PYTHON EDA / Descriptive Statistics

Stakeholder:

One of the leading vehicle fuel pump manufacturers. The pumps they manufacture are used to take fuel as input and push fuel as output at a high velocity. More the velocity, more is the speed at which vehicle will move.

Business Insight:

Unplanned machine downtime is causing productivity loss in the machines that make the pumps. We need to reduce unscheduled machine downtime and maintenance costs. Reduce unplanned downtime by at least 10% and save at least $1 million in costs.

## Introduction:

Machine downtime refers to the period during which a piece of machinery or equipment is not operational or productive. Common reasons for downtime include: Mechanical Failures, Maintenance, Repairs, Changeovers and Setup, Material Shortages, Power outages, Environmental Factors or other unforeseen issues.

Exploratory Data Analysis (EDA) is a critical step in understanding machine downtime data. It involves the process of summarizing and visualizing data to uncover patterns, anomalies, and insights. When dealing with machine downtime data, we can formulate various queries to perform EDA effectively.

## Overall design strategy

1. Data Collection: Collect historical data on machine downtime, including timestamps, reasons for downtime, factors affecting machines, maintenance records, and any other relevant information.
2. Data Cleaning: Clean and preprocess the data to handle missing values, outliers, and inconsistencies.
3. Exploratory Data Analysis (EDA): Perform comprehensive EDA to understand the characteristics of machine downtime and its relationship with various factors.
4. Descriptive Statistics: Calculate key summary statistics and create visualizations to communicate insights effectively.
5. Identify Patterns: Uncover patterns, trends, and correlations within the data that might be contributing to unplanned machine downtime.
6. Business Recommendations: Based on the analysis, provide recommendations on how to reduce unplanned machine downtime while considering the constraint of minimizing maintenance costs.

## Data Overview

Our dataset is a comma separated file – **Machine Downtime.csv** which is imported to mySQL.

Then data is pulled into python file using mySQL connector.

Performing Exploratory Data Analysis (EDA) on machine downtime data can provide valuable insights.

1. **Data Overview:**

* What is the size of the dataset (number of records and features)
* What are the data types of each feature (numeric, categorical, datetime)?
* Are there any missing values in the dataset?

1. **Descriptive Statistics:**
   * Calculate summary statistics for numeric variables (mean, median, standard deviation, etc.).
   * For categorical variables, find the unique categories and their counts.
2. **Time Series Analysis:**
   * Plot a time series chart of machine downtime over time to identify trends or seasonality.
3. **Distribution of Downtime:**
   * Create a histogram or box plot to visualize the distribution of downtime durations.
   * Identify outliers or extreme values.
4. **Categorical Analysis:**
   * Group downtime data by machine type or category and calculate downtime statistics for each group.
5. **Correlation Analysis:**
   * Create a correlation heatmap to visualize relationships.

## EDA before pre-processed data

* **What is the size of the dataset (number of records and features)?**

Ans : There are total 2500 rows in the dataset with 16 columns or features.

* **What are the data types of each feature (numeric, categorical, datetime)?**

Ans: Column names with datatypes:

1. **Date (Date format)** : This column represents the timestamp when an event or observation occurred. It will be crucial for time-based analysis to identify patterns in downtime over time.
2. **Machine\_ID(text)** : Categorical Machine Identifier. This column contains identifiers for different machines involved in the manufacturing process. It will be essential for identifying which machines are most affected by downtime.
3. **Assembly\_Line\_No(text)** : Categorical Assembly Line Identifier. This column indicates the assembly line number where the machine is located. It can help in understanding if specific assembly lines experience more downtime than others.
4. **Hydraulic\_Pressure(bar) (double)** : Numeric pressure measurement. This column indicates the assembly line number where the machine is located. It can help in understanding if specific assembly lines experience more downtime than others.
5. **Coolant\_Pressure(bar)(double)** : Numeric pressure measurement. Similar to hydraulic pressure, this column contains data on coolant pressure, which might also be linked to machine downtime.
6. **Air\_System\_Pressure(bar)(double)** : Numeric pressure measurement . This column represents air system pressure measurements. It's essential to examine how variations in air system pressure relate to downtime incidents.
7. **Coolant\_Temperature(double)** : Numeric temperature measurement. This column contains data on coolant temperature. Analyzing temperature fluctuations can help determine if overheating or cooling issues are associated with downtime.
8. **Hydraulic\_Oil\_Temperature** **(Â°C) (double)** : Numeric temperature measurement. This column provides information about hydraulic oil temperature. It's crucial to investigate how oil temperature affects machine performance.
9. **Spindle\_Bearing\_Temperature** **(Â°C) (double)** : Numeric temperature measurement. This column contains data on spindle bearing temperatures. Examining bearing temperatures can help detect potential wear and tear issues that lead to downtime.
10. **Spindle\_Vibration(double)** **(Âµm)** : Numeric vibration measurement. This column represents spindle vibration measurements. High spindle vibrations could indicate problems with the machine's components.
11. **Tool\_Vibration** **(Âµm) (double)** : Numeric vibration measurement. Similar to spindle vibration, this column contains data on tool vibration. Excessive tool vibration can affect product quality and lead to downtime.
12. **Spindle\_Speed** **(RPM) (int)** : Numerical Rotational speed. This column provides information on spindle speed. Variations in spindle speed could impact manufacturing processes and downtime.
13. **Voltage** **(volts) (int)** : Numeric Electrical measurement. This column contains voltage measurements. Fluctuations in voltage can affect machine operation and productivity.
14. **Torque(Nm)** **(double)**: Numeric Mechanical measurement. This column represents torque measurements. Changes in torque may be indicative of mechanical issues that lead to downtime.
15. **Cutting(kN)** **(double)**: Numeric Process force parameter. This column likely represents a parameter related to the cutting process. Analyzing this variable can help understand its relationship with downtime.
16. **Downtime(text)** : Categorical target variable. It represents the category of unplanned machine downtime. Analyzing downtime will be the primary focus of EDA to identify patterns, causes, and potential solutions.

* **Are there any missing values in the dataset?**

Ans: Date, Machine\_ID, Assembly\_Line\_No, Downtime columns have no null values. Hydraulic\_Pressure has 10 null values,

Coolant\_Pressure has 19 null values

Air\_System\_Pressure has 17 null values

Coolant\_Temperature has 12 null values

Hydraulic\_Oil\_Temperature has 16 null values

Spindle\_Bearing\_Temperature has 7 null values

Spindle\_Vibration has 11 null values

Tool\_Vibration has 11 null values

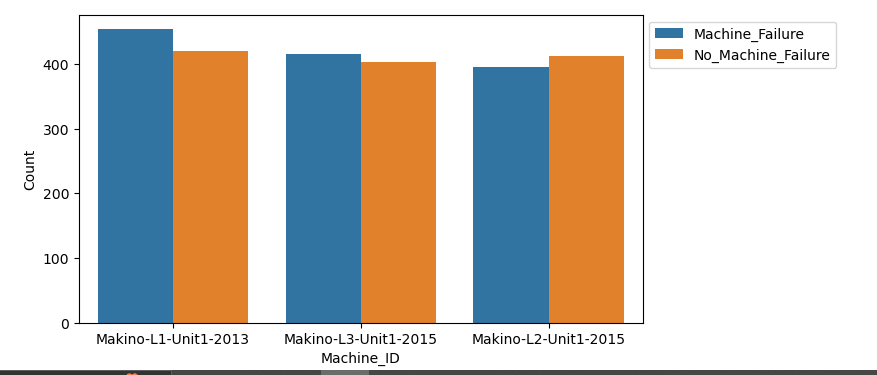
Spindle\_Speed has 13 null values

Voltage has 6 null values

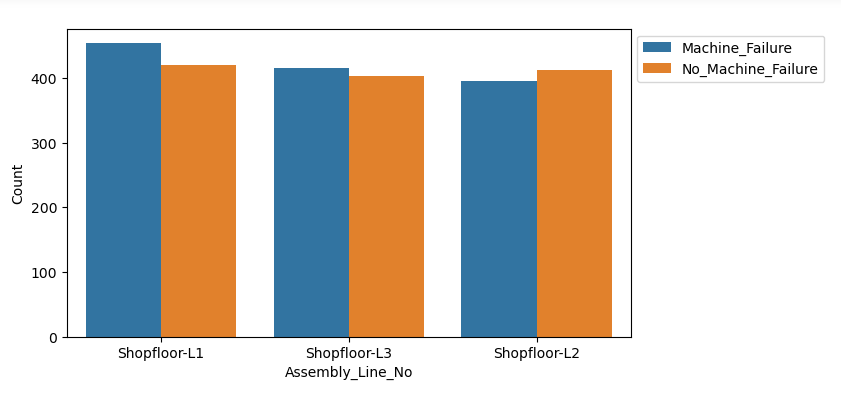
Torque has 23 null values

Cutting has 7 null values

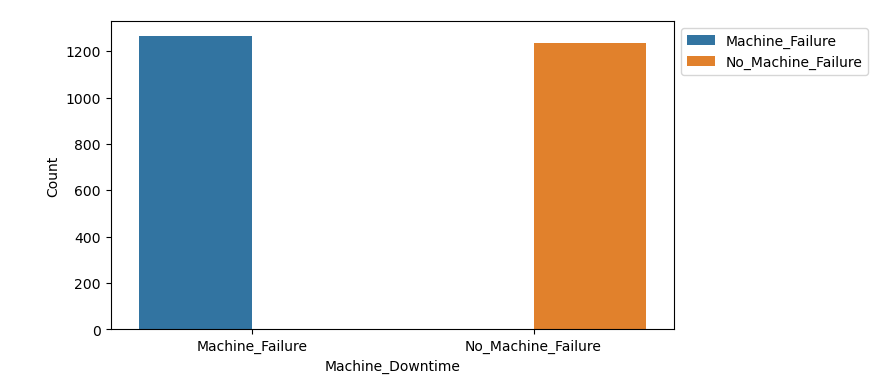
* **For categorical variables, find the unique categories and their counts.**

 Ans : 1. There are 3 types of machines.

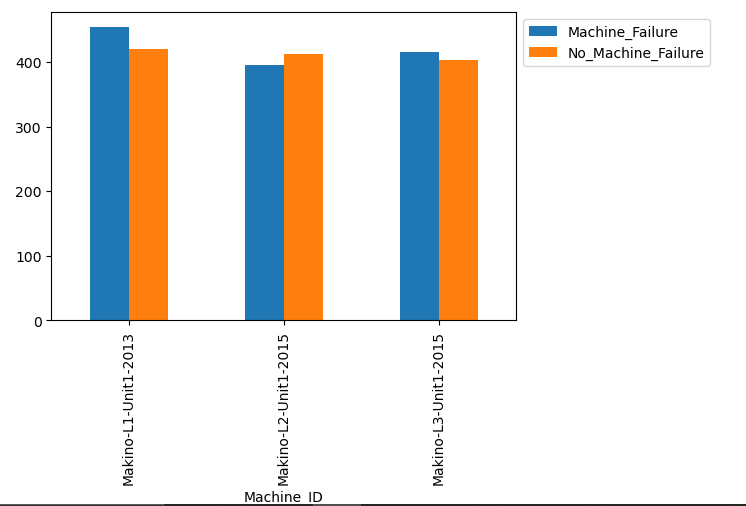
1. There are 3 assembly line numbers



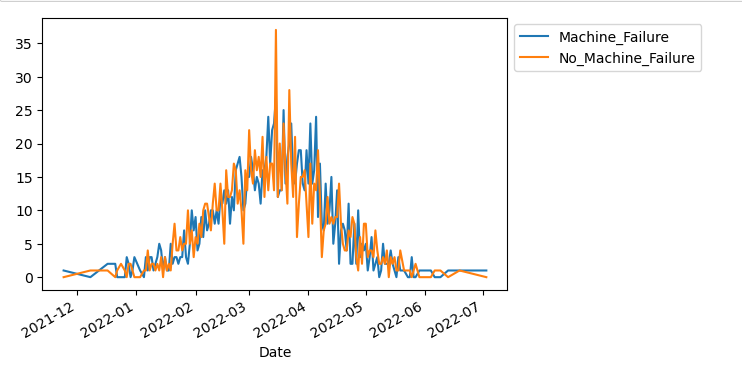
1. There are 2 categories with Machine\_Failure and No\_Machine\_Failure categories

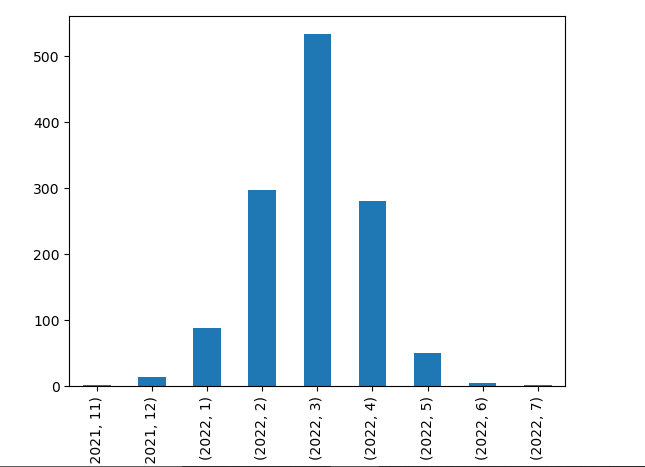


1. Count of Machine Failures and No Machine Failures

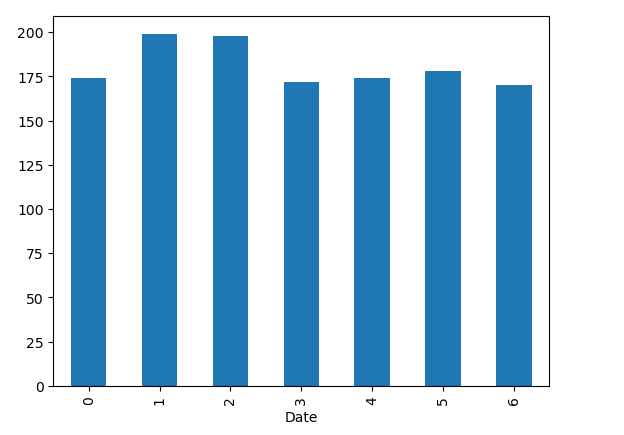


* **Find a time series of machine downtime over time to identify trends or seasonality.**

Ans: 1. Data is collected for 2 years 2021, 2022.

 2. March, February, April are the months machines most down.

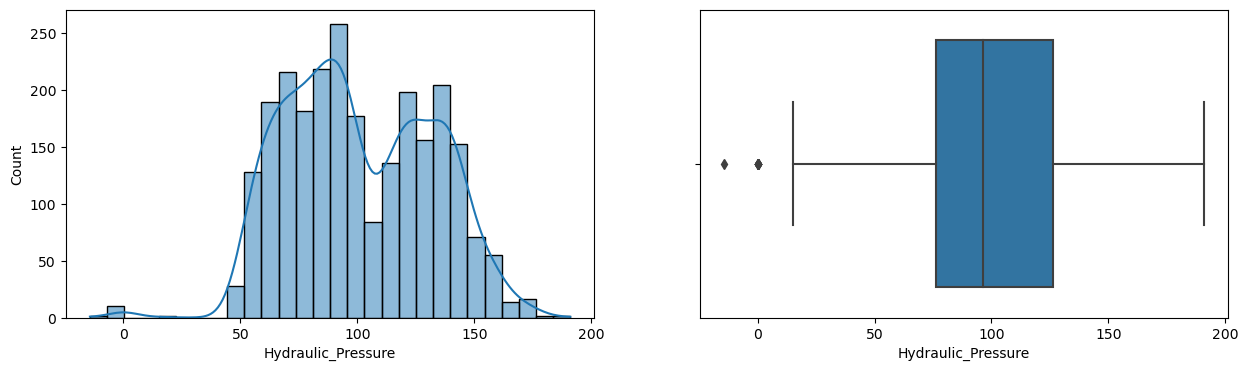
3. Tuesday and Wednesday are the days most machines are down.

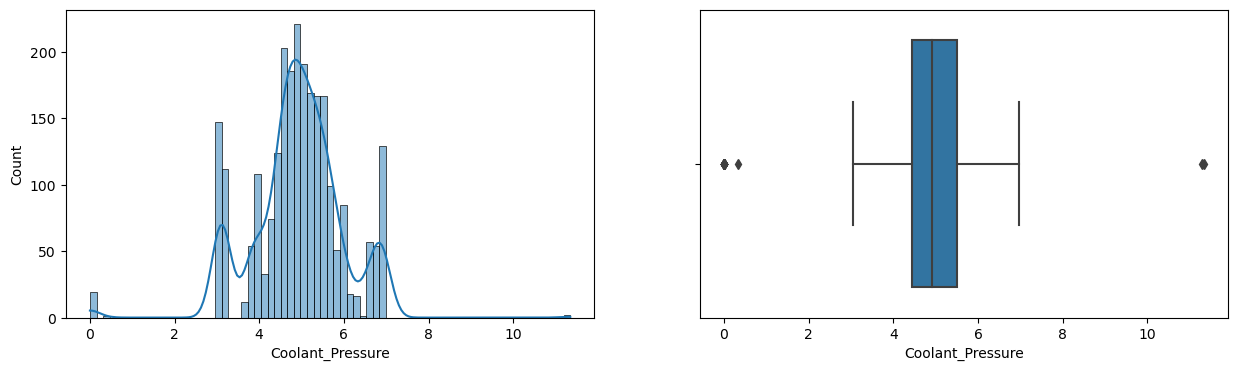


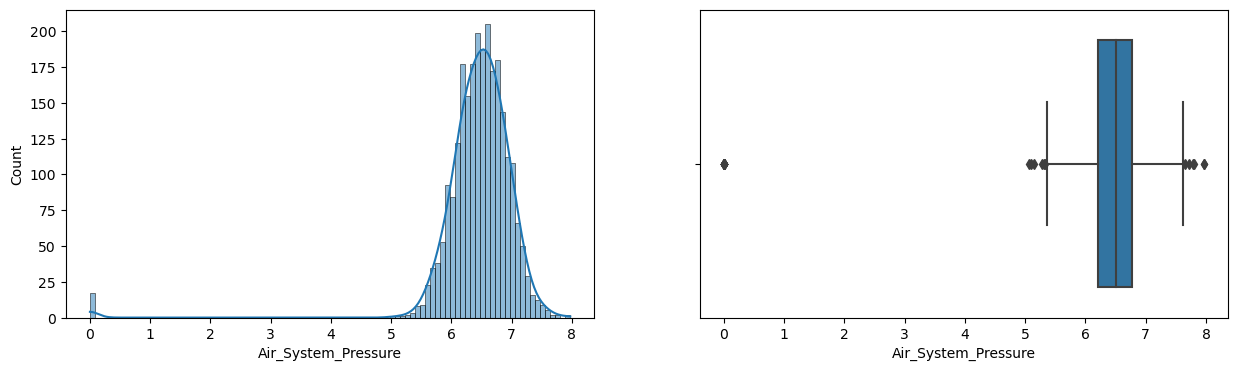
* **Create a histogram or box plot to visualize the distribution of variables.**
* **Identify outliers or extreme values.**

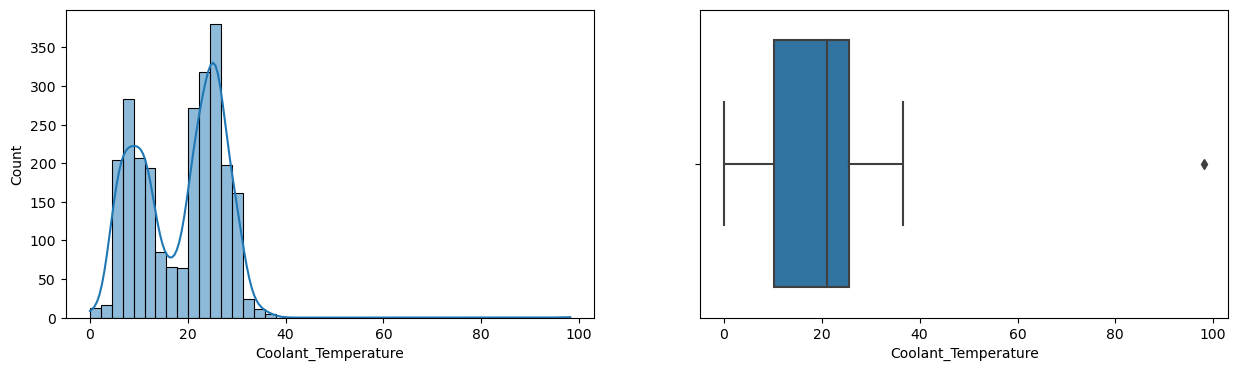
Ans: Identifying outliers as values more than 2 standard deviations away from the mean.

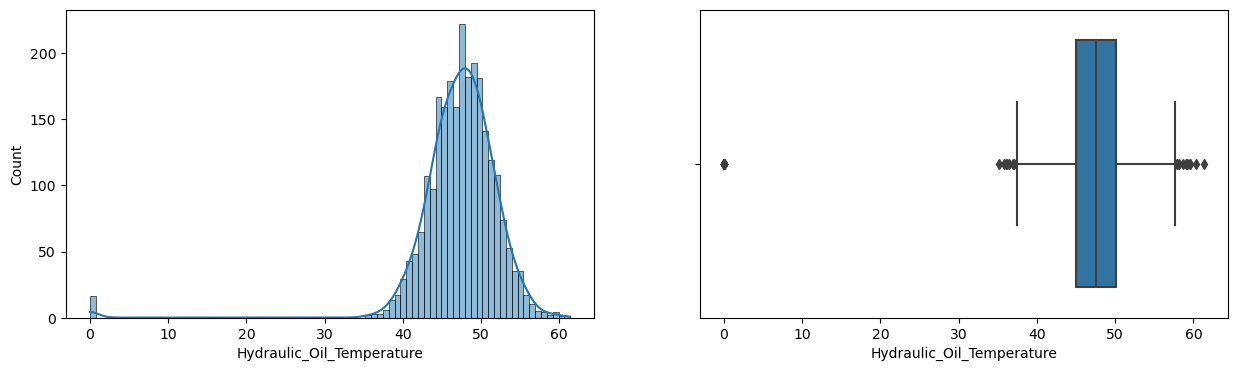
1. Hydraulic\_Pressure is symmetric, and kurtosis is moderate with fewer outliers
2. Coolant\_Pressure is approximately symmetric and kurtosis is laptokurtic with fewer outliers
3. Air\_System\_Pressure is highly negatively skewed and kurtosis is laptokurtic with many outliers
4. Coolant\_Temperature is symmetric, and kurtosis is laptokurtic with very few outliers
5. Hydraulic\_Oil\_Temperature is highly negatively skewed and kurtosis is laptokurtic with many outliers
6. Spindle\_Bearing\_Temperature is highly negatively skewed and kurtosis is laptokurtic with many outliers
7. Spindle\_Vibration is symmetric and kurtosis is mesokurtic with many outliers
8. Tool\_Vibration is approximately symmetric and kurtosis is mesokurtic with many outliers
9. Spindle\_Speed is approximately symmetric and kurtosis is laptokurtic with fewer outliers
10. Voltage is moderately negatively skewed and kurtosis is laptokurtic with many outliers
11. Torque is approximately symmetric and kurtosis is mesokurtic with few outliers
12. Cutting is approximately symmetric and kurtosis is mesokurtic with few outliers

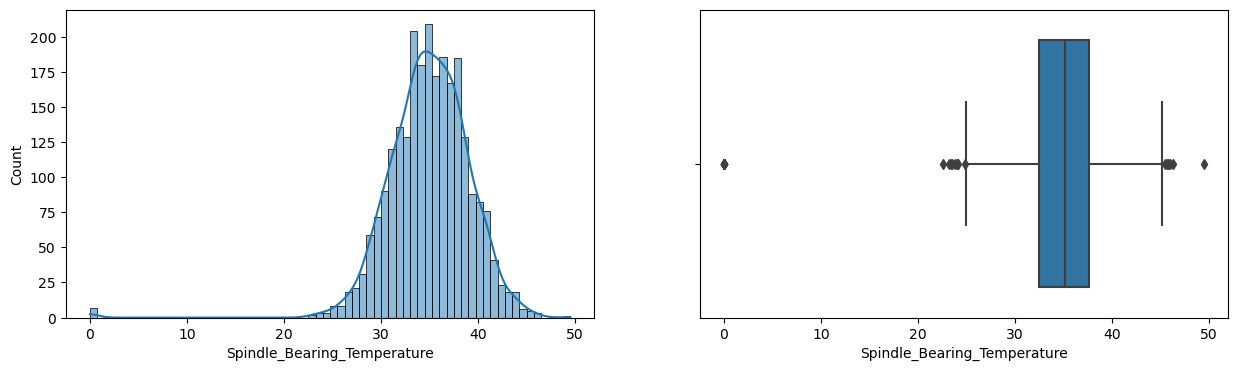


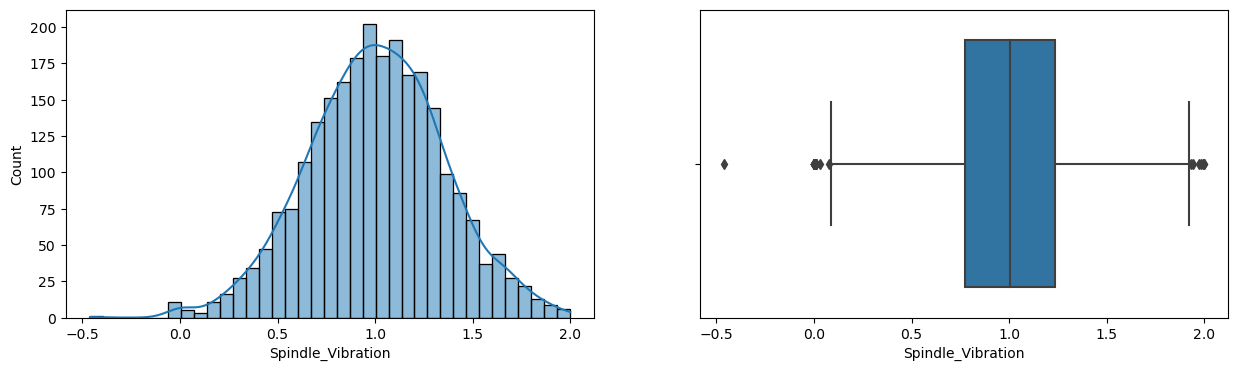


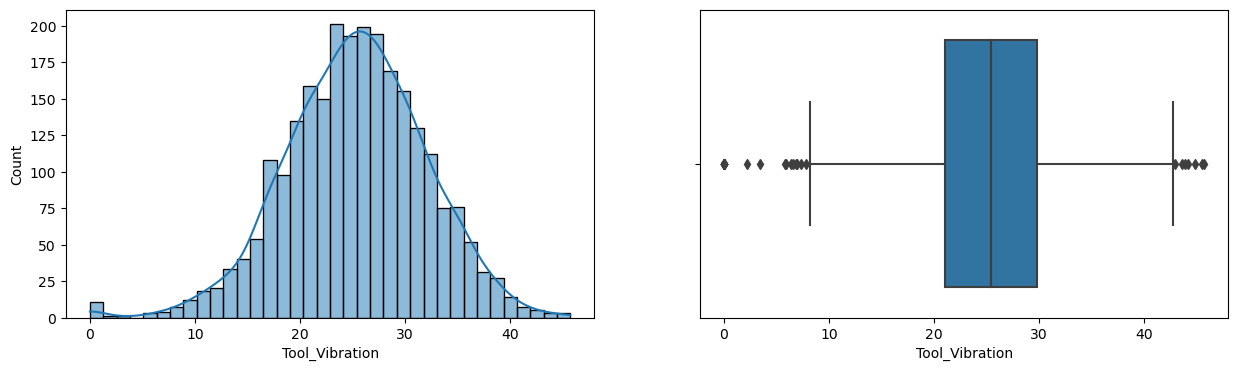


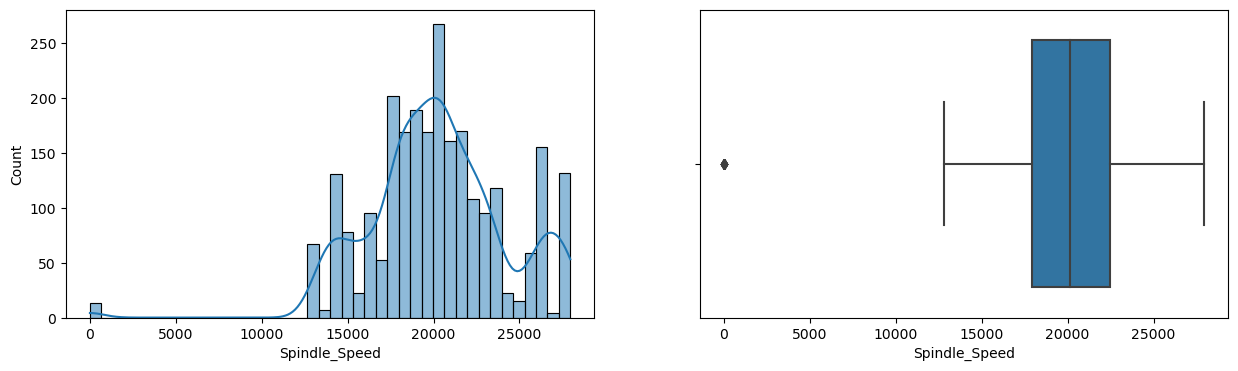


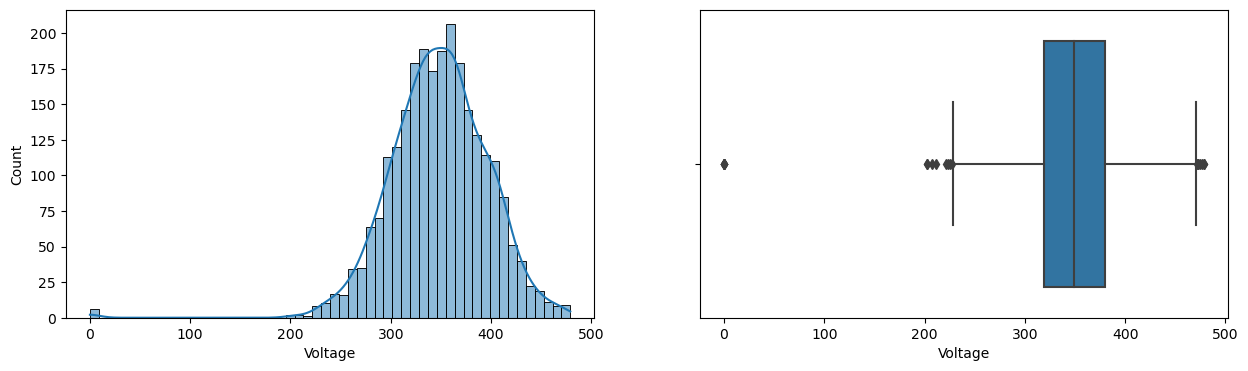


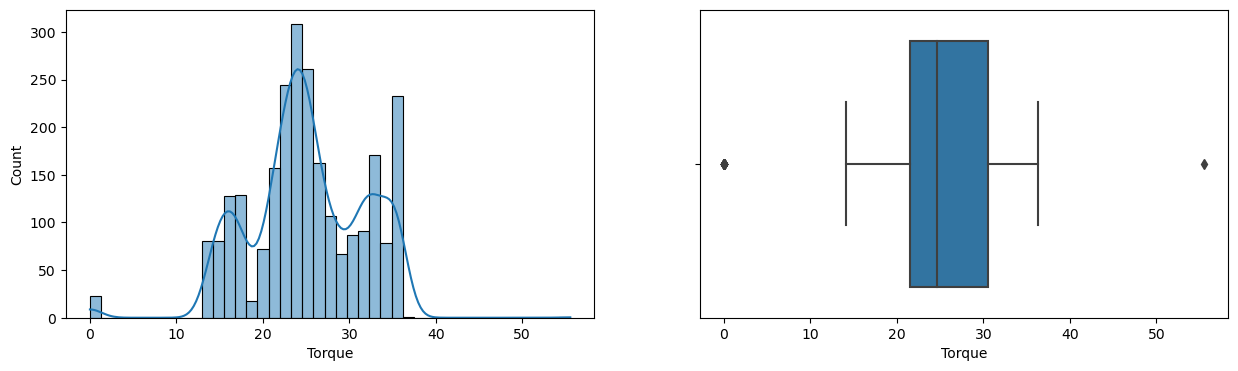


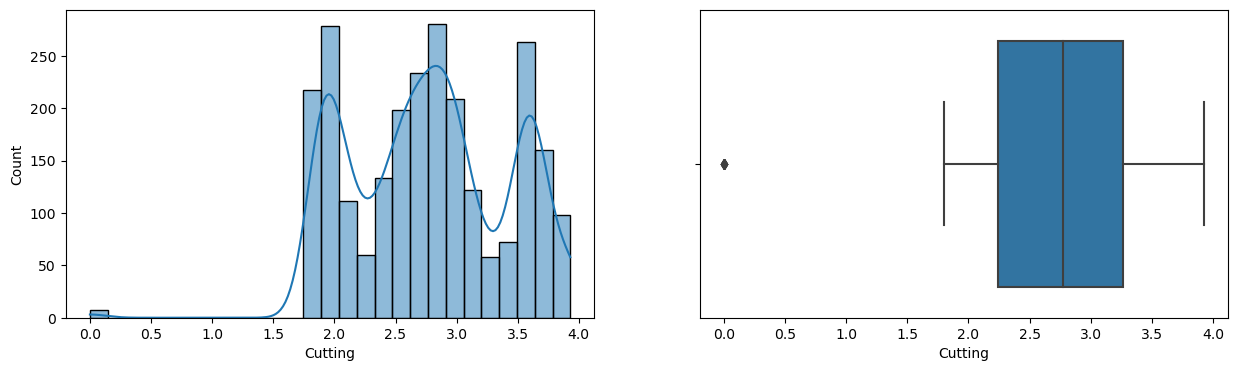




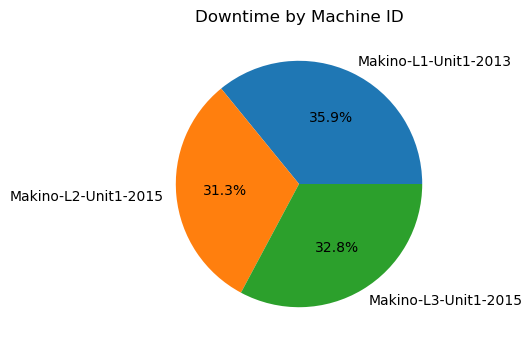




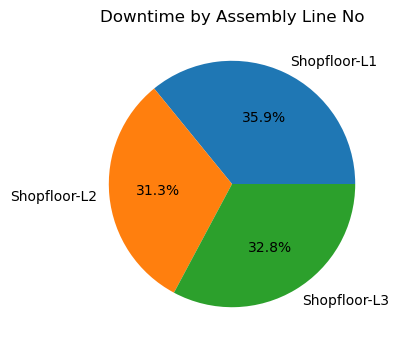




* **Which machines experience the highest frequency of unplanned downtime?**

****

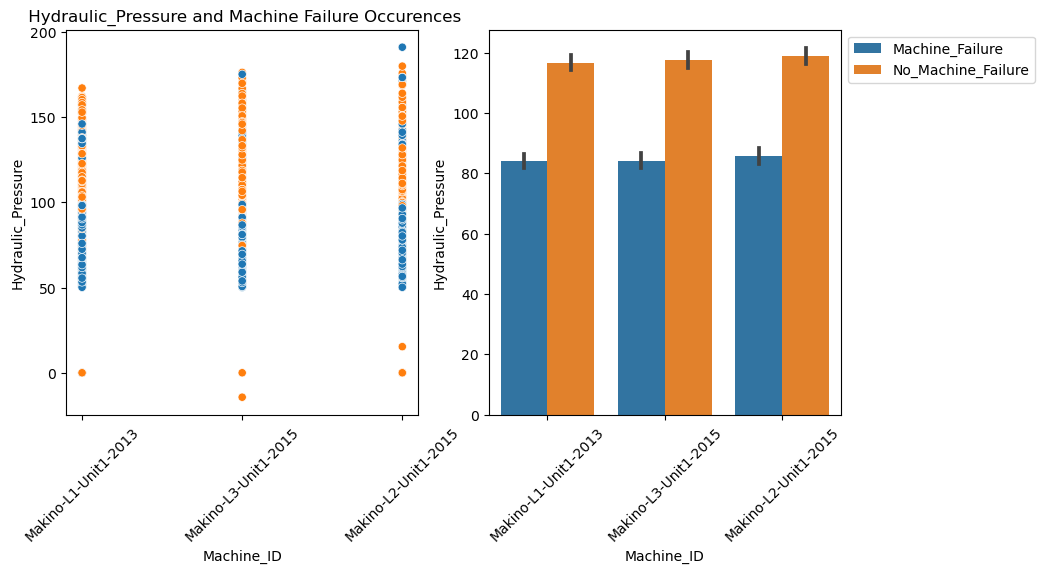
* **Which Assembly line experience the highest frequency of unplanned downtime?**

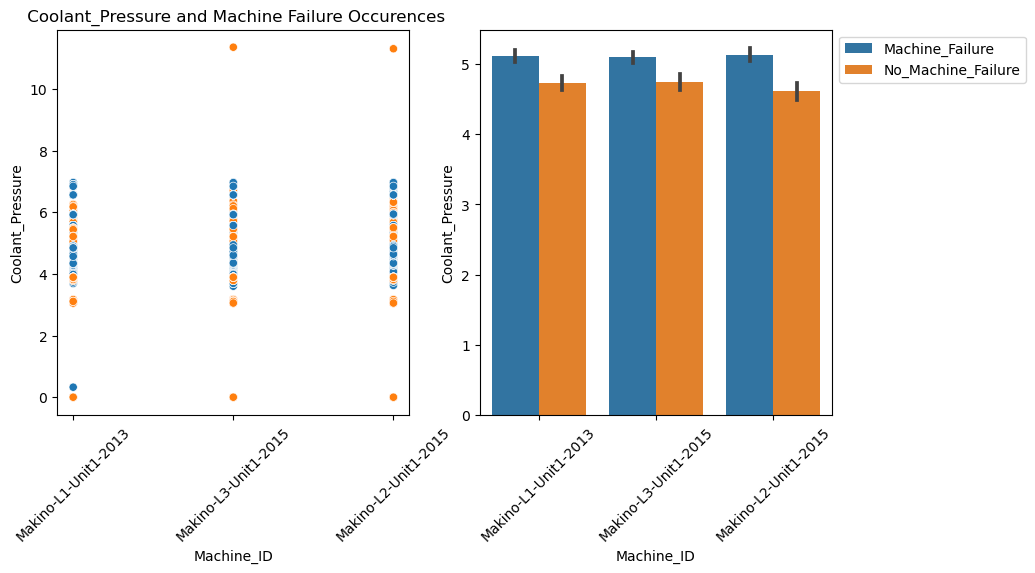
****

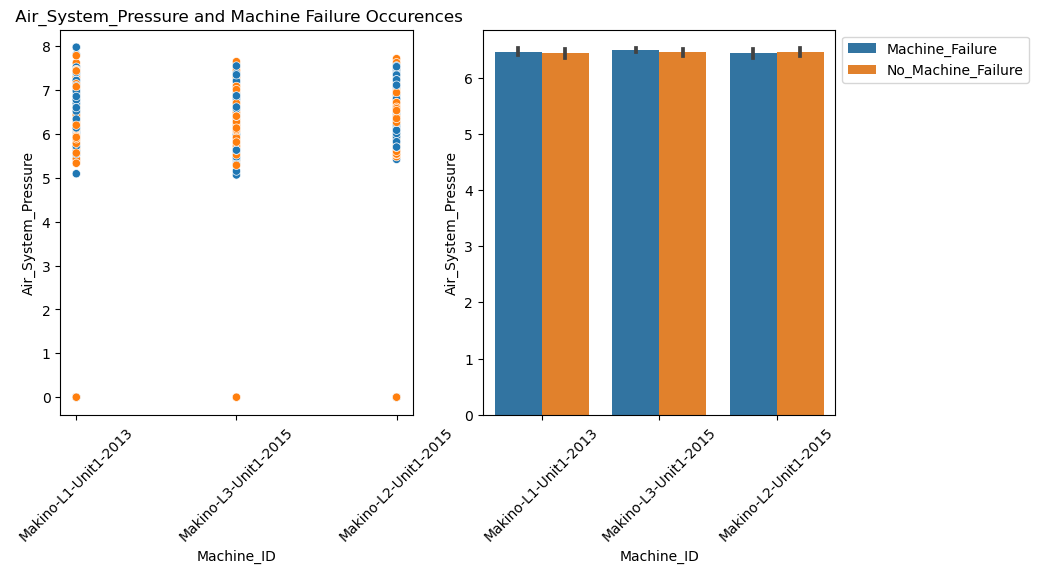
* **Group downtime data by machine type or category and calculate downtime statistics for each group.**

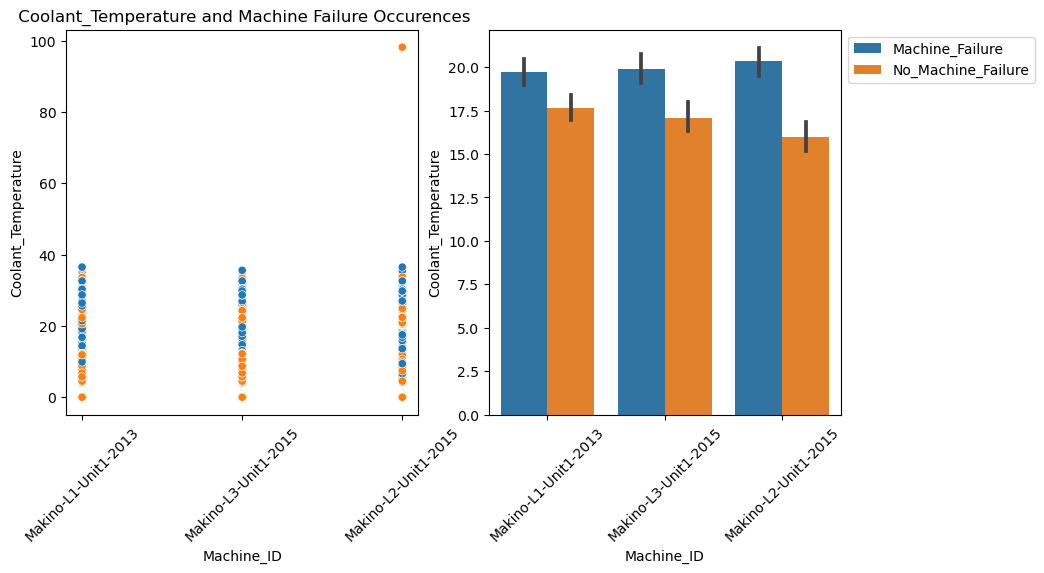
Ans: Most machines are down when

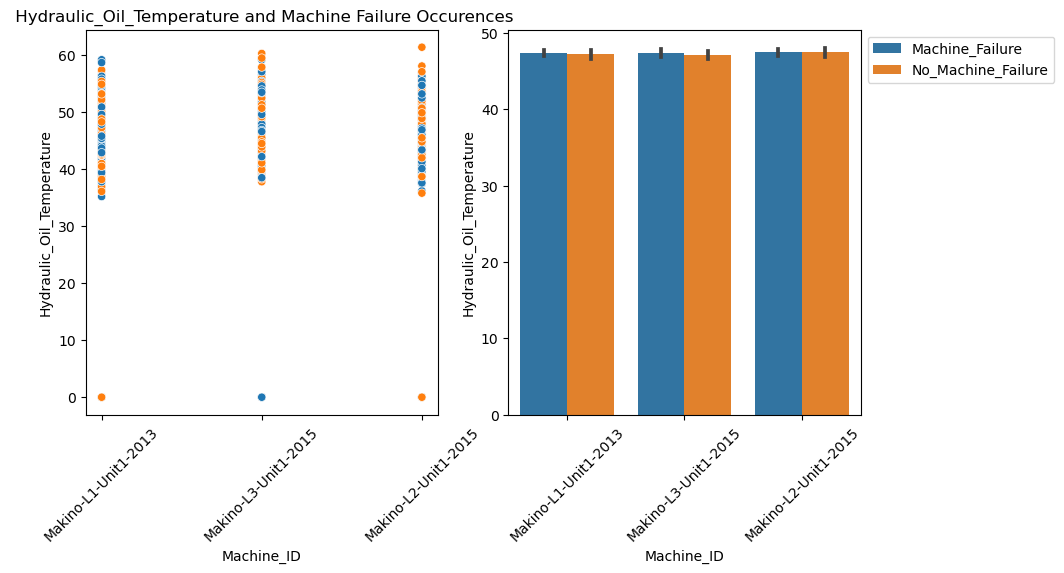
1. Hydraulic\_Pressure is less than 90
2. Coolant\_Pressure is between 4 - 4.8 and above 6.5
3. Air\_System\_Pressure is not affecting the machine downtime
4. Coolant temperature is between 25 -35 degree C
5. Hydraulic\_Oil\_Temperature is no effect
6. Spindle\_Bearing\_Temperature is no effect
7. Spindle\_Vibration is is no effect
8. Tool\_Vibration is not affecting machines to be down
9. Spindle speed is greater than 20500
10. Voltage is no effect
11. Cutting force is between 2.2-2.8 kN and is above 3.5kN
12. Torque is less than 18Nm and 25-30Nm

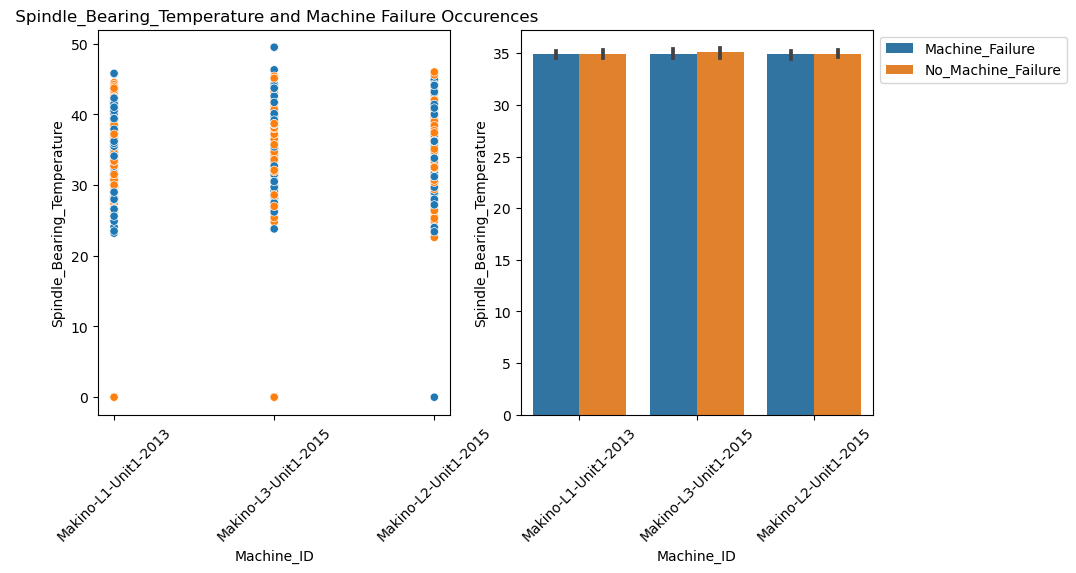


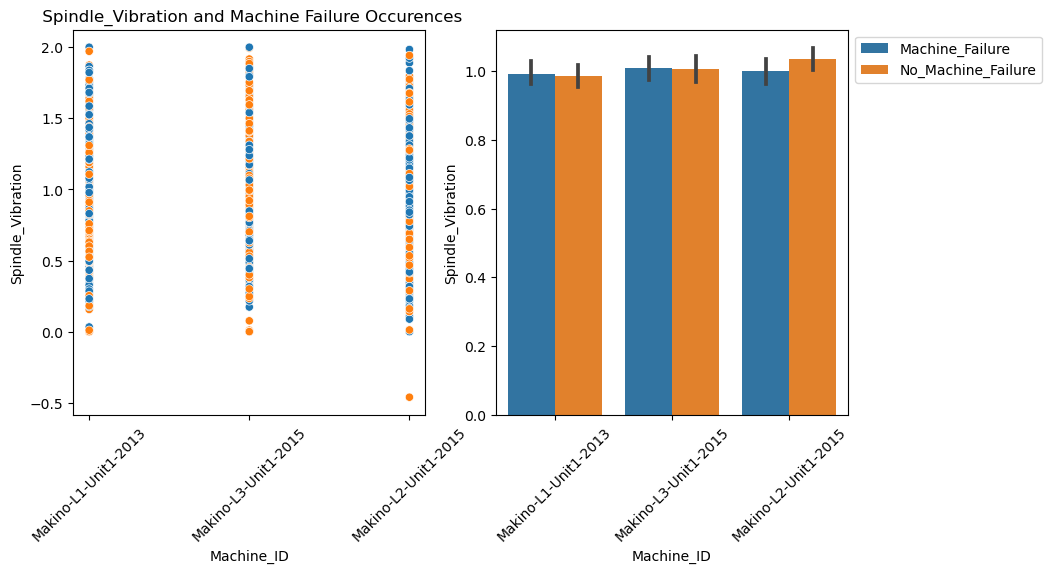


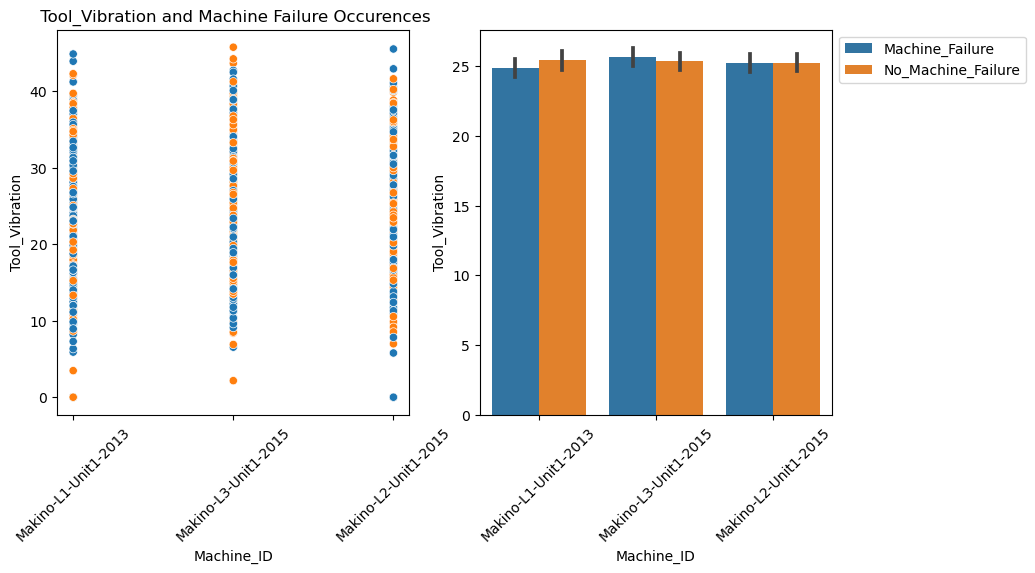


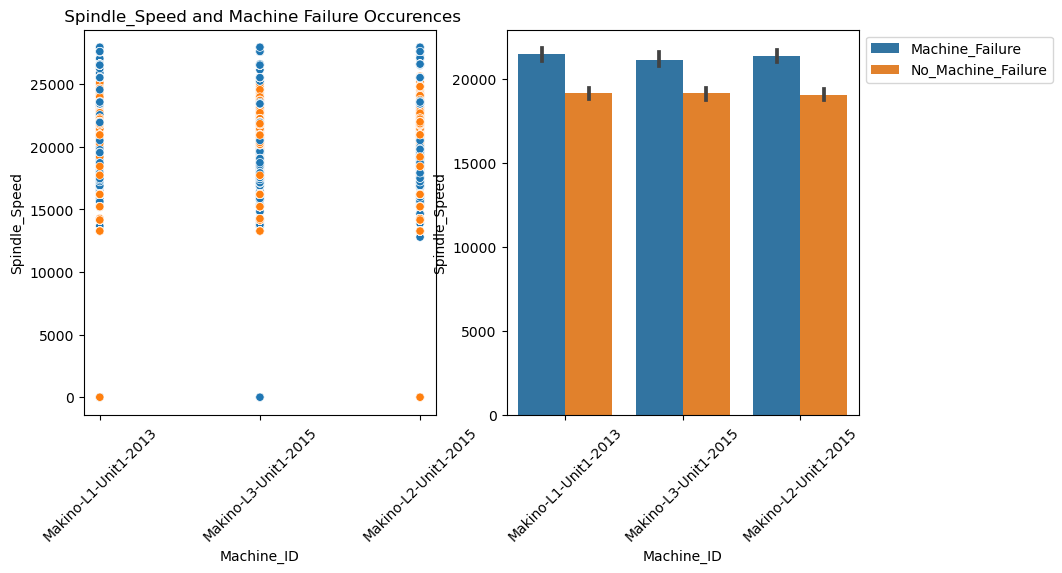


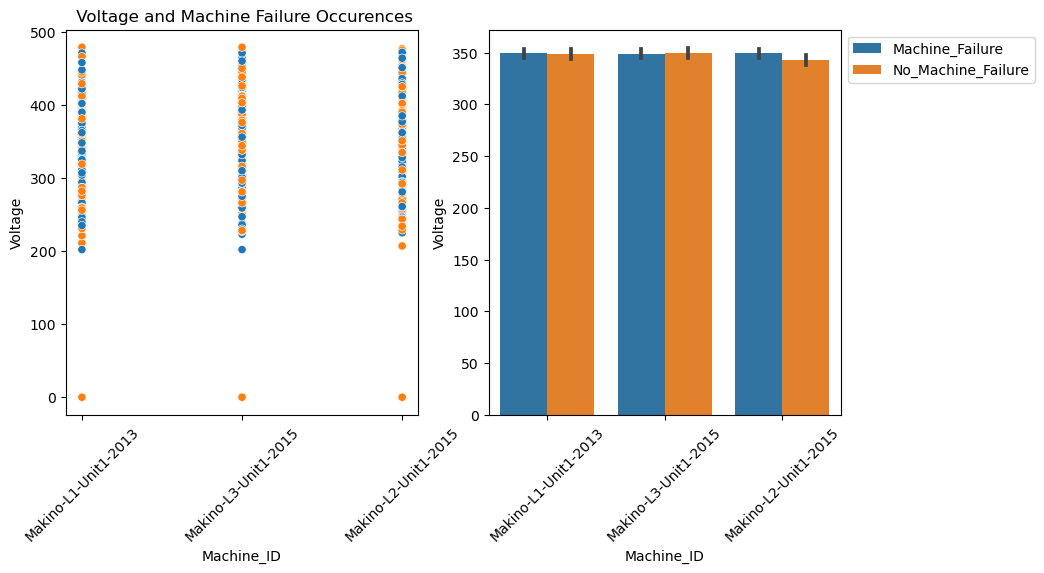


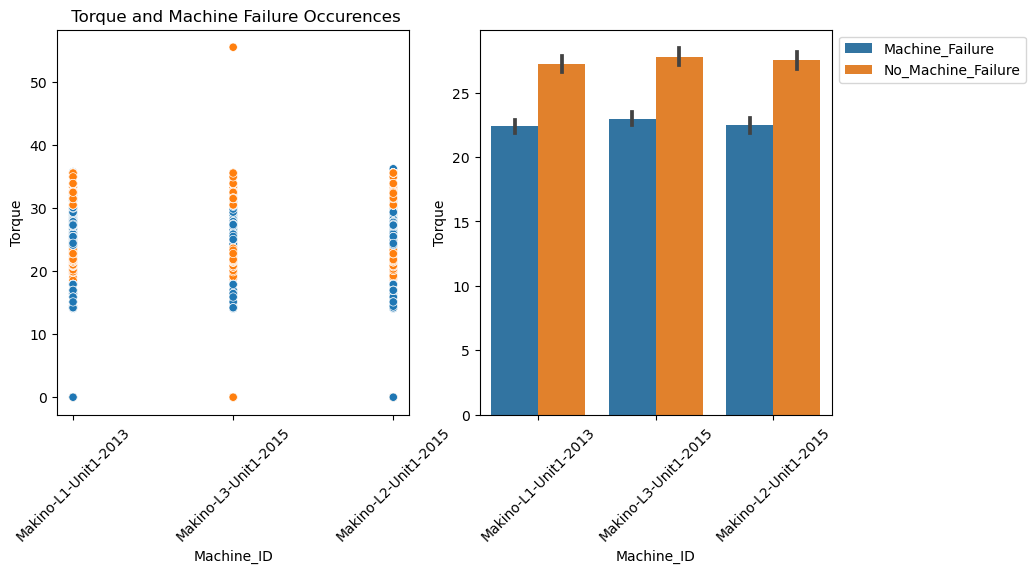


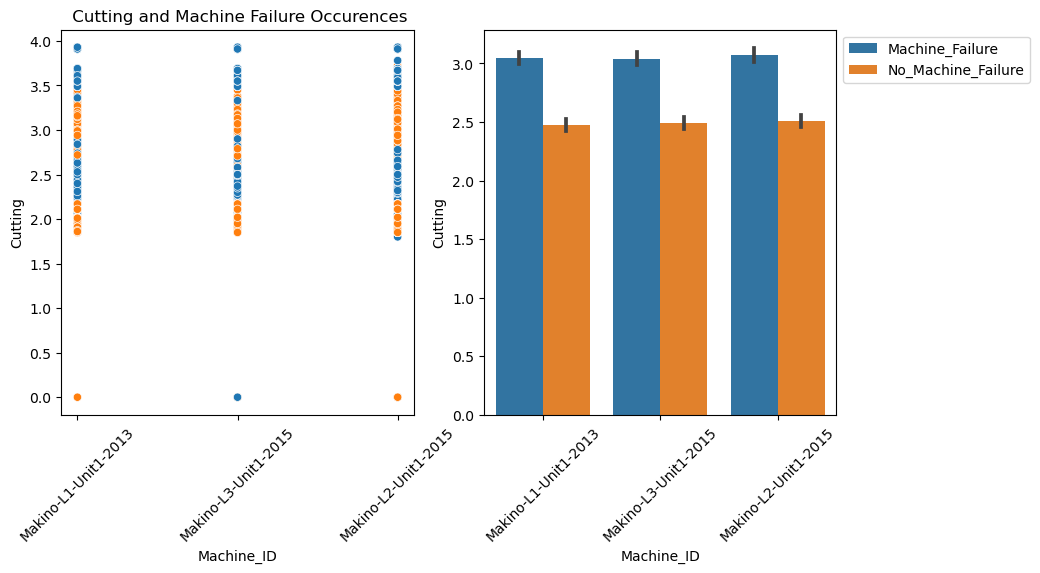












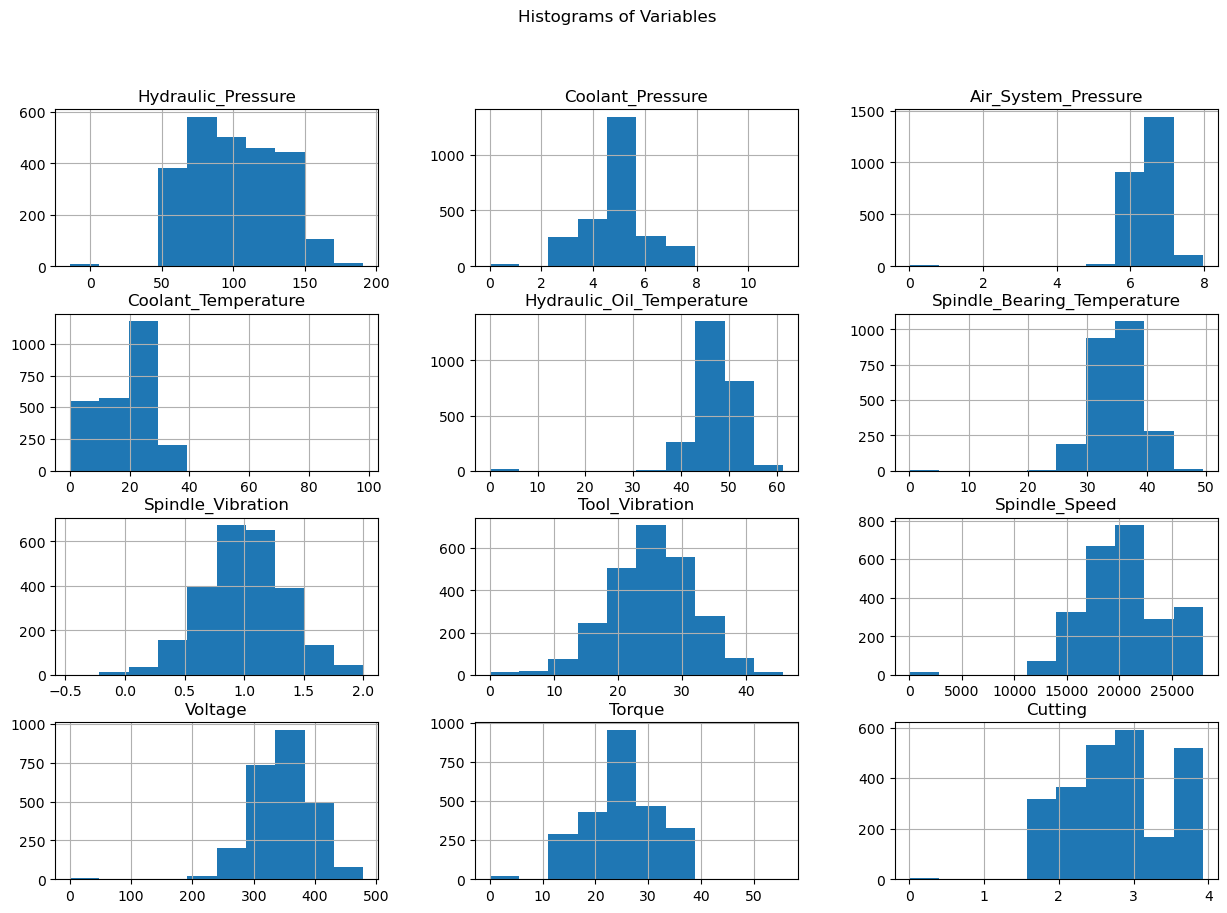
* + Create a correlation heatmap to visualize relationships.

Ans : 1. There is no high correlations between variables

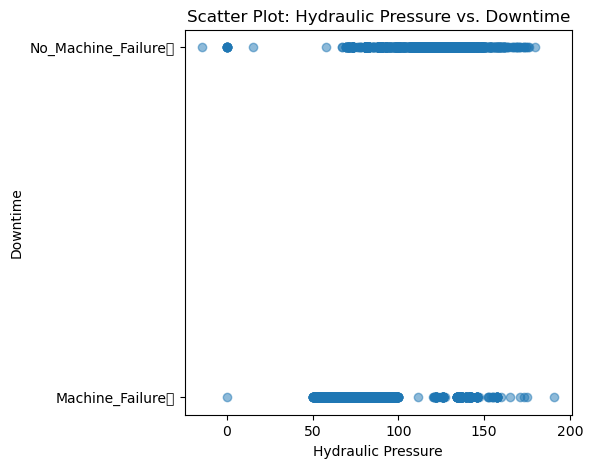
1. Only Spindle speed and cutting has correlation of 23%

## download (25).png

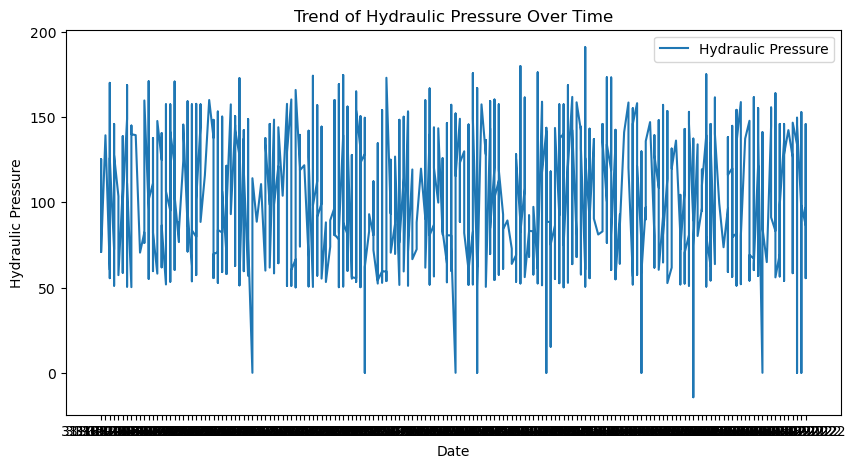
* + Histograms (Univariate Analysis)



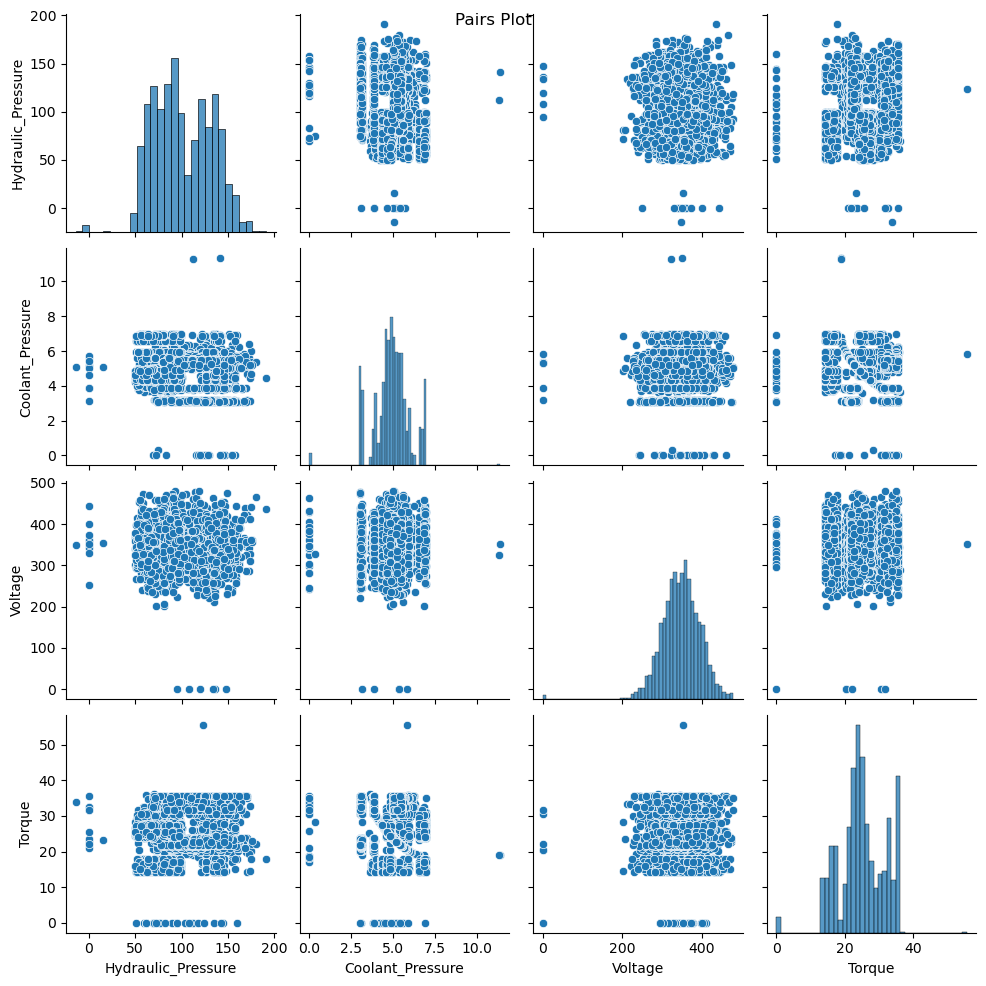
* + create a scatter plot between 'Hydraulic\_Pressure' and 'Downtime':



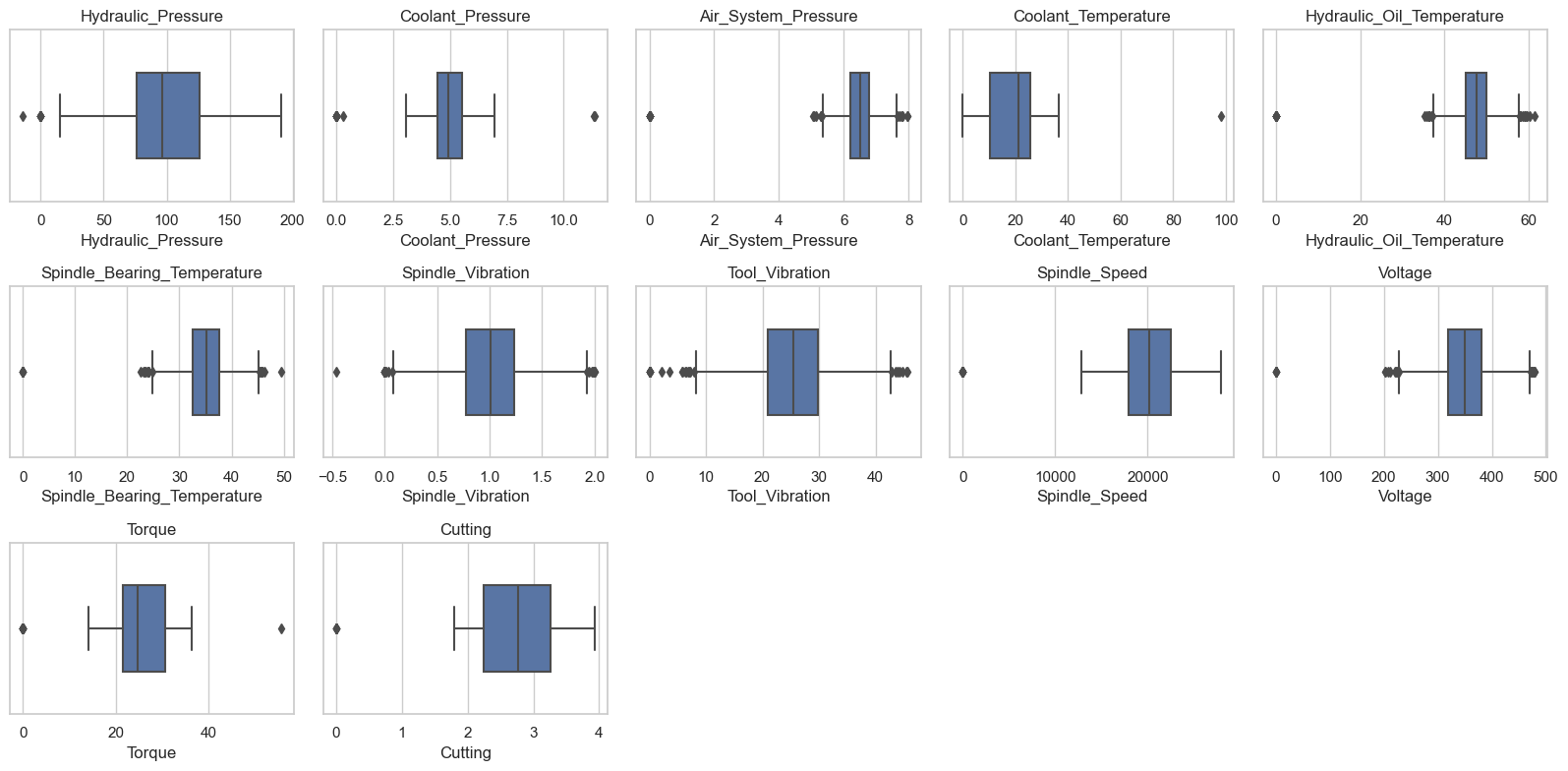
* + Business-Focused Visualizations



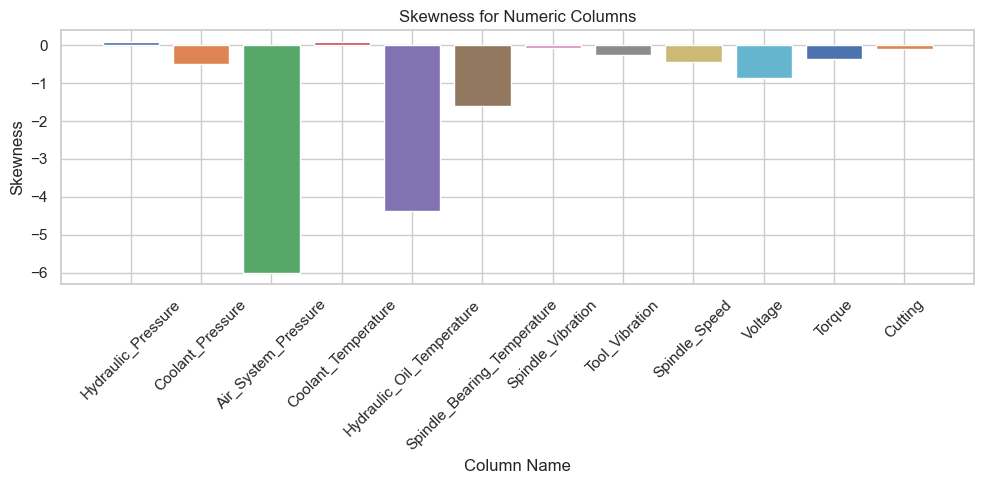
* + Pairs Plot (for numeric columns)



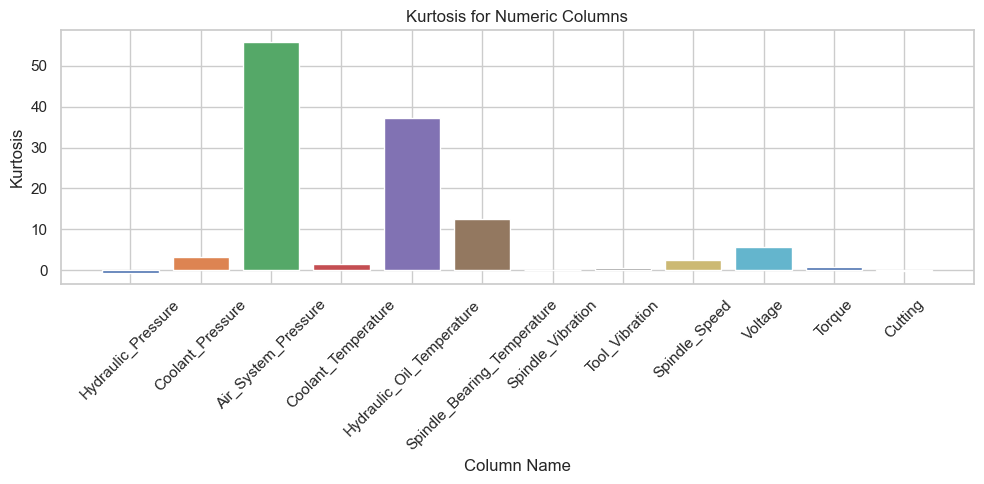
* + UNIVARIATE



* + SKEWNESS



* + kurtosis



## EDA of pre-processed data

* **What is the size of the dataset (number of records and features)?**

Ans : There are total 2500 rows in the dataset with 16 columns or features.

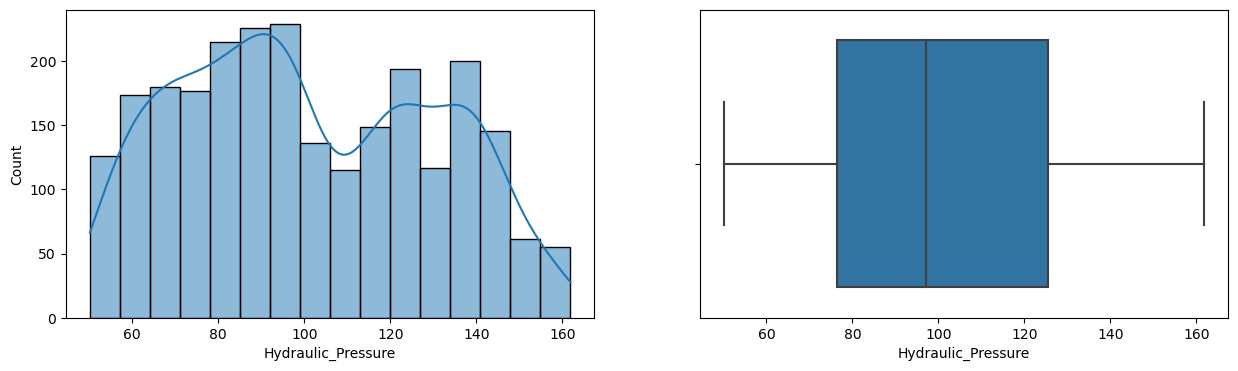
* **Are there any missing values in the dataset?**

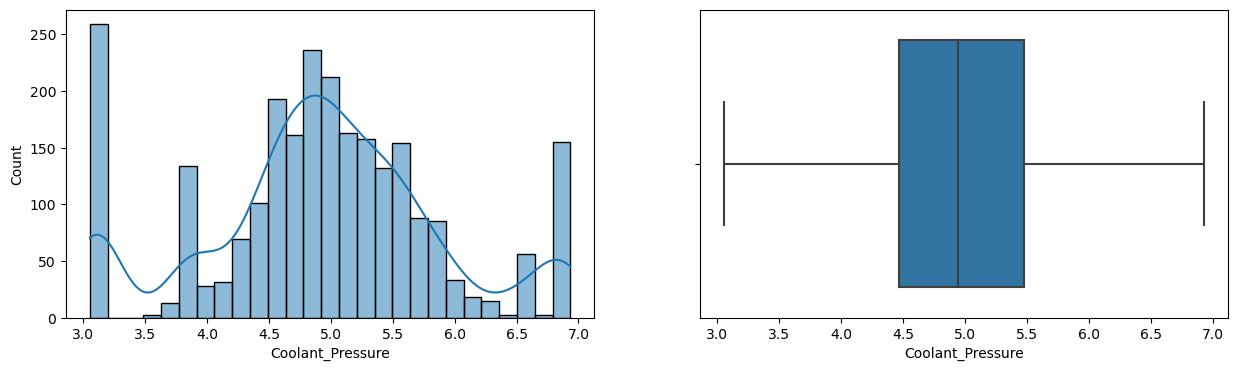
Ans: There are no null values. All missing values are imputed with mean values.

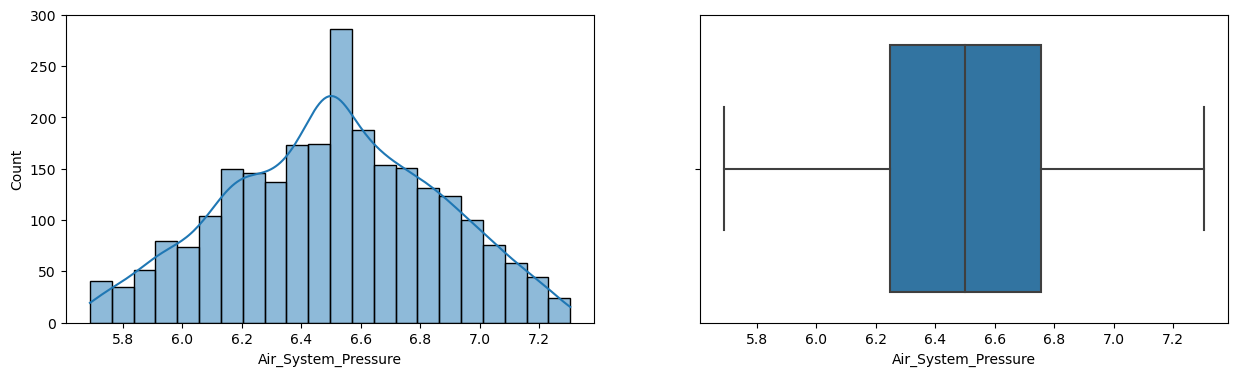
* **Create a histogram or box plot to visualize the distribution of variables.**
* **Identify outliers or extreme values.**

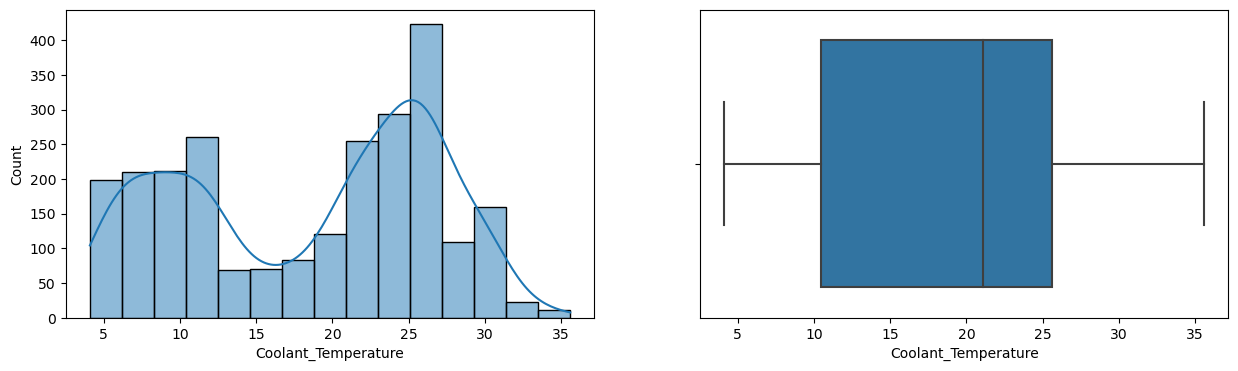
Ans: There are no outliers values. All outliers values are imputed with mean values.

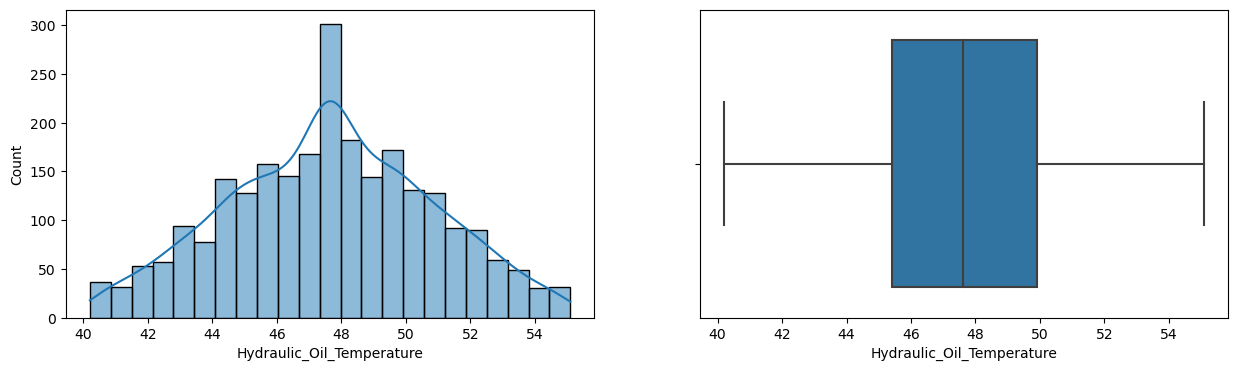
1. Hydraulic\_Pressure is approximately symmetric, and kurtosis is platykurtic
2. Coolant\_Pressure is symmetric and kurtosis is mesokurtic
3. Air\_System\_Pressure is symmetric and kurtosis is mesokurtic
4. Coolant\_Temperature is approximately symmetric, and kurtosis is platykurtic
5. Hydraulic\_Oil\_Temperature is symmetric and kurtosis is mesokurtic
6. Spindle\_Bearing\_Temperature is symmetric and kurtosis is mesokurtic
7. Spindle\_Vibration is symmetric and kurtosis is mesokurtic
8. Tool\_Vibration is symmetric and kurtosis is mesokurtic
9. Spindle\_Speed is approximately symmetric and kurtosis is mesokurtic
10. Voltage is symmetric and kurtosis is mesokurtic
11. Torque is symmetric and kurtosis is mesokurtic
12. Cutting is approximately symmetric and kurtosis is platykurtic

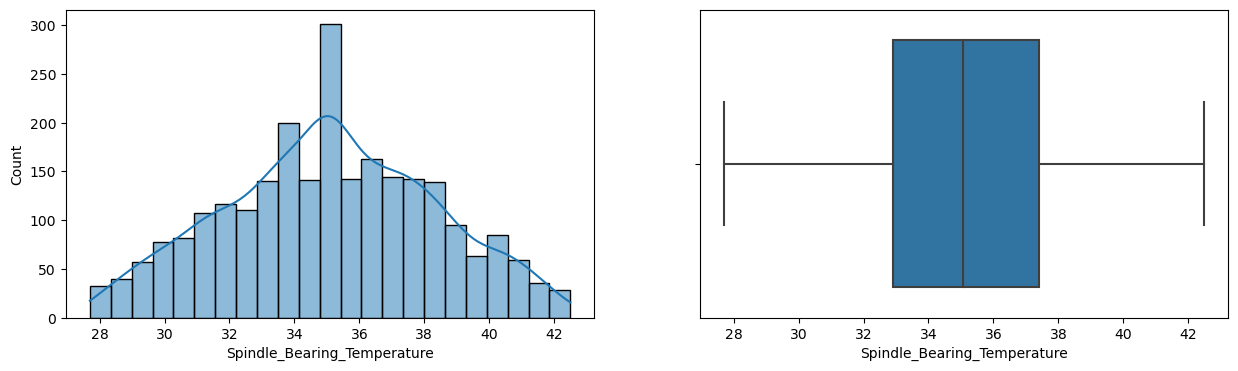


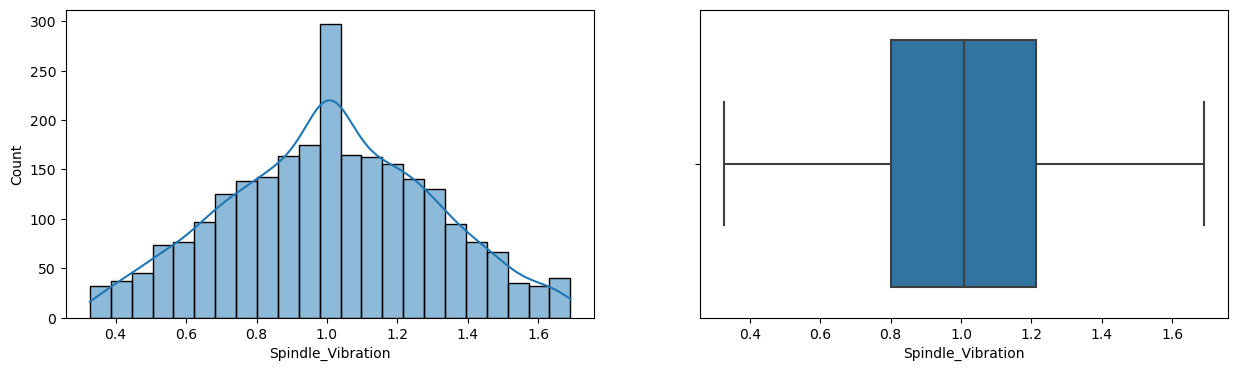


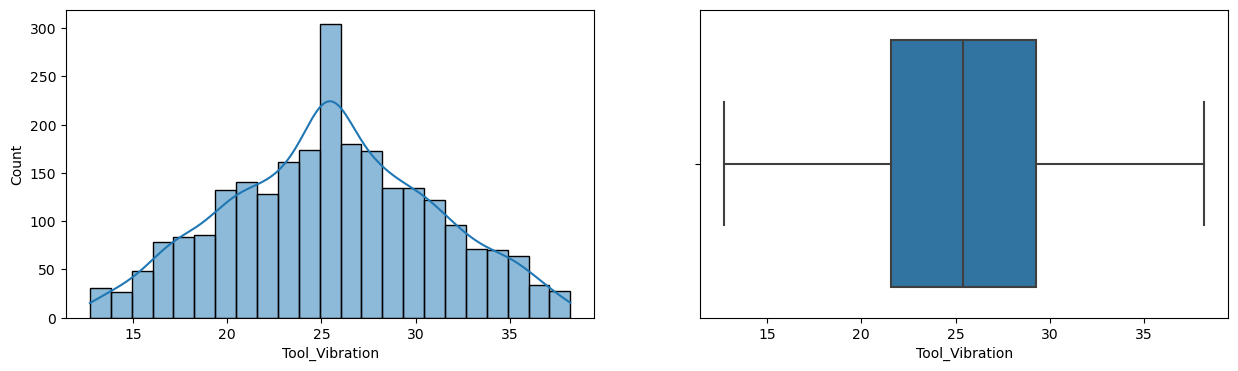


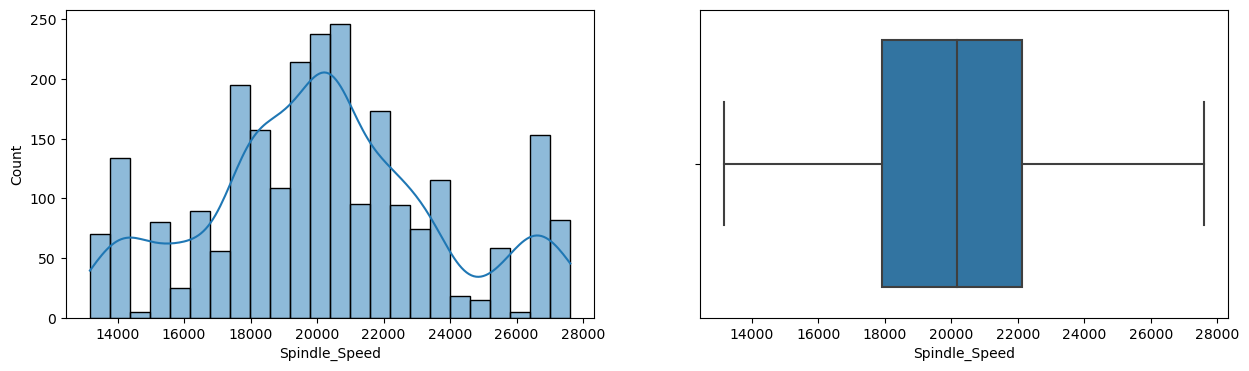


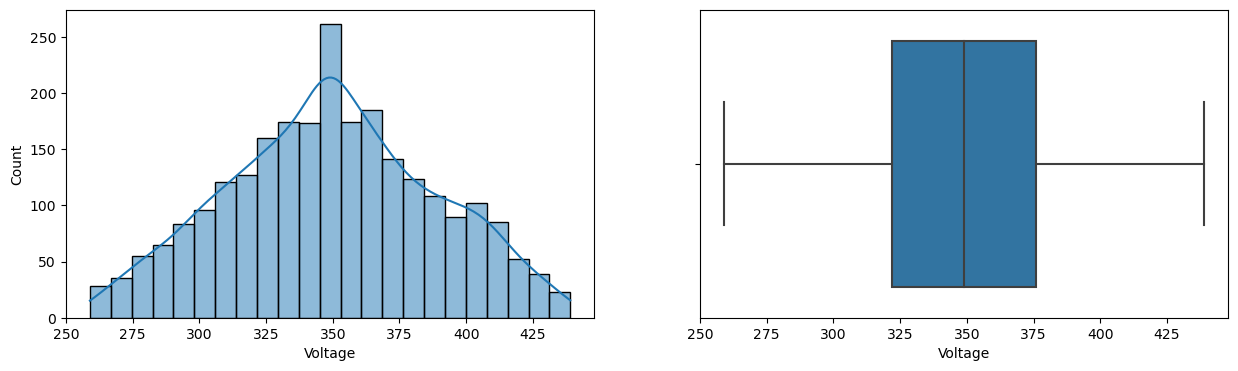


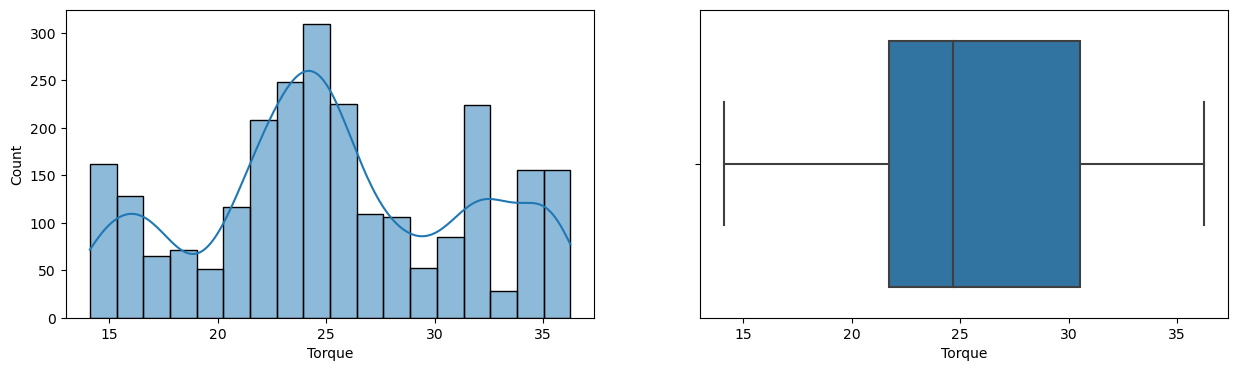


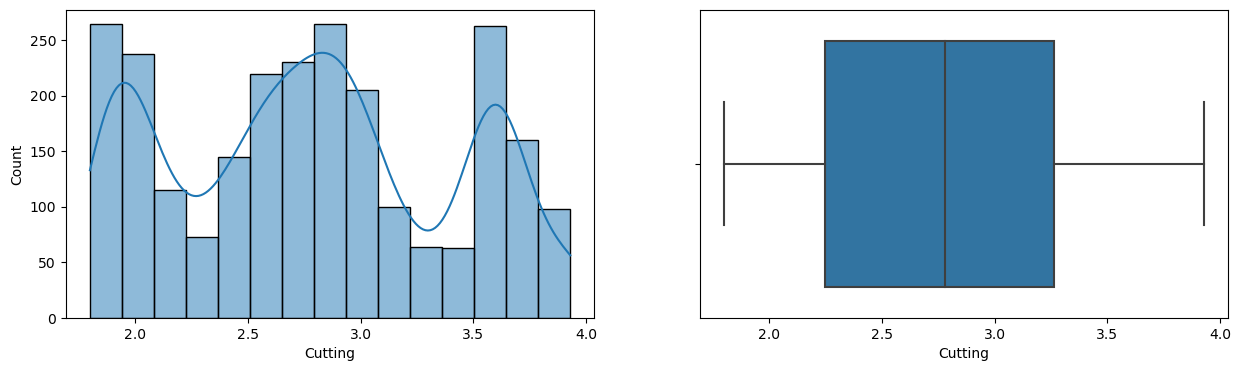












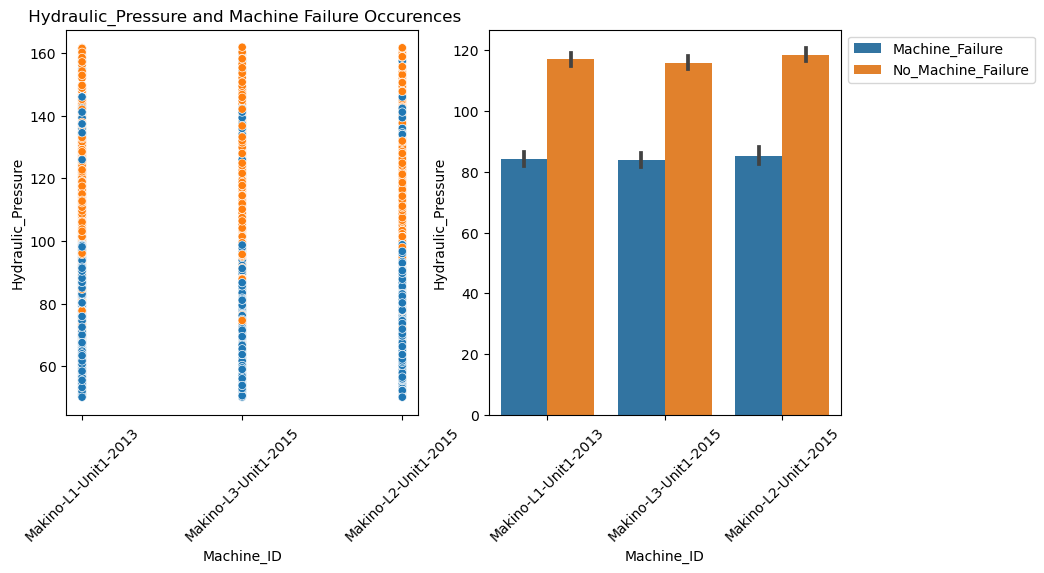
* **Find a time series of machine downtime over time to identify trends or seasonality.**

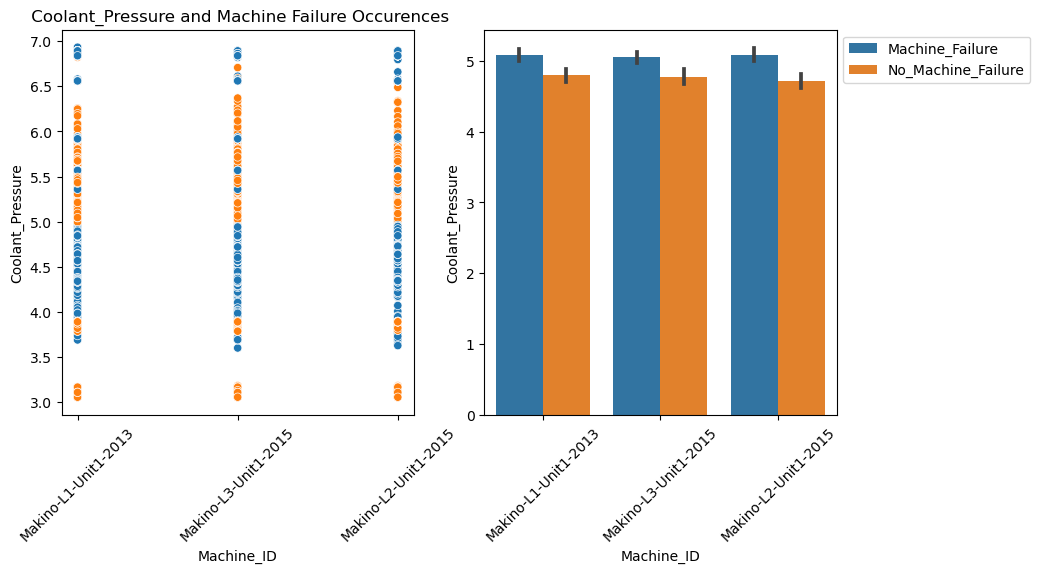
Ans. Data is not so huge to find difference. Same as before preprocessing

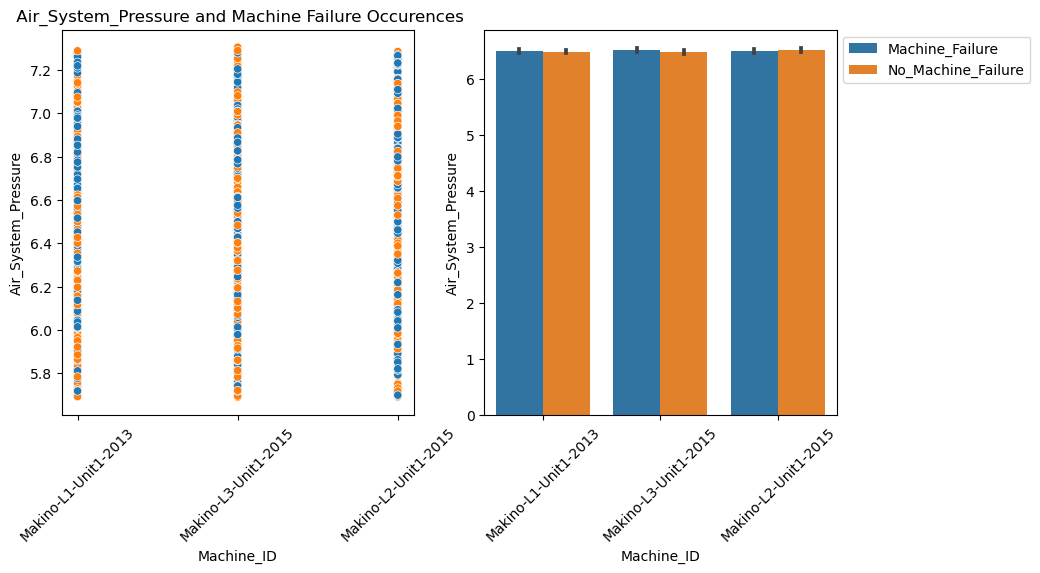
* **Group downtime data by machine type or category and calculate downtime statistics for each group.**

Ans. Most machines are down when

1. Hydraulic\_Pressure is less than 90
2. Coolant\_Pressure is between 4 - 4.8 and above 6.5
3. Air\_System\_Pressure is not affecting the machine downtime
4. Coolant temperature is between 13-20 and 25 -35 degree C
5. Hydraulic\_Oil\_Temperature is no effect
6. Spindle\_Bearing\_Temperature is no effect
7. Spindle\_Vibration is is no effect
8. Tool\_Vibration is not affecting machines to be down
9. Spindle speed is greater than 20500
10. Voltage is no effect
11. Cutting force is between 2.2-2.7 kN and is above 3.5kN
12. Torque is less than 18Nm and 23-30Nm



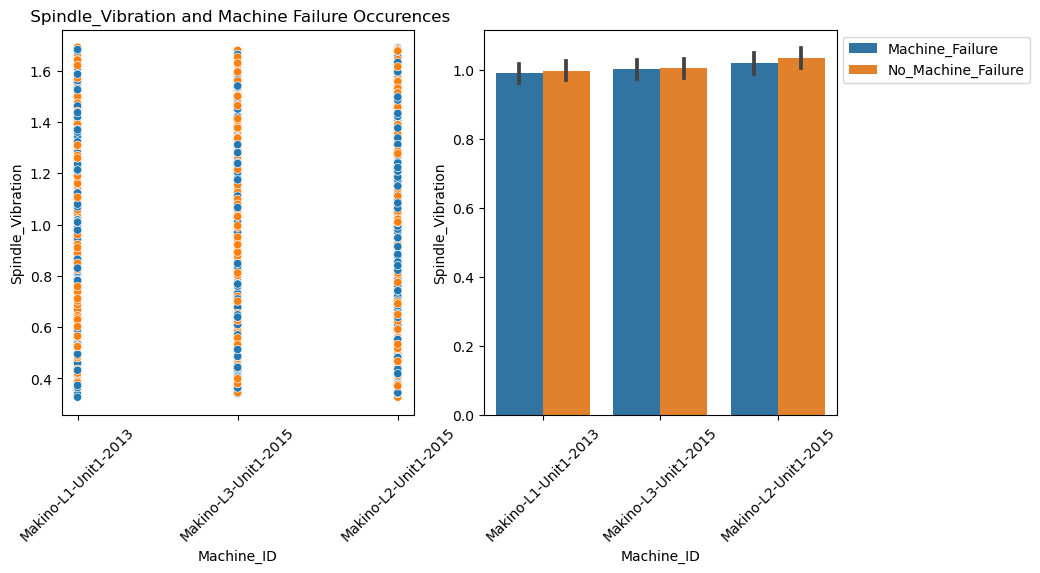




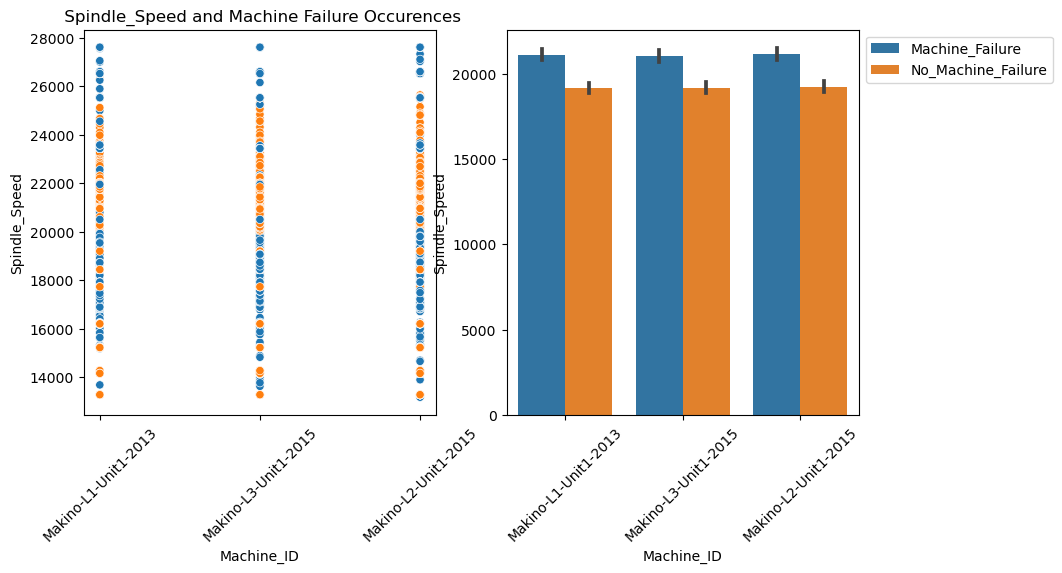
## 

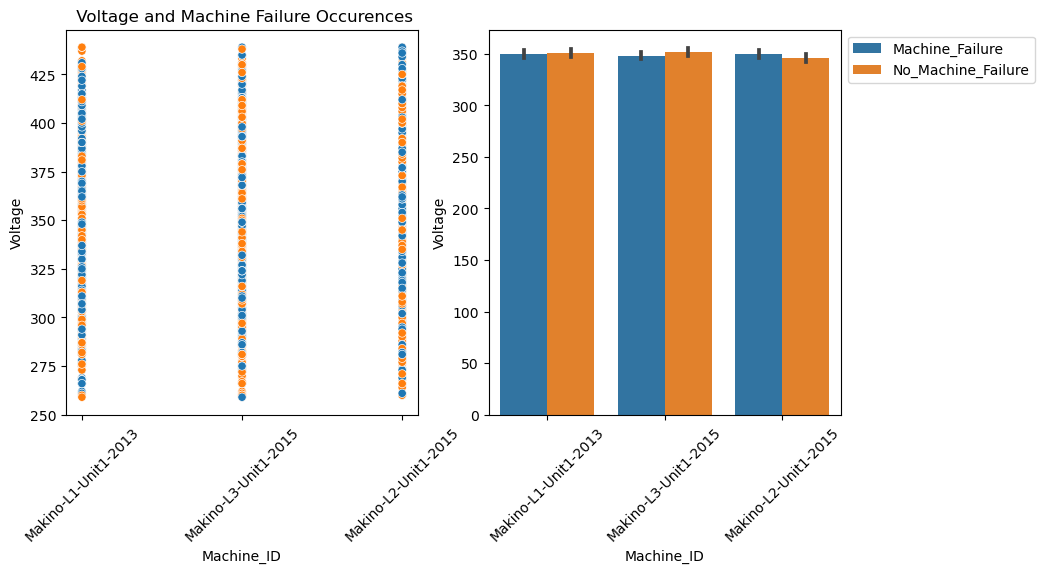
## 

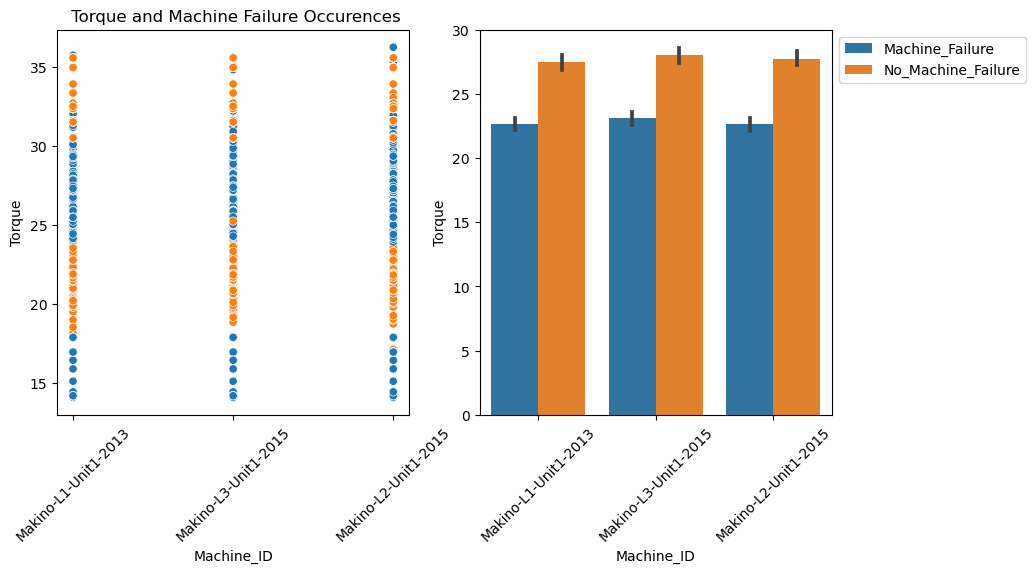
## 

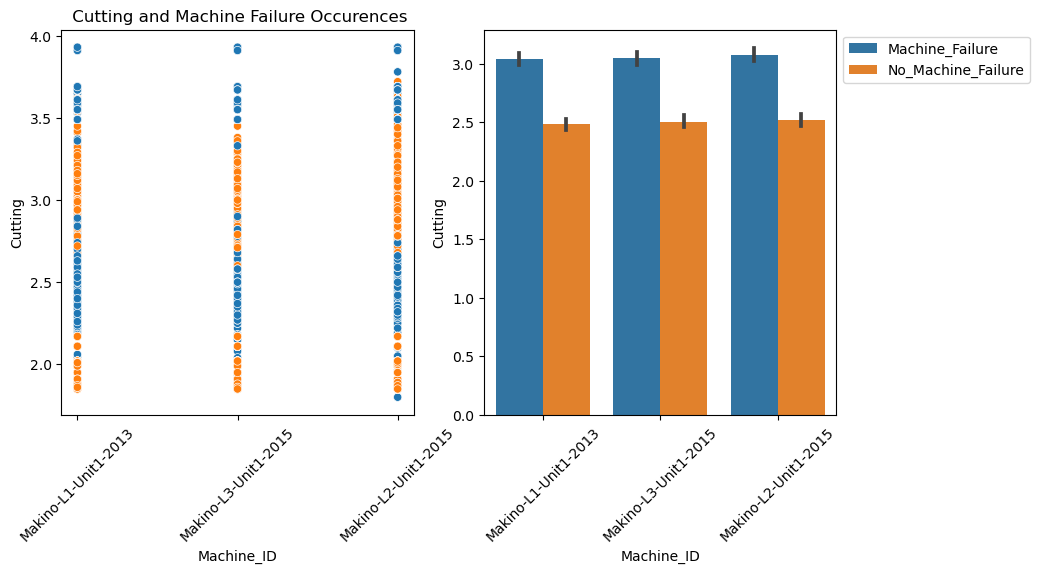


## 









## Business Insights after preprocessing:

If we consider all parameter metrics which are affecting the machines downtime within specified range downtime can be reduced.