**Executive Summary**

Airbnb is a community-based, two-sided online marketplace that facilitates and connects people who want to rent out their home with people who are looking for lodging in a specific location around the world. On the Airbnb platform, there are 2.9 million hosts worldwide and more than 7 million listings in over 200 countries, with more than 150 million users. This means that their biggest assets are the individuals and the hosts, where their revenue is from: host commission and guest transaction fee.

**Introduction & Business Problem**

We selected the Airbnb data because we were inspired by the database that we did in class and we wanted to explore further. From the business model Airbnb need to focus on their guests and hosts. We came up with important query questions that we believe would be beneficial for Airbnb’s:

| * How does pricing differ by region and over time? | * What is the availability by region and over time? |
| --- | --- |
| * How long is the booking time by region? | * What are the popular type of listings? |
| * How does the number of listings vary for each region over time? | * What are the trends and seasonality of the listing in each region? |

**Data Exploration**

Our data used is comprise of data from six states from the Western United States, totaling 2.47 GB of data. These six states were California, Colorado, Hawaii, Nevada, Oregon, and Washington, and across these states we had data from 14 cities in total. The original Kaggle posting where we found our data contained data on more states, but we settled on these western states for feasibility and to analyze listings in the region where Airbnb was founded. Finally, the data contained four tables for each city in the dataset: listings, a table containing information about each unique listing in the city; calendar, a table containing a listing’s availability and price on different dates; reviews, a table containing reviews for each listing; and neighborhoods, a table containing information on the neighborhoods within a city. For our analysis, we looked to explore variations in price and availability across cities, so our final dataset contained only the listings and calendar tables.

We used pySpark to load and query the data, as this proved to be the quickest and most efficient process for working with this size of data. Compared to using Hive or Pig, pySpark offered more complex queries and connection to SQL, giving us a familiar environment to work with. We uploaded the necessary table to an S3 bucket, read them into pySpark individually, then unioned them all together using pySpark’s unionByName command. We used this process, instead of reading all files at once, to account for the possibility of tables having differing columns and schema. At the end of the wrangling phase, we had 2 tables: listings\_union and calendar\_union, each comprised of their respective table from each of the 14 cities. In the preliminary stages of data wrangling, we experienced issues with pySpark recognizing the difference between column values that contained commas and commas that split each column. We solved this by specifying quotes as the escape character when reading in the data and that this escape character could carry over to multiple lines. After final data wrangling, our data contained 118,107 observations in the listings\_union table, with an average yearly availability of 175 days per year and average monthly availability of 8.7 days per month. The calendars\_union table contained 43,100,840 observations with an average price of $230.52 per night.

**Data Querying**

We utilized pySpark’s powerful querying ability to explore pricing and availability across western cities. First, we looked at the cities with the highest average price per night, and found that Malibu, California was the highest at $997.86. Malibu was followed by South Kohala, Hawaii and Los Angeles, California with rates of $961 and $950 respectively. A screenshot of this query, and all our queries, is shown in the appendix. Next, we looked at areas with the highest average monthly and yearly availability, finding that some areas have availability everyday day of each month and year. Among these cities were Los Angeles, California, San Pedro, California, and Silverado, California. Next, we queried to find the area-year-month combination with the highest average price to find during which times of the year prices spike. We found that Kauai, Hawaii in March 2023 has the highest average price of $998.24 per night, followed by Malibu, California in several months. Finally, we queried to find the number of listings by area and month to further identify how availability varies over the year. This query revealed that regardless of the month, Los Angeles, California has the most listings out of the western cities in our data.

**Data Visualization**

After having explored the dataset and answering our business questions on a city level using basic Exploratory Data Analysis on pySpark we decided to do the same but this time visually and our main focus was on a state level to see how do the states compare with respect to parameters like pricing, availability and the popularity of listings to name a few. For this purpose, we chose Tableau as our primary tool for Data Visualisations because of the flexibility and efficiency it provides us in regards to plotting visualisations.

For the Data Visualisation process we chose a bunch of different graphs that were able to capture the insights we were looking for. We chose to plot some bar graphs that gave us an idea about the distribution of pricing, availability and total listings across states. We also plotted some charts like Map and Treemap to compare the magnitude of the difference in listing frequency across states. Furthemore, one of our major reasons for plotting the results we got from our pySpark queries was the fact that it was not the ideal option to get insights into questions that involve time series analysis whereas graphs are the best way to represent a change over time taking into account multiple factors. We wanted to see the change in pricing and availability of the listings in regards to months/seasons so we can check if there is any correlation between the aforementioned factors with the time of the year.

While doing so, we ran into a couple of challenges both were primarily related to the size of the dataset. Firstly, It was quite a hassle to upload the datasets onto tableau and join them because of the size of the resultant dataset which was huge with close to 20 million observations. The data loading and joining took some time. Secondly, the size of the dataset also kind of limited us in the experiments we could do with our visualisations. Plotting graphs that involved disaggregated measures was not possible.

**Insight and Analytics**

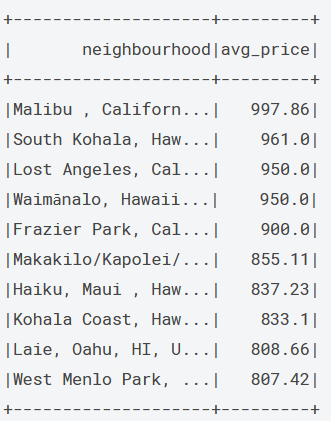
From the queries and visualisations, we focused on the listing, pricing and availability of each host’s properties. We see the state of California have the most listing state wise at 51,929 listings, while Oregon have the least listing state wise at 3,517 listings. For pricing we see that Hawaii have the highest average price of $461.1, while Oregon have the lowest average price of $150.4. We also see that most state show a constant average price over the year and Nevada have the most variation in price, with peaks in March to May and in October. For availability, Nevada have the most monthly availability with 18.145 day per listings available on average, while Washington have the least monthly availability with 4.841 days per listings available on average. Other states average monthly availability is between 5 to 6 days per listings available on average. Looking at the seasonality of booking, we see spike in November and May and a decrease in December and ther is a general increase after June across the states, while the beginning of the year seems to be stable.

Utilizing the queries and visulizations we were able to gain insights into pricing, listing popularity, availability, and location-specific details for Airbnb to use to maximize their business reach.

**Business Application and Future Exploration**

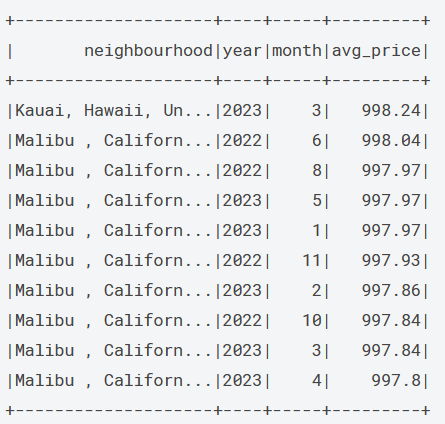
Overall, by understanding Airbnb business model, revenue stream and the insight collected we were able to discover different trends and information regarding the hosts to answer our business questions and generate business strategy. We were able to help Airbnb determine how to generate more revenue through promoting popular listings and leveraging targeted advertising in different cities during peaks and dips of the vacation seasons.

To explore this project further, we can utilized other region data to create comparison and explore trends in different region. We also would like to do sentiment analysis on the reviews data to help hosts improve their services and give personalized feedback to them. This data when combine with the rating data we would be able to accurately explore popularity of the Airbnb. For a more advance project we are looking to build clustering models to explore listing’s characteristic.

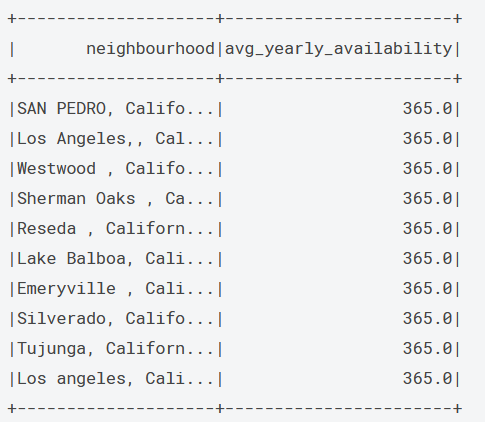
**Appendix A: Data Querying**

**How does pricing different in different cities and states?**

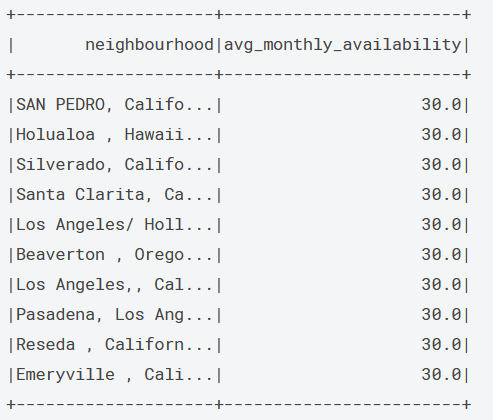
* spark.sql("SELECT l.neighbourhood, ROUND(AVG(c.price), 2) avg\_price FROM listings l LEFT JOIN calendar c on l.id = c.listing\_id GROUP BY l.neighbourhood ORDER BY avg\_price DESC")

**How does pricing differ annually and monthly in different cities and states?**

* spark.sql("SELECT neighbourhood, year(to\_date(c.date)) year, month(to\_date(c.date)) month, ROUND(Avg(c.price), 2) avg\_price FROM listings l LEFT JOIN calendar c on l.id = c.listing\_id WHERE neighbourhood IS NOT NULL GROUP BY l.neighbourhood, year, month ORDER BY avg\_price DESC")

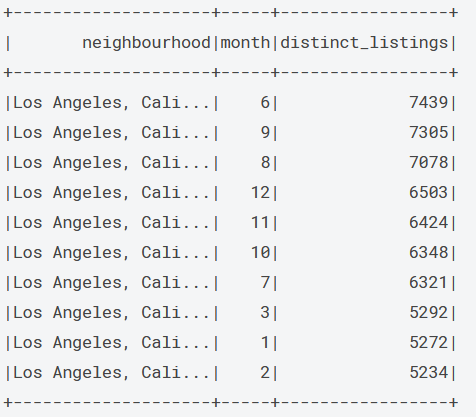
**What is the annual availability by each cities and states?**

* spark.sql("SELECT l.neighbourhood, ROUND(AVG(l.availability\_365), 0) avg\_yearly\_availability FROM listings l LEFT JOIN calendar c on l.id = c.listing\_id GROUP BY l.neighbourhood ORDER BY avg\_yearly\_availability DESC")

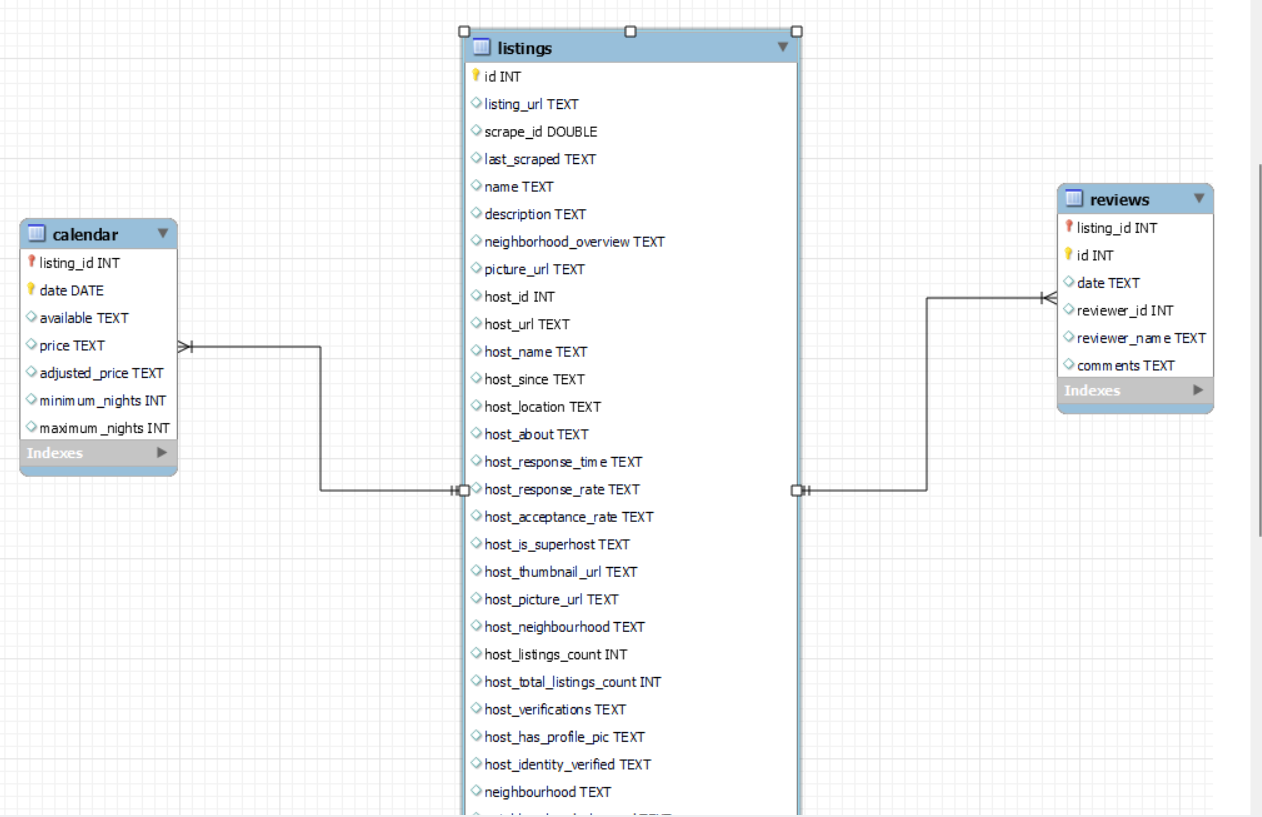
**What is the monthly availability by each cities and states?**

* spark.sql("SELECT l.neighbourhood, ROUND(AVG(availability\_30), 0) avg\_monthly\_availability FROM listings l LEFT JOIN calendar c on l.id = c.listing\_id GROUP BY l.neighbourhood ORDER BY avg\_monthly\_availability DESC")

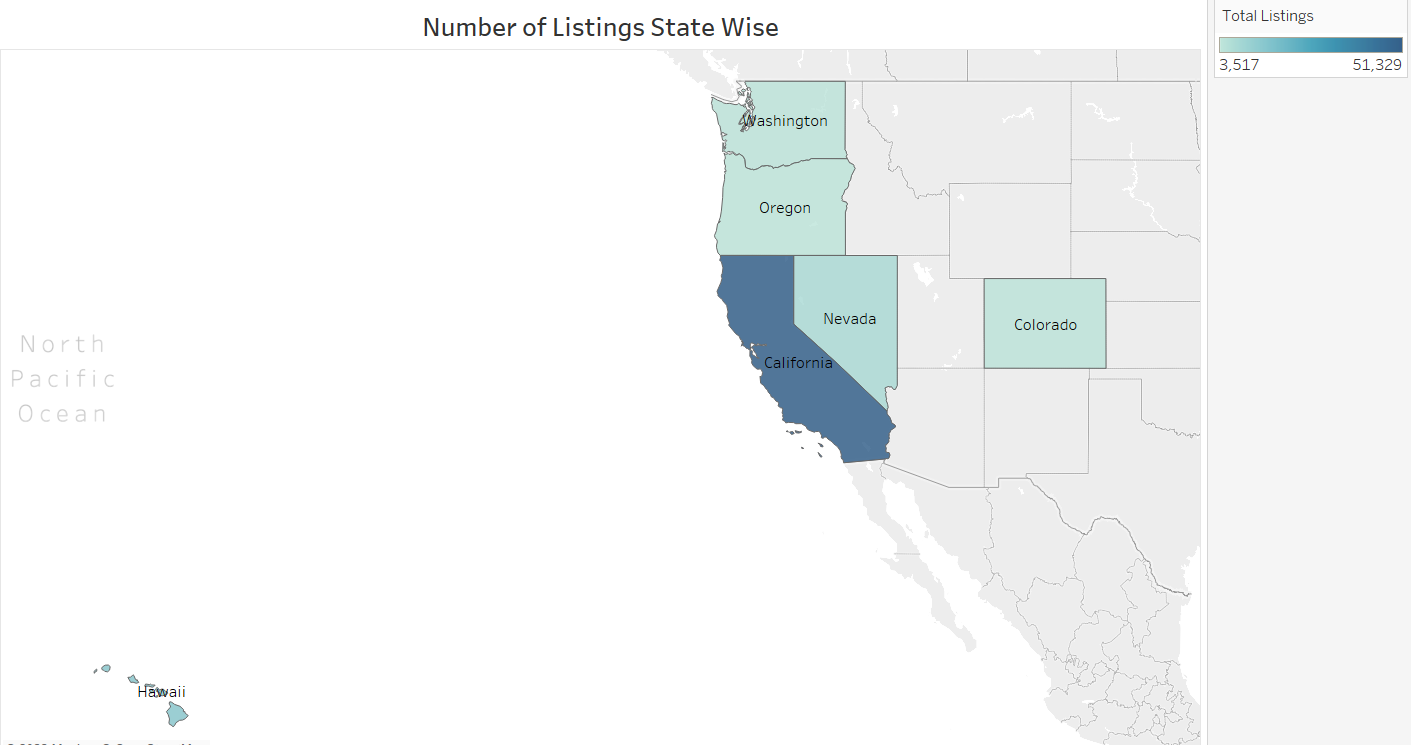
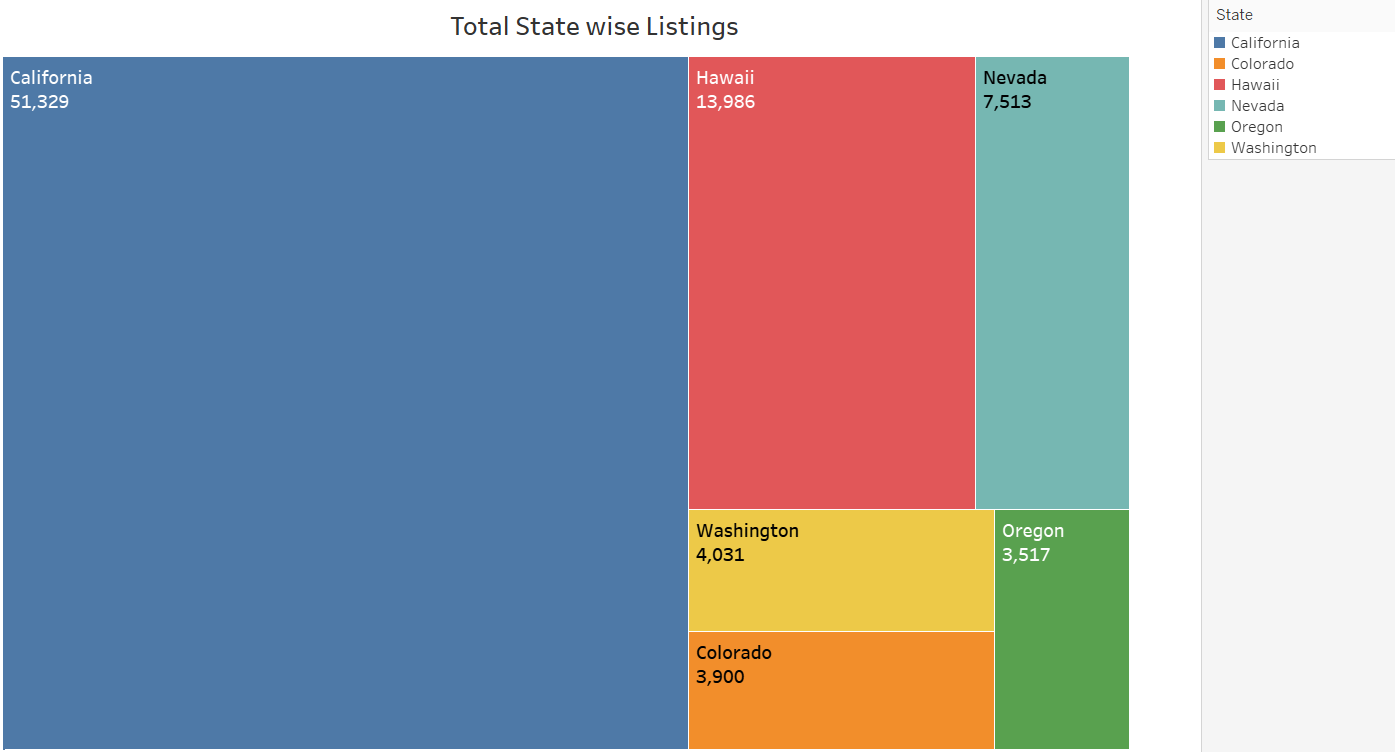
**How does the number of listings vary in different cities and state over time?**

* spark.sql("SELECT l.neighbourhood, month(to\_date(c.date)) month, count(DISTINCT(c.listing\_id)) distinct\_listings FROM listings l LEFT JOIN calendar c ON l.id = c.listing\_id WHERE instr(c.available, 't') >=1 AND neighbourhood IS NOT NULL GROUP BY l.neighbourhood, month ORDER BY distinct\_listings DESC")

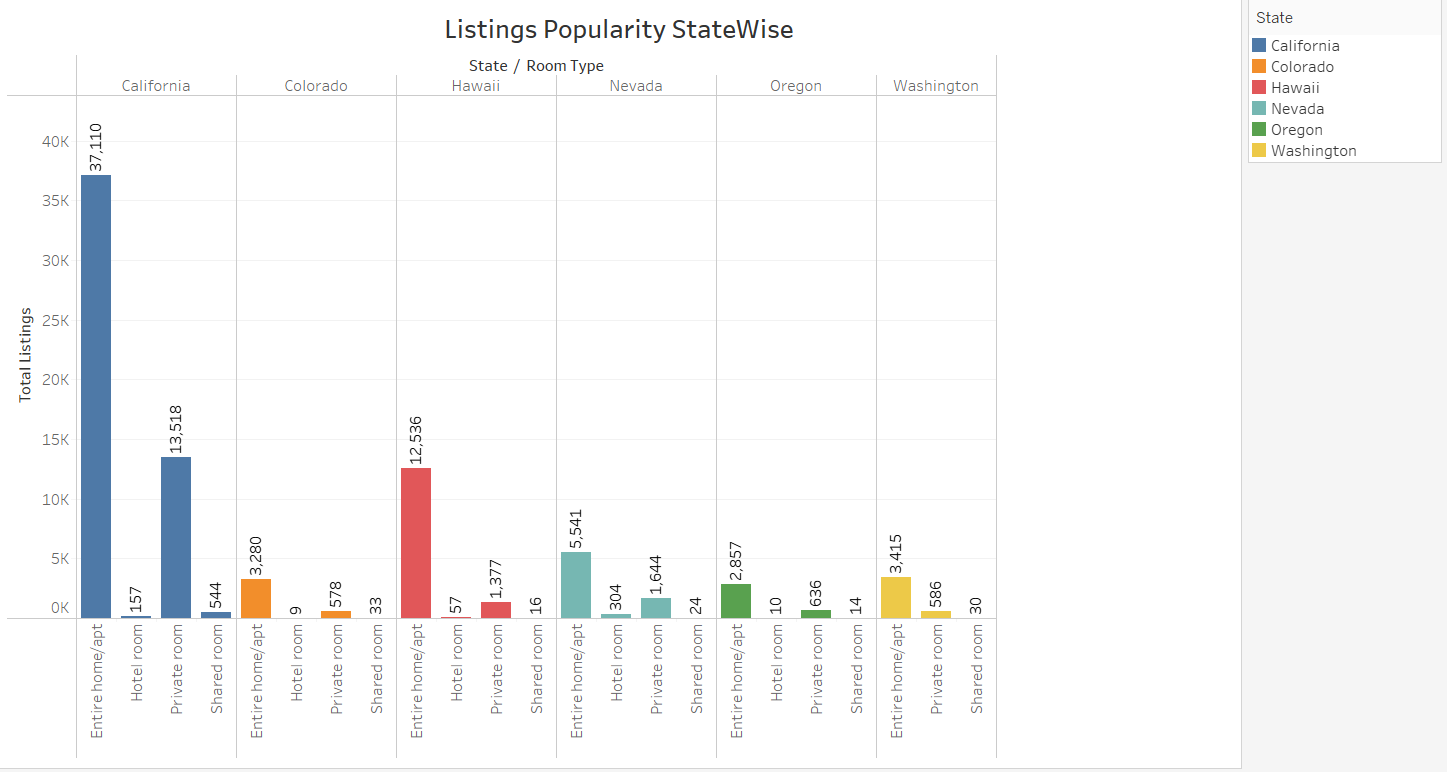
**Appendix B: ER Diagram**



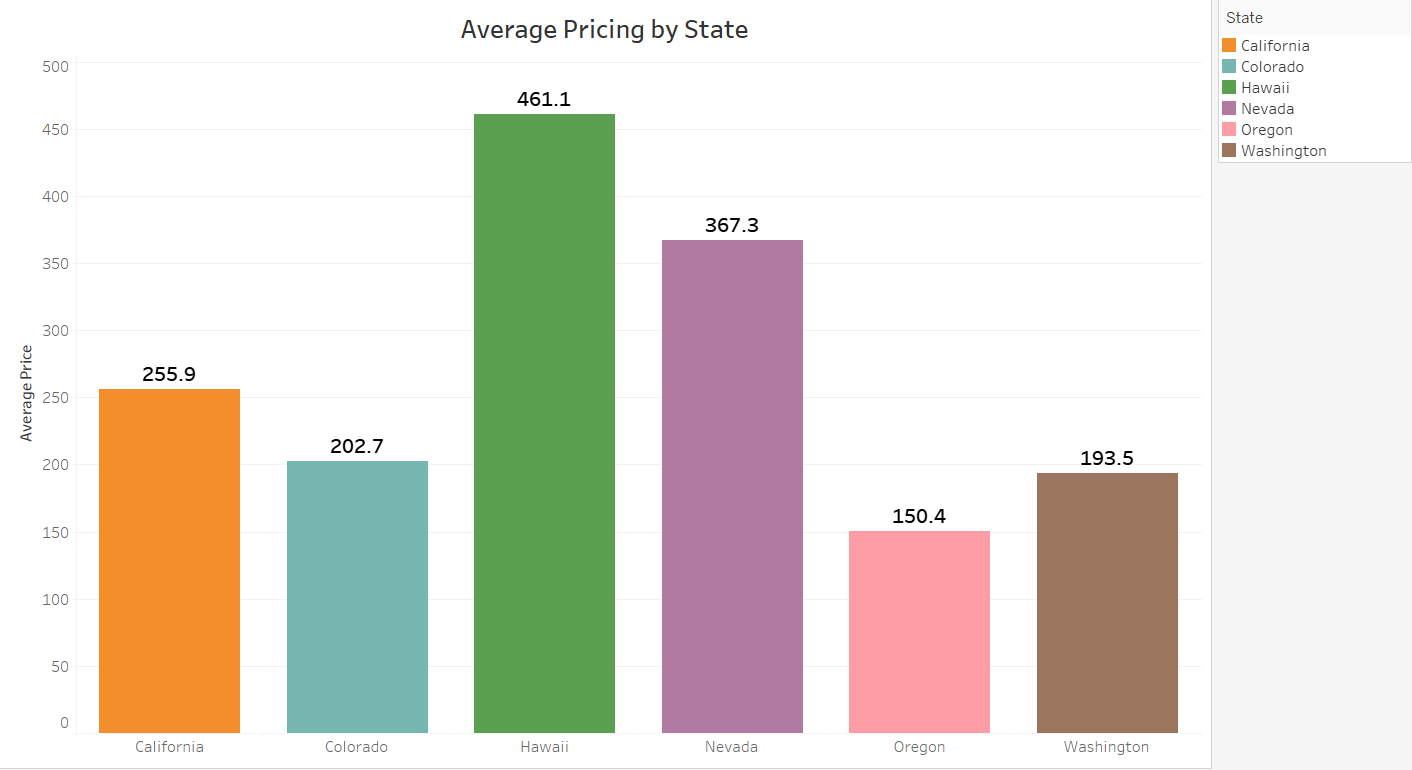
**Appendix C: Data Visualization**

How does the number of listings vary across each state?

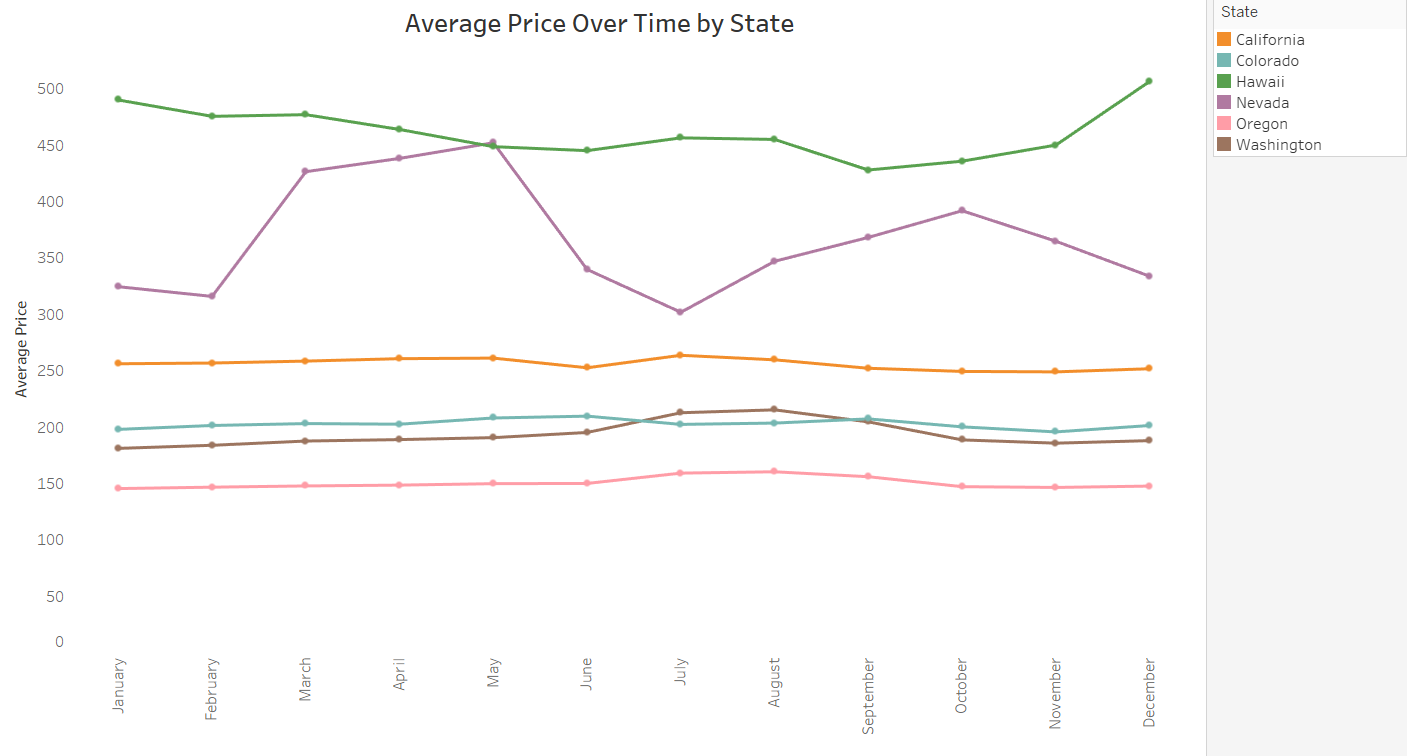
What are the popular listings in different location/region?



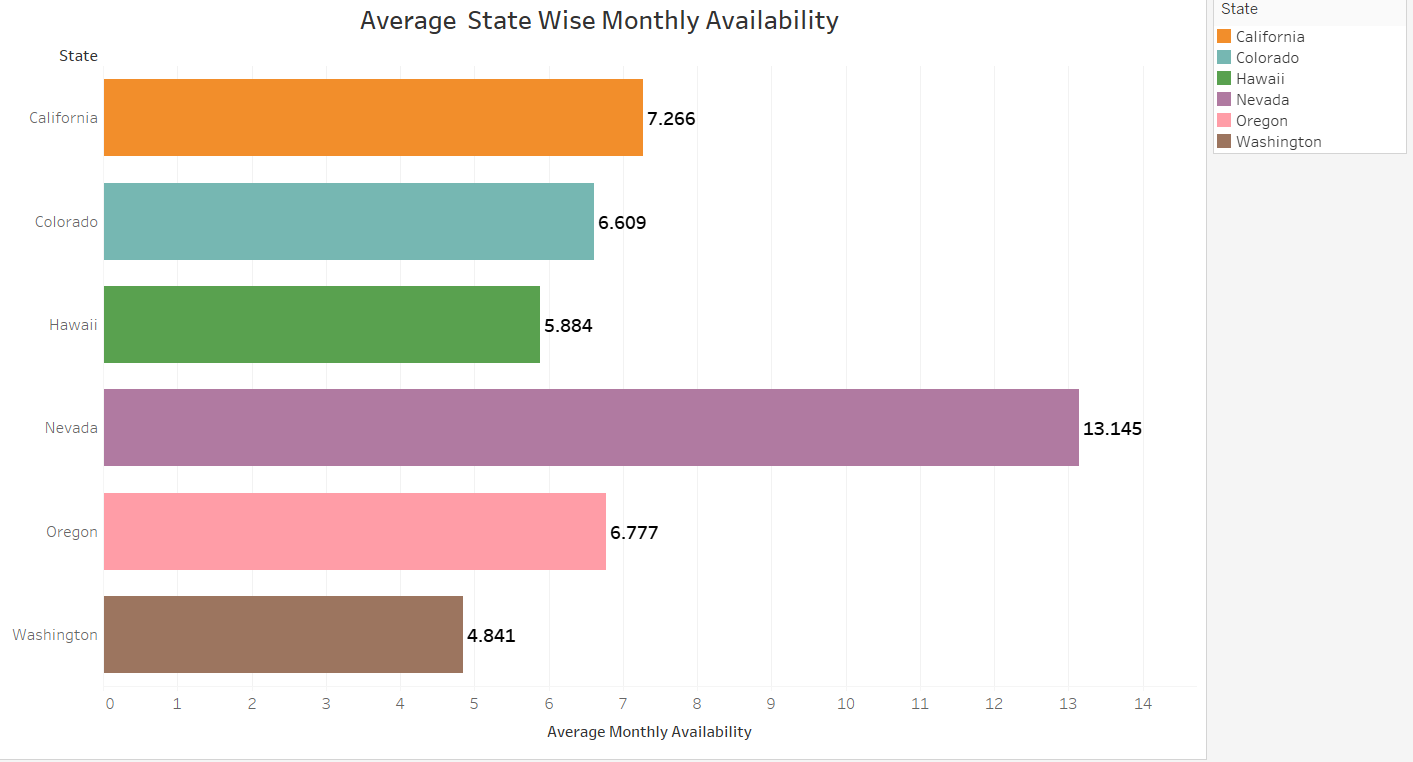
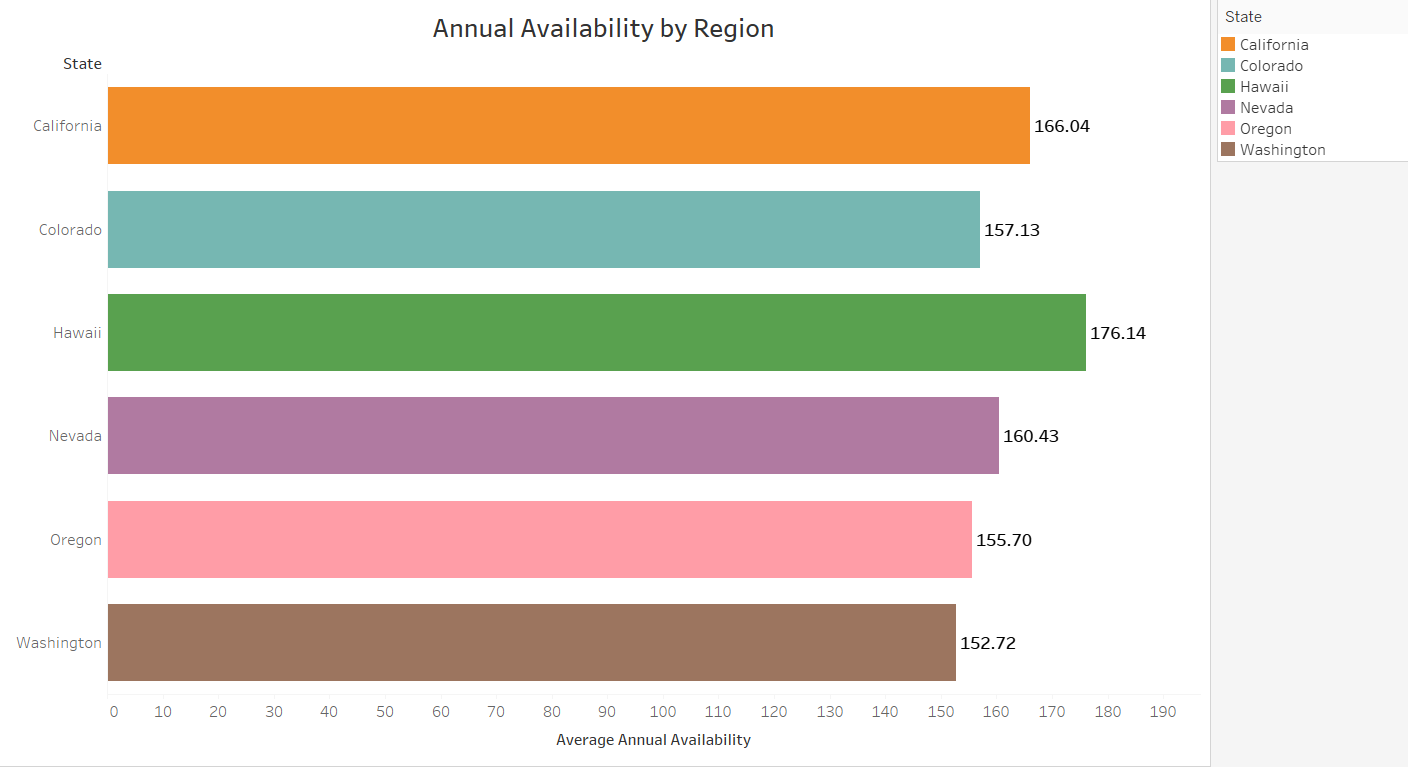
How does pricing differ in different states?



How does pricing differ over time in different states?



What is the monthly and annual availability by state?



* How do prices vary over time by region?

How does availability differ over time in different states?

