Injury Analysis 2025

Introduction

Due to a recent unfortunate accident in South America public concerns about our company's workplace safety practicies have been voiced and are requesting answers.

As a global industrial manufacturing company operating internationally, ensuring compliance with local industrial regulations has split the company into unique safety regimes, all of which are under the microscope of the public eye.

Purpose and Research Questions

The purpose of this report is to perform an analysis of workplace injury data to help inform our response to these questions:

1. Of the various safety regimes in place across your company, which one would you recommend become the international standard for your company, based solely on injury prevention performance?
2. It has been suggested by senior management that industry experience is more important than the safety regime when it comes to preventing injuries. His idea is that a policy should be developed that is directly related to lowering employee turnover will reduce injury rates. Does the available data support this assertion?
3. Is there any relationship between:

- Injuries and the annual bonuses a proportion of employees received
- Injuries and whether staff have received any formal external qualifications e.g. external safety training or a university degree.

Summary of the Available Data

The data contains the counts of injuries and hours worked aggregated by the experience level of the workers for the last 12 months of operation. Each column of the dataset is:

- `record_id` - a **unique key** for the group of workers, by experience level
- `Injuries` - count of injuries in the group
- `Safety` - the safety regime in place for the group
- `Hours` - the total hours worked by the group
- `Experience` - the experience level of the group
- `bonus` - proportion of the group that recieved an annual bonus last year
- `training` - proportion of the group who have completed external safety training
- `university` - proportion of the group who have at least one university degree

Install and load required R packages:

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.3.3
```

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.3.3
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Open and read CSV data:

```r
data <- read.csv("injury-3.csv")
```

Data Processing and Exploratory Data Analysis (EDA)

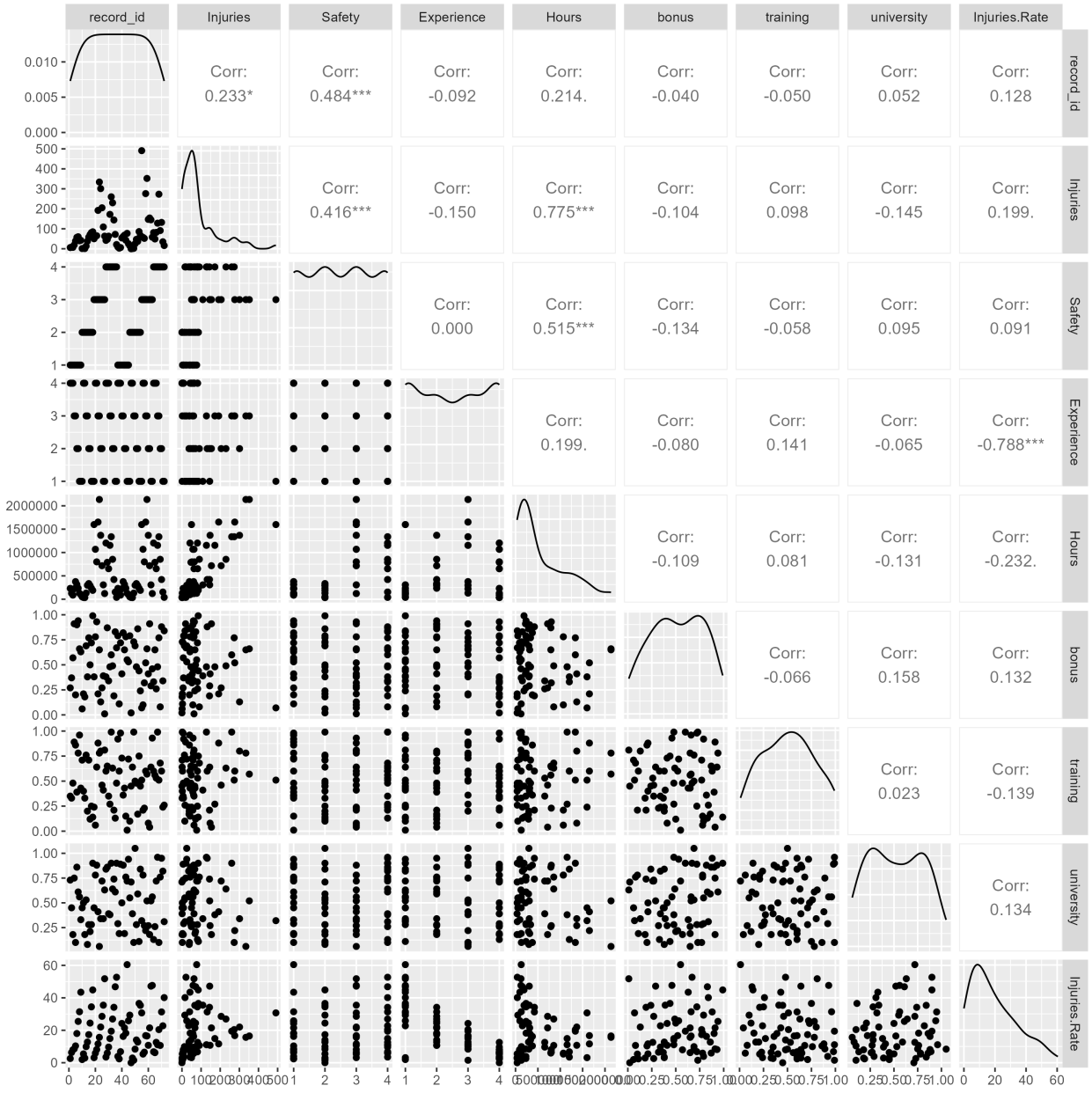
After viewing the dataset in its tabular form, it was theorized that there could be an interaction between `Injuries` and `Hours` due to colision theory where the longer the workers work within the `record_id` the more `Injuries` will occur.

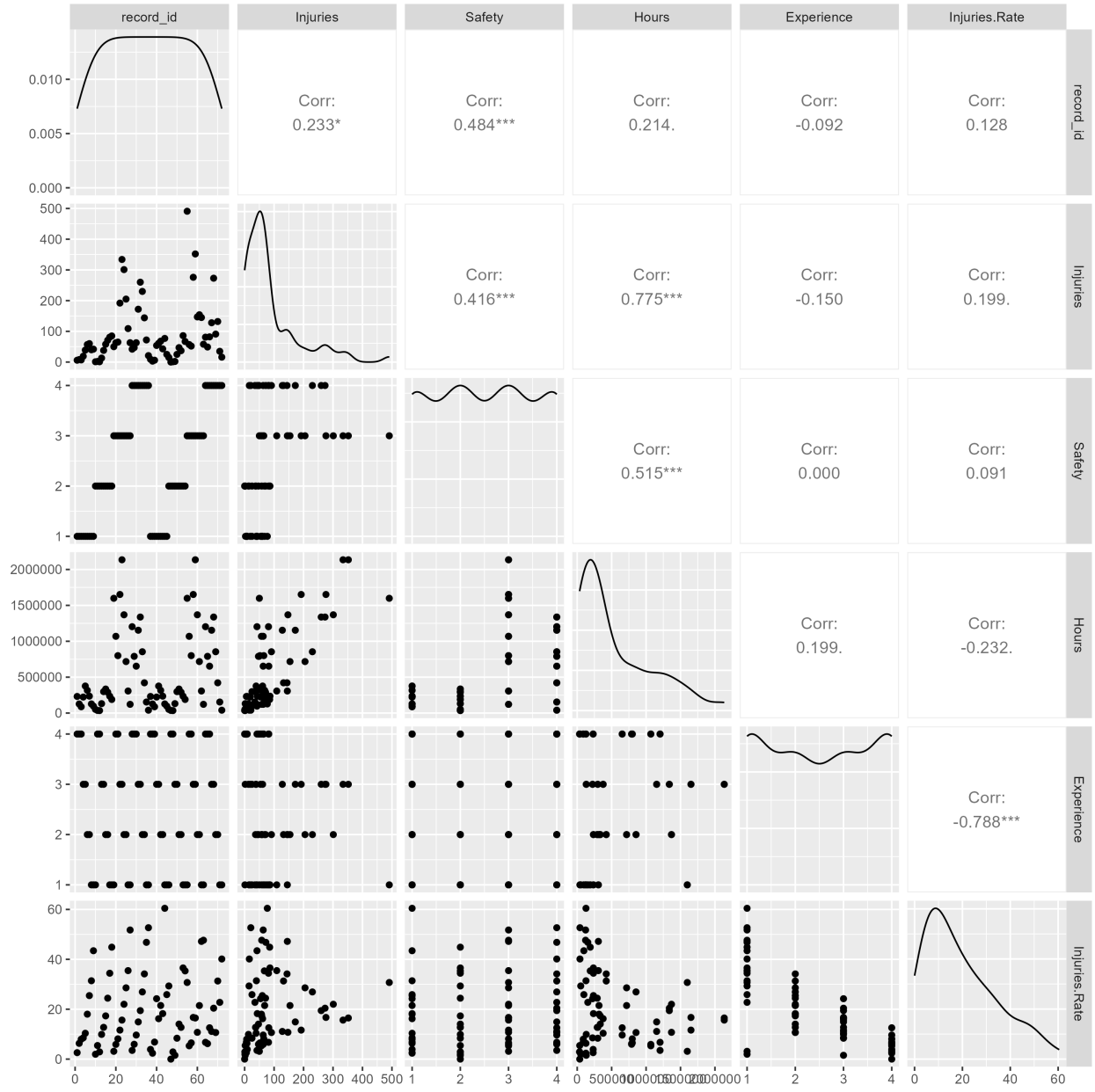
Therefore, in addition to the provided dataset, the `Injury` column will be mutated with the `Hours` column to provide a standardized, `Injury.Rate` column, equivalant to:

$$\text{Injuries.Rate} = \frac{\text{Injuries}}{\text{Hours}} \times 100\ 000$$

```r
data <- data %>% # inplace modification of data dataset
  mutate(Injuries.Rate = (Injuries / Hours) * 100000) # Injuries.Rate = No. Injuries per 100 000 Hours
```

```r
p <- ggpairs(data)
ggsave("figures/ggpairs_data.png", plot = p, dpi = 300, width = 10, height = 10)
p <- ggpairs(data, columns=c(
  "record_id",
  "Injuries",
  "Safety",
  "Hours",
  "Experience",
  "Injuries.Rate"
))
ggsave("figures/ggpairs_data_sigcorr.png", plot = p, dpi = 300, width = 10, height = 10)
```


record_id
Injuries
Safety
Experience
Hours
bonus
training
university
Injuries.Rate
Corr: 0.233*
Corr: 0.484***
Corr: -0.092
Corr: 0.214.
Corr: -0.040
Corr: -0.050
Corr: 0.052
Corr: 0.128
Corr: 0.416***
Corr: -0.150
Corr: 0.775***
Corr: -0.104
Corr: 0.098
Corr: -0.145
Corr: 0.199.
Corr: 0.000
Corr: 0.515***
Corr: -0.134
Corr: -0.058
Corr: 0.095
Corr: 0.091
Corr: 0.199.
Corr: -0.080
Corr: 0.141
Corr: -0.065
Corr: -0.788***
Corr: -0.109
Corr: 0.081
Corr: -0.131
Corr: -0.232.
Corr: -0.066
Corr: 0.158
Corr: 0.132
Corr: 0.023
Corr: -0.139
Corr: 0.134



From the above graphs the following interactions have the highest correlation:

- Injuries:record_id*

  Where the number of `Injuries` tends to increase as the `record_id` increases as the `record_id`. As the `record_id` is derived from the `Safety` regime and `Experience` of the workers this means that the interaction visible here is coming from both or either `Safety` and `Experience`.

- Safety:record_id***

  As expected, there is an interaction between `Safety` and `record_id` this is due to the `record_id` being an aggregated result from `Safety` and `Experience`.

- Safety:Injuries***

  From this, there is a correlation between the `Safety` and the number of `Injuries` from the rough correlation scale '***', this is a significant correlation that may indicate that some Safety regimes are

better than others at preventing injury, or it could also indicate that some Safety regimes may classify different things as an injury and as such will be reported or not reported as such.

- Hours:Injuries***

  This somewhat credits the assumption that as `Hours` increased the `Injuries` would also increase, which is why `Injuries.Rate` was created to reduce the affectes of collision theory, where the longer the workers worked in a record, the more `Injuries` would occur.

- Hours:Safety***

  This is interesting, `ggpairs` has detected a correlation between the number of `Hours` in a `record_id` and the `Safety` regime in-place, suggesting that some `Safety` reigme's have different requirments on how long workers can
  work.

- Injuries.Rate:Experience***

  There is also a correlation between the calculated `Injuries.Rate` and the workers `Experience`, referring to the image, this is a strong negative correlation, therefore, as expected, as the workers are more experienced within the group, the less injuries occur.

`Injuries` and `record_id` interaction

As identified above, the number of `Injuries` appears to interact with `record_id`, however, it is the least significant of the whole list of interactions. After manually inspecting the tabular raw data a weird pattern emerges with how the `record_id` was selected, the pattern is as follows:

- `Safety` is broken into 2 repeated groups 1 - 4 and 1 - 4 again
- for each value of `Safety`, `Experience` goes from 4 - 1
- as `record_id` increases, `Safety` increases until 4 and then starts again at 1

This pattern can also be seen from the ggpairs plots where the `Injuries` increases in repeating patterns for the `Injuries:record_id` interaction.

While, `record_id` isn't that significant in itself, the patterns emerging from it suggest there is a 2-way interaction for:

$$\text{Injuries} = \beta_0 + \beta_1 \cdot \text{Experience} + \beta_2 \cdot \text{Safety} + \beta_3 \cdot (\text{Experience} \cdot \text{Safety})$$

This can be tested with the `lm()` function:

```
model <- lm(Injuries ~ Experience * Safety, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Injuries ~ Experience * Safety, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -134.06  -48.62  -11.65   15.26  370.29
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         54.724     62.838   0.871    0.387
## Experience         -22.037     21.987  -1.002    0.320
## Safety              25.450     23.124   1.101    0.275
## Experience:Safety    3.892      8.037   0.484    0.630
##
```

```
## Residual standard error: 88.37 on 68 degrees of freedom
## Multiple R-squared:  0.1983, Adjusted R-squared:  0.163
## F-statistic: 5.607 on 3 and 68 DF,  p-value: 0.001702
```