King County Housing Data

Presented by:
Dharun Selvan
Rosario Fabian
Ashay Kargaonkar

AGENDA

- Introduction
- Project Info
- Data Preparation
- Modeling
- Validation
- Implementation

INTRODUCTION

- We are going to predict the price of a house depending upon the different variables/factors.
- King county is in the state Washington. This dataset also covers few parts of Seattle.
- Data set was created on 2017 and contains data about the houses built from 1900 to 2015
- The dataset contains 21 variables and 21613 observations.
- The original data set was taken from Kaggle.com

Project Info

- The Model we used : Regression Model
- The software we used in this project : SAS 9.4 ,IBM SPSS Statistics 26, MS EXCEL 2019 , Jupyter Notebook (Python V3.7) , Visual Studio 2019 (Visual Basic)
- Data Variables in the Data Set:

VARIABLE NAME	DESCRIPTION
ID	Id of the Housing
Date	Date of the observation made (1900-2015)
Price	Price of the house when observations made
Bedrooms	Number of bedrooms in the house (0-11)
Bathrooms	Number of Bathrooms in the house (0-8)
Sqft_living	Area of the house built in Sqft
Sqft_lot	Total area of the lot/ground
Floors	Number of floors (1-3.5)

Project Info - Cont.

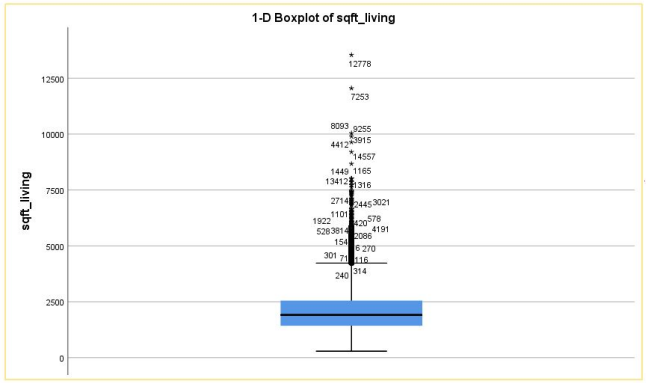
Waterfront	Lake View (0 or 1)
View	Number of Views from the house (0-4)
Condition	Condition of the house (0-5)
Grade	Evaluation of construction materials and level of craftmanship used to build the house (1-13)
Sqft_above	Area of the house built above the basement
Sqft_basement	Area of the basement in Sqft
Yr_built	Year of the house when build (1900-2015)
Yr_renovate	Year of renovation of the house
Zipcode	Zipcode of the area where the house is located
Lat	Latitude of the house location
Long	Longitude of the house location
Sqft_living15	The average house square footage of the 15 closest houses
Sqft_lot15	The average lot square footage of the closest houses

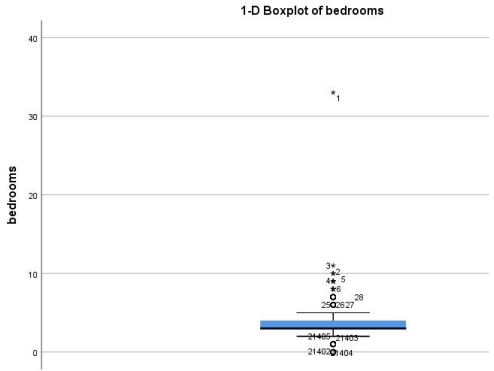
DATA PREPROCESSING

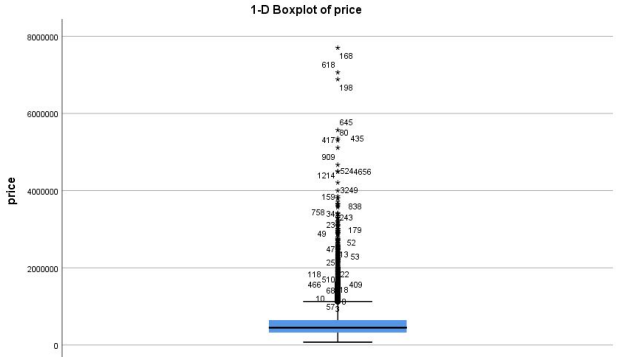
- Data Preprocessing was done using Python,SPSS and Excel.
- Python was used to find the missing values. Those missing values were replaced with the mean value.
- Dummy Variables were created for zipcode, basements, renovations and views using SPSS.
- With the help of Data Exploration, we identified the outliers and we came up with two final data set
    - Data Set without any outliers
    - Data Set with Transformed values of Outliers

DATA PREPROCESSING-Cont.

Outliers found in sqft_living, bedrooms and price

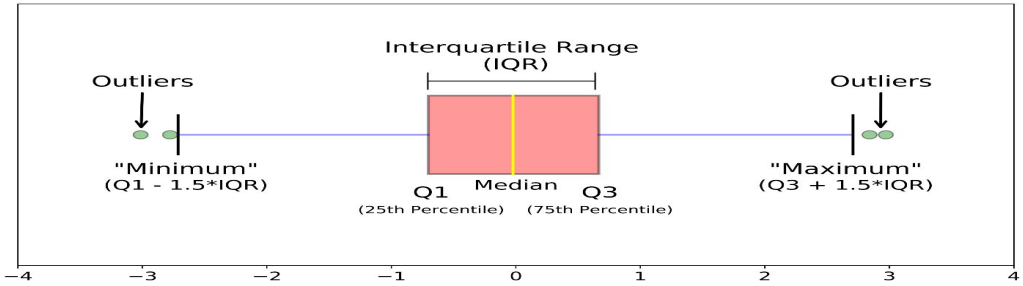






DATA PREPROCESSING-Cont.

- Outlier were calculated using the formula shown in the right side image.



- Both data set splitted into two as train and test with 80%-20% using following python code.
- The final train data set has 17289 rows and 29 columns.

```
from sklearn.model_selection import train_test_split
final_mb_t,final_mb_v=train_test_split(data,test_size=0.2,random_state=33)
```

DATA PREPROCESSING-Cont.

Variables Created and Transformed:

VARIABLES	DESCRIPTION
Zip_range	4 bins were created on the zip code and Dummy variables are created in the data set as zip1,zip2,zip3
year_range	3 bins were created on the yr_built and Dummy variables are created in the data set as yr1, yr2. Bin 1 is from 2013-2015. Bin 2 is from 2003-2012. Bin 3 is from 1900 to 2012
Log_sqft_lot	Log value of Sqft_lot
Yr_age	Age of the house since built
View_t	Dummy : 0- No views : 1- has view(s)
Sqft_basement_t	Dummy : 0- No basement 1- Has basement
Yr_renovated_age	Age of the house since renovated
Yr_renovated_d	Dummy: 0- Not renovated 1- renovated

Modeling: Exploration: Univariate

The SAS System

The UNIVARIATE Procedure
Variable: sqft_living

Moments

N	21613	**Sum Weights**	21613
Mean	2079.89974	**Sum Observations**	44952873
Std Deviation	918.440897	**Variance**	843533.681
Skewness	1.47155543	**Kurtosis**	5.24309299
Uncorrected SS	1.11728E11	**Corrected SS**	1.82304E10
Coeff Variation	44.1579409	**Std Error Mean**	6.24731907

Basic Statistical Measures

Location		**Variability**	
Mean	2079.900	**Std Deviation**	918.44090
Median	1910.000	**Variance**	843534
Mode	1300.000	**Range**	13250
		Interquartile Range	1123

Tests for Location: Mu0=0

Test		**Statistic**	**p Value**	
Student's t	**t**	332.9268	**Pr** > |t|	<.0001
Sign	**M**	10806.5	**Pr** >= |M|	<.0001
Signed Rank	**S**	1.1679E8	**Pr** >= |S|	<.0001

Quantiles (Definition 5)

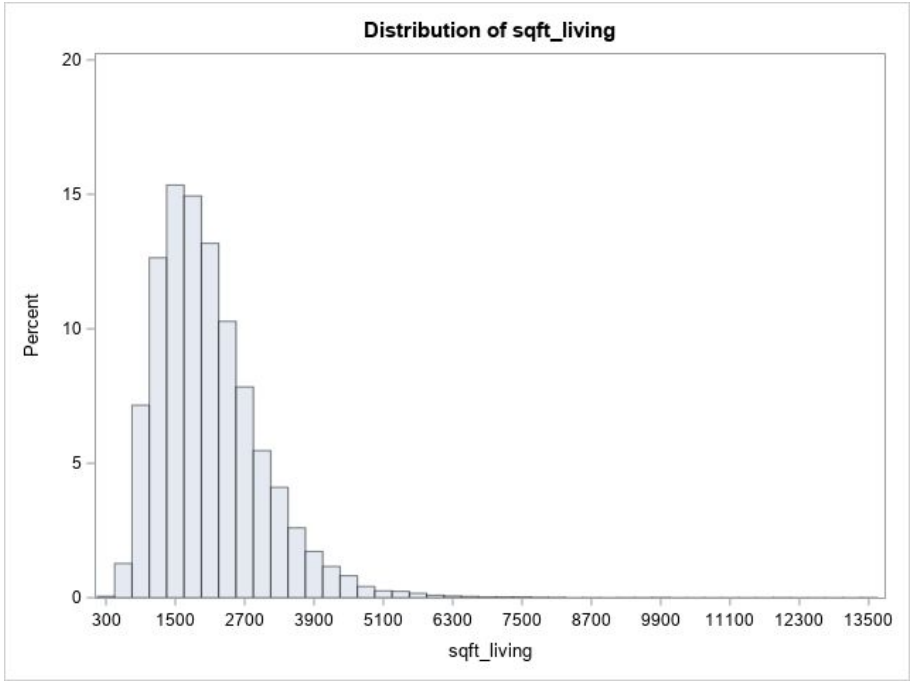
Level	**Quantile**
100% Max	13540
99%	4980
95%	3760
90%	3250
75% Q3	2550
50% Median	1910
25% Q1	1427
10%	1090
5%	940
1%	720
0% Min	290

Extreme Observations

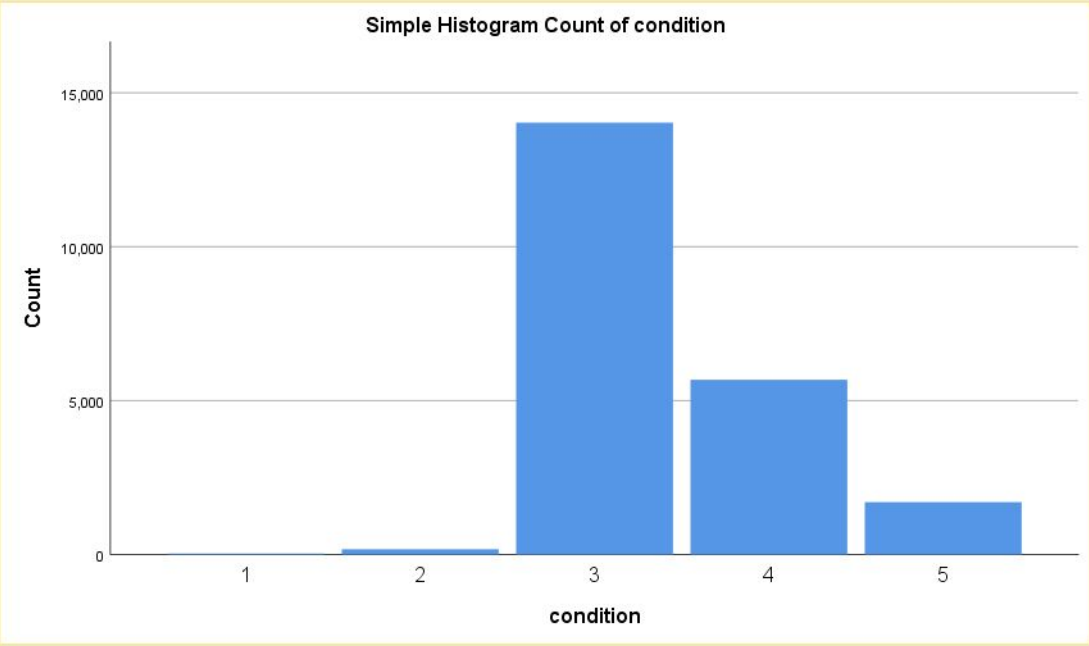
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
290	19453	9640	8093

Modeling: Exploration Univariate

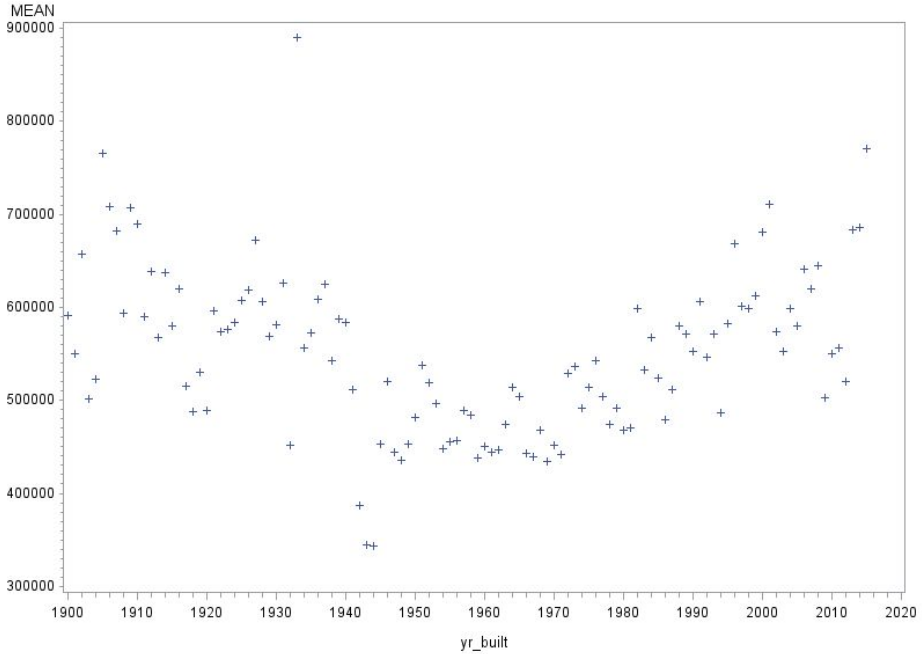
Size of building



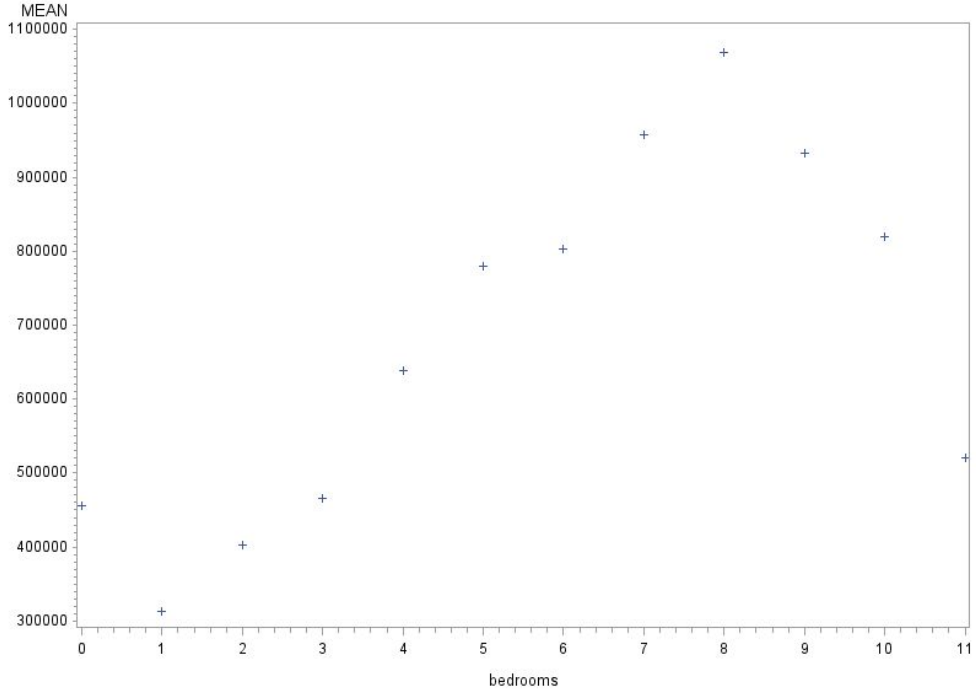
Modeling: Exploration Univariate



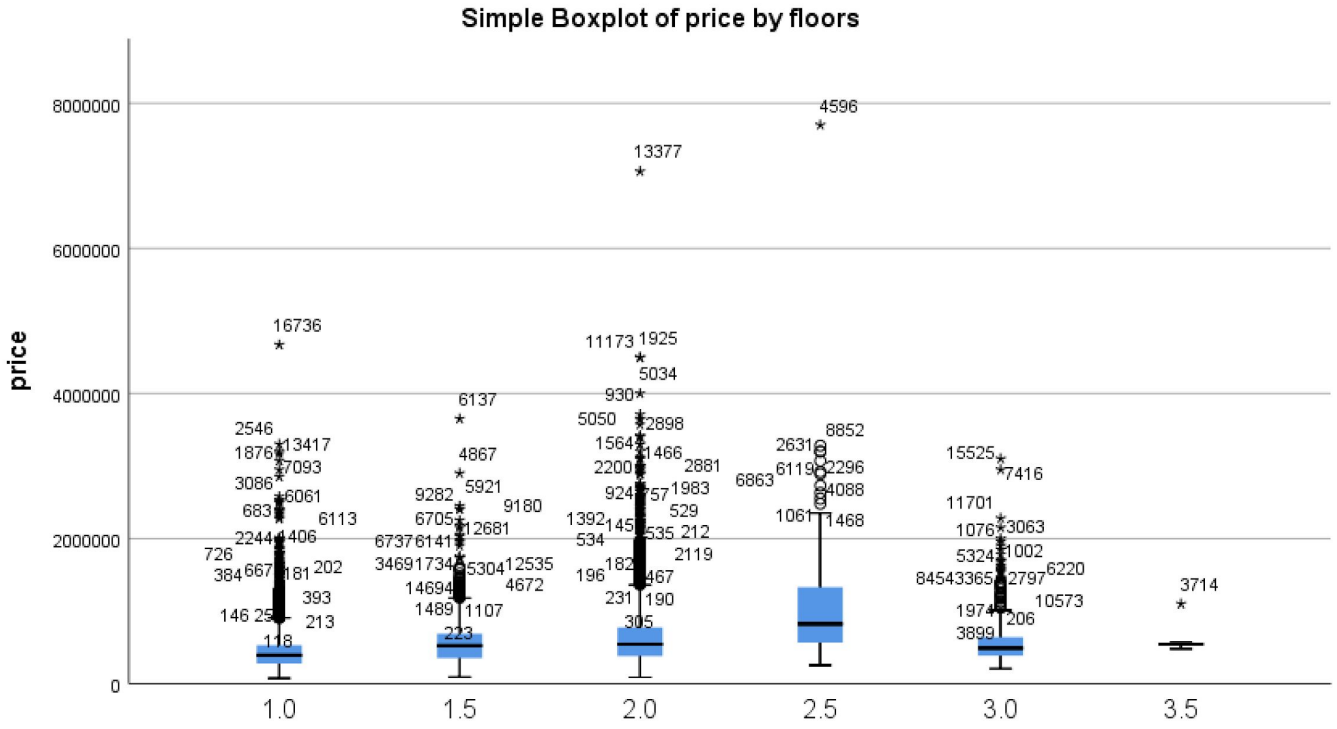
Modeling: Exploration Bivariate



Modeling: Exploration Bivariate



Modeling: Exploration Bivariate



Modeling: Different Models

	Model 1	Model 2	Model 3	Model 4	Model 5
Age	X	X	X	X	X
Renovation					X
Size	X	X	X	X	X
Bedrooms	X	X	X	X	X
View					X
Condition	X				
Grade	X	X	X		X
Zone		X	X		
Floors			X	X	
Living					X
Water front				X	
Basement				X	
Bathroom	X				
Rsquare	**0.54**	**0.58**	**0.58**	**0.62**	**0.64**

Modeling: Final Model

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 (x_2 - k) + \beta_5 x_3 + \beta_6 x_1 x_3 + \beta_7 \log_2 x_4 + \beta_8 x_5 + \beta_9 x_6 + \beta_{10} x_7$$

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	17289
Number of Observations Used	17289

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1.422403E15	1.422403E14	3057.01	<.0001
Error	17278	8.03933E14	46529283939		
Corrected Total	17288	2.226336E15			

Root MSE	215706	**R-Square**	**0.6389**
Dependent Mean	540198	**Adj R-Sq**	0.6387
Coeff Var	39.93104		

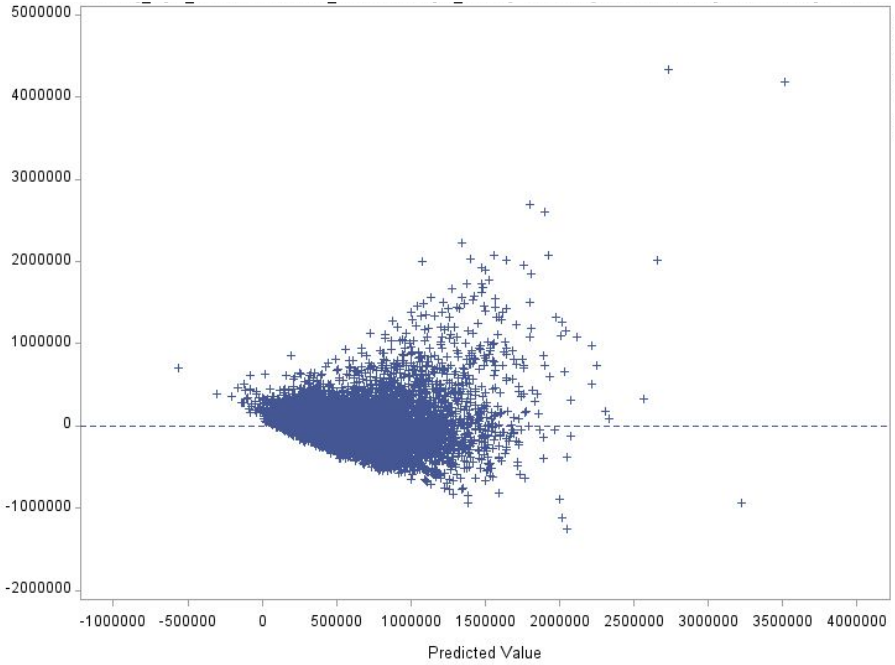
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-566374	24545	-23.07	<.0001
yr_age	1	8643.41071	626.24416	13.80	<.0001
YR_AGE_CUADRA	1	55.69137	6.77634	-8.22	<.0001
bedrooms	1	36286	2328.87542	-15.58	<.0001
X2STAR	1	38793	42948	0.90	0.0488
yr_renovated_age	1	-6133.55030	629.94248	-9.74	<.0001
INT_YR_AGE_YR_RENOVATED	1	64.61776	6.96050	9.28	<.0001
log_sqft_lot	1	-36169	2193.32857	-16.49	<.0001
view_t	1	137908	5922.08304	23.29	<.0001
sqft_living	1	204.91411	3.51002	58.38	<.0001
grade	1	127454	2350.14427	54.23	<.0001

Modeling: Validation

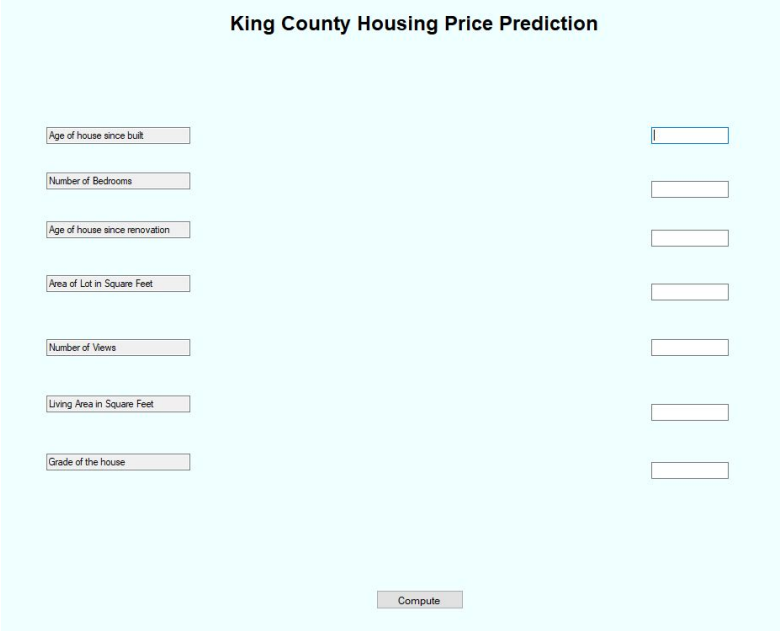
Statistics	Value
R^2	0.64
F value p value	3057.01 0.0001
Parameters	✓
Multicollinearity	✓
Error Analysis	✓
Residual Analysis	✓

Modeling: Validation



Implementation: Prediction Software

- We created a prediction software using Visual Studio with the best model we created.



IT'S TIME FOR DEMO

THANK YOU!!