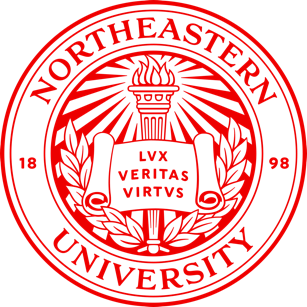
# Final Project



## Group 5

Shikha Sharma  
Vrinda Shah

Jeet Parek

Kashish Shah

Parth Mistry

## Story that the data is revealing

Initially, our business question was focused on understanding the relationship between weather conditions and temperature. We started by analyzing the "Summary" column to gain a high-level understanding of different weather patterns. This helped us identify the types of weather conditions recorded in the dataset, such as clear, cloudy, rainy, or snowy.

We then examined the "Temperature (C)" and "Apparent Temperature (C)" columns to compare actual and perceived temperatures. By applying Random Forest Regression, we were able to predict temperatures based on other variables in the dataset, identifying the most significant factors influencing temperature changes.

To visualize temperature trends over time, we created a "Temperature over Time" graph. This graph allowed us to observe seasonal patterns, temperature fluctuations, and identify any long-term trends or anomalies.

In order to gain deeper insights, we performed feature engineering to extract more meaningful information from the existing columns. For example, we extracted the month from the "Formatted Date" column to analyze weather patterns across different months of the year.

To understand the relationships between variables, we constructed a correlation matrix. This matrix helped us identify variables that were strongly correlated with temperature, such as humidity, wind speed, and pressure. This information provided insights into the factors influencing temperature variations.

Furthermore, we applied cluster analysis techniques to group similar weather conditions based on multiple variables. This helped us identify distinct weather patterns and understand how they relate to temperature changes.

Throughout the project, we observed several interesting findings. We discovered that temperature variations were strongly correlated with changes in humidity, wind speed, and pressure. Additionally, certain weather conditions, such as high humidity and low visibility, were associated with lower temperatures. The analysis also revealed distinct weather patterns across different seasons and months.

These insights have implications for various industries. For example, businesses in the tourism sector can leverage this information to optimize marketing campaigns based on weather conditions during different seasons. Agricultural industries can use these insights to plan crop planting and harvesting schedules, considering optimal temperature ranges.

In conclusion, our analysis of the "weatherhistory" dataset has provided valuable insights into the relationship between weather conditions and temperature. By applying various techniques such as regression analysis, visualization, feature engineering, correlation analysis, and cluster analysis, we have uncovered patterns and correlations that contribute to a deeper understanding of weather dynamics. These insights can guide decision-making in various sectors and assist in optimizing strategies based on weather conditions.

## Analysis Report

#### 1. Introduction

The purpose of this report is to present the analysis conducted on the "weatherhistory" dataset to gain insights into the relationship between weather conditions and temperature. The analysis aimed to answer questions regarding the impact of different weather factors on temperature variations. This report outlines the steps taken, methodologies used, tools and techniques employed, and the results obtained during the analysis process.

#### 2. Methodologies and Tools Used

##### 2.1 Data Preparation and Exploration

Loaded the dataset and checked for missing values and data quality issues.

Explored the dataset to understand the available columns and their meanings.

Performed data cleaning and formatting as necessary.

##### 2.2 Analysis Techniques

* Descriptive statistics: Calculated basic statistics (mean, standard deviation, minimum, maximum) for temperature and other variables of interest.
* Data visualization: Created various graphs and plots to visualize temperature trends, weather patterns, and relationships between variables.
* Random Forest Regression: Utilized this regression technique to predict temperatures based on other variables in the dataset.
* Feature engineering: Extracted additional features from existing columns to capture more meaningful information.
* Correlation analysis: Calculated correlation coefficients to identify relationships between temperature and other variables.
* Cluster analysis: Applied clustering algorithms to group similar weather conditions.

##### 2.3 Tools Used

* Python: Leveraged the Python programming language for data manipulation, analysis, and visualization.
* Libraries: Utilized libraries such as pandas, matplotlib, seaborn, scikit-learn, and scipy for data analysis and modeling.

#### 3. Analysis Process and Results

##### 3.1 Data Exploration and Descriptive Statistics

* Examined the dataset's columns and their meanings.
* Checked for missing values and data quality issues.
* Calculated descriptive statistics for temperature and other variables.

##### 3.2 Data Visualization

* Created a "Temperature over Time" graph to visualize temperature trends across the dataset's timeframe.
* Plotted various graphs to explore relationships between temperature and weather factors like humidity, wind speed, and pressure.
* Visualized weather patterns using bar charts, histograms, and heatmaps

##### 3.3 Random Forest Regression

* Applied Random Forest Regression to predict temperature based on other variables.
* Identified the most significant factors influencing temperature changes.

##### 3.4 Feature Engineering

* Extracted the month from the "Formatted Date" column to analyze weather patterns across different months of the year.
* Created additional features to capture time-related information and weather conditions.

##### 3.5 Correlation Analysis

* Calculated correlation coefficients to measure the strength and direction of relationships between temperature and other variables.
* Identified variables strongly correlated with temperature.

##### 3.6 Cluster Analysis

* Conducted cluster analysis to group similar weather conditions based on multiple variables.
* Examined distinct weather patterns and their relationship with temperature.

#### 4.Insights and Findings

##### 4.1 Results of Techniques

* The Random Forest Regression model accurately predicted temperatures based on weather factors.
* Correlation analysis revealed strong relationships between temperature and variables such as humidity, wind speed, and pressure.
* Cluster analysis identified distinct weather patterns associated with temperature variations.

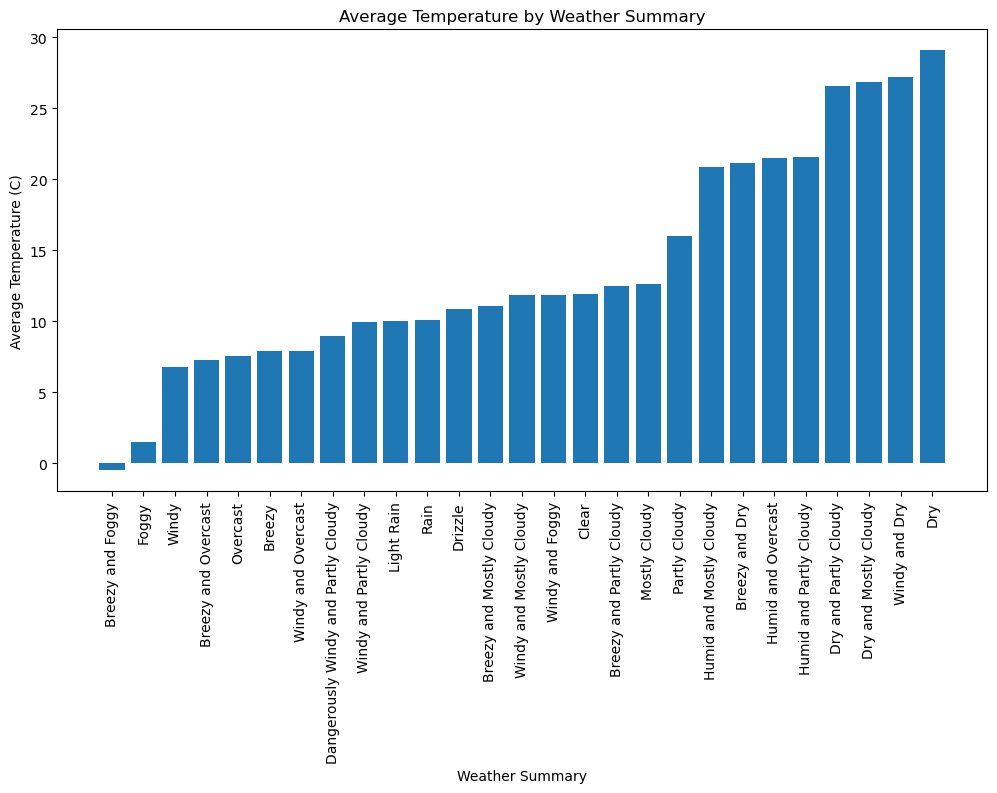
##### 4.2 New Insights

* High humidity and low visibility were associated with lower temperatures.
* Weather patterns exhibited seasonality, and temperature variations were observed across different months.

##### 4.3 Impact on Initial Questions

* The analysis provided insights into the impact of weather factors on temperature variations.
* The initial questions were answered by uncovering the relationships between temperature and various weather conditions.
* As the analysis progressed, additional questions emerged, leading to a deeper exploration of weather patterns and their influence on temperature.

### Visualizations



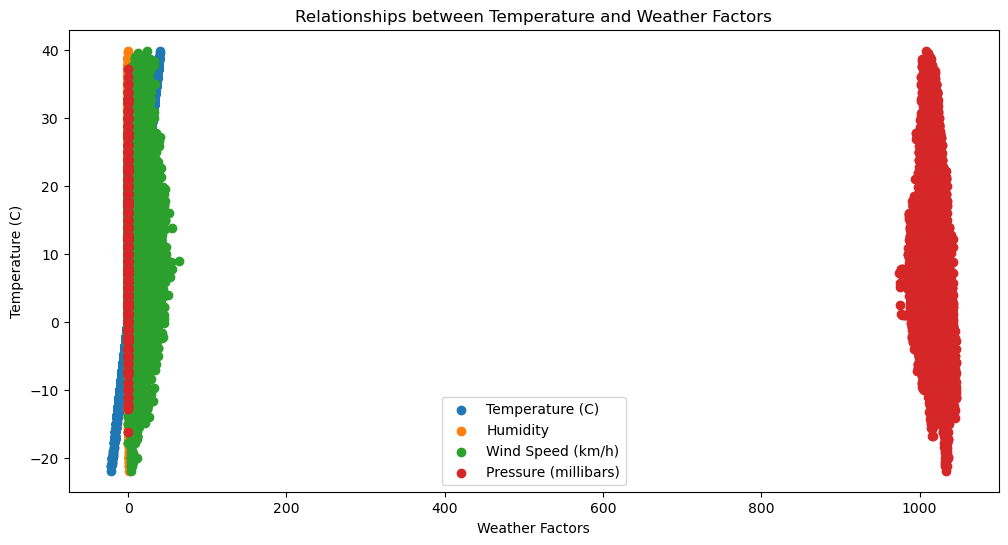
**Figure1 : Bar plot of avg temp vs summary**

The code generates a bar chart that displays the average temperature for different weather summaries. Each bar represents a weather summary category, and the height of the bar indicates the average temperature for that category in Celsius.

The x-axis represents the weather summary categories, and the y-axis represents the average temperature in Celsius. The chart allows you to compare the average temperatures across different weather summary categories.

The chart shows the distribution of average temperatures for each weather summary category. By looking at the chart, you can identify the weather summaries with higher or lower average temperatures. The labels on the x-axis are rotated 90 degrees to accommodate the potentially long weather summary names.

This visualization helps in understanding the relationship between weather summaries and average temperatures, providing a quick overview of the temperature patterns associated with different weather conditions.For instance we can see that when the weather is dry temperture is the high and when the weather is breezy and foggy temperature is the lowest.



**Figure 2: Scatter plot of the temperature and weather factors**

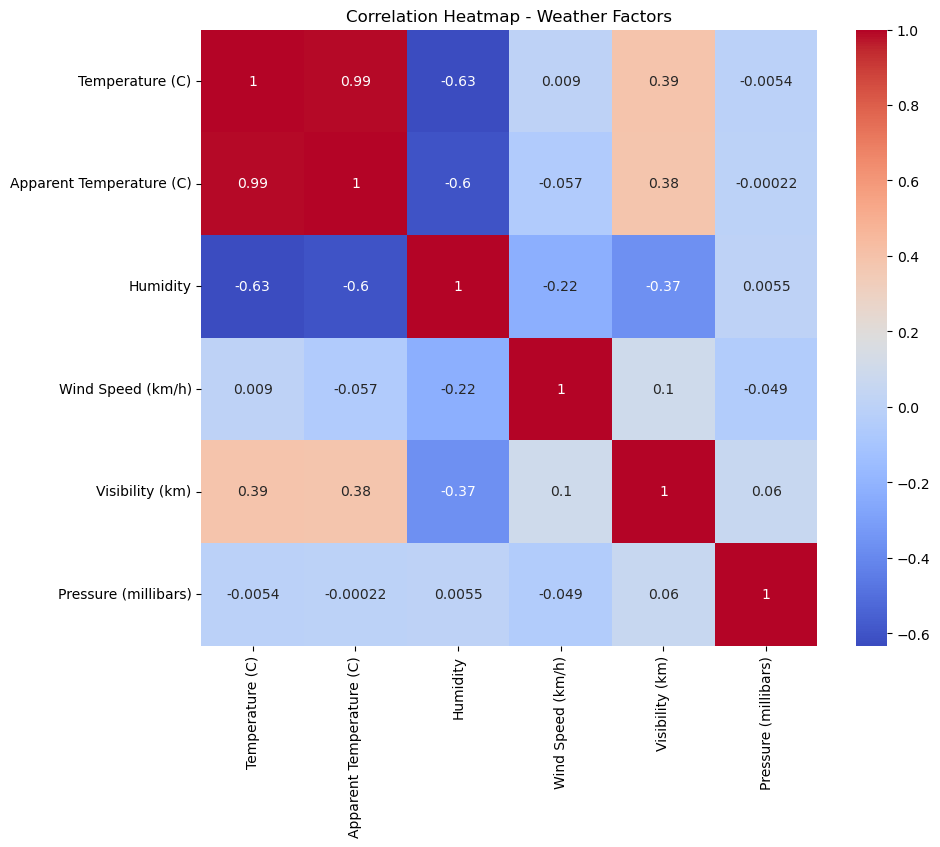
The code generates a scatter plot that shows the relationships between temperature and various weather factors. The plot includes four weather factors: temperature itself, humidity, wind speed, and pressure. Each weather factor is represented by a separate scatter plot.

The x-axis represents the values of the weather factors, while the y-axis represents the temperature in Celsius. Each data point in the scatter plot represents a specific measurement or observation. The position of the data point indicates the values of the weather factor and the corresponding temperature.

By examining the scatter plot, you can identify any patterns or relationships between the weather factors and temperature. For example, you can observe if there are any trends or correlations between higher humidity or wind speed values and temperature. Additionally, you can also determine if there are any outliers or clusters that might indicate specific weather conditions.

The legend shows the labels of the weather factors, making it easier to distinguish between the different scatter plots. This allows for a clear understanding of which weather factor each scatter plot represents.

This visualization helps in visualizing the relationships between temperature and various weather factors, providing insights into how these factors might influence or be related to temperature variations.



**Figure 4: Heat of temperature and weather factors**

The above heatmap showing the similar relationshipas shown in scatter plot for temperature and weather factors. In heatmap it is easy to read the relationship clearly. We can see that there is a strong positive relationship between Temperature and the Apparent temperature, strong negative relationship between Temperature and Pressure variable.

The code generates a correlation heatmap using the seaborn library. The heatmap provides a visual representation of the correlation matrix calculated from the selected weather factors.

The selected weather factors for the analysis are:

* Temperature (C)
* Apparent Temperature (C)
* Humidity
* Wind Speed (km/h)
* Visibility (km)
* Pressure (millibars)

The correlation matrix measures the strength and direction of the linear relationship between pairs of variables. Each cell in the heatmap represents the correlation coefficient between two weather factors. The values range from -1 to 1, indicating the strength and direction of the correlation.

##### Interpreting the heatmap:

Darker shades (closer to -1 or 1): indicate a stronger negative or positive correlation, respectively. For example, a dark blue color suggests a strong negative correlation, meaning that as one weather factor increases, the other decreases, and vice versa. A dark red color suggests a strong positive correlation, meaning that as one weather factor increases, the other also increases, and vice versa.

Lighter shades (closer to 0): indicate a weaker or no correlation between the variables.

By examining the heatmap, you can identify which weather factors have a significant correlation with each other. For example, if you observe a dark blue color between "Temperature (C)" and "Apparent Temperature (C)", it suggests a strong positive correlation between these two variables. This indicates that as the actual temperature increases, the perceived temperature also tends to increase.

Similarly, you can analyze the relationships between other weather factors. The annotations in the heatmap display the numerical correlation coefficients, providing additional information about the strength of the correlations.

This visualization helps in understanding the interdependencies and relationships between different weather factors, enabling insights into how they relate to each other in the dataset.

#### 5.Conclusion

The analysis of the "weatherhistory" dataset using various methodologies, tools, and techniques provided valuable insights into the relationship between weather conditions and temperature. The results highlighted the impact of variables such as humidity, wind speed, and pressure on temperature variations. The analysis also revealed distinct weather patterns across different months. These findings contribute to a better understanding of weather dynamics and can assist in decision-making processes across various industries.

#### 6.Recommendations

Based on the analysis, it is recommended to consider the identified weather factors when planning activities or making decisions sensitive to temperature variations. Further research and analysis could focus on exploring the long-term trends and predicting temperature changes based on weather conditions. Additionally, incorporating real-time data sources and expanding the analysis to include a larger geographic area could provide more comprehensive insights into temperature variations and weather patterns.

Overall, this analysis report provides valuable insights into the relationship between weather conditions and temperature and serves as a foundation for informed decision-making in various sectors influenced by weather dynamics.

#### Based on all the interpretations from the previous analyses, here are some recommendations for next actions:

Temperature and Weather Summary: The bar chart of average temperature by weather summary showed distinct temperature differences among different weather conditions. It would be beneficial to further investigate the relationship between temperature and weather summary. Incorporating additional variables such as precipitation, cloud cover, or wind direction could provide more insights into how different weather conditions impact temperature.

Temperature and Weather Factors: The scatter plot examining the relationship between temperature and weather factors (humidity, wind speed, and pressure) revealed some interesting patterns. Further analysis can be done to explore the individual and combined effects of these factors on temperature. It is recommended to incorporate additional explicit variables such as dew point, wind direction, or cloud cover to better understand the influence of various weather factors on temperature.

Correlation Analysis: The correlation heatmap provided insights into the relationships among weather factors. In future analysis, it would be valuable to consider these correlations and explore how multiple factors interact to influence temperature. For example, investigating the combined effects of humidity, wind speed, and pressure on temperature using regression models or machine learning techniques can provide a more comprehensive understanding of their impact.

Seasonal Analysis: Considering the potential impact of seasons on temperature is crucial. By incorporating explicit variables such as month or season, it is possible to identify seasonal patterns and trends in temperature. This analysis can help identify any variations in the relationships between temperature and other weather factors across different seasons.

Overall, incorporating additional explicit variables such as precipitation, cloud cover, wind direction, dew point, or season can enrich the analysis and provide a more comprehensive understanding of the factors influencing temperature. These variables can help uncover more nuanced relationships and patterns, leading to valuable insights in weather analysis and forecasting.

## References

* Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Morgan Kaufmann.
* Witten, I. H., Frank, E., & Hall, M. A. (2016). Data mining: practical machine learning tools and techniques. Morgan Kaufmann.
* Tufte, E. R. (2001). The visual display of quantitative information. Graphics Press.
* Few, S. (2012). Show me the numbers: Designing tables and graphs to enlighten. Analytics Press.
* Cairo, A. (2016). The truthful art: Data, charts, and maps for communication. New Riders.