15.459 Financial Data Science and Computing - Project F

Devin Connolly

December 10, 2019

1. **Question 1 - OHLC Data**

   (a) The OHLC values were calculated from the dataset. For these calculations, the raw dataset was truncated to only include trades between 9.30am and 4pm, as these are market hours. This was judged to be more accurate, particularly for open and close data. For securities that had multiple opening or closing prices (the last recorded trade time featured more than one price), the mode was used. If there was still a tie, the bigger order was used.

Asset	Date	Open	High	Low	Close	Volume (million)
AIG	5th May 2010	37.00	38.20	36.09	37.79	9.163
AIG	6th May 2010	37.78	38.62	33.37	36.88	12.973
AIG	24th August 2015	54.79	58.87	22.39	56.98	23.876
DIA	5th May 2010	108.5	109.56	108.25	108.8	18.514
DIA	6th May 2010	108.36	109.15	99.16	105.23	34.131
DIA	24th August 2015	156.00	163.32	150.79	158.41	39.609
KO	5th May 2010	53.29	53.9	53.20	53.67	11.074
KO	6th May 2010	53.67	54.20	51.21	52.27	22.546
KO	24th August 2015	37.84	38.91	36.56	38.40	80.677
PG	5th May 2010	61.71	62.57	61.50	62.15	11.012
PG	6th May 2010	61.91	62.67	39.37	60.74	27.085
PG	24th August 2015	68.37	71.06	65.05	69.14	24.175

(b) The Bloomberg data for the same securities is as follows:

Asset	Date	Open	High	Low	Close	Volume (millions)
AIG	5th May 2010	30.99	32.00	30.23	31.58	9.430
AIG	6th May 2010	31.65	32.35	27.95	30.79	13.265
AIG	24th August 2015	54.22	58.89	54.00	56.94	20.043
DIA	5th May 2010	108.55	109.56	108.25	108.8	18.900
DIA	6th May 2010	108.37	109.15	99.16	105.26	34.425
DIA	24th August 2015	155.90	163.45	150.57	158.38	35.102
KO	5th May 2010	26.72	26.95	26.60	26.83	23.671
KO	6th May 2010	26.84	27.10	25.61	26.15	47.917
KO	24th August 2015	37.99	38.94	36.56	38.38	44.065
PG	5th May 2010	61.62	62.57	61.50	62.16	11.612
PG	6th May 2010	61.91	62.67	39.37	60.75	28.566
PG	24th August 2015	68.57	71.07	65.02	69.14	23.696

- **AIG:** The values recorded on Bloomberg for May 2010 are approximately 6 dollars lower than those found in the data. This could be the result of a corporate action in the past that one source is accounting for and the other is not. To double check, I used Yahoo Finance, which reports the OHLC for AIG on May 5th and 6th 2010 as:

Asset	Date	Open	High	Low	Close
AIG	5th May 2010	37.00	38.20	36.09	37.70
AIG	6th May 2010	37.78	38.62	33.37	36.75

Which are almost exactly matching the values calculated from the data. My research shows there were no stock splits for AIG in the period between 2010 and 2015, so it must be a different form of corporate action (or the data scraped from Bloomberg is incorrect).

The volume figures for AIG were similar, with Bloomberg recording slightly higher volumes on each day in 2010 (approximately 300k difference), but a value 3 million lower in 2015. This could be due to a large volume of cancelled trades that have entered the dataset and are not recorded by Bloomberg.

- **DIA:** The results are very similar. For the 5th of May 2010, they are identical barring a 0.05 difference in opening. The largest discrepancy is in the low price for August 2015, which Bloomberg reports as 150.57, versus the 150.79 in the data. This 22 cent error is the largest.

  Once again, volume figures are very similar for 2010, but there is a difference of over 4 million shares in 2015.

- **KO:** The values for KO do not match at all for 2010. They are very similar, however, for 2015. Interestingly, though, the prices reported for Bloomberg are almost exactly half of those prices calculated in the data for May 2010. Doubling every price given on Bloomberg:

Asset	**Date**	**Open**	**High**	**Low**	**Close**
KO	5th May 2010	53.44	53.90	53.20	53.66
KO	6th May 2010	53.68	54.20	51.22	52.30

  A large portion of the values are now within a couple of cent of each other. Going to the Coca-Cola website in their investor relations section, I found evidence of a 2-for-1 stock split on the 27th of July 2012 (https://www.coca-colacompany.com/investors/stock-history/investors-info-splits). This stock split has obviously been backwardly imputed in the Bloomberg data, but not so in the data used (as it is actual trade and quote data).

  This same issue comes into play regarding volume as well, as the volume figures for KO for 2010 are approximately double on Bloomberg than those calculated with the data. However, the Bloomberg volume figure for 2015 is significantly smaller than the volume recorded in the data. Examining further, there are a number of large trades seem to occur numerous times in the data. For example, there is a trade for 2893 lots (289300 shares) that appears four times at the same time: 12:46:42.5850370. Interestingly, these trades are listed as having originated on difference exchanges. This occurs multiple times for numerous big values: 2385, 1153, 940.

- **PG:** For PG, the reported values are once again very similar. It is noteworthy that even the extremely low value of 39.37 (down from an open of 61.91) on the 6th of May was the same in both datasets. The volumes across the data and Bloomberg are similar for PG.

2. **Question 2 - Bid-Ask Spread and Order Imbalance**

The bid-ask spread was calculated by taking the difference between the ask price and bid price for every quote in the dataset, on the day, for each security ($P_{ask} - P_{bid}$).

The order-book imbalance was calculated using the quote data. For each observation, the following formula was used to calculate imbalance:

$$I = \frac{Q_{bid} - Q_{ask}}{Q_{bid} + Q_{ask}}$$

This gave the imbalance at every period throughout the day. This vector of imbalances was then used to calculate the mean, minimum, and maximum. This approach assumes that the bid size and ask size fields in the data give an accurate number of quotes at the bid and ask on the order book at that period (which according to the documentation is true).

I would say that this result can only be called an estimate. As the data is so high-frequency, and order is imperative when calculating intra-second order-book imbalance, we cannot reliably say that this is the exact result. There is no guarantee by the data provider that order is maintained perfectly. I do however expect the mean value to be relatively similar to the true mean value, due to the large number of observations that fed into the calculation.

I would also argue that using solely NBBO data does not give an accurate representation of order book imbalance. Take for example the following case, where data beyond the best bid and best offer are included (first entry is price, second is volume):

$$\begin{cases} \$103\ 500 \\ \$102\ 1000 \\ \$101\ 1100 \\ \$100\ \ \text{Price} \\ \$99\ 1100 \\ \$98\ 0 \\ \$97\ 0 \end{cases}$$

Here, by the NBBO data, the order book is perfectly balanced. However, using the other data, we can clearly see there is a sell-heavy market here (plenty of people willing to sell not many willing to buy). The counter-argument is that NBBO data offers an accurate representation of the prices investors are currently trading at, and only these prices should be considered when measuring imbalance. However, because the data changes so rapidly, maybe it is wise to extend slightly beyond NBBO.

- AIG - There is an average bid-ask spread of 5.68 cent, while there is on average a positive order imbalance, indicating it is a slightly sell-heavy market on this day.

Metric	**Minimum**	**Maximum**	**Mean**
Bid-Ask Spread	0.01	10.3	0.056815
Order Imbalance	-0.990	0.9847	0.01147

- DIA - The average bid-ask spread for DIA is 3.21 cent. It makes intuitive sense that DIA has a narrower range, as it is a much more liquid security. The minimum spread was -159.73, but analysing the two observations where this occurred shows that the ask at this times was 0 - an error in the data, most likely. Ignoring those two observations, the actual minimum was 1 cent. The maximum spread of almost $44 seems high, and may be an error in the data. The 99% percentile for spread was $1.39. Once again, there is a buy-heavy order imbalance on DIA on this day.

Metric	Minimum	Maximum	Mean
Bid-Ask Spread	-159.73 (0.01)	43.84	0.03209
Order Imbalance	-0.975	1	0.06766

- KO - The value for KO are relatively straightforward, once again with a positive order book imbalance (buy-heavy) and average bid-ask spread of 3.46 cents.

Metric	Minimum	Maximum	Mean
Bid-Ask Spread	0.01	10.86	0.03464
Order Imbalance	-0.9854	0.9953	0.002016

- PG - PG yields interesting results compared to the other securities. PG has the largest mean bid-ask spread of 8.7 cents, while being the only security with a negative order-book imbalance (a sell-heavy market).

Metric	Minimum	Maximum	Mean
Bid-Ask Spread	0.01	10.81	0.08732
Order Imbalance	0.9902	0.9684	-0.02529

3. **Question 3 - Five-Minute Times Series**

The five-minute time series plots were constructed as follows:

(a) Create a time-stamp column in each dataset (security and day) by concatenating the date and time. Order the dataset by this time-stamp.

(b) Calculate the end points of each five-minute window using the 'endpoints' function in the xts package. Then use period.apply to calculate the mean price between each of these endpoints.

(c) Cast each mean to the next end-point time. For example, the mean between 9:30 and 9:34:59 goes to 9:35:00.

(d) This results in a value for 15:59:59 (for period 15:55 to 15:59:59) and 16:04:59 (for all prices at 16:00:00 exactly) . To work around this, the average between these two values was cast and applied to the last endpoint. This assumes that the number of observations in both periods is similar. In reality, this is not true. A weighted average could be used, or the final value could be added as a single number to the previous observations, and the average subsequently calculated. However, these two last price values are usually very similar, so the method is unlikely to change the result drastically.

The five-minute time series are shown below.

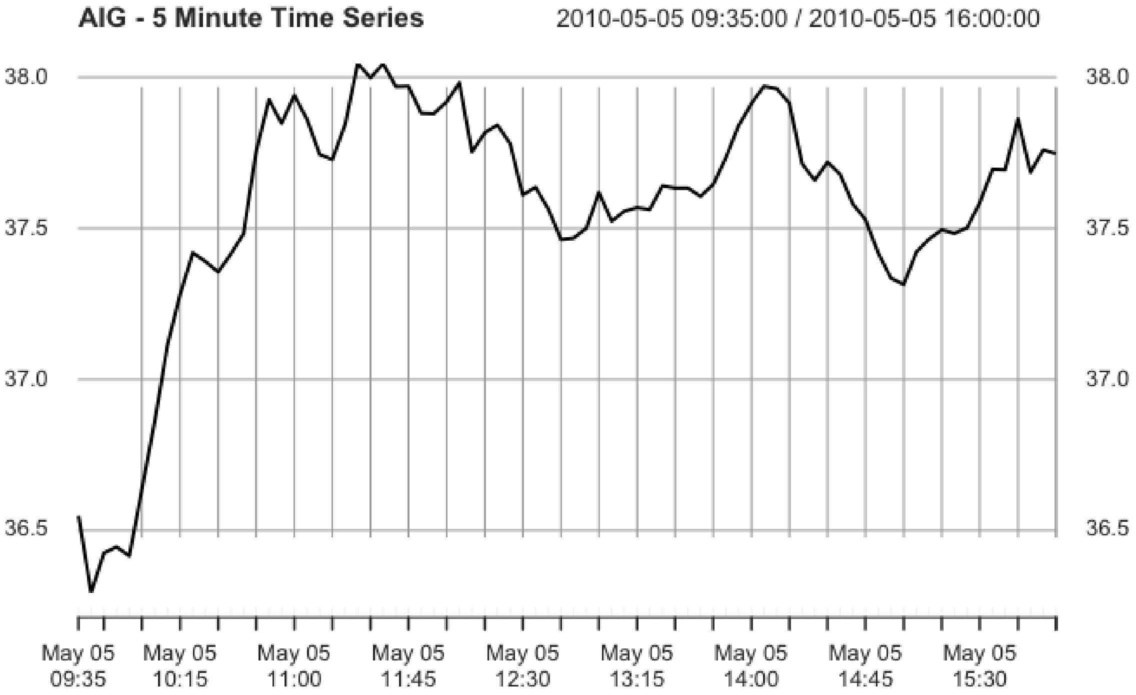

Figure 1: AIG 5th of May 5-Minute Series

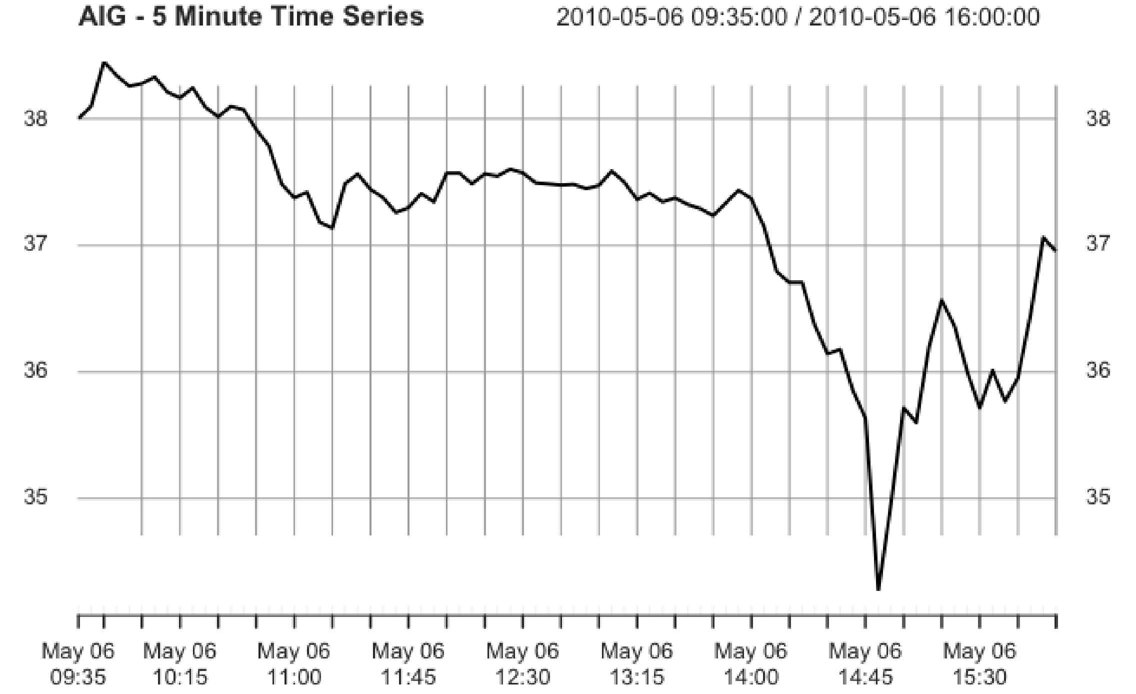

Figure 2: AIG 6th of May 5-Minute Series

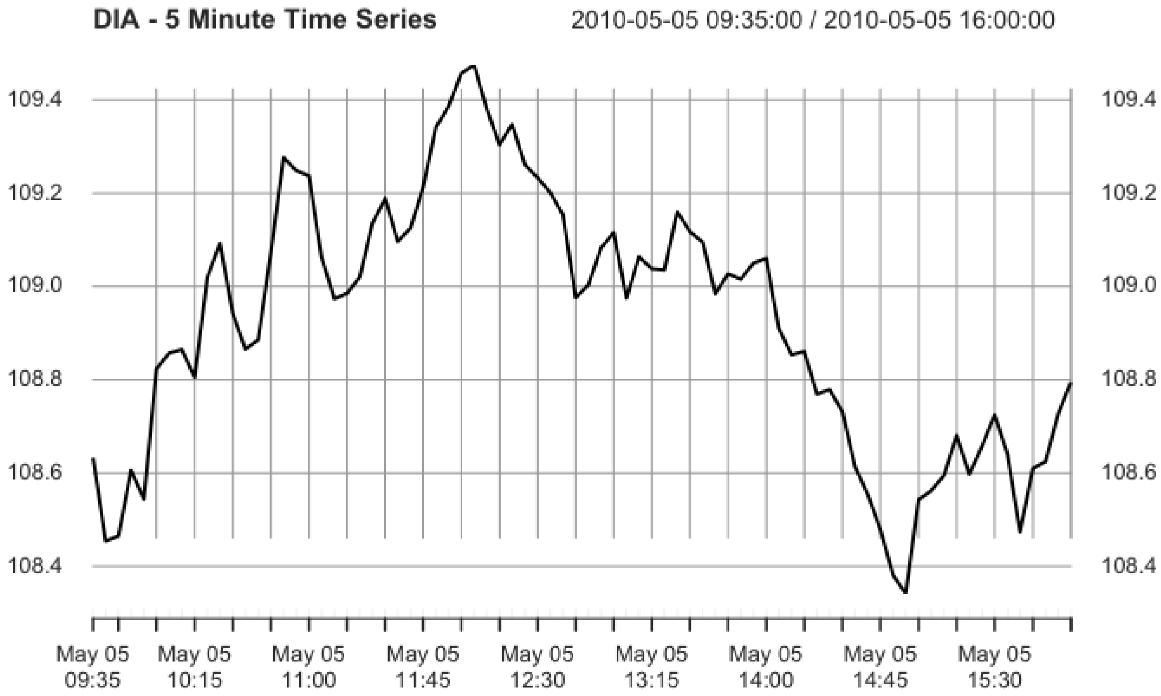

Figure 3: DIA 5th of May 5-Minute Series

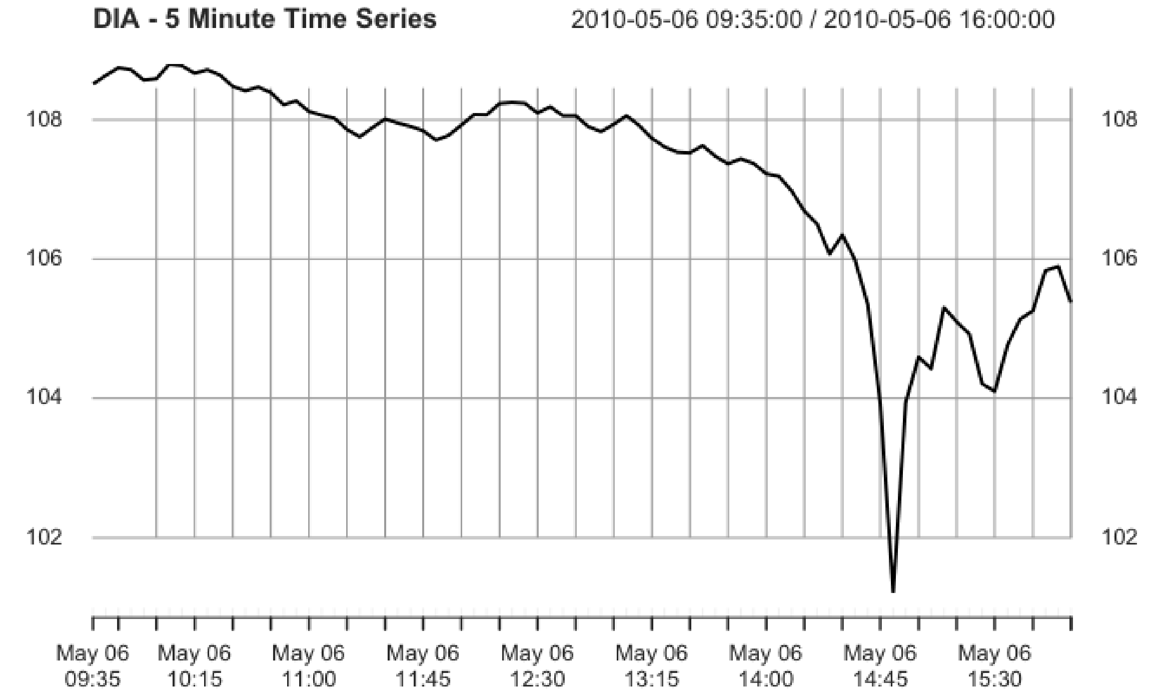

Figure 4: DIA 6th of May 5-Minute Series

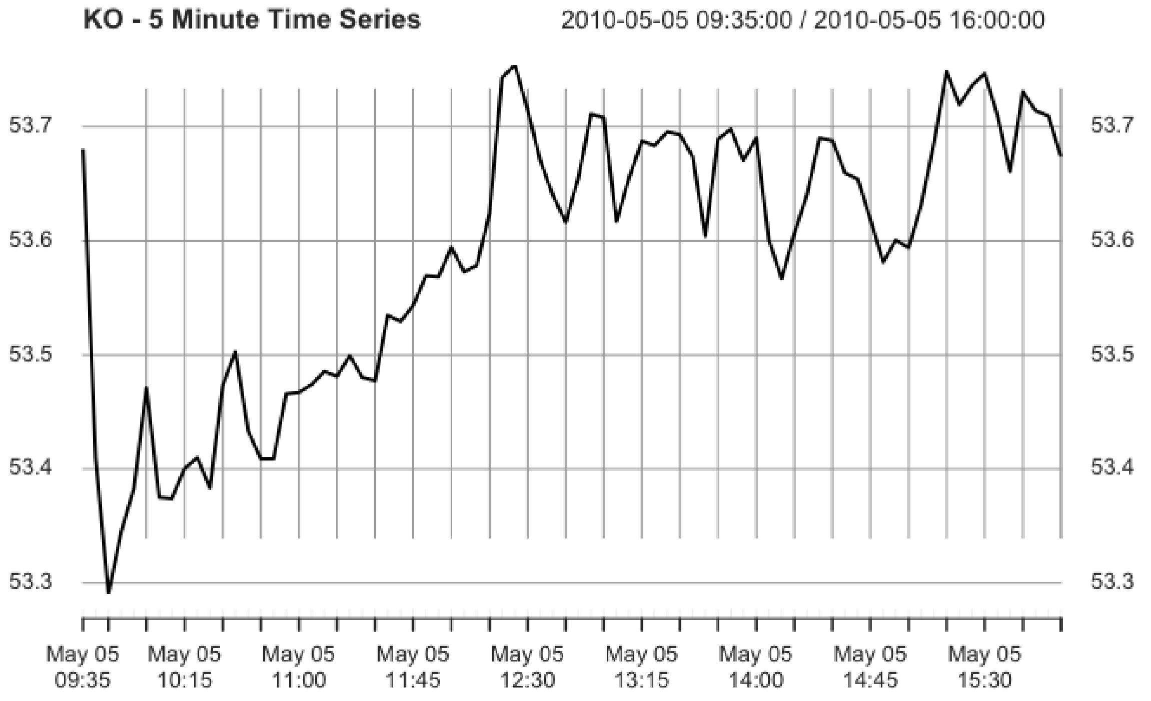

Figure 5: KO 5th of May 5-Minute Series

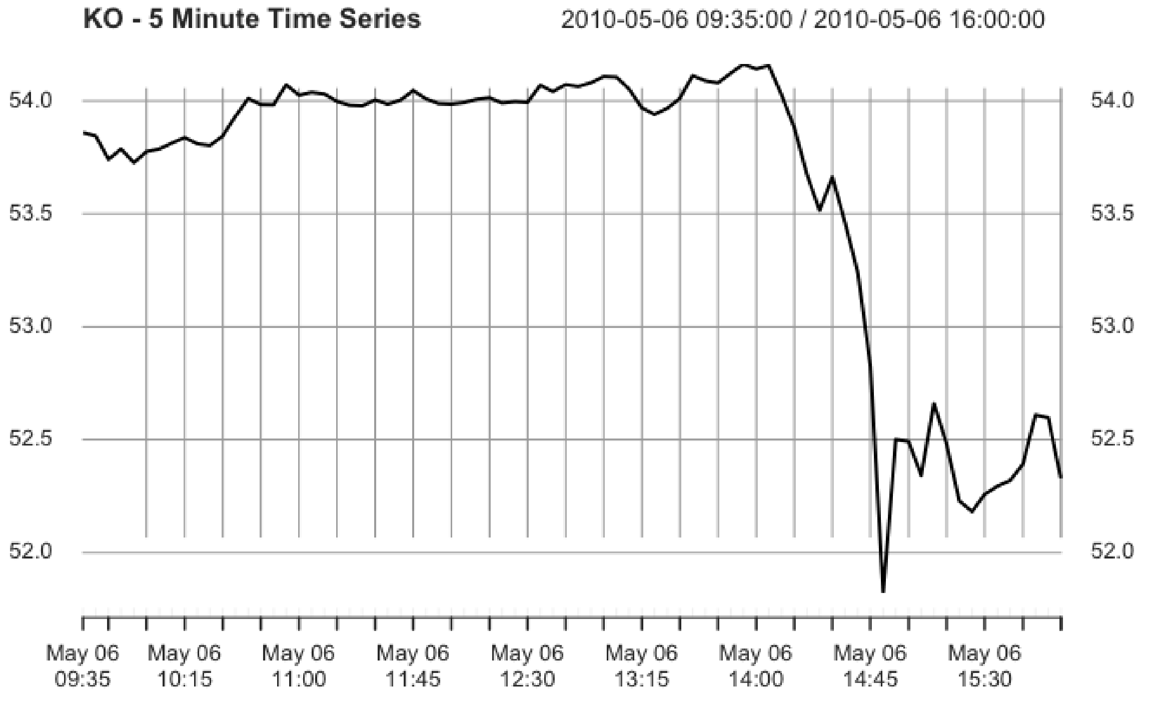

Figure 6: KO 6th of May 5-Minute Series

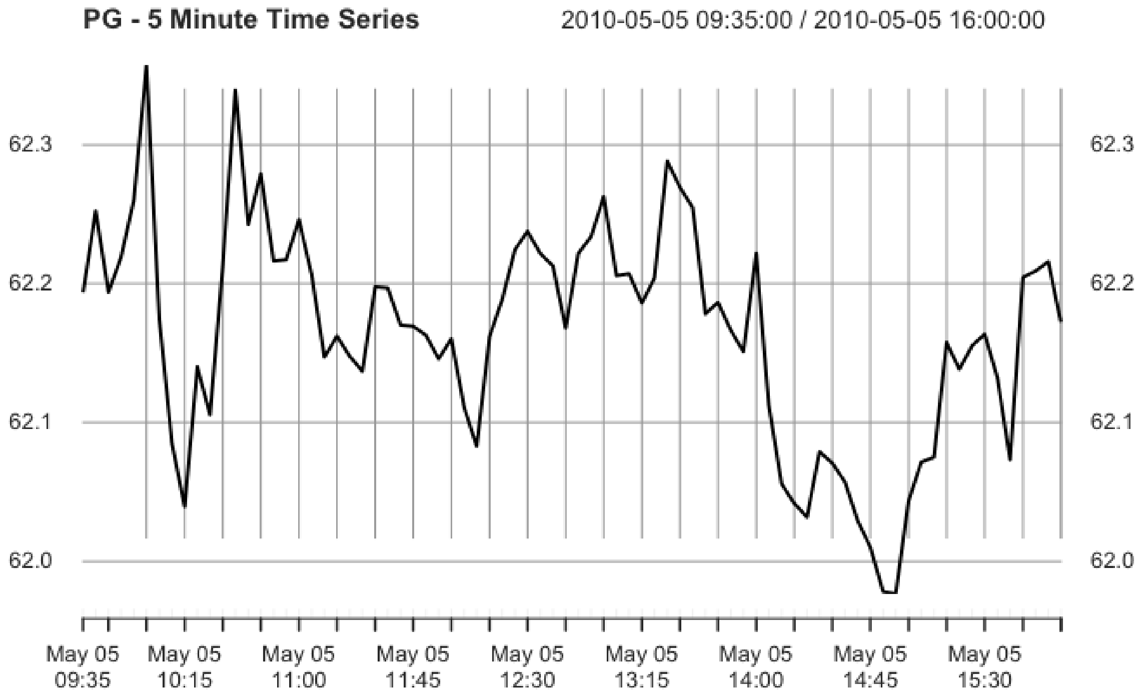

Figure 7: PG 5th of May 5-Minute Series

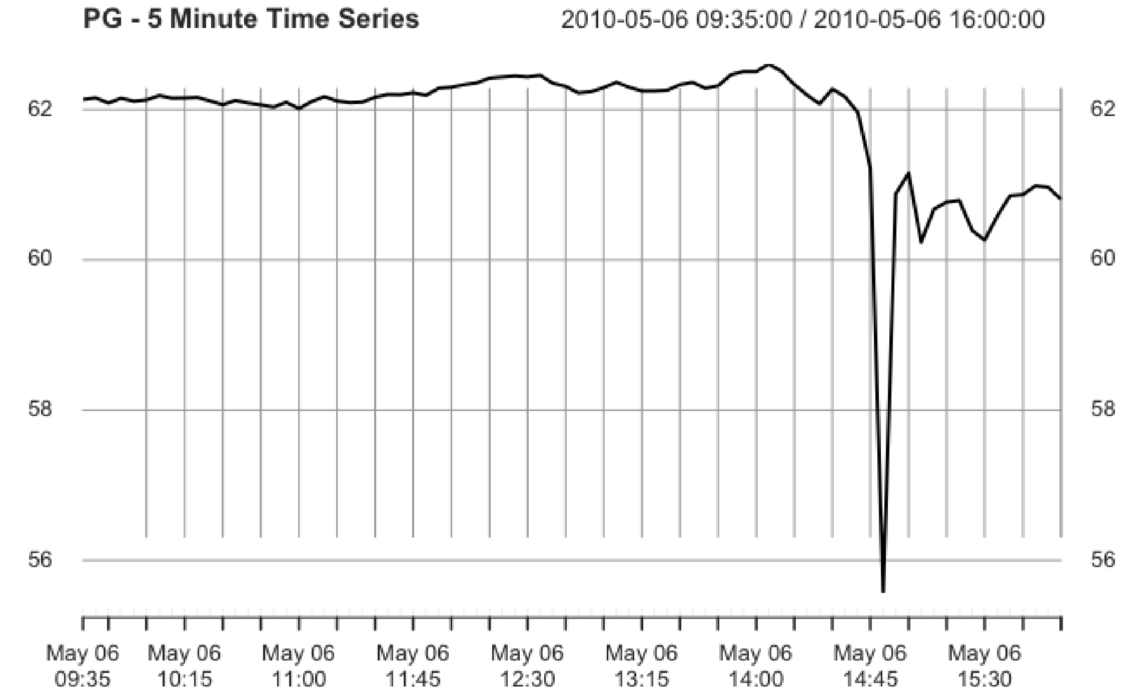

Figure 8: PG 6th of May 5-Minute Series

(c) The minimum values comparison between the five-minute time series and the overall data is shown in the table below.

Asset	Date	Min Time (5)	Min Value (5)	Actual Min
AIG	5th May 2010	09:40:00	36.29	36.09
AIG	6th May 2010	14:50:00	34.27	33.37
DIA	5th May 2010	14:55:00	108.34	108.25
DIA	6th May 2010	14:50:00	101.21	99.16
KO	5th May 2010	09:45:00	53.29	53.20
KO	6th May 2010	14:50:00	51.82	51.21
PG	5th May 2010	14:55:00	61.98	61.50
PG	6th May 2010	14:50:00	55.56	39.37

Of course, the five-minute time series minimum is higher than the true minimum in every case. This is because the five-minute series takes an over a low period (approaching the true low, and bouncing back from). The values were relatively similar, apart for PG on the 6th, where they differed by over \$16. This indicates that the flash-crash had a particularly drastic effect on PG. All of the five-minute lows on the 6th occurred at 14:50, which is when the flash-crash happened.

4. **Question 4 - Volatilities**

Volatility was calculated using the returns over the five-minute time period (return from one period in the five-minute series to the next). This was then multiplied by the square root of 365. When deciding whether to use 252 or 365, I choose 252 as it is the standard number of trade days in a year.

The implied volatility was found using OLAP, by using the May 10th call with strike closest to price. The realized volatility was found by gathering historical price data for each security for the 42 days previous to May 4th (inclusive) and calculating volatility of returns over that period. The same realized return is displayed for both days for each security.

The results were as follows:

Asset	Date	Five-Minute Vol	Realized Vol	Implied Vol
AIG	5th May 2010	4.49%	74.82%	59.06%
AIG	6th May 2010	12.15%	74.82%	62.39%
DIA	5th May 2010	1.42%	13.30%	6.78%
DIA	6th May 2010	8.15%	13.30%	60.03%
KO	5th May 2010	1.60%	13.99%	13.07%
KO	6th May 2010	5.21%	13.99%	36.93%
PG	5th May 2010	1.39%	10.98%	14.94%
PG	6th May 2010	24.65%	10.98%	35.21%

Strikes Used: AIG (38.86), DIA (103), KO (52.50, Stock-Split Adjustment), PG (59.167)

There is significant difference in volatilities across all three metrics. The five-minute time series volatilities are understandably low, as they are computed using prices that are averaged over five-minute periods, which will remove a large amount of variance. This is true even in the case of May 6th, when the flash-crash occurred (although PG recorded five-minute volatility of 24.65% here - the highest by far). The realized volatilities are good measures of short-term historic volatility, but the value for AIG show that even these are susceptible to noisy periods. The IV values were very large for AIG (a volatile time for the stock, judging by the realised volatility), and significantly increased from the 5th to the 6th for each of the other stocks. The IV calculation includes pricing data from the 5th and 6th, while realized volatility was calculated using data up until the 4th, which could explain a portion of the difference. IV increased tenfold in the case of DIA (a stable stock usually). It is clear the impact that the flash-crash had on IV at the time.

5. Appendix - SQL Queries

Query for taq database (2010 data):

```
use taq

select * from dbo.trade T
inner join dbo.dim_time DT on (T.time  = DT.t)
where T.date between ’2010-05-05’ and ’2010-05-06’
and T.symbol in (’AIG’,’DIA’, ’KO’,’PG’)
```

Query for taq 04 database (2015 data):

```
use taq04

select
Q.date,
Q.time_m,
Q.ex,
Q.sym_root,
Q.sym_suffix,
Q.bid,
Q.bidsiz,
Q.ask,
Q.qu_cond,
Q.bidex,
Q.askex,
Q.natbbo_ind,
Q.nasdbbo_ind,
T.tr_scond,
T.size,
T.price,
T.tr_source
from dbo.quote_msec Q
inner join dbo.trade_msec T on (Q.time_m = T.time_m) and
                (Q.date = t.date) and (Q.sym_root = T.sym_root)
where Q.date = ’2015-08-24’
and Q.sym_root IN (’AIG’,’DIA’, ’KO’,’PG’)
```