TAIC Legacy report example report

James Thompson

2023-08-21

Abstract

This is an example report for the TAIC Legacy project. It takes the summary.csv file and does normal EDA on it. This should not be considered an extensive analysis of the data created by the engine. Simply a demonstration of things you might do with it.

Data

The data we have received from the engine looks somthing like this:

ReportID	Fatiuge	Substance use	Inexperience	Crew Resource Management	Maintenance	Safety Management Systems	Interfaces between modes	Navigation in pilotage waters
2010_201	16.0	15.0	10.0	23.0	13.0	10.0	7.0	6.0
2010_202	8.0	27.0	14.0	18.0	10.0	5.0	4.0	14.0
2010_203	4.0	5.0	46.0	24.0	15.0	5.0	0.0	1.0
2010_204	9.2	23.4	13.4	19.2	9.2	12.8	4.6	8.2
2010_206	8.0	32.0	22.0	18.0	6.0	0.0	9.0	5.0
2011_201	2.0	4.0	18.0	38.0	22.0	4.0	6.0	6.0

It provides themes weightings for all 40 reports.

Important themes

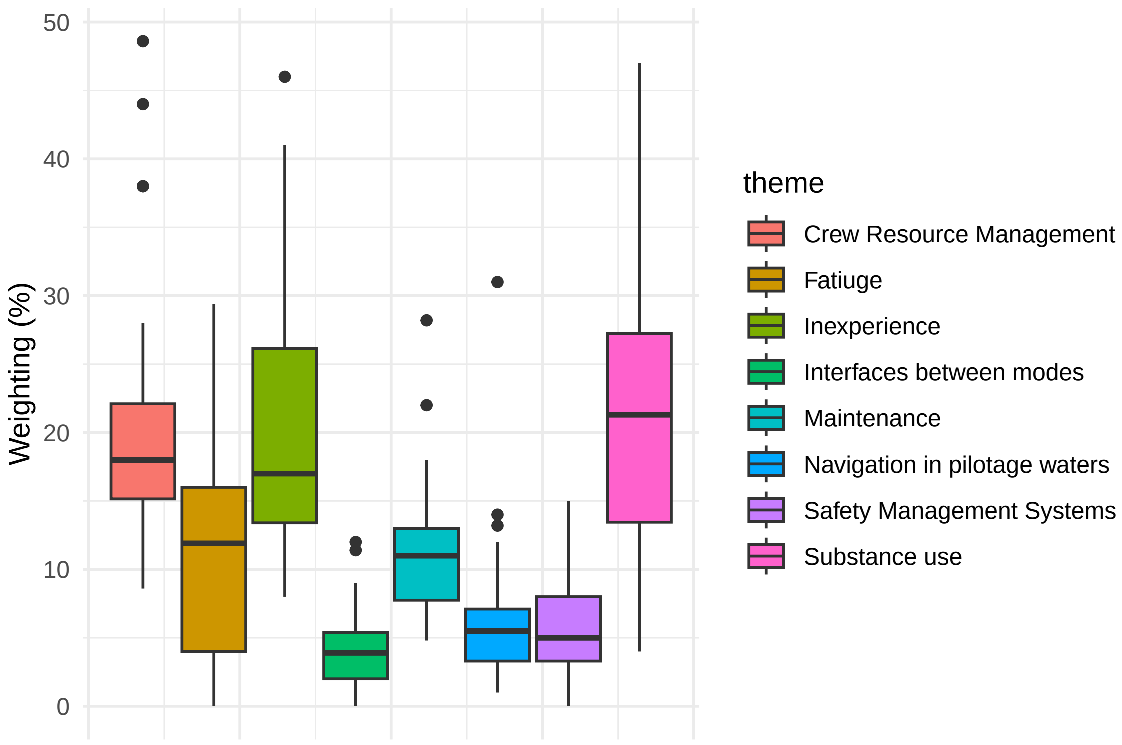
The themes and descriptions provided to the model are:

theme	description
Fatiuge	Tiredness or sleep deprivation affecting performance.
Substance use	Alcohol or Drug use that affects performance.
Inexperience	Missing of essential knowledge and skills.
Crew Resource Management	Crew Resource Management (CRM) is a comprehensive training program and philosophy developed to enhance the teamwork, communication, decision-making, and overall effectiveness of teams, particularly in high-stakes environments.
Maintenance	Mainentance of equipment and systems.
Safety Management Systems	A collection of processes and strucutres that provide effective risk-based decision making.
Interfaces between modes	The interaction between different modes of transportation. I.e level railway crossing, or a ferry terminal.
Navigation in pilotage waters	Pilotage waters are those areas in which a ship is usually required to use the services of a maritime pilot (there are sometimes exemptions). A maritime pilot is an experienced and highly skilled sailor who has detailed knowledge of a particular waterway.

Analysis

Distrubtion

The first thing we want to do is look at the distribution of the weightings for each theme.

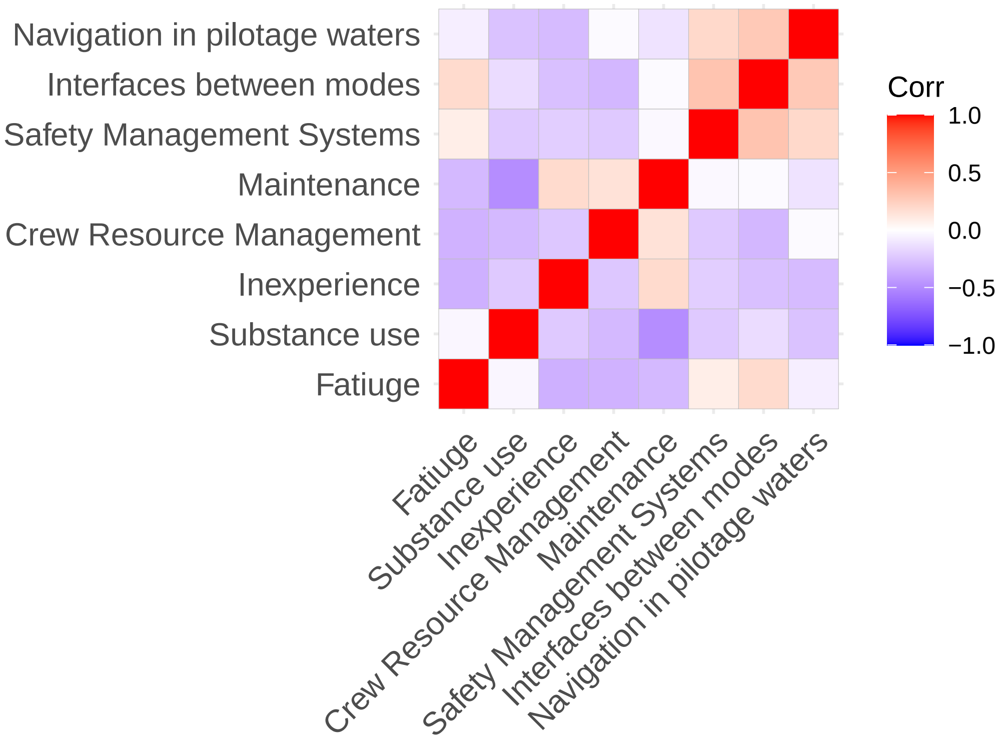


We can see that some of the themes are more common like Substance abuse, while Interfaces between modes are less common. The themes have some large variety in the average and variance of weighting.

It is worth noting that the themes are not well defined at the moment. Small changes in the description could have large consequences in the weighting. For example, the theme "Substance abuse" could be changed to "Substance use" and the weighting would change dramatically.

Correlations

What about correlations between themes?

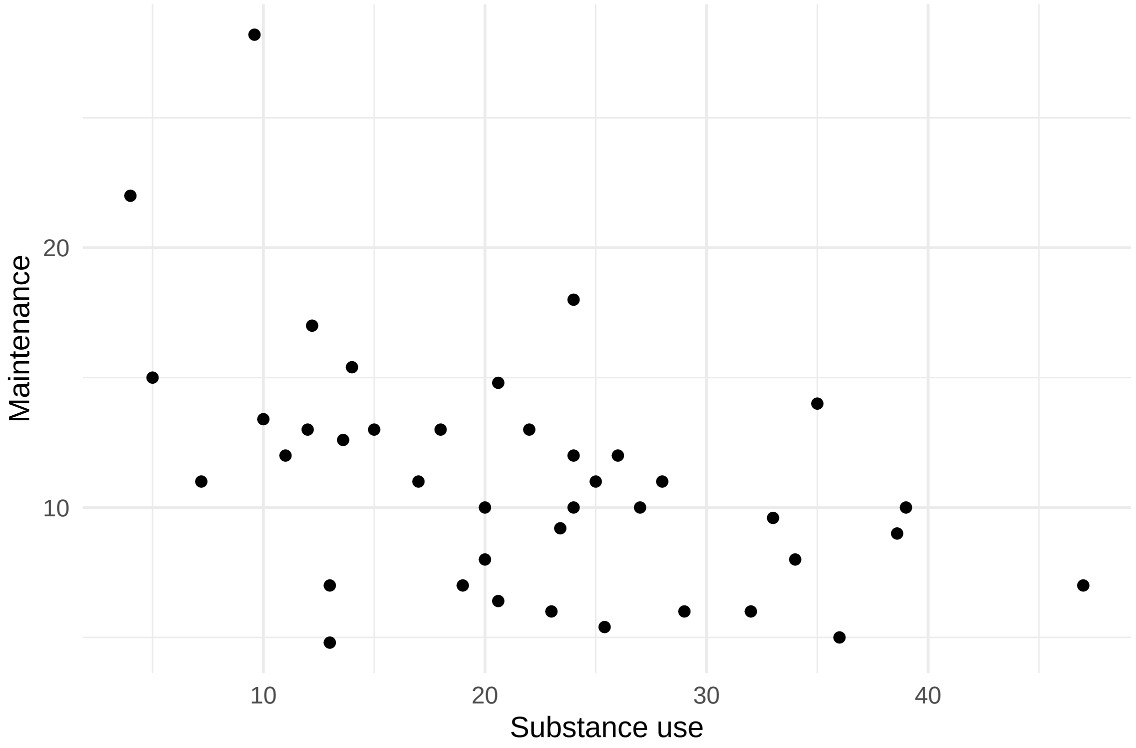


Interpreting this plot: The closer to blue this is the more a high weighting in one theme is correlated with a high weighting in another theme. The closer to red the more a high weighting in one theme is correlated with a low weighting in another theme.

We get some interesting observation from this plot. Firstly most almost all of the themes have some non trivial correlation with the other themes meaning further investigation could be fruitful.

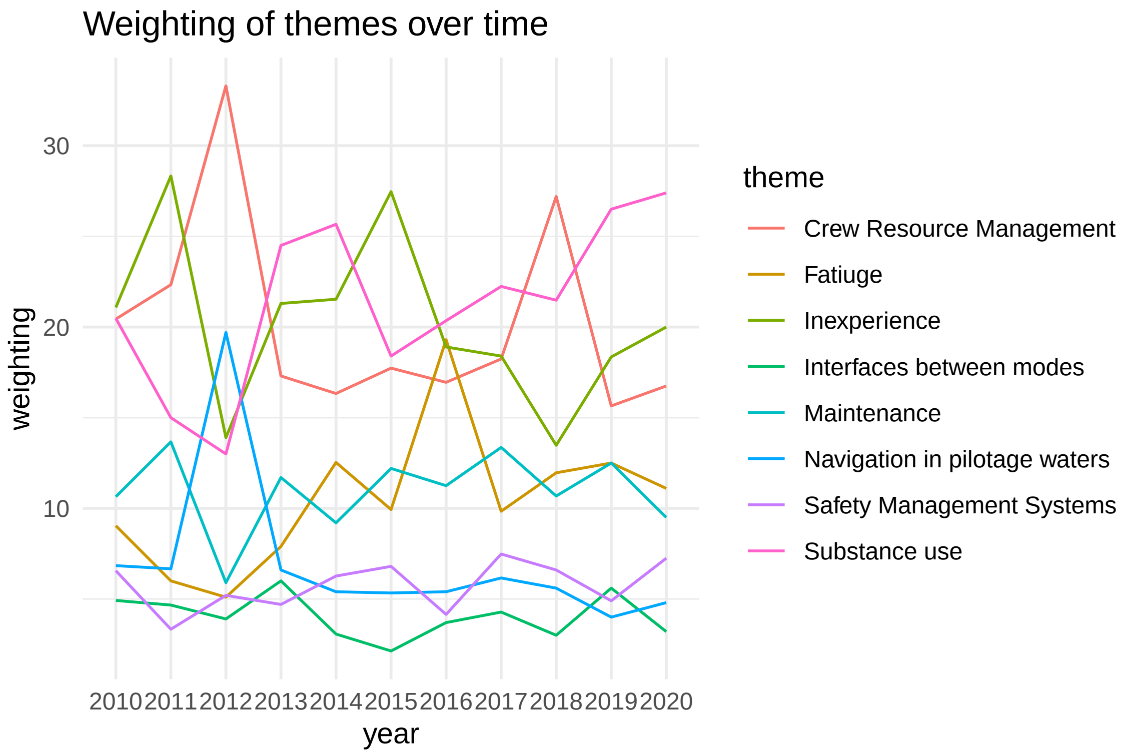
Substance abuse and Maintenance

An interesting correlation is between Substance abuse and Maintenance. This is a negative correlation meaning that an accident with a high weighting in Substance abuse is likely to have a low weighting. This could be considered intuitive as substance abuse is likey to be the main cause of an accident.



Trends over time

We can also look at trends over time. For example, how has the weighting of the themes changed over time?



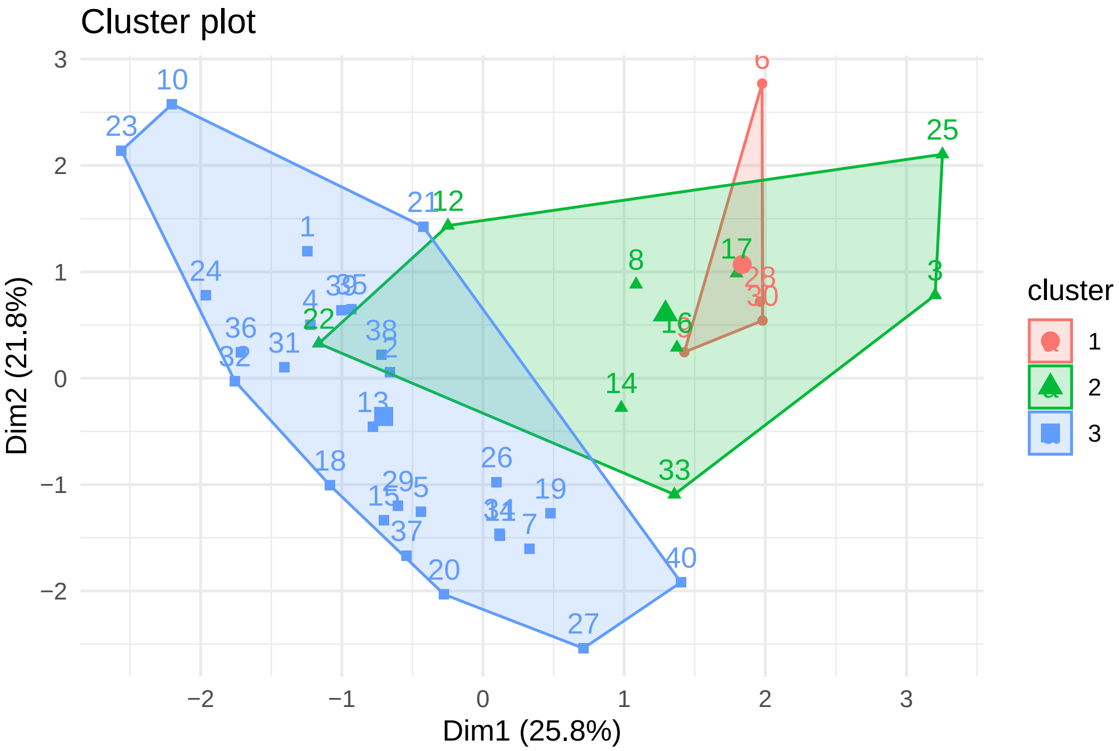
There is quite alot of variation between the yeras and clear trends are hard to find. We do see that consitantly some themes are more common than others. For example, Substance use is always more common than Interfaces between modes.

Clusters

Lets try and cluster the reports based on the themes. Having clusters of the reports would allow us to further dig into specific "types" of accidents.

We are going to do k means clustering. We need to decide how many clusters we want. We can do this by looking at the silhouette plot.

This suggests that there are about 3 different types of reports.



However looking more closely at the data we can see that the clusters are not very well defined. Further analysis on these clusters might help uncover the type or accidents that we have.

Conclusion

Overall we can see that there is some interesting analysis that can be done on this data. However, there is also alot of work that needs to be done to make it more useful. For example, the themes are not very well defined and there is alot of variation in the data.

Adding more variables to the data will help with deeper analysis. Currently theme weightings are void of any context. Adding in the actual text of the report will help with this. For example whether the accident was fatal. This will allow us to do more interesting analysis like "What themes are most common in fatal accidents?".