SI 618 Project 2: Rental Price in the United States

by Seaborn and Sklearn in Python

hyfrankl@umich.edu

I. Motivation

Rentler.com is an online platform for landlords and tenants, which provides such convenient services like background screening and faster move-in. It has ”over 1.3 million users in over 3000 cities” in the United States [1]. It is praised as it is a good help for landlords and a bridge between tenants and landlords.

As an international student, it is my first time being exposed to renting as I live in the off-campus houses. Some of my friends once discussed the rental prices of different off-campus housing. It stimulates my interest in investigating the housing rental price. After some research, I have found that over 30% of Americans rent the houses; especially for younger householders, renting can provide flexibility and low cost of maintainence [2]. Therefore, I intend to find some important factors of rental prices in this project. I will mainly investigate the location, the amenities, and the neighborhoods with rental prices. Besides, I'd like to check anomaly rental prices in the dataset. In brief, the project is to understand the general picture of rental prices in the United States. It can help many Americans understand the rental price of apartments.

Here are some questions I will solve in this project.

1) Box plot: detect any anomaly data points and investigate the distribution of rental price and square feet.
2) Bar plot: observe the mean and the variance of the rental price per square feet in each state.
3) Violinplot: draw the distribution of the rental price per square feet for different population density
4) Decision tree: use decision tree to predict the rental price per square feet and visualize the decision tree.

II. Dataset

Although the dataset was published on the kaggle in 2021, it is still reasonable to use as it was only a year ago. These columns provide a brief summary of the apartment.

1) **Source**: I use the dataset in the kaggle US Rental Listings Summer 2021 accessed from https://www.kaggle.com/datasets/elizabethveillon/us-rental-listings-summer-2021. The data was pulled from Rentler.com and combined with population density data scraped from mapzipcode.com. The merge has already completed in the accessed data.
2) **Format**: It contains a csv file with around 450 MB. It includes some important information like price, number of bedrooms, number of bathrooms, parking, city, state, house type, population density, and room size.
3) **Important Variables**: I use the following columns in the analysis
   a) price: float
   b) num_beds: float
   c) num_baths: float
   d) state: string
   e) house_type: string
   f) ac (air conditioner): float (0.0 or 1.0)
   g) sqft (square feet): float
   h) PopulationDensity: float
4) **Records and Time Periods**: it contains US Rental Listings Summer 2021.

III. Data Manipulation Methods

Summary: the whole project is related to feature engineering to achieve reasonable prediction in the end. However, some entries contain too much null values. Therefore, it uses cleaned data with one hot encoding.

Code: the related code is in the .ipynb file with comments. It is easy to jump to each section.

A. Data Pre-processing

1) Drop all the null values and meaningless data like the null values in price, num_baths, num_beds, and ac.
2) Clean the rows where the population density is not larger than zero and number of bedrooms is larger than zero.
3) Apply one hot encoding to house type.

B. Feature Engineering

Q1: detect any anomaly data points and investigate the distribution of rental price and square feet

1) **Cleaning:** as this question is to find the noisy/anomaly data in rental price and square feet, I only drop the missing and incomplete data.
2) **Manipulation:** as the distribution of the rental price and square feet is similar to a log distribution, I take the log value of the rental price and square feet and then use the box plot with rug to see any anomaly points. Besides, I also show the distribution of the rental price per square feet which is more meaningful with kdeplot because the rental price may rise as the square feet of the house increases.
3) **Challenge:** the main challenge I encounter is that the direct box plot of the rental price will only show the anomaly points. So, it is wise to use the log transform.

Q2: observe the mean and the variance of the rental price per square feet in each state

1) **Cleaning:** the data cleaning including missingincompleteunreasonable data is only based on the pre-processing.
2) **Manipulation:** I use the bar plot to show the mean and std of log rental price per square feet.
3) **Challenge:** the main challenge is that the label of state in the x-axis overlaps each other, so I adjust the size of figure.

Q3: draw the distribution of the rental price per square feet for different population density

1) **Cleaning:** choose the house with the population density larger than zero.
2) **Manipulation:** quantize the population density into three categories.
3) **Challenge:** the main challenge is to find a proper way to quantize the population density. Finally, I select three – ”low”, ”median”, and ”high” to not lose the generality.

Q4: use decision tree to predict the rental price per square feet and visualize the decision tree

1) **Cleaning:** the data cleaning including missingincompleteunreasonable data is only based on the pre-processing.
2) **Manipulation:** I use the columns with numbers and select important columns with correlation. Next, I use PCA and silhouette score to decide the best number of clusters. Then, with this, I use the singular value decomposition to visualize the data and quantize the log price per square feet into clusters to see the relationship between the selected columns and the log rental price per square feet, In the end, I apply the decision tree to predict the log rental price per square feet.
3) **Challenge:** the main challenge is that it is hard to find a better way to select the important columns that can well predict the rental price because we can find that many outliers in the data.

IV. Analysis and Visualization

Q1: detect any anomaly data points and investigate the distribution of rental price and square feet

1) **Workflow:** First, use boxplot and rugplot to show the distribution of rental price and square feet. Then, compute the log value of rental price divided by the square feet and store it in the log_price_per_sqft column and feed this column to kdeplot.
2) **Finding:** From Fig. 1, Fig. 2, and Fig. 3, one can find that there are many outliers in the rental price and square feet, while the log price per square feet is almost normal distribution. So, it is more valuable to examine the log price per square feet than rental price.

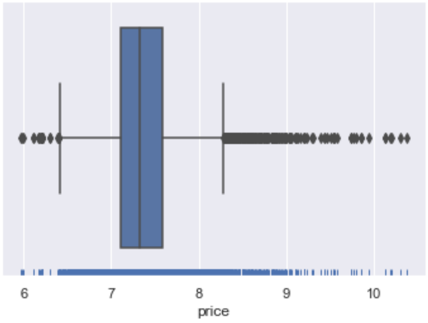

Fig. 1. The Distribution of Log Price

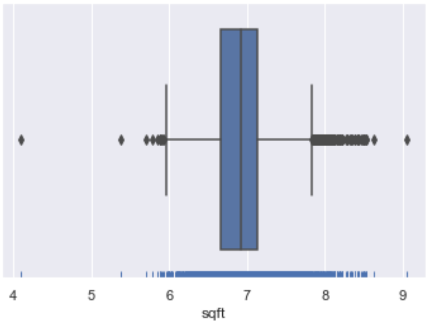

Fig. 2. The Distribution of Log Square Feet

Q2: observe the mean and the variance of the rental price per square feet in each state

1) **Workflow:** Compute the log value of rental price divided by the square feet and store it in the log_price_per_sqft column and feed this column with the state information to the barplot.
2) **Finding:** One can find that the mean of the log rental price per square feet differs for different states, and meanwhile the standard deviation also differs in Fig. 7.

Q3: draw the distribution of the rental price per square feet for different population density

1) **Workflow:** First, I choose only the rows with the population density larger than zero. Then, I quantize the population density into three pieces with equal size. After that, I label each piece with low, median, and high. Next, I use the pd.cut to mark each house with the three defined labels. Finally, I use the violin plot to see the distribution of log rental price per square feet in different levels of population density.
2) **Finding:** The median of the log rental price per square feet is similar for each level of population density shown in Fig. 4. But the distribution of the log rental price per square feet like the quantiles is quite different.

Q4: use decision tree to predict the rental price per square feet and visualize the decision tree:

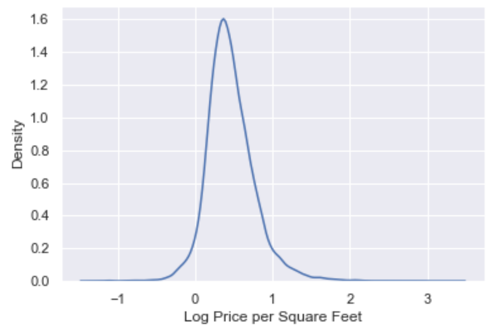

Fig. 3. The Distribution of Log Price Per Square Feet

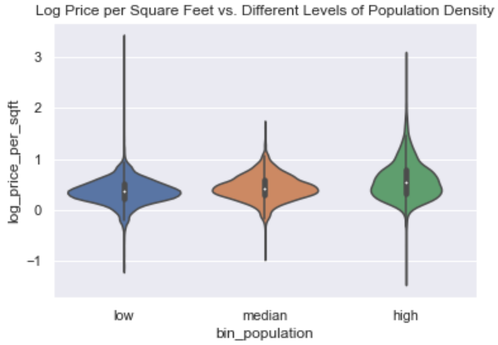

Fig. 4. The Distribution of Log Price Per Square Feet for Different Levels of Population Density

1) **Workflow:** I first use the heatmap to visualize the correlation of each two columns in the data. Next, I use PCA and silhouette score to decide the best number of clusters is 3. Then, I use the correlation to choose important columns from data. Then, I use these (normalized) columns to extract the embeddings from the first column of the U matrix in singular value decomposition. Next, I draw the scatter plot of the embeddings and the log rental price per square feet to see whether it can see clearly relationship. Finally, I use the selected columns to predict the rental price per square feet and visualize the decision tree. As for the decision tree, I limit the max depth as 3 and the proportion of min samples in the leaf is 0.01.
2) **Finding 1:** the log price per square feet is similar to the population density, while the log price is similar to the number of bedrooms, the number of baths, and the square feet. So, I may choose square feet, the number of bedrooms, number of baths, population density to further analysis shown in Fig. 5.
3) **Finding 2:** the best number of clusters is 3 because the silhouette score is the highest shown in Fig. 9.
4) **Finding 3:** with the help of PCA in previous manipulation, we can find that the data can be more easily divided in the scatter plot, so it represents that the columns may well predict the target values. The ratio of low, median, high rental price per square feet is around 0.5, 0.45, 0.05 defined by the previous clustering shown in Fig. 6.
5) **Finding 4:** I use the mean square root of the rental price per square feet (not log) as the criteria. I use the mean value as the baseline 0.49. Then, I use the cross validation and obtain the average train score 0.38 and the average test score 0.39. One of the decision tree is shown in Fig. 8. We can see the population density is the main factor shown in Fig. 8. As the rental data involve many outliers, the accuracy of prediction is acceptable.

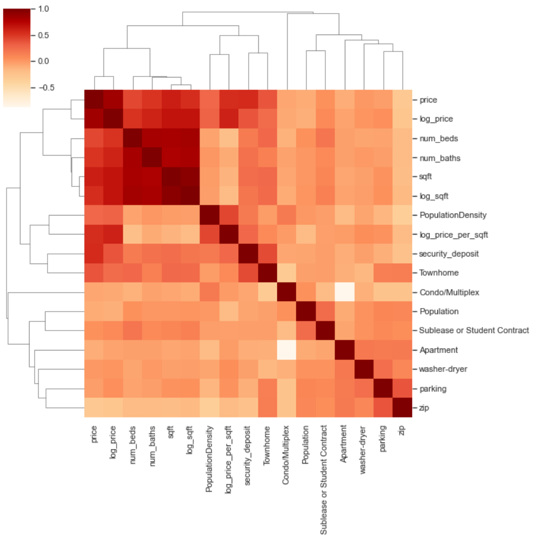

Fig. 5. Heatmap of the correlation between each two columns

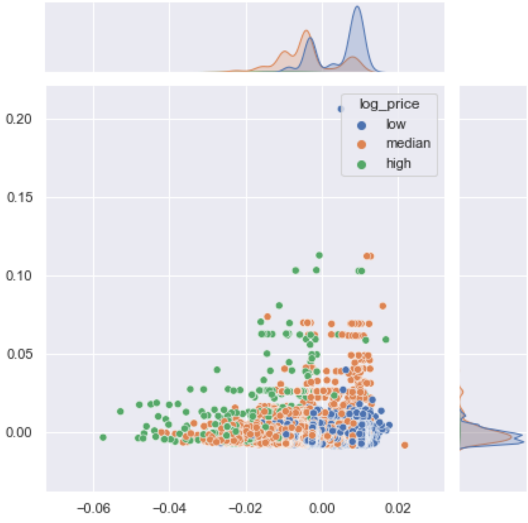

Fig. 6. Data Visualization with Dimension Reduction by SVD

References

[1] R. Social, “Rentler.com reviews 2022: Details, pricing & features,” 2022. [Online]. Available: https://www.g2.com/products/rentler-com/reviews

[2] W. Fargo, “Harvard joint center for housing studies.” *Harvard JCHS Americas Rental Housing 2022, Joint Center for Housing Studies of Harvard University,*, 2022. [Online]. Available: www.jchs.harvard.edu/sites/default/files/reports/files/Harvard_JCHS_Americas_Rental_Housing_2022.pdf

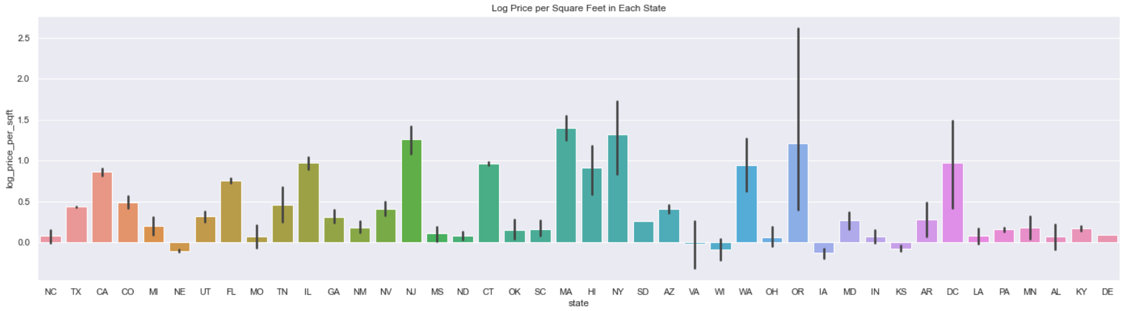

Fig. 7. Log Price Per Square Feet in Each State

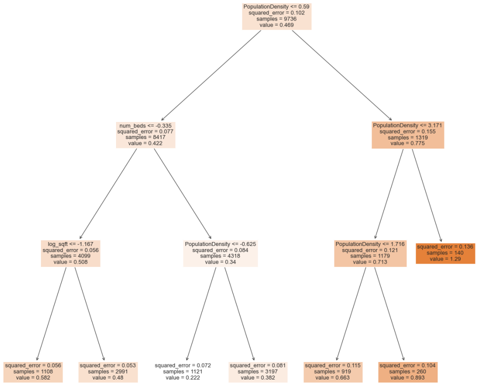

Fig. 8. Prediction Logics with Decision Tree

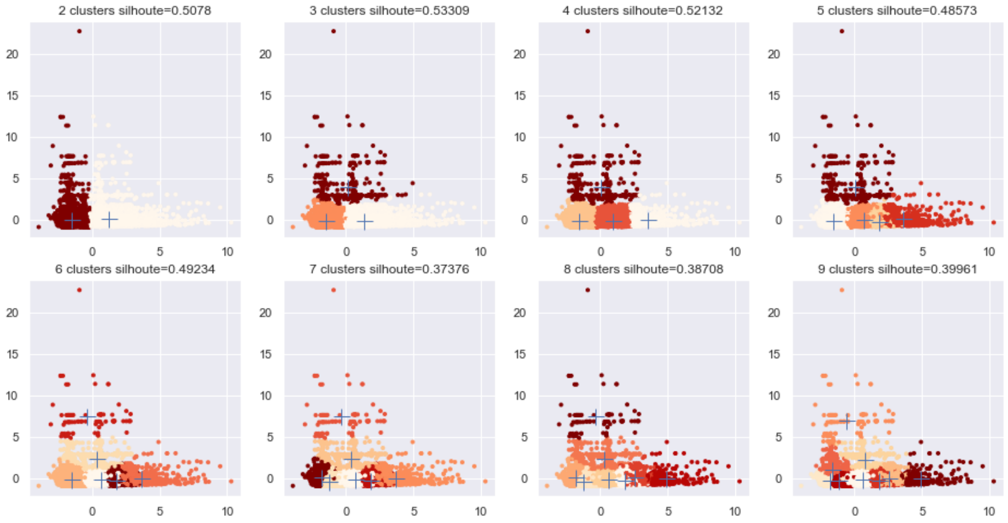

Fig. 9. Silhouette Score of Different Clusters