

KING



SALES

BOUNTY



King County

BUSINESS PROBLEM

- The stakeholders in King County who are the homeowners are looking for insights on how to increase the estimated value of their houses.

BUSINESS OBJECTIVE



King County

Using the King County Sales dataset we looked to address the following objectives;

1. How the house design in respect to number of bedrooms, number of bathrooms and number of floors has influence on the sale value of the houses in King County?
2. How the dimensions of the house and lot size have on price of houses in King County?
3. How the combination of the numeric variable with the highest correlation and viable categorical variable influences the price of houses in King county?
4. How the location of the houses influences the price of houses in King county?

DATA AND METHODS



King County

We used King County dataset which contained sales from 2014-2015 of 21,597 homes to carry out our analysis.

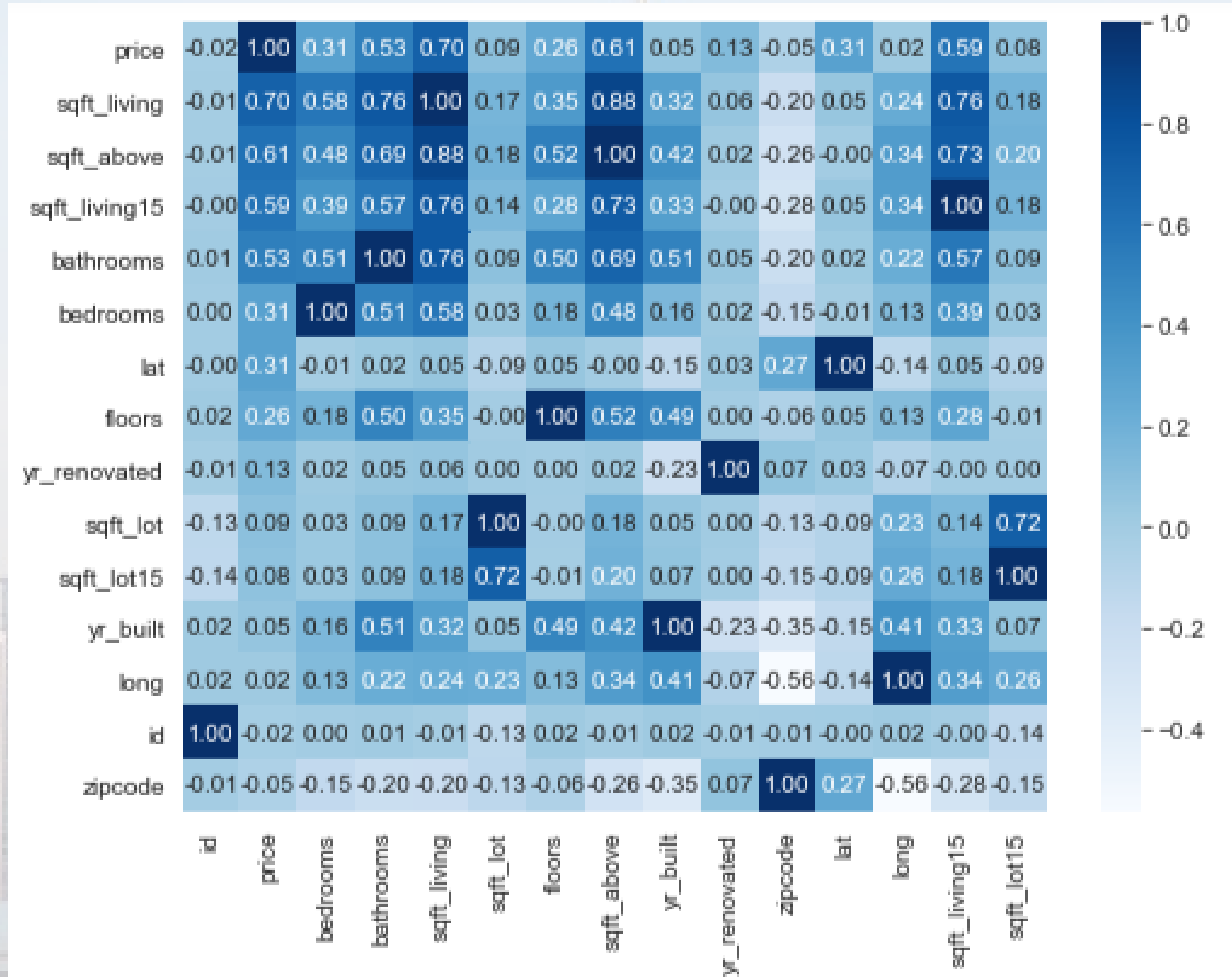
Our methods to analyse the data include;

1. Data cleaning
2. Exploratory data analysis (EDA)
3. Normality testing
4. Hypothesis testing
5. Modeling - simple and multiple linear regression
6. Ridge and Lasso regularization techniques

RELATIONSHIP OF ALL VARIABLES



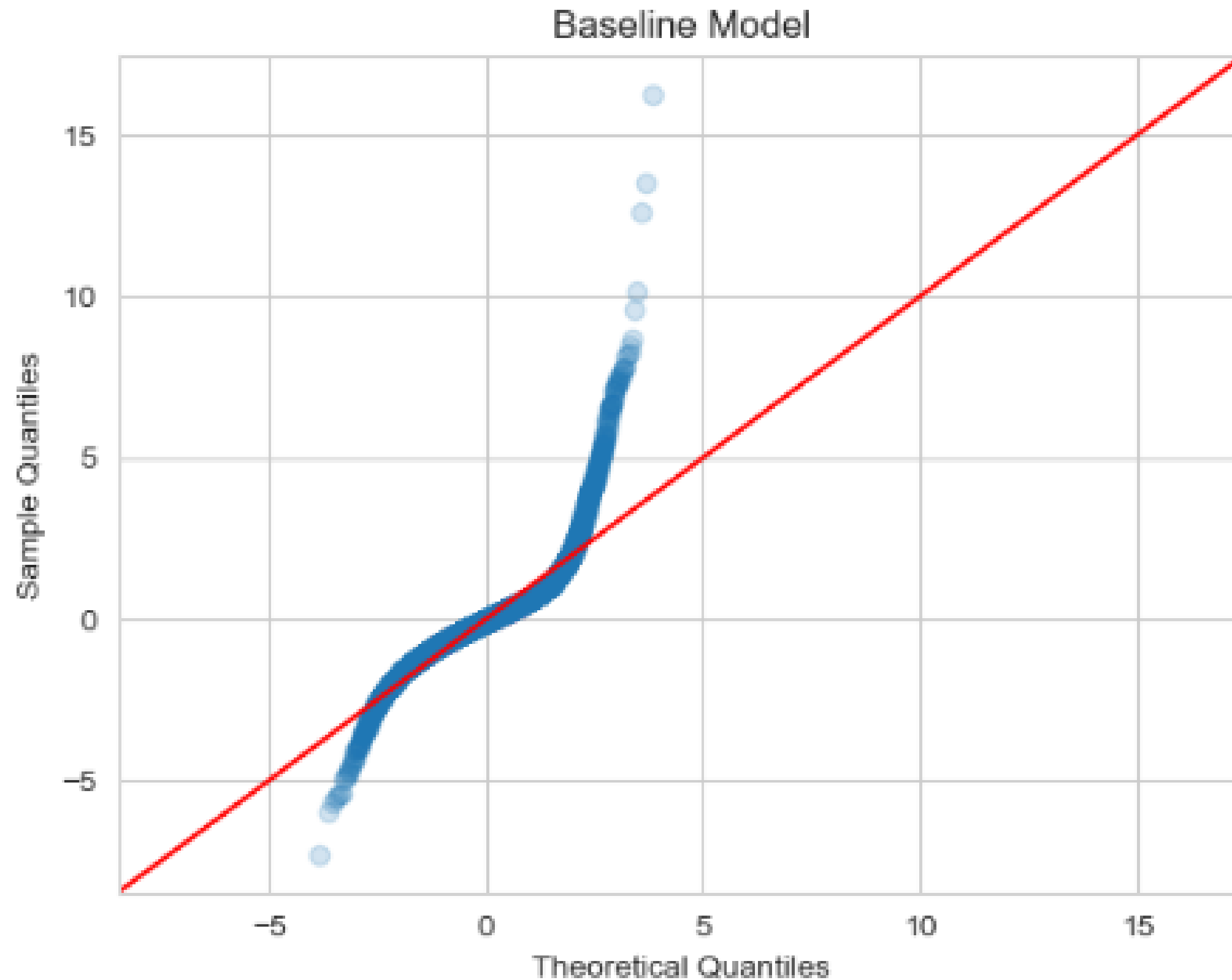
King County



NORMALITY TEST



King County



Our data has heavier tails indicating more extremes thus the data is deviating from normality therefore influencing method choice and exploration.

HYPOTHESIS TESTING



King County

We conducted a hypothesis testing using waterfront and view ratings where our null and alternative hypotheses were as follows;

- Null hypothesis - Waterfront does not have a significant influence on view ratings
- Alternative hypothesis - Waterfront has a significant influence on view ratings

Results;

After conducting a chi-test and ANOVA test we rejected our null hypothesis as results indicated that waterfront had a significant influence on view ratings.

Objective 1;
The effect of the house design i.e (number of bedrooms, number of bathrooms and number of floors) have on the sale value of the houses in King County.



OLS Regression Results						
=====						
Dep. Variable:	price		R-squared:	0.278		
Model:	OLS		Adj. R-squared:	0.278		
Method:	Least Squares		F-statistic:	2743.		
Date:	Fri, 08 Sep 2023		Prob (F-statistic):	0.00		
Time:	21:04:26		Log-Likelihood:	-3.0037e+05		
No. Observations:	21345		AIC:	6.008e+05		
Df Residuals:	21341		BIC:	6.008e+05		
Df Model:	3					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

Intercept	-3.768e+04	9492.700	-3.970	0.000	-5.63e+04	-1.91e+04
bathrooms	2.404e+05	3756.698	63.987	0.000	2.33e+05	2.48e+05
bedrooms	2.185e+04	2815.796	7.759	0.000	1.63e+04	2.74e+04
floors	-3272.6132	4606.887	-0.710	0.477	-1.23e+04	5757.232
=====						
Omnibus:	17099.619	Durbin-Watson:	1.969			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	885044.980			
Skew:	3.464	Prob(JB):	0.00			
Kurtosis:	33.776	Cond. No.	20.8			
=====						

- Multiple regression (R-squared: 27.8%) explains more price variation than a simple model (R-squared: 9.9%).
- Bathrooms contribute significantly (~\$240,400 each), and bedrooms add ~\$21,850 each.
- Number of floors lacks statistical significance (p-value: 0.477), suggesting no impact on prices.



King County

Objective 2;

The effect of dimensions of the house and lot size on the price of houses.

OLS Regression Results

```
=====
Dep. Variable:          price    R-squared:                0.503
Model:                  OLS      Adj. R-squared:            0.503
Method:                 Least Squares    F-statistic:          3603.
Date:                  Fri, 08 Sep 2023    Prob (F-statistic):      0.00
Time:                  21:04:45    Log-Likelihood:         -2.9638e+05
No. Observations:      21345    AIC:                    5.928e+05
Df Residuals:          21338    BIC:                    5.928e+05
Df Model:               6
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	-1.026e+05	5462.508	-18.786	0.000	-1.13e+05	-9.19e+04
sqft_living	230.6043	20.011	11.524	0.000	191.380	269.828
sqft_lot	0.0798	0.062	1.296	0.195	-0.041	0.200
sqft_above	3.9167	20.097	0.195	0.845	-35.474	43.308
sqft_basement	39.4588	19.990	1.974	0.048	0.276	78.641
sqft_living15	76.9778	4.071	18.907	0.000	68.998	84.958
sqft_lot15	-0.7910	0.094	-8.417	0.000	-0.975	-0.607

```
=====
Omnibus:                15309.267    Durbin-Watson:          1.988
Prob(Omnibus):           0.000    Jarque-Bera (JB):       670696.606
Skew:                    2.960    Prob(JB):               0.00
Kurtosis:                29.816    Cond. No.                1.56e+05
=====
```

- Model's R-squared: 50.3%, slightly better than the simple model's 49.2%.
- Sqft_living is strong, adding \$230.6 per square foot.
- Sqft_living15 increases prices by ~\$76.96 with more nearby living space.
- Sqft_lot15 suggests a minor price decrease (~\$0.79) with larger lot space.
- Sqft_above and sqft_basement may not significantly affect prices (p-values > 0.05) and should be considered for removal.

Objective 3;

The effect of the numeric variable with highest correlation and the viable categorical variable to prices of the house



OLS Regression Results						
=====						
Dep. Variable:	price	R-squared:	0.582			
Model:	OLS	Adj. R-squared:	0.582			
Method:	Least Squares	F-statistic:	2968.			
Date:	Fri, 08 Sep 2023	Prob (F-statistic):	0.00			
Time:	21:14:23	Log-Likelihood:	-2.9455e+05			
No. Observations:	21345	AIC:	5.891e+05			
Df Residuals:	21334	BIC:	5.892e+05			
Df Model:	10					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	1.36e+05	5364.116	25.347	0.000	1.25e+05	1.46e+05
sqft_living	157.4288	2.799	56.243	0.000	151.942	162.915
grade_10 Very Good	3.817e+05	9103.068	41.932	0.000	3.64e+05	4e+05
grade_11 Excellent	6.708e+05	1.44e+04	46.618	0.000	6.43e+05	6.99e+05
grade_12 Luxury	1.214e+06	2.76e+04	43.909	0.000	1.16e+06	1.27e+06
grade_13 Mansion	2.397e+06	6.81e+04	35.213	0.000	2.26e+06	2.53e+06
grade_4 Low	-6.941e+04	6.38e+04	-1.089	0.276	-1.94e+05	5.56e+04
grade_5 Fair	-4.599e+04	1.64e+04	-2.804	0.005	-7.81e+04	-1.38e+04
grade_6 Low Average	-2.216e+04	6094.474	-3.637	0.000	-3.41e+04	-1.02e+04
grade_8 Good	6.293e+04	4208.281	14.954	0.000	5.47e+04	7.12e+04
grade_9 Better	1.858e+05	6248.370	29.737	0.000	1.74e+05	1.98e+05
=====						
Omnibus:	13686.863	Durbin-Watson:	1.995			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	465725.595			
Skew:	2.573	Prob(JB):	0.00			
Kurtosis:	25.298	Cond. No.	9.54e+04			

- Grade indicates construction quality, a robust categorical variable.
- Model's R-squared improved to 58.2% from the baseline's 49.2%.
- Baseline price: ~\$136,000.
- Each extra sqft_living adds ~\$157.43.
- Higher grades (10-13) increase prices; Grade 13 adds ~\$2,397,000.
- Grade_4 (Low) effect uncertain with p-value 0.276.

Objective 4;

The effect of location with the other variables on price of houses



King County

OLS Regression Results

=====						
Dep. Variable:	price	R-squared:	0.801			
Model:	OLS	Adj. R-squared:	0.801			
Method:	Least Squares	F-statistic:	1086.			
Date:	Fri, 08 Sep 2023	Prob (F-statistic):	0.00			
Time:	21:15:25	Log-Likelihood:	-2.8660e+05			
No. Observations:	21345	AIC:	5.734e+05			
Df Residuals:	21265	BIC:	5.740e+05			
Df Model:	79					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	-6.728e+05	1.52e+04	-44.201	0.000	-7.03e+05	-6.43e+05
age	1084.6468	59.775	18.146	0.000	967.484	1201.810
numeric_condition	2.421e+04	1952.656	12.399	0.000	2.04e+04	2.8e+04
numeric_grade	6.462e+04	1865.065	34.647	0.000	6.1e+04	6.83e+04
sqft_living	128.2788	3.327	38.558	0.000	121.758	134.800
bathrooms	1.067e+04	2573.653	4.144	0.000	5621.025	1.57e+04
view_rating_numeric	6.65e+04	1964.184	33.856	0.000	6.26e+04	7.03e+04
numeric_waterfront	6.227e+05	1.45e+04	42.919	0.000	5.94e+05	6.51e+05
sqft_lot	0.2383	0.030	8.002	0.000	0.180	0.297
sqft_above	53.8706	3.270	16.474	0.000	47.461	60.280
sqft_living15	12.4718	2.954	4.222	0.000	6.682	18.262
zipcode_98002	2.576e+04	1.47e+04	1.757	0.079	-2980.675	5.45e+04
zipcode_98003	-1.582e+04	1.32e+04	-1.199	0.231	-4.17e+04	1e+04
zipcode_98004	7.656e+05	1.29e+04	59.353	0.000	7.4e+05	7.91e+05
zipcode_98005	2.952e+05	1.55e+04	18.996	0.000	2.65e+05	3.26e+05
zipcode_98006	2.629e+05	1.17e+04	22.482	0.000	2.4e+05	2.86e+05
zipcode_98007	2.292e+05	1.65e+04	13.893	0.000	1.97e+05	2.62e+05
zipcode_98008	2.142e+05	1.32e+04	16.233	0.000	1.98e+05	2.30e+05

Omnibus:	21318.296
Prob(Omnibus):	0.000
Skew:	4.394
Kurtosis:	77.261

Durbin-Watson:	1.986
Jarque-Bera (JB):	4973359.080
Prob(JB):	0.00
Cond. No.	2.91e+06

- Model's R-squared: 80.1%, highest coefficient for zipcode 98004: \$1,330,000 increase.
- Concern: Multicollinearity affecting results.
- Solution: Employed Ridge and Lasso regularization.



Ridge and Lasso regression techniques

- Ridge and Lasso models: MSE of 0.002478 each.
- This implies small squared differences between predicted and actual values.
- Ridge and Lasso make reasonably accurate predictions; lower MSE suggests better model performance.

Conclusion



King County

- OLS model impressively explains 80.1% of price variance (R-squared: 0.801).
- Multicollinearity addressed with Ridge and Lasso regularization, enhancing stability.
- Ridge: R-squared 0.0025, Lasso: Low MSE 0.0025.
- Stick with OLS model for reliable price predictions, backed by strong performance and regularization



Recommendations

1. Bathrooms and bedrooms significantly impact house value (Bath: +\$240,400, Bed: +\$21,850 each).
2. More square footage (sqft_living) adds \$230.6 per sqft, nearby living space (sqft_living15) adds \$76.96.
3. Combining grade and sqft_living boosts price (+\$157.43 per sqft, Grade 13: +\$2,397,000).
4. Location, especially 98004 (Bellevue), plays a vital role, contributing \$1,330,000, with an 80.1% R-squared.



King County