# Introduction

Previously, we explored the practical use of unsupervised machine learning (ML) in HR in an article titled “[A Tutorial on People Analytics Using R – Clustering](https://www.aihr.com/blog/people-analytics-r-clustering/)”. In this follow-up article, we will provide another practical application of an unsupervised ML algorithm—Association Rules. We will demonstrate how we use Association Rules in the context of Safety. Specifically, how Association Rules can assist in the synthesis of large amounts of safety data to identify intervention points to prevent future incidents, and new ways of visualising Association Rule outputs to make interpretation, and more importantly use of the findings, easier for HR leaders. We have included some practical code examples written in both Python and R. Let’s get started!

# Business Challenge

A large manufacturing company is committed to providing the safest working conditions for their employees. There are many initiatives to track and prevent injuries across their facilities, but safety incidents are still happening more frequently than they’d like. They capture very detailed data for each incident and at a recent meeting, the VP of Manufacturing asked HR if they can analyse the safety data to determine whether there are conditions commonly associated with incidents, which if addressed, could prevent future incidents. HR turned to their HR Analytics team to analyse the safety data and identify preventable conditions.

(Workplace safety datasets are not readily available, so we’ve decided to use a dataset of UK vehicle incidents to demonstrate how this analysis could be done.)

# Why Association Rules

There are multiple approaches you could use to identify conditions associated with safety incidents. However, Association Rules provide a means of rapidly identifying conditions that frequently occur together–a key priority for safety and incident prevention.

First, let’s do a quick review of the key components of Association Rules:

1. **Support:**

Support represents the probability of two conditions appearing together (i.e., number of incidents containing both Condition A and Condition B relative to the size of the dataset).

Support = Probability of Condition A and Condition B happening

1. **Confidence:**

Confidence represents the probability of Condition B happening given Condition A happened.

Confidence = Probability of Condition A and Condition B divided by the probability of Condition A

1. **Lift:**

Lift represents how certain we are that we can predict Condition B will happen if Condition A happens. It is generally agreed that a lift value greater than 1 reflects stronger associations.

Lift = Probability of Condition B given Condition A divided by the probability of Condition B

# Workflow

The workflow for addressing this business challenge is as follows:

1. Ingest and prepare the data. We will not spend too much time discussing the steps taken and instead focus on the following two steps.

2. Analyse the data. Under normal conditions we would spend considerable time performing exploratory data analysis prior to performing more sophisticated analyses, such as Association Rules. While this exploratory analysis was performed, it will not be discussed, and instead we will focus on the use of Association Rules in this article.

3. Visualise the results to inform preventive action. We provided two non-traditional ways of visualising Association Rules for two reasons:

1. The key tenets of Association Rules (i.e., Support, Confidence and Lift), do not always resonate with business stakeholders, resulting in confusion and a subsequent lack of action. Thus, we sought more intuitive forms of visualisation that better facilitate understanding and the call to action.
2. We believed that addressing preventable incident conditions is unlikely to have an adverse impact, and were therefore less rigid in our interpretation and visualisation than may be acceptable in other contexts.

The workflow provides a mix of R and Python, reflecting the background of the two authors.

# Ingest and Prepare the Data

First, we downloaded the necessary libraries for both coding languages. Next, we loaded the UK Road Safety Data (available on the [Kaggle website](https://www.kaggle.com/tsiaras/uk-road-safety-accidents-and-vehicles)). There are two data sets, accident information and vehicle information. Both sets are quite large, so we focussed the analysis on 2015 data only before we merged the two datasets to combine accident and vehicle conditions.

**[Code Snippet] - 1**

Using the Apriori Algorithm to find associations between different accident conditions requires categorical data, so we prepared and binned some of the data before we ran the analyses.

**[Code Snippet] - 2**

Finally, we focussed on the variables that represented pre-accident conditions. These variables were then one-hot encoded, a method that allows analysts to identify which conditions were and were not present during each incident (represented as a ‘1’ and ‘0’ respectively) .

**[Code Snippet] - 3**

# Analyse the Data

Our goal was to synthesise the large safety dataset to rapidly identify the pre-incident conditions (known as antecedents in Association Rules) that often result in different incident severity outcomes (known as the consequents). Therefore, we decided to run the analysis three times, once for each accident severity (i.e., slight, serious, and fatal). Our rationale was twofold;

1. The different incident severity outcomes occurred at different rates and thus needed different Support and consequent parameters applied to identify preventable conditions, and
2. To potentially enable prioritisation of results obtained from the analyses.

**Associations Rules for Slight Incidents**

Descriptive analyses revealed the majority of incidents to be ‘slight’. Therefore, we decided upon a higher Support parameter of 0.1, meaning that the combination of conditions occur at least 10 times in 100 incidents. We chose combinations of only 2 antecedents due to the visualisation methods we elected to use, and held the Lift parameter at 1 to identify stronger associations.

**[Code Snippet] - 4**

The analysis identified 144 association rules, or 144 combinations of two conditions that are associated with a slight incident, at least 10 times out of 100 incidents. These rules are visualised below.

**Associations for Serious Incidents**

There were fewer serious accidents than slight. Therefore, we decided to set the Support parameter at 0.05, meaning the combination of conditions identified had to occur at least 5 out of 100 times across the data set. The remaining parameters were held constant (i.e., antecedent combinations was 2, and Lift was 1).

**[Code Snippet] - 5**

Twenty-two associations, or 22 combinations of conditions that are associated with a serious accident, at least 5 times out of 100 incidents (total) were identified. These associations are visualised below.

**Associations for Fatal Incidents**

Descriptive analyses revealed far fewer fatal incidents compared to serious and slight. Despite a lower Support parameter of 0.01, no associations among antecedents were identified for fatal incidents, perhaps suggestive of very random events.

# Visualise Results

Following some experimentation, we chose two methods to visualise the Association Rules for slight and serious Incidents. These methods were:

1. Expanding Trees; and
2. Rosetype Diagram.

Below are Expanding Tree Diagrams for both slight and serious Incidents. The diagrams start from the left hand side with the incident type and can then be expanded to the right, showing the conditions that contribute. The size of the circles are based on the number of times the condition occurred in the full set of Association Rules identified.The key benefit of the Expanding Tree is the ease with which it can be both explained and understood by non-technical audiences.

**[Code Snippet] - 6**

The Rosetype Diagram (part of the [echarts4r](https://echarts4r.john-coene.com/articles/chart_types.html#rosetype-1) library) is a spin on the classic Pie Chart. In our context, it represents a further simplification of our generated Association Rules. Two conditions were identified in each rule (i.e., think two columns in a dataset), then each rule was reduced to a single column (i.e., one put on top of the other). Finally, a count of each unique incident condition was calculated. This count of incident conditions was visualised.

The overwhelmingly clear benefit of the Rosetype Diagram is the speed and clarity with which you can depict and understand the underlying rules. In a matter of minutes we were able to take over 250,000 recorded incidents and determine with confidence the co-occurring conditions most commonly associated with those incidents. Arguably a powerful example of rapid insight generation.

In the case of both the Expanding Tree Diagrams and the Rosetype Diagram, the downside is that the visualisations are a simplistic interpretation of the underlying rule identified, and a loss of the Support, Confidence and Lift detail from the algorithm. Given this reality, we used the core metrics of the Association Rules to generate a short-list of rules, which we then visualised.While these visualisations may not appeal to statistical purists, they have the potential to invite participation from non-technical audiences, particularly business leaders, at minimal cost to the “truth” of the underlying algorithm.

# Closing

In this article, we used Association Rules to analyse a large road-traffic incident dataset to identify preventable incident conditions. We found the approach to be helpful in rapidly synthesising safety data to identify intervention points, thereby getting to that “tell me what I need to know” point. While we used road-traffic incidents, the approach readily lends itself to organisational safety data, and one of the authors has used it to analyse military injury datasets to inform prevention.

Additionally, we sought to explore novel ways of visualising the output of Association Rules. We used Expanding Tree’s and Rosetype diagrams, both of which provided a clear and more readily accessible interpretation of Association Rules output than traditional methods. The visualisation methods used have the advantage of being readily interpretable by non-technical audiences, which may be of benefit when socialising results with safety and/or organisational leaders. However, they both overlooked technical metrics generated by the Association Rules, which may disenfranchise statistical purists.

If you are interested in using Association Rules, this work led us to believe that the method may be applied for other HR use cases, including:

* Attrition - identification of conditions that lead to employees staying or leaving an organization. You can find an example of this analysis [here](https://towardsdatascience.com/using-association-rules-with-categorical-data-e984f8bb8ee4).
* Performance - identification of conditions that lead to low or high performance; and
* Learning - identifying training courses/modules that employees often complete together, to feed a recommendation engine (i.e., people that completed this course also did X, Y, and Z courses).