Modeling New York State Traffic: Which Roadways are Most Congested?

Ryker Dolese, Greyson Elyaderani, Ian Kleppinger, Evan Shope

Modeling New York State Traffic

The dataset under consideration pertains to Annual Average Daily Traffic (AADT), a metric estimating the average daily traffic flow along specific roadway segments. These estimates are derived from short-term traffic counts conducted on the same sections, which are subsequently adjusted to produce the AADT estimate. It's important to note that the AADT value for a particular roadway pertains to the previous year due to this calculation process. This dataset encompasses data for all New York State Routes and roads within the Federal Aid System.

The dataset includes several columns, each providing valuable information about the traffic data:

- **Year**: The year for which the AADT volume value is recorded.

- **Station ID**: A unique identifier for each traffic count station, generated by concatenating the region county code and the station number.

- **County**: The county where the count station is located.

- **Signing**: Indicates the type of route or roadway, such as Interstate, U.S. Route (US), New York State Highway (NY), or 'Local' for town, municipality, or city roadways.

- **State Route**: The route number for roads designated as Interstates or New York State routes.

- **County Road**: The road number for county-level roads.

- **Road Name**: The name of the road.

- **Beginning Description**: A textual description of the start of the roadway segment to which the AADT applies.

- **Ending Description**: A textual description of the end of the roadway segment to which the AADT applies.

- **Municipality**: The name of the municipality where the AADT calculation was performed.

- **Length**: The length of the road segment associated with the count station.

- **Functional Class**: The functional class of the roadway segment, providing information about its role in the transportation network.

- **Ramp**: Indicates whether a ramp is included in the roadway segment (Yes or No).

- **Bridge**: Indicates whether a bridge is included in the roadway segment (Yes or No).

- **Railroad Crossing**: Indicates whether a railroad crossing is included in the roadway segment (Yes or No).

- **One Way**: Indicates whether the roadway segment has one-way traffic (Yes or No).

- **Count**: The AADT volume value, representing the average number of vehicles passing the specified roadway segment daily, as defined by the Begin/End Descriptions. These figures are based on actual counts or statistically validated estimated and forecasted values for the respective year.

Exploratory Data Analysis

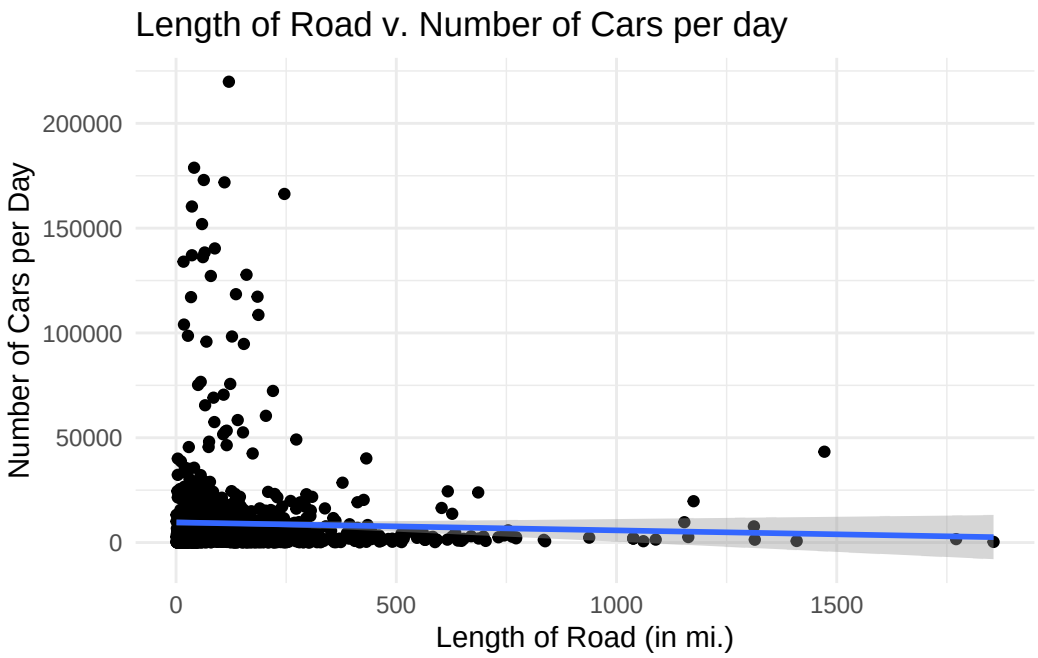

Figure 1: Traffic by Length of Road

In Figure 1, we have plotted length of road (x) and number of cars per day (y). There doesn't appear to be much of a relationship, but we tested this with a linear regression model. The statistics are shown. In summary, based on this regression model, there doesn't appear to be a significant linear relationship between the length of the road (x) and the number of cars per day (y). The p-value for the coefficient of `x` is high, and the R-squared values are close to zero, indicating that the model does not provide a good fit to the data, and road length is not a statistically significant predictor of traffic.

Table 1: Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	9601.454118	713.636520	13.45426	0.0000000
x	-3.783292	3.086136	-1.22590	0.2204752

Table 1 summarizes our model. The p-value is much greater than 0.05, so we reject any possible null hypothesis. This suggests there is no apparent linear relationship between length of roads and traffic congestion.

Traffic Congestion by Signing and Time

Since we couldn't find a strong relationship between road length and traffic congestion from a wide lense, we've decided to explore it on a logarithmically transformed scale while simultaneously deciphering between different road types.

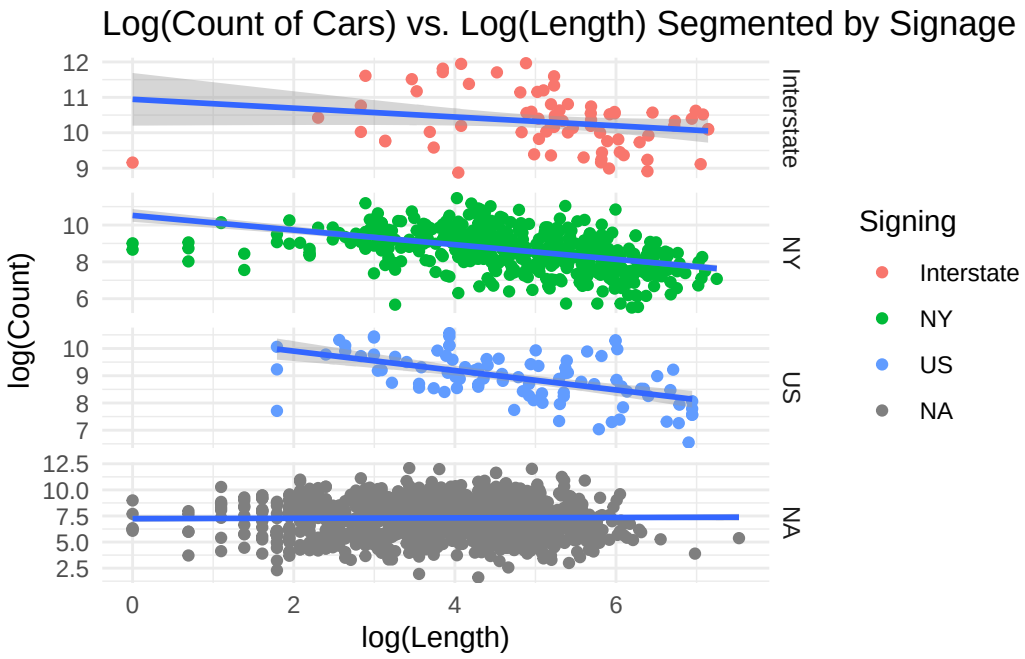

Figure 2: Traffic by Length and Type of Road

Figure 2 presents a visual representation of the relationship between the logarithmically transformed variables of "Length" and "Count" in the context of different types of roadways in New York. The data is grouped by road signing categories (e.g., "NY," "US," "Interstate," and "Local"), allowing for a comparison of these relationships across various road classifications.

Upon examination, it appears that there is a notable negative relationship between the logarithmically transformed "Length" and "Count" variables for New York (NY) and U.S. (US) roadways. This suggests that, on average, as the length of roadways increases, the traffic count tends to decrease. In contrast, the graph shows a relatively flat trend for Interstate roadways, indicating that the relationship between length and traffic count for Interstates is less pronounced. Lastly, the data reveals a slight positive linear relationship for Local roadways, suggesting that, for this category, as the length of local roads increases, there is a modest increase in traffic count.

These observations may imply that different types of roadways in New York exhibit distinct traffic patterns. The negative relationship for NY and US roadways could suggest that longer roads in these categories may experience less traffic congestion, possibly due to their design or purpose. Conversely, the positive relationship for Local roadways may indicate that longer local roads are associated with slightly higher traffic volumes, potentially due to localized factors or urban development. Further analysis and context would be needed to draw definitive conclusions about the reasons behind these observed trends.

Here, in Figure 3, we take a more focused approach and look at the traffic patterns of one individual road, the George Washington Bridge. This road is the busiest by number of cars traveled in the entire state of New York. The above graph shows that at specific points in time the number of cars on the bridge has increased significantly relative to the surrounding years. The late 1980s and early-mid 2000s are the relatively busiest times in the bridge's history. The traffic on the bridge took a large dip into into the mid 2010s, one from which it has not fully recovered.

In Figure 4, the boxplot graph offers valuable insights into the distribution of traffic congestion, represented by the "Count" variable, across different signing categories in roadways. The x-axis measures traffic congestion in terms of daily car counts, while the color aesthetic distinguishes between various signing categories.

Upon closer examination, several noteworthy patterns emerge from the data:

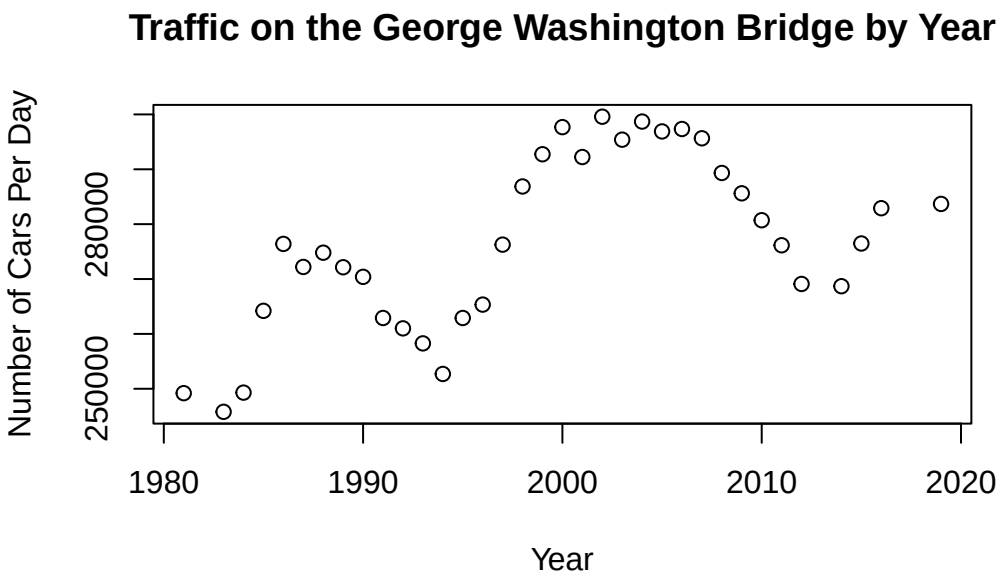

Figure 3: Traffic on the George Washington Bridge

Figure 4: Traffic Congestion by Signage

1. Interstate roadways face the highest congestion among all types. Their median congestion level is approximately 30000 daily cars, suggesting substantial traffic. However, the wide interquartile range indicates varied congestion levels along different sections—some with lower congestion and others experiencing significantly higher traffic.
2. Local roads experience the least congestion among these categories. With a small interquartile range, congestion levels remain relatively consistent. The median congestion on Local roads is the lowest among all categories.
3. US and NY roadways show moderate congestion levels. US roads have a slightly higher median of about 2000 daily cars compared to NY roads, which rank third. Both categories exhibit rightward skewness, indicating moderate median congestion but instances of much higher traffic leading to a positively skewed distribution.

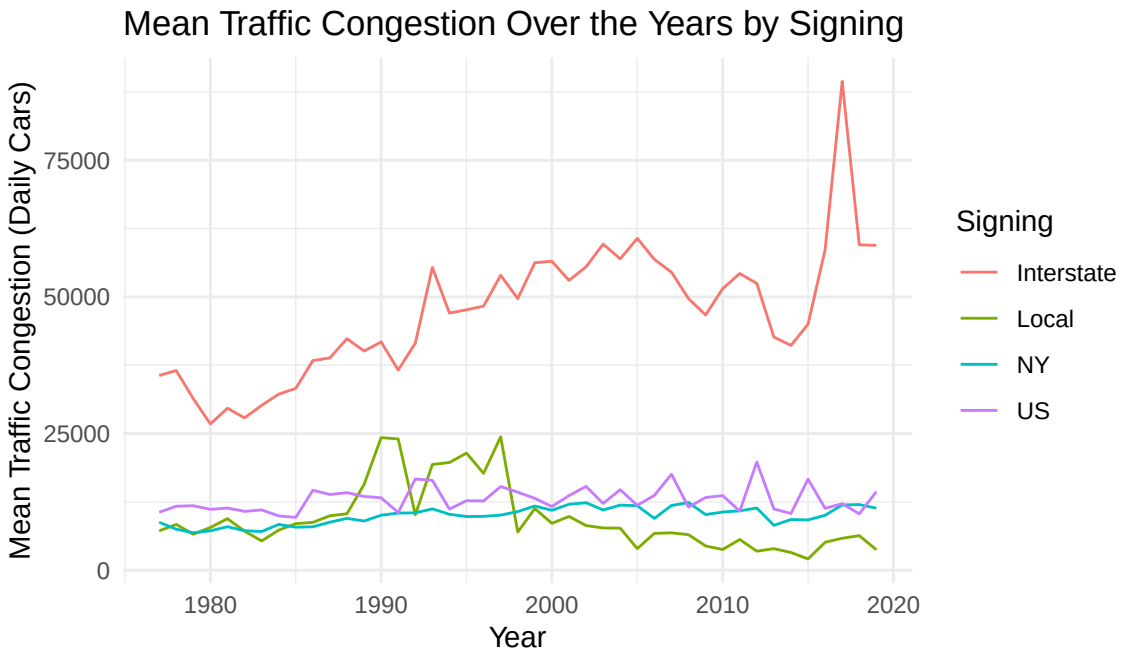

Figure 5: Mean Traffic Over Time by Signage

Figure 5 illustrates the mean traffic congestion trends over time for various types of roadways, revealing consistently higher congestion levels for Interstates, a stable pattern for US and NY roadways, and significant variability in congestion on local roads. Interstate roadways consistently exhibit higher mean traffic congestion levels year over year. The data reveals a significant peak in congestion around 2017, with congestion levels reaching approximately 90,000 daily cars. Since then, congestion on Interstate roadways has steadily increased. This suggests that these roadways face ongoing challenges in managing and alleviating congestion.

In contrast to Interstates, both US and NY roadways demonstrate relatively stable mean traffic congestion levels, hovering around 12,000 daily cars over the years. There are no significant spikes or sharp changes in congestion patterns for these categories. This stability suggests a consistent traffic volume on these roadways, likely due to their regional significance and relatively predictable traffic patterns.

Local roadways display high variability in mean traffic congestion year over year. At certain points, local roadways have experienced congestion levels surpassing those of US and NY roadways, indicating sporadic periods of increased traffic volume. This variability may be attributed to localized factors, such as events, construction, or changing infrastructure, impacting traffic patterns on local roads.

The analysis highlights distinct trends in traffic congestion across different signing categories of roadways. Interstates face a continuous increase in congestion, with a notable peak in 2017. US and NY roadways maintain stable congestion levels, while local roadways exhibit considerable variability. These insights can inform transportation planning and policy decisions, emphasizing the need for targeted strategies to manage congestion effectively in different roadway types, especially for Interstates experiencing persistent growth in traffic volume.

Examinining the Effect of One-Way Streets and Bridges

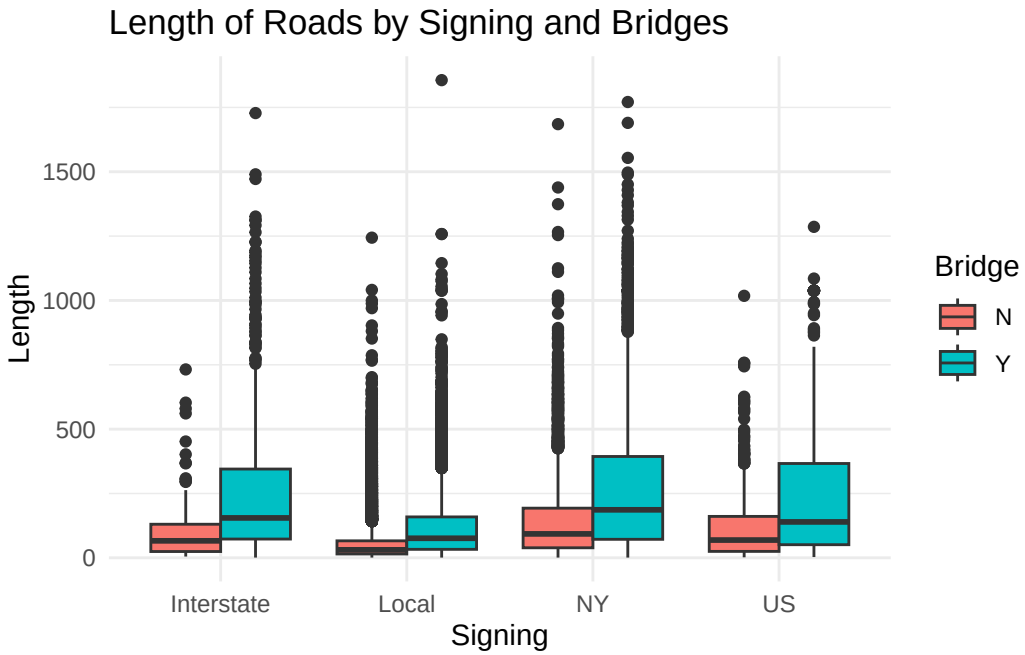

Figure 6: Roads by Signing and One-Way

In Figure 6, based on data from the year 2019, we compare the length of roadways across different signing categories, with a focus on whether each roadway has bridges.

It reveals that, on average, roadways with bridges tend to be longer, with a higher median length for each signing category, but they also exhibit greater variability as indicated by larger interquartile ranges (IQR). Specifically, New York (NY) roadways have the longest average length, with a median around 375 miles, followed by Interstates, US roadways, and local roads. However, local roads display significantly less variability in length, with most roads condensed to less than 100 miles.

All signing categories exhibit rightward skewness, indicating the presence of longer roads, although they vary in terms of median length and the spread of data.

Figure 7, based on filtered data where traffic congestion (Count) is less than or equal to 15,000, compares the distribution of traffic congestion between one-way and non-one-way roadways.

It reveals that roadways classified as “One Way” tend to have higher traffic congestion, and the frequency of lower congestion values is lower for this category. Surprisingly, non-one-way roadways have lower congestion on average, with a higher frequency of lower congestion values. This observation challenges the common expectation that traffic congestion would increase on two-way streets. Additionally, one-way streets exhibit a less pronounced rightward skew in traffic congestion compared to non-one-way streets, further highlighting the differences in congestion patterns between these roadway types.

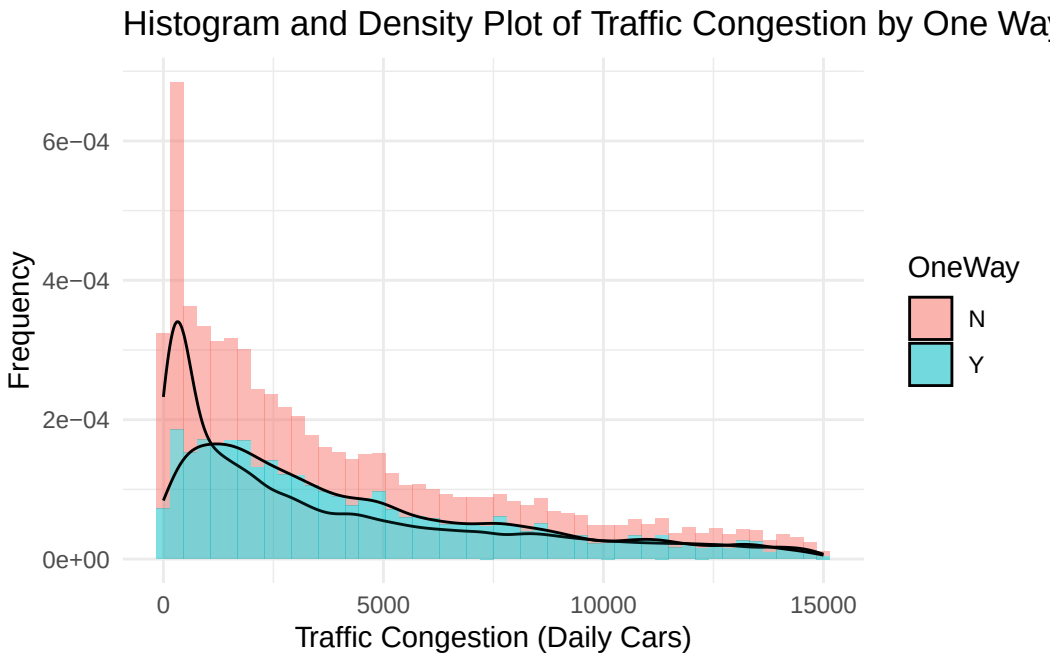

Figure 7: Traffic Congestion on One Way Streets

Assessing Normality of Traffic Congestion

A normal Q-Q plot graphs the theoretical quantiles of a normal distribution on the x-axis and the sample quantiles (taken from observed data) on the y-axis. If the data roughly follows a normal distribution, then the scatter plot should roughly follow the plotted line. In the faceted grid (Figure 8), which displays a normal Q-Q plot for each type of road in NY state based on their daily count of traffic throughout 2019, the graphs display the non-normality of each distribution. Some look more normal than others, such as the roads with the "US" or "Interstate" signings, but deviate at the extremes. Some, such as the local roads, have massive discrepancies on the right end of the extremes. Regardless, the Q-Q plots alone are not enough to test for normality, which is why Shapiro-Wilks tests are performed below. Additionally, we have plotted a histogram, and there appears to be a strong right-skew when it comes to yearly traffic congestion. This information will likely inform our modeling in the future.

Table 2: Shapiro-Wilk Test Results

Test_Statistic	P_Value	Test	Road_Type
0.8979345	0	Shapiro-Wilk normality test	Interstate
0.8682380	0	Shapiro-Wilk normality test	US
0.6413259	0	Shapiro-Wilk normality test	NY
0.3328159	0	Shapiro-Wilk normality test	Local

The Shapiro-Wilks tests (results shown in Table 2) work to certify assumptions made about the traffic's distribution from the normal Q-Q plots. The test is a statistical test with a null hypothesis that the distribution of the sampled data is normal, whereas the alternative hypothesis is that the sampled data does not reflect a normal distribution. Since the limit for the test is 5000 observations, the local and NY roads had to be randomly sampled, as there were more than 5000 observations. Once done, the data could be compared to a normal distribution to test the null hypothesis. In each case, the p-value is near zero. This means that we are almost certain that the roads do not follow a normal distribution, as a low enough p-value

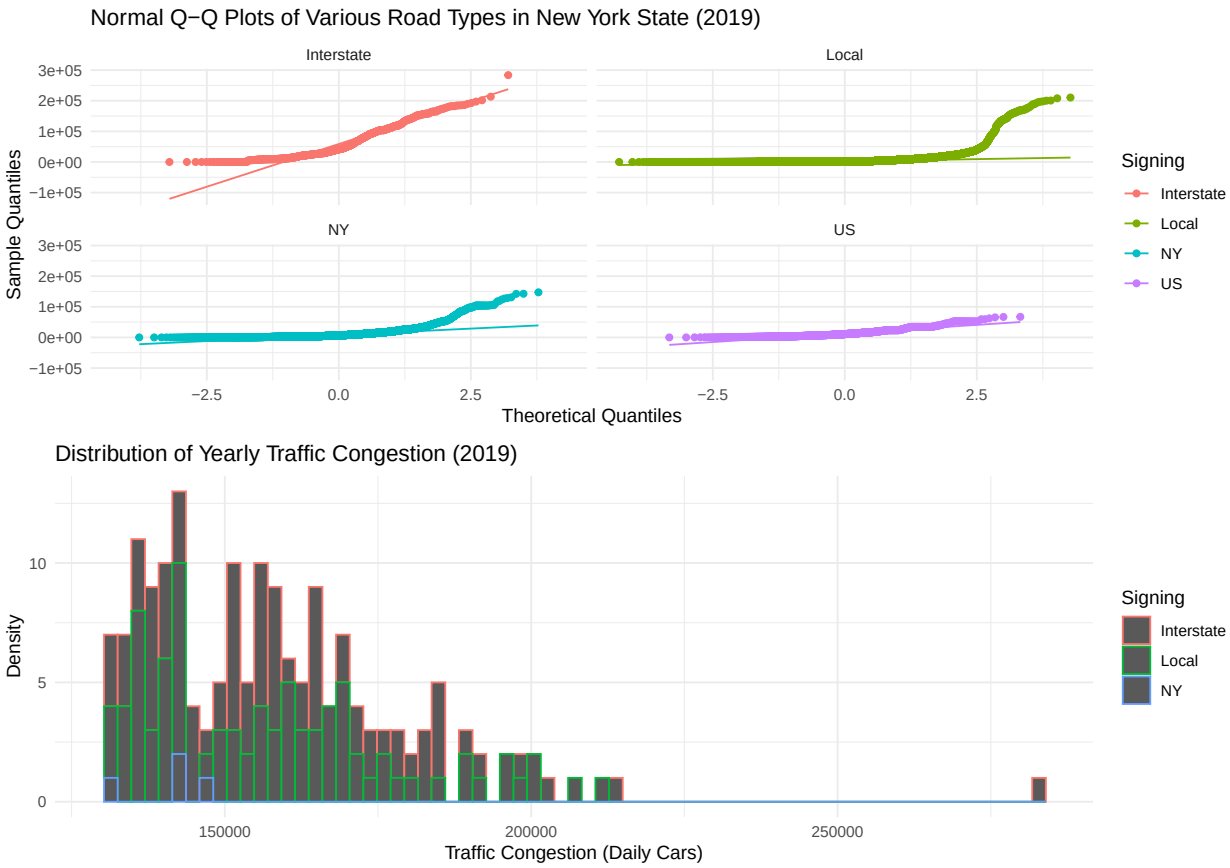

Figure 8: Assessing Normality of Traffic Congestion

allows one to reject the null hypothesis - in this case that the sampled data do follow a normal distribution. Knowledge about the distribution of the data allows one to more accurately model the data and is thus an extremely important step in analyzing any data set.

Introduction of Gas Price Data in New York State

- Here, we have introduced data on yearly gas prices in counties in New York. We are interested in exploring the relationship between traffic congestion and gas prices.

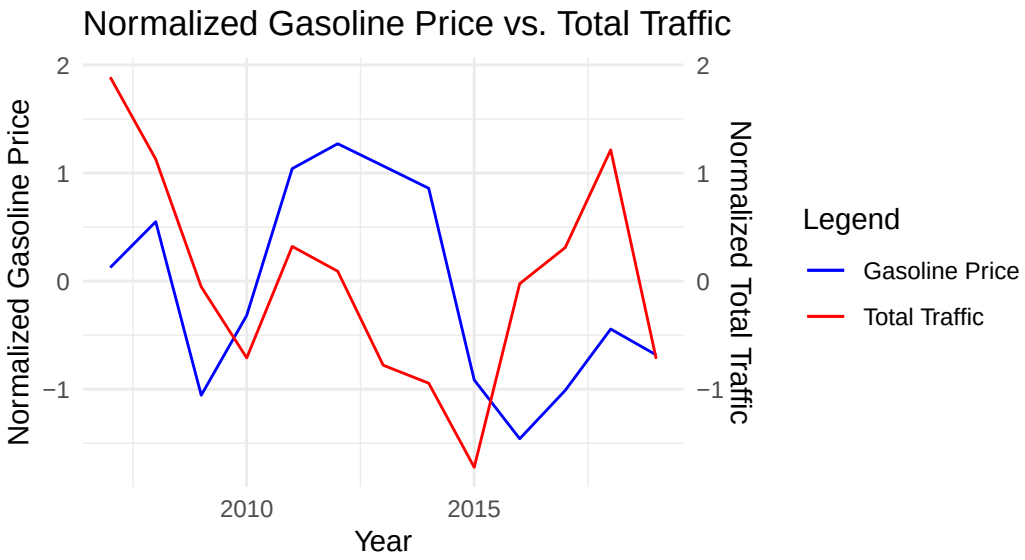

Figure 9: Trends Between Gas Prices and Traffic

Figure 9 displays a compelling time series trend between the number of cars on the road (normalized total traffic) and gas prices (normalized average gasoline price) over several years. It reveals that as gas prices tend to increase over time, there is a corresponding increase in the number of cars on the road. The same generally holds for decreases in gas prices and traffic volume. This correlation suggests that economic and behavioral factors play a significant role in shaping both gas prices and traffic patterns. However, it's essential to note that the relationship is not perfectly synchronized; there are instances where fluctuations in gas prices do not directly mirror changes in traffic volumes.

There are some possible explanations:

- **Economic Conditions**: During periods of economic growth and prosperity, individuals often have more disposable income, leading to an increased demand for personal vehicles and, subsequently, higher traffic volumes. Rising gas prices can be indicative of strong economic activity, which tends to stimulate both vehicle ownership and usage.
- **Consumer Behavior**: Fluctuations in gas prices can influence consumer choices regarding vehicle usage and fuel consumption. As gas prices rise, consumers may seek out more fuel-efficient vehicles or explore alternative transportation options, which can temporarily mitigate the increase in traffic.

Figure 10 extends the analysis of the relationship between normalized gasoline prices and normalized total traffic, this time by facetting the data by county. The trends observed within each county generally align with the overall trend, with Albany, Broome, and Nassau Counties exhibiting particularly close alignment. This suggests that while there are variations among counties, the underlying factors driving the correlation between gas prices and traffic remain consistent. Possible explanations for these trends include economic conditions, consumer behavior, and supply and demand dynamics.

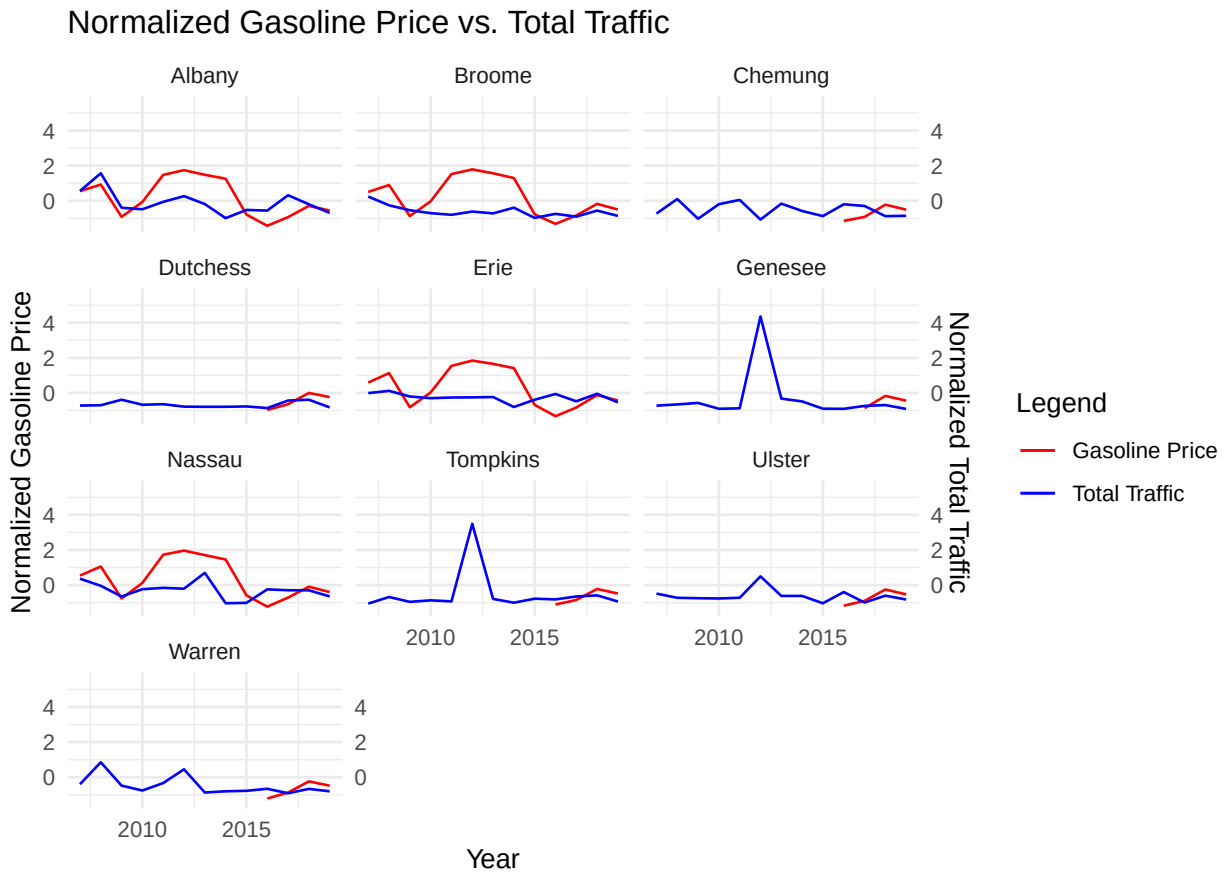

Figure 10: Effect of Gas Prices on Traffic in Each County

String Analysis for Road Type Identification

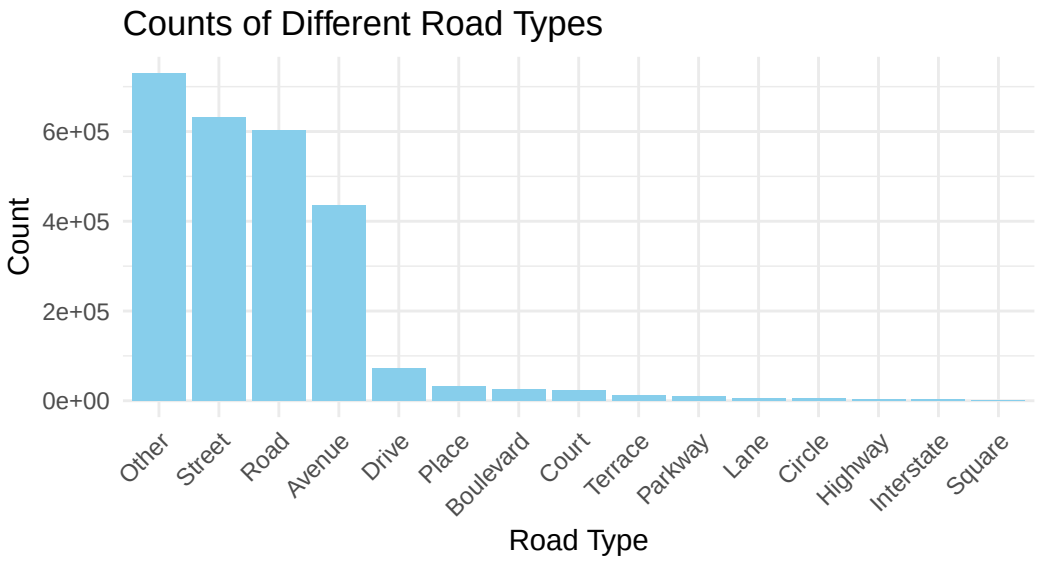

Figure 11: Road Types

Figure 11 shows the counts of different road types. Streets and roads are the most common, with over 600,000 occurrences each. Avenues are the next most common at around 400,000, followed by less common road types like Drive, Boulevard, Place, and Terrace. The rest of the road types are even less common, while plenty of road types do not fall under any typical category.

Modeling

Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

Our goal is to predict traffic volume on any given NY road for an entire year. The first model we will consider is linear regression. We have several regressors available to us, but it is crucial to pick the best ones for our model. Our "full model" might include: Lag, Lag1, Lag2, FC, County, OneWay, RoadType, Length, Signing, Year, RC, Ramp, Bridge, AvgGas.

"Lag, Lag2, Lag3" are feature engineered fields that represent the previous daily traffic total in a year on any given road. Therefore, these measure the 3 previous years. Our first objective is to use a metric, such as AIC, to decide which regressors are most suitable.

Although we understand the limitations of a stepwise algorithm approach (including biased and high r-squared values, f statistics and p-values being overestimated due to multiple-comparison, and the inability to identify the ideal subsets of regressors), we have used step-wise regression as a start to identify some possible regressors. We have decided to use a step-wise forward algorithm to find the best MLR model by AIC. Using the stepwise algorithm to minimize AIC, we find that the best linear regression model includes the following variables: Lag + Lag2 + Lag3 + OneWay + Bridge + RoadType+ FC + County. The results of this process can be seen in table 3. These are the most powerful variables, and the order indicates which ones are most pertinent to our model. We have shown the results of our regression in Tables 4 and 5.

Table 3: Step-Wise Selection

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	2714	1.529991e+12	54708.53
+ Lag3	-1	1.524364e+12	2713	5.627602e+09	39492.05
+ Lag2	-1	1.252831e+09	2712	4.374771e+09	38810.34
+ County	-48	6.207628e+08	2664	3.754008e+09	38490.86
+ Lag	-1	8.302378e+07	2663	3.670984e+09	38432.14
+ Year	-1	4.555165e+07	2662	3.625433e+09	38400.24
+ FC	-13	6.130094e+07	2649	3.564132e+09	38379.94
+ RoadType	-11	5.167192e+07	2638	3.512460e+09	38362.29
+ Bridge	-1	4.428441e+06	2637	3.508031e+09	38360.87

Table 4: Linear Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	986.5673257	301.2061713	3.275389	0.0010688
Lag	0.0175149	0.0023218	7.543831	0.0000000
Lag2	0.4031735	0.0161097	25.026694	0.0000000
Lag3	0.5652962	0.0158354	35.698178	0.0000000
OneWayY	295.7089130	174.1592070	1.697923	0.0896403
BridgeY	189.9434093	93.8319505	2.024294	0.0430406
RoadTypeBoulevard	-458.2405561	325.8483593	-1.406300	0.1597529

Table 5: Additional Statistics

r.squared	adj.r.squared	sigma	statistic	p.value	df	df.residual	nobs	MSE
0.9976756	0.9976077	1161.312	14699.07	0	77	2637	2715	1348645

Our model has performed particularly well, with an R-squared value of 0.99, suggesting much of our data can be explained by our linear regression model. But we *need* to be this precise when comparing roadways with extremely varied values of traffic flow. For this reason, it might be best to look at MSE and sigma, providing a more practical look at the model accuracy. Additionally, since we used the AIC step-wise algorithm, we can be certain these are good regressors with low autocorrelation. Please note that some of the categorical variables have been excluded from the coefficient summary because of the vast amount of options (e.g. "County").

Poisson Regression

$$\mu = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

Although our model above performed respectfully, we will explore other possibilities. Since traffic in roadways can be modeled as a poisson, right-skewed, random variable, we have opted for a poisson regression. This should be more valid for the data we have. Using most of the same regressors, we have fit a new model. Please note we have taken the natural log of the Lag variables since we have log-transformed the dependent variable due to the poisson process. The results can be seen in Table 6 and 7.

Table 6: Poisson Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	-3.6501638	0.0667507	-54.6835135	0.0000000
log(Lag)	0.0067856	0.0003338	20.3298262	0.0000000
log(Lag2)	0.4500604	0.0032924	136.6969583	0.0000000
log(Lag3)	0.5447867	0.0032770	166.2460406	0.0000000
OneWayY	0.0108828	0.0009410	11.5650793	0.0000000
Length	-0.0000001	0.0000006	-0.1644385	0.8693860
SigningLocal	-0.0020430	0.0007837	-2.6068885	0.0091369
SigningNY	-0.0018105	0.0006776	-2.6717747	0.0075451

Table 7: Additional Statistics

null.deviance	df.null	logLik	AIC	BIC	deviance	pseudo_r_squared	sigma	MSE
54803664	2714	-47579.23	95324.46	95814.71	64863.93	0.9982646	1104.642	1220233

The Poisson model has a high R-squared value, very similar to our linear model, showing it is an accurate predictor of traffic given the variables provided in the data. This makes sense, because Poisson distributions are designed to measure discrete counting data, like a count of cars on a road, while linear models best used describing continuous data, which is not how our measures of traffic are formatted. This representation can also be seen in Figure 12, as the points in the Poisson graph are more closely grouped around the trend line and the outlier points are visibly closer to their predicted values than those in the linear model. Additionally, with sigma being around 1100, it suggests the model is typically inaccurate by around 1100 vehicles per day. This is extremely accurate and slightly better than our linear model, which makes sense.

Extreme Gradient Boosting

Our linear regression and poisson models have performed extremely well, but we have determined that a Machine Learning (ML) model might perform even better. For our third model, we have decided to employ an extreme gradient boosting model; this is an ensemble model.

It is designed to be both computationally efficient (e.g. fast to execute) and highly effective, perhaps more effective than other open-source implementations. The name xgboost, though, actually refers to the engineering goal to push the limit of computations resources for boosted tree algorithms.

Table 8: Three Model Summary

Model	SQRT.MSE.	Mean.Absolute.Error	R.squared
Linear	1144.509	457.7138	0.9976756
Poisson	1104.642	457.6234	0.9982646
XGB	1191.425	330.3005	0.9973607

As seen in the summary (Table 8), our model performs almost identically to our other models, especially with a high R-squared value. We can be confident it accurately predicts traffic volume, but the lack of interpretability is something to consider. However, it is important to validate our model. The first step is residual analysis, which we have begun to analyze in Figure 12.

Figure 12 displays the performance of three models' performance on predicting the average amount of daily traffic in a given year based on a set of variables. The variables were chosen using information criteria to select

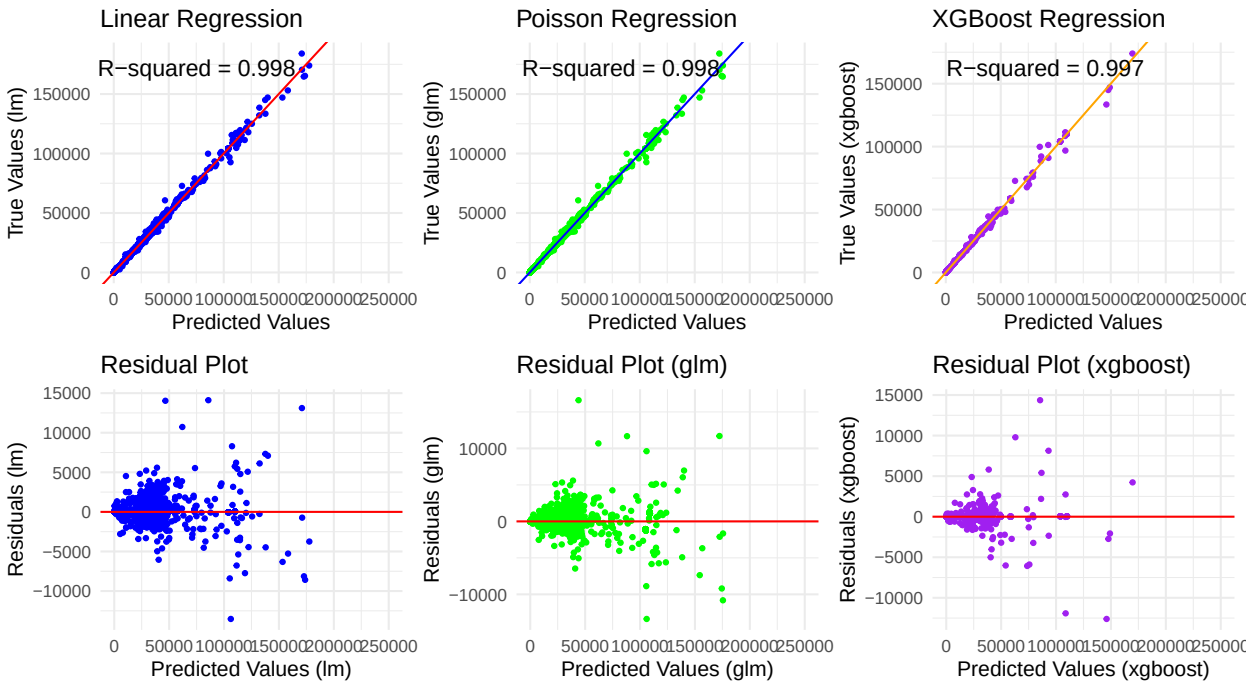

Figure 12: Model Validation

the model that performed the best without the risk of overfitting. The first model is a basic multivariate linear regression. The second model includes a Poisson distribution to predict values. The third model is an XGBoost machine learning model. While the models appear to predict traffic extremely well, there appears to be some heteroskedacity of residuals, especially for high predicted values – but it is not worth discarding the models.

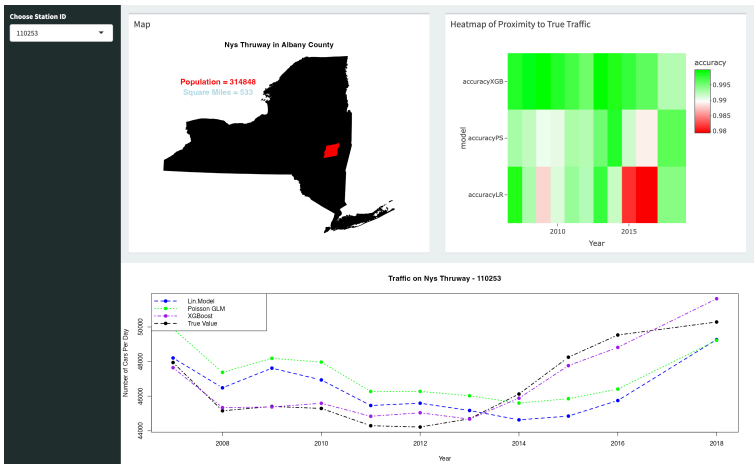

Figure 13: County Analysis and Model Accuracies

In Figure 13, we have once again taken a more focused approach – this time assessing model accuracy for a given road. In the dashboard (which is available through r shiny), you are able to select the road and three distinct plots appear: one identifies the county in which the road appears (also providing population and square miles for the county). This provides insight into the road's surrounding areas. To the right, you can see an accuracy heatmap, allowing for a visual quantification of our three models' accuracies through the years. We have also plotted the predicted traffic volume by our models and true traffic volume in the

bottom plot.

For the example in Figure 13, we have chosen the NYS Thruway in Albany County. Here, we acquire some important information. For instance, while Albany's square mileage is low (signified by the blue coloring), its population is high (signified by the red color). Thus, traffic volumes mght be a bit higher on a highway like NYS Thruway. Taking a look at our model predictions, it appears the XGB model outperforms the others and comes very close to the truth for each year. The poisson and linear models are a bit more varied. While this is a very specific example, the versatility of the dashboard can be extended further.

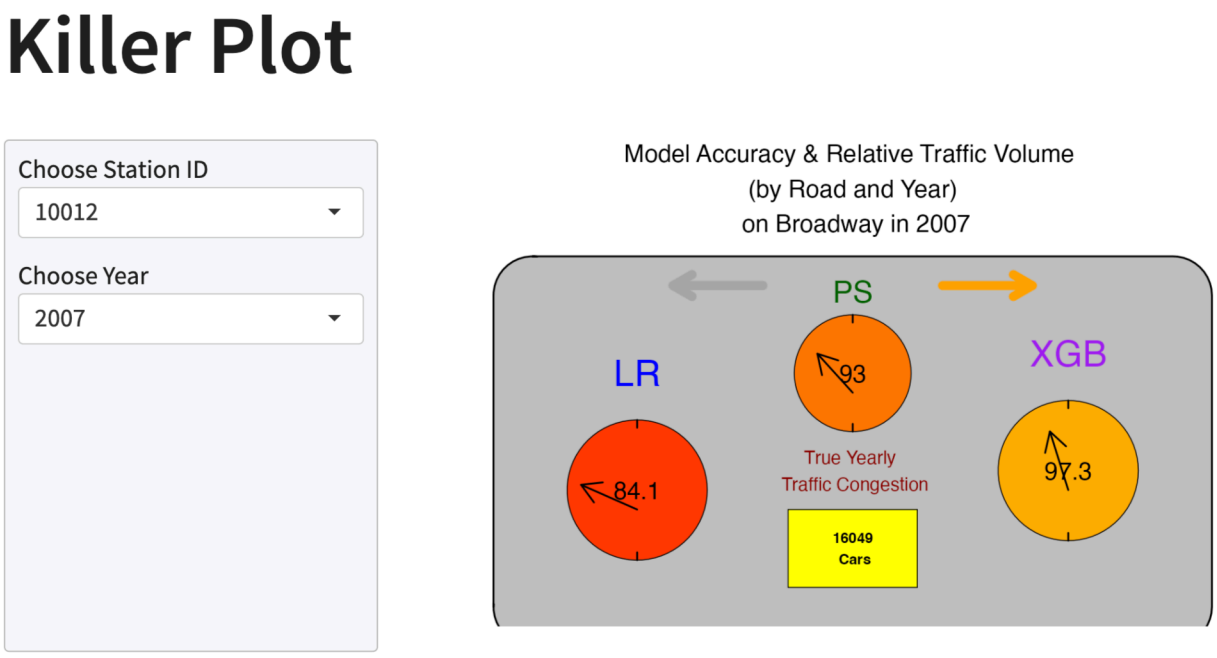

Figure 14: Killer Plot

For our killer plot, we recreate a car's dashboard – keeping with the theme of cars and traffic. In the r-shiny component, you can select the road and the year. Within the dashboard are 3 gauges. Each measures the accuracy of one of our 3 models to predict traffic volume. The gauges move based on how well the model performed for the given road and year, and the performance metric appears within the gauge. Additionally, the color of the gauges changes and is based on the relative prediction accuracy. We make green signify high accuracy and red low accuracy. Additionally, there is another component of the dashboard that says the yearly traffic flow on that road. This also includes a color scale component and changes color based on the traffic flow compared to other roads. The legend for the scale is included in the plot.

In the above example (Figure 14), we have selected Broadway in year 2007 and can see that the XGB predicted the traffic congestion (true value being 16049, which is a moderate amount specified by the yellow) with over 99% accuracy. The poisson regression also did particularly well. As such, this plot provides us with a narrow look into model performance.

Conclusion

Our project successfully predicted traffic volume using XGBoost, Linear Regression (LR), and Poisson Regression. All three preformed similarly. They particularly excelled in forecasting lower traffic values. LR and Poisson Regression offer valuable interpretability when compared to the XGB model. And with the lowest

MSE, we have decided that Poisson Regression might be our best model. This makes sense considering the distribution of traffic flow. The presence of heteroskedastic residuals indicates some variability in prediction accuracy across traffic volumes.

Future steps include exploring finer time-scale data for time series analysis, uncovering seasonality patterns in traffic variations. Despite these potential avenues, we are satisfied with the current model accuracy. Further refinements could involve hyperparameter tuning, ensemble methods, and addressing heteroskedasticity issues for more robust predictions – possibly including further log transformations.

In conclusion, our project provides accurate traffic predictions, highlighting the strengths and trade-offs of different models. As we progress, integrating more granular data and enhancing model interpretability will contribute to a comprehensive traffic prediction framework.

References

1. Annual Average Daily Traffic (AADT) Beginning 1977 (New York State Department of Transportation)

- URL: https://data.ny.gov/Transportation/Annual-Average-Daily-Traffic-AADT-Beginning-1977/6amx-2pbv

New York State Department of Transportation. (2019). Annual Average Daily Traffic (AADT) Beginning 1977* [Data set]. Data.gov. URL

2. Gasoline Retail Prices Weekly Average by Region Beginning 2007 (New York State Energy Research and Development Authority)

- URL: https://data.ny.gov/Energy-Environment/Gasoline-Retail-Prices-Weekly-Average-by-Region-Be/nqur-w4p7

New York State Energy Research and Development Authority. (2020). *Gasoline Retail Prices Weekly Average by Region Beginning 2007* [Data set]. Data.gov. URL

Link to Code