**Analysis of Travel and Ratings**

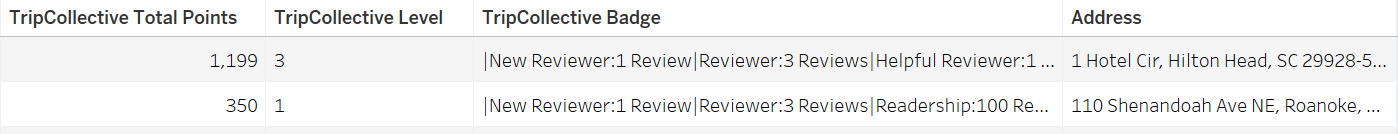
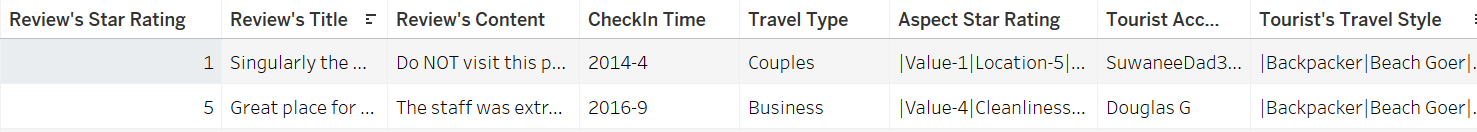
Group 3

Anh Nguyen, Mark Mondville, Mary Thanam, Sai Nalluri

**Introduction**

TripAdvisor is a travel company which operates on web or mobile apps. It offers reservations for the hotels, travel experience advice, transportation and restaurants. It collaborates with a large number of hotels, restaurants and transport to give the customer various options to choose from. Customers who use TripAdvisor to plan their trip can review the previous guests’ experience from specific hotels. In this project we mainly focused on the hotel ratings, how they trend over time and what different factors that can impact the ratings. Here we focused on comparing the star ratings to the average ratings, average and individual aspects over time and checking for the external factors that may have influenced the overall timeline (Superbowl). Our primary intention was to provide sufficient insight so as to allow for relevant businesses to determine further directions for analysis.

**Dataset**

****

We used a TripAdvisor Hotel Rating dataset that was provided by Dr. Chih Hao Ku. The dataset originally had 634,277 records but we filtered it down to 617,281 records. Data records from the years prior to 2006 and after 2016 were removed due to incompleteness and low number of records. The modified dataset’s timeline ranged for 11 years from 2006 to 2016, which had peak entries in 2015. This dataset was put together by combining the data from 752 hotels across the US continent. The ratings from the dataset primarily focus on data provided by two sources, Giata (a provider of non-bookable content) and TripCollectives (TripAdvisor statistics). The ratings from reviewers were also considered by averaging available individual aspect ratings along with Giata star ratings. The images above show an overview of the dataset’s attributes and some records as examples.

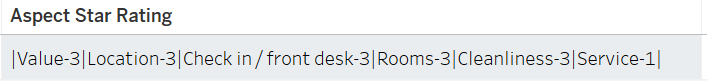
**Data Preprocessing**

From the TripAdvisor’s Hotel Ratings dataset, we primarily focused on these variables as we did time-based analysis:

* Aspect Star Ratings
* CheckIn Time
* Travel Type
* Review’s Star Rating
* Tourist’s Travel Style
* TripCollective Level

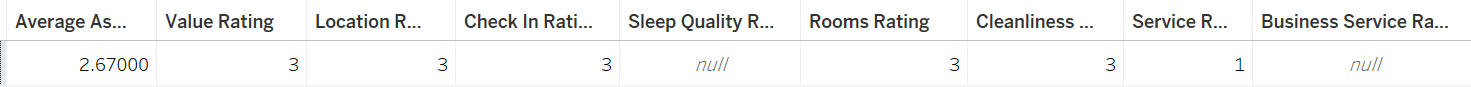
In the original dataset, the Aspect Star Ratings field included multiple values such as Overall Value, Location, Check In/Front Desk, Sleep Quality, Rooms, Cleanliness, Service, Business Service. To make it easier for us to do analysis on these different aspects, we used Tableau to separate them out as individual respective columns and then created a calculating field named Average Aspect Ratings (we will refer this as just Aspect Ratings throughout the rest of this report), which is the average of all these values within the Aspect Star Rating field. Below is the demonstration on how we split the Aspect Star Ratings column.

Before restructuring Aspect Star Ratings:

****

**↓**

After restructuring Aspect Star Ratings:

****

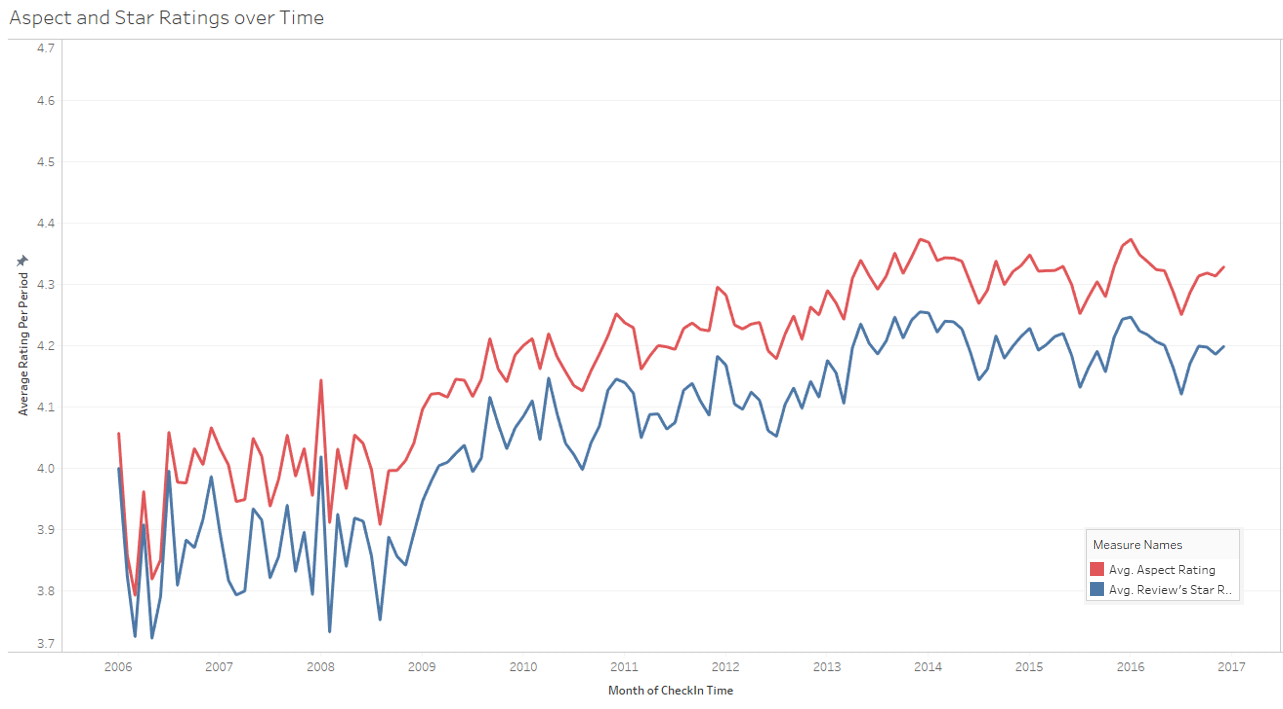
We also used Jupyter Notebook and for more data preprocessing, analysis, and visualization for the report. Python libraries that are used in our project include:

* Numpy
  + Input data as arrays
  + Column Stack
* Scikit-learn
  + Preprocessing data
  + Label Encoder
  + Gaussian Naive Bayes classifier
* Matplotlib
  + Visualize data by plotting

**Data Analysis:**

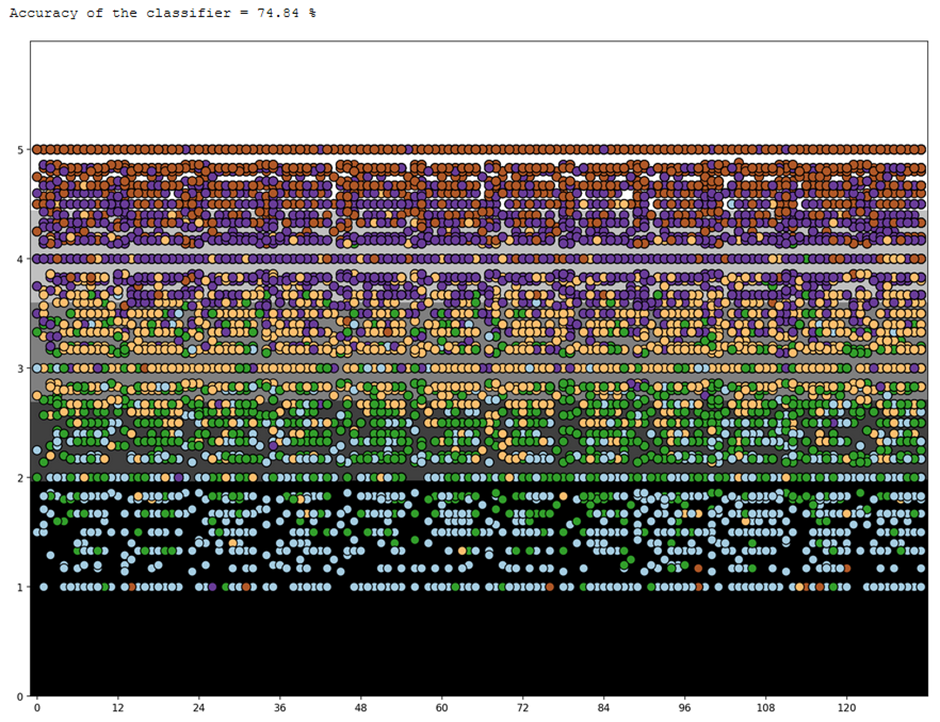
We used the top-down approach for this project where we focused on the proposed business questions and analyzed data to answer those questions. By considering various business questions, our research has led us to various solutions on how to analyze and predict the business growth.

1. How accurate are the Aspect Ratings in predicting overall Review’s Star Ratings?



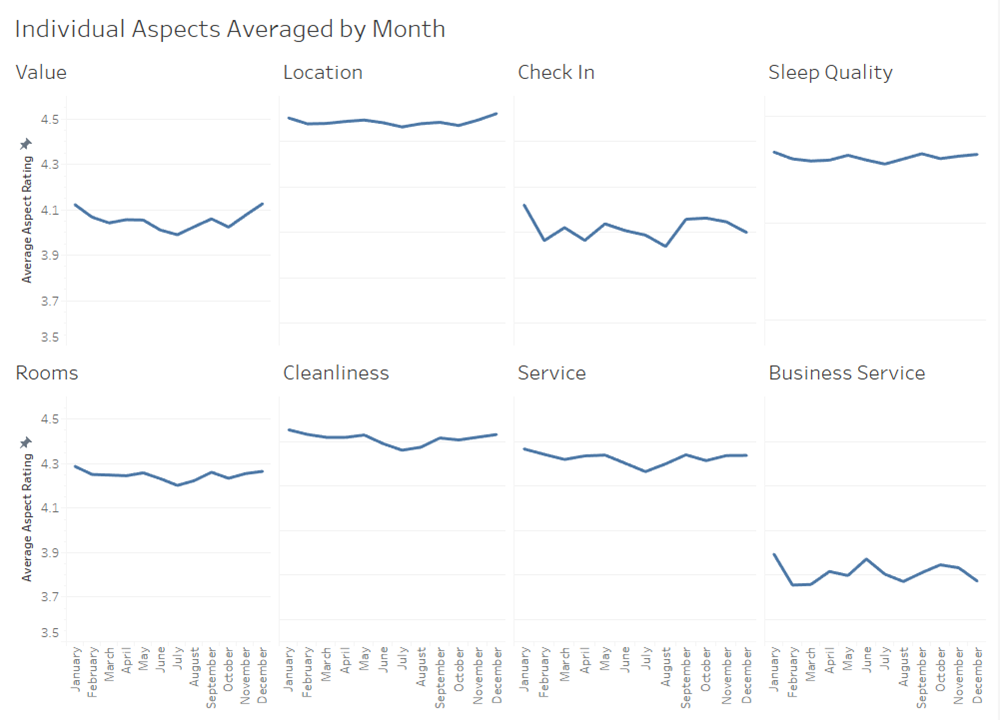
The graph above shows the average of Review’s Star Ratings and Aspect Ratings over time. The Review’s Star Rating is the overall star rating provided by Giata; the Aspect Rating is the rating computed by averaging ratings in different reviewer based aspects. In this graph we discovered a strong correlation between these two types of ratings.

In order to further analyze this correlation, we built a Gaussian Naive Bayes classifier:



By using Naïve Bayes classifier on the aspect ratings, we predicted the individual star ratings with an accuracy of 74.84%. This shows that there is a high correlation between the two types of ratings. This can be utilized by hotel owners whose businesses are not big enough or not incorporated to have their own star ratings to have a general idea of what star ratings they would have based on any reviews that they may have received at that point. Giata could also use this analysis to further refine their process of ratings to approach existing user aspect ratings, making their product more valuable.

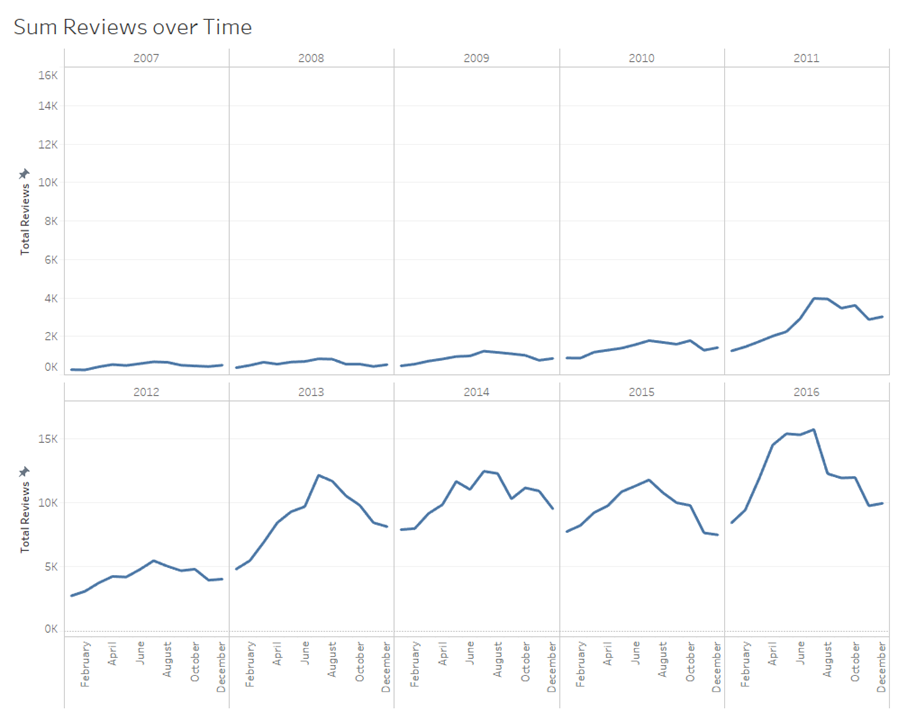
1. How do individual Aspect Ratings compare and do they show trends over time?



The above image compares the average of different individual Aspect Ratings over the months. We could see ratings from various aspects, but within every aspect we can see a dip in average aspect ratings in summertime. This finding can be useful for hotels as they can focus more on providing better services especially during the summertime. For example, hiring more staff to assist customers, cleaning more frequently, or offering special room services.

The aspects that are most harshly rated are Value, Check-In, and Business Services. This can be helpful for hotel management as they know in what aspects they should consider more in improving their services. For example, to improve Value, they can offer services at a lower price, or offer more special services and utilities at a fixed price. To improve Check-In, they can train employees to deliver better customer services. To improve Business Services, they can provide customers with better facilities such as better wifi connection, computers, printers, TVs, etc.

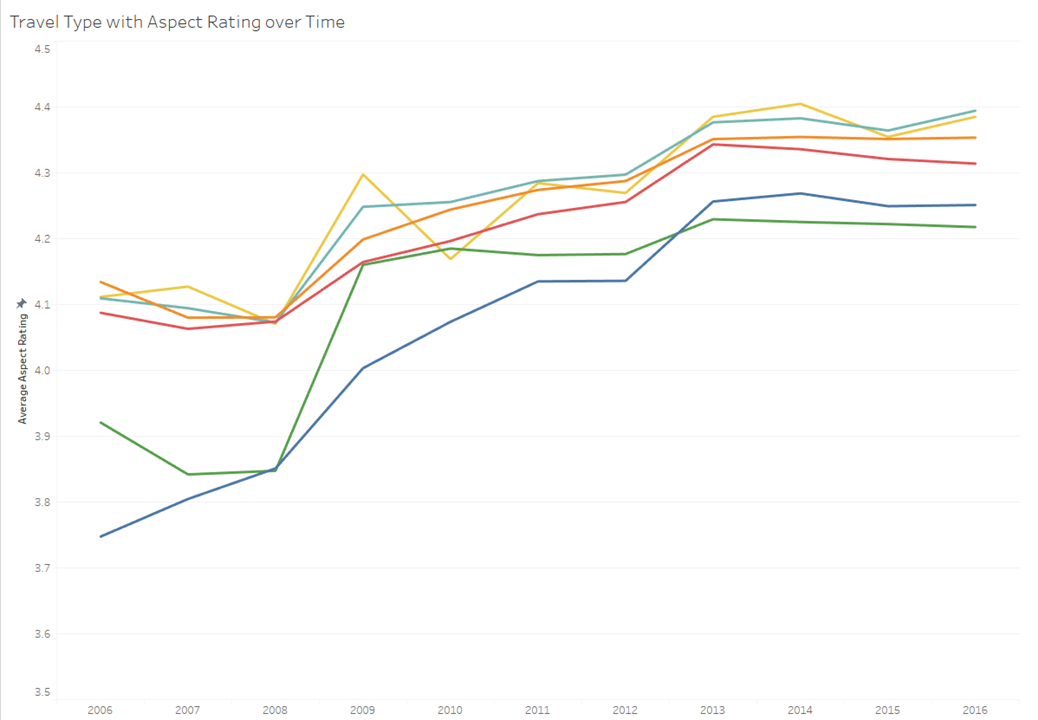
1. Do the number of reviews reveal trends over time?



The above graph shows the total reviews that were received over time. We can see that the total amount of reviews peaked during the summertime (especially in July), whereas in the previous question, the ratings dropped in around the same time.

With this finding, the hotel owners can expect to have a rising number of customers during summertime and may also expect a drop in average ratings if they fail to provide good services to the customers.

1. Does Travel Type reveal trends over time?

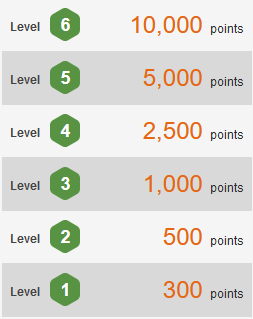
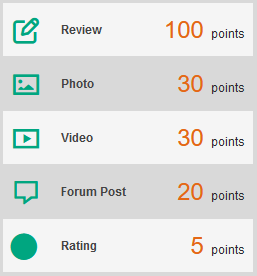


This graph shows Aspect Ratings of different Travel Types over the years. We can see that overall, the lowest line is the blue one that represents the Business type. This is probably because unlike other types of travel, the Business travels are typically not done for pleasure, and are done more frequently, thus making these types of customers more sensitive and harsher when it comes to hotel services.

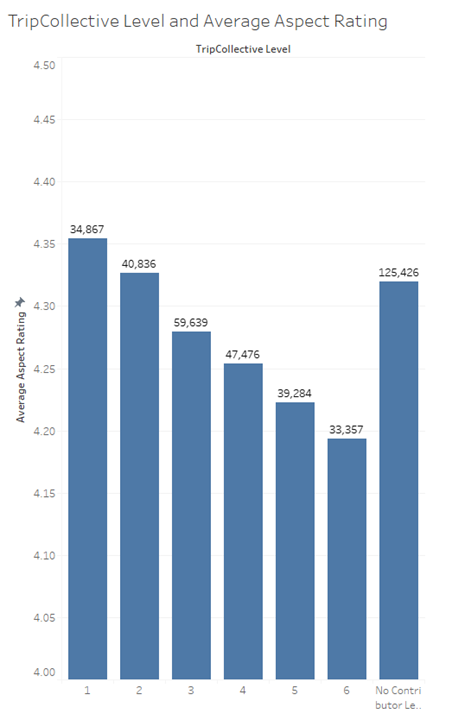
This finding can be utilized by hotels to find ways to improve their services for the business travelers. For example, asking customers in advance to see what the purpose of the stay is, and then offering specific services that may satisfy business travelers, or any other travel types.

We can also observe that Aspect Ratings of all Travel Types were increasing over time. This implies that the overall hotel service quality has been improving throughout the years.

1. Does TripCollective level impact aspect ratings?

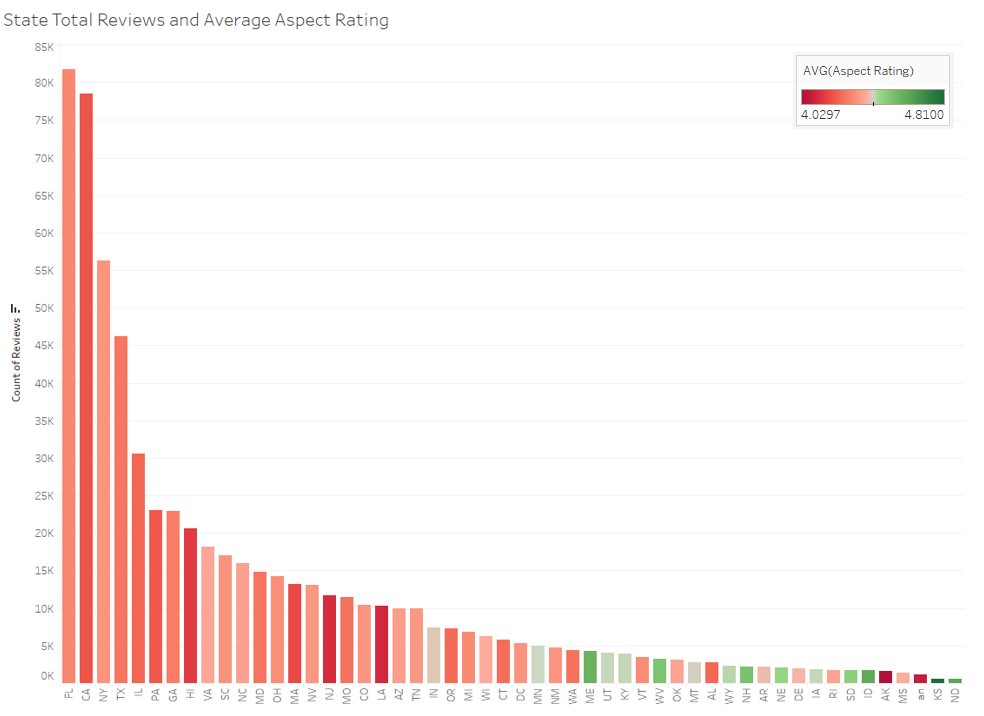


These two images explain how a user can earn points and achieve a level. When they complete different tasks, they will receive a certain amount of points for each certain task. And when they reach a certain amount of total points, they can achieve a higher level.



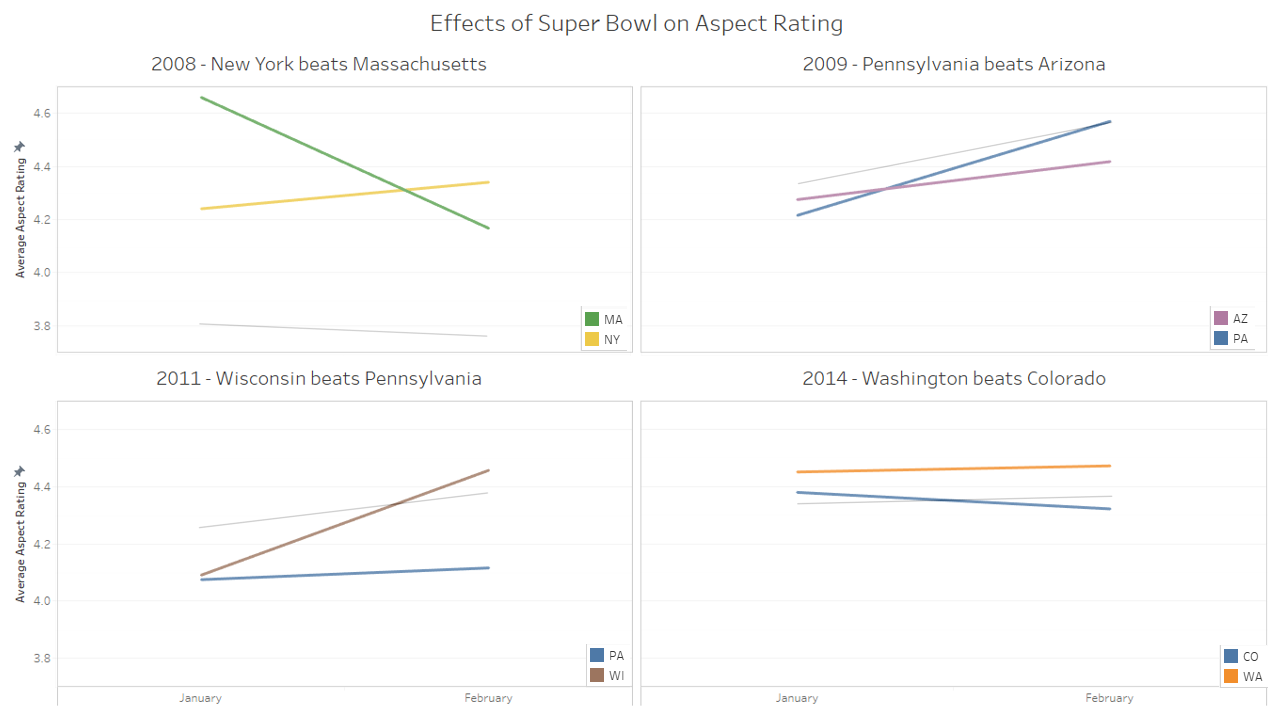
Through this graph, we can observe that, as users gain more experiences, write more reviews and achieve higher levels, they tend to give lower reviews compared to other lower levels’ users. This also raises an issue that some users might be affected by the benefits that come with high TripCollective points, thus trying to make more unnecessary harsh reviews or even fake reviews to raise their points, which somehow affects the average rating

1. How do states rank for total reviews and aspects ratings?



This graph compares the total amount of reviews and the Aspect Ratings from all states in the United States of America. As we can see, the bigger, more populated states have a greater number of reviews but a relatively low average Aspect Rating, while in the smaller states which have fewer ratings, the Aspect Rating is higher than big states. One possible reason for this is that because there are more hotels in the big states, the market is more competitive compared to that of smaller states. It also suggests that the overall (nation-wide) average is lower than that shown for smaller states. The lower numbers of reviews in smaller, less populated states implies a false sense of superior ratings.

1. Are aspect ratings affected by real world events?



The graphs above show the Aspect Rating, from January and February of a particular year, of the states that participated in the Super Bowl. We found that the majority of the years that we reviewed showed a decrease in ratings in the losing teams and an increase in winning teams as compared to the overall general trend over the same period. The four graphs shown above are the strongest examples of what we found.

The gray line represents the overall trend of average Aspect Rating nationwide. In most cases, the ratings of the winning team during this period tended to increase and trend over the overall nationwide ratings. Whereas the ratings of losing states tend to decrease or increase much slower than the winning states. It also shows the winning states ending with higher ratings than the losing states, even when the winning state began rated lower.

The state team who won the Super Bowl has not only brought the cup but also many benefits to the travel industry. The hotels in these states can utilize this event to improve their business, such as hold celebrations for winning the Super Bowl, sell merchandise of winning teams, or offer discounts for the team’s fans.

In contrast, the hotels in states of losing teams may want to be more cautious and aware of the actions that may unintentionally upset the customers. They may want to take down Super Bowl posters and avoid mentioning the results that can negatively impact the customer’s mood.

**Conclusion**

In this project, we have mainly used Tableau and Python to preprocess, analyze and visualize data and discovered and discussed many different findings. We found that there is a strong correlation between the Review’s Star Rating and Aspect Star Ratings. We also built a 74.77% accurate Naive Bayes classifier that proves the relationship of the two types of ratings. We found that the number of reviews tend to increase while ratings tend to decrease during the summertime. The aspects that are most harshly rated are Value, Check-In, and Business Service, and the travel type that usually gives harsher reviews is the Business travel type. We also found that the users with higher TripCollective level tend to give lower ratings, and the more populated states have a higher number of reviews but a relatively lower average rating than those of smaller states. In addition, we discussed how aspect ratings can be impacted by real world events. In this project, we mainly focused on the impact of Super Bowl results to the average aspect ratings of winning and losing states. States of teams that won the Super Bowl might be positively impacted and experience a significant increase in ratings, whereas the state of losing teams might be negatively impacted and even show a decrease in ratings.

Further research could be aided by having more complete and up-to-date data including more recent years and a daily check-in time breakdown. Additional analysis could also be conducted by considering specific locations where the Super Bowl was played and the home city’s of the playing teams as focal points. Aspect Ratings could be viewed in relation to the proximity of the focal points using zip codes.