CA Project 2023–2024

Introduction to Artificial Intelligence and Machine Learning (BSHC3A)

CA Project 2023–2024

Timofey Sneyd

(x21755195)

National College of Ireland

Abstract, Introduction and Previous Related Work to the topic of interest

Tools used for project

-Machine Learning Project

-Python, Jupyter Notebook

-Kaggle for datasets

Topic under study

In this study my decision was to study applying programming/machine learning skills in order to analyse global weather conditions and different global locations weather characteristics in order to make predictions while using conditions that would find and provide statistics on amount of locations globally which are under weather warning. By using daily updated dataset of weather around the world which consists of 18320 rows of information which is nearly double of dataset size required for this project's task.

Reason to select current topic

As for today global situation, shows us that climate around the world are changing facing different global situations and challenges more weather or weather forecasting analytics are required in order to not only predict upcoming changes of weather but also what else may come with the changes of weather and where people are required to be aware of the weather situation.

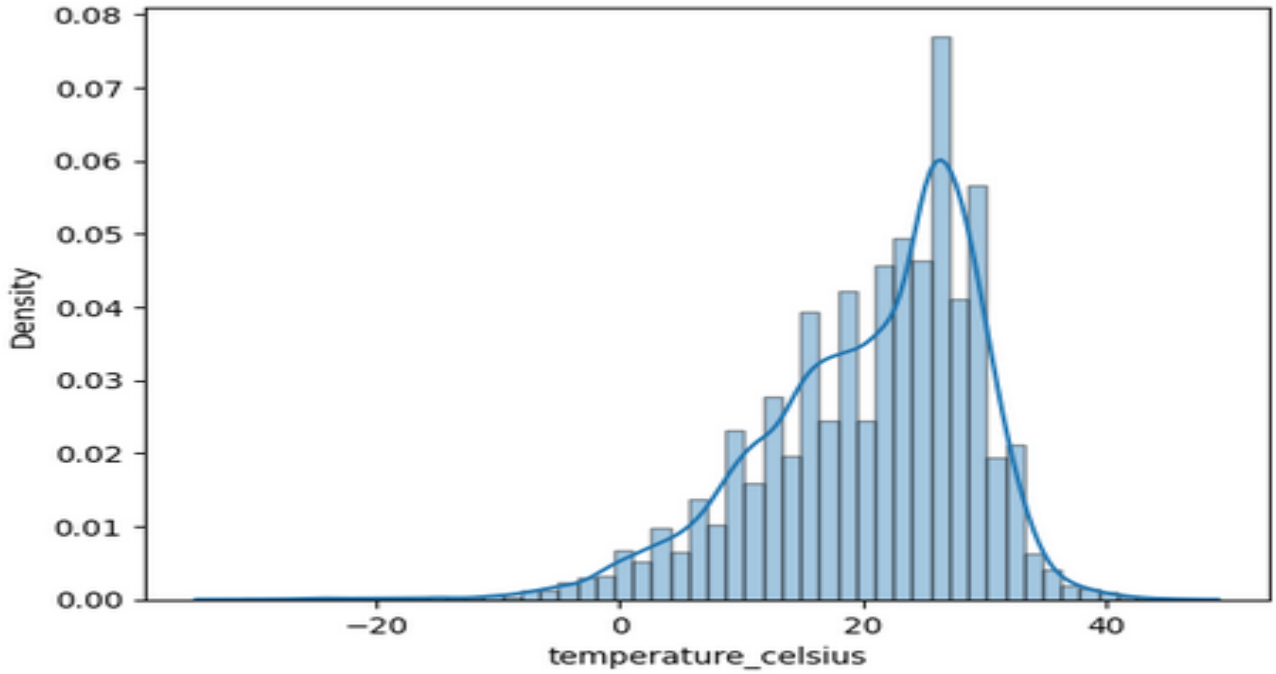
Related Work from Year 2 in BSc in Computing

In previous year for Data Structures and Algorithms subject I was developing a Java application which would use a custom csv dataset and would use algorithm in order to analyse the given data of coast sea levels and highlight in which locations sea levels go over fixed conditions and state that those locations require close observation as an example idea of how would some application would be useful in order to prevent for example natural disaster as flood.

Structure and Methodology

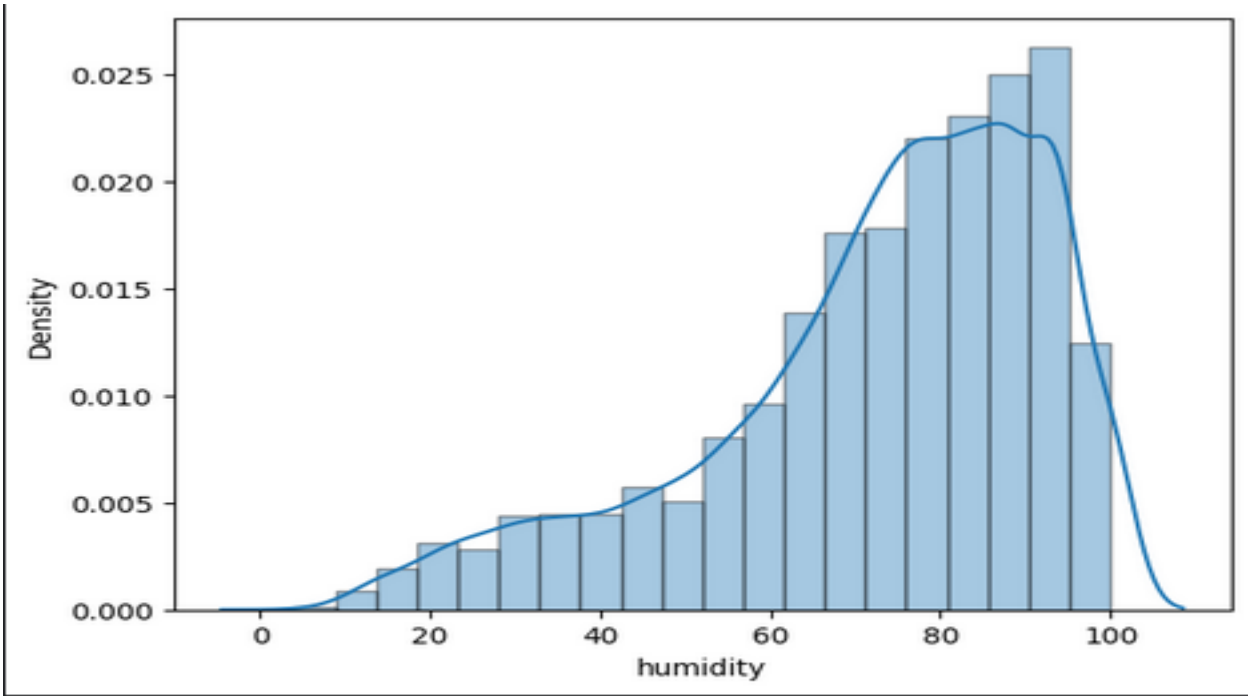
- **Importing dataset**
- **Initial analysis on the dataset**
- **Examining Missing Values**
- **Separating variables (Numeric or Categorical)**
- **Examining Statistics of Variables**
- **Analysis Outputs**
- **Numerical Variables(Analysis with Distplot)**
- **Simplifying the columns from dataset and their values**
- **Replacing high descriptive categories into more simplified to prevent high recursion**
- **Analysing our categoric variables from the dataset after replacing descriptive values(for weather conditions and wind directions) on more simplified**
- **Numerical Variables - Target Variable (Analysis with FaceGrid) and First intergration of our Target column for Weather Warning category into the dataset manually**
- **Categoric Variables - Target Variable + Creating Weather Warning (Analysis with Count Plot)**
- **Numeric Variables Among Themselves (Analysis with Pair Plot)**
- **Finding Logistic Regression Performance**
- **Comparing Logistic Regression and K-Nearest Neighbors (KNN) on a Classification Dataset**
- **Finding Linear Regression Performance**

Graphs Analysis



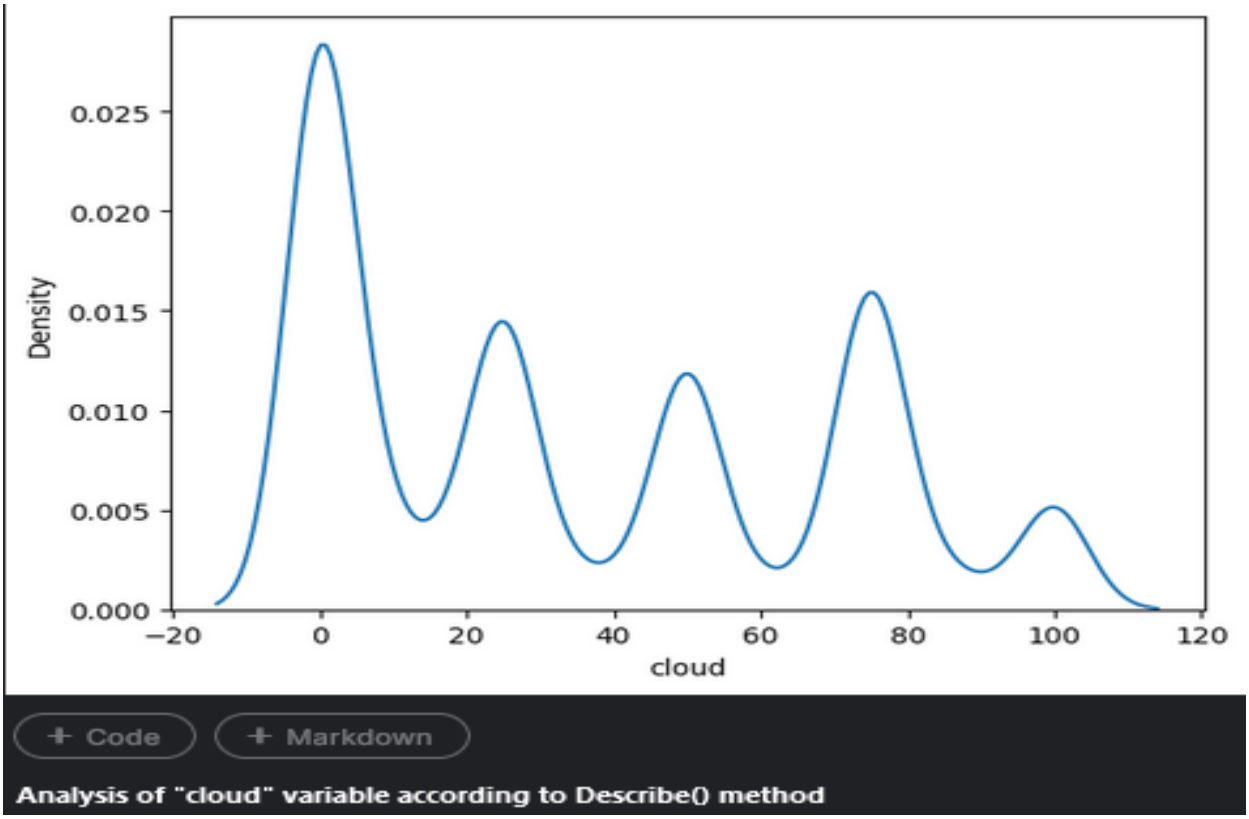
Analysis of "temperature_celsius" variable according to Describe() method

- **Central Tendency**: The peak of the KDE suggests that the most common temperature range is slightly below 20°C, indicating that the mode of the data is around this temperature.
- **Spread and Variability**: The data is spread from below -20°C to above 40°C, showing a wide range of temperatures across the locations. However, there's a significant concentration between approximately 10°C and 30°C.

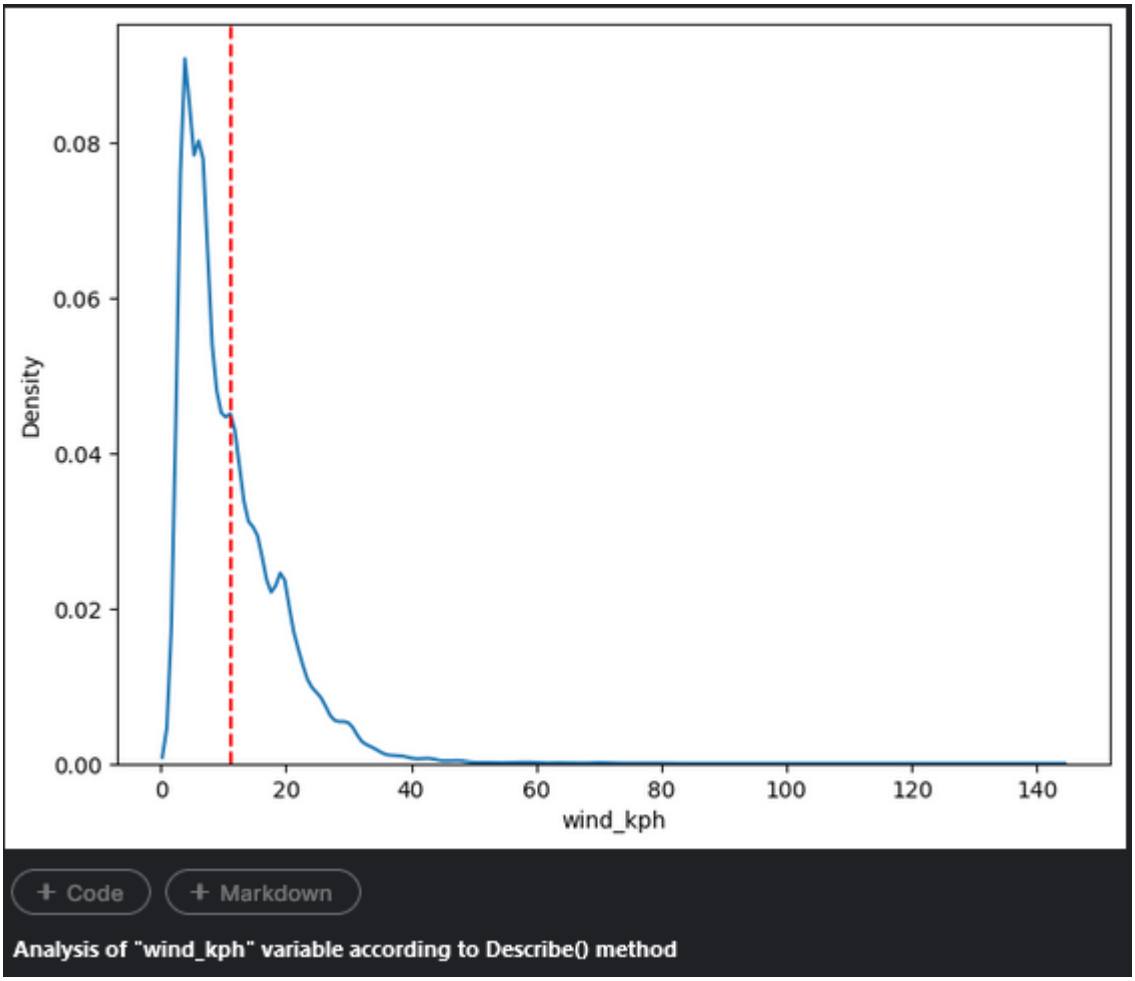


Analysis of "humidity" variable according to Describe() method

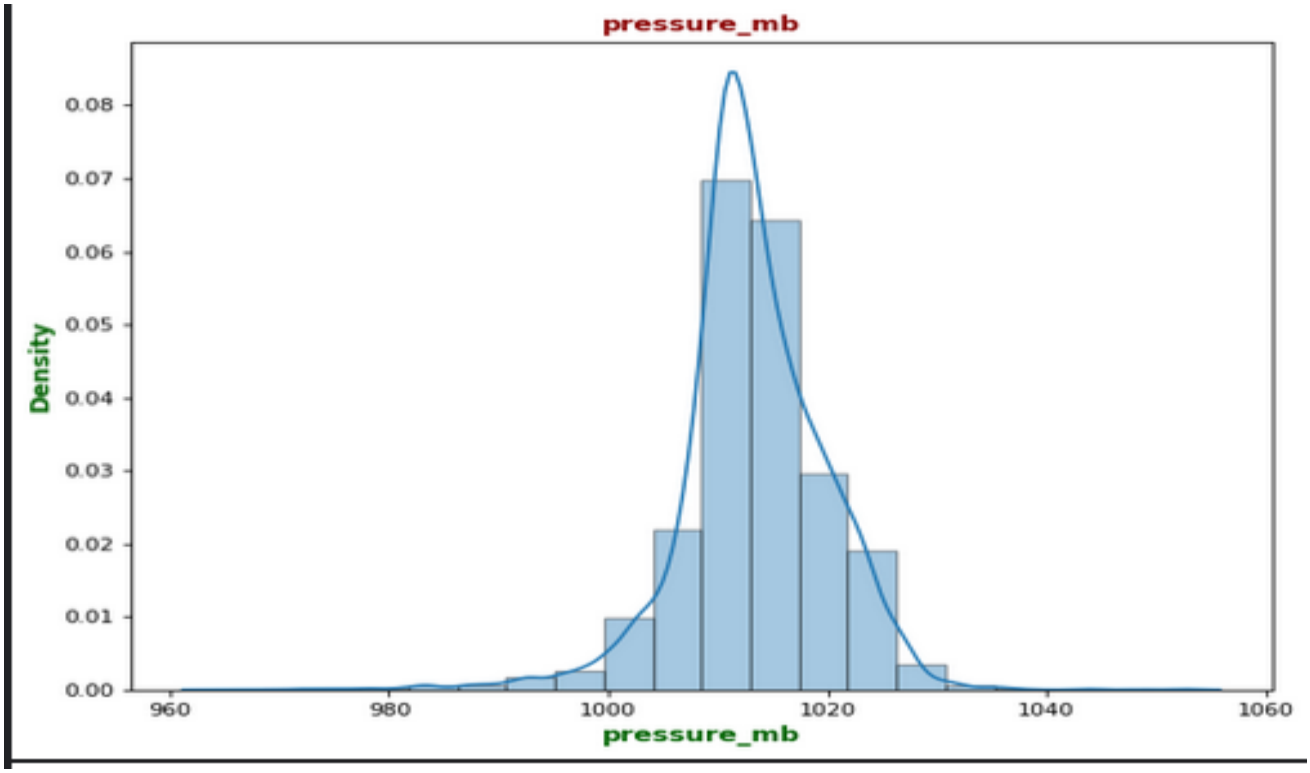
- **Central Tendency**: The distribution has its peak around 80% humidity, indicating that this is the most common humidity level among the sampled locations.
- **Spread and Variability**: The data spans the entire possible range of humidity from 0% to 100%, showing that the sampled locations have a broad range of humidity levels.



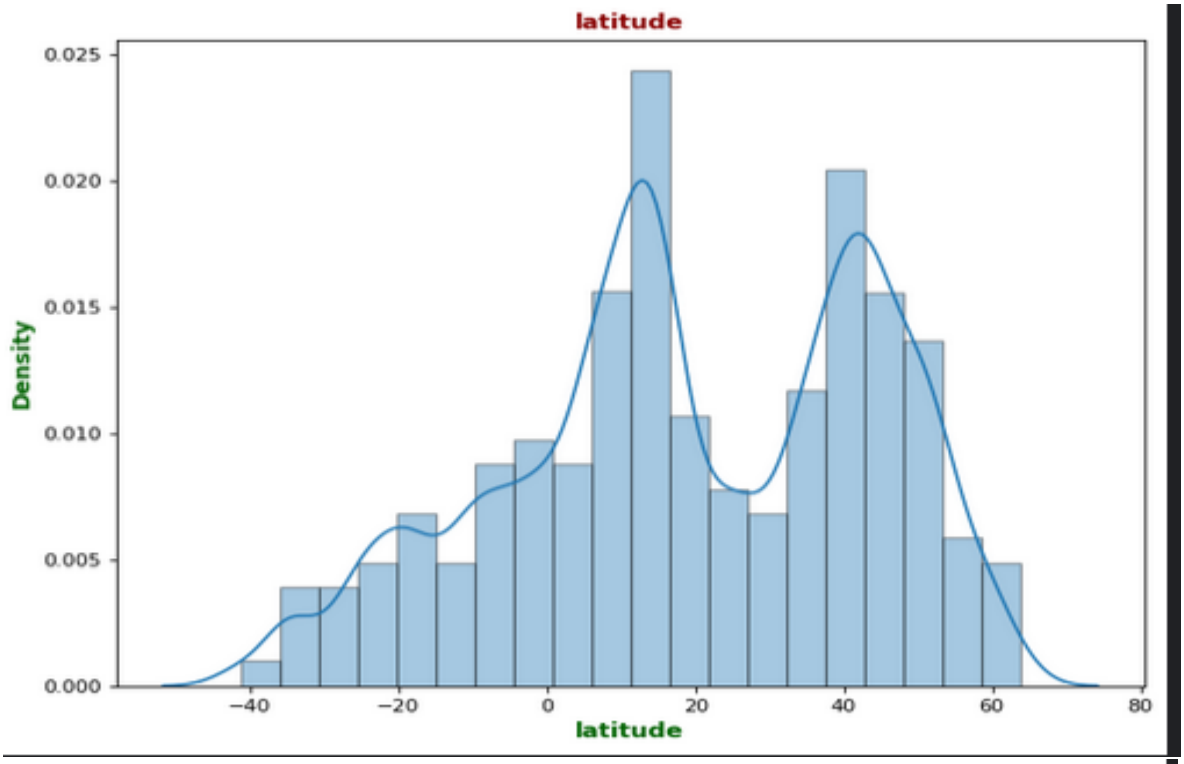
- **Central Tendency**: Unlike the previous plots for temperature and humidity, this graph does not show a clear single mode. Instead, it features several peaks of different heights, which suggests that the data might be multimodal.
- **Periodicity**: The graph shows a periodic-like pattern with peaks and troughs at regular intervals. This could indicate recurring conditions in cloud cover across different locations or it could be an artifact of how the data is collected or processed.
- **Spread and Variability**: The values for cloud cover are spread across the entire range, but there are noticeable concentrations of data at regular intervals. This could imply that cloud cover is commonly reported or measured in specific increments (e.g., 0%, 20%, 40%, etc.).

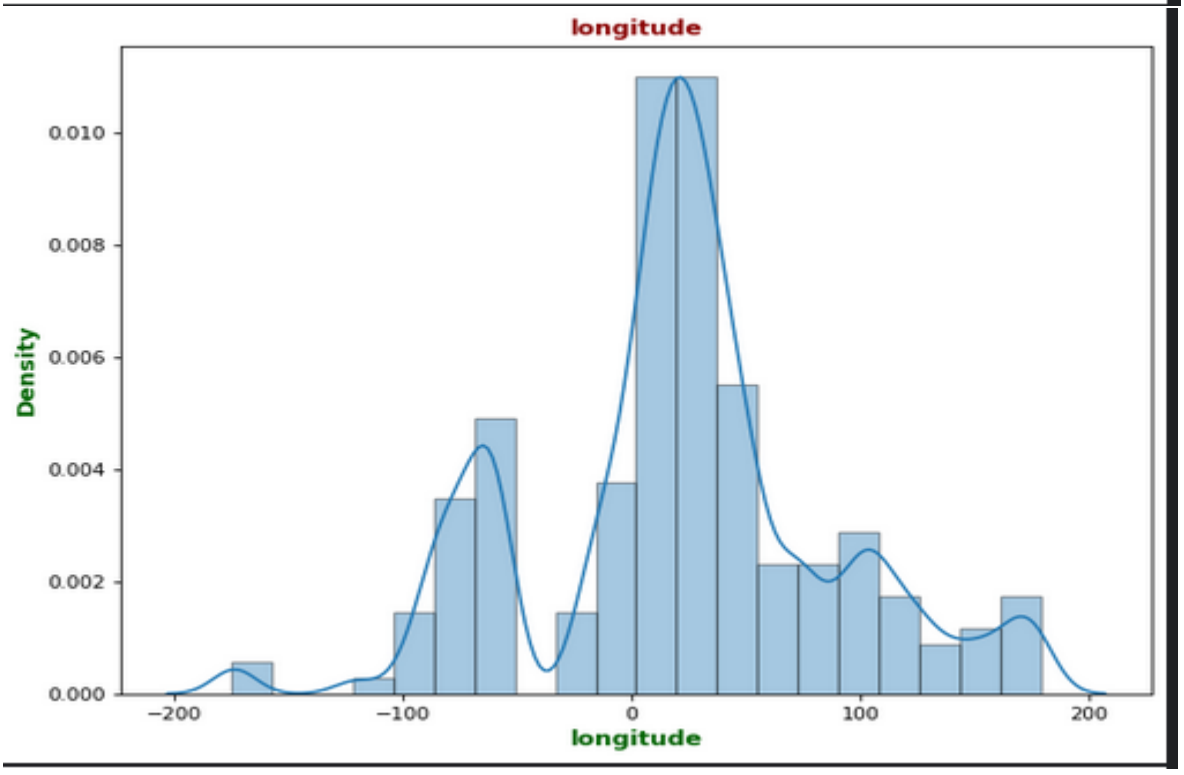


- **Central Tendency**: The peak of the KDE is near 0 kph, which suggests that calm conditions (little to no wind) are the most common across the sampled locations.
- **Skewness**: The distribution is heavily right-skewed, with a steep drop-off after the peak and a long tail extending to the right, indicating that higher wind speeds are less common.
- **Potential Outliers**: Given the nature of wind speed distributions, it's not unexpected to have a tail; however, data points extending to very high wind speeds might be outliers or represent extreme weather conditions such as storms or cyclones.
- **Anomalous Values**: There are no values shown below 0 kph, which is appropriate as negative wind speeds are not physically meaningful.



- **Central Tendency**: The peak of the KDE suggests that the most common atmospheric pressure reading is around 1010 mb, which is close to the average sea-level pressure of 1013.25 mb.
- **Spread and Variability**: The data is primarily concentrated between approximately 980 mb and 1040 mb, showing a relatively tight range of atmospheric pressure readings across the locations.
- **Skewness**: The distribution is fairly symmetrical with a slight left skew, indicating a minor tail with lower pressure values. This symmetry around a central peak is typical for atmospheric pressure data.
- This graph helps in visualizing the overall distribution of atmospheric pressure across the sampled global locations. It indicates that most of the readings are clustered around what is considered normal atmospheric pressure, with fewer instances of very high or very low readings. The symmetry and slight skew toward lower values suggest that extreme low-pressure conditions (which are associated with stormy weather) are less common but still present in the data set.





- **Multimodality**: The distribution is multimodal, with peaks around -40, 0, and 40 degrees latitude. This could indicate a higher concentration of data points (or weather stations) around these latitudinal bands.
- **Spread**: The data is spread out across the latitude range, with notable gaps between the peaks, possibly indicating less data from those regions.

Longitude Graph:

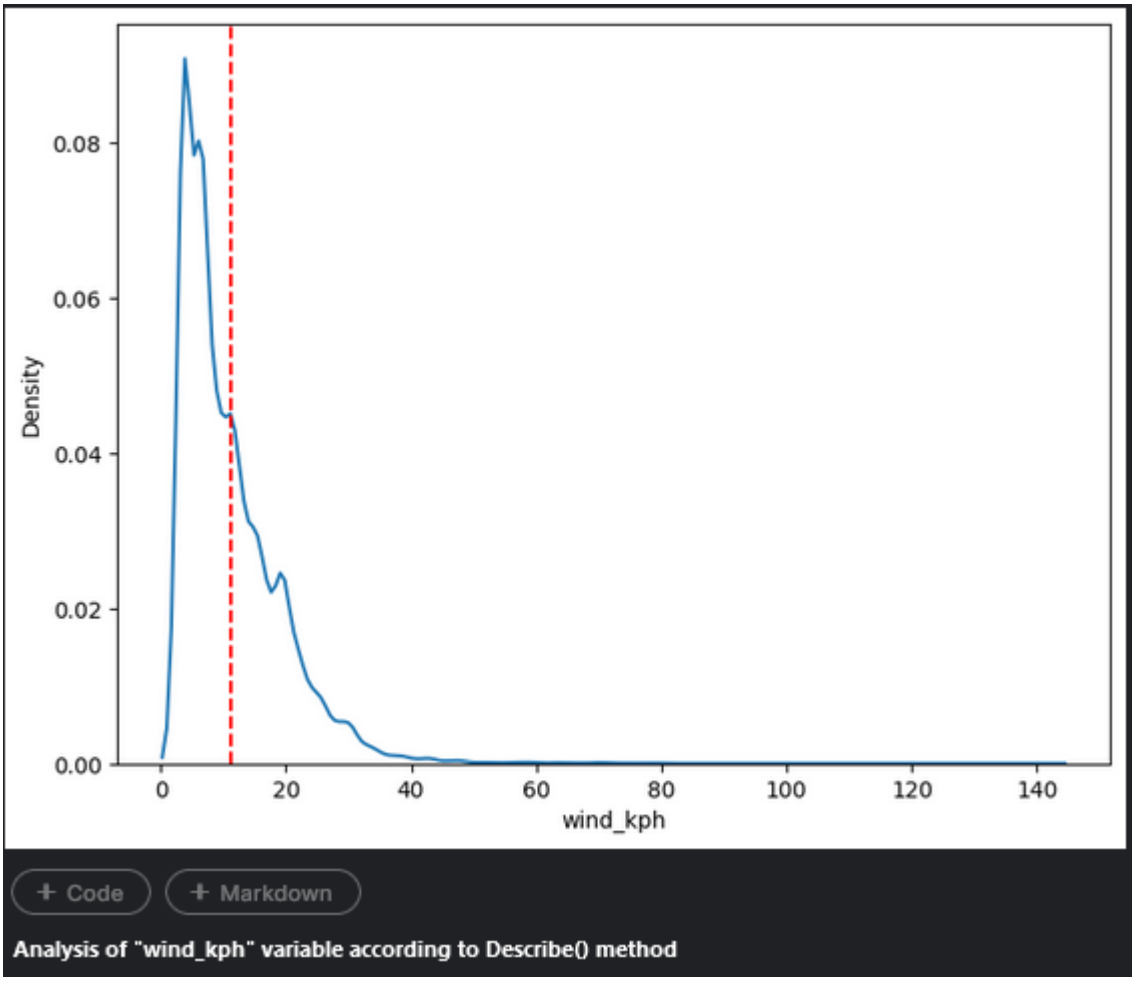
- **Histogram**: This histogram shows the frequency of longitude readings. The x-axis represents longitude values from about -180 to 180 degrees, and the y-axis indicates the density.
- **KDE**: The KDE curve suggests the probability distribution of longitude readings across the sampled locations.

Observations from the Longitude Graph:

- **Central Tendency**: There's a significant peak around 0 degrees, likely corresponding to the Prime Meridian, which passes through Greenwich, England.
- **Spread**: The data points are spread across the entire range of longitude values, with multiple peaks indicating more frequent data points or weather stations along those longitudinal lines.
- **Multimodality**: The graph is also multimodal with additional, albeit smaller, peaks occurring to the left and right of the central peak.

Combined Conclusion:

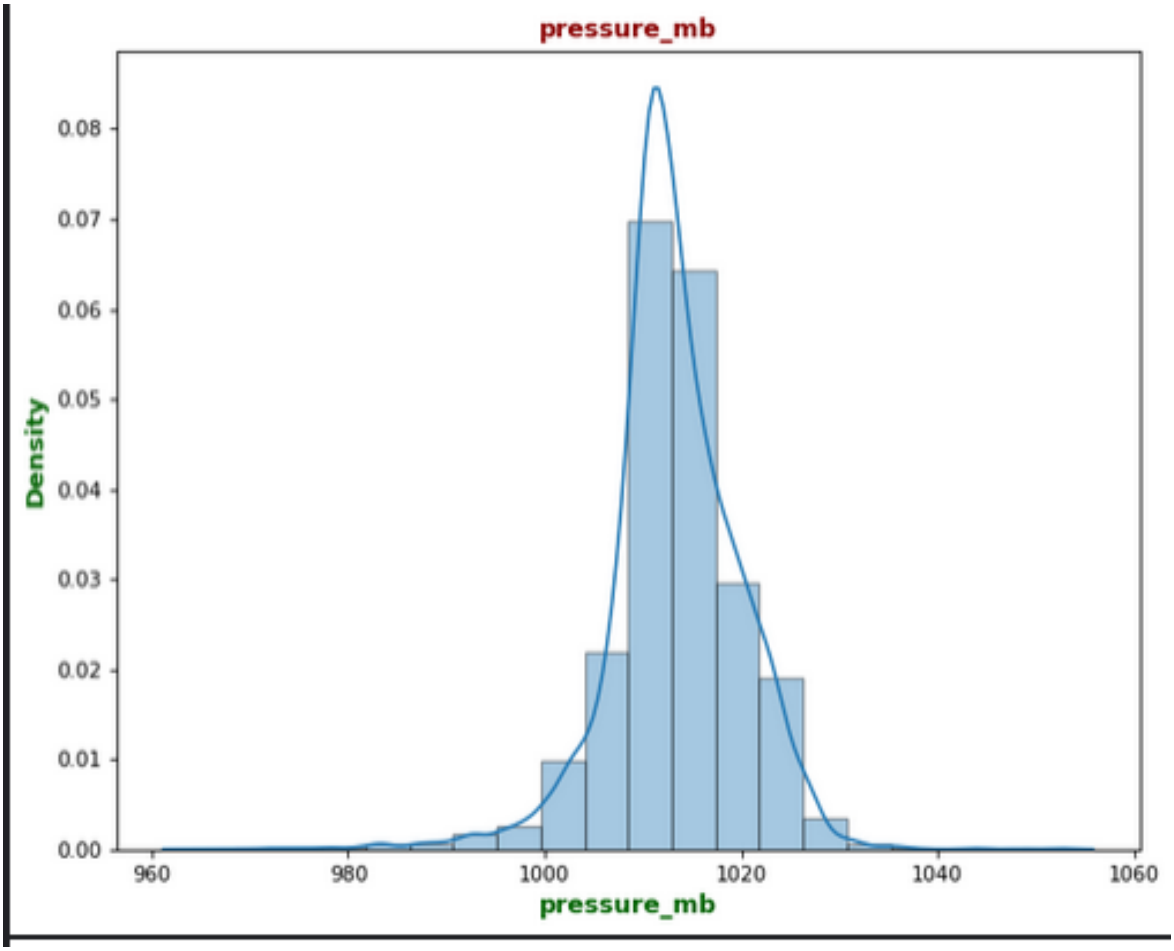
- Both latitude and longitude distributions show multimodal behavior, which suggests that the data points are not evenly spread across the globe but are clustered around certain latitudes and longitudes. This clustering could be due to the higher density of weather stations in populated areas or along certain latitudinal belts that correspond to specific climatic zones. The longitude graph, in particular, shows a strong preference for locations along the Prime Meridian, and possibly other lines that align with the organization of the global weather dataset or the geographic distribution of reporting weather stations.



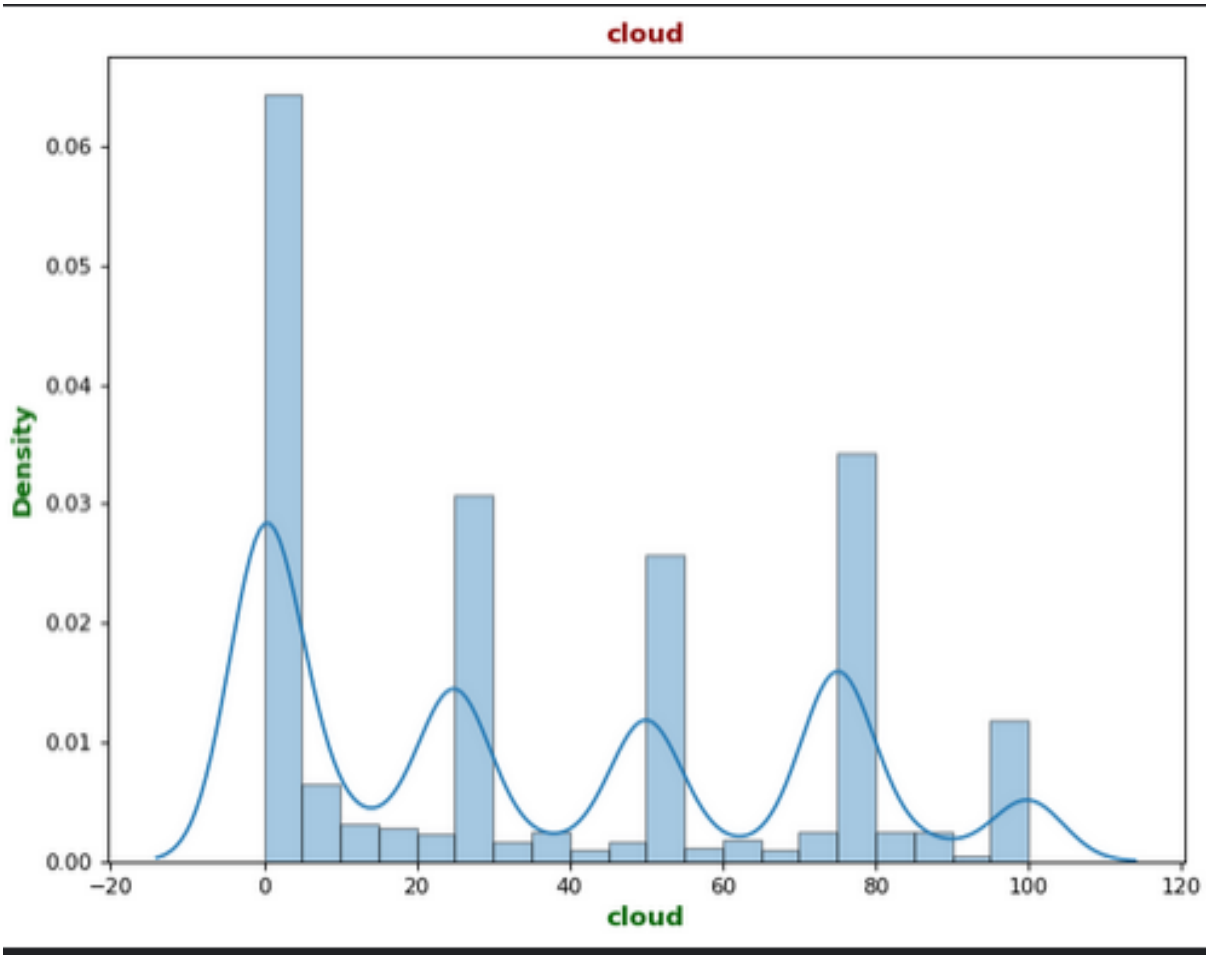
- **Central Tendency**: The peak of the KDE is near 0 kph, which suggests that calm conditions (little to no wind) are the most common across the sampled locations.
- **Skewness**: The distribution is heavily right-skewed, with a steep drop-off after the peak and a long tail extending to the right, indicating that higher wind speeds are less common.
- **Tail Behavior**: The long tail towards the higher wind speeds suggests that while most locations experience lower wind speeds, there are occurrences of much higher speeds, although these are less frequent.
- **Potential Outliers**: Given the nature of wind speed distributions, it's not unexpected to have a tail; however, data points extending to very high wind speeds might be outliers or represent extreme weather conditions such as storms or cyclones.

Conclusion:

The KDE plot indicates that most of the locations in the dataset typically experience low wind speeds, with a significant drop in density as the wind speed increases. This is consistent with what one might expect in a general distribution of wind speeds where high wind speeds are relatively rare. The red dashed line could indicate an average or typical wind speed that can be used as a reference point for comparing individual measurements. It is likely close to the mode of the distribution given the shape of the curve.

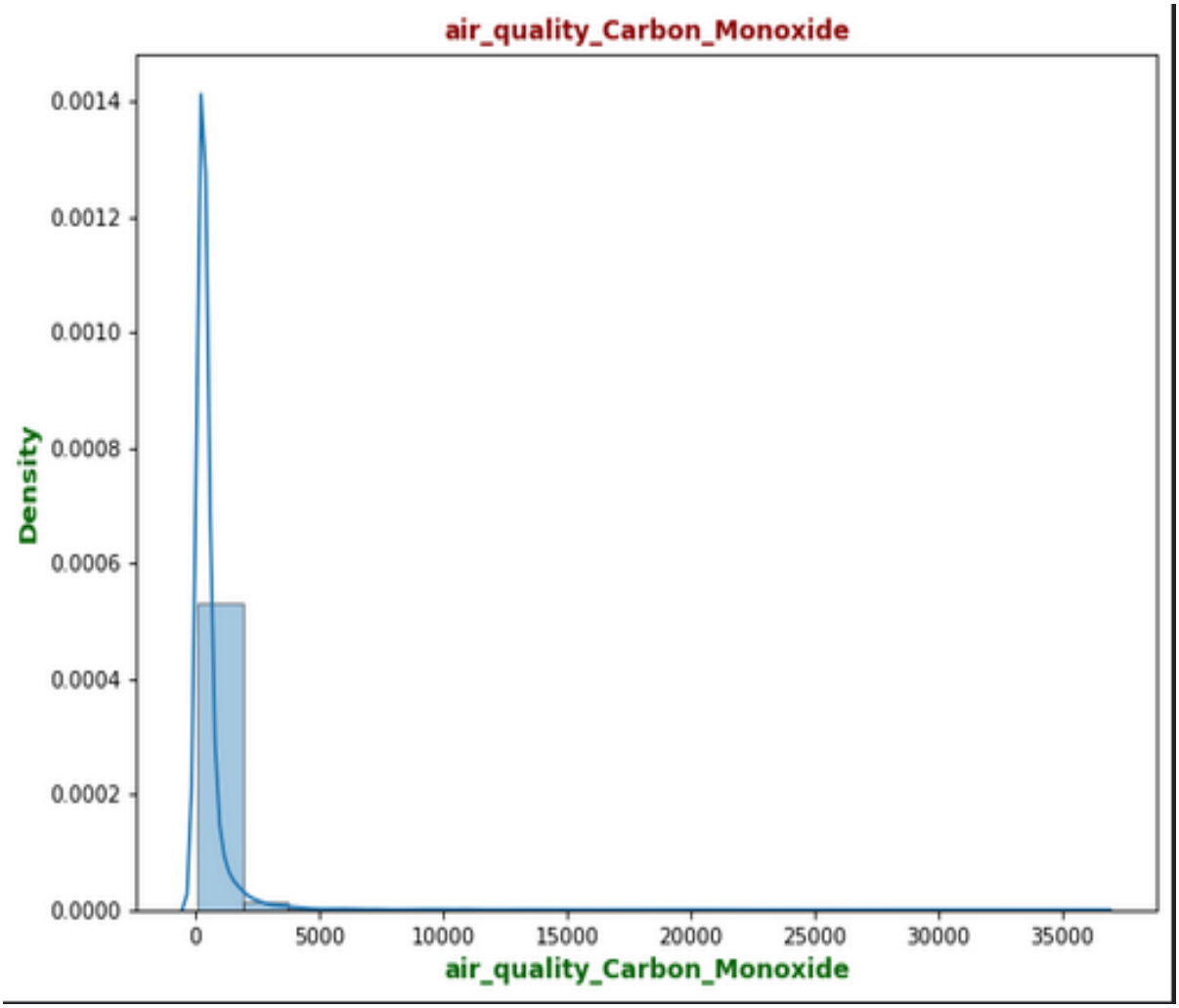


- **Central Tendency**: The peak of the KDE suggests that the most common atmospheric pressure reading is around 1010 mb, which is close to the average sea-level pressure of 1013.25 mb.
- **Spread and Variability**: The data is primarily concentrated between approximately 980 mb and 1040 mb, showing a relatively tight range of atmospheric pressure readings across the locations.
- **Skewness**: The distribution is fairly symmetrical with a slight left skew, indicating a minor tail with lower pressure values. This symmetry around a central peak is typical for atmospheric pressure data.

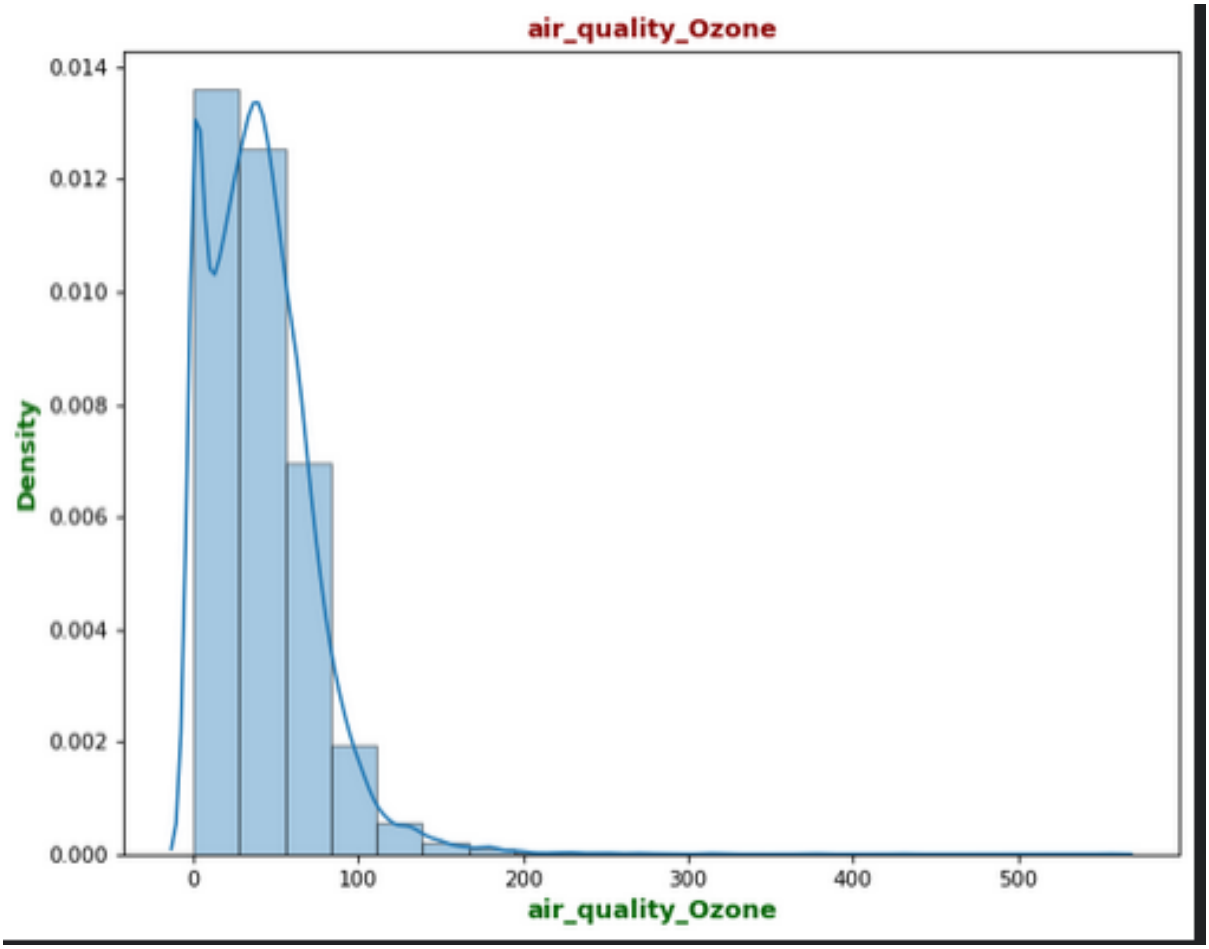


- **Histogram**: The bars represent the distribution of cloud observations from the global weather dataset across 18,320 locations around the globe. The x-axis, labeled 'cloud', likely represents the cloud cover or cloud index at these locations, with the scale running from -20 to 120. This scale is unusual because cloud cover is typically measured as a percentage from 0 to 100, so the meaning of values less than 0 or greater than 100 is unclear without additional context.

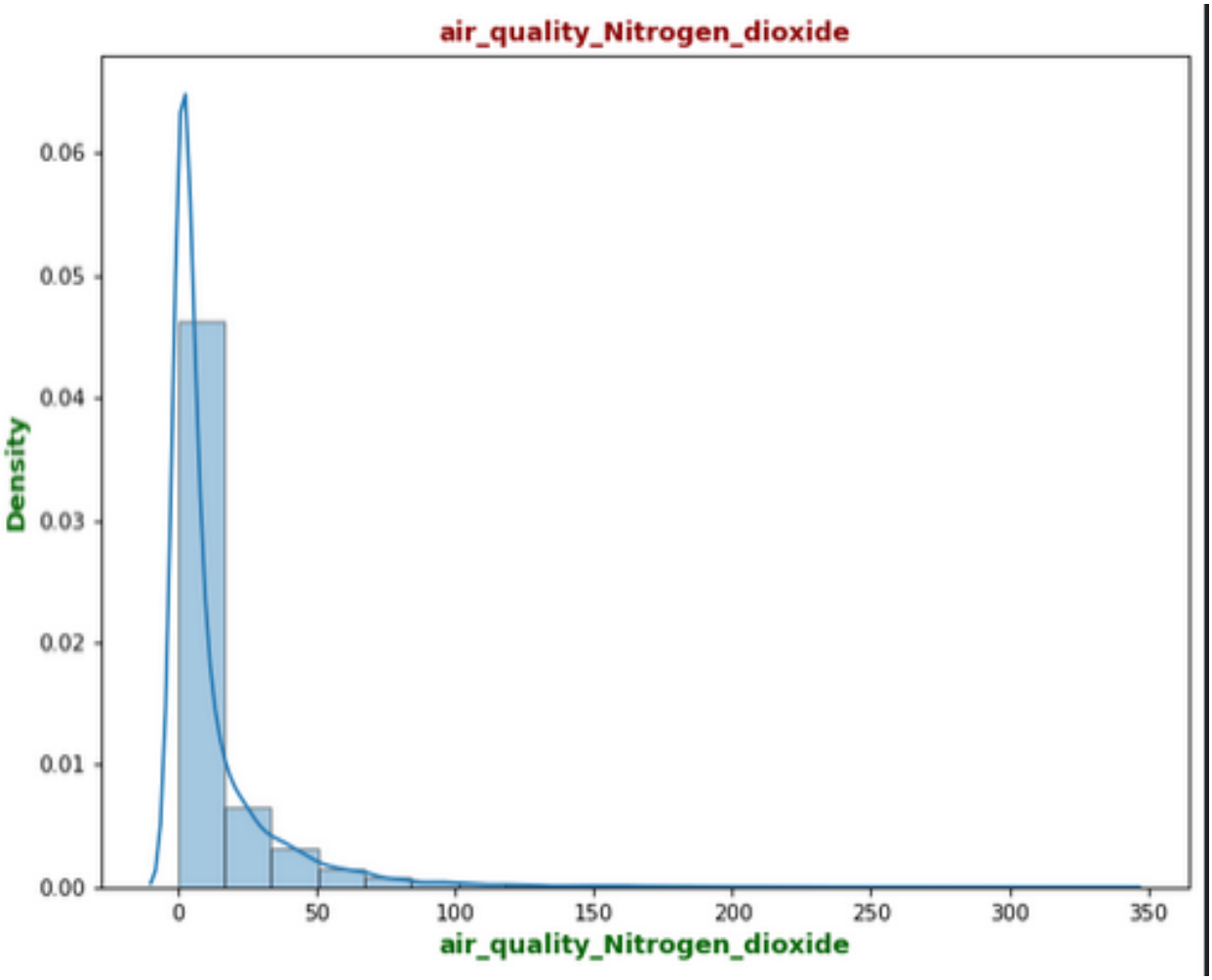
- **Kernel Density Estimate (KDE)**: The smooth blue line represents the KDE, which is a way to estimate the probability density function of a continuous random variable. It provides a smooth curve that represents the density of data points along the range of cloud cover values.
- **Density**: On the y-axis, we see 'density', which in the context of a KDE plot, represents the probability density function for the variable 'cloud'. The values on the y-axis indicate the density of observations for particular values of 'cloud'.



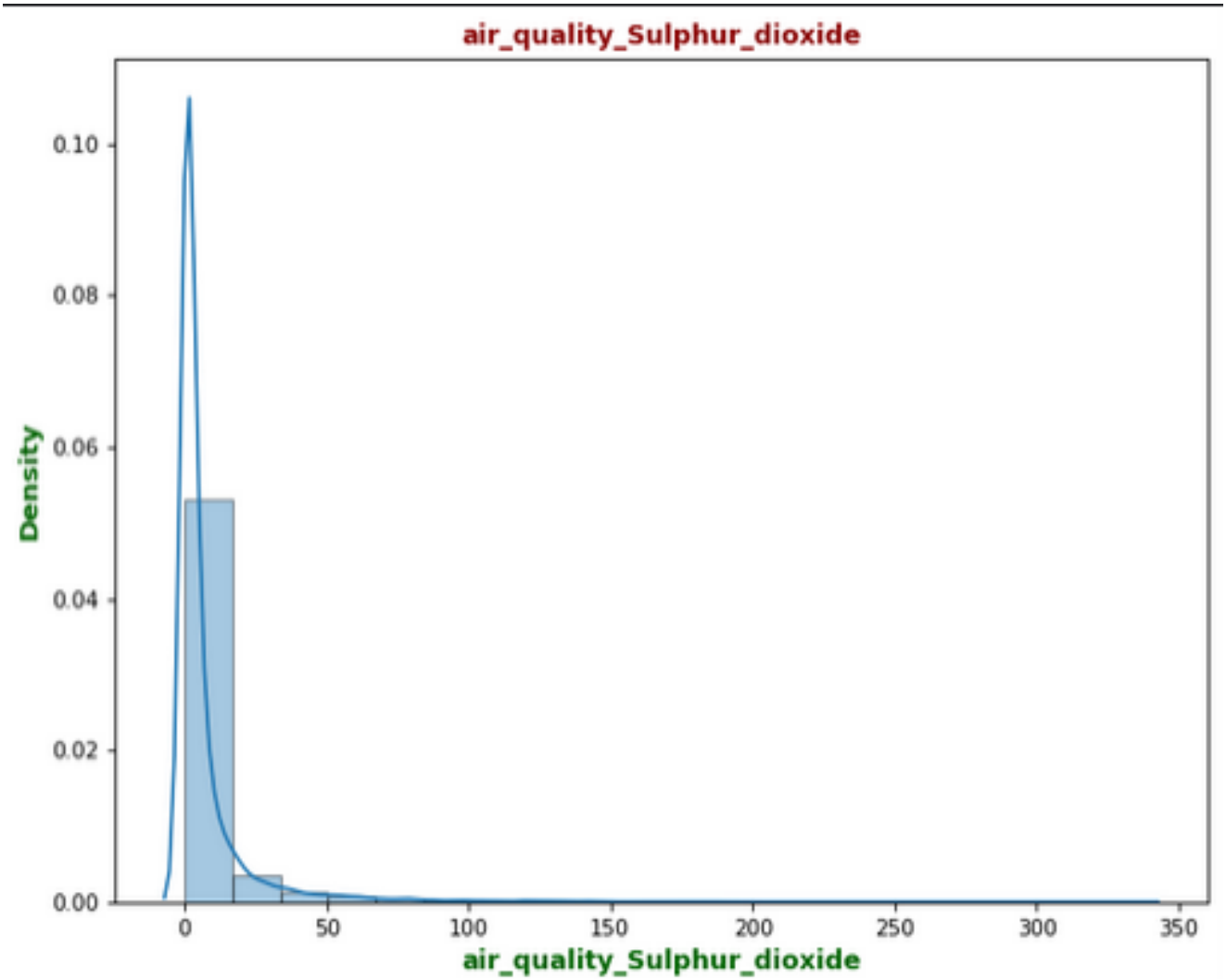
- **Central Tendency**: The sharp peak near the origin indicates that the most common concentration of carbon monoxide across the sampled locations is very low, close to zero.
- **Skewness**: The distribution is heavily right-skewed, showing that higher concentrations of CO are much less common than lower ones.
- **Outliers**: There are data points that extend to very high CO concentrations, suggesting either the presence of outliers, measurement errors, or areas with extremely poor air quality.



- **Central Tendency**: The peak near the origin suggests that the most common ozone concentration levels in the dataset are low.
- **Spread and Variability**: The distribution extends to approximately 500 on the x-axis, with most of the data points concentrated below 200.
- **Skewness**: The distribution is right-skewed, with a tail stretching towards the higher ozone concentrations, indicating that while most of the locations have lower concentrations, there are a few locations with much higher concentrations.
- **Outliers**: There seem to be a few data points with very high ozone concentrations which might be considered outliers or might indicate areas with particularly poor air quality.

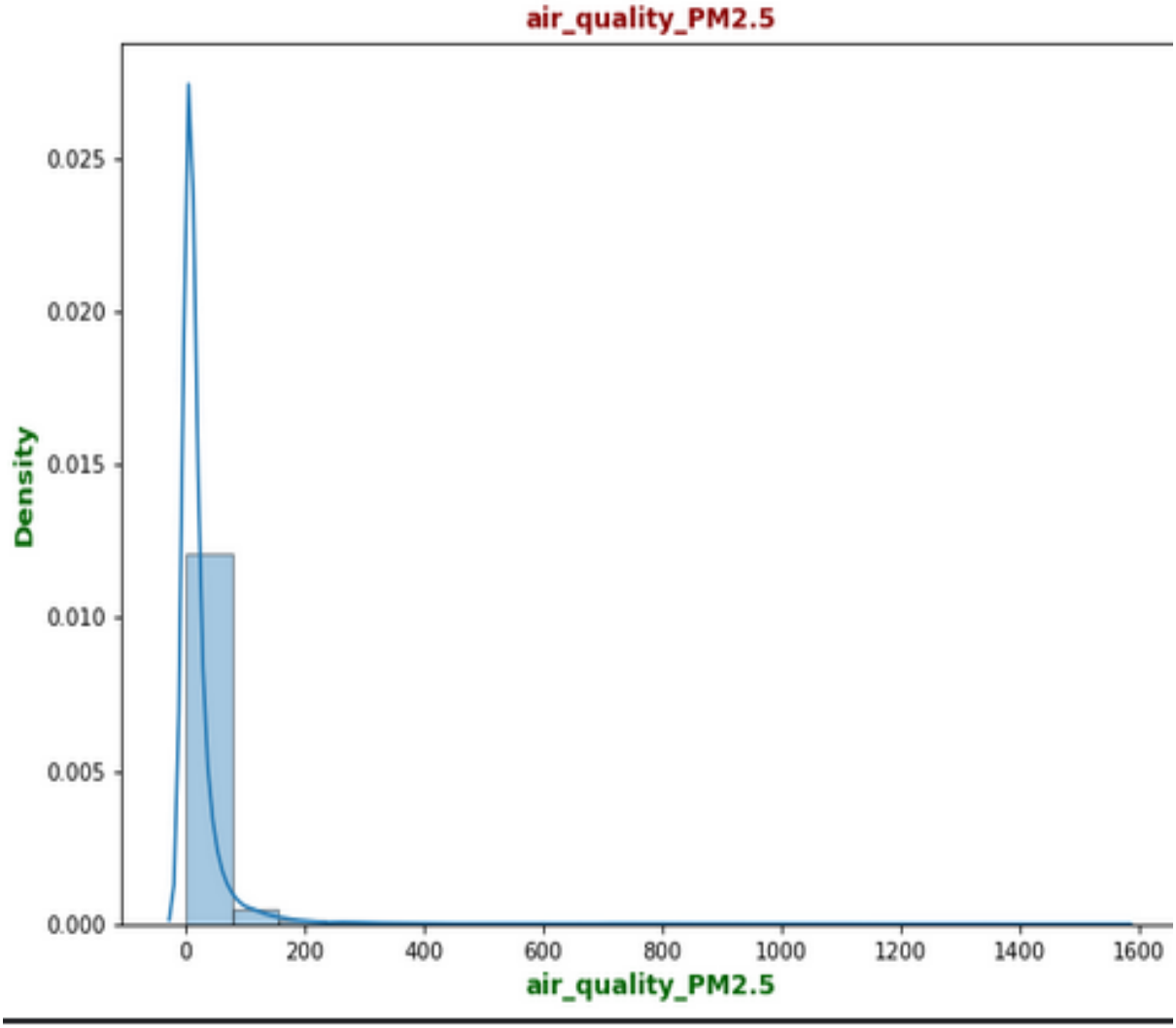


- **Central Tendency**: There is a pronounced peak close to zero, suggesting that lower NO2 concentration levels are the most common across the dataset's locations.
- **Spread and Variability**: The data points are spread out up to 350, with a noticeable drop in density beyond approximately 50.
- **Skewness**: The distribution is right-skewed, indicating that while most places have lower NO2 levels, a tail of data points reflects higher NO2 concentrations, albeit less frequently.
- **Outliers**: Some data points represent high NO2 levels, which could indicate areas with significant air pollution or occasional spikes in NO2 due to specific environmental factors or events.



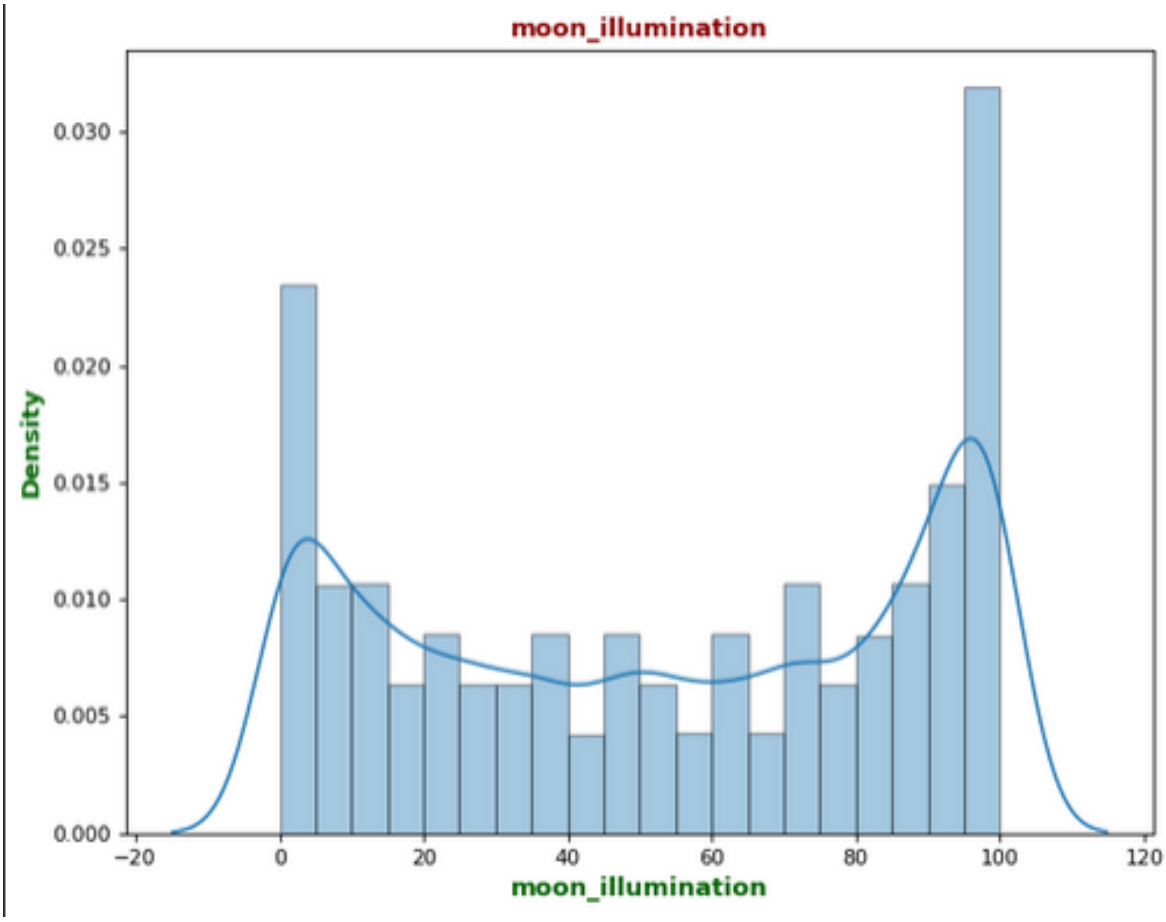
- **Central Tendency**: The peak near the origin indicates that the most frequent sulphur dioxide concentration levels are very low, almost close to zero.
- **Spread and Variability**: The data points are mostly concentrated within the lower range of the x-axis, with density rapidly declining as the concentration increases.
- **Skewness**: The distribution is right-skewed, meaning there is a longer tail on the right side of the peak, which shows that while most places have lower SO2 levels, there are few areas with much higher concentrations.
- **Outliers**: The graph suggests the presence of a few readings with unusually high SO2 concentrations, which may represent areas with specific pollution sources or industrial activities.

- The x-axis represents the concentration lev



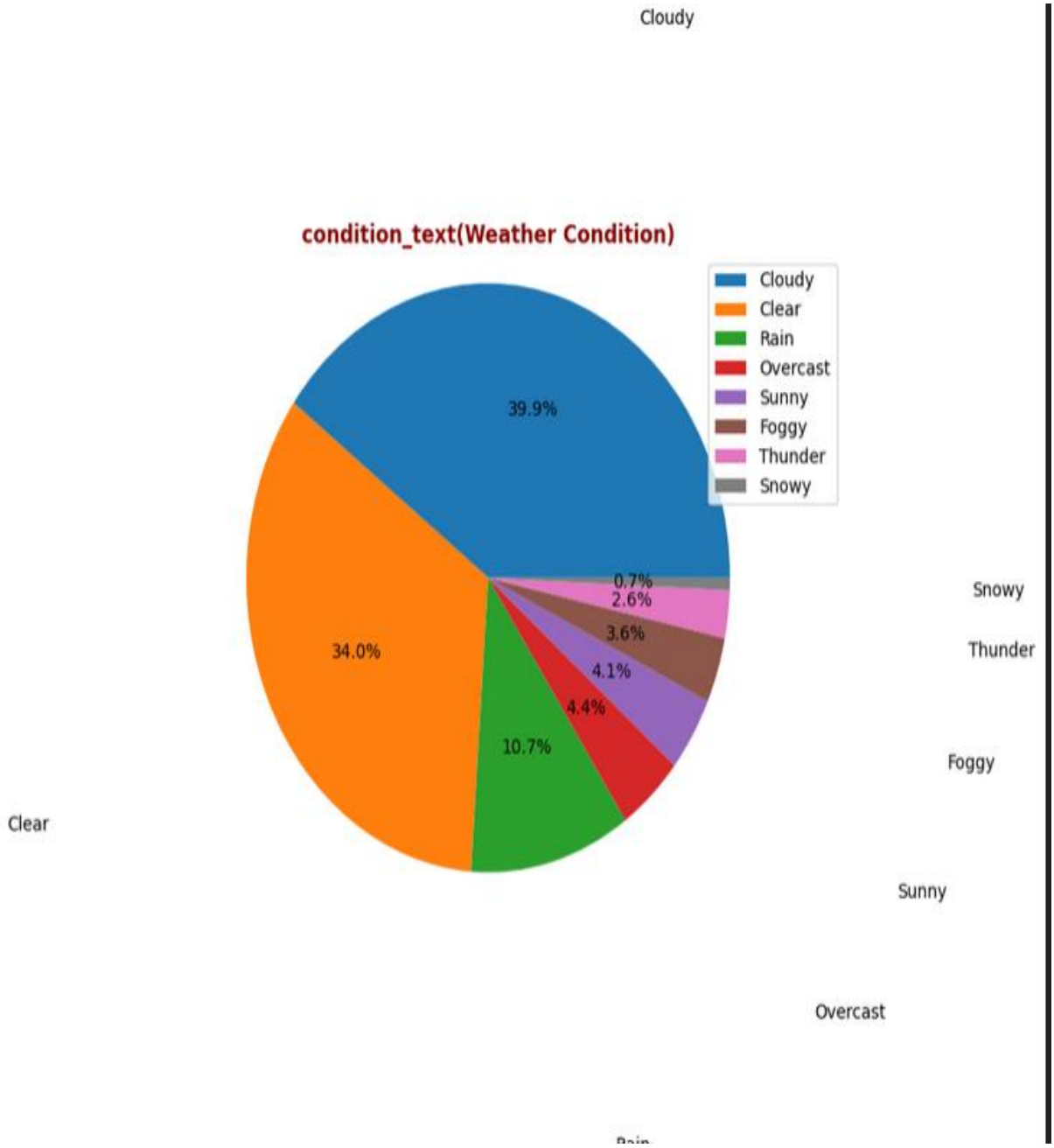
els of PM2.5, which range from 0 to over 1600. PM2.5 concentrations are typically measured in micrograms per cubic meter (µg/m³).
- The y-axis displays the density of the dataset for the given PM2.5 concentrations.

- **Central Tendency**: There is a significant peak at the lower end of the x-axis, suggesting that most locations in the dataset have low PM2.5 concentration levels.
- **Spread and Variability**: While the majority of data points are clustered near the lower concentrations, the range extends up to 1600 µg/m³, indicating some occurrences of very high PM2.5 levels.

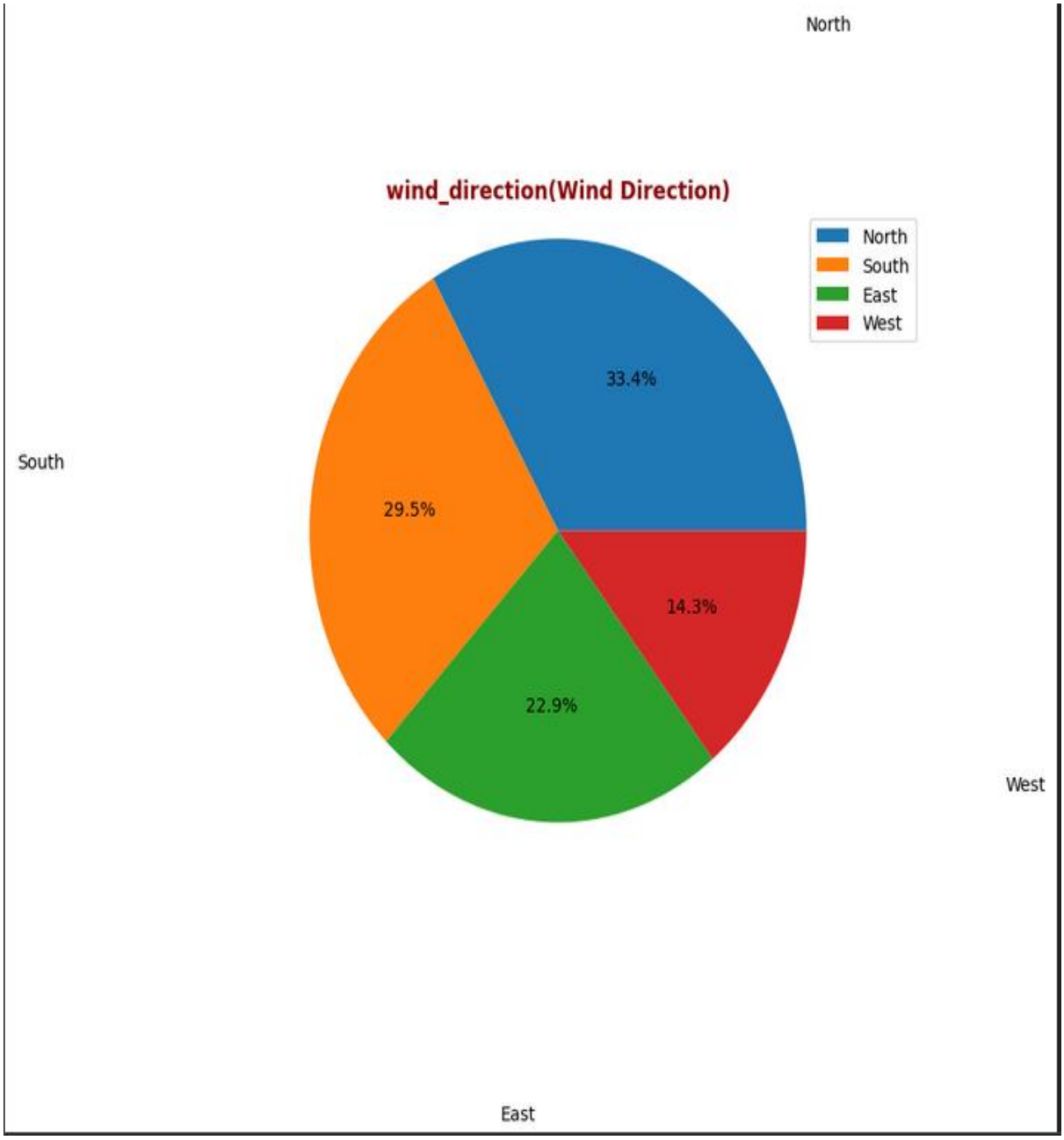


- The x-axis represents the percentage of moon illumination, ranging from below 0% to above 100%. It is expected that the values should be between 0% and 100%, as these represent the possible range of moon illumination from new moon (0%) to full moon (100%).
- The y-axis shows the density of the data points for the given moon illumination percentages.

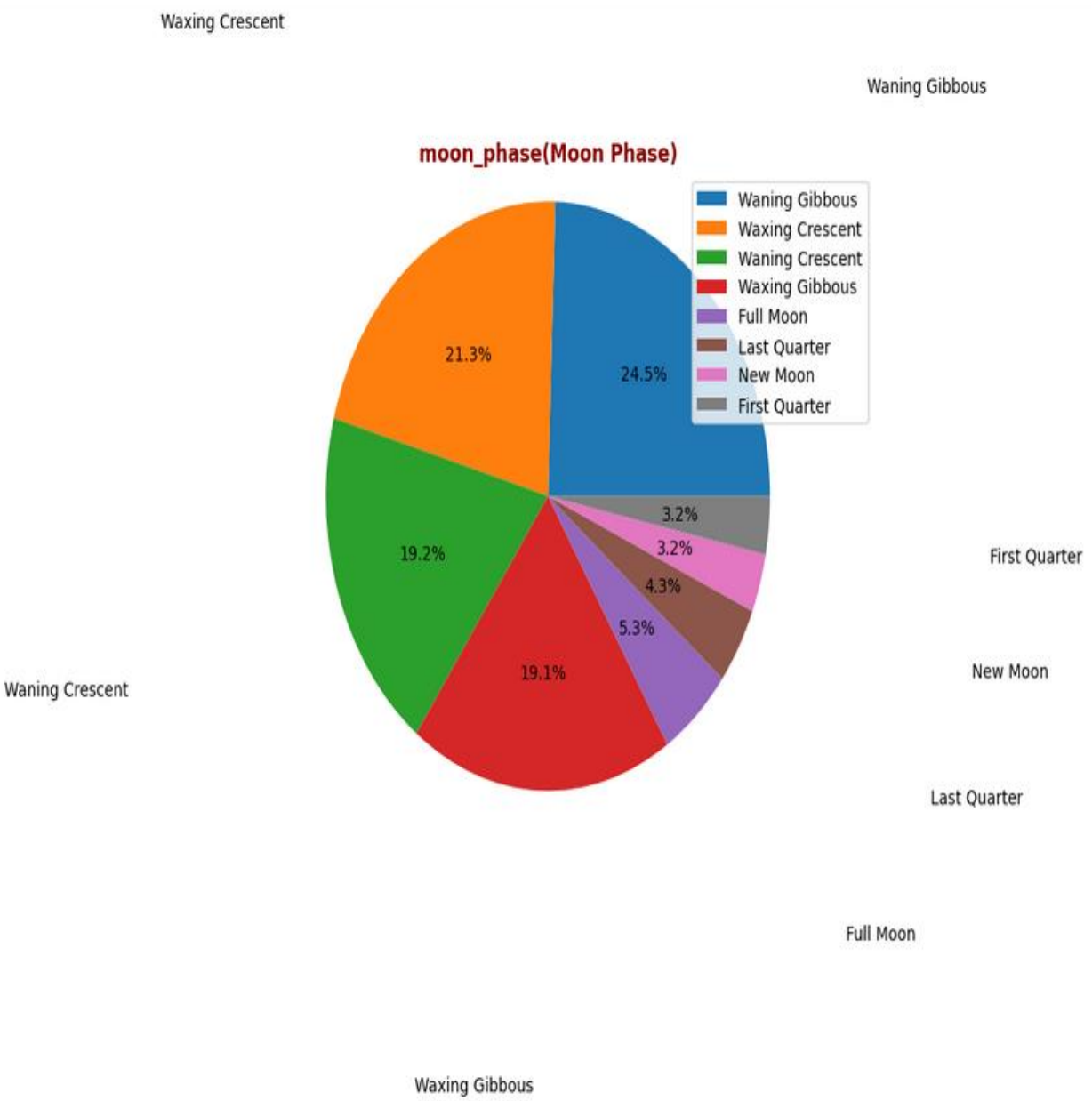
- **Central Tendency**: The graph has multiple peaks, suggesting that the data might be multimodal. Peaks near 0% and 100% likely represent new moons and full moons, respectively.
- **Spread and Variability**: The data is spread across the entire range of moon illumination but shows peaks at intervals that likely correspond to specific phases of the moon.
- **Periodicity**: The pattern of peaks is somewhat periodic and symmetric around 50%, which could indicate the waxing and waning phases of the moon.



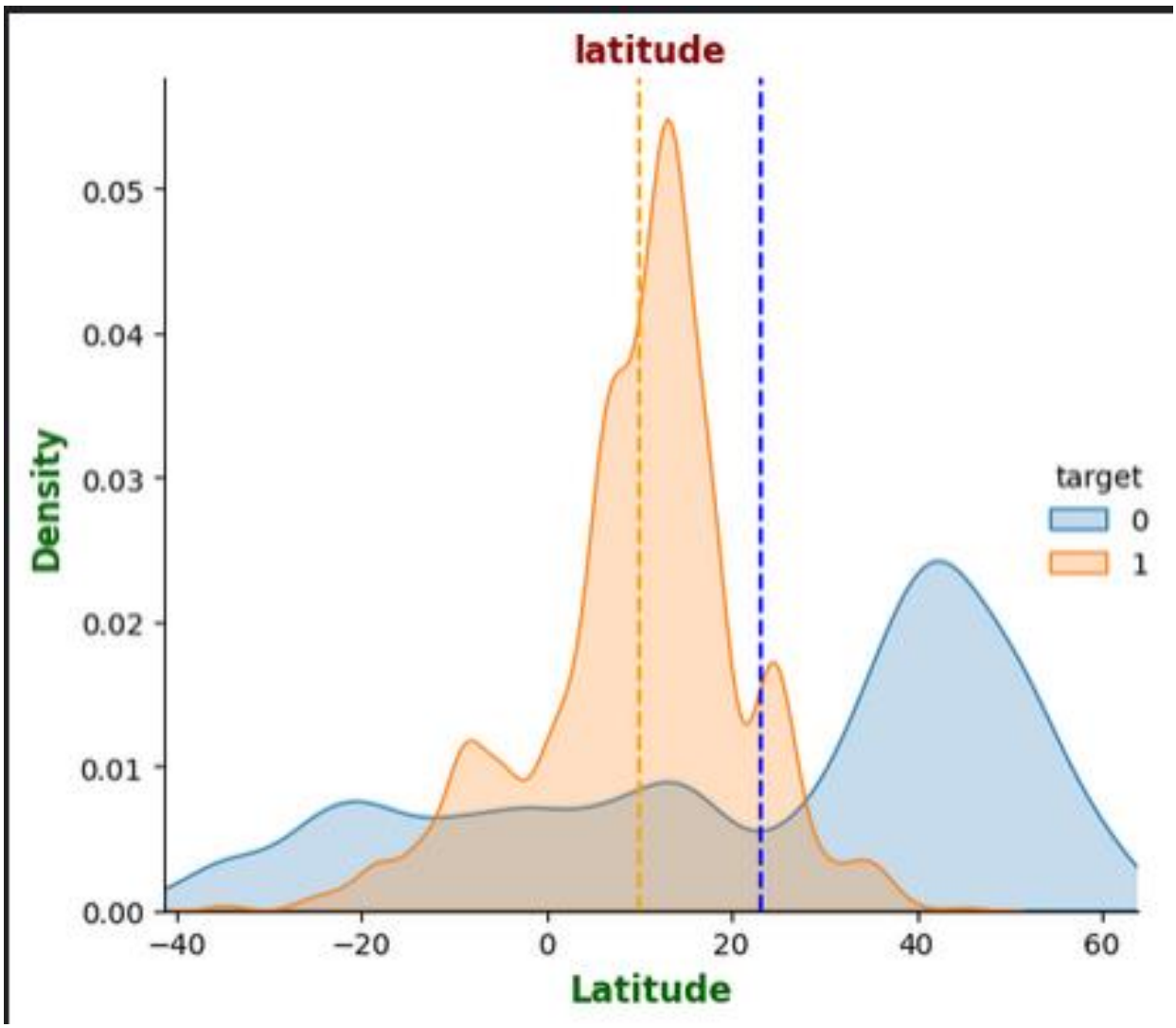
- **Most Common Conditions**: The two most common weather conditions are Cloudy (39.9%) and Clear (34.0%), which together make up the majority of the dataset.
- **Less Common Conditions**: Rainy and Overcast conditions are also represented but to a lesser extent, with Rain at 10.7% and Overcast at 4.4%.
- **Rare Conditions**: Sunny (3.6%), Foggy (2.6%), Thunder (0.7%), and Snowy conditions are the least represented in the dataset.



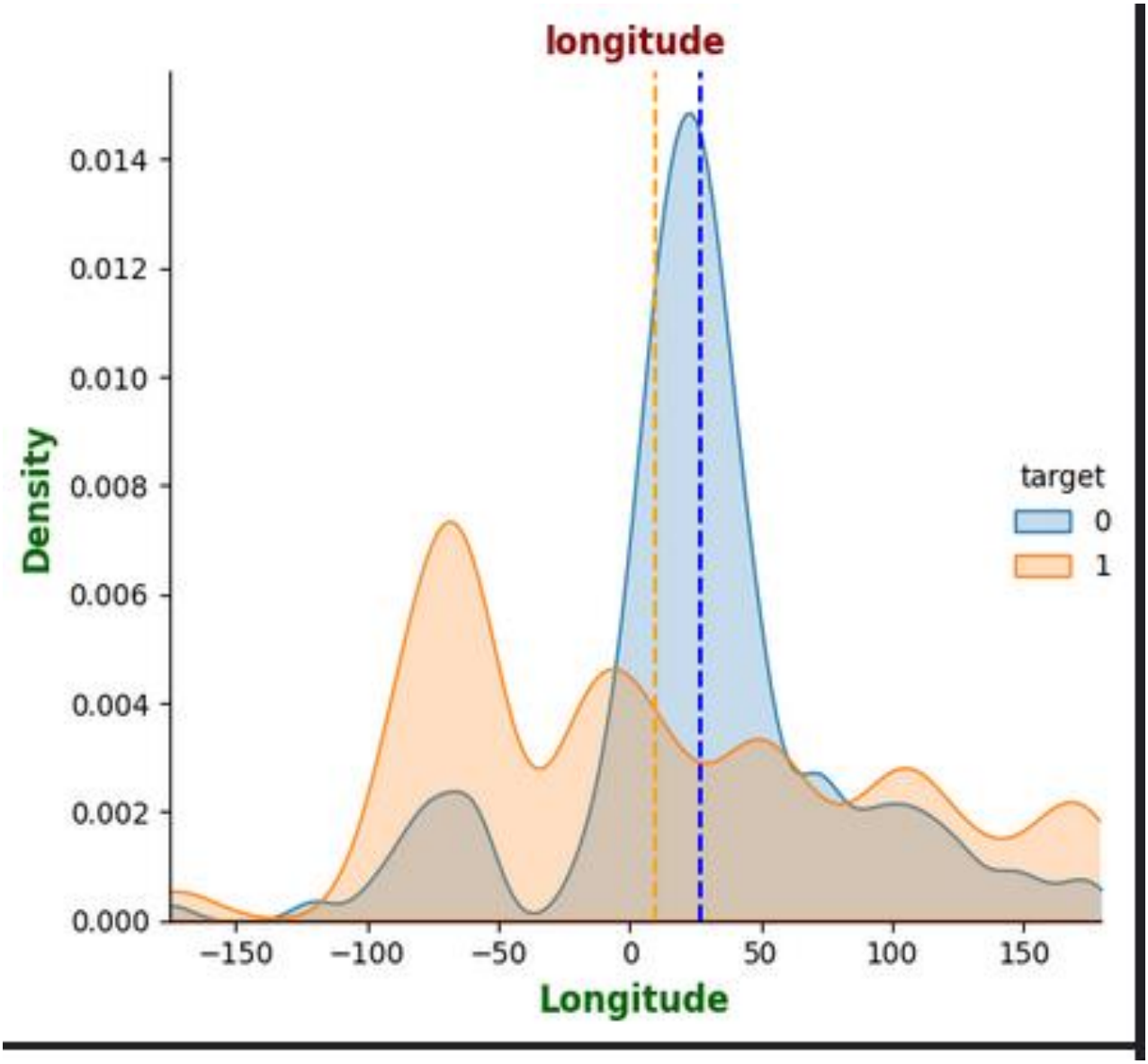
- **Most Common Wind Directions**: The North and South wind directions are the most prevalent, with North winds accounting for 33.4% and South winds accounting for 29.5% of the dataset.
- **Less Common Wind Directions**: East and West winds are less common, with East winds making up 22.9% and West winds the least at 14.3%.



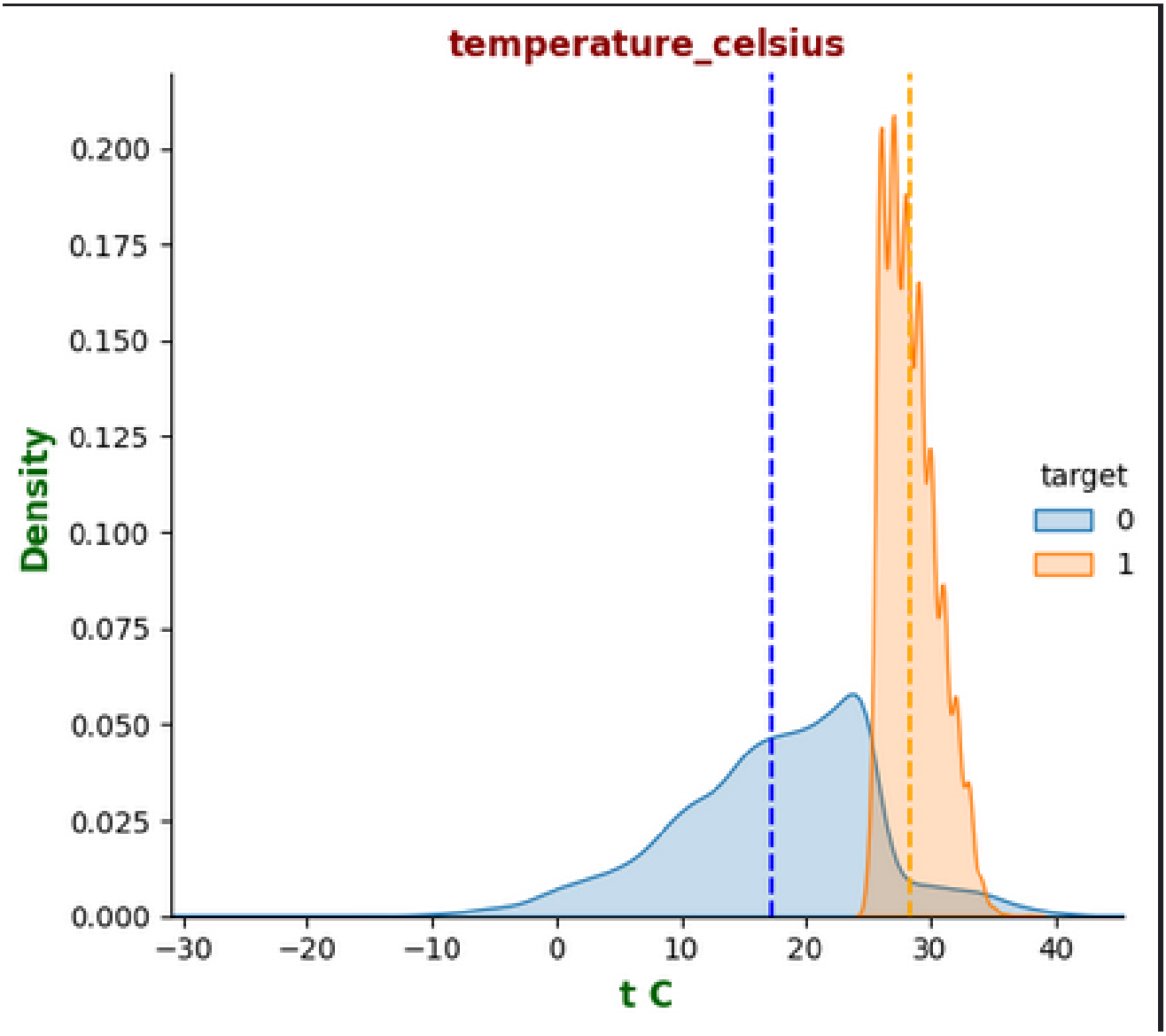
- **Most Common Phases**: The Waxing Gibbous (24.5%) and Waning Gibbous (21.3%) phases are the most represented in the dataset.
- **Next Common Phases**: The Waxing Crescent and Waning Crescent phases are also common, each accounting for just over 19% of the dataset.
- **Full and New Moons**: Full Moon and New Moon phases are less common, showing up in 5.3% and 4.3% of the dataset, respectively.
- **Quarter Phases**: The First and Last Quarter phases are the least represented, each constituting 3.2% of the dataset



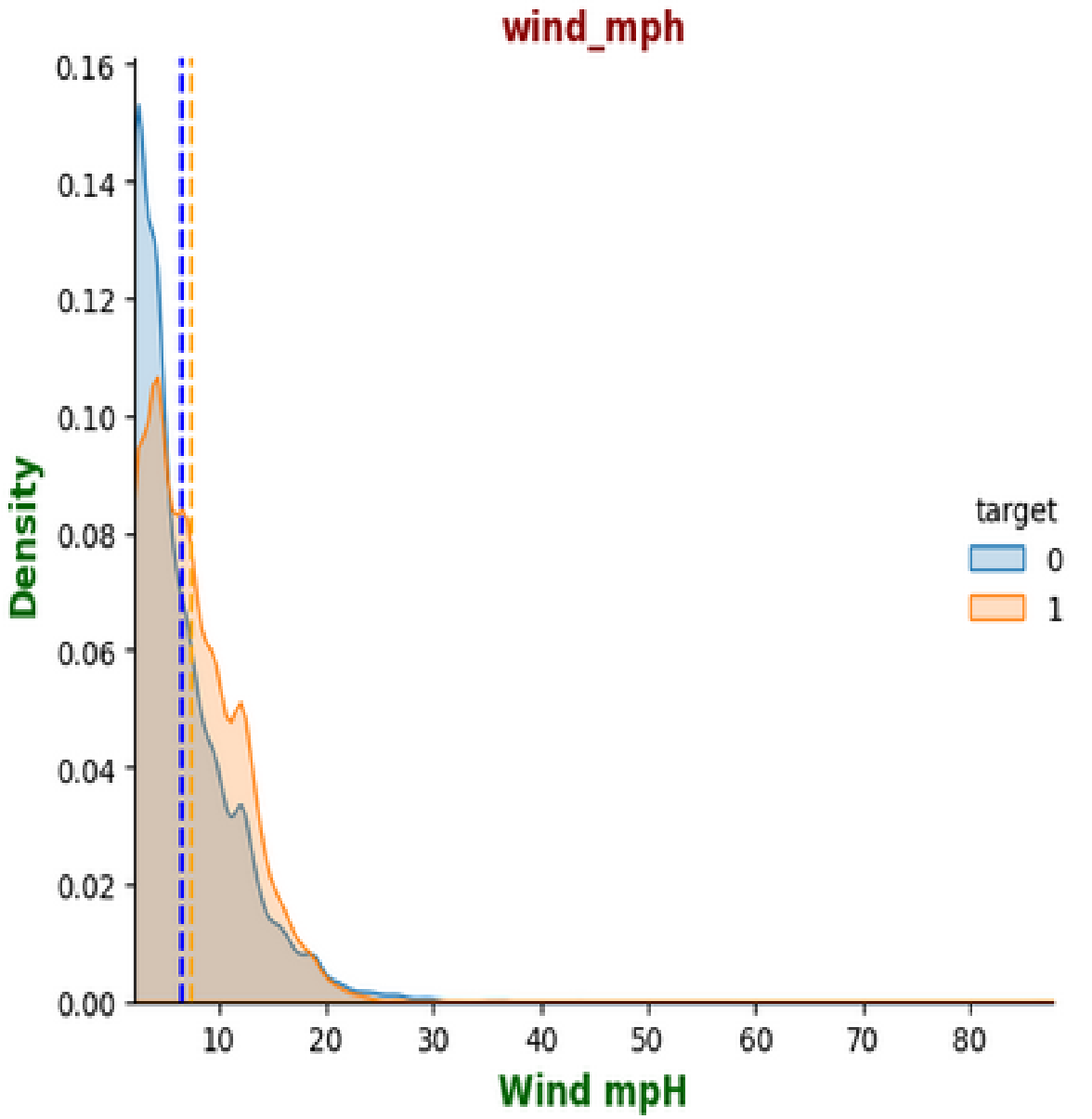
- **Central Tendency**: Each distribution has a central peak, with the peak for 'target 0' at around 10 degrees latitude and the peak for 'target 1' at around 50 degrees latitude.
- **Spread and Variability**: The distribution for 'target 0' is wider, suggesting a greater spread of latitude values, while 'target 1' is more narrowly concentrated around its central peak.
- **Skewness**: Both distributions appear slightly skewed, with 'target 0' skewed to the right and 'target 1' skewed to the left.
- **Overlap**: There is some overlap between the two distributions, particularly around the equator (0 degrees latitude).
- **0 is where weather warning is 0**
- **1 is where there is a weather warning appearing**



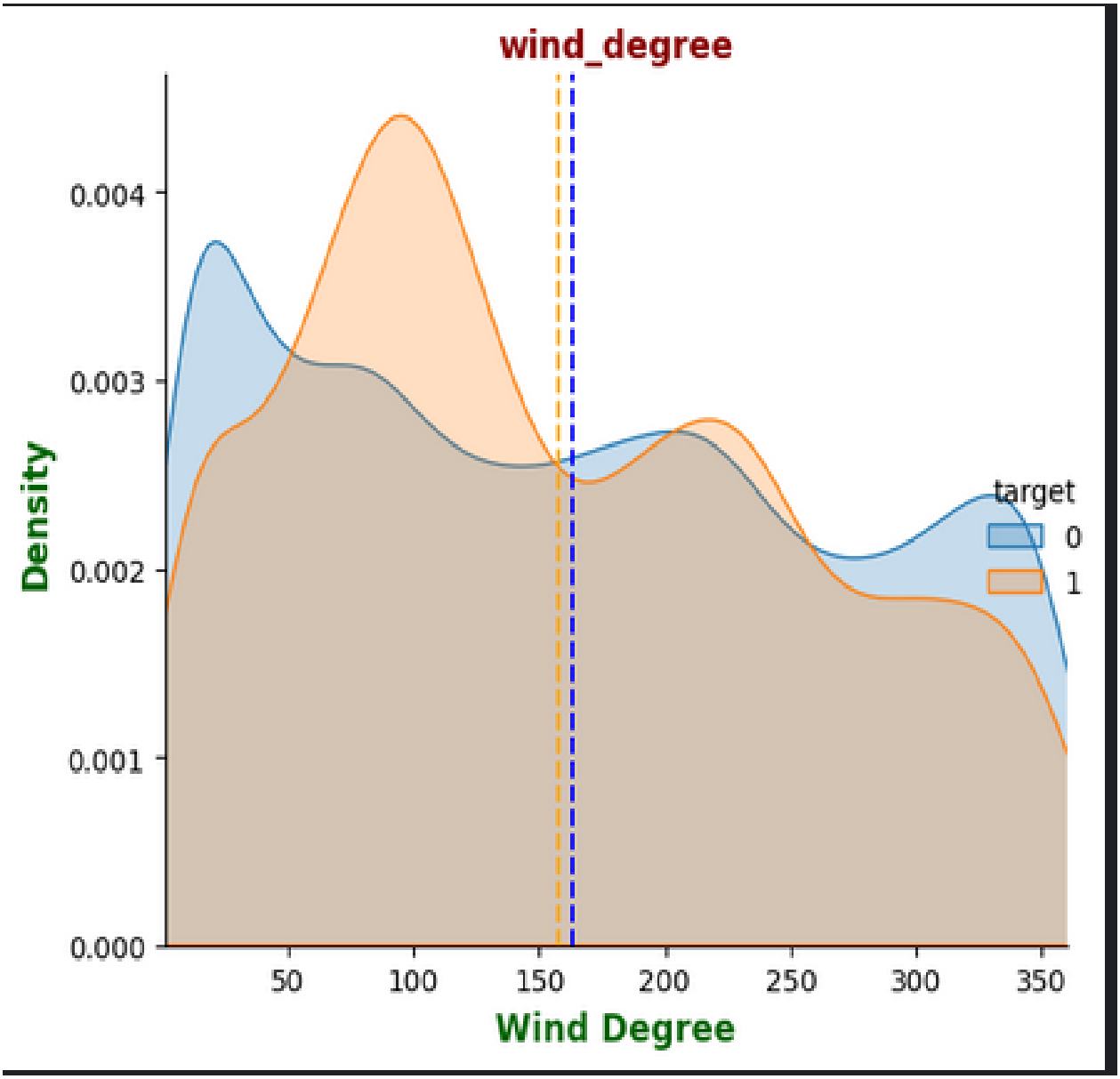
- **Central Tendency**: The distribution for 'target 1' peaks sharply around 0 degrees longitude, whereas 'target 0' has a broader peak that also centers near 0 degrees but with a much wider spread.
- **Spread and Variability**: 'Target 0' has a more extended distribution across longitude values, indicating a global spread, while 'target 1' is concentrated around the Prime Meridian.
- **Skewness**: Both distributions appear slightly skewed; 'target 0' shows a long tail stretching towards both ends of the longitude scale, while 'target 1' has a steep peak and a quick drop-off, indicating less variability.
- **Overlap**: There is overlap between the two distributions around the Prime Meridian, suggesting that for both targets, this longitudinal range is significant.



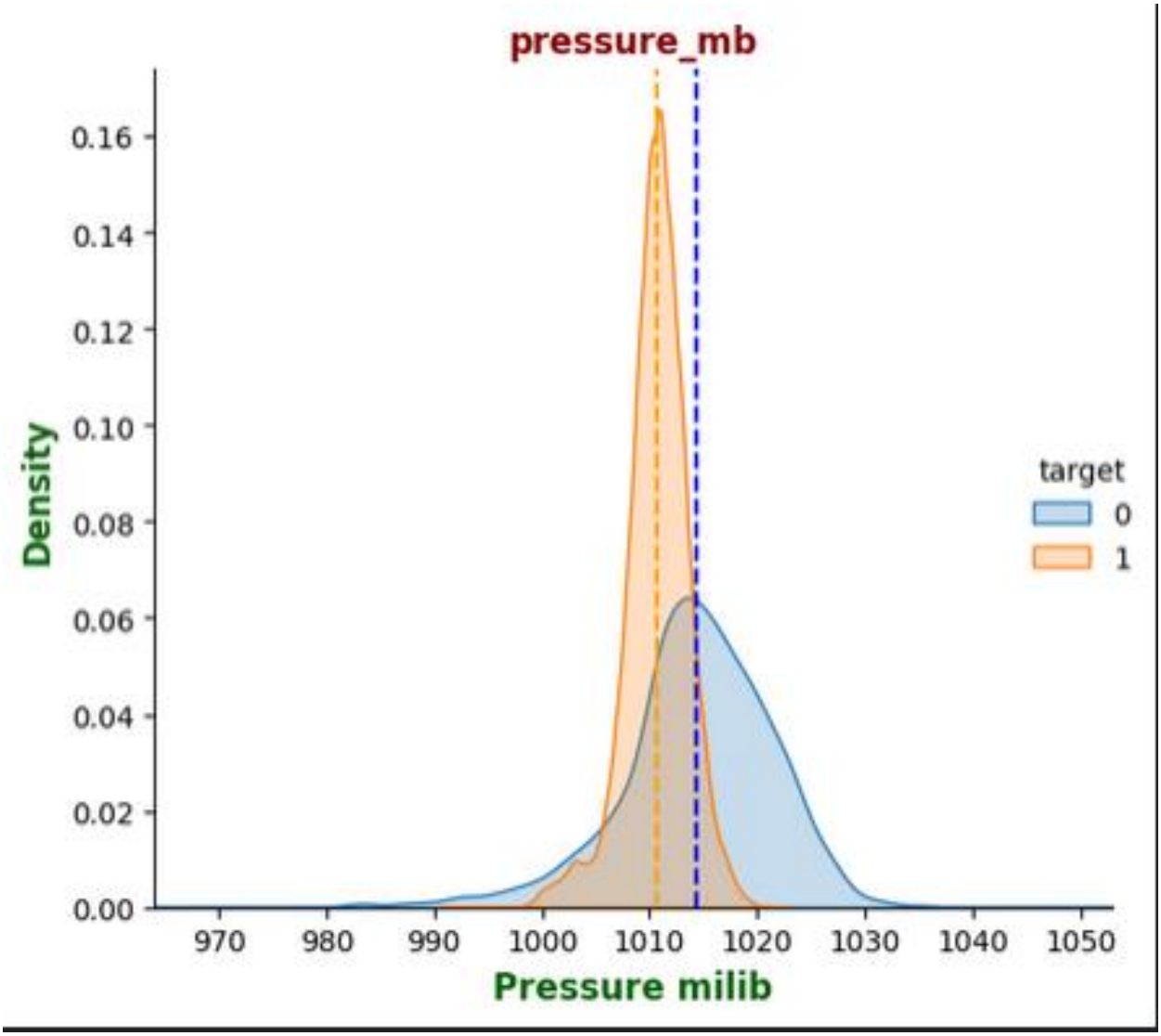
- **Central Tendency**: The median for 'target 1' is much higher than for 'target 0', indicated by the dashed lines; 'target 1' has its median around 20°C, while 'target 0' is around 0°C.
- **Spread and Variability**: 'Target 0' shows a broad distribution, suggesting a wide range of temperature occurrences, while 'target 1' has a narrower and taller peak, indicating more consistent temperature readings centered around its median.
- **Skewness**: 'Target 0' is skewed to the right, with temperatures mostly above its median, while 'target 1' has a slight left skew.
- **Range**: 'Target 0' encompasses a wider temperature range, potentially indicating diverse geographical locations or times of year, while 'target 1' is more concentrated, which could suggest a specific climate or season.



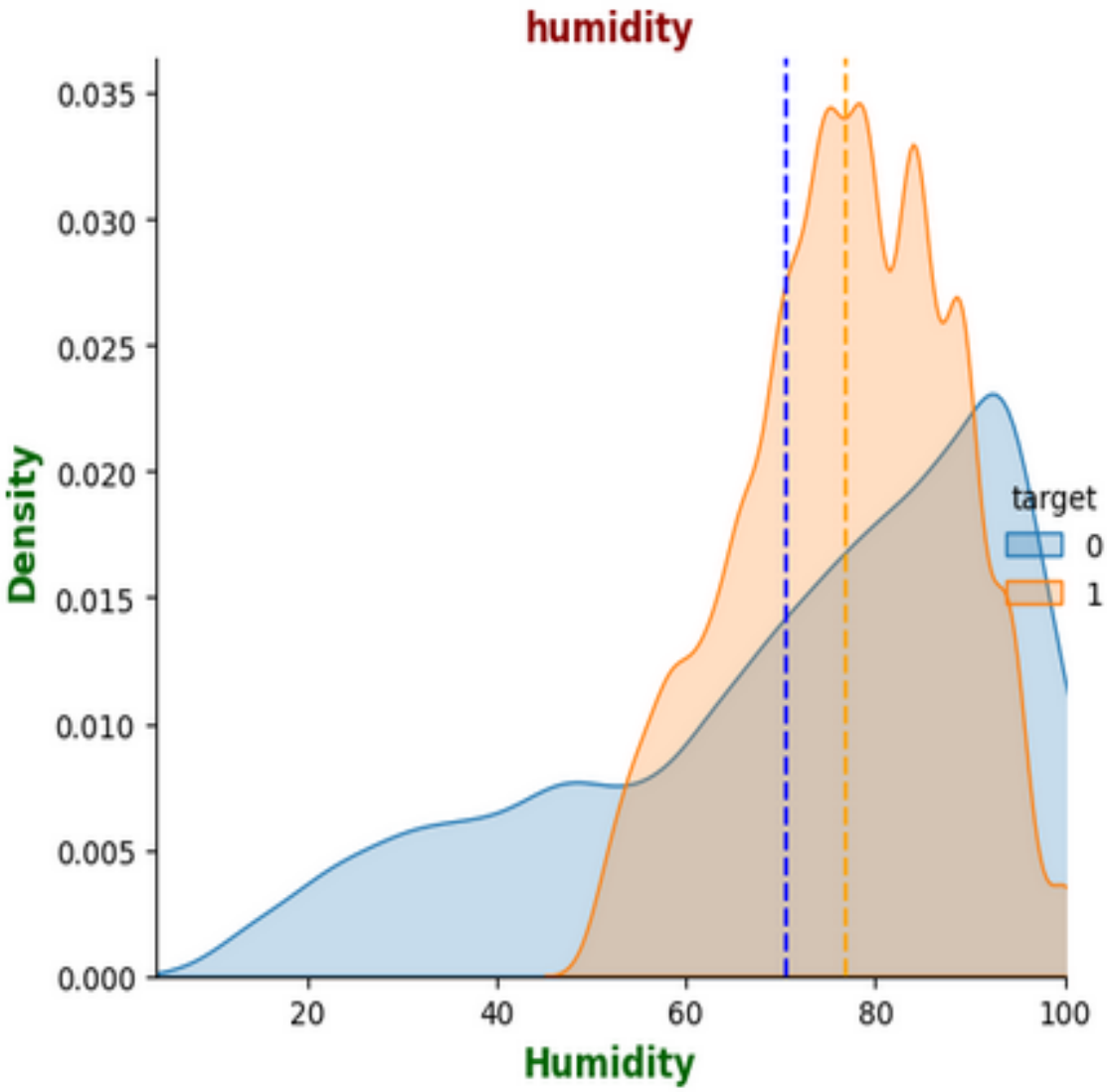
- **Central Tendency**: The median for 'target 1' is closer to 0 mph compared to 'target 0', which has a median around 10 mph.
- **Spread and Variability**: 'Target 0' has a broader distribution with a wider range of wind speeds, while 'target 1' has a steeper and more concentrated distribution near lower wind speeds.
- **Skewness**: Both distributions are right-skewed, indicating that lower wind speeds are more common than higher wind speeds for both targets. However, 'target 0' has a longer tail, suggesting that it occasionally experiences much higher wind speeds.
- **Range**: 'Target 0' spans a greater range of wind speeds, potentially indicating a variety of conditions such as calm days and stormy weather, whereas 'target 1' is concentrated in lower wind speeds, likely representing calmer conditions.



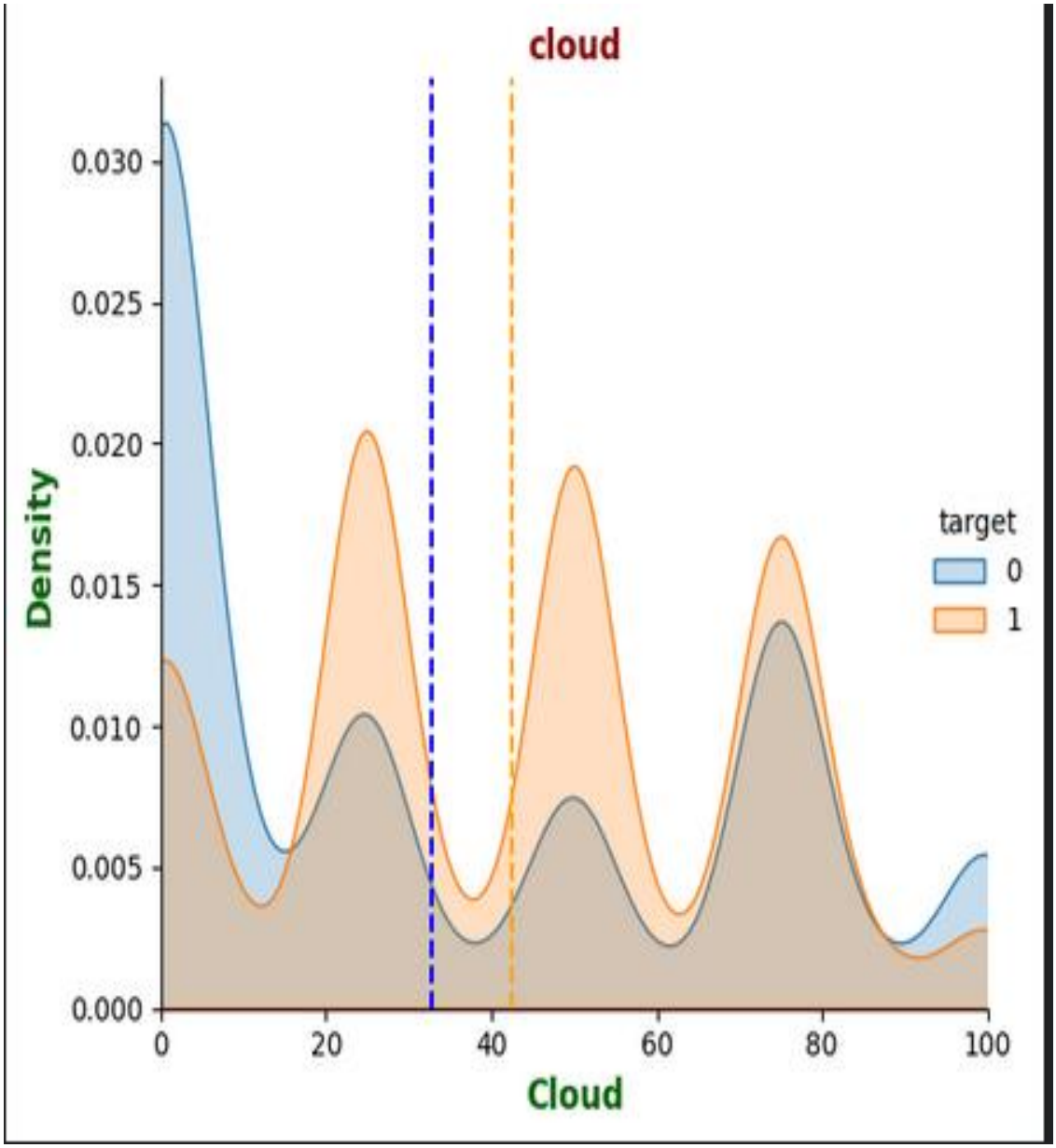
- **Central Tendency**: It's difficult to determine the central tendency for 'target 0' as its median or mean is not marked, but for 'target 1', it appears to be around 180 degrees (south).
- **Spread and Variability**: The distribution for 'target 0' is broad, with a more uniform spread across all directions, while 'target 1' has a narrower distribution with a peak around 180 degrees.
- **Directional Preference**: 'Target 1' shows a preference for southerly winds, while 'target 0' seems to have a more even distribution without a strong directional preference.



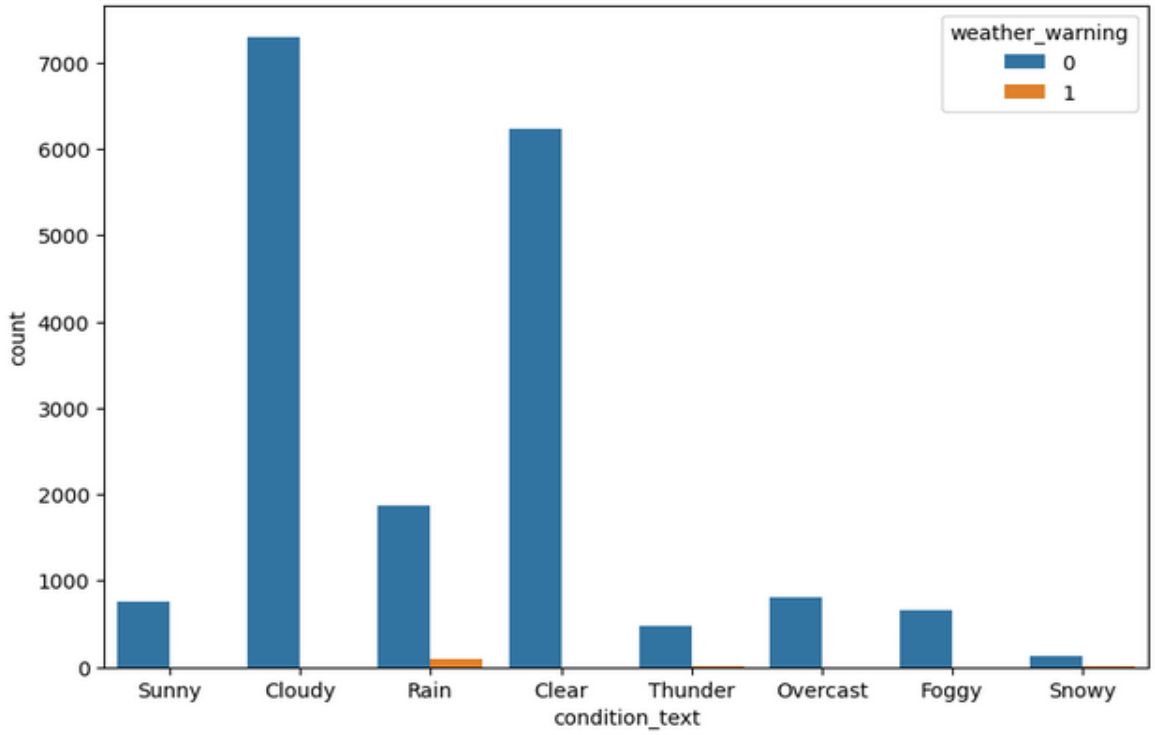
- **Central Tendency**: The median/mean for 'target 1' is around 1010 mb, typical for standard atmospheric pressure at sea level, while 'target 0' has a median/mean slightly lower than 1000 mb.
- **Spread and Variability**: 'Target 1' shows a steeper and taller distribution, indicating that its pressure values are more concentrated around the median. In contrast, 'target 0' has a broader distribution, suggesting a wider range of pressure values.
- **Symmetry**: The distribution for 'target 1' is symmetric around its median, while 'target 0' shows a slight skew towards lower pressures.
- **Range**: The range for 'target 0' suggests that it includes locations or conditions with both relatively high and low atmospheric pressures, while 'target 1' is more narrowly focused around a common atmospheric pressure.



- **Central Tendency**: For 'target 1', the central value indicated by the dashed line is higher, around 80% humidity, while for 'target 0', it is closer to 60% humidity.
- **Spread and Variability**: 'Target 0' has a broader distribution, indicating variability in humidity levels. In contrast, 'target 1' has a narrower peak, suggesting a more consistent range of higher humidity levels.
- **Skewness**: Both distributions are right-skewed, but 'target 1' has a sharper peak and a more pronounced decline after its median, indicating a concentration of data around higher humidity levels.



- **Central Tendency**: 'Target 1' has a median/mean value indicated by the dashed line around 50%, whereas 'target 0' has its median/mean closer to 20%.
- **Spread and Variability**: Both distributions show multiple peaks, suggesting periodicity in cloud cover data, which could correspond to the cyclic nature of cloud formation.



And finally,

- ***Sunny and Clear Conditions***: These conditions have the highest counts, indicating that they are the most frequently observed conditions in the dataset. However, there are very few instances where these conditions have coincided with a weather warning. This could suggest that sunny and clear weather rarely becomes excessively hot and humid.
- ***Cloudy and Overcast Conditions:*** These conditions are also common but less so than sunny and clear conditions. There are a few instances of weather warnings under cloudy conditions, implying that sometimes, even without direct sunlight, the temperature and humidity can exceed the specified thresholds for a weather warning.
- ***Rain and Thunder Conditions:*** Rainy weather and thunder conditions have fewer occurrences compared to sunny and clear conditions. Weather warnings are almost non-existent for thunder conditions and not very common for rain, which suggests that rain and thunder conditions may not often lead to high temperatures and humidity levels above the stated thresholds.
- ***Foggy and Snowy Conditions:*** These have the lowest counts, indicating they are the least common weather conditions in this dataset. There are very few or no weather warnings associated with these conditions, likely because both foggy and snowy weather are associated with lower temperatures, which would not typically trigger a warning based on the criteria of high temperature and humidity.
- ***Weather Warning Analysis:*** The almost uniform absence of weather warnings across most weather conditions except sunny and cloudy may indicate that the criteria for a weather warning (temperature above 25 degrees and humidity above 50%) are not commonly met. It might be interesting to look at the distribution of temperature and humidity values to understand how often these conditions are approached or exceeded.
- ***Statistical Consideration:*** From a statistical perspective, one might analyse the probability of a weather warning given the weather condition, or conversely, the likely weather condition given a weather warning. This would involve calculating conditional probabilities and could yield insights into the weather patterns most associated with extreme temperature and humidity.

Methodologies and algorithms

In the global weather report project, three machine learning algorithms were employed to analyse and predict various weather-related outcomes. Here's how you might describe their usage, based on our conversation:

K-Nearest Neighbours (KNN): The KNN algorithm was applied as a classification tool to predict specific weather events, potentially hazardous conditions, or categorical weather statuses. Given its non-parametric nature, KNN made predictions based on the similarity of the input features to the nearest examples in the training data. It was particularly useful for classifying weather patterns where the relationships between data points were more spatially defined rather than linear, making it well-suited for complex patterns that are common in meteorological data. KNN's performance was evaluated against the Logistic Regression model, with accuracy being the primary metric. It showed a high degree of accuracy, suggesting a strong capability to capture the nuances in the weather dataset.

Linear Regression: Linear Regression was utilized for regression tasks within the dataset, such as predicting continuous variables like temperature or pressure based on other features. It worked on the principle of fitting the best linear relationship between the dependent variable and one or more independent variables. This approach was valuable for understanding which factors most strongly influenced a particular weather outcome and for making quantitative forecasts. The model's effectiveness was measured using the Mean Squared Error (MSE), which quantified the average squared difference between the predicted and actual values, providing a clear metric for the precision of the regression model's predictions.

Logistic Regression: Logistic Regression was employed as a classification technique to predict binary outcomes, such as the presence or absence of a weather warning. This model was particularly relevant for scenarios where the response variable was categorical. Logistic Regression made predictions based on the probability of the occurrence of an event, which was useful for decision-making processes in weather forecasting, such as issuing warnings for extreme conditions. The algorithm's performance was quantified using accuracy, and in our discussion, it demonstrated an excellent capability, achieving a perfect score in the context provided, though such a high score warrants further investigation to confirm it wasn't due to overfitting or data leakage.

Each model provided unique insights into the dataset and contributed to a robust analysis of the weather conditions. The KNN and Logistic Regression models helped classify conditions based on historical patterns, while Linear Regression offered predictions of specific weather metrics. By employing these models in tandem, a comprehensive understanding of the dataset's characteristics was achieved, ensuring robust predictive capabilities for both classification and regression tasks in the realm of global weather forecasting.

Results and Discussion

- From our results we will find out that the logistic regression is 100% accurate and through different analysis following latitude and longitude information given through graphs you can see that humidity is more close to regions where the are sea or in other words equatorial regions have more humidity and temperature.
- While comparing Logistic Regression and K-Nearest Neighbours I've found that the different between both is 2,54% and the the difference is small but it is possible that their close due to such wide dataset provided and so much information been processed.
- From one of the fixed conditions for the weather to find out how much global locations (18320) are under weather warning where we stated that the temperature should be above 25 degrees and above 50% humidity and there are 5387 locations under weather warning! Therefore proofing how weather forecasts/analytics are important. Of course, more countries are below the conditions and not under weather warning conditions (which are 12932) and still we did not added additional conditions for example for cold regions only for more warm regions and tropical regions or seasonally warm/hot regions.
- After running and finding Linear Regression performance the Linear Regression Mean Squared Error is 0.11. That was required search as linear regression in our case uses the "mean_squared_error" function from scikit-learn to calculate the mean squared error between the actual target values ("y_test_reg") and the predicted values ("y_pred_reg") obtained from the linear regression model. (May be found through out the code).
- Through the analysis by using pie charts we founded out that most of the wind direction are heading north or to backwards south from all locations globally, around the whole globe the maximum percentage of weather condition is cloudiness at nearly 40%(39,9%) but at the same time on 2nd place in the list of weather conditions is the clear weather around the world.