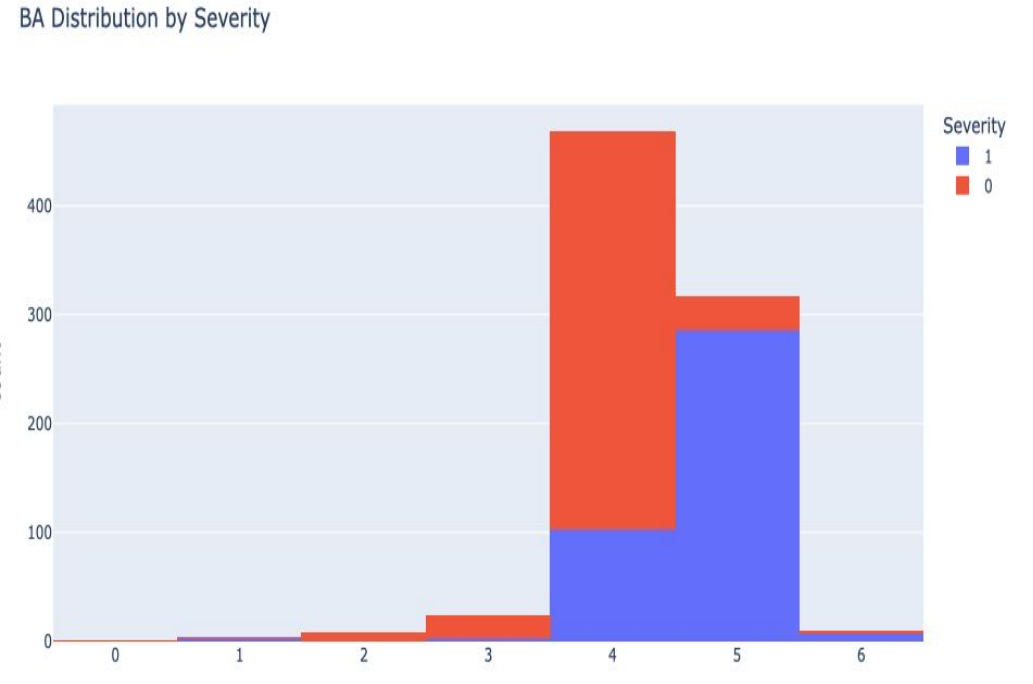
INFOB2DA | Practical Assignment 1

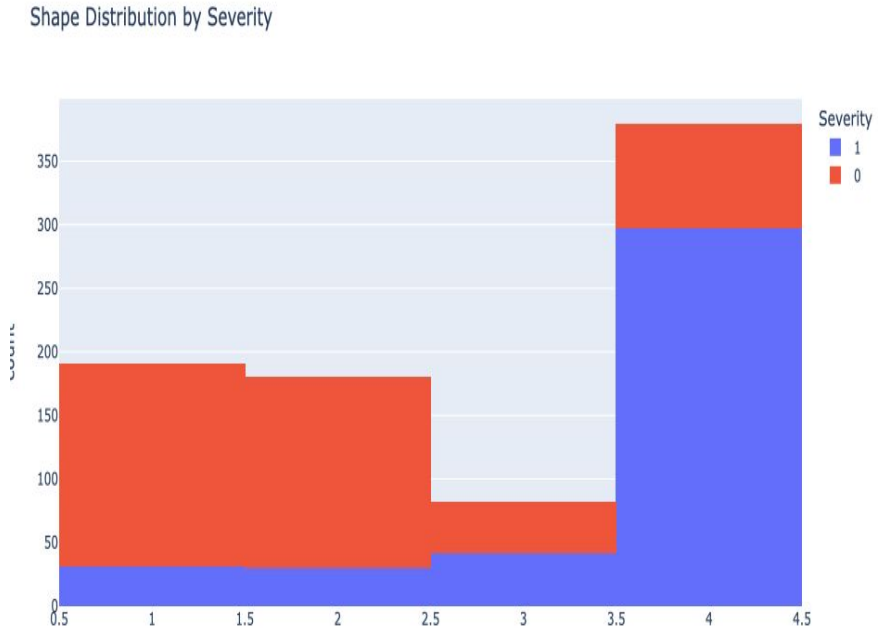
Data understanding and preprocessing

Jack, Felicia and Joost

Overview Dataset

	BA	Age	Shape	Margin	Density	Severity
count	959.000000	956.000000	930.000000	913.000000	885.000000	961.000000
mean	4.300313	55.487448	2.721505	2.796276	2.910734	0.463059
std	0.683469	14.480131	1.242792	1.566546	0.380444	0.498893
min	0.000000	18.000000	1.000000	1.000000	1.000000	0.000000
25%	4.000000	45.000000	2.000000	1.000000	3.000000	0.000000
50%	4.000000	57.000000	3.000000	3.000000	3.000000	0.000000
75%	5.000000	66.000000	4.000000	4.000000	3.000000	1.000000
max	6.000000	96.000000	4.000000	5.000000	4.000000	1.000000





Why

Visualization

- Presentation
- Confirmatory Analysis
- Exploratory Analysis

Reasons for the

Data cleaning procedure

- Ignore the tuple
- + easy
- - loss of information

Data Preprocessing

Why?

- High quality and reliable data
- Enhancing model interpretability
- Smooth out noise

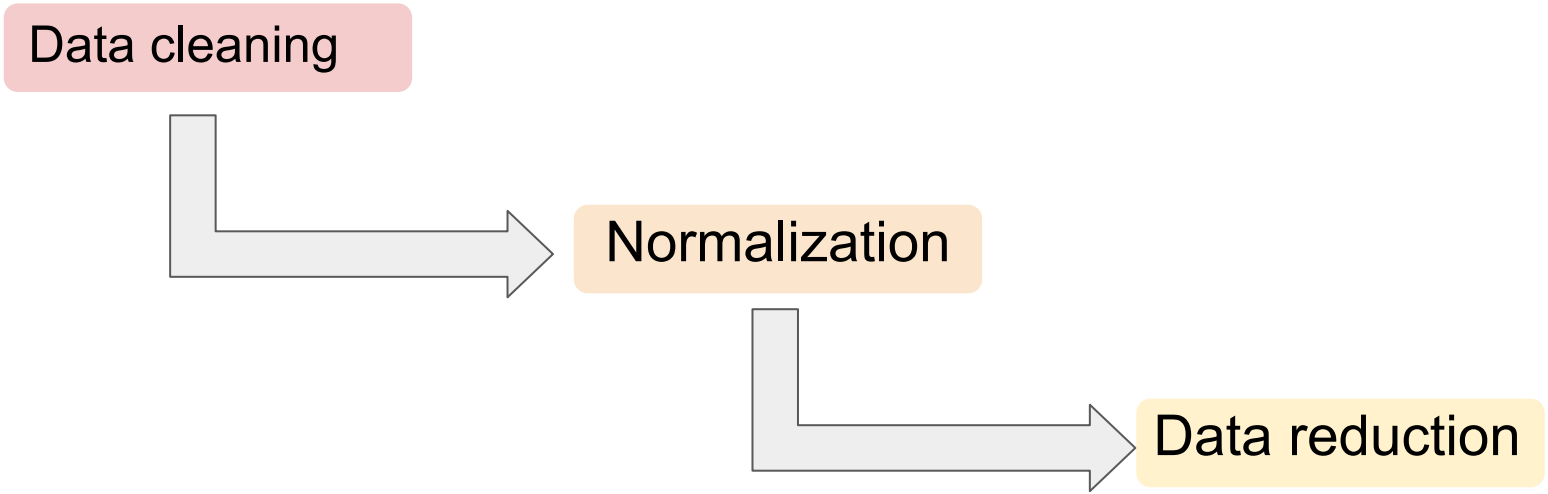
.

.

.

Data Preprocessing

We took three steps to do data preprocessing



Data Cleaning

	BA	Age	Shape	Margin	Density	Severity
count	959.000000	956.000000	930.000000	913.000000	885.000000	961.000000
mean	4.300313	55.487448	2.721505	2.796276	2.910734	0.463059
std	0.683469	14.480131	1.242792	1.566546	0.380444	0.498893
min	0.000000	18.000000	1.000000	1.000000	1.000000	0.000000
25%	4.000000	45.000000	2.000000	1.000000	3.000000	0.000000
50%	4.000000	57.000000	3.000000	3.000000	3.000000	0.000000
75%	5.000000	66.000000	4.000000	4.000000	3.000000	1.000000
max	6.000000	96.000000	4.000000	5.000000	4.000000	1.000000

Same size, more adequate mean and standard

	BA	Age	Shape	Margin	Density	Severity
count	830.000000	830.000000	830.000000	830.000000	830.000000	830.000000
mean	4.338554	55.781928	2.781928	2.813253	2.915663	0.485542
std	0.660689	14.671782	1.242361	1.567175	0.350936	0.500092
min	0.000000	18.000000	1.000000	1.000000	1.000000	0.000000
25%	4.000000	46.000000	2.000000	1.000000	3.000000	0.000000
50%	4.000000	57.000000	3.000000	3.000000	3.000000	0.000000
75%	5.000000	66.000000	4.000000	4.000000	3.000000	1.000000
max	6.000000	96.000000	4.000000	5.000000	4.000000	1.000000

Normalization

We do normalize for…

- Scaling features
- Mitigating numerical instabilities
- Easier to visualize

Normalization method we applied

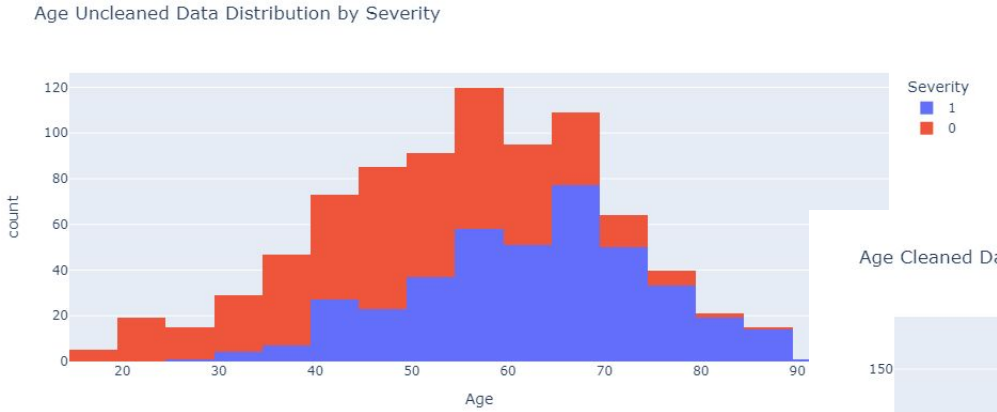
Min-max normalization

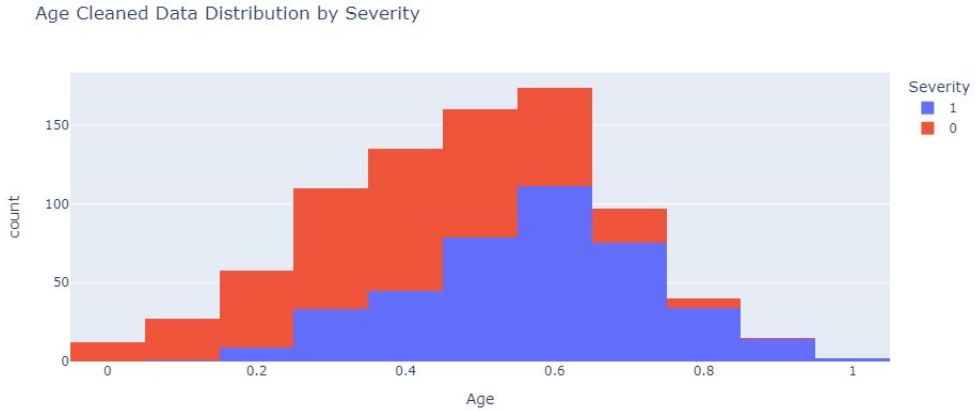
$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

```
def normalize_column(column):
    min_val = column.min()
    max_val = column.max()
    return (column - min_val) / (max_val - min_val)
```

Normalization method we applied

Min-max normalization on the variable-age



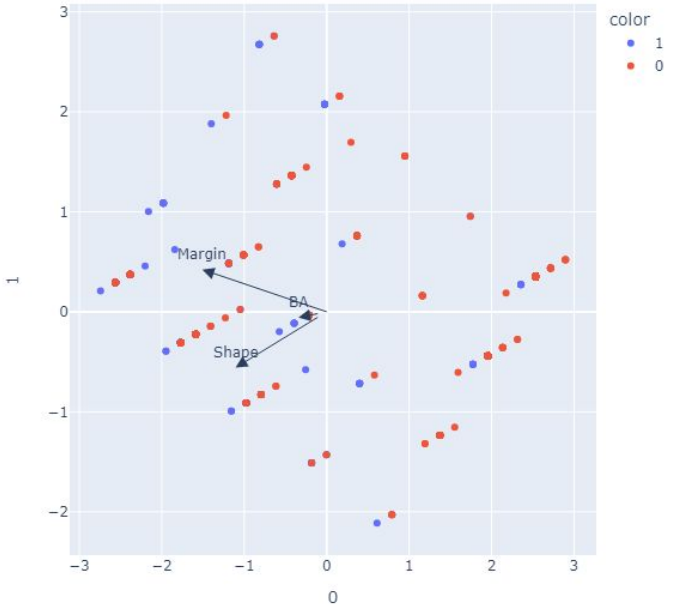


Feature Reduction

- ANOVA F-value, similar to statistical significance
- Comparies variation between the group means to the variation within each group
- Determine random chance or genuine correlation
- Reduce features to reduce processing overhead
- Best features, not necessarily good features

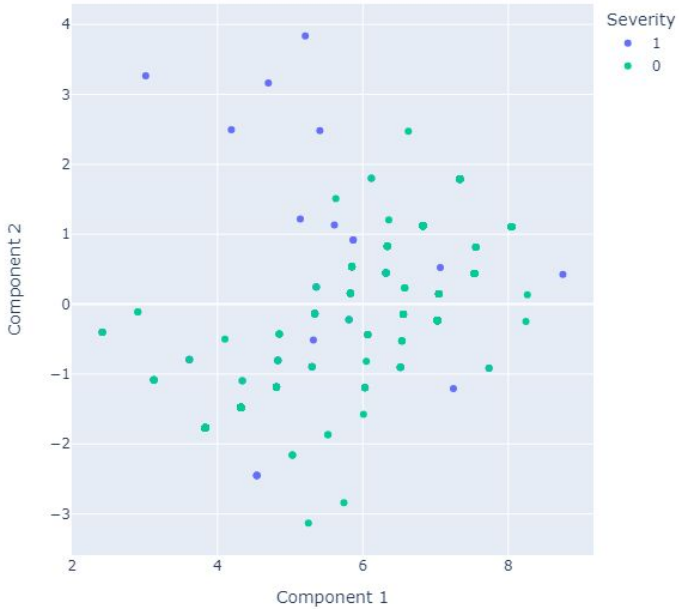
PCA & tSVD

- Dimensionality reduction techniques
- PCA: Reduces the number of features in data while preserving important information.
- PCA: It finds new axes (principal components) where the data varies the most.
- PCA: Projects data onto these new axes.



PCA & tSVD

- tSVD: Decomposes data into a few important components.
- tSVD: Factorizes the data matrix directly into components.
- tSVD: Keeps only the most significant ones.
- PCA finds orthogonal components based on data's covariance.
- TSVD directly decomposes data matrix, not necessarily orthogonal.
- Choose PCA for variance preservation and orthogonality.
- Choose TSVD for direct decomposition and simplicity.



Conclusion

- Margin & Shape most correlated features
- Usefulness of extra features drops off immensely afterwards
- No immediately noticeable clustering of the Severity feature ->
- Prediction of the severity is very unlikely going to be very deterministic
- We can however the trend between these three features