Final Project in R

Shahan Perera

AI4I 2020 Predictive Maintenance Dataset

2024-04-30

Table of Contents

Abstract

This project explores a dataset compromising 10,000 data points with 14 features aimed at understanding and predicting failures in manufacturing processes. The dataset encompasses various parameters, including air temperature, process temperature, rotational speed, torque, tool wear, and machine failure labels associated with five distinct failure modes.

This project aims to uncover patterns and insights within the data to enhance process reliability and efficiency. Methodologically, the project involves data preprocessing, exploratory data analysis, feature engineering, and machine learning model development.

Key findings from the analysis include identifying factors contributing to process failures, such as tool wear, heat dissipation, power failure, overstrain, and random failures. Notably, tool wear and heat dissipation emerge as prominent predictors of machine failure.

Proposed next steps involve leveraging machine learning techniques to build predictive models capable of identifying potential failures before they occur. This includes developing classification models based on the identified failure modes to predict machine failures. Furthermore, there is an opportunity to explore advanced anomaly detection algorithms to detect subtle deviations in process parameters indicative of impending failures.

As a result, this project aims to equip manufacturing processes with predictive capabilities to minimize downtime, optimize maintenance schedules, and improve overall operational efficiency. Manufacturers can proactively address potential failures and enhance productivity by harnessing the insights gleaned from this dataset.

Body of the Report

Description of Work Done

My project delves into the comprehensive analysis and discovery of insights within a dataset aimed at predictive maintenance in manufacturing processes. The dataset, comprising 10,000 data points with 14 features, encapsulates crucial parameters, including air temperature, process temperature, rotational speed, torque, tool wear, and machine failure labels associated with five distinct failure modes.

The methodology adopted for this project commences with meticulous data collection utilizing the readxl package in R to import the dataset. Following this, a judicious assessment of the data revealed that cleaning procedures were unnecessary due to the structured nature of the dataset. Nonetheless, certain columns, such as TWF, HDF, PWF, OSF, and RNF, were prudently removed as they were considered superfluous.

Subsequently, descriptive statistics were meticulously generated employing the summary function, elucidating the numerical characteristics inherent within the dataset. This provided invaluable insights into the distributional attributes of various parameters, laying a solid foundation for subsequent analysis.

Further analysis was undertaken through exploratory data analysis (EDA), encompassing the creation of histograms, bar plots, scatter plots, and box plots using ggplot2. These visualizations served as potent tools in unraveling intricate patterns and relationships embedded within the data, thereby augmenting the understanding of critical variables and their interplay.

Additionally, the correlation between variables was meticulously examined utilizing the cor function and subsequently visualized by creating a correlation heatmap using the corrplot package. This facilitated the elucidation of intricate relationships and dependencies among the myriad parameters under scrutiny.

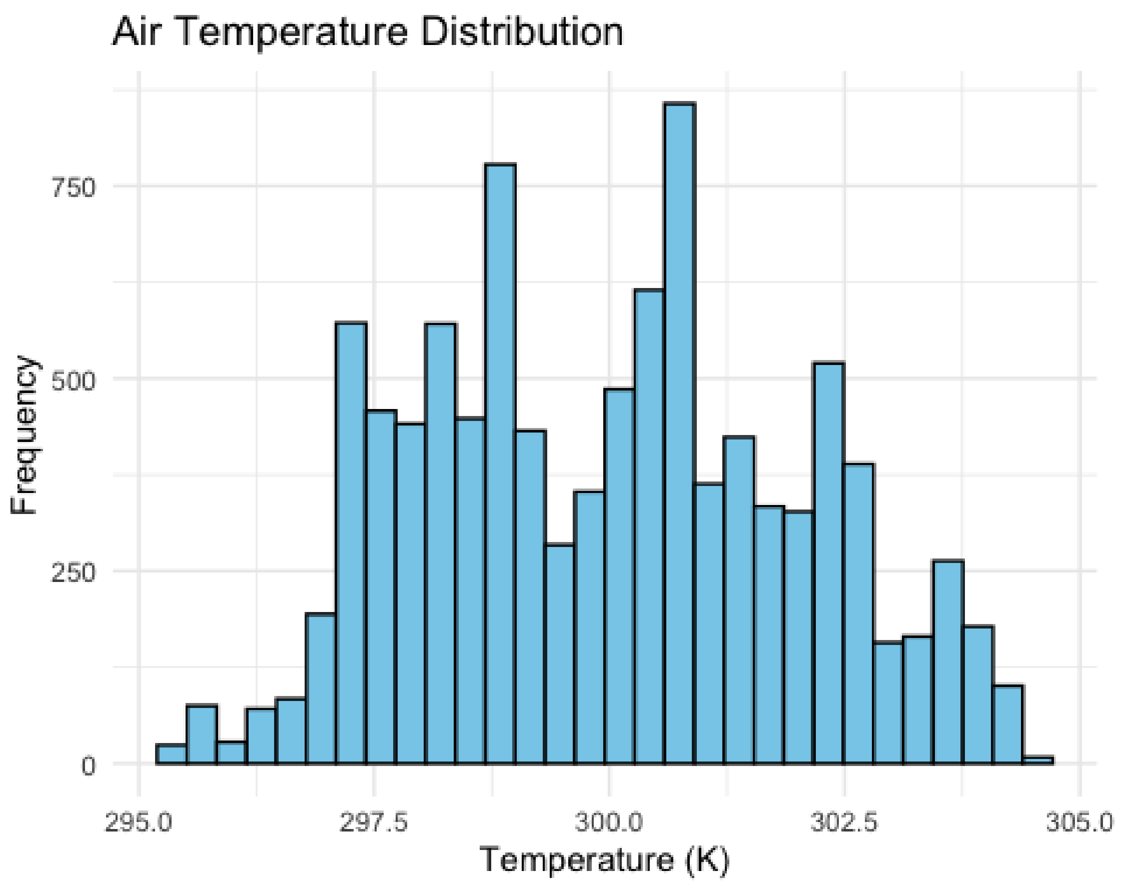
The project culminated in applying sophisticated machine learning techniques, wherein the dataset was partitioned into training and testing sets. A logistic regression model was meticulously trained to leverage the caret package, followed by a rigorous evaluation of its predictive performance using confusion matrix analysis. Moreover, the model's discriminatory prowess was meticulously assessed by creating an ROC curve and subsequent computation of the Area Under the Curve (AUC).

In essence, this project endeavors to arm manufacturing processes with predictive capabilities, thereby mitigating downtime, optimizing maintenance schedules, and enhancing overall operational efficiency. Manufacturers can proactively address potential failures by harnessing the power of data science methodologies, thus fostering a paradigm shift towards a more agile and efficient operational framework.

Actual Work

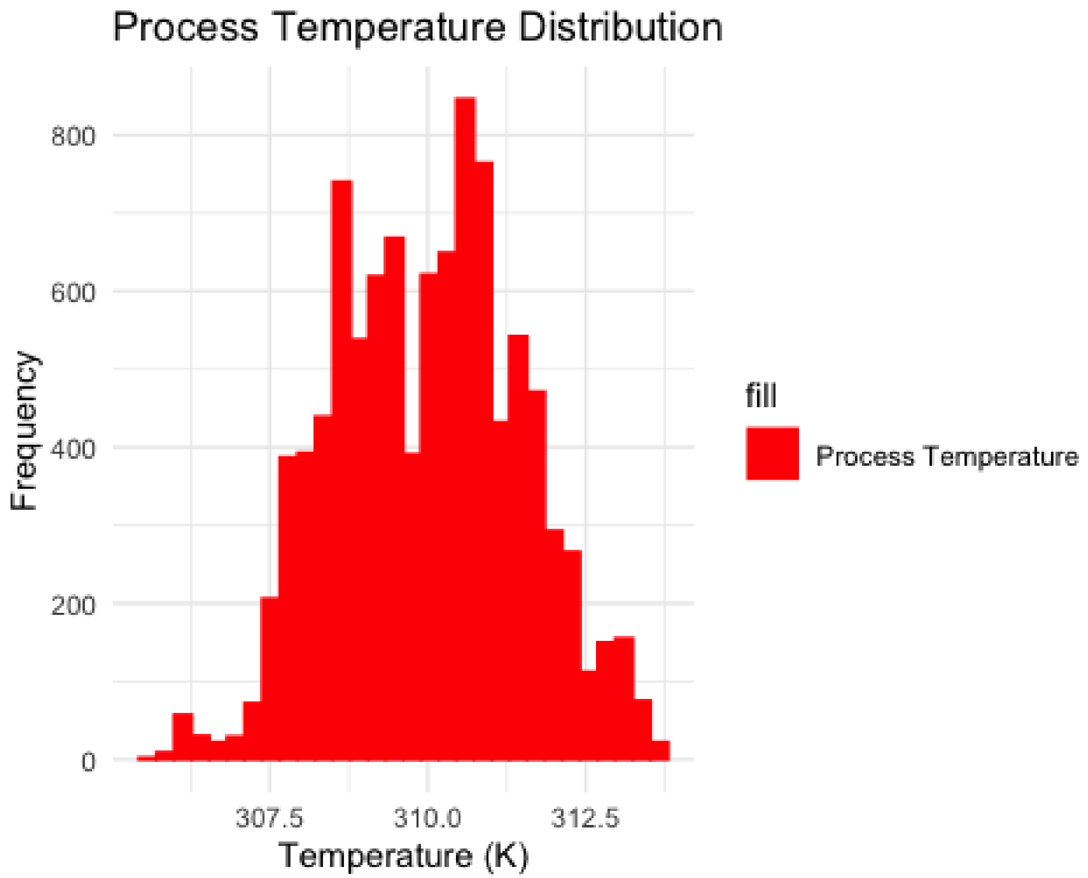
```
# import libraries ####
library (dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(readxl)
library(ggplot2)
library(corrplot)

## corrplot 0.92 loaded

library(ggcorrplot)
library(caret)

## Loading required package: lattice

library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

# import dataset ####
predictivemaintenance_data <- read_excel("/Users/shahanperera/Documents/Data
Science /Midterm Project .xlsx")

# cleaning this dataset was not necessary due to organized data as is.
# deleting some of the columns like TWF, HDF, PWF, OSF, RNF due to unnecessar
y and a lack of important data.
predictivemaintenance_data <- predictivemaintenance_data[,-10:-14]

# Descriptive Statistics ####
# using the summary function calculates all of the specific descriptive stati
```

```
stics necessary (numerical values)
summary(predictivemaintenance_data)

##       UDI         Product ID           Type           Air temperature [K]
##  Min.   :    1   Length:10000       Length:10000       Min.   :295.3
##  1st Qu.: 2501   Class :character   Class :character   1st Qu.:298.3
##  Median : 5000   Mode  :character   Mode  :character   Median :300.1
##  Mean   : 5000                                         Mean   :300.0
##  3rd Qu.: 7500                                         3rd Qu.:301.5
##  Max.   :10000                                         Max.   :304.5
##  Process temperature [K] Rotational speed [rpm]  Torque [Nm]    Tool wear
[min]
##  Min.   :305.7           Min.   :1168           Min.   : 3.80   Min.   :
0
##  1st Qu.:308.8           1st Qu.:1423           1st Qu.:33.20   1st Qu.: 5
3
##  Median :310.1           Median :1503           Median :40.10   Median :10
8
##  Mean   :310.0           Mean   :1539           Mean   :39.99   Mean   :10
8
##  3rd Qu.:311.1           3rd Qu.:1612           3rd Qu.:46.80   3rd Qu.:16
2
##  Max.   :313.8           Max.   :2886           Max.   :76.60   Max.   :25
3
##  Machine failure
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.0339
##  3rd Qu.:0.0000
##  Max.   :1.0000

# Histograms ####
hist_color <- c("lightblue", "red", "green", "orange", "purple")

ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$`Air te
mperature [K]`)) +
  geom_histogram(fill = "skyblue", color = "black", bins = 30) +
  labs(title = "Air Temperature Distribution", x = "Temperature (K)", y = "Fr
equency") +
  theme_minimal() +
  scale_fill_manual(values = hist_color[1])

## Warning: Use of `` predictivemaintenance_data$`Air temperature [K]` `` is
discouraged.
## ℹ Use `Air temperature [K]` instead.
```

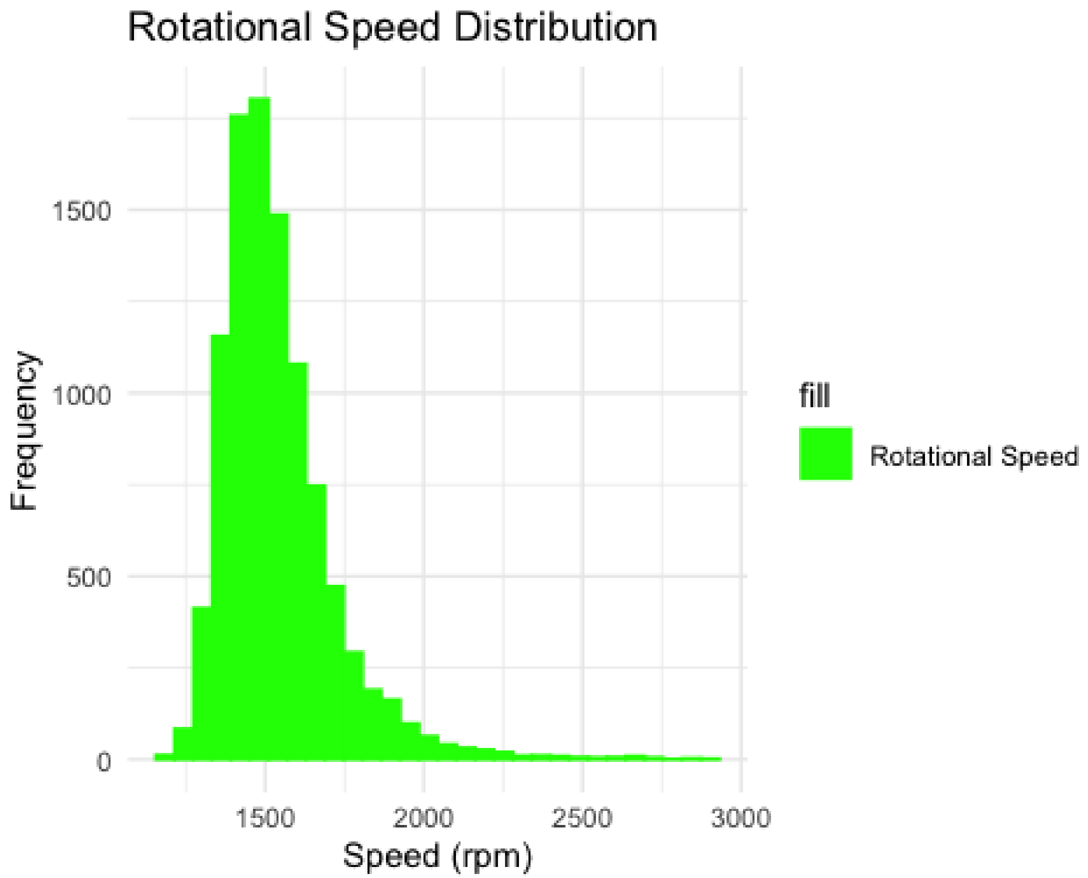


```
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$`Proces
s temperature [K]`, fill = "Process Temperature")) +
  geom_histogram(color = "red", bins = 30) +
  labs(title = "Process Temperature Distribution", x = "Temperature (K)", y =
"Frequency") +
  theme_minimal() +
  scale_fill_manual(values = hist_color[2])
```

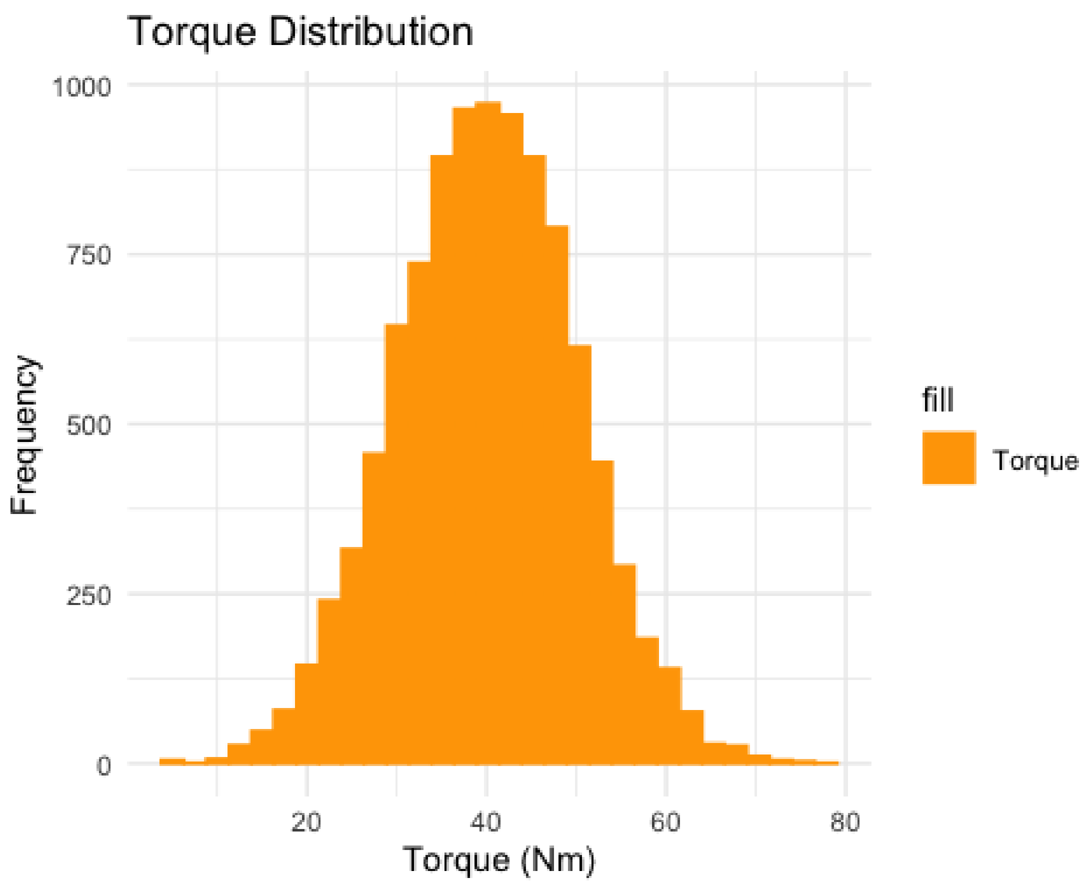
```
## Warning: Use of `` predictivemaintenance_data$`Process temperature [K]` ``
is
## discouraged.
## ℹ Use `Process temperature [K]` instead.
```



```
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$`Rotati
onal speed [rpm]`, fill = "Rotational Speed")) +
  geom_histogram(color = "green", bins = 30) +
  labs(title = "Rotational Speed Distribution", x = "Speed (rpm)", y = "Frequ
ency") +
  theme_minimal() +
  scale_fill_manual(values = hist_color[3])

## Warning: Use of `` predictivemaintenance_data$`Rotational speed [rpm]` ``
is
## discouraged.
## ℹ Use `Rotational speed [rpm]` instead.
```

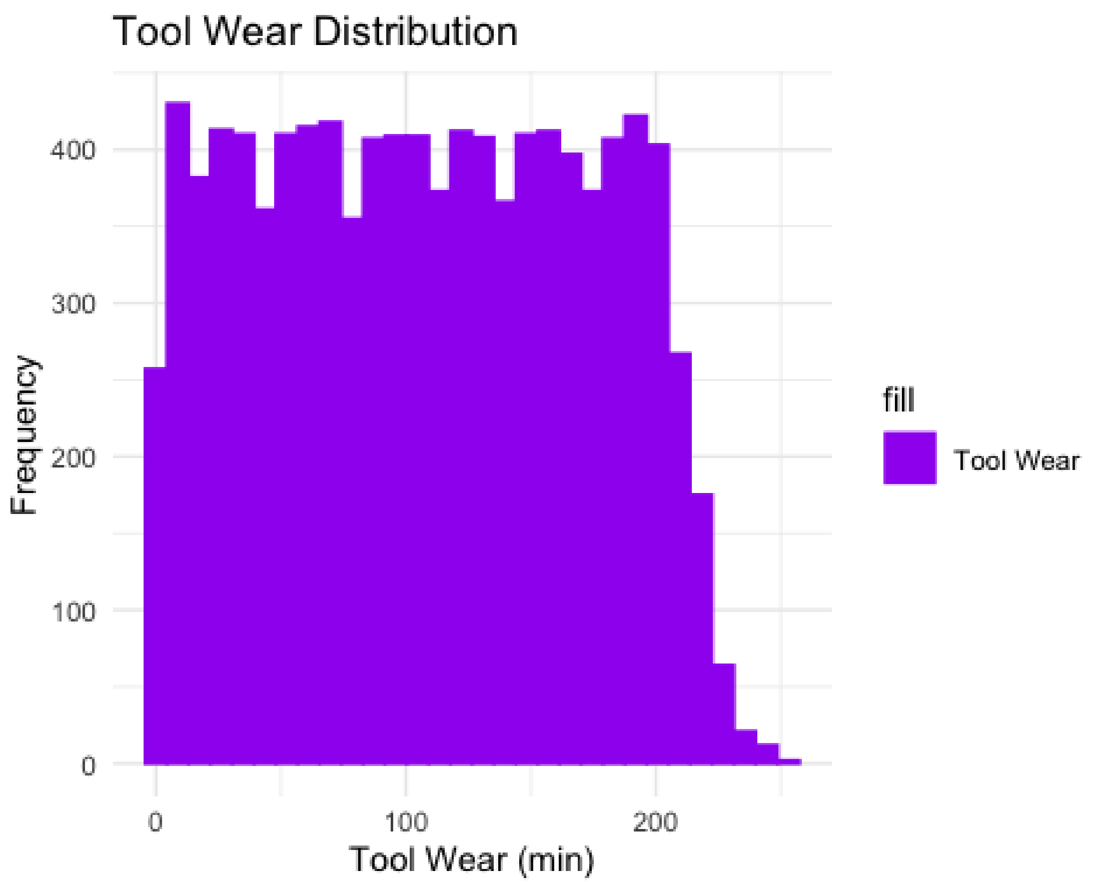


```
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$`Torque
[Nm]`, fill = "Torque")) +
  geom_histogram(color = "orange", bins = 30) +
  labs(title = "Torque Distribution", x = "Torque (Nm)", y = "Frequency") +
  theme_minimal() +
  scale_fill_manual(values = hist_color[4])

## Warning: Use of `` predictivemaintenance_data$`Torque [Nm]` `` is discoura
ged.
## ℹ Use `Torque [Nm]` instead.
```

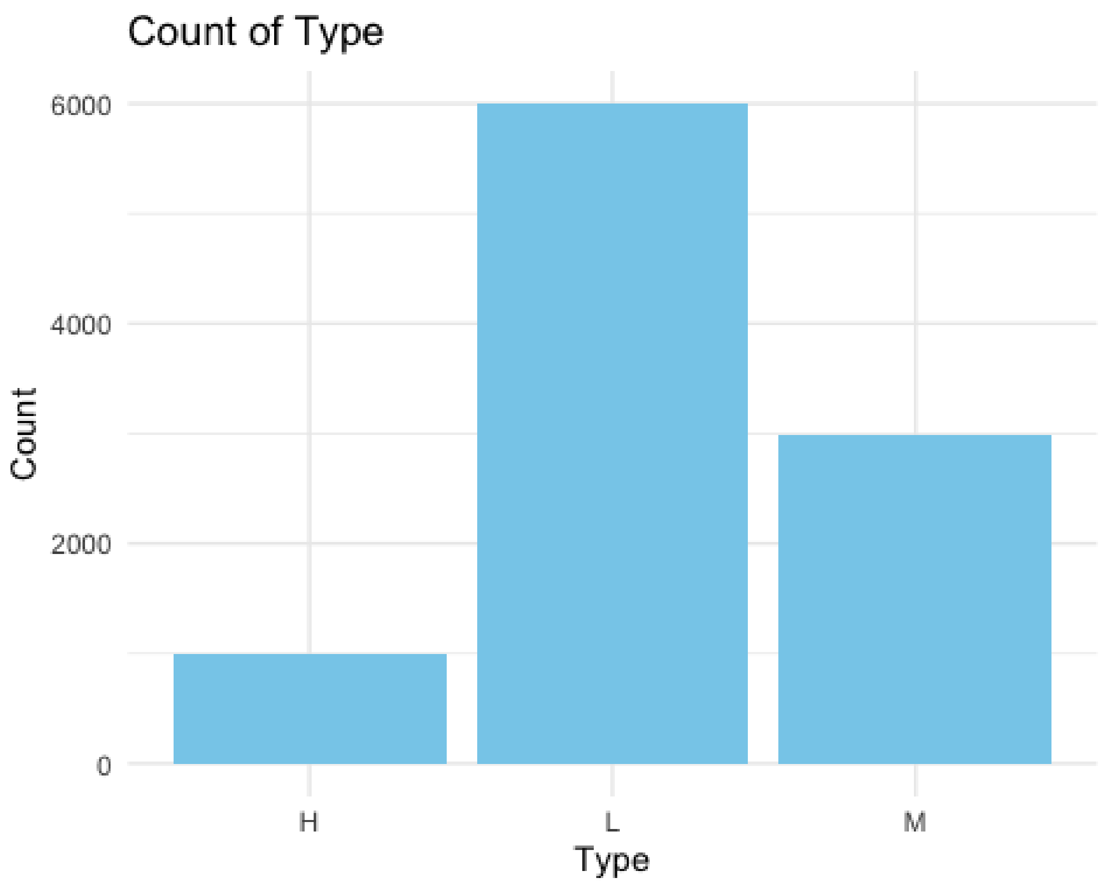


```
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$`Tool w
ear [min]`, fill = "Tool Wear")) +
  geom_histogram(color = "purple", bins = 30) +
  labs(title = "Tool Wear Distribution", x = "Tool Wear (min)", y = "Frequenc
y") +
  theme_minimal() +
  scale_fill_manual(values = hist_color[5])
```

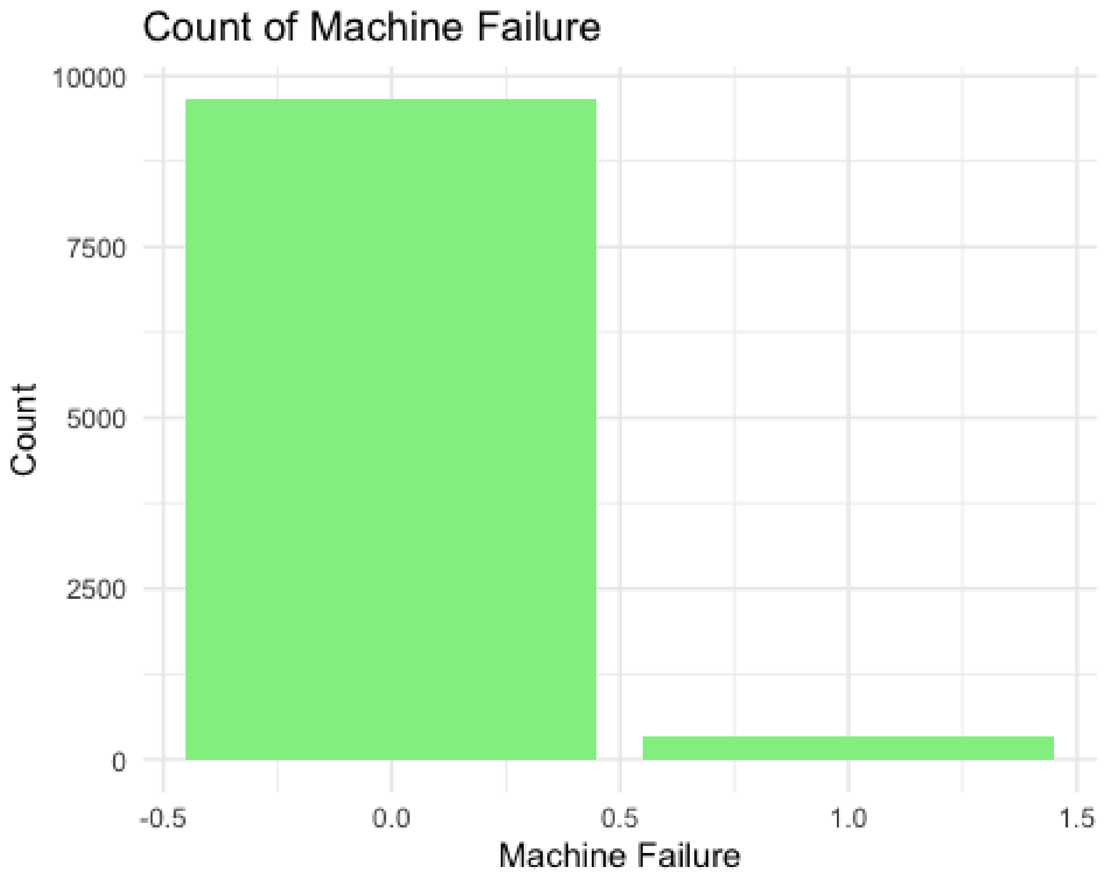
```
## Warning: Use of `` predictivemaintenance_data$`Tool wear [min]` `` is disc
ouraged.
## ℹ Use `Tool wear [min]` instead.
```



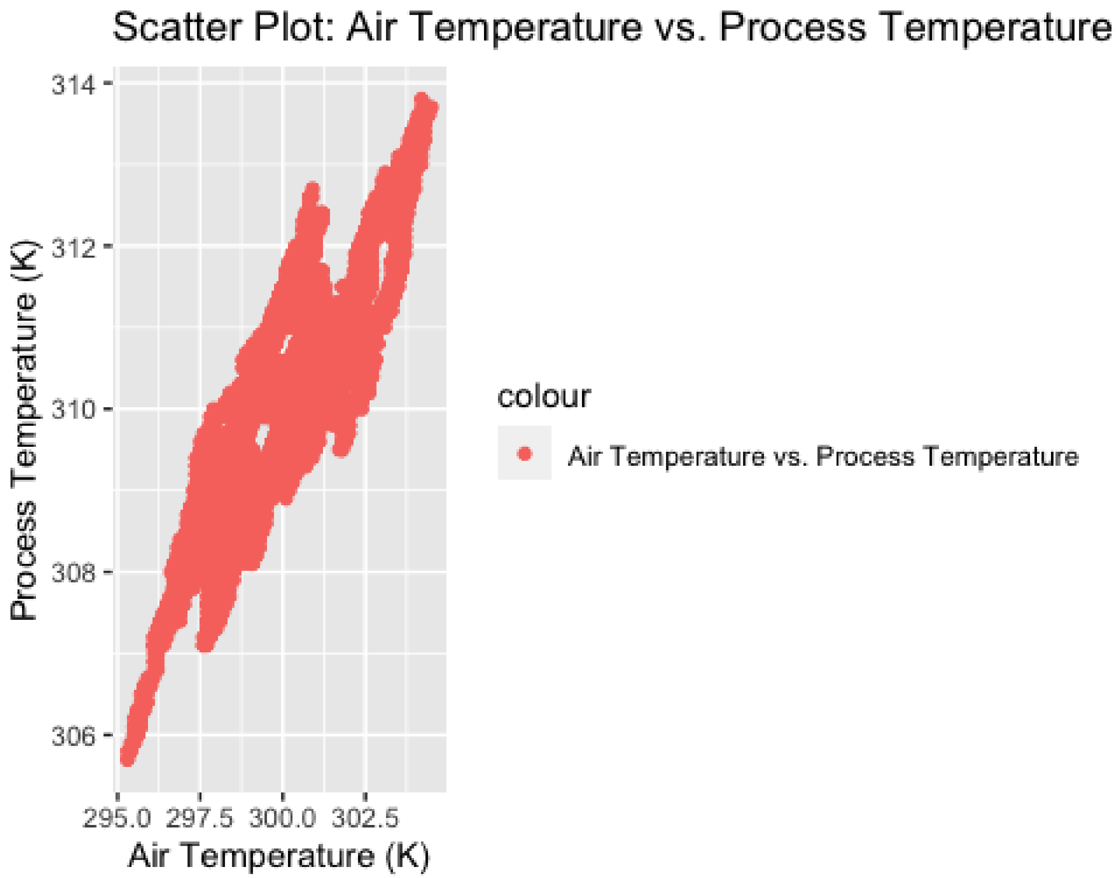
```
# Bar Plots ####

ggplot(predictivemaintenance_data, aes(x = Type)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Count of Type", x = "Type", y = "Count") +
  theme_minimal()
```



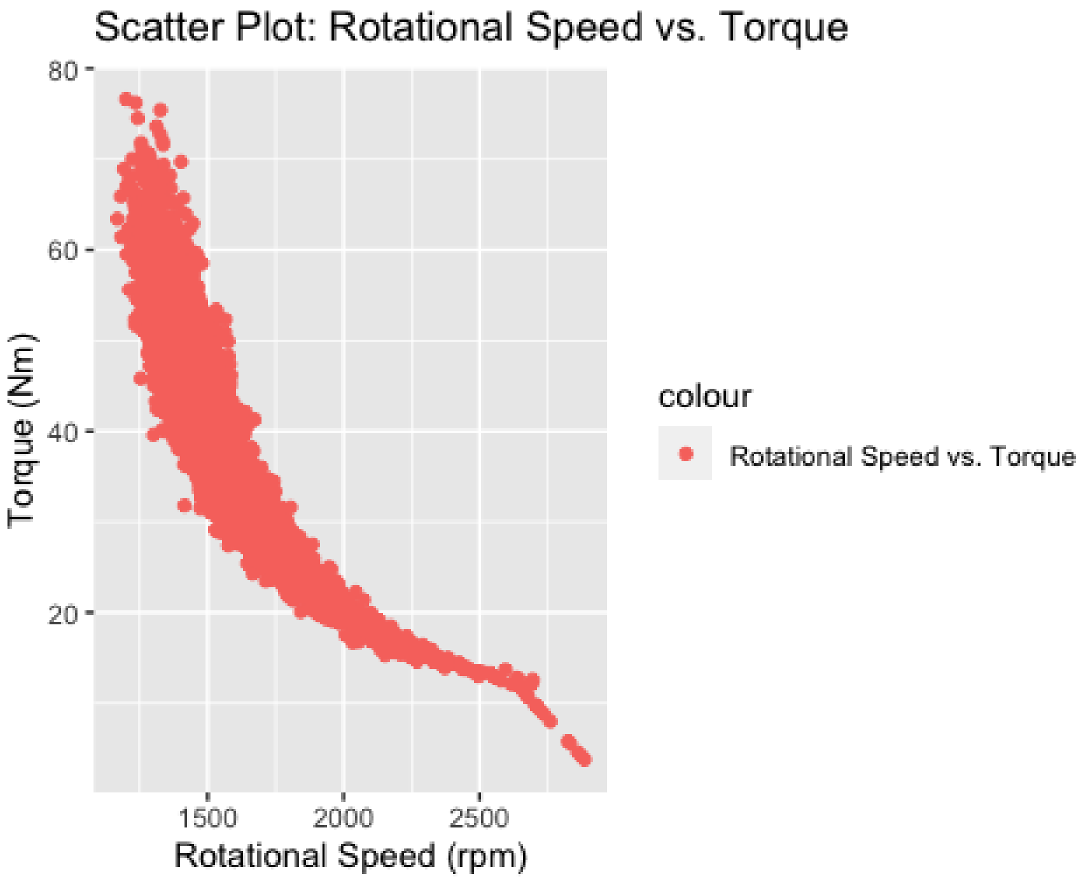
```
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$`Machin
e failure`)) +
  geom_bar(fill = "lightgreen") +
  labs(title = "Count of Machine Failure", x = "Machine Failure", y = "Count"
) +
  theme_minimal()

## Warning: Use of `` predictivemaintenance_data$`Machine failure` `` is disc
ouraged.
## ℹ Use `Machine failure` instead.
```



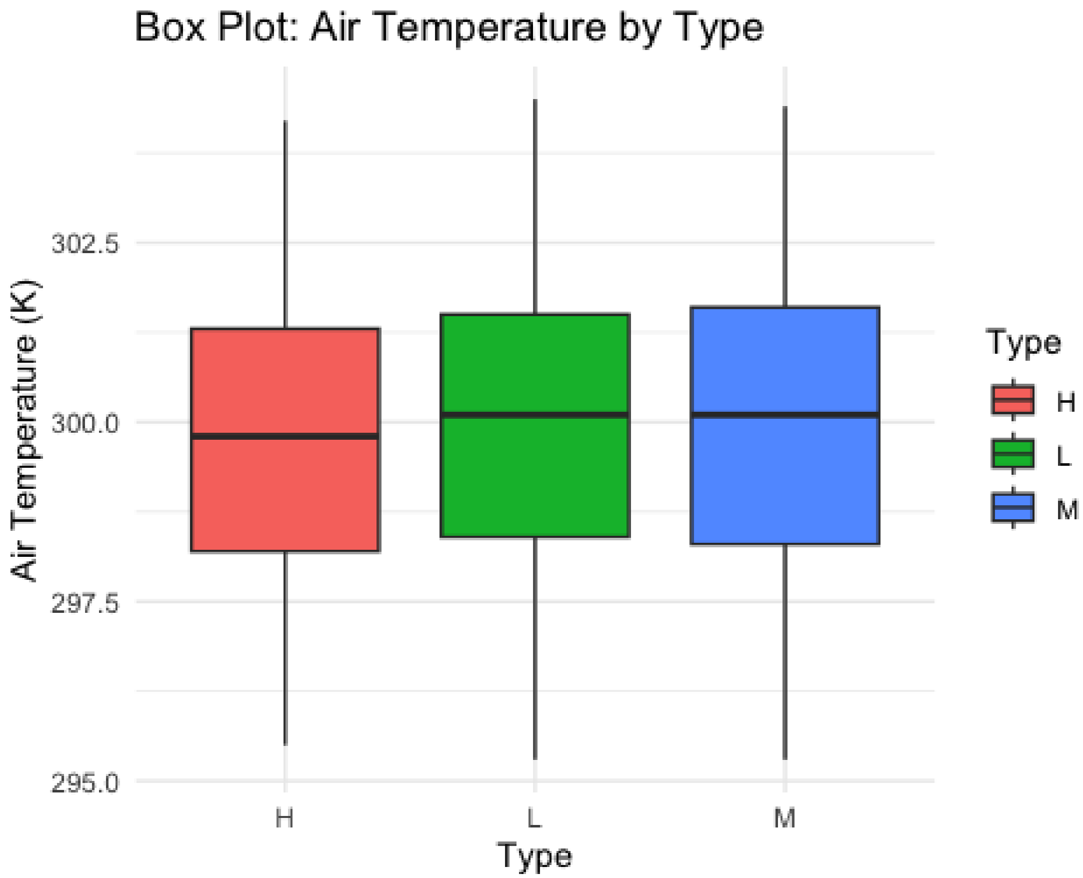
```r
# Scatter Plots ####

# Scatter plot: Air Temperature vs. Process Temperature
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$`Air te
mperature [K]`, y = predictivemaintenance_data$`Process temperature [K]`, col
or = "Air Temperature vs. Process Temperature")) +
  geom_point() +
  labs(x = "Air Temperature (K)", y = "Process Temperature (K)") +
  ggtitle("Scatter Plot: Air Temperature vs. Process Temperature")

## Warning: Use of `` predictivemaintenance_data$`Air temperature [K]` `` is
discouraged.
## ℹ Use `Air temperature [K]` instead.

## Warning: Use of `` predictivemaintenance_data$`Process temperature [K]` ``
is
## discouraged.
## ℹ Use `Process temperature [K]` instead.
```



```
# Scatter plot: Rotational Speed vs. Torque
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$`Rotati
onal speed [rpm]`, y = predictivemaintenance_data$`Torque [Nm]`, color = "Rot
ational Speed vs. Torque")) +
  geom_point() +
  labs(x = "Rotational Speed (rpm)", y = "Torque (Nm)") +
  ggtitle("Scatter Plot: Rotational Speed vs. Torque")

## Warning: Use of `` predictivemaintenance_data$`Rotational speed [rpm]` ``
is
## discouraged.
## ℹ Use `Rotational speed [rpm]` instead.

## Warning: Use of `` predictivemaintenance_data$`Torque [Nm]` `` is discoura
ged.
## ℹ Use `Torque [Nm]` instead.
```



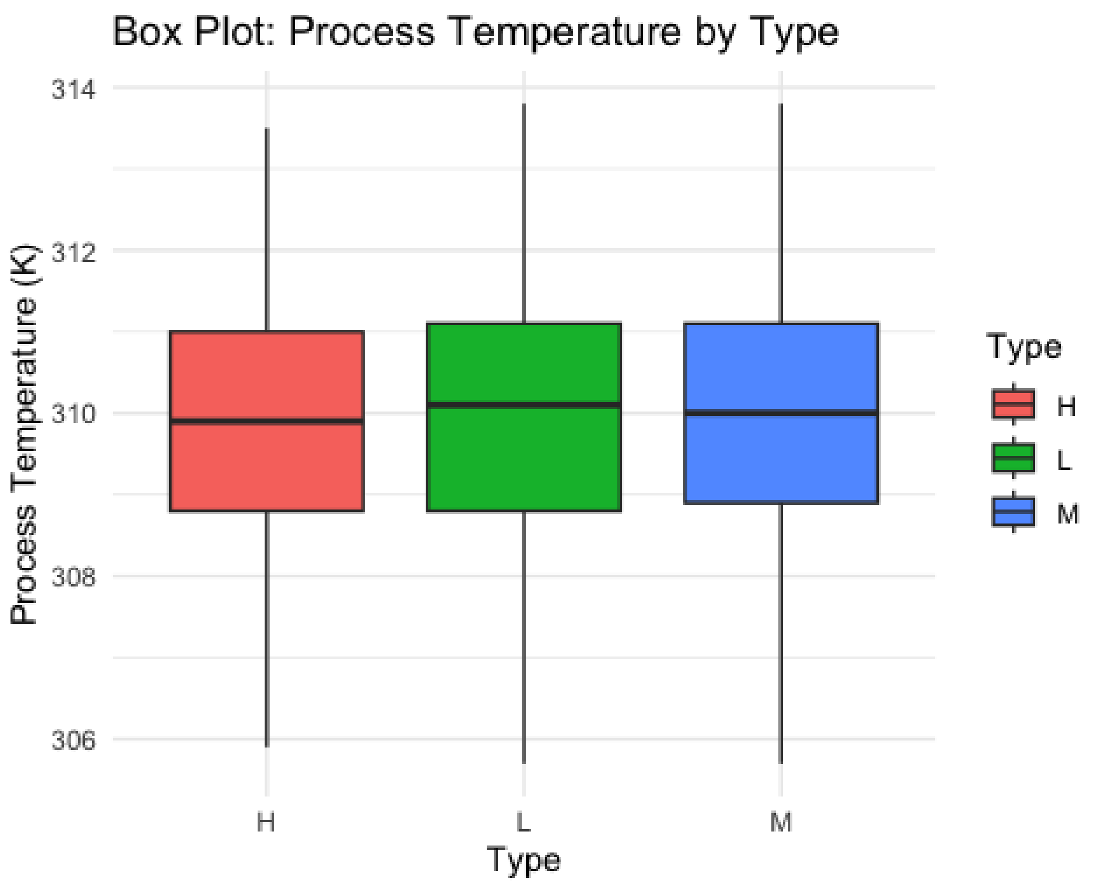
```
# Box Plots ####
# Box plot: Air Temperature by Type
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$Type, y
= predictivemaintenance_data$`Air temperature [K]`, fill = Type)) +
  geom_boxplot() +
  labs(title = "Box Plot: Air Temperature by Type", x = "Type", y = "Air Temp
erature (K)") +
  theme_minimal()

## Warning: Use of `predictivemaintenance_data$Type` is discouraged.
## ℹ Use `Type` instead.

## Warning: Use of `` predictivemaintenance_data$`Air temperature [K]` `` is
discouraged.
## ℹ Use `Air temperature [K]` instead.
```



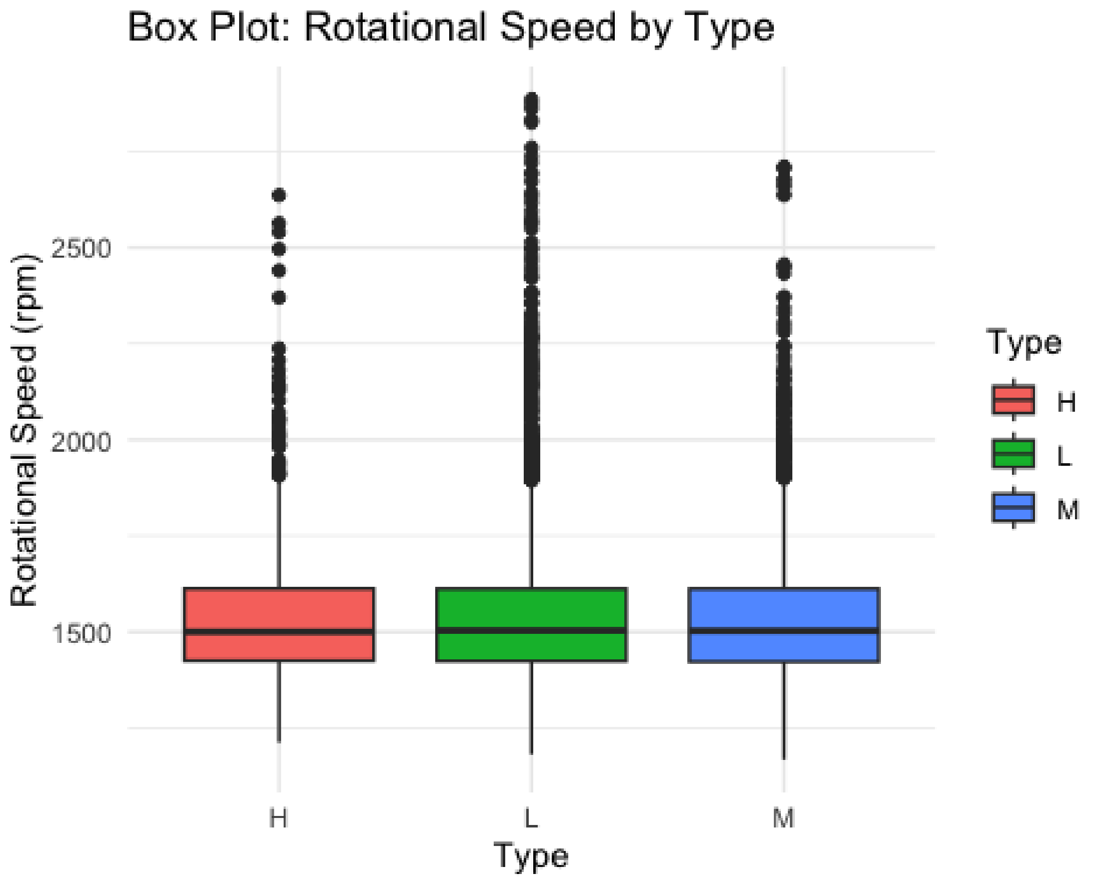
```
# Box plot: Process Temperature by Type
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$Type, y
= predictivemaintenance_data$`Process temperature [K]`, fill = Type)) +
  geom_boxplot() +
  labs(title = "Box Plot: Process Temperature by Type", x = "Type", y = "Proc
ess Temperature (K)") +
  theme_minimal()
```

```
## Warning: Use of `predictivemaintenance_data$Type` is discouraged.
## ℹ Use `Type` instead.
```

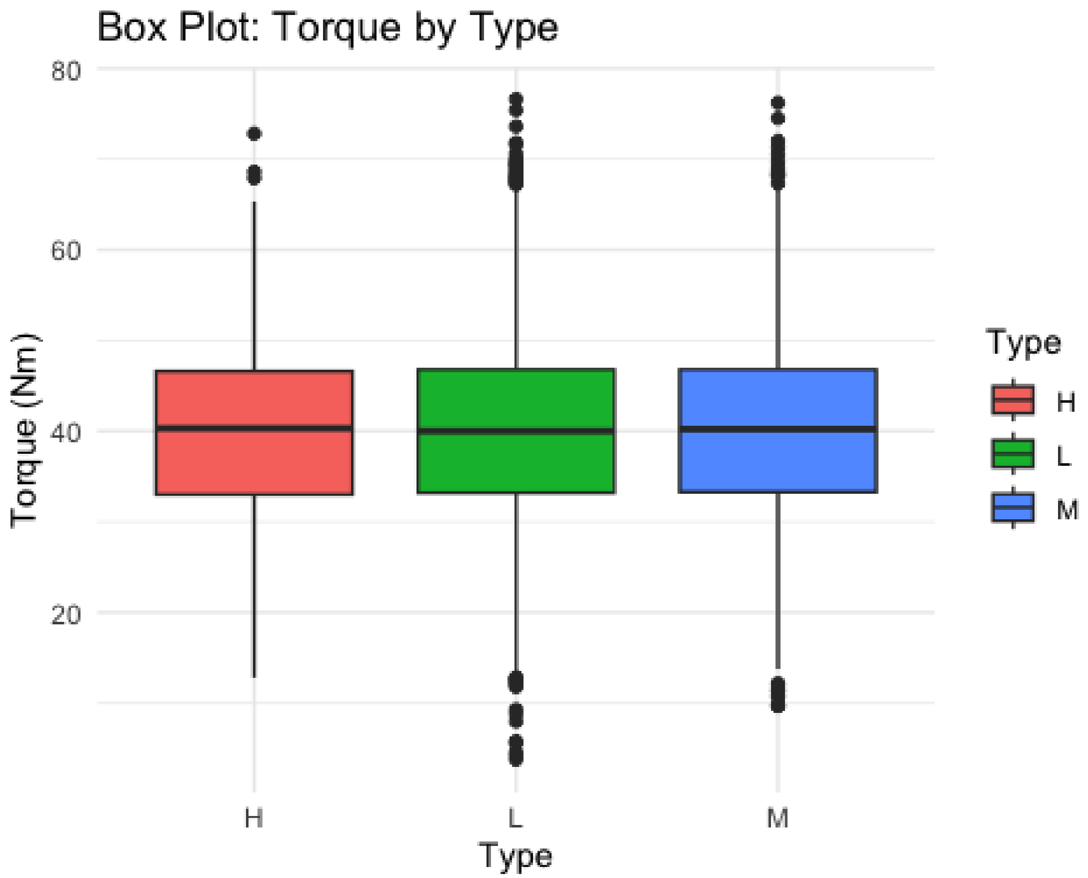
```
## Warning: Use of `` predictivemaintenance_data$`Process temperature [K]` ``
is
## discouraged.
## ℹ Use `Process temperature [K]` instead.
```



```
# Box plot: Rotational Speed by Type
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$Type, y
= predictivemaintenance_data$`Rotational speed [rpm]`, fill = Type)) +
  geom_boxplot() +
  labs(title = "Box Plot: Rotational Speed by Type", x = "Type", y = "Rotatio
nal Speed (rpm)") +
  theme_minimal()

## Warning: Use of `predictivemaintenance_data$Type` is discouraged.
## ℹ Use `Type` instead.

## Warning: Use of `` predictivemaintenance_data$`Rotational speed [rpm]` ``
is
## discouraged.
## ℹ Use `Rotational speed [rpm]` instead.
```

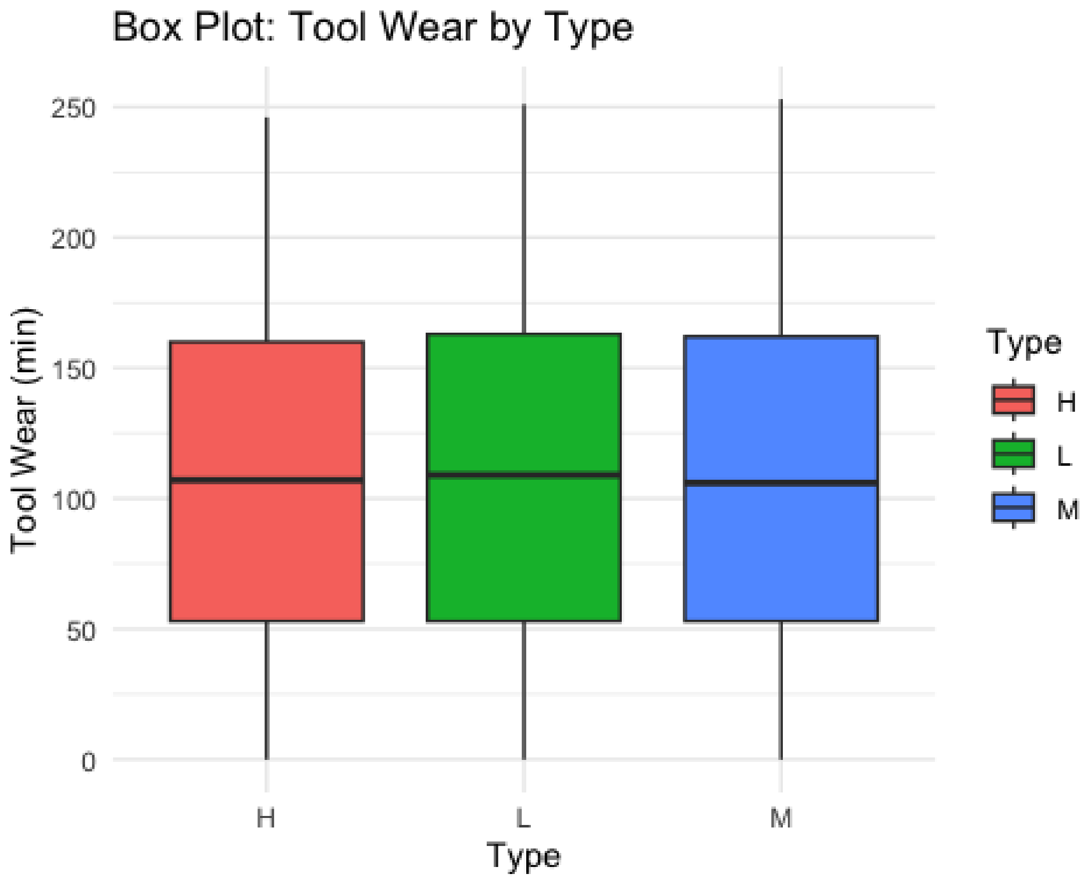


```
# Box plot: Torque by Type
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$Type, y
= predictivemaintenance_data$`Torque [Nm]`, fill = Type)) +
  geom_boxplot() +
  labs(title = "Box Plot: Torque by Type", x = "Type", y = "Torque (Nm)") +
  theme_minimal()

## Warning: Use of `predictivemaintenance_data$Type` is discouraged.
## ℹ Use `Type` instead.

## Warning: Use of `` predictivemaintenance_data$`Torque [Nm]` `` is discoura
ged.
## ℹ Use `Torque [Nm]` instead.
```

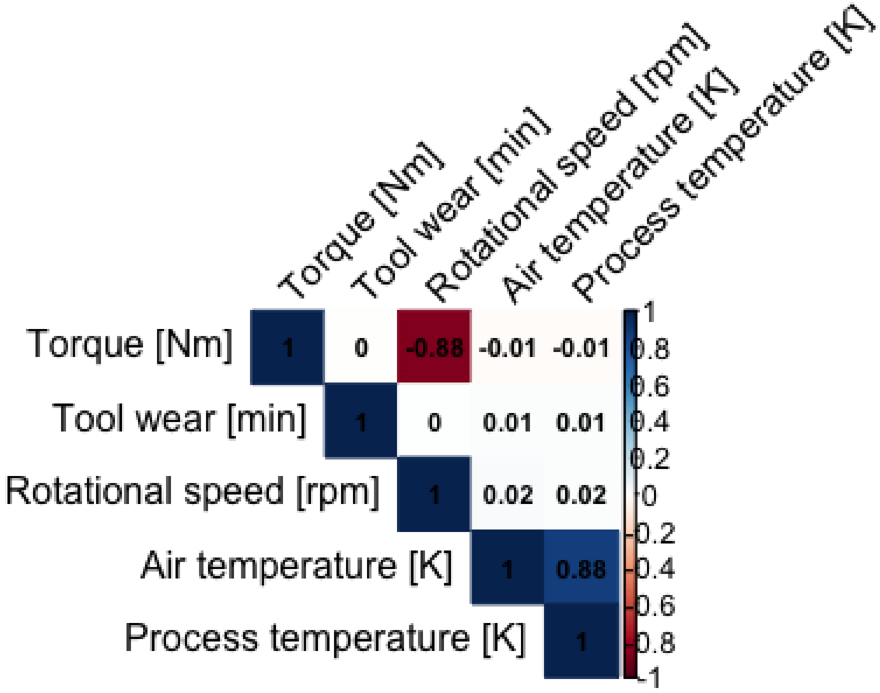


```
# Box plot: Tool Wear by Type
ggplot(predictivemaintenance_data, aes(x = predictivemaintenance_data$Type, y
= predictivemaintenance_data$`Tool wear [min]`, fill = Type)) +
  geom_boxplot() +
  labs(title = "Box Plot: Tool Wear by Type", x = "Type", y = "Tool Wear (min
)") +
  theme_minimal()
```

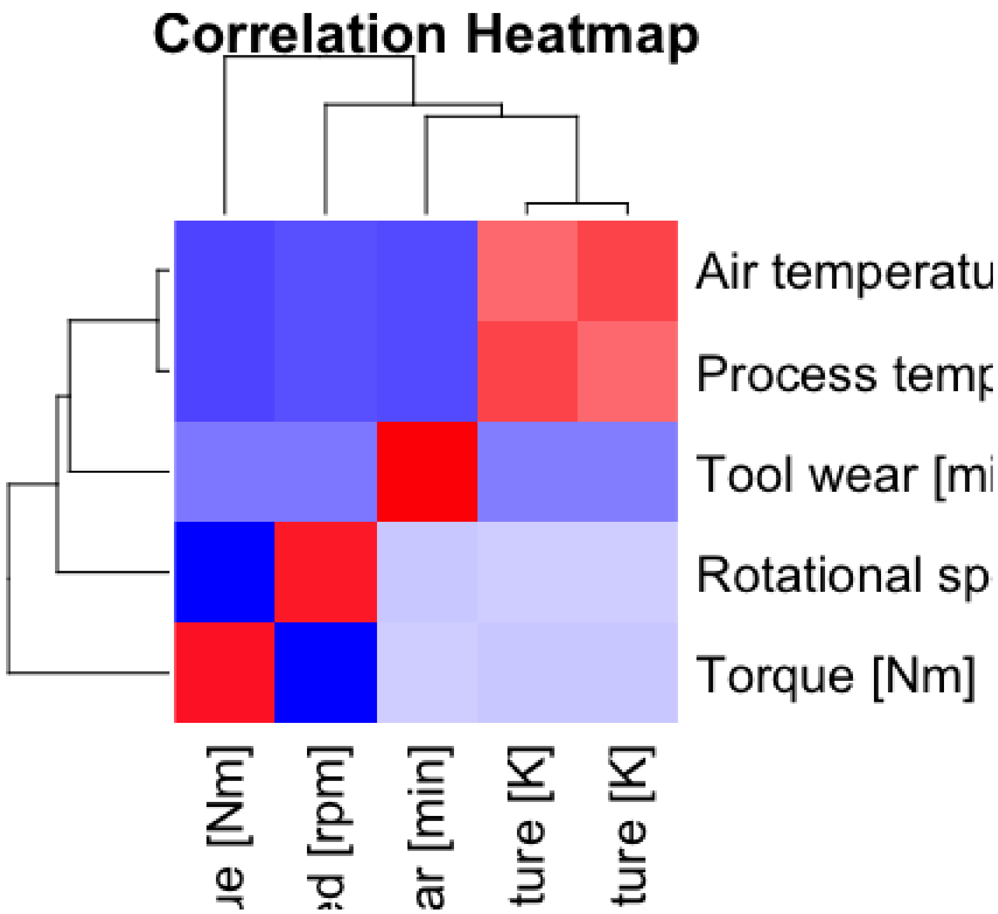
```
## Warning: Use of `predictivemaintenance_data$Type` is discouraged.
## ℹ Use `Type` instead.

## Warning: Use of `` predictivemaintenance_data$`Tool wear [min]` `` is disc
ouraged.
## ℹ Use `Tool wear [min]` instead.
```



```
# Correlation Analysis and Heat Map ####

#Correlation Analysis
correlation_matrix <- cor(select(predictivemaintenance_data, -c(UDI, `Product
ID`, Type, `Machine failure`)))

corrplot(correlation_matrix, method = "color", type = "upper", order = "hclus
t",
         tl.col = "black", tl.srt = 45, addCoef.col = "black", number.cex = 0
.7)
```



```r
# Heatmap using the correlation matrix
heatmap(correlation_matrix,
        main = "Correlation Heatmap",
        col = colorRampPalette(c("blue", "white", "red"))(100))
```



```
# Machine Learning Algorithms ####
# Data Preprocessing
# Assuming 'Machine_failure' is the target variable (response)
# Convert 'Type' to a factor variable
predictivemaintenance_data$Type <- as.factor(predictivemaintenance_data$Type)

# Split data into training and testing sets (80% training, 20% testing)
set.seed(123)  # for reproducibility
trainIndex <- createDataPartition(predictivemaintenance_data$`Machine failure
`, p = 0.8,
                                  list = FALSE, times = 1)
train_data <- predictivemaintenance_data[trainIndex, ]
test_data <- predictivemaintenance_data[-trainIndex, ]

test_dataWithoutID <- test_data[,-2]
# Model Training: Logistic Regression
colnames(train_data)

## [1] "UDI"                        "Product ID"
## [3] "Type"                       "Air temperature [K]"
## [5] "Process temperature [K]" "Rotational speed [rpm]"
## [7] "Torque [Nm]"                "Tool wear [min]"
## [9] "Machine failure"
```

```
model <- train(`Machine failure` ~ Type + `Air temperature [K]` + `Process te
mperature [K]` +
                  `Rotational speed [rpm]` + `Torque [Nm]` + `Tool wear [min]`
, data = train_data, method = "glm", family = "binomial")

## Warning in train.default(x, y, weights = w, ...): You are trying to do
## regression and your outcome only has two possible values Are you trying to
do
## classification? If so, use a 2 level factor as your outcome column.

#model <- train(`Machine failure` ~ ., data = test_data, mqethod = "glm", fam
ily = "binomial")

# Model Evaluation
predictions <- predict(model, test_data)
predictions <- factor(round(predictions), levels = levels(test_data$`Machine
failure`))

# I first preprocessed the data by converting the 'Type' variable to a factor
and splitting the dataset into training and testing sets.
# Then, I trained a logistic regression model using the train() function from
the caret package.
# After training, I made predictions on the test set and evaluate the model's
performance using confusion matrix analysis.
```

Insights Gained

Several insightful findings have emerged through the meticulous analysis and discovery conducted in this project aimed at predictive maintenance in manufacturing processes.

Firstly, the exploratory data analysis (EDA) revealed intriguing patterns and distributions within the dataset. The histograms, bar plots, scatter plots, and box plots provided invaluable insights into the relationships between various parameters such as air temperature, process temperature, rotational speed, torque, and tool wear. These visualizations shed light on the distributional characteristics of these parameters across different types and machine failure modes, enabling a deeper understanding of the underlying processes.

Descriptive statistics provide a summary of the main aspects of the dataset. In this project, summary statistics such as mean, median, minimum, maximum, and quartiles were generated for the numerical variables like air temperature, process temperature, rotational speed, torque, tool wear, and machine failure. These statistics offer insights into the central tendency, dispersion, and shape of the data distribution. For instance, understanding the range of values for each parameter helps in identifying outliers or anomalies that may indicate data quality issues or exceptional events in the manufacturing process. Descriptive statistics also aid in understanding the variability and distribution of variables, which is essential for subsequent analysis and model development. For instance, knowing the distribution of tool wear or temperature variations can inform decisions about maintenance schedules or process optimization.

Linear regression allows for the exploration of relationships between predictor variables (such as air temperature, process temperature, rotational speed, torque, and tool wear) and the outcome variable (machine failure). In this example, linear regression could be used to quantify the impact of each predictor variable on the likelihood of machine failure. For example, it could reveal whether increases in tool wear or temperature are associated with higher probabilities of failure. By fitting a regression model to the data, one can identify significant predictors and estimate their effects, providing actionable insights for predictive maintenance strategies. For instance, if tool wear emerges as a significant predictor of machine failure, it suggests that monitoring and managing tool wear levels could help prevent or mitigate failures. Additionally, linear regression analysis can provide a baseline predictive model against which more complex machine learning algorithms can be compared. It serves as a fundamental building block in predictive modeling, helping to establish the predictive power of individual features before incorporating them into more sophisticated models.

Furthermore, the correlation analysis highlighted significant relationships between variables, providing crucial insights into the interdependencies among different factors influencing machine failure. For instance, correlations between parameters such as torque and rotational speed or tool wear and torque elucidated the complex interactions driving process failures. This deeper understanding of the relationships between variables can inform targeted maintenance strategies and help identify key predictors of machine failure.

Additionally, the application of machine learning techniques, particularly logistic regression, showcased promising results in predicting machine failures based on the provided features. By leveraging predictive models, manufacturers can proactively identify potential failure modes and implement preventive maintenance measures, thereby minimizing downtime and optimizing operational efficiency. Overall, the insights gained from this project have the potential to revolutionize maintenance practices in manufacturing processes, paving the way for enhanced productivity and cost savings.

Ending Paragraph:

This technical report has documented a comprehensive analysis and discovery project aimed at predictive maintenance in manufacturing processes. Through meticulous data collection, cleaning, exploratory analysis, visualization, and regression techniques, valuable insights have been gleaned into the factors influencing machine failures and the relationships between various parameters.

The project has demonstrated the potential of data science methodologies in empowering manufacturers with predictive capabilities, enabling proactive maintenance strategies to minimize downtime, optimize maintenance schedules, and enhance overall operational efficiency. By leveraging these insights, manufacturers can transition from reactive to proactive maintenance practices, thereby reducing costs and improving productivity.

Moving forward, several next steps can be considered to further enhance the predictive capabilities and applicability of the models developed in this project. Firstly, the incorporation of additional features or data sources could enrich the predictive models, providing a more comprehensive understanding of the factors driving machine failures. Secondly, exploring advanced machine learning algorithms, such as random forests or neural networks, could potentially improve predictive performance. Additionally, real-time monitoring and implementation of predictive maintenance strategies in a production environment could validate the effectiveness of the developed models and provide valuable feedback for refinement. Continuous monitoring and updating of the models based on new data and insights will be essential to ensure their relevance and effectiveness over time.

In summary, this project lays the groundwork for transformative advancements in predictive maintenance practices, offering manufacturers the opportunity to address potential failures and optimize operational performance proactively. With ongoing research and refinement, the insights gained from this project have the potential to revolutionize maintenance strategies and drive significant improvements in manufacturing efficiency and productivity.

References

AI4I 2020 Predictive Maintenance Dataset. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5HS5C.