Data Collection and Preprocessing Phase

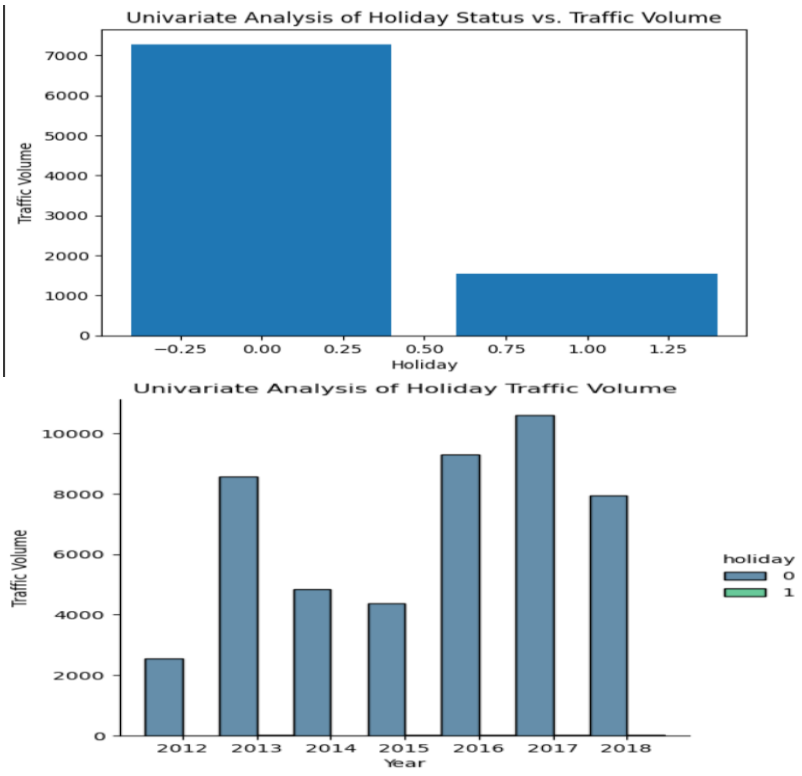
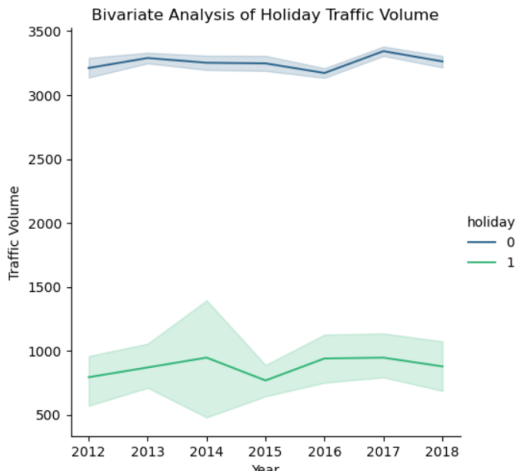
Date	20 JULY 2024
Team ID	SWTID1720014187
Project Title	Traffic Telligence: Advanced Traffic Volume Estimation With Machine
Maximum Marks	6 Marks

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
Data Overview	**Dimension:** 48204 rows × 8 columns **Descriptive statistics:**

	temp	rain	snow	traffic_volume
count	48151.000000	48202.000000	48192.000000	48204.000000
mean	281.205351	0.334278	0.000222	3259.818355
std	13.343675	44.790062	0.008169	1986.860670
min	0.000000	0.000000	0.000000	0.000000
25%	272.160000	0.000000	0.000000	1193.000000
50%	282.460000	0.000000	0.000000	3380.000000
75%	291.810000	0.000000	0.000000	4933.000000
max	310.070000	9831.300000	0.510000	7280.000000

Univariate Analysis	
Bivariate Analysis	
Multivariate Analysis	--
Outliers and Anomalies	--
Data Preprocessing Code Screenshots	

Loading Data

	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

Handling Missing Data

```
#fill the missing cells with the mean of the whole column
data['temp']=data['temp'].fillna(data['temp'].mean())
data['rain']=data['rain'].fillna(data['rain'].mean())
data['snow']=data['snow'].fillna(data['snow'].mean())
data['weather'].fillna('Clouds',inplace=True)
data['weather'].fillna('NaN',inplace=True)
```

Data Transformation

```
holiday_list = ['Labor Day', 'Thanksgiving Day', 'Christmas Day', 'New Years Day',
                'Martin Luther King Jr Day', 'Columbus Day', 'Veterans Day',
               'Memorial Day', 'Independence Day', 'State Fair']

data['holiday'] = data['holiday'].apply(lambda x: '1' if x in holiday_list else '0')

data[["day","month","year"]] = data["date"].str.split("-", expand=True)
data[["hours","minutes","seconds"]] = data["Time"].str.split(":", expand=True)
data.drop(columns=['date','Time'],axis=1,inplace=True)
```

Feature Engineering

```
from sklearn.preprocessing import LabelEncoder

# Assuming your data is in a DataFrame called 'data'
# Assuming the weather column is named 'weather'

# Create a LabelEncoder object
le = LabelEncoder()

# Fit the LabelEncoder to the weather data (learn the categories)
le.fit(data['weather'])

# Transform the 'weather' column to numerical labels
data['weather'] = le.transform(data['weather'])
```

Save Processed Data

	holiday	temp	rain	snow	weather	traffic_volume	day	month	year	hours	minutes	seconds
39346	0	277.44	0.0	0.0	0	2859	29	11	2017	20	00	00
23628	0	296.46	0.0	0.0	1	4603	25	05	2016	18	00	00
6563	0	294.84	0.0	0.0	1	5635	31	05	2013	13	00	00
44041	0	279.69	0.0	0.0	0	622	13	05	2018	02	00	00
43918	0	290.46	0.0	0.0	6	3274	08	05	2018	19	00	00