Process Tracking and Next Steps

Everything you have done to date

- Revised Hypothesis, Introduction, Related Works, and Data Preprocessing section based on the feedback received from Professor and the TA. (By Hanna and Yi)
- Downloaded hygiene data from NYC Open Data and computed data cleansing by removing missing and duplicates, etc. Extracted 5000 samples. (By Chenyue and Yi)
- Create visualization for the distribution of Hygiene grade, borough, and cuisine description from hygiene data. (By Chenyue, Hanna, Yi)
- Based on the updated hygiene dataset (5000), extracted restaurant information using Yelp Fusion. (By Hanna and Chenyue)

Steps from this point to completion

- Extract reviews for each restaurant using Yelp Fusion. Line up the dates of the reviews with the dates of the inspection. (By Hanna and Yi)
- To minimize the influence of confounding factors, when analyzing hygiene scores and ratings, create a filtering algorithm that could select comments that specifically talk about hygiene. (By Chenyue, Hanna, Yi)
- Run statistical analysis for the dataset to test our hypotheses of hygiene scores and other factors (price, cushion type, location) on ratings. Potential tests would be regression analysis and ANOVA. (By Chenyue and Yi)
- [IF TIME PERMITS] Create a machine learning model to predict customer rating based on hygiene score. (By Chenyue, Hanna, Yi)
- Create visualization and analyze the result. (By Hanna and Chenyue)
- Revise the paper and prepare for the presentation. (By Chenyue, Hanna, Yi)

Link to dataset

https://github.com/serenawcy/restaurant_ratings