

MACHINE LEARNING

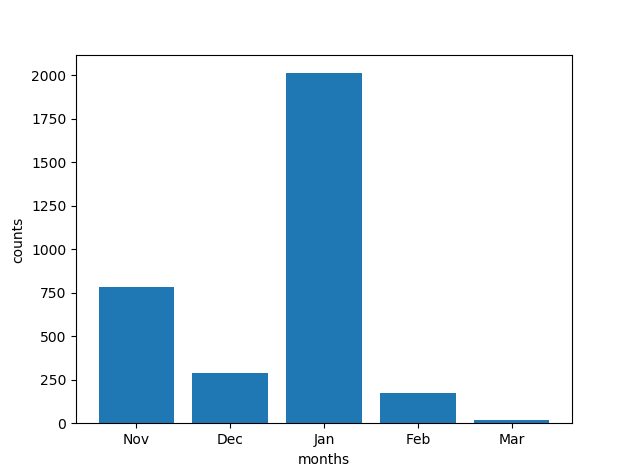
PROJECT 1

### 

# Question:

Using the web server logs,  
Which is the most frequent site that is required to use in this server?  
What is the model that we can fit (linear regression) for this model when we use average frequency per hours (in different days) for the most required site?  
When you point the data, is it linear?  
  
  
What is the rush hours fort his server?  
This time x is hours in a day and y are average usage per hours (in different days).  
What is the model that we can fit (linear regression) for this model?  
When you point the data, is it linear?  
  
Can we use different time series algorithm to predict the most frequent site or rush hour?  
Write the necessary codes that solve this problem; use python  
DATA: <https://www.kaggle.com/shawon10/web-log-dataset#webLog.csv>

# Histogram for most frequent site:

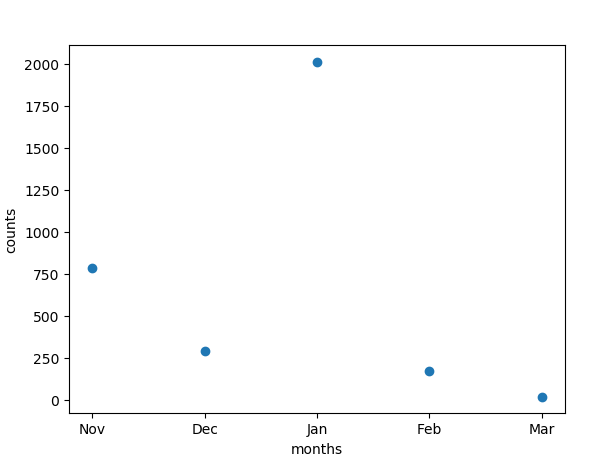


Note: X axis is the time where we have considered months for simplicity. Y axis is the number of frequent sites

GET /login.php HTTP/1.1

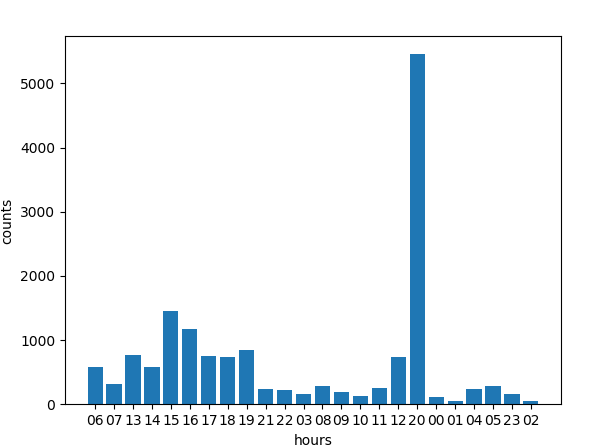
We have plotted some more graphs for the better analyzing.

# Scatter diagram:



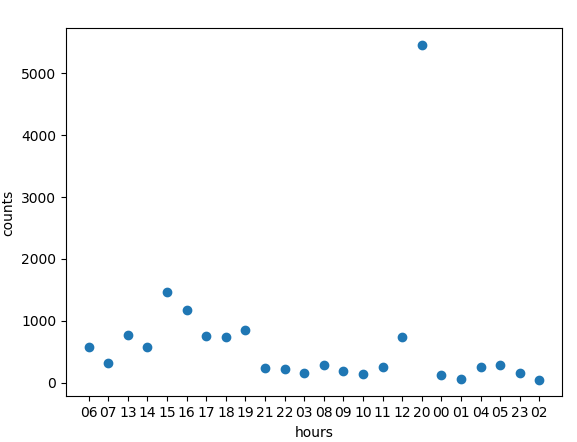
Our model is not linear.

# **Histogram of the most frequent time:**



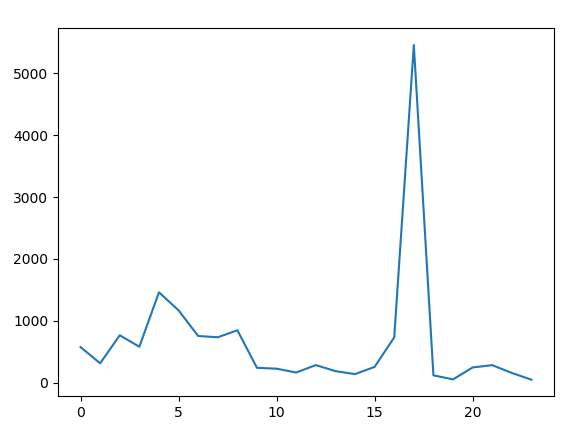
Note: 20 is the most frequent time appeared on the dataset which is above 5000.

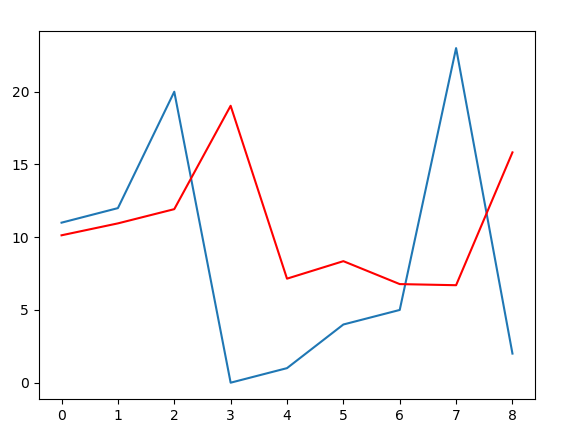
# Scatter diagram:



As observed, points are close to each other so we can say that the model is linear.

# Series algorithm:





We have used Arima time series algorithm for the prediction. The red line indicates the prediction and blue indicates the expected one.

Output data and error is printed in the python code.