Letter Recognition Analysis

-MADHAVI RAO

Introduction-Problem Statement

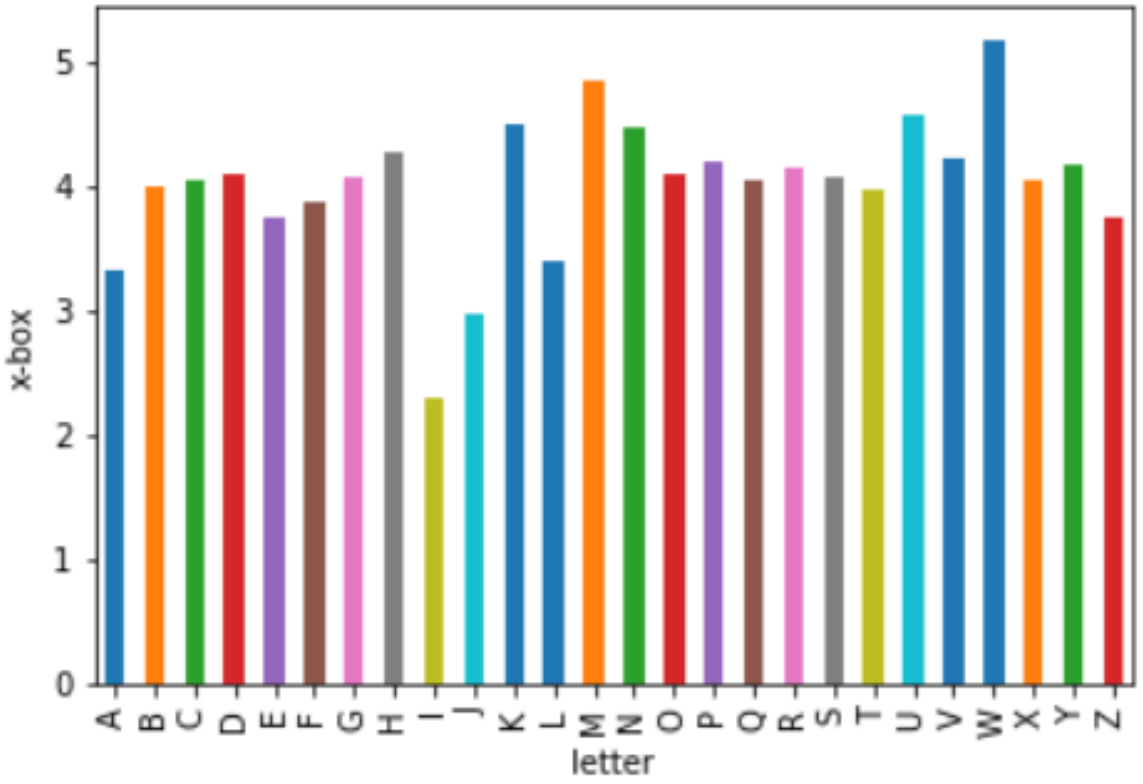
The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet.

The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli.

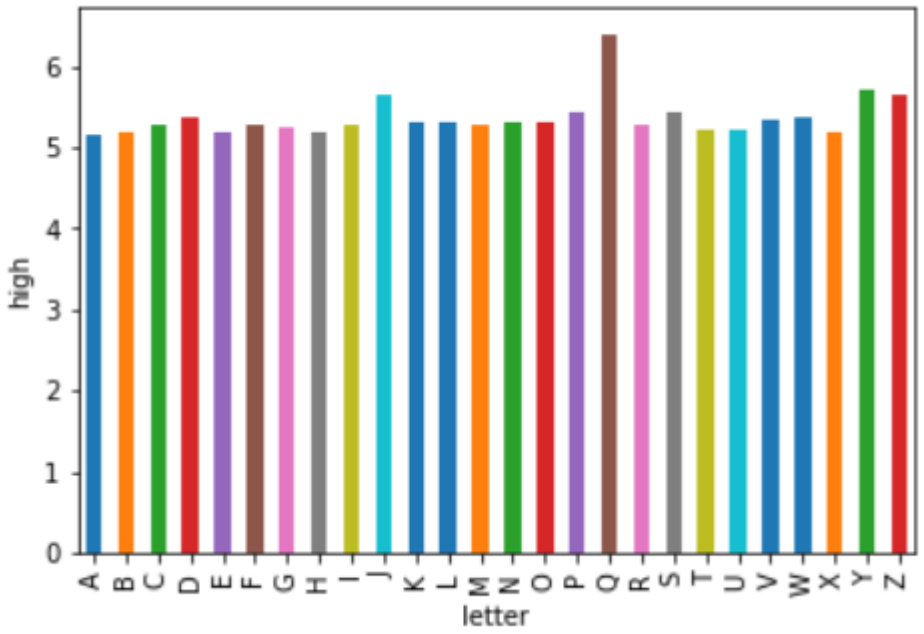
Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15.

Exploratory Data Analysis

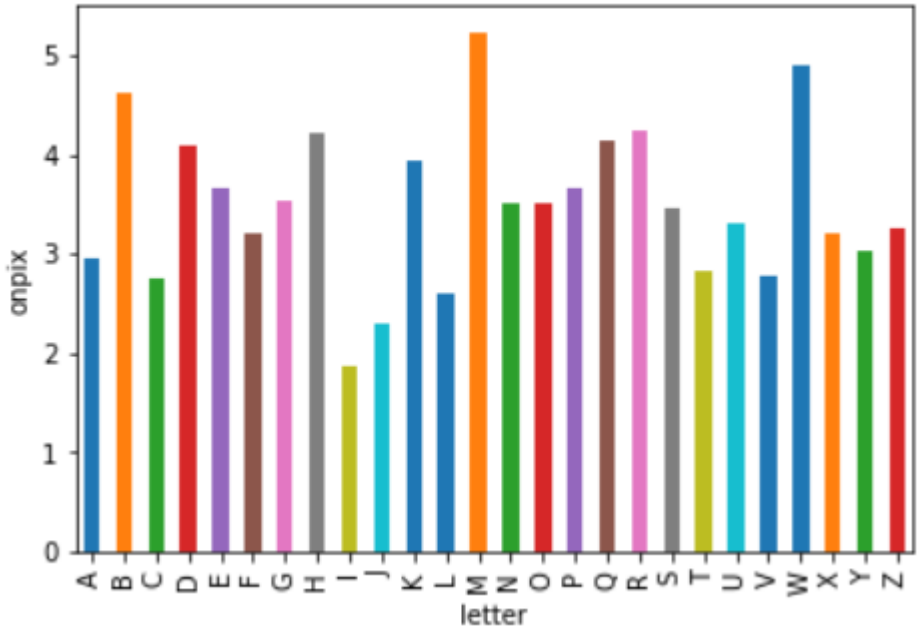
Distribution of attributes



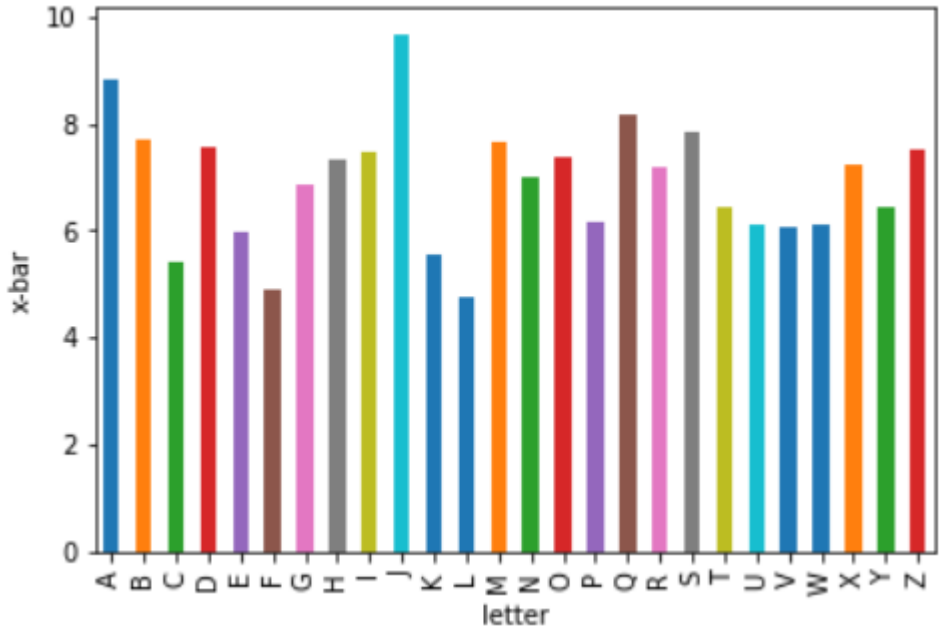
x-box values is highest for W and then for M for obvious reasons of them being wider letters. This value is lowest for I.



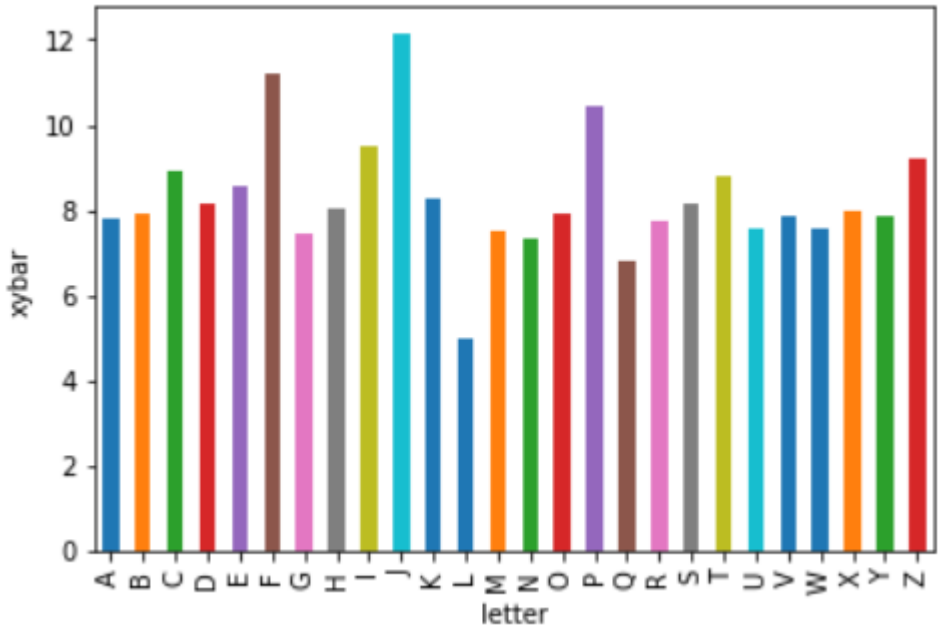
high value is the highest for Q as this letter extends a little below the line when compared to all other letters.



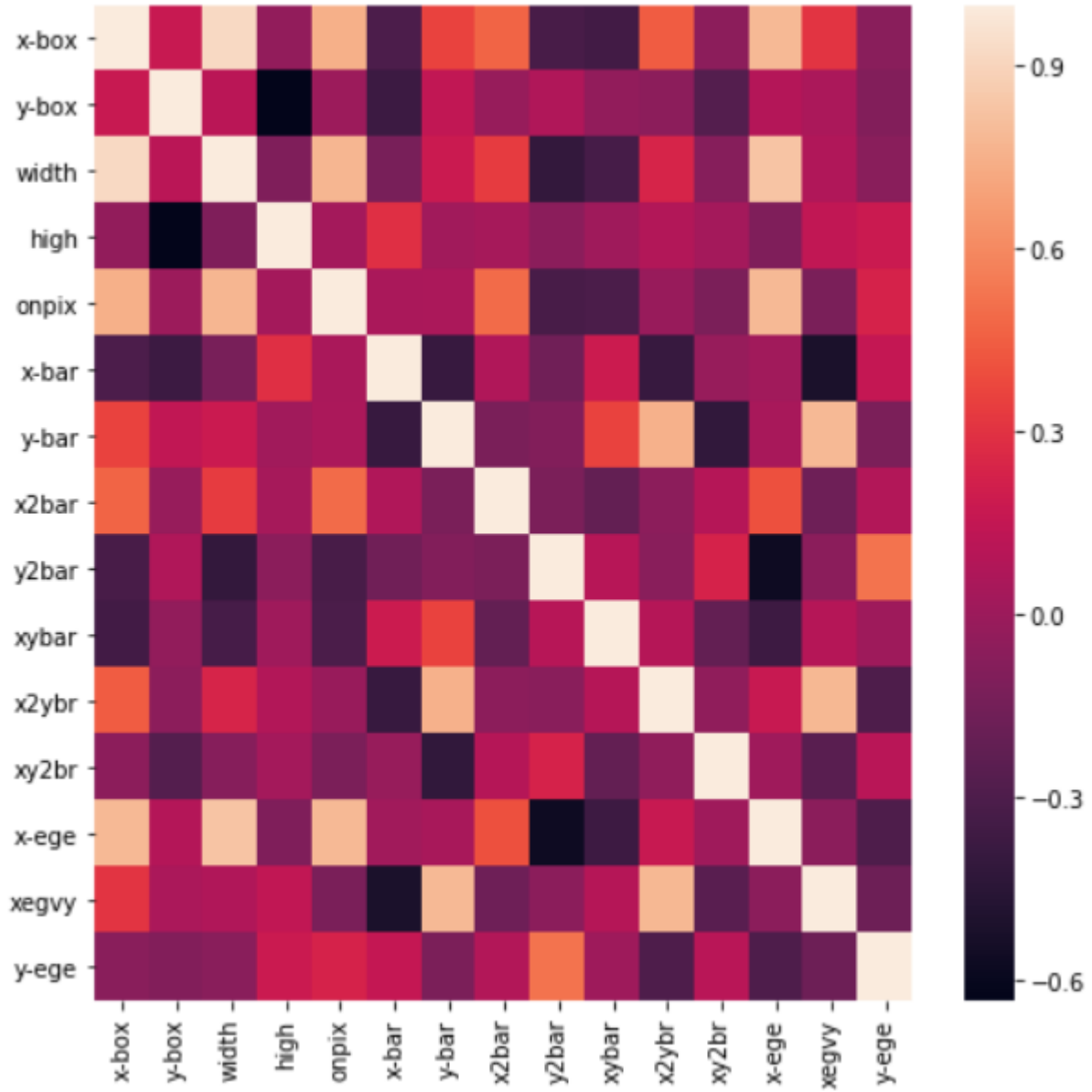
Across all fonts letter M has the maximum no of pixels.



Highest mean x of on pixels in box (x-bar) is letter J and A.



The correlation between x and y is highest for letter J.

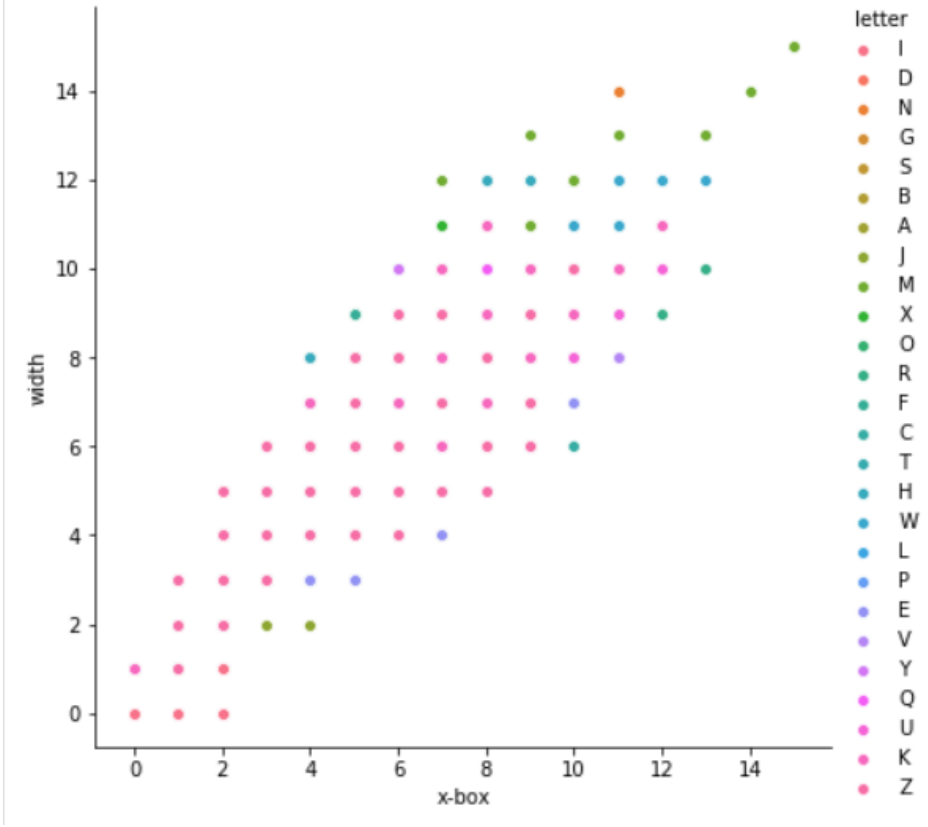


Correlations

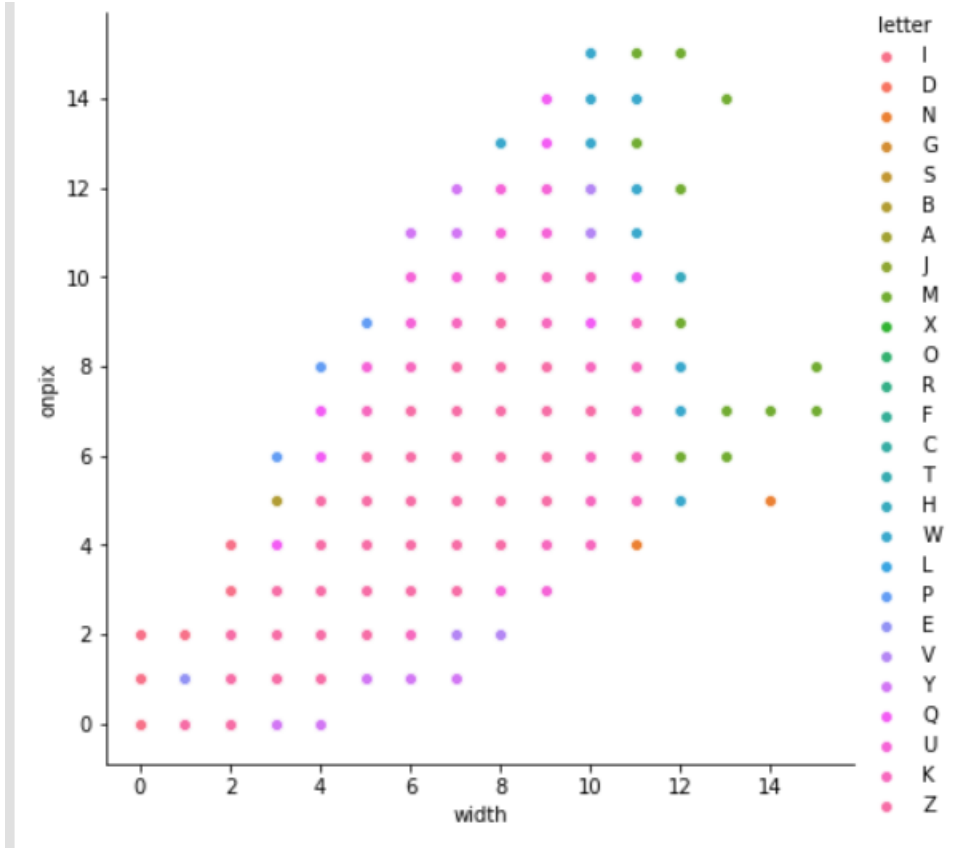
x-box and width = 0.92
width and onpix = 0.78
x-box and onpix = 0.75
y-box and high = -0.63

x-box and width, y-box and high are highly correlated. We can remove width and high features and try to run our models.

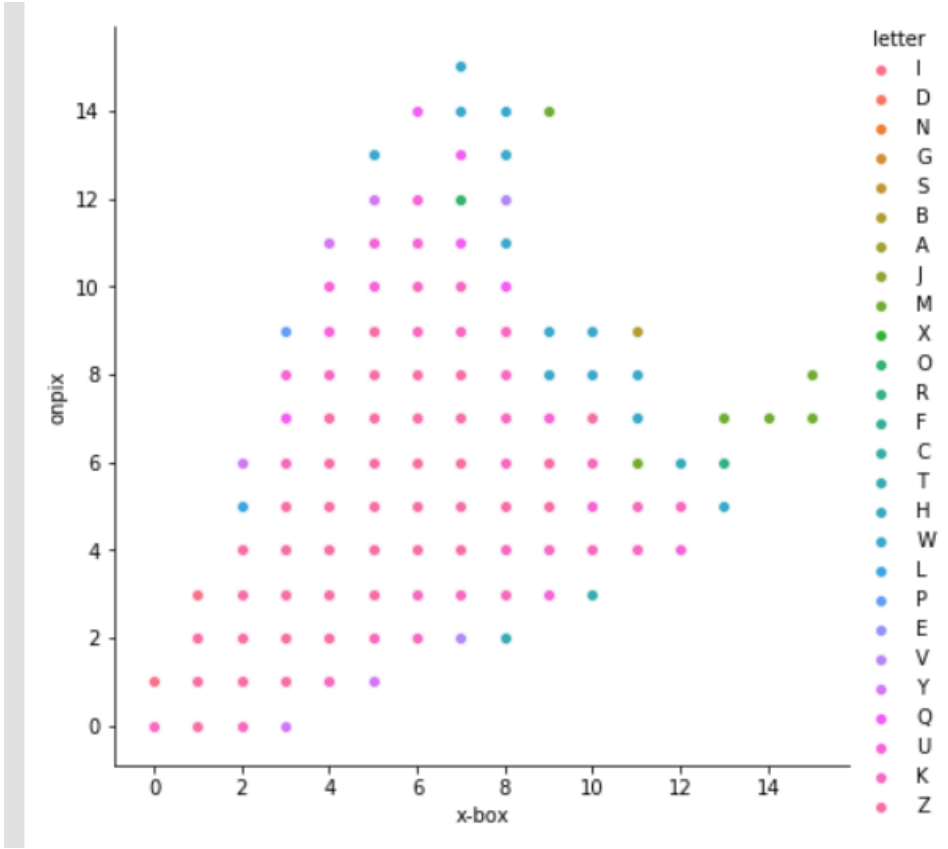
Width and x-box



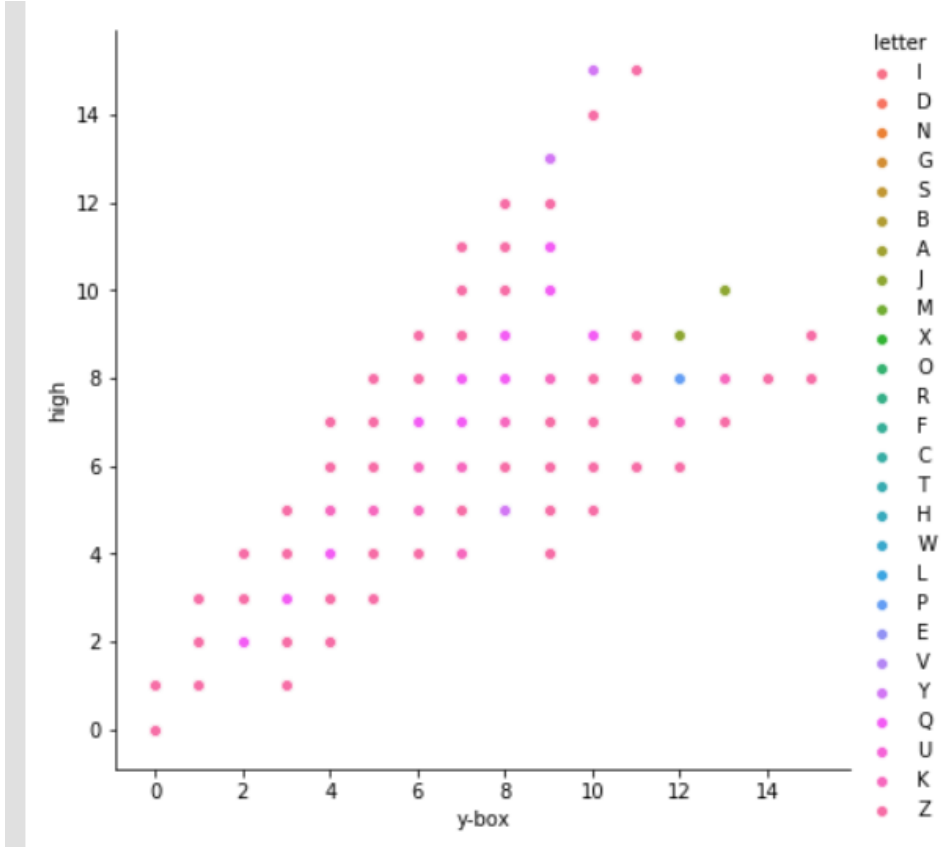
Onpix and width



Onpix and x-box



High and y-box



Model Creation and Evaluation

Algorithms applied:

Multinomial Logistic Regression
Decision trees and hyper parameter tuning
Random Forest and hyper parameter tuning
K Nearest neighbors
Support vector Classifier
Naïve Bayes
XGBoost
ensembleVoteClassifier

Summary

Support Vector Classifer is giving the highest accuracy compared to other models with score of "**96 %**" hence we choose the algorithm for letter prediction.

Hyper parameters used

C=1000
gamma=0.01
kernel="rbf"