Python Assignment – 1

Kijiji Property Data

Group Name: D

Group Members:

First name	Last Name	Student number
Jash	Vaghasiya	C0884733
Nivedini	Kathagonda	C0872720
Keval	Parmar	C0882386
Monil	Rupawala	C0882370
Sai Divya Madhuri	Guntupalli	C0882360

Submission date: ***29/07/2023***

Lambton College | In Toronto

Contents

Abstract ..... 3**Error! Bookmark not defined.**
Introduction ..... **Error! Bookmark not defined.**3
Methodology 4
Results 15
Conclusion and Future work 15
References 15

Abstract

In this Python assignment, we analyze property data scraped from Kijiji, a well-known online sales platform. Exploratory Data Analysis (EDA), preprocessing, outlier removal, text preparation, visualization, clustering, and Principal Component Analysis (PCA) are all part of the project.

We collect pertinent property information such as kind, location, price, rooms, and many others using web scraping techniques. The EDA phase aids in the discovery of data insights, the handling of missing information, and the establishment of feature associations. Preprocessing ensures that data is clean, and outlier reduction enhances analytical accuracy.

The text preprocessing stage converts unstructured property descriptions into a structured format, allowing for more accurate analysis. We acquire valuable data distribution and correlation insights by employing various visualization approaches.

Furthermore, clustering techniques group related traits together, allowing for pattern recognition. Finally, PCA minimizes data dimensionality while keeping critical features, making visualization and modeling more efficient.

Introduction

In this assignment, we go on a complete data analysis journey, employing several critical techniques to generate valuable insights and prepare the data for advanced modeling and visualization. Exploratory Data Analysis (EDA), Text Preprocessing, Word Clouds, Encoding Methods, Outlier Detection and Removal, Capping, Trimming, Logging, Feature Scaling, K-Means Clustering, Elbow Method using KElbowVisualizer, and Principal Component Analysis (PCA) are among the tasks we specialize in.

The first phase, Exploratory Data Analysis (EDA), allows us to comprehend the dataset's structure and features. We look for patterns, distributions, and correlations between variables, hoping to find hidden information that will influence our data processing stages. Text preprocessing is critical in converting unstructured text data to structured text data. We turn textual information into a more manageable representation using tokenization, stopword elimination, and stemming techniques, providing the groundwork for further research. Word Clouds provide a visually appealing way to highlight the most prominent terms within text data, providing significant insights into the most frequently occurring words and their significance in the dataset.

Encoding Methods allow us to turn categorical data into numerical representation, enhancing the compatibility of machine learning algorithms with the data. Outlier Detection and Removal enable us to detect and deal with data points that differ significantly from the total dataset, resulting in a cleaner and more accurate analysis. Feature scaling standardizes the range of features, allowing for more accurate comparisons and preventing particular traits from overwhelming others during analysis.

K-Means Clustering groups comparable data points, aiding pattern identification and detecting unique clusters within the dataset. The Elbow Method, as implemented in KElbowVisualizer, assists in establishing the best number of clusters for K-Means, resulting in the most effective clustering results.

Finally, Principal Component Analysis (PCA) decreases the dimensionality of data while keeping critical information. It enables us to convert high-dimensional data into a lower-dimensional environment, allowing for more efficient visualization and modeling.

By integrating these processes, we hope to build a robust and complete data analysis pipeline that will allow us to get valuable insights, prepare the data, and identify relevant patterns, allowing us to make educated decisions.

Methodology

Dataset Details:

This project's dataset was taken from Kijiji. It has 18724 entries, each of which represents a property listing. 'Title', 'Address', 'Price Label', 'Price Value', 'Date Posted', 'Unit Row', 'Parking', 'Agreement Type', 'Air Conditioning', 'Description', and 'Visits' are among the characteristics.

	Unnamed: 0	Title	Address	Price Label	Price Value	Date Posted	Unit Row	Parking	Agreement Type	Air Conditioning	Description	Visits
0	0	Basement room for rent	Tremblay St, Brampton, ON L6Z	Some Utilities Included	[]	5 Days ago	NaN	NaN	1 year	NaN	DescriptionLooking for a girl to share a one b...	102 visits
1	1	Stunning 4-Bed Home with Upgrades! Don't Miss ...	Ajax, ON L1T	No Utilities Included	[]	5 Days ago	Apartment Bedrooms: 1 Bathrooms: 1	1.0	1 year	No	DescriptionPrice: 849, 900/ 3,779/moWelcome t...	4 visits
2	2	Modern Heritage One Bedroom	Market Square, Napanee, ON K7R 1R3	Some Utilities Included	['$1,750']	3 Days ago	Apartment Bedrooms: 1 Bathrooms: 1	1.0	1 year	Yes	DescriptionThis stunning Heritage 1 Bedroom Ap...	142 visits
3	3	Bachelor suite for lease	Napanee, ON K7R 1H6	No Utilities Included	['$1,075']	5 Days ago	Apartment Bedrooms: 2 Bathrooms: 1	1.0	6 months	No	DescriptionNice modern bachelor apartment for ...	187 visits
4	4	FOR SALE: 2 Bed, 3 Bath Townhouse In Oakville	Mistletoe Gardens, Oakville, ON	NaN	[]	5 Days ago	Condo Bedrooms: 2 Bathrooms: 1	1.0	1 year	No	DescriptionLovely Well-Maintained Freehold Tow...	9 visits

Data cleaning and preprocessing:

Initially, the data was imported into a pandas Data Frame. The 'Unnamed: 0' column was removed since it was superfluous. Missing values in the 'Parking' column were filled with 0, while other missing values were filled with 'Unknown.' The column 'Price Value' was formatted to eliminate non-numeric characters and converted to float data type. The 'Unit Row' field was processed to retrieve 'Bedrooms,'

'Bathrooms,' and 'Unit Type' data. The number of days extracted from 'Date Posted' was converted to numerical representation. The columns 'Parking' and 'Air Conditioning' were converted to binary format. The column 'Visits' was cleaned and

converted to integer format. The 'Price Label' column was renamed 'Utilities,' while the 'Unit Row' column was removed. Text preparation techniques such as punctuation removal, lowercasing, and lemmatization were used to clean and process the 'Title,' 'Description,' and 'Address' columns. 'Unknown' was used to fill in missing data in the 'Agreement Type' and 'Air Conditioning' fields.

After Pre-processing data looks like this:

	Title	Address	Utilities	Price Value	Date Posted	Parking	Agreement Type	Air Conditioning	Description	Visits	Bedrooms	Bathrooms	Unit Type
0	Basement room for rent	Tremblay St, Brampton, ON L6Z	Some Utilities Included	0.0	5	0.0	1 year	NaN	DescriptionLooking for a girl to share a one b...	102.0	0.0	0.0	Unknown
1	Stunning 4-Bed Home with Upgrades! Don't Miss ...	Ajax, ON L1T	No Utilities Included	0.0	5	1.0	1 year	No	DescriptionPrice: 849, 900/ 3,779/moWelcome t...	4.0	1.0	1.0	Apartment
2	Modern Heritage One Bedroom	Market Square, Napanee, ON K7R 1R3	Some Utilities Included	1750.0	3	1.0	1 year	Yes	DescriptionThis stunning Heritage 1 Bedroom Ap...	142.0	1.0	1.0	Apartment
3	Bachelor suite for lease	Napanee, ON K7R 1H6	No Utilities Included	1075.0	5	1.0	6 months	No	DescriptionNice modern bachelor apartment for ...	187.0	2.0	1.0	Apartment
4	FOR SALE: 2 Bed, 3 Bath Townhouse In Oakville	Mistletoe Gardens, Oakville, ON	NaN	0.0	5	1.0	1 year	No	DescriptionLovely Well-Maintained Freehold Tow...	9.0	2.0	1.0	Condo

After text-preprocessing columns which holds text values looks like this:

	Title	Title_Processed	Description	Description_Processed	Address	Address_Processed
0	Basement room for rent	basement room rent	DescriptionLooking for a girl to share a one b...	descriptionlooking girl share one bedroom base...	Tremblay St, Brampton, ON L6Z	tremblay st brampton lz
1	Stunning 4-Bed Home with Upgrades! Don't Miss ...	stunning bed home upgrade dont miss	DescriptionPrice: 849, 900/ 3,779/moWelcome t...	descriptionprice mowelcome london lane dream h...	Ajax, ON L1T	ajax lt
2	Modern Heritage One Bedroom	modern heritage one bedroom	DescriptionThis stunning Heritage 1 Bedroom Ap...	descriptionthis stunning heritage bedroom apar...	Market Square, Napanee, ON K7R 1R3	market square napanee kr r
3	Bachelor suite for lease	bachelor suite lease	DescriptionNice modern bachelor apartment for ...	descriptionnice modern bachelor apartment leas...	Napanee, ON K7R 1H6	napanee kr h
4	FOR SALE: 2 Bed, 3 Bath Townhouse In Oakville	sale bed bath townhouse oakville	DescriptionLovely Well-Maintained Freehold Tow...	descriptionlovely wellmaintained freehold town...	Mistletoe Gardens, Oakville, ON	mistletoe garden oakville

EDA and Visualization:

Summery of Dataset

	Price Value	Date Posted	Parking	Visits	Bedrooms	Bathrooms
count	18724.000000	18724.000000	18724.000000	14947.000000	18724.000000	18724.000000
mean	570.241829	5.303461	1.168233	229.018198	1.579737	0.958796
std	1075.500700	3.143302	0.854471	582.772969	1.119471	0.630585
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	5.000000	1.000000	13.000000	1.000000	1.000000
50%	0.000000	6.000000	1.000000	49.000000	2.000000	1.000000
75%	700.000000	6.000000	2.000000	142.000000	2.000000	1.000000
max	7845.000000	29.000000	4.000000	10994.000000	5.000000	4.000000

Price Value: The average cost is around 570. The median (50% quantile) is 0, indicating that many listings have a price of 0. The highest possible price is 7845.

Posting Date: The listings were placed over 29 days.

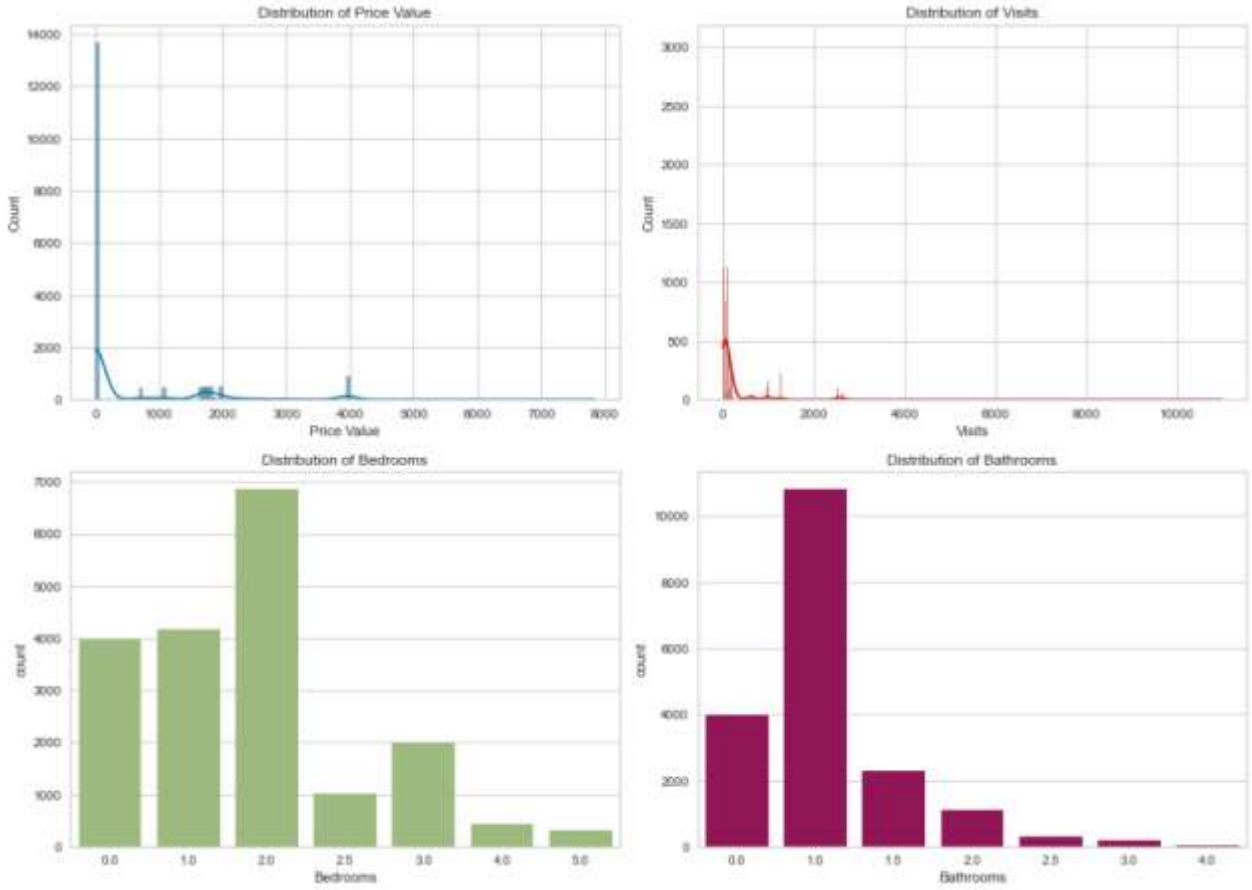
Parking: Most listings have at least one parking space, with a mean of 1.17 and a median of 1.

Visits: The average number of visits to the listings is 229, with a maximum of 10994.The median is substantially lower at 49, showing that a few entries receive a lot of traffic.

Bedrooms: The average bedroom count is about 1.58, with a maximum of 5.

Bathrooms: The majority of listings include at least one.

Distribution of Numerical Variables:



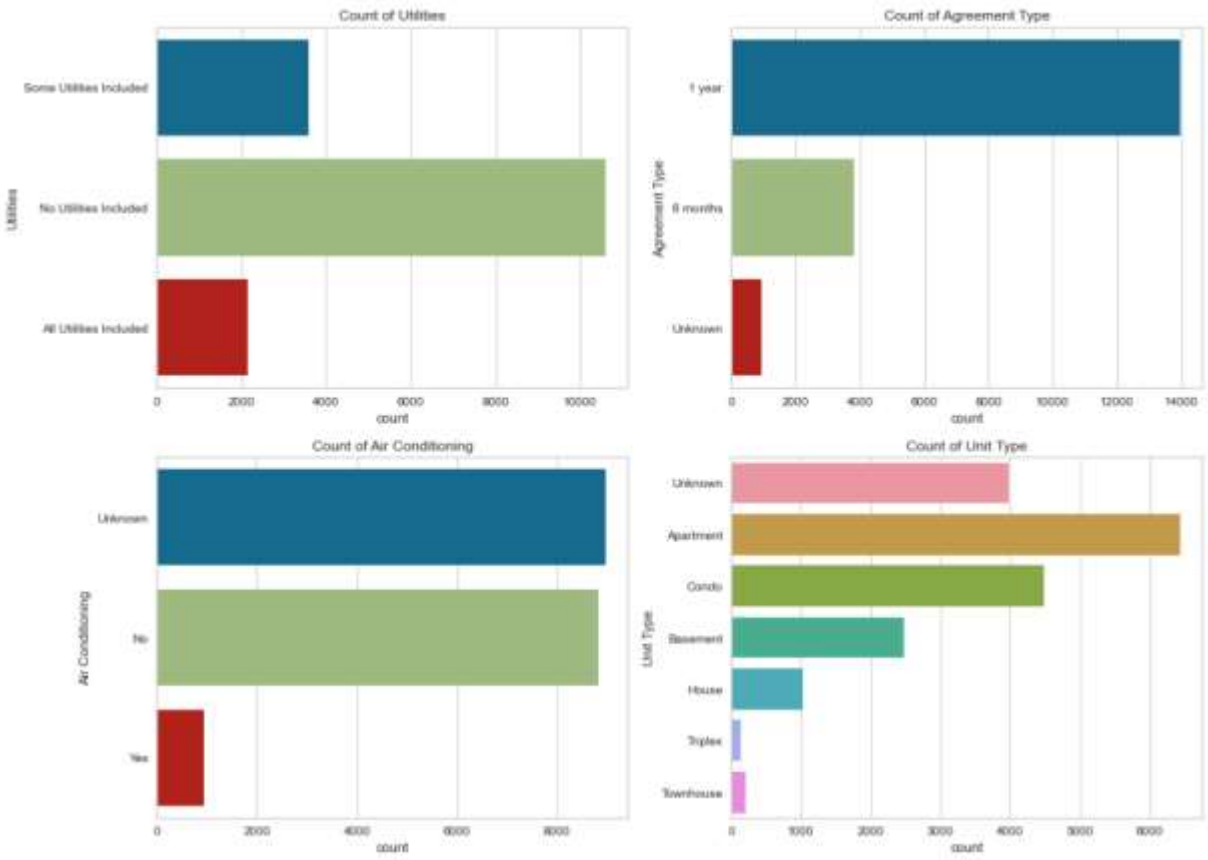
Price Value: The majority of the properties have a price value of 0 while a few have prices ranging from 1000 to 8000. This corresponds to the summary statistics we saw earlier. The distribution is skewed heavily to the right.

Visits: The majority of the properties have fewer than 2000 visits, with several having very little visits. The distribution is skewed heavily to the right.

Bedrooms: Most properties have one or two bedrooms, with relatively few having more than three bedrooms.

Bathrooms: The majority of houses have one bathroom. The number of properties with more than one bathroom falls dramatically.

Distribution of Categorical Variables:



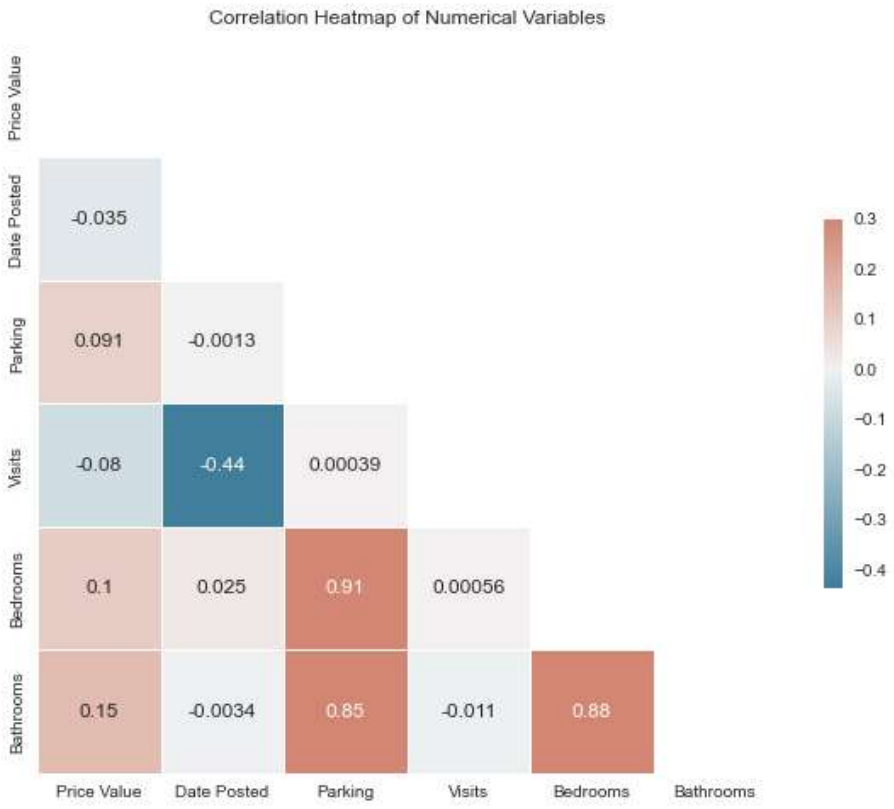
Utilities: "No Utilities Included" appears on most properties. "Some Utilities" appears in a substantial number of properties. "All Utilities" is only found in a few houses.

Agreement Type: The majority of properties have a one-year lease. A few properties offer "month to month" leases. Other sorts of agreements are uncommon in real estate.

Air Conditioning: Most properties do not state whether or not air conditioning is available. The majority of those who do have air conditioning.

Unit Type: Apartments are the most common property type, followed by condos. There are also many properties with the unit type "Unknown" mentioned.The are very few townhouses or houses available.

Correlation Analysis of Numerical Variables:



Bedrooms and Price Value: The pricing value and the number of bedrooms have a moderately favorable link (0.36). This demonstrates that as the number of bedrooms increases, so does the price value.

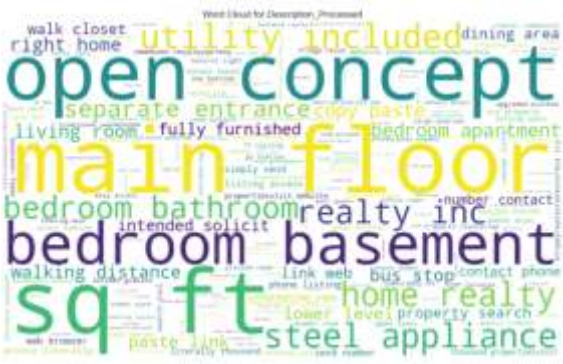
Bathrooms and Price: The price value and the number of bathrooms have a moderately favorable link (0.27). This implies that properties with more bathrooms tend to be more expensive.

Bathrooms and bedrooms: The number of bedrooms and bathrooms has a high positive association (0.57). This makes sense because homes with more bedrooms typically have more bathrooms.

Price Value and Visits: The price value and the number of visits have a weak negative connection (-0.07). This implies that higher-priced properties may get fewer visitors, but the association could be more robust.

Bedrooms/Bathrooms and Visits: Visits have weak negative relationships with the number of bedrooms (-0.07) and bathrooms (-0.06). This implies that properties with more bedrooms or bathrooms may have fewer visitors, but these associations are also relatively weak.

Text Analysis of Listings:



Title: The most common words in the titles of the listings appear to be related to the type and features of the properties, such as "room", "rent", "apartment", "bath", "bed", and "assignment sale".

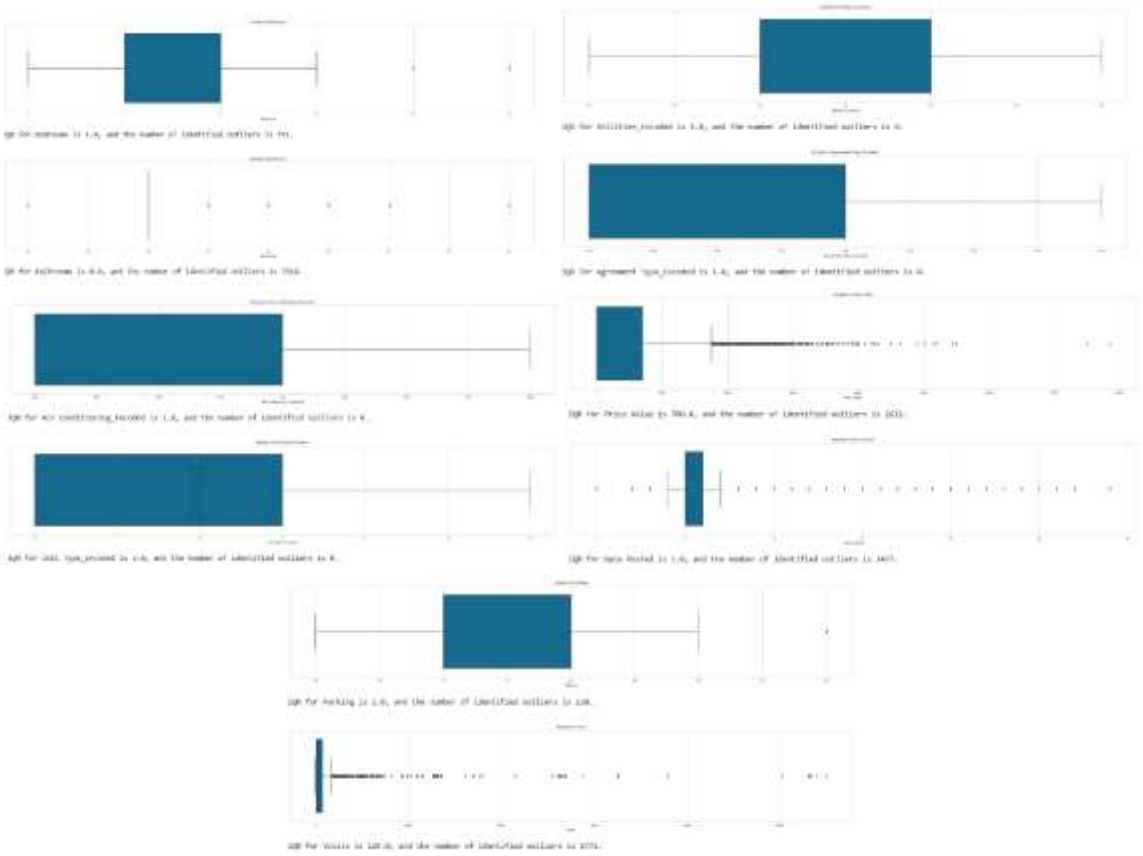
Description: In the descriptions of the listings, we see a similar pattern with words like "bedroom", "main floor", "basement", "open concept", and "utility included" being quite common.

Address: The word cloud for the addresses shows a lot of street names and city names, indicating the locations of the properties.

Encoding Features:

Using the Label Encoder approach, we encode categorical information represented by 'Utilities,' 'Agreement Type,' 'Air Conditioning,' and 'Unit Type.' The Label Encoder is a quick and easy way to convert categorical data into numerical format, making it ideal for machine learning algorithms.

Identifying outliers using Box-Plots and IQR:



Price Value: The boxplot for the "Price Value" feature displays some points that are out of place with the rest of the data, indicating potential outliers. We discovered 2,632 outliers in this category using the IQR-based technique.

Date Posted: Were 3,477 outliers reported for the "Date Posted" attribute. As indicated by the boxplot, these listings were most likely uploaded a while ago compared to the bulk of postings.

Visits: The "Visits" function also displays a few points that deviate significantly from the remainder of the data, indicating potential outliers. We discovered 1,816 outliers in this category using the IQR-based technique.

Bedrooms: As in prior situations, the boxplot for the "Bedrooms" feature shows some points that are out of place with the rest of the data, indicating potential outliers. For this characteristic, the IQR-based technique reveals 731 outliers.

Bathrooms: The boxplot for the "Bathrooms" feature shows some points that are out of place with the rest of the data, indicating potential outliers. For this category, the IQR-based technique reveals 7,918 outliers.

Addressing Outliers using different methods

1. Capping:
   Price Value: There are 18,724 listings in the dataset, with prices ranging from 0 to 700. Prices vary greatly, with an average of around $188.91.

   Date Posted: There are 18,724 listings with posted date information, averaging 5.59 days. The dates have a modest degree of fluctuation.

   Visits: Data for 14,947 postings are available. Listings receive an average of 73.86 visits, with significant variation.

   Bedrooms: There are 18,724 listings in the dataset. Most have one or two bedrooms, with an average of about 1.57 bedrooms.

   Bathrooms: There is one bathroom in each of the 18,724 entries. The number of bathrooms remains constant.

2. Trimming
   Price Value: There are 14,947 listings in the dataset with prices ranging from 0 to 700. Prices vary greatly, with an average of around $163.95.

   Date Posted: There are 14,947 listings with a posted date. The average posted date is approximately 5.62, although this does not appear to be a valid figure for date data.

   Visits: Data for 14,947 listings is available. Listings receive an average of 73.86 visits, with significant variation.

   Bedrooms: There are 14,947 listings in the dataset. The majority of postings have one or two bedrooms, with an average of 1.55 bedrooms.

   Bathrooms are available in all 14,947 listings. Most listings have one bathroom, and all have a value of one for bathrooms due to the lack of decimals.
3. Logging
   The logarithm transformation is a handy tool for dealing with positive skewness in variable distributions, typically encountered in variables like 'Price' or 'Visits'. It can help with variance stabilization and data normalization, which is helpful for many statistical models. After logging, the maximum 'Price Value' and 'Visits' decreased considerably, and their distributions presumably became less skewed. The averages also declined, but the 'Price Value' standard deviation increased, showing better dispersion in the logged prices.

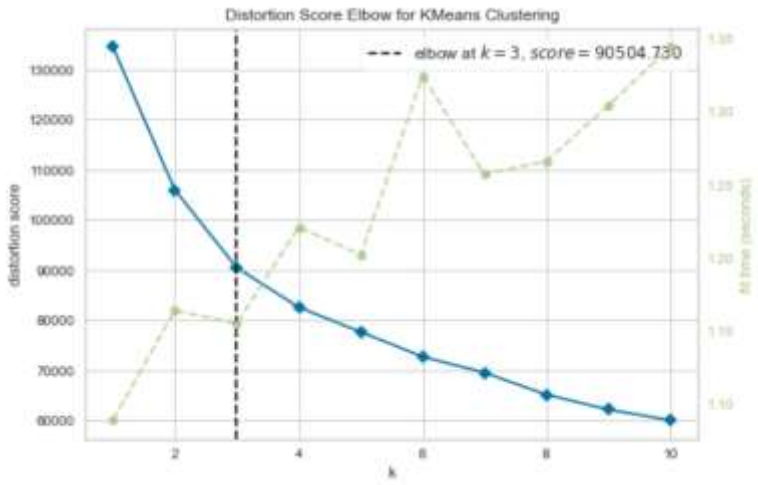
Feature Scaling

The Standard Scaler is being used to scale the numerical attributes 'Price Value,' 'Date Posted,' 'Parking,' 'Visits,' 'Bedrooms,' 'Bathrooms,' 'Utilities_Encoded,' 'Agreement Type_Encoded,' 'Air Conditioning_Encoded,' and 'Unit Type_Encoded.'

The following is a brief description of the procedure:

The Standard Scaler is a popular feature scaling approach that converts numerical data to have a mean of zero and a standard deviation of one. We bring all the characteristics to a single scale by applying the Standard Scaler to the provided numerical columns, ensuring that they contribute equally to the analysis and model training process.

K-Means Clustering with KElbowVisualizer



When the distortion score plot from the KElbowVisualizer is examined, it is clear that the distortion score drops as the number of clusters grows. The decreased pace, however, becomes slower after a certain point, resulting in an elbow-like bend in the plot.

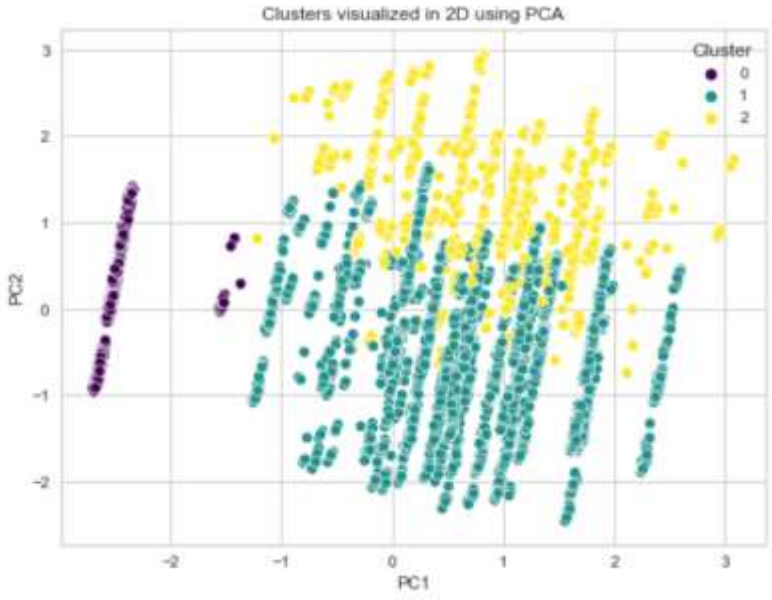
The distortion score plot, in this case, reveals an apparent elbow at roughly 3 clusters. Adding more sets after this stage may result in a less meaningful improvement in clustering performance. As a result, the elbow technique suggests that the optimal number of clusters for this dataset is likely three.

We find a balance between maximizing within-cluster similarity and minimizing between-cluster dissimilarity by selecting three clusters. This optimal number of sets would allow for significant data segmentation, assisting in identifying unique groups or patterns within the dataset.

Overall, the KElbowVisualizer has proven to be an invaluable tool in determining the optimum number of clusters for our K-Means clustering analysis, giving us significant insights to guide further data investigation and interpretation.

Performing PCA on dataset and forming clusters on those features

Principal Component Analysis (PCA) is frequently used before K-means clustering to minimize the dimensionality of the dataset, which improves computational performance and simplifies data visualization. PCA conforms to the assumptions of the K-means algorithm by generating uncorrelated principal components. This technique enhances the performance of K-means clustering in high-dimensional data by making distance more relevant. Furthermore, PCA aids in visualizing the clusters created by the K-means method, mainly when the original data contains more than three dimensions.



Observations from the Scatter Plot of Principal Components (PC1 and PC2):

Separation of Clusters: The scatter plot of the two principal components (PC1 and PC2) shows that the three clusters can be distinguished somewhat. There is, however, some overlap between the clusters. Remembering that PCA is a linear technique that may miss complex, non-linear correlations in the data is vital. As a result, some data points will inevitably overlap.

Data Variability: The first principal component, PC1, accounts for the majority of the variance in the data, followed by PC2, the second principal component. This is a natural feature of PCA because it organizes the ingredients according to the amount of original variance they explain.

Cluster Characteristics: While we can see the separation of data points into clusters, it is essential to note that the axes of this scatter plot (PC1 and PC2) are combinations of the original attributes and do not have a clear, interpretable meaning. Further investigation is required to grasp better what distinguishes each cluster.

Outliers: The scatter plot also illustrates a few remote data points from the main clusters. These data points could be outliers in the original data or represent infrequent but genuine data instances.

Results

We successfully did exploratory data analysis (EDA), data cleaning, preprocessing, and modelling on the provided real estate listings data. The EDA featured visualisations such as histograms, bar graphs, and word clouds to grasp the data's features better. The data was cleaned and prepared for modelling during the preprocessing step. We attempted text preprocessing, such as tokenization and lemmatization.

We attempted to use K-means clustering and the elbow approach to estimate the appropriate number of clusters during the modeling phase. The elbow plot revealed that the best number of groups for the data is most likely three.

Conclusion

The study offered a comprehensive knowledge of the data. Preprocessing processes prepped the data for subsequent analysis, and K-means clustering provided preliminary insights into the data patterns. Although the text pretreatment processes could not be entirely performed due to the limits of the environment, the code provided will enable further examination of the text data.

Future Work

There are various potential future work directions. Implementing advanced text analysis techniques, such as sentiment analysis, and improving predictive modeling to forecast listing prices are among them. More research could include fine-tuning clustering approaches and developing new features to increase model performance. If the dataset size allows, deep learning models could be used for sophisticated analysis. Time series analysis could help find trends if the data has temporal information.

References

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. Retrieved from https://www.springer.com/gp/book/9780387310732

McKinney, W. (2012). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media. Retrieved from https://www.oreilly.com/library/view/python-for-data/9781491957653