Hotel Reviews

Kim Luong 014385096

Daniel Ly 01748170

Nathan Nguyen 012855724

James Takeda 014691311

Nam Vo 012440725

**Table of Contents**

Data 3

Problem 4

Analysis 4 - 5

Results 6

Conclusion 8

Code 9

**Data**

In this project, our group decided to utilize the hotel reviews data set. This data set includes 12 different variables: hotel address, categories, city, country, name of the hotel, postal code, province, date of review, when the review was added onto the list, review rating, the review itself, and the user’s username. For the hotel’s location; the variables composed consists of the name, address, categories, city, country, postal code, and province. Regarding to the hotels’ reviews and ratings itself, the essential variables would be the textual review, date of the review, date added onto the file, and review rating. We asked ourselves a couple questions; How can we provide meaning to a dataset as large as the hotel reviews Excel file? This was the main question we wanted to tackle when deciding how we would go about analyzing the file. The main problem we faced with this dataset was trying to figure out how to do an analysis with such a large set of data. Initially we decided to try and analyze all hotels in the US, but realized even that set of data was still too large. Ultimately, we settled on conducting several different analyses on all hotels in the California provinces. 200/300 provinces , 4000 data points.

When compiling the Hotel reviews on the data sheet for sentiment analysis we broke it down into descriptive statistic ie. finding the average, standard deviation, minimum, and maximum. We’ve translated all reviews into english as well as performing sentiment analysis as is. We also broke down the descriptive statistics on the ratings of the hotels. When we find the province with the best ratings, we narrowed down the ratings even more to find best cities within the province.

**Problem**

Businesses can make a lot out of big data, making it an important resource, much like oil, that you have to dig into and use what you can get. Data about sales date, sales amount, inventory, payment method, payment amount, customer information, and all other pertinent information that are involved in a transaction all flow and fill up a reservoir of data and it is the business’ role to interpret these data and make decisions that can affect higher income generation.When having such a large data set, we need to identify what we want out of it and therefore we chose to isolate California due to the fact that we live in California and it’ll be most usefully to recommend our out of state friends. We want to look for the best hotels (that includes best ratings and reviews)

In research, there is this thing called the margin of error or that leeway for miscalculations, exceptions, and other factors that cause a relatively dismissible amount of error. Analysis of the Hotel data, some had 5 stars rating and a horrible review and some had amazing review with low stars rating. As it is very difficult, if not impossible, to check the sentiment analysis manually, our best bet to ensure that our analytics will not provide gravely inaccurate data is to use a trusted data analysis tool that guarantees the highest level of accuracy.

**Analysis**

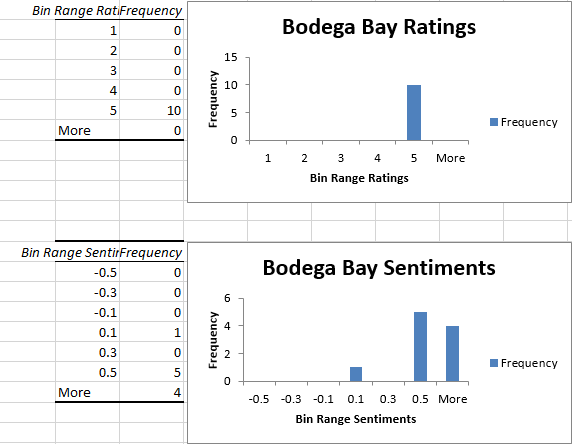
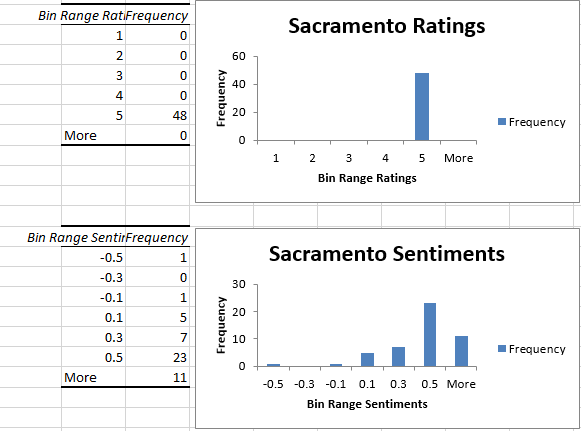
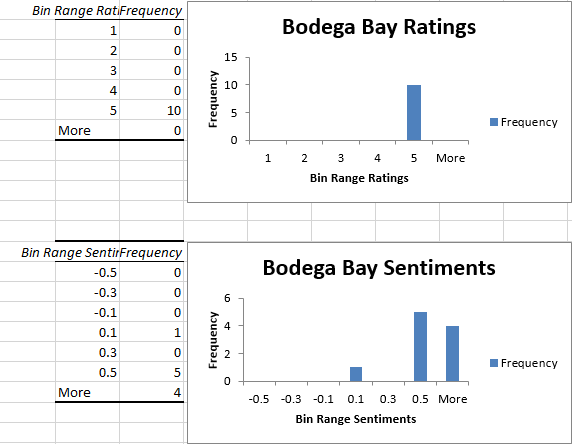
We decided to used various types of functions in Python to calculate different statistics in our variables. First, we classified a user-defined function and created a statistical table including the mean, standard deviation, sample size, maximum, minimum for the hotels’ reviews and sentiments. In order to create a user-friendly environment, we used the functions to create a table because it would be the best option to test and present our statistical outcomes in a straightforward manner. Next, we created different types of lists for the review ratings and sentiments to organize and separate California from other locations’ data points. The first list was particularly designated to count the amount of California provinces. The following two lists were for the review ratings and review sentiments from the hotel clients. Although lists are a changeable and ordered sequence of elements, there needs to be a method of counting and identifying the California cities just once without showing multiples. Through this technique, we developed a set so that it will only view and count California cities once.

While creating a loop to read the whole text file, multiple if-conditions are created to narrow down the specific data we want to analyze. As provinces are initialized as “province”, this approach filters out among the data to only California provinces. Another variable called rating is declared as “rating” and this filters out all the null items on the sheet. Initializing these two variables narrows down the data set to California and non-null ratings. Therefore, we save the names of the California cities in a variable and add it onto a set list so that it prints out the unique city names within California. Then, an if-else condition is created to use for each specific city so that that city’s corresponding rating and sentiment is added onto the list which the descriptive statistical operators can be called and performed later.

If this application is released to the public, we’ve also included a user input function so that the hotel client can see the California city of their choice’s ratings and sentiments. We created a loop asking the user first if they’d like to see a city’s rating and sentiment. Through this, we check for input validation if the user followed the directions. If the user does not enter the number 1 or 2, an error will occurred causing the application to not run anymore. Afterwards, an if-else condition is created to validate the hotel client’s city of their choice. This user input function allows anyone that’s interested in knowing the city’s star ratings and sentiment through gathering this information before their visit.

**Results**

The distribution graphs for the sentiments and ratings are positively correlated for Bodega Bay and Sacramento hotels. The distribution graphs specifically for ratings centers around four to five stars while their corresponding sentimental values centers in the positive range as well. We came to notice that most of the results from our code, specifically the average sentiment and average star rating correlated with keywords in guests reviews. For example, reviews that used positive words such as “great”, “good”, and “best” tended to have higher average star ratings and higher sentiment ratings as well. Conversely, guests who used words in their reviews such as “worst”, “bad”, or “never” resulted in lower average star ratings and lower average sentiment ratings. However, we did come across some outliers to this finding. Some hotels only had a small amount of ratings or even just one rating, making their average ratings either very positive or very negative. For example, there was only one hotel review at Willows, but it was an extremely positive review, making it have a high average star rating and a high average sentiment rating. The same happened for hotels that received single negative reviews, skewing the averages. In our analysis, the two highest averages we found for hotels found in Bodega Bay and Sacramento. From this we were able to deduct that these were some of the best two cities with the best hotels in the state of California. Likewise, we were able to also find the cities with the worst hotels based on guest reviews.



**Conclusion:**

Our group came into conclusion and utilized the methods of loops, user defined functions, if-statements, lists, sets, and dictionaries to translate the dataset of 35,912 holistic hotel reviews to just California provinces and cities. Ultimately, it all comes down to the guest experience. Companies that utilize data analysis to drive business decisions will outperform competitors. Especially in the hotel industry, data analytics and detailed guest reviews can be used in numerous different ways to improve business operations, occupancy rates and overall customer satisfactions.