Sabreena Abedin

Question 1:

I chose train accident damage (ACCDMG) and Casualty (TOTINJ + TOTKLD).

Accident damage provides the total recordable damage, which appeared to be more than simply the sum of equipment damage, car damage, and track damage. This was the best summary metric relating to monetary damage.

Casualty provides information on human life. Since deaths are rare, adding injuries allows for a more nuanced approach while not ignoring the death metric. In the future, this might be scaled, for example, Casualty might be recalculated as TOTINJ + 10*TOTKLD.

Looking at the biplot of these two metrics, we see that the metrics are almost perpendicular, meaning they won't be providing redundant data.
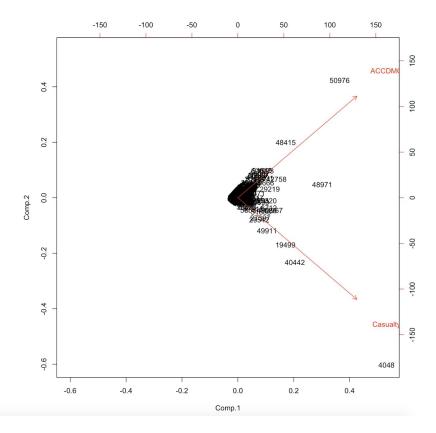
```
Importance of components:
                          Comp.1    Comp.2
Standard deviation     1.0717233 0.9227184
Proportion of Variance 0.5742954 0.4257046
Cumulative Proportion  0.5742954 1.0000000
> totacts.pcaMetrics$loadings

Loadings:
         Comp.1 Comp.2
Casualty  0.707 -0.707
ACCDMG    0.707  0.707

               Comp.1 Comp.2
SS loadings       1.0    1.0
Proportion Var    0.5    0.5
Cumulative Var    0.5    1.0
```

Question 2:

I chose train speed (TRNSPD), high speed (HIGHSPD), alcohol, drug, and number of engineers (ENGRS) as quantitative predictors to look into. Looking at the scatterplots, we can see that train speed and high speed are strongly correlated, meaning they will provide redundant information. This is supported by the PCA biplot, showing the two factors almost on top of each other. They account for the same amount of variance, meaning we should only keep one. Number of engineers isn't highly correlated to either of our primary metrics, and this is supported by the fact that has minimal loadings in most of the principal components and doesn't make up a large amount of the variance until the 4th component.

Train speed should be kept as a quantitative metric, because train speed and high speed are highly correlated but train speed has higher loadings. Number of engineers doesn't provide a lot of information, however, alcohol and drugs both provide significantly to the variance, as shown by the loadings in principal component #2.

```
> summary(totacts.pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6     Comp.7
Standard deviation     1.3915063 1.0630536 1.0293227 1.0026811 0.9506354 0.9160615 0.35479043
Proportion of Variance 0.2766128 0.1614404 0.1513579 0.1436242 0.1291011 0.1198812 0.01798232
Cumulative Proportion  0.2766128 0.4380532 0.5894111 0.7330353 0.8621364 0.9820177 1.00000000
> totacts.pca$loadings

Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
TRNSPD  -0.679 -0.136 -0.134                       0.704
Casualty -0.137  0.491  0.514 -0.247  0.161  0.623
ACCDMG  -0.246  0.432  0.426               -0.752
HIGHSPD -0.666 -0.122         0.236               -0.690
ALCOHOL         0.487 -0.494  0.220  0.683
DRUG            0.547 -0.434        -0.705  0.122
ENGRS   -0.132        -0.319 -0.911        -0.112 -0.167

               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000
Proportion Var  0.143  0.143  0.143  0.143  0.143  0.143  0.143
Cumulative Var  0.143  0.286  0.429  0.571  0.714  0.857  1.000
```