Visualizing a dashboard for Airbnb Data in New York

(ISTE 608 Database Design Implementation)

by

Gulnaaz Shaikh
Parth Keyur Gawande
Vikashini Sabarinathan

Rochester Institute of Technology

B. Thomas Golisano College
of
Computing and Information Sciences

School of Information

April 4, 2023

Table of Contents

Introduction

Background

With the help of Airbnb, people are making money by renting out their houses, flats, and even individual rooms to tourists. Airbnb has developed into one of the greatest hospitality businesses in the world. Travelers can choose from a variety of lodging alternatives on Airbnb, including unusual homes like treehouses, yurts, and houseboats. In short, Airbnb has become the Uber of the housing market today. The system enables hosts to list their properties while enabling visitors to look for and reserve lodging based on location, availability, and cost. By giving guests more individualized and affordable lodging alternatives while also giving hosts a second source of income, Airbnb has upended the conventional hotel sector.

Problem Statement

The growing demand for Airbnb has led to room for improvement. Identifying areas of the business that are underperforming, overutilized, and need to be focused on cannot be done using traditional methods. Data visualization, on the other hand, can provide meaningful insights and determine trends, patterns, and outliers which cannot be identified just by looking at raw data.

Relevance

Business decisions can be made based on visualizing data in the form of a dashboard. The key performance indicators (KPIs), such as reviews, rates, occupancy demands, etc., can be visualized. Progress of businesses and tracking of goals in real-time can be made using dashboards. The areas that require attention and lack performance are also identified using dashboards.

Goals

The aim of this proposal is to visualize data and create a dashboard that provides information that is useful for a better understanding of raw data, which helps in making better decisions for the Airbnb business. The goal is to design a data visualization dashboard for Airbnb data which can help business owners make data-driven decisions and improve overall performance

Literature review

The paper "A Study on the Performance of Airbnb Listings in New York City" by Walton et al. used a quantitative research method. The authors analyzed data from Inside Airbnb, a website that scrapes and makes available Airbnb data and conducted statistical analyses to examine the performance of Airbnb listings in New York City. The researchers gathered information on price, location, host specifics, and reviews from 44,687 Airbnb listings in the metropolis. They used multiple regression analysis to identify the factors that influence Airbnb listing prices, occupancy rates, and review scores. The research discovered that the most important variables affecting the success of an Airbnb advertisement were location, the number of bedrooms, and the host response rate. The research also discovered that listings with instant booking options had higher occupancy rates than listings without. The authors also performed Naive Baye’s classifier analysis on price_comaprison to investigate the effect of price in relation to the typical expense of a hotel bed in each of New York City's boroughs as they compete directly with Airbnb listings. The researchers also performed text analysis on the reviews given by users and turned that into useful information by getting review scores.The researchers came to the conclusion that a number of variables, including location, pricing, and host responsiveness, influence the success of an Airbnb listing in New York City.

In the context of Airbnb, the paper titled "Towards a Machine Learning and Data Mining Approach to identify customer satisfaction factors" by Chiny et al. (2021) suggests a method for identifying variables that influence customer satisfaction. Data from Airbnb reviews from December 2009 to April 2020 in London were gathered for the research, and it was then examined using a number of methodologies, using NLP to analyze the reviews and then training the Gradient Boosting Regression and Multiple Linear Regression to determine the weights of the six elementary scores noted on Airbnb. According to the findings, the truthfulness of the listing description, the host's communication skills, the cleanliness of the property, and the location of the property are the factors that have the biggest impact on guest happiness on Airbnb. The authors contend that by assisting hosts, these results may be used to raise the overall quality of the Airbnb experience.

This article uses ordinary least squares (OLS) methods and geographically-weighted regression (GWR) to investigate the factors that influence Airbnb prices in Bristol. The dataset consists of 2056 Airbnb listings in Bristol, United Kingdom. It is conceivable that not all room types' prices can be explained by the same set of price factors because the estimated models have dramatically different levels of goodness-of-fit. The research also reveals statistically significant differences between the price determinants of apartment and home listings, as well as spatial patterns in price effects.These findings have an effect on both the evaluation of competition and price setting, and future study should account for potential variations in properties and room types as well as the spatial variability of the estimated coefficients.

Furthermore, the paper "Airbnb Pricing Based on Statistical Machine Learning Models" by Liu (2021) proposes a method for predicting the price of Airbnb listings using statistical machine learning models. The study uses data collected from Airbnb listings in three cities: Boston, San Francisco, and Seattle. The author proposes a novel approach for data preprocessing by creating new features, such as distance to tourist attractions and availability of public transportation, to improve the accuracy of the models. The study then compares the performance of several machine learning algorithms, including random forest, linear regression, support vector regression, and decision tree, to predict the price of Airbnb listings. The results show that the random forest model outperforms the other models in terms of accuracy, with an R-squared value of 0.76. The study also highlights the importance of including location-based features in the model, as they are significant predictors of price. Overall, the paper provides a useful framework for predicting the price of Airbnb listings using statistical machine learning models, which could help hosts optimize their pricing strategies and improve the user experience for guests.

Similarly, an Airbnb Data case study by Sinthong and Carey (2021) presents a study on using database-backed data frames for exploratory data analysis of Airbnb data. The study focuses on the data from Airbnb listings in Los Angeles and uses Apache Spark and Databricks as the computing platform. The authors demonstrate how to perform exploratory data analysis using SQL and data frames in Databricks, which enables faster and more efficient data processing compared to traditional methods. They also illustrate the benefits of using database-backed data frames for exploratory data analysis, such as the ability to handle large datasets and the ability to perform complex data transformations. The study provides insights into the characteristics of Airbnb listings in Los Angeles, such as the distribution of listing types, the average price by neighborhood, and the availability of amenities. The goal of this paper is quite similar to the idea proposed by us. The intention is to more likely create a dashboard and analyze Airbnb data through visualization.

Another case study by Subroyen et al. (2023) analyses and visualizes topic trends in Airbnb reviews using natural language processing techniques. The target city is Amsterdam, where the study focuses on reviews from the market and uses topic modeling and visualization tools to extract and visualize the main topics and trends in the reviews. The authors use Latent Dirichlet Allocation (LDA) to identify the underlying topics in the reviews and then visualize the results using a network graph, which shows the relationships between the topics. The study identifies several key topics, including location, cleanliness, communication, and value for money, and visualizes their relationships. The paper demonstrates the usefulness of topic analysis and visualization for understanding customer feedback and identifying trends in peer-to-peer platform data. The authors argue that such insights can help platform operators and hosts improve the customer experience and make more informed decisions.

Coles, Egesdal, et al. (2017), in the paper Airbnb Usage Across New York City Neighborhoods: Geographic Patterns and Regulatory Implications, the authors examined the usage of Airbnb in New York City and to know how dispersed they are in New York City. They have used the data collected from Airbnb, used data from the city government sources, and surveys to know how spread Airbnb usage is. The rapid growth of Airbnb has become challenging for all other local housing markets and hotel industries. They have found that Airbnb is often used in high-density, tourist-oriented neighborhoods like Manhattan and Brooklyn. Also, in this study, they have found that the landlords are looking for more profits by renting the property for a short period of time. Also, people are mostly looking to rent the entire house or apartment rather than shared rooms, because of this, the landlords are trying to make more profits out of it. These are the few insights we are getting from this paper.

Yang, García, et al. (2022) wrote a paper, Competitors or Complements: A Meta-analysis of the Effect of Airbnb on Hotel Performance, in which they conducted a meta-analysis of existing papers to know the impact of Airbnb on the hotel industry. The authors have collected around fifty-five existing papers to examine the impact of Airbnb on hotel performance. There are factors like price, quality, and location that influence the relationship between Airbnb and the hotels. In the study, they found that Airbnb has a more negative effect on the markets with lower hotel prices and lower hotel quality because Airbnb is a budget-friendly option compared to other hotels. Also, there is only less negative impact of Airbnb on higher-end hotels.

The paper “A socio-economic analysis of Airbnb in New York City” by Dudás et al. (2017) has analyzed the socio-economic factors influencing the dispersion of Airbnb in America. They have collected data from Airbnb and many other sources. They also tried to learn about Airbnb's impact on the hotel industry, the housing market, and the economy. The authors have done a statistical analysis (correlation matrix and regression analysis) to know the socio-economic condition of Airbnb in New York City. The author is trying to say that there are only a few rental properties around New York because the landlords are trying to convert most of the long-term rental properties to short-term Airbnb properties in order to get more profit. Also, when comparing Airbnb to the hotel industry, there is not much impact on the hotel industry. Airbnb has a significant effect on the economy. They also found that Airbnb is more popular among the young population.

Data Sources

A dataset on Airbnb by Sandeep Majumdar is taken from Kaggle.com which will be used for creating and visualizing a dashboard on Airbnb. This dataset will be visualized to help business owners analyze and gain insights that cannot be interpreted by looking at the raw data alone.

Dataset Description

The "airbnbnyccleaned" dataset (Airbnb-NYC-Cleaned, 2022) contains information about Airbnb listings in New York City, USA. It includes data on the listings' location, host information, availability, prices, and reviews.

Here are the column names and a brief description of each:

- **id**: the unique identifier of the listing
- **name**: the name of the listing
- **host_id**: the unique identifier of the host
- **host_name**: the name of the host
- **host_identity_verified**: whether the host is verified by Airbnb or not.
- **neighbourhood_group**: the borough where the listing is located (Bronx, Brooklyn, Manhattan, Queens, or Staten Island)
- **neighbourhood**: the name of the neighborhood where the listing is located
- **lat**: the latitude of the listing's location
- **long**: the longitude of the listing's location
- **instant_booking**: TRUE if yes, FALSE if no.
- **cancellation_policy**: the type of cancellation, strict, moderate, or flexible.
- **room_type**: the type of room being listed (Private room, Entire home/apt, or Shared room)
- **construction_year**: The year in which the property was constructed.
- **price**: the nightly price of the listing in USD
- **service _fee**: The service fee charged by the property in USD.
- **minimum_nights**: the minimum number of nights required to book the listing
- **number_of_reviews**: the number of reviews the listing has received
- **last_review**: the date of the most recent review
- **reviews_per_month**: the number of reviews per month for the listing
- **review_rate_number**: Gives the rate number of reviews
- **calculated_host_listings_count**: the number of listings that the host has
- **availability_365:** the number of days the listing is available for booking in the next 365 days
- **house_rule**: Rules given by property owners that should be followed by all visitors

Methodology

Procedure

The project intends to visualize a dashboard for the Airbnb dataset from New York. The data visualization will be done using Google Big Query. Various visualizations, such as box plots, line charts, bar graphs, pivot tables, scatter graphs, histograms, pie charts, etc., would be used.

Based on the "airbnb_nyc_cleaned" dataset, some dashboard slides that can be created using Google BigQuery and Data Looker Studio are:

- Overview of Airbnb listings in NYC: Use a map visualization to show the distribution of listings across the different boroughs and neighborhoods in NYC.
- Price distribution of listings: Use a histogram or box plot to show the distribution of listing prices in NYC, broken down by borough or neighborhood.
- Availability of listings: Use a bar chart or line chart to show the number of available listings over time, broken down by borough or neighborhood.
- Room types: Use a pie chart or bar chart to show the breakdown of different types of listings (private room, entire home/apartment, shared room) in NYC.
- Host verification: Use a bar chart or pie chart to show the proportion of hosts who have been verified by Airbnb, broken down by borough or neighborhood.
- Reviews and ratings: Use a scatter plot or bubble chart to show the relationship between the number of reviews and the overall rating of listings in NYC.
- Cancellation policies: Use a pie chart or bar chart to show the breakdown of different types of cancellation policies (strict, moderate, flexible) for listings in NYC.
- Property construction year: Use a histogram or box plot to show the distribution of property construction years for listings in NYC, broken down by borough or neighborhood.
- Hosts with multiple listings: Use a bar chart or line chart to show the distribution of the number of listings per host in NYC, broken down by borough or neighborhood.
- House rules: Use a word cloud or bar chart to show the most common house rules across all listings in NYC or broken down by borough or neighborhood

These dashboard slides can provide insights into the performance of listings, hosts, and neighborhoods on Airbnb in New York City and help users make data-driven decisions related to pricing, marketing, and customer service.

Deliverables

The following will be delivered at the end of the project

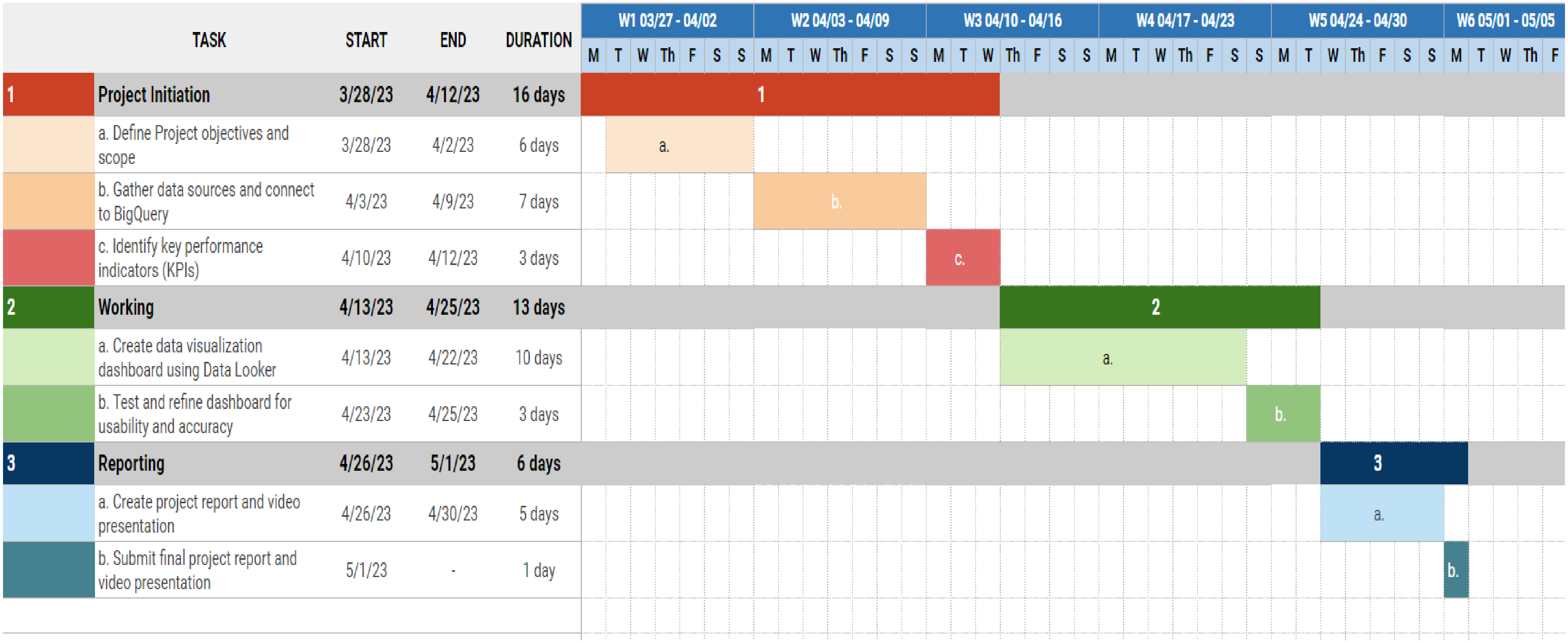
a. Project Report
b. Final PowerPoint Presentation
c. A video post on Youtube explaining the code and each individual dashboard slide.
d. Source code of the project with the necessary comments.

Project Plan

The following table depicts the timeline to be followed for this project

Task Description	Start Date	End Date	Duration
Define project objectives and scope	March 28, 2023	April 2, 2023	6 days
Gather data sources and connect to BigQuery	April 3, 2023	April 9, 2023	7 days
Identify key performance indicators (KPIs)	April 10, 2023	April 12, 2023	3 days
Create a data visualization dashboard using Datalooker	April 13, 2023	April 22, 2023	10 days
Test and refine the dashboard for usability and accuracy	April 23, 2023	April 25, 2023	3 days
Create project report and video presentation	April 26, 2023	April 30, 2023	5 days
Submit the final project report and video presentation	May 1, 2023	May 1, 2023	1 day

Anticipated timeline represented by a gantt chart



	TASK	START	END	DURATION
1	**Project Initiation**	**3/28/23**	**4/12/23**	**16 days**
	a. Define Project objectives and scope	3/28/23	4/2/23	6 days
	b. Gather data sources and connect to BigQuery	4/3/23	4/9/23	7 days
	c. Identify key performance indicators (KPIs)	4/10/23	4/12/23	3 days
2	**Working**	**4/13/23**	**4/25/23**	**13 days**
	a. Create data visualization dashboard using Data Looker	4/13/23	4/22/23	10 days
	b. Test and refine dashboard for usability and accuracy	4/23/23	4/25/23	3 days
3	**Reporting**	**4/26/23**	**5/1/23**	**6 days**
	a. Create project report and video presentation	4/26/23	4/30/23	5 days
	b. Submit final project report and video presentation	5/1/23	-	1 day

Meta- Analysis

Based on the information provided, here is a breakdown of what parts of the meta-analysis are easier or harder:

Easy:

- Identifying the dataset columns and their descriptions
- Describing the types of dashboard slides that can be created using the dataset and visualizations
- Listing the different types of charts and graphs that can be used for each slide

Hard:

- Actually creating the dashboard slides using Google BigQuery and Data Looker Studio (this would require technical skills and familiarity with these tools)
- Analyzing the data in more depth, beyond what is suggested in the list of potential dashboard slides (this would require a deeper understanding of statistical analysis and data science techniques)
- Drawing conclusions or making recommendations based on the analysis (this would require expertise in the specific field or industry being analyzed)

References

Airbnb-NYC-Cleaned. (2022b, August 25). Kaggle. https://www.kaggle.com/datasets/sandeepmajumdar/airbnbnyccleaned

Chiny, M., Bencharef, O., & Chihab, Y. (2021). Towards a Machine Learning and Data Mining approach to identify customer satisfaction factors on Airbnb. 2021 7th International Conference on Optimization and Applications (ICOA), 1–5. https://doi.org/10.1109/ICOA51614.2021.9442657

Coles, P. A., Egesdal, M., Ellen, I. G., Li, X., & Sundararajan, A. (2017). Airbnb Usage Across New York City Neighborhoods: Geographic Patterns and Regulatory Implications. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3048397

Dudás, G., Vida, G., Kovalcsik, T., Boros, L. (2017). A socio-economic analysis of Airbnb in New York City. Regional Statistics. https://doi.org/10.15196/RS07108

Liu, Y. (2021). Airbnb Pricing Based on Statistical Machine Learning Models. *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*, 175–185. https://doi.org/10.1109/CONF-SPML54095.2021.00042

Sinthong, P., & Carey, M. J. (2021). Exploratory Data Analysis with Database-backed Dataframes: A Case Study on Airbnb Data. *2021 IEEE International Conference on Big Data (Big Data)*, 3119–3129. https://doi.org/10.1109/BigData52589.2021.9671603

Subroyen, J., Turpin, M., de Waal, A., & Van Belle, J.-P. (2023). Topic Analysis and Visualisation of Peer-to-Peer Platform Data: An Airbnb Case Study. In A. Shukla, B. K. Murthy, N. Hasteer, & J.-P. Van Belle (Eds.), *Computational Intelligence*

(Vol. 968, pp. 157–166). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-7346-8_14

Voltes-Dorta, A., & Sánchez-Medina, A. (2020). Drivers of Airbnb prices according to property/room type, season and location: A regression approach. *Journal of Hospitality and Tourism Management*, *45*, 266–275. https://doi.org/10.1016/j.jhtm.2020.08.015

Walton, C., Williams, T., & Sari, T. (n.d.). A Study on the Performance of Airbnb Listings in New York City. https://ursa.mercer.edu/bitstream/handle/10898/12383/P24_Walto_C_BUS.pdf?sequence=1&isAllowed=y

Yang, Y., Nieto García, M., Viglia, G., & Nicolau, J. L. (2022). Competitors or Complements: A Meta-analysis of the Effect of Airbnb on Hotel Performance. Journal of Travel Research, 61(7), 1508–1527. https://doi.org/10.1177/00472875211042670