Author: Micah Simmerman

Group: 5

Subject: Numeric PCA Report on ‘collisions’ and ‘parties’ tables

Date: August 11, 2023

**Group 5: Numeric PCA Report**

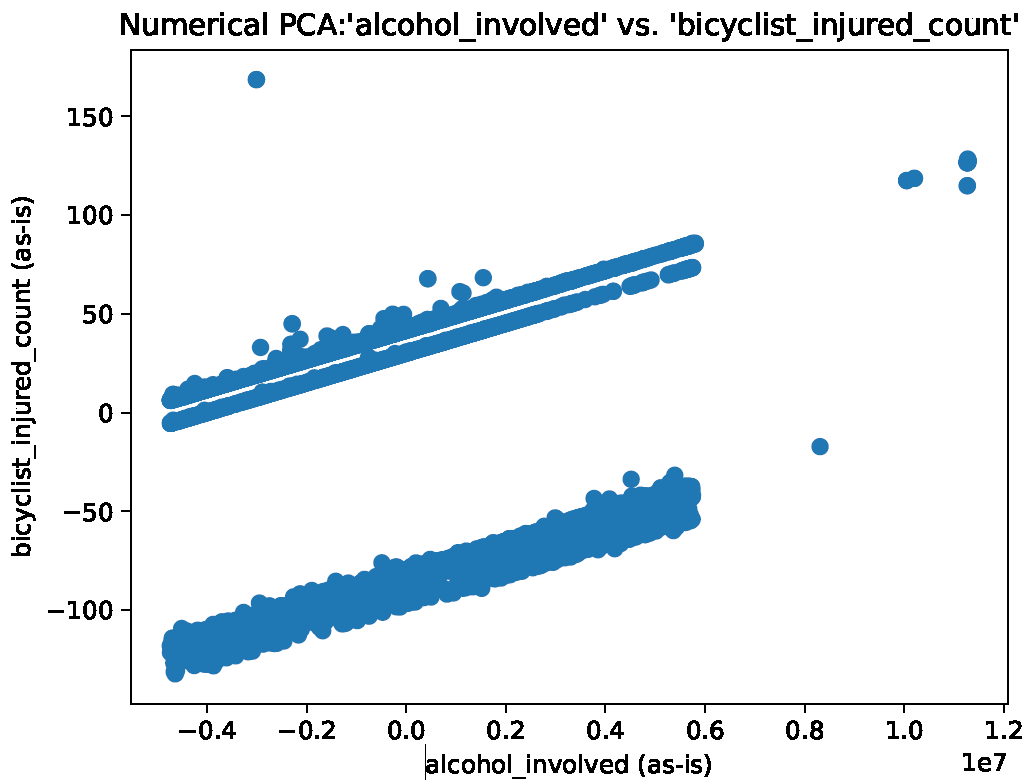
Here, we take a visualization-based data mining approach to PCA by examining numerical attribute columns using a series of cross-plots to observe correlations and interesting clusters throughout the training set data frames collected from the ‘collisions’, ‘parties’, and ‘victims’ tables of the switrs.sqlite 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 under different attribute pairs, we can observe useful correlations and clustering patterns that might otherwise be obscured by noise or hidden within the massive data set. Insights into these interesting trends and distributions provide us with useful guidance that can assist in constructing classifiers capable of distinguishing the SWITRS common driver classes by their varying attribute qualities. The PCA plots themselves can be the source of rich information that can provide the viewer with a enhanced 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 trend is not surprising, but it is exceptionally well correlated for a data set of this size. This plot can easily be fitted to a linear regression function that can be used to predict numeric outcomes within the SWITRS data set.

**Plot:**

**A blue dots on a white background

Description automatically generated**

**Observations:** A very high-quality separation is observed, based on whether the accident involved a bicycle-related injury or not. From the look of this figure, bicyclists appear to be injured a lot!

**Action:** This set of attributes does a great job of splitting the data set into roughly equal sets. As a result, ‘bicylist\_injured’ is an attribute that is likely to appear in many of our classification models!

**Plot:**

A graph of blue dots

Description automatically generated

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

**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. This means that each of these attributes is likely a strong candidate for becoming a decision in our decision tree classification models.

**Plot:**

A blue dot diagram with white background

Description automatically generated

**Observations:** killed\_victims also appears to be a reasonable class identifier in its own right, although some null response and bias appears to 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 (i.e., with or without a bicyclist fatality).

**Action:** Parents of young drivers sensitive to the fact that novice drivers can be dangerous. AI systems that tutor novice drivers are likely to become a lucrative sub-topic in the future 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 within the ‘not\_private\_property’ attribute column. (i.e., those that hover about the 0.0 y-axis marker appear to be no-responses). Generally, the ‘not\_private\_property’ attribute appears to be used to apply an additional penalty for roadway infractions occurring on public property. As such, the usefulness of this attribute may be limited.

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

**Action:** cell phone intervention 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 enjoy being separated from them. This presents a conundrum in terms of designing AI features to reduce the use of cell phones while driving. Keeping the driver distracted and/or engaged with some form of forward-facing camera system (often fitted with some form of object recognition) is considered by many researchers to be the best solution for improving user engagement with AI collision detection systems. (see the Prior\_Works folder for more information.) This opinion appears driven by the assumption that presenting evidence of dangerous driving conditions to the user may help to condition his/her future response.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** people appear to use the telephone most while driving alone.

**Action:** Observe that the relative support of single driver class may outweigh that of non-solo drivers. As a result, we should be careful when observing rules concerning party\_count, as the relative support for solo-drivers may “tilt the scales”, causing us to adopt a conclusion that is not as strong as we were once lead to believe. We can determine the weighted effect of any rules we uncover under this attribute by incorporating knowledge about the relative distribution of competing classes. In other words, the strength of each derived rule will be weighted by the set membership of its’ respective class.

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** somewhat surprisingly, ‘cellphone\_use’ does not appear to be correlated with ‘pedestrian\_collisions’.

**Action:** Myth busted!

**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:** latitude and longitude are ideal attribute candidates for composing regional data cubes.

**Plot:**

A diagram of blue dots

Description automatically generated with medium confidence

**Observations:** Desirable clustering.

**Action:** The attributes examined in this cross-plot make ideal decision candidates for a DTC classifier.

**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:** Another potentially interesting data cube concept that may explain regional trends in motocycle collisions. Perhaps there is a hot bed of activity waiting to be uncovered.

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** the majority of motorcycle accidents appear to involve younger drivers.

**Action:** Teenagers and parents of young drivers appear to be the primary target class(es) under examination here.

**Plot:**

A graph of blue dots

Description automatically generated

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

**Action:** This graphic is important because it shows that the probability of colliding with a bicycle is somewhat higher in domestic settings than in urban environments. An AI system can be trained to detect environmental cues to ascertain the relative likelihood of encountering different road obstacles under an active setting, which should extend to bicycles as well.

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** unexpected distribution

**Action:** younger drivers tend to be involved in collisions near private residences. As ‘party\_age’ increases so does the likelihood that the collision took place in an urban environment.

**Plot:**

A diagram of a motorcycle accident

Description automatically generated with medium confidence

**Observations:** Bimodal distribution observing the probable injury count of a collision involving a motorcycle compared to other, non-motorcycle collisions. We can see that the relative density of points decreases as we move to the right, indicating an estimate of the relative severity of these accidents when compared side-by-side. The ‘other\_visible\_injury\_count’ attribute variable appears to reach a maximum value of 1.0, whereas non-motorcycle collisions appear to extend much further.

**Action:** The root cause of this pattern could be explained by frequent use of rider safety equipment or an increased rate of motorcyclist fatalities compared to other driver classes.

**Plot:**

A blue and red dot pattern

Description automatically generated with medium confidence

**Observations:** Dead motorcyclists tell no lies, but the coroner’s office might.

**Action:** Interested in determining the angle of impact in some of the deadliest SWITR motorcycle collisions? Examining the sparse clusters to the right of the highlighted cluster may lead to interesting reports, including public hospital records and coroners’ certificates.

**Plot:**

A graph with blue and red dots

Description automatically generated

**Observations:** Older victims tend to suffer a greater number of injuries in an automobile accident.

**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) AI safety systems.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** vehicles with additional passengers do not appear to add any additional safety risk to the passengers themselves, as evidenced by ‘other\_visible\_injury\_count’.

**Action:** vehicles with additional passengers are likely to be a contributing factor to many collisions due to the increased number of distractions experienced by the driver under these conditions. The visible injury count of these passengers. However, visible\_injury\_count does not appear to be correlated with the number of passengers inside the vehicle.

**Plot:**

A diagram of a number of dots

Description automatically generated

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

**Action:** If the AI system is aware of the driver’s approximate age, enhanced safety features may be applied to assist in correcting deficits in novice driving patterns. Such features would surely come in handy when teaching a new driver how to decelerate at an intersection, for example.

**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 would likely test well with audiences of concerned parents of young adults.

**Plot:**

A diagram of a party age

Description automatically generated with medium confidence

**Observations:** odds of every type of collision decrease with respect to party\_age.

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

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** A distribution demonstrating the relative likelihood of dying in a motorcycle collision.

**Action:** This is an interesting statistic that can be examined under a Bayesian Belief Network (BBN).

**Plot:**

A graph with blue dots and red circle

Description automatically generated

**Observations:** This cluster appears to present a set of data points describing reckless teenage drivers.

**Action:** AI systems can monitor environmental conditions to determine the probability of being involved in an accident under the circumstances using an intuition pattern that is similar to a BNN. Hardline intervention measures are usually avoided at all costs for the sake of customer convenience. In a life-or-death situation however, the best policy may be for the AI system to intervene to prevent potential loss of life and “ask for forgiveness” later. Reckless teenage drivers are a major safety concern on US roadways.

**Plot:**

A diagram of a person's body

Description automatically generated

**Observations:** pedestrians are more likely to be involved in a collision when a minor party is driving.

**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:** observe the figure.

**Plot:**

A graph of blue dots

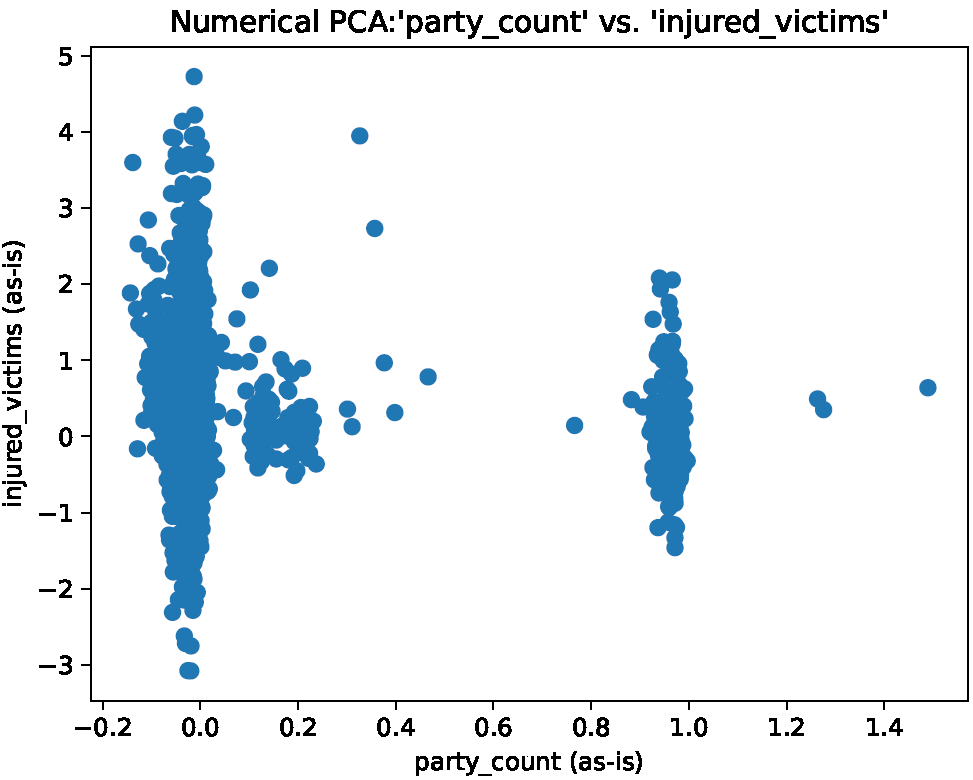
Description automatically generated

**Observations:** people are more likely to use a cell phone when they are alone than they are 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. Alternatively, the driver can be presented with “good distractions”, i.e., those that enhance the safety of the passengers. For example, if the vehicle detects that the driver is performing a google search, it may prompt “I see that you are looking for something on the web, can I help you with that search”, or “your wife has sent you a message. Would you like me to read it to you?”, or even “I’ve noticed that your phone is creating a distraction. Is there anything I can assist you with?”

Any of these intervention techniques is likely to succeed in refocusing the driver. It’s best not to be *too* annoying with these features, however.

**Plot:**



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

**Action:** The clustering presented in this figure might indicate a need for AI-features that are tailored to individual drivers. In many ways, an ideal AI system would act as this second passenger and provide many of the same features as a “backseat driver” would (albeit in a much less intrusive manner). Highlighted forward-facing dash cameras are one of the most promising technological developments that could be tailored to suit the purpose of an AI-copilot.

**Plot:**

A graph of blue dots

Description automatically generated

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

**Action:** A “silent ride-along parter” idea could be just the right safety feature for more *private* consumers. Investigate these trends but be wary of the support bias of the ‘party\_count’ attribute that we mentioned earlier.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** Again, single drivers appear more likely to be involved in an accident than a driver with a co-pilot, whether the collision involves fatalities or not.

**Action:** Examine this distribution to derive a probabilistic rule regarding this observation. E.g., “single drivers are \_\_% more likely to be involved in a motor vehicle accident than those with a co-pilot”. The strength of this weighted rule should provide an early indication of how effective an AI-copilot system may be expected to be in avoiding collisions.

**Plot:**

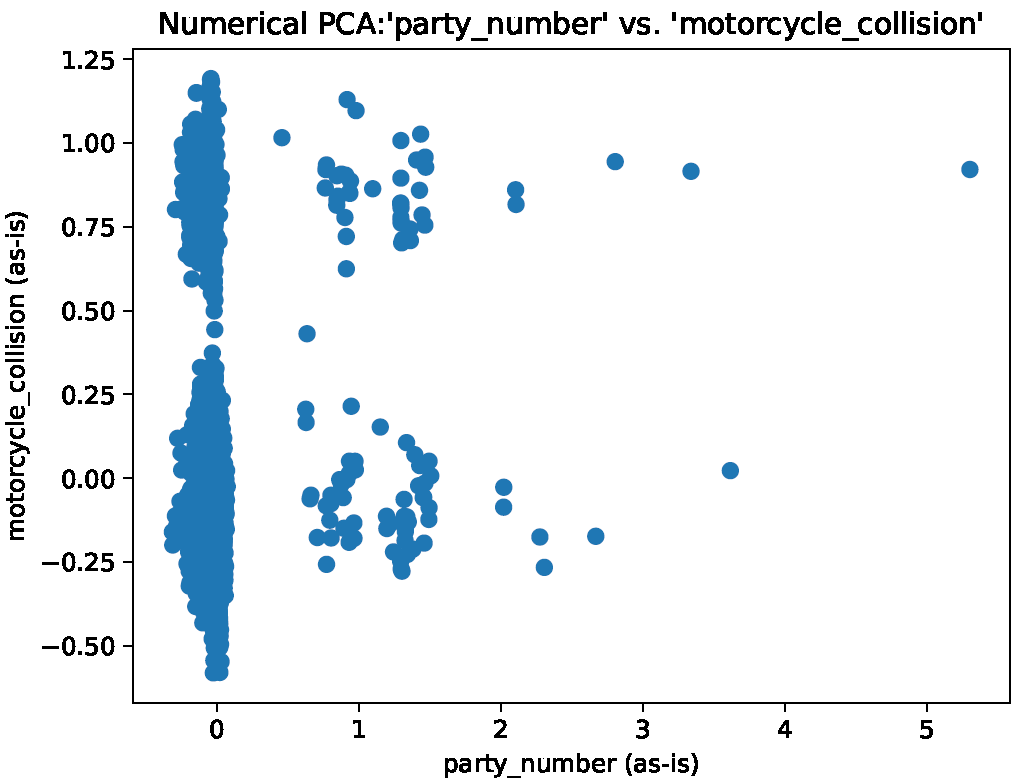
A graph of a number of bicyclists

Description automatically generated

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

**Action:** This is an interesting finding that adds further evidence to our AI-copilot hypothesis.

**Plot:**



**Observations:** solo-drivers would appear to be more likely to hit motorcycles.

**Action:** similar trends and patterns appear to be observed across all collision types.

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** A negative correlation is observed between ‘party\_number’ and ‘pedestrian\_collision’. This time the overwhelming support for single-party drivers does not appear to cloud the model (as before).

**Action:** Determine if this trend is meaningful by observing the integrity of the underlying data points. It appears that the plot is telling us that “the likelihood of a pedestrian collision decreases with respect to party\_number”. Although this trend agrees with our AI-copilot theory, we should investigate these data points further to uncover the source of this trend.

**Plot:**

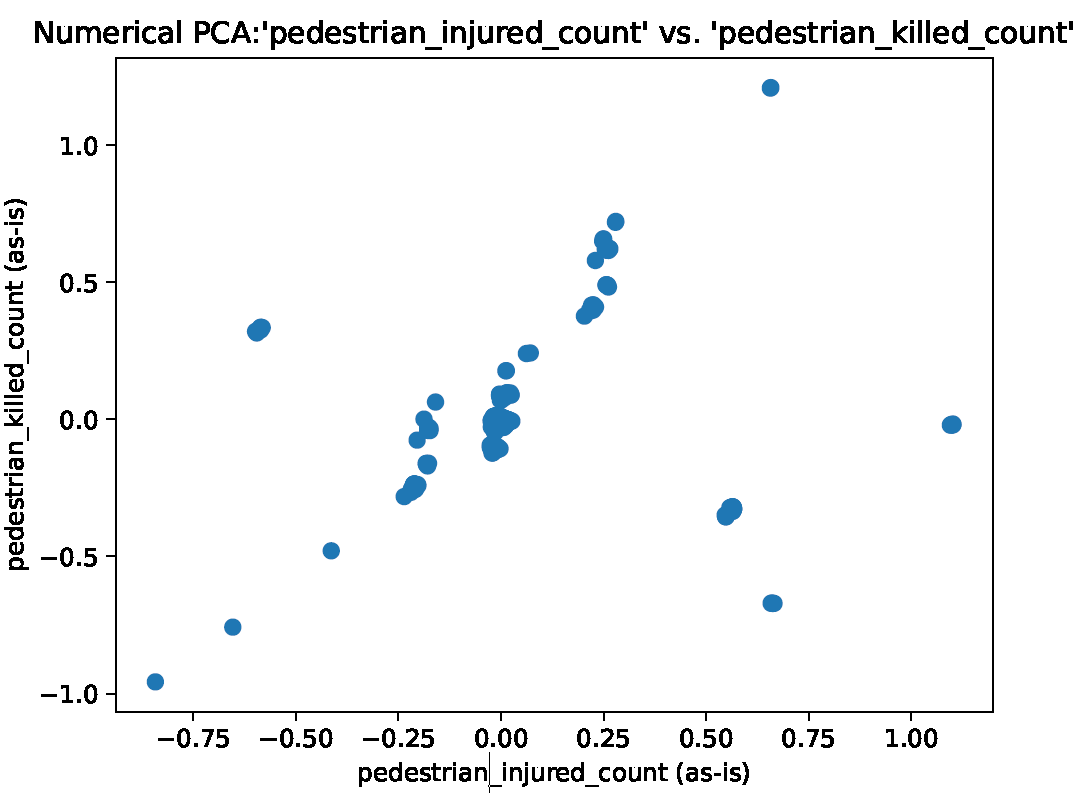
A graph with blue dots

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. (Hopefully the kids were alright!)

**Action:** Life gets 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 strongly correlated with pedestrian deaths.

**Action:** the only surprising thing about this observation is just how well correlated these variables are. Correlations like these can be difficult to locate in large datasets.

**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 goes up, with a slight inflection near the 0.0 x-axis mark. There appear to be a handful of outliers near the top of this plot, I’m assuming that these individuals most likely totaled their vehicles. This trend seems surprising, as one might expect damage to the vehicle and passenger to be positively correlated. Perhaps this finding is the result of a California state traffic policy?

**Action:** It is difficult to know what to make of this plot without examining these data points further. While the relevance of this finding is not yet clear, we will save it here for future reference.

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