Linear Regression Analysis on Hostel Prices in Japan

Liam Daly

2024-06-07

Contents

List of Figures

Introduction

Subject Matter

I'm interested in examining whether regression analysis can effectively be applied to determine which factors may influence higher hostel prices in Japan.

Research Question

Which predictors are positively linearly related with minimum nightly hostel prices in Japan?

Analysis Goal and Motivation

My main goal is to create a linear regression model which may be used to make inferences on statistically significant predictors with prices as the response variable. I'm personally interested in factors which may influence hostel prices because I'm planning a grad trip to Japan and I'm currently weighing accommodation options.

Stakeholder Interpretation

Given my subject matter, stakeholders will likely be hostel managers interested in increasing earnings. Using results of my analysis they may be able to:

- Identify key factors that justify changes to prices.
- Adjust pricing & accommodation strategy accordingly.

Data Description

This dataset "Hostel.csv" was created by Koki Ando, who scraped data from HostelWorld.com: "a global database with over 16,000 hostels." According to Ando, the data was scraped before the 2020 Tokyo Olympic games.

```
ncol(hostel)
```

```
## [1] 16
```

```
nrow(hostel)
```

```
## [1] 342
```

There are a total of 342 hostel observations included with 16 distinct variables.

Each row denotes a unique Hostel in Japan. I'll explain each variable below.

- "X": categorical integer variable to identify the row number
- "Hostel Name": categorical character variable for the name of the hostel
- "City": categorical character variable for the city the hostel is based at
- "price.from": continuous integer for minimum nightly price per night in Japanese yen
- "Distance": character variable describing how far the hostel is from the city center in km
- "rating.band": character variable variable describing how the customer rated the hostel
- "lon": hotel's longitude
- "lat": hotel's latitude

All following rating score variables are on a scale from 0.0 to 10.0. They are collective scoring variables based on a variety of criteria.

- “summary.score”: double ordinal variable for the mean the hostel’s rating score based on all criteria
- “atmosphere”: double ordinal variable for score of hostel’s atmosphere
- “cleanliness”: double ordinal variable for score of hostel’s cleanliness
- “facilities”: double ordinal variable for score of hostel’s facilities
- “locaition.y”: double ordinal variable for score of hostel’s location
- “security”: double ordinal variable for score of hostel’s security
- “staff”: double ordinal score for hostel’s staff
- “valueformoney”: double ordinal score for hostel’s value for money

I will be selecting price.from as my response/outcome variable because I believe several relevant variables may be used as predictors related to it. Additionally, summary statistics for all of my variables are available in Appendix section A1.1

Exploratory Data Analysis

Data Cleaning

Upon looking through the entire dataset, it’s clear that it wasn’t fit for analysis straight away. I began this process by removing NA’s and duplicates. I also filtered out X, hostel.name longitude, and latitude, since I had no use for them. Distance was initially listed as a character variable. For analysis, I needed it to be double so I converted it while cutting out “km from city centre.” City and rating.band only contained five categories so they were fit to include as a categorical variables.

Variable Distributions

Upon analyzing the distribution of hostel prices, I found it was severely right skewed. After this, I decided to filter out prices greater than 1000000 yen. The distribution still was right skewed, so I removed outliers over 4500 yen in order to preserve normality (A2.1).

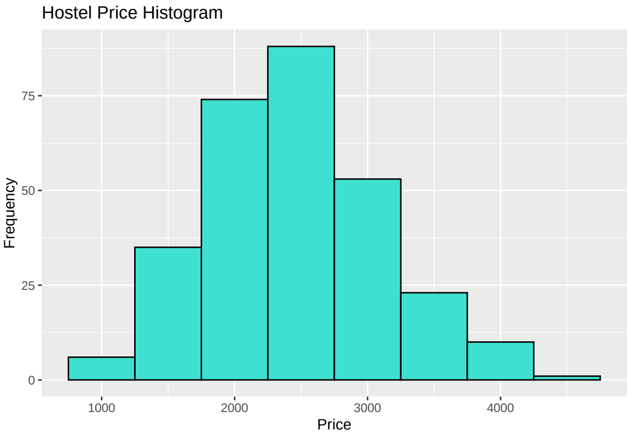

Figure 1: Prices Distribution without Outliers

Below, my city pie chart indicates a vast majority of the observations are from Tokyo, Osaka, and Kyoto. There's not much representation from Fukuoka or Hiroshima.

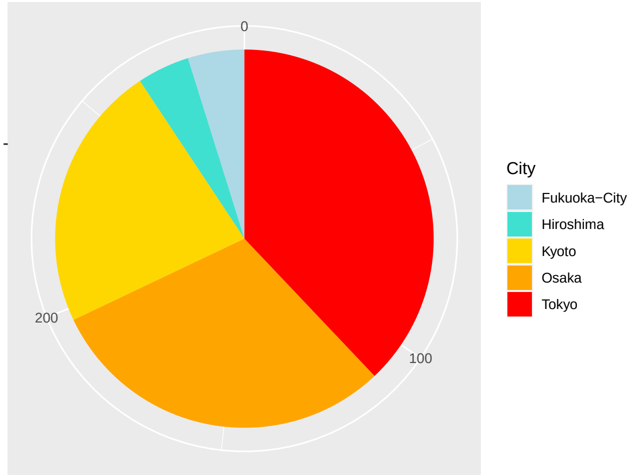

Figure 2: City Distribution Pie Chart

Furthermore, my rating band bar chart indicates of the hostels are either rated Superb or Fabulous. There are few rated Very Good or Good and an extreme few that aren't rated.

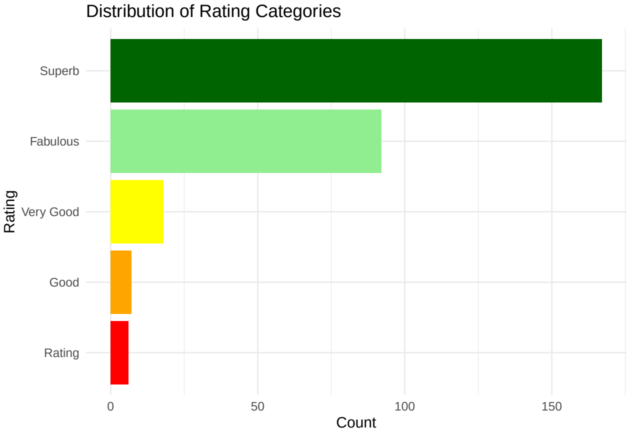

Figure 3: Rating Band Distribution

The below chart shows that the vast majority of hostels are located within 10km (6.2 miles) of their respective city's center. The distribution is right skewed with few outliers. I'll remove the two outliers that have a distance greater than 21km.

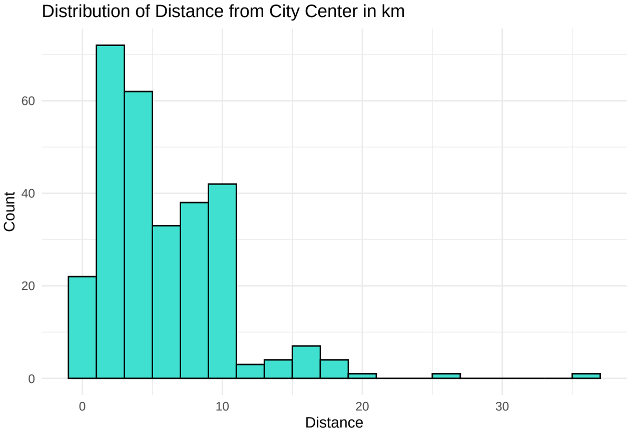

Figure 4: Distance Distribution

Below, you'll notice all distributions of scores appear extremely similarly left skewed. I needed to keep this in mind when checking for multicollinearity..

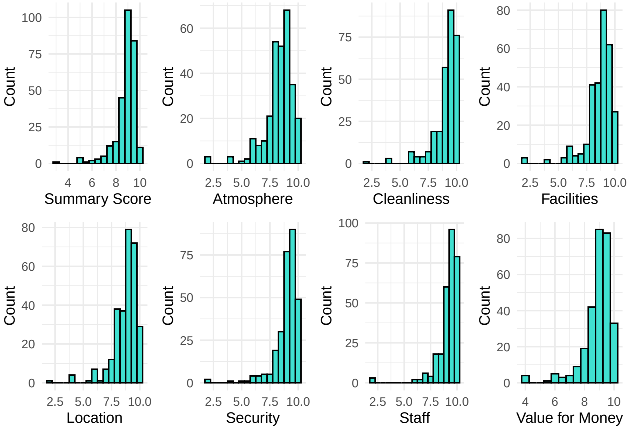

Figure 5: Score Variables Distribution

Multicollinearity Exploration

The correlation heatmap below indicates extremely high high correlation between summary.score and all other rating score predictors, which is expected since it serves as the average of them. I'll remove summary.score to prevent multicollinearity.

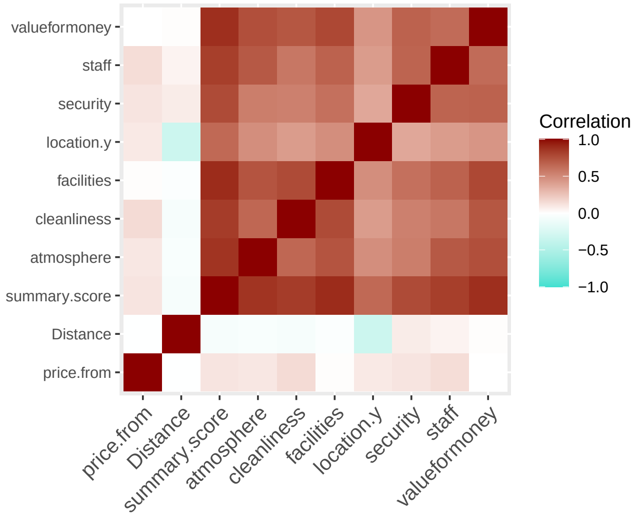

Figure 6: Correlation Heatmap with Summary.Score

The plot below illustrates an improved degree of correlation after removing summary score.

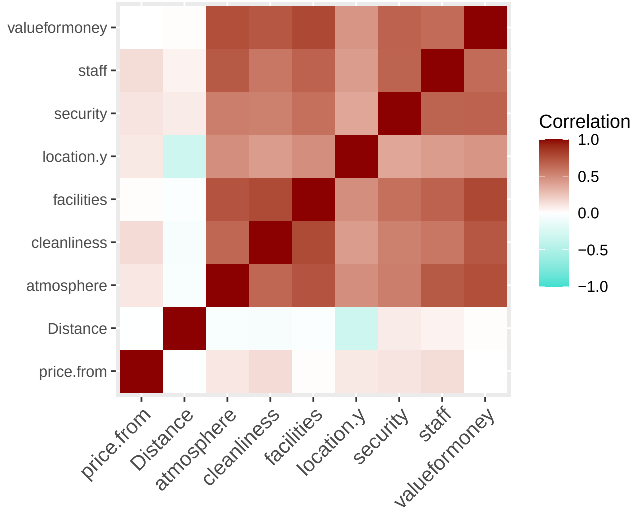

Figure 7: Correlation Heatmap without Summary.Score

Score variables appear relatively highly correlated to each other, but I decided to keep them and later ran vif() during model building to assess whether or not they needed to be removed (A2.2).

Regression Analysis

Model 1: Full Linear Model

```r
m1 <- lm(price.from ~ City + Distance + rating.band + atmosphere +
           cleanliness + facilities + location.y + security + staff +
           valueformoney, data = hostel_cleaned)
summary(m1)
```

```
##
## Call:
## lm(formula = price.from ~ City + Distance + rating.band + atmosphere +
##     cleanliness + facilities + location.y + security + staff +
##     valueformoney, data = hostel_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1185.16  -456.18   -45.05   366.36  1629.56
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1663.439    727.749   2.286 0.023042 *
## CityHiroshima           -78.226    240.868  -0.325 0.745609
## CityKyoto              -424.973    182.528  -2.328 0.020634 *
## CityOsaka              -181.896    182.450  -0.997 0.319672
## CityTokyo                74.323    192.025   0.387 0.699025
## Distance                -28.044     12.955  -2.165 0.031286 *
## rating.bandGood        -131.742    418.700  -0.315 0.753273
## rating.bandVery Good -223.453    476.343  -0.469 0.639375
## rating.bandFabulous  -449.762    573.531  -0.784 0.433609
## rating.bandSuperb    -334.369    656.744  -0.509 0.611074
## atmosphere               74.992     55.204   1.358 0.175451
## cleanliness             224.807     58.235   3.860 0.000142 ***
## facilities             -162.706     61.690  -2.637 0.008835 **
## location.y                8.434     45.029   0.187 0.851569
## security                 95.395     56.102   1.700 0.090205 .
## staff                    82.533     58.923   1.401 0.162453
## valueformoney          -172.769     74.586  -2.316 0.021284 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 614.1 on 271 degrees of freedom
## Multiple R-squared:  0.1692, Adjusted R-squared:  0.1201
## F-statistic: 3.449 on 16 and 271 DF,  p-value: 1.317e-05
```

Fitting a model with price.from as the response and all other variables in hostel_cleaned as predictors resulted in an R squared of .1692 and an adjusted R squared of .1201. Only 16.92% of variance in minimum nightly hostel prices can be explained by this model. Given how this data is real, the relatively low R squared value isn't unexpected.

Using an alpha value of .1, significant predictors are revealed to be CityKyoto, Distance, cleanliness, facilities, security, and valueformoney. The most significant predictor is cleanliness and the least significant predictor is location.y. The model has a statistically significant F value, indicating the model is somewhat effective for explaining variation in minimum nightly hostel price.

Moving forward, my goal was to determine if this model could be improved.

Model Building

I ran vif() on my full model, and rating.band's score of 12.629 indicated it contributed to some multicollinearity. I decided remove rating.band due to this result (A3.1).

Despite how some of the score variables were moderately correlated to each other, the low GVIF values indicated they did not contribute to multicollinearity. I decided to keep them in my model moving forward. I then proceeded to initialize a second model with rating.band removed.

Model 2: Rating.Band Removed

```
m2 <- lm(price.from ~ City + Distance + atmosphere +
           cleanliness + facilities + location.y + security + staff +
           valueformoney, data = hostel_cleaned)
summary(m2)
```

```
##
## Call:
## lm(formula = price.from ~ City + Distance + atmosphere + cleanliness +
##     facilities + location.y + security + staff + valueformoney,
##     data = hostel_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1266.28  -445.12   -44.65   401.74  1699.06
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1768.555    456.675   3.873 0.000135 ***
## CityHiroshima  -45.028    239.830  -0.188 0.851210
## CityKyoto     -387.925    181.760  -2.134 0.033705 *
## CityOsaka     -165.656    182.149  -0.909 0.363905
## CityTokyo       79.485    190.924   0.416 0.677503
## Distance       -25.282     12.843  -1.969 0.050009 .
## atmosphere      82.512     51.740   1.595 0.111918
## cleanliness    218.583     55.353   3.949 9.98e-05 ***
## facilities    -159.860     58.364  -2.739 0.006565 **
## location.y       6.841     42.433   0.161 0.872038
## security        86.822     50.431   1.722 0.086263 .
## staff           65.481     56.297   1.163 0.245783
## valueformoney -203.572     70.609  -2.883 0.004249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 614.4 on 275 degrees of freedom
## Multiple R-squared:  0.1559, Adjusted R-squared:  0.1191
## F-statistic: 4.233 on 12 and 275 DF,  p-value: 3.906e-06
```

```
vif(m2)
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## City          1.921951  4        1.085095
## Distance      2.015598  1        1.419717
## atmosphere    3.246492  1        1.801803
## cleanliness   2.888692  1        1.699615
## facilities    3.914165  1        1.978425
## location.y    1.694583  1        1.301762
## security      2.274689  1        1.508207
## staff         2.715300  1        1.647817
## valueformoney 3.853342  1        1.962993
```

Model 2's R squared and adjusted R squared were slightly lower, as expected. The vif() output indicates multicollinearity was no longer an issue and I could make more progress with model improvement. The above output indicates the same predictors are statistically significant, so below I'll examine interactions.

Variable Interactions

I noticed different intercepts and slopes depending on city in the visualization below. I decided to include the City x Distance interaction in my model to test if it contributed to improvements. I didn't notice any further interactions between City and the other numerical predictors (A3.2).

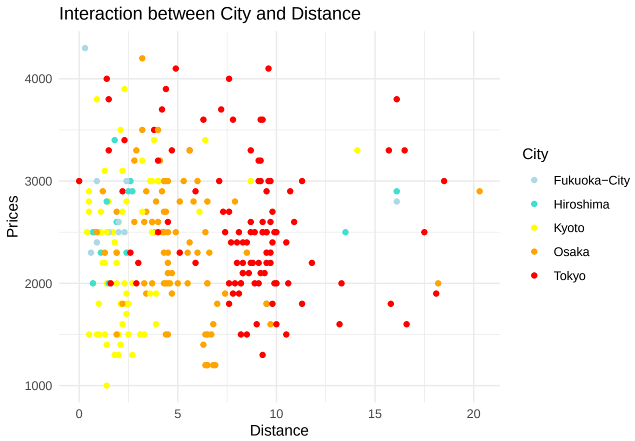

Figure 8: Plot of Interaction between Distance and City

Model 3: Adding City x Distance Interaction

```
m3 <- lm(price.from ~ City + Distance + City:Distance + atmosphere +
           cleanliness + facilities + location.y + security + staff +
           valueformoney, data = hostel_cleaned)
summary(m3)
```

```
##
## Call:
## lm(formula = price.from ~ City + Distance + City:Distance + atmosphere +
##     cleanliness + facilities + location.y + security + staff +
##     valueformoney, data = hostel_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1242.54  -434.59   -11.46   360.55  1795.47
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1654.636    454.668   3.639 0.000327 ***
## CityHiroshima             -136.066    290.125  -0.469 0.639454
## CityKyoto                 -581.546    225.054  -2.584 0.010289 *
## CityOsaka                  -39.779    236.233  -0.168 0.866402
## CityTokyo               410.834    247.620   1.659 0.098246 .  
## Distance                  4.106     41.935   0.098 0.922075    
## atmosphere               69.506     51.032   1.362 0.174328    
## cleanliness             210.025     54.602   3.846 0.000149 ***
## facilities             -157.166     57.775  -2.720 0.006944 ** 
## location.y                1.306     42.322   0.031 0.975413    
## security                105.849     50.889   2.080 0.038465 *  
## staff                    67.442     55.534   1.214 0.225643    
## valueformoney          -195.346     69.800  -2.799 0.005500 ** 
## CityHiroshima:Distance   11.124     54.838   0.203 0.839395    
## CityKyoto:Distance       75.686     56.222   1.346 0.179365    
## CityOsaka:Distance      -41.432     48.688  -0.851 0.395539    
## CityTokyo:Distance      -61.209     45.454  -1.347 0.179230    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 603.7 on 271 degrees of freedom
## Multiple R-squared:  0.1969, Adjusted R-squared:  0.1495 
## F-statistic: 4.153 on 16 and 271 DF,  p-value: 3.655e-07
```

This output seemed extremely promising. Model 3 improved R squared by 4.3% and adjusted R squared by 3.04%. Significant predictors were identified as CityKyoto, CityTokyo, cleanliness, facilities, security, and valueformoney. I decided to run vif() on this model to check for multicollinearity.

```
vif(m3)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## City           51.342232  4        1.636098
## Distance       22.257938  1        4.717832
## atmosphere      3.271221  1        1.808652
## cleanliness     2.911268  1        1.706244
## facilities      3.972724  1        1.993169
## location.y      1.745990  1        1.321359
## security        2.399056  1        1.548889
## staff           2.736703  1        1.654298
## valueformoney   3.900223  1        1.974898
## City:Distance 525.720345  4        2.188237
```

Above, there are extremely high GVIF scores for City, Distance, and the interaction between city and distance. This indicates adding City:Distance inflated the variance of my estimates, and made my model unstable. To avoid multicollinearity, I decided to remove this interaction and proceed forward with model 2.

Assumption Tests and Transformations

I found model 2 satisfied the linearity and constant variance assumptions, but it violated normality and independence (A3.3). I decided to test various transformations, and I found using sqrt(prices.from) was my best option. This transformation satisfied the normality assumption and best-satisfied constant variance and linearity compared to the other models I tested. It also slightly improved my R squared and adjusted R squared.

However, the low p value on the Durbin-Watson test indicated autocorrelation, therefore independence was not satisfied. Notably, none of the transformations satisfied this assumption so I concluded the data was autocorrelated by default (A3.4)

Stepwise Selection

My goal was to determine if I could further increase my R squared values to improve interpretability. I then used forward/backward stepwise regression to see if reducing predictors will work.

The stepwise model displayed a slightly lower R squared but slightly greater adjusted R squared. CityKyoto, cleanliness, valueformoney, security, Distance, facilities, and atmosphere are statistically significant predictors with a = .10. There also was no indication of multicollineairty.

Adjusted R squared improved and more predictors were statistically significant, so I decided select the stepwise model as my final model due to its simplicity (A3.4).

Final Model

```
final_model <- lm(formula = sqrt(price.from) ~ City + cleanliness + valueformoney +
    security + Distance + facilities + atmosphere, data = hostel_cleaned)
summary(final_model)
```

```
##
## Call:
## lm(formula = sqrt(price.from) ~ City + cleanliness + valueformoney +
##     security + Distance + facilities + atmosphere, data = hostel_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2008  -4.3535  -0.1133   4.2733  15.6465
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    43.4162     4.0873  10.622  < 2e-16 ***
## CityHiroshima  -0.2725     2.4289  -0.112 0.910749
## CityKyoto      -4.3114     1.8443  -2.338 0.020115 *
## CityOsaka      -1.7986     1.8480  -0.973 0.331275
## CityTokyo       0.6754     1.9257   0.351 0.726079
## cleanliness     2.2096     0.5611   3.938 0.000104 ***
## valueformoney  -2.2791     0.7110  -3.206 0.001506 **
## security        1.2186     0.4719   2.582 0.010330 *
## Distance       -0.2553     0.1185  -2.153 0.032145 *
## facilities     -1.3527     0.5795  -2.334 0.020296 *
## atmosphere      1.0992     0.4753   2.313 0.021483 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.238 on 277 degrees of freedom
## Multiple R-squared:  0.1569, Adjusted R-squared:  0.1265
## F-statistic: 5.155 on 10 and 277 DF,  p-value: 6.355e-07
```

F Test:

H_0: All model terms are unimportant for predicting the square root of minimum hostel prices

H_A: At least one model term is useful for predicting minimum hostel prices

The p value is statistically significant so reject H_0.

Interpretations

The R squared value of .1569 and adjusted R squared of .1265 indicate a minimal degree of variance in the square root of minimum nightly hostel prices in Japan can be explained by this model. Specifically, the data in my model explains only 15.69% of variance in prices.from. The adjusted R squared improved compared to model 2, indicating my final model is a better fit with respect to the amount of relevant predictors used. These values are relatively low, but they're still interpretable. My overall goal was never to make a highly accurate model. I simply wanted to make initial inferences on predictors given my results.

Predictors

Out of all predictors included the following are statistically significant:

- The City of Kyoto
- Cleanliness
- Value for Money
- Security
- Distance
- Facilities
- Atmosphere

Cleanliness, Security, and Atmosphere are positively linearly associated with the square root of price.

The City of Kyoto, Value for Money, Distance, and Facilities are negatively linearly associated with the square root of price.

The most important predictors are the City of Kyoto, cleanliness, and value for money since they are associated with the highest absolute t values and lowest p values.

Coefficient Interpretations

- on average, the square root of minimum nightly hostel prices in Hiroshima is expected to be 0.2725 units lower compared to Fukuoka
- on average, the square root of minimum nightly hostel prices in Kyoto is expected to be 4.3114 units lower compared to Fukuoka
- on average, the square root of minimum nightly hostel prices in Osaka is expected to be 1.7986 units lower compared to Fukuoka
- on average, the square root of minimum nightly hostel prices in Tokyo is expected to be .6754 units higher compared to Fukuoka
- for every one-unit increase in cleanliness rating, the expected square root of minimum nightly hostel prices increases by 2.2096 units, holding all other variables constant.
- for every one-unit increase in the value for money score, the expected square root of minimum nightly hostel prices decreases by 2.2791 units, holding all other variables constant.
- for every one-unit increase in security score, the expected square root of minimum nightly hostel prices increases by 1.2186 units, holding all other variables constant
- for every one-unit increase in distance from city center in km, the expected square root of minimum nightly hostel prices decreases by approximately 0.2553 units, holding all other variables constant.
- for every one-unit increase in facilities score, the expected square root of minimum nightly hostel prices decreases by 1.3527 units, holding all other variables constant
- for every one-unit increase in atmosphere rating, the expected square root of minimum nightly hostel prices increases by 1.0992 units, holding all other variables constant.

Final Model Assumptions

After testing assumptions on my final model, I found it satisfied normality, linearity, and constant variance. It expectedly failed the independence assumption (A3.5).

Conclusion

Essentially, I found that the following factors are significantly linearly related to the square root of hostel prices:

- The City of Kyoto
- Cleanliness
- Value for Money
- Security
- Distance
- Facilities
- Atmosphere

Of these variables, cleanliness, security, and atmosphere are positively linearly associated with the square root of price while the city of Kyoto, value for money, distance, and facilities are negatively linearly associated with the square root of price.

I determined that hostels further from the city center with higher facility and valueformoney scores tend to be associated with lower prices. Additionally, hostels in Kyoto tend to display lower rates compared to other cities.

Most importantly: stakeholders, or hostel managers, should allocate resources towards maintaining safe, clean spaces with appealing atmosphere in order to justify charging higher rates.

Limitations and Future Work

I need to emphasize that these interpretations should be taken with a grain of salt. My final model has a low R squared value, meaning only a very small portion of variability in hostel prices is explained by my predictors. My analysis can be used to make initial inferences, but more work is definitely needed in order to make definitive concrete determinations.

The hostel data was scraped before 2020, so it's possible prices are extremely different today. Additionally, this dataset is heavily dependent on subjective opinion-based score data. I'd like to conduct further analysis using objective data such as occupancy per room in order to hopefully improve my model accuracy.

Appendix

Data Description

A1.1 Summary Statistics for all Variables

```
summary(hostelA)
```

```
##        X          hostel.name            City            price.from
##  Min.   :  1.00   Length:342         Length:342         Min.   :   1000
##  1st Qu.: 86.25   Class :character   Class :character   1st Qu.:   2000
##  Median :171.50   Mode  :character   Mode  :character   Median :   2500
##  Mean   :171.50                                         Mean   :   8388
##  3rd Qu.:256.75                                         3rd Qu.:   2900
##  Max.   :342.00                                         Max.   :1003200
##
##    Distance         summary.score    rating.band          atmosphere
##  Length:342         Min.   : 3.100   Length:342         Min.   : 2.000
##  Class :character   1st Qu.: 8.600   Class :character   1st Qu.: 7.800
##  Mode  :character   Median : 9.000   Mode  :character   Median : 8.600
##                     Mean   : 8.783                      Mean   : 8.239
##                     3rd Qu.: 9.400                      3rd Qu.: 9.000
##                     Max.   :10.000                      Max.   :10.000
##                     NA's   :15                          NA's   :15
##   cleanliness      facilities      location.y       security
##  Min.   : 2.000   Min.   : 2.000   Min.   : 2.000   Min.   : 2.000
##  1st Qu.: 8.800   1st Qu.: 8.000   1st Qu.: 8.000   1st Qu.: 8.700
##  Median : 9.300   Median : 9.000   Median : 9.000   Median : 9.200
##  Mean   : 9.012   Mean   : 8.598   Mean   : 8.695   Mean   : 8.947
##  3rd Qu.: 9.800   3rd Qu.: 9.300   3rd Qu.: 9.400   3rd Qu.: 9.600
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##  NA's   :15       NA's   :15       NA's   :15       NA's   :15
##      staff        valueformoney         lon            lat
##  Min.   : 2.000   Min.   : 4.000   Min.   :103.9   Min.   : 1.311
##  1st Qu.: 9.000   1st Qu.: 8.600   1st Qu.:135.5   1st Qu.:34.669
##  Median : 9.400   Median : 9.000   Median :135.8   Median :34.998
##  Mean   : 9.133   Mean   : 8.848   Mean   :136.8   Mean   :34.977
##  3rd Qu.: 9.800   3rd Qu.: 9.500   3rd Qu.:139.8   3rd Qu.:35.697
##  Max.   :10.000   Max.   :10.000   Max.   :139.9   Max.   :36.205
##  NA's   :15       NA's   :15       NA's   :44      NA's   :44
```

```
hostelA <- hostel
```

Exploratory Data Analysis

A2.1 Response Plots

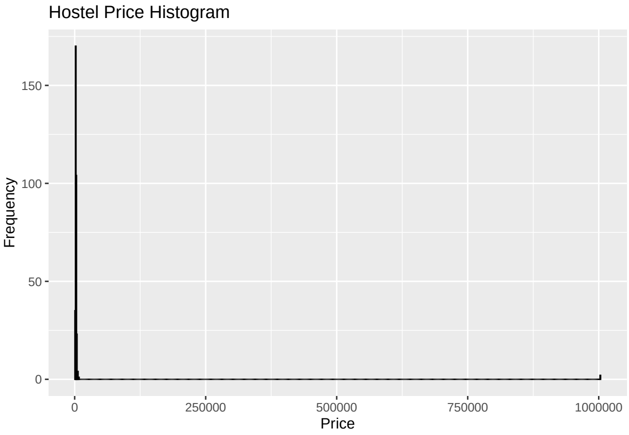

Figure 9: Prices Distribution with Outliers

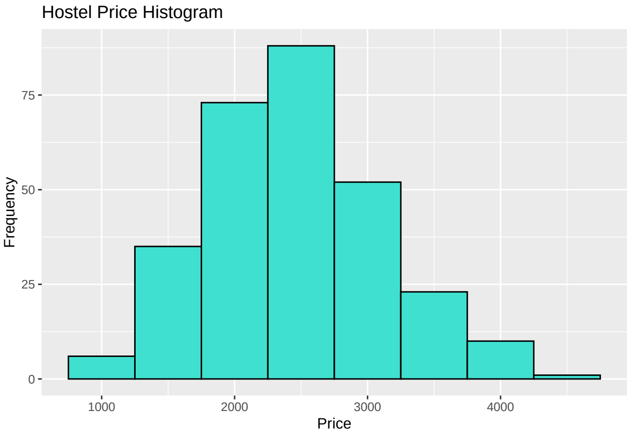

Figure 10: Prices Distribution without Outliers

A2.2 Checking Correlation

```
cor2 <- cor(hostel_num2)
```

```
cor2
```

```
##                  price.from    Distance  atmosphere cleanliness  facilities
## price.from     1.0000000000 -0.01438057  0.10434082  0.15023013  0.01234368
## Distance      -0.0143805663  1.00000000 -0.04928061 -0.06048418 -0.03175701
## atmosphere     0.1043408238 -0.04928061  1.00000000  0.65460548  0.72591971
## cleanliness    0.1502301326 -0.06048418  0.65460548  1.00000000  0.77006872
## facilities     0.0123436821 -0.03175701  0.72591971  0.77006872  1.00000000
## location.y     0.0861653679 -0.33312376  0.48357062  0.41728564  0.48393197
## security       0.1137938107  0.07523649  0.55090734  0.53869962  0.61233198
## staff          0.1419065009  0.05086251  0.71048724  0.57890418  0.66710284
## valueformoney -0.0002243443  0.01188128  0.75194442  0.72294358  0.78376522
##                 location.y   security      staff valueformoney
## price.from     0.08616537 0.11379381 0.14190650 -0.0002243443
## Distance      -0.33312376 0.07523649 0.05086251  0.0118812813
## atmosphere     0.48357062 0.55090734 0.71048724  0.7519444176
## cleanliness    0.41728564 0.53869962 0.57890418  0.7229435770
## facilities     0.48393197 0.61233198 0.66710284  0.7837652248
## location.y     1.00000000 0.36731162 0.42449630  0.4521053724
## security       0.36731162 1.00000000 0.65785620  0.6652750929
## staff          0.42449630 0.65785620 1.00000000  0.6252385224
## valueformoney  0.45210537 0.66527509 0.62523852  1.0000000000
```

I notice somewhat high degrees of correlation between certain score variables, however removing or combining them will weaken potential interpretability. My goal is to how individual score categories may contribute to price, so it'll be difficult to make conclusions if I remove or combine too many of them. Including scores despite moderately high correlation to each other will allow me to be more specific in my interpretations. I'll move forward and create a full regression model, then I'll analyze VIF scores to justify whether or not more predictors may interfere with analysis.

Regression Analysis

A3.1 VIF Analysis on Full Model

```
vif(m1)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## City           2.054005  4        1.094146
## Distance       2.053540  1        1.433018
## rating.band   12.629214  4        1.373005
## atmosphere     3.700232  1        1.923599
## cleanliness    3.201119  1        1.789167
## facilities     4.378317  1        2.092443
## location.y     1.910557  1        1.382229
## security       2.818456  1        1.678826
## staff          2.978178  1        1.725740
## valueformoney  4.304849  1        2.074813
```

A3.2 City x Rating Score Predictors

I'll now examine possible interaction between City and each rating score predictor.

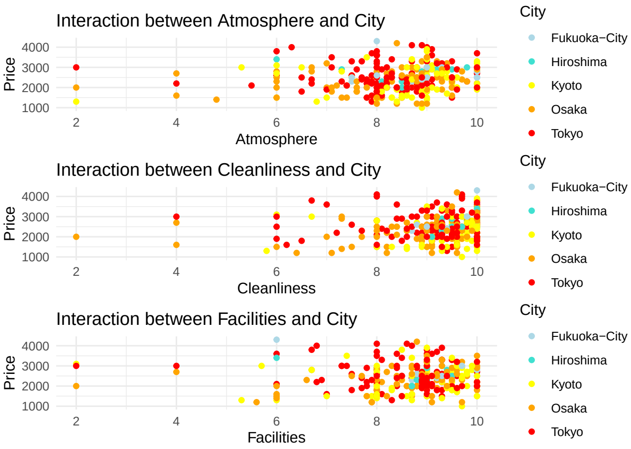

Figure 11: Other Interactions with City I

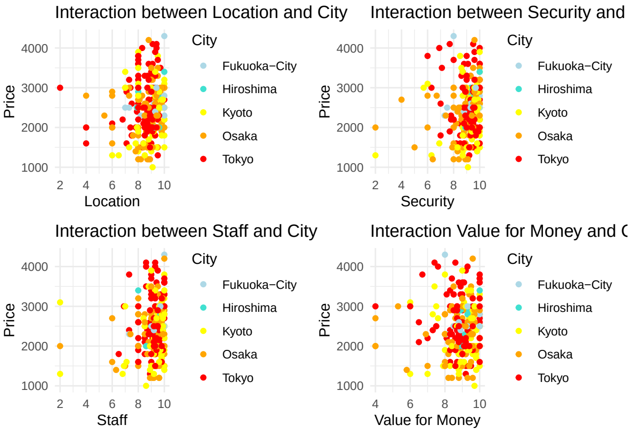

Figure 12: Other Interactions with City II

A3.3 Checking Model 2 Assumptions

Normality:

```
shapiro.test(m2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m2$residuals
## W = 0.98729, p-value = 0.01226
```

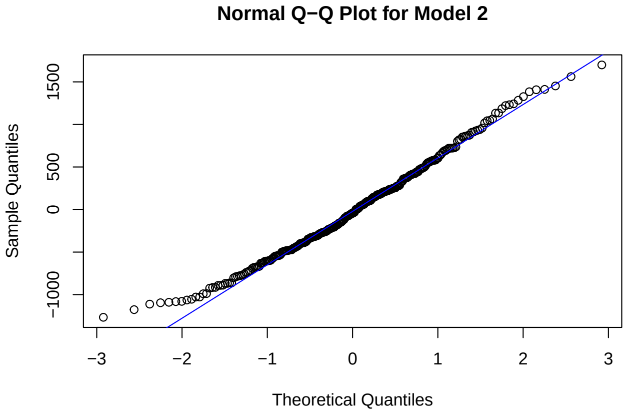

Figure 13: Model 2 QQPlot

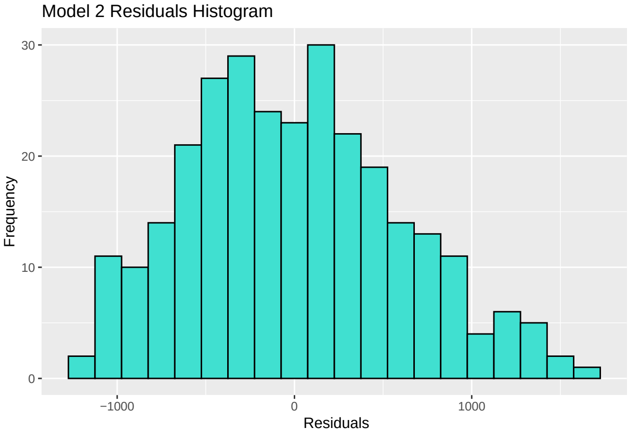

Figure 14: Model 2 Residuals

The low p value indicates normality is violated. However, the points on the qqplot are relatively close to the qqline and the residual bar plot appears nearly normal. I believe transformations on my response may improve this assumption.

Linearity and Constant Variance:

There is somewhat random scattering of points. There seems to be slightly visible downward linear patterns concentrated toward the middle of the plot. Given how there are no blatant patterns, I conclude constant variance is approximately satisfied.

Additionally, the red fitted line appears reasonably close to the dashed line. Therefore, this indicates linearity approximately holds.

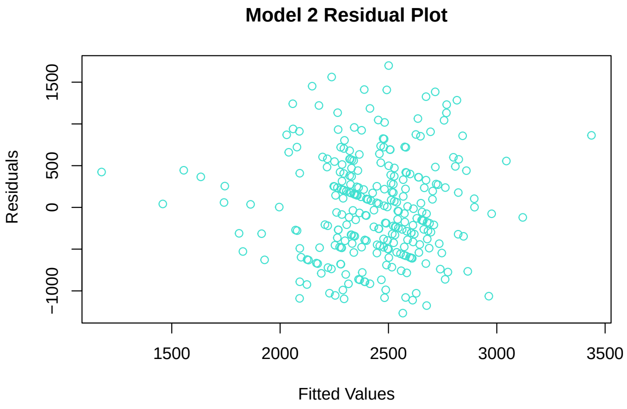

Figure 15: Model 2 Residuals v Fitted

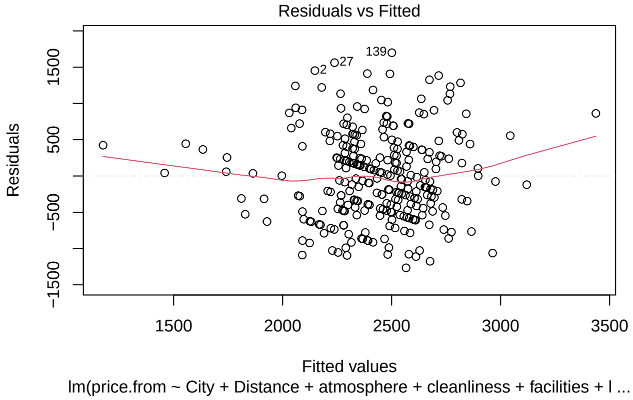

Figure 16: Other Model 2 Assumption Plots

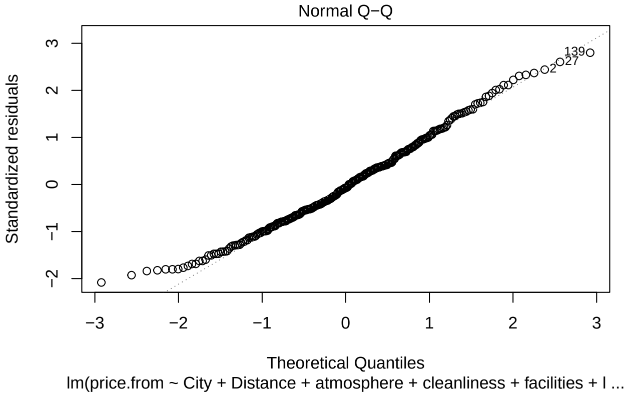

Figure 17: Other Model 2 Assumption Plots

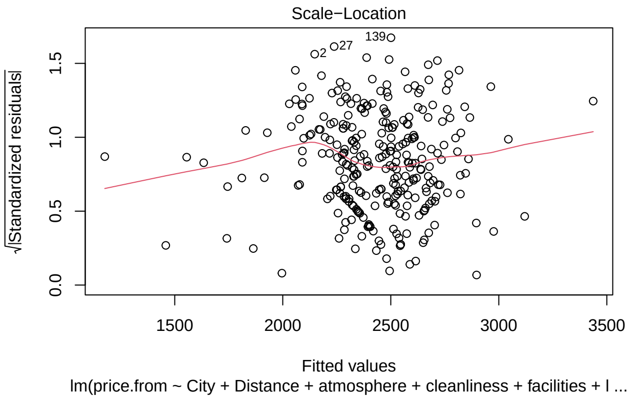

Figure 18: Other Model 2 Assumption Plots

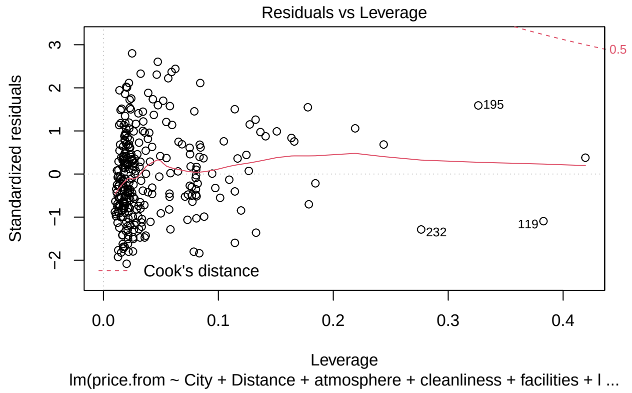

Figure 19: Other Model 2 Assumption Plots

Independence:

```
dwt(m2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1192014      1.753007   0.028
##  Alternative hypothesis: rho != 0
```

The p value is barely less than a = .05, indicating residuals may be correlated. Independence is not satisfied.

These results suggests a transformation on price.from is warranted.

A3.4 Transformations on Prices

```
summary(m2)
```

```
##
## Call:
## lm(formula = price.from ~ City + Distance + atmosphere + cleanliness +
##     facilities + location.y + security + staff + valueformoney,
##     data = hostel_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1266.28  -445.12   -44.65   401.74  1699.06
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1768.555    456.675   3.873 0.000135 ***
## CityHiroshima  -45.028    239.830  -0.188 0.851210
## CityKyoto     -387.925    181.760  -2.134 0.033705 *
## CityOsaka     -165.656    182.149  -0.909 0.363905
## CityTokyo       79.485    190.924   0.416 0.677503
## Distance       -25.282     12.843  -1.969 0.050009 .
## atmosphere      82.512     51.740   1.595 0.111918
## cleanliness    218.583     55.353   3.949 9.98e-05 ***
## facilities    -159.860     58.364  -2.739 0.006565 **
## location.y       6.841     42.433   0.161 0.872038
## security        86.822     50.431   1.722 0.086263 .
## staff           65.481     56.297   1.163 0.245783
## valueformoney -203.572     70.609  -2.883 0.004249 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 614.4 on 275 degrees of freedom
## Multiple R-squared:  0.1559, Adjusted R-squared:  0.1191
## F-statistic: 4.233 on 12 and 275 DF,  p-value: 3.906e-06
```

Square root:

```
m2_t1 <- lm(formula = sqrt(price.from) ~ City + Distance + atmosphere + cleanliness +
    facilities + location.y + security + staff + valueformoney,
    data = hostel_cleaned)
summary(m2_t1)
```

```
##
## Call:
```

```
## lm(formula = sqrt(price.from) ~ City + Distance + atmosphere +
##     cleanliness + facilities + location.y + security + staff +
##     valueformoney, data = hostel_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1053  -4.3349  -0.0972   4.2743  15.2807
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   41.78116    4.64260   9.000  < 2e-16 ***
## CityHiroshima -0.35400    2.43813  -0.145 0.884664
## CityKyoto     -4.24722    1.84779  -2.299 0.022281 *
## CityOsaka     -1.84049    1.85174  -0.994 0.321131
## CityTokyo      0.63192    1.94095   0.326 0.744995
## Distance      -0.25417    0.13056  -1.947 0.052583 .
## atmosphere     0.85541    0.52599   1.626 0.105035
## cleanliness    2.17787    0.56273   3.870 0.000136 ***
## facilities    -1.46989    0.59333  -2.477 0.013838 *
## location.y     0.06055    0.43138   0.140 0.888483
## security       0.99898    0.51268   1.949 0.052368 .
## staff          0.61819    0.57232   1.080 0.281022
## valueformoney -2.19411    0.71782  -3.057 0.002459 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.246 on 275 degrees of freedom
## Multiple R-squared:  0.1606, Adjusted R-squared:  0.124
## F-statistic: 4.385 on 12 and 275 DF,  p-value: 2.084e-06
```

```
vif(m2_t1)
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## City          1.921951  4        1.085095
## Distance      2.015598  1        1.419717
## atmosphere    3.246492  1        1.801803
## cleanliness   2.888692  1        1.699615
## facilities    3.914165  1        1.978425
## location.y    1.694583  1        1.301762
## security      2.274689  1        1.508207
## staff         2.715300  1        1.647817
## valueformoney 3.853342  1        1.962993
```

There is slight improvement on R squared and adj. R squared. VIF indicates no multicollinearity.

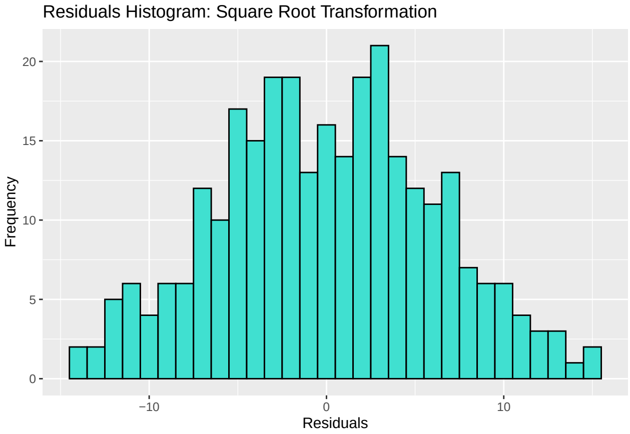

Figure 20: Residuals Histogram: Square Root Transformation

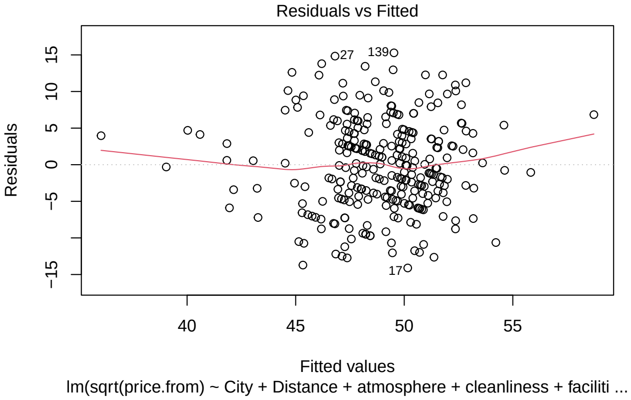

Figure 21: Square Root Transformation Assumption Plots

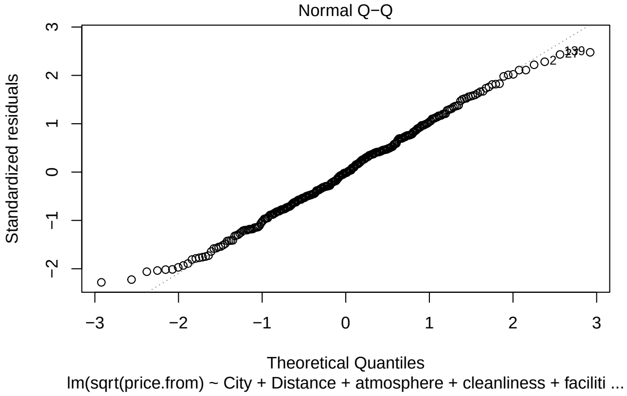

Figure 22: Square Root Transformation Assumption Plots

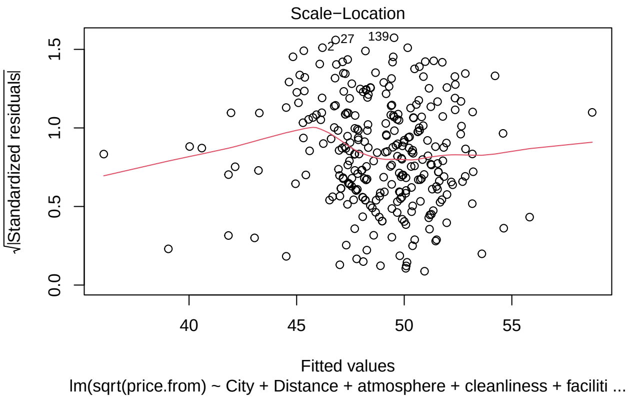

Figure 23: Square Root Transformation Assumption Plots

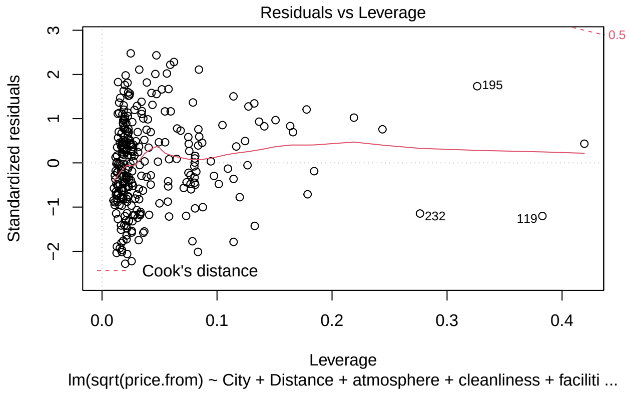

Figure 24: Square Root Transformation Assumption Plots

```
shapiro.test(m2_t1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m2_t1$residuals
## W = 0.99438, p-value = 0.3684
```

```
dwt(m2_t1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1225927      1.746584    0.02
##  Alternative hypothesis: rho != 0
```

Normality is improved, given the plots along with the fact that the p is greater than .05. Constant variance and linearity assumptions still approximately hold. Independence is worsened, given how its p value is lower than model 2.

Log:

```
m2_t2 <- lm(formula = log(price.from) ~ City + Distance + atmosphere + cleanliness +
    facilities + location.y + security + staff + valueformoney,
    data = hostel_cleaned)
summary(m2_t2)
```

```
##
## Call:
## lm(formula = log(price.from) ~ City + Distance + atmosphere +
##     cleanliness + facilities + location.y + security + staff +
##     valueformoney, data = hostel_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70242 -0.17092  0.01364  0.18272  0.57150
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.450396   0.193587  38.486  < 2e-16 ***
## CityHiroshima -0.011208   0.101665  -0.110 0.912297
## CityKyoto     -0.188931   0.077049  -2.452 0.014825 *
## CityOsaka     -0.083045   0.077214  -1.076 0.283082
## CityTokyo      0.019421   0.080934   0.240 0.810538
## Distance      -0.010433   0.005444  -1.916 0.056358 .
## atmosphere     0.035774   0.021933   1.631 0.104023
## cleanliness    0.088375   0.023465   3.766 0.000203 ***
## facilities    -0.054382   0.024741  -2.198 0.028777 *
## location.y     0.002202   0.017988   0.122 0.902666
## security       0.046227   0.021378   2.162 0.031452 *
## staff          0.023819   0.023864   0.998 0.319109
## valueformoney -0.095911   0.029932  -3.204 0.001513 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2605 on 275 degrees of freedom
## Multiple R-squared:  0.165,  Adjusted R-squared:  0.1286
## F-statistic: 4.528 on 12 and 275 DF,  p-value: 1.155e-06
```

```
vif(m2_t2)
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## City          1.921951  4        1.085095
## Distance      2.015598  1        1.419717
## atmosphere    3.246492  1        1.801803
## cleanliness   2.888692  1        1.699615
## facilities    3.914165  1        1.978425
## location.y    1.694583  1        1.301762
## security      2.274689  1        1.508207
## staff         2.715300  1        1.647817
## valueformoney 3.853342  1        1.962993
```

There is greater improvement on R squared and adj. R squared. VIF indicates no multicollinearity.

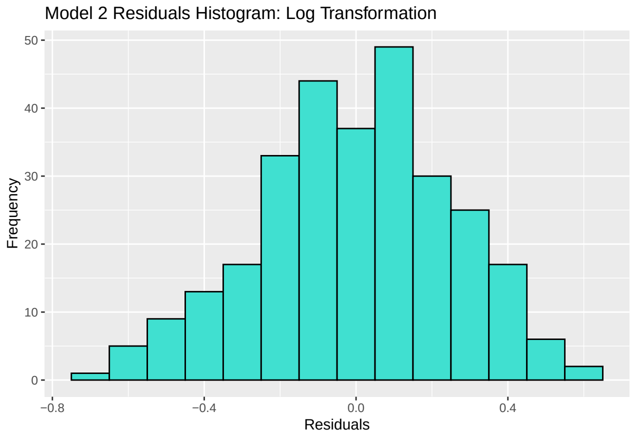

Figure 25: Residuals Histogram: Inverse Transformation

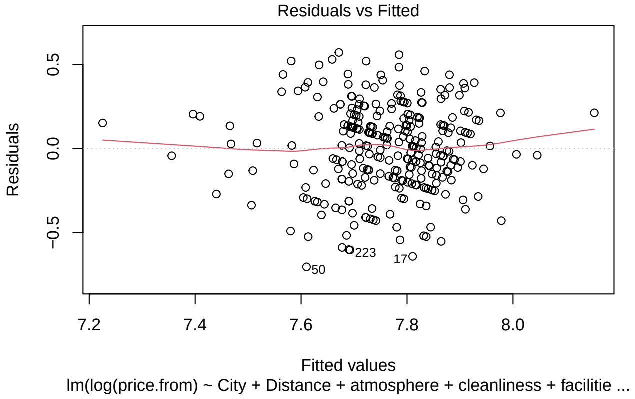

Figure 26: Log Transformation Assumption Plots

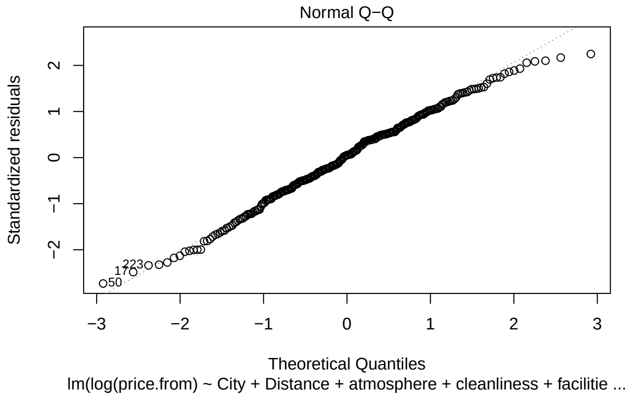

Figure 27: Log Transformation Assumption Plots

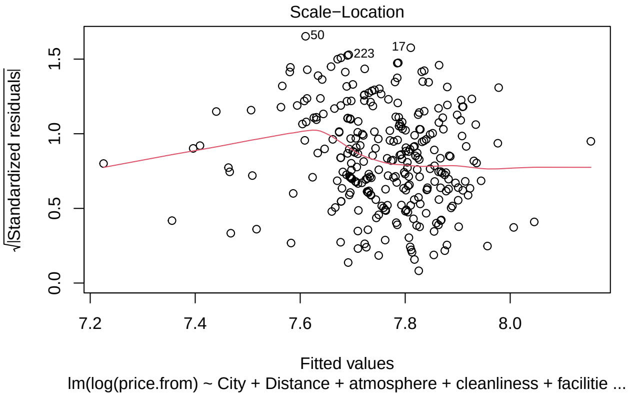

Figure 28: Log Transformation Assumption Plots

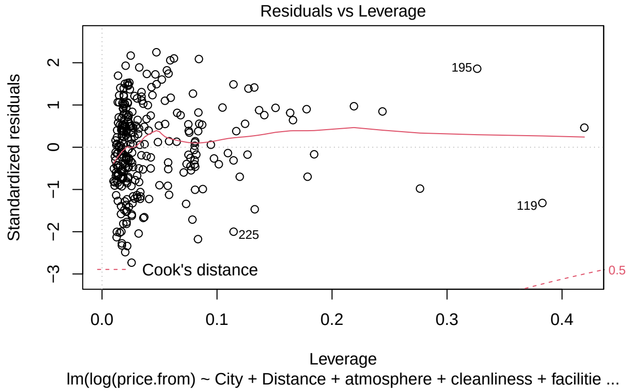

Figure 29: Log Transformation Assumption Plots

```
shapiro.test(m2_t2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m2_t2$residuals
## W = 0.99263, p-value = 0.1652
```

```
dwt(m2_t2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1278282      1.736677   0.026
##  Alternative hypothesis: rho != 0
```

Normality is improved, given the plots along with the fact that the p is greater than .05. However, the square root transformation better-maintained this assumption. Constant variance and linearity assumptions still approximately hold. Independence is even further worsened, given how its p value is lower than model 2.

Inverse:

```
m2_t3 <- lm(formula = 1/(price.from) ~ City + Distance + atmosphere + cleanliness +
    facilities + location.y + security + staff + valueformoney,
    data = hostel_cleaned)
summary(m2_t3)
```

```
##
## Call:
## lm(formula = 1/(price.from) ~ City + Distance + atmosphere +
##     cleanliness + facilities + location.y + security + staff +
##     valueformoney, data = hostel_cleaned)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.323e-04 -8.037e-05 -1.492e-05  6.470e-05  4.852e-04
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.995e-04  9.113e-05   6.579 2.38e-10 ***
## CityHiroshima  2.867e-06  4.786e-05   0.060 0.952271
## CityKyoto      9.734e-05  3.627e-05   2.684 0.007720 **
## CityOsaka      4.385e-05  3.635e-05   1.206 0.228673
## CityTokyo     -3.932e-06  3.810e-05  -0.103 0.917866
## Distance       4.669e-06  2.563e-06   1.822 0.069542 .
## atmosphere    -1.600e-05  1.032e-05  -1.550 0.122266
## cleanliness   -3.849e-05  1.105e-05  -3.485 0.000573 ***
## facilities     1.890e-05  1.165e-05   1.623 0.105825
## location.y    -7.786e-07  8.467e-06  -0.092 0.926807
## security      -2.534e-05  1.006e-05  -2.518 0.012364 *
## staff         -9.523e-06  1.123e-05  -0.848 0.397317
## valueformoney  4.778e-05  1.409e-05   3.391 0.000798 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0001226 on 275 degrees of freedom
## Multiple R-squared:  0.1704, Adjusted R-squared:  0.1342
## F-statistic: 4.708 on 12 and 275 DF,  p-value: 5.487e-07
```

```
vif(m2_t3)
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## City          1.921951  4        1.085095
## Distance      2.015598  1        1.419717
## atmosphere    3.246492  1        1.801803
## cleanliness   2.888692  1        1.699615
## facilities    3.914165  1        1.978425
## location.y    1.694583  1        1.301762
## security      2.274689  1        1.508207
## staff         2.715300  1        1.647817
## valueformoney 3.853342  1        1.962993
```

This displays the greatest improvement on R squared and adj. R squared. VIF indicates no multicollinearity.

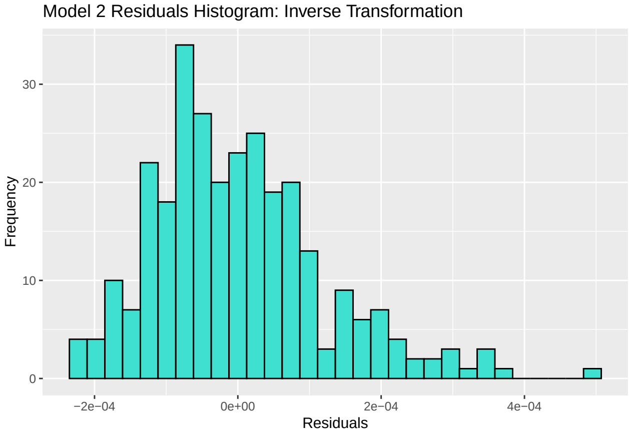

Figure 30: Residuals Histogram: Inverse Transformation

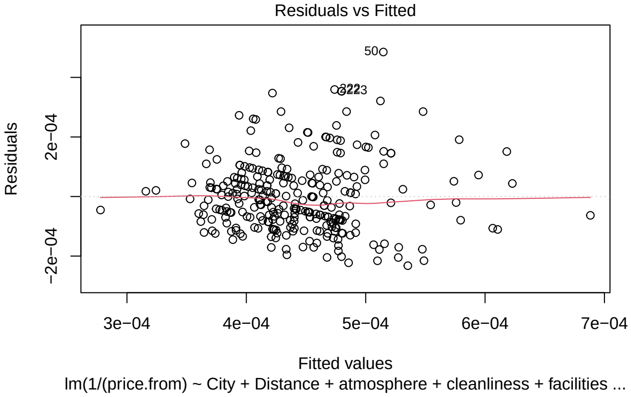

Figure 31: Inverse Transformation Asssumption Plots

```
shapiro.test(m2_t3$residuals)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  m2_t3$residuals
## W = 0.95886, p-value = 2.853e-07
```

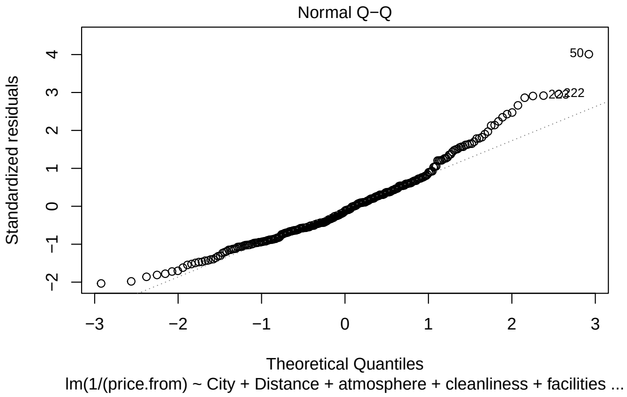

Figure 32: Inverse Transformation Asssumption Plots

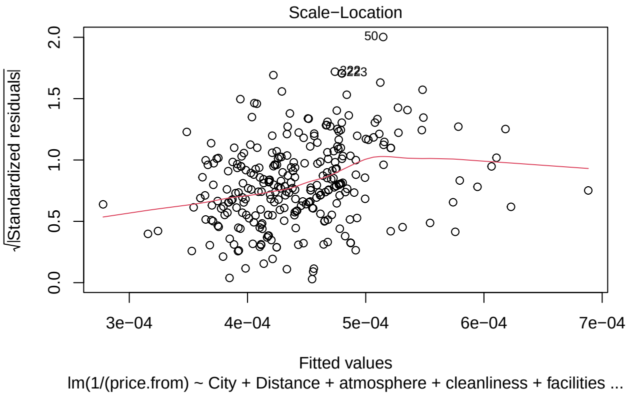

Figure 33: Inverse Transformation Asssumption Plots

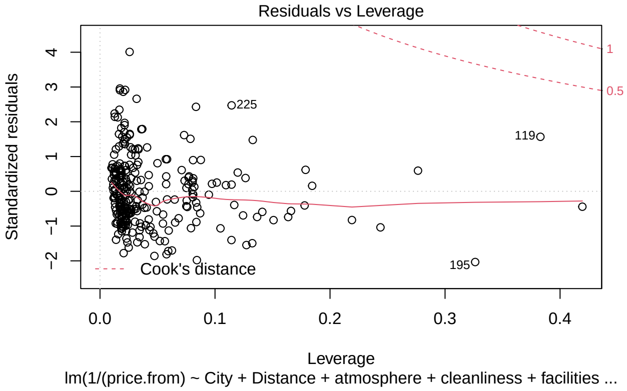

Figure 34: Inverse Transformation Asssumption Plots

```
dwt(m2_t3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1415227      1.710859   0.012
##  Alternative hypothesis: rho != 0
```

Normality is not satisfied since there is visible deviation from the qqline, and the Shapiro-Wilk p value is furthest from .05. Constant Variance and Linearity still approximately hold. The p value for the Durbin-Watson test is still less than .05, indicating independence is violated.

When considering these results, I believe a square root transformation is the best option since it upholds normality, constant variance, and linearity as best as possible. I will move forward with m2_t1.

A3.4 Stepwise Regression

```
#null model with no predictors
stepwise_null_model = lm(sqrt(price.from) ~ 1, hostel_cleaned)

#full model with all relevant predictors
stepwise_full_model <- lm(sqrt(price.from) ~ City + Distance + atmosphere +
    cleanliness + facilities + location.y + security + staff +
    valueformoney, data = hostel_cleaned)
```

```
stepwise_model <- step(stepwise_null_model, scope = list(lower = stepwise_null_model, upper =
stepwise_full_model), direction = "both")
```

```
## Start:  AIC=1094.35
## sqrt(price.from) ~ 1
##
##                 Df Sum of Sq   RSS    AIC
## + City           4    632.77 12150 1087.7
## + cleanliness    1    317.85 12465 1089.1
## + staff          1    294.93 12488 1089.6
## + security       1    215.83 12567 1091.5
## + atmosphere     1    164.41 12618 1092.6
## + location.y     1    107.70 12675 1093.9
## <none>                       12783 1094.3
## + facilities     1     10.36 12772 1096.1
## + Distance       1      1.41 12781 1096.3
## + valueformoney  1      0.79 12782 1096.3
##
## Step:  AIC=1087.73
## sqrt(price.from) ~ City
##
##                 Df Sum of Sq   RSS    AIC
## + cleanliness    1    315.97 11834 1082.1
## + staff          1    233.25 11917 1084.2
## + Distance       1    180.98 11969 1085.4
## + security       1    156.81 11993 1086.0
## + atmosphere     1    155.10 11995 1086.0
## + location.y     1    108.81 12041 1087.1
## <none>                       12150 1087.7
## + facilities     1     10.84 12139 1089.5
## + valueformoney  1      0.03 12150 1089.7
## - City           4    632.77 12783 1094.3
```

```
##
## Step:  AIC=1082.14
## sqrt(price.from) ~ City + cleanliness
##
##                 Df Sum of Sq   RSS    AIC
## + valueformoney  1    334.64 11499 1075.9
## + facilities     1    271.29 11563 1077.5
## + Distance       1    187.68 11646 1079.5
## <none>                       11834 1082.1
## + staff          1     36.71 11797 1083.2
## + security       1     13.27 11821 1083.8
## + location.y     1     13.07 11821 1083.8
## + atmosphere     1      1.38 11833 1084.1
## - cleanliness    1    315.97 12150 1087.7
## - City           4    630.89 12465 1089.1
##
## Step:  AIC=1075.88
## sqrt(price.from) ~ City + cleanliness + valueformoney
##
##                 Df Sum of Sq   RSS    AIC
## + security       1    198.72 11301 1072.9
## + staff          1    185.05 11314 1073.2
## + atmosphere     1    170.38 11329 1073.6
## + Distance       1    159.23 11340 1073.9
## <none>                       11499 1075.9
## + location.y     1     69.78 11430 1076.1
## + facilities     1     67.78 11432 1076.2
## - valueformoney  1    334.64 11834 1082.1
## - City           4    663.91 12163 1084.0
## - cleanliness    1    650.57 12150 1089.7
##
## Step:  AIC=1072.86
## sqrt(price.from) ~ City + cleanliness + valueformoney + security
##
##                 Df Sum of Sq   RSS    AIC
## + Distance       1    183.97 11117 1070.1
## + atmosphere     1    140.68 11160 1071.2
## + facilities     1    120.63 11180 1071.8
## <none>                       11301 1072.9
## + staff          1     74.48 11226 1073.0
## + location.y     1     50.62 11250 1073.6
## - security       1    198.72 11499 1075.9
## - City           4    600.13 11901 1079.8
## - valueformoney  1    520.08 11821 1083.8
## - cleanliness    1    574.31 11875 1085.1
##
## Step:  AIC=1070.13
## sqrt(price.from) ~ City + cleanliness + valueformoney + security +
##     Distance
##
##                 Df Sum of Sq   RSS    AIC
## + facilities     1    131.39 10985 1068.7
## + atmosphere     1    127.46 10989 1068.8
## + staff          1     79.54 11037 1070.1
## <none>                       11117 1070.1
## + location.y     1      2.96 11114 1072.1
## - Distance       1    183.97 11301 1072.9
## - security       1    223.45 11340 1073.9
## - valueformoney  1    506.24 11623 1081.0
## - City           4    783.96 11901 1081.8
## - cleanliness    1    547.39 11664 1082.0
##
## Step:  AIC=1068.71
## sqrt(price.from) ~ City + cleanliness + valueformoney + security +
##     Distance + facilities
##
##                 Df Sum of Sq   RSS    AIC
## + atmosphere     1    208.06 10777 1065.2
## + staff          1    147.84 10837 1066.8
## <none>                       10985 1068.7
## - facilities     1    131.39 11117 1070.1
## + location.y     1     15.41 10970 1070.3
## - Distance       1    194.74 11180 1071.8
## - valueformoney  1    250.89 11236 1073.2
## - security       1    282.30 11268 1074.0
## - City           4    780.53 11766 1080.5
## - cleanliness    1    677.85 11663 1084.0
##
## Step:  AIC=1065.2
## sqrt(price.from) ~ City + cleanliness + valueformoney + security +
##     Distance + facilities + atmosphere
##
##                 Df Sum of Sq   RSS    AIC
## <none>                       10777 1065.2
## + staff          1     46.87 10730 1066.0
## + location.y     1      2.11 10775 1067.2
## - Distance       1    180.43 10958 1068.0
## - atmosphere     1    208.06 10985 1068.7
## - facilities     1    212.00 10989 1068.8
## - security       1    259.43 11037 1070.0
## - valueformoney  1    399.81 11177 1073.7
## - City           4    769.06 11546 1077.0
## - cleanliness    1    603.30 11380 1078.9
```

A3.5 Final Model Assumptions

```
shapiro.test(stepwise_model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  stepwise_model$residuals
## W = 0.995, p-value = 0.4752
```

```
dwt(stepwise_model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1275398      1.736661   0.016
##  Alternative hypothesis: rho != 0
```

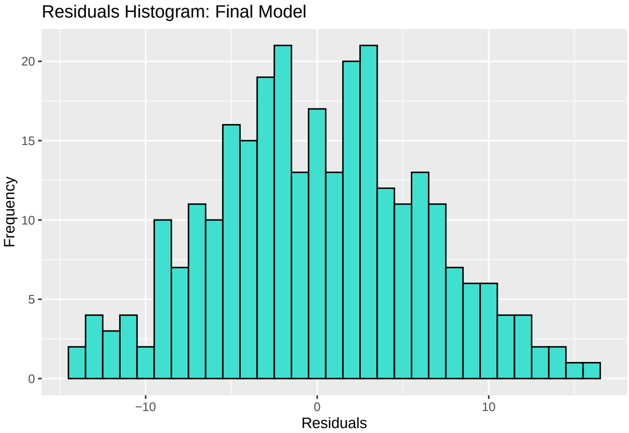

Figure 35: Residuals Histogram: Final Model

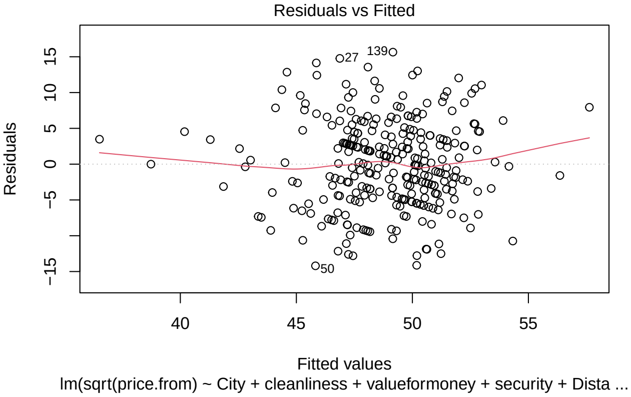

Figure 36: Final Model Assumption Plots

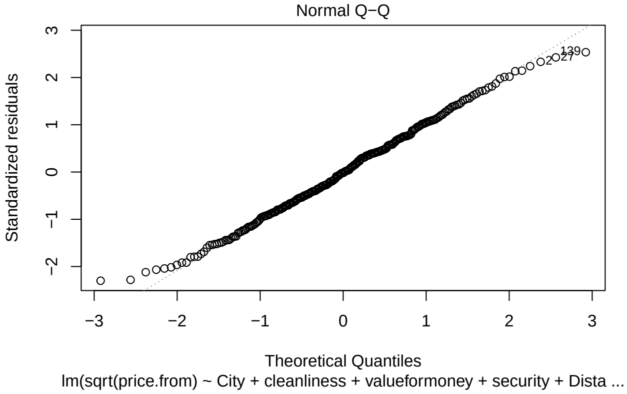

Figure 37: Final Model Assumption Plots

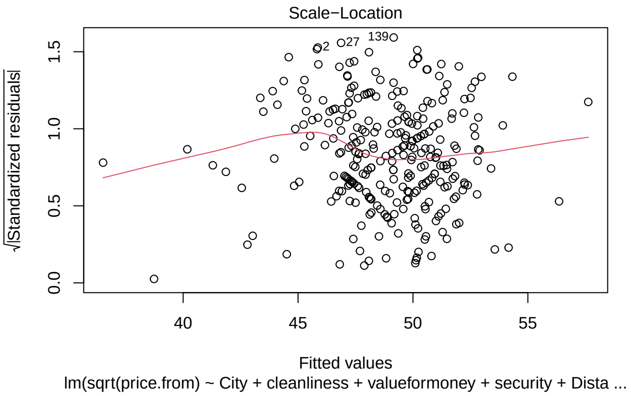

Figure 38: Final Model Assumption Plots

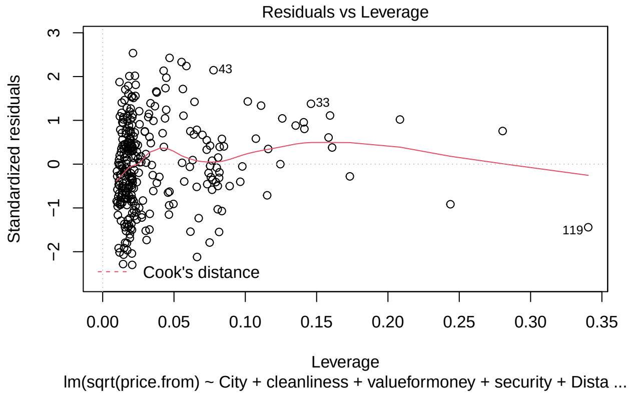

Figure 39: Final Model Assumption Plots

References

https://www.kaggle.com/datasets/koki25ando/hostel-world-dataset

https://www.hostelworld.com