Final project in Module 2

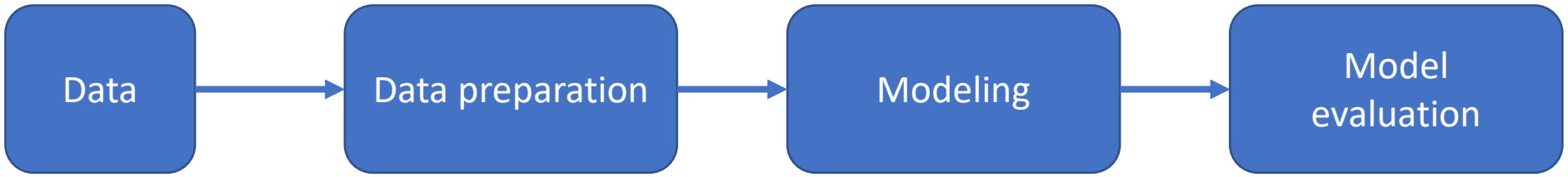
Guofa Shou

Self-paced

Outline

- The aim of the project:
  - To provide suggestions regarding the price of house
    - For house buyer, they will know the approximate price of the house and the investment value of the house for future sale, based on the characteristics of the house
    - For house seller, they will know what they need to do to sell the house with a better price
- A linear regression model is built based on the selected house characteristics to achieve the aim of the project
- Conclusion

A linear regression model



Data

- Data source: kc_house_data.csv
- Data understanding:
  - A total of 19 predictors available after excluding the id and the target (price)
  - A total of 21597 rows, while some rows have null values in some predictors
  - Several predictors' data type need to be changed, e.g., date and sqft_basement

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21597 entries, 0 to 21596
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             21597 non-null  int64
 1   date           21597 non-null  object
 2   price          21597 non-null  float64
 3   bedrooms       21597 non-null  int64
 4   bathrooms      21597 non-null  float64
 5   sqft_living    21597 non-null  int64
 6   sqft_lot       21597 non-null  int64
 7   floors         21597 non-null  float64
 8   waterfront     19221 non-null  float64
 9   view           21534 non-null  float64
 10  condition      21597 non-null  int64
 11  grade          21597 non-null  int64
 12  sqft_above     21597 non-null  int64
 13  sqft_basement  21597 non-null  object
 14  yr_built       21597 non-null  int64
 15  yr_renovated   17755 non-null  float64
 16  zipcode        21597 non-null  int64
 17  lat            21597 non-null  float64
 18  long           21597 non-null  float64
 19  sqft_living15  21597 non-null  int64
 20  sqft_lot15     21597 non-null  int64
dtypes: float64(8), int64(11), object(2)
memory usage: 3.5+ MB
```

Data Preparation

- Deal with data types
- Deal with missing values
- Deal with outliers
- Deal with categorical variables
- Deal with target variable: price

Modeling

- Final regression model after excluding some predictors with high collinearity and non-significance

Predictors: 56 -> 30

OLS Regression Results

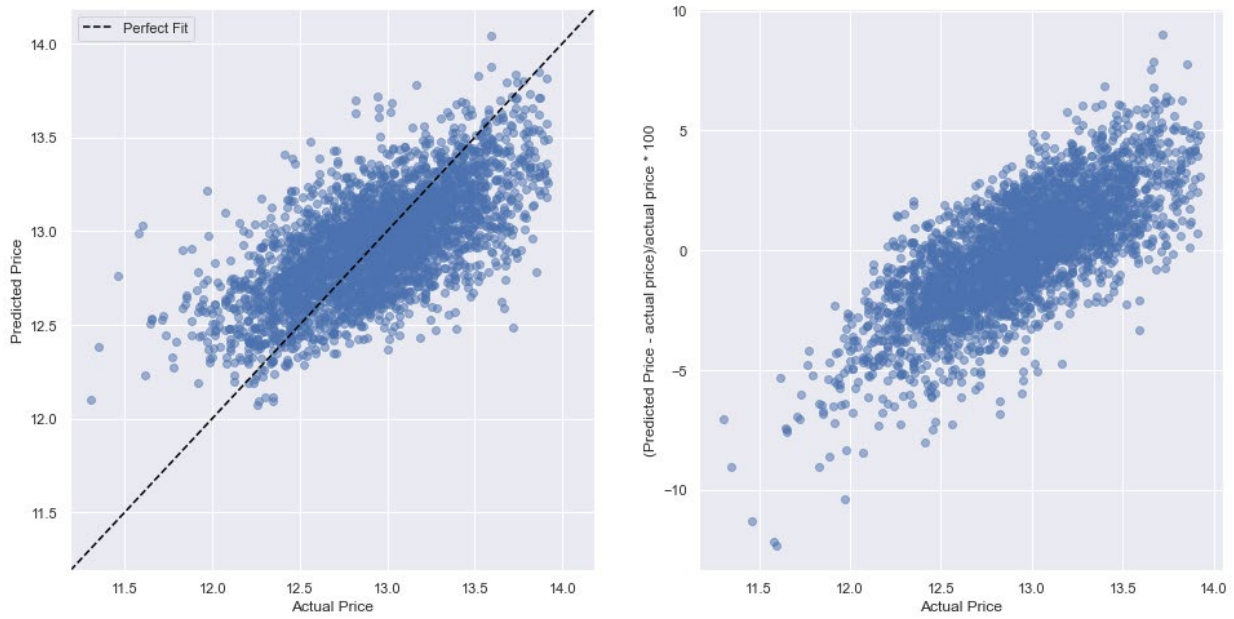
Dep. Variable:	price_log1p	**R-squared:**	0.449
Model:	OLS	**Adj. R-squared:**	0.447
Method:	Least Squares	**F-statistic:**	342.6
Date:	Thu, 18 Nov 2021	**Prob (F-statistic):**	0.00
Time:	23:18:42	**Log-Likelihood:**	-3325.9
No. Observations:	12664	**AIC:**	6714.
Df Residuals:	12633	**BIC:**	6945.
Df Model:	30		
Covariance Type:	nonrobust		

```
['sqft_lot',
 'sqft_living15',
 'sqft_lot15',
 'bedrooms_3',
 'bedrooms_4',
 'bedrooms_5',
 'floors_2.5',
 'floors_3.0',
 'waterfront_1.0',
 'is_renovated_1.0'
 'view_1.0',
 'view_2.0',
 'view_3.0',
 'view_4.0',
 'grade_10',
 'grade_11',
 'zipcode4_98020',
 'zipcode4_98030',
 'zipcode4_98040',
 'zipcode4_98060',
 'zipcode4_98070',
 'zipcode4_98090',
 'zipcode4_98100',
 'zipcode4_98110',
 'zipcode4_98120',
 'zipcode4_98130',
 'zipcode4_98150',
 'zipcode4_98160',
 'zipcode4_98170',
 'zipcode4_98180']
```

Model evaluation

- Cross-validation
  - Train score: 0.4465
  - Test score: 0.4501

Predicted values vs. the real values



The fitted regression model can predict house price very well

Conclusion

- Observations from coefficients
  - The grade and sqft_living15 have the strongest relationship with the house price
  - It is interesting to see sqft_lot15 has a negative relationship with the house price
  - Waterfront_1.0 and grade_11 also have a positive relationship with the price
  - For some zip codes, e.g., 98100 and 98110, they have high positive relationships with the price
- To address the business questions:
  - For buyer, they will know that the house price is higher for a house with the high grade and sqrt_living15 values
  - For seller, if they want to sell their house with a higher price, they could add waterfront and improve the grade.

```
sqft_lot            0.001943
sqft_living15       1.096717
sqft_lot15         -0.144789
bedrooms_3          0.083760
bedrooms_4          0.149979
bedrooms_5          0.163189
floors_2.5          0.226109
floors_3.0          0.056801
waterfront_1.0      0.398927
is_renovated_1.0    0.192348
view_1.0            0.137734
view_2.0            0.148326
view_3.0            0.117277
view_4.0            0.243554
grade_10            0.258860
grade_11            0.418129
zipcode4_98020     -0.112151
zipcode4_98030     -0.086005
zipcode4_98040     -0.156698
zipcode4_98060     -0.055208
zipcode4_98070      0.128485
zipcode4_98090     -0.420847
zipcode4_98100      0.304544
zipcode4_98110      0.355810
zipcode4_98120      0.233377
zipcode4_98130      0.149527
zipcode4_98150      0.056281
zipcode4_98160     -0.225556
zipcode4_98170     -0.097728
zipcode4_98180     -0.279241
Name: Coefficients, dtype: float64

Intercept: 12.30344679761995
```