University of Cape Town

STA5091Z

Data-analysis for High-Frequency Trading

Assignment 1

Data Workflow and Initial EDA

Authors:
Matthew Dicks (DCKMAT004).

August 29, 2021

Contents

1 Data Workflow

This section will discuss how the data was prepared before the Exploratory Data Analysis (EDA) was performed. It will detail how the dataset was downloaded, how the cleaning and compacting was performed, how the full dataframe was created as well as how the trades were classified.

1.1 Downloading the Data

The first part of this project was to download 6 months of top-of-book (TAQ) data for 10 stocks on the Johannesburg Stock Exchange (JSE) from Bloomberg. Given that we are currently off campus and do not have access to the library's Bloomberg Terminal service I could not use Bloombergs's Desktop API to fetch the data. After doing some research I was not able to find an alternative method to get the data. Therefore, I used the test data from Jericevich *et al.* (2020) to do my analysis. This dataset consists of a weeks worth of TAQ data from 2019-07-08 to 2019-07-12 for 10 JSE stocks. The dataset can be found here, and the 10 equities used in the analysis are listed in Appendix A.

1.2 Cleaning and Compacting the Data

For this analysis I used the same cleaning methods that were used by Jericevich *et al.* (2020) to clean their data. Since we only care about the continuous trading data, for this assignment, the first issue was to remove all the trading activity from the opening and closing auctions. Therefore, only activity that occurred between 9:00 and 16:50 was considered. All other data was removed. It is also vital to remove any events related to intra-day volatility auctions and the impact of various futures close-outs. Given that the futures close-outs occur only in March, June, September and December we do not need to consider this case. To remove most of the unwanted trades such as after-hour trades (LT), correction of previous days trades (LC) and auction uncrossing price trades (IP) I only considered automated trades (AT).

Now as discussed in the notes, often there are larger trades that get executed against a set of smaller trades. This means that a single trade, at a given timestamp, gets split up into smaller trades that change the best bid and best ask as it moves up and down the order book. This event is represented in the data as multiple trades when it is actually a single trade. Also there are timestamps that have multiple quotes. To deal with these events I performed trade and quote compacting. This is the process of changing the data to better represent the fact that a single trade has occurred, or to ensure that there is only a single quote per timestamp. For quotes with the same timestamp the most recent quote in the sequence of quotes is kept and the rest are removed. For trades with the same timestamp, and the same order type, the volume of the trade is calculated to be the aggregated volume for that timestamp and the price of the trade is the volume weighted average price.

1.3 Computing Columns

This subsection will briefly detail how each of the additional quantities were calculated from the TAQ data after cleaning and compacting was performed. Before any of the quantities are defined lets use i to denote the i-th event and j to denote the j-th day. The first quantity that was computed was the mid-price. The mid price for the i-th event on the j-th day is defined as the average of the best bid and the best ask

$$m_{ij} = \frac{1}{2}\left(b_{ij} + a_{ij}\right) \tag{1}$$

where b_{ij} and a_{ij} are the best bid and best ask for the i-th event on the j-th day. Now using the mid-price we can compute the mid-price change. This is defined as the log difference between mid prices

$$\Delta m_{ij} = \log(m_{i+1,j}) - \log(m_{i,j}) \tag{2}$$

The third quantity that needed to be calculated was the micro-price. This is the volume weighted average of the best bid and best ask, and is defined as

$$s_{ij} = \frac{v^a_{ij}}{v^a_{ij} + v^b_{ij}} a_{ij} + \frac{v^b_{ij}}{v^a_{ij} + v^b_{ij}} b_{ij} \tag{3}$$

where v^b_{ij} and v^a_{ij} are the volumes for the best bid and best ask respectively. In assignment 2 we are going to be investigating price impact. To do so we will need to compute the normalized trade volume. The normalized trade volume is defined as

$$\omega_{ij} = \frac{v_{ij}}{\sum_{k=1}^{T_j} v_{kj}} \left[\frac{\sum_{j=1}^{N} T_j}{N} \right] \tag{4}$$

where there are T_j trading events on the j-th day, v_{ij} is the traded volume and there are N trading days. The normalized trade volume can also be used to compare different equities with different liquidity. The final quantity that needed to be computed was the inter-arrival times. These are defined as

$$\tau_{ij} = t_{i+1,j} - t_{i,j} \tag{5}$$

where $t_{i,j}$ is the arrival time of the i-th event on the j-th day. For this assignment I have computed the inter-arrival times in seconds.

1.4 Classifying Trades

To perform trade classification there are three prevailing methods: the quote-rule (Hasbrouck 1988), the tick rule (Blume *et al.* 1989), and the Lee/Ready rule (Lee & Ready 1991) (a combination of the quote and tick rule). As Jericevich *et al.* (2020) did, I will be using the Lee/Ready rule as the method by which I will be classifying trades. I will be using the method nicely outlined in Theissen (2001), which is as follows:

1. Transactions which occur at prices higher (lower) than the prevailing mid-price are classified as buyer-initiated (seller-initiated) trades.
2. Transactions that occur at a price that is equal to the mid-price but is higher (lower) than the previous transaction price are classified as buyer-initiated (seller-initiated) trades.
3. Transactions that occur at a price that is equal to the mid-price and equal to the previous transaction price but is higher (lower) than the last different transaction price are classified as buyer-initiated (seller-initiated) trades.

From these rules you can see that the first rule is just the quote rule and the second and third rules are the tick rule. You can also notice that the quote rule is taking precedence over the tick rule.

2 Initial EDA

The second part of the assignment involves doing EDA to sanity check the data. This section is broken up into three parts: Section 2.1 involves visualizing two hours of trading for two different stocks. Section 2.2 presents the order-flow auto-correlation found in these same two stocks over the week of trading. Section 2.3 visualizes the inter-arrival times for all trades and compares them to distributions. Anglo American (AGL) and Naspers (NPN) are the two stocks used in this analysis.

2.1 Visualizing Trading

In Figures 1 and 2, the blue bubbles represent the best bids, the red bubbles represent the best asks, and the yellow bubbles represent the trades. The size of the bubbles indicate the size of the volume of the trade. To scale the volumes so that they can be seen nicely on the plot, I used the mean of the means of the best bid volumes, best ask volumes and the traded volumes.

2.1.1 Naspers - NPN

In Figure 1, some of the trade sizes, compared to the bids and asks, were a little surprising. This is because they are far larger than I expected. In each of the time slots there are multiple large orders that have hit the order book and it is likely that they have consumed the entirety of the best bid or ask volumes. These trades are going to have a large impact on the mid-price as it will cause it to drastically move up and down.

For the plot on the left, the price is going down and we observe more liquidity on the ask side. This could be an indication that more people may be trying to get rid of the stock. The plot on the right shows more liquidity on the bid side and the stock is rising. A possible reason for this is that during this time period trader(s) are trying to buy the stock and therefore there are more bids to buy.

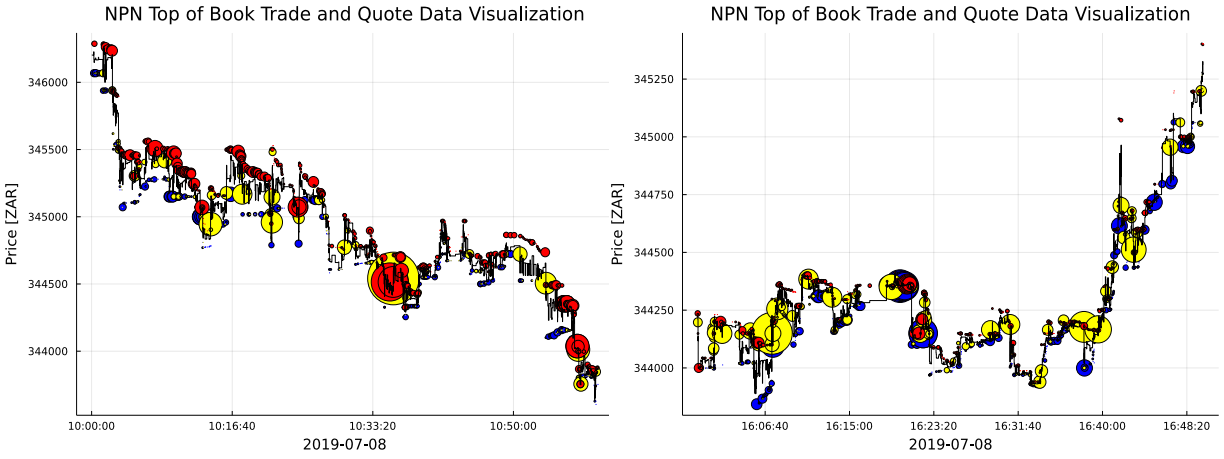

Figure 1: Visualizing the trading, for Naspers (NPN), from 10:00 to 11:00 as well as from 16:00 to 17:00 on 2019-07-08. The red bubbles represent the best asks, the blue bubbles are for the best bids, and the yellow bubbles represent the trades. The size of the bubbles are proportional to the volume.

2.1.2 Anglo American - AGL

Now considering Figure 2. The plot on the left shows a lot less variability in the volume when compared with NPN. Over this time period there seems to be a steady flow of medium sized quotes and trades. This changes when looking at the plot on the right. There are far more large trades that are hitting the order book. An interesting artifact to note is that the large trade volumes are very close to the peaks and troughs of the price. Now this could be a random artifact but it could also mean that some trader is being very accurate at targeting a mean reversion tactic.

Comparing AGL to NPN, we see that in AGL there seem to be periods where there is not much activity or there is activity but it is happening at very low volumes. For example, in the figure on the left we see that there is a period from approximately 10:27 to 10:40 where there are trades and quotes happening but they are occurring at relatively low volumes. In the plot on the right there is a similar period from 16:17 to 16:25.

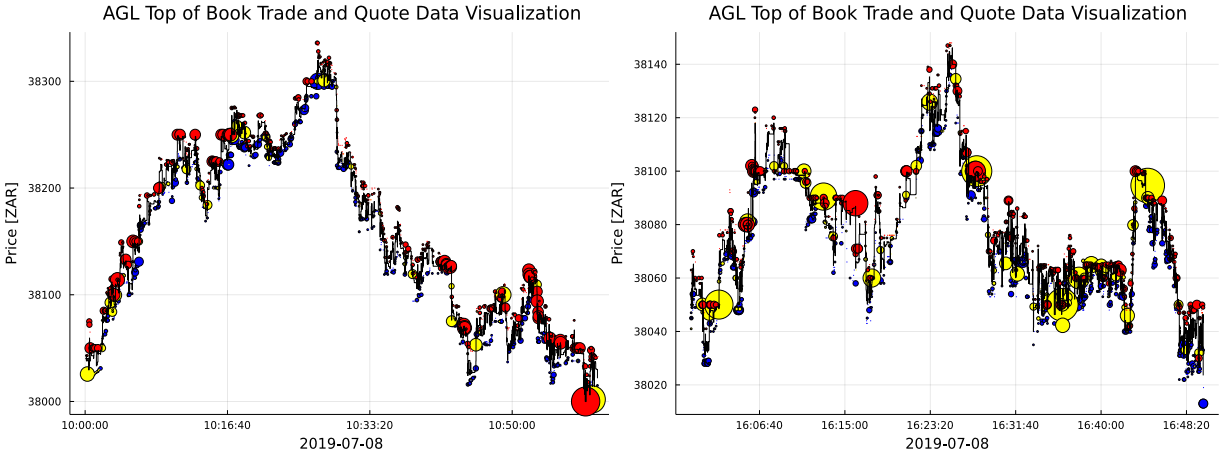

Figure 2: Visualizing the trading, for Anglo American (AGL), from 10:00 to 11:00 as well as from 16:00 to 17:00 on 2019-07-08. The red bubbles represent the best asks, the blue bubbles are for the best bids, and the yellow bubbles represent the trades. The size of the bubbles are proportional to the volume.

2.2 Order-flow Auto-correlation

One of the stylised facts in high-frequency trading that I try to recover is that there is auto-correlation in the order flow. Toth *et al.* (2015) showed that on the London Stock Exchange there is a clear presence of the correlation in the trade signs (buyer/seller) for time scales of up to thousands of lags. They showed that this behaviour is possibly due to two factors; herding and order splitting. In herding traders follow the actions of other traders. If a trader observes lots of buying (selling) then they too will start to buy (sell). Order splitting is the process of splitting a parent order into smaller child orders to try and limit price impact. Toth *et al.* (2015) found that the auto-correlation in the order-flow was largely due to the the latter case of order splitting.

2.2.1 Naspers - NPN

Figure 3 visualizes the order-flow auto-correlation for NPN. The plot on the left shows the x-axis on regular scale and the plot on the right shows the results with the x-axis on $\log_{10}$ scale. In the plot on the

left the red bars represent the 95% confidence interval for the significance of the auto-correlation. These values were computed based on the asymptotic assumption that

$$\rho_j \overset{a}{\sim} \mathrm{N}(0, \frac{1}{N}) \tag{6}$$

where ρ_j is the auto-correlation for the j-th lag and N is the number of observations.

The plot on the left shows that the order-flow is significantly correlated up to a lag of approximately 200. The significant correlations, for the most part, are positive indicating that the trades have the same sign. This corroborates the findings of Toth *et al.* (2015) who stated that the persistence in the order-flow is due to buy orders being followed by more buy orders or sell orders being followed by more sell orders. I have plotted the auto-correlation with lags in log scale to show the exponential decrease in the auto-correlation. The stylised fact of order-flow auto-correlation has been shown to occur for NPN over the week of trading.

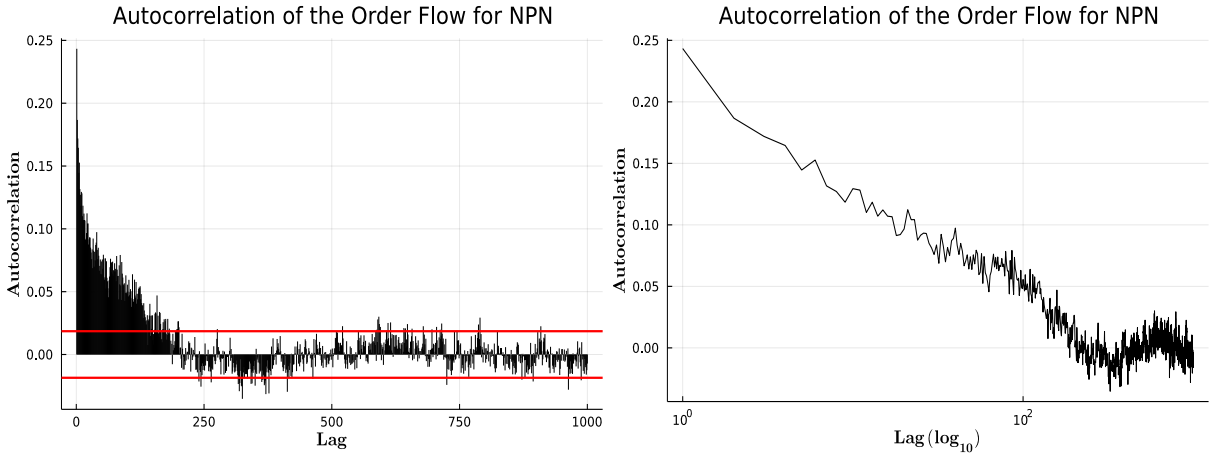

Figure 3: This figure visualizes the order-flow auto-correlation aggregated over the week from 2019-07-08 to 2019-07-12 for the equity NPN. The plot on the left shows the auto-correlation on regular scale and the plot on the right shows the auto-correlation on log scale. The red bars represent the confidence interval for the auto-correlation.

2.2.2 Anglo American - AGL

Now considering the order-flow auto-correlation in AGL. This plot looks slightly different when compared with NPN. We are getting a much lower value for the auto-correlation, and thus a lower significance, but we see that the significance of the auto-correlation persists for a longer period of time. The significant auto-correlations, in NPN, fell away after approximately 200 lags. However, in AGL we observe that significant auto-correlations are present up to 500 lags. The exponential decrease is much slower than NPN as evidenced by the plot on the right. This is likely due to the lower initial values for the auto-correlation. Even though the auto-correlations are much smaller than what we observed for NPN, we do recover the stylised fact because there is definitely a clear presence of a significant auto-correlation for the first 150 lags.

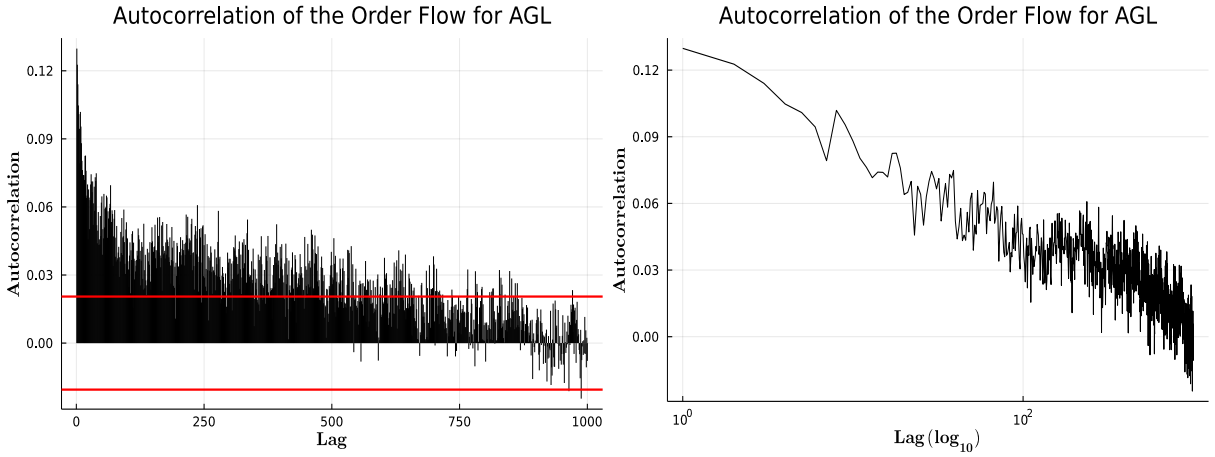

Figure 4: This figure visualizes the order-flow auto-correlation aggregated over the week from 2019-07-08 to 2019-07-12 for the equity AGL. The plot on the left shows the auto-correlation on regular scale and the plot on the right shows the auto-correlation on log scale. The red bars represent the confidence interval for the auto-correlation.

2.3 Inter-arrival Times

2.3.1 Naspers - NPN

Figure 5 shows the distribution of inter-arrival times between trades for NPN. The plot on the right shows the inter-arrival time frequencies on $\log_{10}$ scale. The histogram on the left shows that the distribution of inter-arrivals is likely going to be heavier tailed than an exponential distribution. However, the log scaled plot shows that the inter-arrivals should not follow a power law distribution.

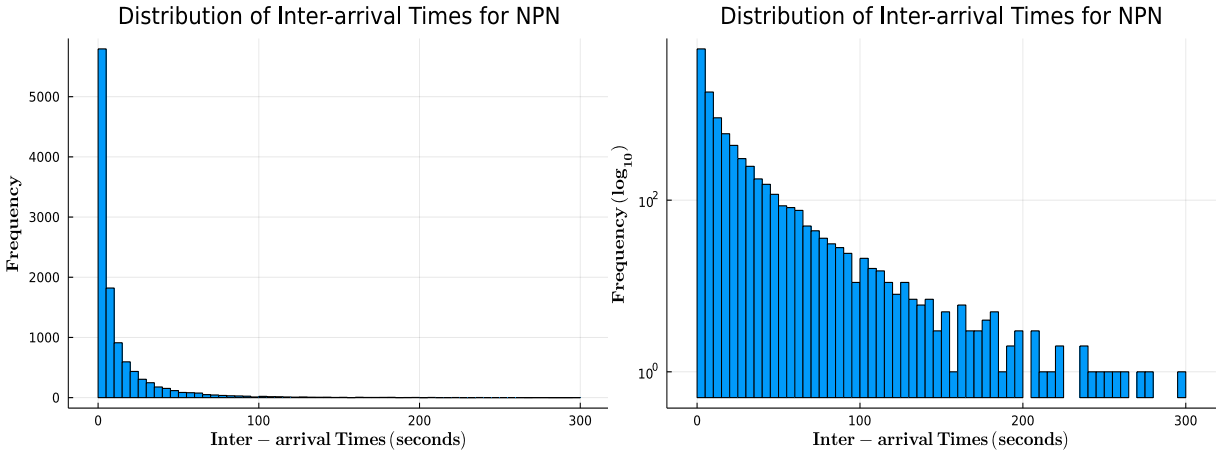

Figure 5: This figure shows the distributions of the inter-arrival times for NPN. The plot on the left shows the distribution in normal scale and the plot on the right shows the distribution in $\log_{10}$ scale.

To find out if the statements above were true I created QQ-plots to compare the inter-arrival times to

the exponential distribution and a power law distribution. In the assignment brief we were told to use a power law distribution with a CDF of the following form

$$\mathrm{P}(\tau_{ij} < \tau) = 1 - \left(\frac{k}{\tau}\right)^{\alpha} \tag{7}$$

The power law distribution I have decided to use is the Pareto distribution. Considering the QQ-plots below we can see that the inter-arrival times are fatter tailed than the exponential distribution but far lighter tailed than the Pareto distribution. The distribution of the inter-arrival times seems to fall in-between the Pareto and exponential distributions.

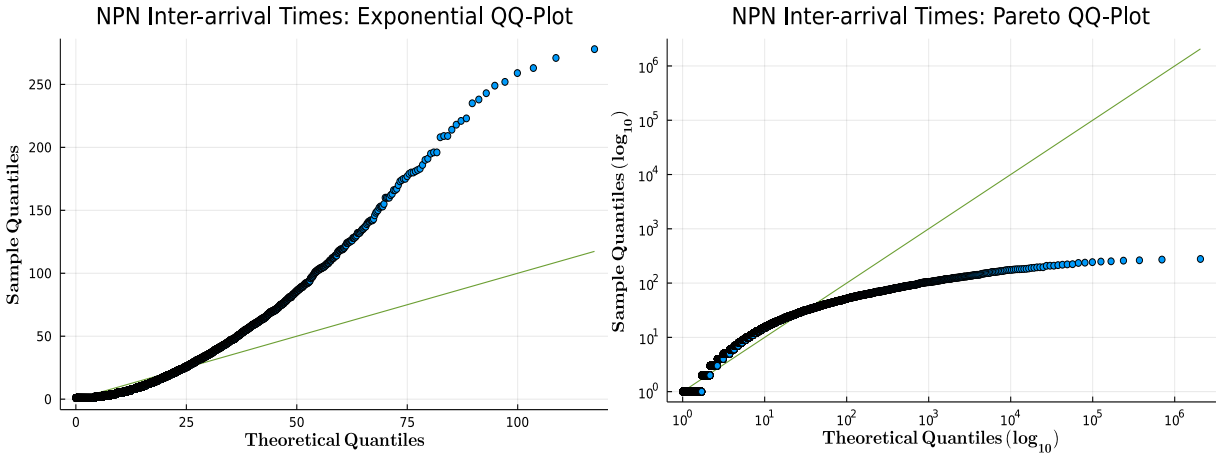

Figure 6: This figure uses QQ-plots to compare the inter-arrival times, for NPN, to the exponential and Pareto distributions. The scale for the Pareto distribution is $\log_{10}$.

Figure 7 shows the auto-correlations in the inter-arrival times across the five trading days. The correlations in the inter-arrival times clearly form a wave and there is a clear period in the wave of auto-correlations. This wave pattern means that there should be a wave pattern in the inter-arrival times. After plotting the inter-arrival times (not shown here) the reason for this pattern became clear. Initially trades are happening really quickly and there are low inter-arrival times. The inter-arrival times start to increase until approximately midday, where they reach their peak. After the peak the inter-arrival times start to decrease indicating more frequent trading activity. This wave pattern occurs with a period of approximately 2500 trading events. The average number of trades per day that occurred in this week of trading, for NPN, was 2237. Therefore, this wave of inter-arrival times seems to be a daily pattern, which becomes clear in the correlation plot.

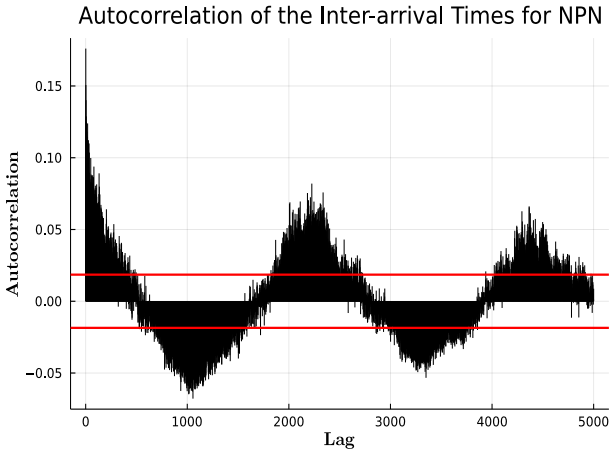

Figure 7: This figure gives the auto-correlation between the inter-arrival times for NPN. The red bars represent the upper and lower bounds on the confidence interval for the auto-correlations.

2.3.2 Anglo American - AGL

What was observed for NPN is also observed for AGL. Except looking at the log scaled plot, in Figure 8, we have that there is a steeper decrease in the frequency of inter-arrival times for higher seconds. This indicates that we are likely to see a slightly lighter tailed distribution when compared with NPN.

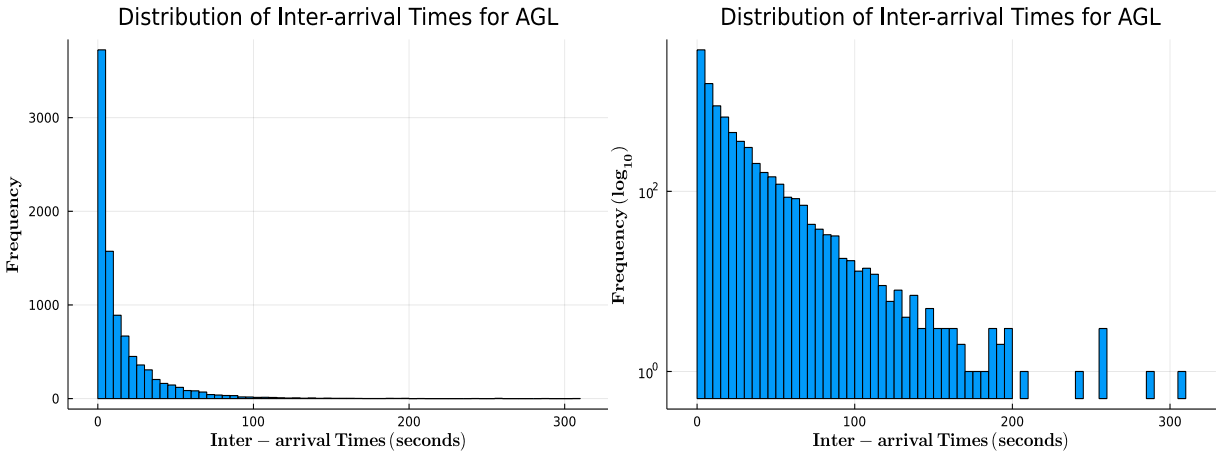

Figure 8: This figure shows the distributions of the inter-arrival times for AGL. The plot on the left shows the distribution in normal scale and the plot on the right shows the distribution in $\log_{10}$ scale..

What was seen in the histograms can also be seen in the QQ-plots in Figure 9. Even though it is clear that the inter-arrival times are heavier tailed when compared to the exponential distribution. We see that the distribution looks far closer to exponential, when compared with NPN. You can also see that the QQ-plot for the Pareto distribution shows that the Pareto distribution is far heavier tailed when compared to the distribution of inter-arrival times.

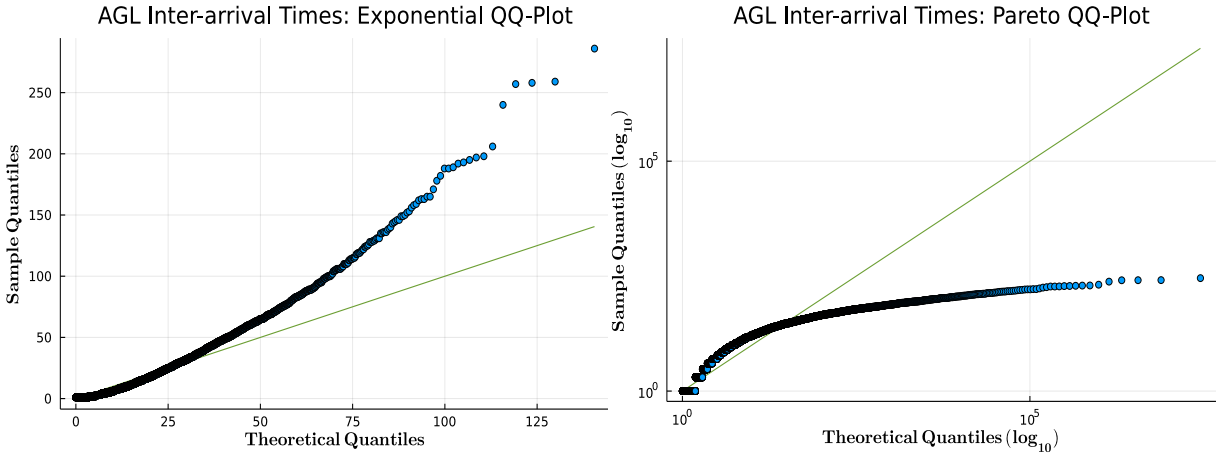

Figure 9: This figure uses QQ-plots to compare the inter-arrival times, for AGL, to the exponential and Pareto distributions. The scale for the Pareto distribution is $\log_{10}$..

In Figure 10 We see similar behavior, in the auto-correlation of the inter-arrival times, when compared to NPN. There also seems to be a wave pattern that is occurring in the inter-arrivals for AGL. However, this pattern seems to be far less pronounced when compared to what we saw for NPN. Once again I plotted the inter-arrivals (not shown here) and saw that there was a wave pattern in the inter-arrivals. The pattern is not as clear as it was for NPN. That is why we see the smaller correlations and the dissipating of the correlations.

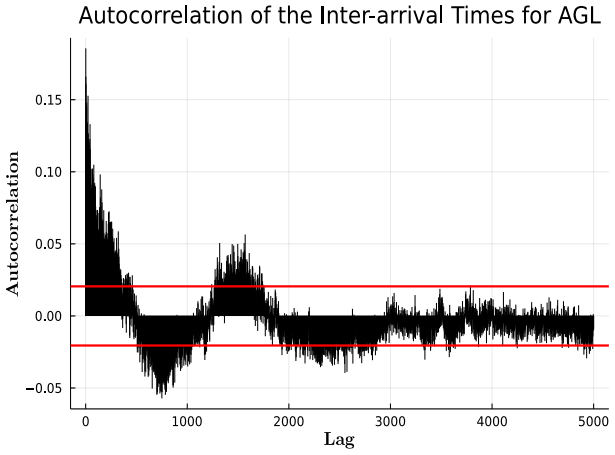

Figure 10: This figure gives the auto-correlation between the inter-arrival times for AGL. The red bars represent the upper and lower bounds on the confidence interval for the auto-correlations..

3 Code

The GitHub repository for this assignment can be found here.

References

1. Blume, M. E., MacKinlay, A. C. & Terker, B. Order imbalances and stock price movements on October 19 and 20, 1987. *The Journal of Finance* **44,** 827–848 (1989).
2. Hasbrouck, J. Trades, quotes, inventories, and information. *Journal of financial economics* **22,** 229–252 (1988).
3. Jericevich, I., Chang, P. & Gebbie, T. Comparing the market microstructure between two South African exchanges. *arXiv preprint arXiv:2011.04367* (2020).
4. Lee, C. M. & Ready, M. J. Inferring trade direction from intraday data. *The Journal of Finance* **46,** 733–746 (1991).
5. Theissen, E. A test of the accuracy of the Lee/Ready trade classification algorithm. *Journal of International Financial Markets, Institutions and Money* **11,** 147–165 (2001).
6. Toth, B., Palit, I., Lillo, F. & Farmer, J. D. Why is equity order flow so persistent? *Journal of Economic Dynamics and Control* **51,** 218–239 (2015).

4 Appendix A - JSE Equities

Security name	Security code
Absa Group Ltd	ABG
Anglo American Plc	AGL
British American Tobacco Plc	BTI
FirstRand Ltd	FSR
Nedbank Group Ltd	NED
Naspers Ltd	NPN
Standard Bank Group Ltd	SBK
Shoprite Holdings Ltd	SHP
Sanlam Ltd	SLM
Sasol Ltd	SOL

Table 1: JSE equities used in this analysis.