MGT 6203 Data Analytics in Business
Group Project Proposal

TEAM INFORMATION

Team #:

95

Team Members:

Daniel Forcade
dforcade@gatech.edu
Quechee, VT

Background in Statistics and Computer Science, with a focus on Complex Systems. Some analytical projects I've worked on are Allotaxonometry-themed social media predictive engines, and U.S. Election sentiment analysis.

Ryan Hopkins
ryanhop@gatech.edu
Union, KY

Background in Mathematics and Physics, with a focus on theory. Some experience in Computer Science and Quantum Computing Theory. Previously worked as a Data Analyst/ Project Manager role for a Clinical Research Org.

Soheil Sameti
ssameti3@gatech.edu
Atlanta, GA

Background in Transportation Data Science, with a focus on GIS and modeling. I work on data visualization and analyze large datasets of trips using R, Python and GIS.

Steven Wasserman
swasserman9@gatech.edu
St. Louis, MO

Data analyst and systems engineer for The MITRE Corporation, supporting defense logistics and transportation. Previous UVA graduate in systems engineering and computer science and homeland security intern. Quantum computing and biomimetic design enthusiast!

OBJECTIVE/PROBLEM

Project Title

Food: An Analysis on the Future of Crop Production in the United States

Background Information

American agronomist and leader of the Green Revolution Norman Borlaug once said, "Civilization as it is known today could not have evolved, nor can it survive, without an adequate food supply." Every major civilization throughout history was stimulated by the cultivation of a major agricultural system that could provide food for its citizens and support the development of an empire. Technological innovations led to the compartmentalization of food production processes by farmers, manufacturers, and sellers alike, supported by a global supply chain, making one thing clear, in perpetuity: food is a necessity.

While crop production in the United States began the moment English settlers landed on the East Coast, major government intervention through different areas has been largely prevalent since World War 2. New Deal farm programs from the 1930s were further advanced with federal funds, including farm loans, commodity subsidies, and price support. Technology created for the war was then repositioned as fertilizers, and later, pesticides. Other technologies were repurposed for supporting labor requirements in harvesting as well as major innovations in animal husbandry, like milking

parlors, grain elevators, and confined animal-feeding operations. Government advancements in technology also supported the development of plant breeding research, leading to modern-day genetically modified organisms (GMOs) with a cadre of benefits and drawbacks.

Food and its production has evolved over hundreds of years and will continue to change well into the future. Much of this change is impacted or encouraged by government intervention, both direct and indirect. Given the ranging perspectives held by the public at large, it is essential to provide informed perspectives on different areas of interest in food production and analyze the future of these perspectives on food cultivation.

Problem Statement

Government stakeholders require data-informed recommendations on crop production that consider a variety of competing policy priorities – including climate change, public health, and environmental conservation – to support decision-making on agricultural subsidies that are future-facing and support both producers and consumers.

Primary Research Question (RQ)

How can data-driven insights from precision agriculture support government recommendations for crop production that best benefit the public?

Supporting Research Questions

1. Where and how are crops currently grown throughout the United States?
2. How might changes to regional climates impact farmland in the future, and can risks to crop production be predicted?
3. Can localized weather, climate, and ecology data be used to recommend regional crop production practices?
4. What impact does accessibility to affordable and nutritious foods have on public health?
5. What role does the production of different crops and produce have in changes to the environment? Can ecological sustainability be ameliorated through more concerted efforts in growing food?
6. Can economic controls be effectively leveraged to promote growth of selected crops that best benefit farmers and consumers?

Business Justification

Food has played a significant role in government administration, positioned at the cornerstone of impactful policy decisions and widespread public perspectives and opinions on many interrelated topics. Agricultural subsidies administered by the U.S. Department of Agriculture (USDA) drive support for farmers and assist in regulating commodities markets to allow for fair competition in a consistent economic ecosystem. Food stamps and the cadre of school lunch programs provide food benefits to low-income families and extend the accessibility of nutritious foods to those who might otherwise not be able to afford them. Furthermore, the security, safety, and efficacy of food and other biological products are tasked to the U.S. Food and Drug Administration (FDA), the federal watchdog responsible for the protection of public health.

Food, as evidenced by the past, will also continue to transform and change in response to a variety of factors, many of which are contemplated in this project. Ecological systems and the environment at large have been in constant evolution since the dawn of domesticated agriculture almost 12,000 years ago, and climate change sits ready to threaten farmland availability and food security as the planet warms. Agricultural markets and the business of crop cultivation rest upon capitalistic forces

that drive production in the United States, supported in large by the federal and state governments in concert with the private sector to generate supply to demand from the masses. Public health and perceptions of food in general also play an important role in promoting and cultivating a healthy flow of produce from farm to table.

As the watchdog of food and the people at-large, the U.S. government must be well-informed on the future of food cultivation in the United States as well as which foods can best support different policy targets, whether environmental, economic, or health-related. Moreover, this research can apprise private industries on how to posture and establish businesses that can adapt to future needs now, benefitting themselves and hopefully, through reduction of inefficiency, the public.

DATASET/PLAN FOR DATA

Data Sources

Links, attachments, etc.

The aggregated list of suggested datasets found during initial data discovery are summarized in the table below, including elements from the next two sections on data descriptions and key variables descriptions.

Name: National Agricultural Statistics Service – CroplandCROS
Link: https://croplandcros.scinet.usda.gov/

Description	Key Variables
USDA's Agricultural Research Service aims to enhance research capabilities by giving scientists access to powerful computing resources, fast networks for sharing data, and training in how to use these technologies effectively.	Area Selection, Filter Crops, Layer Selection, Year Selection and Animation.

Name: NOAA Storm Events Database
Link: https://catalog.data.gov/dataset/ncdc-storm-events-database2

Description	Key Variables
This dataset comes from the National Oceanic and Atmospheric Administration and contains extensive data on Storm events in the United States, including but not limited to location, date, precipitation, estimated crop damage, and episodic narrative. It contains three associated datasets for each year, which are split between Details, Fatalities (Not required in our use case), and Location.	Latitude, Longitude, Event Type, Event Narrative, Crop Damage, Date

Name: NOAA Temperature and Precipitation
Link: https://www.ncei.noaa.gov/cdo-web/search

Description	Key Variables
This dataset comes from the National Oceanic and Atmospheric Administration and contains location-paired data on a daily average, maximum, and minimum temperatures across the United States, as well as precipitation. It will be used to supplement the NOAA Storm Events Data listed above, as that data does not contain temperatures, which are crucial to crop selection.	Latitude, Longitude, Temp Max, Temp Min, Temp Avg, Precipitation, Wind Gust, Date

Name: NASS CropScape Cropland Data
Link: https://nassgeodata.gmu.edu/CropScape/

Description
Same data foundation as the CroplandCROS, but with additional parameters and features, maintained by George Mason University.

Key Variables
Crop, Location, Year

Name: FAO United Nations Crop Information
Link: https://www.fao.org/land-water/databases-and-software/crop-information/en/

Description
United Nations Food and Agriculture synopsis format data and information on various crops and their water usage and yield throughout various growth stages.

Key Variables
Water Requirements, Yield, Crop Growth Stages

Name: FNDDS Food Nutrient Data
Link: https://fdc.nal.usda.gov/download-datasets.html#bkmk-3

Description
A Food and Nutrient Database for Dietary Studies database that provides nutrient profiles for specific foods and beverages and their associated portions and recipes. Specifically, we'll use *nutrient.csv* and *fndds_ingredient_nutrient_value.csv* datasets.

Key Variables
Ingredient Description, Nutrient Code/Id, Nutrient Value, Unit_Name

Name: USDA ARMS Farm Financial and Crop Production Practices
Link: https://www.ers.usda.gov/data-products/arms-farm-financial-and-crop-production-practices/ and https://catalog.data.gov/dataset/arms-farm-financial-and-crop-production-practices

Description
The Agricultural Resource Management Survey (ARMS) is the U.S. Department of Agriculture's primary source of information on the production practices, resource use, and economic well-being of America's farms and ranches. The results of this survey give farmers, ranchers, and many others factual insights into many aspects of farming, ranching, and conditions in agricultural communities.

Key Variables
variable_name, estimate, category2_value, category_value, year, state

Found here

Name: Fruits And Vegetables Prices In USA
Link: https://www.kaggle.com/datasets/anshikakashyap12/fruits-and-vegetables-prices-in-usa

Description
This dataset contains information about the 'Fruits and Vegetables Prices In USA In The Year 2020'. The dataset contains 8 columns and 156 rows.

Key Variables
Item, Form, Retail Price, Retail Price Unit, Yield, Cup Equivalent Size, Cup Equivalent Unit, Cup Equivalent Price

Name: Commodity Costs and Returns
Link: https://www.ers.usda.gov/data-products/commodity-costs-and-returns/documentation.aspx and https://catalog.data.gov/dataset/commodity-costs-and-returns

Description
USDA has estimated annual production costs and returns and published accounts for major field crop and

Key Variables
Various variables for operating costs, allocated overhead, and

livestock enterprises since 1975. Cost and return estimates are a primary data source reported for the United States and for Farm Resource Regions for corn, soybeans, wheat, cotton, grain sorghum, rice, peanuts, oats, barley, milk, hogs, and cow-calf	total costs listed; Various variables for net value and production practices

Found here

Name: Dairy Data
Link: https://catalog.data.gov/dataset/dairy-data and http://www.ers.usda.gov/data-products/dairy-data/documentation.aspx

Description	**Key Variables**
These data are from several USDA agencies. They were previously included in the Meat Statistics page in the Livestock, Dairy, and Poultry Outlook tables and may contain revisions not included in previous releases of the LDP tables.	Various variables for milk production, prices, consumer indices, output, and more

Found here

Name: Vegetables and Pulses Data
Link: https://www.ers.usda.gov/data-products/vegetables-and-pulses-data/trade-and-prices-by-category/ and https://catalog.data.gov/dataset/vegetables-and-pulses-data

Description	**Key Variables**
These tables provide vegetable and pulses data by category (e.g. price, imports, exports, etc.) for fresh and processed products. Tables are updated as new information is available throughout the year (typically monthly).	Year, Production/Imports, Total Supply, Per Capita Availability, Current Dollars

Also look here: https://www.ers.usda.gov/data-products/vegetables-and-pulses-data/vegetables-and-pulses-yearbook-tables/

Found here

Name: Feed Grains Database
Link: https://catalog.data.gov/dataset/feed-grains-database and https://www.ers.usda.gov/data-products/feed-grains-database/feed-grains-yearbook-tables/

Description	**Key Variables**
The Feed Grains Database contains statistics on four feed grains (corn, grain sorghum, barley, and oats), foreign coarse grains (feed grains plus rye, millet, and mixed grains), hay, and related items.	Year/Q#, Beginning Stocks, Production, Imports, Total Supply, Domestic Use, Exports, Ending Stocks

Found here

Name: National Agriculture Imagery Program (NAIP)
Link: https://catalog.data.gov/dataset/national-agriculture-imagery-program-naip and https://earthexplorer.usgs.gov/

Description	Key Variables
The National Agriculture Imagery Program (NAIP) acquires aerial imagery during the agricultural growing seasons in the continental U.S. A primary goal of the NAIP program is to make digital ortho photography available to governmental agencies and the public within a year of acquisition. NAIP is administered by the USDA's Farm Service Agency (FSA) through the Aerial Photography Field Office in Salt Lake City. This "leaf-on" imagery is used as a base layer for GIS programs in FSA's County Service Centers, and is used to maintain the Common Land Unit (CLU) boundaries.	Polygon variables that describe data

Found here

Name: Agricultural Productivity in the U.S.
Link: https://catalog.data.gov/dataset/agricultural-productivity-in-the-u-s and https://www.ers.usda.gov/data-products/agricultural-productivity-in-the-u-s/

Description	Key Variables
This data product provides agricultural output, inputs, and total factor productivity indices for the aggregate farm sector for the period 1948 to 2021 (table 1), along with price and quantity indices for 10 outputs and 12 inputs (table 1a). It also contains estimates of output, inputs, and the growth and relative levels of productivity across U.S. States for 1960–2004 (tables 3–23).	Year, Total Agr. Output, Output by commodity, capital Investments inputs, and Labor Investment inputs

Found here

Name: Trade and Prices by Category
Link: https://www.ers.usda.gov/data-products/vegetables-and-pulses-data/trade-and-prices-by-category/

Description	Key Variables
Described above in *Vegetables and Pulses Data* entry	Described above in Vegetables and Pulses Data entry

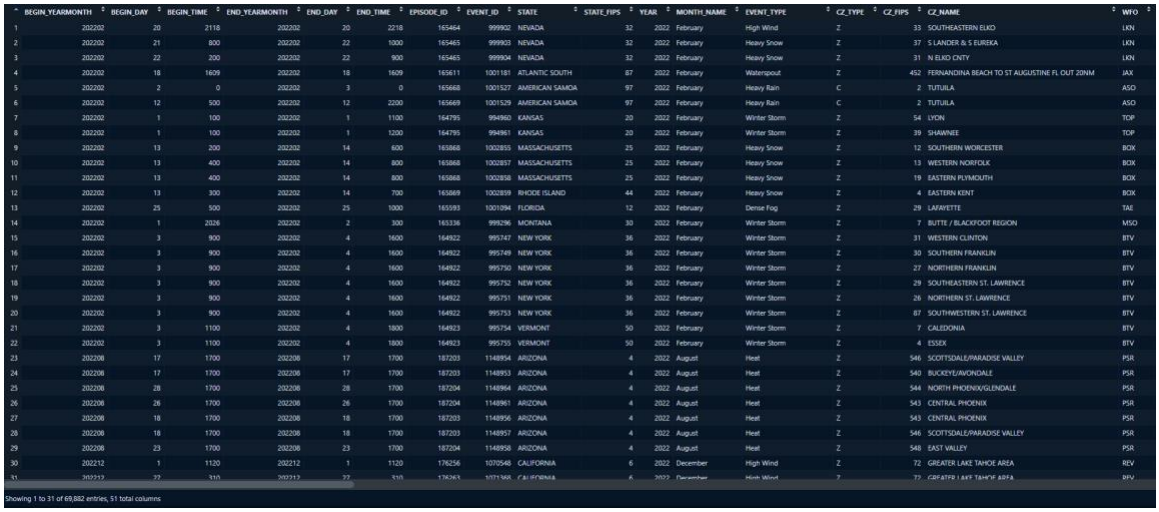
Found here

Name: Nutrition Information for Raw Vegetables
Link: https://www.fda.gov/food/food-labeling-nutrition/nutrition-information-raw-vegetables

Description	Key Variables
Nutrition facts for various macronutrient by vegetable Found here	Calories, Fat, Potassium, Carbohydrates, Sugars, Protein, Vitamin A, Vitamin C, Calcium, Iron

Data Description
Describe each of your data sources, include a few screenshots of a few rows of data

The comprehensive list of initial data sources is included above, with descriptions of each data source. Several sources are extrapolated below with screenshots displaying data from each.



	BEGIN_YEARMONTH	BEGIN_DAY	BEGIN_TIME	END_YEARMONTH	END_DAY	END_TIME	EPISODE_ID	EVENT_ID	STATE	STATE_FIPS	YEAR	MONTH_NAME	EVENT_TYPE	CZ_TYPE	CZ_FIPS	CZ_NAME	WFO
1	202202	20	2118	202202	20	2218	165464	999902	NEVADA	32	2022	February	High Wind	Z	33	SOUTHEASTERN ELKO	LKN
2	202202	21	800	202202	22	1000	165465	999903	NEVADA	32	2022	February	Heavy Snow	Z	37	S LANDER & S EUREKA	LKN
3	202202	22	200	202202	22	900	165465	999904	NEVADA	32	2022	February	Heavy Snow	Z	31	N ELKO CNTY	LKN
4	202202	18	1609	202202	18	1609	165611	1001181	ATLANTIC SOUTH	87	2022	February	Waterspout	Z	452	FERNANDINA BEACH TO ST AUGUSTINE FL OUT 20NM	JAX
5	202202	2	0	202202	3	0	165668	1001527	AMERICAN SAMOA	97	2022	February	Heavy Rain	C	2	TUTUILA	ASO
6	202202	12	500	202202	12	2200	165669	1001529	AMERICAN SAMOA	97	2022	February	Heavy Rain	C	2	TUTUILA	ASO
7	202202	1	100	202202	1	1100	164795	994960	KANSAS	20	2022	February	Winter Storm	Z	54	LYON	TOP
8	202202	1	100	202202	1	1200	164795	994961	KANSAS	20	2022	February	Winter Storm	Z	39	SHAWNEE	TOP
9	202202	13	200	202202	14	600	165868	1002855	MASSACHUSETTS	25	2022	February	Heavy Snow	Z	12	SOUTHERN WORCESTER	BOX
10	202202	13	400	202202	14	800	165868	1002857	MASSACHUSETTS	25	2022	February	Heavy Snow	Z	13	WESTERN NORFOLK	BOX
11	202202	13	400	202202	14	800	165868	1002858	MASSACHUSETTS	25	2022	February	Heavy Snow	Z	19	EASTERN PLYMOUTH	BOX
12	202202	13	300	202202	14	700	165869	1002859	RHODE ISLAND	44	2022	February	Heavy Snow	Z	4	EASTERN KENT	BOX
13	202202	25	500	202202	25	1000	165593	1001094	FLORIDA	12	2022	February	Dense Fog	Z	29	LAFAYETTE	TAE
14	202202	1	2026	202202	2	300	165336	999296	MONTANA	30	2022	February	Winter Storm	Z	7	BUTTE / BLACKFOOT REGION	MSO
15	202202	3	900	202202	4	1600	164922	995747	NEW YORK	36	2022	February	Winter Storm	Z	31	WESTERN CLINTON	BTV
16	202202	3	900	202202	4	1600	164922	995749	NEW YORK	36	2022	February	Winter Storm	Z	30	SOUTHERN FRANKLIN	BTV
17	202202	3	900	202202	4	1600	164922	995750	NEW YORK	36	2022	February	Winter Storm	Z	27	NORTHERN FRANKLIN	BTV
18	202202	3	900	202202	4	1600	164922	995752	NEW YORK	36	2022	February	Winter Storm	Z	29	SOUTHEASTERN ST. LAWRENCE	BTV
19	202202	3	900	202202	4	1600	164922	995751	NEW YORK	36	2022	February	Winter Storm	Z	26	NORTHERN ST. LAWRENCE	BTV
20	202202	3	900	202202	4	1600	164922	995753	NEW YORK	36	2022	February	Winter Storm	Z	87	SOUTHWESTERN ST. LAWRENCE	BTV
21	202202	3	1100	202202	4	1800	164923	995754	VERMONT	50	2022	February	Winter Storm	Z	7	CALEDONIA	BTV
22	202202	3	1100	202202	4	1800	164923	995755	VERMONT	50	2022	February	Winter Storm	Z	4	ESSEX	BTV
23	202208	17	1700	202208	17	1700	187203	1148954	ARIZONA	4	2022	August	Heat	Z	546	SCOTTSDALE/PARADISE VALLEY	PSR
24	202208	17	1700	202208	17	1700	187203	1148953	ARIZONA	4	2022	August	Heat	Z	540	BUCKEYE/AVONDALE	PSR
25	202208	28	1700	202208	28	1700	187204	1148964	ARIZONA	4	2022	August	Heat	Z	544	NORTH PHOENIX/GLENDALE	PSR
26	202208	26	1700	202208	26	1700	187204	1148961	ARIZONA	4	2022	August	Heat	Z	543	CENTRAL PHOENIX	PSR
27	202208	18	1700	202208	18	1700	187203	1148956	ARIZONA	4	2022	August	Heat	Z	543	CENTRAL PHOENIX	PSR
28	202208	18	1700	202208	18	1700	187203	1148957	ARIZONA	4	2022	August	Heat	Z	546	SCOTTSDALE/PARADISE VALLEY	PSR
29	202208	23	1700	202208	23	1700	187204	1148958	ARIZONA	4	2022	August	Heat	Z	548	EAST VALLEY	PSR
30	202212	1	1120	202212	1	1120	176256	1070548	CALIFORNIA	6	2022	December	High Wind	Z	72	GREATER LAKE TAHOE AREA	REV
31	202212	22	310	202212	22	310	176263	1071368	CALIFORNIA	6	2022	December	High Wind	Z	72	GREATER LAKE TAHOE AREA	REV

Showing 1 to 31 of 69,882 entries, 51 total columns

Figure 1. *Screenshot of data from the 'NOAA Storm Events Database' entry*

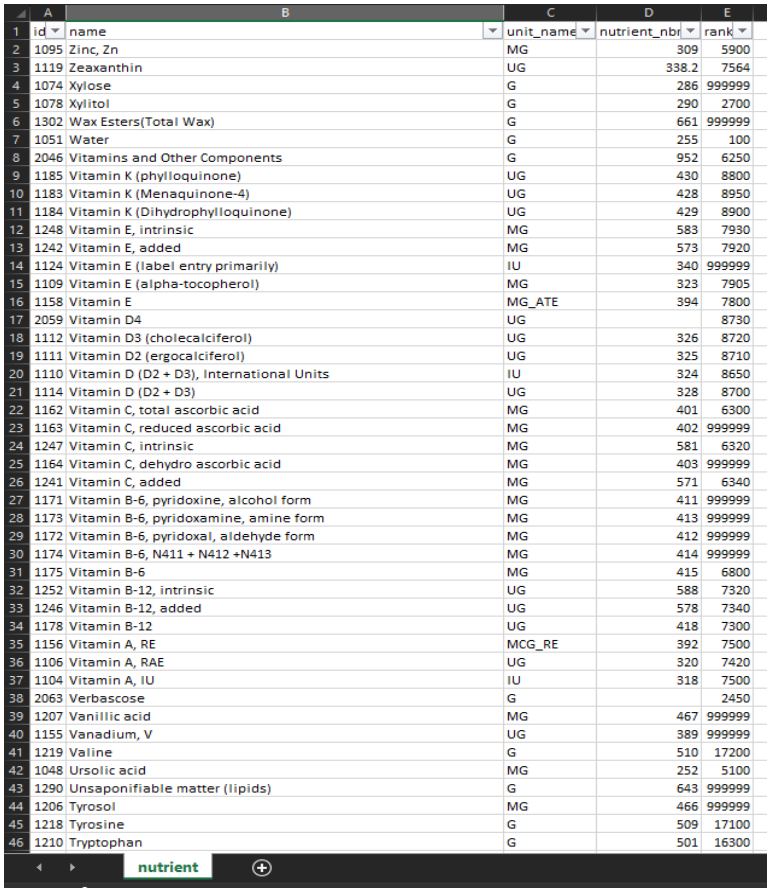
	LATITUDE	LONGITUDE	ELEVATION	DATE	PRCP	SNOW	TAVG	TMAX	TMIN
691	44.54310	-72.52940	365.8	2024-01-20	NA	NA	4	8	0
57	44.77000	-71.70170	379.5	2024-01-20	NA	NA	4	9	0
58	44.77000	-71.70170	379.5	2024-01-21	NA	NA	6	14	-5
692	44.54310	-72.52940	365.8	2024-01-21	NA	NA	7	13	-4
56	44.77000	-71.70170	379.5	2024-01-19	NA	NA	7	15	-1
847	44.46825	-73.14990	101.1	2024-01-20	0.02	0.6	8	14	6
690	44.54310	-72.52940	365.8	2024-01-19	NA	NA	8	17	0
20	44.50780	-73.11560	103.6	2024-01-20	NA	NA	9	13	5

Figure 2. *Screenshot of data from the 'NOAA Temperature and Precipitation' entry*

	A	B	C	D	E	F	G	H	I	J	K
1	Ingredient code	Ingredient description	Nutrient code	Nutrient value	Nutrient value source	FDC ID	Derivati	SR AddMod year	Foundation year acquired	Start date	End date
7983	2005	Apples, fuji, with skin, raw	614	0.11	SR Legacy	170918		1977	0	2019-01-0	12/3/2020
7984	2005	Apples, fuji, with skin, raw	617	7.035	SR Legacy	170918		1977	0	2019-01-0	12/3/2020
7985	2005	Apples, fuji, with skin, raw	618	3.122	SR Legacy	170918		1977	0	2019-01-0	12/3/2020
7986	2005	Apples, fuji, with skin, raw	619	0.15	SR Legacy	170918		1977	0	2019-01-0	12/3/2020
7987	2005	Apples, fuji, with skin, raw	620	0	SR Legacy	170918		1995	0	2019-01-0	12/3/2020
7988	2005	Apples, fuji, with skin, raw	621	0	SR Legacy	170918		1995	0	2019-01-0	12/3/2020
7989	2005	Apples, fuji, with skin, raw	626	0.09	SR Legacy	170918		1977	0	2019-01-0	12/3/2020
7990	2005	Apples, fuji, with skin, raw	627	0	SR Legacy	170918		1995	0	2019-01-0	12/3/2020
7991	2005	Apples, fuji, with skin, raw	628	0	SR Legacy	170918		1995	0	2019-01-0	12/3/2020
7992	2005	Apples, fuji, with skin, raw	629	0	SR Legacy	170918		1995	0	2019-01-0	12/3/2020
7993	2005	Apples, fuji, with skin, raw	630	0	SR Legacy	170918		1995	0	2019-01-0	12/3/2020
7994	2005	Apples, fuji, with skin, raw	631	0	SR Legacy	170918		1995	0	2019-01-0	12/3/2020
7995	2005	Apples, fuji, with skin, raw	645	7.125	SR Legacy	170918		1977	0	2019-01-0	12/3/2020
7996	2005	Apples, fuji, with skin, raw	646	3.272	SR Legacy	170918		1977	0	2019-01-0	12/3/2020
30422	9003	Apples, gala, with skin, raw	203	0.26	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30423	9003	Apples, gala, with skin, raw	204	0.17	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30424	9003	Apples, gala, with skin, raw	205	13.81	SR Legacy	171688	NC	2007	0	2019-01-0	12/3/2020
30425	9003	Apples, gala, with skin, raw	208	52	SR Legacy	171688	NC	2007	0	2019-01-0	12/3/2020
30426	9003	Apples, gala, with skin, raw	221	0	SR Legacy	171688		1985	0	2019-01-0	12/3/2020
30427	9003	Apples, gala, with skin, raw	255	85.56	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30428	9003	Apples, gala, with skin, raw	262	0	SR Legacy	171688	Z	2001	0	2019-01-0	12/3/2020
30429	9003	Apples, gala, with skin, raw	263	0	SR Legacy	171688	Z	2001	0	2019-01-0	12/3/2020
30430	9003	Apples, gala, with skin, raw	269	10.39	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30431	9003	Apples, gala, with skin, raw	291	2.4	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30432	9003	Apples, gala, with skin, raw	301	6	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30433	9003	Apples, gala, with skin, raw	303	0.12	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30434	9003	Apples, gala, with skin, raw	304	5	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30435	9003	Apples, gala, with skin, raw	305	11	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30436	9003	Apples, gala, with skin, raw	306	107	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30437	9003	Apples, gala, with skin, raw	307	1	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30438	9003	Apples, gala, with skin, raw	309	0.04	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30439	9003	Apples, gala, with skin, raw	312	0.027	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30440	9003	Apples, gala, with skin, raw	317	0	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30441	9003	Apples, gala, with skin, raw	319	0	SR Legacy	171688	Z	2002	0	2019-01-0	12/3/2020
30442	9003	Apples, gala, with skin, raw	320	3	SR Legacy	171688		2003	0	2019-01-0	12/3/2020
30443	9003	Apples, gala, with skin, raw	321	27	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30444	9003	Apples, gala, with skin, raw	322	0	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020
30445	9003	Apples, gala, with skin, raw	323	0.18	SR Legacy	171688	A	2003	0	2019-01-0	12/3/2020

fndds_ingredient_nutrient_value

Figure 3. *Screenshot of data from the* fndds_ingredient_nutrient_value.csv *described above*



	A	B	C	D	E
1	id	name	unit_name	nutrient_nbr	rank
2	1095	Zinc, Zn	MG	309	5900
3	1119	Zeaxanthin	UG	338.2	7564
4	1074	Xylose	G	286	999999
5	1078	Xylitol	G	290	2700
6	1302	Wax Esters(Total Wax)	G	661	999999
7	1051	Water	G	255	100
8	2046	Vitamins and Other Components	G	952	6250
9	1185	Vitamin K (phylloquinone)	UG	430	8800
10	1183	Vitamin K (Menaquinone-4)	UG	428	8950
11	1184	Vitamin K (Dihydrophylloquinone)	UG	429	8900
12	1248	Vitamin E, intrinsic	MG	583	7930
13	1242	Vitamin E, added	MG	573	7920
14	1124	Vitamin E (label entry primarily)	IU	340	999999
15	1109	Vitamin E (alpha-tocopherol)	MG	323	7905
16	1158	Vitamin E	MG_ATE	394	7800
17	2059	Vitamin D4	UG		8730
18	1112	Vitamin D3 (cholecalciferol)	UG	326	8720
19	1111	Vitamin D2 (ergocalciferol)	UG	325	8710
20	1110	Vitamin D (D2 + D3), International Units	IU	324	8650
21	1114	Vitamin D (D2 + D3)	UG	328	8700
22	1162	Vitamin C, total ascorbic acid	MG	401	6300
23	1163	Vitamin C, reduced ascorbic acid	MG	402	999999
24	1247	Vitamin C, intrinsic	MG	581	6320
25	1164	Vitamin C, dehydro ascorbic acid	MG	403	999999
26	1241	Vitamin C, added	MG	571	6340
27	1171	Vitamin B-6, pyridoxine, alcohol form	MG	411	999999
28	1173	Vitamin B-6, pyridoxamine, amine form	MG	413	999999
29	1172	Vitamin B-6, pyridoxal, aldehyde form	MG	412	999999
30	1174	Vitamin B-6, N411 + N412 +N413	MG	414	999999
31	1175	Vitamin B-6	MG	415	6800
32	1252	Vitamin B-12, intrinsic	UG	588	7320
33	1246	Vitamin B-12, added	UG	578	7340
34	1178	Vitamin B-12	UG	418	7300
35	1156	Vitamin A, RE	MCG_RE	392	7500
36	1106	Vitamin A, RAE	UG	320	7420
37	1104	Vitamin A, IU	IU	318	7500
38	2063	Verbascose	G		2450
39	1207	Vanillic acid	MG	467	999999
40	1155	Vanadium, V	UG	389	999999
41	1219	Valine	G	510	17200
42	1048	Ursolic acid	MG	252	5100
43	1290	Unsaponifiable matter (lipids)	G	643	999999
44	1206	Tyrosol	MG	466	999999
45	1218	Tyrosine	G	509	17100
46	1210	Tryptophan	G	501	16300

nutrient

Figure 4. *Screenshot of data from the* nutrient.csv *described above*

Key Variables

Which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?

The comprehensive list of initial data sources is depicted above, including descriptions of key variables from each source. A general approach to incorporating key variables from each dataset as is pertinent for analysis is described here.

Some of our key variables in the Weather Events analysis, outside of geographic location, are Event Type, Event Description, and Precipitation. For the crop nutrient data, key variables include Nutrient Value and Nutrient Type.

We plan to create several new variables based on pre-existing and transformed features within the established data, to help fit with our selected analytical and modeling approaches. Some transformations we are considering are scaling, reclassification, rank, and indexing. For example, ranking or indexing weather events by severity may be more beneficial for predicting Crop Suitability in terms of stakeholder interpretability. One of our aims in this project is to create both simple new variables (such as temperature range on each day), and complex or nuanced new variables such as event sentiment analysis indexes, or Crop Health Score.

Variables hypothesized to be most important:

- Location Variables
- Precipitation
- Event Severity
- Temperature

- Crop Hardiness
- Crop Health Score

In this context, some of our independent variables are:

- Temperature
- Precipitation
- Wind Gust
- Nutrition Content

Our dependent variables are:

- Crop Health Score
- Crop Suitability

APPROACH/METHODOLOGY

Planned Approach

In paragraph(s), describe the approach you will take and what are the models you will try to use. Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?

Our approach to this project follows the core principles of established data science pipelines:

1. Objective Identification
2. Research Domain for Familiarity
3. Data Collection and Integration
4. Data Cleaning and Preprocessing
5. Exploratory Data Analysis
6. Modeling
7. Model Evaluation and Selection

We plan to continue to build out our domain knowledge and research over the first phase of the project while diving deeper into the datasets we have already collected. Through this phase, we are committed to finding auxiliary data on crop profiles to supplement our currently established data to enhance predictive and analytical models and outputs. We are using a combination of multiple datasets, so there will be many joins, merges, and transformations that will need to be completed in the data preprocessing phase – specifically on the Longitude/Latitude coordinate system, ZIP Code, and FIPS columns.

We have several ideas for models to use, with our leading models being Clustering and Random Forest models, each for separate use cases. We believe that k-means Clustering is a suitable candidate for bucketing Event severity relative to crop. We intend to build a Random Forest model to predict what type of crop is growing in specific locations as of 2022, and into the future based on our climate and crop data. For a more experimental approach, we are considering using a scaled Term Frequency-Inverse Document Frequency (TF-IDF) Natural Language Processing (NLP) method combined with Principal Component Analysis (PCA) for dimensionality reduction on the Descriptive Text data from the NOAA Events database, and using a DBSCAN Clustering algorithm on that transformed output to search for patterns or additional insights into the human description and interpretation of events, if time allows.

Special consideration will be given to a GIS-based approach to map and dashboard our preprocessed latitude and longitude Weather and Events data and integrate it with the NASS

CropScape Cropland dataset. Within this approach, we plan to produce the following assessments and methods:

1. Precipitation Analysis: Identify ideal crop recommendations based on precipitation.
2. Soil Erosion Assessment: Identify and compare which crops are best suited for current and projected soil erosion.
3. Storm Frequency and Severity: Recommend crops with proper resistance to significant weather effects
4. Spatial Interpolation: Use an interpolation technique such as Inverse Distance Weighting (IDW) or Kriging (Gaussian) to fill in the gaps in data and impute missing GIS values.
5. Weighted Overlay: Create a rank-based weight scoring system based on these factors to determine which crops should be grown in which areas, when matched with our nutrition and health analysis. This will be represented in an overlay on our GIS dashboard.

In terms of model validation and optimization, we plan to evaluate more advanced models with a Training-Validation-Test split, and then use k-fold cross-validation to determine optimal parameters and model selection. For classification models, F1, ROC, and accuracy/precision may be used. For simpler models such as linear regression and multiple regression, R-squared, p-value, and MAE will be looked at.

Due to the multifaceted and diverse nature of this analysis, we expect many localized results and key insights to be found within the data as we work through our outline and schedule. For instance, we anticipate finding stakeholder relevant insights into the storm and event profiling, precipitation trends, crop nutritional profiles, and public health information; all which can individually deliver insight and benefit to stakeholders before any meta-analysis is conducted. Methods such as linear and multiple regression will be used to demonstrate simple, yet effective relationships; while methodologies such as Random Forest and Clustering will deliver more complex outputs and groupings that can be of tremendous value to stakeholder crop decision-making. By considering both the individual variable analysis and having a plan for larger, combined analytical deliverables such as the Crop Suitability Score, we aim to create a holistic analysis that strives for interpretable results and deliverables, that can be digested effectively by various stakeholders throughout the technical spectrum.

Anticipated Conclusions/Hypothesis

What results do you expect? How will your approach lead you to determine the final conclusion of your analysis? Note: At the end of the project, you do not have to be correct or have acceptable accuracy – the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement.

Our multi-modal approach will enable a comprehensive conclusion on both the predictability and estimation of crop cultivation throughout the United States as well as the prioritization of production by crop along several laterals, including health benefit, environmental impact, and economic sustainability. Through employment of various GIS techniques alongside analysis of weather, climate, and other geologic data, the team will be able to provide insight into the weather conditions that support or hinder effective farming as well as demonstrate capabilities for estimating such conditions into the future, for consideration by government and private stakeholders alike. Through analysis of various datasets that support a variety of different policy bases, the team can also provide reasonable estimates for crop production that can benefit a number of different policy platforms, including food availability impacts on public health, crop production impacts on regional ecosystems, and cost/benefit analyses for producers and consumers of crops. This dynamic, comprehensive approach

aims to underpin a system-of-systems strategy that can synchronize foresighted information on food security from disciplines that are generally siloed.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

Direct business decisions that may be affected include regionally based crop recommendations based on seasonality trends of flood resistance, temperature, wind-gust tolerance, storm severity, and weather-event description. Additional consideration will also be given to public health, nutritional content and environmental impact.

In large-scale agriculture, immediate business impact may be difficult to quantify, as many producers and operations are locked into multi-year contracts - involving significant capital investments in loans, grants, and subsidies. However, we believe that this analysis will provide valuable insight in assisting decisions for targeted crop selection based on climate trends and event severity in the designated regions. Further, this analysis will provide immediate auxiliary business impact for those farmers who are not locked into multi-year crop contracts, such as small scale and independent farmers.

PROJECT TIMELINE/PLANNING

Project timeline; mention key dates you hope to achieve certain milestones by

Key Dates

- Project Proposal – *due February 25th*
- Progress Report – *due March 17th*
- Final Report – *due April 21st*
- Teammate Evaluation – *due April 25th*

Schedule

Week #	Dates	Goal	Outcomes

1	February 25th → March 3rd	Project Setup and Data Collection	• Team Kickoff: Define roles, establish communication protocols, and set up project management tools.
			• Literature Review: Dive into existing research on precision agriculture, focusing on the application of data analytics in sustainable farming.
			• Data Collection:
			  ○ Download the USDA Agriculture, Land, and Farm Management Database for foundational agricultural data.
			  ○ Access the Phenology Database for insights into plant growth stages and their environmental interactions.
			  ○ Obtain USGS Landsat Imagery to analyze geographical and crop health information over time.

2	March 3rd → March 10th	Data Preparation	• Data Cleaning: Address anomalies in datasets such as the Soil Erosion and Organic Matter for Great Central Plains and the ACPF Database, ensuring consistency and reliability. • Data Integration: Combine datasets like US Weather Events and the NREL National Solar Radiation Database with agricultural data to establish a

			multifaceted view of the factors affecting sustainable farming.
3	March 10th → March 17th	Exploratory Data Analysis	• Descriptive Statistics: Apply to all datasets to understand basic trends and outliers, particularly focusing on crop yield and soil health data. • Visualization: Map data from the USGS Landsat Imagery and ACPF Database to visualize spatial patterns in agriculture and conservation practices. • Correlation Analysis: Examine relationships between weather patterns from the US Weather Events dataset and agricultural outcomes in the USDA database.
4-5	March 17th → March 31st	Model Development	• Feature Engineering: Use insights from the Phenology Database and Soil Erosion data to create features that encapsulate growth stages and soil health impacts on yield. • Model Selection: Choose models suited to the diverse nature of our datasets, such as using ARIMA models for forecasting based on the weather patterns dataset. • Model Training: Leverage the integrated dataset, ensuring a comprehensive approach that accounts for variables across all sources.
5-6	March 31st → April 7th	Model Evaluation and Refinement	• Performance Metrics: Apply metrics relevant to each model type, considering the specificity of datasets like crop yield and soil health data. • Cross-Validation: Implement using the integrated dataset to ensure models are robust across different agricultural contexts. • Model Tuning: Refine models by considering the unique characteristics of each dataset, such as the spatial resolution in satellite imagery or temporal resolution in weather data.
7-8	April 7th → April 21st	Insights and Reporting	• Insight Generation: Draw on the comprehensive analysis of datasets like the ACPF Database and Landsat Imagery to inform sustainable practices. • Recommendations: Base actionable strategies on insights derived from datasets, ensuring recommendations are data-driven and tailored to the nuances captured in our analysis. • Final Report and Presentation: Synthesize findings from all datasets, highlighting how integrated data analysis can inform precision agriculture practices.

Appendix

Any preliminary figures or charts that you would like to include

None Currently