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'
 - o '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

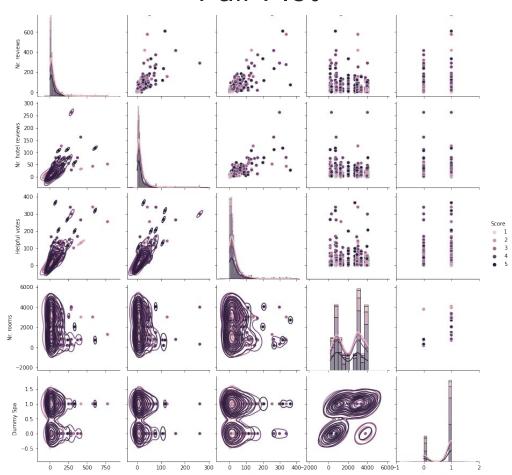
Analysis

- 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

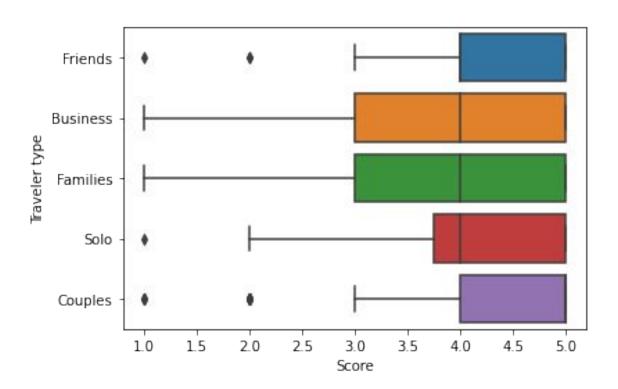
Heat Map

Nr. reviews -	1	0.66	0.79	-0.024	-0.14	0.023	-0.03	0.02	0.047	0.056	-0.07	-0.029	-0.098	-0.041	-0.071	-0.094	-0.058		- 1.00
Nr. hotel reviews -	0.66	1		0.014	-0.13	0.024	-0.04	0.054	0.044	0.077	-0.046	-0.021	-0.097	-0.065	-0.013	-0.02	0.0058		
Helpful votes -				0.015	-0.099	0.024	-0.015	0.021	0.042	0.061	-0.062	7.3e-05	-0.074	-0.021	-0.035	-0.09	-0.037		- 0.75
Score -	-0.024	0.014	0.015	1	-0.08	-0.036	0.017	-0.043	0.18	0.2	0.0088	0.066	0.05	-0.014	0.18	-0.0091	-0.048		
Nr. rooms -	-0.14	-0.13	-0.099	-0.08	1	-0.024	0.056	-0.045	-0.37	-0.27	0.35	-0.21	0.45	0.42	-0.14	2.4e-18	0.088		- 0.50
Member years -	0.023	0.024	0.024	-0.036	-0.024	1	-0.072	0.024	0.053	-0.0085	-0.0078	-0.079	-0.023	-0.014	-0.0078	0.057	0.028		
Num Traveler type -	-0.03	-0.04	-0.015	0.017	0.056	-0.072		-0.031	0.042	0.056	0.025	-0.022	0.093	0.047	-0.062	0.019	0.012		- 0.25
Num User country -	0.02	0.054	0.021	-0.043	-0.045	0.024	-0.031	1	0.012	0.085	0.028	-0.024	-0.034	0.0044	-0.034	0.059	0.41		
Num Hotel name -	0.047	0.044	0.042	0.18	-0.37	0.053	0.042	0.012	1	0.37	-2e-18	-0.092	-0.074	-0.29	0.3	0	-0.11		- 0.00
Dummy Pool -	0.056	0.077	0.061	0.2	-0.27	-0.0085	0.056	0.085	0.37	1	-0.05	0.12	0.4	-0.073	-0.05	1.4e-17	0.016		
Dummy Gym -	-0.07	-0.046	-0.062	0.0088	0.35	-0.0078	0.025	0.028	-2e-18	-0.05	1	0.12	0.4	-0.073	-0.05	1.4e-17	0.082		0.25
Dummy Tennis -	-0.029	-0.021	7.3e-05	0.066	-0.21	-0.079	-0.022	-0.024	-0.092	0.12	0.12	1	0.05	-0.2	0.12	4.2e-18	0.043		
Dummy Spa -	-0.098	-0.097	-0.074	0.05	0.45	-0.023	0.093	-0.034	-0.074	0.4	0.4	0.05	1	0.58	-0.12	2.4e-18	0.093		0.50
Dummy Casino -	-0.041	-0.065	-0.021	-0.014	0.42	-0.014	0.047	0.0044	-0.29	-0.073	-0.073	-0.2	0.58	1	-0.073	-7.4e-18	0.071		
Dummy Free Internet -	-0.071	-0.013	-0.035	0.18	-0.14	-0.0078	-0.062	-0.034	0.3	-0.05	-0.05	0.12	-0.12	-0.073	1	1.4e-17	-0.05		0.75
Num Review month -	-0.094	-0.02	-0.09	-0.0091	2.4e-18	0.057	0.019	0.059	0	1.4e-17	1.4e-17	4.2e-18	2.4e-18	-7.4e-18	31.4e-17	1	0.086		0.75
Num User continent -	-0.058	0.0058	-0.037	-0.048	0.088	0.028	0.012	0.41	-0.11	0.016	0.082	0.043	0.093	0.071	-0.05	0.086	1		
	Nr. reviews -	Nr. hotel reviews	Helpful votes -	Score -	Nr. rooms –	Member years -	Num Traveler type -	Num User country -	Num Hotel name -	Dummy Pool -	Dummy Gym -	Dummy Tennis -	Dummy Spa -	Dummy Casino	Dummy Free Internet -	Num Review month -	Num User continent -		1.00

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
 - o 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