DATA-390B: Model and Data Usage (Version 2)

Maha Mapara

25th November, 2020

1 How is version 2 different from version 1?

Version 2 describes how the data sets will be used individually, given that I will not be able to merge them. It also provides clarity on the data that I was looking for last week and what I have found now.

The reason the data can not be merged is that they do not have the variables where I can join them; and the data was collected during different time periods.

Additionally, I have narrowed down my modeling approach to two ensemble models for obesity prediction. I may still use other modeling approaches for data analysis, like using support vector machine.

The sections below detail the information on the data sets, modeling and show some of the exploratory analysis I have done.

2 The data

2.1 Louisiana County Health Ranking Data

Overview: This is the county health ranking data from Louisiana's Department of Health. The data set contains the ranks for each county in Louisiana and the underlying data details for the measures used in calculating the yearly county health rankings [4]. Beside the main county rankings, there are sub ranks for variables like quality of life, length of life, clinical care, socio-economic factors and environmental factors. Then there is data that is used to calculate the rankings using information on smoking, alcohol use, physical exercise, nutrition, obesity, insurance, education, income, and poverty.

I am using this data as the primary data for predict obesity rates, as this is the only data set measuring obesity. At the moment, I am merging all the ranking data for Louisiana from 2010 to 2020. The data is not available as an aggregate on the website and due to its formatting, I cannot use R or python functions join to merge data sets. Due to this I have to manually merge the data sets and I am still in the process of doing that. This is time consuming because of the manual work but also because the definitions of certain measures change from year to year, or new variables are being measured and older ones are not measured which leads to some confusion on how to merge the data.

2.2 Food Research Atlas Data

Overview: This data set is from the US Department of Agriculture; it presents a spatial overview of food access indicators for low-income and other census tracts using different measures of supermarket accessibility, it provides food access data for populations within census tracts; and offers census-tract-level data on food access that can be downloaded for community planning or research purposes [5].

I am using this data to expand on some variables on food access from the county health ranking data. The idea is to magnify correlations between food access, access to transportation, poverty, and family income. This data is also a good supplement to the county ranking data as the time frame for the data collection overlaps. This data was collected from 2013 to 2017 and the county data is from 2010 to 2020. I will be re-downloading this data set to account for the new update made on November 13, 2020.

2.3 Walmart opening data

The only data I could find on Walmart was a store opening data set, covering openings from 1970 to 2006. The data set does not have any useful variables except for store opening dates and where they opened. I may use the zip code data in it and map out where the Walmart stores opened in the states I research [1].

2.4 Quick service restaurant data

Statista has useful data on quick service restaurants including information on restaurant density and per capita. I am working with LITS and the Computer Science department to access it [2].

The Economic Census County Business Patterns data for numbers of Limited Service Restaurants establishments is also available by county in Louisiana [3]. The Bureau of Labor Statistics Consumer Expenditures Survey also has data on food eaten outside the home. It is not broken down by restaurant or county, but a state overview should be useful to understand these eating patterns in relation to obesity.

3 Modeling ideas

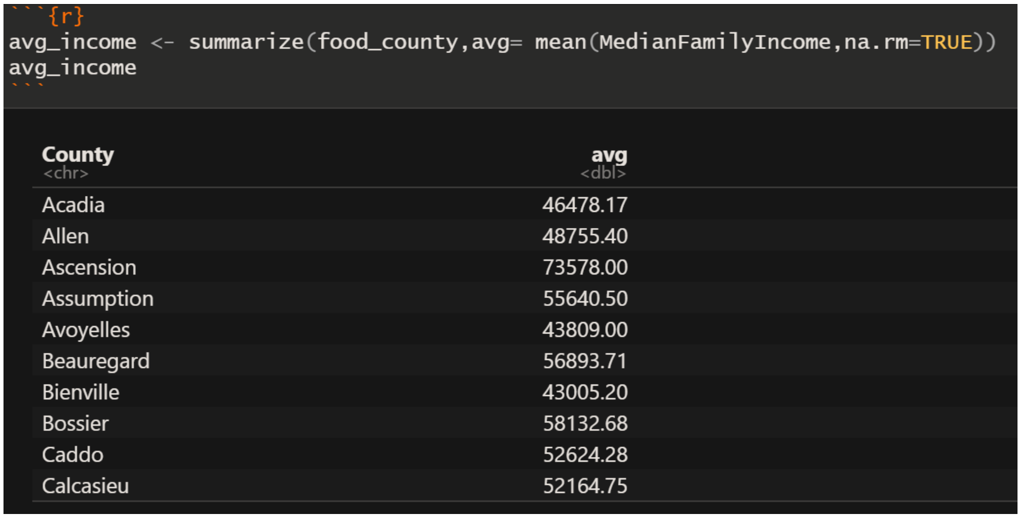
Recap from last week: I am predicting obesity rates using the County Health Rankings data.

For the prediction, I am focusing on ensemble models. The two I will use are random forest and extreme gradient boosting; and will see which performs the best.

4 Exploratory Data Analysis on Food Access data

Each county has multiple row entries depending on the county size.

Figure 1 shows a snapshot of the average median family income for each county and the code:

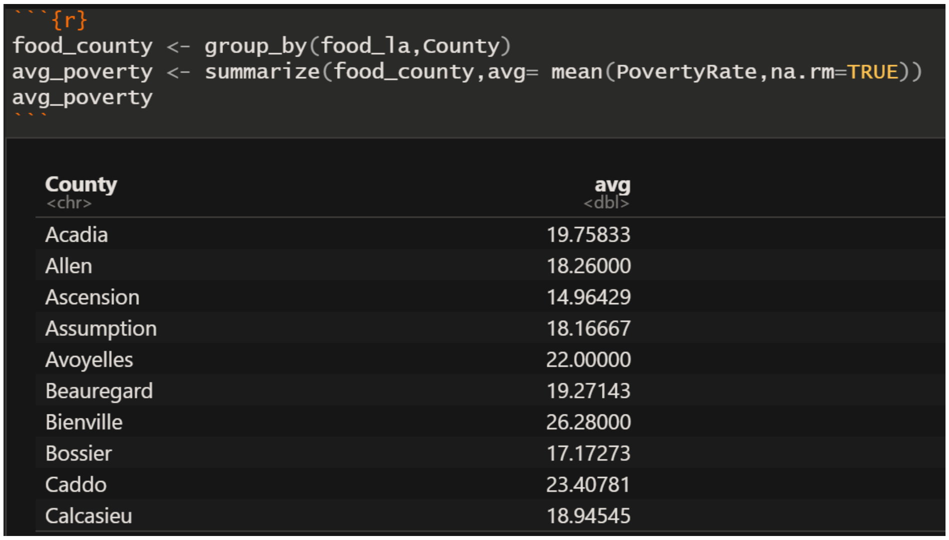


~~~
```{r}
avg_income <- summarize(food_county,avg= mean(MedianFamilyIncome,na.rm=TRUE))
avg_income
```
~~~

```
County                          avg
<chr>                         <dbl>
Acadia                     46478.17
Allen                      48755.40
Ascension                  73578.00
Assumption                 55640.50
Avoyelles                  43809.00
Beauregard                 56893.71
Bienville                  43005.20
Bossier                    58132.68
Caddo                      52624.28
Calcasieu                  52164.75
```

Figure 1: Average median family income by county

Figure 2 shows a snapshot of the average poverty rate for each county and the code:



~~~
```{r}
food_county <- group_by(food_la,County)
avg_poverty <- summarize(food_county,avg= mean(PovertyRate,na.rm=TRUE))
avg_poverty
```
~~~

```
County                                    avg
<chr>                                   <dbl>
Acadia                               19.75833
Allen                                18.26000
Ascension                            14.96429
Assumption                           18.16667
Avoyelles                            22.00000
Beauregard                           19.27143
Bienville                            26.28000
Bossier                              17.17273
Caddo                                23.40781
Calcasieu                            18.94545
```

Figure 2: Average poverty rate by county

Figure 3 is a plot of the poverty rate against the median household income for each county. The points are colored by whether people with low income and low access have vehicular access or not (LILATracts_Vehicle).

Figure 4 is a plot of population count beyond 1/2 mile from supermarket against low income population count beyond 1/2 mile from supermarket. The points are colored by LATractsVehicle_20 which is a flag for tract where greater or equal to 100 of households do not have a vehicle, and beyond 1/2 mile from the supermarket; or greater or equal to 500 individuals are beyond 20 miles from the supermarket ; or greater or equal to 33% of individuals are beyond 20 miles from the supermarket. [Note: Figure 4 appears after the reference section despite not being placed there. I don't know how to fix this so we'll have to work with this.]

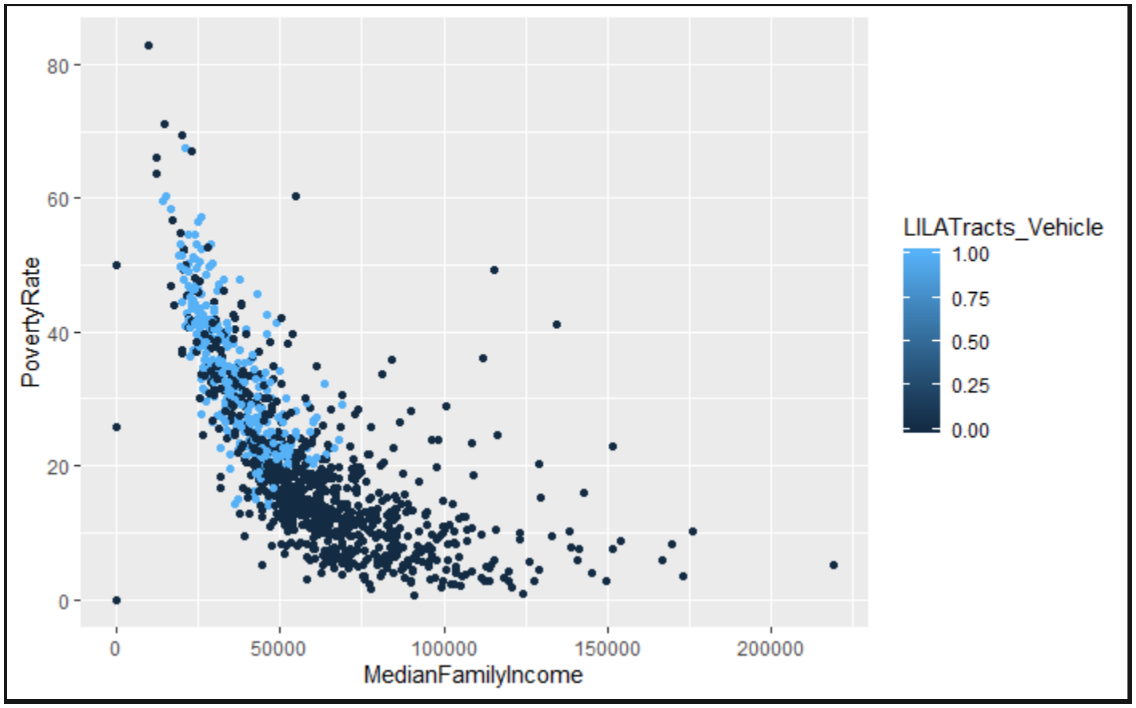


Figure 3: Poverty rate against the median household income for each county

References

[1] Thomas J. Holmes. The diffusion of wal-mart and economies of density. 2011.

[2] S. Lock. Number of quick service restaurant (qsr) franchise establishments in the united states from 2007 to 2020. 2020.

[3] United States Census of Bureau. County business patterns. 2020.

[4] LA Department of Health. Louisiana county health ranking data. 2020.

[5] USDA. Food access research atlas. 2017.

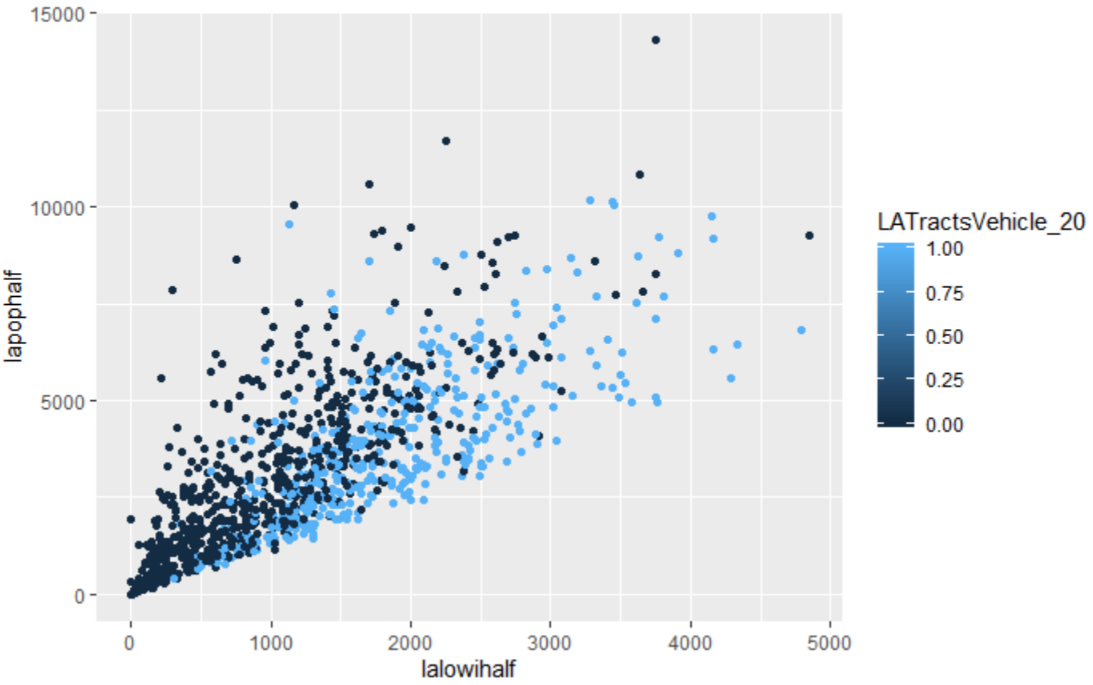


Figure 4: Population count beyond 1/2 mile from supermarket vs. low income population count beyond 1/2 mile from supermarket