Final Project Data Management and Visualization

Executive Summary: Data Visualization and Airbnb Market Insights by Jørgen Leiros

This project involved an initial analysis of Airbnb's business model. Based on the business' key objectives we performed a market analysis and identified potential areas of expansion as well as causes of low customer satisfaction. Throughout this project I performed database design according to a star schema as this was an analytical database and I was in charge of the creating the databases as well as the data pipeline to clean the data, transport it into the database and then further into Tableau for visualization.

Key Focus Areas:

1. **Airbnb Market Analysis and Strategic Priorities** –
    - Focused on **two key Airbnb strategic goals**:
        - **"Make hosting mainstream"** (increasing host adoption).
        - **"Perfect the core service"** (optimizing customer experience and pricing).
    - Target audience: **Regional managers in Denmark**, aiming to identify **growth opportunities and operational improvements**.
2. **Data Preprocessing and Cleaning for Analysis** –
    - Conducted **exploratory data analysis (EDA)** on a dataset of **12,495 Airbnb listings in Copenhagen**.
    - Addressed **missing values, structural inconsistencies, and outliers** to ensure data reliability.
    - **Dimensional modeling (star schema approach)** was applied for efficient database structuring and querying in **Tableau**.
3. **Airbnb Dashboard Development and Insights** –
    - Designed an **interactive Tableau dashboard** to visualize Airbnb's market trends.
    - **Key dashboard elements:**
        - **Growth of hosts over time** – Showing stagnation in new host adoption post-2017.
        - **Customer satisfaction ratings by room type** – Private rooms and entire homes performed well, while **hotel rooms and shared spaces rated lower**.
        - **Neighborhood analysis** – Revealed **uneven distribution of listings**, with opportunities in the **private room segment**.
        - **Pricing and availability insights** – Identified potential pricing inefficiencies, as **19.1% of listings remained available despite peak tourist season**.
4. **Strategic Findings and Business Recommendations:**
    - **Opportunity for growth in private room listings** – Encouraging **students to rent rooms short-term** could **increase supply and revenue**.
    - **Potential legal barriers** – Identified possible **regulatory concerns** that may prevent renters from listing properties.

    - **Need for improved pricing strategies** – Suggested enhancements to **Airbnb's pricing recommendation tools** to optimize occupancy rates.

Key Takeaways:

- **Data visualization enhances strategic decision-making**, allowing Airbnb to refine its **growth and customer experience strategies**.
- **Private room rentals offer an untapped opportunity**, particularly among students and short-term renters.
- **Pricing inconsistencies highlight a need for better AI-driven host guidance**, ensuring optimal revenue generation.

Through this research, Jørgen has demonstrated **expertise in data visualization, business analytics, and strategic decision-making**, providing actionable insights for **Airbnb's regional operations and market growth strategies**.

Executive Summary

In this final project, we have made a dashboard with the objective to get an overview on how Airbnb are performing on two of their strategic priorities, (1) Make hosting mainstream, and (2) Perfect the core service. In short, we want insight on adoption of the platform for homeowners and how good the Airbnb platform really is, measured in customer satisfaction and operational efficiency with an emphasis on pricing.

What is interesting diving into this area is that we can get a look into main value drivers for Airbnb as a company. With a relatively low variable cost per booking, higher volume (occupancy and number of hosts) and higher prices are what Airbnb strives for. We are picturing a regional manager with local knowledge of Denmark and Copenhagen as a rental market as the target audience for our dashboard and analysis. With this dashboard we aim to highlight areas for potential growth in income from the market in Copenhagen, providing actionable insight to local executives.

Through data analysis and visualization, one of our most important findings was the potential for growth in the private room market. These listings make up a small portion of the total, while still having relatively satisfied customers on the listings that are present. Seeing that the number of hosts is reaching what looks like an equilibrium state, focusing on potential hosts for private rooms could be a way to facilitate new growth. Combined with local knowledge one can come to the conclusion

that there is great potential in making it “mainstream” for students to rent out their rooms in their shared flats while for instance being on vacation in their hometown. We hypothesize that a potential bottleneck in this market could be legal issues regarding renting out properties that the residents themselves rent. Our recommendation is therefore to deploy a team that can investigate how these insecurities and potential conflicts can be resolved between homeowners and renters. While our dashboard offers insightful data on Airbnb's Copenhagen market and growth opportunities, it's important to acknowledge the limitations of our dataset. We would advise Airbnb's management to obtain a more extensive data set for a deeper and more accurate analysis of market trends and performance.

Introduction

Among Airbnb's six core values, two are particularly focused on enhancing customer experiences: "Champion the mission" and "Be a host” (Airbnb, 2023). The former supports Airbnb's mission of creating a world where anyone can feel a sense of belonging, achieved by fostering a community and network on the platform that builds trust and drives company growth. The latter, "Be a host," emphasizes the importance of exceptional hospitality, encouraging hosts to provide unforgettable experiences to their guests. These goals are further elaborated on in the company’s quarterly reports, highlighting three main strategic priorities:

1) Make hosting mainstream.
2) Perfect the core service.
3) Expand beyond the core.

Airbnb is focused on increasing the number of active listings, a strategy that aligns well with their primary revenue drivers: booking volume and pricing. Enhancing the core service is another key area, with a particular emphasis on refining the platform. This includes the development of advanced tools to assist hosts in setting competitive prices. Additionally, the company is turning its attention to markets that have not yet fully embraced its services, aiming to expand its reach in these less saturated areas.

Given the data at our disposal, it seems most pertinent to concentrate on the first two strategic priorities while examining the dataset. The objective for the dashboard we're developing is to offer

insights into how well these strategic priorities are being met, potentially revealing opportunities for further optimization and growth.

The report is written with an objective to please both managers wanting to dive deeper into our reasoning, as well as other technical experts wanting to validate our process. The final dashboard is purely intended for managers/executives.

Data preprocessing

The value of insights and analysis hinges significantly on the quality of the data used. Managers recognize that poor-quality data leads to a range of problems, including inefficient time management, increased costs, and compromised decision-making, which collectively hinder the formulation of effective business strategies (Redman. T, 2017). Analytics relies heavily on accurate data to identify patterns and trends. When data is fraught with errors, inconsistencies, or is outdated, it can yield misleading conclusions and render reporting ineffective. Our dashboard, designed to highlight key metrics, depends on the integrity of its input data. Inaccuracies, such as duplication, missing values, or structural errors, can undermine its reliability and lead to misrepresented facts, thus affecting the overall integrity of the insights provided.

Cleaning Process Guideline

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Our approach includes key steps such as exploratory data analysis (EDA), which involves looking at the data before making assumptions. It helps pinpointing clear mistakes and gaining a deeper understanding of the patterns present in the data. (IBM, n.d.) This is followed by fixing structural errors and introducing new variables based on the insights gained from our EDA. Next, we focus on filtering out unwanted outliers to enhance the reliability of the data we are working with. Handling missing data is another crucial step where we apply different imputation strategies based on the context of the missing information. The final step in the data cleaning process is the validation of data and quality assurance, ensuring the accuracy of our dataset. During the process, we primarily utilized Python for the data cleaning tasks. However, Tableau was also helpful in the data exploration process and was used to create some calculated fields. This applies to the review variables which will be discussed later.

Exploratory Data Analysis

We started by loading the complete dataset ‘AirbnbListings.csv‘ which consisted of 74 columns and 12495 rows. The dataset contained detailed information on listings for properties available for rent on Airbnb in Copenhagen, including location, price, room type, number of reviews, and host details. It provided insights into rental pricing, popularity, and availability patterns across different neighborhoods, helping us to understand the market.

Utilizing the info(), isnull(), and describe() functions revealed key information of the dataset, such as the data types which are integer, floats and objects. The presence of null values was notable, particularly in columns like *neighbourhood_group_cleansed*, *bathrooms*, *calendar_updated*, and *license*, which were entirely empty. More interesting fields such as *host_response_time* and review-related scores also exhibited numerous missing entries. Analysis of descriptive statistics highlighted a generally high average for review scores. The *calculated_host_listings_count* showed considerable variation, with an average of 7.9 and a peak at 346. The *price* column displayed substantial fluctuation, and with a standard deviation of approximately 1,870, it indicated that the prices of listings were not consistent and varied greatly from one listing to another, with some prices being extremely high.

From briefly investigating the dataset, we identified several variables that could be interesting for us in constructing an insightful dashboard. Location-related variables which could provide insight into the geographical distribution of listings and popular areas; room type which could give us an understanding of the types of accommodation (e.g., entire home, private room); host information such as *host_since*, *host_response_time* and *host_is_superhost* could give an idea about the experience and responsiveness of hosts; availability and booking variables could offer insights into how often properties are booked and the booking policies; reviews and ratings were important for understanding guest satisfaction and the popularity of a listing; amenities could give insights in the importance of features and facilities in a listing; and price-related variables could give us commercial understanding of the listings.

By analyzing and investigating the distributions of a few of these variables, we got an overview of their main characteristics, which helped us discover patterns, spot anomalies and check assumptions.

We decided to look at the distribution of prices, number of reviews, review score rating and the different room types. This is illustrated below:

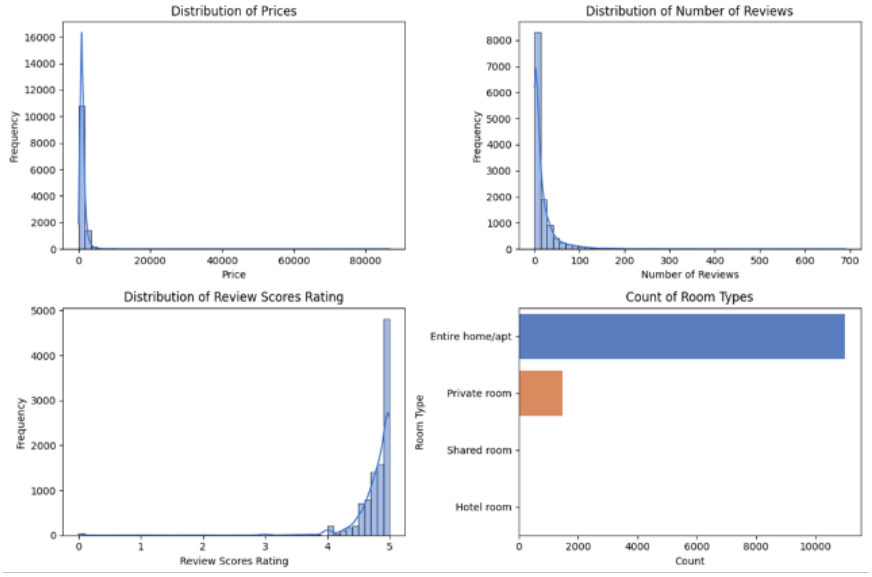

Figure 1: Distributions of Prices, Number of Reviews, Review Scores Rating and Room Types.

From the *Distribution of Prices* histogram, we observed that most listings are priced at the lower end of the spectrum, suggesting an affordability trend in offerings, however it also confirms that a few listings are of significant high pricing, which is something we chose to investigate further. The *Distribution of Number of Reviews* histogram revealed a skewed pattern where a substantial portion of listings have only a handful of reviews, and listings with a higher count of reviews became progressively rarer. Turning to the *Distribution of Review Score Rating*, the data was predominantly weighted towards the upper range, which implied that the listings typically received favorable feedback from guests. Lastly, examining the *Distribution of Room Types* reveals that entire homes/apartments are by far the most common room types on the market, with private rooms being the second most common type. This trend highlighted consumer preferences and market offerings on the platform, which could be interesting to investigate further.

We decided to get a brief insight of the variables of Copenhagen's neighborhoods, with a focus on the average pricing of listings and the total number of listings in the area. Neighborhoods with higher average prices might be more desirable or could have properties with better amenities or more space. Neighborhoods with more listings could suggest they are popular destinations or have a higher

density of available properties. Moreover, this data could shed light on market demand and supply dynamics, investment potential, and strategic hosting opportunities within the city. The visualization below illustrates these points:

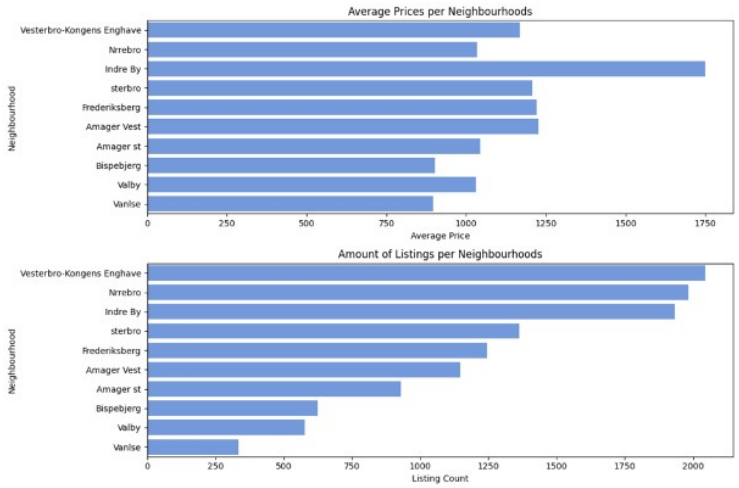

Figure 2: Bar charts of Average Prices per Neighborhood and Amount of Listings per Neighborhood.

Lastly, a few feature variables were integrated into the dataset to enhance our analysis: *host_since* calculates the total days a host has been active on Airbnb; *days_since_last_review* measures the time elapsed since the most recent review. Exploring these variables made us better prepared for the calculated fields used in the dashboard. Dummy variables for the amenities located on Airbnb's filterpage were also added to the dataset, in case we were to visualize this later in the dashboard.

Dictionary of Data

The following is a list of variables that we found to be most important in the initial faze, and therefore included going into further analysis:

Variable	**Type**	**Measurement unit**	**Description**	**% Missing Data**
id	Integer	Identifier	Unique identifier for the listing.	
host_id	Integer	Identifier	Unique identifier for the host.	

host_since	Date	Date (YYYY-MM-DD)	The day the host/user signed up.	
host_response_time	String	Categorical	The duration it takes for a host to answer users.	22%
host_response_rate	Numeric	%	The percentage of inquiries the	22%
			host responds to.	
host_acceptance_rate	Numeric	%	The percentage of booking requests the host accepts.	12%
neighbourhood_cleansed	String	Categorical	The neighborhood the host is located in.	
room_type	String	Categorical	The type of room offered (e.g., Entire home/apt, Private room).	
accommodates	Integer	People	The number of people the listing accommodates.	
amenities	List	Categorical	Detailed list of features and services provided at the listing, such as Wi-Fi, towels, etc.	
price	Numeric	Currency (DKK)	Listing price per night.	
availability	Integer	Days	The number of days a listing is available for booking. This is recorded for various time frames: within the next 30, 60, 90, and 365 days.	
last_review	Date	Date (YYYY-MM-DD)	The date of the last review the listing received.	18%
days_since_last_review	Integer	Days	The number of days since the last review.	18%
review_scores (all 7)	Numeric	Rating Scale	Aggregate of all review scores (rating, accuracy, cleanliness, checkin, communication, location, value).	18%

Converting Data and Fixing Structural Errors

Following the EDA, we specifically targeted the columns we found interesting, and that could be utilized in our dashboard later. Two of the variables had to be converted, such as the objects *host_since* and *last_review* which had to be in datetime format to be recognized by pandas. Some of the variables had 't/f' values, which were changed to numeric '1/0' to make it more convenient in statistical analysis which often require numerical input. This was also done when we introduced a new categorical variable that made the values in *host_response_time* numerical. Dollar signs and commas in the price column were removed, converting the values into floats, and the rate variables were transformed from percentage strings into float values.

Filtering of Unwanted Outliers

The EPA revealed certain observations that were inconsistent with the general trends of the data. In our analysis, the 'price' variable is a key metric. We wanted to make sure that the very high, and possibly unrealistic, prices in a few listings did not skew our overall understanding of the Airbnb market. Considering the visualizations provided below and our knowledge of the pricing of listings in Copenhagen, we decided to exclude listings with nightly rates exceeding 15,000 DKK to maintain a realistic representation of the market.

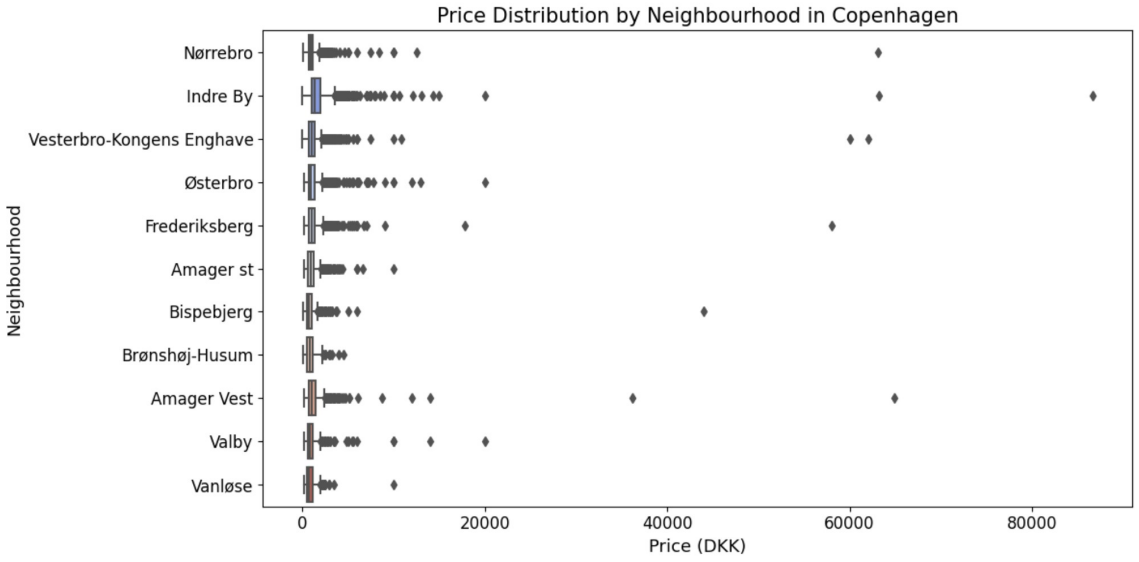

Figure 3: Price Distribution of the different neighborhoods in Copenhagen

Handling Missing Data and Its Challenges

The pandas library built on top of the Python language provides great methods for handling missing values. Detecting missing values in a dataset is a crucial step in data preprocessing, and functions like isna(), notna(), and isnull() in pandas provide a straightforward way to identify them. These functions can be applied to each column to determine the presence of missing data. More importantly, calculating the percentage of missing values in each column helps to assess the extent of the issue. This information is vital because it influences the strategy for handling missing values, such as deciding whether to fill them in, drop them, or use more sophisticated imputation techniques. Most of the variables had missing values less than 1-2%. However, 22% of the values in *host_response_time* and *host_response_rate* were missing, and approximately 18% of the values in all review variables were empty.

Addressing the *host_response_time* posed a challenge. As a categorical variable, the typical method would be to substitute missing data with the mode (most common value), but this risked skewing the data, potentially assigning one value to a whole cluster of data points. The variable's distribution and further examination didn't offer clear guidance. Consequently, we opted to create a distinct category, ‘unknown’, for the null values. On the other hand, handling *host_response_rate* was more straightforward. We used the median to fill in missing values. This method was used based on the distribution of the variables and selected because it is less sensitive to outliers than the mean. Handling the review columns was more difficult. Within the dataset, a total of 2,304 listings were identified where entries across all review score columns were null. The distribution of these null values appears to be random, with no immediate pattern as to their occurrence. This poses a challenge for data imputation. Our decision was to handle the missing values of these columns in Tableau, where it is easy to filter out the null values if needed. As for the rest of the variables, the few null values were replaced with the use of mean or replaced with 0, all based on the specific nature of the data in the column.

Data modeling

The initial exploratory analysis and data cleaning gave us valuable insight which was important for the data modeling task that followed. After the data cleaning and transformation was done, we had the choice of breaking the data down into smaller tables in a database structure or loading the csv

file directly into tableau for visual presentation. There would be no significant technical difficulties with loading the flat csv file directly into tableau, but as a good database structure can provide benefits in many ways, it is an important consideration to make.

When deciding on a database management approach, it is important to consider the nature of the data we are storing and the purpose of the database. Data processing is done to support both operational and analytical efforts in an organization. The operational processes often entail focusing on single records and carrying out repeated transactional tasks, adding loads of information to the database, one record at a time. To support such operations, the database needs to be highly normalized, often leading to high complexity. Normalization is the process of organizing data in a database in order to remove redundancy and improve database integrity (Microsoft, 2023). This integrity comes at a cost, making querying a challenging task due to increased complexity of the database structure. Analytical systems on the other hand, have significantly differing needs from the operational systems (Kimball & Ross 2002, p. 2 - 10). Dimensional modeling focuses on simplicity and is the preferred approach for structuring a database to support analytical efforts.

As our dataset does not contain transactional data, and the purpose of our database will be to carry out analysis and visualize findings, the database structure should follow a dimensional modeling approach. Dimensional modeling entails identifying a subject of analysis and separating the data into tables where the subject is stored in a central fact table with surrounding dimensions tables. The fact refers to the measurement data that we strive to keep in the central table as these make up the majority of the data and the most frequently changing data, and you want to avoid duplicating it (Kimball & Ross, 2002, p. 19). The idea of dimension tables is that they provide the textual descriptions of the data, and as such, they serve as useful sources of query constraints, groupings and report labels.

In dimensional modeling the database schema can take different shapes, and we can differentiate between star schema, snowflake schema and fact constellation schema. Both star schema and snowflake schema have dimension tables surrounding one fact table, snowflake schema being an extension of the star schema as it can have additional dimension tables connected to other dimension tables in a many to one relationship (IBM, 2021). A fact constellation schema differs from the former

two schemas in that it can have more than one fact table (Jukic, 2019, p. 260-266). Benefits of structuring data in tables according to a database schema is that it makes querying simpler as you need fewer joins and provides valuable constraints (Kimberly & Ross, 2002, p. 22). Furthermore, dimensional schemas are more intuitive and provide simple relations that makes it more accessible to non-technical report builders. Even though some of the benefits of dimensional modeling will not apply to our use case, we have decided to structure our data further according to a dimensional modeling approach instead of loading the flat csv file directly into tableau. This is to remove redundancy and increase usability for all report builders with varying insight into the data. Additionally, even though our goal is not to build an operational database, some normalization has been performed to remove redundancy and simplify our analysis. This includes removing repeated attributes, removing attributes with low cardinality, splitting attributes with multiple values into separate attributes in its own table, and ensuring that an attribute is fully defined by its primary key and no non-key attributes.

In the case of our dataset, we have found that the most prominent subject for our analysis will be 'Listing'. Another possible subject could have been 'Host', but we have found that 'Host' is better suited as a dimension table for 'Listing'. For this reason we are only going to have one fact table, and as we see no need for further normalization, we have chosen to structure our database according to a star schema. One relationship we did identify is that one of the review attributes are more so related to a host than listing, namely the *review_scores_communication* attribute. Through our initial analysis of the data however, we identified different communication ratings for the same host indicating that these ratings are not calculated across listings. Additionally, if we go by Kimball and Ross' criteria for separating attributes into dimension or fact tables, there are very few attributes in the dataset that can be naturally separated into descriptive categories whilst there are a lot of variable measurement attributes that could fit into the 'Listings' fact table (Kimberly & Ross, 2002, p. 20). On the other hand, as all of the data is collected from one scrape, we can assume that the data will not change at all for the use case of our database. As all values in our database are constant, any descriptive attributes that naturally belong to one of the dimension tables, can be placed there.

A detailed examination of the dataset reveals minimal redundancy and shared attributes across potential dimensions. A notable exception is the host dimension, where attributes such as

host_response_rate, *host_response_time*, and *host_is_superhost* are replicated across a host's portfolio of listings. The separation of host data into its own dimension does not significantly reduce the dataset size, lowering it from 12,436 to 11,157 rows. Despite this modest reduction, separation is deemed a good practice. Attributes related to location are also replicated across several entities. Other natural dimensions we have identified are 'Availability', 'Reviews' and 'Amenities'. By separating these categories into dimension tables we get an intuitive folder structure when transferring the tables into Tableau. Furthermore, amenities were stored as a list in a single attribute in the initial flat csv file, and by splitting different amenities into their own attribute we have the ability to better analyze impacts of different amenities while also adhering to the first normal form by removing composite values. Still, these operations remove minimal redundancy as most of the dimension tables are just as long as the fact table, while doing little to improve change anomalies compared to 3NF. However as the dataset is relatively small and we don't expect changes to the dataset during our analysis, these concerns will not impact our data or analysis significantly.

If the use case for the database was different, and we were expecting information to be added and changed during our analysis, possible changes to the structure could be to include more of the frequently changing measurement attributes in the fact table, and follow normalization rules more strictly. One issue with normalization is the availability of listings in the future. For every booking within the next 30 days, all of the other availability attributes will also need to be updated. Other than this, it will be difficult to normalize the attributes we have decided to include more than what has already been done.

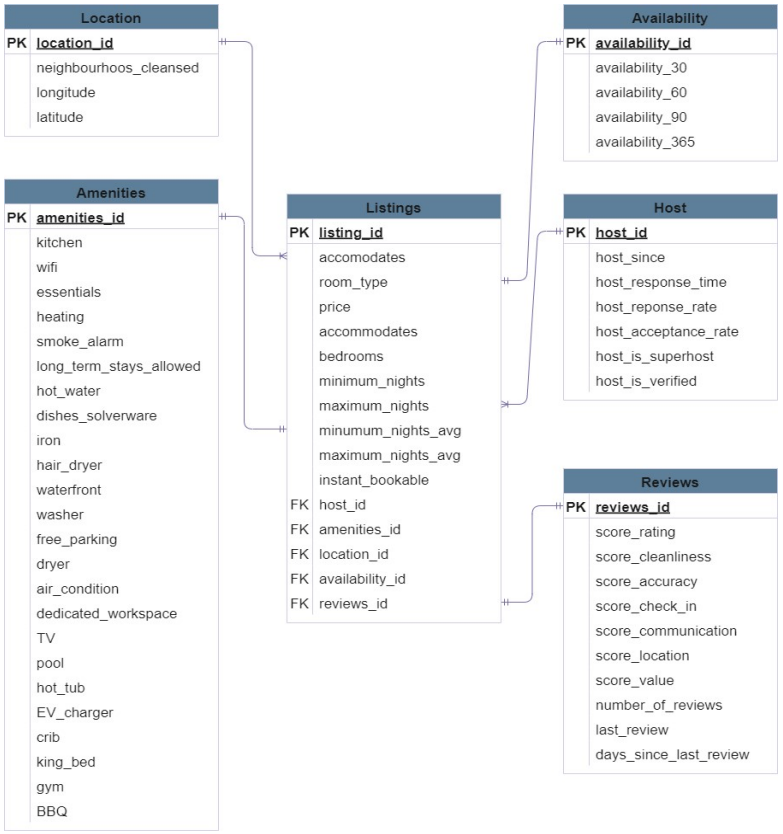

Figure 4: Star schema ER-diagram

Above is the final design of our analytical database with 'Listings' as the central fact table. What is notable is that three of the dimension tables have one to one relations to the fact table. The reason is that these tables have exactly the same amount of rows as the fact table. For instance 'Listings' has only one combination of 'Reviews', and a combination of 'Reviews' can only belong to one 'Listing'**.** More classical one to many relationships are spotted in 'Host' and 'Location' with 'Listings'**.** These relationships occur because a host can have several listings, as well as several listings are recorded with the same location.

Dashboard

Target Audience and Goals of the Dashboard

The primary aim, as outlined in the introduction, is to provide Airbnb with analytical insights pertaining to their 2023 strategic goals. Our target audience is a regional manager responsible for Airbnb's performance in Denmark, particularly in Copenhagen. This manager is expected to have a comprehensive understanding of the city and its neighborhoods. The dashboard is designed to enable

the manager to guide efforts towards attracting and helping hosts, as well as providing feedback to product teams working on the platform.

To achieve this, our focus is concentrated on a handful of metrics: number of listings, hosts, pricing, and ratings. We find these important as they are defining for Airbnb's success. They want right pricing, securing high occupancy while being sure that the highest price possible is taken. The ratings again are securing future revenues, proving the quality of their product. The goal is to contextualize these metrics in various ways, allowing for comparative analysis across Copenhagen's eleven distinct neighborhoods.

Dashboard Design

Dashboards provide a way to communicate complex and vast data sets in a time-efficient and comprehensible way to non expert audiences (Bach et al., 2022). The goal of the dashboard is to provide valuable insight while the activity of exploring the dashboard remains relatively effortless. One of the reasons for dashboards' increasing popularity is that it capitalizes on the human perceptual capabilities (Yigitbasioglu & Velcu, 2012). With this dashboard we try to leverage cognitive science further to impress upon our audience the most important findings of our analysis. Visualization techniques have been identified and researched, and in our dashboard we utilize recommended techniques when it comes to dashboard layout, color choices and interactivity.

Designing a dashboard can be a discipline of balance between too little and too much. The dashboard is successful if it manages to convey large amounts of data efficiently, but too much information in a visualization can take attention away from the more important findings. Edward Tufte has suggested maximizing the data-ink ratio as a means of ensuring that the attention is aimed at the data rather than the visualization (Tufte, 2006, p. 91). Following this suggestion, we have removed any redundant information from our visualizations, rather supplying additional information when the audience actively seeks it. According to Stephen Few, such fragmentation should not undermine the audience's ability to gain desired insight at a glance, so we have been careful to only include additional information that would have otherwise been distracting if included in the dashboard (Few, 2006, p. 39).

It is common practice to use colors to get the attention of the audience and distinguish entities from one another, but too much color can overwhelm the audience. We have used colors to relate the dashboard to the company that the data pertains to, make the dashboard more inviting, and highlight differences in our data. We imported the color scheme that Airbnb utilizes to ensure a consistent theme. This ensures that we use the same colors throughout the dashboard, without too much variety. Additionally, we have used Adobe's contrast checker to ensure that the contrast at least adheres to an AA accessibility standard.

Dashboards leverage human perceptual capabilities, and for our dashboard to be successful, we have utilized a range of techniques that leverage cognitive science. Patterson et al. have suggested a human cognition approach to data visualization that has influenced our design significantly (Patterson et al., 2014). Considering the information that we are aiming at communicating, and the target audience for this dashboard, we have not considered cues for triggering exogenous attention important for our visualizations. As we are presenting an overview of the Airbnb market, it is more important for us to prevent one visualization from taking too much attention away from others. When it comes to endogenous attention, we have attempted to assist the audience in keeping information in their working memory by using clear labels and providing the audience with filters and tooltip functionality. The endogenous attention is further assisted by the principle of maximizing the data-ink ratio. The third leverage point suggested by Patterson et al. is to facilitate chunking of information by grouping them together to communicate some form of self-similarity. We have utilized a treemap to provide retrieval cues in an effort to provoke activation of the target audience's knowledge structures. Additionally the dashboard is tailored to our target audience by following the conventional reading direction in Europe, placing the most significant data for our analysis in easily accessible numeric presentations in the top left corner. This approach has also been used in arranging the different visualizations in our dashboard.

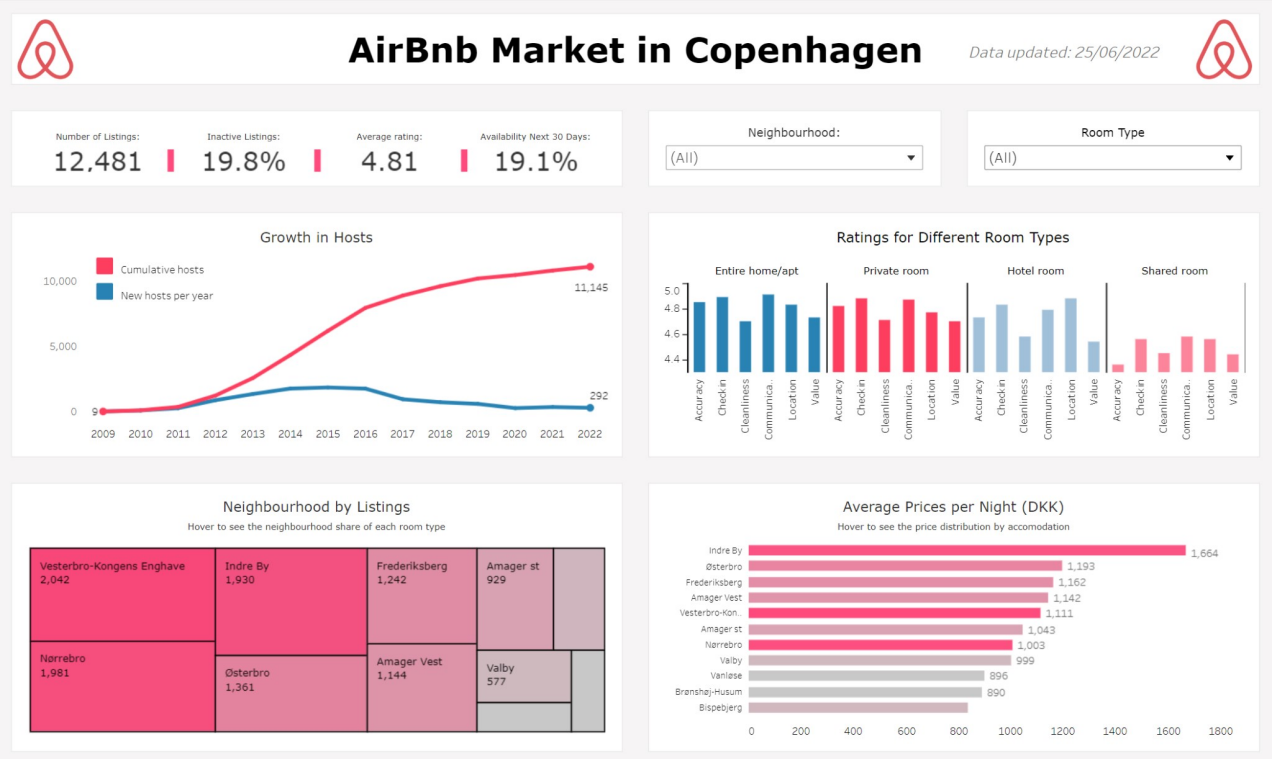

Figure 5: Final Dashboard Design

Dashboard Elements

Key Metrics

The key metrics are used in the top left of our dashboard to display important numbers for the Airbnb market in Copenhagen. The first key metric represents the number of Airbnb listings in Copenhagen, and is created simply by placing the count of *listing_id* in the worksheet 'text' mark and formatting it to the appropriate size. Additionally, the number of listings is subject to both the 'Neighborhood' and 'Room Type' filter, thus changing when either or both the filters are used. The next key metric in our dashboard, 'Average rating', displays the average rating of the Airbnb listings in Copenhagen, and is created by placing the measure *review_score_rating* in the text mark, similarly to the key metric 'Number of listings'. Without a proper explanation in the provided data dictionary, we assumed that 'review score rating' was a calculated average based on the other reviews to show a total average rating of a listing. Upon further investigation, however, we discovered that the *review_score_rating* does not take the other reviews into account, but is rather a separate review that the guests leave. Since this is the method used by Airbnb to showcase the rating of a listing, we also decided to use the metric *review_score_rating* for our own visualization.

The third key metric shows the percentage of inactive listings, and is based on listings without a review in the last 12 months. This is an important number as it provides a better overview on the amount of actual active listings in the market. This can be of importance for the regional manager for Airbnb in Copenhagen because it could be an indicator of the churn rate among its hosts. The key metric is produced by making use of two calculated fields. The first one is called 'Inactive Listing' and is calculated with an 'IF' statement that turns all listings with more than 365 'Days since last review' into '1', and the rest to '0', making it easy to count the total amount of listings with more than 365 days since their last review. The second calculated field is called 'Inactive listings %', and as the name indicates, it calculates the percentage of listings that are inactive. It is done by dividing the sum of 'Inactive Listing' (the first calculated field) with the count of listings (total amount) and multiplying it by 100. This key metric is, similarly to the others, subject of filtering for both neighborhood and room type.

The fourth and final key metric in our dashboard represents the percentage of availability in the next 30 days. Our dataset provided the total availability for the next 30, 60, 90 and 365 days as separate measures. We chose to highlight the availability for the next 30 days. As we explored the dataset and compared future availability measures with each other, we found it interesting that the availability for the next 30 days was as high as 19.1%, considering the availability for the next 60 and 90 days were 17.7% and 18.3%, respectively. These time frames correspond to Copenhagen's most active tourist season. As the data was gathered in mid-June, it covers the period of heightened tourism activity in the city, which typically spans from June through September. The reasoning for such high availability could be a pricing problem among the hosts. This is a problem directly connected to Airbnb's second strategic priority, perfecting the core service, where they seek to develop advanced tools to help hosts set competitive prices.

Originally we thought it would be a good idea to have a key metric describing the future revenue for hosts in Copenhagen, further showcasing the potential income for Airbnb itself in the Copenhagen area. Our initial thought was to multiply the price per night with the nights booked (i.e. number of days minus availability for chosen number of days). When exploring and further investigating this, however, we found that the availability is not necessarily a measure of nights booked, but it could rather be a combination of nights booked and nights blocked by the hosts. This assumption is made

on personal experience with being a host for Airbnb, in addition to examining the dataset and finding listings with zero availability the next 90 and even 365 days. The latter finding is further supported by the fact that it conflicts with listing regulations in Denmark. Since the difference is not possible to detect, leading to potential unrealistic and misleading results, we decided to distance ourselves from using this metric in our dashboard.

Charts

The first chart, meaning the top left corner, shows the growth of hosts. This chart is considered the most important because it is a simple and easy way to give a good picture on the development of the market. It is also highly related to the strategic priority of making hosting mainstream. This information is derived from the variable "host since" creating a calculated field summing up the amount of hosts that entered the market for a given year. This is shown as accumulated hosts, meaning the total for each year, as well as how many new hosts entered each year. This approach simplifies the process of understanding why growth is decelerating, while also offering the ability to access precise data on the number of entities that joined each year.

Secondly, we have the ratings chart with the objective to give an overview on customer satisfaction in the top right corner, aligned with the strategic objective of "perfecting the core service". This chart is split into the different listening types to visualize the difference in satisfaction between the different products. The ratings that are included are the "subcategories" of the overall rating shown in the key metrics pane. This is to provide deeper insight in where the differences in ratings come from.

The third chart, in the bottom left corner, provides a general overview of the listing distribution across the neighborhoods in Copenhagen. The idea is that this chart provides additional information to the first graph, showing where the hosts have their listings. The visualization is presented in a treemap format and takes the count of the measure 'Listing ID' to display the volume of listings in proportionate sizes for each neighborhood in comparison to the rest. When hovering the different neighborhoods, the tooltip shows the selected neighborhood's distribution of room types. This functionality was used by creating a separate worksheet, which displays a vertical barchart with the room type distribution, and further adding this sheet to the 'Tooltip' in the treemap worksheet. In

addition to this, the treemap chart of listing distribution is, as previously mentioned, under the influence of the filter ‘Room type’.

Lastly, we have placed a bar chart on prices in the bottom right corner. The reasoning is that it is deemed one of the least important features of our dashboard. Additionally, by placing the bar chart side by side with the listing distribution, these can be seen in combination to get a rough overview of the market sizes. The ease of combining these graphs is enhanced by coloring the neighborhoods, so that each neighborhood has the same color for both charts. When you hover over these bars, it reveals more detailed insights into the pricing dynamics, including the average price per accommodation type and a distribution breakdown. Additionally, the tooltip accessed while hovering offers information on the availability in each neighborhood. This design approach is implemented to prevent misunderstandings about pricing that could arise from variations in listing sizes among different neighborhoods. The chart maintains a similar format as the one on listings, involving the creation of a distinct sheet for the distribution data. This sheet is then incorporated into the tooltip of the "main sheet."

Filters

It was important for us to include filters for the users to be able to zoom in on subjects of interest. Given our dashboard elements that are explained above, we found it natural to give the users the opportunity to filter on neighborhood and room type. The two filters affect the charts differently, with the neighborhood filter having a direct effect on the two charts on top in addition to the key metrics. Since the two charts on the bottom show the neighborhoods in a sorted manner with the intention of providing an overview, we decided to exclude them from the neighborhood filter as this would disturb the overview itself and possibly lead to confusion for the user. When exploring this filter, the user can gain insight into the different neighborhoods’ development over time, as well as differences in ratings.

The second filter lets the user choose a room type to explore further, and it is targeted at three of the four charts, in addition to all of the key metrics. This approach enables the user to get a detailed overview of the popularity, prices and most prominent neighborhoods for the chosen room type. The filter naturally does not apply to the chart ‘Ratings for different room types’, as we wanted to keep

an overview of the rating distribution so the user can compare and have an idea of the overall standings.

Summary and Findings

The primary goal of our Airbnb dashboard is to give executives in the organization an overview on the performance of their strategic priorities. The ones we chose to highlight were “perfect the core product” and “make hosting mainstream”. In this closing chapter we will discuss how this dashboard can be used to gain insight on these points.

One of the objectives was to look into how Airbnb is performing on the goal to “perfect the core service”. As discussed in the introduction, this involves happy customers, but also a platform designed to help hosts set the “right” price. As for the latter, pricing distributions on neighborhoods can be an interesting insight. This is found hovering over the bar chart on prices, where you can see to what degree listings are centered around a mean, or have a bigger variation. Combining this with availability can give an overview on whether or not hosts are setting prices in their, and Airbnb’s best interest. The other part of this strategic priority is the quality of the product. This is also visualized in the upper right graph showing how the different products as listing types perform on the reviews. Here, we can quickly see that in large, private rooms and entire homes are quite similar in customer satisfaction, while hotel rooms and shared rooms are scoring significantly lower.

Going into the other key strategic priority, “make hosting mainstream”, information on the development of the number of hosts is obviously important. It seems prominent that Airbnb wants to continue to accelerate the number of hosts. Looking at the graph showing accumulated and new hosts, adoption seems to slow down. Finding exponential growth from the first hosts in 2010 to 2015, new hosts fell all the way to 2021. As managers and executives, finding the underlying reason is key. Market demand can have been met already, while also finding resistance in other exogenous variables. Explaining low adoption in 2020 and 2021 will have the pandemic as a central cause, while also regulations on the rental market in Denmark, imposed in 2017 could have cooling effects (Nielson, 2018) . It should be noted that the number of new hosts in 2022 is not complete, as the scrape of the data is from mid june 2022, and with a more “stable” 2023, pandemic wise, could show propelling growth yet again.

Potential growth could be further amplified by exploring the areas highlighted by the insights presented in this dashboard. Hovering over the listing chart, you can see that the number of private rooms are quite low for many neighborhoods. Seeing that customers seem just as happy with this type of listing, the potential in this segment can be huge. Combined with local knowledge on the significant amount of students living in the city, short term rental with students being home on holidays should be a lucrative market.

Limitations

As a concluding remark we want to point to some limitations of the data and analysis as a whole. One frustrating aspect is the inability to distinguish between a date that is already reserved and one that is simply unavailable when considering availability. This makes it impossible to do reliable estimates on earnings, as well as doing proper correlations between pricing and availability on listings. We also see it as a weakness that we lack time series data, positioning us to look at the development of different metrics. We were able to build some time series data on calendar data for *host_since*, but we lack data on potential customers that have opted out of the platform along the way. Including the publication dates for each listing would significantly enhance the analysis. We hypothesize that existing hosts are more likely to introduce new properties on the platform compared to new hosts, suggesting a trend towards increased market concentration. Although we attempted to validate this by analyzing the number of hosts and listings, we relied on the *first_review* date as a proxy. However, initial analysis showed certain years with four times as many hosts as listings, proving the weakness of this assumption.

References

Airbnb. (2023, 11 01). *Shareholder letter Q3 2023*. nordnet.dk. Retrieved 12 17, 2023, from https://www.nordnet.dk/markedet/aktiekurser/17400014-Airbnb-a

IBM. (n.d.). Exploratory data analysis. IBM. Retrieved 16. 12. 2023 from https://www.ibm.com/topics/exploratory-data-analysis

IBM. (2021). Dimensional Schemas. IBM. Retrieved 16. 12. 2023 from https://www.ibm.com/docs/en/ida/9.1.2?topic=design-dimensional-schemas

Jukic, N., Vrbsky, S., & Nestorov, S. (2019). *Database Systems: Introduction to Databases and Data Warehouses* (2nd ed.). Prospect Press.

Kimball R. & Ross M. (2002). The data warehouse toolkit : the complete guide to dimensional modeling (2nd ed.). Wiley.

Microsoft (2023), Description of the database normalization basics. Retrieved from https://learn.microsoft.com/en-us/office/troubleshoot/access/database-normalization-description

Nielson, E. G. (2018, May 17). Airbnb to report homeowners' income to Danish tax authorities. *Reuters*. https://www.reuters.com/article/us-Airbnb-denmark/Airbnb-to-report-homeowners-income-to-dani sh-tax-authorities-idUSKCN1II1HV/

Patterson, R. E., Blaha, L. M., Grinstein, G. G., Liggett, K. K., Kaveney, D. E., Sheldon, K. C., Havig, P. R., & Moore, J. A. (2014). A human cognition framework for information visualization. *Computers & Graphics*, *42*, 42–58. https://doi.org/10.1016/j.cag.2014.03.002

Redman, T. (2017, September). Only 3% of companies' data meets basic quality standards.

Harvard Business Review.

https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards

Tufte, E. R. (2006). The visual display of quantitative information. *Information Design Journal*, *4*(3). https://doi.org/10.1075/idj.4.3.12cos

Yigitbasioglu, O. M., & Velcu, O. (2012). A review of dashboards in performance management: Implications for design and research. *International Journal of Accounting Information Systems*, *13*(1), 41–59. https://doi.org/10.1016/j.accinf.2011.08.002