Slika na kojoj se prikazuje tekst

Opis je automatski generiran

Department of Computer Science

CS460 Artificial Intelligence

Exploring and Clustering AirBnB Data by Availability

December 2022

Professor: prof. dr. Jasminka Hasić

Student: Begić Faris

# Introduction

AirBnB is a popular online platform that connects individuals looking for short-term accommodations with hosts who have space to rent out. The platform offers a wide variety of rental options, from private rooms in a host's home to entire apartments or houses. AirBnB was founded in 2008 and has since grown to become a major player in the hospitality industry. It is used by millions of people around the world, both to find affordable and unique accommodations while traveling, and as a way for hosts to earn extra income by renting out their space. In addition to offering a wide range of accommodations, AirBnB also provides a range of services for both hosts and guests. These include secure payment systems, 24/7 customer support, and insurance for hosts. Overall, AirBnB has had a major impact on the hospitality industry and has provided travellers with a new and unique way to book accommodations. It has also given hosts the opportunity to earn money from their unused space and has helped boost local economies around the world.

The goal of this project is to gain an understanding of what this dataset contains and to draw conclusions about AirBnB rental data in Istanbul based on the data provided. Exploratory data analysis is an important step toward comprehending and fully understanding a dataset. In this project, we will conduct exploratory data analysis on AirBnB data from Istanbul. Cleaning and preparing the data will be the first step, followed by the use of various data visualization techniques to uncover insights and trends in the data. Our goal is to better understand the city's rental market, as well as the factors that influence pricing and availability. Both travellers and policymakers can benefit from this information. To accomplish the project's objective, data availability will be clustered. The conclusions concerning price and other related features will be drawn based on the clusters created.

# Data Cleansing

The process of identifying and addressing any issues with a dataset is known as data cleansing, also known as data cleaning or data scrubbing. This is an important step in any data analysis project because incorrect or dirty data can lead to inaccurate or misleading results. Data cleansing entails several tasks, such as identifying missing or incorrect values, correcting inconsistencies or errors in the data, and removing duplicates or irrelevant information. Data cleansing is an iterative process, and multiple passes may be required to ensure that the data is clean and ready for analysis. After cleansing the data, it can be used to build machine learning models.

To clean the data, we need to know what kind of data we're dealing with, so we used commands that display the data and count the number of null values in it, because null values can cause problems later in the project when we use aggregate functions. We quickly notice that the neighbourhood group attribute has all null values and does not provide us with any additional information, so we remove the column entirely. Following that, we can see that last review and reviews per month have a large number of null values, and because we decided not to use last review in our exploratory data analysis, we can remove it. Reviews per month, on the other hand, will be used later in our exploratory data analysis and clustering, so we decide to fix null values. There are many ways to recover null values: either take the mean of all similar accommodations and assign it, or simply put 0. In this case, it is more logical to enter 0 as null in reviews per month to make it appear as if there are no reviews per month. Following that, we discovered that we have quite a few hosts with names equal to null, and because many graphs in EDA are based on names, those records are useless to us, so we remove them entirely because there aren't many of them. Following that, we check for duplicate values. Records with identical attributes are considered duplicates, and the function's output shows that there are no duplicate values, so no action is required in this case. Finally, we check the data to see if everything is in order, which it appears to be at first, but as we write a function to get analytical information about the price, we notice that there are some accommodations with prices of 0, which is not normal, and we remove them.

After this, we wanted to check the analysis of prices to see what the max, min and average price and we is concluded that there are some enormous high prices which don’t seem real for accommodations, which implies that we have outliers in our data. To get rid of the outliers we use IQR method. The IQR is calculated by subtracting the 25th percentile from the 75th percentile and represents the range of values that most of the data falls within. To identify and remove outliers from a dataset using the IQR method, first compute the IQR and then define a range of acceptable values by adding and subtracting the IQR from the 25th and 75th percentiles, respectively. Any values that fall outside of this range are considered outliers and may be removed from the dataset. After application of IQR method, we check the analytical data of the price and now the prices seem more realistic than before. We discovered certain problems after undertaking additional data exploration using different graph visualisations to understand our data. Because the minimum nights parameter had some really high values, we decided to plot the values for this attribute in order to give it a closer look. On the plot, we can see that there are a lot of high values that are lower than 400, but there are only a few that are greater than 400. Since those values would alter our results—specifically, normalized price—we had to remove them because this characteristic is so important to our clustering method. We plotted minimum nights once more after clearing up this property, and the graph looked better. Additionally, we realized that lodgings with availability of 0 are not pertinent to us because they are either closed or may be renovating their unit, so we opted to ignore those results in order to obtain the most accurate information. We have completed the data cleaning step, and our data is now ready for exploratory data analysis.

# Exploratory Data Analysis

Exploratory data analysis is a method for analysing and comprehending large amounts of data. It is an important step in the data science process because it allows analysts to discover hidden patterns and relationships in the data as well as develop hypotheses about the data that can be tested and refined further. This can be accomplished using a variety of techniques, including plotting and graphing the data, summarizing the data with descriptive statistics, and applying statistical tests to the data to identify significant relationships. Another advantage of exploratory data analysis is that it allows analysts to gain a better understanding of the data, which can help them develop more advanced analysis methods.

To start exploratory data analysis, we have decided to look again at the statistics of the price, which shows that the cheapest apartment is only 16, whereas the most expensive one is 659, which tells us that Istanbul provides people with accommodation of variety of prices, from the cheap and not so good, to expensive and luxury. Later in this exploratory data analysis we are going to see corelation of busyness to the price which will give us a better insight. There is one thing that is not known to us, and that is whether this price is in Turkish lira or some other currency, as it obviously makes a big difference but considering that the cheapest accommodation is only 16, it is most likely in dollars since 16 Turkish lira is very cheap, but on the other side 659 dollars seems realistic for the luxury apartment. Following that, we checked how many types of apartments do we have and counted which one is the most occurring one and we figured out that Istanbul accommodations on AirBnB are mostly private rooms. After that, we listed the accommodations with most listings and found out that most appearing name for the accommodation is “İstanbul Birden fazla bölümden oluşan bina” which, when translated means “An Istanbul building consisting of more than one part”. After that we wanted to find out which neighbourhood in Istanbul is the biggest on AirBnB in terms of accommodation. The results showed that Beyoglu is the neighbourhood with most accommodations followed by Sisli, Fatih and Kadikoy. Having some previous knowledge about Istanbul, it can be confirmed that this information was not suprising and therefore accurate. After this, we have proceeded to calculate the mean price per neighbourhood and represent the results in a bar graph. Before the conclusion, important thing to mention is that this bar graph can indicate to us whether we have outliers or not. If this bar graph had one value that is drastically higher or lower than others, we would need to do additional checking, but clearly from this visual representation, there are no two adjacent regions that have massive difference, as it would have been if we haven’t performed data cleansing. From this graph, it can be concluded that Arnavutkoy is the region that has the biggest mean price followed by Basaksehir, Catalca and Sile. Based on this and previous graph, it can be seen that the top three most expensive neighbourhoods by the mean are not even in the top 10 neighbourhoods with the most number of accommodations, which means that not many “luxury” apartments are rented on AirBnB. The reason we avoided removing rows that have number of reviews as NaN value is because it would give completely different results in the following graph, and the graph talked about is a bar chart representing busyest hosts. Counting reviews might not be a great indication because the people who stay at the AirBnBs don’t have to leave the rating. However, most of the people are doing that because it helps the community grow, but also some of renters modify price for a good review or the people that leave the rating get some credits by AirBnB, so it Is safe to say that this gives us good insights. From this graph we can see that The Ataturk Airport accommodation is the busiest amongst all, appearing more than three hundred times. It is followed by “Sweet flat in Cihangir Cukurcuma” and “Sultanahmet Gardens” having less than three hundred appearances. The next graph is a scatter plot showing us all apartments on a map, where the accommodations in the same neighbourhood are colored with the same color. This graph is a better visual representation of the bar graph showing neighbourhoods with most accommodations. The next graph is showing us a comparison of two attributes for each row. We compared the price with the number of reviews in order to get an insight how does price affect the busyness of the hosts. This graph clearly indicates that some accommodations in the price range up to the $300 have more than 300 reviews, after which the number of reviews starts decreasing. In general, it can be seen that as the price increases the number of reviews falls, which indicates that the AirBnBs with higher prices have less customers.

# Clustering

For this research, the topic of clustering by availability was selected. This means that we will cluster or group the data according to availability and price, resulting in n groups of AirBnBs that will be clustered according to availability and price. What use can we make of this? The homes in a certain neighborhood with consistently high availability and cheaper prices than homes in other neighborhoods will be included in the final dataframe. As a result, we may locate flats in areas that provide us the best value for our money while they are also available. The K-Means approach is used in this project to perform clustering. The K-means algorithm is a well-liked and straightforward technique for data grouping. Given that it is an unsupervised learning technique, it can be used to identify patterns and clusters in data without the need for pre-established labels or classifications. You must first choose the number of clusters you want to identify in the data before you can utilize the k-means algorithm (this is represented by the "k" in the name of the algorithm). Based on how similar each data point is to other points in the same cluster, the algorithm will then iteratively assign each data point to one of the k clusters. Until the assignments stop changing or a maximum number of iterations is reached, the algorithm will keep updating the cluster assignments and the cluster centres (which are the average of all the points in the cluster). How do you calculate k? The elbow method is a way for figuring out what number of clusters, k, is best to utilize in a k-means clustering analysis. It is based on the hypothesis that the within-cluster sum of squares (WCSS) will drop as the number of clusters rises, but that the decline will eventually become less significant, and when the rate of WCSS decline slows, is regarded as the ideal value of k. We first conducted a k-means clustering analysis for a variety of different values of k before applying the elbow approach. The WCSS for the resulting clusters is determined for each value of k. Then, using a graph with k on the x-axis and WCSS on the y-axis, we plotted the WCSS for each value of k. The elbow point, is the value of k at which the plot begins to slow down bending. We plotted the data after the K-Means method was run, and it is clear from the plot that our clusters are good because there is a clear separation between them and no points that are going from one group to the next. The results may be further improved by eliminating some of the lodgings with high prices and ratings, but doing so would reduce the quantity of records, so we leave it as is.

# Conclusion

AirBnB is a website that connects people who need a place to stay with people who have a room or house to rent out. It offers many different types of rentals, from a single room in someone's home to a whole apartment or house. As every other Machine Leaning project, this project starts with data cleansing as well, as it is important because incorrect or dirty data can lead to inaccurate or misleading results. In this project, various methods were used to clean the data, including removing unnecessary columns and fixing null values. IQR method and statistic functions were used to identify and remove outliers. Following that, performed exploratory data analysis, which is a technique for analysing and fully understanding enormous amounts of information in order to uncover hidden patterns and relationships and generate theories about the records. The exploratory data analysis was used to study the prices of accommodations in Istanbul, and to look for correlations between busyness and price. We also looked at the types of accommodations available and the most popular neighbourhoods on AirBnB in Istanbul. The conclusion is that “Arnavutkoy” is the region with the highest mean price, but that the most expensive neighbourhoods are not among the busiest. It was also found that the busiest host accommodations on AirBnB in Istanbul are "The Ataturk Airport" and "Sweet flat in Cihangir Cukurcuma". Following that, we used the clustering method based on availability and price, and used it to discover the cheapest accommodations in the specified location. To summarize, after locating an appropriate dataset, we performed data cleansing, which enabled us to obta in the most accurate results and draw sound conclusions from the next step, exploratory data analysis.

Sources:

Fehmifratpolat, “Istanbul Airbnb data analysis and visualization,” *Kaggle*, 26-Sep-2020. [Online]. Available: https://www.kaggle.com/code/fehmifratpolat/istanbul-airbnb-data-analysis-and-visualization/data. [Accessed: 08-Dec-2022].