I have neither given or received, nor have I tolerated others' use of unauthorized aid.
Peyton Camden, Christopher Barua, Brandon Cook

Data Collection and Reduction Methodology:

The data set used in this project was gathered from the UC Irvine Machine Learning Repository, found at https://archive.ics.uci.edu/ml/datasets/Las+Vegas+Strip#. The data was extracted from Trip Advisor online reviews between January and August of 2015 by S.Moro, P.Rita, and J.Coelho, and amounted to 504 total reviews. This data set will hereafter be referred to as "the data" or "the dataset" for simplification purposes, as this is the only data being used in this analysis.

The first thing done in the data transformation was dropping variable columns thought to be insignificant by the researchers. For example, the 'Casino' column was dropped because every hotel in the study had a casino amenity attached to it as well. However, the researchers quickly realized that this introduced a large amount of bias into the study, and quickly moved to find a more quantitative way to eliminate potential predictor variables. It was decided that positive correlations with high hotel review scores would be a good way to determine predictors, and then to sort the data from each predictor by the traveler type category, possibly using colored kernel density plots in combination with pair plots.

Analysis:

After we had cleaned the data, next came properly analyzing it in order to come up with conclusions and results for our hypothesis. The first part of this involved turning every column that contained categorical data into numerical data. We achieved this by using a for loop to redefine and store the values from a specific column into a dictionary, and then map them to a new column. After completing this we went on to complete a 5 number summary using df.describe() and took a correlation, using the base correlation method df.corr(), of the cleaned data. In order to better visualize the correlation between specific columns, we also created a heatmap as seen below:

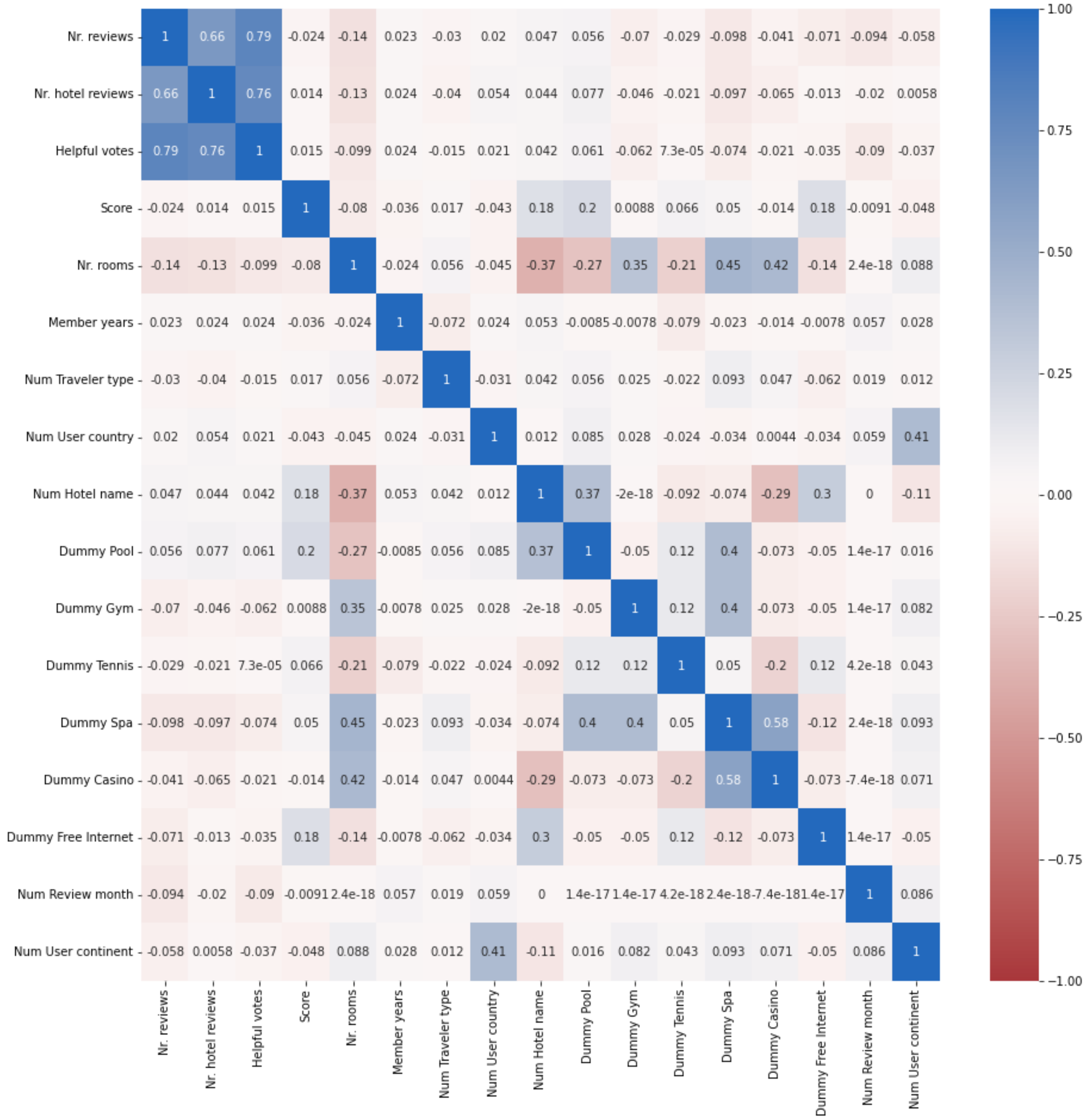

Fig: Heatmap of Numeric Columns

This heat map gave us a better understanding of the variables that are highly correlated, such as number of reviews, number of hotel reviews, and number of helpful votes. Taking these highly correlated variables into consideration, we then proceeded to create a pair plot that included these variables, as well as two other not as highly correlated variables, these being dummy spa and number of rooms. This pair plot can be seen below:

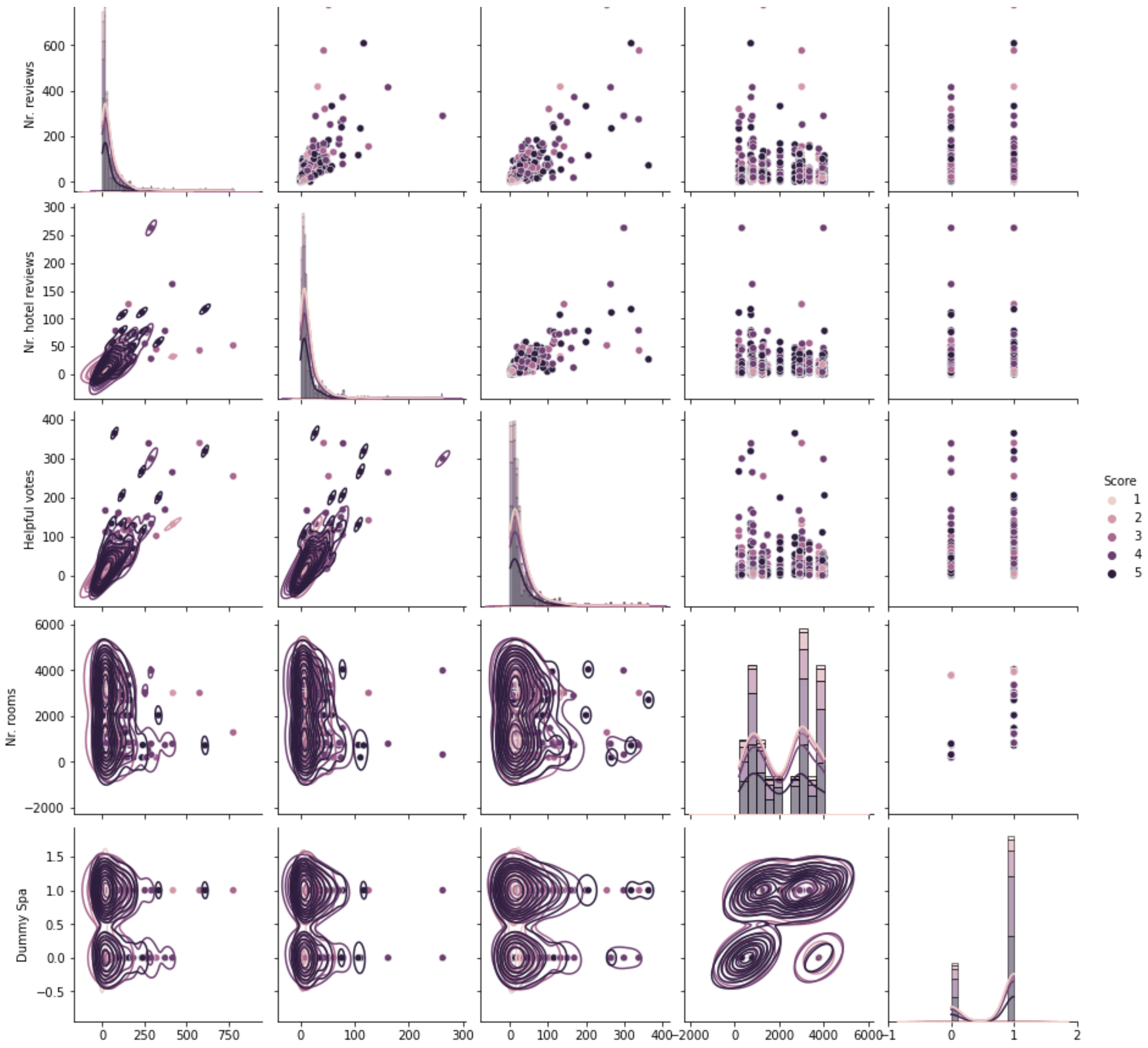

Fig: Pair Plot of Variables in Relation to Traveller Type

This pair plot specifically showed the highly correlated variables, and not as highly correlated variables, in relation to the traveller type. With the more highly correlated variables, the plots seem much more linear, as well the histogram curves forming much more of a normal curve than the other not as correlated variables. After creating this pair plot we then also created specific diagrams such as histograms for each individual variable included within the pair plot, as well as a box plot that compares the Traveller Type in relation to the score variable. These can be seen below:

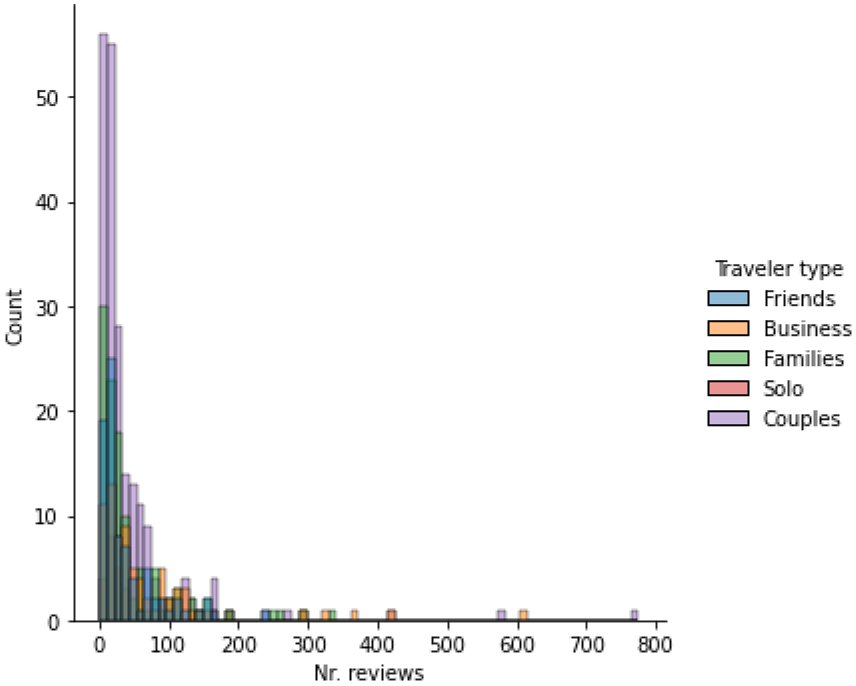

Fig: Histogram of Number of Reviews

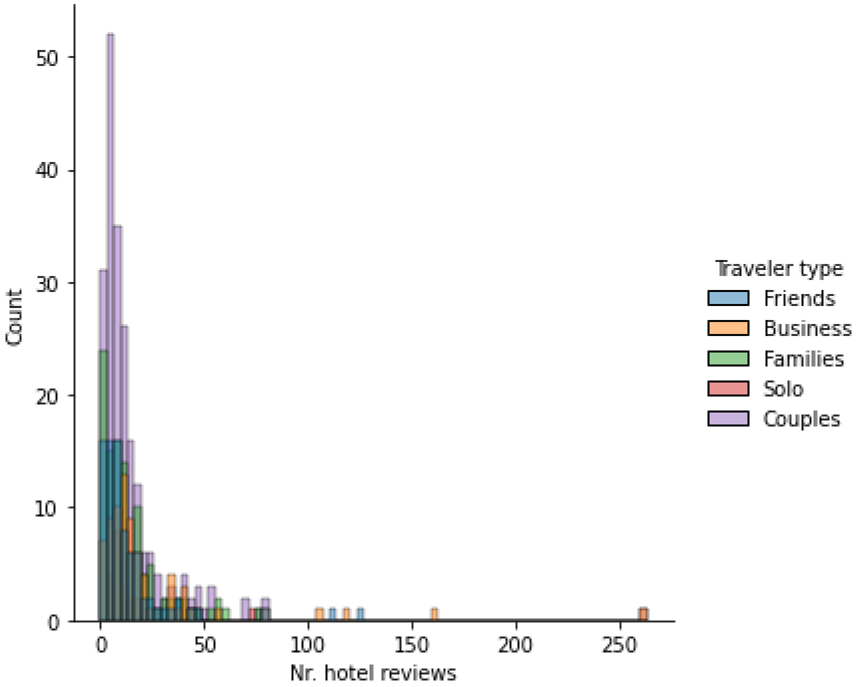

Fig: Histogram of Number of Hotel Reviews

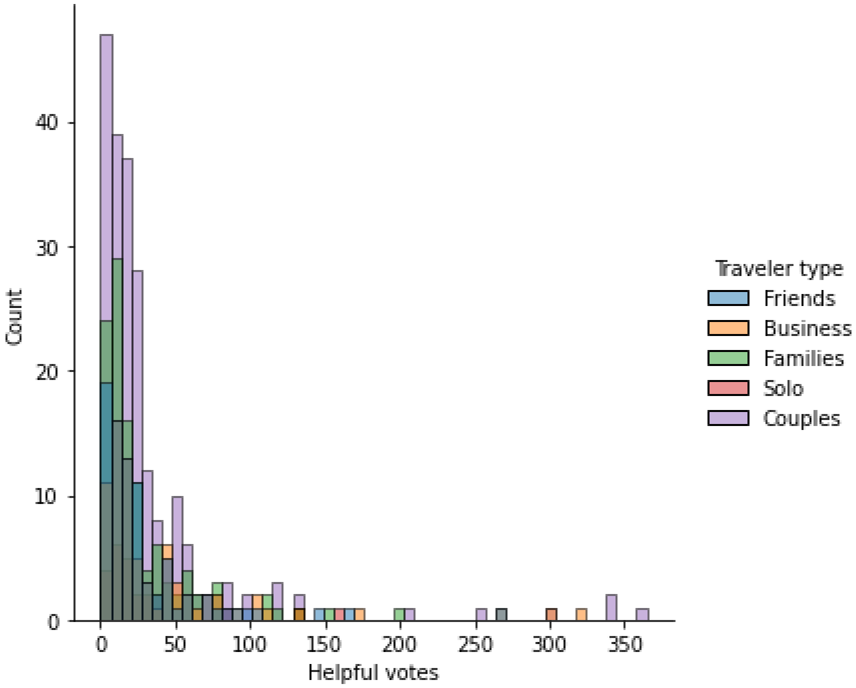

Fig: Histogram of Number of Helpful Votes

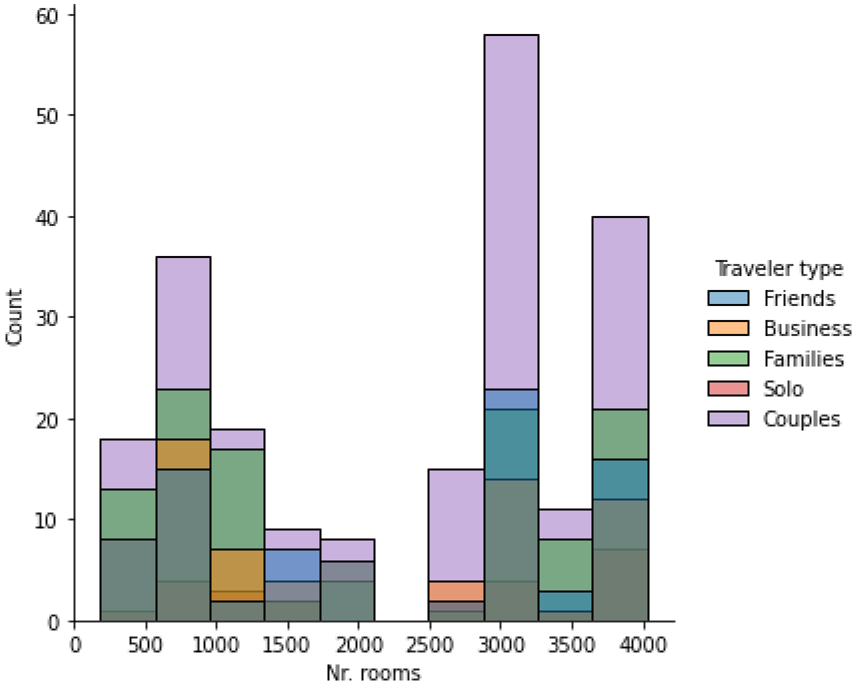

Fig: Histogram of Number of Rooms

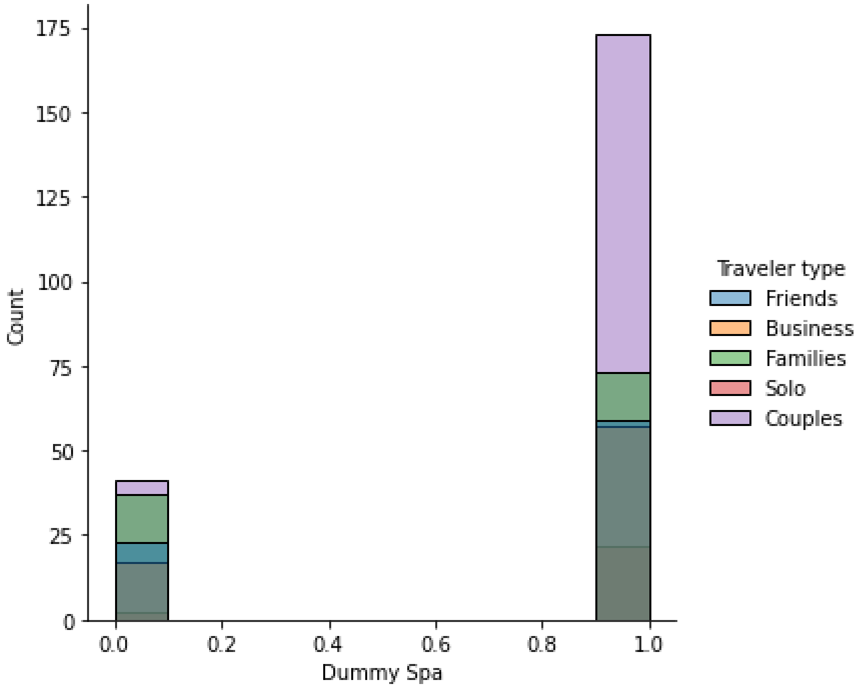

Fig: Histogram of if Hotels Included a Spa

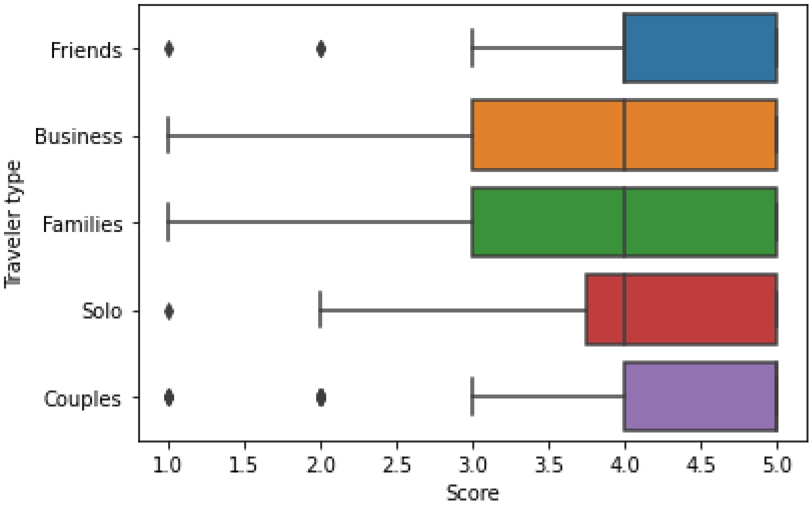

Fig: Box Plot of Traveler Type with the Scores They Gave

Unexpected Challenges Thus Far:

Until this point, the thing that has proven hardest is comparing the numerical variables to the boolean variables. While two numerical variables create a scatter plot that is easy to understand, a boolean value and a numerical variable do not make an easily discernible graph. With so much data in the data set, the researchers initially thought that eliminating columns with no noticeable connection to review scores or amenities would allow the data set to be reduced and easier to work with. However, it was quickly evident that without quantitative data to support the elimination of data columns, a large amount of bias would be introduced to the study, and the researchers decided to find a quantitative value to decide what predictor values would be valuable. Because much of the data is still being evaluated, there is much to learn about the relationships between the traveller type variable and each other variable. There is much more to explore about this data set than initially understood. Although the final goal is to create a prediction model, much of the efforts of this project are still focused on finding appropriate and defendable predictor variables and reducing the data appropriately for each score type. Subsetting the data appropriately by score and traveller type to determine accurate predictor variables is the next step and the biggest challenge in creating the prediction model.

Preliminary Results:

After looking at the heatmap that was created, it is clear that overall there are weak to dismissable correlations for our main predictand of score. The variables that stand out for potential predictors are hotel name (r = 0.18), whether the hotel contains a pool (r = 0.2), and if the hotel has free internet connection (r = 0.18). Potentially, backwards elimination could be used to try and create a model so all of the variables are run through to see if some are more significant than they seem to be based on correlation.

Schedule Continuing Forward:

- Part 3: Due 16 November 2021
  - Paper draft
    - Finished by November 9
    - Introduction and References: Peyton
    - Data and methods: Christopher
    - Results: Brandon (Christopher + Peyton as needed)
  - Demonstration
    - Practice date November 12
    - Walk/talk through of code/results
    - How the model works: Christopher
    - Preliminary results on “predictive power”: Brandon
    - Trade-offs and challenges: Peyton
- Part 4: Due 16 December 2021
  - Completed research paper
    - Done by Dec. 13
    - Discussion: Christopher + Brandoln
    - Conclusion: Peyton
  - Presentation
    - Work on leading into Dec. 16
    - Parts to be determined
  - Demonstration
    - Practice walk/talk through
    - Practice date to be determined

References:

Moro, S., Rita, P., & Coelho, J. (2017). Stripping customers' feedback on hotels through data mining: The case of Las Vegas Strip. Tourism Management Perspectives, 23, 41-52.