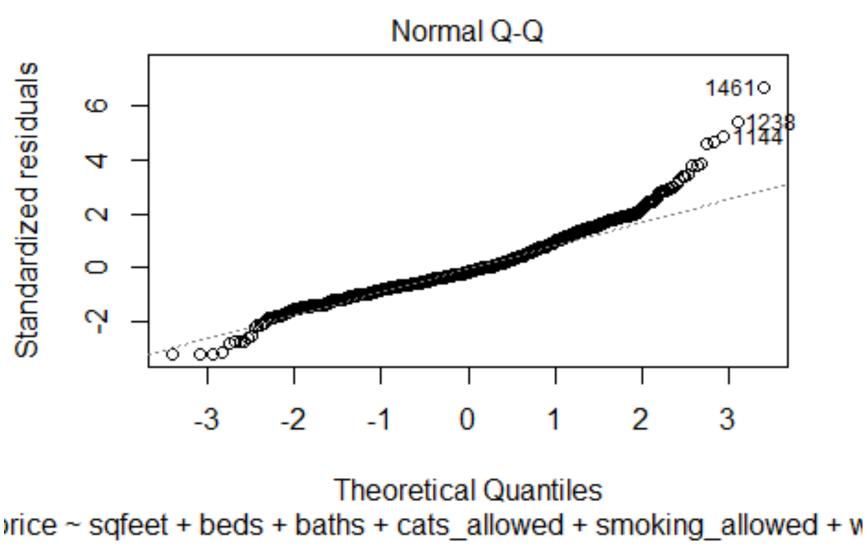
Our final regression model output is as follows:

$$\begin{aligned}&\boldsymbol{priceBoston}\\&\quad = \mathbf{844.64165 + 0.35938} * \boldsymbol{sqfeet} + \mathbf{274.40820} * \boldsymbol{beds} + \mathbf{333.85874}\\&\quad * \boldsymbol{baths} + \mathbf{125.95422} * \boldsymbol{cat\ allowed} + \mathbf{122.92883} * \boldsymbol{somking\ allowed}\\&\quad + \mathbf{346.38051} * \boldsymbol{wheelchair\ access} + \mathbf{463.48403} * \boldsymbol{comes\ furnished}\end{aligned}$$

To Assess the reliability of our model, we employed two residual plots: the Tukey-Anscombe (Residual vs. Fitted), and the Normal Q-Q (Quantile-Quantile) plots



Figures 17: Boston rental price Tukey-Anscomb



Figures 18: Boston rental price Normal Q-Q plot

From our 'Tukey-Anscombe' plot (left). We can see that the data displays a good until price $\leq$ 3000, and becomes worse as it increases.

From our Quantile-Quantille plot, we can see that the data is fairly normally distributed between quantiles of -2 to +2, until the values become extreme.

From the plots, we conclude that our regression model should be fairly reliable for housing less than $3000.

Model Prediction Using Real-World Examples:

Now that we have our model, to test how well our multiple regression model is, we found recent apartment listings and fitted it against our model. We ran 2 examples to predict the rental price with specific conditions.

Example 1:

When sqfeet=1000, beds=2,bath=1,cats-allowed = 1, smoking-allowed =1, wheelchair-access = 0, comes-furnished = 0

$$\boldsymbol{Price = 844.64 + 0.3594 * 1000 + 273.40 * 2 + 333.86 * 1 + 125.95 * 1 + 122.93 * 1 = 2333.58}$$

Since there are several cases with the exact same values but different prices, we computed the mean price for these cases = 2263.33, comparing with the predicted price 2333.58, the error is 70.25 (approximate 3% error).

Example 2:

When sqfeet=498, beds=1, bath=1, cats-allowed = 0, smoking-allowed = 0, wheelchair-access = 0, comes-furnished = 0

$$\boldsymbol{Price = 844.64 + 0.3594 * 498 + 273.40 * 1 + 333.86 * 1 = 1630.88}$$

The actual price is 1850, so the error is 219.12 (approximate 11% error), which is relatively higher than the example 1.

Discussion

Our assumptions were correct for the most part – although some of our datasets consisted of small sample sizes (n_{loft} = 2, for example), these samples were removed from our final analysis to preserve the normality of our data, so the reduced sets of data actually utilized for our final analyses were all large enough ($n \geq 30$) to assume normal approximation.

Our reduced dataset is deemed as an observed dataset, as opposed to a controlled dataset.Since our dataset was created by collecting individual listings across various websites independently, we believe our data remains independent, with robust-enough randomization in sampling that is accurately representative of the statistic analyzed (Overall U.S., and Boston City).

Furthermore, our resultant regression model yielded accurate results with as low as 3% error, when tested against random, independent real-life data unused in our analysis; therefore, we believe the scope of inference for our analysis encompasses the entire population studied – U.S. Overall, and Boston city, respectively.

Conclusion

Comparison between populations

The initial data analysis shows that the mean prices for the following three pairs of housing types are close: condo and flat, cottage and townhouse, house and loft. To determine if there are significant differences between the means of each group, the two sample t-test was applied to validate null hypothesis: H_0: $\mu_1 = \mu_2$. From the calculated p-values, it is clear that we fail to reject all null hypothesis. It may be concluded that the true mean for each group: condo and flat, cottage and townhouse, house and loft, is the same.

In addition, it is observed that the mean prices for 2 triplets of housing types, group 1: cottage, duplex and townhouse, group 2: house, inLaw and loft are close. Given that the prices of 3 housing types have mean μ_1, μ_2 *and* μ_3 respectively, we would like to test the null hypothesis that they identical. We used the extension of the two-sample t-test, ANOVA test to validate null hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3$. we failed to reject the null hypothesis for the in-law, loft, and house, so we can conclude that the true means of in-law housing, loft housing, and house are the same. In the cottage, duplex, and townhouse ANOVA test, we reject the null hypothesis. Thus, we conclude that the true price means of these three housing types are different.

We then pinpoint the difference since the means are different. We use Tukey to control the FWER here. The p-value for duplex and cottage/cabin is 0.1298737; the p-value for townhouse and cottage/cabin is 0.9574927; the p-value for townhouse and duplex is 0.000 which indicates the difference between townhouse and duplex is significant.

Correlation

To determine the most important factor that affects prices in the continental US and Boston, we ran spearman correlation test for each set of numerical variables (the categorical variables were transferred to numerical with 0 indicates no and 1 indicates yes). The null hypothesis for each test is that there is no linear relationship between x and y. Null hypothesis: H_0: $\rho = 0$. Then the absolute value of correlation coefficients were ranked, and the factor with largest correlation coefficient is the most important factor. For continental US, the factor is the number of bathrooms and for Boston, the most critical factor is number of bedrooms.

<u>Multiple regression Model</u>

After running the linear regression model, we removed the parameters with largest p-value for both the continental US and Boston. The R squared value decreased a bit because of the modification. T=The final linear regression models for the continental US is:

$$\boldsymbol{priceUS = 889.3 + 0.01858 * sqfeet + 24.39 * beds + 267.4 * baths - 55.74 * dogs\ allowed - 188.1 * somking\ allowed + 68.17 * wheelchair\ access + 580.8 * electric\ vehicle\ charge - 18.11 * comes\ furnished}$$

The linear regression model for Boston is:

$$\boldsymbol{priceBoston = 844.64165 + 0.35938 * sqfeet + 274.40820 * beds + 333.85874 * baths + 125.95422 * cat\ allowed + 122.92883 * somking\ allowed + 346.38051 * wheelchair\ access + 463.48403 * comes\ furnished}$$

Team Member Synergy

We divided our team into two major groups: three members for the mathematical analysis, and three for the report.

Ning, Wanning, and Bowen performed data clean-up, analysis, designed the test procedures, and wrote the R-code - they met in-person several times and performed the analysis together, comparing results to ensure the accuracy of individual conclusions.

Chenyu, Phil, and Lu worked together with the analysis team to write this report, each member authoring a specific section of equal proportions.

All members of the team assumed personal ownership of this report and held themselves responsible for their portion(s), and each showed their utmost effort and leadership qualities dedication to the completion of this report.

Appendix

```
install.packages("readxl")
library("readxl")
data <- read_excel("/Users/apple/Desktop/MATH7343 Applied Statistics/final project/related files/housing clean.xlsx")

## calculate mean and standard deviation for all prices
meanPriceAll <- mean(data$price) #1192.198 point estimator
meanPriceAll
sdPriceAll <- sd(data$price) #592.9392
sdPriceAll
alpha <- 0.05
meanPriceAll + c(-1,1) * qt(1-alpha/2, df = length(data$price) - 1) * sdPriceAll/sqrt(length(data$price))
#--------------------------------------------------------------------------------------------------------
## calculate mean and standard deviation for each type of housing, and 95% C.I(t-interval)
apartment <- data[data$type == 'apartment',]
meanApart <- mean(apartment$price)
meanApart #1160.792
sdApart <- sd(apartment$price)
sdApart #536.9006
meanApart + c(-1,1) * qt(1-alpha/2,df = nrow(apartment)-1) * sdApart/sqrt(nrow(apartment)) #(1158.877, 1162.707)

assisted_living <- data[data$type == 'assisted living',]
meanAssi <- mean(assisted_living$price)
meanAssi #1787.5
sdAssi <- sd(assisted_living$price)
sdAssi #2280.419
meanAssi + c(-1,1) * qt(1-alpha/2,df=nrow(assisted_living)-1) * sdAssi/sqrt(nrow(assisted_living)) #(-18701.26,22276.26)

condo <- data[data$type == 'condo',]
meanCon<-mean(condo$price) #1595.137
meanCon
sdCon<-sd(condo$price)
sdCon #854.7913
meanCon + c(-1,1) * qt(1-alpha/2,df=nrow(condo)-1) * sdCon/sqrt(nrow(condo)) #(1573.513, 1616.762)

cottage <- data[data$type == 'cottage/cabin',]
meanCot <- mean(cottage$price)
meanCot #1279.6
sdCot <-sd(cottage$price)
sdCot #707.5114
meanCot+ c(-1,1) * qt(1-alpha/2,df=nrow(cottage)-1) * sdCot/sqrt(nrow(cottage)) #(1225.441, 1333.759)

duplex <- data[data$type == 'duplex',]
meanDu<-mean(duplex $price)
meanDu #1230.995
sdDu<-sd(duplex$price)
sdDu # 603.004
meanDu+ c(-1,1) * qt(1-alpha/2,df=nrow(duplex)-1) * sdDu/sqrt(nrow(duplex)) #(1214.172 ,1247.818)

flat <- data[data$type == 'flat',]
meanFlat<-mean(flat$price) #1597.846
meanFlat
sdFlat<-sd(flat$price) #860.9716
sdFlat
meanFlat+ c(-1,1) * qt(1-alpha/2,df=nrow(flat)-1) * sdFlat/sqrt(nrow(flat)) #(1522.648, 1673.043)

house <- data[data$type == 'house',]
meanHouse<-mean(house$price) #1380.349
meanHouse
sdHouse<-sd(house$price)
sdHouse #887.0839
meanHouse+ c(-1,1) * qt(1-alpha/2,df=nrow(house)-1) * sdHouse/sqrt(nrow(house)) #(1370.718 ,1389.980)

inLaw <- data[data$type == 'in-law',]
meanLaw<-mean(inLaw$price) #1330.08
sdLaw<-sd(inLaw$price) #497.3621
alpha <- 0.05
meanLaw+ c(-1,1) * qt(1-alpha/2,df=nrow(inLaw)-1) * sdLaw/sqrt(nrow(inLaw)) #(1246.049, 1414.112)

land <- data[data$type == 'land',]
meanLand<-mean(land$price) #530
sdLand<-sd(land$price) #144.0486
meanLand+ c(-1,1) * qt(1-alpha/2,df=nrow(land)-1) * sdLand/sqrt(nrow(land)) #(351.14, 708.86)

loft <- data[data$type == 'loft',]
meanLoft<-mean(loft$price) #1376.787
sdLoft<-sd(loft$price) #700.3527
meanLoft+ c(-1,1) * qt(1-alpha/2,df=nrow(loft)-1) * sdLoft/sqrt(nrow(loft)) #(1321.506 ,1432.067)

manu <- data[data$type == 'manufactured',]
meanManu<-mean(manu$price) #917.2199
sdManu<-sd(manu$price) #345.4665
meanManu+ c(-1,1) * qt(1-alpha/2,df=nrow(manu)-1) * sdManu/sqrt(nrow(manu)) #(906.6564, 927.7834)
```

```r
townhouse <- data[data$type == 'townhouse',]
meanTown<-mean(townhouse$price) #1286.374
sdTown<-sd(townhouse$price) #602.8269
meanTown+ c(-1,1) * qt(1-alpha/2,nrow(townhouse)-1) * sdTown/sqrt(nrow(townhouse)) #(1276.940, 1295.808)

#-------------------------------------------------------------------
# Compare the true mean between different housing types

# condo and flat:
t.test(condo$price, flat$price)

# cottage and townhouse
t.test(cottage$price, townhouse$price)

# house and loft
t.test(house$price, loft$price)
# -------------------------------------------------------------------------
# Compare the true mean between different housing types

# cottage, duplex, townhouse:
df1 <- rbind(cottage, duplex, townhouse)
df1$type <- as.factor(df1$type)
df1.fit <- aov(price~type, data = df1)
summary(df1.fit)
TukeyHSD(df1.fit, conf.level = 0.95)

# house, inLaw, loft:
df2 <- rbind(house, inLaw, loft)
df2$type <- as.factor(df2$type)
df2.fit <- aov(price~type, data = df2)
summary(df2.fit)
#-------------------------------------------------------------------
## Calculate correlation between all numerical variables and price

#sqfeet and prices
cor.test(data$sqfeet, data$price, method = 'spearman')

#beds and prices
cor.test(data$beds, data$price, method = 'spearman')

# baths and prices
cor.test(data$baths, data$price, method = 'spearman')

#cats_allowed and price
cor.test(data$cats_allowed, data$price, method = 'spearman')

#dogs_allowed and price
cor.test(data$dogs_allowed, data$price,method = 'spearman')

#smoking allowed and price
cor.test(data$smoking_allowed, data$price,method = 'spearman')

# wheelchair_access and price
cor.test(data$wheelchair_access, data$price,method = 'spearman')

# electric_vehicle_charge and price
cor.test(data$electric_vehicle_charge, data$price,method = 'spearman')

#comes_furnished and price
cor.test(data$comes_furnished, data$price,method = 'spearman')
#---------------------------------------------------------------------------
## Multiple Linear Regression
data$cats_allowed <- as.factor(data$cats_allowed)
data$dogs_allowed <- as.factor(data$dogs_allowed)
data$smoking_allowed <- as.factor(data$smoking_allowed)
data$wheelchair_access <- as.factor(data$wheelchair_access)
data$electric_vehicle_charge <- as.factor(data$electric_vehicle_charge)
data$comes_furnished <- as.factor(data$comes_furnished)

full_weight_model <- lm(price ~ sqfeet + beds + baths + cats_allowed
                        + dogs_allowed + smoking_allowed + wheelchair_access + electric_vehicle_charge
                        + comes_furnished, data = data)

summary(full_weight_model)

## Depending on full_weight_model, remove cats_allowed1 since its p value 0.81 is biggest, adjusted R^2 remains unchanged
step.model1 <- lm(price ~ sqfeet + beds + baths + dogs_allowed + smoking_allowed + wheelchair_access + electric_vehicle_charge
                  + comes_furnished, data = data)
summary(step.model1)
plot(step.model1, which=1)
plot(step.model1, which=2)
```

Figure 19. R commands for continental US

```r
## Data for Boston only
data <- read.csv("/Users/apple/Desktop/MATH7343 Applied Statistics/final project/related files/housing boston.csv")

## calculate mean and standard deviation for all prices
meanPriceAll <- mean(data$price) #2411.648, point estimator
sdPriceAll <- sd(data$price) #878.4851
alpha <- 0.05
meanPriceAll + c(-1,1) * qt(1-alpha/2, df = length(data$price) - 1) * sdPriceAll/sqrt(length(data$price))
#------------------------------------------------------------------------------------------------------
## calculate mean and standard deviation for each type of housing, and 95% C.I(t-interval)
apartment <- data[data$type == 'apartment',]
meanApart <- mean(apartment$price) #2369.043
sdApart <- sd(apartment$price) #837.2677
meanApart + c(-1,1) * qt(1-alpha/2,df = nrow(apartment)-1) * sdApart/sqrt(nrow(apartment)) #(2323.221 2414.864)

condo <- data[data$type == 'condo',]
meanCon<-mean(condo$price) #2556.846
sdCon<-sd(condo$price) # 835.7147
meanCon + c(-1,1) * qt(1-alpha/2,df=nrow(condo)-1) * sdCon/sqrt(nrow(condo)) #(2368.422 2745.271)

duplex <- data[data$type == 'duplex',]
meanDu<-mean(duplex $price) #3037.5
sdDu<-sd(duplex$price) #1809.449
meanDu+ c(-1,1) * qt(1-alpha/2,df=nrow(duplex)-1) * sdDu/sqrt(nrow(duplex)) #(1524.762 4550.238)

flat <- data[data$type == 'flat',]
meanFlat<-mean(flat$price) #2376.429
sdFlat<-sd(flat$price) #364.4026
meanFlat+ c(-1,1) * qt(1-alpha/2,df=nrow(flat)-1) * sdFlat/sqrt(nrow(flat)) #(2039.412 2713.445)

house <- data[data$type == 'house',]
meanHouse<-mean(house$price) #2843.434
meanHouse
sdHouse<-sd(house$price)
sdHouse #1258.461
meanHouse+ c(-1,1) * qt(1-alpha/2,df=nrow(house)-1) * sdHouse/sqrt(nrow(house)) #(2568.641, 3118.226)

loft <- data[data$type == 'loft',]
meanLoft<-mean(loft$price)
meanLoft #3200
sdLoft<-sd(loft$price)
```


meanLoft #3200
sdLoft<-sd(loft$price)


```r
sdLoft #0
meanLoft+ c(-1,1) * qt(1-alpha/2,df=nrow(loft)-1) * sdLoft/sqrt(nrow(loft)) #(3200, 3200)

townhouse <- data[data$type == 'townhouse',]
meanTown<-mean(townhouse$price)
meanTown #2500.353
sdTown<-sd(townhouse$price)
sdTown #816.4448
meanTown+ c(-1,1) * qt(1-alpha/2,nrow(townhouse)-1) * sdTown/sqrt(nrow(townhouse)) #(2080.576 ,2920.130)

#---------------------------------------------------------------------------
## Calculate correlation between all numerical variables and price
#sqfeet and prices
cor.test(data$sqfeet, data$price, method = 'spearman')

#beds and prices
cor.test(data$beds, data$price, method = 'spearman')

# baths and prices
cor.test(data$baths, data$price, method = 'spearman')

#cats_allowed and price
cor.test(data$cats_allowed, data$price, method = 'spearman')

#dogs_allowed and price
cor.test(data$dogs_allowed, data$price,method = 'spearman')

#smoking allowed and price
cor.test(data$smoking_allowed, data$price,method = 'spearman')

# wheelchair_access and price
cor.test(data$wheelchair_access, data$price,method = 'spearman')

# electric_vehicle_charge and price
cor.test(data$electric_vehicle_charge, data$price,method = 'spearman')

#comes_furnished and price
cor.test(data$comes_furnished, data$price,method = 'spearman')
#-------------------------------------------------------------------------------------
```

```r
#-------------------------------------------------------------------------------
## Multiple Linear Regression
data$cats_allowed <- as.factor(data$cats_allowed)
data$dogs_allowed <- as.factor(data$dogs_allowed)
data$smoking_allowed <- as.factor(data$smoking_allowed)
data$wheelchair_access <- as.factor(data$wheelchair_access)
data$electric_vehicle_charge <- as.factor(data$electric_vehicle_charge)
data$comes_furnished <- as.factor(data$comes_furnished)

full_weight_model <- lm(price ~ sqfeet + beds + baths + cats_allowed
                        + dogs_allowed + smoking_allowed + wheelchair_access + electric_vehicle_charge
                        + comes_furnished, data = data)

summary(full_weight_model)

## Depending on full_weight_model, remove electric_vehicle_charge1 since its p value 0.62 is biggest, adjusted R^2 increases to 0.3465
step.model1 <- lm(price ~ sqfeet + beds + baths + cats_allowed + dogs_allowed + smoking_allowed + wheelchair_access + comes_furnished, data = data)
summary(step.model1)

## Depending on step.model1, remove dogs_allowed1 since its p value 0.19 is biggest, adjusted R^2 decreased to 0.3462
step.model2 <- lm(price ~ sqfeet + beds + baths + cats_allowed + smoking_allowed + wheelchair_access + comes_furnished, data = data)
summary(step.model2)
plot(step.model2, which=1)
plot(step.model2, which=2)

```

Figure 20. R commands for Boston

Reference

A. Reese, “USA housing listings,” *Kaggle*, 17-Jun-2020. [Online]. Available: https://www.kaggle.com/datasets/austinreese/usa-housing-listings?resource=download. [Accessed: 25-Apr-2022].