**Group 5 Sub-Report: PCA Observations and Outcomes**

Definition of Principal Component Analysis according to Wikipedia:

“Principal component analysis is a popular technique for analyzing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while preserving the maximum amount of information, and enabling the visualization of multidimensional data.”

In this document, we will look at categorical attribute associations mined from running chi2 tests on the two randomly sampled data frames. Even though every combination of categorical attributes was tested, the only results kept were those with p values such that 0 < p < .05. This keeps only potentially interesting associations for us to analyze.

Included below is a list of the most interesting associations, after further eliminating results from tests that compare irrelevant associations. For example, anything relating caseid and another attribute is removed.

Dataframe containing joined Collisions and Parties dataframes:

Columns: ('weather\_1', 'location\_type') Chi-squared: 22.1926, p-value: 0.0354

Columns: ('weather\_1', 'collision\_severity') Chi-squared: 56.1568, p-value: 0.0002

Columns: ('weather\_1', 'pedestrian\_action') Chi-squared: 66.0544, p-value: 0.0002

Columns: ('weather\_1', 'party\_safety\_equipment\_1') Chi-squared: 151.3247, p-value: 0.0011

Columns: ('weather\_1', 'movement\_preceding\_collision') Chi-squared: 150.0624, p-value: 0.0046

Columns: ('location\_type', 'collision\_severity') Chi-squared: 31.4739, p-value: 0.0001

Columns: ('location\_type', 'road\_surface') Chi-squared: 15.3754, p-value: 0.0175

Columns: ('location\_type', 'road\_condition\_1') Chi-squared: 35.6775, p-value: 0.0004

Columns: ('location\_type', 'party\_sobriety') Chi-squared: 32.4249, p-value: 0.0012

Columns: ('location\_type', 'party\_drug\_physical') Chi-squared: 24.2555, p-value: 0.0070

Columns: ('ramp\_intersection', 'road\_condition\_1') Chi-squared: 82.7321, p-value: 0.0018

Columns: ('collision\_severity', 'primary\_collision\_factor') Chi-squared: 43.5448, p-value: 0.0002

Columns: ('collision\_severity', 'cellphone\_use\_type') Chi-squared: 39.3109, p-value: 0.0253

Interesting FP Tree findings:

Support: 0.22857142857142856, Itemsets: location\_type: highway

Support: 0.06666666666666667, Itemsets: location\_type: highway, party\_safety\_equipment\_1: air bag not deployed

Support: 0.06666666666666667, Itemsets: 0, movement\_preceding\_collision: ran off road

Support: 0.06666666666666667, Itemsets: location\_type: highway, party\_sobriety: had been drinking, under influence

Support: 0.05714285714285714, Itemsets: movement\_preceding\_collision: making left turn