Northwestern University

Data Engineering 200, Winter 2023

Final Project Report

Andrew Yang and Dietmar Krause

Introduction

Our project was a data analysis on the Fast Food Industry, and more specifically on its restaurants. We wanted to analyze trends between Fast Food Restaurants, focusing on how restaurants compare to each other in terms of popularity, as well as how restaurants correlate with population count and income level. Our project would be helpful for a wide variety of practices, though most notably it could help restaurant owners and real estate businesses. Restaurant owners could see our data analysis on popularity and composition, and determine where opening new restaurants would have the most impact. Those who have their hands in real estate managing/building would see use in our population/income data analysis, as they might better understand where properties are projected to have more/less monetary value.

Dataset. Introduce and describe in detail about the datasets that you are using. e.g., attributes, size, instances. Clearly describe the datasets and the amount of data you are using in your project. etc.

The “Fast Food Restaurants in America” from Kaggle was used as the main dataset of the project. It Consisted on 10.000 datapoints with information about the location, the name, and the website of Fast food restaurants in the US;provided by Datafiniti's Business Database. . We used the entire dataset, without removing any row. In addition we worked with other three datasets, also from Kaggle. The first one was the population by state, which has the population estimate for each state for the year 2019, containing 50 rows, one for each state, which was originally created from the dataset of the US Governmanet Census Bureau. The second one was the Annual state-level income of USA, which contains 15 columns and 550 rows for datapoints describing the level of income, equity indicators regarding racial, socioeconomical and ethnical implications, where the one we focused was in the GDP of the state . The third one was the “US Population By Zip Code”. The United States census dataset includes nationwide population counts from the 2000 and 2010 censuses. Data is broken out by gender, age and location using zip code tabular areas (ZCTAs) and GEOIDs. ZCTAs are generalized representations of zip codes, and often, though not always, are the same as the zip code for an area. GEOIDs are numeric codes that uniquely identify all administrative, legal, and statistical geographic areas for which the Census Bureau tabulates data. GEOIDs are useful for correlating census data with other censuses and surveys., being the latter the one we used.

Data Processing and Analysis. Describe what did you do to understand the data. Any cleaning? aggregation or analysis? Explain details and how they are relevant to your project topic.

We did multiple cleaning to our data. The first thing we realized was that the latitude and longitude provided in the main dataset was not correctly labeled, so we had discrepancies with the zipcodes of the same row. After plotting the data and the map, we could saw the the zipcodes were consistent with the restaurant locations, but the latitude and longitudes were not, so we drop that columns. In the same dataset, when trying to filter by name we realized that there were misspelled and duplicate names for the same brand, so we aggregated the names by creating unique labels and matching those names with their “parent labels” (e.g Mcdonalds and MC donals were two different tags that after the cleaning were tagged as the same McDonalds.) . For the population by zipcode we had to aggregate the data since it was not clearly structured within the dataset, there were rows in the middle of the data containing Totals (i.e the sum of the previous rows) and other ones that were duplicated, so we unified them so we could have an accurate count that match the ones that we could see in the Burous's page. Another thing that was also relevant, was that the names of the state varies through the datasets, so we had to unified them by convert state names in various formats into unified state codes.

Summary of Insights

From our Tableau analysis, we primarily saw that McDonald's was the most popular Fast Food Restaurant by far, constituting 21% of all restaurants in the United States. This was more evident in our per-state breakdown of fast food restaurants, where we found McDonald's to almost consistently be the most popular restaurant in each state. Another interesting trend we discovered was that for states with fewer total number of restaurants, there was a more even breakdown of restaurants in terms of popularity. Following the trend, we saw that states with more total restaurants had a less even breakdown of restaurant popularity. To give an example, for Illinois, which is the state with the 6th highest total number of restaurants, 52% of all its restaurants are one of the following: McDonald's, Burger King, or Taco Bell. However, Montana, which is the state with the 5th lowest total number of restaurants, has a much more even distribution of restaurants, evident in its state-composition bar graph.

For our Python analysis, we found out that there was a strong relationship (as we expected) between the wealth of each states and the number of restaurants, being New York not only the state with the greater number of people per restaurants, but also the one with each restaurants receive more proportional to the GDP of the state. Also, in the map we generated using Matplotly, we realize that the density of fast food restaurants is much more dense in the east coast, where most of the fast food restaurant seems to be located. We expected to see the same population/restaurant ratio trend in the GDP/ restaurant, but we found out that there's state with much more population that have even less restaurants, which implies that the income distribution is more decisive than the population distribution when it comes to the number of restaurants in the area.

Teamwork

Our team worked swiftly and effectively on this final project. We began by reading the document and brainstorming ideas for questions to answer, as well as how we would organize the project. Once we had four questions that we thought were insightful and valuable, we assigned two of them to be done in Python, and the other two in Tableau. Working with our strengths, we decided to both have a hand in cleaning/exploring the data, but left the Python work to Dietmar, and the Tableau section to Andrew. After we finished our respective parts, we cross-checked our work and ensured everything was correct. Our group utilized a hybrid system of meetings, coming together to work in person at University Library, as well as organizing consistent Zoom meetings to check in. Overall, the team dynamic was great, and both of us worked well with each other to complete our project.