**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.”

Here, we take a visualization-based data mining approach to PCA, plotting numerical attribute columns in a series of cross-plots to observe correlations and interesting data clusters throughout the training data frames collected from the ‘collisions’, ‘parties’, and ‘victims’ tables of the switrs.sqlite database file.

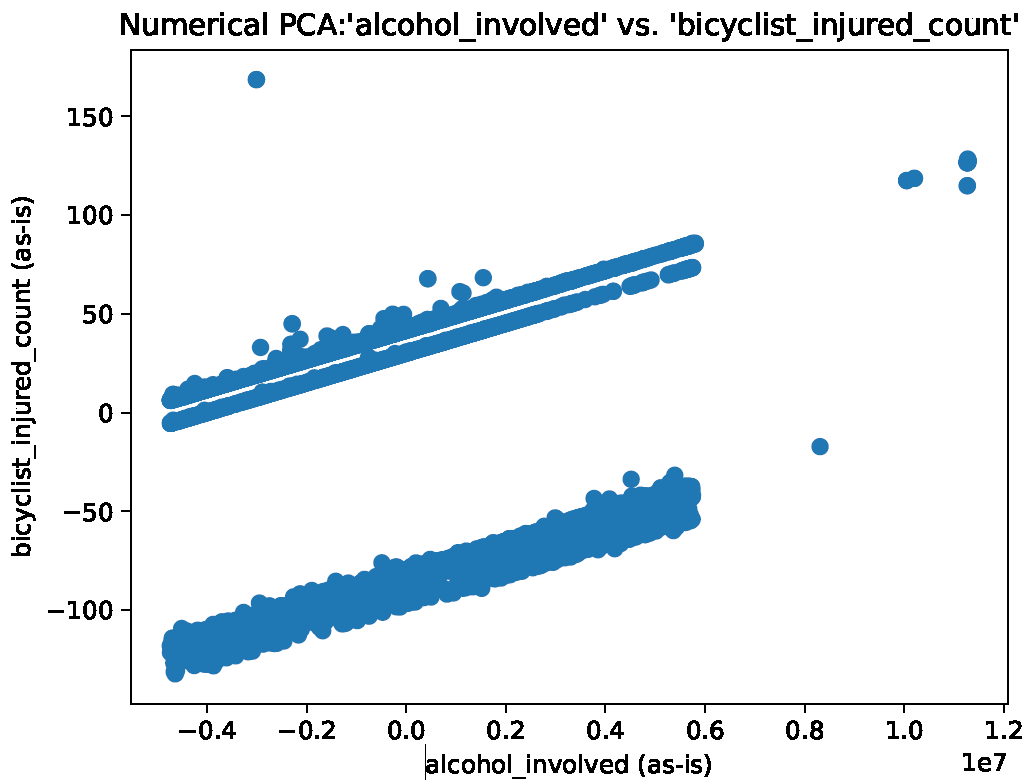
We begin with a cross-comparison of the numeric attributes under two alternate data preparation schemes: “as-is”, and “min-max normalized”. “As-is” attribute comparisons reveal general data trends, as well as correlation patterns between attributes with the units presented as they exist in the source data. “Min-max normalization” standardizes numeric attribute columns with respect to the minimum and maximum values in each attribute column.

We begin by viewing cross-plots of non-standardized data to observe general correlations and patterns, then view an identical set of cross-plots produced with min-max normalized data to reveal meaningful clusters highlighted by the normal vector space.

Note: Please disregard the yellow points, they represent a code artifact and are not semantically meaningful under this analysis.

BEGIN: AS-IS NUMERIC PCA CROSS-PLOTS

**Plot:**



**Observations:**

**Action:**

**Plot:**

**A blue dots on a white background

Description automatically generated**

**Observations:**

**Action:**

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:**

**Action:**

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:**

**Action:**

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:**

**Action:**

**Plot:**

A blue dot diagram with white background

Description automatically generated

**Observations:**

**Action:**

**Plot:**

A blue dots on a white background

Description automatically generated

**Observations:**

**Action:**

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:**

**Action:**

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:**

**Action:**

**Plot:**

A diagram of a cellphone

Description automatically generated

**Observations:**

**Action:**

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:**

**Action:**

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:**

**Action:**

**Plot:**

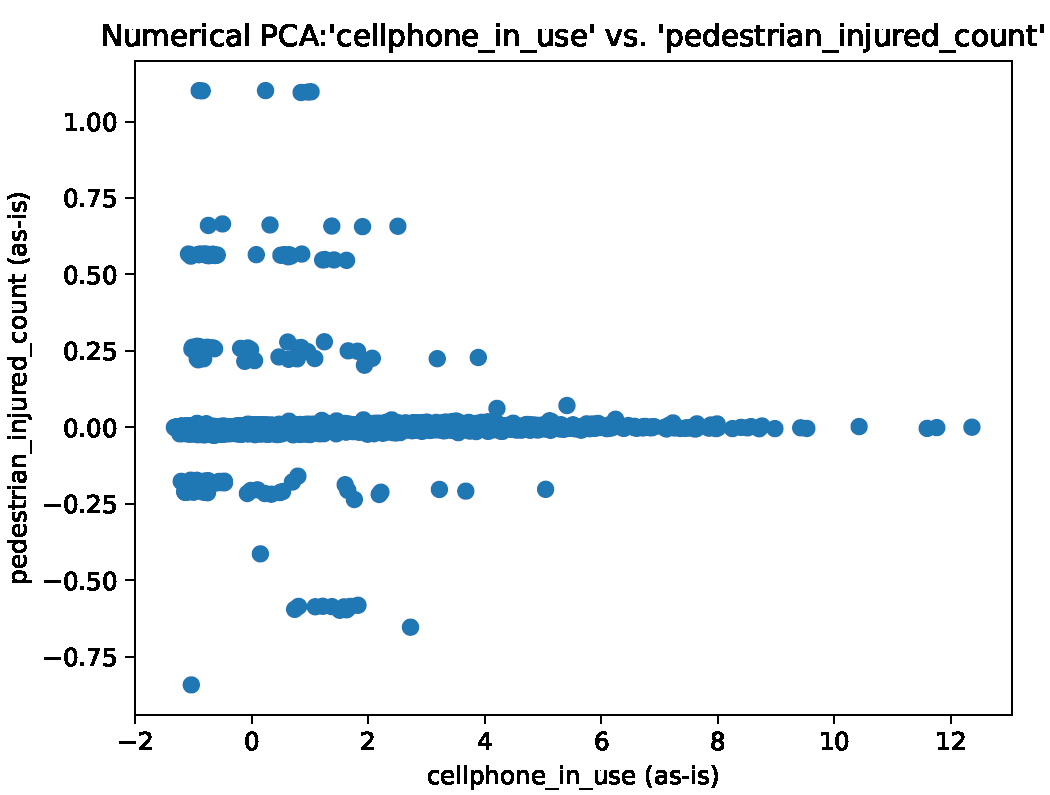
A graph with blue dots

Description automatically generated

**Observations:** ‘cellphone\_use’ does not appear correlated with ‘pedestrian\_collisions’

**Action:**

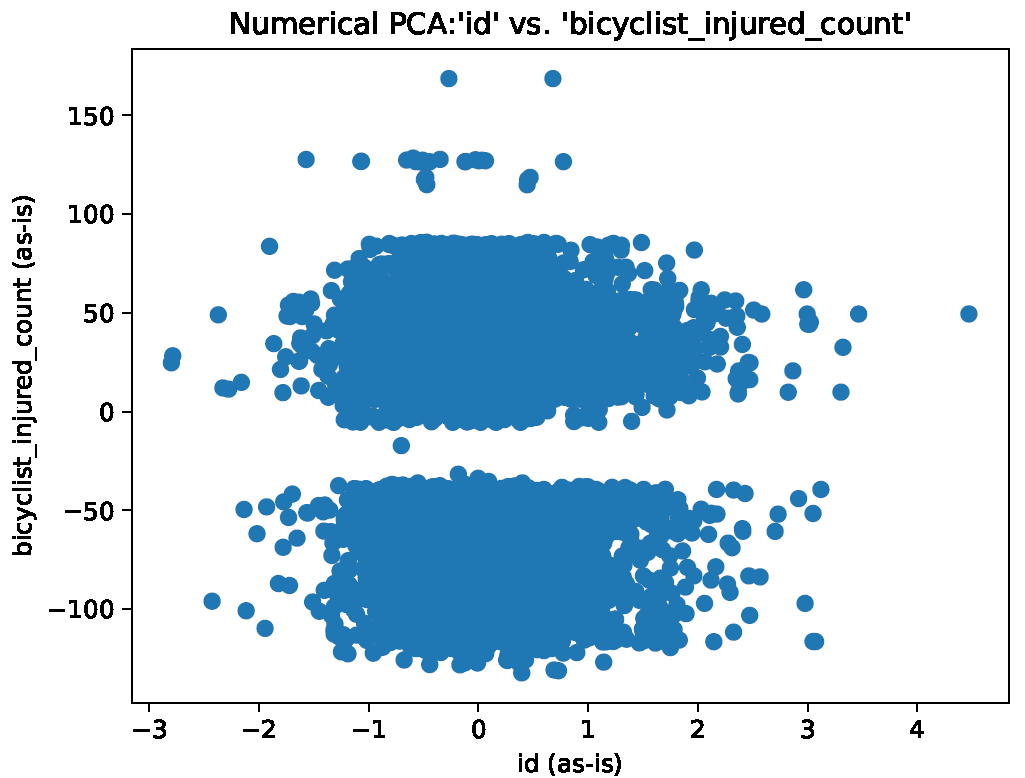
**Plot:**



**Observations:**

**Action:**

**Plot:**



**Observations:** the ratio of injured to uninjured bicyclists looks to be about 50%

**Action:**

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** combining this cross-section with the latitude dimension should produce a meaningful data cube structure. Use ‘pedestrian\_killed\_count’ as the dependent variable.

**Action:**

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** combining this cross-section with the latitude dimension should produce a meaningful data cube structure. Use ‘severe\_injury\_count’ as the dependent variable.

**Action:**

**Plot:**

A diagram of blue dots

Description automatically generated with medium confidence

**Observations:** Desirable clustering.

**Action:**

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** another latitude+longitude spatial data cube snapshot (motorcyclists killed by region).

**Action:** make a data cube

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** most motorcycle accidents involve younger drivers.

**Action:**

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** most bicycle collisions occur away from public property, such as in a residential area.

**Action:** This graphic is important because it shows that the chances of colliding with a bicycle are higher in domestic settings than in urban settings. AI systems can easily detect and avoid bicyclists, using contextual clues about the environment to ascertain the relative likelihood of encountering a bicyclist.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:**

**Action:** the distribution shown here indicates that cell phone involvement in domestic areas is likely under-reported. ‘not\_private\_property’ responses near 0.0 represent a “no response”, indicating that the cellphone usage in !not\_private\_property = private\_property occurs much less frequently than in urban settings. Cell phone usage is still dangerous in domestic settings, the lack of phone control in these environments could explain some of the other distributions we have seen.

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** unexpected distribution

**Action:** younger drivers tend to be involved in residential accidents, as ‘party\_age’ increases so does the likelihood of a collision in an urban area.

**Plot:**

A diagram of a motorcycle accident

Description automatically generated with medium confidence

**Observations:** Bimodal pattern observed.

**Action:** good candidate for regression fitting.

**Plot:**

A blue and red dot pattern

Description automatically generated with medium confidence

**Observations:** dead motorcyclists tell no lies.

**Action:** examine cluster to determine underlying causes of deadly motorcycle accidents. The sparse cluster to the right may link to autopsy reports.

**Plot:**

A graph with blue and red dots

Description automatically generated

**Observations:** Older victims tend to get more mashed up in automobile accidents.

**Action:** AI system features that provide enhanced vehicle road support for senior citizens would probably do very well under certain demographics, as elderly people are known to have more disposable income to spend on enhanced (software based) safety systems.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** vehicles with additional passengers do not appear to add to the safety risk.

**Action:** vehicles with many passengers are more likely to be a contributing factor to many collisions due to the increased number of distractions experienced by the driver. The visible injury count of these passengers does not appear to increase in correlation to the number of passengers in the vehicle. Therefore, distraction prevention systems tailored to a younger crowd would probably be the best choice for this demographic. Younger people are more likely to purchase “high value” vehicles, meaning vehicles that are reliable, easy to afford, and as feature rich as possible within certain budgetary restrictions.

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** Guess who uses their cell phones while driving the most!

**Action:** younger parties are more likely to be involved in a phone-related collision than their parents or grandparents.

**Plot:**

A diagram of a graph

Description automatically generated with medium confidence

**Observations:**

**Action:**

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** (CLUSTERING)

**Action:**

END AS-IS NUMERICAL PCA ON COLLISIONS SET

BEGIN MIN-MAX NORMALIZED PCA ON COLLISIONS DATA

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:**

**Action:**

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** (CLUSTERING)

**Action:**

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** (GOOD CLUSTERING)

**Action:**

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** (GOOD CLUSTERING)

**Action:**

**Plot:**

A diagram of a graph

Description automatically generated with medium confidence

**Observations:** (GOOD CLUSTERING)

**Action:**

**Plot:**

A diagram of a graph

Description automatically generated with medium confidence

**Observations:** (GOOD CLUSTERING)

**Action:**

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** (INTERESTING SUBJECT)

**Action:**

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:**

**Action:** collisions at intersections tend to involve more injuries.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:**

**Action:**

**Plot:**

**Observations:**

**Action:**

**Plot:**

**Observations:**

**Action:**

**Plot:**

**Observations:**

**Action:**

**Plot:**

**Observations:**

**Action:**

**Plot:**

**Observations:**

**Action:**

**Plot:**

**Observations:**

**Action:**

**Plot:**

**Observations:**

**Action:**