# Data Mining And Discovery

Report

Taissir Boukrouba - 22084758
**Subject N°8 ( Group 48 )**
University Of Hertfordshire

April 12, 2024

# Contents

# List of Figures

# List of Tables

## 1.1 Choosing The dataset

The chosen dataset is called **”Apartment for Rent Classified”** [1] which is business dataset sourced from the University of California, Irvine’s Machine Learning Repository (UCI) that has information about apartments in the USA and its features. This table will summarize the information we need :

| | |
|---|---|
| Source | UCI |
| Problem | Anomaly Detection Clustering |
| Data Size | 100K |
| Features | 21 |
| Main Features | Price , Square feet , Bathrooms , Bedrooms |
| Datatypes | Object (17) , Float32 (4) |
| Sample Size | 8.3K |

Table 1.1: Data Information Summary

Please note that we dropped 3 meta-features which are features that has information about the dataset itself such as *time* (when data was created) , *source* (the source of each data point ) , *price_display* (the same price column but in integer format) upon loading our data as a pandas dataframe.

## 1.2 Data Preprocessing

Due to the initial poor data quality, this cleaning process was extensive. Therefore, we will not cover all the steps in detail here.

### 1.2.1 Cleaning Data

This process addressed two key data cleaning aspects : **Missing data** and **Column uniqueness** where we removed columns with a very high proportion of missing values, as these likely wouldn’t contribute meaningfully to the analysis. Additionally, columns with either very high or very low cardinality (the number of unique values) were dropped.These columns with low/high cardinality likely wouldn’t provide sufficient patterns for effective model training. Finally, to further improve data quality, we **removed duplicates** ( about 73 rows ). [3]

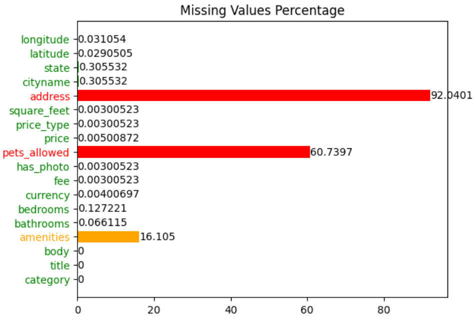


Figure 1.1: Missing Values Barplot

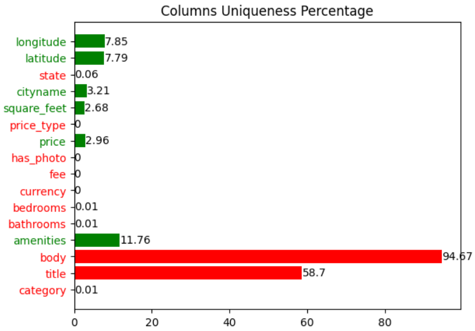
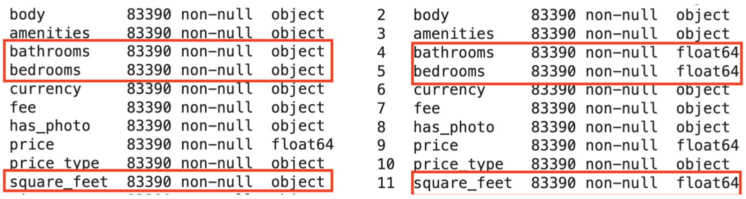


Figure 1.2: Column Uniqueness Barplot

### 1.2.2 Datatype Formatting

Data cleaning revealed a weird finding where columns like ”*bathrooms*” and ”*bedrooms*,” meant for room counts, were in text format (object), not numbers. Similarly, ”*square_feet*” (meant for numerical values) was also text. To fix this and enable proper analysis, we converted all three to usable numerical formats.



```
body         83390 non-null  object        2   body         83390 non-null  object
amenities    83390 non-null  object        3   amenities    83390 non-null  object
bathrooms    83390 non-null  object        4   bathrooms    83390 non-null  float64
bedrooms     83390 non-null  object        5   bedrooms     83390 non-null  float64
currency     83390 non-null  object        6   currency     83390 non-null  object
fee          83390 non-null  object        7   fee          83390 non-null  object
has_photo    83390 non-null  object        8   has_photo    83390 non-null  object
price        83390 non-null  float64       9   price        83390 non-null  float64
price_type   83390 non-null  object        10  price_type   83390 non-null  object
square_feet  83390 non-null  object        11  square_feet  83390 non-null  float64
```

Figure 1.3: Before And After Datatype Formatting

### 1.2.3 Handling Outliers Using Winsorizing

Upon investigation of the numerical features revealed a significant number of outliers in the ”price” feature (3,929 data points to be exact). The distribution of this feature exhibits positive skew (right-skewness), indicating a clear asymmetry towards higher values, as illustrated in the figures below.

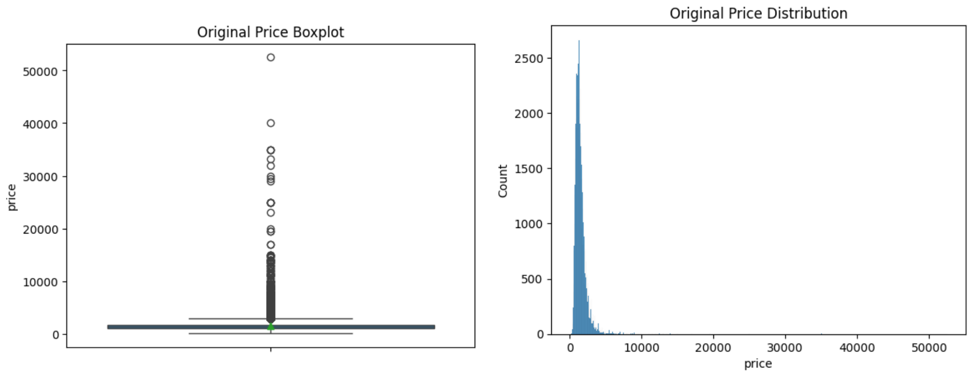


Figure 1.4: Original Price Distribution

In the process we tried to to get rid of them using winsorising [3] but instead using manual upper and lower limits we used quartiles (q1,q3) and IQR to calculate these limits for more accurate results

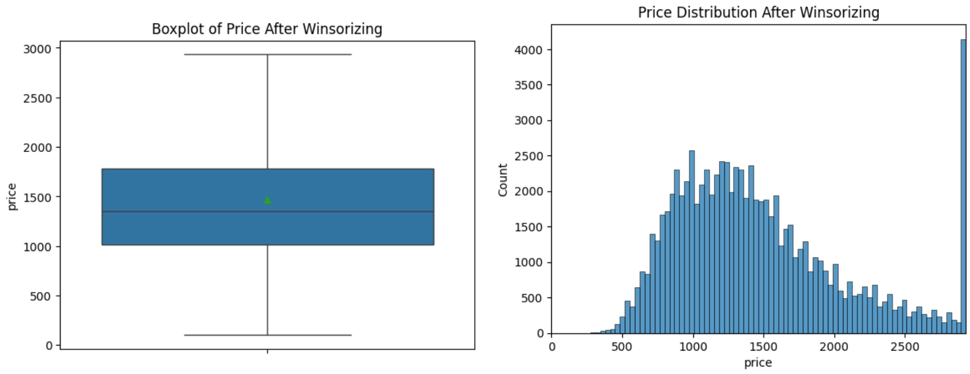


Figure 1.5: Price Distribution After Winsorizing

## 1.3 Clustering Using K-means

After Using **Stratified Random Sampling** [2] to sample our data (8K samples) we used the **elbow method** to investigate the optimal number of clusters for our **k-means** [4] model which we found it to be 3 , then we applied clustering and investigated the results where we found that the clusters have divided our data into 3 categories where the first one is **low rent prices** ranging from (0-1.5K$) and the second one for **moderate rent prices** from (1.5K$ -   2.3K$) and then from 2.3K$ onwards it’s the **expensive rent prices** which is the true interpretation of our data.

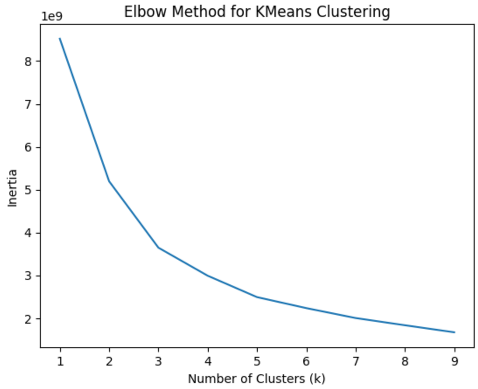


Figure 1.6: Elbow Method Plot

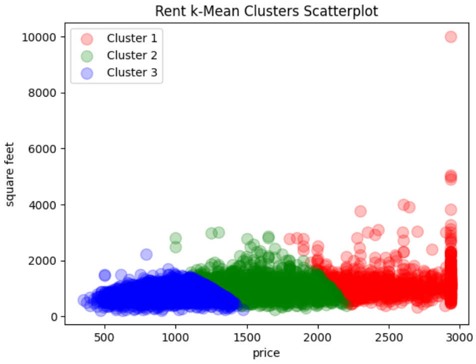


Figure 1.7: K-Means Clustering Results

## 1.4 Anomaly Detection

During preprocessing, we removed outliers.This might seem counter-intuitive for anomaly detection, but our goal was to identify genuine anomalies, not simply large values that deviate from the average. To achieve this, we employed a **distance-based approach using KNN** [5] with a maximum of 4 neighbors. While we didn't detect a significant number of outliers, the ones with the highest anomaly scores were concentrated in the uppermost values. Notably, this included a data point with 10K sqft listed as a 0-bedroom (studio) apartment for $2.9K per month. This instance clearly exemplifies an anomaly, as 10K sqft is far outside the expected range for a studio, and the price further reinforces this conclusion

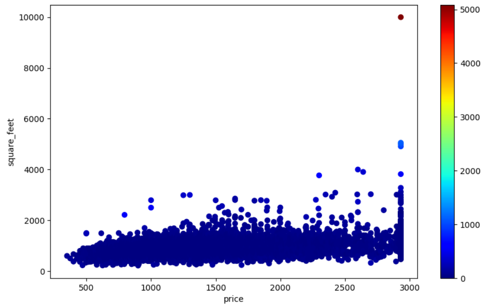


Figure 1.8: Anomalies I

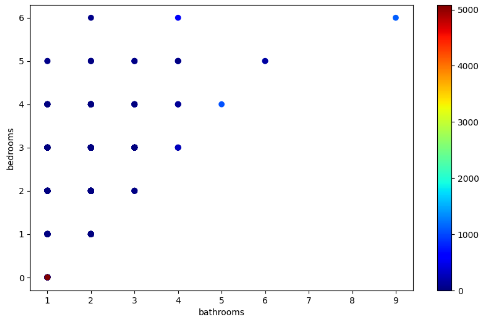


Figure 1.9: Anomalies II

## 1.5 Comparisons & Final Notes

Clustering and Anomaly detection approach data points differently. In clustering, data points are grouped together based on similarities, forming clusters. This means even outliers can be categorized into the closest cluster, as you might have observed in the clustering visualizations. Conversely, Anomaly detection treats each data point as an individual entity and assesses whether it deviates significantly from the norm. In our rent price analysis, this approach effectively highlighted specific listings that were significantly different from the norm.

# Bibliography

[1] UCI Machine Learning Repository,Apartment for Rent Classified. 2019. DOI: https://doi.org/10.24432/C5X623.

[2] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. Survey methodology. volume 561, pages 189–234. John Wiley & Sons, 2009.

[3] Sang Kyu Kwak and Jong Hae Kim. Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology*, 70(4):407, 2017.

[4] Abdulhamit Subasi. Chapter 7 - clustering examples. In Abdulhamit Subasi, editor, *Practical Machine Learning for Data Analysis Using Python*, pages 465–511. Academic Press, 2020.

[5] Ming Zhao, Jingchao Chen, and Yang Li. A review of anomaly detection techniques based on nearest neighbor. In *Proceedings of the 2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018)*, pages 290–292. Atlantis Press, 2018/04.