**Solution of BXB data science challenge:**

I used python to solve this challenge.

1. **Data Visualization:**

First I loaded the dataset into python using pandas. To visualize the effect of different variables on IssuesInMonth, I plotted different variables with IssuesInMonth. It was observed that TransfersInMonth varies almost linearly with IssuesInMonth. It is basically obvious because as the number of issues(demand) increases, the number of Transfers also should Increase to fulfil the demand of manufacturing companies.

BusinessDaysInMonth also plays an important role as demand depends on number of business days.

1. **Feature Engineering:**

I observed that number of transfers in a day is also an important feature to predict the total number of issues in a month. Therefore, I added a new column to the dataframe namely ‘per day transfer’ which is equal to TransfersInMonth/ BusinessDaysInMonth.

I also divided the training dataset into training and validation dataset to check the validity of the model using model selection.train\_test\_split library of python. I took 20% of the training dataset to validate the model.

1. **Model Selection:**

I used linear regression to predict the number of issues in a month. This model can be used because as number of transfers increases, number of issues also increases.

Therefore, my input feature consists of TransfersInMonth, BusinessDaysInMonth and PerdayTransfer. The output feature is IssuesInMonth.

I first fitted the model using training data and then predicted the IssuesInMonth for validation dataset. I, then used the prediction of the model for validation dataset to check the accuracy of model.

1. **Error Observation and conclusion:**

For validation data set, the **mean absolute percentage error** was measured to be **2.901 %.** Thus, the model can be used to predict the output of the test dataset.

The **R-squared** value of the validation dataset comes out be **0.83**, which is close to 1. Hence, I concluded that the model predicts the output quite accurately.

1. **Predicting the output of test data:**

To predict the output of test data, I used the linear regression model fitted above and predicted the output values.