Airbnb Reservation Data Analysis and Visualization

Introduction

The data is about the Airbnb Entity Travel team. This team is in charge of adding and on-boarding entities where they track the overall success and trends of all entities combined as well as each entity through metrics such as nights booked and booking value. In this project, I will understand the dataset, prepare data, deep into the dataset to understand different scenarios and business problems, do research and define metrics and finally find trends, put forward recommendations and give insights on how to improve the entity travel team’s performance.

Data Tables

There are four data tables with each column description below. All dates and timestamps are in UTC. In this repository, fct\_bookings.csv, fct\_cancellations.csv and dim\_entities.csv are uploaded, but dim\_listings.csv is too big to be uploaded.

fct\_bookings.csv

Column | Type | Comment

--------------------------+---------+----------------------------------------------

id\_entity | varchar | ID of the business entity

id\_reservation | varchar | ID of the reservation

dim\_source | varchar | Source of reservation

id\_guest | varchar | ID of the user that made the booking

id\_traveler | varchar | ID of the guest travelling on the booking

id\_listing | varchar | ID of the listing

m\_nights\_booked | bigint | Number of nights booked in the reservation

m\_guests | bigint | Number of guests in the reservation

m\_booking\_value | varchar | Total value of the reservation

ds\_checkin | varchar | Datestamp of the start of the reservation

ds\_checkout | varchar | Datestamp of the end of the reservation

is\_business\_travel\_ready | bigint | 1 if the listing is business ready, 0 otherwise

is\_self\_checkin | bigint | 1 if the listing is self checkin ready, 0 otherwise

ts\_booking | varchar | Unix timestamp of when the booking was made

ds | varchar | Date stamp of when booking was made

fct\_cancellations.csv

Column | Type | Comment

-------------------------+---------+-----------------------------------------------

id\_reservation | varchar | ID of the reservation

id\_listing | varchar | ID of the listing

id\_guest | varchar | ID of the user that made the booking

id\_traveler | varchar | ID of the guest travelling on the booking

id\_host | varchar | ID of the host on the booking

dim\_cancellation\_reason | varchar | Reason for cancellation

ts\_cancelled | varchar | Unix timestamp of when the booking was cancelled

ds | varchar | Date stamp of when booking was cancelled

dim\_entities.csv

Column | Type | Comment

--------------------------+---------+----------------------------------------------

id\_entity | varchar | ID of the business entity

dim\_company\_size | varchar | Entity size, as entered by user

dim\_preferred\_currency | varchar | Preferred currency for reporting and receipts

dim\_region | varchar | Region of the entity

dim\_country | varchar | Country of the entity

dim\_state | varchar | State of the entity

dim\_payment\_type | varchar | Preferred payment type of the entity

dim\_status | varchar | Marks the current invite status of the entity

dim\_is\_setup\_for\_groups | bigint | Flag if the entity is setup for groups

dim\_is\_active | bigint | Flag if the entity is not deleted

dim\_is\_signed\_up\_company | bigint | Flag if the entity is signed-up (has at least 1

verified admin (travel manager))

ts\_created | varchar | Unix timestamp of creation

dim\_listings.csv

Column | Type | Comment

-------------------------+---------+----------------------------------------

id\_listing | varchar | ID of the listing

id\_user | varchar | ID of the user who manages the listing

dim\_region | varchar | Region of the listing

dim\_country | varchar | Country of the listing

dim\_state | varchar | State of the listing

dim\_market | varchar | Market of the listing

dim\_is\_active | varchar | 1 if the listing is an active listing, 0 otherwise

dim\_room\_type | varchar | The type of room (Private room or Entire home/apt)

dim\_listing\_tier | varchar | The tier of the listing (select, luxury or marketplace)

dim\_total\_bookings | bigint | Total number of nights with calendar blocked

dim\_total\_reviews | bigint | Total number of reviews

dim\_cancellation\_policy | varchar | Cancellation policy for the listing

dim\_person\_capacity | bigint | Number of people the listings can accommodate

dim\_bedrooms | bigint | Number of bedrooms

ts\_listing\_created | varchar | Unix timestamp of listing creation

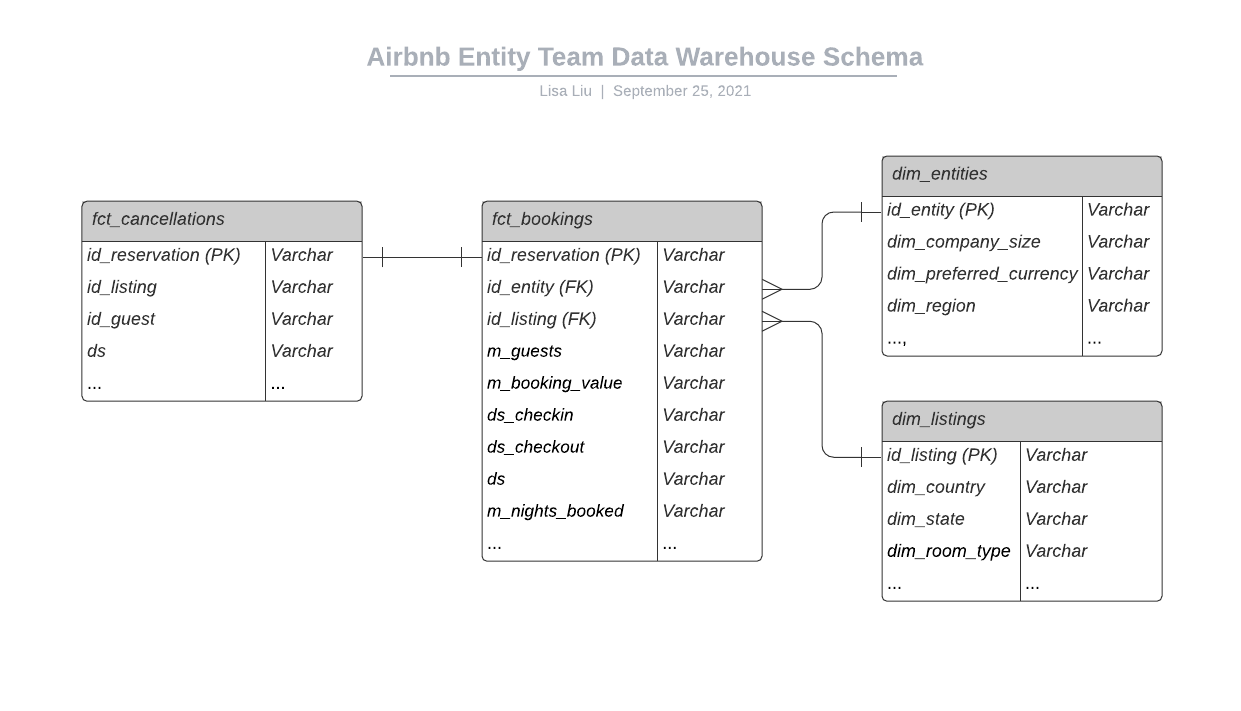
Business problem

Scenario 1: The VP of the Entity Travel team notices two daily spikes in total booking value (before cancellations) based on checkin dates throughout 2017 and 2018 for listings in the United States. The VP wondered what are the root causes to the top two main spikes and what is the financial impact compared to a normal day. The VP also wondered whether any findings or recommendations could be put forward.

Scenario 2: A sales team manages the top 10 entities based on 2018 YTD booking values by entity. The team would like a report for these 10 entities that show findings and trends in key metrics since 2017. A report for the team is expected which highlights key findings and recommendations.

Data Warehouse Schema

I used MySQL to build up a warehouse schema and related four tables by defining the primary keys and foreign keys. The Data Warehouse Schema is shown in the picture below.



Dataset Understanding and Metrics Defining

Table fct\_booking is a fact table which can be related easily by inner joining with dimensional tables dim\_entities and dim\_listings. I cannot use ‘inner join’ to join fct\_bookings and fct\_cancellations, since it will remove those reservations which are not canceled. Table fct\_bookings has a narrower time span (587 days) than table fct\_cancellations (698 days). It can be concluded that some rows of fct\_canncellation are not common with fct\_bookings, so I cannot use ‘full join’ to join these two tables. I finally chose to use ‘left join’. After running ‘left join’, I found that 512 out of 70772 reservations are cancelled, which is reasonable. At this stage, the joins between all tables were established.

There are a total of 141 entities and the top five in the number of bookings are 52, B98, 8BE, 267 and 53, which account for 57% of bookings. There are a total of 147 countries where listings are located and the top five are US, GB, DE, FR and AU and they account for 70% of listings.

The next step is to deep into the variables in table fct\_bookings and to define metrics which will be used to analyze the performance. Based on my research, I found a few high-level matrices below:

● Bookings: The count of id\_reservation from fct\_bookings.csv

● Nights Booked: The sum of m\_nights\_booked

● Booking Value: The sum of m\_booking\_value

● Cancellation Rate: count of id\_reservation from fct\_cancellations.csv / count of id\_reservation from fct\_bookings.csv

● Price Per Night: The sum of m\_booking\_value / the sum of m\_nights\_booked

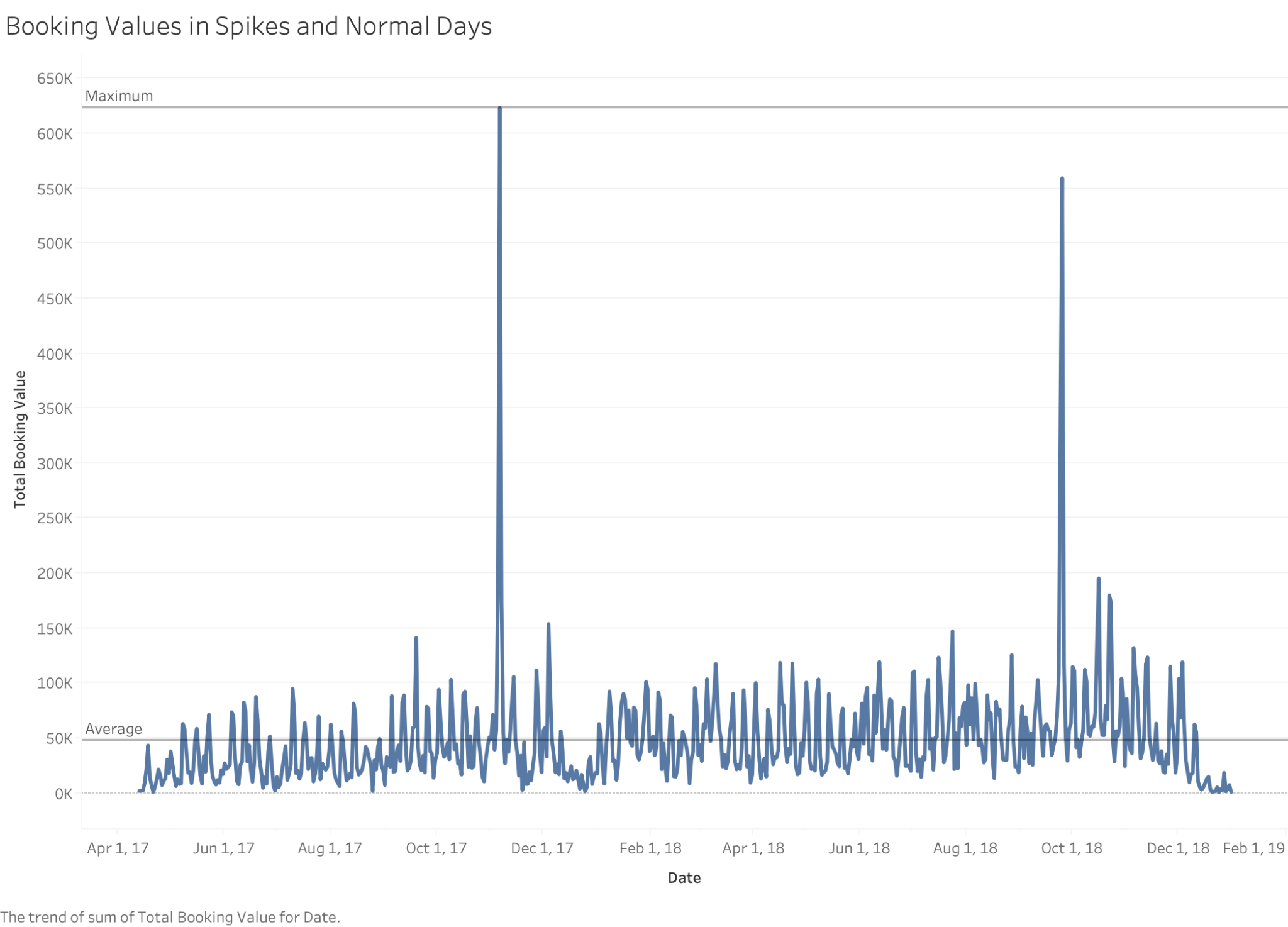
● Nights Per Booking: The sum of m\_nights\_booked / count of id\_reservation from fct\_bookings.csv

● Budget per Person per Night: The sum of m\_booking\_value / ( the sum of m\_guests \* the sum of m\_nights\_booked)

Data Analysis and Visualization

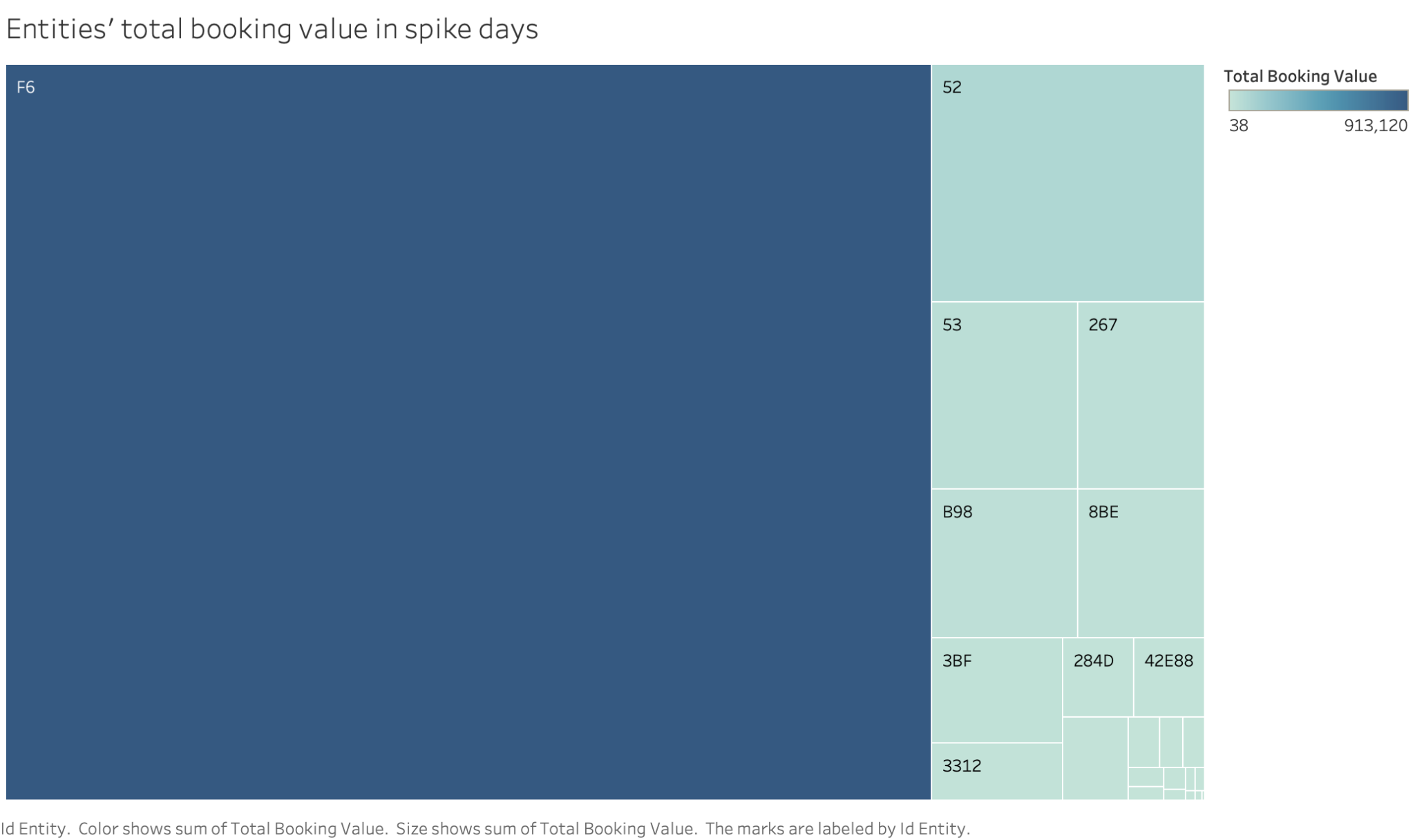
I would like to deal with Scenario 1 first. To find two spikes, I need to know the sum of booking values for each day in the US and in 2017 and 2018. I found two days have the highest booking values- 2017-11-06 $623,355 and 2018-09-25 $559,418.

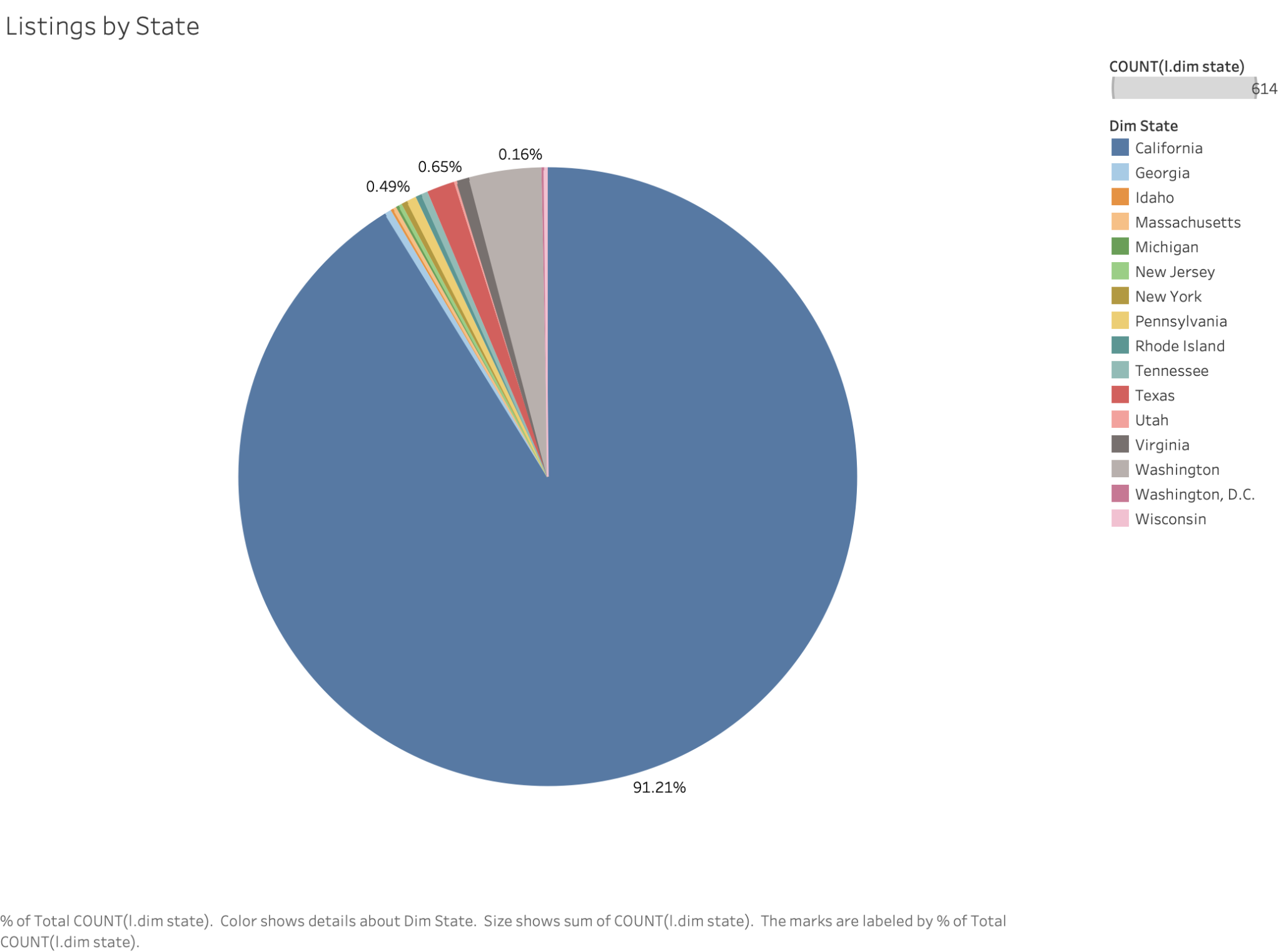
As shown in the picture below, the booking values in these two spikes are 10-15 times normal days ($47,598).



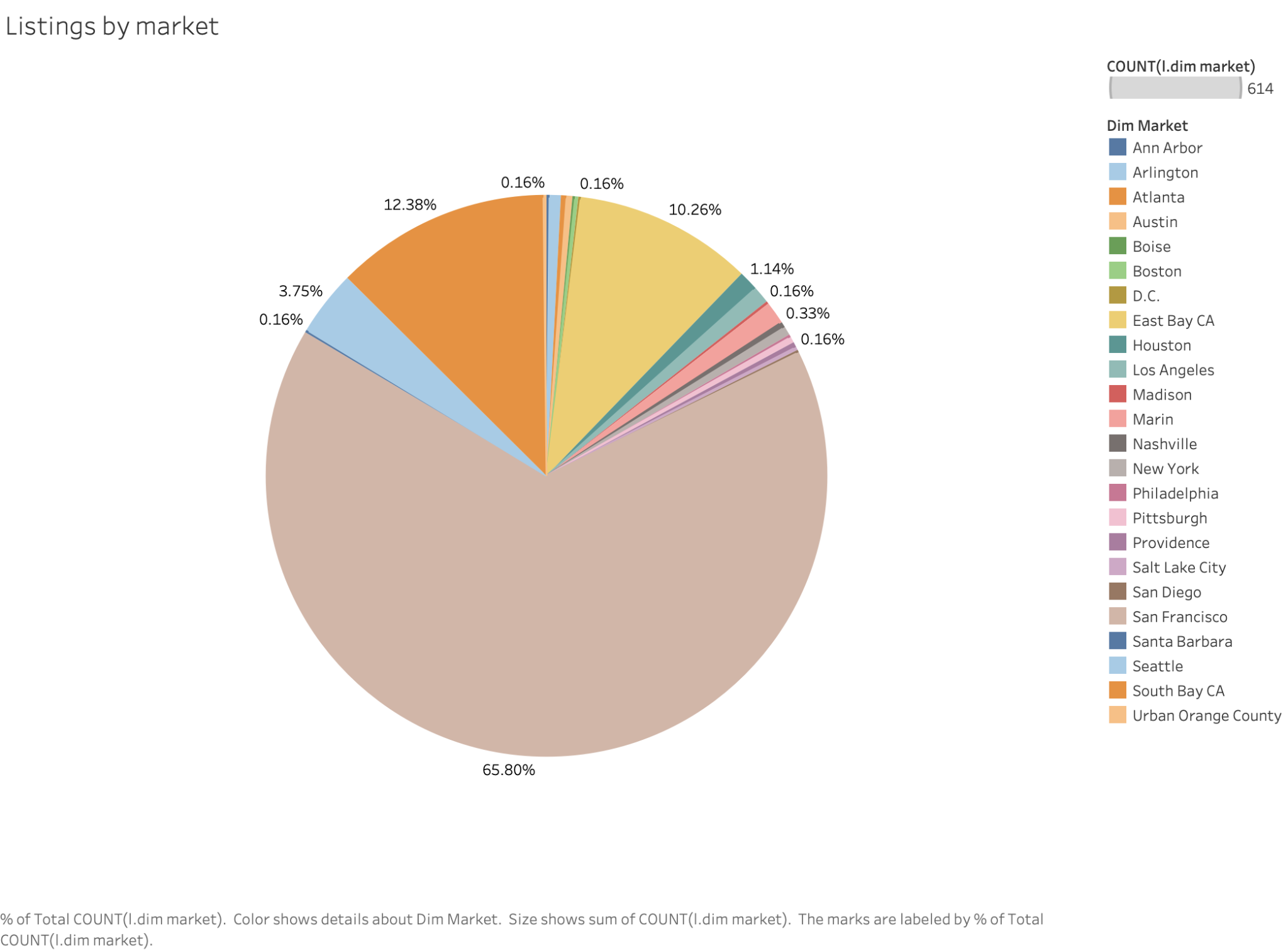
To figure out the root cause of the spikes in these two days, I searched these two days, but they did not turn out to be holidays. I suppose the spikes may be caused by the behaviors of an entity or a group of entities.

First, I tried to find out what entities reserve rooms on these two days. I found 23 entities have reservations on these two days. Then I tried to find out whether a specific entity/ a group of entities reserved a majority of all the rooms. Entity 'F6' reserve 77.2% out of all 23 entities (no 1: 77.2% F6, no 2: 7.4% 52, no 3: 3.1% 53, no 4: 2.7% B98)



Then I tried to figure out where the listings are located. By querying the data for listings in which the above top 4 entities (Top 4 entities account for 90.4% of all total booking value.) live, I found that most of the listings (91.21%) are in California, North America. 

And the majority of the listings (88.44%) are located in San Francisco and near San Francisco (South Bay CA and South Bay CA ).



In conclusion, the root cause of two main spikes is an annual event in San Francisco, CA. Since silicon valley is famous for its information and technology industry, I assume the event should be related to IT.

*Insights from data-Recommendation*

In order to make suggestions on these two spike days, I will dig deep into customer behaviors and find the pattern related to the entity itself, preferred listings and reservation behavior. To be specific, the metrics include entity’s country, room price, cancellation rate, # guest per room, # staty days, business or not, how early they would like to cancel, company size, group or not, preferred room type, listing tier, reviews and cancellation policy.

*Reservation Behavior*

In this part, I will explore reservation and cancellation behaviors for all entities on these two spike days. I dip further into the data and added a few low-level metrics here:

Demand-room: how many rooms are demanded on that check-in day, SUM(m\_nights\_booked from fct\_bookings \* dim\_bedrooms from dim\_listings)

Price per room per night: how much it costs per room per night, SUM(m\_booking\_value from fct\_bookings) / SUM(m\_nights\_booked from fct\_bookings\* dim\_bedrooms from dim\_listings)

Demand-guest: how many guests need to stay in Airbnb at that day, SUM((m\_guests from fct\_bookings) \* (m\_nights\_booked from fct\_bookings))

Price per Person per night: how much it costs per person per night, SUM(m\_booking\_value from fct\_bookings) / SUM((m\_guests from fct\_bookings) \* (m\_nights\_booked from fct\_bookings))

Reservation beforehand: how many days rooms are reserved ahead of check-in date, AVG(DATEDIFF(ds\_checkin from fct\_bookings, ds from fct\_bookings))

Cancellation beforehand: how many days reservations are cancelled ahead of check in date, AVG(DATEDIFF(ds\_checkin from fct\_bookings, ds from fct\_cancellations))

The table below is a comparison in metrics between the average of all days in 2017 and 2018 and two spike days.

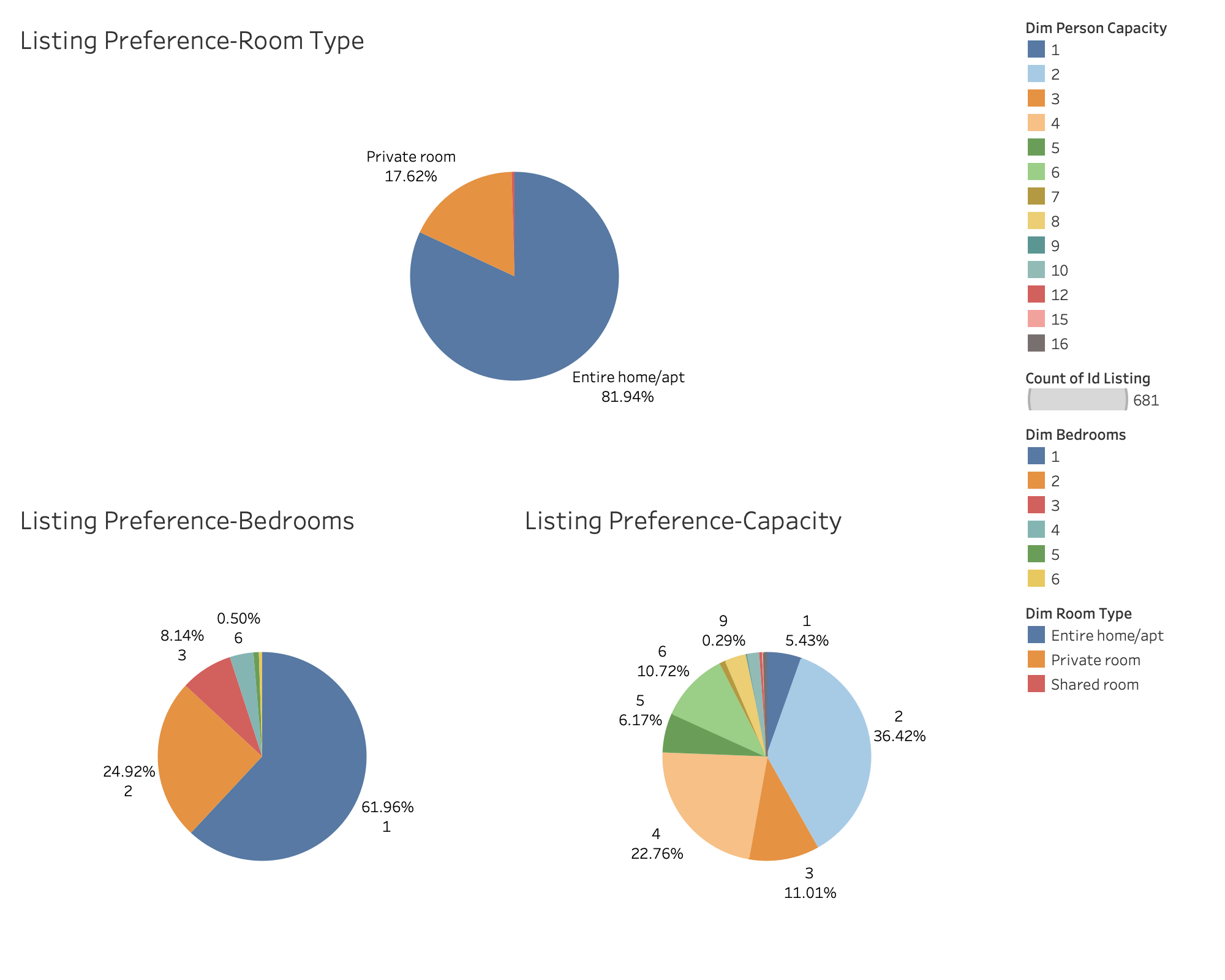
|  | **Total Bookings** | **Nights per Booking** | **Total Booking Value** | **Price per Night** | **demand-room** | **Price per room per night** | **demand-guest** | **Price per Person per Night** | **Reservation beforehand** | **Total Cancellations** | **Cancellation beforehand** | **Cancellation Rate(%)** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **AVG** | 45 | 6.2 | 47598 | 179 | 426 | 116 | 488 | 99 | 19.3 | 0.34 | 15.9 | 0.78 |
| **2017-11-06** | 356 | 5.014 | 623355 | 349.2185 | 2539 | 245.512 | 2804 | 222.3092 | 49.0225 | 4 | 44.75 | 1.1 |
| **2018-09-25** | 351 | 4.1795 | 559418 | 381.3347 | 2219 | 252.1037 | 2309 | 242.2772 | 46.339 | 8 | 67.625 | 2.3 |

The comparisons are analyzed below:

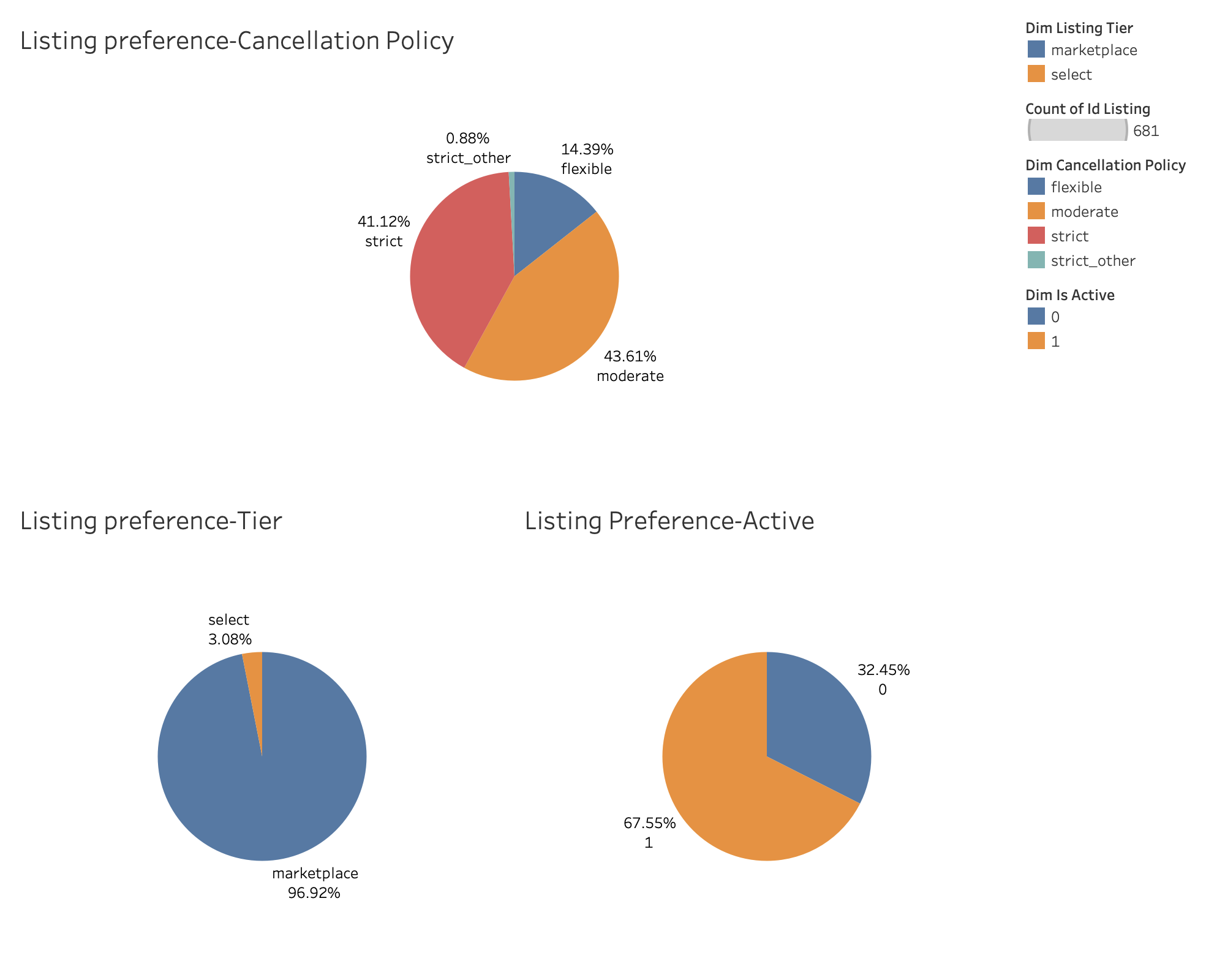
1. Total Bookings: The number of spike days is around 7 times that of the average.
2. Nights per booking: The number of spike days is around 0.6-0.8 times that of the average.
3. Total Booking Value: The number of spike days is around 10-12 times that of the average.
4. Price per Night: The number of spike days is around 2-2.5 times that of the average. It means price bumps a lot on spike day.
5. Demand-room: The number of spike days is around 5-6 times that of the average. It means much more market demand on these two days.
6. *Price per room per night: The number of spike days is around 2-2.5 times that of the average. It is consistent with Price per Night.*
7. *Demand-guest: The number of spike days is around 5-6 times that of the average. It is consistent with Demand-room above. It means that the number of guests per room is around 1.*
8. *Price per Person per night: The number of spike days is around 2-2.5 times that of the average. It is consistent with Price per Night.*
9. Reservation beforehand: The number of spike days is around 1.5-3 times that of the average. It means that customers reserve much earlier than normal days.
10. *Total Cancellations: The number of spike days is around 10-20 times that of the average. It means that there are much more cancellations in spike days than in normal days. But one of the reasons is that there are more reservations.*
11. Cancellation beforehand: The number of spike days is around 3-5 times that of the average. It means that customers cancel reservations much earlier than normal days.
12. Cancellation rate: The number of spike days is around 3-5 times that of the average.

All in all, in terms of revenue generation, The events on these two spike days account for 3.96% of the US market in two years' revenue (2017 and 2018) and 2.34% of the global market in two years' revenue (2017 and 2018). It could be concluded that these two events are very important for corporate financial status. And, one spike day is equal to around 10 normal days and spike days should be paid much emphasis. However, guests tend to stay fewer days, which is backward in revenue generation, so efforts should be made on improving the revenue per spike day. The prices on spike days are much higher than on a normal day and prices are positively affected by demand. If demand for the following year keeps increasing, it may be forecasted that the price will increase along with it. Airbnb should collect the information of estimated event attendance and then make reasonable pricing strategies. Customers have good habits of reserve and cancel rooms as early as possible and Airbnb should keep on it. In addition, the cancellation rate is a little bit higher on spike days than normal days. Airbnb listings should use a more strict cancellation policy to avoid revenue loss. The night before these two days are also good, although around half of these two days. Hosts should get well prepared two days before the first event day. My suggestion is to keep connected with the clients at least half a year ahead to plan for the event. Besides the room, Airbnb can also provide other services, such as transportation, activities, food, etc. to help them successfully host the events and generate additional revenue.

*Listing*

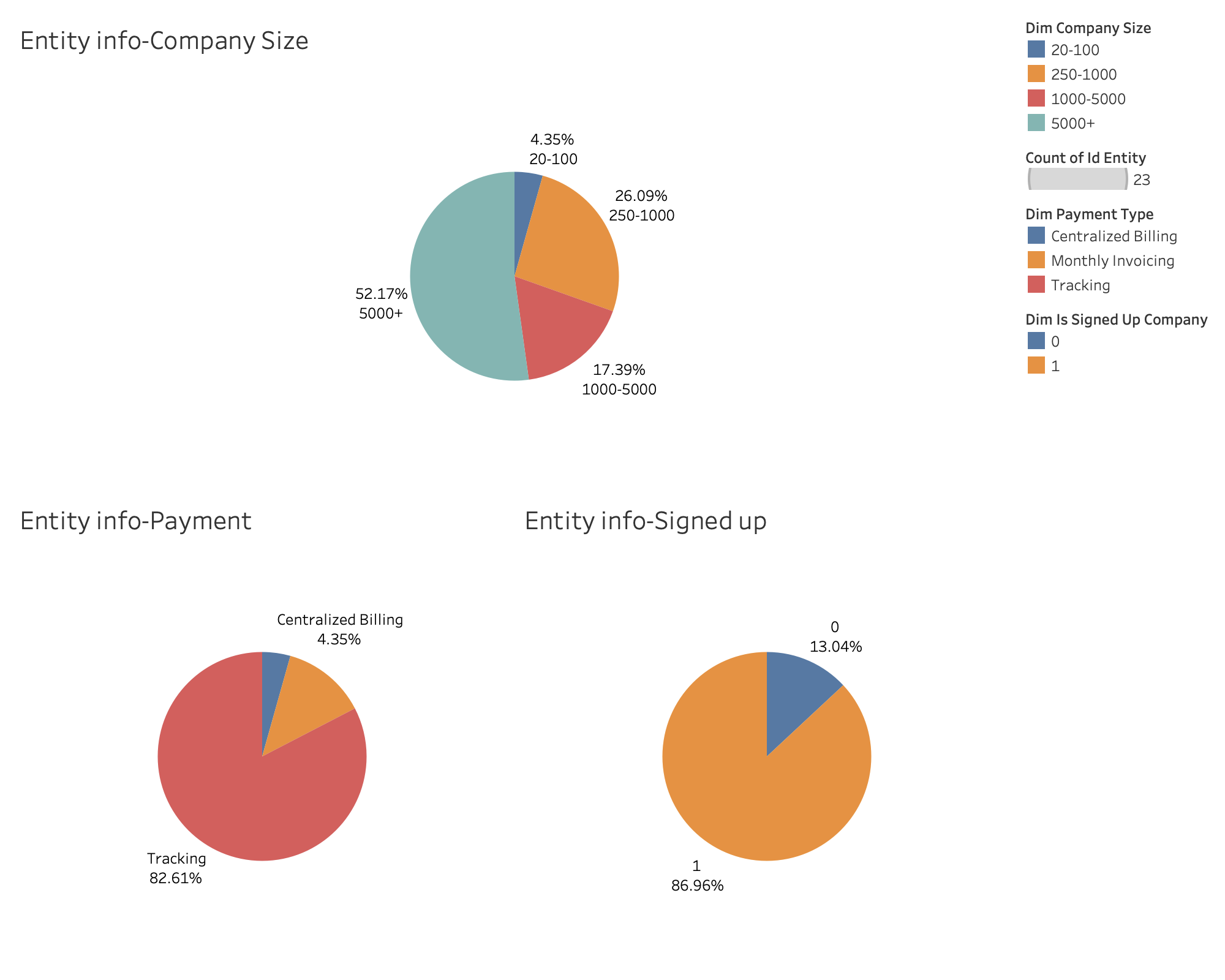
As shown in the diagrams below, entire home/apt occupies 81.94% of all room types, ‘1-2 rooms’ occupy 86% of all # bedrooms, ‘2-4 capacity’ occupy 69% of all capacity. The recommendation would be to prepare a large number of entire homes/apartments with 1-2 rooms and 2-4 capacity and promote these listings to customers in facing competition with hotels. 

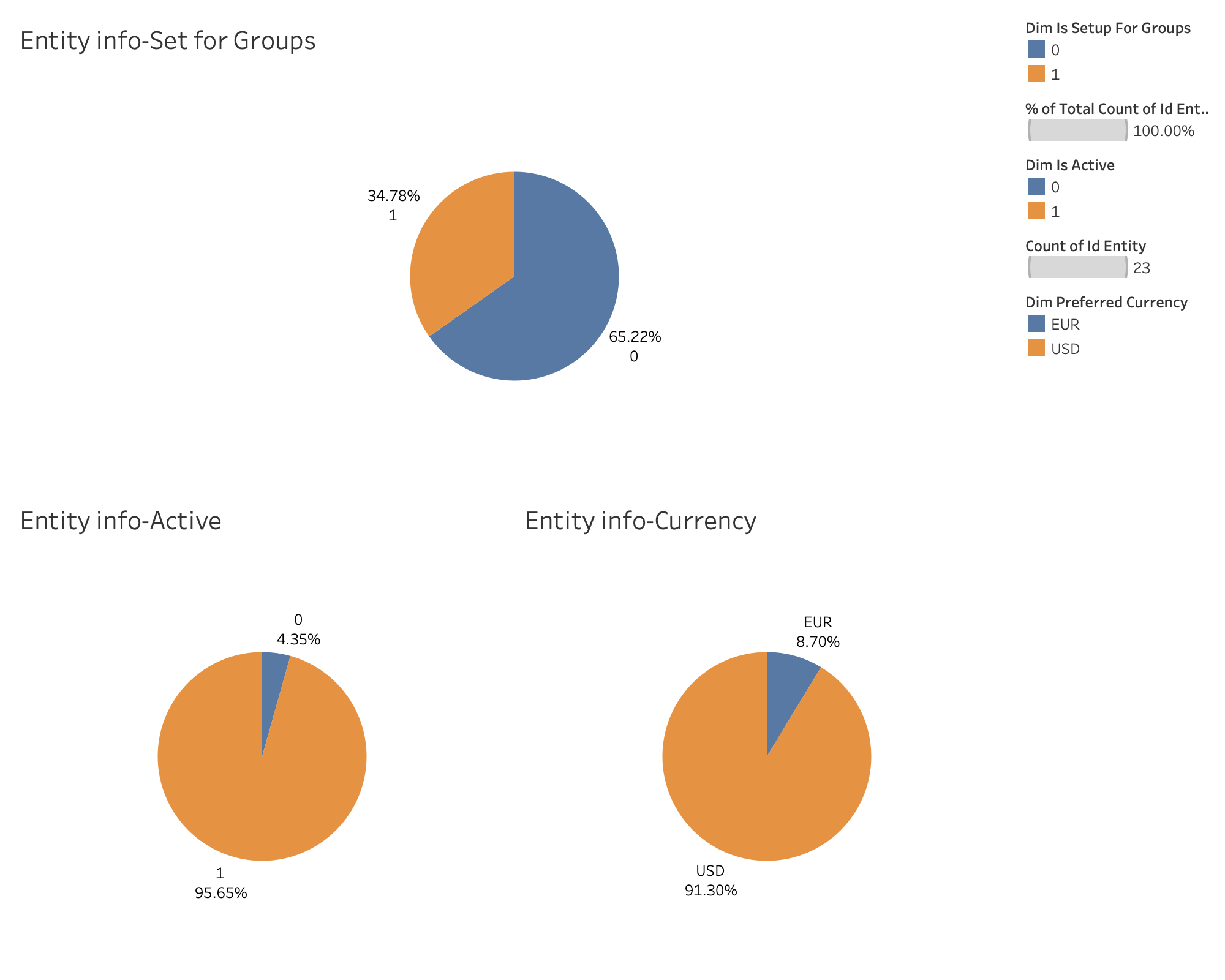
As shown in the diagrams below, ‘strict and moderate’ cancellation policy occupies 85% of all cancellation policies, ‘market place’ occupies 96.9% of all listings, ‘active listing’ occupies 68% of all listings. The recommendation would be to prepare a large number of active listings which are ‘marketplace’ tier in facing competition with hotels. And it can be concluded that customers do not care too much about the cancellation policy, so cancellation policy is not a priority when reaching out to customers.



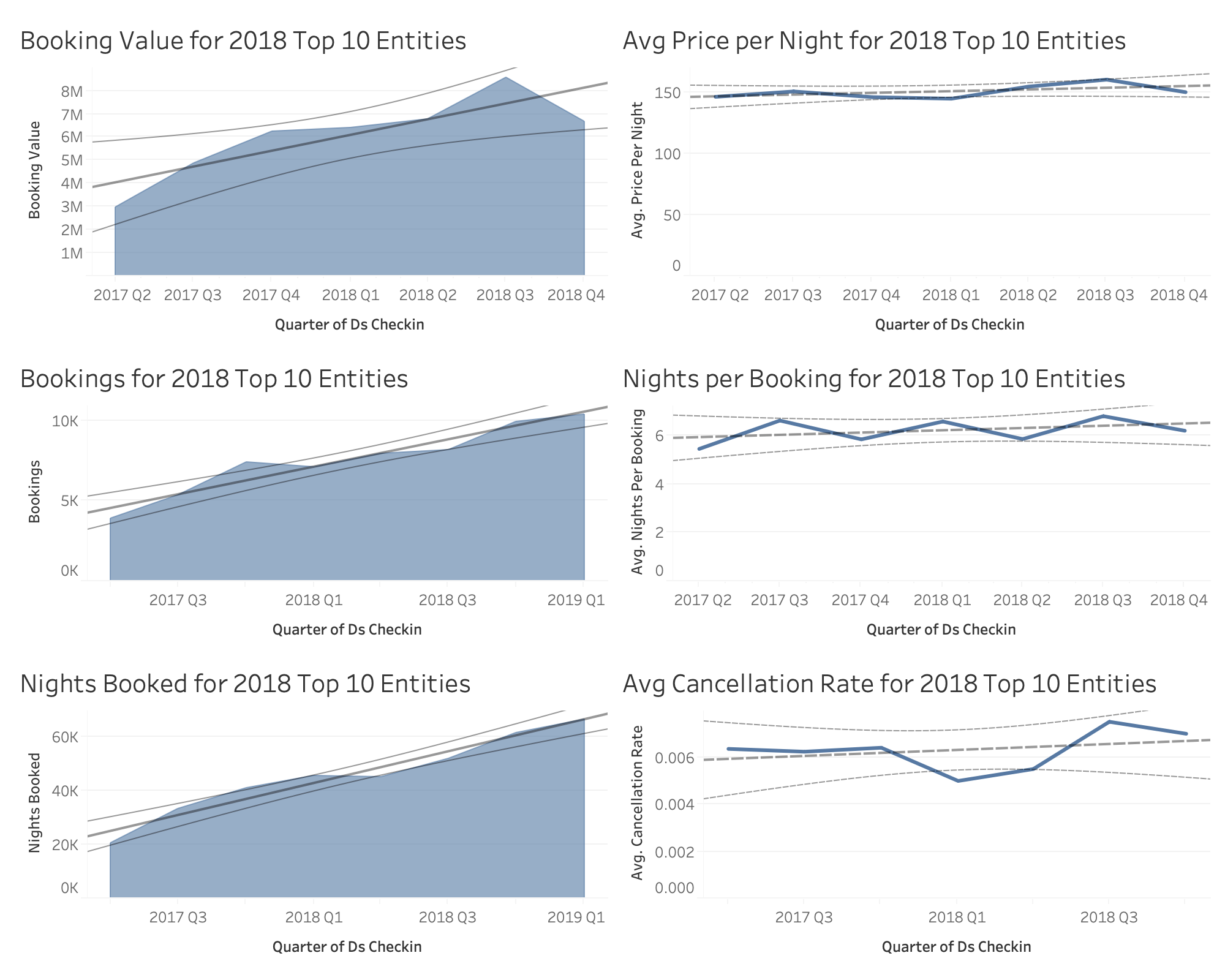
*Entities*

The last part will focus on entities. As shown in the two charts below, when Airbnb is promoting listings on spike days. They should emphasize active companies with 1000-5000 and 5000+ people, prioritize ‘tracking’ payment methods and prioritize ‘USD’ currency. And also, Airbnb should try to sign up entities and set up groups for them because there is still a small portion of entities who are not signed up yet or have not set up for groups yet.





Next, I will start scenario two with high-level metrics first. I manipulated the key metrics for the top 10 entities first and visualized them in Tableau. In Tableau, I figured out the trend along with accuracy metrics the R squared value and p-value ( the R squared value which is higher than 70 is regarded as a good accuracy; p-value which is lower than 1% is regarded as a good accuracy.) Please see the diagrams below.



All key metrics kept increasing from 2017 to 2018. To be detailed,

* booking values kept steadily increasing with the R squared value as high as 71% and p-value as low as 1.7%;
* bookings kept steadily increasing with the R squared value as high as 93% and p-value as low as 0.01%;
* nights booked kept steadily increasing with the R squared value as high as 96% and p-value as low as 0.01%;
* cancellation rate, price per night, nights per bookings kept slightly increasing, however, the R squared value is low and p-value is high. As a result, I may not trust the trends in these three metrics.

Then I will dip into low-level metrics. To be detailed,

* Room demand kept steadily increasing with the R squared value as high as 80% and p-value as low as 0.6%;
* Reservation beforehand kept steadily increasing with the R squared value as high as 81% and p-value as low as 0.5%;
* Price per room per night kept slightly decreasing and cancellation beforehand kept slightly increasing, however the R squared value is too low and p-value is too high. As a result, I may not trust the trends in these two metrics.



In conclusion, since the accuracy of nights booked is much higher than price per night. I would conclude the increase in booking values is due to the increase in nights booked. I would forecast an increase in nights booked in 2019 and entities may reserve rooms even earlier which is a very good trend. Airbnb should seize this business opportunity and get well prepared for increased market demand.