Comparing Orange County, CA and Hennepin County, MN

Erick Campos

December 20, 2020

1. Introduction

Disclaimer – this business report is for the final project for Coursera Course "Applied Data Science Capstone." While the data compiled is real, the proposed business case is entirely fictitious. Any real-world resemblance to this report is purely coincidental. Any thoughts and opinions presented are my own.

1.1 Background

Disneyland Resort is located in Orange County, CA and Walt Disney World is located in Orange County, FL. These two locations are on opposite coasts of the United States. These properties bring a large volume of economic value for its host city and the surrounding cities in its county. It could be advantageous to place a third Disney property somewhere in the middle of the United States. This project aims to compare Hennepin County, Minnesota to the location of Disneyland (the first Disney theme park) in Orange County, California to decide whether or not Hennepin County is a viable option for a third Disney property in the US. The comparison will be made using mapping with k-means clustering.

1.2 Problem

Data is restricted to the zip code data available for both counties, as well as the data recorded on the Foursquare API. Anything that exists in either county but is not available on the Foursquare database will be missed. The location of Hennepin County was picked based on an estimate of distance. The distance between Orange County, CA, Hennepin County, MN, and Orange County, FL would more or less create a triangle with sides of equal distance. The k-means factor selected is arbitrary in order to keep things consistent between the two counties.

1.3 Interest

The potential stakeholders of interest include The Walt Disney Company, as well as the residents of Hennepin. Things to consider are monetary benefits or costs as well as how adding a large theme park would affect the way of life in Hennepin County.

2. Data Acquisition and Cleaning

2.1 Data Sources

The two data sources used to extract the zip codes for Orange County, CA and Hennepin County, MN are found at https://www.zipcodestogo.com/California/ and https://www.zipcodestogo.com/Minnesota/ . The latitude and longitude data for the zip codes come from a function that calls upon the ArcGIS mapping API. The data source for the venues of interest for both counties will come from the Foursquare API.

2.2 Data Cleaning

The data from the two webpages listed in the previous section were transposed into two tables, one for California and one for Minnesota. The data tables created include columns for zip codes, cities, and counties. From there, the data tables were filtered on the counties of interest: Orange County and

Hennepin County. The data was then mapped to ensure that there were no data discrepancies. After that, the location data was matched to its corresponding venue data from the Foursquare API. Any rows containing NaN values were once again removed to avoid errors. The final step was clustering the location and venue data based on similar venues using a k-means clustering with a factor of 5.

2.3 Feature Selection

The information to be examined is the way the venues for each county were clustered. The comparison between the two counties will be made using a k-means clustering with a factor of 5. The decision to place a Disney theme park in Hennepin County will be made based on the comparison between the two different sets of clusters. The clusters do not have to be an exact match – rather we must make assumptions as to why it would be justifiable to place a Disney property in Hennepin County based on the story that those clusters are telling us. We must keep in mind that the clusters in Orange County may have been influenced by Disneyland and not the other way around.