590V Final Project Report

Chenhao Huang, Ethan Johnson, Renjie Li, Danyang Liu

Selection of Data and Problem

Dataset

We plan on using the dataset of 515K hotel reviews data in Europe on Kaggle. The dataset was scraped from Booking.com, and it was released under CC0 public domain from Kaggle in 2018.

The dataset contains 515,000 customer reviews and scoring of 1493 luxury hotels across Europe. The data contents and metadata of this CSV file are listed below:

1. Hotel_Address: Address of hotel.
2. Additional_Number_of_Scoring: There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there.
3. Review_Date: Date when reviewer posted the corresponding review.
4. Average_Score: Average Score of the hotel, calculated based on the latest comment in the last year.
5. Hotel_Name: Name of Hotel
6. Reviewer_Nationality: Nationality of Reviewer
7. Negative_Review: Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
8. Review_Total_Negative_Word_Counts: Total number of words in the negative review.
9. Positive_Review: Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
10. Review_Total_Positive_Word_Counts: Total number of words in the positive review.
11. Reviewer_Score: Score the reviewer has given to the hotel, based on his/her experience
12. Total_Number_of_Reviews_Reviewer_Has_Given: Number of Reviews the reviewers has given in the past.
13. Total_Number_of_Reviews: Total number of valid reviews the hotel has.
14. Tags: Tags reviewer gave the hotel.
15. days_since_review: Duration between the review date and scrape date.
16. lat: Latitude of the hotel

17. lng: Longitude of the hotel

Problem

We plan to solve the problem of which variables in the data affect users' reviews of the given hotels. We plan to do this by looking into aspects of the data like relationships between reviews and reviewer's nationality, determining if certain mentions of words in reviews correlate to the polarity of the review, or determine if there is a correlation between hotel location and review scores. The tags for the review may be also meaningful to find trends related to the trip purpose and customer type.

Ref:

1. https://www.kaggle.com/jiashenliu/515k-hotel-reviews-data-in-europe
2. https://www.booking.com/

Initial Project Description

What is the problem being addressed?

- We plan to solve the problem of which variables in the data affect users' reviews of the given hotels. We plan to do this by looking into aspects of the data like relationships between reviews and reviewer's nationality, determining if certain mentions of words in reviews correlate to the polarity of the review, or determine if there is a correlation between hotel location and review scores.
    - What is the distribution of reviewers' scores?
    - What is the relationship between the frequency of reviews and scores?
    - What are some tags or keywords mentioned commonly in different kinds of reviews?

Who would be interested in understanding this data better (users)?

1. Travelers who would like to visit Europe and would like to pick the best/suitable hotel among many based on the reviews and ratings
2. Hotel booking app/website (ex. TripAdvisor) or Agents booking for other people
3. Hotel management
4. Hotel investors
5. Hotel marketing team
6. Third-party data analyst

What would they want to see in the data (tasks, why is the user looking at it)?

1. For the travelers, they would like to see the location of the hotel with its general information, the positive and negative reviews of the hotel, the average score of the hotel, etc. They go through all the information and determine which hotel they prefer to stay at. They are the customers and they want to spend money in the best way possible for the best service solution so they would be interested in seeing the data. Customers can also get a general idea of other customers' trip purpose and type during the stay.
2. For booking app/websites and agents who are booking hotels for their customers, they would like to see all the details and reviews from their own

professional perspective. They might be interested in which nationality/trip type of their customer prefers based on their budget and find the best solution for their customers.

3. For hotel management, they would like to see the positive and negative reviews and the average score of the hotel, etc. The main purpose for the management to review the visualizations is that they can adjust their services, examine the facility's environment, and evaluate employee performances based on the detailed reviews. They can then improve their score and performance in the future while attracting more customers in order to gain more profit. They can look at their customers' nationality/tags potentially to accommodate them.
4. For hotel investors, they would like to see the review and scores of the hotels that they invested in. They want to evaluate the hotel so they would be interested in the data to aid in their investing decisions. They don't directly interact with the customers, that is why the online data is important to see from a customer's point of view. They would like to see the locations with trends of negative or positive rating to determine places where they can invest a new hotel to make money.
5. For hotel marketing team, they can make investing decisions and potentially target advertisement towards certain nationalities. They can come up with marketing and sales strategies using information about how long people are staying during what time of the year with what type of the room.
6. For third-party data analyzers, they would like to see all the information and try to find any correlation between the review, score and location / hotel names / reviewer's nationality; determining if certain mentions of words in reviews correlate to the polarity of the review; finding trends in the industry and make reports for different purposes, etc.

Where is the data coming from and what are the characteristics?

The data is pulling from booking.com, and the dataset is downloadable from Kaggle. It has 17 fields, 8 of them contain dimensions data, 9 of them contain measures data.

The data gives a detailed description of the individual reviews for hotels in Europe including information such as the review given, the reviewer's score, nationality, the number of nights stayed, etc. The data also gives the location of the hotel, relevant tags the reviewer gave, whether the review was positive or negative, and more general information such as the average score of the hotel.

Why is this a visualization problem and cannot be solved with ML/statistical analysis/etc.(why Visualization is needed to solve this problem)?

1. In this problem, we want to find out location of hotels and their distribution on the map. This information can be delivered efficiently by data visualization but not by statistical analysis because it is not necessarily enough to know which hotel to pick, but also to understand why certain hotels are better than others nearby for reasons other than rating.
2. Since the data users are mainly travelers, investors, and hotel management etc., the statistical analysis is too implicit to understand. Data visualization could help them identify the problems they interested in immediately while at the same time understanding more about the hotel and area nearby through other customers' reviews.
3. Personal preferences are different from customer to customer, there might be different hotels satisfying similar standard by a customer, but the customer would have to make the final decision based on their personal experiences and expectations. Not everyone is trying to find the very best hotel in the area but find the best suitable hotel for the travelling plan.

Studies in European Hotel Reviews Proposed Solutions

Chenhao Huang, Ethan Johnson, Renjie Li, Danyang Liu

1. Overview

We started thinking because it is geographic information the main interface should intended to display and select hotels with different reviews. Once a hotel is selected, there should be more informations about the reviews for the hotel being displayed. The workflow for the user should be simple and intuitive to see different representations behind the hotel reviews once selected. Here we had several ideas for the data visualization and proposed solutions for implementation.

2. Map

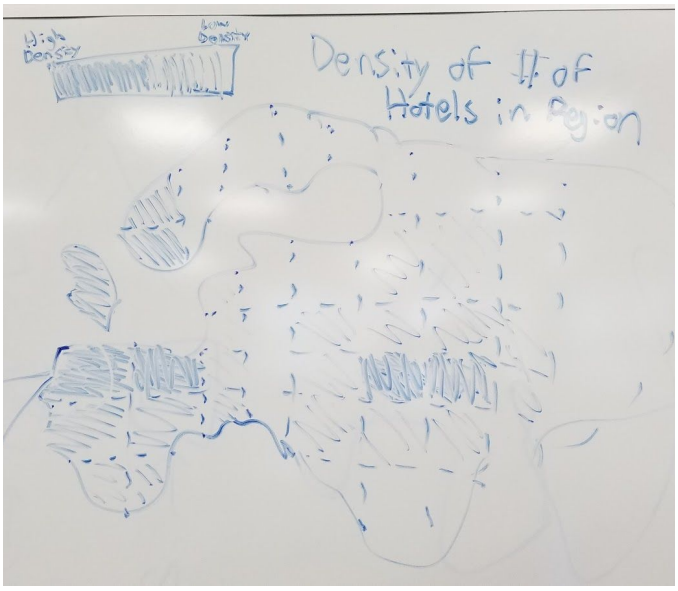


Figure 2.1

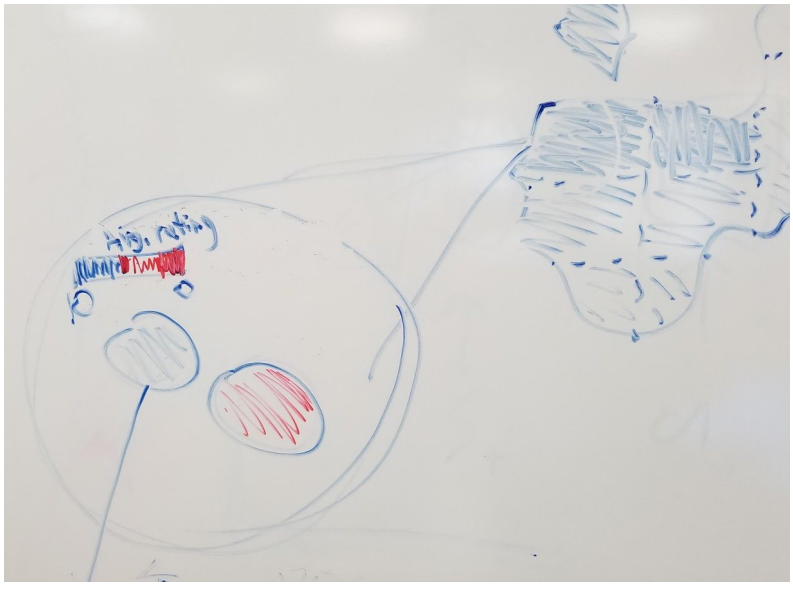
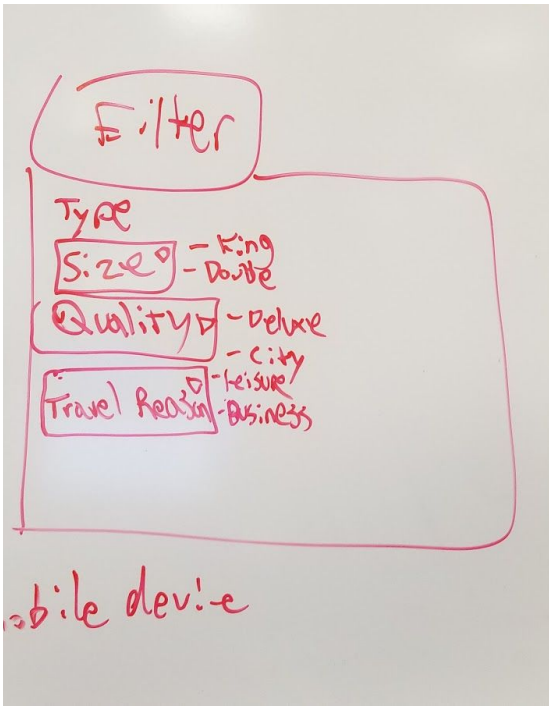


Figure 2.2

This design idea(Figure 2.1) is a density grid where each grid on the map was a geographic location on the map and the saturation of each cell in the grid is the number of hotels in that region. If you were to zoom in, it would look similar to the area in the circle in the right image where each hotel is a dot, where the color describes the

average review score of that hotel using a sequential color map. We also considered using this same design to show the density of the number of reviews in each cell, but ended up altering the design as although it was relatively somewhat unique, it did not provide any useful information, as the number of hotels or reviews would likely be due to the population of that area and more likely the presence of certain countries or cities.

We decided instead that a bubble map (Figure 2.1) would be more useful, where each bubble is a country or city depending on the level of zoom and the size of the bubble is the number of hotels or reviews in that country or city. This would also work to show density while avoiding the issue of not representing different countries or cities equally. In addition, we found that using sequential color map with lower average review scores being red and higher average review scores being blue or green as shown in the right image would allow us to determine relations between areas of the map and their average review scores nicely at a glance, as in practice this has shown to scale well if the map were to be zoomed in.



This image (Figure 2.3) shows a design idea for a filter system we are considering implementing that would filter the data points in the geographic map visualization to help show the relations between different tags and the average scores of reviews. We consider expanding the filter system to include the months the reviews are given as well as the nationality of the reviewer, as these are all attributes that could affect a user's review score.

Figure 2.3

3. Tags

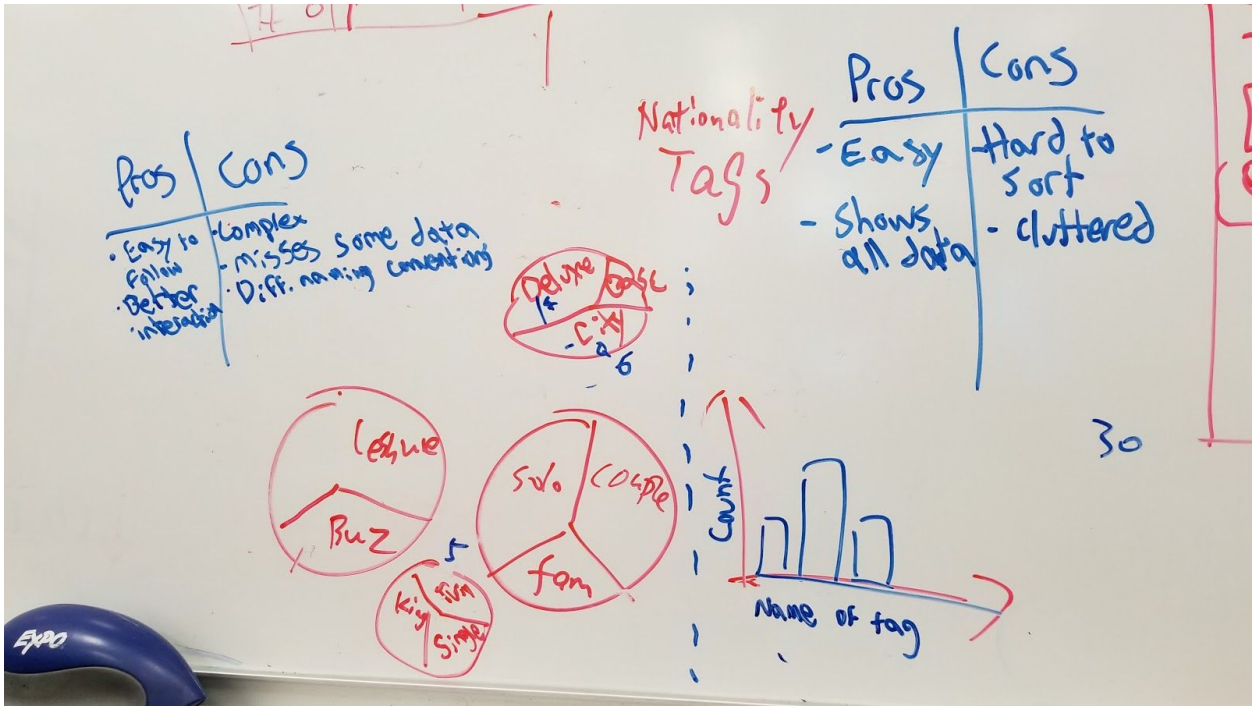


Figure 3.1

Here we have two different ideas on how to visualize the “tags” field of the data. The option on the left was to show each of the different types of tags as a separate pie chart. The option on the right displayed the same data, but as a bar graph for the entire dataset, not categorized.

We settled on using something in between where we show a separate bar graph for each of the categories as the bar graph showed a better comparison between the different tags than the pie chart, which is what we are trying to accomplish. In addition, we decided that showing all of the data for all of the tags at once was a bad idea, as comparisons between tags of different categories could occur, which would cause confusion.

4. The Big Picture

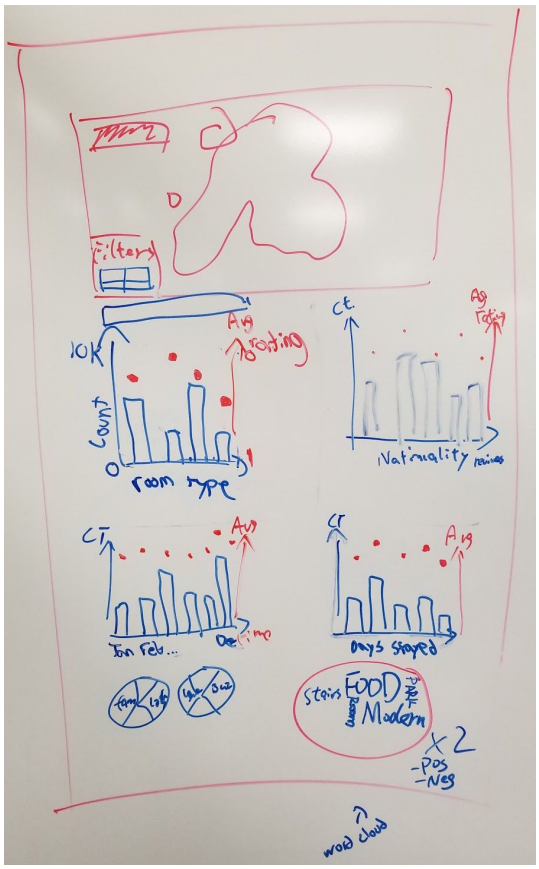


Figure 4.1 is our sketch of the general design of our planned page. The top shows the geographic bubble map on the selected data, which is filtered by the area on the bottom-left of the bubble map. These filters determines what is being counted (number of hotels or number of reviews) as well as which categories of the data to show. The bubble map would also have its color dependent on a sequential color map similar to the zoom-in in the first image. Also, the map would be interactive and you can zoom in on different parts of the map and the bubbles would change size and clusters appropriately, changing from country to city, city to hotel, etc.

Figure 4.1

Below the map are bar charts of several different features on the x-axis such as the room type or days stayed in the hotel, and on the y-axis for all of them are the total count of the reviews that fit the category described as well as the average review score in that category for the selected data.

Other visualizations we think would be useful to help interpret which features affect user reviews are which words are most prominent in positive or negative reviews (not including stopwords) which is shown in the bottom-right of the page. Another possible visualization would be to show pie charts of the data using the same information as the the bar/dot graphs, except make the pie slices equal to the the percent count of reviews in that slice and the color would be a sequential color map with one end being low review scores and the other end being high review scores, as talked about above.

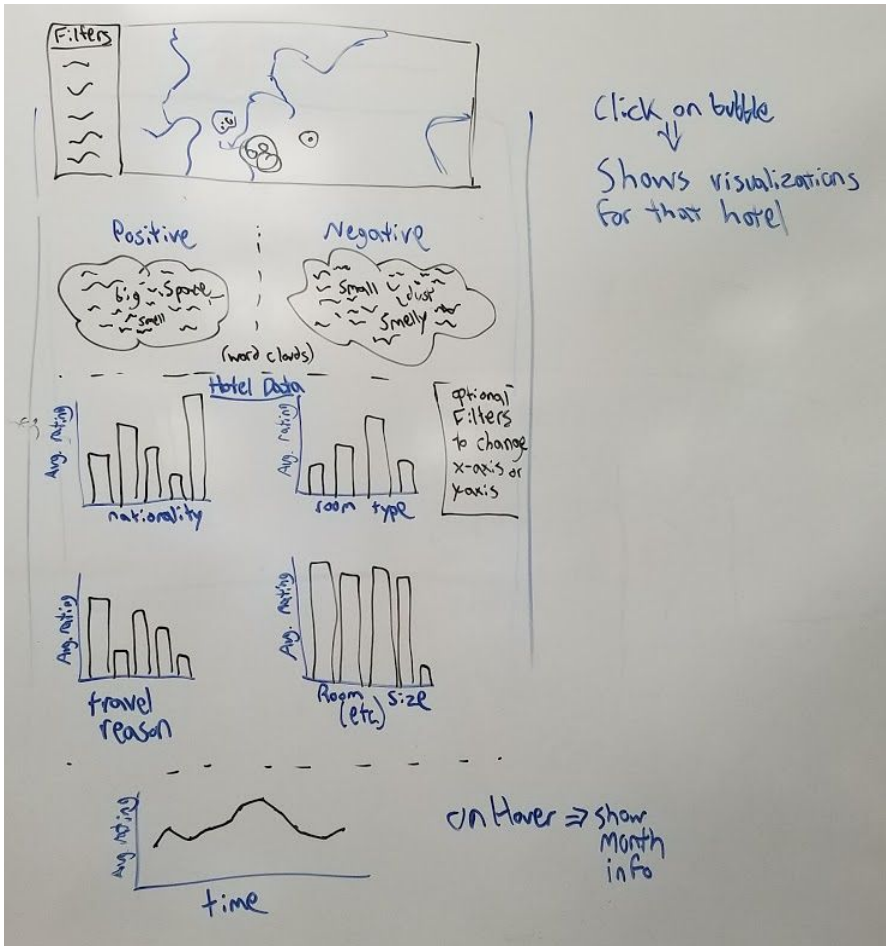


Figure 4.2 is an overview of what we wanted the final design of the page to generally look like. We removed the counts for each bar graph as we thought that it was not important information and added in a line chart that shows the average review score of the given hotel for each month. The map would also have zoom functionality.

Figure 4.2

Implementations

1. Summary

Our implementation is composed of several different languages from Python to JS. Our implementation process was decided by the size of our initial dataset and the features specified in the dataset. We knew initially that the geographic information would be a key part of our visualization and pairing this with the review scores would prove important.

Our final presentation consist of one geographic visualization, two word clouds, one pie chart, one line chart, and one bar chart. We chose such a variety of different visualization techniques because we found those techniques to be the best way to visualize the specific information represented in each of the visualizations.

2. Technologies used

- Python scripts for reformatting the data
- D3 for bar charts and geographic visualization
- JS to filter the data in separate ways
- Bootstrap for layout
- Leaflet map module for interactive map
- Jason Davies word cloud library

3. Platform

Any modern web browser

- Chrome
- Firefox
- Etc.

4. Data Handling

The data handling process was completed in two parts.

The first part includes split single tag column with clustered information into separate columns; delete unwanted columns like “Review_Total_Negative_Word_Counts” and “Review_Total_Positive_Word_Counts”; delete rows have null value; combine longitude column and latitude column into

coordinates column; delete hotels with less than 500 review records, and etc. This part is completed mainly by using Excel.

The second part includes build a json file with no duplicate hotels for displaying unique hotels on the map; assign a new column “group” to each hotel based on the average score it has; assign and calculate a new column “Avg_Score_Per_Month” to each hotel based on averaging the reviewers’ score it receives per month; build two json files to store both positive and negative review words for each hotel with specific format for displaying word cloud, and etc. This part is completed mainly by using python.

5. Implementation Details

We had an initial brainstorming session, writing our visualizations on a whiteboard. We thought about how easy or difficult each aspect of the visualization would be and split up the work, starting primarily with ways to reduce the amount of data we were working with to just the data we needed. We had one person work on splitting the tags into separate columns, others worked to reformat the data of individual columns or remove them entirely, and one person began research on different approaches to making the interactive map and bubble map for word cloud.

As we got the data into the format that we wanted, we learned that the size of the database was still too large and decided to work with a subset of the data while figuring out ways to further reduce the size of the overall dataset. One person began working on filtering the data using python. we were able to filter the data so that either only positive comments, negative comments, reviews for specific hotels, or other similar groups would be filtered on.

We then started to visualize the location by the position of circle. We first add svg layer on the map and set the axis--’cx’ and ’cy’-- based on coordinates of the hotel. To reflect the average score by the color of circle, we write function to input the average score to color sequential. To display number of reviews, we make the radius of the circle proportional to the number. Since there will be hundreds of hotel on the bubble map, we add a filter, which users could select hotels within certain range of score. To do so, we add checkbox in the div and assign the value to the checkbox. Then we define function to update circle displayed each time that checkbox was select or undetected. We also add mouseover, mousemove and mouseleave to our svg. In this case, a tooltip with hotel information will popup if you place your mouse on the circle. Besides that, click function is added to the svg, if you want to know more about this hotel, clicking it will provide you more information.This visualization could help users quickly identify which hotels are more suitable for them.

We implemented word cloud based on the click function. By clicking a specific circle on the map, the hotel name will be passed to word cloud function as an input parameter. In word cloud function, we read the input hotel name and search its reviews in our formatted word cloud json file accordingly. By using jason davies word cloud library, we can construct word cloud SVG element, and draw the word cloud by entering the existing words or exiting the existing word cloud and updating it for a new input hotel by re-computing the review words. To distinguish the difference between word cloud for positive words and for negative words, we added the title for each different word cloud and assign different type of color sets to them.

For the pie chart of different room type, we used D3 with multiple paths. There are two different room types in our dataset. When the user selected the hotel by clicking, the chart will be initialized. We obtain the data grouped by hotel and extract the review count for each room type. Then it splits the pie into two parts based on the count, and uses two different color for display. Then the legend shows the count for each type as well as the percentage based on the total reviews. And it will remove and update when a new hotel is selected.

The line chart is based on time and the review. After clicking on the hotel, the hotel name will be passed. Then we parsed the time format and grouped by month. Then the updateTimeChart function was called. Then it finds the minimum and maximums for the count and the time and then display the axises based on the extremes, the append lines based on the average scores. Finally add the circles with color for each entry along with the tooltips. The legend shows how the darkness of blue represents how many reviews there are.

Our bar graph implementation used D3 similar to the pie chart to display the information. An xScale and yScale were added like in the lab examples to make the bar graph more adaptive in its bar width and height. After extracting the necessary information from the given hotel data, the visualization was very similar to the labs. Afterwards, the same technique was used to show the dots on the graph instead changing the value of the y-axis to be average review score. The object's shape was changed to dots but the positional scaling was kept the same which made showing both the bars and lines at the same time simple.

Evaluation

1. Process

The initial step in our evaluation process was creating the google form that participants would use to evaluate our project. The form consisted of four sections. The first section asked participants to rate different aspects of the page overall. Section 2 asked about how participants liked individual components of the visualization. Section 3 asked about which bar chart of a selection of bar charts was preferred to each participant. Lastly, Section 4 asked participants for any additional comments.

We then created a link directly to the html page of our project visualizations that we would use to send to participants. We then asked participants to use our web page to try to inform them on which hotels they found preferable. We told them that when they when they were finished using the page, they were to fill out the evaluation form. Google forms has a straightforward way to view overall responses using histograms and pie charts as well as individual responses, which were used to look at the feedback we were given.

2. Participants

The final count of participants for our evaluation is 18 undergraduate and graduate students from UMASS. Almost all of the participants were computer science majors, therefore were familiar with html. All participants had at least general computer science knowledge.

3. Findings

Based on the questionnaire we sent out, the users have more positive feedback when using the visualization system we built up compared to negative feedback. The reasons include but are not limited to how easy it is to learn about different influences on hotel reviews in Europe, the ease of using the user interactive map, how easily the information for each component of the graph is represented, etc.

There is also some negative feedback we got from the users. Some of the major points that most of the users are not satisfied with are how the loading speed is slow and some color choices do not have a strong contrast which makes it hard to read. Also, the last graph, which is the bar graph, is not working to properly visualize the information as expected. We did some improvement on the colors we used to make everything easier to read and we picked the best graph from all the designs we had by

evaluating the users' feedback. However, the loading speed is very difficult to improve. Since the initial dataset we had was massive, we still got more than 80,000 records after data cleaning.

Based on the scores from 1 to 4 for each individual visualization showing how much participant's liked each component, we found that the word cloud was the least liked visualization followed by the bar chart. The map visualization was in the middle of the pack in terms of appeal while the pie chart and line graph were the most appealing. Our responses unfortunately generally did not include details as to why the participants liked or disliked certain components.

For one of the questions we asked the participants which of four bar graphs they though conveyed information the best. The purpose of this question was to help us determine the best course of action to take for our bar graph, whether it was best to make the bars in the bar graph the total count of ratings for each category or the average rating for each category. Additionally, we were curious if changing the color of the dots to the color of the associated bars was preferable to users.

The most popular selection was determined by the participants to be graph A, which told us that the most popular visualization was the one where the bars were the count of the number of ratings for each category and the dots were the average rating, kept as black. However, upon reviewing the comments by the participants as to why they chose the answers they did, we found that we did not control for all the necessary variables as graph C was the only one with x and y axis labels which influenced the selection of some of our participants. Despite this, graph A was still more popular so we were able to get some important information out of this section of the form despite this flaw.