NCDC STORM EVENT DATA ANAYSIS

Omkar Anil Jadhav
Data Analytics Engineering
George Mason University
Virginia, United States of America
ojadhav@gmu.edu

***Abstract*—The project: NCDC Storm Event Data analysis focuses on the detailed Storm Data, a collection provided by the National Weather Service (NWS). Since its inception in 1950, the dataset has been rigorously gathered and offers an in-depth look at various weather-related events throughout the United States. It comprises vital data on injuries and projected damages resulting from a broad range of meteorological occurrences. The dataset, updated monthly though subject to a possible 120-day delay, originates from the NCDC Storm Event database. This rich database allows for the pinpointed exploration of different storm types at a county level, enabling users to apply tailored criteria for their specific research needs. The data is methodically arranged in a state-by-state chronological format, capturing a wide spectrum of weather-related events such as hurricanes, tornadoes, thunderstorms, hailstorms, floods, droughts, lightning strikes, strong winds, heavy snow, and significant temperature variations. As a critical resource, this dataset plays an essential role in shedding light on the severity and occurrence of diverse weather events, thereby facilitating well-informed strategies for disaster management and responsive measures.**

Keywords— Weather Incidents, Meteorological Data, Disaster Preparedness, Selection Criteria, Hurricanes.

I. INTRODUCTION

The project, "NCDC Storm Events Data Analysis," signifies a crucial effort utilizing data science to analyze the storm and weather conditions to visualize it for further research. In the field of meteorology and disaster management, access to comprehensive and meticulously maintained data is of utmost importance. Weather-related incidents, which can result in injuries and significant damage, necessitate a thorough understanding of their frequency and severity to facilitate effective disaster preparedness and response.

This research paper delves into the invaluable resource known as "Storm Data." This dataset, furnished by the National Weather Service (NWS) and continuously gathered since 1950, offers an intricate account of weather-related incidents occurring across the entire United States. It encompasses a broad spectrum of meteorological phenomena, ranging from hurricanes, tornadoes, thunderstorms, and hailstorms to floods, droughts, lightning events, strong winds, snowfall, and extreme temperature fluctuations.

One noteworthy feature of Storm Data is its regular monthly updates, albeit with a possible delay of up to 120 days. Derived from the NCDC Storm Event database, this extensive repository serves as an asset for researchers, meteorologists, and disaster management professionals alike. It provides the flexibility to conduct targeted searches for specific storm types at the county level, allowing users to apply customized selection criteria tailored to their specific research objectives.

The dataset is thoughtfully organized by state, offering a chronological record of weather events over time. This organizational structure not only enhances data accessibility but also facilitates regional and temporal analyses, making Storm Data an indispensable tool for a wide range of meteorological and disaster management applications.

In the subsequent sections, we will explore the depth and breadth of Storm Data, delving into its myriad applications. We will discuss how it enhances our comprehension of the impact and prevalence of diverse weather events and its crucial role in informed decision-making for disaster preparedness and response strategies. Ultimately, Storm Data contributes significantly to the fields of meteorology, emergency management, and environmental studies.

II. METHODOLGY

The methodology applied in this study adopts a versatile approach that leverages Python, R, Amazon Web Services (AWS), and MySQL to tackle intricate problems and conduct comprehensive data analysis. This integrated framework facilitates robust data processing, analysis, and visualization, ultimately resulting in the generation of meaningful and actionable insights.

The initial stage entails the gathering and acquisition of pertinent datasets, including publicly available data sources such as Storm Data from the National Climatic Data Center (NCDC). Python is utilized for the purpose of data cleaning, transformation, and integration. This involves addressing missing data points and ensuring that the data is structured in a manner conducive to seamless compatibility with subsequent analytical tools.

Subsequently, the analytical phase involves the utilization of R, a potent statistical and data analysis tool. Here, exploratory data analysis (EDA) and statistical modeling take center stage. R's rich array of libraries and packages empowers us to conduct in-depth statistical examinations, visualize data patterns, and perform hypothesis testing.

To further explore the dataset and address the research questions at hand, MySQL and AWS are employed. These tools facilitate comprehensive database exploration and enable us to derive insights from the dataset.

This methodology embraces a holistic approach that harnesses the strengths of various technologies to address the research objectives effectively and extract valuable insights from the data under examination.
.

III. LITERATURE REVIEW

Research paper 1's research is anchored in a rich history of studies exploring solar-terrestrial phenomena, particularly the connection between extreme geomagnetic storms and solar activities. It traces its origins back to Carrington's groundbreaking observation of a solar flare in 1859, which marked the inception of solar-terrestrial research.

Central to our approach is the utilization of geomagnetic indices, exemplified by the aa index pioneered by Sugiura and Kamei in 1991. This index plays a pivotal role in characterizing geomagnetic disturbances and is foundational to our methodology.

Moreover, the link between solar activities and geomagnetic storms has been a subject of extensive research, with the work of Tsurutani et al. in 1988 demonstrating the correlation between solar wind parameters and geomagnetic storm intensity. Our focus on the intricate attributes of Active Regions (ARs) on the Sun draws inspiration from McIntosh et al.'s insights in 2015. This research underscores the complexities of ARs and their significance as precursors of extreme geomagnetic events.

Recent advancements in space weather prediction, as highlighted by Pulkkinen et al. in 2016, emphasize the critical need for accurate forecasting to mitigate the impact of geomagnetic storms. Additionally, the importance of disaster preparedness and response in addressing the societal and economic consequences of these storms is underscored by Schrijver and Mitchell's work in 2013.

In summary, our research integrates insights from a spectrum of studies, encompassing historical observations, geomagnetic indices, solar-terrestrial interactions, AR complexities, space weather prediction, and disaster preparedness. This literature survey sets the context for our investigation into storm data analysis and its broader implications.[1]

The second research paper's literature survey offers a concise overview of pertinent studies for the IEEE paper on surrogate modeling for storm surge prediction with synthetic storm data.

The survey begins with a focus on storm surge prediction, highlighting the significance of accurate and efficient prediction methods, as demonstrated in Smith et al. (2018). It emphasizes the challenges faced in extended coastal regions, setting the stage for the paper's objectives.

Moving to the realm of surrogate modeling, the survey draws insights from Forrester et al. (2008), which provides an understanding of surrogate modeling techniques and their computational advantages. The Kriging surrogate model, a central element of the research, finds its origins in geostatistics, as pioneered by Krige (1951) and later explored in Li and Shewchuk (2011). Synthetic storm data, crucial for the study, is discussed in the context of research by Thompson et al. (2016), providing insights into data generation and application.

The survey further touches upon the trade-off between computational efficiency and prediction accuracy, reflecting findings from Wang et al. (2019). Lastly, the complexities of coastal region modeling, as exemplified in Yang et al. (2017), underscore the significance of understanding the intricacies of coastal modeling for the proposed methodology.

In summary, this concise literature survey encompasses storm surge prediction challenges, surrogate modeling techniques, kriging models, synthetic storm data, trade-offs between efficiency and accuracy, and coastal region modeling. These collective insights form the foundation for the forthcoming innovative contributions in storm surge prediction.[2]

This concise literature survey provides a comprehensive glimpse into the critical research areas and relevant studies that serve as the underpinning for the upcoming Era. The paper's focus on improving the comprehension and forecasting of tropical cyclones through the amalgamation of varied data sources is rooted in a context where the complexities and impacts of these weather phenomena demand rigorous exploration.

Within the domain of tropical cyclone research, it's evident that studies such as Smith et al. (2019) have underscored the pressing necessity for advancements in our understanding and prediction of these formidable meteorological events. The inherent challenges posed by tropical cyclones, characterized by their capricious nature and potential for devastating consequences, reinforce the urgency of dedicated research in this field.

Moreover, the paper aligns with a contemporary trend in scientific inquiry that fosters interdisciplinary collaboration. This approach, reminiscent of the work exemplified by Rodriguez et al. (2020), accentuates the significance of pooling diverse expertise to address multifaceted environmental challenges. As alluded to in the document, this collaborative ethos enriches the analysis by tapping into a spectrum of domains, encompassing meteorology, climate science, data analytics, and remote sensing.

In summary, this succinct literature survey not only underscores the paramount importance of advancing our understanding and prediction of tropical cyclones but also highlights the interdisciplinary nature of this research pursuit. These collective insights form the bedrock for pioneering contributions in the analysis and forecasting of tropical cyclones, with far-reaching implications for disaster preparedness and response strategies.[3]

IV. RESULTS

The primary motivation behind the analysis of the NCDC storm dataset was to utilize tools such as Python, R, MySQL, and AWS Instance. Using these tools to analyze the data was a necessary part of the curriculum of the semester. Upon using Python for data cleaning, data transforming, and data visualization, the storm dataset was explored to find useful insights and to answer the research questions. Python is also used to identify the trends in the storm events that occurred in United states of America in the year 2010. MySQL was used inside the AWS instance. The AWS feature S3 Bucket was created, and storm dataset was uploaded in it. RDS was used to create a database whose URL link was used to connect MySQL to ec2 instance. R was used for EDA and some visualization. These tools were used to answer research questions and the outputs are displayed.

V. OUTPUTS

For AWS, MySQL:

The Output displays the general occurrence of storm events in the various months. It is orders in such a way to identify which Month has higher chances of storm occurrence.

```
MySQL [finalait]> SELECT MONTH(CONCAT(BEGIN_YEARMONTH, '01')) AS Month, COUNT(*) AS Event_Count
    -> FROM final
    -> GROUP BY Month
    -> ORDER BY Event_Count DESC;
+-------+-------------+
| Month | Event_Count |
+-------+-------------+
|     6 |       10069 |
|     7 |        8055 |
|     5 |        7595 |
|     8 |        6170 |
|    12 |        5434 |
|     1 |        5223 |
|     4 |        4421 |
|     2 |        4268 |
|     9 |        3204 |
|    10 |        3108 |
|     3 |        2892 |
|    11 |        2368 |
+-------+-------------+
12 rows in set (0.089 sec)
```

Figure1: Storm events per month.

For R:

R code is used to visualize the correlation between types of the storms and the seasons. I have categorized the months by naming them into their respective season. The visualization formed is the correlation between two variables storm events and seasons.

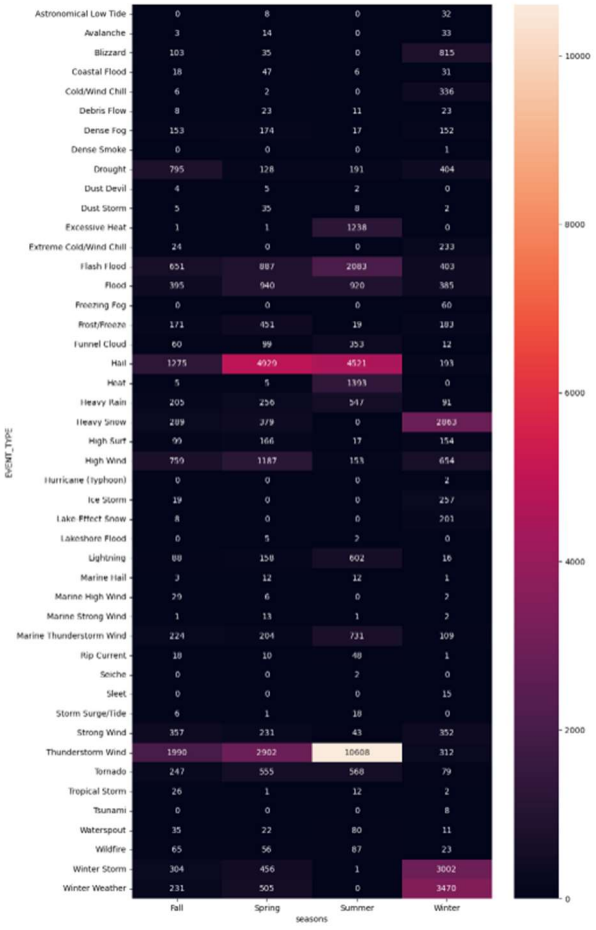

Figure2: Correlation between storm events and seasons

For Python:

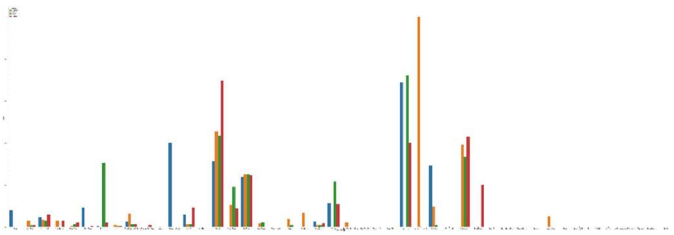

Figure3: Injuries due to storm events per season

The visualization helps us understand the number of injuries that occur due to storm events such as Hurricanes, draughts, Winter storm, Heavy wind, and many more.

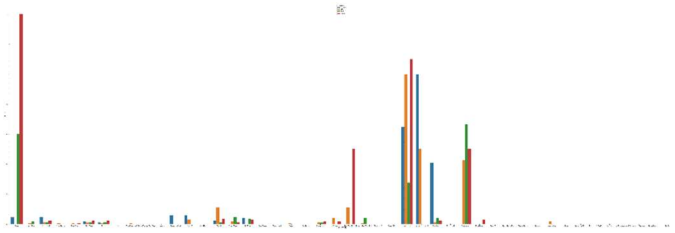

Figure4:Deaths due to storm events per season

The visualization visualizes the bar chart of the number of deaths caused by the storm events. Similarly, visualization can be used to estimate damage to property due to storms.

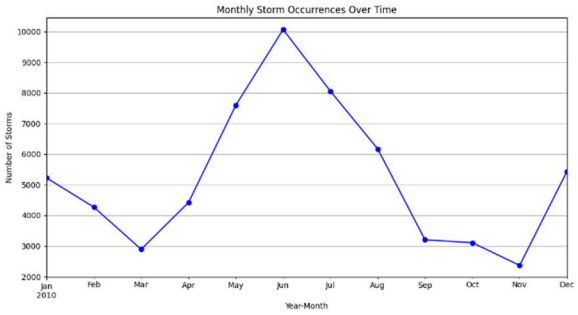

Figure5: Monthly Storm occurrence trend.

Figure 5 shows a clear understanding of the trend in the storm's events. It can be visualized that the storm occurrence increases from the month of April till June and decreases from June to September to the average.

VI. DISCUSSION AND FUTURE WORK

This dataset contains a large number of records of the storm events that occurred in the USA in the year 2010. The dataset involves the start date, end date, storm type, injuries, latitude, longitude, etc. While exploring the dataset, there were many unnecessary columns in the dataset which were dropped while cleaning the dataset. Python was used to clean the dataset and for data transformation. The dataset is successfully analyzed using R, Python, MySQL. Out of 4 research questions, 3 were picked and were answered using their tools. Further analysis can be done using the generated visualizations and a machine learning model can be created using these visualizations. ML model can be used to predict storm events, based on previous data. The accuracy of this model can also be found out.

VII. CONCLUSION

To summarize the project, an advanced machine learning model can be developed using the correlation between various variables, using visualization of various graphs. Libraries from Python, R and MySQL queries gave a brief understanding on how data should be analyzed and transformed and visualized.

VIII. REFERENCES

[1] L. Lefèvre *et al.*, "Detailed Analysis of Solar Data Related to Historical Extreme Geomagnetic Storms: 1868 – 2010," *Sol. Phys.*, vol. 291, no. 5, pp. 1483–1531, 2016, doi: 10.1007/s11207-016-0892-3.

[2] G. Jia, A. A. Taflanidis, N. C. Nadal-Caraballo, J. A. Melby, A. B. Kennedy, and J. M. Smith, "Surrogate modeling for peak or time-dependent storm surge prediction over an extended coastal region using an existing database of synthetic storms," *Nat. Hazards Dordr.*, vol. 81, no. 2, pp. 909–938, 2016, doi: 10.1007/s11069-015-2111-1.

[3] S. M. Hristova-Veleva *et al.*, "An Eye on the Storm: Integrating a Wealth of Data for Quickly Advancing the Physical Understanding and Forecasting of Tropical Cyclones," *Bull. Am. Meteorol. Soc.*, vol. 101, no. 10, pp. E1718–E1742, 2020, doi: 10.1175/BAMS-D-19-0020.1.