GLM: Coursework Assignment 1

By: Saif Rehman

1. **Introduction**

I am working with sales data for the city of Northampton, Massachusetts, which is a city located in the United States with a population of more than 28,500 people for the 2014 year. Geographically, this city lies beside the Connecticut River and is about 100 miles west of Boston, Massachusetts which is a major city in the United States. The city of Northampton used to have a train route running through it which connected Boston and New Haven, Connecticut until the passenger service was discontinued in the 1920’s and freight service was also discontinued in the 1960’s.

This train route was converted in to a three-mile-long section in 1984 to make a rail trail which has multiple uses for the citizens such as: cycling, walking, and running. The main purpose of this analysis is to identify whether house prices are higher for houses closer to the trail versus the houses that are more distant from the trail. More generally, it is to determine whether the trail brings economic

1. **Overview of the Data**

The data that was in the dataset contained no missing values as it was already a simplified dataset and consisted of these original variables:

housenum: a unique identifying house number

acre: Number of acres that the property encompasses

bedgroup: States how many beds there are (either “1-2”, “3”, or “4+”)

bikescore: A score given from 0-100 for the bike friendliness (higher the score the better)

distance: Distance (in feet) to the nearest entry point to the rail trail

garage\_spaces: Number of garage spaces (0-4)

no\_full\_baths: Number of full baths (includes shower or bathtub)

no\_half\_baths: number of half baths (no shower or bath present)

no\_rooms: Number of rooms

price2014: Zillow price estimate from 2014 (in thousands of dollars)

squarefeet: square footage of interior finished space (in thousands of square feet)

walkscore: A score given from 0-100 for walking friendliness (higher the score the better)

zip: zipcode of the house location (1060 = Northampton, 1062 = Florence)

Walkscore and bikescore are both scores that range from 0-100 and are neighborhood characteristics. A high walkscore such as one in the 90’s is described as a “Walker’s paradise” as stated by Professor Brian Francis where “daily errands do not require a car”. Similarly, a high bikescore such as in the 90’s is described by Professor Brian Francis as “Biker’s paradise: flat as a pancake, excellent bike lanes”.

Those described above are the original variables in the dataset prior to any recoding and transformation of the dataset. After looking carefully at each of the variables, I have transformed some of the original variables into factors due to them either having groups or because they had small enough outputs that it seemed inappropriate to consider them continuous:

bedF: Factor from the bedgroup variable where it has groups of “1-2”, “3”, and “4+”

garageF: Factor from the garage\_spaces variable with its individual values

fullbathF: Factor from no\_full\_baths variable with its individual values

halfbathF: Factor from no\_halfs\_baths variable with its individual values

roomsF: Factor from the no\_rooms variable with its individual values

zipcodeF: Factor from the zip variable with 1060=Northampton and 1062=Florence

Due to changing the variables to factors, this may result in a different resulting model and should be taken under consideration when further analyzing this dataset.

1. **Data Exploration**

Here are the summary statistics of the various variables to get an idea of their means, ranges, and counts (for factors).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Price2014 | Acre | Bikescore | Distance | Garage\_spaces | No\_full\_baths |
| Mean | 293.1 | 0.2574 | 57.28 | 1.11432 | 0.7596 | 1.452 |
| 1st – 3rd Quartile | [212.9, 334.2] | [0.1675, 0.3300] | [36, 77.25] | [0.32879, 1.89579] | [0, 1] | [1, 2] |
| Median | 272.9 | 0.2500 | 54.50 | 0.76042 | 1 | 1 |
| Range | [132.1, 879.3] | [0.05, 0.56] | [18, 97] | [0.03883, 3.97678] | [0, 4] | [1, 4] |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | No\_half\_baths | No\_rooms | Squarefeet | Walkscore | Zip |
| Mean | 0.2212 | 6.615 | 1.566 | 38.88 | 1061 |
| 1st – 3rd Quartile | [0, 0] | [5, 7.25] | [1.206, 1.832] | [14.75, 60.75] | [1060, 1062] |
| Median | 0 | 6.5 | 1.516 | 36 | 1062 |
| Range | [0, 1] | [4, 14] | [0.524, 4.030] | [2, 94] | [1060, 1062] |

|  |  |
| --- | --- |
|  | garageF |
| 0 garages | 51 |
| 1 garage | 29 |
| 2 garages | 23 |
| 4 garages | 1 |

|  |  |
| --- | --- |
|  | fullbathF |
| 1 full bath | 63 |
| 2 full bath | 36 |
| 3 full bath | 4 |
| 4 full bath | 1 |

Factors which show the frequency at each level:

|  |  |
| --- | --- |
|  | roomF |
| 4 rooms | 7 |
| 5 rooms | 20 |
| 6 rooms | 25 |
| 7 rooms | 26 |
| 8 rooms | 16 |
| 9 rooms | 7 |
| 11 rooms | 1 |
| 12 rooms | 1 |
| 14 rooms | 1 |

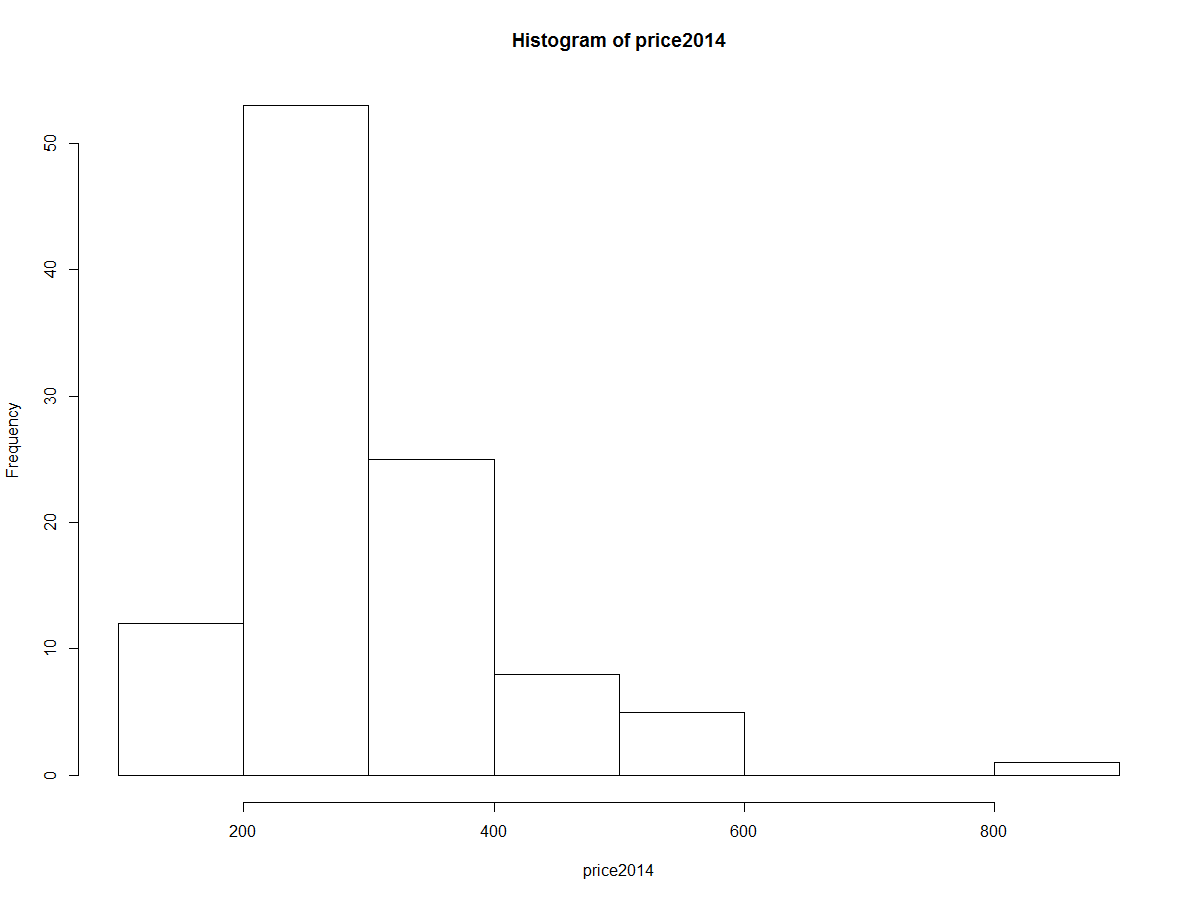
|  |  |
| --- | --- |
|  | bedF |
| 1-2 beds | 16 |
| 3 beds | 52 |
| 4+ beds | 36 |

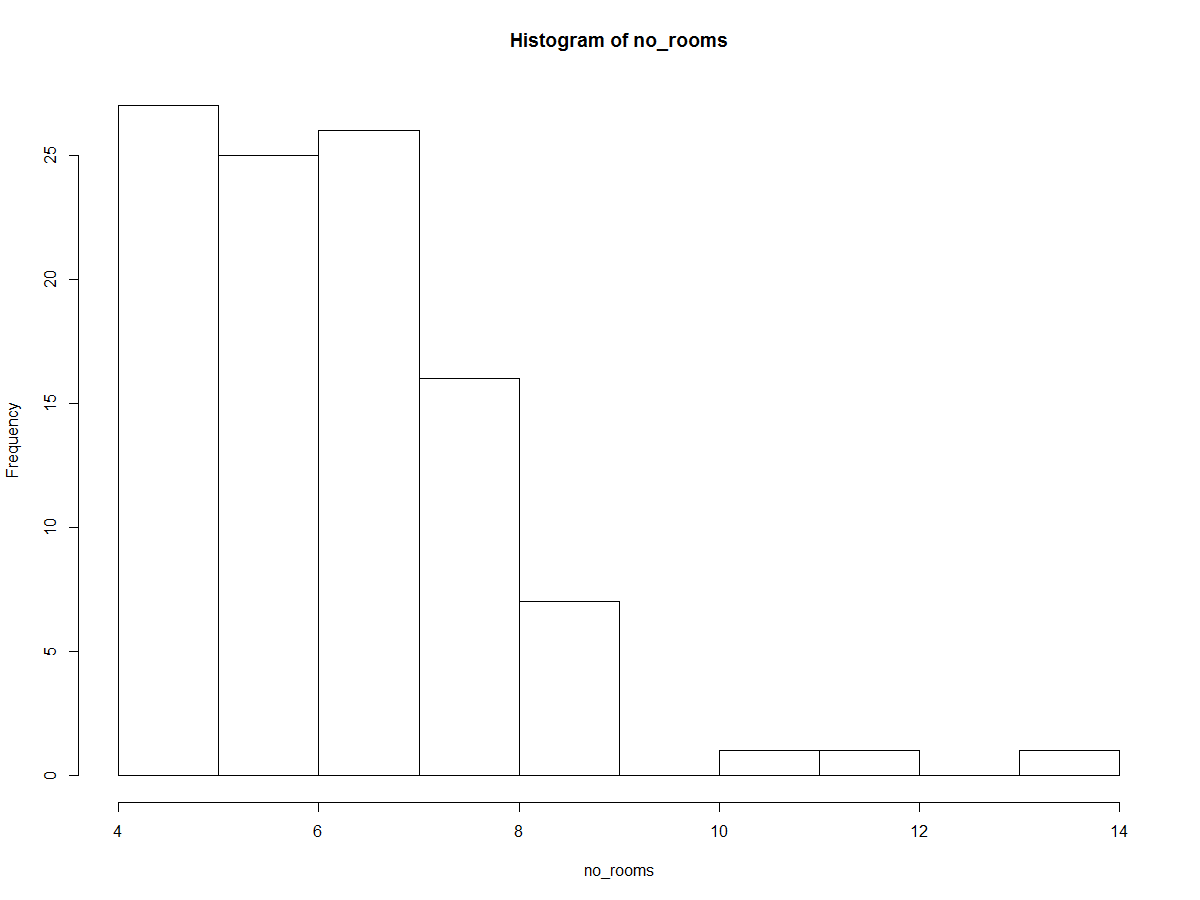
|  |  |
| --- | --- |
|  | zipF |
| 1060 zip | 43 |
| 1062 zip | 61 |

|  |  |
| --- | --- |
|  | No\_half\_baths |
| 0 half baths | 81 |
| 1 half bath | 23 |

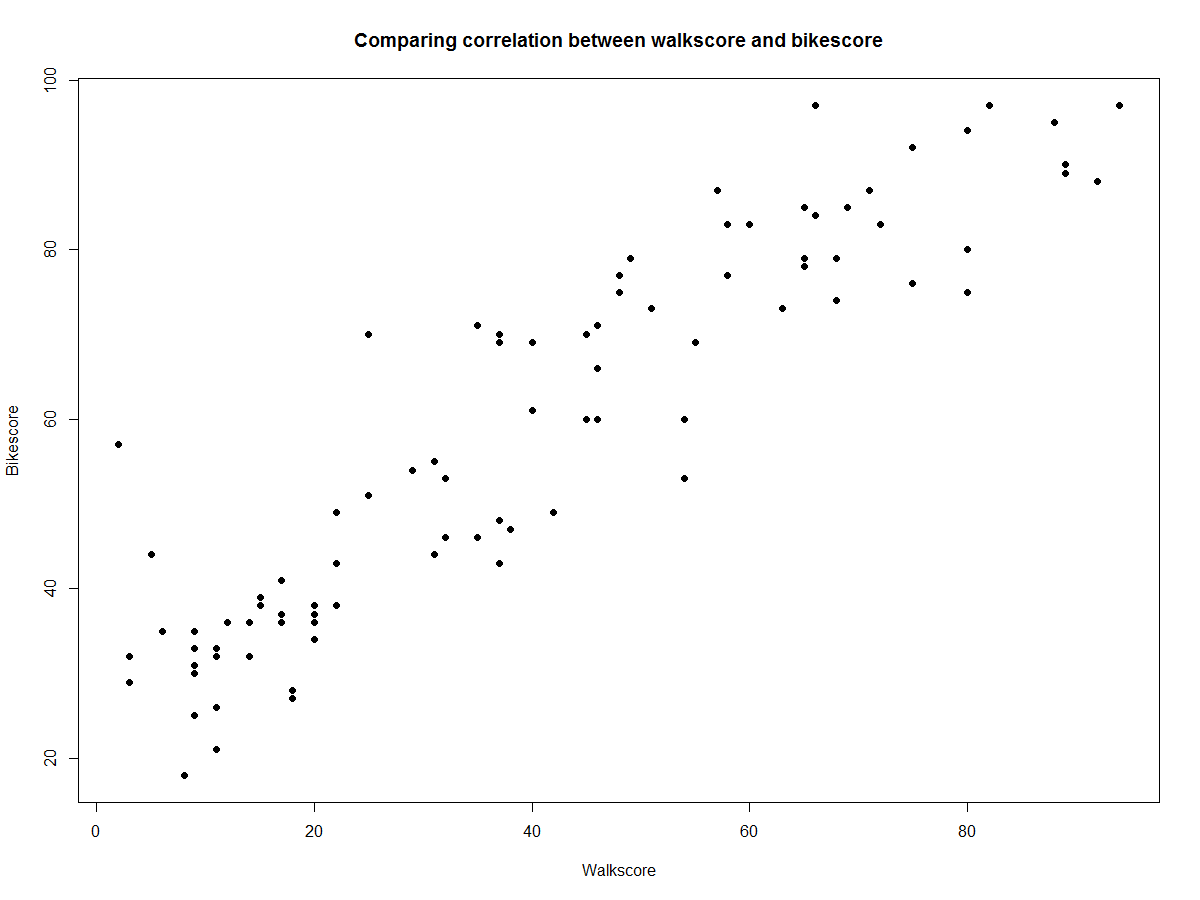
Here is some exploratory analysis to see the distribution of some variables and to see any possible relationships between them:

The histogram of price2014 does show that there are some houses that are extremely high priced as compared with the rest of the houses in the area:

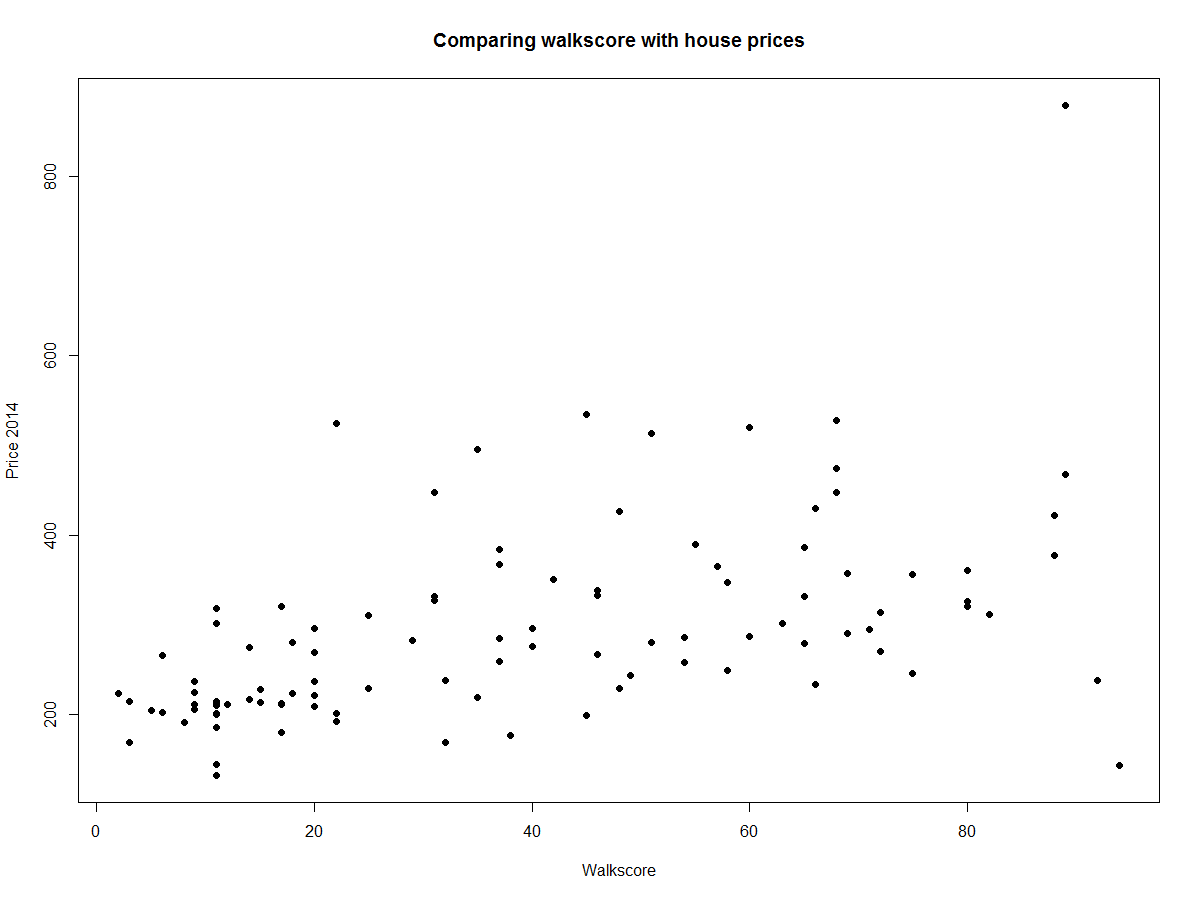


The histogram of number of rooms shows that a majority of the houses have 4, 5, or 6 rooms, while a very small number of houses have a large number of rooms (11, 12, 14):

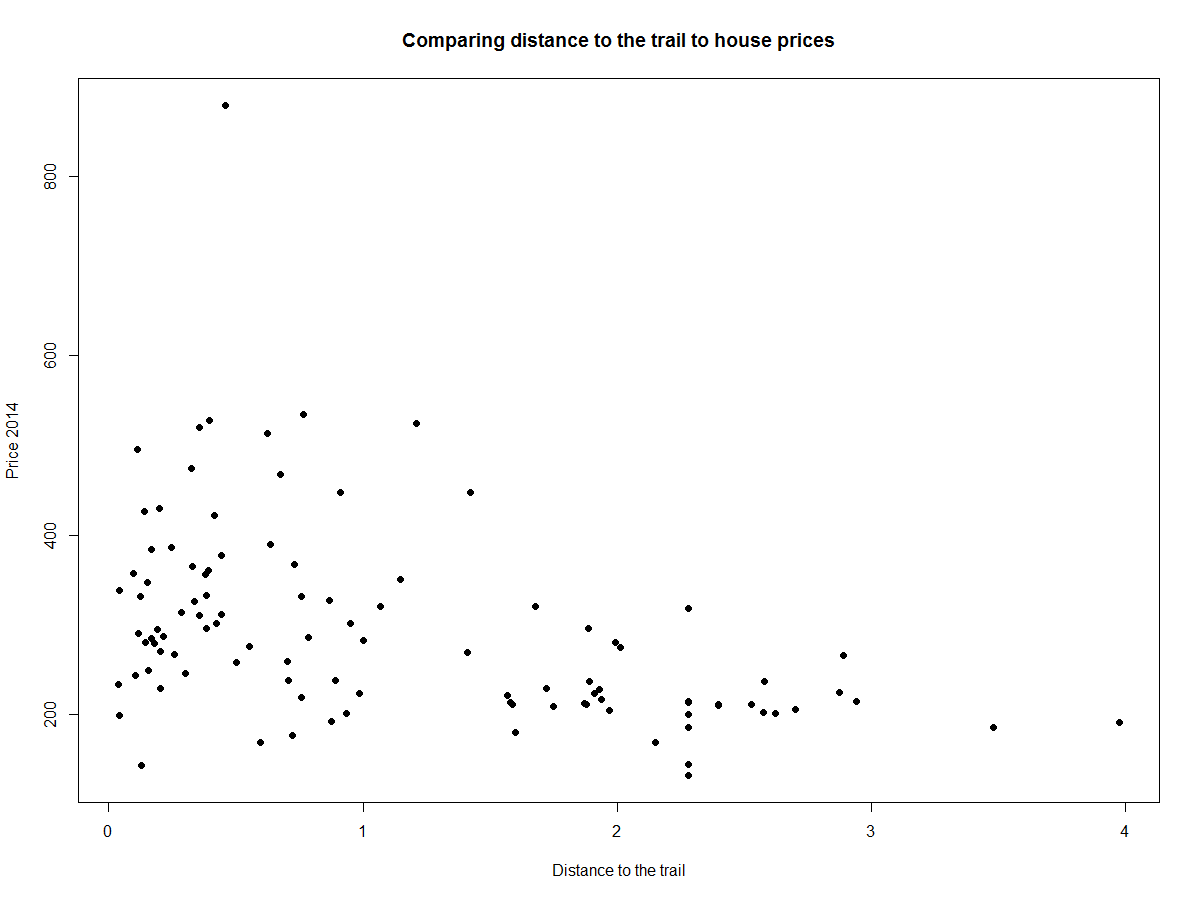
Plotting walkscore and bikescore shows that Walkscore and bikescore are highly positively correlated:



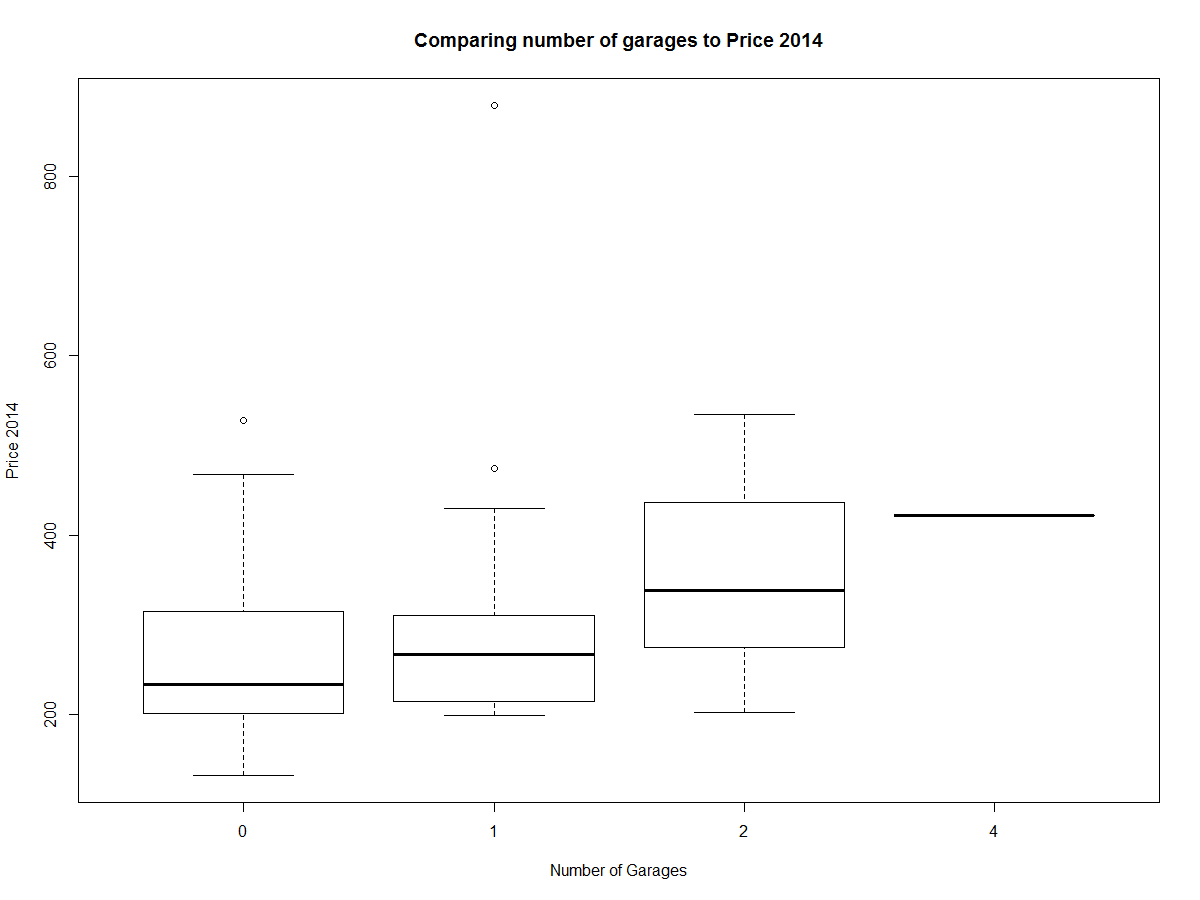
Plotting walkscore and house prices shows that Walkscore is positively correlated with price:



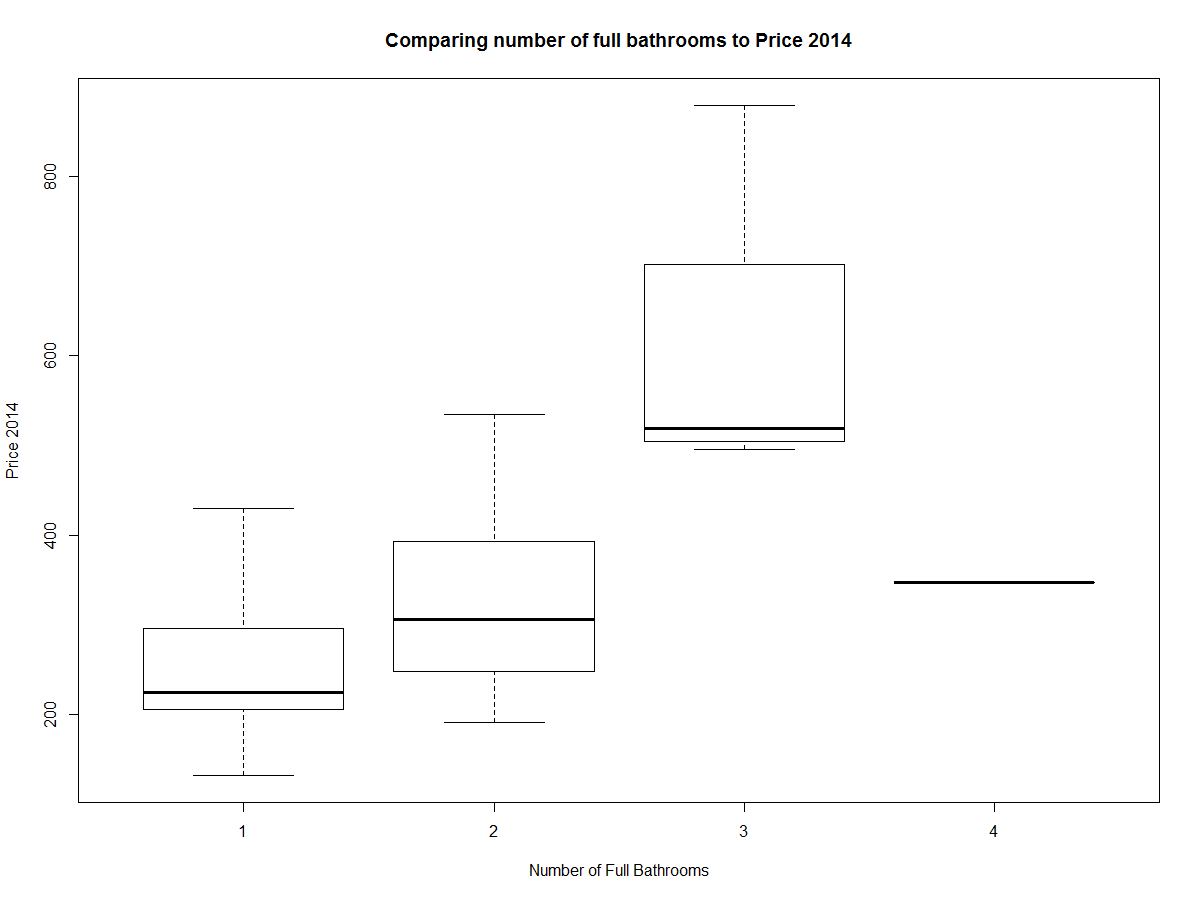
Plotting distance to the trail against house prices shows that there is some correlation which shows that as the distance to the trail gets close to 0, house prices do tend to be higher:



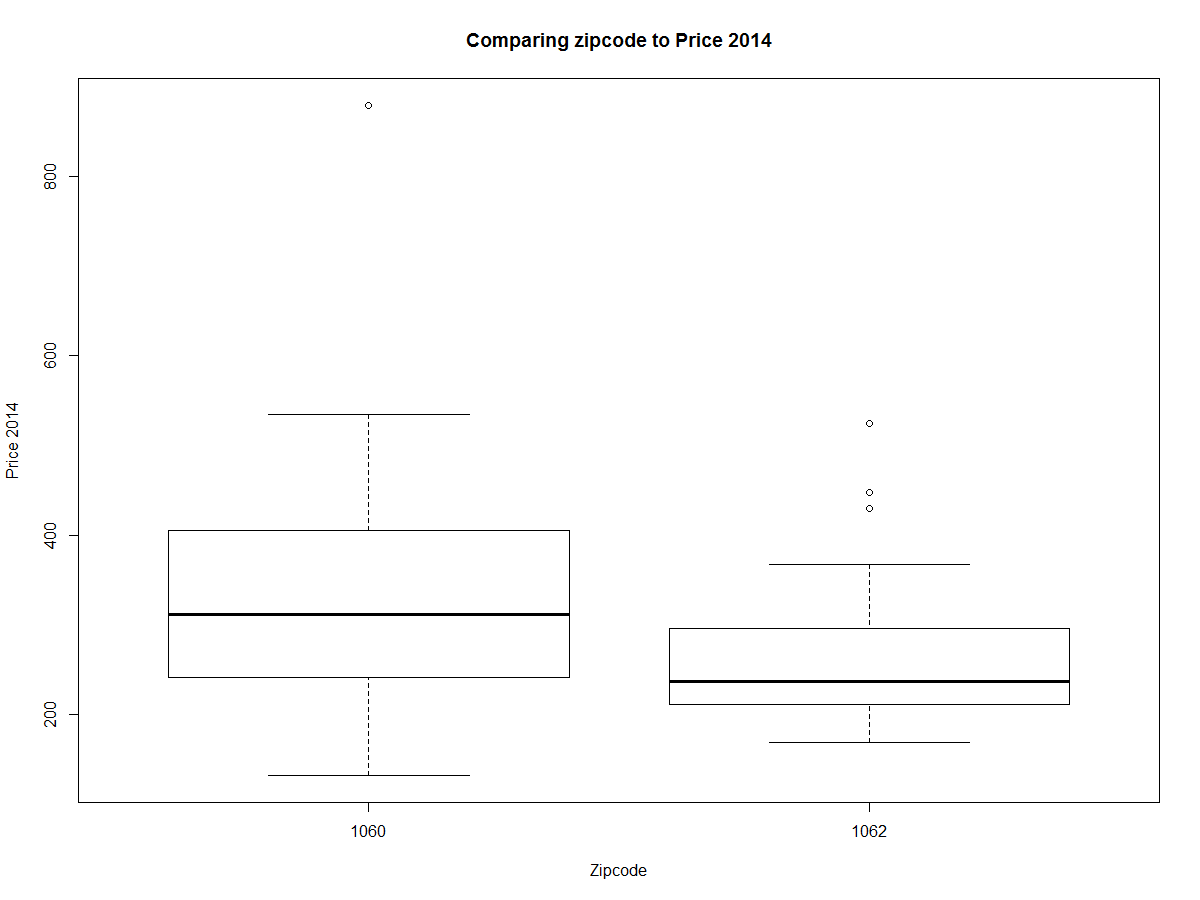
Plotting number of garages with house prices shows that the house price does vary according to the number of garages and it is positively correlated:



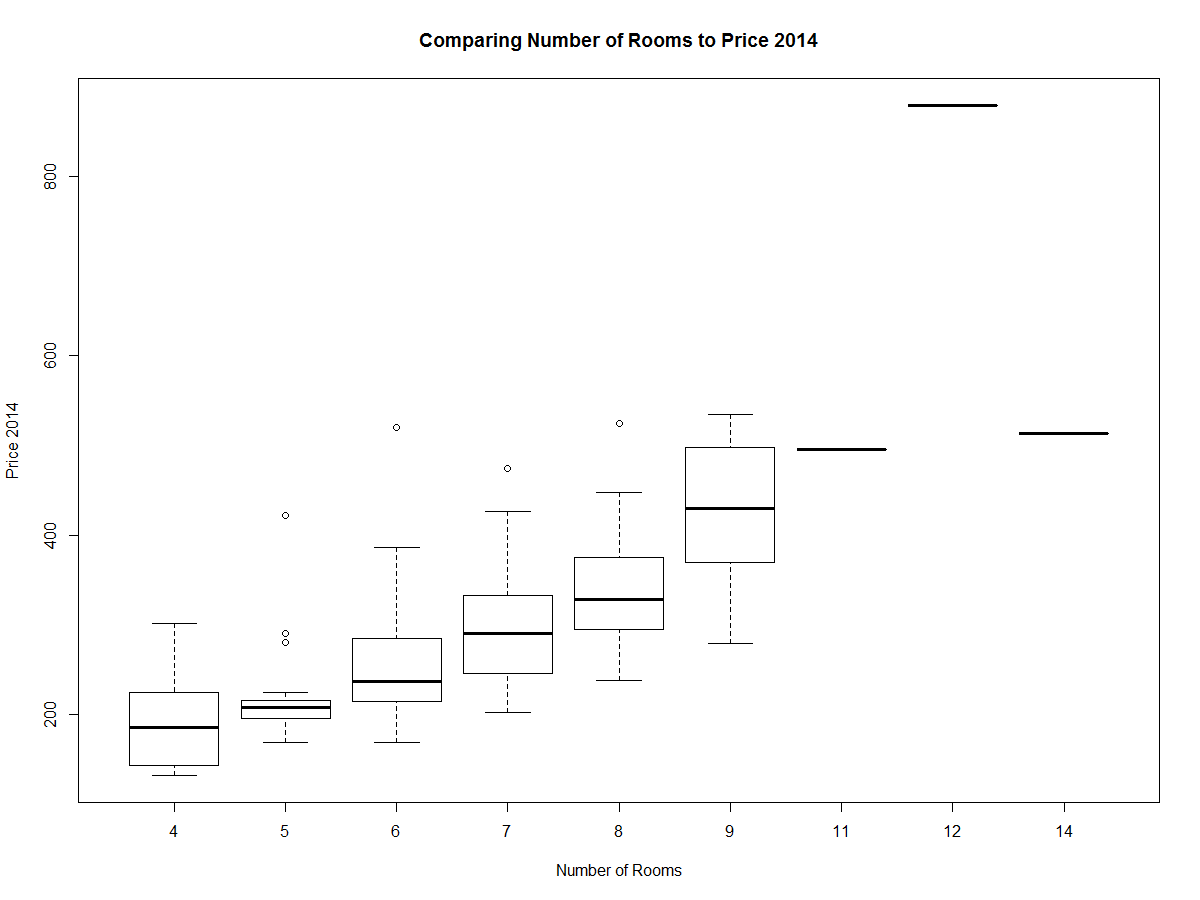
Plotting the number of full bathrooms with house prices shows that, in general, house prices are higher with a higher number of full bathrooms:



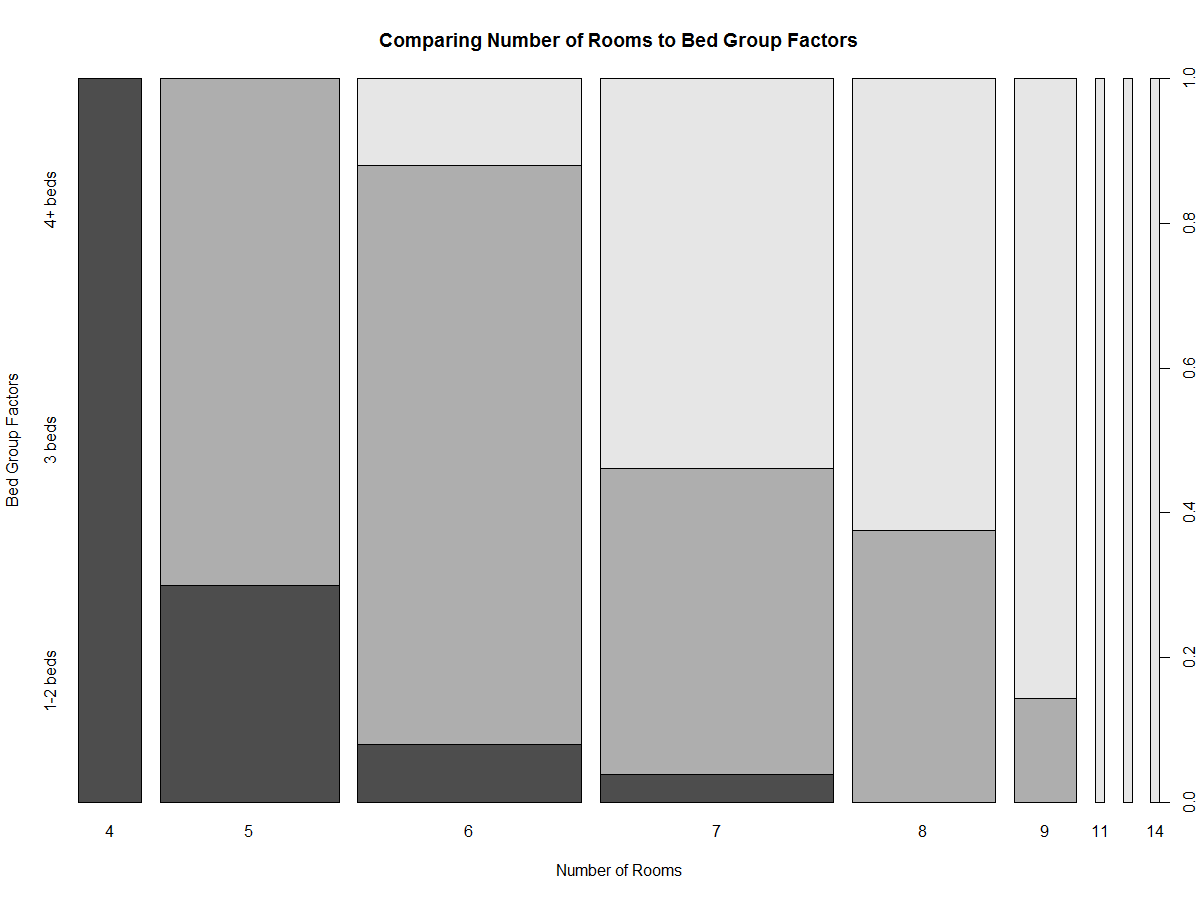
Plotting zipcode with house prices shows that there are some price differences between the two zipcodes and that zipcode 1060 does on average have higher house prices than 1062 zipcode:



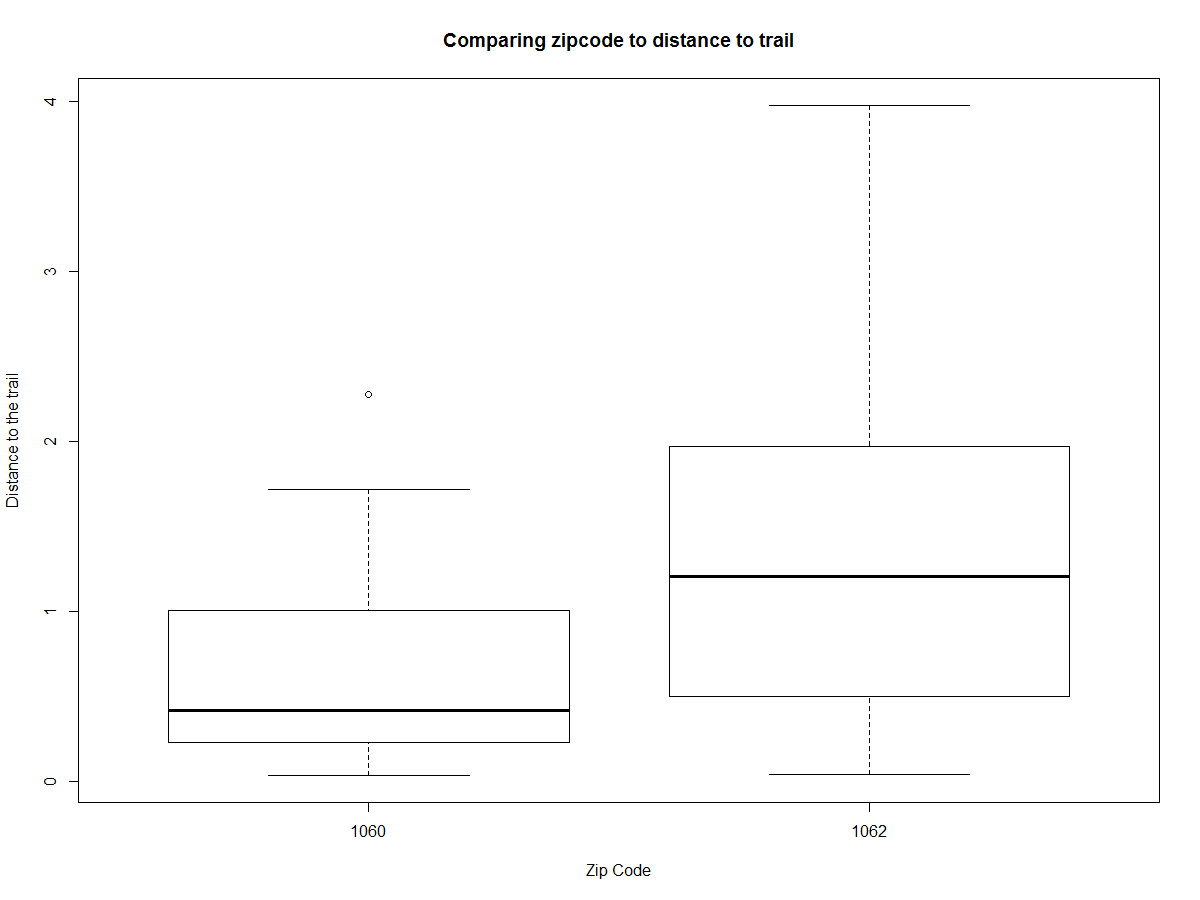
Plotting number of rooms to house prices shows that, in general, house price does indeed increase as the number of rooms increases:



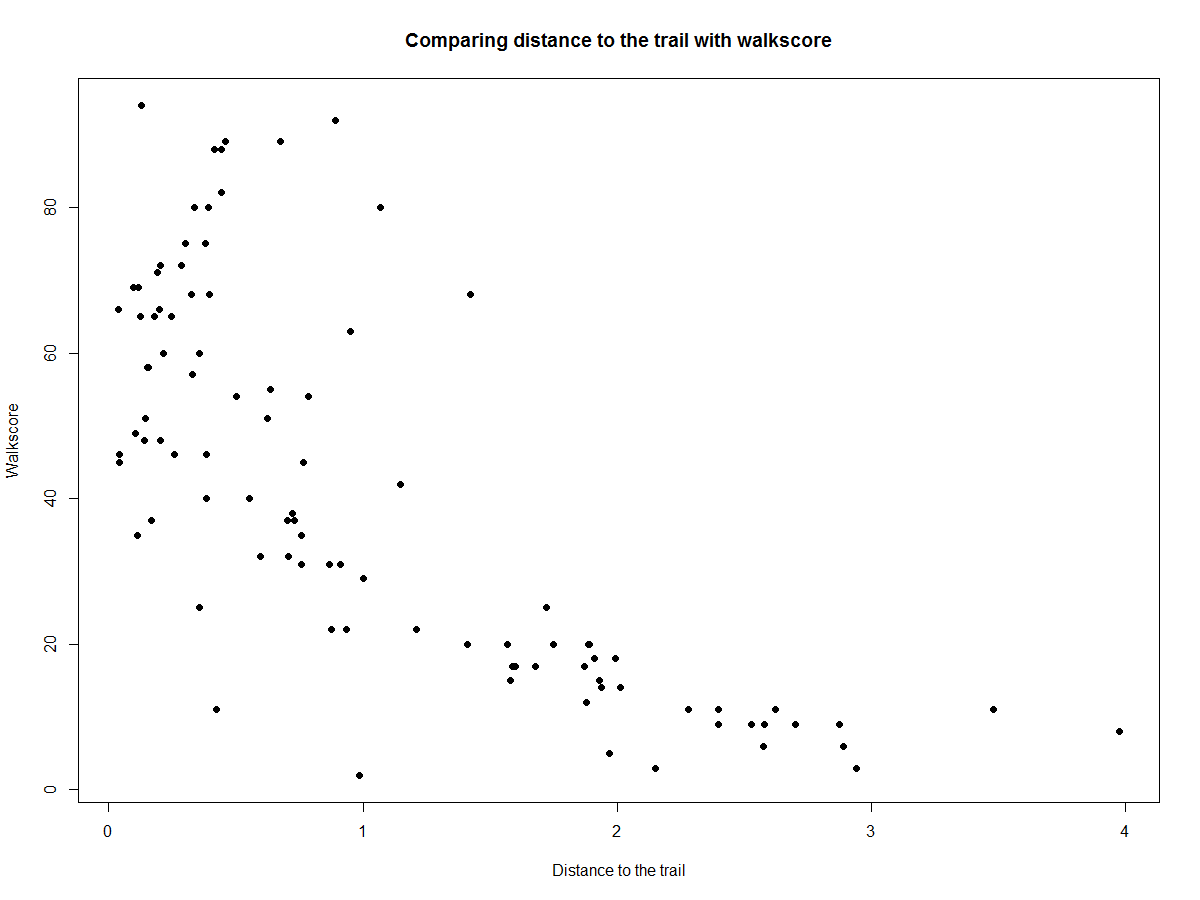
Plotting number of rooms to bed group factors shows that as the number of rooms increases, the number of bedrooms also increases as the bedgroupF is changing as the number of rooms is increasing:



Plotting zipcode against distance to the trail shows that on average, zipcode 1060 does have a smaller distance to the trail as compared with zipcode 1062:



Plotting distance to the trail and walkscore shows that as distance to the trail decreases, walkscore increases so there does seem to be an inverse relationship here:



1. **Model Fitting**

I performed a series of Normal linear regression models where I treated Price2014 as the response variable and treated acre, bedF, bikescore, distance, garageF, fullbathF, halfbathF, roomF, squarefeet, walkscore, and zip as the covariates.

In formal mathematical terms, if *yἰ* is the value of Price2014 for respondent *ἰ*, then the model was fitted as:

*yἰ* = *β0* +

with

Where *xij* is the value of the jth covariate for respondent *ἰ*.

1. **Statistical Analysis**
2. **Simple linear regression between distance and house prices in 2014**

I started with a simple linear regression to examine the relationship between distance and the house prices in 2014 to check if they are significantly related.

> mdistance <- lm(price2014 ~ distance , data = houseprices)

> Anova(mdistance)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

distance 255635 1 25.793 1.72e-06 \*\*\*

Residuals 1010912 102

---

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

As shown by the Anova table, distance to the trail appears to be highly significant to house prices. Using a 5% confidence level, distance to the trail is significantly related to house prices as its P-value of 1.72e-06 < 0.05.

1. **Multiple linear regression with all covariates and distance**

Now that it is known that distance is highly correlated with house prices, I have created a multiple linear regression analysis to include all other covariates apart from *housenum* including distance as explanatory variables for the response variable of *price2014*:

> m1 = lm(price2014 ~ acre + bedF + bikescore + distance + garageF + fullbathF + halfbathF + roomF + squarefeet + walkscore + zipF, data = houseprices)

> Anova(m1)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

acre 1452 1 0.5673 0.453527

bedF 4360 2 0.8519 0.430445

bikescore 4459 1 1.7425 0.190584

distance 4252 1 1.6616 0.201109

garageF 16611 3 2.1639 0.098737 .

fullbathF 36231 3 4.7198 0.004391 \*\*

halfbathF 4681 1 1.8292 0.180033

roomF 115745 8 5.6542 1.088e-05 \*\*\*

squarefeet 18650 1 7.2884 0.008466 \*\*

walkscore 5904 1 2.3072 0.132714

zipF 17790 1 6.9525 0.010053 \*

Residuals 204706 80

---

> summary(m1)

Call:

lm(formula = price2014 ~ acre + bedF + bikescore + distance +

garageF + fullbathF + halfbathF + roomF + squarefeet + walkscore +

zipF, data = houseprices)

Residuals:

Min 1Q Median 3Q Max

-130.62 -24.13 0.00 20.69 115.00

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 178.9657 56.9232 3.144 0.002339 \*\*

acre -40.8413 54.2221 -0.753 0.453527

bedF3 beds 5.2230 20.6191 0.253 0.800680

bedF4+ beds -14.4781 24.7382 -0.585 0.560025

bikescore -1.0861 0.8228 -1.320 0.190584

distance -14.2450 11.0511 -1.289 0.201109

garageF1 12.4936 14.0725 0.888 0.377309

garageF2 42.3717 16.8585 2.513 0.013966 \*

garageF4 44.7955 63.7296 0.703 0.484160

fullbathF2 41.5424 13.0178 3.191 0.002025 \*\*

fullbathF3 149.9404 63.7956 2.350 0.021221 \*

fullbathF4 98.5210 54.8879 1.795 0.076439 .

halfbathF1 21.6174 15.9835 1.352 0.180033

roomF5 33.5050 28.8070 1.163 0.248253

roomF6 37.6781 32.1829 1.171 0.245176

roomF7 37.5860 35.6625 1.054 0.295085

roomF8 49.6205 40.9388 1.212 0.229055

roomF9 101.5504 49.1486 2.066 0.042047 \*

roomF11 -10.9981 81.2648 -0.135 0.892685

roomF12 354.9726 93.0250 3.816 0.000266 \*\*\*

roomF14 -100.6652 93.4837 -1.077 0.284797

squarefeet 72.1372 26.7204 2.700 0.008466 \*\*

walkscore 0.9237 0.6081 1.519 0.132714

zipF1062 -42.0149 15.9343 -2.637 0.010053 \*

---

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

Residual standard error: 50.58 on 80 degrees of freedom

Multiple R-squared: 0.8384, Adjusted R-squared: 0.7919

F-statistic: 18.04 on 23 and 80 DF, p-value: < 2.2e-16

Running the summary command on m1 shows that the Multiple R-squared of the multiple linear regression is 0.8384. This means that this model with all of the explanatory variables included in it explains 83.84% of the variance. This model also has a residual standard error of 50.58 on 80 degrees of freedom. This model also has a residual value of 204706. This is our base line understanding as the model hasn’t undergone backwards elimination to eliminate the least significant terms contributing to the model.

1. **Backwards elimination from the main effects model**

Now that we have the model with all of the explanatory variables inside the model, I will carry out backwards elimination using Anova F-tests to determine the least significant term to remove at each stage. I will be using a critical value of p=0.05 where the least significant value is removed at each stage until all the terms are found to be significant at the 5% level.

So the first term that I eliminate is the “acre” variable as it is the least significant, now the model becomes:

> m2 = lm(price2014 ~ bedF + bikescore + distance + garageF + fullbathF + halfbathF + roomF + squarefeet + walkscore + zipF, data = houseprices)

> Anova(m2)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

bedF 3570 2 0.7013 0.498915

bikescore 4557 1 1.7904 0.184623

distance 4433 1 1.7418 0.190629

garageF 16116 3 2.1106 0.105280

fullbathF 36979 3 4.8430 0.003768 \*\*

halfbathF 4371 1 1.7172 0.193753

roomF 114741 8 5.6352 1.1e-05 \*\*\*

squarefeet 18008 1 7.0752 0.009419 \*\*

walkscore 6400 1 2.5146 0.116692

zipF 29082 1 11.4263 0.001116 \*\*

Residuals 206158 81

---

Now the “bedF” variable is the least significant so it is thus eliminated, now the model is:

> m3 = lm(price2014 ~ bikescore + distance + garageF + fullbathF + halfbathF + roomF + squarefeet + walkscore + zipF, data = houseprices)

> Anova(m3)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

bikescore 4186 1 1.6567 0.201631

distance 4059 1 1.6063 0.208551

garageF 17699 3 2.3349 0.079789 .

fullbathF 36861 3 4.8625 0.003640 \*\*

halfbathF 6766 1 2.6778 0.105544

roomF 117512 8 5.8132 6.922e-06 \*\*\*

squarefeet 15041 1 5.9526 0.016825 \*

walkscore 6504 1 2.5740 0.112431

zipF 28650 1 11.3384 0.001153 \*\*

Residuals 209728 83

---

Now the “distance” variable is removed, resulting in:

> m4 = lm(price2014 ~ bikescore + garageF + fullbathF + halfbathF + roomF + squarefeet + walkscore + zipF, data = houseprices)

> Anova(m4)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

bikescore 1354 1 0.5319 0.467836

garageF 15565 3 2.0386 0.114653

fullbathF 35200 3 4.6102 0.004908 \*\*

halfbathF 7936 1 3.1181 0.081059 .

roomF 113700 8 5.5843 1.129e-05 \*\*\*

squarefeet 21559 1 8.4710 0.004618 \*\*

walkscore 6548 1 2.5728 0.112472

zipF 25408 1 9.9832 0.002196 \*\*

Residuals 213787 84

---

Now the “bikescore” variable is removed, leading to:

> m5 = lm(price2014 ~ garageF + fullbathF + halfbathF + roomF + squarefeet + walkscore + zipF, data = houseprices)

> Anova(m5)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

garageF 14307 3 1.8842 0.138372

fullbathF 34003 3 4.4781 0.005732 \*\*

halfbathF 7662 1 3.0272 0.085498 .

roomF 121767 8 6.0136 4.131e-06 \*\*\*

squarefeet 21111 1 8.3408 0.004916 \*\*

walkscore 11632 1 4.5958 0.034907 \*

zipF 24426 1 9.6503 0.002573 \*\*

Residuals 215141 85

---

Finally, I drop out “garageF” out of the model:

> m6 = lm(price2014 ~ fullbathF + halfbathF + roomF + squarefeet + walkscore + zipF, data = houseprices)

> Anova(m6)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

fullbathF 31755 3 4.0596 0.0094358 \*\*

halfbathF 11889 1 4.5599 0.0355082 \*

roomF 118504 8 5.6812 8.07e-06 \*\*\*

squarefeet 40312 1 15.4607 0.0001676 \*\*\*

walkscore 13452 1 5.1590 0.0255649 \*

zipF 19924 1 7.6416 0.0069470 \*\*

Residuals 229448 88

---

This is the final model as all of the variables are now significant and are under the 5% level. Now I will run a summary output to see how the model well the model fits.

> summary(m6)

Call:

lm(formula = price2014 ~ fullbathF + halfbathF + roomF + squarefeet +

walkscore + zipF, data = houseprices)

Residuals:

Min 1Q Median 3Q Max

-124.25 -24.21 0.00 24.28 154.81

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 90.8951 28.0002 3.246 0.001655 \*\*

fullbathF2 37.8287 12.9256 2.927 0.004360 \*\*

fullbathF3 158.5714 63.3312 2.504 0.014130 \*

fullbathF4 66.6912 53.3686 1.250 0.214747

halfbathF1 32.3801 15.1636 2.135 0.035508 \*

roomF5 41.0764 24.8895 1.650 0.102437

roomF6 31.8066 25.1112 1.267 0.208630

roomF7 38.3790 28.6235 1.341 0.183429

roomF8 28.5184 32.0302 0.890 0.375701

roomF9 70.7727 38.9051 1.819 0.072295 .

roomF11 -25.0274 77.8199 -0.322 0.748513

roomF12 298.6364 81.3263 3.672 0.000412 \*\*\*

roomF14 -152.5335 85.1504 -1.791 0.076676 .

squarefeet 87.7491 22.3166 3.932 0.000168 \*\*\*

walkscore 0.5913 0.2603 2.271 0.025565 \*

zipF1062 -35.2703 12.7590 -2.764 0.006947 \*\*

---

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

Residual standard error: 51.06 on 88 degrees of freedom

Multiple R-squared: 0.8188, Adjusted R-squared: 0.788

F-statistic: 26.52 on 15 and 88 DF, p-value: < 2.2e-16

In this final model, it shows that distance is not significantly correlated with Price 2014. The R-squared of this model is 0.8188 which means that the model can explain around 81.88% of the data reasonably well while the adjusted R-squared is 0.788 which is comparable to the m1 model with all of the explanatory variables included. We do obtain a residual standard error of 51.06 with a total of 88 degrees of freedom with the residuals increasing to 229448.

To understand the coefficients of this model and how it affects price:

Intercept: The price of standard/initial/base house with 1 Full Bathroom, 0 Half Baths, and 4 rooms with a zipcode of 1060 is approximately 90.8951 (which correlates to $90,895.10)

fullbathF2: With all other variables held constant, if there are two full bathrooms in the house (as opposed to 1 full bathroom since that is the initial) then the house price increases 37.8287 ($37828.8)

fullbathF3: Similarly, 3 Full Bathrooms in the house is associated with an increase in the house price by 158.5714 ($158,571.4)

fullbathF4: 4 Full Bathrooms in the house is associated with an increase in the house price by 66.6912 ($66,691.2)

halfbathF1: 1 Half Bathrooms in the house is associated with an increase in the house price by 32.3801 ($32,380.1)

roomF5: Having 5 rooms (as compared to 4 in the standard) in the house is associated with an increase in the house price by 41.0764 ($41,076.4)

roomF6: Similarly, Having 6 rooms in the house is associated with an increase in the house price by 31.8066 ($31,806.6)

roomF7: Having 7 rooms in the house is associated with an increase in the house price by 38.3790 ($38,379)

roomF8: Having 8 rooms in the house is associated with an increase in the house price by 28.5184 ($28,518.4)

roomF9: Having 9 rooms in the house is associated with an increase in the house price by 70.7727 ($70,772.7)

roomF11: Having 11 rooms in the house is associated with a decrease in the house price by -25.0274 (-$25,027.4)

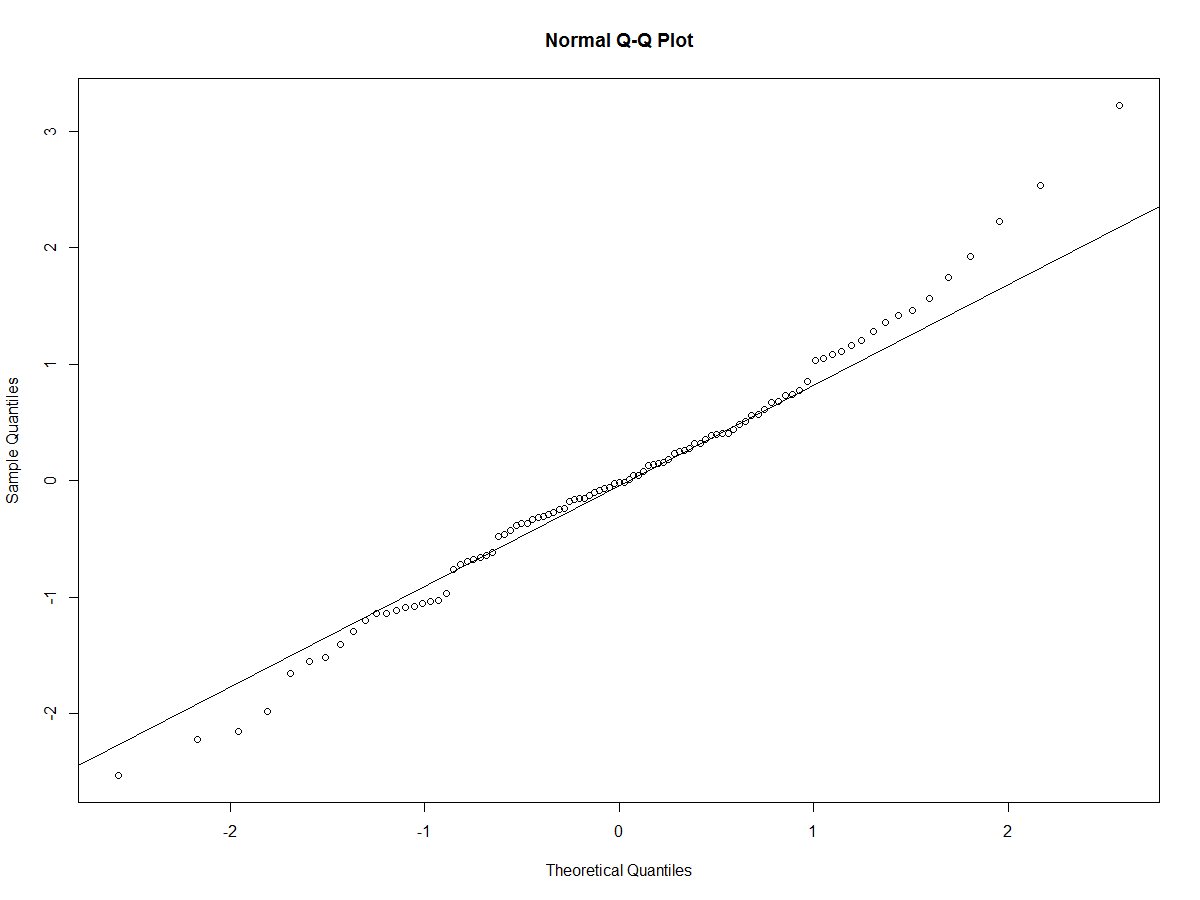
roomF12: Having 12 rooms in the house is associated with an increase in the house price by 298.6364 ($298,636.4)

roomF14: Having 14 rooms in the house is associated with a decrease in the house price by -152.5225 (-$152,522.5)

squarefeet: An incremental increase of 1 unit of squarefeet (which corresponds to 1000 actual squarefeet because squarefeet is represented in thousands in the dataset) is associated with a price increase 87.7491 ($87,749.1)

zipF1062: Living in zipcode 1062 (as opposed to 1060) is associated with a decrease in houseprice of -35.2703 (-$35,270.3)

Running a QQ-plot on the residuals for the final m6 model:



As shown by the plot, the residuals do not appear to be normal as they are deviating at both tails significantly. This means that the model can be further improved. Considering the introductory statistical analysis that I ran before, there do seem to be some variables that are highly correlated with each other such as Walkscore and bikescore, and zipF and distance, and bedF and roomF so I have adjusted the model and added the two-way interactions between these variables in addition to the previous covariates:

> m60 <- lm(price2014 ~ squarefeet + acre + (bikescore + walkscore)^2 + (zipF + distance)^2 + (bedF + roomF)^2 + garageF + fullbathF + halfbathF, data = houseprices)

> Anova(m60)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

squarefeet 12995 1 5.1958 0.025566 \*

acre 7876 1 3.1492 0.080135 .

bikescore 3786 1 1.5136 0.222534

walkscore 5687 1 2.2740 0.135877

zipF 10146 1 4.0569 0.047676 \*

distance 558 1 0.2231 0.638121

bedF 5213 2 1.0423 0.357845

roomF 124685 8 6.2318 4.009e-06 \*\*\*

garageF 22694 3 3.0247 0.034942 \*

fullbathF 31630 3 4.2157 0.008321 \*\*

halfbathF 2829 1 1.1312 0.291027

bikescore:walkscore 8842 1 3.5355 0.064058 .

zipF:distance 3858 1 1.5427 0.218186

bedF:roomF 12580 5 1.0060 0.420432

Residuals 182571 73

---

Now I performed backwards elimination similar to previously where I will drop the least significant explanatory variable in the model until all of the variables are significant at the 5% level. Due to this, I first drop distance from the model which results in:

> m61 <- lm(price2014 ~ squarefeet + acre + (bikescore + walkscore)^2 + zipF + zipF:distance + (bedF + roomF)^2 + garageF + fullbathF + halfbathF, data = houseprices)

> Anova(m61)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

squarefeet 12995 1 5.1958 0.025566 \*

acre 7876 1 3.1492 0.080135 .

bikescore 3786 1 1.5136 0.222534

walkscore 5687 1 2.2740 0.135877

zipF 12547 1 5.0168 0.028147 \*

bedF 5213 2 1.0423 0.357845

roomF 124685 8 6.2318 4.009e-06 \*\*\*

garageF 22694 3 3.0247 0.034942 \*

fullbathF 31630 3 4.2157 0.008321 \*\*

halfbathF 2829 1 1.1312 0.291027

bikescore:walkscore 8842 1 3.5355 0.064058 .

zipF:distance 4416 2 0.8829 0.417950

bedF:roomF 12580 5 1.0060 0.420432

Residuals 182571 73

---

Similarly, now I drop bedF:roomF resulting in:

> m62 <- lm(price2014 ~ squarefeet + acre + (bikescore + walkscore)^2 + zipF + zipF:distance + bedF + roomF + garageF + fullbathF + halfbathF, data = houseprices)

> Anova(m62)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

squarefeet 17676 1 7.0649 0.009533 \*\*

acre 4395 1 1.7566 0.188916

bikescore 5555 1 2.2202 0.140246

walkscore 6362 1 2.5429 0.114836

zipF 15323 1 6.1245 0.015502 \*

bedF 5213 2 1.0419 0.357655

roomF 124685 8 6.2294 3.286e-06 \*\*\*

garageF 16842 3 2.2439 0.089754 .

fullbathF 32038 3 4.2684 0.007623 \*\*

halfbathF 4513 1 1.8038 0.183150

bikescore:walkscore 8079 1 3.2293 0.076204 .

zipF:distance 3932 2 0.7857 0.459365

Residuals 195150 78

---

Now I drop off zipF:distance resulting in:

> m63 <- lm(price2014 ~ squarefeet + acre + (bikescore + walkscore)^2 + zipF + bedF + roomF + garageF + fullbathF + halfbathF, data = houseprices)

> Anova(m63)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

squarefeet 19485 1 7.8299 0.006436 \*\*

acre 2518 1 1.0119 0.317480

bikescore 1464 1 0.5883 0.445324

walkscore 5865 1 2.3570 0.128668

zipF 15323 1 6.1575 0.015182 \*

bedF 4464 2 0.8970 0.411863

roomF 121155 8 6.0857 4.173e-06 \*\*\*

garageF 16333 3 2.1878 0.095897 .

fullbathF 32560 3 4.3614 0.006752 \*\*

halfbathF 5295 1 2.1278 0.148565

bikescore:walkscore 9876 1 3.9686 0.049767 \*

Residuals 199082 80

---

Next, I dropped of bikescore as it is the least significant resulting in:

> m64 <- lm(price2014 ~ squarefeet + acre + walkscore + bikescore:walkscore + zipF + bedF + roomF + garageF + fullbathF + halfbathF, data = houseprices)

> Anova(m64)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

squarefeet 20410 1 8.2930 0.005090 \*\*

acre 2443 1 0.9925 0.322087

walkscore 8950 1 3.6366 0.060069 .

zipF 17540 1 7.1271 0.009172 \*\*

bedF 4519 2 0.9182 0.403361

roomF 127495 8 6.4756 1.712e-06 \*\*\*

garageF 17014 3 2.3044 0.083036 .

fullbathF 34146 3 4.6248 0.004894 \*\*

halfbathF 5453 1 2.2156 0.140506

walkscore:bikescore 11075 1 4.5002 0.036948 \*

Residuals 199347 81

---

Dropped of bedF resulting in:

> m65 <- lm(price2014 ~ squarefeet + acre + walkscore + bikescore:walkscore + zipF + roomF + garageF + fullbathF + halfbathF, data = houseprices)

> Anova(m65)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

squarefeet 16907 1 6.8833 0.010352 \*

acre 1345 1 0.5476 0.461381

walkscore 10370 1 4.2218 0.043053 \*

zipF 19304 1 7.8592 0.006294 \*\*

roomF 128095 8 6.5189 1.441e-06 \*\*\*

garageF 18844 3 2.5573 0.060682 .

fullbathF 34303 3 4.6553 0.004670 \*\*

halfbathF 7894 1 3.2137 0.076668 .

walkscore:bikescore 10441 1 4.2510 0.042357 \*

Residuals 203866 83

---

Now dropped off acre :

> m66 <- lm(price2014 ~ squarefeet + walkscore + bikescore:walkscore + zipF + roomF + garageF + fullbathF + halfbathF, data = houseprices)

> Anova(m66)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

squarefeet 16954 1 6.9399 0.0100344 \*

walkscore 11632 1 4.7615 0.0318965 \*

zipF 30799 1 12.6072 0.0006327 \*\*\*

roomF 126988 8 6.4976 1.454e-06 \*\*\*

garageF 18089 3 2.4681 0.0676338 .

fullbathF 34965 3 4.7708 0.0040435 \*\*

halfbathF 7310 1 2.9924 0.0873310 .

walkscore:bikescore 9930 1 4.0645 0.0469896 \*

Residuals 205211 84

---

Since I am testing at the 5% level, I dropped off halfbathF since it is the least significant:

> m67 <- lm(price2014 ~ squarefeet + walkscore + bikescore:walkscore + zipF + roomF + garageF + fullbathF, data = houseprices)

> Anova(m67)

Anova Table (Type II tests)

Response: price2014

Sum Sq Df F value Pr(>F)

squarefeet 27751 1 11.0991 0.001280 \*\*

walkscore 11777 1 4.7105 0.032769 \*

zipF 28428 1 11.3701 0.001125 \*\*

roomF 127813 8 6.3900 1.778e-06 \*\*\*

garageF 22351 3 2.9799 0.035922 \*

fullbathF 27862 3 3.7146 0.014544 \*

walkscore:bikescore 10281 1 4.1121 0.045709 \*

Residuals 212522 85

---

This leads the m67 model to be the final resulting model where all of the variables are significant at the 5% level.

> summary(m67)

Call:

lm(formula = price2014 ~ squarefeet + walkscore + bikescore:walkscore +

zipF + roomF + garageF + fullbathF, data = houseprices)

Residuals:

Min 1Q Median 3Q Max

-144.95 -22.92 0.00 18.81 113.17

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 98.62374 29.31185 3.365 0.001152 \*\*

squarefeet 77.20899 23.17525 3.332 0.001280 \*\*

walkscore 2.75304 1.10815 2.484 0.014940 \*

zipF1062 -44.25686 13.12496 -3.372 0.001125 \*\*

roomF5 23.07565 24.48585 0.942 0.348655

roomF6 30.29179 25.48050 1.189 0.237820

roomF7 13.02298 29.33074 0.444 0.658168

roomF8 25.40077 32.81661 0.774 0.441065

roomF9 81.42821 40.92261 1.990 0.049827 \*

roomF11 -34.42878 77.03827 -0.447 0.656080

roomF12 335.64715 84.62201 3.966 0.000152 \*\*\*

roomF14 -122.55994 86.43112 -1.418 0.159844

garageF1 15.10219 13.20178 1.144 0.255855

garageF2 46.84605 15.83111 2.959 0.003996 \*\*

garageF4 57.83288 61.05051 0.947 0.346174

fullbathF2 31.48316 11.98670 2.627 0.010231 \*

fullbathF3 126.27903 59.99203 2.105 0.038250 \*

fullbathF4 84.46306 53.16482 1.589 0.115841

walkscore:bikescore -0.02354 0.01161 -2.028 0.045709 \*

---

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

Residual standard error: 50 on 85 degrees of freedom

Multiple R-squared: 0.8322, Adjusted R-squared: 0.7967

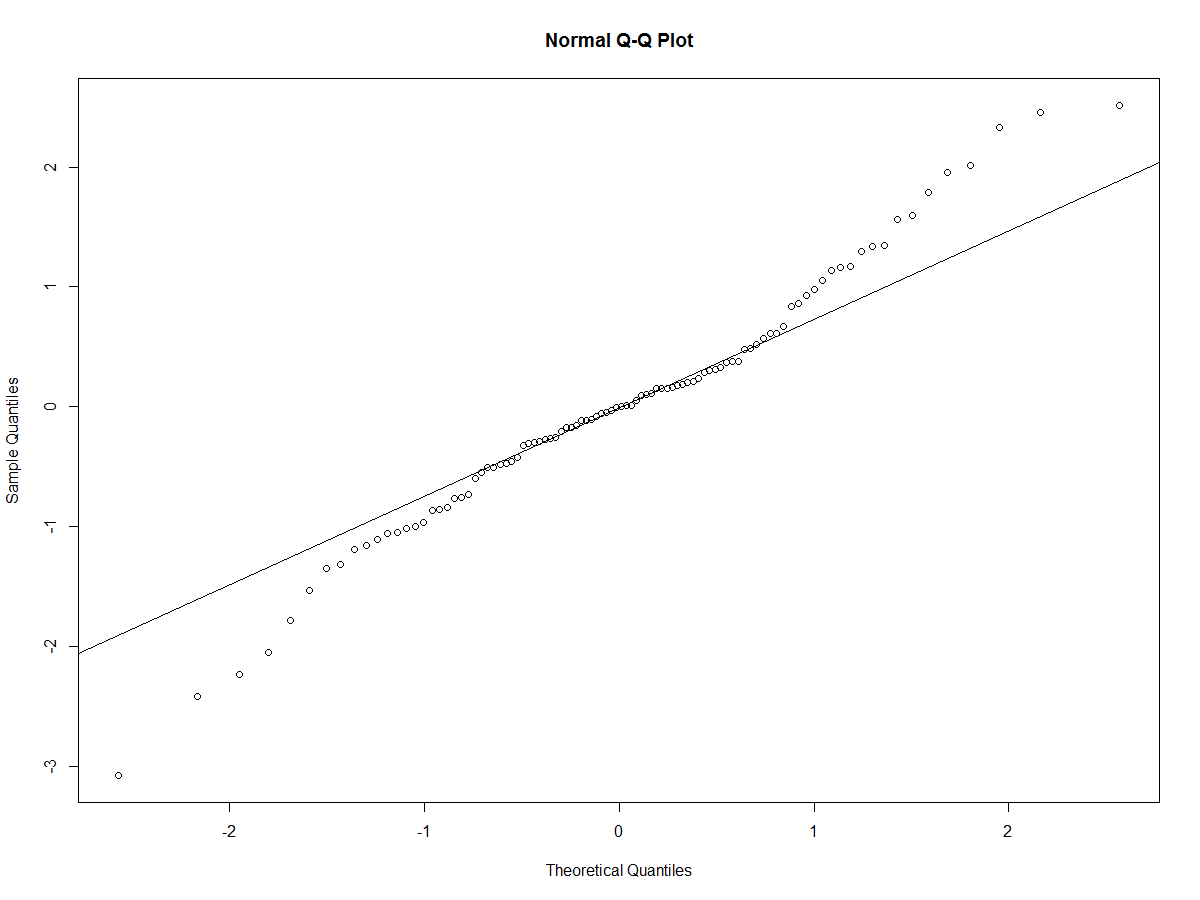
F-statistic: 23.42 on 18 and 85 DF, p-value: < 2.2e-16

Running a summary command on the final m67 model shows that the multiple R-squared is 0.8322 which means that 83.22% of the variance can be explained using this model. This is pretty good when comparing it to the original m1 model which had all of the variables involved and still only marginally improved the variance and shows that this is a better model then the final m6 model as well. This has a residual standard error of 50 with 85 degrees of freedom.

The coefficients are explained in the same manner as they were explained with the final m6 model, for example, having 5 rooms is associated with a 23.07565 increase in price2014 which means an increase of $23,075.65 as compared with the 4 rooms in the standard house.

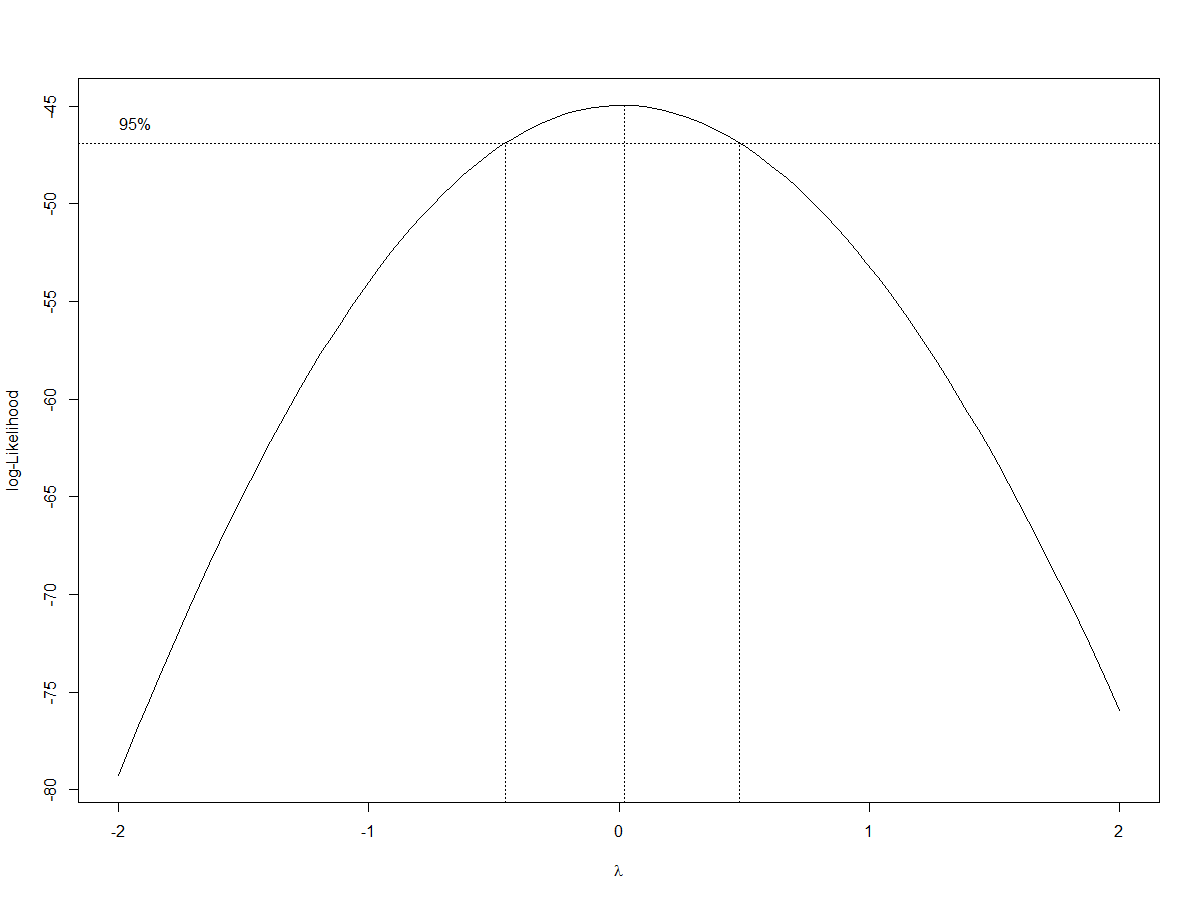
1. **Checking standardized residuals, model diagnostics, and model updating**

Running a QQ plot for this model:



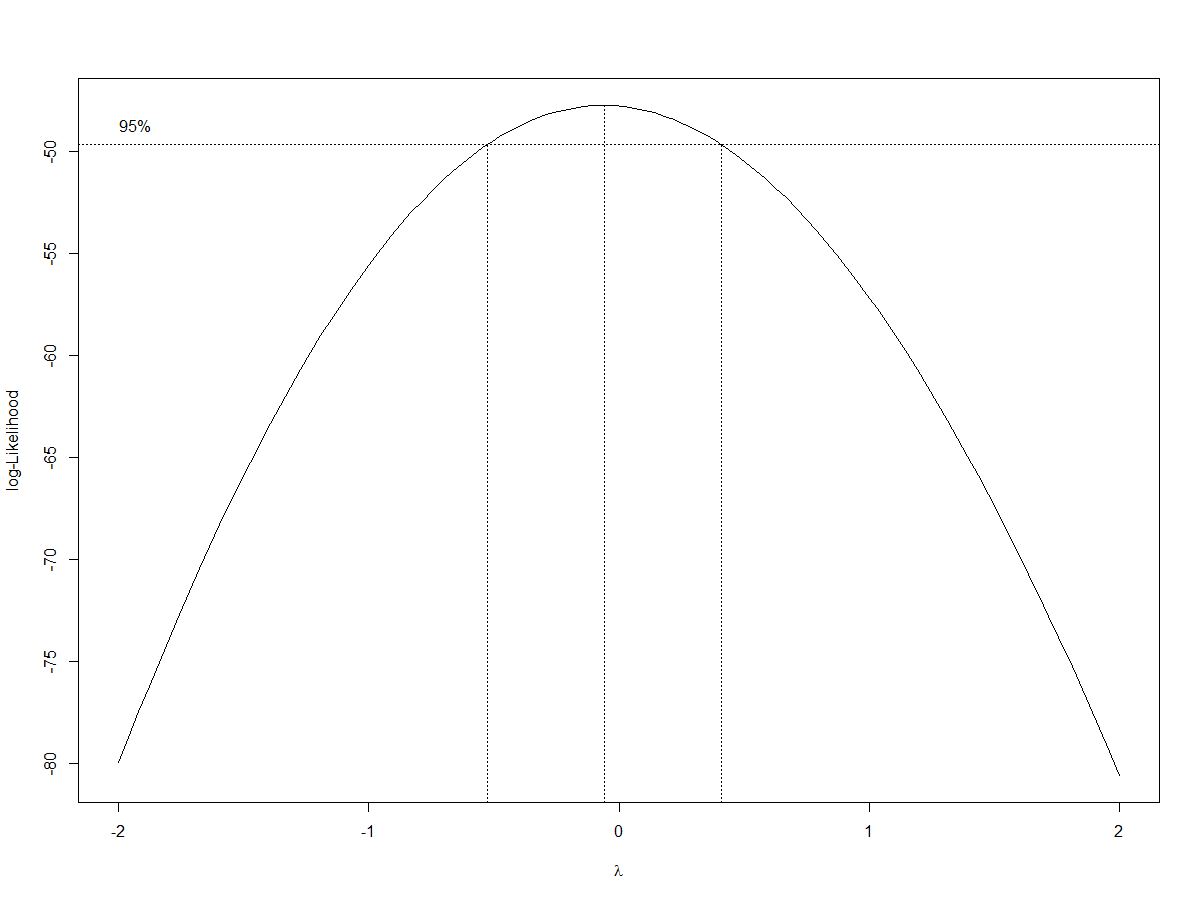
Upon inspection of the QQ-plot, the residuals still do not seem normal which means that the model can be further improved. Comparing the m67 residuals to the m6 residuals, the m67 residuals seem to be better predicted in the middle but have more deviation towards both ends.

Doing a boxcox analysis for the m67 model results in:



This boxcox analysis shows that it is more appropriate to do a log transformation.

Doing a boxcox analysis for the m6 model:



The boxcox analysis for the m6 model also suggests to perform a log transformation on the m6 model as well.

Performing a log transformation on the m67 model leads to:

> m68 <- lm(log(price2014) ~ squarefeet + walkscore + bikescore:walkscore + zipF + roomF + garageF + fullbathF, data = houseprices)

> Anova(m68)

Anova Table (Type II tests)

Response: log(price2014)

Sum Sq Df F value Pr(>F)

squarefeet 0.40205 1 14.3952 0.0002768 \*\*\*

walkscore 0.18419 1 6.5947 0.0119764 \*

zipF 0.21485 1 7.6927 0.0068139 \*\*

roomF 0.50980 8 2.2816 0.0290465 \*

garageF 0.19881 3 2.3727 0.0759743 .

fullbathF 0.20386 3 2.4330 0.0705326 .

walkscore:bikescore 0.08997 1 3.2215 0.0762324 .

Residuals 2.37401 85

---

After the log transformation, it seems that some of the values have turned non significant so I will continue backwards elimination until all of them are significant at the 5% level.

So first I drop off Walkscore:bikescore as it is the least significant:

> m69 <- lm(log(price2014) ~ squarefeet + walkscore + zipF + roomF + garageF + fullbathF, data = houseprices)

> Anova(m69)

Anova Table (Type II tests)

Response: log(price2014)

Sum Sq Df F value Pr(>F)

squarefeet 0.47299 1 16.5085 0.0001066 \*\*\*

walkscore 0.18419 1 6.4286 0.0130378 \*

zipF 0.16258 1 5.6745 0.0194185 \*

roomF 0.49294 8 2.1506 0.0393715 \*

garageF 0.16597 3 1.9309 0.1306029

fullbathF 0.19512 3 2.2700 0.0861197 .

Residuals 2.46399 86

---

Next, I drop off garageF as it is now the least significant:

> Anova(m70)

Anova Table (Type II tests)

Response: log(price2014)

Sum Sq Df F value Pr(>F)

squarefeet 0.86703 1 29.3412 5.113e-07 \*\*\*

walkscore 0.19759 1 6.6866 0.01134 \*

zipF 0.11117 1 3.7623 0.05559 .

roomF 0.47820 8 2.0228 0.05254 .

fullbathF 0.14428 3 1.6275 0.18865

Residuals 2.62995 89

---

Now I drop off fullbathF as it is now the least significant:

> Anova(m71)

Anova Table (Type II tests)

Response: log(price2014)

Sum Sq Df F value Pr(>F)

squarefeet 1.51643 1 50.2883 2.661e-10 \*\*\*

walkscore 0.18564 1 6.1561 0.01491 \*

zipF 0.09683 1 3.2112 0.07642 .

roomF 0.50375 8 2.0882 0.04473 \*

Residuals 2.77423 92

---

zipF is dropped:

> m72 <- lm(log(price2014) ~ squarefeet + walkscore + roomF, data = houseprices)

> Anova(m72)

Anova Table (Type II tests)

Response: log(price2014)

Sum Sq Df F value Pr(>F)

squarefeet 1.62380 1 52.5982 1.196e-10 \*\*\*

walkscore 0.34699 1 11.2396 0.001159 \*\*

roomF 0.43671 8 1.7682 0.093270 .

Residuals 2.87106 93

---

RoomF is dropped, resulting in the final model m73:

> m73 <- lm(log(price2014) ~ squarefeet + walkscore, data = houseprices)

> Anova(m73)

Anova Table (Type II tests)

Response: log(price2014)

Sum Sq Df F value Pr(>F)

squarefeet 4.7844 1 146.088 < 2.2e-16 \*\*\*

walkscore 0.5285 1 16.136 0.0001135 \*\*\*

Residuals 3.3078 101

---

This is now the final model with all of the variables significant at the 5% level.

> summary(m73)

Call:

lm(formula = log(price2014) ~ squarefeet + walkscore, data = houseprices)

Residuals:

Min 1Q Median 3Q Max

-0.47425 -0.10088 0.01419 0.10910 0.43352

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.835378 0.053351 90.634 < 2e-16 \*\*\*

squarefeet 0.427116 0.035338 12.087 < 2e-16 \*\*\*

walkscore 0.003021 0.000752 4.017 0.000114 \*\*\*

---

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

Residual standard error: 0.181 on 101 degrees of freedom

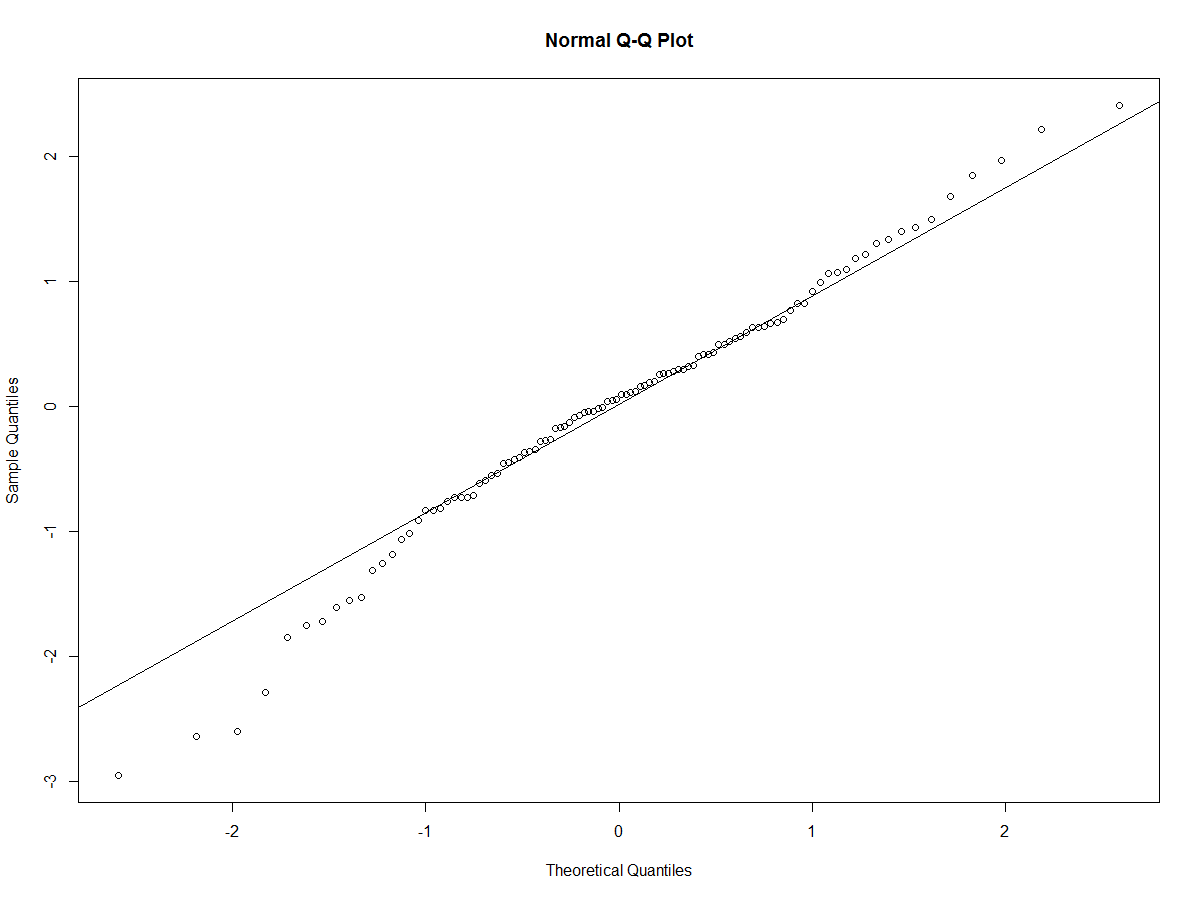
Multiple R-squared: 0.712, Adjusted R-squared: 0.7063

F-statistic: 124.9 on 2 and 101 DF, p-value: < 2.2e-16

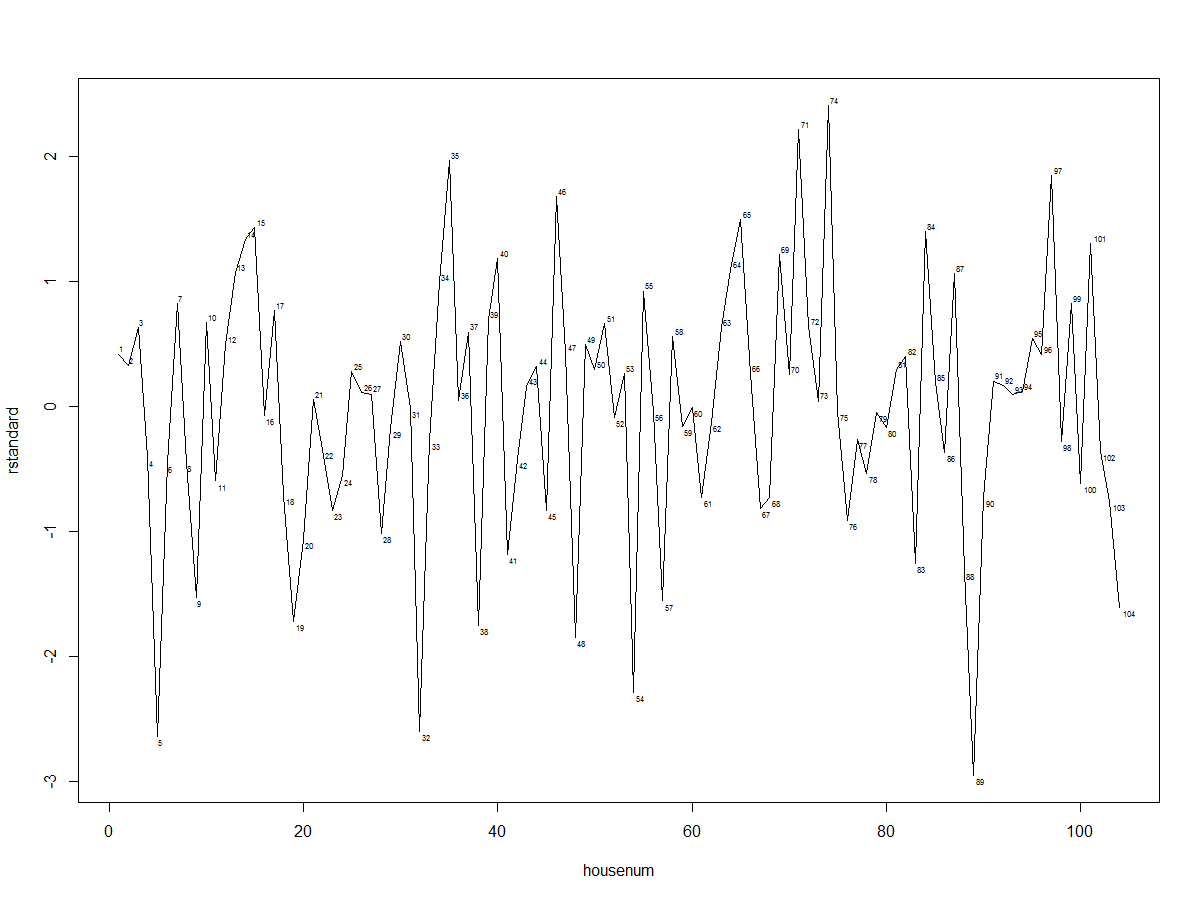
This model has a multiple R-squared of 0.712 which means it can explain only 71.2% of the variance, however, it does have higher number of degrees of freedom of 101 df as lots of variables were dropped due to non-significance.

Interpreting the coefficients of this model means that median house price will be 53.283% higher for an increase in Squarefeet (because exp(0.427116) is 1.53283) and correspondingly the median house price will be 0.3026% higher for an increase in Walkscore (because exp(0.003021) is 1.003026).

Finally, checking the normality of residuals for this model:



The residuals do seem to have improved their normality as the line is now a better approximation, however, it could still use some improving as mainly the left end has bigger deviance.



I created a plot to identify any outliers with large standardized residuals for model m73 plotting against house number. A large residual value consists of the residuals being above 2 or less than -2. Using this, there are 6 outliers (house number’s 5, 32, 54, 71, 74, and 89). So even though there are 6 apparent outliers, noticing that there are 104 data points, and that 1 in 20 residuals will have a residual value of greater than 2 or less than -2, I have ignored the outliers. However, the model may change if the outliers are removed.

1. **Conclusion**

When plotting a graph, and using a simple linear regression using only distance as the explanatory variable for house prices in Northampton, Massachusetts, distance to the trail was found to be significantly correlated with house prices. However, involving other explanatory variables and performing multiple linear regressions several times, distance did not end up being a significant factor in determining house prices.

When doing a simple multiple linear regression, including all of the covariates and performing backwards elimination, it resulted in model m6.

In model m6, fullbathF is positively associated with price, halfbathF is also positively associated with price, roomF (on average) is positively associated with a higher price, Squarefeet is also positively associated with a higher price, Walkscore is also positively associated with a higher price, but zipF is associated with a lower house price (zipcode 1062 has a lower price than zipcode 1060). All other variables, including distance, were found to be not significant to house price. RoomF, fullbathF, and Squarefeet seem to have the biggest impact on house price in terms of magnitude, however, that is subjective as since the factors are included, the magnitude is hard to decipher. However, when checking the normality of the residuals of the model m6, they did seem to be non-normal so the model was updated to include interaction terms that were distinguished using the exploratory analysis.

Model m67 is the updated version of model m6 which included all covariates, but also included interaction terms between bikescore:walkscore, zipF:distance, and bedF:roomF. Similarly, backwards elimination was performed with the least significant variable dropped from the model until all of the variables were found to be significant at the 5% level. The final resulting model, m67, shows squarefeet, walkscore, zipF, roomF, garageF, fullbathF, and walkscore:bikescore as significant to the model. All of these variables are positively correlated with an increase in house price, however, zipF 1062 is associated with a lower house price, and interestingly, walkscore:bikescore is correlated with a lower house price (albeit to a small degree). This is interesting as Walkscore itself is positively associated with an increase in house price, so maybe houses with already high walkscores, but an increased bikescore results in a lower price because of some unknown relationship between walkers and bikers. Maybe neighborhoods designed for walkers specifically are priced at a premium, but neighborhoods designed for both walkers and bikers are discounted a bit because walkers and bikers don’t get along.

However, plotting the residuals for model m67 only resulted in a marginally improved normality. Performing a boxcox analysis on the model showed that there should be a log transformation.

After performing the log transformation and similarly performing backwards elimination as usual, resulted in model m73 with only two covariates: Squarefeet and Walkscore. Both of these are positively correlated with an increase in price, however, Squarefeet has a bigger magnitude of effect on house price, but that may be related to the fact that Squarefeet is measured in thousands, so an increase of 1 unit of Squarefeet in the model actually means an increase of 1000 actual square footage of the house. Model m73 did provide a better approximate normality of residuals, even though not perfect, but significantly improved from the first two final models. However, since there are only two covariates in the m73 model, it did only account for 71.2% of the variance while the previous models (m6 and m67) did account for greater than 80% of the variance, albeit with lower degree of freedoms.

I did want to point out that for model m67, I did perform multiple complex models before finally settling on those interaction terms such as:

m11 <- lm(price2014 ~ squarefeet\*acre + roomF\*fullbathF\*halfbathF\*bedF\*garageF + bikescore\*walkscore\*zipF\*distance, data = houseprices)

m31 <- lm(price2014 ~ (squarefeet+acre)^2 + (roomF+fullbathF+halfbathF+bedF+garageF)^2 + (bikescore+walkscore+zipF+distance)^2, data = houseprices)

However, I did not receive drastically different results in the final resulting models after backwards elimination and they were too lengthy and complex to be included in the report. The residuals and r-squared did not change drastically from the results obtained from model m67.

Furthermore, some further considerations to be considered with this model are that the factorizations of number of rooms, number of full baths, half baths, and garage spaces could be argued. So, it may result in a different model if treating those as continuous. RoomF especially was a difficult decision as there were only one occurrences where the house had either 11 rooms, 13 rooms, and 14 rooms, so the factorization grouping could have been allocated better to result in a more even distribution for roomF although it is out of the scope of this paper. Also, there are some outliers in the data, however, since there were only 6 out of the dataset, I decided to ignore them so excluding them might significantly impact the results of the model.

Finally, regarding the economic benefit of the rail trail in Northampton, Massachusetts, the distance to the rail trail does not seem directly related to house prices after including all other covariates. However, that does not mean that there is no economic benefit of the rail trail as the other covariates might be colluding the effect of the distance variable. This is evident because in all of the models, Walkscore was considered significant to house prices, and as shown in the data exploratory analysis, it shows that the smaller distance to the trail correlates to a higher Walkscore. So in that sense, the trail may have a significant impact on the house price, however, the effect is combined with Walkscore (and correspondingly bikescore as Walkscore and bikescore are positively correlated). All in all, even though distance is not significant at the model level for affecting house prices, there may still be economic benefit of the rail trail because other variables such as walkscore which have high correlation to distance are significant in affecting house prices.