MSc in Business Administration and Data Science

Final Project Data Management and Visualization

Report Data Management and Visualization

Alexandros Kyriakopoulos; 167290

Arthur Söhler; 167336

Francesca Salute; 167284

Jan Bendix Portius; 167605

Course Coordinator: Abid Hussain

Date: 05/01/2024

Pages: 24

Characters: 35645

Dashboard: Click here to visit the Dashboard

Contents

Executive Summary

This project aims to enhance the understanding of Airbnb guests' preferences and ultimately increase guest satisfaction. It provides insights for management and corresponding business recommendations based on an exploratory data analysis (EDA) and a dashboard. The report provides detailed information about the methods used, an in-depth discussion of the findings, as well as clear business recommendations. The dashboard provides condensed information for management, facilitating decision-making to increase guest satisfaction.

The workflow of data cleaning before exporting the data to SQL consisted of getting an overview, creating a data dictionary, converting data types, and replacing invalid values. Subsequently, the data was exported to PostgreSQL, and used for building the dashboard. Further preprocessing steps for the EDA consisted of handling missing values, dropping duplicates and unnecessary columns, performing additional transformations on the data, and developing additional attributes.

Key Result	**Key Business Recommendation**
Market gap for one-night stays (mean minimum stay is 4.6 nights, mode is 2 nights)	Expand offerings for one-night stays, targeting business travelers to increase sales and guest satisfaction
Mode of reviews per listing is 0	Encourage more guest reviews, especially for new listings, to build trust and enhance guest satisfaction
Shared accommodations have lower satisfaction ratings compared to private ones	Improve and refine strategies for shared accommodations to enhance guest satisfaction
Amenities don't significantly impact overall ratings	Reassess the emphasis on amenities; conduct further research to understand guest preferences accurately
Long-time hosts have higher overall satisfaction ratings than more recent ones	Keep host retention as high as possible to increase the quality of guests' stays
Overall rating influenced by a complex interplay of variables	Develop personalized strategies for hosts and guests to utilize unique value propositions and cater to specific preferences

Table 1: Key Results and Business Recommendations

Facilitating decision-making concerning guest satisfaction, the business recommendations are reflected in the dashboard design. The dashboard is divided into the Overview, Host, and Neighborhood sections, each with its visualisations and functionalities, tailored to the information to be conveyed.

1 Introduction

Airbnb is a platform built to link travellers with hosts all over the world. It is defined as a 'community built for belonging' (Airbnb, n.d. - a), with unique stays, experiences, adventures, and services users can trust. Various factors influence guest satisfaction with a listing on Airbnb, potentially the neighbourhood, price, type of accommodation, and amenities. To improve the understanding of guest satisfaction, we analysed a dataset containing a scrape of Airbnb listings in Copenhagen, Denmark. Based on our analysis, we provide possible future business suggestions and implementations to increase guest satisfaction. To describe our process, we provide details for every important step: from the initial data cleaning, the export of the data to SQL, and the exploratory data analysis, to the creation of a functional dashboard. The remainder of our report is structured as follows:

- The data used for the project is briefly described, followed by a thorough explanation of the analytical techniques used in the workflow of data cleaning as well as the exploratory analysis.
- Implementation of the operational database using PostgreSQL.
- Results of the explanatory analysis are shown and discussed, linked with the functionalities and design choices made for the dashboard.
- Business recommendations conclude the report intending to foster business growth and development.

2 Data Overview and Characteristics

The dataset used for this project contains Airbnb data of about 12,500 listings in Copenhagen, collected over two days from June 24th, 2022 to June 25th, 2022. Each data point consists of 74 attributes. The dataset can be roughly grouped into the following categories:

1. Scrape information, such as IDs and scrape date.

2. Host information, such as name, self-description, and response rate.

3. General listing information, such as name, price, and description of the listing.

4. Detailed listing information, such as number of beds and amenities.

5. Location information, such as longitude and latitude.
6. Review scores ranging from 1 to 5, such as the score for overall satisfaction, cleanliness, and score of value, as perceived by the guests.
7. Further review information, such as the total number of reviews, and the number of reviews in the last month.

To gather a detailed overview of the data, we have created a data dictionary. This includes all attributes, a description of them, data types, missing values, and the number of unique values, as well as further comments on the attributes. Summary statistics are also provided for the numeric columns. The data dictionary we created can be accessed via this link: Data Dictionary

3 Analytical Procedures and Tools

The conducted analysis consists of three pillars: data loading, data cleaning, and data exploration. The first one handles the import of necessary libraries, as well as loading the data into a *pandas* DataFrame. The data cleaning involves eight steps, including the export of raw data to SQL:

1. Overview of the data
2. Data dictionary
3. Type conversions
4. Replacing Invalid Values
5. Export to SQL
6. Dropping Duplicates
7. Missing values and Dropping Columns
8. Data transformation
9. Further feature development

Based on the clean data, the exploratory analysis is conducted. As introduced, the overall goal is to improve the understanding of guest satisfaction. However, the exploratory analysis aims in multiple directions, to ensure the identification of all important relationships and patterns. The exploratory analysis consists of three main steps:

1. Mean, Median, and Mode.
2. Linear correlation, including Pearson's Correlation Coefficient, Point Biserial Correlation Coefficient, as well as Spearman's Rank Correlation Coefficient.
3. Further exploration covering selected visual checks for non-linear correlation and further exploration of the price, location, neighbourhood, superhost, as well as property types.

Subsequently, the procedure is explained in detail, emphasising the reasons the particular methods were used.

3.1 Data Loading

As a first step in our process, we load the CSV file containing the Airbnb data into a DataFrame using the data manipulation library *pandas*. Additional libraries are imported to aid in our data analysis and visualisation:

- *NumPy* is essential for numerical operations,
- *Seaborn* and *matplotlib.pyplot* are used for data visualisation,
- *Collections.Counter* for counting occurrences of elements,
- *Scipy.stats.pointbiserialr* for correlation calculations,
- *Sqlalchemy* for creating and updating the PostgreSQL database,
- *Missingno* library is particularly useful for visualising missing data within the dataset.

This comprehensive setup ensures to effectively manage, analyse, and visualise the data.

3.2 Data Cleaning

Overview of the Data

After importing all necessary libraries and the dataset, we started cleaning the data. To better understand how it is structured, we looked at the shape, column names, head and tail, as well as the data types of the dataset.

Data Dictionary

Even though a data dictionary was provided, we decided to create our own one with more extensive information. This allows for a better overview of the dataset and acts as a central source of information.

Type Conversions

As a preprocessing step, several data types are converted. Binary attributes denoted with "t" and "f" are converted to boolean values, replacing the strings with *True* and *False.* The price attribute is cleaned by removing the currency symbol from it and subsequently converting it to float. Further, attributes such as the date of the first review are converted to *datetime* objects. Finally, attributes representing rates are converted to float. These type conversions make the attributes suitable for insightful analysis, facilitating the generation of valuable business recommendations.

Replacing Invalid Values

As a next step invalid values are replaced. Scores of zero are transformed to NaN, the names in the 'neighbourhoods_cleansed' column are changed to correct spelling mistakes, and the column 'id' is renamed to avoid any confusion between the host ID and listing ID.

Export to SQL

We incorporate this process into the Data Cleaning section of the report because it aligns with the stage in the data cleaning pipeline where data is inserted into the database, and we are using the results of the previous steps to insert cleaned data into our database. It should be noted that we are including missing values and empty columns in the database since the process of removing them in the database or in the dashboarding tool itself is straightforward and effortless; this also ensures that all of the data is incorporated into the database and future scrapes are fully captured. The ‘license’ column for instance is empty, but in future data scrapes, it could have values in it that have high potential for analysis.

Since the dashboard will be used by management to make decisions, we need to ensure that the data used in the dashboard is up to date. This is why our goal is to integrate the dashboard with a centralised data source, enabling continuous updates with current data. To allow this, we establish a PostgreSQL database, which contains all of the data provided by the dataset after the initial cleaning process, and subsequently connect it directly to Tableau. Ultimately we have created our own ETL (Extract, Transform, Load) Process; first, we extract the data, then we transform it by cleaning it, and then we load it into our operational database.

The main reason we chose to store the data in a Postgres database is speed and efficiency. Since we are opting to provide a dashboard that displays current data that is constantly updated, we are potentially dealing with a large dataset. SQL is specifically used to manage and query large datasets with speed and efficiency. In contrast, Python, while a versatile programming language, is not optimised for processing large datasets with the same level of speed and efficiency as SQL. Also, having a database makes it easier to work with the data in Tableau, as it is split into smaller tables and therefore gives a better overview.

Before creating our Postgres database, we first split up all the columns into different tables, according to the entities they represent and made an ER-Diagram to give an overview of our structure as seen in Figure 1. This design was chosen after some careful consideration since it separates the data into smaller, more specific tables relating to different topics, such as ‘listing’, ‘host’, and ‘review’.

The implementation of the database is fully done in Python. The table creation as well as the data insertion is implemented in Python using the *sqlalchemy* library. Having created an empty Postgres database in pgAdmin, the first step is to establish a connection to the database. To do this we use the *sqlalchemy* engine, which is responsible for connecting to the database. To do so, a connection path must be created, which utilises the user, password, host, port, and database parameters to establish a connection. Finally, we open a connection to the database using the *engine.connect()*

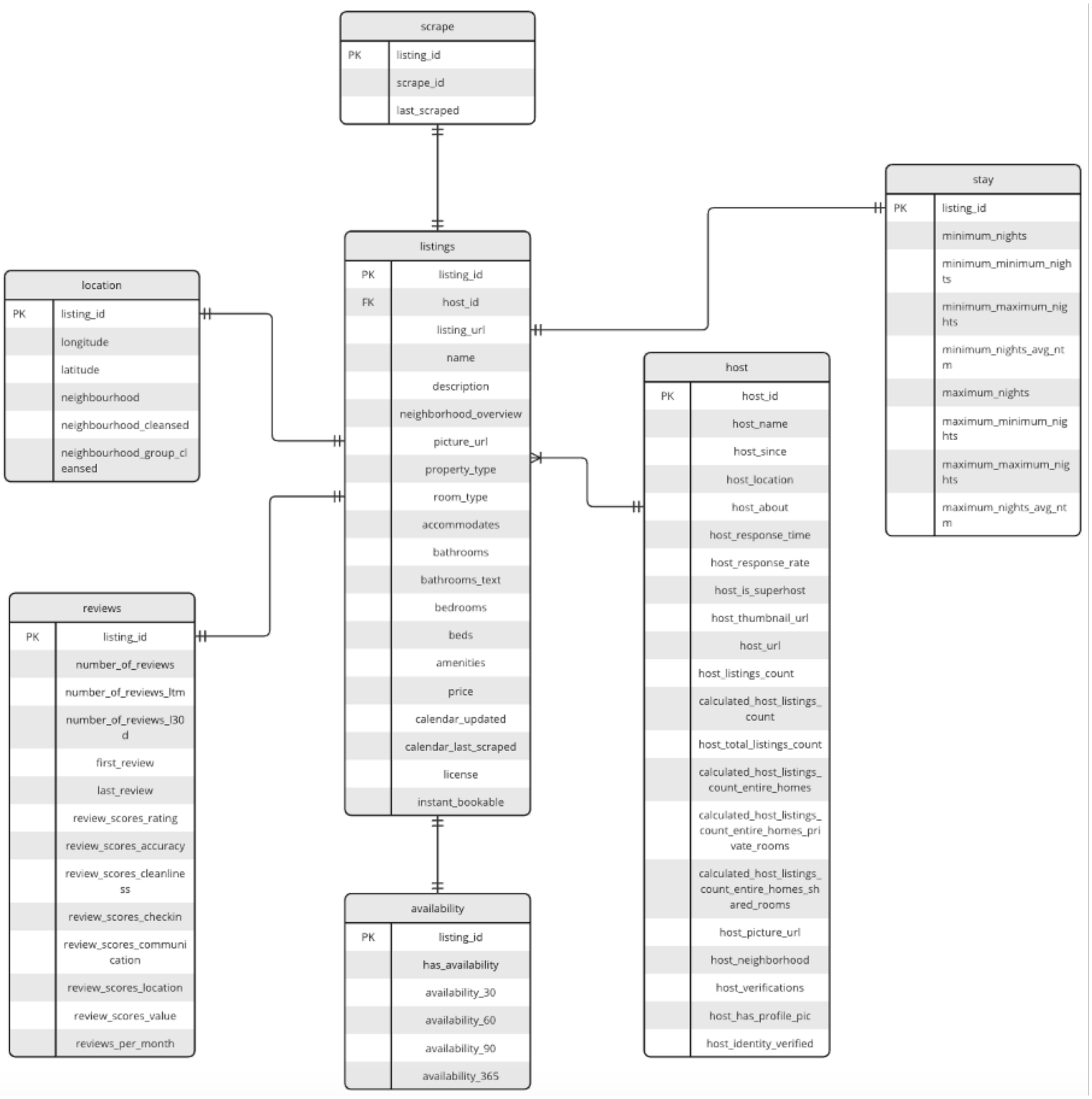


Figure 1: Entity Relationship Diagram of the operational database

command. In this part, informative messages have been implemented to display to the user whether the connection was successful or raise an exception and display error information if it has not been successful. Then, the Python code necessary to create the tables is provided. First, *sqlalchemy* Table objects are created for each table using the *Table()* command. For each *Table()* command, Column objects are created, specifying column names, data types, and foreign and primary keys. After the tables are specified, the *metadata.create_all(engine)* is executed; the *create_all* method uses the engine to establish a connection to the database and create the tables.

The last step of this process is to insert the data into the previously defined tables. This is done by specifying a list of columns to be included in the table, for each table. Then the *to_sql* command is used to insert the data into the specified table. If the table already exists, then the specified data shall be appended to the table and if it does not exist then an error will be raised. Finally, a method is provided to drop all tables from the database for testing purposes.

Dropping Duplicates

After exporting the data to a database using PostgreSQL, we returned to Python, where we proceeded to drop duplicates, which revealed that there were no duplicates in the data, meaning that there was no row that perfectly matched another across all fields.

Missing Values and Dropping Columns

Missing values are further analysed utilising the *missingno* library. First, an overview is provided, showing that missing values are not problematic for most attributes. However, the analysis also reveals that some attributes have substantial proportions of missing values. In particular, not accounting for the four empty columns, five attributes have more than 20 percent missing values, and nine attributes have more than 10 percent missing values. A closer look at the distribution through the distribution matrix shows more information. For the columns 'neighborhood_overview' and 'host_about', the null values seem randomly distributed and they have a high probability of not carrying bias. The pattern of the column 'host_neighborhood' reveals a high chance of introducing a bias, based on the cluster of missing values. A similar but less prominent pattern can be observed also for the 'neighbourhood' column.

With 'reviews_per_month' and all rating other attributes we can see a pattern that probably also signals an error in the data collection. Most values were correctly recorded. However, there is an empty cluster of rows in the middle section, indicating that something went wrong during the data collection.

The decision to retain as many data points as possible was made considering that, although not all rows are complete, the comprehensive nature of the dataset as a whole provides valuable information. Therefore, avoiding the removal of rows is crucial, as doing so would lead to a substantial loss of up to 77% of the database's content. Further, the decision was made against imputation of values, because imputing values would be grounded on too many assumptions. Finally, considering that the neighbourhood-related attributes convey similar information, but consist of different proportions of missing values, the decision was made to only continue with the complete attribute neighbourhood_cleansed and drop the columns we do not use; 'neighbourhood', 'host_neighbourhood' and 'neighborhood_overview'. Additionally, as shown by the missing values overview, there were four empty columns: 'neighbourhood_group_cleansed', 'bathrooms', 'calendar_updated', and 'license'. While we kept these columns in our database, for the Python exploration, we decided to drop them, since there was no way in which they could affect the results of the exploratory analysis.

In essence, the missing values do not dominate the dataset. For the purpose of an exploratory data analysis, we decided to barely modify the dataset further. Despite that, the risk of a bias in the data must be kept in mind, especially for potential future use cases, such as predictive models based on this data.

Data Transformation

Further engaging in the preprocessing the next step is to perform necessary data transformations. Amenities are defined as all the different services a host on Airbnb provides for its guests. Some essential amenities are often expected for a comfortable stay, such as pillows, and linens, and then there are amenities guests love to have such as Wi-Fi, a pool, a kitchen, and more (Airbnb, 2023). To facilitate the analysis of the provided amenities, as a first step, a list with all amenities is created. Subsequently, *Collections.Counter* is used for counting occurrences of amenities. Covering about 90 percent of amenities, the prominent ones were encoded with binary values. Thus, further analysis can be conducted on the impact of the presence or absence of certain amenities, leading to potential insights into guest satisfaction. Subsequently, we proceed with the host verification, creating new columns with binary results indicating whether the host used email, work email, phone, photographer, or none of these as verification methods. The number of bathrooms is determined based on ‘bathroom_text’ by creating new columns for the number of own bathrooms, and the number of shared bathrooms, respectively.

Further Attribute Development

To uncover potentially important relationships that might not be immediately evident in the existing dataset, additional attributes were created. For instance, the attributes ‘price_per_bed’ and ‘price_per_bedrooms’ were created. These newly created attributes are included in our successive exploratory analysis. This approach is in line with our objective to thoroughly explore the data.

3.3 Data Exploration

Mean, Median, Mode

As a first step to explore the data, the mean, median, and mode of selected attributes are calculated. These measurements are important to understand the nature of the dataset and its distribution. To ensure insightful results, only selected columns were taken into consideration, disregarding non-numerical attributes and constant attributes.

Linear Correlation

Analysing the correlation of the respective attributes of the dataset allows us to assess the relationships between them, improving the understanding of the data and ultimately the business. Discovering unexpected correlations can lead to new hypotheses, building the grounds for further analysis and experiments to discover causality. In addition, by conducting a correlation analysis we can potentially discover redundant features that convey similar information, leading to a better understanding of the dataset. The underlying dataset contains different types of variables, namely continuous variables, binary variables, as well as categorical variables. To ensure accurate results, every combination of types of variables requires different measurements. Khamis, H. (2008) suggests using the Pearson Correlation Coefficient for continuous-continuous pairs, the Point Biserial Correlation Coefficient for continuous-binary pairs, and Spearman's Rank Correlation Coefficient for continuous-ordinal pairs. For better visibility, a heat map is created for the continuous-continuous pairs. This allows for a rapid overview of the relationships. Due to the high dimensionality of the attributes, only for selected continuous-binary pairs the correlation coefficient is calculated. The decision was made to include all binary attributes, but only continuous attributes related to review scores and the price. Similarly, for the continuous-ordinal pairs the same continuous attributes are used, combined with attributes related to bathrooms. This selection was made in the prospect of the most promising insights.

Further Exploration

For a complete exploration of the data, different pairs of attributes are visualised, using different techniques including scatter plots, box plots, bar charts, empirical cumulative distribution functions, and tables summarising data. This enhances the overall understanding of the data. First, we complement the linear correlation analysis with a variety of scatter plots, each featuring selected attributes. This allows us to visually identify potentially relevant non-linear relationships in addition to linear relationships within the data. Second, we use a boxplot to understand the distribution of overall ratings. This visualisation provides insights into the distribution and central tendency of ratings. For a deeper exploration of price distribution, we plot the empirical cumulative distribution function, offering a comprehensive view of prices across the dataset. Moreover, to enhance our understanding of pricing, we map prices to their locations using a scatter plot. In this visualisation, high prices are represented by large red bubbles, while low prices are shown with small blue bubbles. We also dive into neighbourhood-specific analysis. One bar chart is used to display the mean price in each neighbourhood, while another bar chart is created with the count of listings, providing different perspectives on pricing as well as popularity. A bar plot is also used to compare the pricing strategies of superhosts versus non-superhosts, highlighting potential pricing differences based on host status. Finally, the different property types are counted and displayed with their respective share. In addition, two bar charts are used to visualise the mean price per property type as well as the overall rating for each property type. This approach sheds light on how different property types are valued and rated in the market.

4 Results and Discussion

Our analysis of the dataset led to several findings, reported and discussed subsequently. It is important to emphasise that only relevant results are highlighted in this report to convey only important information.

Overall guest satisfaction is very high, with the mode for ratings being 5 in every category. The mean rating exceeds 4.8, while the median rating of 4.8 further confirms this high level of satisfaction, while at the same time, highlighting how this average rating is substantially higher compared to a hotel firm's rating (Zervas et al., 2015) and therefore reviews and rating don't appear to have a great effect on the listing price (Ert et al., 2015). The boxplot of overall ratings highlights a few outliers, suggesting areas for potential improvement.

The mean price across listings is 1209 DKK, with a median of 960 DKK. However, the Empirical Cumulative Distribution Function (ECDF) suggests that the majority of prices were on the lower end. A few significantly highly-priced outliers skew the mean. Further educating hosts about Airbnb's dynamic pricing tools can reinforce a competitive environment in prices, and potentially facilitate hosts' efforts to balance profitability and utilisation of the properties. Unjustified outliers could be set to appropriate prices. It must be emphasised that the currency is given as dollars ($) in the CSV data source, but due to the high prices, we assumed that it was in DKK as it is the local currency. The minimum nights required for a booking show a mean of 4.60, a median of 3.00, and a mode of 2.0. This shows a potential gap in the market for one-night stays. More data from the demand side is necessary to further investigate this hypothesis. If it turns out that guests demand one-night stays, or that new guests could be generated with such offers, incentivizing hosts to also offer one-night stays is important. Potential gaps in the utilisation of properties could be closed. A potential target group for such short-term stays is guests on business trips. Increasing the offer of short-term stays could not only lead to additional sales but also the goal of increased guest satisfaction due to higher flexibility when choosing the desired length of stay.

Additionally, the mean number of reviews per listing is 18.07, with a median of 6.00 and a surprisingly low mode of 0.0. This underscores the need to encourage more guest reviews, but it must be kept in mind that recent listings are unlikely to immediately receive them. The number of reviews per listing does not show a notable correlation with the overall rating, however, increasing the number of reviews, especially for listings without any review can help to identify properties that are perceived as unsatisfactory. In addition, providing reviews for all listings is likely to improve the overall guest experience and ultimately satisfaction by building trust and transparency. Thus, there is a need for a refined strategy, incentivizing guests to leave reviews.

The correlation analysis leads to a few notable findings, but first, it's important to highlight that there's no precise way of interpreting the correlation coefficient. In social sciences studies the commonly used spectrum of correlation considers a correlation coefficient of 0.20 to be a very weak correlation, up until 0.35 a weak correlation, until 0.50 a fair correlation, until 0.70 a strong correlation, and over 0.70 a very strong correlation (Senthilnathan, 2019). First, the availability of the listing and its price show a very weak positive correlation (r = 0.16). Second, the detailed review ratings are, not surprisingly, moderate and strongly correlated with the overall review score. However, the overall review score shows the least correlation with location rating (r = 0.40), suggesting that location might not be as critical as other factors in determining overall satisfaction. Third, a weak negative correlation (r = -0.21) is observed between the availability and host response rate, indicating that quicker host responses might lead to better utilisation of properties. Fourth, a weak correlation (r = 0.26) is found between the length of host descriptions and property descriptions, suggesting that more detailed hosts are thorough in all aspects of listing information. Fifth, review scores for communication show a very weak negative correlation with the host's total listings count (r = -0.15), indicating that hosts with more listings might spend less time for detailed communication on each listing. As expected the review score value presented a very weak negative correlation with availability, potentially because as the scores are higher, the places will be fully booked and no longer available (r = -0.14). Finally, more very weak correlations were found but not considered to be meaningful.

Furthermore, causality must be ensured, for example by setting up experiments, to strengthen the understanding of the relationships. However, the awareness of these correlations builds a foundation for hypotheses for such experiments, potentially leading to a better understanding of guest preferences and ultimately guiding decision-making. Moreover, for some of the correlations, the significance remains unclear, which should be assessed in further research. Prices vary significantly by location, with Indre By having the highest prices and Bispebjerg, Vanløse, and Brønshøj-Husum being at the lower end. Therefore, a high valuation of proximity to the city centre is indicated.

Vesterbro and Nørrebro show the most listings, while Vanløse and Brønshøj-Husum show the least, pointing to a strong correlation between high prices and higher listing frequency (r = 0.67). Furthermore, the analysis of the correlation between the binary variable 'has_free_street_parking' and the continuous variable 'review_scores_location' indicates a negative correlation (r = -0.25). This suggests that locations without free street parking tend to have higher location review scores; another indication that guests value proximity to the city centre, which is reflected by the discrepancy in 'review_scores_location' and the presence or absence of free street parking. This is supported by the fact that free street parking spots in city centres are scarce (Copenhagen Citizen Service, n.d.). The type of accommodation influences pricing and satisfaction. Shared accommodations receive worse overall ratings compared to private ones while also being the cheapest, indicating a difference in guest satisfaction. Entire rental units account for about half of the offerings, and shared accommodations are significantly cheaper than private ones. Superhosts tend to have slightly lower prices.

Contrary to expectations, amenities did not emerge as a significant driver of overall ratings. Specific amenities like irons, TVs, dishwashers, high chairs, and fire extinguishers show very weak positive correlations with the price but do not show notable correlation with overall satisfaction. This raises questions about Airbnb's 2023 article - and promotion - of amenities guests want, which displays the most frequently searched amenities worldwide. It is important to refine the understanding of the effects of amenities because if amenities only drive up prices and do not contribute to guest satisfaction, it must be considered whether amenities lead to upselling or unnecessarily high prices with downside effects on demand. It would be beneficial to conduct additional research or data analysis to further understand guest preferences. This could include guest surveys or experimental listings with varied amenities to empirically test their impact on satisfaction and demand. It must also be emphasised that the analysed dataset only contains data about listings in Copenhagen, potentially leading to a bias in the analysis.

Regarding data quality, the correlation analysis shows clusters with high correlation, indicating multicollinearity within the dataset. Additionally, 41 attributes contain missing values, with 4 attributes only containing missing values. Further analysis can focus on isolating key variables impacting guest satisfaction. These insights can then be used for targeted improvements.

Finally, the overall rating appeared to be influenced by multiple factors rather than a few isolated ones, suggesting a complex interplay of variables affecting guest satisfaction. This complexity calls for a personalised strategy, tailored to each host and guest, utilising the value proposition of each host and accounting for specific guest preferences.

5 Guest Satisfaction Dashboard

Dashboard: Click here to visit the Dashboard

5.1 Target audience

The dashboard we have created is specifically tailored towards displaying guest satisfaction. Its primary users are Airbnb's management teams, including those in various departments. For instance, a guest satisfaction team within the marketing department might utilise the dashboard to monitor performance and gain insights into the factors that impact guest ratings. This information is also valuable to host community managers. These professionals engage with and support the host community by addressing their questions and promoting a supportive and informative environment on platforms such as Airbnb's forums. With a deeper understanding of what influences ratings, host community managers can improve their guidance on hosts toward achieving higher review scores. To illustrate, a campaign might aim to increase guest satisfaction metrics. In this scenario, the dashboard would enable management to comprehend the factors influencing review scores, directing the campaign toward the neighbourhoods most in need of assistance. The dashboard could also be a necessary tool for guest support team managers. By understanding the specific areas that have issues, these managers can plan accordingly, hiring the right amount of staff to ensure a good guest experience, even in cases where problems arise. While highly specialised, the dashboard offers various management teams diverse ways to generate value. Even though the data available for this project is limited to Copenhagen, this dashboard could be effective on a global scale. Given that many teams can benefit from the information it provides, making this tool available worldwide would be valuable for the company.

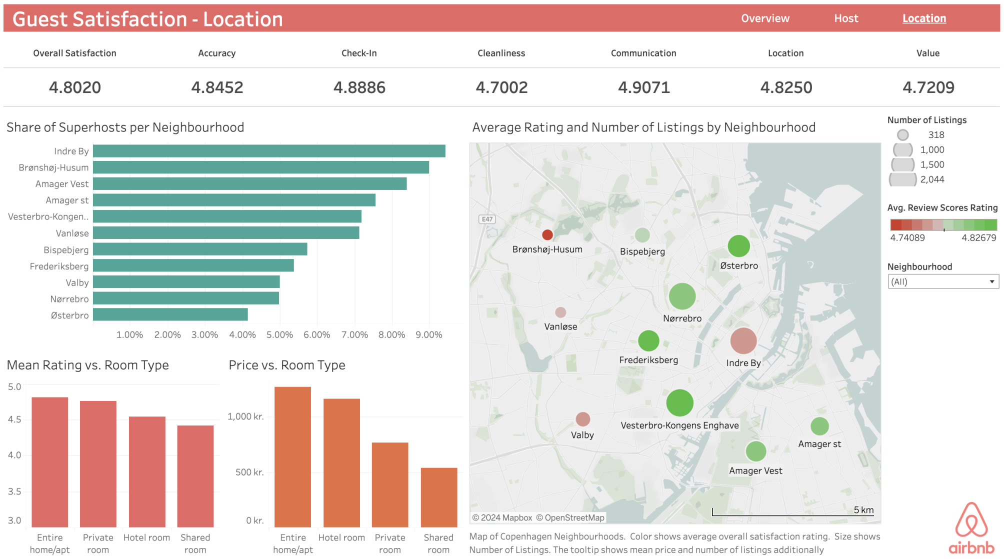
5.2 Design Choices

According to Bach et al. (2023), using colours consistently makes the dashboard easier to read and more familiar to the user, therefore we chose to persistently use Airbnb's formal company colours. Additionally, we made this decision to be consistent with other dashboards that the management might already be using or might use in the future. Further, numbers are displayed in neutral colours to avoid semantic colouring in meaningful values, so that the user's attention is focused on the number's meaning, without cognitive bias. Finally, as common practice dictates, we designed the KPIs font size to be substantially larger than its header text. Considering that the dashboard is designed for use by various Airbnb management teams, we have optimised it for display on standard desktop computers. The company's employees commonly use these devices for daily tasks. Additionally, the dashboard is optimized for display in PowerPoint presentations. This is due to its compatibility with the display size of generic desktop screens, which allows employees to effectively present and share data findings. We opted for Tableau containers over floating layouts for our dashboard objects, a decision driven by the need for clear visibility of graphs. Using containers ensures greater responsiveness, facilitating to maintain a consistent layout. This is particularly important as we use various filters and the sizes of objects may vary during user interaction with the dashboard. Tiled layouts, as opposed to floating, ensure that each object remains fully visible, a critical factor when dealing with changing dimensions and filters. This choice aligns with best practices, where tiled layouts are recommended for maintaining clear visibility and proportional adjustments of objects, especially in dynamic and interactive dashboard environments. As already discussed, the focus of this project is on guest satisfaction, therefore the dashboard aims to visualise relevant information, such as rating scores. The dashboard has been further broken down into three pages: Overview, Host, and Location. Each page contains information that is relevant to guest satisfaction. This approach ensures that visual clutter is avoided and users are not overwhelmed, while at the same time providing a cornucopia of information to the management.

5.3 Description of Pages

One key aspect of every page of the dashboard is the depicted review scores at the top of each page. The decision to include them on every page was made because the different review scores are arguably the most important measures of guest satisfaction. By providing the users with visual access to such KPIs on every page, we make sure that while the user is browsing pages with different themes, the important measures are easily visible and no back-and-forth between pages is required. This approach maximises readability, clarity, consistency, and user focus on the most important metrics. Such consistency in the presentation of key data points helps decision-makers assimilate the information (Bach et al., 2023), thereby reducing the cognitive effort required to navigate changing sets of data across different sections of the dashboard. Additionally, on the right side of the dashboard's pages page, we present explanations of the legends such as the circles for the number of listings and the colours for average pricing, as for example shown in Figure 2. Furthermore, when applying filters by clicking on certain graphs or by using the neighbourhood filter on the right side, the review scores at the top change depending on what is selected. For example, as displayed in Figure 3 below, clicking on a neighbourhood in the neighbourhood bar chart changes the values of the review scores.

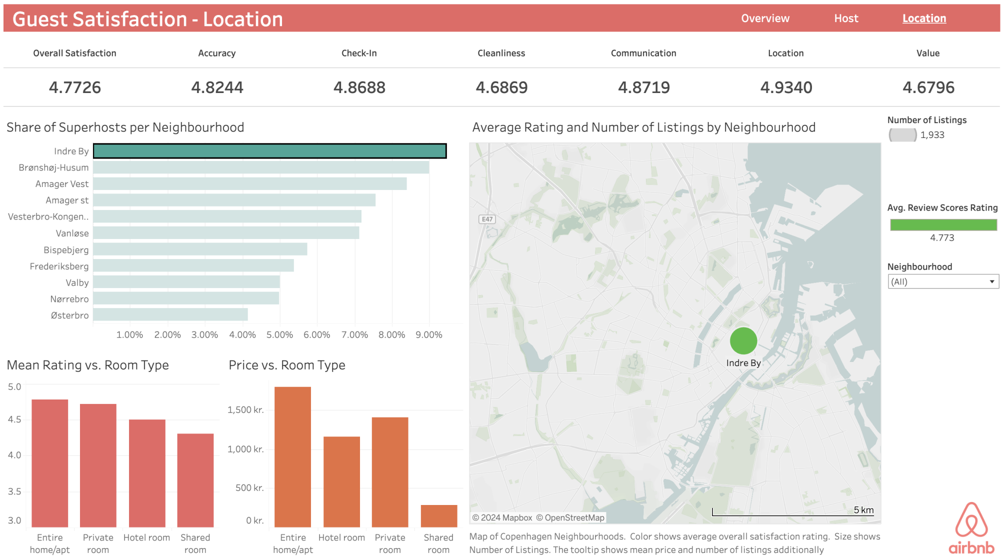
As seen in the figure, this is also the case for other visuals in the dashboard, making the dashboard interactive and easy to filter. One could also filter for other features such as "Private room" or the number of listings a host has by clicking on the respective graphs. Subsequently, the individual pages are explained in detail.



Overall Satisfaction	Accuracy	Check-In	Cleanliness	Communication	Location	Value
4.8020	4.8452	4.8886	4.7002	4.9071	4.8250	4.7209

Map of Copenhagen Neighbourhoods. Color shows average overall satisfaction rating. Size shows Number of Listings. The tooltip shows mean price and number of listings additionally

Figure 2: Location page of the dashboard



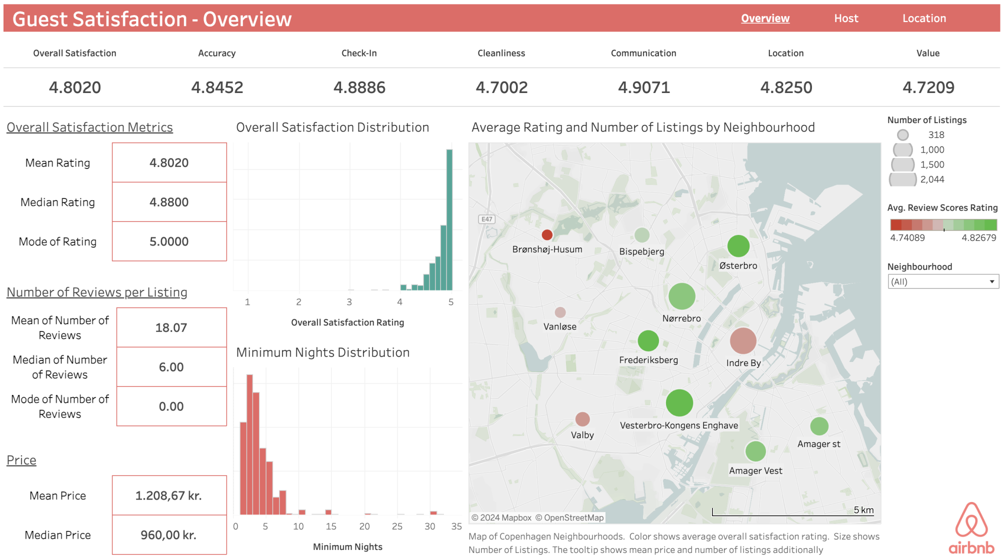
Overall Satisfaction	Accuracy	Check-In	Cleanliness	Communication	Location	Value
4.7726	4.8244	4.8688	4.6869	4.8719	4.9340	4.6796

Map of Copenhagen Neighbourhoods. Color shows average overall satisfaction rating. Size shows Number of Listings. The tooltip shows mean price and number of listings additionally

Figure 3: Location page with a filter applied on the Share of Superhosts per Neighborhood graph

5.4 Overview

The Overview page contains general guest satisfaction measures and is designed to provide a general overview of guest satisfaction to Airbnb's management. On the left-hand side, it contains the mean, median, and mode of overall satisfaction rating, number of reviews as well as price - the most important metrics to get an overview of the listings. As well as this, it contains the distribution of ratings for the audience. This way it can be understood that nearly all ratings lie between four and five. Below there is another, similar bar chart, showing the distribution of minimum nights a guest needs to stay at a listing. This shows that Airbnb does not particularly cater to guests looking for one-night stays, which we discussed in our findings. Finally, on the right side, there is a map to show the different neighbourhoods of Copenhagen. It displays average overall satisfaction and the number of listings per neighbourhood with colour and size of the bubble respectively. Also, when hovering over the bubbles, the user can see the mean price and number of listings in the tooltip. This design makes it possible to see at a glance the neighbourhoods with the best reviews and the ones with the most listings, as well as the price for each.



Guest Satisfaction - Overview

Overall Satisfaction	Accuracy	Check-In	Cleanliness	Communication	Location	Value
4.8020	4.8452	4.8886	4.7002	4.9071	4.8250	4.7209

Overall Satisfaction Metrics

Mean Rating	4.8020
Median Rating	4.8800
Mode of Rating	5.0000

Number of Reviews per Listing

Mean of Number of Reviews	18.07
Median of Number of Reviews	6.00
Mode of Number of Reviews	0.00

Price

Mean Price	1.208,67 kr.
Median Price	960,00 kr.

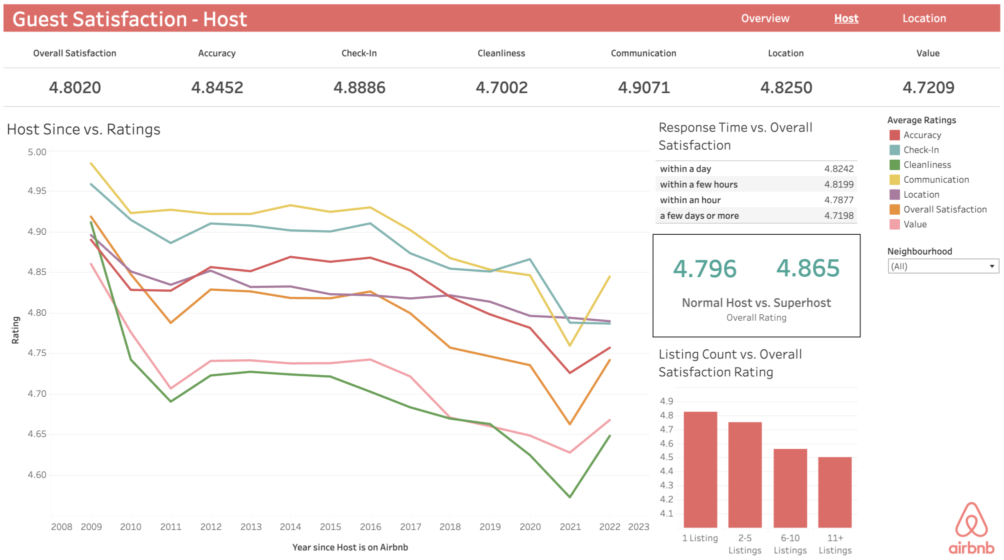
Neighbourhood: (All)

Map of Copenhagen Neighbourhoods. Color shows average overall satisfaction rating. Size shows Number of Listings. The tooltip shows mean price and number of listings additionally

Figure 4: Overview page of the dashboard

5.5 Host

The Host page displays relevant guest satisfaction measures relating to hosts. It shows the rating for different host attributes. Noticeably, it displays the overall satisfaction with the listing per year when the host created their profile. It shows that as a general rule, hosts that have been participating for longer have higher ratings. Additionally, we included a graph showing the average rating versus the number of listings a host manages. Here, as mentioned previously, fewer listings signify a better rating, indicating that a host should be cautious to maintain a high level of service when managing multiple listings. Another field on the left shows the difference in overall satisfaction between hosts and superhosts for all listings. Only a small difference is revealed. Thus, being a superhost does not necessarily cause higher ratings. Although there is not a strong correlation between review scores and being a superhost, becoming a superhost is particularly difficult for a host. To become a superhost, having very high ratings, a low response time, and cancellation rates are conditions (Airbnb, n.d. - b); these are factors that can directly influence the guest experience. Finally, we have included a table that shows the overall satisfaction rating depending on the host's response time, showing that being responsive is generally better for a host. Taking a few days or more to respond results in significantly worse ratings.



Overall Satisfaction	Accuracy	Check-In	Cleanliness	Communication	Location	Value
4.8020	4.8452	4.8886	4.7002	4.9071	4.8250	4.7209

Response Time vs. Overall Satisfaction

within a day	4.8242
within a few hours	4.8199
within an hour	4.7877
a few days or more	4.7198

4.796 4.865

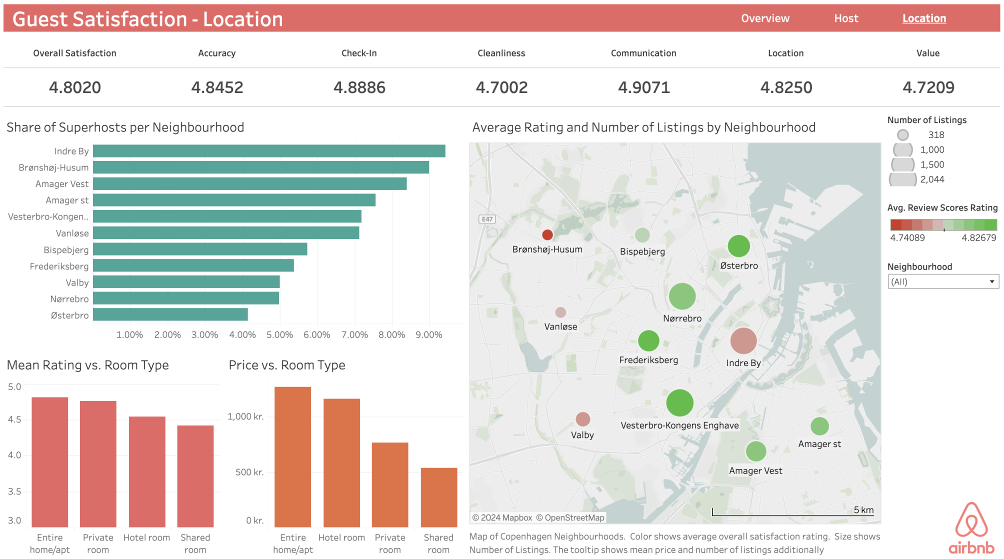
Normal Host vs. Superhost

Overall Rating

Figure 5: Host page of the dashboard

5.6 Location

The Location page provides insights into guest satisfaction in different Copenhagen neighbourhoods. It compares the measures relevant to review scores of different neighbourhoods. When the user opens the Location page the first thing that will be noticed are the KPIs. Further, a bar chart on the top left has been created to provide insights into the distribution of superhosts and every neighbourhood. Also, two bar charts have been created at the bottom left to showcase the difference in price and rating for shared and entire apartments. Finally, this page contains a multiple selection drop-down filter, where the user can choose one or multiple neighbourhoods, for which information will be displayed in the dashboard.



Guest Satisfaction - Location

Overall Satisfaction	Accuracy	Check-In	Cleanliness	Communication	Location	Value
4.8020	4.8452	4.8886	4.7002	4.9071	4.8250	4.7209

Neighbourhood: (All)

Map of Copenhagen Neighbourhoods. Color shows average overall satisfaction rating. Size shows Number of Listings. The tooltip shows mean price and number of listings additionally

Figure 6: Location page of the dashboard

6 Business Recommendations

Based on the conducted analysis and further discoveries revealed while building the dashboard, the following specific business recommendations are made:

1. Dynamic pricing tool education: There is an opportunity to further educate hosts about dynamic pricing tools to optimise pricing, balance profitability, and increase property

utilisation.

2. Market opportunity for one-night stays: The data reveals a potential market gap for one-night stays, for example for business travellers. Expanding offerings to include such short-term stays could lead to increased sales and guest satisfaction.

3. Importance of encouraging guest reviews: Encouraging more guest reviews is essential for building trust and enhancing guest satisfaction. Strategies to incentivize reviews, particularly for new listings, are important.

4. Shared Accommodations and Satisfaction: Shared accommodations, while being more affordable, receive lower satisfaction ratings, indicating a need for improvements and refined strategies in this segment.

5. Reconsidering the effect of amenities: Contrary to expectations and guest search data, amenities do not seem to be a significant driver of overall ratings. This challenges the current emphasis on amenities and points to a need for further research to understand guest preferences more accurately.

6. Data quality and analysis: The indication of multicollinearity and missing values in the dataset calls for a sophisticated approach to data analysis, focusing on isolating key variables that impact guest satisfaction.

7. Need for personalised strategies: The overall rating is influenced by a complex interplay of variables, suggesting the effectiveness of personalised strategies. Tailoring strategies to each host and guest can help utilise the unique value proposition of hosts, and account for guest preferences, which is an opportunity to increase satisfaction.

8. Retention of hosts: Hosts that have been on the platform for longer, tend to have higher ratings than ones that are new to the platform. Focusing on keeping long-time hosts on the platform can therefore prove valuable for maximising guest satisfaction.

9. Service Quality and Listings: Having fewer listings as a host appears to be linked to higher ratings for a listing. Therefore hosts should be urged by Airbnb's management

to be mindful to maintain a high level of service quality when managing more than a few listings so that guests remain satisfied.

7 References

About Airbnb: What it is and how it works - Airbnb Help Center. (n.d.). Airbnb. Retrieved 2 January 2024, from https://www.airbnb.com/help/article/2503

Airbnb Superhost details for guests. (n.d.). Airbnb. Retrieved 29 December 2023, from https://www.airbnb.com/d/superhost-guest

Bach, B., Freeman, E., Abdul-Rahman, A., Turkay, C., Khan, S., Fan, Y., & Chen, M. (2023). Dashboard Design Patterns. IEEE Transactions on Visualization and Computer Graphics, 29(1), 342–352. https://doi.org/10.1109/TVCG.2022.3209448

Ert, E., Fleischer, A., & Magen, N. (2015). Trust and Reputation in the Sharing Economy: The Role of Personal Photos on Airbnb. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2624181

Khamis, H. (2008). Measures of Association: How to Choose? Journal of Diagnostic Medical Sonography, 24(3), 155–162. https://doi.org/10.1177/8756479308317006

Public parking in Copenhagen — International.kk.dk. (n.d.). Retrieved 30 December 2023, from https://international.kk.dk/live/transport-and-parking/parking-in-copenhagen/public-parking-in-copenhagen

Senthilnathan, S. (2019). Usefulness of Correlation Analysis (SSRN Scholarly Paper 3416918). https://doi.org/10.2139/ssrn.3416918

The amenities guests want—Resource Center. (n.d.). Airbnb. Retrieved 27 December 2024, from https://www.airbnb.com/resources/hosting-homes/a/the-amenities-guests-want-25

Zervas, G., Proserpio, D., & Byers, J. W. (2017). The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry. Journal of Marketing Research, 54(5), 687–705. https://doi.org/10.1509/jmr.15.0204