Working Title: U.S. Food Environment Analysis

The primary project goal is to examine the relationships between food insecurity and health indicators like diabetes and obesity, and to forecast trends.

Additional relationships that will be examined are those between food insecurity and use of SNAP benefits, food access, diabetes and obesity rates, recreation and fitness facilities, and socioeconomic factors such as income, race, age, and metro/ non-metro areas.

Data Source

Dataset is the USDA Food Environment Atlas which includes the following components:

- Variable list of metadata about all the variables mapped in the atlas
- Nine spreadsheets with data for each of the atlas categories, which address three categories of community food environment factors: food choices, health and well-being, and community characteristics.
- Supplemental state and county data of population estimates used to calculate atlas data.

Data was obtained from a high-quality, reliable governmental source, the Economic Research Service of the U.S. Department of Agriculture.

The USDA Food Environment Atlas contains data aggregated from collaboration with reliable governmental, nonprofit, and academic entities. The Centers for Disease Control and Prevention provided the statistics on obesity, diabetes, and physical activity; the U.S. Census Bureau provided indicators on recreation centers and businesses in **County Business Patterns**; USDA's Agricultural Marketing Service provided indicators on farmers' markets and food hubs; USDA's Food and Nutrition Service provided information on State-level food and nutrition assistance program participation rates and farm to school activities. The information on State beverage and snack taxes are from the **Bridging the Gap Program**, University of Illinois at Chicago. The information on food banks are from **Feeding America's** nationwide network of food banks.

The dataset is available publicly for download. The current version was posted in 2020 and there are six archived versions that are also available. There were two main objectives:

- To gather statistics for research into the relationship between food environment indicators and diet quality
- To provide insight into community access to healthy food.

I have chosen this data source for the assurance of its quality and the depth of information it provides. The USDA has taken an in-depth approach to assembling statistics on the myriad of factors at play in the food environments of communities across the U.S. I want to use it exactly for what it was intended: to facilitate closer examination of the interconnection of these factors.

Data Profile

Data Collection	The dataset contains a mixture of survey, usage, and administrative data collected by the entities outlined above.

Data Contents	Excluding the supplemental data for counties and states, the nine spreadsheets had 282 discrete columns. I eliminated data prior to 2013 to include only date from the last ten years in the project. The exception to this is the data in the SOCIOECONOMIC table, which has been used with the Supplemental County Data to establish estimated demographic percentages for the years between the 2010 and 2020 Census counts. The five spreadsheets ACCESS, ASSISTANCE, HEALTH, INSECURITY, and SOCIOECONOMIC contain aggregate data from the other sheets, and will be the focus of the project. I combined the 34 remaining columns from these five spreadsheets into one table called Merged Data, which contains the columns indicated in Table 1. The tables Supplemental County Data and Merged Data will be the two datasets for the project. I converted them from Excel files into CSV to work in Python.
Data Limitations	The calculations in the SOCIOECONOMIC table are from the 2010 U.S. Census count, and the numbers in the Supplemental Data - County table are based on the estimates as they were calculated for years between the Census counts. Calculations in these tables utilized the estimated population counts. Eliminating data from before 2013 means that I may not have enough data for forecasting or for temporal analyses requiring multiple years.
Data Ethics	There is the potential for bias in the interpretation or presentation of results of this project, particularly in examining and presenting demographic relationships that may show correlation to health outcomes. I will be vigilant in remaining aware of and accounting for bias in the presentation of key findings.

Data Consistency Checks and Cleaning

1. Merged Data dataset

Missing Values	Missing values were found that represented 20 counties with missing information about access to food sources. One county was missing data for adult diabetes rate, and 4 counties were missing information on median income. The USDA's description of the data indicates that "Data that were not available, not applicable, or suppressed for specific counties are denoted with a blank cell or -9999." Based on this explanation, I used the assumption that the values for the data in these counties are not "0" but missing for other reasons, possibly including suppression for privacy concerns. There were also 9 counties with "0" values listed for all access data. Based on the Supplemental County Data from the estimated population counts and the data on SNAP benefit recipients in these counties, I used the assumption that the values for these counties were missing rather than "0."

	Therefore, I chose to impute values for these counties rather than eliminate them or change the values to "0." Because of the high standard deviations across the columns for this dataset, I chose to use imputation with median values to fill in the missing values. I imputed median values of the state in which the county was located. I used this same imputation method to replace the "0" values in the additional 9 counties. Because of the relatively small size of the dataset and the large number of imputed values, I completed the other cleaning and data consistency tasks listed below in Python, exported the cleaned data file, and then did the imputation in Excel.
Renaming columns	Columns were renamed for brevity and clarity. The original and replaced names are listed in Table 1.
Data types	I changed the datatype of the FIPS column from "int64"to "object" because it is a code identifier.
Duplicate records	No duplicates found

2. Supplemental County Data dataset

Missing values	No missing values were found. However, there was one fewer row in the Supplemental County Data (3142 rows) than in the Merged Data files (3143 rows).

Renaming columns	I renamed the columns to follow the Merged Data naming conventions, for brevity and clarity:
	
	`Population_Estimate 2013 to pop_13`
	`Population_Estimate 2014  pop_14`
	`Population_Estimate 2015  pop_15`
	`Population_Estimate 2016  pop_16`
	`Population_Estimate 2017  pop_17`
	`Population_Estimate 2015  pop_18`

Data types	I changed the datatype of the FIPS column to "object" from "int64" because it is a code used as an index identifier. I also changed each of the Population Estimate columns to "int64" datatypes from "object".
	
	`pop_13  int64`
	`pop_14  int64`
	`pop_15  int64`
	`pop_17  int64`
	`pop_16  int64`
	`pop_18  int64`

Duplicate records	No duplicates found

Project Questions

- How does food insecurity relate to the health outcomes of diabetes and obesity?
- How does food insecurity affect different segments of the population by race and age?
- How does access to food sources relate to food insecurity?
- How does median income relate to food insecurity?
- How does the eligibility for and usage of federal food benefit programs vary by geography in the U.S.?
- Are areas with more vulnerable populations (children, seniors) accessing federal food benefits for which they are eligible?

Main hypotheses:

- If food is difficult to access, rates of food insecurity increase.
- If food insecurity increases, rates of diabetes and obesity also increase.
- If a county has a high rate of food insecurity, they will also have high percentages of people with diabetes and obesity.

Table 1. Merged Data

Original Table	Variable Explanation	Original Variable Name	New Variable Name	Type of Variable	Nominal/ Ordinal/Binary
ACCESS	Federal Information Processing System (FIPS) Codes for States and Counties	FIPS	FIPS	Categorical	Nominal
ACCESS	State Abbreviation	State	State	Categorical	
ACCESS	County	County	County	Categorical	
ACCESS	Population, low access to store, 2015	LACCESS_POP15	laccess	Continuous	Nominal
ACCESS	Population, low access to store (%), 2015	PCT_LACCESS_POP15	pct_laccess	Continuous	Nominal
ACCESS	Low income & low access to store, 2015	LACCESS_LOWI15	lowinc_laccess	Continuous	Nominal
ACCESS	Low income & low access to store (%), 2015	PCT_LACCESS_LOWI15	pct_lowinc_laccess	Continuous	Nominal
ACCESS	SNAP households, low access to store, 2015	LACCESS_SNAP15	snap_laccess	Continuous	Nominal
ACCESS	SNAP households, low access to store (%), 2015	PCT_LACCESS_SNAP15	pct_snap_laccess	Continuous	Nominal
ACCESS	Children, low access to store, 2015	LACCESS_CHILD15	child_laccess	Continuous	Nominal
ACCESS	Children, low access to store (%), 2015	PCT_LACCESS_CHILD15	pct_child_laccess	Continuous	Nominal
ACCESS	Seniors, low access to store, 2015	LACCESS_SENIORS15	senior_laccess	Continuous	Nominal
ACCESS	Seniors, low access to store (%), 2015	PCT_LACCESS_SENIORS15	pct_senior_laccess	Continuous	Nominal

ACCESS	White, low access to store, 2015	LACCESS_WHITE15	white_laccess	Continuous	Nominal
ACCESS	White, low access to store (%), 2015	PCT_LACCESS_WHITE15	pct_white_laccess	Continuous	Nominal
ACCESS	Black, low access to store, 2015	LACCESS_BLACK15	black_laccess	Continuous	Nominal
ACCESS	Black, low access to store (%), 2015	PCT_LACCESS_BLACK15	pct_black_laccess	Continuous	Nominal
ACCESS	Hispanic ethnicity, low access to store, 2015	LACCESS_HISP15	hisp_laccess	Continuous	Nominal
ACCESS	Hispanic ethnicity, low access to store (%), 2015	PCT_LACCESS_HISP15	pct_hisp_laccess	Continuous	Nominal
ACCESS	Asian, low access to store, 2015	LACCESS_NHASIAN15	asian_laccess	Continuous	Nominal
ACCESS	Asian, low access to store (%), 2015	PCT_LACCESS_NHASIAN15	pct_asian_laccess	Continuous	Nominal
ACCESS	American Indian or Alaska Native, low access to store, 2015	LACCESS_NHNA15	natamer_laccess	Continuous	Nominal
ACCESS	American Indian or Alaska Native, low access to store (%), 2015	PCT_LACCESS_NHNA15	pct_natamer_laccess	Continuous	Nominal
ACCESS	Hawaiian or Pacific Islander, low access to store, 2015	LACCESS_NHPI15	pacific_laccess	Continuous	Nominal
ACCESS	Hawaiian or Pacific Islander, low access to store (%), 2015	PCT_LACCESS_NHPI15	pct_pacific_laccess	Continuous	Nominal
ACCESS	Multiracial, low access to store, 2015	LACCESS_MULTIR15	multi_laccess	Continuous	Nominal
ACCESS	Multiracial, low access to store (%), 2015	PCT_LACCESS_MULTIR15	pct_multi_laccess	Continuous	Nominal
ASSISTANCE	SNAP participants (% pop), 2017*	PCT_SNAP17	pct_snap	Continuous	Nominal
ASSISTANCE	SNAP participants (% eligible pop), 2016*	SNAP_PART_RATE16	pct_snap_participation	Continuous	Nominal
HEALTH	Adult diabetes rate, 2013	PCT_DIABETES_ADULTS13	pct_adult_diabetes	Continuous	Nominal
HEALTH	Adult obesity rate, 2017*	PCT_OBESE_ADULTS17	pct_adult_obese	Continuous	Nominal
INSECURITY	Household food insecurity (%, three-year average), 2015-17*	FOODINSEC_15_17	pct_food_insec	Continuous	Nominal
INSECURITY	Household very low food security (%, three-year average), 2015-17*	VLFOODSEC_15_17	pct_vlow_foodsecure	Continuous	Nominal
SOCIOECONOMIC	Median household income, 2015	MEDHHINC15	med_income	Continuous	Nominal

Table 2. Supplemental County Data

Original Excel Table	Variable Explanation	Original Variable Name	New Variable Name	Type of Variable	Nominal/ Ordinal/Binary

Supplemental Data - County	Federal Information Processing System (FIPS) Codes for States and Counties	FIPS	FIPS	Categorical	Nominal
Supplemental Data - County	County name	County	County	Categorical	
Supplemental Data - County	State name	State	State	Categorical	
Supplemental Data - County	Population Estimate 2013	Population_Estimate_2013	pop_13	Continuous	Nominal
Supplemental Data - County	Population Estimate 2014	Population_Estimate_2014	pop_14	Continuous	Nominal
Supplemental Data - County	Population Estimate 2015	Population_Estimate_2015	pop_15	Continuous	Nominal
Supplemental Data - County	Population Estimate 2016	Population_Estimate_2016	pop_16	Continuous	Nominal
Supplemental Data - County	Population Estimate 2017	Population_Estimate_2017	pop_17	Continuous	Nominal
Supplemental Data - County	Population Estimate 2018	Population_Estimate_2018	pop_18	Continuous	Nominal