Project 1

Data Warehousing for Business Intelligence

Business domain: Hotel and Accommodation

By

6388030	Kulawut	Makkamoltham
6388040	Ariya	Phengphon
6388104	Peerawat	Sorosthunyapong
6388133	Pitchaya	Teerawongpairoj
6388196	Sasima	Srijanya

Present to

Asst. Prof. Dr. Songsri Tangsripairoj

A Report Submitted in Partial Fulfillment of

the Requirements for

ITCS 453: Data Warehousing and Data Mining

Faculty of Information and Communication Technology

Mahidol University

Semester 2/2022

Executive Summary

This report studies the business requirements of Airbnb and the importance of data warehouse and Business Intelligence to Airbnb, Inc. Methods of Analysis includes the

business requirements extraction to create ERD, star schema, and data dictionary; data collection; ETL process to extract, join, and clean data with Tableau Prep; and data visualizations in various form such as bar chart, map, pie chart, etc. with Tableau Desktop. Results of the study show that Airbnb uses a data warehouse and Business intelligence to make data-driven decisions, which help it to understand the user needs and improve the user experience accordingly. There are 8 visualizations created according to the design of the fact table in the star schema.

The report finds that the decision-making process of Airbnb heavily depends on the data generated on the platform. Airbnb uses a data warehouse and BI to analyze the data about accommodation, booking and reservation, reviews, and revenue.
Recommendations include:

- Using visualization generated to understand the user's needs.
- Using the insight from the analysis to improve the user experience on the platform.
- Using the full dataset to make use of the full potential of the data warehouse and BI.

The report also investigates the fact that the analysis conducted has limitations. Some of the limitations include Lacking a real-world dataset, and the data about revenue and bookings being confidential. Thus, they are not available on the internet and the team needs to generate them. Unfortunately, the data is also generated randomly since the trends and behaviors of the users are unknown.

Table of Contents

Introduction

Business Domain Overview

The Hotel and Accommodation sector focuses on providing lodging and hospitality to guests. The main functions of this business domain include

1. Property management - Managing the physical facilities that serve the customers.
2. Reservation and bookings - Currently, most accommodations provide an online reservation and booking to the customers.
3. Marketing and Promotion - Hotels and accommodations attract guests to stay, so they have to attract customers by creating a promotion or marketing strategy.

In addition, It is a crucial part of the tourism industry, which is one of the top 10 fastest-growing industries with approximately 10% YoY. [1] However, the covid-19 pandemic causes a great loss to the whole tourism industry which covers hotels and accommodations since people are afraid of the disease and avoid traveling. As a result, all the businesses in the Hotel and Accommodation sector have their revenue and the number of visitors diminished. Fortunately, many countries declared covid-19 as an endemic disease. They reduced the strictness of the policy, allowing people to have a free vacation, so the sector started to recover globally in the last quarter of 2021.

Objectives and scope of the project

The objective of this project is to analyze the database and data warehouse schema and analyze the data of Airbnb to discover the benefits of Business Intelligence and a Data warehouse in the company. In addition, the scope of the project covers

- Business requirements extraction
- The data collection process by importing the open datasets on the internet.
- Generating data to analyze, Extract, Transform, and Load (ETL) data with Tableau Prep.
- Using Tableau to analyze and visualize the data to gain insights from the dataset.

The expected outcomes of these processes are

- ERD and Star Schema design
- Data dictionary
- Cleaned Dataset
- Visualizations, a Dashboard, and a storyboard to provide insight to stakeholders.
- Video showing ETL process and visualization.

Business requirements for building a data warehouse

Airbnb is an online marketplace that allows hosts to rent their places and guests to book accommodations. It generates around 20 Terabytes daily and stores 1.4 petabytes. The data is generated from the customers that use the platform and processed by the data scientists to provide the best matching experience between guests and hosts. AirBnb needs a high-performance data warehouse to handle a large amount of data for these business requirements:

1. Accommodation - Hosts list their accommodations including information about the number of rooms, description, and amenities. AirBnb needs to analyze Accommodation data for the recommendation system that helps customers find preferred rental places.
2. Booking and Reservation - Guests can book accommodation on the platform. This data helps Airbnb to gain insights into user behavior such as how long they want to stay, the high season and low season, and what kind of accommodations people prefer.
3. Reviews - Guests can create a review by commenting and choosing the score of each category of the place they stayed. AirBnb can determine the popularity and quality of the hosts from the reviews. This information will be used to rank the search along with the hundreds of criteria. [2]
4. Revenue Management - The main source of revenue of AirBnb is transaction fee which costs from 14.2% to 20% of a transaction for the guests and 3% for the hosts. [3] This data will be used to analyze the profit and adjust the price of the place to increase the booking chance.

How data warehouse and BI are important to the company

Data warehouse and Business Intelligence are the keys to Airbnb's success. Airbnb uses Superset as a BI solution mainly because it has low migration cost and it is an open-source project that Airbnb contributes efforts to develop features. [4] Since Airbnb's main revenue is from transaction fees, the booking and reservation traffic is considered the lifeline of Airbnb. Data warehouses and BI are crucial tools that help Airbnb to find insight into user behaviors and be able to adapt the business model to meet the hosts' and guests' expectations.

1. Improving features - By using data warehouse and BI, Airbnb can create features to help improve the user's decisions. For example, the Airbnb price tips feature helps the hosts to determine the likelihood of being booked with the price, so the hosts can adjust their prices according to the tip. In addition, it is found that the hosts that set

the price within 5% of the suggestion are likely to get booked 4 times more than the ones whose prices are more than 5% away. [5] Another example of a feature that is improved with big data is the search feature, Airbnb collects and analyzes the interaction between hosts and guests to provide a personalized search feature to the customers.[6]

2. Data-driven decisions - According to the blog [6], thousands of employees need data to make decisions on their daily job. In addition, 90% of dashboards are viewed more than once daily. With the BI tools, employees can do the analytics tasks by themselves and be able to use data-driven decisions to ideate the products and services that improve the users' experiences.
3. Discovering the real user needs - Many hotels and accommodations suffer from the COVID-19 pandemic because people chose to stay at home to prevent infections from late 2019 until 2021. The number of bookings on Airbnb dropped by 70% after a few months of the pandemic. [11] Later, Airbnb found that people are eager to travel as they are staying at home for almost 2 years. [7], [8] Moreover, most people are allowed to work from home, so they are likely to look for a long-term stay in different places. In the 2021 report, it is found that people stayed 28 nights or more, and 60% of them were guests who work or study during their rental. [9] Airbnb decided to shift its model to focus on long-term stays to meet customer needs by creating a promotion for long-term stays.

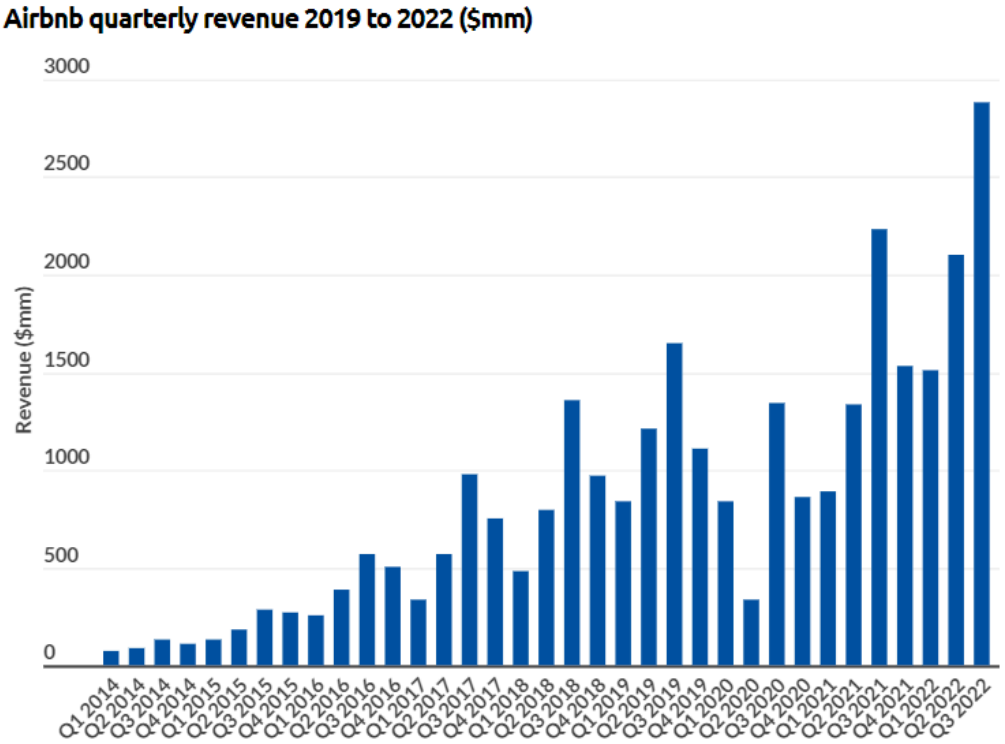


Figure 1: Airbnb quarterly revenue 2019 to 2022 [10]

Figure 1 illustrates the Airbnb quarterly revenue from 2019 to 2022 in million USD. It shows that the revenue of Airbnb is in a downtrend from Q3 2019 until Q2 2021 and the trend started to change after Q3 2021, which is a few months after Airbnb implemented the model that focuses on a long-term stay. As a result of discovering user habits, Airbnb can adapt its business model causing it to regain its value from the crisis in a short time.

Data Sources

ER Diagram

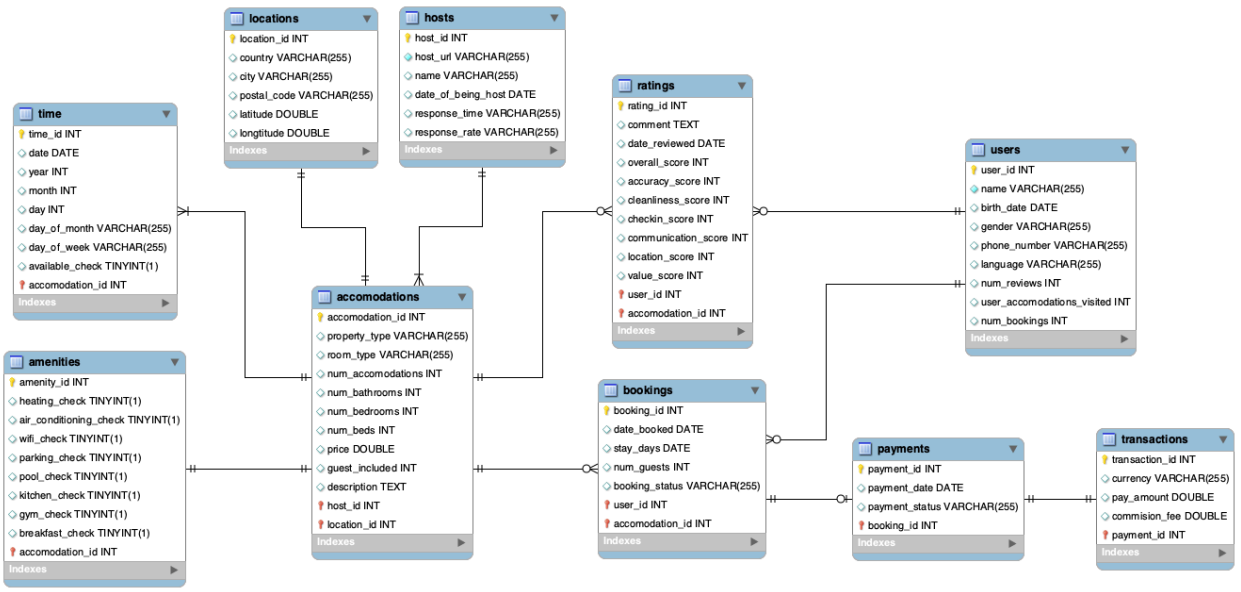


Figure 2: ER Diagram of Airbnb

Figure 2 shows the entity relationship diagram (ERD) of the Airbnb database schema. The database schema defines several tables that together represent the entities, attributes, and relationships of an accommodation booking system which is the major business process of Airbnb. This diagram consists of 10 tables which are users, hosts, locations, accommodations, amenities, bookings, time, ratings, payments, and transactions. The tables are related through foreign keys and contain various attributes such as user information, accommodation details, location data, ratings, and payment information.

Data Dictionary of Operational Database (OLTP)

Table: users

ttribute Name	Description	Data Type	Data Format	Constraint	Example
user_id	Unique identifier for the user	INT	NNNNNNNN	PRIMARY KEY	39674640
name	User's name	VARCHAR(255)		NOT NULL	David
birth_date	User's date of birth	DATE	MM/DD/YYYY		06/24/1986
gender	User's gender	VARCHAR(255)			Male, Femal
hone_number	User's phone number	VARCHAR(255)	+NN(NN)NNNN-NNN		+49(30)1570-9
language	User's preferred language	VARCHAR(255)			English

ttribute Name	Description	Data Type	Data Format	Constraint	Example
num_reviews	Number of reviews the user has written	INT			20
r_accomodations _visited	Number of accommodations the user has visited	INT			33
um_bookings	Number of bookings the user has made	INT			121

The users table stores information about users who have signed up on Airbnb. Each user is identified by a unique user_id attribute, and the table includes details such as their name, birth date, gender, phone number, and language preference together with user activity information such as the number of reviews they have given, the number of accommodations they have visited, and the number of bookings they have made. This table is used to track user activity on the platform and to provide personalized recommendations and communication.

Table: hosts

ttribute Name	Description	Data Type	Data Format	Constraint	Example
host_id	Unique identifier for the host	INT	NNNNNNNN	PRIMARY KEY	51484542
host_url	Host's URL on Airbnb website	VARCHAR (255)		NOT NULL	https://www.airbnb.co sers/show/5148454
name	Host's name	VARCHAR (255)			Julia
e_of_being_host	Date when the host started being active on Airbnb	DATE	MM/DD/YYYY		12/16/2015
esponse_time	Average response time of the host	VARCHAR (255)			Within a few hours Within a day, etc.
esponse_rate	Average response rate of the host	VARCHAR (255)	N%		100%

The host table stores information about hosts who have listed accommodations on Airbnb. Each host is identified by a unique host_id attribute, and the table includes details

such as their name, host URL, the date they started hosting on the platform, their average response time, and response rate. This table is used to track host activity and to provide a way for guests to communicate with their hosts.

Table: locations

Attribute Name	Description	Data Type	Data Format	Constraint
location_id	Unique identifier for the location	INT		PRIMARY KEY
country	Country where the location is	VARCHAR(255)		
city	City where the location is	VARCHAR(255)		
postal_code	Postal code of the location	VARCHAR(255)		
latitude	Latitude of the location	DOUBLE	N.NNNNN	
longitude	Longitude of the location	DOUBLE	N.NNNNN	

The locations table stores information about the locations of accommodations listed on Airbnb. Each location is identified by a unique location_id attribute, and the table includes details such as the country, city, postal code, latitude, and longitude of the location. This table is used to provide accurate search results and to help guests make informed decisions about where to book their accommodations.

Table: accommodations

Attribute Name	Description	Data Type	Data Format	Constraint
accommodation_id	Unique identifier for the accommodation	INT	NNNNNNNN	PRIMARY KEY
property_type	Type of the accommodation	VARCHAR(255)		
room_type	Type of the room	VARCHAR(255)		
num_accomodations	Number of accommodations available in the property	INT		

Attribute Name	Description	Data Type	Data Format	Constraint
num_bathrooms	Number of bathrooms in the property	INT		
num_bedrooms	Number of bedrooms in the property	INT		
num_beds	Number of beds in the property	INT		
price	Price per night for the property	DOUBLE		
guest_included	Maximum number of guests allowed in the property	INT		
description	Description of the property	TEXT		

Attribute Name	Description	Data Type	Data Format	Constraint
host_id	Unique identifier of the host who owns the property	INT	NNNNNNNN	FOREIGN KEY
location_id	Unique identifier of the location of the property	INT		FOREIGN KEY

The accomodations table stores information about individual accommodations listed on Airbnb. Each accommodation is identified by a unique accomodation_id attribute, and the table includes details such as the property type, room type, the number of accommodations, bathrooms, bedrooms, and beds, the price per night, the number of guests included, and a description of the property. This table also includes foreign keys to the hosts and locations tables to link the accommodation to the host and location. This table is used to display accommodation listings to guests and to manage bookings.

Table: amenities

Attribute Name	Description	Data Type	Data Format	Constraint
amenity_id	Unique identifier for the amenity	INT		PRIMARY KEY
heating_check	A flag indicating whether amenity has heater	BOOLEAN		
air_conditioning_check	A flag indicating whether amenity has air conditioner	BOOLEAN		
wifi_check	A flag indicating whether amenity provides wifi service	BOOLEAN		
parking_check	A flag indicating whether amenity	BOOLEAN		

Attribute Name	Description	Data Type	Data Format	Constraint
	provides parking service			
pool_check	A flag indicating whether amenity has swimming pool	BOOLEAN		
kitchen_check	A flag indicating whether amenity has kitchen	BOOLEAN		
gym_check	A flag indicating whether amenity has gym	BOOLEAN		
breakfast_check	A flag indicating whether amenity provides breakfast	BOOLEAN		

Attribute Name	Description	Data Type	Data Format	Constraint
accomodation_id	Unique identifier of the accommodation where provided	INT	NNNNNNNN	FOREIGN KEY

An amenities table is used to indicate the amenity in an accommodation such as air conditioner, wifi service, parking service, swimming pool, etc. The accommodation_id is the foreign key from the accommodation table, which is for identifying the accommodation. The available amenities are indicated by the boolean value, True means the accommodation has the amenity while False means the amenity is not available at the accommodation.

Table: bookings

Attribute Name	Description	Data Type	Data Format	Constraint
booking_id	Unique identifier for a booking	INT		PRIMARY KEY
date_booked	The date when the booking was made	DATE	MM/DD/YYYY	

Attribute Name	Description	Data Type	Data Format	Constraint
stay_days	The duration of the stay	DATE		
num_guests	The number of guests included in the booking	INT		
booking_status	The status of the booking	VARCHAR(255)		
user_id	The user who made the booking	INT	NNNNNNNN	FOREIGN KEY
accomodation_id	The accommodation that was booked	INT	NNNNNNNN	FOREIGN KEY

The bookings table stores information about individual bookings made by guests on Airbnb. Each booking is identified by a unique booking_id attribute, and the table includes details such as the date the booking was made, the dates of the stay, the number of guests, the status of the booking, and foreign keys to the users and accomodations tables to link the

booking to a specific user and accommodation. This table is used to track guest bookings and to manage payments and transactions related to those bookings.

Table: time

Attribute Name	Description	Data Type	Data Format	Constraint
time_id	Unique identifier for a time record	INT		PRIMARY KEY
date	The date represented by the time record	DATE	MM/DD/YYYY	
year	The year represented by the time record	INT	YYYY	
month	The month represented by the time record	INT	MM	
day	The day represented by the time record	INT	DD	

Attribute Name	Description	Data Type	Data Format	Constraint
day_of_month	The day of the month represented by the time record	VARCHAR(255)		
day_of_week	The day of the week represented by the time record	VARCHAR(255)		
available_check	A flag indicating whether the accommodation is available on the given date	BOOLEAN		
accomodation_id	The accommodation associated with the time record	INT	NNNNNNNN	PRIMARY KEY, FOREIGN KEY

The time table stores information about the availability of individual accommodations listed on Airbnb. Each entry in the table includes a unique time_id attribute, along with the date, year, month, day, day of the month, and day of the week that the accommodation is available. The table also includes a boolean value indicating whether the accommodation is available on that date and a foreign key to the accomodations table to link each entry to a

specific accommodation. This table is used to display availability calendars for accommodations and to manage bookings.

Table: ratings

Attribute Name	Description	Data Type	Data Format	Constraint
rating_id	Unique identifier for a rating	INT	NNNNNNNN	PRIMARY KEY
comment	The comment associated with the rating	TEXT		
date_reviewed	The date when the rating was made	DATE	MM/DD/YYYY	
overall_score	The overall score given in the rating	INT		RANGE(1-100)
accuracy_score	The accuracy score given in the rating	INT		RANGE(1-10)

Attribute Name	Description	Data Type	Data Format	Constraint
cleanliness_score	The cleanliness score given in the rating	INT		RANGE(1-10)
checkin_score	The check-in score given in the rating	INT		RANGE(1-10)
communication_score	The communication score given in the rating	INT		RANGE(1-10)
location_score	The location score given in the rating	INT		RANGE(1-10)
value_score	The value score given in the rating	INT		RANGE(1-10)
user_id	The user who made the rating	INT	NNNNNNNN	FOREIGN KEY

Attribute Name	Description	Data Type	Data Format	Constraint
accomodation_id	The accommodation associated with the rating	INT	NNNNNNNN	FOREIGN KEY

The ratings table stores information about reviews that guests leave for accommodations after their stay. The table includes attributes such as the comment left by the guest, the date the review was submitted, and scores for various aspects of the accommodation (e.g. accuracy, cleanliness, communication). The ratings table is linked to the users and accommodations tables through foreign keys, which allows us to associate reviews with the guest who left them and the accommodation they stayed in. This information can be used to provide valuable feedback to hosts and to help guests make informed decisions about which accommodations to book.

Table: payments

Attribute	Description	Data Type	Data Format	Constraint
payment_id	Unique identifier for a payment	INT	NNNNNN	PRIMARY KEY
payment_date	Date of the payment	DATE	MM/DD/YYYY	

Attribute	Description	Data Type	Data Format	Constraint
payment_status	Status of the payment	VARCHAR(255)		
booking_id	ID of the booking associated with the payment	INT		

The payments table stores information about payments made by guests to book accommodations. This includes details such as the date the payment was made, the payment status, and the booking ID associated with the payment. The payments table is linked to the bookings table through a foreign key, which allows us to associate payments with the bookings they are associated with. This information can be used to keep track of payments made by guests and to ensure that hosts receive payment for their bookings.

Table: transactions

tribute Name	**Description**	**Data Type**	Data Format	**Constraint**	Example
ansaction_id	Unique identifier for a transaction	INT		PRIMARY KEY	50
currency	Currency used in the transaction	VARCHAR(255)			€, $, 𝔹, etc.
ay_amount	Amount paid in the transaction (i.e. days of stay * price)	DOUBLE			455
mmission_fee	Commission fee charged for the transaction (i.e. 10% of paid amount)	DOUBLE			45.5
ayment_id	ID of the payment associated with the transaction	INT	NNNNNN	FOREIGN KEY	457353

The transactions table stores information about financial transactions related to bookings. This includes details such as the currency used in the transaction, the amount paid, and any commission fees charged. The transactions table is linked to the payments table

through a foreign key, which allows us to associate transactions with the payments they are associated with. This information can be used to track the financial aspects of bookings and to ensure that Airbnb receives appropriate compensation for bookings made through their platform. Since the dataset used for the case study is from Berlin, the currency is Euro. In addition, pay_amount is calculated from the price and the days of stay while the commission fee is 10% of the pay_amount.

Data Warehouse Design

Star/Snowflake Schema Diagram

From our observation and analysis through the Airbnb website, our group decided to create three fact tables including reservation, review, and revenue fact tables.

Reservation fact table

Reservation fact table is about users reserving accommodation via the Airbnb website or application. Users will see the detail of accommodations such as room type, number of rooms of each accommodation, and amenities that the accommodation provided for users that are available on Airbnb's platform for users to decide to reserve the accommodation. There are six dimensions including user, booking, time, location, amenity, and accommodation. Our group analyzes the amount of each accommodation booking, the amount of availability of each accommodation, the summation of users that stay at the accommodation, and the highest booking of each accommodation per week as the fact measure.

Review fact table

Review fact table is about reviewing host accommodation reviews by users. Users will review by rating the accommodation in each part including overall accommodation, accuracy, cleanliness, check-in, location, value, and comment and the host will respond to users' ratings. There are four dimensions including time, user, host, and rating. Our group analyzes the number of reviews that users review of each accommodation, the average rating of each service including accuracy, cleanliness, check-in, communication, location, and value, and the overall rating of each accommodation as the fact measure.

Revenue fact table

Revenue fact table is about Airbnb's revenue by collecting commission fees from each transaction which is 10% of the amount of payment. (This is only the assumption from the research since the actual revenue is not accessible) For example, if the user paid €1404, Airbnb will receive (a commission fee) €140.4 and the host will receive €1263.6 for this transaction. Also, users can check the status of the payment if it is complete or not. There are six dimensions including time, user, host, accommodation, payment, and transaction. Our group analyzes the revenue of Airbnb, the number of completed payments, and the amount of not completed payments.

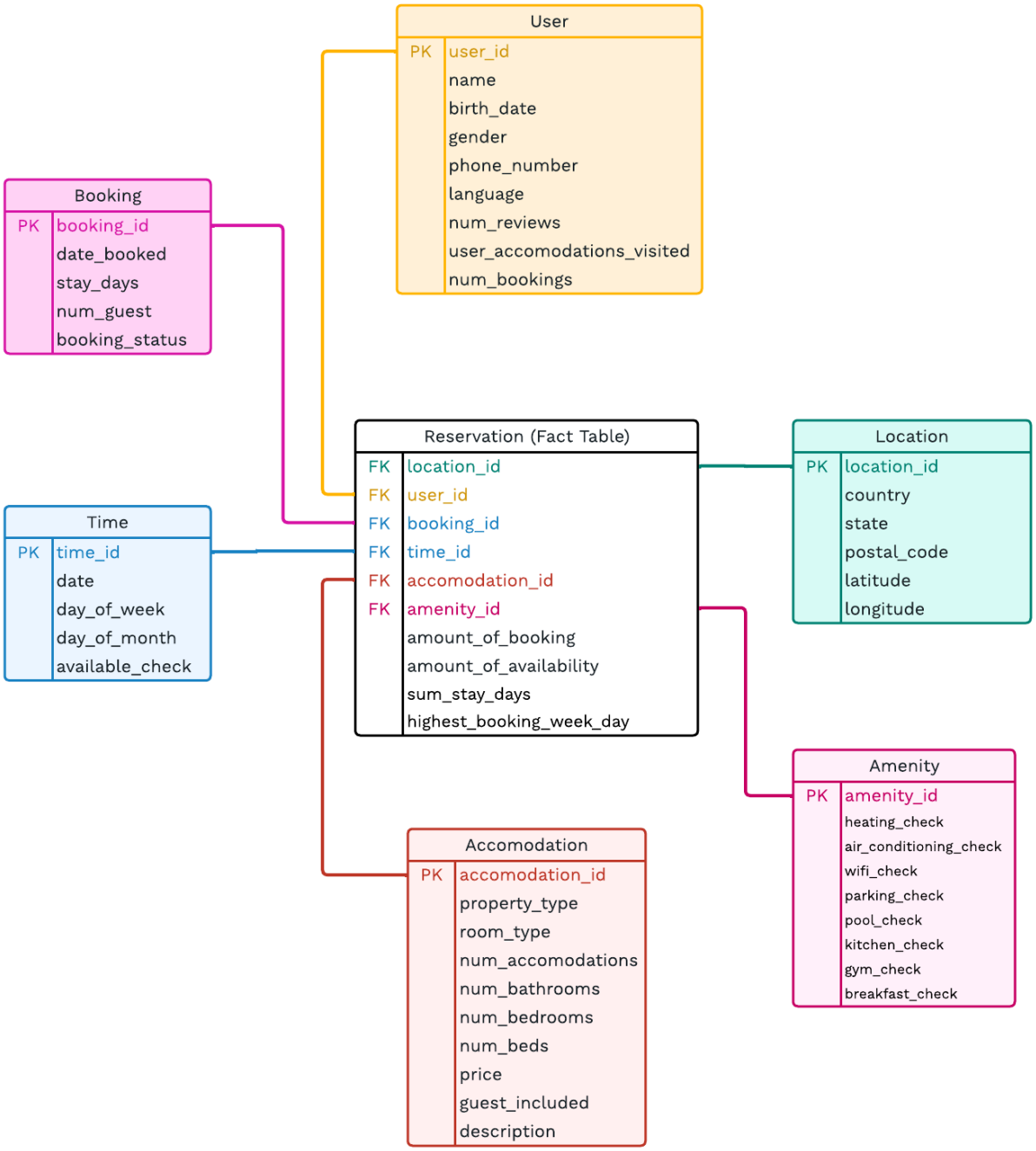


Figure 3: Reservation Fact Table

Figure 3 shows the reservation fact table that has the FK mapping to the PK. For instance, amount_of_booking is the number of all bookings that appear in the Airbnb system. To calculate this fact measure, count the number of booking_id required similar to Visualization 3.

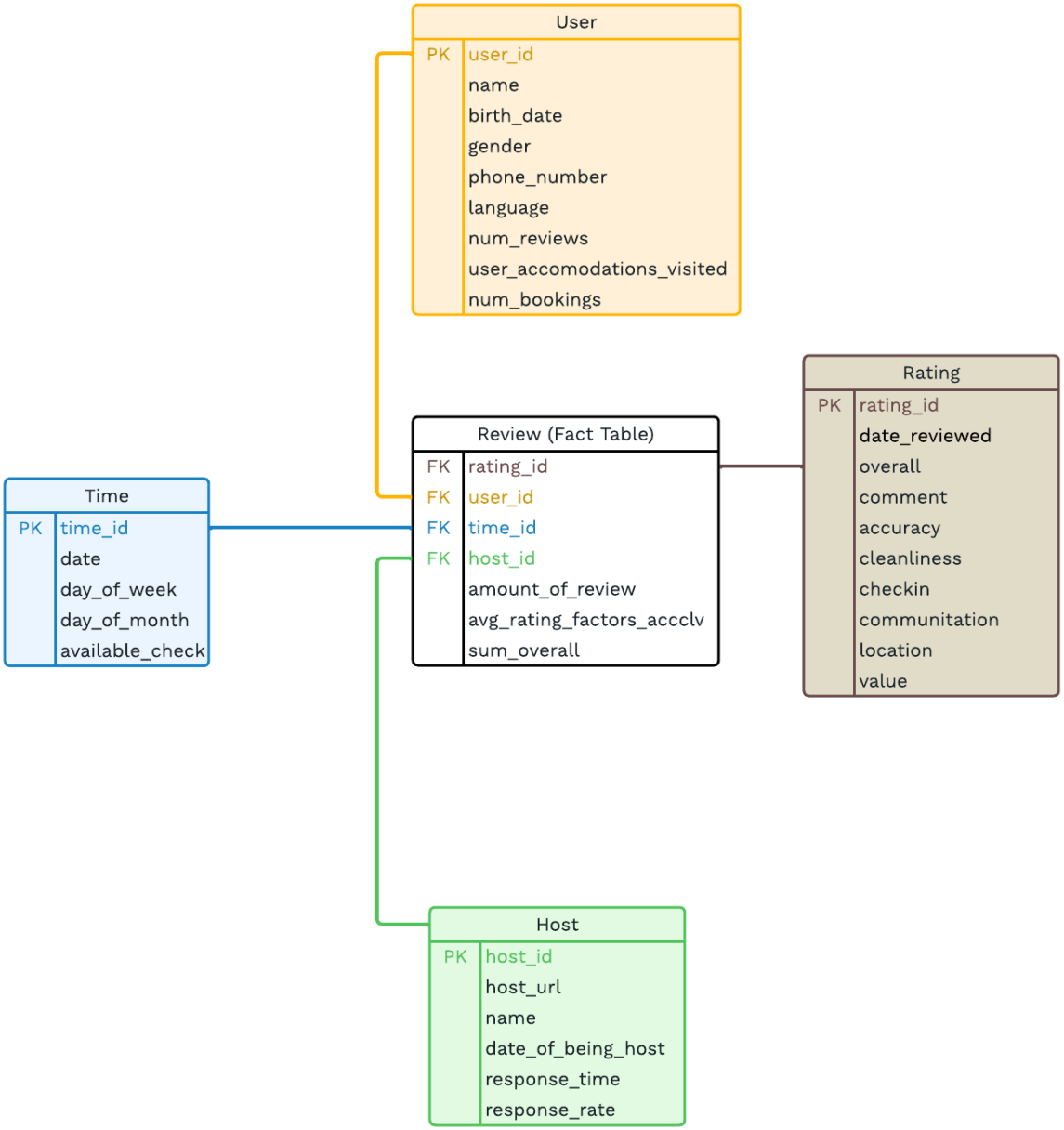


Figure 4: Review Fact Table

Figure 4 shows the review fact table that has the FK mapping to the PK. For instance, amount_of_review is the number of all reviews that appear in the Airbnb system. To calculate this fact measure, it needs to count the number of rating_id in the Rating dimension table similar to Visualization 1.

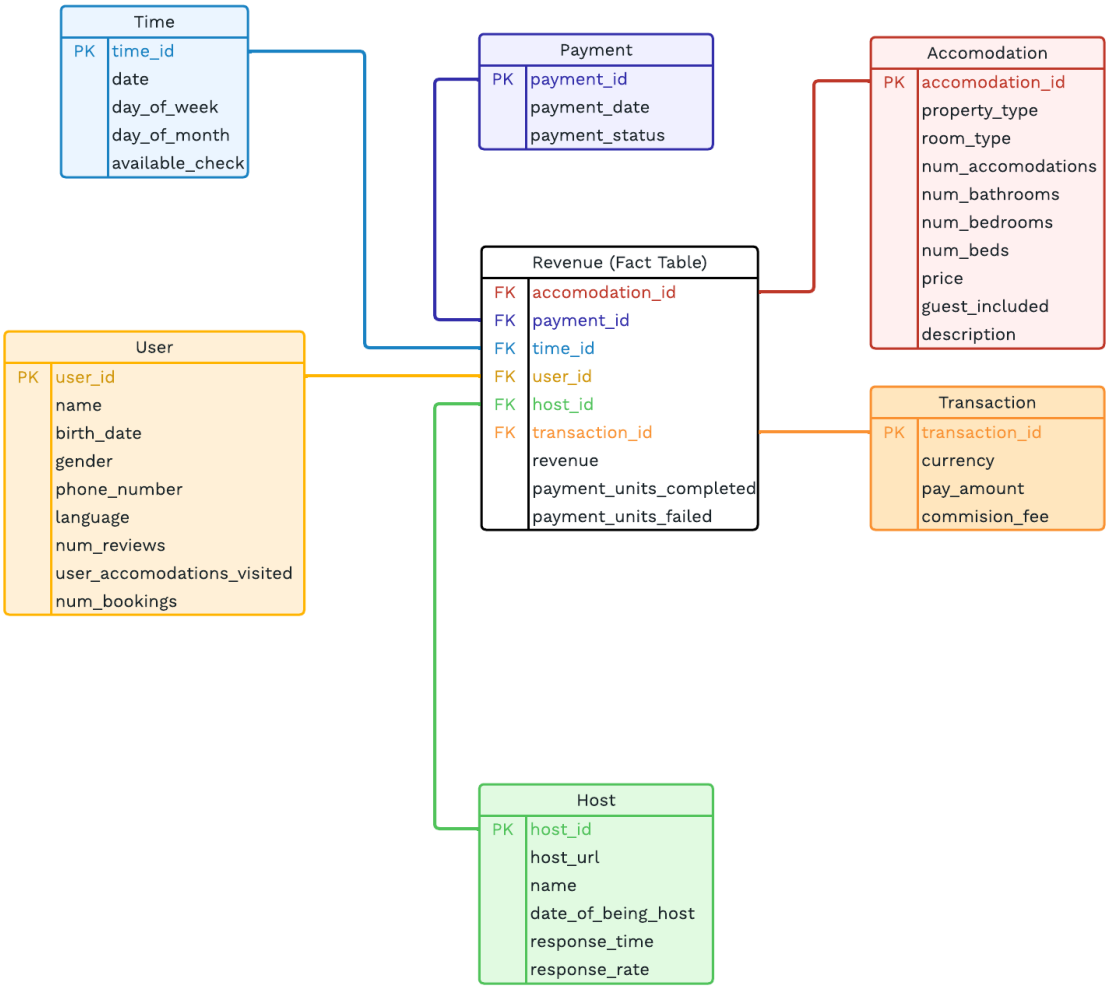


Figure 5: Revenue Fact Table

Figure 5 shows the revenue fact table that has the FK mapping to the PK. For instance, revenue is the total of Airbnb's revenue. To calculate this fact measure, it needs to sum up all of the records in the commission_fee attribute in the Transaction dimension table.

ETL Process

OLTP, OLAP, and ETL are three crucial concepts in data management that are closely related to each other.

OLTP stands for Online Transaction Processing, a type of database system used for managing and processing transactional data. It involves the collection and processing of individual transactions. OLTP systems are commonly used in operational environments, where real-time processing and responsiveness are critical as OLTP databases are read, written, and updated frequently.

OLAP, Online Analytical Processing, is a type of data processing utilized for analytical data management. It involves the analysis of large volumes of data to derive insights and support decision-making. OLAP systems are typically used in reporting and analysis environments, where historical data and trends are important [14].

ETL, which stands for Extract, Transform, and Load, is a process of integrating data from multiple sources into a single destination system. ETL is an essential process for organizations that integrate data from multiple sources and make it available for analysis and reporting. The process is often automated using specialized tools such as ETL software like Tableau Prep.

Extract

- In the extraction step, raw data is extracted from various sources, which can be both structured or unstructured data, such as databases, spreadsheets, or web services. The goal is to collect the necessary data that will be used in the subsequent steps of the ETL process [12].

Transform

- In this stage, the raw data that has been extracted will be transformed into a format suitable for analysis or storage in the destination system. Transformations may include cleaning the data, removing duplicates, merging multiple sources, or changing the data format [12].

Load

- Lastly, once the transformed data has been loaded into the destination system, such as a data warehouse or a business intelligence tool, it becomes readily available for various applications. For instance, it can be used for reporting and analysis purposes, allowing users to derive valuable insights and make informed decisions based on the data.

Tableau Prep

- Tableau Prep is an ETL tool that is commonly used for small and some medium-sized projects. However, it may be challenging or costly to utilize for production flows that require periodic refreshing. The tool's performance may suffer from a limitation when dealing with larger datasets or more complicated transformations [13].

The relationship between OLTP, OLAP, and ETL is that they can work together to support the data needs of an organization, with OLTP serving as a source system for ETL, and OLAP providing analytical capabilities for the transformed data.

Screenshots of Tableau Prep step-by-step

This section will show the screenshot step-by-step way to perform an ETL data cleaning in Tableau Prep.

1) Importing the data set

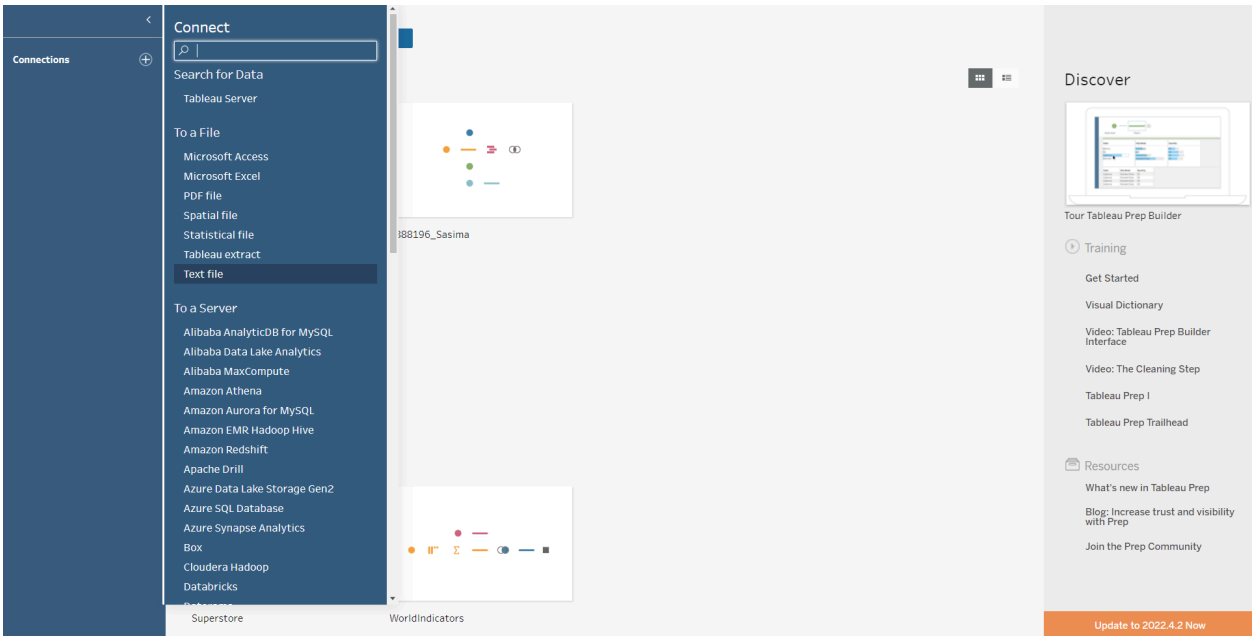


Figure 6: Text File Connection

Initially, upon launching Tableau Prep, we will navigate to the "Connections" tab to establish a connection to the file that we wish to use. For this particular project, we have chosen to connect to a "Text file".

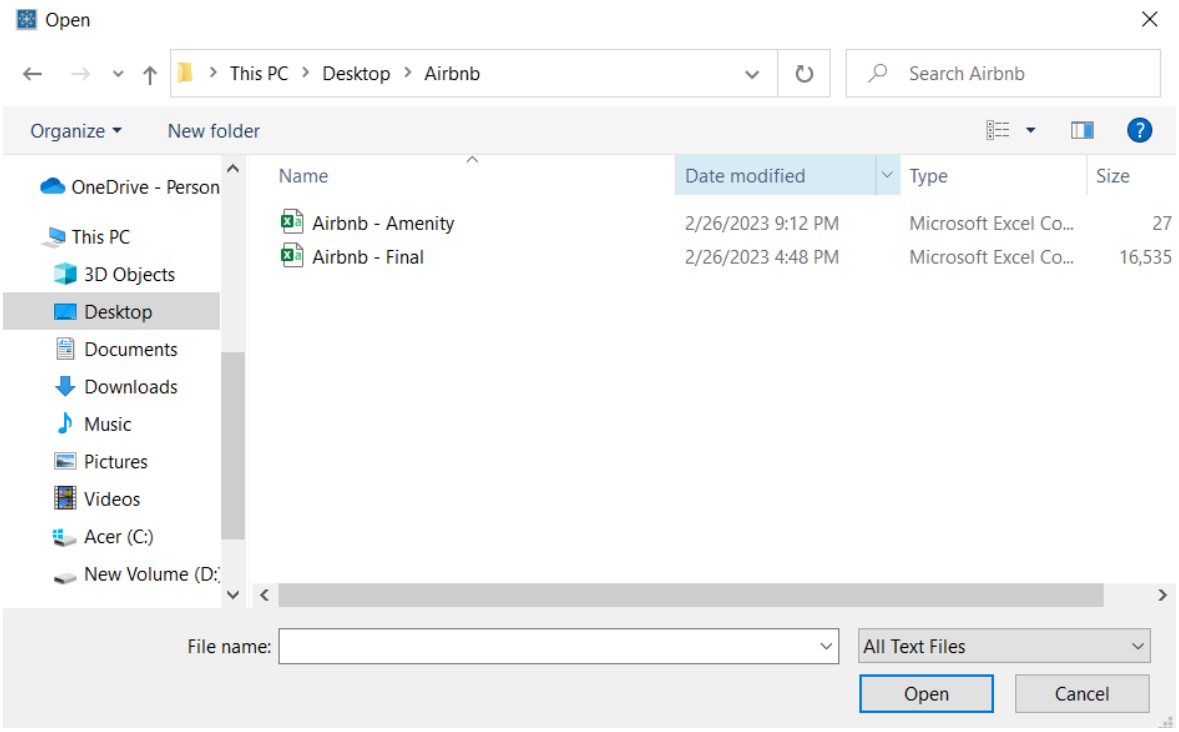


Figure 7: Airbnb Data Sets

Afterward, we can select the desired folder and files to be imported into Tableau Prep for data processing. Our team has gathered two files as our data sets for Airbnb. We will begin the data cleaning process by selecting the Airbnb-Final and Airbnb-Amenity files.

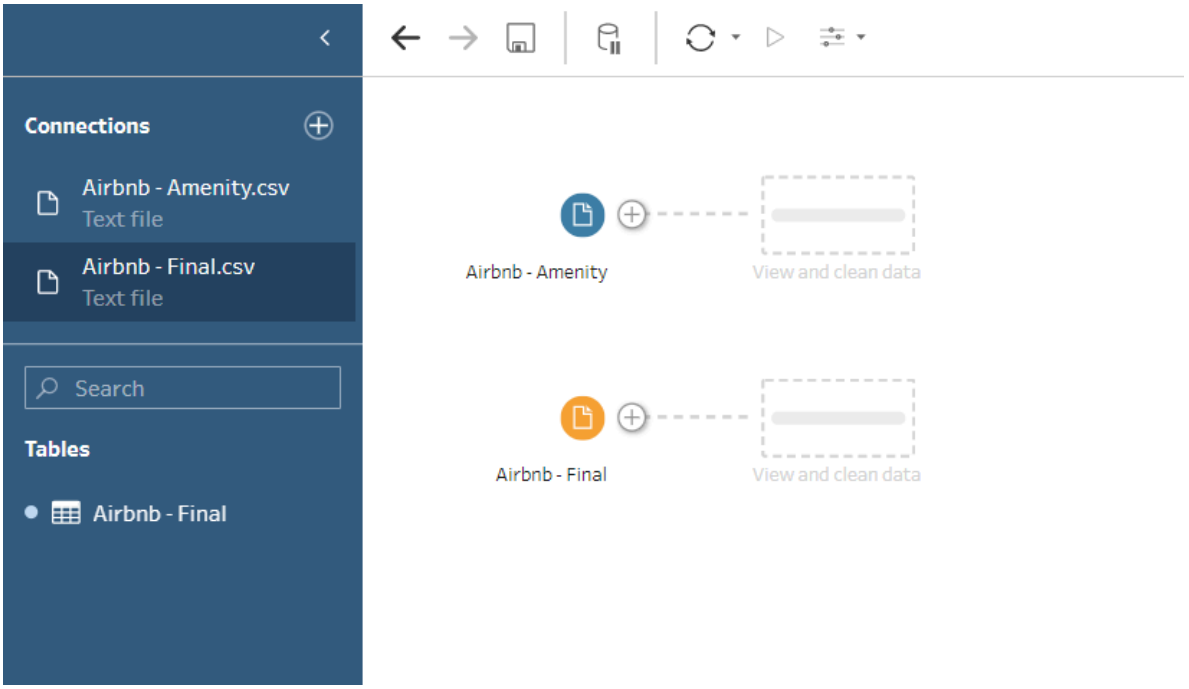


Figure 8: Data Sets Connections Result

2) Data Cleaning

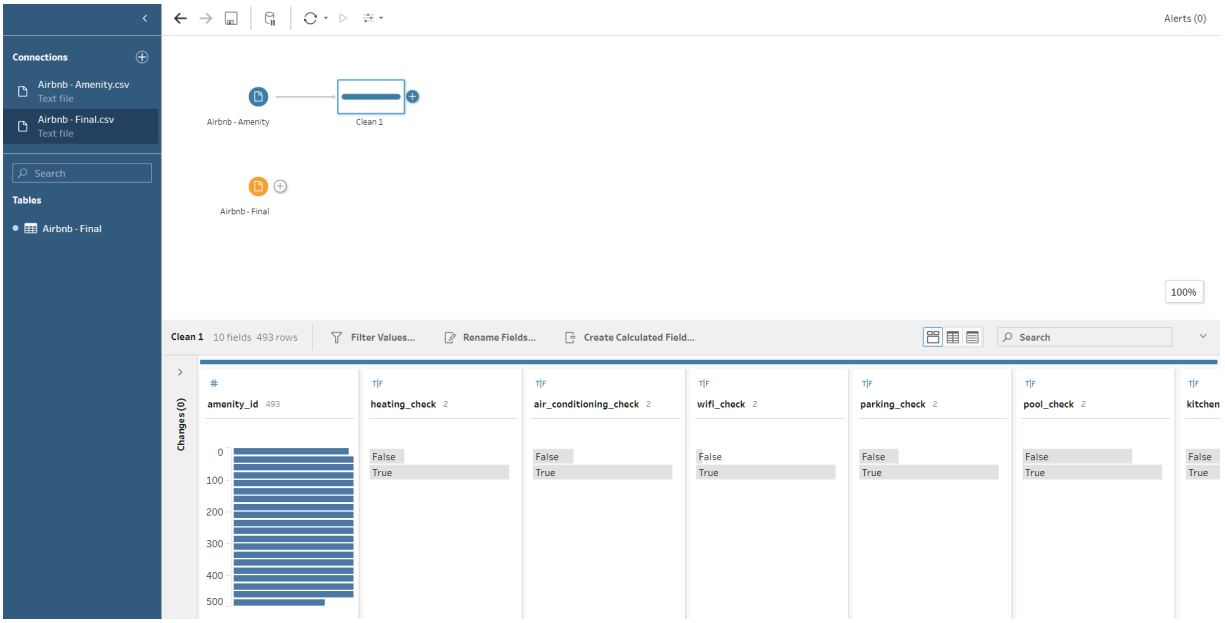


Figure 9: Add Clean Step

To add a data cleaning step for each file, we can easily do so by clicking on the "+" symbol and dragging it onto the canvas. This will enable us to perform data cleaning operations on each file individually, such as removing duplicates, renaming columns, changing data types, or filtering rows based on certain criteria. By adding these cleaning steps to the workflow, we can ensure that our data is accurate, consistent, and formatted correctly, which in turn will help us create more insightful and meaningful visualizations in Tableau.

2.1) Change Formatting

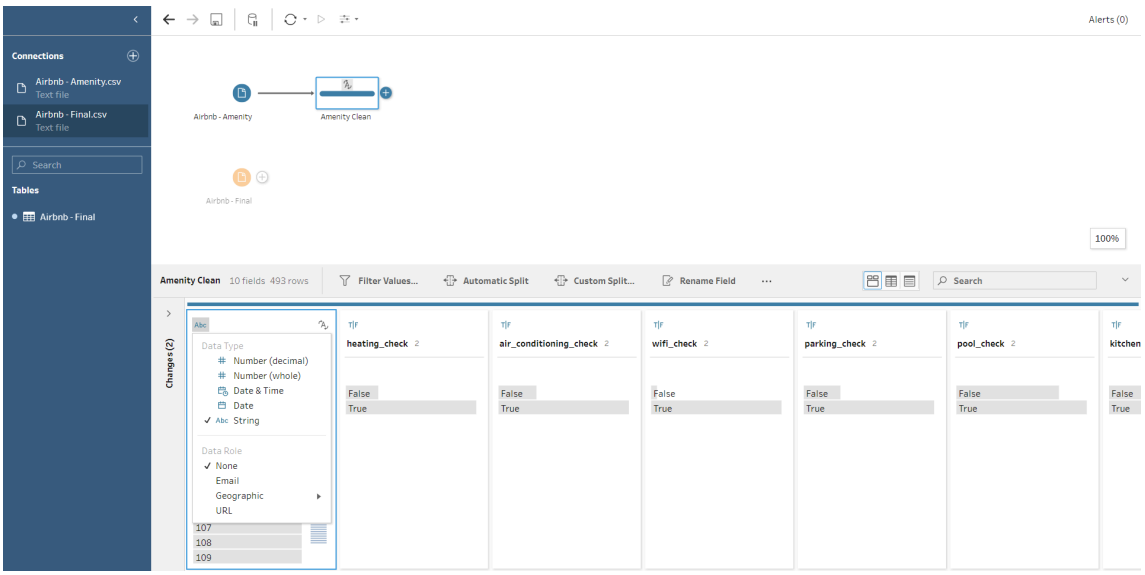


Figure 10: Changing Data Type

In Tableau Prep, changing data types is a straightforward process that can be done through the "Data Type" option in the "Profile" pane. To change a data type, we need to select the field we want to modify, then click on the "Data Type" dropdown menu and select the new data type we want to apply. We can choose from a variety of data types, including string, date, datetime, boolean, and numeric. After selecting the new data type, we can preview the changes in the "Preview" pane to ensure that the data is transformed correctly.

2.2) Dealing with missing or incorrect data

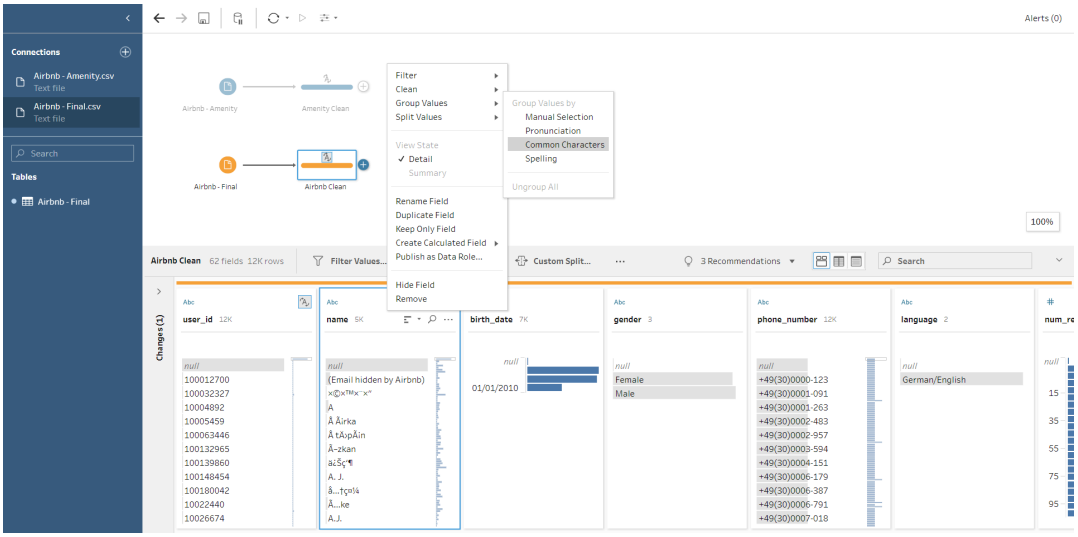


Figure 11: Replacing Incorrect Data

Dealing with missing or incorrect data is an important step in data cleaning. One way to address this is by using group values or common characters to replace missing or incorrect data. For example, if the dataset contains inconsistent values, we can replace them with the desired value.

2.3) Remove Unwanted Column

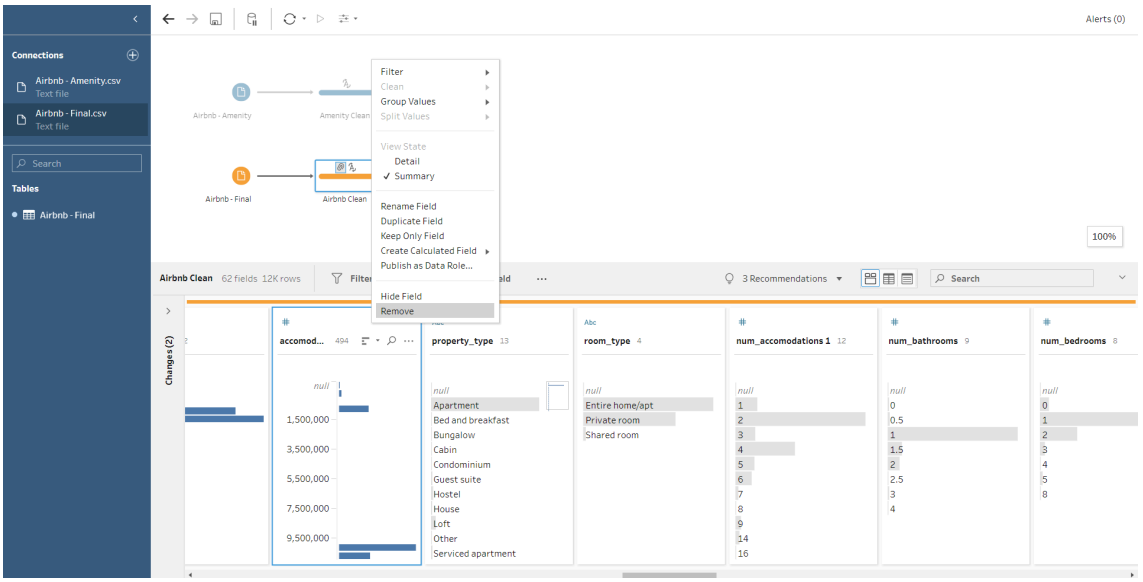


Figure 12: Removing Data Fields

When cleaning data, sometimes it is necessary to remove unnecessary or irrelevant fields from the dataset. In Tableau Prep, removing any data fields is simple. First, we have to click on the field name to highlight it. Then, right-click and press the "remove" button, which appears on the screen to remove the field from the dataset.

2.4) Renaming Field

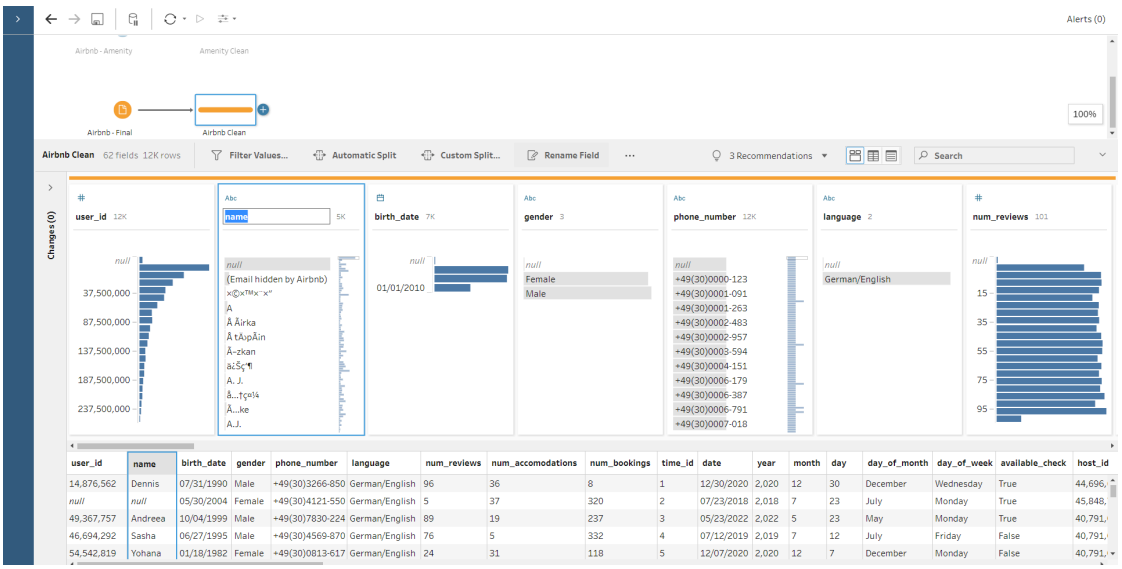


Figure 13: Renaming Data Fields

In Tableau Prep, we can easily change the name of a field during the cleaning step. This can be done by selecting the field in the data grid or the Fields pane, and then clicking on the "Rename Field" option. We can then enter a new name for the field, and the changes will be automatically applied to all the data in the workflow.

2.5) Changes Tab

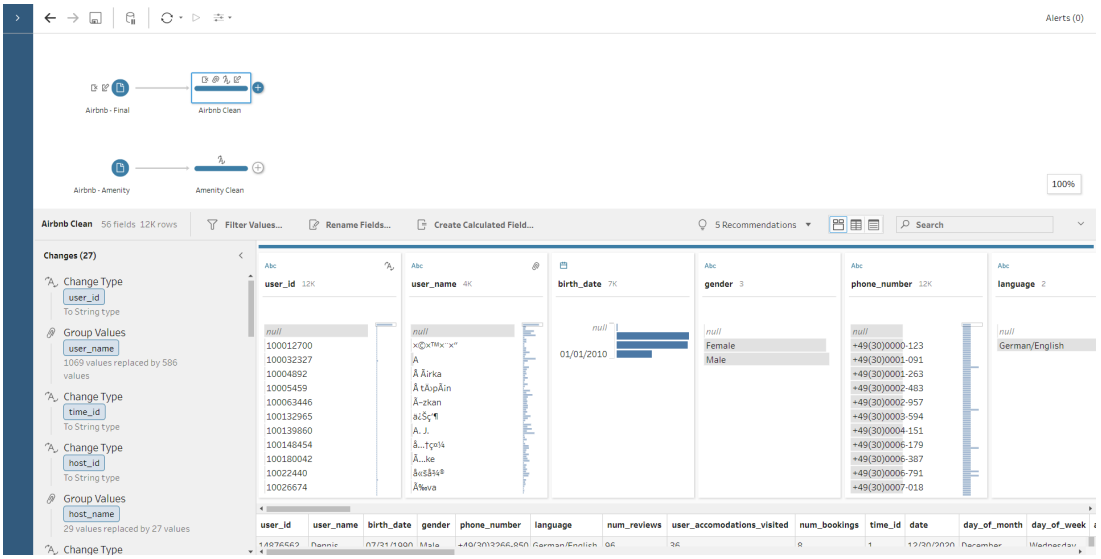


Figure 14: Airbnb Clean Changes

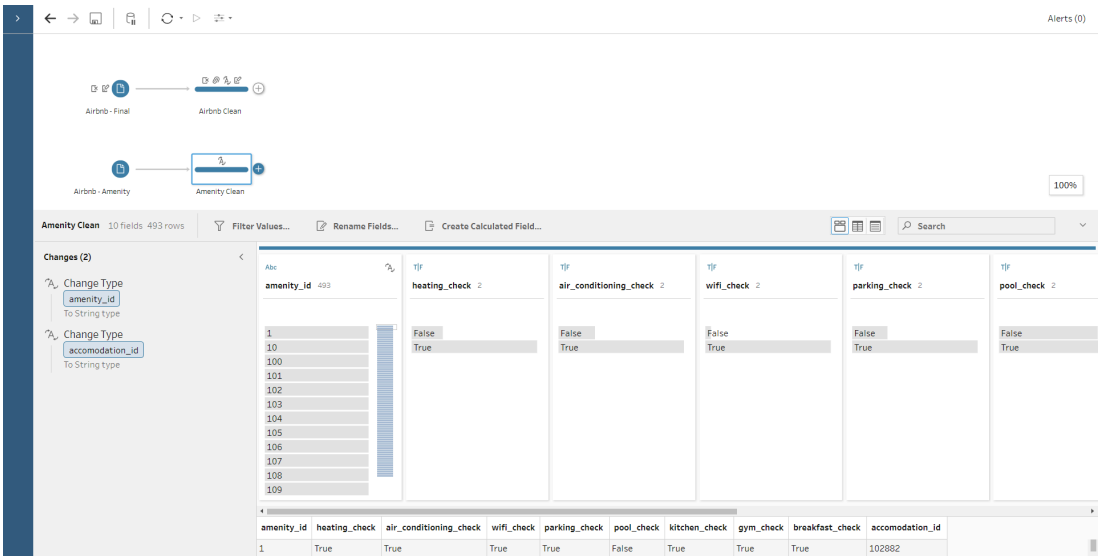


Figure 15: Amenity Clean Changes

The "Changes" tab in Tableau Prep's cleaning step displays and allows users to change the values of specific fields. It can be accessed by clicking on the "Changes" button located on the left side of the cleaning step pane. Once inside, users can see and choose the field they want to modify and apply changes such as renaming, replacing, or deleting values, and even converting values to different data types.

3) Join Method

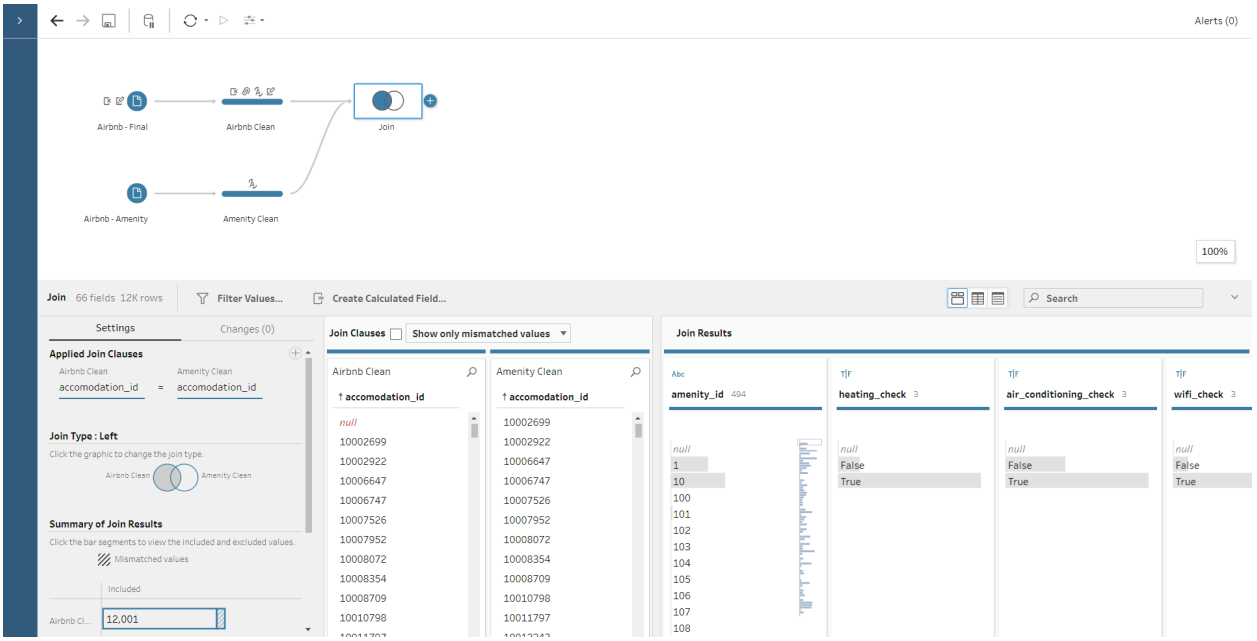


Figure 16: Left Join Method

Tableau Prep offers a convenient way to combine data from multiple sources into a single dataset using the Join step. The Join step requires you to select the two data sources that you want to join, and specify the join key(s) that will be used to match records between the two sources. The join key(s) are typically the fields that have common values in both datasets. In this project, we have used accommodation id to be our key.

Once the two data sources have been selected and specified the join key(s), we can choose the type of join we want to perform. Tableau Prep offers several types of join, including Inner Join, Left Join, Right Join, and Full Outer Join. Each type of join produces a different result, depending on how it matches records from the two datasets. Our team has selected Left Join to be our approach since we want to return all the records from the left table (Airbnb Clean) and the matching records from the right table (Amenity Clean).

In addition to the join type, we can choose to keep all records or only match records. Keeping all records will result in a larger dataset that includes all records from both sources while keeping only matching records will result in a smaller dataset that includes only the records that match between the two sources.

4) Export the Output

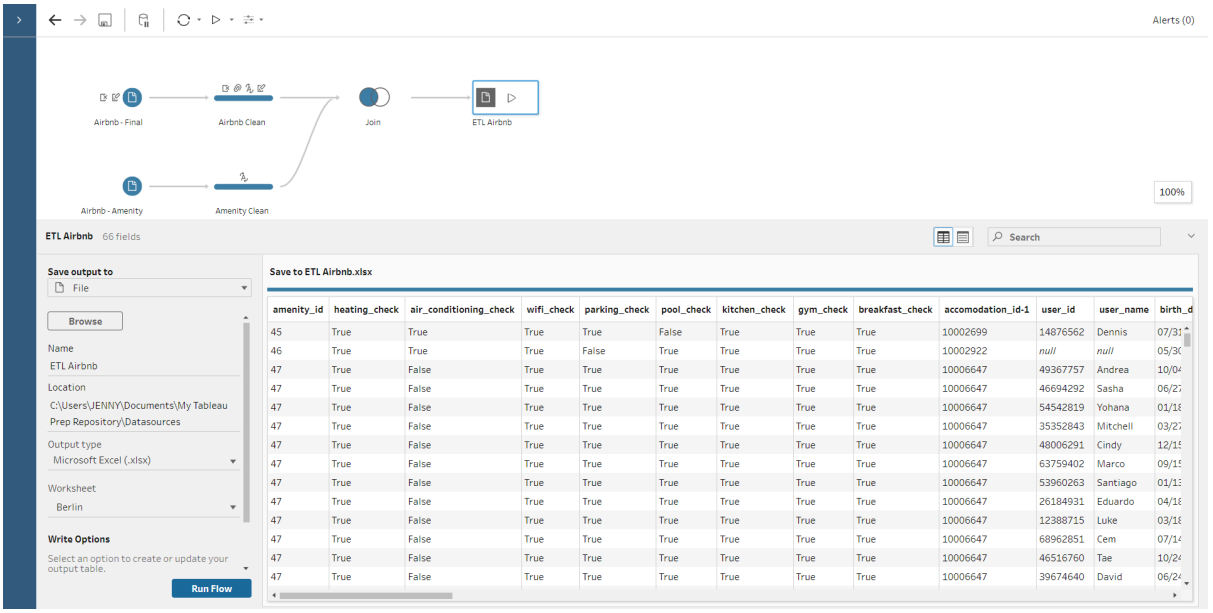


Figure 17: Exporting File

Exporting data from Tableau Prep is a straightforward process that allows you to save the cleaned and transformed data in a format that can be easily shared or further analyzed in other tools or platforms. To export the data, you need to switch to the "Output" tab, where you can choose the desired file type, such as CSV or Excel, and specify the file name and location where you want to save the file. Additionally, you can include field names in the output and select which fields to include or exclude, depending on your requirements.

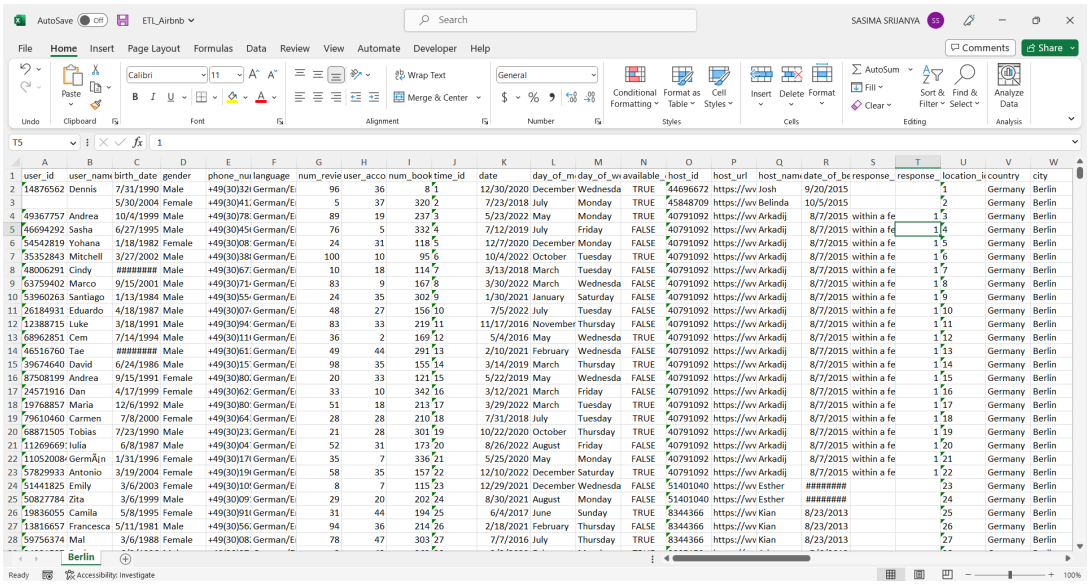
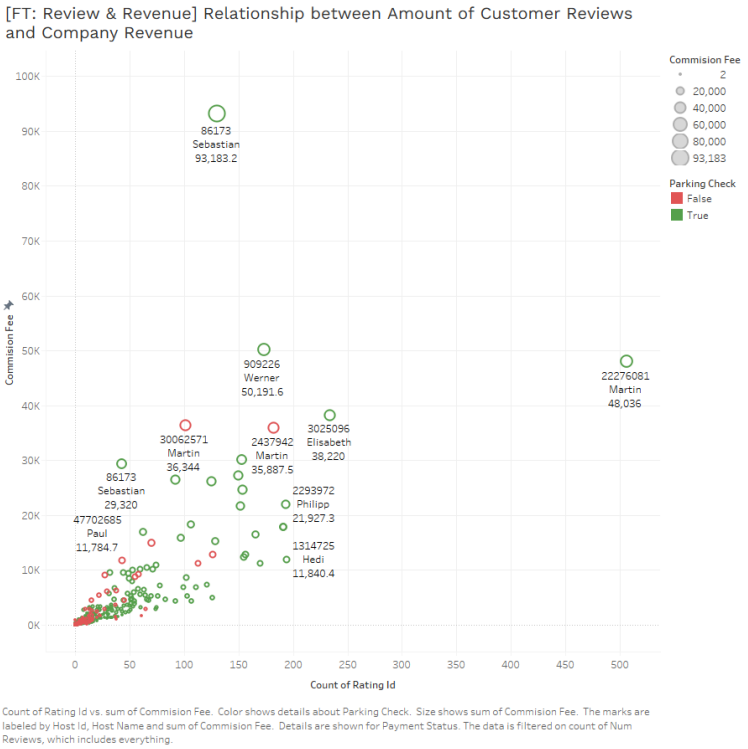


Figure 18: ETL Airbnb Microsoft Excel File Result

Analysis and Visualization Report

Business Requirements: Reviews, Revenue Management

Fact Table and Measures: Review Fact Table (amount_of_review), Revenue Fact Table (revenue)



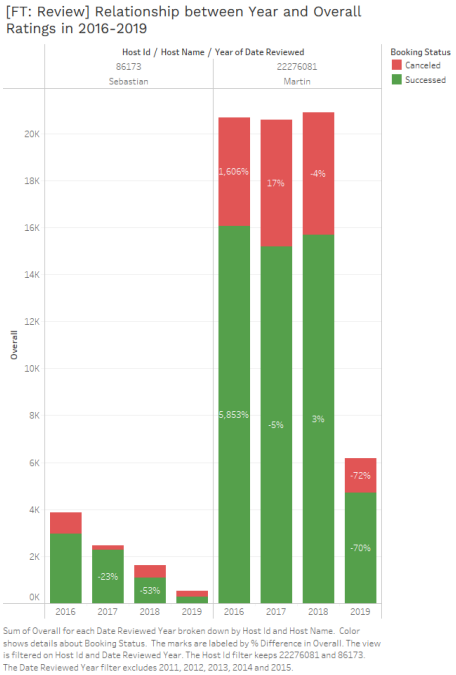
Visualization 1: Relationship between Amount of Customer Reviews and Company Revenue

Visualization 1 illustrates the relationship between two fact measures, amount of review and revenue from the Review Fact Table and Revenue Fact Table respectively. It appears that the host id who makes the highest revenue to the organization is “Sebastian”, however, the number of reviewers is significantly lower than “Martin” who receives the highest amount of reviews from the users. Moreover, on the left bottom corner, there are many host id that does not include parking reservation with lower organization revenue and customer reviews.

This visualization would help the organization see the revenue and host id who makes the highest revenue to the organization. They can use this information to boost the view rate of the popular accommodation provided by this host id and recognize what factor makes each accommodation gains interest from the users.

Business Requirements: Reviews, Booking, and Reservation

Fact Table and Measures: Review Fact Table (sum_overall)



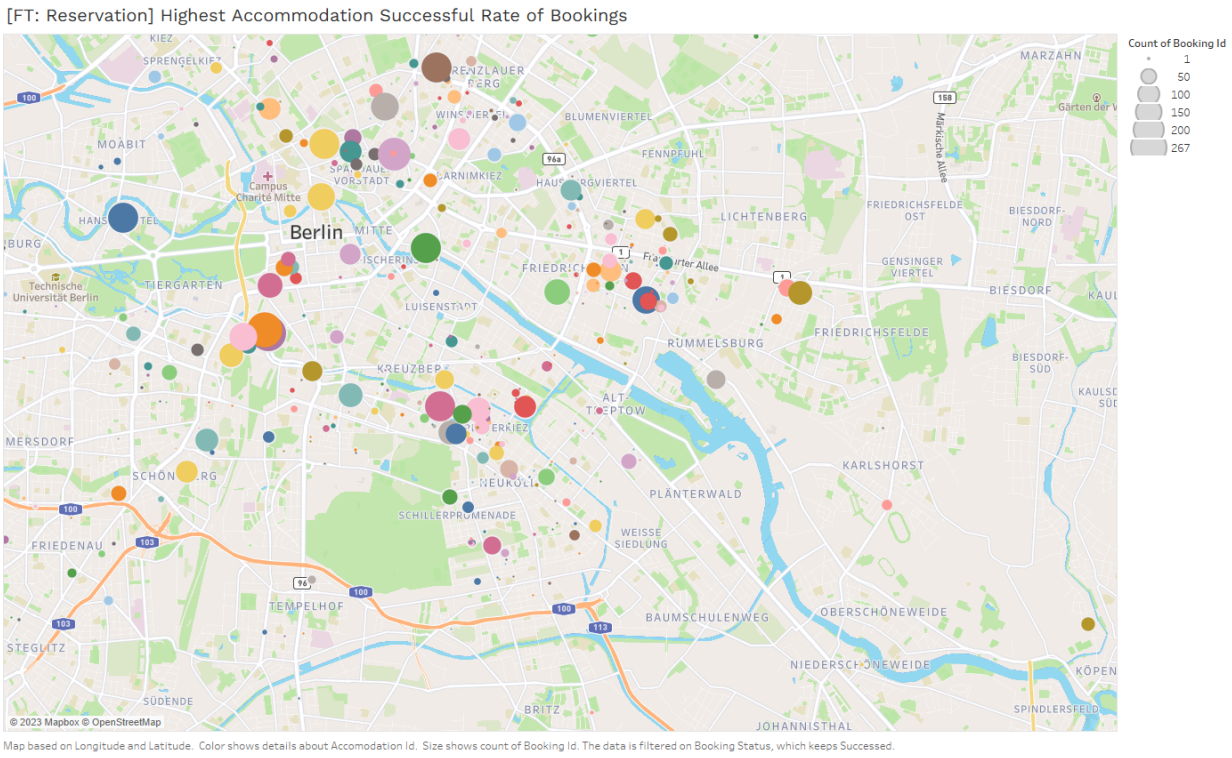
Visualization 2: Relationship between Year and Overall Ratings in 2016-2019

Visualization 2 illustrates the relationship between reviews' overall from the Review Fact Table. This chart is the continuation of visualization 1 since our group spots the host id that makes the most revenue to the organization and the host id that gains the highest number of users' reviews is different. Therefore, our group drilled down into the review scores of those two mentioned host id to see the difference and trends between years. It appears that even though Sebastian makes the highest revenue to the organization, the customer reviews are decreasing every year, especially for the year 2019 when both host id receives lower customer reviews. Moreover, for the color masking, it is the payment status (Complete and Fail). It shows that users are typically satisfied and tend to give a higher score when the payment is completed (Green color), however, it is just only our assumption and there might be other several factors to analyze.

This visualization would help the organization understand the previous visualization better, since the previous one does not provide enough information on the review scores, and trends between each year.

Business Requirements: Accommodation, Booking and Reservation

Fact Table and Measures: Reservation Fact Table (amount_of_booking)



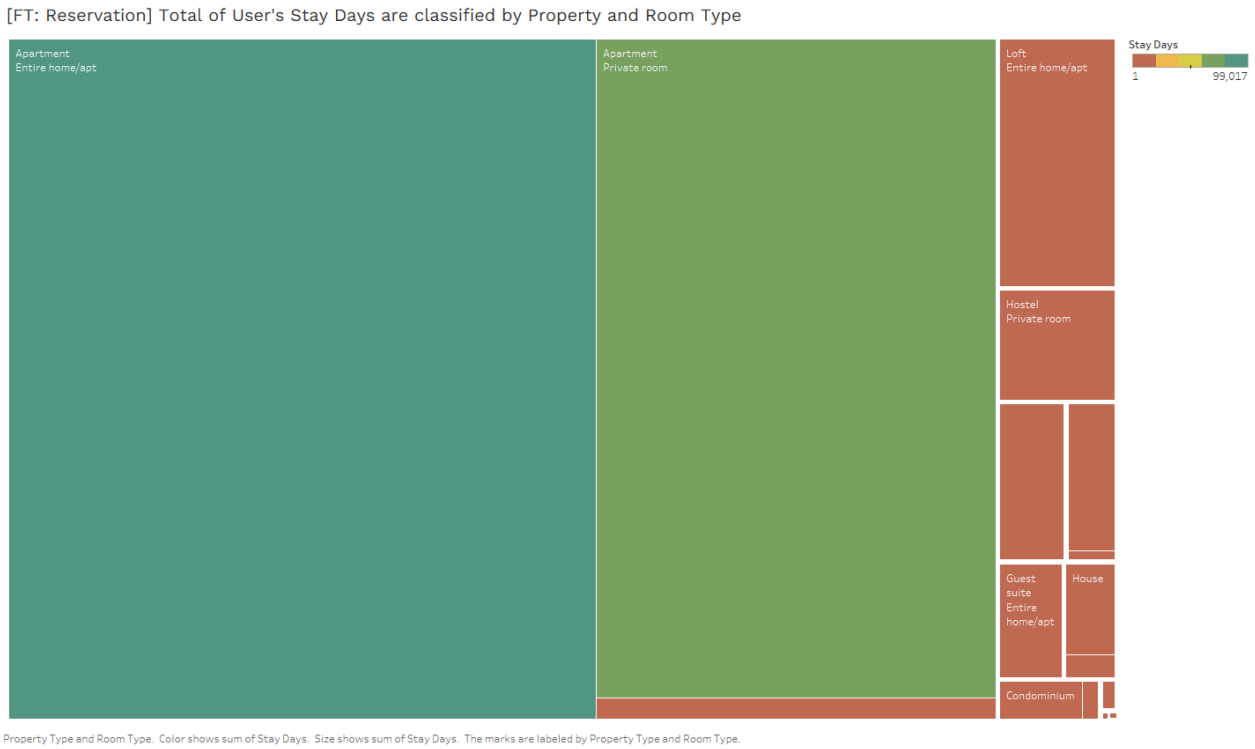
Visualization 3: Highest Accommodation Successful Rate of Bookings

Visualization 3 illustrates the amount booking accommodation from the Reservation Fact Table. It shows the accommodation that receives the most successful bookings by the users. The accommodation that has the highest will be represented with the biggest circle. Since the map has a large scale, it would be better to display this chart on the Tableau Desktop for zooming in on the details.

This visualization would help the organization see the accommodation's success rate of bookings and accommodations' dispersion so that the organization can see the accommodation scattered or the overall zone the accommodation is located in. For example, the figure shows that the accommodation that receives the successful rate of booking is most likely to be located in the center of Berlin.

Business Requirements: Accommodation, Booking, and Reservation

Fact Table and Measures: Reservation Fact Table (sum_stay_days)



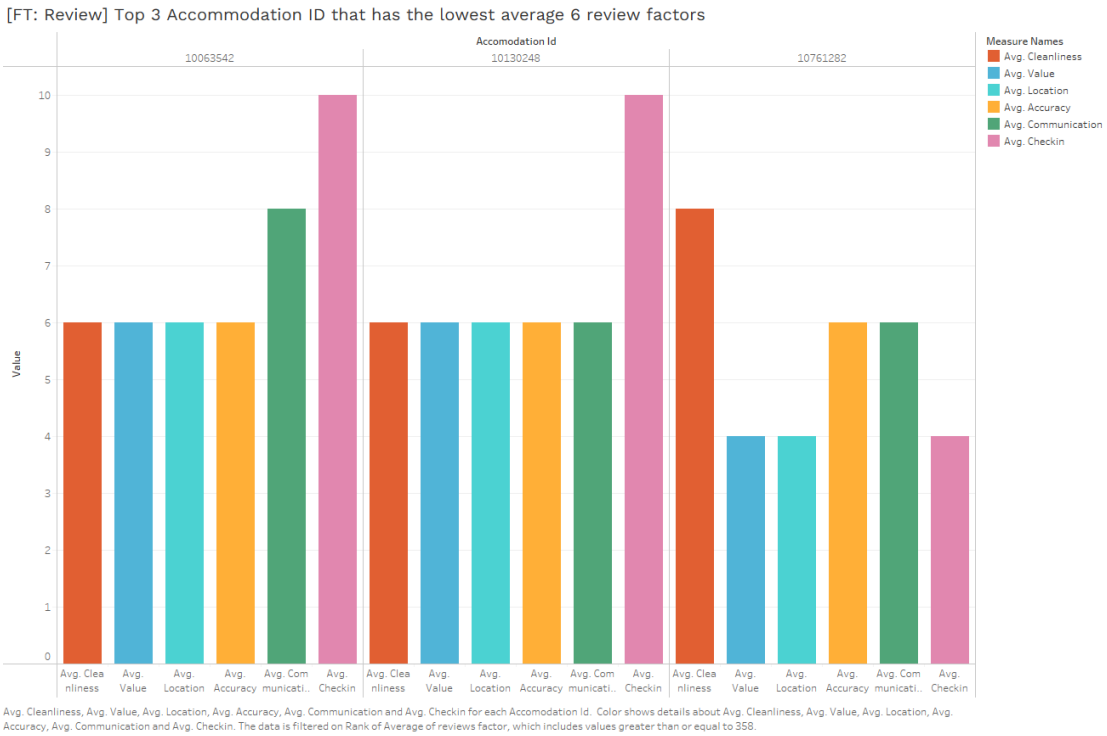
Visualization 4: Total of User's Stay Days are classified by Property and Room Type

Visualization 4 illustrates the property and room type stayed by the user from the Reservation Fact Table. This figure has two levels of classification: Property and Room Type. It shows that the Apartment that is an Entire home/apartment is the most used and stayed by the customers. It can imply that the customer might prefer staying in this type of property or book it for personal purposes for a long period (since it has the highest stay days).

This visualization shows the customer accommodation preferences, which types of property, and room types they tend to prefer and choose to stay so that the organization can provide more of this type of accommodation for them to choose from. In contrast, it might help the organization to rethink the lower stay days accommodation. It cannot be concluded that this type of property is not satisfactory, since it has several different factors such as price range. For example, the price of a House might be more expensive than the Apartment, so the user chooses the lower price to stay instead.

Business Requirements: Accommodation, Reviews

Fact Table and Measures: Reservation Fact Table (avg_rating_factors_accclv)



Visualization 5: Top 3 Accommodation ID that has the lowest average 6 review factors

Visualization 5 illustrates the average ratings received (it has 6 main review factors for each accommodation: Accuracy, Cleanliness, Check In, Communication, Location, and Value) from the users from the Review Fact Table. Although it has the overall column to represent the total score of each record, it is not enough for analyzing the individual factor. Therefore, this visualization shows the top 3 accommodations that receive the lowest average 6 review factors (calculated by the sum of each column divided by a total number of factors). However, this analysis hides the NULL values, so that every record has its values to compute.

This visualization shows the top 3 accommodations that have the lowest average review factors. It would be useful for the organization to know the quality of their provided accommodation in detail for future improvement. Organizations can use this information to take action for any changes to the accommodation owner so that they would increase a higher chance of providing a good service to the user.

Business Requirements: Booking and Reservation, Accommodation

Fact Table and Measures: Reservation Fact Table (amount_of_availability)

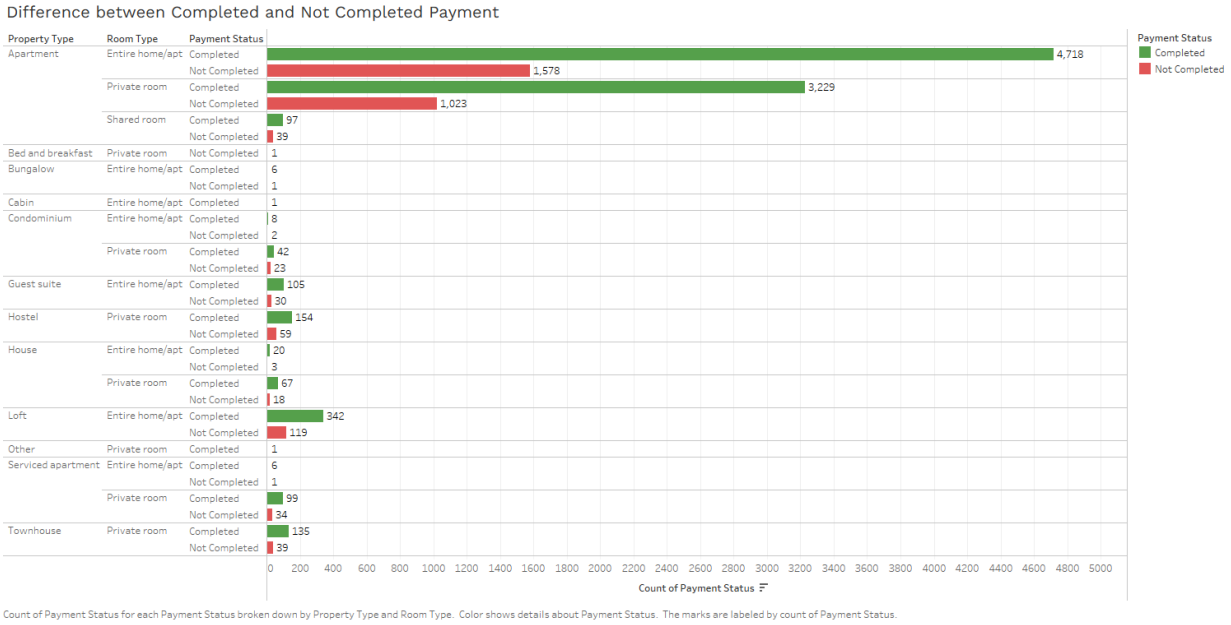
Visualization 6: Number of Available and Unavailable Accommodation in 2016-2022

Visualization 6 illustrates the number of available and unavailable accommodations from the Reservation Fact Table. It appears that in 2021, the number of both available and unavailable accommodations significantly dropped. It can show the difference between both attributes between years, and see the trend in bookings. If the availability is greater than that, then it might give a sign to the organization to announce some promotions or campaigns for the customer to use their services, however, it is not always the case since the amount of accommodation might be increased and not equal compared to the previous year.

This visualization shows the number of available and unavailable accommodations between 2016 and 2022. It has an amount of difference between years to show if the number has a big difference or not compared to the previous year. It would be useful for organizations to know the number of available and unavailable accommodations as they can provide some additional factors to increase the unavailability of rooms and provide enough rooms for users to book in the future.

Business Requirements: Booking and Reservation, Accommodation

Fact Table and Measures: Revenue Fact Table (payment_units_completed, payment_units_failed)



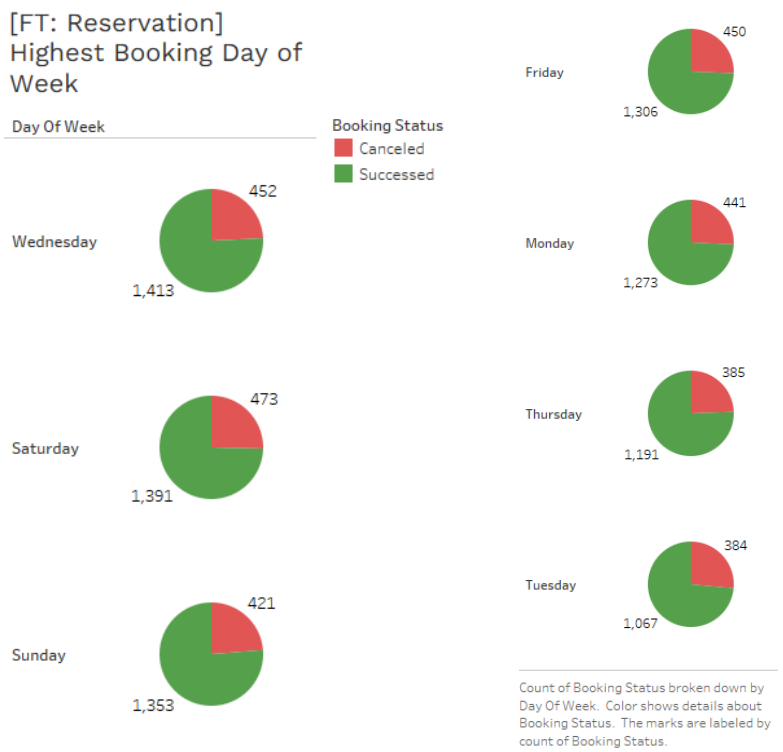
Visualization 7: Difference between Completed and Not Completed Payment

Visualization 7 illustrates the completed and failed payment status units from the Revenue Fact Table. The payment is categorized by the property type and room type respectively. It appears that Apartment that has an entire home/apartment and a private room has a significant amount of payments. Most of the records have a completed payment status which is a good sign for the organization.

This visualization shows the amount of completed and not completed payments, it would be useful for an organization to know this information. First, they can improve their payment services if the not completed payment status is continually increasing. If the same accommodation types have the same common mistakes (high rate of not completed payment), then it might give information for the organization to analyze the root cause of the problem and fix it. Second, they can see which types of accommodation (property type and room type) the user prefers.

Business Requirements: Booking and Reservation, Accommodation

Fact Table and Measures: Reservation Fact Table (highest_booking_week_day)



Visualization 8: Highest Booking Day of Week

(note: image is cropped to 2 pictures to fit the page)

Visualization 8 illustrates the highest booking day of the week from the Reservation Fact Table. It appears that Wednesday has the highest booking, followed by Saturday and Sunday respectively. Moreover, it has two color masking (Succeed and Canceled) to show the proportion of payment status. It shows that most of the weekday has a similar amount of payment status.

This visualization shows the highest booking day of the week. It would be useful for the organization to see the customer behavior of which days the customer prefers the most so that they would announce the promotion on the most customers' favorable day and it would help increase the organization's revenue and stand out from their competitors.

Discussion and Conclusion

Discussion

The outputs of the project are Entity Relationship Diagram, Star Schema, Data dictionary, Tableau ETL flow, Cleaned dataset, and Analysis and Visualizations in Tableau Desktop. The analysis shows the importance of the data warehouse and BI to Airbnb. For example, Airbnb can understand some user behaviors or trends by interpreting the Highest Accommodation Successful Rate of Bookings (Visualization 3), the Total of User Stay days classified by Property and Room Type (Visualization 4), the Top 3 Accommodation IDs with the lowest average 6 review factors (visualization 5), and Highest booking day of the week (Visualization 8). Those visualizations can determine the characteristics of good and bad accommodation, and which proprietor and room type are in trend at a time. As a result, Airbnb can effectively implement the search algorithm to map the guests and the hosts, which results in better user experiences. Unfortunately, the dataset has a limitation in that the dataset doesn't have all the information required, so some of them are generated randomly based on the hypothesis such as a successful booking should be more than a canceled booking, and it will generate incorrect insight from the analysis. Moreover, the dataset used in the case study is sampled from the full dataset, so many accommodations are not included in the analysis. Although the analysis might not be useful in the real world, it shows how the analysis can be beneficial to Airbnb. Since Airbnb needs to attract users consisting of Hosts and Guests, it is crucial for the company to understand the users and implement functionalities to improve the user experience that makes them stick to the platform and create strategies to deal with unexpected situations like the COVID-19 pandemic.

Conclusion

In conclusion, Airbnb is in the hotel and accommodation sector, which is an important part of the tourism industry. It is a marketplace that maps hosts and guests together and mainly gets the revenue from the transaction fee. As a result, Airbnb needs to understand user behavior and trends to create strategies, policies, and features to improve the user experience and attract users to use the platform. The study found that Airbnb uses a data warehouse and Business Intelligence to analyze and visualize the data on a daily basis, which helps them to understand the data easily and accurately. In the data collection step, the team observes the Airbnb platform to extract the business requirements to create schema designs which are Entity Relationship diagram and Star Schema in the collection step. Moreover, The dataset is

from the data.world, which is a website that provides open datasets and is used as a case study with Google Sheets, Tableau Prep, and Tableau Desktop. The processing step is mainly about joining and cleaning the data with Tableau Prep by removing unwanted columns, changing data types, replacing null values with valid values, and renaming the attribute names. In the analysis and visualization step, the Tableau Desktop is used to create the visualizations from the clean dataset. There are 8 visualizations created to analyze the trends and user behaviors to satisfy the business requirements based on the facts designed in the Star Schema. For future research, the data used to analyze should be from the real dataset, so the analysis will be correct and accurate, and use all the tuples in the dataset to make use of the full potential of data warehouse and BI.

References

[1] “Industry market research, reports, and Statistics,” *IBISWorld*. [Online]. Available: https://www.ibisworld.com/global/industry-trends/fastest-growing-industries/. [Accessed: 27-Feb-2023].

[2] Vacasa, “How do airbnb reviews work?,” *Vacasa*. [Online]. Available: https://www.vacasa.com/homeowner-guides/how-airbnb-reviews-work. [Accessed: 27-Feb-2023].

[3] “Airbnb service fees - airbnb help center,” Airbnb. [Online]. Available: https://www.airbnb.com/help/article/1857. [Accessed: 27-Feb-2023].

[4] E. Ritter, “Supercharging apache superset,” *Medium*, 24-May-2021. [Online]. Available: https://medium.com/airbnb-engineering/supercharging-apache-superset-b1a2393278bd. [Accessed: 27-Feb-2023].

[5] Airbnb, “Using data to help set your price,” *The Airbnb Blog - Belong Anywhere*, 04-Jun-2015. [Online]. Available: https://blog.atairbnb.com/using-data-to-help-set-your-price/. [Accessed: 27-Feb-2023].

[6] ProjectPro, “How data science increased Airbnb's valuation to $25.5 bn?,” *ProjectPro*, 18-Jan-2022. [Online]. Available: https://www.projectpro.io/article/how-data-science-increased-airbnbs-valuation-to-25-5-bn/199. [Accessed: 27-Feb-2023].

[7] C. Keown, “How airbnb went from pandemic loser to long-term winner,” *Barron's*, 16-Feb-2022. [Online]. Available: https://www.barrons.com/articles/airbnb-stock-price-earnings-travel-covid-51645027041. [Accessed: 28-Feb-2023].

[8] J. Menze, “People are staying longer in Airbnbs - for now,” *PhocusWire*, 20-May-2020. [Online]. Available: https://www.phocuswire.com/airbnb-long-term-stays-on-rise. [Accessed: 28-Feb-2023].

[9] M. Hines, “As work and school can be done from anywhere, Airbnb sees more monthly stays,” *USA Today*, 09-Feb-2021. [Online]. Available: https://www.usatoday.com/story/travel/hotels/2021/02/09/airbnb-rentals-monthly-stays-up-covid-travel-meets-virtual-life/4375982001/. [Accessed: 28-Feb-2023].

[10] “Airbnb revenue and Usage Statistics (2023),” *Business of Apps*, 19-Jan-2023. [Online]. Available: https://www.businessofapps.com/data/airbnb-statistics/. [Accessed: 28-Feb-2023].

[11] D. L. Yohn, “How airbnb survived the pandemic-and how you can too,” *Forbes*, 10-Nov-2020. [Online]. Available: https://www.forbes.com/sites/deniselyohn/2020/11/10/how-airbnb-survived-the-pandemic--and-how-you-can-too/. [Accessed: 28-Feb-2023].

[12] *What is ETL (extract, transform, load)?* (no date) *IBM*. Available at: https://www.ibm.com/th-en/topics/etl#:~:text=brief%20(169%20KB)-,What%20is%20ETL%3F,warehouse%20or%20other%20target%20system. (Accessed: February 28, 2023).

[13] Garlowski, I. (2021) *A Tableau Prep Primer: Its ideal use cases*, *InterWorks*. Available at: https://interworks.com/blog/2021/05/11/a-tableau-prep-primer-its-ideal-use-cases/#:~:text=Tableau%20Prep%20is%20an%20ETL,hoc%20tool%20for%20data%20transformation. (Accessed: February 28, 2023).

[14] *OLTP and OLAP: A practical comparison* (no date) *Stitch*. Available at: https://www.stitchdata.com/resources/oltp-vs-olap/ (Accessed: February 28, 2023).

Appendix

This is the shared link to the demonstration of the Tableau Prep to implement the ETL process and the Tableau Desktop to do the analysis and visualisation reports step-by-step based on the business case study.

Link: (We have provided the back up link in case the shared drive is not accessible)
https://drive.google.com/file/d/1SM7IBRQivVknfipxq9jU3bYUJmjUbEf6/view?usp=sharing

Back up link: https://youtu.be/gub8tH857Uo