DataCorp.™

House Price Analysis & Insights

John Paul Hernandez Alcala

Hello, my name is John Paul, and I am a data scientist with DataCorp. Today my goal is to give you a better understanding of house pricing. Let's get started!

Before we dive into the analysis of house pricing, we have to present ourselves with three questions that make up our problem statement: Investigate which house features affect house price, derive a model that can predict house pricing, and uncover 3 concrete features that highly influence house price.

Business Value

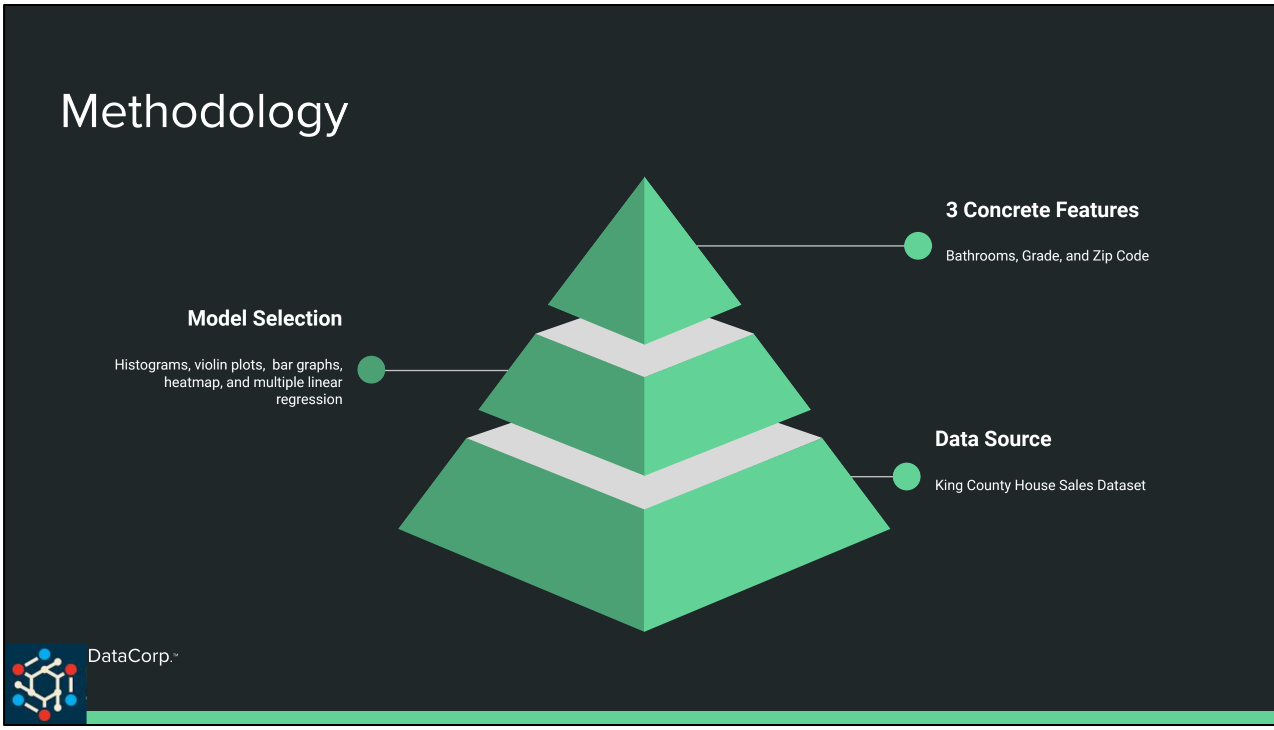
01	Investigate which house features could be price predictors	• Heatmap for Price vs Features • Average Price vs Categorical Features • Violin Plots for Price vs Categorical Features • Histogram for Continuous Features
02	Derive a model that can predict house pricing	• Multiple linear regression

03	Uncover 3 concrete features that highly influence house price	• Top Predictors(coefficients) from model
		    ○ Bathrooms
		    ○ Grade
		    ○ Zip Code

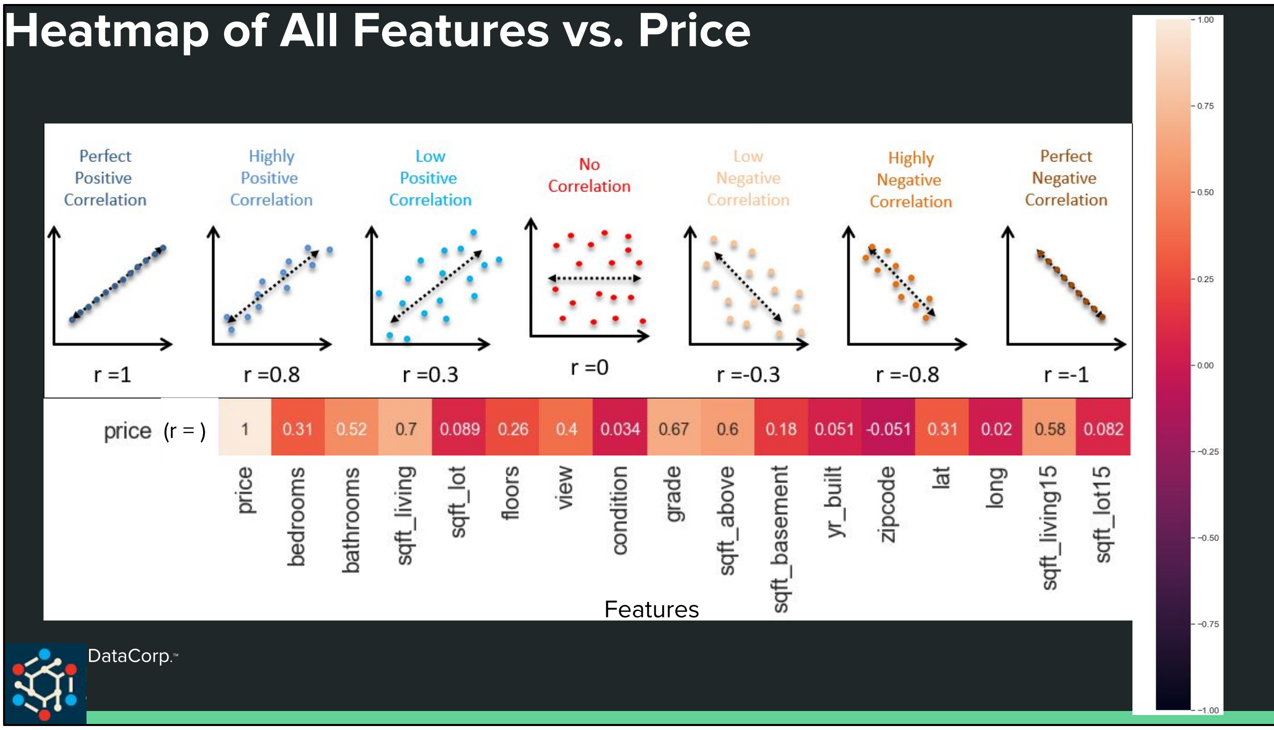
DataCorp.™

We address these questions to guide a buyer or seller on which features to look for or showcase. For the 1st question, we will look at multiple graphs, so we know which features result in a high or low house price.

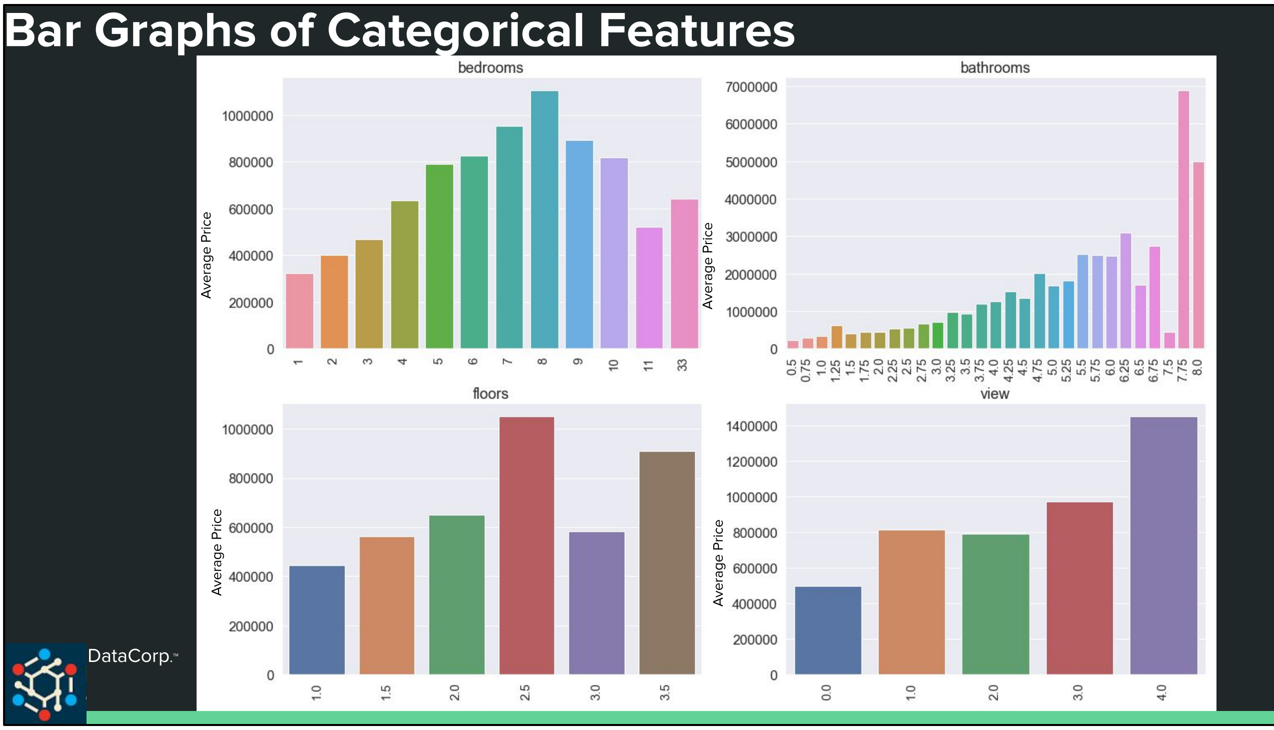
For the 2nd question, we will use our data to make a multiple linear regression model that can accurately predict house pricing. Finally, based on the results of our model, we can identify three concrete features that highly influence house prices which are Bathrooms, Grade and Zip Code.



With the previous questions answered, we are able to start off with data we obtained from King County, work our way up with statistical models such as histograms, violin plots, bar graphs, heatmap, and multiple linear regression, and finally arrive to our three concrete features that highly influence housing prices.

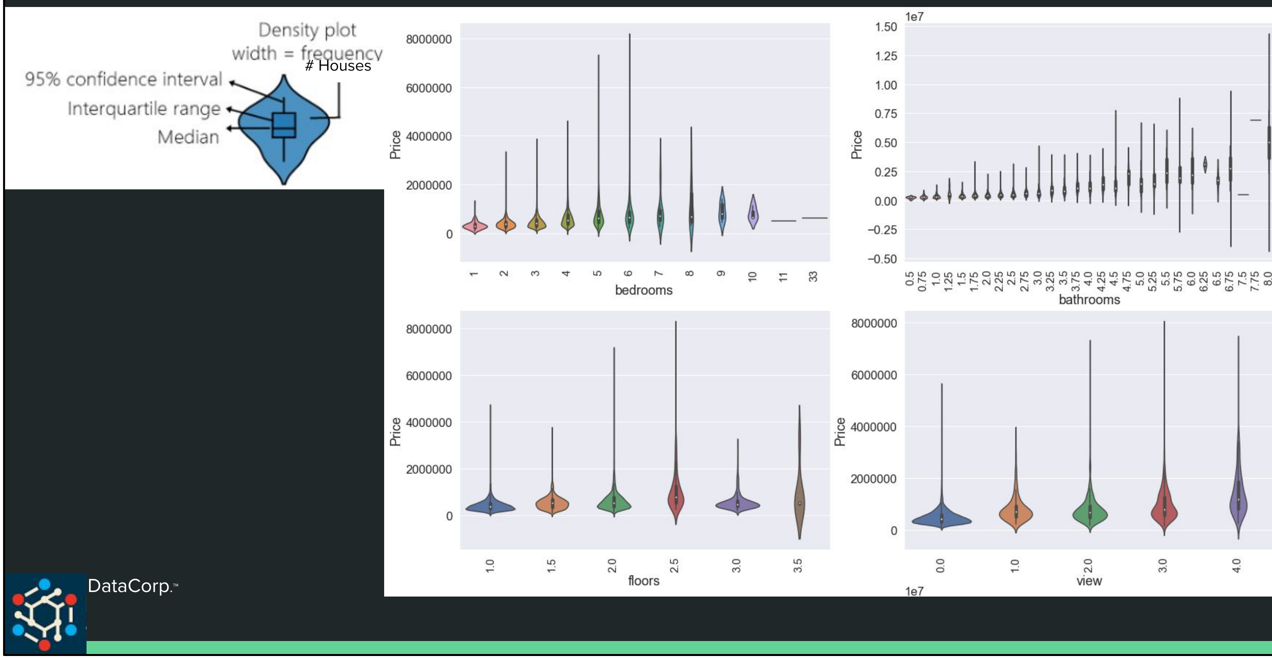


To start our analysis, we look at the general correlation of our features to ‘price’. We identify 'bedrooms', 'bathrooms', 'floors', 'view', 'condition', 'grade', 'sqft_basement', 'yr_built', 'zipcode' as categorical features, and ‘sqft_lot’, ‘lat’, ‘sqft_living15’, and ‘sqft_lot15’ as continuous features.



From these graphs we can see within each feature which sub-features have a higher average price. So for example, it looks like 8 bedrooms result in the highest Average Price.

Violin Plots of Categorical Features



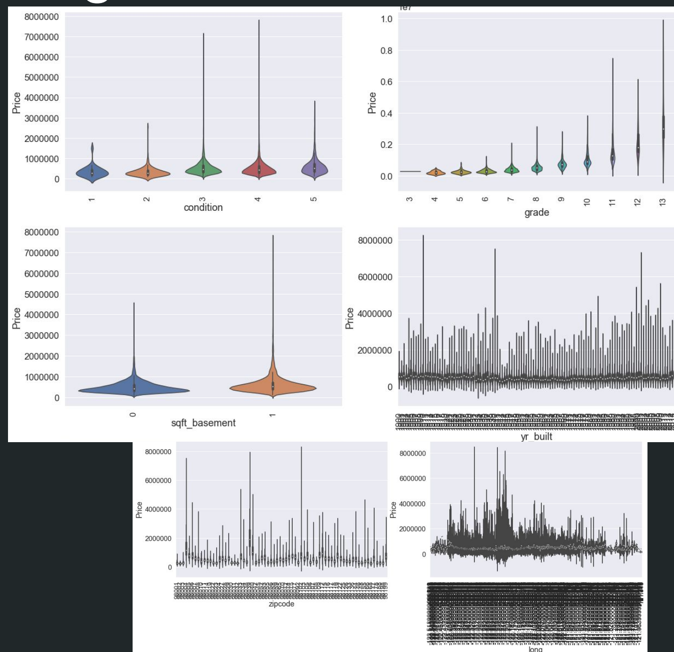
From the violin plots of the previous features, we can see under which sub-feature in each feature are houses concentrated; this is identified by looking at the width of the each violin plot. So for example, it is frequent for houses to have 1-4 bedrooms.

Bar Graphs of Categorical Features

DataCorp.™

Here are the bar graphs for the other features.

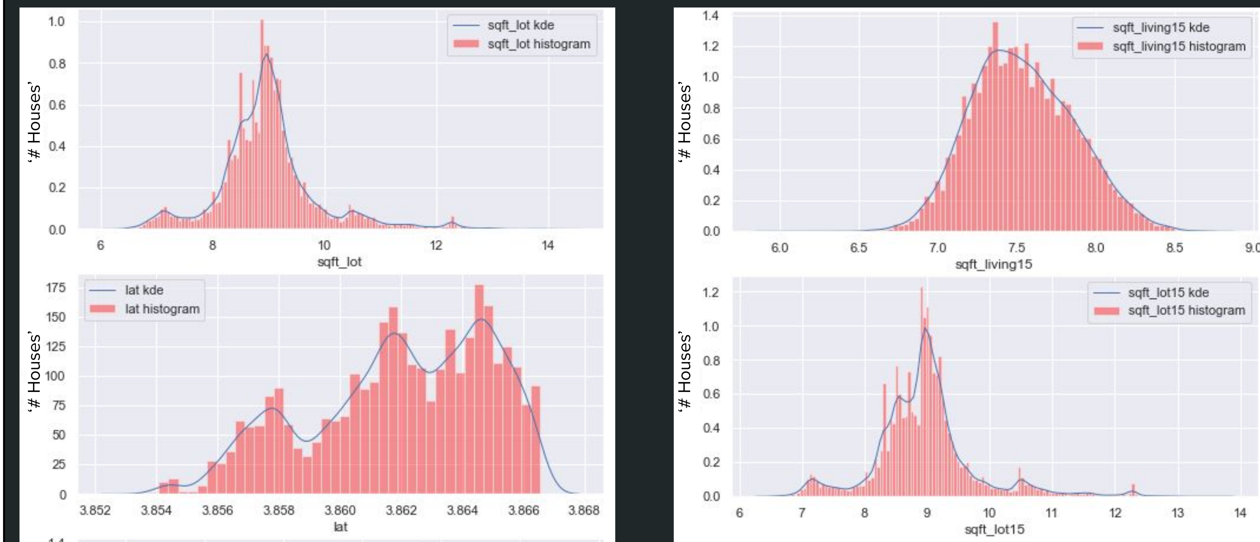
Violin Plots of Categorical Features



DataCorp.™

And here are the violin plots for the previous features. To use these features in our model, we one-hot encoding all these variables first.

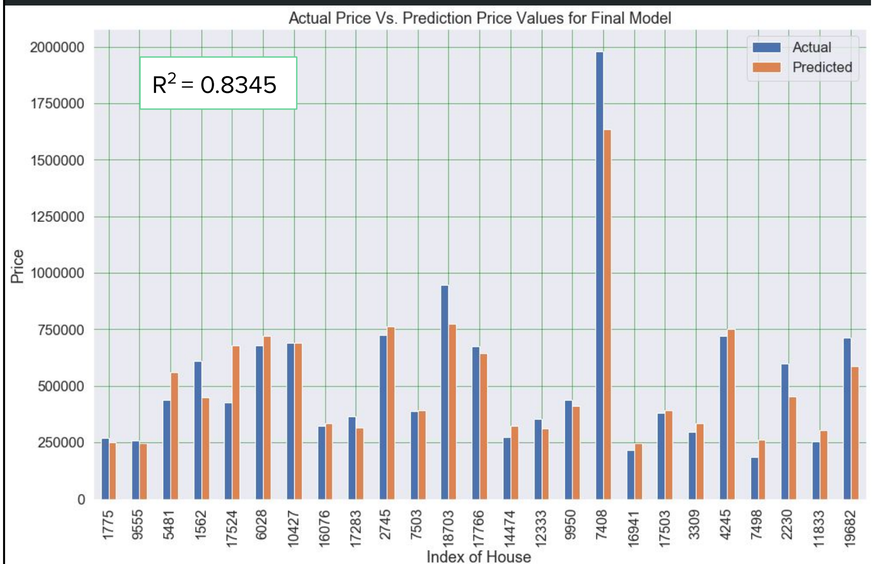
Distribution of Continuous Features



DataCorp.™

We can next look at how our continuous features are distributed. We can see that 'sqft_living15' is nicely distributed after transforming the data. Note: the values have been transformed logarithmically and normalized, so they will look weird and not really understandable. From this and other results, we determine that only 'sqft_living15' should be used in our model (only one that follows Assumptions for Linear Regression)

Final Model and 3 Concrete Features



DataCorp.™

```
Mean Absolute Error: 0.3076337245356352
Mean Squared Error: 0.16653983429979327
Root Mean Squared Error: 0.4080929236090639
coefficient of determination R^2 of the prediction: 0.8344757279089347
```

	Coefficeint
bathrooms_6.25	2.905211
zipcode_98039	2.534385
zipcode_98004	2.148283
grade_13	2.113758
zipcode_98112	1.886598
bathrooms_8.0	1.773682
zipcode_98109	1.741787
zipcode_98040	1.664505
zipcode_98119	1.652526
zipcode_98105	1.648524
zipcode_98102	1.577912
grade_12	1.514799
zipcode_98199	1.457809
bathrooms_6.5	1.442780
zipcode_98115	1.434040
zipcode_98103	1.411139
zipcode_98005	1.410718
bathrooms_5.5	1.406927
zipcode_98033	1.396852
zipcode_98117	1.378351
zipcode_98107	1.364407
bathrooms_4.75	1.316179
zipcode_98122	1.306652
bathrooms_5.0	1.238727
zipcode_98116	1.198955
zipcode_98007	1.179953
zipcode_98053	1.162380
zipcode_98008	1.160271
zipcode_98052	1.159679
bathrooms_4.25	1.148817
bathrooms_4.5	1.140382
zipcode_98006	1.135421
bathrooms_5.25	1.111341
zipcode_98075	1.106814
zipcode_98136	1.098633
grade_11	1.089995
bathrooms_4.0	1.070637
view_4.0	1.067938
zipcode_98144	1.054376
zipcode_98177	1.044663

	Coefficeint
bathrooms_6.25	2.905211
zipcode_98039	2.534385
zipcode_98004	2.148283
grade_13	2.113758
zipcode_98112	1.886598
...	...
floors_3.0	-0.461790
floors_3.5	-0.551197
grade_6	-0.584346
grade_5	-0.882641
grade_4	-1.162832

With all the features, we built a model with multiple linear regression. This model was evaluated for performance and found to have an accuracy of 83.45% and a high precision (low error). Because of the high accuracy and precision of this model, we can identify three concrete features (also known as coefficients) that highly influence housing prices: Bathrooms, Grade and Zip Code.

Recommendations Derived from Model

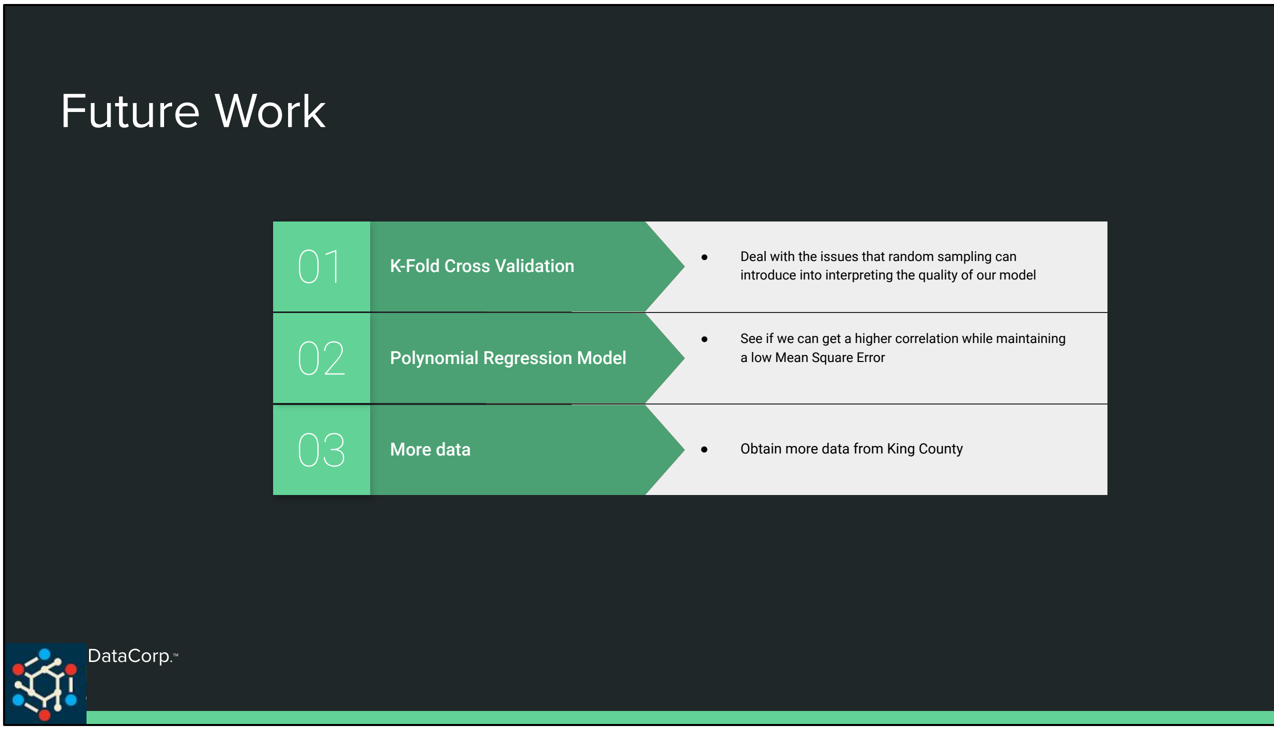
1
- Allows housing developments to see which features in houses will see higher, lower, or negligible effects on house price

2
- Allows sellers to accurately and precisely place prices on houses

3
- Allows placements of houses with specific features in specific budget ranges for buyers

DataCorp.™

With our model, we are able to see whether certain features will affect house price (or not), to accurately and precisely place prices on houses, and to place houses with specific features in specific budget ranges.



Now that we have identified three concrete features that highly influence housing prices, are there ways we can have more confidence with our predictors? YES! We can use K-Fold Cross Validation to deal with the issues that random sampling (splitting train-testing data) can introduce into interpreting the quality of our model. We can use a polynomial regression model that may model the data better and thus render a better correlation value (accuracy) while maintaining high precision (low error). Finally, we can get more data to train and test our current model on which will also render a better correlation value (accuracy) while maintaining high precision (low error).

Thank you for your time! Please, feel free to ask me any questions at this time.

Appendix 1a. Features from Dataset

```
id** - unique identified for a house
dateDate** - house was sold
pricePrice** -  is prediction target
bedroomsNumber** -  of Bedrooms/House
bathroomsNumber** -  of bathrooms/bedrooms
sqft_livingsquare** -  footage of the home
sqft_lotsquare** -  footage of the lot
floorsTotal** -  floors (levels) in house
waterfront** - House which has a view to a waterfront
view** - Has been viewed
condition** - How good the condition is ( Overall )
grade** - overall grade given to the housing unit, based on King County grading system
sqft_above** - square footage of house apart from basement
sqft_basement** - square footage of the basement
yr_built** - Built Year
yr_renovated** - Year when house was renovated
zipcode** - zip
lat** - Latitude coordinate
long** - Longitude coordinate
sqft_living15** - The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15** - The square footage of the land lots of the nearest 15 neighbors
```

DataCorp.™

Key for features

Appendix 1b. Model Equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = expanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

```
Mean Absolute Error: 0.3076337245356352
Mean Squared Error: 0.16653983429979327
Root Mean Squared Error: 0.4080929236090639
coefficient of determination R^2 of the prediction: 0.8344757279089347
```

```
B_0 = -1.6189635235482411
```

	Coefficeint
bathrooms_6.25	2.905211
zipcode_98039	2.534385
zipcode_98004	2.148283
grade_13	2.113758
zipcode_98112	1.886598
bathrooms_8.0	1.773682
zipcode_98109	1.741787
zipcode_98040	1.664505
zipcode_98119	1.652526
zipcode_98105	1.648524
zipcode_98102	1.577912
grade_12	1.514799
zipcode_98199	1.457809
bathrooms_6.5	1.442780
zipcode_98115	1.434040
zipcode_98103	1.411139
zipcode_98005	1.410718
bathrooms_5.5	1.406927
zipcode_98033	1.396852
zipcode_98117	1.378351
zipcode_98107	1.364407
bathrooms_4.75	1.316179
zipcode_98122	1.306652
bathrooms_5.0	1.238727
zipcode_98116	1.198955
zipcode_98007	1.179953
zipcode_98053	1.162380
zipcode_98008	1.160271
zipcode_98052	1.159679
bathrooms_4.25	1.148817
bathrooms_4.5	1.140382
zipcode_98006	1.135421
bathrooms_5.25	1.111341
zipcode_98075	1.106814
zipcode_98136	1.098633
grade_11	1.089995
bathrooms_4.0	1.070637
view_4.0	1.067938
zipcode_98144	1.054376
zipcode_98177	1.044663

	Coefficeint
bathrooms_6.25	2.905211
zipcode_98039	2.534385
zipcode_98004	2.148283
grade_13	2.113758
zipcode_98112	1.886598
...	...
floors_3.0	-0.461790
floors_3.5	-0.551197
grade_6	-0.584346
grade_5	-0.882641
grade_4	-1.162832

DataCorp.™

Equation with coefficients (103 different ones), y-intercept, and epsilon = sqrt(MSE)