BANA 7038 – DATA ANALYSIS METHODS

Final Project – Predicting House Prices in Cincinnati

Team Members

Pawar, Aishwarya (M13481763)

Marar, Rinisha (M13439960)

Nimbalkar, Sharang (M13262376)

Summary

We observed that the real estate market in Cincinnati is exposed to many fluctuations in prices because of the existing correlation between many variables, some of which cannot be controlled or might even be unknown. For our analysis we used sold price of the property as the response variable and identified 20 potential Covariates; Address, Zip code, Neighborhood, Sold date, Month Sold, Season, No. of bedrooms, No. of bathrooms, Sq. ft, Built Year, No. of stories, Lot size, Parking type, No. of Parking spaces, Market land value, Exterior, Basement, Fireplace, Floor and Roof. Considering this, we developed a regression model predicting the sold price of the property to identify the opportunities in the real estate.
The factors that affect the prices of houses sold in Cincinnati was studied and modelled using the dataset HousingDataset. This dataset was created manually, by looking up websites like www.trulia.com and www.zillow.com. The dataset was cleaned of all the missing values, and a final dataset of 202 records were used.

The Model consists of:

Sold Price – Response variable
Square feet – Covariate/Independent Variable
Neighbourhood – Covariate/Independent Variable

1.1 Data Exploration and Data Cleaning

Goal: The intension of data exploration and cleaning is to condense the raw data to a more usable form. The process of data cleaning is instrumental in revealing insights for the user. Perform exploratory data analysis for the manually collected data from www.trulia.com and www.zillow.com.

R code and Output:

```
> str(HousingData)
'data.frame':   202 obs. of  22 variables:
 $ Address          : Factor w/ 201 levels "1022 Portway Dr",..: 59 118 125 41 21 123 130 128 126 124 ...
 $ Sold_price       : int  162600 97000 235000 260000 110000 130000 114000 139500 107500 134900 ...
 $ Zipcode          : int  45239 45223 45239 45223 45223 45239 45239 45239 45239 45239 ...
 $ Month            : int  5 10 2 7 8 4 9 5 8 7 ...
 $ Num_of_bedrooms  : int  3 3 5 3 1 2 2 2 2 2 ...
 $ Num_of_bathrooms : num  1.5 1 3.5 2 1 2 2 2 2 2 ...
 $ Squared_feet     : int  1247 1348 2530 2468 806 1520 1153 1453 1280 1526 ...
 $ Year_built       : int  1963 1920 1967 1972 1880 2005 1991 1992 1990 2004 ...
 $ Stories          : int  1 1 2 1 1 1 2 2 2 1 ...
 $ Lot_size         : int  10890 17424 22215 47045 40075 1520 1568 1742 1873 1898 ...
 $ Parking_Spaces   : int  2 1 2 2 1 1 1 1 1 1 ...
 $ Market_Land_Value: int  130 72 93 105 136 86 99 96 84 88 ...
 $ Sold_date        : Factor w/ 125 levels "Apr 1, 2019",..: 74 101 28 43 26 4 115 75 24 32 ...
 $ Season           : Factor w/ 4 levels "Fall","Spring",..: 2 1 4 3 3 2 1 2 3 3 ...
 $ Parking          : Factor w/ 7 levels "Attached","Both off and on",..: 1 5 1 6 1 1 1 1 1 1 ...
 $ Exterior_Wall_Type: Factor w/ 13 levels "Aluminium","Brick",..: 2 6 4 2 2 4 2 2 2 2 ...
 $ Neighborhood     : Factor w/ 30 levels "Camp Washington",..: 14 14 14 14 17 14 14 14 14 14 ...
 $ Basement         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 1 2 2 2 1 ...
 $ Fireplace        : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 2 2 2 1 ...
 $ Floor            : Factor w/ 16 levels "","Carpet","Carpet, Hardwood",..: 15 2 1 2 1 1 12 12 2 2 ...
 $ Roof             : Factor w/ 14 levels "","Asphalt","Attic",..: 10 4 1 6 1 12 8 8 8 1 ...
 $ Link             : Factor w/ 201 levels "https://www.trulia.com/p/oh/cincinnati/1022-portway-dr-cincinnati-oh-4525
5--2047182048",..: 58 116 123 41 20 121 128 126 124 122 ...
```

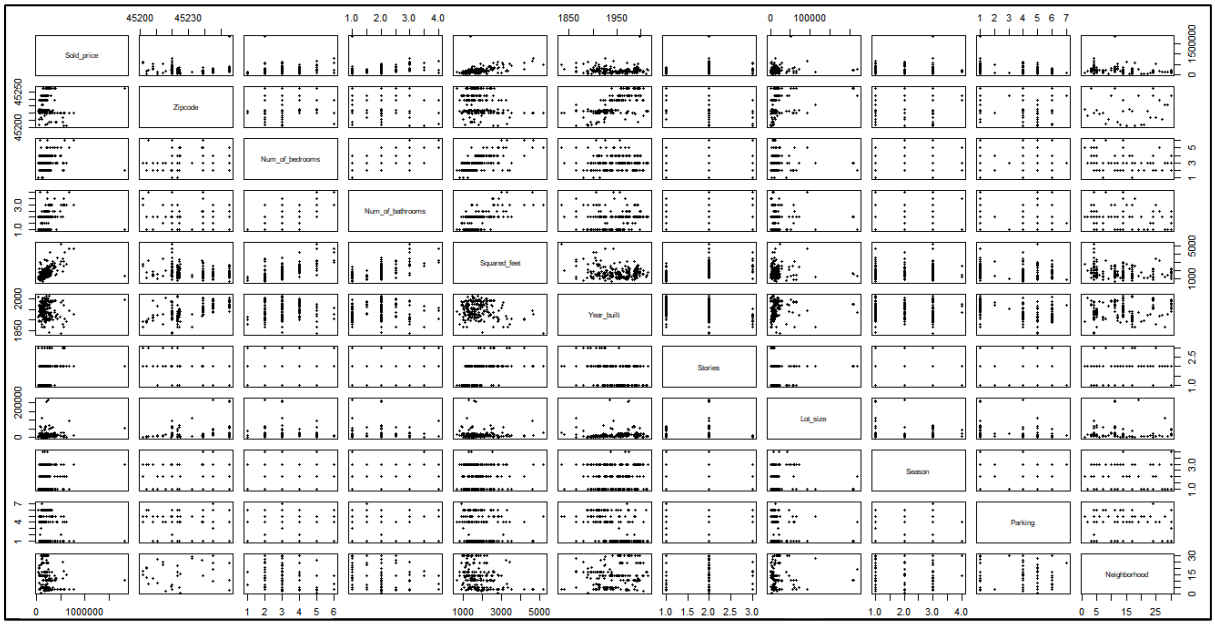
```
> sum(is.na(HousingData))
[1] 210
> sum(is.na(HousingData$Floor))
[1] 97
> sum(is.na(HousingData$Roof))
[1] 113
```

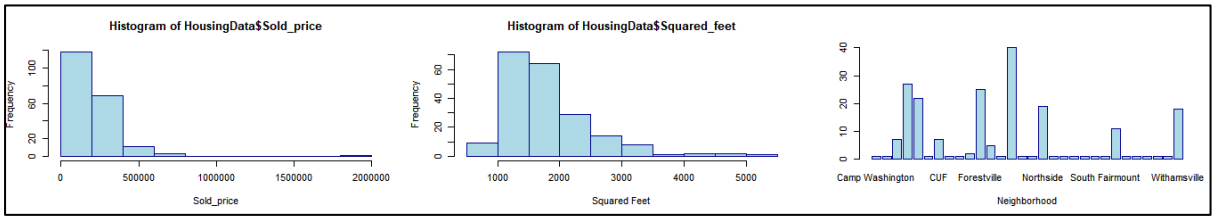
Observation & Conclusion: The dataset consisting of 202 records was read into R for further analysis. This data was checked for missing values and inconsistencies, if any. It was observed that the Floor and Roof covariates had 97 and 113 missing values respectively, which constituted for all the missing values in the dataset, and hence we decided to drop these covariates. We analysed the data type and significance of 22 covariates, and its influence in the prediction of the prices. We also decided to take Seasons as a covariate along with month, derived from the sold date.

1.2 Data Visualization

Goal: Visualize and infer from the dataset for correlation, skewedness and normality.

R code and Output:





Observation & Conclusion: The histograms obtained for sold_price and square feet are both right skewed, since the mean is greater than the median. Sold_price is right skewed whereas a squared foot is slightly skewed.Sold_price. We also see that square feet has a positive correlation with the response variable, and hence must be used as a covariate for the prediction of house prices.

1.3 Modelling

Goal: The objective of modelling is to find the right combination of independent variables that can be good predictors for dependant variables. We do so, by doing regression and comparing the analysis of variance for each of those parameters.

R code and Output:

```
model_bkwd <- lm((Sold_price) ~ (Squared_feet) + Neighborhood, data=HousingData)
summary(model_bkwd)
```

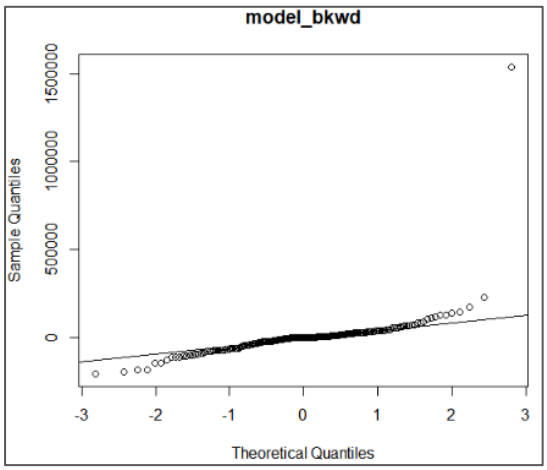
```
Call:
lm(formula = (Sold_price) ~ (Squared_feet) + Neighborhood, data = HousingData)

Residuals:
    Min      1Q  Median      3Q     Max
-207588  -36353       0   23704 1539446

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -171739.18  142582.90  -1.204  0.23007
Squared_feet                     96.74      16.13   5.997 1.16e-08 ***
NeighborhoodCarthage         145096.43  191657.95   0.757  0.45006
NeighborhoodCherry Grove     222987.99  145404.99   1.534  0.12698
NeighborhoodClifton          262828.03  137738.33   1.908  0.05804 .
NeighborhoodCollege Hill     160004.82  139103.09   1.150  0.25164
NeighborhoodCorryville       238517.99  192360.34   1.240  0.21669
NeighborhoodCUF              173637.23  146186.57   1.188  0.23657
NeighborhoodEast Price Hill   90500.17  191334.84   0.473  0.63682
NeighborhoodEast Westwood    139226.51  193385.28   0.720  0.47254
NeighborhoodEastgate         194074.31  165705.96   1.171  0.24315
NeighborhoodForestville      323600.18  138864.98   2.330  0.02096 *
NeighborhoodHyde Park        403218.89  149445.12   2.698  0.00767 **
NeighborhoodLower Price Hill 253487.41  193195.65   1.312  0.19125
NeighborhoodMt. Airy         152461.30  138152.72   1.104  0.27133
NeighborhoodMt. Auburn       521322.83  191190.95   2.727  0.00706 **
NeighborhoodNorth Fairmount    3980.23  191402.65   0.021  0.98343
NeighborhoodNorthside        207852.05  140213.47   1.482  0.14008
NeighborhoodOver-The-Rhine   493992.31  191222.95   2.583  0.01062 *
NeighborhoodPierce TWP       296495.10  192723.86   1.538  0.12579
NeighborhoodRiverside         60239.86  191828.73   0.314  0.75388
NeighborhoodSedamsville      119391.77  193122.17   0.618  0.53725
NeighborhoodSouth Cumminsville 122779.82 193061.21   0.636  0.52565
NeighborhoodSouth Fairmount  136335.33  192637.78   0.708  0.48008
NeighborhoodUnion TWP        206568.63  142524.11   1.449  0.14907
NeighborhoodWest End         381844.80  191458.22   1.994  0.04770 *
NeighborhoodWest Price HIll   59659.94  191309.19   0.312  0.75553
NeighborhoodWestwood         139182.16  192220.88   0.724  0.47001
NeighborhoodWinton Hills     176378.76  192927.05   0.914  0.36189
NeighborhoodWinton Place     172141.89  192458.18   0.894  0.37234
NeighborhoodWithamsville     188900.25  140453.79   1.345  0.18043
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 135200 on 171 degrees of freedom
Multiple R-squared:  0.4411,    Adjusted R-squared:  0.3431
F-statistic: 4.499 on 30 and 171 DF,  p-value: 1.345e-10
```

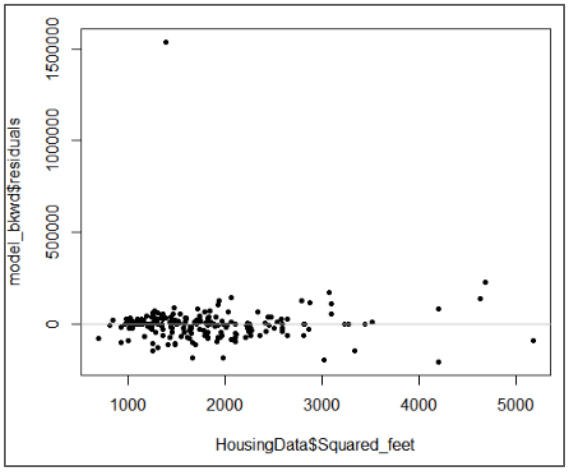
Observation & Conclusion: We used the Backward and Forward Selection methods to decide on our covariates. This process was iterative, by seeing the highest and lowest correlations with the response variable and accordingly adding or removing them. In this process, we obtained two covariates – Squared feet and neighbourhood. It was observed that these two covariates were the result of both forward as well as backward selection. The model with these variables is significant with a p-value less than the threshold

(0.05). The adjusted R-squared value is observed to be only 0.3431 which means only 34 % of the variability in the Sold price can be explained by the variation in the covariates.

1.4 Model Adequacy checking and Model Validation

Goal: Check the LINE assumptions and also check the goodness of fit of the model
R code and Output:

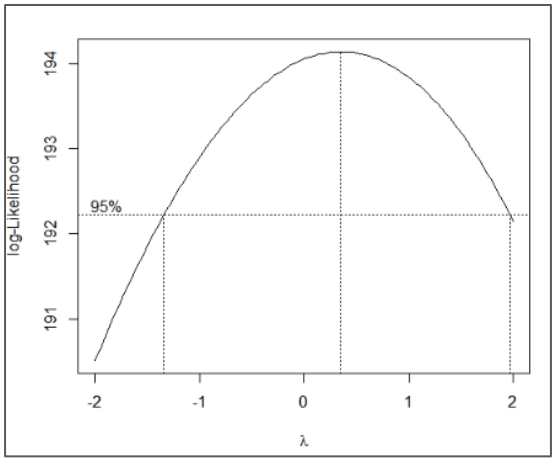


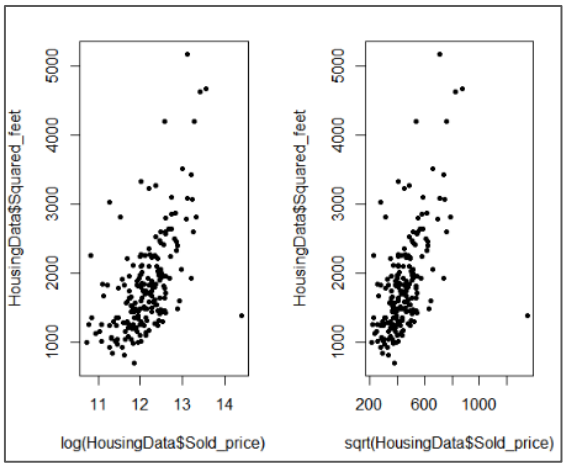


Observation & Conclusion: The Q-Q plot of residuals generally checks the normality assumptions and checks out for potential outliers. It is observed here, that the Q-Q plot is not along the 45 degree line, and appears to be heavy a tailed distribution. Thus, it defies the first assumption of normality. The scatter plot of residual against fitted values, checks for linearity and equal variance assumptions.

1.5 Transformation

Goal: The aim over here is to transform the dataset as the previous model violates the LINE assumptions.
R code and Output:





Observation & Conclusion: Since the model faces non-normality and/or unequal variance issue we decided on transforming the response y values only. By applying the BoxCox transformation we observed the value of lambda to be equal to zero which suggests us that the transformation to be applied on y is log transformation. In order to be sure with this step, we also transformed the y response value using sqrt() but the results obtained from log transformation was more accurate and satisfactory and hence we decide to go ahead with log transformation on the response variable y (Sold price).

1.6 Multicollinearity

Goal: To check for multicollinearity or near linear dependence amongst the regressors.

R code and Output:

```
> vif(model1)
                 GVIF Df GVIF^(1/(2*Df))
Squared_feet 1.528603  1        1.236367
Parking      3.324992  6        1.105306
Lot_size     1.180505  1        1.086510
Neighborhood 4.604177  9        1.088533
>
```

Observation & Conclusion: Multicollinearity is the instance in a multiple linear regression wherein one predictor variable can be linearly predicted from the others with a certain degree of accuracy. In our case we do not observe any values of variation inflation factors that are between 5 to 10 or greater than 10 and hence we conclude that there is no multicollinearity situation observed in our multiple linear regression.

1.7 Finalizing the model

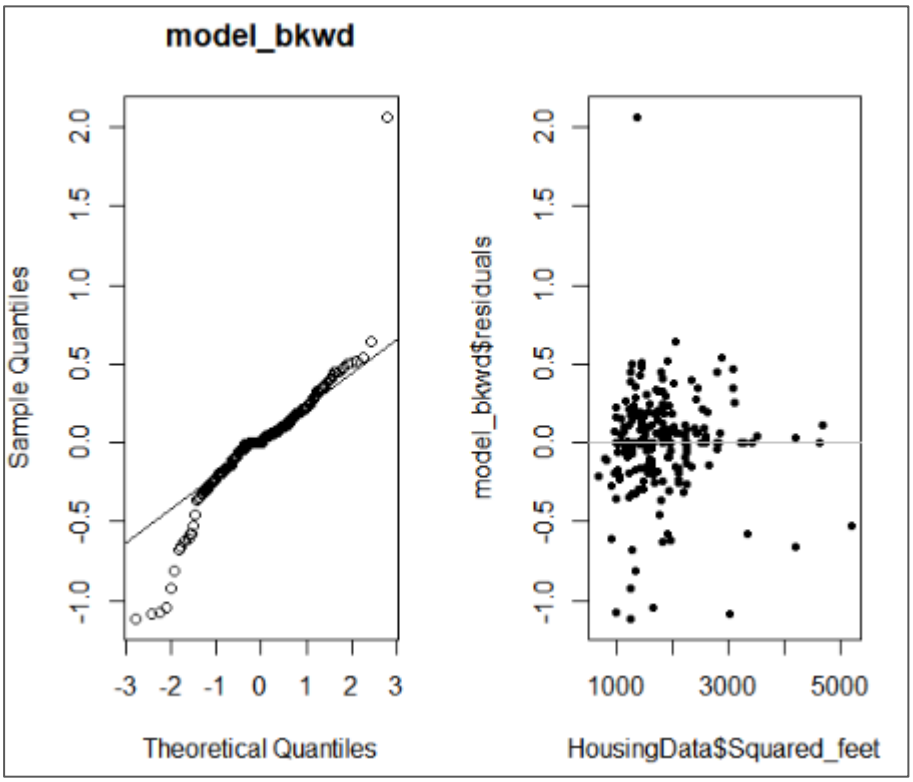
Goal: After correcting the cons of the previous model we aim at obtaining a more accurate and appropriate model for predicting the housing prices.

R code and Output:

```
model_bkwd <- lm(log(Sold_price) ~ (Squared_feet) + Neighborhood, data=HousingData)
summary(model_bkwd)
```

0

```
Call:
lm(formula = log(Sold_price) ~ (Squared_feet) + Neighborhood,
    data = HousingData)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11429 -0.13441  0.00257  0.15421  2.06368

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                10.3386173  0.3739361  27.648  < 2e-16 ***
Squared_feet                0.0004180  0.0000423   9.883  < 2e-16 ***
NeighborhoodCarthage        0.8467342  0.5026397   1.685 0.093895 .
NeighborhoodCherry Grove    1.2138599  0.3813373   3.183 0.001731 **
NeighborhoodClifton         1.1432723  0.3612308   3.165 0.001837 **
NeighborhoodCollege Hill    0.8579755  0.3648100   2.352 0.019819 *
NeighborhoodCorryville      1.2979572  0.5044818   2.573 0.010936 *
NeighborhoodCUF             0.8841372  0.3833871   2.306 0.022304 *
NeighborhoodEast Price Hill 0.6621114  0.5017923   1.319 0.188768
NeighborhoodEast Westwood   0.3221288  0.5071698   0.635 0.526180
NeighborhoodEastgate        1.0917661  0.4345784   2.512 0.012923 *
NeighborhoodForestville     1.4370435  0.3641855   3.946 0.000116 ***
NeighborhoodHyde Park       1.8276797  0.3919329   4.663 6.25e-06 ***
NeighborhoodLower Price Hill 1.3454060  0.5066725   2.655 0.008671 **
NeighborhoodMt. Airy        0.7514942  0.3623176   2.074 0.039566 *
NeighborhoodMt. Auburn      1.8248435  0.5014150   3.639 0.000362 ***
NeighborhoodNorth Fairmount -0.4598741  0.5019702  -0.916 0.360885
NeighborhoodNorthside       1.0308137  0.3677221   2.803 0.005644 **
NeighborhoodOver-The-Rhine  1.8364328  0.5014989   3.662 0.000334 ***
NeighborhoodPierce TWP      1.5490330  0.5054352   3.065 0.002532 **
NeighborhoodRiverside      -0.0028758  0.5030876  -0.006 0.995446
NeighborhoodSedamsville     0.1262714  0.5064798   0.249 0.803419
NeighborhoodSouth Cumminsville 0.2167667  0.5063199   0.428 0.669101
NeighborhoodSouth Fairmount 0.5594807  0.5052094   1.107 0.269666
NeighborhoodUnion TWP       1.1048188  0.3737819   2.956 0.003559 **
NeighborhoodWest End        1.4288993  0.5021159   2.846 0.004973 **
NeighborhoodWest Price HIll 0.5188195  0.5017251   1.034 0.302563
NeighborhoodWestwood        0.6973617  0.5041161   1.383 0.168365
NeighborhoodWinton Hills    0.8640454  0.5059680   1.708 0.089505 .
NeighborhoodWinton Place    0.9072349  0.5047384   1.797 0.074031 .
NeighborhoodWithamsville    1.0054276  0.3683523   2.730 0.007006 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3546 on 171 degrees of freedom
Multiple R-squared:  0.6627,    Adjusted R-squared:  0.6036
F-statistic:  11.2 on 30 and 171 DF,  p-value: < 2.2e-16
```




```
> AIC(model_bkwd)
[1] 184.6916
> BIC(model_bkwd)
[1] 290.5561
```

Observation & Conclusion:

The final model obtained has an adjusted R-square value of 60.36%, which means that 60.36 % variability can be explained by variable square feet and neighbourhood, which are two key predictors of the house prices. The Q-Q shows a more inclined 45 degree line and the scatter plot is evenly distributed with equal variance, and hence the log transformation was useful. It also helped the R-value to shoot up by 20-25%, which is an incredible increase.