ALY6010 PROBABILITY THEORY INTRODUCTORY STATISTICS
NORTHEASTERN UNIVERSITY
REPORT
PROMISE GBAMA
JUN 08, 2023

INTRODUCTION

The dataset originates from the official records managed by the National Hurricane Center (NHC), which can be accessed directly from the NHC's website. The version available on Kaggle, created by Utkarsh Singh in 2022[1], is a carefully curated adaptation of the original data provided by NHC.

The dataset is a valuable tool for researchers, meteorologists, climatologists, and individuals with an interest in Atlantic hurricanes. It provides a wealth of information that enables the study and analysis of historical hurricane patterns, the tracking of storm intensities over time, the assessment of hurricane impacts, and the identification of trends and shifts in hurricane behavior.

The NOAA Atlantic Hurricane dataset offers extensive information regarding Atlantic hurricanes spanning from 1975 to 2021. It encompasses a wide range of attributes and characteristics associated with storms, including storm names, dates, times, locations, classifications, wind speeds, air pressures, and other relevant details. The dataset comprises 19,067 rows and 13 columns, each representing a specific aspect of the hurricane data. These fields provide researchers and analysts with a comprehensive resource for studying and understanding Atlantic hurricanes over the specified time frame.

The description of each field is as follows:

- **Name**: This field contains the name of the storm in text format.
- **Year**: This field represents the year when the report was made, using numerical values.
- **Month**: It denotes the month when the report was recorded, using numerical values.
- **Day**: This field indicates the specific day of the report, using numerical values.
- **Hour**: It represents the hour of the report in Coordinated Universal Time (UTC), using numerical values.
- **Latitude** (Lat): This field provides the latitude of the storm center, expressed as numerical values.
- **Longitude** (Long): It indicates the longitude of the storm center, using numerical values.
- **Status**: This field describes the storm's classification, which can be Tropical Depression, Tropical Storm, or Hurricane. It is represented in text format.
- **Category**: It represents the Saffir-Simpson hurricane category calculated based on the storm's wind speed. The category can range from NA (Not a hurricane) to 1, 2, 3, 4, or 5, indicating increasing intensity. The category is represented using numerical values.
- **Wind**: This field specifies the storm's maximum sustained wind speed in knots, using numerical values.
- **Pressure**: It provides the air pressure at the center of the storm in millibars, using numerical values.
- **Tropical Storm Force Diameter**: This field denotes the diameter of the area experiencing tropical storm strength winds, using numerical values. This information is available starting from 2004.
- **Hurricane Force Diameter**: It represents the diameter of the area experiencing hurricane strength winds, using numerical values. This information is available starting from 2004.

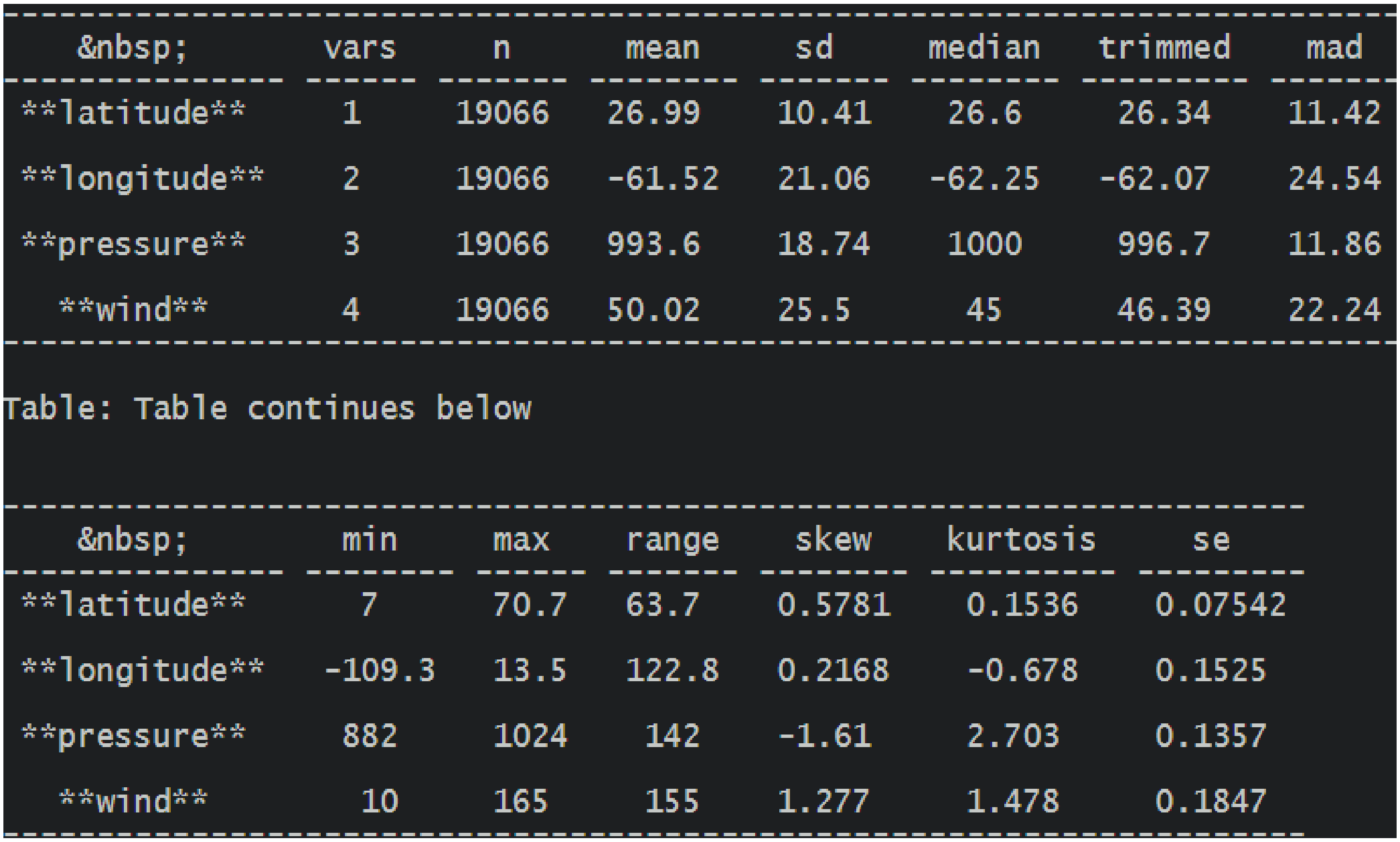
ANALYSIS PLAN

Data Cleaning

The provided code performs several common operations on the storms_data dataset. The colSums(is.na(storms_data)) calculates the count of missing values (nulls) in each column by creating a logical matrix with is.na(storms_data) and summing the logical values column-wise. The newdf dataframe is created by excluding three columns (-tropicalstorm_force_diameter, -hurricane_force_diameter, -category) from the original storms_data dataset using the select() function from the dplyr package. The duplicated(newdf) checks for duplicate rows in newdf and returns a logical vector indicating the duplicated rows. The glimpse(newdf) displays a concise overview of newdf's structure, including column names, data types, and a preview of the data. The summary(newdf) provides a summary of the variables in newdf, presenting statistical measures such as minimum, maximum, quartiles, means, and counts for numeric and categorical variables.

Additionally, the code demonstrates variable renaming in the newdf dataframe. By utilizing the rename() function from the dplyr package, the code updates the variable names for improved clarity and interpretation. Specifically, the latitude variable is renamed from lat, longitude from long, and storm_name from name. This renaming process enhances data understanding and facilitates easier analysis by using clear and descriptive variable names.

Chart: Table Shows the Descriptive Statistics of Numeric Variables.

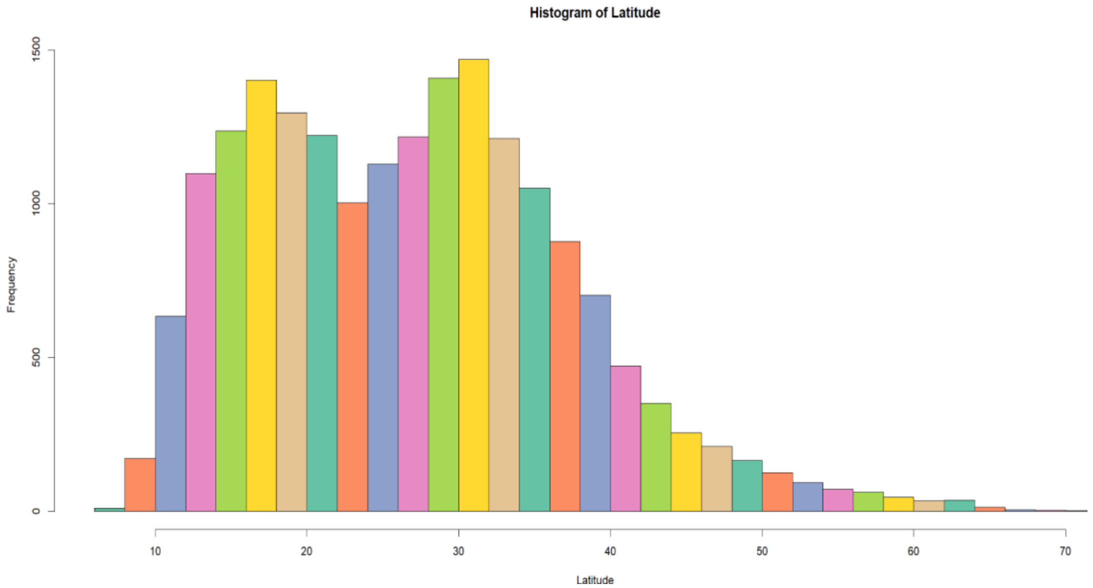


```
-----------------------------------------------------------------------------
    &nbsp;        vars     n       mean      sd     median   trimmed    mad
-------------- ------ ------- -------- ------- -------- --------- -------
 **latitude**     1    19066   26.99    10.41    26.6     26.34    11.42

 **longitude**    2    19066   -61.52   21.06   -62.25   -62.07    24.54

 **pressure**     3    19066   993.6    18.74    1000     996.7    11.86

   **wind**       4    19066   50.02    25.5      45      46.39    22.24
-----------------------------------------------------------------------------

Table: Table continues below


----------------------------------------------------------------------
    &nbsp;         min      max    range    skew    kurtosis     se
-------------- -------- ------ ------- -------- ---------- ---------
 **latitude**      7      70.7    63.7   0.5781    0.1536    0.07542

 **longitude**  -109.3    13.5   122.8   0.2168    -0.678    0.1525

 **pressure**     882     1024    142    -1.61     2.703     0.1357

   **wind**        10     165     155    1.277     1.478     0.1847
----------------------------------------------------------------------
```

Observation: The table displays descriptive statistics for the selected numerical variables in the dataset. It provides information such as the number of observations, average value, spread of the data, middle value, trimmed mean, median absolute deviation, minimum and maximum values, range, skewness, and kurtosis. These statistics give insights into the central tendency, variability, shape of the distribution, and presence of extreme values for each variable. For example, mean and median provide measures of the average and middle values, standard deviation indicates the extent of variation, and skewness helps assess the symmetry of the data. Kurtosis indicates whether the data is more peaked or flat compared to a normal distribution. Overall, the table shows the descriptive statistics of numeric varaiables.

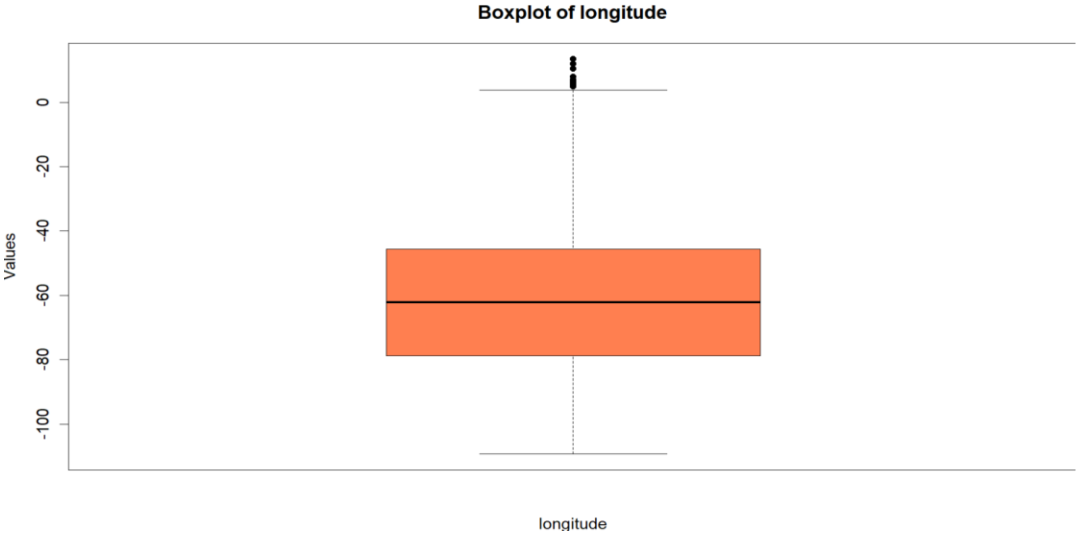
Data Visualization

Chart 1: Distribution of Latitude Variable



Observation: The histogram of latitude is displayed with the frequency on the y-axis and the latitude on the x-axis, which ranges from 10 to 70 and 0 to 1500 respectively. This histogram shows a high frequency between latitudes 10 and 40, a slow start that leads to a decline after latitude 40, and a minimum frequency after latitude 60.

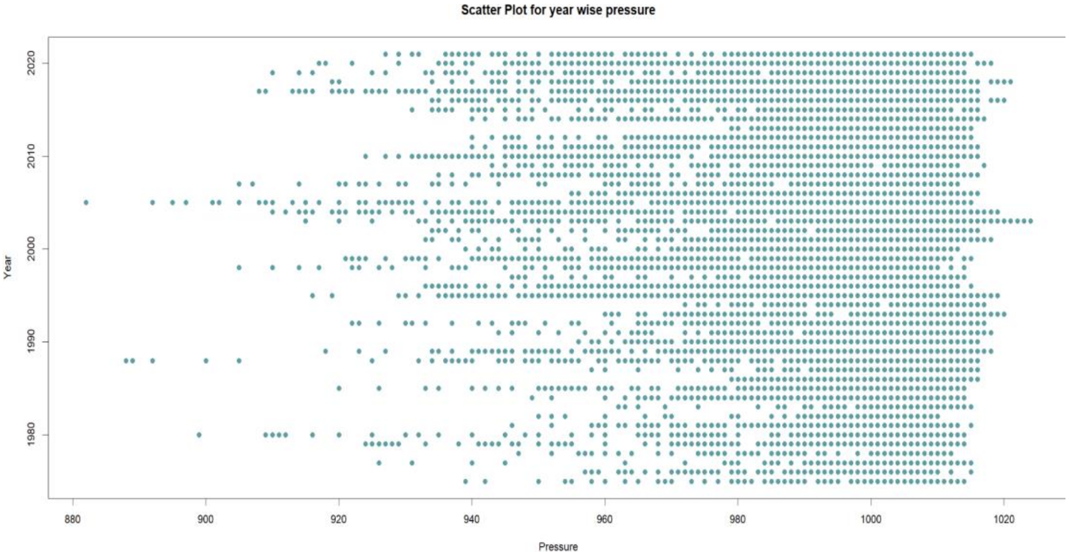
Chart 2 : Longitude Distribution Visualization



Observation: The boxplot focuses on analyzing the variable "longitude." It shows the spread and central tendency of the data using a box, whiskers, and median line. The box represents the middle 50% of the data, while the whiskers extend to show the range beyond that. Outliers, marked as individual points, represent potential extreme values. The plot's styling, including colors and line widths, enhances its appearance. By examining the boxplot, we can understand the distribution of longitude values and identify any unusual observations.

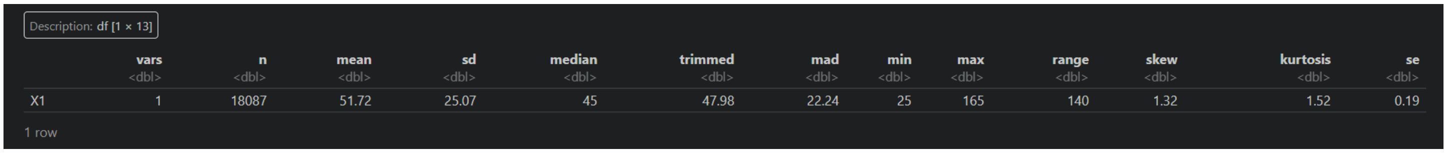
In conclusion, the boxplot offers information on the distribution of longitude values. It depicts the center 50% of the data, as well as outliers and the median. We may acquire a better grasp of the distribution features and probable extremes in the longitude values by reviewing the plot's parts and assessing its layout.

Chart 3: Pressure Distribution by storm year in a scatter plot.



Observation: The scatter plot shows how pressure and year variables relate to one another. Every point reflects a pressure measurement and the year it was recorded. The pressure variable's values are depicted on the x axis, which runs from 880 to 1020. The year variable's values are displayed on a y-axis that runs from 1980 to 2020. We can find patterns or trends in the plot that point to a relationship between pressure and year. The scatter plot's name, "Scatter Plot for Year-Wise Pressure," highlights its emphasis on examining the connection between these two variables[2].

Chart 4: Descriptive statistics of Subset Wind Variable



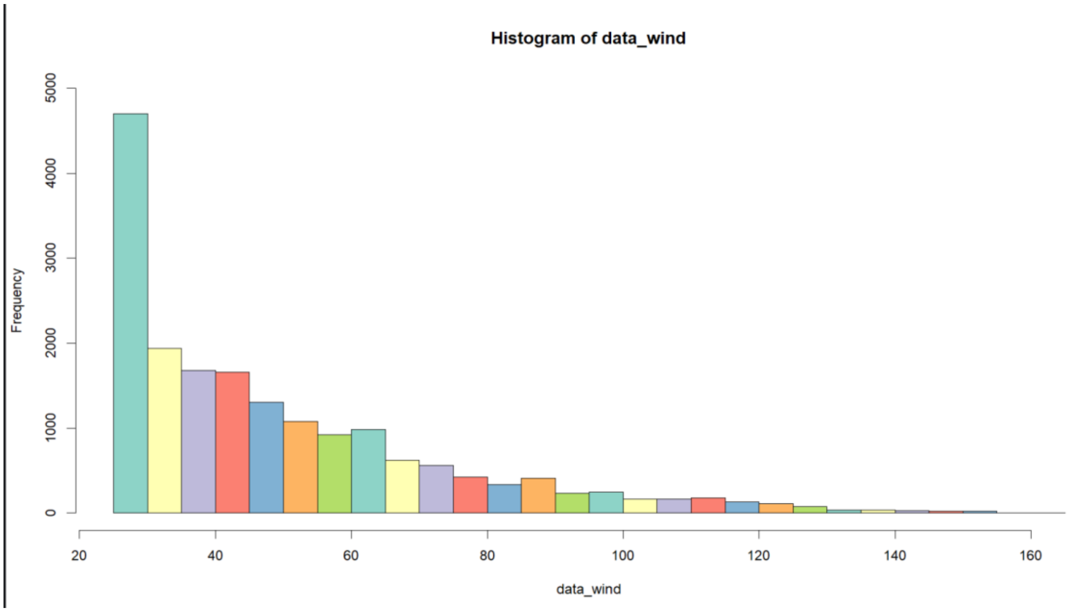
Description: df [1 × 13]

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
X1	1	18087	51.72	25.07	45	47.98	22.24	25	165	140	1.32	1.52	0.19

1 row

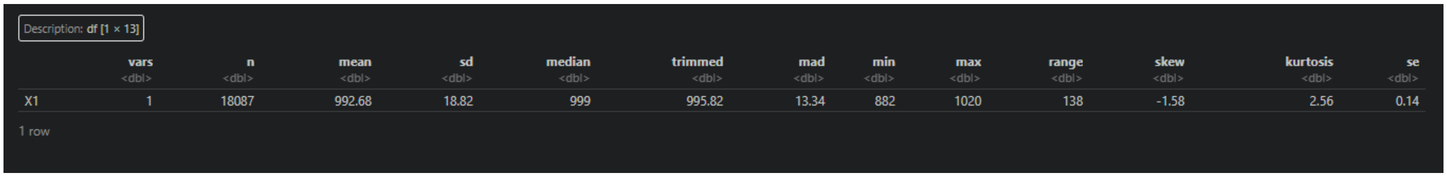
Observation: Here, we calculated the descriptive statistics for the Wind Variable. The mean and median values for the pressure variable are 51.72 and 45, respectively. When the mean value exceeds the median value, the data distribution is usually skewed to the right or positively skewed. This indicates that the mean is being pulled up by a few outlier high values, causing the mean to be higher than the median. Additionally, you can see that the skew value is positive as a result of this. The range equals 140 when the minimum value and maximum value are subtracted from each other[3].

Chart 5: Histogram of Wind Speed(Subset)



Observation: The frequency distribution of the data wind variable is displayed in the histogram of data visualization. The wind values, which range from 20 to 160, are represented on the x axis. The highest frequency, at around 4800, is seen when the wind speed is close to 20. The distribution is positively skewed.

Chart 6: Descriptive statistics of Subset Pressure Variable.



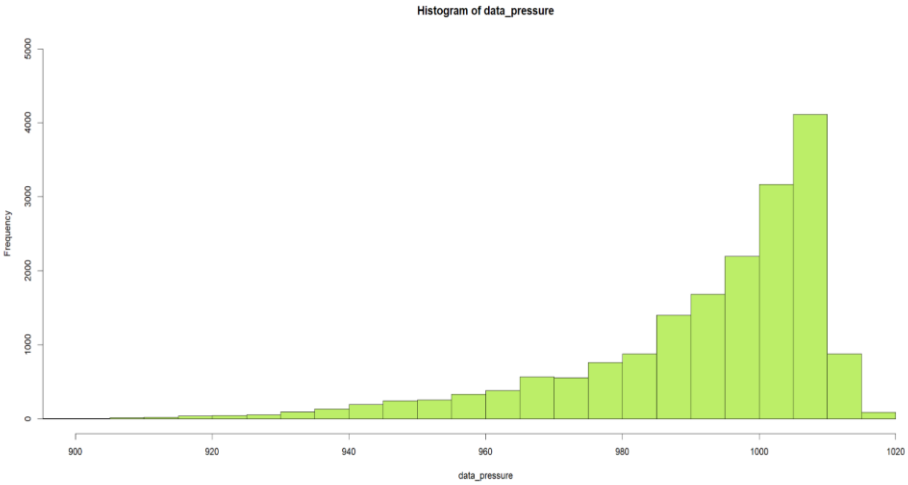
Description: df [1 × 13]

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
X1	1	18087	992.68	18.82	999	995.82	13.34	882	1020	138	-1.58	2.56	0.14

1 row

Observation: Here, we determined the Pressure Variable's descriptive statistics. The pressure variable has a mean value of 992.68 and a median value of 999. Since the median value is higher than the mean value, the distribution is said to be negatively skewed. This indicates that the mean is being negatively impacted by a small number of exceptionally low values, causing the mean to be lower than the median. You can also observe that the skew value is negative because of this. By subtracting the least value and the greatest value, the range is equal to 138[3].

Chart 7: Histogram Pressure Subset



Observation: The histogram provided displays the distribution of air pressure values at the center of the storm. The x-axis represents the air pressure values, specifically those where the corresponding wind speed is greater than or equal to 25. The y-axis represents the frequency or the number of occurrences of each air pressure value in the dataset.

From the histogram, we observe that the distribution is negatively skewed or left-skewed. This indicates that the majority of air pressure values are concentrated towards the higher end, while fewer occurrences of lower pressure values are observed. The histogram suggests that the air pressure at the center of the storm is most frequently found between 960 millibars and 1020 millibars.

The concentration of air pressure values in the higher range suggests that the storms in the dataset tend to have relatively strong air pressure.

CONCLUSION

The study examines the NOAA Atlantic Hurricane dataset, which offers useful information regarding Atlantic storms from 1975 through 2021. We acquired insights into the dataset by cleaning the data, computing descriptive statistics, and making visualizations.

Based on our findings, the dataset is a great resource for researching past hurricane trends, tracking storm intensities, and comprehending hurricane behavior. We learned about the average values, variability, and distribution of essential variables like wind speed and air pressure thanks to descriptive statistics.

We were able to display the distribution of data and investigate potential links between them using visualizations such as histograms and scatter plots. Hurricanes are most prevalent between latitudes 10 and 40, and the distribution of air pressure values is negatively biased, with higher values being more often.

Moving further, we may investigate the link between hurricane strength and other factors, as well as long-term trends in hurricanes and the geographical distribution of storms depending on latitude and longitude. Answering these questions will allow us to gain a better understanding of Atlantic storm behavior and its consequences on certain places.

We can increase our readiness and establish effective ways for coping with natural catastrophes by learning more about storms and their features. This analyses' findings can help with improved decision-making and mitigation actions in the event of storms.

REFERENCES

1. Utkarsh Singh. (2022). NOAA Atlantic Hurricane dataset. Retrieved from kaggle.com:
https://www.kaggle.com/datasets/utkarshx27/noaa-atlantic-hurricane-database

2. Dee Chiluiza, PhD. (2022, March). Scatter Plots. Retrieved from Scatter Plots:
https://rpubs.com/Dee_Chiluiza/scatterplot

3. Naveen. (2022, June 22). R subset() Function.Retrieved from https://sparkbyexamples.com/r-programming/r-subset-function-usage/