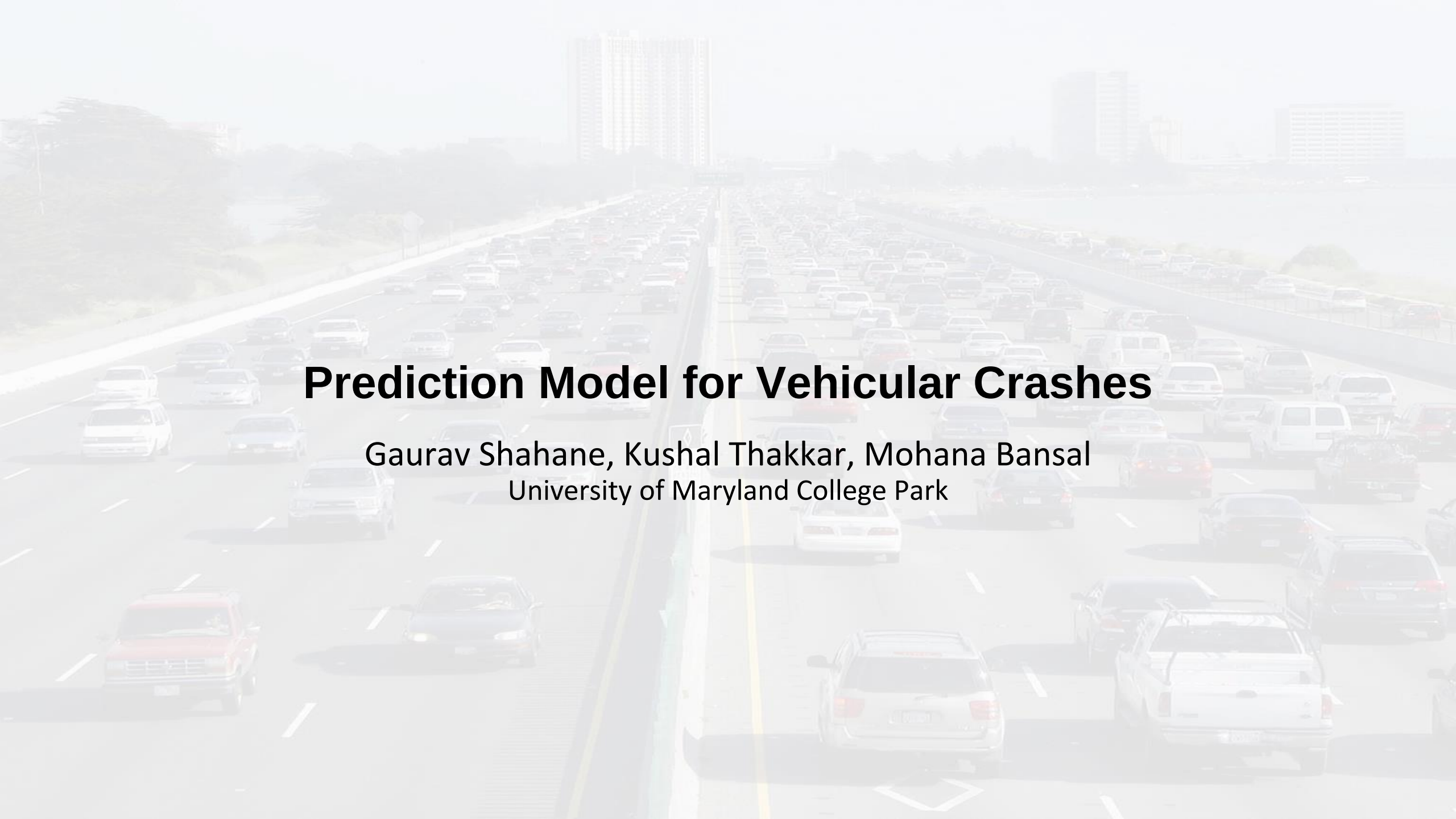


Prediction Model for Vehicular Crashes

Gaurav Shahane, Kushal Thakkar, Mohana Bansal
University of Maryland College Park

Research Question

Predicting the occurrences of vehicular crashes on roadways of the State of Iowa based on spatial and temporal features of weather and traffic data.

Data: Road Weather data, Traffic data and Crash data

Literature Review

Abdel-Aty, M. A., Pemmanaboina, R. (2006). Calibrating a Real- Time Traffic Crash-Prediction Model Using Archived Weather and ITS Traffic Data, vol. 7(2)

Chen, F., Chen, S., Ma, X. (2016). Crash Frequency Modeling Using Real Time Environmental and Traffic Data and Unbalanced Panel Data Models, International Journal of Environmental Research and Public Health

Goodwin, L. C. (2002). Analysis of Weather Related Crashes on U.S. Highways Lee, C., Hellinga, Bruce., Saccomanno, F. (2003). Real-Time Crash Prediction Model for the Application to Crash prevention in Freeway Traffic, pp. 03-2749

Yu, R., Abdel-Aty, M. A., Ahmed, M. M., Wang, X. (2014). Utilizing Microscopic Traffic and Weather Data to Analyze Real-Time Crash Patterns in the Context of Active Traffic Management, vol. 15(1)

Literature Review Contd…

Incorporate traffic safety improvement and suggest active traffic management (ATM)strategies by identifying real time crash patterns on I-70 freeway

Binary logistic regression models compared single with multi vehicle crashes and side-swipe with rear end crashes

Hierarchical logistic regression model to accommodate the above two regression models

Different ATM strategies should be designed for two lane and three lane roadways considering seasonal effects

Reference: Yu, R., Abdel-Aty, M. A., Ahmed, M. M., Wang, X. (2014). Utilizing Microscopic Traffic and Weather Data to Analyze Real-Time Crash Patterns in the Context of Active Traffic Management, vol. 15(1)

Data Set:

State of Iowa - Road Weather Information System

station	obtime	Air Temperature	Dew Point Temperature	Wind Speed	Wind Direction	Wind Gust	Pavement Sensor: Temperature	Road Condition
RSNI4	1/1/2016 0:01	25.1	17.4	6.79803	259	9.16256	30.5	
RPFI4	1/1/2016 0:01	13.82	10.22	5.93953	260	9.71923	20.3	Dry
RJFI4	1/1/2016 0:01	17.6	12.2	12.959	240	15.1188	21.92	Dry
RPLI4	1/1/2016 0:01	21.92	12.92	5.93953	235	7.5594	24.8	Dry
RFDI4	1/1/2016 0:01	16.52	10.58	9.17927	260	12.419	16.88	Dry

- Data from 71 Road Weather Information Stations across the State of Iowa
- Time interval of 10 mins between successive observations
- **Main Features:** Temperature, Wind Speed, Road Condition

Data Cleaning Continues..

Weather Data

- Aggregated **1.7 Million Records**
- Aggregated the data based on **Per Station Per Date Per Hour** Records
- Merged the data by taking **Median** across sensors for the same station
- Used Merge, and Aggregate Functions in R

Data Set:

State of Iowa - Historic Traffic Data

	station	obtime	lane_id	avg_speed	avg_headway	normal_vol	long_vol	occupancy
0	RDVI4	1/1/2016 0:01	0	69.5934	65535	26	23	3
1	RDVI4	1/1/2016 0:01	1	76.4285	65535	21	18	1
2	RPLI4	1/1/2016 0:01	0	65.8652	65535	8	2	2

- Data from 71 Road Weather Information Stations across the State of Iowa
- Time interval of 10 mins between successive observations
- **Main Features:** Average Speed, Avg headway, Volume, Occupancy

Data Cleaning Continues..

Traffic Data

- Aggregated **1.2 Million Records**
- Aggregated the data based on **Per Station Per Date Per Hour** Records
- Merged the data by taking **Median** across sensors for the same station
- Used Merge, and Aggregate Functions in R

Merging Road Pavement Conditions..

- Challenge was to aggregate Text Records such as:
  - Wet, Dry, Snow Watch, Ice/Snow Watch, Error, No Reports, Other

- Methodology used:
  - Coded Text Labels as numbers, such as: Dry: 1, Wet: 2, and Etc.
  - Replaced Errors with Blanks
  - Calculated the highest frequent condition for a given hour
  - Assigned this Pavement condition for the Station for that hour

Merging Crash Numbers, Stations, and Distance

- Challenge was to match Crash site with one of the Weather Stations
  - Latitude-Longitudes were available for the both
  - Station on the same Road type was required, to maintain correct traffic data

- Methodology used:
  - Calculated the distance between Crash site and each Weather Station.
  - Used Haversine Formula for the calculation, Implemented the same in R
  - Identified the Road Type at the Crash Site using OpenStreetMap API
  - Matched Crash record with the Nearest Station on the same Road Type
  - Recorded Crash Numbers Per Hour, Station, and Distance from the Crash Site

Distance Between two Geo Locations

```
dlon = lon2 - lon1
dlat = lat2 - lat1
a = (sin(dlat/2))^2 + cos(lat1) * cos(lat2) * (sin(dlon/2))^2
c = 2 * atan2( sqrt(a), sqrt(1-a) )
d = R * c (where R is the radius of the Earth)
```

Source: http://andrew.hedges.name/experiments/haversine/

Data Merging

Merging Weather, Traffic and Crash Data

- Merged Weather and Traffic data based on **Per Station Per Date Per Hour**
- Added Road Pavement Conditions
- Added Number of Crashes, Closest Weather Station on same Road Type, and Distance
- Cleaning, Aggregating, and Merging data from multiple sources reduced the size of the Final Dataset to 1/3rd of the original records.

Exploratory Data Analysis: Records from 2016

Dimensions: 125330 Records * 19 Variables
Data Types: Characters (Station), Numbers (Avg Speed, Wind Gust, Occupancy)

The values of the five numbers for the final dataset

1. Min
2. 1st Quantile
3. Median
4. 2nd Quantile
5. Max

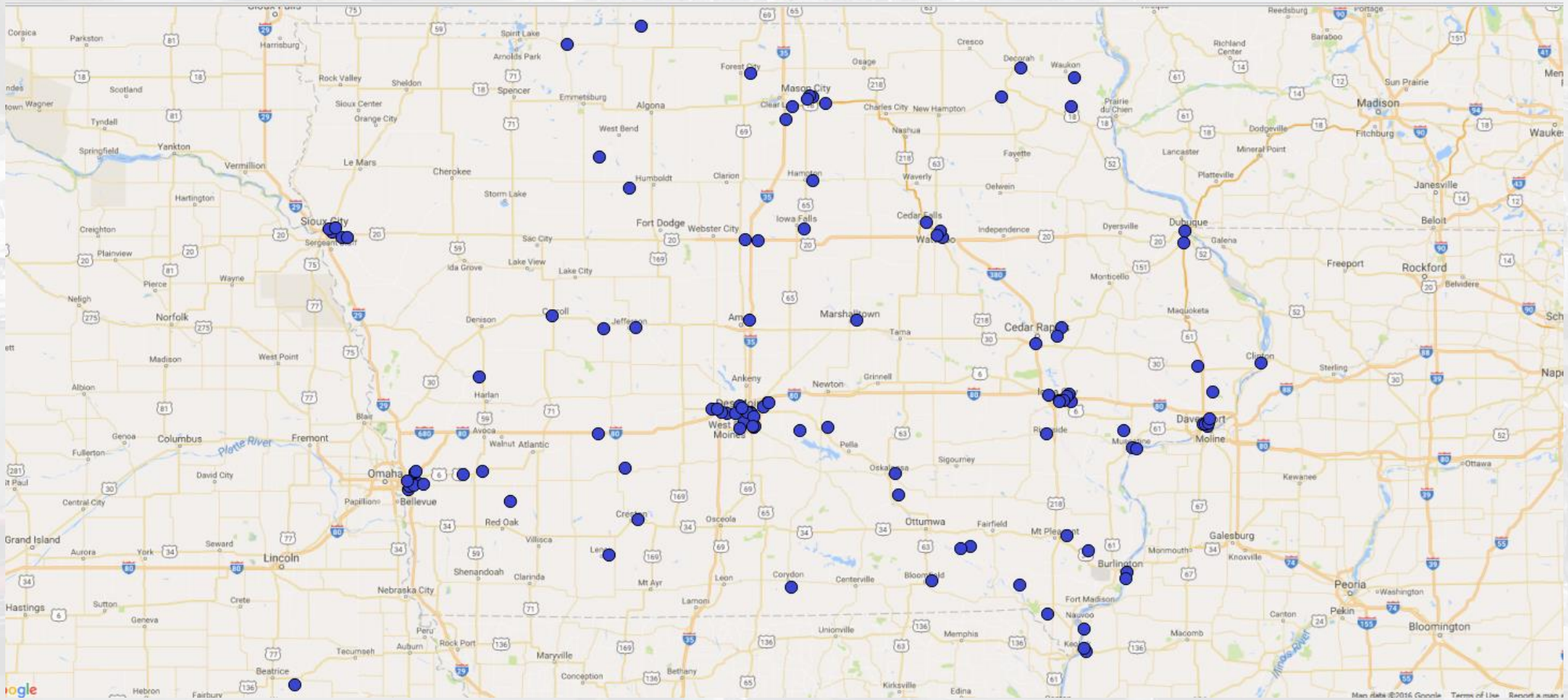
```
> fivenum(data$tmpf)
[1] -19.75  31.37  45.86  62.24  99.05
> fivenum(data$dwpf)
[1] -24.70  25.43  36.32  51.08  88.25
> fivenum(data$tmpf)
[1] -19.75  31.37  45.86  62.24  99.05
> fivenum(data$dwpf)
[1] -24.70  25.43  36.32  51.08  88.25
> fivenum(data$sknt)
[1]  0.000000  3.509720  6.209510  9.719235 36.987100
> fivenum(data$drct)
[1]   0.0 107.5 180.0 277.5 355.0
> fivenum(data$gust)
[1]  0.000000  5.939530  9.989215 15.118800 92.062850
> fivenum(data$tfs)
[1] -16.15  33.71  51.17  70.97 136.76
> fivenum(data$avg_speed)
[1]  0.00000 55.30195 61.51560 69.28275 83.88490
> fivenum(data$avg_headway)
[1] 65535 65535 65535 65535 65535
> fivenum(data$normal_vol)
[1]  0.0  1.0  3.0  6.0 73.5
> fivenum(data$long_vol)
[1]  0.0  0.0  0.5  2.0 63.0
> fivenum(data$occupancy)
[1]  0  0  0  1 99
```

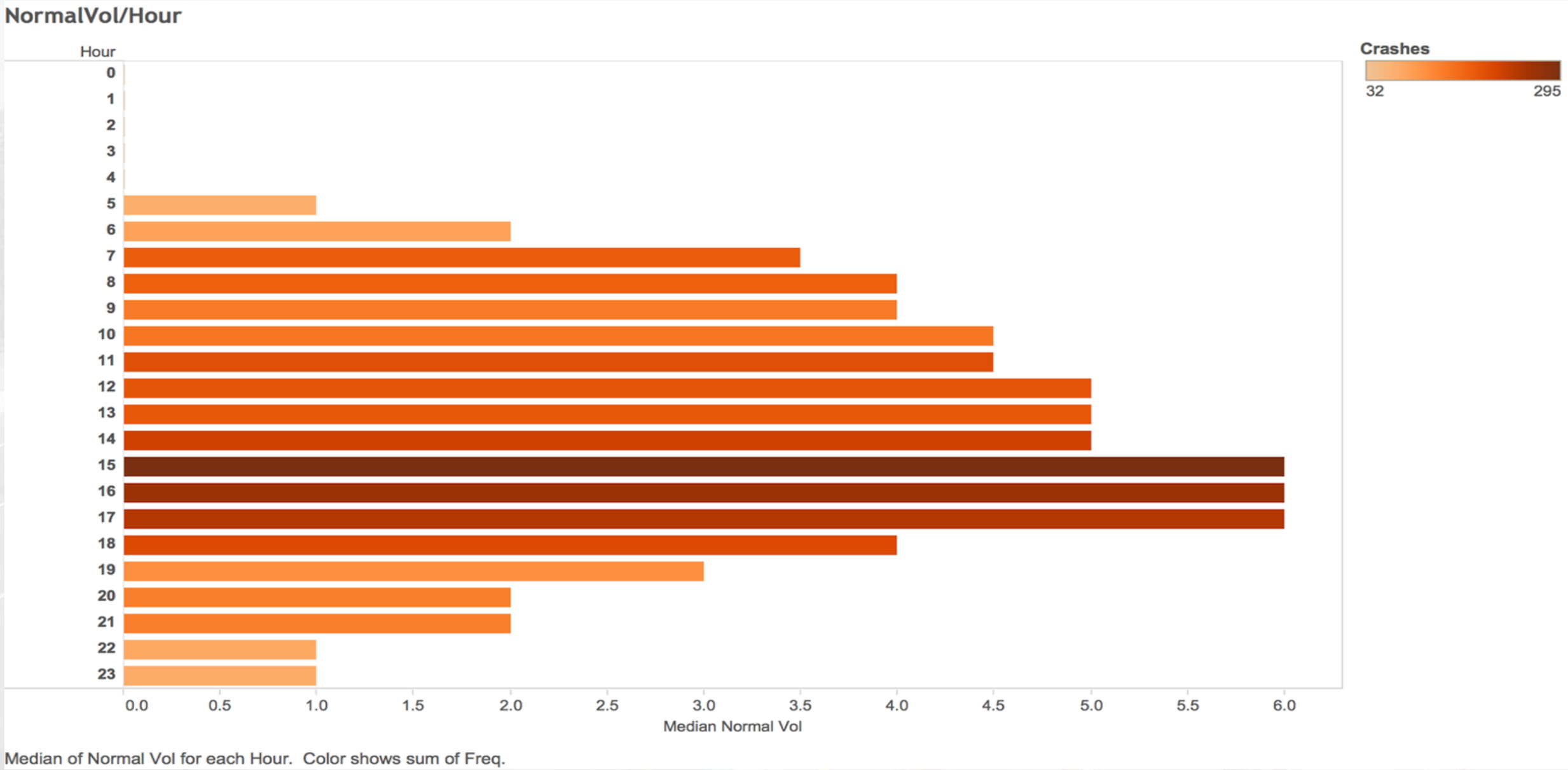
The values Standard Deviation (SD) for the final dataset

- Realized the Avg_headway is a column with constant values
- Need to eliminate the same from further analysis

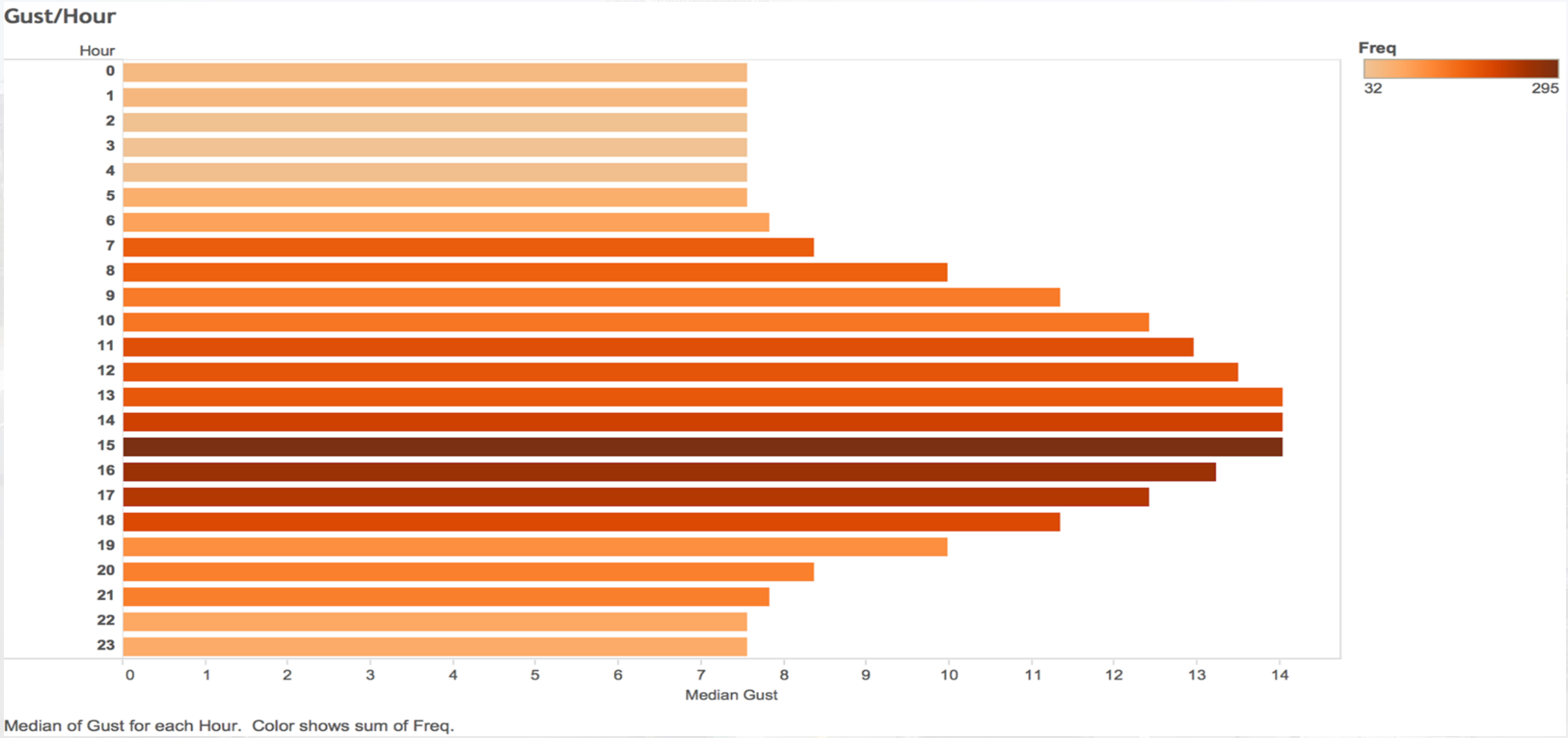
```
> #Standard Deviation
> sd(data$tmpf)
[1] 20.8961
> sd(data$dwpf)
[1] 18.44673
> sd(data$sknt)
[1] 4.817836
> sd(data$drct)
[1] 101.8534
> sd(data$gust)
[1] 6.696386
> sd(data$tfs)
[1] 26.21229
> sd(data$avg_speed)
[1] 10.1849
> sd(data$avg_headway)
[1] 0
> sd(data$normal_vol)
[1] 7.431384
> sd(data$long_vol)
[1] 4.745973
> sd(data$occupancy)
[1] 3.864226
```

Accident Crash Screenshot 01/01/2016





Traffic Volume by Hour. #Crashes indicated with color



Median Gust by Hour. #Crashes indicated with color

1.Multivariate Linear Regression

- Considered all variables for modelling
- No variables were significant enough

Correlation between Predictions & Crashes without Regularization:
0.2128064

```
Call:
lm(formula = Freq ~ ., data = dataLm)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3440 -0.1486 -0.0814 -0.0242  4.7486

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.2700558  0.0620200  20.478  < 2e-16 ***
hour         0.0016878  0.0012672   1.332 0.182988
tmpf        -0.0038650  0.0018863  -2.049 0.040550 *
dwpf         0.0027470  0.0010296   2.668 0.007670 **
sknt        -0.0174114  0.0050906  -3.420 0.000634 ***
drct        -0.0000450  0.0000690  -0.652 0.514329
gust         0.0139828  0.0035320   3.959 7.70e-05 ***
tfs          0.0003880  0.0010928   0.355 0.722604
lane_id      0.0076856  0.0111770   0.688 0.491741
avg_speed   -0.0040257  0.0007474  -5.386 7.74e-08 ***
avg_headway         NA         NA      NA       NA
normal_vol   0.0058192  0.0014985   3.883 0.000105 ***
long_vol    -0.0018703  0.0020873  -0.896 0.370305
occupancy   -0.0029733  0.0012729  -2.336 0.019562 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3656 on 3019 degrees of freedom
Multiple R-squared:  0.0417,    Adjusted R-squared:  0.03789
F-statistic: 10.95 on 12 and 3019 DF,  p-value: < 2.2e-16
```

2. Multivariate Linear Regression with Regularization

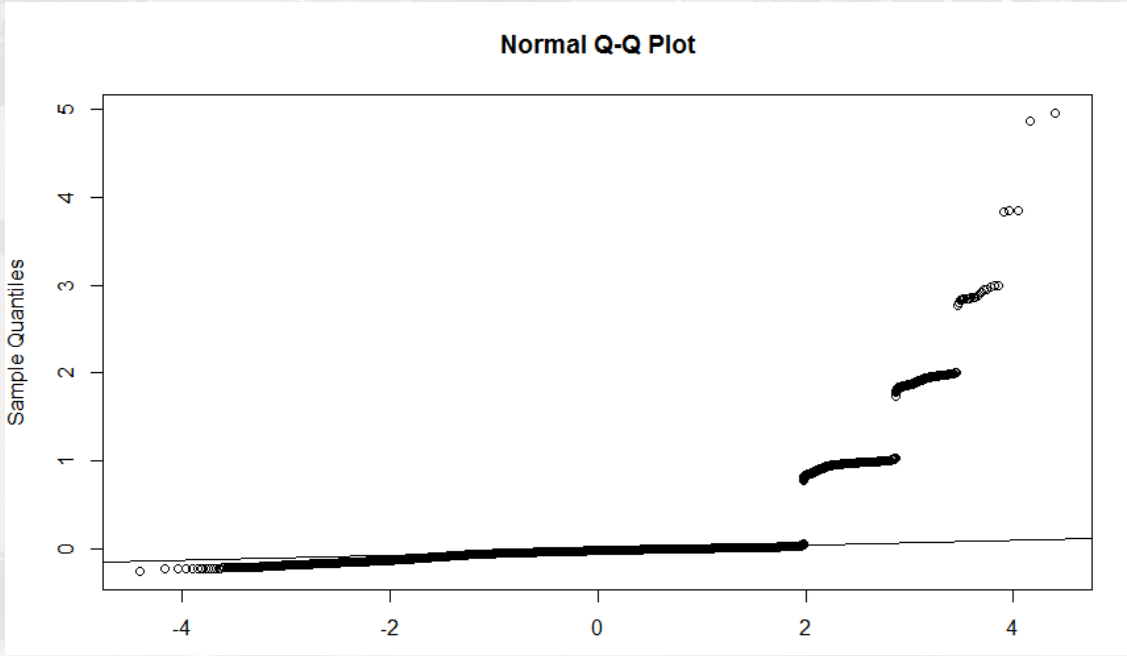
```
Call:
lm(formula = Freq ~ ., data = crash_train)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2483 -0.0385 -0.0175 -0.0004  4.9524

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.075e-01  4.364e-03  24.637  < 2e-16 ***
hour         4.654e-04  8.652e-05   5.379 7.52e-08 ***
tmpf         1.900e-04  1.505e-04   1.263 0.206719
dwpf        -2.394e-04  8.954e-05  -2.674 0.007503 **
sknt        -6.358e-03  4.453e-04 -14.278  < 2e-16 ***
drct        -3.753e-05  5.824e-06  -6.444 1.17e-10 ***
gust         4.423e-03  3.249e-04  13.614  < 2e-16 ***
tfs         -1.810e-04  8.736e-05  -2.071 0.038334 *
lane_id      3.146e-02  9.411e-04  33.432  < 2e-16 ***
avg_speed   -2.113e-03  6.128e-05 -34.487  < 2e-16 ***
avg_headway          NA         NA      NA       NA
normal_vol   3.031e-03  1.634e-04  18.547  < 2e-16 ***
long_vol    -8.646e-04  2.410e-04  -3.587 0.000334 ***
occupancy   -2.683e-04  1.601e-04  -1.676 0.093741 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1738 on 93985 degrees of freedom
Multiple R-squared:  0.04281,   Adjusted R-squared:  0.04269
F-statistic: 350.3 on 12 and 93985 DF,  p-value: < 2.2e-16
```

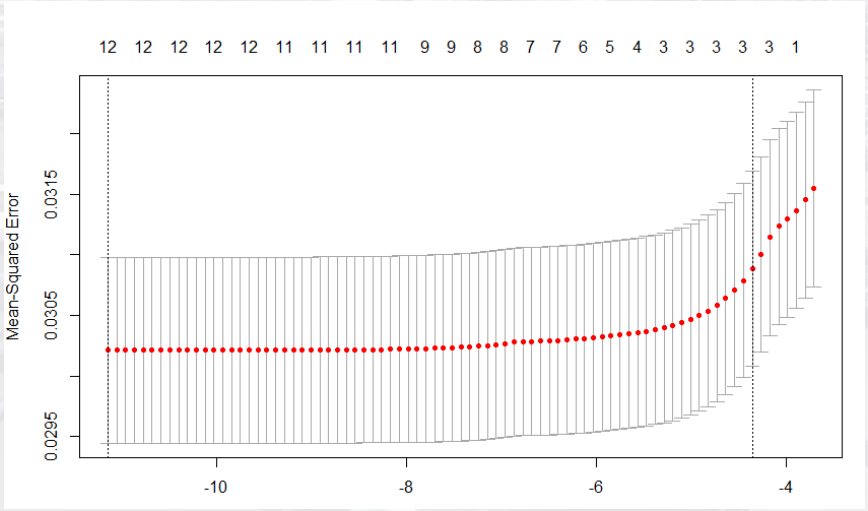
Residuals were not normal.



Cross Validation Results:

Results After Cross Validation: lambda value: 1.44359e-05

	1
(Intercept)	0.0297120845
hour	.
tmpf	.
dwpf	.
sknt	.
drct	.
gust	.
tfs	.
lane_id	0.0174244313
avg_speed	-0.0004638787
avg_headway	.
normal_vol	0.0004661011
long_vol	.
occupancy	.



Correlation after Regularization: 0.1924064

3. Zero Inflated Negative Binomial Model

- Result based on Occurrence of the crashes
- Very low accuracy as only 3 crashes were correctly predicted.

Actual	Predicted 0	1	Row Total
0	30599 0.977	3 0.000	30602
1	730 0.023	1 0.000	731
Column Total	31329	4	31333

- Result based on Number of crashes occurred
- Very low accuracy as only 4 crashes were correctly predicted

Actual	Predicted 0	1	12	746	Row Total
0	30586 0.976	14 0.000	1 0.000	1 0.000	30602
1	661 0.021	7 0.000	0 0.000	0 0.000	668
2	51 0.002	1 0.000	0 0.000	0 0.000	52
3	8 0.000	1 0.000	0 0.000	0 0.000	9
4	2 0.000	0 0.000	0 0.000	0 0.000	2
Column Total	31308	23	1	1	31333

4. Random Forest

- Converted Number of Crashes as a Binary Dependent Variable with 0: No Crashes, 1: Crashes
- 97.5% Records with No Crashes, 2.5% Records with Crashes, Hence it was an unbalanced Dataset.
- Training and Testing Dataset created with equal proportion of Crash and Non-Crash Records
- Variables used: "hour", "tmpf", "dwpf", "sknt", "drct", "gust", "tfs", "avg_speed", "avg_headway", "normal_vol", "long_vol", "occupancy"

Result

- Accuracy of the model is 97.45%
- However the ROC curve for the model is poor and the accurate prediction are likely by chance.
- We are working on other models and techniques to address the issue.

Total Observations in Table: 31333

	Predicted		
Actual	0	1	Row Total
0	30504	96	30600
	0.974	0.003	
1	700	33	733
	0.022	0.001	
Column Total	31204	129	31333

Random Forest after Down Sampling

Results:

- Accuracy for Class 0: 77.52%
- Accuracy for Class 1: 66.58%
- Overall Accuracy: 77.25%

Model has a higher false positive, which means the model is stringent.

	Predicted		
Actual	0	1	Row Total
0	23692 0.756	6869 0.219	30561
1	258 0.008	514 0.016	772
Column Total	23950	7383	31333

Random Forest with 3 Classes

3 Classes were generated based on the number of crashes in an hour

- No Crash, 1 Crash, 1+ Crashes

Accuracy for different classes:

- **No Crash: 73.3%**
- **1 Crash: 68.26%**
- 1+ Crashes: 4.7%

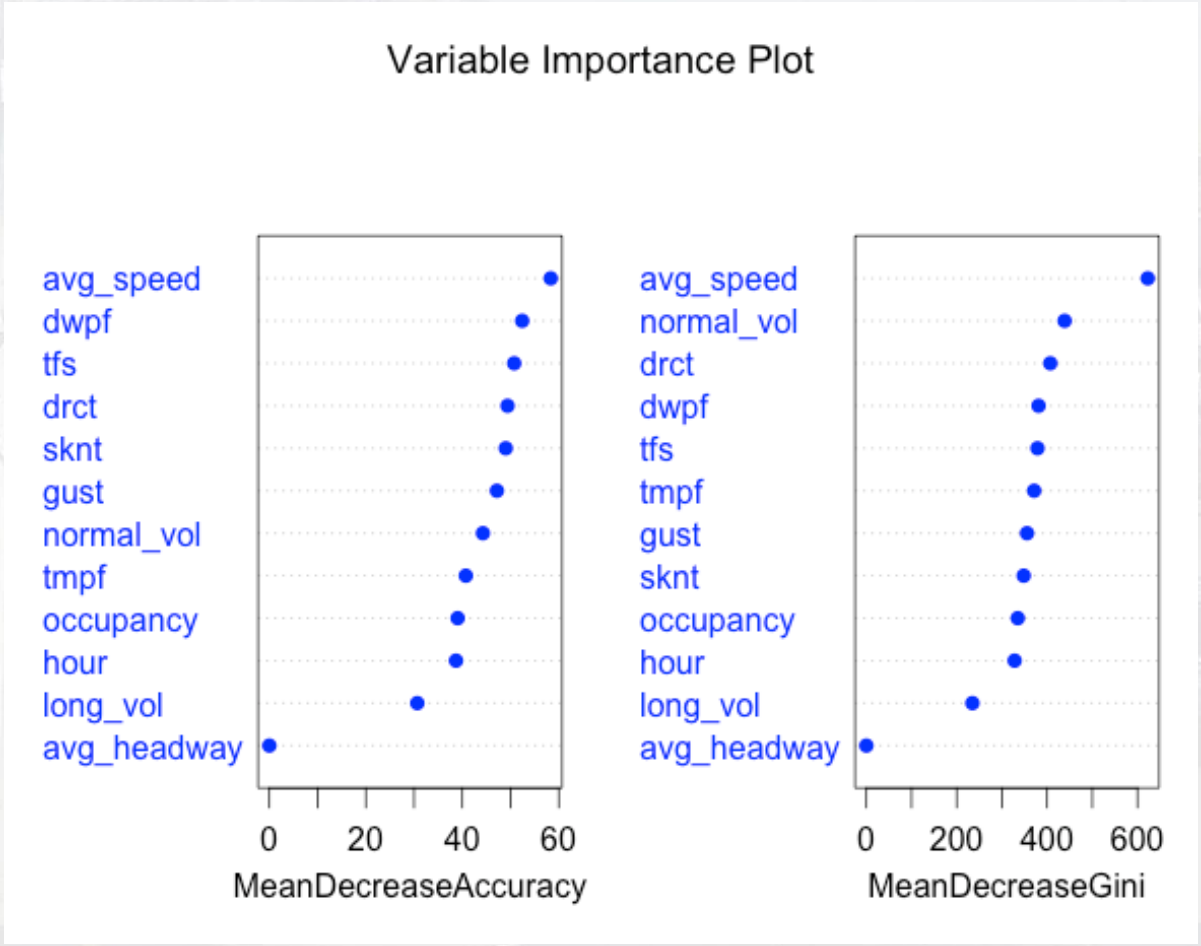
Total Observations in Table: 31333

	Predicted			
Actual	0	1	2	Row Total
0	23048	7505	49	30602
	0.736	0.240	0.002	
1	200	456	12	668
	0.006	0.015	0.000	
2	6	54	3	63
	0.000	0.002	0.000	
Column Total	23254	8015	64	31333

Random Forest with 3 Classes

Important Variables:

- Average Speed
- Dew Point Temperature
- Pavement Temperature
- Direction of wind
- Wind Speed in knots
- Wind gust in knots
- Traffic volume



Conclusion

Best Predictive model: Random Forest with Accuracy:

No Crash: **73.3%**
1 Crash: **68.26%**
1+ Crashes: 4.7%

Most Predictive Features for Vehicular crash prediction in the order of importance:

- Average Speed
- Dew Point Temperature
- Pavement Temperature
- Direction of wind
- Wind Speed in knots
- Wind gust in knots
- Traffic volume