DS 5100 Project: Food Deserts

Max Jones, Kaia Lindberg, Grace Lyons, Mani Shanmugavel
December 2021

Food Deserts

- In this project, we wished to examine the relationships of food deserts across the US

- Food deserts: geographic areas where residents have few to no convenient options for securing affordable and healthy foods — especially fresh fruits and vegetables

Research Question 1

<u>Where do food deserts occur?</u>

To answer, we wished to compare the occurrence of food deserts to the following:

- Population
- Metro Status (city vs non-city)
- Region (Northeast, Midwest, West, South)
- Ethnicity

Research Question 2

How could our project be used to analyze specific areas?

To answer, we will demonstrate our project with an analysis into a single state.

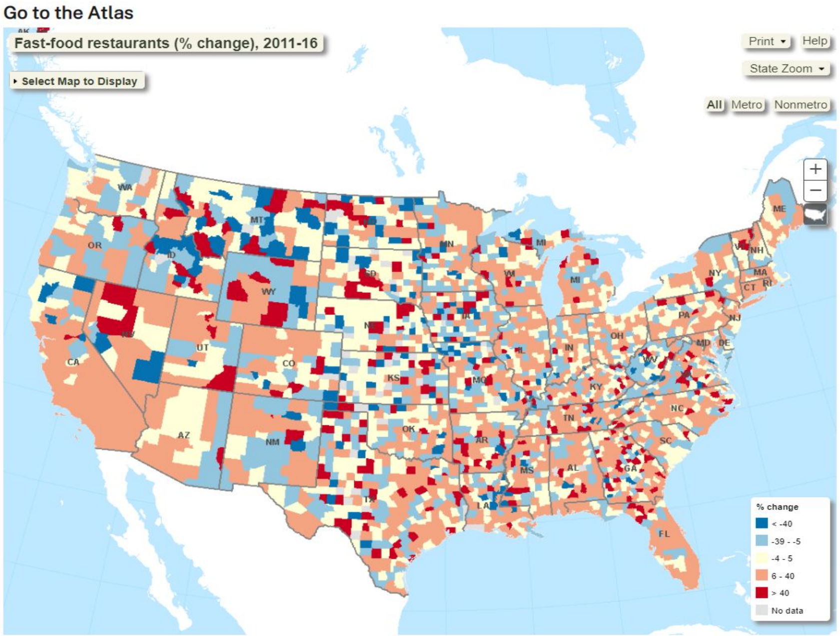
Project Scenario

--Data set source: USDA Food Environment Atlas

--Contains various statistics for each county in the US

--Over 250 variables, including:

- Food Choices
  - Food store accessibility (access/proximity)
  - Type of stores
  - Availability of fresh and local foods
- Health and Wellbeing
  - Obesity and diabetes rates
  - Recreation and physical activity levels
- Community characteristics
  - Metro/non-metro
  - Ethnicity percentages
  - Income and poverty



Variables of Interest

- 2015 Census Population
- Metro Status
- Region
- Ethnicity Percentages

- <u>Percentage of low income, low access to food</u>

  --Variable that measures what percent of the population is classified as having a low income and low access to food

  --Used this to measure food deserts

Project Pipeline

Acquisition

Source data from USDA, load into Python, preliminary analysis, evaluate quality

Cleaning

Transform data, clean columns, and address inconsistencies in the data

Analysis

Perform data analysis to process the data and identify variables of interest

Results

Create visualizations to make meaning of and communicate results

Future Work...

Data Cleaning

Wrote a class with methods for each data cleaning step to improve readability and facilitate testing

	FIPS	State	County	Variable_Code	Value
0	1001	AL	Autauga	LACCESS_POP10	18428.439690
1	1001	AL	Autauga	LACCESS_POP15	17496.693040
2	1001	AL	Autauga	PCH_LACCESS_POP_10_15	-5.056026
3	1001	AL	Autauga	PCT_LACCESS_POP10	33.769657
4	1001	AL	Autauga	PCT_LACCESS_POP15	32.062255

1. Clean state column to handle extra spaces — “VA” versus “ VA”
2. Webscrape FIPS lookup table to resolve inconsistencies in county names — “Raleigh” versus “Raleigh County”
3. Reformat data so 1 row for each FIPS code and each variable is a column
4. Split county and state level data into two separate dataframes

Data Cleaning

Wrote a class with methods for each data cleaning step to improve readability and facilitate testing

	FIPS	State	County	Variable_Code	Value
0	1001	AL	Autauga	LACCESS_POP10	18428.439690
1	1001	AL	Autauga	LACCESS_POP15	17496.693040
2	1001	AL	Autauga	PCH_LACCESS_POP_10_15	-5.056026
3	1001	AL	Autauga	PCT_LACCESS_POP10	33.769657
4	1001	AL	Autauga	PCT_LACCESS_POP15	32.062255

	FIPS	State	County	2010_Census_Population	AGRITRSM_OPS07	AGRITRSM_OPS12	AGRITRSM_RCT07	AGRITRSM_RCT12
0	1001	AL	Autauga	54571.0	7.0	10.0	228000.0	146000.0
1	1003	AL	Baldwin	182265.0	18.0	16.0	124000.0	204000.0
2	1005	AL	Barbour	27457.0	27.0	32.0	163000.0	304000.0
3	1007	AL	Bibb	22915.0	5.0	6.0	NaN	21000.0

	FIPS	State	County	CACFP_PART_2012	CACFP_PART_2013	CACFP_PART_2014	CACFP_PART_2015	CACFP_PART_2016
0	1	AL	Total	42903.75	41862.50	45038.75	51263.75	59447.50
1	2	AK	Total	10540.75	10962.75	11429.50	10786.50	10039.00
2	4	AZ	Total	41670.75	42945.25	40344.00	42551.50	42618.00

Data Analysis

Examples

1. Filtering down to a single state
2. Remove missing values based on threshold
3. User interaction to make selections for analysis
4. Calculate correlations between variables of interest
5. Keep only the most recent column
6. Remove “hidden” state level variables

Varia

```
Please select a variable of interest:
PCT_DIABETES_ADULTS13
PCT_DI
```

```
df_county.groupby('State')[['PCT_OBESE_ADULTS17']].std()
```

Popu

```
PCH_VEG_ACRESPTH_07_12      0.464768
WICSPTH16                   0.450395
```

	PCT_OBESE_ADULTS17
State	
AK	0.0
AL	0.0

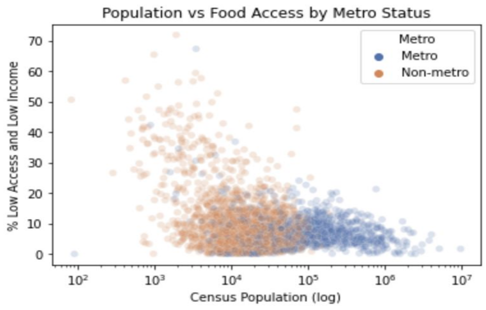
Testing

- Developed Classes to clean and analyze data
- Thorough unit testing of cleaning and analysis class using assertEqual()
- Used test fixtures like setup() to set up the test cases
- Data gathered for State and County FIPS codes through Web scraping was also tested

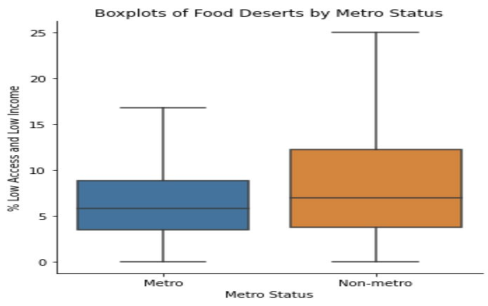
```python
#@title Default title text
# Create class that inherits from unittest.TestCase
class DataCleaning_Test(unittest.TestCase):
  def setUp(self):
    # Set up
    # Create class with small dataset to use for testing
    sample_df = pd.DataFrame({'FIPS':[1,1,20,20,1001,1001,99999],
                              'State': ['AL','AL','KS','KS ', 'AL', 'AL ', 'ZZ'],
                              'County': ['Total','Total','Total','Total','Autauga','Autauga County','Fake'],
                              'Variable_Code':['Var1', 'Var2', 'Var1', 'Var2', 'Var1', 'Var2', 'Var1'],
                              'Value': [5,np.nan,10,np.nan,np.nan,20,np.nan]})
    self.clean1 = DataCleaning(sample_df)

  def test_clean_state_col(self):
    # Is any white space removed from states?
    # Clean state column
    self.clean1.clean_state_column()
    # Test: using assetEqual() method
    self.assertEqual(list(self.clean1.df.State), ['AL','AL','KS','KS', 'AL', 'AL', 'ZZ'])

  def test_state_name_col(self):
    # Is the state code correct based on FIPS code?
    self.clean1.full_data_cleaning()
    df = self.clean1.df_state
    result = df[df['FIPS'] == 20].State.to_string(index=False)
    # Test: using assetEqual() method
    self.assertEqual(result.strip(), 'KS')

  def test_county_name_col(self):
    # Is the County name correct based on FIPS code?
    self.clean1.full_data_cleaning()
    df = self.clean1.df_county
    result = df[df['FIPS'] == 1001].County.to_string(index=False)
    print(result)
    # Test: using assetEqual() method
    self.assertEqual(result.strip(), 'Autauga')
```

Food Deserts are more common in...

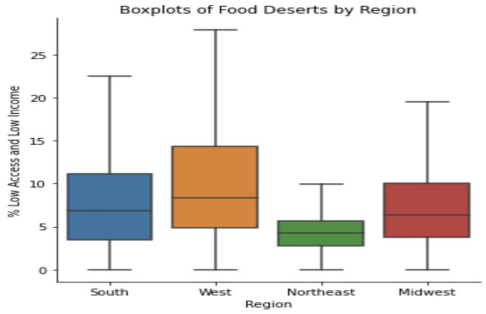
Low Population



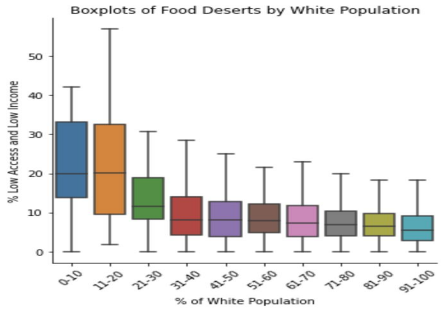
Non-metro Areas



West Region

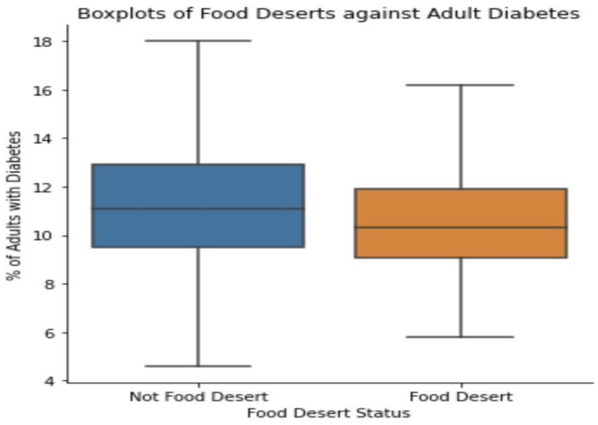


Areas with Lower White Population

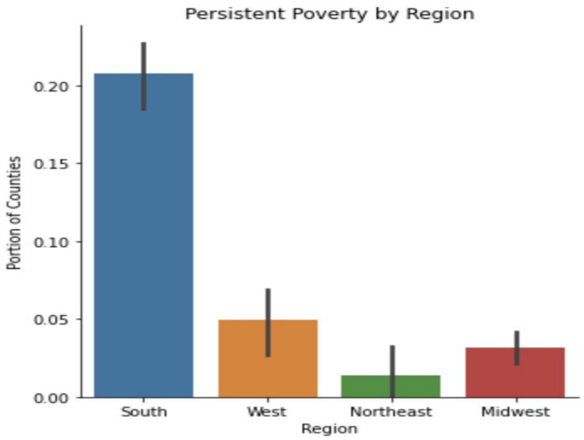


Food Deserts and other Variables

Diabetes Rate

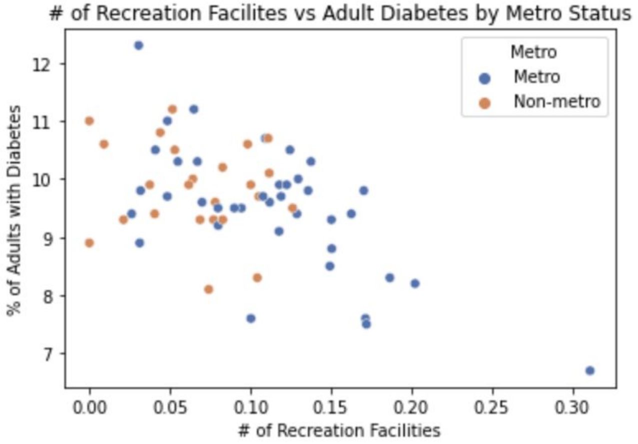


Poverty Persistence Rate

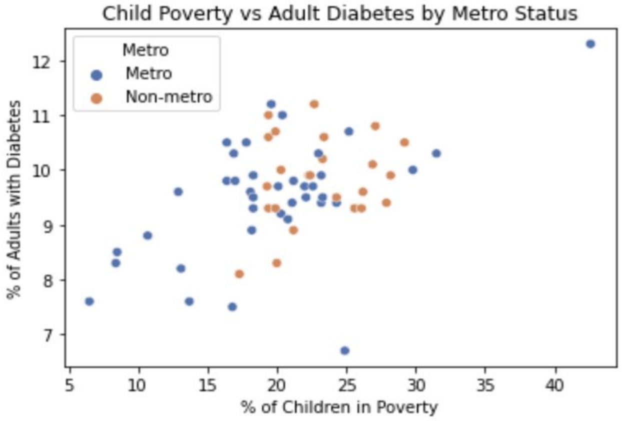


We've also analyzed relationships between variables within a single state

of Recreation Facilities vs Adult Diabetes



Child Poverty vs Adult Diabetes



Conclusions--Question 1

<u>Where do food deserts occur?</u>

- Small population
- Non-metro
- West, Midwest, and South regions
- Low White population

Conclusions--Question 2

How could our project be used to analyze specific areas?

For New York:

- Negative correlation between adult diabetes rate and number of recreational facilities
- Positive correlation between child poverty rate and adult diabetes rate

Significance of Results

In areas we determined to be food deserts, one could:

- Incentivize grocery stores and supermarkets
- Fund city-wide programs to encourage healthier eating
- Extend support for small, corner-type stores and neighborhood-based farmers markets

For Further Investigation

- Additional research into all 250 variables
- Machine learning model to more accurately predict food desert locations
- Perhaps obtain more accurate food desert variable that emphasized healthy food over all food

Works Cited

www.aecf.org/blog/exploring-americas-food-deserts

www.ers.usda.gov/data-products/food-environment-atlas/

https://www.ers.usda.gov/webdocs/publications/45014/30940_err140.pdf