Determining Which Hotel in Times Square is Best for Tourists

Bradley Betts

June 26, 2020

1. Introduction

1.1 Background

New York City is a very popular location for both tourists visiting the United States, and American citizens that wish to explore another state. In fact, New York City boasts about one third of the tourists in North America. A vast majority of these visitors will explore and/or reside in the Times Square area. It's likely that when someone is going to another location that isn't their home, they choose to stay in a hotel. In doing so, they might also use programs like Trivago, to inform them of the prices of staying at these different hotels. Certain services like this will also inform the user how far the hotels are from the closest major city. On top of that, people can simply google sites of interest in their designated city of choice. However, most people choose which hotel to stay at not only because of their cost or only because of what's nearby; it is usually a mixture of both variables.

1.2 Problem

This project seeks to determine which hotel within Times Square is the best choice based on cost per night, and number of nearby venues. The chosen hotel will alter based on a limitation to how much the user would spend and what venues are within a specific distance. Another goal of the project is to see if there is a correlation between cost and nearby venues.

1.3 Interest

Tourists venturing to New York City would for sure be interested in this project because it will make travelling to this bumbling city much easier to plan. Hotel executives would also find this information useful, as it would show them the cost of competition nearby, with the addition of which venues they could possibly market to their customers to increase their likelihood of staying in their hotel.

2. Data Acquisition and Cleaning

2.1 Data sources

The location of all the venues in New York City can be found using the Foursquare API, and the location of specific hotels will be gathered from an open-source CSV file on Kaggle shown here.

2.2 Data cleaning

There main problem that came with gathering the data of each hotel was finding their respective prices, as there was not a dataset containing all of these of these hotel costs to be found. This is partially since many hotels have different kinds of rooms, each with a different cost, and partially due to costs constantly changing. Therefore, the cost was determined by browsing each of the 41 hotels on Tripadvisor and marking the lowest cost. If the hotel was unable to be found on the website, or if there was not cost available, that hotel was excluded from the dataframe. The cost of each of the remaining 29 hotels were concatenated to the CSV file and were plotted out on a bar chart (Figure 1).

Different venues near Times Square were located using the Foursquare API, with their latitudes and longitudes recorded in a separate dataframe. Using this API in the project returned 30 venues of varying categories.

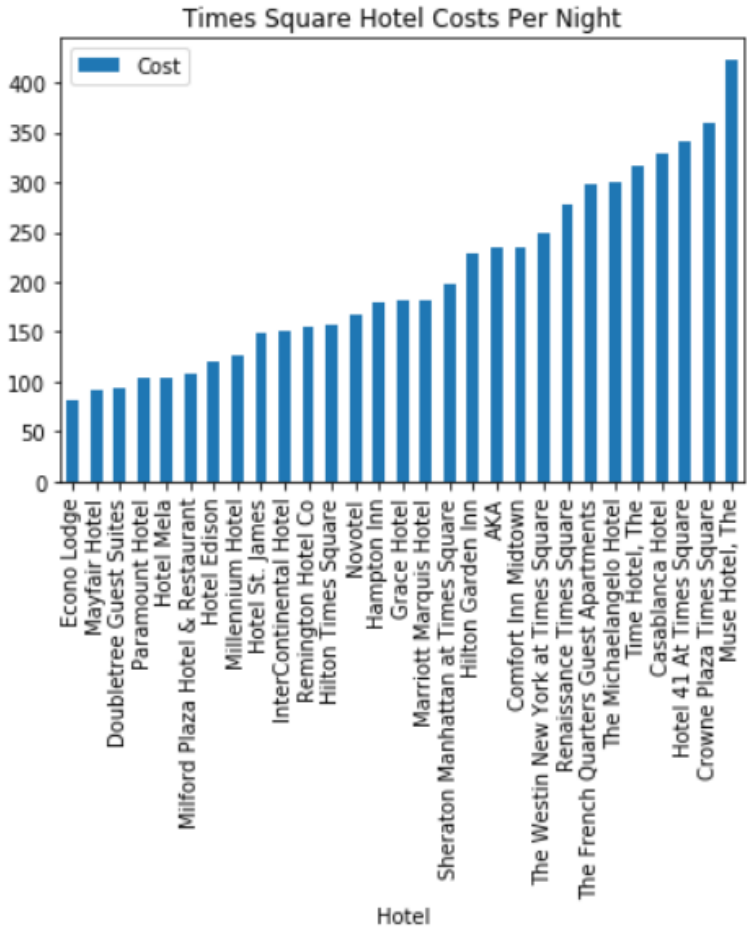

Figure 1. Bar chart of cost of different hotels

2.3 Feature selection

There were some features in the CSV file that proved to be unnecessary for data analysis, such as address, phone number, website and borough. The only features that remained in this dataframe were hotel name, latitude, longitude and cost. The Foursquare dataframe only contains the name,

category, latitude and longitude, as the rest of the features from the dataframe that was originally created with the Foursquare API are unneccessary.

3. Exploratory Data Analysis

3.1 Calculation of venues close to hotels

To find how many venues are within walking distance of the hotels, the geodesic distance was calculated for each venue to each hotel using geopy. Most of the hotels had almost every venue within 500 meters, so the function using geopy was modified to also calculate hotels within 250 meters on top of the original 500. This made the data significantly more varied (Table 1).

	Hotel	Latitude	Longitude	Cost	Venues_within_500m	Venues_within_250m
0	Paramount Hotel	40.759132	-73.986348	104	27	21
2	Millennium Hotel	40.756667	-73.984396	127	27	13
4	Novotel	40.762897	-73.983683	168	8	0
7	Doubletree Guest Suites	40.759055	-73.984710	93	27	20
9	Hotel 41 At Times Square	40.755597	-73.987999	340	23	6
10	Time Hotel, The	40.761014	-73.985052	316	25	7
11	Hilton Times Square	40.756747	-73.988659	157	25	8
12	The Westin New York at Times Square	40.757482	-73.988309	249	27	16
13	Crowne Plaza Times Square	40.760537	-73.984644	359	26	9
14	Renaissance Times Square	40.759581	-73.984366	277	27	11
16	Casablanca Hotel	40.756022	-73.984808	328	27	12
17	The French Quarters Guest Apartments	40.760359	-73.989286	299	22	2
18	Hotel St. James	40.757026	-73.983291	148	27	10
20	Hotel Mela	40.756461	-73.983952	104	27	11
21	AKA	40.756491	-73.983977	234	27	11
22	Grace Hotel	40.757163	-73.983613	181	27	13
23	Comfort Inn Midtown	40.757811	-73.983230	235	27	13
25	Muse Hotel, The	40.757814	-73.983280	423	27	13
28	Econo Lodge	40.760315	-73.987554	81	26	6
29	Hilton Garden Inn	40.761127	-73.986940	229	23	4

Table 1. Subsidized table of hotels with cost and number of venues within 250 and 500 meters

3.2 Relationship between cost and nearby venues

It is understood that houses and hotels alike are usually more expensive the closer they are to cities and other interesting attractions. In this case, there was not a significant correlation between cost and nearby venues. When 500 meters was used for the distance variable, there was almost no correlation at all (Figure 2). A slightly negative correlation was seen when 250 meters was used, but it isn't significant enough to say that the cost per night in a hotel is decided by the venues nearby (Figure 3).

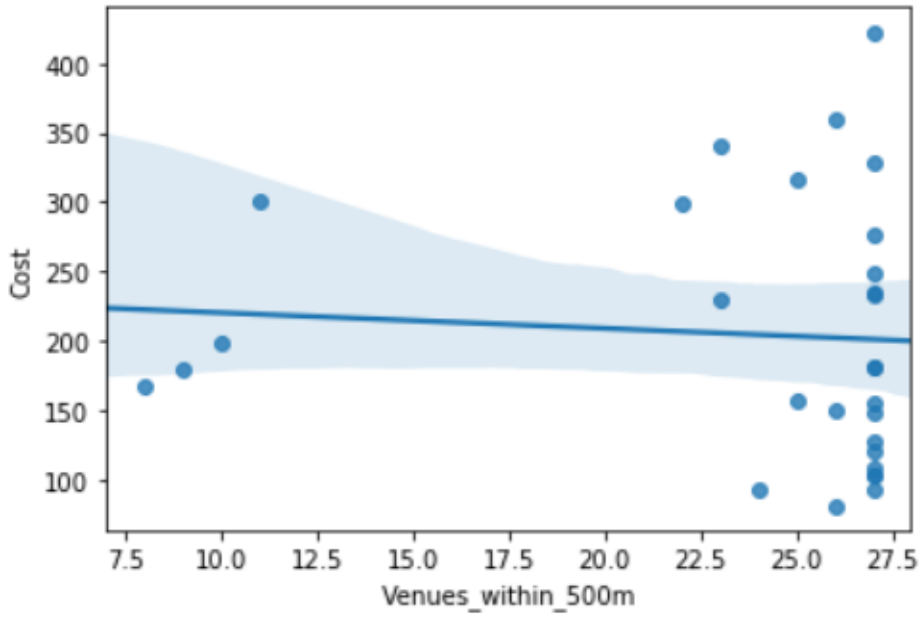

Figure 2. Scatter plot of cost and venues within 500 meters

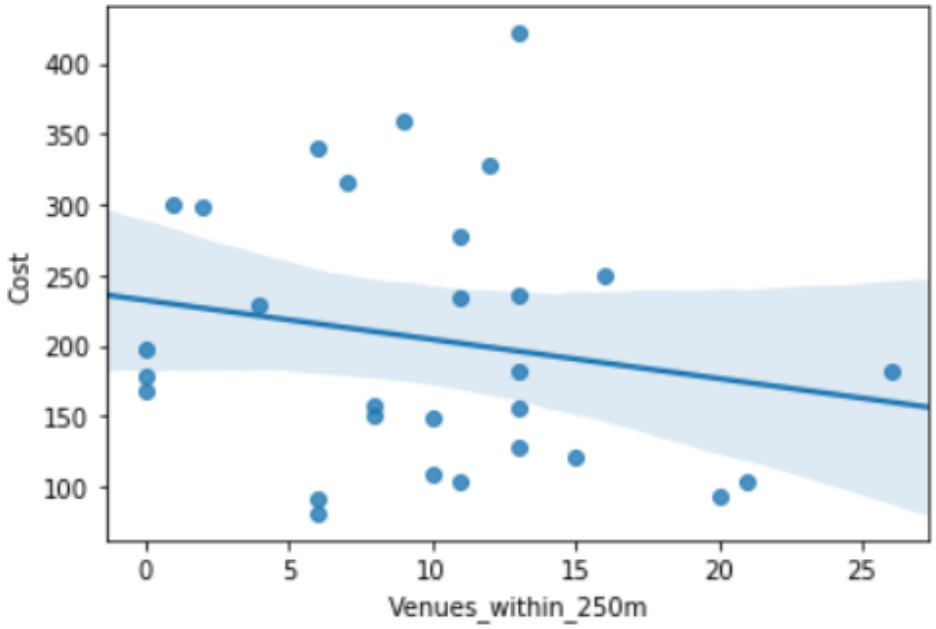

Figure 3. Scatter plot of cost and venues within 250 meters

4. Results

The data from this project shows that of the listed hotels, Doubletree Guest Suites is the best decision if that choice is based on the cost per night in that hotel and the nearby venues. The price of staying at this hotel is only $93/night with 20 nearby attractions. Another good candidate is Paramount Hotel, with similar results of $104/night and 21 nearby attractions. Due to the lack of correlation between cost and nearby venues, there are certainly other variables to influence the

cost of the hotel. Some of those variables would be hotel quality and size; however, while the size is easily scraped, the quality of the hotel and its service requires more advanced machine learning techniques.

5. Conclusions

In this project, the relationship between the cost of a hotel and the venues that surround it was analyzed. I created a table of each hotel with their respective cost and number of venues nearby as well as a regression model to show the lack of correlation between price and venue quantity. These models could be most useful for tourists deciding on where to stay in Times Square. Though the models can be of use to anyone looking to stay in a hotel to help them understand that nearby venues does not affect the hotel's cost whatsoever, so they could save money on gas by doing more research on what's in the area of the hotels they're looking into. The final table of hotel prices can be useful to hotel executives looking to further understand their nearby competition (which is far more prevalent in places, like Times Square, where hotels are clustered together).

6. Future directions

This study could be improved upon by analyzing more than just the nearby venues against the cost of a hotel, by using multiple regression and correlation. With this expanded study, it could show which independent variables weigh more on the dependent variable of cost.

Another psychological study of this data could analyze why tourists pick specific hotels, rating from 1-10 their opinion on the importance of cost, nearby venues, quality, etcetera. However, this would be a much longer study that would require a survey to be given to thousands, or possibly millions, of participants.