Exploratory analysis for achieving optimal environmental and operational parameter settings for making quality crossmember castings.

Abstract:

Crossmember is a structural component that undergoes strict X-ray inspection to ensure its quality. Porosity is the major defects that are associated with this part. Based on the product conditions and related scrap rates data collected by FCA, with additional local humidity and temperature records, an exploratory data analysis has been conducted, determine the statistical properties of each factor. The optimal environmental and operational parameter settings are identified for making quality crossmember castings through a novel optimization algorithm.

Introduction:

The front crossmember is a structural element in front of the vehicle that acts like a structural link between driver and passenger side, also providing mounting locations for suspension components. (skip this part for now)

Design of experiment

Data description:

The data used in this research is provided by FCA, containing product conditions and related scrap rates. There are files in original data, alloy chemistry, and other 3 op (figure this out). According to timestamps of each data point, a single file is compiled consisting of 500,000 data points and over 70 variables (here cites Yang’s paper). Besides timestamps, these variables incorporate part number furnace type, melt number, alloy compositions, parameters for vacuum, die lube, cooling water, air pressure, shot profiles, biscuit length, cycle time, cavity number, and X-ray examination results. Among all variables, there are 29 are considered to have important impact on the internal quality of the casting. These variables are listed in table1. (a table here)

Note that within the crucial 29 variables, those related to alloy composition are not adjustable. Thus, this analysis only focuses the contribution of the other controllable variables to the passing rate.

In addition to the compiled data collected by FCA, the data indicating the local environmental condition are employed. The humidity and temperature information of Toronto airport area is extracted from (dataset) from Jan 4th, 2015 to Jul 31st, 2015, and added to each data point in the aforementioned dataset. The time interval in the environmental condition data is an hour, so between integer hours, all the data points share the same humidity and temperature. The passing rate is denoted by the ratio of number of passed and the total number.

Data preprocessing: (outlier detecting and taking out)

First look at data importance. (random forest feature importance)

Prior to analysis, a data cleaning operation is needed. There are two parts in this process:

1. taking out data points that contain NA values.

2. conducting an outlier detection and removal operation.

Due to the incompleteness in the process of collecting data, there are some data points containing NA values. These data points are taken out so that the analysis can be performed using relatively complete information in the data.

The outlier detection and a removal operation are conducted based on the results of boxplots of the first 5 most important variables. They are intensification stroke, intensification rise time, intensification pressure, biscuit length, and fill time. A sample is considered an outlier when its value of any of those variables is out of 1.5 times interquartile range (IQR) of the corresponding boxplot. The boxplots of the original data points for those 5 variables are shown in Figure 1.

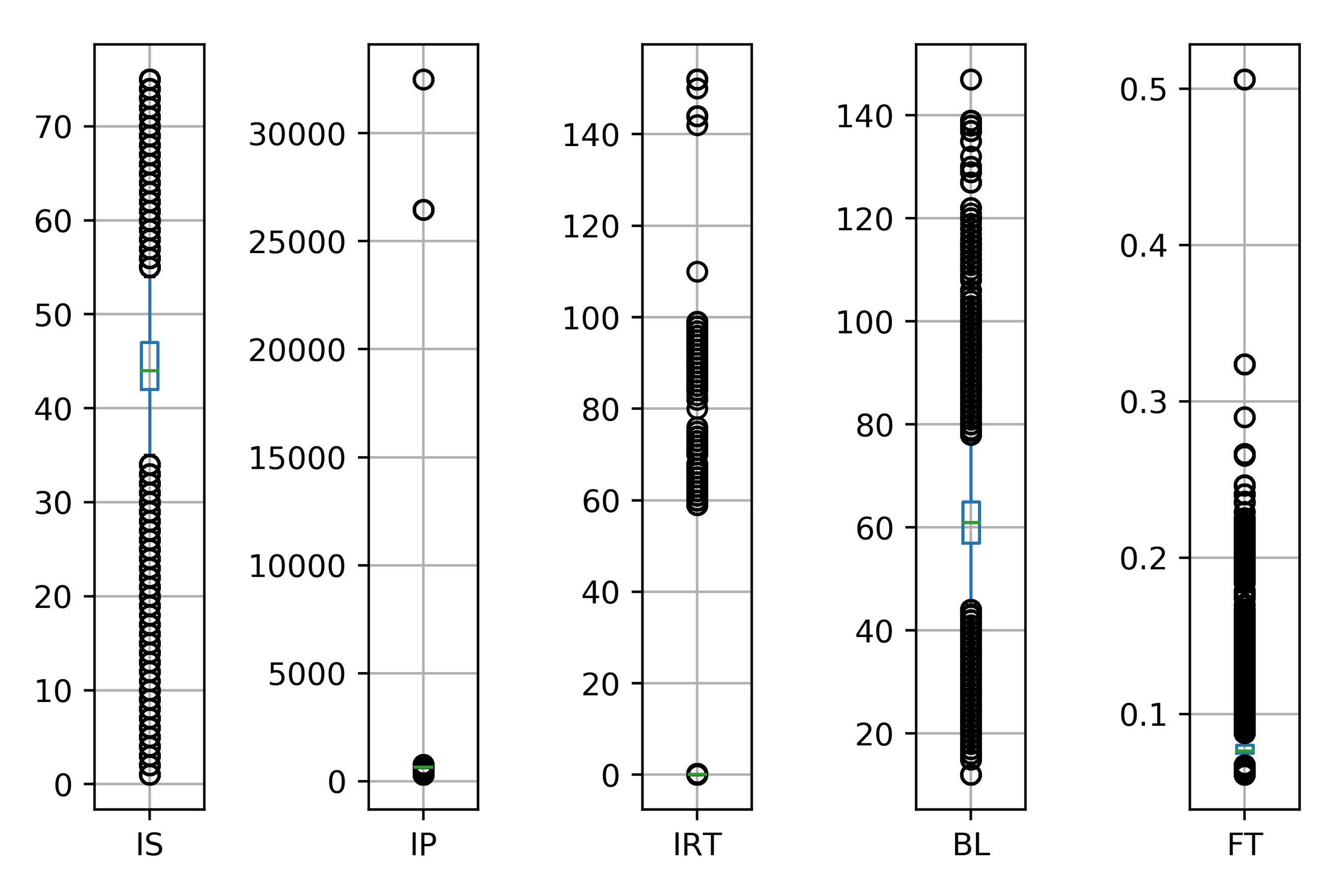


Figure 1 - Boxplots of intensification stroke, intensification rise time, intensification pressure, biscuit length, and fill time from the raw data

The black circles, which are values exceeding the 1.5IQR boundaries in the boxplots, are considered outliers for each factor. For intensification stroke and biscuit length, the extreme values exist on both sides. The extreme values in intensification pressure, fill time, and intensification rise time are mostly extremely large values. They cause sever oscillations in passing rate and variable plots. A data point will be fully taken out when it is considered an outlier. After data cleaning, the number of data points becomes 250,866.

The next step is to convert categorical results into numerical. The examination results for all the data points are labeled as “pass/fail”. In order to make the data fit for statistical models and analyses, the “pass” is replaced by the number 1, while “fail” is replaced by the number 0. The passing rate based on the cleaned data is (passing rate)

All the operation and manipulation are done by using Pandas in Python programming language.

Exploratory analysis (boxplot, scatterplot…. Quantitively and graphically)

The purpose of this analysis is to find correlations between certain features and the passing rate, so that we can limit the values of those features within some range to achieve the highest passing rate.

The effect of humidity and temperature:

In this research, the local environmental information has been taken into account. The histogram plots of humidity and temperature for the entire dataset are shown in Figure 2.

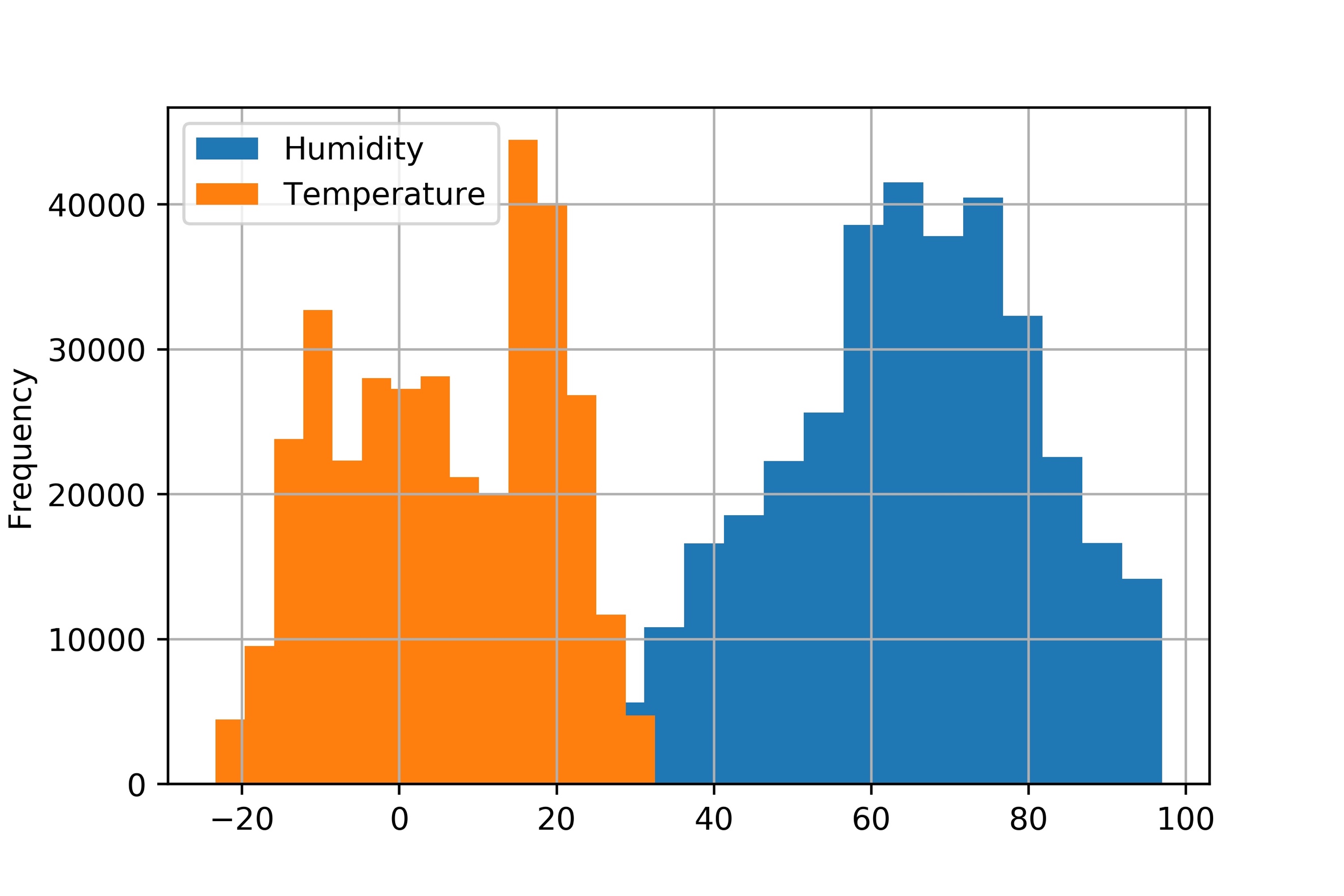


Figure 2 - Histogram plots of humidity and temperature.

The histogram plots show that both humidity and temperature do not have outliers in the data. The spread ranges for both features coincide with the fact. It indicates the humidity and temperature data is reliable and appropriate for analysis. The humidity data is nearly symmetrical and normally distributed, implying the operation happened at humidity of 60% to 80% the most. Meanwhile, the temperature data has two peaks or modes, resulting from the lasting time of winter and summer is longer than spring and fall.