Analysis and Way Forward

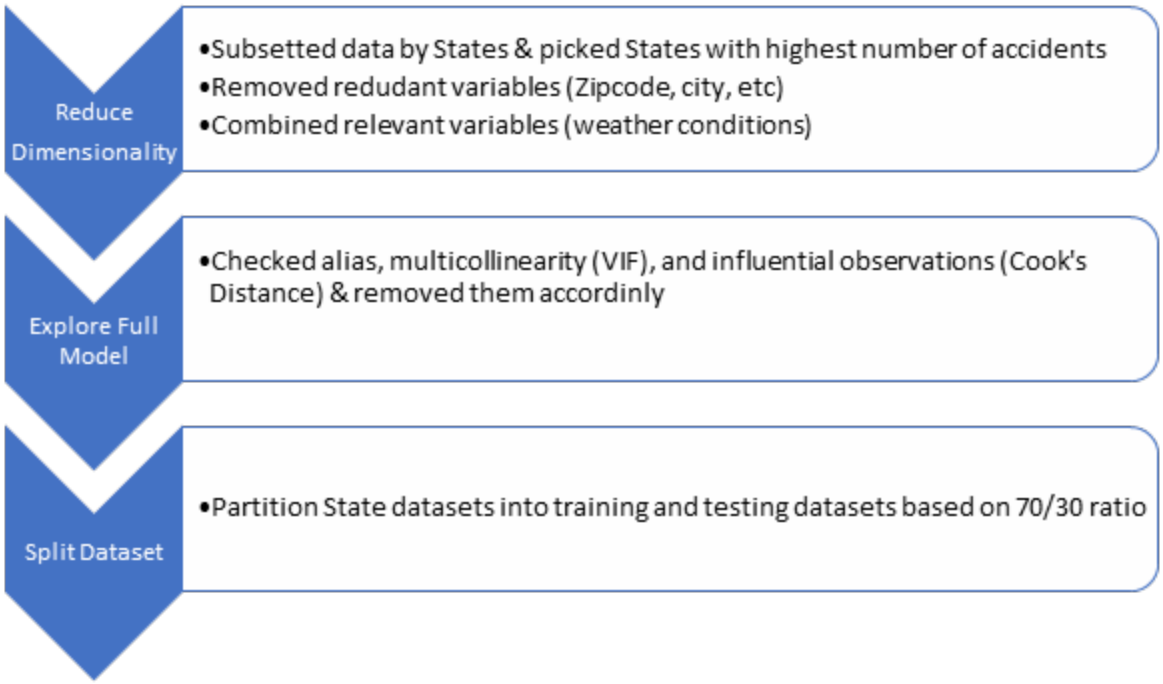
Goals of Analysis

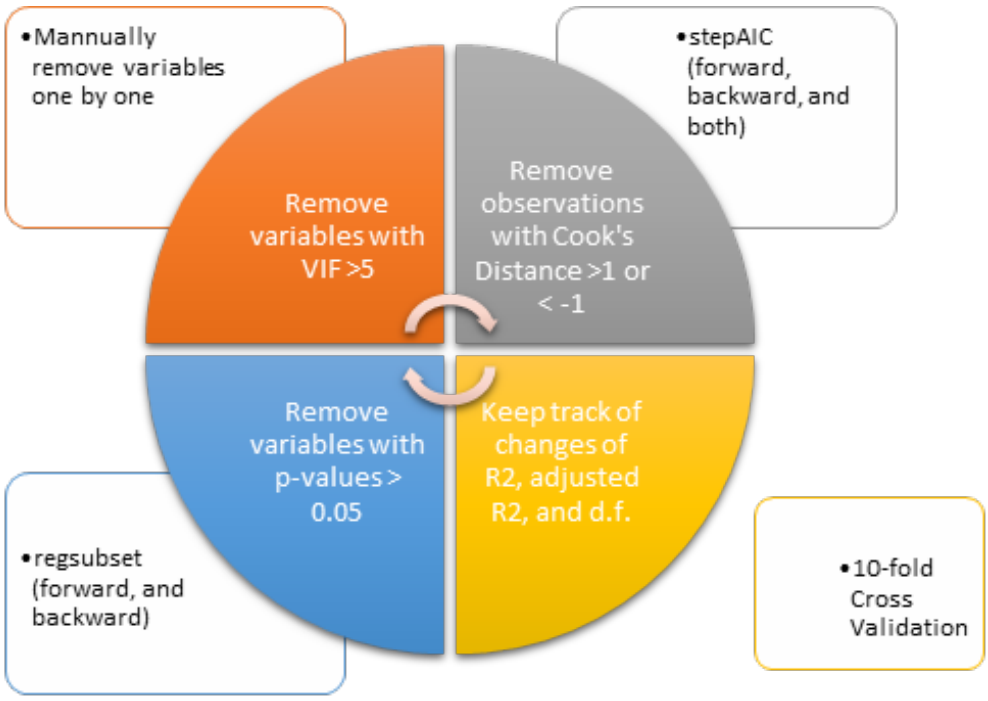
Based on our exploratory data analysis, we identified 'Duration of Impact' (reflected by the difference between 'Start Time' of the accident and 'End Time' of the impact from the original dataset) to be our variable of interest, which I would like to predict.

The goal of this project is to identify factors pertaining to geographical locations, time and weather, and road conditions that are associated with accidents in the United States, in particular California, Texas, Florida, South Carolina, New York, and Pennsylvania. We looked into the three categories of factors figuring out the ones with highest correlation with accidents. Limited by the enormous size of national dataset, we zoomed into the states with highest number of accidents to study the patterns of accidents instead of conducting a national investigation. For example, we explored whether the time and weather; and road conditions factors have different relationships with accidents in states with highest number of accidents in the past four years.

Being able to predict impact of accidents and understand the major drivers of accidents are not only beneficial for accident prevention but also mitigation of the damage caused by accidents. There are many use cases of the potential knowledge generated from this project. For instance, from a city planning perspective, the conclusion of the project will shed light on safe road designs. The algorithm derived from this project could be useful to make accurate prediction of impact duration of accidents. If such information can be broadcasted through radio, GoogleMaps, and other popular applications, this will save the general public from traffic jams.

Methodology





Data Cleaning & Pre-Processing?

Given the variety of variables, we made a huge effort in feature selection. A flowchart (Figure XX - arrow) demonstrates the steps we took to remove variables of less concern or interest. Before adopting any specific measurement tools, we assessed the feasibility of using the variables available as predictor variables based on our understanding of the dataset as well as the techniques that we used. We determined that some variables can be excluded from the regression analysis given the information they carry and the complexity of the variable. For example, ZIP code was eliminated as it would add tremendous complexity to the regression analysis and it had overlapping information with other geographical variables. Other variables removed at the early stage because of redundancy, missingness and complexity included ID, Source, TMC, City, County, State, Country, End latitude, End longitude, Description, (Street) Number, Time zone, Airport Code, Weather Timestamp, Precipitation, Turning Loop (variable only had one unique value), Sunrise/ Sunset, Nautical Twilight, and Astronical Twilight. Additionally, Pearson correlation coefficients were calculated among numeric variables to pinpoint correlated variables. We dropped Wind Chill because it had very high correlation (over 0.7) with Temperature. Wind speed was considered redundant as the data did not include driving speed to allow reasonable interpretation of the data. Other treatments were adopted to certain variables to extract useful info or reduce complexity. Variable Street was turned into Highway/ City variable to show whether the accidents were located on a highway as such information was considered more informational. Weather conditions originally contained 151 unique values where some were identical or highly similar (e.g. Fog VS Patches of Fog, T-Storm VS Thunderstorm). To reduce complexity of the model, we grouped the values and turned them into 19 values for analysis.

After making sensible judgement of the features, we relied on various statistical measures to study model adequacy. After generating a base regression model, we examined variance inflation factor (VIF) and Cook's distance to rule out multicollinearity and spot outliers.

Model building and feature selection is a reiterative process, as shown in Figure XXX (the circle) especially for a big dataset like ours. Residual analysis was run to check whether the basic assumptions of regression were met and to further identify outliers. Normality Probability Plot, Residual Against Fitted Values Plot, and PRESS Residuals came in place of assessment at this stage. Cook's distance can also be employed to identify influential data point (i.e., outliers). Outliers were then removed from the dataset. The aforementioned processes were repeated as we removed insignificant variables based on their p-value in the t-test. 95% confidence interval were set as the baseline for variable retention throughout the process.

A good regression model captures significant factors while minimizing the number of regressors involved. To come up with a good regression model, our decision-making based on various statistical measures including Coefficient of Multiple Determination (R^2), Adjusted Coefficient of Multiple Determination (Adjusted R^2), Akaike's Information Criterion (AIC), Schwarz Bayesian Criteria (represented by BIC in our R packages), PRESS Criterion and Mallow's Criterion (Cp). These measures were tracked and compared as we chose our candidate models.

Having selected a few candidate models, we entered the model validation stage. With an abundance of observations (over 130K observations in Pennsylvania State subset, which is the smallest subset out of all chosen States), we ran a 10-fold cross validation to ensure model accuracy and consistency. K-fold cross validation outputs provided a glance into root mean square error (RMSE), coefficient of determination (R^2), mean absolute error (MAE) and their respective standard deviations. The lower the RMSE, MAE and their standard deviations, the stronger the models in terms of model consistency.

In summary, the feature selection and model validation processes were highly intertwined, and we were conducting both steps back and forth before a desirable model was concluded for each selected State.

Method and Analysis

STATE: Pennsylvania (PA)

Dataset:

Dataset created for PA contains 130026 observations and 88 variables. Checked aliased coefficients for the variables and removed them. PA final dataset contained 65 variables. Further, the dataset was partitioned into test and train dataset in the ratio of 30:70 respectively.

Analysis and Insights

Initial Analysis

Duration Variable: Duration:
Upon initial exploration, data transformation appeared to be necessary to utilize the variables "Start Time" and "End Time". The output variable "Duration of Impact" (denoted as "Duration") was derived from variables "Start Time" of the accident and "End Time" of the impact in the original dataset.

$$Duration = End\ Time - Start\ Time$$

"Start Time" and "End Time" were both in character data-type in the imported dataset and it meant that we had to convert the variable into Time data-type for calculation.

Transformation of "Duration" (hour)
The boxplot of "Duration" (Figure xxx) shows that there were a lot of outliers in the data set. The values of "Duration" ranges from 0.02 hour (i.e., 1.2 min) to 25465.8 hour (i.e., 1061 days). The extremely high impact duration could be because of construction work on highways led by the accidents.

When "Duration" was taken as the target variable in the regression, the result was undesirable. Transformation was apparently the solution. Standardization, Box-cox transformation, and various others means of transformation were tested to normalize the variable. It was found that taking reciprocal exponential of the variable achieved perfect transformation. We can see the change of the shape in the boxplots. y' was adopted throughout the regression analysis.

$$y' = {}^{1}/_{e^{y}} = e^{-y}$$

Wind Condition: 151 unique weather conditions were classified and clubbed under 19 weather conditions in order to reduce the dimensionality post creation of dummy variables.

Fair	Clear	Cloudy	Partly cludy	Light Drizzle	Light Rain	Rain	Heavy Rain	Light Snow	Snow
Fair	Clear	Mostly Cloudy	Partly Cloudy	Light Drizzle	Light Rain	Rain	Heavy Rain	Light Snow	Snow
Fair / Windy		Mostly Cloudy / Windy	Partly Cloudy / Windy	Light Freezing Drizzle	Drizzle	Rain / Windy	Heavy Rain / Windy	Light Snow / Windy	Blowing Snow
		Overcast	Scattered Clouds	Light Drizzle / Windy	Light Rain / Windy	Heavy Drizzle	Heavy Rain Showers	Light Snow Showers	Snow / Windy
		Cloudy			Showers in the Vicinity	Rain Showers	Heavy Rain Shower	Light Snow with Thunder	Blowing Snow / Windy
		Cloudy / Windy			Light Rain Showers	Rain Shower		Light Snow Grains	Snow Grains
		Funnel Cloud			Light Rain Shower			Low Drifting Snow	Snow Showers
		N/A Precipitation			Light Rain Shower / Windy			Light Snow Shower	Drifting Snow
					Drizzle / Windy			Light Blowing Snow	

Heavy Snow	Fog	Smoke	Thunderstorm	Light Thunderstorm	Heavy Thunderstorm	Thunder	Wintry Mix	Others
Heavy Snow	Haze	Smoke	T-Storm	Light Thunderstorms and Rain	Heavy T-Storm	Thunder in the Vicinity	Wintry Mix	Dust
Heavy Snow / Windy	Fog	Smoke / Windy	Thunderstorm	Light Rain with Thunder	Heavy Thunderstorms and Rain	Thunder	Wintry Mix / Windy	Hail
Heavy Snow with Thunder	Patches of Fog	Heavy Smoke	Thunderstorms and Rain	Light Thunderstorms and Snow	Heavy T-Storm / Windy	Thunder / Windy	Thunder / Wintry Mix	Volcanic Ash
Heavy Blowing Snow	Shallow Fog		T-Storm / Windy	Light Thunderstorm	Heavy Thunderstorms with Small Hail		Thunder / Wintry Mix / Windy	Tornado
Thunderstorms and Snow	Light Freezing Fog		Snow and Thunder		Heavy Thunderstorms and Snow			
	Haze / Windy		Thunder and Hail					
	Drizzle and Fog		Thunder and Hail / Windy					
	Fog / Windy		Squalls / Windy					
	Partial Fog		Squalls					
	Light Haze							
	Patches of Fog / Windy							
	Light Fog							
	Partial Fog / Windy							
	Mist							

Lower frequency weather conditions were not removed as it would have taken away the unique characteristic of weather conditions pertaining to that area.

Wind Direction:

Detailed Analysis of regression model:

Regression analysis was performed as per the steps mentioned in the methodology and 13 models were shortlisted using various techniques like full regression, stepAIC (forward, backward, both), regsubset minimum BIC (forward, backward), regsubset minimum Cp (forward, backward). Further these models were compared on various statistical measures like PRESS, Adjusted R square, Residual Analysis etc.

#Final Models

Model 1: Model2Q #Model from full regression
Model 2: Model3H #Model from stepAIC forward
Model 3: Model4D #Model from stepAIC backward
Model 4: Model5F #Model from stepAIC both
Model 5: Model_minBIC_26 #Model from min BIC Forward
Model 6: Model_minBIC_27 #Model from min BIC Forward
Model 7: Model_minBIC_28 #Model from min BIC Forward
Model 8: Model_minCP_49J #Model from min CP Forward
Model 9: Model_min1BIC_32 #Model from min BIC backward
Model 10: Model_min1BIC_33 #Model from min BIC backward
Model 11: Model_min1BIC_34 #Model from min BIC backward
Model 12: Model_minCP1_43A #Model from min CP backward
Model 13: Model_minCP1_44B #Model from min CP backward

Model 1: Created by running full regression with all the 65 variables on training dataset. Removed 1 variable with high VIF and checked for cook's distance. Cook's distance for all the

observations was under the range of -1 to 1, suggesting no outliers in the dataset. 19 variables were removed one by one as the p-value was higher than alpha (at 95% level of confidence). Final, model 1 contained 39 significant variables with VIF less than 5 and observations with cook's distance in the range of -1 to 1.

```
Call:
lm(formula = y ~ x2 + x3 + x4 + x10 + x11 + x12 + x13 + x15 +
    x17 + x19 + x25 + x29 + x33 + x35 + x39 + x41 + x43 + x44 +
    x47 + x48 + x49 + x51 + x54 + x57 + x58 + x60 + x61 + x66 +
    x70 + x71 + x72 + x75 + x77 + x79 + x80 + x83 + x84 + x85 +
    x86, data = train_PA_new)
```

Model 2: Model from stepAIC forward
Model contained 38 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model3H)

Call:
lm(formula = y ~ x12 + x17 + x47 + x86 + x85 + x15 + x84 + x19 +
    x44 + x2 + x33 + x43 + x10 + x3 + x48 + x13 + x4 + x29 +
    x49 + x25 + x39 + x61 + x79 + x57 + x11 + x66 + x83 + x72 +
    x35 + x71 + x75 + x78 + x70 + x77 + x41 + x80 + x60 + x54,
    data = train_PA_new)
```

Model 3: Model from stepAIC backward
Model contained 44 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model4D)

Call:
lm(formula = y ~ x2 + x3 + x4 + x10 + x11 + x12 + x13 + x15 +
    x17 + x19 + x25 + x29 + x33 + x35 + x39 + x41 + x43 + x44 +
    x45 + x48 + x49 + x50 + x52 + x53 + x54 + x55 + x56 + x57 +
    x58 + x60 + x62 + x66 + x70 + x71 + x72 + x75 + x77 + x78 +
    x79 + x80 + x83 + x84 + x85 + x86, data = train_PA_new)
```

Model 4: Model from stepAIC both
Model contained 38 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model5F)

Call:
lm(formula = y ~ x12 + x17 + x47 + x86 + x85 + x15 + x84 + x19 +
    x44 + x2 + x33 + x43 + x10 + x3 + x48 + x13 + x4 + x29 +
    x49 + x25 + x39 + x61 + x79 + x57 + x11 + x66 + x83 + x72 +
    x35 + x71 + x75 + x78 + x70 + x77 + x41 + x80 + x60 + x54,
    data = train_PA_new)
```

Model 5: Model from forward regsubset with minimum BIC
Model contained 26 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model_minBIC_26)

Call:
lm(formula = y ~ x2 + x3 + x4 + x10 + x12 + x13 + x15 + x17 +
    x19 + x25 + x29 + x33 + x39 + x43 + x44 + x45 + x47 + x48 +
    x49 + x57 + x79 + x82 + x83 + x84 + x85 + x86, data = train_PA_new)
```

Model 6: Model from forward regsubset with minimum BIC
Model contained 27 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model_minBIC_27)

Call:
lm(formula = y ~ x2 + x3 + x4 + x10 + x12 + x13 + x15 + x17 +
    x19 + x25 + x29 + x33 + x39 + x43 + x44 + x45 + x47 + x48 +
    x49 + x57 + x72 + x79 + x82 + x83 + x84 + x85 + x86, data = train_PA_new)
```

Model 7: Model from forward regsubset with minimum BIC
Model contained 28 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model_minBIC_28)

Call:
lm(formula = y ~ x2 + x3 + x4 + x10 + x12 + x13 + x15 + x17 +
    x19 + x25 + x29 + x33 + x35 + x39 + x43 + x44 + x45 + x47 +
    x48 + x49 + x57 + x72 + x79 + x82 + x83 + x84 + x85 + x86,
    data = train_PA_new)
```

Model 8: Model from forward regsubset with minimum Cp
Model contained 39 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model_minCP_49J)

Call:
lm(formula = y ~ x2 + x3 + x4 + x10 + x11 + x12 + x13 + x15 +
    x17 + x19 + x25 + x29 + x33 + x35 + x39 + x41 + x43 + x44 +
    x45 + x47 + x48 + x49 + x50 + x52 + x54 + x55 + x57 + x58 +
    x67 + x68 + x72 + x73 + x76 + x79 + x81 + x83 + x84 + x85 +
    x86, data = train_PA_new)
```

Model 9: Model from backward regsubset with minimum BIC
Model contained 31 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model_min1BIC_32)

Call:
lm(formula = y ~ x3 + x4 + x10 + x12 + x13 + x15 + x17 + x19 +
    x25 + x29 + x33 + x39 + x43 + x44 + x45 + x48 + x49 + x50 +
    x52 + x53 + x54 + x55 + x57 + x58 + x62 + x79 + x82 + x83 +
    x84 + x85 + x86, data = train_PA_new)
```

Model 10: Model from backward regsubset with minimum BIC
Model contained 33 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model_min1BIC_33)

Call:
lm(formula = y ~ x2 + x3 + x4 + x10 + x12 + x13 + x15 + x17 +
    x19 + x25 + x29 + x33 + x39 + x43 + x44 + x45 + x48 + x49 +
    x50 + x52 + x53 + x54 + x55 + x57 + x58 + x62 + x72 + x79 +
    x82 + x83 + x84 + x85 + x86, data = train_PA_new)
```

Model 11: Model from backward regsubset with minimum BIC
Model contained 34 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model_min1BIC_34)

Call:
lm(formula = y ~ x2 + x3 + x4 + x10 + x12 + x13 + x15 + x17 +
    x19 + x25 + x29 + x33 + x35 + x39 + x43 + x44 + x45 + x48 +
    x49 + x50 + x52 + x53 + x54 + x55 + x57 + x58 + x62 + x72 +
    x79 + x82 + x83 + x84 + x85 + x86, data = train_PA_new)
```

Model 12: Model from backward regsubset with minimum BIC
Model contained 42 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model_minCP1_43A)

Call:
lm(formula = y ~ x2 + x3 + x4 + x10 + x11 + x12 + x13 + x15 +
    x17 + x19 + x25 + x29 + x33 + x35 + x39 + x41 + x43 + x44 +
    x45 + x48 + x49 + x50 + x52 + x53 + x54 + x55 + x56 + x57 +
    x58 + x60 + x62 + x67 + x68 + x72 + x73 + x76 + x79 + x82 +
    x83 + x84 + x85 + x86, data = train_PA_new)
```

Model 13: Model from backward regsubset with minimum BIC
Model contained 42 significant predictors of dependent variable 'y'. Removed the variables with high VIF (above 5). Cook's distance for all observations was in the range of -1 to 1.
Removed non-significant variables with p-value higher than alpha (at 95% level of confidence).

```
> summary(Model_minCP1_44B)

Call:
lm(formula = y ~ x2 + x3 + x4 + x10 + x11 + x12 + x13 + x15 +
    x17 + x19 + x25 + x29 + x33 + x35 + x39 + x41 + x43 + x44 +
    x45 + x48 + x49 + x50 + x52 + x53 + x54 + x55 + x56 + x57 +
    x58 + x60 + x62 + x67 + x68 + x72 + x73 + x76 + x79 + x82 +
    x83 + x84 + x85 + x86, data = train_PA_new)
```

Further statistical analysis:

1. Comparison of basic statistical parameters:
   Below table compares all the shortlisted model from the various techniques we have generated and applied regression methodology

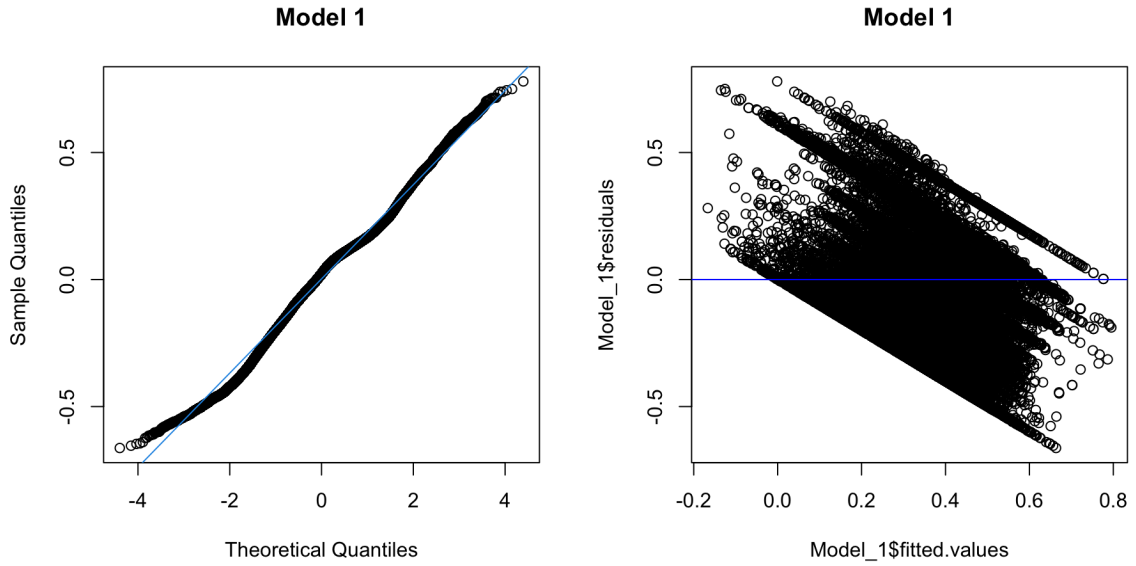
   M2 and M4 are similar (39 variable) models with lowest PRESS statistics while M9 has the lowest BIC value.
   M9 seems to be the best model as we are not losing much on Adjusted Rsqaure and prediction R square values, but it significantly reduces the number of variables from 39 to 32 out of the two models.

PA	Technique	Number of Variables	R square	Adj R Square	F-statistic	AIC	BIC	PRESS	Sum of Square Pred Error	Pred R Square
M1	Full Regression	40	25.71%	25.68%	807.5	-41455.1	-41068.9	3379.7	1438.1	25.64%
M2	Step AIC Forward	39	25.72%	25.69%	829.0	-41464.1	-41087.3	3379.4	1437.9	25.65%
M3	Step AIC Backward	45	25.74%	25.71%	716.8	-41482.6	-41049.3	3378.8	1437.8	25.66%
M4	Step AIC Both	39	25.72%	25.69%	829.0	-41464.1	-41087.3	3379.4	1437.9	25.65%
M5	Min BIC Forward-1	27	25.62%	25.60%	1205.5	-41367.2	-41103.5	3383.0	1438.5	25.57%
M6	Min BIC Forward-2	28	25.63%	25.61%	1161.4	-41377.5	-41104.4	3382.6	1438.3	25.58%
M7	Min BIC Forward-3	29	25.64%	25.62%	1120.5	-41386.9	-41104.3	3382.2	1438.4	25.59%
M8	Min Cp Forward	40	25.69%	25.65%	806.4	-41423.3	-41037.1	3380.9	1438.5	25.62%
M9	Min BIC Backward-1	33	25.64%	25.61%	980.3	-41377.0	-41056.7	3382.7	1439.1	25.58%
M10	Min BIC Backward-2	34	25.65%	25.62%	951.0	-41386.0	-41056.3	3382.3	1439.0	25.59%

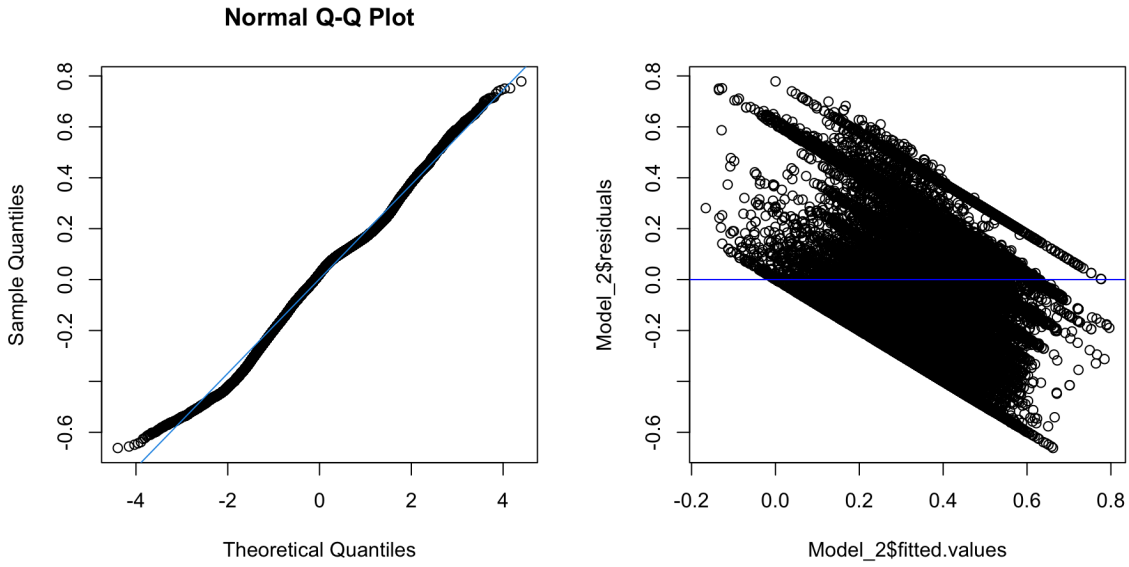
M11	**Min BIC Backward-3**	35	25.66%	25.63%	923.5	-41394.9	-41055.8	3382.0	1439.0	25.59%
M12	**Min Cp Backward-1**	43	25.70%	25.67%	749.3	-41433.8	-41019.4	3380.6	1438.3	25.62%
M13	**Min Cp Backward-2**	43	25.70%	25.67%	749.3	-41433.8	-41019.4	3380.6	1438.3	25.62%

1. Residual Plots for all models

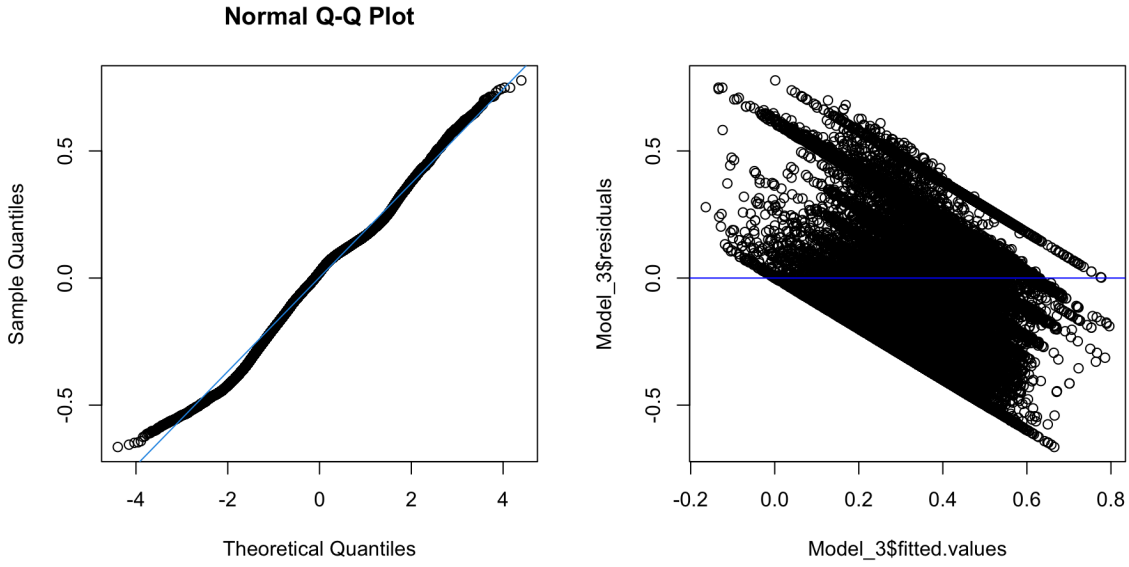
   Model 1:



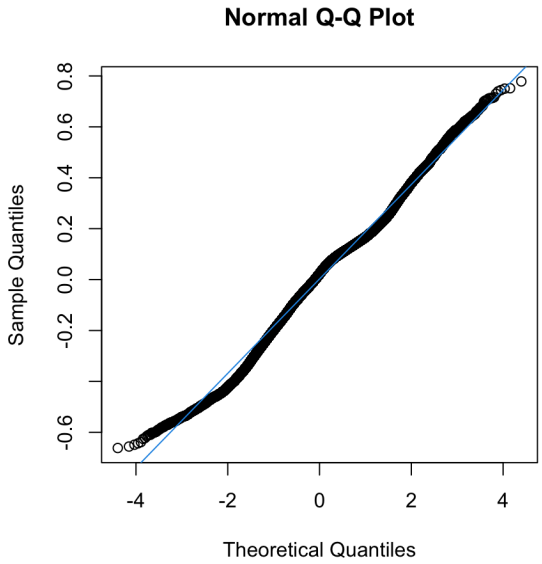
   Model 2:

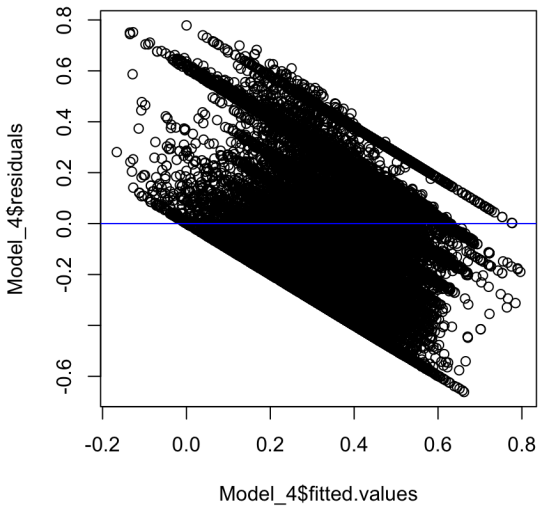


Model 3:

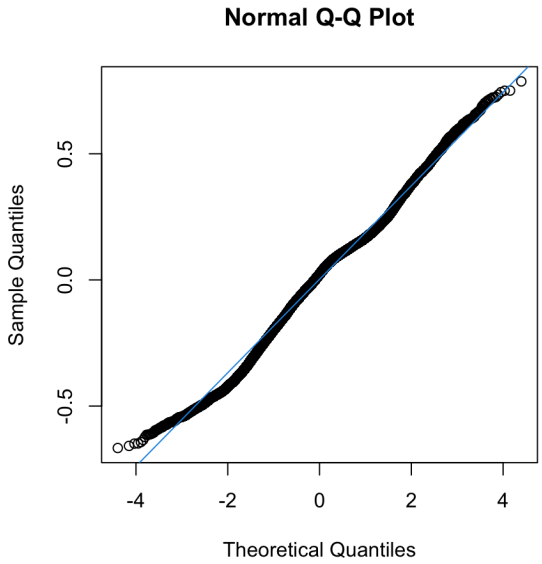


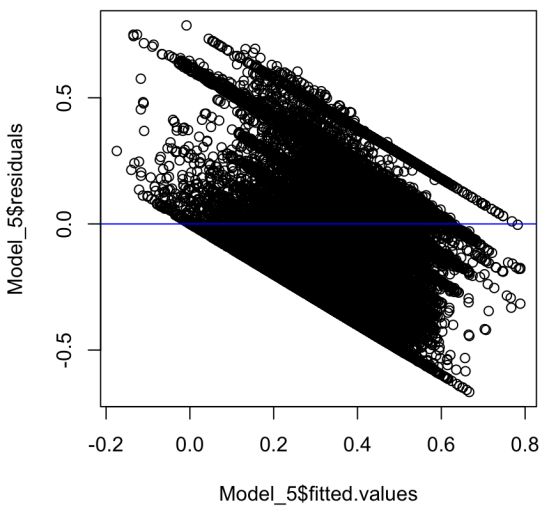
Model 4:



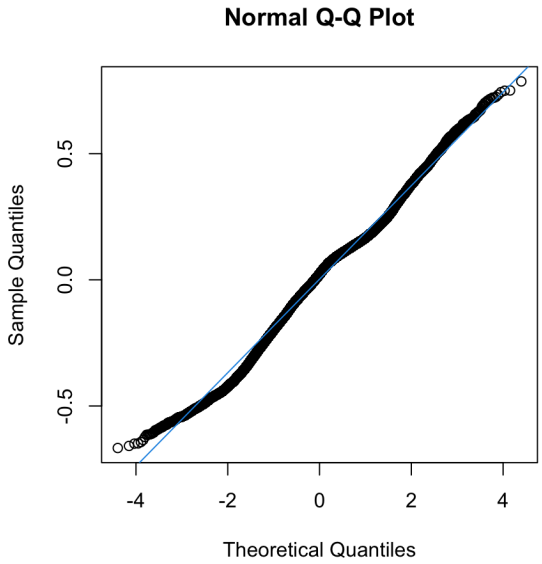


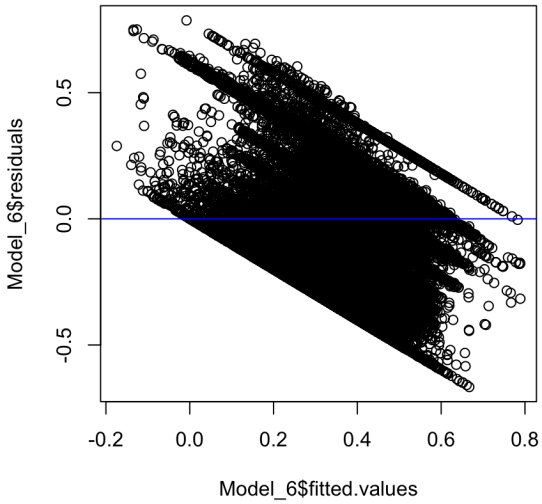
Model 5:



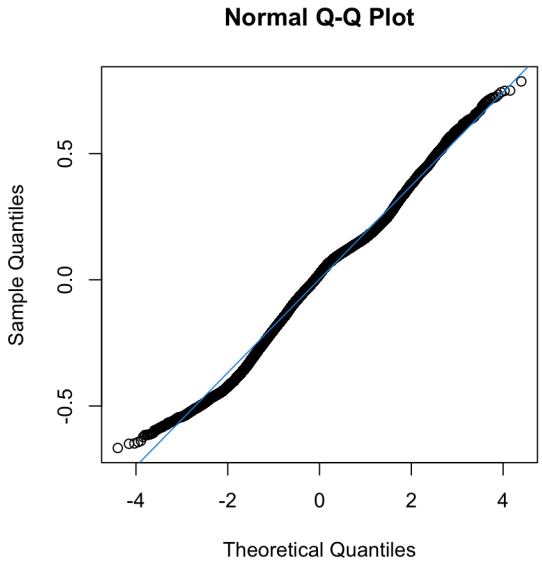


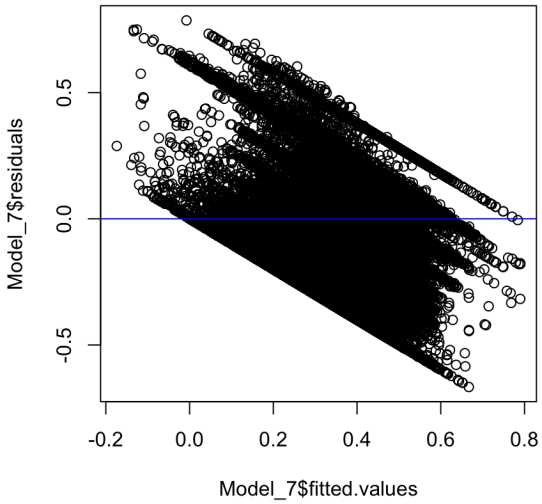
Model 6:



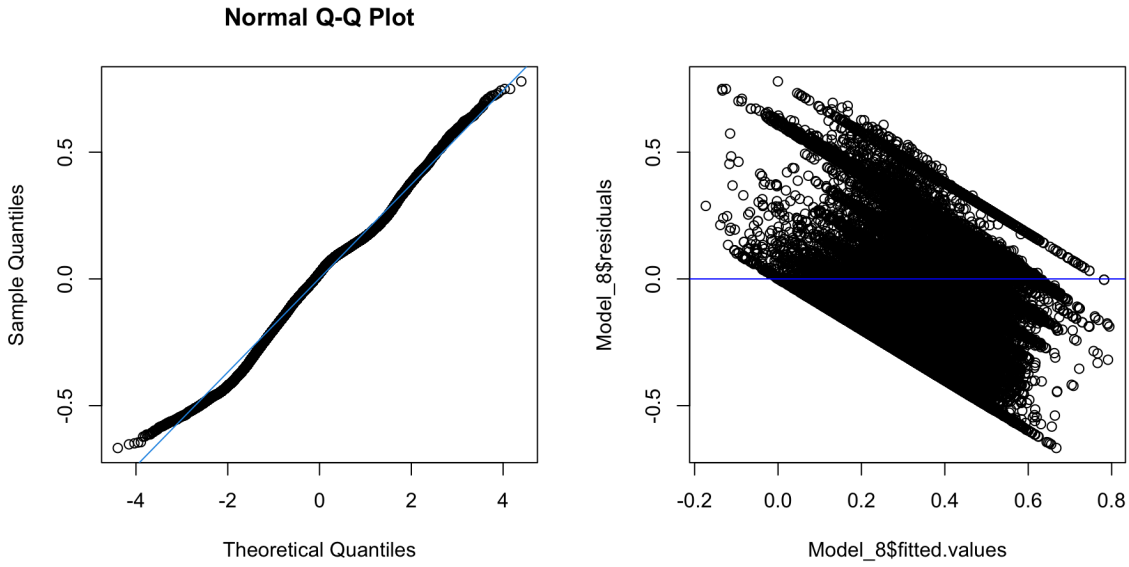


Model 7:

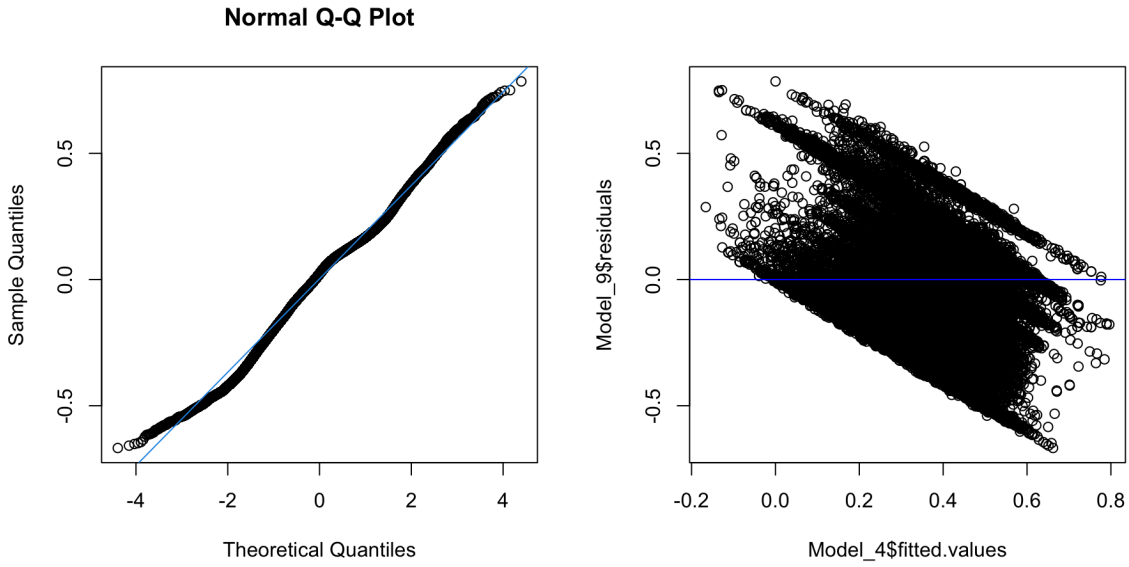




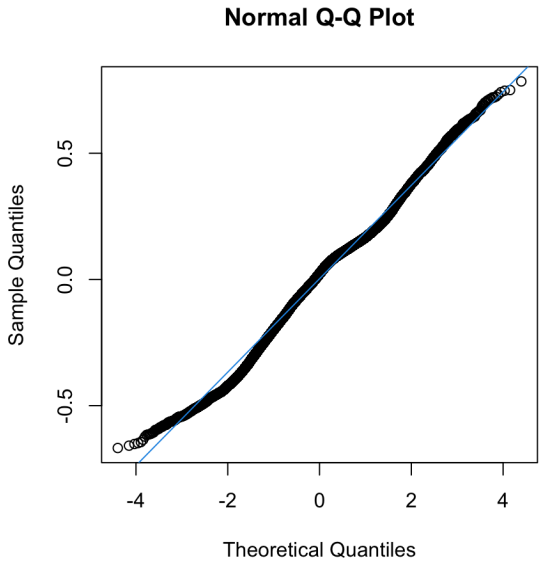
Model 8:

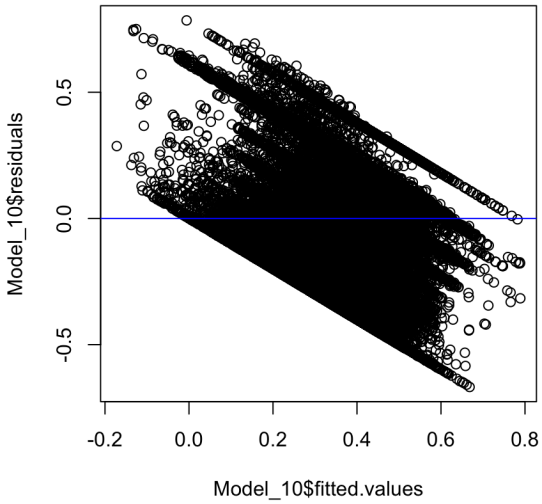


Model 9:

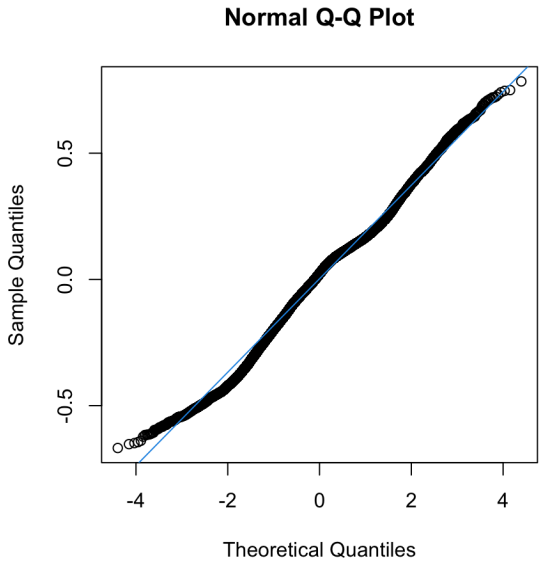


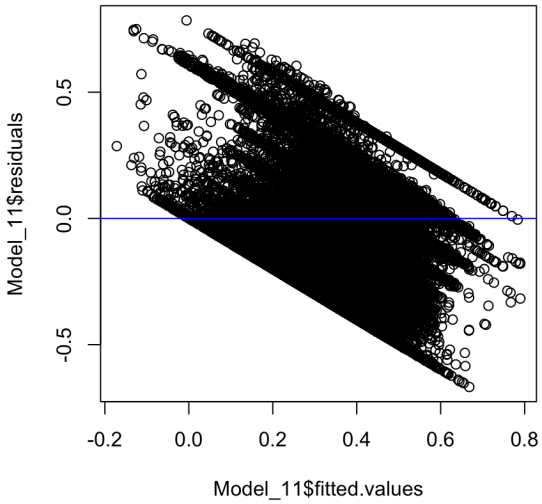
Model 10:



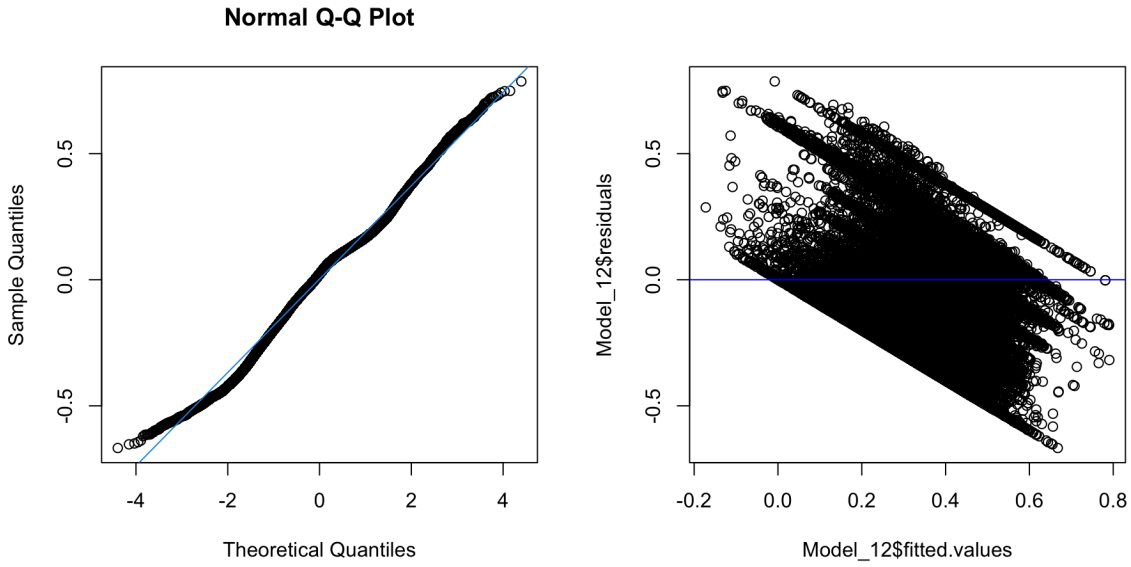


Model 11:

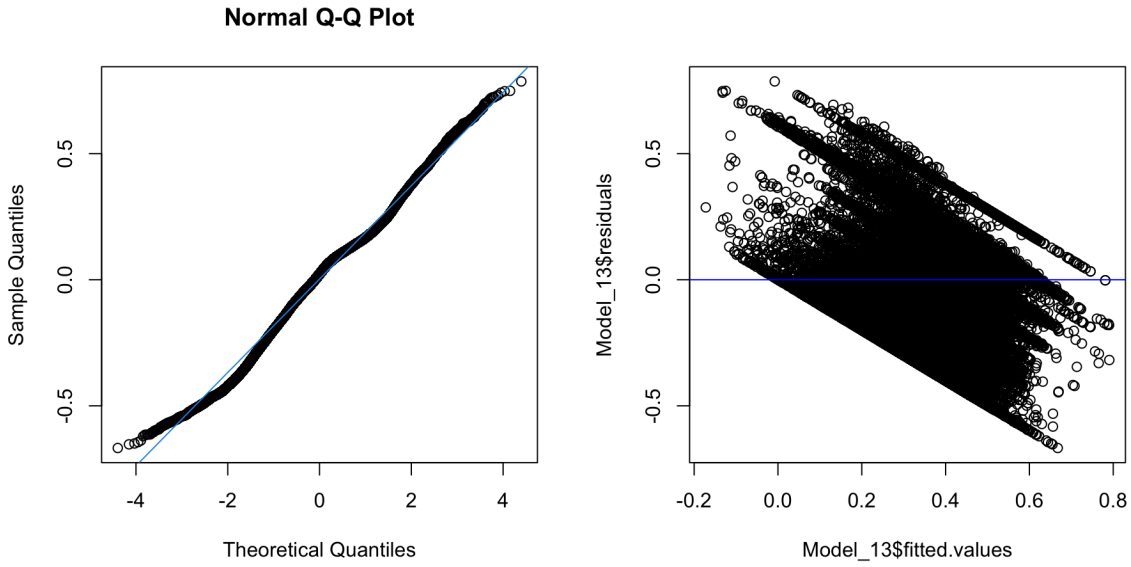




Model 12:



Model 13:



2. K-Fold Validation Table showing all models

PA	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
M1	0.19244	25.74%	0.1530	0.000868	0.00781	0.00073
M2	0.19246	25.73%	0.1531	0.001267	0.01017	0.00093

M3	0.19245	25.74%	0.1531	0.000830	0.00469	0.00056
M4	0.19246	25.73%	0.1531	0.001273	0.00959	0.00083
M5	0.19255	25.66%	0.1532	0.001171	0.00719	0.00102
M6	0.19254	25.67%	0.1532	0.001008	0.00607	0.00070
M7	0.19253	25.68%	0.1532	0.001313	0.00864	0.00090
M8	0.19251	25.69%	0.1531	0.000874	0.00555	0.00074
M9	0.19251	25.69%	0.1531	0.000874	0.00555	0.00074
M10	0.19251	25.69%	0.1531	0.000874	0.00555	0.00074
M11	0.19251	25.69%	0.1531	0.000874	0.00555	0.00074
M12	0.19251	25.69%	0.1531	0.000874	0.00555	0.00074
M13	0.19251	25.69%	0.1531	0.000874	0.00555	0.00074

Finalized Model:
Model_9/M9 has proven to be the best model to predict the duration of the accident for Pennsylvania. 32 variables.

Regression Equation:
Exponential of Duration is primarily dependent on 32 variables with an intercept of -0.69 and duration (exponential) of accident has a maximum negative impact due to a bump on the road.

```
Call:
lm(formula = y ~ x2 + x3 + x4 + x10 + x12 + x13 + x15 + x17 +
    x19 + x25 + x29 + x33 + x39 + x43 + x44 + x45 + x48 + x49 +
    x50 + x52 + x53 + x54 + x55 + x57 + x58 + x62 + x79 + x82 +
    x83 + x84 + x85 + x86, data = train_PA_new)
```

Exponential Duration, y = β_0 +	β_1Start Lat +	β_2 Start Long +	β_3 Distance +	β_4 Temperature +
	β_5 Pressure +	β_6 Visibility +	β_7 Highway +	β_8 Night +
	β_9 Traffic Signal +	β_{10} Station +	β_{11} Railway +	β_{12} Junction +
	β_{13} Bump +	β_{14} Right Side +	β_{15} WC Clear +	β_{16} WC Cloudy +
	β_{17} WC Fog +	β_{18} WC Hail +	β_{19} WC Heavy Rain +	β_{20} WC Heavy Thunderstorm +
	β_{21} WC Light Drizzle +	β_{22} WC Light Rain +	β_{23} WC Light Snow +	β_{24} WC Partly cloudy +
	β_{25} WC Rain +	β_{26} WC Thunderstorm +	β_{27} WD SW +	β_{28} WD WNW +
	β_{29} WD WSW +	β_{30} Severity1 +	β_{31} Severity2 +	β_{32} Severity3

Exponential Duration, y =	-0.69	-0.04	Start Lat	+	-0.005	StartLong	+	0.00	Distance	+	0.0004	Temperature	+
		0.07	Pressure	+	-0.004	Visibility	+	-0.06	Highway	+	-0.08	Night	+
		0.04	TrafficSignal	+	0.02	Station	+	-0.04	Railway	+	-0.05	Junction	+
		-0.22	Bump	+	-0.02	RightSide	+	0.13	WCClear	+	0.08	WCCloudy	+
		0.02	WCFog	+	0.14	WCHail	+	0.07	WCHeavyRain	+	0.10	WCHeavyThunderstorm	+
		0.06	WCLightDrizzle	+	0.07	WCLightRain	+	0.07	WCLightSnow	+	0.09	WCPartlycloudy	+
		0.06	WCRain	+	0.07	WCThunderstorm	+	-0.01	WDSW	+	0.01	WDWNW	+
		-0.01	WDWSW	+	0.44	Severity1	+	0.19	Severity2	+	0.30	Severity3	+

Regression statistics:
Overall model is significant as the p-value for F-Stat is less than alpha (at 95% level of confidence). The above model explained 25.6% of variation in the data, suggesting it as decent model.

```
Residual standard error: 0.1927 on 90987 degrees of freedom
Multiple R-squared:  0.2564,    Adjusted R-squared:  0.2561
F-statistic: 980.3 on 32 and 90987 DF,  p-value: < 2.2e-16
```

Conclusion:

Although duration of accidents varies significantly, and it is possible to establish important determinants of duration of accidents. The recommended regression model conclude that time required to clear the impacted area due to accidents can be mostly explained in terms of the bump on the road, starting latitude, right side of the road, night, junction, presence of railway in vicinity.