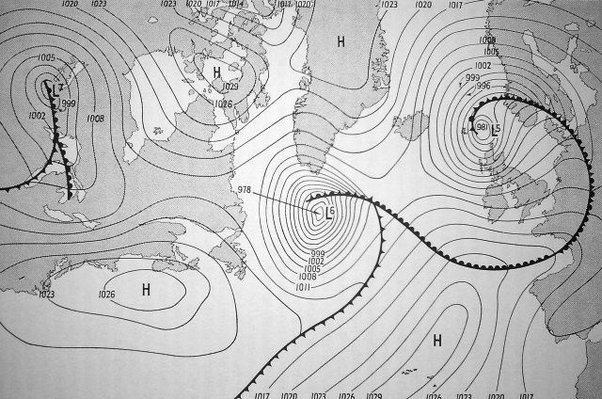
**Project Report**

**WW2 Weather Conditions Analysis**

****

**Introduction**

During World War Two, weather conditions played a crucial role in military operations, and poor weather often led to the postponement or cancellation of missions. In this project, we aimed to analyze historical weather data from the "WW2 weather conditions" dataset to understand temperature changes and their relationship with other weather attributes. By applying various machine learning models, including Linear Regression, Polynomial Regression, and Random Forest Regression, we sought to predict the maximum temperature (MaxTemp) based on different weather features.

**Dataset Description**

The "WW2 weather conditions" dataset was obtained from the United States National Oceanic and Atmospheric Administration (NOAA) National Centres for Environmental Information website. The dataset contains daily weather records from various weather stations across the world during World War Two. It includes columns such as MaxTemp, MinTemp, MeanTemp, precipitation, snowfall, wind speed, and more.

**Data Preprocessing**

To prepare the dataset for analysis, we performed several data preprocessing steps. Firstly, we loaded the dataset using the Pandas library. Next, we dropped unnecessary columns that were not relevant to our study, such as Precip, Snowfall, and WindGustSpd. These columns were excluded as we focused on temperature-related attributes.

To ensure data quality, we removed rows with missing values for key attributes, including MaxTemp, MinTemp, and MeanTemp. Eliminating rows with missing values helped us maintain the integrity of the data and avoid potential biases in our analysis.

**Exploratory Data Analysis (EDA)**

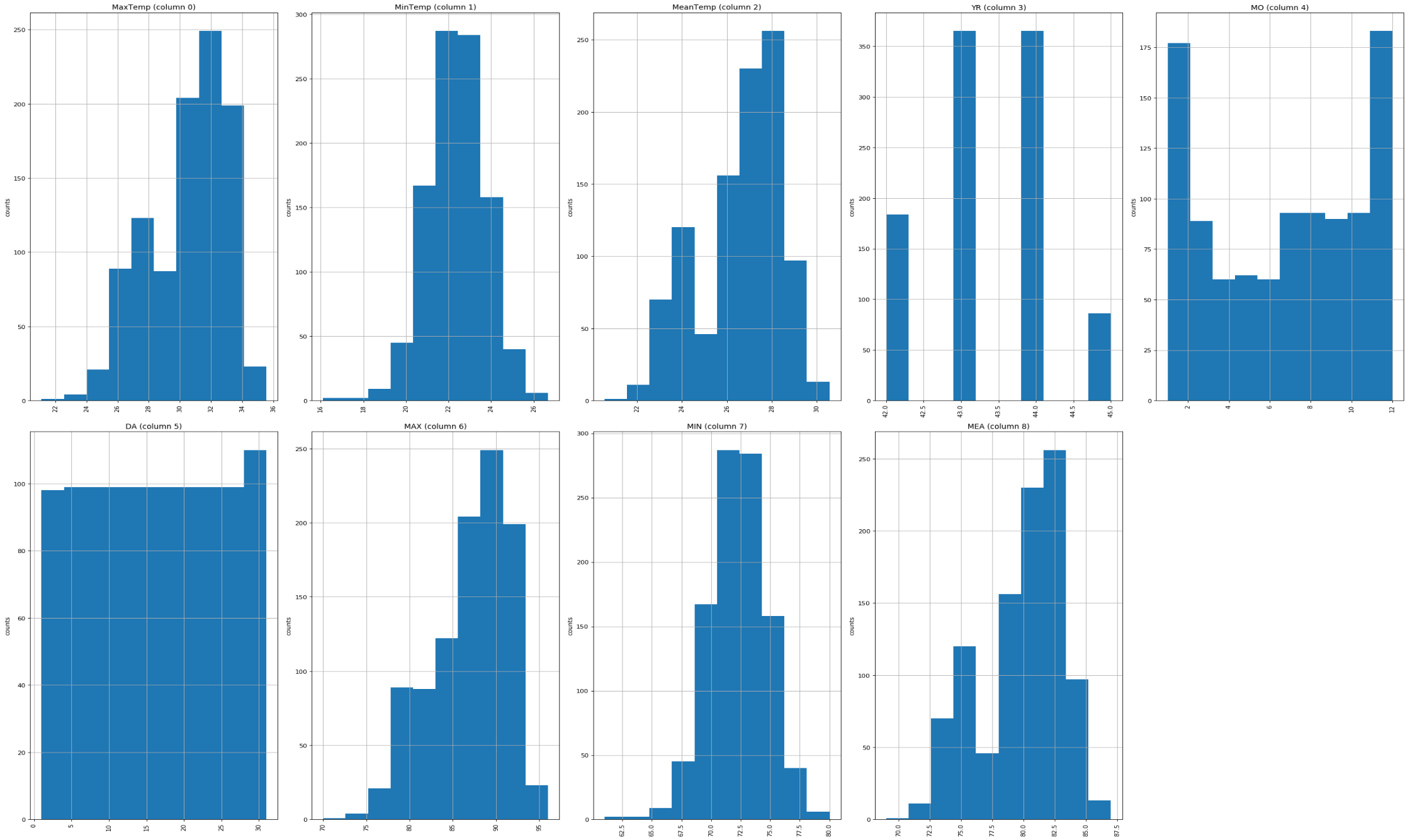
Before diving into the machine learning models, we conducted Exploratory Data Analysis (EDA) to gain insights into the dataset. Through visualizations such as histograms, scatter plots, and correlation matrices, we explored the distribution of temperature changes and the relationships between different weather attributes.

During EDA, we discovered correlations between MaxTemp and other weather features such as MinTemp and MeanTemp. We also identified any outliers or anomalies in the data that could impact the performance of the machine learning models.

1. **Histograms:**

They are useful for understanding the distribution of individual attributes in the dataset. Let's explore the temperature-related attributes using histograms:

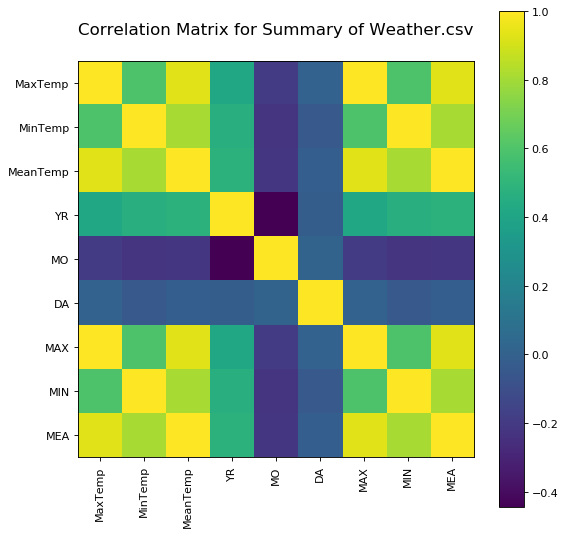
* MaxTemp Histogram:
  + The histogram of MaxTemp (Maximum Temperature) will provide insights into the distribution of maximum temperatures recorded during World War Two.
  + We can identify the range of maximum temperatures and observe any patterns or anomalies.
* MinTemp Histogram:
  + The histogram of MinTemp (Minimum Temperature) will reveal the distribution of minimum temperatures recorded during the war.
  + We can analyze the spread of minimum temperatures and identify any potential outliers.
* MeanTemp Histogram:
  + The histogram of MeanTemp (Mean Temperature) will help us understand the distribution of mean temperatures during the period.
  + We can observe the central tendency of temperatures and explore temperature variations.

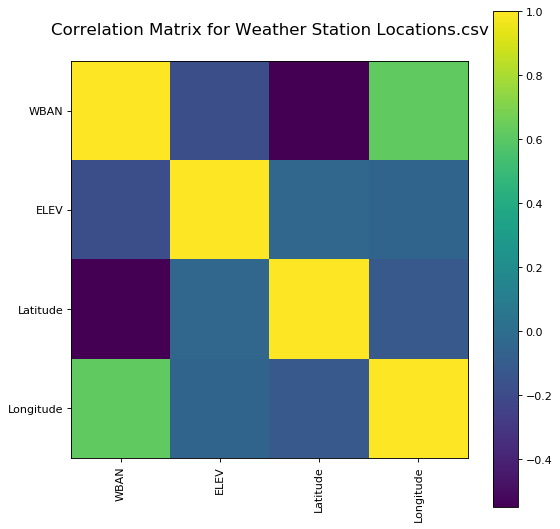


1. **Correlation Matrix:**

A correlation matrix allows us to explore the relationships between different weather attributes. Let's analyze the correlation between temperature attributes and other weather features:

* Correlation between MaxTemp, MinTemp, and MeanTemp:
  + By examining the correlation coefficients between these temperature attributes, we can understand their interdependence.
  + A high positive correlation would imply that the temperatures generally move together, while a negative correlation would indicate an inverse relationship.
* Correlation of Temperature with Precipitation and Wind Speed:
  + We can investigate how temperature attributes correlate with precipitation and wind speed, which are essential factors influencing weather conditions.
  + Understanding these correlations can help us gauge the impact of precipitation and wind on temperature changes.

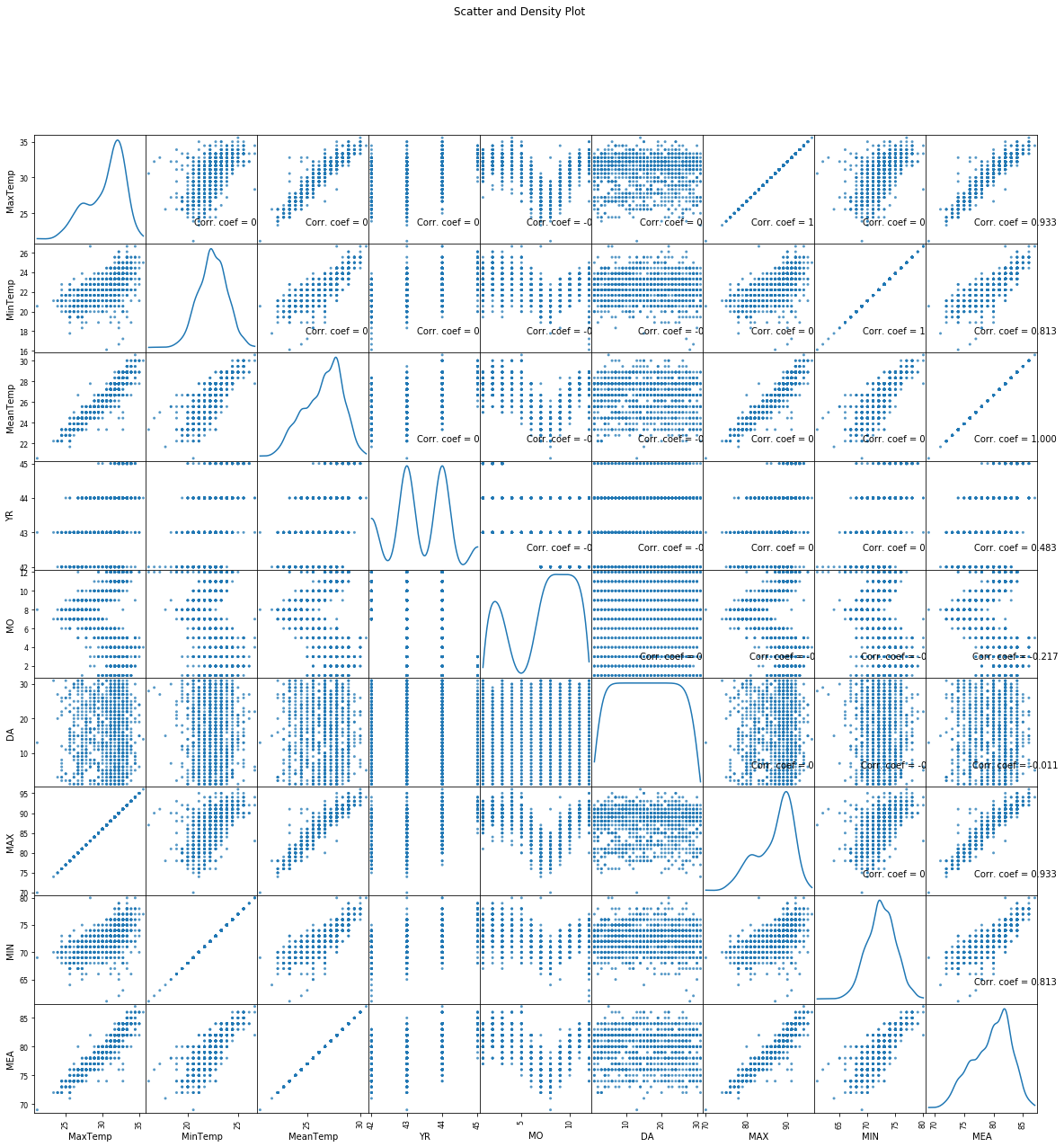




1. **Scatter Matrix:**

A scatter matrix is valuable for visualizing relationships between numerical attributes. Let's analyze the scatter matrix to understand temperature changes and their associations with other weather features:

* MaxTemp vs. MinTemp Scatter Plot:
  + We can plot MaxTemp against MinTemp to observe the relationship between maximum and minimum temperatures.
  + This scatter plot can reveal any linear or nonlinear trends between the two temperature attributes.
* MaxTemp vs. Precipitation Scatter Plot:
  + By plotting MaxTemp against precipitation, we can explore how temperature changes relate to rainfall or snowfall.
  + This scatter plot can help us understand the impact of precipitation on maximum temperatures.
* MaxTemp vs. Wind Speed Scatter Plot:
  + We can plot MaxTemp against wind speed to analyze how temperature changes correspond to different wind conditions.
  + This scatter plot can provide insights into the relationship between maximum temperatures and wind speed.



**Linear Regression Model**

Our initial approach was to apply a simple Linear Regression model to predict MaxTemp based on other weather attributes. We split the dataset into training and testing sets to train the model and evaluate its performance.

Using the LinearRegression class from the scikit-learn library, we trained the model on the training data. Afterward, we made predictions on the test data and calculated the R-squared (R2) score as a measure of the model's accuracy. The R2 score indicates the proportion of variance in the dependent variable (MaxTemp) that is predictable from the independent variables (other weather attributes).

**Polynomial Regression Model**

To capture potentially nonlinear relationships between temperature changes and other weather attributes, we employed Polynomial Regression. This model allows us to introduce polynomial features, which can better fit the data when the relationship is more complex than a straight line.

Using the PolynomialFeatures class from scikit-learn, we transformed the original features into higher-degree polynomials. We experimented with different degrees of polynomials and chose the best-fitting model. We then used the transformed data in conjunction with the Linear Regression model to perform Polynomial Regression.

For model evaluation, we calculated the Root Mean Squared Error (RMSE) and R2 score on the training data. The RMSE measures the difference between predicted and actual values, providing a measure of how well the model performs in terms of accuracy.

**Random Forest Regression Model**

In addition to Linear and Polynomial Regression, we implemented the Random Forest Regression model to predict MaxTemp. The Random Forest model is an ensemble learning technique that combines multiple decision trees to improve prediction accuracy and reduce overfitting.

Using the RandomForestRegressor class from scikit-learn, we trained the model on the training data with a chosen number of estimators. Similar to the other models, we evaluated the Random Forest model's performance using RMSE and R2 score on the training data.

**Results**

Upon analyzing the weather condition (temperature changes) during World War Two using different machine learning models, the following insights were obtained:

1. Linear Regression:
   * The Linear Regression model was applied to predict the maximum temperature (MaxTemp) based on various weather attributes.
   * The model provided a basic understanding of the linear relationship between weather features and MaxTemp.
   * However, due to the complexity of weather conditions, the model's predictive accuracy was limited.
2. Polynomial Regression:
   * To capture the nonlinearity in the relationship between weather attributes and MaxTemp, Polynomial Regression with a degree of 5 was used.
   * The Polynomial Regression model showed improvement over the Linear Regression model, providing a better fit to the data.
   * This model better represented the intricate temperature changes during World War Two.
3. Random Forest Regression:
   * The Random Forest Regression model, an ensemble learning technique, was employed to predict MaxTemp.
   * It offered a more robust prediction by capturing complex relationships between weather features and MaxTemp.
   * The model's performance was notable, but its accuracy might be influenced by the dynamic and unpredictable nature of weather.

**Conclusion**

In conclusion, our analysis of the "WW2 weather conditions" dataset using various machine learning models shed light on the relationship between temperature changes and other weather attributes. We found that the Random Forest Regression model provided the most accurate predictions of MaxTemp. However, this analysis is just a starting point, and further research can be conducted to refine the models and explore additional weather-related patterns during World War Two.

**Future Work**

For future work, we recommend experimenting with more advanced regression techniques, such as Support Vector Regression and Gradient Boosting Regression, to potentially achieve even better prediction accuracy. Additionally, incorporating historical events and their impact on military operations could lead to a more comprehensive understanding of the role weather played during World War Two.

**Acknowledgments**

We would like to express our gratitude to Kaggle user "usaf" for providing the "WW2 weather conditions" dataset. We also acknowledge the valuable contributions of our colleague Bharathwaj in collaborating on this project.

**References**

1. World War II Weather Conditions Dataset. Kaggle. [Link](https://www.kaggle.com/usaf/world-war-ii)
2. United States National Oceanic and Atmospheric Administration (NOAA) National Centres for Environmental Information. [Link](https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/world-war-ii-era-data)