process_book.md

Process Book

Overview

The aim of our project is to derive insights about food distribution and the availability of healthy eating options in the US using data from the USDA Food Environment Atlas, supplemented with data from the USDA Branded Food Products Database.

The Food Environment Atlas contains data at the county level and has many features, including access to restaurants, grocery stores, farmers markets and welfare programs, as well as demographic and health data. The Food Products Database contains the nutritional composition of branded and private-label foods.

We plan to produce a web app that features an interactive map of the US to visually display food access data, and a prediction calculator.

Task Timeline

We've decided to work concurrently on the various tasks required for our project, and have allocated the next four weeks to:

- [X] **Week 1**: We plan to spend the majority of week one exploring the datasets, refining the questions we'd like the data to answer, and identifying the variables that we'll use in our model. We'll also spend some time building the basic US map visualization and scaffolding out the web app.

☑ **Week 2**: This week will be spent working with Tableau to help us visualize the data, and building the model that that app will use. We'll also decide on the exact D3 visualizations that we will include.

- [X] **Week 3**: Populate the map visualization and build and test the prediction calculator.

☐ **Week 4**: Finalize the visualizations, trying to use the data to tell a cohesive story.

Roles and Responsibilities

We plan to use a trello board to keep track of our assigned tasks. We'll also check in during weekly meetings, and communicate via Slack.

As a group, we are all very interested in diving into the data science aspect of this project, so all four of us will collaborate on that. Additionally, Tushal Desai and Gabe Mansur will handle back-end development, and Rachael Serur and Rebecca Larson will focus on data visualization.

Ideas & Strategies

- Design a map for policymakers
- Will have to do it by state because each data point (row) is a county
- Have a model for each state and predict how the obesity rate (any health disease) would change if they changed X about the state
    - Ex: Would the obesity rate decrease if 20 more grocery stores were added?
- IF TIME (likely not): can we get more granular, for example, if a whole foods specifically was added?
- IF TIME: can we get better data for massachusetts, so we could do it by county?
- Can we improve predictions by augmenting with additional county data and grouping via demographic correlates? (e.g. http://www.statsamerica.org/uscp/) What if we asked for predictions for counties with median incomes > 90K, 80K-89K, etc... ? Or find some other way to group counties together besides the fact that they share a state?

Backend

Database

Schema

Features List:

Variable name	Variable code	Comments
Population, low access to store (%), 2015	PCT_LACCESS_POP15	Demographic (maybe can use in prediction calc?)
Low income & low access to store (%), 2015	PCT_LACCESS_LOWI15	Demographic (maybe can use in prediction calc?)
Households, no car & low access to store (%), 2015	PCT_LACCESS_HHNV15	Demographic (maybe can use in prediction calc?)
SNAP households, low access to store (%), 2015	PCT_LACCESS_SNAP15	Demographic (maybe can use in prediction calc?)
Children, low access to store (%), 2015	PCT_LACCESS_CHILD15	Demographic (maybe can use in prediction calc?)
Seniors, low access to store (%), 2015	PCT_LACCESS_SENIORS15	Demographic (maybe can use in prediction calc?)
White, low access to store (%), 2015	PCT_LACCESS_WHITE15	Demographic (maybe can use in prediction calc?)
Black, low access to store (%), 2015	PCT_LACCESS_BLACK15	Demographic (maybe can use in prediction calc?)
Hispanic ethnicity, low access to store (%), 2015	PCT_LACCESS_HISP15	Demographic (maybe can use in prediction calc?)
Asian, low access to store (%), 2015	PCT_LACCESS_NHASIAN15	Demographic (maybe can use in prediction calc?)
American Indian or Alaska Native, low access to store (%), 2015	PCT_LACCESS_NHNA15	Demographic (maybe can use in prediction calc?)
Hawaiian or Pacific Islander, low access to store (%), 2015	PCT_LACCESS_NHPI15	Demographic (maybe can use in prediction calc?)
Multiracial, low access to store (%), 2015	PCT_LACCESS_MULTIR15	Demographic (maybe can use in prediction calc?)

Grocery stores/1,000 pop, 2014	GROCPTH14	
Supercenters & club stores/1,000 pop, 2014	SUPERCPTH14	
Convenience stores/1,000 pop, 2014	CONVSPTH14	
Specialized food stores/1,000 pop, 2014	SPECSPTH14	
SNAP-authorized stores/1,000 pop, 2016	SNAPSPTH16	
Fast-food restaurants/1,000 pop, 2014	FFRPTH14	
Full-service restaurants/1,000 pop, 2014	FSRPTH14	
SNAP participants (% pop), 2016*	PCT_SNAP16	Demographic (maybe can use in prediction calc?)
National School Lunch Program participants (% pop), 2015*	PCT_NSLP15	Demographic (maybe can use in prediction calc?)
School Breakfast Program participants (% pop), 2015*	PCT_SBP15	Demographic (maybe can use in prediction calc?)
~~Summer Food Program participants (change % pop), 2009-15*~~ Summer Food Service Program participants (% pop), 2015*	~~PCH_SFSP_09_15~~ PCT_SFSP15	Demographic (maybe can use in prediction calc?)
WIC participants (% pop), 2015*	PCT_WIC15	Demographic (maybe can use in prediction calc?)
Soda sales tax, retail stores, 2014*	SODATAX_STORES14	
Soda sales tax, vending, 2014*	SODATAX_VENDM14	
Chip & pretzel sales tax, retail stores, 2014*	CHIPSTAX_STORES14	
Chip & pretzel sales tax, vending, 2014*	CHIPSTAX_VENDM14	
General food sales tax, retail stores, 2014*	FOOD_TAX14	
Farmers' markets/1,000 pop, 2016	FMRKTPTH16	
Farmers' markets that report accepting SNAP (%), 2016	PCT_FMRKT_SNAP16	
Farmers' markets that report accepting WIC (%), 2016	PCT_FMRKT_WIC16	
Farmers' markets that report accepting WIC Cash (%), 2016	PCT_FMRKT_WICCASH16	
Farmers' markets that report accepting SFMNP (%), 2016	PCT_FMRKT_SFMNP16	
Farmers' markets that report accepting credit cards (%), 2016	PCT_FMRKT_CREDIT16	
Farmers' markets that report selling fruit & vegetables (%), 2016	PCT_FMRKT_FRVEG16	
Farmers' markets that report selling animal products (%), 2016	PCT_FMRKT_ANMLPROD16	
Farmers' markets that report selling baked/prepared food products (%), 2016	PCT_FMRKT_BAKED16	
Farmers' markets that report selling other food	PCT_FMRKT_OTHERFOOD16	

products (%), 2016		
Food hubs, 2016	FOODHUB16	
Adult diabetes rate, 2013	PCT_DIABETES_ADULTS13	Response
Adult obesity rate, 2013	PCT_OBESE_ADULTS13	Response
High schoolers physically active (%), 2015*	PCT_HSPA15	
Recreation & fitness facilities/1,000 pop, 2014	RECFACPTH14	
% White, 2010	PCT_NHWHITE10	Demographic
% Black, 2010	PCT_NHBLACK10	Demographic
% Hispanic, 2010	PCT_HISP10	Demographic
% Asian, 2010	PCT_NHASIAN10	Demographic
% American Indian or Alaska Native, 2010	PCT_NHNA10	Demographic
% Hawaiian or Pacific Islander, 2010	PCT_NHPI10	Demographic
% Population 65 years or older, 2010	PCT_65OLDER10	Demographic
% Population under age 18, 2010	PCT_18YOUNGER10	Demographic
Median household income, 2015	MEDHHINC15	Demographic
Poverty rate, 2015	POVRATE15	Demographic
Metro/nonmetro counties, 2010	METRO13	Demographic
FIPS		Key
State		Key
County		Key

Data to get separately:

- CSA farms per county
- ***Current diabetes and obesity rates per county
  - By state obesity: https://stateofobesity.org/states/ma/
  - By county diabetes 2013: https://www.cdc.gov/diabetes/atlas/countydata/atlas.html?detectflash=false
- More recent population data - race, income (current is from 2010 census)
  - https://statisticalatlas.com/state/Massachusetts/Race-and-Ethnicity#data-map/county
  - https://statisticalatlas.com/state/Massachusetts/Household-Income#data-map/county

Extra Data:

- Food stamps by county in MA: https://statisticalatlas.com/state/Massachusetts/Food-Stamps
- Population by county in MA: https://statisticalatlas.com/state/Massachusetts/Population

TODO/Current Status (updated 12/17/18)

(Trello Board tracking progress)

Finished:

- [X] Presentation script
- [X] Connect prediction calculator to data & backend

- [X] Connect "Deep Dive" visualizations to map
- [X] "Storytelling" prose & frontend polish
- [X] Try to strengthen prediction model per recommended techniques from TAs
- [X] Use Tableau to illustrate and build out the model
- [X] Build (code) prediction calculator frontend
- [X] Code out and populate the US map with data
- [X] Explore Food Environment Atlas, USDA branded foods, and what we eat in America datasets - brainstorm predictors
- [X] Finalize and organize predictors, datasets
- [X] Scaffold basic US map (per county) in D3
- [X] Plan/design D3 visualizations based on data findings
- [X] Evaluate project proposal options: food distribution vs. MBTA
- [X] Upload data to GH repo
- [X] Organize Project Plan notes from meeting and add to GH repo
- [X] Submit Project Plan on Canvas
- [X] Scaffold basic web application

☑ Presentation recording

☑ README

Database Schema

Table "food_atlas"

Column	Type	Collation	Nullable	Default
id	integer		not null	nextval('food_atlas_id_seq'::regclass)
fips	integer		not null	
county	character varying(255)		not null	
state	character varying(255)		not null	
pct_laccess_pop15	double precision		not null	
pct_laccess_lowi15	double precision		not null	
pct_laccess_hhnv15	double precision		not null	
pct_laccess_snap15	double precision		not null	
pct_laccess_child15	double precision		not null	
pct_laccess_seniors15	double precision		not null	
pct_laccess_white15	double precision		not null	
pct_laccess_black15	double precision		not null	
pct_laccess_hisp15	double precision		not null	
pct_laccess_nhasian15	double precision		not null	
pct_laccess_nhna15	double precision		not null	
pct_laccess_nhpi15	double precision		not null	
pct_laccess_multir15	double precision		not null	
grocpth14	double precision		not null	

supercpth14	double precision		not null	
convspth14	double precision		not null	
specspth14	double precision		not null	
snapspth16	double precision		not null	
ffrpth14	double precision		not null	
fsrpth14	double precision		not null	
pct_snap16	double precision		not null	
pct_nslp15	double precision		not null	
pct_sbp15	double precision		not null	
pch_sfsp_09_15	double precision		not null	
pct_wic15	double precision		not null	
sodatax_stores14	double precision		not null	
sodatax_vendm14	double precision		not null	
chipstax_stores14	double precision		not null	
chipstax_vendm14	double precision		not null	
food_tax14	double precision		not null	
fmrktpth16	double precision		not null	
pct_fmrkt_snap16	double precision		not null	
pct_fmrkt_wic16	double precision		not null	
pct_fmrkt_wiccash16	double precision		not null	
pct_fmrkt_sfmnp16	double precision		not null	
pct_fmrkt_credit16	double precision		not null	
pct_fmrkt_frveg16	double precision		not null	
pct_fmrkt_anmlprod16	double precision		not null	
pct_fmrkt_baked16	double precision		not null	
pct_fmrkt_otherfood16	double precision		not null	
foodhub16	integer		not null	
pct_diabetes_adults13	double precision		not null	
pct_obese_adults13	double precision		not null	
pct_hspa15	double precision		not null	
recfacpth14	double precision		not null	
pct_nhwhite10	double precision		not null	
pct_nhblack10	double precision		not null	
pct_hisp10	double precision		not null	
pct_nhasian10	double precision		not null	
pct_nhna10	double precision		not null	

pct_nhpi10	double precision		not null	
pct_65older10	double precision		not null	
pct_18younger10	double precision		not null	
medhhinc15	double precision		not null	
povrate15	double precision		not null	
metro13	boolean		not null	

Frontend & Visualizations

- *US Map*: Takes d3-geo map county and uses this to render choropleths based on selected Food Atlas data topics. This will be connected to the prediction calculator and deep dive graphs. As time permits, we may also explore the data as a bubble map using the same map object code.
- *Fast Food per 1000 & Obesity*: This is a scatterplot looking at the connection of the number of fast food restaurants per 1000/pop and percent of the population with low access to food (radius) to obesity. This chart will also be iterated on to include visual distinctions between metro and non metro counties. The goal is to give the user an understanding of demographics of a county that may influence the presence of fast food restaurants, and how that impacts the county's health.
- *Farmer's Markets per 1000 & Obesity:* This is a scatterplot looking at the number of Farmer's Markets per 1000/pop and obesity. The size of the circle radius reflects the percent of the population with low access to food sources.
- *Food Tax Rate & Obesity:* This is a scatterplot looking at the food tax rate and obesity. The size of the circle radius reflects the percent of the population with low access to food sources.

Presentation

Screencast Script

Overview:

Our project aims to derive insights about food access and attributes across the US by evaluating how those factors affect the health of a community. Our team is Tushal Desai, Rebecca Larson, Gabe Mansur, and Rachael Serur. We worked to solve the question, 'How can policy makers introduce policies that are the most effective at bettering the health of a county?' This project is designed for policy makers to better understand what changes could make better the health of their county. There are many factors that influence the health of a community -- ranging from culture, income, food access, healthy food options, welfare programs, etc. Each county differs across these variables and therefore changing certain variables (for example, adding more grocery stores or reducing the number of fast food restaurants) affect the health of the county uniquely from others. Our project allows policy makers to see a visual summary of the US according to these variables, both through a full map of all counties and three more linked visualizations that show deeper detail on specific veriables. The user can make further discoveries on a narrowed county level. By clicking on a county, the policy maker can use our Obesity Prediction Calculator to test how changing a couple variables can influence the health of that county. The three linked visualizations also update their context to show the policy maker only data from that county's state. This application allows policy makers comprehensive insight into the trends and specifics of the country and each county and provides guidance in policies they should target to best help a specific county.

Process:

To accomplish our goal, we used data from the USDA Food Environment Atlas. This dataset contains data at the county level and has many features, including access to restaurants, grocery stores, farmers markets and welfare programs, as well as demographic and health data.

- looking through data
- EDA
- narrowing our scope and defining goal

- feature engineering and model creation
- map creation
- deep dive visualization creations
- prediction calculator creation
- putting it all together

Demo:

- The different variables of the map
- Tooltips on counties
- The bottom visualizations
  - Brushing the first graphs filters the rest
  - tooltips
- Clicking on a county
- Use the prediction calculator
- See how the bottom visualizations filtered and the last one changed to be better suited to the data
- Reset the visualizations

Challenges:

One of the most immediate challenges in this project was our dataset. After selecting features, we found that about half of our data had at least one null value in each county, so the entire county had to be dropped from the prediction model. Having all of the county data would have made our model more accurate and could show the user more accurate trends. Another challenge was figuring out at what level our model should predict. We originally wanted to have separate models for each state with the understanding that there were many cultural factors not captured in the data, and were hoping to account for this by narrowing our focus to a state level. However, we only had county level data and most states have too few counties to build an accurate model around.

- there weren't really trends among features and response variable when did EDA, all a large bunch

Future Questions / Next Steps:

This first iteration of the Food Access Project helped bring the data to life and raise other questions that could be potentially answered by seeking ways to augment or further explore county data, such as:

- What other measures of health could we explore with this data? Exercise habits, diabetes, access to community recreational facilities or gyms?
- Can we drill into mediating factors that paint a clearer picture between obesity rates, farmer's markets, and fast food? Demographics such as socioeconomic status, race/ethnicity, and education levels are likely to play interesting roles.
- When socioeconomic status does play a role, does increased access to food assistance programs at farmer's markets and grocery stores adequately help health rates in a community? The Food Atlas has data on Supplemental Nutrition Assistance Program (SNAP) acceptance, school programs, and Women, Infants, & Children (WIC) services that could shed light on program effectiveness and varying levels of access country-wide.
- Are there datasets on legislature that could help further explain trends and help policymakers make data-driven decisions?

Conclusion:

Overall, this project provides a good jumping off point to further investigate this idea of county-specific food policy. To have better accuracy, one would need more recent data with more relevant variables. The inaccuracy of this obesity prediction calculator also indicates that there are likely many factors that influence the health of a community that can't be perfectly captured by a number, such as culture, reinforcing that policy makers should not rely solely on data to introduce policy into a community. They should incorporate the voice of the community to have a fuller understanding and work with the community to find effective solutions.