KING COUNTY HOUSING MODEL

BY JUSTIN WEIBLE AND NAOMI WEINBERGER

AN ANALYSIS ON FACTORS THAT CONTRIBUE TO HOUSING PRICES

FOR PROSPECTIVE BUYERS IN THE KING COUNTY AREA

DATASET: KAGGLE

21,597 HOUSES IN THE SEATTLE AREA WERE ANALYZED

PARAMETERS:

- UNDER ONE MILLION DOLLARS
- ONE TO SIX BEDROOMS
- ONE TO FOUR BATHROOMS
- LESS THAN 3.5 FLOORS
- LESS THAN 5,000 SQUARE FEET

Key Factors in Price Determination

PROXIMITY TO DOWNTOWN

The closer a house is to downtown, the higher a price will be. Each mile outside the city reduces the price by about $10,400

YEAR BUILT

The age of the house decreases the value by approximately 1,172 per year

NUMBER OF BEDROOMS

Each bedroom adds an approximate value of 15,580

NUMBER OF BATHROOMS

Each bathroom adds an approximate value of 15,660

SQUARE FOOTAGE

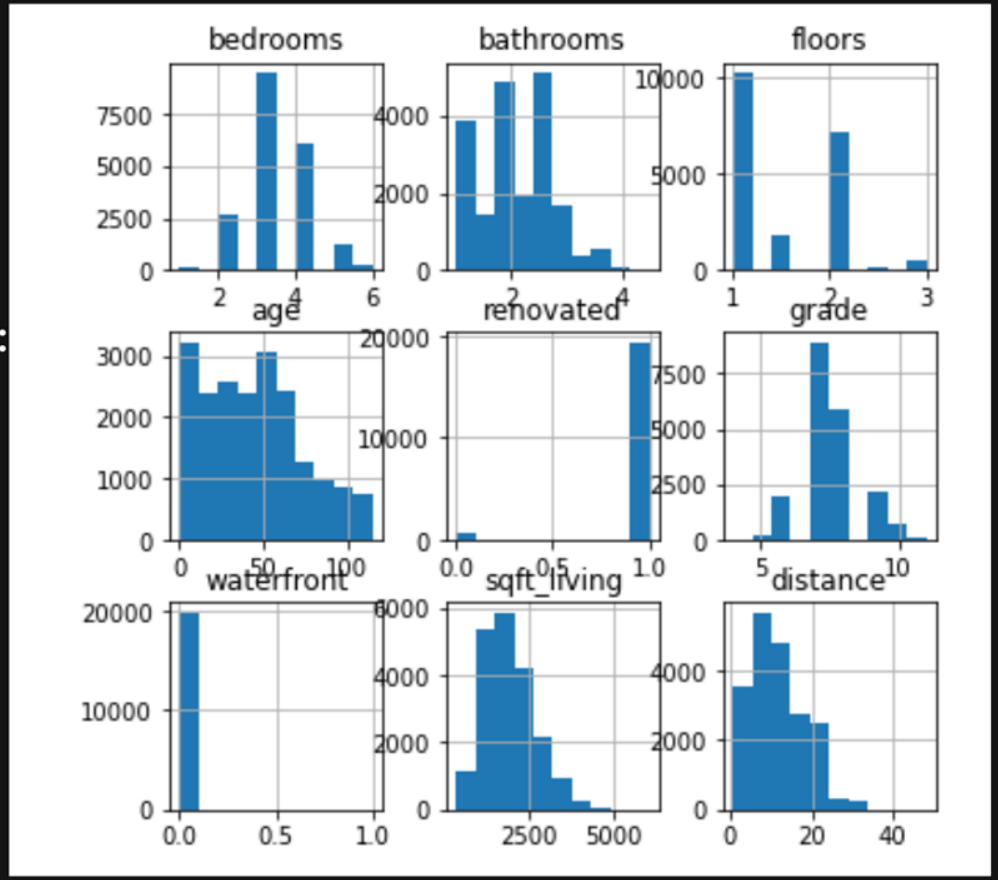
Each square foot adds an approximate value of $115.84

GRADE

King County's assigned grades on the houses are good indicators as to what the final price will be

EDA

WHEN LOOKING AT OUR DATA SET, WE NARROWED DOWN OUR FEATURES TO INCLUDE: SQUAREFEET, NUMBER OF BEDROOMS, NUMBER OF BATHROOMS,NUMBER OF FLOORS,GRADE,CONDITION,AGE OF THE HOUSE, WATERFRONT, DISTANCE FROM DOWNTOWN, AND WHETHER OR NOT IT WAS RENOVATED



Models

```
Test RMSE: 185212.7798740588
Train RMSE: 183591.80606517757
Training Score: 0.69
Test Score: 0.68
Coefficients: [ 1.80491024e+02 -3.43265701e+04  1.92287809e+04 -5.58838958e+03
  1.10775024e+05  3.11113403e+04 -6.68291343e+04  7.19920196e+05
 -1.40294460e+04  1.55210948e+03]
```

OLS Regression Results

Dep. Variable:	price	R-squared:	0.692
Model:	OLS	Adj. R-squared:	0.692
Method:	Least Squares	F-statistic:	3831.
Date:	Thu, 08 Jul 2021	Prob (F-statistic):	0.00
Time:	21:14:19	Log-Likelihood:	-2.3060e+05
No. Observations:	17032	AIC:	4.612e+05
Df Residuals:	17021	BIC:	4.613e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.58e+05	2.03e+04	-27.496	0.000	-5.98e+05	-5.18e+05
sqft_living	180.4910	3.223	56.000	0.000	174.173	186.809
bedrooms	-3.433e+04	2101.895	-16.331	0.000	-3.84e+04	-3.02e+04
bathrooms	1.923e+04	3405.375	5.647	0.000	1.26e+04	2.59e+04
floors	-5588.3896	3328.424	-1.679	0.093	-1.21e+04	935.666
grade	1.108e+05	2103.795	52.655	0.000	1.07e+05	1.15e+05
condition	3.111e+04	2390.025	13.017	0.000	2.64e+04	3.58e+04
renovated	-6.683e+04	8165.683	-8.184	0.000	-8.28e+04	-5.08e+04
waterfront	7.199e+05	1.82e+04	39.578	0.000	6.84e+05	7.56e+05
distance	-1.403e+04	245.235	-57.208	0.000	-1.45e+04	-1.35e+04
age	1552.1095	75.033	20.686	0.000	1405.037	1699.182

Omnibus:	10975.804	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	442833.076
Skew:	2.532	Prob(JB):	0.00
Kurtosis:	27.462	Cond. No.	3.28e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 3.28e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

```
Test RMSE: 118646.0026142218
Train RMSE: 117197.17757928869
Training Score: 0.64
Test Score: 0.63
Coefficients: [ 1.15992431e+02 -1.53245118e+04  1.43793254e+04  1.32901955e+04
  8.11923233e+04  2.37732207e+04 -2.52287317e+04  1.93059501e+05
 -1.04491442e+04  1.11060368e+03]
```

OLS Regression Results

Dep. Variable:	price	R-squared:	0.637
Model:	OLS	Adj. R-squared:	0.637
Method:	Least Squares	F-statistic:	2790.
Date:	Thu, 08 Jul 2021	Prob (F-statistic):	0.00
Time:	21:52:06	Log-Likelihood:	-2.0797e+05
No. Observations:	15887	AIC:	4.160e+05
Df Residuals:	15876	BIC:	4.160e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.496e+05	1.4e+04	-24.897	0.000	-3.77e+05	-3.22e+05
sqft_living	115.9924	2.325	49.880	0.000	111.434	120.551
bedrooms	-1.532e+04	1420.464	-10.788	0.000	-1.81e+04	-1.25e+04
bathrooms	1.438e+04	2345.644	6.130	0.000	9781.598	1.9e+04
floors	1.329e+04	2220.293	5.986	0.000	8938.169	1.76e+04
grade	8.119e+04	1445.622	56.164	0.000	7.84e+04	8.4e+04
condition	2.377e+04	1587.591	14.974	0.000	2.07e+04	2.69e+04
renovated	-2.523e+04	5731.320	-4.402	0.000	-3.65e+04	-1.4e+04
waterfront	1.931e+05	1.96e+04	9.847	0.000	1.55e+05	2.31e+05
distance	-1.045e+04	160.349	-65.165	0.000	-1.08e+04	-1.01e+04
age	1110.6037	51.042	21.758	0.000	1010.555	1210.653

Omnibus:	434.620	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	525.304
Skew:	0.352	Prob(JB):	8.55e-115
Kurtosis:	3.545	Cond. No.	4.38e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 4.38e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

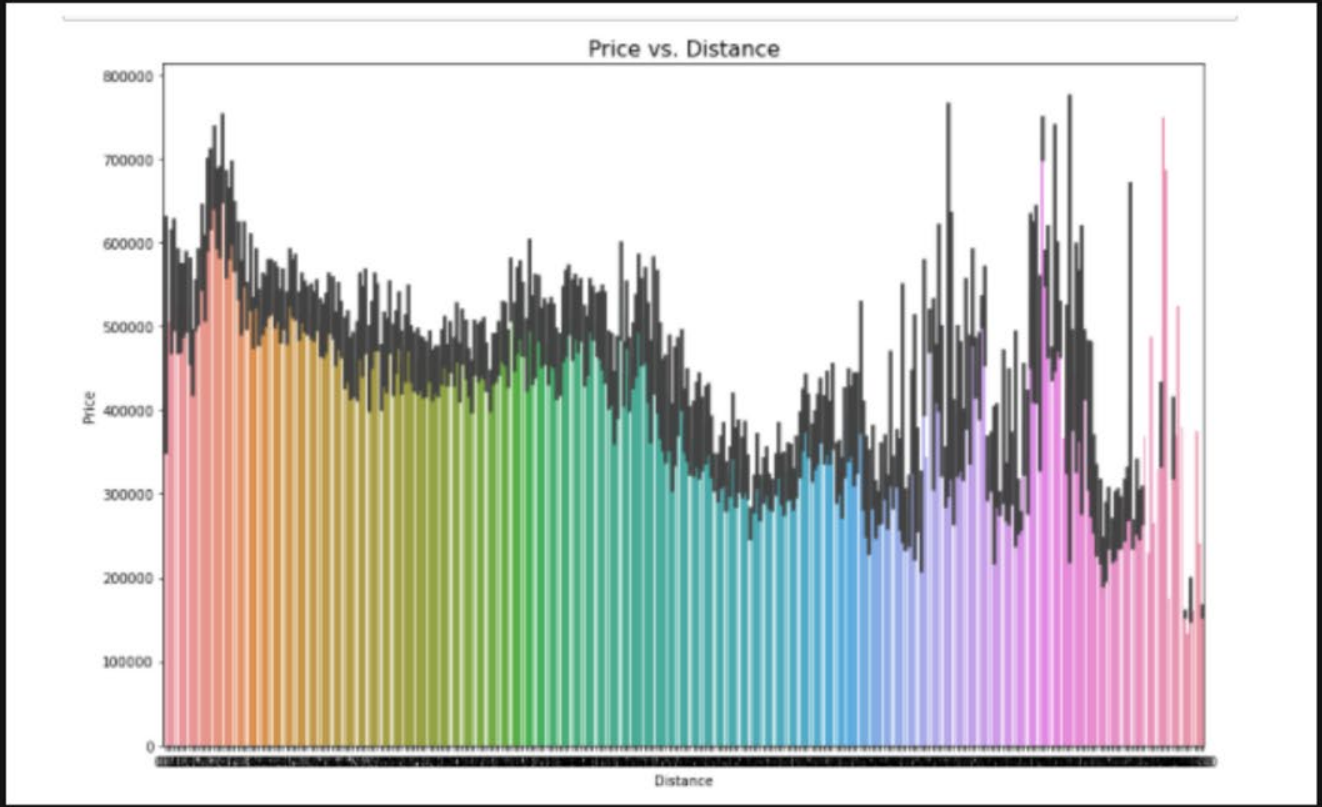
MODEL EXPLAINATION

Our initial model looked at all of the columns from our EDA.

As we refined the model, we only needed to remove significant outliers in price, square footage of living space, bathrooms, and floors

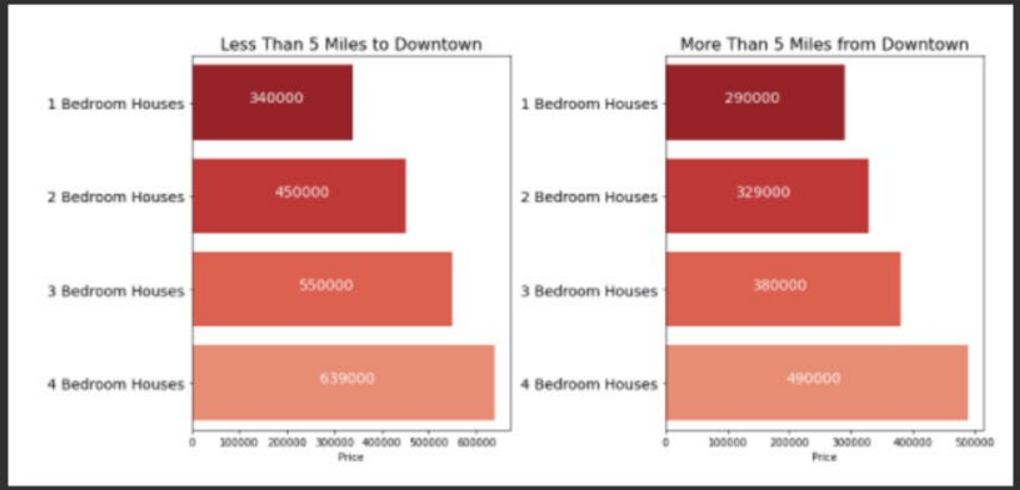
Our final model is able to predict 63.7% of the variance in housing prices with an error of $118,646.

Results


Price vs. Distance
800000
700000
600000
500000
400000
300000
200000
100000
0
Price
Distance

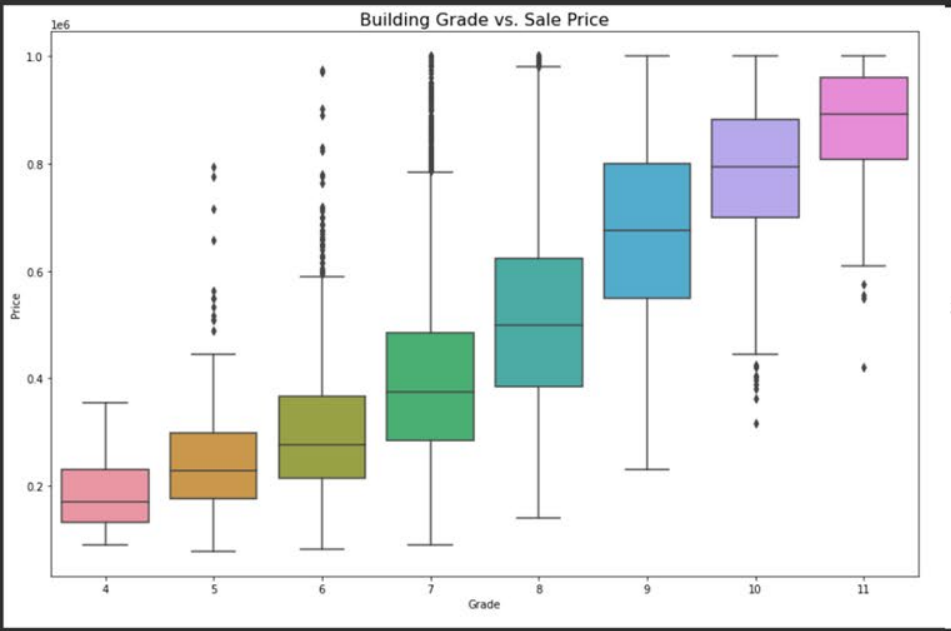
→

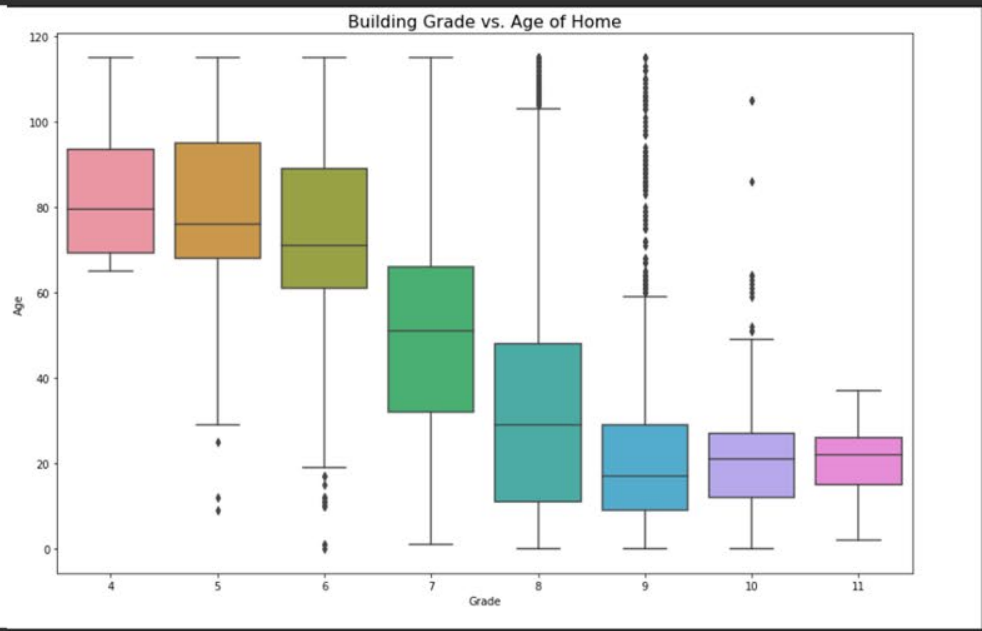
Results


Less Than 5 Miles to Downtown
1 Bedroom Houses
340000
2 Bedroom Houses
450000
3 Bedroom Houses
550000
4 Bedroom Houses
639000
0
100000
200000
300000
400000
500000
600000
Price
More Than 5 Miles from Downtown
1 Bedroom Houses
290000
2 Bedroom Houses
329000
3 Bedroom Houses
380000
4 Bedroom Houses
490000
0
100000
200000
300000
400000
500000
Price


Less Than 5 Miles to Downtown
Under 1250 Square Feet
402000
Under 2500 Square Feet
563000
Under 3250 Square Feet
775000
Under 5000 Square Feet
760000
0
200000
400000
600000
800000
Price
More Than 5 Miles from Downtown
Under 1250 Square Feet
278000
Under 2500 Square Feet
392000
Under 3250 Square Feet
625000
Under 5000 Square Feet
812500
0
200000
400000
600000
800000
Square Feet

Results


Building Grade vs. Sale Price
1e6
1.0
0.8
0.6
0.4
0.2
Price
4
5
6
7
8
9
10
11
Grade


Building Grade vs. Age of Home
120
100
80
60
40
20
0
Age
4
5
6
7
8
9
10
11
Grade

Next Steps:

We would want to analyze the information using other models because linear regression doesn't seem to be the best method of predicting houses.

Thank you for your time.
Please reach out for more information

Contact details:
Justin Weible
jweible23@gmail.com
https://github.com/justinweible

Naomi Weinberger
weinberger.naomi@gmail.com
https://github.com/Naomiweinberger