Tasks

1. Predict the review score attributes (accuracy, cleanliness, check-in, communication, location, value, and overall score) based on natural language processing of the reviews.

2. Analyze factors that contribute to higher or lower ratings (e.g. price, cleanliness, location, etc).

Both? Daniel seemed to indicate NLP would be complex enough.

1. Figure out what data needs to be pulled from which tables
2. Analyze reviews with NLP; predict different scores – put in bins per category?
3. Check if predictions match with actual review score
4. Visualizations
5. Literature Review
6. Create Binder/Finalize Notebook

Final Project Details and Rubric

In Tools 1, we are concerned with data cleaning and exploratory analysis. Please select a project that has enough scope for the following activities.

For the final project, you will create a Github repository for your project and tag it with the label TOOL1\_FINAL\_PROJECT by the due date. The github repository must have a .ipynb notebook file with output and associated code. Having output in the notebook cell is very important if your dataset is big or we won’t be able to run the notebook in a reasonable time. Also, we should be able to run your project with a Binder link. The binder link should be in the README.md file. Please check https://mybinder.org/ to see how to create a Binder link. If this service is down, this step is not required.

Your final report should read like a data-driven story/scientific study (data science). This is really important if you want to publish your story as a blog on the web or share with stakeholders. Scientific publications have their own style and content requirements.

Use code cells and markdown cells to carry out your analysis. Please write the report using the following section format guidelines. You can create more sections if it is more natural to do so, depending on your project. Please write each section like a report and address the points mentioned in the following rubric. Try to make your report more enjoyable to read.

• Proper tagging of Github repository for final report as per deadlines (0.5 = 0.25 + 0.25 points)

• Dataset and motivation slide (1 points)

How/why the dataset was collected and a description of the metadata of your dataset.

• Actual task definition/research question (2 points)

What real-world problem are you trying to solve? What are the input and output of your

analysis?

• Literature review (2 points)

What other work has been done in this area, and how is your work novel compared to others?

• Quality of cleaning (6 points, 2 points each)

- Data cleaning and type conversion activity. Please share anything unusual you faced during this activity.

- What did you do about missing values and why? Handling missing values properly is very important.

* New feature/attribute creation and data summary statistics and interpretation.

• Visualization (8 points, 2 points each)

- Data visualization activity (box plot, bar plot, violin plot, and pairplot to see relationships and distribution, etc.).

- Describe anything you find in the data after each visualization.

- What data visualization helped you understand about data distribution.

- What you did about possible outlier as per data distribution visualization. (Did you confirm with your client whether it is actually an outlier or put a disclosure statement in your notebook if you decided to remove it?)