1. Introduction

The Metro Interstate Traffic Volume Data Set, sourced from the UCI Machine Learning Repository (UCI, 2019), comprises 48,204 hourly traffic instances spanning nine attributes, including holiday type, temperature, rainfall, snowfall, cloud cover, weather descriptions, date-time information, and traffic volume. The data spans from October 2012 to September 2018, covering the region between Minneapolis and St. Paul, MN. Of particular concern in Minneapolis is the alarming rise in highway accidents, despite a 5% decrease in average daily traffic within the city (Lee, 2019). This report aims to investigate whether traffic volume is genuinely increasing. For the purposes of this study, detailed weather and seasonal conditions are omitted, focusing solely on traffic volume. The data is aggregated into monthly traffic volume, disregarding hourly fluctuations to assess the overall volume increase.

Although the original dataset contains a substantial 48,204 instances, the aggregation to monthly volume reduces the dataset to a mere 63 instances.

2. Objective

The project aims to understand how the variables in “Metro Interstate Traffic Volume Data Set " affect the traffic volume. For this, we choose “traffic_volume” as Dependent Variable and all the other variables except “traffic_volume” as Independent variables.

3. Gathering Data

Since our focus lies solely on the monthly traffic volume, we isolate the date_time column from the dataset and compute the monthly volume. Subsequently, the data is processed and applied to various models to determine the most accurate prediction for the upcoming two years.

	holiday	temp	rain_1h	snow_1h	clouds_all	weather_main	weather_description	date_time	traffic_volume
1	None	288.28	0	0	40	Clouds	scattered clouds	2012-10-02 09:00:00	5545
2	None	289.36	0	0	75	Clouds	broken clouds	2012-10-02 10:00:00	4516
3	None	289.58	0	0	90	Clouds	overcast clouds	2012-10-02 11:00:00	4767
4	None	290.13	0	0	90	Clouds	overcast clouds	2012-10-02 12:00:00	5026
5	None	291.14	0	0	75	Clouds	broken clouds	2012-10-02 13:00:00	4918
6	None	291.72	0	0	1	Clear	sky is clear	2012-10-02 14:00:00	5181

The initial dataset was structured as follows.

3.1. Data Cleaning

We need to validate and verify the data before further processing. Our dataset doesn't contain any NA values. The date_time column was divided into separate date and time components. Following this, the date field was further segmented into year, month, and day. The traffic flow data was then aggregated on a monthly basis.

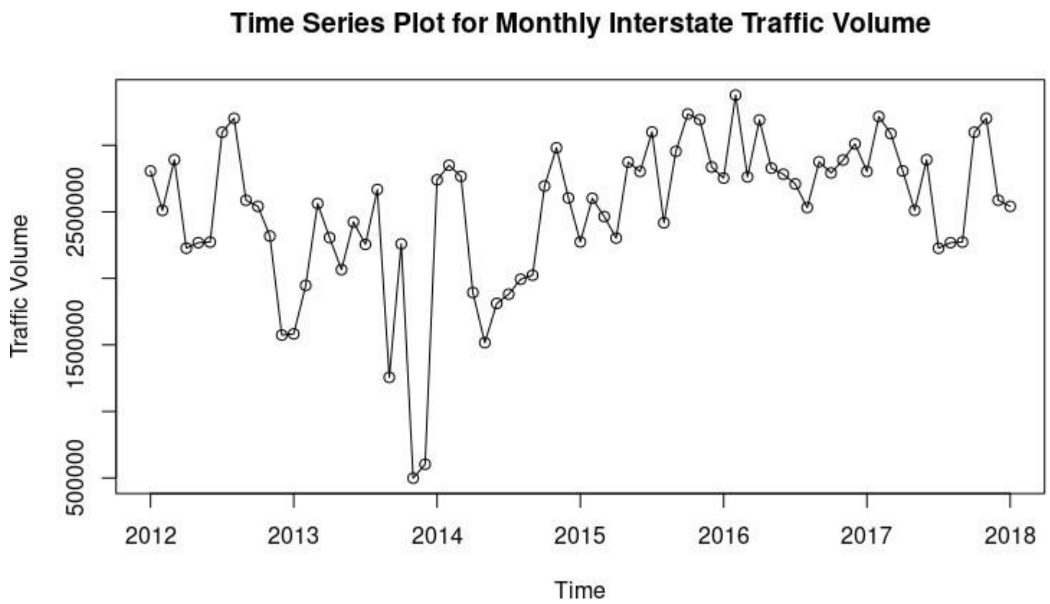
Aggregated new data looks like below:

	year	month	volume_by_month
1	2012	10	2806826
2	2012	11	2510769
3	2012	12	2891172
4	2013	01	2226480
5	2013	02	2268026
6	2013	03	2272416
7	2013	04	3097942
8	2013	05	3202233
9	2013	06	2586575
10	2013	07	2541007

4. Descriptive Analysis

Time Series Plot

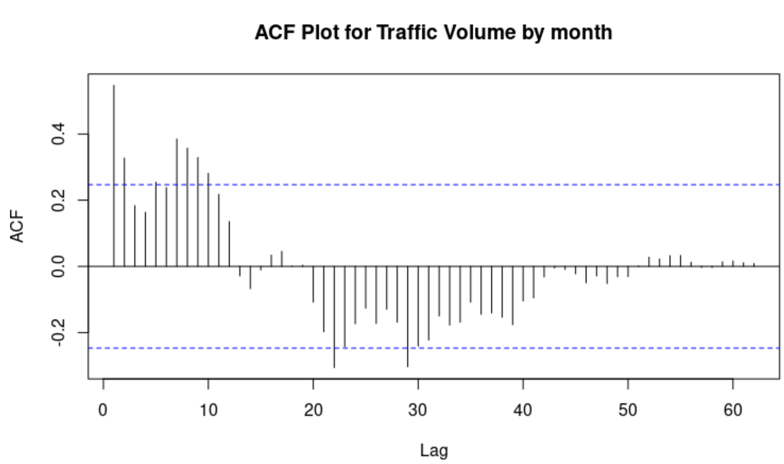
A time series plot serves as a visual representation of data points collected or recorded at specific time intervals, arranged in chronological order. It is employed to observe patterns, trends, and fluctuations within the data over time. Time series plots are especially valuable for examining time-related data such as stock prices, weather conditions, sales figures, and other variables dependent on time.



The data displays recurring seasonal patterns, and a significant decline in traffic volume is evident towards the end of 2014, even during the holiday season. The overall trend does not distinctly indicate whether there is a consistent increase in traffic volume. Subsequently, we investigated whether the traffic volume of the previous month has an impact on the traffic volume of the subsequent month.

ACF Plot

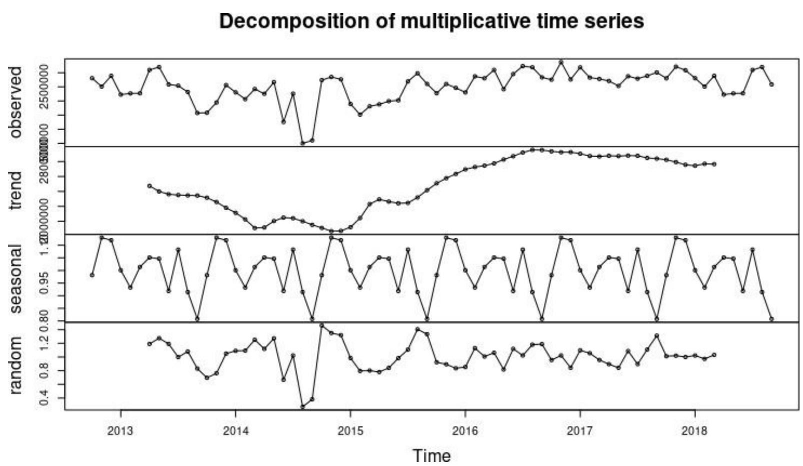
The ACF plot is a specialized tool designed to unveil the autocorrelation or self-similarity within traffic volume data at various time lags. Essentially, it allows us to explore the potential repetition of periodic patterns, which might indicate the presence of seasonality in the data. If our time series exhibits a strong seasonal pattern, we might observe significant autocorrelation values at specific lags corresponding to the season's length.

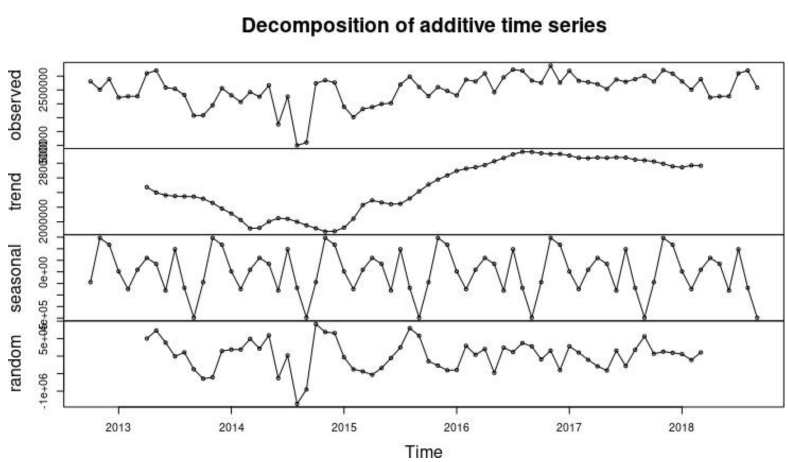


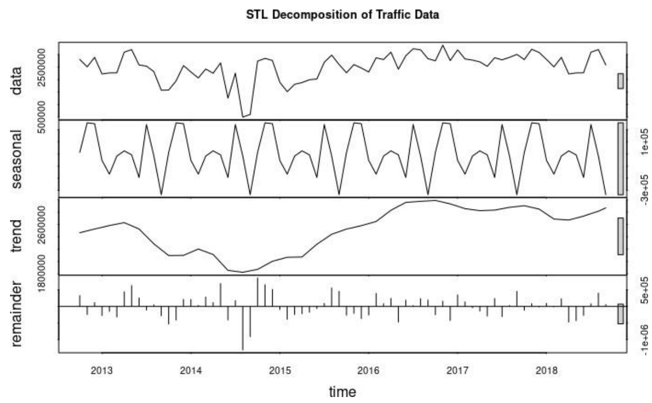
There is a weak positive correlation between the previous month's traffic volume to the next month.Only few values are statistically significant.

5. Time Series Decomposition

Time Series Decomposition involves breaking down time-based data into components like trend, seasonality, and noise. Analyzing these elements enables predictions and insights. Methods like additive,multiplicative,STL models are crucial for accurate forecasts and strategic decisions.



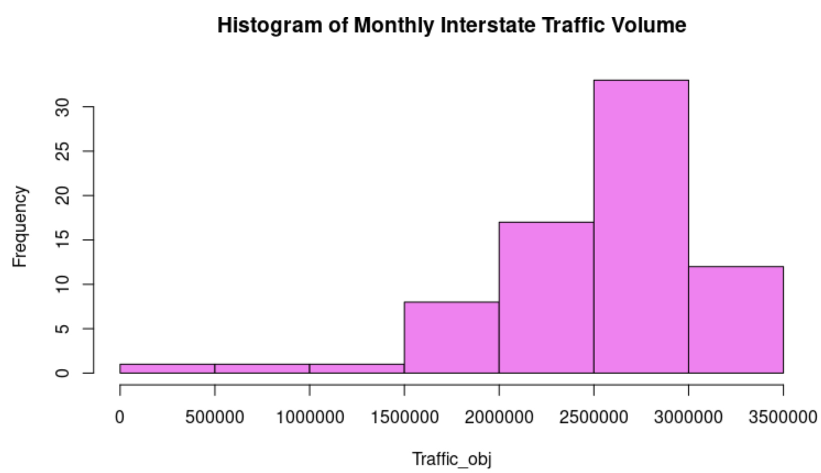




Consistent findings from additive,multiplicative and STL decomposition methods indicate stable and predictable traffic patterns. These reliable trends, like increased holiday traffic, allow accurate short-term predictions and enhance forecasting models. In essence, stable patterns improve forecasting reliability and inform practical decision-making in traffic management.

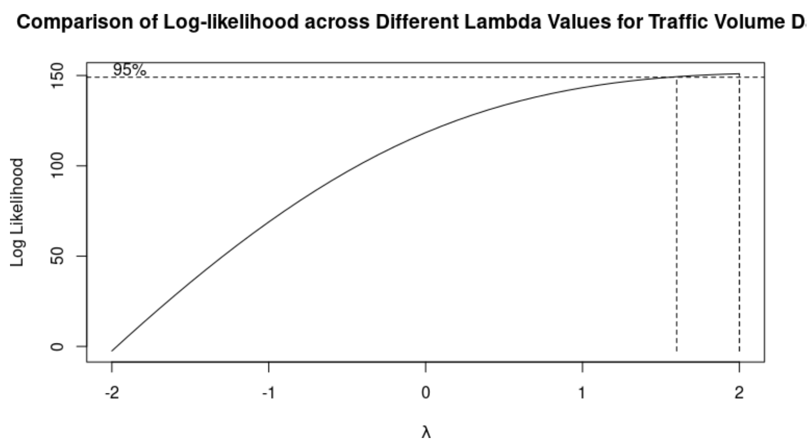
6. Transformations

We assess the data's shape before any transformations to determine the optimal normalization method.

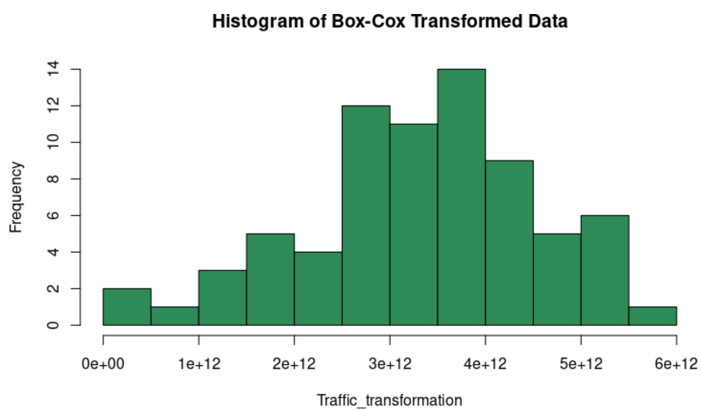


The above histogram indicates that the data doesn't follow a normal distribution. In these situations, you could explore various transformation methods like logarithmic, square root, or Box-Cox transformations to make the data align more closely with a normal distribution.

Box-cox transformation is applied to the series to help make the series stationary.



After applying Box-Cox the histogram looks closer to the normal distribution



7. Forecasting Evaluations

Drift Forecasting

Drift forecasting includes a linear trend, allowing gradual increases or decreases over time. It suits data with consistent trends, offering more accurate predictions than naive methods.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Oct 2018	2792896	2139057	3446734	1792936	3792855
Nov 2018	2806583	2128014	3485153	1768801	3844366
Dec 2018	2820271	2117824	3522718	1745971	3894571
Jan 2019	2833959	2108404	3559515	1724318	3943601
Feb 2019	2847647	2099681	3595613	1703731	3991563
Mar 2019	2861335	2091595	3631075	1684119	4038551
Apr 2019	2875023	2084093	3665953	1665399	4084646
May 2019	2888711	2077129	3700292	1647504	4129918
Jun 2019	2902399	2070664	3734133	1630370	4174427
Jul 2019	2916086	2064661	3767511	1613944	4218228
Aug 2019	2929774	2059091	3800458	1598179	4261370
Sep 2019	2943462	2053923	3833001	1583030	4303894
Oct 2019	2957150	2049134	3865166	1568460	4345840
Nov 2019	2970838	2044701	3896975	1554434	4387242
Dec 2019	2984526	2040603	3928449	1540920	4428131
Jan 2020	2998214	2036821	3959606	1527891	4468536
Feb 2020	3011901	2033339	3990464	1515320	4508483
Mar 2020	3025589	2030142	4021037	1503183	4547995
Apr 2020	3039277	2027214	4051341	1491460	4587095
May 2020	3052965	2024542	4081388	1480128	4625802
Jun 2020	3066653	2022116	4111190	1469172	4664134
Jul 2020	3080341	2019923	4140758	1458572	4702109
Aug 2020	3094029	2017954	4170103	1448314	4739743
Sep 2020	3107716	2016198	4199235	1438383	4777050

Above is the Forecasted Traffic volume for the years 2019 and 2020 using Drift Forecasting.

Naïve Forecasting

Naive forecasting predicts future values using the latest observation, assuming the next value will be similar.

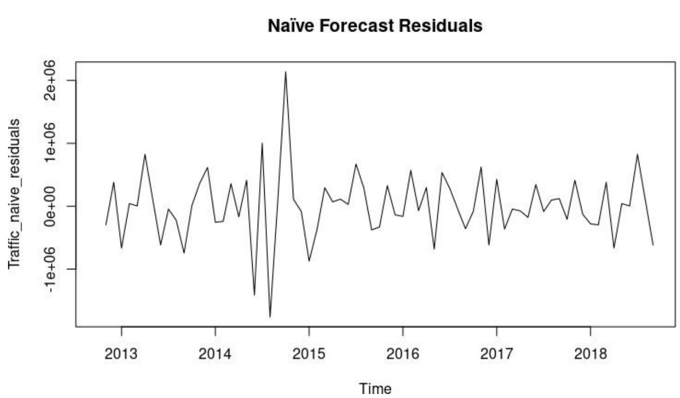
It's straightforward but not ideal for complex data patterns.

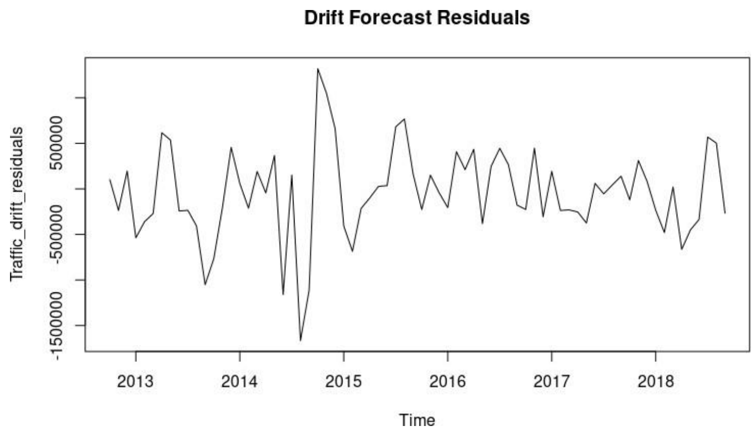
	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Oct 2018	2586575	1895127.8774	3278022	1529097.881	3644052
Nov 2018	2586575	1608721.1016	3564429	1091076.516	4082073
Dec 2018	2586575	1388953.4530	3784197	754970.902	4418179
Jan 2019	2586575	1203680.7549	3969469	471620.762	4701529
Feb 2019	2586575	1040452.2311	4132698	221984.277	4951166
Mar 2019	2586575	892882.3656	4280268	-3704.356	5176854
Apr 2019	2586575	757177.8690	4415972	-211246.474	5384396
May 2019	2586575	630867.2032	4542283	-404421.967	5577572
Jun 2019	2586575	512233.6323	4660916	-585856.357	5759006
Jul 2019	2586575	400027.2112	4773123	-757461.270	5930611
Aug 2019	2586575	293304.3321	4879846	-920679.828	6093830
Sep 2019	2586575	191331.9060	4981818	-1076633.196	6249783
Oct 2019	2586575	93526.9454	5079623	-1226212.975	6399363
Nov 2019	2586575	-583.2337	5173733	-1370142.074	6543292
Dec 2019	2586575	-91388.1904	5264538	-1509016.271	6682166
Jan 2020	2586575	-179213.4902	5352363	-1643333.476	6816483
Feb 2020	2586575	-264334.5208	5437485	-1773514.858	6946665
Mar 2020	2586575	-346986.6951	5520137	-1899920.451	7073070
Apr 2020	2586575	-427373.1320	5600523	-2022860.897	7196011
May 2020	2586575	-505670.5378	5678821	-2142606.446	7315756
Jun 2020	2586575	-582033.7782	5755184	-2259393.944	7432544
Jul 2020	2586575	-656599.4807	5829749	-2373432.345	7546582
Aug 2020	2586575	-729488.9071	5902639	-2484907.103	7658057
Sep 2020	2586575	-800810.2688	5973960	-2593983.713	7767134

Above is the Forecasted Traffic volume for the years 2019 and 2020 using Naïve Forecasting. We

check the model correctness using several models:I have used **MAE** and **Residuals** here.

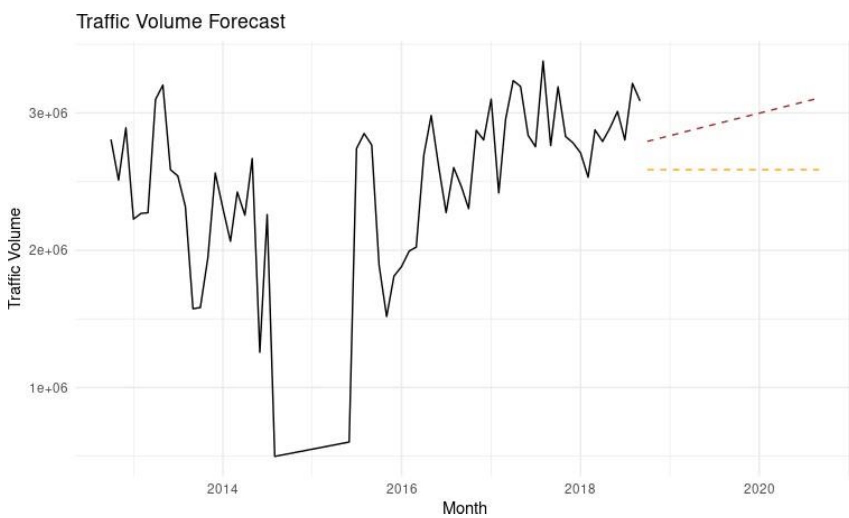
```
> # Print MAE values
> print(paste("MAE for Drift Forecasting:", MAE_drift))
[1] "MAE for Drift Forecasting: 182698.280574598"
> print(paste("MAE for Naive Forecasting:", MAE_naive))
[1] "MAE for Naive Forecasting: 181032.769230769"
```





The model’s standardized residual plots shows no trend nor changing variance meaning that the both of the models are supported.

Plot the Forecast



8. Conclusion

In the long term, the traffic volume appears relatively stable, showing neither a significant increase nor decrease despite seasonal fluctuations. However, the accuracy of predictions might be affected by the small dataset size, leading to unexpected results. Additionally, translating predictions from Box-Cox transformed values back to the original data is challenging. Using a larger dataset and assuming a normal distribution could mitigate these issues. Notably, a substantial drop in traffic volume at the end of 2014 influences the overall trend. Removing the data from that year might result in a more accurate model and precise predictions.