This project consists of three parts: **Basic EDA, Assignment Solution, and Model Building(XG Booster).**

To load the dataset, I used Pandas and imported both CSV files: one containing temperature data and the other containing bin size information. I then merged both CSV files on the ID column to combine relevant data.After merging, I performed basic exploratory data analysis (EDA). The temperature values were recorded in tenths of degrees Celsius, so I converted them back to Celsius by dividing by 10. I then extracted the year, month, and day from the date column to facilitate further analysis.  
Next, I examined the distribution of temperature, latitude, longitude, and elevation to understand the dataset better and identify potential patterns or anomalies.

The mean temperature is 9.5°C, with a high standard deviation of 12.35°C, indicating significant seasonal variation. The temperature ranges from -34.3°C to 40.6°C, a difference of over 70°C, reflecting extreme fluctuations.Regarding latitude, the mean is approximately 42.1°N, with all stations clustered between 41.56°N and 42.44°N. The small standard deviation of 0.2 suggests that the locations are situated within a narrow geographical range, indicating proximity between them.The minimum temperature for February is -10°C, while the maximum temperature recorded is 0°C.

After analyzing the relationship between latitude and temperature, an interesting pattern emerged. Typically, temperature decreases with increasing elevation, which is expected. In lower elevations (260m - 280m), this trend holds true. However, beyond 280m, the temperature starts to rise instead of continuing to drop. This unusual behavior could be due to warm air getting trapped above cooler air, creating an inversion layer that acts like a blanket, preventing heat from escaping.

After that, I summarized the dataset by identifying locations with the highest and lowest temperatures and ordering them accordingly. Toledo and Metcalf recorded the highest temperature at 40.6°C. Milan4Ese has an average temperature of 10.883°C. Temperature fluctuations across seasons are evident, showing periods of increase and decrease. This insight can help in planning flight services based on temperature variations.

The analysis starts by removing leap days to ensure consistency in yearly comparisons, keeping the number of days the same across all years. Historical data from 2005 to 2014 is then extracted to determine temperature records for this period. The dataset is grouped by month and day to calculate daily historical highs and lows, identifying the highest recorded temperature (TMAX) and the lowest recorded temperature (TMIN) for each day. A continuous date column is created to facilitate plotting. A line plot is generated to visualize temperature trends, with the red line representing record high temperatures and the blue line representing record lows. To improve readability, the area between these lines is shaded in grey, highlighting the daily temperature range and making seasonal variations more visible. The x-axis is formatted to show months instead of specific dates, making it easier to track temperature changes throughout the year. Labels and a legend are added to provide context to the visualization. This analysis reveals clear seasonal temperature patterns, with high temperatures peaking in the middle of the year (summer) and low temperatures dropping at the beginning and end (winter). The highest recorded temperature is 43°C, while the lowest is -31°C.

The dataset is grouped by month and day to calculate daily historical highs and lows, identifying the highest recorded temperature (TMAX) and the lowest recorded temperature (TMIN) for each day. To facilitate analysis, the data from 2005 to 2014 is consolidated into a single representative year, which is defined as 2013.

This analysis extends the historical temperature study by identifying outliers in 2015 compared to records from 2005 to 2014. The historical data is first filtered for the years 2005 to 2014, and daily record high (TMAX) and low (TMIN) temperatures are computed for each day of the year. To maintain consistency in plotting, all dates are mapped to the year 2013.

The 2015 data is then processed separately and merged with historical records to compare daily temperatures. Outliers are identified where 2015 temperatures exceed previous record highs or fall below record lows. These extreme values indicate new high or low temperatures recorded in 2015.

The visualization includes a line plot of historical record highs and lows, with a shaded region between them to show the typical temperature range. Outliers from 2015 are highlighted using scatter points—orange for new record highs and green for new record lows. The x-axis is formatted to display months, making seasonal patterns and anomalies easily interpretable. This analysis helps identify significant climate variations and trends over time.

Some analysis are, In 2015 the region experienced its lowest recorded temperatures. Additionally, the highest temperature ranged between 33.5-34°C, which did not surpass the average summer temperature from 2005-2014. This indicates a temperature drop of 7°C during the summer season. Further analysis suggests that this change was influenced by environmental factors(Polar Vortex,Lake Effect Snow)

**Polar Vortex (Super Cold Winter)**   
- A polar vortex is when super cold air from the Arctic moves south. This can lead to extremely low temperatures in winter.

The **Great Lakes** can create lake effect snow, which happens when cold air picks up moisture from the lakes and dumps heavy snow. Report said it was strong during 2015

**1. Familiarize yourself with the dataset, then write some python code which returns a line graph of the record high and record low temperatures by day of the year over the period 2005-2014. The area between the record high and record low temperatures for each day should be shaded.**

This analysis extends the previous temperature study by incorporating data from both 2015 and 2016 to determine whether 2015 was an anomaly. First, temperature data is cleaned by removing leap days for consistency. The dataset is then merged with bin size information and converted into appropriate date formats. Historical temperature records (2005-2014) are computed by identifying the highest (TMAX) and lowest (TMIN) recorded temperatures for each day of the year. For uniform visualization, all dates are mapped to 2012.

Next, 2015 and 2016 temperature data are separately merged with historical records to identify outliers—instances where temperatures exceeded previous highs or dropped below previous lows. These outliers are extracted and labeled by year for comparison. The visualization plots historical high and low records as red and blue lines, respectively, with a shaded region in between. Any new record highs or lows in 2015 and 2016 are marked with distinct colors—orange for 2015 and blue for 2016. I have used T test for this to compare means of both years

Upon comparison, no new record-breaking temperatures are found in 2016, confirming that the extreme temperature fluctuations in 2015 were not part of an ongoing trend. Instead, they likely resulted from specific environmental factors such as the Polar Vortex or Lake Effect Snow, making 2015 an anomaly rather than part of a long-term pattern.

**Handling leap years**

Leap days (February 29th) can create inconsistencies when analyzing historical temperature records since they occur only once every four years. To ensure uniformity across the dataset and maintain a consistent comparison between years, these points were removed. This was done using a filtering step where rows with dates corresponding to February 29th were excluded.

**Legends**

To make the visualization clear and informative, proper legends, axis labels, and titles were added. The legend differentiates between temperature records for different stations and years, while labels on the x-axis and y-axis specify the date and temperature (°C), respectively. Additionally, a title summarizing the dataset and location provides context. Gridlines and reference lines were included for readability, and unnecessary chart elements (chart junk) were minimized to avoid distractions.

**5. The data you have been given is near \*\*Ann Arbor, Michigan, United States\*\*, and visualize on map the stations the data.**

The map visualization plots weather stations in Ann Arbor using Folium and MarkerCluster, allowing for an interactive view of their locations and elevations. The base map is generated using Stamen Terrain tiles, providing a clear topographic representation. Each station is marked with a blue icon, displaying its name and elevation when clicked.

Ann Arbor is classified under Category 5 in the dataset. The recorded elevations of its different stations are as follows:

* **ANN ARBOR U OF MICH:** 274.3m
* **ANN ARBOR 1W:** 263.7m
* **ANN ARBOR SE:** 253.6m
* **ANN ARBOR MUNI AP:** 255.7m

Notably, ANN ARBOR 1W and ANN ARBOR MUNI AP have nearly identical elevations, and both belong to Category 2 in the dataset. This analysis helps in understanding the geographical distribution and elevation differences of weather stations in the region.

**6. Plot Temperature Summary near Ann Arbor, Michigan, United States (Year 2015).**

The temperature profile for Ann Arbor, MI stations in 2015 highlights daily high (TMAX) and low (TMIN) temperatures across three key locations: ANN ARBOR U OF MICH, ANN ARBOR MUNI AP, and ANN ARBOR SE. Solid lines represent the daily high temperatures, while dashed lines indicate the daily lows, with different colors distinguishing the stations. To provide a seasonal climate context, a blue dotted line marks the average low temperature for January, while a red dotted line represents the average high temperature for July. This helps in understanding how temperatures fluctuated throughout the year and whether any anomalies occurred.

The temperature difference could be due to the contrast between urban and rural areas. Ann Arbor Municipal Airport and Michigan are urban locations, which tend to be warmer due to the presence of buildings and reduced vegetation. Additionally, traffic and vehicle emissions contribute to heat buildup, which might explain a slight increase in temperature.

Another possible factor is wind exposure—some areas may be more affected by winds, which can influence temperature variations. Additionally, ocean currents transport heat from the equator toward the poles, which could also play a role in these temperature differences.

**summary is given below**

2015 Temperature Summary for Ann Arbor Stations

Warmest Day: 33.3°C (2015-07-29)

Coldest Night: -34.3°C (2015-02-20)

January Average Low: -10.6°C

July Average High: 27.6°C

**Model Building**

In the Model Building phase, both XGBoost and Linear Regression were used. Among these, XGBoost outperformed Linear Regression, achieving an accuracy of over 72%. Feature engineering was applied by removing certain columns, extracting the month from the date column, and converting it into seasons. For Linear Regression, values were standardized, and multicollinearity was checked using VIF. Columns like longitude and elevation, which showed high collinearity, were removed. Skewness in temperature was minimal and acceptable. One-hot encoding was applied to the TMAX and Season columns.

After analyzing the data, it was evident that there was no linear relationship between the dependent and independent variables, which led to Linear Regression achieving an accuracy of around 47%, indicating underfitting. Ridge regression produced similar results. To address this, an ensemble method like XGBoost was implemented, which improved the model’s performance significantly, achieving an accuracy of around 72%. This demonstrated that the model generalizes well on both training and testing data, and XGBoost effectively reduced bias in the model.

**Future Plans (Deferred Due to Time Constraints):**

Moving forward, I plan to conduct a more in-depth statistical analysis to gain further insights. I aim to apply **ANOVA (Analysis of Variance)** to examine the variability in temperature across different Ann Arbor, Michigan, stations, helping to determine whether significant differences exist between them. Additionally, I will use the **Chi-Square test** to explore associations between categorical variables, such as station locations and extreme weather occurrences.