Capstone Project 1 Data Wrangling Report by Hye Joo Han

Overview

The main dataset, House Sales in King County, USA, was from Kaggle (Link: https://www.kaggle.com/harlfoxem/housesalesprediction). This data contains house sale prices for King County included in the Seattle–Tacoma–Bellevue metropolitan area in Washington. The dataset has one year of records for houses sold between May 2014 and May 2015. It consists of 19 house features, house id and sale price for 21,613 observations. The house features include the number of bedrooms and bathrooms, square footages of home and lot, year built, zip code, latitude and longitude. Zip codes or latitude and longitude can give some pieces of information about the area where each house is located. They can help house price predictions since location of a house is one of the most important factors when buying houses. I found niche.com (link: https://www.niche.com) provides grades for public school, safety, cost of living, jobs, commute, etc for each entered zip code. I collected those grades for relevant zip codes using web scraping and made a structured dataset. After cleaning both datasets, the collected zip code dataset was merged into the main dataset. In other words, additional columns derived from zip codes were added to the original house dataset.

Answers to the key questions

1. What kind of cleaning steps did you perform?

I took 3 major cleaning steps.

(1) House data
In this step, I cleaned the main dataset which has house price and features for each sale. First, I checked overall information, description and plots of the dataset. There was no missing value, but some columns had unideal data types and suspicious values. I checked each of those columns to change data types and take care of outliers.

(2) Zip code data
In this step, I gathered grades (from A+ to D-) for 12 categories including public school, safety, cost of living, jobs, commute, etc for each zip code by scraping Niche.com. I made a dataframe with column names in the format consistent to the main dataset. I also removed one column with the same value for all zip codes. Finally, I transformed the alphabet grades into score grades for later analysis. I took care of some missing grades.

(3) Merge zip code data into house data
After cleaning those dataframes in step (1) and (2), I merged the zip code dataset into the main dataset. The final dataframe as well as the cleaned dataframes in step(1) and (2) were saved as csv files.

2. How did you deal with missing values, if any?

In step (2), I found the diversity category has the grade NG for 4 zip codes. Niche.com says they do not grade places with insufficient data and NG means the area was ungraded. When I transformed alphabet grades into score grades, I mapped NG to None. Finally, the missing values with None were filled with the median diversity grade because the zip code areas with those missing values have normal grades for all other categories.

3. Were there outliers, and how did you handle them?

I found some outliers in the main dataset using the describe() function and box plots. Most of them were plausible outliers except for one with 33 bedrooms. The house only has 1.75 bathrooms and 1,620 square footage and 1 floor. My best guess was that 33 was a typo made by pressing 3 twice instead of once. Therefore, I changed the number of bedrooms from 33 to 3.

More detailed comments and python codes are in the Jupyter notebook in this link
https://github.com/math470/Springboard_Capstone_Project_1