King County Housing Analysis

Paul Gitonga Njoki

paulgitonga58@gmail.com

Overview

BUSINESS PROBLEM

BUSINESS OBJECTIVES

DATA STRUCTURE AND DATA UNDERSTANDING:

01 - Overview

King County is in the state of Washington in the United States. It is Washington's most populous county and the 13th most populous in the United States. Seattle, the state's most populated city, serves as the county seat.

Problem statement

01 - Introduction

BUSINESS PROBLEM

A Seattle real estate agent wants to discover which elements have a substantial impact on the price of a house in King County. This will help in strategizing on the best criterion to use to maximise profit. The company has tasked me with developing a model that will be used to estimate property prices in King County and obtaining substantial advice on activities that they should take to ensure the business's success.

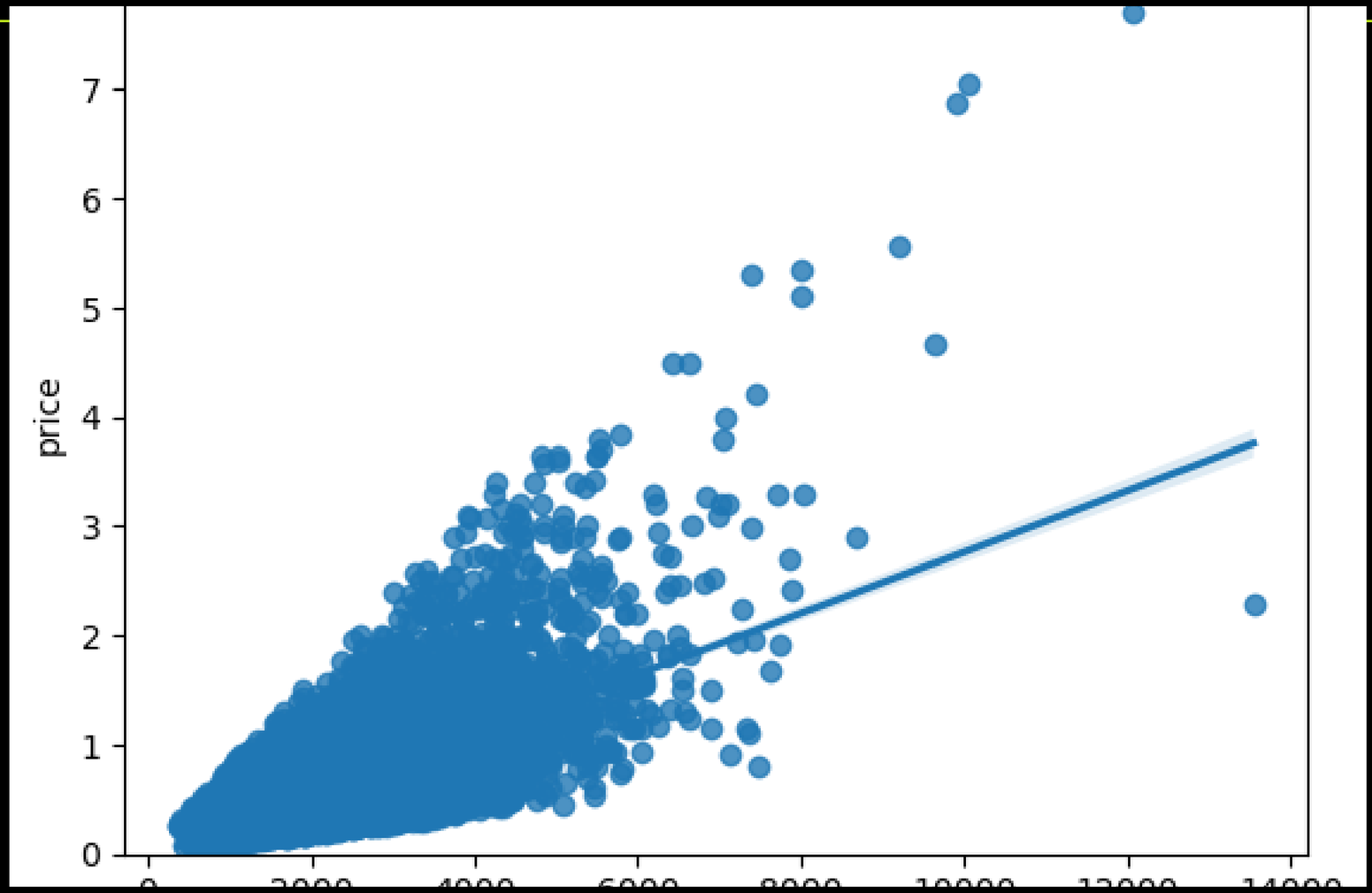
BUSINESS OBJECTIVES

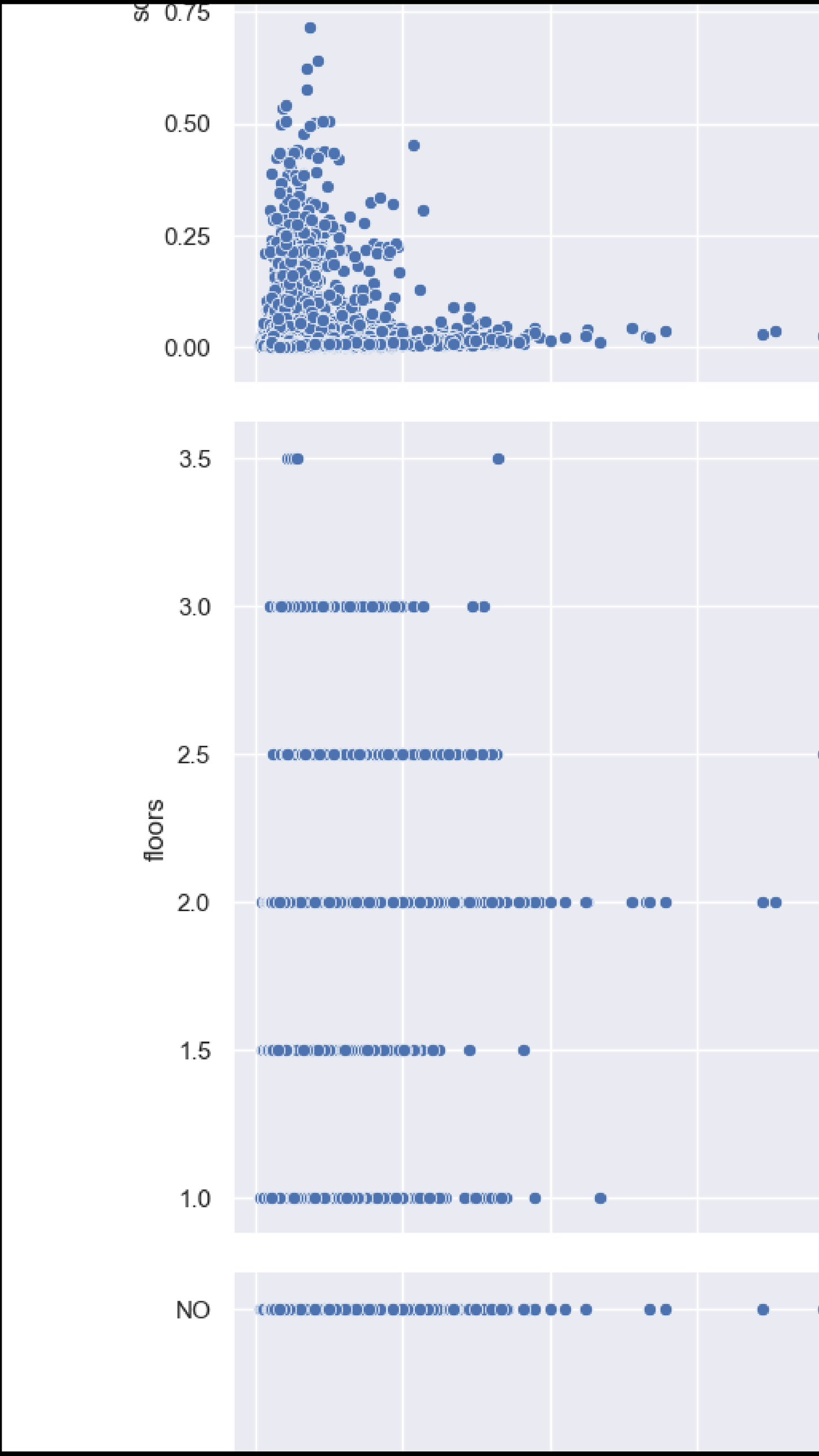
- To understand factors that are most predictive of price.
- Which house features will give the best deals.
- Obtain a model that will be of use when predicting the price of a property.

02 - History

Data visualization simplifies the communication of analysis findings

03 - Data Visualization




0.75
0.50
0.25
0.00
3.5
3.0
2.5
floors
2.0
1.5
1.0
NO

Data Preparation Report

We first dropped the columns that won't be needed for the analysis.

Our target variable is price which will be used in creating a model.

Removed outliers in our independent variables.

Examining the distributions of price it contained outliers which were removed and performed a log transformation.

Missing values were removed by dropping rows with the nulls.

In correlation between price and other features, we have sqft_living, sqft_above, sqft_living15 and bathrooms having high positive correlation. This will be used in creating our model.

we will perform one-hot encoding for the categorical variables waterfront, grade, condition so as to be used in creating a model.

Interpretation of our Multiple Linear Regression Model

The model built is: price = -1.021e+05 + 249.0055sqft_living + 62.9802sqft_living15

The model explains 50.2 % variation in price.

The model coefficients (const, sqft_living, and sqft_living15) are all statistically significant, with t-statistic p-values well below 0.05

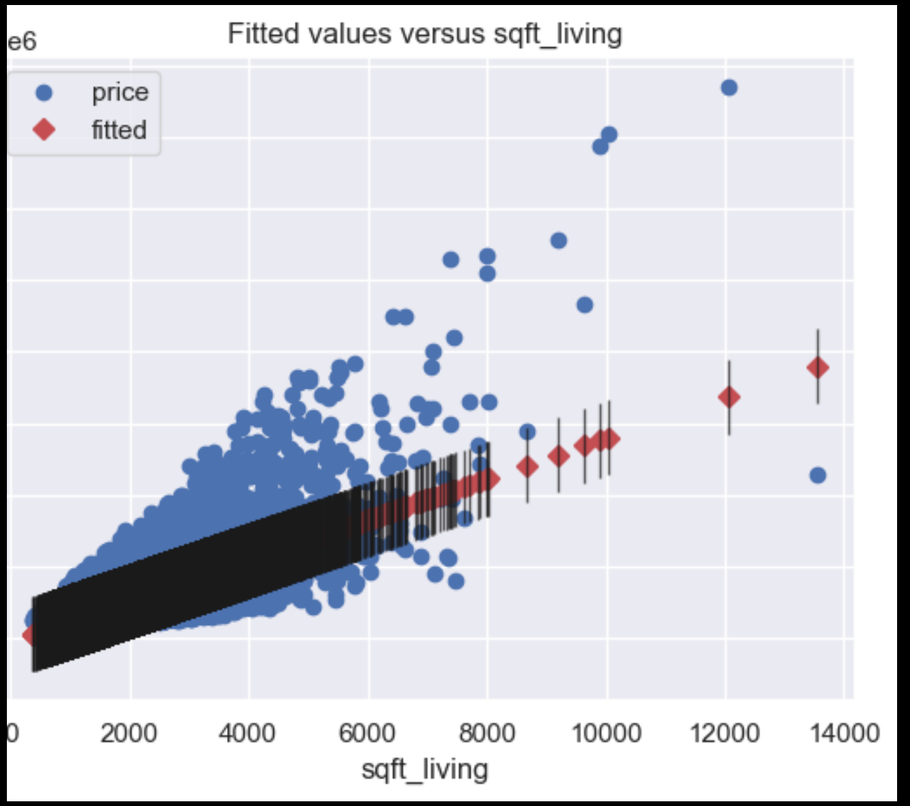
For each increase of squarefoot, we see an associated increase in price of about 249.0055.

This is a little bit smaller of a increase than we saw with the simple model, but not a big change.

For each increase of 1 in sqft_living15, we see an associated increase in price of about 62.9802.

All of the p-values round to 0.

#relationship between price and sqft_living


Fitted values versus sqft_living
e6
price
fitted
0
2000
4000
6000
8000
10000
12000
14000
sqft_living

Thanks.

Paul Gitonga
Njoki