

Las Vegas Hotel Rating Layer 2: Collection, Cleaning, and Analysis

Peyton Camden, Christopher Barua, Brandon Cook

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.

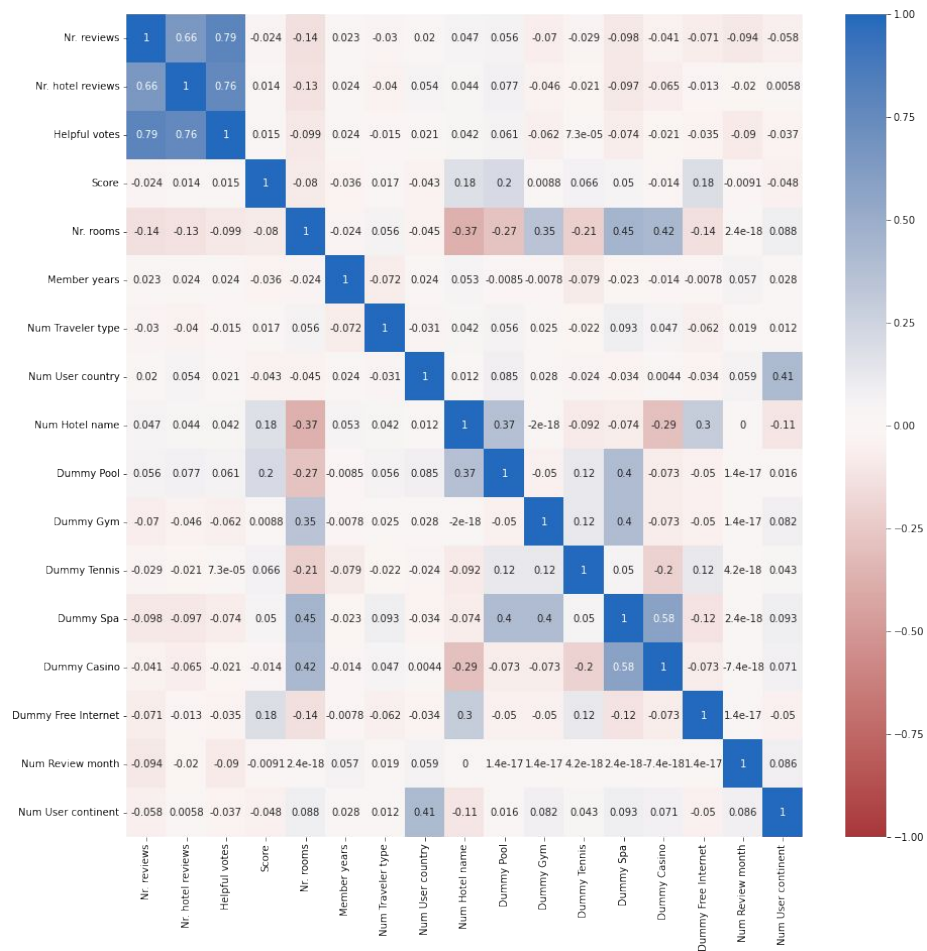
Wrangling/Cleaning

- Dropped columns
 - 'User continent'
 - 'User country'
 - 'Review month'
 - 'Review weekday'
 - 'Casino'
 - 'Free internet'
- Un-dropped columns and started over with fresh data set- misled at start
- Created dictionary for remaining categorical variables to make numerical
 - Used for loop to run through all unique variable names
- Turned all 'Yes'/ 'No' variables into dummy variables to be used as booleans
- Ready for analysis

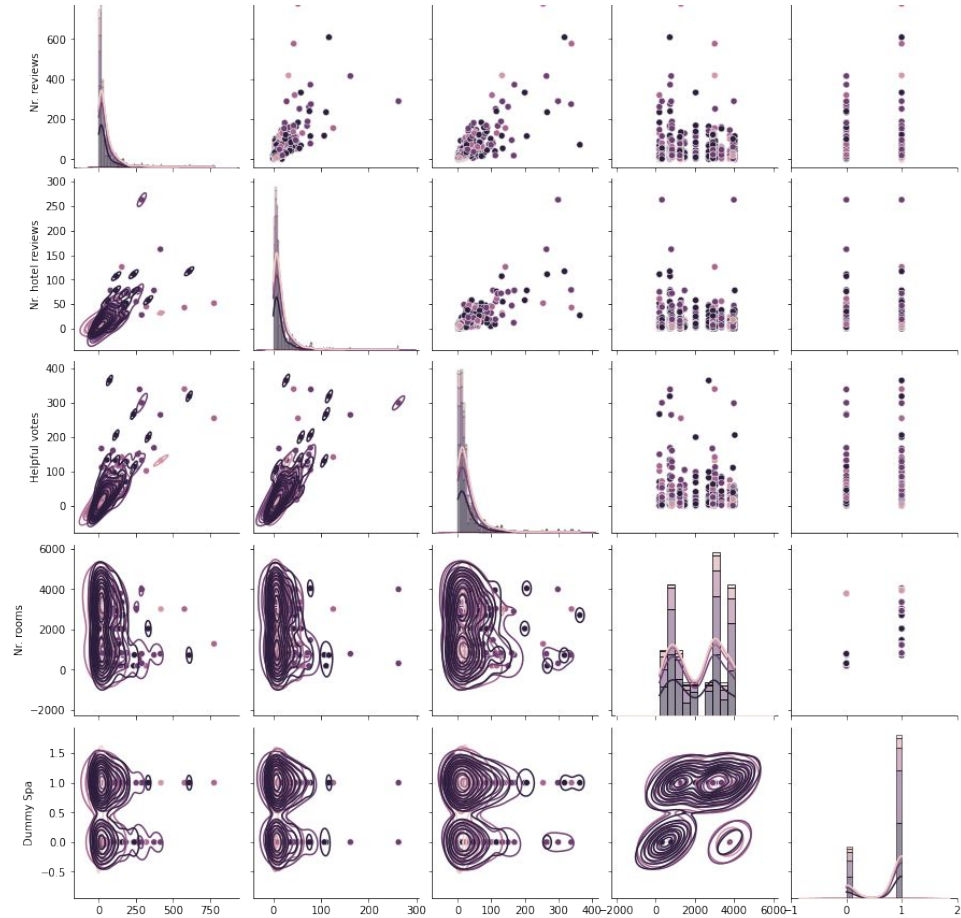
Analysis

- Did a Five Number Summary and Correlation
 - Used `df.describe()` and `df.corr()`
- Created a heatmap to better visualize correlation between variables
- Created a pair plot with the 5 highest correlated values and two other not as correlated variables
 - These variables included Nr. Reviews, Nr. Hotel Reviews, Helpful votes, Nr. rooms, Dummy Spa
- Created individual histograms for each variable included in pair plot
- Created a box-plot that showed distribution of scores for different traveller types

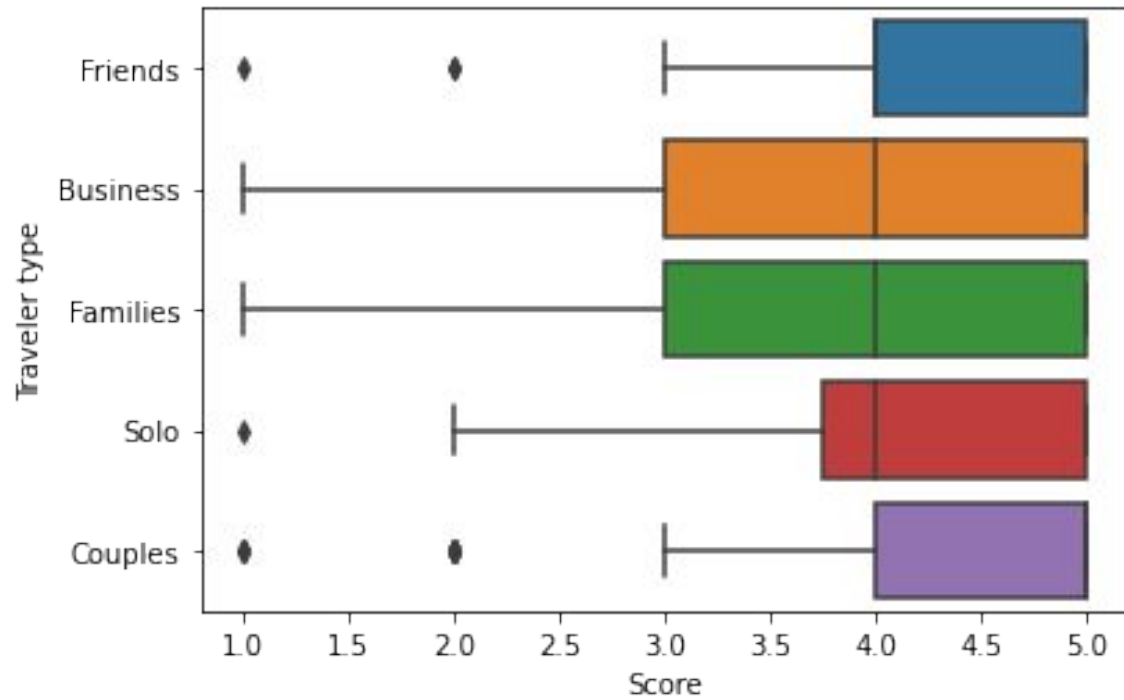
Heat Map



Pair Plot



Box Plot



Preliminary Results

- Weak correlations for score
 - Hotel name, Pool, Free internet
 - Backwards elimination
-
- What has proved harder than expected to this point?
 - ◆ Determining how to subset the data, and comparing boolean values to numeric values
 - What revisions have you made to your expected modelling efforts?
 - ◆ Based on quantitative data, not common sense/our ideas
 - What have you learned about your data and the topic in general?
 - ◆ There is a lot of comparisons that can be made and a lot of potential predictors

Next Steps

- Examine 'Traveler type' variable more closely
 - Correlations with each other column
 - Sample size for each traveler type
- Look at correlations for each traveler type individually
- Determine which graphical display demonstrates data best
- Determine correlation with each boolean variable
- Subset score and traveller type and compare correlation values to determine best predictor variables

Schedule

- **Part 3: Due 16 November 2021**

- Paper draft
- Finished by November 9
 - Introduction: Peyton
 - Data and methods: Christopher
 - Results: Brandon (Christopher + Peyton as needed)
- Demonstration
- Practice on 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 + Brandon
 - Conclusion : Peyton
- Presentation
 - Work on leading into Dec. 16
- Demonstration
 - Walk/talk through
 - Date to be determined

*I have neither given or received,
nor have I tolerated others' use of
unauthorized aid.*

*Peyton Camden, Christopher
Barua, Brandon Cook*