KING COUNTY HOUSING

Analysis by Christopher Freyre

CONTENTS

BUSINESS PROBLEM

DATA UNDERSTANDING

DATA CLEANING

MODEL AND VALIDATION

BUSINESS PROBLEM

- Real-State agencies are looking to advise home owners on how performing a renovation might increase price of home sales and if so, what factors are relevant to this renovation.
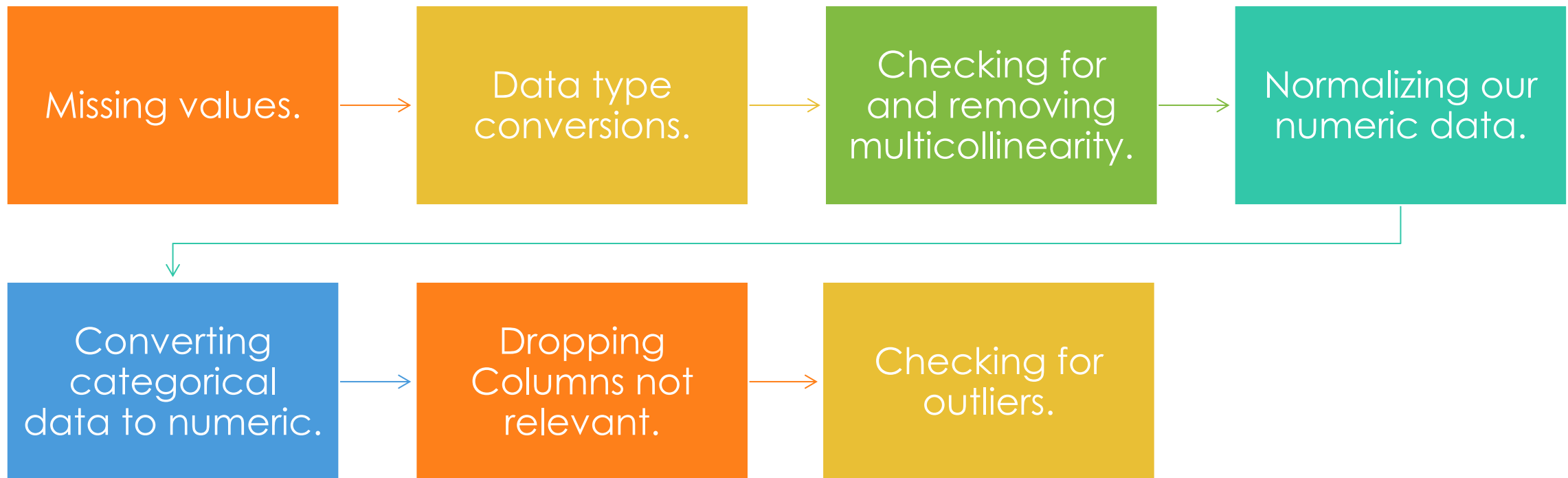
DATA UNDERSTANDING

- House sales data in King County in 2014-2015
- Price
- Id & date
- Bedrooms, Bathrooms & Floors
- Square foot: Living, Lot, Above, Basement, 15
- Grade & Condition
- Area: Zip code, Lat & Long
- Year: built & renovated
- Waterfront
- Views

Out[4]:

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	yr_built	yr_re
0	7129300520	10/13/2014	221900.0	3	1.00	1180	5650	1.0	NaN	0.0	...	7	1180	0.0	1955	
1	6414100192	12/9/2014	538000.0	3	2.25	2570	7242	2.0	0.0	0.0	...	7	2170	400.0	1951	
2	5631500400	2/25/2015	180000.0	2	1.00	770	10000	1.0	0.0	0.0	...	6	770	0.0	1933	
3	2487200875	12/9/2014	604000.0	4	3.00	1960	5000	1.0	0.0	0.0	...	7	1050	910.0	1965	
4	1954400510	2/18/2015	510000.0	3	2.00	1680	8080	1.0	0.0	0.0	...	8	1680	0.0	1987	
5	7237550310	5/12/2014	1230000.0	4	4.50	5420	101930	1.0	0.0	0.0	...	11	3890	1530.0	2001	
6	1321400060	6/27/2014	257500.0	3	2.25	1715	6819	2.0	0.0	0.0	...	7	1715	?	1995	
7	2008000270	1/15/2015	291850.0	3	1.50	1060	9711	1.0	0.0	NaN	...	7	1060	0.0	1963	
8	2414600126	4/15/2015	229500.0	3	1.00	1780	7470	1.0	0.0	0.0	...	7	1050	730.0	1960	
9	3793500160	3/12/2015	323000.0	3	2.50	1890	6560	2.0	0.0	0.0	...	7	1890	0.0	2003	

10 rows × 21 columns

DATA CLEANING PROCESS

Missing values. → Data type conversions. → Checking for and removing multicollinearity. → Normalizing our numeric data.

Converting categorical data to numeric. → Dropping Columns not relevant. → Checking for outliers.
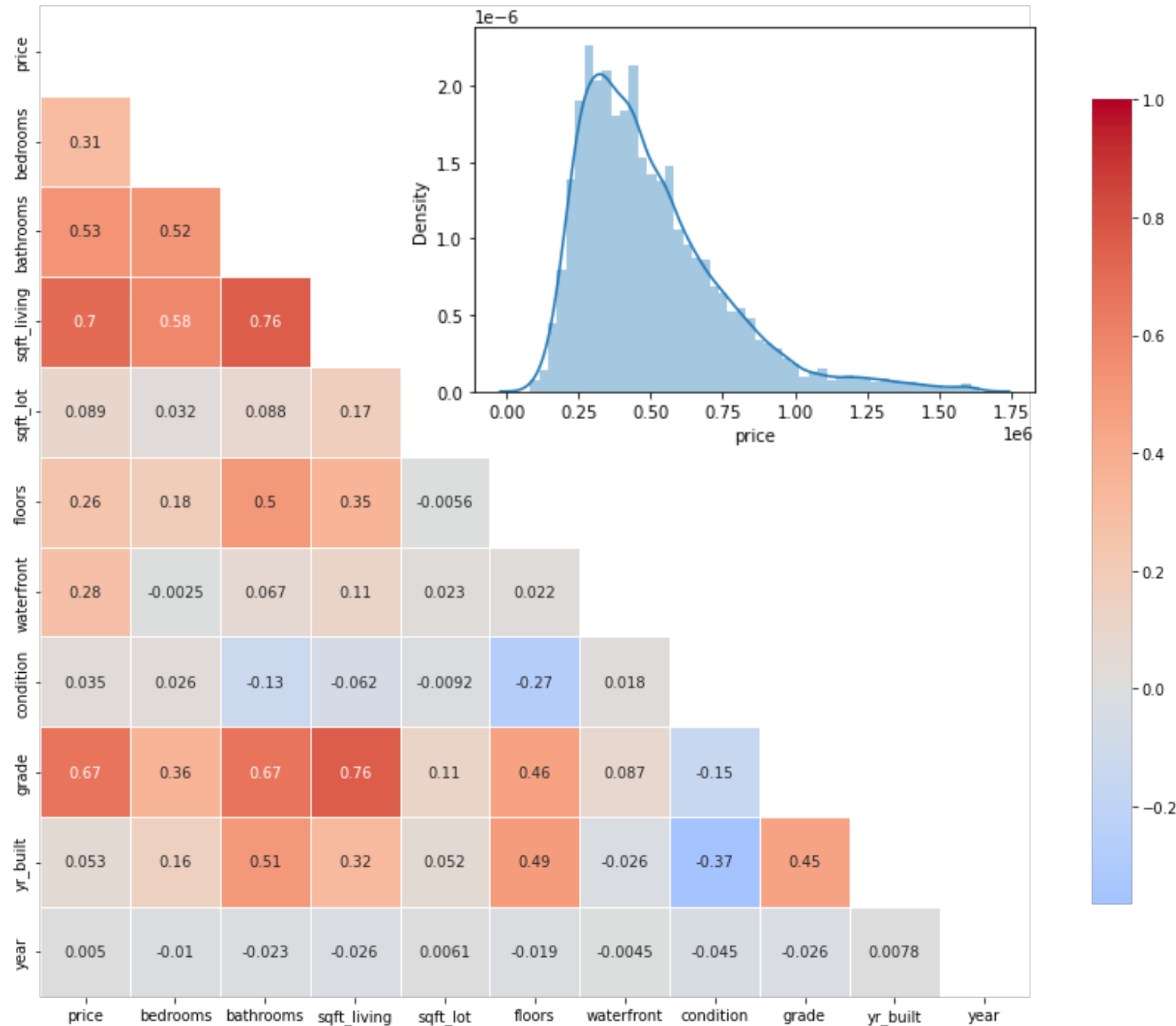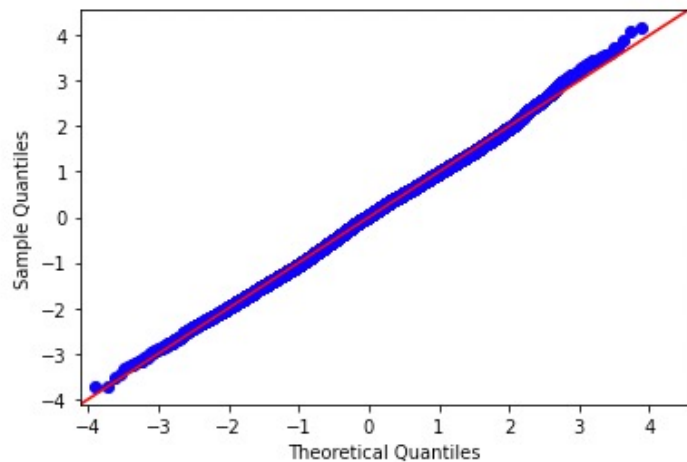
MODEL AND VALIDATION

- Correlations
- Dependent and independent variables
- Model 1, 2 and 3
- Dummy variables
- Log Transformation

VALIDATION

- Checking R-squared value
- Coeff values
- T-Test / Mean square error
- Cross validation
- Assumption of linear regression

Dep. Variable:	price_log	R-squared:	0.526
Model:	OLS	Adj. R-squared:	0.526
Method:	Least Squares	F-statistic:	1130.
Date:	Sun, 23 Oct 2022	Prob (F-statistic):	0.00
Time:	18:10:15	Log-Likelihood:	-6220.7
No. Observations:	20389	AIC:	1.248e+04
Df Residuals:	20368	BIC:	1.265e+04
Df Model:	20		
Covariance Type:	nonrobust		

CONCLUSION

- Model 3 provided the best fit for linear regression with an R- Squared value of 0.526. Meaning that it represents 52.6% of the data.

- Validating it with T-test, provided a Mean Squared Error value: 0.10 (MSE)

- Cross validation resulted in 0.10

- Highest Coeff contributing to a high price by high condition and grade, implying a quality renovation is key to sell at a higher price :
  - Condition 5 = 1.5164 Coeff
  - Grade 11 = 1.2728 Coeff

RECOMMENDATION

I would recommend Real Estate agencies advise their clients that:

- Price of properties increase by square foot living, bedrooms and bathrooms.

- Renovations that provide a high condition rate of 5, or at least in a range 2 to 5.

- Renovations that provide a high grade rate of 11, or at least in a range 5 to 11.

THANK YOU

- Email: Christopher.Freyre@gmail.com
- Github: @Chrisfreyre
- Linkedin: https://www.linkedin.com/in/christopher-freyre-56760b136