University of Cape Town

STA5091Z

Data-analysis for High-Frequency Trading

Assignment 2

Price Impact and Seasonality

Authors:
Matthew Dicks (DCKMAT004).

September 19, 2021

1 Data Workflow

The data used in this assignment is the same as that used for assignment 1. This data was collected by Jericevich *et al.* (2020) and can be found here. The cleaning of the data followed the same process as what was outlined in assignment 1. However, I did make a change to the compacting of the data. Instead of only keeping the last bid, last ask and compacted trades, for a given timestamp. If there was a bid or ask that came before the trade in the same timestamp I have included those quotes as well. I have done this to try and increase the accuracy of my price impact curves. Adding the quotes that came before the trade will allow me to accurately account for the which quotes the trades were executed against.

2 Bar Data and Candle-stick Plots

Figure 1 plots the Open-High-Low- Close-Volume (OHLCV) bar data for AGL in 1 minute and 10 minute bars. These candle-stick plots are created using the following rules: the top of top of candle-stick is the maximum traded price in that bar, the bottom is minimum, the top of the bar is the maximum of the opening and closing price, and the bottom is the minimum. The bar is blue if the opening price is less than the closing price and red if the closing price is less than the opening price. For the transaction data the black line represents the Volume Weighted Average Price (VWAP) price for that bar. Figure 2 creates the OHLCV plots but instead of using the transaction prices I have used the mid-price to create the bars. For this figure the black line is the closing micro-price for that bar.

From these plots you can see that the mid-prices and the transaction prices follow similar patterns. Therefore, the mid-price and the transaction price contain similar information for AGL. This means that there are minimal transactions occurring at prices far away from the mid-price.

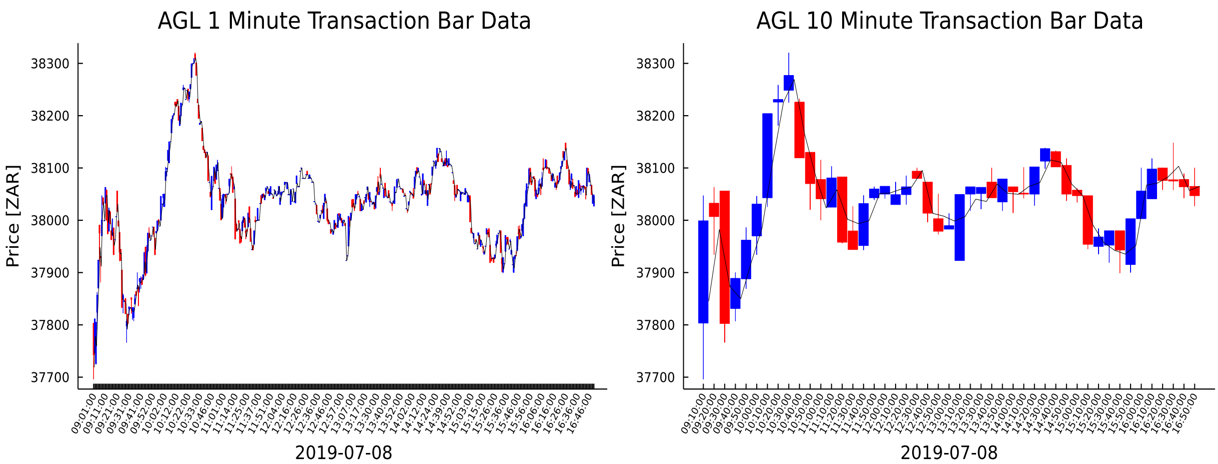

Figure 1: This figure plots the OHLCV bar data for AGL's transactions in 1 minute and 10 minute bars. The black line in both of these plots is the VWAP price obtained over the bar.

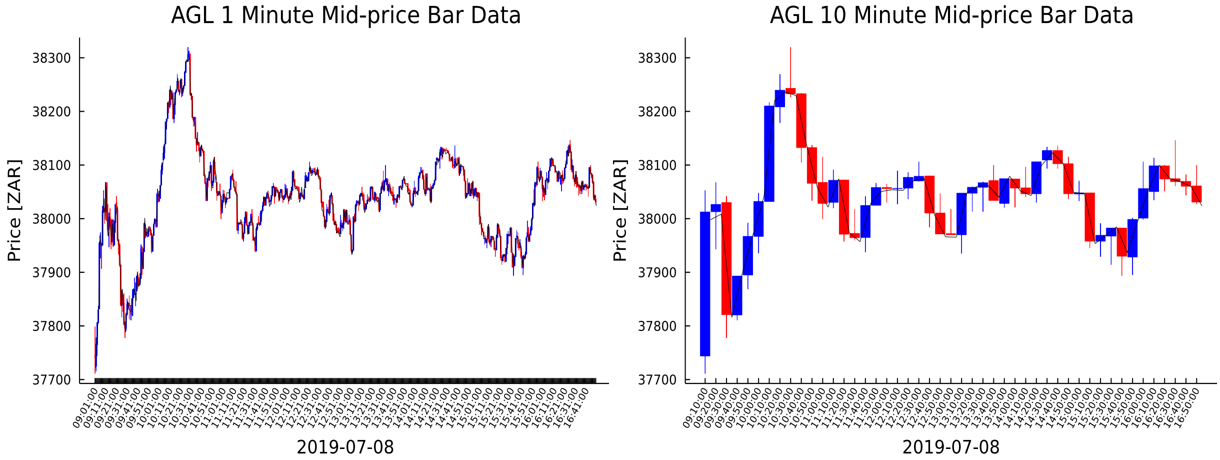

Figure 2: This figure plots the OHLCV bar data for AGL's mid-prices in 1 minute and 10 minute bars. The black line in both of these plots is the closing micro-price for the bar.

3 Time-series Stylized Facts

In Figure 3 I plot the distributions for the VWAP and micro-prices for AGL over the week of trading. For the micro-price distribution I have used the closing price from the 1 minute bar. From the candle-stick plots, shown in Figures 1 and 2, it is not surprising to see that the distributions are very similar. Both distributions are skewed and most of the mass is located near the higher prices. The distribution indicates that the price mainly stayed between 37700 and 38100 with a few increases above 38100 but more frequent and larger dips below 37700.

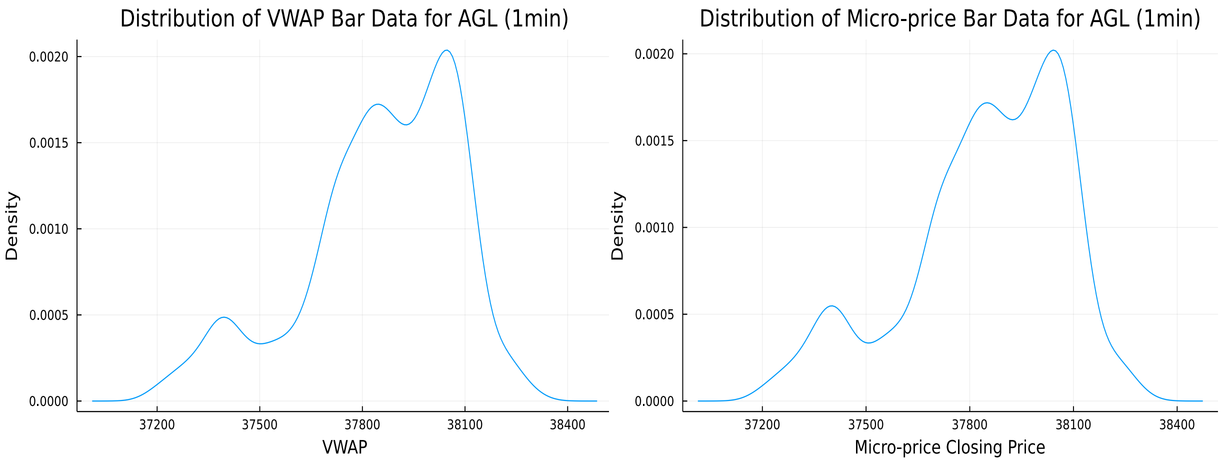

Figure 3: The plot on the left shows the distribution of the 1 minute VWAPs across all trading days in the dataset. The plot on the right shows the distribution of the 1 minute closing micro-prices across all days in the dataset.

The QQ-plots given in Figure 4 also show the similarity between the distributions of the VWAPs and closing micro-prices. The QQ-plots compare the price distributions to the normal distribution with parameters fit to the data using using the Maximum Likelihood method. The distributions are approximately normal but with lighter tails on the right. This pattern was evident from the full distribution plots provided in Figure 3.

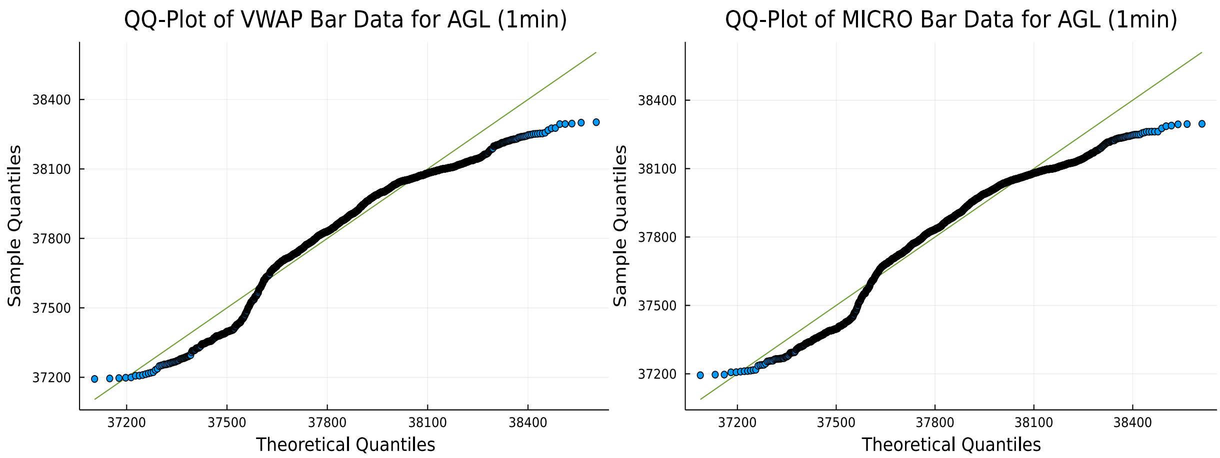

Figure 4: This Figure compares the distributions of the VWAPs and closing micro-prices to normal distributions. The parameters of the normal distributions are estimated from the data.

The plots given in Figure 5 show the left and right tails of the price distribution for the VWAP bar data. The tails are constructed using data from above the 95-th percentile. There is not much information in these plots but we can see that they make sense. In the left tail plot we have that most of the density is located towards the right and the right tail shows us that most of the density is located towards the left. This is consistent to what we have seen in the full distribution.

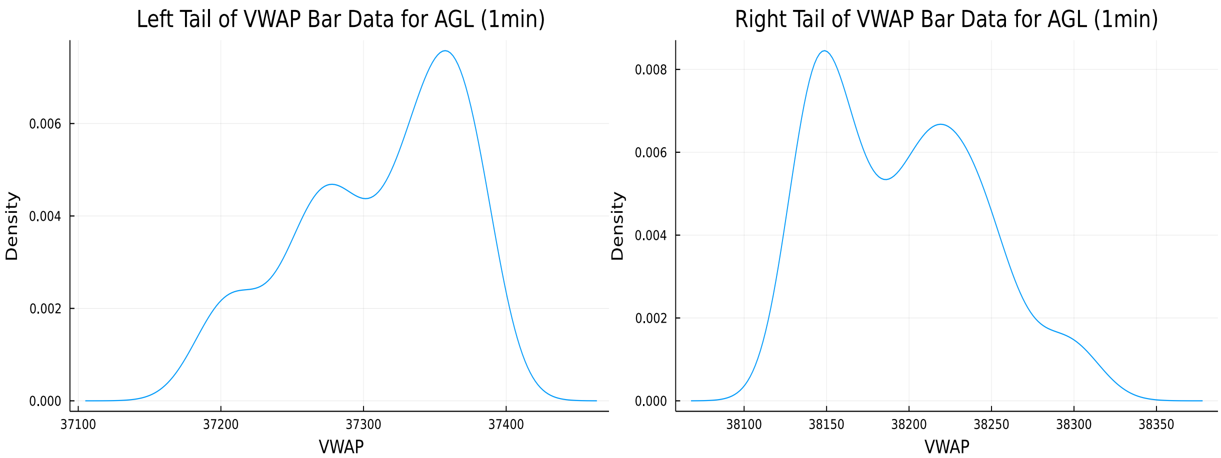

Figure 5: Plots the left and right tails of the VWAP distribution. The tails are constructed using data from above the 95-th percentile.

Figure 6 depicts the left and right tail distributions for the closing micro-prices computed using 1 minute bars. The tails are computed by using data above the 95-th percentile. The left and right tail distributions show the same features found in the VWAP distributions given in Figure 5.

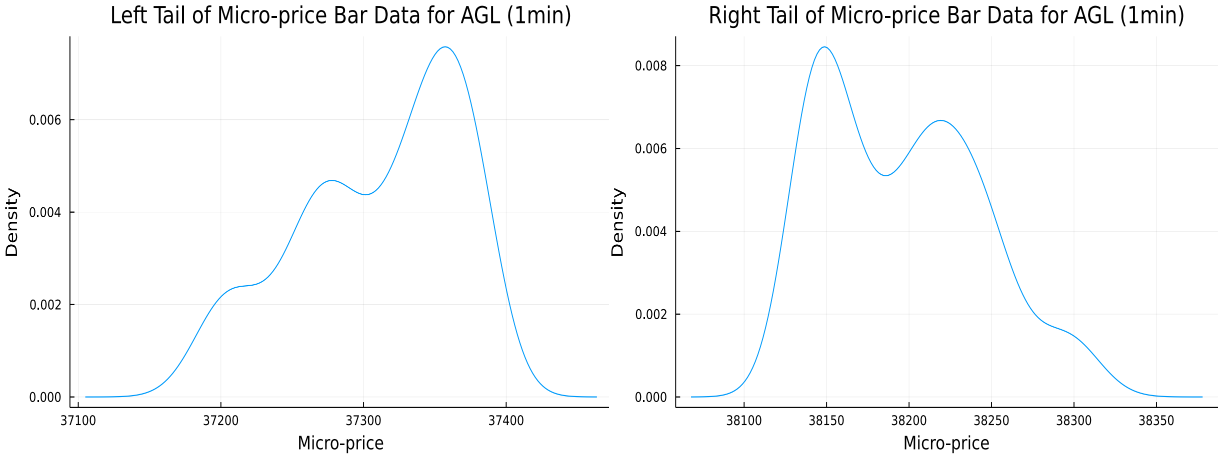

Figure 6: Plots the left and right tails of the closing micro-price distribution computed using 1 minute bars. The tails are constructed using data from above the 95-th percentile.

Figure 7 plots the auto-correlation of the VWAP and micro-price bar data, computed on 1 minute bars. The red lines, in this figure, give the upper and lower bounds on the 95% confidence interval. These values were computed based on the asymptotic assumption that

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

where ρ_j is the auto-correlation for the j-th lag and N is the number of observations. For both prices we see that for low lags there is a strong auto-correlation between the prices. This auto-correlation fades away but then changes sign and becomes negative. This indicates a period in the prices. The auto-correlations can also be seen to be significant for most of the lags.

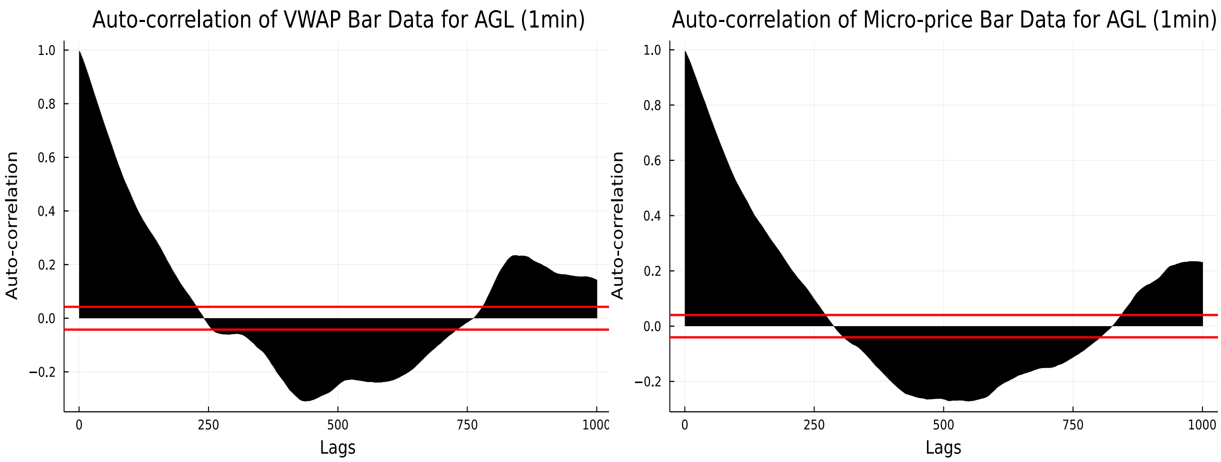

Figure 7: Plots the auto-correlations for the VWAP and closing micro-prices. The auto-correlations are computed using data from all the trading days.

4 Price Impact

To compute the price impact of a trade, I used the method outlined by Lillo *et al.* (2003), Harvey *et al.* (2017), and Jericevich *et al.* (2020). If we let m_t be the mid price at time t we have the price impact of a trade at time t is given by the following formula:

$$\Delta p_{t_k} = \log(m_{t_{k+1}}) - \log(m_{t_k}) \tag{2}$$

where $m_{t_{k+1}}$ is the mid-price after the trade and m_{t_k} is the mid price before the trade. To create the price impact curves I created 20 logarithmically spaced normalized volume bins on $[10^{-3}, 10^{0.5}]$. Each trade was placed into one of these bins depending on its normalized volume. Then for each bin I computed the average price impact, Δp^* and the average normalized volume, ω^*, that occurred in that bin over the course of the trading period. This method was applied to all 10 stocks. The price impact curves were then created by plotting the average price impact, Δp^* as a function of the average normalized volume, ω^*, for all the stocks. The price impact curves were plotted on a log-log scale. For each stock, different price impact curves were plotted, one for buyer-initiated trades and one for seller-initiated trades. The trade classifications were done using the Lee/Ready rule (Lee & Ready 1991).

The price impact curves for the buyer-initiated and seller-initiated trades can be seen in Figure 8. The price impact curves follow a similar pattern to those observed by Jericevich *et al.* (2020). As they noted, for smaller volumes there is a deviation from the expected linear relationship between the average normalized volume and the average price impact on the log-log scale. However, for volumes larger than 10^{-1} the linear relationship is clearly evident. For the larger volumes the relationship looks to follow the relationship given by:

$$\Delta p^* = \frac{\text{sign}(\omega^*)|\omega^*|^\alpha}{\lambda} \tag{3}$$

where λ is a liquidity parameter. This effect is consistent with the results of Jericevich *et al.* (2020) and Harvey *et al.* (2017), who observed the similar effects on the JSE. From (3) we see an inverse relationship between liquidity and price impact. The effect of liquidity on the price impact is also evident in these plots. If we look at two stocks ABG and NPN we can see that the price impact for ABG is higher than that of NPN. If we use average daily value traded as a proxy for liquidity it can be seen that NPN has an average daily value traded that is approximately 5 times larger than ABG. This shows that a stock with higher liquidity will have, on average, a lower price impact for the same normalized volume. Another feature to take note of is that the price impact curves for the buyer-initiated trades and seller-initiated trades are not very different from each other. This means that the price impact from selling and the price impact from buying are approximately the same for each stock.

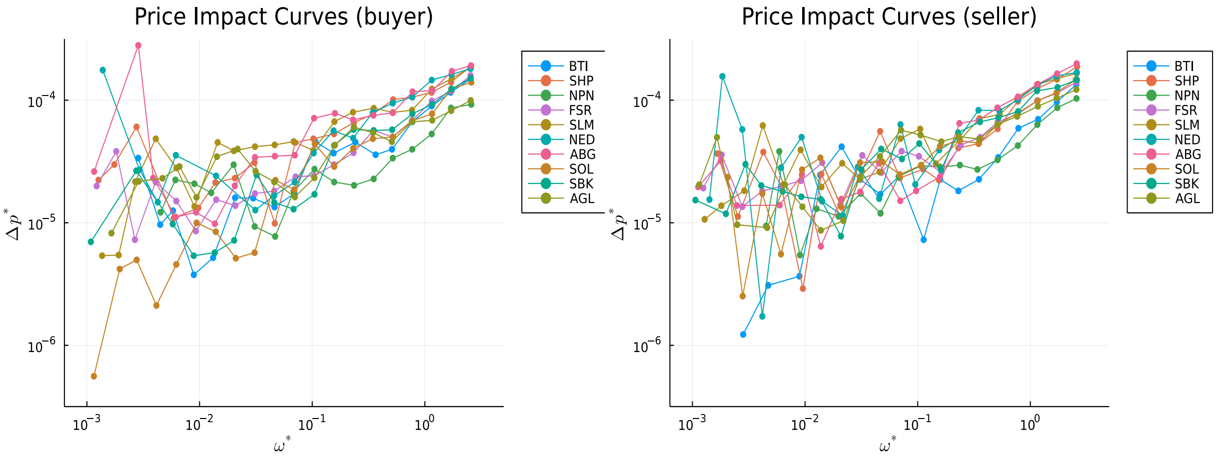

Figure 8: Plots the price impact curves for all the stocks. The plot on the left gives the price impact curves for buyer-initiated trades and the plot on the rights shows the price impact curves for seller initiated trades.

5 Order-book Seasonality

To create the intraday volume curves I split up each trading day into 10 minute intervals. For each of the intervals I computed the total volume traded in that interval normalized by the total volume traded in that day. This was done for each day. The average aggregate normalized volume for a given stock was then computed as the average, across the days, of the aggregate normalized volumes. The volume curves were constructed using all the trading days in the dataset.

Empirical studies have found that the intraday volume patterns should follow a U-shape (Cartea *et al.* 2015; Jericevich *et al.* 2020). With large volumes occurring at the beginning and the end of the day and smaller volumes occurring during the middle of the day. The volume curves for each stock, and the average volume curve for all stocks, can be seen in Figure 9. It is clear from this plot that I have managed to recover the U-shape volume curve found previously by Jericevich *et al.* (2020), for the JSE. Initially there is a large amount of volume being traded, then there is a gradual slow down in the amount of volume traded. After approximately 13:50 there is an increase in the volume traded. This increase accelerates until the volume traded reaches a peak at the end of the day. This pattern is evident for all stocks as they all follow the U-shaped volume curve.

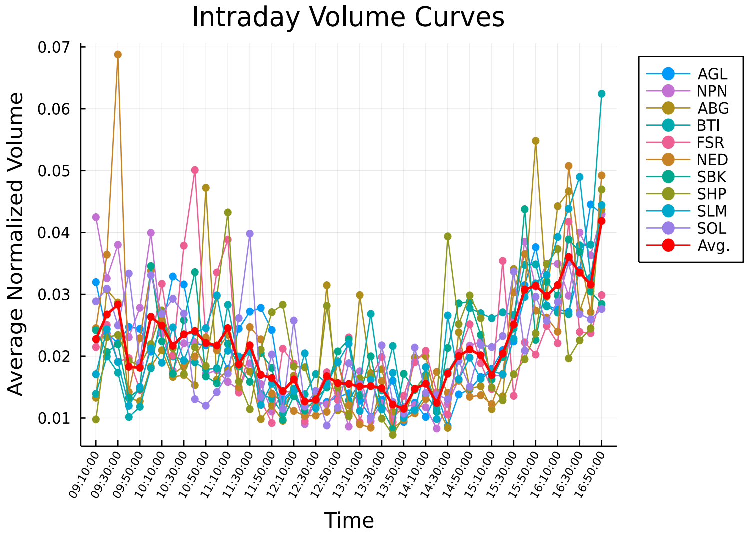

Figure 9: Plots the volume curves for each of the stocks computed using Top-of-Book Trade-and-Quote data from 2019-07-08 to 2019-07-12. The red line represents the average volume curve averaged over all stocks.

As a proxy for volatility I will use the average absolute intraday returns. This measure for volatility was proposed by Cartea *et al.* (2015). If we let p_t be the trade price that occurred at time t. The following formula shows how I computed the returns for a trade at time t:

$$r_t = \log(p_t) - \log(p_{t-1}) \tag{4}$$

This method was used by Jericevich *et al.* (2020). To compute the average absolute intraday returns I once again split up the trading day into intervals of 10 minutes. For each day and each stock I computed the average absolute returns in that bin, normalized by the average absolute returns computed over the entire day. To create the intraday return curves for a single stock I averaged the returns in each bin across the trading days. The average intraday return curve was created by averaging the return curves for each stock.

Figure 10 gives the intraday return curves for each stock, and the average intraday return curve. For all the stocks, there is initially a large amount of volatility, but this volatility decreases rapidly until it hits a floor and then it stays relatively constant for the rest of the day. This left slanted volatility smile is consistent with what was observed by Cartea *et al.* (2015) using Apple's stock and Jericevich *et al.* (2020) using the same 10 stocks used in this analysis on the JSE. As noted by Cartea *et al.* (2015), this pattern is seen due to trading at the beginning of the day being subject to greater uncertainty and being more informationally driven.

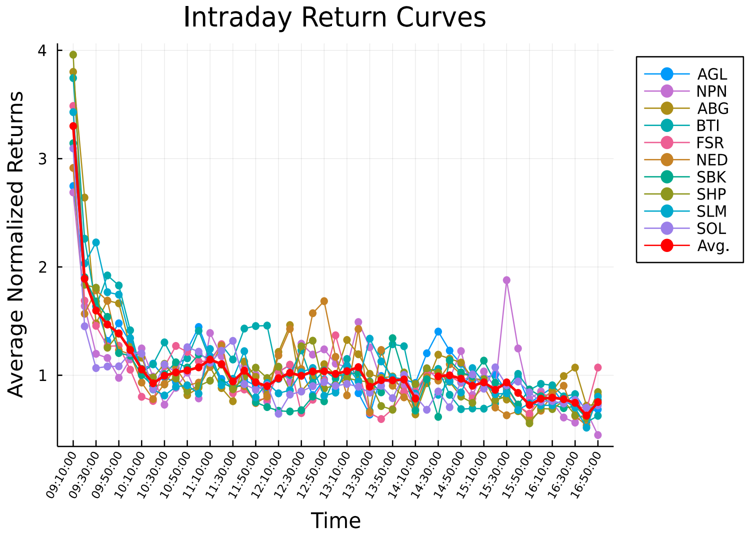

Figure 10: Plots the return curves for each of the stocks computed using Top-of-Book Trade-and-Quote data from 2019-07-08 to 2019-07-12. The red line represents the average return curve averaged over all stocks.

The final curve is the intraday spread curve. This curve is used as a proxy for liquidity. To compute the average spreads I used 10 minute bins and calculated the average spread in that bin, normalized by the average spread calculated for the entire day. The spread I have chosen to use is the quoted spread given by the following formula:

$$\mathrm{QS}_t = a_t - b_t \tag{5}$$

where a_t and b_t are the best ask and bid at time t. I have used the quoted spread in this form because it can be calculated at all time steps and not just time steps when a trade occurs.

Figure 11 shows the intraday spread curves for all the stocks as well as the average intraday spread curve. From the plot we can see that the quoted spreads are initially high but decline rapidly during the first half and hour of trading. The quoted spreads then stay relatively constant for the rest of the day. This pattern in the spreads was observed by Cartea *et al.* (2015) and Jericevich *et al.* (2020). This means that these results are consistent with previous empirical studies. Cartea *et al.* (2015) says that the large spreads observed during the first 30 minutes of trading is due to the increased uncertainty in the associated with the same period (this uncertainty was seen in Figure 10). They note that under greater uncertainty it is better to place wider bid-ask spreads. What is interseting to note in this curve is that there seems to be greater variation between the stocks' curves during the initial period. Some stocks have average spreads close to 10 while others have average spreads closer to 4. This could be due to traders of different stocks observing different degrees of risk aversion under the uncertainty in the initial 30 minutes.

In this section I have investigated the seasonality of the order-book. The volume, return and spread curves were presented and have all been able to recover the stylized facts found by Cartea *et al.* (2015) and Jericevich *et al.* (2020).

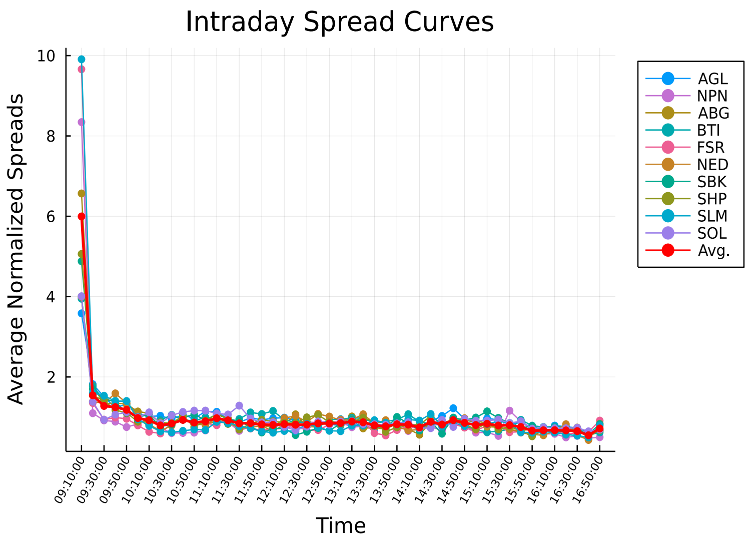

Figure 11: Plots the spread curves for each of the stocks computed using Top-of-Book Trade-and-Quote data from 2019-07-08 to 2019-07-12. The red line represents the average spread curve averaged over all stocks.

6 Code

The GitHub repository for this assignment can be found here.

References

1. Cartea, Á., Jaimungal, S. & Penalva, J. *Algorithmic and high-frequency trading* (Cambridge University Press, 2015).
2. Harvey, M., Hendricks, D., Gebbie, T. & Wilcox, D. Deviations in expected price impact for small transaction volumes under fee restructuring. *Physica A: Statistical Mechanics and its Applications* **471,** 416–426 (2017).
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. Lillo, F., Farmer, J. D. & Mantegna, R. N. Master curve for price-impact function. *Nature* **421,** 129–130 (2003).

Appendix: Data Snap-shot

Figure 12 shows a snap-shot of the cleaned TAQ data for AGL. One thing to note in this figure is that the time column is only showing the minutes and seconds. This is just an artifact of how excel is viewing the data. When reading into Julia the time column gives the full time of the trade.

timeStamp	date	time	eventType	bid	bidVol	ask	askVol	trade	tradeVol	normTradeVol	midPrice	midPriceChange	microPrice	interArrivals	tradeSign	id
2019-07-08T09:00:10.0	08/07/2019	00:10.0	ASK	NaN	NaN	37803	500	NaN	NaN	NaN	NaN	NaN	NaN	NaN		1
2019-07-08T09:00:10.0	08/07/2019	00:10.0	BID	37685	500	NaN	NaN	NaN	NaN	NaN	37744	-6.62E-05	37744	NaN		2
2019-07-08T09:00:11.0	08/07/2019	00:11.0	ASK	NaN	NaN	37798	500	NaN	NaN	NaN	37741.5	1.32E-05	37741.5	NaN		3
2019-07-08T09:00:11.0	08/07/2019	00:11.0	BID	37686	500	NaN	NaN	NaN	NaN	NaN	37742	2.65E-05	37742	NaN		4
2019-07-08T09:00:12.0	08/07/2019	00:12.0	ASK	NaN	NaN	37800	396	NaN	NaN	NaN	37743	0.000198693	37736.38393	NaN		5
2019-07-08T09:00:12.0	08/07/2019	00:12.0	BID	37701	500	NaN	NaN	NaN	NaN	NaN	37750.5	-3.97E-05	37744.75446	NaN		6
2019-07-08T09:00:19.0	08/07/2019	00:19.0	BID	37698	500	NaN	NaN	NaN	NaN	NaN	37749	-2.65E-05	37743.08036	NaN		7
2019-07-08T09:00:22.0	08/07/2019	00:22.0	ASK	NaN	NaN	37798	896	NaN	NaN	NaN	37748	-1.32E-05	37762.18338	NaN		8
2019-07-08T09:00:23.0	08/07/2019	00:23.0	ASK	NaN	NaN	37797	396	NaN	NaN	NaN	37747.5	0	37741.75446	NaN		9
2019-07-08T09:00:23.0	08/07/2019	00:23.0	BID	37698	500	NaN	NaN	NaN	NaN	NaN	37747.5	5.30E-05	37741.75446	NaN		10
2019-07-08T09:00:24.0	08/07/2019	00:24.0	BID	37702	374	NaN	NaN	NaN	NaN	NaN	37749.5	3.97E-05	37750.85714	NaN		11
2019-07-08T09:00:24.0	08/07/2019	00:24.0	ASK	NaN	NaN	37800	396	NaN	NaN	NaN	37751	0	37752.4	NaN		12
2019-07-08T09:00:24.0	08/07/2019	00:24.0	TRADE	NaN	NaN	NaN	NaN	37803.38	562	0.933670922	37751	0.000331062	37752.4	9	1	13
2019-07-08T09:00:24.0	08/07/2019	00:24.0	BID	37727	400	NaN	NaN	NaN	NaN	NaN	37763.5	0.000926392	37763.31658	NaN		14
2019-07-08T09:00:24.0	08/07/2019	00:24.0	ASK	NaN	NaN	37870	890	NaN	NaN	NaN	37798.5	-7.94E-05	37825.65891	NaN		15
2019-07-08T09:00:25.0	08/07/2019	00:25.0	ASK	NaN	NaN	37864	950	NaN	NaN	NaN	37795.5	-2.65E-05	37823.40741	NaN		16
2019-07-08T09:00:26.0	08/07/2019	00:26.0	ASK	NaN	NaN	37862	500	NaN	NaN	NaN	37794.5	0	37802	NaN		17
2019-07-08T09:00:31.0	08/07/2019	00:31.0	ASK	NaN	NaN	37862	1000	NaN	NaN	NaN	37794.5	0	37823.42857	NaN		18
2019-07-08T09:00:33.0	08/07/2019	00:33.0	TRADE	NaN	NaN	NaN	NaN	37704.34	7149	11.87689221	37794.5	-0.001072159	37823.42857	1	-1	19
2019-07-08T09:00:33.0	08/07/2019	00:33.0	ASK	NaN	NaN	37781	278	NaN	NaN	NaN	37754	-0.000423886	37749.14159	NaN		20
2019-07-08T09:00:33.0	08/07/2019	00:33.0	BID	37695	500	NaN	NaN	NaN	NaN	NaN	37738	1.32E-05	37725.73008	NaN		21
2019-07-08T09:00:34.0	08/07/2019	00:34.0	BID	37696	500	NaN	NaN	NaN	NaN	NaN	37738.5	-1.32E-05	37726.37275	NaN		22

Figure 12: Snap-shot of the cleaned trade-and-quote data for AGL.