Project 3: Solar Equity

Introduction

This ***"Solar Panel Installation"*** Project explores energy injustice by determining whether there exist statistically significant differences in the rooftop solar adoption between racial/ethnic groups. There is a long history of energy injustice in the United States. 75% of African American live within 30 miles of coal-fired power plants, a radius within which the health effects of the emissions are most harmful. It is important to recognize that injustice exists not only when a group is marginalized with uneven burdens, such as exposure to pollution, but also when a group is excluded from benefits. Rooftop solar allows the household to receive a variety of benefits such as tax credits, rebates, lower cost of electricity, etc. This analysis was performed at the national level in ***"Disparities in Rooftop Photovoltaic Deployment in the United States by Race and Ethnicity"*** by Sunter, Castellanos, and Kammen (ref). For this module, you will be using the census data from the *American Community Survey* and rooftop solar data from *Google Project Sunroof*. While these datasets provide rich nation-wide data, for this first part, you will specifically analyze the state of California

Project Learning Objectives

1. Learning how to apply non-linear regression techniques
    1. Polynomial Regression
    2. Local Regression
2. Recognizing advantages and disadvantages of the different non-linear regression techniques, with particular emphasis on handling outliers and leverage points.
3. Introducing uncertainty methods, specifically bootstrapping.
4. Learning how to justify your choices regarding which data science techniques to use.
5. Learning how to justify feature engineering decisions.
6. Learning how to draw justifiable conclusions based on your analysis and communicate the limitations of the analysis.

Additional Resources

To assist in understanding the project and completing the tasks, here are some additional resources:

- "Disparities in rooftop photovoltaics deployment in the United States by race and ethnicity" by Deborah Sunter, Sergio Castellanos, and Daniel Kammen
- Google Project Sunroof site: https://sunroof.withgoogle.com/data-explorer/
- United States Census Bureau "2015 Planning Database"
- Bootstrapping data analysis technique (10-min video)

- Local Regression (LOWESS) data analysis technique (10-min video)

Day 1: Overview

After an introduction, your coding tasks today are to:

- Read in and prepare the data
- Perform a polynomial regression
- Perform bootstrapping to determine confidence intervals

This analysis example of processing and analyzing data will be done on a subset of the national dataset looking specifically at the state of California.

Task 1: Data Preparation

As detailed in the article ("*Disparities in Rooftop Solar Deployment in the United States by Race and Ethnicity*" by Sunter, Kammen and Castellanos) there are several steps of data pre-processing and preparation needed prior to analysis.

Task 1.1: User Inputs

Majority Level

The majority level sets the percentage of the census tract population that must self-identify as the same race/ethnicity in order for the census tract to be classified as that race/ethnicity. We have defined *majority* as (over) 50% and *strong-majority* as (over) 75%.

```
majority_level=50;
```

Variable of Interest: Median Household Income (Med_HHD_Inc_ACS_09_13)

For the analysis, we will initially explore racial disparity in rooftop PV deployment as a function of **Median Household Income** and the percentage of households occupied by renters. If you would like to generate plots with the median household income along the x-axis (similar to Figures 2, 5, S1, and S5 from the paper), use 'median income' as your variable of interest.

```
variable_of_interest='median_income';
```

Add both the ***Majority Level*** and ***Variable of Interest*** lines to your code.

Task 1.2: Load Data

The original data for this project can be found at these sources:

- https://www.google.com/get/sunroof/data-explorer/
- https://www.census.gov/data.html

The data for this project has been uploaded to Canvas as `CensusSunroofMerged.csv`. The data provided is the same data used for the publication in *Nature Sustainability*. It was formed by merging data from these two sources by matching census tracts to create this new file. Within ProjectSolarV1.m is a function to load the solar data: solarLoadData().

```
data= readtable('CensusSunroofMerged.csv');
```

Task 1.3: Look at Data

Now that data have been loaded, look at the data (confirm it loaded correctly) and answer the question after the IDE.

Note that the data are loaded as a table, rather than a matrix. You can learn more about tables on the MathWorks web site.

Question 1.1: How many columns are there in our dataset? How many rows of data are there in our dataset?

Question 1.2: How are tables different from matrices? How are they similar? When might it be preferable to use tables rather than matrices?

Task 1.4: Clean the Data

We are going to do three steps to clean the data: (1) Remove census tracts without enough Google Project Sunroof data, (2) Remove census tracts with no potential for solar panels, and (3) Remove census tracts with low median household income.

These steps have been added to the `ProjectSolarV1.m` file.

1. Remove census tracts where Google Project Sunroof analyzed < 95% of all buildings in the census tract:

```
data = data((data.percent_covered >= 95),:);
```

2. Remove census tracts where there is no potential to install rooftop PV:

```
data = data((data.count_qualified ~= 0),:);
```

3. Remove census tracts with median household income below the 2013 poverty threshold for a 4-person household

```
data = data(data.Med_HHD_Inc_ACS_09_13 >= 23834,:);
```

See: https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-thresholds.html

After removing data from our dataset, we will "reset the index" on our data:

```
data.var1 = (1:1:length(data.Med_HHD_Inc_ACS_09_13))';
```

Question 1.3: How many rows were removed during this process? You may wish to add some code to figure this out.

Task 1.5: Normalize Solar Deployment

As described in the article, we are going to ***normalize*** the data. Doing so requires two equations that are automatically included/performed in the code.

Equation 1 in the *Nature Sustainability Analysis*: calculate and create new column current_installs_relative_to_total_installs

Equation 2 in the *Nature Sustainability* Analysis: to normalize the solar deployment rate, create and use new column state_mean_existing_installation_relative_to_potential in order to create new column normalized_existing_installation

Details:

See **Equation 1 in the Nature Sustainability Analysis:** Compute the percentage of buildings with existing rooftop PV relative to the total number of buildings with roofs that qualify to support PV based on the algorithm and criteria to identify appropriate potential rooftop space for PV deployment set forth by Google Project Sunroof. This is the solar deployment rate.

Census tracts (CTs) were categorized by how well they reach their rooftop PV potential. The number of buildings with installed PV systems in each census tract ($N_{\text{ExistingRooftopPV}}$) was divided by the total number of buildings in that tract (N_{CT}), as shown in equation (1):

$$\text{SolarDeployment}_{\text{CT}} = \frac{N_{\text{ExistingRooftopPV}}}{N_{\text{CT}}} \quad (1)$$

where both the numerator and denominator entries were obtained from the Project Sunroof dataset (https://www.google.com/get/sunroof/data-explorer/), following their detection algorithm and criteria to identify appropriate potential rooftop space for PV deployment[36].

Here is the code that calculates this value and creates a new column:

```
data.current_installs_relative_to_total_installs = ...
    data.existing_installs_count./data.count_qualified.*100;
```

See **Equation 2 in the Nature Sustainability Analysis:** Normalize the solar deployment rate by the average solar deployment rate in each state. This mitigates the effects of variations across states, such as available solar resources, incentive programs and policies, and electricity prices.

Variations across states, such as available solar resources[25], incentive programmes and policies (http://www.dsireusa.org/), electricity prices[26] and state racial compositions[27], were mitigated by normalizing the census tract solar deployment performance by the population (P)-weighted census tract solar deployment performance average in each state, as shown in equation (2). Hence, any value greater than 1 indicates that the census tract has installed more rooftop PV relative to the state average installation, and the opposite is the case for values less than one:

$$\text{StateNormalizedSolarDeployment}_{\text{CT}} = \frac{\text{SolarDeployment}_{\text{CT}}}{\sum_{\text{CT}_{\text{State}}} \frac{P_{\text{CT}}}{P_{\text{State}}} \text{SolarDeployment}_{\text{CT}}} \quad (2)$$

Here is the code that creates a new dataframe (of the state means), populates it with the state-level data, merges that state-level data back into the original dataframe, and uses it to calculate the new normalized_existing_installation column.

```
for Num = 1:size(unique(data.state_name))
    idx_state = find(strcmp(data.state_name, States(Num)));
    mean_state = ...
        mean(data.current_installs_relative_to_total_installs(idx_state));
    df_state_mean(idx_state) = mean_state;
end

data.state_mean_existing_installation_relative_to_potential = df_state_mean;
```

```
data.normalized_existing_installation = ...
    data.current_installs_relative_to_total_installs./...
    data.state_mean_existing_installation_relative_to_potential;
```

To do: Add some code to see the columns "before" and "after" to see the additional new columns added to the table by this step. For clarity write these tables to variables `dataBefore` and `dataAfter`. (Keep the code for grading but you may comment it out once you've completed this comparison exercise.)

Task 1.6: Select California Census Tracts for Analysis

Some of the data science algorithms take quite some time to run when there are many data points being considered. For this first analysis you will focus your study on ***the state of California***. This will serve as a point of comparison to the larger analysis that was done at the national level (***"Disparities in Rooftop Solar Deployment in the United States by Race and Ethnicity"*** by Sunter, et al.).

As an example, to select ***JUST*** the data associated with California from the data dataframe, one would use:

```
data = data(strcmp(data.state_name ,'California'),:);
```

Question 1.4: How many census trats are there in California that will be included in our study? To help answer this question you may wish to make a new table called `dataCalifornia.`

Task 1.7: Group Census Tracts by their Majority Race/Ethnicity

We will now create separate variables to hold data for ***each race/ethnicity in the study***.

The following table lists the Race/Ethnicity and the associated column headers.

Race/Ethnicity	Column name
Asian	`pct_NH_Asian_alone_ACS_09_13`
Black	`pct_NH_Asian_alone_ACS_09_13`
Hispanic	`pct_Hispanic_ACS_09_13`
White	`pct_NH_White_alone_ACS_09_13`
No Majority	

For example, to retrieve the census tracts with an Asian majority (e.g. census tracts in California where 50% or more of the population self-identified as Asian based on self-reporting in the American Community Survey), we would use the following code (that generates a variable `AsianMajority` to hold this subset):

```
AsianMajorityIndex = find (data.pct_NH_Asian_alone_ACS_09_13 >= 50 );
AsianMajority = data(AsianMajorityIndex, :);
```

Question 1.5: How many census tracts are in each racial/ethnic majority group in California?

Question 1.6: Are your results concerning? Why or why not?