Weather-Driven Traffic Crash Analysis for Enhanced Traffic Management

Yuxiang Huang, Xiao Tan, Ruochen Xiao, Ying Liu

New York University, yh5047, xt2256, rx2238, yl12065@nyu.edu

***Abstract* - This project investigates the relationship between traffic crashes and weather conditions to provide actionable insights for traffic management. Using a 5GB dataset of U.S. traffic accident and weather records from 2016 to 2024, we conducted data cleaning using Hadoop MapReduce and integrated the datasets by matching crash dates with corresponding weather conditions. Key features retained include accident date, temperature, humidity, visibility, and rainfall. Through Hive-based analytics, we explored how these weather variables influence crash frequencies, revealing critical patterns that could inform traffic management and safety policies.**

***Keywords – traffic crash, weather condition, big data processing, MapReduce, Hive.***

**Ⅰ. Introduction**

On a foggy winter morning in Chicago, a multi-car crash brought traffic to a standstill on the city’s busiest expressway. Emergency responders struggled with low visibility and icy roads. This incident highlights how weather conditions like rain, snow, and fog critically impact road safety beyond driver error.

To better understand this relationship, we gathered 5GB of traffic accident and weather data from major U.S. cities from 2016 to 2024. Using Hadoop MapReduce, we joined crash records with weather data by matching accident dates, ensuring compatibility by cleaning and formatting datasets. We retained five key features: accident date, temperature, humidity, visibility, and rainfall conditions.

With the cleaned dataset, we conducted extensive analyses using Hive. We investigated the impact of various weather factors on traffic crash frequencies, finding patterns linked to specific conditions. Our findings could guide traffic management teams in optimizing police deployment, particularly when weather forecasts suggest increased crash risks. This study offers data-driven recommendations to improve traffic safety policies.

**II. Data Cleaning**

In our project, we needed to extract five key features: accident date, temperature, humidity, visibility, and rainfall. Since each of our four team members worked on different raw datasets, we had to clean and format the data to combine it into one consistent dataset for analysis.

As for the data collection part, we obtained the datasets from several sources. Firstly, we collected all the weather data from visualcrossing.com [1]. The raw weather data contains 34 features. Secondly, as for the crash data, we collected 400MB of traffic crash data from Chicago through data.cityofchicago.org [2], 3GB of nationwide traffic crash data from kaggle.com [3], and additional crash data from New York via catalog.data.gov [4], among other sources. In total, all the traffic crash datasets amount to approximately 5GB in size and cover the period from 2016 to 2024.

Regardless of the data source, our cleaning process followed similar steps. For traffic crash data, we needed to extract the accident date, while for weather data, key features like date, temperature, and rainfall were required. After that, we joined them together according to the accident date. To demonstrate this process, we use Chicago’s data cleaning as an example, showing how we applied the Join Pattern to develop a MapReduce program for merging traffic and weather datasets.

1. *Mapper for Weather Data*

The CHIJoinWeatherMapper class processes weather data from Chicago, indicated by the prefix CHI in the class name. During the map phase, each line of the weather dataset is parsed, skipping the header if detected. The mapper extracts the following key features:

* **Date**: Used as the join key.
* **Temperature**: The recorded temperature on the date.
* **Humidity**: The air's moisture content.
* **Visibility**: The distance drivers could see on the road.
* **Rain or Not**: A binary value ("1" if rainy, "0" otherwise).

If any of these fields are missing, the record is discarded to ensure data integrity. The extracted values are formatted as a comma-separated string. The mapper outputs the crash date as the key and the weather data prefixed with "W" to indicate that the record comes from the weather dataset. This prefix helps distinguish weather records from crash records during the reduce phase.

1. *Mapper for Crash Data*

The CHIJoinCrashMapper class processes traffic crash data from Chicago, indicated by the CHI prefix in its name. During the map phase, each line of the crash dataset is parsed, skipping the header line if detected. The mapper performs the following steps:

* **Extracting Crash Date**: The crash date is extracted from the third column of the input line.
* **Formatting the Date**: Since the original date format is MM/dd/yyyy, it is converted to the standard format yyyy-MM-dd using Java’s DateTimeFormatter class. This ensures compatibility with the weather data format.

The mapper then outputs the crash date as the key and the string "Crash" as the value. The "C" prefix helps distinguish crash records from weather records during the reduce phase.

1. *Reducer*

The CHIJoinReducer merges Chicago’s traffic crash and weather data based on the crash date. It starts by extracting the crash date from the key and initializing two variables: crashCount to track the number of crashes and weatherLine to store the relevant weather data.

During the reduce phase, it processes each value. If a record starts with 'C', it represents a crash, so crashCount is incremented. If a record starts with 'W', it contains weather data, and the prefix 'W' is removed to extract weather details.

If corresponding weather data exists, the reducer formats the output as crashDate, temperature, humidity, visibility, rain or not. It writes this combined information for each recorded crash, ensuring only matched records are output for analysis.

1. *Driver*

The CHIJoin class sets up and runs the MapReduce job for joining Chicago’s crash and weather data. It takes three command-line arguments: crash data path, weather data path, and output path. The job uses MultipleInputs to read both datasets, assigning CHIJoinCrashMapper for crash data and CHIJoinWeatherMapper for weather data. The reducer is set to CHIJoinReducer, and the output format is NullWritable as the key and Text as the value.

So far, the data cleaning and joining process has completed, the final dataset contains five key features: Date, Temperature, Humidity, Visibility, and RainOrNot. These features form the basis for our analysis.

In the next section, we will create a Hive table using the cleaned and integrated dataset. All subsequent analyses will be conducted using Hive to explore the relationship between weather conditions and traffic crash occurrences.

**III. Data Analytics**

Data analytics were completed in five parts.

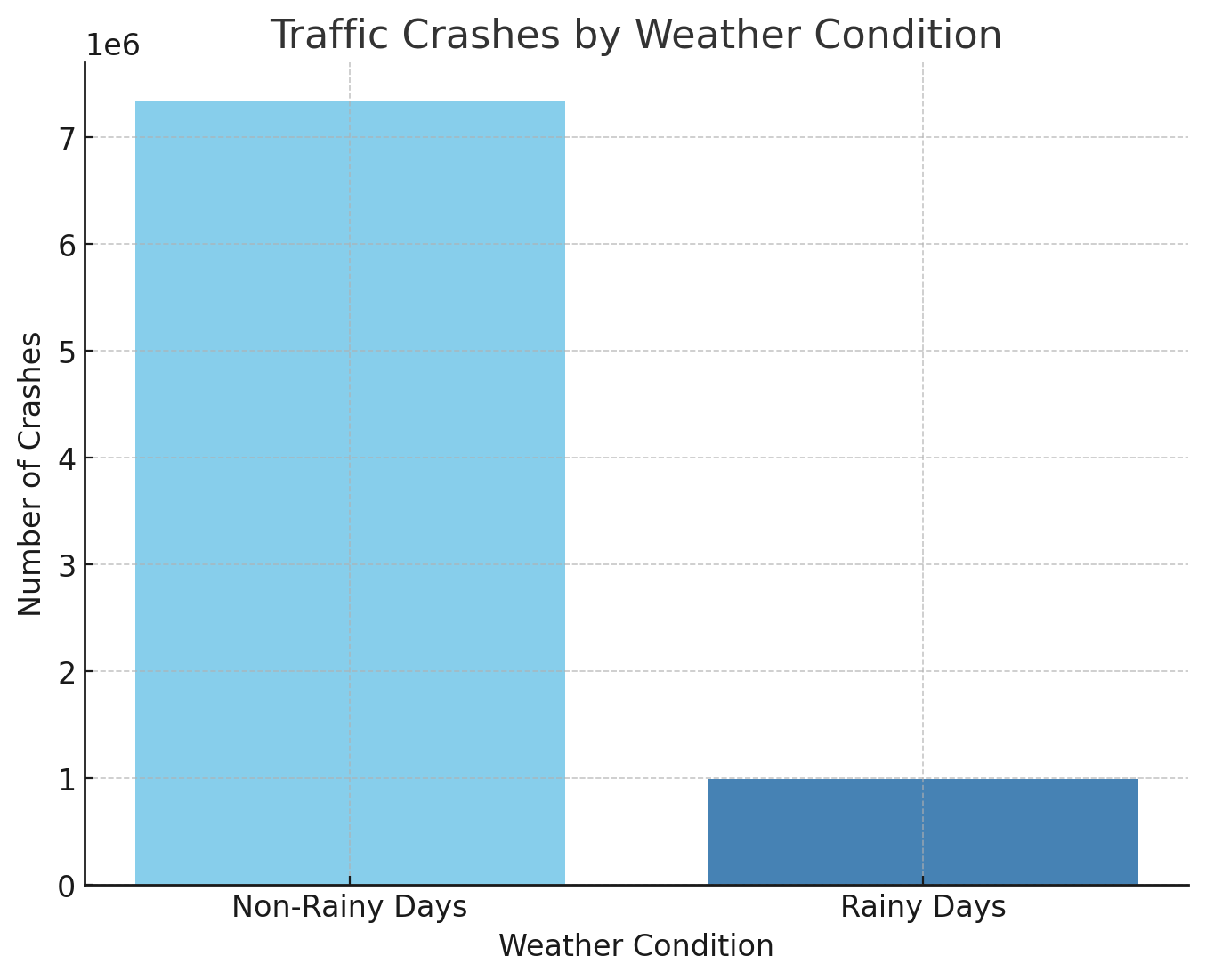
* Impact of Rainfall
* Impact of Temperature
* Impact of Humidity
* Impact of Visibility
* Impact of Season

We put all the cleaned data from the previous phase into the same folder, and used Hive to create a table based on the data. Notice that the table has the same five feature columns with those cleaned datasets.

1. Impact of Rainfall

To analyze the impact of rainfall on traffic crashes, we queried the cleaned dataset using Hive.

Surprisingly, the number of traffic crashes on rainy days is much lower than on clear days. According to the query result, 996,860 crashes occurred on rainy days (rain\_or\_not = 1), while 7,332,772 crashes happened on non-rainy days (rain\_or\_not = 0).

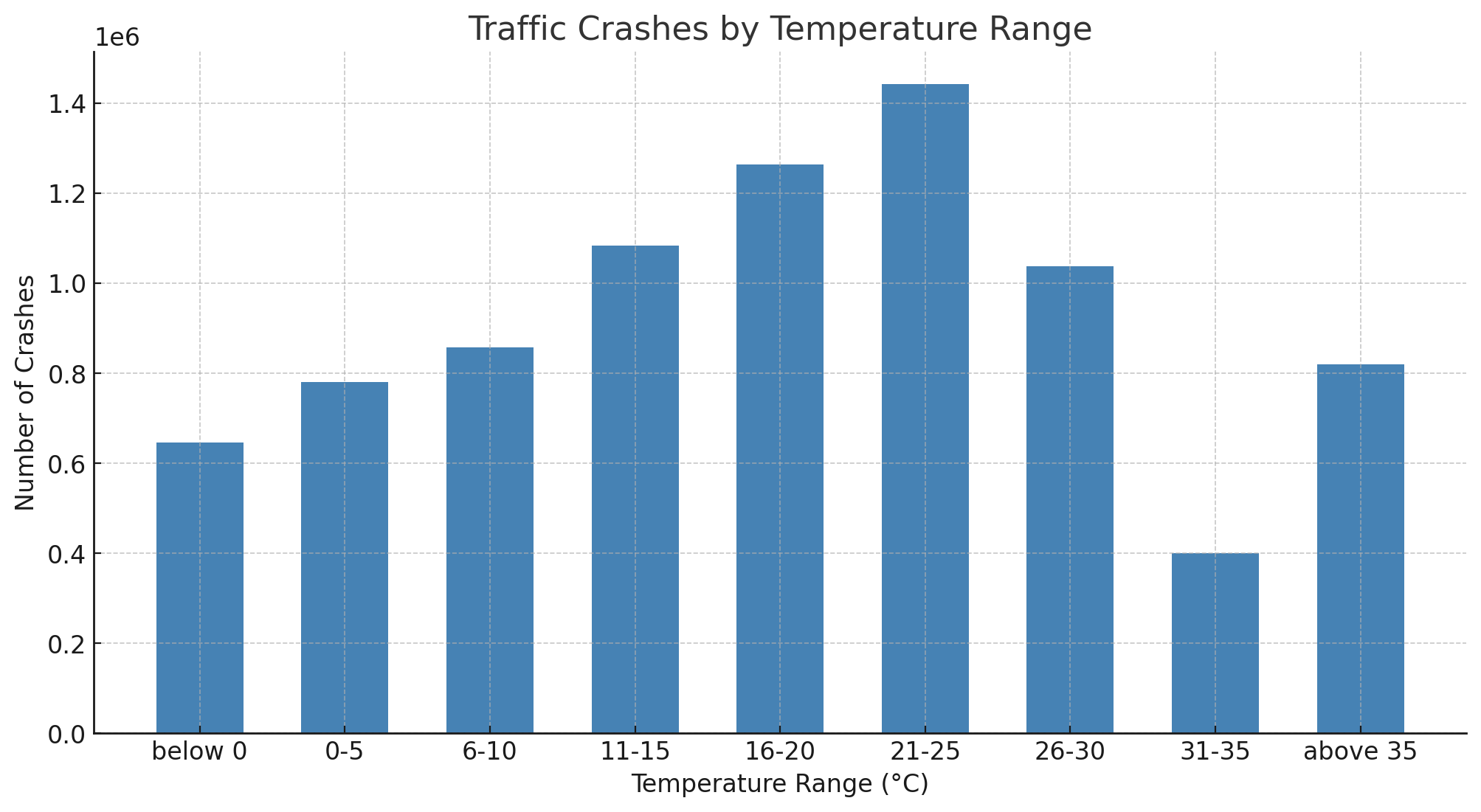


However, to draw a more accurate conclusion, we may need additional data on the rainfall frequency across different regions in the U.S. This would allow us to compare the crash rate relative to the actual occurrence of rainy weather, providing a clearer understanding of rainfall’s true impact on traffic crashes.

1. Impact of Temperature

To examine how temperature affects traffic crashes, we -grouped accidents by temperature ranges using Hive. The query categorized temperatures into eight ranges, from "below 0°C" to "above 35°C," and counted crashes in each range.

The result shows that the highest number of crashes occurred in the 21-25°C range (1,442,198 crashes), followed by the 16-20°C range (1,263,967 crashes). Interestingly, even the "above 35°C" range had a significant number of crashes (820,431). On the lower end, the below 0°C range recorded the fewest crashes (645,257).

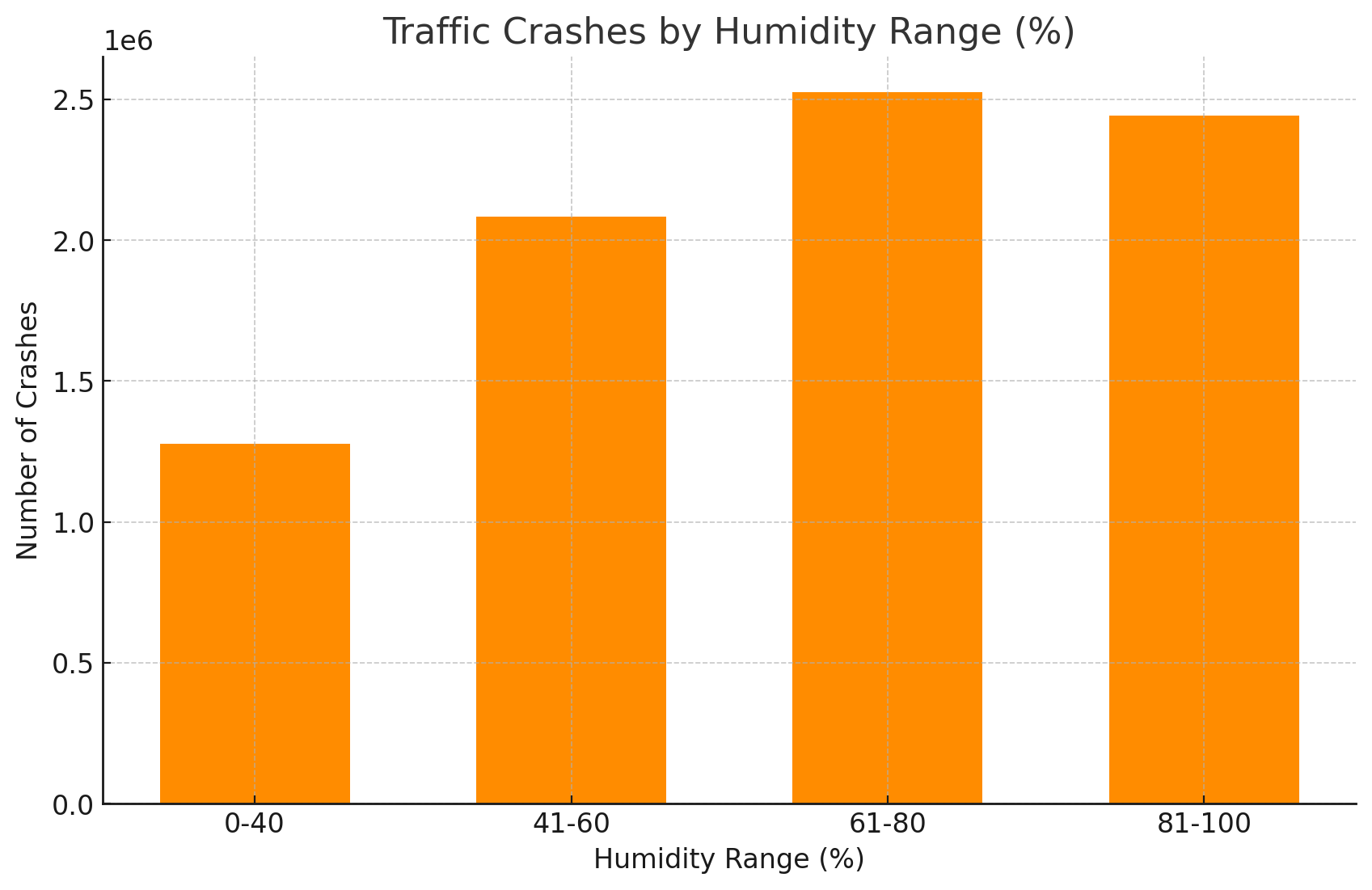


These findings suggest that moderate temperatures (around 16-25°C) may coincide with increased traffic activity, possibly due to favorable driving conditions that encourage more vehicles on the road. However, extreme temperatures, whether cold or hot, seem to be associated with fewer crashes, potentially due to reduced travel in such weather. Further analysis involving traffic volume data could provide deeper insight.

1. Impact of Humidity

To explore the impact of humidity on traffic crashes, we divided the humidity values into four ranges using Hive: 0-40%, 41-60%, 61-80%, and 81-100%. We then counted the number of crashes occurring in each range.

The query results revealed that the majority of crashes happened in the 61-80% humidity range (2,525,333 crashes), followed closely by the 81-100% range (2,443,413 crashes). The 41-60% range recorded 2,084,056 crashes, while the 0-40% range saw the fewest crashes (1,276,830).

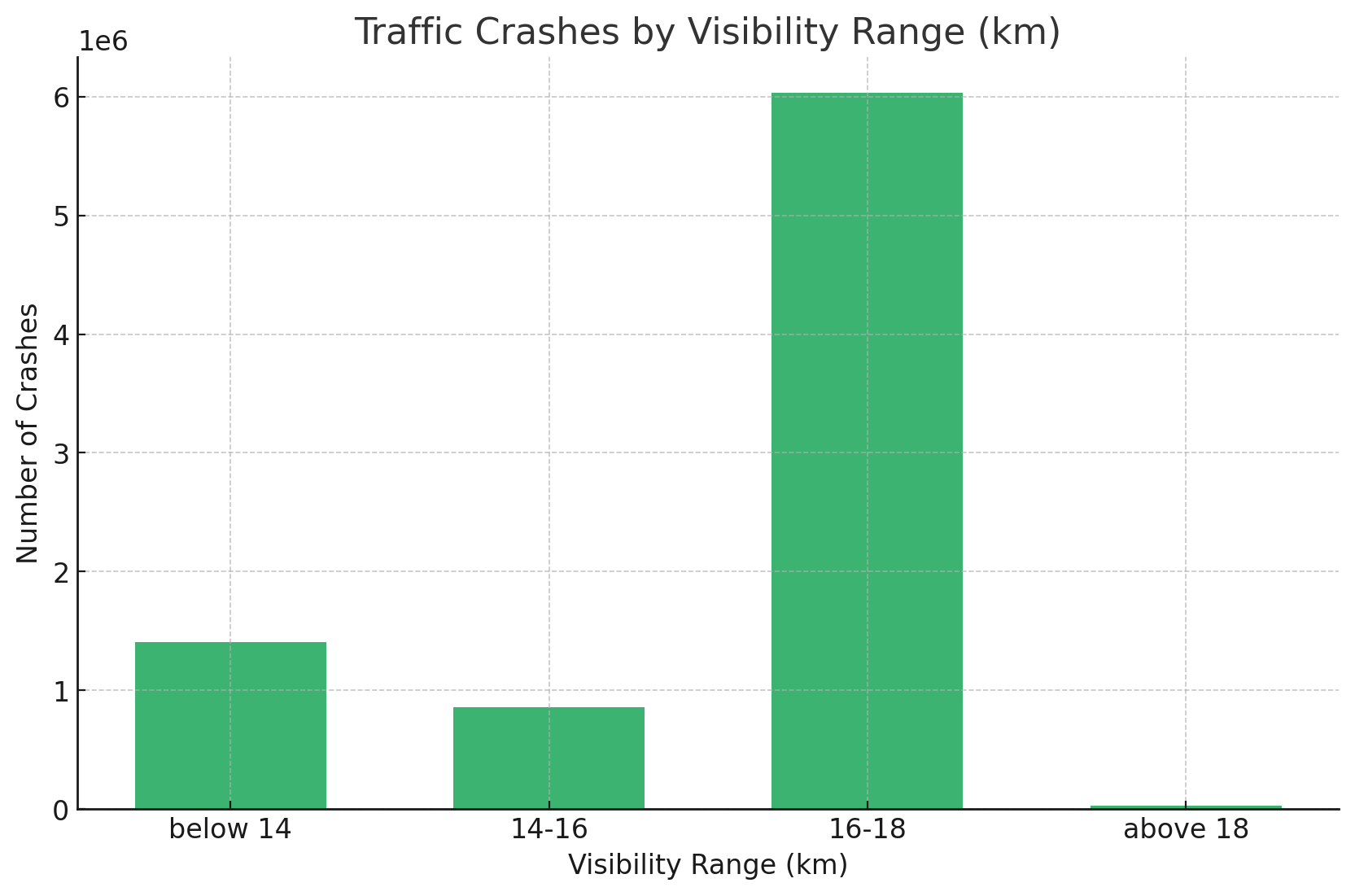


This trend suggests that higher humidity levels, especially between 61-100%, might be linked to more traffic crashes. This could be due to adverse weather conditions like fog, rain, or slippery roads typically associated with high humidity.

1. Impact of Visibility

To explore how visibility affects traffic crashes, we divided visibility into four ranges using Hive: below 14km, 14-16km, 16-18km, and above 18km. We then counted the number of crashes occurring in each range.

The results revealed that the majority of crashes happened in the 16-18km visibility range (6,031,937 crashes), followed by below 14km (1,406,454 crashes) and 14-16km (861,052 crashes). Interestingly, the above 18km range recorded only 30,189 crashes, making it the least common category.

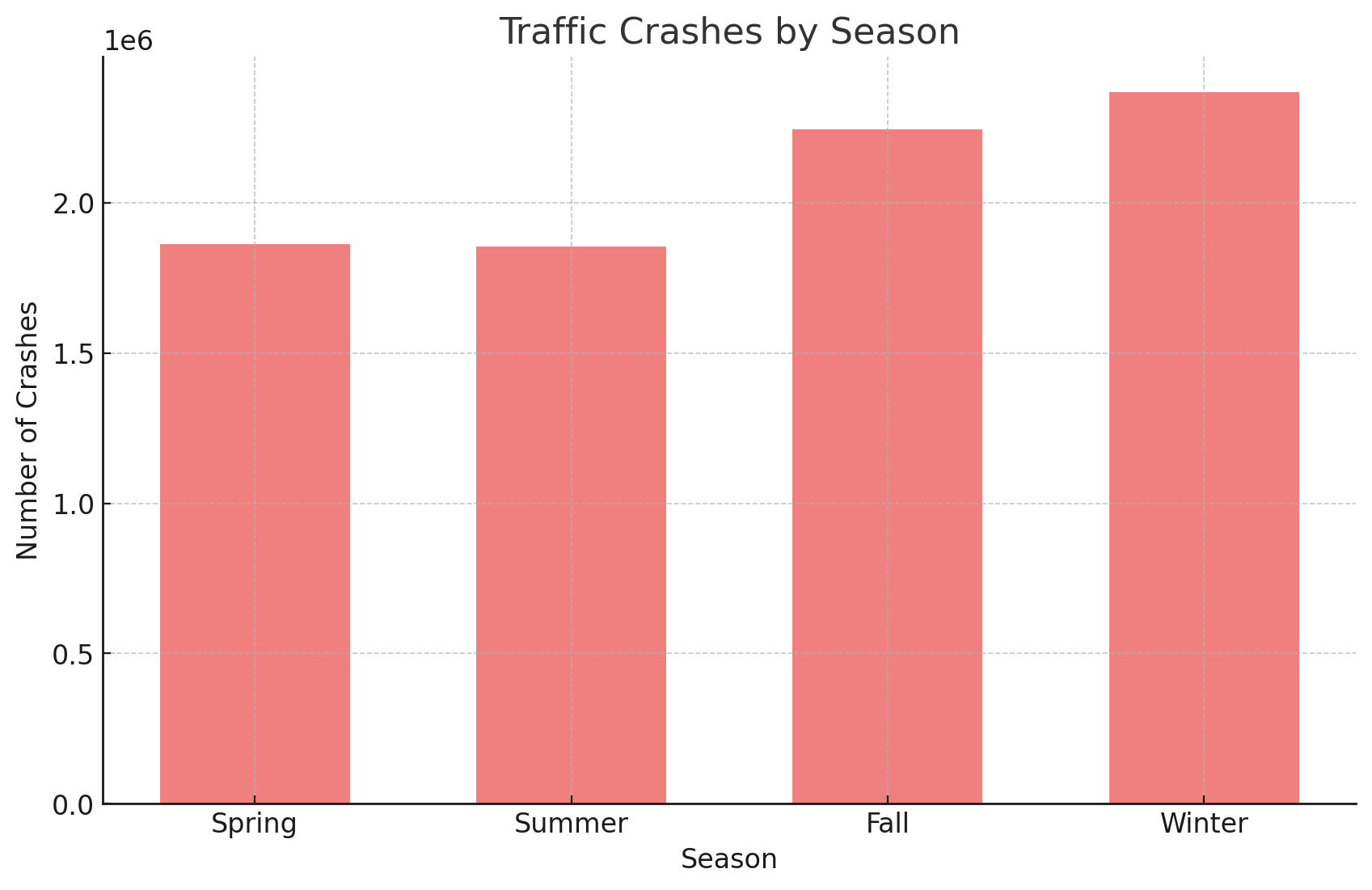


This trend suggests that moderate visibility conditions (around 16-18km) see the highest number of crashes, possibly due to higher traffic volumes during typical driving conditions. In contrast, extremely high visibility (above 18km) corresponds to significantly fewer crashes, possibly due to fewer driving occurrences in remote or less populated areas.

1. Impact of Season

To analyze the seasonal distribution of traffic crashes, we grouped crashes by season using the crash\_date column. Seasons were defined as follows: Winter (Dec-Feb), Spring (Mar-May), Summer (Jun-Aug), and Fall (Sep-Nov).

The results show that the highest number of crashes occurred in Winter, with 2,368,003 accidents, followed by Fall (2,244,210). Spring and Summer had comparatively fewer crashes, with 1,862,363 and 1,855,056 accidents, respectively.



These findings suggest that winter conditions, such as snow and ice, may contribute to increased traffic crashes. Additionally, the high number of crashes in fall might relate to changing weather patterns, such as rain or shorter daylight hours.

**IⅤ. Conclusion**

Our analysis revealed significant correlations between weather conditions and traffic crashes. Rainfall, temperature, humidity, visibility, and seasonal changes all demonstrated distinct impacts on crash occurrences. Notably, crashes were more frequent in moderate temperatures (16-25°C), high humidity levels (61-100%), and moderate visibility (16-18km). Winter saw the highest crash count, likely due to hazardous weather such as snow and ice.

Surprisingly, rainy days accounted for fewer crashes than clear days. This could be due to reduced driving activity during adverse weather, but it may also result from fewer rainy days compared to clear days in the dataset. Additionally, the high crash frequency during fall and winter suggests that traffic authorities should remain vigilant even in seasons without extreme weather.

These insights underscore the importance of weather-aware traffic management strategies. Future research could incorporate traffic volume data, detailed rainfall frequency, and specific road conditions for more comprehensive crash-risk prediction models.

**References**

[1] Weather Data, Visual Crossing Weather, 2016-2024. Accessed: Dec. 10, 2024. [Online]. https://www.visualcrossing.com

[2] Chicago Traffic Crash Data, City of Chicago Data Portal, 2016-2024. Accessed: Dec. 10, 2024. [Online]. https://data.cityofchicago.org

[3] U.S. Nationwide Traffic Accident Data, Kaggle, 2016-2024. Accessed: Dec. 10, 2024. [Online]. https://www.kaggle.com

[4] New York Traffic Crash Data, Data.gov, 2016-2024. Accessed: Dec. 10, 2024. [Online]. Available: https://catalog.data.gov