Introduction

In today's world, data analysis is an essential aspect of every business operation. Companies need to keep track of various parameters such as sales, production, and marketing, to stay ahead of the competition. In this context, the analysis of production data is crucial for any business to improve efficiency and plan for future production. Production data provides insights into the various aspects of the manufacturing process, including product quality, labor efficiency, and production speed. This information can help businesses optimize their production process, reduce costs, and improve their bottom line.

In this article, we will explore a dataset that consists of 118 files, representing actual production data for a period of 6 months. Each file corresponds to a single day's production, providing detailed information on various parameters such as Sales Order, Line Item, Standard Minute Value, Product, and Efficiency. The dataset provides a wealth of information that can be used to analyze production trends, identify patterns, and make informed decisions.

Sales Order (S/O) and Line Item (LI) are two critical parameters in the dataset. The combination of S/O and LI forms a unique key that distinguishes one product from another. This information can help businesses identify which products are popular, which products are not selling well, and which products need to be improved. Another essential parameter in the dataset is the Standard Minute Value (SMV), which represents the standard time taken to finish a particular product. This information can be used to optimize the production process, reduce production time, and increase efficiency. Additionally, the dataset provides information on product styles, which can be used to identify popular styles and improve less popular styles.

Efficiency is another crucial parameter in the dataset, which represents the standard hours divided by work hours. This information can be used to analyze the efficiency of the production process and identify areas that need improvement. By analyzing efficiency data, businesses can optimize their production process and increase their bottom line. The dataset also provides information on production trends over time. By analyzing production data for a specific period, businesses can identify patterns, such as which products are selling well during specific months or seasons. This information can help businesses adjust their production process to meet the demand for specific products.

Furthermore, by analyzing production data, businesses can identify areas of inefficiency in the production process. For example, if a particular product is taking longer to produce than others, this information can be used to identify bottlenecks in the production process and optimize it. Similarly, if a particular product is of lower quality than others, this information can be used to identify areas where improvements can be made.

In conclusion, production data analysis is a crucial aspect of any business operation. The dataset we have discussed in this article provides a wealth of information that can be used to analyze production trends, identify patterns, and make informed decisions. By analyzing this data, businesses can optimize their production process, reduce costs, increase efficiency, and improve their bottom line.

Dataset

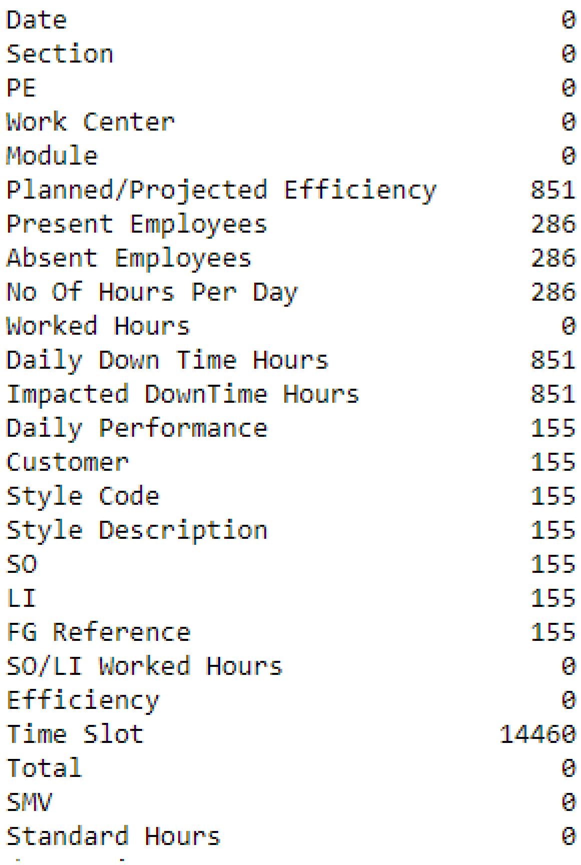
The data is contained in 118 files, with each file containing planning data for a specific time period and line section. The dataset consists of following 25 columns

- Date
- Section
- PE
- Work Center
- Module
- Planned/Projected Efficiency
- Present Employees
- Absent Employees
- No of Hours per Day
- Worked Hours
- Daily Down Time Hours
- Impacted Downtime Hours
- Daily Performance

- Customer
- Style Code
- Style Description
- SO
- LI
- FG Reference
- SO/LI Worked Hours
- Efficiency
- Time Slot

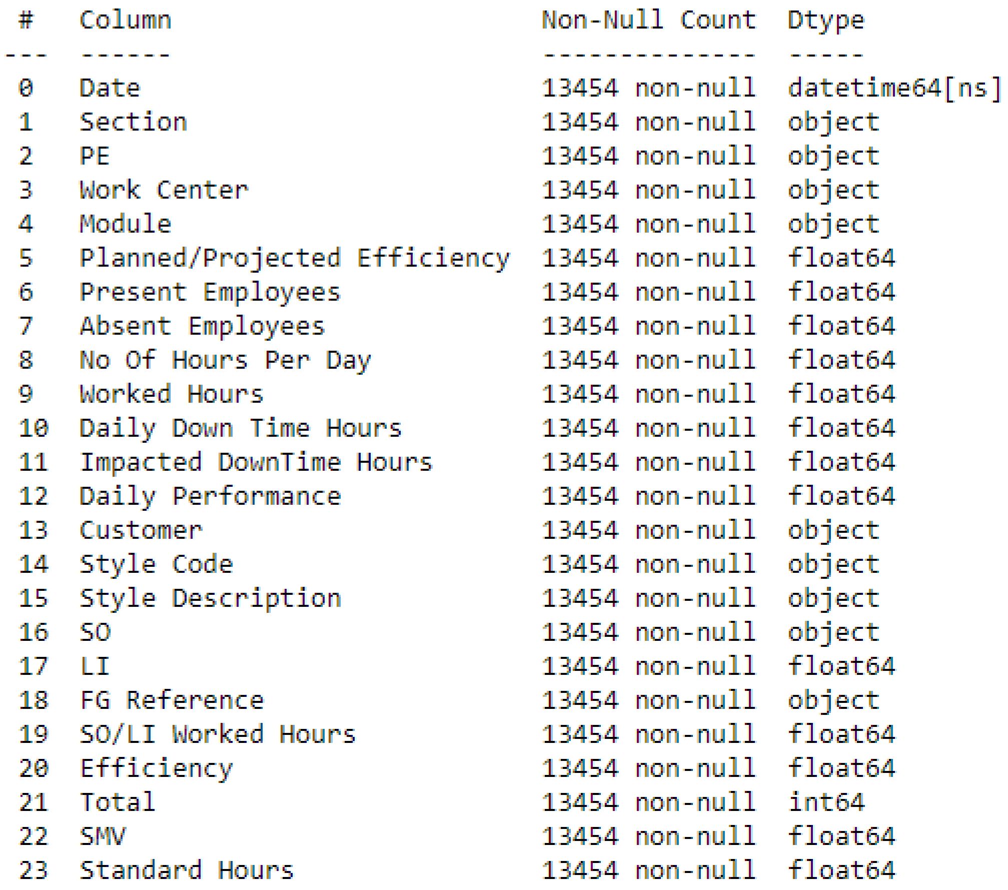
Information of Dataset

Initially, the dataset contains 14460 instances of 25 columns. If we look at information of dataset then figure 1 show the information of dataset columns containing null values. Highest number of null values were found in “Time Slot” it has 14460 which equal to all row has null value which mean this column doesn’t have single value in each row so we drop this column from dataset. Figure 2 show the type of information columns hold mean its type whether its object type or float or int type. There were 8 column that have type object and 15 columns are of type float and one is of datetime type.



```
Date                              0
Section                           0
PE                                0
Work Center                       0
Module                            0
Planned/Projected Efficiency    851
Present Employees               286
Absent Employees                286
No Of Hours Per Day             286
Worked Hours                      0
Daily Down Time Hours           851
Impacted DownTime Hours         851
Daily Performance               155
Customer                        155
Style Code                      155
Style Description               155
SO                              155
LI                              155
FG Reference                    155
SO/LI Worked Hours                0
Efficiency                        0
Time Slot                     14460
Total                             0
SMV                               0
Standard Hours                    0
```

Figure 1 Information of columns containing null values



```
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Date                          13454 non-null  datetime64[ns]
 1   Section                       13454 non-null  object
 2   PE                            13454 non-null  object
 3   Work Center                   13454 non-null  object
 4   Module                        13454 non-null  object
 5   Planned/Projected Efficiency  13454 non-null  float64
 6   Present Employees             13454 non-null  float64
 7   Absent Employees              13454 non-null  float64
 8   No Of Hours Per Day           13454 non-null  float64
 9   Worked Hours                  13454 non-null  float64
 10  Daily Down Time Hours         13454 non-null  float64
 11  Impacted DownTime Hours       13454 non-null  float64
 12  Daily Performance             13454 non-null  float64
 13  Customer                      13454 non-null  object
 14  Style Code                    13454 non-null  object
 15  Style Description             13454 non-null  object
 16  SO                            13454 non-null  object
 17  LI                            13454 non-null  float64
 18  FG Reference                  13454 non-null  object
 19  SO/LI Worked Hours            13454 non-null  float64
 20  Efficiency                    13454 non-null  float64
 21  Total                         13454 non-null  int64
 22  SMV                           13454 non-null  float64
 23  Standard Hours                13454 non-null  float64
```

Figure 2 Information about type of columns

Pre-Processing of Data

Preprocessing of data is an essential step in any data analysis project, including the analysis of production planning data. The preprocessing step involves cleaning and transforming the data to make it suitable for analysis. This may involve removing duplicates, correcting errors, handling missing values, and transforming the data into a suitable format for analysis. Other preprocessing techniques may include data normalization, feature scaling, and dimensionality reduction. The goal of preprocessing is to ensure that the data is accurate, complete, and in a format that can be easily analyzed to gain insights into the production planning process.

- **Data Cleaning:** This step involves identifying and removing or correcting any errors, inconsistencies, or missing values in the production planning data.

- **Data Transformation:** This step involves transforming the production planning data into a format that is suitable for analysis. This may include converting categorical variables into numerical ones, or scaling the data to a common range.
- **Data Normalization:** This step involves scaling the production planning data to a standard range or distribution, making it easier to compare across different line sections and time periods.
- **Feature Scaling:** This step involves scaling the different features or variables in the production planning data to a similar range, so that they can be compared more accurately.
- **Dimensionality Reduction:** This step involves reducing the number of features or variables in the production planning data to a smaller, more manageable set. This can help to improve the accuracy and speed of analysis.
- **Data Aggregation:** This step involves combining the production planning data into a smaller, more manageable set by aggregating data from different time periods or line sections.
- **Data Sampling:** This step involves selecting a subset of the production planning data for analysis, to reduce the computational requirements of the analysis.

These are just some of the steps that can be taken to preprocess production planning data, depending on the specific requirements of the analysis. In this case study, we first remove the outlier from the columns of the dataset and the shape of the dataset after removing outlier will become (7891,24) where column reduce because there are all null value in “Time Slot” column.

Analysis

The analysis of the production planning dataset can provide valuable insights into the performance of different line sections and time periods, as well as identify potential areas for improvement in the production planning process. The analysis may involve exploratory data analysis techniques such as data visualization, descriptive statistics, and hypothesis testing. It may also involve more advanced statistical and machine learning techniques, such as regression analysis, time series analysis, clustering, and classification. The goal of the analysis is to uncover patterns, trends, and anomalies in the production planning data and use this information to make data-driven decisions to improve the production planning process. The results of the analysis may be presented in the form of charts, graphs, tables, and reports, along with recommendations for improvement.

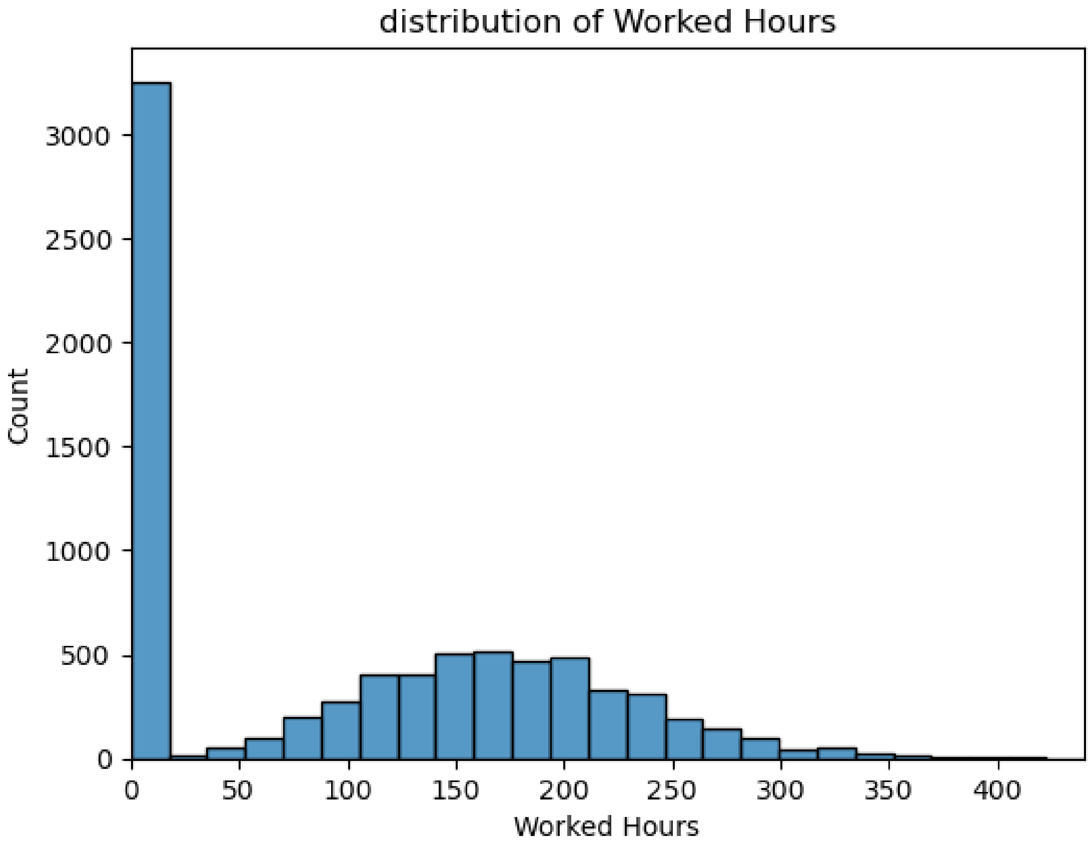


Figure 3 Distribution of Work Hours

Figure 3 shows the distribution of work hours (target variable) which shows that most of emp have work hours between 0 to 50 and average peoples have work hours between 50 to 250. The number of such people are below 500. Figure 4 shows there are seven work center and it can be seen that “0048-A” has lowest number of present employee overall and “0071-A” with “0024-A” has most number of present employee around 1600 while if we look at average then most of the work center has on average 200 to 1200 number of present employees.

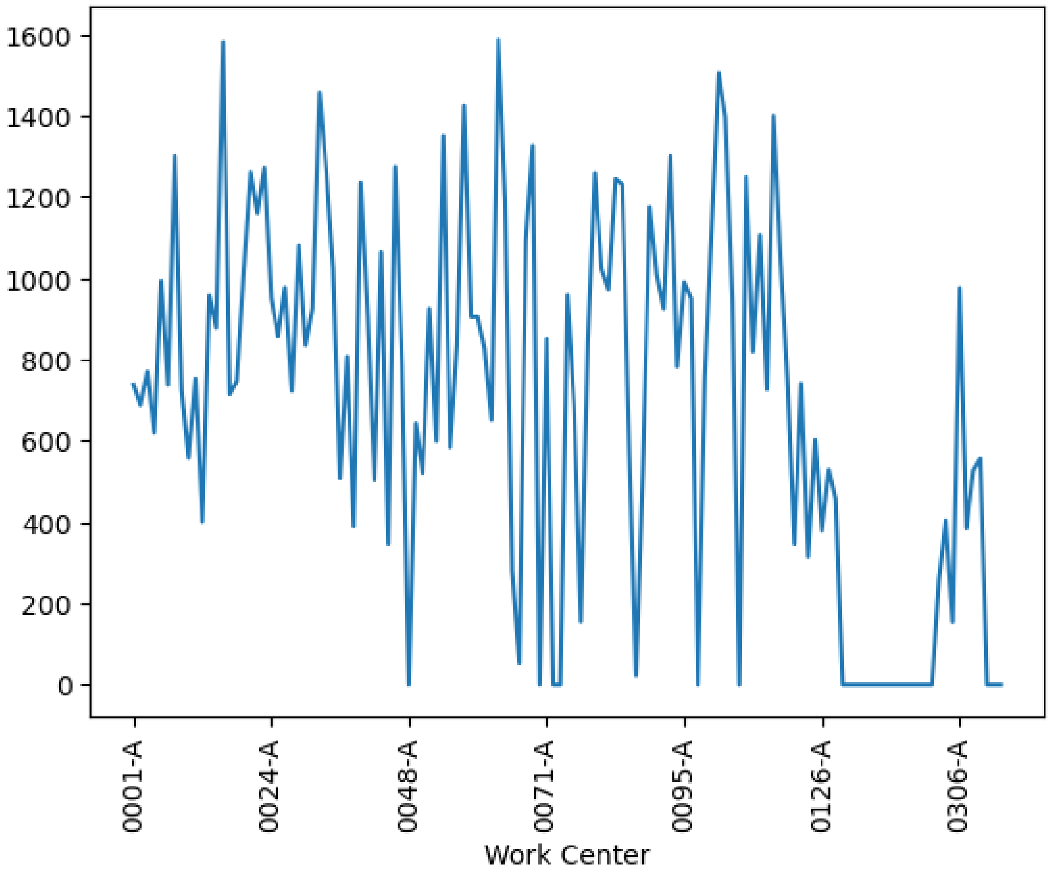


Figure 4 Present employee in a work center

Figure 5 shows Down time of the centers available for work and it was found that two center “0001-A” and “0024-A” have down time while other all center has zero down time which mean performance of all center are good they have very few time consider as down.

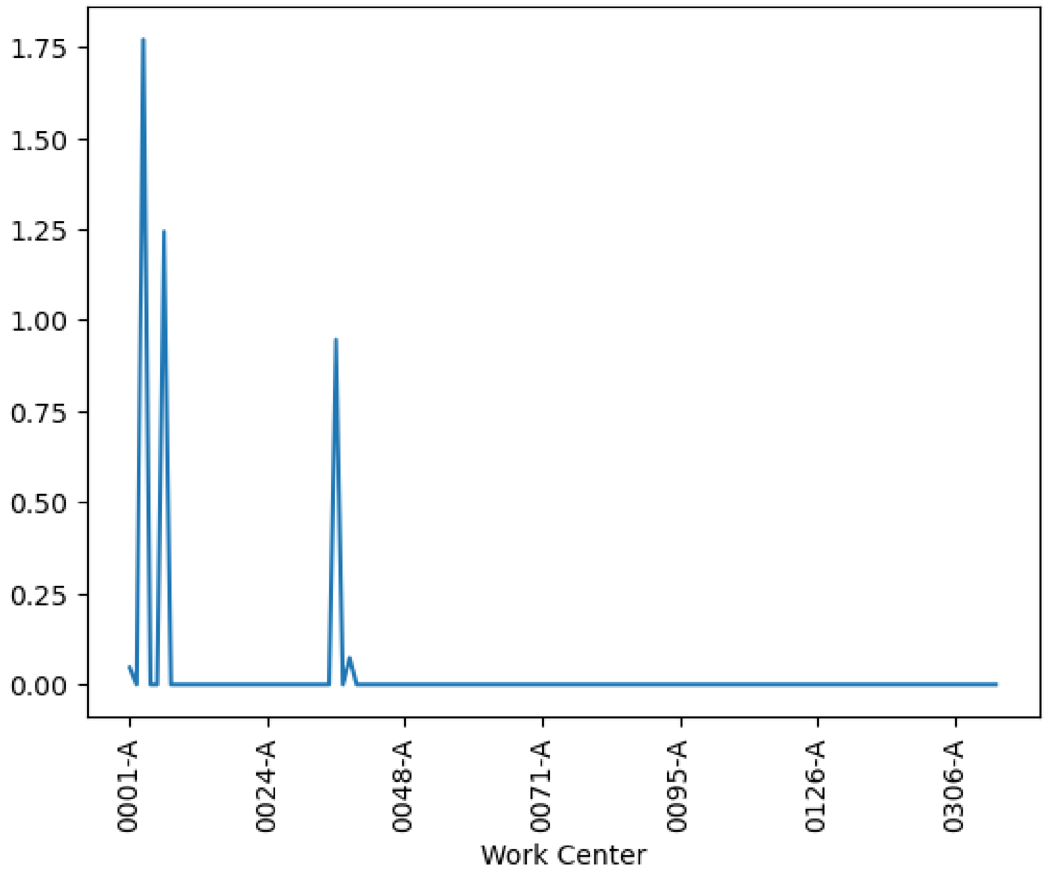


Figure 5 Down time of work center

Correlation Graph

A correlation graph is a visual representation of the relationship between two or more variables in a production planning dataset. The graph can be used to determine the strength and direction of the relationship between the variables. A positive correlation indicates that an increase in one variable is associated with an increase in the other variable, while a negative correlation indicates that an increase in one variable is associated with a decrease in the other variable. A correlation coefficient is often calculated to provide a quantitative measure of the strength of the correlation.

Correlation graphs can provide valuable insights into the production planning dataset by helping to identify relationships between different variables. For example, a correlation graph may reveal that there is a strong positive correlation between the number of raw materials used and the number of units produced. This information can be used to optimize the production process by ensuring that the appropriate number of raw materials is available for each line section to meet production goals. Figure 6 shows the correlation graph between columns of the dataset and Qty and work hour have most column correlation other column that have correlation are SMV, Standard Hours, Total, Efficiency and SO/LI Worked Hours.

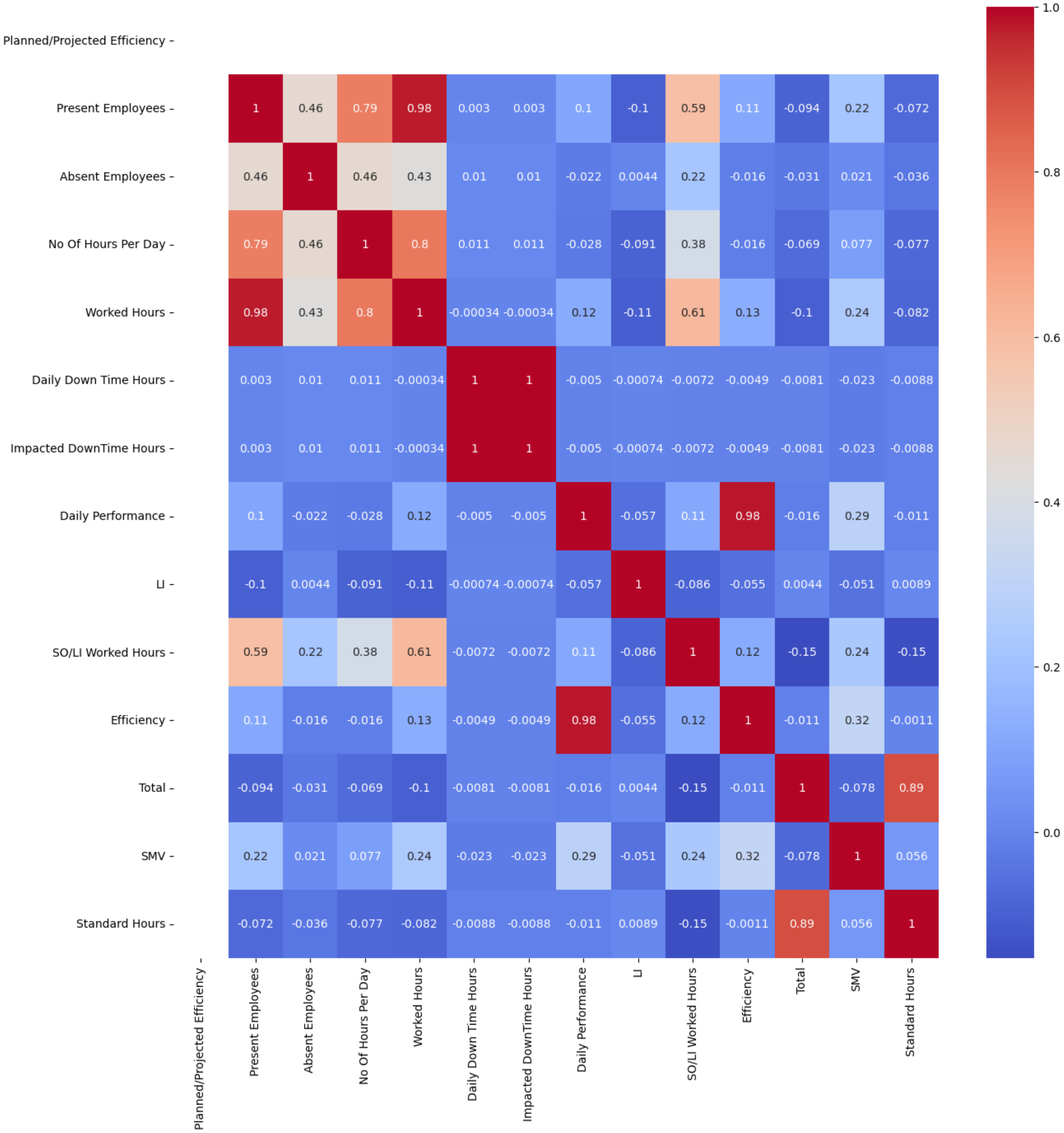


Figure 6 Correlation graph

Methodology

Label encoding is a common technique used in data preprocessing to convert categorical variables into numerical ones. In production planning data, categorical variables may include information such as line section names or product types. Label encoding assigns a numerical value to each category in a variable, making it easier to use the variable in statistical analysis or machine learning models. Label encoding can be useful in cases where there is a clear ordinal relationship between the categories in a variable. For example, if there are product types named "low", "medium", and "high", label encoding might assign the values 1, 2, and 3 to these categories, respectively. This allows statistical analyses to take into account the relative ordering of the categories.

However, it's important to note that label encoding can also have limitations. In cases where there is no clear ordering between the categories in a variable, such as with product names or colors, label encoding may not be appropriate. In these cases, one-hot encoding or other techniques may be more suitable. In this scenario, “Section”, ”PE”, ”Customer” and “Work Center” are the columns on which label encoding is applied on it.

Algorithm

Linear Regression is a simple yet powerful algorithm that is widely used in production planning and other applications for predicting continuous values. It works by finding the best linear relationship between a set of input features and the output variable. The algorithm learns the coefficients of the linear equation through a process called optimization, typically using a method such as gradient descent. Once the coefficients are learned, the model can make predictions on new data by simply multiplying the input features by the coefficients and adding a constant term.

Linear regression has several advantages over other machine learning algorithms. It is easy to implement, interpretable, and computationally efficient. However, it also has some limitations. For example, linear regression assumes a linear relationship between the input features and the output variable, which may not always be the case in production planning data. Additionally, linear regression can be sensitive to outliers and may not perform well with multicollinear input features.

To address these limitations, several regularization techniques have been developed. One such technique is Ridge Regression, which adds a penalty term to the linear regression cost function. This penalty term forces the algorithm to keep the weights of the input features small, which helps to prevent overfitting. Ridge regression is particularly useful in cases where there are many input features and some of them are highly correlated.

Lasso Regression is another regularization technique used in linear regression. It works by adding a penalty term to the cost function that forces some of the input features to be set to zero. This results in a sparse model that only uses the most important input features. Lasso regression is particularly useful in cases where there are many input features and only a few of them are relevant for predicting the output variable.

Elastic Net is a combination of Ridge and Lasso regression. It adds both L1 and L2 penalties to the cost function, which results in a model that is both sparse and with small weights on input features. Elastic Net is particularly useful in cases where there are many input features and some of them are highly correlated, but not all of them are relevant for predicting the output variable.

Finally, Bayesian Ridge Regression is a probabilistic model used in linear regression. It assumes that the weights of the input features are normally distributed, which allows it to make probabilistic predictions on new data. Bayesian Ridge Regression can handle noisy and multicollinear input features and is less prone to overfitting than other linear regression models. These five algorithms were applied to the dataset.

Results

The results columns of a production planning dataset contain the output of the prediction models that were applied to the input features. These columns are essential for evaluating the performance of the models and determining their accuracy in predicting the target variable. Depending on the type of production planning problem, the results columns may include continuous values, binary values, or categorical values. Result of the algorithm shown in figure 7 shows that all algorithms have 0.9+ accuracy which is good accuracy.

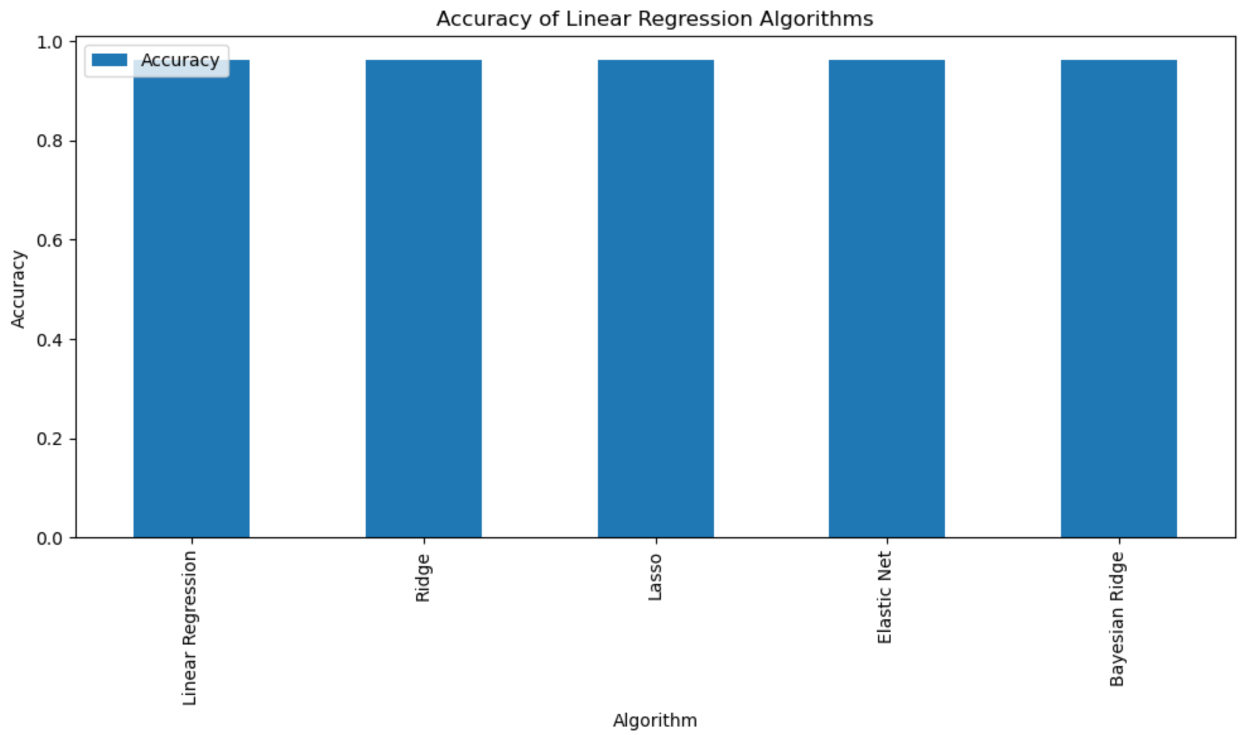


Figure 7 Result of Algorithm