SAS Final Project:
PA Vehicle Accidents Report

Group 3: Matt Chylack, Ben Phillippy, Jayden Carlucci, and Matt Granato

Statistical Analysis and Predictive Modeling

5/1/2023

Executive Summary

Our group analyzed vehicle accidents in Pennsylvania from the years 2016-2021. Our goal was to define and categorize the variables that most influenced vehicle crashes and their severity. Using the variables we found most influential, we then implemented them into predictive modeling techniques to foresee their predictability using the variables given in our dataset. Our ultimate goal is to establish and conclude what variables are most influential in predicting severe crashes. This information can then help public officials and drivers within Pennsylvania to ultimently prevent vehicle crashes and minimize the severity of these incidents.

Why It Matters

What we hope to achieve is to find a way to accurately predict the highest risk times and conditions that car crashes occur. With that information we want to make people more aware of crash risk in order to help decrease the amount of crashes. We believe that by sharing our findings we can help people become more aware of dangerous road conditions and what signs to look out for. We hope to influence future drivers into being safer and driving in good conditions. With our new data we want to influence law to make more road safety laws in order to decrease the accident death toll.

Our Findings

After running various different tests, we were unable to ultimately construct a model that had a significant level of predictability given the variables of our data

set. We constructed many different models, and discovered that the best model given our variables would be a decision tree. Through this model along with various regression models we were able to discover the variables that had the greatest correlation and predictive power for our models. However, these variables were never able to formulate in a model that produced a model with a R-squared value of greater than 5% or having an average squared error less than 2.37. In the end we were able to discover some relationships within the data, and determine more important variables to monitor in the future, however we never were able to construct a model that would explain the additional 95% of reasons for the traffic getting delayed.

What It Means

Based on the high levels of predicting error in our models, what we now know is that there are other variables unseen in our dataset that contribute more to the duration in which traffic gets delayed due to an accident. We saw variables like specific weather conditions, time of day, and location play a slight role in impacting the traffic delays that could be expected. However, because we saw low overall testing scores for these models, we understand that there are high levels of randomness and possible other variables that more significantly impact the expected delays in traffic when an accident occurs.

Project Motivation/Background

Our project motivation came from our interest in the root causes of car crashes throughout the US and which ones disrupted traffic the most. Specifically,

we were interested in data from our home state of Pennsylvania. Car crashes are an important topic to study because they can result in significant vehicle damage, traffic congestion as well as being the cause of injuries and fatalities. Specifically from our data, we will be analyzing the distance of backup from accidents which will help us to better interpret what factors are causing the most severe crashes. Our data ranks the severity of car crashes from 1-4 which indicates the level of traffic impact that the accident caused. While our data does not include whether each accident resulted in injuries or fatalities, it is logical to assume that accidents with such casualties would be ranked at higher levels of severity. We want to create diagrams that will generate results to help drivers understand the causes of accidents and better understand the risks to consider before/when driving. Our hope is that if more information is found and emphasized related to car crashes, the bad drivers of PA will better understand some of the leading causes of car accidents. Results from such findings could be used by many different organizations who work to make Pennsylvania roads safer. For example, local news stations or municipal officials could talk about a finding that many severe crashes occur when there is light freezing rain so that people understand the risk of driving in such weather. There could be tests done that could be used by PennDot such as which sections of interstate highways see the most crashes.

Data Description

The dataset looks at US Accident Reports from 2016-2021 in the United States. Our original set had information for 49 States and over 2.7 million records listed. Multiple APIs are recorded for each incident using traffic cameras, police

reports, and traffic sensors to gather all the data. This is then reported to the Department of Transportation, and ultimately seen within our dataset then.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

Both of these sources are also cited for our dataset. Sobhan Moosavi, a data scientist at Lyft, was the creator of the final dataset that we are using. The variables and their descriptions were posted along with the dataset and can be seen below. Two of the main variables we plan to target are the severity and the distance. Both of these variables we feel are two of the more important factors in determining the significance of the crash. By targeting these variables, we can then use all of the other various columns to determine what is most key in influencing these car crashes so we can better predict and prepare around them.

We also consolidated the original dataset down to being just incidents in Pennsylvania. The 2.7 million rows of data proved to be extremely difficult to clean, organize, and run within SaS; so we ultimately decided to just focus within our home state.

We also partook in other data cleaning steps within the original set so that we could get the most accurate results out of our project. Slight differences in spellings, capitalizations, and abbreviations, were addressed at this stage of the project.

Variables

#	Attribute	Description	Variable Insight (range, scale)	Nullable
1	ID	This is a unique identifier of the accident record.	Range: A-1 to A-41849 Scale: The ID is just an identifying number assigned to the crash, there is no scale.	No
2	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).	Range: 1-4 Scale: 1 Least significant delay in traffic (minor crash) 4 Most significant delay in traffic (major crash, heavy delays in traffic, standstill, long recovery time)	No
3	Start_Time	Shows start time of the accident in local time zone.	Range: Date -: 1/1/2016 - 12/31/2021 Time: 00:00 - 23:59 Scale Date - there is no scale for the date as it is just an identifier to when the accident occurred. Scale Time - 0000 represents 12AM, signifying the accident began in the early hours of the morning. 23:59 represents 11:59PM. As the numerical value increases, it means the accident ended later on in the day.	No
4	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	Range: Date -: 1/1/2016 - 12/31/2021 Time: 00:00 - 23:59 Scale Date - there is no scale for the date as it is just an identifier to when the accident occurred. Scale Time - 0000 represents 12AM, signifying the accident began in the early hours of the morning. 23:59 represents 11:59PM. As the numerical value increases, it means the accident ended later on in the day.	No
5	Start_Lat	Latitude represents distance between north and south of the equator. Latitude in this data set shows latitude in GPS coordinate of the start point of the crash. Latitude is measured from (-)90° to (+)90°, the closer to (-)90° means coordinates distance is located closer to the south. 0° represents the	Range: 39.72126 to 42.252765. Our latitude is only ranged between these numbers rather than (+,-)90° because our data is only for accidents located in PA Scale: The closer to 39° a latitude is, means the accident was located closer to southern PA. The	No

		equator line. The closer a latitude is (+)90° represents that the distance is located in the north	closer to 42° a latitude is, means that the accident was closer to northern PA.	
6	Start_Lng	Longitude represents distance between east and west of the equator. In this data set starting longitude in GPS coordinate represents of the start point of the crash. Longitude is measured from (-)180° to (+)180°, with a coordinate that is negative, the closer to (-)90° a coordinate is, represents the location is more west. The closer longitude is (+)90° represents that the location is more east.	Range: (-)74.709 to (-)80.59. Our longitude is only ranged between these numbers rather than (+,-)180° because our data is only for accidents located in PA Scale: The closer to (-)74° a longitude is, means the accident was located closer to eastern PA. The closer to (-)80° a longitude is, means that the accident was closer to western PA.	No
7	End_Lat	Latitude represents distance between north and south of the equator. Latitude in this data set shows latitude in GPS coordinate of the end point of the crash. Latitude is measured from (-)90° to (+)90°, the closer to (-)90° means coordinates distance is located closer to the south. 0° represents the equator line. The closer a latitude is (+)90° represents that the distance is located in the north	Range: 39.709836 to 42.252751. Our latitude is only ranged between these numbers rather than (+,-)90° because our data is only for accidents located in PA Scale: The closer to 39° a latitude is, means the accident was located closer to southern PA. The closer to 42° a latitude is, means that the accident was closer to northern PA.	Yes
8	End_Lng	Longitude represents distance between east and west of the equator. In this data set starting longitude in GPS coordinate represents of the start point of the crash. Longitude is measured from (-)180° to (+)180°, with a coordinate that is negative, the closer to (-)90° a coordinate is, represents the location is more west. The closer longitude is (+)90° represents that the location is more east.	Range: (-)74.708982 to (-)80.57173. Our longitude is only ranged between these numbers rather than (+,-)180° because our data is only for accidents located in PA Scale: The closer to (-)74° a longitude is, means the accident was located closer to eastern PA. The closer to (-)80° a longitude is, means that the accident was closer to western PA.	Yes
9	Distance(mi)	The length of the road extent affected by the accident.	Range: 0 to 44.132 Scale: 0 represents that the road was not affected at all, as that number increases, it means that a higher distance in miles of the road were affected	No
10	Description	Shows natural language description of the accident.	Range & Scale - There is no range or scale for this as all descriptions vary	No

11	Number	Shows the street number in address field.	Range: 1 to 48,974 Scale: There is no scale, Pennsylvania street numbers have no correlation to geographic location, they are just identifiers	Yes
12	Street	Shows the street name in address field.	Range: Various Scale: There is no scale, Pennsylvania street numbers have no correlation to geographic location, they are just identifiers	Yes
13	Side	Shows the relative side of the street (Right/Left) in address field.	Range: L & R Scale: L indicates a crash on the left side of the street, R indicates a crash on the right side of the street	Yes
14	City	Shows the city name of the location of the crash in address field.	Range: All cities located within the state of Pennsylvania (2,568 cities total) Scale: There is no scale, as the city name is just an identifier	Yes
15	County	Shows the county of the accident in address field.	Range: US Scale: There is no scale, as all accidents included occurred in the US	Yes
16	State	Shows the state in address field.	Range: PA Scale: There is no scale, as all accidents included occurred in PA	Yes
17	Zipcode	Shows the zipcode in address field.	Range: 15001 to 19611	Yes
18	Country	Shows the country in address field.	Range: US	Yes
19	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	Range: US/Eastern	Yes
20	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	Form: KPNE	Yes
21	Weather_Timestamp	Shows the time-stamp of the weather observation record (in local time).	Form: 3/25/2021 5:17	Yes

22	Temperature(F)	Shows the temperature (in Fahrenheit).	This gauges the temperature at the time of the accident. The range had a maximum of 98 degrees fahrenheit and a minimum of 1 degree fahrenheit.	Yes
23	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	This gauged the wind chill during the time of the accident. The range had a maximum of 98 degrees fahrenheit and a minimum of -12.2 degrees fahrenheit.	Yes
24	Humidity(%)	Shows the humidity (in percentage).	This gauged the humidity during the time of the accident. The maximum in the range was 100% and a minimum of 10%.	Yes
25	Pressure(in)	Shows the air pressure (in inches).	This accounts for the air pressure during the time of the accident. The range had a maximum of 56.54 inches and a minimum of 26.76 inches.	Yes
26	Visibility(mi)	Shows visibility (in miles).	For visibility it accounts for the visibility at the time of the accident. The range had a maximum of 20 miles and a minimum of 0 miles.	Yes
27	Wind_Direction	Shows wind direction.	This shows the wind direction during the accident. The range is in the form of the cardinal direction symbols (N,S,E,W) and sub directions like south south west (SSW).	Yes
28	Wind_Speed(mph)	Shows wind speed (in miles per hour).	The highest wind speed recorded was 38 mph and the lowest was 0 mph. These indicated the wind speed during the accident.	Yes
29	Precipitation(in)	Shows precipitation amount in inches, if there is any.	The amount of precipitation varied a lot, but the lowest amount was 0 inches and the highest amount was 2.42 inches.	Yes
30	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	Variables: Light Snow, Overcast, Clear, Raining, etc.	Yes
31	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	Range: 2 (True or False)	No
32	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	Range: 2 (True or False)	No
33	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	Range: 2 (True or False)	No

34	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	Range: 2 (True or False)	No
35	Junction	A POI annotation which indicates presence of junction in a nearby location.	Range: 2 (True or False)	No
36	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	Range: 2 (True or False)	No
37	Railway	A POI annotation which indicates presence of railway in a nearby location.	Range: 2 (True or False)	No
38	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	Range: 2 (True or False)	No
39	Station	A POI annotation which indicates presence of station in a nearby location.	Range: 2 (True or False)	No
40	Stop	A POI annotation which indicates presence of stop in a nearby location.	Range: 2 (True or False)	No
41	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	Range: 2 (True or False)	No
42	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.	Range: 2 (True or False)	No
43	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	Range: 2 (True or False)	No
44	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	Range: 2 (Day or Night)	Yes
45	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	Range: 2 (Day or Night)	Yes
46	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	Range: 2 (Day or Night)	Yes
47	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	Range: 2 (Day or Night)	Yes

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Amenity	INPUT	0.01619	0.126208	16739	1	0	0	1	7.667759	56.80131
Crossing	INPUT	0.065297	0.247056	16739	1	0	0	1	3.51949	10.38805
Distance_mi_	INPUT	0.975537	1.643074	16739	1	0	0.406	30.513	5.134041	46.59298
End_Lat	INPUT	40.36646	0.459439	16739	1	39.71433	40.20926	42.25174	1.394674	1.582793
End_Lng	INPUT	-76.7096	1.631433	16739	1	-80.5218	-76.2478	-74.7128	-0.88017	-0.51937
Give_Way	INPUT	0.011948	0.108656	16739	1	0	0	1	8.984518	78.73097
Junction	INPUT	0.078619	0.269151	16739	1	0	0	1	3.131563	7.807622
No_Exit	INPUT	0.001673	0.040866	16739	1	0	0	1	24.39118	593.0006
Number	INPUT	2170.202	3603.25	8926	7814	1	961	48974	4.387692	30.22338
Precipitation_in_	INPUT	0.006462	0.044943	14377	2363	0	0	1.67	16.49579	387.2133
Pressure_in_	INPUT	29.42895	0.668238	16480	260	26.76	29.59	56.54	3.173518	164.4584
Railway	INPUT	0.007886	0.088454	16739	1	0	0	1	11.12838	121.8553
Start_Lat	INPUT	40.36647	0.459391	16739	1	39.72126	40.20718	42.23135	1.395395	1.586287
Start_Lng	INPUT	-76.7094	1.631487	16739	1	-80.5191	-76.2484	-74.7211	-0.88031	-0.52022
Station	INPUT	0.006333	0.079327	16739	1	0	0	1	12.44786	152.9675
Stop	INPUT	0.041102	0.198531	16739	1	0	0	1	4.623487	19.37895
Temperature_F_	INPUT	58.83783	17.38892	16472	268	3.9	60	98	-0.14917	-0.88388
Traffic_Signal	INPUT	0.097377	0.29648	16739	1	0	0	1	2.716345	5.379173
VAR24	INPUT	65.70478	20.36973	16452	288	11	66	100	-0.12069	-0.98495
Visibility_mi_	INPUT	9.122044	2.431373	16377	363	0	10	20	-1.71752	5.711737
Wind_Chill_F_	INPUT	56.5996	19.68888	14635	2105	-8.1	59	98	-0.29709	-0.79098
Wind_Speed_mph_	INPUT	7.027862	5.062627	16083	657	0	6.9	38	0.810257	1.445534
Severity	TARGET	2.200012	0.595519	16740	0	2	2	4	2.666646	5.15379

FIGURE 1

Data Preparation Activities

We took part in several activities to prepare our data before creating models with SaS. We first filtered our data down to include only data from the state of Pennsylvania in Excel. Our original data set included more entries than a standard excel sheet could handle. By narrowing it down to being only PA car crash incidents, we were able to reduce our run time of models and also specify our research to the region that we all live in currently.

Our original dataset also had given a variable list which stated if it was possible for each of the variables to have null values or not. We left missing variables in our dataset for the variable categories that were considered “nullable”,

meaning that null values could be handled. We made sure to go through the missing values within variables that were considered non-nullable and remove all the rows from our data. We did this because having missing values that are not able to be ignored would have left noise, meaning less data within these categories which would have made our results inaccurate.

Our next step involved importing the file into SaS. We created a new model and imported the Excel file with PA car crashes. We then created a diagram and used the File Import node to insert the PA car crashes data file. We used Edit Variables in order to create the appropriate Role and Level for each variable within our dataset. We assigned Distance_mi_ as our target variable which is the distance in miles that traffic was backed up during the car crash. We rejected the corresponding variable of Severity since this variable was related to Distance_mi_ and the severity of the accident would only be known afterwards given the distance_mi_ tha traffic was backed up. For this reason we needed to reject this variable to allow our model to use unknown variables to try and create a predictive model.

We initially ran a Stat Explorer and a Regression to better interpret our data. SaS failed to generate output for the Regression and displayed an error message pop up saying “target levels exceed 512.” We were able to fix this by going to the File Import node and setting the maximum number of rows to import to 1,000,000 and the maximum number of the columns to 10,000.

First Step: Cluster Analysis

Our first step that we ran was a cluster analysis. Using the cluster analysis we hoped to identify what were some of the possible common combinations of variables that could be seen in creating groups for our average distance that traffic was getting delayed per accident. We ended up using the clustering method average for our clustering node, as we felt it ultimately produced some of the better groupings compared to the other methods we tried. The cluster ultimately analyzed a specific group highlighted in FIGURE 2 that constantly recorded a higher distance of delayed traffic. Analyzing this cluster vs. the others we then discovered a few common trends within this particular set.

- More significant incidents occurred in the daytime hours. (Noted: This was the most common time for crashes overall in our dataset, leading us to believe if there was more travel occurring during these times and consequently traffic would back up further during these times as more cars would be on the road.)
- An accident is more likely to be significant when the sun is setting or rising.
- An accident is more likely to be significant when the weather condition is foggy or during a thunderstorm.

Some of these facts can be backed up by looking at our variables list and seeing each variable's significance. The top two variables of highest significance is the distance_mi, representing the distance traffic got backed up from, and also the

true/false sunrise_sunset category. Between these two variables having the highest importance levels we can see that they are accurate predictors for determining the severity of incidents.

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
Distance_mi_		2	8	1.00000
Sunrise_Sunset	Sunrise_Sunset	0	2	0.86592
Pressure_in_		2	4	0.73157
Civil_Twilight	Civil_Twilight	1	0	0.70890
Temperature_F_		3	2	0.68775
Astronomical_Twilight	Astronomical_Twilight	1	3	0.68010
Visibility_mi_		5	3	0.62379
Weather_Condition	Weather_Condition	1	7	0.62299
Give_Way	Give_Way	1	0	0.17916
Precipitation_in_		4	2	0.15029

Frequency of Cluster ▼	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Distance_mi_	Precipitation_in_	Pressure_in_	Temperature_F_	Visibility_mi_
10635	0.168607	17.51297	6	1.772007	0.455364	.0004731	29.67635	76.21837	9.80038
9958	0.293106	14.97431	6	3.485089	0.755494	.0007814	29.44101	50.39643	9.578563
7790	0.183498	12.66587	1	1.772007	0.460964	.0004601	29.76023	45.61048	9.752142
4733	0.237715	14.56003	1	2.21747	0.633576	0.000688	28.4535	60.62519	9.668727
2977	0.2801	13.08783	1	2.885551	3.300992	0.000572	29.53975	63.59457	9.67664
2297	0.580904	30.34441	6	4.313358	0.864281	0.033475	29.27987	54.40227	3.851199
460	0.415146	30.31451	6	9.401933	0.801993	0.002095	29.44022	56.74573	9.281938
157	.	0.432349	4	28.83545	0.068688	1.26E-18	30.14	75	3
66	.	3.962106	4	29.2965	0.254652	0.113333	29.63231	60.38	5.4
44	.	3.469149	4	28.77976	1.552159	0.005	29.92615	34.15	3.45
40	.	5.410877	3	20.4234	3.84905	1.48E-19	29.87875	20.68571	8.571429
23	.	2.85324	4	28.84492	0.615348	1.48E-19	28.22	58.5	3
23	.	0.706679	3	28.77957	2.29687	1.48E-19	29.095	64	5

FIGURE 2

Models

After running this cluster we then looked to input these same exact variables into different models to determine which one could predict the severity and distance that traffic was getting delayed to the best degree. Originally we only wanted to input information on weather and the time of the incident. We felt that these were very standard measurements that could be easily collected and used as inputs into a model. By creating a model that simply uses variables that can easily be recorded, it would allow for our model to then be used by public officials in advising roadway safety protocols, without the need for a complex listing of many variables. The model could then describe the risk of a particular time and weather conditions at any given point in time if accurate. This being our ultimate goal is what led us to this decision. The variables we Imputed and Rejected and Targeted can be seen in FIGURE 3 below.

From here, we ran three separate tests to determine which one could take our imputed variables to respond with the highest predictive accuracy on the severity of a car crash. The three tests that we tried out were a Neural Network, Decision Tree, and Regression. Using a model comparison Node, we found our decision tree to be able to accurately predict the distance of a crash using the imputed time and weather variables better than the other model types. This was due to it having the lowest average squared error measuring of all three models being at 2.40 compared to 2.42 for our neural network, and 2.43 for our regression model.

Now with an understanding that our decision tree was likely the model of choice given our inputs, we began to test different decision trees against each other to determine the best setup for our model. We adjusted the maximum number of

branches and also the maximum depth for each branch. With playing around with these different numbers we landed on our most successful model being that of a maximum branch of 3 and a maximum depth of 6. This particular decision tree model resulted in us getting a 2.37 average squared error, which ultimately proved to be our best model overall.

From here we wished to test out models that included all the variables that were given. We wished to see if things like the location of the incident could also play any role in being able to determine how backed up traffic got. We once again ran a group of different models with the attempt at comparing them to determine the best fitting model for our data then. This time once again running a Decision tree, Neural Network, Regression model, and an additional MBR test, our decision tree came out with the highest rated score for all models. The variables included for this test can be seen below in FIGURE 4

When we took this Decision Tree and compared it with our original model, it actually came out to be less accurate than the original model despite all the additional inputs. We found this to be quite interesting and surprising as we would have naturally thought that having more inputs would result in better models. This did not end up being the case however, which somewhat makes sense given that many of the alternative inputs were so specific and based on distinct locations that the model may find it difficult to accurately predict future occurrences using such specific inputs. This was also good news for us, as it meant that simply knowing the time and weather could help in predicting the severity of a crash, as these inputs are very easy to record and actually implement in a hypothetical real life use of this model.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Airport_Code	Rejected	Nominal	No		No	.	.
Amenity	Rejected	Interval	No		No	.	.
Astronomical_Tw	Input	Binary	No		No	.	.
Bump	Rejected	Binary	No		No	.	.
City	Rejected	Nominal	No		No	.	.
Civil_Twilight	Input	Binary	No		No	.	.
Country	Rejected	Nominal	No		No	.	.
County	Rejected	Nominal	No		No	.	.
Crossing	Rejected	Binary	No		No	.	.
Description	Rejected	Nominal	No		No	.	.
Distance_mi_	Target	Interval	No		No	.	.
End_Lat	Rejected	Interval	No		No	.	.
End_Lng	Rejected	Interval	No		No	.	.
End_Time	Time ID	Interval	No		No	.	.
Give_Way	Input	Binary	No		No	.	.
VAR24	Rejected	Interval	No		No	.	.
ID	ID	Nominal	No		No	.	.
Junction	Rejected	Binary	No		No	.	.
Nautical_Twilight	Rejected	Binary	No		No	.	.
No_Exit	Rejected	Binary	No		No	.	.
Number	Rejected	Interval	No		No	.	.
Precipitation_in_	Input	Interval	No		No	.	.
Pressure_in_	Input	Interval	No		No	.	.
Railway	Rejected	Binary	No		No	.	.
Roundabout	Rejected	Binary	No		No	.	.
Severity	Rejected	Interval	No		No	.	.
Side	Rejected	Binary	No		No	.	.
Start_Lat	Rejected	Interval	No		No	.	.
Start_Lng	Rejected	Interval	No		No	.	.
Start_Time	Time ID	Interval	No		No	.	.
State	Rejected	Nominal	No		No	.	.
Station	Rejected	Binary	No		No	.	.
Stop	Rejected	Binary	No		No	.	.
Street	Rejected	Nominal	No		No	.	.
Sunrise_Sunset	Input	Binary	No		No	.	.
Temperature_F_	Input	Interval	No		No	.	.
Timezone	Rejected	Nominal	No		No	.	.
Traffic_Calming	Rejected	Binary	No		No	.	.
Traffic_Signal	Rejected	Binary	No		No	.	.
Turning_Loop	Rejected	Binary	No		No	.	.
Visibility_mi_	Input	Interval	No		No	.	.
Weather_Condit	Input	Nominal	No		No	.	.
Weather_Timest	Rejected	Interval	No		No	.	.
Wind_Chill_F_	Rejected	Interval	No		No	.	.
Wind_Direction	Rejected	Nominal	No		No	.	.
Wind_Speed_mp	Rejected	Interval	No		No	.	.
Zipcode	Rejected	Nominal	No		No	.	.

FIGURE 3

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Airport_Code	Rejected	Nominal	No		No	.	.
Amenity	Rejected	Interval	No		No	.	.
Astronomical_Twilight	Input	Binary	No		No	.	.
Bump	Input	Binary	No		No	.	.
City	Text Location	Nominal	No		No	.	.
Civil_Twilight	Input	Binary	No		No	.	.
Country	Rejected	Nominal	No		No	.	.
County	Input	Nominal	No		No	.	.
Crossing	Input	Binary	No		No	.	.
Description	Text	Nominal	No		No	.	.
Distance_mi_	Prediction	Interval	No		No	.	.
End_Lat	Rejected	Interval	No		No	.	.
End_Lng	Rejected	Interval	No		No	.	.
End_Time	Time ID	Interval	No		No	.	.
Give_Way	Input	Binary	No		No	.	.
ID	ID	Nominal	No		No	.	.
Junction	Input	Binary	No		No	.	.
Nautical_Twilight	Input	Binary	No		No	.	.
No_Exit	Input	Binary	No		No	.	.
Number	Rejected	Interval	No		No	.	.
Precipitation_in_	Input	Interval	No		No	.	.
Pressure_in_	Input	Interval	No		No	.	.
Railway	Input	Binary	No		No	.	.
Roundabout	Input	Binary	No		No	.	.
Severity	Target	Interval	No		No	.	.
Side	Input	Binary	No		No	.	.
Start_Lat	Rejected	Interval	No		No	.	.
Start_Lng	Rejected	Interval	No		No	.	.
Start_Time	Time ID	Interval	No		No	.	.
State	Rejected	Nominal	No		No	.	.
Station	Input	Binary	No		No	.	.
Stop	Input	Binary	No		No	.	.
Street	Text	Nominal	No		No	.	.
Sunrise_Sunset	Input	Binary	No		No	.	.
Temperature_F_	Input	Interval	No		No	.	.
Timezone	Rejected	Nominal	No		No	.	.
Traffic_Calming	Input	Binary	No		No	.	.
Traffic_Signal	Input	Binary	No		No	.	.
Turning_Loop	Input	Binary	No		No	.	.
VAR24	Rejected	Interval	No		No	.	.
Visibility_mi_	Input	Interval	No		No	.	.
Weather_Condition	Input	Nominal	No		No	.	.
Weather_Timestamp	Time ID	Interval	No		No	.	.
Wind_Chill_F_	Input	Interval	No		No	.	.
Wind_Direction	Input	Nominal	No		No	.	.
Wind_Speed_mph_	Input	Interval	No		No	.	.
Zipcode	Rejected	Nominal	No		No	.	.

FIGURE 4

Through evaluating alternative tests and clusters we also came across an interesting grouping done by one of our cluster tests. A separate cluster analysis prior to running all of our variables as inputs created a group of accidents that occurred during a weather condition of light freezing drizzle. This particular group averaged a very large average distance of delayed traffic of 3.8 miles. Using this clustering node, we were able to now see the importance of some of the additional variables when determining the distance traffic might get delayed. What this meant was that the weather condition variable itself wasn't necessarily all that significant in each and every test, however, some of the particular weather condition variable options within this column could be. This same statement could also be true of some of the other variables that are listed within the dataset. Example of this can be seen in the highlighted row of FIGURE 5

Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Distance_mi_	Weather_Condition=Light Freezing Drizzle
43	0.276777	6.173319	2	20.93151	3.822279	1
45	.	4.289197	2	29.04488	1.965022	-6.2E-22
2576	0.387606	30.59291	2	3.944778	1.129694	2.64E-18
2848	0.46056	30.72449	2	6.359271	1.072763	3.33E-18
110	0.329608	4.740848	2	20.89763	0.986509	-6.2E-22
418	0.521909	29.86617	2	9.611634	0.808617	-1.6E-19
29252	0.403005	29.97952	5	3.944778	0.757811	-2E-17
33	0.290687	4.743099	2	11.92014	0.69497	-6.2E-22
11	0.32626	4.753838	7	17.12548	0.612091	-6.2E-22
1532	0.424952	11.50776	2	5.019017	0.609726	-5.6E-19
61	0.109366	5.576854	2	17.62873	0.467721	-6.2E-22
25	.	6.676665	2	29.23228	0.311	-6.2E-22
173	0.23069	5.515512	2	16.83274	0.26526	-6.2E-22
43	0.391397	6.327376	2	15.03712	0.202791	-6.2E-22

FIGURE 5

Due to this fact, we then decided to look further into our regression model and each specific response a variable could be. In doing so we could determine which of the variables did end up being significant vs. those that were not. These variables are laid out in the following three charts labeled FIGURE 6.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	-0.8850	1.4206	-0.62	0.5333
Astronomical_Twilight	Day	1	0.00915	0.0309	0.30	0.7670
Civil_Twilight	Day	1	-0.0208	0.0369	-0.56	0.5736
County	Adams	1	-0.5578	0.1332	-4.19	<.0001
County	Allegheny	1	-0.2271	0.0588	-3.86	0.0001
County	Armstrong	1	0.0540	0.3127	0.17	0.8629
County	Beaver	1	0.4042	0.3128	1.29	0.1963
County	Bedford	1	-0.3234	0.1179	-2.74	0.0061
County	Berks	1	0.1252	0.0897	1.40	0.1626
County	Blair	1	-0.5091	0.0914	-5.57	<.0001
County	Bradford	1	-0.3651	0.2399	-1.52	0.1280
County	Bucks	1	-0.0504	0.1024	-0.49	0.6224
County	Butler	1	0.8464	0.2250	3.76	0.0002
County	Cambria	1	-0.6290	0.1108	-5.68	<.0001
County	Cameron	1	-0.0839	0.4075	-0.21	0.8368
County	Carbon	1	3.1730	0.3487	9.10	<.0001
County	Centre	1	0.0267	0.0818	0.33	0.7442
County	Chester	1	-0.6890	0.0561	-12.27	<.0001
County	Clarion	1	0.3244	0.1967	1.65	0.0991
County	Clearfield	1	-0.2447	0.0963	-2.54	0.0111
County	Clinton	1	0.4199	0.1478	2.84	0.0045
County	Columbia	1	0.1274	0.1326	0.96	0.3369
County	Crawford	1	1.2675	0.4066	3.12	0.0018
County	Cumberland	1	0.0806	0.0913	0.88	0.3773
County	Dauphin	1	-0.1757	0.0828	-2.12	0.0338
County	Delaware	1	-0.3869	0.0805	-4.80	<.0001
County	Elk	1	0.1359	0.1519	0.89	0.3709
County	Erie	1	-0.3298	0.1775	-1.86	0.0633
County	Fayette	1	-0.2588	0.2747	-0.94	0.3461
County	Forest	1	0.5099	0.7398	0.69	0.4907
County	Franklin	1	0.5396	0.1383	3.90	<.0001
County	Fulton	1	-0.2096	0.1652	-1.27	0.2044
County	Greene	1	-0.2492	0.3433	-0.73	0.4678
County	Huntingdon	1	-0.3868	0.1391	-2.78	0.0054
County	Indiana	1	-0.00740	0.2535	-0.03	0.9767
County	Jefferson	1	0.3551	0.1625	2.19	0.0289
County	Juniata	1	0.2354	0.1631	1.44	0.1489
County	Lackawanna	1	0.3109	0.1667	1.86	0.0622
County	Lancaster	1	-0.7409	0.0573	-12.92	<.0001
County	Lawrence	1	-0.5659	0.1687	-3.35	0.0008
County	Lebanon	1	-0.0549	0.1297	-0.42	0.6721
County	Lehigh	1	0.1844	0.0978	1.89	0.0593
County	Luzerne	1	0.9279	0.1368	6.78	<.0001
County	Lycoming	1	-0.5845	0.0907	-6.44	<.0001
County	Mckean	1	-0.3389	0.1633	-2.08	0.0380
County	Mercer	1	1.1825	0.2485	4.76	<.0001
County	Mifflin	1	-0.1108	0.1247	-0.89	0.3744

155	County	Monroe	1	0.0965	0.1655	0.58	0.5599
156	County	Montgomery	1	-0.8910	0.0548	-16.25	<.0001
157	County	Montour	1	-0.4504	0.2233	-2.02	0.0437
158	County	Northampton	1	0.2676	0.1112	2.41	0.0162
159	County	Northumberland	1	-0.7161	0.1096	-6.53	<.0001
160	County	Perry	1	-0.0582	0.1768	-0.33	0.7420
161	County	Philadelphia	1	-0.2319	0.0753	-3.08	0.0021
162	County	Pike	1	0.2993	0.1689	1.77	0.0764
163	County	Potter	1	-0.2926	0.3268	-0.90	0.3707
164	County	Schuylkill	1	0.4676	0.1674	2.79	0.0052
165	County	Snyder	1	-0.5305	0.1269	-4.18	<.0001
166	County	Somerset	1	-0.4333	0.1207	-3.59	0.0003
167	County	Sullivan	1	-0.4518	0.3712	-1.22	0.2235
168	County	Susquehanna	1	0.5833	0.2133	2.74	0.0062
169	County	Tioga	1	-0.4067	0.1369	-2.97	0.0030
170	County	Union	1	-0.1361	0.1306	-1.04	0.2976
171	County	Venango	1	0.8253	0.2087	3.95	<.0001
172	County	Warren	1	-0.6994	0.7404	-0.94	0.3449
173	County	Washington	1	0.1375	0.1392	0.99	0.3232
174	County	Wayne	1	0.4809	0.4072	1.18	0.2376
175	County	Westmoreland	1	-0.4737	0.0825	-5.74	<.0001
176	County	Wyoming	1	-0.3995	0.4277	-0.93	0.3503
177	Crossing	0	1	0.1021	0.0270	3.78	0.0002
178	Give_Way	0	1	-0.0158	0.0580	-0.27	0.7853
179	Junction	0	1	-0.0419	0.0249	-1.69	0.0917
180	Nautical_Twilight	Day	1	0.00880	0.0372	0.24	0.8133
181	Precipitation_in_		1	0.00346	1.1276	0.00	0.9975
182	Pressure_in_		1	0.0639	0.0487	1.31	0.1896
183	Side	L	1	-0.1941	0.0144	-13.49	<.0001
184	Stop	0	1	0.1134	0.0287	3.95	<.0001
185	Sunrise_Sunset	Day	1	0.0182	0.0286	0.64	0.5247
186	Temperature_F_		1	-0.00501	0.00721	-0.70	0.4869
187	Traffic_Signal	0	1	0.1789	0.0215	8.34	<.0001
188	Visibility_mi_		1	0.00172	0.00911	0.19	0.8502
189	Weather_Condition	Clear	1	-3.4909	1.3359	-2.61	0.0090
190	Weather_Condition	Cloudy	1	-0.1261	0.1241	-1.02	0.3096
191	Weather_Condition	Cloudy / Windy	1	-0.0430	0.2972	-0.14	0.8851
192	Weather_Condition	Drizzle	1	4.6938	0.9045	5.19	<.0001
193	Weather_Condition	Fair	1	-0.0720	0.1239	-0.58	0.5608
194	Weather_Condition	Fair / Windy	1	-0.3289	0.2365	-1.39	0.1643
195	Weather_Condition	Fog	1	-0.00576	0.1859	-0.03	0.9753
196	Weather_Condition	Haze	1	-0.1264	0.1718	-0.74	0.4619
197	Weather_Condition	Haze / Windy	1	-0.5077	1.2740	-0.40	0.6903
198	Weather_Condition	Heavy Rain	1	0.3977	0.2651	1.50	0.1336
199	Weather_Condition	Heavy Rain / Windy	1	-0.0988	1.2686	-0.08	0.9379
200	Weather_Condition	Heavy T-Storm	1	0.6298	0.5817	1.08	0.2790
201	Weather_Condition	Heavy Thunderstorms and Rain	0	0	.	.	.
202	Weather_Condition	Light Drizzle	1	-0.2862	0.2515	-1.14	0.2551
203	Weather_Condition	Light Freezing Drizzle	1	2.3504	1.2696	1.85	0.0642
204	Weather_Condition	Light Freezing Rain	1	-0.1222	0.4997	-0.24	0.8068
205	Weather_Condition	Light Rain	1	-0.0395	0.1293	-0.31	0.7598

FIGURE 6b

206	Weather_Condition	Light Rain / Windy	1	-0.4780	0.4202	-1.14	0.2554
207	Weather_Condition	Light Rain with Thunder	1	-0.3739	0.2789	-1.34	0.1801
208	Weather_Condition	Light Snow	1	0.5331	0.1712	3.11	0.0018
209	Weather_Condition	Light Snow / Windy	1	-0.4816	1.3137	-0.37	0.7139
210	Weather_Condition	Light Thunderstorms and Rain	0	0	.	.	.
211	Weather_Condition	Mostly Cloudy	1	-0.0508	0.1265	-0.40	0.6881
212	Weather_Condition	Mostly Cloudy / Windy	1	-0.4140	0.3893	-1.06	0.2876
213	Weather_Condition	N/A Precipitation	1	-0.3446	0.9018	-0.38	0.7024
214	Weather_Condition	Overcast	1	-0.3706	0.2999	-1.24	0.2166
215	Weather_Condition	Partly Cloudy	1	-0.0940	0.1281	-0.73	0.4631
216	Weather_Condition	Partly Cloudy / Windy	1	-0.6270	0.3883	-1.61	0.1064
217	Weather_Condition	Patches of Fog	1	-0.7946	1.2695	-0.63	0.5314
218	Weather_Condition	Rain	1	0.0232	0.1651	0.14	0.8883
219	Weather_Condition	Rain / Windy	1	-0.2790	0.9036	-0.31	0.7575
220	Weather_Condition	Scattered Clouds	1	-0.3927	1.2722	-0.31	0.7576
221	Weather_Condition	Showers in the Vicinity	1	-0.0989	0.4208	-0.24	0.8141
222	Weather_Condition	Snow	1	1.0703	1.2748	0.84	0.4012
223	Weather_Condition	T-Storm	1	0.0900	0.2345	0.38	0.7010
224	Weather_Condition	Thunder	1	0.1366	0.3023	0.45	0.6514
225	Weather_Condition	Thunder / Windy	1	-0.3976	0.9018	-0.44	0.6593
226	Weather_Condition	Thunder in the Vicinity	1	0.1767	0.2616	0.68	0.4995
227	Weather_Condition	Thunderstorm	0	0	.	.	.
228	Weather_Condition	Thunderstorms and Rain	0	0	.	.	.
229	Wind_Chill_F_		1	0.00559	0.00657	0.85	0.3948
230	Wind_Direction	Calm	1	-0.0631	0.0451	-1.40	0.1619
231	Wind_Direction	E	1	-0.1781	0.0584	-3.05	0.0023
232	Wind_Direction	ENE	1	-0.0159	0.0768	-0.21	0.8355
233	Wind_Direction	ESE	1	-0.0688	0.0735	-0.94	0.3494
234	Wind_Direction	N	1	-0.1535	0.0630	-2.44	0.0148
235	Wind_Direction	NE	1	0.0203	0.0825	0.25	0.8053
236	Wind_Direction	NNE	1	-0.1067	0.0968	-1.10	0.2706
237	Wind_Direction	NNW	1	-0.1069	0.0574	-1.86	0.0625
238	Wind_Direction	NW	1	-0.1343	0.0491	-2.74	0.0062
239	Wind_Direction	S	1	-0.1034	0.0514	-2.01	0.0443
240	Wind_Direction	SE	1	-0.1132	0.0765	-1.48	0.1389
241	Wind_Direction	SSE	1	-0.1172	0.0690	-1.70	0.0896
242	Wind_Direction	SSW	1	-0.2687	0.0531	-5.06	<.0001
243	Wind_Direction	SW	1	-0.2283	0.0507	-4.50	<.0001
244	Wind_Direction	VAR	1	-0.1829	0.0559	-3.27	0.0011
245	Wind_Direction	W	1	-0.2267	0.0420	-5.39	<.0001
246	Wind_Direction	WNW	1	-0.1794	0.0458	-3.92	<.0001
247	Wind_Direction	WSW	1	-0.0687	0.0521	-1.32	0.1872
248	Wind_Speed_mph_		1	0.00718	0.00407	1.76	0.0778

FIGURE 6c

Findings

Our findings conclude that with the variables that we have given our dataset, we were unable to accurately predict the majority reasoning for why traffic gets delayed per accident. There are many different variables that unfortunately did not correlate well to the predictability of these delays in traffic. Some of the reasoning we determined for this was probable non correlation in some of the variables. For example, when we were expecting the possibility for more accidents and them to be more severe, mostly from bad weather, we found that there also were less cars on the roads at these times and therefore less traffic that would get delayed. These contradicting effects made it so that the actual relationship between some of the variables was very difficult to discover. What was concluded was that when there was more traffic on the roads, accidents had a higher likelihood of delaying traffic for longer. This meant most of our higher delaying incidents occurred during the daytime hours. Sunset and Sunrise also generated slightly higher rates of crashes that also resulted in longer delayed traffic.

In the end we found that our decision trees were the best of the models tested as they were able to group certain commonalities of variables together in order to best determine the likelihood of delayed traffic. Things like bad weather paired with times of high traffic would be able to be grouped together in order to predict possible extra delays in traffic. This ability for our decision trees is what made them the best model as they could also piece apart some of the different columns and look at the different combinations of variable listings separately.

Unfortunately, as mentioned before, this model still did not come without a higher level of error at times. This was the case for all our models, and really proved that the variables that we had were not capable of generating a model that could

on its own accurately predict the resulting distance that crashes would back traffic up.

Managerial Implications/Concluding Remarks

As a group we were able to get a few things out of the tests, but the variables just never ended being able to accurately predict what we wanted too to a high degree. We would need to expand our variable base in order to obtain a stronger model and convince policy makers on which variables to address and make changes and improve road safety from. We would need to create a more complex model to get exactly what we need to facilitate any sort of change going forward that could accurately predict traffic delays to a high enough degree. We would still recommend policy makers to look at this report, as it provides interesting insights and analysis on some specific variables, but really highlights that additional/ more complex analysis may need to be done in order to discover a better predicting model. Like we said, many of the variables in the end contradict themselves and left us with a low level of correlation. Grouping these variables through models like decision trees was found to be the strongest method, and would be recommended for moving forward, but understanding the difficulty in predicting delays in traffic from just location, time, and weather proved to be difficult due to high levels of variance in the reports and their traffic backups.

References

Link: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents