Name: Keane Dominic Travasso                                            Student No: S10202630(G)

Class: DV\_P04                                                                       Assignment: 1

DV Assignment 1 Report

**Introduction**

I am a data analyst and have been hired by NP Management Ltd, a hotel conglomerate that owns the Metropolis Grand Hotel and Island Grand Resort to explore the hotel\_bookings\_0421.csv and see what kind of data driven exploratory questions can be asked and confirmed and what kind of data driven declarative statements can be made.

The Dashboard will be made for the higher-ups at NP Management Ltd. It will be viewed on a Laptop or Desktop, It is being made to discover trends through the use of exploratory questions and will allow the users to answer many exploratory questions on the dataset from the two locations.

1. Project Objectives

For this project, I will be coming up with 3 dashboards, one which will explore the details of the customers for the two hotels, the second which will explore the booking details and finally, the third which will explore the hotel details.

**The questions I will be creating visualizations on are:**

* Which country are the greatest number of customers from?
* What is the percentage of returning customers as compared to that of the one-time customers?
* What is the average length of stay for customers on both weekdays and weekends across each month?
* What are the top market segments from which the customers booked their stay at the hotel?
* What are the top customer types, customers use when booking their stay at the hotels across the years?
* What percentage of all bookings were canceled?
* Which days of each month had the greatest number of customer arrivals?
* How does the lead time vary based on the customer type and the distribution channel used?
* Which room types have the highest average ADR and the greatest number of customers?
* How has the number of customers and the total revenue made changed over the months?
* Which hotel has a higher number of customers checking-out as compared to those who canceled their stay or did not show up?

**Why should NP Management Ltd. care about the Exploratory Questions I have identified?**

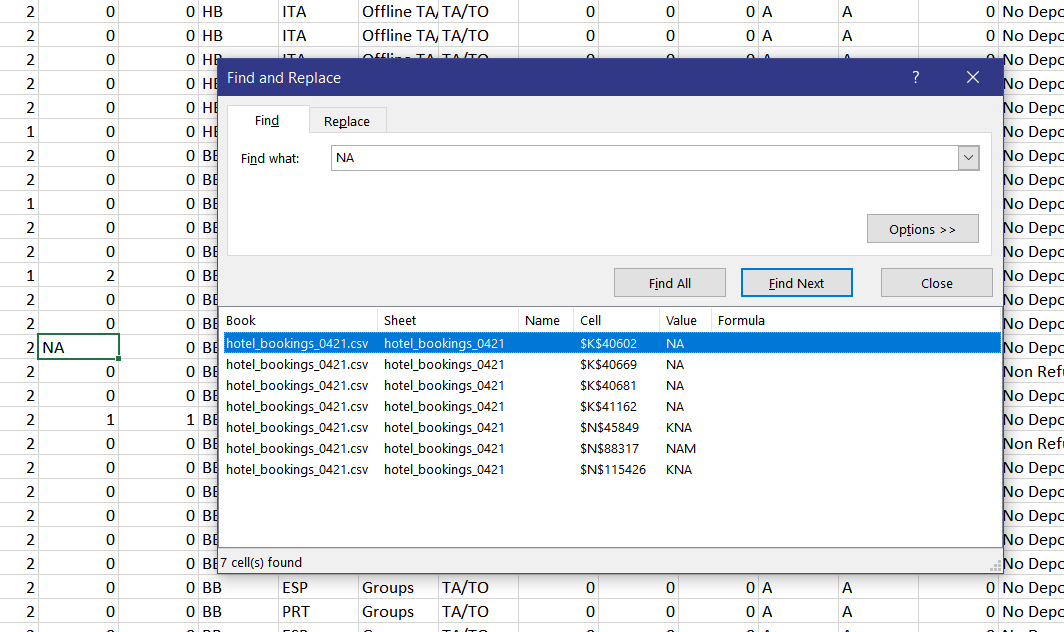
I believe that NP Management Ltd should care about the Exploratory Questions I have identified as all of these questions explore the key values of the hotel bookings dataset. These questions give an insight on the customers visiting the hotels so that NP Management Ltd can use this data in their marketing so as to attract more customers to their hotels. The questions also give an insight into how the hotels are performing over time as well as what factors lead to a higher revenue. Through this NP Management Ltd will understand what works and does not work for each hotel.

2. Data Preparation

The Data hotel\_bookings\_0421 is for the most part a complete set of data with only a few cells needing to be cleaned. The data has 32 columns with 119391 rows of data entries. To create this dashboard with its visualizations, I do not expect to conduct that much of a cleanup to the data.

**Modifying the Dataset:**

In order to ensure that the data is suitable for exploration and analysis, the first thing I did was to take a brief look at the data columns and the values in them. While looking through the ‘children’ column I saw that one of the cells had the value of ‘NA’. After seeing this, I used ‘Ctrl + f’ to search through the data to find all the cells that had the value of ‘NA’. I then replaced all the cells with the value of 0 as there were already cells under children with the value of 0 and hence the value ‘NA’ could be replaced by 0 instead.

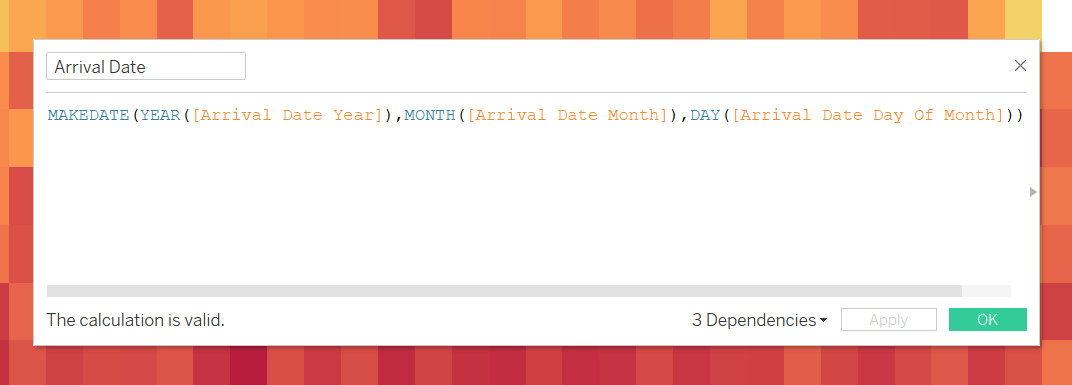


**Sorting the Measures and Dimensions:**

After cleaning the dataset, I opened the CSV file in Tableau in order to sort out the Measures and Dimensions in order for me to utilize the data to create the visualizations to answer my Exploratory Questions.

I first changed ‘Company’ to a dimension, ‘Is Cancelled’ to a dimension and ‘Is Repeated Guest’ to a dimension. I then changed the Data Type of ‘Arrival Date Day of Month’, ‘Arrival Date Month’ and ‘Arrival Date Year’ to the Date value.

In order to work with the Arrival Date better, I created a new calculation called, ‘Arrival Date’ and used the Make Date function to create a Date variable with the values of the year, month and day combined. The command I used to do this was, “MAKEDATE(YEAR([Arrival Date Year]),MONTH([Arrival Date Month]),DAY([Arrival Date Day Of Month]))”.



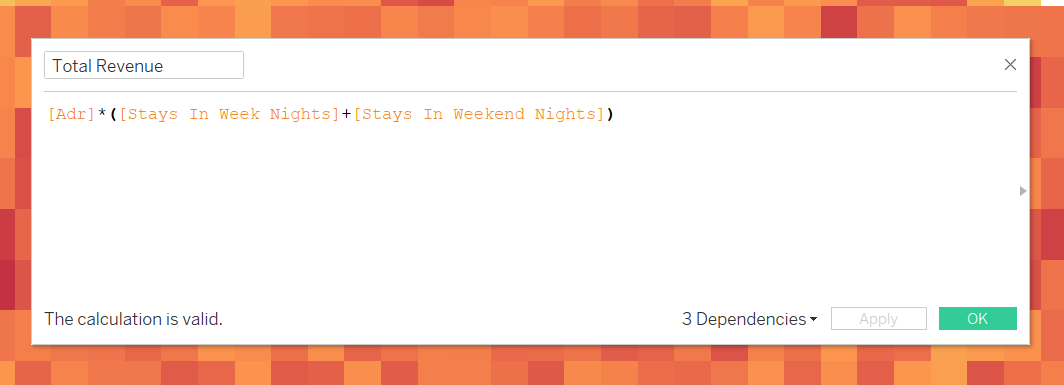
I also created a new calculation called, ‘Total Customers’ which takes the sum of the number of ‘Adults’, ‘Children’ and ‘Babies’. This was done so that the visualizations I create reflect all the customers rather than a select few. The command I used to do this was, “SUM([Adults])+SUM([Children])+SUM([Babies])”.



Next, I created a new calculation called, ‘Stays in Total Nights’ which takes the sum of ‘Stays in Week Nights’ and ‘Stays in Weekend Nights’. This was done so that I can create visualizations that can easily tell how many nights a customer stayed in the hotels. The command I used to do this was, “[Stays In Week Nights]+[Stays In Weekend Nights]”.



Lastly, I created a new calculation called, ‘Total Revenue’ which takes the Average Daily Rate (adr) and multiplies it with the sum of Stays in Week Nights and Stays in Weekend Nights. This equation was taken from the metadata. This was done so that I could tell how much Revenue and money the hotels made. The command I used to do this was, “[Adr] \* ([Stays In Week Nights] + [Stays In Weekend Nights])”.



**Further Research on the Data and Metadata:**

In order to better understand the data, I am working with, so that I can create better visualizations for exploration, I decided to conduct some further research on some of the terminology that I did not fully understand.

The first term I searched up was **Lead Time**. From what I read, I learnt that Lead Time is generally defined as the amount of time that passes from the start to the conclusion of a process. This is used by companies in their processes in order to determine if their process has any flaws and inefficiencies. I also learnt that by reducing the Lead Time, processes can be streamlined and productivity can be increased. In the context of the hotel industry, Lead Time is defined as the time taken between the date of booking and the arrival date. Hotels tend to use Lead Time in order to manage room availability so that customers will have a room whenever they arrive.

The next term I searched up was **Market Segment**. From what I read, I learnt that Market Segment refers to a population of people who all share a similar characteristic. This data is usually used for marketing purposes and also most of all Market Segments are unique and tend to differ from one another.

I also searched up what **Distribution Channel** was. From what I read, I learnt that Distribution Channels are ways in which the hotel promotes and sells its rooms to customers. In the context of this dataset, it is used to determine the ways that the customers booked their rooms.

Finally, the last term I searched up was the **Average Daily Rate**. From what I read, I learnt that Average Daily Rate is a simple calculation that calculates the average rental revenue earned per day for an occupied room. The Average Daily Rate is usually used to determine the performance of a business.

Sources of information used:

Landman, P. (2020). *Lead Time Restriction Definition / Meaning - Xotels*. [online] Xotels. Available at: <https://www.xotels.com/en/glossary/lead-time-restriction>

Investopedia. (2021). *What Is Lead Time?* [online] Available at: <https://www.investopedia.com/terms/l/leadtime.asp>

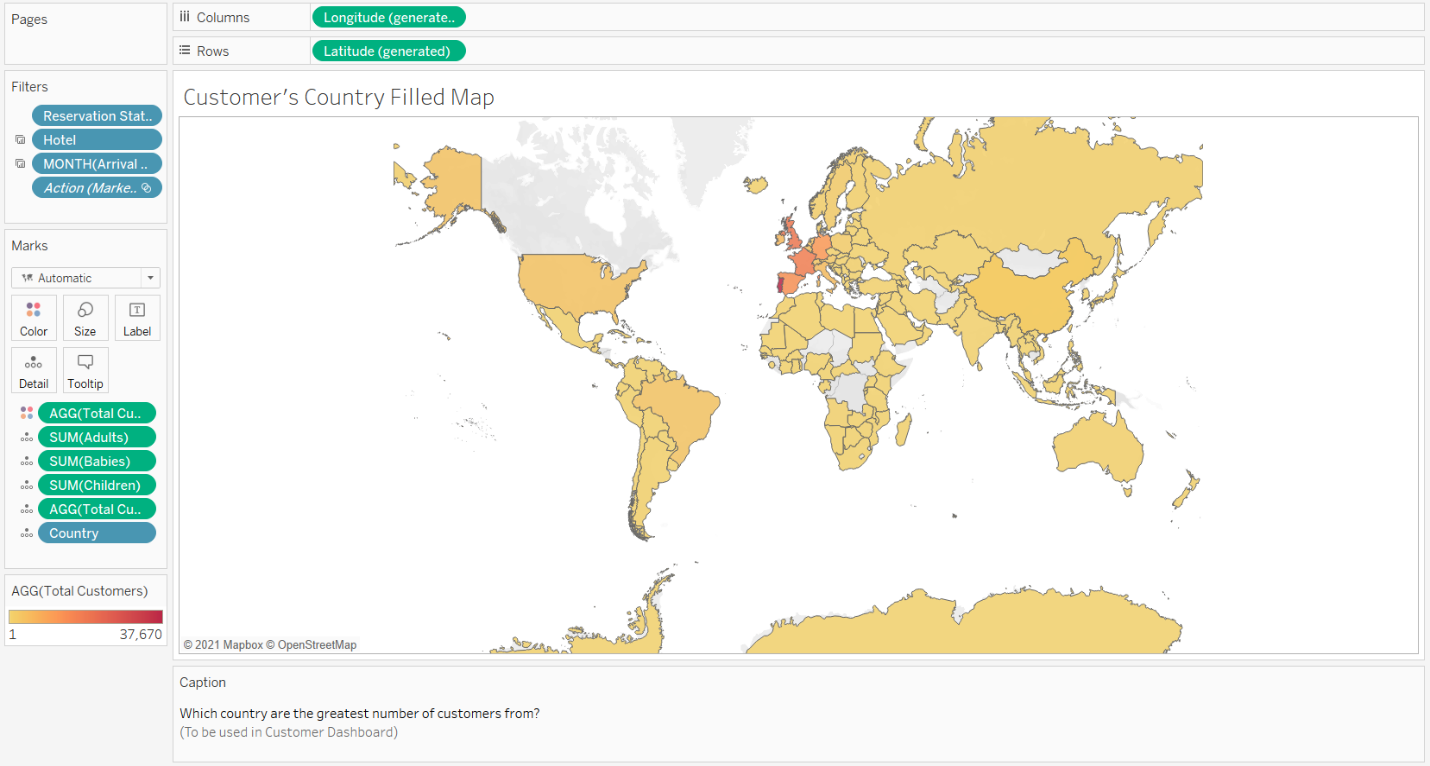
Landman, P. (2021). *Hotel Market Segmentation - Hotel Target Market Segments*. [online] Xotels. Available at: <https://www.xotels.com/en/revenue-management/revenue-management-book/hotel-market-segmentation>

Investopedia. (2021). *How Distribution Channels Work*. [online] Available at: <https://www.investopedia.com/terms/d/distribution-channel.asp>

Investopedia. (2021). *Average Daily Rate (ADR)*. [online] Available at: <https://www.investopedia.com/terms/a/average-daily-rate.asp>

3. Exploratory Data Analysis and Visualization

**Which country are the greatest number of customers from?**

**Main Visualization:**

**Reason for picking Visualization:**

In order to answer this Exploratory Question, I decided to use a filled (polygon) map. The reason why I chose to visualize the data in this way is because location data is best represented in a map format as it allows the user to understand where exactly the data is from. Another reason why a filled map is used is because the question asked is a spatial one.

**What the Visualization shows the user:**

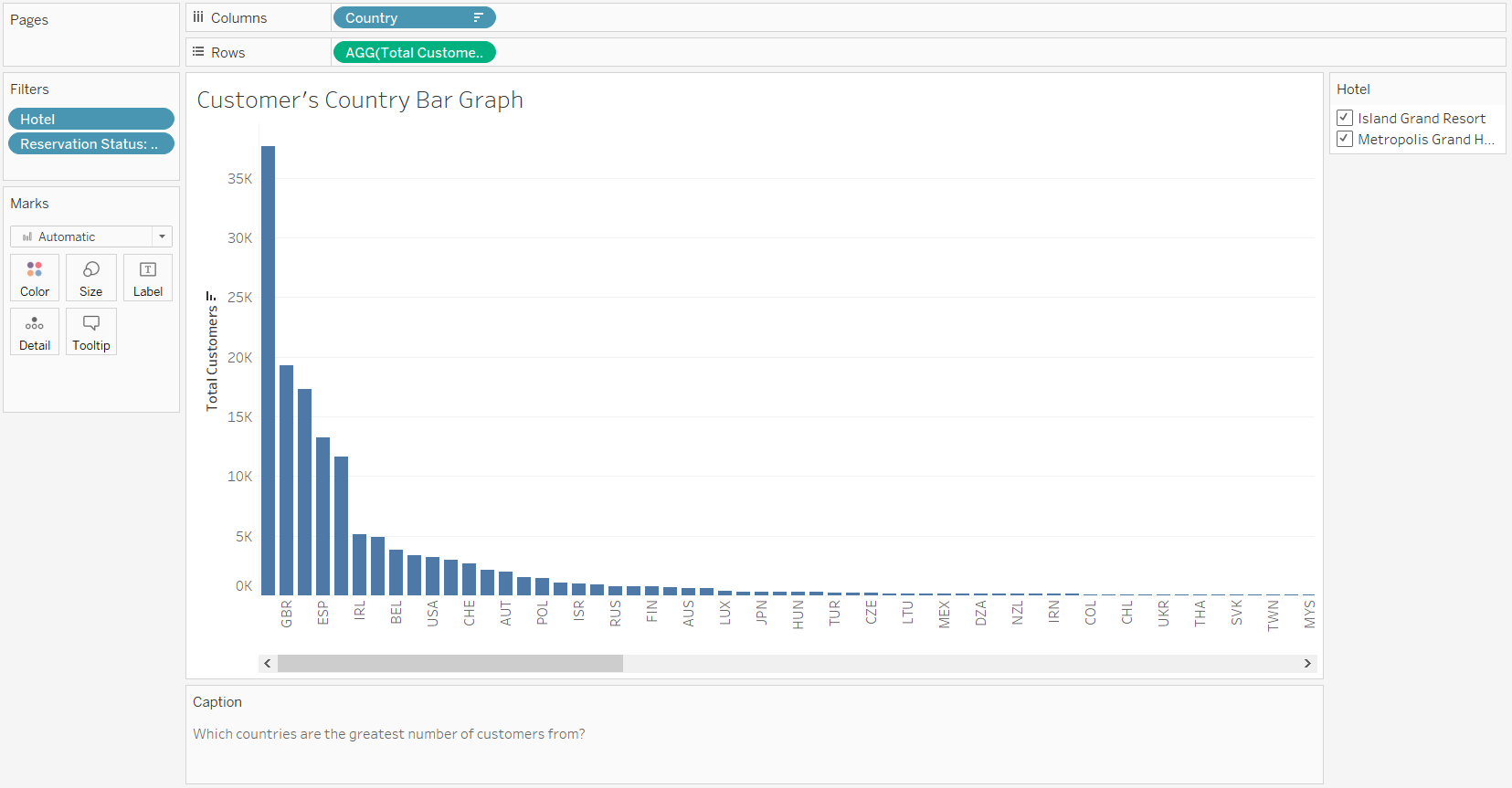
The filled map shows the countries where the customers are from. It also is sorted by color to show which countries have the larger numbers of customers.

Hovering over a country will present the user with the country code followed by the number of adults, the number of babies, the number of children and lastly the total number of customers. This was done so that the user can view the breakdown of the customers from the country they are hovering over.

**The Core Findings and Insights:**

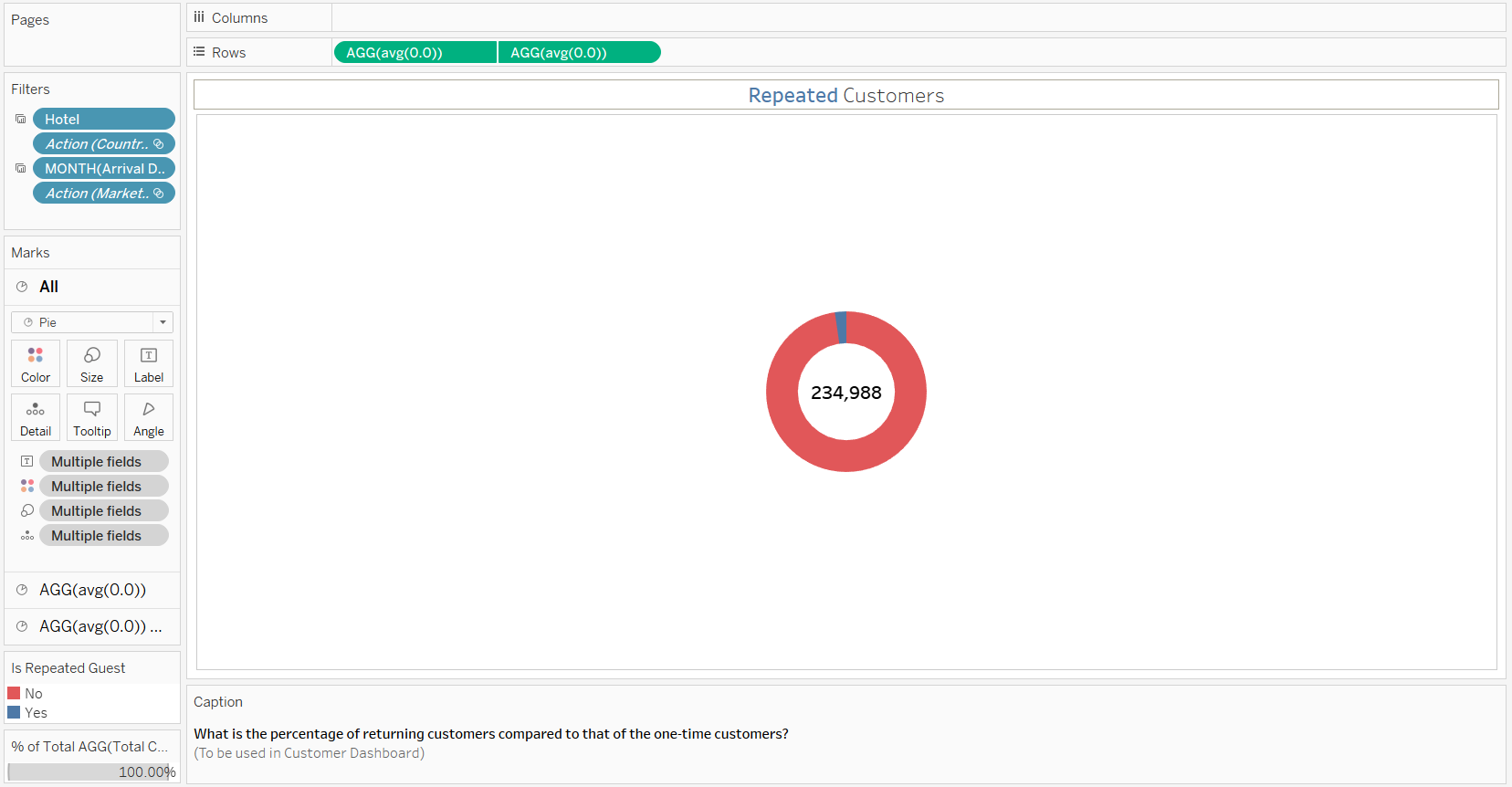
* Upon first glance, all users will be able to tell that the majority of customers the hotel receives comes from Europe. This is evident as the darkest countries on the filled map are Portugal, The United Kingdom, France, Spain, Germany and Italy.
* Another finding that is interesting is that the hotels have received customers from almost every country with a few exceptions like Canada, Mongolia, etc.

**Alternative Visualizations (Rejected):**



Another visualization I was considering was a bar chart that shows the country code along with the total number of customers. In the end I decided against the bar chart as it did not give any perspective on where the countries were.

**What is the percentage of returning customers as compared to that of the one-time customers?**

**Main Visualization: **

**Reason for picking Visualization:**

In order to answer this Exploratory Question, I decided to use a doughnut chart. The reason why I chose to visualize the data in this way is because I am trying to compare parts of a meaningful whole, the repeated customers and not repeated customers from the total number of customers. Another reason why I chose the doughnut chart is because it can display the total number of customers in the middle, which makes it even more useful.

**What the Visualization shows the user:**

The doughnut chart shows the user the proportion of customers that are repeat customers as compared to those who are not. The chart also shows the total number of customers in the middle.

Hovering over any part of the doughnut chart will present the user with whether the part shown is for repeat guest or not followed by the percentage the part covers as well as the number of customers that percentage represents. This was done so that the user can easily get a breakdown of the number of customers as well as what percentage of the total that amount is.

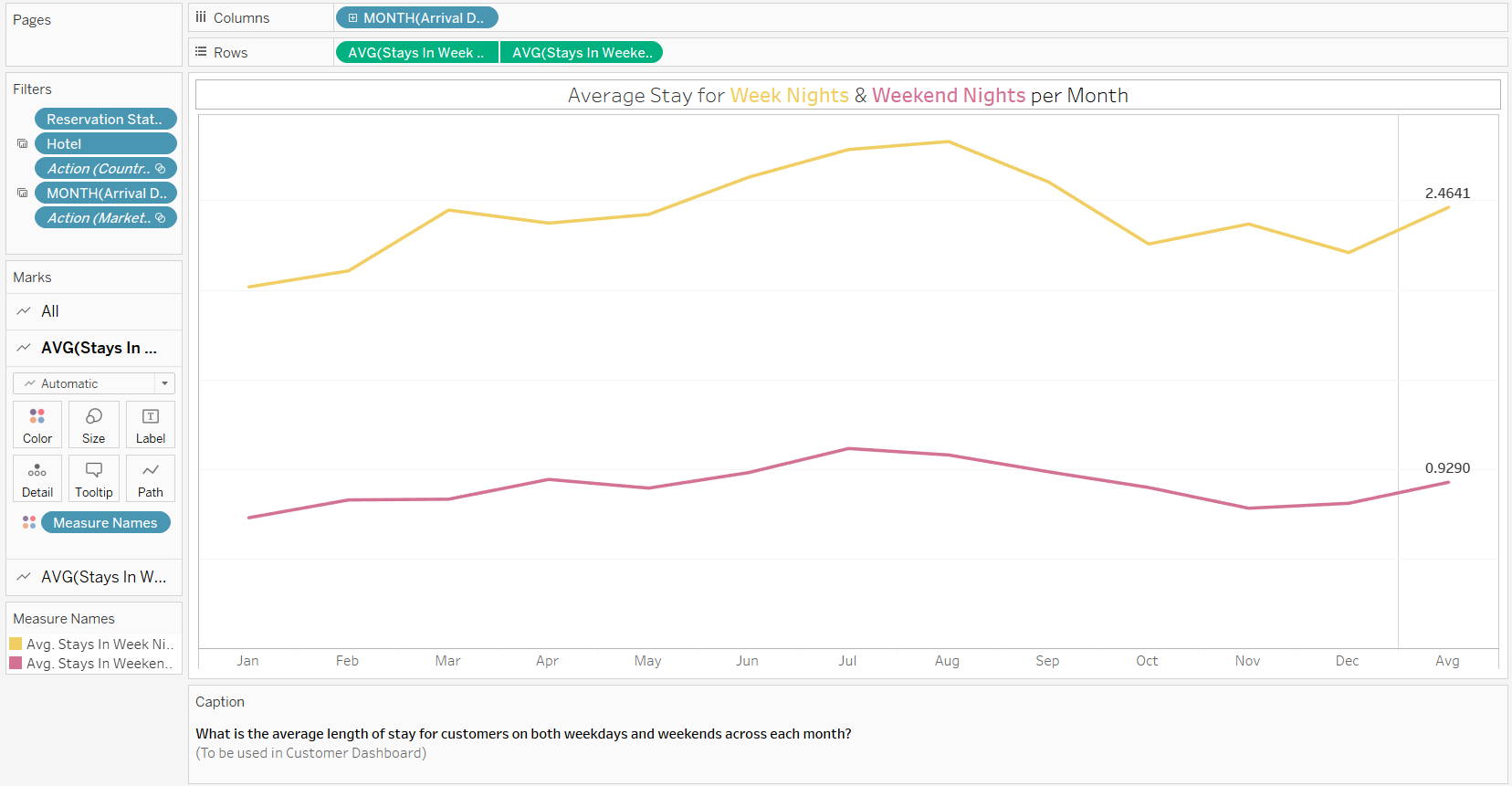
**The Core Findings and Insights:**

* Island Grand Resort has a higher customer repeat rate, 3.32% while Metropolis Grand hotel is 1.78%
* The month of January has the highest repeated customer rate, 5.20%.

**Alternative Visualizations (Rejected): **

Another Visualization I was considering to answer this exploratory question was a bar chart.In the end, I decided against using the bar chart as I felt like the doughnut chart best expressed the message, I wanted to convey all while looking much better and cleaner.

**What is the average length of stay for customers on both weekdays and weekends across each month?**

**Main Visualization:** 

**Reason for picking Visualization:**

In order to answer this Exploratory Question, I decided to use a dual line chart. The reason why I chose to visualize the data in this way is because I wanted to compare the average of the two measures, ‘Stays in Week Nights’ and ‘Stays in Weekend Nights’ across all the months. Making the visualization a dual line chart made the most sense as it is easy to compare the values and view the trends over the months.

**What the Visualization shows the user:**

The dual line chart shows the user the trend in the average stay for both weekday nights and weekend nights across the 12 months. It also has an extra column at the end to show the user the average stay for all the months.

Hovering over any point in the chart presents the user with the Month name, the Hotel name and the average stay in either week nights or weekend nights for that month.

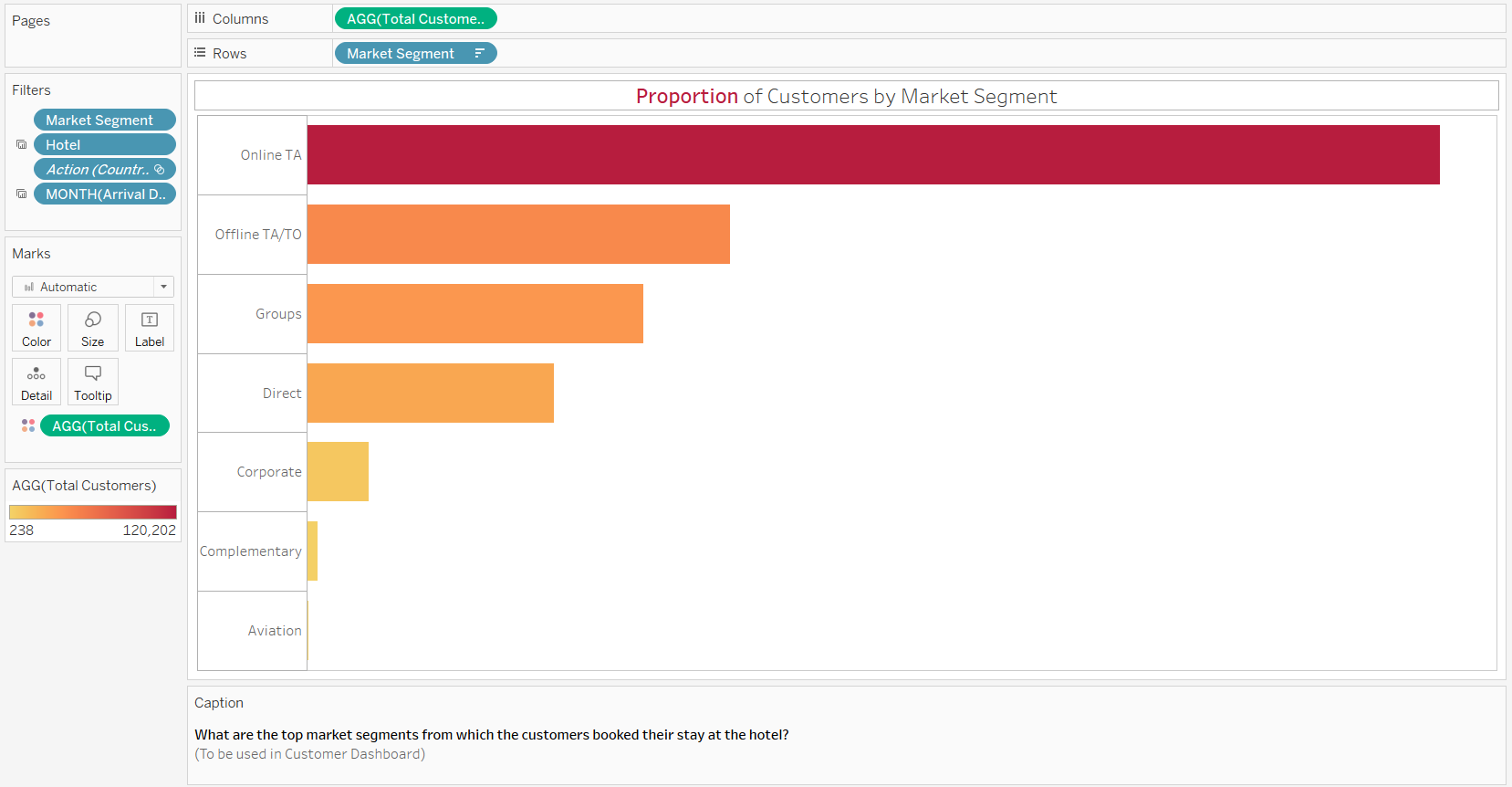
**The Core Findings and Insights:**

* From first glance, users can tell that the Island Grand Resort has more of a variation in the average stay for both week nights and weekend nights. The average stay for week nights starts going up from January to June where it reaches a plateau up till September. From there to December, the average stay drops to a point that is slightly higher than January. The average stay in weekend nights follows a similar pattern but to a smaller extent.
* Metropolis Grand Hotel has a more consistent line for the average stay in week nights and weekend nights with small peaks in March, August and November.

**Alternative Visualizations (Rejected):**

An alternative visualization I was considering was an area chart but I decided against using it since it would not show the exact amount on the line and would require the user to subtract and find the number themselves.

**What are the top market segments from which the customers booked their stay at the hotel?**

**Main Visualization: **

**Reason for picking Visualization:**

In order to answer this Exploratory Question, I decided to use a bar chart. The reason why I chose to visualize the data in this way is because I wanted to show the distribution of Total Customers across all the different Market Segments and make comparing the values between all the rows easier.

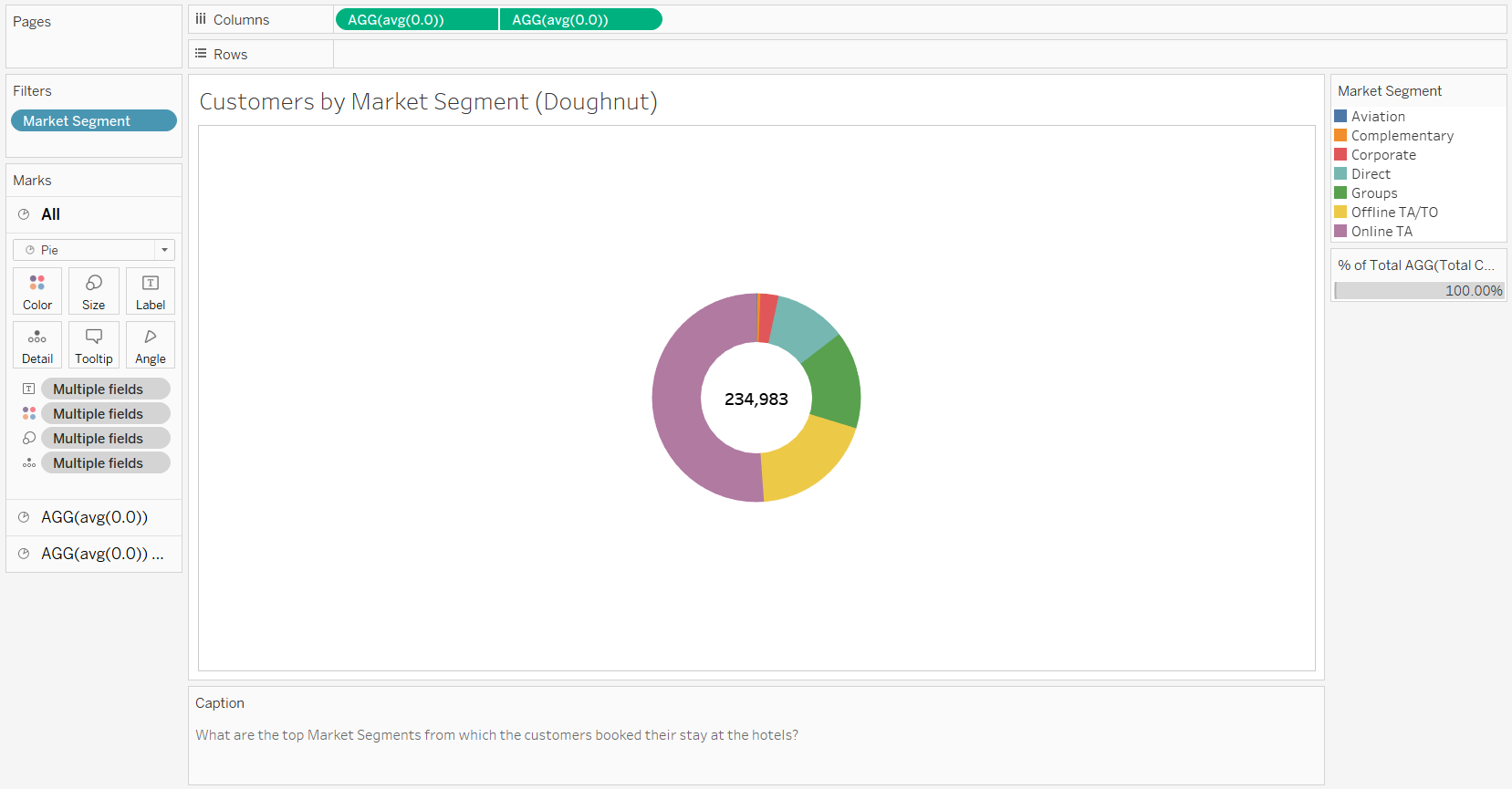
**What the Visualization shows the user:**

The bar chart shows the user the distribution of customers across the various market segments. The bars are sorted and colored by the total number of customers, with the darker color meaning more customers.

Hovering over any point in the chart presents the user with the Market Segment Name as well as the number of customers belonging to that segment. The Market Segment bars are also colored and organized by number of customers.

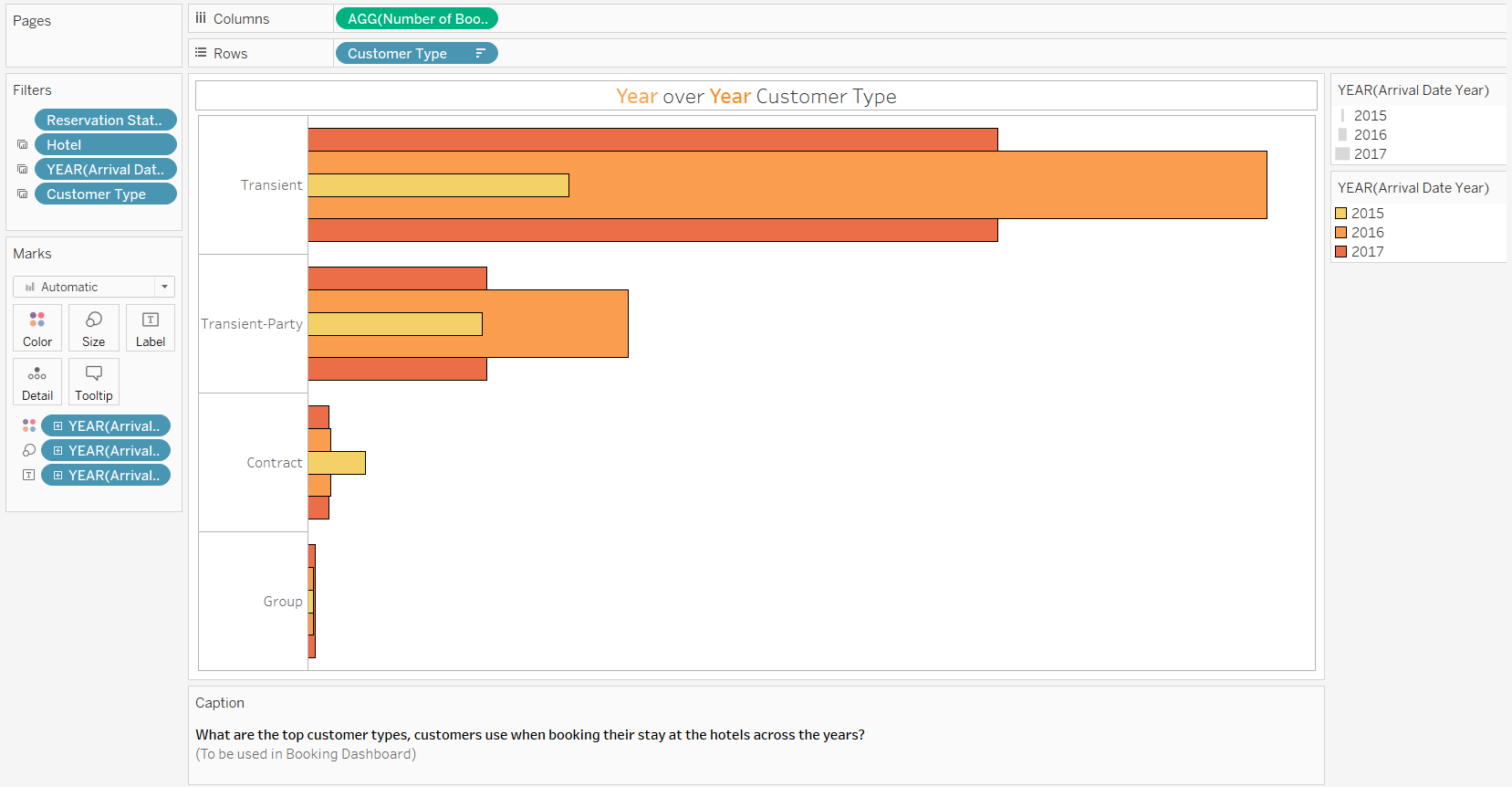
**The Core Findings and Insights:**

* Online TA is the Market Segment with the greatest number of customers for both Island Grand Resort and Metropolis Grand Hotel.
* Only Metropolis Grand Hotel has customers from the Aviation Market Segment, having 238 customers in total.
* The country with the greatest number of customers, Portugal, has Groups as their most used Market Segment.

**Alternative Visualizations (Rejected): **

Another way of visualizing this exploratory question, that I considered, was a doughnut chart. The doughnut chart also shows the distribution of customers across the Market segments and makes it easy to compare the values. In the end, I decided against using the doughnut chart to represent the proportion of customers in each market segment as I already had a doughnut chart created for repeated guest and I did not want too many of the same type of graph.

**What are the top customer types, customers use when booking their stay at the hotels across the years?**

**Main Visualization: **

**Reason for picking Visualization:**

In order to answer this Exploratory Question, I decided to use a bar in bar chart. The reason why I chose to visualize the data in this way is because I wanted to show the distribution of number of bookings by the customer type they used when booking. The bar in bar chart also allows me to add another dimension, which in this case is the year. This allows me to compare the variation in customer type over the years.

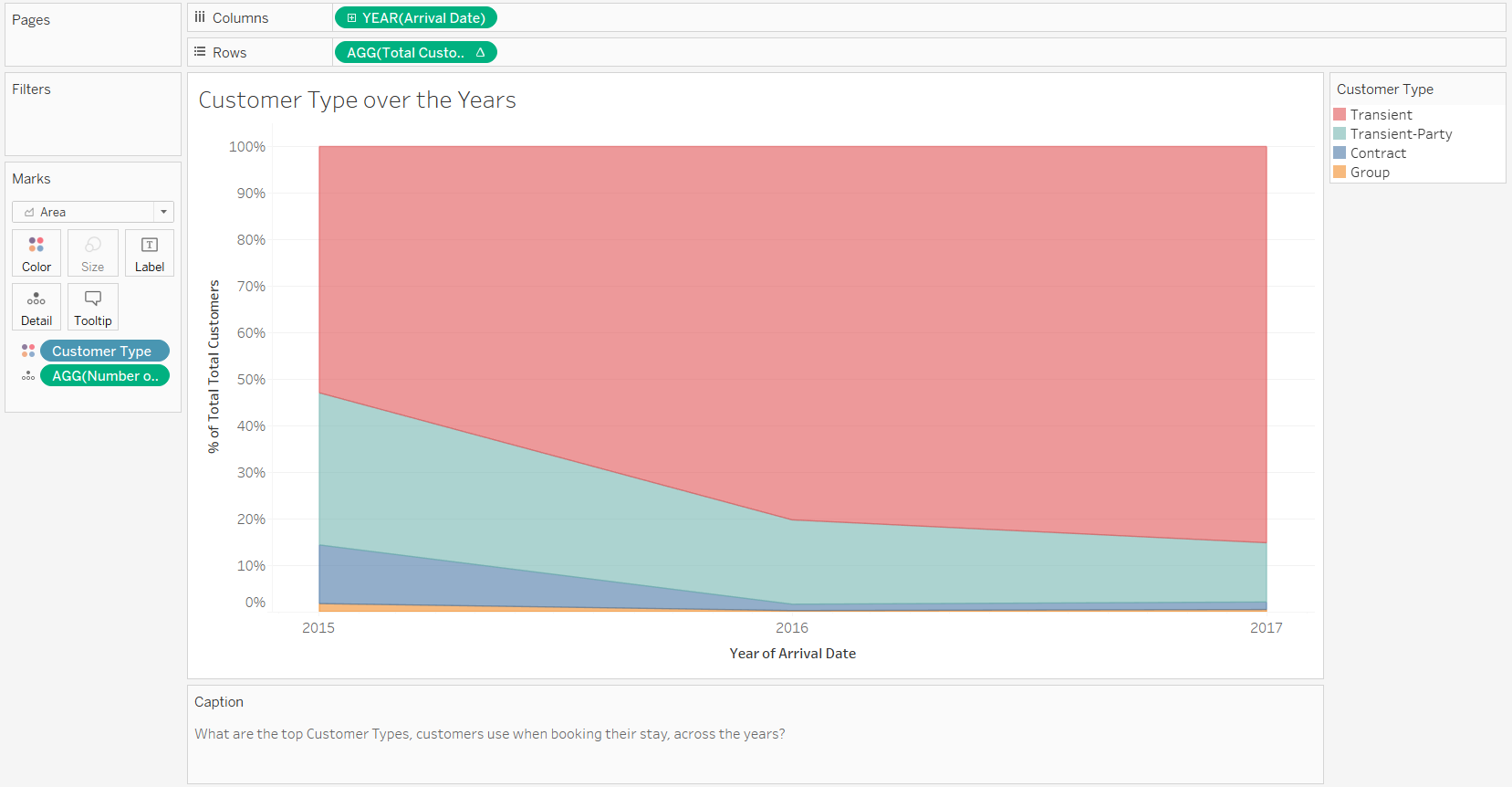
**What the Visualization shows the user:**

The bar in bar chart shows the user the variation in the customer type based on the number of customers, over the years. The bar in bar chart is sorted by color and size based on the year, with 2015 being the small yellow bar, 2016 being the medium orange bar and 2017 being the large dark orange bar.

Hovering over any point in the chart presents the user with the Customer Type, the Year and the number of bookings.

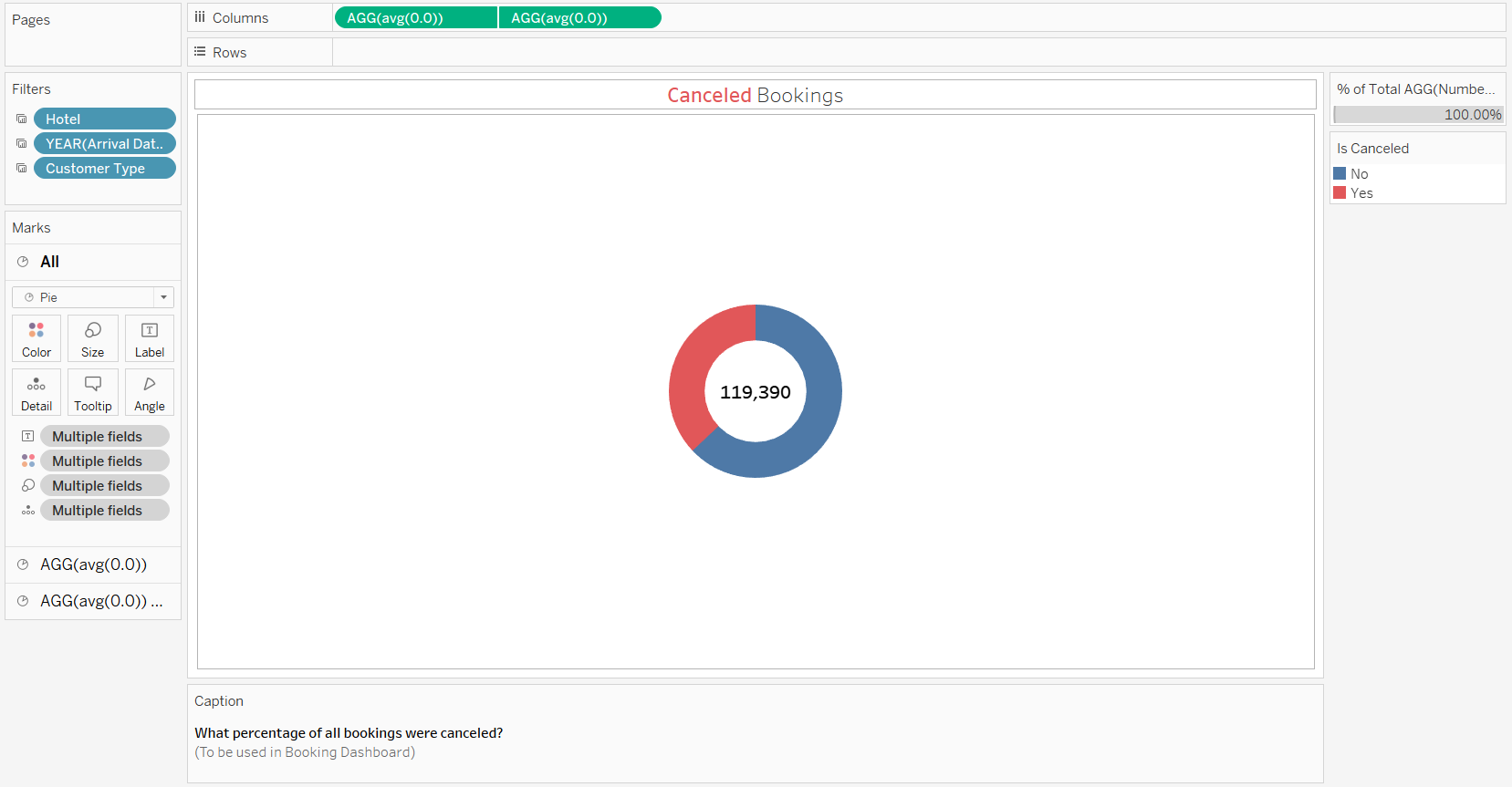
**The Core Findings and Insights:**

* Transient was the most popular customer type for all 3 years in Island Grand Resort while in Metropolis Grand Hotel, it was the most popular customer type of 2016 and 2017.
* Group is the least popular customer type overall.

**Alternative Visualizations (Rejected): **

An alternative to the bar in bar chart that I considered was a proportionate area graph. The area graph does all of what the bar in bar chart does. In the end, I decided against using the proportionate area graph as I felt like it was a harder to compare the variation of number of bookings for all the customer types as compared to the bar in bar chart.

**What percentage of all bookings were canceled?**

**Main Visualization:** 

**Reason for picking Visualization:**

In order to answer this Exploratory Question, I decided to use a doughnut chart. The reason why I chose to visualize the data in this way is because I am trying to compare parts of a meaningful whole, whether the booking was canceled or not from the total number of bookings. Another reason why I chose the doughnut chart is because it can display the total number of bookings in the middle, which makes it even more useful.

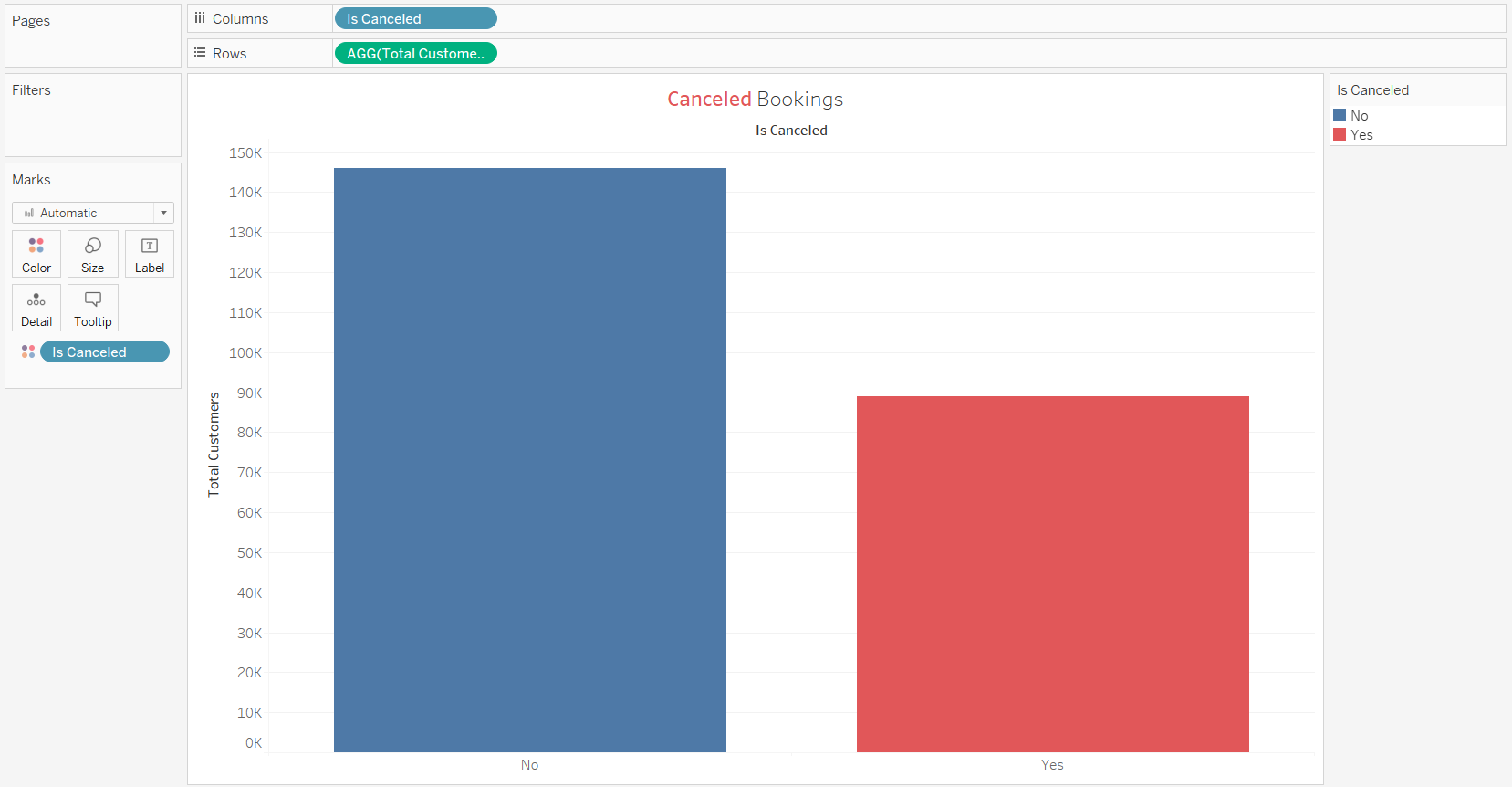
**What the Visualization shows the user:**

The doughnut chart shows the user the proportion of bookings that were cancelled to those that were not canceled. The doughnut chart also shows the number of bookings in the middle.

Hovering over any part of the doughnut chart will present the user with whether the part shown is for canceled booking or not followed by the percentage the part covers as well as the number of bookings that percentage represents. This was done so that the user can easily get a breakdown of the number of bookings as well as what percentage of the total that amount is.

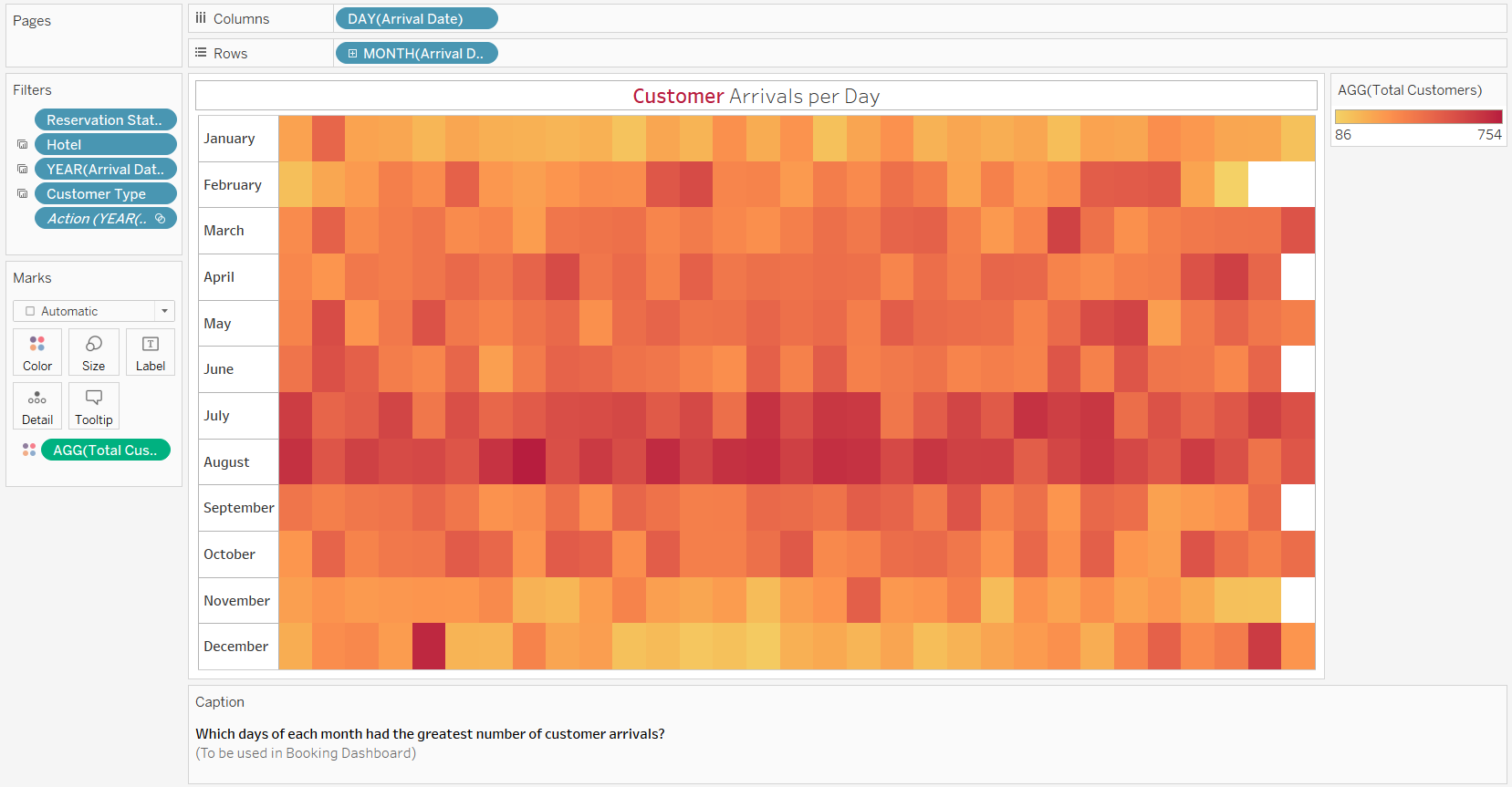
**The Core Findings and Insights:**

* Metropolis Grand Hotel has the most booking cancelations with, 33102, 41.73% of bookings being canceled. Meanwhile, Island Grand Resort has 11122,27.76% of bookings being canceled.
* The Transient Customer Type has the most cancellations among all the customer types with, 36514, 40.75% of all bookings from the Transient customer type being canceled.

**Alternative Visualizations (Rejected): **

Another Visualization I was considering to answer this exploratory question a bar chart. In the end, I decided against using the bar chart as I felt like the doughnut chart best expressed the message, I wanted to convey all while looking much better and cleaner.

**Which days of each month had the greatest number of customer arrivals?**

**Main Visualization: **

**Reason for picking Visualization:**

In order to answer this Exploratory Question, I decided to use a heatmap. The reason why I chose to visualize the data in this way is because I am trying to view the days of the month with the greatest number of customers. The heatmap does this easily by taking the sum of all the customers that arrived on a specific day of the month, throughout the years and puts it in an easy-to-read graph.

**What the Visualization shows the user:**

The visualization shows the sum of customers that arrived on a specific day of the month from all the years. The heatmap is sorted by color based on the sum of the customers, with the darker color meaning more customers. Due to this, the user can easily tell which days are the most popular arrival days in the month.

Hovering over any box in the heatmap gives the day of arrival, the month of arrival and the total number of customers that arrived on that day. The darker the color on the heatmap, the more the number of customers that arrived on that day.

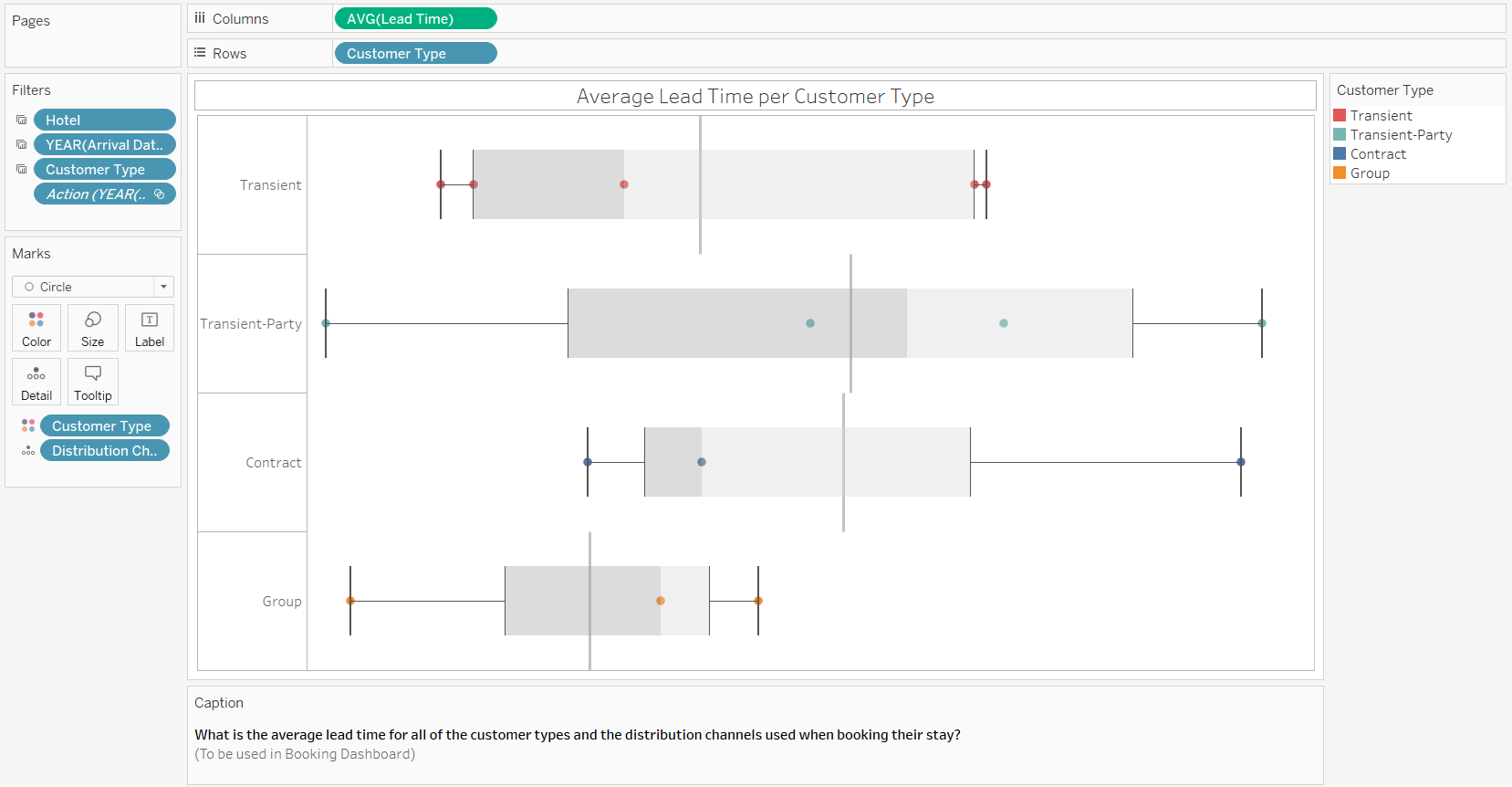
**The Core Findings and Insights:**

* The months of July and August are when the hotels receive the greatest number of guests.
* Metropolis Grand Hotel gets more guests from February to August while Island Grand Resort gets more guests during July and August.

**Alternative Visualizations (Rejected):**

I could not come up with any alternative visualization for this question as I believe that the heatmap visualizes the total customers per day of the month data the best.

**How does the lead time vary based on the customer type and the distribution channel used?**

**Main Visualization: **

**Reason for picking Visualization:**

In order to answer this Exploratory Question, box and whisker chart. The reason why I chose to visualize the data in this way is because I can use both Customer Type and Distribution Channels as well as the Average Lead Time. Using the box and whisker plot allows me to show the average lead time amongst all the distribution channel based on the customer type. Using the box and whisker plot allows me to show the interquartile range which tell the user how spread out the lead time is based on the distribution channels.

**What the Visualization shows the user:**

The visualization shows the user the average lead time for all customer types based on the booking distribution channel used. The visualization also has a box plot which gives more information on data like how big the range is and how big the interquartile range is.

Hovering over any circle on the plot tells the user what customer type the circle belongs to, what distribution channel the circle belongs to and lastly, what the average lead time is for that circle. Hovering over the grey part in the middle tell the user the box plot statistics.

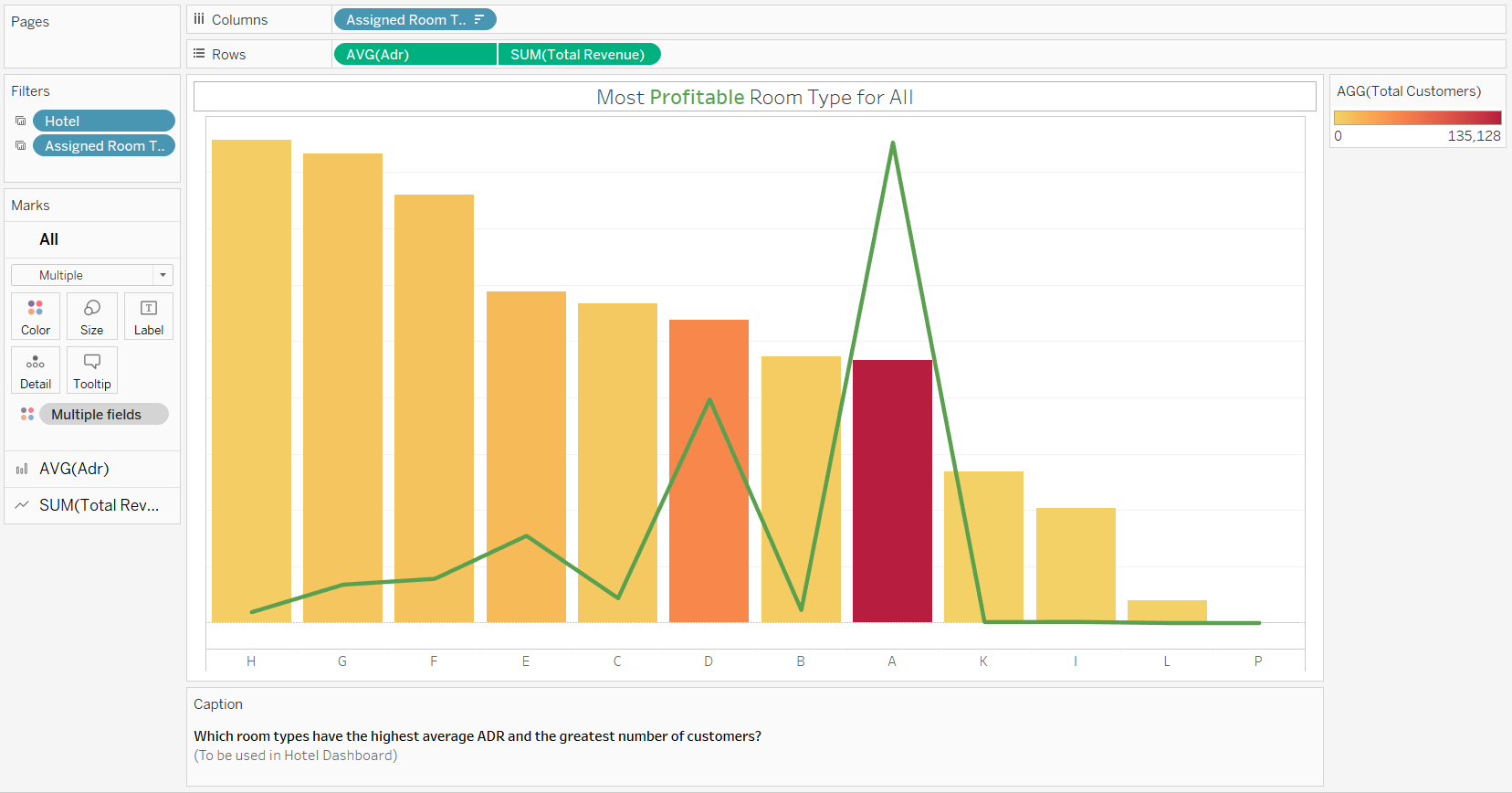
**The Core Findings and Insights:**

* Group Customer Type had the shortest lead time at 43.7 while Transient-Party had the longest lead time at 83.9.
* Transient-Party has the largest range while Group has the shortest.

**Alternative Visualizations (Rejected):**

I could not come up with any alternative visualization for this question as I believe that the box-and-whisker plot visualizes the lead time data the best.

**Which room types have the highest average ADR and the greatest number of customers?**

**Main Visualization: **

**Reason for picking Visualization:**

In order to answer this exploratory question, I decided to use a bar graph and line chart dual axis. The reason why I chose to use a dual axis is because, I wanted to show both the average adr per room type and the total revenue each room type makes. The room types are also colored by the total number of customers hence darker colors means that more customers stay in the relevant room type. Another reason why I chose the dual axis chart was so that I could show the relationship between average adr, number of customers and the total revenue.

**What the Visualization shows the user:**

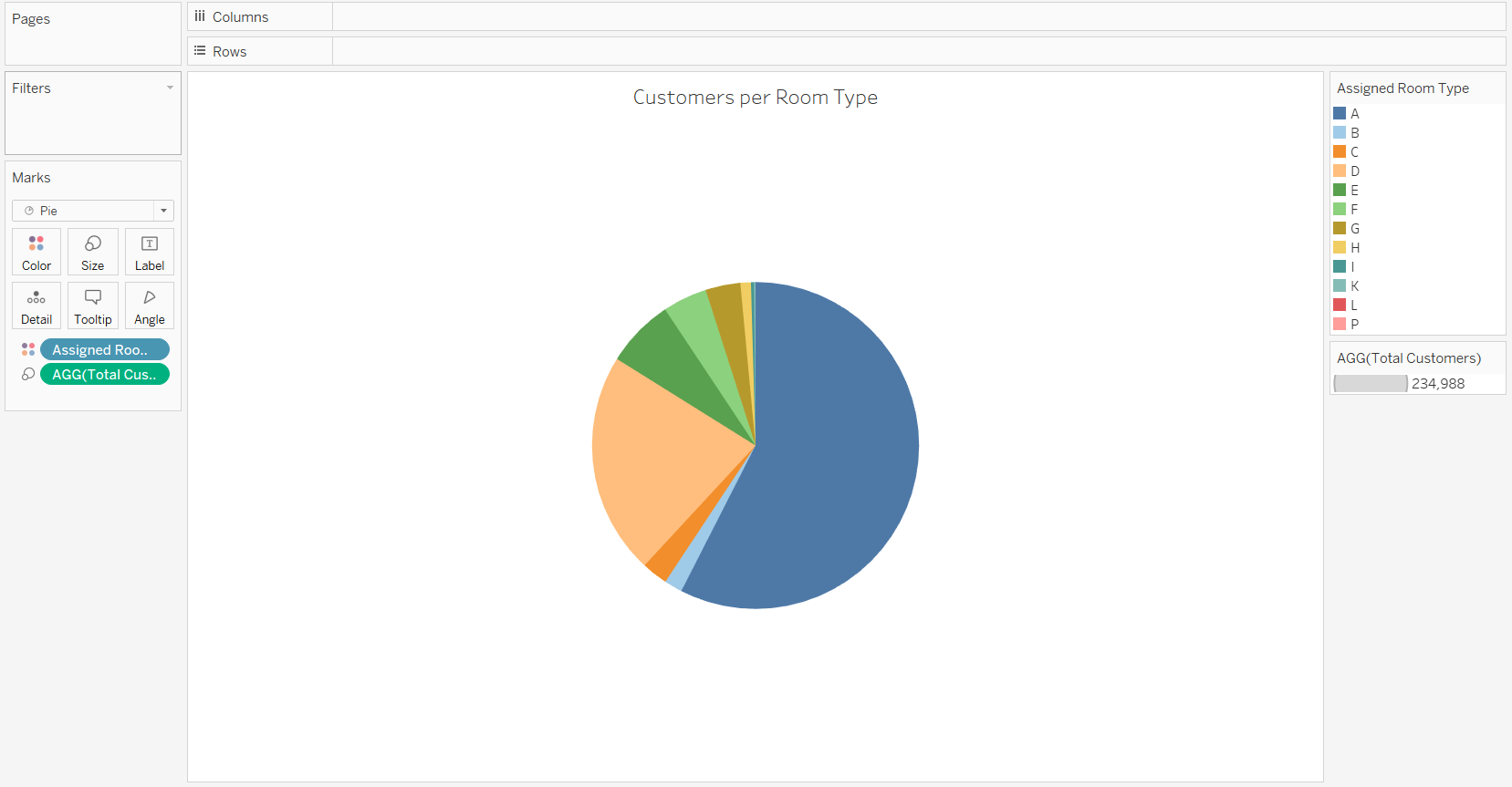
The visualization shows the user which room type has the highest average adr, which room type has the greatest number of customers and which room type makes the most revenue. Bar chart shows the number of customers through color, with the darker color meaning more customers.

Hovering over any part of the chart presents the user with the relevant room type, the average adr for the room type, the total revenue for the room type and lastly, the total number of customers who have stayed at the relevant room type.

**The Core Findings and Insights:**

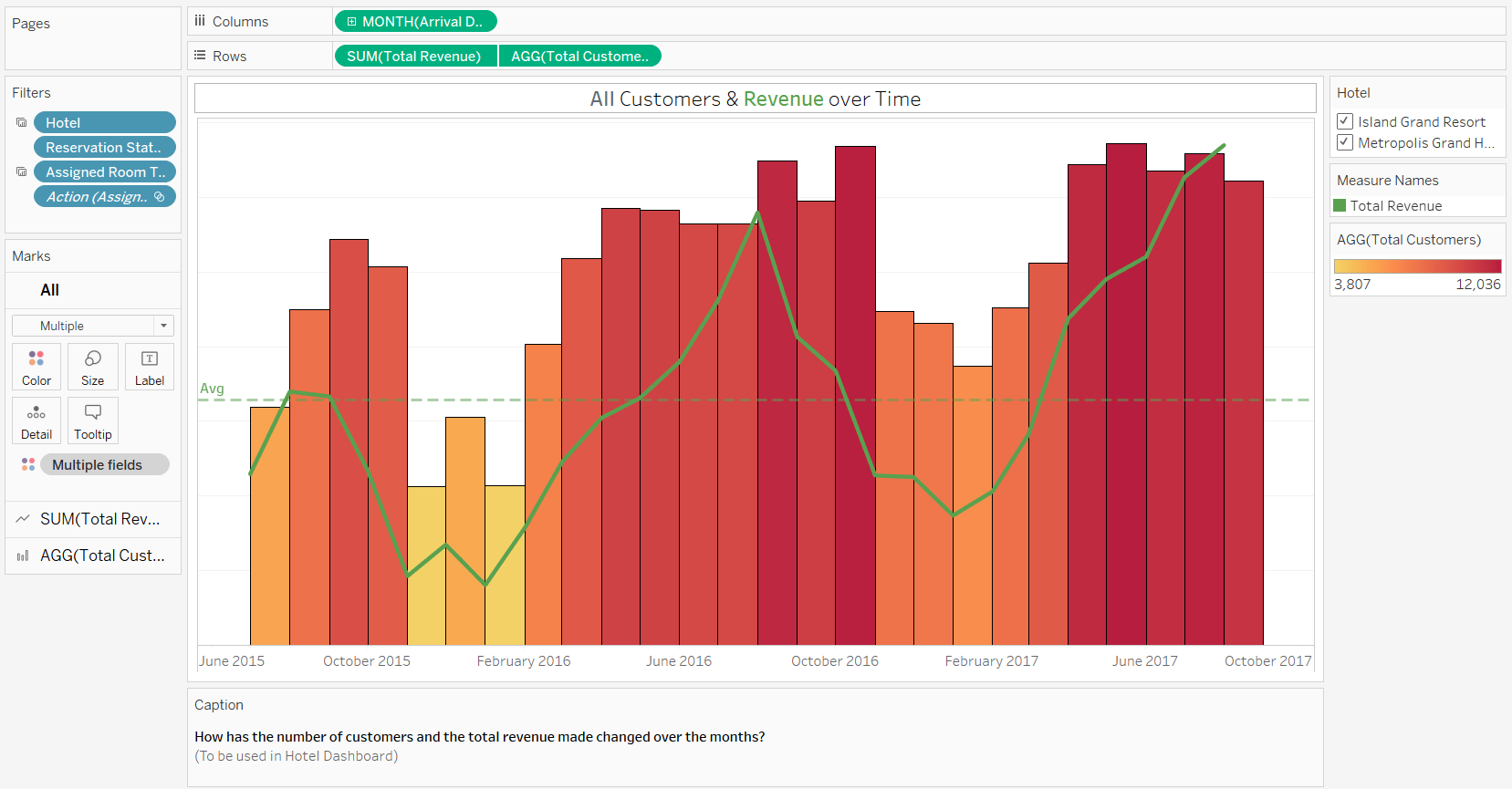
* Room type H has the highest average adr of, $171.38 but made lower a lower revenue of $499573.67 due to the total number of customers using room type H being 2439.
* Room type A has the highest number of customers, 135128, and the highest revenue made, $22202693.95 despite having a lower average adr of $93.14

**Alternative Visualizations (Rejected):**



A visualization that I was considering to use to try and attempt to answer this question was a pie chart based on the room type and sum of customers for the room type. I decided to reject this visualization as I felt it only answered what the most popular room types were. Eventually I came up with the main dual axis chart which answered a multitude of questions and looked better at the same time.

**How has the number of customers and the total revenue made changed over the months?**

**Main Visualization: **

**Reason for picking Visualization:**

In order to answer this exploratory question, I decided to use a bar graph and line chart dual axis. The reason why I chose to use a dual axis is because, the dual axis allowed me to compare the customers over time and the total revenue over time. This graph allows the user to check to see if there is any relation ship between the 2 axes.

**What the Visualization shows the user:**

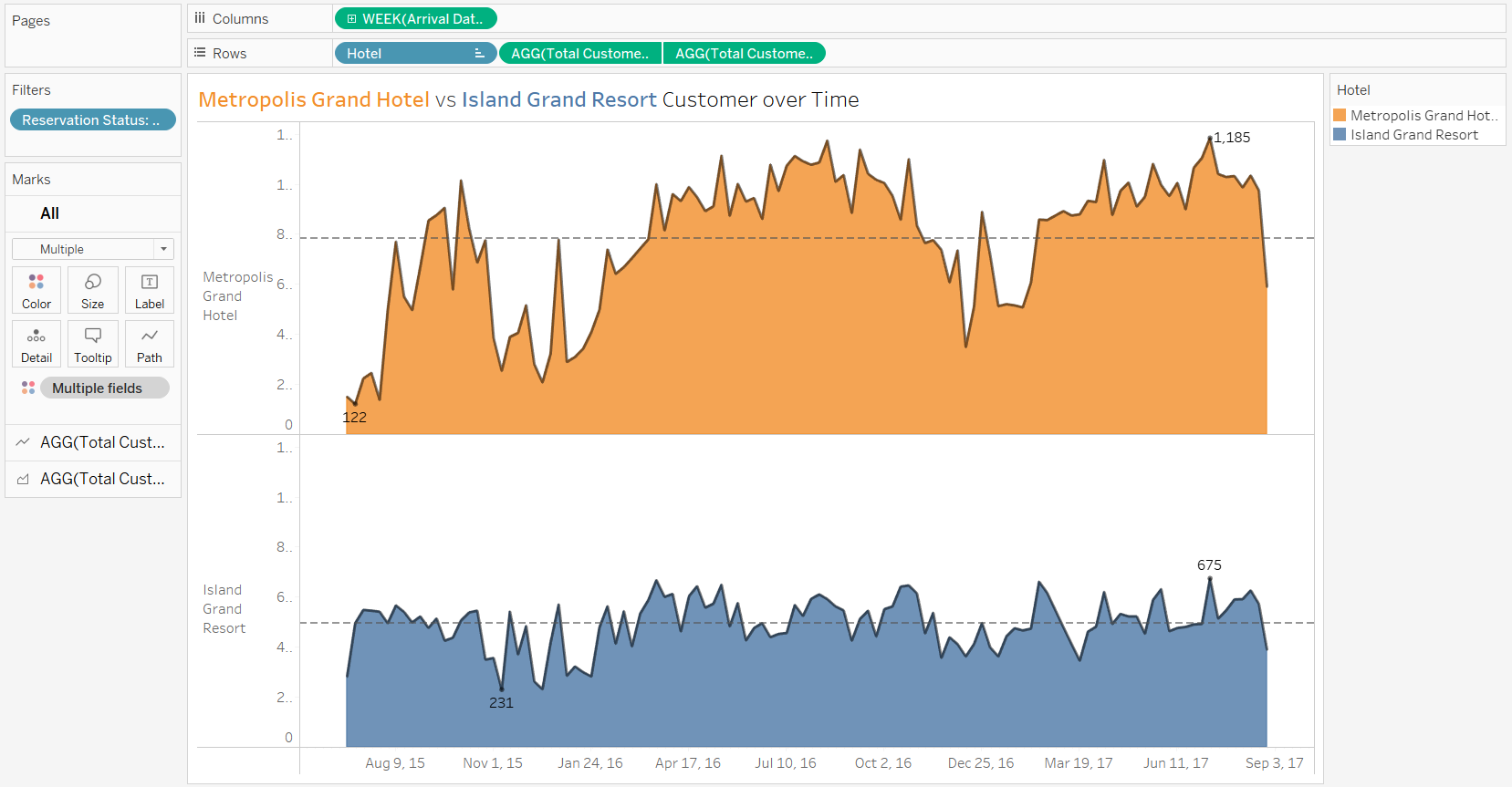
The visualization shows the user the variation in the number of customers that visited the hotels over the months from 2015 to 2017. It shows the user the general period of time when there is a spike in customers and the period of time where there is a drop in customers. Alongside this, the dual axis graph shows the user how the revenue has varied along with the number of customers. There also is an average line for the revenue that shows the user when the hotels were making more revenue than the average and when they were not.

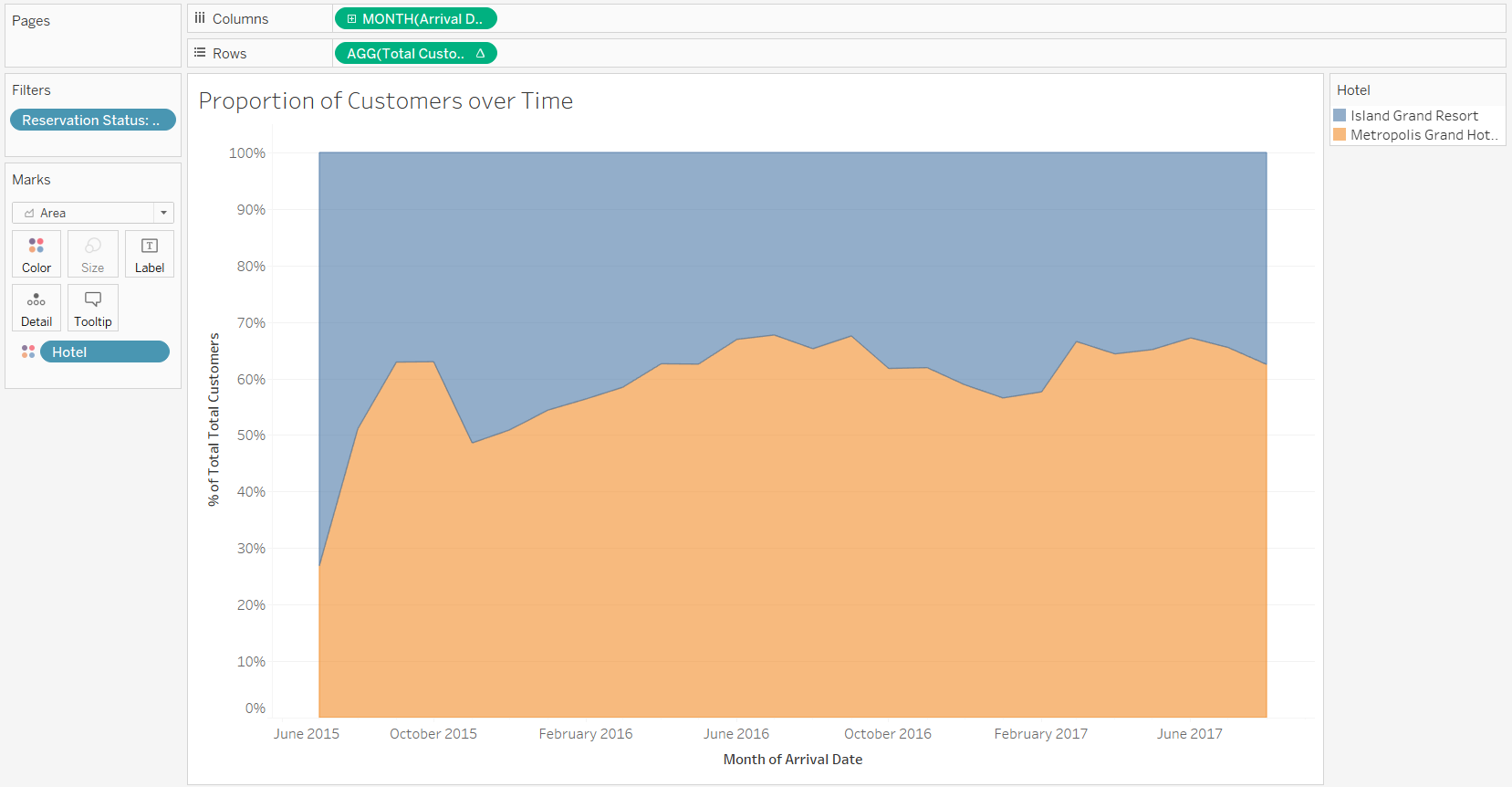
Hovering over any part of the dual axis chart, presents the user with the month and year, the total revenue for that month and year and lastly, the total number of customers who were staying at the hotel during the month and year.

**The Core Findings and Insights:**

* The total revenue made per month is related to the number of customers per month.
* The most profitable month for NP Management Ltd was August 2017 where the total revenue was $3348626.53. The least profitable month was January 2016 where the total revenue was $403694.85
* The months where there is a spike in Customers and Revenue is from February to October while the drops are from October to January.

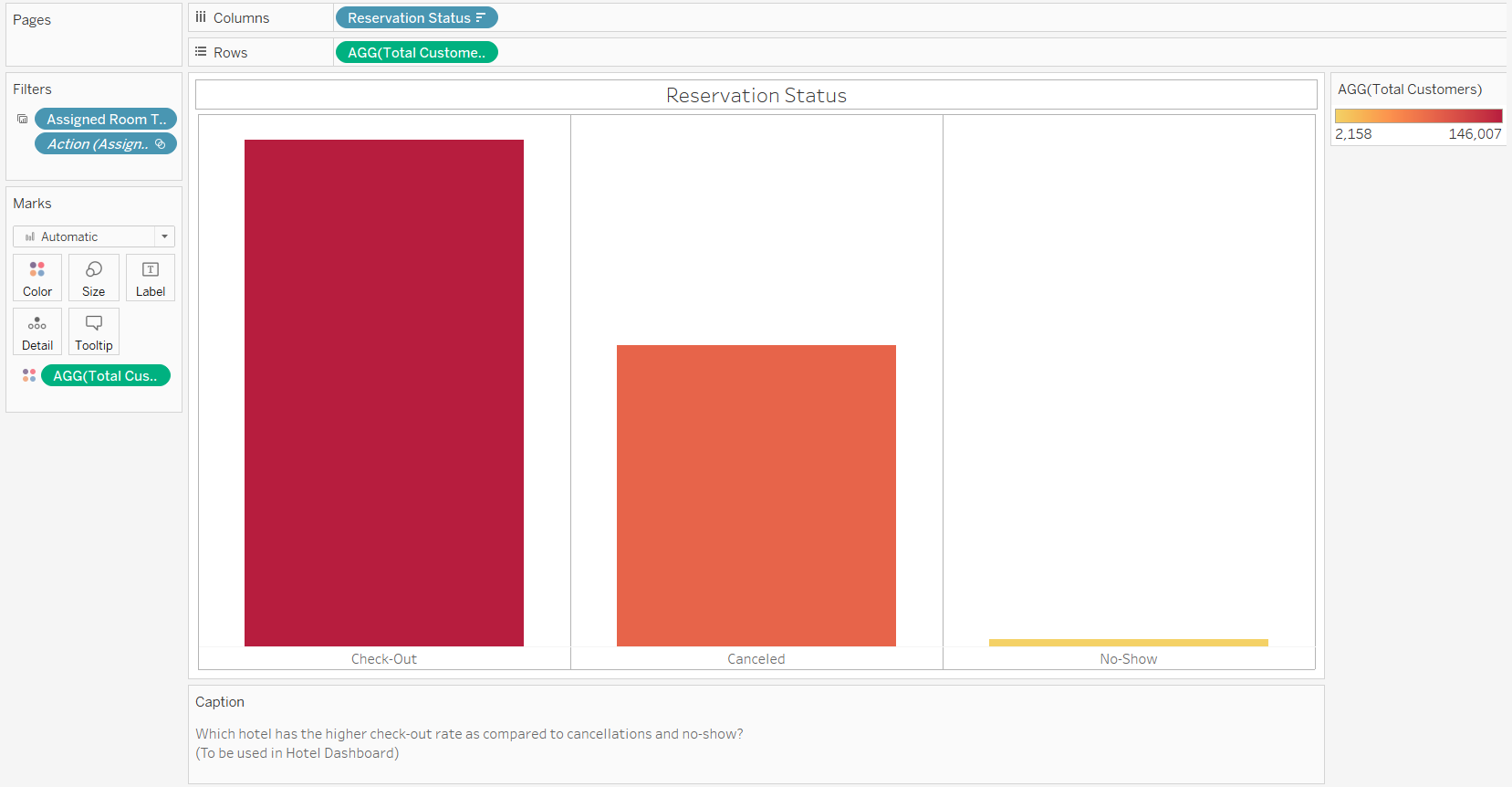
**Alternative Visualizations (Rejected):**





Alternative Visualizations I came up with while trying to attempt to answer this question was a Weekly Area chart based on the Total Customers over time and a Proportionate Area graph that showed the proportion of customers over time that belonged to both hotels. I ended up rejecting both of these visualizations as I felt like they did not answer the question and did not give much of an insight into the data. I ended up using the dual axis as it looked better, answered more questions and gives more of an insight into the data.

**Which hotel has a higher number of customers checking-out as compared to those who canceled their stay or did not show up?**

**Main Visualization: **

**Reason for picking Visualization:**

In order to answer this exploratory question, I decided to use a bar graph. The reason why I chose to use a bar graph is because, it shows the distribution of customers across the reservation status columns. The bar chart also allows a better and easier comparison.

**What the Visualization shows the user:**

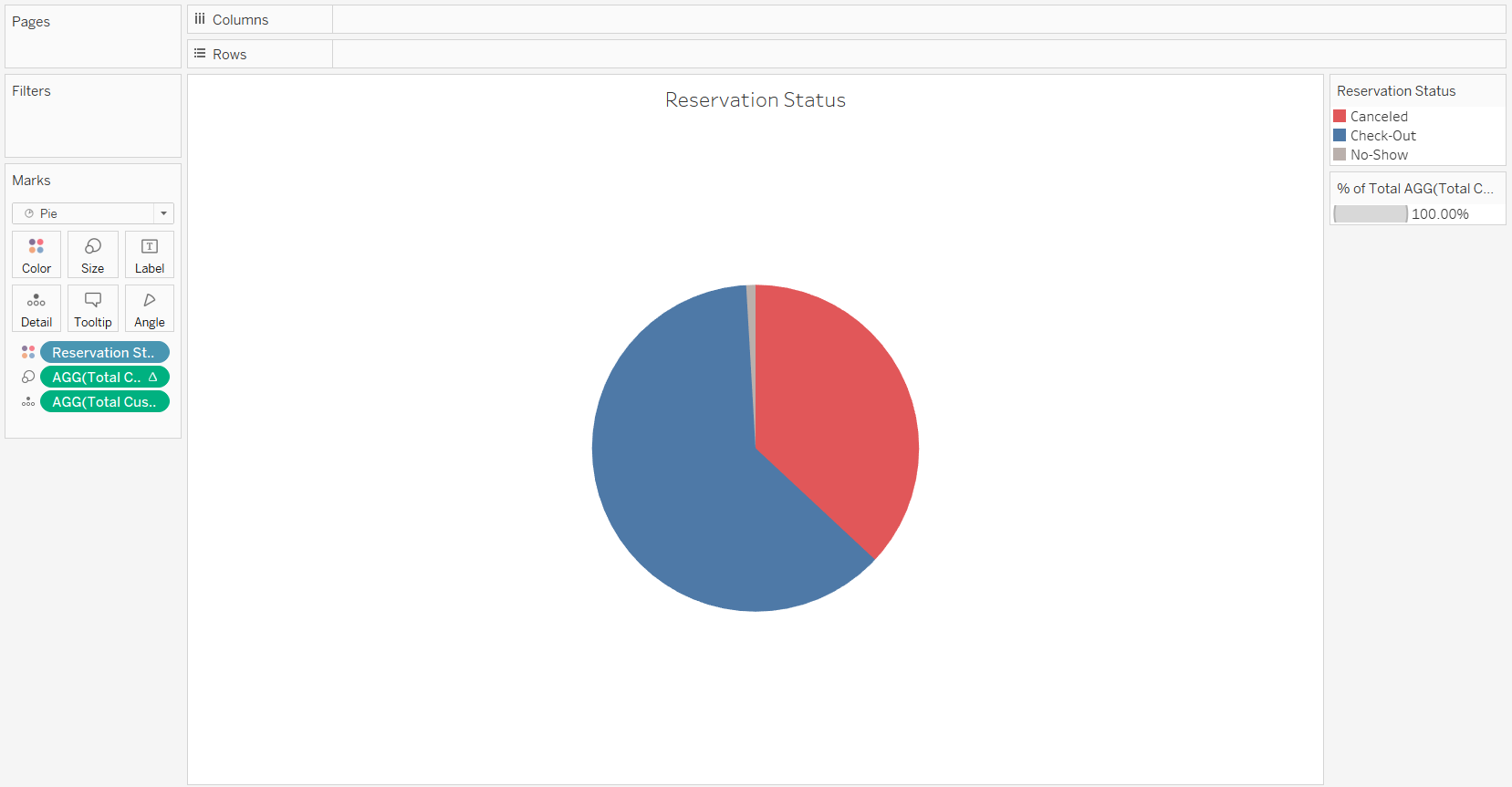
The visualization shows the user the distribution of customers across the reservation type. It allows the user to observe how many customers checked-out, how many canceled their stay and how many did not show up at all. The bars are colored by the total number of customers hence the darker colors mean that there are more customers.

Hovering over any bar in the bar chart presents the user with the name of the reservation status, the percentage of customers the bar is representing and the total number of customers the bar is representing.

**The Core Findings and Insights:**

* 0.92% of all customers did now show up to the hotel for their stay.

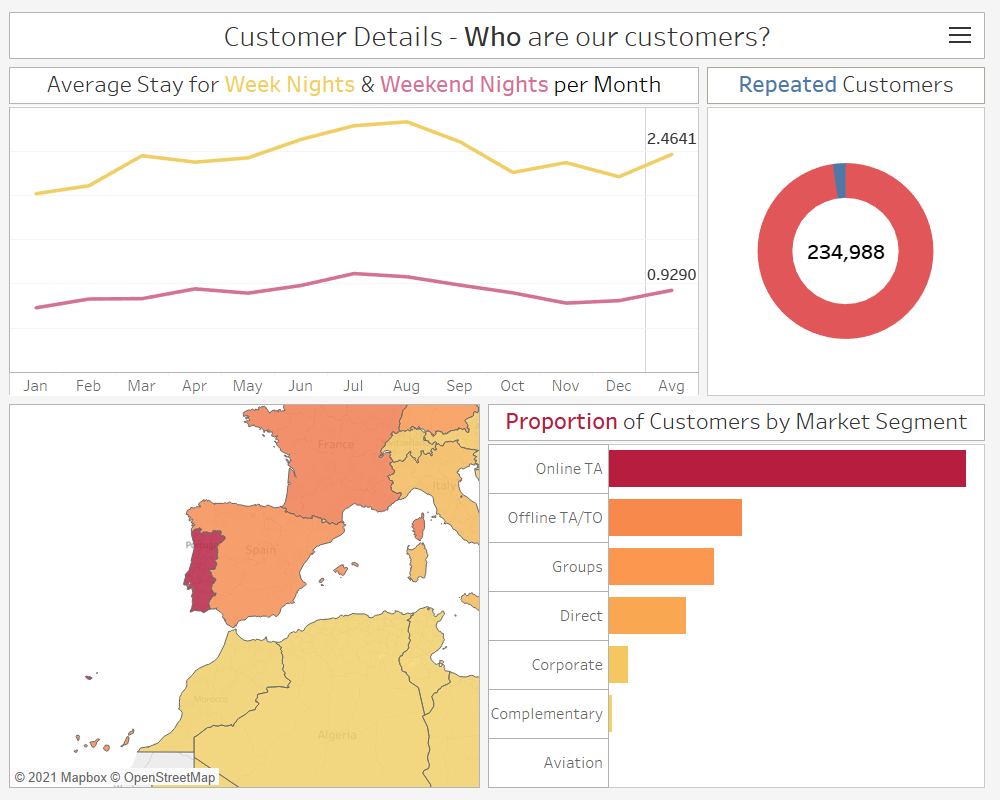
**Alternative Visualizations (Rejected):**

****

An alternative visualization I was considering to use was a pie chart for the reservation status and the total number of customers. This chart works as well as the main bar chart however, I felt like it was harder to compare the proportion of the customers. I also felt like the bar chart would fit in better with the dashboard I had in mind. Hence this pie chart was rejected.

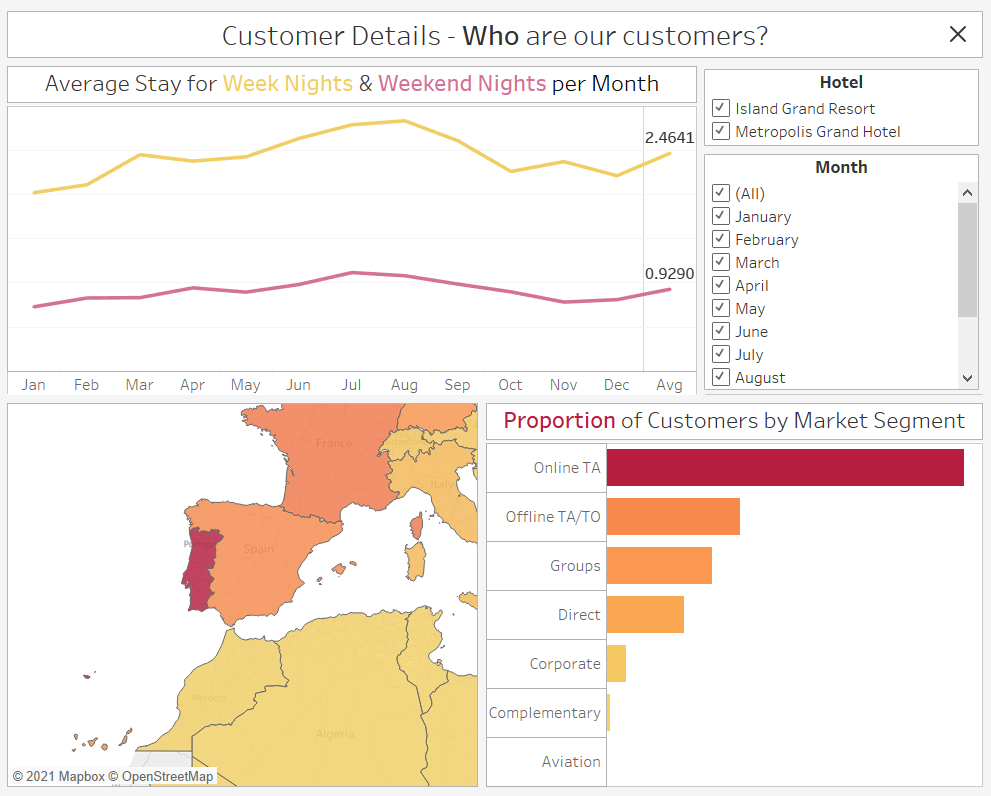
4. Dashboards

**Dashboard 1: Customer Details**

****

(Without Filters)

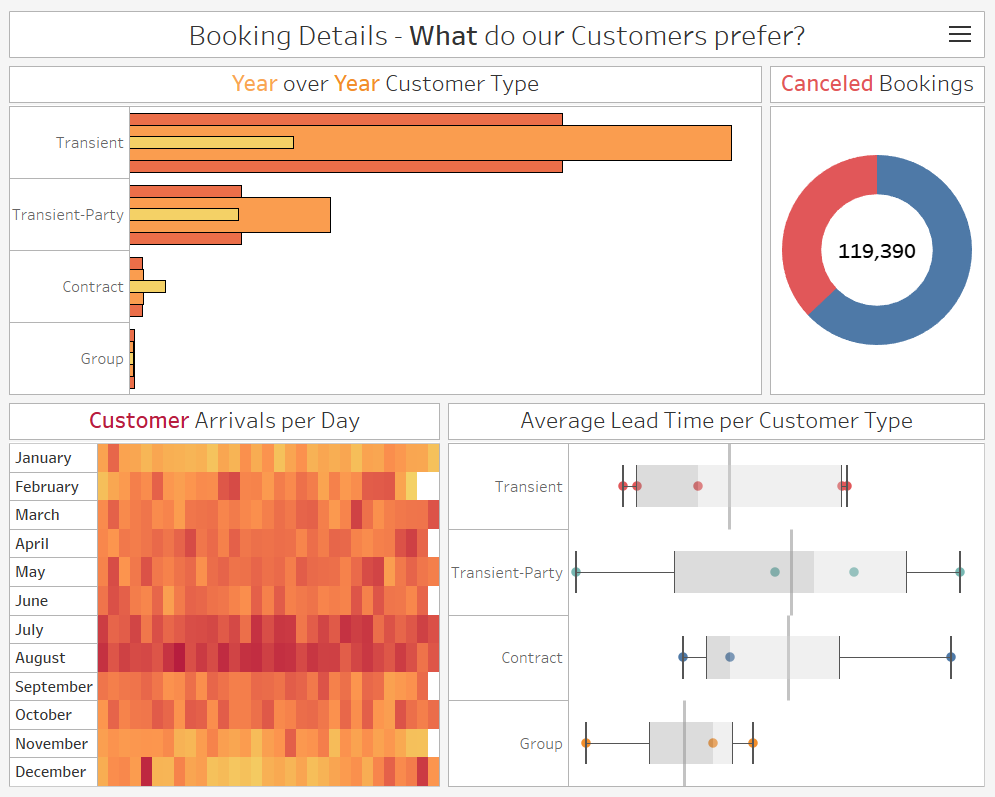
The customer dashboard takes the 4 visualizations that give an insight into who the customers are and combines it nicely. This dashboard shows the user what the average stay is in terms of weekday nights and weekend nights across all the months, as well as the average. It also tells the user about how many of the customers the hotels have received are repeat customers. Another thing the dashboard tells the user is where the customers are from and how many customers come from a certain country. Lastly, the dashboard tells the user about the proportion of the customers by the market segment. This tells the user about the ways in which the customers booked their stay at the hotels.



(With Filters)

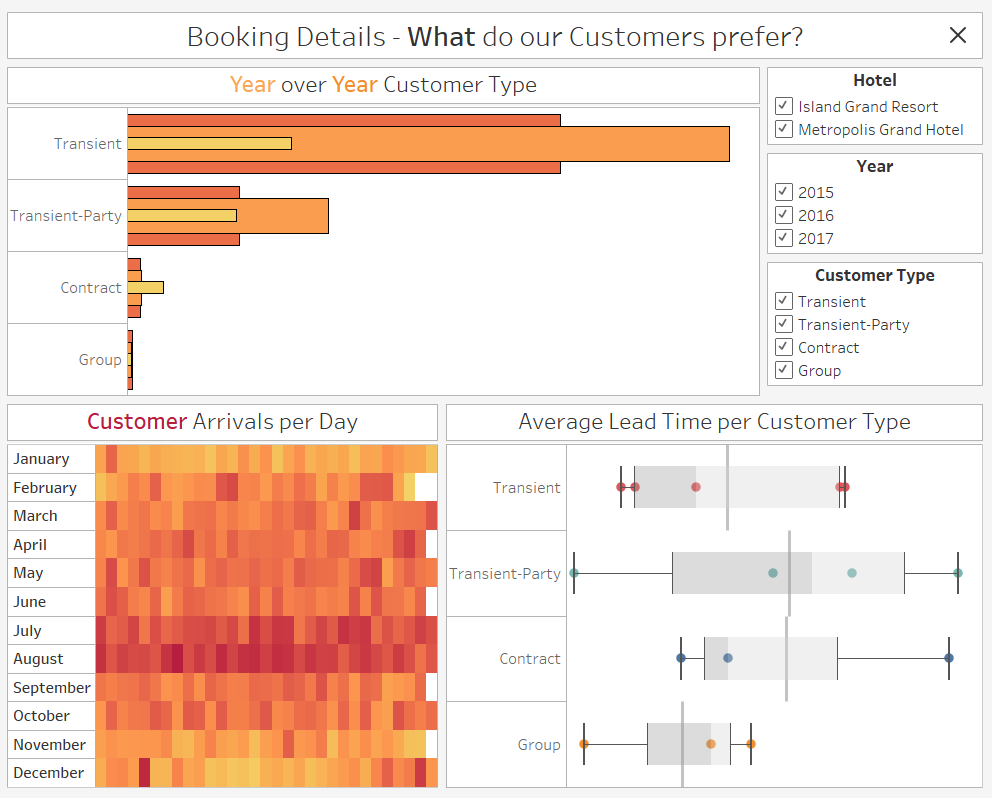
With the filters, the user can filter the customer details based on the hotel name and the name of the month. Clicking on a country in the filled map also filters the results of the other visualizations to match the country data. Lastly, clicking on the name of the market segment filters the result of the other visualizations to match the market segment data.

**Dashboard 2: Booking Details**

****

(Without Filters)

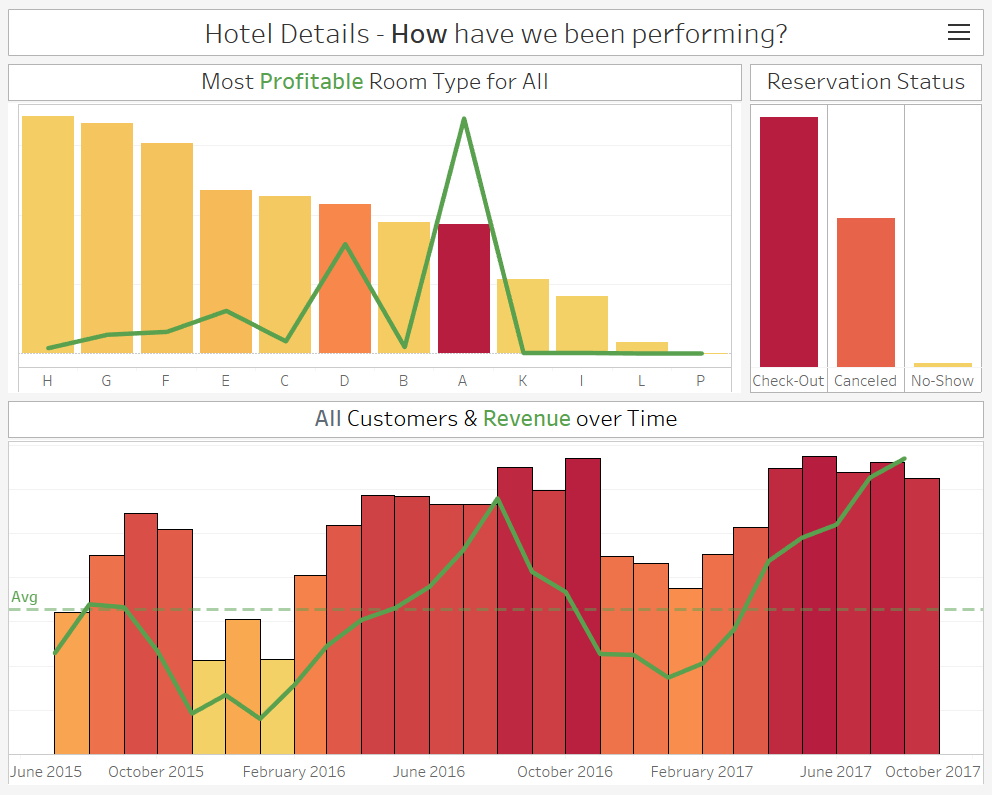
The booking dashboard takes the 4 visualizations that tell the user more about the customers preferences while booking and combines it into a cleanly. The dashboard shows the user the variation in the number of customers that choose the customer type over the 3 years, with the yellow bar being 2015, orange bar being 2016 and the darker orange bar being 2017. Next, the dashboard tells the user about how many bookings were cancelled compared to those that were not. It also tells the user the total number of bookings. Following this, the dashboard tells the user about the most popular days in each month for customer arrivals. The darker the cell on the graph, the more popular the day. Lastly, the dashboard provides an analysis on the average lead time by customer type, depending on the booking distribution channel used.



(With Filters)

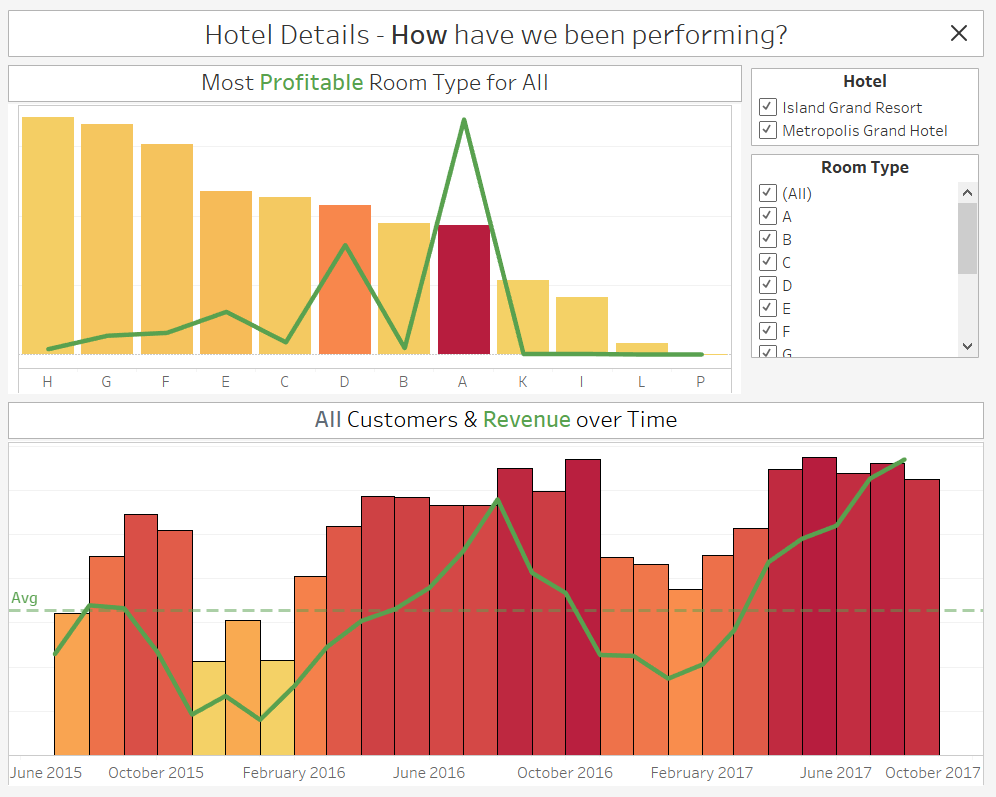
With the filters, the user can filter the dashboard by the hotel name, the year and the customer type. By clicking on any day or month in the heatmap, the user can filter the results of the other visualizations to match the day or month selected.

**Dashboard 3: Hotel Details**



(Without Filters)

The hotel dashboard takes 3 visualizations that provide an insight into the hotel performance and combines it perfectly. In the dashboard, the user can view how profitable each room is by looking at the average adr which is the bar chart and the total revenue which is the green line chart. The bar chart for room type is sorted by colour based on the sum of the total number of customers per room type. The dashboard also tells the user about the hotel’s proportion of customer reservation status, how many customers checked-out, how many cancelled and how many did not show up. Lastly, the dashboard shows the user the total number of customers over the months sorted by colour based on the sum of customers for that month. Alongside this, there is a green line chart on top of the customers over the months bar chart which represents the total revenue for that month. This makes it easy for the user to discover the trends between number of customers and the total revenue.



(With Filters)

The filters allow the user to filter the visualizations based on hotel name and room type. Clicking on the month will also filter the rest of the visualizations to match the month selected.