Capstone 1 - Milestone Report

Introduction

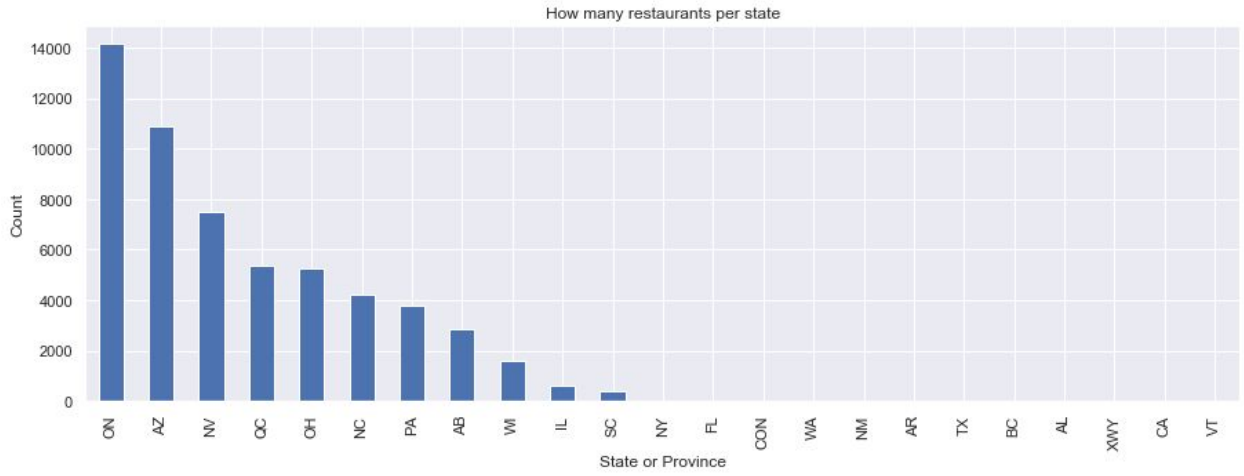
Cultural exportation and importation has often been talked about such as Hollywood's and by extension the US's global media influence. This cultural export and import is not really measured and my capstone aims to provide a methodology and analysis on cultural export. Various products have different cultural weight such as unprocessed wheat having little to no cultural weight, but prepared meals often carrying high cultural significance. A burger, pasta, or dim sum is more representative of America, Italy, and China respectively than flour and meat from each one for example. While I don't know what products have how much cultural weight (movies vs toys) they have compared to each other, I can compare amounts of one product which should have similar weights to each other thus I am using restaurants as a proxy for cultural influence.

The parties interested in this data would be companies in industries that are heavily influenced by cultural factors such as entertainment, travel services, and food. If a company is in an industry that is more sensitive to cultural sentiment, they could use the data or methodology to decide whether to enter into a market depending on whether or not there is enough demand for those goods. A company selling Japanese translated books would do much better in areas that are receptive to Japanese culture for example.

The Dataset

The dataset is from Yelp's dataset challenge which was also uploaded to Kaggle (https://www.kaggle.com/yelp-dataset/yelp-dataset). This dataset has several parts: Business information, tip information, check in information, reviews, tips, user information, and photos of businesses in JSON format. The only sections we will be using is the business information and the check in information as we want to know what their business is, business tags, and how many times each business has been checked in.

The dataset needed to be flattened from the nested JSON into more workable Pandas dataframes and CSV files. From then, I had to take only the restaurant businesses and separate the business tags into their own separate categories by national origin(Chinese, Ethiopian, Mexican, ect.) with dummy variables. A small caveat shown below was found after working on the data for a bit; The dataset is only a small subset of Yelp's data, only showing data in several western states in the United States and a few provinces in Canada. What this means is that the analysis only really applies to cultural importation in these areas.

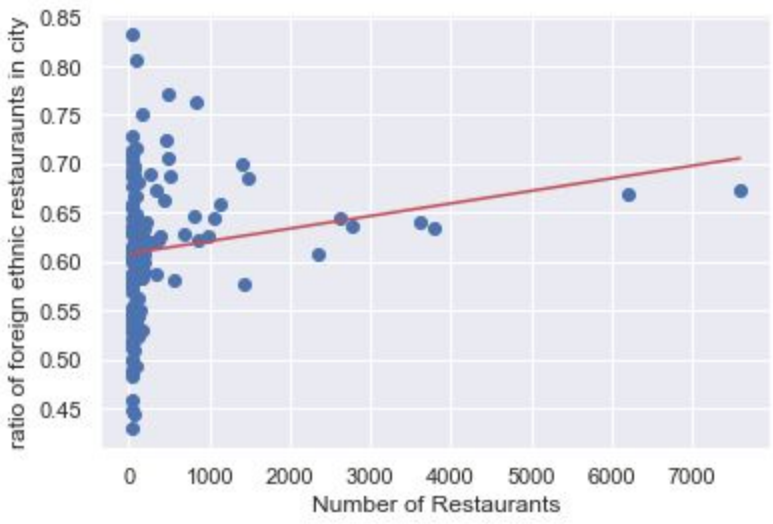


Analysis

Regression among ratio of restaurants in cities

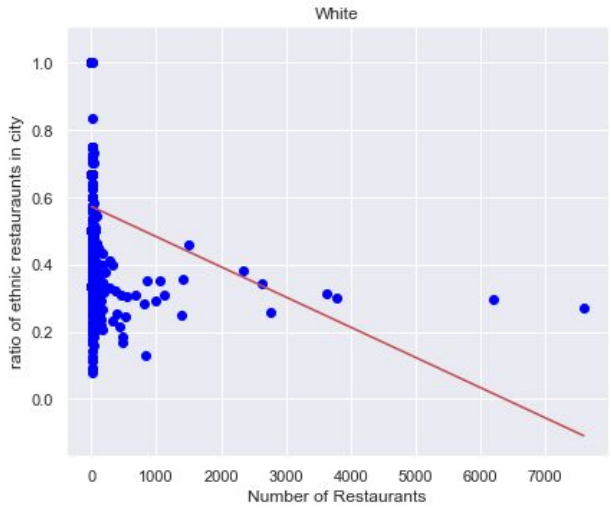
A lot of towns had only a singular observed restaurant in the dataset which skewed the data, making the r-squared value (0.02943) incredibly low meaning the data points don't relate to the regression well. To reduce the effect on towns with a small amount of restaurants skewing the data, I took the cities in the top 80th percentile of number of restaurants and the r-value became 0.17880 which is still relatively low, but shows that it's not an irrelevant link.

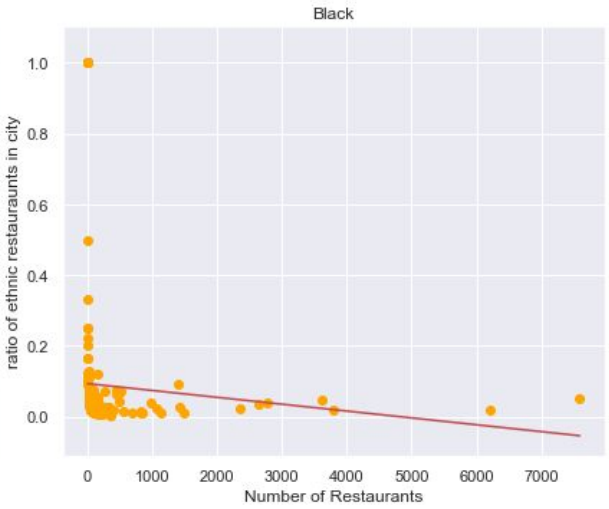
What the results mean is that as a city grows, there are more foreign restaurants as a percentage of restaurants in the city indicating that there is more cultural influence and imports in larger cities. The R-value is not that high, but in a measurement of both the high population cities and all cities, the intercept is around .6 indicating that there is a high base influence from foreign cultures in our dataset.

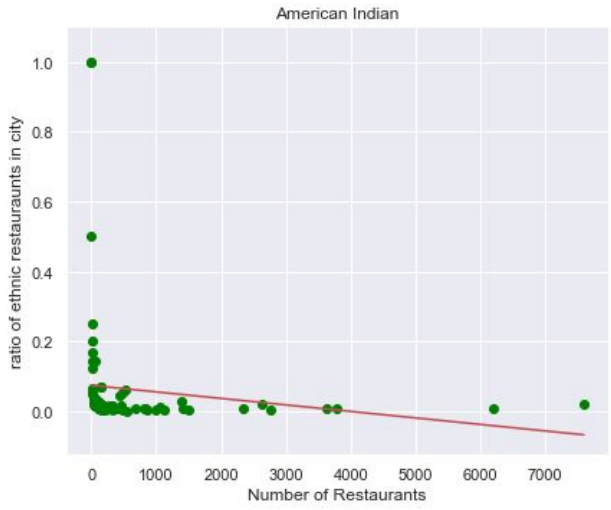


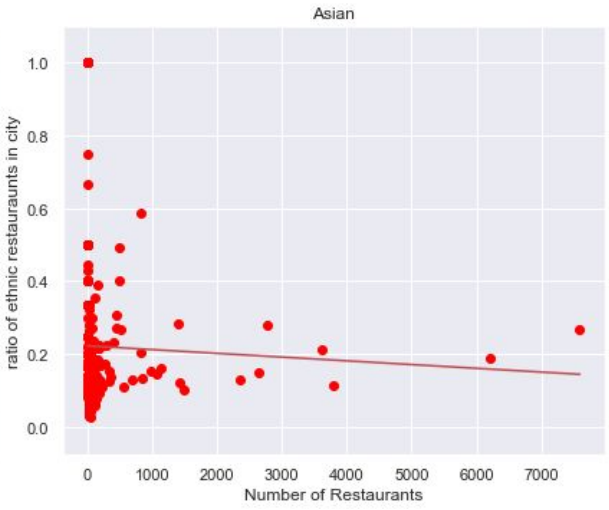
Regression among different ethnic restaurants

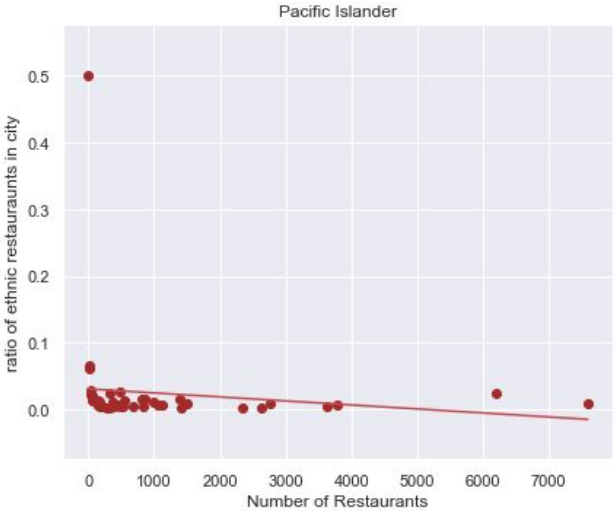
I have split up this figure into six separate charts to show the six racial types that the US census bureau checks. These plots have the count of restaurants in the X axis and the Y axis being the ratio of ethnic restaurants in that racial type compared to all restaurants in that city with each point being a city. All of these plots have a negative relationship as not all restaurants are going to be tagged, but the less negative the R in the regression is the higher cultural influence that racial category exerts.

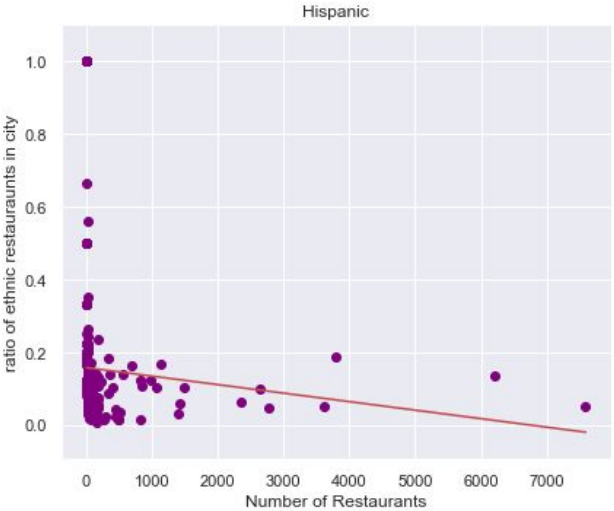

White
ratio of ethnic restaurauants in city
Number of Restaurants


Black
ratio of ethnic restaurauants in city
Number of Restaurants


American Indian
ratio of ethnic restaurauants in city
Number of Restaurants

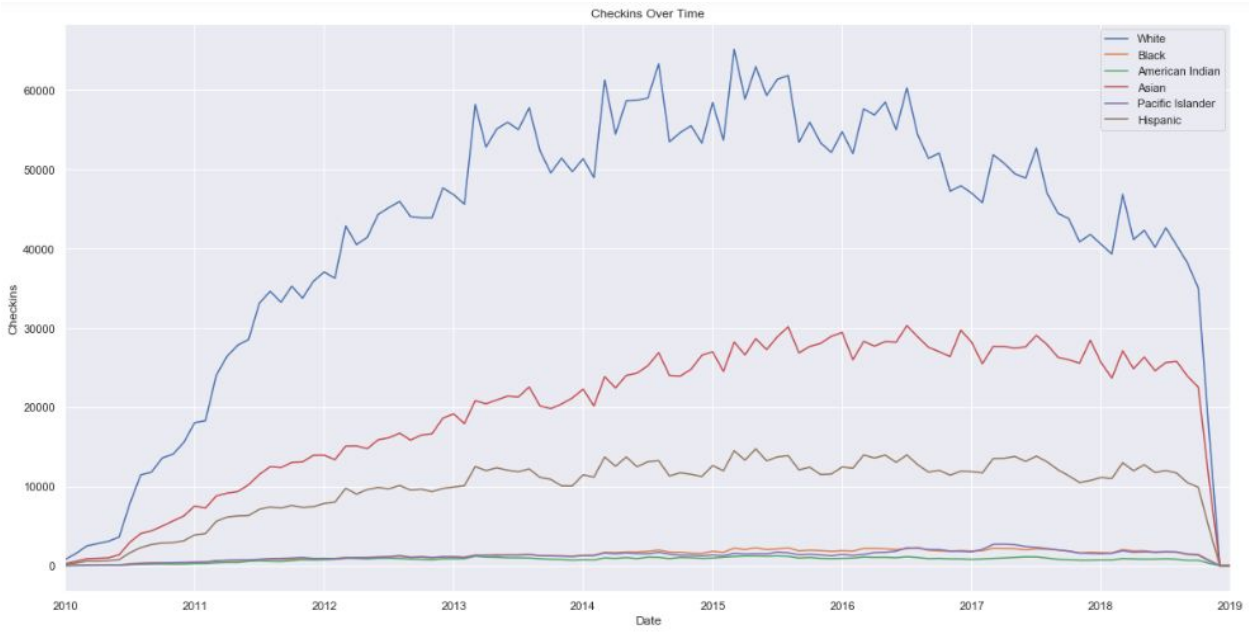

Asian
ratio of ethnic restaurauants in city
Number of Restaurants


Pacific Islander
ratio of ethnic restaurauants in city
Number of Restaurants


Hispanic
ratio of ethnic restaurauants in city
Number of Restaurants

Time Series analysis of check ins to restaurants of different racial origins

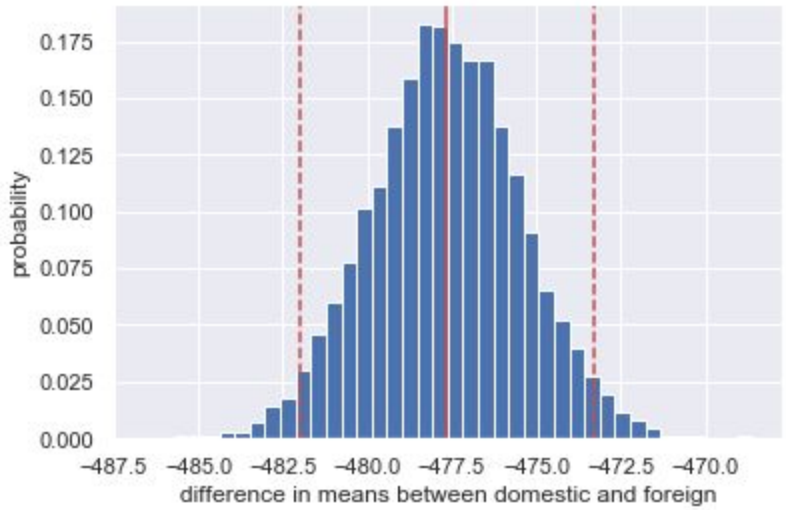
From the dataset, all of the racial categories kept gaining check ins to 2013 and 2014 where they plateaued. The dataset is composed of checkins from restaurants throughout Canada and the USA, the racial types are those that are defined by the US census bureau. The US census bureau does not check for ethnicity(Greek, Cambodian, Mexican, ect.) but racial types(White, black, Asian, ect.) so restaurants of different ethnicities were bucketed into the corresponding racial type. Check Ins to a "white" restaurant is the effect of domestic culture as the dataset is from the US and Canada with a predominantly white population.



From 2010 to mid 2013 check ins for restaurants of all racial types rose before mainly plateauing off. This raise in checkins is likely due to Yelp's growth as a platform as it grows for all racial types of restaurants in this period. Less people went to restaurants that are coded to white ethnicities in the middle of 2015 yet the amount of restaurant goers to asian, hispanic, and other restaurant types stayed relatively constant. Though white restaurants lost restaurant goers overall, the other restaurant types did not absorb the loss in cultural sway. This indicates that the cultural influence of foreign cultures is relatively stable, but domestic influence is lowering.

Inferential Statistical analysis

As the data I was working with is mainly categorical data, I could only really do analysis on the counts of each. I bootstrapped the dataset and got a few thousand replicates to see if the true means between the count of domestic and foreign restaurants is the same. The lower limit of the 95% confidence interval for the true mean of domestic origin restaurants is 17769. The lower limit of the 95% confidence interval of the true mean for foreign restaurants is 18246.



With a zero being outside of the 95% confidence interval, we will have to reject the null hypothesis that there is no difference in the means of domestic and foreign foods. What this implies is that foreign restaurants and by proxy foreign culture have a stronger presence than domestic culture in the US and Canada where the dataset is taken from.