School of Computing

Lower Kent Ridge Road, Singapore 119260

BT3017

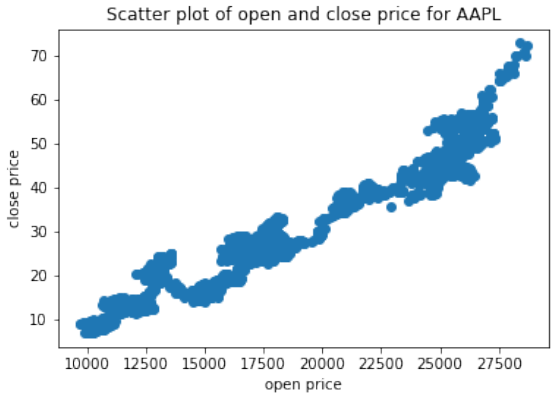
Feature Engineering for Machine Learning

Noah Teo Rui-Sheng - A0222800X

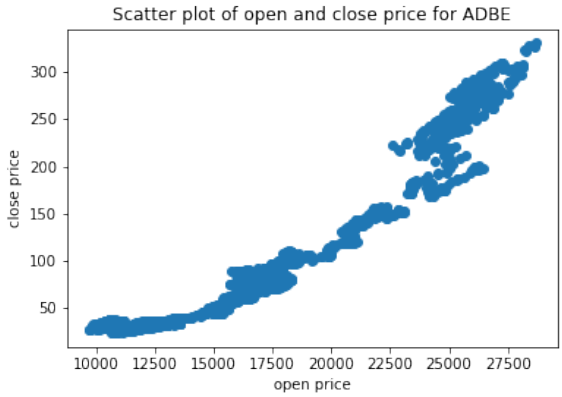
Q1a) The following Dataframe is Com12:

	Date	Open	High	Low	Close	Adj Close	Volume	Name
0	2010-01-04	7.622500	7.660714	7.585000	7.643214	6.562591	493729600	AAPL
1	2010-01-05	7.664286	7.699643	7.616071	7.656429	6.573935	601904800	AAPL
2	2010-01-06	7.656429	7.686786	7.526786	7.534643	6.469369	552160000	AAPL
3	2010-01-07	7.562500	7.571429	7.466071	7.520714	6.457407	477131200	AAPL
4	2010-01-08	7.510714	7.571429	7.466429	7.570714	6.500339	447610800	AAPL
...	...	...	...	...	...	...	...	...
248227	2021-09-03	732.250000	734.000000	724.200012	733.570007	733.570007	15246100	TSLA
248228	2021-09-07	740.000000	760.200012	739.260010	752.919983	752.919983	20039800	TSLA
248229	2021-09-08	761.580017	764.450012	740.770020	753.869995	753.869995	18793000	TSLA
248230	2021-09-09	753.409973	762.099976	751.630005	754.859985	754.859985	14077700	TSLA
248231	2021-09-10	759.599976	762.609985	734.520020	736.270020	736.270020	15114300	TSLA

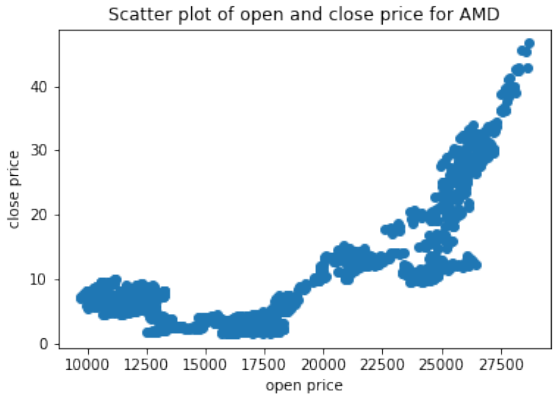
Q1b) The following are scatter charts of the close stock price of the company against the 12 listed with the next day open price of the DJIA.



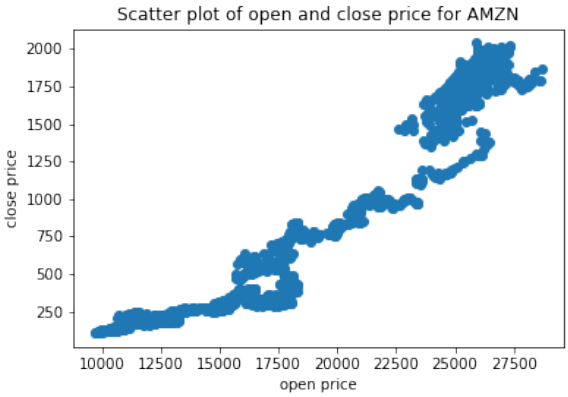
Correlation coefficient of AAPL open and close price is 0.9641974195836246



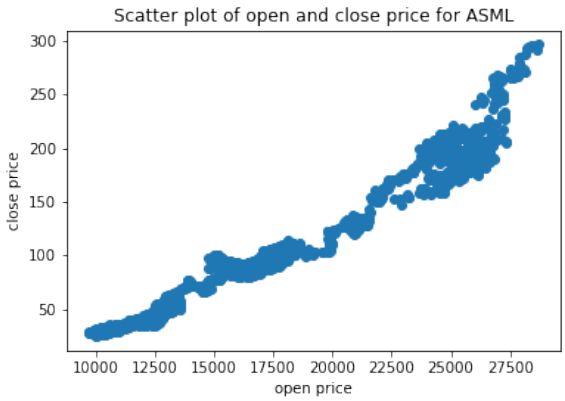
Correlation coefficient of ADBE open and close price is 0.954952937196089



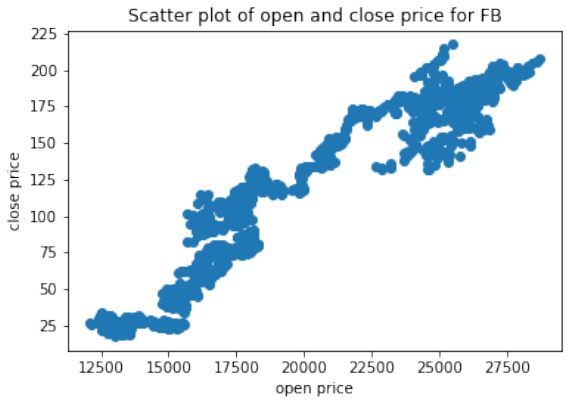
Correlation coefficient of AMD open and close price is 0.7278684915240142



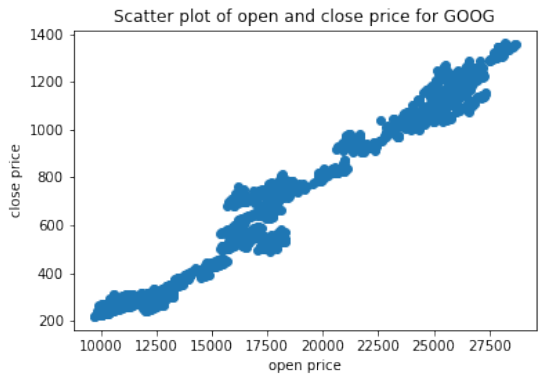
Correlation coefficient of AMZN open and close price is 0.9481892712341211



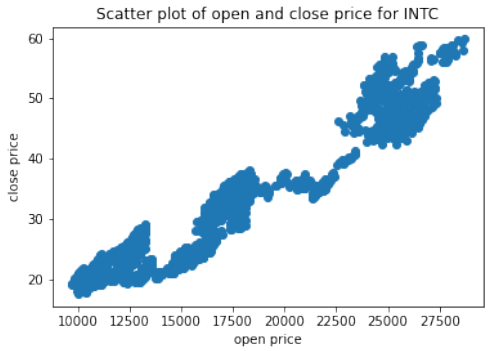
Correlation coefficient of ASML open and close price is 0.978306400377268



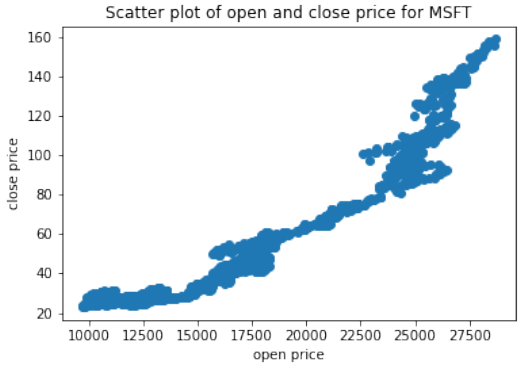
Correlation coefficient of FB open and close price is 0.9440182919212817



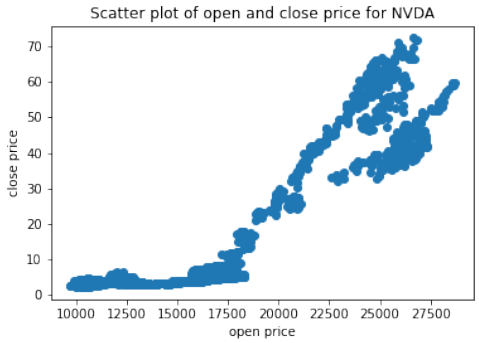
Correlation coefficient of GOOG open and close price is 0.9820904499813787



Correlation coefficient of INTC open and close price is 0.9525125409829914

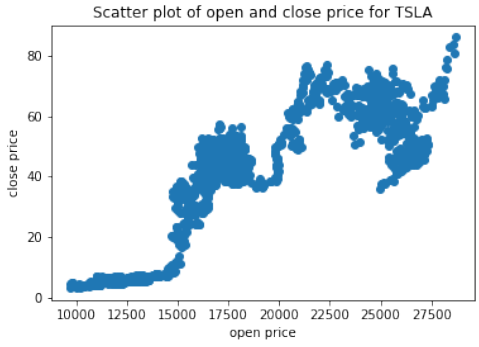


Correlation coefficient of MSFT open and close price is 0.948034855975061



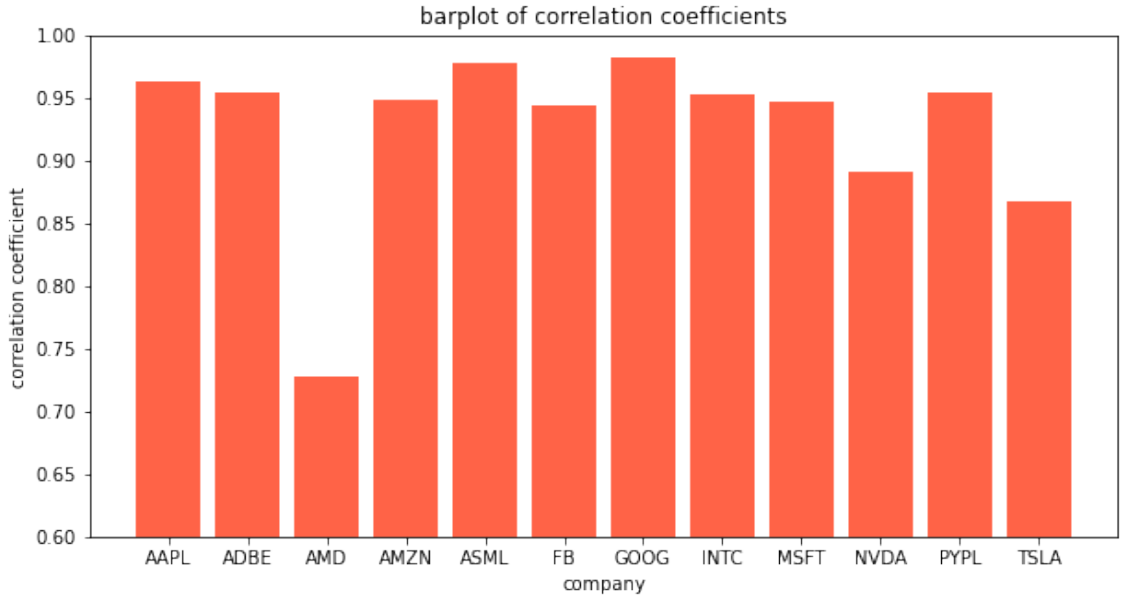
Correlation coefficient of NVDA open and close price is 0.8914887162174195

Correlation coefficient of PYPL open and close price is 0.9542218504261344



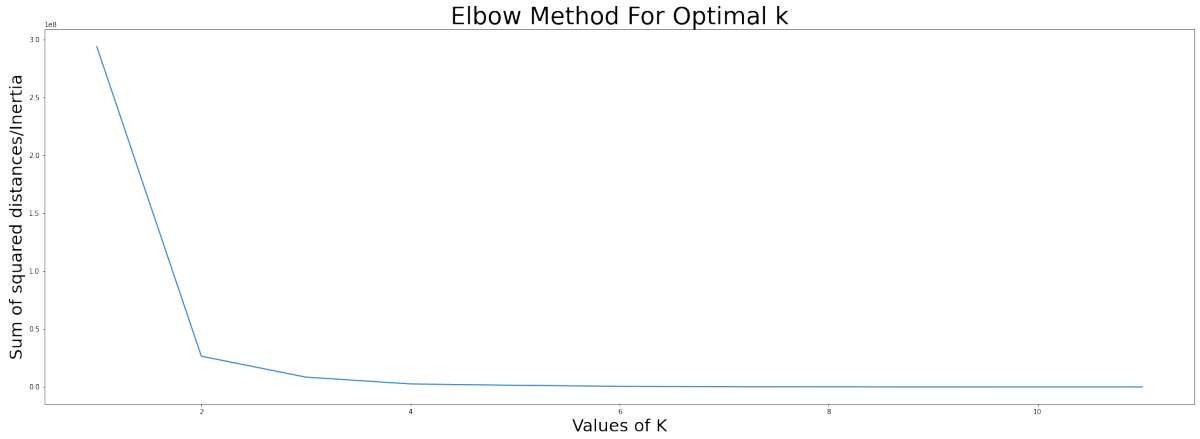
Correlation coefficient of TSLA open and close price is 0.8672492932693232

1b.1) The following is a bar graph of the correlation coefficients between the close stock price of a company amongst the 12 listed, and the next day open price of the DJIA.



Q1ci/ii) The following is the dataframe for com12month:

Date	AAPL	ADBE	AMD	AMZN	ASML	FB	GOOG	INTC	MSFT	NVDA	PYPL	TSLA
2015-07	31.333864	81.071818	1.960000	478.709095	101.088182	92.055000	590.093636	29.272273	45.611818	4.958182	37.338500	53.319909
2015-08	28.348691	81.752381	1.880952	518.464765	93.138096	91.778572	636.838097	28.366190	45.506667	5.555000	36.945714	48.910095
2015-09	28.199405	80.371904	1.796190	520.955718	89.487143	91.446191	617.934756	29.100952	43.561428	5.730833	33.593333	50.581239
2015-10	28.340000	86.005909	1.991364	566.743181	90.160000	97.129545	663.592718	32.938182	48.700909	6.765909	33.995455	44.396546
2015-11	29.540625	90.896500	2.164500	657.695499	93.130500	105.968999	735.388498	33.686000	53.885000	7.601125	36.127000	44.176400
...	...	...	...	...	...	...	...	...	...	...	...	...
2021-05	126.784000	489.344499	76.976000	3246.260010	644.137503	317.335997	2352.595496	56.151000	247.395498	146.881999	250.618501	616.753000
2021-06	129.958636	545.707729	83.331363	3367.725431	688.425462	336.425909	2501.394098	56.813636	259.018181	182.186704	275.742730	626.919550
2021-07	145.139524	609.197144	92.046667	3616.006185	713.769520	353.377145	2646.785703	55.245238	281.502385	196.463452	296.913811	659.134760
2021-08	148.177727	639.683638	108.794091	3312.917725	795.797727	363.016821	2786.406827	53.512727	294.314090	207.121819	276.529545	705.243172
2021-09	153.614286	663.192862	107.968570	3486.895717	856.241429	378.575714	2891.637102	53.624286	299.638567	224.765714	287.300005	742.567147



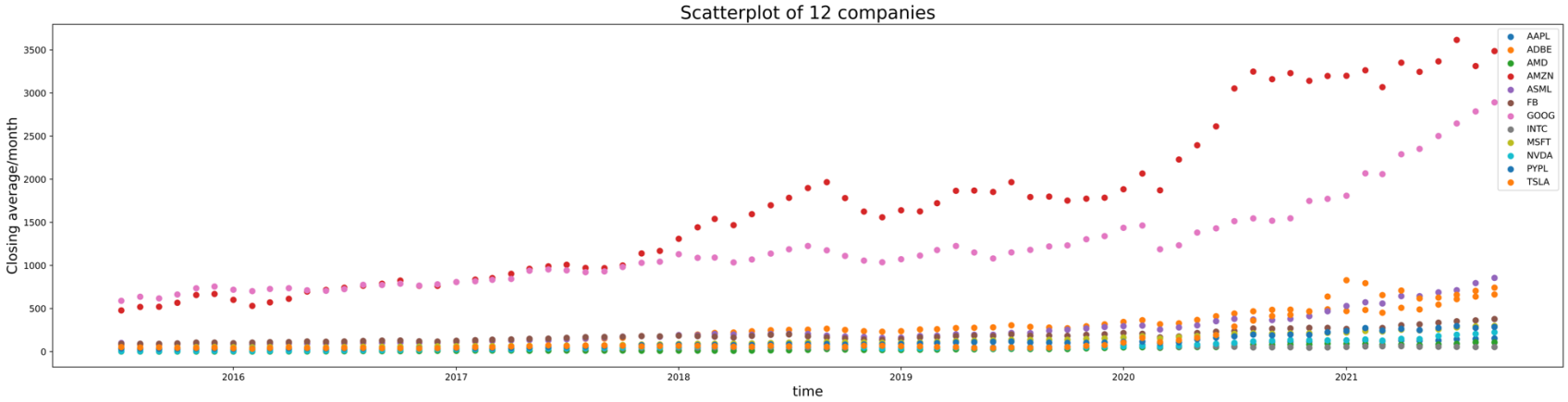
Using K means clustering, a K value of 4 is where the graph starts to taper off, hence we can set 4 to be the number of clusters using the elbow method.

From clustering the vectors together, we get this index: [0 3 0 1 3 0 2 0 0 0 0 3]
corresponding to the companies: ['AAPL', 'ADBE', 'AMD', 'AMZN', 'ASML', 'FB', 'GOOG', 'INTC', 'MSFT', 'NVDA', 'PYPL', 'TSLA'].

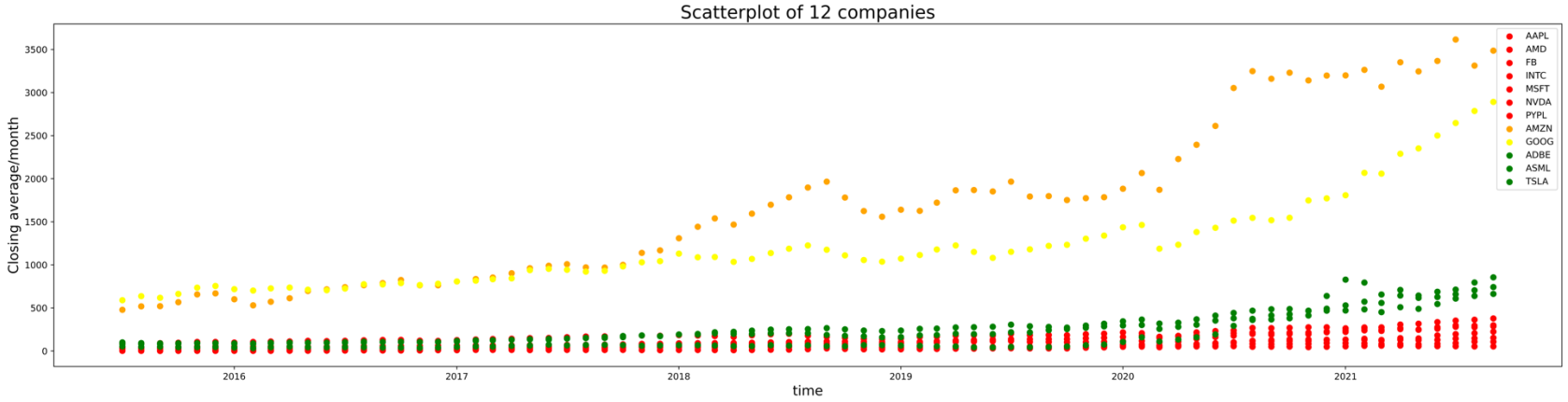
Hence, we can determine that:
AAPL AMD FB INTC MSFT NVDA PYPL belong to one cluster 1
AMZN belongs to one cluster 2
GOOG belongs to one cluster and 3
ADBE ASML TSLA belong to one cluster 4

(labelled cluster 1 through 4 for simplicity sake)

1cii1) The following is a scatter chart of the 12 companies price against time before clustering



1cii2) he following is a scatter chart of the 12 companies price against time after performing k means clustering



Q2a/b)

The following is the Nas2020 dataframe:

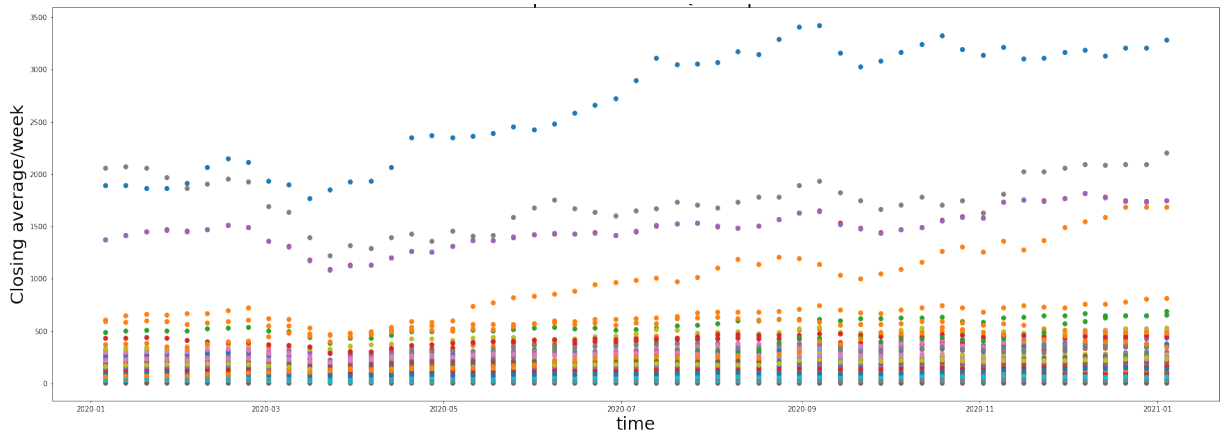
	Date	Close	Name
2516	2020-01-02	75.087502	AAPL
2517	2020-01-03	74.357498	AAPL
2518	2020-01-06	74.949997	AAPL
2519	2020-01-07	74.597504	AAPL
2520	2020-01-08	75.797501	AAPL
...	...	...	...
271501	2020-12-24	375.170013	ZM
271502	2020-12-28	351.390015	ZM
271503	2020-12-29	353.750000	ZM
271504	2020-12-30	353.399994	ZM
271505	2020-12-31	337.320007	ZM

The following is the Nas2020week dataframe:

Date	AAPL	ADBE	ADI	ADP	ADSK	...	WBA	WDAY	XEL	XLNX	ZM
2020-01-06	74.798332	333.316661	118.553332	170.156672	186.633331	...	59.250001	168.463338	62.550001	99.399999	68.773333
2020-01-13	76.925000	339.430005	119.706001	170.500003	190.839999	...	55.616001	178.634000	62.682000	99.714000	72.838000
2020-01-20	78.624374	345.672501	118.910000	173.947498	191.937500	...	54.469999	180.749996	64.585001	100.672501	75.442501
2020-01-27	79.038002	350.017999	118.073999	177.062003	196.845999	...	52.971999	183.826004	66.642000	100.630002	74.036002
2020-02-03	79.203500	354.434003	111.932001	173.935999	198.974002	...	51.762000	186.328000	68.408000	88.680000	76.718002
...	...	...	...	...	...	...	...	...	...	...	...
2020-12-07	122.948001	484.588000	141.764001	173.539999	278.451996	...	41.392000	224.951996	67.352000	146.696002	410.114001
2020-12-14	122.717999	483.644000	142.534000	173.409998	282.594000	...	41.684000	221.109998	65.364000	144.862000	397.663995
2020-12-21	127.855998	493.848004	143.557999	177.082001	296.586005	...	40.802000	234.872000	66.021999	150.038000	404.006000
2020-12-28	132.875004	499.972504	144.162498	176.174999	301.269997	...	39.605000	249.287495	64.772499	142.415005	379.860008
2021-01-04	133.760000	499.893331	145.856664	174.959997	300.953328	...	39.543334	238.086665	65.913333	141.256668	348.156667

Q2c)

Scatterplot of NASDAQ company's average closing price/week over the year 2020

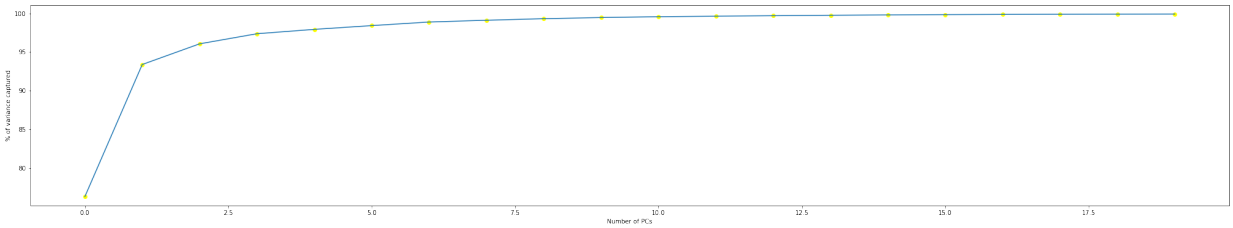


After normalization, the following PCA data chart was derived, ordered by decreasing significance:

	0	1	2	3	4	...	97	98	99	100	101
0	-0.002238	0.026411	-0.008531	-0.024216	-0.003500	...	0.013466	0.013466	0.008300	0.002197	0.0
1	-0.043831	0.005351	0.052460	-0.085327	0.088759	...	-0.000043	-0.000043	0.003496	-0.001512	0.0
2	-0.031586	0.002673	0.023375	0.015342	0.011699	...	0.001551	0.001551	-0.001201	-0.000003	0.0
3	-0.067985	-0.008135	0.048130	-0.025298	-0.018830	...	-0.023208	-0.023208	0.018948	0.007650	0.0
4	-0.030941	0.039542	0.036241	0.069458	0.035812	...	0.055950	0.055950	0.029382	0.000361	0.0
...	...	...	...	...	...	...	...	...	...	...	...
97	-0.026722	0.028028	-0.006700	0.023052	0.016647	...	0.203324	0.203324	-0.060807	0.061420	0.0
98	-0.065364	0.114456	0.083764	0.014382	-0.216975	...	-0.016334	-0.016334	-0.016359	0.009392	0.0
99	-0.004411	-0.004101	-0.001766	0.015328	-0.001789	...	0.158033	0.158033	-0.391459	-0.245252	0.0
100	-0.031934	0.036526	0.033170	0.055831	-0.027137	...	-0.029536	-0.029536	-0.016366	0.006810	0.0
101	-0.005671	-0.004381	0.005473	0.001536	-0.005198	...	0.418922	0.418922	0.099448	0.201026	0.0

Q2d) To reduce the dimension of the feature vectors, we wish to project the data vectors on the fewest principal components while capturing the most variance within the data. I use a similar method to the K-Means clustering technique by plotting percentage of total variance covered over principal components, to derive the fewest principal components that capture the most variance.

Graph of total percentage of variance captured over principal components. (capped at 20)



From graphical representation of the percentage of total variation captured over number of principal components, the variation seems to be stable at the 11th principal component, therefore we will only take the top 11 principal components.

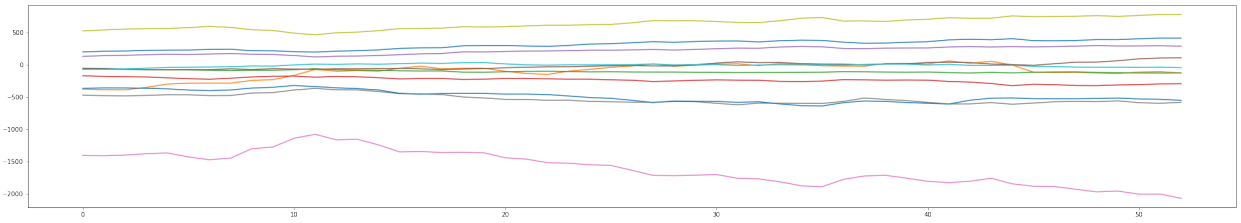
Using the top 11 principal component results in the following dimensionality reduced principal components:

	0	1	2	3	4	...	97	98	99	100	101
0	-0.002238	0.026411	-0.008531	-0.024216	-0.003500	...	0.013466	0.013466	0.008300	0.002197	0.0
1	-0.043831	0.005351	0.052460	-0.085327	0.088759	...	-0.000043	-0.000043	0.003496	-0.001512	0.0
2	-0.031586	0.002673	0.023375	0.015342	0.011699	...	0.001551	0.001551	-0.001201	-0.000003	0.0
3	-0.067985	-0.008135	0.048130	-0.025298	-0.018830	...	-0.023208	-0.023208	0.018948	0.007650	0.0
4	-0.030941	0.039542	0.036241	0.069458	0.035812	...	0.055950	0.055950	0.029382	0.000361	0.0
5	-0.038135	-0.024150	0.027625	-0.046023	0.002083	...	-0.001929	-0.001929	0.000196	-0.005002	0.0
6	-0.035323	0.239101	0.013512	-0.048386	0.054540	...	0.006129	0.006129	0.009135	0.003879	0.0
7	-0.015506	0.015381	0.014846	0.015075	-0.031226	...	0.143266	0.143266	0.084738	0.097287	0.0
8	0.002664	0.020863	0.001411	-0.034548	-0.040915	...	0.088946	0.088946	0.010238	0.077831	0.0
9	-0.063145	-0.079541	0.007265	-0.005227	-0.022309	...	-0.002619	-0.002619	0.029842	-0.006062	0.0
10	0.058927	-0.011896	0.364636	-0.109857	-0.185696	...	-0.008982	-0.008982	0.002657	-0.002712	0.0

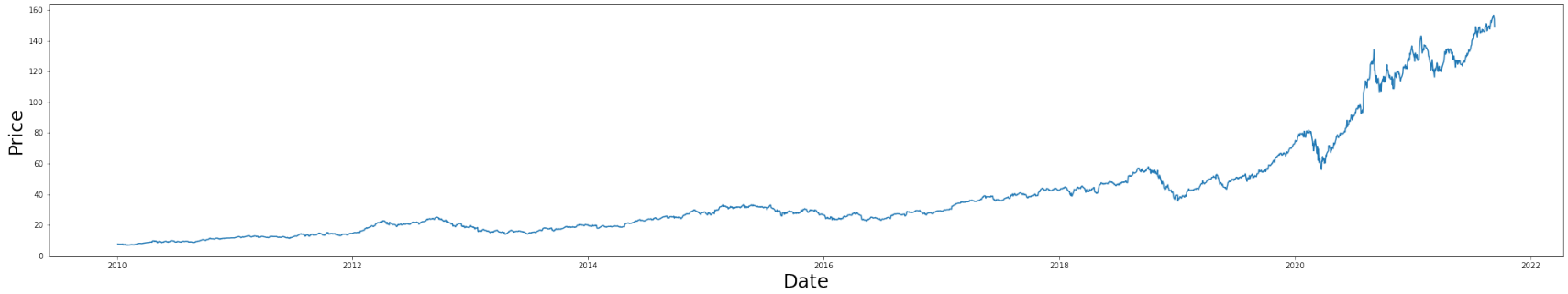
2.1) From projecting the data vectors onto the principal components, the following finaldataset table was created (columns represent weeks, rows represent reduced dimension data points):

	0	1	2	3	4	...	6	7	8	9	10
0	159.463133	-322.418443	127.078819	-371.138476	-99.763721	...	-1310.297072	500.669071	507.287379	-119.712510	-402.277454
1	162.894754	-326.886166	131.356393	-375.753950	-92.749448	...	-1312.066969	519.243056	524.739301	-121.020345	-398.259572
2	165.866302	-329.826004	131.175958	-382.377913	-87.570945	...	-1298.679228	532.463934	537.195475	-112.249004	-399.864892
3	175.916806	-291.982491	127.628514	-384.754587	-80.962688	...	-1277.530927	521.615285	547.134181	-103.804290	-405.579179
4	179.723209	-240.446209	121.670700	-387.997084	-69.243983	...	-1267.000937	515.412645	545.548797	-86.827225	-415.693835
...	...	...	...	...	...	...	...	...	...	...	...
48	263.154784	7.546757	157.212641	-549.133662	-24.496005	...	-1832.316392	740.664151	751.634870	-131.207766	-629.507870
49	256.715250	0.502230	157.958590	-537.707718	-34.798165	...	-1817.265390	743.433615	743.047981	-136.274227	-624.196431
50	263.417080	29.323542	159.944924	-534.991939	-43.488295	...	-1864.582519	775.009097	744.917329	-137.282166	-641.684247
51	263.497797	36.296308	163.207107	-536.244192	-42.815078	...	-1865.213172	790.592650	748.022819	-133.029900	-643.447543
52	259.420291	12.863469	171.298245	-536.689976	-45.504946	...	-1923.499925	790.113220	750.763778	-143.221801	-662.224681

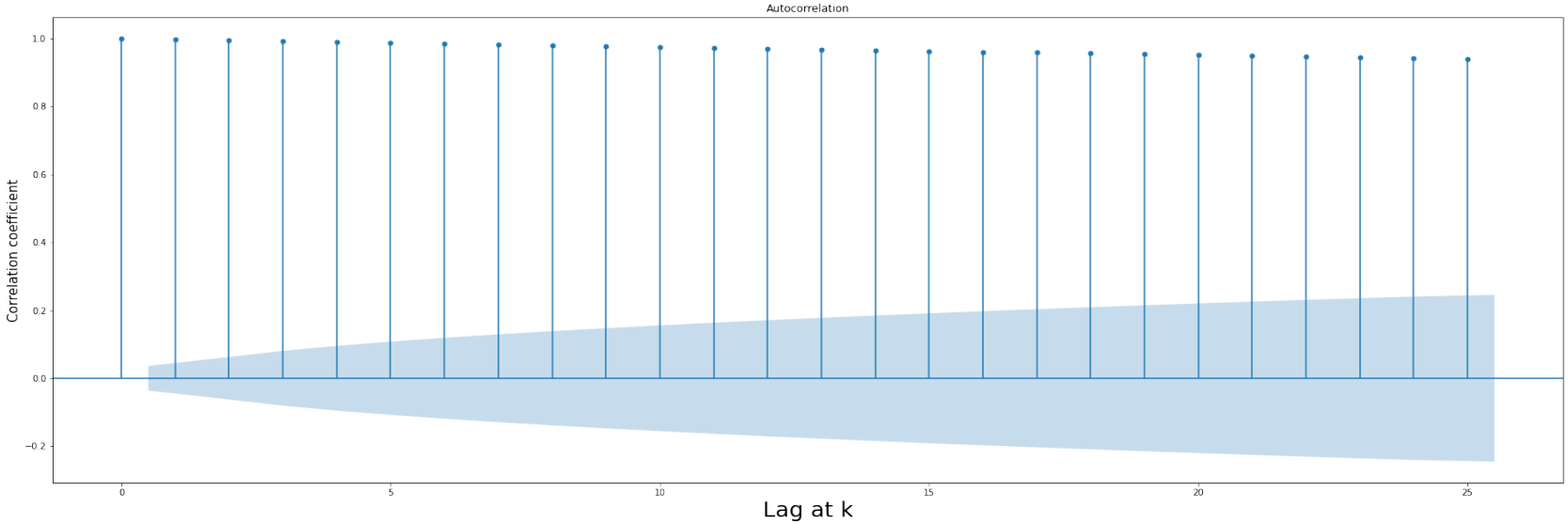
Dimension reduced close prices of NASDAQ companies over time



Q3) Graph of APPL close stock prices over time



3.1) Autocorrelation graph of APPL close stock prices over time

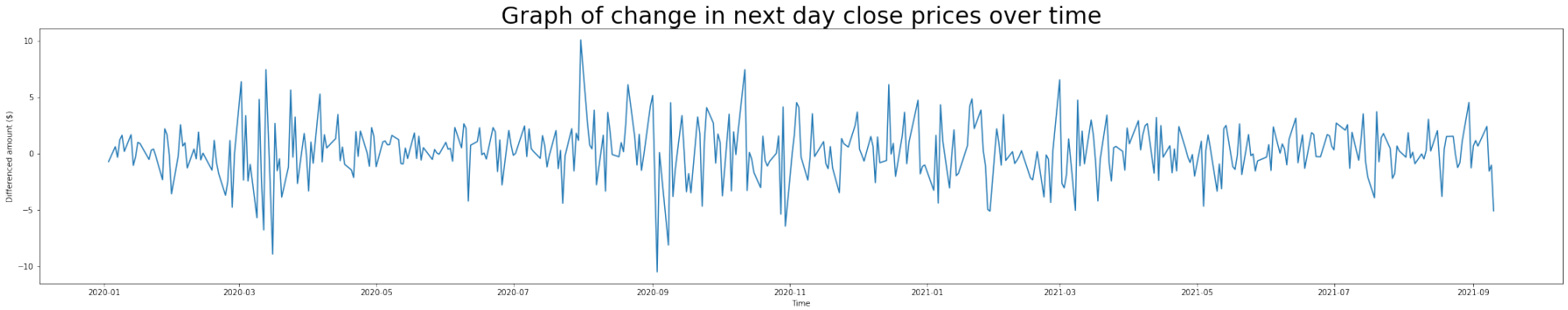


According to the graph of AAPL stock prices over time, the year 2020 seems to be the year where new trends start to arise. Furthermore, 2020 was when coronavirus pandemic started (Coronavirus disease (COVID-19) update, 2020), ushering a new era in the stock market. Thus, I only take values from past 2020 to study the Fourier trends that are relatable to today's context.

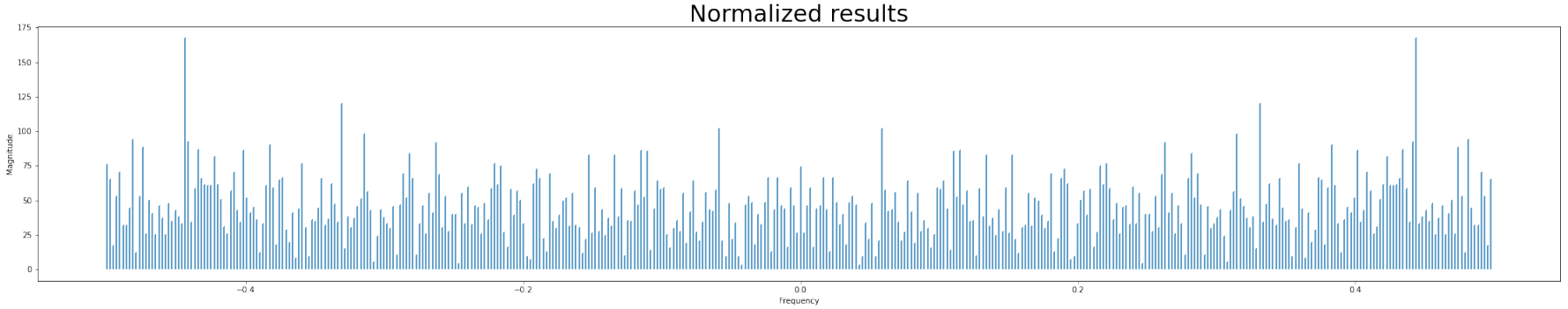
I choose to run the Fourier analysis on the difference between the current day value and its previous day value to gather a feature that determines how much fluctuation a closing stock price has over time.

Setting 0.95 as an accepted autocorrelation value, looking at the autocorrelation graph, we take an approximate k value of 25. The correlation between values that are 25 days apart are still high with a coefficient of 0.95 thus we set the window size of 25 as we wish to set a window size large enough to capture as much relevant information as possible that we can extrapolate it to the next day's potential change in frequency price and therefore determine the following day's close price, which will be done in 3.3.

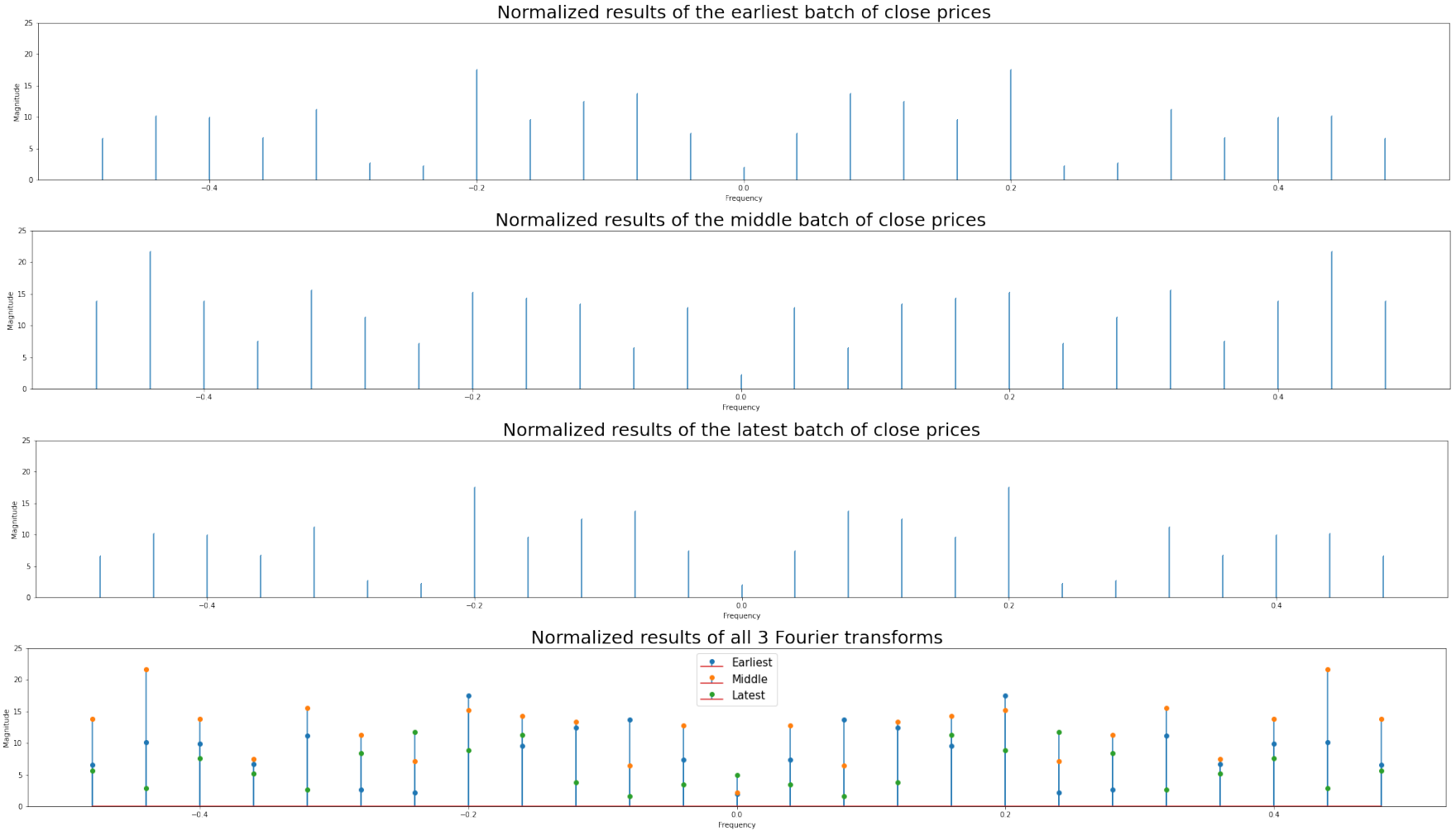
3.2) We take the overall Fourier Transform from 2020 onwards to explore what the overall frequencies of change in prices look like. according to the graph of change in next day price over time, we can infer that the change in price changes vary over time.



Graph of Fourier transform over the time period from 2020 onwards shows that there are varying frequencies of price changes shows that the frequencies are fairly spread out with exception to the anomalous peak at frequency value of 0.444.



3.3) To test if the frequency of price change is consistent throughout the time span of later than 2020, I tabulated the Fourier transform of 3 time periods of 25 days from 2020 to now, with no overlap, to see if the Fourier charts differ over the new season that is post 2020. Taking the start of 2020, the middle point, and the latest days in the dataset, the following charts were derived:



The normalized result of all 3 Fourier transforms show that the frequency trends change with time. The frequency of change in prices can help us forecast future prices. Thus, Fourier transform can be used as a feature in Q4.

Q4)

In order to develop a machine learning model to predict the next day's close price, we have to derive a set of relevant features through manual feature selection. I will use a three-step method to preparing the features necessary for the predictive machine learning model.

First, I will specify how features should be used for the machine learning model.

Secondly, I will decide how to pre-treat the data.

Finally, I will validate that the features should be included in the machine learning model through checking if it fulfills 3 requirements (Vickery, 2020),

If the feature doesn't hit at least one of the three requirements, we will reject the feature from the model.

Requirement 1 (Common correlation)	Ensure that the feature being chosen is not correlated with another feature in the set, to avoid two features capturing the same information
Requirement 2 (Variation)	Ensure that the feature has sufficient variation to prevent them from providing no information to the model
Requirement 3 (Relationship)	Ensure that a feature has a relationship with the target variable (next day's close price in this case)

4.1 Machine learning specification)

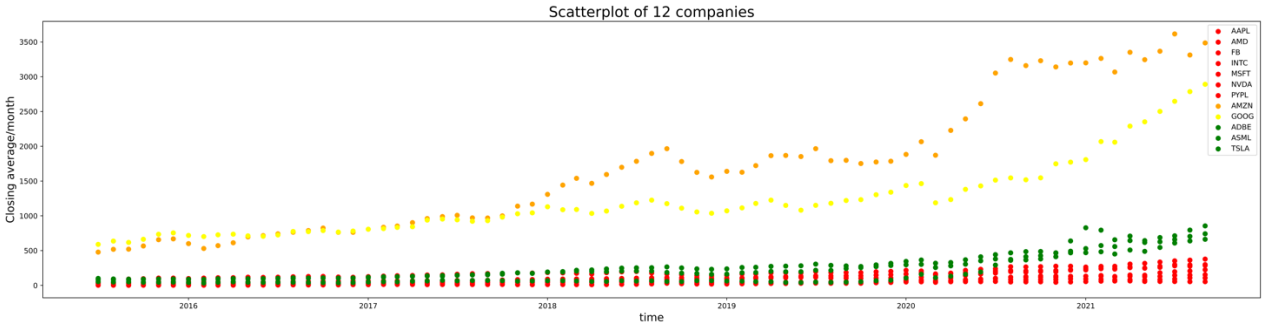
To determine the next day close price of a company given the previous history, I use these features.

Fourier trends	I will first run a K-Means clustering together with the 12 companies to determine which cluster that company belongs to. Since the companies within that cluster should have similar attributes, I will input Fourier transforms of all the companies latest frequency trends in a 25-day window, as per stated in 3.1. I will then average out the Fourier transform trends amongst all the companies within that cluster.
Dimension reduced overall stock trends	Assuming that the NASDAQ trends describe general market trends, I will take PCA from 2.1 to derive the general market trends to predict for market changing events.

Taking TSLA stocks as an example, I will generate the relevant features:

Fourier trends:

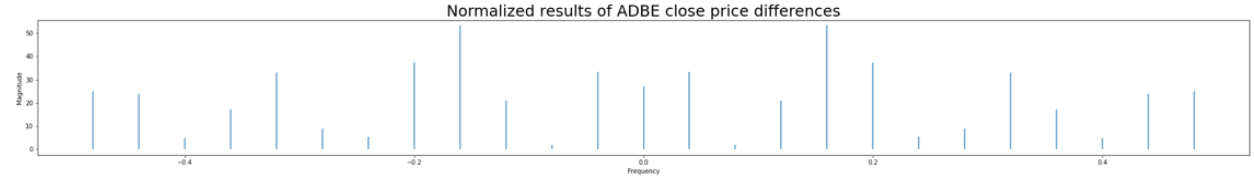
Taking K-means:



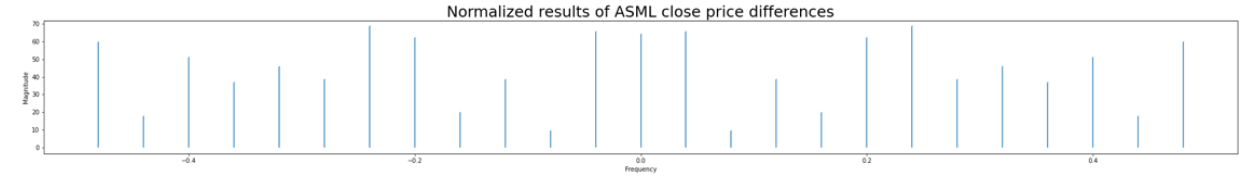
Fourier averaging

Since ADBE ASML TSLA belong to one cluster, we tabulate the mean Fourier transform of all 3 stocks in the past 25 days by summing up the individual Fourier vectors and dividing it by 3.

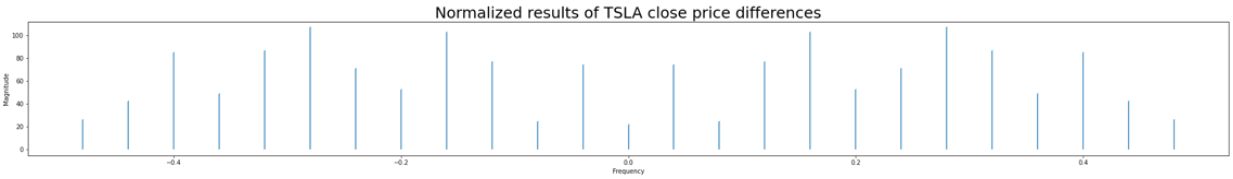
ADBE Fourier Transform:



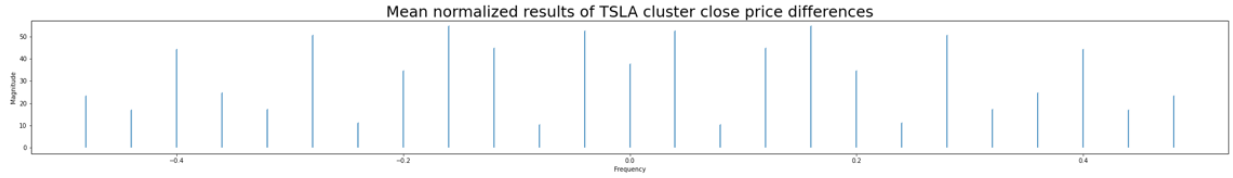
ASML Fourier Transform:



TSLA Fourier Transform:



Average Fourier Transform:

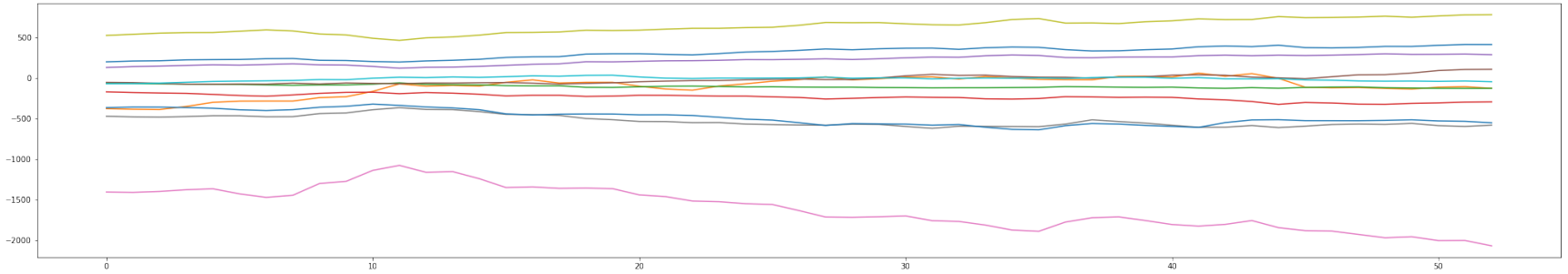


Dimension reduced overall stock trends:

Dimension reduced data for overall market trends across the year.

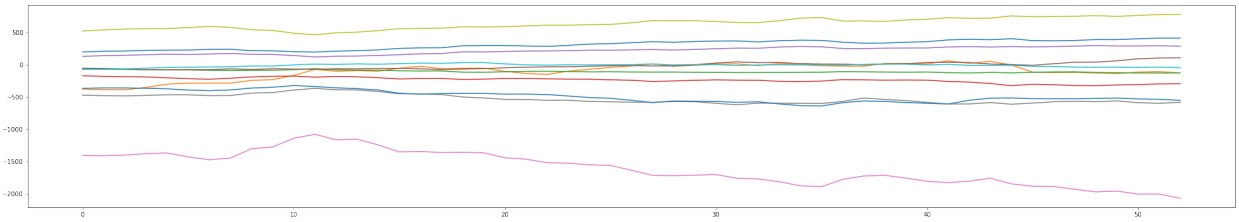
	0	1	2	3	4	...	48	49	50	51	52
0	200.962428	211.097388	214.437432	225.342647	228.149062	...	391.193381	390.589793	404.798411	414.763025	414.276659
1	-375.988210	-382.746235	-386.267382	-348.504536	-297.702524	...	-125.625621	-134.101960	-111.534549	-105.104493	-125.358712
2	-65.000553	-67.749537	-70.067949	-72.944682	-75.337536	...	-117.345652	-121.082053	-125.035743	-126.011501	-125.247157
3	-168.684523	-177.290009	-182.818240	-187.172154	-199.875709	...	-321.827907	-310.999079	-304.728058	-294.813150	-291.998736
4	131.750606	142.920801	148.887400	156.703385	163.944882	...	299.078018	292.356880	293.085388	295.993999	288.978916
5	-51.888911	-54.661888	-63.860034	-73.397492	-73.882041	...	43.130400	63.458435	93.705109	106.762207	109.251090
6	-1403.187968	-1407.410001	-1395.684922	-1374.752688	-1362.968634	...	-1966.277212	-1953.609756	-2000.807100	-1999.502204	-2065.530831
7	-469.422865	-476.310378	-479.725306	-472.758743	-462.818851	...	-569.765561	-557.058849	-584.956089	-594.598961	-579.230983
8	526.298919	540.625727	554.381286	561.271749	562.015519	...	763.356013	750.708336	766.000159	779.401482	782.305700
9	-68.790607	-67.148061	-60.875776	-50.525486	-40.630239	...	-37.128395	-35.478757	-39.766488	-34.736123	-44.045182
10	-363.125390	-356.510601	-357.122704	-361.927252	-370.672914	...	-520.398826	-513.381805	-527.392152	-531.906007	-550.169817

Graph of dimension reduced overall market trends across the year

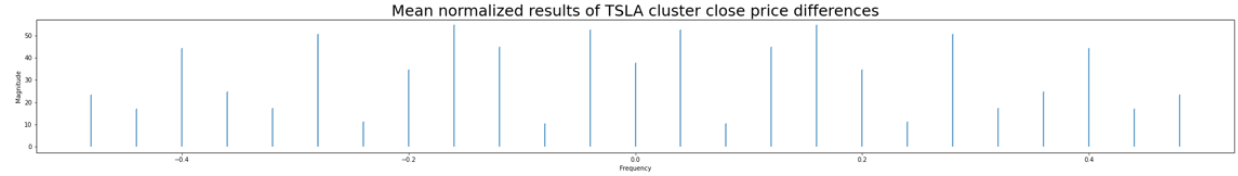


4.2 Pre-treating data)
Features before normalized:

Dimension reduced close prices over time:



Mean frequency of close prices in the past 25-days:

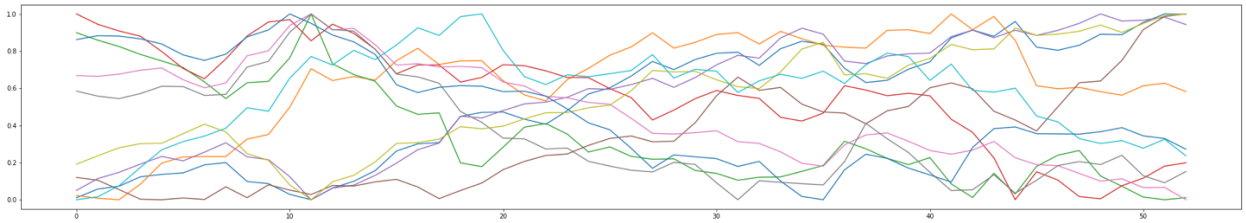


Features after normalized:

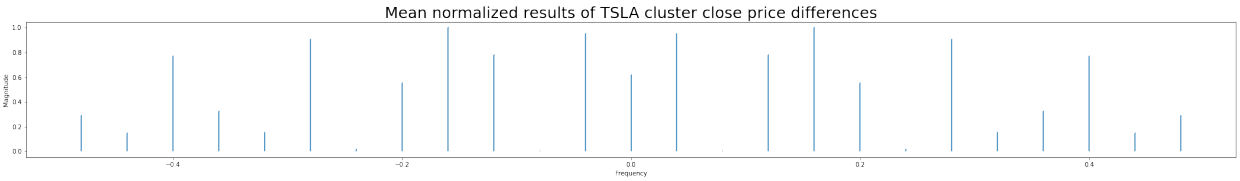
After which, I will normalize both features using minmax. I will not standardize as the above features do not display a normal distribution as seen form above.

Then I will run the features through the predictive machine learning model.

Normalized dimension reduced close prices over time:



Normalized mean frequency of close prices in the past 25-days:

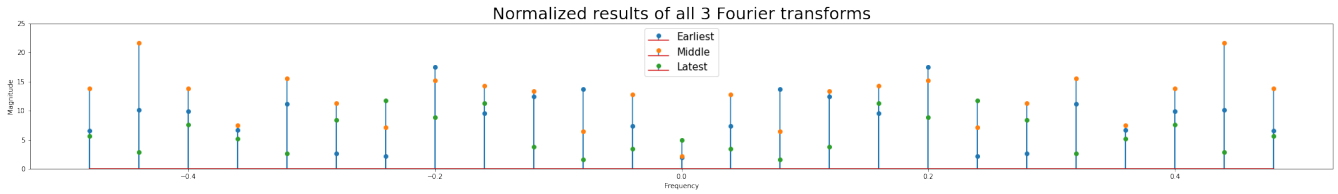


4.3 Feature validation)

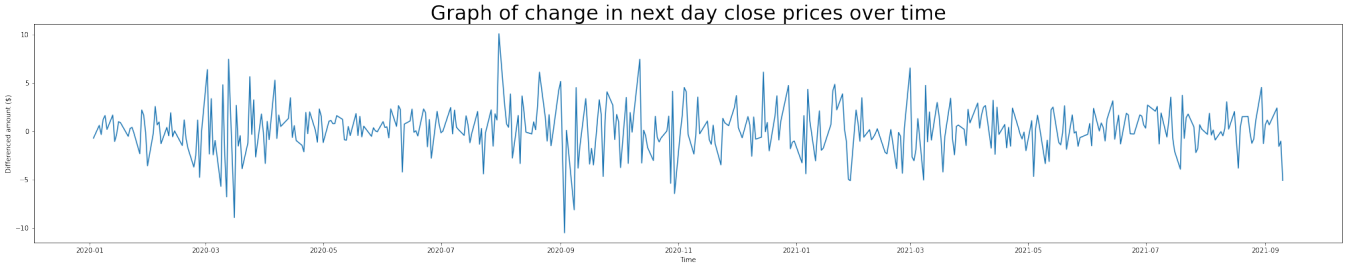
1. Fourier Feature

Common correlation	This is our first feature.
Variation	According to 3.3, frequencies of Fourier change over time, hence proving variation. According to 3.2, there is a change in frequency over time, further proving variation.

3.3



3.2

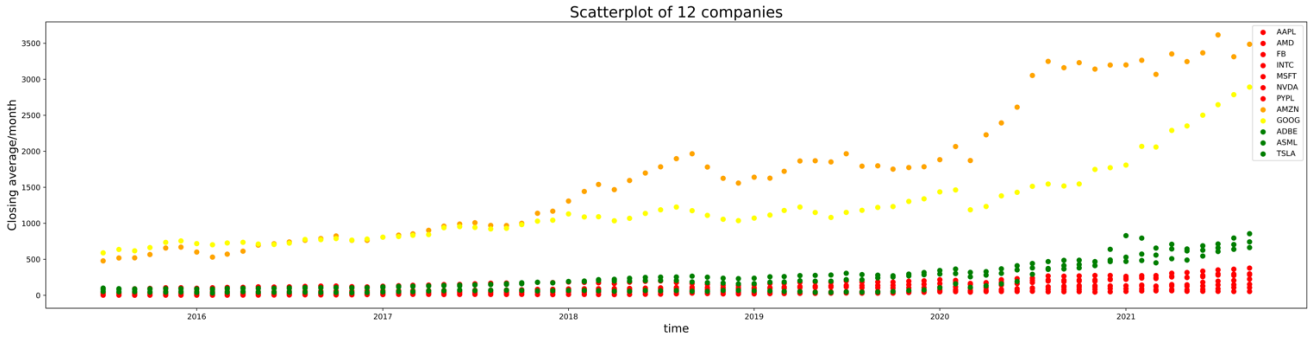


Relationship	Momentum and the frequency of a change in close price affects the next day price. (yates, 2021)Thus, it can be extrapolated to show us the possibilities of price changes in the future and therefore the probable close price.

2. K-Means clusters

Common correlation	K-Means clusters will be used to decide which category of companies the company we are looking to predict will fall under, to provide more samples for the current company to be studied under. KMeans is used to define clusters and thus do not have a correlation to the prior features explored.
Variation	According to 1cii2 different company vectors belong to different clusters with different centroids, thus proving variance.

1cii2



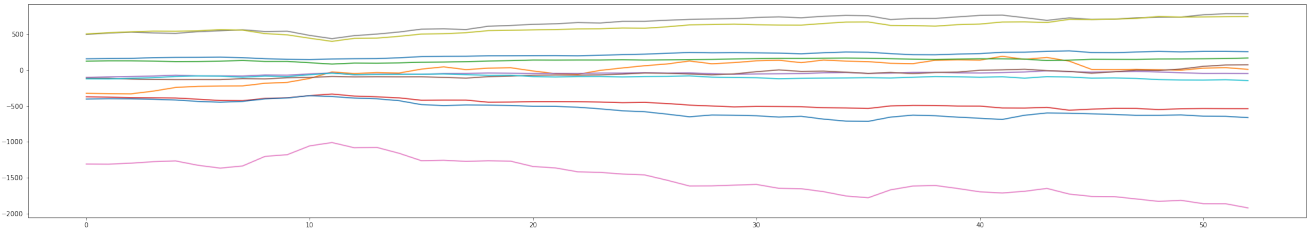
Relationship	Each company belongs to a category, with its price moving in relative tangent with the other companies similar to itself. Hence when we study all the companies within the same category (cluster) it allows us to predict with a higher degree of accuracy the close price of the following day.

3. PCA

Common correlation	PCA reduces the variation from anomalous data of different companies to capture overall market trends and is run on all companies in the NASDAQ data set whereas the other features are run only on the companies within their cluster group. This is not correlated to any other feature in our model.
Variation	According to the following graph, the prices of the reduced dimensions change over time. Hence there is variation in this feature.

Plot of dimension reduced companies prices over time:

4.4



Relationship	The dimensionality reduced close prices over time provide information on the overall market trends. When extrapolated, this can give us information on what the overall market close prices will be the following day, therefore giving us a gauge on what the close price of a next day's specific company.

Bibliography

Coronavirus disease (COVID-19) update. (2020). Retrieved from WHO: https://www.who.int/bangladesh/emergencies/coronavirus-disease-(covid-19)-update#:~:text=On%20this%20website%20you%20can,on%2031%20December%202019.

Vickery, R. (2020, Apr 29). *The Art of Finding the Best Features for Machine Learning*. Retrieved from Medium: https://towardsdatascience.com/the-art-of-finding-the-best-features-for-machine-learning-a9074e2ca60d

yates, t. (2021, June 12). *Investopedia*. Retrieved from 4 Ways to Predict Market Performance: https://www.investopedia.com/articles/07/mean_reversion_martingale.asp