Dimensions
The dataset contains a total of 10 column variables namely:

1. Observation number
2. Product ID
3. Type
4. Air Temperature
5. Process Temperature
6. Rotation Speed
7. Torque in Newton Meters
8. Tool wear in Minute
9. Target
10. Failure Type

Character Variables
On understanding the data types of the columns included in the dataset, there are 4 character variable, which are follows:

1. Product ID
2. Types
3. **Target**
4. Failure Type

We also understand that Product ID is the unique column of the dataset, which can be later used to join tables with a similar kind of data. Therefore this checkpoint can be used later for an analysis.

Failures
The target variable of the target contains 0's and 1's, no failure and failure respectively. A deeper analysis into the failure type reveals that there are 2 main failures::

1. Failure
2. No failure

Types of Failures
We can understand from this that there are 5 different types of failures:

1. Heat Dissipation Failure
2. Overstrain failure
3. Power Failure
4. Random Failure
5. Tool Wear Failure

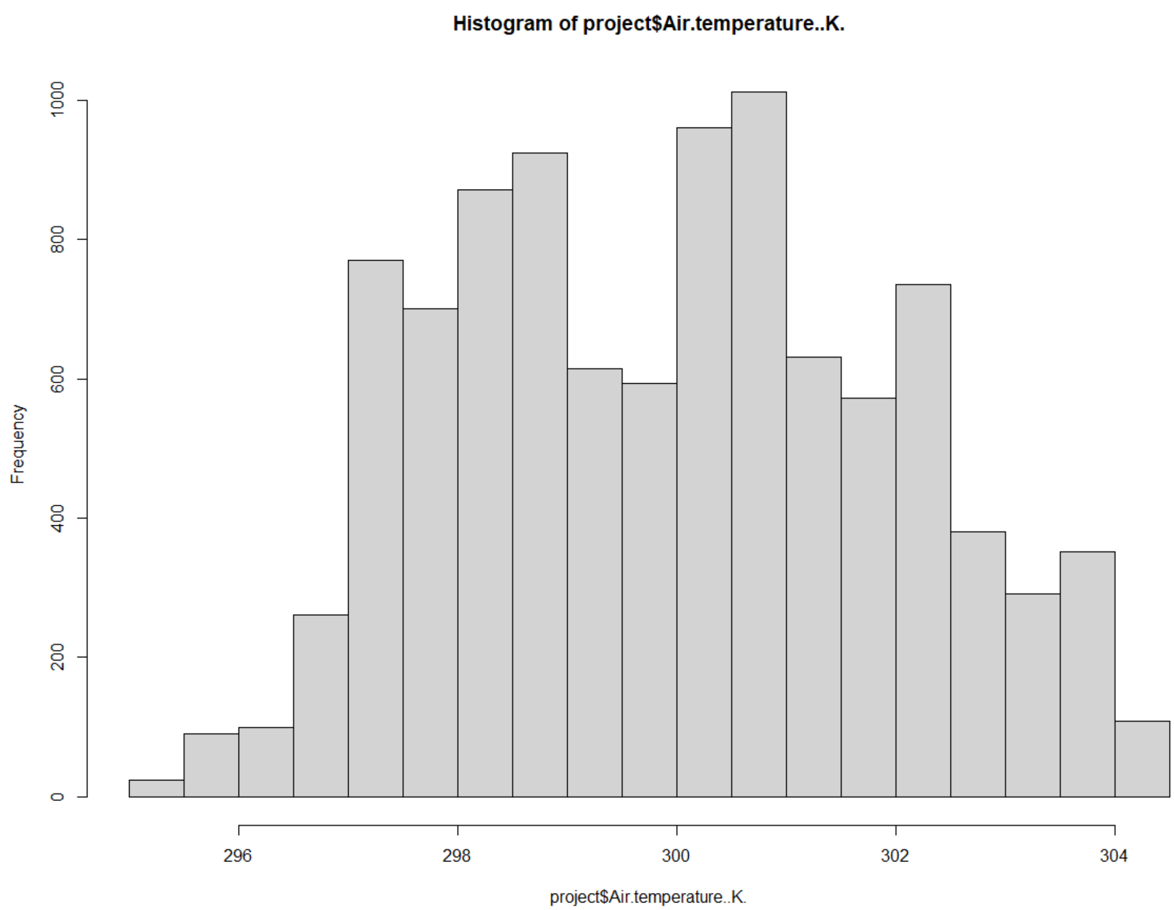
Category Types
On converting the category_type of each observation we identify that there are 3 distinct types of machine types which may be realized as part types as per work instruction; they are as follows:

1. H
2. L
3. M

Histograms

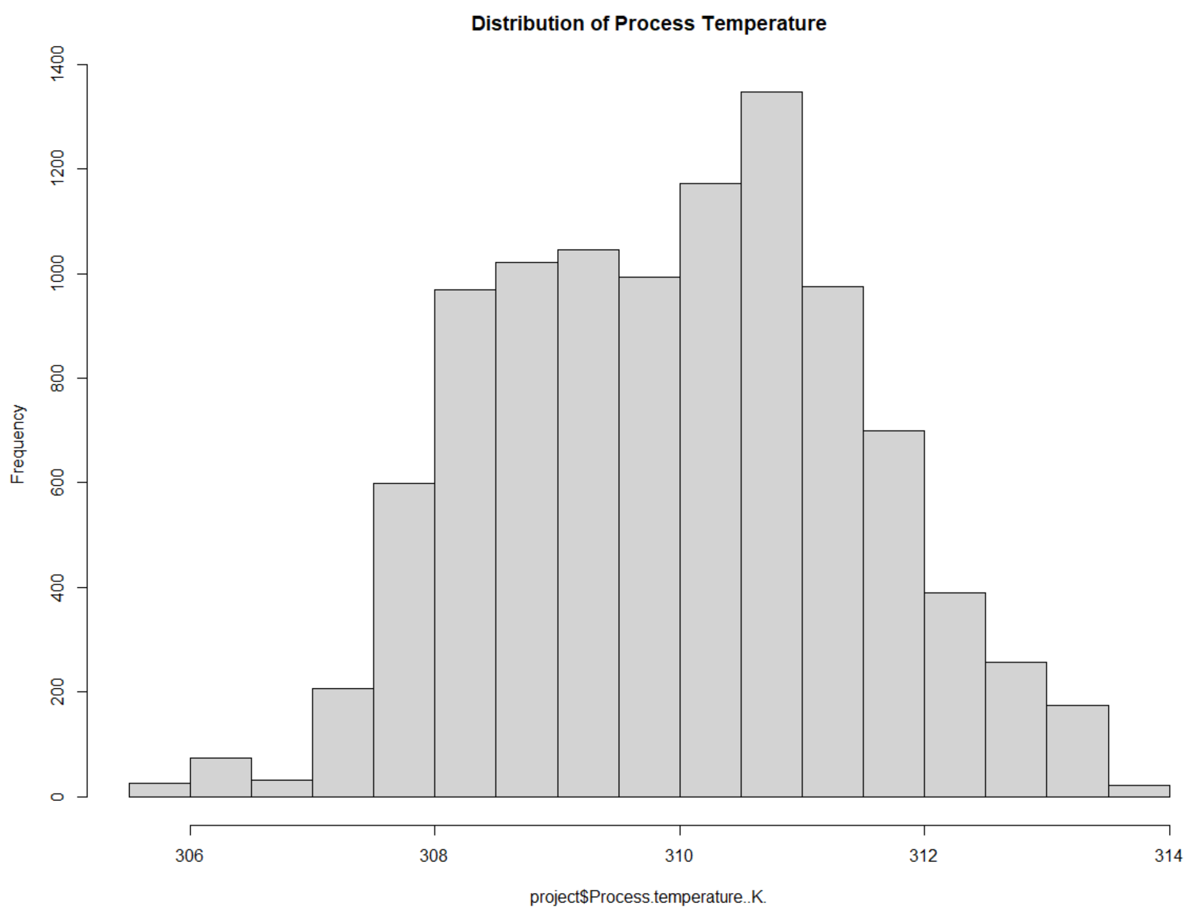
A quick summary statistics on the dataset(summary on R) for understanding hte dataset better, the numerical variables are as follows:

1. Air Temperature



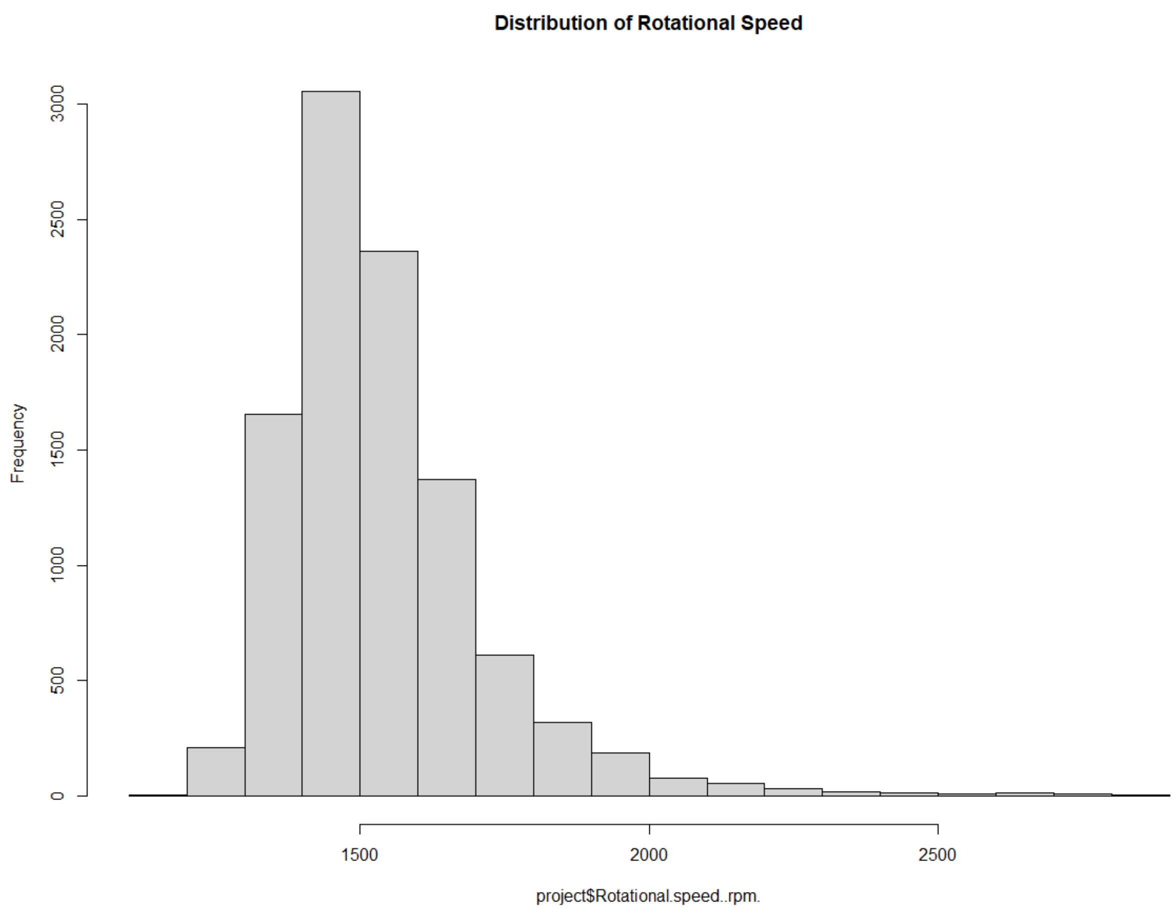
From the histogram we understand that the air temperature data is spread in quite a normal distribution, but has a slight skew towards the left. Would require standardization.

2. Process Temperature



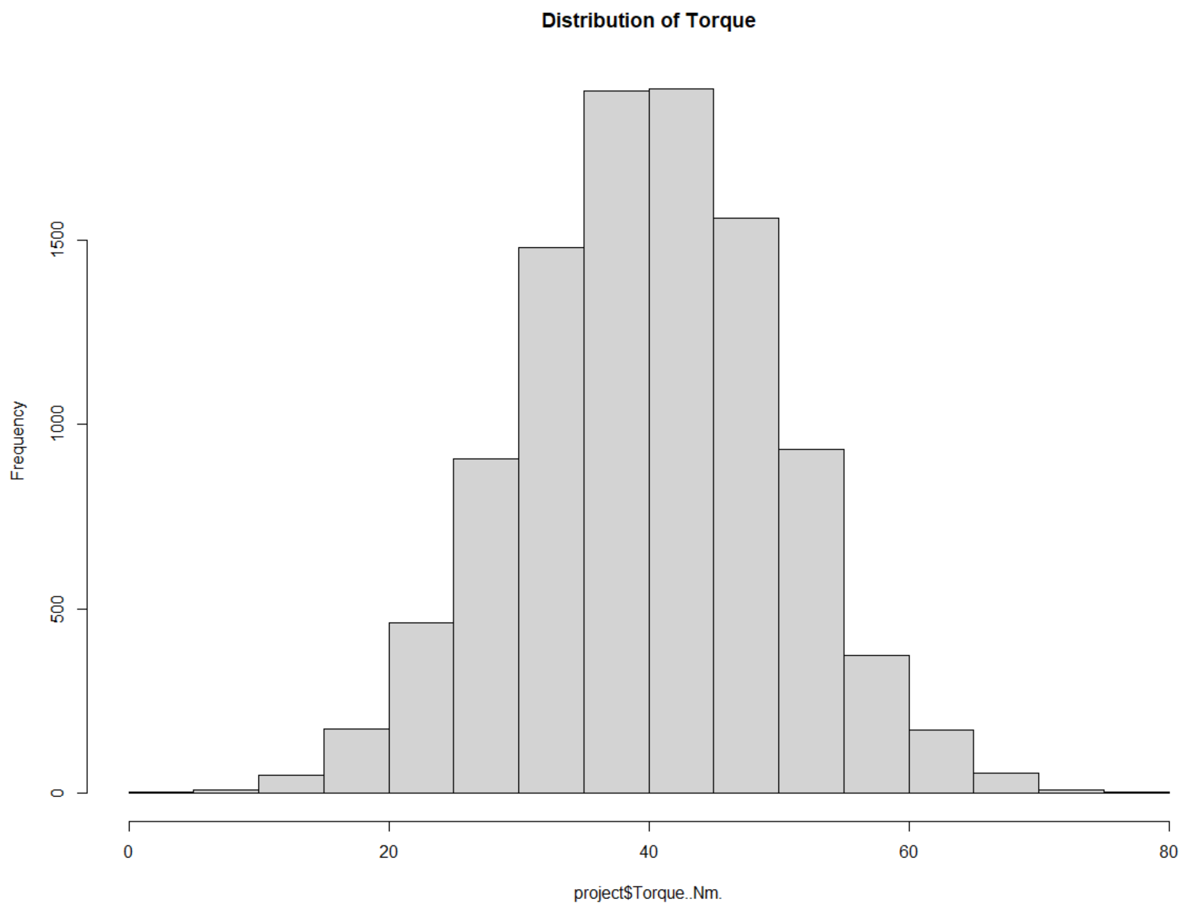
The histogram of the process temperature has a slight skew towards rhe left, with a higher average process temperature towards the maximum end of the temperature. A The mean of process temperature is around 310.0 in Kelvin i.e. towards the end of the distribution.

3. Rotational Speed in Rotations per minute



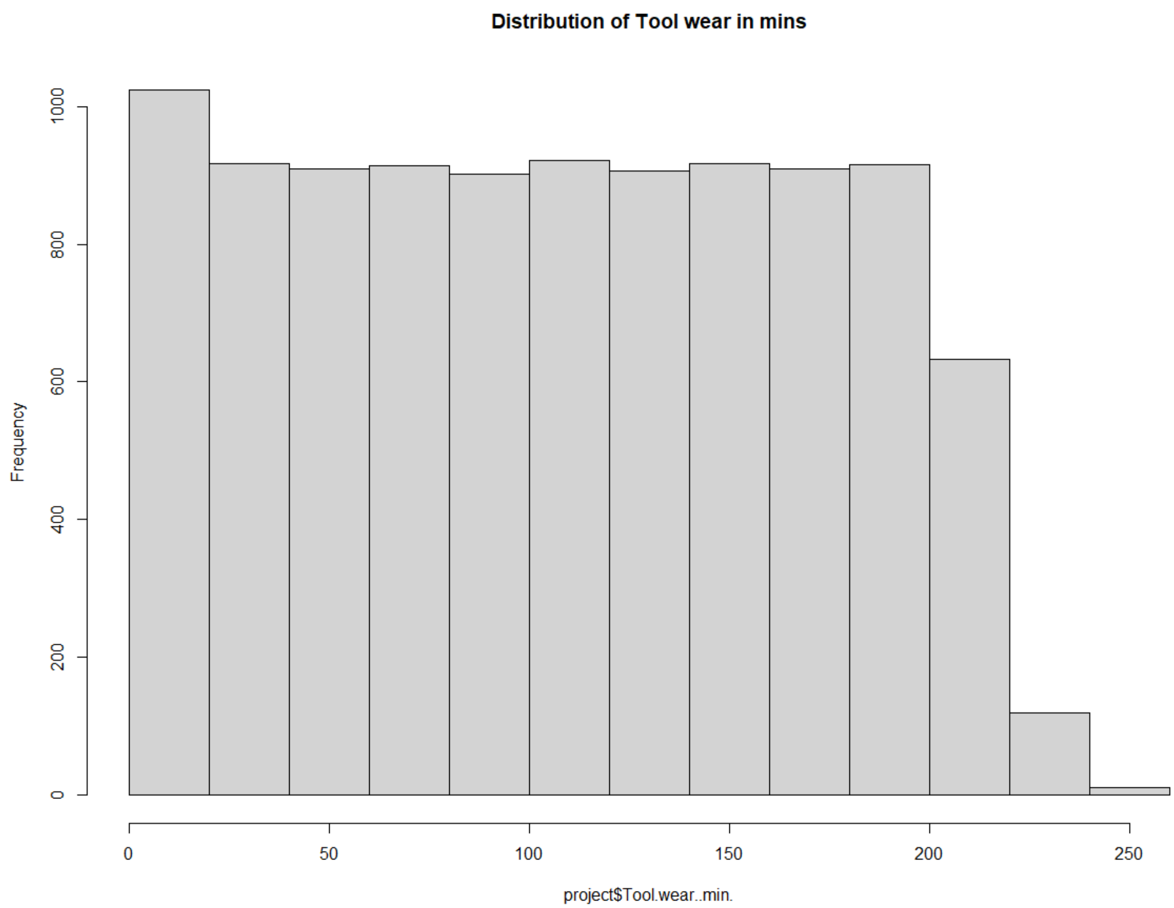
The histogram of the rotational speed depicts that the histogram has an extreme right skew, which means that the mean is closer towards the start of the distribution, the summary statistics confirm and reveal that the mean is 1539 N/m.

4. Torquie in Newton Meter



The torque data is almost normally distributed.

5. Tool Wear in Minutes



The tool wear almost represents a uniform distribution with high values just towards the maximum.

Correlations

```
> cor(project[sapply(project, is.numeric)])
                           ï..UDI Air.temperature..K. Process.temperature..K. Rotational.speed..rpm.  Torque..Nm. Tool.wear..min.
ï..UDI                  1.000000000          0.11742795              0.32442815          -0.0066148683  0.003206586   -0.0107020066
Air.temperature..K.     0.117427946          1.00000000              0.87610716           0.0226704588 -0.013777823    0.0138528277
Process.temperature..K. 0.324428145          0.87610716              1.00000000           0.0192767139 -0.014060613    0.0134875171
Rotational.speed..rpm. -0.006614868          0.02267046              0.01927671           1.0000000000 -0.875027086    0.0002230848
Torque..Nm.             0.003206586         -0.01377782             -0.01406061          -0.8750270863  1.000000000   -0.0030927814
Tool.wear..min.        -0.010702007          0.01385283              0.01348752           0.0002230848 -0.003092781    1.0000000000
```

Pair of variables with a high coefficient of correlation(can cause homoskedasticity in the data)

1. Air temperature and **process temperature**
2. **Torque** and Rotational Speed

From domain knowledge, this implies that:

a. When any machine is affected by changes in the ambient temperature of the factory or the surrounding working environments, friction is caused due to motor heating and mechanical movement as well as cutting heat and cooling medium, all of which causes temperature rise in various parts of the machine tool, resulting in changes in working efficiency and the accuracy of the machine tool, thereby rendering it out of use sooner.

b. A high correlation between variables is also possible when either one is a linear function or is related to another dimension in the dataset. Torque is the rotational equivalence of linear force. The relation between torque and speed are inversely proportional to each other. The torque of a rotating object can be mathematically written as the ratio of power and angular velocity.

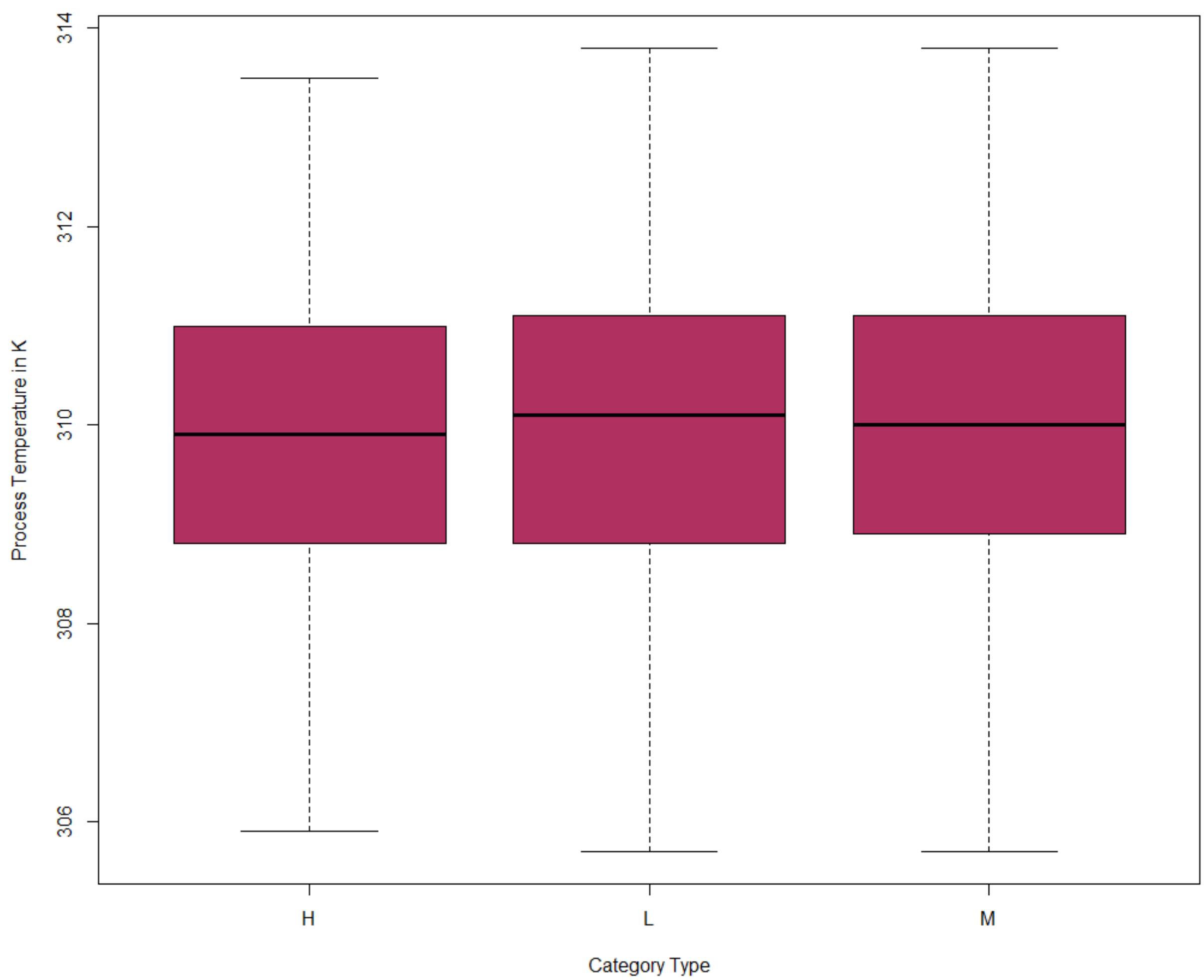
Replacing a pair of 2 highly correlated variables can be easily resolved by using either one of the pair of the dimensions.

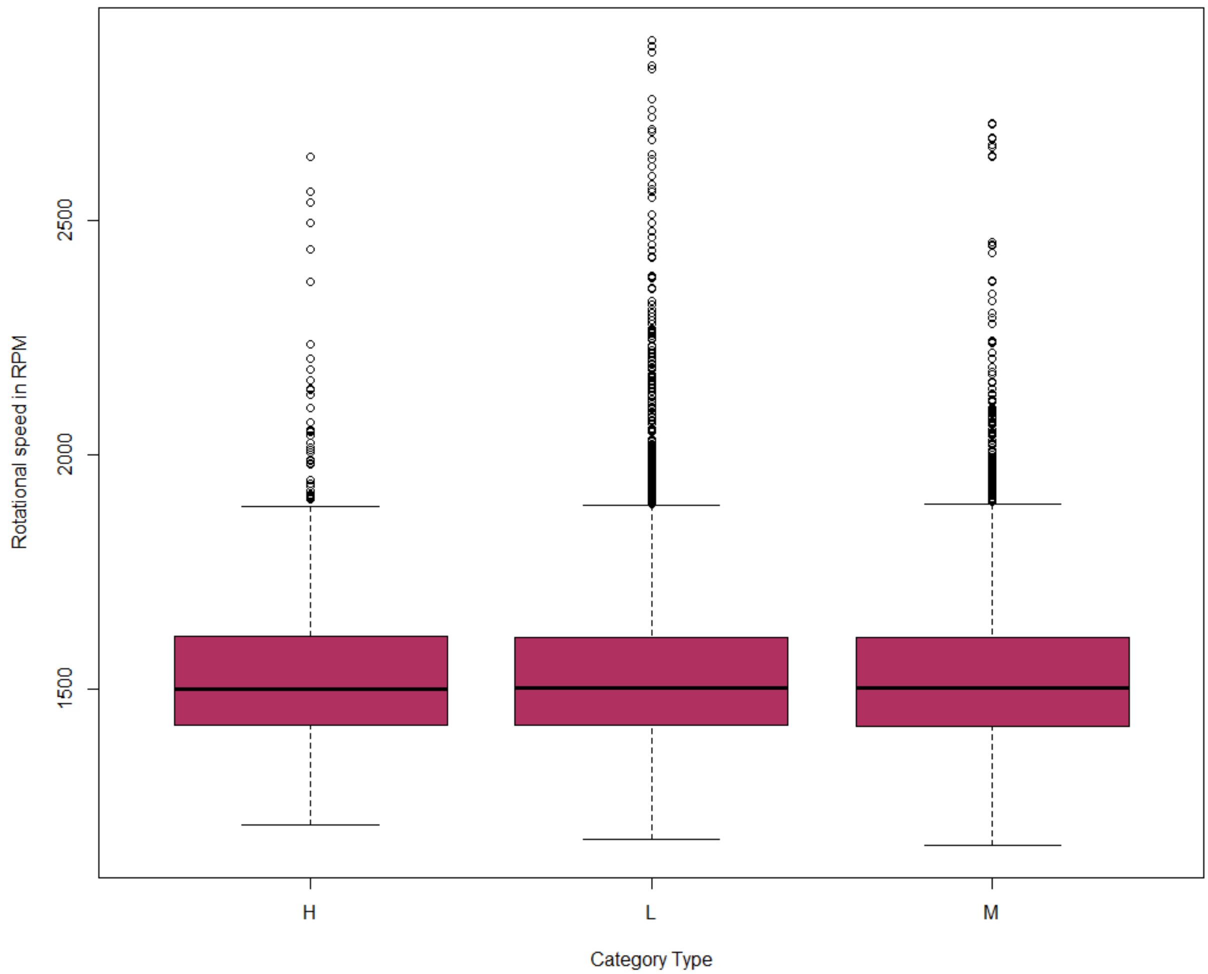
Summary statistics of the variables are as follows:

```
> summary(project)
    ï..UDI        Product.ID           Type           Air.temperature..K. Process.temperature..K. Rotational.speed..rpm.
 Min.   :    1   Length:10000       Length:10000       Min.   :295.3       Min.   :305.7           Min.   :1168
 1st Qu.: 2501   Class :character   Class :character   1st Qu.:298.3       1st Qu.:308.8           1st Qu.:1423
 Median : 5000   Mode  :character   Mode  :character   Median :300.1       Median :310.1           Median :1503
 Mean   : 5000                                         Mean   :300.0       Mean   :310.0           Mean   :1539
 3rd Qu.: 7500                                         3rd Qu.:301.5       3rd Qu.:311.1           3rd Qu.:1612
 Max.   :10000                                         Max.   :304.5       Max.   :313.8           Max.   :2886
  Torque..Nm.    Tool.wear..min. Target    Failure.Type
 Min.   : 3.80   Min.   :  0     0:9661   Length:10000
 1st Qu.:33.20   1st Qu.: 53     1: 339   Class :character
 Median :40.10   Median :108              Mode  :character
 Mean   :39.99   Mean   :108
 3rd Qu.:46.80   3rd Qu.:162
 Max.   :76.60   Max.   :253
```

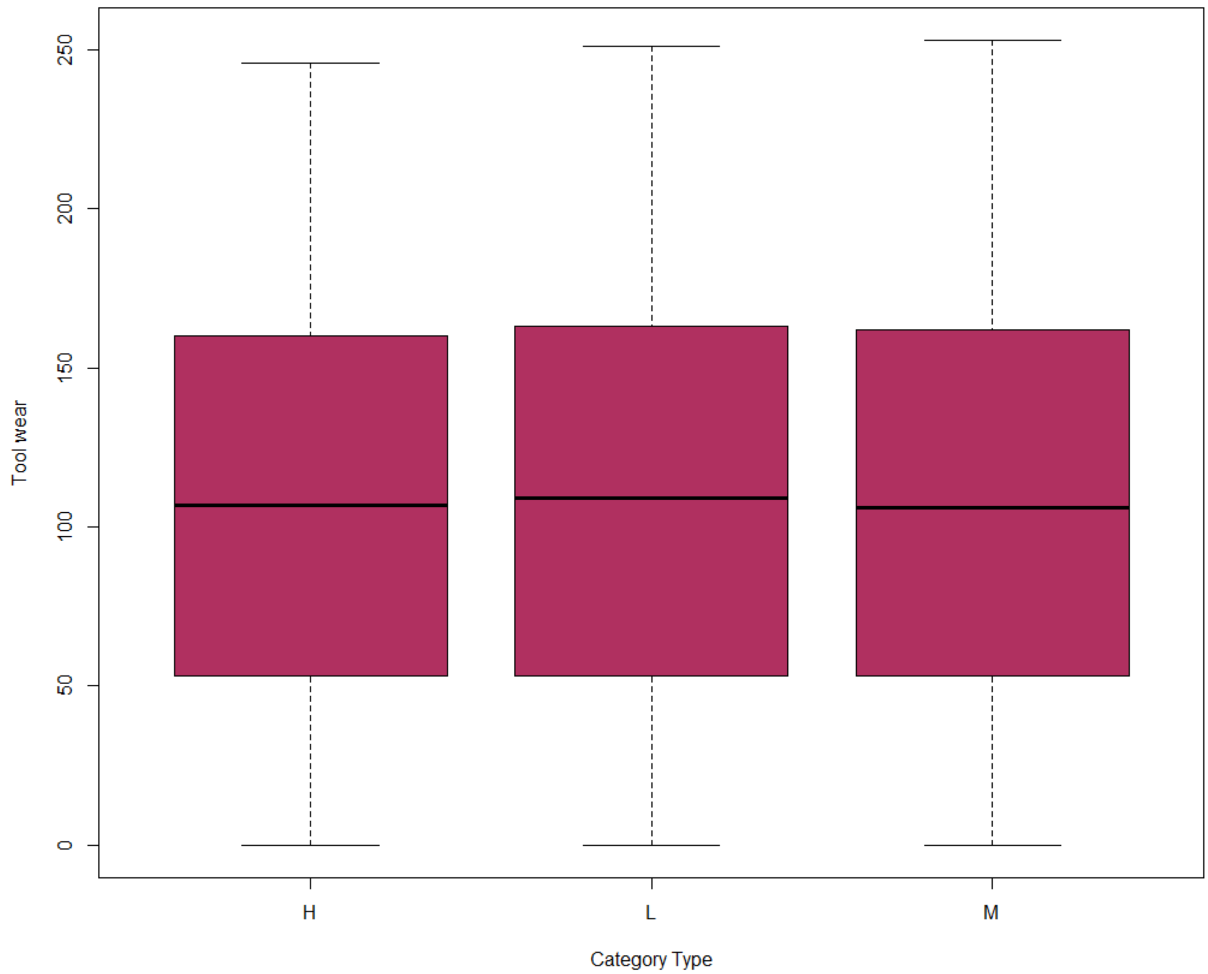
Comparing the behavior of the machine with respect to the category type and the other parameters.

Important boxplots



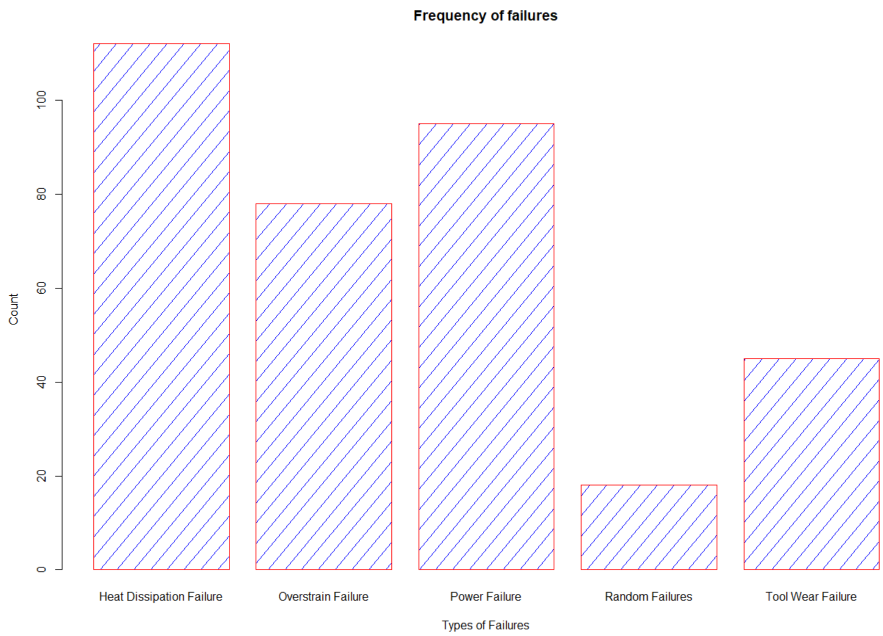


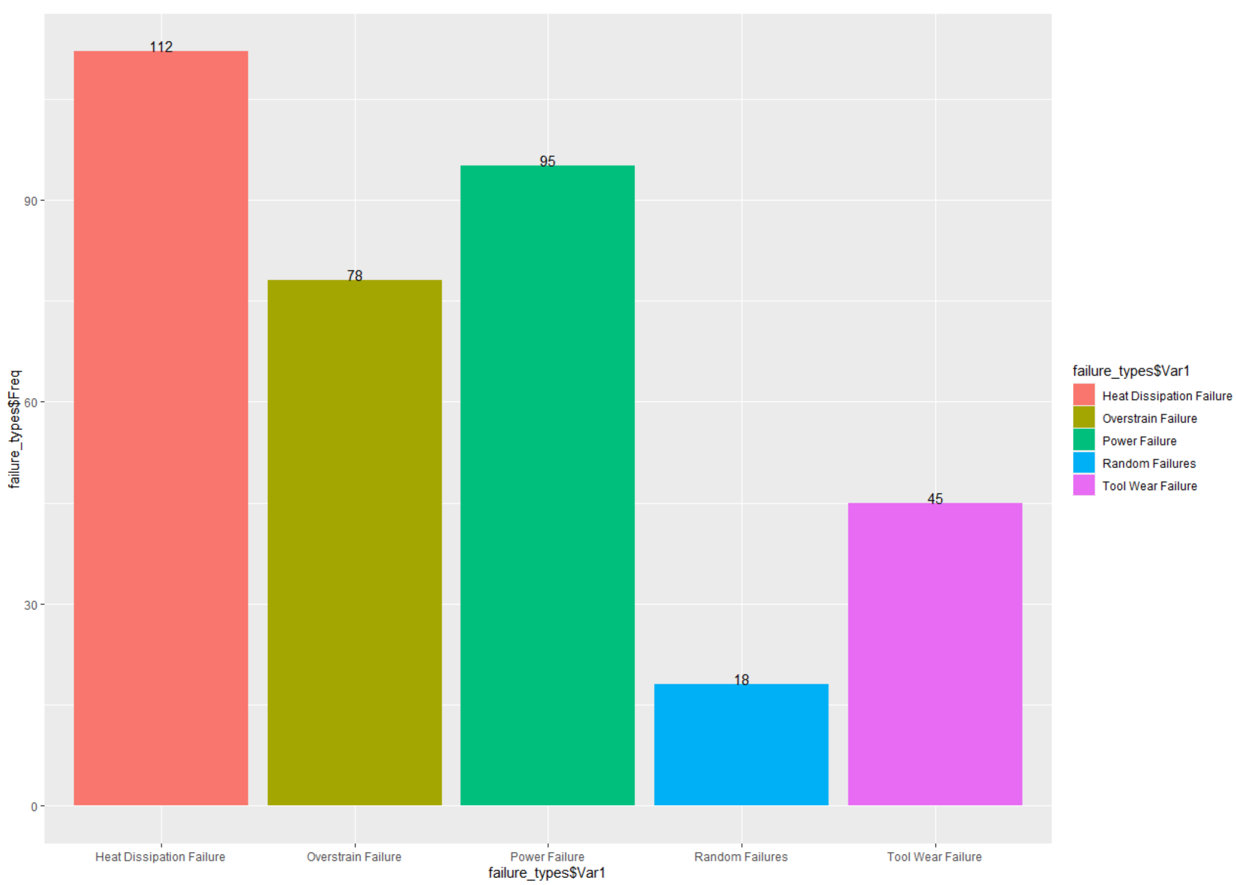
Category type L has a higher number of outliers as compared to H and M. Which means that L is used for processes with a higher rotational speeds, followed by M and then H.



Tool wear
0
50
100
150
200
250
H
L
M
Category Type

Failure Types in Detail





The types of failure are ranked as follows:

1. Heat Dissipation Failure
2. Power Failure
3. Overstrain Failure
4. Tool wear failure
5. Random Failure

A detailed explanation of the types of failure types:

a. **Heat dissipation** is a type of heat transfer. This type of failure occurs when an object that is hotter than other objects is placed in an environment where the heat of the hotter object is transferred to the colder objects to the colder objects and the surrounding environment. Heat dissipation is unavoidable and sometimes is considered a negative process in a range of industries and applications. Temperature difference is the factor that determines the mode and rate of heat transfer that in a given application.

b. **Power failure** is a short or long term loss of the electric power to an area. In more specific terms it is period of time when the electricity supply to a particular abuilding or area is interrupted, which might be due to the damage of the cables included in the machine.

c. **Overstrain failure** is a weighted average of the tool wear and torque. If the product of tool wear and torque exceeds a certain amount i.e. 11,000 min/Nm the process fails due to overstrain.

d. **Tool Wear** can be caused due to various other factors such as mechanical. Corrosive and adhesive. Wear is a process that place through contact. As a result of this wear, particles of material detach and the created debris quickens it. Another thing to keep in mind is that the wear mechanisms vary. They are both chemical and physical. The reasons are classified as mechanical, corrosive and adhesive.

e. **Random Failure** is a failure that occur at not set time. It can occur randomly at any time during a process. A failure that is known to happen but not when it will happen.

Principal Component Analysis

Satisfying the regulations for PCA by removing the character variables from the dataset. The following variables were removed from the dataset, which are as follows:

a. IDs
b. Product ID
c. Category Type
d. Failure Type

```
> pcs <-  prcomp(na.omit(project[,-c(1,2,3,10)]))
> pcs
Standard deviations (1, .., p=6):
[1] 179.4963440  63.6541633   4.8203457   2.4182458   0.5916080   0.1692858

Rotation (n x k) = (6 x 6):
                                 PC1           PC2           PC3           PC4           PC5           PC6
Air.temperature..K.       2.525001e-04  0.0004357414 -0.0067319628  0.8144928048 -5.797778e-01  0.0203382948
Process.temperature..K.   1.593153e-04  0.0003147123 -0.0028443178  0.5801091603  8.142834e-01 -0.0201910096
Rotational.speed..rpm.    9.988167e-01 -0.0001158906 -0.0486262936 -0.0006471459  1.088803e-04  0.0005271211
Torque..Nm.              -4.863269e-02 -0.0004507068 -0.9987194777 -0.0071673092  1.919591e-03  0.0117901697
Tool.wear..min.           9.368770e-05  0.9999997023 -0.0004483729 -0.0005422112  5.707974e-06  0.0003025412
Target                   -4.496492e-05  0.0002998181 -0.0118860375  0.0047696879 -2.822376e-02 -0.9995195343
> summary(pcs)
Importance of components:
                            PC1     PC2     PC3     PC4     PC5    PC6
Standard deviation     179.4963 63.6542 4.82035 2.41825 0.59161 0.1693
Proportion of Variance   0.8876  0.1116 0.00064 0.00016 0.00001 0.0000
Cumulative Proportion    0.8876  0.9992 0.99983 0.99999 1.00000 1.0000
> 
```

According to the PCA analysis we can understand that PC1 explains 88.76% of the total variance, which means that nearly four fifths of the information in the dataset can be epsulated by just one principal component. PC2 explains 99.92% of the variance.

Classification based on the **target** variable

Other variables to be removed include the following:

1. ï..UDI
2. Product.ID
3. Failure.Type
4. Air temperature and **process temperature**
5. **Torque** and Rotational Speed: Torque and speed are inversely proportional to each other. The torque of a rotating object is the ratio of power and angular velocity. In our dataset, Rotational…speed…rpm and Torque..Nm columns are highly correlated as well.

   Torque = Power/Speed

   Creating a new column

6. Torque is the rotational equivalence of linear force. Speed measures the distance covered in unit time. The relation between torque and speed are inversely proportional to each other. The torque of a rotating object can be mathematically written as the ratio of power and angular velocity.

10, 6 7