Data understanding

1. Gathering data

- **Data requirements:** Every entry of our data should have the following parameters: Date(Date object), when the trade was made, Open(a float value), the value the stock was at when the stock market opened on that specific date, High(float value), the max value of the stock during that specified date, Low(float value), the lowest value of the stock during that specific date, Close(float value), the value the stock ended at the end of that date, Volume(Integer value), the amount of stock that was traded during that date.
- **Data availability:** We have a few thousand files with 5 thousand to 10 thousand entries per file. Each file is about a different company so in case our scope is too broad, we can filter the dataset with a specific criterion like IT companies or predict the stocks for a specific company. And as the data is in a chronological format then we could narrow the scope to a certain period.
- **Selection criteria:** Our data is in text files that we downloaded from ics.uci.edu archives. The relevant fields in the dataset are the previously mentioned Date, Open, High, Low, Close and Volume fields that are in every file.

2. Describing data

Our data's brief explanation is "Company stock price changes during several years". Our dataset is from ics.uci.edu archives and the dataset is meant for machine learning activities, which reduces the need for exorbitant data preparation to get the data into a state that we can run classifiers or neural networks on. The fields in the dataset are already data types that we can use straight away. The fields in this dataset are : Date, Open, High, Low, Close, OpenInt. The OpenInt field is the only field that has no real meaning for us. The dataset is made up of over seven thousand files each describing the stock price changes by days. Averaging multiple thousand entries per file means that we have plenty of data to work with and the dataset is described as a Time-Series dataset meaning that the entries are formatted chronologically. This gives us quite a deal of freedom to analyse

periods of time for the different companies and deduce what kind of events in the world may have influenced the stock prices of those companies and find trends in the prices of the stock by date. This also allows us to gather more data for neural networks to deduce whether certain events in the world had any effect on the stock prices in our dataset.

3. Exploring data

Regarding the available data that we currently have, it seems that the earliest entry was in the 1970s, and the latest in the 2010s, 2017 to be correct. But, since our dataset is massive, more than a thousand files of content, we should have a really good chance of applying machine learning and neural networks correctly to predict the stock market in the future. Each company has its own text file and in it all the data we need. From the amount of stock traded to the highest and lowest price of that particular date. Because every change in the stock market is shown, there really are not problems with data quality.

Because our dataset is taken from archive.ics.uci.edu archives as stated before, the data is already in the format we need, so it saves us the time needed to set up the info in the wanted form and the stage for data preparation is set.

4. Verifying data quality

For our dataset the quality is good, there does not seem to be any problems regarding the given content, only 1 field, OpenInt, we will not be using, since it has no actual use for us. Better yet, every company is in its own text file and in chronological order, with every change in the given date recorded. Because of this, we do not need to look for alternatives to get data, more likely if the project's scope is too large, we may need to cut down on the companys we look through, for example filter out non-IT companies and the like. And so we have verified that the given data is enough to complete our goals.