Project Proposal

Team 19 Members:
- Xiang Zhou
- Maro Derhovanessians
- Chiebuka Onwuzurike

Project Dataset Link:
https://www.kaggle.com/datafiniti/hotel-reviews/activity

Our Motivation:
What drives us looking into this dataset is mainly because of the capstone project. Since we're investigating online reviews of a company and analyzing sentiments from those reviews so that we can possess a more comprehensive customer experience evaluation for the company. Furthermore, aside from the capstone project, we consider customer reviews analysis is an extremely significant part to the company development. It would provide business with feedback with exactly what customers really want. So businesses may be able to improve their customer service by an efficient and effective method to resolve the issue that customers faced.

Dataset Introduction:
The project we're diving in is more about customer reviews for the hotel industry in North America. The dataset contains a list of 1,000 hotels and their reviews provided by Datafiniti's Business Database. The dataset includes hotel location, name, rating, review data, title, username, and more. Thus, we can use the dataset to compare hotel reviews by state dimension, and testing on sentiment scoring and natural language processing techniques.

- Address: the address of a hotel
- Categories: Hotel, Motel, Casino
- City: city name
- Name: hotel name
- Province: state name
- Reviews.rating: review rating, from 1-5, 5 as the highest
- Reviews.text: comments collected from customers
- Reviews.title: title of reviews from customers
- Reviews.username: customer username, can be used to check the same user left multiple comments
- Reviews.date: the date that customer left review

What Methodology to Apply?

- NLP Sentiment Analysis - we can do some unsupervised analysis to see what are the patterns and correlations between review text and review score
- K-Means & Hierarchical Clustering - we can see if there are any clustering of good reviews or hotels given our different dimensions

Other Comments:

Since the objective is customer segmentation, we've gotten a clearer vision on how to process the data. Our initial idea is downsizing the data to different dimensions, like: review texts corresponding to each username and analyze sentiment by using NLP so that we can know more details of hotels based on keywords; in addition, review rating received from each user to have a quick view of hotel quality; moreover, categorizing based on hotel types.