Crash Severity Prediction Study

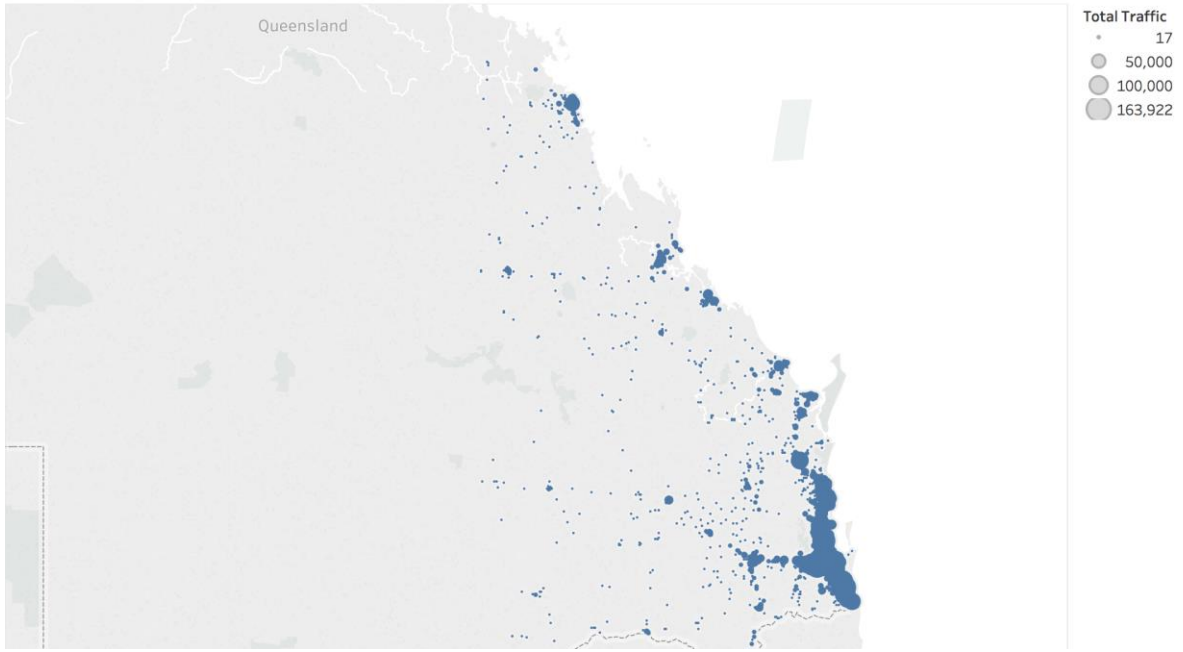
PRESENTED BY GROUP #4

- BADR AL-MAHRUQI (44831132)
- SANSKRUTI SHINDE (45238989)
- SHREY SHAH (45207417)
- VINCENT ABBOSH (45019218)

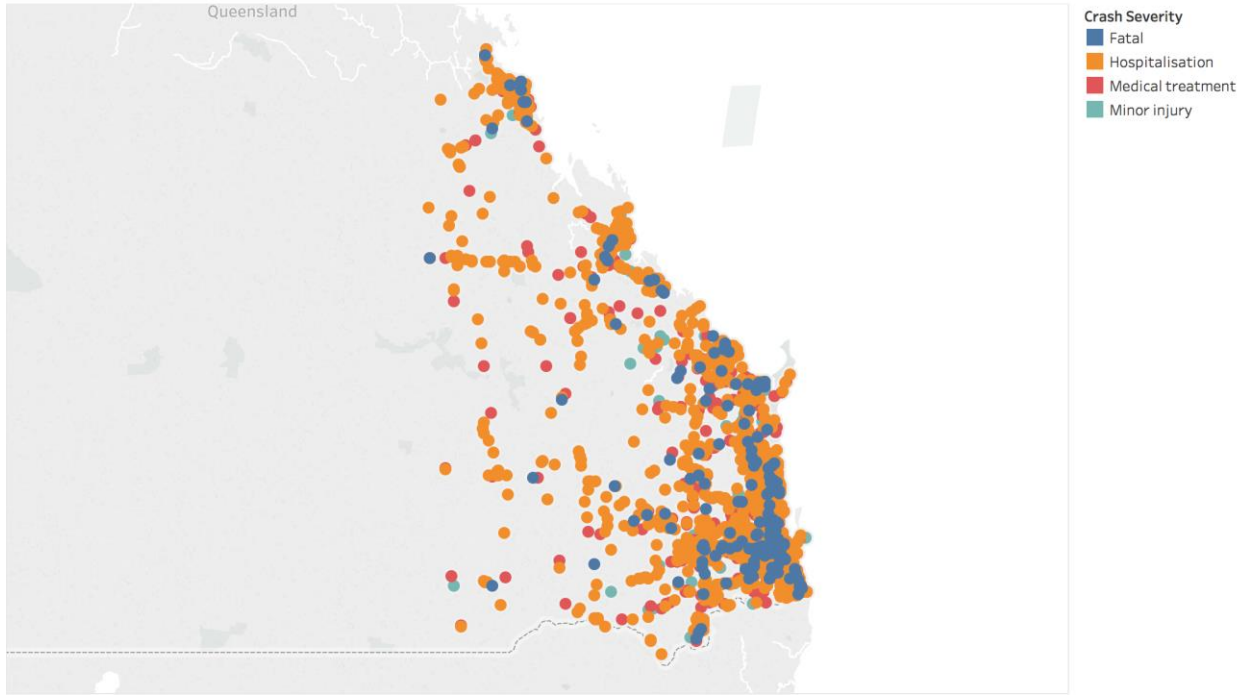
Problem Solving with Data

- Aim – To show that severity of a crash is predictable.
- Classifying and recognizing an accident as Severe / Non-Severe.
- Predicting crash severity with reasonable error margin.
- Enable to recognize the influential predictors of crash severity.

SE Qld and Sunshine Coast Average Traffic



SE QLD and Sunshine Coast Crashes



Getting the Data I need

Is my Data Fit for Use

- **Carried out Data Quality process like,**

1. Discarding Incomplete Rows

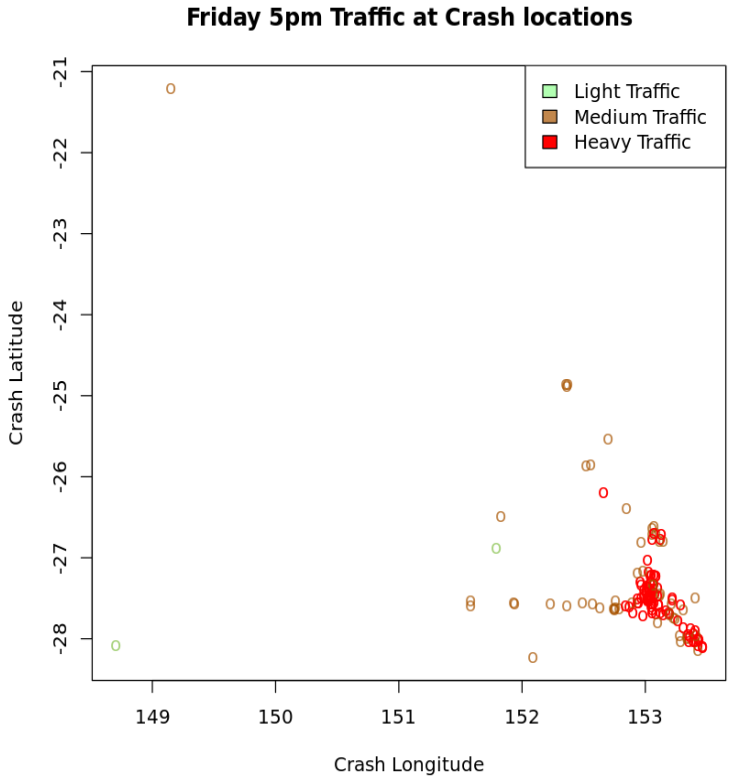
2. Location wise bidirectional Traffic data aggregation.

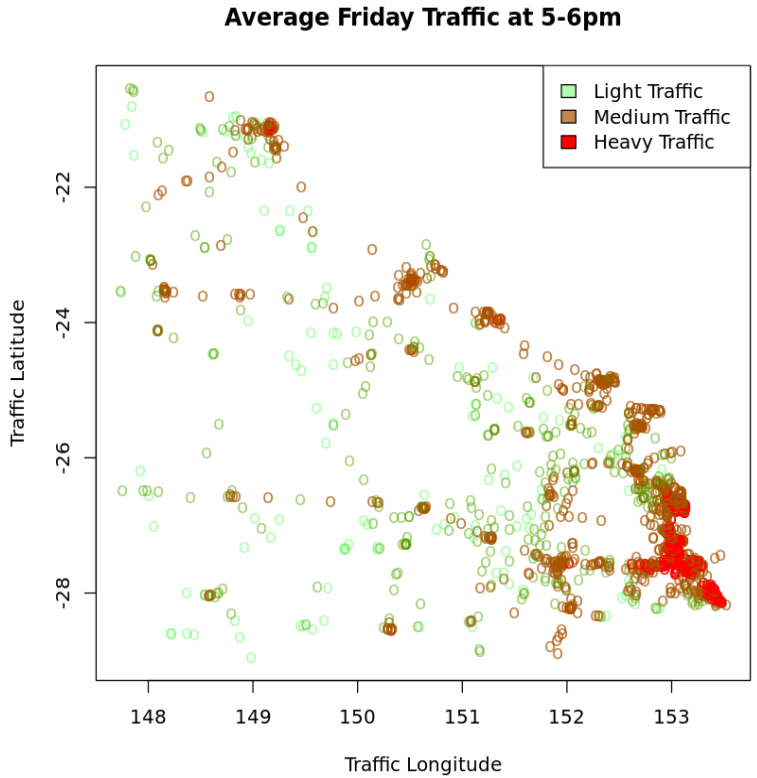
3. Cleansing of speed dimension.

- **Carried out Data Integration process like,**

1. Estimated traffic at crash locations using K-NN Regression technique

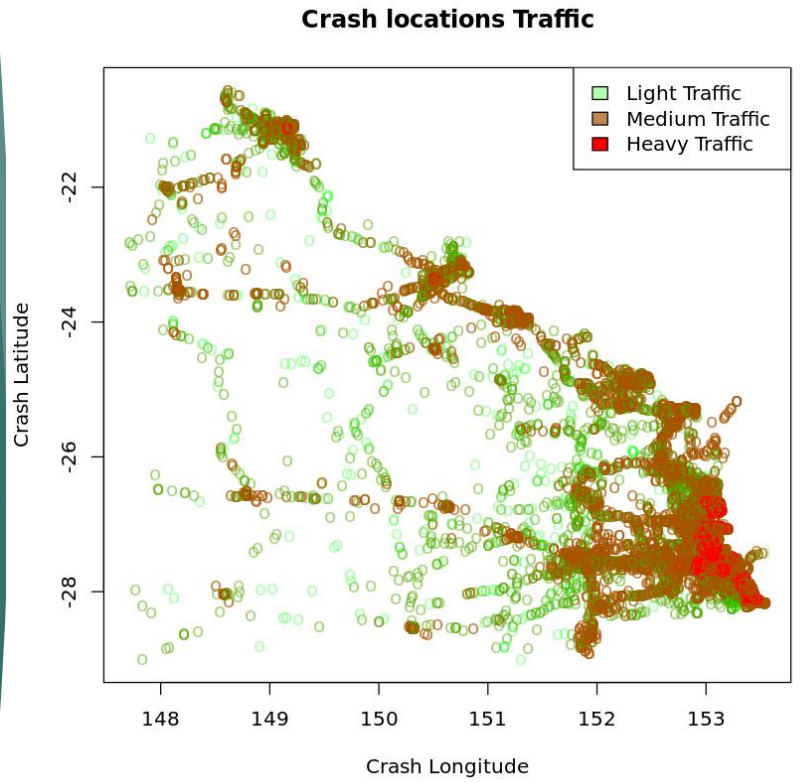
2. Generalized Multi Valued attribute ‘Severity’ to binary valued attribute.

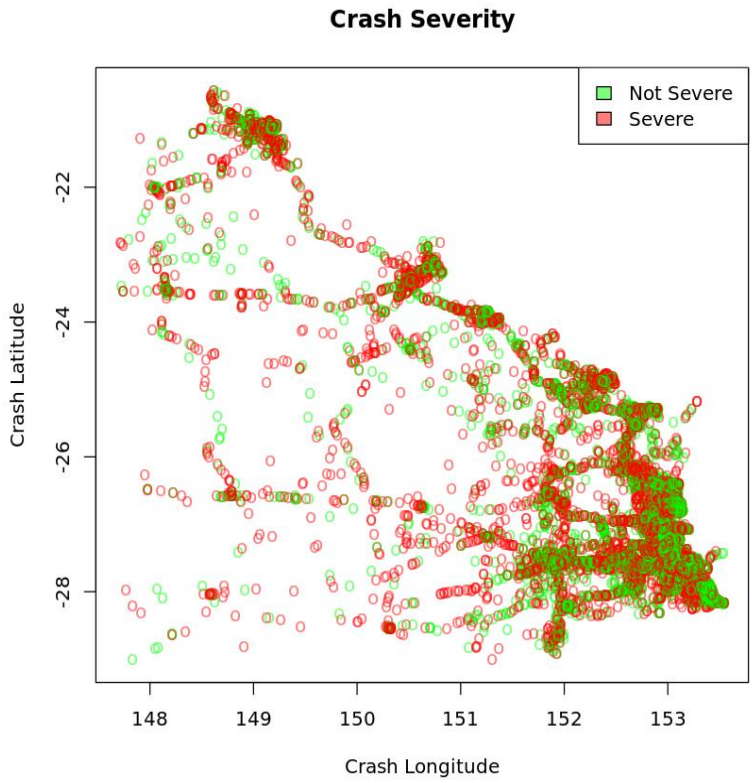


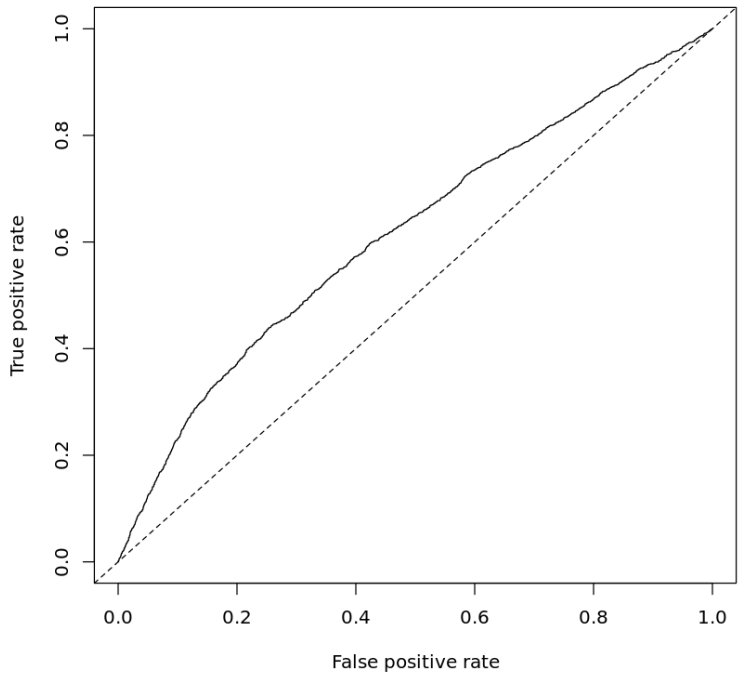


Data Exploration

- Traffic at all crash locations post (left)
- Crash Severity at all locations (right)







AUC on testing data: 0.6154
Prediction Accuracy: 62.3%

```
logfit <- glm(sev ~
              longitude + month + hour + speed_limit*traffic + I(traffic^2) +
              road_hor_al + lighting_cond,
              data = Crash_Sev_train, family=binomial)
summary(logfit)
```

```
Call:
glm(formula = sev ~ longitude + month + hour + speed_limit *
    traffic + I(traffic^2) + road_hor_al + lighting_cond, family = binomial,
    data = Crash_Sev_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7571  -1.0077  -0.8723   1.2441   1.6158

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             1.577e+01  2.232e+00   7.067 1.58e-12 ***
longitude                              -1.025e-01  1.452e-02  -7.055 1.72e-12 ***
monthAugust                            -6.899e-02  6.214e-02  -1.110  0.26689
monthDecember                           2.286e-02  6.357e-02   0.360  0.71914
monthFebruary                          -9.035e-02  6.412e-02  -1.409  0.15882
monthJanuary                            4.541e-02  6.484e-02   0.700  0.48373
monthJuly                               1.046e-01  6.202e-02   1.686  0.09183 .
monthJune                              -7.247e-02  6.283e-02  -1.153  0.24873
monthMarch                             -1.320e-01  6.226e-02  -2.120  0.03400 *
monthMay                                3.353e-03  6.088e-02   0.055  0.95608
monthNovember                          -6.374e-02  6.249e-02  -1.020  0.30777
monthOctober                           -7.581e-02  6.210e-02  -1.221  0.22220
monthSeptember                         -1.244e-01  6.427e-02  -1.936  0.05285 .
hour                                   -4.143e-03  2.673e-03  -1.550  0.12119
speed_limit                             8.421e-03  8.844e-04   9.522  < 2e-16 ***
traffic                                -2.774e-04  6.816e-05  -4.069 4.72e-05 ***
I(traffic^2)                            8.710e-08  1.093e-08   7.969 1.61e-15 ***
road_hor_alCurved - view open          -2.845e-01  6.605e-02  -4.307 1.65e-05 ***
road_hor_alStraight                    -5.058e-01  6.198e-02  -8.160 3.34e-16 ***
lighting_condDarkness - Not lighted     3.414e-02  6.264e-02   0.545  0.58576
lighting_condDawn/Dusk                 -1.716e-01  5.921e-02  -2.898  0.00375 **
lighting_condDaylight                  -2.350e-01  3.981e-02  -5.904 3.56e-09 ***
speed_limit:traffic                    -3.065e-06  7.141e-07  -4.292 1.77e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

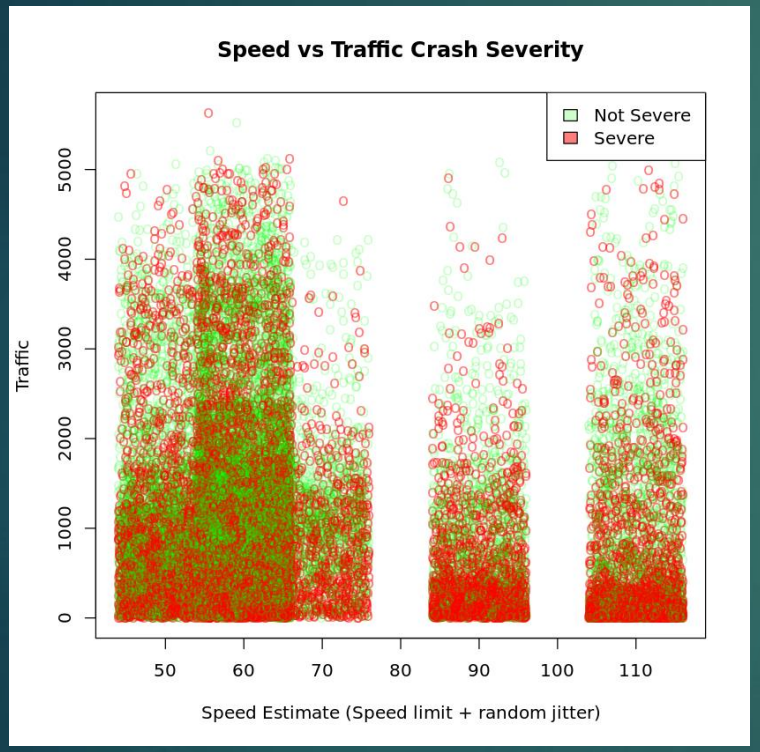
Making the Data Confess

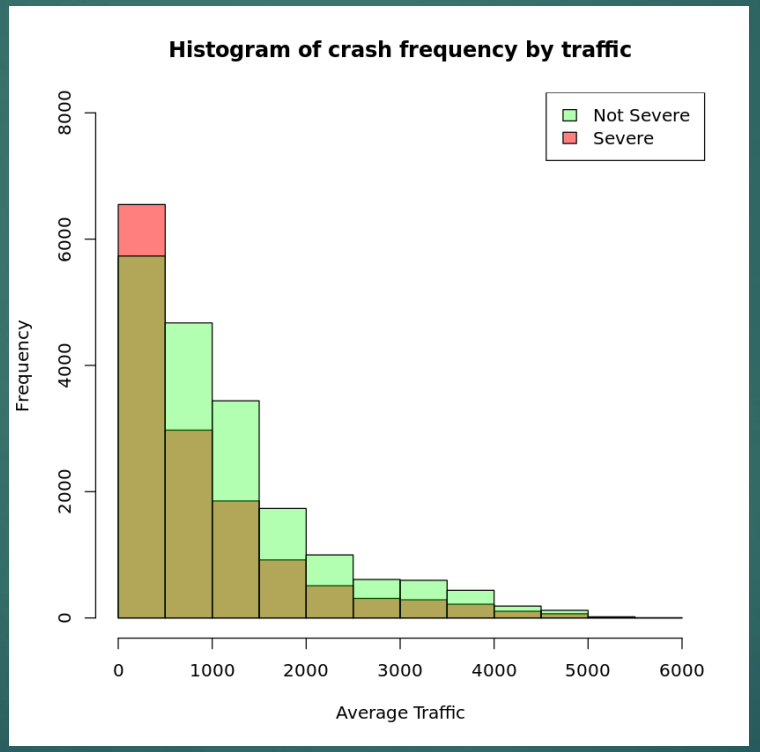
using Logistic Regression

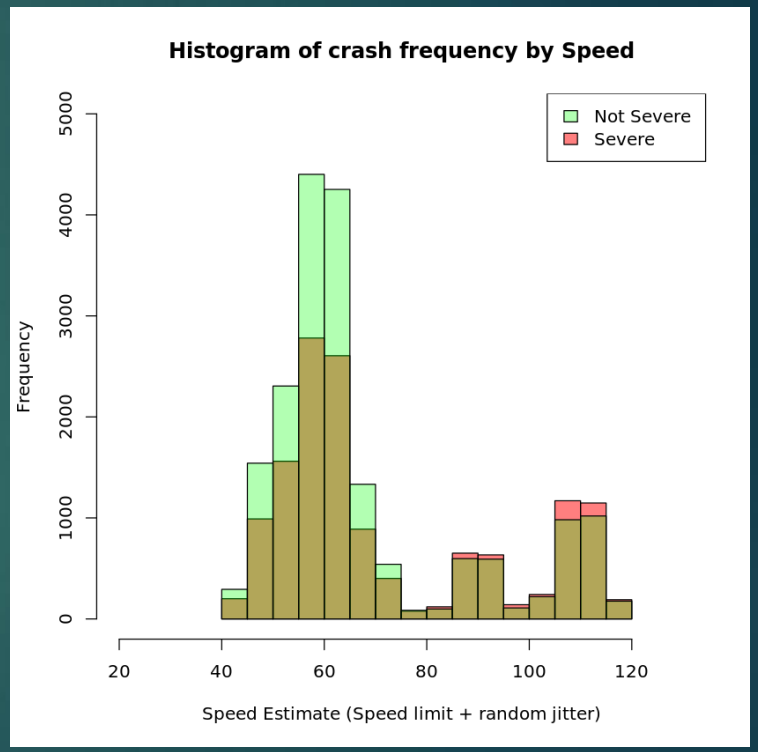
Story Telling with Data

Traffic and Estimated Speed

- Severe crashes tend to occur at High Speed but in Low Traffic conditions (Red cluster).
- None Severe crashes are concentrated on Low speed but High traffic regions (Green cluster).



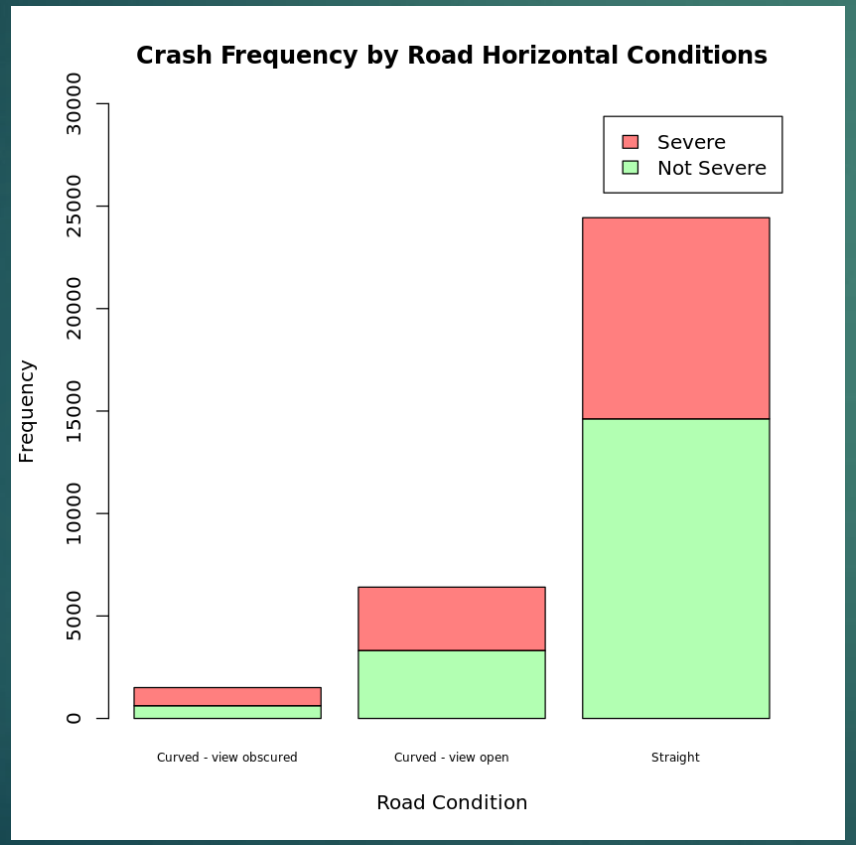


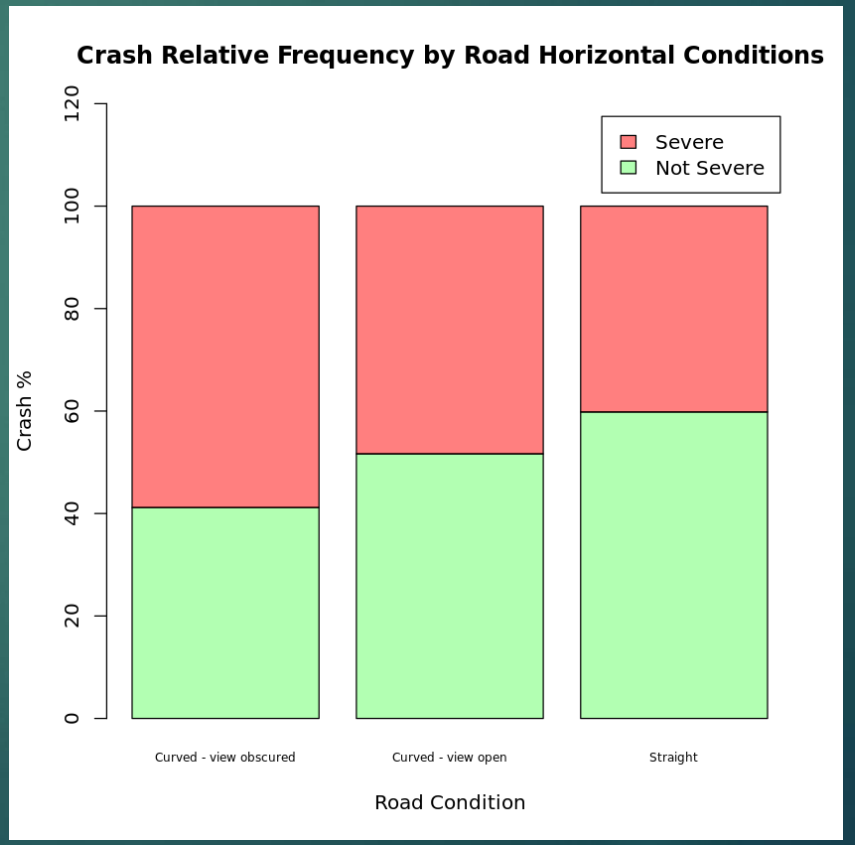


Story Telling with Data

Road Condition

- Severity is clearly influenced by the horizontal shape of the road.
- Crashes on curved roads are more severe.
- Around 60% of crashes on Curved-Obscured view roads are Severe compared to 40% on Straight roads.

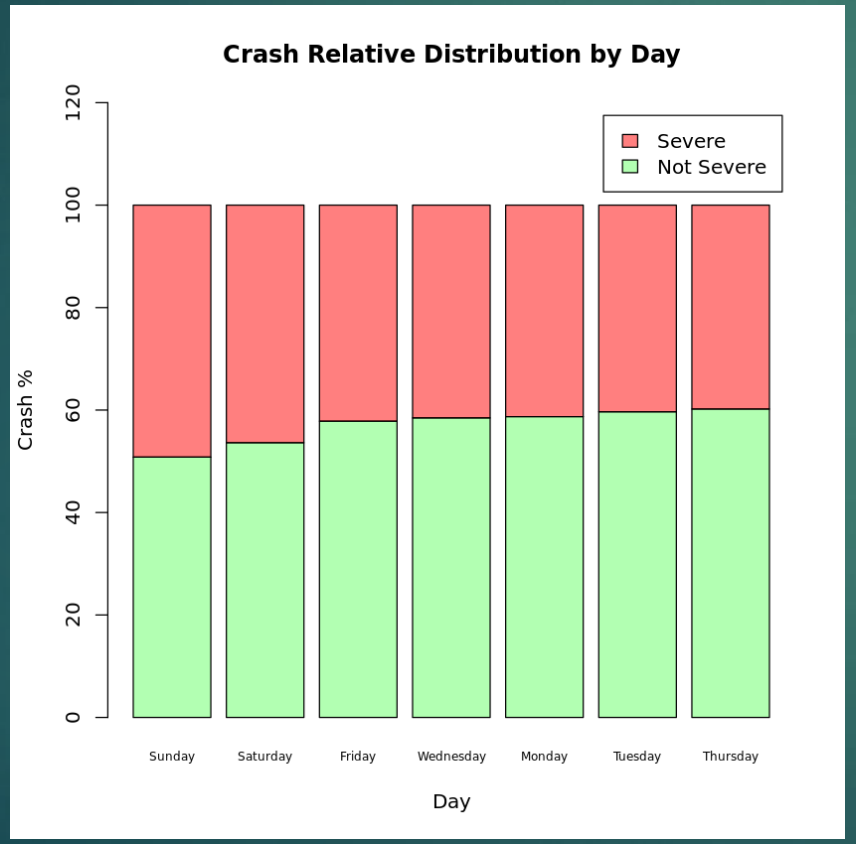


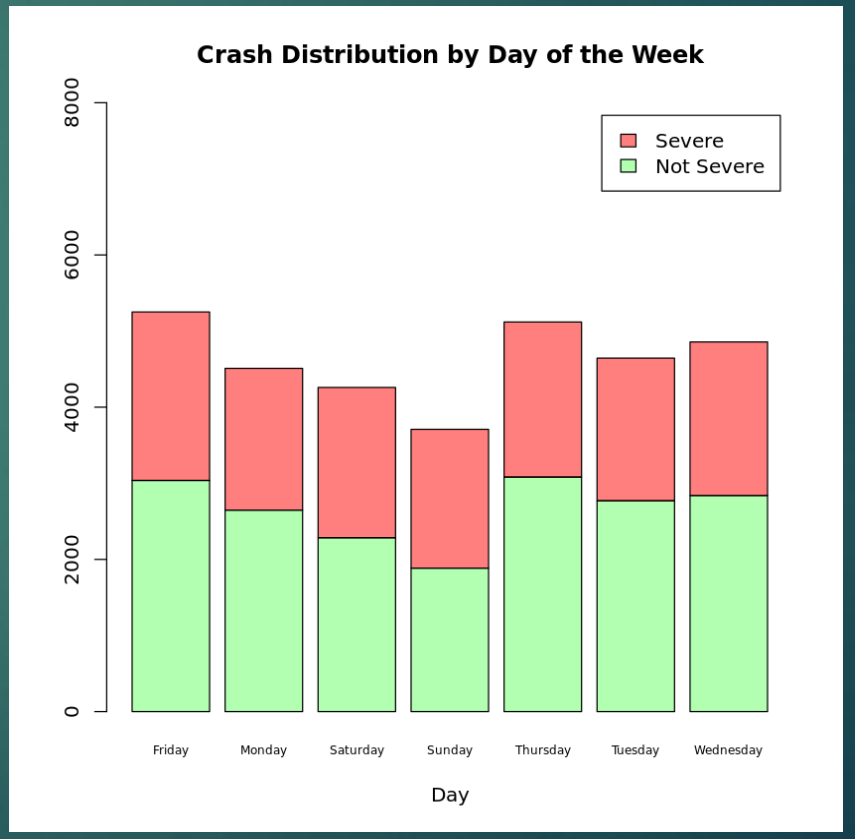


Story Telling with Data

Day of Week

- Day of the week is NOT very influential to predict Severe and Not Severe crashes.
- Amongst the total # crashes, Friday and Thursday lead the chart.
- But for Severe crashes, Sunday and Saturday are the highest with 49.2% and 46.4% respectively.

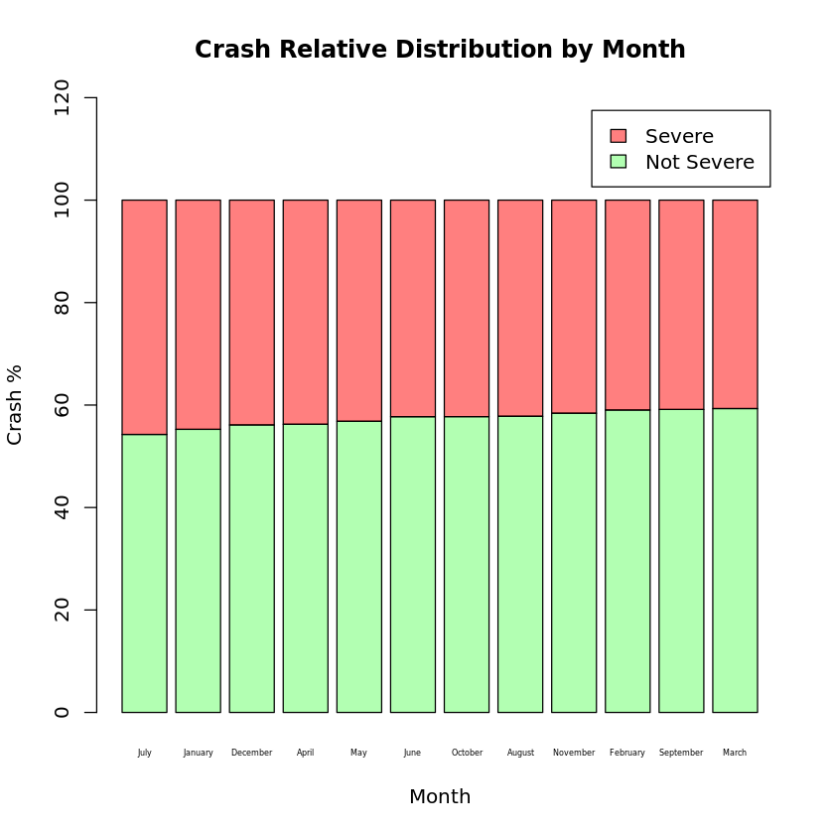


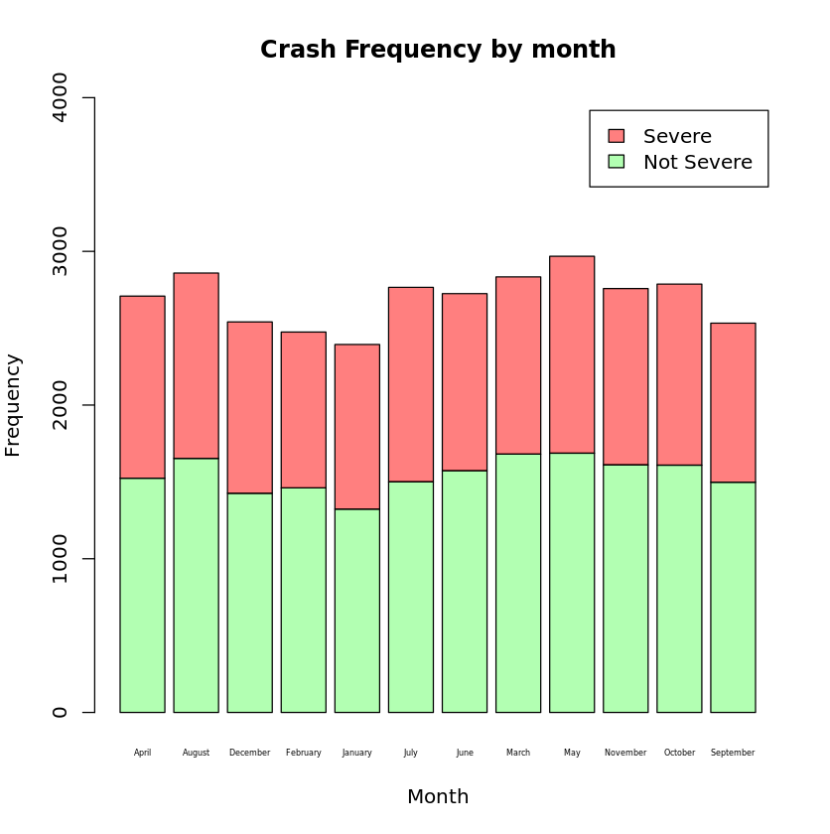


Story Telling with Data

Month of Year

- Month of the year is NOT very influential to predict Severe and Not Severe crashes.
- Amongst the total # crashes, May and August lead the chart.
- But for Severe crashes, July and January are the highest with 45.7% and 44.7% respectively.

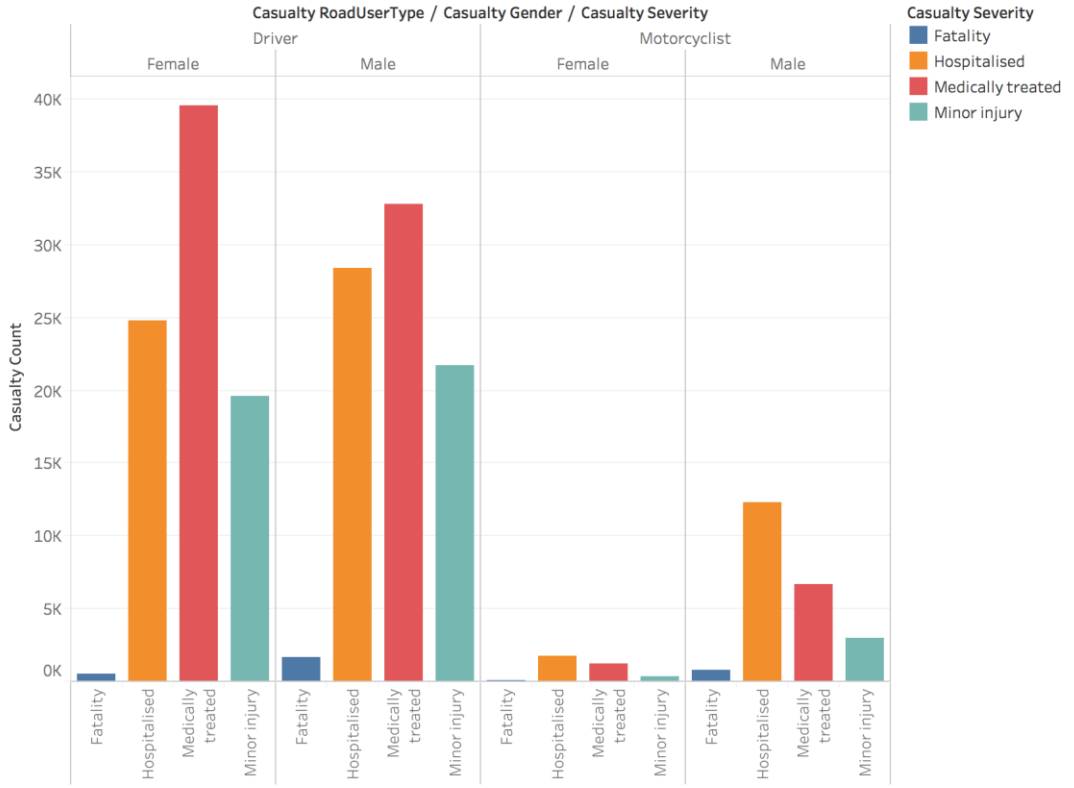




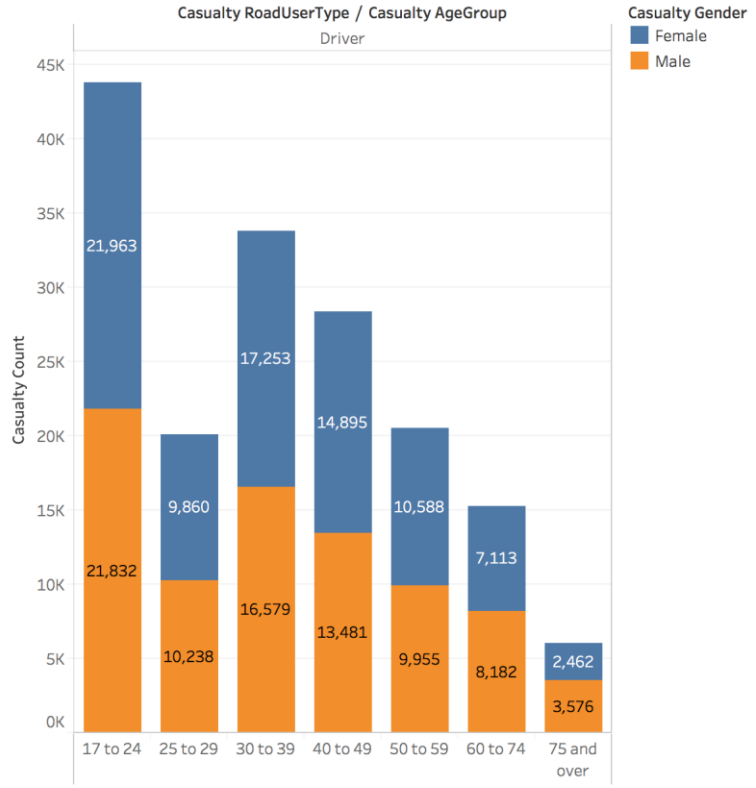
Conclusion

- Many crash factors are NOT captured in the dataset.
- Reliability on Average Traffic and Speed NOT enough to predict crash severity.
- External dataset show how Gender (left) and Age (right) influence crash severity.

Casualty by Road User/Gender/Severity



Driver Casualty by Gender/Age group



Thanks