Course Project: Phase 1

Taoufic Gandi

2025-06-14

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(caret)

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(ggplot2)  
library(corrplot)

## corrplot 0.95 loaded

library(GGally)

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

# Load data  
train <- read\_csv("~/BAN-502-802-803/train.csv")

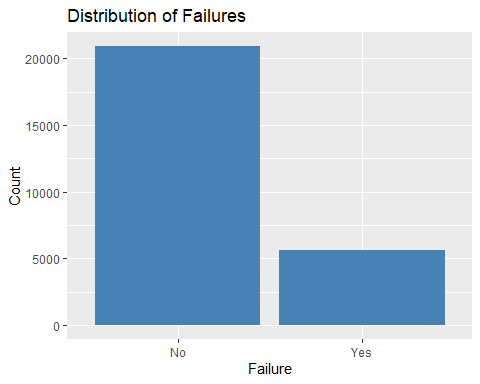
## Rows: 26570 Columns: 26  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): product\_code, attribute\_0, attribute\_1, failure  
## dbl (22): id, loading, attribute\_2, attribute\_3, measurement\_0, measurement\_...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

test <- read\_csv("~/BAN-502-802-803/test.csv")

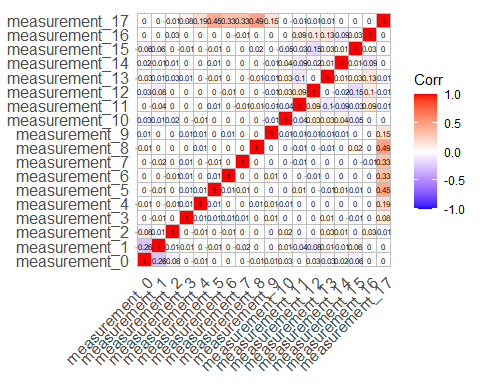
## Rows: 20775 Columns: 25  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): product\_code, attribute\_0, attribute\_1  
## dbl (22): id, loading, attribute\_2, attribute\_3, measurement\_0, measurement\_...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#Cleaning Missing Values  
missing\_summary <- train %>%  
 summarise(across(everything(), ~ sum(is.na(.)))) %>%  
 pivot\_longer(cols = everything(), names\_to = "Variable", values\_to = "Missing") %>%  
 arrange(desc(Missing)) %>%  
 filter(Missing > 0)

train$failure <- as.factor(train$failure)  
  
ggplot(train, aes(x = failure)) +  
 geom\_bar(fill = "steelblue") +  
 labs(title = "Distribution of Failures", x = "Failure", y = "Count")



# Correlation analysis  
numeric\_vars <- train %>% select(starts\_with("measurement"))  
cor\_matrix <- cor(numeric\_vars, use = "pairwise.complete.obs")  
ggcorrplot::ggcorrplot(cor\_matrix, lab = TRUE, lab\_size = 2)



# Create an average of all measurements as a new feature  
train$avg\_measurement <- rowMeans(select(train, starts\_with("measurement")), na.rm = TRUE)

ggplot(train, aes(x = failure, y = avg\_measurement)) +  
 geom\_boxplot(fill = "orange") +  
 labs(title = "Avg Measurement vs Failure", y = "Average of Measurements")

