Comparing Fast Food Restaurant Density to Median Household Income

Group 5
Michael Battaglia, Sarah Gearardo, Deep Ghosh,
Julie Magro, Melissa Nondorf

Our Topic

- In the US, fast food restaurants are generally thought of as "cheap food" and are often associated with low-income areas in people's minds. **We will use machine learning to investigate if there is indeed a significant correlation between median income and density of chain fast food restaurants (CFFRs) by zip code**.
- We chose a subset of eight states (CA, CO, FL, IL, NY, TN, TX, and VA) to run our analysis on in order to keep the scope of the project manageable, given the project's time constraints.

- We chose this topic because it sounded interesting and offered a good way to test assumptions around socio-economic status and food options.

Data

- The data was collated from primarily 3 sources:
  - Census Data on US Household Income Statistics for 2011-2015 sourced from Kaggle.com
  - Restaurant locations scraped from Yelp.
  - Wikipedia's list of the largest fast food chains globally, which we reduced to only companies with a US presence.

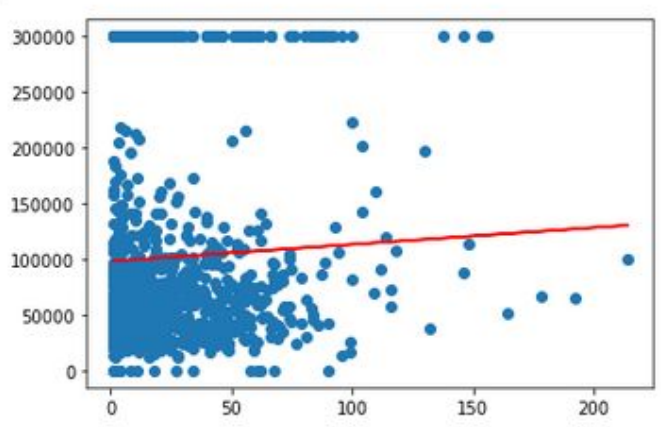
- Our hypothesis was that there is a strong negative correlation between density of CFFRs and median income level per zip code.
- In the data exploratory phase, the team had to contend with multiple challenges pertaining to relevancy and accuracy of the data to arrive at the right data source.
- Further, the data had to be cleaned and prepped into separate data tables for further analysis.

Resources

- The project involved primarily the use of,
  - Python in Jupyter Notebook for data analysis,
  - Postgres and PgAdmin to join and query the data,
  - GitHub, Slack, Google Docs and Zoom as collaboration tools

Analysis

- We will use supervised learning for this project, specifically using a linear regression model to identify relationships between the data.
  - For example, from the initial view of the income data and the number of restaurants, it seems like there is a cluster of zip-codes with the same income of $300K.
- The hypothesis we want to test is whether density of fast food restaurants (x) has an effect on the median income level (y) of the zip code we are considering.
- If the data and time permit, we will also attempt to parse out types of restaurants to see if there are any differences in preference between income levels for certain chains or types of food.




```
[150.85486846]
98046.2566762438
```

InnerGrouped_Results.csv

```
"zip_code","count","median"
"90680",12,"$47,046.00"
"90755",22,"$36,506.00"
"22182",52,"$113,602.00"
"90043",47,"$31,219.00"
"91730",4,"$85,634.00"
"90247",54,"$35,604.00"
"11210",32,"$48,829.00"
"90260",43,"$45,831.00"
"60189",3,"$83,140.00"
"10002",112,"$91,469.00"
"91202",6,"$113,952.00"
"93536",9,"$57,343.00"
```

Results

- The initial correlation results suggests a slope of 150.85 with a y-intercept of 98,046.
- This suggests that for a zip-code, $98K is the base median income and for every additional restaurant, the model expects and increase of $150 in income for that zip.
- The results suggest that income levels for a particular zip would increase with the number of fast food restaurants which goes against our initial hypothesis.

In [12]:

```
y_pred = classifier.predict(X_test)
results=pd.DataFrame({"Prediction":y_pred, "Actual":y_test}).reset_index(drop=True)
results.head(30)
```

Out[12]:

	Prediction	Actual
0	103738.114106	25700
1	113528.083564	300000
2	100868.640299	33741
3	107620.343374	109102
4	101881.395760	0
5	100193.469992	20792
6	108801.891412	300000
7	128888.208058	65963
8	106438.795335	300000
9	99011.921953	300000
10	103400.528952	37333
11	102387.773491	160021
12	100193.469992	195953

Recommendations

- Next steps would include,
  - Incorporate the findings into the machine learning model to predict income levels for a zip based on the number of fast food restaurants.
  - Work to understand ways to work towards a greater R-squared value in the model to increase its accuracy.
-