Lloyds Banking Group

Data Analyst Incubation

Sprint 3 : Transaction Time Series Analysis & Forecasting

Contents

Background

It is 2017 – business analysts have noticed an uptick in the total transactions made by our clients over the last 2 years. Last January, client transactions increased beyond expectation and Quay Bank had concerns about reliably maintaining a robust liquidity ratio.

The Markets and Finance team are investigating how a time-series analysis and model that can reliably forecasts transaction totals could be utilised to help maintain effective cashflow and liquidity management for daily operations at Quay Bank. This is the second aspect of transactional data analysis they want your help with after looking at fraud flagging. The data is the same so you should be able to use what you have already learnt to dig deeper into this requirement.

This document has been put together by a senior data analyst in your team, who has included a lot of guidance on how to meet the business needs through your analysis.

Business Requirements

You have been tasked with delivering a Summary Report and forecasting model on the Transactions table in the Quay Bank database. The report can take the form of a Word document with screenshots of code & plots where appropriate. Alongside the report, you should reference the python notebook which includes the forecast model you have built.

The summary report should answer the following questions, which are only possible after exploration and analysis of the time-series:

- What is driving the increase in summed transaction amounts?

- What patterns and insights are evident across?
  - Weeks – Select either - Monday to Sunday OR Business days
  - Months – evident across the month.
  - Years – Seasonal patterns
- What periods during the year do we see peaks in transaction totals? Do these peaks trend similarly to the rest of the time-series?
- Based on the patterns evident in the data, can we develop an optimal forecasting model for one of the following:
  - Daily values projected 4 weeks ahead?
  - Weekly values projected 3 months ahead?
  - Monthly values projected 2 years ahead?

Deliverables:

1. Time-Series Analysis Notebook showing forecast model
2. Any relevant analysis **reports** produced in tools like PowerBI, Tableau
3. Summary report, which must include:

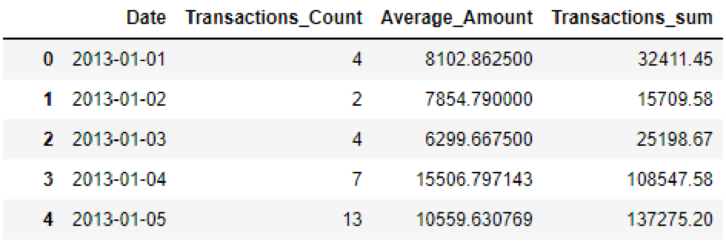
- Analytical Insights structured by Task.
- All models and plots included in analysis.
- Future Recommendations
- Data Sourcing documentation.
- Challenges experienced and opportunities for future investigations.

Assignment Detail:

Question 1 – What's driving the steady increase in summed transactions?

To complete this task the steps outlined below are expected :

- Query the Quay Database using SQL, extract the transactions data and load it into a Python IDE for analysis.
- Clean and aggregate your data to produce a daily count of transactions, the average of each transaction amount and the daily sum of Transactions. Expected output -



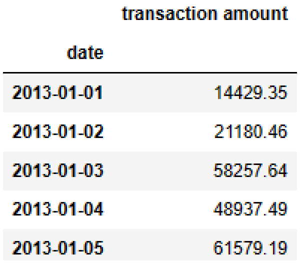
	Date	Transactions_Count	Average_Amount	Transactions_sum
0	2013-01-01	4	8102.862500	32411.45
1	2013-01-02	2	7854.790000	15709.58
2	2013-01-03	4	6299.667500	25198.67
3	2013-01-04	7	15506.797143	108547.58
4	2013-01-05	13	10559.630769	137275.20

- Analyse the average and count to determine which has contributed the most towards the steady increase in the summed transaction amounts. Discuss and document this.

Question 2 – What patterns and trends are evident across:

- Weeks - Monday to Sunday OR Business Day patterns?
- Months - Pattern evident across a single month? Are transaction fluctuations in Feb the same as July?
- Years - Seasonal pattern evident in each year?

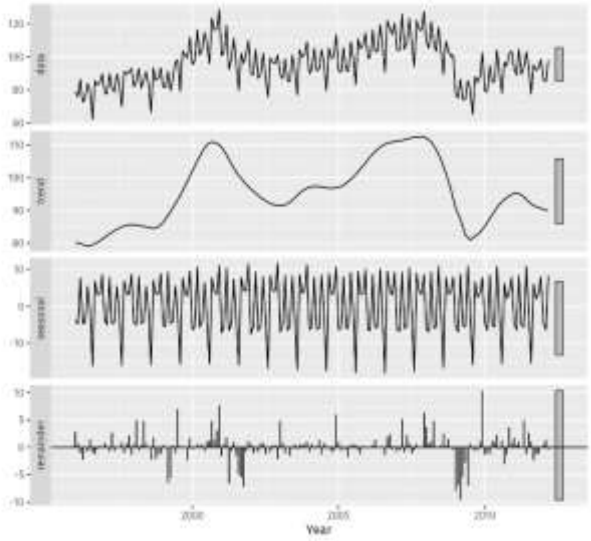
HINT: Once your data is aggregated by day (see right), the next step will be to aggregate it by week and by month. Consult Pandas documentation and your previous workbooks to support datetime operations as needed.



	transaction amount
date	
2013-01-01	14429.35
2013-01-02	21180.46
2013-01-03	58257.64
2013-01-04	48937.49
2013-01-05	61579.19

HINT: Typically for this analysis it is necessary to aggregate from transaction level data to a chosen datetime level of resolution (daily, weekly, monthly), before decomposing the time-series into its component parts: trend, seasonality and random(error).

Reminder : Before leaving this question behind you, consider documenting the significance of the observed trend(s) and any visible seasonality (look at their impact on the y-axis) across the week, month and year, on the business context for the Bank.



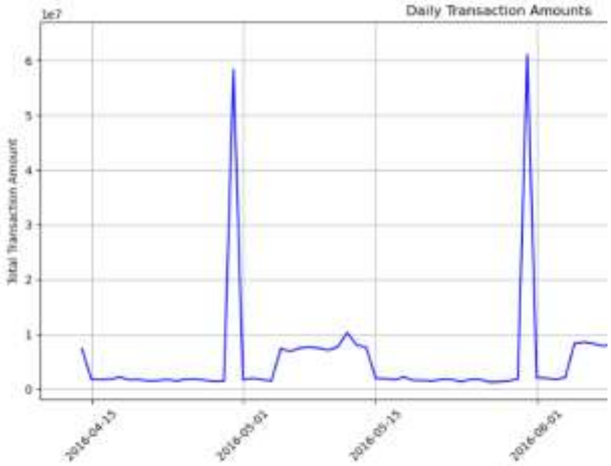
Following on from Question 2 – Data Quality Analysis and Validation

The exploratory data analysis you have completed so far, where you have organised and visualised the data, is an opportunity to flag errors, outliers and values that do not meet your data quality standards.

- Outline in your report, which quality standards you are measuring the data against and how does the data perform
- Identify and take note of observed errors, outliers and missing values. What steps will you take to improve the quality of your time-series?

Question 3 – What periods during the year does the bank see peaks in transactions (which would demand higher than normal liquidity/ reserves)? Do these peaks trend similarly to the rest of the time-series?

To answer this question, you are expected to identify the peak days or even weeks with the highest transaction amounts, relative to rest of the year. You can make this identification using statistical methods or keep it as a subjective interpretation of the data when it is plotted.



HINT: Plot partitions of data with a slicer, which allows you to zoom in on a particular week. You can also try plotting partitions of a stationary time-series, to view peaks without the linear trend.

Before leaving this question behind you, ask yourself whether you can observe these peaks increasing steadily in line with the linear trend? This is because when planning a time series forecast model you will typically need to decide whether you want the model to be fit to peaks and troughs or if instead it should be forced to smooth them over.

Question 4 - Based on the patterns evident in the data, can we develop an optimal forecasting model for one of the following:

- o Daily values projected 4 weeks ahead?
- o Weekly values projected 3 months ahead?
- o Monthly values projected 2 years ahead?

To answer this question, the tools available to you are ARIMA models in the stats library and the prophet procedure/ library, both available in python. You can also use the functionality of reporting tools like Tableau or Power BI. It is necessary to conduct a time-series analysis in Python to decompose the time series and assess data quality before attempting the model because you will design your model on the insights you found during the exploratory data analysis phase.

Hint for using ARIMA –

- It is crucial that you plot and interpret ACF and PACF plots. *First, be sure that you understand lags!*
- It is recommended to build several models for comparison.

Hint for using Prophet –

- It would be good to see the inclusion of holiday modelling, even of it is merely an assessment of its inclusion.
- Tune the flexibility parameters for trend or seasonality, before analysing the different models produced.

Remember! Alongside any model, evaluation and summary, you should also provide your interpretation the model's predictions of the unseen(test) data AND its metrics (e.g. p-value, AIC, BIC, RMSE). This will in effect answer the question – is it possible to develop a reliable model.