**METADATA FILE**

**Data Pre-Processing – Check Down Sheet**

Data Cleaning – Tasks (STAGE 1)

* Fill in missing values
* Identify outliers, noisy data – e.g. linear regression of values
* Correct inconsistent data

Data Integration (STAGE 2)

* Combine data from multiple sources using primary keys, etc.
* Avoid data redundancy

Data Transformation (STAGE 2)

* Aggregate data into variable subsets (low, medium, high)

Data Reduction (STAGE 1)

* Reduce the dataset size (smallest amount of data to accomplish the task)

Data Discretization – “Divide into intervals” e.g. Histogram (STAGE 2)

* Interval labels can replace actual data values (low, medium, high) priorities

**Datasets (5 files total)**

**Upload all files to Github.com**

[**https://github.com/CRobbins77/Project/upload/master**](https://github.com/CRobbins77/Project/upload/master)

**In R Studio, perform a Git Pull to bring files into Project**

**Dataset 1 - Wholesale Distributor Data by County FIPS (1 file)**

U.S. Census Bureau – American Fact Finder

2012 Economic Census of the United States

**SOURCE:https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmkk**

DESCRIPTION: EC1242A1 - Wholesale Trade: Geographic Area Series - Summary Statistics for the U.S., States, Metro Areas, Counties, and Places: 2012

Select “Table View” from the menu and use the table tools to show hidden rows and columns.

Verify that the following boxes are checked for FIPS state code, FIPS county code, 2012 NAICS code, meaning of 2012 NAICS code and number of establishments.

Research NAICS Codes that are related to merchant wholesalers tied to farmers coops,farmers markets,local wholesalers, distributors, retail grocery stores, etc.

**NAICS SIC CODES: http://siccode.com/en/naicscodes/445230/fruit-and-vegetable-markets**

NAICS 42448 - Fresh Fruit and Vegetable Merchant Wholesalers

NAICS 42459 - Other Farm Product Raw Material Merchant Wholesalers

Download .csv file and save as ny\_county\_wholesale

Open file in MS Excel and rename columns accordingly (delete all other fields): County\_FIPS, Name, NAICS\_Code, Description, Year and Total\_Establishments

**Pre-Processing Steps in R:**

1. **Examine the data**
2. **Run general descriptive statistics**
3. **Reduce the dataset size to include only NAICS Codes 42448 and 42459**
4. **Rollup codes by county to calculate the total number of wholesale trade establishments by FIPS**
5. **Identify any outliers – Kernal Density Plot of wholesale establishments**

**Dataset 2 - Poverty Data by County FIPS (1 file)**

U.S. Census Bureau - Small Area Income and Poverty Estimates for 2014

**SOURCE:https://www.census.gov/did/www/saipe/data/statecounty/data/2014.html**

DESCRIPTION: The U.S. Census Bureau's Small Area Income and Poverty Estimates(SAIPE)program provides annual estimates of income and poverty statistics for all school districts, counties (including those with populations <65K) and states.

SAIPE represents the most recent estimates of income and poverty for the administration of federal programs and the allocation of federal funds to local jurisdictions.

**Download .csv file and save as US\_county\_poverty**

Open file in MS Excel and rename columns accordingly (delete all other fields): State\_FIPS, County\_FIPS, State, County, Poverty\_Est\_All\_Ages, Poverty\_Per\_All\_Ages, Poverty\_Est\_Age\_0-17, Poverty\_Per\_Age\_ 0-17, Poverty\_Est\_Age\_5-17, Poverty\_Per\_Age\_5-17 and Med\_HH\_Inc

**Pre-Processing Steps in R:**

1. **Examine the data**
2. **Run general descriptive statistics**
3. **Reduce the dataset size to include only records for the state of NY**
4. **Remove New York State total record = 62 counties**
5. **Combine State\_FIPS and County\_FIPS to form a five digit FIPS code**
6. **Identify duplicate records or missing values**
7. **Identify any outliers – Kernal Density Plot of poverty percent all ages**

**Dataset 3 - Drought Data by County FIPS (2 files)**

USDA – National Drought Mitigation Center

FSA Eligibility Tool: Summary Data (2014 Criteria)

**SOURCE:http://droughtmonitor.unl.edu/fsa/FsaEligibilityState2014.aspx**

DESCRIPTION: Provides county-level data by state to determine, which counties meet the Livestock Forage Disaster Program requirements.

Verify U.S. Drought Monitor Classification Scheme to pull only those counties with severe (D2), extreme (D3) or exceptional drought (D4)conditions.

**SOURCE:http://droughtmonitor.unl.edu/aboutus/classificationscheme.aspx**

Following the DMC Scheme, two datasets are available (D2) and (D3), (D4) does not apply to any counties in New York State.

D2 DEFINITION – Severe drought conditions for at least eight consecutive weeks during the grazing period.

Location: New York State

Grazing Period: Start (01/01/2016) / End (12/31/2016)

Download .csv file and name ny\_county\_D2

Open file in MSExcel and rename columns accordingly (delete all other fields): FIPS, State, County, D2\_Class (ConsecWeeks)

D3 DEFINITION – Extreme drought conditions for at least four (nonconsecutive) weeks during the grazing period.

Location: New York State

Grazing Period: Start (01/01/2016) / End (12/31/2016)

Download .csv file and name ny\_county\_D3

Open file in MSExcel and rename columns accordingly (delete all other fields): FIPS, State, County, D3\_Class (NonConsecWeeks)

**Pre-Processing Steps in R:**

**Both datasets are small enough that all pre-processing steps can be done in MS Excel.**

**Dataset 4 - Farm Data by County (1 file)**

Data.NY.Gov - County Agricultural Districts Profile

**SOURCE:https://data.ny.gov/Economic-Development/County-Agricultural-Districts-Profile/9bc8-mx4a**

DESCRIPTION: Includes data on towns affected, acres in agricultural districts, cropped acres and number of farms.

The database is quite robust since the user can filter information prior to exporting it, saving time in pre-processing.

**TIP: Use the Filter feature in the menu to roll-up the pertinent variables (function “sum”) grouped by county.**

FILE ACCESS: Two available options

1. Export file as a standard .csv
2. Access the API – Requires an app token. To sign-up for an app token, click on this link: [**https://data.ny.gov/login**](https://data.ny.gov/login)

It will require you to first set-up an account

**NYSTATE Socrata ID**

**e-mail:** [**crobbins@oishei.org**](mailto:crobbins@oishei.org)

**password: Nysdata2016**

Next you will need to apply for an app token by clicking on the following link:

**https://dev.socrata.com/foundry/data.ny.gov/8jaw-iviy**

App Token

hpiUm07LyAXAeLZk97XayMxS4

**#R CODE:**

#install.packages("RSocrata")

#library("RSocrata")

#df <- read.socrata("https://data.ny.gov/resource/8jaw- iviy?$$app\_token= hpiUm07LyAXAeLZk97XayMxS4")

#API Endpoint: [**https://data.ny.gov/resource/8jaw-iviy.json**](https://data.ny.gov/resource/8jaw-iviy.json)

If option 1, then download .csv file and name ny\_county\_farms

Open file in MSExcel and rename columns accordingly (delete all other fields): County, Towns\_Affected, Total\_Acres, Farmed\_Acres, Cropped\_Acres, Acres\_Owned, Acres\_Rented and Total\_Farms

**TIP: Since the FIPS code is not included in the file, one must use a NY County FIPS Codes lookup table to append the codes for the 62 counties.**

**SOURCE:https://data.ny.gov/Government-Finance/New-York-State-ZIP-Codes-County-FIPS-Cross-Referen/juva-r6g2/data**

**Download .csv file and name ny\_county\_FIPS**

Open file in MSExcel and rename columns accordingly (delete all other fields): FIPS and County

**Pre-Processing Steps in R:**

1. **Examine the data**
2. **Run general descriptive statistics**
3. **Append County FIPS to County Name**
4. **Identify duplicate records and replace missing values (see below)**

* **Westchester County – Cropped Acres and Acres Rented**
* **Seneca County – Cropped Acres**
* **Putnam County – Cropped Acres and Acres Rented**

1. **Identify any outliers – Kernal Density Plot of small farms**