**Electronic Assignment Cover sheet**

**Student (s) Number as per your student card:** 10505324

**Course Title:** Master of Science in Data Analytics

**Lecturer Name:** Terri Hoare

**Module/Subject Title:** Data Mining

**Assignment Title:** CA 2

**Time-series Analysis Foreign Exchange Rates with ARIMA model**

Foreign Exchange rate is rate of currency of one country expressed in the terms of currency of any other country. This exchange rate keeps changing depending on the demand and supply of a currency for a time frame. There are many political as well as economical decisions which affect rates of currencies with respect to other countries. But this document describes an attempt to forecast currencies with the help of time series as well as depicts my learning outcomes from the CA. For the prediction of Foreign Exchange rates, CRISP-DM methodology is adapted, and all the steps of CRISP-DM are followed. CRISP-DM is Cross Industry Standard Process for Data Mining. It has been proved and adapted to be most effective process for Data Mining. Following are the steps involved in CRISP-DM: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment.

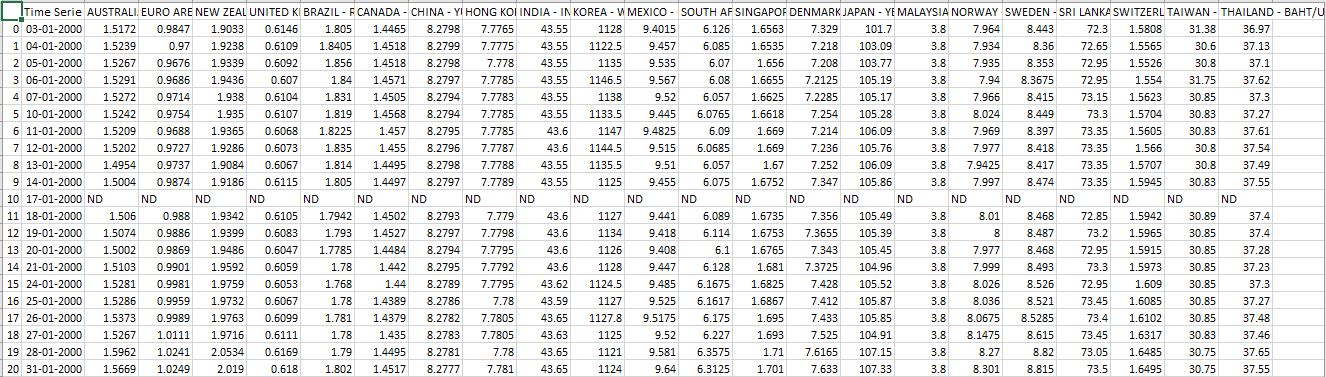
**Business Understanding:**

Business understanding is the main step of any data mining project, as it defines the goals of the project, lay out a preliminary plan that should be followed. Thus, business understanding contains the main reason why the project should be carried out and how it should be carried out. In case of forecasting Foreign Exchange rates there can be several reasons. Some traders use foreign currency as an investment. If such traders can predict the rates of currencies beforehand, they can plan out the more precise time to buy any currency or to sell the currency. No doubt, rates of currency do depend on other factors, such as political decisions, economical decisions, etc. but if there is an idea when the rates are low or when are they at peak during the year it can help the investors to plan and arrange money accordingly. Therefore, in this step business goals, objectives, requirements or how the business works can be learned.

**Data Understanding:**

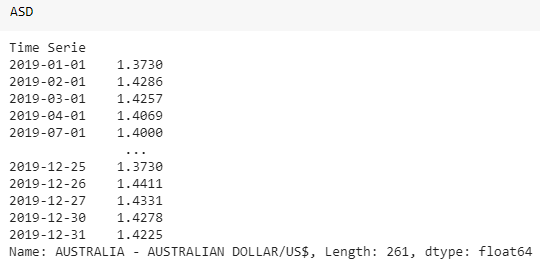
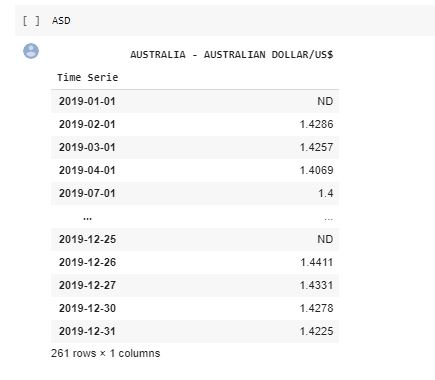
Data understanding involves tasks like data collection, exploration of data, looking if the available data is enough for the analysis and can the business goals be achieved from the available data. Data Understanding basically is the process of making judgement about the data before starting any of the analysis work and deciding whether the data is appropriate for the analysis. As CRISP-DM is cyclic in nature, if it is found out that the data is not suitable according to the business goals then the data or the source of data can be changed. For the forecasting of currency using a time series analysis, data was collected from Kaggle.com and was in the csv format. The data had values of currencies of many countries in the term of dollars. From initial analysis it was observed that data was in the normalized form with no outliers, but few missing values which required data preparation. Thus, from initial analysis it is found that data is suitable to proceed with next step of data preparation.

The csv file had data foreign exchange rates converted in US dollars format for 24 countries from year 2000 to 2019. Below fig is the screenshot of CSV file.



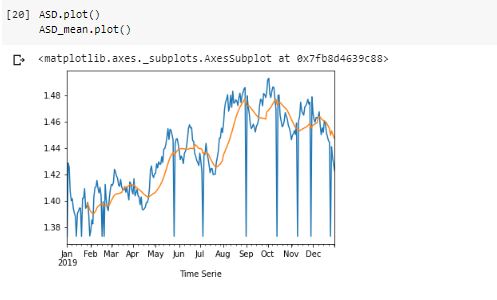
**Data Preparation:**

Data preparation involves tasks like data cleaning, selecting a dataset from huge data source which is necessary for analysis, removing unnecessary columns, checking whether all the columns are in the normalized state or not. Basically, data preparation involves making data ready for model learning. Machine Learning model can learn the trends and patterns in the data more accurately when the data is well prepared. If the model fails to learn with the prepared data, then again, the data is passed to the data preparation state. In the data of foreign exchange currencies focus was on the ‘Australian Dollars’, so from entire dataset that column was selected, and a dataset was formed. From the dataset all the missing values for holidays was replaced with the mean value. Then for model application only latest data of year 2019 was considered. Below figure gives the comparison of dataset before and after preparation



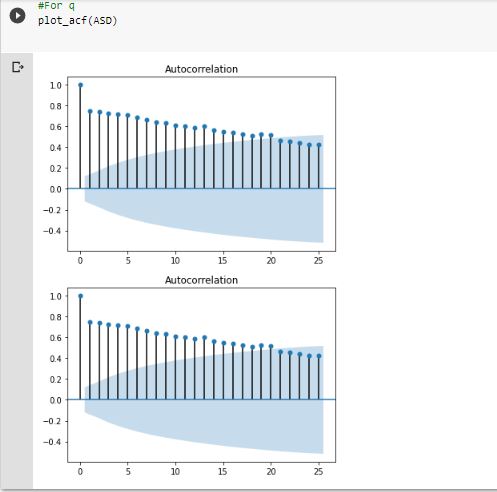
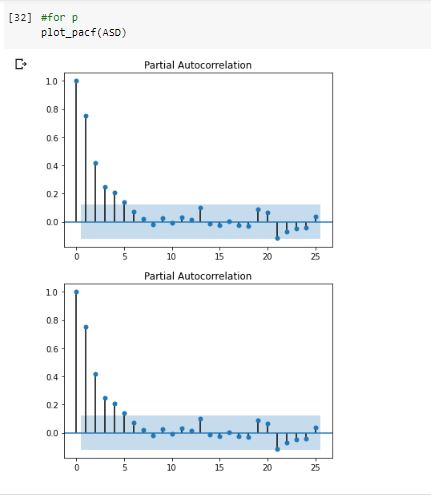
Before Data Preparation After Data Preparation

Before applying the model if the dataset contains any trends or seasonality was checked using moving average technique. But there were no trends or seasonality found. The data was stationary, so the data was passed further to the data modelling part. Below figure gives the graph of moving average.



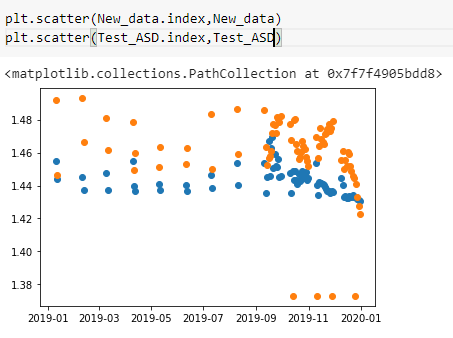
**Modelling:**

This step of CRISP-DM includes the selection of models to be used for the purpose of analysis. For example, if the business goal is to classify the dataset in to sets then classification model is used otherwise for forecasting prediction models are used. Similarly, if unsupervised learning is to be carried out then clustering or association models are selected. When the model selection is done then parameters required for the model are identified. Further the dataset is split into training and testing dataset. Model is built using this training dataset. For forecasting the currency rate, ARIMA model was used. ARIMA is time series analysis model which includes 3 parts Auto Regressive, Integrated and Moving average. Therefore, it has three parameters, p, d, q, respectively. p is for auto regressive model which is calculated using pacf plot; d is for integrated difference which depends on the trends or seasonality presence in dataset, as there was no trend or seasonality d = 0 was used. Finally, third parameter is q which for Moving average and is calculated using acf. Further, when the parameters were configured model was built on 70 % of the dataset and 30% was used for model testing. The testing is carried out using accuracy metric RMSE that is root mean squared error. The RMSE is found to be 0.021 which is very low. PACF and ACF graphs are mentioned below

ACF plot PACF plot

Comparison between actual and predicted values.



**Evaluation:**

Evaluation involves evaluating the model built and checking whether it reaches the success criteria defined in business understanding. If the success criteria are reached, then the model is passed for the deployment otherwise the process is again started in the cyclic manner. The entire process is again reviewed to see if any improvements are possible. If the model is to be passed to the deployment phase, then plans are made accordingly for deployment process. In our project Evaluation was carried out by calculating RMSE and MAPE values. RMSE finds the variance of the residuals in the prediction. Residual is the difference between the regression line created by model and the actual point. It is calculated by formula:

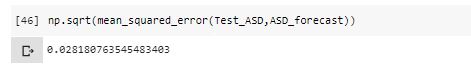
RMSE= √{[(y1-f1) ^2+(y2-f2) ^2+…. (yk-fk) ^2]/K}

Yk-Actual value

fK-Forecasted value

K-Number of observations

The value for RMSE was found as:



MAPE or Mean Absolute Percentage Error is percentage of the mean value calculated by taking the absolute of difference between actual values and the predicted values.

It is calcuted by formula:

A screenshot of a cell phone

Description automatically generated

And the MAPE value for our model was:

A screenshot of a cell phone

Description automatically generated

**Deployment:**

Deployment phase starts with the planning of deployment process. Whenever the project is deployed it should be provided with support maintenance and monitoring. Clients may not be able to use the model built thus, documentation of all the process should be created and provided with the deployment. A final report and presentation should be created which includes all the insights gained from the model, how the insights are relevant to the business requirements and how much model is successful. There are various ways to present the insights like traditional PowerPoint presentation. Nowadays due to increase in the use of Business Intelligence many people prefer the BI tools like Tableau, which provides interactivity to the presentation. This entire project was built on some sample data collected from Kaggle. As we have observed that there is very less error in our time series forecasting, we can deploy this model. For deployment we can take data from data source where every day current currency rate would be collected. Then this will be passed through the model and for some time frame, the prediction can be made. The predicted data and the original data can be visualized using tableau dashboards, this dashboard can be designed to view data of various years and compare if the predicted rates have some similarity with past year data.