Team Members:

Tony Wilson, Rajeev Daithankar,

Matthew Harper, Anand Punwani

AI Boot Camp **Project 1**

Traffic Volume Analysis of I-94 Westbound between Minneapolis-St.Paul, Minnesota

Project Purpose / Description

This project will analyze a series of data points collected from the West Bound lane of I-94 in the Minneapolis-St Paul, MN area between 2012 and 2018. This data was collected on an hourly basis and includes the following:

- Traffic Volume
- Temperature
- Rainfall
- Snowfall
- Percentage of Cloud cover
- Holidays
- and much much more!!

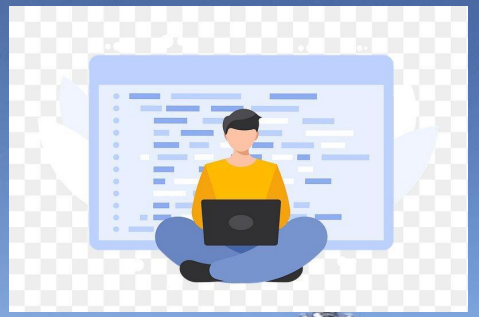
This project will analyze that data to evaluate correlations between the traffic volume and the other various data points.

Goals to be addressed

- Determine impact of holidays on traffic volume
- Determine impact of rain on traffic volume
- Determine impact of snow on traffic volume
- Determine impact of clouds on traffic volume
- Determine impact of temperature on traffic volume

Data Exploration

- **What is Data Exploration?**
  - The data exploration is done to gain understanding / insight about the selected data
- **Typical steps involved are**
  - Understanding of variables of the dataset, number of rows, columns, data types.
  - Identifying missing values, duplicates, outliers and handling them.
  - Understanding of Central tendency(Statistics), Mean, Mode, Median and dispersion, distribution.
  - Identifying pattern, trends and relationships in data through visualization techniques such as histograms, scatter plots and heatmaps.



Data Collection

Data Source : UC Irvine Machine Learning Repository

https://archive.ics.uci.edu/dataset/492/metro+interstate+traffic+volume

Understanding Data

Variable Name	Role	Type	Description	Units
holiday	Feature	Categorical	US National holidays plus regional holiday, Minnesota State Fair	
temp	Feature	Continuous	Average temp in kelvin	Kelvin
rain_1h	Feature	Continuous	Amount in mm of rain that occurred in the hour	mm
snow_1h	Feature	Continuous	Amount in mm of snow that occurred in the hour	mm
clouds_all	Feature	Integer	Percentage of cloud cover	%
weather_main	Feature	Categorical	Short textual description of the current weather	
weather_description	Feature	Categorical	Longer textual description of the current weather	
date_time	Feature	Date	Hour of the data collected in local CST time	
traffic_volume	Target	Integer	Hourly I-94 ATR 301 reported westbound traffic volume	

Data Cleanup

Identifying and Handling missing values:

- Holiday values are present only at 00:00 hour of the date, missing at other hours for the date. Created a program to fill in missing holidays.
- Missing hour sequences for a particular date. To handle this, data can be manufactured.
  - Pros and Cons of manufacturing data to fill in missing data
    - Pros: Completeness, Preserves structure
    - Cons: Not real data, leading to bias, Incorrect results.

  We have decided not use manufactured data for this assignment, in order avoid biased results

- **Checking Duplicates:**
  - Found duplicates for date_time and traffic volume combination. Removed these duplicates.

holiday	date_time
Columbus Day	10/8/2012 0:00
None	10/8/2012 1:00
None	10/8/2012 2:00

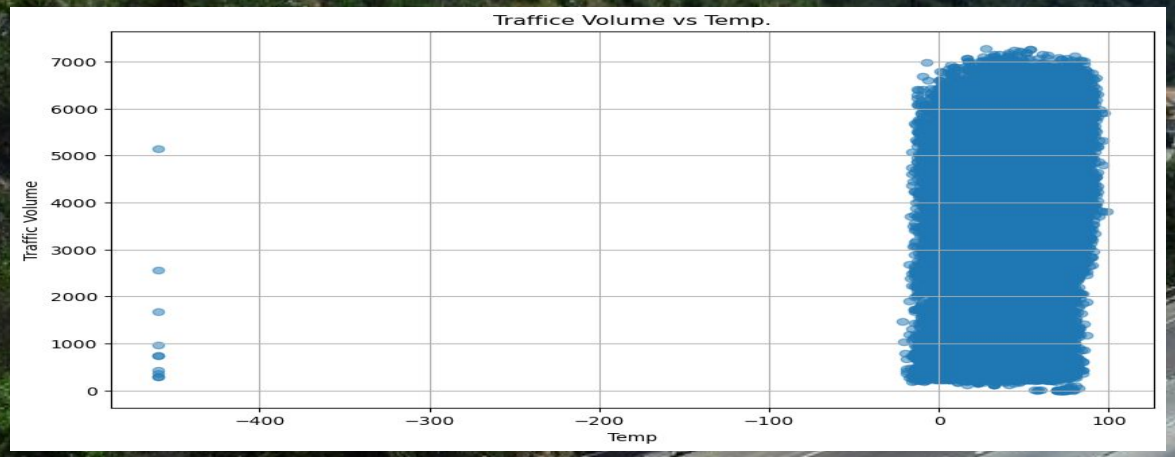
holiday	date_time
Columbus Day	10/8/2012 0:00
Columbus Day	10/8/2012 1:00
Columbus Day	10/8/2012 2:00

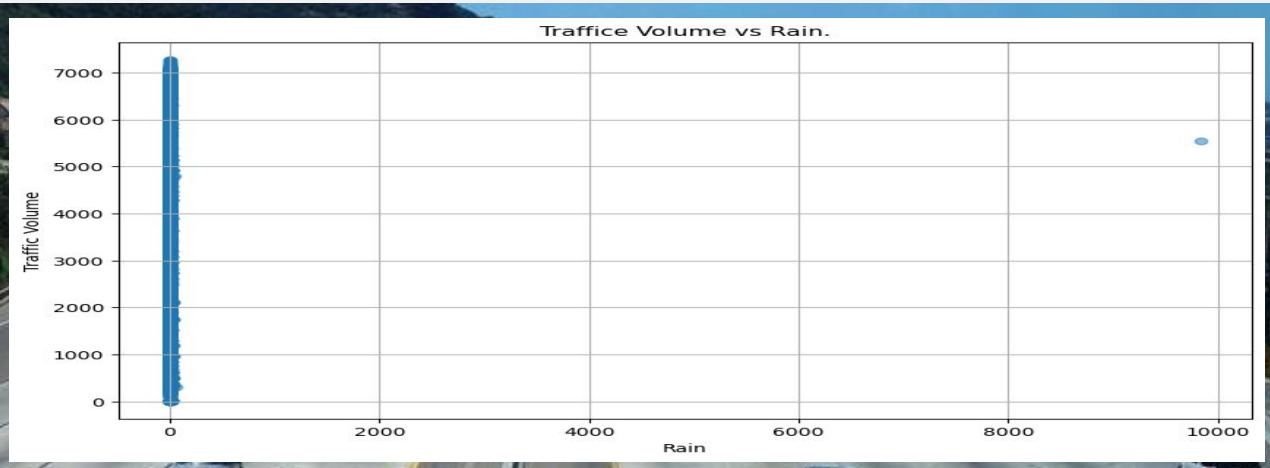
Original Seq.	Missing Seq.
9/21/2017 9:00	
9/21/2017 13:00	9/21/2017 10:00
9/21/2017 14:00	9/21/2017 11:00
9/21/2017 15:00	9/21/2017 12:00

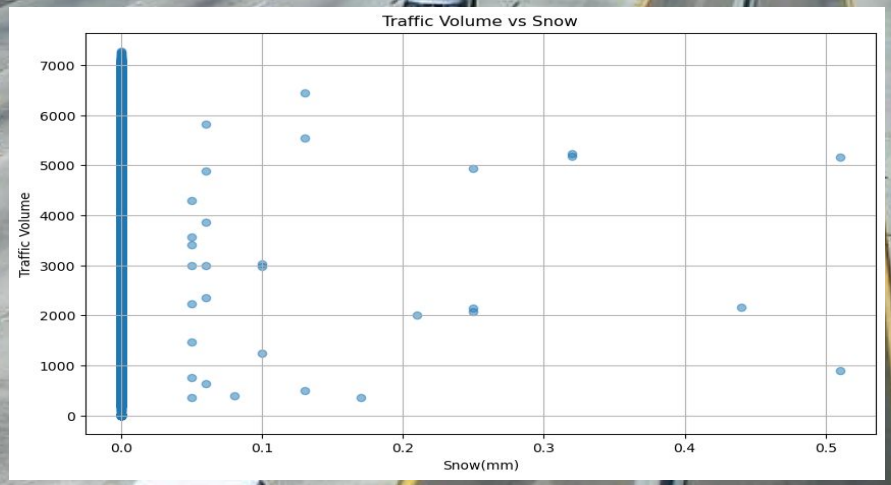
temp	weather_main	weather_description	date_time	traffic_volume
59	clear	sky is clear	9/21/2017 10:00	2500
59	clear	sky is clear	9/21/2017 11:00	2500
59	clear	sky is clear	9/21/2017 12:00	2500

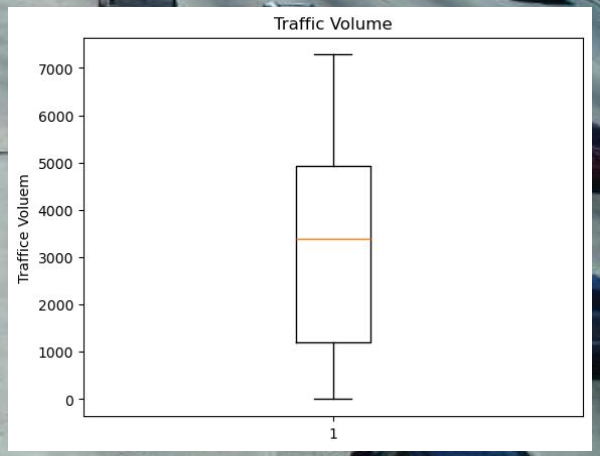
Data Cleanup

- **Identifying and removing outliers:**
  - With the help of scatter plots we found that there are outliers in variables Temp,Rain
  - Removed outliers : rain <407 mm and temp temp > -50 Deg. F (converted temp Kelvin to Fahrenheit)
  - No outliers for Cloud % and Snow mm
  - Box plot for Traffic Volume - No outliers for traffic volume
    - Q1: 1194.0, Q3 : 4933.0. IQR: 3739.0
      - Values below -4414.5 could be outliers , Values above 10541.5 could be outliers..





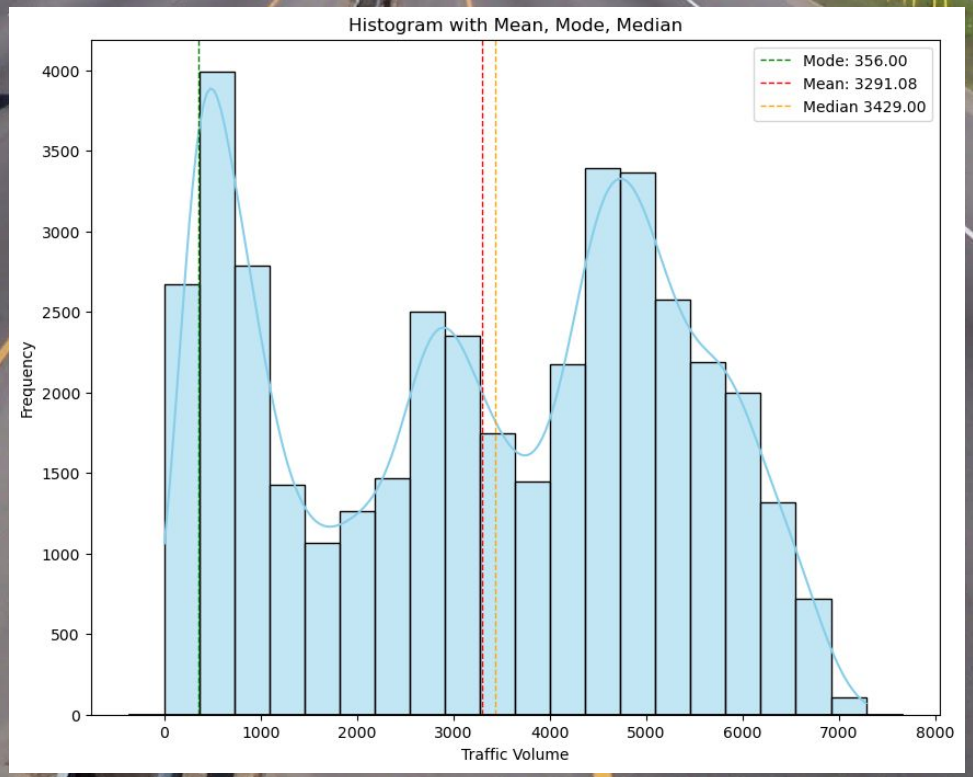




Data Exploration

- **Central Tendency**
  - **Mean, Mode, Median** of traffic_volume
  - Mean: 3291.08, Median: 3429.00 Mode: 356, count=40

  For this dataset
  - The mode is significantly lower than both the mean and median, indicating that there is a peak towards lower end of distribution.
  - In this case, the distribution is **negatively skewed** (mean is less than median).This means that the tail of the distribution extends towards lower values, while bulk of data is concentrated towards higher values.



Data Exploration

Measures of Dispersion

- Range, Variance, Standard Deviation, Interquartile Range (IQR)
- The range of Traffic Volume is : 7280
- The population variance is : 3938694.26
- The population standard deviation is : 1984.61

Mean and Std. Deviation

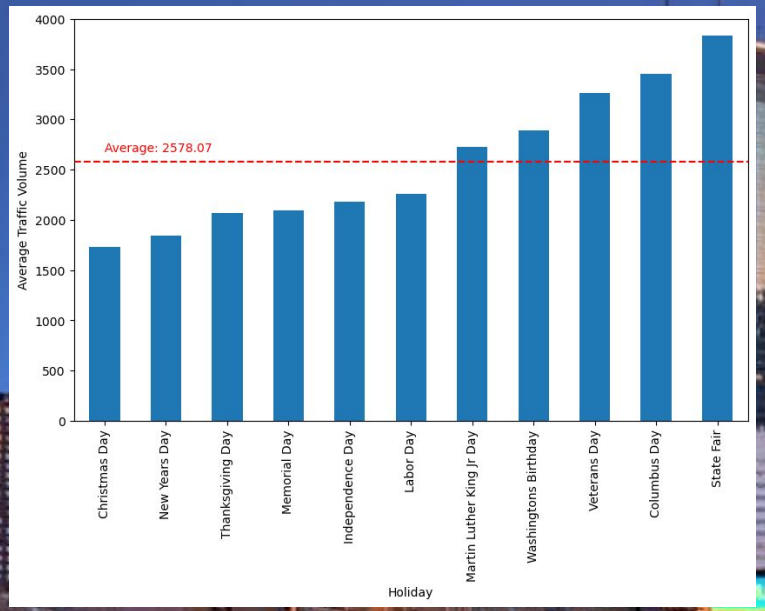
Mean of traffic volume is 3291.08 (SD=1984.61)
SD of 1984.61 indicates how much variation there is from mean 3291.08

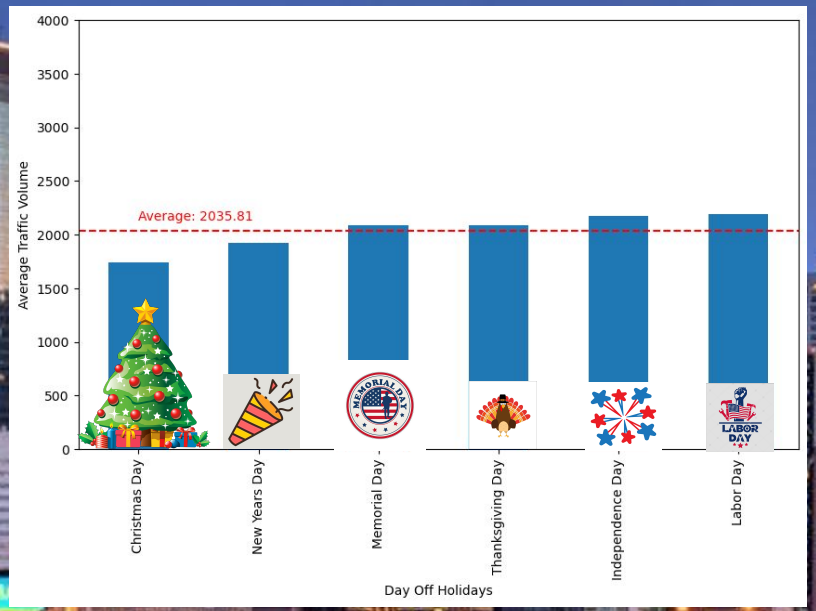
Interpretation:

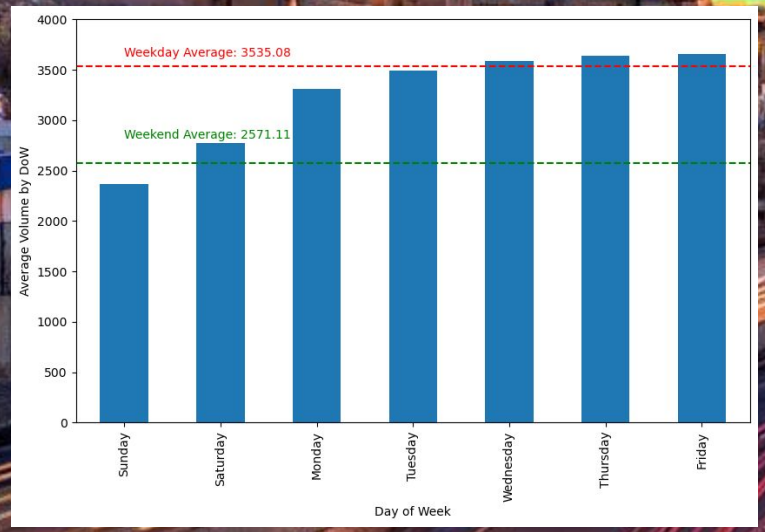
1. Large SD indicates that data points are spread out widely around mean.
2. Mean is greater than SD, implies that there is significant variability or dispersion in dataset, with data points scattered over a wide range.

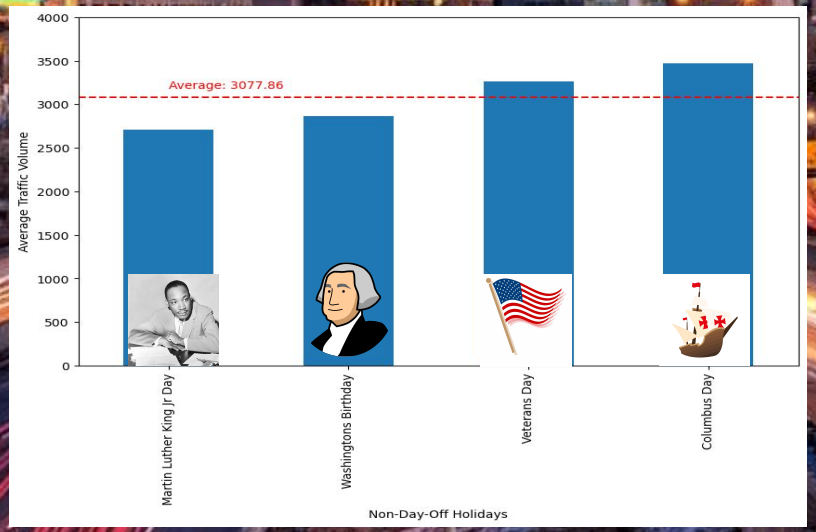
The Holiday Impact

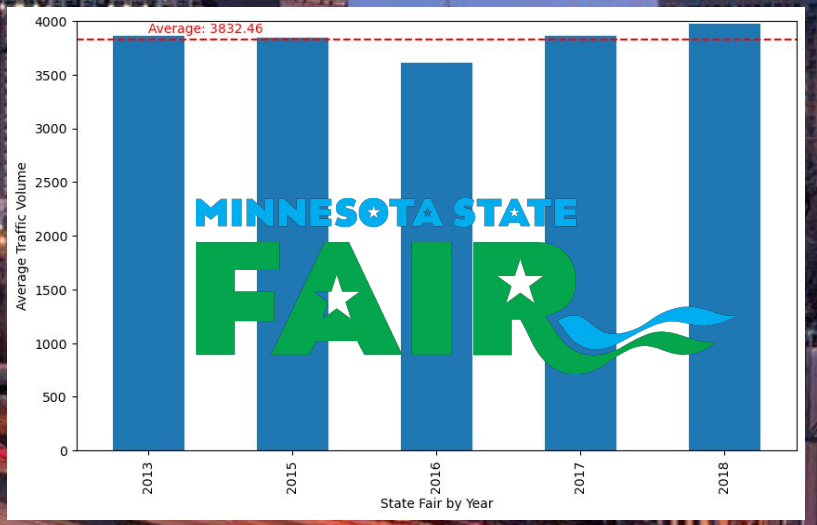
Exploring the impact of holidays on traffic volume as compared to average (non-holiday) days of the week.







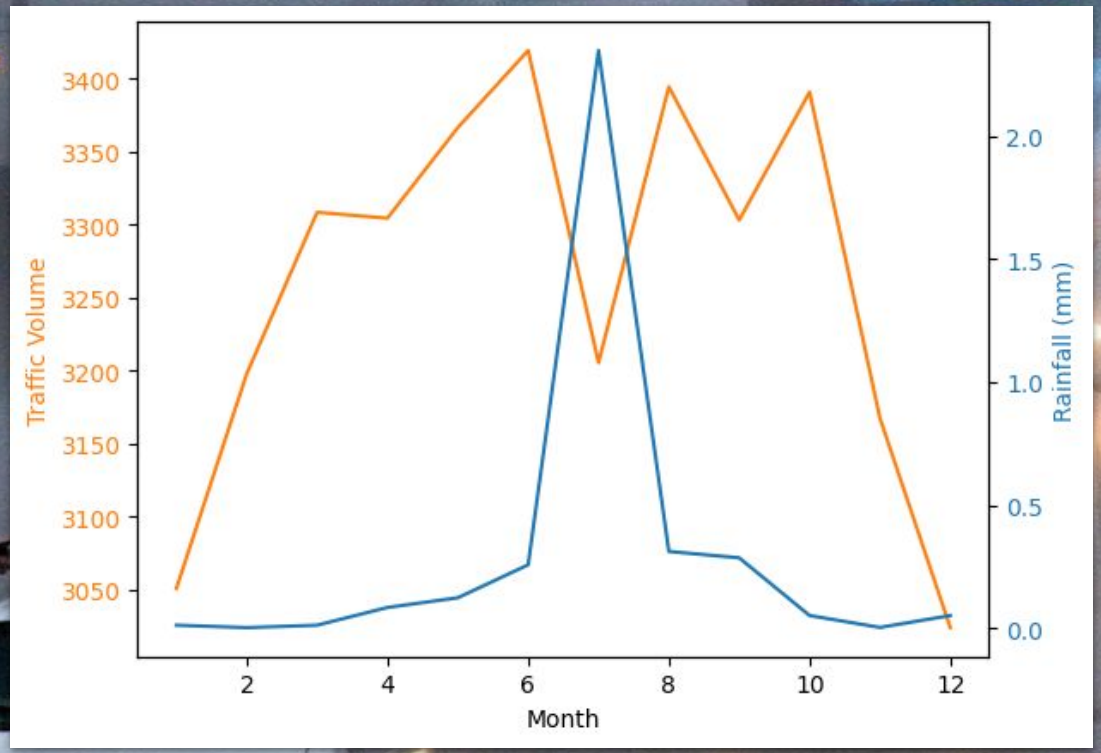


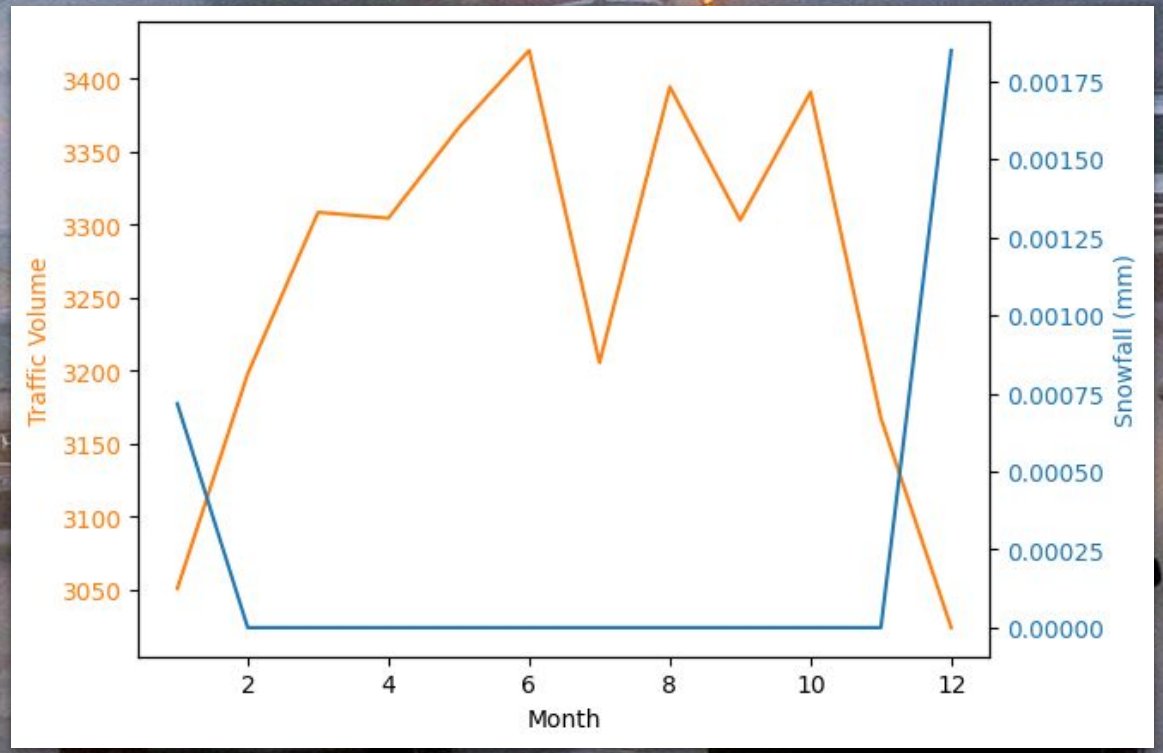


Precipitation Impact

Monthly Average Hourly Traffic Volume graphed against Monthly Average Hourly Rainfall and Snowfall

- There is a very slight negative correlation between traffic volume and rainfall (-0.01)
  - Most noticeable around July although this could be for a number of reasons (construction/vacations)
- Virtually no correlation between traffic volume and snowfall (0.00)





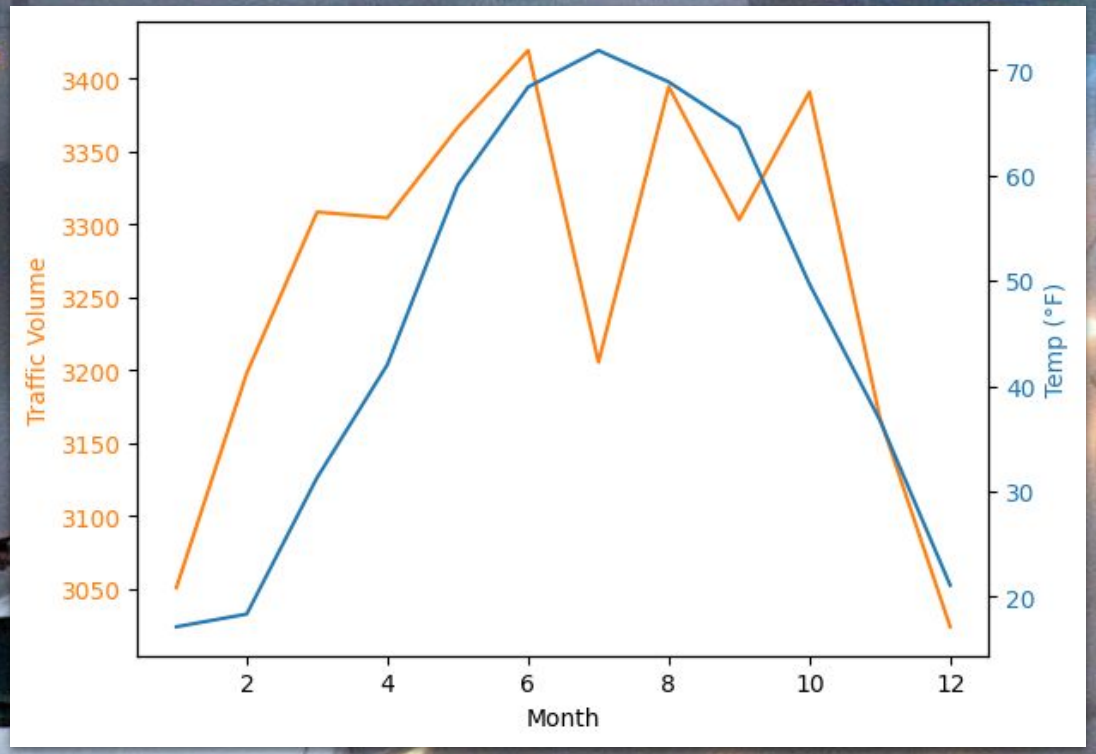
Temperature & Cloud Cover Impact

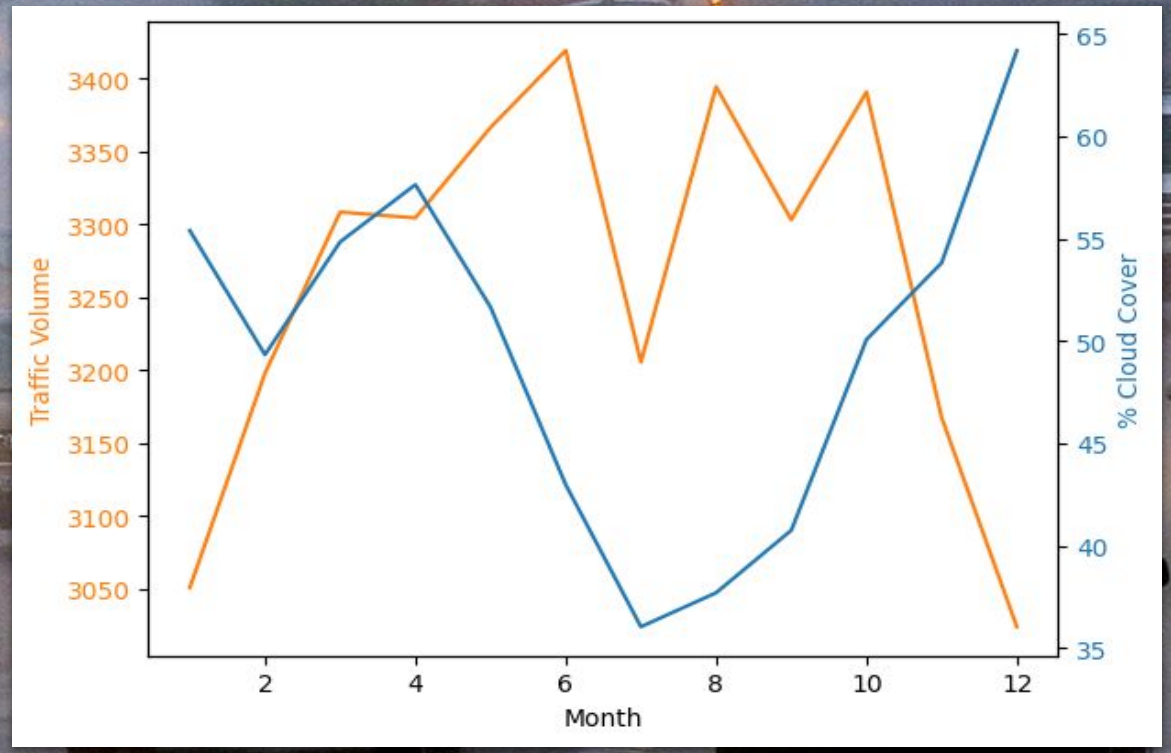
Monthly Average Hourly Traffic Volume graphed against Monthly Average Hourly Temperature

- This was our strongest correlation (0.14) but still very weak.

Monthly Average Hourly Traffic Volume graphed against Monthly Average Hourly Cloud Cover

- Correlation was .08 which again suggests very little correlation



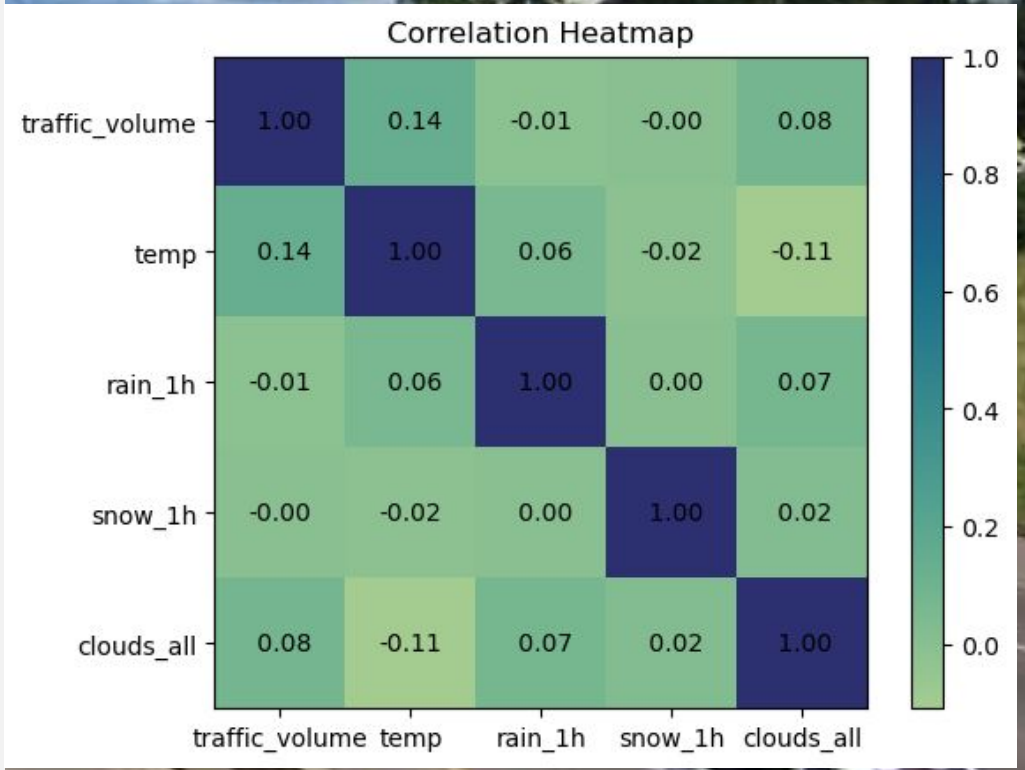


What is the correlation between variables and increased traffic volume?

There have been two types of variables given, whether it was a holiday, or whether there were weather based variables.

- Based on the data, National holidays do not show a strong enough correlation overall to provide much towards the increase or decrease of traffic volume. The holiday data does show that there may be some other cause during these times to influence the traffic volume based on the specific holiday

- Overall the temperature itself had a higher correlation to traffic volume over any specific weather condition but was still not high enough to be a strong correlation and thus suggests other reasons as to why they seem to correlate.



Summary

- First: Based on our analysis of the data there is no strong correlation between the weather and whether or not the traffic volume increases. Any correlation between weather based variables must be due to other circumstances such as job attendance and similar.

- Second: Based on our analysis of the data, although there is no strong correlation between there being a national holiday and the traffic volume increasing, what is apparent is that depending on how a national holiday is viewed and their general practices, traffic volume will raise or lower from the median accordingly

Problems Encountered

1. Use of Seaborn library to plot histogram for correlation matrix. All values for correlation matrix are not displayed in cell.
Resolution: Instead of Seaborn, used matplotlib and it worked.

2. Data had rows with duplicate date_times but other differing data which caused certain data to receive duplicate values which may have skewed the data slightly.

Future Considerations

Questions for Future Development

- If the data would be available, we believe it would be interesting to look at the same data but during the COVID lockdowns, to see if the same trends showed up even with the decrease in traffic volume. In the same vein it would be interesting to look at other roadways traffic volume data to compare and see if the same trends appear in different areas of the United States

- As an additional inquiry, we would like to find the data for the year after our data ends and test it against a Prophet forecast to see how current events may affect the true data vs the prediction