**NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES**

**Peshawar Campus**

**

**DATA SCIENCE**

**PROJECT**

**Name:**

**HASNAT AHMAD (20P-0079)**

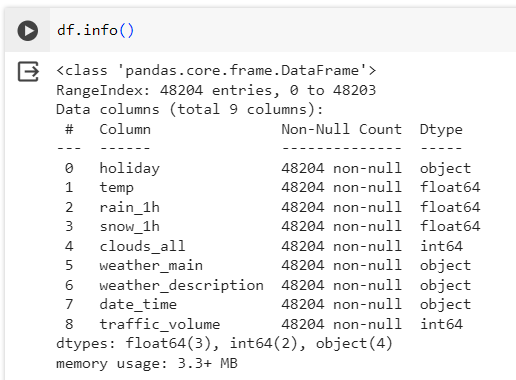
**1.INTRODUCTION:**

The project aims to analyze the Metro Interstate Traffic Volume.Hourly dataset, focusing on westbound I-94 traffic in Minneapolis-St Paul, MN. The goal of our analysis is to determine a few indicators of heavy traffic on I-94. These indicators can be weather type, time of the day, time of the week, etc.

**2.DATASET DESCRIPTION:**

The dataset has 48204 instances and 9 columns. The attributes are:

1. **holiday (Categorical):**
   1. Includes US National holidays and regional holidays such as the Minnesota State Fair.
2. **temp (Numeric):**
   1. Average temperature in Kelvin.
3. **rain\_1h (Numeric):**
   1. Amount of rain in mm that occurred in the hour.
4. **snow\_1h (Numeric):**
   1. Amount of snow in mm that occurred in the hour.
5. **clouds\_all (Numeric):**
   1. Percentage of cloud cover.
6. **weather\_main (Categorical):**
   1. Short textual description of the current weather.
7. **weather\_description (Categorical):**
   1. Longer textual description of the current weather.
8. **date\_time (DateTime):**
   1. Hour of the data collected in local CST time.
9. **year (Integer):**
   1. Year extracted from the date.
10. **month (Integer):**
    1. Month extracted from the date.
11. **day (Integer):**
    1. Day extracted from the date.
12. **traffic\_volume (Numeric):**
    1. Hourly I-94 ATR 301 reported westbound traffic volume (target variable).



**3.DATA LIFE CYCLE:**

**1. QUESTION:**

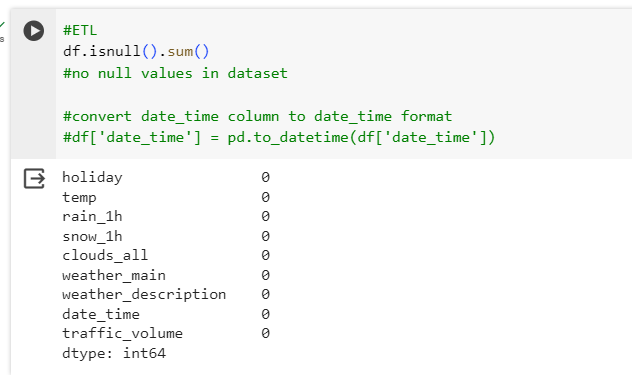
* Objective:
  + Understand and analyze factors influencing westbound I-94 traffic volume in Minneapolis-St Paul, MN.
* Key Questions:
  + What are the main drivers of traffic volume? How do weather conditions and time of day impact traffic?

**2. ACQUIRE:**

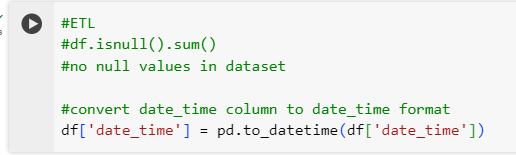
* + Dataset was obtained from Kaggle website

**3. ETL:**

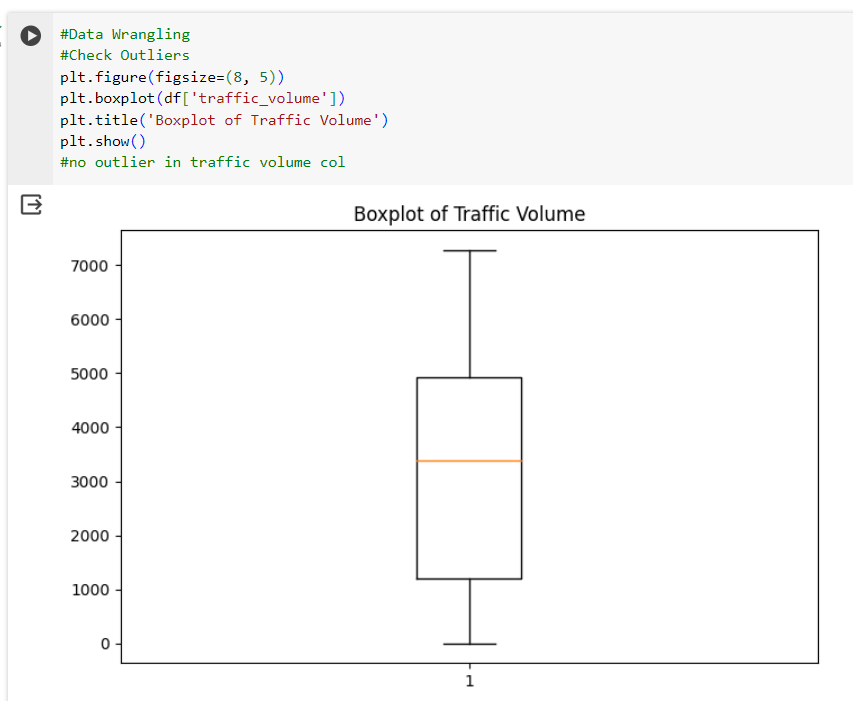
* + Checked for missing values (no null values were found).



* + Converted the 'date\_time' column to the datetime format.

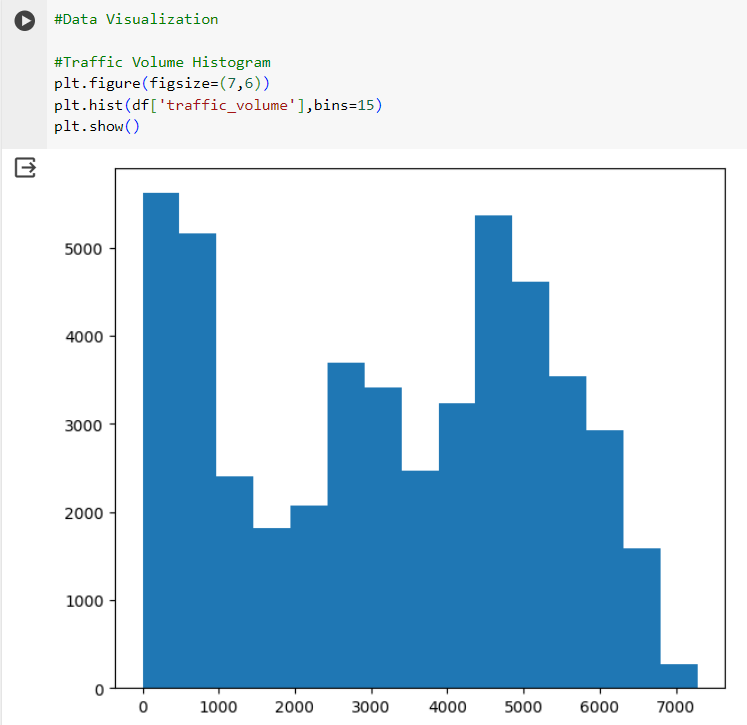


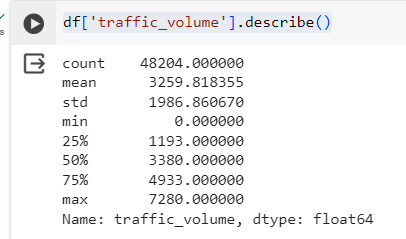
**4. DATA WRANGLING:**

* + Examined outliers in the 'traffic\_volume' column using boxplots.
  + No outliers were detected.

**5. DATA VISUALIZATION:**

* + Start by examining the distribution of the traffic\_volume column.

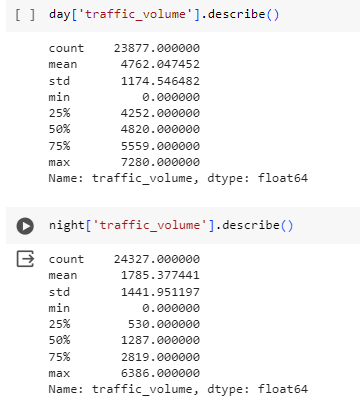




* + Between 2012-10-02 09:00:00 and 2018-09-30 23:00:00, the hourly traffic volume varied from 0 to 7,280 cars, with an average of 3,260 cars.About 25% of the time, there were only 1,193 cars or fewer passing the station each hour.
  + Let’s compare traffic volume by day and night time.

1. We'll start by dividing the dataset into two parts:

* Daytime data: hours from 7 AM to 7 PM (12 hours)
* Nighttime data: hours from 7 PM to 7 AM (12 hours)

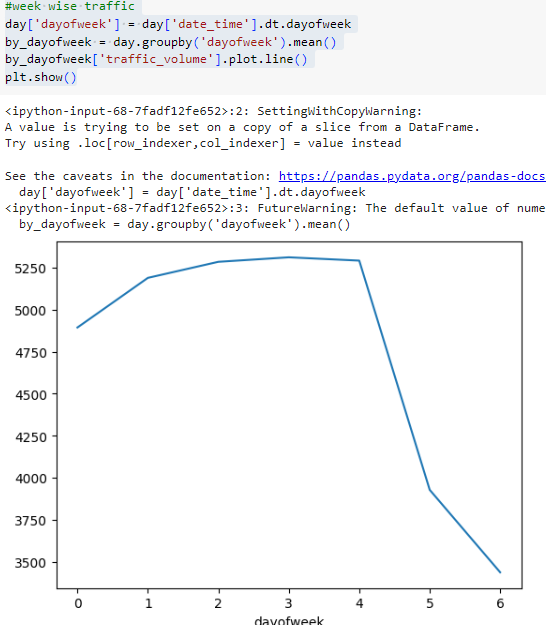


* + The histogram that shows the distribution of traffic volume during the day is left skewed. This means that most of the traffic volume values are high.There are 4,252 or more cars passing the station each hour 75% of the time (because 25% of values are less than 4,252).
  + The histogram displaying the nighttime data is right skewed. This means that most of the traffic volume values are low — 75% of the time, the number of cars that passed the station each hour was less than 2,819.
  + As our goal is to find indicators of heavy traffic, we will only focus on the daytime data moving forward.

**5.1. INDICATING TRAFFIC VOLUME THROUGH TIME:**

* One of the possible indicators of heavy traffic is time. There might be more people on the road in a certain month, on a certain day, or at a certain time of day.
* We're going to look at a few line plots showing how the traffic volume changes.

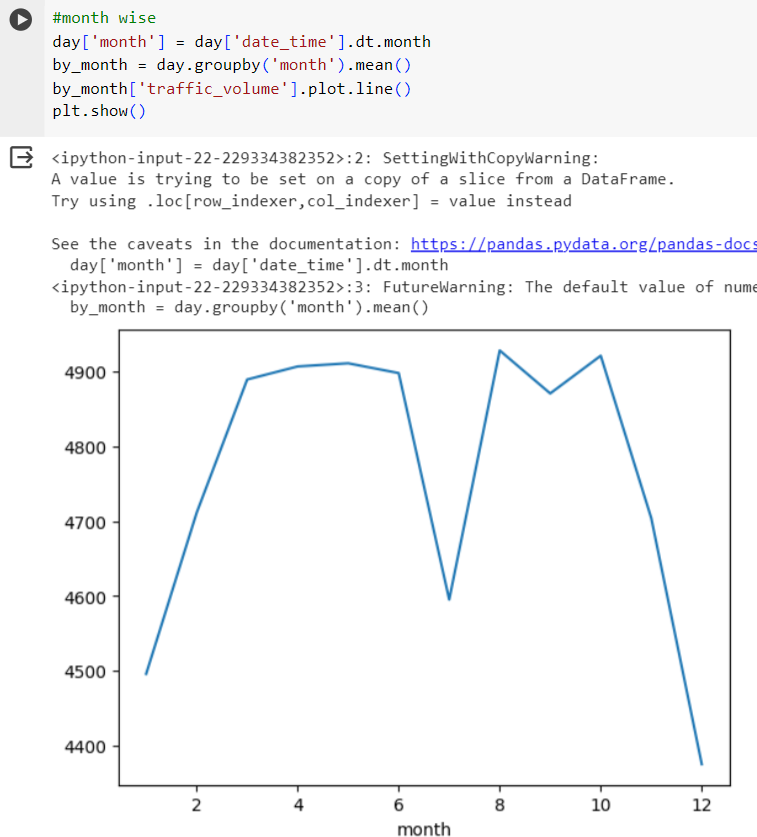
**5.1.1. DAILY TRAFFIC VOLUME**



* The analysis suggests a noticeable increase in traffic volume from Monday to Thursday, reaching the highest point on Thursday.
* Friday maintains a high average traffic volume, while weekends (Saturday and Sunday) show a substantial decrease in traffic, with Sunday having the lowest volume.

**5.1.2. HOURLY TRAFFIC VOLUME**

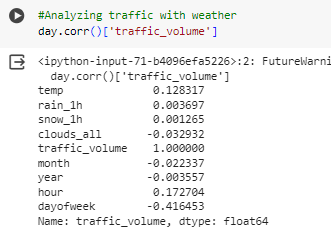
* At each hour of the day, the traffic volume is generally higher during business days compared to the weekends. As somehow expected, the rush hours are around 7 and 16 — when most people travel from home to work and back. We see volumes of over 6,000 cars at rush hours.

**5.1.3. MONTHLY TRAFFIC VOLUME**

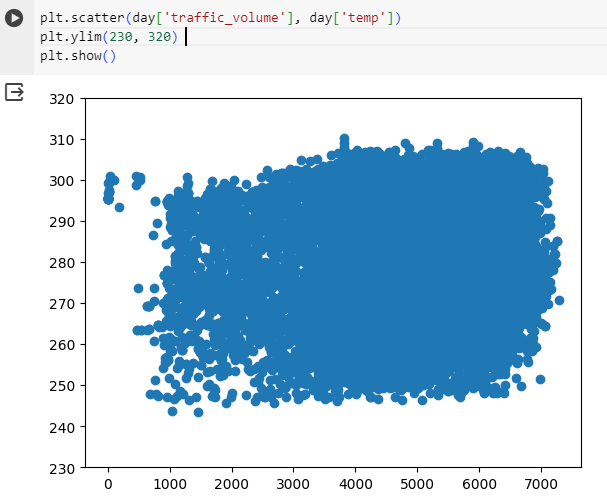
* **Winter (December to February):**
  + **December:** The mean traffic volume is approximately 4,375 cars per hour, indicating a lower volume, possibly due to the winter holiday season.
  + **January:** The mean traffic volume increases to approximately 4,496 cars per hour, suggesting a rise in traffic as the winter season progresses.
  + **February:** The mean traffic volume further increases to approximately 4,711 cars per hour, indicating continued growth and possibly a transition towards spring.
* Spring (March to May):
  + **March to May:** The mean traffic volume stabilizes or increases during these months, characteristic of the spring season when weather conditions improve, and travel activity tends to rise.
* Summer (June to August):
  + **June:** The mean traffic volume remains stable.
  + **July:** A noticeable decrease in mean traffic volume may indicate a summer holiday period with reduced commuter traffic.
  + **August:** The mean traffic volume increases, possibly indicating the end of the summer break.
* Fall (September to November):
  + **September and October:** The mean traffic volume shows stability or a slight increase during the fall season.
  + **November:** A slight decline in mean traffic volume may signify a transition toward winter.

**5.2. INDICATING TRAFFIC VOLUME THROUGH WEATHER:**

* Another possible indicator of heavy traffic is weather. The dataset provides us with a few useful columns about weather: temp, rain\_1h, snow\_1h, clouds\_all, weather\_main, weather\_description.
* A few of these columns are numerical, so let's start by looking up their correlation values with traffic\_volume.

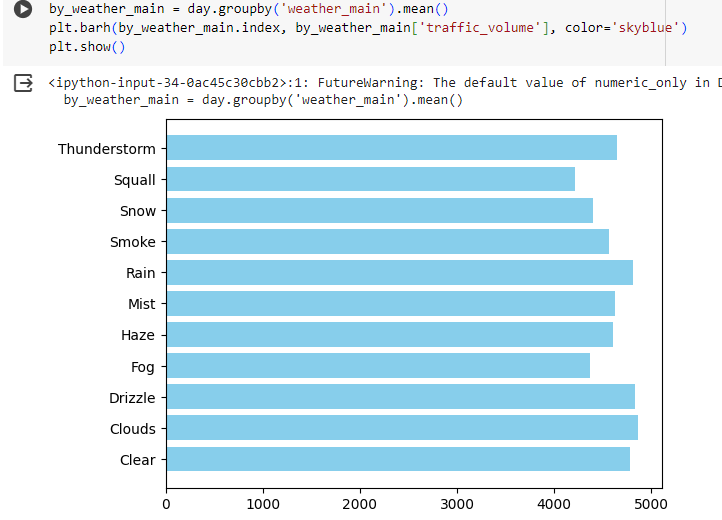


* Temperature shows the strongest correlation with a value of just +0.13.
* Let's generate a scatter plot to visualize the correlation between temp and traffic\_volume



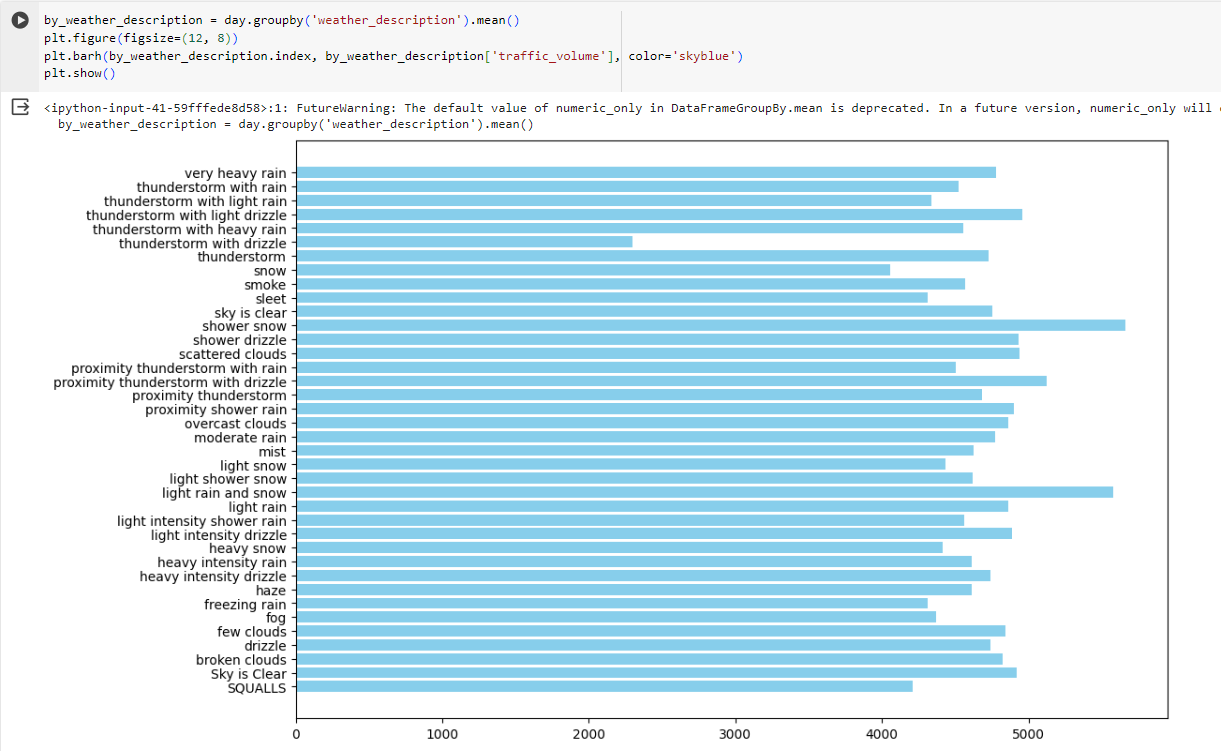
* We can clearly see that temperature doesn't look like a solid indicator of heavy traffic.
* Let's now look at the other weather-related columns: weather\_main and weather\_description.

**5.2.1. WEATHER TYPES:**

**5.2.1.1 WEATHER MAIN:**

* It looks like there's no weather type where traffic volume exceeds 5,000 cars. This makes finding a heavy traffic indicator more difficult.

**5.2.1.1 WEATHER DESCRIPTION:**

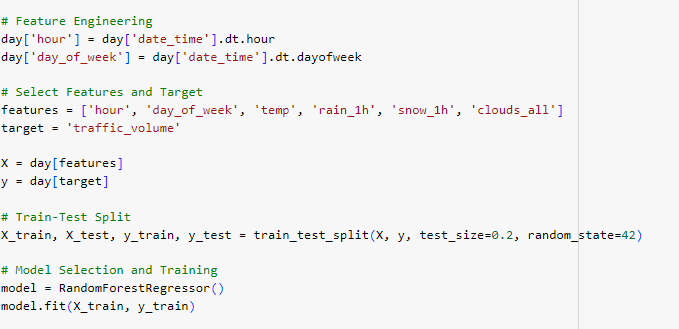


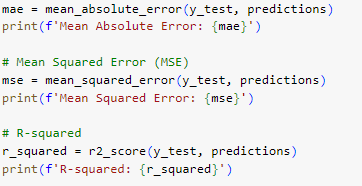
It looks like there are three weather types where traffic volume exceeds 5,000:

* Shower snow
* Light rain and snow
* Proximity thunderstorm with drizzle

**6. CHOOSE:**

Selected Random Forest as the machine learning model due to its ability to handle complex relationships.

**7. BUILD:**

**8. VALIDATE:**



**9. STORY TELLING:**

In this project, we tried to find a few indicators of heavy traffic on the I-94 Interstate highway. We managed to find two types of indicators:

* Time indicators
  + The traffic is usually heavier during warm months (March–October) compared to cold months (November–February).
  + The traffic is usually heavier on business days compared to the weekends.
  + On business days, the rush hours are around 7 and 16.
* Weather indicators
  + Shower snow
  + Light rain and snow
  + Proximity thunderstorm with drizzle