# YELP REVIEWS ANALYSIS

TEAM NAME: Five Stars
TEAM MEMBERS: Jianan Liang, Tryphena Hu, Aileen Wu, Kaiwen Li, Adrianna Hohil

**DATA SCIENCE QUESTION(S) & HYPOTHESIS**:

Customers mention various positive features in their reviews on Yelp (such as “fast”, “delicious”, “friendly”, “healthy,”, etc). What is the effect of “mentioning a certain feature in review” on a customer’s star-rating?

Since it’s restaurant industry, we hypothesize that the feature of “delicious” has the biggest positive impact on customers’ star-ratings. However, is it possible that some other factors have larger impacts?

Note: We will figure out the most frequent positive features to look at after some processing of our data.

**BACKGROUND**:

With technology being readily available through smartphones and the internet, people often refer to online forums to discuss quality of service in the food industry. And today, one of the most popular forums that customers use to review restaurants is Yelp. Yelp allows users to discuss their overall experience at a venue by rating the venue and writing about it. Users can rate on a scale of five stars (one being the worst, five being the best) and on a scale of four dollar signs (one being the cheapest, four being the most expensive), while also writing out their experience in a textbox to include service, ambience, wait time, and preferred method of payment. And although some users may take advantage of the platform by submitting overly negative or positive reviews about a restaurant, Yelp regulates this to ensure it’s a reliable source, and 85% of the restaurants on Yelp have ratings that are three stars or higher. (https://computer.howstuffworks.com/internet/social-networking/networks/yelp2.htm)

Considering all of this information, our project group realized that it was possible to use Yelp to help better restaurant service and advertisement. When using Yelp, customers can use filters or the search bar to find a restaurant by name, or a particular type of restaurant they desire. However, within different types of restaurants (Mediterranean, Italian, Chinese, Mexican, etc.), customers have different preferences for what makes that type of restaurant desirable. For example, at an Italian restaurant one may want free bread and olive oil at the table before the meal versus at a Mexican restaurant, one may want a live music performance. To identify what is considered a “good” type of restaurant in a specific category, our group will be designing a linear regression model that analyzes a number of reviews in a specific category and see what words are used in “good” and “bad” reviews. Each word will be assigned a value. We will be analyzing a number of Italian and Chinese restaurants, and when analyzing reviews for a restaurant in a category, the sum of the number of times both “good” and “bad” words appear will assign that restaurant an overall score. And within each category, we’ll see which Italian restaurants were considered the best, and which Chinese restaurants were considered the best according to their scores and why (depending on the aspects

that the customers favored). From there, companies can view their results and compare it to their competitors to see how they can better their overall dining experience and attract more customers.

Previous groups in Cogs 108 have used Yelp, as well as Linear Regression Models for their final projects.

Group 006 from Spring 2017 investigated whether the quality of facilities in Chicago correlated with the crime rate present in the area of those facilities. The group decided that the best way to evaluate this was by looking at restaurants in Chicago because the city is large enough to have a fine-grained data set. The group used Yelp's open API to collect their data because there were no JSON files available at the time. However, that has since changed and we will be using Yelp's JSON files. The group mainly used the latitude and longitude data to identify the general location area of a restaurant and identify trends between areas of the city. They also hypothesized that people of varying demographic groups reviewed restaurants on Yelp, and that the number of reviews correlated to foot traffic. After analyzing Yelp data, restaurant health data, crime data, and census data, the group did not find a conclusive correlation between restaurant health and crime. This was due to the fact that restaurant health did not prove reliable - restaurants either passed or failed, and the factors that go into that are ambiguous - and that their scatter plots showed no correlation between the two. Also, when doing linear regressions, their R^2 values were 0 or close to 0, and their P > |t| values were large, so the correlations did not prove meaningful.

Group 033 from Spring 2017 investigated why research accidents occur in San Diego, as it proves to be a danger to infrastructure and individuals in society. The group sought to do data analysis on the conditions surrounding a Motor Vehicle Accident (MVA) in San Diego to help the community take action in improving road conditions and prevent MVAs from happening. The group hypothesized that accidents were more likely to occur on roads that had poor or less favorable conditions than those that did not. They defined these conditions using an "Overall Condition Index" or OCI, in which there were three categories of road condition quality. A "good" street condition had "little to no cracking and minor potholes, had excellent drivability, and needed little maintenance or remedial repair." These roads were assigned OCI values between 70 and 100. A "fair" street condition had "moderate cracking, minor potholes, adequate driveability, and is in need of remedial repairs and a slurry seal." These roads were assigned OCI values between 40 and 69. A "poor" street condition had "severe cracking, numerous areas of failed pavement with possible sub base failure, and exhibits a rough ride. It qualifies for a comprehensive repair or total reconstruction if conditions continue." These roads were assigned OCI values between 0 and 39. If a strong correlation was found between poor street condition and MVAs, meaning the group's hypothesis was correct, they would investigate the incidents further since MVAs are due to numerous factors that should be taken into account (weather, poor lighting, etc.). The group analyzed OCI data, created graphs, and did a t-value test to investigate data regarding road conditions and created a linear regression model and a chi-test to investigate the severity of MVAs in relation to lighting, weather, and reaction time. Overall, the group found that despite their confidence in their hypothesis, there was little to no correlation between OCIs and MVAs because the p-value resulting from their t-test was so low. Also, their chi-square statistical analysis showed their was most likely very little correlation between OCI, weather, lighting, and collision time. Using the chi-squared values, they created a linear regression model that compared MVAs, weather, lighting, and collision time, which yielded p-values 0.069, 0.000, and 0.031, respectively. And based on those results, the MVAs correlated to collision time was the only category that appeared promising for further investigation.

**ETHICAL CONSIDERATIONS**:

The source from which we acquired data for this project is the Yelp Open Dataset which according to the "subset of our businesses, reviews, and user data for user in personal, education, and academic purposes". This dataset draws from data from Yelp, a public, online platform where users can search for, find, and review businesses. On one hand, the data is publicly available and Yelp gives permission to use this data for academic purposes. However because the data is also user crowd sourced and user generated, it is possible that Yelp has questionable rights to the data. While legally, Yelp requires users to agree to a certain set of terms and conditions before using the site, it is likely that its users can be potential unclear or misinformed about how their data can be used. Within their privacy policy they are also transparent about their ability to share public information and user information in the aggregate with third parties. However, this could potentially be a problem if users are not aware of agreeing to the privacy policy.

Yelp operates by allowing customers to write reviews for particular businesses, where they are able to write text summarizing their experience and also provide a rating (one to five stars). These reviews create a community where potential customers can filter their searches and read reviews before making a decision. In this way, Yelp has created a reputation for itself for providing quality, reliable information and also is known for its commitment to being an open and public source of information. Because this trust they have built with the public is important to their success, one ethical concern we might have for our project is how identifying certain patterns with words or phrases might lead to filtering of certain reviews. For example right now, yelp filters through reviews they believe might be false or deceptive and highlights information that they believe is useful which they then label as "recommended" or "not recommended".

Because yelp also has a tremendous amount of influence in the world of businesses by directly impacting customer decisions and business revenue, we also need to ensure that our project does not allow businesses to take advantage of certain patterns that might be revealed through our analysis of the data. For example recognizing certain patterns as to how customers might respond, could open the door for businesses to capitalize or manipulate certain sentiments in a way that would allow them to cut corners or provide a lesser quality experience.

Considering Yelp's commitment and reputation also leads us to some of its practices could pose potential ethical concerns. While the data we are taking is public information, Yelp filters reviews based on a person's profile. They use an algorithm to determine which features have the most influence such as textual classifiers, sentiment analysis and people who have more reviews, a profile photo, larger number of friends are more likely to be recommended. IN order to highlight useful information, Yelp has to filter out what appears deceptive and suspect. However because the issue of free speech on Yelp's platform, non-recommended reviews are still accessible to users. Yelp also has a unique responsibility to maintain the integrity of the service in a way that rejects bribery of any kind. However it would be necessary to disclose the use of filtering or algorithms in order to prevent conflicts of interest to arise.

**DATA**:

We will be downloading dataset from https://www.yelp.com/dataset. The downloaded data have 6 JSON file (business.json, review.json, user.json, checkin.json, tip.json, photo.json), but we will be only using 2 of them (business.json and review.json) for this project.

From the business.json file, we will extract 3 variables including "business_id" (identify which restaurant it is), "review count" (number of reviews left), and "categories" (the type of food it serves). After that, to form a complete set of data, we will match the "business_id" in the business.json file with the "text" (review itself) and "stars" (stars rated) sections in the review.json file. Hence, our data set will include the following variables {"business_id", "stars", "review_count", "categories", "text"}

The number of observations in business.json is 192609
The number of observations in review.json is 1048576

## Team Expectations Agreement

Read over the COGS108 Team Policies individually. Then, include your group's expectations of one another for successful completion of your COGS108 project below. Discuss and agree on what all of your expectations are. Discuss how your team will communicate throughout the quarter and consider how you will communicate respectfully should conflicts arise. By including each member's name above and by adding their name to the Gradescope submission, you are indicating that you have read the COGS108 Team Policies, accept your team's expectations below, and have every intention to fulfill them.

These expectations are for your team's use and benefit—they won't be graded for their details. Goals should be realistic: "No group member will never miss a meeting and everyone will always show up early" is probably unrealistic, but "Group members will attend almost every meeting and will communicate their absence at least a day in advance of the group meeting" and "When group members are unable to attend a meeting, they will submit their notes and progress ahead of the group meeting" are realistic expectations. Expectations for deadlines, how you'll work together, meeting attendance and participation, and project completion should all be considered and details included below.

**INCLUDE YOUR TEAM'S EXPECTATIONS HERE**

- Split up project & be responsible for your part but aid others
- Don't be afraid to ask for help
    - In case your section is more work than expected and you need help
    - If you're stuck
- Work together for final product
- Meet bi-weekly to update on project
- Communicate with team members through messenger in regards to schedule availability, conflicts, needs, etc.
- Meet with TAs, IAs, or Prof Ellis on a need basis

# Project Timeline Proposal

Include actual dates and times for due dates and meetings below, not just what week they'll be completed

| | Draft Text? | Write Code? | Proposed due date (Friday) | Discuss at team meeting (Friday) | Edit? |
|---|---|---|---|---|---|
| **Initial team meeting** | NA | NA | NA | week 2 | NA |
| **Background Research** | Adrianna | NA | week 3 | week 4 | Adrianna |
| **Question & Hypothesis** | Jianan | NA | week 3 | week 4 | Jianan |
| **Ethical Considerations** | Tryphena | NA | Week 3 | Week 4 | Tryphena |
| **Dataset** | Aileen, Kaiwen | NA | week 3 | week 4 | Aileen, Kaiwen |
| **Data Wrangling** | Aileen, Jianan | Aileen, Jianan | week 3 | week 4 | Aileen, Jianan |
| **Descriptive** | Tryphena, Kaiwen | Tryphena, Kaiwen | week 5 | week 6 | Tryphena, Kaiwen |
| **Exploratory** | Adrianna | Adrianna | week 5 | week 6 | Adrianna |
| **Analysis - Part I** | Adrianna, Tryphena | Adrianna, Tryphena | week 6 | week 7 | Adrianna, Tryphena |
| **Analysis - Part II** | Jianan, Aileen | Jianan, Aileen | week 6 | week 7 | Jianan, Aileen |
| **Analysis - Part III** | Tryphena, Kaiwen | Tryphena, Kaiwen | week 6 | week 7 | Tryphena, Kaiwen |
| **Summarize Results** | Everyone | NA | week 7 | week 8 | Everyone |
| **Conclusions** | Everyone | NA | week 7 | week 8 | Everyone |

**Once completed, save this document as a PDF & submit on Gradescope. Be sure to add each team member's name to the Gradescope submission.**