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 an intriguing 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.

Task:

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

Case Study related Resources - link

Solution Approach:

We'll analyze the data and based on the insights that we get, we will suggest an optimal and minimum number of images along with supporting evidence and assumptions that we made during the analysis.

Data:

Note: This is not real data and is only provided to you for the purpose of this case.

The 3 datasets provided are as follows:

1. **Listing:**
    - 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)

2. **Open listing:**
    - 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.

3. **Redundant listing:**
    - 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:

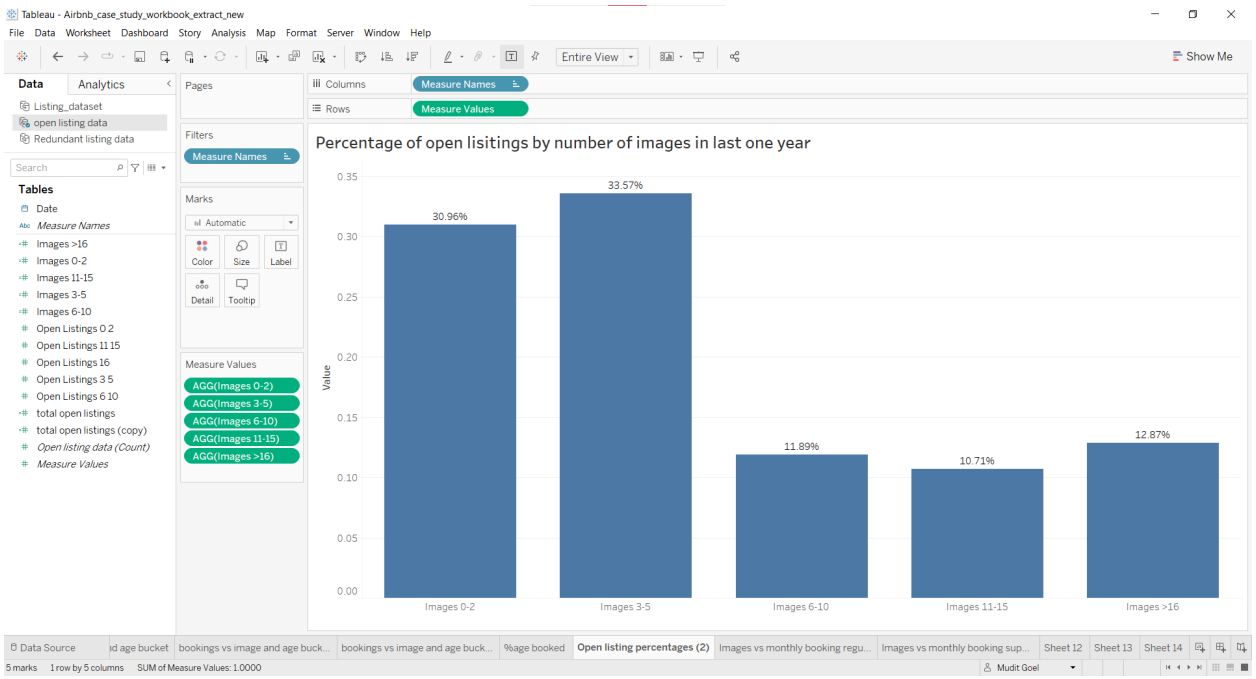
Variable	Description
Property_Images	Range for the number of associated images for the property, posted along with the listing.
Total_Listings	Total number of listings, with the associated number of property images in the specified range, that are active.
Redundant_Listings:	Number of listings that are active, having the associated number of property images, but did not attract even a single booking in the last 1 year.

Solving the problem:

Let's first analyze the past year's data using **open listings** and **redundant listing datasets.**
Tableau Workbook with solution**: Link**

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
    a. 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
    a. Example: SUM([Open Listings 0 2])/[total open listings]
3. Create visualization.

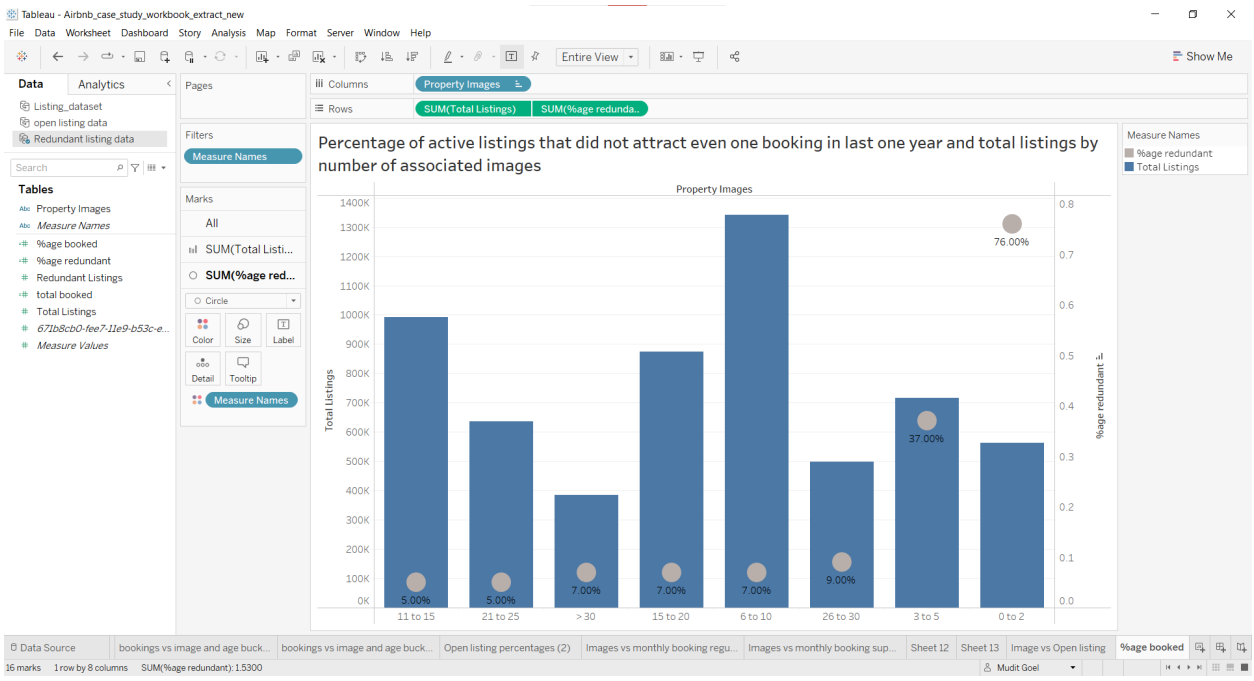
Insight from the plot:

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.

Ask Learners

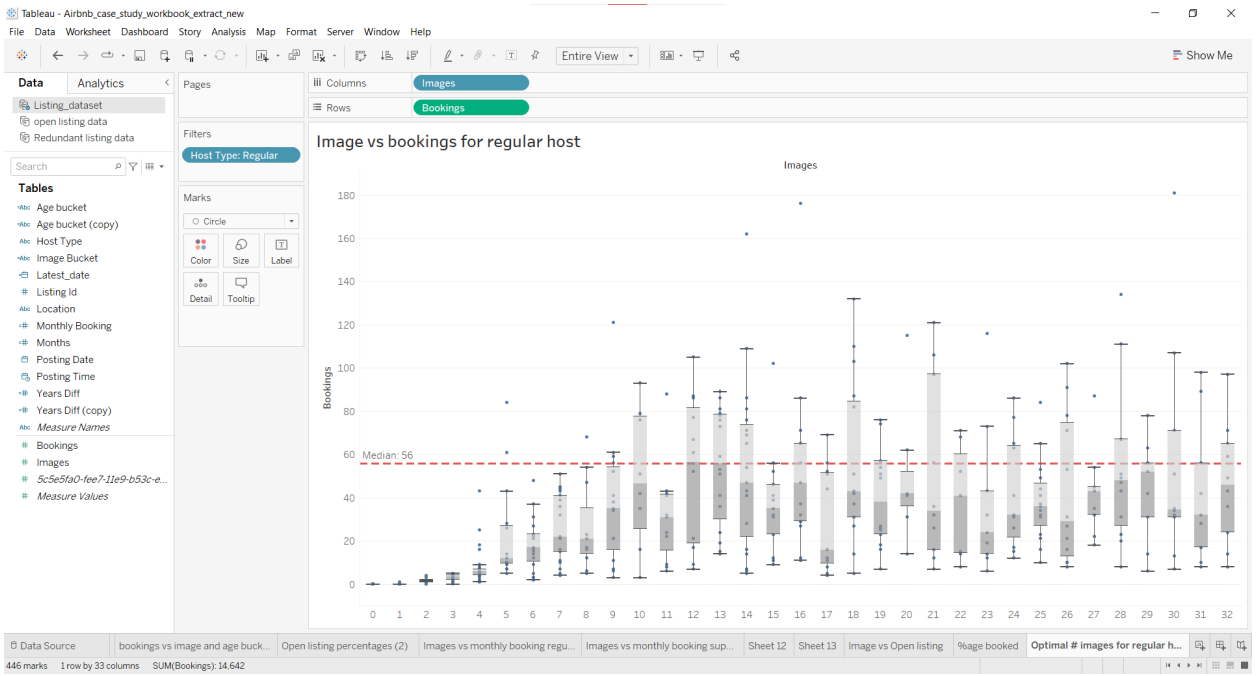
- Which plot would you use if we want to compare the distribution of data points across different values of a categorical variable in a single plot?

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**.

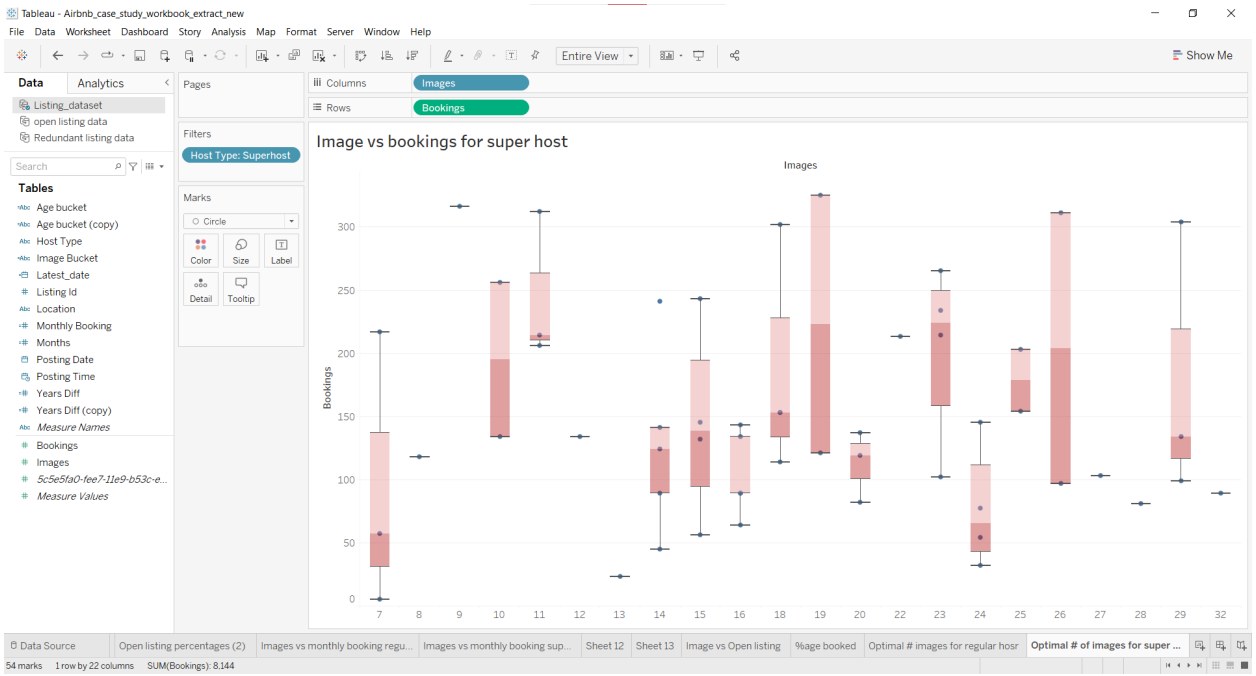
Note: We're looking into different host types separately because of their different nature.



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 maximise their bookings and also because it has the third highest median value.

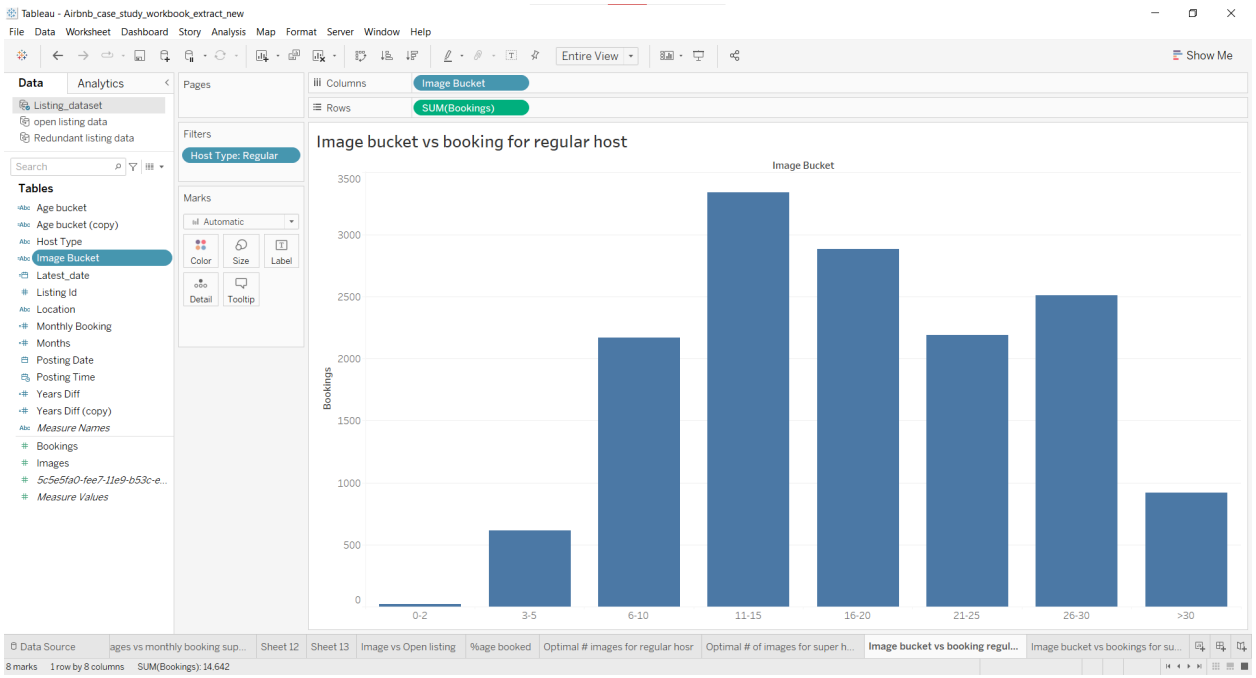
Let us analyze the data further to back the above findings.

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:

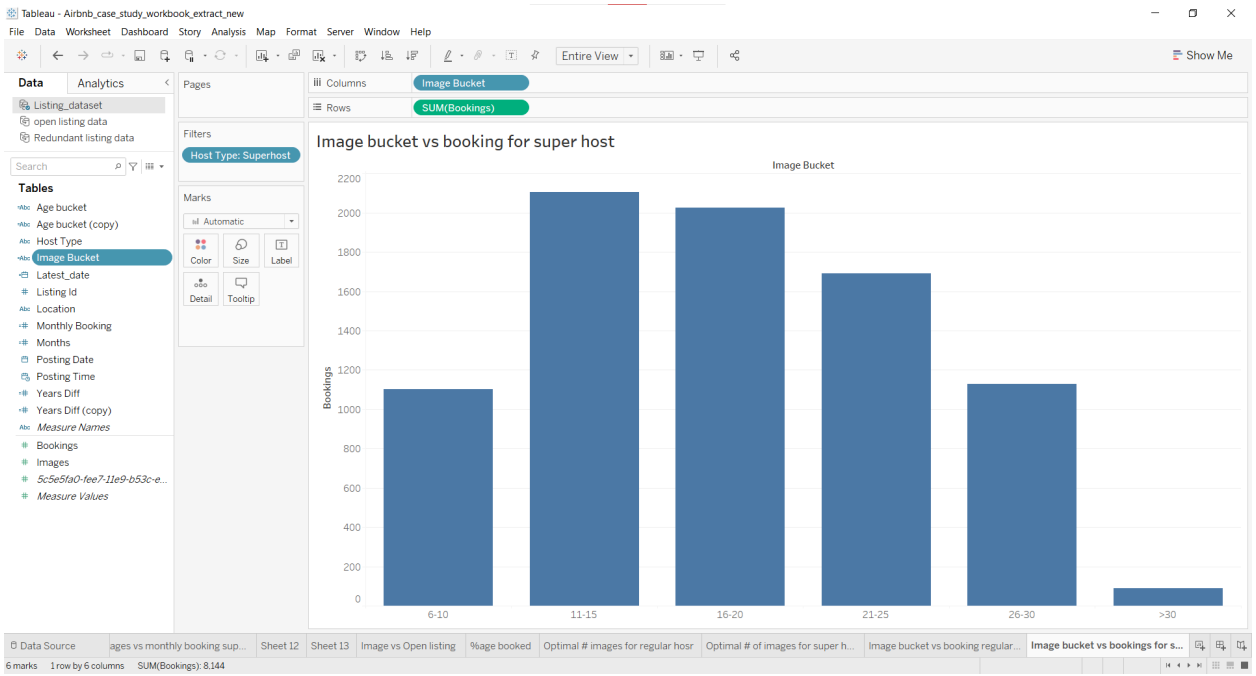
```
IF [Images]<=2
THEN
"0-2"
ELSEIF [Images]>2 AND [Images]<=5
THEN
"3-5"
ELSEIF [Images]>5 AND [Images]<=10
THEN
"6-10"
ELSEIF [Images]>10 AND [Images]<=15
```

```
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:

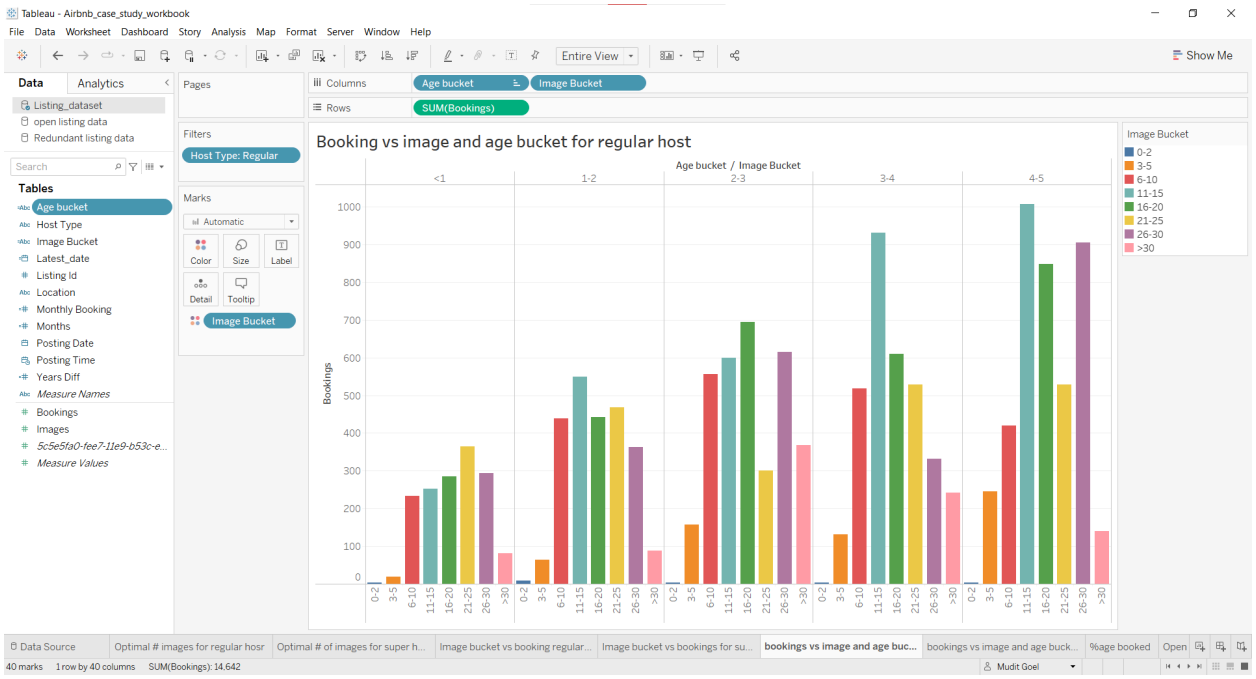
- <1 (property is less than 1 year old)
- 1-2 (Property age is between 1 to 2 years old)
- 2-3 (Property age is between 2 to 3 years old)
- 3-4 (Property age is between 3 to 4 years old)
- 4-5 (Property age is between 4 to 5 years old)

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:

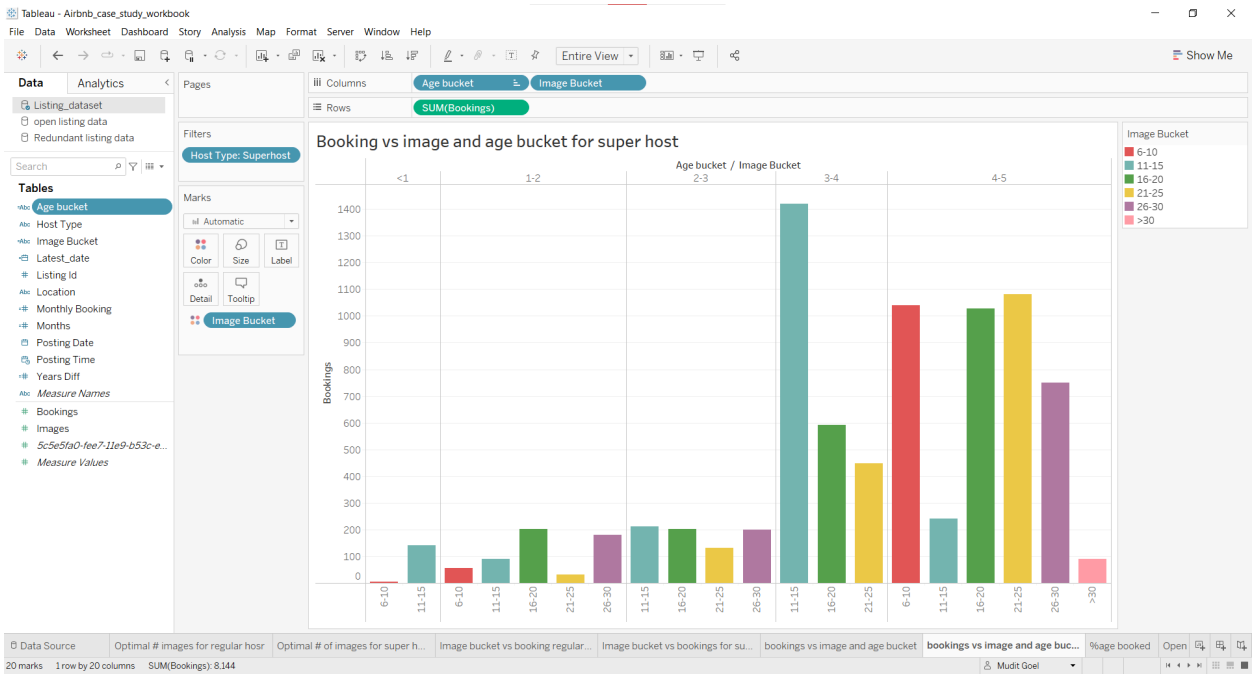
- If the posting date is 1-9-2018 then there are 11 months till August 2019 and when we divide it by 12, we will have a value of less than 1 hence the listing is less than a year old.
- If the posting date is 1-9-2017 now here we have 23 months till August 2019. Hence when we divide it by 12, the value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

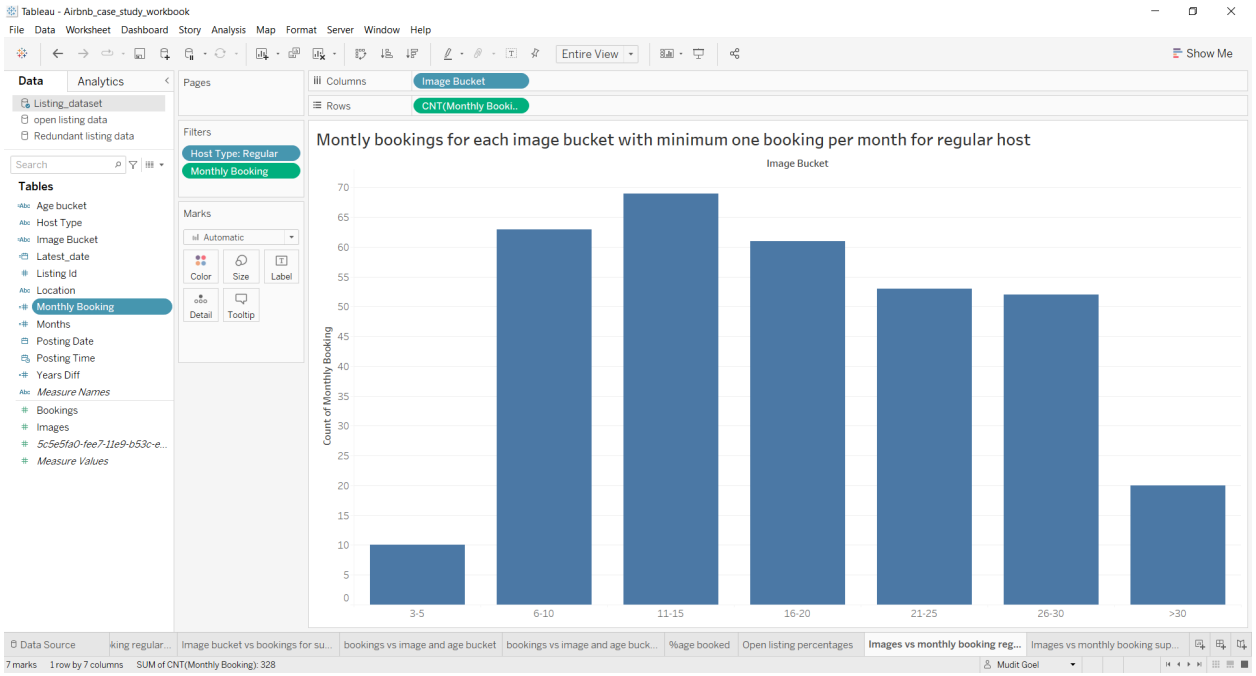
Steps to get listing age in months -

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

Example:

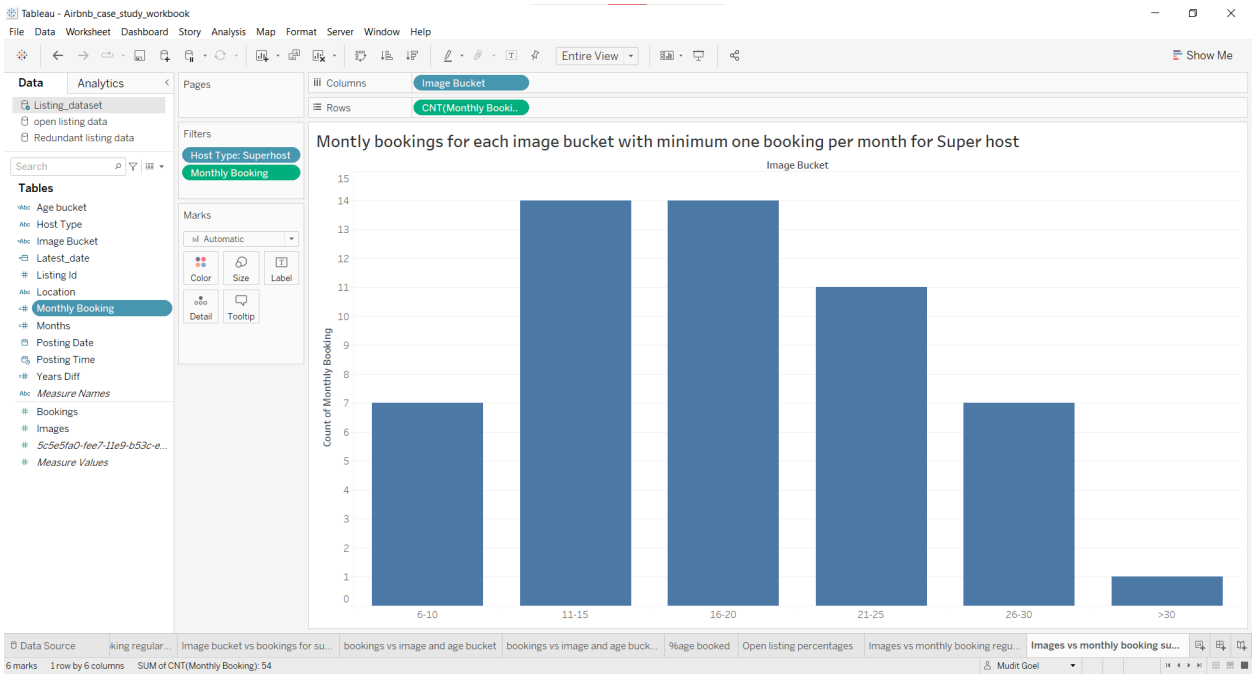
- If the posting date is 31-8-2014 then the number of days till 31st August 2019 is 1826 (365*5+1[current date]).

- 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**.

Assumptions:

- Regular and super hosts have been analyzed separately due to their different nature.
- Although bookings are dependent on various factors such as review volume, location, pricing, amenities, brand, neighborhood listings, rules & regulations, etc., we have considered '**image count**' as one of the main factors in this analysis.
- While calculating the minimum number of images to ensure bookings, we are keeping the target of at least 1 booking per month. This target can be changed from 1 booking per month to 1 booking per quarter/year and the corresponding minimum number of images required would change.