Phase 2 Project
Housing Regression Analysis

Presented by: Carla Kirby

Overview

In this linear regression modelling project, we take a deeper dive into housing data for King County a region of Seattle Washington, in the United States. The business owners are eager to understand what favourably drives home purchasing in the King County area.

Agenda Discussion Points

Using an OSEMN analysis model (Obtain, Scrub, Explore, Model, Interpret)

Business Problem

Data

Process Methods

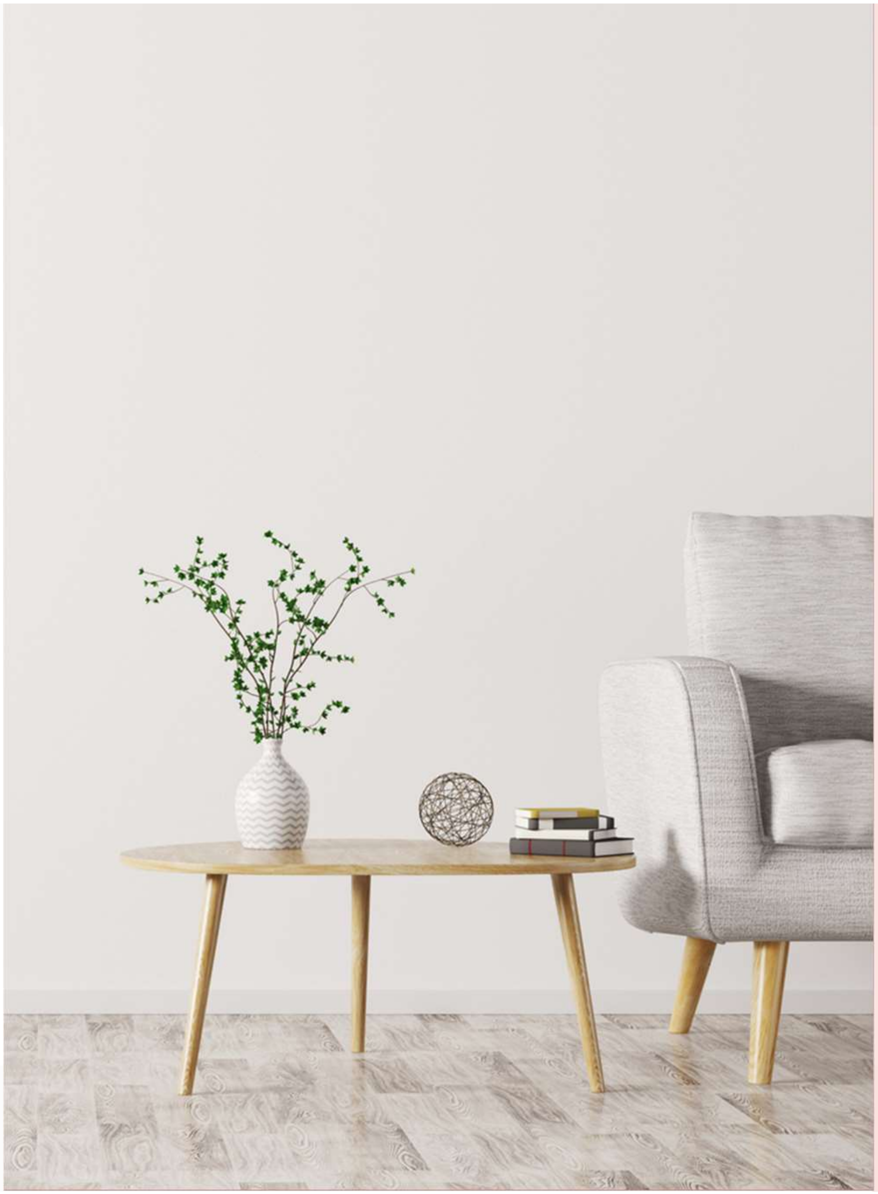
Results

Conclusion

Business Problem

King County Real Estate would like to build an effective strategy to start up a new real estate business in the well known and established community of king county.

The stakeholders need to understand what drives purchases of homes in this area so they can cater to the community and ensure that they market effectively for their new business.



Data - Obtain & Scrub

- .CSV kc_housing_data – sourced from the business owner
- Removed null values, duplicates, unnecessary columns, excessive outliers, changed the data types
- Wrangled exponential data by creating lambda functions
- Inspected data for strong correlations to price, created samples & tested those samples
- Obtained necessary statistical values for analysis

Process Methods - Explore

- Predicted Correlations
- Mean Standard Error - MSE
- Kernal Density Estimate - KDE
- Plotted and graphically represented Outliers for further analysis
- Explored K-folds and cross validation tests
- Types of graphical interpretations used to Model: Bell Curve, box plot, multi-scatter plots, histograms

Graphical Data Interpretation - Modelling

Kernal Density Plot

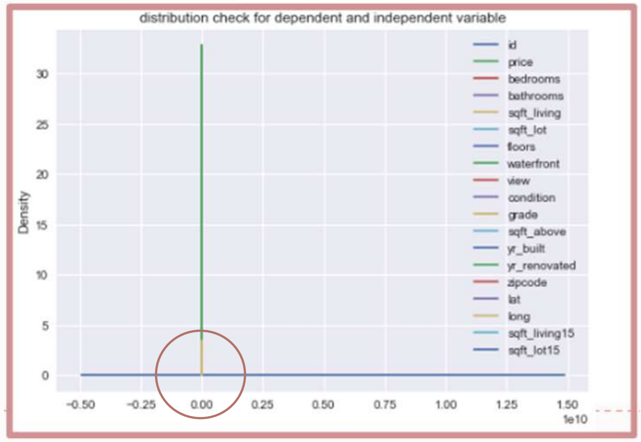
- Non-parametric way to estimate the probability density function of a random variable
- Determine that **the sqft_living** and **grade** are closest to the expected value 0.00

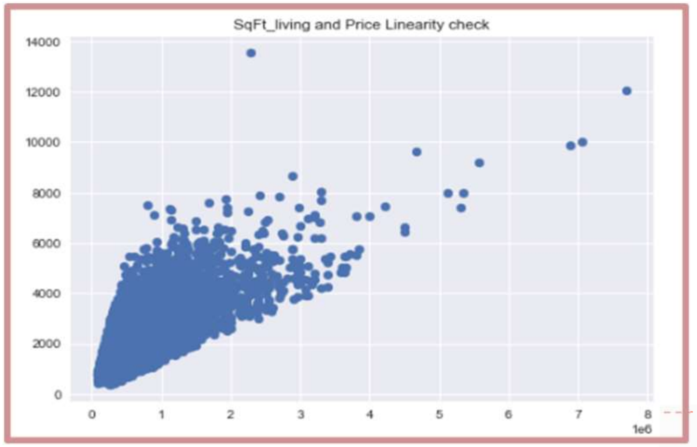
Linearity Check

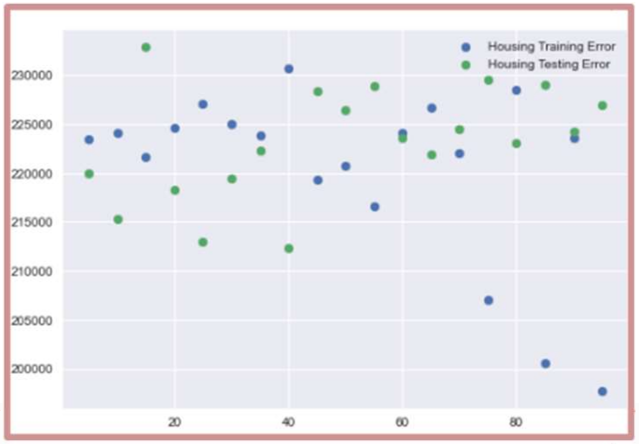
- Strongest Linearity shown was with the sqft_living at 70%– despite the strength of the sqft_lot correlation of 89%

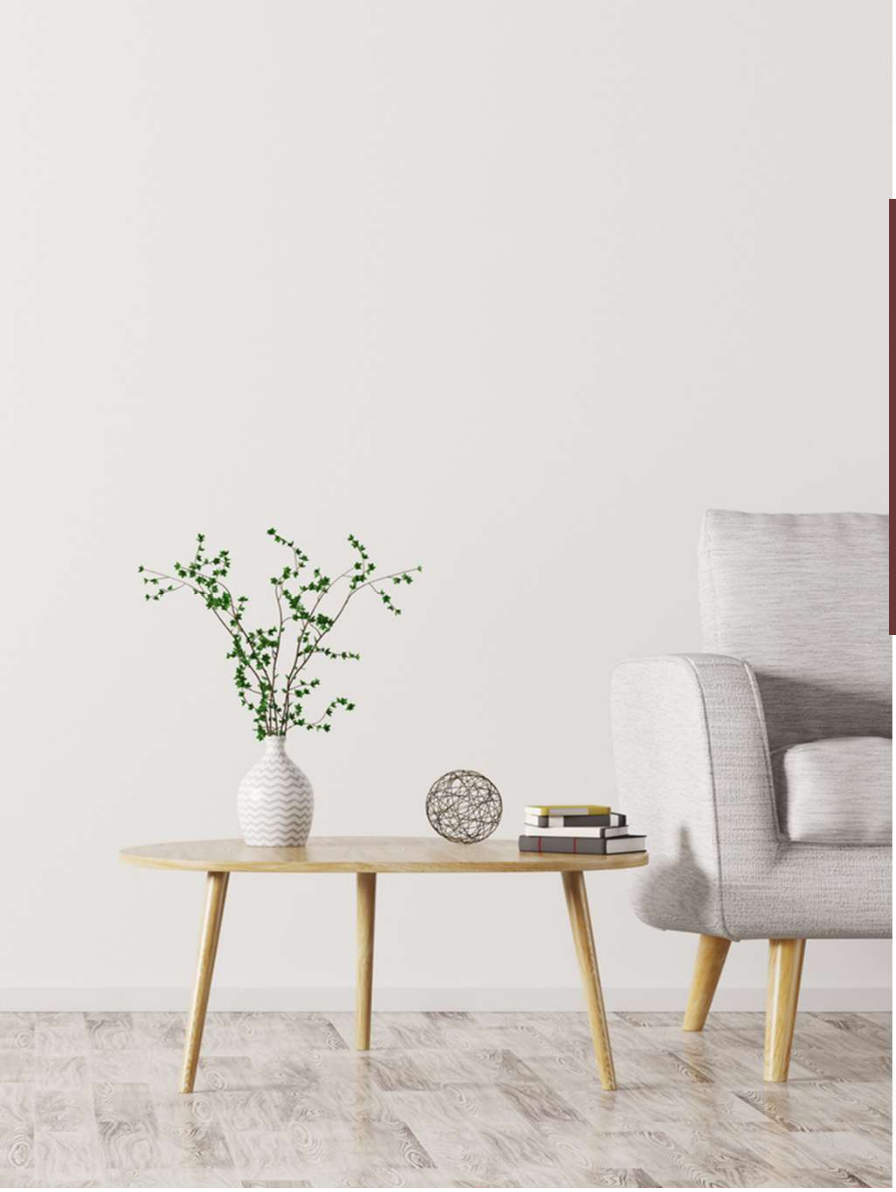
Training / Testing Error Variances

- Plot of the residuals appears to be fitted between both data sets with a level of homoscedasticity present.

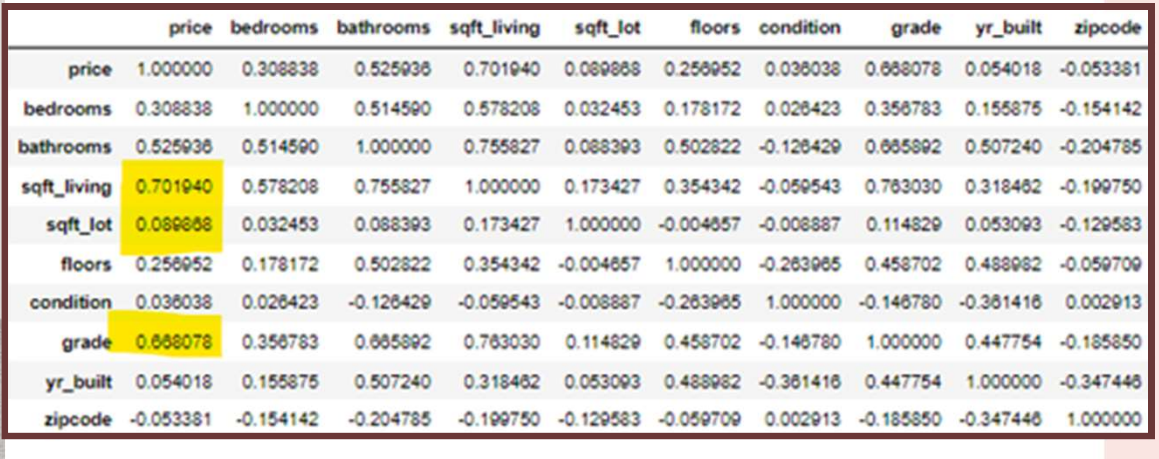








Strongest Linearity shown was with the sqft_living at 70%–
despite the strength of the sqft_lot correlation of 89%



	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	condition	grade	yr_built	zipcode
price	1.000000	0.308838	0.525936	0.701940	0.089868	0.256952	0.036038	0.668078	0.054018	-0.053381
bedrooms	0.308838	1.000000	0.514590	0.578208	0.032453	0.178172	0.026423	0.356783	0.155875	-0.154142
bathrooms	0.525936	0.514590	1.000000	0.755827	0.088393	0.502822	-0.126429	0.665892	0.507240	-0.204785
sqft_living	0.701940	0.578208	0.755827	1.000000	0.173427	0.354342	-0.059543	0.763030	0.318462	-0.199750
sqft_lot	0.089868	0.032453	0.088393	0.173427	1.000000	-0.004657	-0.008887	0.114829	0.053093	-0.129583
floors	0.256952	0.178172	0.502822	0.354342	-0.004657	1.000000	-0.263965	0.458702	0.488982	-0.059709
condition	0.036038	0.026423	-0.126429	-0.059543	-0.008887	-0.263965	1.000000	-0.146780	-0.361416	0.002913
grade	0.668078	0.356783	0.665892	0.763030	0.114829	0.458702	-0.146780	1.000000	0.447754	-0.185850
yr_built	0.054018	0.155875	0.507240	0.318462	0.053093	0.488982	-0.361416	0.447754	1.000000	-0.347446
zipcode	-0.053381	-0.154142	-0.204785	-0.199750	-0.129583	-0.059709	0.002913	-0.185850	-0.347446	1.000000

Results

Based on the data set and through rigorous testing measures we have determined that ft^2 of the living area has the strongest linear correlation to increased sales of housing price.

Conclusion – Interpretation

- R squared values were between 0 and 1 for both the sqft_lot and the sqft_living – indicating strong correlation to increased sale prices.
- Null hypothesis was validated, confirmed that sqft_living has a strong connection to the prices. Elevated size of sqft-Living meant higher cost of the home
- Sqft_living presented with the best line of fit when compared to grade and sqft_lot, irrespective of the fact that sqft_lot had a higher correlation value.