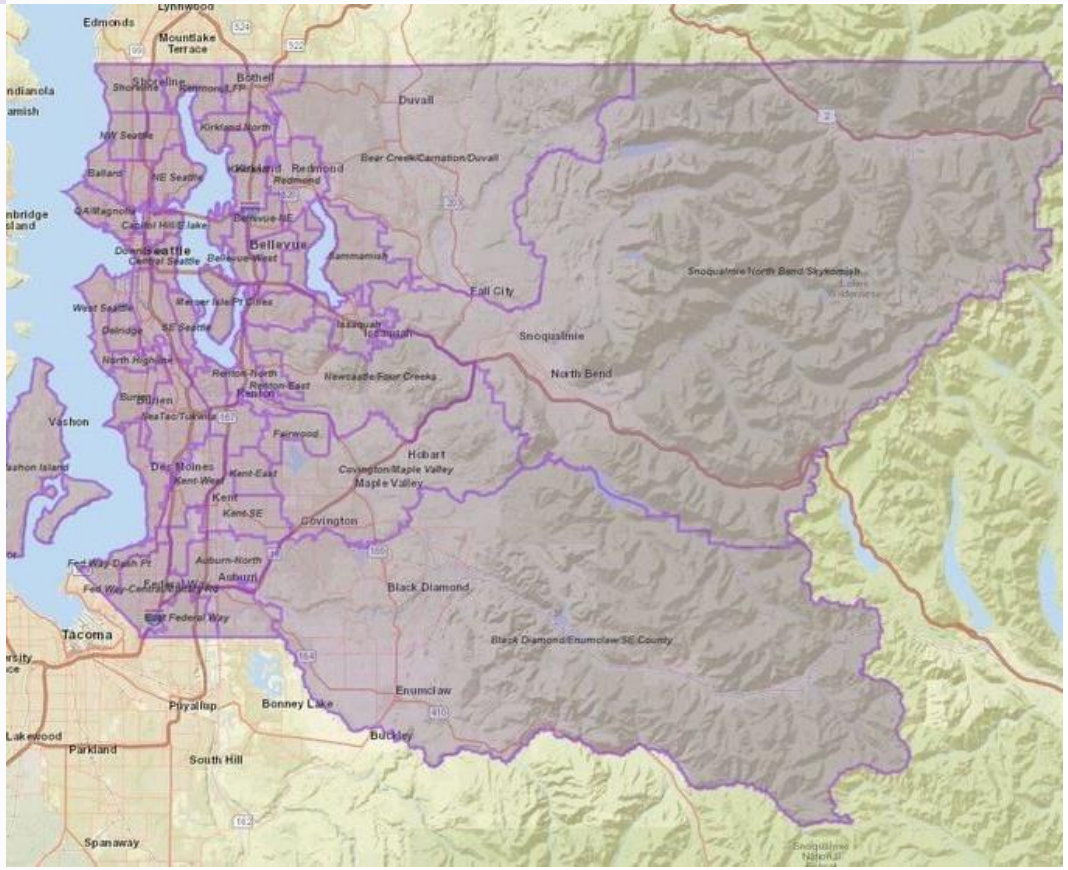
King County Data Analysis

Kevin Gross



Agenda

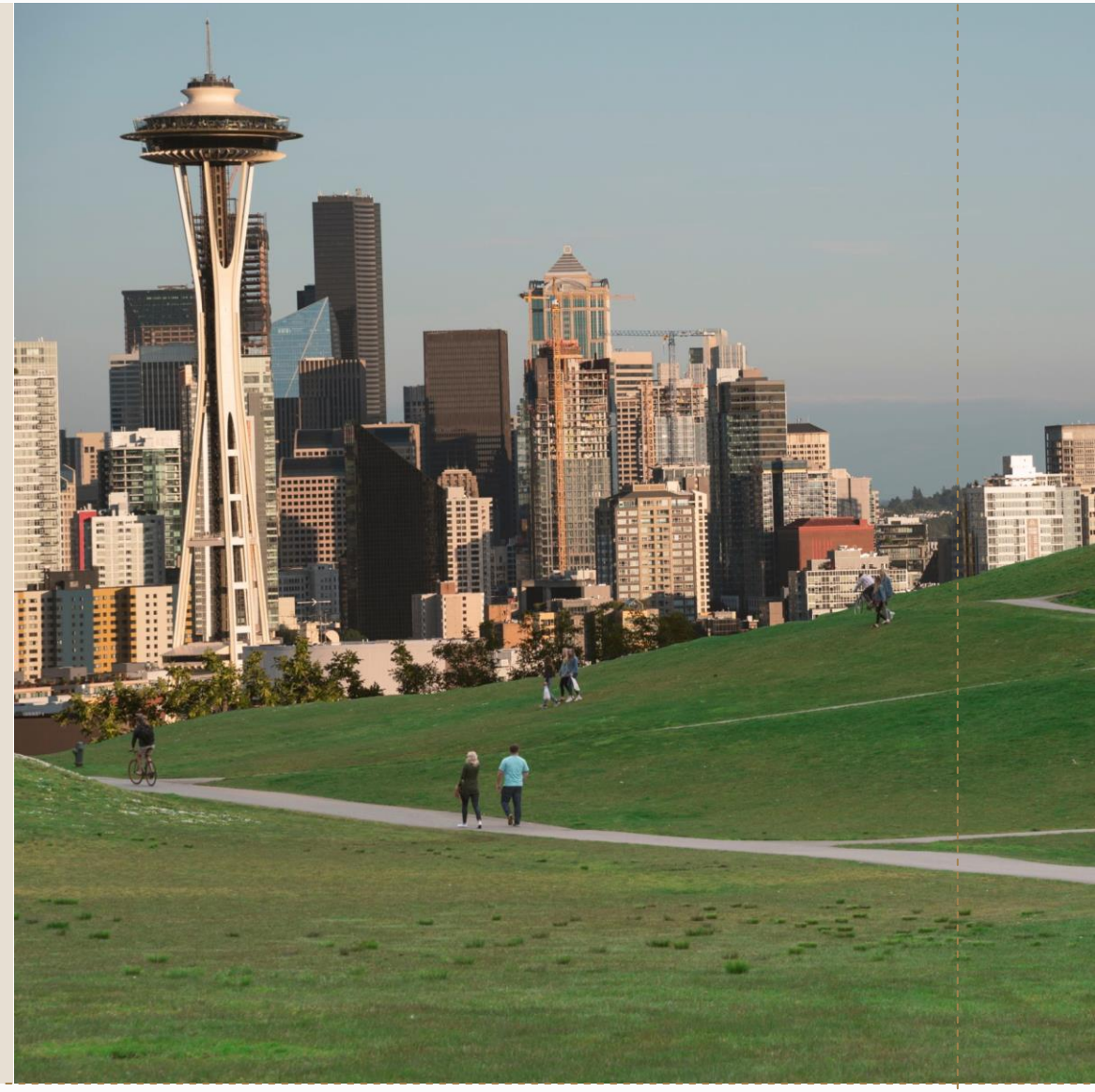
- Overview
- Business & Data Understanding
- Modeling and Terminology
- Regression Results
- Recommendations
- Summary

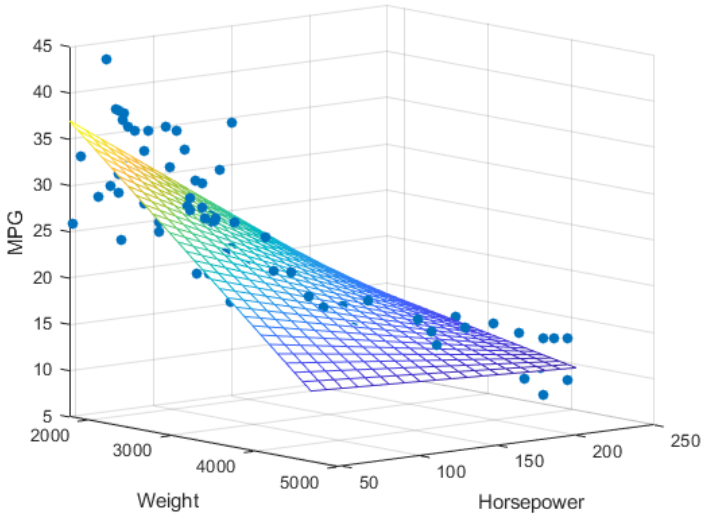
Overview

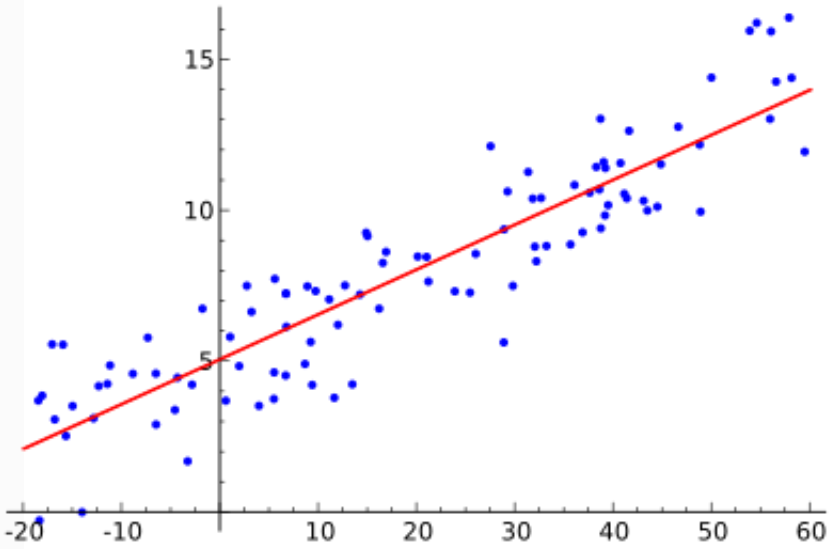
This project analyzes housing data from King County, Washington to help **interpret and predict future house sale prices**.

Business and Data Understanding

The stakeholder for this project is **Berkshire Hathaway's clients**, also known as their home buyers and home sellers. Berkshire Hathaway's goal is to use the data provided to **provide advice to their clients on how to increase the value of their home**. The regression analysis starts off at a basic level with 1 dependent variable predicting the price of a home and then builds using multiple iterations of linear regression.







Modeling and Terminology

This project uses **linear regression** to formulate a **baseline model** to predict King County housing prices.

Linear Regression: a model that assumes a linear relationship between the input variables (x) and single output variable (y)

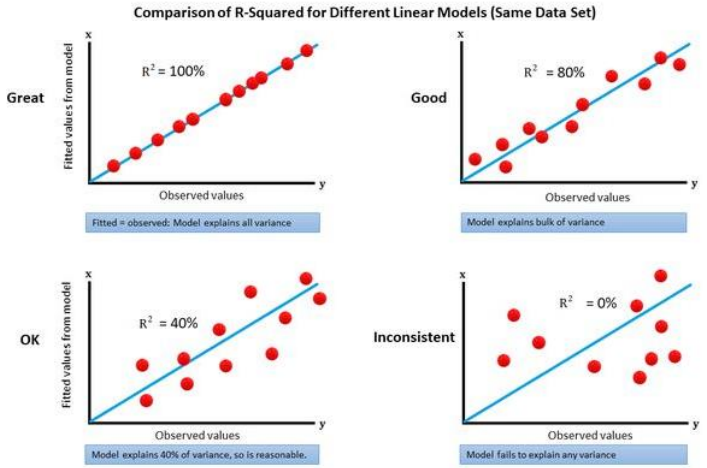
Modeling and Terminology

Modeling

- Next, additional **numeric variables** are added into the model to increase the **r-squared** value.
- Lastly, **categorical variables** are added to increase the **r-squared** value even more.

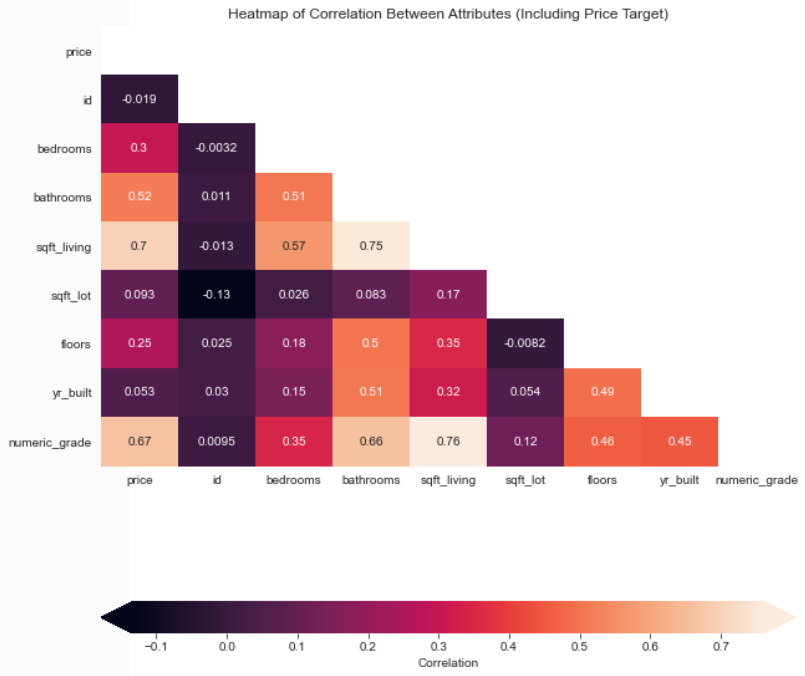
Terminology

- **R-Squared:** the proportion of the variation in the dependent variable that is predictable from the independent variable(s).
- **Numeric Variables:** values that describe a measurable quantity as a number, like 'how many' or 'how much'.
- **Categorical Variables:** values that have two or more categories, but there is no intrinsic ordering to the categories

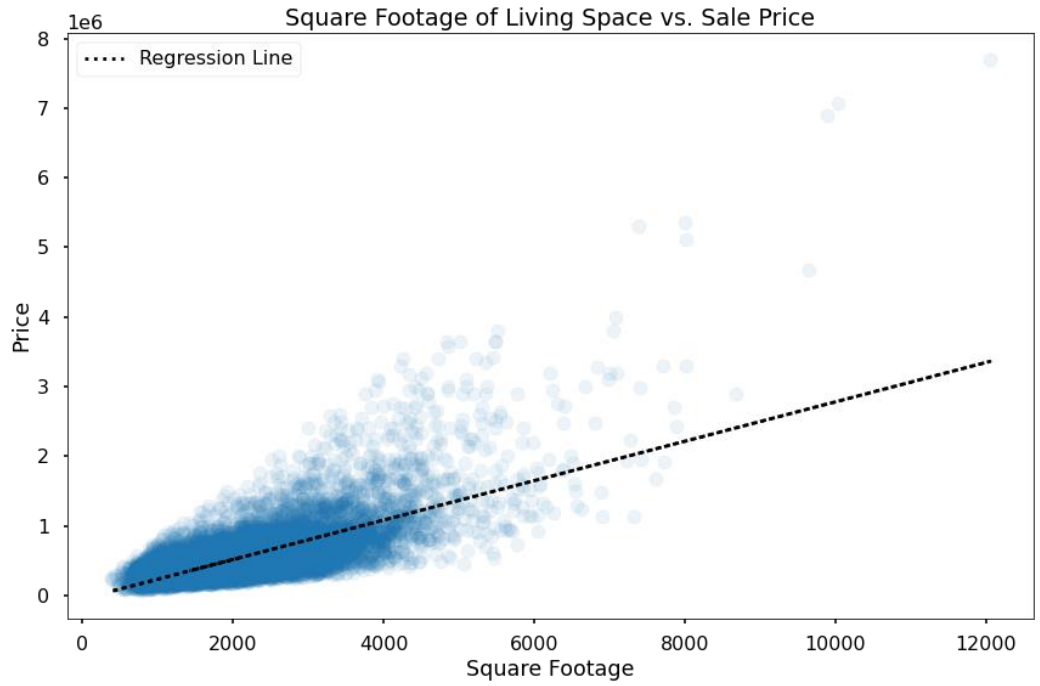


Regression Results

Pearson Correlation



R-Squared: 0.49 or 49%

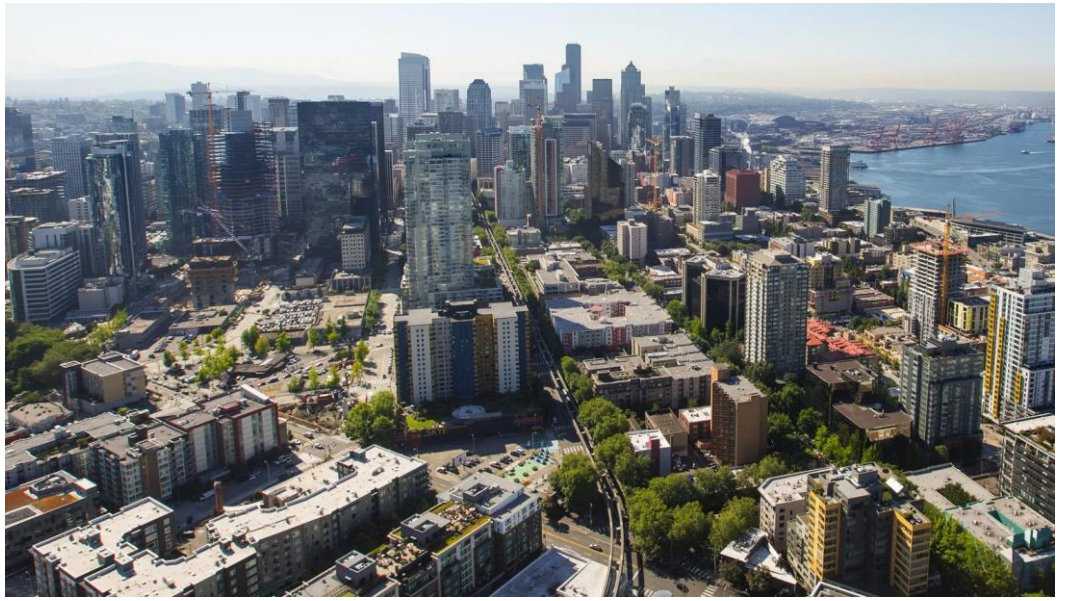


Linear Regression Model #2

Input Variables

- Number of Bedrooms
- Number of Bathrooms
- Square Footage of the Lot
- Number of Floors
- Year House was Constructed
- Numeric Grade House Received
- Square Footage of House

R-Squared: 0.614 or 61.4%



Linear Regression Model #3

R-Squared: 0.644 or 64.4%

Input Variables

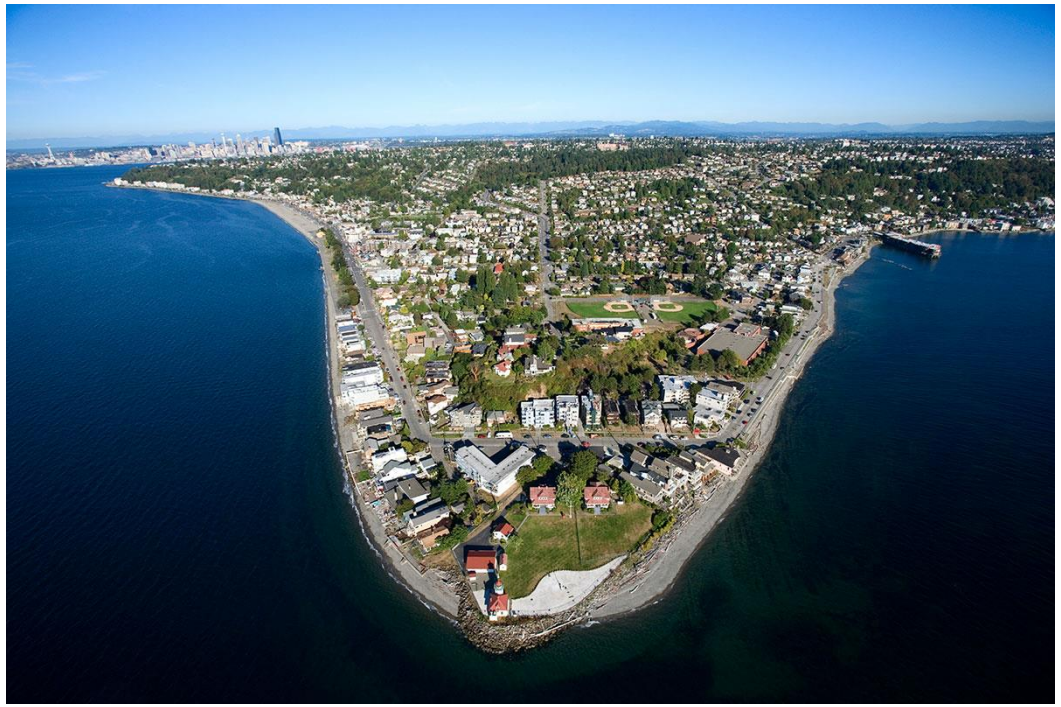
Square Footage of House	Number of Bedrooms	Number of Bathrooms
Square Footage of the Lot	Number of Floors	Year House was Constructed
Numeric Grade House Received	Fair Condition (Y/N)	Good Condition (Y/N)
Poor Condition (Y/N)	Very Good Condition (Y/N)	Waterfront (Y/N)

Linear Regression Model #4

Input Variables

- Square Footage of House
- Number of Bedrooms
- Number of Bathrooms
- Year Built
- Numeric Grade
- Waterfront (Y/N)

R-Squared: 0.642 or 64.2%



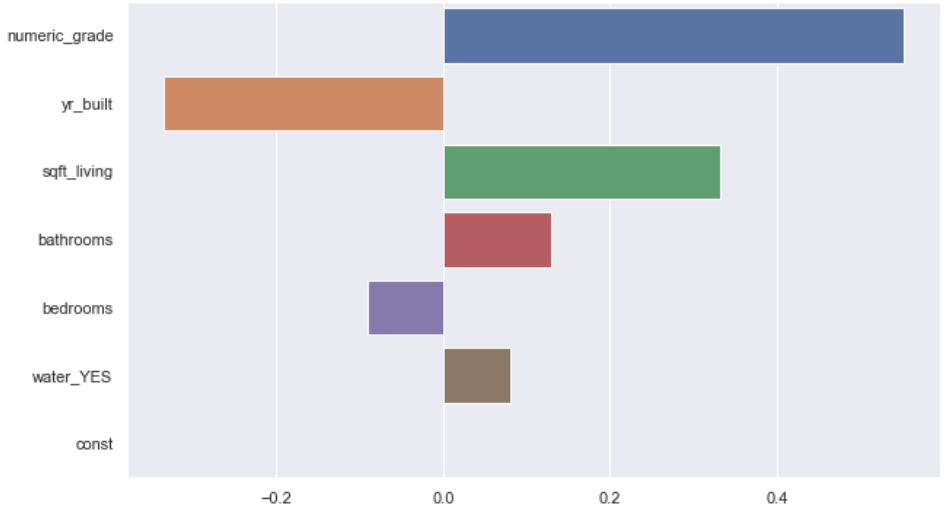
Linear Regression Model #5

Same variables as model #4

Meets all 4 assumptions of linear regression:

Linearity	Multicollinearity	Normality	Homoscedasticity

R-Squared: 0.634 or 63.4%



Business Recommendation #1

- The best way to increase the value of one's home is to increase the **Numeric Grade**.
- This is indicated by an index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high-quality level of construction and design.
- Since this is on a scale of 1 to 13, if we increase our Numeric Grade by 1 notch (1/13 = 7.69%), we expect the value of our home to increase by 4.2%.

Business Recommendation #2

- Another way to increase the value of one's home is to **increase the square footage of the house**.
- If we increase the square footage of the house by 20%, we expect the value of the home to increase by 6.2%.

Business Recommendation #3

- A third way to increase the value of one's home is to **add more bathrooms**.
- If we add 1 bathroom to a 2.5 bathroom house (1/2.5 = 40%), we expect the value of the home to increase by 4.4%.

Summary

Baseline Model

R-Squared: 0.49

Input Variables: 1 (Square Footage)

Second Model

R-Squared: 0.614

Input Variables: 7 (All Numeric Variables)

Third Model

R-Squared: 0.644

Input Variables: 12 (All Numeric and Categorical Variables)

Fourth Model

R-Squared: 0.642

Input Variables: 6 (Combination of Numeric and Categorical Variables)

Final Model

R-Squared: 0.634

Input Variables: 6

Meets all linear regression assumptions

Questions?

- Kevin Gross
- kevgross89@gmail.com
- github.com/kevgross89