# 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.

|  |
| --- |
| According to the cleaned data and the generated chart, the line graph shows that the average score of all Subway stores showed an upward trend and peaked in 2006. However, starting in 2008, there was a significant decline in the average score and it continued to decline for the following ~13 years, with only a weak recovery beginning in 2021, but still far from the 4.5/5 target.  From the bar chart, it can be observed that the number of comments fluctuates significantly in different periods. The very small number of early (e.g., 2006 and before) and recent (2022) reviews may cause the average ratings in these years to be inflated due to insufficient samples. However, 2019 has the largest number of reviews, so the average rating in this year is relatively more representative and reliable.  Taken together, these data do not effectively support the view of the head of customer service, as Subway's average score did not improve significantly over most of the time. Moreover, since the data may exhibit different trends across states (such as differences in the distribution of the number of reviews), this conclusion cannot be applied to all states. |

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.

|  |
| --- |
| To test the point of view of the head of store operations, I selected the scoring data of Subway and its two main competitors for comparison. As can be seen from the figure, the average scores of these three sandwich chain fast food restaurants are concentrated in the range of 2.0-2.5.  This suggests that the sandwich industry does have an overall low rating. Thus, the data do kind of support the view of the head of store operations. On the other hand, ratings are more likely to be influenced by other factors such as service quality, restaurant size, or regional preferences. |

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.

|  |
| --- |
| The overall average rating for national chains is about 2.5, while the overall average rating for local restaurants is close to 4.0. The data indicate that local restaurants are rated significantly higher than chain restaurants. This may be related to the fact that local restaurants can provide more personalized service, so I think it can support the Head of Social Media's view that "only local restaurants can achieve higher scores". As for the point of view of part B, the data show that only one restaurant has chain stores in cities between 1 and 50, which is not enough to support the argument. |

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.

|  |
| --- |
| It is evident from the first and second graphs that the distribution of ratings shows a clear trend towards polarization. In each year between 2018 and 2021, the number of reviews rated 1 and 5 stars was significantly higher than the number of reviews rated 2 to 4 stars. This distribution trend indicates that customers are more inclined to leave reviews when they are extremely dissatisfied or very satisfied. In particular, the high proportion of 5-star ratings further supports the chief data scientist's view that ratings tend to come from customers with extreme emotions. |

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.

|  |
| --- |
| These data suggest that Subway's score peaked in 2006 but declined significantly after 2008 before showing a weak recovery through 2021, which is cause for concern. Reaching 4.5/5 is unrealistic for national chains, as their average rating is about 2.5, compared to the average rating of 4.0 for local restaurants. Ratings are heavily polarized, and I recommend focusing on realistic goals such as improving to 3.5/5 within two years, addressing service quality issues, and targeting underperforming areas for incremental improvements in customer satisfaction and ratings. |

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.

|  |
| --- |
| An analysis of Subway's average rating across states reveals that no state has a rating higher than 4.0. Among them, Arizona has an average rating of 3.0, while the rest of the states have average ratings clustered between 1.0 and 2.74, with generally low performance. This suggests that there is room for improvement in Subway's overall service and customer satisfaction across states.  Based on this result, my recommendation is to focus on low-scoring states as a priority and analyze in depth the main causes of low scores, such as problems with service quality, food taste, or store management. Improving customer satisfaction through targeted improvement measures can gradually improve the overall score. |

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