Safety-Aware Airbnb Dashboard for Travelers

Team Members: Jasmit Gill, Pietro Del Bianco, Pranay Reddy

March 10th, 2024

BUDT737: Enterprise Cloud Computing and Big Data

Executive Summary

For many people traveling to new cities is a once-in-a-lifetime experience but most cannot do it due to safety concerns. For people visiting from out of the country who might not know the local DC area we wanted to help solve that problem. For our project we decided we wanted to develop a dashboard that is user-oriented and helps solve the problem by combining Airbnb listings with neighborhood safety information. This tool is designed to help travelers make informed decisions about where to stay in case they are new to the city or are a tourist visiting the area.

We worked with a large dataset containing over 77,000 Airbnb listings across six major US cities, with a special focus on Washington DC. Using PySpark, we processed this data and organized it into individual files for each city. Our main challenge was to define neighborhoods in a way that made sense for both travelers and our analysis.

To do this, we used a machine learning technique called K-Means Clustering. We started with 146 clusters of Airbnb listings in DC and narrowed them down to 30 distinct neighborhoods. We applied the same technique to over 26,000 crime records to match each crime to a neighborhood. This allowed us to create safety indicators for each area.

We developed five key indicators to give travelers a clear picture of neighborhood safety: danger score, police surveillance, theft score, property crime, and crime rate. We then scaled these indicators to make them easier to compare across different neighborhoods.

The final step was to create an interactive map using the Folium Python library. This map lets users click on neighborhoods to see detailed safety information and Airbnb listing details. We used HTML and CSS to make the map look good and ensure that the information was easy to understand.

1. Introduction

1.1 Motivation Behind the Project

The motivation behind this project stems from the recognition that safety is a top priority for travelers when exploring unfamiliar cities but also a great limiting factor when traveling to unfamiliar places. However, finding reliable and comprehensive safety information can be challenging. Many existing tools focus either on accommodation details or safety metrics, but there is a gap in the market for a tool that integrates both. Our project aims to fill this gap by creating a dashboard that provides a holistic view of both Airbnb listings and neighborhood safety data as a one stop solution for anyone traveling.

1.2 Research Objective

The primary objective of our research is to enhance travelers' experiences by offering a useful tool that combines safety information with Airbnb options. Our research objective started from the idea to give travelers a complete picture of how safe different neighborhoods are because a neighborhood can be completely different based on a difference if just a few miles, so they can choose where to stay with confidence. This way, they can find a safe place to stay and have a more enjoyable and worry-free trip. This tool is built upon using a static airbnb data set however this data set is always being constantly updated for everyone's convenience which keeps our tool as updated as well.

2. Dataset Description

Our project utilizes two main datasets: the Airbnb dataset and the crime dataset. Both datasets play a crucial role in providing the necessary information for our safety-aware dashboard.

2.1 Airbnb Dataset

This Airbnb dataset that we sourced from Kaggle contains over 77,000 listings from six major US cities. Each listing includes details such as the location, price, number of bedrooms, and host information. We knew for this project we wanted to reduce our scope area by focusing and perfecting one city. As UMD students we wanted to create a local

impact so we decided to data clean/prep this data set to only include the data for Washington D.C airbnb listings and Washington D.C Crime data. But too manage this large dataset efficiently, we used Spark streaming to process the data and create individual CSV files for each city. This approach allowed us to handle the data in a scalable and organized manner, making it easier to analyze and integrate into our dashboard.

2.2 Crime Dataset

The crime dataset consists of more than 26,000 instances of crime of various different kinds, each representing a reported crime in the same cities covered by the Airbnb dataset. These dataset includes many different features such as the type of crime (e.g., theft, assault), the offense (e.g., robbery, burglary), the weapon involved, the police shift during which the crime occurred, and the geographical coordinates of the crime scene. These features provide a comprehensive view of the crime landscape in each city, which is very important to analyze neighborhood safety.

3. Methodology

Our methodology involves several key steps to process and analyze the data, define neighborhoods, and associate crime data with these neighborhoods.

3.1 Neighborhood Definition (Unsupervised ML)

To define neighborhoods, we used K-Means Clustering, an unsupervised machine learning algorithm. We started with 146 clusters based on the latitude and longitude of Airbnb listings in Washington DC and reduced them to 30 distinct neighborhoods. This reduction helped simplify the analysis and make the dashboard more user-friendly.

3.2 Cluster Definition on Crime Data

We applied the same K-Means Clustering process to the crime data, creating 30 clusters based on the geographical coordinates of reported crimes. This step ensured that the crime data could be accurately associated with the defined neighborhoods.

3.3 Cluster Distance from Neighborhood Center

To associate crime clusters with neighborhoods, we calculated the Euclidean distance between the center of each neighborhood and the center of each crime cluster. We used a custom User-Defined Function (UDF) in PySpark to perform this calculation, allowing us to match each crime to the nearest neighborhood.

3,4 Data Cleaning

To streamline the datasets and improve the efficiency of our analysis, we removed unnecessary columns that did not contribute to the dashboard's functionality. This step helped reduce the complexity of the data and focus on the most relevant information.

3.5 Final Dataset Preparation (SQL)

Finally, we used SQL functions to create a comprehensive dataset for the dashboard. This involved grouping the data by neighborhood, aggregating key statistics (such as average price, total listings, and crime rates), and joining the Airbnb and crime datasets. The result was a dataset where each row represented a neighborhood, with columns displaying both Airbnb and crime features. This dataset served as the foundation for our interactive dashboard.

4. Development of Indicators

To provide travelers with a clear understanding of the safety landscape in different neighborhoods, we developed a set of crime indicators.

4.1 Indicator Overview

The purpose of developing crime indicators is to summarize complex crime data into easily understandable metrics. These indicators help travelers quickly assess the safety of a neighborhood and make informed decisions about their accommodation.

4.2 Indicator Details

We created five crime indicators:

- Danger Score: A weighted sum of violent crimes, including assault, homicide, and sex abuse. This score reflects the overall level of danger in a neighborhood.
- Police Surveillance: This crime indicator focuses on how many shifts police take each day which actually gives an idea of the police presence and monitoring in the neighborhood. More shifts can in correlation to more crime can be inferred that there is more crime in that area relative to a area that has less crime and less monitoring,
- Theft Score: This indicator focuses on different types of thefts, such as auto theft, burglary, and robbery, to provide a measure of the risk of theft in the area.
- Property Crime Score: Combines property related crimes to indicate the likelihood of property damage or loss.
- Crime Rate: The ratio of total crimes to the number of Airbnb listings, giving a sense of how common crime is relative to the number of available accommodations.

4.3 Scaling of Indicators

To make these indicators more interpretable in the context of our dashboard and allow for fair comparisons between neighborhoods, we applied MinMaxScaler to scale each indicator between 0 and 1. This scaling makes sure that the indicators are on the same scale, making it easier for travelers or tourists to understand and compare the safety of different neighborhoods.

5. Interactive Map Creation

A crucial part of our dashboard is the interactive map, which provides a visual representation of the neighborhoods and their safety indicators.

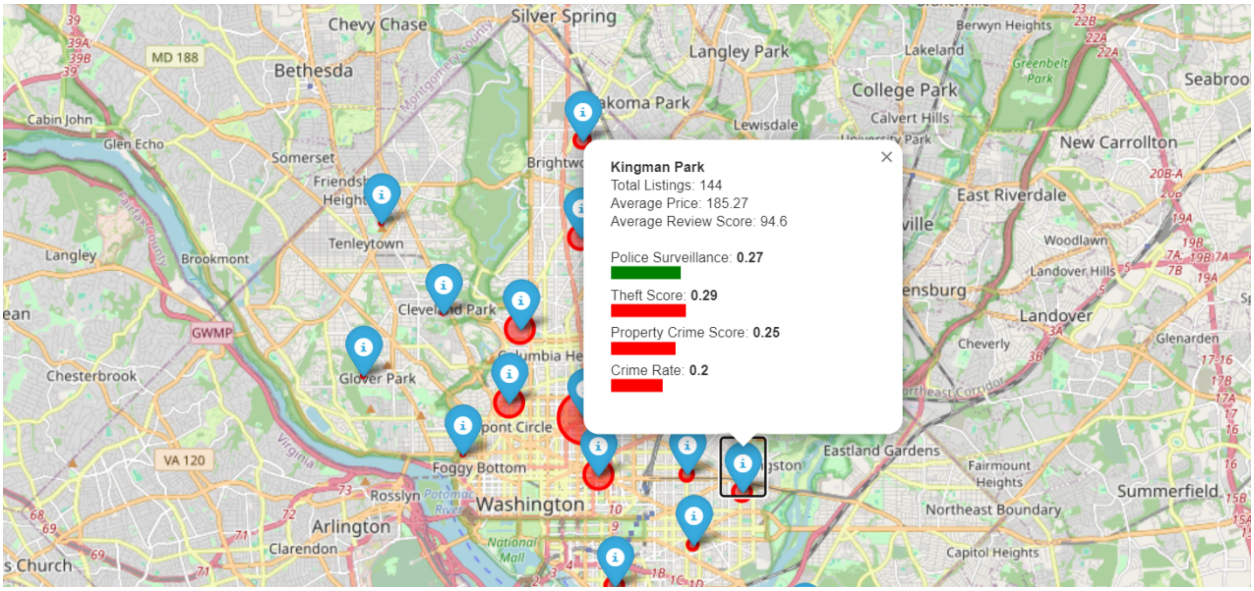
5.1 Leveraging Folium Library

We used the Folium Python library to build the interactive map. Folium allows for the creation of interactive maps with ease and supports various map styles and overlays. We

utilized Folium to plot each neighborhood as a marker on the map, with the location based on the average latitude and longitude of the Airbnb listings in that neighborhood.

5.2 HTML and CSS Integration

To enhance the user experience, we integrated HTML and CSS into the Folium map. This allowed us to add pop-up windows to each marker, displaying detailed information about the neighborhood, such as the name, total listings, average price, and safety indicators. We used HTML to structure the content of the pop-ups and CSS to style the text, create div elements, and format graphs. This customization ensured that the information was presented in a clear, visually appealing manner, making it easy for users to interact with the map and access the data they need to make informed decisions about their stay. For the scope of our project all this data and utilization will be useless if the tool in itself is unreadable so we developed our focus onto the HTML integration for a good quality dashboard.



6. Results

The final dashboard and its features are the final results of our project for future tourists and travelers to D.C who aim to stay safe during their stay, showcasing the integration of Airbnb listings with neighborhood safety indicators. TAn overview of the dashboard and highlights some key findings derived from the data analysis.

6.1 Dashboard Overview

The dashboard is an interactive map built using the Folium library, which displays each neighborhood in Washington DC with markers. By clicking on a marker, users can access a pop-up window that provides detailed information about the neighborhood, including the total number of Airbnb listings, average price, average review score, and the normalized safety indicators: police surveillance, theft score, property crime score, and crime rate. Additionally, a circle marker indicates the danger score, with the radius and color intensity representing the level of danger.

6.2 Key Findings

Neighborhood Clustering: By applying K-Means Clustering to both Airbnb listings and crime data, we successfully identified 30 distinct neighborhoods. This clustering allowed us to analyze safety and Airbnb data at a granular level, providing more relevant and localized insights.

- Safety Indicators: The development of safety indicators, such as the danger score and theft score, offers users a quick and easy way to assess the safety of a neighborhood. These indicators are particularly useful for travelers unfamiliar with the city.
- Price and Safety Correlation: The dashboard reveals interesting correlations between Airbnb prices and safety indicators. For example, neighborhoods with higher danger scores tend to have lower average Airbnb prices, suggesting that safety is a significant factor in pricing.
- User-Friendly Interface: The interactive map and pop-up windows make the dashboard user-friendly and accessible. Travelers can easily navigate the map and access the information they need to make informed decisions about their accommodation.