Candlestick Classification

Classifies the candles into five major categories viz. Marubozu, Doji, Dragonfly Doji, Gravestone Doji and Spinning Top. The candlesticks that can not be classified into above five categories are categorized as ‘undefined’. Support Vector Machine algorithm is used as classification model and gave around 80% of accuracy on test dataset.

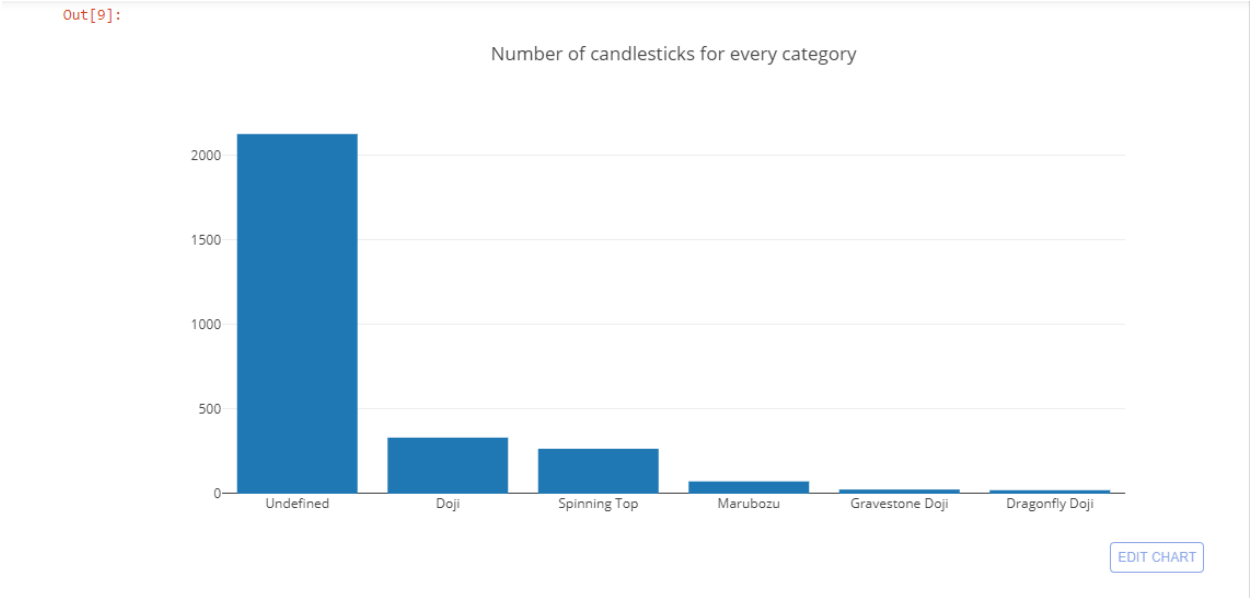
Data

Initially there is the data for daily stock data i.e. open, high, low and close. Ta-Lib library is used to assign labels to the data. Ta-Lib in python is the library for technical analysis that have several functions like “talib.CDLMARUBOZU(open, high, low, close)” that return 100 or -100 if there is marubozu candlestick, and 0 if no marubozu pattern is there. Ta-Lib assigned 75% of data to ‘undefined’ category. So not getting dependent on results of Ta-Lib, more samples for each candlestick were generated for training purpose.

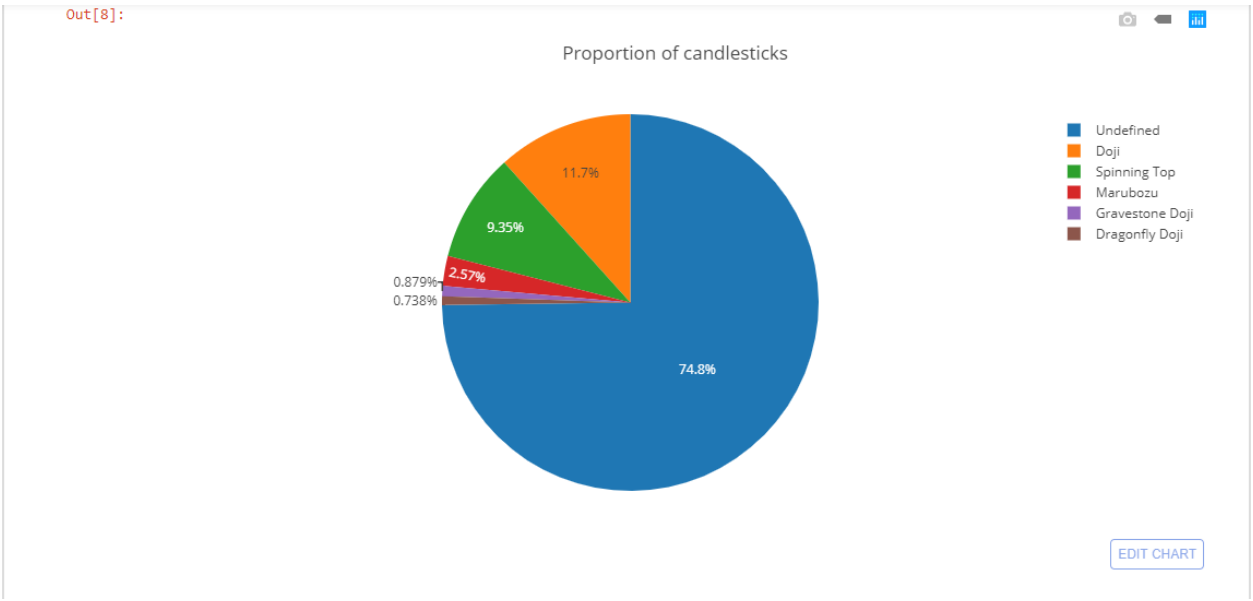
Getting started

The unlabeled data is labeled by Ta-Lib library and analyzed. The proportion and counts for each category were visualized

Out[9]:

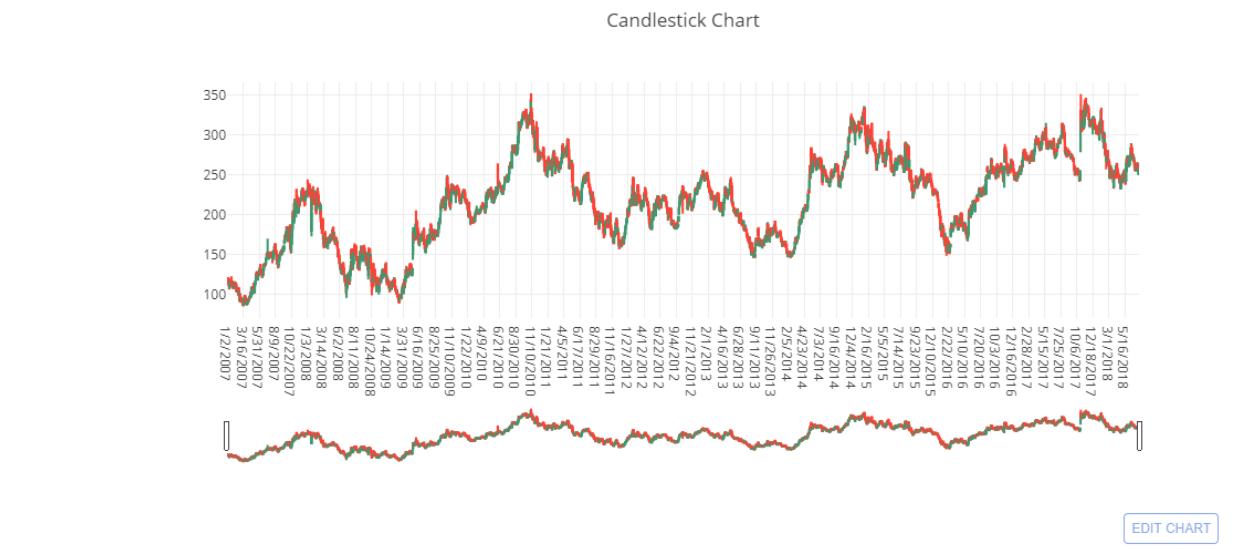


Out[8]:



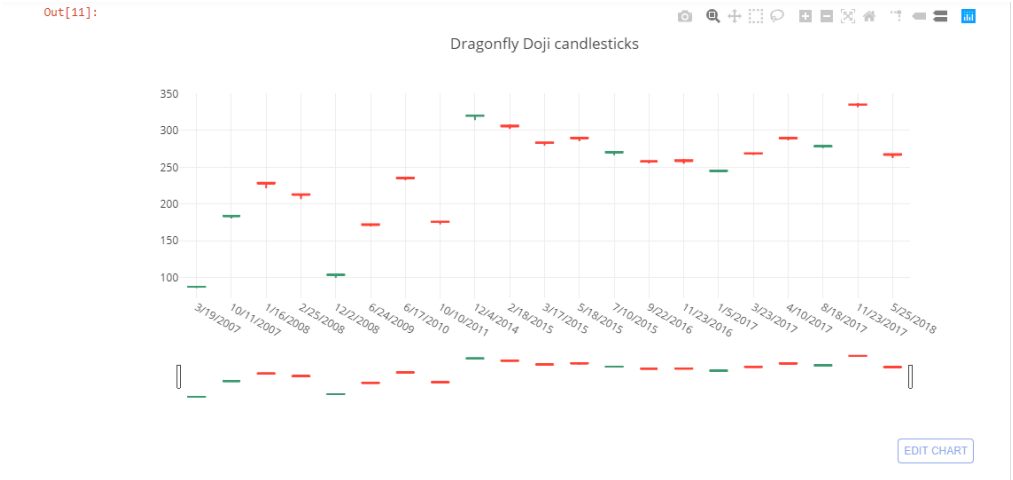
Candlestick chart

Out[10]:

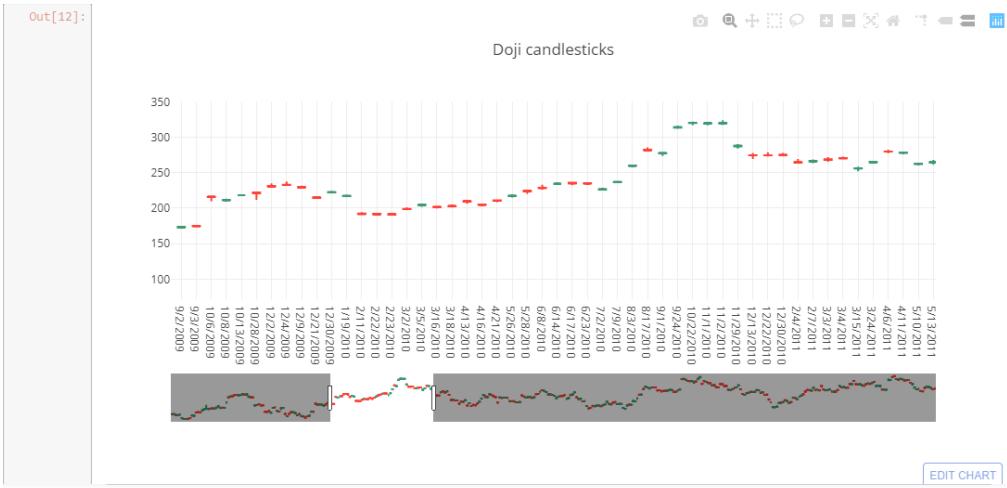


Plotting individual candlestick category to check.

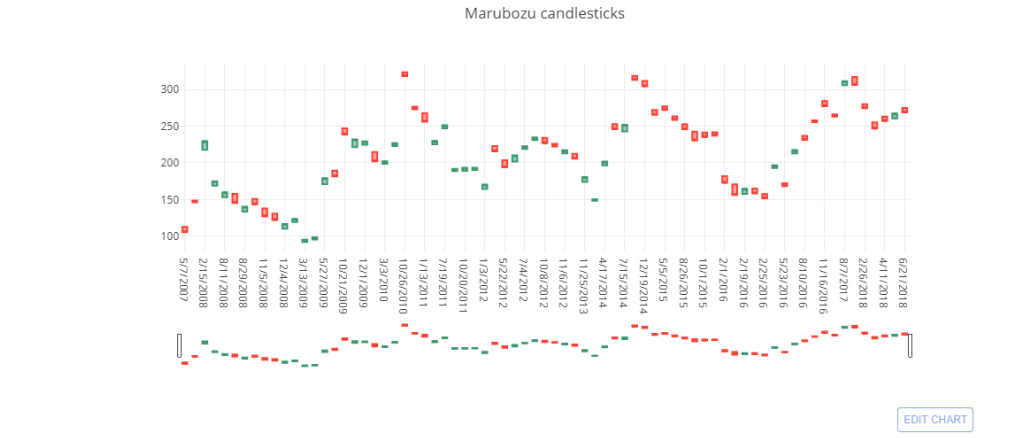
Out[11]:

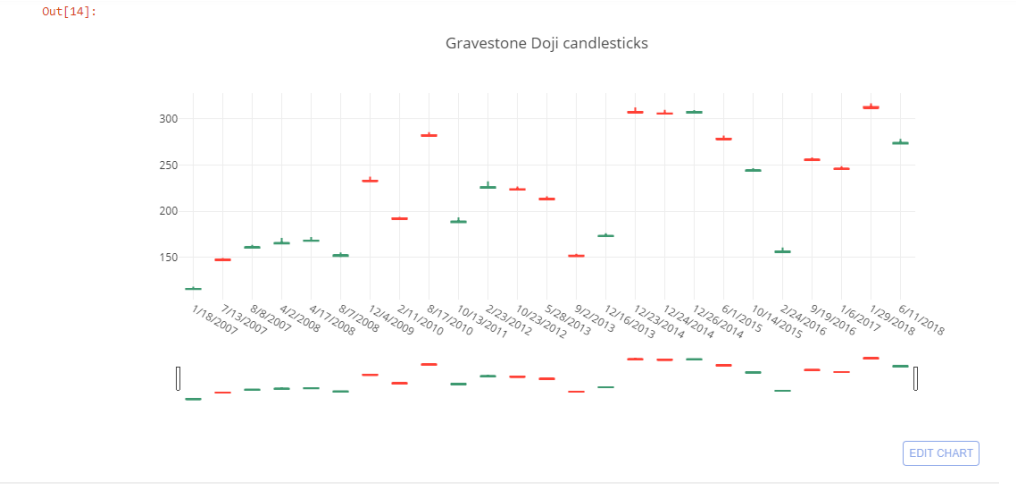


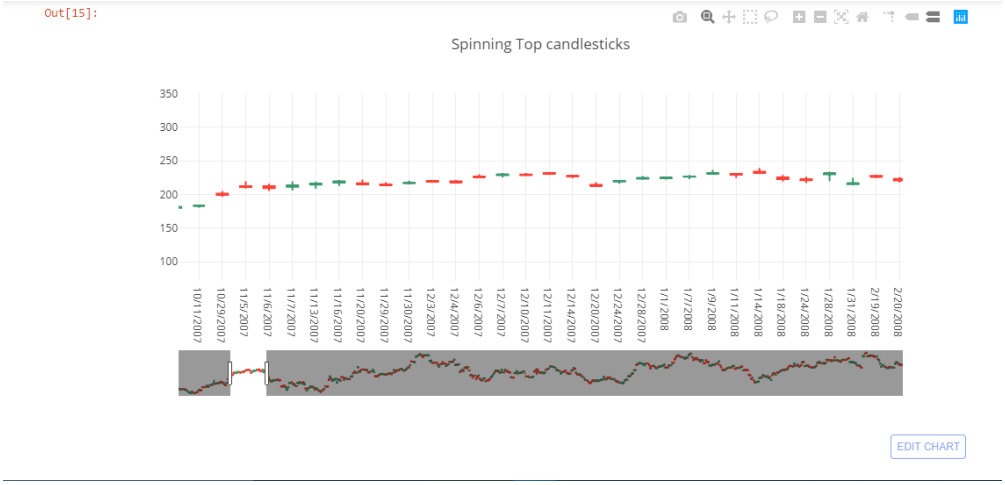
Out[12]:



Out[13]:







Now as we can see the categorized data is good enough. But there are huge number of undefined candlesticks that creates imbalanced data. We treat the imbalanced data and generate balanced training data.

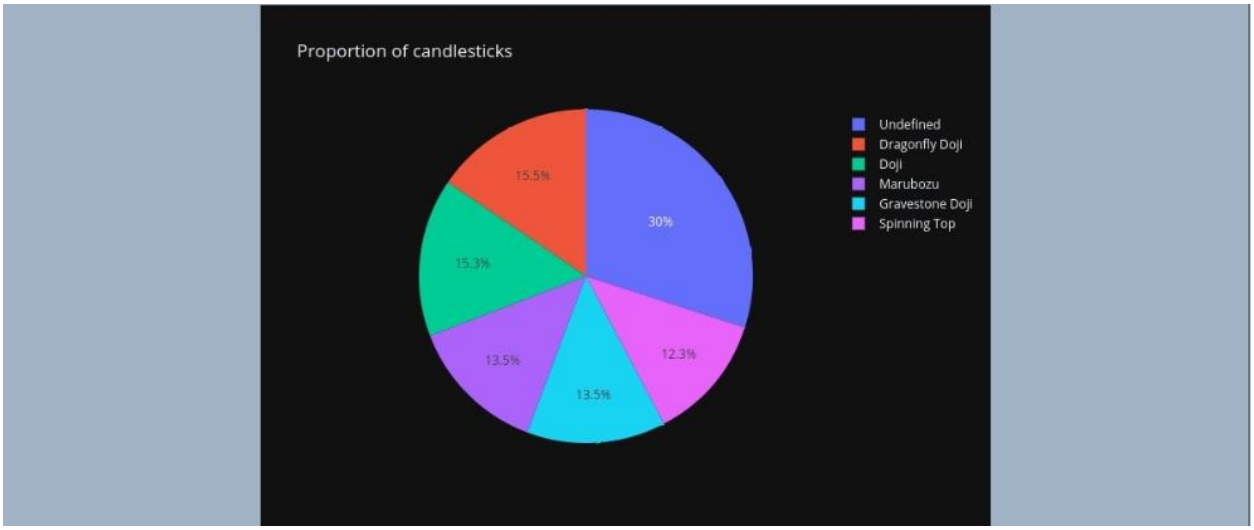
Training data

Training data is generated by adding data samples for the candlesticks in less proportion
Data samples for respective candlestick are created by adding or subtracting very small constant values from open, high, low, close stocks.

new_sample = [open, high, low, close] + 0.01
or
new_sample = [open, high, low, close] - 0.01

This remains the candlestick pattern while changes the data values.

Now as we can see we have the balanced data for training. We can continue to out model.

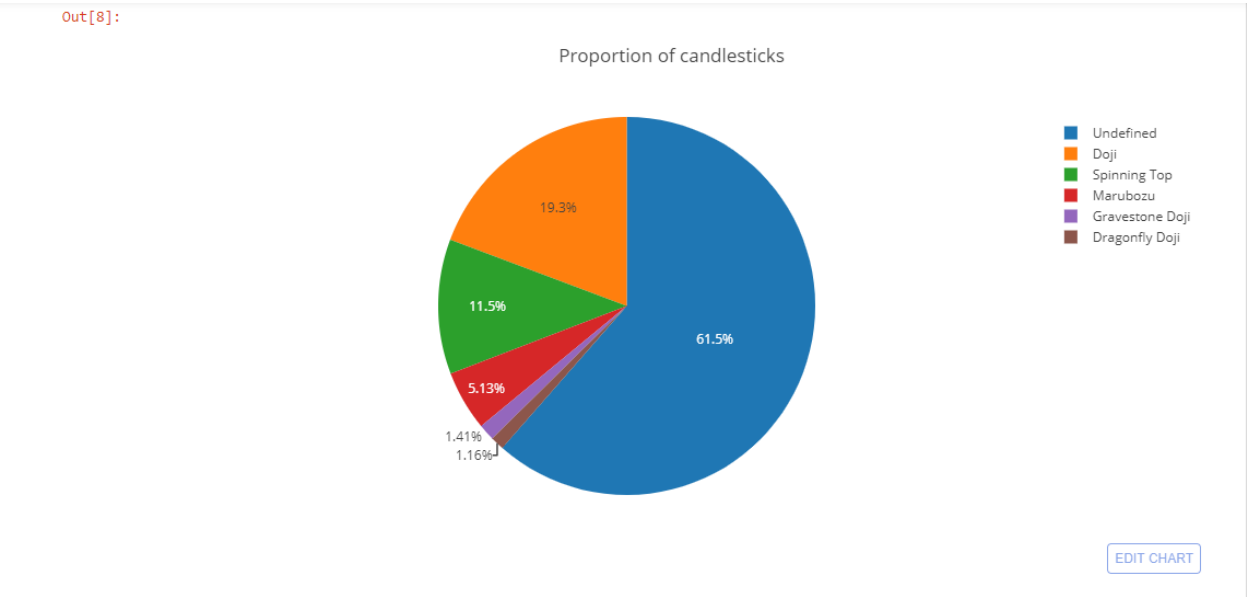


Classification model

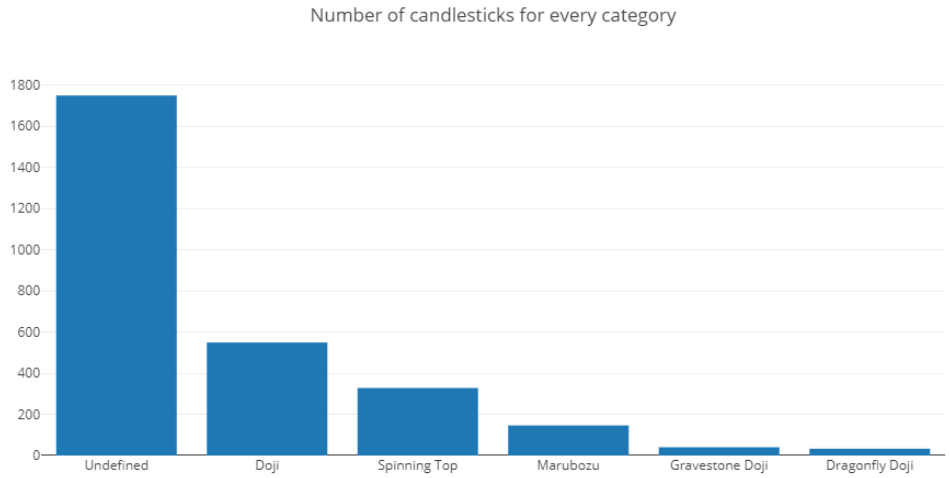
SVM classification model is trained on generated training data and tested on the data labeled by Ta-Lib library. It gave around 88% of accuracy.

Analysis on predicted data

Out[8]:

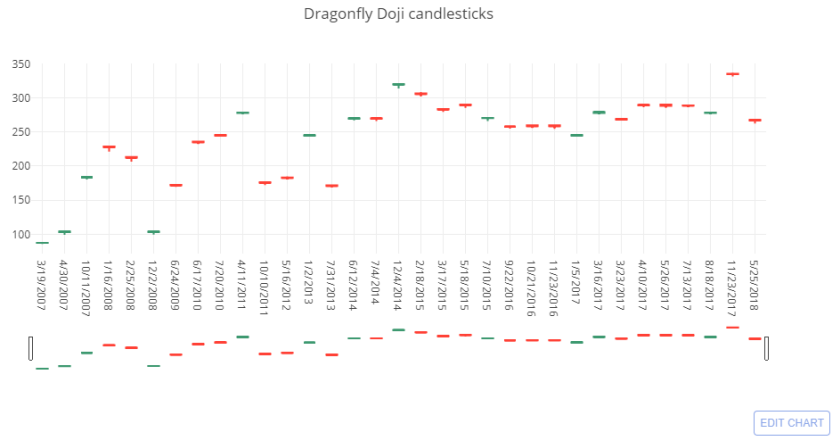


Out[9]:

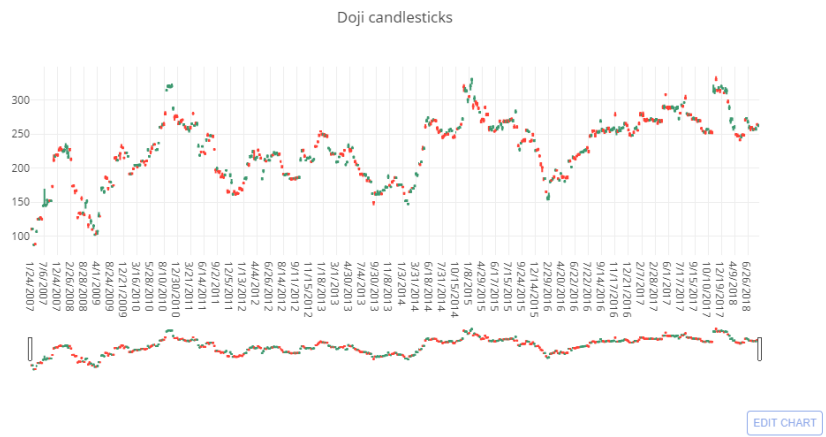


Plotting individual candlestick category on predicted data to check.

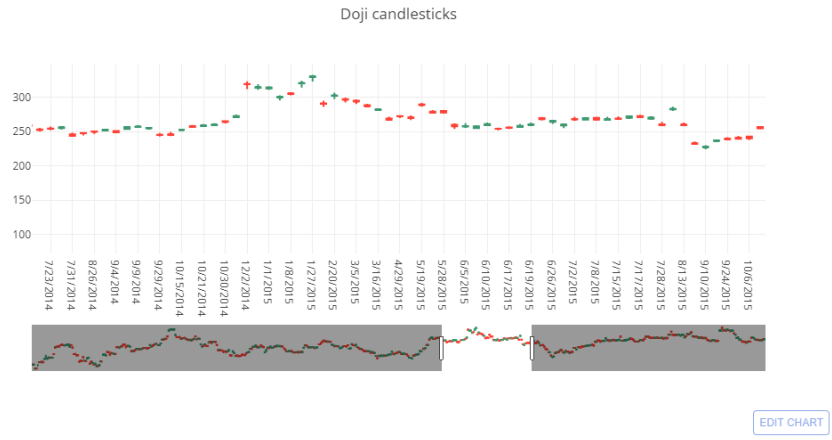
Out[11]:



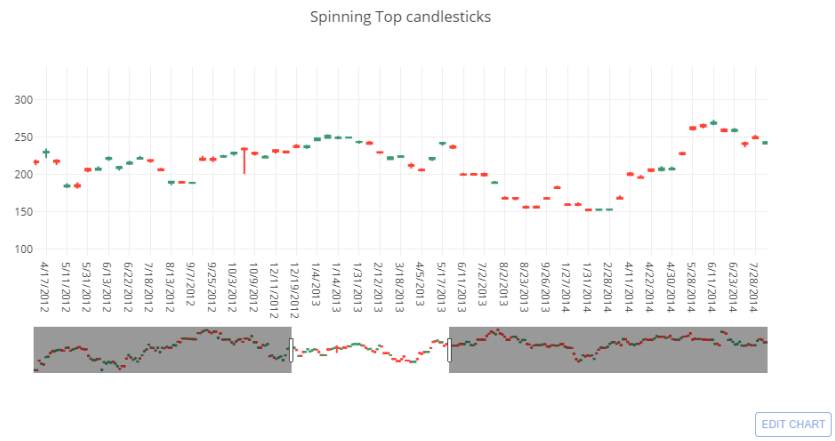
Out[12]:



Out[12]:



Out[15]:



Out[14]: