**Week 1 Task Sheet: Data Understanding, Cleaning & Exploration**

**1. Understand the Project Purpose**  
The goal of this project is to understand how climate variables affect agricultural crop yields, particularly in the context of climate change. By analyzing climate data alongside USDA corn yield data, we aim to identify which environmental factors—like temperature, rainfall, and humidity—are most strongly associated with crop productivity. This understanding is vital for improving food security and developing climate-resilient agriculture, especially in the face of increasing weather extremes.

**Key Climate Variables:** Temperature (max, min, avg), Precipitation, Relative Humidity, Vapor Pressure Deficit, Wind Speed, Radiation Flux.

**Importance:** This research can inform policy, guide future crop planning, and help farmers adapt to a changing climate with better decision-making tools.

**2. Load and Review the Datasets**

* **Crop Yield Data:** USDA Corn County 2022
  + 1516 rows, 14 columns
  + Includes both numerical (e.g., yield, production) and categorical (e.g., state, county) data
  + No significant missing values after cleaning
* **Climate Data:** HRRR March 2022 - Alabama
  + 48,799 rows, 23 columns
  + Primarily numerical values with categorical identifiers (e.g., State, County, Date)
  + Minor inconsistencies were fixed (e.g., unified FIPS codes and column formats)

**3. Clean and Prepare the Data**

* Column names converted to snake\_case
* County and state names standardized
* Date column created by merging year, month, and day
* Temperature converted from Kelvin to Celsius
* Merged datasets using FIPS code (left join); dropped missing values for analysis

**4. Generate Summary Statistics**

* Crop Yield: Mean = ~139, Std Dev = varies by county
* Max Temp: Mean ~22.5°C, wide variability
* Min Temp: Mean ~8.7°C, lower variability
* Precipitation: Highly variable, with extremes

**Most Variable:** Radiation Flux, Max Temperature  
**Least Variable:** Vapor Pressure Deficit, Wind Components  
**Unusual Values:** Precipitation spikes, extreme radiation values, high temp outliers

**5. Create Visualizations**

* **Max Temperature Trend:** Fluctuates daily with observable peaks
* **Precipitation Trend:** Large spikes on certain days, indicating storm events
* **Crop Yield Histogram:** Centered around 130–140 bu/acre, some lower outliers

**6. Explore Relationships Between Datasets**

* Averaged climate data by FIPS and merged with crop yield data
* Correlation matrix computed using only numeric features
* **Most positively correlated:** Avg/Max Temp, Radiation Flux
* **Negatively correlated:** High Relative Humidity, Wind Speed

**7. Summary Report (300–400 words)**  
This week’s analysis focused on understanding and preparing USDA crop yield data and HRRR climate data to identify early trends in how environmental variables affect corn yield. The crop dataset was clean, structured, and provided county-level yield information. The HRRR dataset, while larger and more complex, offered rich climate metrics at the daily level. Key challenges included standardizing the FIPS codes and unifying the date and location formats for merging.

We calculated descriptive statistics which highlighted radiation flux and maximum temperature as the most variable features. Variables like vapor pressure deficit and wind components showed minimal fluctuation. The precipitation variable displayed high spikes, likely due to storm events, which are important stressors in crop growth.

Visualizations revealed patterns that confirmed our assumptions: daily max temperature trends aligned with heat events, and precipitation spikes corresponded with low yield observations. The crop yield histogram showed a concentration around 130–140 bushels per acre with fewer high-yield counties.

Correlation analysis revealed that temperature and solar radiation positively influence yield, while high humidity and excessive wind may have negative or negligible effects. The correlation matrix helped pinpoint environmental factors most worth investigating in the next stages, such as radiation flux and VPD.

Moving forward, I want to explore how specific extreme events (e.g., consecutive hot days or storms) affect yield and examine lag effects (e.g., conditions in March affecting harvest in September). These insights can improve crop forecasting models and guide adaptive farming strategies under climate uncertainty.