MSIS 2621 | PROJECT REPORT

Group 3
Dhanashree Mane
Faliha Zikra Naushad Hussain
Fui Sim Yap
Sanya Purwar
Xiao Zhang

TABLE OF CONTENTS

Executive Summary

Airbnb has revolutionized the lodging industry and with the increase in number of properties , it has become increasingly competitive for hosts to get guests. Guests take into account many aspects when hunting for a place to stay. Customers choose an accommodation no only on the attributes that are provided, such as the physical (i.e. space, location, amenities etc.) but also on the non-physical (i.e. sociability, trustworthiness, friendliness, etc.) attributes. These attributes play a crucial role on customers' decision making, therefore, understanding what drives consumers to book an accommodation becomes vital for our company in developing strategies to compete in the market.

Airbnb introduced a new feature' Experiences' a few months ago where customers can purchase tickets to concert, events, tours, meetups etc. As a result, being the VP of Experience at Airbnb, we want to bring in new experiences to cities and to be able to understand what is the best time or best place to host an event it is important for us to understand trends from previously hosted events. We examined data from America's 5 most populated cities: San Francisco, washington DC, New york, Los Angeles and Chicago.

By analyzing the availability and pricing and reviews , we will have a better understanding on how successful the events might be , which can be used to determine the times and destinations that are more popular for bringing in new experiences. Bringing new experiences to a city not only adds value to a guest's experience but also helps hosts boost their profits. This feature will also help our business keep a competitive advantage by allowing us to expand our customer base, since we are aiming to provide more than just an accommodation service.In future we can study the correlation between listing process and the events held in the city . This information can be exploited to help promote tourism in cities which are not very popular during some seasons.

Based on our analysis, the following are some of the insights we have come up with:

1. **Which is the most popular neighborhoods in 5 cities?**
   This gives us an idea on which neighborhoods in each city would be best for bringing in new experiences. The more popular a neighborhood is, the higher the likelihood of attracting more customers, and therefore allowing Airbnb to benefit as a business.

2. **What are the revenue trends for each state?**
   This helps us understand which is the highest performing state in terms of revenue and at which time of the year are they most profitable, so as to give us an idea on what the best time is to attract customers for each state.

3. **Which is the most and least vacant periods for each state in each quarter?**
   This gives us an idea on which specific locations and in which time period that we should bring in more experiences (i.e. the least vacant, the higher the population the area has, and hence customer attraction rate would be significantly higher)

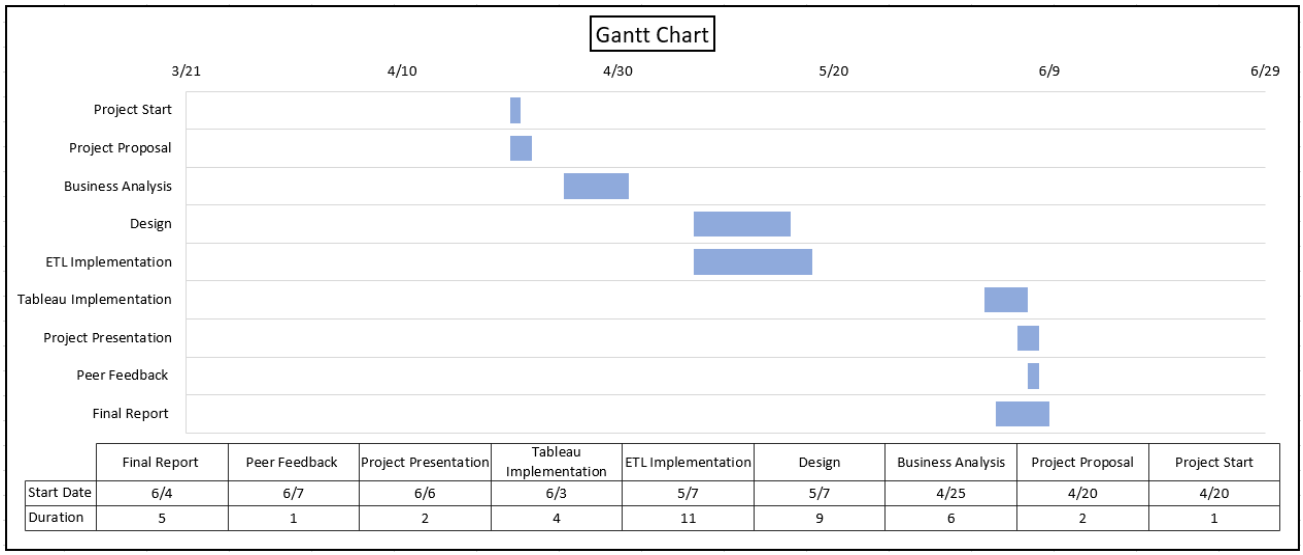
4. **Which is the most commonly chosen property type by customers in each state?**
   This allows hosts to better plan their strategy on the type of property to be rented out with respect to the states, in order to boost their profits.

5. **What is the overall listings availability throughout the year for all 5 cities?**
   This will help both Airbnb and Airdna recognize the booking trends of the Airbnb listings, and use this information to further analyze the correlation between the availability of the listings and the time of the year, and make linkages as to what might be the reason for such correlations.

Ultimately, we hope that all our insights, which include, but are not limited to, the ones mentioned above, would aid us in coming up with the appropriate business metrics that can help solve our problem statement.

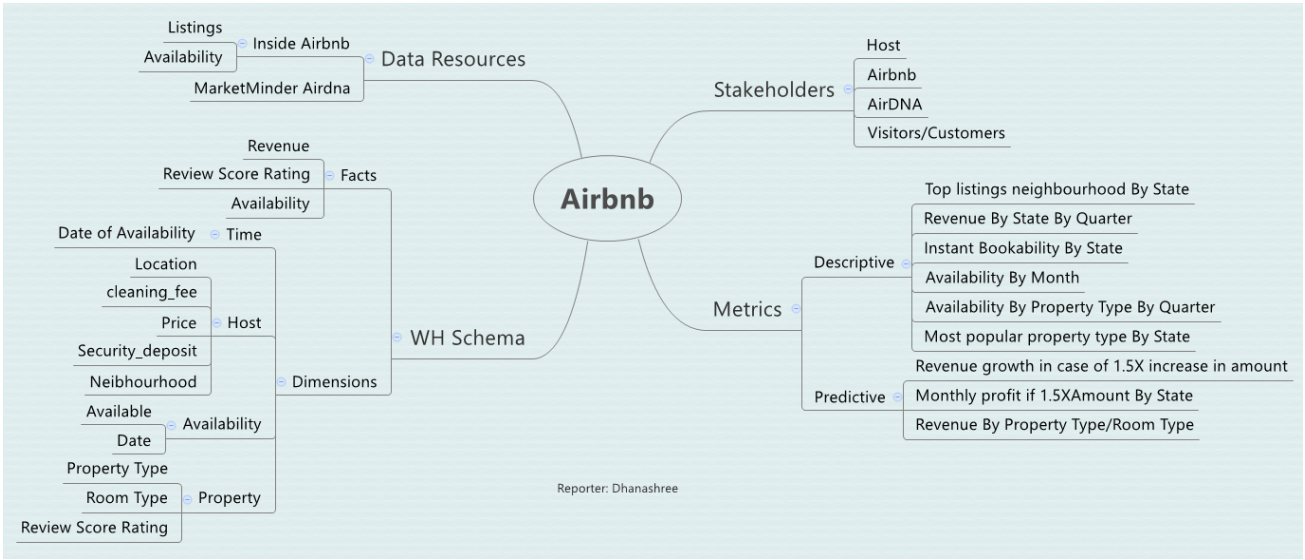
Project Description

Project Timeline



	Final Report	Peer Feedback	Project Presentation	Tableau Implementation	ETL Implementation	Design	Business Analysis	Project Proposal	Project Start
Start Date	6/4	6/7	6/6	6/3	5/7	5/7	4/25	4/20	4/20
Duration	5	1	2	4	11	9	6	2	1

Project MindMap



Domain Identification

Project Assumptions

- Business requirements and conditions remain stable after data collection is complete.
- Post data warehouse conception the scope of the project doesn't change.
- All the tasks and objective will be achieved in accordance to our predefined schedule.
- All the resources will gather new skills during the course of the project and will contribute to all aspects of the project.

Project Risks

Before initiating the implementation of our project, we have taken into consideration multiple possible risks that could occur, so that in the case where such risks occur, we are prepared ahead of time to handle mitigating the risks and avoid project failures. We have highlighted the main possible risks below:

- **Data quality**: We had to ensure that we had the correct and appropriate data that could provide us with the information that we needed and that the data was able to be cleansed and made consistent. This is so that we could avoid facing any data integration related issues.
- **Scope Creep**: We wanted to make sure that the scope of our project was realistically large enough to be handled by our team and so as to avoid the risk of scope creep.
- **Requirements gathering**: Our project should be able to meet or exceed the requirements proposed prior to the start of our project implementation. Our project's success metrics will not be met if our desired outcomes are not fulfilled.
- **Technology:** We wanted to ensure that we use the right technologies to implement our project, by taking into consideration the amount of work that is needed to be computed and the available resources we have at hand.

Challenges

- **Domain Selection and Scope Definition**: The initial task of this project was to identify which domain to focus on and what kind of business metrics it can provide that could help solve a specific set of problem statements. We had to ensure that the domain we selected had enough data that would fit the scope of our project.
- **Data Gathering and Data Cleansing**: Based on the domain that we selected, there were multiple data sources that were available, so we had to find the one that was appropriate for our project and had the right data quality that is manageable and could deliver meaningful insights that could help solve our problem statement. As the dataset contained multiple fields and a large number of rows, we had to cleanse the data by removing the unnecessary fields that were unrelated to the topic of our project and deleting redundant rows so that the data can be reduced to a reasonable size.
- **Data Modelling**: Based on what was taught in the lectures, we understood that it is vital that our data model was designed correctly. Hence, time was invested in creating our fact and dimension tables and ensuring that they were properly linked together.
- **Data Integration**: As our team was not very familiar and experienced with Pentaho, it took us quite some time to ensure that the data integration process was successful, and that all the errors that were identified in each transformation and job were fixed. We also made sure that the transformations were successful by cross-checking with our SQL servers to confirm that the correct output was shown.
- **Managing Distributed Tasks**: To ensure that our project was executed smoothly, we distributed the tasks based on each team member's skill set and availability, and constantly set deadlines to ensure that each step and process were met in accordance to our project timeline.

Information and Data Architecture

The information and data architecture below identifies the “WHAT”, “WHO”, “WHERE” and “WHY” components of our project, and is essentially a summary of how we have laid out the data in our project, and what the data means based on our analysis.

WHAT	What business processes will be supported?	Revenue Growth, Property Investments, Customer Experience Improvement
	What type of analytics to be performed?	Trend analysis, Correlation analysis, Future prediction
	What decisions to be taken?	- What is the correlation between two or more different factors? - What’s the future prediction of price in which state at what time? - What’s the overall strategy we should place based on our analysis to boost profit?
WHO	Who will benefit?	Customers, Airbnb, Stakeholders
WHERE	Where is the data now?	Excel and CSV files, with inconsistent granularities
	Where will it be cleansed and refined?	Microsoft Excel (filtered)

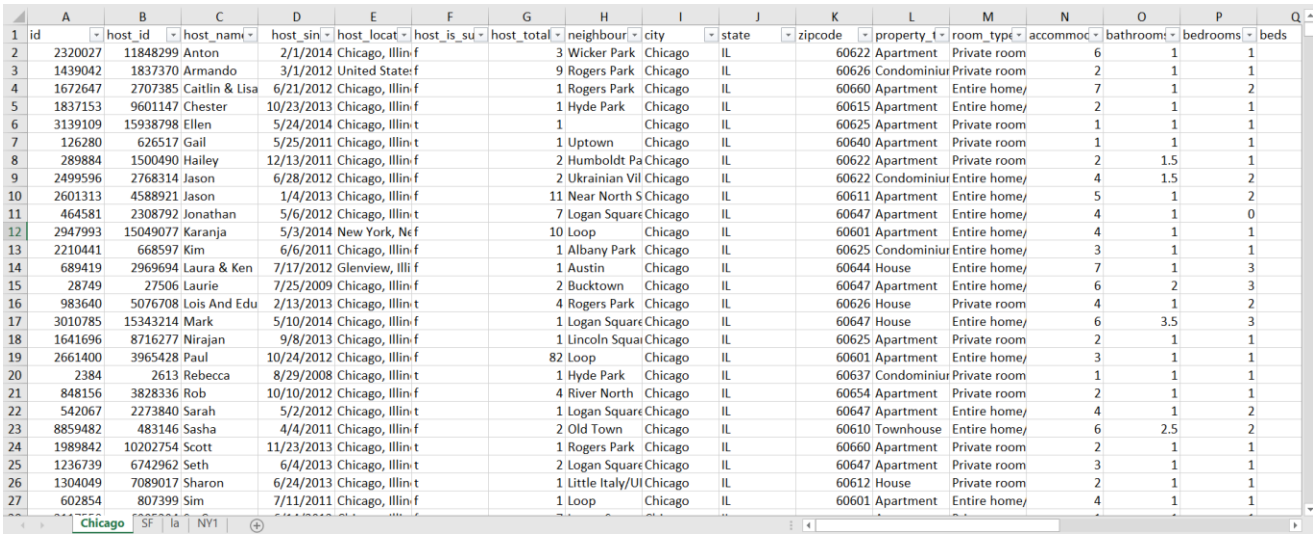
	Where will it be transformed and segregated in data marts?	Pentaho
	Where will analytics be consumed?	Tableau
WHY	Why will dashboards be built?	To provide a visual representation of our analysis in a more clear and defined way, allowing for users to draw conclusions on what the data means and hence create useful recommendations that could help drive Airbnb's business

Data Identification

Our data source comes from http://insideairbnb.com/get-the-data.html. In order to stay within our scope range, we decided to pick from the data source the datasets of the top five most populated cities (viz. Chicago, Washington, Los Angeles, New York, and San Francisco) where demand for lodging would be highest in America.

Data Cleaning

After understanding our perspective and demands, we first came up with a draft of fact & dimension tables that we needed. Based on those, we sliced out the csv data file we downloaded from 'Inside Airbnb' with only columns we need and filtering out rows to keep our data size within a reasonable scope. Below are screenshots of the steps we took and a description of each step:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	host_id	host_nam	host_sin	host_locat	host_is_su	host_total	neighbour	city	state	zipcode	property_t	room_type	accommod	bathroom	bedrooms	beds
2	2320027	11848299	Anton	2/1/2014	Chicago, Illin	f	3	Wicker Park	Chicago	IL	60622	Apartment	Private room	6	1	1	
3	1439042	1837370	Armando	3/1/2012	United State	f	9	Rogers Park	Chicago	IL	60626	Condominiur	Private room	2	1	1	
4	1672647	2707385	Caitlin & Lisa	6/21/2012	Chicago, Illin	f	1	Rogers Park	Chicago	IL	60660	Apartment	Entire home/	7	1	2	
5	1837153	9601147	Chester	10/23/2013	Chicago, Illin	f	1	Hyde Park	Chicago	IL	60615	Apartment	Entire home/	2	1	1	
6	3139109	15938798	Ellen	5/24/2014	Chicago, Illin	t	1		Chicago	IL	60625	Apartment	Private room	1	1	1	
7	126280	626517	Gail	5/25/2011	Chicago, Illin	t	1	Uptown	Chicago	IL	60640	Apartment	Private room	1	1	1	
8	289884	1500490	Hailey	12/13/2011	Chicago, Illin	f	2	Humboldt Pa	Chicago	IL	60622	Apartment	Private room	2	1.5	1	
9	2499596	2768314	Jason	6/28/2012	Chicago, Illin	f	2	Ukrainian Vil	Chicago	IL	60622	Condominiur	Entire home/	4	1.5	2	
10	2601313	4588921	Jason	1/4/2013	Chicago, Illin	f	11	Near North S	Chicago	IL	60611	Apartment	Entire home/	5	1	2	
11	464581	2308792	Jonathan	5/6/2012	Chicago, Illin	t	7	Logan Square	Chicago	IL	60647	Apartment	Entire home/	4	1	0	
12	2947993	15049077	Karanja	5/3/2014	New York, Ne	f	10	Loop	Chicago	IL	60601	Apartment	Entire home/	4	1	1	
13	2210441	668597	Kim	6/6/2011	Chicago, Illin	f	1	Albany Park	Chicago	IL	60625	Condominiur	Entire home/	3	1	1	
14	689419	2969694	Laura & Ken	7/17/2012	Glenview, Illi	f	1	Austin	Chicago	IL	60644	House	Entire home/	7	1	3	
15	28749	27506	Laurie	7/25/2009	Chicago, Illin	f	2	Bucktown	Chicago	IL	60647	Apartment	Entire home/	6	2	3	
16	983640	5076708	Lois And Edu	2/13/2013	Chicago, Illin	t	4	Rogers Park	Chicago	IL	60626	House	Private room	4	1	2	
17	3010785	15343214	Mark	5/10/2014	Chicago, Illin	f	1	Logan Square	Chicago	IL	60647	House	Entire home/	6	3.5	3	
18	1641696	8716277	Nirajan	9/8/2013	Chicago, Illin	f	1	Lincoln Squa	Chicago	IL	60625	Apartment	Private room	2	1	1	
19	2661400	3965428	Paul	10/24/2012	Chicago, Illin	f	82	Loop	Chicago	IL	60601	Apartment	Entire home/	3	1	1	
20	2384	2613	Rebecca	8/29/2008	Chicago, Illin	t	1	Hyde Park	Chicago	IL	60637	Condominiur	Private room	1	1	1	
21	848156	3828336	Rob	10/10/2012	Chicago, Illin	f	4	River North	Chicago	IL	60654	Apartment	Private room	2	1	1	
22	542067	2273840	Sarah	5/2/2012	Chicago, Illin	t	1	Logan Square	Chicago	IL	60647	Apartment	Entire home/	4	1	2	
23	8859482	483146	Sasha	4/4/2011	Chicago, Illin	f	2	Old Town	Chicago	IL	60610	Townhouse	Entire home/	6	2.5	2	
24	1989842	10202754	Scott	11/23/2013	Chicago, Illin	t	1	Rogers Park	Chicago	IL	60660	Apartment	Private room	2	1	1	
25	1236739	6742962	Seth	6/4/2013	Chicago, Illin	t	2	Logan Square	Chicago	IL	60647	Apartment	Private room	3	1	1	
26	1304049	7089017	Sharon	6/24/2013	Chicago, Illin	t	1	Little Italy/UI	Chicago	IL	60612	House	Private room	2	1	1	
27	602854	807399	Sim	7/11/2011	Chicago, Illin	f	1	Loop	Chicago	IL	60601	Apartment	Entire home/	4	1	1	

Chicago | SF | la | NY1

(Figure 1 Listing worksheet)

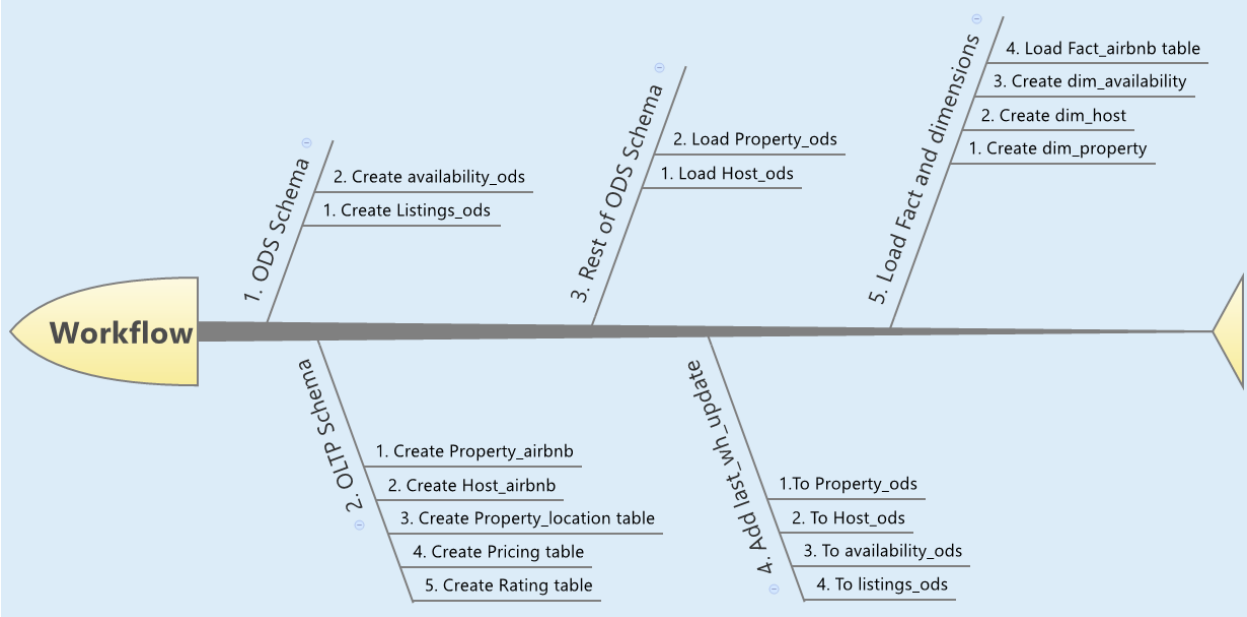
	A	B	C	D
1	listing_id	date	available	price
2	3139109	4/13/2018	1	75.00
3	3139109	4/12/2018	1	75.00
4	3139109	4/11/2018	1	75.00
5	3139109	4/10/2018	1	75.00
6	3139109	4/9/2018	1	75.00
7	3139109	4/8/2018	1	75.00
8	3139109	4/7/2018	1	75.00
9	3139109	4/6/2018	1	75.00
10	3139109	4/5/2018	1	75.00
11	3139109	4/4/2018	1	75.00
12	3139109	4/3/2018	1	75.00
13	3139109	4/2/2018	1	75.00
14	3139109	4/1/2018	1	75.00
15	3139109	3/31/2018	1	75.00
16	3139109	3/30/2018	1	75.00
17	3139109	3/29/2018	1	75.00
18	3139109	3/28/2018	1	75.00
19	3139109	3/27/2018	1	75.00
20	3139109	3/26/2018	1	75.00
21	3139109	3/25/2018	1	75.00
22	3139109	3/24/2018	1	75.00
23	3139109	3/23/2018	1	75.00
24	3139109	3/22/2018	1	75.00
25	3139109	3/21/2018	1	75.00
26	3139109	3/20/2018	1	75.00
27	3139109	3/19/2018	1	75.00
28	3139109	3/18/2018	1	75.00
29	3139109	3/17/2018	1	75.00
30	3139109	3/16/2018	1	75.00

Chicago

(Figure 2 availability sheet)

The listings sheet consists of the Airbnb listings in each of the top 5 cities that we picked, and the availability represents the availability of each listing on each day in 2018.

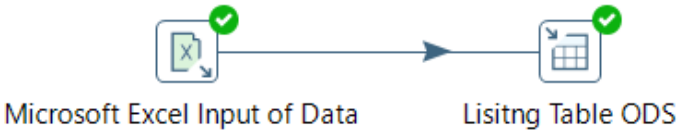
Execution Workflow



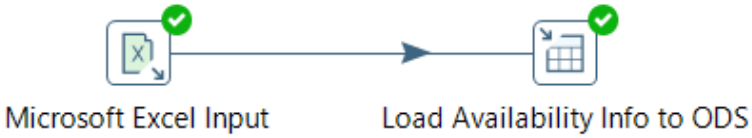
Data Transformation

ODS:

After finishing the raw data process in excel, we loaded the Listing and Availability sheets into Pentaho as our ODS tables.



(Figure 3 Listings ODS)

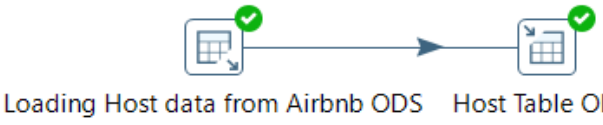


(Figure 4 Availability ODS)

OLTP:

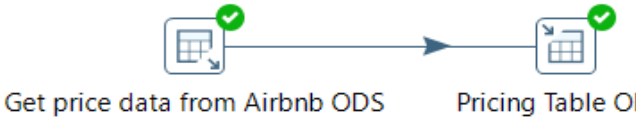
In Pentaho, we used Availability and Listing ODS tables that we previously created in the database to generate the Host_OLTP, Property_Location_OLTP, Pricing_OLTP,

Property_airbnb_OLTP & Rating_OLTP tables in our Airbnb_OLTP database. We did this through the process of reverse-engineering.

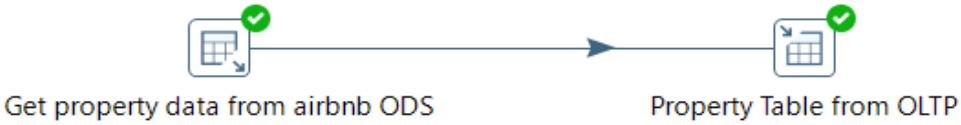


(Figure 5 Host_OLTP)

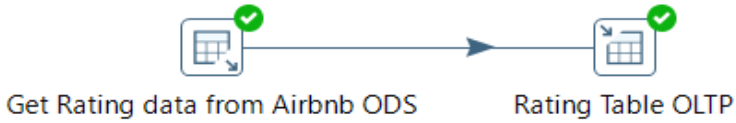
(Figure 6 Property_Location_OLTP)



(Figure 7 Pricing_OLTP)



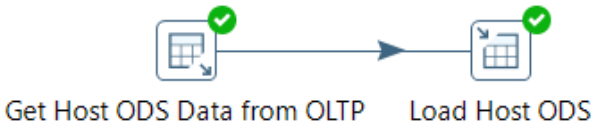
(Figure 8 Property_airbnb_OLTP)



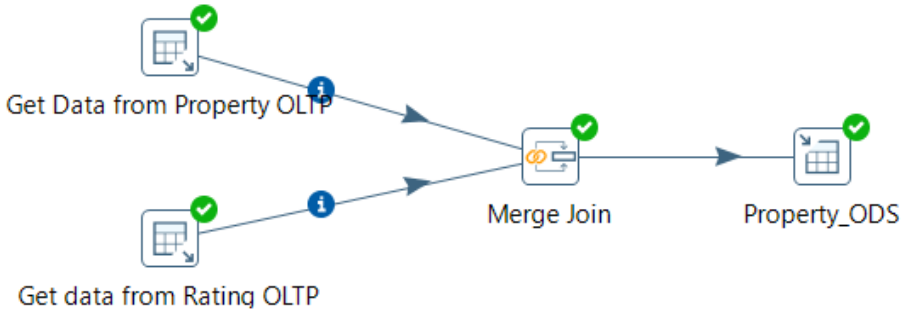
(Figure 9 Rating_OLTP)

ODS 2:

Once we created the OLTP database, we used that database to create the remaining tables in our ODS database. We then created a transformation to load Host data into Host_ODS table. The transformation for Property_ODS is a little different, because it consists of two parts including Property_OLTP and Rating_OLTP, so we merged the two tables before transforming it into the final Property_ODS table.



(Figure 10 Host_ODS)



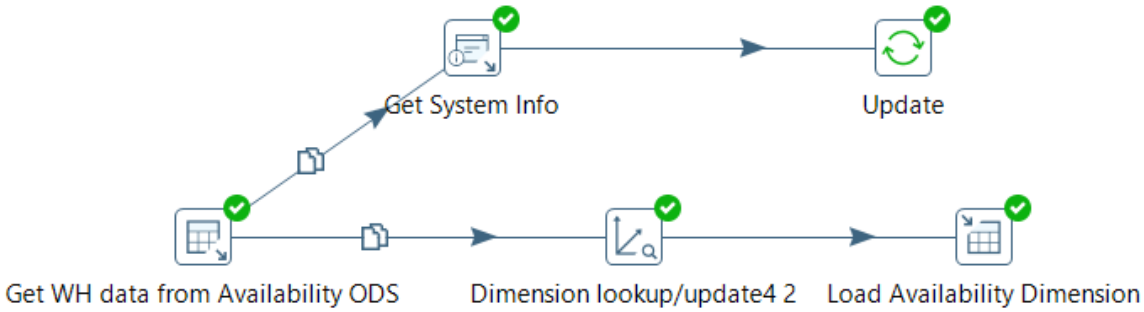
(Figure 11 Property_ODS)

Dimension Modelling

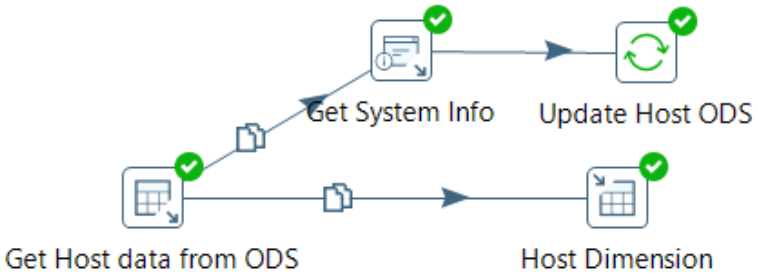
Load Fact Table and Dimensions:

To create our Fact & Dimension tables, we made a connection to the Airbnb_ODS database. We then created a transformation to load the Availability dimension. We first used table input to get the data from the Availability ODS and then a dimension lookup to create a key for each listing_id. Once we created the key, we used table output to create a dim_availability table in the database. Similarly, we created a transformation for Property and Host dimensions as well.

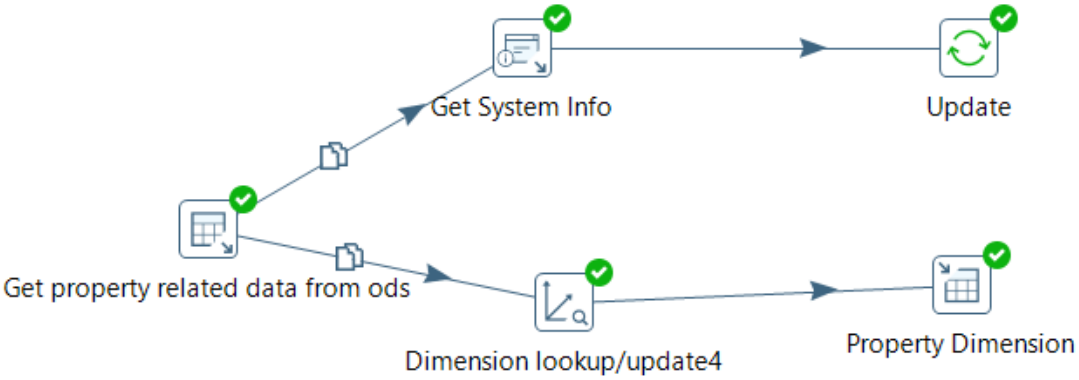
To create the fact table we got the data from the lising_ODS and used dimension lookup to lookup necessary values from the host table, property table and the availability table. Once we got all the attributes that were needed, we used calculator to calculate the total price as the amount. We calculated that by summing the price, security_deposit and cleaning_fee. Once that was done, we then used table output to create the fact table in the Airbnb_wh database.
Below are the screenshots of the Fact and & Dimension transformations:



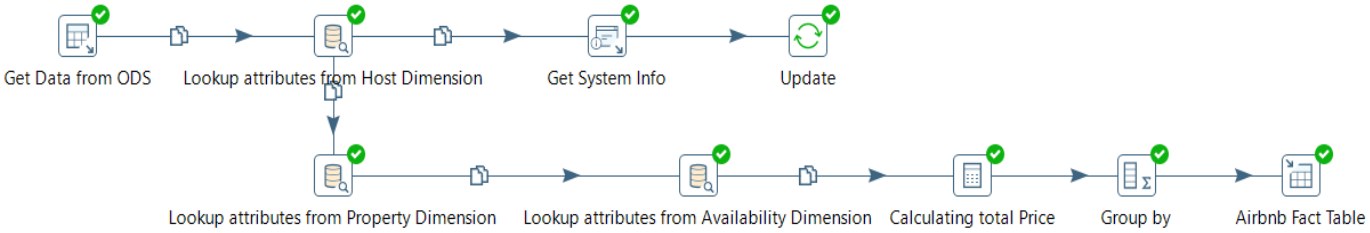
(Figure 12 Availability Dimension)



(Figure 13 Host Dimension)

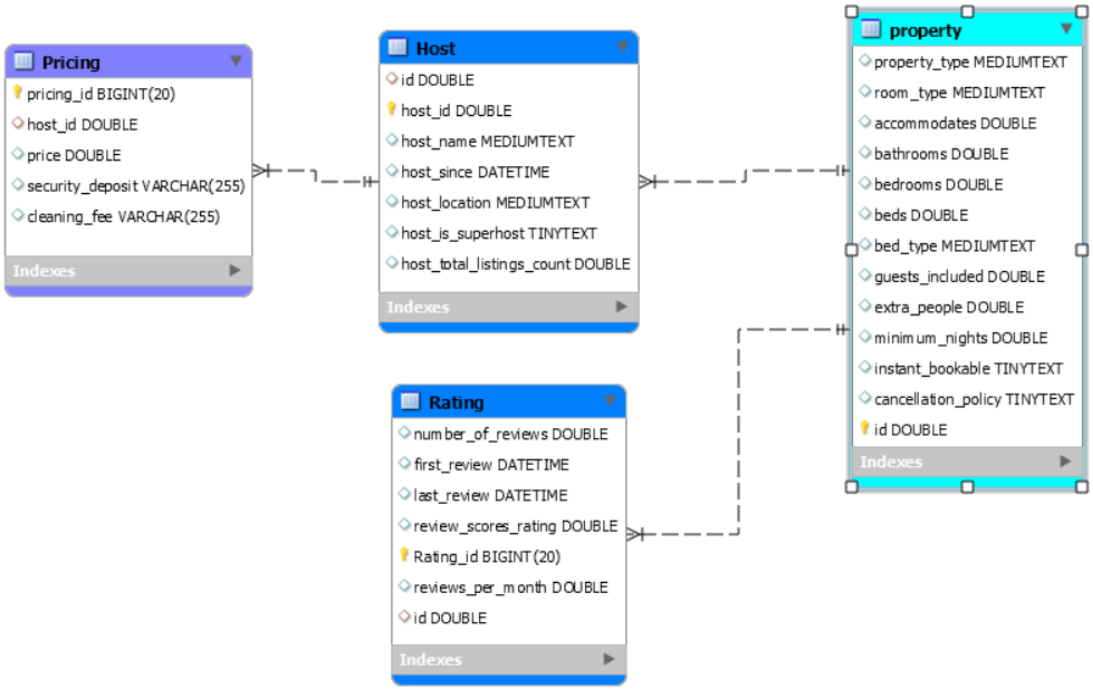


(Figure 14 Property Dimension)

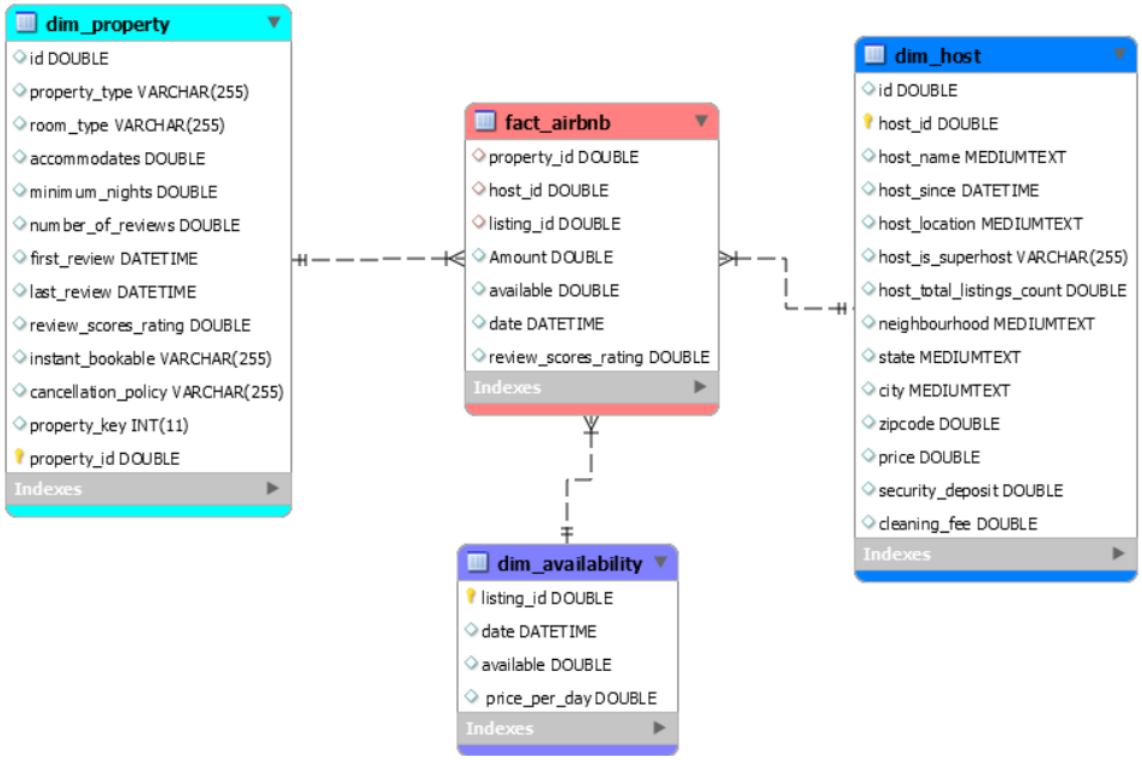


(Figure 15 Fact Table)

Online Transaction Processing Schema:

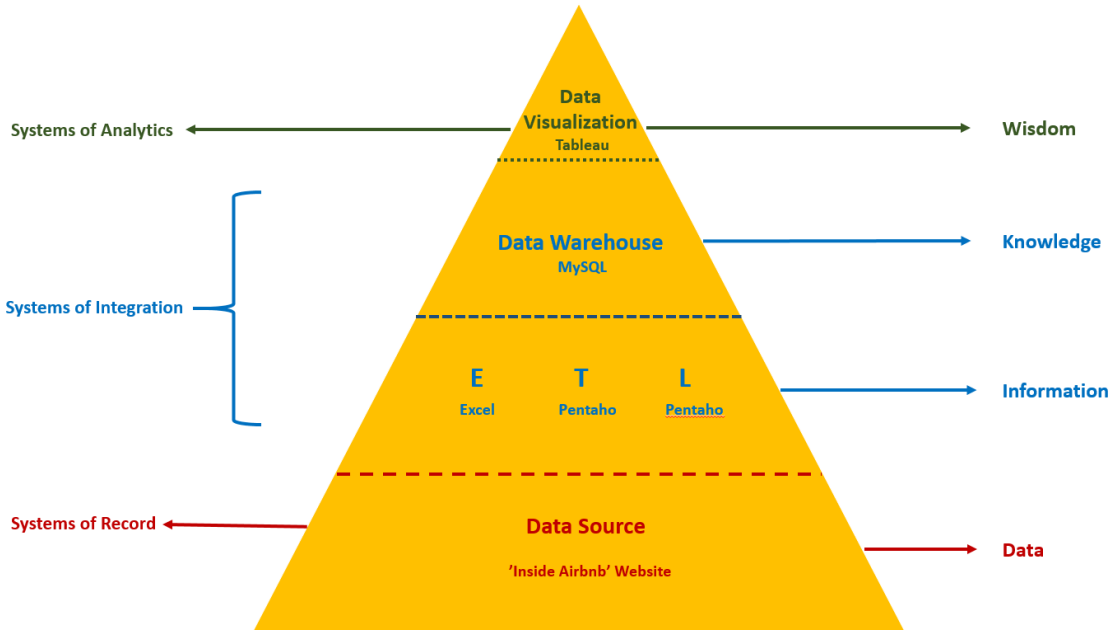


Data Warehouse Schema:



Technology Architecture

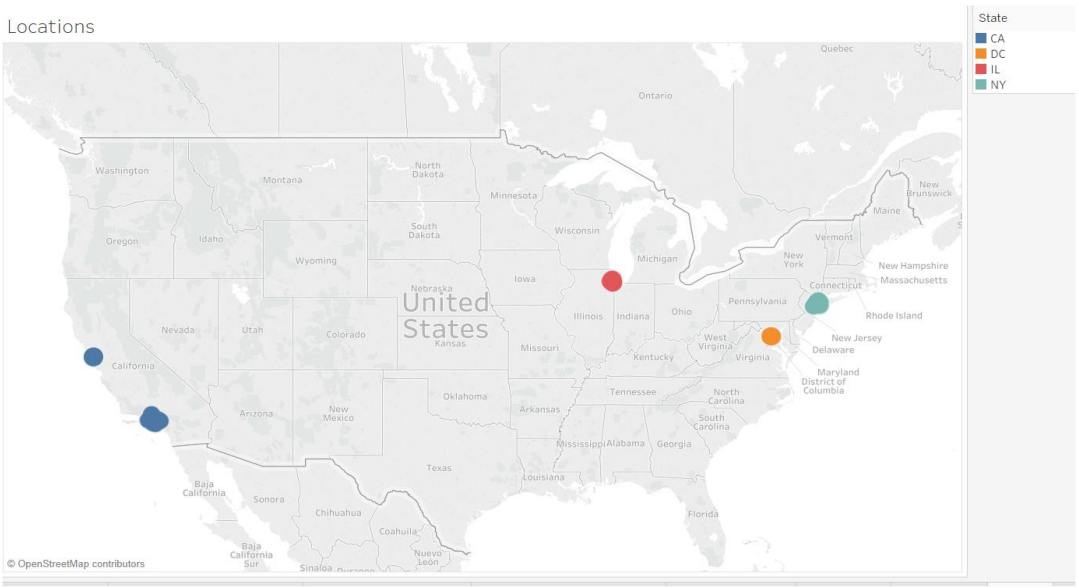
Our technology architecture from data phase to ETL phase to visualization phase can be summarized in the diagram below:



Dashboards

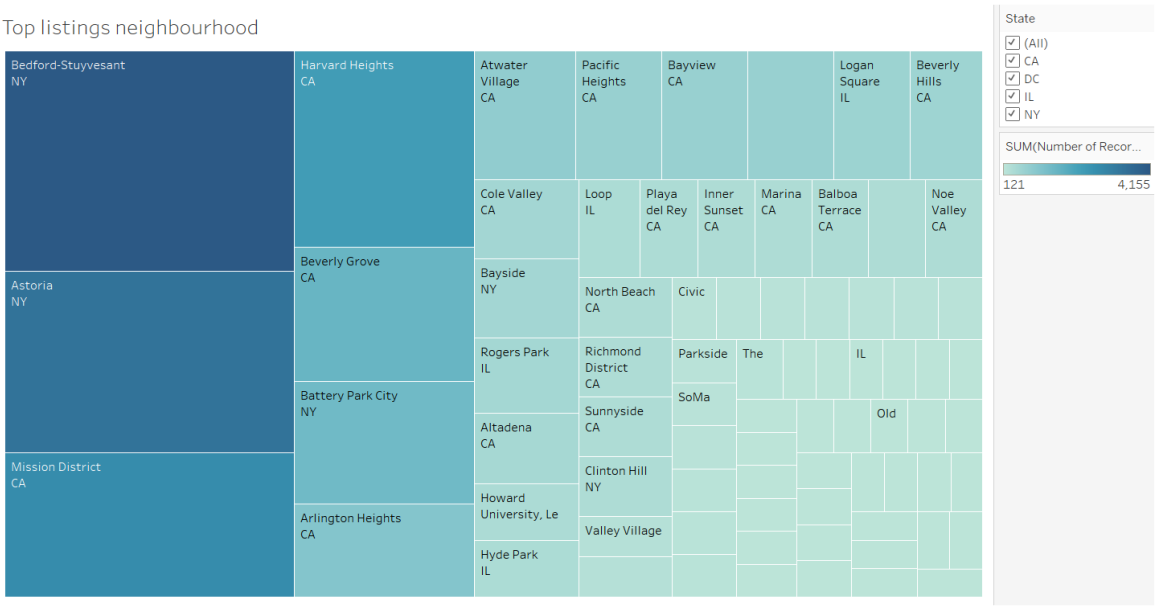
Descriptive Analysis:

1. We have done analysis for the following locations:
    a. Los Angeles, CA
    b. San Francisco, CA
    c. Chicago, IL
    d. Washington, DC
    e. New York, NY



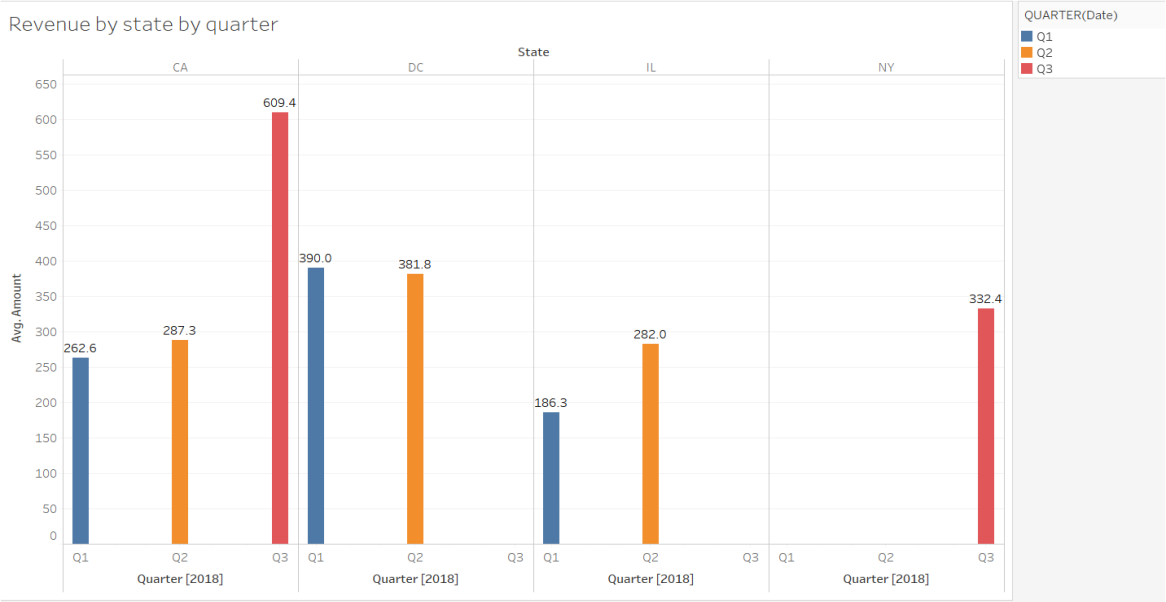
The diagram above gives us an overview of the location distributions of the 5 cities in the map of the United States.

2.



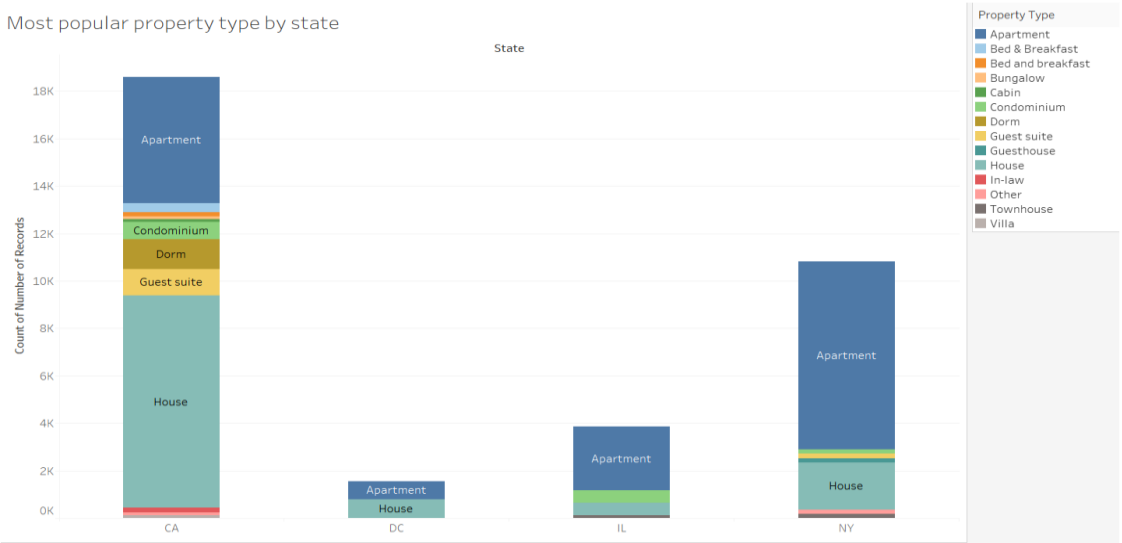
This chart lists out all the neighborhood in all the five cities in the order of most popular to least popular; the darker color being the more popular neighborhood, and vice-versa. We can see that generally the popularity ranking of the neighborhood can be laid out in the format: NY>CA>IL>DC, with Bedford-Stuyvesant NY, Astoria NY, and Mission District CA being the top 3 popular neighborhoods. This can help Airbnb in understanding which areas are more popular and hence, target those areas to bring in new experiences.

3.



As the title suggests, the dashboard shows the quarterly revenue generated from the listings in each of the states. Based on the results shown, we can identify the states that are generating the highest revenue each quarter and from that, we can analyse the changes in trend and the possible reasons for those changes. For example, we can see the in the 3rd quarter, the listings in California appears to be making the highest revenue in comparison to the other states, this could be due to the increasing amount of international tourism during the Summer vacations in California.

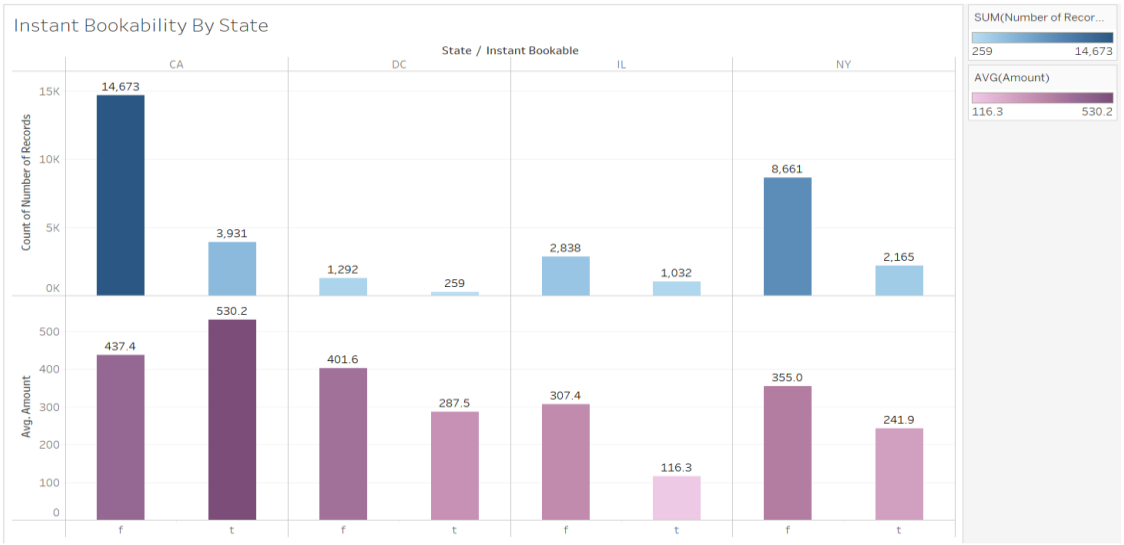
4.



In this chart, we have identified which are the most popular property types in each states (House for CA, Apartment for IL and NY). This information is important for hosts who are looking

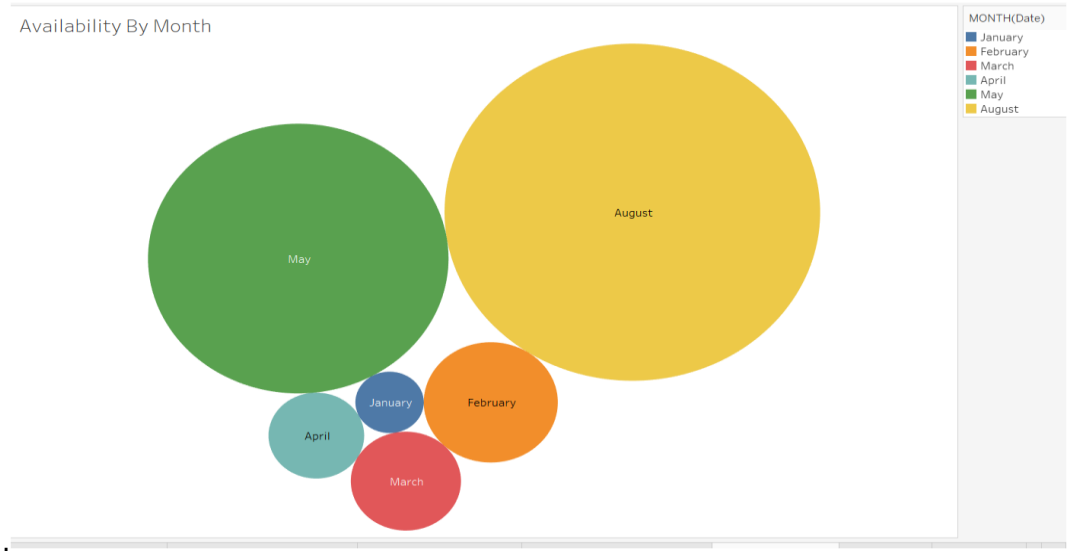
to rent out their properties on Airbnb, because this gives them an idea on which type of property to rent out that can attract more customers. By doing so, this not only benefits the hosts, but also Airbnb as this could in turn drive Airbnb's business.

5.



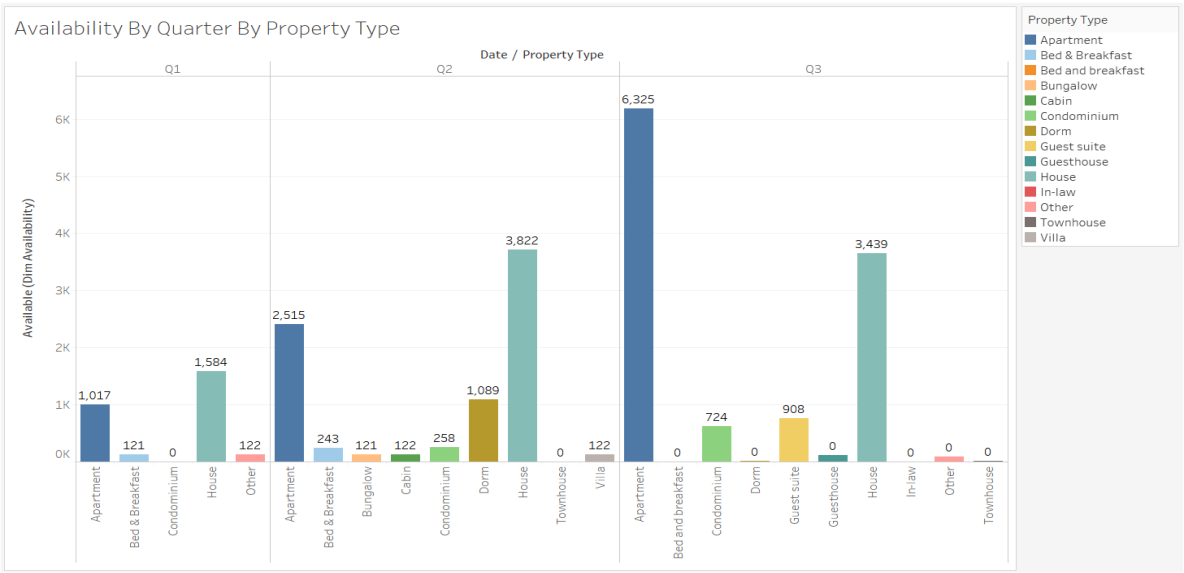
This 'Instant Bookability by State' dashboard shows the total number and average amount of instant bookings in each state. This information lets hosts know of the most common method of booking that customers are more likely to set as a filter on Airbnb. That way, hosts can decide on the type of booking policy to go for if they are looking to attract more customers. Generally the notion is people prefer instant bookable listings over waiting for approval. But as per this chart, we can see that maximum revenue is generated from 'wait for approval' type of listings.

6.



This chart shows the months where most listings are available in all of the cities, the largest circle being the the month with the most listing availability and the smallest circle being the opposite. Clearly, we can see that August, May and February are the top 3 months with the most listing availabilities, which are the months of the Summer, Spring and Winter breaks, where travelling rates are high.

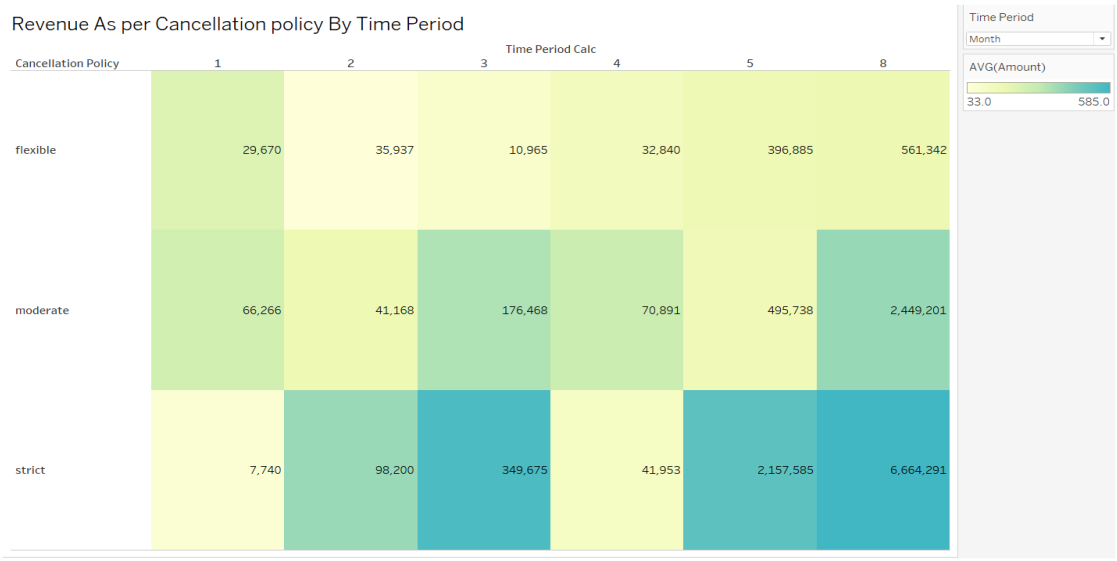
7.

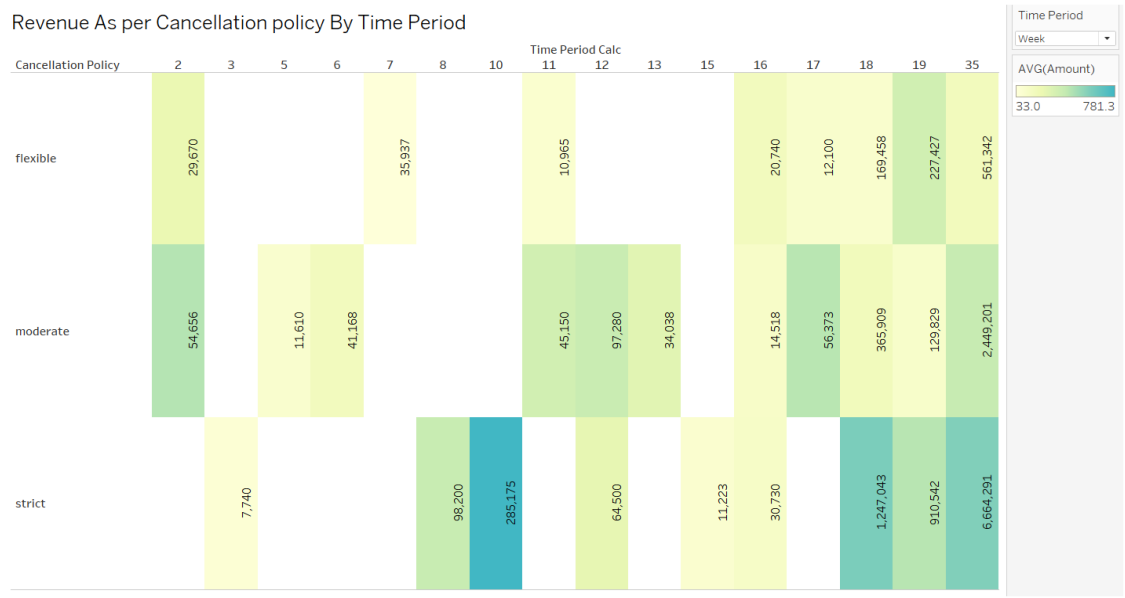


In order to determine the best time to rent out a certain type of property, hosts need to understand the demands for each property type throughout the year. This ‘Availability by Quarter by Property Type’ dashboard gives host exactly the information they need in this case. Based on the result, we can see that for every quarter, Apartments and Houses are easily the top 2 choices.

Predictive Analysis:

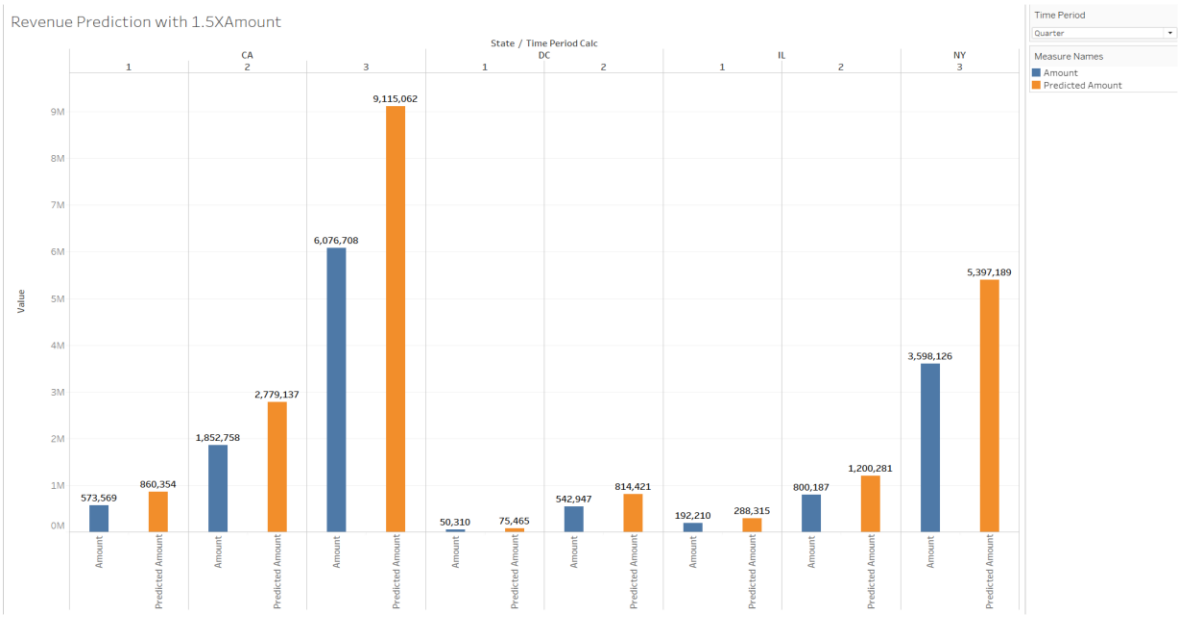
1.





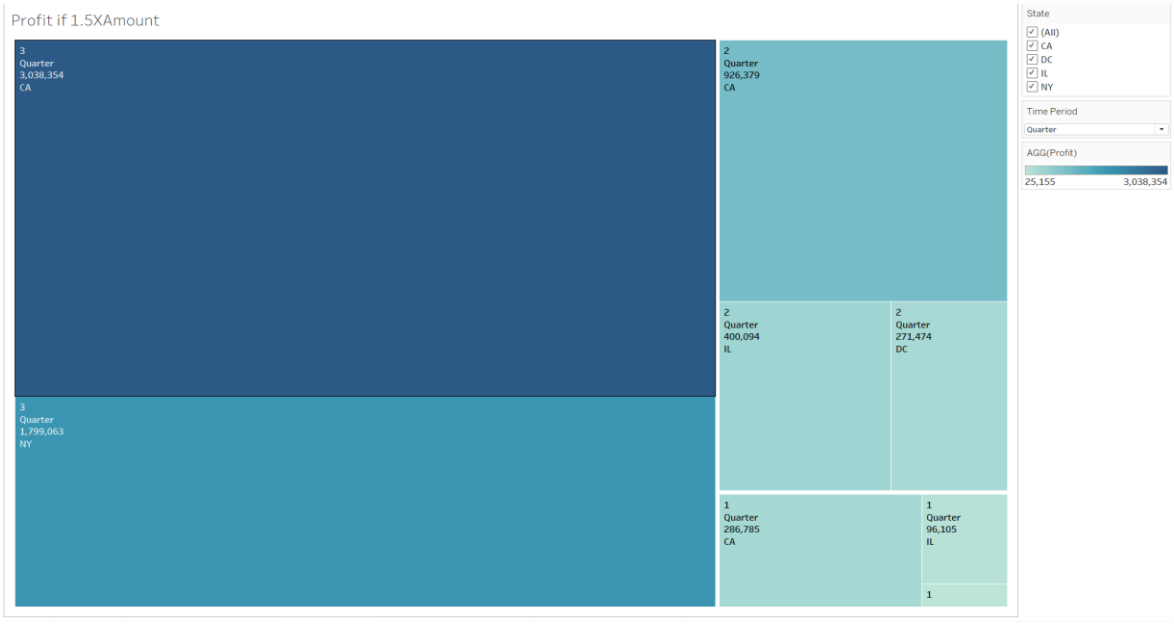
This ‘Revenue as per cancellation policy by time period’ chart shows the revenue result based on different cancellation policies during different time periods. From this, we can see that the cancellation type generates the most revenue in month 8 (first picture). By changing different time periods, we can track the revenue changes in terms of days, weeks, months or quarters.

2.



This ‘Revenue Prediction with 1.5 X Amount’ dashboard shows the difference between the amount and the predictive amount; by increasing the amount by 1.5, we can see how much profit will be generated in each state according to the time period chosen.

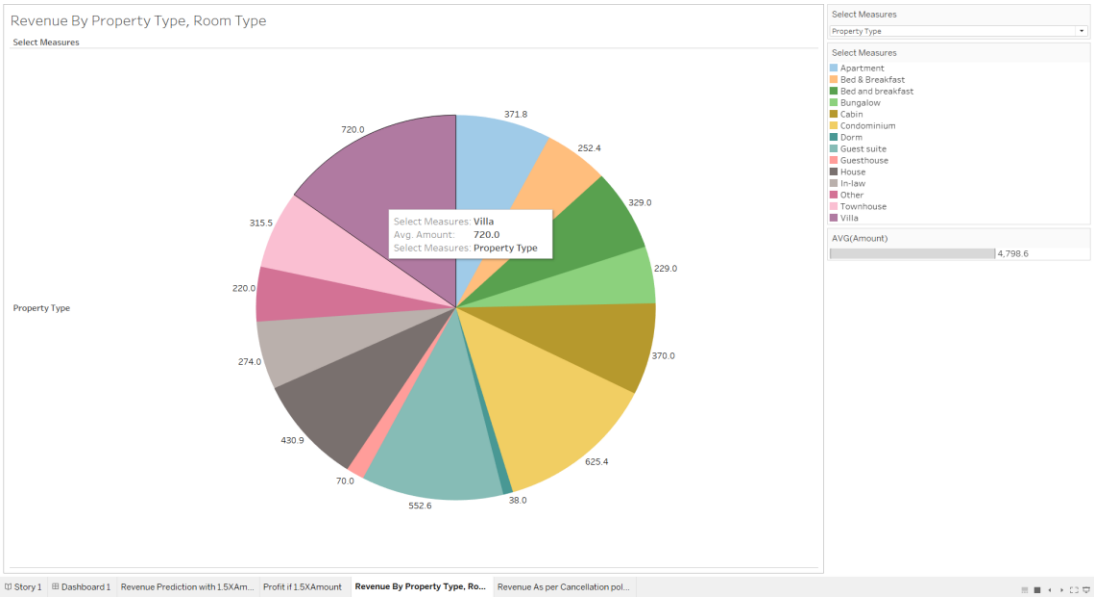
3.

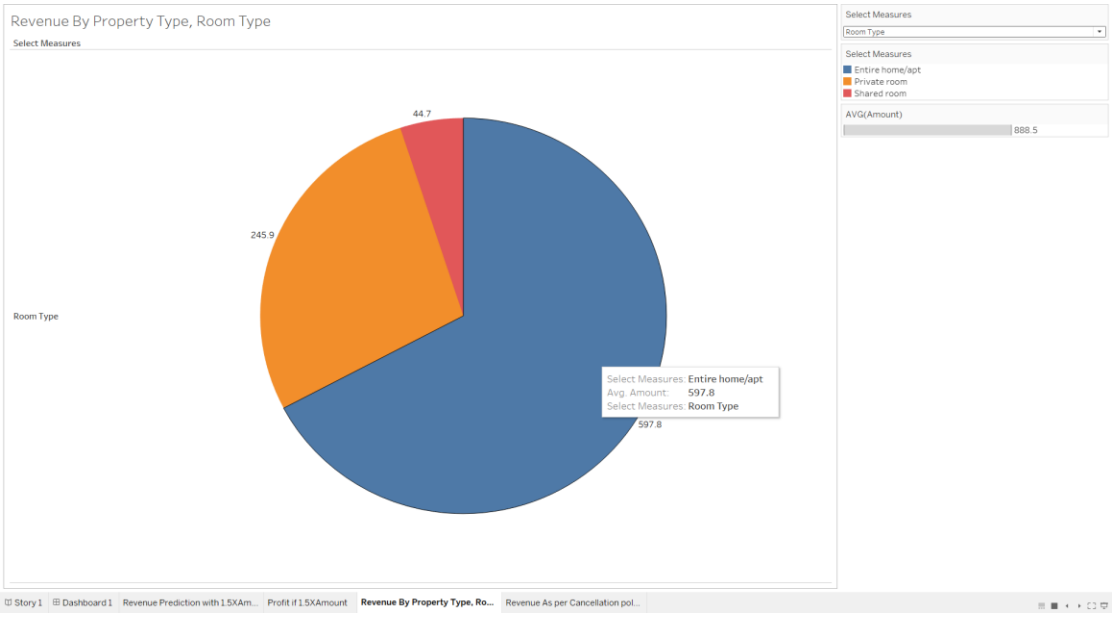


Similarly, in this ‘Profit if 1.5 X Amount’ chart, it shows how much profit will be achieved by increasing amount by 1.5 times in each state during the specified quarter, in the order from

highest profit to the lowest profit (the darker colour being the higher profit). Based on the results, we can see that CA has clearly the highest profits.

5.





This ‘Revenue by Property Type and Room Type’ dashboard takes into account the revenue generated based on 2 different factors: the Property Type, and the Room Type. By analysing this chart, hosts can determine the type of room type and property type to rent out, in order to earn maximum profits. Hosts can filter out specific select types of room type or property type by changing the measures to fit the host’s preference.

Key Learnings

We have highlighted in the following the key learnings that this project had provided us with:

- We learnt that before processing the raw data acquired from any open source, it is vital that we have a clear understanding of what our business demands are, therefore it was very important that we select the right domain to fit the scope of our project.
- This project provided us with a lot of exposure in handling data the correct way, by giving us a great deal of practice in:
    - Data modelling (i.e. creating fact and dimension tables)
    - Data integration through Pentaho
    - Creating dashboards, visualizations, by setting parameters and filters through Tableau.
- We also learnt the importance of time management in a project - by assigning tasks according to each member's skill set and availability, as well as setting deadlines for each task to be accomplished, we were able to complete the project and meet the requirements according to our predefined schedule.
- Overall, this project was a great opportunity for us to apply the knowledge and tools that we have learnt in class to a real life issue.

References

[1]: https://theislandnow.com/news-98/new-hyde-park-cracks-short-term-rentals-like-airbnb/
[2]: http://www.databaseanswers.org/data_models/
[3]: http://www.bpmn.org/
[4]: http://insideairbnb.com/get-the-data.html.
[5]: https://www.airdna.co/market-data/app/us/california/san-francisco/overview
[6]: MSIS 2621 Course handouts and class notes