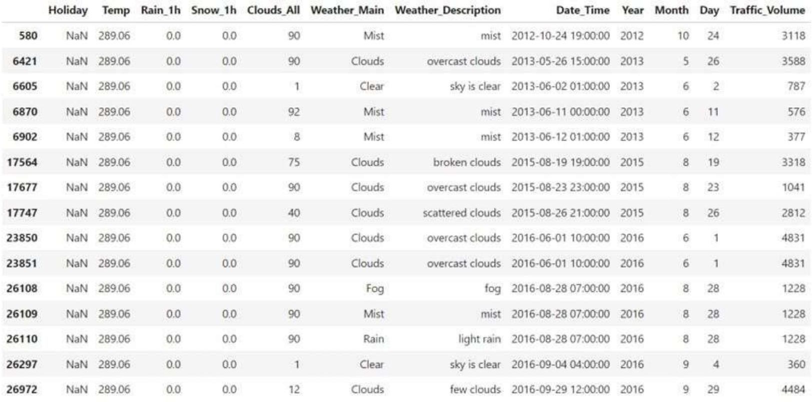
Data Collection and Preprocessing Phase

Date	20 June 2025
Team ID	LTVIP2025TMID43963
Project Title	Traffictelligence-Advanced-Traffic-Volume Estimation-With-Machine-Learning
Maximum Marks	6 Marks

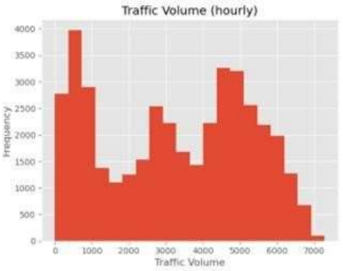
Data Exploration and Preprocessing Template

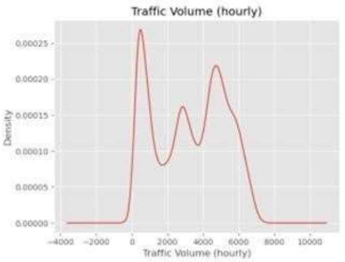
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	**Description**

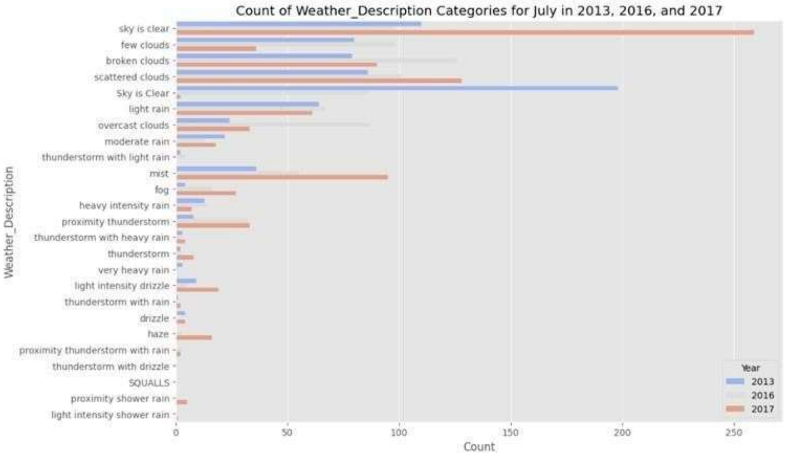
<table>
<tr><td>Data Overview</td><td>Dimension:- 40632 rows*12 columns</td></tr>
<tr><td></td><td>
<table>
<tr><th></th><th>Holiday</th><th>Temp</th><th>Rain_1h</th><th>Snow_1h</th><th>Clouds_All</th><th>Weather_Main</th><th>Weather_Description</th><th>Date_Time</th><th>Year</th><th>Month</th><th>Day</th><th>Traffic_Volume</th></tr>
<tr><td>580</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>90</td><td>Mist</td><td>mist</td><td>2012-10-24 19:00:00</td><td>2012</td><td>10</td><td>24</td><td>3118</td></tr>
<tr><td>6421</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>90</td><td>Clouds</td><td>overcast clouds</td><td>2013-05-26 15:00:00</td><td>2013</td><td>5</td><td>26</td><td>3588</td></tr>
<tr><td>6605</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>1</td><td>Clear</td><td>sky is clear</td><td>2013-06-02 01:00:00</td><td>2013</td><td>6</td><td>2</td><td>787</td></tr>
<tr><td>6870</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>92</td><td>Mist</td><td>mist</td><td>2013-06-11 00:00:00</td><td>2013</td><td>6</td><td>11</td><td>576</td></tr>
<tr><td>6902</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>8</td><td>Mist</td><td>mist</td><td>2013-06-12 01:00:00</td><td>2013</td><td>6</td><td>12</td><td>377</td></tr>
<tr><td>17564</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>75</td><td>Clouds</td><td>broken clouds</td><td>2015-08-19 19:00:00</td><td>2015</td><td>8</td><td>19</td><td>3318</td></tr>
<tr><td>17677</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>90</td><td>Clouds</td><td>overcast clouds</td><td>2015-08-23 23:00:00</td><td>2015</td><td>8</td><td>23</td><td>1041</td></tr>
<tr><td>17747</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>40</td><td>Clouds</td><td>scattered clouds</td><td>2015-08-26 21:00:00</td><td>2015</td><td>8</td><td>26</td><td>2812</td></tr>
<tr><td>23850</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>90</td><td>Clouds</td><td>overcast clouds</td><td>2016-06-01 10:00:00</td><td>2016</td><td>6</td><td>1</td><td>4831</td></tr>
<tr><td>23851</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>90</td><td>Clouds</td><td>overcast clouds</td><td>2016-06-01 10:00:00</td><td>2016</td><td>6</td><td>1</td><td>4831</td></tr>
<tr><td>26108</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>90</td><td>Fog</td><td>fog</td><td>2016-08-28 07:00:00</td><td>2016</td><td>8</td><td>28</td><td>1228</td></tr>
<tr><td>26109</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>90</td><td>Mist</td><td>mist</td><td>2016-08-28 07:00:00</td><td>2016</td><td>8</td><td>28</td><td>1228</td></tr>
<tr><td>26110</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>90</td><td>Rain</td><td>light rain</td><td>2016-08-28 07:00:00</td><td>2016</td><td>8</td><td>28</td><td>1228</td></tr>
<tr><td>26297</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>1</td><td>Clear</td><td>sky is clear</td><td>2016-09-04 04:00:00</td><td>2016</td><td>9</td><td>4</td><td>360</td></tr>
<tr><td>26972</td><td>NaN</td><td>289.06</td><td>0.0</td><td>0.0</td><td>12</td><td>Clouds</td><td>few clouds</td><td>2016-09-29 12:00:00</td><td>2016</td><td>9</td><td>29</td><td>4484</td></tr>
</table>
Descriptive Statistics:-</td></tr>
</table>

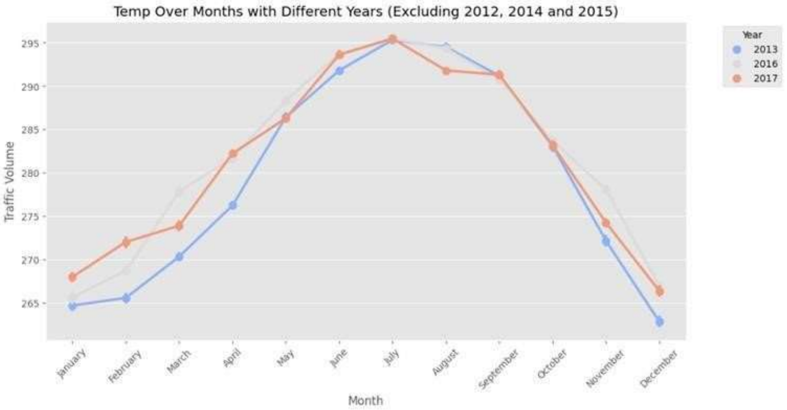
Univariate Analysis

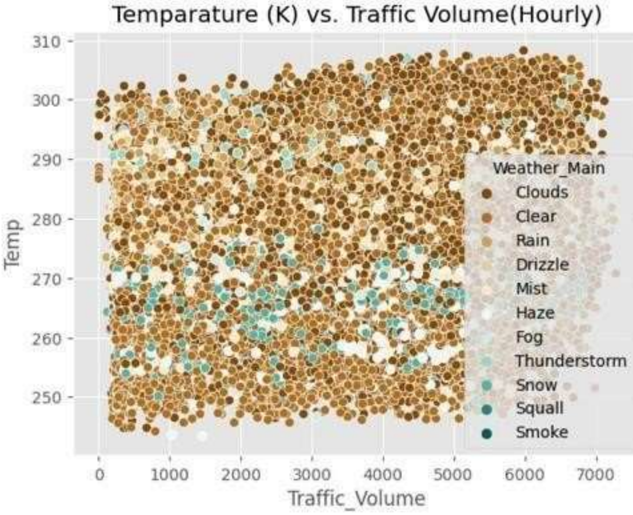


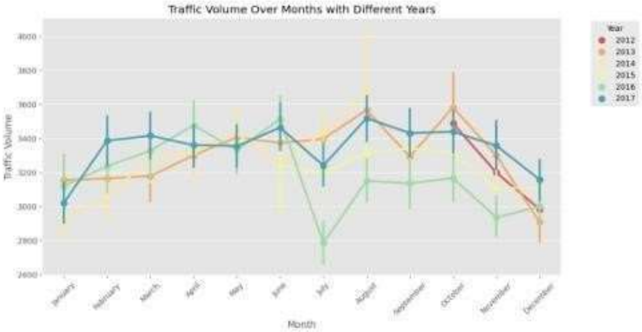


Bivariate Analysis

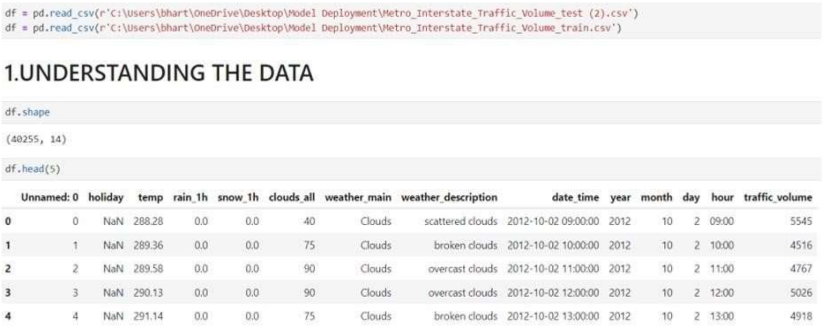




Multivariate Analysis	

	
Outliers and Anomalies	extreme weather, special events, accidents, data errors, or unusual traffic patterns
Data Preprocessing Code Sc reenshots	

Loading Data



```
df = pd.read_csv(r'C:\Users\bhart\OneDrive\Desktop\Model Deployment\Metro_Interstate_Traffic_Volume_test (2).csv')
df = pd.read_csv(r'C:\Users\bhart\OneDrive\Desktop\Model Deployment\Metro_Interstate_Traffic_Volume_train.csv')
```

1.UNDERSTANDING THE DATA

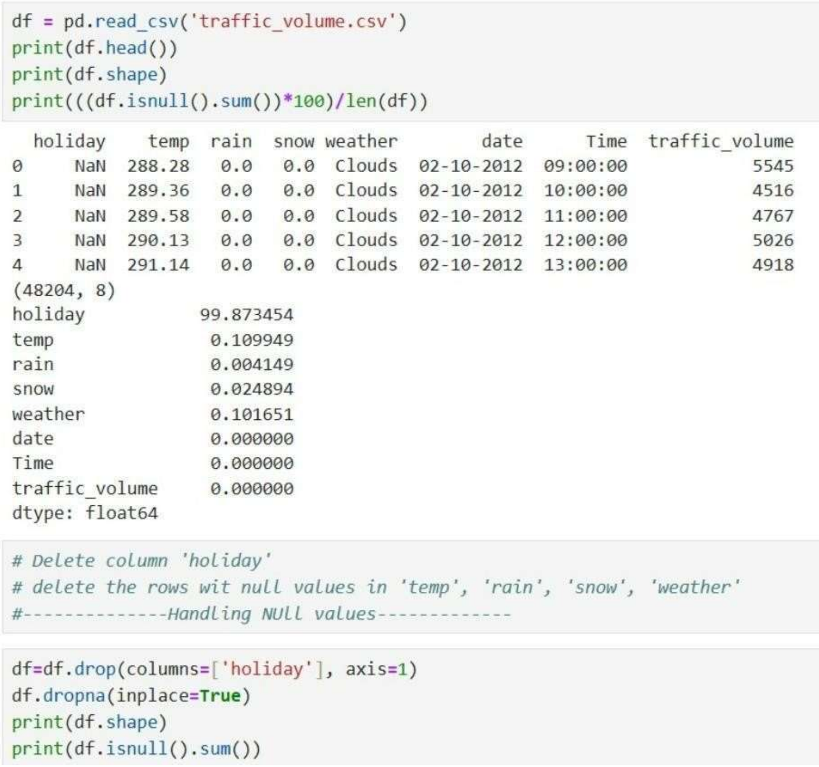
```
df.shape
```

```
(40255, 14)
```

```
df.head(5)
```

	Unnamed: 0	holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_time	year	month	day	hour	traffic_volume
0	0	NaN	288.28	0.0	0.0	40	Clouds	scattered clouds	2012-10-02 09:00:00	2012	10	2	09:00	5545
1	1	NaN	289.36	0.0	0.0	75	Clouds	broken clouds	2012-10-02 10:00:00	2012	10	2	10:00	4516
2	2	NaN	289.58	0.0	0.0	90	Clouds	overcast clouds	2012-10-02 11:00:00	2012	10	2	11:00	4767
3	3	NaN	290.13	0.0	0.0	90	Clouds	overcast clouds	2012-10-02 12:00:00	2012	10	2	12:00	5026
4	4	NaN	291.14	0.0	0.0	75	Clouds	broken clouds	2012-10-02 13:00:00	2012	10	2	13:00	4918

Handling Missing Data



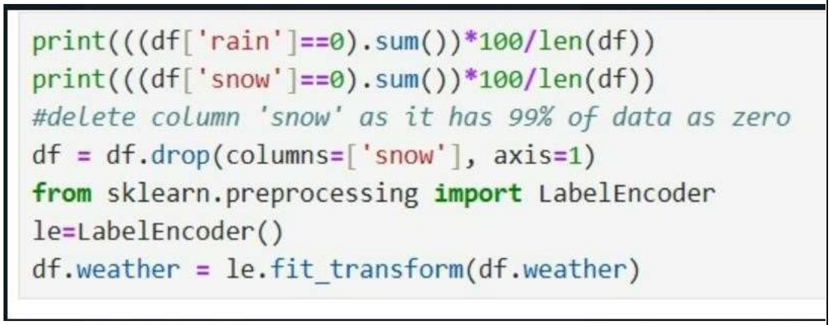
```
df = pd.read_csv('traffic_volume.csv')
print(df.head())
print(df.shape)
print(((df.isnull().sum())*100)/len(df))
```

```
  holiday    temp  rain  snow weather        date      Time  traffic_volume
0     NaN  288.28   0.0   0.0  Clouds  02-10-2012  09:00:00            5545
1     NaN  289.36   0.0   0.0  Clouds  02-10-2012  10:00:00            4516
2     NaN  289.58   0.0   0.0  Clouds  02-10-2012  11:00:00            4767
3     NaN  290.13   0.0   0.0  Clouds  02-10-2012  12:00:00            5026
4     NaN  291.14   0.0   0.0  Clouds  02-10-2012  13:00:00            4918
(48204, 8)
holiday           99.873454
temp               0.109949
rain               0.004149
snow               0.024894
weather            0.101651
date               0.000000
Time               0.000000
traffic_volume     0.000000
dtype: float64
```

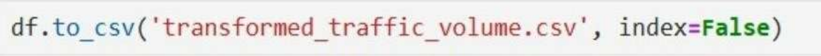
```
# Delete column 'holiday'
# delete the rows wit null values in 'temp', 'rain', 'snow', 'weather'
#-------------Handling NUll values-------------
```

```
df=df.drop(columns=['holiday'], axis=1)
df.dropna(inplace=True)
print(df.shape)
print(df.isnull().sum())
```

Data Transformation



```
print(((df['rain']==0).sum())*100/len(df))
print(((df['snow']==0).sum())*100/len(df))
#delete column 'snow' as it has 99% of data as zero
df = df.drop(columns=['snow'], axis=1)
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
df.weather = le.fit_transform(df.weather)
```

Feature Engineering	Attached the codes in final Submission
Save Processed Data	 `df.to_csv('transformed_traffic_volume.csv', index=False)`