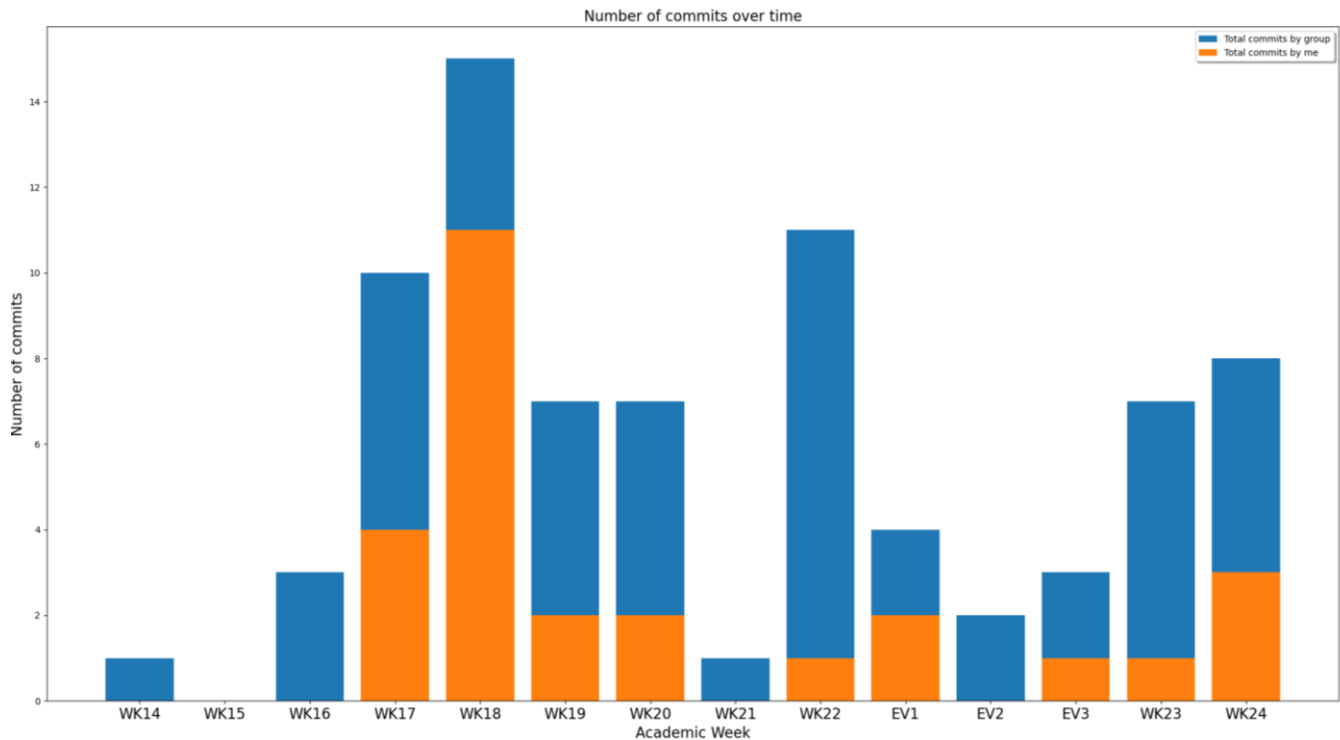**Individual Reflective Report**

individual project activity and visualisations

Below is a data visualisation of the number of my commits versus the rest of the group, throughout the project; it should serve to visualise my productivity throughout the project.

My tasks throughout the project:

During weeks 14 – 15 the group was setting up our Github repository and setting up our Trello Kanban board. As you can see from the graph, I hadn't made any git commits between weeks 14-16; this was because I was reading around the academic literature. While I did not make any git commits at this time, I did distribute an early draft of the literature review so that the group could start making long term plans for our project.

From weeks 16 – 17 I was working on a python script to convert the LOB txt files into CSV files. In week 17 I committed the LOB text file to CSV converter. However, I would continue to make changes to this file in weeks 17 and 18. These changes were made so that this solution could scale up to all the data in our large dataset.

During week 18 I would create a few utility scripts for managing the large data files. During that week I also spent a lot of time making a LOB data compressor. As you can see from the graph, I made a lot of commits during this week 18 and that is because I was constantly changing the data compressor so that it would scale up to more data. In the end the development of this file was abandoned, and it turns out that there were over ways to reduce the size of the dataset. In week 18 I also started converting all the LOB text data into CSV files and then uploading them to our S3 bucket.

In week 19 I committed a new file that would automatically add LOB features, that I had read about in academic literature, to the dataset. During this time, I also was continuing to convert the LOB text files into CSV.
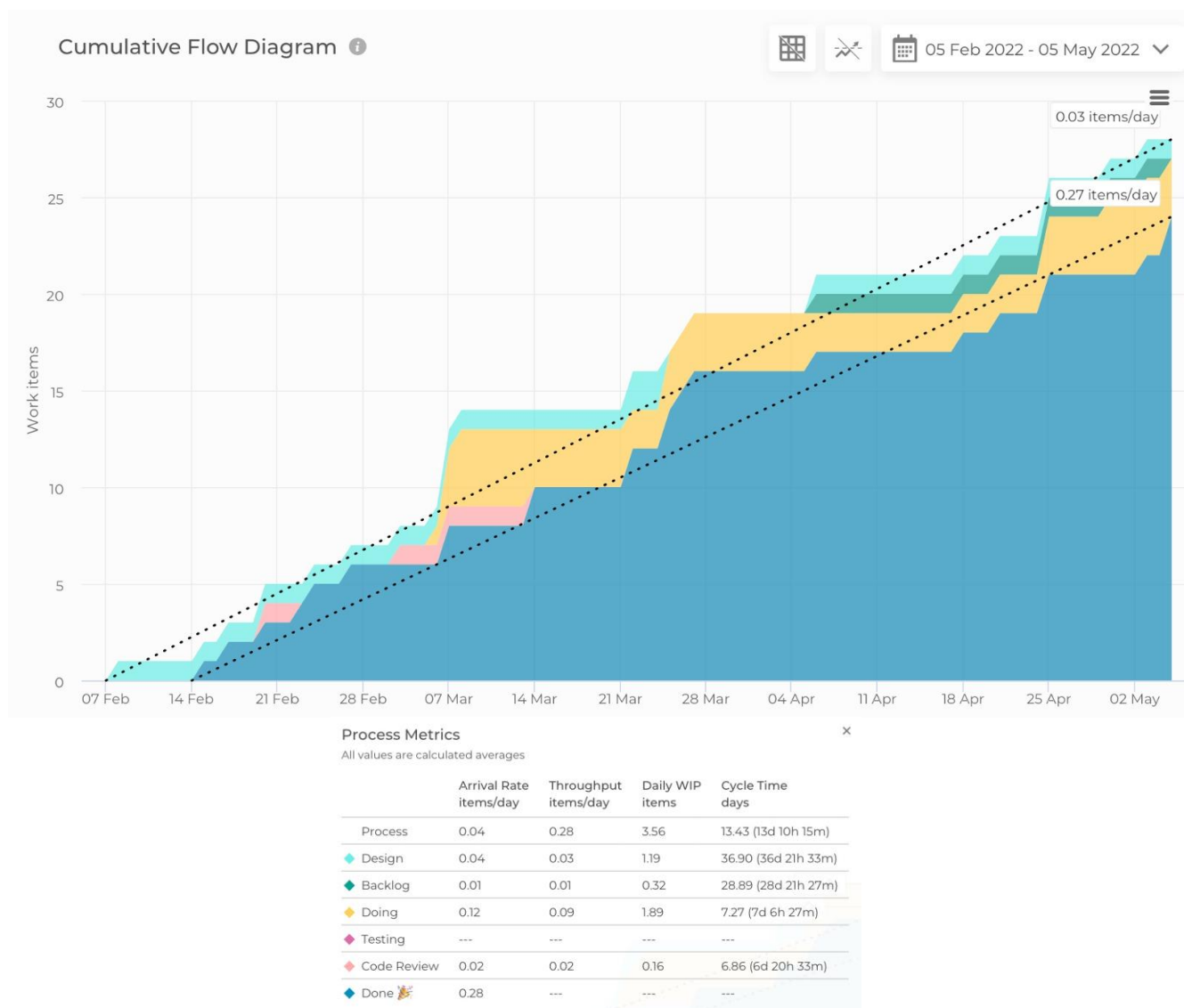
For week 20 I explored the LOB data and committed a script that would visualise the insights that I had learnt about. These visualisations were then used in the group formative presentation, which I helped create, with the rest of the group

In week 21 the group was experiencing a bottleneck in productivity. The group wanted to start designing multivariate models, but the data was not available yet, as it would take literal days of non-stop computation to generate this data. As you can see from the graph, I did not commit anything to the git repository. That was because no one else in the group had enough secondary storage capacity to store all the data necessary to convert all the data into feature data and so I alone was converting all the data in the dataset. This took two solid days of computation to do.

In week 22 all our data, that I had generated from the week before, was wiped from our S3 account. We confirmed that this was of no fault of anyone in the group. Due to this massive data loss, I repeated what I did in week 21 to generate all the data again. I also committed one file so that I could sequentially download the file rather than downloading them in parallel, which would take a lot of time and the downloads would occasionally fail.
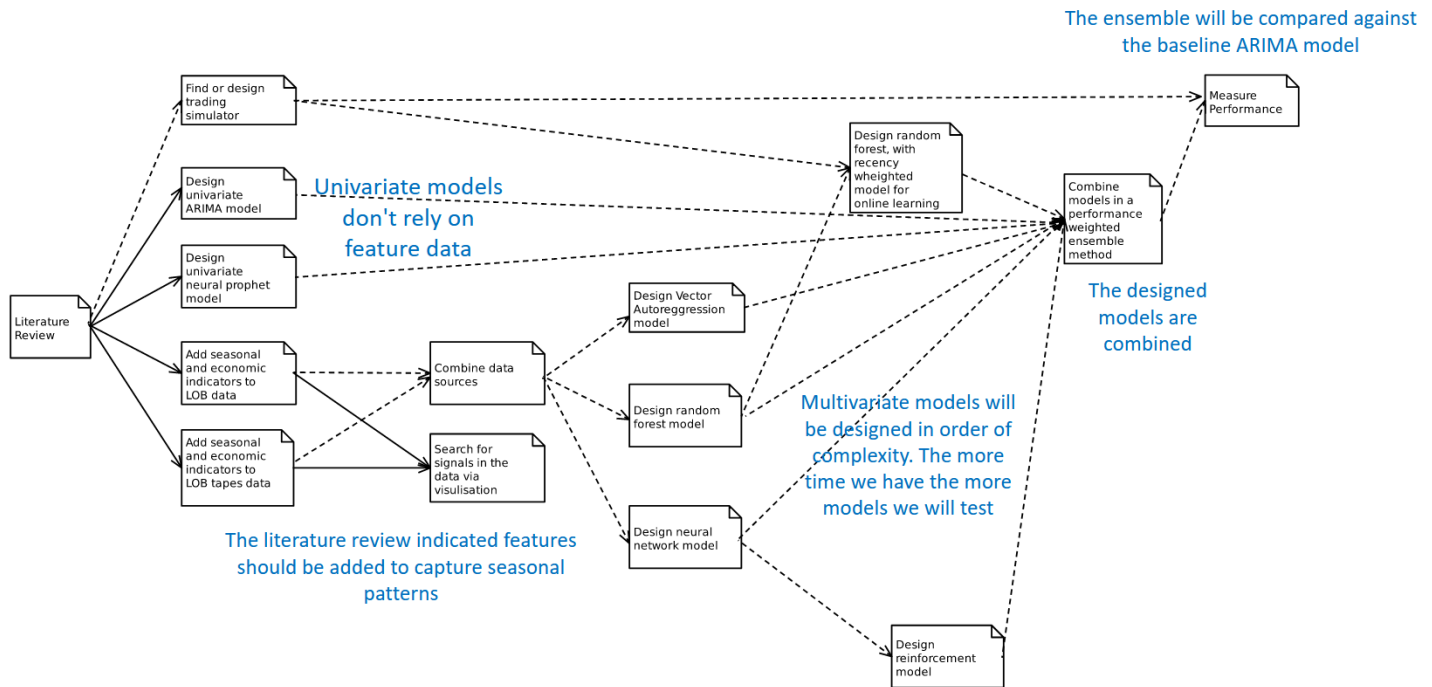
During the easter holidays I created a script that would combine the LOB and tapes features together, so that they could be used in multivariate models. At the end of the easter and the start of WK23. I designed an algorithm that would convert the regression output into transaction decisions.

WK24 was most assigned to help write the group report but I wanted to be able to quote our test results, which we had not generated yet. To do this I worked with Aditya to get his trading simulator to use my transaction decisions.

Cumulative Flow Diagram ⓘ

🗓 05 Feb 2022 - 05 May 2022 ⌄

0.03 items/day

0.27 items/day

Process Metrics ✕
All values are calculated averages

	Arrival Rate items/day	Throughput items/day	Daily WIP items	Cycle Time days
Process	0.04	0.28	3.56	13.43 (13d 10h 15m)
◆ Design	0.04	0.03	1.19	36.90 (36d 21h 33m)
◆ Backlog	0.01	0.01	0.32	28.89 (28d 21h 27m)
◆ Doing	0.12	0.09	1.89	7.27 (7d 6h 27m)
◆ Testing	---	---	---	---
◆ Code Review	0.02	0.02	0.16	6.86 (6d 20h 33m)
◆ Done 🎉	0.28	---	---	---

Above you can see the group cumulative flow diagram, which show the groups Trello progress over time. As you can see from the CFD we made regular progress, apart from the moments were the group's productivity was bottlenecked by not having the data ready.

I would make sure that everyone in the group had something to do in the week so that we could maintain our productivity. I constructed a graph for the formative presentation about our projects plan; it shows the dependencies of various stages. This illustrates how our group was able to work in parrallel to achieve a high level of productivity.

Literature Review

Find or design trading simulator

Design univariate ARIMA model

Design univariate neural prophet model

Add seasonal and economic indicators to LOB data

Add seasonal and economic indicators to LOB tapes data

Combine data sources

Search for signals in the data via visulisation

Design Vector Autoregression model

Design random forest model

Design neural network model

Design random forest, with recency wheighted model for online learning

Combine models in a performance weighted ensemble method

Measure Performance

Design reinforcement model

The ensemble will be compared against the baseline ARIMA model

Univariate models don't rely on feature data

The designed models are combined

The literature review indicated features should be added to capture seasonal patterns

Multivariate models will be designed in order of complexity. The more time we have the more models we will test

Individual weekly log

Student Name: Michael McCoubrey

UoB Username: nc21482

Week	Topic	Notes
15	Individual tasks:	- Read literature discussed in this week's collaborative tasks
	Collaborative tasks:	- Set up Kanban board using Trello - Set some initial tasks/tickets on Trello - Set up a common git repository both locally and on GitHub - Set up a teams group and a WhatsApp group - The team met up and discussed: o We will be using Jupyter notebooks as they will help the group organise and present our work using markdown and Latex. They will also allow for partial computation of the programs o We will all aim to read ▪ Project leader/provider's (Dr Ash Booth) blog • https://www.ashbooth.com/blog/ ▪ Project leader/provider's (Dr Ash Booth) PhD paper • https://www.ashbooth.com/publications-talks-and-presentations/ ▪ The references provided in the project description o If we have enough time we should inspect the data and if required we should start considering how to clean the data

	Challenges:	- Initially we were considering using PyCharm as an IDE we would all use but due to experiencing errors in MacOS we chose to use Jupyter Notebooks. Jupyter Notebooks were chosen for the reasons discussed in this week's collaborative tasks but we also knew that it worked for everyone in the group as we had used it in current course modules.

16	Individual tasks:	- Further research on Ash Booths publications
		- Work on the data cleaning and compression of the limit order books data
		   o New data only shows transactions to the limit order book
		   o New data will be converted into csv tabular data

	Collaborative tasks:	- Had a video call with team members, discussing on how we were going to go about copying the large amount of files to our S3 bucket
		   o We decided that we would each be responsible for a proportion or the files to download and then upload to our S3 bucket
		   o In the end we found an easier way of just copy the contents of one S3 bucket to another S3 bucket

	Challenges:	- The amount of data that we had to upload to our S3 bucket was too big for anyone group member to download so we either planned to split the download up but in the end copying the data from s3 bucket to s3 bucket was quicker and simpler
17	Individual tasks:	- I converted the LOB data into a csv format, but more work still needs to be done to compress this data so that only the beginning and end times of the data are recorded

	Collaborative tasks:	- Had a team meeting, on Thursday, discussing progress:
		   o Amy presented her progress on using the ARIMA model. She asked that she take the average of the data in each day, which the group agreed to
		   o Aditya was having issue mapping the text files to a csv document as they were not strictly in a JSON format. I said I would also work on this.

	Challenges:	- To get the start and end time of LOB data you need to look for occurrences of the same amount and quantity throughout the history of the data, meaning that multiple orders of the same value and quantity could be present. To fix this I will have to look for gaps in time periods in the data.
18	Individual tasks:	- I attempted to compress the LOB csv data so that only the beginning and end times of the data are recorded. I was able to do this but the solution wouldn't scale to a larger amount of data - I continued to use the LOB txt to csv converter for the the majority of the text files and stored them in the s3 bucket

	Collaborative tasks:	-	The team met up and talked about are current progress and our plans for ensemble methods
	Challenges:	-	The program I wrote that compressed the LOB csv data worked but did not scale to the amount of data. I tried to reduce the big O notation of the algorithm, but it did not help. The program would have taken 18 days to convert 1 days' worth of data.
19	Individual tasks:	- -	I wrote a program which takes in the LOB csv data and creates features for each unit of time. I again used the LOB txt to csv converter and stored the files in the s3 bucket
	Collaborative tasks:		
	Challenges:		
20	Individual tasks:	- - - -	I've been working on converting LOB csv data into LOB features and then merging it with the tapes data I explored the data to look for feature in the LOB data I've been looking for signals in LOB data so that I can make useful visualisations for the presentation we have this week I've contributing to the presentation slides to talk about the literature review and the our future plans
	Collaborative tasks:	-	We met up on the Monday and discussed our current progress and plans for the presentation as part of the formative assessment
	Challenges:	-	Amy and Brooke, have been working on an ARIMA and Neural Prophet models but they have poor prediction ability, due to a lack of seasonality from the data they were training it on.
21	Individual tasks:	- - -	I've been working on converting all the LOB csv data into LOB features and storing it in S3 merging some of the tapes and LOB data, via using tableau prep Experimenting with alternative ways of merging the LOB data with the tapes data so that data models will predict the future price rather than the current price
	Collaborative tasks:	-	I've needed to work with Aditya so that I could use his tapes feature files, so that I could combine them into one dataset
	Challenges:		
22	Individual tasks:	- -	I worked on converting all the data again and then uploading it to S3 I made an offline backup of all the data, in case we experience another data loss
	Collaborative tasks:	- -	We have started to write our group report The group met up and discussed our plan for the next five weeks
	Challenges:	-	Our AWS S3 bucket was deleted (not by a team member) and so we lost all our data and data that we had converted.

		- It takes days of computation to convert all the data we were provided with so due to the bucket being deleted this took a long time. - It would take a long time for files to be downloaded from s3 as they were being downloaded in parallel, so I wrote a program to sequentially download the files so that I could start converting individual files
23	Individual tasks:	- I wrote a program to merge all the tapes and lob files together - I wrote a program that took the output of regression models and produced transaction decisions. This also allowed me to consider our trading strategy for maximising profit
	Collaborative tasks:	- I again used Aditya's tapes feature files which were again uploaded to S3 - I used the output of Amy's rForest.py a random forest regression model for deciding transactions - The group has had many group call for planning and discussing the project.
	Challenges:	- Tableau prep was merging the data into an inconsistent format and so I made a program that would do it more consistently
24	Individual tasks:	- Writing literature review section of the group report - Writing LOB subsection of the data preparation section of the group report - Writing transaction decision algorithm section of the group report - I noticed that GitHub only provided data visualisations of project progress to either public repos or pro accounts. o Instead of using this I wrote a python script to visualise a group member's progress in relation to the rest of the overall group's progress. As this python script was simple and general in use (and not specific to this project), I made it publicly available on my own GitHub repository and allowed my group members to use it for their individual reports.
	Collaborative tasks:	- Updating transaction decision algorithm to be more accurate and in the format that Aditya's trading simulator required. - Talking with Aditya to find out where the source of the inaccuracy was in the transaction decision algorithm - The group met up to make final changes to the group report.
	Challenges:	- Initially the transaction decision algorithm was inaccurate. Luckily though the algorithm was not ineffective but was erroneously causing a statistically significant loss in profit. Noticing this I was able to change the area that I knew was wrong in my algorithm, allowing it to perform as intended. The performance of this strategy was surprisingly strong. - The general lack of time due to the need to be productive in other units of my degree.