Module 2 Individual Project Presentation

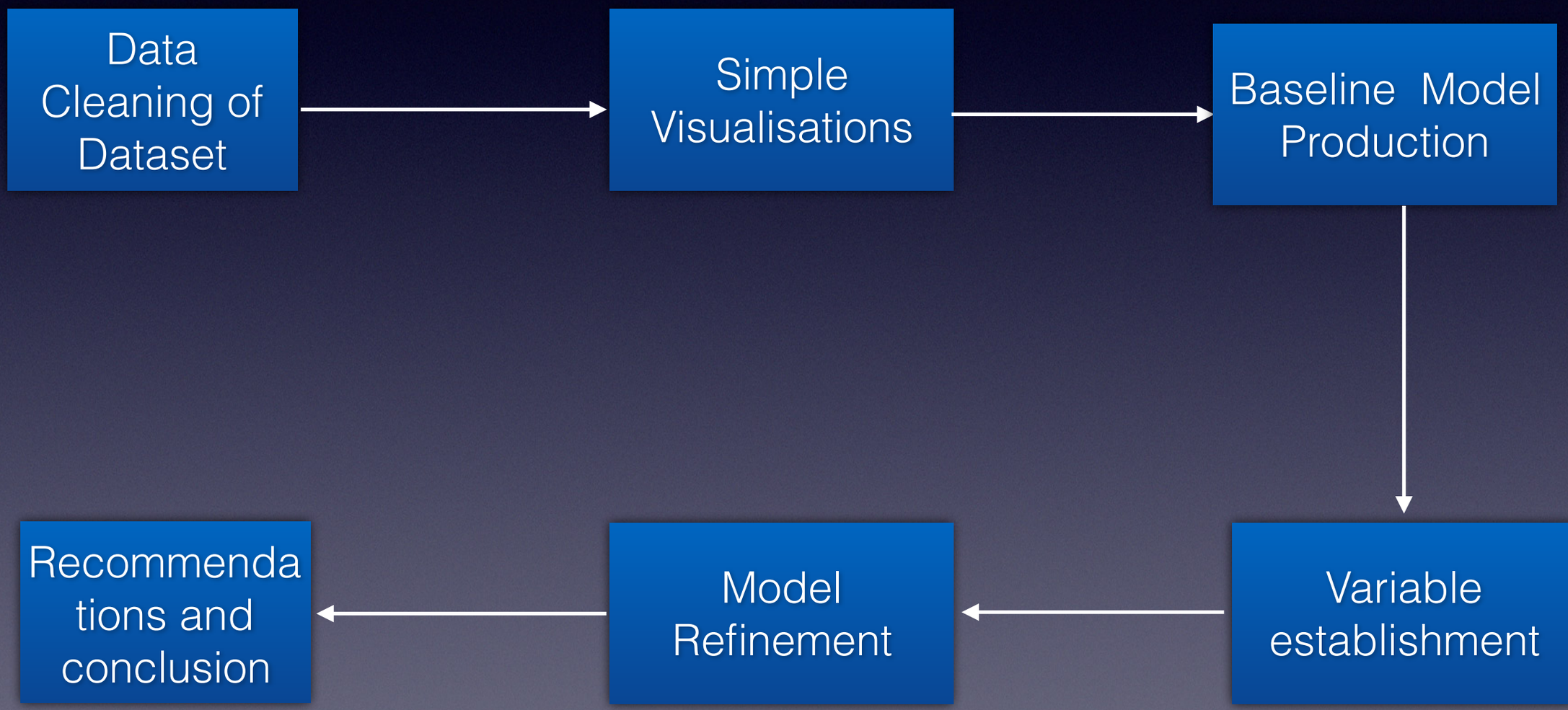
Arman Hussain

Introduction

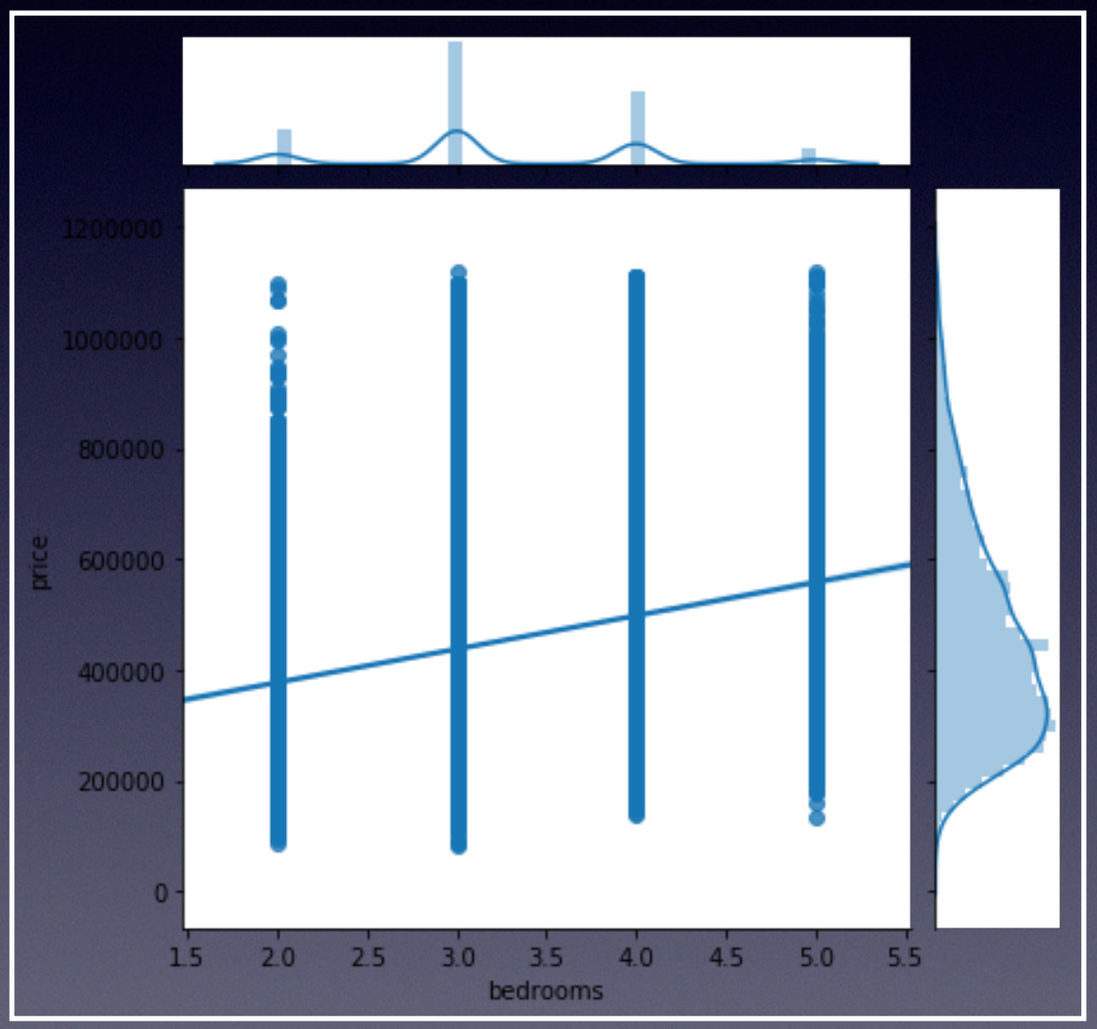
- Working with the King County House Sales Dataset model.

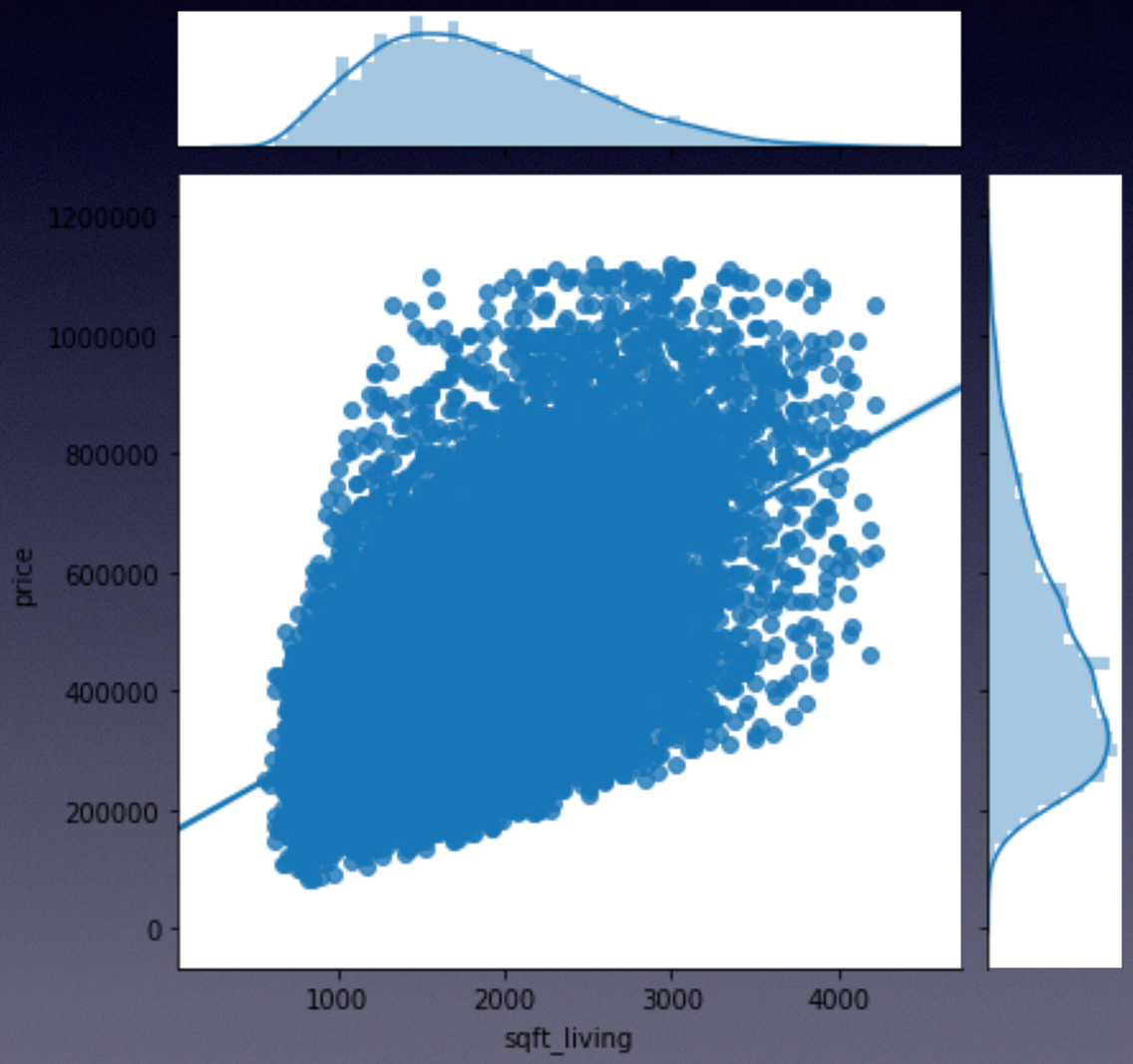
- Clean, explore, and model the data set to predict the sale price of houses

Method


Data Cleaning of Dataset
Simple Visualisations
Baseline Model Production
Variable establishment
Model Refinement
Recommendations and conclusion

Initial Visualizations





Business Questions

- 1) How is the price of property affected by standard factors customers typically look for including baths/bed/square foot coverage/ square foot of the lot/and grade of the property?
- 2) What is the relationship between the number of bedrooms and price of property?
- 3) How does the year of the property built effect the price of the property?

Solutions

Dep. Variable:	price	**R-squared:**	0.908
Model:	OLS	**Adj. R-squared:**	0.908
Method:	Least Squares	**F-statistic:**	1.361e+04
Date:	Fri, 27 Mar 2020	**Prob (F-statistic):**	0.00
Time:	04:26:37	**Log-Likelihood:**	-1.2881e+05
No. Observations:	9661	**AIC:**	2.576e+05
Df Residuals:	9654	**BIC:**	2.577e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
bedrooms	2189.9987	2194.231	0.998	0.318	-2111.154	6491.152
bathrooms	-1.92e+04	2570.667	-7.467	0.000	-2.42e+04	-1.42e+04
sqft_living	253.8077	7.855	32.314	0.000	238.411	269.204
sqft_lot	-9.2313	0.530	-17.404	0.000	-10.271	-8.192
waterfront	-1.199e-12	5.76e-13	-2.083	0.037	-2.33e-12	-7.08e-14
grade	7.958e+04	2438.017	32.643	0.000	7.48e+04	8.44e+04
rooms	-1.701e+04	1344.248	-12.651	0.000	-1.96e+04	-1.44e+04
square_foot_living	-9600.7794	778.076	-12.339	0.000	-1.11e+04	-8075.586

Question 1

Dep. Variable:	y	**R-squared:**	0.363
Model:	OLS	**Adj. R-squared:**	0.363
Method:	Least Squares	**F-statistic:**	2869.
Date:	Fri, 27 Mar 2020	**Prob (F-statistic):**	0.00
Time:	04:27:35	**Log-Likelihood:**	-18016.
No. Observations:	15097	**AIC:**	3.604e+04
Df Residuals:	15093	**BIC:**	3.607e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.633e-17	0.006	-1.18e-14	1.000	-0.013	0.013
z_bedrooms	-0.0888	0.008	-10.693	0.000	-0.105	-0.073
z_sqft_living	0.3850	0.010	37.860	0.000	0.365	0.405
z_grade	0.3287	0.008	39.047	0.000	0.312	0.345

Question 2

Dep. Variable:	price	**R-squared:**	0.292
Model:	OLS	**Adj. R-squared:**	0.292
Method:	Least Squares	**F-statistic:**	3112.
Date:	Fri, 27 Mar 2020	**Prob (F-statistic):**	0.00
Time:	01:34:06	**Log-Likelihood:**	-2.0238e+05
No. Observations:	15097	**AIC:**	4.048e+05
Df Residuals:	15094	**BIC:**	4.048e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.327e+05	7603.421	-17.453	0.000	-1.48e+05	-1.18e+05
time_since_1960	1151.4350	77.711	14.817	0.000	999.112	1303.758
square_foot_living	1.312e+04	179.218	73.219	0.000	1.28e+04	1.35e+04

Question 3

Recommendations

- Ensure high standards of quality control in newly built homes
- Give customers a better insight of the year built of the property
- Houses with more bedrooms and lower property value require refurbishment.

Future Work

- Be able to use the zip code column as areas with greater wealth and opportunity would typically yield better house prices.

- Be able to use the yr_renovated column and this would similarly yield better house price data.

- Be able to use the lat/long column to map out the geographical location of the homes in King County and display the effect on house price data.

Thank You For Your Time