

Weather Dataset Analysis Documentation:

Approach and Methodologies for Weather Dataset Analysis

1. Data Preprocessing:

- Missing Values: Identified missing values in the 'Precip Type' column (517 missing values).
- Data Types: Ensured appropriate data types; noted 'Wind Bearing (degrees)' has integer values and 'Formatted Date' is in string format.
- Handling Zero Values: Acknowledged that minimum values of 'Humidity', 'Wind Speed (km/h)', 'Wind Bearing (degrees)', and 'Visibility (km)' are zero, which could be valid.

2. Exploratory Data Analysis (EDA):

- Statistical Summary: Utilized `data.info()` function to get an overview of data types and summary statistics.
- Correlation Analysis: Observed high correlation between 'Apparent Temperature' and 'Humidity'.
- Pairwise Correlation Chart: Identified the strong negative correlation between 'Apparent Temperature' and 'Humidity', validating a hypothesis.
- Column Selection: Narrowed down analysis to key columns: 'Formatted Date', 'Apparent Temperature(c)', 'Humidity', ignoring others and missing values.

3. Regression Analysis:

- Linear Relation: Established a linear relationship between 'Apparent Temperature' and 'Humidity' with a negative slope.

4. Patterns Identified:

- Temperature and Humidity Relation: Noted that as air temperature increases, humidity decreases. Humidity remained constant over the observed period while 'Apparent Temperature' changed at regular intervals with constant amplitude.
- Temporal Trends: Observed fluctuations in 'Apparent Temperature' over time, with sharp rises in 2008-2009 followed by decreases in 2009-2010. Average 'Apparent Temperature' peaked in 2009 and decreased to its lowest in 2015.
- Seasonal Effects: Recognized that precipitation and weather summary mainly depend on the month (season), with no significant effect from year or day. Days with different weather summaries were noted, with rain being the most common type of precipitation.
- Impact of Timezone: Anticipated that changes in timezone may affect temperature and overall daily weather conditions.

5. Outlier Detection:

- Identified outliers in temperature, apparent temperature, humidity, and wind speed.
- Noted that pressure is highly skewed.

6. Weather Summary Analysis:

- Observed that most days are partly cloudy, followed by mostly cloudy, overcast, and foggy days.
- Noted that the dataset contains only two types of precipitation: rain and snow, with rain being more common.
- Recognized that precipitation type may differ based on weather summary, with rain being possible in all weather summaries while snowing shows some variation.

7. Other Relationships Identified:

- Relationship between visibility and humidity.
- Relationship between temperature and humidity.
- Variation in temperature based on weather conditions (e.g., highest on dry days, lowest on foggy and breezy days).
- Expectation of low pressure on windy and breezy days.

This comprehensive approach and analysis provide insights into the weather patterns over the observed period in Finland, contributing to a better understanding of the dataset and its implications.