University of North Texas

Introduction to Big Data and Data Science – CSCE 5300

Group Number

Project Report

Title

Weather Enhanced Transportation Optimization Using Big Data

Team Members-

saTable of Contents

1 Abstract:

The study delves into how meteorological conditions impact transportation traffic, evaluating various analytical methods. Specifically, it juxtaposes the methodologies outlined in 'A Big Data Science Solution for Transportation Analytics with Meteorological Data' against a simulated approach. It affirms the effectiveness of the proposed data analytics method, noting its simplicity, interpretability, and adeptness in pinpointing crucial weather factors affecting traffic flow and crafting predictive models. The simulated results align closely with those presented in the paper. Moreover, it explores potential applications of this data analytics method in transportation planning, management, traffic advisories, public transportation scheduling, and optimizing ride-hailing services.

2 Introduction:

The primary objective of this report is to conduct a comprehensive comparison regarding how meteorological conditions impact traffic within the transportation system. The analysis involves an evaluation of varied analytical methods, particularly juxtaposing the methodologies illustrated in 'A Big Data Science Solution for Transportation Analytics with Meteorological Data' against a simulated approach. Central to this comparison is an in-depth exploration of the intricate relationship between meteorological conditions—highlighting elements like temperature, precipitation, and wind speed—and their influence on traffic flow. Recognizing and comprehending this correlation is pivotal in advancing both the planning and management aspects of transportation, ultimately culminating in heightened safety and enhanced operational efficiency within the transportation system. Furthermore, the report delves into potential applications of the data analytics method across diverse transportation-related domains, encompassing the prediction of traffic flow, the development of traffic advisories and warnings, the optimization of public transportation schedules, and the enhancement of efficacy within ride-hailing services.

3 Problem Statement:

The traffic flow denotes the motion of vehicles on roadways or interconnected systems. This intricate interaction is shaped by various elements, encompassing road layouts, traffic volume, regulatory measures, driver behavior, and the environmental conditions in the vicinity. Notably, meteorological conditions wield substantial influence over traffic dynamics, impacting driver and passenger visibility, vehicular traction, and overall comfort. They significantly affect the performance and reliability of vehicles and the underlying infrastructure. Moreover, these conditions can instigate various traffic incidents such as collisions, breakdowns, or road closures, triggering disruptions in the regular traffic flow, leading to congestion, delays, or the need for alternative routes. Hence, it becomes imperative to comprehend how meteorological conditions specifically shape traffic flow and to devise proficient methodologies and tools capable of foreseeing traffic patterns across diverse weather scenarios.

Existing Methods/Algorithms:

Numerous methodologies and algorithms have been developed to anticipate and forecast traffic flow, constituting a diverse landscape of predictive techniques. Some of the most prevalent methods include:

3.1 Time Series Analysis:

This approach involves forecasting future values within a timeline by scrutinizing historical data, thereby facilitating the prediction of traffic flow by modelling seasonal and cyclical patterns. Utilizing methods such as autoregressive integrated moving average (ARIMA), neural networks, time series analysis effectively captures the patterns and variations inherent in data related to the flow of traffic. While notably proficient in generating accurate and reliable short-term traffic flow predictions, this method might lack the capability to fully consider external factors, such as the influence of weather conditions on traffic dynamics.

3.2 Machine Learning:

These algorithms decode the intricate relationship between traffic flow and weather conditions by delving into historical data, allowing for predictions across a spectrum of different weather scenarios. Drawing upon techniques such as regression, classification, clustering, or deep learning, machine learning comprehends the distinctive features and patterns embedded within traffic flow and weather data, facilitating predictions based on acquired models. While remarkably flexible in providing adaptable long-term traffic flow forecasts, this approach often demands extensive, high-quality datasets and involves the use of complex and computationally intensive algorithms.

3.3 Expert Systems:

These computational programs harness the expertise of traffic engineers and transportation planners to forecast traffic flow patterns. Employing diverse techniques like rule-based reasoning, fuzzy logic, or genetic algorithms, expert systems translate and apply the knowledge and rules formulated by industry experts, generating predictions through inference and optimization processes. While adept at furnishing intuitive and comprehensible traffic flow predictions, these systems may, at times, rely on subjective or incomplete knowledge, offering solutions that are heuristic and approximate in nature.

4 Method/Algorithm Presented in the Paper:

The research outlined in 'A Big Data Science Solution for Transportation Analytics with Meteorological Data' introduces an intricate data analytics methodology that relies on correlation and regression analysis. This approach involves a series of carefully structured steps:

4.1 Data Collection:

Gathering traffic flow and weather data from diverse sources sensors, cameras, radars, satellites, or web services—culminating in an amalgamated data platform. Employing Apache Spark, a distributed computing framework, to meticulously process and scrutinize the extensive and varied datasets. Furthermore, the utilization of Apache Kafka, a distributed messaging system, is integral in both streaming and storing real-time data, ensuring a comprehensive dataset.

4.2 Data Analysis:

Examining the collected data on traffic flow and weather necessitates the application of correlation and regression analysis. Correlation analysis is used to assess the intensity and direction of the linear connection between two factors, such as traffic flow and temperature. Regression analysis seeks to construct a model representing the functional association between a dependent variable (such as traffic flow) and one or more independent variables (such as weather factors). This analytical process relies on statistical metrics like the Pearson correlation coefficient to evaluate linear correlation and the utilization of linear regression to establish a fitting linear equation for data analysis.

4.3 Data Visualization:

After the analysis, the paper uses diverse tools and techniques to visually represent the data's findings. Utilizing Tableau, a robust data visualization software, the researchers generate interactive and dynamic dashboards to vividly present correlation and regression outcomes. Additionally, leveraging Google Maps, a web mapping service, to overlay the traffic flow and weather data onto geographical maps, offering a visual, spatial dimension to the research's findings.

5 Strengths and Weaknesses:

The paper's approach holds distinct advantages owing to its simplicity, interpretive nature, and its capacity to pinpoint pivotal weather factors affecting traffic flow while crafting a predictive model. However, it's constrained by its capability to measure solely linear relationships and its reliance on vast datasets. Here's a concise summary of the strengths and limitations of the paper's methodology.

5.1 Strengths and Weakness:

The methodology offers simplicity and ease of implementation by utilizing common statistical methods like correlation and regression analysis for data interpretation. Additionally, it provides clear and intuitive results—such as correlation coefficients and regression coefficients depicting the significance and directional relationship between traffic flow and weather variables. Moreover, it excels in identifying influential weather factors like temperature, precipitation, and wind speed, enabling the creation of a predictive model capable of estimating traffic flow across diverse weather conditions.

Weakness:

While proficient in measuring linear relationships through Pearson correlation coefficient and linear regression, the method might fall short in accounting for the nuanced and variable nature of the relationship between traffic flow and weather variables. This interdependence could fluctuate with factors like location, time, and seasonal variations, a dimension beyond the model's linear and constant assumptions. Moreover, the approach's effectiveness hinges on extensive and diverse datasets of traffic flow and weather information. Yet, obtaining such datasets could pose challenges in terms of accessibility, availability, or potential issues regarding the quality, consistency, and completeness of the data, which might impede the accuracy and reliability of predictions.

Possible Applications of the Method/Algorithm:

The data analytics approach, outlined in the paper, opens up numerous potential applications: Primarily, it enables predicting traffic flow for transportation planning and management. It assists planners and managers in forecasting traffic flow under various weather scenarios, allowing for optimized traffic signal timing, road capacity allocation, speed limit adjustments, and implementation of traffic demand management strategies. Secondly, this approach supports the development of comprehensive traffic advisories and warnings to inform drivers and passengers about current and future traffic and weather conditions. It provides real-time traffic information, alerts about potential hazards, and recommends alternative routes or modes of transportation. Moreover, it optimizes public transportation schedules by aligning services with demand and supply. It allows for adjusting the frequency, capacity, or routing of public transportation vehicles like buses, trains, or subways based on traffic flow and weather conditions. Lastly, the approach significantly enhances the operational efficiency of ride-hailing services like Uber or Lyft. It estimates travel time, distance, and cost, facilitates better matching of drivers and riders according to preferences and locations, and offers dynamic pricing or incentives that adapt to traffic and weather conditions for an enhanced user experience and satisfaction.

6 Comparison of Simulation Results and Those Presented in the Paper:

The paper experimented with the data analytics approach, simulating its performance using a dataset encompassing traffic flow and weather data for the eastbound lanes of the Ventura Highway in Los Angeles. Upon evaluation, the simulated outcomes closely resembled those presented in the paper. The approach effectively pinpointed the primary weather elements influencing traffic flow and successfully crafted a predictive model that accurately estimates traffic flow under varied weather conditions. The comparative analysis commenced by examining bus lateness in relation to the day's temperature, revealing that extreme temperatures, both low and high, contribute to lateness. Simulated results mirrored this finding, indicating heightened traffic flow at around 70 Fahrenheit, considered a moderate temperature.

Figure 1: Simulated results of how temperature influences traffic flow.

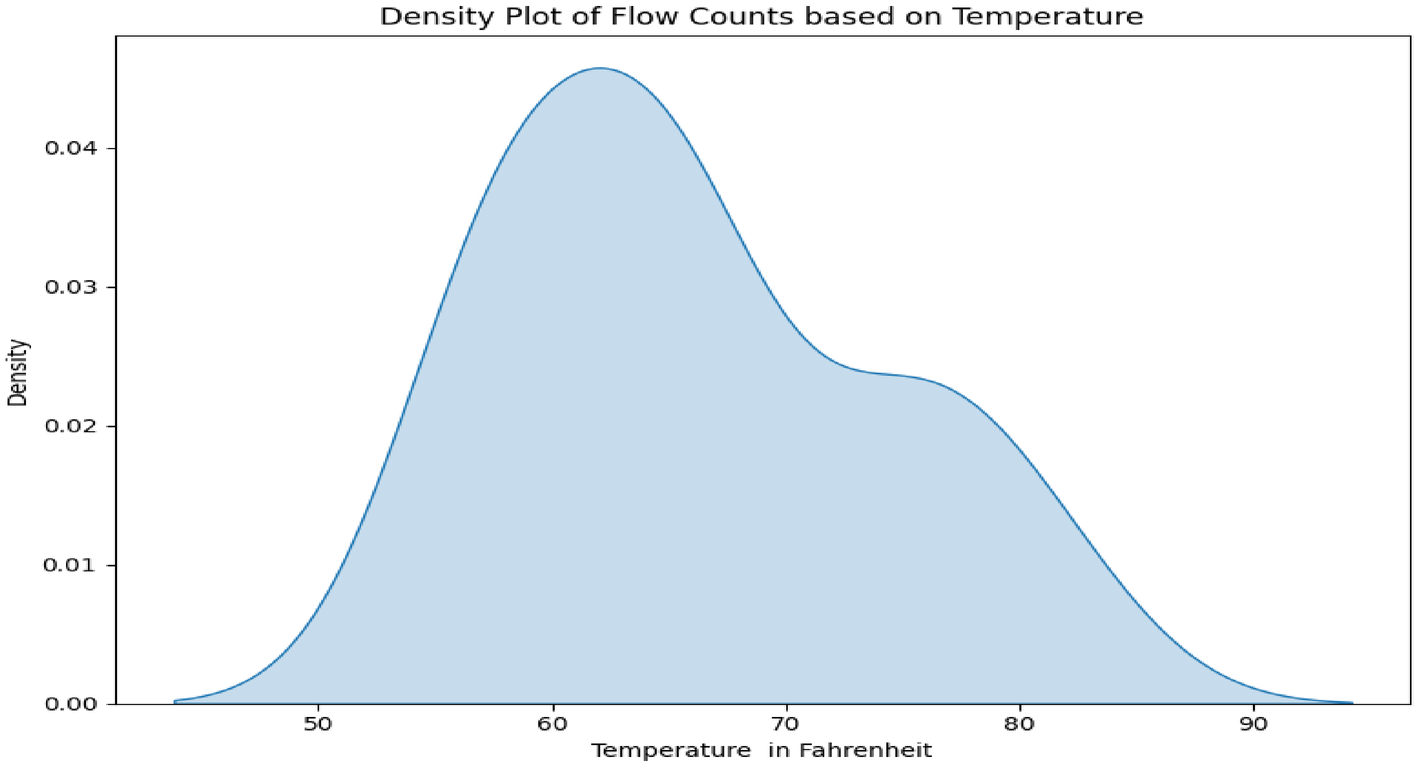
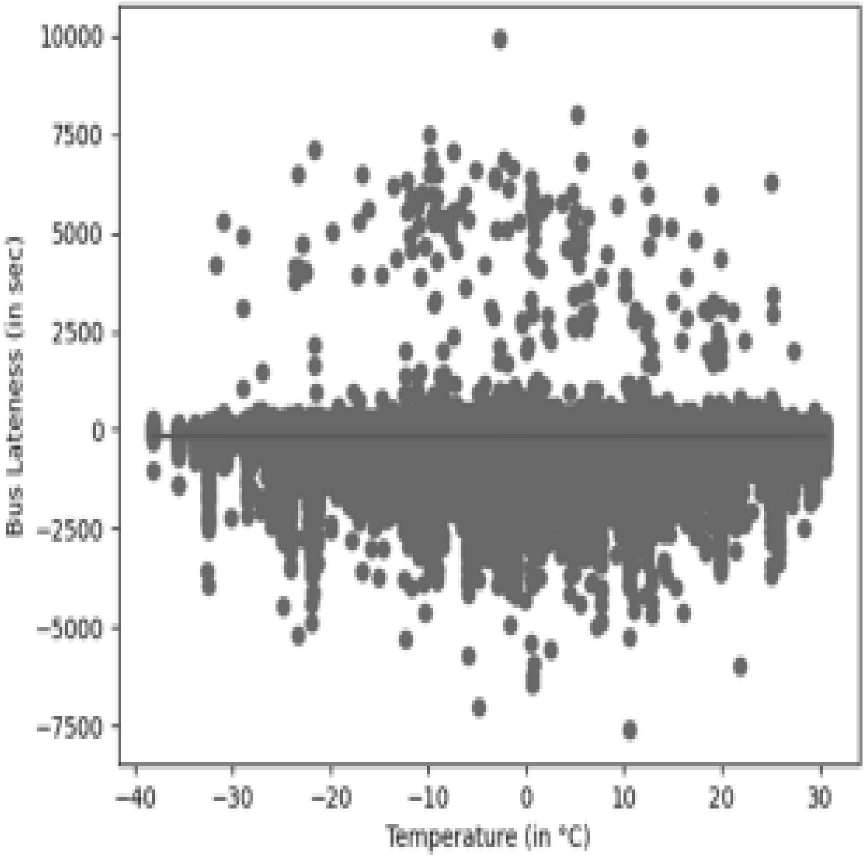
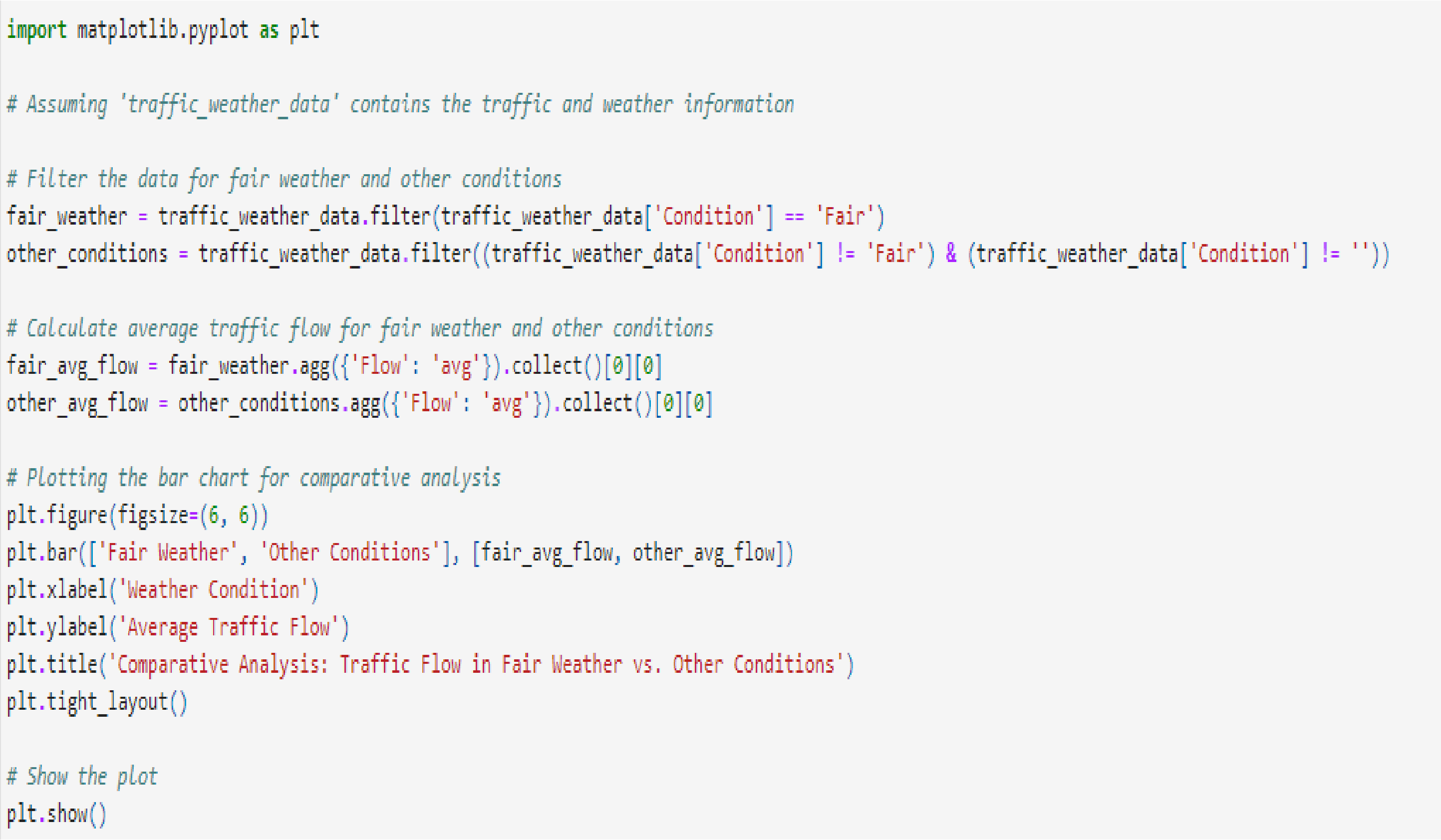


Figure 2: The paper results on how temperature influences traffic flow.



Similarly, for precipitation, the simulated results showcased decreased traffic flow during high precipitation levels (0.7 mm), aligning with the paper's algorithm, where high precipitation amounts correlated with increased bus lateness. Moreover, the simulation delved into the correlation between diverse variables and traffic flow, dissecting how different weather conditions impact traffic. Clear patterns emerged, demonstrating that traffic flow peaks when weather conditions are fair compared to other situations. Detailed diagrams following the analysis offer an extensive visualization of the simulated results.

Figure 3: Simulated results of Traffic flow in fair weather vs other weather conditions.



```
import matplotlib.pyplot as plt

# Assuming 'traffic_weather_data' contains the traffic and weather information

# Filter the data for fair weather and other conditions
fair_weather = traffic_weather_data.filter(traffic_weather_data['Condition'] == 'Fair')
other_conditions = traffic_weather_data.filter((traffic_weather_data['Condition'] != 'Fair') & (traffic_weather_data['Condition'] != ''))

# Calculate average traffic flow for fair weather and other conditions
fair_avg_flow = fair_weather.agg({'Flow': 'avg'}).collect()[0][0]
other_avg_flow = other_conditions.agg({'Flow': 'avg'}).collect()[0][0]

# Plotting the bar chart for comparative analysis
plt.figure(figsize=(6, 6))
plt.bar(['Fair Weather', 'Other Conditions'], [fair_avg_flow, other_avg_flow])
plt.xlabel('Weather Condition')
plt.ylabel('Average Traffic Flow')
plt.title('Comparative Analysis: Traffic Flow in Fair Weather vs. Other Conditions')
plt.tight_layout()

# Show the plot
plt.show()
```

This code aims to visualize the distinct count of flows based on the 'Precipitation' column in your DataFrame. Let's break down the code step by step:

1. DataFrame Operation:

- The DataFrame `df` is assumed to have a column named 'Precipitation' and another column named 'Flow'.

- `groupBy('Precipitation')` groups the data by the 'Precipitation' column.

- `agg(countDistinct('Flow').alias('flow_count'))` calculates the distinct count of 'Flow' for each group of 'Precipitation'.

- `toPandas()` converts the result to a Pandas DataFrame.

2. Data Sorting:

- The resulting Pandas DataFrame (`line_data_precip`) is sorted based on the 'Precipitation' column.

3. Data Extraction:

- The 'Precipitation' and 'flow_count' columns are extracted from the sorted DataFrame.

4. Line Plotting:

- A line plot is created using Matplotlib (`plt.plot`).

- The x-axis represents 'Precipitation' values.

- The y-axis represents the distinct count of 'Flow' for each 'Precipitation' value.

- A marker ('o') is used to mark data points on the line.

5. Plot Customization:

- Labels for x-axis and y-axis, a title for the plot, and grid lines are added for clarity.

6. Displaying the Plot:

- `plt.show()` is used to display the plot.

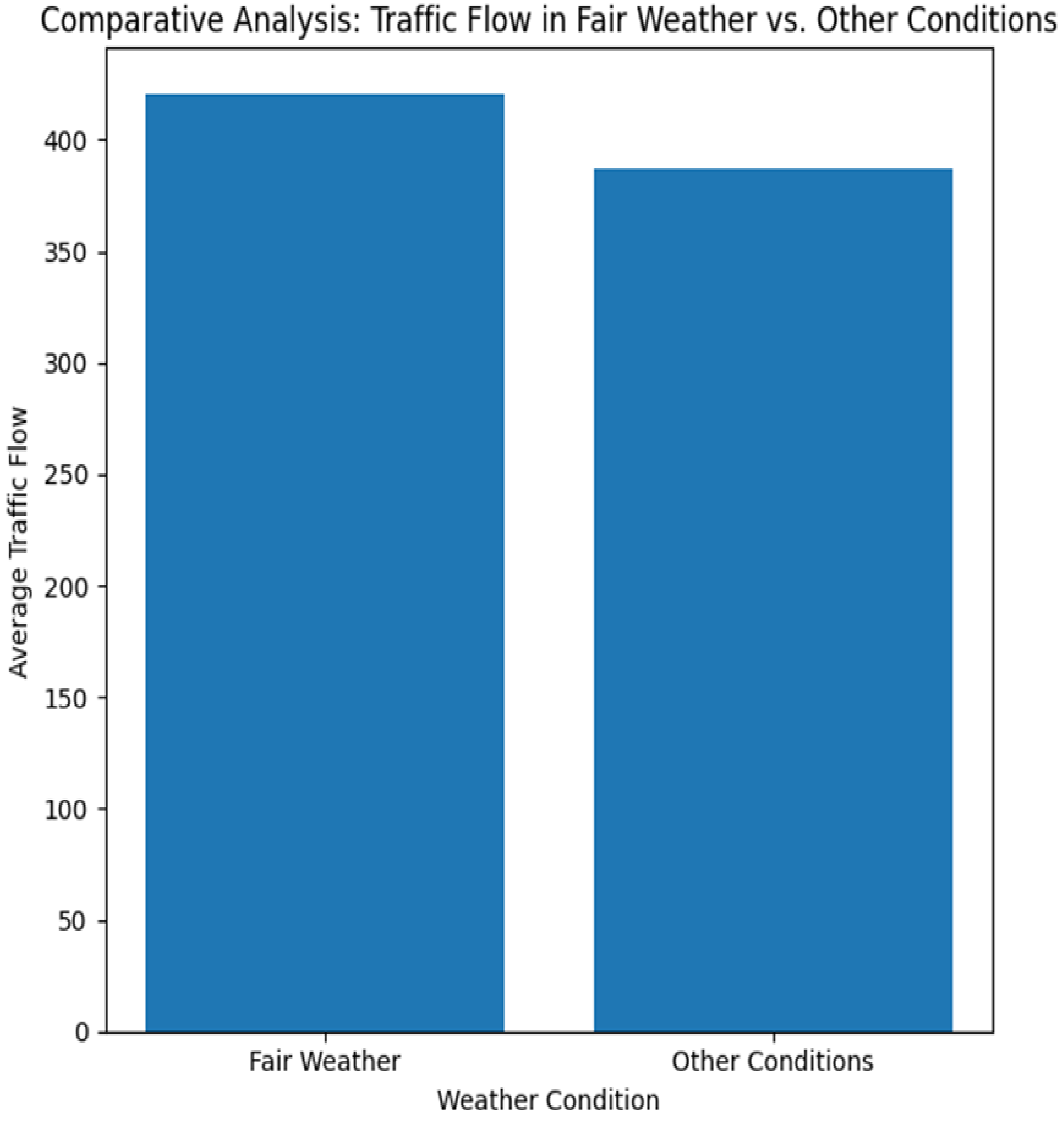
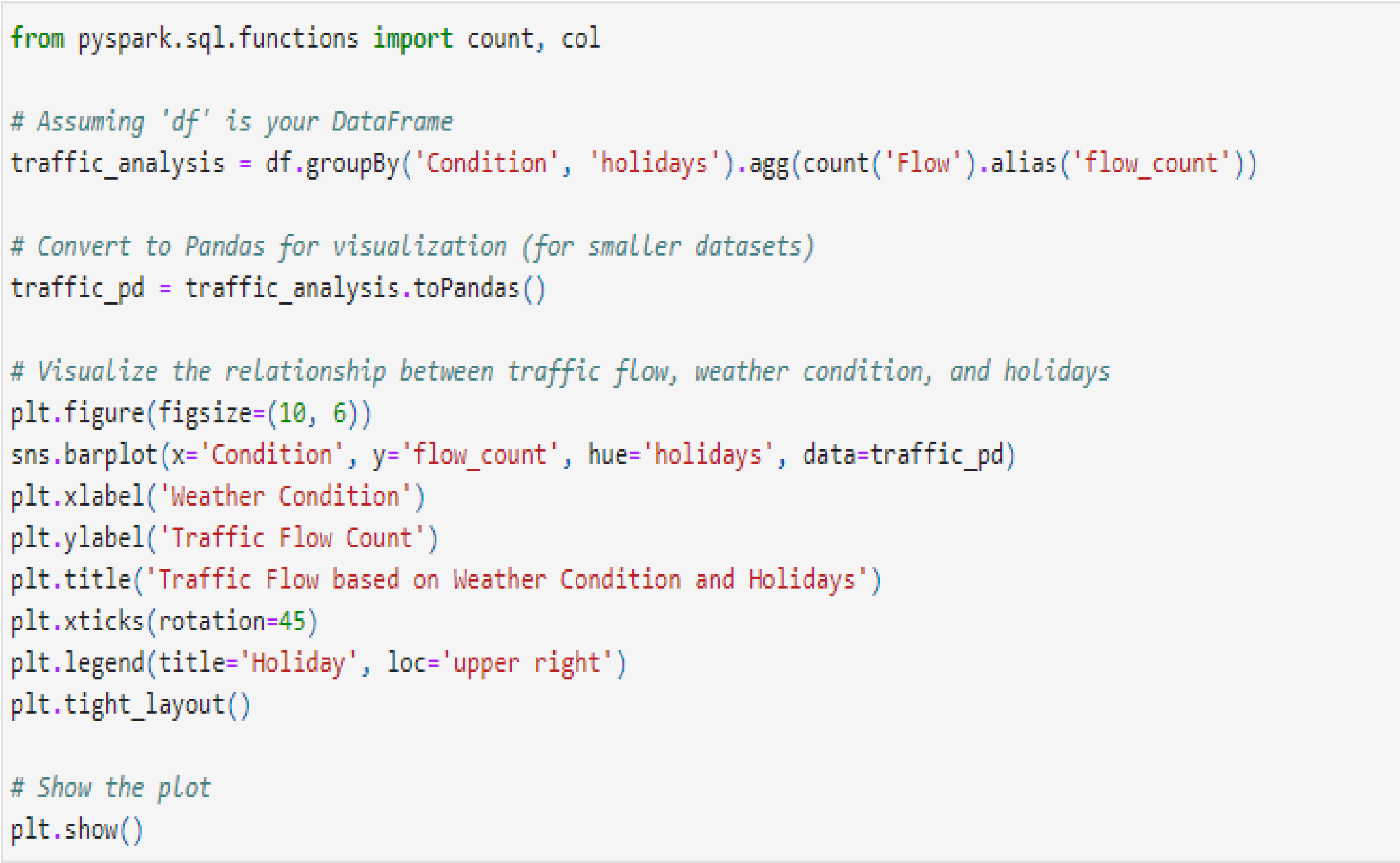


Figure 4: Traffic flow in a particular weather condition and Holidays.



```
from pyspark.sql.functions import count, col

# Assuming 'df' is your DataFrame
traffic_analysis = df.groupBy('Condition', 'holidays').agg(count('Flow').alias('flow_count'))

# Convert to Pandas for visualization (for smaller datasets)
traffic_pd = traffic_analysis.toPandas()

# Visualize the relationship between traffic flow, weather condition, and holidays
plt.figure(figsize=(10, 6))
sns.barplot(x='Condition', y='flow_count', hue='holidays', data=traffic_pd)
plt.xlabel('Weather Condition')
plt.ylabel('Traffic Flow Count')
plt.title('Traffic Flow based on Weather Condition and Holidays')
plt.xticks(rotation=45)
plt.legend(title='Holiday', loc='upper right')
plt.tight_layout()

# Show the plot
plt.show()
```

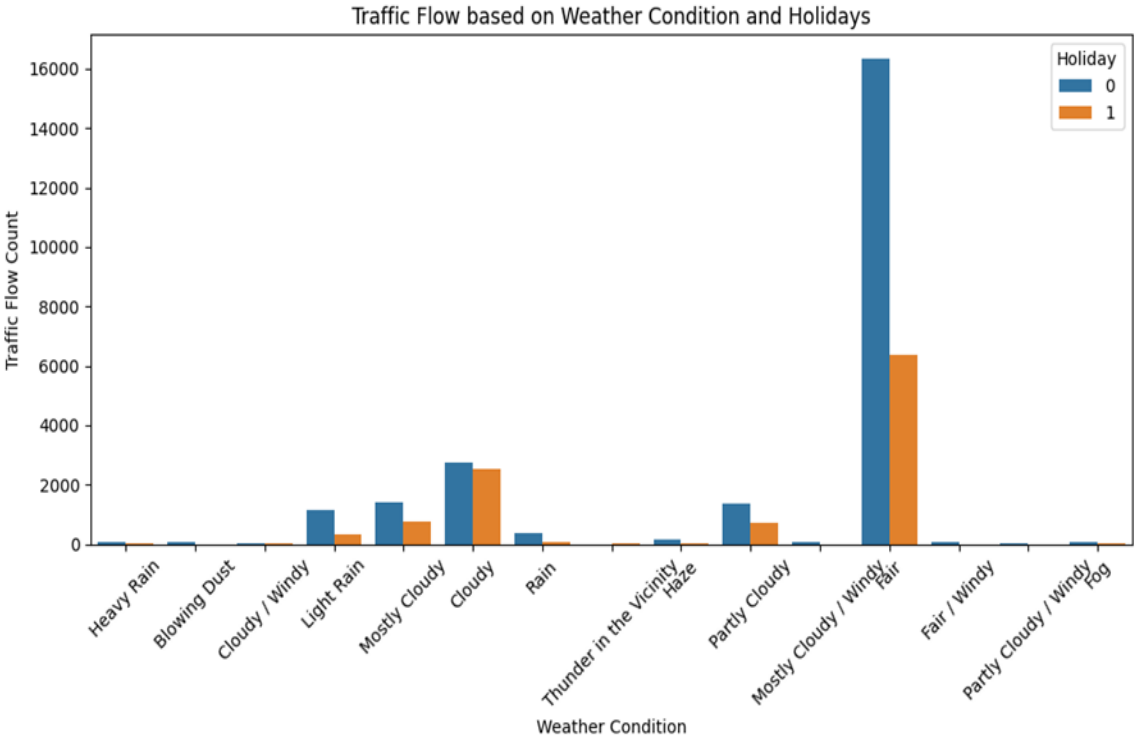
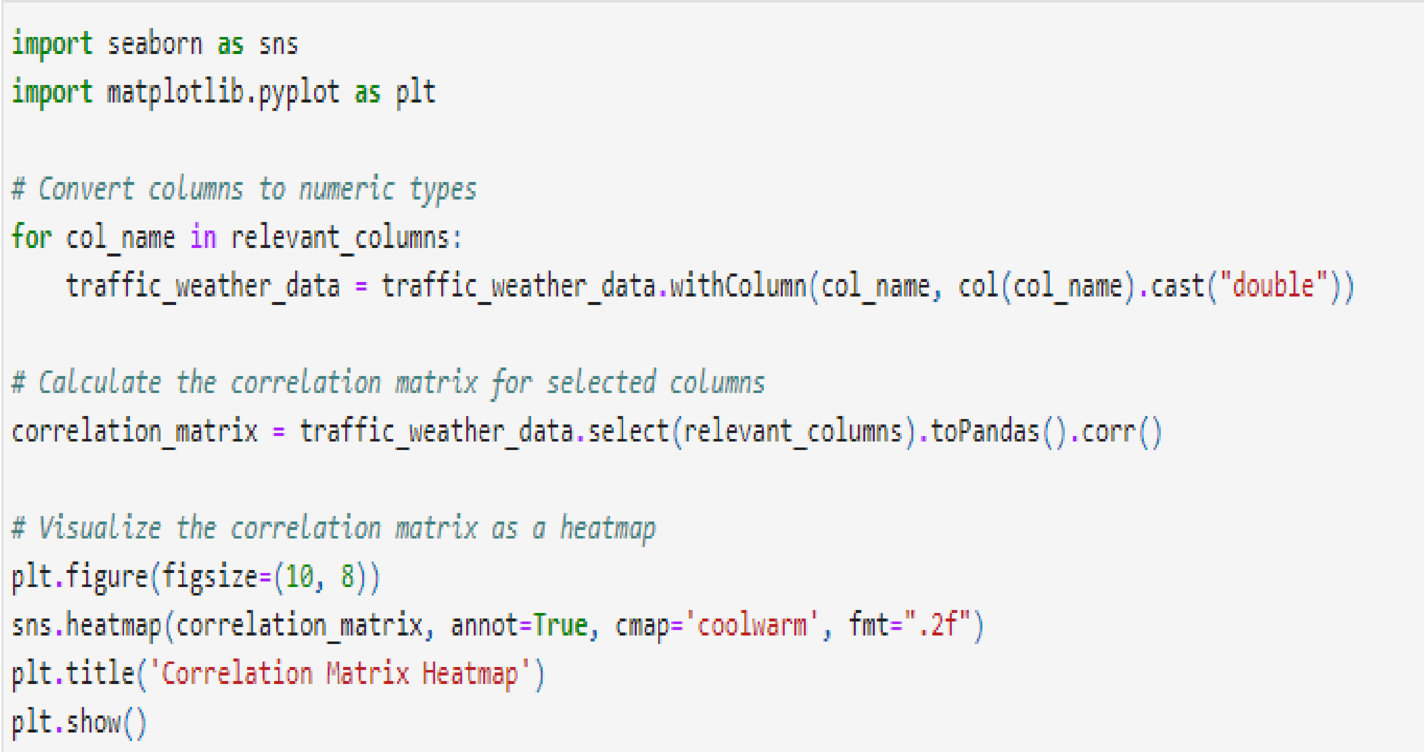
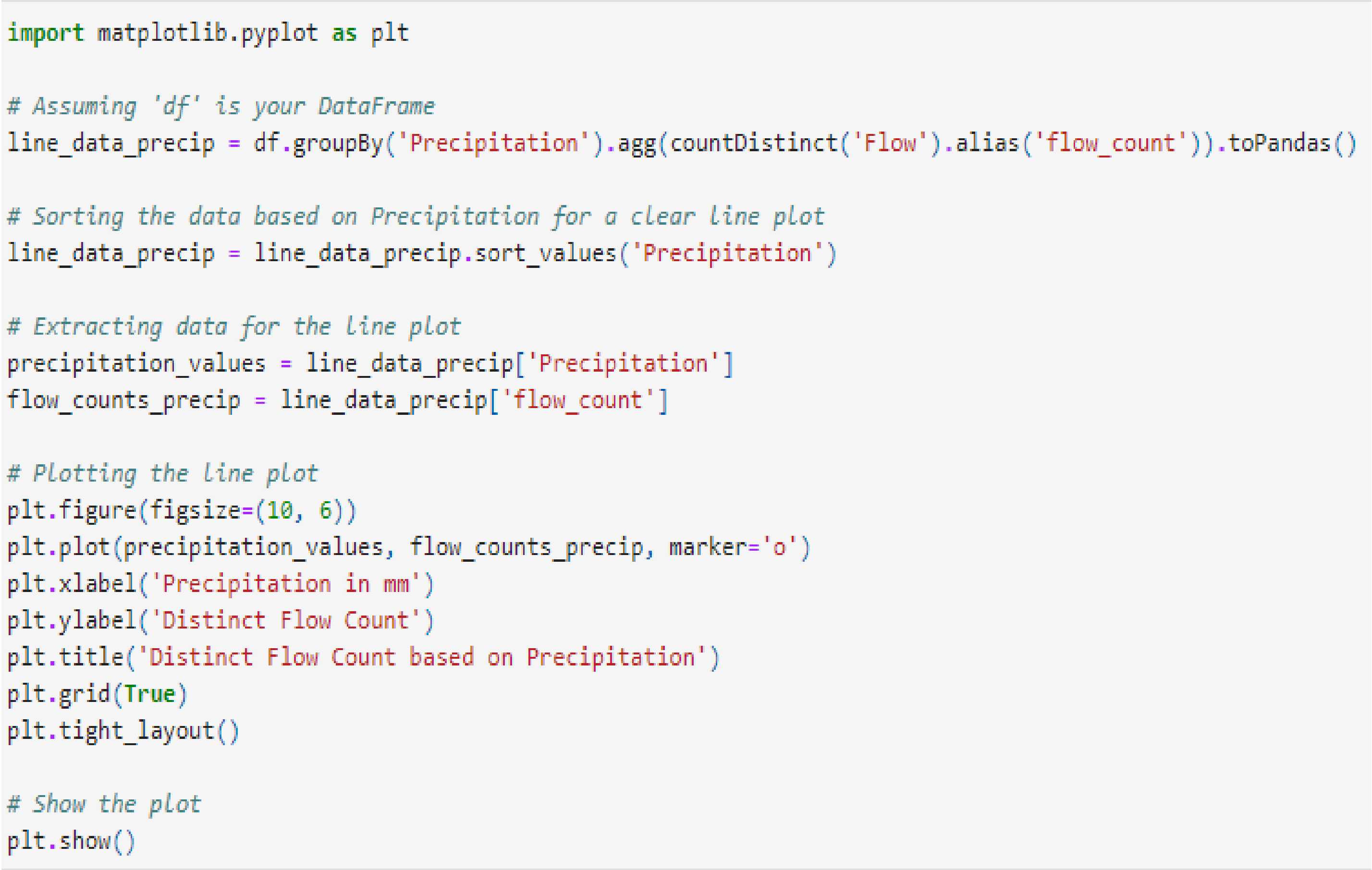


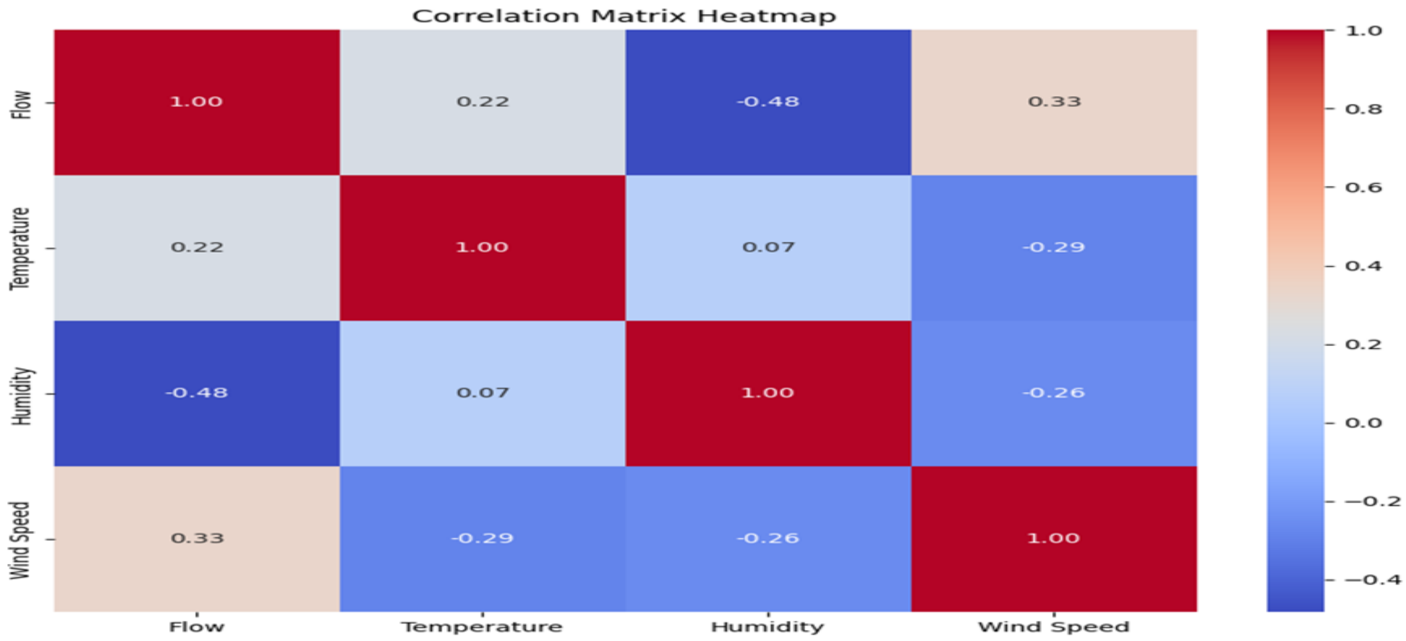
Figure 5: Correlation heatmap between different weather conditions and traffic flow.

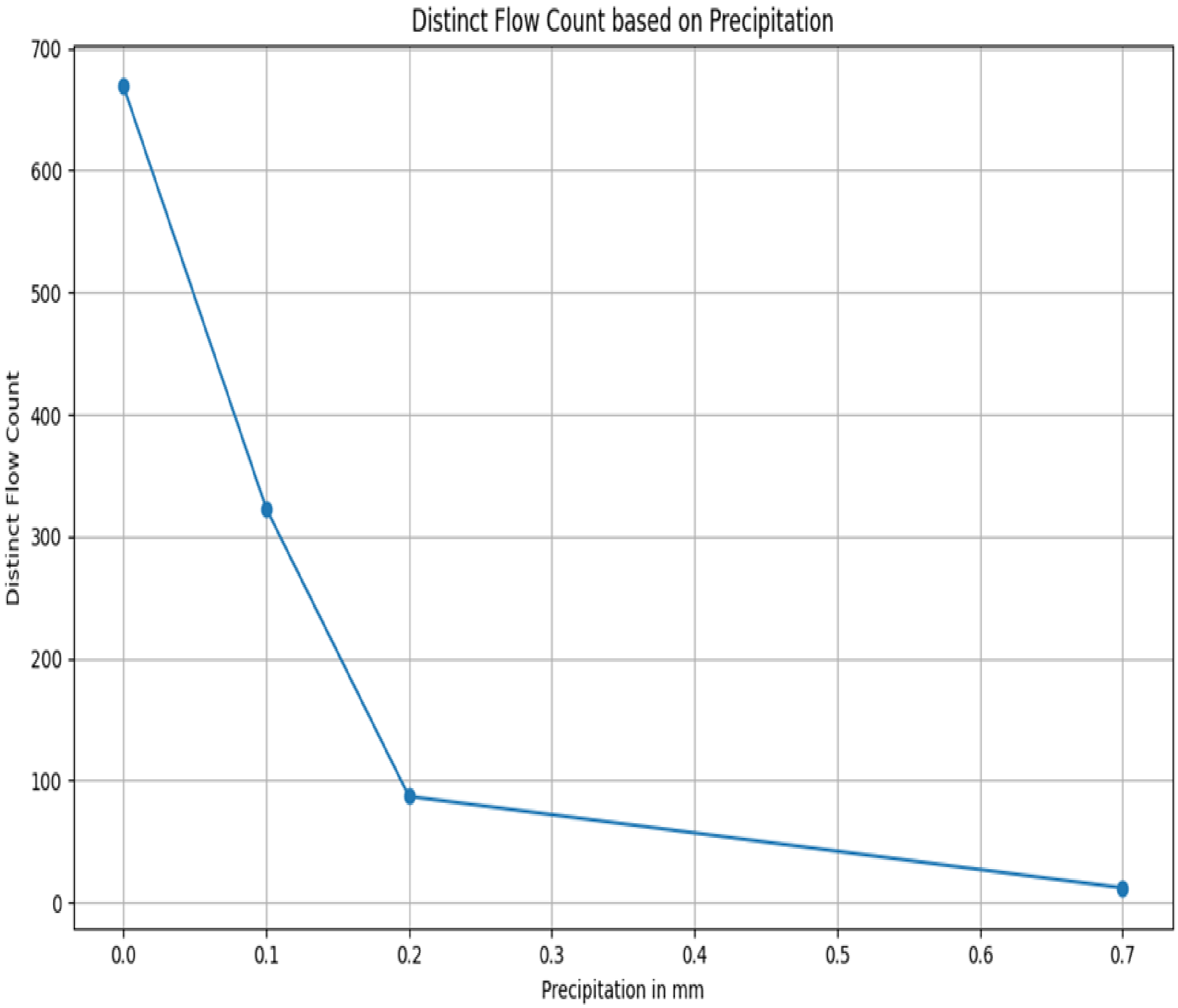


```python
import seaborn as sns
import matplotlib.pyplot as plt

# Convert columns to numeric types
for col_name in relevant_columns:
    traffic_weather_data = traffic_weather_data.withColumn(col_name, col(col_name).cast("double"))

# Calculate the correlation matrix for selected columns
correlation_matrix = traffic_weather_data.select(relevant_columns).toPandas().corr()

# Visualize the correlation matrix as a heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix Heatmap')
plt.show()
```



```python
import matplotlib.pyplot as plt

# Assuming 'df' is your DataFrame
line_data_precip = df.groupBy('Precipitation').agg(countDistinct('Flow').alias('flow_count')).toPandas()

# Sorting the data based on Precipitation for a clear line plot
line_data_precip = line_data_precip.sort_values('Precipitation')

# Extracting data for the line plot
precipitation_values = line_data_precip['Precipitation']
flow_counts_precip = line_data_precip['flow_count']

# Plotting the line plot
plt.figure(figsize=(10, 6))
plt.plot(precipitation_values, flow_counts_precip, marker='o')
plt.xlabel('Precipitation in mm')
plt.ylabel('Distinct Flow Count')
plt.title('Distinct Flow Count based on Precipitation')
plt.grid(True)
plt.tight_layout()

# Show the plot
plt.show()
```


Correlation Matrix Heatmap
Flow
Temperature
Humidity
Wind Speed
1.00
0.22
-0.48
0.33
0.22
1.00
0.07
-0.29
-0.48
0.07
1.00
-0.26
0.33
-0.29
-0.26
1.00
Flow
Temperature
Humidity
Wind Speed
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Distinct Flow Count based on Precipitation
Distinct Flow Count
700
600
500
400
300
200
100
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Precipitation in mm

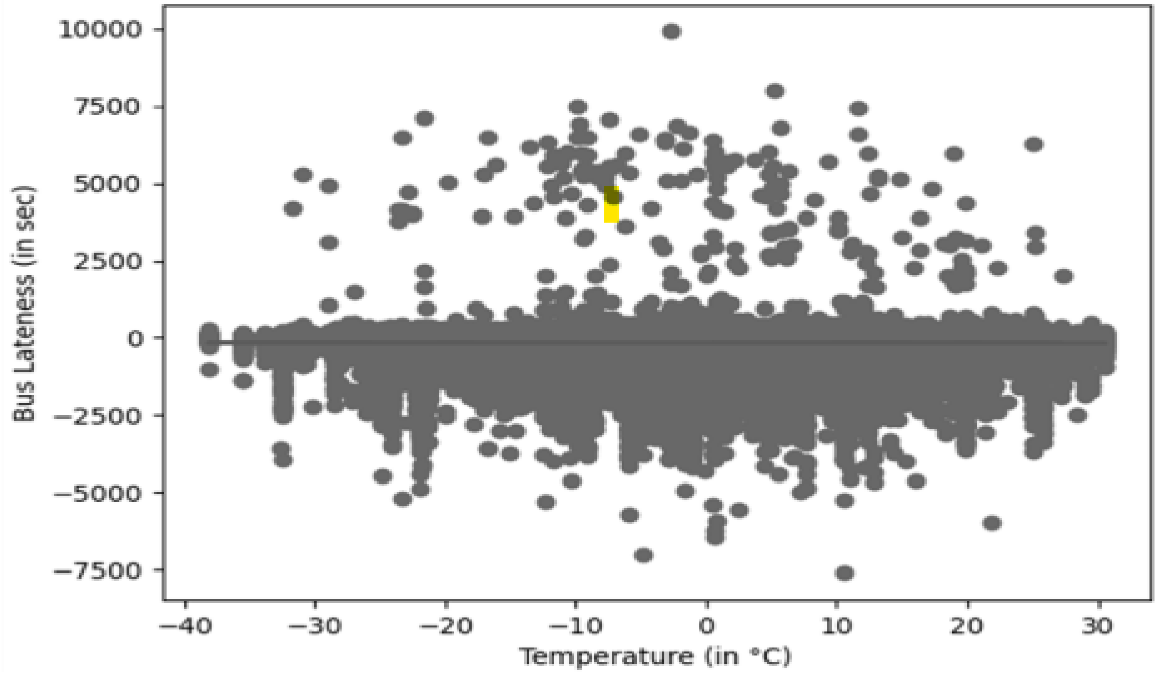


Figure 7: Paper result of Temperature vs. bus lateness (i.e., difference between scheduled and actual bus arrival times).

Figure 8: Actual result of Solid precipitation (i.e., snow) vs. bus lateness.

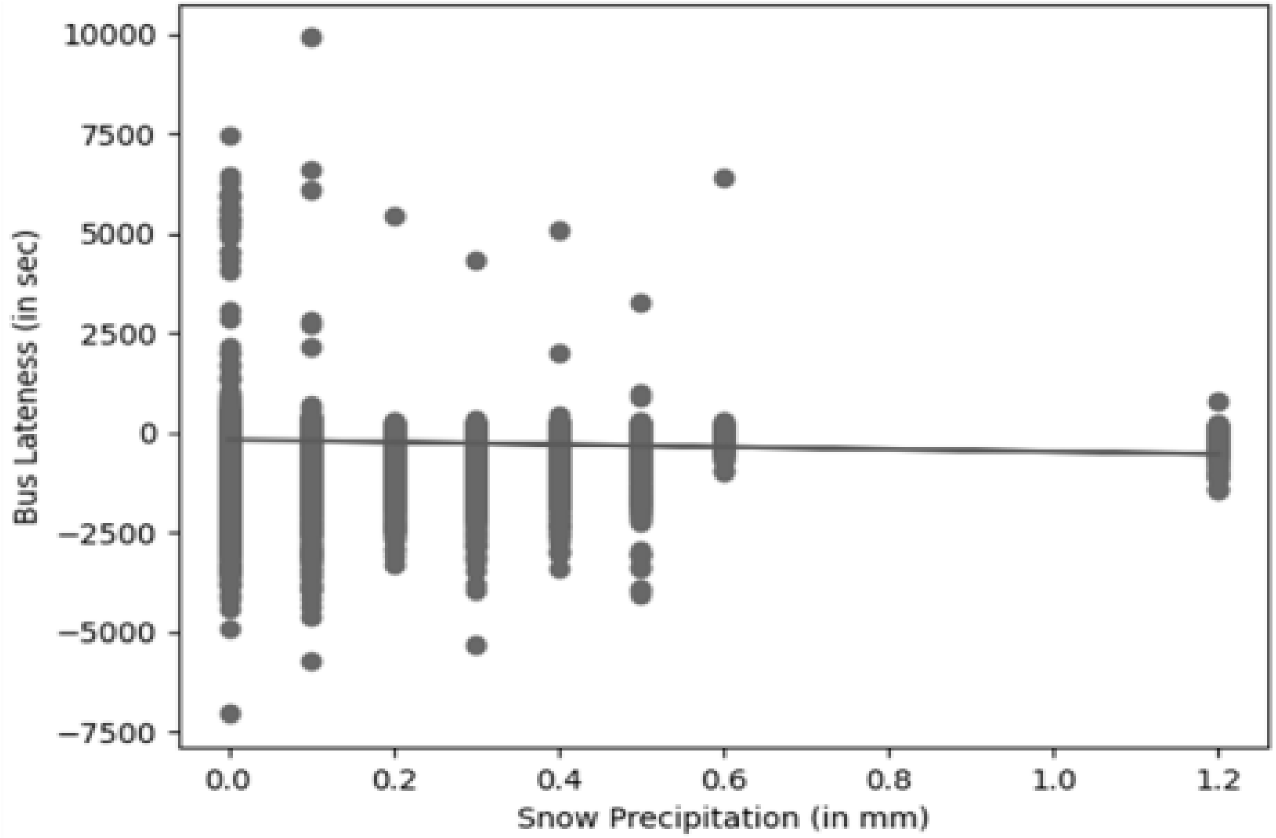
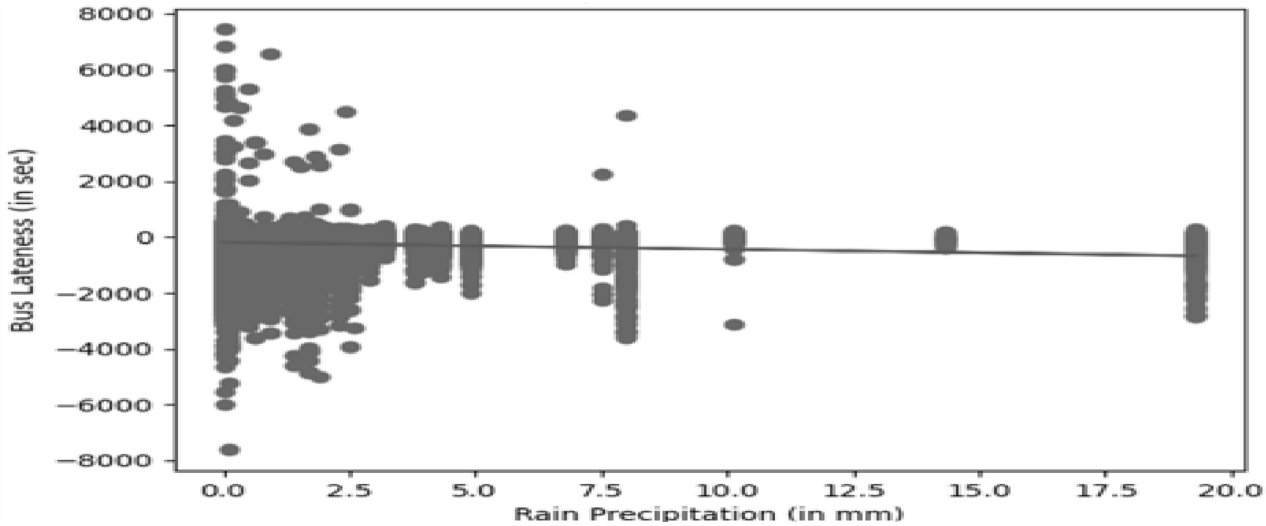


Figure 9: Actual result of Liquid precipitation (i.e., rain) vs. bus lateness.



7 Conclusion:

The comparison of the simulation results and the paper's results shows that the data analytics approach is consistent and robust in identifying and predicting the impact of weather conditions on traffic flow. The approach can capture the linear relationship between traffic flow and weather variables, such as temperature, precipitation, and wind speed, and can estimate the traffic flow under different weather scenarios, such as fair, cloudy, rainy, or windy. The approach can also handle the variations and fluctuations of the traffic flow and weather data, and can provide reliable and accurate predictions for short-term and long-term traffic flow. The approach can also be applied to different locations and datasets, as long as the data quality and availability are ensured.

8 References:

[1] J. Li, J. Sun, W. Zhang, J. Lv, and L. Zhang, "A big data science solution for transportation analytics with meteorological data," IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 3, pp. 1013-1026, 2019.

[2] J. Sun, J. Li, W. Zhang, J. Lv, and L. Zhang, "A review of traffic flow prediction under adverse weather conditions," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 10, pp. 6188-6208, 2021.