Understanding Traffic

Anvita Nagireddi, Jenna Odom, Rankin Odister, Vaish Pulla

AGENDA

- Overview of the dataset
- Exploratory analysis
- Evaluation metrics
- K Nearest Neighbors
- Multiple Regression Models
- Arima Model

Traffic Dataset

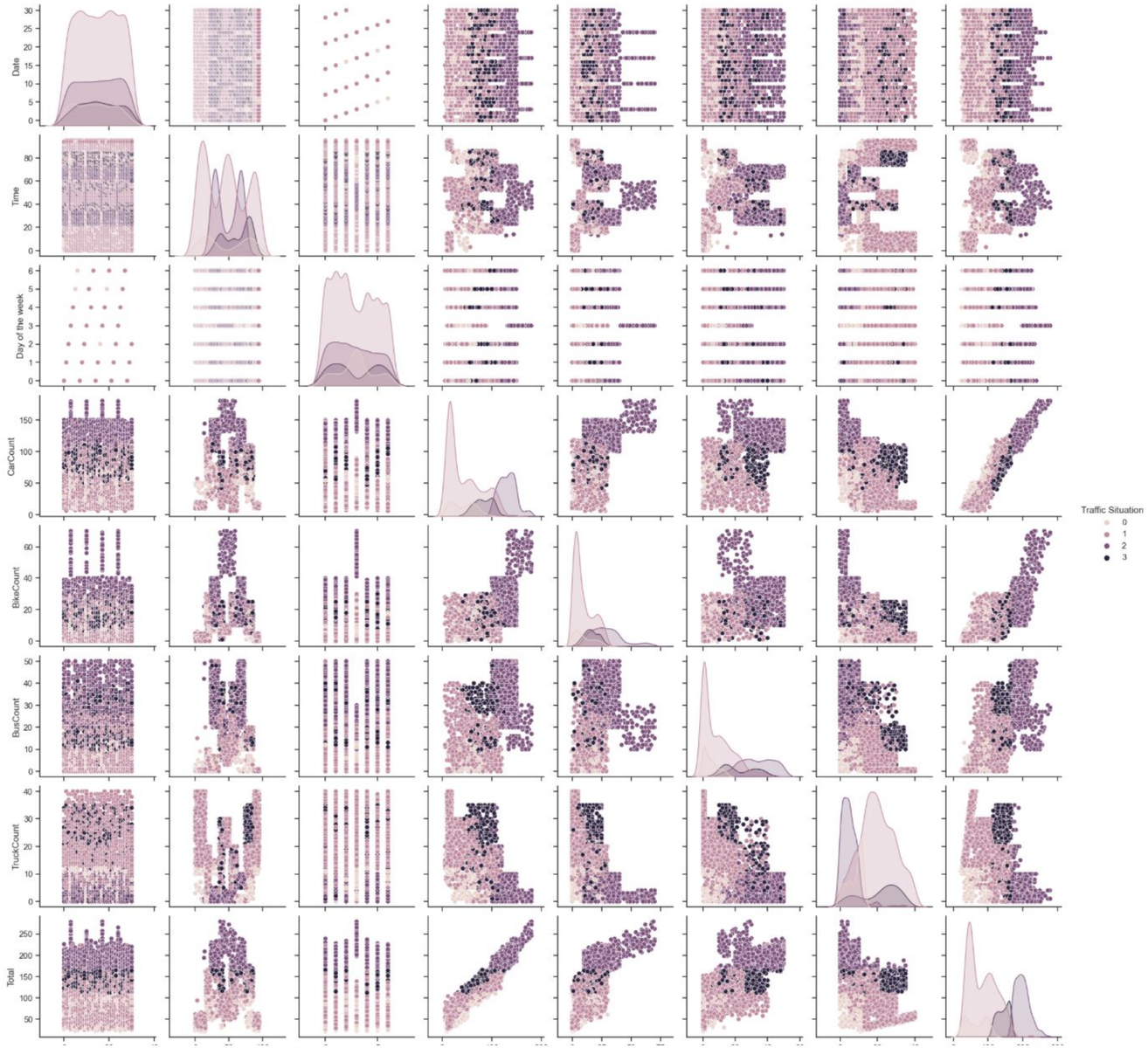
- Traffic Prediction Data Set
- Columns:

Time	Date	Day of the week	CarCount	BikeCount	BusCount	TruckCount	Total	Traffic Situation

- Potential Uses: transportation planning, congestion management, and traffic flow analysis.
- Important Notes: Traffic often changes based on the time of day on a specific day of the week. Thus, for this reason some form of categorization of the data may be beneficial.
- Kaggle description:

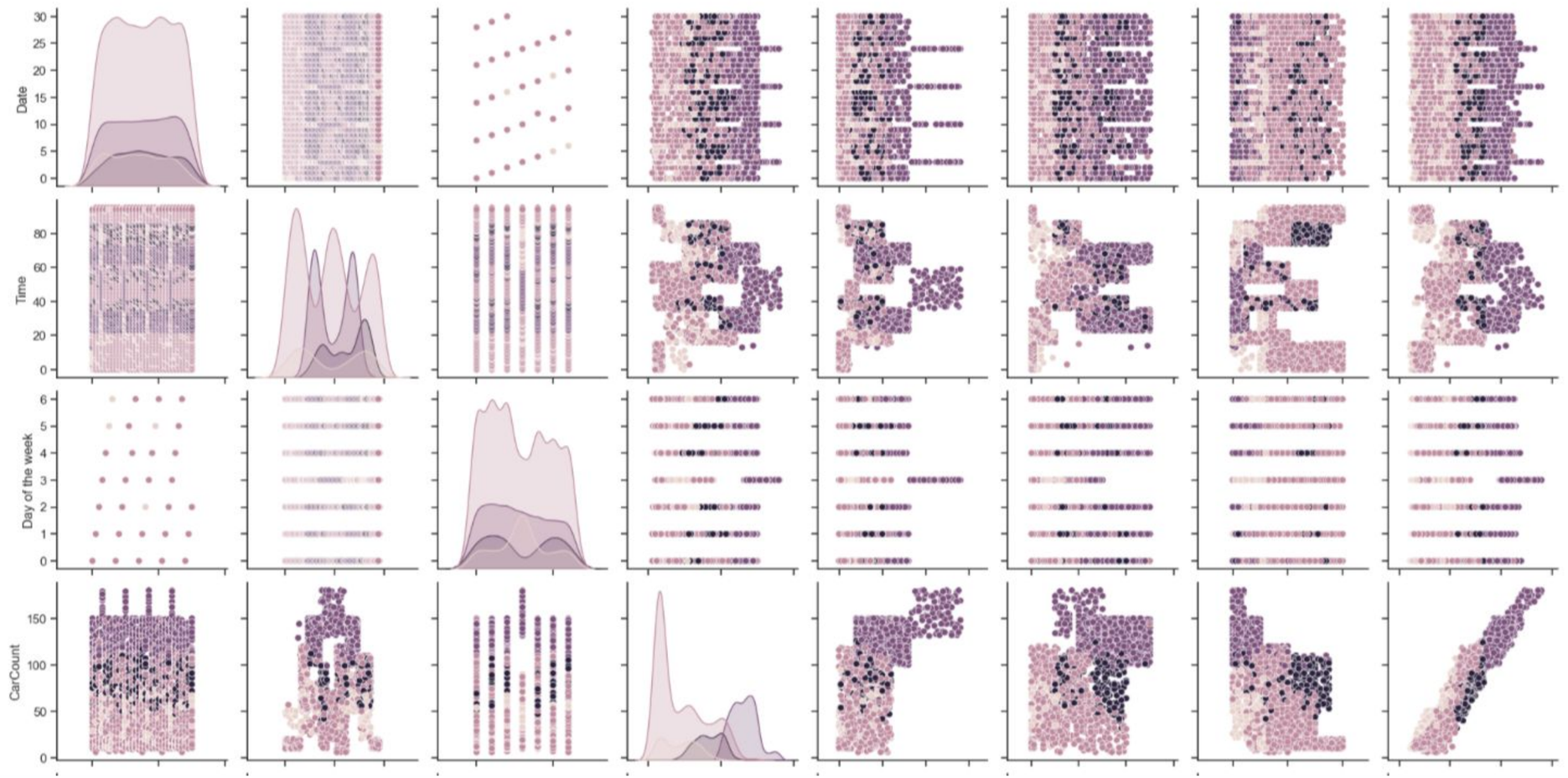
Exploratory Data Analysis

- Covariance
- Correlation
- Mean
- Median
- Standard deviation

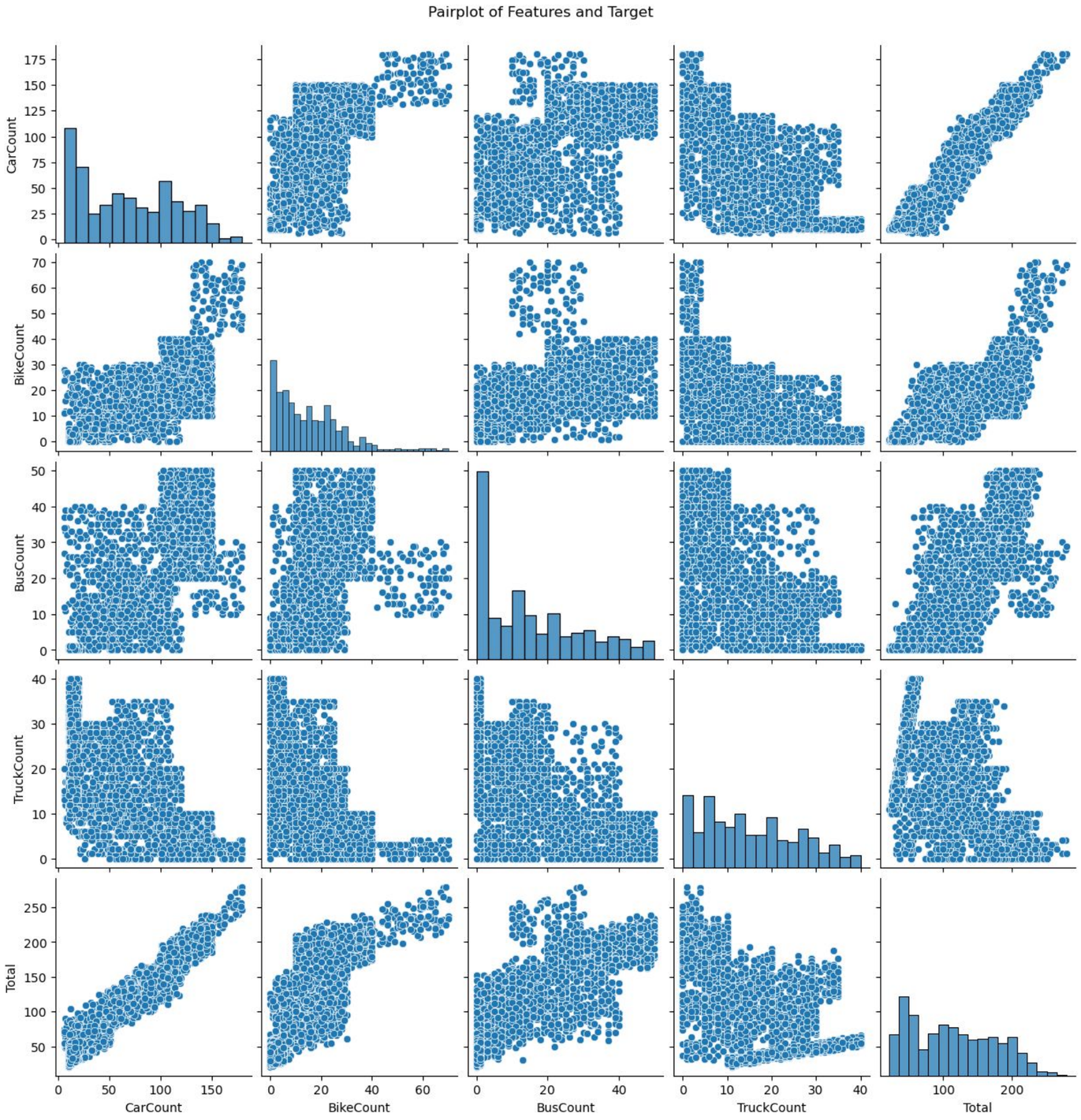


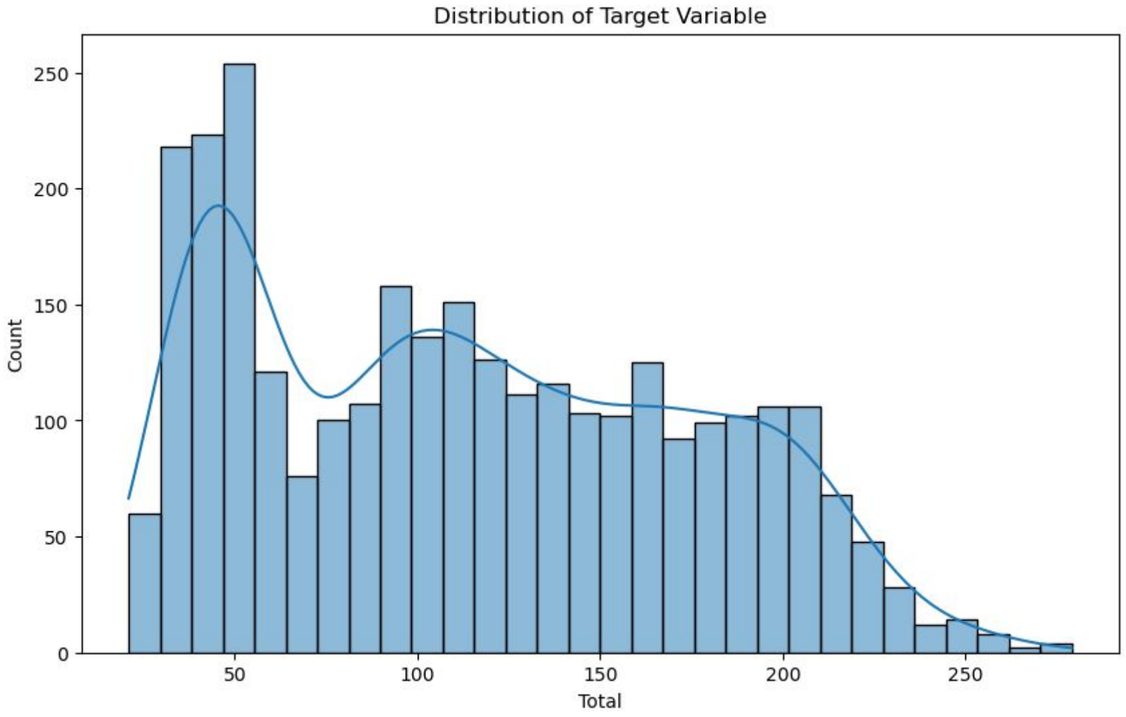
Exploratory Data Analysis

- Correlation (Continued)
- Time variables
- Clear Correlation between variables that are related to time with regards to how it impacts traffic situation.
- Indicates that the time feature may be of importance considering different models



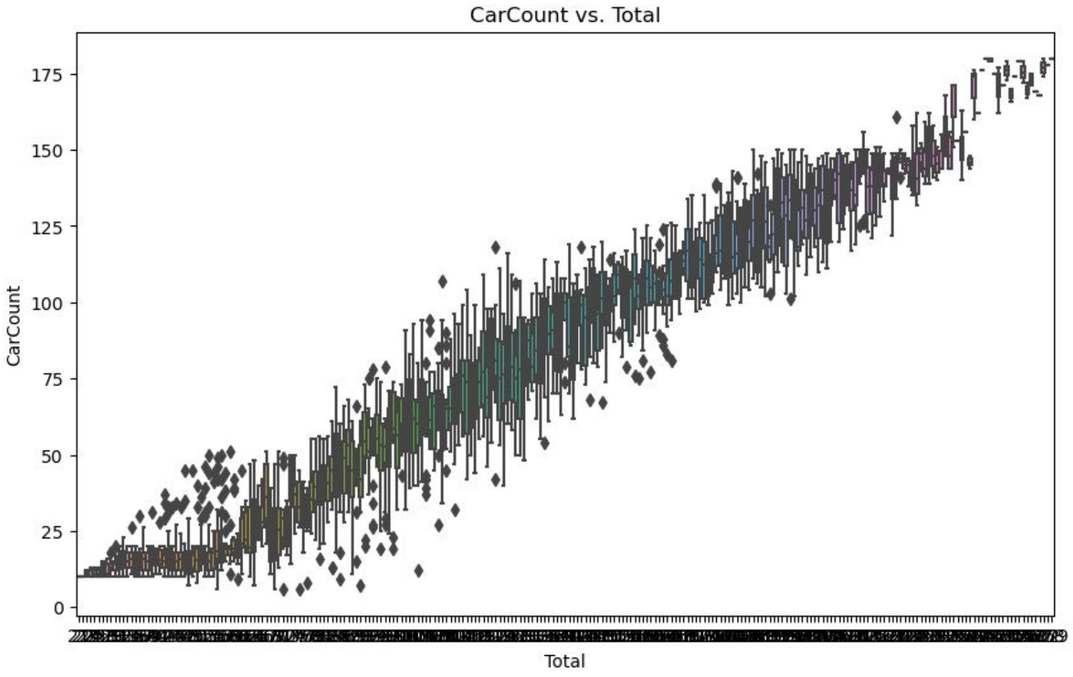
K Nearest Neighbors

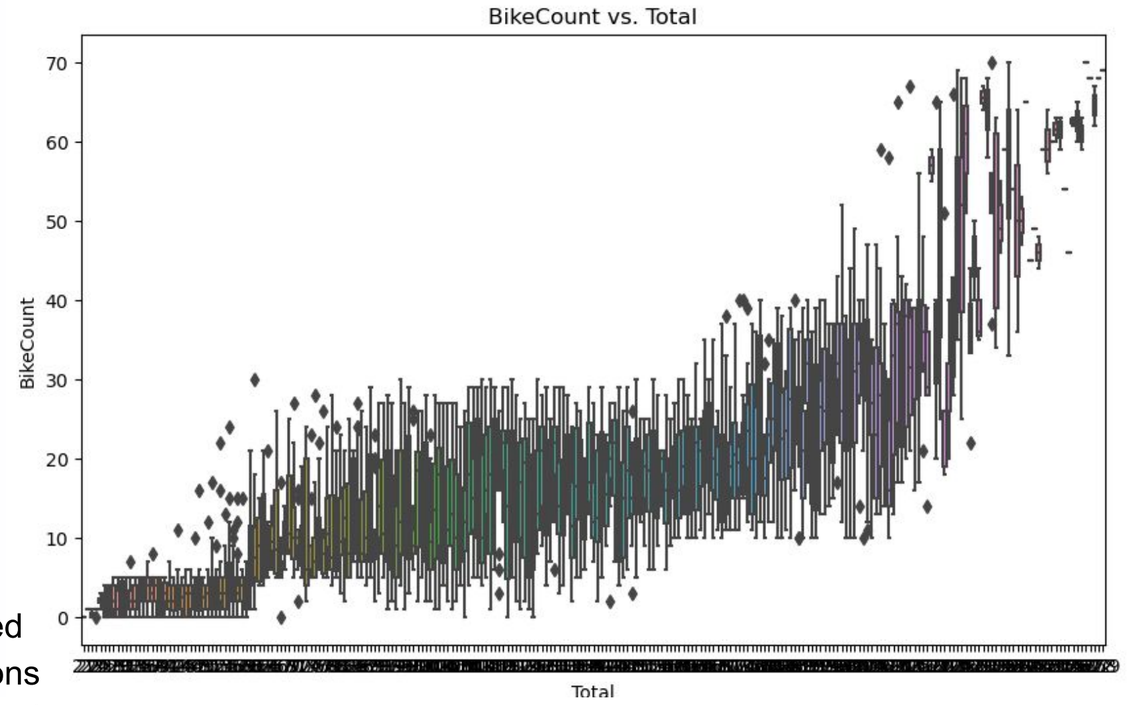




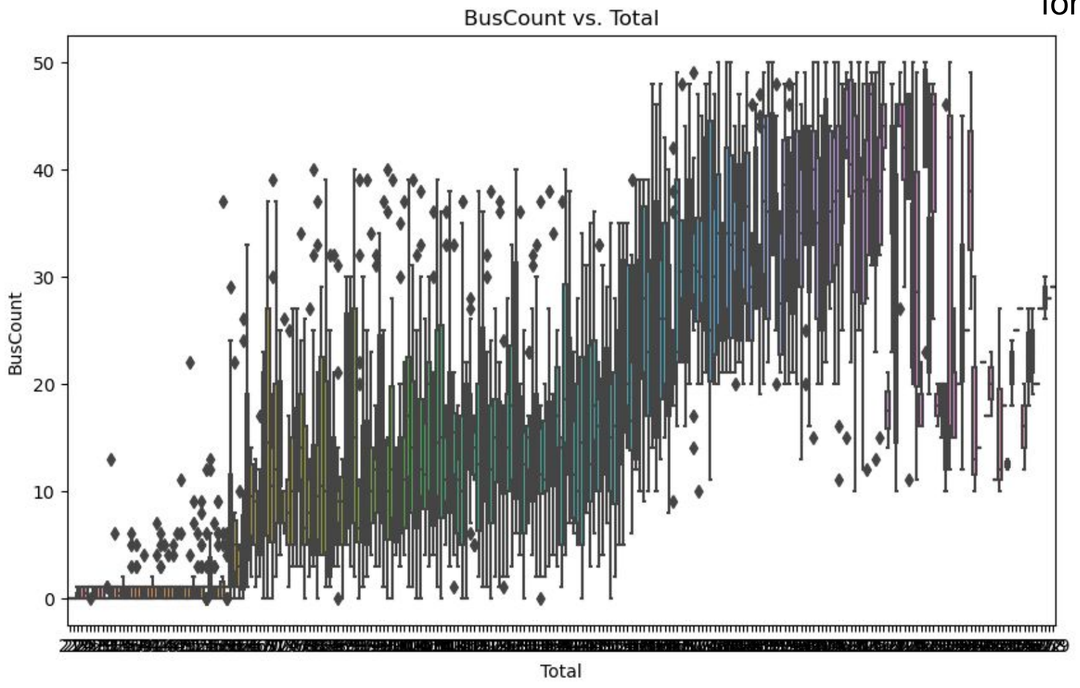
Pairplot: Explored pairwise relationships among features and the target variable ('Total').

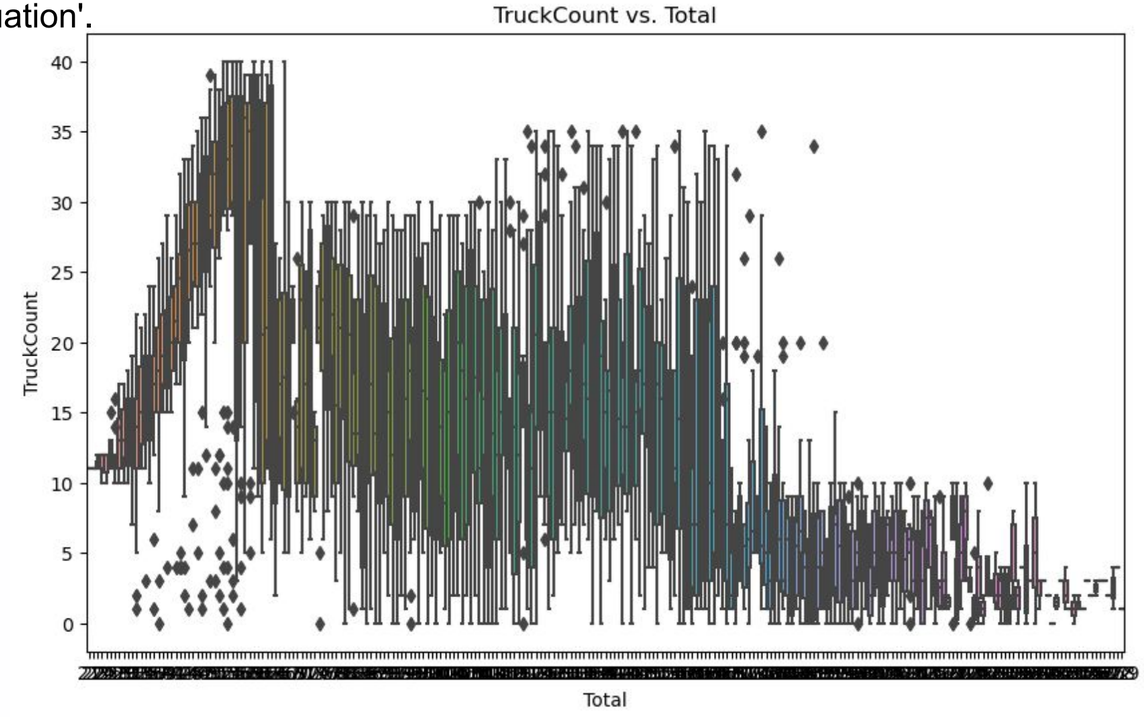
Distribution Plot: Examined the distribution of the target variable.



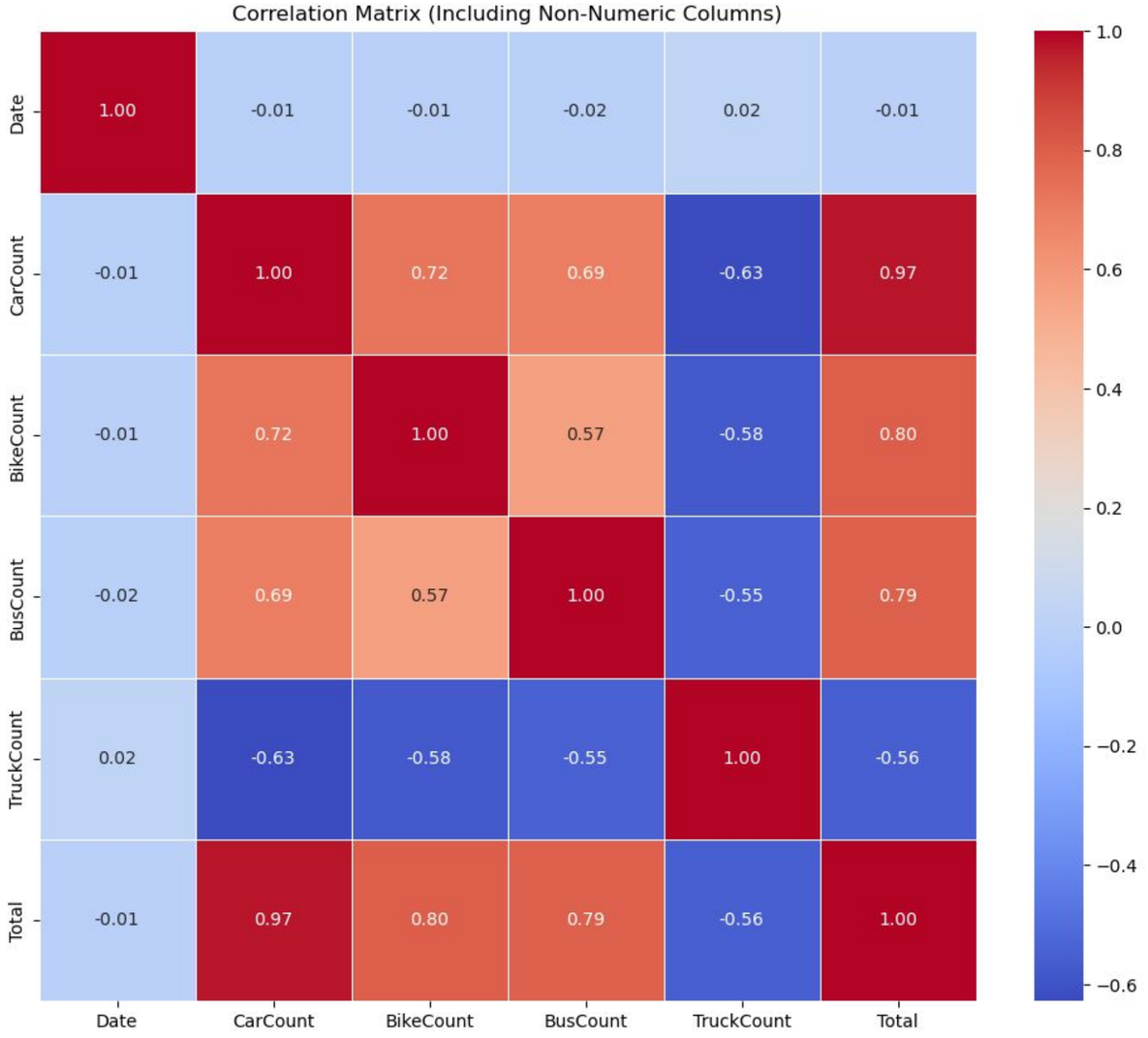


Count Plots: Displayed the count of observations for each 'Traffic Situation'.

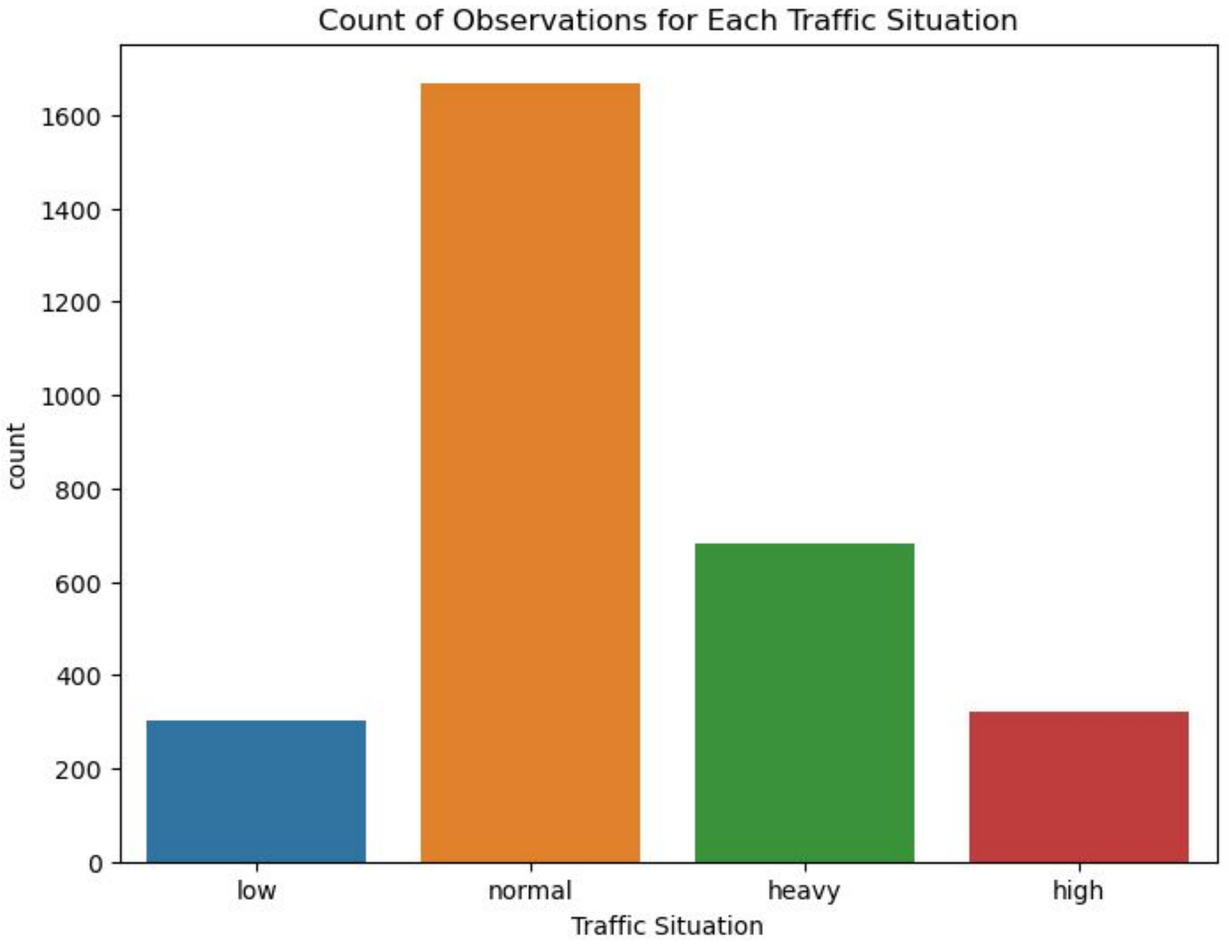




Correlation Heatmap: Visualized correlations among numeric columns, including non-numeric ones.



Box Plots: Investigated the relationship between each feature and the target variable.



Made predictions on both the training and testing sets.

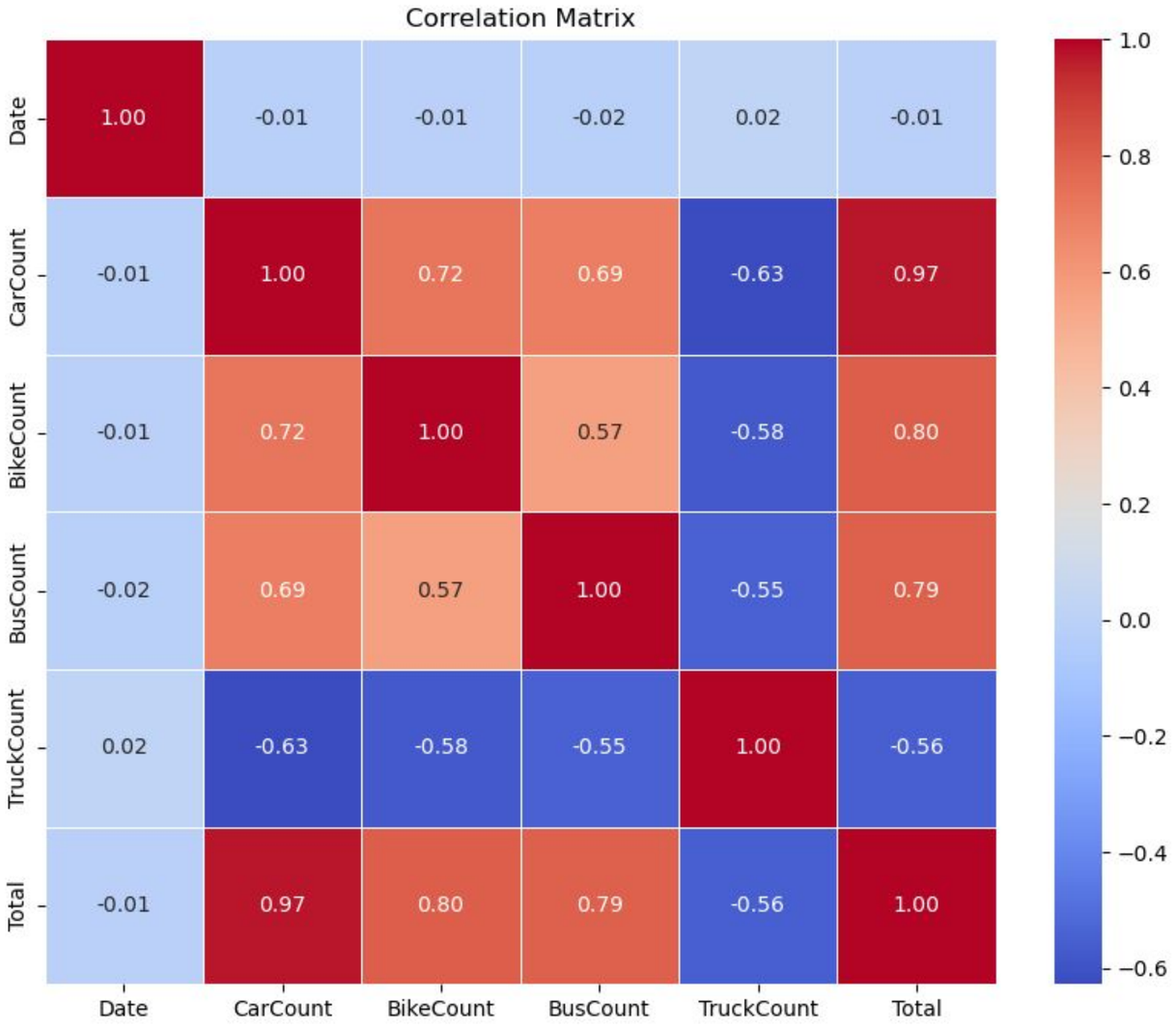
Calculated R2 Scores for training and testing to evaluate model performance. (Training R2 Score: 0.9985725315666825 Testing R2 Score: 0.9977704145531208)

Displayed the evaluation metrics, showing the goodness of fit of the model.

Computed and visualized the correlation matrix.

Visualized the covariance matrix.

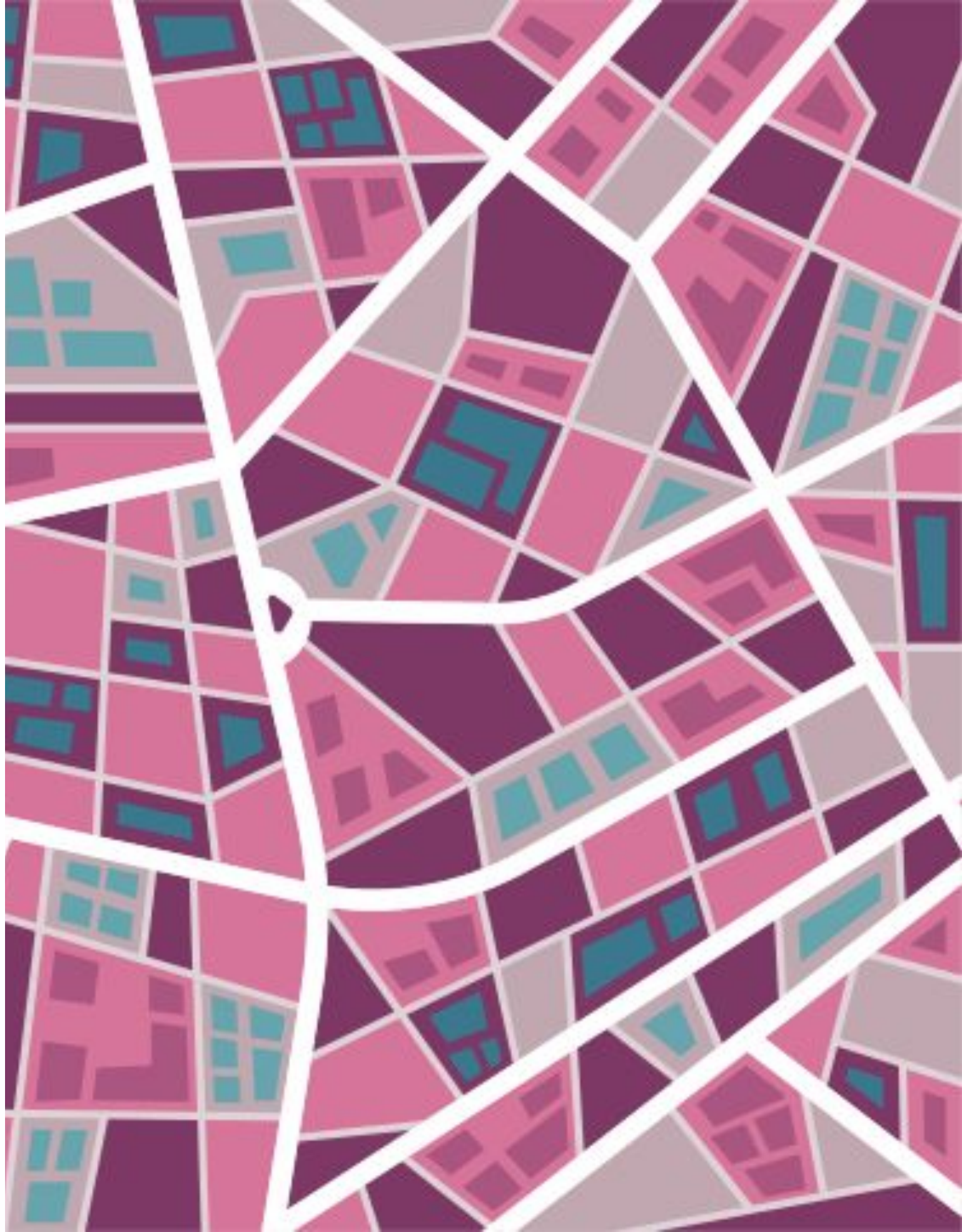
Analyzed relationships and patterns among variables



Multiple Regression Models

Linear, Quadratic, Cubic

Comparing count of automobile types on road vs. traffic situation
In the creation of these models, time was held constant which may explain poor fit.



Linear Regression

Coefficient indicates that 'CarCount', 'BikeCount', 'BusCount', 'TruckCount' all have an impact upon the Y-variable (‘Traffic Situation’), but BusCount' and 'TruckCount' have a significantly larger impact per unit on the road in relation to the Traffic Variable

R^2 Values indicate that only a little over half of the x-variables’ variation is captured by the model, in either training or testing format

```
Linear Regression
Coef = [0.00817537 0.00794831 0.02938121 0.04661421]
Intercept = -0.5084693064704002
Training R^2 = 0.5871946462014692
Testing R^2 = 0.5666406064788108
```

Quadratic Regression

R^2 Values indicate that only a little over half of the x-variables' variation is captured by the model, in either training or testing format

This is very close to the previous model's, but it is minutely lesser in capturing the variation of the X-variables.

```
Quadratic Regression
Training R^2 = 0.5629358150134136
Testing R^2  = 0.5511084542908786
```

Cubic Regression

R^2 Values indicate that only a little over two-thirds of the x-variables' variation is captured by the model, in either training or testing format.

This is definitely better than the previous two regression models.

```
Cubic Regression
Training R^2 = 0.6671459670285883
Testing R^2  = 0.6637118406257296
```

ARIMA Model

Autoregressive integrated moving average (ARIMA)

- It has been historically used to predict traffic
- Uses times series past values to predict future values
-

ARIMA (p,d,q)

AR (Auto Regressive) - model linearly regressed on its own past values allows for the consideration of seasons (week to week)

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t$$

I (Integrated) - considers the difference between each time period

$$\delta y_t = y_t + y_{t-1}$$

MA (Moving Average): Considers the errors from previous prediction errors

ARIMA (p,d,q)

- pmdarima package
- Coefficient of determination is not appropriate because it is a time series.
- used stepwise function to optimize pdq and minimize AIC
- seasonality measured into the model
- To interpret the effectiveness - used the AIC - Akaike Information Criterion

Summary of ARIMA Model

Training

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	2231
Model:	SARIMAX(0, 0, 1)	**Log Likelihood**	-2652.773
Date:	Tue, 28 Nov 2023	**AIC**	5311.547
Time:	12:12:24	**BIC**	5328.677
Sample:	0	**HQIC**	5317.802
	- 2231		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	1.3263	0.019	69.494	0.000	1.289	1.364
ma.L1	0.0361	0.021	1.720	0.085	-0.005	0.077
sigma2	0.6314	0.021	30.069	0.000	0.590	0.673

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	118.50
Prob(Q):	0.96	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.08	**Skew:**	0.56
Prob(H) (two-sided):	0.30	**Kurtosis:**	2.94

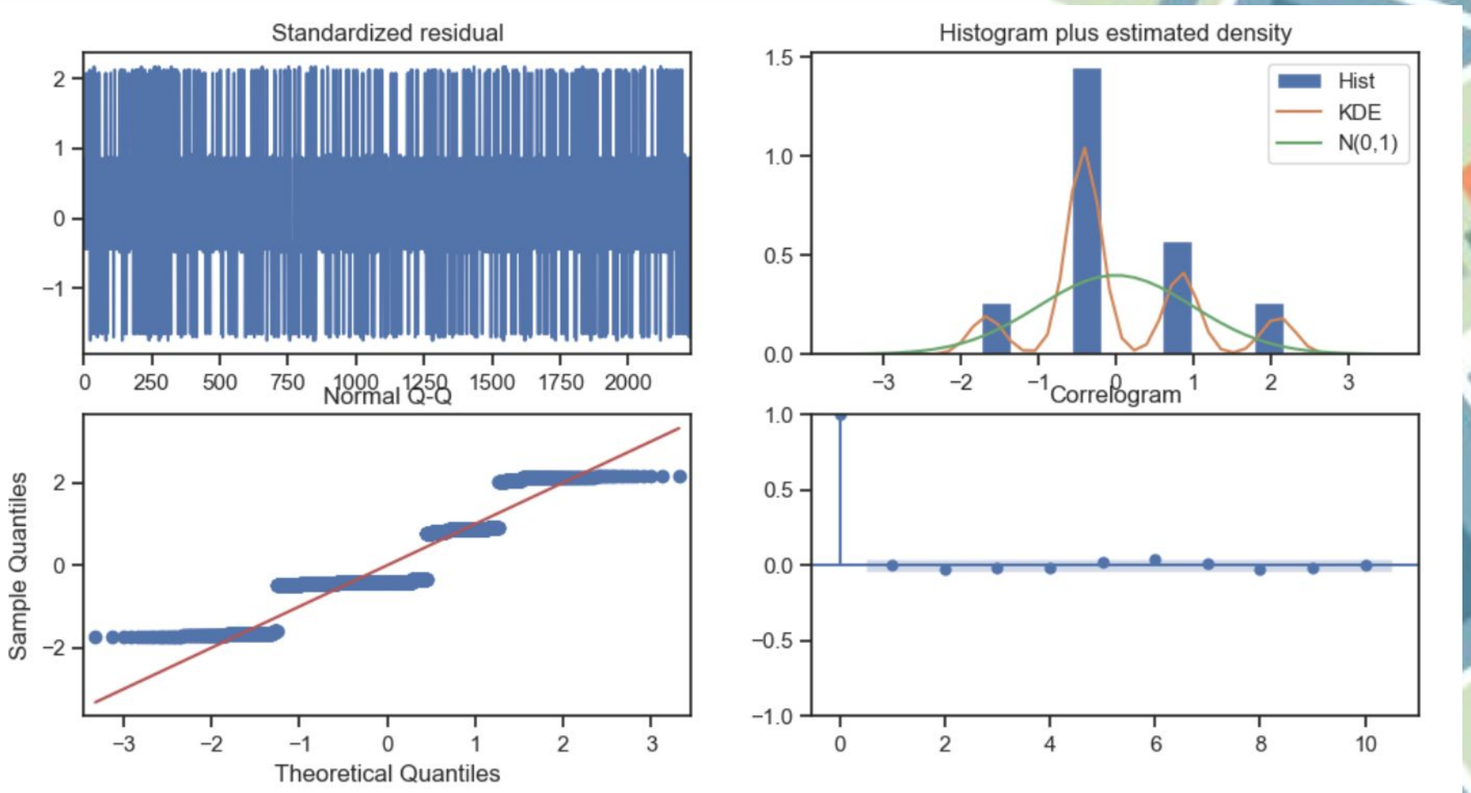
Testing

SARIMAX Results

Dep. Variable:	Traffic Situation	**No. Observations:**	744
Model:	SARIMAX(0, 0, 1)	**Log Likelihood**	-1251.800
Date:	Tue, 28 Nov 2023	**AIC**	2507.601
Time:	12:41:06	**BIC**	2516.820
Sample:	0	**HQIC**	2511.155
	- 744		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	0.5172	0.037	13.920	0.000	0.444	0.590
sigma2	1.7089	0.097	17.607	0.000	1.519	1.899

Ljung-Box (L1) (Q):	161.38	**Jarque-Bera (JB):**	12.23
Prob(Q):	0.00	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.99	**Skew:**	0.26
Prob(H) (two-sided):	0.92	**Kurtosis:**	2.64



Training AIC: 5311.547

Testing AIC: 2516.820

ARIMA

- Notes about this model
- It seems to be a poor fit to the model as it is arrange right now as the Q-Q Norm Plot is not linear.
- Further improvements to the way seasonality is measured within the model may be beneficial.
- Currently with the way the model is fit it doesn’t predict future data effectively.

Conclusion

In conclusion the model with the best fit was the KNeighbors model was the best fit based on our decided measurement of fit, however the AIC and BIC on ARIMA model indicate that it could be advantageous as well if adapted to better account for seasonality.

-

Thank you!

See you again soon!