DESCRIPTIVE ANALYTICS | EXPLORATORY DATA ANALYSIS

Since my dataset contains categorical and numeric features, I conducted three different kinds of analysis to learn more about my dataset. Univariate analysis, Bivariate Analysis and Multivariate Analysis.

Data Acquisition

Data for this project is acquired from kaggle. The datasets are open-sourced and compliant with the MRP requirements and have been collected and provided by Playground Series. The dataset we'll be using includes various features related to machine operations and failures. Each row represents a unique machine operation, and includes measurements such as air temperature, process temperature, rotational speed, and torque. It also includes a binary indicator of whether the machine failed during that operation, along with indicators for different types of failures.

Data Source & Data Files

The datasets have been collected by Playground Series(Kaggle).This dataset was split into two subsets for training and evaluation of the classifier. Details about the main training/testing dataset are given below:

a. For training we have 136429 records, 14 features.
b. For testing we have 90954 records, 13 features.

We have multiple data types lille categorical. Numerical and floating in our training and testing data.

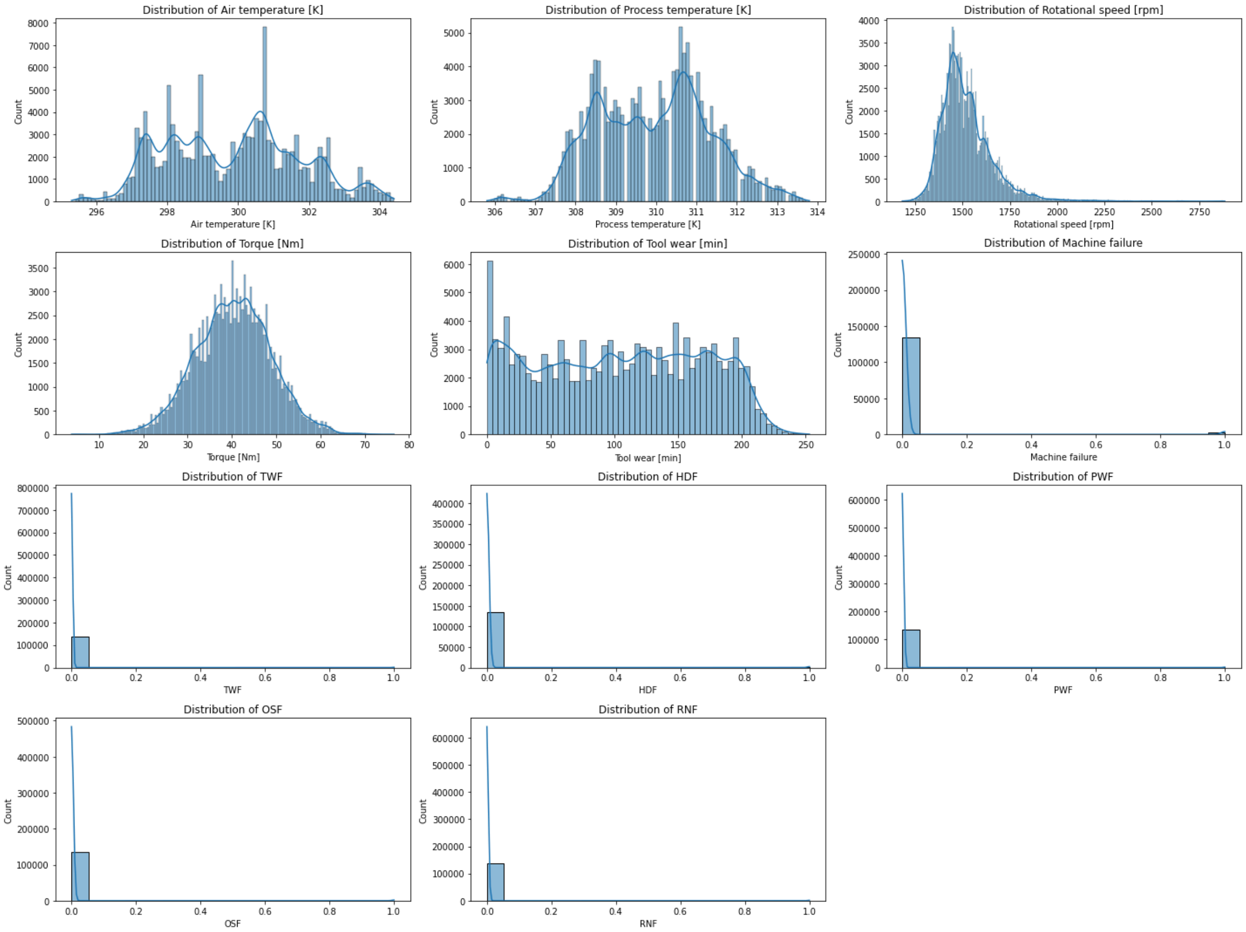
Training Data statistics Summary Insights

- The dataset contains `136,429` records with features such as temperatures, rotational speed, torque, and tool wear.
- The target variable, "Machine failure", and specific failure types (TWF, HDF, PWF, OSF, RNF) are binary indicators.
- The air temperature ranges from `295.3K` to `304.4K`, with a mean of approximately `299.9K`.
- The process temperature ranges from 305.8K to 313.8K, with a mean of approximately `309.9K`.
- The rotational speed varies from `1,181 rpm` to `2,886 rpm`, with a mean of about `1,520 rpm`.
- Torque ranges from `3.8 Nm` to `76.6 Nm`, with a mean of `40.3 Nm`.
- Tool wear spans from `0` to `253 minutes`, with an average of about `104.4 minutes`.

Exploratory Data Analysis

Univariate Analysis

- Examine the distribution of each feature.

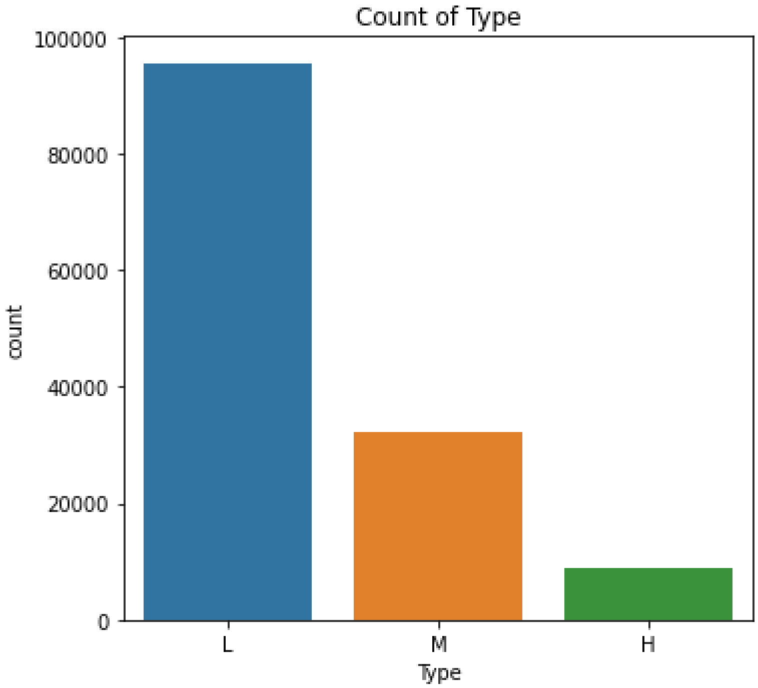


In this Analysis the air temperature shows a general trend of fluctuating around 300 Kelvin over time. This indicates a relatively stable environment with minor variations in temperature, which could be due to natural environmental changes or operational factors. In contrast, the process temperature demonstrates a clear upward trend over time, suggesting that the system or process being monitored is gradually heating up. This could be due to the accumulation of heat from continuous operation or an increase in workload or intensity of the process.

The rotational speed, on the other hand, shows significant fluctuations, indicating variability in the speed at which the system operates. Notably, there are two prominent peaks observed in the data: one around 1750 rpm and another around 2500 rpm. These peaks suggest that the system reaches these rotational speeds more frequently or maintains these speeds for longer periods, possibly due to operational requirements or specific phases in the process cycle.

Lastly, the torque also fluctuates over time, generally centering around 30 Nm. This indicates that while there are variations, the torque remains relatively consistent on average, reflecting the system's ability to maintain a steady force during operation despite fluctuations in other parameters like rotational speed and temperature. Together, these insights provide a comprehensive understanding of the system's operational characteristics and potential areas for further analysis or optimization.

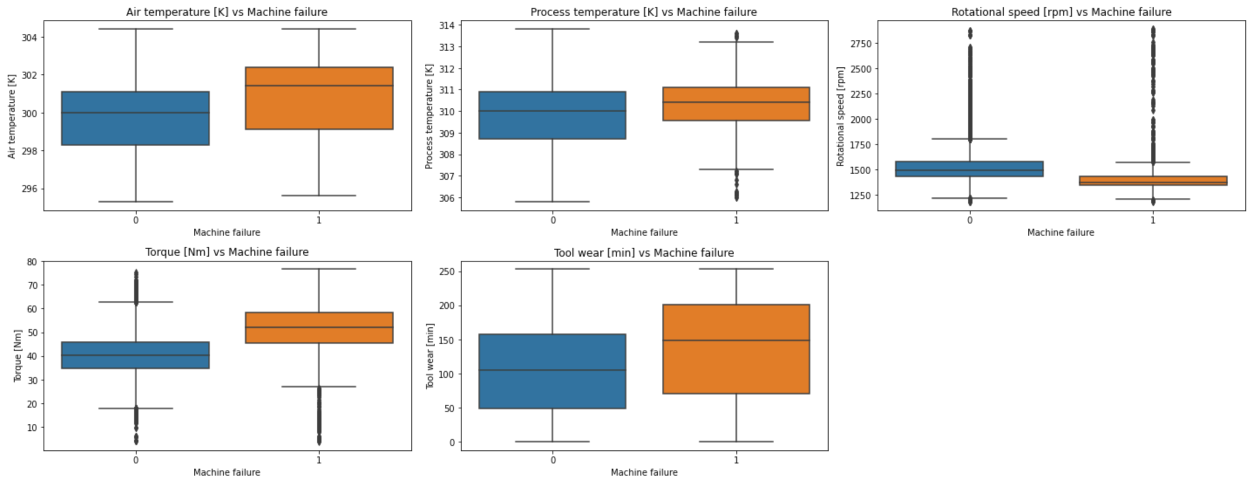
Distribution of each categorical feature.



The bar graph shows the number of types of machines. The text labels on the x-axis (H, M, L) likely correspond to different failure rates (High, Medium, Low).

- **Complexity:** Machines with more moving parts or more complex designs are generally more likely to fail than simpler machines. This is because there are more opportunities for things to go wrong.
- **Stress:** Machines that are under a lot of stress are more likely to fail than machines that are not. For example, a machine that is constantly vibrating or that is exposed to extreme temperatures is more likely to fail than a machine that is operated in a more controlled environment.
- **Age:** As machines age, they are more likely to fail. This is because the parts of the machine wear out over time.

Bivariate Analysis
Relationships between the target variable (machine failure) and other features.

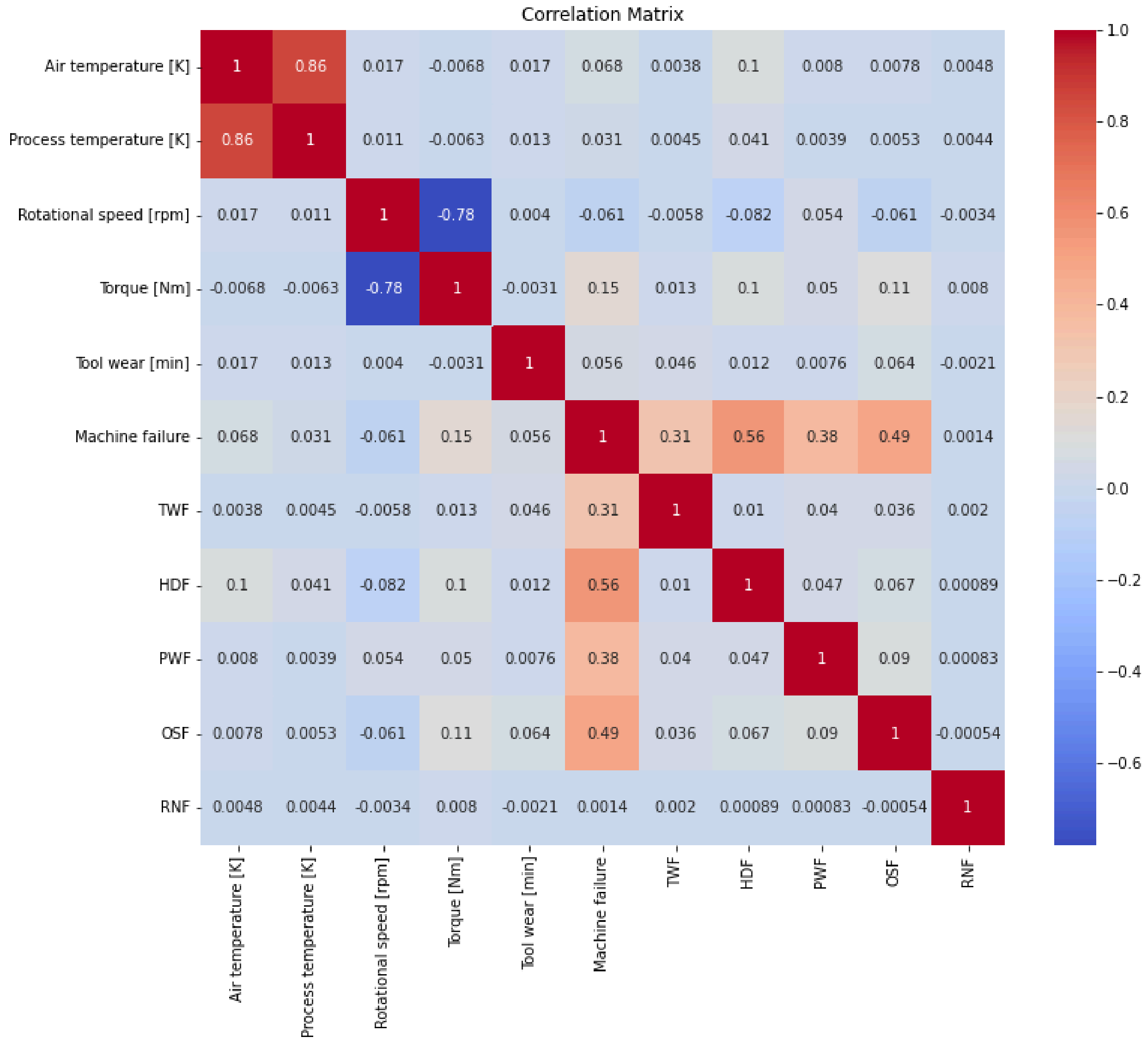


The presence of overstrain is clearly evident in the torque and tool wear box plots. When analyzing these box plots, we see that there are instances where the torque values deviate significantly from the norm, indicating that the system is experiencing excessive force beyond its usual operating range. This overstrain is not only reflected in the torque measurements but also in the increased wear on the tools, as depicted in the corresponding box plot for tool wear. The tool wear box plot shows higher wear rates and possibly more outliers, suggesting that the tools are subjected to conditions that accelerate their degradation.

Examining the relationship between torque and rotational speed provides further insights into the overstrain conditions. When plotting these two variables against each other, we observe that overstrain is often associated with low variance and lower values of rotational speed. This means that when the system is running at lower speeds, the torque tends to be more consistent but higher, indicating overstrain. The low variance in rotational speed suggests that the system operates at a relatively constant, lower speed during these instances, which could be due to operational constraints or specific tasks that require high torque at lower speeds.

This relationship is crucial because it helps identify the conditions under which overstrain occurs. By understanding that overstrain is linked to lower rotational speeds and consistent but high torque values, we can better monitor and control the operational parameters to prevent excessive wear and potential damage to the tools and machinery. This insight can lead to improved maintenance schedules, operational adjustments, and overall better management of the system to avoid the detrimental effects of overstrain.

Multivariate Analysis
Investigate interactions between multiple features.

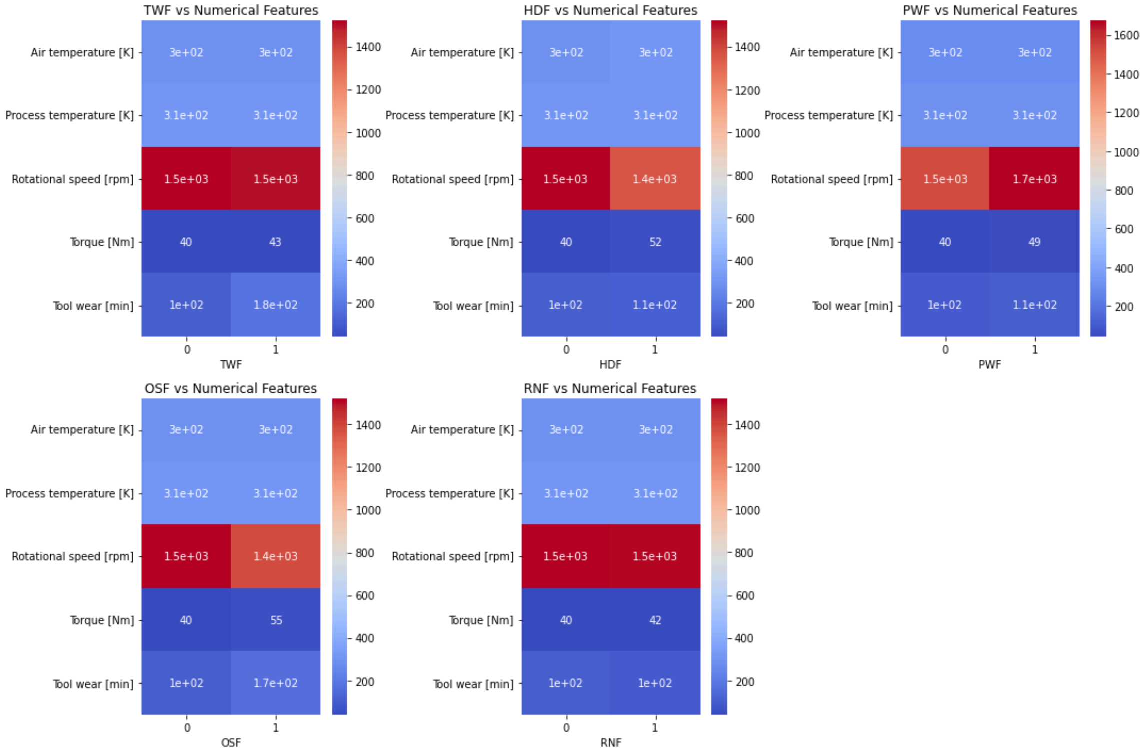


- `Strong positive` correlations were observed between machine failure and specific failure types, including HDF (Heat Dissipation Failure), OSF (Overstrain Failure), and PWF (Power Failure). These failure types are strongly associated with overall machine failure occurrences.
- A `moderate positive` correlation was found between machine failure and TWF (Tool Wear Failure), indicating that tool wear failures contribute to the likelihood of machine failures.
- `Weak positive` correlations were identified between machine failure and features such as Torque [Nm], Air temperature [K], Tool wear [min], and Process temperature [K].

While these correlations exist, their influence on machine failure occurrences is relatively weak.

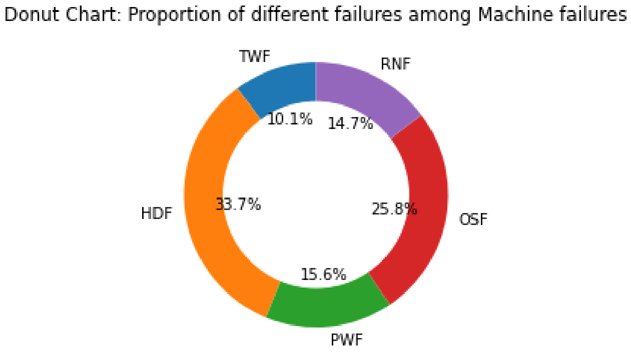
- A `weak negative` correlation was observed between machine failure and Rotational speed [rpm], suggesting that higher rotational speeds may slightly decrease the likelihood of machine failures.

Correlation between failure types and numerical features



- These features likely represent different failure modes in a machine.
- The top left chart titled `TWF vs Numerical Features` shows the distribution of air temperature (on the y-axis) for machine failures categorized as TWF (Tool Wear Failure).
- It is difficult to make specific comparisons between the failure modes (TWF, HDF, PWF, OSF, RNF) because the x-axis titles (Air temperature, Process temperature, Rotational speed, Torque, Tool wear) are not consistent across the charts.

Proportion of different failures among Machine failures.



we can observe the following:

1. TWF (Tool Wear Failure)

- The proportion of `tool wear failures` among machine failures is represented by the corresponding segment in the donut chart. There are a total of 208 tool wear failures.

2. HDF (Heat Dissipation Failure)

- The proportion of `heat dissipation failures` among machine failures is represented by the corresponding segment in the donut chart. There are a total of 701 heat dissipation failures.

3. PWF (Power Failure)

- The proportion of `power failures` among machine failures is represented by the corresponding segment in the donut chart. There are a total of 320 power failures.

OSF (Overstrain Failure)

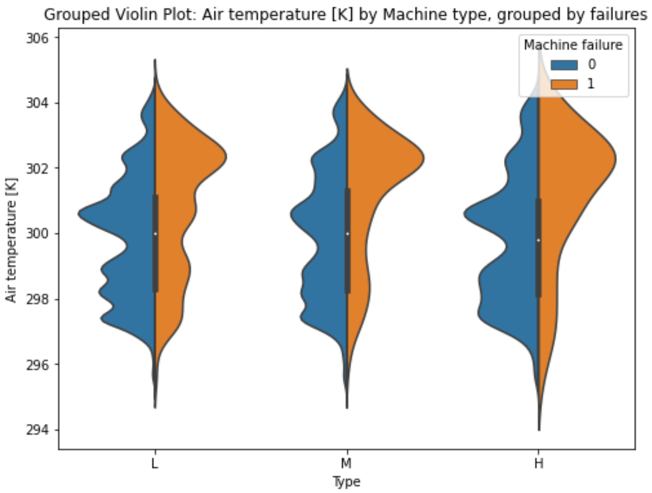
- The proportion of `overstrain failures` among machine failures is represented by the corresponding segment in the donut chart. There are a total of 533 overstrain failures.

RNF (Random Failure)

- The proportion of `random failures` among machine failures is represented by the corresponding segment in the donut chart. There are a total of 306 random failures.

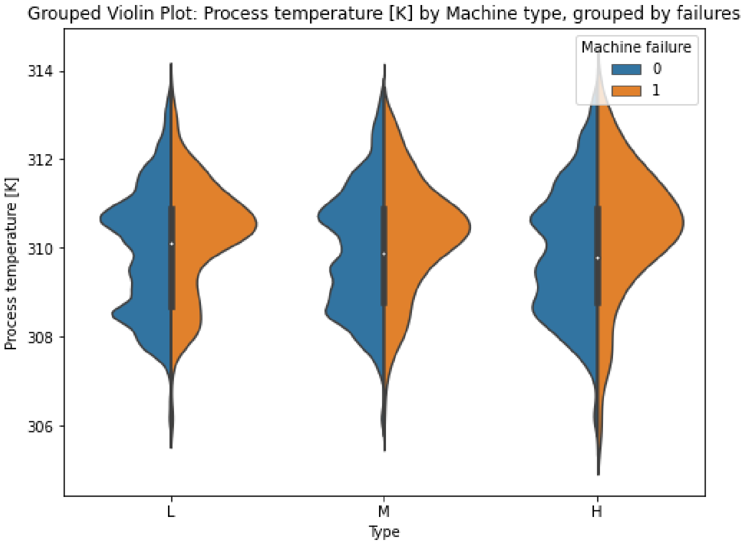
Distribution of Tool wear duration by Machine type, grouped by different failures.

1. Air temperature [K]



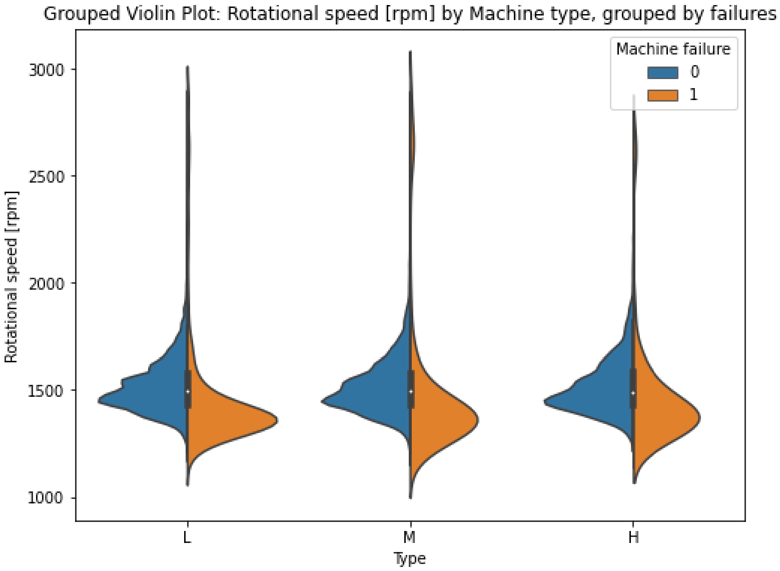
2. Process temperature [K]

- The distribution of air temperature for machines that have experienced `failure` tends to be `higher` compared to those that have not. This pattern is observed across different machine types. This could suggest that higher air temperatures are associated with machine failure, regardless of the machine type.



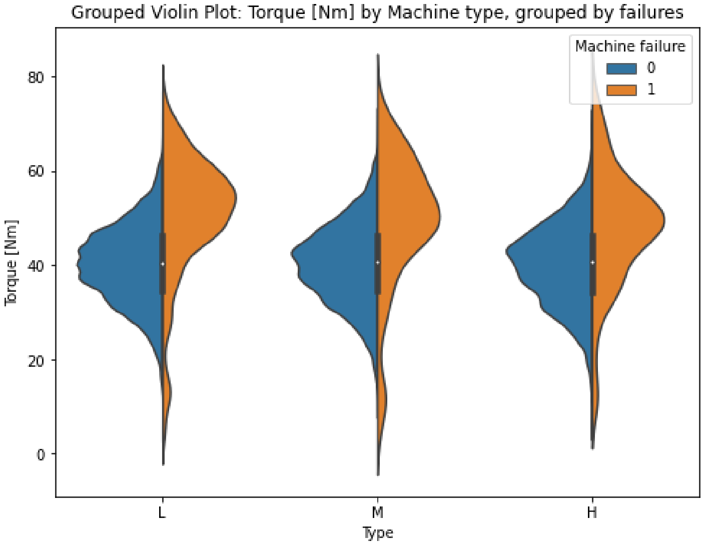
- The distribution of process temperature for machines that have experienced `failure` is also `higher`, across different machine types. This could indicate that higher process temperatures are associated with machine failure, regardless of the machine type.

3. Rotational speed [rpm]



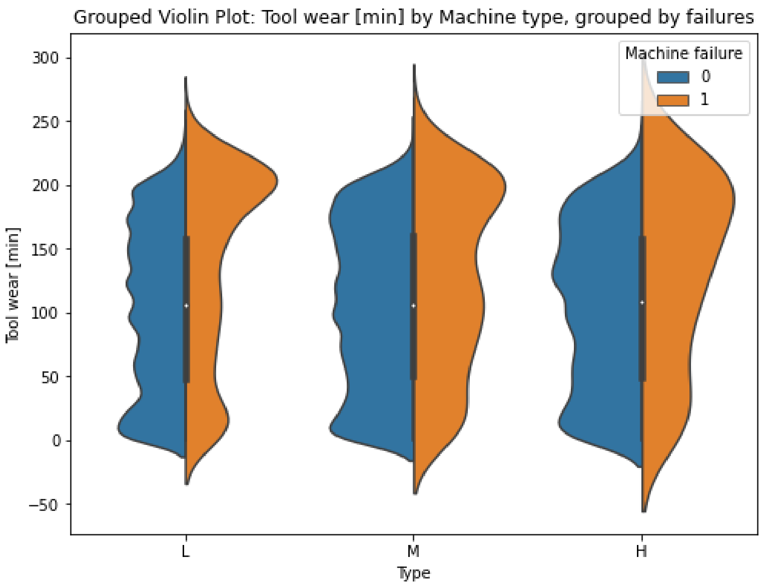
- The distribution of rotational speed for machines that have experienced `failure` appears to be `lower`, across different machine types. This could suggest that lower rotational speeds are associated with machine failure, regardless of the machine type.

4. Torque [Nm]



- The distribution of torque for machines that have experienced `failure` is `higher`, across different machine types. This could suggest that higher torque is associated with machine failure, regardless of the machine type.

5. Tool wear [min]



- The distribution of tool wear for machines that have experienced `failure` is `higher`, across different machine types. This could suggest that longer tool usage is associated with machine failure, regardless of the machine type.

Our analysis of the dataset reveals critical insights into machine operations and failures. Higher air and process temperatures, as well as increased torque and tool wear, are associated with machine failures across different machine types. Conversely, lower rotational speeds are linked to higher failure rates. These findings highlight key factors influencing machine reliability and can guide targeted maintenance and operational strategies to mitigate failures.