MAT 303-Applied Statistics | Project One Summary Report

Ari Meier

1. Introduction

This project explores a housing data set titled *housing_v2.csv*. The analysis will focus on the relationship between the square footage, or number of bedrooms, and its selling price. Three models will be employed:

1. First-order regression: This model will examine the connection between the selling price and one or more independent variables (likely square footage or bedrooms). It will consider both quantitative and qualitative variables if present in the data set.

2. Complete second-order multiple regression with quantitative variables: This model builds upon the first by incorporating squared terms of the independent variables. This allows for the exploration of more intricate relationships between the variables and the selling price. It's important to note that this model will only focus on quantitative variables.

3. Nested models F-test: This test will statistically compare the first two models to determine if the added complexity of the second-order model significantly improves the explanation of the selling price.

2. Data Preparation

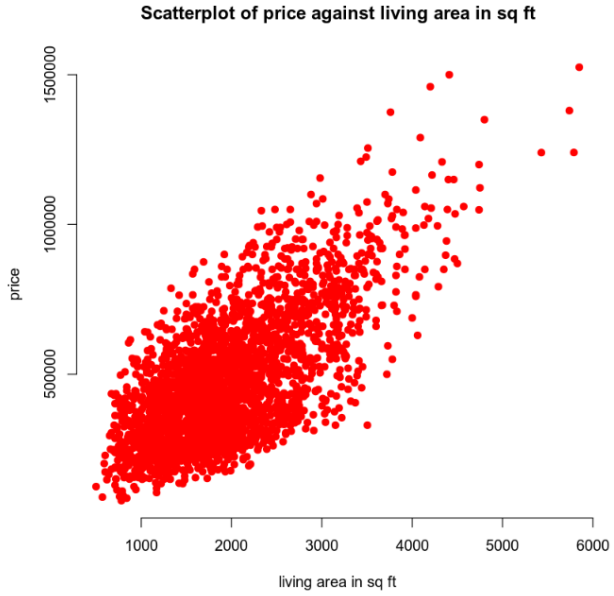
The *housing_v2.csv* data set contains information on 2,692 houses with 23 variables each. Key variables of interest for this analysis include:

- **Price:** Selling price of the house
- **Housing Characteristics:** Number of bedrooms, bathrooms, square footage (upper and lower levels), lot size (square feet)
- **Property Condition:** Age of the home, quality grade
- **Appliance Age:** Average age of all appliances in the home
- **Neighborhood Characteristics:** Crime rate per 100,000 people, average school rating
- **Property Features:** Backyard presence (binary variable)
- **View:** Categorical variable indicating the view (lake, trees, or road)

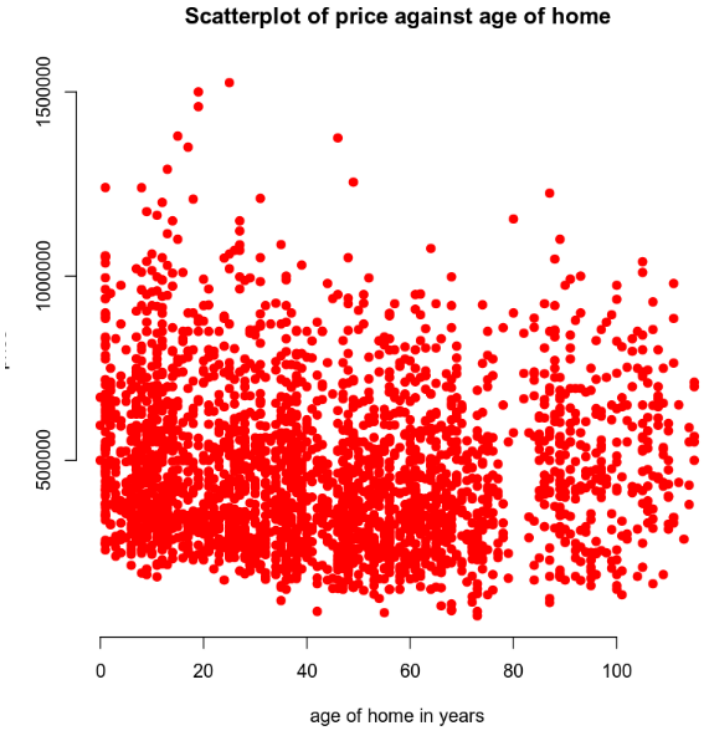
3. Model #1 - First Order Regression Model with Quantitative and Qualitative Variables

Correlation Analysis

Price (price) vs. the living area (sqft_living)



Price (price) vs. the age of the home (age)



There is a positive relationship between the size of the living area and home sales price. This relationship is sales price increase as the size of the house increases. There was no strong relationship between the age of the homes and sales price.

Report the correlation coefficients between the following variables:
The correlation between price (price) vs. living area (sqft_living) is strong positive (0.6895) and the correlation between price (price) vs. the age of the home (age) is weak, negative (-0.0746).

Reporting Results

The general form equation for the first order regression model with price as the response variable and living area, and age of the home variables is:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

The β_0 is the intercept. The $\beta_1 \text{ and } \beta_2$ values are the terms for living area, age of the home. The $X_1, \text{ and } X_2 \text{ terms}$ are the variables for living area, and age of the house.

$\hat{Y} = -11470.393 + 215.829X_1 + 1439.334X_2$

The value of R^2(R-squared) is 0.5084 and R_a^2(Adjusted R-squared) is 0.5081. This shows that approximately 51% of the variance in price can be explained by a model that uses living area, and age of the house as predictor variables.

The beta estimate for living area is 215.829. This means that for every 1 unit increase in living area, the price will increase by this much.