CSE6242 Final Report – Team 24

Michael Crist, Saajid Haque, Albert Hsueh, and Sergio Roca

Introduction:

When searching for an Airbnb, neighborhood safety can be one of the most important factors to a user. However, Airbnb does not provide any information about safety, and often it is very difficult to independently determine the safety of a particular neighborhood. Our project provides an interactive map of crime rate overlayed with Airbnb property data to quickly inform renters and homeowners and help them choose their most desirable properties based on both crime statistics and listing price. We implemented this solution for 5 major U.S cities: Boston, Chicago, LA, NYC, and San Franscisco. Additionally, our project provides a regression model to allow a prospective Airbnb property owner to estimate the rental price of his/her property. The impact of the project is to provide potential reduction in crime for renters through awareness, reduce the time to book desirable properties, and increase revenue for property owners.

Problem Definition:

Safety is a major concern for a user booking an Airbnb in an unfamiliar area. Presently, users must parse through various data sources to evaluate the safety of a particular neighborhood. Data sources are not centralized, the current practice is time consuming for the user, and may result in inaccurate conclusions. Another problem is that Airbnb property owners do not know what price to list their property to maximize profit. Often, it requires trial and error to determine a price that is consistent with the market. Our project addresses these problems by providing an interactive map of crime rate overlayed with Airbnb property data, thus allowing a user to quickly identify the relative safety of the area in which they are booking an Airbnb. The regression model provided by our project allows an Airbnb property owner to determine what price to list their rental property.

Literature Survey:

Voltes and Sánchez (2020) investigated how locations, property types, and room types affect Airbnb prices. They concluded combining different property types when performing regression analysis is likely not a good approach. Also using regression, Xu, Kim, and Pennington (2018) studied the spatial relationships between Airbnb facilities and crimes. They found Airbnb locations are positively correlated to property crime while negatively to violent crime. Both these papers are useful to guide us in factors and types of regression we considered but have limitations of focusing only in a specific area, and they did not explore how crime rates affect Airbnb revenue. Another article focusing on Airbnb was done by exploring the comments left in reviews (Cheng and Jin 2018). They performed sentiment analysis and found location, amenities and the host were the three key attributes that affect the experience. We take into consideration the key factors found, but a shortcoming was the validity of the sentiment while we focused more on factors that are not subjective.

Maiellaro and Varasano (2017) investigated how an interactive map, using GeoJSON, provides a preferable user interface versus text search. This is useful because it details the use of GeoJSON for building an interactive map. One shortcoming was that it only allows a user to filter categorically, but our tool allows thea user to filter by range, hence providing more precise data. Brantingham and Brantingham (1998) compared violent crimes across cities using crime count, rate, and location quotients (LQCs). This is useful because LQCs provides an additional metric to analyze crime. One shortcoming is that it focuses on crime only in British Columbia, which may not reflect crime patterns in the US. Our project uses crime data for 5 major cities in the US. Ozgur, Hughes, Rogers, and Parveen (2016) identified 7 features that most significantly impact home price and built a linear model to predict home price. This is useful because it provides a framework for determining the most important factors for predicting home price (or Airbnb price). A potential shortcoming is that it is specific to Indiana, which may not characterize major US cities.

Utilizing the same base Insider Airbnb dataset, Ding et al. (2021) investigated what Airbnb attributes contribute customer satisfaction or dissatisfaction using Latent Dirichlet Allocation (LDA) and supervised LDA (sLDA) approaches on consumer reviews. Likewise, McNeil (2020) investigates different host controlled and non-host controlled variables that can influence the price of the rental unit. Both papers show that the location of the unit influences either price or customer satisfaction. To further build off both papers, this project takes a more focused approach on 5 specific cities to visualize this relationship between location and Airbnb listing price. Ihlanfeldt and Mayock (2009) researched the influence of local crime rate and resulting impacts on house prices showing that violent crimes can slightly decrease house value. Similarly, we intend to demonstrate the effects of local crime rate and potential effects on Airbnb rentals. The shortcoming of this paper is that the analysis looks specifically at Miami, Florida while we are looking at multiple cities throughout the US and our project intends to focus on the short-term rental prices versus home purchase prices.

To take into consideration the causal/effect relationship of the variables that we are going to use, we take on the work of Ke, O'Brien, and Heydari (2021) which explores the relationship between increasing Airbnb listings and crime in the city of Boston. This paper helped us in tuning our model by providing us with existing relationships between the parameters that we are going to use, particularly in this case the likelihood that an increase of Airbnb listings in a city will be positively correlated with a potential increase in crime. The paper uses a Difference of Differences regression as the basis of their findings. We took inspiration from their model as a starting point to build upon it by looking at the geospatial and temporal trends (Bate, S. 1987) and use other methods of regression and clustering, such as the multivariate regression model inspired by the one used in Reinhard D (2021).

Proposed Methods:

Our approach was to consolidate, in a single visualization, both crime data and Airbnb listings of interest to the user. One traditional way of seeing the relationship between crime and Airbnb listing would have been to look at the results in long lists of tables per city. Our approach aimed to create a unified visual solution so that the user can interact and interpret, in a distinct visual manner, any relationships of interest. The interactive visualization we created is user friendly, and easy to navigate, it allows the user to search through Airbnb listings in different neighborhoods within a city. During our literature survey, we did not come across a solution that both combined Airbnb listing with crime statistics at the city level. Part of the novelty in our approach was to use analytics to show in one single visual solution both crime and Airbnb listings, in a way that can inform the user of certain relationships between past crime occurrences and prospective listings of interest.

The overall process flow of our application is described as follows: the app first provides the user with a front-end interface, in which the user inputs a city of interest from our 5 options, as well as desired Airbnb property features (see Figure 1). Once the user hits "Submit", the interactive map is opened, revealing a map of the city (see Figure 2). Our regression model provides a nightly price prediction for the Airbnb, based upon the user inputs. The map displays locations of Airbnb listings that are in a similar price range to the predicted price. The user can click on any of the Airbnb listings, and a hyperlink to the listing and some select Airbnb listing features will be displayed (see Figure 3). Additionally, the map overlays the Airbnb locations with a heat map of crime rates for each neighborhood in the city. Areas with higher crime are represented by darker color in the heat map. When the user hovers the cursor over a particular neighborhood, the neighborhood name and crime rate is displayed, as shown in Figure 4. The sections below provide a detailed explanation of the approach taken for each building block of our project.

Data Cleaning and Regression: Our project provides a regression model to allow a prospective Airbnb property owner to estimate the rental price of his/her property. The user inputs select features of his/her property, and the model provides an estimated nightly rental price for the property, as shown in figures 1 and 2. This allows a property owner to quickly identify an appropriate listing price, without requiring trial and error to determine what the market will accept. To build the regression model, we first utilized OpenRefine to clean the Airbnb datasets used to train the regression model. Next, we created several linear regression models, using differing input features and optimization methods. We evaluated the performance of each model based upon R-squared and mean-squared error values, in addition to model complexity. Our final regression model used the following input features: city, room type, # of accommodates, # of bathrooms, # of beds, and neighborhood. The model was created using the *statsmodels* package in Python. Specifically, we used the OLS() function to generate our linear regression model. The model was optimized by removing highly influential outliers via cook's distance. Highly influential outliers were identified as any datapoints with a cook's distance greater than 2x the mean cook's distance for all datapoints. More information on model performance can be found in the *Experiments / Evaluation* section. Additionally, to extract the amenities from our raw data (which were all together in a single column) we used the R language and its base packages to clean, extract, and rank the amenities to create word clouds.

Interactive Map for Airbnbs and Crime: Our strategy is to provide an easy-to-use interface overlaying crime data and Airbnb listings. The Python libraries *Folium*, *GeoPandas*, *Pandas*, and *Numpy* were used to build the visualization for spatial analysis and aggregation of crime data. *Folium* is a Python wrapper for Leaflet.js which is an open-source JavaScript library for plotting interactive maps. The interactive visualization includes a choropleth map that indicates the crime rates in different neighborhoods in a city. Crime statistics dating back one year for the local area are used to create this part of the visualization. Based off the latitude and longitude coordinate for each crime, the crime is mapped to a neighborhood in a city. This mapping is completed using functions in the *GeoPandas* library. Once all the crimes are mapped, the heat map is created using the counts of crime in each neighborhood. The darker colored neighborhoods indicate a higher count of crime. Once the choropleth map is created, we add markers for the Airbnb properties that are likely of interest to the user with a hyperlink to the Airbnb listing. In addition, with the cleaned amenities data we utilized the *wordcloud* package in Python to create word cloud visual representations of popular amenities unique to each city. The larger text size represents higher ranking amenity, as shown in figure 5.

Web Application: In order to incorporate the regression model, map scripts, and take the user inputs from HTML, Flask was used to develop a web application. After the user inputs its desired Airbnb features on the webpage, the server sends this information to a python script which runs the regression model. The model returns an estimated nightly rental price. After estimating the price, the server uses another python script to create the visualization with the heat map and Airbnb listing layers. This map is then embedded into the bottom of the main web page along with a word cloud for the most common amenities for the selected city. Figures 1 through 5 show an example of our application.

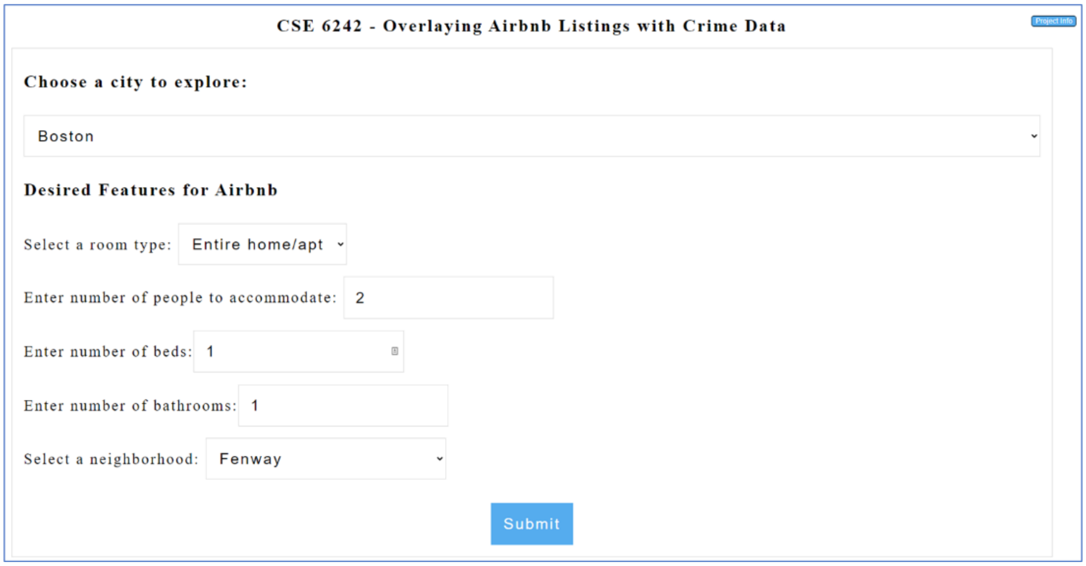

Figure 1. Front end: user selects city and desired Airbnb features

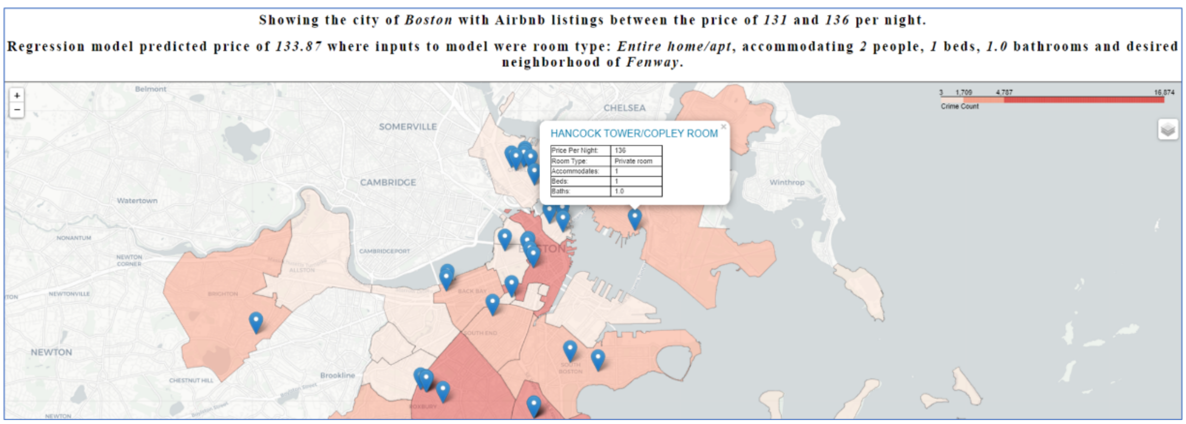

Figure 2. Interactive Map: User is taken to Interactive Map displaying heat map of crimes overlayed with Airbnb properties filtered by the regression model price prediction for the given user inputs

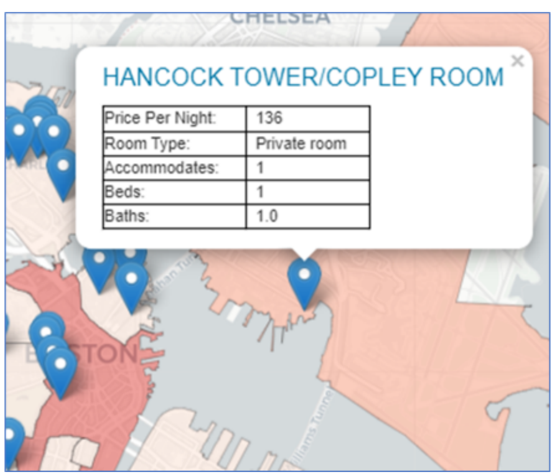

Figure 3. Zoomed capture displaying Airbnb property feature data that is displayed when a user clicks on a listing

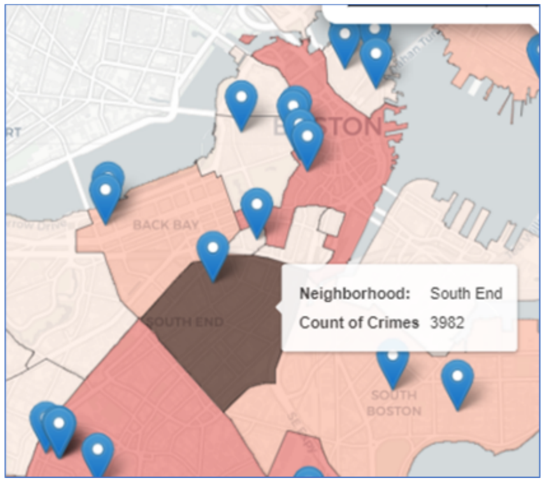

Figure 4. Zoomed capture displaying crime and neighborhood data that is displayed when a user hovers cursor over a particular neighborhood

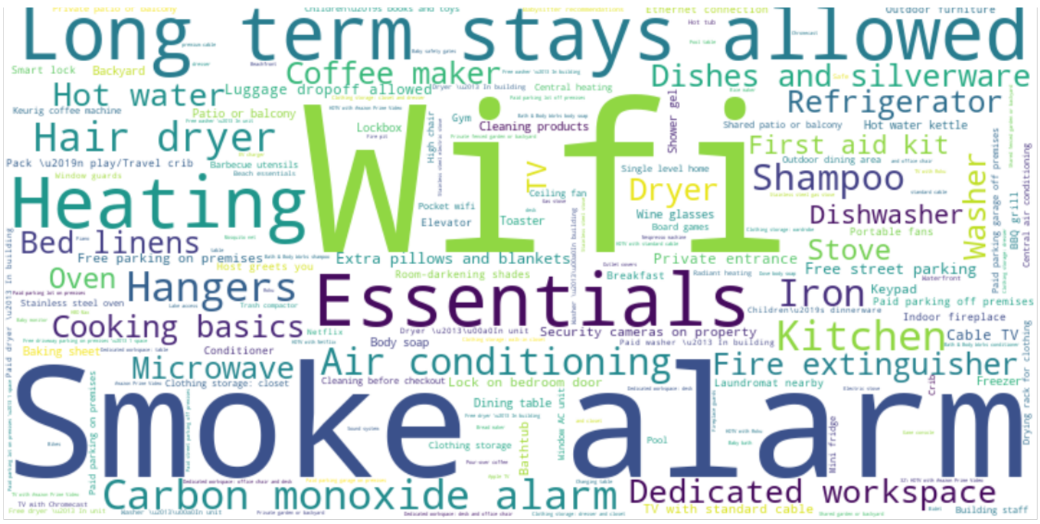

Figure 5. Word cloud based upon common amenities

Experiments / Evaluation:

When building our regression model, we attempted several approaches. We varied the independent variables used in the model to optimize model performance, and we also took different approaches to the model including detection of highly influential outliers using cook's distance. We evaluated each regression approach using R-squared and mean squared error values and selected the best performing model to use for the final project based upon these values in addition to practicality (i.e., we need to use a small enough feature set such that it is reasonable for a user to input all the necessary features). For the visualization, we evaluated our approaches based upon load/execution time and user experience. In a real-world application, we would execute the tests below for additional evaluation of our visualization approach. However, for the scope of this educational exercise, we are not actually able to run these tests.

- Compare a user's average time spent using our tool vs Airbnb website + manually searching crime data. This could be performed by selecting participants for a randomized controlled trial. This would provide us with a quantifiable metric to show how our tool saves a user's time.

- Perform an A/B test to measure changes in Airbnb revenue and crime incidence upon deploying our tool. This would be done by deploying our tool for only select cities and evaluating the treatment results in comparison to the control.
- Survey users to understand what filterable attributes are desired and which can be removed.

Results: Our final regression model achieved an average R-squared value of 0.55, and an average MSE value of 13395, averaged between the 5 cities. The model performance for each city can be found below:

City	Boston	Chicago	LA	NYC	San Fran
Model R-squared	0.64	0.47	0.57	0.56	0.49
Model MSE	2672.80	8573.69	45392.19	4481.69	5855.21

The result of the visualization is an interactive map for each city with a choropleth map of the crime per neighborhood overlayed with Airbnb listings that the user is likely interested in. The map loads quickly once the inputs are submitted due to optimization efforts. Mapping each crime to a neighborhood using *GeoPandas* is the most time-consuming process for the Python scripts. Since the crime data is static for this project, we precomputed this information once and used it in our application leading to faster load times and a better user experience. Before performing this optimization, the load time for each map was roughly a minute on average (load times vary per city). Now with the performance tuning, the maps load in a few seconds on average. The testbed for the timing was running the application locally.

Comparison to other approaches: Our solution improves upon other approaches in this problem space in that it provides a unique display of both crime and Airbnb listings, as well as only showing listings in the price range of our regression results. This provides a level of custom results to each user, along with the crime heatmap that does not exist in other solutions.

Conclusions and discussion:

Our project provides a user-friendly map interface that overlays Airbnb listings with crime data. This is an innovative tool that allows a user to quickly identify the relative safety of the area in which they are listing or renting an Airbnb. Additionally, our project provides a regression model to allow a prospective Airbnb property owner or renter to estimate the prospective rental price of a property, see similarly priced listings around the city of interest, and view most common amenities. All team members contributed a similar amount of effort to this project.

Bibliography

Bates, S. “Spatial and Temporal Analysis of Crime” US Department of Justice Research Bulleting (1987)

Brantingham, Patricia L., and Paul J. Brantingham. "Mapping crime for analytic purposes: location quotients, counts and rates." Crime mapping and crime prevention 8 (1998): 263-288.

Cheng, Mingming, and Xin Jin. “What Do Airbnb Users Care about? An Analysis of Online Review Comments.” International Journal of Hospitality Management. Pergamon, May 1, 2018.

Ding, Kai, Wei Chong Choo, Keng Yap Ng, Siew Imm Ng, and Pu Song. “Exploring Sources of Satisfaction and Dissatisfaction in Airbnb Accommodation Using Unsupervised and Supervised Topic Modeling.” *Frontiers in Psychology* 12 (2021). https://doi.org/10.3389/fpsyg.2021.659481.

Ihlanfeldt, Keith, and Thomas Mayock. "Crime and housing prices." In Handbook on the Economics of Crime. Edward Elgar Publishing, 2010.

Ke L, T. O’Brien D, Heydari B (2021) Airbnb and neighborhood crime: The incursion of tourists or the erosion of local social dynamics? PLoS ONE 16(7): e0253315.

Maiellaro, Nicola, and Antonietta Varasano. 2017. "One-Page Multimedia Interactive Map" ISPRS International Journal of Geo-Information 6, no. 2: 34. https://doi.org/10.3390/ijgi6020034

McNeil, Brandon, "Price Prediction in the Sharing Economy: A Case Study with Airbnb data" (2020). Honors Theses and Capstones. 504.

Ozgur, Ceyhun, Zachariah Hughes, Grace Rogers, and Sufia Parveen. "Multiple linear regression applications in real estate pricing." International Journal of Mathematics and Statistics Invention (IJMSI) 4, no. 8 (2016).

Reinhard, D. “The Influence of Vacation Home Rentals on Neighborhood Crime and Disorder.” Am J Crim Just (2021).

Voltes-Dorta, Augusto, and Agustín Sánchez-Medina. "Drivers of Airbnb prices according to property/room type, season and location: A regression approach." Journal of Hospitality and Tourism Management 45 (2020): 266-275.

Xu, Y., J. Kim, and L. Pennington-Gray. "Explore the spatial relationship between Airbnb rental and crime. Tourism Travel and Research Association." In 2017 International Conference. Retrieved April. 2018.