Module 2 Data Science Project

King County property price predictor

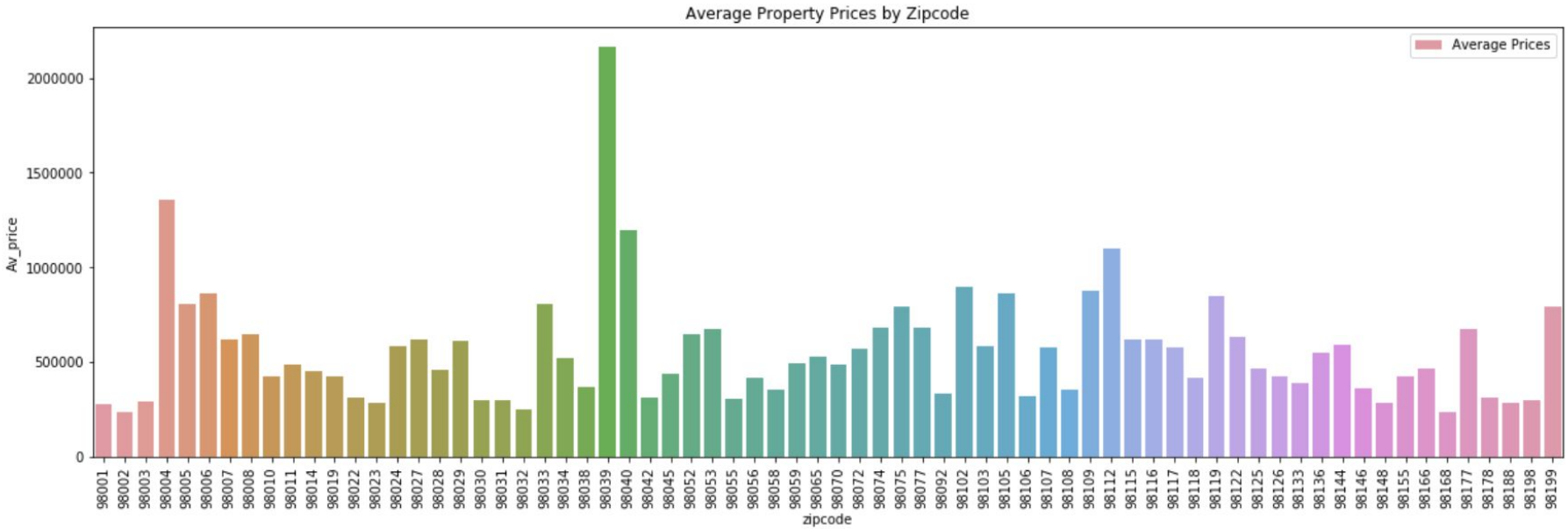
By Naweed and Jim

Intro & Project Outcomes

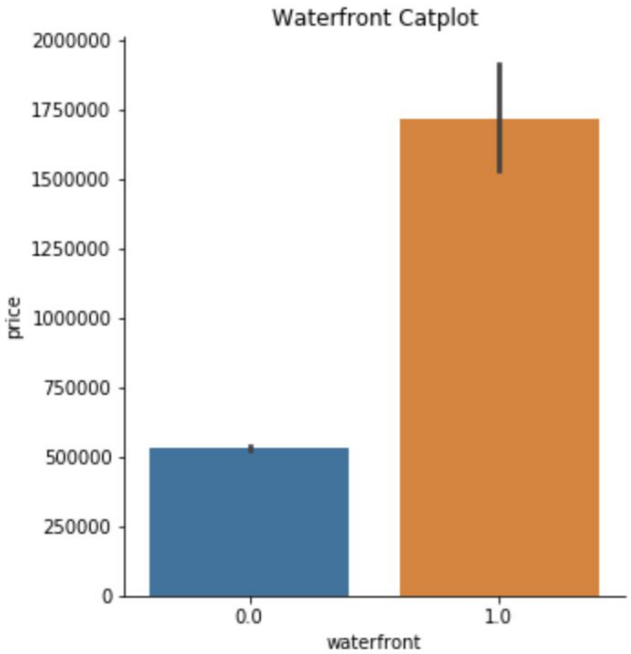
Our aim with this project is to produce effective pricing models for the domestic housing market within the King County area. We have investigated a number of parameters available to us to give accurate predictions of house prices in different cases. We have been guided by the demand for information by different sections of the domestic housing market - homeowners, developers, investors and agents looking for more insight. We developed the following guide to inform our work:

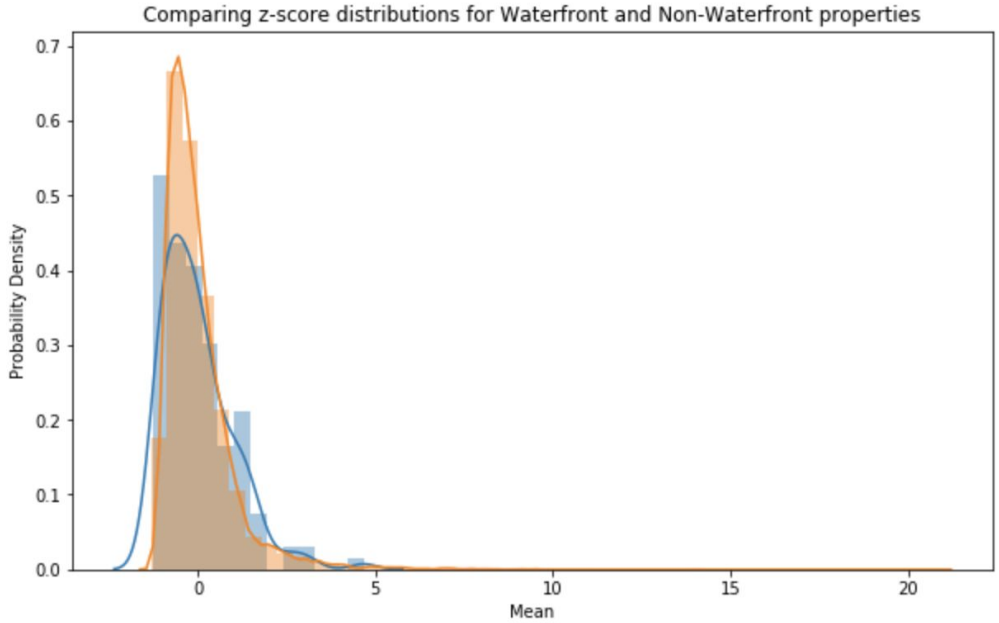
1. Create and demonstrate an accurate tool to predict house prices within the King County Area by zip-code.
2. Provide a guide to current homeowners who are considering adding value to their property and would like to understand if it is worth the initial outlay.
3. Provide a guide to prospective homeowners who would like to know whether a property is under or over-valued.

How do average property prices rank across King County?

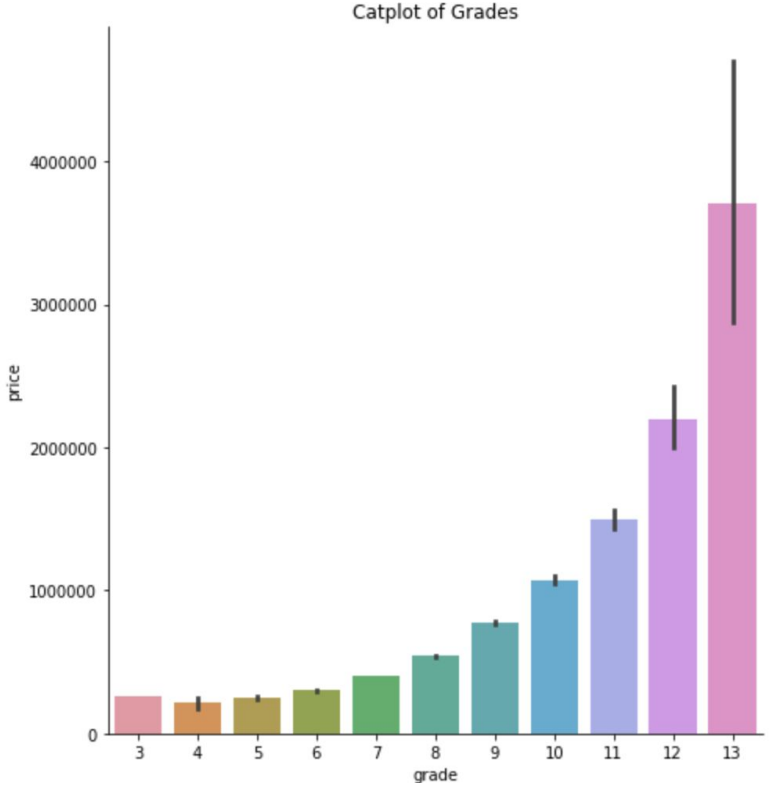


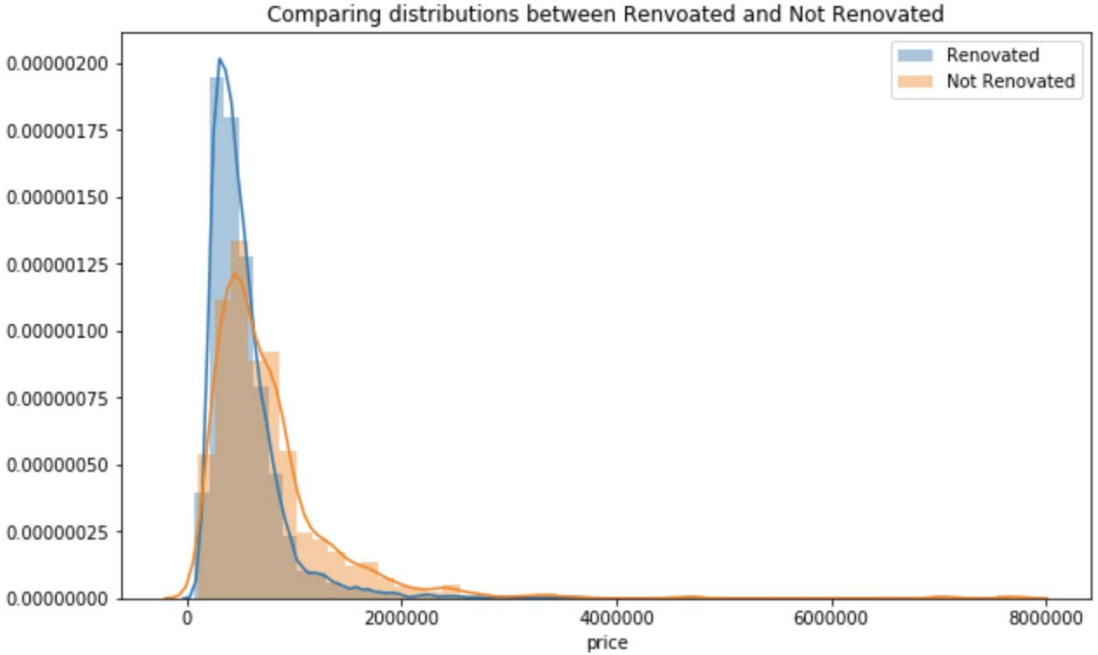
Does living by the waterfront add value?





Does Grade add value to property? If so is it worth buying low grade properties and renovating?





Final Regression Model

$$price = 293 * \beta_{sqft_living} + 124,077 * \beta_{renovate_15} + 26,368 * \beta_{waterfront} - 26,368 * \beta_{yr_built} + 4,100,000$$

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.559
Model:	OLS	**Adj. R-squared:**	0.559
Method:	Least Squares	**F-statistic:**	6853.
Date:	Fri, 27 Mar 2020	**Prob (F-statistic):**	0.00
Time:	03:33:12	**Log-Likelihood:**	-2.9851e+05
No. Observations:	21595	**AIC:**	5.970e+05
Df Residuals:	21590	**BIC:**	5.971e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.103e+06	1.18e+05	34.828	0.000	3.87e+06	4.33e+06
sqft_living	293.9264	1.929	152.348	0.000	290.145	297.708
renovate_15	1.241e+05	1.28e+04	9.692	0.000	9.9e+04	1.49e+05
waterfront	8.188e+05	2.04e+04	40.132	0.000	7.79e+05	8.59e+05
yr_built	-2121.9655	60.396	-35.134	0.000	-2240.347	-2003.584

Baseline Regression Model

$$price = 270 * \beta_{sqft\_living} + 862,256 * \beta_{waterfront} + 26,363 * \beta_{basement} - 37,390$$

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.531
Model:	OLS	**Adj. R-squared:**	0.531
Method:	Least Squares	**F-statistic:**	8139.
Date:	Fri, 27 Mar 2020	**Prob (F-statistic):**	0.00
Time:	10:43:45	**Log-Likelihood:**	-2.9921e+05
No. Observations:	21596	**AIC:**	5.984e+05
Df Residuals:	21592	**BIC:**	5.985e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.739e+04	4291.221	-8.713	0.000	-4.58e+04	-2.9e+04
sqft_living	270.0042	1.914	141.083	0.000	266.253	273.755
waterfront	8.623e+05	2.1e+04	41.021	0.000	8.21e+05	9.03e+05
basement	2.636e+04	3593.738	7.336	0.000	1.93e+04	3.34e+04

Test - Train Split (80/20 split)

Train Dataset

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.568
Model:	OLS	**Adj. R-squared:**	0.568
Method:	Least Squares	**F-statistic:**	5681.
Date:	Fri, 27 Mar 2020	**Prob (F-statistic):**	0.00
Time:	03:42:47	**Log-Likelihood:**	-2.3895e+05
No. Observations:	17276	**AIC:**	4.779e+05
Df Residuals:	17271	**BIC:**	4.779e+05
Df Model:	4		
Covariance Type:	nonrobust		

Test Dataset

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.524
Model:	OLS	**Adj. R-squared:**	0.524
Method:	Least Squares	**F-statistic:**	1189.
Date:	Fri, 27 Mar 2020	**Prob (F-statistic):**	0.00
Time:	03:45:05	**Log-Likelihood:**	-59523.
No. Observations:	4319	**AIC:**	1.191e+05
Df Residuals:	4314	**BIC:**	1.191e+05
Df Model:	4		
Covariance Type:	nonrobust		

Recommendations

1. For anyone looking to invest in property in King County, it seems that waterfront properties do command a premium.

2. There seems to be a case for buying lower grade properties, renovating it and then selling it on.

Future Work

In order to expand our work going forward, it would be good to get access to data from a longer period to assess trends and make stronger recommendations.

It would be good to be able to get access to more location data and incorporate it into the model. As slide 3 shows, there does seem to be a few locations in King County, namely around downtown Seattle where prices are higher on average.