COMP SCI 7209 Big Data Analysis and Project
Project Title: Big Data: House Price Analysis and Prediction (Part A)
Ning Ni (a1869549)
Project Coordinator: Bernard Evans
1 June 2023

1. Introduction

Real estate market plays a significant role in the economic field. Congressional Research Service (2023, p.1) has reported that in the United States of 2021, the volume on property investment and services stood at approximately $3.9 trillion, constituting around 16.7% of GDP. Therefore, it's crucial for investors to gain a profound insight of the house price trends.

This project aims to analyze and evaluate the impact of attributes of a house on the price of the house. Build appropriate models to predict housing prices.

2. Initial Objective

RQ1: Identify house features that exhibit obvious changes along with variations in house prices. Visualize and explain the correlation between prices and house attributes.

RQ2: See if a predictive model of house price can be made with or without house structure. Observe if there are differences in outcome of the model.

RQ3: Utilize models to quantify the potential impact on house price by adding additional functional zones like bedrooms, while keeping other features in the same condition.

RQ4: Put house prices into low, medium and high level. Utilize house location and geographic attributes to classify house with price level.

3. Data Source

3.1 Data Description

The data source used in this study is the Ames, Iowa housing dataset edited by Dean De Cock (2011, p.4).

It includes the sale records of properties from 2006 to 2010. It comprises 2,930 observations. According to Wikipedia (2023), the sample size represents approximately 13% of the total number of households in Ames. It contains 80 explanatory variables, including 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables.

3.2 Data Assessment

3.2.1 Data Feasibility

- Dean De Cock (ibid., p.3) mentions the dataset was originally obtained from the Ames City Assessor's Office. He preliminarily performed data cleaning and made edits to ensure that it’s easily comprehensible for individuals of all skill levels.
- Ames is a typical small to medium-sized city in the United States. The dataset could reflect the general real estate situation in similar cities.
- The dataset contains various of house features including location, structure, materials, geographical factors, amenities, etc. Thus, the dataset can support the research goals of the project.

3.2.2 Data Limitation

- The dataset only covers the period from 2006 to 2010, which may not fully capture recent market dynamics or changes in housing trends.
- The dataset represents a specific size city, and the findings based on this dataset may not directly generalize to metropolitan region.

3.3 Data Limitation Processing

To address the data limitations, the project will incorporate an additional data source called "Melbourne Housing Market" (Kaggle, 2018). It comprises 34,858 records of housing sales in Melbourne from 2016 to 2017. It contains 21 attributes including location and structural information. Although the Melbourne dataset cannot completely address the limitation of Ames dataset, analyzing it will provide a supplementary opportunity to investigate how the location and structural characteristics of houses have influenced housing prices in large urban areas in recent years.

4. Refined Objective

Through section 3.2, the refined questions are listed as follows:

- Setting temporal and geographic prerequisites for RQ1 to RQ4, the project focuses on the time period between 2006 and 2010 and specifically targets small and medium-sized cities.

Section 3.3 also leads to a backup question:

- Has the impact of housing location and structure on housing prices differed over the past 10 years and in recent years, and between large cities and small to medium-sized cities?

5. Detailed Plan

1. Tools:

Sebastian & Vahid (2019) assert that Python is a powerful and accessible language for big data analysis. Therefore, this project will utilize Python3.9 as the main development language, along with packages such as NumPy, Pandas, Matplotlib, and Scikit-learn.

2. Data Cleaning:

Perform data cleaning and processing, including handling missing values, outliers and duplicates.

3. Exploratory Data Analysis (EDA):

Knaflic (2015, p.18-42) emphasizes EDA enables researchers to gain valuable insights into the data, address data quality issues, generate new questions and hypotheses, and support data-driven decision-making. Hence, this project will conduct summary statistics and visualization methods to explore the relationship between house price and house attributes.

4. Variable Selection:

As noted by Cai et al. (2018, p.70), feature selection aims at eliminating irrelevant and redundant features, which results in significant advantages when dealing with high-dimensional datasets. Accordingly, the project will employ feature selection techniques to extract pertinent house attributes for analysis.

5. Data Split:

Split the dataset into training and testing sets for fitting and testing models.

6. Fit Model:

RQ2 and RQ3 are regression problems. RQ4 is a classification problem. In term of different problem, the project may use lasso regression, knn, random forest, SVM, and XGBoost, etc.

7. Model Evaluation:

Calculate error metrics, such as MSE, MAE or R-squared, to assess the prediction models.

8. Result Interpretation and Reporting:

Answer the question based on the outcomes of analysis. Provide key statistical indicators to support conclusions and decision-making.

Reference

Congressional Research Service, 2023. *Introduction to U.S. Economy: Housing Market*.[pdf] Available at: <https://sgp.fas.org/crs/misc/IF11327.pdf> [Accessed 3 June 2023].

De Cock, D. 2011. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3).

Wikipedia: The free encyclopedia, 2023. *Ames, Iowa*. [online] Available at: <https://en.wikipedia.org/wiki/Ames,_Iowa#Demographics > [Accessed 3 June 2023].

Kaggle, 2018. *Melbourne Housing Market*. [online] Available at: <https://www.kaggle.com/anthonypino/melbourne-housing-market.> [Accessed 6 June 2023].

Hilber, C. A., & Vermeulen, W. 2016. The impact of supply constraints on house prices in England. *The Economic Journal*, 126(591), pp.358-405.

Sebastian R., Vahid M., 2019. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition*. Birmingham, England ; Mumbai: Packt.

Knaflic, C. N., 2015. *Storytelling with Data : A Data Visualization Guide for Business Professionals*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Cai, J., Luo, J., Wang, S., & Yang, S. 2018. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300, pp.70-79.