Foreign Exchange Normality Analysis and Modeling

A Data Science Project by James Solomon

Introduction

The global stock exchange has about $1.5 Trillion in traded daily volume, spread out amongst thousands of companies and tickers… Compare that to $6 Trillion in daily traded volume on the foreign exchange market, spread across about 12 currency pairs. This vast amount of volume leads to quick, and numerous amount price movement. If this price movement can be proven to stay within a consistent range, as well as consistently revert to it's mean (in a sense… follow a normal distribution), an automatic trading bot can profit using proper investment strategy and probability theory. This report aims to provide support to the hypothesis: 'The top traded foreign exchange rates follow periods of normality, separated by fundamental economic states.'

Why Care About Periodic Normality?

Normality is a profitable notion to prove since it can help as a heuristic when creating an automatic trading bot. A machine is be able to identify distributions over many timeframes incredibly efficiently. An investor with a proper outlook on global dynamics as well as an understanding of how normality applies to probability, can passively profit by applying this machine during long term periods of normality.

*This report modeled hourly data ranging from 2006 – 2021 for the USD/CAD exchange rate.

The long-term periods of normality identified in this report are:

- June 2006 - November 2014
- November 2014 - December 2021

The separating event between the two periods seems to be the Republicans taking control of the United States senate in November 2014. However, the reader of this report should take notice that a shift in economic state is not always evident until it has already shifted. Therefore, anyone using this hypothesis as a trading heuristic should following the fundamental status of the economies of focus closely to be able to halt trades during a shift and resume once the new distribution has been identified.

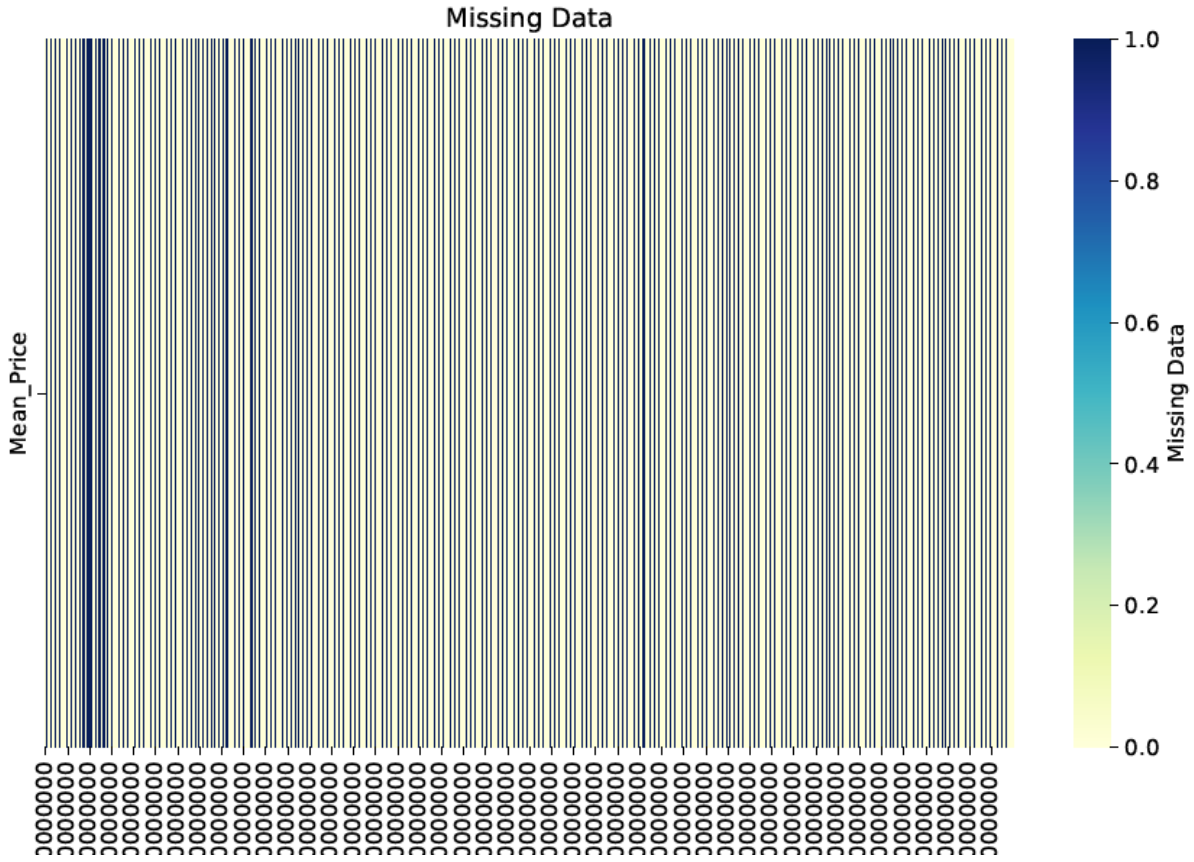
Data Wrangling

The raw data used in this report was downloaded off of: histdata.com. It came formatted as 20 csv files, each representing one years worth of minutely tick data for the USD/CAD foreign exchange rate. Initially the columns were unlabeled, so the author of this report manually labeled the columns according to their respective

titles: DateTime_Stamp, Bar_OPEN_Bid_Quote, Bar_HIGH_Bid_Quote, Bar_LOW_Bid_Quote, Bar_CLOSE_Bid_Quote, Volume. Through further analysis, the author of this report found that all values under the 'Volume' column were 0, therefore this column was immediately dropped as it contained no useful information. Afterwards, the 'Mean_Price' column was added by taking the mean of the Bar_HIGH_Bid_Quote and Bar_LOW_Bid_Quote columns. All columns except for 'Mean_Price' and 'DateTime_Stamp' were subsequently dropped, as these are the only columns left that contain relevant information that can be used for modeling. Finally, the DateTime_Stamp column was set as the index for the data frame.

Data Cleaning

Before feature engineering can begin, the author of this report must remove all nulls from the data set. The set initially came in 6 million rows of minutely ticks. This is reasonable as it is representing the minutely price movement of a foreign exchange pairing over a 20-year period. However, there is no data being recorded on weekends on holidays, therefore there are many hours missing within the data frame once the data was grouped into hourly rows. The missing data is represented by the blue streaks in the visual below:

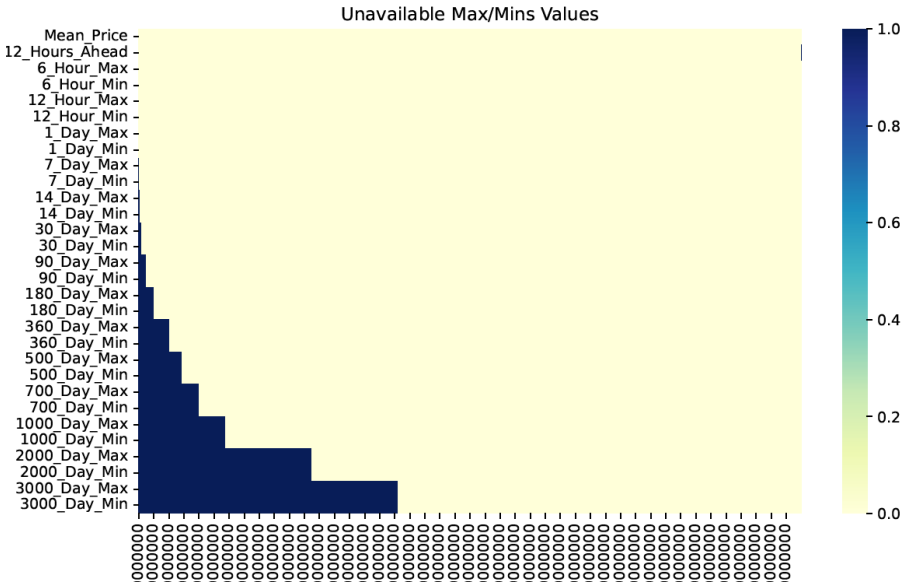


As shown above, there are numerous holidays and weekends that have missing data, and therefore will be forward filled as this is the most realistic representation of the real world data. Since most often, after hours prices are close to if not the same as closing prices. After forward filling the null data, there are officially no more nulls in the raw data set. The data frame started with 6 million rows of minutely data and is now condensed down to 180 000 rows of hourly data, made up of 2 columns: DateTime_Stamp, and Mean_Price.

Feature Engineering

The author of this report decided to use a 12 hour forward price prediction as the prediction target for this report's models, and the maximums and minimums of various ranges as the main features of the data frames. In order to record the train values for the forward price predictions, the author of this report shifted the Mean_Price column upwards by 12 columns (the amount of hours ahead the model is predicting). Now the data frame has three columns: DateTime_Stamp, Mean_Price, 12_Hours_Ahead. The features were engineered by applying a rolling window to various ranges, and using a max and min function respectively on those windows.

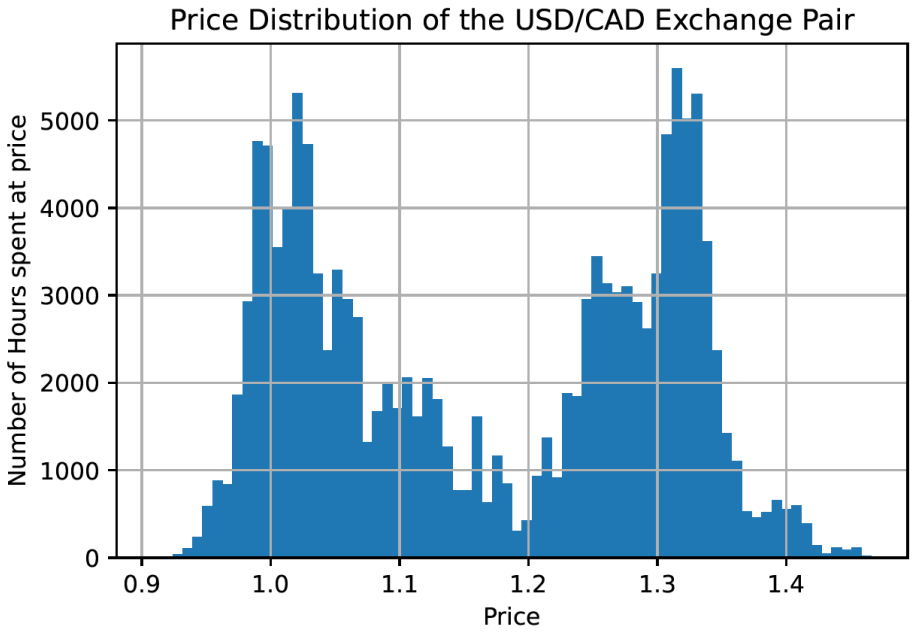
Once all the features have been engineered, implicitly there will be nulls for ranges that have not finished yet… as it is impossible to have a 300 day maximum before 300 days have passed. These missing data points are represented by the blue bars in the following graph:



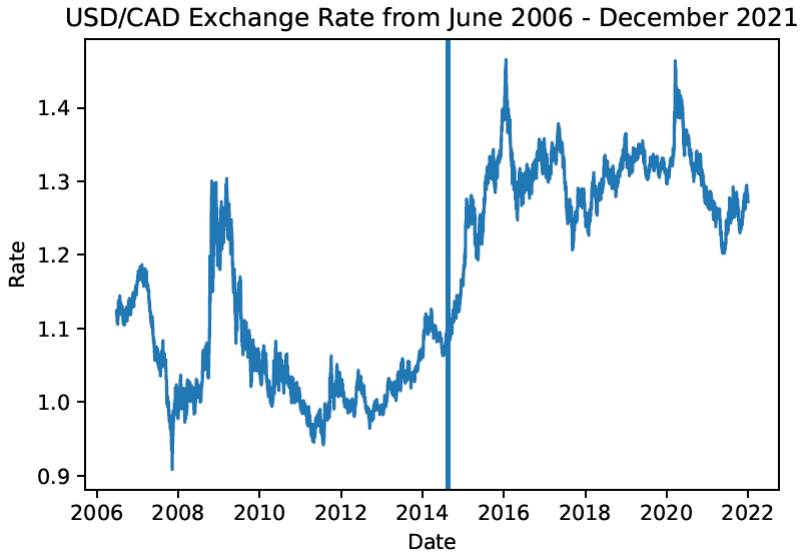
Therefore this report concluded to drop the columns 3000_Day_Max and 3000_Day_Min, as the nulls within this column account for 40% of the dataset. The rest are kept to help the machine better identify long term normal distributions. After dropping the above columns, this report shaved off the top 48 000 rows, resulting in 136 000 rows. This process allows the model to start off knowing every value it needs to know.

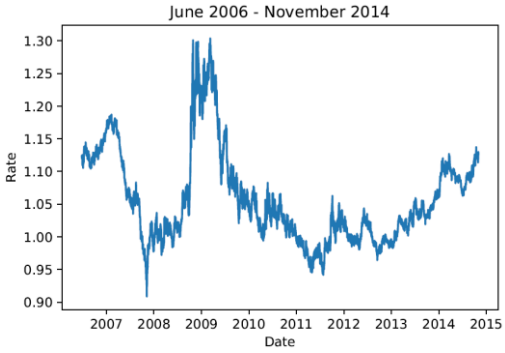
Distribution Analysis

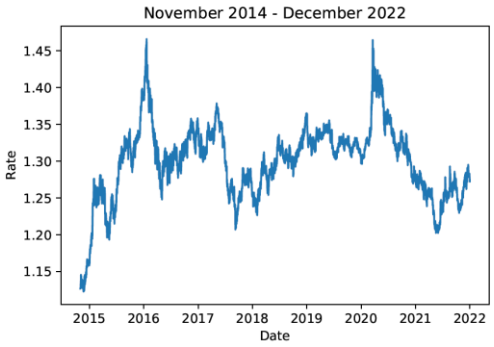
Now that the data frame is completely clean, and all the features have been engineered, some analysis can be done. As shown below, when the entire data set is put into a histogram, two clear modes seam to appear: 1.03, and 1.32. This report will attempt to separate the dataset by date in order to achieve two subsets of normally distributed prices.



The separation date identified is November 2014, as this is the date where the economic state of the United Stated and Canada seems to shift considerably.

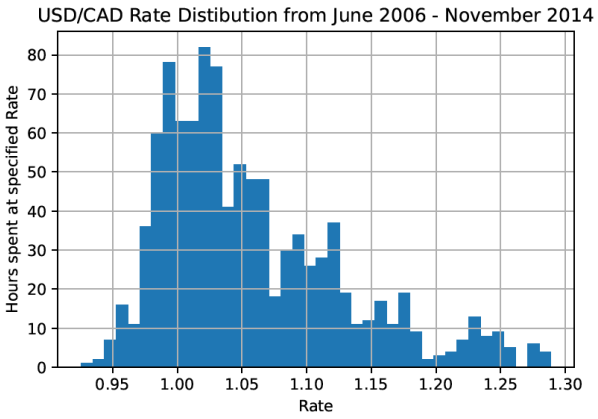


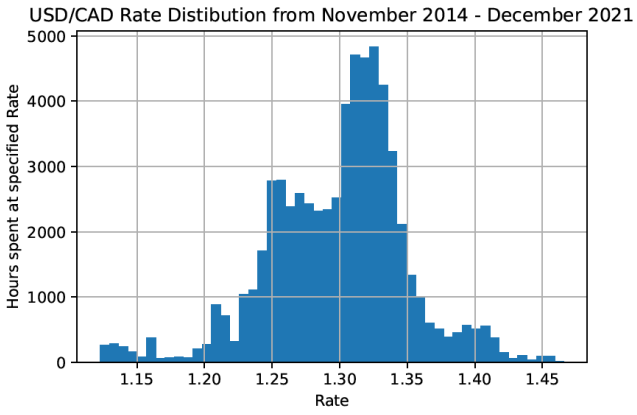




Since this shift occurs for fundamental reasons it cannot not be immediately picked up by this machine. However, once the shift does occur and the new mean is set, the rates starts a new long term pattern reverting back to the mean 1.32.

To further prove the normality, a histogram is plotted, and a Shapiro-Wilk test is applied to both subsets:





Shapiro-Wilk P-Value: 0.23

Filtered outliers above 1.13, as during this time the rate was only above this price about 5% of the time. Representing a separate distribution for about 6 months during 2008 with a mean of 1.23. And 3 months in 2007 with a mean of 1.17. Given more time, this report would further analyse these short normal distributions.

Shapiro-Wilk P-Value: 0.24

No filter applied.

As shown, both cases exhibit P-Values well above the generally accepted threshold of 0.05, and the histograms both display traits of slightly skewed normality. Both of these normally distributed subsets were taken as large consecutive chunks of time, supporting the hypothesis: The top traded foreign exchange rates follow periods of normality, separated by fundamental economic states.

Modeling

*Please note that the models implemented in this report are strictly experimental, and meant to display the relevance of multiple range max/mins as features when predicting foreign exchange rates. The Max/Min features are meant to act as a heuristic for the model in determining the range of distribution.

Linear Regression Model:

The first model this report attempted to fit was a linear regression model. Linear regression was a natural first choice, as the data this report is working with is completely continuous. After confirming that all the columns are heavily correlated with the target column through multiple scatter plots, the linear model is instantiated. The initial X train and test sets need weights to be assigned to each value for the model to work.

Therefore, this report adds constants to each value in the test set so they can be fit to the linear model. Once the model is fit, the reader can notice a variety of features with p values > 0.05. In order to remedy these ineffective features, this report applied PCA to optimize down to 6 features. The final p values are now all below 0.01, and the R-squared value is 99.7%, making this an accurate model with incredibly low residuals.

Fully Connected Feed-Forward Neural Network:

The next model this report attempted to fit was a neural network. Using the already curated PCA components, this report instantiated a fully connected feed-forward neural network with 1 input layer, 3 hidden Dense layers (consisting of 30, 60, and 80 nodes respectively) and an output Dense layer consisting of 1 node indicating the 12 hour forward price prediction. Resulting in an R-Squared score of 99.7%.

Recurrent Neural Network:

The last model implanted was a recurrent neural network with 1 hidden LSTM layer and 2 hidden Dense layers (1 of which with a drop out factor of 0.15). The dropout layer is meant to reduce the amount of learning done at each iteration within the Dense layer, making the model less overfit. This model resulted in an R-Squared score of 99.5%, which demonstrates how predictive multi range max/mins can be.

Conclusion

The resulting recurrent neural network received an R-squared score of 99.6% when tested against the unseen test data. The author of this report considers this accuracy score a success, as it show that various ranges of max/min data can be used to intelligently predict future prices. This is only possibly because the USD/CAD foreign exchange rate follows periods of normality as proven in this report. If an economist is able to determine when two countries are going through a change in dynamic, they can (and should) turn this strategy off and wait for a new normal distribution to set.