# ASSIGNMENT 2 (INDIVIDUAL)

The Subway Reviews Problem

## Instructions

This graded assignment is based on reviews shared by customers on the Yelp platform. The data is released by Yelp for academic use and can be found at [Yelp Open Dataset](https://www.yelp.com/dataset). A subset   
of data is being shared with you for academic use. Please read the [Yelp Dataset Terms of Use](https://s3-media0.fl.yelpcdn.com/assets/srv0/engineering_pages/dc1cabe7cb95/assets/vendor/Dataset_User_Agreement.pdf) before proceeding with the assignment.

The assignment uses names of real companies and individuals as they appear in the dataset. The context laid out, however, is purely fictitious.

Except as indicated, use this document to record all your assignment work and responses to any questions. At a minimum, you will need to turn in a digital copy of this document to your instructor as part of your assignment completion. You may also have additional supporting documents that you will need to submit. Your instructor will provide feedback to help you work through your findings.

**Note:** Though your work will only be seen by those grading the course and will not be used   
or shared outside the course, you should take care to obscure any information you feel might be of a sensitive or confidential nature.

*Complete each assignment part below. Wait to submit the assignment until all parts are complete. Directions to submit your assignment can be found on the assignment page in   
the unit. Information about the grading rubric is available on the assignment page online.   
Do not hesitate to contact your instructor if you have any questions about the assignment.*

# Part One

Problem Setup

Subway CEO John Chidsey received a report from his field officer that customers are unhappy with one of its stores in Milford, CT, where Subway is headquartered. Upon checking the online reviews for this store, he finds that the average rating received by the store is 3.2/5. John is concerned that if a store next to the headquarters is at 3.2/5, stores farther away might be performing even more poorly.

John calls for an urgent meeting of the Head of Customer Service, the Head of Store Operations, the Head of Social Media, and the Chief Data Scientist (your boss). He expresses his concern and urges the team to take measures to improve the average ratings received across all stores in the U.S. to 4.5/5.

His team tells him not to worry by making the following statements:

* Head of Customer Service: “Our ratings are gradually improving, and we will soon reach 4.5/5.”
* Head of Store Operations: “Sandwiches are a tricky business. All sandwich chains suffer from poor customer ratings.”
* Head of Social Media: “The goal of 4.5/5 is unreasonable for national chains like us. Only small, local, and boutique restaurants can achieve such high ratings.”
* Chief Data Scientist: “It is well known that customers make the effort to give a rating only when they are either extremely angry or absolutely delighted with the service. So online ratings are not reliable.”

Your boss, the Chief Data Scientist, gives you two datasets

1. Review Dataset: This data contains details of ratings given by users to restaurants on an online platform. It contains ~400K ratings given to ~2,000 different restaurants along with the date of posting the rating.
2. Restaurant Dataset: This data contains details of restaurants such as name, category, city, and state.

# Part Two

Problem Formulation and Analysis

Answer the following questions.

1. **Part A:** Does the data support the statement made by the Head of Customer Service?

**Part B:** Is this statement valid across all states?

Hint: Construct a plot with year on the x-axis and average rating received on the primary y-axis. Plot a bar chart of number of ratings received on the secondary y- axis.

|  |
| --- |
| **PART A:**  No. The data does not support his statement saying “Our ratings are gradually improving, and we will soon reach 4.5/5”.  Consider this graph:  Not only can we clearly see the trend is clearly negative, but also, we can see that the average rating is nowhere near 4.5 stars regardless.   One might ask “what about that uptick at the end?”  Although this might look like an upward trend, we also must consider that the dataset for this year is incomplete as we can see from the blue bar chart. I predict that with complete data the negative trend would continue.  **PART B:**      The statement made by the head of customer service is wrong across many states. As we can see from the graphs, all the states plotted here have a negative outlook. AZ appears to have a positive upturn in 2022, however the last datapoint is invalid because there are too few reviews in that period thus resulting in an outlier.  Regardless, none of the average ratings for any state I plotted are even close to 4.5 stars, so the statement is false according to the data. |

1. Does the data support the statement made by the Head of Store Operations?

**Part A:** Are sandwiches the only tricky business?

Hint: Identify one or two national competitors of Subway. Create a plot to compare   
the mean and standard deviation of reviews received by Subway with its competitors.

|  |
| --- |
| The data does not directly refute the statement made by the head of store operations that “sandwiches are a tricky business” because this is a subjective statement.   However, if we define “poor” in the same way the CEO did in the original problem (~3/5 or lower) then the statement that all sandwich chains suffer from poor ratings appears to be supported by the data. Consider this competitive analysis:  Here we can see that only one competitor (Jersey Mike’s Subs) averages just barely above a 3-star rating. I think what is more important, however, is that we can clearly see Subway is being outperformed by 2 direct competitors in the sandwich business, which I would argue is more a cause for concern. Therefore, the statement “all sandwich chains suffer from poor ratings” is true if we define poor ratings as < ~ 3/5 stars, however Subway’s particularly low ratings in comparison to competitors (according to the chart) is much more important to focus on.  **PART A:** Sandwiches are not the only “tricky” business. Consider the following chart:  Here we compare a sandwich restaurant (Subway) to a pizza chain (Papa John’s) and a coffee shop/breakfast chain (Dunkin’). As we can see, both of these businesses score lower than Subway in terms of average star rating. This allows us to infer that sandwiches may not be the only “tricky” business. More comprehensive analysis would be needed to determine the overall most “tricky” industry. |

1. **Part A:** Does the data support the statement made by the Head of Social Media?

**Part B:** Is it true that average rating decreases as the size of restaurant increases?

Hint: Categorize restaurants with a presence in more than 50 cities as national chains and a presence in only one city as local. Construct a plot to compare ratings received by national chains and local chains.

|  |
| --- |
| **PART A:**  The data does support the head of social media’s statement to an extent as we can see from the chart below: Here we can see the local restaurants (located in 49 cities or less) plotted in purple, where the national restaurants (located in 50 cities or more) are plotted in blue. We can clearly see that local restaurants have higher ratings than national chains, so the head of social media’s stance is partially supported.  However, we can also clearly see that the average rating of local restaurants is below the target value of 4.5 stars hovering at around 4.0 stars. Therefore, the head of social media’s statement is partially rejected.  **PART B:**  No, it is **not** true that average rating decreases as the *size* of a restaurant increases.   I want to start with the premise that the larger a restaurant is, then the more customers they will have.   If we agree on this premise, then it then follows logically that the more customers a restaurant has, the more reviews they will receive.   Therefore, If the “size of the restaurant” is defined by how many reviews they receive, then there is actually a positive correlation between review score and restaurant size, as denoted by the scatterplot below.  If the question had instead asked “is it true that average rating decreases as the number of cities a restaurant has locations in increases?” the answer would probably be different.    **Correlation between number of reviews and average rating: 0.2572645296011175** |

1. **Part A:** Does the data support the statement made by the Chief Data Scientist?

**Part B:** Is the statement true across all years from 2018 to 2021?

Hint:Create a plot with the ratings on the x-axis and the number of reviews with that rating on the y-axis.

|  |
| --- |
| **PART A:** “It is well known that customers make the effort to give a rating only when they are either extremely angry or absolutely delighted with the service. So online ratings are not reliable.”  Well, this statement is barely falsifiable given how it is currently phrased, but I will do my best.   So, do customers *only* give ratings when they are extremely angry or delighted? Well no, there are very clearly people who rate things 2, 3, or 4 stars as evidenced by the data.   However, do the *majority* of reviewers gravitate towards 1 or 5 star ratings? Yes, this is supported by the data and can be seen in the chart below:    Does this mean online ratings are not reliable?   Well, we should first define “reliable” in this context. If he is saying that online reviews are not a reliable indicator of food quality, service quality, building quality, location quality, etc. I don’t think we can adequately prove or disprove this premise with the data we were given for this assignment.   If he instead defines “unreliable” merely as having skewed data, then yes, I think there is a high probability that data is skewed by selection bias when considering online reviews from Yelp as the source.  **PART B:**  Yes, according to the chart below it appears that 1 and 5 star ratings vastly outweigh the other ratings from the years 2018-2021 |

1. Based on the insights generated from Questions 1 through 4, what would be your recommendation to John and the team (at most 100 words)? Your response should include, but may not be limited to the following:

* Is there cause for John to worry? Justify your answer.
* Is the target of 4.5/5 reasonable? If “yes,” justify your answer. If “no,” what do you think would be reasonable and why?
* Support your recommendations by quoting appropriate numbers computed from the data.

|  |
| --- |
| There is cause for John to worry, but it is not because of Subway’s average overall Yelp scores. Rather, Subway is not outcompeting direct competitors in the sandwich industry. A target yelp score of 4.5/5 is definitely unreasonable, considering that even local restaurants average around 4.0 stars as evidenced by the prior “average star ratings of national vs. local restaurants over the years” graph. Instead, I think a more reasonable goal would be to raise the average star rating to above 3.357. If they are able to get an average review score of above 3.357 they would become the highest rated nation-wide sandwich chain and push Jersey Mike’s down to second (they average 3.356 stars). As long as Subway can be the highest rated in their niche, they should be able to outcompete the field and maintain growth and profitability. |

1. Please upload an additional single .pdf file along with this document with three sections:

* Your Python code appropriately commented
* Additional instructions for running your code, if necessary
* Snapshots of the output of your implementation

1. **Bonus question:** What additional insight on ratings can you provide John?

* Please write a brief comment (at most 50 words).
* Support your argument with up to three data visualizations.

|  |
| --- |
| As we can see from the graphs above, Subway is not the only national sandwich vendor that has seen steadily decreasing ratings over time. Because Subway is not the only chain with a negative trendline, we can infer that the problem more likely comes from yelp itself or may be an inherent problem to the sandwich industry rather than Subway providing uniquely poor food and services. |

*To submit this assignment, please refer to the instructions in the course*.