1. Introduction

This document serves as the Data Appendix for the time series project "Spring Arrival Weather Data". It provides a structured overview of the datasets and transformations applied, from raw input data to analysis data.

2. Data Pipeline Workflow

Step 1: Input Data Files

Charlottesville Weather Data (charlottesville_weather.csv)

- **Unit of Observation:**
  Each row represents a **daily weather observation** in Charlottesville, Virginia.

- **Key Variables:**
  - time – The date of the weather observation
  - Precipitation (mm) – Daily precipitation in millimeters
  - Wind Direction (°) – Wind direction in degrees
  - Wind Speed (km/h) – Wind speed in kilometers per hour
  - Pressure (hPa) – Atmospheric pressure in hectopascals
  - Avg Temp (°F) – Average daily temperature (converted from Celsius)
  - Min Temp (°F) – Minimum daily temperature (converted from Celsius)
  - Max Temp (°F) – Maximum daily temperature (converted from Celsius)
- **Purpose:**
  This **raw dataset** is the primary source of daily weather observations used to investigate winter temperature patterns and define whether Charlottesville reaches 60°F by March 20.

- **Processing Steps:** Data was acquired from a weather API (e.g., Meteostat), columns were renamed, temperatures were converted from Celsius to Fahrenheit, then the file was sorted by date and saved as charlottesville_weather.csv.

Step 2: Cleaned Data

Cleaned Weather Data (cleaned_charlottesville_weather.csv)

- **Unit of Observation:**
  Each row is still a **daily weather observation** in Charlottesville, but now cleaned and preprocessed.

- **Key Variables:**
  - time – Date of observation (sorted)
  - Precipitation (mm), Wind Direction (°), Wind Speed (km/h), Pressure (hPa) – Same as raw data, but cleaned
  - Avg Temp (°F), Min Temp (°F), Max Temp (°F) – Temperature columns, with missing values interpolated
- **Purpose:**
  This intermediate dataset ensures missing values are handled (via interpolation/backfilling) and that data is sorted and ready for further aggregation.

- **Processing Steps:** Missing values in the raw daily data were addressed via linear interpolation and backfill, and the cleaned, sorted dataset was saved as cleaned_charlottesville_weather.csv.

Step 3: Analysis Data

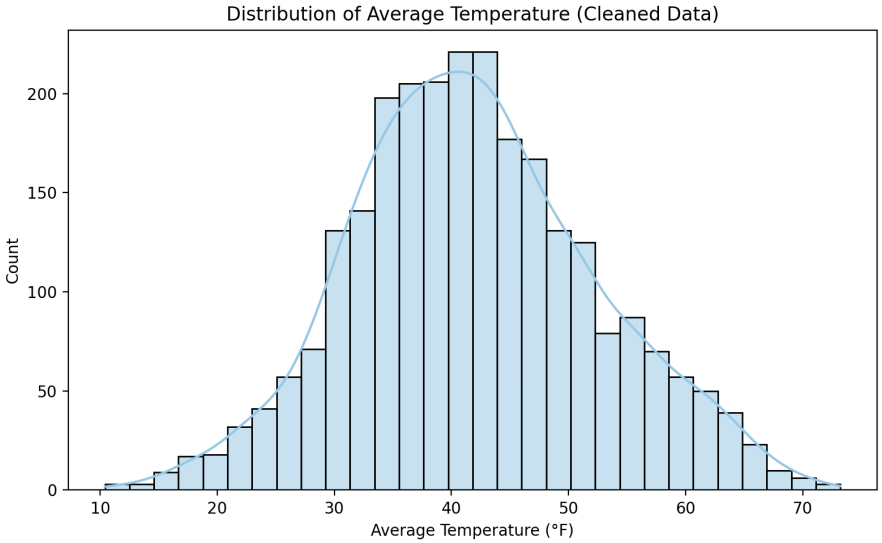
Logistic Regression Data (merged_charlottesville_weather.csv)

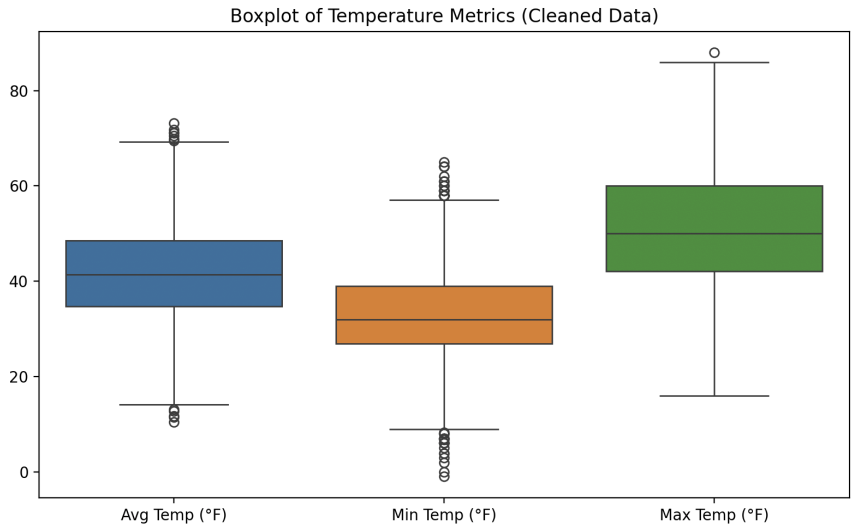
- **Unit of Observation:**
  Each row corresponds to **one winter season** (December through March).
  - For December, the year remains the same.
  - For January, February, and March, the season is assigned to the previous December's year.
- **Key Variables:**
  - Season_Year – Identifier for the winter season (e.g., 2024 if it's December 2024 to March 2025)
  - AvgTemp_Mean, AvgTemp_Min, AvgTemp_Max, AvgTemp_Std – Aggregated temperature statistics
  - Prcp_Sum, Prcp_Mean – Total and average precipitation over the winter months
  - Pressure_Mean – Average atmospheric pressure for the season
  - WindSpeed_Mean – Average wind speed
  - Spring_Arrival – Binary variable: 1 if any day on or before March 20 reached ≥60°F, else 0
- **Purpose:**
  This dataset is used for **logistic regression analysis** to predict whether Charlottesville

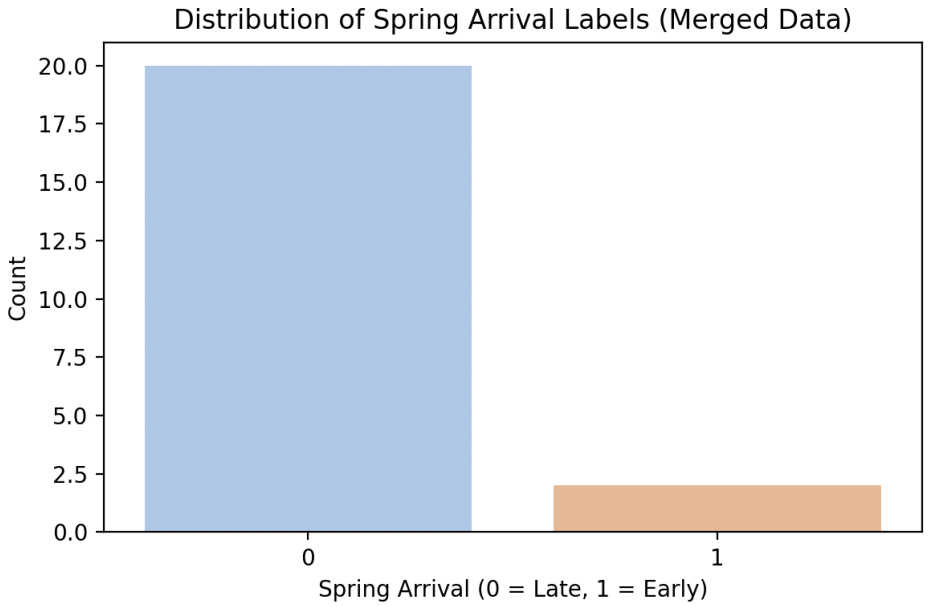
reaches 60°F by March 20 based on aggregated winter weather features.

- **Processing Steps:** The cleaned data was aggregated by winter season (December–March) to compute summary statistics (mean, min, max, std, etc.), a binary target was defined (≥60°F by March 20 or not), and the resulting season-level dataset was saved as merged_charlottesville_weather.csv.

3. Summary Statistics and Visualizations







Summary Statistics for Weather Features (Merged Data)

	count	mean	std	min	max
AvgTemp_Mean	22.0	41.84	2.99	37.47	46.4
AvgTemp_Min	22.0	18.96	4.84	10.4	28.76
AvgTemp_Max	22.0	67.08	3.74	59.72	73.22
AvgTemp_Std	22.0	10.05	1.15	8.08	12.04
Prcp_Sum	22.0	144.16	103.9	0.0	373.0
Prcp_Mean	22.0	1.21	0.88	0.0	3.06
Pressure_Mean	22.0	1018.7	1.27	1015.43	1020.68
WindSpeed_Mean	22.0	8.52	0.87	6.97	9.93