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

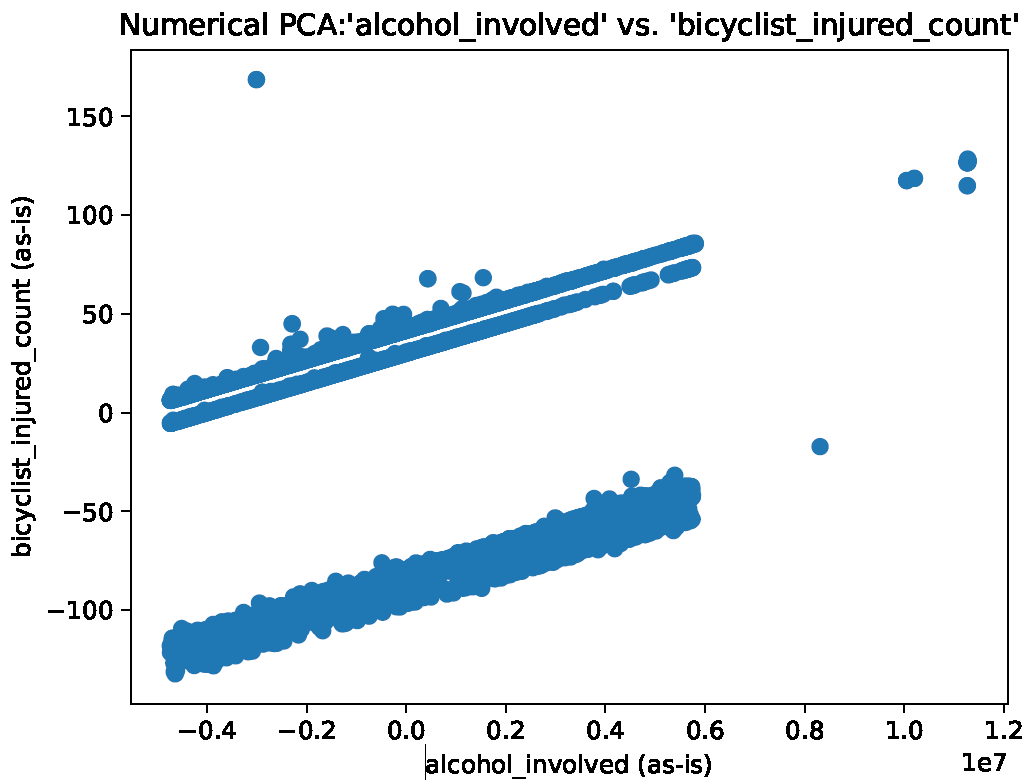
Here, we take a visualization-based data mining approach to PCA by examining numerical attribute columns using a series of cross-plots to observe correlations and interesting clusters throughout the training set data frames collected from the ‘collisions’, ‘parties’, and ‘victims’ tables of the switrs.sqlite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

“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.” -current definition of PCA, according to Wikipedia.

Team 5’s numerical PCA was conducted using the pre-processed data training tuples, collected from the switrs.sqlite source file, by way of PANDAS data frames. The algorithm that we developed for this task (see ‘SWITRS\_Python\_Notebook.ipynb’ for more details about the numeric PCA algorithm) performs an exhaustive cross-comparison of 86 original numeric and transformed-binary nominal attribute columns and generates 992 plots for the df\_collision\_parties data frame, which is itself composed of an inner-join of matching tuple sets collected from the ‘collisions’ and ‘parties’ tables of the switrs source file. The algorithm has been proven to complete this task in under 9 minutes (performed on a Lenovo x86-64 machine), in Θ(n2) computational complexity. This algorithm generates a cross-plot for every attribute intersection that can be observed in the ‘df\_collision\_parties’ joined data frame object composed of 10,000 random samples.

By viewing cross-plots of the non-standardized data in these frames, we can observe many useful correlations and clustering patterns that might be obscured by noise or otherwise hidden within the wide scope of the massive data set. Insights into these underlying trends and distributions provide us with useful guidance that can assist us in constructing classifiers capable of distinguishing the switrs common driver classes by their varying attribute qualities. The PCA plots themselves can also be the source of much information and provide the viewer with a general intuition about the data set being examined.

**Plot:**



**Observations:** a linear trend is observed between alcohol-involvement and the rate of bicyclist injuries.

**Action:** This pattern can easily be fitted to a linear regression model and used to predict numeric outcomes in the SWITRS data set.

**Plot:**

**A blue dots on a white background

Description automatically generated**

**Observations:** good separation is observed, based on whether the accident involved a bicycle related injury or not. From the looks of it, bicyclists appear to get injured a lot!

**Action:** this is likely a good categorization to split data points on in active classifier and/or decision models.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** Again, a good clustering pattern is observed here based on the bicyclist\_injured\_count numerical attribute.

**Action:** We can leverage this separation in most of the classification models that we plan on using in this assignment.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** A four-way split.

**Action:** this cross-section of attributes appears to achieve a near optimal class separation based on ‘bicyclist\_injured\_count’ and the ‘killed\_victims’ binary attributes. We can leverage this separation in the assumptions of our forthcoming classification models.

**Plot:**

A blue dot diagram with white background

Description automatically generated

**Observations:** killed\_victims appears to be a reasonable class identifier in its own right, although some null response and bias may be present.

**Action:** none.

**Plot:**

A blue dots on a white background

Description automatically generated

**Observations:** younger drivers are more likely to be involved in collisions of any kind. (with or without bicyclists)

**Action:** Parents of young drivers are a particularly sensitive class in regard to this figure. AI systems that guide or tutor novice drivers is likely to be a lucrative sub-category of automotive AI technologies.

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** it looks like many of the data points within this set may constitute some form of non-response in the ‘not\_private\_property’ attribute column. (i.e., those that hover about the 0.0 y-axis marker).

**Action:** It is important to be aware of the uneven distribution that is present in the ‘not\_private\_property’ attribute so that we are not misled when analyzing subsequent patterns involving this attribute.

**Plot:**

A diagram of a cellphone

Description automatically generated

**Observations:** It appears that younger generations do indeed spend more time on their cellphones, apparently even when driving.

**Action:** cell phone intervention while driving is a tricky subject. On the one hand, cell phones almost certainly contribute to traffic collisions. On the other hand, drivers very much enjoy their cell phones and do not like to be separated from them. This presents a conundrum in terms of designing AI features that reduce the use of cell phones while driving. Keeping the driver engaged with some form of forward-facing camera system, fitted with object recognition system, is considered by many researchers (see sources cited for this project) to be the best user interface for engaging users with AI collision detection and avoidance systems. This presumption is based on the conclusion that presenting evidence of dangerous driving to the user may condition his or her response.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:**

**Action:**

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:**

**Action:**

**Plot:**

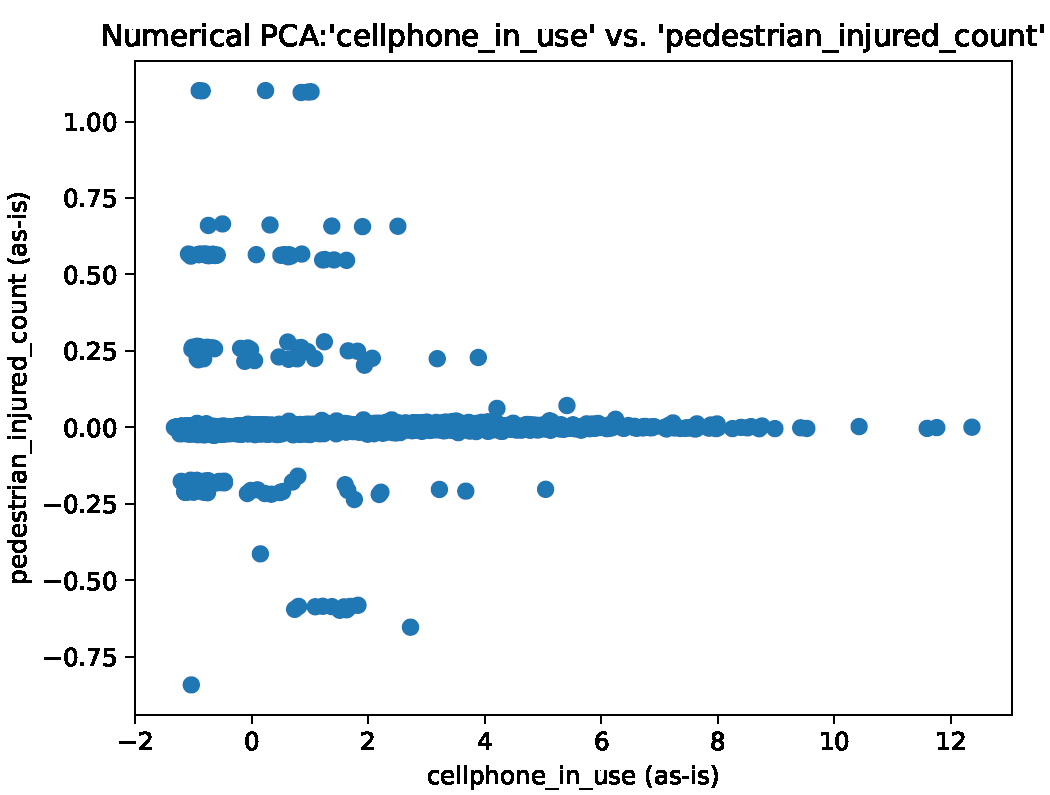
A graph with blue dots

Description automatically generated

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

**Action:**

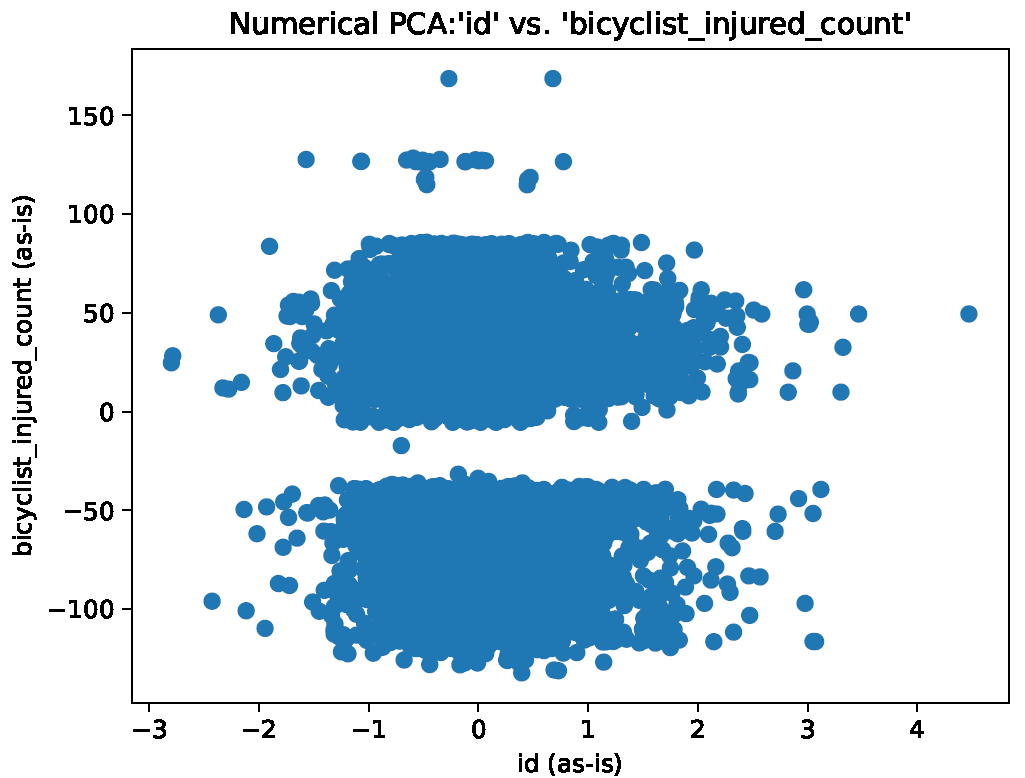
**Plot:**



**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 number of dots

Description automatically generated

**Observations:** younger drivers tend to have more difficulty at intersections.

**Action:** If the AI system is aware of the driver’s approximate age, enhanced safety features can be turned on whenever (perhaps the family car) is approaching an intersection. An AI system can also act as a driving coach!

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** collisions involving young drivers are more likely to be fatal.

**Action:** Parents are often sentimental to the fact that teenage driving can be dangerous. AI based safety features may test well with audiences of concerned parents.

**Plot:**

A diagram of a party age

Description automatically generated with medium confidence

**Observations:** collision with a motorcycles (as well as non-motorcycles) decreases with party\_age.

**Action:** Note that this is a cross-section of a binomial attribute ‘motorcycle\_collision’, which fits under a yes or no category, and ‘party\_age’ on the x-axis, demonstrating the decrease in collisions of both types as party\_age increases (as evidenced by the clustering and relative support shown in the figure).

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** distribution showing the relative likelihood of dying in a motorcycle accident.

**Action:** This is an interesting statistic.

**Plot:**

A graph with blue dots and red circle

Description automatically generated

**Observations:** this cluster appears to represent a cluster of reckless teenage drivers.

**Action:** AI systems may be able to monitor environmental conditions and determine the probability of being involved in an accident under the circumstances. Hardline intervention measures are usually avoided at all costs for the sake of customer convenience. In a life-or-death situation or when it comes to avoiding (e.g., a bicycle collision at an intersection) however, the best policy may be for the AI system to act fast and ask forgiveness later. Reckless teenage drivers are a major safety concern on US roadways.

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** the odds of a pedestrian being struck by a young driver also decreases with party\_age.

**Action:**

**Plot:**

A diagram of a person's body

Description automatically generated

**Observations:** more pedestrians are likely to be involved in a pedestrian collision involving a minor party.

**Action:** The evidence for this claim is observed by the relative support shown above.

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** there appear to be more injuries associated with collisions involving minor parties.

**Action:** notice the density of the plot on either side of the 0.0 y-axis. There are clearly more injuries observed when the party\_age is low.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** people who have been drinking tend to drive alone, or with a single partner.

**Action:**

**Plot:**

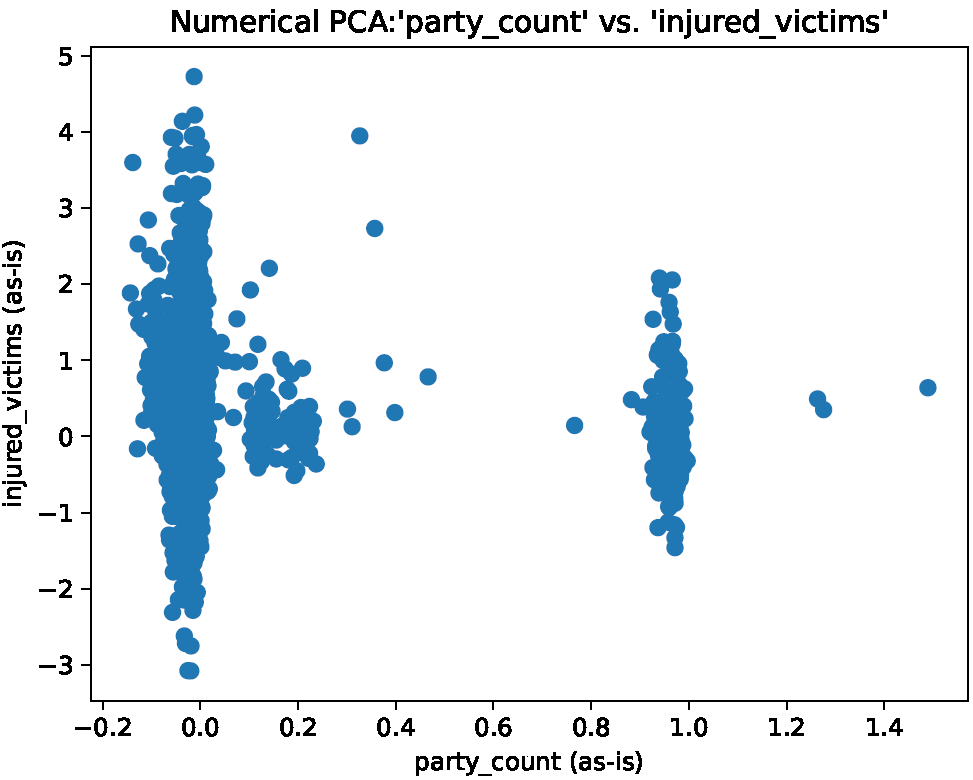
A graph of blue dots

Description automatically generated

**Observations:** people are more likely to use a cell phone when they are alone than with passengers.

**Action:** A safety reminder can be set by the driver of the vehicle to remind them not to use the phone while driving, as cellphone use has negative implications on traffic safety. It’s best not to be *too* annoying with these features, however.

**Plot:**



**Observations:** it is safer to drive with a co-pilot

**Action:** this is not surprising but could indicate a need for AI-features or modes suited to individual drivers. In many ways, the ideal AI system would act as this second passenger, without distracting or annoying the driver. Highlighted forward-facing dash cameras are one of the most promising technological developments that could be tailored to suit this purpose.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** individual drivers have more difficulty at intersections (although they’ll never admit it).

**Action:** A “silent ride-along parter” could be just the right safety feature for more *private* consumers.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** Again, single drivers are more likely to be involved in an accident, whether it involves fatalities or not.

**Action:** The real trick to developing an AI ride-along system is ***stealth***. Since solitary drivers are not likely to tolerate annoying system faults or interruptions for very long, before taking a solitary ride back to the dealership. Stealth is the reason why image-recognition dashboard camera systems has been so popular in recent publications. (See prior works for more details.)

**Plot:**

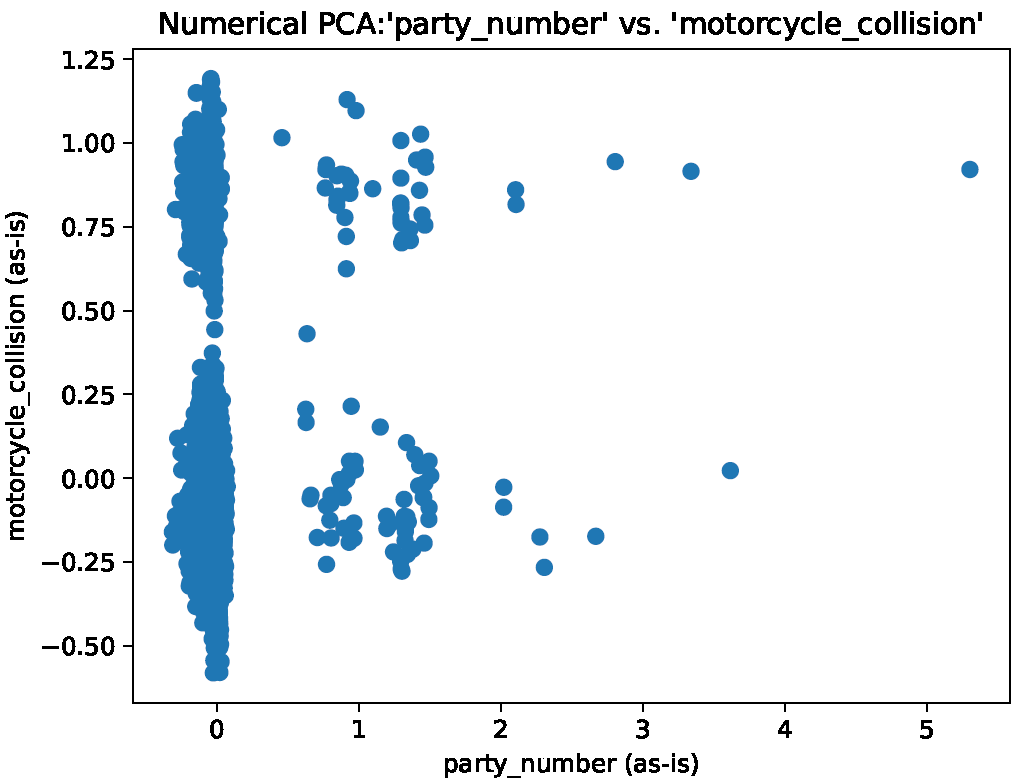
A graph of a number of bicyclists

Description automatically generated

**Observations:** solo-drivers are more likely to hit bicylists.

**Action:**

**Plot:**



**Observations:** solo-drivers would appear to be more likely to hit motorcycles based on this observation.

**Action:** this observation (solo-drivers are more dangerous) is likely to have been affected by the relative abundance of solo-drivers compared with other drivers. This does not however, detract from the numerous safety benefits of driving with a passenger. This effect of relative support on rule generation can be assessed using a chi-square table to compare expected vs. observed members in each category created by the intersection of ‘party\_number’ and ‘motorcycle\_collision’ (for example). A similar analysis can be performed anytime the distribution of relative support is not even, and when the validity of the rule may be affected.

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** a negative correlation between ‘party\_number’ and ‘pedestrian\_collision’. This time the support for single-party drivers does not appear to be clouding the model.

**Action:** determine if this trend is meaningful by observing the integrity of the underlying data. It seems the plot is most clearly saying “the likelihood of a pedestrian collision decreases with party\_number”. Investigate to examine the source of this trend.

**Plot:**

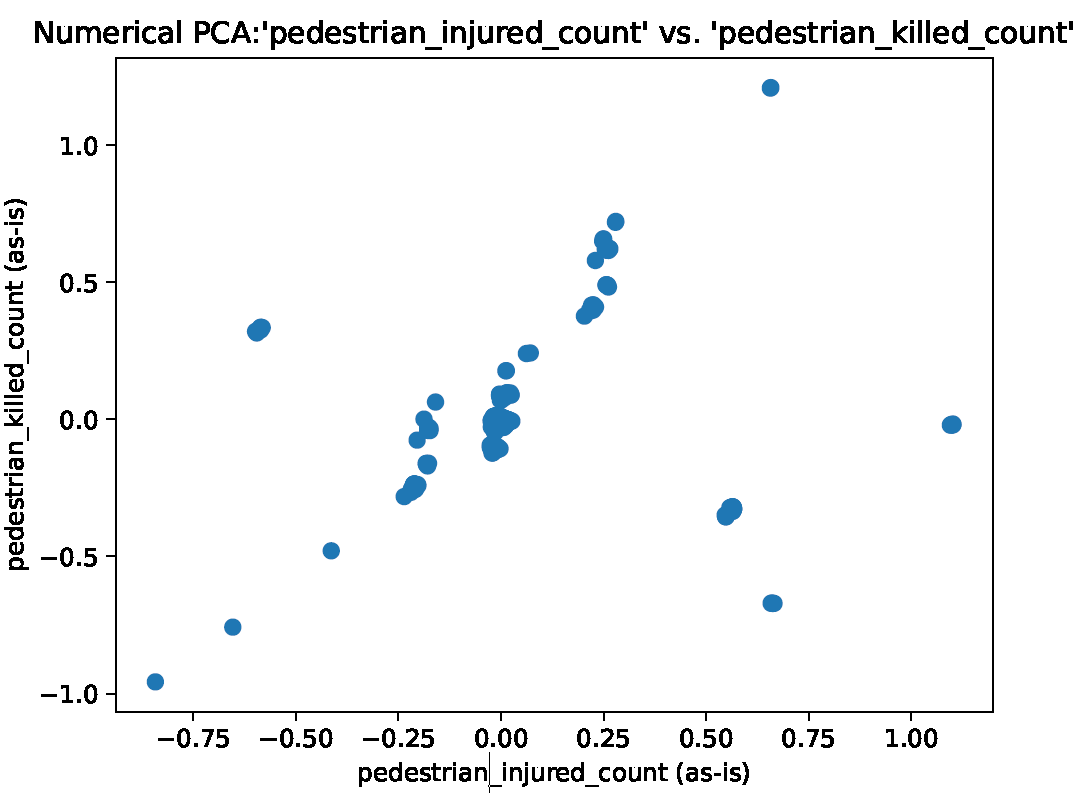
A graph with blue dots

Description automatically generated

**Observations:** Most of the school bus related collisions appear to occur when there are 1-3 people in the car. These poor folks were probably on their way to drop the kids off at school.

**Action:** Life is challenging and messy sometimes. Luckily, a school bus should be an easy object for an AI-guided automobile to detect and avoid.

**Plot:**



**Observations:** pedestrian injuries are correlated with pedestrian deaths.

**Action:** the only surprising thing about this observation (really) is just how well these variables are correlated.

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** It appears that the likelihood of a tow-away decreases as the severe injury count increases, with a slight inflection near the 0.0 x-axis mark. There appear to be a handful of outliers near the top of the plot, these individuals most likely totaled their vehicles. Maybe it is California policy to leave the road on the side of the vehicle if the passenger prefers?

**Action:** I’m not quite sure what to make of this plot. My first suspicion is that the likelihood of a tow\_away would increase with the severe\_injury\_count. We see from the plot, however, that this is not the case.

**Numerical PCA Summary: ‘df\_collisions\_parties’**