IT REVOLUTION IN AGRICULTURE

An analytical approach for deploying irrigation technology in New York State

Abstract

To encourage economic development in New York State, a $10 million investment in the agricultural sector is recommended. Agriculture was identified as a key economic driver, producing a total estimated economic impact of $37.6 billion in 2011. (DiNapoli 1) The investment will support the deployment of cutting edge irrigation technology in a single county for the initial round of funding. To determine the recipient, all 62 counties in New York State were evaluated for the program based on the concentration of small farms, percentage of population living in poverty, proximity to wholesale distributors and most recent drought conditions. Counties with the highest total scores were further analyzed using a linear regression model to determine their capacity to increase cropland acres. **The findings from the analysis identified Wayne County as the top ranked county for investment since it offers both short-term and long-term economic growth opportunities.**

Introduction to Data Analytics (CIS 512) / BSC, Fall 2016

Submitted to: Dr. John Coles / Dr. Joaquin Carbonara

Report Prepared By: Curtis Robbins

Datasets and R code: https://github.com/CRobbins77/Project

Table of Contents

Project Goal

To strengthen small farms located in economically depressed communities in New York State, while systematically increasing the production of locally grown produce for wholesale distribution.

Introduction and Topic Selection

The potential for economic growth from farming is profound. It is responsible for direct economic activity of "more than $5.4 billion in commodity sales in New York during 2012, an increase of more than 22 percent from 2007" (DiNapoli 1). New York is ranked as a national leader for a number of agricultural commodities including milk, yogurt, apples, onions, sweet corn, tomatoes, etc. What makes this industry even more impressive is that they have been able to accomplish this level of growth with minimal technological advancements.

Governor Cuomo has made food access a priority of his administration, promoting locally-grown food in traditionally underserved urban communities through a number of initiatives centered on Farmers' Markets. The idea is those consumers who purchase locally-sourced food will in turn support small "family" farms and the communities they reside in, while expanding access of healthy, nutritious foods to residents. It has resulted in a vast supply chain that is trying to keep pace with demand at a time when agriculture remains a rather outdated industry. Farmers are systematically responding to changing consumer preferences and technological advancements all while dealing with a host of challenges including competition from factory farms, pressure to subdivide property for development and extreme weather conditions. For example, in 2016, New York experienced the most extensive severe drought on record. As a result 24 counties across Upstate New York were designated as a natural disaster area by the federal government.

To help mitigate drought conditions, farmers need better data to generate more tailored solutions and that is where information technology can play a key role. With our initial investment, CropX, a new agtech startup, will strategically deploy advanced irrigation technology in a single county for the initial round of funding. As recently publicized in WIRED

magazine, CropX provides cost-effective technology to assist farmers in determining precisely how much water to use on their crops; it amounts to an “IT revolution in agriculture” (Finley).

Analyst’s Workflow and Results (Summary)

To analytically determine a county for investment, the first phase of the analysis merged the four datasets into a single matrix where counties were ranked based on their total score across a series of categorical variables. In the second phase, counties with the highest total scores were further analyzed using a linear regression model (**Figure 1 – Appendix**) to evaluate their capacity to increase cropland acres. The findings from the analysis indicated that Wayne County would be a prime target for investment because farmers can potentially increase their cropped acres by 360% over the long-term, while in the short-term, drought mitigation can assist the high concentration of small farms exposed to “extreme” drought. Full details of the analyst’s workflow are provided in the appendix along with supporting graphics, R code and metadata.

Conclusion and Discussion

Investing in Wayne County has the potential to stretch our dollars furthest by galvanizing economic growth among small farms in depressed rural communities. The workflow for this analysis was developed in a systematic way to allow for data to be readily updated each year and duplicated for similar initiatives. Once the project commences, transparency around the selection process will be promoted, including sharing the Heatmap **(Figure 2 - Appendix),** with community partners. Sharing visual content such as this may offer additional insight into other variables to improve the results of the analysis or any potential reporting errors in the data.

Works Cited

DiNapoli, Thomas P. “The Importance of Agriculture to the New York State Economy.” Office of the State Comptroller. March 2015. http://www.osc.state.ny.us/reports/importance_agriculture_ny.pdf. Accessed 25 October 2016.

Finley, Klint. “A Smart Sensor to Help Farmers Save Water in a Drought.” WIRED Magazine. June 2015. https://www.wired.com/2015/06/smart-sensor-farmers-dont-waste-water-drought/. Accessed 25 October 2016.

Assessment and Validation of Another Project

I chose to replicate Paul Brennan's project titled: Vermont Economic Stimulus. I decided on this particular project because of the challenge it posed in using both Python and R together to conduct the analysis.

The findings from the analysis identified Information Services as having the most potential for economic growth by way of leveraging the economic stability of the Healthcare and Social Assistance industry. Paul arrived at this conclusion by comparing the linear regressions for GDP by year per industry and filtering out those industries based on high growth at both the state and national level. The selection criteria were further refined by looking at the R-squared values closest to 1.0, to identify the most stable and predictable positive growth sector.

After walking through and executing the code, the results of this analysis reinforce Paul's recommendation for increasing economic impact in Vermont. The ability to leverage one industry's economic performance by establishing a strong level of collaboration with an underperforming industry is an innovative approach and could certainly be a plausible solution given the limited population growth and resources afforded this State.

Taking these findings into account, I would be interested in learning more about what specific areas of the healthcare industry are driving GDP? It may be innovative startups, job growth, a highly-skilled workforce or simply an aging population that is primarily responsible for the sustained growth. Having this type of information may open the door to new opportunities for the Information Services sector while at the same time pinpointing specific areas for investment.

Reproducing the Basic Analysis

Datasets and Python / R code available at: https://github.com/CRobbins77/Brennan_Project

Paul's code, whether it be in Python or R was well documented. In the first stage of the analysis, the only issue I ran into was when executing the Python code (api_data_get.py) in PyCharm, an error on one of the called packages surfaced. I had to manually load the "requests" package under preferences/project interpreter in order for the code to run. For some reason the library "requests" could not be found until it was manually installed. The other two called packages "json" and "os" had no issues.

/Library/Frameworks/Python.framework/Versions/3.5/bin/python3.5
/Users/CRobbins/Desktop/api_data_get.py
Process finished with exit code 0

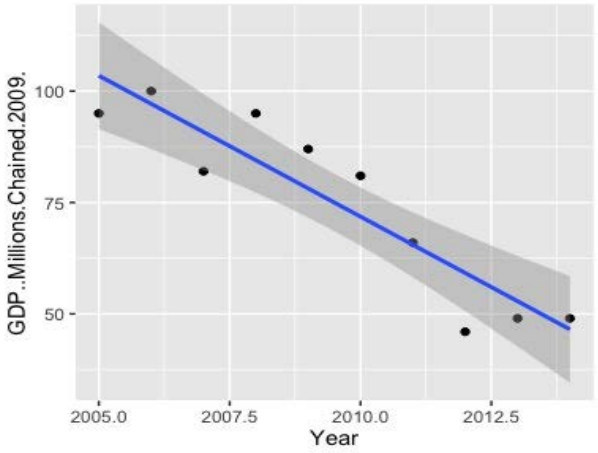
Once the APIs were located using the request package, executing the code in PyCharm produced a series of six datasets. In preparation for the second stage of the analysis, I uploaded the .csv files into the "forked" project on Github for use in RStudio. R was used to further process and analyze the datasets following along with Paul's approach.

R Code Revisions
When importing the files in RStudio, it was not necessary to include the "data" directory for the files. e.g. GDPbyInd_VT <-read.csv ("data/RealGrossOutputbyIndustry_VT.csv") This was a result of the API dumping the results on the desktop in a data folder, however, when running in Github and sharing the files with others, it is not necessary.

State codes in the VT dataset were then matched to the U.S. dataset using fuzzy string matching. From what I concluded, the industry description is read from both sources, R computes a distance matrix between all elements, pairing the elements with the minimum distance using the "stringdist" method. It was highly effective in this case.

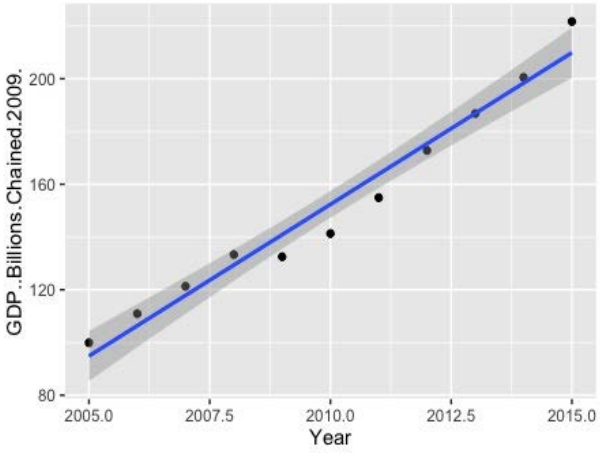
In reviewing the VT dataset, the target industry for economic stimulus was identified as data processing, internet publishing, and other information services (Sector code 514) because of its potential for economic growth in comparison to the U.S. Graphs were then developed to visually identify GDP growth for the information services (IS) sector for both VT and the U.S., using chained dollar GDPs to adjust for inflation with a base year of 2009. (Figure 1.1 and 1.2)

Figure 1.1



Vermont - Information Services
GDP Growth 2005-2015

Figure 1.2



U.S. – Information Services
GDP Growth 2005-2015

The remainder of the analysis performed in R focused on identifying a candidate industry to stimulate economic growth in the IS sector by using the following approach:

- Determine industries with positive/strong growth for VT and the U.S.
- Using the coefficient field, filter out the top ten sectors for growth in each geography
- Apply the intersect command, highlighting where there is overlap between the two areas
- Run descriptive statistics on all four identified sectors
- The Healthcare and Social Assistance sector produced the highest R-squared value (0.986), indicative of a stable industry with year over year positive linear growth

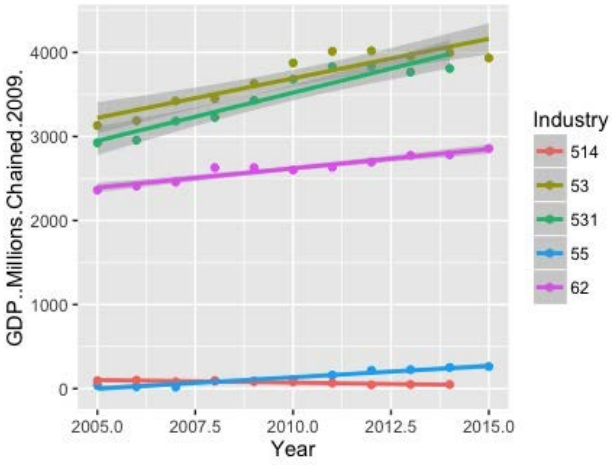
Healthcare and Social Assistance (Sector 62)

Coefficients:

Estimate Std. Error t value Pr(>|t|)
*(Intercept) -85256.500 3451.624 -24.70 1.40e-09 ****
*Year 43.289 1.717 25.21 1.17e-09 ****
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*
Residual standard error: 18.01 on 9 degrees of freedom
***Multiple R-squared: 0.986**, Adjusted R-squared: 0.9845*
F-statistic: 635.5 on 1 and 9 DF, p-value: 1.169e-09

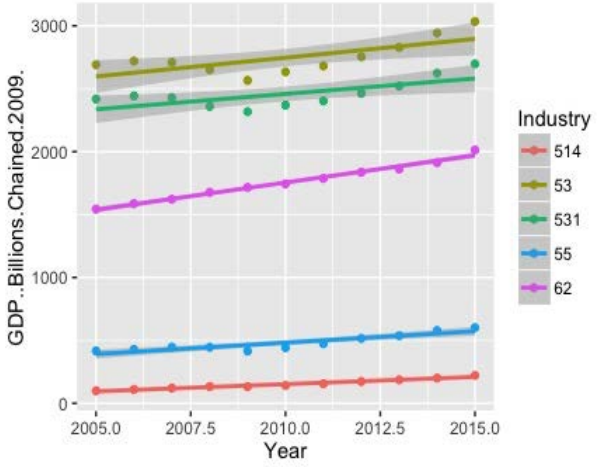
The linear regression graphs (Figure 2.1 and 2.2) represent the top four sectors for GDP growth among both VT and U.S. including the information services sector (514). Sector 62, depicted in "Magenta" represents the Healthcare and Social Assistance industry.

Figure 2.1



Vermont: Top 4 Growth Sectors
GDP Growth 2005-2015

Figure 2.2



U.S.: Top 4 Growth Sectors
GDP Growth 2005-2015

APPENDIX

TABLE OF CONTENTS

A) – ANALYST'S WORKFLOW AND RESULTS

Data Access and Use in Project:

Data for the analysis was obtained from both state and federal sources including:

Dataset 1: Wholesale Distributor Data by County FIPS (2012 Economic Census)
Dataset 2: Poverty Data by County FIPS (2014 U.S. Census Bureau)
Dataset 3: Drought Data by County FIPS (2014 USDA – Drought Mitigation Center)
Dataset 4: Farms Data by County (2016 Data.NY.Gov – Agricultural Districts Profile)

County was determined as the smallest geographic area for which agriculture data was available. Therefore, the primary key or unique identifier common across all four datasets was County FIPS (5 digit code). All four datasets were exported in .csv format to GitHub where they were pulled into RStudio for statistical analysis. Pre-processing steps varied for each dataset; descriptive statistics, subsetting, aggregating, replacing missing values and identifying outliers using Kernal Density plots.

The farms dataset required a linear regression model to calculate missing fields, since initial bootstrapping techniques did not provide accurate results. The regression analysis generated the following equation to describe the statistical relationship between farmed acres and cropped acres: **Cropped_Acres=(.4748)*(Farmed_Acres) + 1617.97**

It is a significant relationship provided that 56% of the variation is explained by the linear model **(R-squared = .5641)** and the **p-value = 3.321e-10** indicates that changes to the farmed acres value are related to the changes in the cropped acres values. Therefore their predictor variables will also be significant. Missing Cropped_Acres fields were populated using the regression equation.

Subjects or Data Source:

Out of 62 counties in New York State, 53 were identified as having an agricultural district. Of these counties, only 27 had a concentration of small farms with average cropped acres <= 100. Cropped acres represent the portion of farmed acres that are used to grow a cash crop. The difference between the two represents land that is currently being used for grazing or is simply left fallow with an opportunity for future crop production.

Experimental Procedure:

Analysis – Phase 1: Categorical Variable Scoring

A scoring matrix was developed to rank each county based on the concentration of small farms, percentage of population living in poverty, proximity to wholesale distributors and most recent drought conditions. Values for each categorical variable were first classified as low, average or high by calculating the mean and standard deviation for each. Scores were then assigned for each classification (low = 1 / average = 3 / high = 5) and aggregated, allowing for the numeric ranking of each county. Counties meeting the conditional statement: total scores >= 16 and drought score >1 (severe or extreme) were identified as potential candidates for investment: Chautauqua County (36013), Chemung County (36015) and Wayne County (36117).

Analysis – Phase 2: Linear Regression Modeling

Applying the regression model (Cropped_Acres=(.4748)*(Farmed_Acres) + 1617.97) to the three potential candidates determined the percentage that cropped acres can proportionally increase in relation to farmed acres. The findings from the analysis indicated that Wayne County would present the best opportunity for investment because farms can potentially increase their cropped acres by 360%, compared to 36% for Chautauqua and 33% for Chemung.

B) – SUPPORTING GRAPHICS

Figure 1

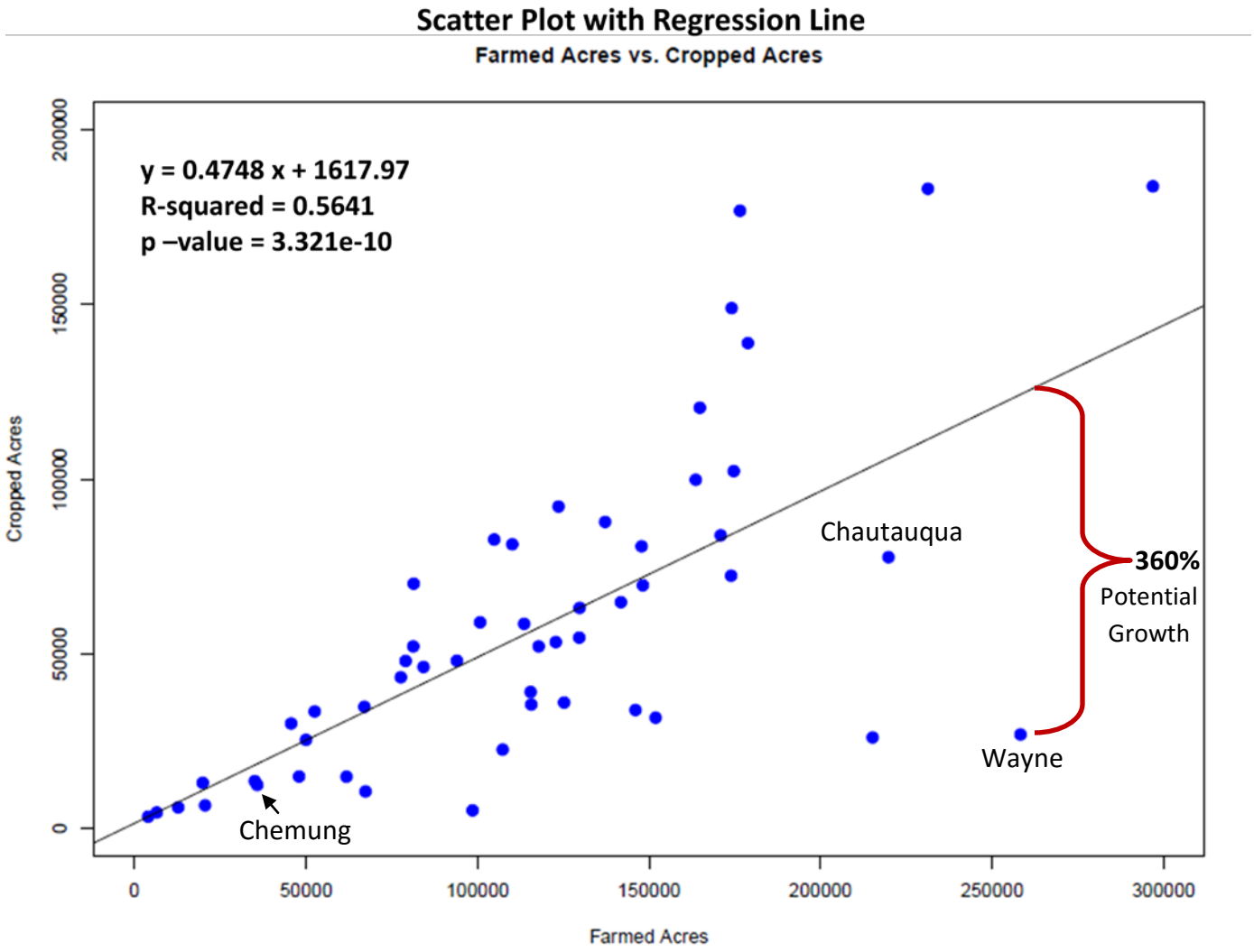
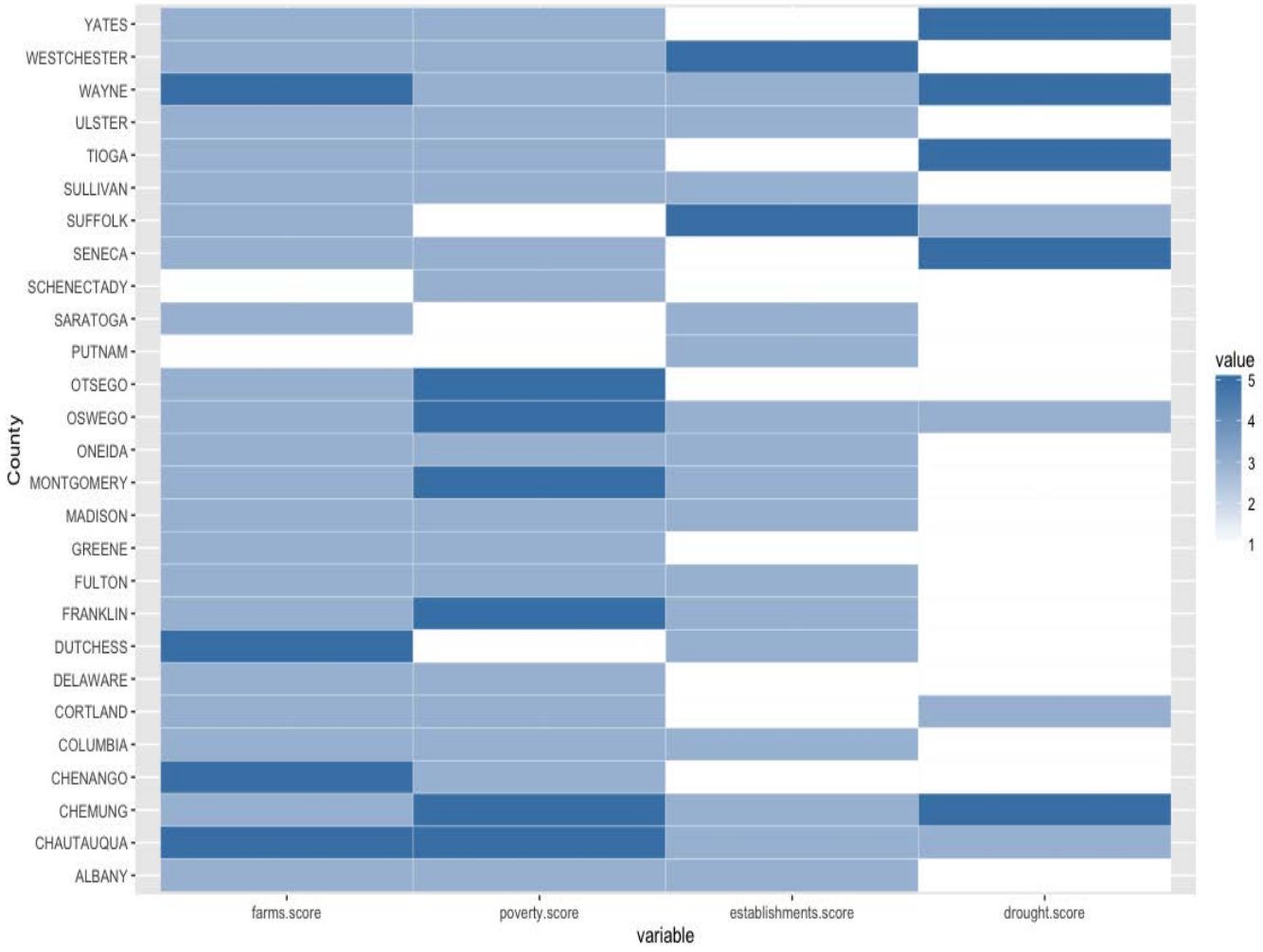


Figure 2


YATES
WESTCHESTER
WAYNE
ULSTER
TIOGA
SULLIVAN
SUFFOLK
SENECA
SCHENECTADY
SARATOGA
PUTNAM
OTSEGO
OSWEGO
ONEIDA
MONTGOMERY
MADISON
GREENE
FULTON
FRANKLIN
DUTCHESS
DELAWARE
CORTLAND
COLUMBIA
CHENANGO
CHEMUNG
CHAUTAUQUA
ALBANY
County
farms.score
poverty.score
establishments.score
drought.score
variable
value
5
4
3
2
1

C) – PROJECT CODE: PROGRAMMED IN R

#%%%%%% LOAD PACKAGES %%%%%%%

#Load the VIM and mice packages in R (used for inputting missing values in Dataset #4)>>>>>

#The following packages should also be loaded in RStudio before running:

#"assertthat", datasets", "dplyr", "ggplot2", "graphics", "grDevices", "methods", "Rcpp", "reshape2", "scales", "stats" and "utils"

#%%%%%% PHASE 1 - DATA CLEANING %%%%%%%

#<<<<<<DATASET 1 - Load raw data for wholesale distribution by County FIPS>>>>>>

wholesale <- read.csv("ny_county_wholesale.csv", header = TRUE)

#------Examine the imported dataset------

#Display the first six records

head(wholesale)

#Descriptive stats - median and mean for each column

summary(wholesale)

#Display structure of the data frame (individual data types)

str(wholesale)

#Retrieve only those records with NAICS Codes 42448 and 42459

wholesale_dist <- subset(wholesale, NAICS_Code %in% c(42448,42459))

#Aggregate codes by county to calculate the total number of wholesale trade establishments by FIPS

wholesale_est <- aggregate (Total_Establishments~County_FIPS,

sum,data=wholesale_dist)

#Use a Kernal Density Plot to view the distribution of wholesale establishments among counties

est_plot <-density(wholesale_est$Total_Establishments)

plot(est_plot)

#FINDINGS - urban counties closer to NYC have a higher number of wholesale trade establishments

#Dataset 1 (wholesale_est) - Wholesale Establishments by County – Cleaned

#<<<<<<DATASET 2 - Load raw data for poverty data by County FIPS>>>>>>

poverty <- read.csv("US_county_poverty.csv", header = TRUE)

#------Examine the imported dataset------

#Display the first six records

head(poverty)

#Descriptive stats - median and mean for each column

summary(poverty)

#Display structure of the data frame (individual data types)

str(poverty)

#Retrieve only poverty data for counties in New York State

ny_poverty <- subset(poverty, State %in% c('NY'))

#Remove New York State total record (row 1), creating file with only 62 counties

ny_poverty <- subset(ny_poverty, County_FIPS >0)

#Add leading zeros to County_FIPS to allow for combining of State and County FIPS codes

ny_poverty$County_FIPS <- formatC(ny_poverty$County_FIPS, width = 3, format = "d", flag = "0")

```
#Combine State_FIPS and County_FIPS to form FIPS field in table
ny_poverty$FIPS <- paste0(ny_poverty$State_FIPS,ny_poverty$County_FIPS)
#Change the order of columns so FIPS is first
ny_poverty <- ny_poverty[,c(12,1:11)]
#List rows of data that have missing values
ny_poverty[!complete.cases(ny_poverty),]
#No rows in the dataset have missing values, next identify any outliers in the dataset by graphing
#First begin by displaying structure of the data frame (individual data types)
str(ny_poverty)
#Since poverty percentage per all ages is a factor, it must be converted to a numeric value before graphing
ny_poverty$Poverty_Per_All_Ages <-
as.numeric(as.character(ny_poverty$Poverty_Per_All_Ages))
#Use a Kernal Density Plot to view the distribution of poverty percentage all ages
pov_plot <-density(ny_poverty$Poverty_Per_All_Ages)
plot(pov_plot)
#Based on graph, look specifically at counties with poverty percentages less than 10 and greater than 20
low_pov <- subset(ny_poverty, Poverty_Per_All_Ages <10)
#With Medium Household Income no less than $70K for these four counties, the low poverty rate is in line
high_pov <- subset(ny_poverty, Poverty_Per_All_Ages >20)
```

#With Medium Household Income of $34K and $48K respectively for Bronx and Kings County, poverty percentages remain very high because

#of the concentration of poverty in this urban area;counties represent the Bronx and Brooklyn in the five boroughs of NYC.

#FINDINGS - There are a total of 62 Counties in NYS

#Dataset 2 (ny_poverty) - Poverty by NYS County – Cleaned

#<<<<<<DATASET 3 - Load Drought data (2 files) by County FIPS>>>>>>

#See metadata file; both datasets are small enough that all pre-processing steps can be done in MS Excel.

```
severe_drought <- read.csv("ny_county_D2.csv", header = TRUE)
extreme_drought <- read.csv("ny_county_D3.csv", header = TRUE)
```

#Append extreme_drought classfication to severe_drought classification by county.

```
ny_drought_con <- merge(severe_drought,extreme_drought, by=c("FIPS", "State", "County"), all=TRUE)
```

#Replace remaining NAs under D3_Class with 0

```
ny_drought_con[is.na(ny_drought_con)] <- 0
```

#Dataset 3 (ny_drought_con) Drought Conditions by County - Cleaned

#<<<<<<DATASET 4 - Load raw data for farm data by County FIPS>>>>>>

```
farms <- read.csv("ny_county_farms.csv", header = TRUE)
```

#------Examine the imported dataset------

#Display the first six records

```
head(farms)

#Descriptive stats - median and mean for each column

summary(farms)

#Display structure of the data frame (individual data types)

str(farms)

#Check if there are any duplicate county records

n_occur <- data.frame(table(farms$County))

n_occur[n_occur$Freq > 1,]

farms[farms$County %in% n_occur$Var1[n_occur$Freq > 1],]

#Load raw data for County FIPS codes

FIPS <- read.csv("ny_county_FIPS.csv", header = TRUE)

#Display structure of the data frame (individual data types)

str(FIPS)

#Change the order of columns so County is first for appending

FIPS <- FIPS[,c("County","FIPS")]

#Change county field from lowercase to uppercase prior to merge

FIPS <- as.data.frame(sapply(FIPS, toupper))

#Append County FIPS to farms dataset

farms_FIPS <- merge(FIPS,farms, by=c("County"))

#List rows of data that have missing values

farms_FIPS[!complete.cases(farms_FIPS),]

#Results of analysis indicate that Putnam, Seneca and Westchester have missing data.
```

#Use the mice function called "md.pattern" to get a better understanding of the pattern of missing data.

md.pattern(farms_FIPS)

#OPTION 1 - Try using a random sampling with replacement method to fill in the missing fields.

#One option is to use the R package: Multivariate Imputation by Chained Equations (mice) to calculate missing values in the dataset.

#Include Random Forest function as a regression and classification method to accommodate interactions and non-linearities.

#Data Source: http://r-statistics.co/Missing-Value-Treatment-With-R.html (Section 4.3 mice)

#miceMod <- mice(farms_FIPS[, !names(farms_FIPS) %in% "medv"], method="rf") # perform mice imputation, based on random forests.

#farms_alldata <- complete(miceMod) # generate the completed data

#anyNA(farms_alldata)

#The missing values are now populated.

#Look at the populated fields, do the numbers make sense? Can cropped acres be more than farmed acres?

#After further review, the cropped acres field does not appear to be an independent variable.

#In fact there exists a distinct relationship between farmed acres and cropped acres.

#Therefore, peforming a bootstrapping analysis with mice (random forests) does not provide accurate results.

#OPTION 2 - Perform a linear regression analysis to determine an equation for calculating the missing fields.

with (farms_FIPS,plot (Farmed_Acres,Cropped_Acres))

```
lm(Cropped_Acres~Farmed_Acres,data=farms_FIPS)

abline(lm(Cropped_Acres~Farmed_Acres,data=farms_FIPS))

#Coefficients:

#(Intercept)  Farmed_Acres

#1617.9660        0.4748

#Equation of a line: y=mx+b

#Cropped_Acres=(.4748)*(Farmed_Acres) + 1617.97

#Check R2 value - This model accounts for 56% of the variance, a more accurate method for

estimating the missing values.

farms.lm <- lm(Cropped_Acres~Farmed_Acres,data=farms_FIPS)

summary(farms.lm)$r.squared

#R-Sq=.5641

summary(farms.lm)

#p-value=2.14e-11

#Populate missing Cropped_Acres fields by County using the regression equation.

#Putnum County = 3490

farms_FIPS[35,5]=(farms_FIPS[35,4] * .4748) + 1617.97

#Seneca County = 63,210

farms_FIPS[41,5]=(farms_FIPS[41,4] * .4748) + 1617.97

#Westchester County = 4,680

farms_FIPS[51,5]=(farms_FIPS[51,4] * .4748) + 1617.97

#Round new values under Cropped_Acres to zero decimal places

farms_FIPS$Cropped_Acres <- round(farms_FIPS$Cropped_Acres, digits=-1)
```

#Replace remaining NAs under Acres_Rented fields with 0 since Acres_Rented = Total Acres

farms_FIPS[is.na(farms_FIPS)] <- 0

#Create new field titled "Avg_CA_Per_Farm" = Average Cropped Acres Per Farm rounded to zero decimal places

farms_FIPS$Avg_CA_Per_Farm <- with(farms_FIPS,

round(Cropped_Acres/Total_Farms),0)

#Next display descriptive stats and structure of the new data frame

summary(farms_FIPS)

str(farms_FIPS)

#Use a Kernal Density Plot to view the distribution of small farms <= 100 cropped acres

sf_plot <-density(farms_FIPS$Avg_CA_Per_Farm)

plot(sf_plot)

small_farms <- subset(farms_FIPS, Avg_CA_Per_Farm <=100)

#27 out of 53 counties with an agricultural district in NYS have been identified as having

#average cropped acres per farm <100 (concentration of small farms)

#Note: Using the bootstrapping method for missing values only 26 counties were identified.

#Dataset 4 (small_farms) - NYS Counties with a High Concentration of Small Farms - Cleaned

#<<<<< All coding complete - All 5 Datasets ready for analysis (11-07-16)>>>>>

#<<<<< Dataset #4 revised after further review of bootstrapping analysis (11-22-16)>>>>>

#%%%%%% PHASE 2 - ANALYSIS STAGE %%%%%%%

#Prepare for the first stage of the analysis begin by merging the following datasets.

#Perform a left outer join of the small_farms dataset and the remaining 3 data frames.

```
farms_alldata <- merge(small_farms,ny_drought_con, by=c("FIPS"), all.x=TRUE)
farms_alldata <- merge(farms_alldata,wholesale_est, by.x=c("FIPS"),
by.y=c("County_FIPS"), all.x=TRUE)
farms_alldata <- merge(farms_alldata,ny_poverty, by=c("FIPS"), all.x=TRUE)
#Create a function to remove all of the unwanted columns after merging and rename columns
accordingly.
col.dont.want <- c("State.x", "County.y", "State_FIPS", "County_FIPS", "State.y",
"County", "Acres_Owned", "Acres_Rented", "Poverty_Est_All_Ages",
"Poverty_Est_Age_0.17","Poverty_Per_Age_.0.17", "Poverty_Est_Age_5.17",
"Poverty_Per_Age_5.17", "Med_HH_Inc")
analysis_dataset <- farms_alldata[,!names(farms_alldata) %in% col.dont.want,drop=F]
names(analysis_dataset)[names(analysis_dataset)=="County.x"]<-"County"
#Replace remaining NAs under farms_alldata fields with 0
analysis_dataset[is.na(analysis_dataset)] <- 0
#Display structure of the data frame (individual data types)
str(analysis_dataset)
summary(analysis_dataset)
#Convert all integers (Total_Acres, Farmed_Acres and Total_Farms) to numeric values before
categorizing.
analysis_dataset$Total_Acres <- as.numeric(as.character(analysis_dataset$Total_Acres))
analysis_dataset$Farmed_Acres <-
as.numeric(as.character(analysis_dataset$Farmed_Acres))
analysis_dataset$Total_Farms <- as.numeric(as.character(analysis_dataset$Total_Farms))
```

#Because the dataset has a unique field for each county (27 different FIPS codes), a decision tree in R could not be used.

#Instead, a scoring matrix will be utilized to rank each county based on the predetermined categorical variables.

#Begin by assigning scores (low = 1 /average = 3 / high = 5) across the 4 categorical variables for each county.

#Based on the counting principle, this will produce 81 different combinations (3x3x3x3).

#Determine distribution ranges for data by calculating the mean and standard deviation for each field to be used in the analysis:

#Total_Farms, Total_Establishments and Poverty_Per_All_Ages (Note: Drought is already classified as extreme and severe)

sd(analysis_dataset$Poverty_Per_All_Ages,na.rm=FALSE) #SD 3.858 and Mean is 14.07 (Class: Low<10.2 / Average 10.2-17.9 / High>17.9)

sd(analysis_dataset$Total_Establishments,na.rm=FALSE) #SD 7.111 and Mean is 3.481 (Class: Low<1 / Average 1-10.6 / High>10.6)

sd(analysis_dataset$Total_Farms,na.rm=FALSE) #SD 350.175 and Mean is 544.9 (Class: Low<194.7 / Average 194.7-895.1 / High>895.1)

#Next classify the above fields based on the calculated distribution ranges.

analysis_dataset$farms.type<-

ordered(cut(analysis_dataset$Total_Farms,c(0,195,895,1600), labels=c(1,3,5)))

analysis_dataset$poverty.type<-

ordered(cut(analysis_dataset$Poverty_Per_All_Ages,c(0,10.2,17.9,20), labels=c(1,3,5)))

```
analysis_dataset$establishments.type<-
ordered(cut(analysis_dataset$Total_Establishments,c(-1,.95,11,35), labels=c(1,3,5)))
#Create a function to convert the rankings to numeric values, retaining the original values (1,3,5)
as.numeric.factor <- function(x) {as.numeric(levels(x))[x]}
analysis_dataset$farms.score <- as.numeric.factor (analysis_dataset$farms.type)
analysis_dataset$poverty.score <- as.numeric.factor (analysis_dataset$poverty.type)
analysis_dataset$establishments.score <- as.numeric.factor
(analysis_dataset$establishments.type)
#Classify the Drought field as Normal = 1 / Severe = 3 / Extreme = 5
analysis_dataset$drought.score<-ifelse(analysis_dataset$D3_Class>0,5,
ifelse(analysis_dataset$D2_Class>0,3,1))
#Create a new total.score field by aggregating all 4 scores in the dataset.
analysis_dataset$total.score <-analysis_dataset$farms.score +
analysis_dataset$poverty.score + analysis_dataset$establishments.score +
analysis_dataset$drought.score
#Determine the capacity for expansion of cropped acres among the three counties.
#Begin by determining the additional capacity for crop acres using the previous regression
equation for cropped acres vs. farmed acres (Dataset 4)
#Cropped_Acres=(.4748)*(Farmed_Acres) + 1617.97
#Create a function to automate the calculation.
regression.eq <- function(x) {as.numeric((.4748)*(x)) + 1617.97}
analysis_dataset$capacity <- regression.eq (analysis_dataset$Farmed_Acres)
```

#Determine the percentage that cropped acres can proportionally increase before increasing farmed acres.

analysis_dataset$per.increase <- ((analysis_dataset$capacity - analysis_dataset$Cropped_Acres)/analysis_dataset$Cropped_Acres) * 100

#Prioritize counties based on their opportunities and challenges.

#Step 1: Identify counties with total scores >= 16 and drought.score >1 (severe and extreme drought conditions)

inv_counties <- subset(analysis_dataset, total.score >= 16 & drought.score >1)

#The analysis identified three prospective counties as potential candidates for investment.

#Chautauqua County (FIPS Code = 36013), Chemung County (FIPS Code = 36015) and Wayne County (FIPS Code = 36117)

#Step 2: From the remaining counties, identify a single county for investment (greatest potential for increasing cropped acres = largest per.increase value)

target_co <- head(inv_counties[order(inv_counties$per.increase, decreasing=T),],n = 1)

#Final Results - Wayne County will be targeted for investment because they can potentially increase their cropped acres by 360%,

#before investing in additional farmed acres, offering both short-term (drought) and long-term (increased production) economic growth opportunities.

#%%%%%%%%%% Code and Analysis Complete %%%%%%%%%%

#%%%%%%%%%% Supporting Graphics for Project %%%%%%%%%%

#Linear Regression Plot

with (farms_FIPS, plot(Farmed_Acres, Cropped_Acres,

```
            pch = 20, cex = 2.0, col="blue", main = "Farmed Acres vs. Cropped Acres",
sub = "(Scatter Plot with Regression Line)",
            xlab = "Farmed Acres", ylab = "Cropped Acres",
            xlim = c(0,300000), ylim = c(0,200000)))
lm(Cropped_Acres~Farmed_Acres,data=farms_FIPS)
abline(lm(Cropped_Acres~Farmed_Acres,data=farms_FIPS))
#Select a subset of variables to create a heatmap
myvars <- c("County", "farms.score", "poverty.score", "establishments.score",
"drought.score")
plotdata <- analysis_dataset[myvars]
plotdata <- arrange(plotdata, desc(County))
#Heatmap for Data Visualization
data.m <- melt(plotdata)
heatmap <- ggplot(data.m, aes(variable,County)) +
  geom_tile(aes(fill=value),colour="white") +
  scale_fill_gradient(low = "white", high = "steelblue")
print(heatmap)
#%%%%%%%%%% Supporting Graphics Complete %%%%%%%%%%
```

D) - METADATA

Dataset 1 - Wholesale Distributor Data by County FIPS

U.S. Census Bureau, American Fact Finder, 2012 Economic Census of the United States

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

Data Source:

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

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

Dataset 2 - Poverty Data by County FIPS

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

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.

Data Source: https://www.census.gov/did/www/saipe/data/statecounty/data/2014.html

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

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

USDA – National Drought Mitigation Center, FSA Eligibility Tool: Summary Data (2014 Criteria)

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

Data Source: http://droughtmonitor.unl.edu/fsa/FsaEligibilityState2014.aspx

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

DROUGHT CLASS: 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)

Dataset 4 - Farm Data by County (1 file)

Data.NY.Gov - County Agricultural Districts Profile

Description: Includes data on agricultural districts in New York State including towns affected, total acres, farmed acres, cropped acres and number of farms.

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

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

TIP #1: 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*
   It will require you to first set-up an account
   NYSTATE Socrata ID
   E-mail: *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

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 #2: 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.

Data 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

E) – END OF PROJECT REFLECTION

The content covered in this course is a perfect primer for anyone interested in data analytics. It's especially true for a working professional looking to gain hands on experience in understanding how to structure databases, develop an analyst workflow, code in various software environments and describe, condense and evaluate data in a timely manner.

In my particular workplace, new skills are only acquired through projects, leaving little time to become fluent in advanced software programs and data science techniques. What this course has afforded me is a sound foundation for expanding my skills. Now that I am over the initial learning curve with coding, I will be more apt to utilize it on a more frequent basis, as they say; it's another tool in my arsenal.