MODELLING HOUSE PRICES

Analysing trends observed with regards to price of property sold in the King County Area

Megan Grant
Flatiron School
Part-Time Data Science
Module 1

OBJECTIVES

During this investigation my aim was to produce data analysis that would provide stakeholders with *key insight into factors affecting house price* in the King County area.

Ultimately I aim then to provide ***a predictive model*** for house prices in the area.

Data was provided from over *20,000 house sales between 5th May 2014 and 27th May 2015* , providing enough information for us to be able to perform multiple different lines of enquiry.

Example 1: Sqft of the largest home sold in the King County area in this time:
9410sqft

Example 2: Worst condition a house was sold in, as per the King County grading system:
Grade 3

Example 3: Largest lot of a house sold:
871,200sqft

This data spanned *all houses* in King County; a consideration when evaluating the conclusions here today is that they were produced with the aim of being applicable to typical homes of the area.

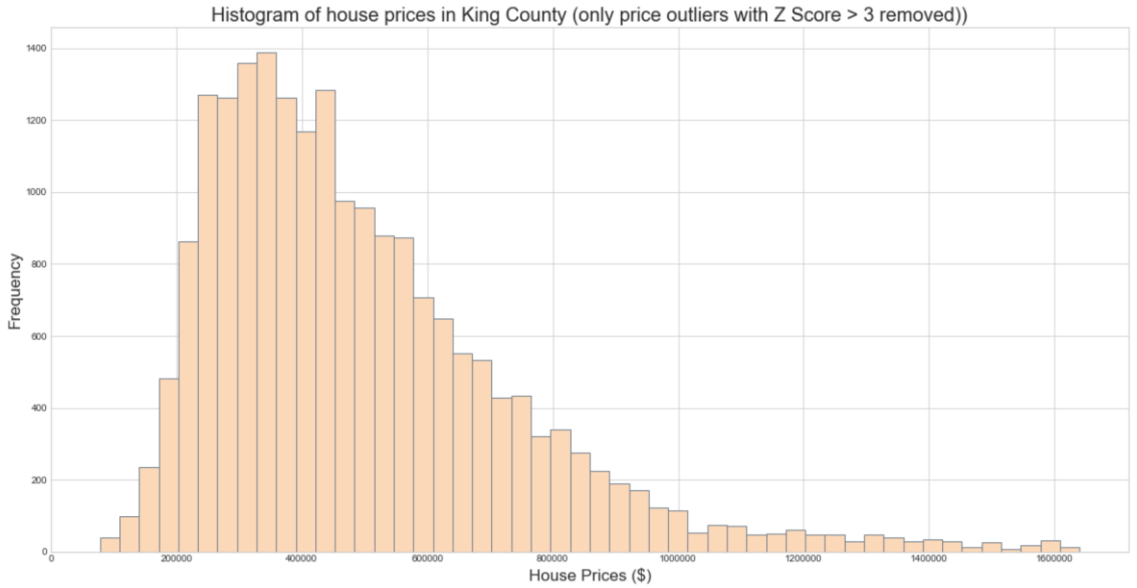
DATA CONSIDERATIONS AND IMPACT ON ANALYSIS

Missing or misformatted data was cleaned initially and outliers removed

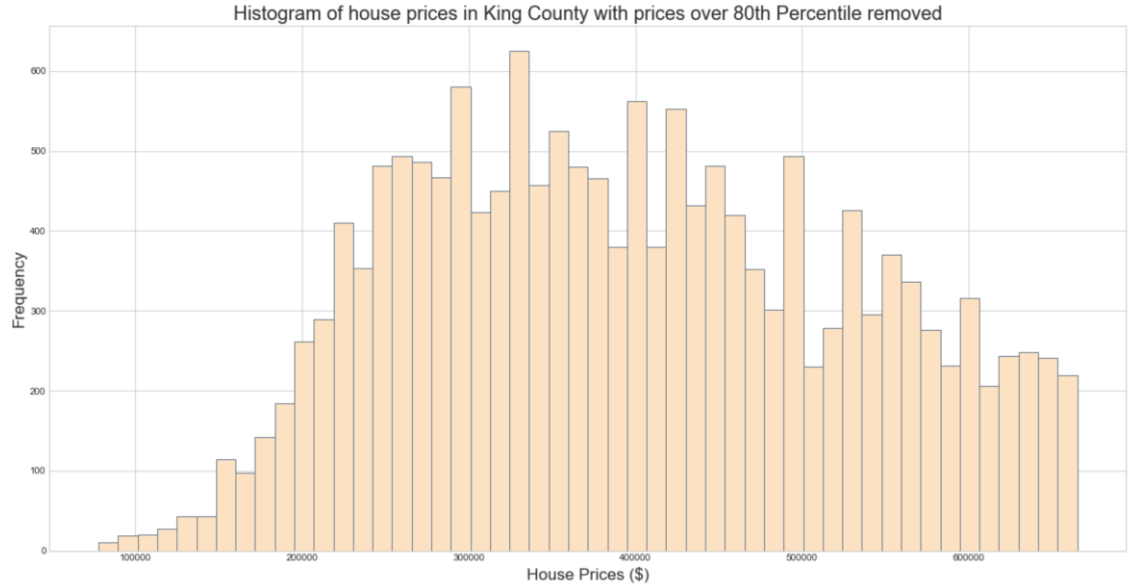
- All prices that were within 3 standard deviations of the mean value were removed

Resulting price data was skewed to lower cost homes, so the decision was taken to use only the bottom 80% of cleaned data.

Our resulting model therefore will predict house prices up to $665,000.



Above: Histogram of cleaned house prices before tail 20% is removed; the skew is evident.



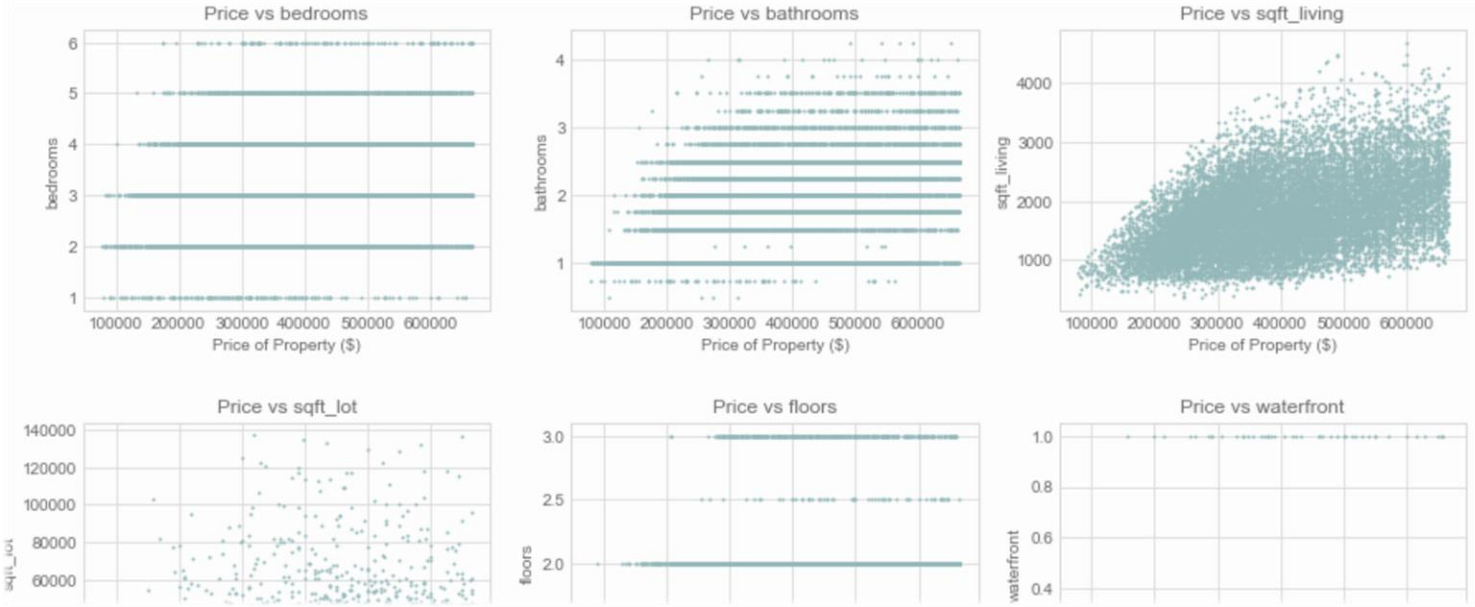
Above: Histogram once the tail 20% is removed; small skew remains but is overall more normal.

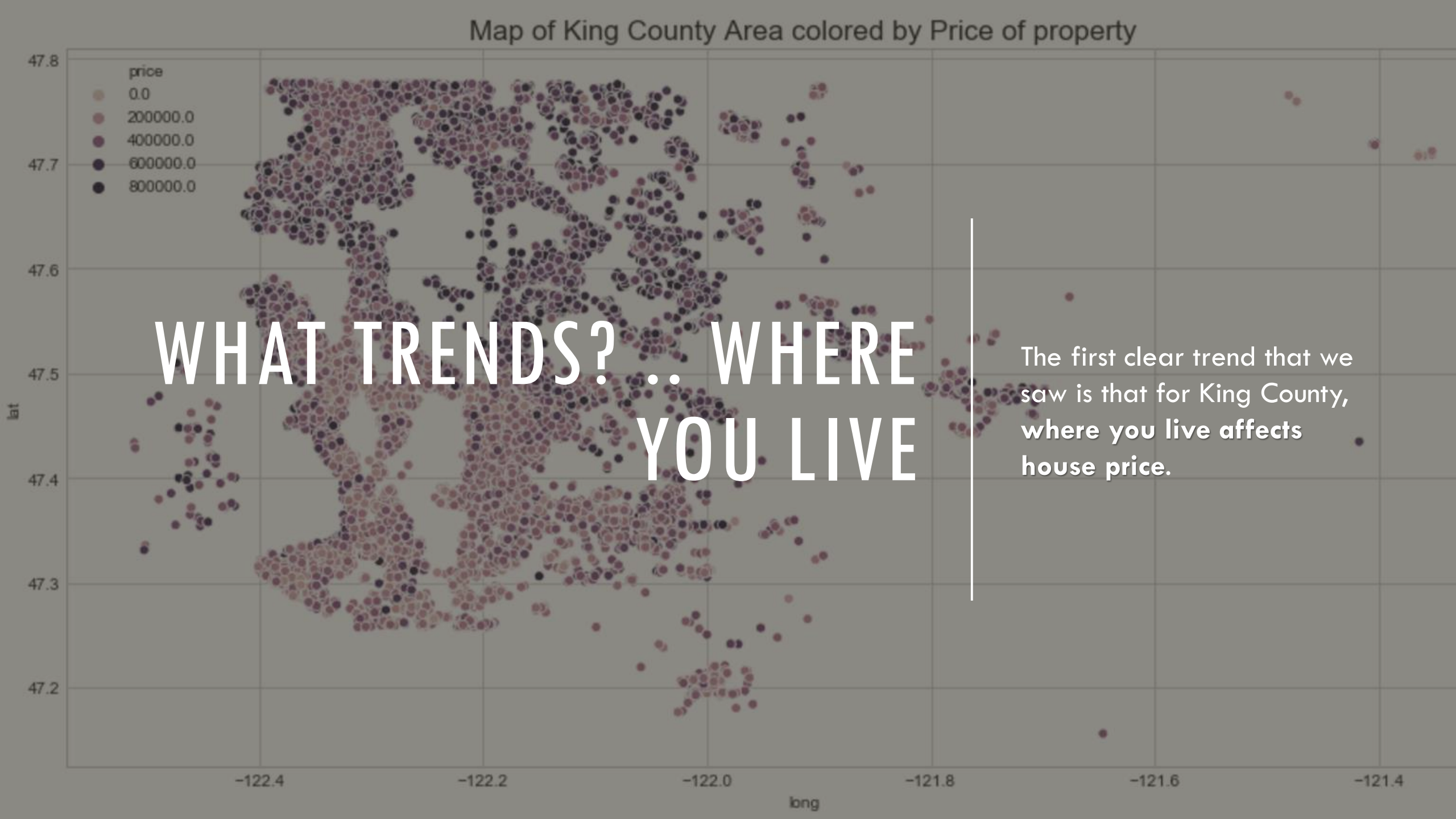
IS LINEAR REGRESSION APPROPRIATE?

One key consideration here is that we will be using linear regression to describe the trends seen in house prices in the King County Area.

All of the assumptions of linear regression have been checked and have been shown to be valid for our concluding remarks.

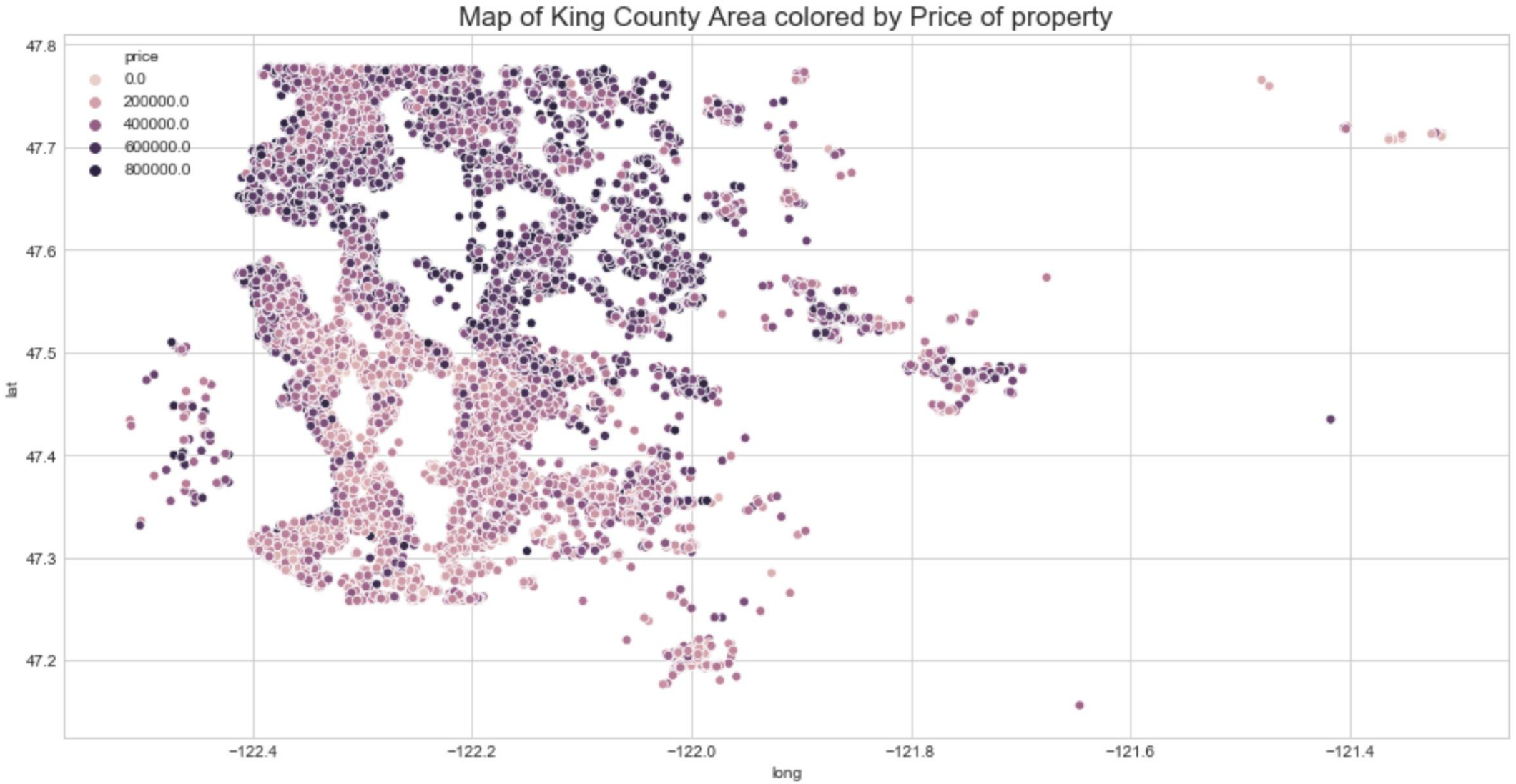
Below: The devil is not in the detail... The below shows one of the assumptions that was checked for our data, **linearity.**





WHAT TRENDS?... WHERE YOU LIVE

The first clear trend that we saw is that for King County, **where you live affects house price.**



We found that there was no relationship between longitude and house prices ***but*** a linear relationship between latitude and increased house price.

The further north you live, the higher the estimated sale price of your home.

WHAT TRENDS? .. GRADE OF THE PROPERTY

As can be seen in the graph to the left, we found that there was a *strong correlation* between the grade of property (as per the King County grading system) and the price that the property sold for.

THE PREDICTIVE MODEL

Overall, the model below shows how the *grade, latitude and sqft of the property* affect the overall sale price of a property in King County.

This model is applicable for house prices up to $665,000 predicted sale price.

The average error of this model, when predicting future sale prices, is ±$87,900.

It can be see that all 3 major variables have a positive relationship with sale price.

House Price = 71***(sqft_living)*** + 480,347***(latitude)*** + 43,779***(Grade)*** -22,884,972

Example: A1,000sqft, Grade 7 home at a latitude of 47.7 is estimated to sell for $405,033.

APPENDIX 1: COLUMN DESCRIPTION

Column Names and descriptions for Kings County Data Set

- **id** - unique identified for a house
- **dateDate** - house was sold
- **pricePrice** - is prediction target
- **bedroomsNumber** - of Bedrooms/House
- **bathroomsNumber** - of bathrooms/bedrooms
- **sqft_livingsquare** - footage of the home
- **sqft_lotsquare** - footage of the lot
- **floorsTotal** - floors (levels) in house
- **waterfront** - House which has a view to a waterfront
- **view** - Has been viewed
- **condition** - How good the condition is (Overall)
- **grade** - overall grade given to the housing unit, based on King County grading system
- **sqft_above** - square footage of house apart from basement
- **sqft_basement** - square footage of the basement
- **yr_built** - Built Year
- **yr_renovated** - Year when house was renovated
- **zipcode** - zip
- **lat** - Latitude coordinate
- **long** - Longitude coordinate
- **sqft_living15** - The square footage of interior housing living space for the nearest 15 neighbors
- **sqft_lot15** - The square footage of the land lots of the nearest 15 neighbors