HOSTEL PRICES IN JAPAN

REGRESSION ANALYSIS

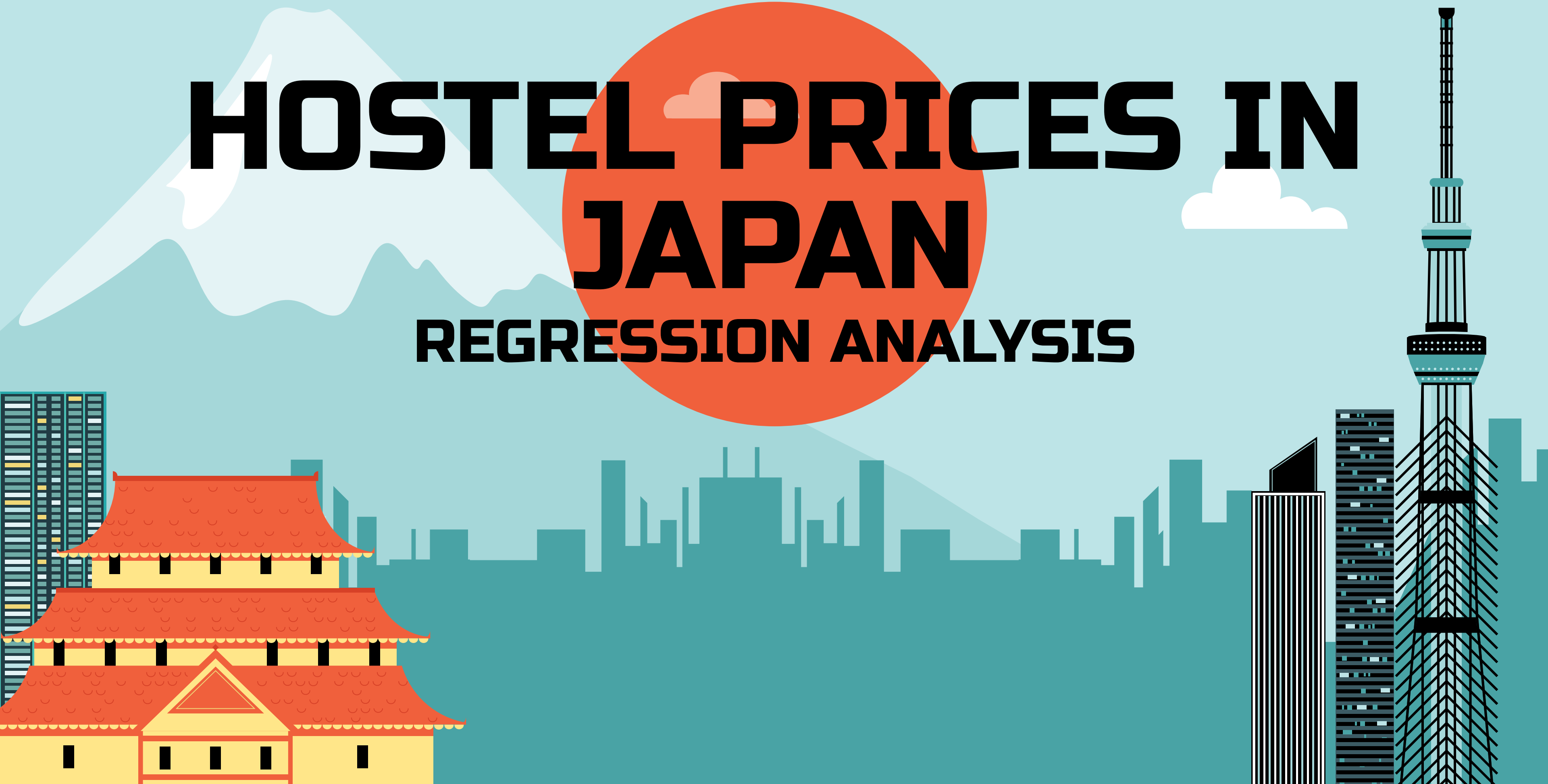
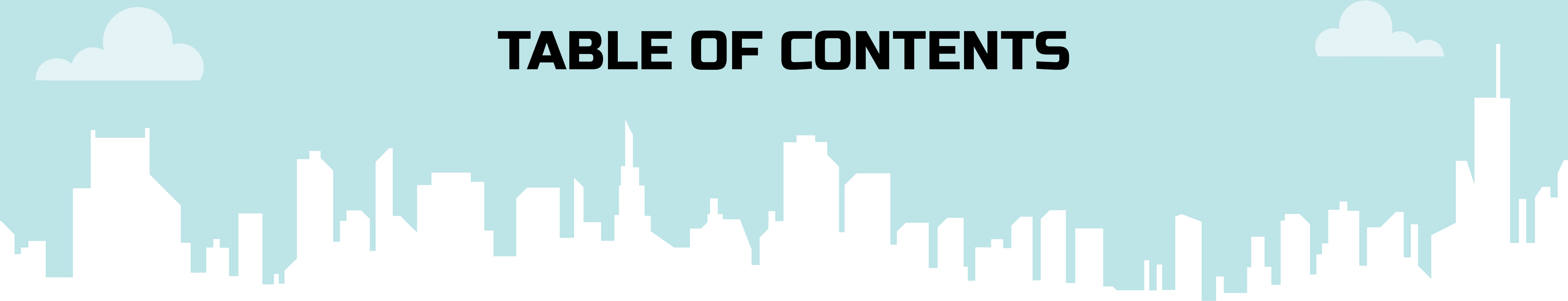


TABLE OF CONTENTS



01.

INTRO + ANALYSIS GOAL

02.

DATA DESCRIPTION

03.

EDA + CLEANING

TABLE OF CONTENTS

04.

MODEL BUILDING

05.

FINAL MODEL

06.

CONCLUSION

WHAT'S A HOSTEL?

A budget-friendly accommodation focused on shared social experiences.

- **MEDIAN NIGHTLY RATE: 2400¥**
- **US DOLLAR EQUIVALENT: $15.26**

Target demographic tends to be younger tourists or solo-travelers.

ANALYSIS GOAL

WHICH PREDICTORS ARE LINEARLY RELATED WITH MINIMUM NIGHTLY HOSTEL PRICES?

STAKEHOLDERS: Hostel Managers

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

DATA DESCRIPTION

- Dataset from Kaggle
- Author scraped 342 real-world observations of 16 variables from HostelWorld.com

DATA DESCRIPTION

Response Variable

- Prices.from: (num) minimum nightly rate in JPY
    - Median = 2400¥
    - Mean = 9228 ¥
    - Min = 1000 ¥
    - Max = 1003200 ¥

DATA DESCRIPTION

Predictors: Score Variables

- Rating Band: (cat, chara) category of rating score
  - Superb, Fabulous, Very Good, Good, Rating
- Summary Score: (ord, num)
  - Atmosphere
  - Cleanliness
  - Facilities
  - Location
  - Security
  - Staff
  - Value for Money

DATA DESCRIPTION

Predictors: Other Variables

- City (cat, chara)
- Distance in km (chara)
- X (num)
- Hostel Name (chara)
- Longitude
- Latitude

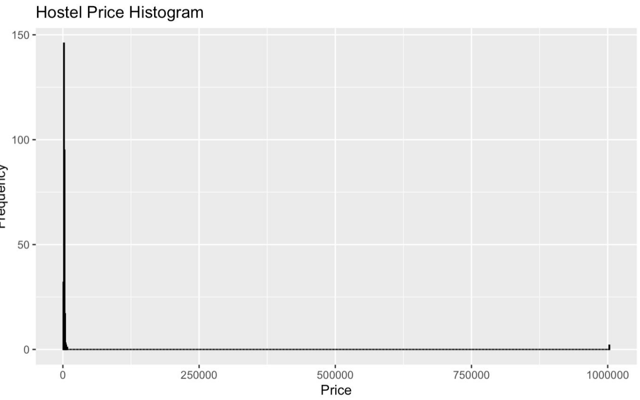
DATA CLEANING

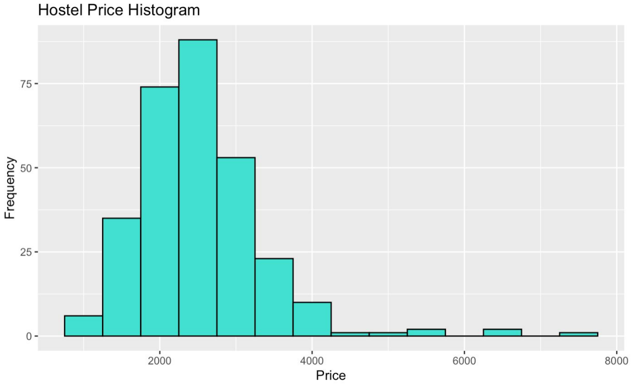
Removed NA's from dataset

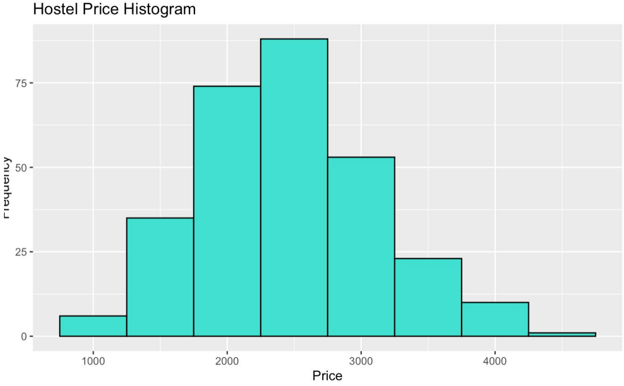
- City (cat, chara)
- Distance in km (num)
- ~~X (num)~~
- ~~Hostel Name (chara)~~
- ~~Longitude~~
- ~~Latitude~~

Converted distance from character to numerical.

EDA: RESPONSE VARIABLE

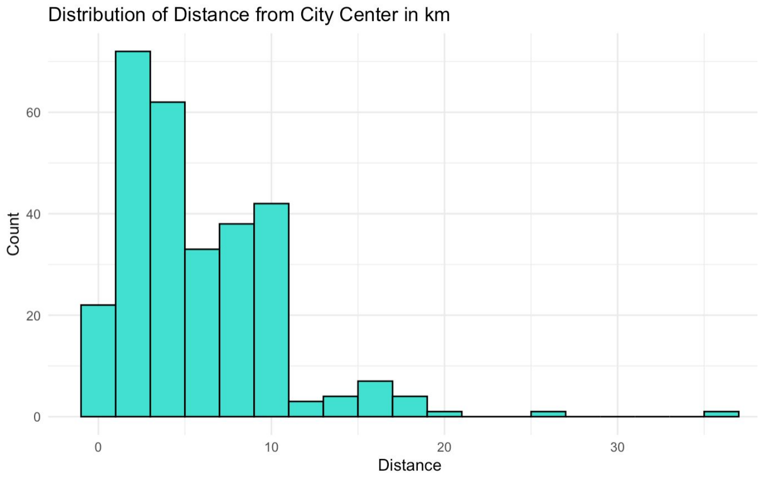


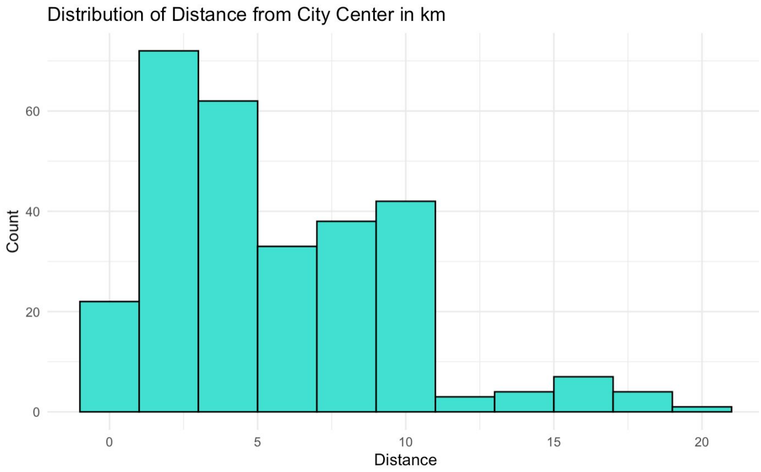




- **Cleaned hostel prices by filtering out expensive outliers**
- **Prices is now somewhat normal with slight right skew.**

EDA: DISTANCE

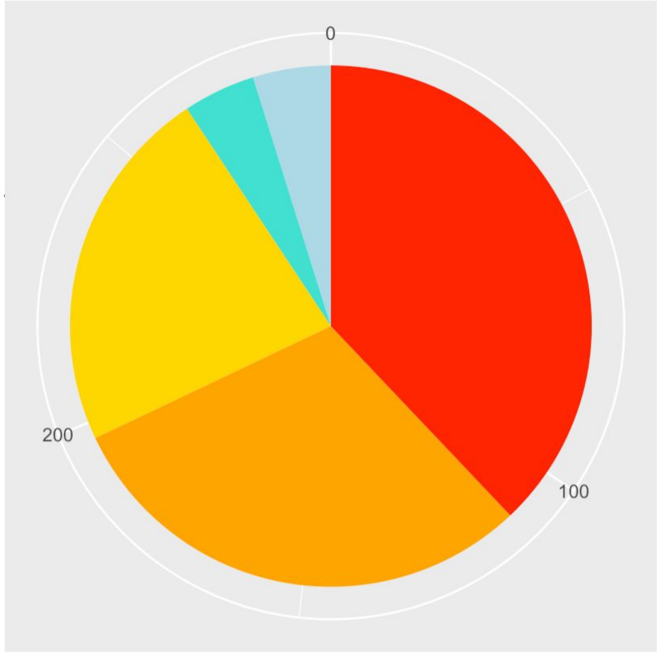




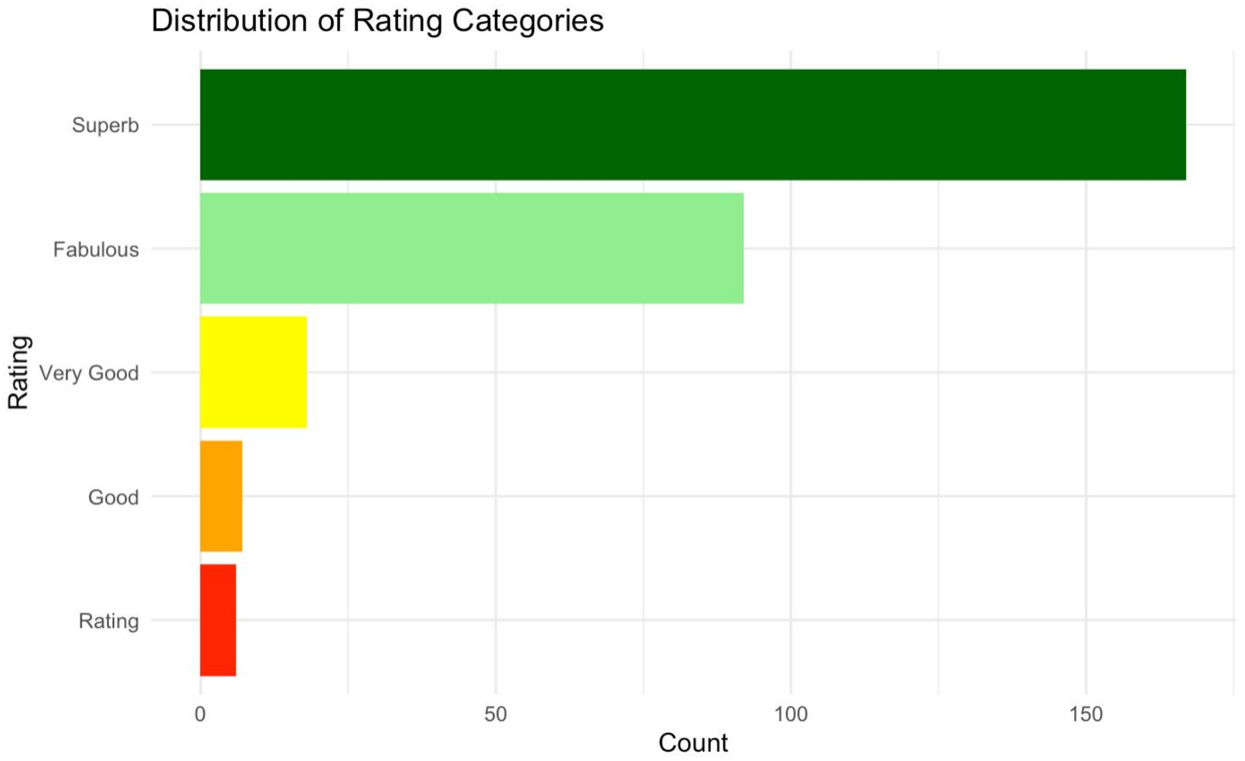
- **Cleaned distance by filtering out entries > 21km**

EDA: CITY DISTRIBUTION

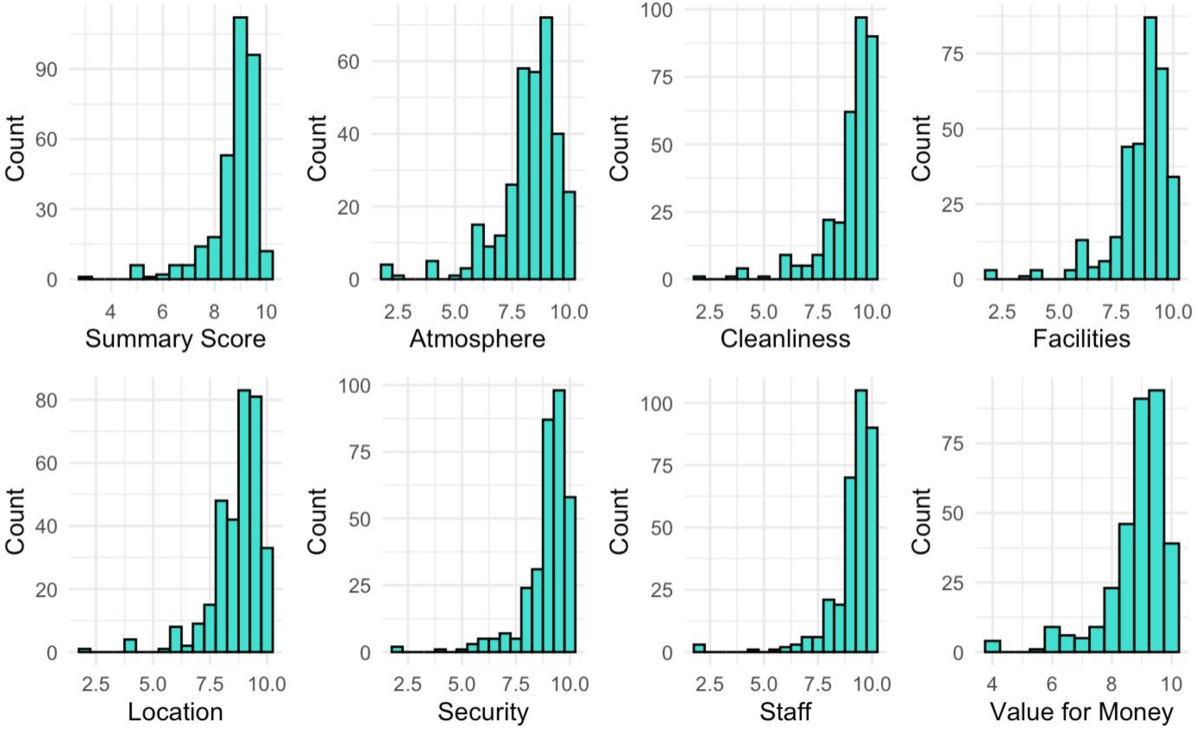
38%	Tokyo
30%	Osaka
22%	Kyoto
4%	Fukuoka-City
4%	Hiroshima



EDA: RATING CATEGORY DISTRIBUTION



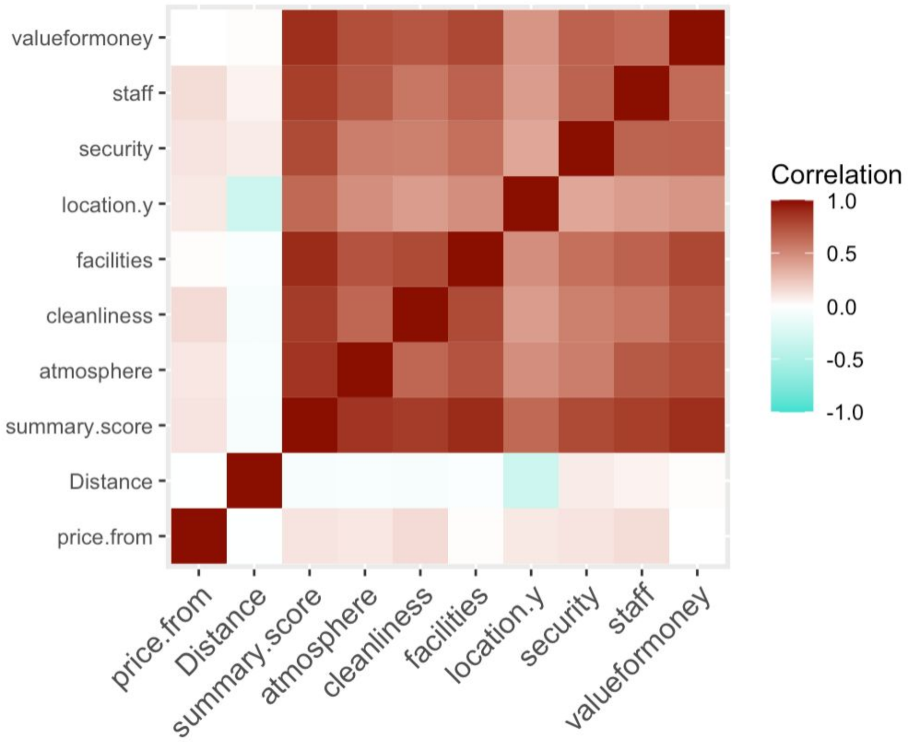
EDA: SCORE VARIABLE DISTRIBUTIONS



EDA: CORRELATION AND MULTICOLLINEARITY

- **Extremely high correlations between summary score and all score predictors.**

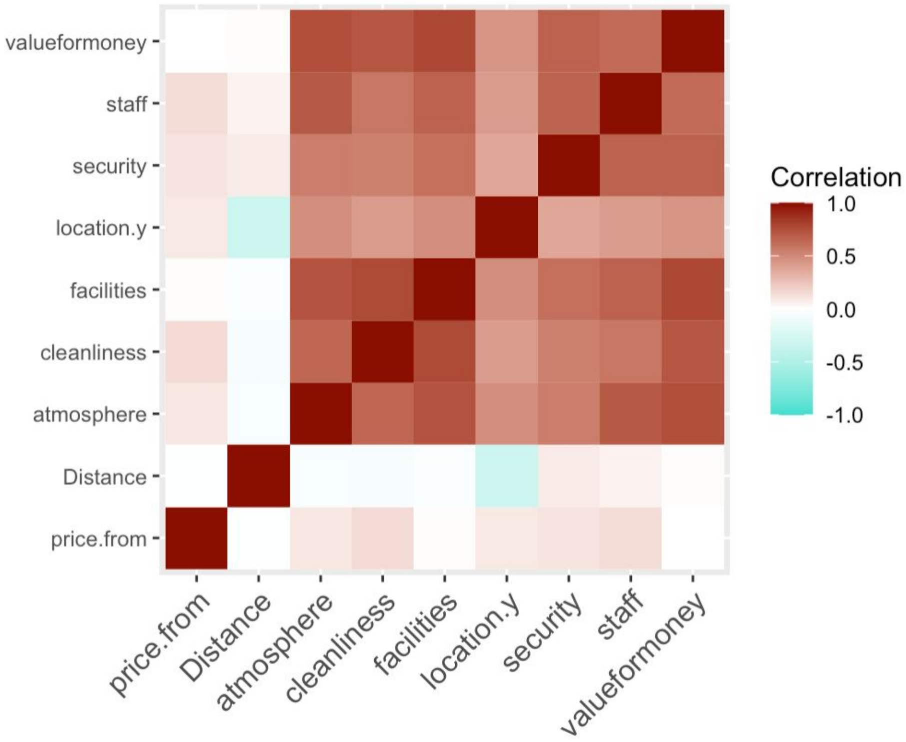
- **Further cleaned data by removing summary score.**



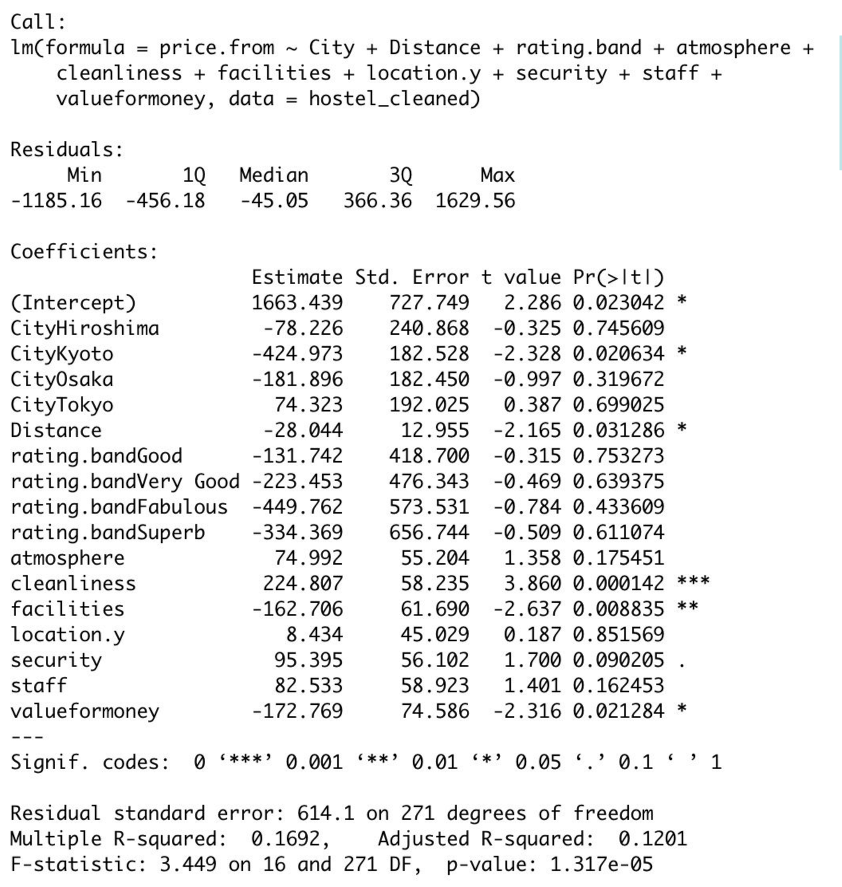
EDA: CORRELATION AND MULTICOLLINEARITY

- **With summary score removed, there's a smaller degree of strong correlation.**

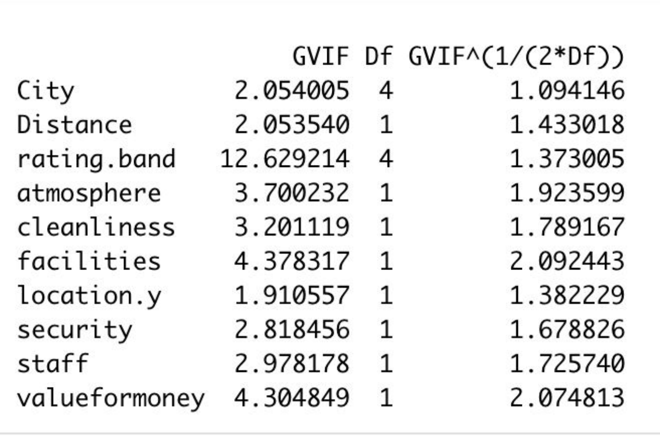
- **Score variables are highly correlated to each other, but I'm keeping them.**



MODEL BUILDING: FULL MODEL



```
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
```

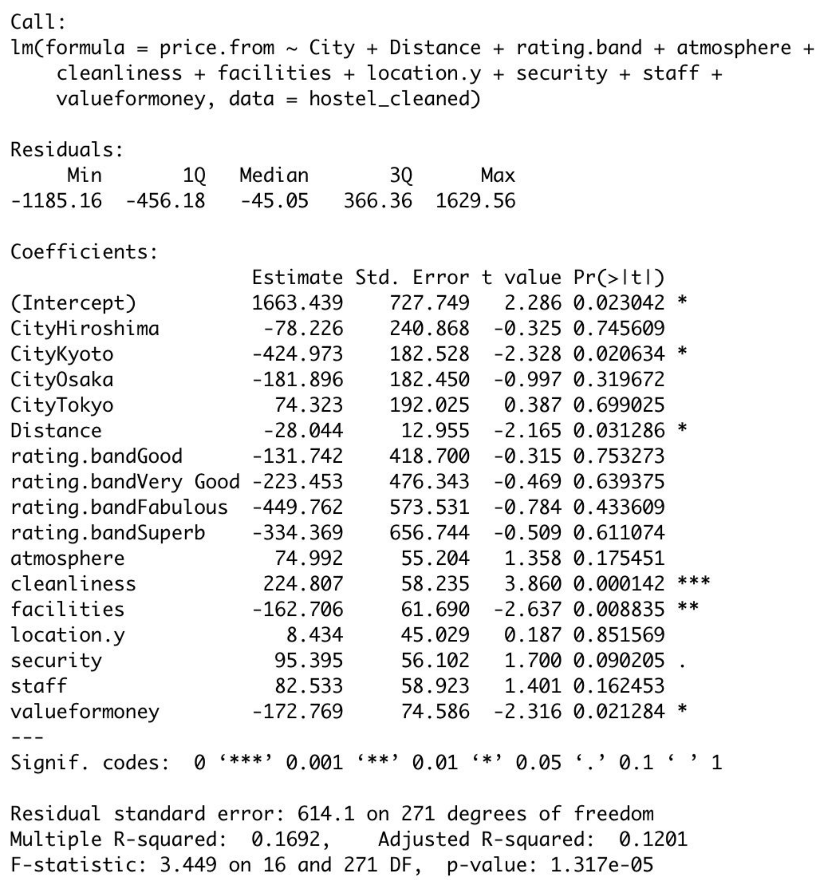


```
                   GVIF Df GVIF^(1/(2*Df))
City           2.054005  4        1.094146
Distance       2.053540  1        1.433018
rating.band   12.629214  4        1.373005
atmosphere     3.700232  1        1.923599
cleanliness    3.201119  1        1.789167
facilities     4.378317  1        2.092443
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
```

- **Full model has an R squared of .1692 and adj. R squared of .1201.**

- **VIF indicates rating.band contributes to multicollinearity**

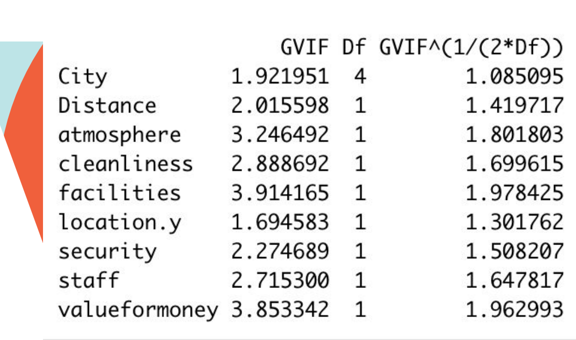
MODEL BUILDING: FULL MODEL



```
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
```

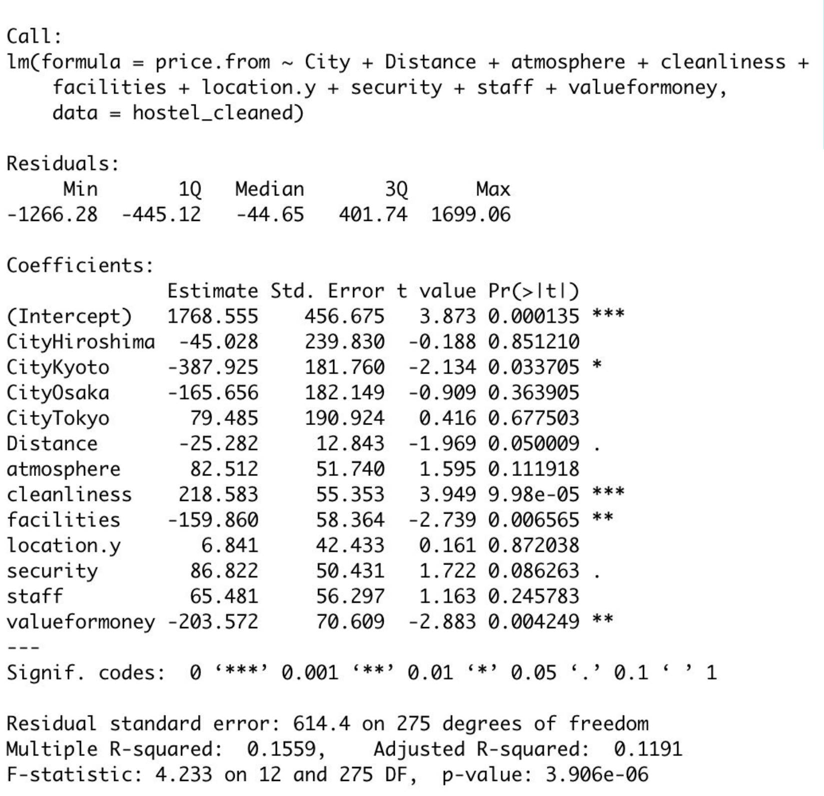
- **Significant Predictors (a = .10)**
  - **City: Kyoto**
  - **Distance**
  - **Cleanliness**
  - **Facilities**
  - **Security**
  - **Value for Money**
- **Most significant predictors for hostel price are cleanliness and facilities**

MODEL BUILDING: REDUCED MODEL 1



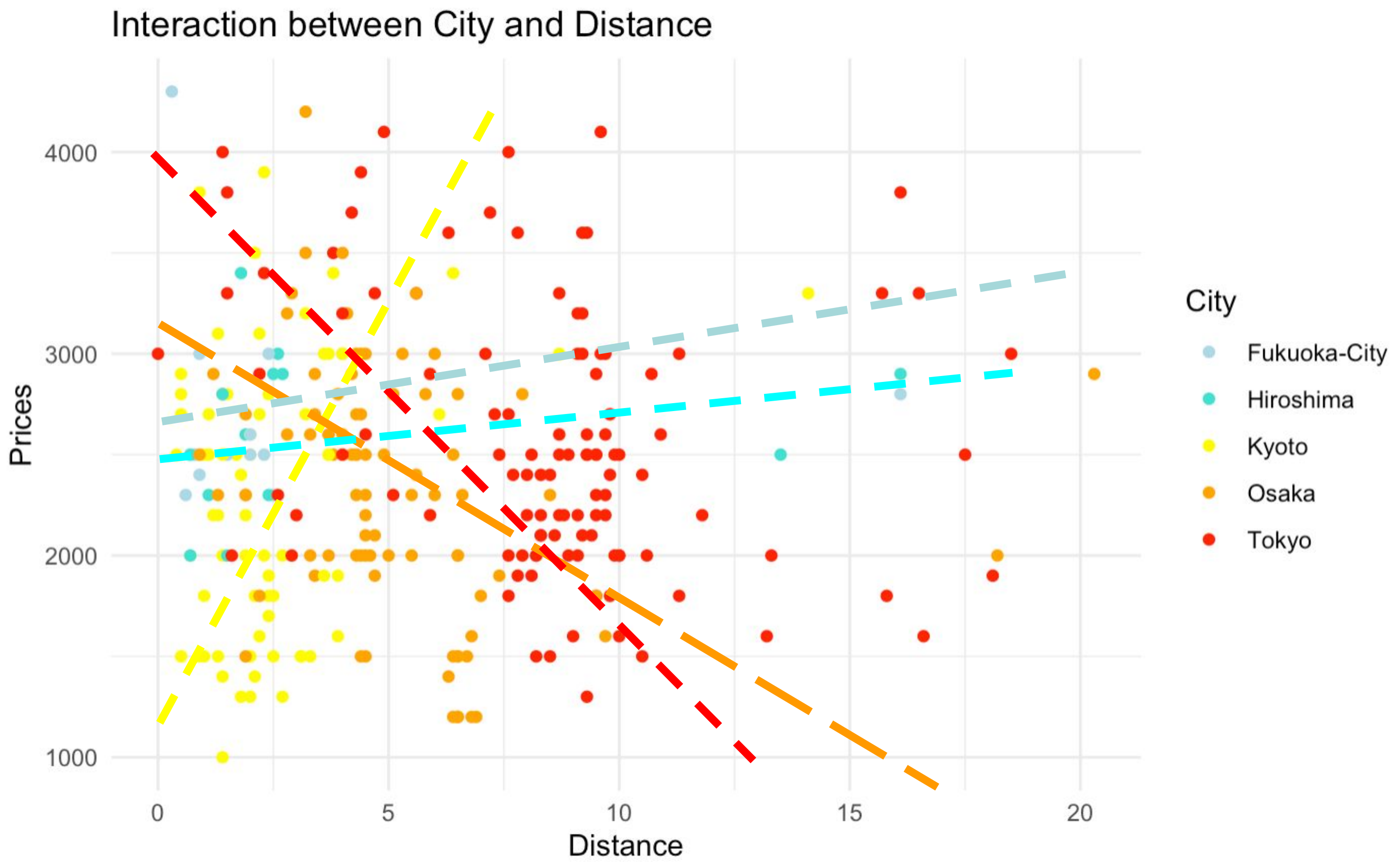
```
                  GVIF Df GVIF^(1/(2*Df))
City          1.921951  4        1.085095
Distance      2.015598  1        1.419717
atmosphere    3.246492  1        1.801803
cleanliness   2.888692  1        1.699615
facilities    3.914165  1        1.978425
location.y    1.694583  1        1.301762
security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

- **R squared decreased by 1.33%.**
- **All GVIF scores below 10, multicollineairty is no longer an issue.**
- **I'm moving forward using this model**

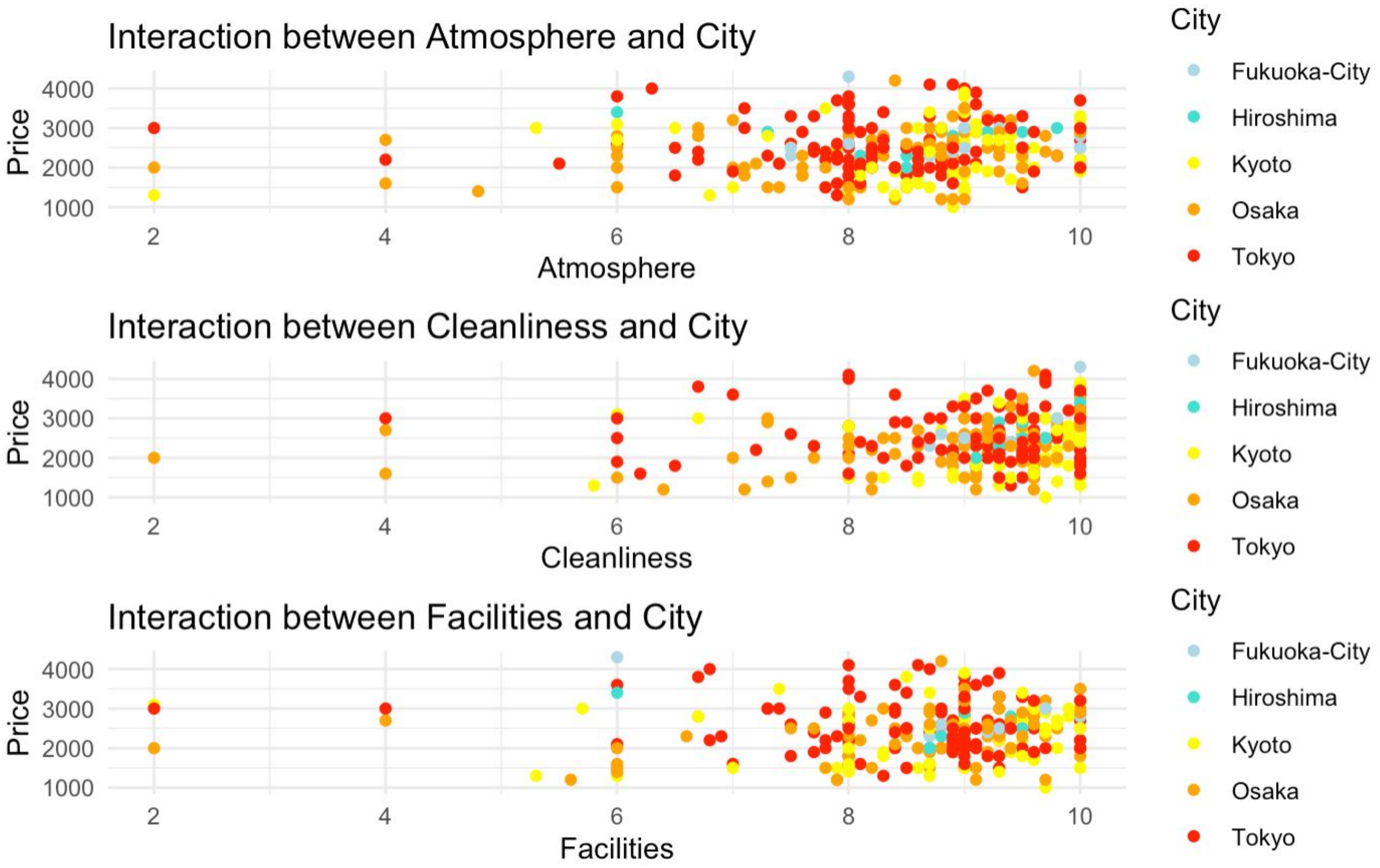


```
Call:
lm(formula = price.from ~ City + Distance + atmosphere + cleanliness + 
    facilities + location.y + security + staff + valueformoney, 
    data = hostel_cleaned)

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
```

MODEL BUILDING: INTERACTIONS

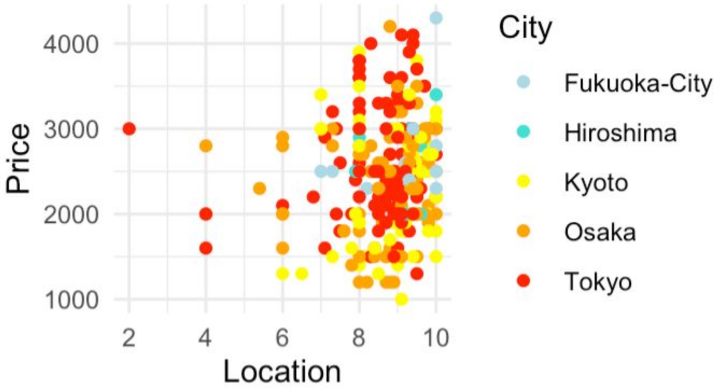


MODEL BUILDING: INTERACTIONS

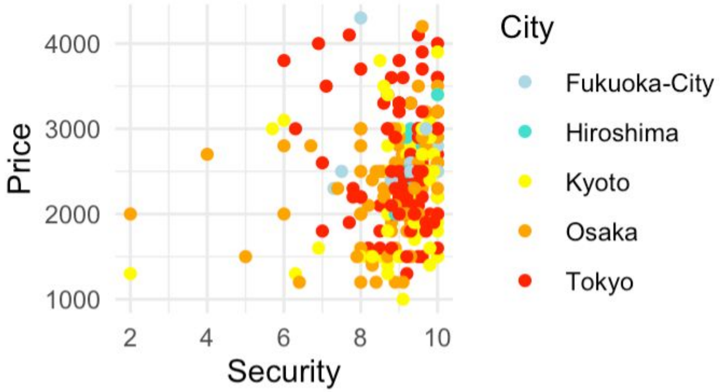


MODEL BUILDING: INTERACTIONS

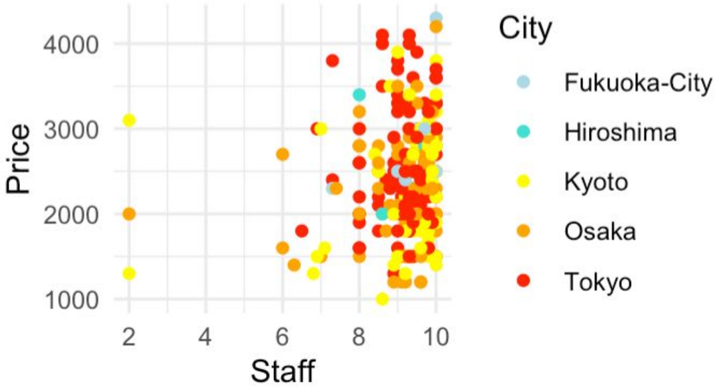
Interaction between Location and City



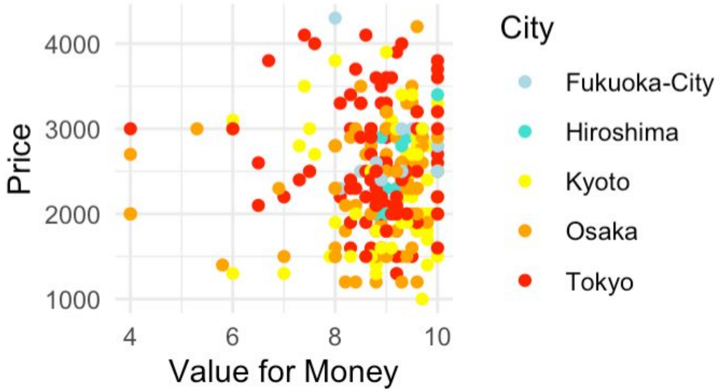
Interaction between Security and City



Interaction between Staff and City

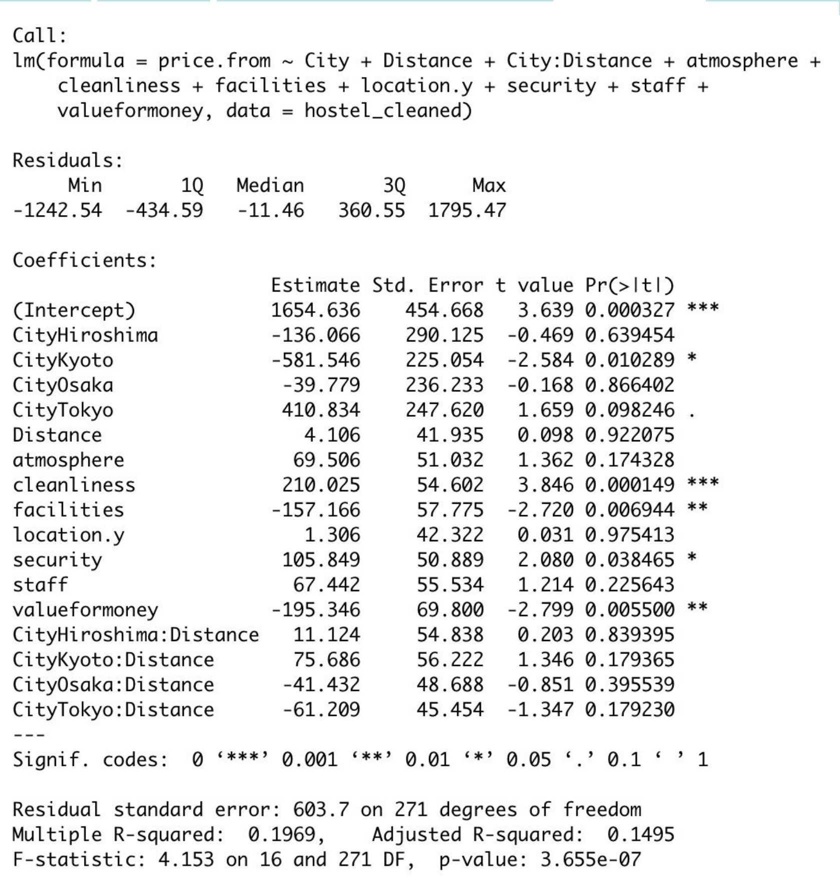


Interaction Value for Money and City



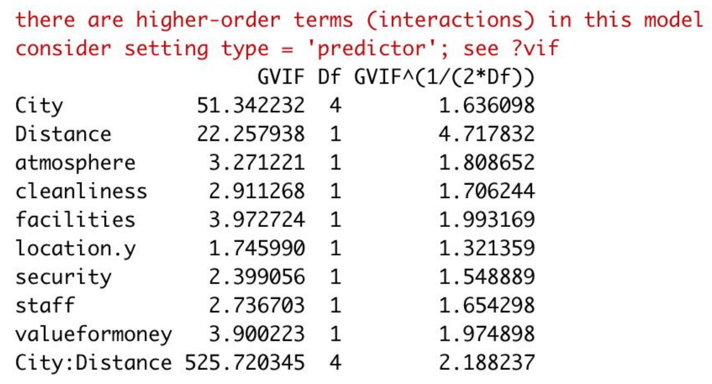
MODEL BUILDING: CITY x DISTANCE INTERACTION

- **R squared improved by 4.1%.**
- **Same predictors are statistically significant**



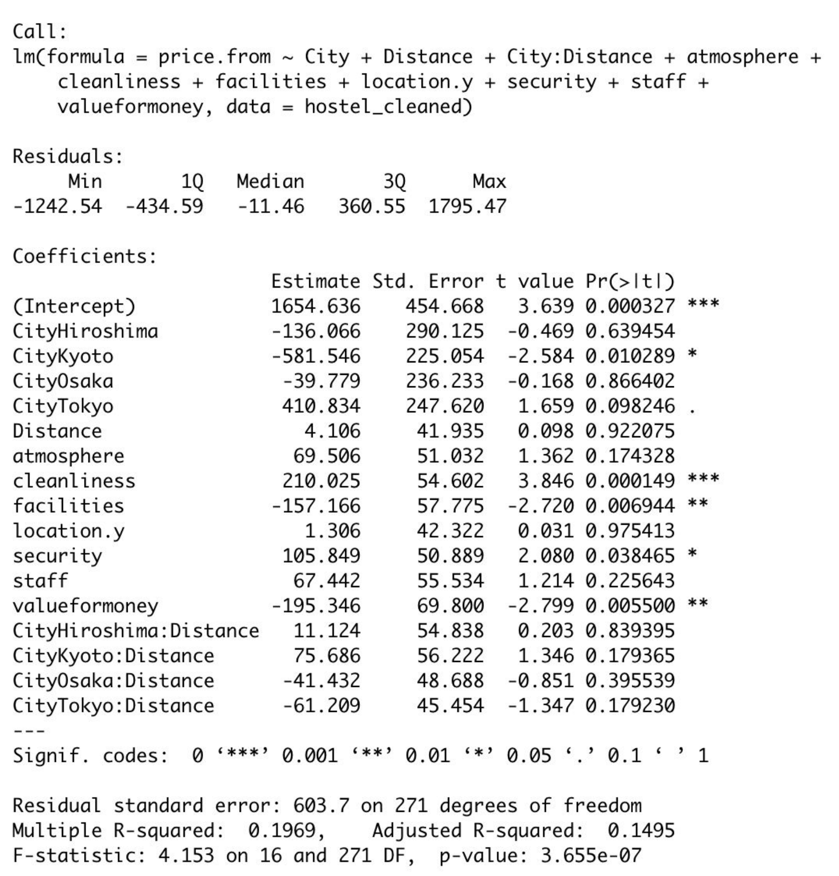
```
Call:
lm(formula = price.from ~ City + Distance + City:Distance + atmosphere +
    cleanliness + facilities + location.y + security + staff +
    valueformoney, data = hostel_cleaned)

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
```

MODEL BUILDING: CITY x DISTANCE INTERACTION



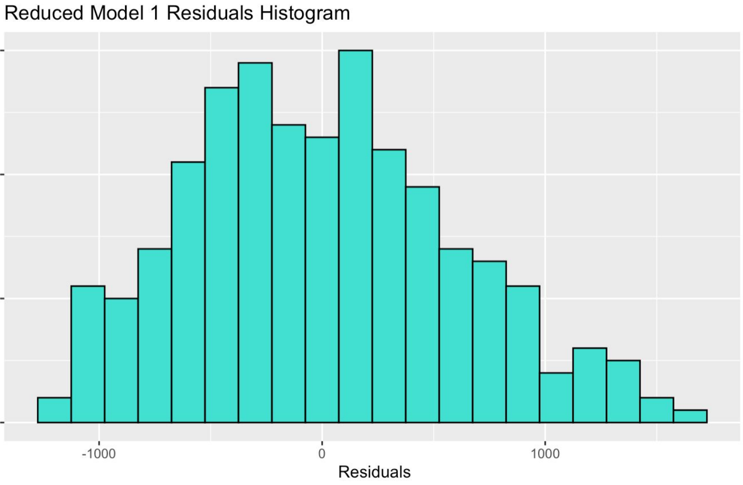
```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
                  GVIF Df GVIF^(1/(2*Df))
City         51.342232  4        1.636098
Distance     22.257938  1        4.717832
atmosphere    3.271221  1        1.808652
cleanliness   2.911268  1        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
```

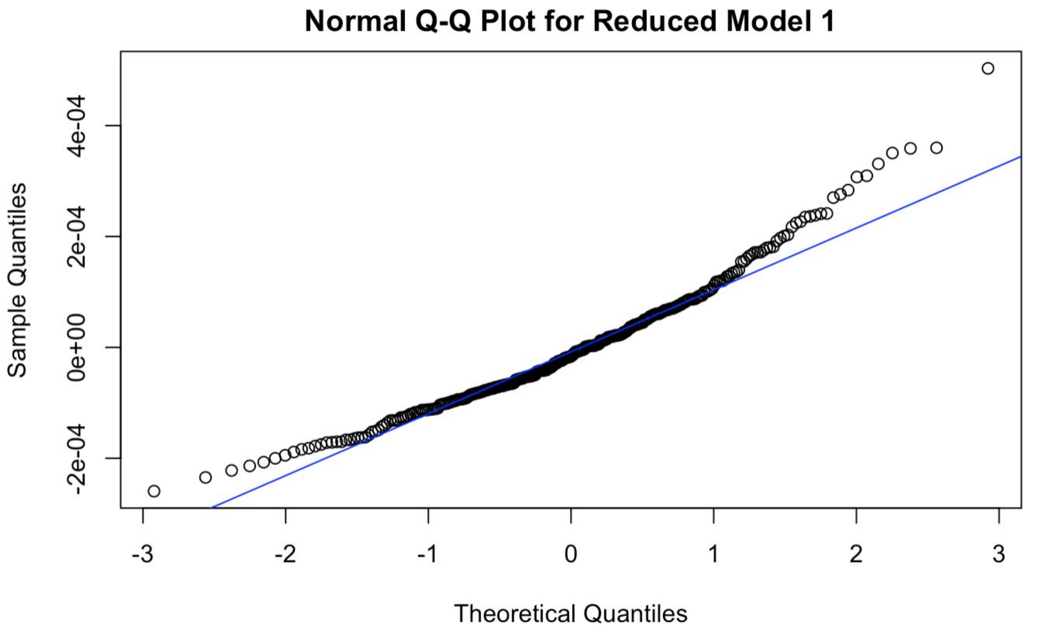
- **EXTREMELY HIGH GVIF SCORES!!**
- **Adding City:Distance inflated the variance of my estimates, and made my model unstable**

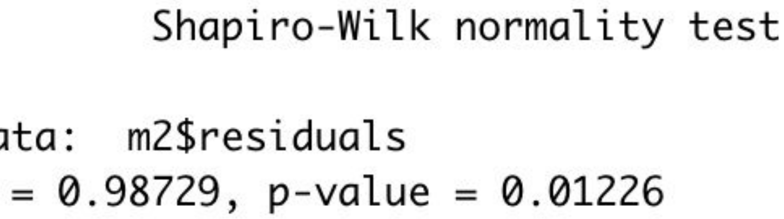


```
Call:
lm(formula = price.from ~ City + Distance + City:Distance + atmosphere +
    cleanliness + facilities + location.y + security + staff +
    valueformoney, data = hostel_cleaned)

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
```

ASSUMPTION CHECKS ON REDUCED MODEL 1



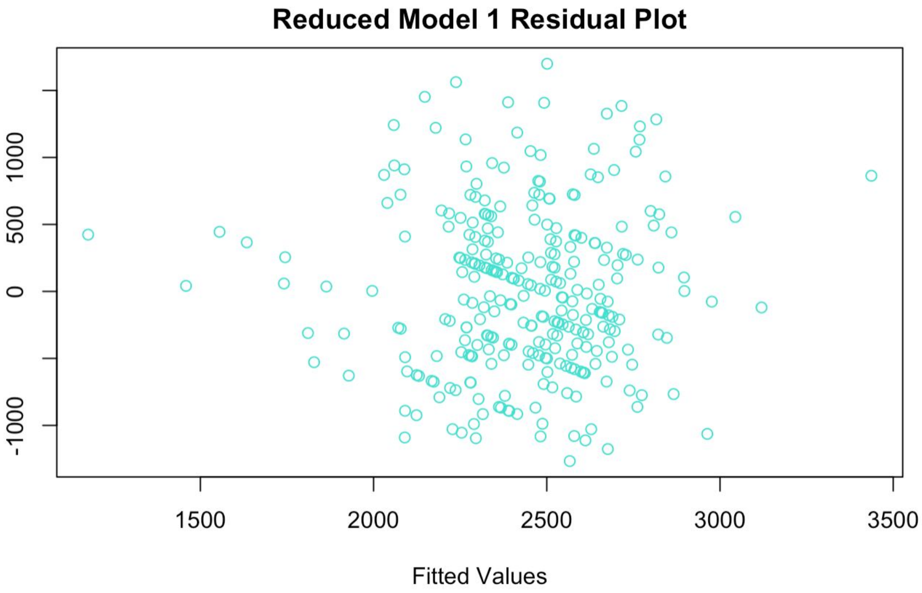


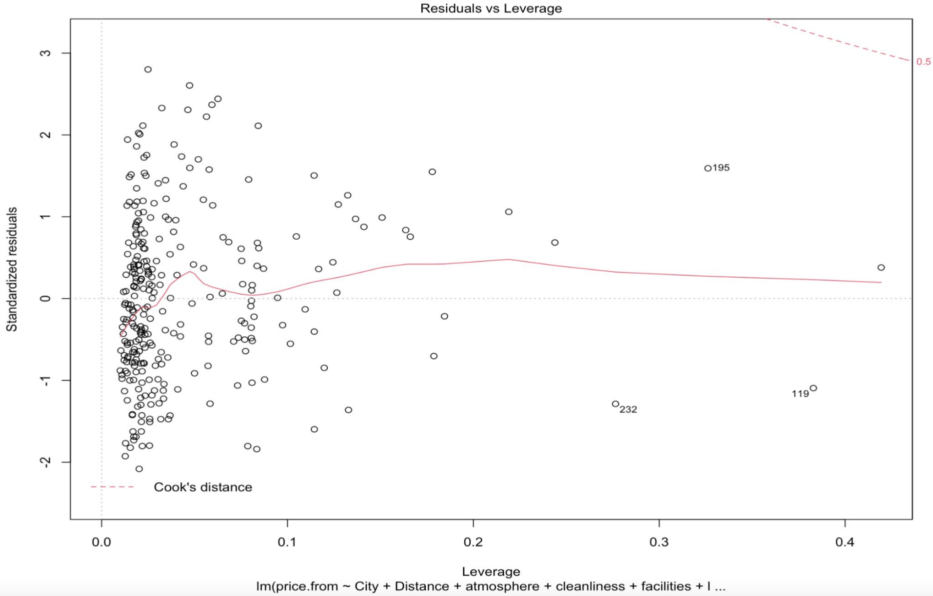


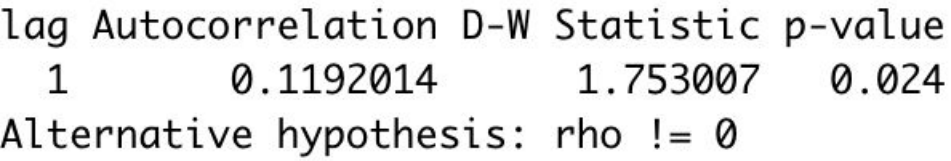
```
          Shapiro-Wilk normality test

ata:  m2$residuals
= 0.98729, p-value = 0.01226
```

Normality is violated

ASSUMPTION CHECKS ON REDUCED MODEL 1







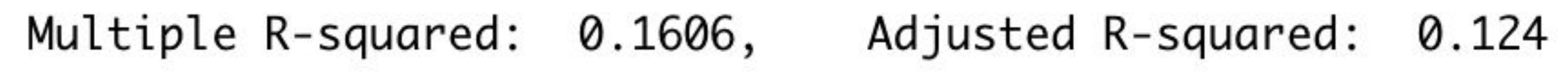
```
lag Autocorrelation D-W Statistic p-value
  1       0.1192014      1.753007   0.024
Alternative hypothesis: rho != 0
```

Linearity and constant variance approximately hold.

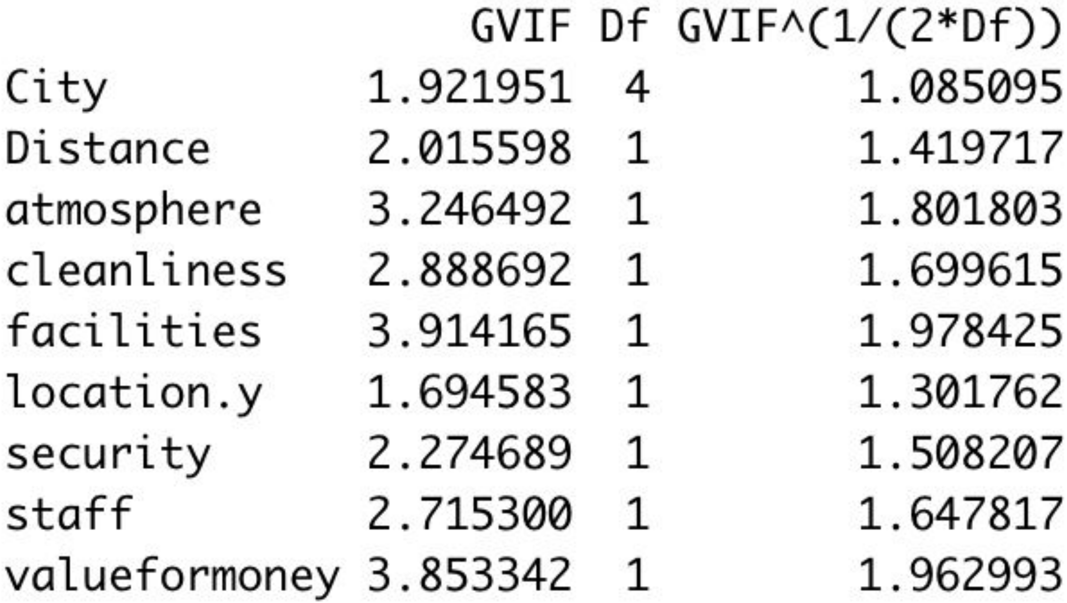
Independence is violated.

MODEL TRANSFORMATIONS

I chose to move forward with a square root transformation on prices after testing.

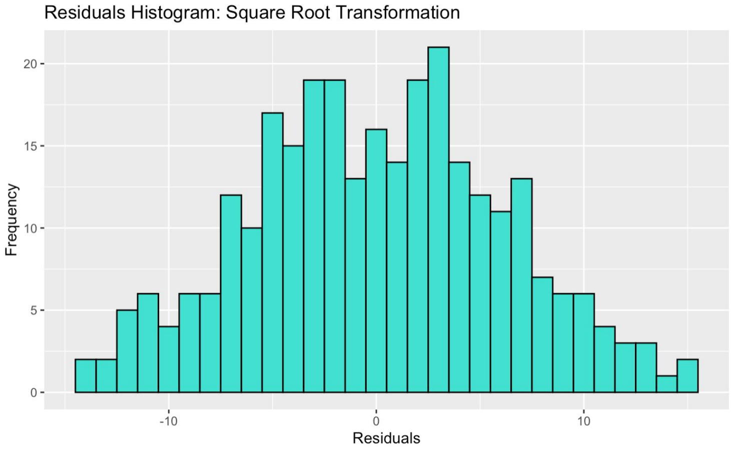


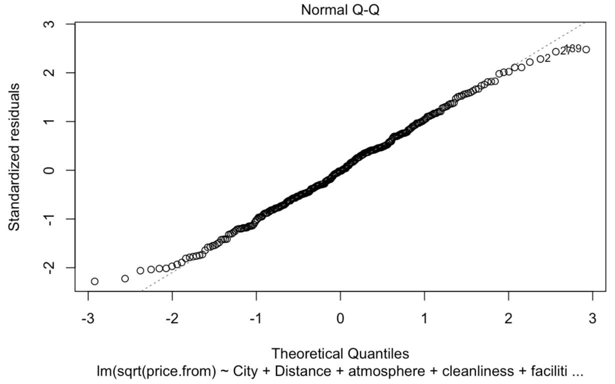
```
Multiple R-squared:  0.1606,	Adjusted R-squared:  0.124
```

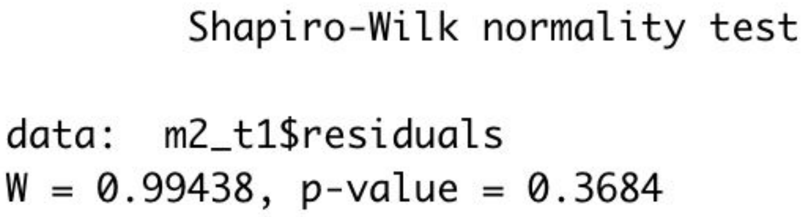


```
                   GVIF Df GVIF^(1/(2*Df))
City           1.921951  4        1.085095
Distance       2.015598  1        1.419717
atmosphere     3.246492  1        1.801803
cleanliness    2.888692  1        1.699615
facilities     3.914165  1        1.978425
location.y     1.694583  1        1.301762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

MODEL TRANSFORMATIONS



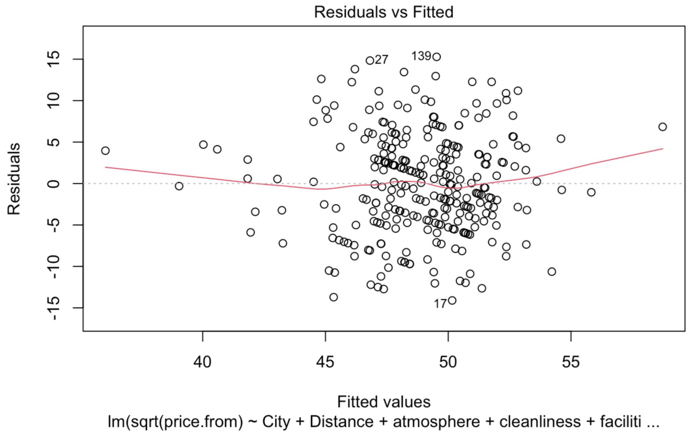


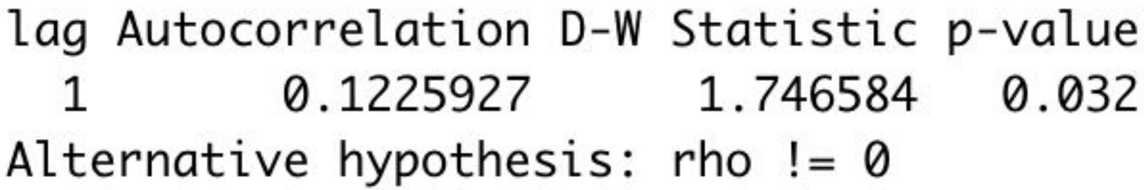


```
Shapiro-Wilk normality test

data:  m2_t1$residuals
W = 0.99438, p-value = 0.3684
```

Normality assumption holds

MODEL TRANSFORMATIONS

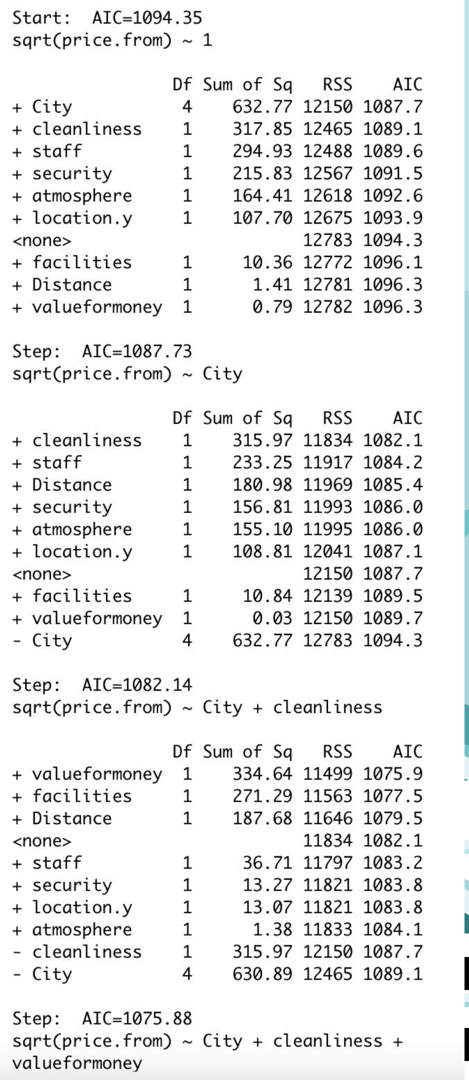




```
lag Autocorrelation D-W Statistic p-value
  1       0.1225927      1.746584   0.032
Alternative hypothesis: rho != 0
```

Constant variance and linearity approximately hold.

Independence violated regardless of transformation



```
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
```

STEPWISE SELECTION

Reduced model uses City, Distance, atmosphere, cleanliness, facilities, location, security, staff, and value for money.

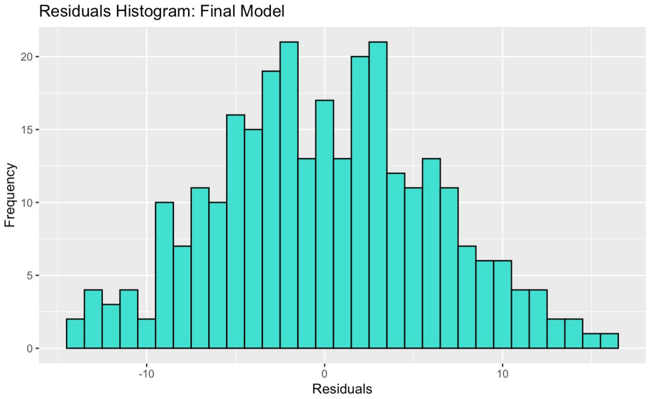
R squared = .1606, adj R squared = .1191

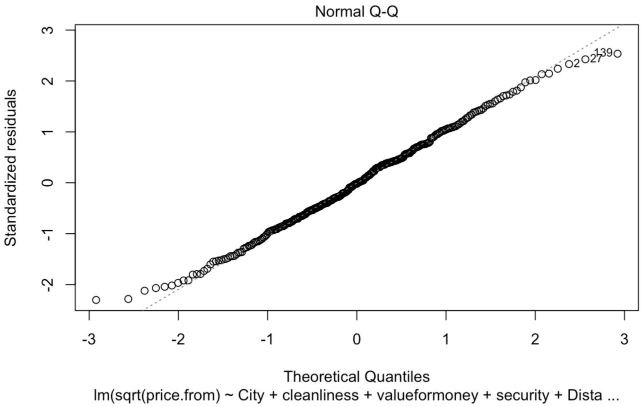
Stepwise model removes staff and location

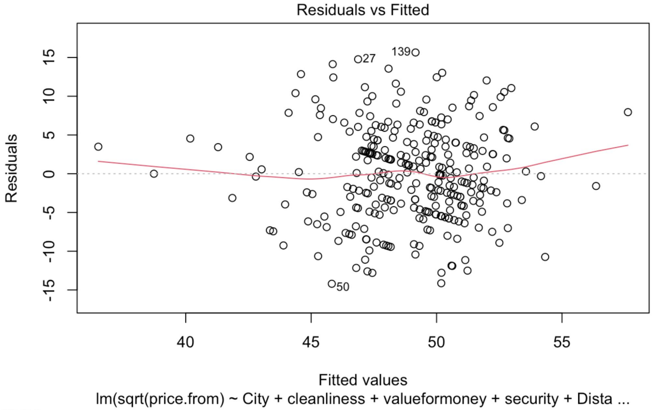
R squared = .1569, adj. R squared .1265

I'm selecting the stepwise model for simplicity.

FINAL MODEL ASSUMPTIONS


Residuals Histogram: Final Model

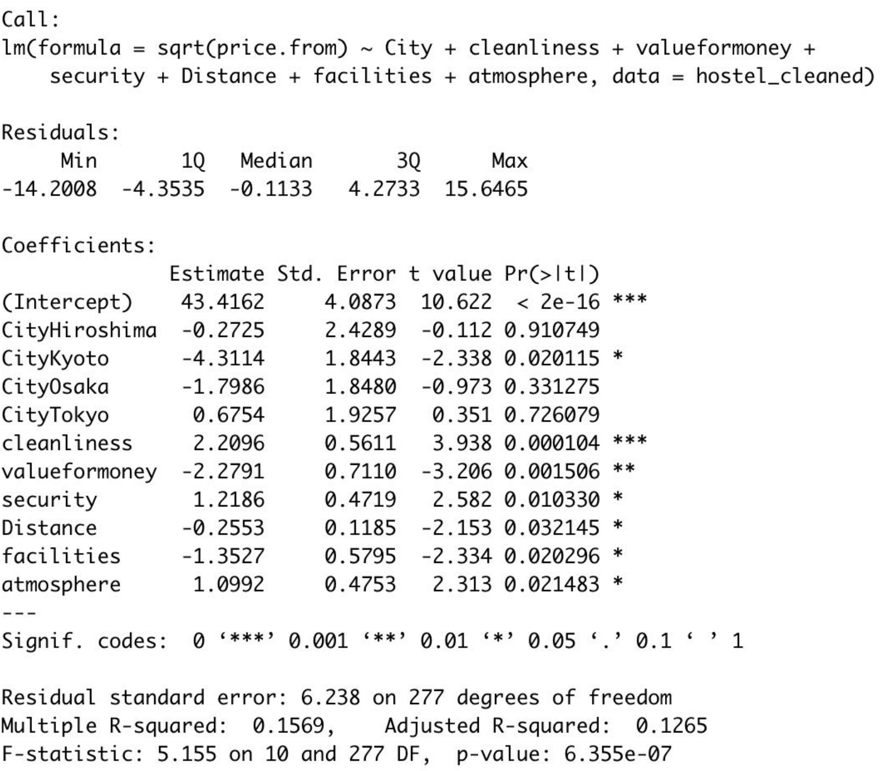

Normal Q-Q


Residuals vs Fitted

```
     Shapiro-Wilk normality test

data:  stepwise_model$residuals
W = 0.995, p-value = 0.4752

 lag Autocorrelation D-W Statistic p-value
   1       0.1275398      1.736661   0.018
 Alternative hypothesis: rho != 0
```

Final model satisfies all assumptions except independence.

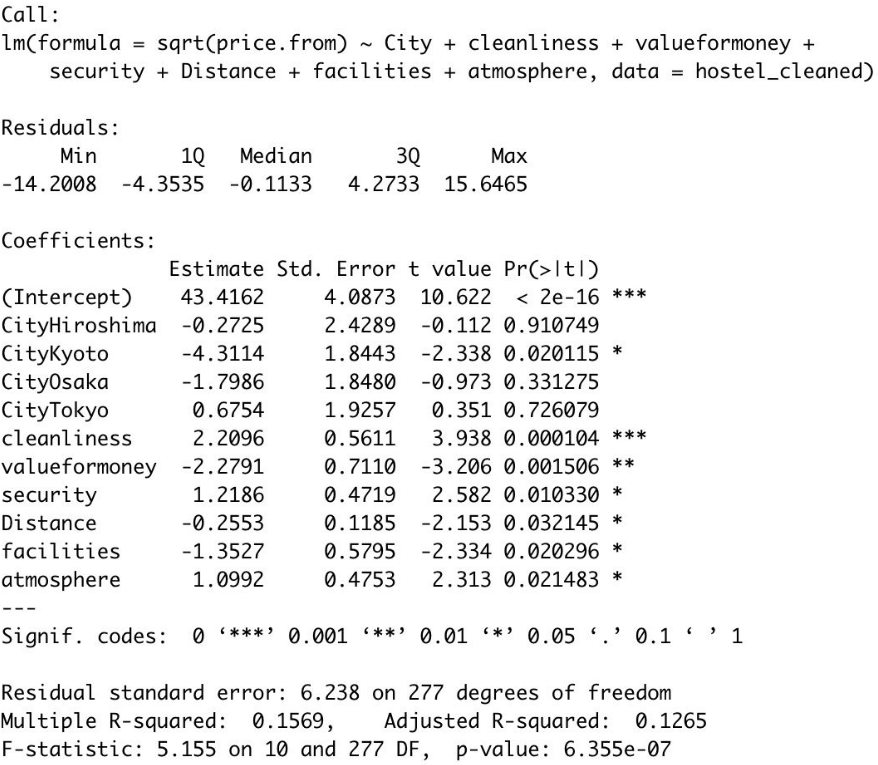
FINAL MODEL OUTPUT



```
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
```

- **Significant Predictors**
  - **City: Kyoto**
  - **Cleanliness**
  - **Value for money**
  - **Security**
  - **Distance**
  - **Facilities**
  - **Atmosphere**
- **The most significant predictors for the square root of hostel price are cleanliness and value for money.**

FINAL MODEL OUTPUT



```
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
```

- **R squared = .1569**
- **Adjusted R squared = .1265**
- **15.69% of variance of the square root of minimum nightly hostel rates in Japan can be explained by this model.**

CONCLUSION

WHICH PREDICTORS ARE LINEARLY RELATED WITH MINIMUM NIGHTLY HOSTEL PRICES?

The city of Kyoto, distance, value for money, and facilities are negatively related to hostel prices.

Cleanliness, security, and atmosphere are positively related to hostel prices.

Allocate resources to improving cleanliness, security, and atmosphere to justify charging higher nightly rates.

LIMITATIONS

Relatively low R squared value only explains a small portion of variability.

Data was scraped before 2020 and prices may be vastly different today.

This dataset is heavily focused on subjective rating scores.

THANK YOU!