Injury Analysis 2025

Context

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.

The purpose of this report is to perform an analysis of workplace injury data to help inform our response to recent journalist questions.

Research Questions

To help inform the CEO's response to this crisis, she has provided data on workplace injury within the company, as well as research questions to guide the analysis.

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 Available Data

The data provided for this report `injury.csv` contains counts of `Injuries` and `Hours` worked for the past 12 months of operation, aggregated by the `Experience` level of the workers and the workplace `Safety` regime in place at their factory as well as other variables that may have an affect on the number of `Injuries` at the factory.

More specifically:

- `Injuries` – count of injuries in the group

- `Safety` – the safety regime in place for the group
- `Hours` – total hours worked by the group, e.g. if 2 workers worked 10 hours, that's 20 hours total
- `Experience` – the experience level of the group
- `bonus` – proportion of the group who received 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

It is also important to note that there is no provided information on how many workers there are per group, or if all the groups have the same amount of workers. Therefore, for the purposes of this report, it is assumed that the number of workers in each group is insignificant to the number of Injuries that happens within a `report_id`

Data Processing / EDA

To inform the CEO with the relevant information to respond to journist's questions, an Exploratory Data Analysis was conducted.

Load data and libraries

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

```
# code credit to: https://statsandr.com/blog/an-efficient-way-to-install-and-load-r-packages/

# Package names
packages <- c("GGally", "ggpubr", "tidyverse", "backports", "patchwork", "MASS")

# Install packages not yet installed
installed_packages <- packages %in% rownames(installed.packages())
if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages])
}

# Packages loading
invisible(lapply(packages, library, character.only = TRUE))
```

```
## Warning: package 'GGally' was built under R version 4.4.3
```

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

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

```
## Warning: package 'ggpubr' was built under R version 4.4.3
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
```

```
## -- 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
##
## Attaching package: 'MASS'
##
##
## The following object is masked from 'package:patchwork':
##
##     area
##
##
## The following object is masked from 'package:dplyr':
##
##     select
```

Uni-variate Observations

Firstly, the raw tabular data was inspected to find any uni-variate distributions of interest.

```r
nrow(data) # get this to use for histogram bins later
```

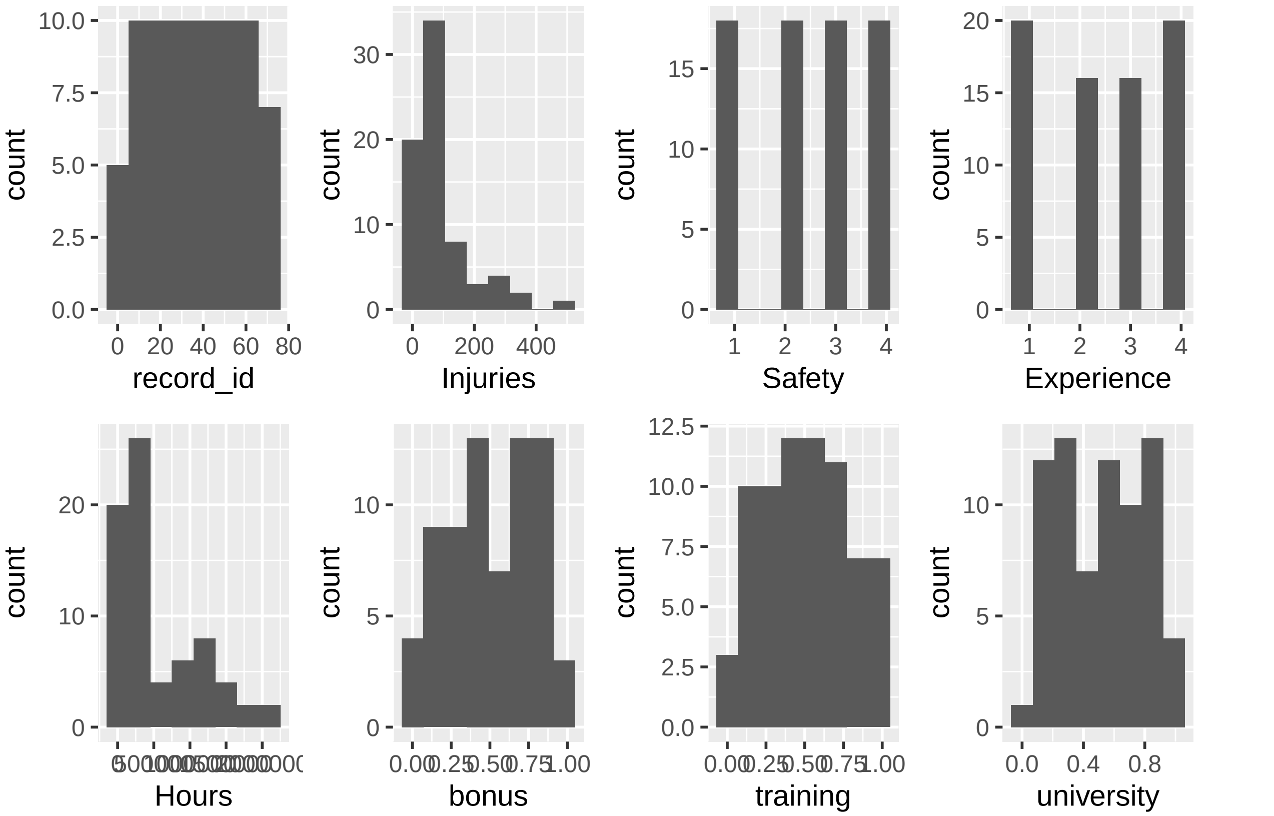
```
## [1] 72
```

```r
vars <- names(data)

plots <- lapply(vars, function(v) {
  ggplot(data, aes(x = .data[[v]])) +
    geom_histogram(bins = (nrow(data) %/% 9)) # 72 / 8 = 9
})

wrap_plots(plots, nrow = 2)
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```



```
min(data$Hours)
```

```
## [1] 34574
```

```
max(data$Hours)
```

```
## [1] 2135146
```

From these plots, the following observations can be made about the data:

- `record_id` - as expected, are roughly sequential and uniform across the histogram, no obvious missing values can be seen.
- `Injuries` - shows a strong right skewed distribution that also may be exponential, suggesting a poisson regression may be used to predict `Injuries`.
- `Safety`- four discrete spikes at 1, 2, 3, 4 show that `Safety` is a **categorical factor** describing 4 different `Safety` regime.
- `Experience` - also show discrete spikes at 1, 2, 3, 4 showing that `Experience` is also a **categorical factor** describing 4 different `Experience` levels, however, 2 and 3 are slightly under represented compared to 1 and 4.
- `Hours` - also shows a strong right skewed distribution, and, using `min` & `max(data$Hours)` can be seen that the range is 3474 – 2135146. With a similar uni-variate distribution to `Injury` - as well as `Hours` worked, intuitively being a factor that causes injury - this supports using an Injury Rate per 100 000 `Hours`, or, due to both following an exponential distribution, using the log (Injury Rate) for further multivariate analysis and log(Hours) as the offset function when fitting a model to the data.

- `bonus`, `training` and `university` - either weak right or left skews can be observed in these graphs. They all appear to be roughly normally distributed from 0 - 1 and all are **continuous factors**, however nothing meaningful can be extracted from just uni-variate analysis and more analysis is required to understand if these factors play any role in understanding injury prevention.

With the uni-variate analysis, it can be seen that the most suited GLM for modelling this data-set would be a Poisson Regression model. It can also be seen that using $\log \frac{\text{Injuries}}{\text{Hours}} \cdot 100\ 000$ to model this Poisson Regression model could be used for easier readability.

Multi-variate Observations

With the uni-variate observations done, $\log \frac{\text{Injuries}}{\text{Hours}} \cdot 100\ 000$ will now be used to conduct a multi-variate analysis to find significant factors.

`boxplots` can be used for the categorical, `Safety` and `Experience` variables in the data-set, whereas `bonus`, `training` and `university` are indiscreet, continuous variables requiring scatter-plots for multivariate analysis. All the variables will be plotted against the log injury rate.

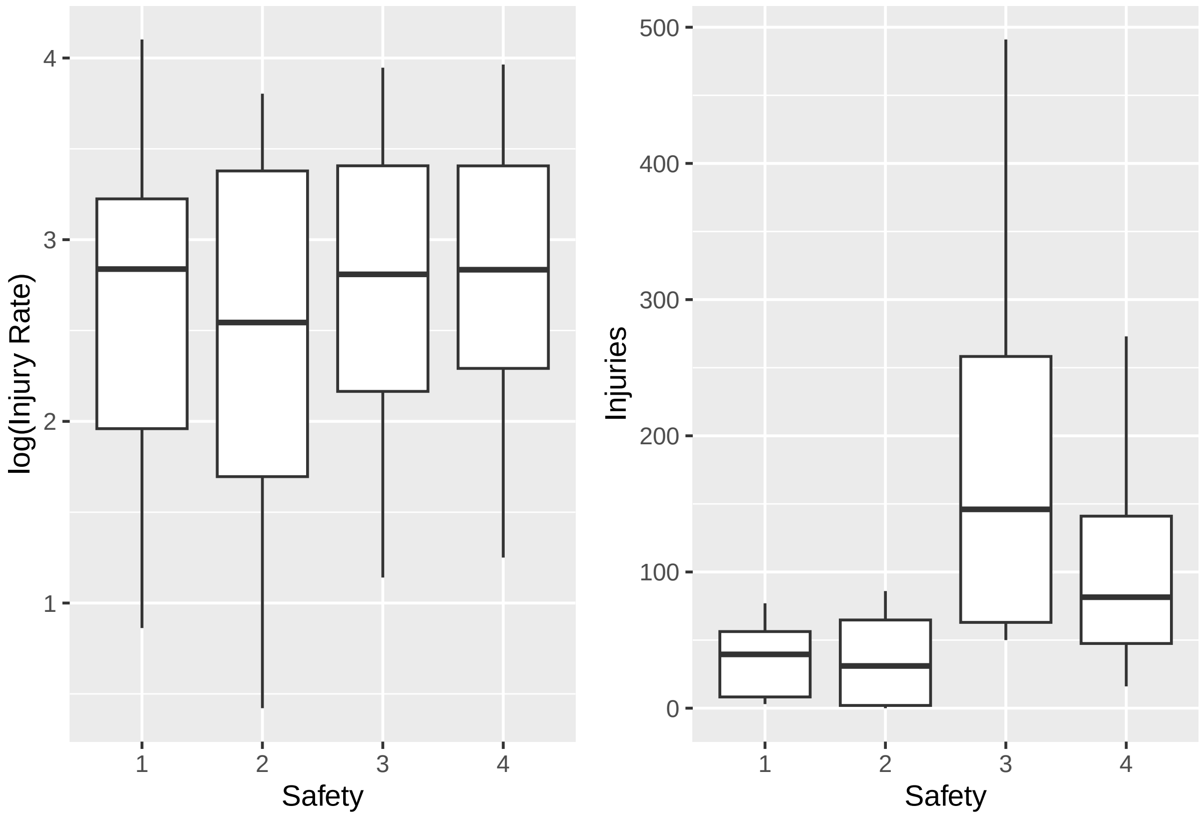
Categorical Variables

```r
data$Safety <- as.factor(data$Safety)

p1 <- ggplot(data, aes(x = Safety, y = log(Injuries / Hours * 100000))) +
  geom_boxplot() +
  labs(x = "Safety", y = "log(Injury Rate)")

p2 <- ggplot(data, aes(x = Safety, y = Injuries)) +
  geom_boxplot() +
  labs(x = "Safety", y = "Injuries")

p1 + p2
```



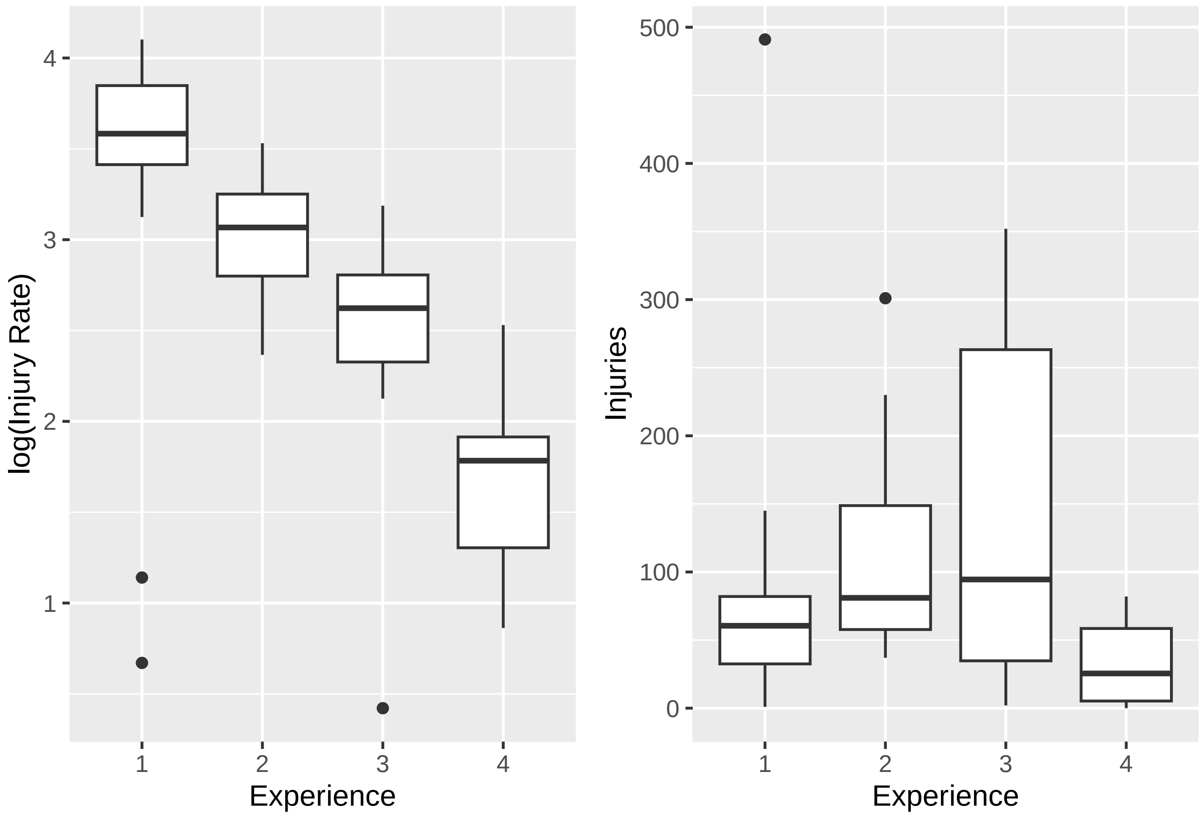
From the above log(Injury Rate) & `Injuries` vs `Safety` plots it can be seen that, while more total Injuries have occurred when `Safety` regime 3 is in-place, it doesn't have a significantly higher log(Injury Rate) meaning that more observation is required when looking at the model to determine whether or not `Safety` regime is relevant.

```
data$Experience <- as.factor(data$Experience)

p1 <- ggplot(data, aes(x = Experience, y = log(Injuries / Hours * 100000))) +
  geom_boxplot() +
  labs(x = "Experience", y = "log(Injury Rate)")

p2 <- ggplot(data, aes(x = Experience, y = Injuries)) +
  geom_boxplot() +
  labs(x = "Experience", y = "Injuries")

p1 + p2
```



From the above log(Injury Rate) & `Injuries` vs `Experience` plots it can be seen that, there is a very strong negative correlation between `Experience` level and the log(Injury Rate), indicating that a large amount of the variance in the Injury Rate can be explained with the `Experience` of the factory workers, supporting senior management's suggestion that industry `Experience` is important for preventing `Injuries` , however, more research is required to say whether `Safety` regime is not useful, as senior management also suggests.

Continuous Variables

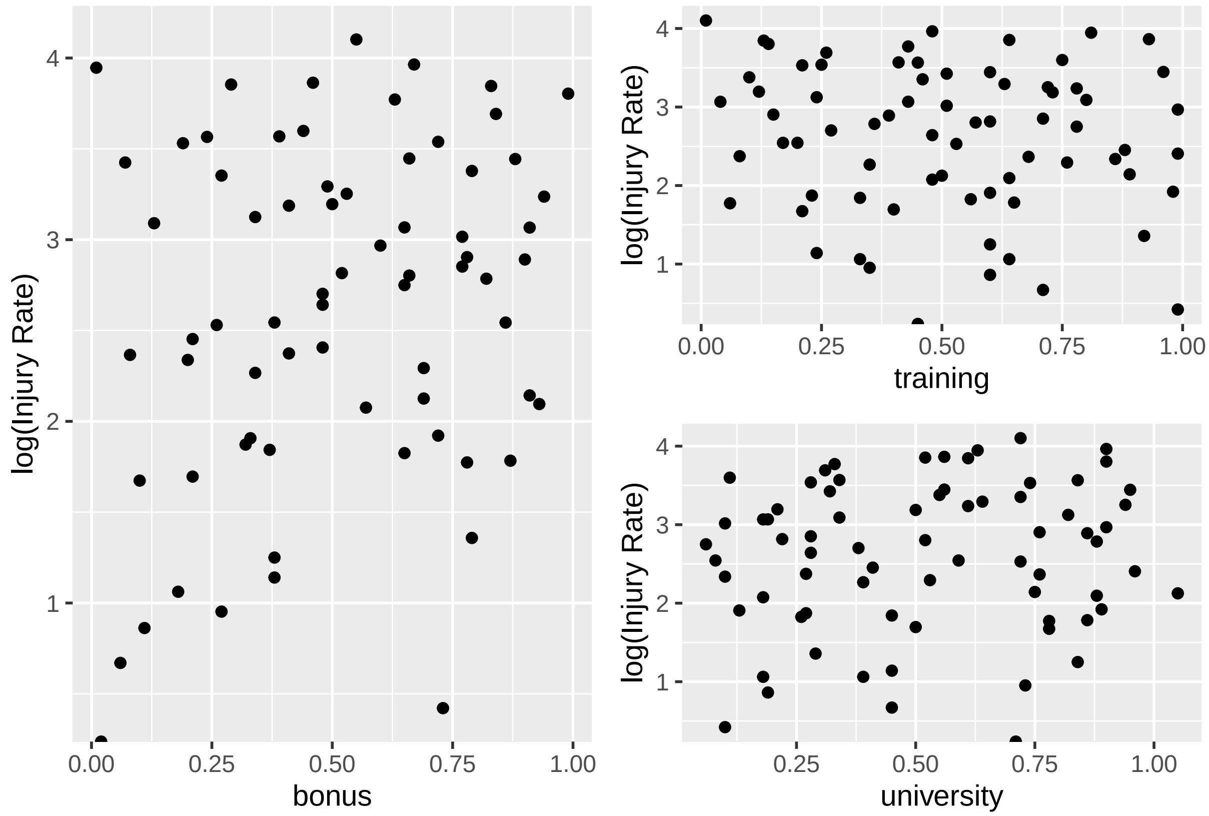
```
p1 <- ggplot(data, aes(x = bonus, y = log(Injuries / Hours * 100000))) +
  geom_point() +
  labs(x = "bonus", y = "log(Injury Rate)")

p2 <- ggplot(data, aes(x = training, y = log(Injuries / Hours * 100000))) +
  geom_point() +
  labs(x = "training", y = "log(Injury Rate)")

p3 <- ggplot(data, aes(x = university, y = log(Injuries / Hours * 100000))) +
  geom_point() +
  labs(x = "university", y = "log(Injury Rate)")

p1 + p2 /
p3
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
```



The continuous variables don't seem to have much meaningful relation to the log(Injury Rate) when directly plotted against each other, more observation is required when modelling to determine if the variables `bonus`, `training` and `university` have any compounded affect on the log(Injury Rate).

Modelling Approach and Justification

Poisson Regression

$$p(Y = y|\lambda) = \frac{\lambda^y \exp{(-\lambda)}}{y!}, \text{for } y = 0, 1, \cdots$$

As previously understood, the number of `Injuries` in a `record_id` follow an exponential function, alongside `Hours` following a similar distribution, these facts support using a Poisson regression algorithm with the $\log{(\text{Hours})}$ for the offset function.

Starting with a basic model with all the main effects

Firstly, a basic model containing all the main effects of the dataset.

```r
fit <- glm(
  Injuries ~ offset(log(Hours)) + Safety + Experience + bonus + training + university,
  family = "poisson",
  data = data
)
summary(fit)
```

```
##
## Call:
## glm(formula = Injuries ~ offset(log(Hours)) + Safety + Experience +
##     bonus + training + university, family = "poisson", data = data)
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -8.38619    0.07166 -117.026  < 2e-16 ***
## Safety2     -0.08419    0.05771   -1.459   0.1446
## Safety3     -0.05319    0.04809   -1.106   0.2687
## Safety4      0.12991    0.04817    2.697   0.0070 **
## Experience2 -0.37301    0.03568  -10.453  < 2e-16 ***
## Experience3 -0.78098    0.04119  -18.959  < 2e-16 ***
## Experience4 -1.58138    0.05007  -31.586  < 2e-16 ***
## bonus        0.32761    0.05919    5.535 3.11e-08 ***
## training     0.32214    0.05700    5.651 1.59e-08 ***
## university  -0.09180    0.05224   -1.758   0.0788 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2449.8  on 70  degrees of freedom
## Residual deviance: 1136.8  on 61  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 1556.8
##
## Number of Fisher Scoring iterations: 5
```

Unexpectedly, the model appears to have pulled reason from the chaos in the variables `bonus` and `training`, and, although much less significant, the effect of `university` is only 7.88% due to chance. This shows that in an all encompassing model, these seemingly random variables do have an effect on the final log (Injury Rate) warranting further analysis.

Let R decide which interactions are important

With the discoveries in the initial model, modelling the main effects, `stepAIC` from `MASS` is used to choose a model by AIC in a Stepwise Algorthm:

```
step_fit <- stepAIC(
  fit,
  scope = ~ (Safety + Experience + bonus + training + university)^2,
  direction = "both"
)
```

```
## Start:  AIC=1556.76
## Injuries ~ offset(log(Hours)) + Safety + Experience + bonus +
##     training + university
##
##                          Df Deviance    AIC
## + Safety:training         3   926.37 1352.3
## + Experience:training     3  1035.21 1461.2
## + Safety:Experience       9  1042.30 1480.3
## + training:university    1  1096.53 1518.5
## + Experience:university  3  1110.61 1536.6
## + Safety:university      3  1110.88 1536.8
## + bonus:training         1  1119.78 1541.7
## + Safety:bonus           3  1119.66 1545.6
## + Experience:bonus       3  1125.73 1551.7
## <none>                      1136.80 1556.8
## - university             1  1139.88 1557.8
## + bonus:university       1  1136.08 1558.0
## - Safety                 3  1169.97 1583.9
## - bonus                  1  1167.37 1585.3
## - training               1  1168.86 1586.8
## - Experience             3  2390.47 2804.4
##
## Step:  AIC=1352.33
## Injuries ~ Safety + Experience + bonus + training + university +
##     Safety:training + offset(log(Hours))
##
##                         Df Deviance    AIC
## + Experience:training    3   799.14 1231.1
## + Safety:Experience      9   827.77 1271.7
## + Experience:bonus       3   895.84 1327.8
## + Safety:bonus           3   902.17 1334.1
## + Safety:university      3   911.72 1343.7
## + training:university    1   920.25 1348.2
## - university             1   926.39 1350.3
## <none>                       926.37 1352.3
## + bonus:university       1   924.64 1352.6
## + Experience:university  3   921.49 1353.5
## + bonus:training         1   926.31 1354.3
## - bonus                  1   979.47 1403.4
## - Safety:training        3  1136.80 1556.8
## - Experience             3  2158.45 2578.4
##
## Step:  AIC=1231.11
## Injuries ~ Safety + Experience + bonus + training + university +
##     Safety:training + Experience:training + offset(log(Hours))
##
##                         Df Deviance    AIC
## + Safety:Experience      9   704.46 1154.4
## + Safety:bonus           3   777.48 1215.4
## + Safety:university      3   787.74 1225.7
## + Experience:bonus       3   788.23 1226.2
## - university             1   799.33 1229.3
## + Experience:university  3   792.75 1230.7
## <none>                       799.14 1231.1
## + training:university    1   798.42 1232.4
## + bonus:training         1   798.58 1232.5
## + bonus:university       1   799.14 1233.1
## - bonus                  1   873.33 1303.3
## - Experience:training    3   926.37 1352.3
## - Safety:training        3  1035.21 1461.2
##
## Step:  AIC=1154.43
## Injuries ~ Safety + Experience + bonus + training + university +
##     Safety:training + Experience:training + Safety:Experience +
##     offset(log(Hours))
##
##                         Df Deviance    AIC
## + Experience:bonus       3   654.03 1110.0
## + bonus:training         1   689.02 1141.0
## + Safety:university      3   687.90 1143.9
## + Safety:bonus           3   691.13 1147.1
## + Experience:university  3   696.86 1152.8
## - university             1   706.22 1154.2
## <none>                       704.46 1154.4
## + training:university    1   702.50 1154.5
## + bonus:university       1   704.34 1156.3
## - bonus                  1   761.71 1209.7
## - Safety:Experience      9   799.14 1231.1
## - Experience:training    3   827.77 1271.7
## - Safety:training        3   943.98 1387.9
##
## Step:  AIC=1109.99
## Injuries ~ Safety + Experience + bonus + training + university +
##     Safety:training + Experience:training + Safety:Experience +
##     Experience:bonus + offset(log(Hours))
##
##                         Df Deviance    AIC
## + Safety:university      3   614.03 1076.0
## + Safety:bonus           3   637.36 1099.3
## + bonus:training         1   645.33 1103.3
## + Experience:university  3   643.37 1105.3
## - university             1   654.52 1108.5
## + training:university    1   651.67 1109.6
## <none>                       654.03 1110.0
## + bonus:university       1   653.39 1111.4
## - Experience:bonus       3   704.46 1154.4
## - Experience:training    3   717.43 1167.4
## - Safety:Experience      9   788.23 1226.2
## - Safety:training        3   931.57 1381.5
##
## Step:  AIC=1076
## Injuries ~ Safety + Experience + bonus + training + university +
##     Safety:training + Experience:training + Safety:Experience +
##     Experience:bonus + Safety:university + offset(log(Hours))
##
##                         Df Deviance    AIC
## + bonus:training         1   607.35 1071.3
## + Experience:university  3   605.54 1073.5
## + bonus:university       1   610.34 1074.3
## <none>                       614.03 1076.0
## + Safety:bonus           3   608.65 1076.6
## + training:university    1   612.93 1076.9
## - Safety:university      3   654.03 1110.0
## - Experience:training    3   674.47 1130.4
## - Experience:bonus       3   687.90 1143.9
## - Safety:Experience      9   769.59 1213.6
## - Safety:training        3   915.41 1371.4
##
## Step:  AIC=1071.32
## Injuries ~ Safety + Experience + bonus + training + university +
##     Safety:training + Experience:training + Safety:Experience +
##     Experience:bonus + Safety:university + bonus:training + offset(log(Hours))
##
##                          Df Deviance    AIC
## + Experience:university  3   598.56 1068.5
## + Safety:bonus           3   599.65 1069.6
## + bonus:university       1   605.12 1071.1
## <none>                       607.35 1071.3
## + training:university     1   606.11 1072.1
## - bonus:training         1   614.03 1076.0
## - Safety:university      3   645.33 1103.3
## - Experience:bonus       3   667.53 1125.5
## - Experience:training    3   674.34 1132.3
## - Safety:Experience      9   769.01 1215.0
## - Safety:training        3   904.63 1362.6
##
## Step:  AIC=1068.52
## Injuries ~ Safety + Experience + bonus + training + university +
##     Safety:training + Experience:training + Safety:Experience +
##     Experience:bonus + Safety:university + bonus:training + Experience:university +
##     offset(log(Hours))
##
##                          Df Deviance    AIC
## + Safety:bonus           3   588.12 1064.1
## <none>                       598.56 1068.5
## + bonus:university       1   598.38 1070.3
## + training:university     1   598.44 1070.4
## - Experience:university  3   607.35 1071.3
## - bonus:training         1   605.54 1073.5
## - Safety:university      3   634.60 1098.6
## - Experience:bonus       3   658.98 1122.9
## - Experience:training    3   661.95 1125.9
## - Safety:Experience      9   765.15 1217.1
## - Safety:training        3   891.54 1355.5
##
## Step:  AIC=1064.08
## Injuries ~ Safety + Experience + bonus + training + university +
##     Safety:training + Experience:training + Safety:Experience +
##     Experience:bonus + Safety:university + bonus:training + Experience:university +
##     Safety:bonus + offset(log(Hours))
##
##                          Df Deviance    AIC
## <none>                       588.12 1064.1
## + training:university     1   587.92 1065.9
## + bonus:university       1   587.96 1065.9
## - Safety:bonus           3   598.56 1068.5
## - Experience:university  3   599.65 1069.6
## - bonus:training         1   597.36 1071.3
## - Safety:university      3   604.21 1074.2
## - Experience:bonus       3   650.58 1120.5
## - Experience:training    3   652.53 1122.5
## - Safety:Experience      9   743.27 1201.2
## - Safety:training        3   836.06 1306.0
```

Compare models

```
AIC(fit, step_fit)
```

```
##          df      AIC
## fit      10 1556.758
## step_fit 38 1064.084
```

```
BIC(fit, step_fit)
```

```
##          df      BIC
## fit      10 1579.385
## step_fit 38 1150.066
```

```
summary(step_fit)
```

```
##
## Call:
## glm(formula = Injuries ~ Safety + Experience + bonus + training +
##     university + Safety:training + Experience:training + Safety:Experience +
##     Experience:bonus + Safety:university + bonus:training + Experience:university +
##     Safety:bonus + offset(log(Hours)), family = "poisson", data = data)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -6.25657    0.27646 -22.631  < 2e-16 ***
## Safety2                 -1.02457    0.27105  -3.780 0.000157 ***
## Safety3                 -3.05171    0.24032 -12.698  < 2e-16 ***
## Safety4                 -0.99902    0.24415  -4.092 4.28e-05 ***
## Experience2             -1.24311    0.24400  -5.095 3.49e-07 ***
## Experience3             -1.53761    0.29214  -5.263 1.42e-07 ***
## Experience4             -1.76016    0.28053  -6.274 3.51e-10 ***
## bonus                   -1.76881    0.43704  -4.047 5.18e-05 ***
## training                -0.62184    0.27776  -2.239 0.025172 *
## university              -0.79562    0.24620  -3.232 0.001231 **
## Safety2:training        -0.34937    0.27085  -1.290 0.197077
## Safety3:training         2.70839    0.22699  11.932  < 2e-16 ***
## Safety4:training         0.60264    0.29781   2.024 0.043013 *
## Experience2:training    -1.19396    0.17238  -6.926 4.32e-12 ***
## Experience3:training    -1.03644    0.22091  -4.692 2.71e-06 ***
## Experience4:training    -2.10002    0.34163  -6.147 7.89e-10 ***
## Safety2:Experience2     -0.25205    0.16559  -1.522 0.127988
## Safety3:Experience2      1.27612    0.16528   7.721 1.16e-14 ***
## Safety4:Experience2      0.47701    0.18495   2.579 0.009906 **
## Safety2:Experience3     -0.49465    0.18239  -2.712 0.006686 **
## Safety3:Experience3      0.47508    0.14873   3.194 0.001402 **
## Safety4:Experience3    -0.05689    0.19680  -0.289 0.772533    
## Safety2:Experience4    -0.15185    0.61855  -0.246 0.806068    
## Safety3:Experience4     1.31572    0.23322   5.642 1.69e-08 ***
## Safety4:Experience4     0.61819    0.23740   2.604 0.009213 ** 
## Experience2:bonus       1.55435    0.21782   7.136 9.60e-13 ***
## Experience3:bonus       1.51104    0.25679   5.884 4.00e-09 ***
## Experience4:bonus       0.66577    0.24773   2.688 0.007198 ** 
## Safety2:university      0.37374    0.27267   1.371 0.170489    
## Safety3:university      0.91288    0.25198   3.623 0.000291 ***
## Safety4:university      0.49216    0.27640   1.781 0.074981 .  
## bonus:training          1.04078    0.34353   3.030 0.002449 ** 
## Experience2:university  0.08206    0.22508   0.365 0.715432    
## Experience3:university  0.49518    0.21165   2.340 0.019303 *  
## Experience4:university -0.01696    0.26376  -0.064 0.948728    
## Safety2:bonus           1.31392    0.41531   3.164 0.001558 ** 
## Safety3:bonus           0.78188    0.36542   2.140 0.032382 *  
## Safety4:bonus           0.83851    0.37462   2.238 0.025203 *  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 2449.83  on 70  degrees of freedom
## Residual deviance:  588.12  on 33  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 1064.1
## 
## Number of Fisher Scoring iterations: 5
```

While including 3-way interactions may reduce the AIC, such complexity offers little practical value. Moreover, none of the potential 3-way relationships in the data-set are theoretically meaningful, and their inclusion would likely just result in over-fitting rather than insight.

Validity of Model and Modelling Results

These results align with the previous relationships observed in the Multi-variate analysis, however, interestingly, `Safety` regime 4 appears to have a statistically significant relationship causing a higher $\log(\text{Injury Rate})$ with `Safety` regime's 1, 2 and 3 causing statistically insignificant differences in $\log(\text{Injury Rate})$. Another expected observation from the fitted model is the `Experience` values, as `Experience` level increases, significant decreases in $\log(\text{Injury Rate})$ can be seen from the increasingly negative coefficient estimates. The final model's fit achieves a good AIC score of 1064.0 a $\Delta - 492.7$ improvement when compared to the previous, 1556.7, achieved by the base model.More explanation behind the statistical significance of `bonus`, `training` and `university` is also uncovered in the final model.

Recommendations and Conclusions

Answers to Research Questions

1. Recommended Safety Regime

Based on the final model, **Safety Regime 3** is linked with the lowest log (Injury Rate) with a coefficient of ~ -3.05 where $-3.05 < \min\{0.00, -1.00, -1.02\}$. This safety regime proves to be effective at using `bonuses` as incentives to reduce injury in high risk jobs, as well as ensuring workers are propery trained and qualified for high risk jobs. As such, it is recommended that **Safety Regime 3** is adopted as an international standard for the company. Moreover, further research into local regulations and potential legal constraints should be conducted to ensure that the international implementation of this Safety Regime complies with all applicable laws in each operating region.

2. Importance of Industry Experience vs Safety Regime

The final model supports senior management's claims that Industry Experience is a very important factor, where **the more experienced the workers are, the less injuries occur**. However, `Safety` Regime still contributes significantly to the log (Injury Rate) within the injury reports.

3. Influence of `bonuses`, `training` and `university`

`bonuses` From the final model it can be seen that `bonuses`, by their-selves, have a slight negative correlation with the log (Injury Rate) showing that giving a worker a bonus, in and of itself reduces injuries. However, two way interactions, such as `Experience:bonus`, `Safety:bonus`, both produce positive correlations for predicting the log (Injury Rate) for all factors of `Experience` and `Safety`. Logically this positive correlation makes sense as more experienced workers would handle difficult, and possibly more risky jobs and different safety regimes may use bonuses as incentives for those more risky jobs. Linking back to `bonuses`'s coefficient is negative, it shows that the incentives in these theoretically high risk jobs do, in fact, work at reducing injuries in these jobs.

`training` Alike `bonuses`, `training` also has a slight negative coefficient in the final model where training proves effective at reducing injury in factories. With the interaction between `training` and `Safety` it can be seen that for regimes 2 and 4, `training` doesn't have much effect on the log (Injury Rate), however, for regime 3, a significantly higher number of injuries occur as `training` increases. As this safety regime is largely effective at reducing injury it can't be simply assumed that the training in this regime is just ineffective, and this could be due to this regime giving more training to those working in higher risk jobs within the factory. More investigation is required to determine if this is true or if the `training` within Safety Regime 3 needs improvement. As expected, the interaction between `training` and `Experience` appears to go hand in hand at preventing injury within the factories, the more experienced a worker is and the more training they receive, the lower their chances are of injuring themselves. The interaction between `training` and `bonus` was also deemed significant by AIC tests, showing that as these variables increase, the rate of injury increases, this observation supports the theory that `bonuses` are given to those in high risk jobs that require more `training`.

`university` The model follows intuition where the more qualified a group of workers are, the less injuries occur with a slightly negative but significant correlation between `university` and log (Injury Rate). A similar trend to `training` can be seen with the interactions between `Safety` and `Experience` where safety regime 3 is giving higher risk jobs to more qualified workers. Where the most experienced workers with qualifications show a negative correlation with injury, lowering injury rates with highly experienced and qualified workers.

Conclusion

This report used a Poisson Regression Algorithm to fit a GLM to workplace injury data, incorporating the whole dataset which includes: the number of `Injuries` and total `Hours` worked by all workers in the record, aggregated by the `Safety` regime of the factory and `Experience` of the workers, including percentages for the proportion of workers with `bonuses`, `training` and `university` degrees or qualifications within the group. The final model achieved a strong fit AIC : 1064.0 confirming key relationships:

- **Safety Regime 3** is most effective at injury prevention compared to other regimes in-place.
- **More experience** significantly reduces injury rates.
- Bonuses, training and university degrees when utilized correctly reduce injury rates.

A comprehensive injury reduction strategy should be put in action to implement Safety Regime 3 internationally, retain experienced workers and refine the usage of bonuses, training and implement education programs to increase the proportion of qualified workers.