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

Principal Component Analysis Wikipedia Definition: “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.”

Group 5 has taken a visualization data mining approach to PCA by plotting numerical attribute columns in a series of cross-plots to observe any indication of correlations between the attributes of the ‘collisions’, ‘parties’, and ‘victims’ tables of the SWITRS SQLite dataset.

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

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

BEGIN: AS-IS NUMERIC PCA CROSS-PLOTS

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** Intersection and distance are extremely correlated for a data set with 10,000 points.

**Action:** Monitor ‘intersection’ and ‘distance’ variables of the ‘collisions’ table as potentially redundant attributes.

**Plot:**

A graph of a number of blue dots

Description automatically generated

**Observations:** The number of injured victims appears to increase linearly with respect to distance (across 3 parallel patterns)

**Action:** Observe the clustering of this data set closely using min-max normalization.

**Plot:**

A graph of a distance between two different states

Description automatically generated with medium confidence

**Observations:** linear trends observed between ‘pedestrian\_collision’ and ‘distance’

**Action:** some bicycle collisions appear to take place at a long stopping distance. This suggests that the car was likely moving fast just before the collision took place.

**Plot:**

**A graph of a graph showing a number of points

Description automatically generated with medium confidence**

**Observations:** positive correlation observed between ‘distance’ and ‘bicycle\_collision’

**Action:** The density of this plot seems to suggest that most collisions with bicycles take place at a short distance (i.e., when the vehicle is probably moving slow). The plot appears to suggest that bicycle collisions can occur over a wide range of collision distances, with the bulk of this data occurring in the [0-20,000] range.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** ‘tow\_away’s appear to take place at predictable distances from nearby intersections. Frequently reported values are also observed in this graph, which may represent heuristic decisions made on the part of the responding officer.

**Action:** See if we can make use of this fact under the context of automated vehicle safety suggestions.

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** Most of the collisions resulting in ‘killed\_victims’ occurred some distance behind an intersection.

**Action:** It is interesting to note that these vehicles were likely moving at relatively high speeds at this distance from an intersection. It may also be deduced that drivers tend to take much greater care at intersections than they do on other parts of the road.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** Note the interesting increase in ‘injured\_victims’ as the distance from the intersection increases.

**Action:** This pattern is likely described by the fact that people’s attention tends to wane as they progress further from an intersection.

**Plot:**

A graph of blue dots

Description automatically generated

**Observations:** No significant correlation observed.

**Action:** it is interesting to observe how few ‘pedestrian\_collisions’ occurred at an intersection. The information contained in this graph appears to suggest that most pedestrian\_collisions involve some form of jay-walking. Pedestrian detection systems were some of the earliest technologies adopted by AI-vehicles.

**Plot:**

A graph of a graph

Description automatically generated with medium confidence

**Observations:** Note the steep drop-off in ‘bicycle\_collisions’ just AFTER the intersection, followed by a steady increase in collisions thereafter.

**Action(s):**

* This data appears to suggest that drivers are more likely to notice a bicyclist **when they are not at an intersection**.
* Conversely, many bicyclists may prefer to execute a safe crossing of their own, one that does **not** involve crossing at an intersection.
* A significant percentage of ‘bicycle\_collisions’ appear to take place approximately -100 feet from an intersection.

**Plot:**

A diagram of a motorcycle collision

Description automatically generated

**Observations:** The distribution is more exciting than the correlation here

**Action:** Observe the proximity of different motorcycle collisions with respect to intersection proximity. Motorcycles are another group of objects that AI systems can detect and avoid.

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** ‘pedestrian\_collision’ appears positively correlated with ‘party\_count’.

**Action:** As the ‘party\_count’ increases, so do the likely number of distractions experienced by the driver of the vehicle involved in a collision. AI-generated ‘attention-assistance’ features could help to reduce the number of distractions experienced by drivers, thereby reducing the number of associated collisions.

**Plot:**

A graph of blue dots

Description automatically generated with medium confidence

**Observations:** seemingly positive correlation(s) between ‘motorcycle\_collision’ and ‘injured\_victims’

**Action:** unsurprisingly, many victims of collisions involving motorcycles are injured.

**Plot:**

A chart of a truck collision

Description automatically generated

**Observations:** positive correlations observed between ‘truck\_collision’ and ‘motorcycle\_collision’

**Action:** This does not tell us much, other than the observations of collisions in different categories tend to be correlated.

**Plot:**

A diagram of a truck collision

Description automatically generated

**Observations:** note the under-supported “DUI correlation” involving truck collisions.

**Action:** This correlation would be a lot more interesting if the support from the other vehicle types was included. **IDEA:** use the principal component ‘alcohol\_involved’ with support from all vehicle types and plot the collision numbers with respect to time or space (i.e., a data cube ‘collisions\_all\_types’ x ‘alcohol\_involved’, over time or city).

**Plot:**

A graph of a collision

Description automatically generated

**Observations:** positive correlations between ‘truck\_collision’ and ‘severe\_injury\_count’

**Action:** Accidents involving trucks cause a disproportionate amount of damage compared to other vehicle types. Avoiding collisions with large objects should weigh disproportionately compared to collisions with lighter objects. AI systems may be able to ascertain the expected weight of an object based on material classification (i.e., density) and volume (i.e., size) of the vehicle and/or object when a

Collision is imminent.

**Plot:**

A graph of a collision

Description automatically generated with medium confidence

**Observations:** one or more observable positive correlations between ‘truck\_collisions’ and ‘other\_visible\_injury\_count’

**Action:** the same principal applies here as in the example above.

**Plot:**

A graph showing a number of blue dots

Description automatically generated

**Observations:** collisions that occur on public property are much more likely to be towed away.

**Action:** AI systems can call for a tow-truck before the police can search or impound it (preservation of consumer’s legal rights). A series of other helpful actions can be initiated by AI systems following a collision, including calls for urgent medical assistance, confirmation of passenger consciousness by eliciting a verbal response (“I’ve detected a collision, do you need me to call an ambulance?”), triggering hazard lights, extending road markers, and more.

**Plot:**

A diagram of a graph

Description automatically generated with medium confidence

**Observations:** perceived positive correlation between ‘not\_private\_property’ and ‘party\_count’

**Action:** this graphic appears to suggest that carpooling is increasingly frequent in public areas. This highlights an opportunity for AI systems that can observe and interact with multiple passengers. (Open to suggestions)

**Plot:**

A graph with blue dots

Description automatically generated

**Observations:** positive correlation between ‘not\_private\_property’ and ‘pedestrian\_collision’.

**Action:** Unsurprisingly, pedestrian collisions occur more frequently on public property.

**Plot:**

A diagram of a number of blue dots

Description automatically generated

**Observations:** negative correlation(s) observed between ‘not\_private\_property’ and ‘alcohol\_involved’

**Action:** This graph appears to suggest that the percentage of alcohol-related collisions is higher near areas constituting “private property” than it is near public property (i.e., “not\_private\_property”).

**Plot:**

A chart of blue dots

Description automatically generated

**Observations:** extending the observation above

**Action:** accidents on private property tend to involve more injuries as well. The trend appears to suggest that drivers tend to drive more cautiously in public areas than they do in rural areas. (They also appear to consume more alcohol near private property.)

**Plot:**

A graph of alcohol and injury

Description automatically generated

**Observations:** positive correlation between ‘alcohol\_involved’ and ‘severe\_injury\_count’

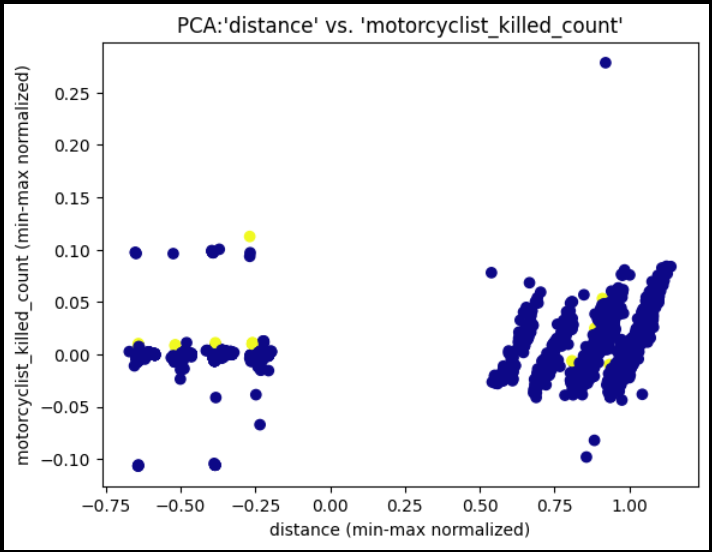
**Action:** This trend is not surprising, given what we know about alcohol and motor vehicles.

BEGIN: MIN-MAX NORMALIZED NUMERIC PCA CROSS-PLOTS

Min-max PCA cross-plots are helpful for identifying behavioral clusters within data. The semantic units range from [-1.0, 1.0] with -1.0 being the lowest possible response and +1.0 being the highest. This method of standardization tends to evoke a higher level of separation between data points than the ‘as-s’ plots observed above. Because of this “resolving power”, min-max numeric PCA can be used to label and identify underlying classes in the training set data distribution(s). Let’s see what we can observe…

Keep in mind that we are not only observing correlations, but also the relative distribution of meaningful clusters of data presented in the training set.

**Plot:**



**Observations:** a series of positive correlations between ‘distance’ and ‘motorcyclist\_killed\_count’, presented at varying ‘distance’ values.

**Action:** The fact that motorcycle collisions are positively correlated with distance suggests that most fatal motorcycle accidents occur on a highway, interstate, or similar road structure that would afford vehicles to maintain large distances from one another. Fortunately, an AI collision-avoidance system should be able to monitor motorcycles with relative ease.

**Plot:**

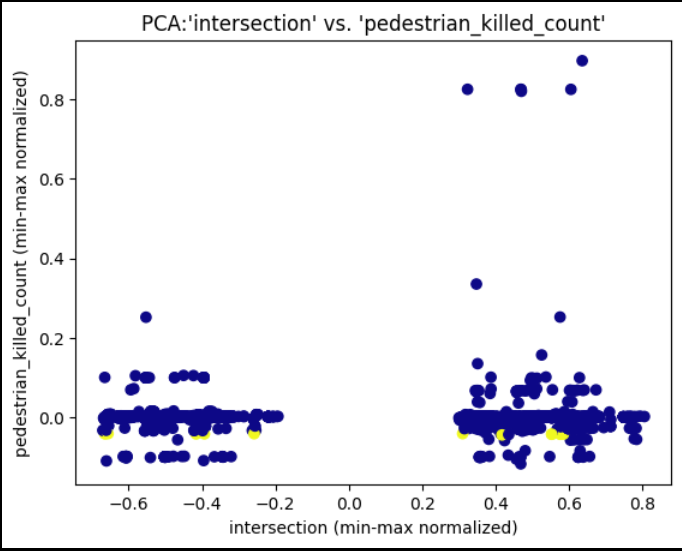
A diagram of a graph with Ice hockey rink in the background

Description automatically generated with medium confidence

**Observations:** four distinct data clusters represented by A, B, C, D.

**Action:** A: ‘injured\_victims’ behind an intersection. B: ‘injured\_victims’ in front of an intersection. C and D: no injured victims.

**Plot:**



**Observations:** pedestrians are more likely to be killed after an intersection based on the relative support shown here.

**Action:**

**Plot:**

A graph with blue and yellow dots

Description automatically generated

**Observations:** bicyclists are also more likely to be killed when struck by a vehicle that has recently crossed an intersection!

**Action:** AI-vehicles should adjust for human weakness; we already know that drivers tend to pay less attention to the road as their distance from a nearby intersection increases. The figure above clearly demonstrates that this is exactly when collisions with pedestrians and bicyclists are most deadly.

If we can add features to the vehicle to alert drivers to dangerous situations such as the one above, then we can reduce collisions under similar circumstances and save lives.

**Plot:**

TO BE CONTINUED…

**Observations:** asdf

**Action:** asdf

**Plot:**

**Observations:** asdf

**Action:** asdf