RYLIE - ALLISON - JOSH - MARIA

FINAL PROJECT

Predicting the market value of single family homes in the Coraopolis municipality

SAMPLE DEVELOPMENT

Population: All single family homes in Coraopolis
Sample: 200 single family homes in Coraopolis

- Subsetting original data to contain only single family homes in Coraopolis and only selecting relevant variables
- Using simple random sampling to pull out 200 observations from this subset without replacement
- Checking if the sample is representative of the population
  - Proportion tables for categorical variables
  - Comparison of continuous variable distributions

RELEVANT VARIABLES

- **NEIGHCODE**
  - Code for the name of the neighborhood
- **LOTAREA**
  - Total square footage of land
- **HOMESTEADFLAG**
  - Owner may apply for a homestead reduction and if granted the owner will receive a standard reduction on their assessment for County taxes.
- **COUNTYTOTAL**
  - The assessed property value (land & building together) for county tax purposes.
- **LOCALTOTAL**
  - The assessed property value (land & building together) for local tax purposes.
- **STYLEDESC**
  - Description for building style.
- **STORIES**
  - Story height of the main dwelling
- **YEARBLT**
  - The original date of construction.
- **EXTFINISH_DESC**
  - Description of building material used for exterior walls.
- **ROOFDESC**
  - Description of roofing material.
- **BASEMENTDESC**
  - Description of basement type, if one exists.
- **GRADEDESC**
  - Quality of construction

RELEVANT VARIABLES CONT.

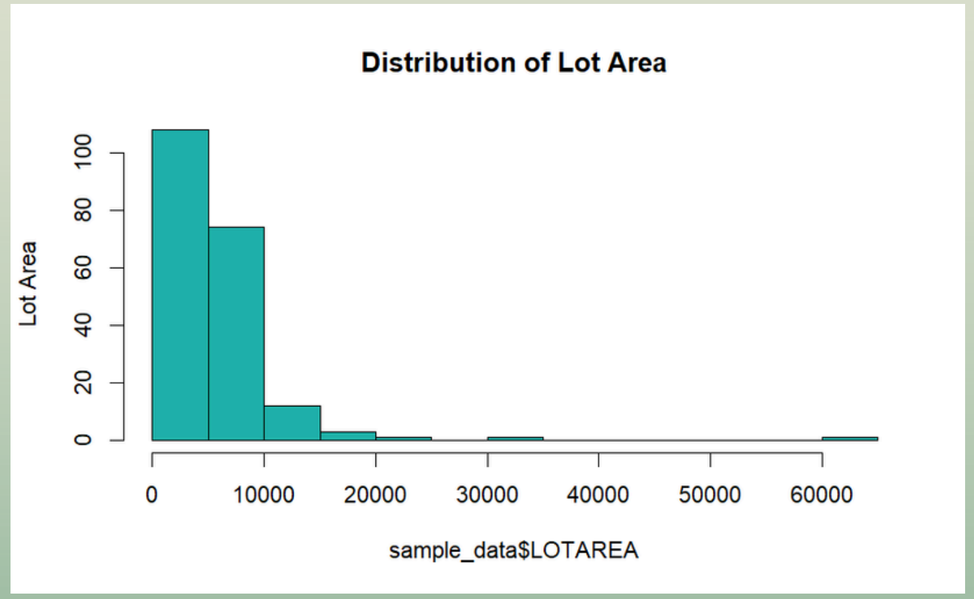
- **CDUDESC**
  - Composite rating for structures measuring: Condition (physical condition relative to age), Desirability (location, style), and Utility (functional obsolescence of layout or design).
- **TOTALROOMS**
  - Total number of rooms in the main dwelling
- **BEDROOMS**
  - Total number of separate rooms designed to be used as bedrooms
- **FULLBATHS**
  - A full bath has a toilet, sink and bathing facility
- **FINISHEDLIVINGAREA**
  - Total Square Feet of Living Area
- **HALFBATHS**
  - A half bath has a toilet and sink only.
- **HEATINGCOOLINGDESC**
  - Description for the type Heating / Cooling system.
- **FIREPLACES**
  - Number of wood-burning fireplace chimneys/vents
- **BSMTGARAGE**
  - Number of vehicle spaces available in a garage that is basement level

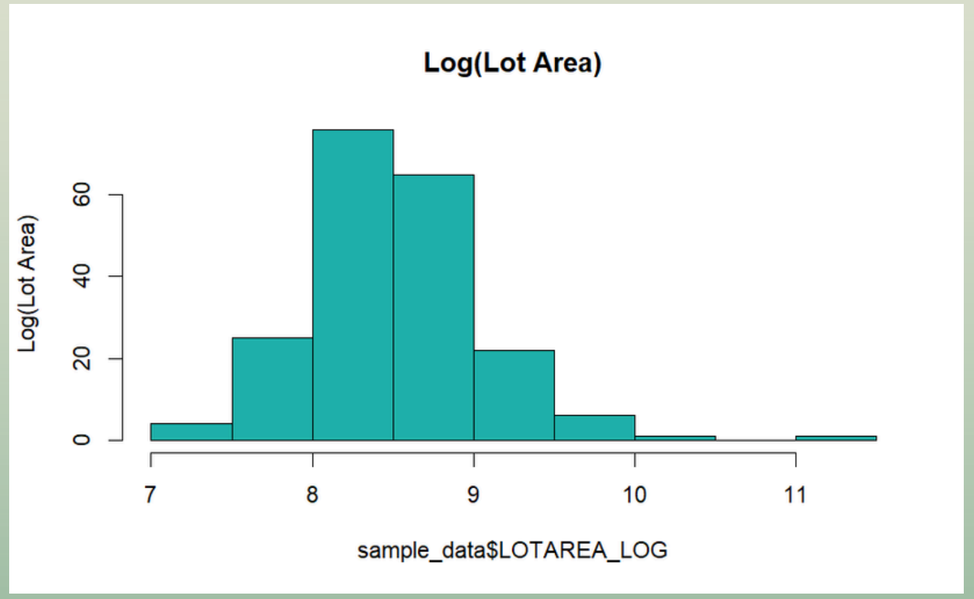
MULTICOLLINEARITY

- Look at correlation matrix
- Removal of highly correlated variables (correlation coefficients more extreme than 0.65 or -0.65)
  - **TOTALROOMS**
    - Correlated with FINISHEDLIVINGAREA and BEDROOMS
  - **COUNTYTOTAL & LOCALTOTAL**
    - Very highly correlated with FAIRMARKETTOTAL
      - Although we want correlation with our target, too much can cause overfitting

VARIABLE MANIPULATION

- Logarithmic transformation of **LOTAREA** in an attempt to make it more normally distributed
  - Some outliers caused for a right-skewed distribution of this variable in the original sample





VARIABLE MANIPULATION

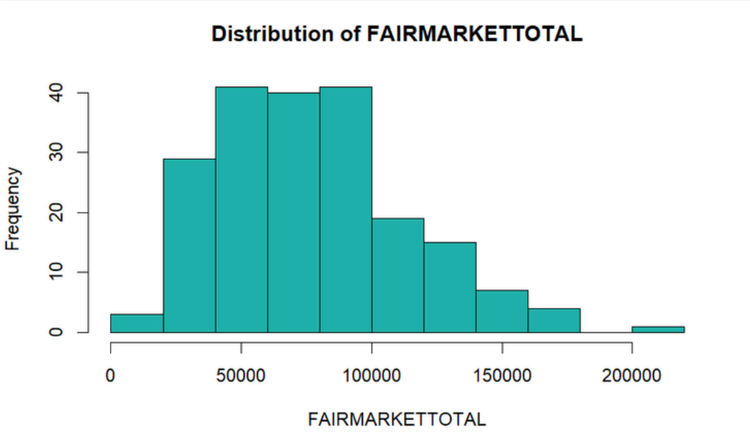
- Almost all observations in both the population and sample fall into the same category for **BASEMENTDESC**
  - Not helpful for prediction of market value
- New variables
  - **AGE**: 2024 - YEARBLT
- Creation of dummy variables
  - **HOMESTEADFLAG**: homestead reduction or not
  - **STYLEDESC_OLD**: old style home or not
  - **ROOFDESC_SHINGLE**: shingle roof or not
  - **NEWGRADEDESC**: quality of construction grouped by below average, average, and above average

FINAL VARIABLES

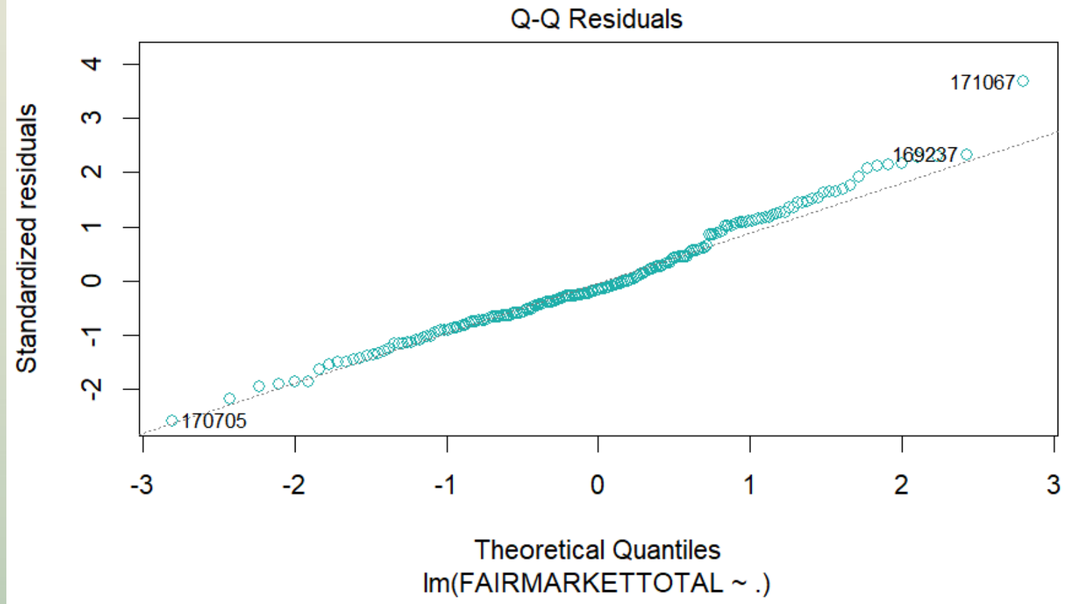
- NEIGHCODE
- LOTAREA_LOG
- HOMESTEADFLAG
- STYLEDESC_OLD
- STORIES
- AGE
- EXTFINISH_DESC
- ROOFDESC_SHINGLE
- NEWGRADEDESC
- CDUDESC
- BEDROOMS
- FULLBATHS
- HALFBATHS
- HEATINGCOOLINGDESC
- FIREPLACES
- BSMTGARAGE
- FINISHEDLIVINGAREA

FULL MODEL

- Adjusted R2 is .5318 - 53.18% of the variability in **FAIRMARKETTOTAL** can be explained by these predictors
- When testing for normality, the QQ-plot was found to not follow the line of the best fit, and the KS-test confirmed this lack of normality with a p-value of 0.08177
- Once we found that there was no normality with our residuals, we decided to perform a logarithmic transformation on **FAIRMARKETTOTAL** to see if we could get a normal distribution



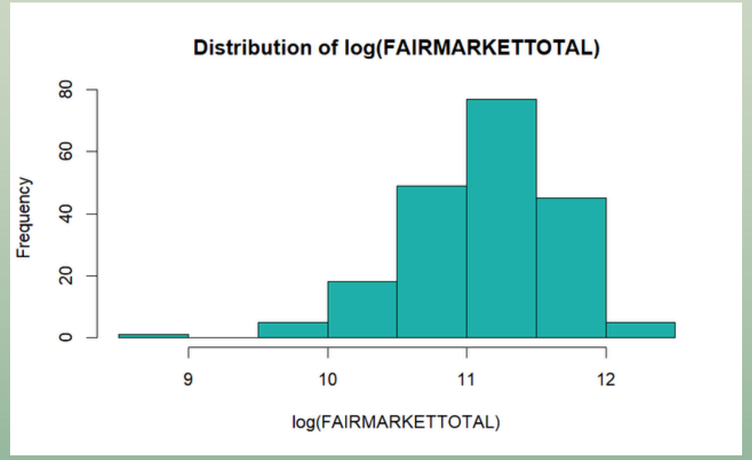

```
Residual standard error: 24720 on 172 degrees of freedom
Multiple R-squared:  0.5953,    Adjusted R-squared:  0.5318
F-statistic:  9.37 on 27 and 172 DF,  p-value: < 0.00000000000000022
```



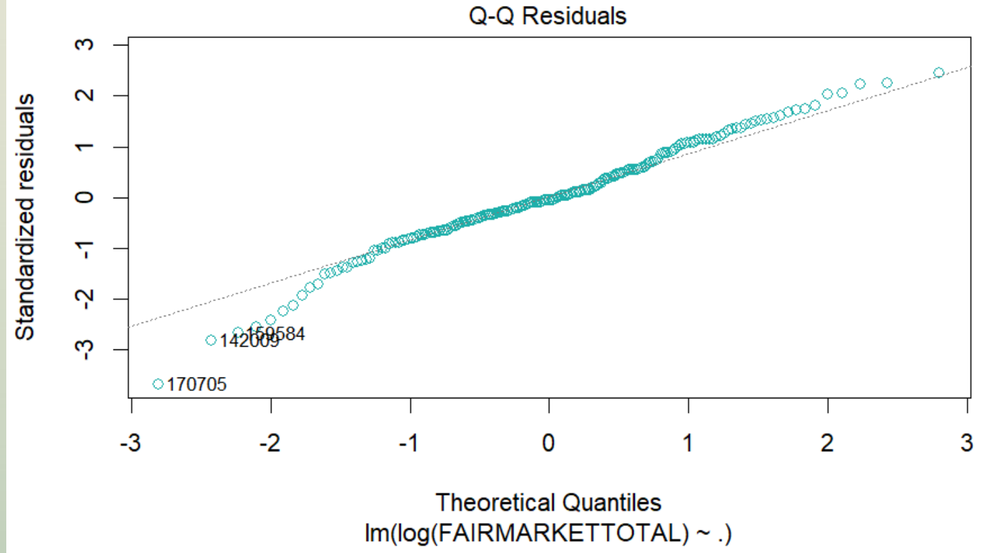

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuals/sd(residuals)
D = 0.0894, p-value = 0.08177
alternative hypothesis: two-sided
```

LOG MODEL

- Adjusted R2 is .5898 - 58.98% of the variability in **FAIRMARKETTOTAL** can be explained by these predictors
- When testing for normality, the QQ-plot was found to not do a great job of following the line of the best fit, but the KS-test showed that we could conclude there was normality with a p-value of 0.4063. The distribution of log(**FAIRMARKETTOTAL**) also supported normality, as the results were normally distributed
- Once we found that there was enough evidence to support there being normality with our residuals, we determined we made our model better, but wanted to find what variables were truly the best predictors of **FAIRMARKETTOTAL**




```
Residual standard error: 0.3346 on 172 degrees of freedom
Multiple R-squared:  0.6454,    Adjusted R-squared:  0.5898
F-statistic:  11.6 on 27 and 172 DF,  p-value: < 0.00000000000000022
```




```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuals/sd(residuals)
D = 0.062953, p-value = 0.4063
alternative hypothesis: two-sided
```

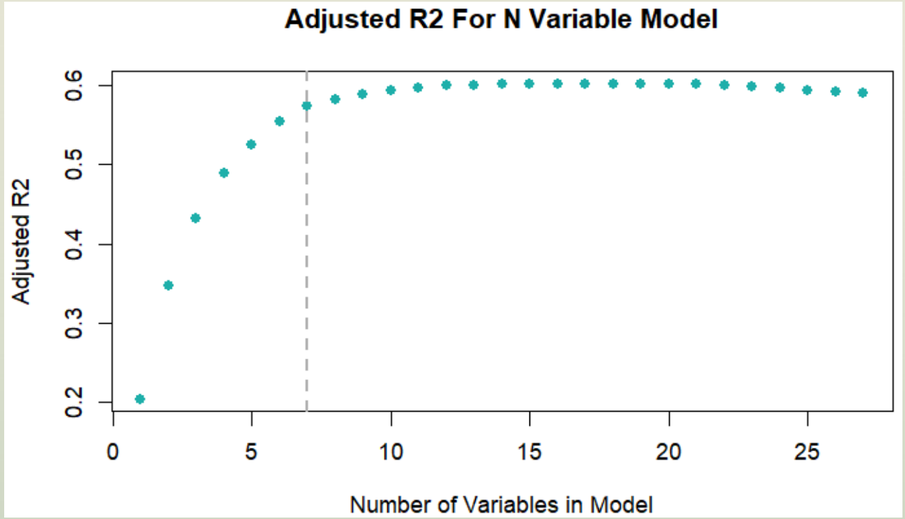
MODEL SELECTION

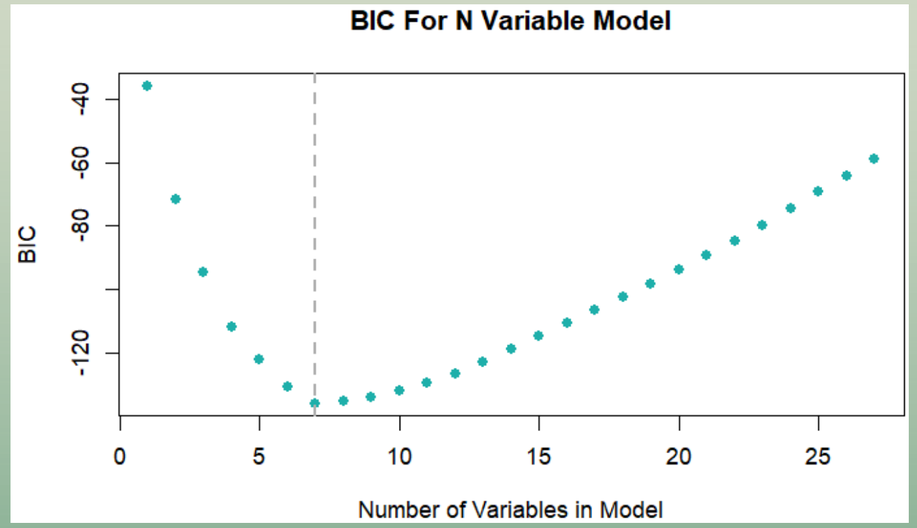
- Used **Best Subset Selection** to perform variable selection on all variables previously discussed to predict log(FAIRMARKETTOTAL)
  - utilized the **regsubsets()** function in the **leaps** package
- Found the best n variable models and what variables were used
- Analyzed **Scree Plots** (Elbow Plots) to determine the optimal number of variables to use
  - **Mallow's Cp** - Balances the tradeoff between complexity and fit using the error and number of parameters
  - **R2** - How much variability is explained by the best n predictors
  - **BIC** - Measures goodness of fit while penalizing for additional parameters

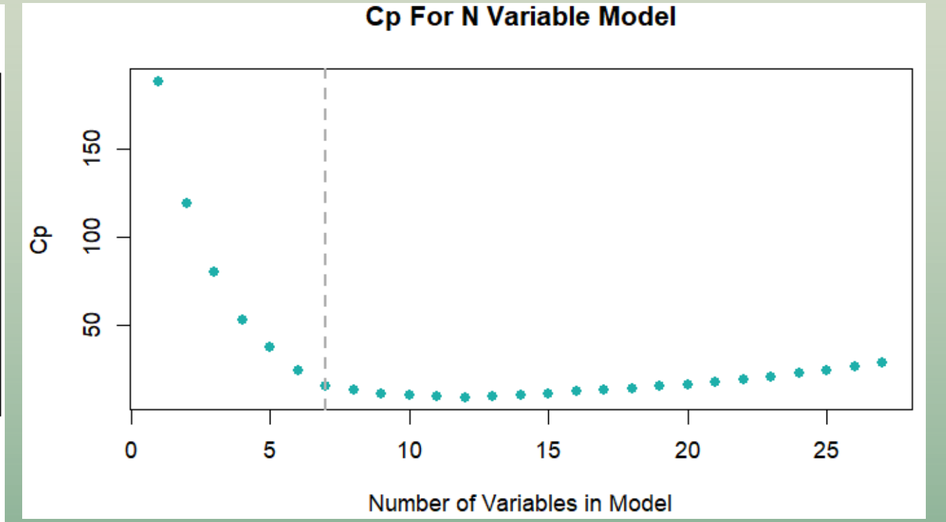
SCREE PLOTS

Decided to use the **7 variable model**. The significant variables were:

- Heating/Cooling - Central Heat with AC
- Neighborhood Code - 81703
- CDU Description - Unsound
- CDU Description - Very Good
- Finished Living Area
- log(Lot Area)
- Homestead Flag







FINAL MODEL

- Adjusted R2 is .5925 - 59.25% of the variability in log(FAIRMARKETTOTAL) can be explained by these predictors
- All variables are significant except HEATINGCOOLINGDESCNone, NEIGHCODE81702, CDUDESCFAIR, CDUDESCGOOD
- Exponentiated coefficients are located in the chart on the right - meaning that a one unit increase in these values will result in a change of the coefficient for FAIRMARKETTOTAL
  - interpretations differ slightly for each variable depending on data type

```
Call:
lm(formula = log(FAIRMARKETTOTAL) ~ LOTAREA_LOG + HEATINGCOOLINGDESC +
    HOMESTEADFLAG + NEIGHCODE + CDUDESC + FINISHEDLIVINGAREA,
    data = model_data)
```

```
Residual standard error: 0.3335 on 187 degrees of freedom
Multiple R-squared:  0.617,     Adjusted R-squared:  0.5925
F-statistic: 25.11 on 12 and 187 DF,  p-value: < 0.00000000000000022
```

Exponentiated Coefficients

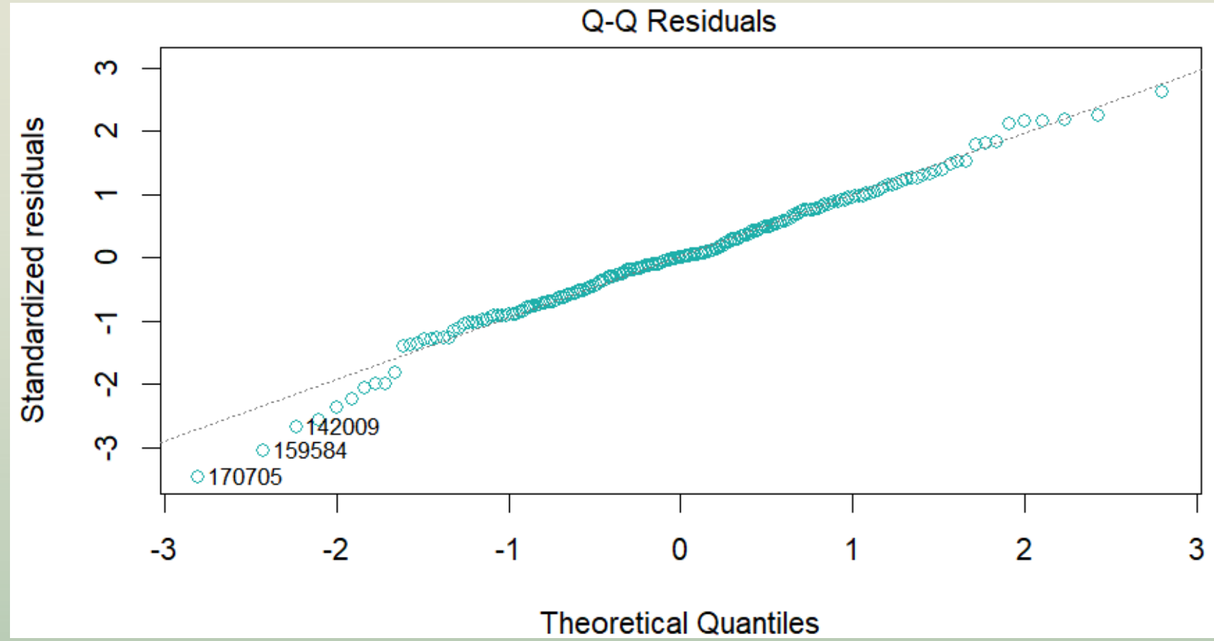
(Intercept)	NEIGHCODE81703	
8.6680916423	-0.3721765184	
LOTAREA_LOG	HOMESTEADFLAG	
0.1998307060	0.1786253123	
CDUDESCUNSOUND	CDUDESCVERY GOOD	
-2.1184789166	-1.1034100928	
HEATINGCOOLINGDESCCentral Heat with AC	FINISHEDLIVINGAREA	
0.2211143263	0.0004616112	

VERIFYING ASSUMPTIONS

Verified the normality assumptions for the error term of our final model

- **QQ plot**
  - Points closely follow the line in the middle of the graph
  - Points stray somewhat significantly at the tail ends of the plot with the extreme deviations labeled at the bottom

- **Kolmogorov-Smirnov test**
  - Test statistic of 0.04 and a p-value of 0.72
  - We would have a 72% chance of getting the results that we did if the null hypothesis were true and the error term was normally distributed. The high p value means that we do not reject the null hypothesis and conclude that the distribution of the error terms is normally distributed.




```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  residuals_1/sd(residuals_1)
D = 0.049046, p-value = 0.7218
alternative hypothesis: two-sided
```

GOING FORWARD

Though our model improved throughout our analysis, the best R2 value that we achieved was 59.25%. The means that 40.75% of the variation is left unexplained by our variables. Other variables that were not included in the dataset that could influence the market value of a home are...

- Whether the house is located on a main road or not
  - could be a categorical variable
- If the house/property includes a pool
- If the house/property includes a garage
  - the dataset only had a variable for garages attached to the home
- The distance of the closest neighboring house
- Current market conditions
  - Could maybe make a ranking scale to give an idea of the market conditions at the time

THANK YOU