[Analysis 2](#_Toc26878069)

[Introduction 2](#_Toc26878070)

[Executive Summary 2](#_Toc26878071)

[Exploratory Data Analysis 2](#_Toc26878072)

[Criteria Match 12](#_Toc26878073)

[Trading Strategy 12](#_Toc26878074)

[Momentum Features / Old School trend following 13](#_Toc26878075)

[Calendar Features 14](#_Toc26878076)

[Quantitative Model Implementations 14](#_Toc26878077)

[References 15](#_Toc26878078)

# Analysis

## Introduction

The purpose of this document is to socialize the train of thought and ideas behind the basic usage of Machine Learning in Applied Finance. This document is divided into 3 parts which includes the data loading, data transformation applied and the conclusion.

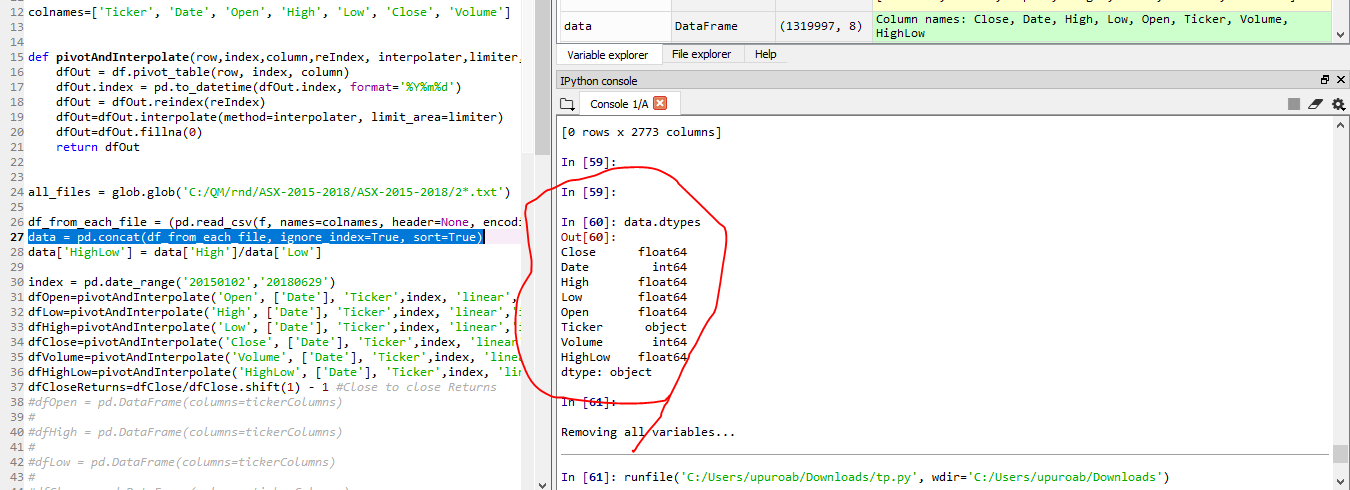
### Executive Summary

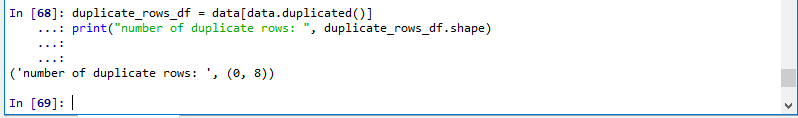
The end goal of any machine learning backed trading system is to make money while managing the risks. This document specifically covers data integrity risks and this risk is managed by objectively looking at the forest as well as trees, outliers and other mathematical abnormality with a data utilitarian mindset, i.e using statistical check over the whole data test rather than eyeball them individually. We have applied the thought that stock being a derivative is either having a 0 price or its price follows a trend, and finally applying a research backed trading strategy for sending buy and sell signal against a security. The 5 basic trading strategies which have been implemented are listed in increasing order of accuracy →

1. Linear Regression [Least Accurate]
2. K-Nearest Neighbour
3. ARIMA
4. Prophet
5. LSTM [Most accurate]

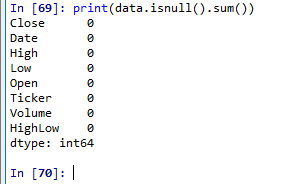
## Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the process of summarizing the data sets by their main characteristics using statistical methods and visual aides. This step is an important step before we arrive at modeling the data

1. Data Type Check - Pandas data frame containing all the asx data was checked for data type, to rule out data type mismatch, like special character or alphanumeric data in place of integer or float for stock price (Open, Close, High, Low, Volume) and int64 for Date.   
   
2. There is no need to drop any irrelevant column.(using dataframe drop)
3. The column names are clear and self descriptive so no need to rename column names. (using dataframe rename).
4. There are no duplicate rows



1. There are no null values or missing values in the csv. Though there are missing values at a functional level by missing data for dates and it will be handled in later stages.



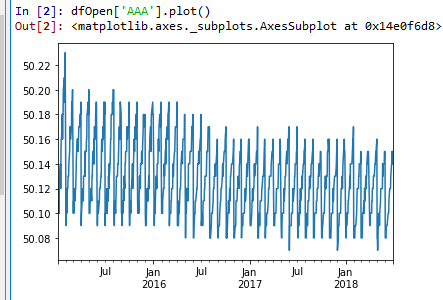
1. Data Transformation - There are 3 types of missing data as a part of time series, the details are as follows along with their resolution
   1. Data not available at the start of the time period i.e. the stock price is not available on 2nd Jan 2015 but is made available at a later data. This could be because the stock price had their IPO later and there is no data history.
   2. Data not available in the middle, i.e. we have public holidays, weekends which may result in missing data
   3. Data not available in the end - this could be due to de-listing of stock.

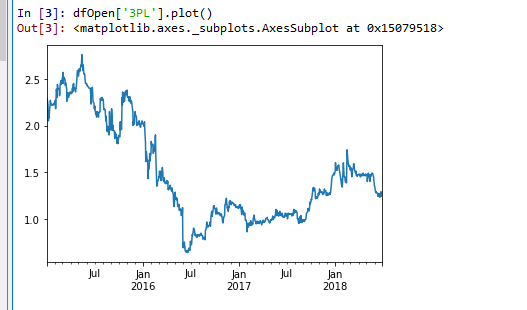
*Resolution - Any stock price which is missing in the middle is interpolated linearly but in not extrapolated.*

*The remaining rows towards the start and the end with no value are set to 0, since the stock is a derivative, it should always have a value 0 or a finite number. The code is updated to reflect this change.*

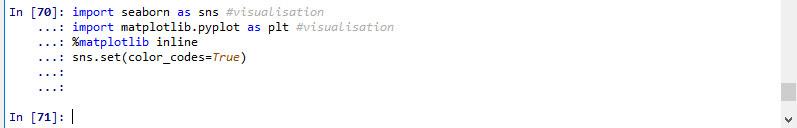
1. Detecting Outliers - The outliers are checked using the following techniques
   1. Price Chart -

Opening price chart of ticker AAA and 3PL

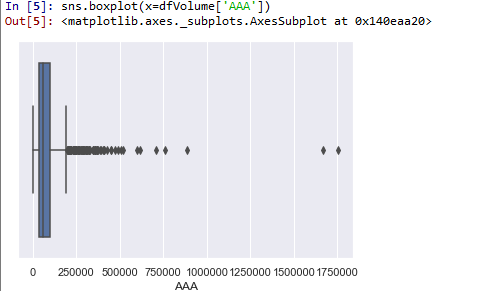




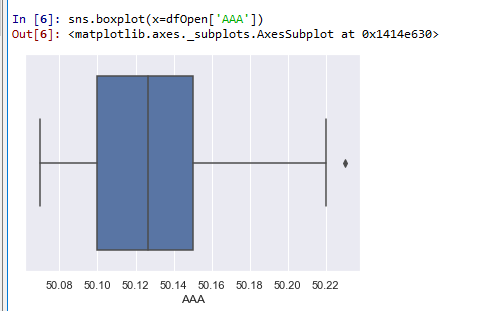
* 1. Box Plot - In order to set the correct baseline for box plot, the Open/High/Low/Close/Volume data will be compared for each stock individually for Box plot.
     1. Importing required libraries



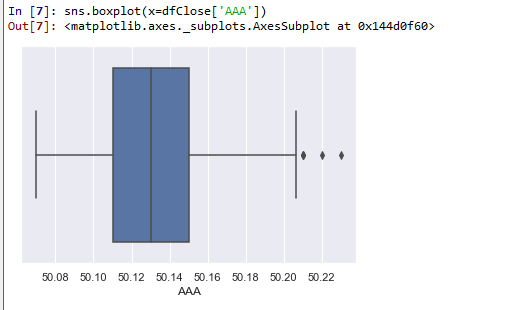
* + 1. Volume



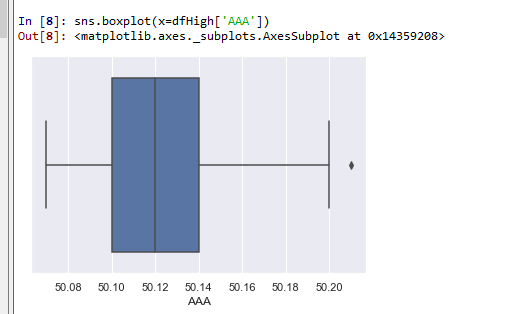
* + 1. Open Price



* + 1. Close Price



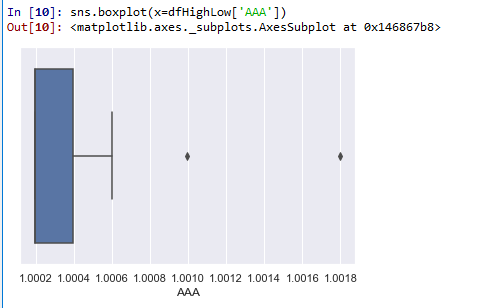
* + 1. High Price



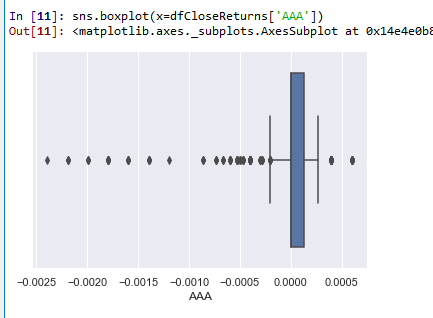
* + 1. Low Price



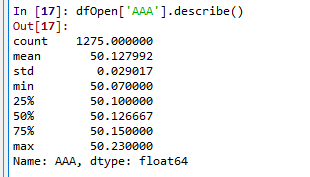
* + 1. High Low ratio



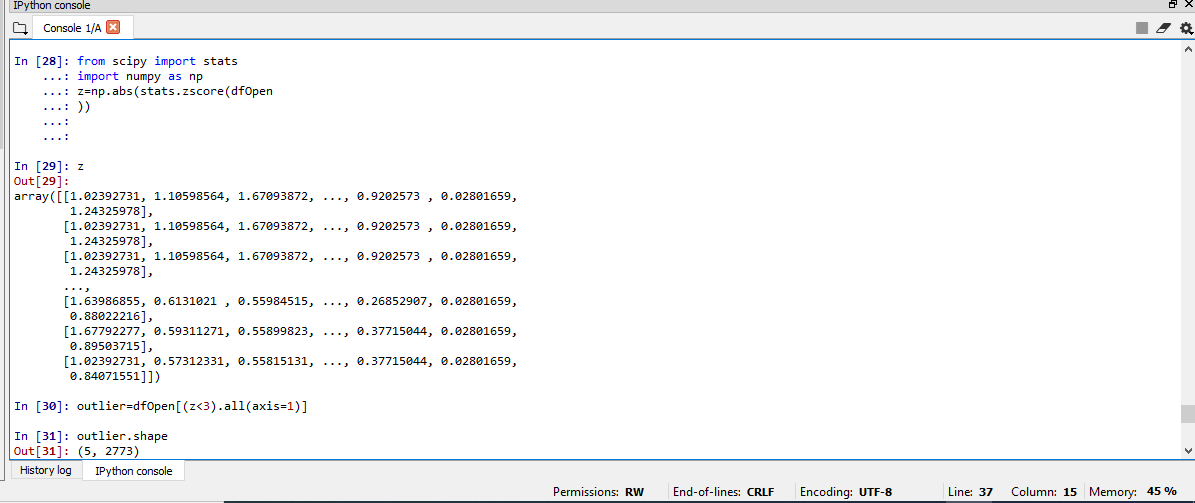
* + 1. Close to Close ratio



* 1. Statistical details



* 1. Scatter plot - requires more than 1 variable to show correlation so not applicable in this data set. For e.g if you are correlating horsepower to price of a car.
  2. Z-score - The Z-score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. While calculating the Z-score we re-scale and center the data and look for data points that are too far from zero. These data points which are way too far from zero will be treated as outliers.



This indicates that open price has few outliers.

* 1. IQR - Wikipedia Definition

The interquartile range (IQR), also called the midspread or middle 50%, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1.

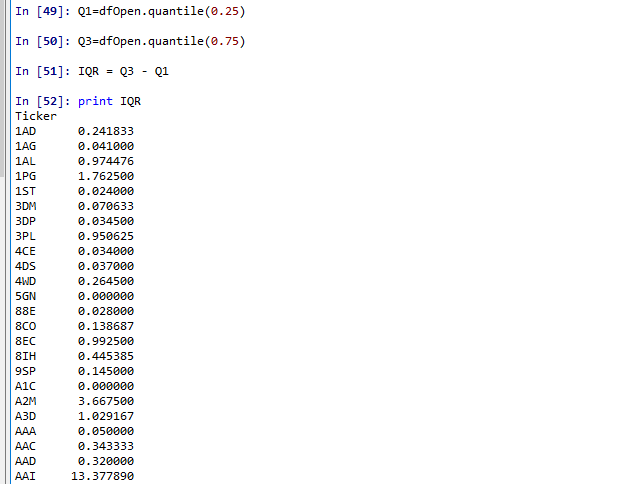
In other words, the IQR is the first quartile subtracted from the third quartile; these quartiles can be clearly seen on a box plot on the data.

It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outliers.

Using IQR score the outliers can be removed from pandas dataframe

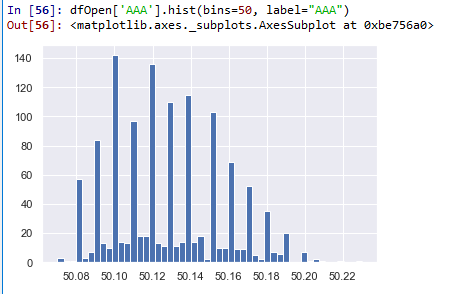
Code Snippet -

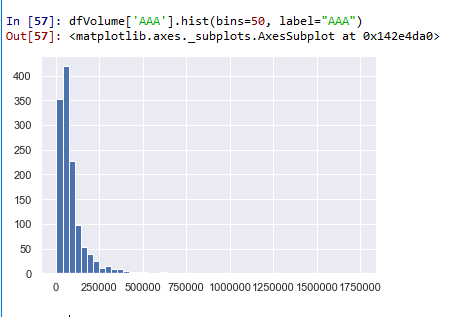
*outlier\_IQR=dfOpen[~((dfOpen < (Q1 - 1.5 \* IQR)) |(dfOpen > (Q3 + 1.5 \* IQR))).any(axis=1)]*



* 1. Histogram of Open and Volume for Ticker AAA

Matplotlib histogram is used to visualize the frequency distribution of numeric array by splitting it into small equal-sized bins, which in layman terms means how many times Open price was 50.14 or how many times AAA trade at a volume of 250k





* 1. The Heat Map procedure shows the distribution of a quantitative variable over all combinations of 2 categorical factors. If one of the 2 factors represents time, then the evolution of the variable can be easily viewed using the map. A gradient color scale is used to represent the values of the quantitative variable. The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

*This however requires a very powerful hardware, but this map helps uncover information not easily available which can be used as an edge while betting on financial markets. In this instance my hardware couldn’t cope with the heatmap for all stock ticker.*

### Criteria Match

1. Coverage and Quality of data

Missing values - Yes! There are 3 types of missing values

* + 1. No data available from Day 1 of the range, 2 Jan 2015 but available from Day n. It is handled by filling 0s
    2. Intermediate data missing - most likely weekend and public holiday data, which is filled by linear interpolation (no extrapolation).
    3. No data available from Day n onwards till the last day of the range, - 29th June 2018. It is handled by filling 0s

Additional Notes - there are various other ways to handle missing values, but they are not free, or computationally exhaustive or programmatically complex, thus time consuming. They are →

* Buy data outright - from Bloomberg or DTCC. So we are 100% sure data is clean
* Missing data imputation techniques / Feature engineering
  + Using Median or Mode replacement
  + Hot Deck (LOCF) or Cold Deck
  + Reconcile data with Quandle or Alphavantage using alpha vantage
  + Use Stock AI package against each stock to verify.

1. Time Series nature of data helps with forecasting and building an effective trading strategy, at a high level there are 3 ways to forecast a time series data
   1. Basic forecasting - rolled mean / Autoregressive model
   2. Machine Learning forecasting - Random Forest / Prophetic forecasting / Support Vector machines
   3. Deep Learning forecasting - Neural network / Long short term memory forecasting

## Trading Strategy

At a high-level systemic trading can be divided into 2 broad categories →

1. Simple trend following systems which used momentum indicators to buy / sell signal. These were made famous by Larry Hite, Richard Dennis (Turtle Trading System) ,Ed Seykota among the list of great traders
2. Trend following was a great money spinner till the widespread use of computers from the late 90s, and this brought the next generation of traders called quants who were using more sophisticated trading system backed by research. This revolution was started by Jim Simons from Renaissance technologies. Renaissance was using BIG Data to predict prices 15-20 years before the term was coined.

Currently the quantitative trading is going through the observer effect, *the theory that the mere observation of a phenomenon inevitably changes that phenomenon*. Which means all the markets have become immune to all vanilla machine learning models. The race is now to find the next generation of more sophisticated and advance trading models. For the purpose of the assignment both the old school trend following system and quantitative model have been discussed in detail.

### Momentum Features / Old School trend following

A moving average (MA) is a widely used indicator in technical analysis that helps smooth out price action by filtering out the “noise” from random short-term price fluctuations. It is a trend-following, or lagging, indicator because it is based on past prices.

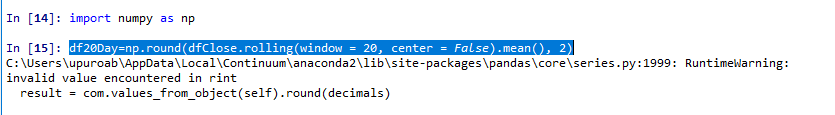
We will demonstrate a *moving average crossover strategy*. We will use two moving averages, one we consider “fast”, and the other “slow”. The strategy is:

1. Trade the asset when the fast moving average crosses over the slow moving average.
2. Exit the trade when the fast moving average crosses over the slow moving average again.

Fast moving averages is made up of average of small number of recent data (for e.g. 20 day) and more closely follow the stock, while slow moving averages is made up of larger number of recent data (for e.g. 200 day), resulting in them responding less to the fluctuations of the stock and being more stable.

20 day average →

df20Day=np.round(dfClose.rolling(window = 20, center = False).mean(), 2)



### Calendar Features

Some of the basic calendar related data like ‘Year’, ‘Month’, ‘Week’, ‘Day’, ‘Dayofweek’, ‘Dayofyear’, ‘Is\_month\_end’, ‘Is\_month\_start’, ‘Is\_quarter\_end’, ‘Is\_quarter\_start’, ‘Is\_year\_end’, and ‘Is\_year\_start’ can be deduced using *fastai* python package.

### Quantitative Model Implementations

The following 5 trading strategies have been implemented. The respective python script files have also been uploaded on google drive.

1. Linear Regression - is a simple technique and quite easy to interpret but requires a lot of data mining and data enrichment with functional, technical and volatility indicators to give accurate predictions.
2. K-nearest neighbour - finds the similarity between new data points and old data points. It works well with indexes or other low volatility instruments or top end real estate prices.
3. ARIMA is a very popular statistical method for time series forecasting. ARIMA models take into account the past values to predict the future values. There are three important parameters in ARIMA:

p (past values used for forecasting the next value)

q (past forecast errors used to predict the future values)

d (order of differencing)

1. Prophet, designed and pioneered by Facebook, is a time series forecasting library that requires no data preprocessing and is extremely simple to implement. The input for Prophet is a dataframe with two columns: date and target (ds and y).Prophet tries to capture the seasonality in the past data and works well when the dataset is large.
2. LSTMs are widely used for sequence prediction problems and have proven to be extremely effective. The reason they work so well is because LSTM is able to store past information that is important, and forget the information that is not. LSTM has three gates:
   1. The input gate: The input gate adds information to the cell state
   2. The forget gate: It removes the information that is no longer required by the model
   3. The output gate: Output Gate at LSTM selects the information to be shown as output

### 

## References

<https://www.investopedia.com/terms/m/movingaverage.asp>

Achelis, Steven B. *Technical Analysis from A to Z*. Probus, Chicago, 1995

Blau, William. *Momentum, Direction, and Divergence.* John Wiley & Sons, New York, 1995.

Bollinger, John. *Bollinger on Bollinger Bands*. McGraw-Hill, New York, 2002

Teleb, Nassim Nicholas. *Fooled by Randomness*. Texere, New York, 2001

Thorp, Edward O. *Beat the Dealer.* Vintage, New York, 1966

Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long‐short term memory. PLOS ONE, 12, 1–24.