King County Real Estate Development Solutions

Solomon Rapoport

TOC

Overview

This project analyzes real estate in King County, Washington, using data-driven solutions from a dataset of 30,155 homes to provide concrete business suggestions for a real estate development company in KC.

Understanding the Business Problem

What is Real Estate Development?

Real estate development is the process of improving real property to increase its value. RE developers acquire property and develop it into commercial or residential buildings, in the hopes of turning a considerable profit.

Why do some developers make money while others lose?

Many factors contribute to the resulting P/L of a real estate investment, including location, environment, economics, accounting, aesthetics, among others.

How to address these issues in KC?

RE developers in KC might be able to utilize data-driven solutions to gain better insight in regards to factors that impact the price of a home, be those more obvious factors (e.g. location, size), or more subtle ones (e.g. number of bathrooms, patio size), and use these insights to improve their sales and profit margins.

Data Understanding

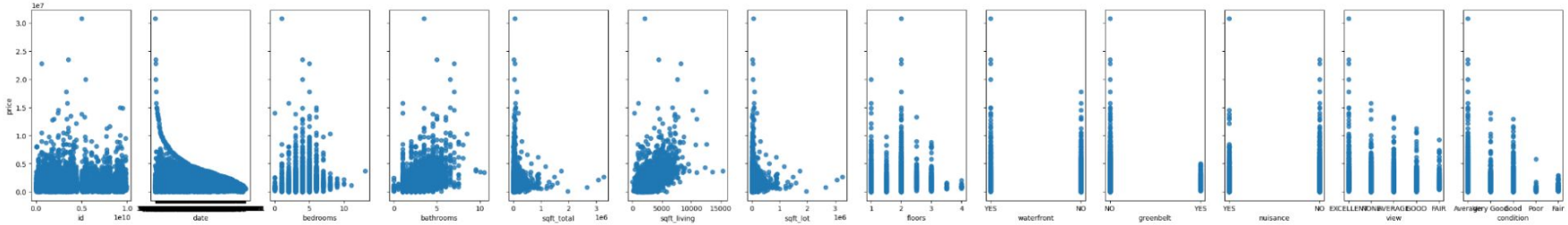
The King County Dataset

- Comprises data on over 30,000 homes in KC, purchased between 2021 and 2022
- Includes information on each home such as:
  - Year built (ranges from 1900 to 2022)
  - Location
  - Price Paid
  - Square footage
  - Year last renovated
  - Number of bedrooms/bathrooms

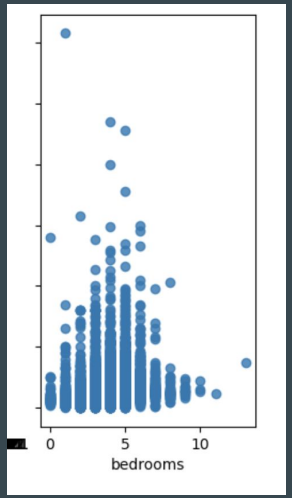
- How can we gain data driven insight on this dataset, which will provide solutions for increasing profits when it comes to real estate development?
- Initial investigation into the correlation between all variables and price show that square foot living space has the highest positive correlation with price, but how could we use further modeling and data analysis to yield concrete business suggestions?

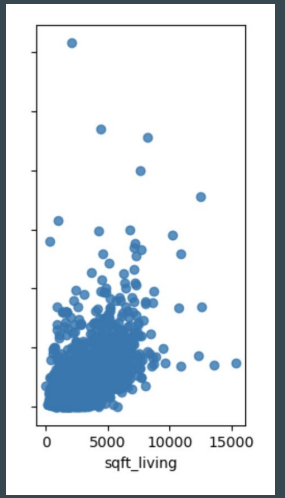
Basic Analysis of Dataset

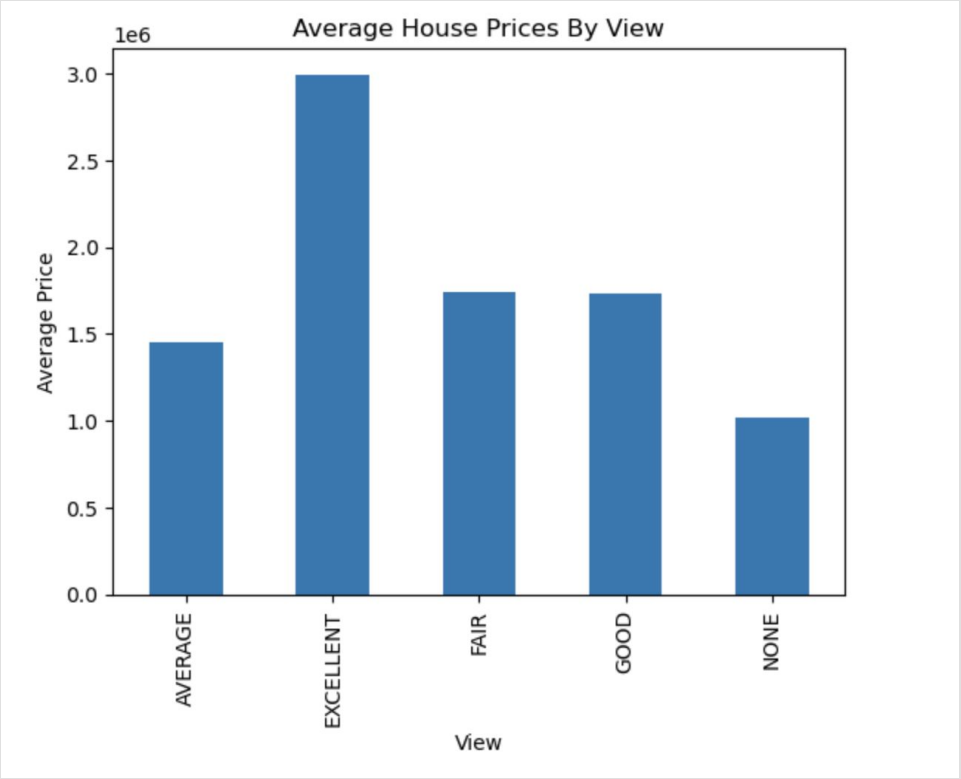
- Below, scatter plots produced comparing each variable with price



- Only square foot living and bathrooms appear to have a somewhat clear visual linear correlation, while others do not. Condition and view might have a somewhat linear relationship, but it is difficult to tell, given their categorical relationship.
- 'Bedrooms' is an interesting variable, as it appears to be normally distributed.
  - Could indicate diminishing marginal returns as the number of bedrooms in a house exceeds 6





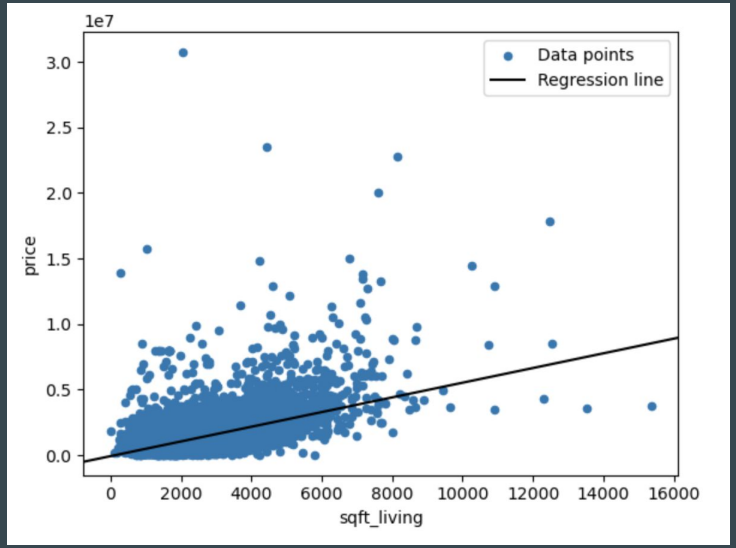


- Initial analysis seems to indicate the clear impact a view (or more specifically, the lack thereof) has on the sale price of a home.
- The jump from no view to an average view alone comes with an average price difference of +$433,386, a massive %42.5 increase!
    - Homes with an “excellent” view are sold at an average of %200 *more* than homes with no view (nearly $2 million more!)

Data Analysis Using Linear Regression Modeling

Baseline Model: Simple Linear Regression

- Target/dependent variable was price, and used square foot living space as our baseline independent variable, given that it had the strongest correlation with price.
- Model overall was statistically significant, as well as each of the coefficients.
- R-squared of 0.37 means that 37% of the variance in price is explained by our independent variable (square feet of living space)
- Independent variable coefficient of 560 implies that every additional square foot of living space correlates to an additional $560 in property sale price



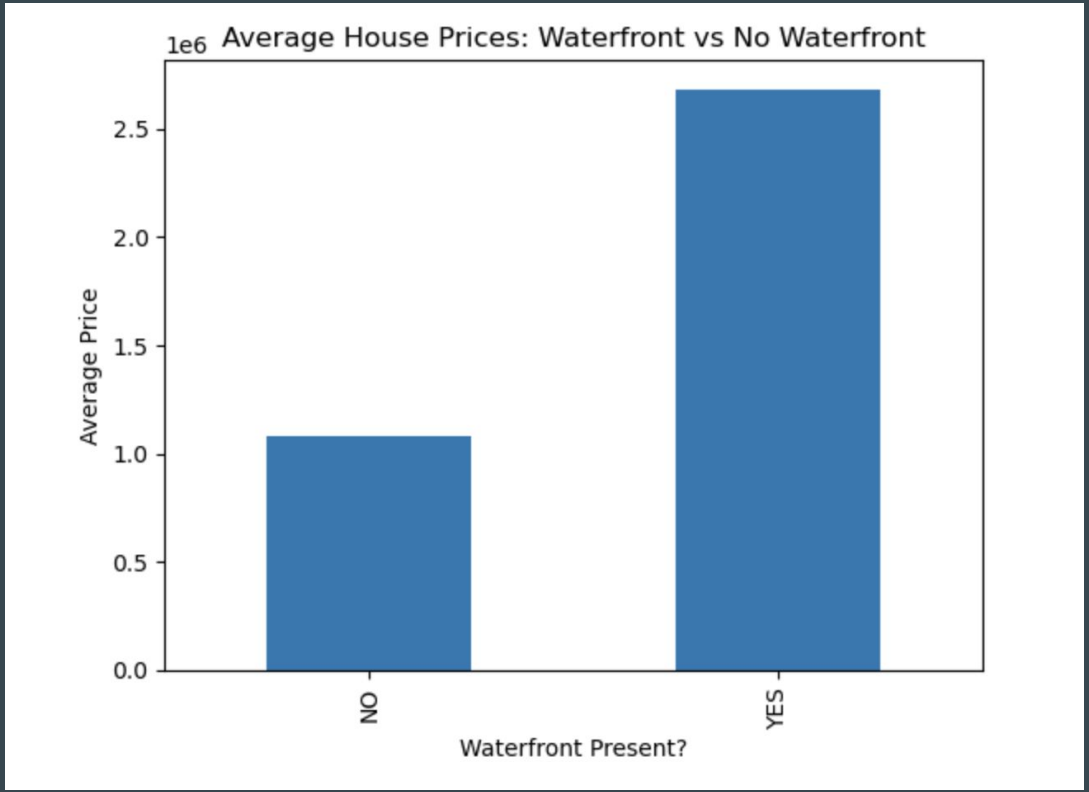
Dep. Variable:	price	**R-squared:**	0.370
Model:	OLS	**Adj. R-squared:**	0.370
Method:	Least Squares	**F-statistic:**	1.773e+04
Date:	Fri, 16 Jun 2023	**Prob (F-statistic):**	0.00
Time:	18:41:58	**Log-Likelihood:**	-4.4912e+05
No. Observations:	30155	**AIC:**	8.982e+05
Df Residuals:	30153	**BIC:**	8.983e+05
Df Model:	1		
Covariance Type:	nonrobust		

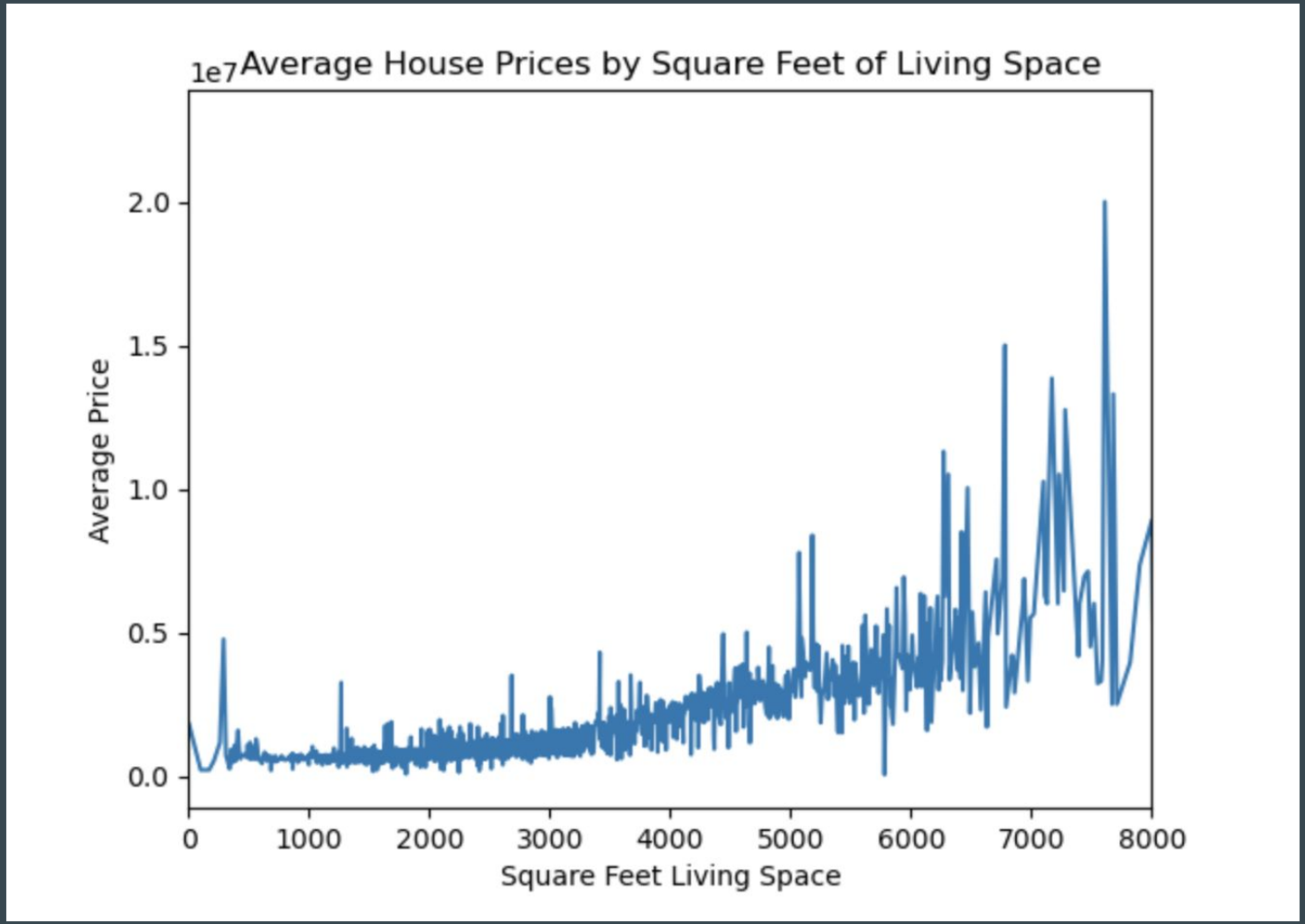
	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.443e+04	9782.728	-7.609	0.000	-9.36e+04	-5.53e+04
sqft_living	560.0050	4.206	133.160	0.000	551.762	568.248

Model 2: Multiple Linear Regression

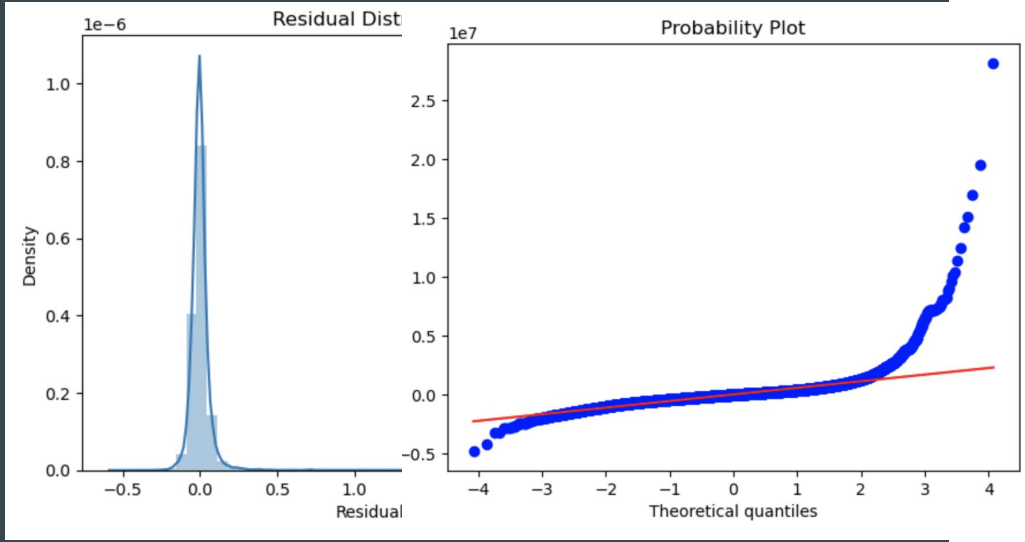
- Added other predictor variables from the dataset. Results:
  - Model is an improvement from baseline, with a **higher R-squared** of 0.429. The model is statistically significant as are the intercept and coefficients.
- Some observations:
  - Does not appear to be any negative correlation between how old a home is and the sale price.
  - Square feet of living space has the highest correlation with price among the various square feet measurements, with a correlation of $442.26 increase in home sale price for every additional square foot of living space
  - The “floors” variable has a high coefficient, as each additional floor is correlated with an increase of $84,840 in sale price
  - Most strikingly, the presence of a waterfront is correlated with an increase of $1,136,000 in value versus a home that is not on some sort of waterfront!
    - Looking back to our raw data, the price of a waterfront home is about 1.755 standard deviations higher than the average home price.

Waterfront Homes: A $1,600,700 difference!



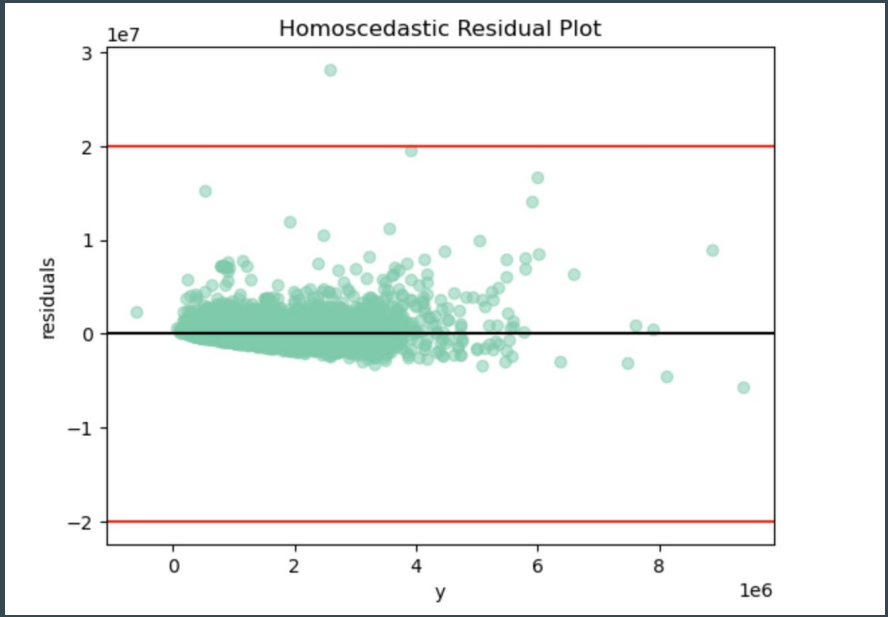

Average House Prices by Square Feet of Living Space
1e7
2.0
1.5
1.0
0.5
0.0
Average Price
0
1000
2000
3000
4000
5000
6000
7000
8000
Square Feet Living Space

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.31e+05	2.28e+04	-5.748	0.000	-1.76e+05	-8.64e+04
bedrooms	-1.297e+05	5245.715	-24.721	0.000	-1.4e+05	-1.19e+05
sqft_living	442.2601	22.288	19.843	0.000	398.574	485.946
sqft_lot	-86.4541	9.948	-8.691	0.000	-105.952	-66.956
floors	8.484e+04	8487.855	9.996	0.000	6.82e+04	1.01e+05
years_old	3161.1364	148.150	21.337	0.000	2870.755	3451.517
sqft_patio	125.2965	20.591	6.085	0.000	84.937	165.656
sqft_total	86.2039	9.947	8.666	0.000	66.707	105.701
waterfront_YES	1.136e+06	3.06e+04	37.134	0.000	1.08e+06	1.2e+06



- Relevant data from the subset model summary, including coefficient, standard error, and p-values

- Normally distributed residuals of the subset model



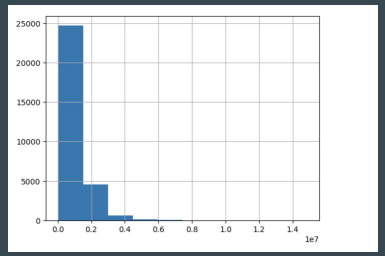

```
from statsmodels.stats.diagnostic import het_goldfeldquandt
het_goldfeldquandt(y_2, X_2, alternative='two-sided')
```

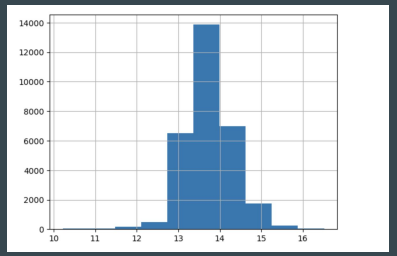
```
(0.03106419716543548, 0.0,
```

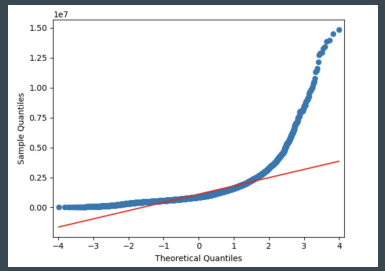
- Scatterplot visualizing the residuals, to examine if the errors are homoscedastic. Looks like it could be homoscedastic, but let's look at a statistical test.
- Since the null hypothesis is homoscedasticity, the G-Q test result of 0.0 rejects the null hypothesis, meaning the residuals are heteroscedastic.
  - How else can we improve the model?

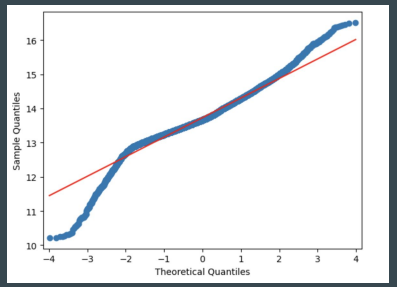
Model 3: Log-Transformed Y variable/Target/Price

- It appears that price follows a logarithmic distribution, and taking the logarithm of our target variable price yields what appears to be a normal distribution of the logarithmically transformed price.
  - Pre vs. post log-transformed price distribution, visualized with histograms and QQ plots
- Results
  - Square foot living coefficient of 0.0002 means that each increase of 1 square foot is correlated with a 0.02% increase in sale price
  - Floors coefficient of 0.0339 means that each increase of one floor is correlated with a 3.39% increase in sale price









Log-Transformed Target Variable Model Summary

Dep. Variable:	log(price)	R-squared:	0.404
Model:	OLS	Adj. R-squared:	0.404
Method:	Least Squares	F-statistic:	1201.
Date:	Thu, 22 Jun 2023	Prob (F-statistic):	0.00
Time:	18:29:08	Log-Likelihood:	-18711.
No. Observations:	30155	AIC:	3.746e+04
Df Residuals:	30137	BIC:	3.761e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.9477	0.029	240.925	0.000	6.891	7.004
bedrooms	-0.0139	0.016	-0.888	0.374	-0.045	0.017
bedrooms_transformed	-0.0749	0.055	-1.353	0.176	-0.183	0.034
sqft_living	0.0002	1.35e-05	15.424	0.000	0.000	0.000
sqft_total	8.207e-05	5.79e-06	14.176	0.000	7.07e-05	9.34e-05
sqft_lot	-8.204e-05	5.79e-06	-14.172	0.000	-9.34e-05	-7.07e-05
sqft_total_transformed	154.6137	124.273	1.244	0.213	-88.968	398.195
sqft_lot_transformed	52.3956	25.035	2.093	0.036	3.326	101.465
floors	0.0339	0.006	5.476	0.000	0.022	0.046
condition_Average	1.3761	0.013	105.572	0.000	1.351	1.402
condition_Fair	1.4093	0.027	52.263	0.000	1.356	1.462
condition_Good	1.3752	0.013	102.916	0.000	1.349	1.401
condition_Poor	1.4173	0.047	30.442	0.000	1.326	1.509
condition_Very Good	1.3697	0.014	96.204	0.000	1.342	1.398
view_AVERAGE	1.3829	0.012	111.738	0.000	1.359	1.407
view_EXCELLENT	1.3779	0.018	77.103	0.000	1.343	1.413
view_FAIR	1.4050	0.026	54.045	0.000	1.354	1.456
view_GOOD	1.4047	0.015	91.809	0.000	1.375	1.435
view_NONE	1.3771	0.009	145.102	0.000	1.359	1.396
NO	3.2868	0.017	195.647	0.000	3.254	3.320
YES	3.6609	0.018	198.629	0.000	3.625	3.697

Conclusions/Recommendations

1. **House with a view, any view:** Merely the presence of any sort of view is correlated with a massive price increase in average sale price compared with homes that have no view. Develop properties with favorable views (be it property location, window placement, etc.

2. **Square feet of living space is paramount**. Our models showed that square feet of living space is correlated with a higher price per square foot than any other part of the property. So in development, skip the patios, big garage, large basement, etc.

3. **Waterfront properties are a gold mine.** The basic data, confirmed by one of the models produced, show that the average home price that is a 'waterfront property' is over $1.5 million more than homes with no waterfront! Prioritizing the development of waterfront properties could bump up sale price significantly.

Next Steps

- What further types of data could we use to provide better business suggestions to our real estate development company?
  - **Development costs:** Of course, it would be great for sale price to develop properties with waterfronts, nice views, and big living spaces, but what are the costs in King County of such properties/modifications? At what point do the costs outweigh the added revenue?
  - Data which breaks down average property prices per neighborhood/town
  - Other independent variables, which may or may not impact price, but would be interesting to take a look at, such as:
    - Exterior paint color/design
    - Presence of home amenities/luxuries (e.g. pool, gym, sports court)
    - Local crime rates
    - Distance from commercial hubs (e.g. Seattle)
    - Property taxes

Thank You

Solomon Rapoport

-Questions?

-LinkedIn: https://www.linkedin.com/in/solomon-rapoport-3939a9136/