# **Flight Overbooking problem**

Most airlines practice overbooking. That is, they’re willing to make more reservations than the number of seats that they have on an airplane.

## **Why do airlines overbook flights?**

* The simple answer to why airlines overbook is to **maximize profit**. Airlines want to avoid empty seats at all costs, and overselling tickets is a great way to do just that.
* On any given flight a few passengers are likely to be “no-shows”. If the airline overbooks slightly, it still may be able to fill the airplane.
  + “**no-shows**” in the aviation industry are people who did not arrive on time for the flight or canceled at the very last moment.
* Another reason tickets are oversold is because airlines expect a percentage of people to be coming from **connecting flights**.
  + Flights that could very possibly be **delayed** or **canceled**.
* These types of events may lead to flights with empty seats even if all physical seats have been sold.

## **Passenger Bumping**

When a flight is oversold i.e. the number of passengers showing up exceeded the seats on the flight, the airline would **rebook** some customers on a later flight.

* If the flight is much later, the bumped passengers are provided with a **free meal**.
* If it was the next day, they were provided with **overnight accommodation**.

In addition,

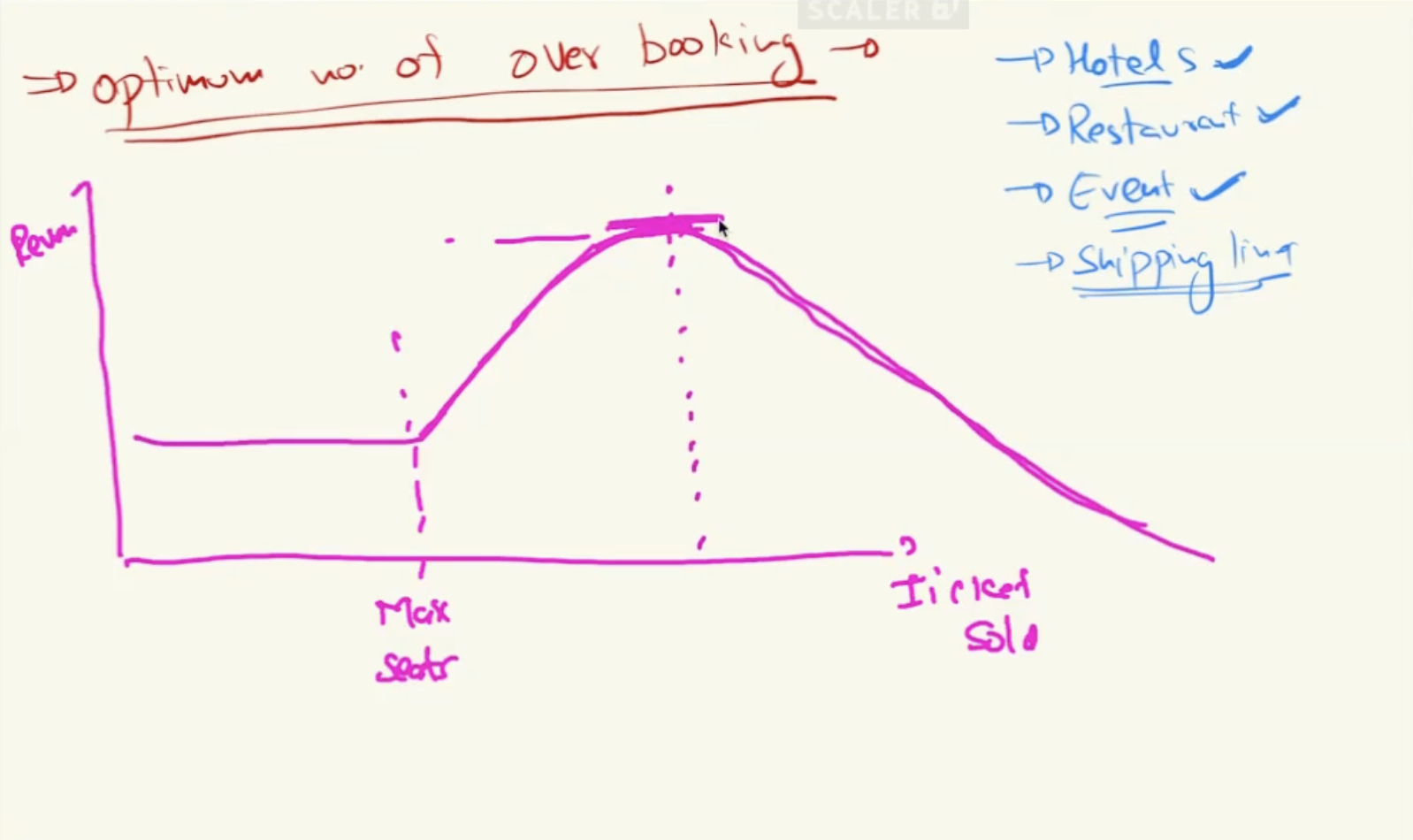
* The airline either pays a **penalty** to each bumped passenger.
* Or they may arrange a **substitute transportation** for the passengers.

## **Problem Statement**

If you are working in the revenue department for an airline and need to find out what is the optimal number of overbooking that can be done for the maximum revenue.

Your goal is to sell as many tickets as possible.

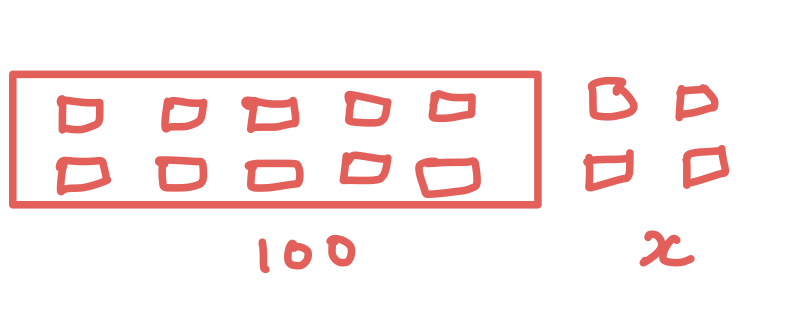
If the overbooking is too high the airline will lose money that it will have to pay to each extra passenger that has been bumped.



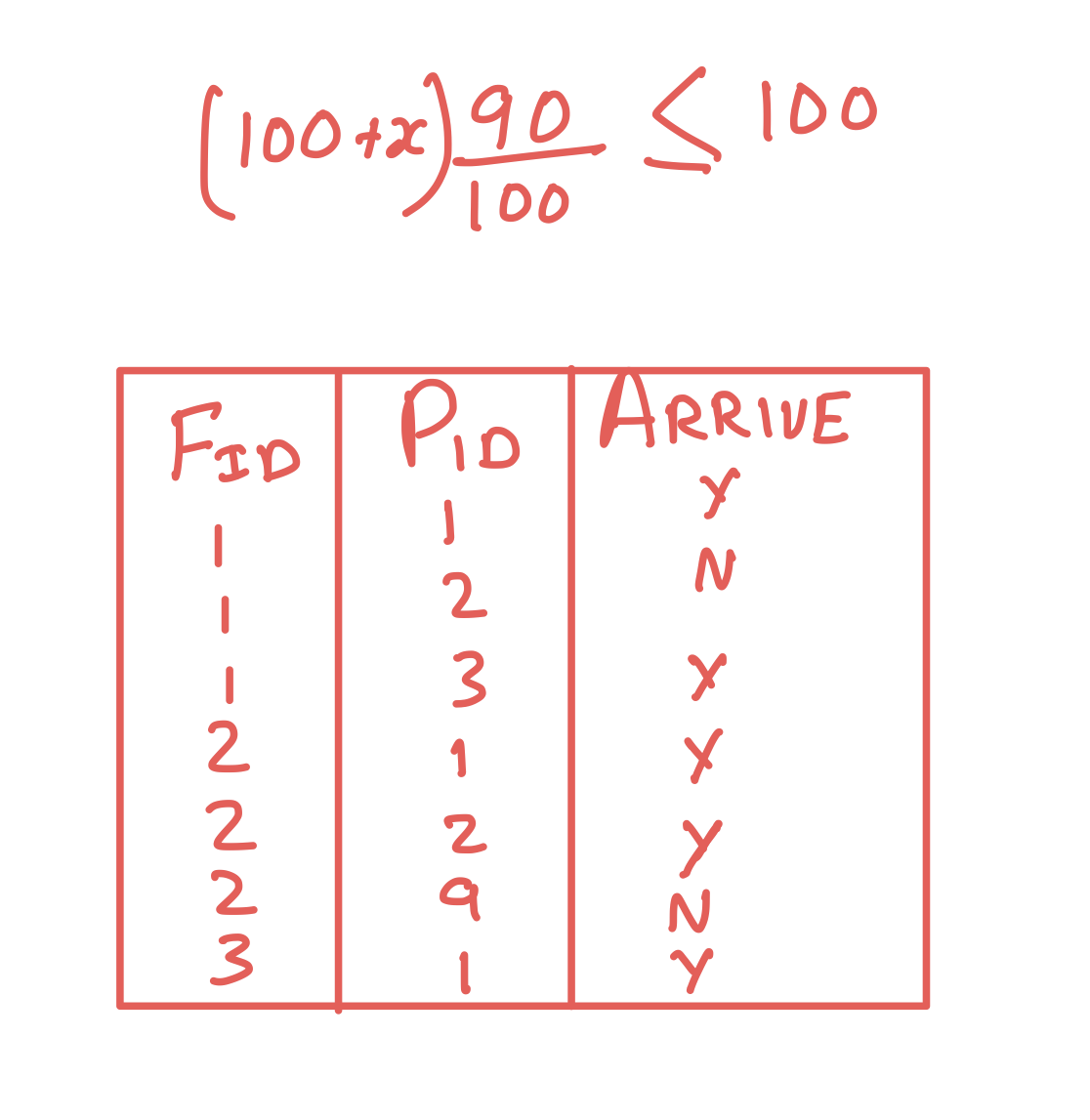
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **Simple probabilistic model :**

Assume the plane has 100 seats, you need to find out the number of extra tickets that can be booked for best results.



You have some historical data that has flight id, passenger id, and the information whether the passenger arrived for that particular flight or not.



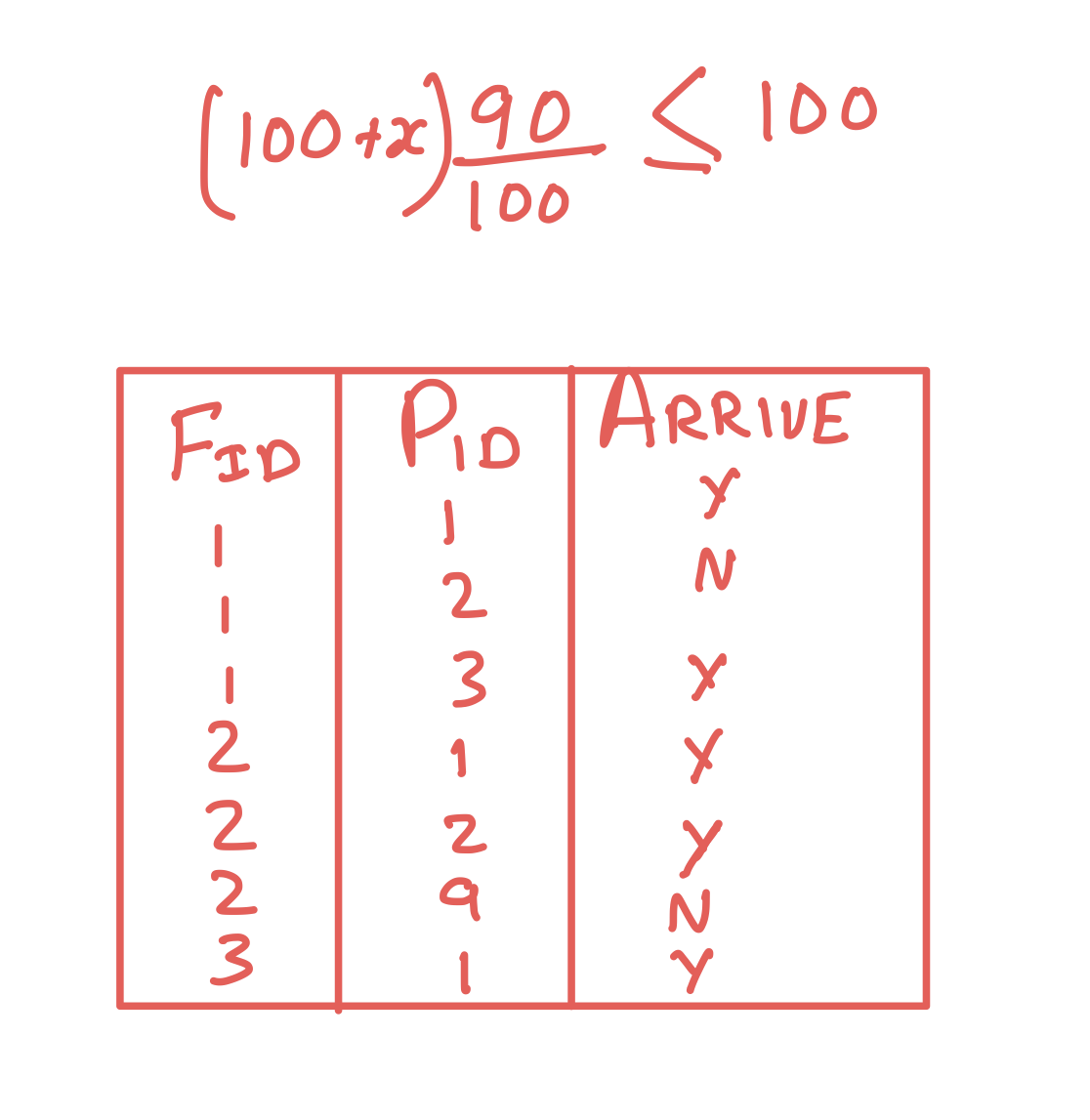
Let's say we have in total 1 lakh rows and in 90k of the cases, the passengers arrived and in 10k they didn't.

* What this means is that there is 90% probability that the passenger will show, i.e. 90000/100000.

So how do we use this data? We know that there are 10% chances of not showing up and we have 100 seats.

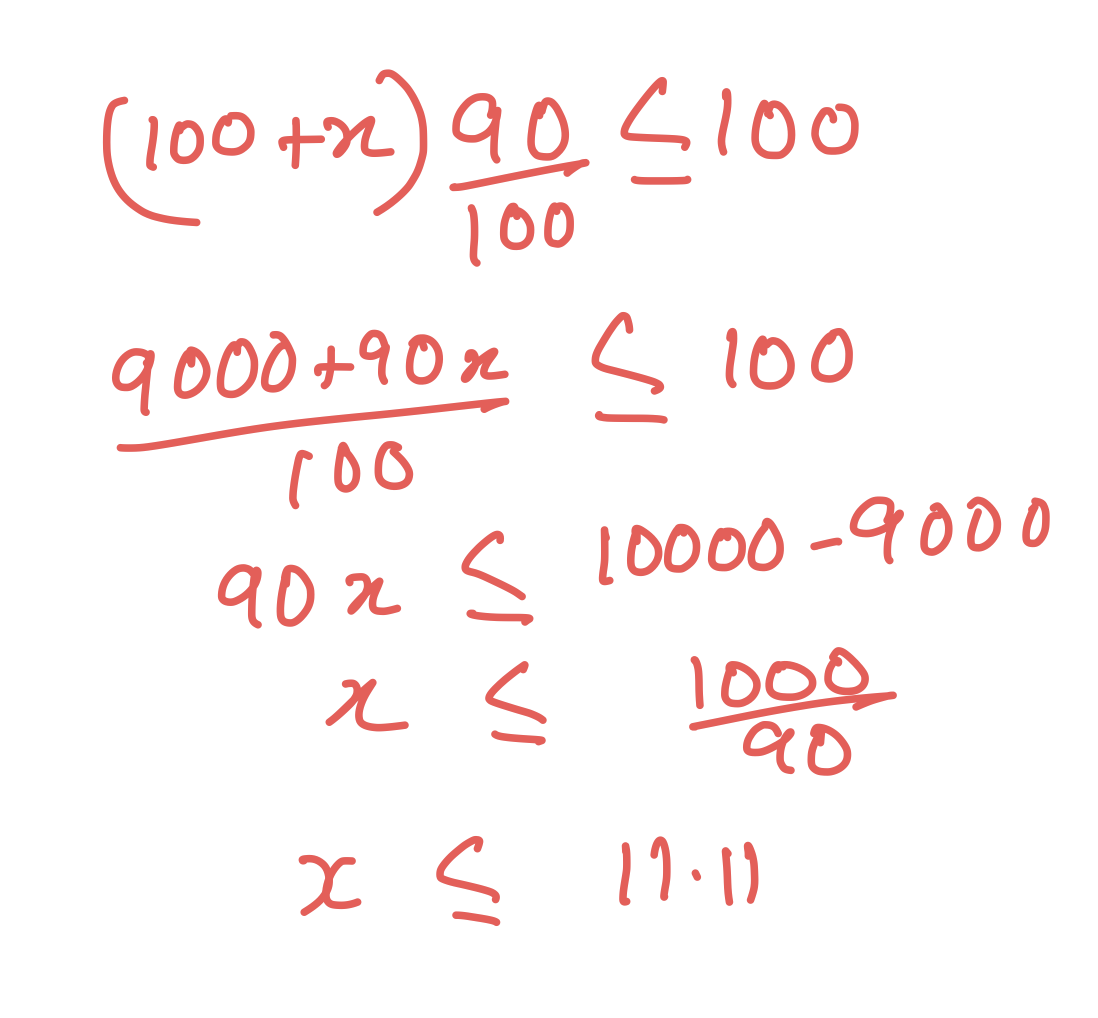
So let's assume that we sell 100+x tickets.

Out of these 100+x tickets how many are expected to show up?



And how many do we want to show up, i.e. less than or equal to 100, because we only have 100 seats.

Now if we solve this equation for x, what do we get?



This was the most simple probabilistic model that tells us based on historical data that if we have 90% chances of passengers arriving we can book additional 11 seats to get a fully booked flight.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **Why does this model sometimes fail?**

For all additional passengers that we do not have a seat for, we have to pay them a penalty, so assume 102 passengers arrive, then we have to pay the extra two passengers.

If the penalty for a particular flight is 5000 or if the penalty is 500000, will you sell the same amount of ticket in both the scenarios?

* If the penalty is zero, theoretically we can sell almost an unlimited number of tickets because there is no risk.
* If the penalty is small then we can sell some additional tickets.
* If the penalty is very large then in that case we may not want to sell any additional ticket.

The above probabilistic model does not take into account the penalty or the risk factor for the booking; it will in all scenarios tell you to book 11 extra tickets.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **How can we improve it?**

The first attempt was a good guess but not a great result.

Let's assume the penalty is 10k, i.e. the net amount given back to the user including everything.

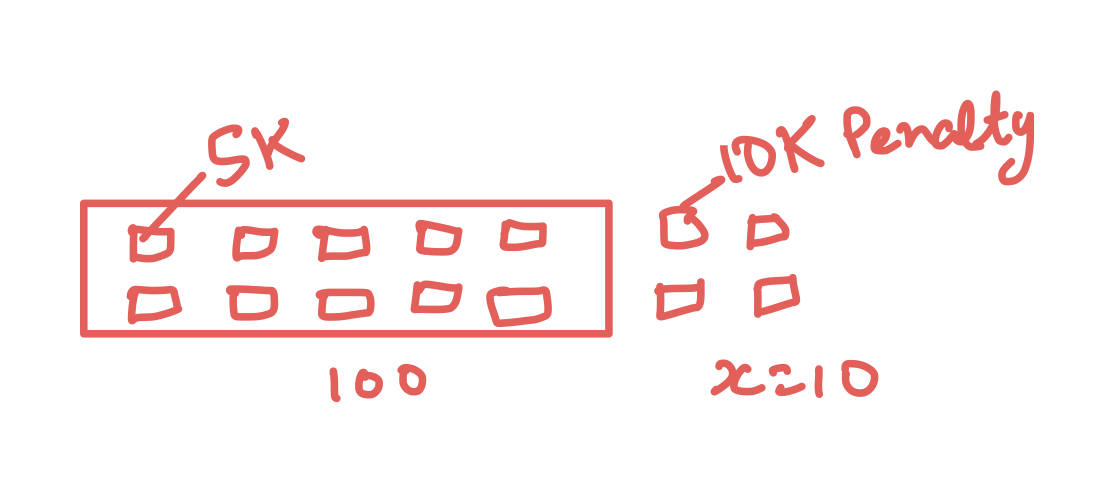
The model should be able to update its recommendation on the basis of the penalty, i.e.

* If the penalty is high it should be able to recommend a lower number of additional tickets.
* If the penalty is low it should recommend a higher number of additional tickets.

We will do the calculations based on 10,000 and then generalize the output for the variables.

* The revenue per seat is 5k
* The penalty is 10k
* Let us also fix the extra number of tickets at 10

For every ticket you sell we get 5,000 and for every extra passenger we lose 10,000.



If zero additional people show up, then

* The revenue is 110\*5000=550000 i.e. five lakh fifty thousand
* No penalty is to be given
* Earning=550000-0

So in this case the additional 50k that we are earning is because of the additional tickets that we sold.

What if 1 extra person shows up. Then,

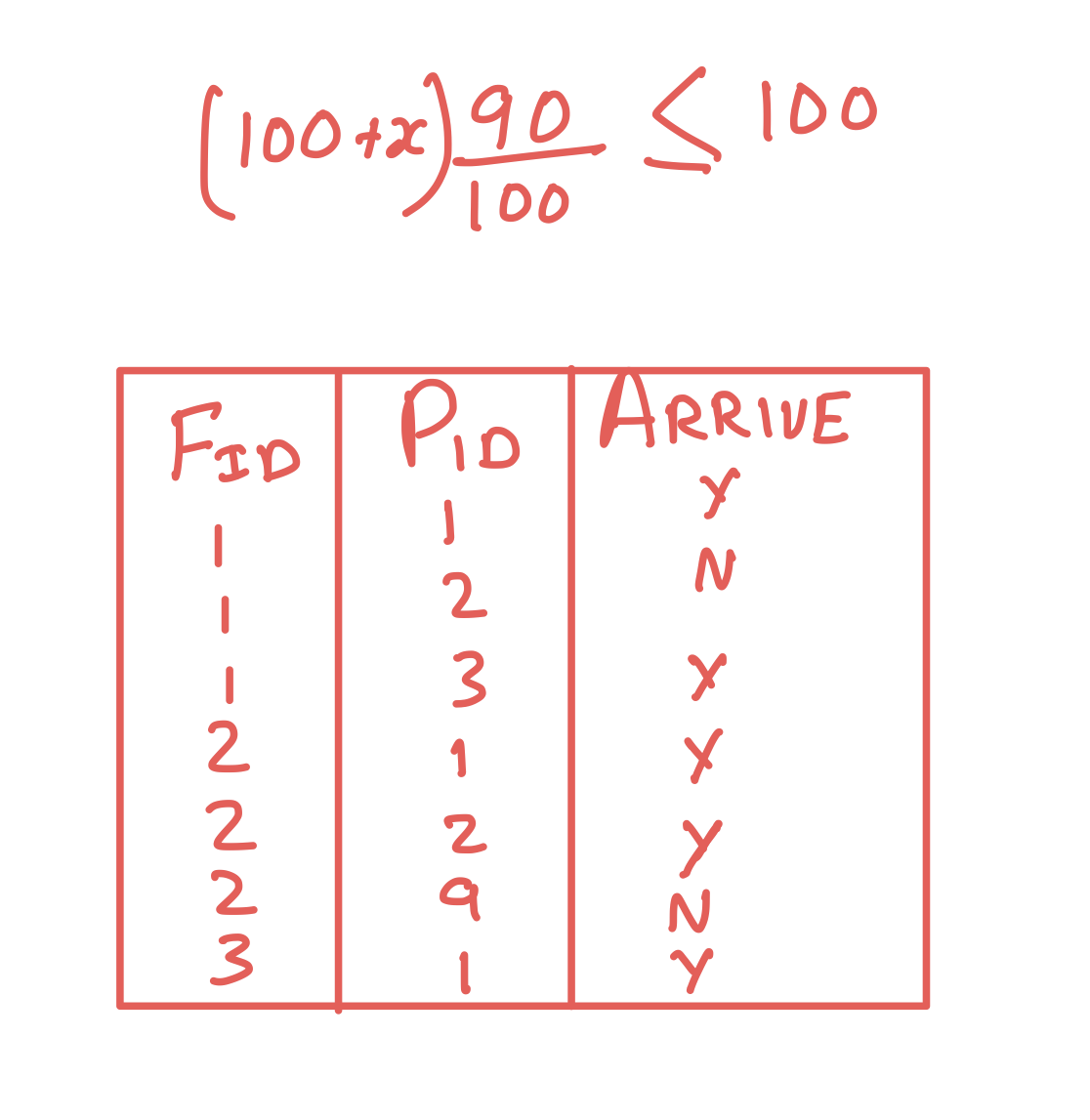
* The revenue is still 550000
* Penalty =10000
* Earning=550000-10000=540000

Let's see what is the probability for this situation...

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

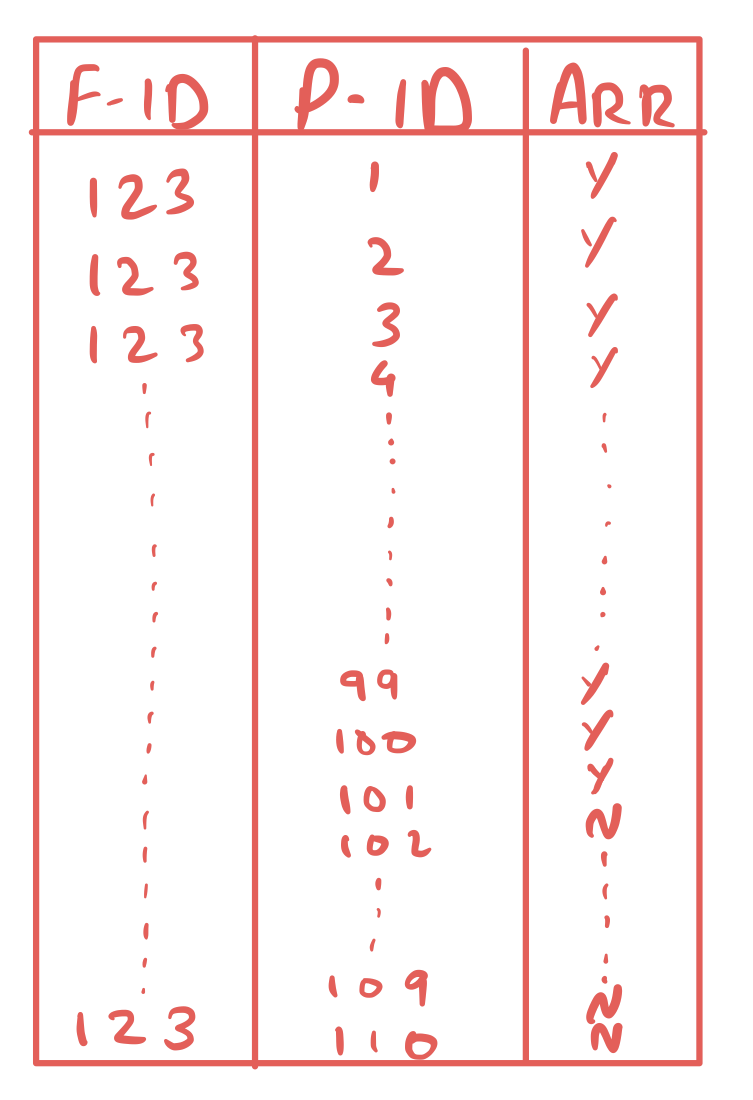
## **Probability of one extra person showing up**

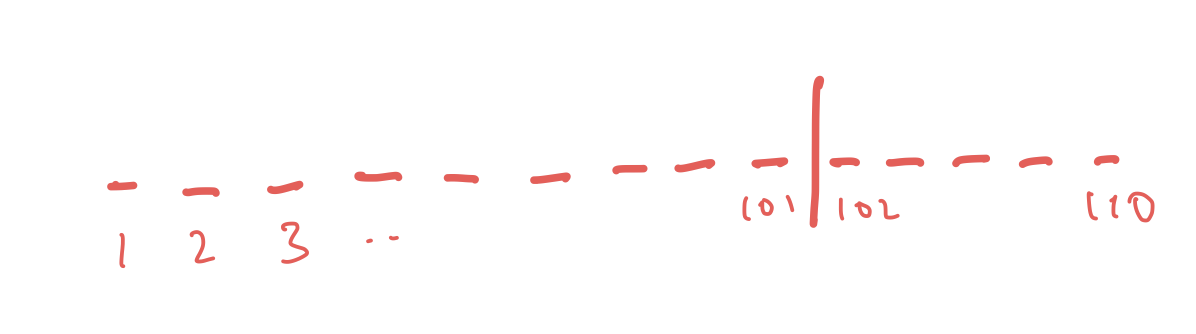
Let's look back to our data -



Now assume we have a certain flight-id (assume it to be 123) and for that we sold 110 tickets.

So according to the data format we have -

￼



Let us write every passenger number 1-110, and we know only 101 people showed up.

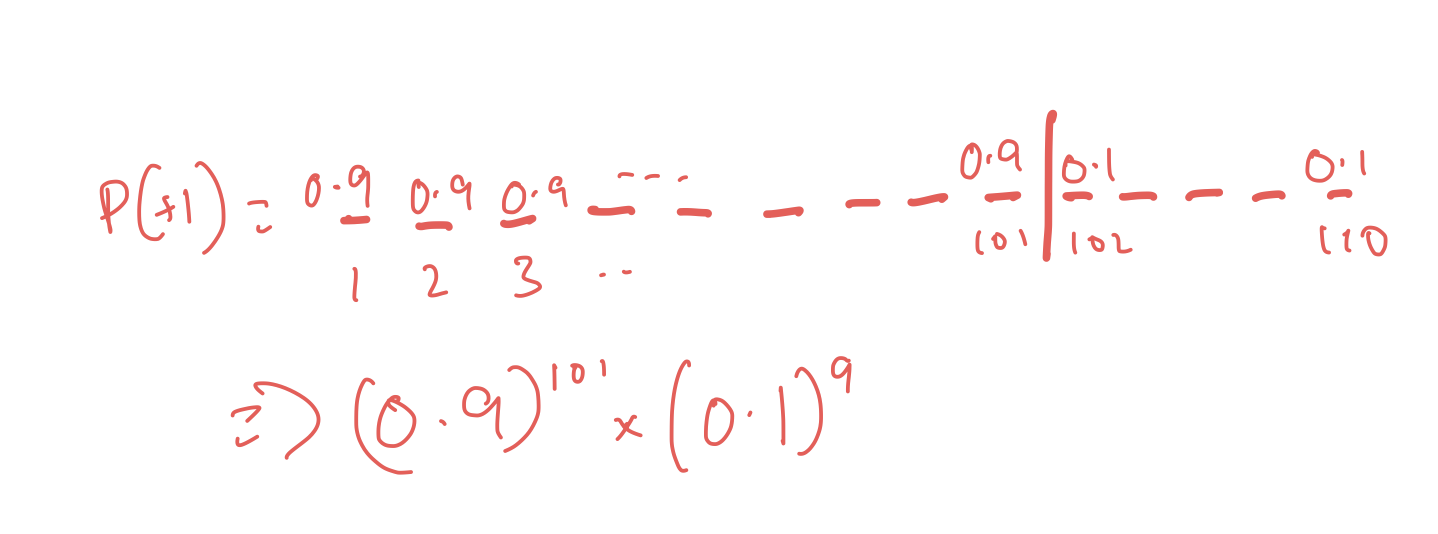
* For now, let us assume P-IDs from 1-100 show up and P-ID 101-110 do not.

Now, what is the probability for each of them who are showing up?

From the past data we know that it is 90% and for not showing up is 10%

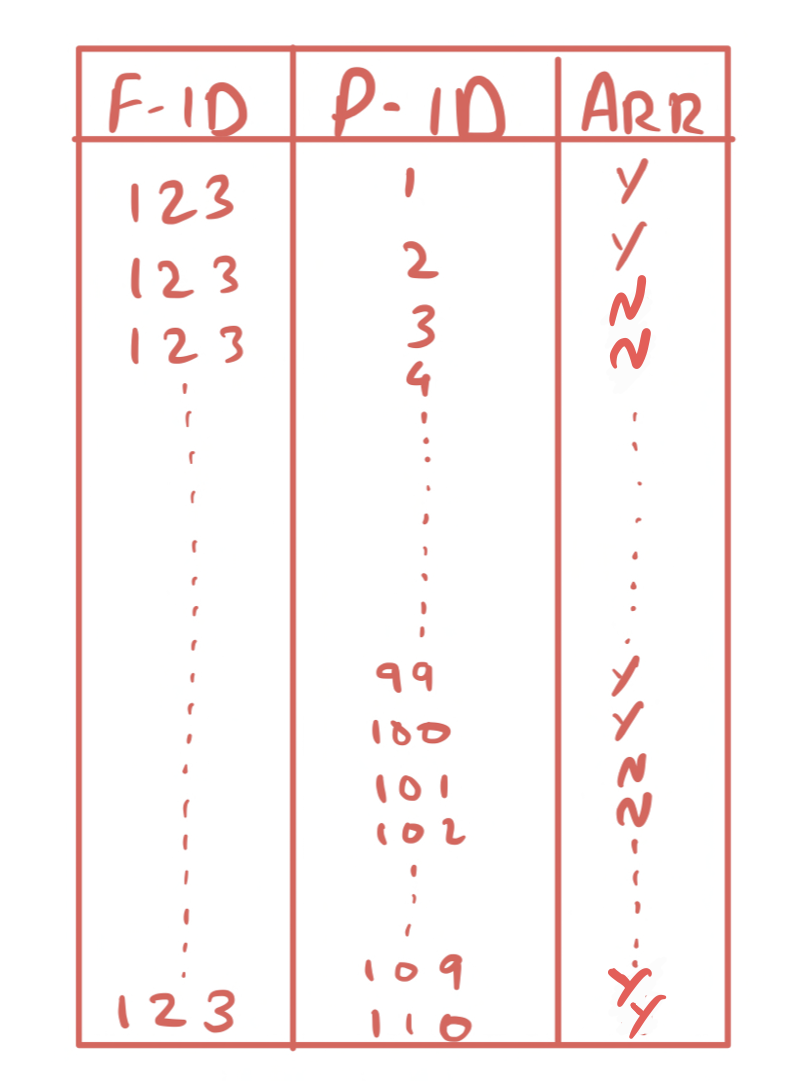
**Let’s calculate the probability of these 101 people showing up -**

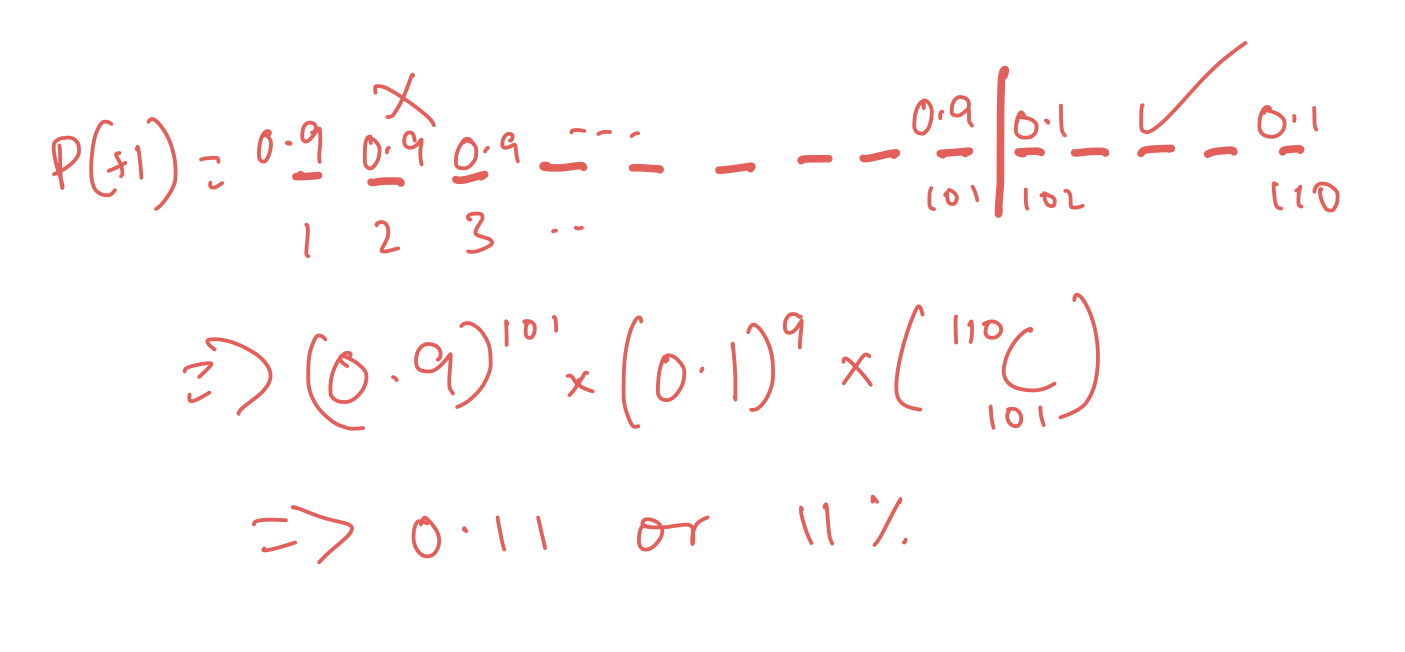
Now for these 101 passengers, we have one extra passenger above our 100 passenger available space, lets see the probability of that happening



So we can multiply all of them together to get the actual probability of these 101 people showing up.

* But there is one more thing that we need to take in consideration, we assumed that P-ID 1-101 showed up
* When calculating these probabilities, which is that it may be the case that ticket number 102 shows up instead of 2 or any other combination of ticket numbers can be the one showing up or not showing up.
  + We only selected a single order or combination of 101 people showing up, but it can be very different.
* The probability for all of them is 90% but still they are very different people any of them could be or not be present.





We can keep shuffling them and the order would change, so we also need to get the combination or the number of arrangements possible.

So out of 110 tickets it can be any 101 passenger ids.

* We need to get the combination to get the probability of one extra person showing up, It is not required for the revenue calculation.
* A better way to understand the need for a combination is that instead of considering tickets, consider them as passengers.
  + Passenger 1, Passenger 2 or 3 and so on. Then the situation when Passenger 1 shows up and Passenger 2 does not will be completely different than 2 showing up but 1 doesn’t.

So when calculated, we get that the probability of one additional person showing up is 11 percent.

And the penalty for an extra person showing up is 10000,

But when do we have to pay this penalty?

* In 11% cases only.

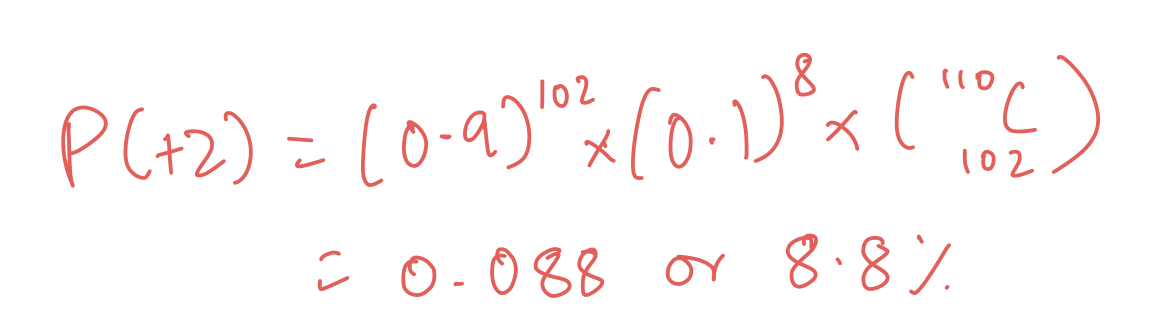
So what is the expected penalty?

The expected penalty is 10k \* 11% i.e. 1100

* So the raw penalty is 10k but the expected penalty is 1100.

**Now, what if 2 additional people show up?**

* We can just use the formula.

****

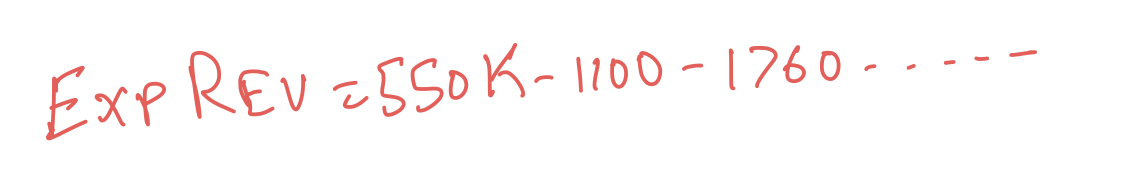
This is 8.8%.

And 8.8% of 20000 is 1760,

So on for the rest of the seats, the percentage of actually showing up goes down and down further.

For 3 extra people it will be 6.1% and so on.

So the total expected revenue for that flight will be the sum of all those results.



# **Airbnb - Analytics Case Study**

**Problem Statement**

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in specific locales.

Over the last five years, Airbnb has witnessed a trend that suggests a correlation between the number of property images associated with a listing and the number of bookings it attracts.

They have also noticed an overwhelming number of listings being redundant (does not attract any bookings), due to the lack of any associated images.

**You need to help the management decide upon a minimum number of images to be made mandatory for a listing that would ensure bookings.**

**Also, come up with an optimal number of images that we can suggest the host to post along with a listing that would attract the most bookings.**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Data:**

The 3 datasets provided are as follows:

1. [**Listings**](https://docs.google.com/spreadsheets/d/1aBrE-8JaUsig4k7eXVWpb-aGF1dVv918/edit?usp=drive_link)**:**
   * Provides a **random data sample** of 500 listings posted by various hosts (including Superhosts) in the last **5 years** from various locations, along with their associated number of property images and the number of bookings they attracted.
   * **Host:** The person who lists the space in their Airbnb account. This is usually the person who owns or lives on the property. They have been segmented into 2 types based on certain criteria they have met -
     + Regular
     + Superhost
   * Variable description:

| **Variable name** | **Description** |
| --- | --- |
| Listing\_Id | Id of the property listing |
| Posting\_Date | Listing posted on a random date in the last 5 years |
| Posting\_Time | UTC time when the listing is posted |
| Location | Location of the property |
| Images | Number of property images associated with the listing |
| Bookings | The number of booking the listing has attracted **until Aug 31, 2019, since it was first posted** |
| Host\_Type | Posted by a regular or a Superhost (Host status as on Aug 31, 2019) |

1. [**Open listings**](https://docs.google.com/spreadsheets/d/1cilMQoWY2ri5d60d9OD3hBMbjcjUUMVK/edit?usp=drive_link)**:**
   * Provides data for **over a year** that shows the number of **open listings** for each date.
   * **Open listings mean the property listings that were available but did not attract any booking by the end of the day.**
   * The listings have been classified according to the number of associated images.
   * Variable description:

| **Variable** | **Description** |
| --- | --- |
| Date | Each date between **Aug 1, 2018, to Aug 31, 2019** |
| Open\_Listings\_0\_2 | Number of listings which were available for the mentioned date but did not attract a booking by end of the day even on that specific date and have 0 to 2 associated property images. |
| Open\_Listings\_3\_5 | Number of listings which were available…date and have 3 to 5 associated property images. |
| Open\_Listings\_6\_10 | Number of listings which were available…date and have 6 to 10 associated property images. |
| Open\_Listings\_11\_15 | Number of listings which were available…date and have 11 to 15 associated property images. |
| Open\_Listings\_16 | Number of listings which were available…date and have more than 16 associated property images. |

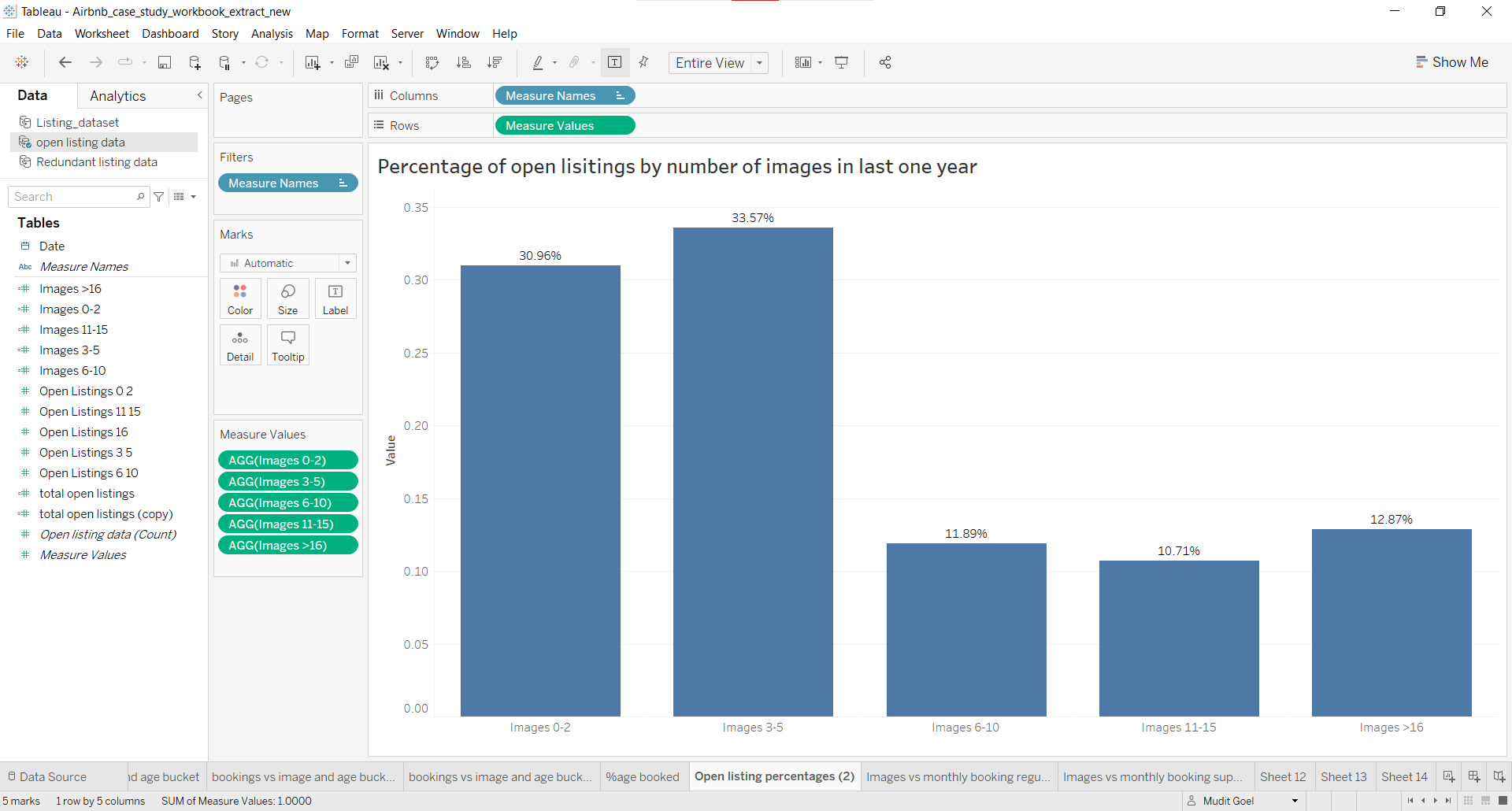
1. [**Redundant listings**](https://docs.google.com/spreadsheets/d/1OqO54nFQ9Afq4wU2UmaYtjCJmBfiW3AQ/edit?usp=drive_link)**:**
   * Provides data as on **August 31, 2019**, for the Total Listings and the Redundant Listings in each category.
   * **Redundant listings here mean the listings that have not attracted even a single booking in the last 1 year.**
   * The categories here are classified according to the associated number of property images.
   * Variable description:

| **Variable** | **Description** |
| --- | --- |
| Property\_Images | Range for the number of associated images for the property, posted along with the listing. |
| Total\_Listings | Total number of listings, with the associated number of property images in the specified range, that are active. |
| Redundant\_Listings: | Number of listings that are active, having the associated number of property images, but did not attract even a single booking in the last 1 year. |

Let’s first analyze the past year's data using **open listings** and **redundant listing datasets.**

**Using Open listing dataset:**

We have data from **1 August 2018 to 31st August 2019** we can analyze and see the number of images that had a higher number of open listings for the past year.



Steps:

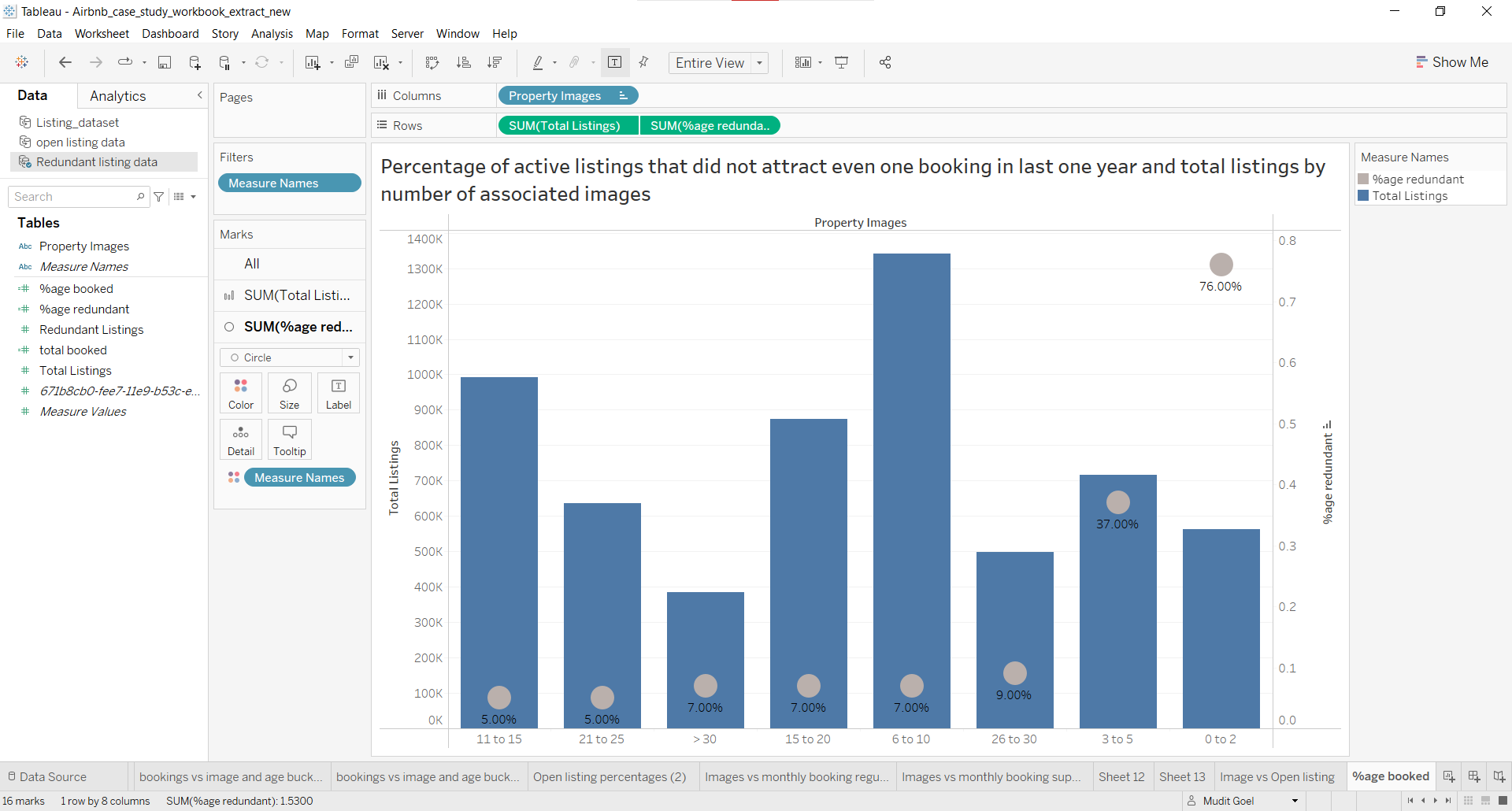
1. Create a calculated field to get the total open listings in the past year
   1. Use formula: **SUM([Open Listings 0 2]+[Open Listings 11 15]+[Open Listings 16]+[Open Listings 3 5]+[Open Listings 6 10])**
2. Divide each open listing by the total open listings to get the percentage
   1. Example: SUM([Open Listings 0 2])/[total open listings]
3. Create visualization.

**Insight from the plot:**

1. We can see that listings with an associated number of images ranging from 11 to 15 had the least open listings.
2. We can also see that listings with an associated number of images ranging from 6 to 10 had the second-lowest number of open listings.
3. Listings with images between 0 to 5 had the highest open listings
4. We can **assume** that images for a listing should be at least 6 to attract bookings and should not be more than 15.

**Using Redundant Listing dataset:**

Here we have Total listings up to **31st August 2019** and redundant listings in the last year.



**Steps:**

* Create a calculated field to get the Percentage of redundant listings
  + Use formula: **ROUND([Redundant Listings]/[Total Listings],2)**
* Create the visualization

**Insight from the plot:**

1. Listings with the associated number of images between **6 to 15** had the highest number of active listings indicating that the hosts prefer to have images in this range.
2. We can also see that the percentage of **redundant listings** in the **past year** was the **lowest** for listings with images between **11 to 15** thus further strengthening our earlier assumption that 11 to 15 should be an optimal number of images to attract bookings and images should range in between 6 to 15.

**Listings Dataset**

Let’s analyze the distribution of bookings across different numbers of images uploaded along with the listings for different host types.

**Ques.**

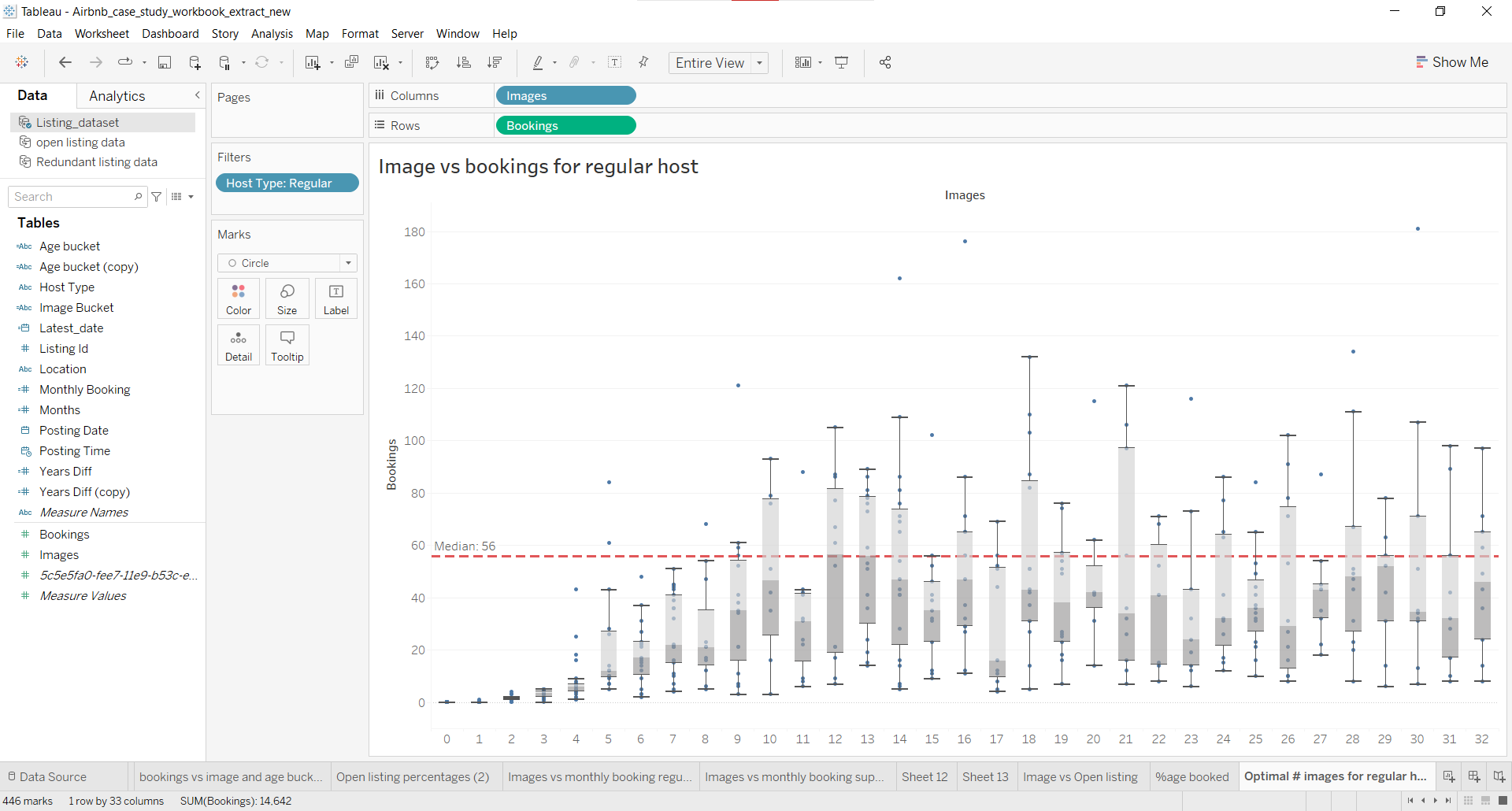
* Which plot would you use if we want to compare the distribution of data points across different values of a categorical variable in a single plot?

**Ans.**

* Box plot because they summarize important statistics of the data using 5 data points which are the minimum, 25th, 50th (median) & 75th percentile and maximum.

Below is a visualization using a **box plot** to compare the distribution of bookings for the different numbers of images for a **regular host**.

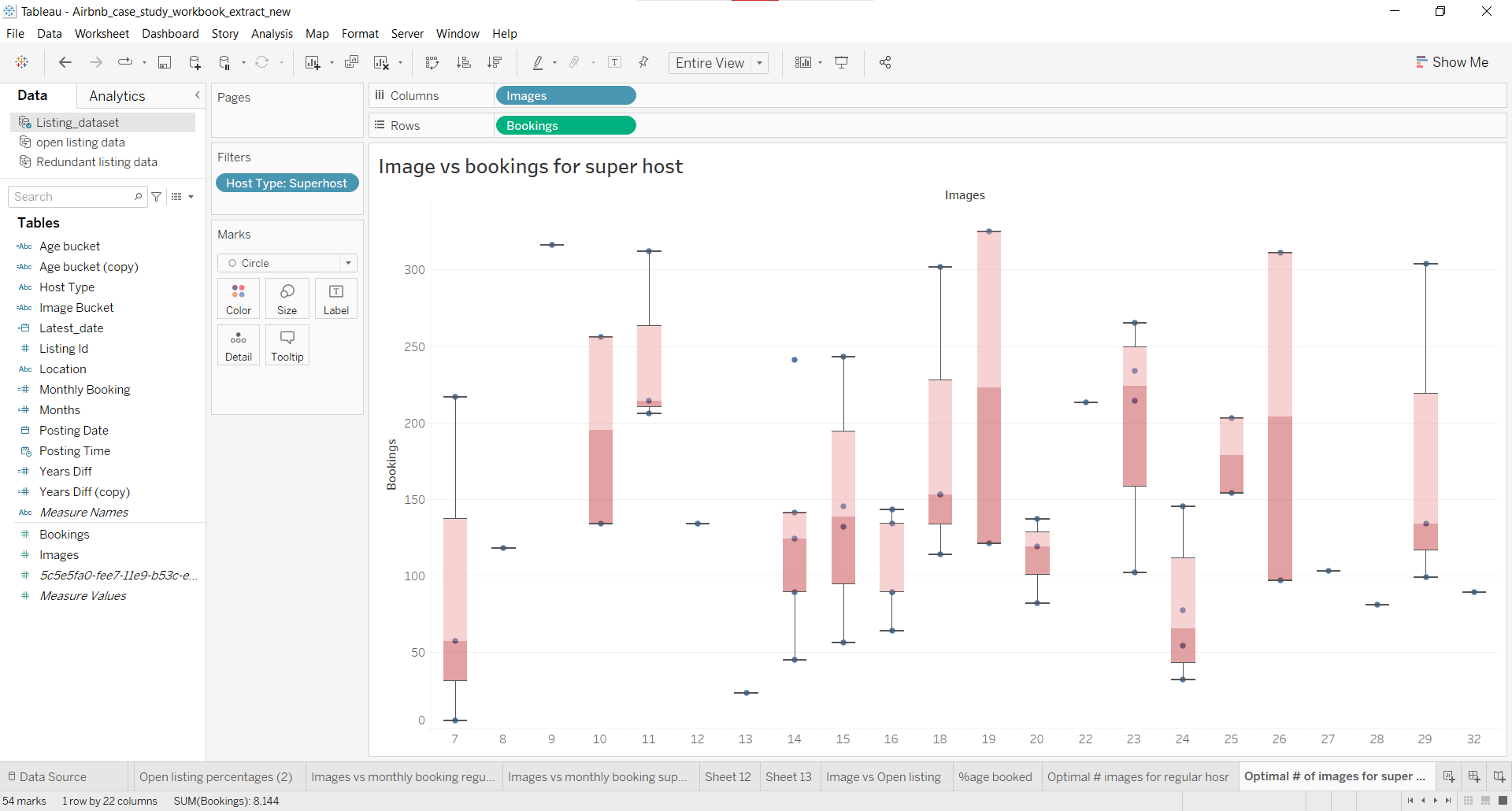
**Note:** We're looking into different host types separately because of their different nature.



**Insights from the plot:**

* We can see that the **median** number of bookings for listings with 12 to 13 images is the highest.
* Median is a simple metric that is not affected by outliers and we can use it to decide on an optimal number of images that would maximize bookings for a host.
* Thus we can say that the optimal number of images for a regular host would be 12 or 13 and backing our prior analyses that images between 11 to 15 are optimal.

Below is a visualization using a box plot to compare the distribution of bookings for the different numbers of images for **Super Host**.



**Insights from the plot:**

* The median value for listings with the number of images equal to 23 is the highest but we also have to keep in mind that 23 images are a lot to click and upload for a host.
* The median value for listings with the number of images equal to 19 is the Second highest but we can see that there are only 2 listings (number of blue dots).
* Keeping the above points in mind we can say that listings with Superhost type should have 11 images for a listing that will maximize their bookings and also because it has the third highest median value.

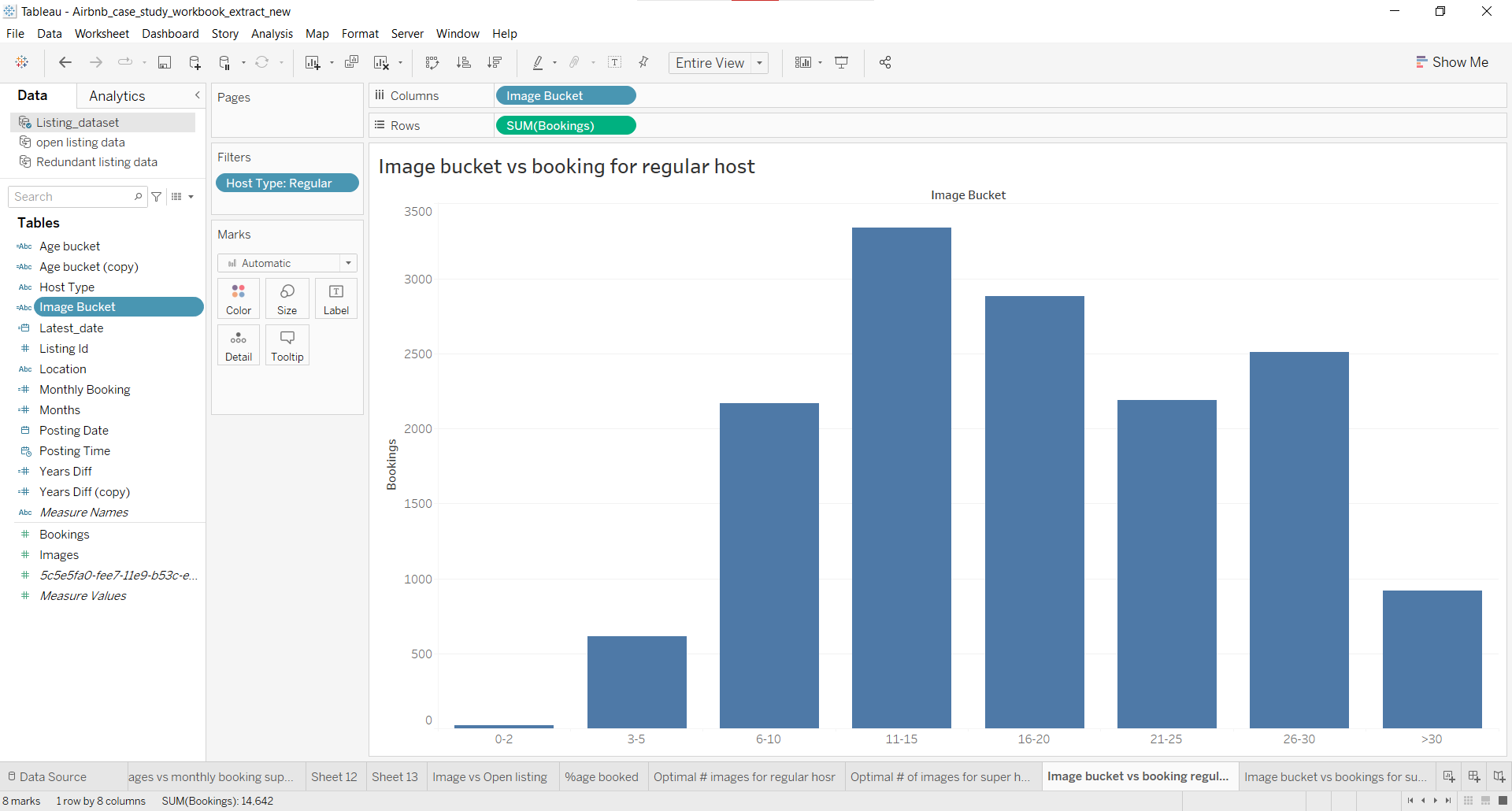
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Let us analyze the data further to back the above findings.**

We have the number of images for different listings. Let’s bin them as follows and analyze the total number of bookings made for each bin for different host types:

* 0 to 2
* 3 to 5
* 6 to 10
* 11 to 15
* 16 to 20
* 21 to 25
* 26 to 30
* >30

**Note:** We are binning them this way so that the number of image buckets (Bin) is consistent across datasets



**Steps:**

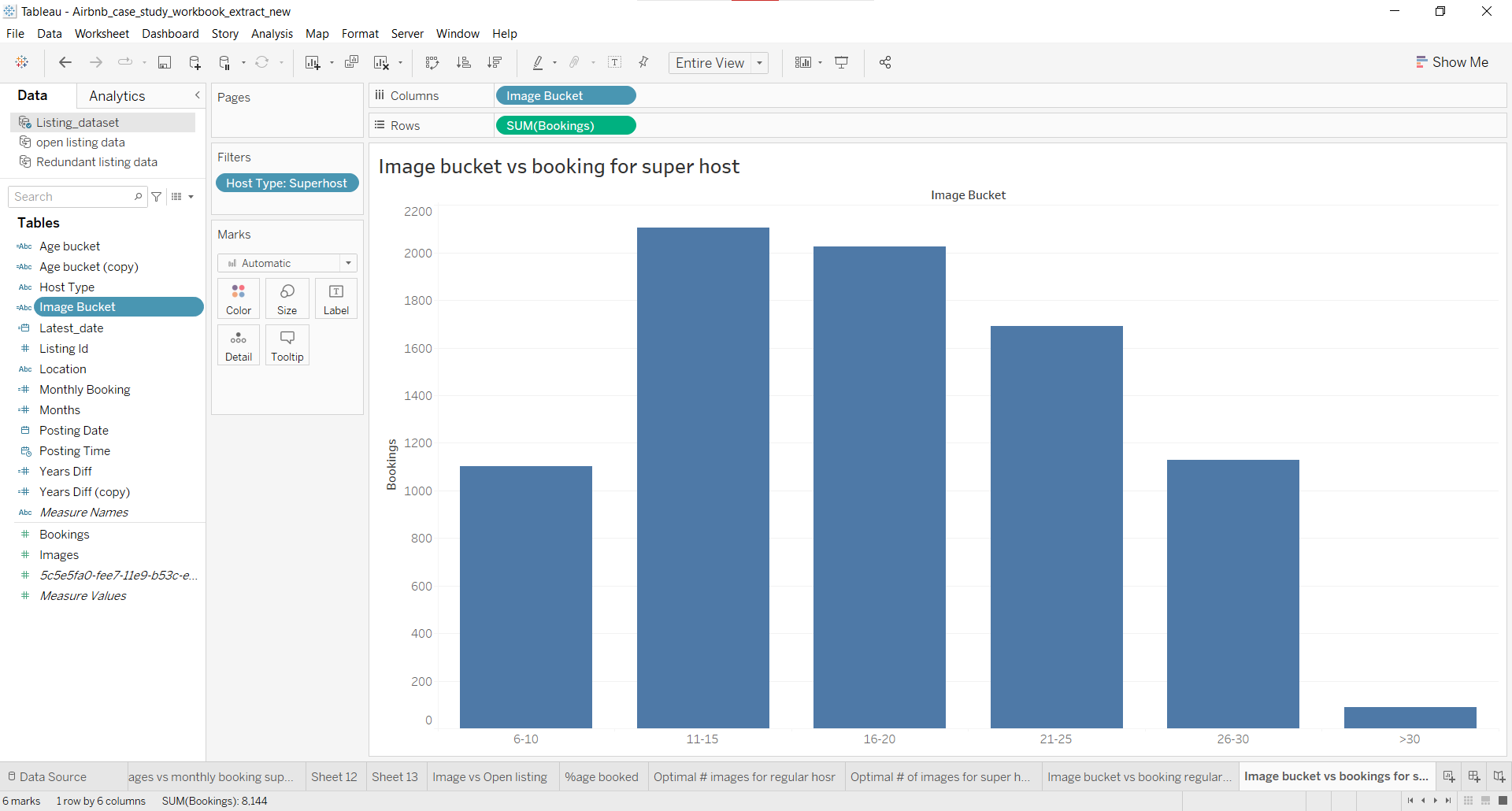
* Create a calculated field named **image bucket** with the following formula:

| IF [Images]<=2 THEN "0-2" ELSEIF [Images]>2 AND [Images]<=5 THEN "3-5" ELSEIF [Images]>5 AND [Images]<=10 THEN "6-10" ELSEIF [Images]>10 AND [Images]<=15 THEN "11-15" ELSEIF [Images]>15 AND [Images]<=20 THEN "16-20" ELSEIF [Images]>20 AND [Images]<=25 THEN "21-25" ELSEIF [Images]>25 AND [Images]<=30 THEN "26-30" ELSEIF [Images]>30 THEN ">30" END |
| --- |

* Create visualization

**Insights from the plot:**

* From the above plot, we can see that listings with a regular host with images between 11 to 15 had received the highest number of bookings which further strengthens our findings from the box plot for a regular host.



**Insights from the plot:**

* We can see that listings with the super host type tend to upload more than 6 images.
* Here too we can see that listings with 11 to 15 images received the most bookings, supporting the findings from the box plot for super hosts.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In the dataset, we have the date when a listing was posted and our data is till 31st August 2019.

Using this we can calculate the property age in years and bin them as follows:

* <1 (property is less than 1 year old)
* 1-2 (Property age is between 1 to 2 years old)
* 2-3 (Property age is between 2 to 3 years old)
* 3-4 (Property age is between 3 to 4 years old)
* 4-5 (Property age is between 4 to 5 years old)

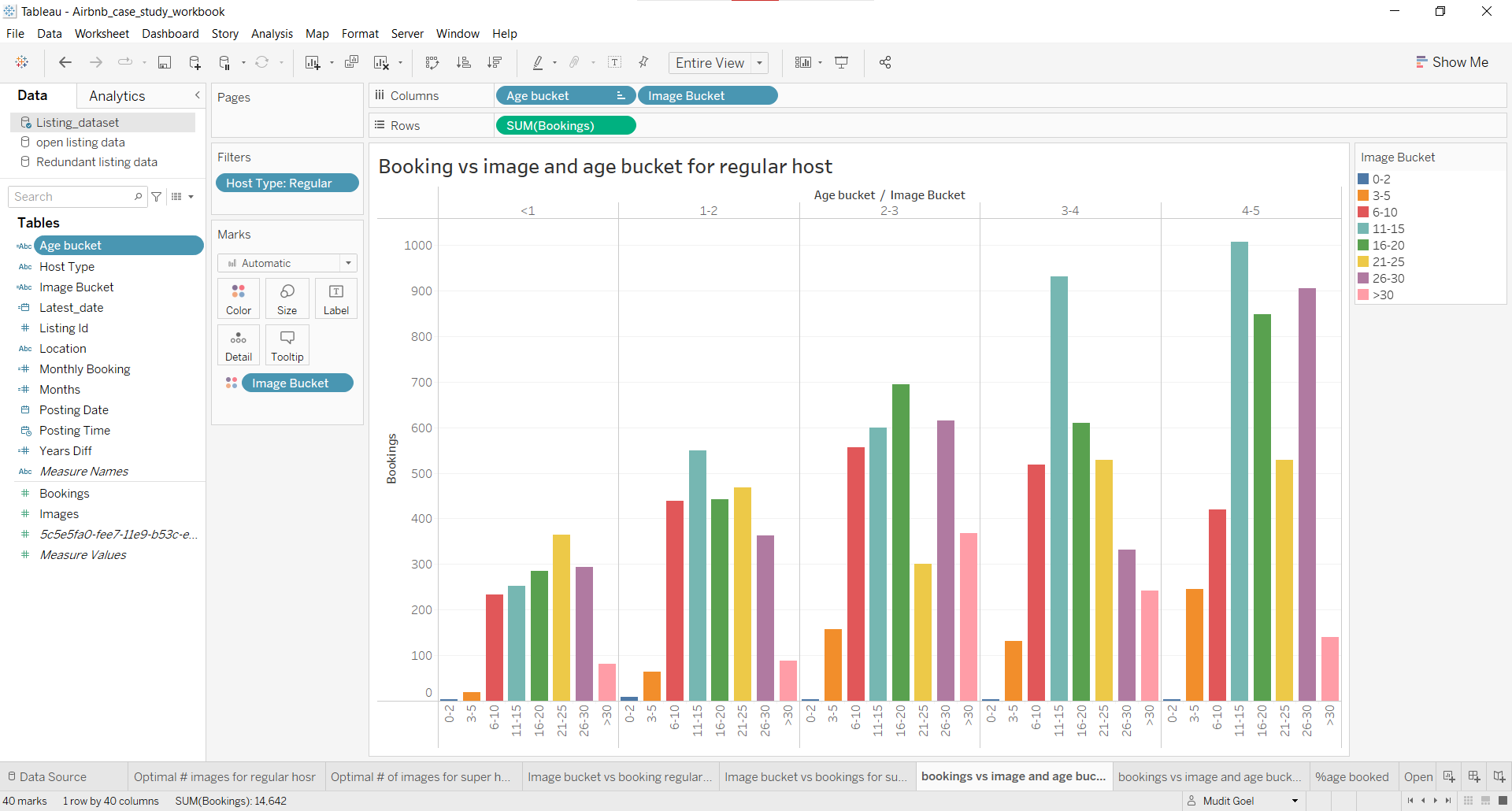
**Note:**

* To calculate the number of years for a listing from the posting date to 31st August 2019 we are considering *August to August* as a year.
* To do that we count the number of months and divide it by 12 to check whether it’s been a year or not from 31st August 2019.

**Example:**

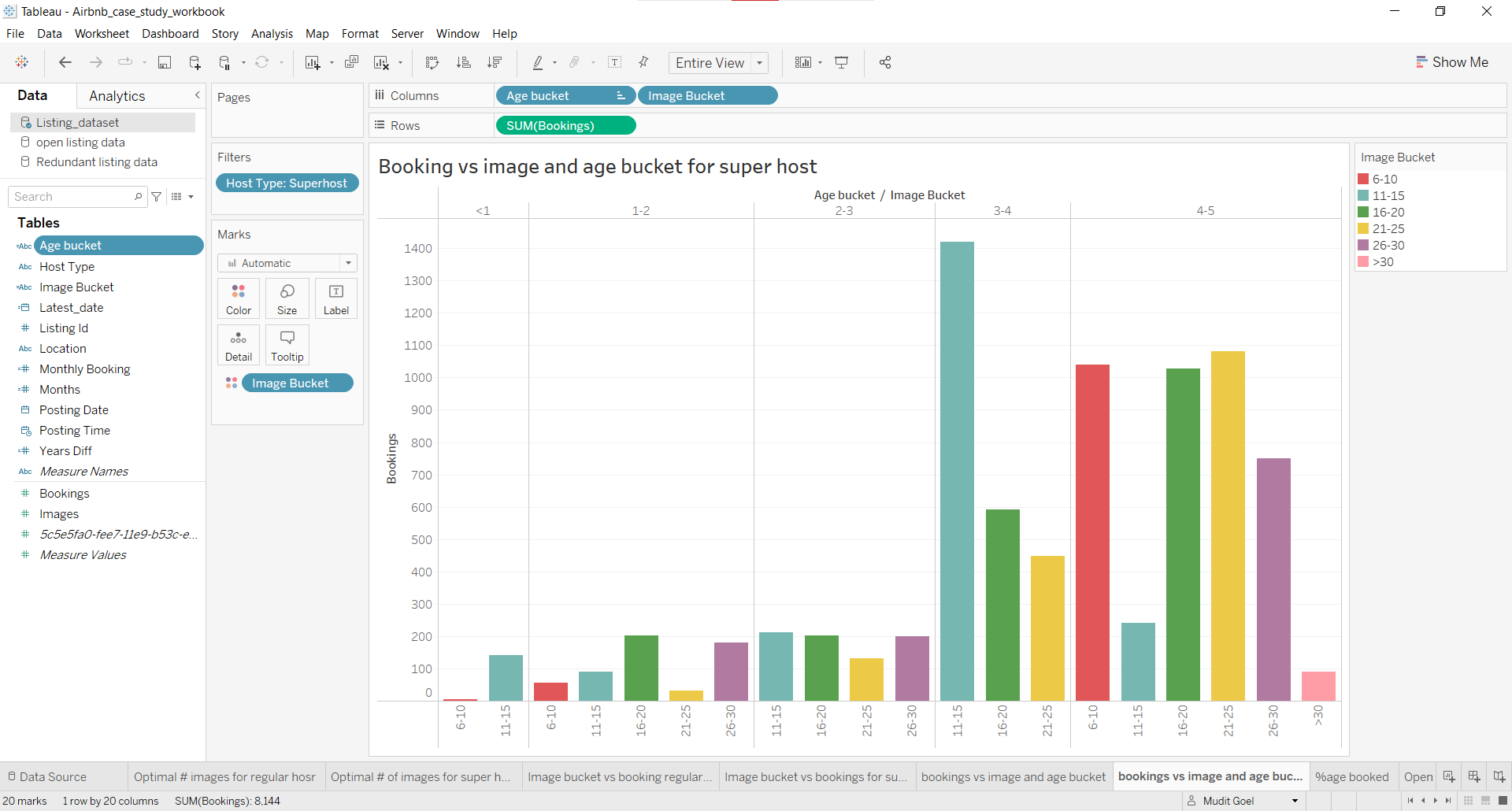
* If the posting date is 1-9-2018 then there are 11 months till August 2019 and when we divide it by 12, we will have a value of less than 1 hence the listing is less than a year old.
* If the posting date is 1-9-2017 now here we have 23 months till August 2019. Hence when we divide it by 12, the value would be less than 2 and we would classify the listing to be 1 to 2 years old.

Now using this let's analyze the total bookings for different Ages and image bins for regular hosts.



**Insights from the plot:**

* We can see from the above plot that listings with regular hosts with 4 to 5, 3 to 4 and 1 to 2 years had the highest number of bookings for listings with 11 to 15 associated images.
* Listings with 2 to 3 and less than 1 year had the second highest and third highest number of bookings with 11 to 15 associated images.
* Thus this further strengthens our analysis that the optimal number of images for the regular host should be between 11 to 15.



**Insight from the plot:**

* From the above plot, we cannot conclude the same as we did for regular hosts.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Using the **listing dataset**, let's analyze and back up the above finding.

Let’s find out the number of monthly bookings each image bucket attracts with a condition that each listing should at least have one booking per month for different host types.

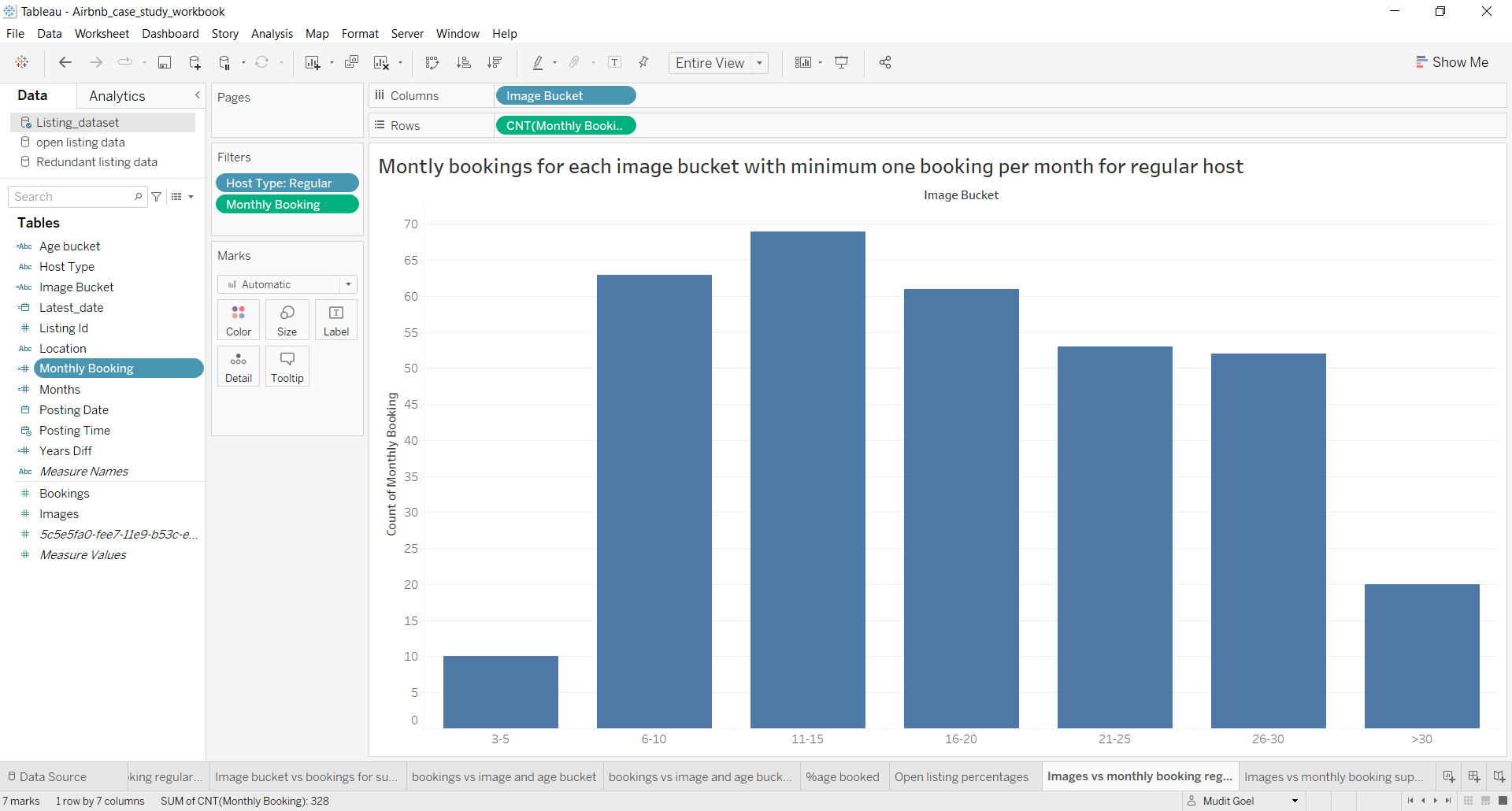
To do this we can calculate the listing age in months till 31st August 2019, and we already have total bookings for each listing so we can divide the total bookings by months to get monthly bookings for each listing.

Steps to get listing age in months -

* Get the number of days between the posting date and 31st august 2019
* Divide it by 30 to get the total months

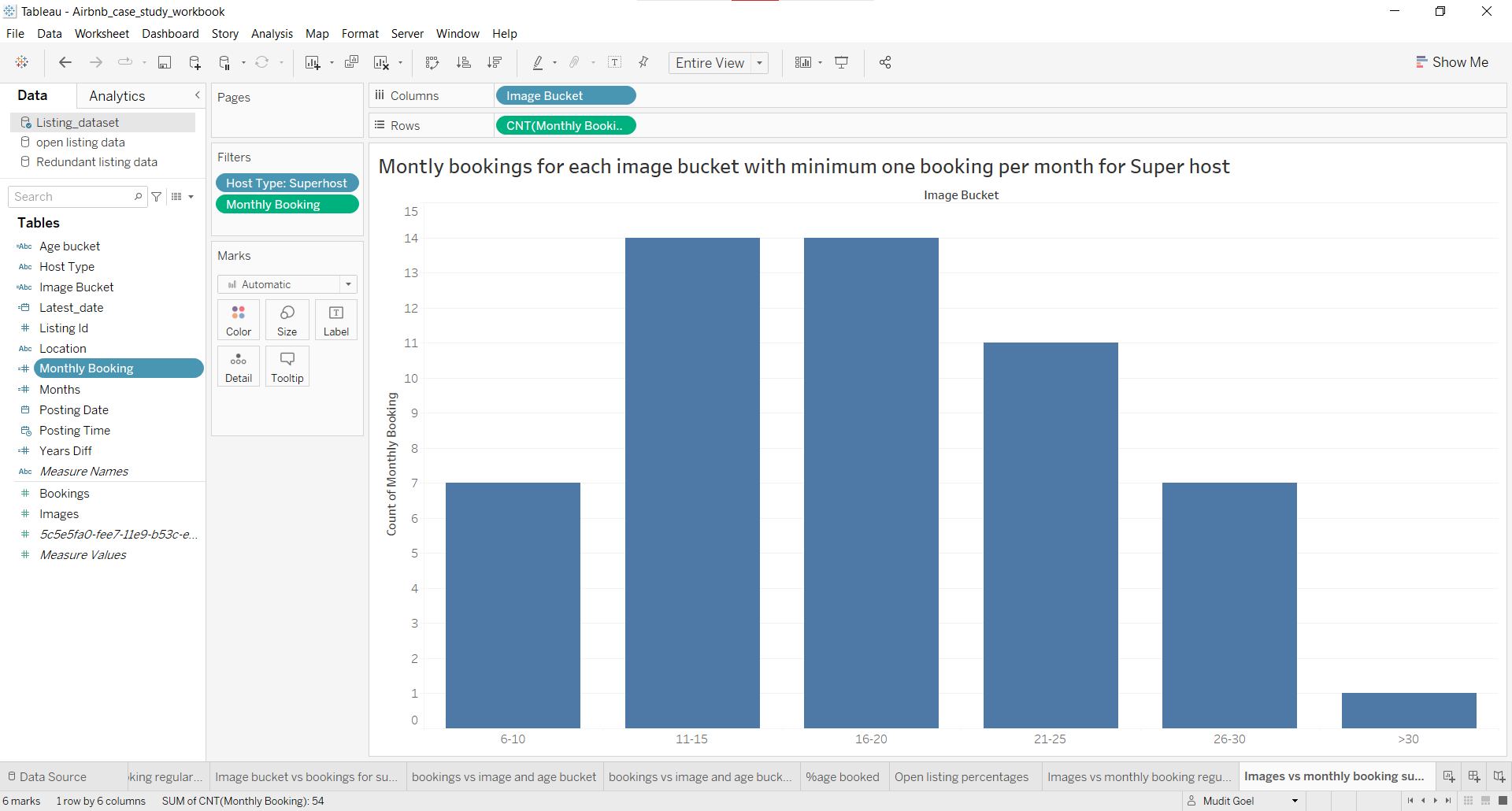
**Example:**

* If the posting date is 31-8-2014 then the number of days till 31st August 2019 is 1826 (365\*5+1[current date]).
* When we divide by 30, we will have 60.86. Round it off, it will be 6.



**Insights from the plot:**

* From the above plot, we can see that the listings with images between 6 to 10 have received the second-highest number of bookings.
* Given that we are suggesting images between 11 to 15 to be an optimal number of images, we can say that for the regular host to attract at least one booking per month they should have a minimum of 6 images. This would ensure that they would have a minimum of 60 bookings in 5 years.



**Insights from the plot:**

* We are already suggesting the optimal number of images to be between 11 to 15. Hence to get the minimum number of images for a listing, we would focus on a number of images less than 11.
* So from the plot, we can see that there are 7 listings with images between 6 to 10 with at least one booking per month.
* Hence we can suggest that for the super host to attract at least one booking per month they should have a minimum of 6 images. This would ensure that they would have a minimum of 60 bookings in 5 years.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Conclusion:**

From the above analysis, we can conclude that -

* The optimal number of images to be suggested to the host in order to maximize bookings is between **11 to 15**.
* The minimum number of bookings to be suggested to the host to attract at least one booking per month would be **6**.