


Machine Learning





PROJECT ASSIGNMENT


STUDENT NAME: GATTUOCH CHAMBANG TENY


ID:1401298

- 
- ▶ Table of Contents
 - ▶ 1 Introduction
 - ▶ 1.1 Problem Statement
 - ▶ 1.2 Data Structure
 - ▶ 1.3 Source of Data
 - ▶ 1.4 Approach and methodology
 - ▶ 2 Import Necessary Library or Obtain Data
 - ▶ 2.1 Load data
 - ▶ 3 Exploratory Data Analysis(EDA)
 - ▶ 3.1 Investigate Data Types
 - ▶ 3.2 Identity missing value
 - ▶ 3.3 histogram to understand the Distribution
 - ▶ 3.4 scatter plot to understand the relationship
 - ▶ 3.5. correlation with headmap to interpret the relation and multicolliniarity

- 
- ▶ 4. Data Preprocessing
 - ▶ 4.1. Handle Missing Value
 - ▶ 4.2. Handle Inconsistencies
 - ▶ 4.3. Initial feature engineering
 - ▶ 4.4. Consider outliers
 - ▶ 4.5. Encode Categorical Features
 - ▶ 4.6. Scale or Normalize Numerical Features
 - ▶ 5 Q1: Location
 - ▶ 5.2 House prices per Zipcode
 - ▶ 5.3 Reverse Geocoding
 - ▶ 5.4 Location of most expensive houses
 - ▶ 5.5 Waterfront Feature
 - ▶ 5.6 Conclusion
 - ▶ 6 Q2: House attributes

- 
- ▶ 5.1 Map of house sales
 - ▶ 6.1 Overview of house features
 - ▶ 6.2 Exploring bedroom count
 - ▶ 6.3 How important is a large lot?
 - ▶ 6.4 What is building grade and how does it affect price?
 - ▶ 6.5 Conclusion
 - ▶ 7 Q3: Time
 - ▶ 7.1 Engineering relevant features
 - ▶ 7.2 Visualisations
 - ▶ 7.3 Conclusion
 - ▶ 8 Preparing data for modelling
 - ▶ 8.1 Investigate linearity assumption
 - ▶ 8.2 Investigate multicollinearity
 - ▶ 8.3 One-hot encoding
 - ▶ 8.4 Remove unnecessary features
 - ▶ 9 model implementation and Training
 - ▶ 9.1. Simple linear regression
 - ▶ 9.2. Train-Test Split

- 
- ▶ 9.3. Hyperparameter Tuning
 - ▶ 10 Model Evaluation and Analysis
 - ▶ 10.1. Make prediction on the testing Data
 - ▶ 10.2. Evaluate the Model's performance
 - ▶ 10.3. Base multiple linear regression model and evaluation
 - ▶ 10.4. Further one-hot encoding
 - ▶ 10.4 Model 1a with neighbourhood data
 - ▶ 10.5 Model 1b with zipcode data
 - ▶ 10.6 Establish zipcode tiers
 - ▶ 10.7 Model 1c with zipcode tiers
 - ▶ 10.8 Model 2 with feature scaling
 - ▶ 10.9 Model 3 with interactions
 - ▶ 10.10 Model 4 with polynomials

- 
- ▶ 10.11 Model 5 with 3 main features
 - ▶ 10.12 Recursive Feature Selection
 - ▶ 11 Interpret
 - ▶ 11.1 Model A
 - ▶ 11.2 Model B
 - ▶ 11.3 Model C
 - ▶ 11.4 Comparing B and C
 - ▶ 11.5 Predictions with test data
 - ▶ 11 Conclusion
 - ▶ 12.1 Summary of Findings and Recommendations
 - ▶ 12.2 Final Model considerations
 - ▶ 12.3 Future Work
 - ▶ 13 Appendix
 - ▶ 13.1 Presentation visualisations

Predicting House Prices

- ▶ **Introduction**

- ▶ **Problem Statement**

- ▶ As a junior data scientist at real estate company PropertiesInc., I have been tasked with investigating house sales in the King County area and building a model to predict the sale price. Key executives are keen to launch an advertising campaign directed towards home owners in that area who might consider selling their house, focusing on higher-end residential properties.
- ▶ Before building the model, we will address the following descriptive questions through data exploration:
 - ▶ **1. Which locations within the King County area have the highest average house prices?**
 - ▶ Understanding what locations to focus the advertising campaign on is key for our stakeholders.
 - ▶ **2. Which house attributes increase sale price?**
 - ▶ Understanding home buyers' preferences can focus our campaign and help us guide clients willing to undertake renovations prior to selling.
 - ▶ **3. Does time of the year have an impact on house sales?**
 - ▶ Understanding seasonal trends will influence when the campaign should be launched.
- ▶ I have been provided with a dataset with house sale prices in King County, Washington State, USA from May 2014 to May 2015 to use for this project.

. Data Structure

- ▶ A dataset has been provided and can be found in the dfclearn.csv file in this repository.
- ▶ The column names and descriptions as provided . For convenience they have been reproduced below
- ▶ Column Names and descriptions for Kings County Data Set
- ▶ id - unique identified for a house
- ▶ dateDate - house was sold
- ▶ pricePrice - is prediction target
- ▶ bedroomsNumber - of Bedrooms/House
- ▶ bathroomsNumber - of bathrooms/bedrooms
- ▶ sqft_livingsquare - footage of the home
- ▶ sqft_lotsquare - footage of the lot
- ▶ floorsTotal - floors (levels) in house
- ▶ waterfront - House which has a view to a waterfront
- ▶ view - Has been viewed
- ▶ condition - How good the condition is (Overall)
- ▶ grade - overall grade given to the housing unit, based on King County grading system
- ▶ sqft_above - square footage of house apart from basement
- ▶ sqft_basement - square footage of the basement
- ▶ yr_built - Built Year
- ▶ yr_renovated - Year when house was renovated
- ▶ zipcode - zip
- ▶ lat - Latitude coordinate
- ▶ long - Longitude coordinate
- ▶ sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbors
- ▶ sqft_lot15 - The square footage of the land lots of the nearest 15 neighbors

1.4. Approach and methodology

- 1 Import Necessary Library or Obtain Data
 - 2 . Exploratory Data Analysis(EDA)
 3. Data Preprocessing (clean the data, deal with missing values and data types)
 4. Explore (answer descriptive questions using EDA)
 5. model implementation and training
 6. Model Evaluation and analysis
- ▶ This notebook has been organised following the above, however note the following:
 - ▶ a holdout set was extracted at the start in section 1 to conduct a final testing of our model.
 - ▶ there is a separate section for each question in the explore phase to improve readability
 - ▶ there is an additional section between steps 3 and 4 in which we checked linear regression assumptions and performed some final data preparation prior to modeling.
 - ▶ feature engineering is performed throughout as and where needed. In the data exploration sections, a separate DataFrame was created for each question to freely manipulate.
 - ▶ we have taken an iterative approach to modelling starting with simple linear regression, then multiple linear regression, adding in interactions and finally polynomial regression.

