Introduction

In today's fast-paced business environment, production planning plays a critical role in ensuring that organizations can meet the demands of their customers efficiently and effectively. Production planning involves the creation of a roadmap that outlines the activities required to produce a product or service, taking into account the available resources, timelines, and other relevant factors.

The purpose of this report is to analyze the production planning data for four-line sections - LC Sec 1, LC Sec 2, LC Sec 3, and LC Sec 4 - for different periods. The data is contained in 49 files, with each file containing planning data for a specific time period and line section. By analyzing this data, we can gain insights into the performance of the different line sections and identify any areas that need improvement.

The production planning data provides information on various aspects of the production process, such as the number of units produced, the quantity of raw materials used, and the time taken to complete each task. By analyzing this data, we can identify any bottlenecks or inefficiencies in the production process and take steps to address them. For example, we may find that a particular line section is consistently falling behind schedule, which could indicate that there are issues with the allocation of resources or the scheduling of tasks.

The data analysis in this report will involve identifying trends and patterns in the production planning data, as well as any anomalies or outliers that may require further investigation. We will also examine the data to identify any correlations between different variables, such as the quantity of raw materials used and the number of units produced.

The findings of this analysis will be presented in this report, along with recommendations for improving the production planning process. These recommendations may include changes to the allocation of resources, adjustments to the scheduling of tasks, or the implementation of new tools and technologies to streamline the production process. Ultimately, the goal of this analysis is to help the organization improve its production planning process and ensure that it can continue to meet the needs of its customers in a timely and efficient manner.

Dataset

The data is contained in 49 files, with each file containing planning data for a specific time period and line section. The dataset consists of following 20 columns

- Module
- Material
- Customer No
- Description
- Customer Dept.
- Gender
- S/O
- L/I
- Order No
- Order Qty
- Emp.
- SMV
- Date
- Eff. %
- Qty.
- Cum Qty.
- Standard Hours
- Cum.Standard Hours
- Work Hours
- Cum.Work Hours

Information of Dataset

Initially, the dataset contains 17479 instances of 20 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 two columns “Gender” and “S/O”. 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 11 columns are of type float and one is of datetime type.

```
Module                  3
Material              797
Customer No            31
Description             8
Customer Dept.          3
Gender               2205
S/O                  2205
L/I                  1411
Order No.               3
Order Qty.              3
Emp.                    3
SMV                     3
Date                    3
Eff. %                  3
Qty.                   56
Cum Qty.              228
Standard Hours.        55
Cum.Standard Hours.   228
Work Hours.            43
Cum.Work Hours.       228
dtype: int64
```

Figure 1 Information of columns containing null values

```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Module               15090 non-null  object
 1   Material             15090 non-null  object
 2   Customer No          15090 non-null  object
 3   Description          15090 non-null  object
 4   Customer Dept.       15090 non-null  object
 5   Gender               15090 non-null  object
 6   S/O                  15090 non-null  object
 7   L/I                  15090 non-null  object
 8   Order No.            15090 non-null  float64
 9   Order Qty.           15090 non-null  float64
 10  Emp.                 15090 non-null  float64
 11  SMV                  15090 non-null  float64
 12  Date                 15090 non-null  datetime64[ns]
 13  Eff. %               15090 non-null  float64
 14  Qty.                 15090 non-null  float64
 15  Cum Qty.             15090 non-null  float64
 16  Standard Hours.      15090 non-null  float64
 17  Cum.Standard Hours.  15090 non-null  float64
 18  Work Hours.          15090 non-null  float64
 19  Cum.Work Hours.      15090 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 (12884,20).

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.

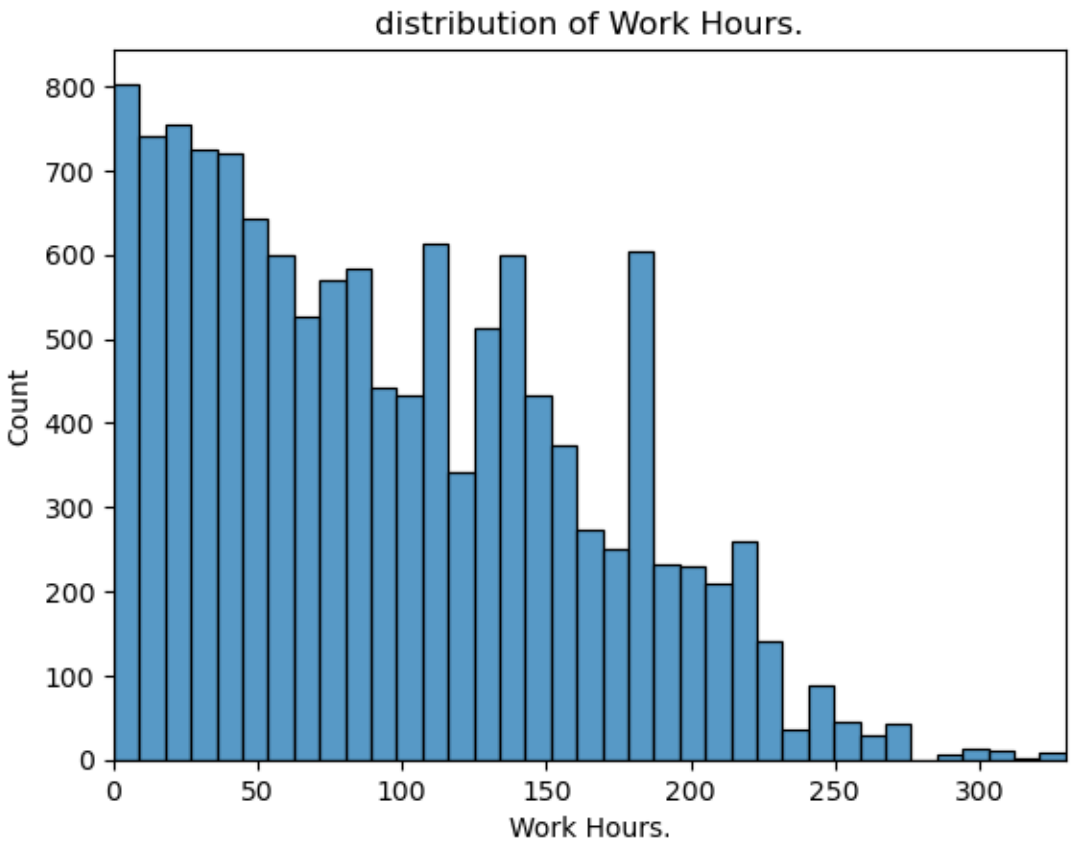

Figure 3 Distribution of Work Hours

Figure 3 shows the distribution of work hours (target variable) which shows that most of emp have work hours between 0 to 150 and very people have work hours between 250 to 300. The number of such people are below 100. Figure 4 shows which gender count has most placed order and it can be seen that women has mostly placed order and there is few order which are from other gender. In figure can be seen that there is huge gap between gender counts of order.

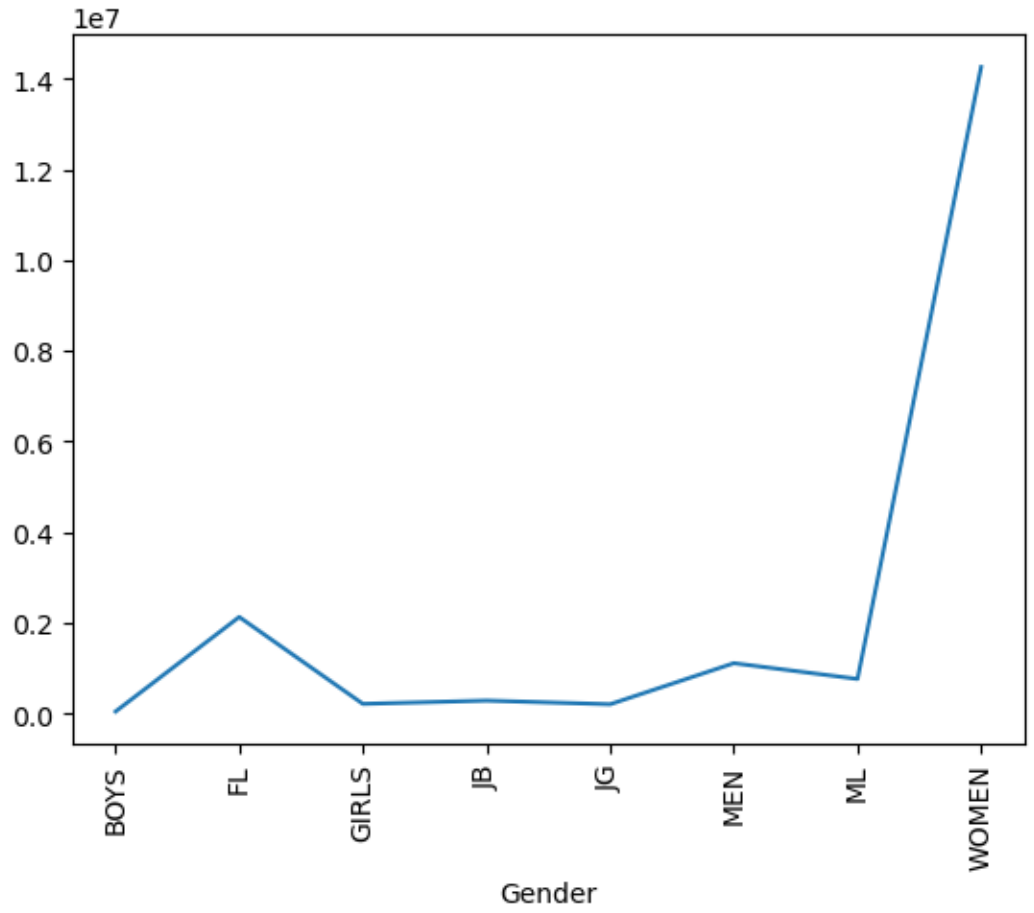

Figure 4 Gender count

Figure 5 shows Date-wise distribution of count of quantity of order placed in month and it can be seen that most order are sold in month of march and it has count of 250k order placed while month of august of 2018 has lowest order quantity which is below 50k.Most of the month have order quantity is above 50k and 200k.

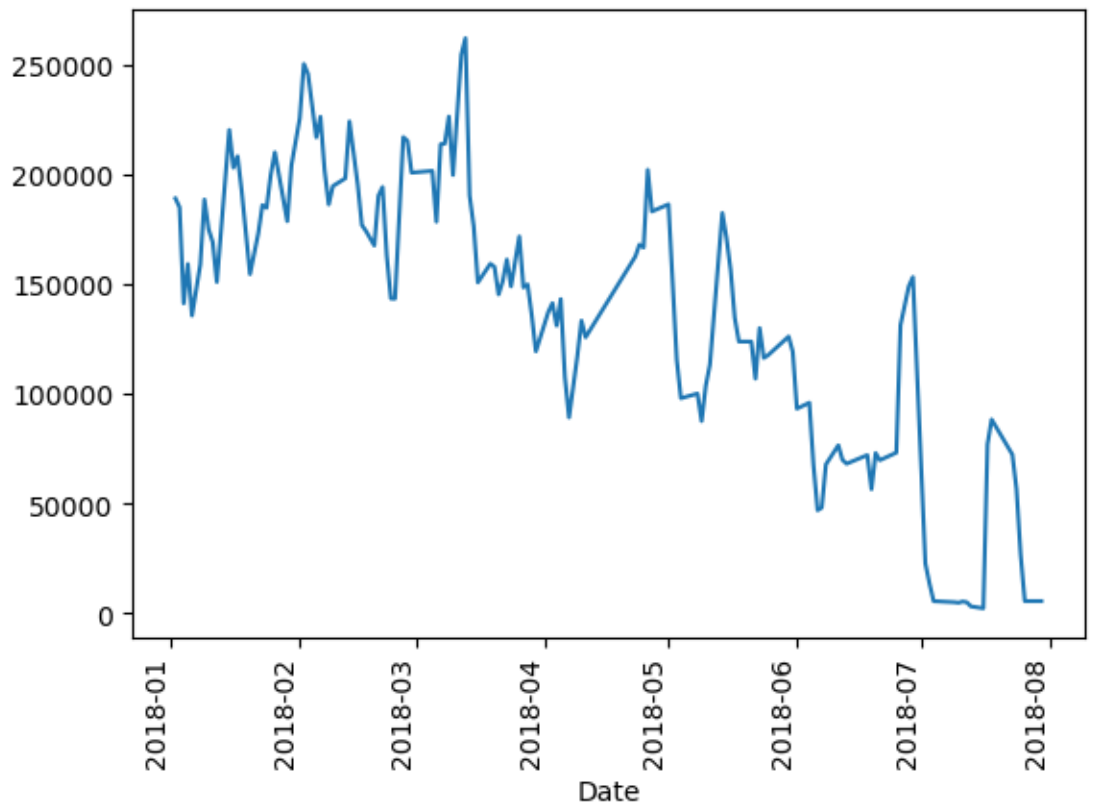

Figure 5 Date-wise count of Order Quantity

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 Cum Qty. , Qty , Standard Hours, Cum. Standard Hours, and Cum. Work Hours.

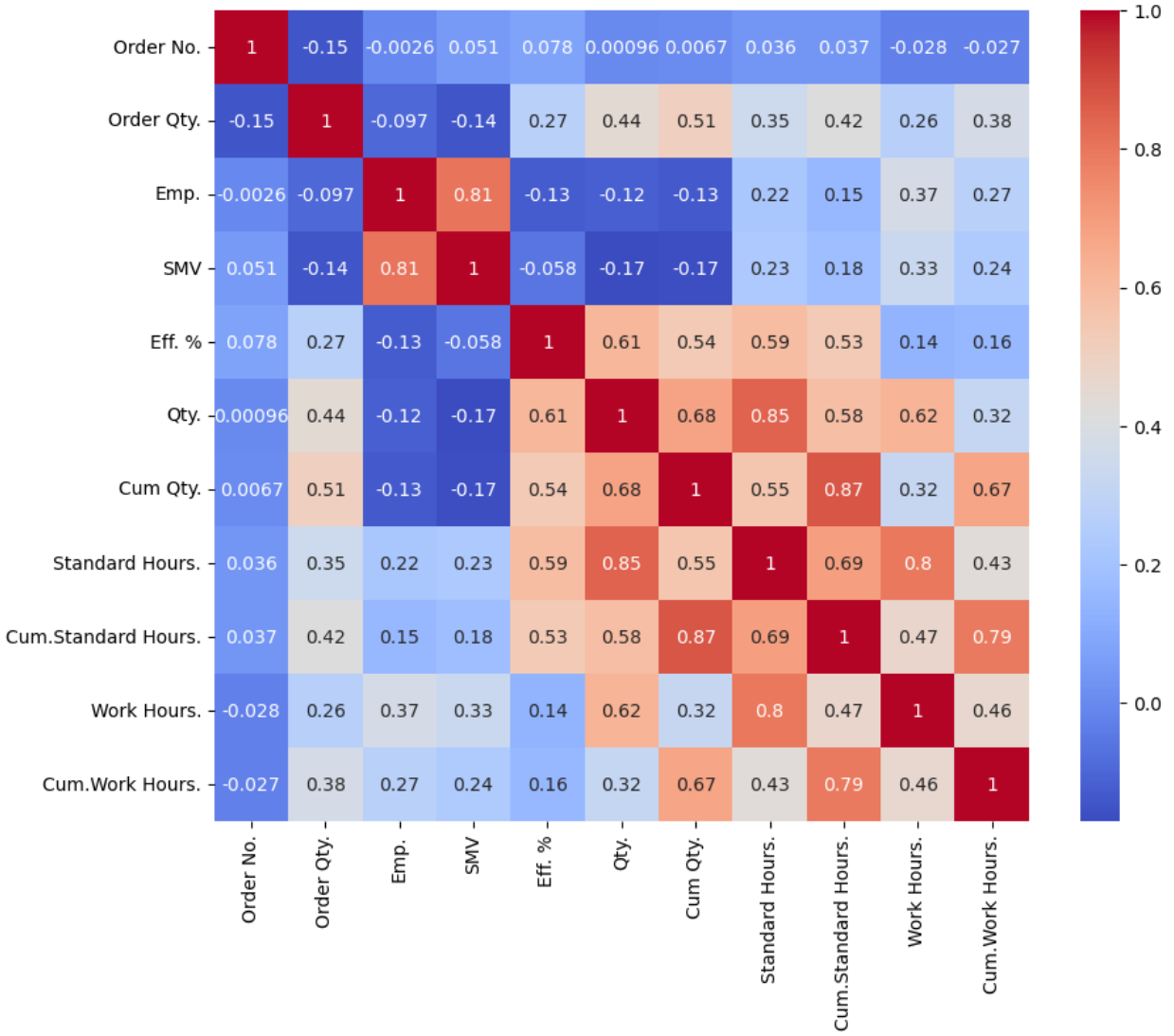

Figure 6 Correlation graph

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, “Gender” is the column 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 0.8+ accuracy which is good accuracy.

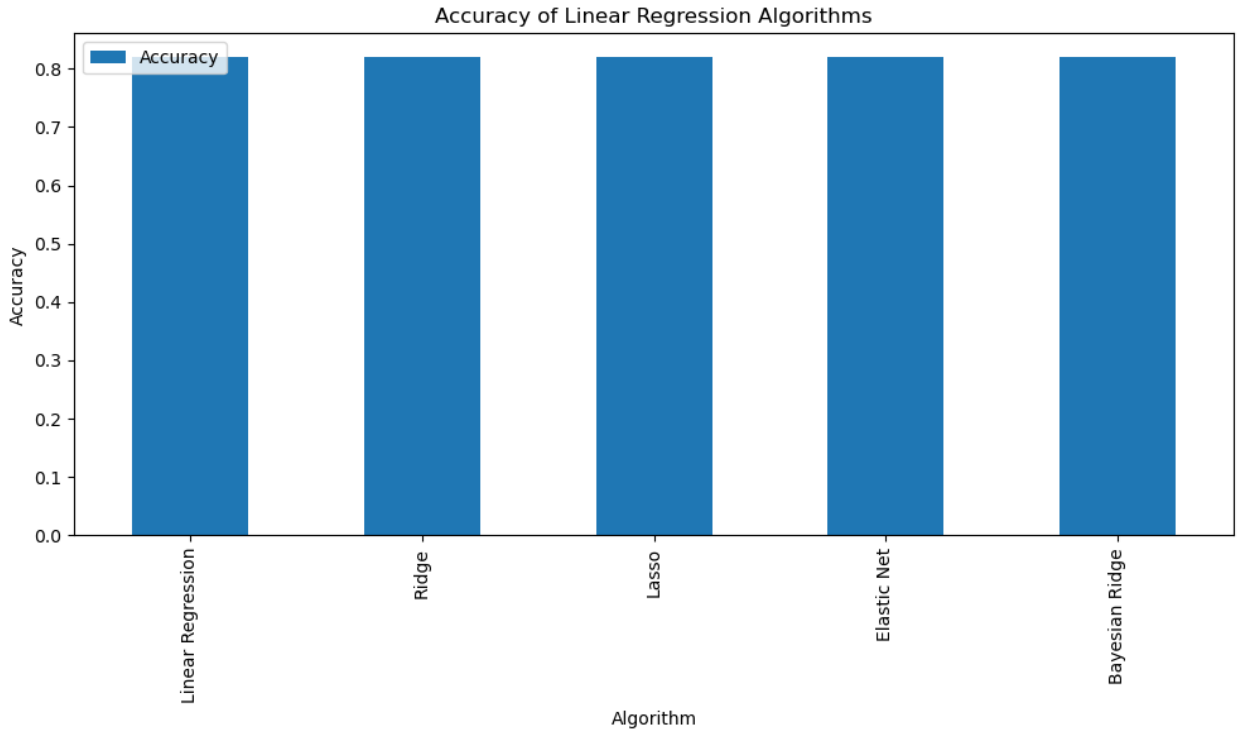

Figure 7 Result of Algorithm