**Long-Term Rental Houses**

**Analysis**

**Group Members:**

|  |  |  |
| --- | --- | --- |
| **First name** | **Last Name** | **Student number** |
| **Alisha** | **Thomas** | **C0872330** |
| **Aswathy** | **Anilkumar** | **C0896280** |
| **Jasmin** | **Joseph** | **C0897352** |
| **Sudhy** | **Sukumaran** | **C0891288** |

**Submission date: *16-03-2024***

Contents

[1.**Abstract** 1](#_Toc161563341)

[2.**Introduction** 1](#_Toc161563342)

[**3.Method** 1](#_Toc161563343)

[**4. Analysis methods** 3](#_Toc161563344)

[**5. Results** 12](#_Toc161563345)

[**6. Conclusion And Future Work** 13](#_Toc161563346)

[**7.References** 14](#_Toc161563347)

# **1.Abstract**

The main focus of the project to do an extensive exploratory data analysis (EDA) and outlier detection on a dataset that was collected via web scraping from Kijiji, an online classified advertising service. The assignment consists of several processes, including encoding, data wrangling, visualization, pandas profiling, outlier identification, and using several approaches to deal with outliers. To ensure appropriate formatting and organization, data is first extracted and cleaned. The distribution of various attributes within the dataset is then displayed using visualizations. Panda’s profiling is then employed to produce an in-depth report that enables additional EDA.

## **2.Introduction**

Through leveraging the capabilities of web scraping, this report examines the detailed analysis of long-term house rentals and advertisements from the Kijiji website. Kijiji is a popular website that widely offers a range of goods and services, including everything from furniture and electronics to job listings and housing rentals. It is a platform that is user-friendly and provides communication between buyers and sellers. Here we are mainly focusing on the analysis of long-term housing rentals in Ontario. The main objective is to obtain trends, facts, and insights about long-term housing rentals in Ontario.

## **3.Method**

**3.1Context and setting of the study**

The main objective of the House renting online platforms is to allow people to rent houses, apartments, basements, condos, townhouses, duplexes/Triplexes. The Kijiji website where the data is scraped, provides a platform for sellers to rent out their properties and buyers to choose from them according to their requirements.

The descriptive analysis enables to uncovering of correlations, trends, and outliers within the dataset with property attributes and rental prices.

**3.2 Study Design**

After the data is scraped, the report moves on to the data-wrangling stage. Also, it includes handling missing values, duplicates, and standardizing column names. The next step is visualizing the data. Histograms, Boxplots, scatterplots, Bar charts, Pie charts, and Pair plots are employed to identify patterns and trends. Following that is the Pandas profiling, using the Pandas profiling we can generate comprehensive reports to facilitate Exploratory Data Analysis. The report is helpful in providing insights into summary statistics, correlation, missing values, and other relevant information.

The critical phase combines all the attributes, by revealing relations and patterns in the dynamics of long-term room rentals that were before hidden.

**3.3 Data set details**

The scraped data contains 11076 rows and 18 columns, but after dropping the duplicates the data contains 6084 rows and 18 columns.

* 1. **Identify the main study variables**

Price\_Condition: conditions of utility price included.

Unit\_type: describing the types of rentals like condos, apartment basements, town houses.

Bedrooms: Number of bedrooms

Bathrooms: Number of bathrooms

Parking: availability of parking space

Pet\_friendly: gives the pet-friendliness type of the property.

Size\_sqft: the size of the property in square feet

Furnished\_info: if furnished or unfurnished

Price: the rent amount

Description: extra comments on the property

* 1. **Data collection instruments and procedures**

The data collection process gathers information on long-term room rentals from online sources, such as classified websites like Kijiji, to perform exploratory data analysis and outlier detection. Then developed a Python script utilizing libraries such as Beautiful Soup and Requests to scrape data from Kijiji. The script will navigate through the website's HTML structure to extract relevant information about long-term room rentals, including rental prices, location, room types, amenities, and additional attributes.

## **4. Analysis methods**

Long-term room rental data analysis starts with careful data wrangling, which includes actions like data scraping from Kijiji cleaning to address missing values and irregularities, formatting to standardize column names and formats, and validating to guarantee data integrity. The next step is called exploratory data analysis (EDA), in which correlation analysis is carried out to find possible patterns, descriptive statistics are computed for important variables, and visualizations such as histograms and boxplots are made to comprehend distributions and relationships. Categorical variables are transformed into numerical representations using encoding techniques, and a thorough report is produced using pandas profiling to reveal information about missing values, correlations, and data properties.

**4.1 Data Scraping**

The data about the long-term housing was downloaded from the Kijiji Website. Library used for scraping the data is “Beautiful Soup”. The library enables the user to scrape the data and parse HTML and XML documents. The data scrapped from the Website using Beautiful soup is converted into the text file and then its converted into the CSV file using .

**4.2 Data Wrangling**

The scraped data contains 11076 rows and 18 columns, but after dropping the duplicates the data contains 6084 rows and 18 columns. The reason behind the occurrence of the duplicate values is due to the time duration of the data scraped. It took around one week to complete the data scrapping during which the Kijiji page contents kept updating. The concise summary of the dataset is shown in figure1.

A screenshot of a computer

Description automatically generated

Figure:1 - initial concise summary of dataset

In the figure:2 black color indicates non null value and white color indicate missing value.

Utilities such as Hydro, Heat, Water, Cable TV, Internet, and moving date have a higher number of missing values.

A close-up of a bar code

Description automatically generated

Figure 2: Visualization of null values.

Moving on, the unique value count of each feature is analyzed. The feature ‘Price\_condition’ has three values No Utilities Included, Some Utilities included, and All Utilities Included. The feature ‘Unit\_Type’ has unique values, Apartment, Basement, House, Condo, Townhouse, Duplex/Triplex, and Not Available.

While analyzing the unique values of the feature ‘Bedrooms’ noise was found in the data.

A table of numbers and a number

Description automatically generated with medium confidence

figure 3: Unique value count of Bedrooms

Hence, the values related to the bathroom present in the feature ‘Bedrooms’ may be because of the human error while saving the scraped data. Therefore, those noises were removed from the data. From the data, it is found that 19.53% of the Sizensqft feature is null value. And found that the value ‘Not Available’ can be considered as a null value, the missing values are imputed using KNNIpmuter. Even after applying KNN imputer, there were 432 missing values. This is maybe due to insufficient neighbors or sparse data. Hence, we dropped the remaining null values.

**4.3 Plotting Technique**

A graph of different colored bars

Description automatically generated with medium confidence

Figure 4: Count of Bedrooms by Unit Type

The majority of the apartments are one-bedroom and Two bedrooms. In houses, the majority are 3 bedrooms and 4 bedrooms. Among them, 3 bedrooms are more. In the case of condos, the maximum number of bedrooms is three. That is less in number. From the data, we can see that apartment counts are higher in number whereas duplexes and triplex are the least.

A graph showing different colored squares

Description automatically generated

Figure 5: Count of parking by Unit Type

The majority of the apartments does not have parking space. In houses the majority have 2 or 3 parking spaces.

A graph of different colored bars

Description automatically generated with medium confidence

Figure 6: Count of Bathrooms by Unit Type

In Figure 6, it is shown that the majority of the units have only one bathroom except the townhouse.

A graph of a number of people

Description automatically generated with medium confidence

Figure 7: Count of Pet Friendly by Unit Type

It is shown that most of the apartments are pet-friendly. The majority of the Basement is not pet-friendly.

**4.4 Pandas Profiling and Making Profiling Report**

Pandas profiling is used to generate a complete and exhaustive report for the dataset. It also helps a lot in exploratory Data Analysis which helps to understand the underlying structure of the data, determine the patterns, and generate insights in a visual format.

**4.5 Encoding Categorical Columns**

The “Furnished\_info” column contains categorical binary values, Which are converted into numerical values by using Label encoding. Columns “Price\_condition” and “Pet\_friendly” also contain 3 categorical ordinal values. So these values are converted to corresponding numerical using Label encoding too.

“Unit\_type” is a nominal categorical variable, For this one hot encoding is suitable for conversion. It contains 6 different value. After the One-Hot Encoding data frame will added by 6 extra columns each containing 0 or 1 value.

**Correlation Matrix**

After above preprocessing by correlation matrix, Identify the relationship between each column and how much each other depends on each other. The correlation matrix is a table that shows the correlation between variables. The values of the correlation matrix specify that none of the components are neither have a strong positive correlation or a negative correlation. So, It concludes that all the data collected is relevant for the insight.

**4.6 Outliers Identification**

A couple of graphs showing a graph

Description automatically generated with medium confidence

Figure 8 : Box plot in Price and Size\_sqft columns

By analyzing the above figure,it is clear that the dataset contains, especially the price column and sizeist columns above the upper viscus level there is a lot of outliers presented. By using IQR, a number of outliers is counted as 425. Since the number of outliers is high. It is not a good practice to simply eliminate all. If all the outliers are removed then the dataset record count will be reduced to 4671.

**4.7 Outliers Treatment**

**4.7.1 Quantile based Flooring and Capping**

A graph of histograms

Description automatically generatedA graph of histograms

Description automatically generatedThe quantile function calculate the percentiles for the specified columns such as Price and Size\_sqft at the first percentile (Floor) starting from 0.01% and ending at 0.99% which is Capping. Any values between this range will be kept, and out-of-the-range values will replaced by upper and lower quantile values. This leads to keeping the dataset size but may introduce noise.

Figure 9. Comparison of Price and Size\_sqft with and without Capping and Flooring

**4.7.2 Trimming**

In trimming the out the range values will removed from the dataset. It will reduce the size of the dataset. The range will define the base upper and lower whisker

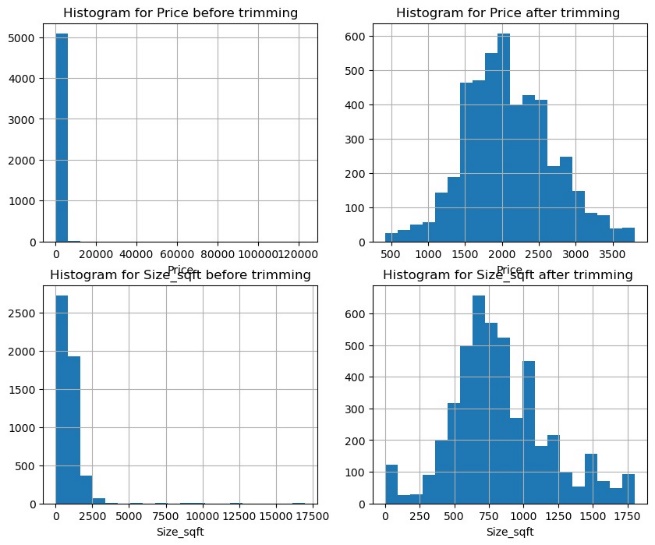


Figure 10. Comparison of Price with and without Trimming

**4.7.3 Log Transformation**

A graph of a diagram

Description automatically generated with medium confidence

Figure 11. Comparison of Price with and without Log Transformation

Comparison of Outlier techniques:

* Cap and floor method, even though the data lose is less, but the column Size\_sqft is highly skewed(1.4). The skewness couldn't removed even after applying log transform. Hence, this method is not using for further analysis.
* Trimming method, the data lose is little bit higher, but the skewness is low.
* 3.Log transformation , the skewness for both data is very high(-4.41 and -4.54).

Finally, Trimming method is chosen for further processes.

**4.8 Clustering Techniques**

**4.8.1 KMeans Clustering**

Before applying KMeans, it is required to standardize the Price and Size\_sqft value. Standardization is a scaling technique in which values are centered at 0 mean with a standard deviation of 1. Those column values are very large when compared to other columns which leads to the model giving extra preference for these two columns. Applying standardization helps to avoid this issue. One of the parameters for KMeans is the number of clusters.To decide the optimum number of clusters/centroid elbow method was used.

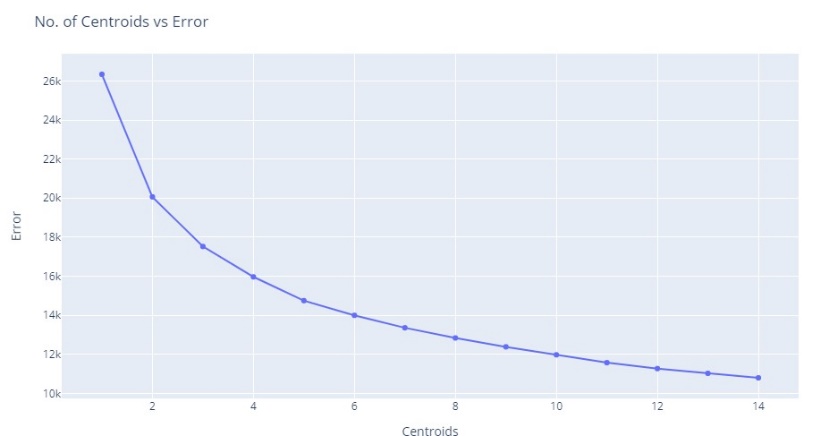


Figure 12. Elbow method output – Centroid vs Error

Using the elbow method 5 or 6 was chosen as the best optimum cluster value. By passing the value to KMeans , The output gives certain patterns.

A colorful dots on a white background

Description automatically generated

Figure 13. Clustered 3D scatter plot between Price vs Size\_sqft vs Bedrooms

The above 3D plot possesses some clusters that show patterns that help to identify the relationships between various factors.

**4.8.2 DBSCAN**

It clearly showing one cluster is concentrated in the center. But not showing any particular pattern. Even the 3D plot does not show any helpful pattern as the previous model output.

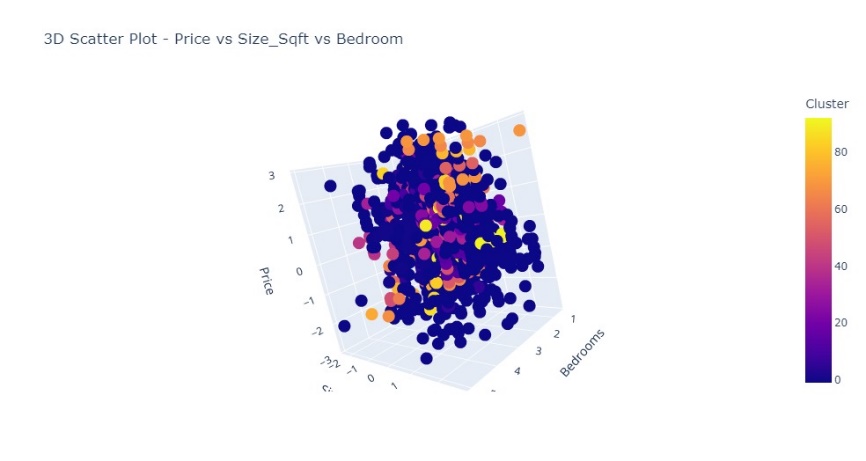


Figure 14 3D Scatter plot after DBSCAN between Price,Size\_sqft and Bedrooms

**4.8.3 Agglomerative clustering**

Like the other clustering techniques agglomerative is a hierarchical clustering technique that groups objects in clusters based on their similarity.

1. Points in the same cluster are closer to each other 2. Points in the different clusters are far apart

A colorful diagram of a diamond

Description automatically generated with medium confidence

Figure 15. 3D Scatter Plot Price vs Size\_sqft vs Bedroom

Among these 3 clustering methods, KMeans and Agglomerative clustering show certain patterns in the dataset.

**4.9 NLP Techniques**

A close-up of words

Description automatically generated

Fig 16: Word Cloud

From the Word Cloud obtained through performing the NLP techniques, we can conclude that the most used keywords are related to the kitchen such as fridge, stove, stainless steel and so on. The significance of these terms indicates that kitchen features are significant factors. Furthermore, we can also conclude that information derived from this dataset will be mostly centered around kitchen-related attributes. Followingly the next most used keyword id pet friendly. The techniques used are;

* Cleaning Special Characters
* Removed Customized list of Stop words
* Removed Punctuation
* Replace emoji with description
* Removed Digits
* Lemmatization
* Ngram, two gram ,trigram

## **5. Results**

The dataset that contains the raw data is scrapped first. Exploratory Data Analysis, Data Wrangling, minimizing the outliers, categorical variable encoding, and transformation which are part of the data analysis process are performed. After all these primary techniques is performed, we then detect the outliers. Flooring, capping, trimming, and log transformation were employed to deal with the outliers. Box plots were used to visualize the effects of the three strategies on treating the outliers. K Means clustering which is a clustering method is used to provide more insights into exploring the different patterns in the data. To sum up, the study provided a comprehensive comprehension of the dataset through a sophisticated analytical approach.

## **6. Conclusion And Future Work**

We can expand the understanding of the long-term house rental data by utilizing more sophisticated outlier identification techniques, more cluster exploration, and predictive modelling. We can obtain more data to increase the range of analysis and prediction models. Summing up the examination of the long-term house rental data enables us to get a significant understanding of the price and features of the home rentals.

## **7.References**

1 Scraping Kijiji Home Rental Advertisements using beautiful Soup.Medium <https://medium.com/analytics-vidhya/scraping-kijiji-home-rental-advertisements-using-beautiful-soup-5e286af9d96>

2.Whats is Exploratory Data Analysis

<https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/>