1. Introduction

In the fast-changing sharing economy, Airbnb is a key platform connecting travellers with unique places to stay worldwide. This assignment or study leverages Airbnb's extensive dataset to interpret the dynamics of California's Airbnb market, focusing on property locations, market trends, and amenity offerings. It aims to provide a comprehensive view of how these elements contribute to the overall experience on the platform for both hosts and guests. The core goal is to employ data visualization techniques for insights into the distribution of Airbnb properties across California, understand how market factors like occupancy rates and pricing affect profitability, and analyse the impact of amenities on a property's appeal and earnings. This analysis seeks to offer valuable perspectives that aid property owners and platform users in maximizing their Airbnb experience. Below are the details of the dataset utilized for this analysis and visualization.

1 Geolocations dataset: It as 49,312 data with 5 features (unified_id, month, street name, longitude and latitude).
2. Amenities dataset: It as same 49,312 data with 4 features (unified_id, month, hot tub and pool).
3. Market dataset: This as same data rows with 14 features, few important features are revenue, host type, occupancy, city, nightly rate and occupancy.

2. Data Visualisation

The dataset utilized was sourced from the "**Kaggle**" website. The primary aim of utilizing this data is to **analyse the evolution of tourism and Airbnb within a specific region, particularly during and after the COVID-19 pandemic**. By examining this data, I hope to gain insights into any fluctuations in travel behaviour during this period, whether there was a decline or increase in people making travel plans. I opted to focus on Airbnb data because I frequently use their platform for booking accommodations during my trips. (**question a**)

There are few codes on EDA and visualisation done on the same data and **they are available on the "Kaggle" website** but all the visualisation performed on this data is **very basic and they are plane visualisation** which provides minimum insight on the data set. (**question b**)

2.1. Geographical Distribution of Airbnb Properties

In this we plot different graphs to understand the spread of the 'AirBnB' properties around the California region. In this visualisation we can get an insight on the distribution of the properties based on longitude and latitude. And also, we can understand which street as major share of AirBnB booking over the years (2020-2022). Below is representation of street with highest number of booking records over the provided years.

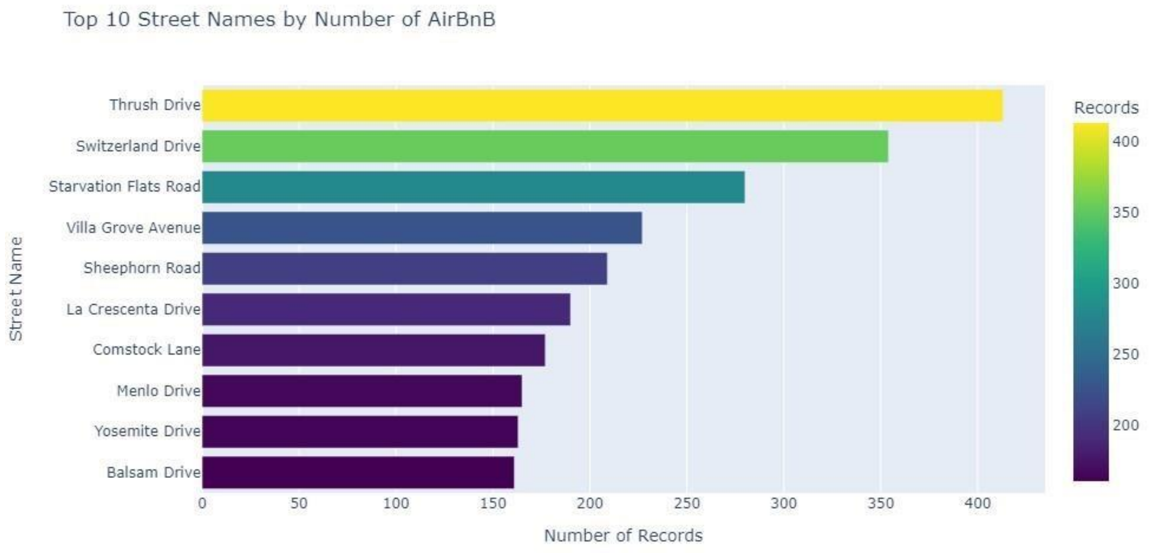

Figure 1. Distribution of AirBnB based on streets

From the above plot we can easily find that the "Thrush Drive" street as the highest number of bookings and the "Balsam Drive" street as the lowest. The reason for using this colour palette is it represents the flag colour of "California" and this is also representing the colour of the regions landscape during the spring and autumn season. In this we can see the implementation of one of the "**Edward Tufte**" principles that is "**Providing context**" – giving proper labels and representation using proper measures thus making the graph self-explanatory.

The provided "**Geo-representation**" showcases the spatial distribution of properties across California using the "**Open-street-map**" style for enhanced geographical context. This visualization focuses on the California region, delimited by specific **latitude (32.5 to 42.0)** and **longitude (-124.5 to -114.1)** coordinates, to ensure a concentrated view of the area. This method highlights notable concentrations of properties, such as the dense clustering around the "Big Bear Lake" region, indicating its popularity for vacations and staycations. The map's interactivity, powered by **Mapbox**, enriches the user experience by providing **precise** property locations and **customizable** hover information, such as latitude and longitude. This choice of Mapbox and map style was deliberate to offer a detailed and **interactive** visual exploration of property distributions, emphasizing key areas of interest within California.

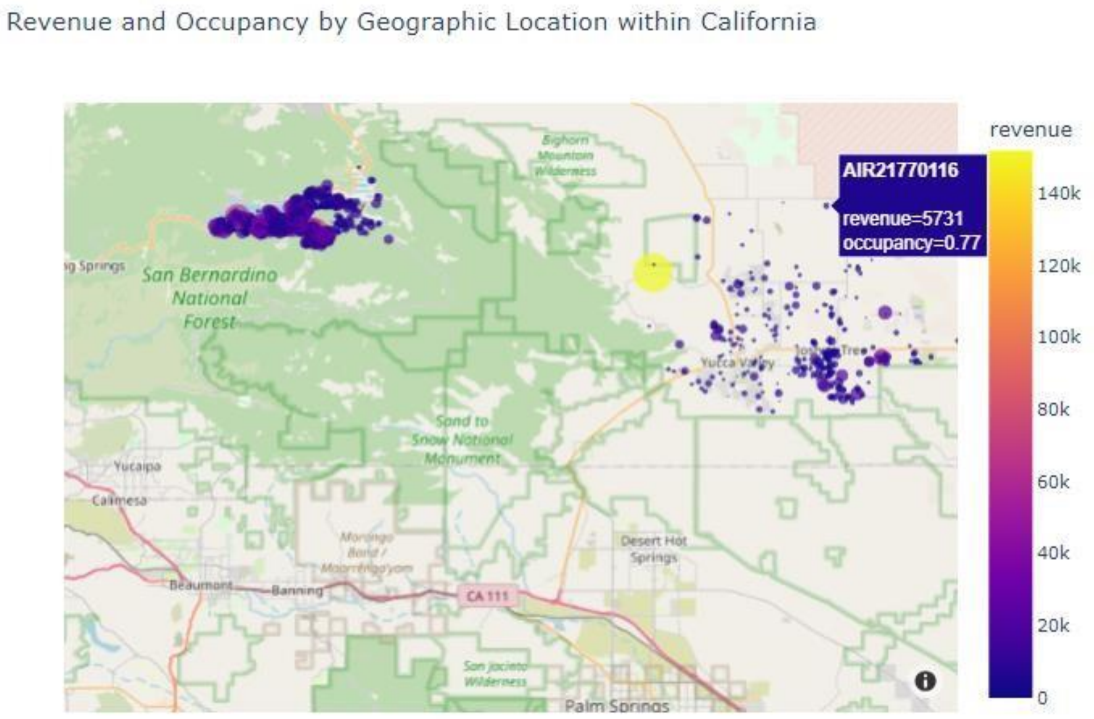

Figure 2. Geo plot for the distribution and revenue of the AirBnB

2.2. Amenities Impact Analysis and Visualisation

In this we are trying to explore and gain insight on the amenities data set and have they effect the revenue and also to understand the pattern of growth or decrease in the amenities (hot tub and pool) over given period of time. Below is a time series plot for understanding the trend in the hot tub and pool over the years from 2020 and 2022. This plot is also **interactive**, as you hover over it you can see the value of number of amenities available during that time. For all this kind of plots have kept the “White grid” as the background.

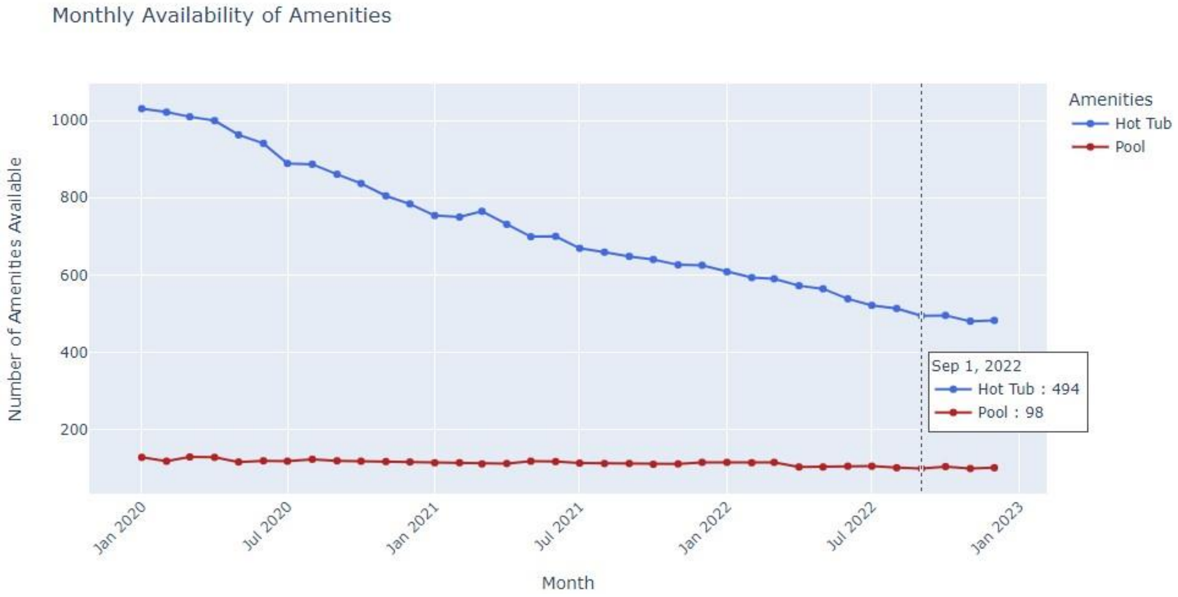

Figure 3. Time series plot for amenities over years

From the above visualisation we can depict that there is a drop in numbers in the hot tub, where as there is a constant number in the pool. This can be due to various reason like introduction better shower experience over the year and also can be due to safety measures during the covid time and also may be due to some revenue growth strategy. In this we can see the implementation of the main principle of the "**Edward Tufte**" that is not using any unnecessary decoration and just keeping the visualisation simple and clean, thus keeping "**chart junk**" away.

2.3. Market dynamic analysis and visualisation

This dataset is one of the primary sources used for analysis as it offers valuable insights into property revenue, city, occupancy rates, and other metrics. One of the visualizations conducted involves analysing revenue based on occupancy across different months of the given year. We opted for a time series plot for this analysis because it provides deeper insights compared to other types of plots like bar plots or histograms, given its dependency on time. Furthermore, we enhanced the plot's **interactivity** by incorporating a **dropdown** menu for selecting years, allowing for a more granular understanding of revenue trends over each year. This approach represents a **novel** method not attempted in existing code for this dataset. Below is the plot for the same.

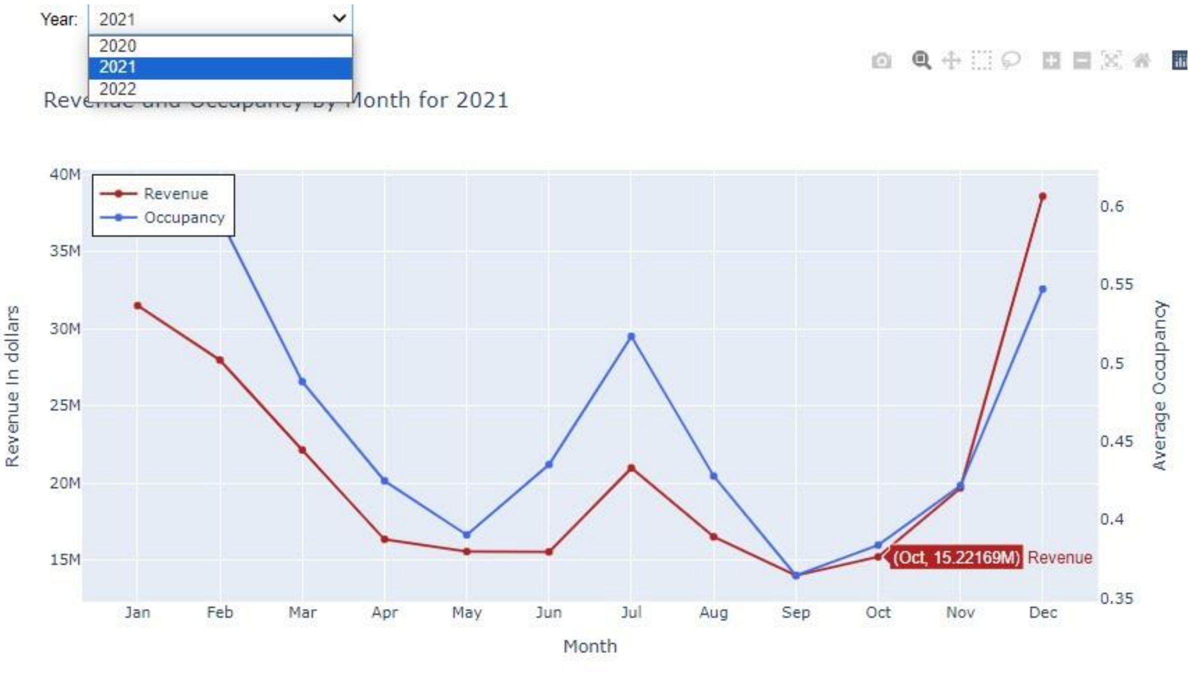

Figure 4. Time series plot for revenue and occupancy over the years

To create this visualization, we initially extracted the year and month individually from the month variable in the dataset. Then, we calculated the mean values for revenue and occupancy by grouping them according to the month and further grouping the months by year. For plotting, we utilized the "plotly-graph object" library. This visualization demonstrates the implementation of Edward Tufte's principle of "**Data Integrity**." We faithfully represent the data without making any alterations or modifications, ensuring that the graph accurately reflects the actual dataset.

We've also created a **3D scatter plot** to explore the relationship between revenue, occupancy, and nightly rate for the properties. We opted for this visualization because it allows us to compare three different variables simultaneously. Additionally, the 3D scatter plot enables us to identify any potential linear relationships among the variables, which may not be as evident in other types of plots. In this we have tried to implement or use the "**Data Density**" principle of Edward Tufte by using dense data that is multiple data and to provide a clear and vivid data visualisation of that. For this we are using the "**plotly.express**" library at this helps to visualise the complex variable by using simpler code in it.

Below is the 3D plot for the same.

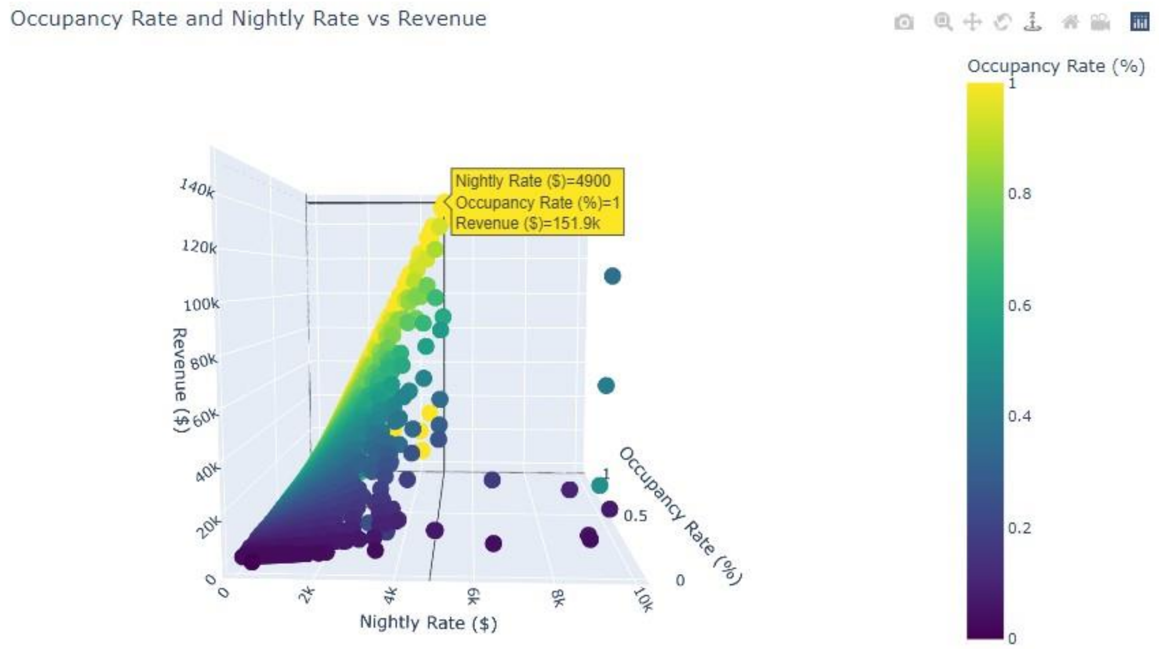

Figure 5. 3D – Scatter plot of Revenue, Occupancy and Nightly rate

One of the **key complex concepts** utilized during the analysis and visualization involved **merging all the datasets** into a unified dataset by extracting relevant variables from each dataset to gain insights. To achieve this, we utilized the 'unified_id' attribute, which serves as a common identifier across all datasets and corresponds to the unique ID associated with each Airbnb property. During the merging process, several exploratory data analysis (EDA) tasks were necessary. For instance, we noticed **that 'AIR' code was missing** in the market dataset, so we had to include it before merging. Additionally, we ensured that all data types were consistent across the datasets to facilitate seamless integration. By making this we can easily gain insight on the overall data set. As the data set after merging had many **duplicate data** in it, so had to perform few tasks to retrieve only the unique data from the merged data to gain the insight of the data. In this we have created a data visualisation by taking the latitude, longitude, host type, revenue, amenities available and city of the property. (**question d**)

In this plot to have used the "**Open-street map**" style to represent the map for pointing different properties based on the longitude and latitude. The colour in the plot represents different host type and also the corresponding amenities available. Below is the graph or visualisation of it.

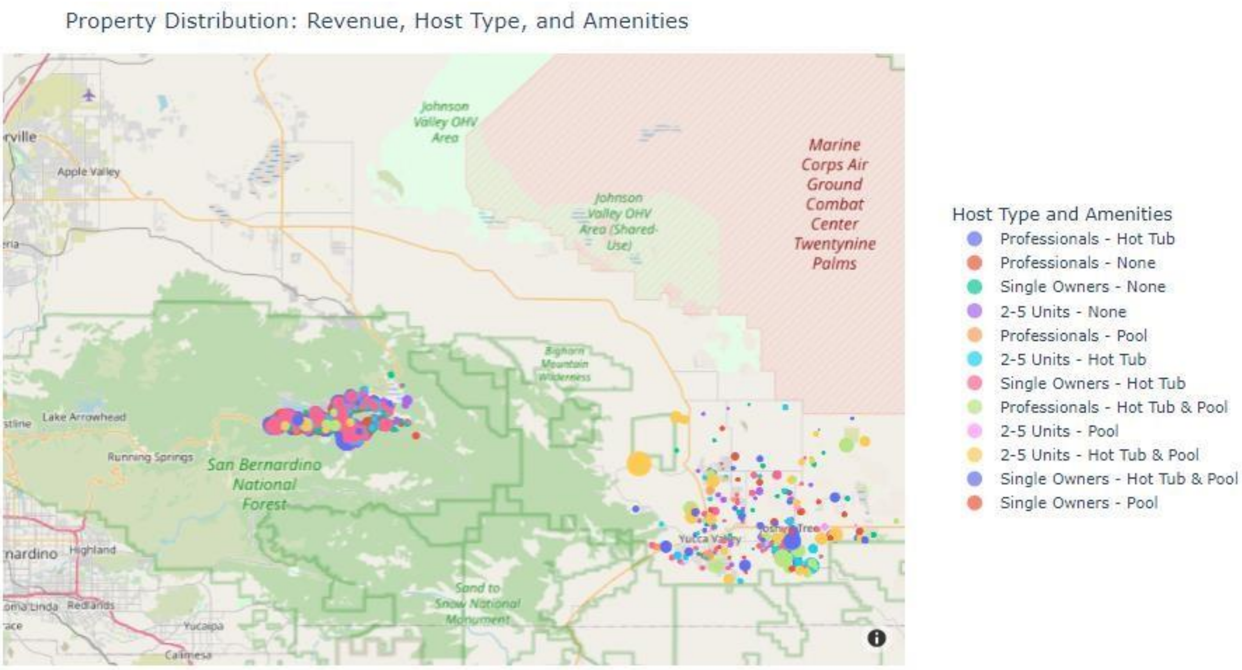
Property Distribution: Revenue, Host Type, and Amenities
Johnson Valley OHV Area
Marine Corps Air Ground Combat Center Twentynine Palms
Johnson Valley OHV Area (Shared-Use)
Apple Valley
Bighorn Mountain Wilderness
Lake Arrowhead
Running Springs
San Bernardino National Forest
Yucca Valley
Joshua Tree
Highland
Redlands
Yucaipa
Calimesa
Sand to Snow National Monument
Host Type and Amenities
● Professionals - Hot Tub
● Professionals - None
● Single Owners - None
● 2-5 Units - None
● Professionals - Pool
● 2-5 Units - Hot Tub
● Single Owners - Hot Tub
● Professionals - Hot Tub & Pool
● 2-5 Units - Pool
● 2-5 Units - Hot Tub & Pool
● Single Owners - Hot Tub & Pool
● Single Owners - Pool

Figure 6. Geo plot for representing the important attribute from all data set

In the depicted plot, you'll notice a 'legend' featuring 'Host Type and Amenities'. This plot is **interactive**; as you hover over the points, you'll receive detailed information about the property. Another **complex** aspect of this plot is that you can **selectively view specific properties** by choosing them based on the 'Host Type and Amenities' available in the legend. (**question d**)

The **novel** approach I adopted for this data analysis and visualization involved **merging** datasets based on the specific requirements for visualization and integrating all the relevant data. Furthermore, I enhanced the **interactivity** of most plots to extract maximum information. Additionally, I utilized various types of plots tailored to the attributes we were examining. This type of visualization is unique because it hasn't been implemented by any other codes available on websites. This approach allows for a comprehensive understanding of the data by presenting it in a visually engaging and interactive manner, ensuring that users can explore and interpret the data more effectively (**question c**).

3. Data Analysis (question - f)

The primary focus of my data analysis on the 'California Airbnb' dataset was to extract insights regarding revenue distribution across various cities and to conduct statistical analysis on revenue generated by different properties based on various attributes. This involved examining how revenue varies across different cities and understanding the statistical trends in revenue generation concerning different property attributes. By delving into these aspects, I aimed to uncover patterns and relationships within the data that could provide valuable insights for stakeholders, such as property owners, policymakers, and Airbnb users.

Below is one the plot that represents the **distribution of the revenue over the years** based on different cities of the California.

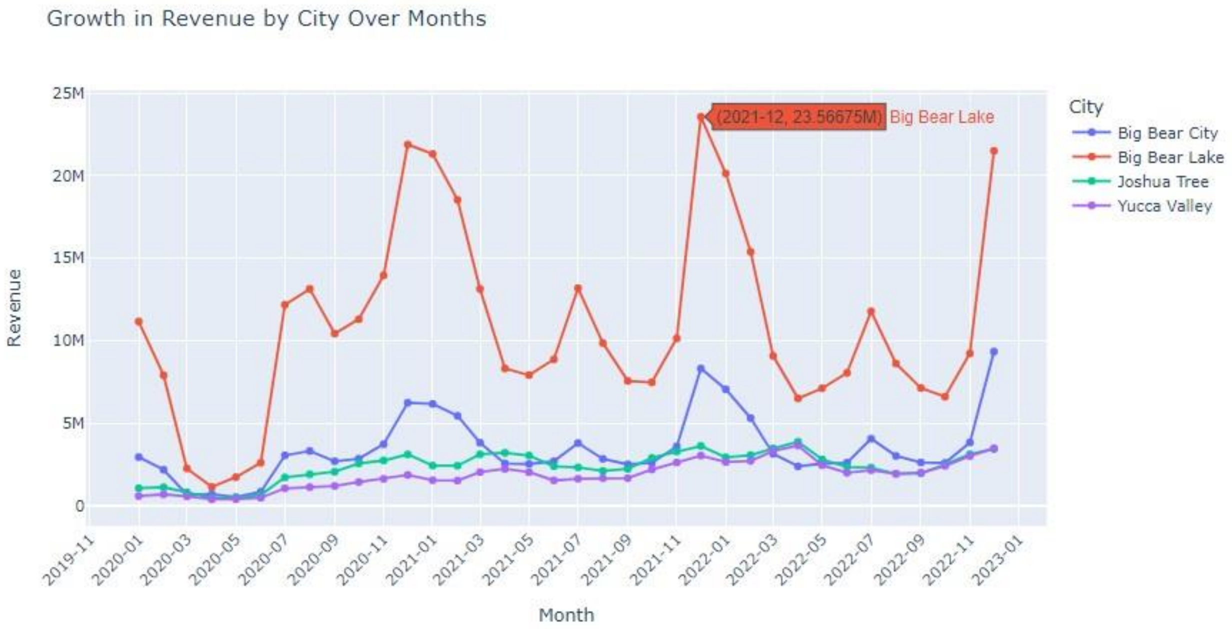

Figure 7. Line graph for representation of revenue vs cities

From this we can easily compare the revenue distribution over different cities of the California. And also from the **figure 5**, we can also depict that there is a linear relationship between the nightly rate and the revenue of the properties.

4. Theoretical and Design Considerations (question - e)

Below are the four main theoretical and design consideration I made while analysing or doing this project,

- **Geospatial Data Visualization:** The includes sophisticated geospatial visualizations, such as the use of "Open Street Map" to filter and display Airbnb locations within California. This choice indicates a focus on providing clear context to the data, allowing users to understand how Airbnb listings are distributed geographically. The visualization of data on maps helps in identifying patterns, such as clustering of listings in urban versus rural areas or proximity to points of interest. This method utilizes human spatial understanding to improve the readability and interpretation of complex data sets.
- **Interactive Widgets for Dynamic Data Exploration:** Incorporating interactive elements, like dropdown menus for selecting particular years or features for study, highlights a focus on engaging and dynamic examination of data. This design choice enables us to customize their data view and engage with the visualization in a more interactive manner. By providing options to refine the dataset according to our preferences, the analysis turns more userfriendly and cater to a wide range of users. Such a method promotes active engagement with the data, moving away from simply viewing fixed charts.

- **Time Series Analysis for Trend Identification:** The visualization showcases time series analysis to investigate temporal trends, like tracking the availability of amenities (for instance, hot tubs and pools) over various months. This aspect underscores the significance of understanding time-related changes within the data, offering perspectives on how specific attributes or trends develop over periods. Through organizing the data to support time series analysis, the visualization empowers users to detect seasonal variations, trends of growth or reduction, and possible cyclical patterns. Such insights are vital for making educated forecasts or decisions rooted in historical data trends.
- **Choice of Colour:** As mentioned earlier too the colour chosen through out the visualisation is mostly gold(yellow), blue and purple shade as it represents the flag of California and also, it's the tropical colour during the spring and autumn season in that region.

Conclusion

From the data analysis and visualisation performed on the taken data set of AirBnB, we can see that the revenue is linearly related to the nightly rate of the stay. The "Big Bear Lake" city is the people's favourite place to travel for there holidays. So, this is the major contributor to the revenue. And also, from the trends we can see that the revenue and occupancy increase during the summer and during Christmas holidays in that region.

Future Scope (question g)

Below are few scopes for future work if we were given more time,

- **Predictive Model:** We can implement machine learning models to predict trends, pricing strategies, and occupancy rates. Utilizing historical data patterns to forecast future demand and pricing can help hosts optimize their listings for maximum profitability and occupancy.
- **Cross-Dataset Analysis for Comprehensive Insights:** Merge Airbnb data with external datasets such as tourism statistics, events, local transportation availability, and economic indicators to explore broader impacts on rental demand and pricing. This holistic view could reveal opportunities and challenges in the short-term rental market.
- **Sentiment Analysis of Reviews:** Perform sentiment analysis on guest reviews to identify factors that significantly impact guest satisfaction. Insights gained can guide hosts in making targeted improvements and help Airbnb in developing features that enhance the guest experience.

Word Count : 2320

References

[1] https://www.kaggle.com/datasets/computingvictor/zillow-market-analysis-and-real-estatesales-data/data - Data source.

[2] Nathan Yau and Hans Gosling, Chapter 3 – Representing data, Data Points: Visualisation.

[3] Edward Tufte’s principles - https://www.youtube.com/watch?v=HfXSltlDfDw

[4] How to make interactive maps using Python - https://www.youtube.com/watch?v=16ndLqsy6M

[5] 3D- Scatter plot using plotly - https://stackoverflow.com/questions/67293189/plotly-3dplot-in-python