**Impact of Climate Indicators on Visibility**

Andrew Fennimore, Birendra Khimding, Muzhgan Rustaqui

Department of Engineering, University of San Diego

AAI-500 Probability and Statistics for Artificial Intelligence

Dallin Munger, M.S.

February 23, 2025

**Summary**

**Impact of Climate Indicators on Visibility**

Predicting visibility is crucial for ensuring safety across various industries, including aviation, transportation, and climate safety. According to Zhang et al. (2022), visibility, simply put, is a measure of atmospheric transparency. Measuring it is vital for safety in industries such as aviation, transportation, and environmental monitoring (Kadam et al., 2023). According to Ortega et al. (2022), poor visibility conditions are linked to approximately 31,500 traffic accidents yearly in the United States, leading to 11,500 injuries and 500 fatalities. In addition, measuring visibility can be an important indicator of air quality and pollutants like aerosols (Liang et al., 2023). This is especially vital in poorer regions where measuring instrumentation and data are unavailable (Liang et al., 2023). By examining the relationships between climate indicators such as weather and humidity, our project aims to help stakeholders implement effective safety measures proactively.

The most common method for measuring visibility is by using scattered visiometers to measure the distance that can be seen (Liang et al., 2023). Airports measure visibility through a system known as RVR, or runway visual range, which provides a consistent method to determine the distance a pilot can expect to see (U.S. Department of Transportation, 2024). Measuring equipment can be seen along runways and is used to determine if pilots can fly safely during possibly unsafe weather conditions (U.S. Department of Transportation, 2024).

This study aims to use easily accessible variables to ensure that our model can be universally applied regardless of location or wealth. The dataset used to build the model was obtained from Ocean County Airport (MJX) and comprises 11,423 samples. This dataset includes various climate indicators, including temperature, humidity, wind speed, precipitation, and atmospheric pressure. We not only aim to determine if there is an association between visibility and the climate indicators but attempt to develop a model to predict future visibility.

**Null Hypothesis (H₀):** Climate indicators do not have any significant impact on visibility.

**Alternative Hypothesis (H₁):** At least one of the climate indicators has a significant impact on visibility.

Table 1

*Columns and description of dataset*



*Note. Iowa State University College of Ag. (2025). ASOS-AWOS-METAR Data Download. Iastate.edu. https://mesonet.agron.iastate.edu/request/download.phtml?network=NJ\_ASOS*

**Cleaning and Preparation**

Of the total variables found in the dataset, 6 of the variables had no datapoints, including skyc4, skyl4, 3 stages of ice\_accretion and snowdepth. These variables can be immediately removed. Other variables that were of no obvious use included “valid” and “metar” which were dates with only unique values and “station” which had only 1 unique value.

There were several areas of difficulty that were encountered when analyzing the data. This includes dealing with missing variables, dealing with a heavy left skew, and addressing the large quantity of variables used for analysis. Not properly dealing with one of these issues could lead to erroneous conclusions. To address these issues, the team needed to use several methods and statistical tools even before analysis.

When dealing with missing values, it is imperative that they are addressed correctly. Not doing so can lead to biased estimates or reduce statistical power, leading to invalid conclusions from the data (Kang, 2013). The data set uses “M” for all missing values and treated as NA. This model assumes MCAR for all variables with a threshold above 15%. MCAR, or missing completely at random, is the ideal assumption and used when missing values are missing only due to random issues (Kang, 2013). This could include issues like instrument failure for a day or power being lost (Kang, 2013). The assumption of MCAR permitted us to omit cases that were null in our data. This is the most common approach when data is missing but may lead to false power of variables, even if the MCAR assumption is true (Kang, 2013). The data set uses “M” for all missing values and treated as NA within the dataset.

We noticed a negative skew with the dependent variable “vsby”. There are no outliers that would impact the distribution of the variables, so thoughts of using a generalized linear model were entertained even before any analysis. Scatterplots would later confirm the need for further data transformation.

Our dataset only consisted of 11,423 samples; however, this is still far larger than others discussed and utilized in class. According to Fan et al. (2014), larger datasets can be useful for determining patterns and providing better oversight than smaller ones. However, some challenges with bigger datasets include scalability, spurious correlations, ad measurement errors (Fan et al., 2014). It also proved to be somewhat taxing on our computational efforts and required careful adjustments to code.

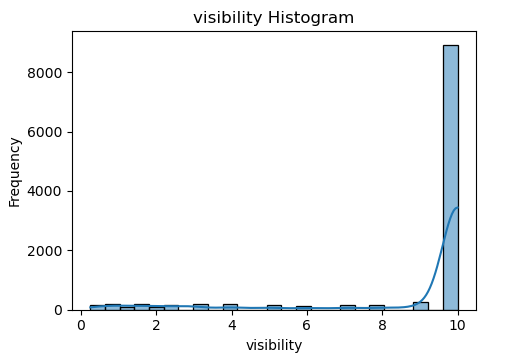
**Exploratory Data Analysis**

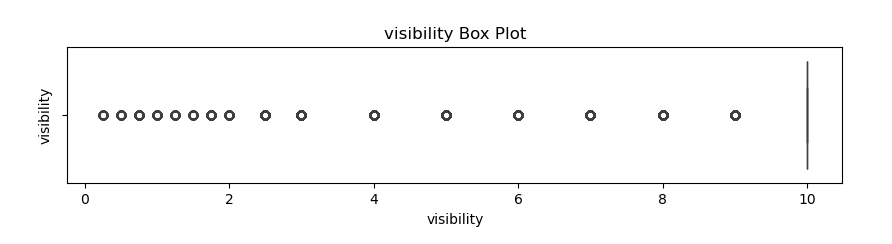
Table 2 provides an overview of the descriptive statistics for the dataset, as well as other pertinent information. Results with greater than 15% null value were removed from analysis, as well as the other variables previously mentioned.

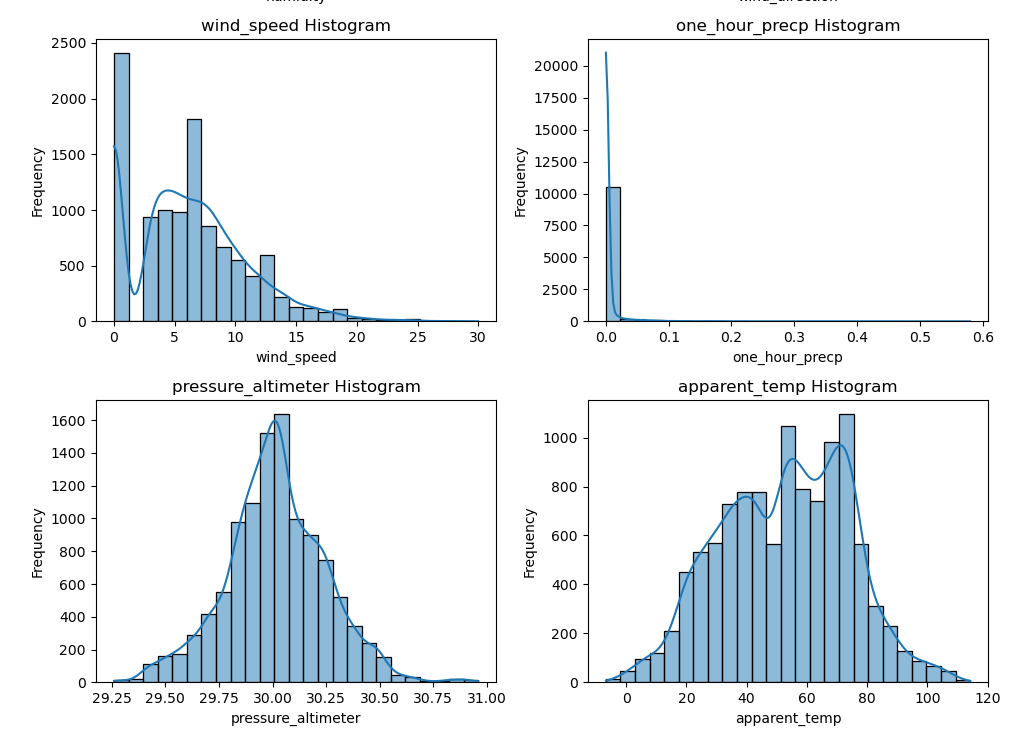
Looking at the dependent variable “vsby” we see a median of 10 miles. This is also equivalent to the highest value. This further demonstrates the need to transform some of our variables when we start to begin analysis.

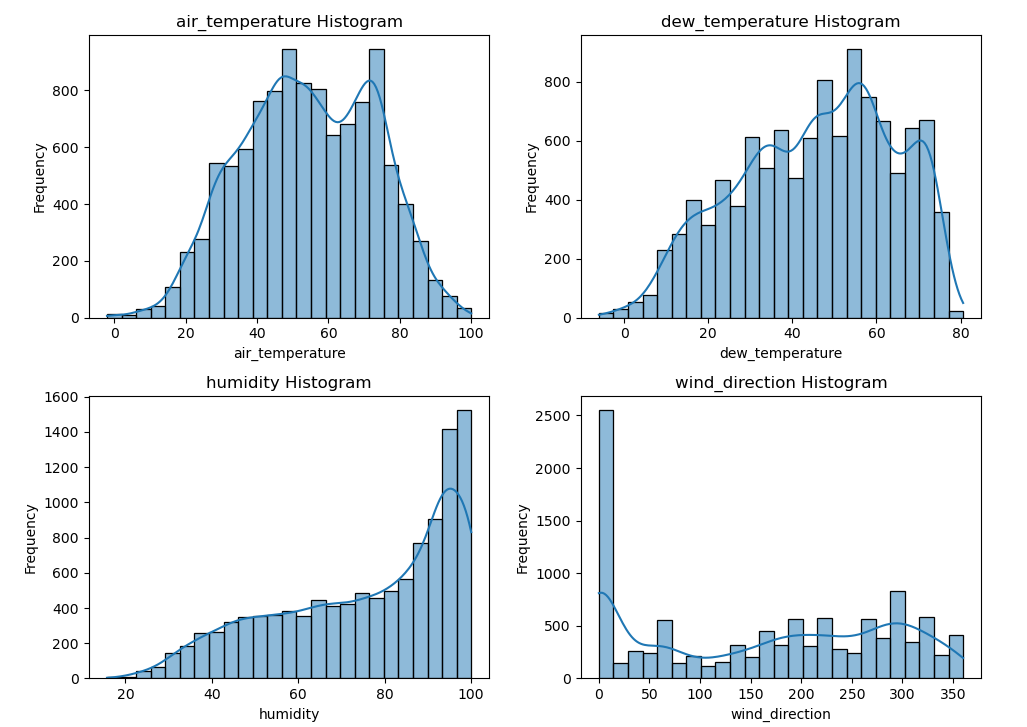
****

Table 2 Descriptive Statistics of variables









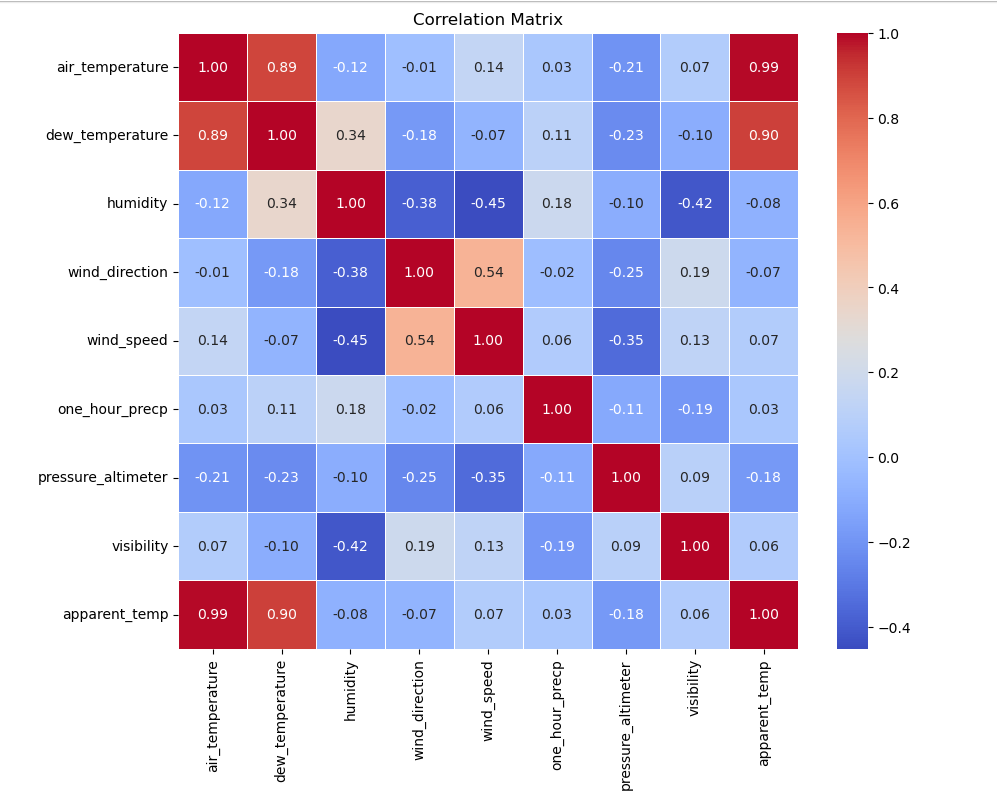
**Visualization**

We can see from graph 3 and graph 4 verification of the negative skew. This will likely make it difficult to create a good fit for a model. According to Dugan and Greyserman (2019), a negative skew may initially give a false impression of an improved model, while a positive skew has the opposite effect. Data transformation will be necessary for a best fit model.

The histograms for the predictor variables show a wide range of distributions. While pressure, apparent temperature, air temperature and wind direction show a fairly bell-shaped normal distribution, wind speed shows a strong positive skew. Humidity demonstrates the opposite distribution and shows a positive skew, and one hour precipitation shows values all clustered at 0. Fortunately, no assumptions need to be made of the distributions of explanatory variables when dealing with linear models and least squares (Agresti & Kateri, 2021).

In the correlation matrix for Table 5, it's evident that certain variables, such as dew temperature, air temperature, and apparent temperature, exhibit multicollinearity. Multicollinearity refers to when explanatory variables showing some overlap and demonstrating redundant values (Agresti & Greyserman, 2019). Its effects can be seen when two predictors that are highly correlated are assessed at the same time in a regression model, and ignoring this can lead to misleading interpretations of results (Vatcheva & Lee, 2016).

Depending on what the research is looking for would depend on how to deal with the issue. According to Vatcheva and Lee (2016), Multicollinearity will not impact the fit of the model. If however, we are looking to investigate associations, multicollinearity can obscure effects of an independent variable on the outcome variable (Vatcheva & Lee, 2016). Given their high correlations, it would be prudent to eliminate two of these variables and retain only one. This approach will provide a more accurate assessment of statistical power.



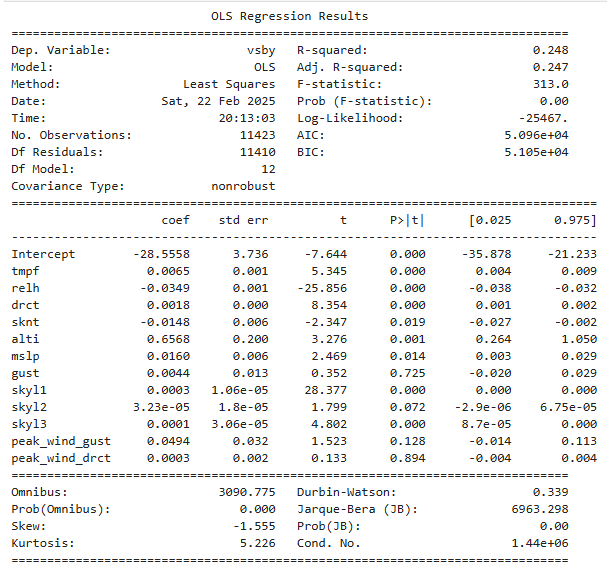
**Creating and Analyzing Models**

**Ordinary Least Squares Model**

Ordinary Least Square Models attempt to find a linear relationship between a dependent variable (vsby) and the independent variables(Agresti & Katari, 2021).Several assumptions are necessary for an ordinary least squares model. According to Williams et al (2013), these include being unbiased, consistent and efficient. Being unbiased according to Williams et al. (2013), is the mean of a sample is the same as the true parameter of the mean of the population. It is basically stating that the samples that are obtained must represent the population as a whole. Consistent means that as a sample size increases, so does its accuracy (Williams et al., 2013). Efficiency then is accuracy of the samples. Normality of the residuals, or the difference between the observed and predicted values, is also a necessity in ordinary least squares modeling (William et al., 2013).

Our model does not fit these assumptions well. We do not see normality of residuals, making it difficult to accurately fit a regression model. However, we can use it to compare to the generalized linear model and determine what transformations are necassary. These results however, should be used with caution, as the assumptions show some

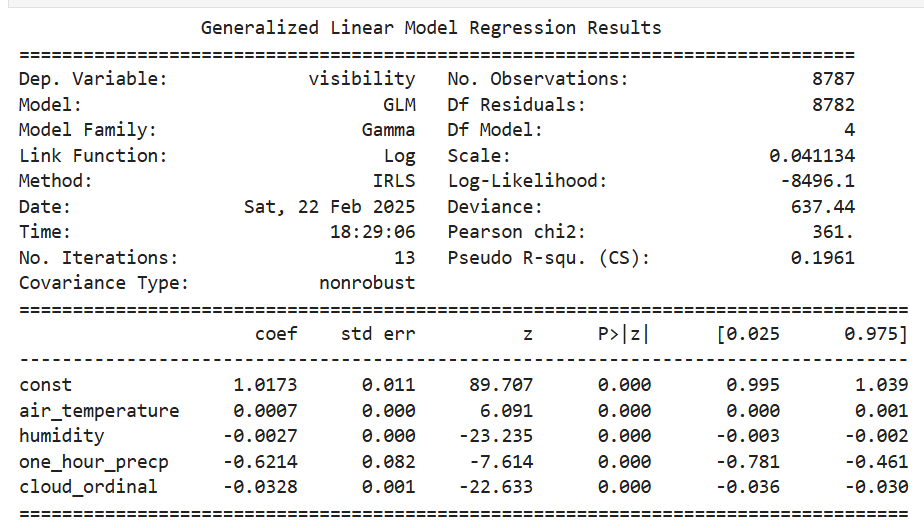
After removing the variables previously discussed, we are left with the following variables "vsby", "tmpf", "relh", "drct", "sknt", "alti", "mslp", "gust", "skyl1", "skyl2", "skyl3", "peak\_wind\_gust" and "peak\_wind\_drct". These are all quantitative values and don’t show any relationship with each other. If we use these to fit a not transformed linear regression, the results are as follows:

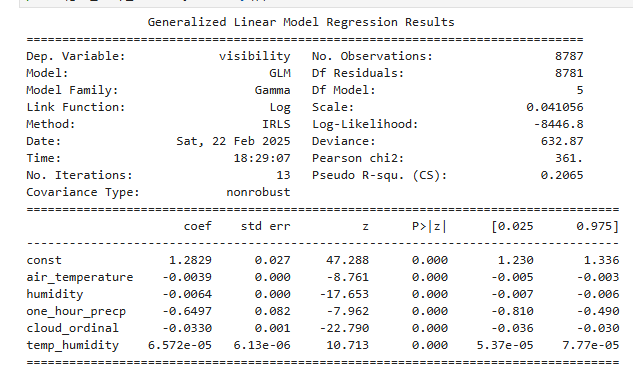


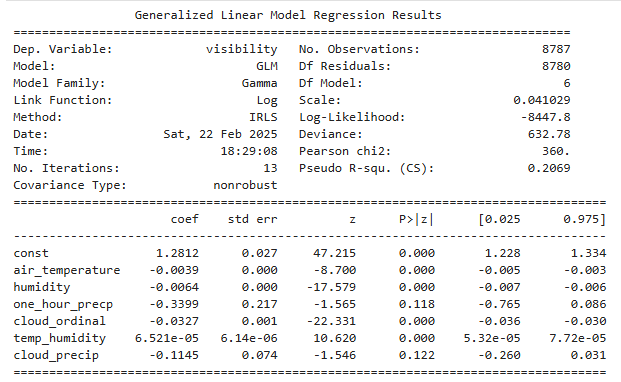
Looking at the results, we see that about 24.8% of the variability for the dependent variables can be explained using the proposed model. The F statistic shows us that the overall model is statistically significant. Of the variables used, tmpf, relh, drct, sknt, alti and mslp demonstrate significant effects on the dependent variable vsby.

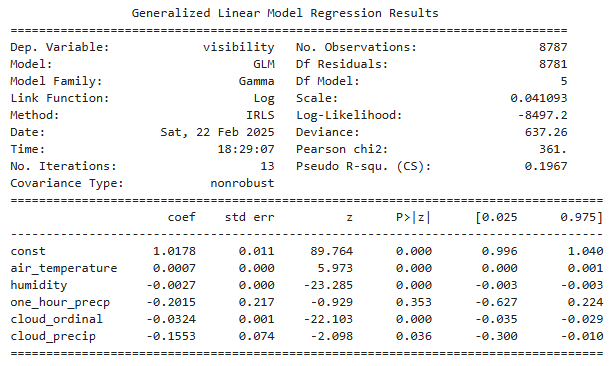
Due to this, we are able to reject our null hypothesis and say that at least one of the indicators has a significant impact on visibility.

**General Linear Model**

****

****

****



**Conclusions**

**References**

Agresti, A., & Kateri, M. (2021). *Foundations of Statistics for Data Scientists*. CRC Press.

Dugan, Z., & Greyserman, A. (2019). The impact of skew on performance and bias: How skew distorts short term performance, triggers bias, and changes drawdowns. *Journal of Behavioral and Experimental Finance*, *22*, 232–238. https://doi.org/10.1016/j.jbef.2019.03.008

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293–314. https://doi.org/10.1093/nsr/nwt032

Iowa State University College of Ag. (2025). *ASOS-AWOS-METAR Data Download*. Iastate.edu. https://mesonet.agron.iastate.edu/request/download.phtml?network=NJ\_ASOS

Kadam, G., Tobaria, A., Arya, S., Prof, A., & Kaushik, J. (2023). Climate Visibility Prediction Using Machine Learning. In *International Research Journal of Engineering and Technology*. https://www.irjet.net/archives/V10/i5/IRJET-V10I5281.pdf

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Liang, C.-W., Chang, C.-C., Hsiao, C.-Y., & Liang, C.-J. (2023). Prediction and analysis of atmospheric visibility in five terrain types with artificial intelligence. *Heliyon*, *9*(8), e19281. https://doi.org/10.1016/j.heliyon.2023.e19281

Ortega, L. C., Otero, L. D., Solomon, M., Otero, C. E., & Fabregas, A. (2022). Deep learning models for visibility forecasting using climatological data. *International Journal of Forecasting*, *39*(2). https://doi.org/10.1016/j.ijforecast.2022.03.009

U.S. Department of Transportation. (2024). *Evaluating the Performance of Runway Visual Range Sensors for FAA | Volpe National Transportation Systems Center*. Dot.gov. https://www.volpe.dot.gov/news/evaluating-performance-runway-visual-range-sensors-faa

Vatcheva, K. P., & Lee, M. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology: Open Access*, *06*(02). https://doi.org/10.4172/2161-1165.1000227

Williams, M., Alberto, C., Grajales, G., & Kurkiewicz, D. (2013). Assumptions of Multiple Regression: Correcting Two Misconceptions. *Practical Assessment, Research & Evaluation*, *18*(11). https://files.eric.ed.gov/fulltext/EJ1015680.pdf

Zhang, Y., Wang, Y., Zhu, Y., Yang, L., Ge, L., & Luo, C. (2022). Visibility Prediction Based on Machine Learning Algorithms. *Atmosphere*, *13*(7), 1125–1125. https://doi.org/10.3390/atmos13071125