Predicting House Prices with the King's County Housing Data Set

Module 1: Exploratory Data Analysis

-Sarabi Eventide

Methodology Overview

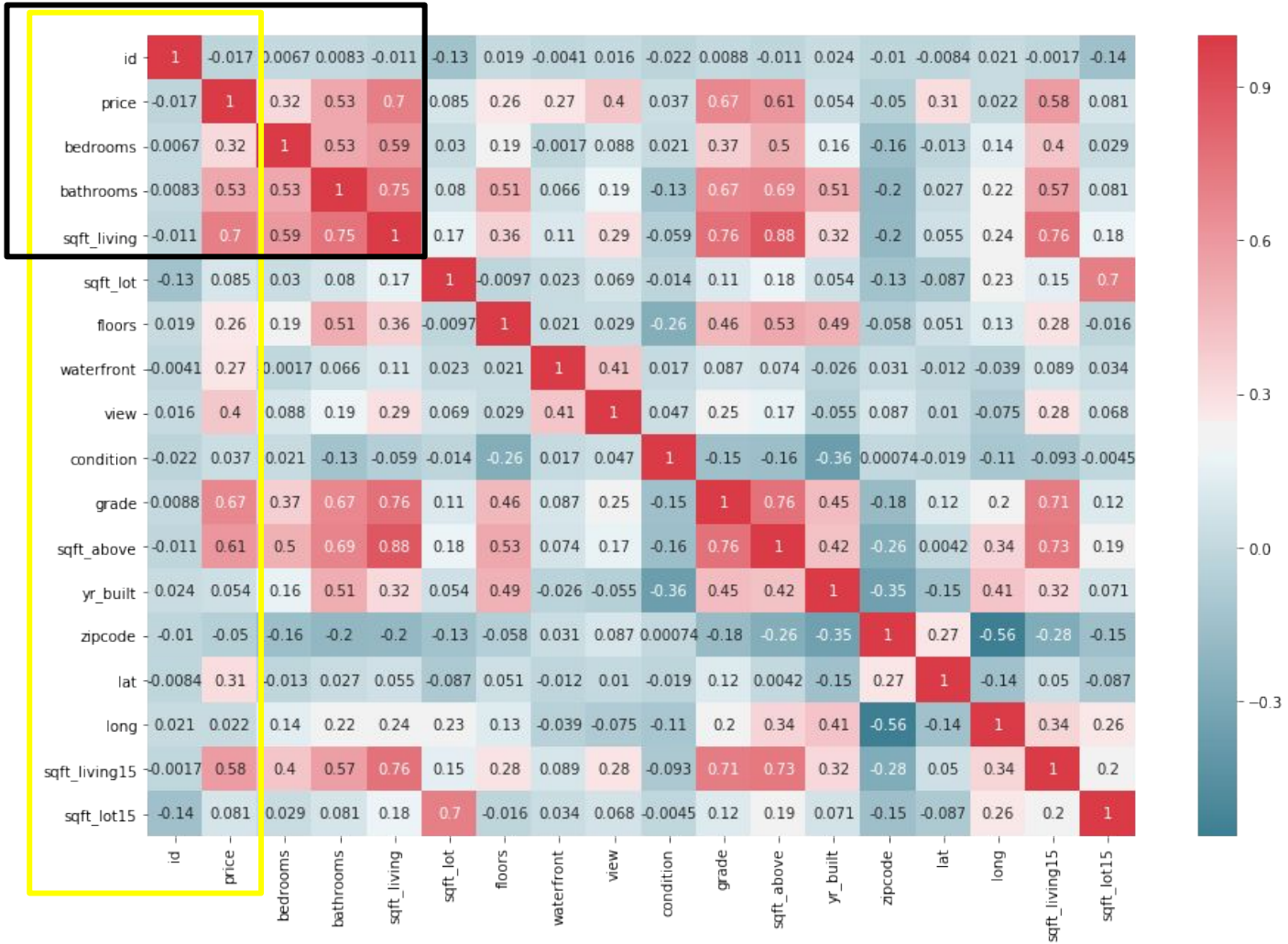
1. Clean the Data
2. Explore the Relationships
3. Normalize the Data
4. Model the Data
5. Make Predictions

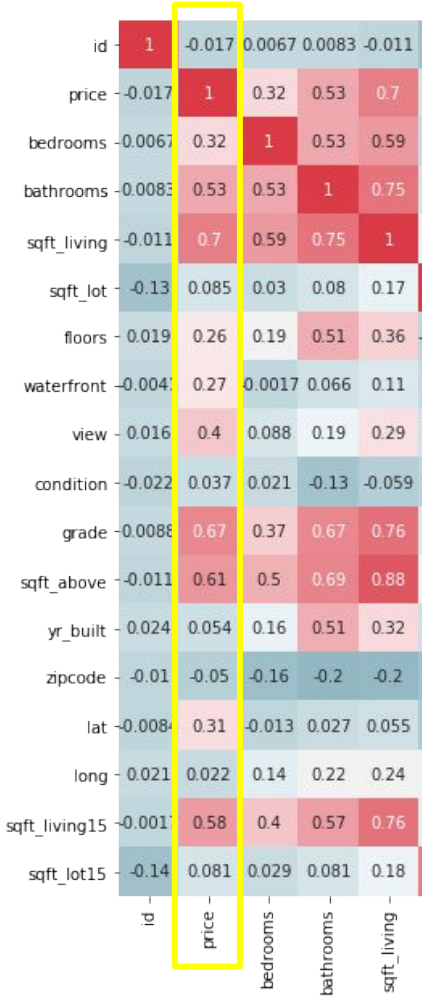
1. Clean the Data

- Eliminate obvious outliers
- Determine the presence of categorical variables
- Eliminate or replace null values
    - Explore “waterfront”

2. Explore the relationships

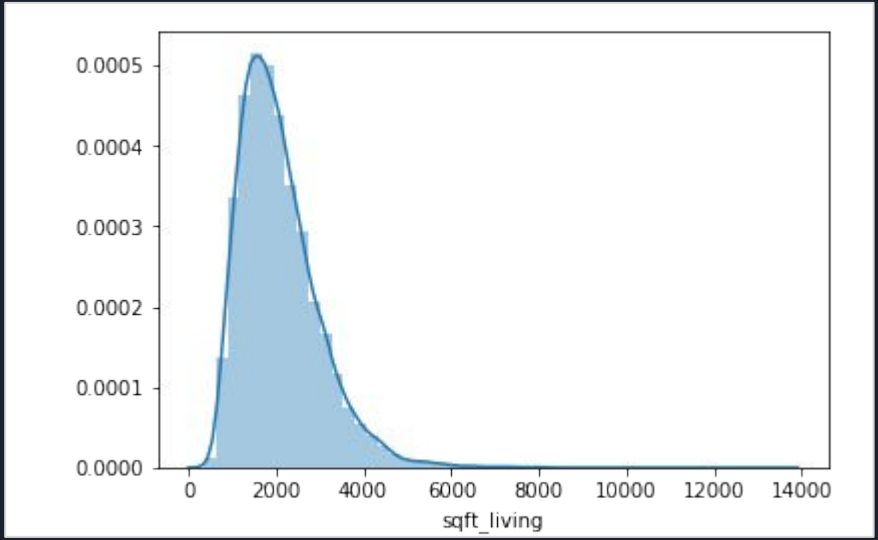
- Check distribution of housing prices for reference
- Correlation heatmap: how features work together

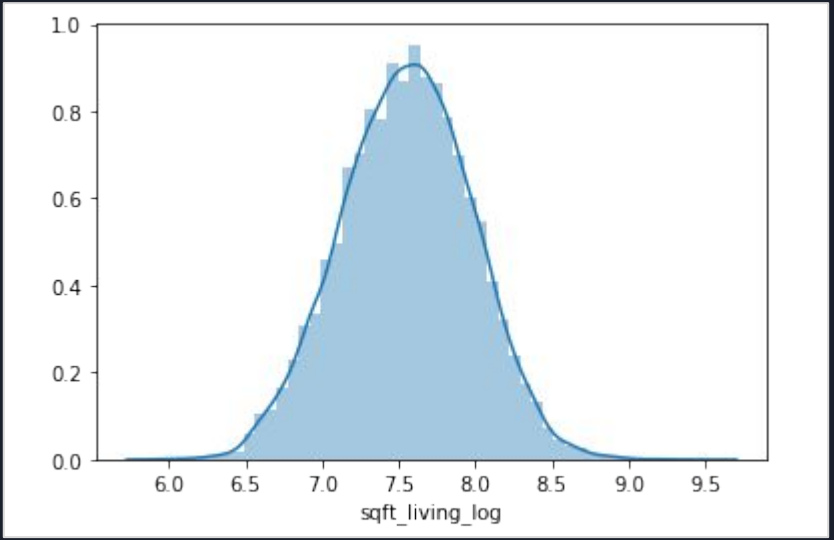

id 1 -0.017 0.0067 0.0083 -0.011 -0.13 0.019 -0.0041 0.016 -0.022 0.0088 -0.011 0.024 -0.01 -0.0084 0.021 -0.0017 -0.14
price -0.017 1 0.32 0.53 0.7 0.085 0.26 0.27 0.4 0.037 0.67 0.61 0.054 -0.05 0.31 0.022 0.58 0.081
bedrooms 0.0067 0.32 1 0.53 0.59 0.03 0.19 -0.0017 0.088 0.021 0.37 0.5 0.16 -0.16 -0.013 0.14 0.4 0.029
bathrooms 0.0083 0.53 0.53 1 0.75 0.08 0.51 0.066 0.19 -0.13 0.67 0.69 0.51 -0.2 0.027 0.22 0.57 0.081
sqft_living -0.011 0.7 0.59 0.75 1 0.17 0.36 0.11 0.29 -0.059 0.76 0.88 0.32 -0.2 0.055 0.24 0.76 0.18
sqft_lot -0.13 0.085 0.03 0.08 0.17 1 -0.0097 0.023 0.069 -0.014 0.11 0.18 0.054 -0.13 -0.087 0.23 0.15 0.7
floors 0.019 0.26 0.19 0.51 0.36 -0.0097 1 0.021 0.029 -0.26 0.46 0.53 0.49 -0.058 0.051 0.13 0.28 -0.016
waterfront 0.0041 0.27 0.0017 0.066 0.11 0.023 0.021 1 0.41 0.017 0.087 0.074 -0.026 0.031 -0.012 -0.039 0.089 0.034
view 0.016 0.4 0.088 0.19 0.29 0.069 0.029 0.41 1 0.047 0.25 0.17 -0.055 0.087 0.01 -0.075 0.28 0.068
condition -0.022 0.037 0.021 -0.13 -0.059 -0.014 -0.26 0.017 0.047 1 -0.15 -0.16 -0.36 0.00074 -0.019 -0.11 -0.093 -0.0045
grade 0.0088 0.67 0.37 0.67 0.76 0.11 0.46 0.087 0.25 -0.15 1 0.76 0.45 -0.18 0.12 0.2 0.71 0.12
sqft_above -0.011 0.61 0.5 0.69 0.88 0.18 0.53 0.074 0.17 -0.16 0.76 1 0.42 -0.26 0.0042 0.34 0.73 0.19
yr_built 0.024 0.054 0.16 0.51 0.32 0.054 0.49 -0.026 -0.055 -0.36 0.45 0.42 1 -0.35 -0.15 0.41 0.32 0.071
zipcode -0.01 -0.05 -0.16 -0.2 -0.2 -0.13 -0.058 0.031 0.087 0.00074 -0.18 -0.26 -0.35 1 0.27 -0.56 -0.28 -0.15
lat -0.0084 0.31 -0.013 0.027 0.055 -0.087 0.051 -0.012 0.01 -0.019 0.12 0.0042 -0.15 0.27 1 -0.14 0.05 -0.087
long 0.021 0.022 0.14 0.22 0.24 0.23 0.13 -0.039 -0.075 -0.11 0.2 0.34 0.41 -0.56 -0.14 1 0.34 0.26
sqft_living15 0.0017 0.58 0.4 0.57 0.76 0.15 0.28 0.089 0.28 -0.093 0.71 0.73 0.32 -0.28 0.05 0.34 1 0.2
sqft_lot15 -0.14 0.081 0.029 0.081 0.18 0.7 -0.016 0.034 0.068 -0.0045 0.12 0.19 0.071 -0.15 -0.087 0.26 0.2 1
id price bedrooms bathrooms sqft_living sqft_lot floors waterfront view condition grade sqft_above yr_built zipcode lat long sqft_living15 sqft_lot15
0.9 0.6 0.3 0.0 −0.3


id 1 -0.017 0.0067 0.0083 -0.011
price -0.017 1 0.32 0.53 0.7
bedrooms -0.0067 0.32 1 0.53 0.59
bathrooms -0.0083 0.53 0.53 1 0.75
sqft_living -0.011 0.7 0.59 0.75 1
sqft_lot -0.13 0.085 0.03 0.08 0.17
floors 0.019 0.26 0.19 0.51 0.36
waterfront -0.004 0.27 -0.0017 0.066 0.11
view 0.016 0.4 0.088 0.19 0.29
condition -0.022 0.037 0.021 -0.13 -0.059
grade -0.0088 0.67 0.37 0.67 0.76
sqft_above -0.011 0.61 0.5 0.69 0.88
yr_built 0.024 0.054 0.16 0.51 0.32
zipcode -0.01 -0.05 -0.16 -0.2 -0.2
lat -0.008 0.31 -0.013 0.027 0.055
long 0.021 0.022 0.14 0.22 0.24
sqft_living15 -0.001 0.58 0.4 0.57 0.76
sqft_lot15 -0.14 0.081 0.029 0.081 0.18
id
price
bedrooms
bathrooms
sqft_living

3. Normalize Data

-Features: square footage of house, number of bathrooms, square footage of the house in 2015 above ground space





4. Model the Data

1. OLS regression
2. RFE
3. Train, test, split

Findings