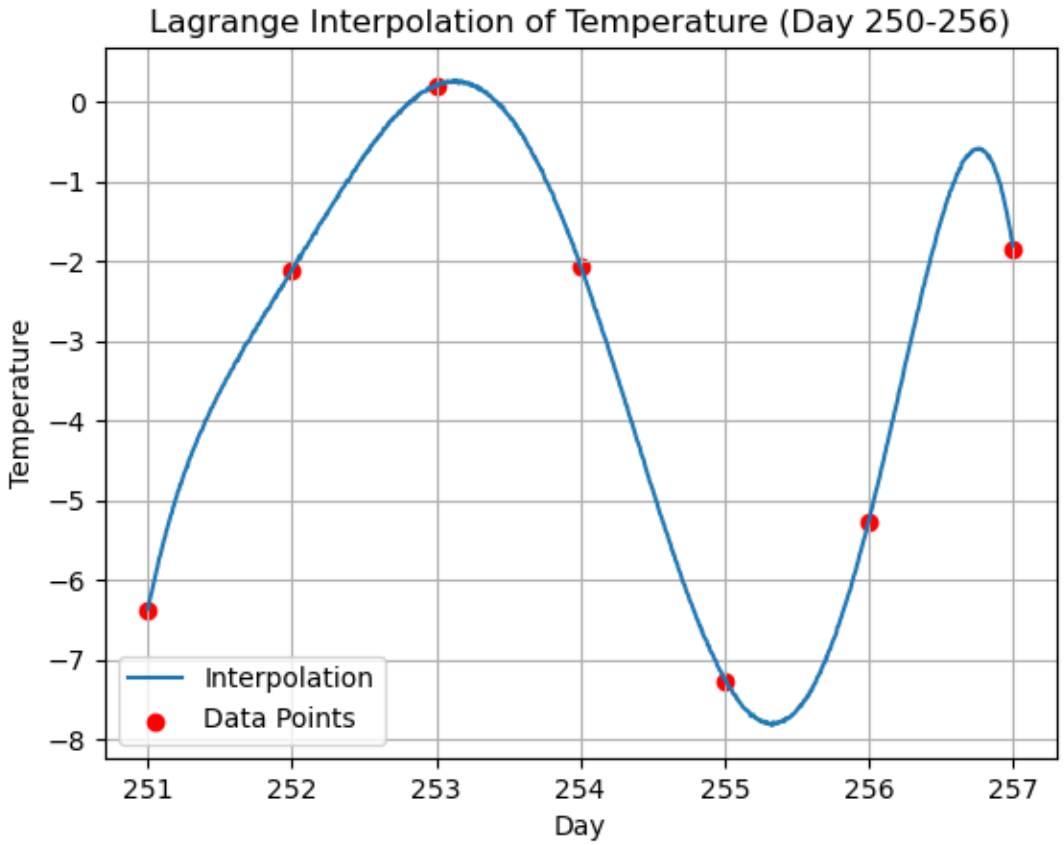
Mathematical Modeling Course Project: Interpolation and Approximation Techniques

Benjamin Nicholson
Seton Hill University



Introduction:
Interpolation is a common statistical technique to better understand the relationship between two variables. Interpolation looks to 'fill in' or complete graphs by estimating expected values and then including them a part of the data. This is different to extrapolation which looks to forecast or predict further data based off the existing data. Interpolation is a technique that is commonly used in Data Science to fill in missing data. For example, if you have existing data between the variable 'Temperature' and 'Ice Cream Sales' you might be missing the number of ice cream sales based on a particular temperature. Interpolation is going to estimate the missing ice cream sales at a particular temperature based off the data that explains the relationship between temperature and ice cream sales.

There are many ways of interpolating data, using many different techniques. In this project we used the following

- Lagrange Interpolation
- Newton's Divided Difference
- Chebyshev Distance
- Cubic Splines

Each one of these techniques can be used on the same data, however understanding how each technique is able to give different insights into the data can help determine which interpolation technique to use.

Lagrange Interpolation
Lagrange interpolation technique creates a polynomial of n-1 to fit the data. The polynomial will go through each of the original data point, and you are able to find any data in between the minimum and maximum values of the dataset. However, as you increase the number of data points the complexity of the interpolation polynomial is going to increase which results in Runge's Phenomenon. Runge's Phenomenon refers to the extreme oscillations that occurs as the degree of the interpolation polynomial increases. From what was observed in the data, when n exceeded 7 there were points in the polynomial that would be far too large and would not appropriately represent the expected behaviour of the data. This is a negative of Lagrange' interpolation technique as the degree of polynomials increases it becomes increasingly difficult to find the expected values.

The following equation is the mathematical equation for finding the Lagrange Interpolation Polynomial.

$$\frac{(x-x_2)(x-x_3)\cdots(x-x_n)}{(x_1-x_2)(x_1-x_3)\cdots(x_1-x_n)}$$

$$y_1+\frac{(x-x_1)(x-x_3)\cdots(x-x_n)}{(x_2-x_1)(x_2-x_3)\cdots(x_2-x_n)}y_2+$$

$$\cdots+\frac{(x-x_1)(x-x_2)\cdots(x-x_{n-1})}{(x_n-x_1)(x_n-x_2)\cdots(x_n-x_{n-1})}y_n.$$

https://mathworld.wolfram.com/LagrangeInterpolatingPolynomial.html

Newton's Divided Difference

Newton's Divided Difference (NDD) is another interpolation technique which takes a recursive approach to finding the interpolation polynomial. The recursive method is explained by the mathematical equation below. Essentially you must go through each layer of differencing before getting to the final value. NDD usually takes less computation that other techniques as it uses simple arithmetic to solve for the coefficients of the polynomial. NDD usually out performs other interpolation techniques when there is an uneven spread of data which was observed in the Stock Data in the experiment. It also constructs a polynomial of degree n-1 and it does suffer from the same limitation of Runge's Phenomenon where the increase in values that are used in the technique results in a far more complex polynomial.

The following is the mathematical equation for finding the Newton's Divided Different Interpolation Polynomial.

xi	fi	F(xi,xj)	F(xi,xj,xk)
x1	f1		
		$f[x1,x2]=\frac{f2-f1}{x2-x1}$	
x2	f2		$f[x1,x2,x3]=\frac{f[x3,x2]-f[x2-x1]}{x3-x1}$
		$f[x2,x3]=\frac{f3-f2}{x3-x2}$	
x3	f3		

https://www.geeksforgeeks.org/newtons-divided-difference-interpolation-formula/

Chebyshev Distance

Chebyshev distance takes a different approach with the interpolation techniques above. It is about finding the largest distance between points in the data. This technique is useful in that it can minimise the impacts of Runge's Phenomenon. This is achieved through strategically placing nodes to minimise the extreme oscillations that can take place when solving for interpolation polynomials.

The following is the mathematical equation for finding the Chebyshev Distance between two sets of data values.

$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

https://medium.com/@balaka2605/distances-in-machine-learning-289afbce8148

Cubic Splines

Cubic splines are very versatile. As more points are added there are different ways that the cubic spline will change. If the new points are added at the end then it creates a new cubic function and is added to the end of the piecewise function. This is very beneficial as the rest of the function can be left alone and does not need to use any computational power. If the

points are added to the middle then the piecewise function it is likely just local adjustment however if there are sparse number of points or quality conditions that need to be met by the spline then there is going to be a global impact. Some of these quality conditions will include the continuity and the smoothness which is usually dependent on the derivative of the functions when they meet. Cubic splines are best when there is going to be a changing number of points.

The following is format that the piecewise function for interpolation takes when using the Cubic Spline technique

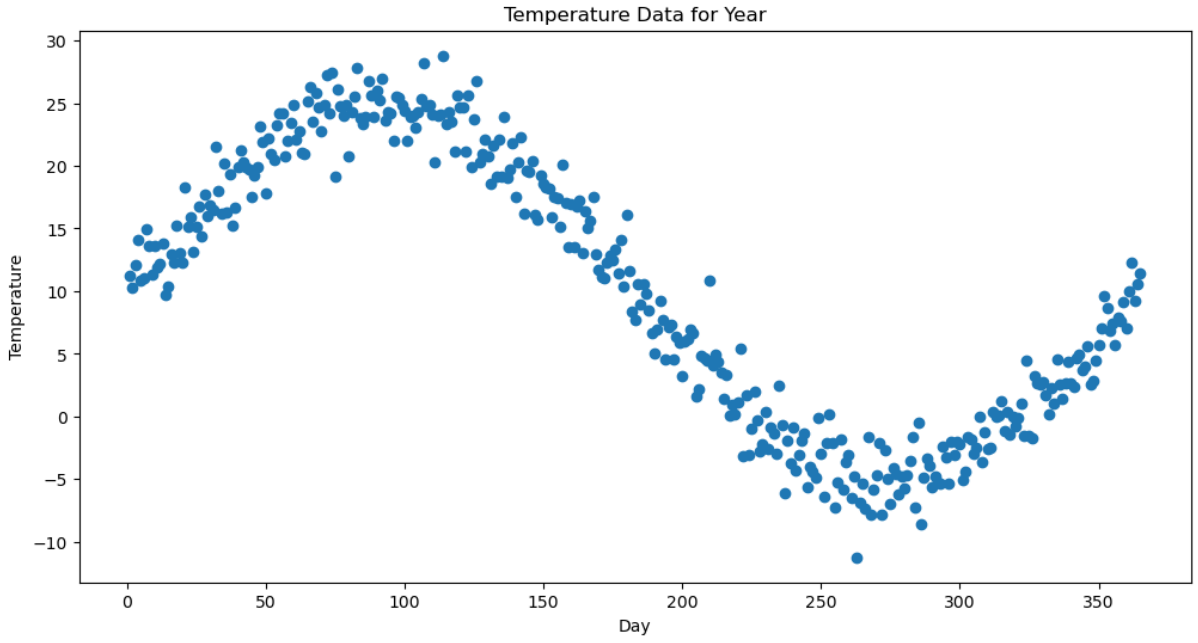
$$S(x) = \begin{cases} C_1(x), & x_0 \le x \le x_1 \\ \quad \dots & \\ C_i(x), & x_{i-1} < x \le x_i \\ \quad \dots & \\ C_n(x), & x_{n-1} < x \le x_n \end{cases}$$

https://en.wikiversity.org/wiki/Cubic_Spline_Interpolation

Problem 1: Temperature Data Interpolation
Given a set of monthly average temperature readings over several years, interpolate missing monthly data to complete the series.

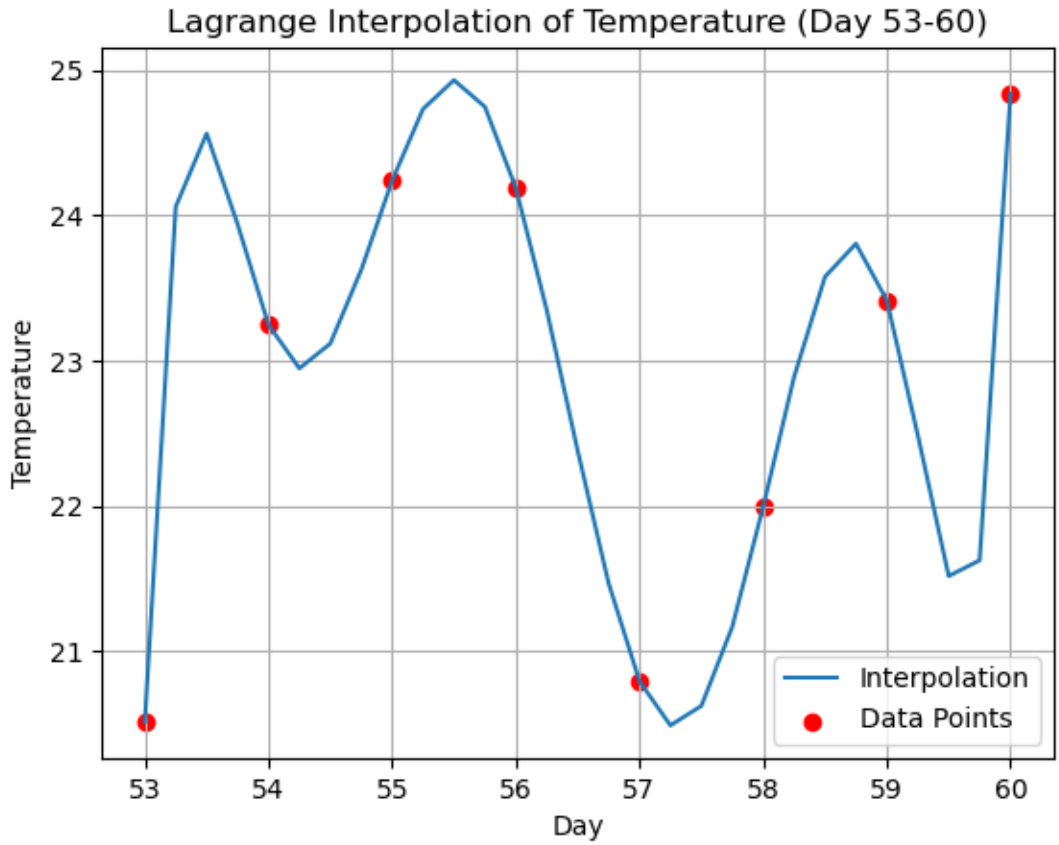
This data takes temperature for a particular calendar year. For this data, it is going to be assumed that the time of the recording of the temperature was taken in the middle of the day. The day column is going to represent how many days past the 1st day of January is. With the temperature recording being the daily maximum temperature of a certain location.

Graph 1: Temperature Data for Year



You can observe that the data follows a sinusoidal pattern. There is no missing values for days so to interpolate using half day values are going to be solved for. This would imply that you are finding the temperature at midnight.
If you were to try and interpolate every data value it would return a function to the 364th degree so I will be taking subsets of 1 week to give different examples how interpolation can take place. To interpolate the temperature data, Lagrange Interpolation technique is going to be utilised.

Graph 2: Lagrange Interpolation of Temperature

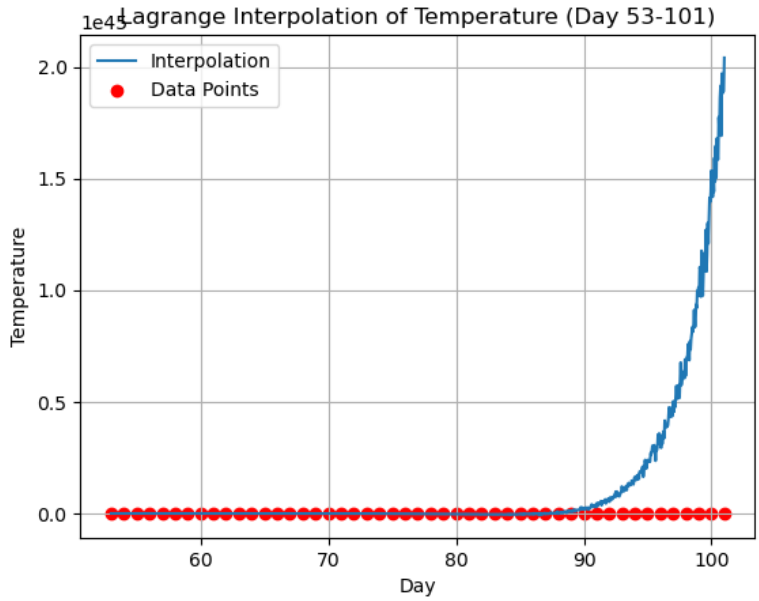


The Lagrange Interpolation technique has created a function with 29 different data points. This number is used to return values at four different points during the day. This would be midnight, 6am, midday and 6pm. The table shows times at 0.25 increments which reflects those quarterly daily intervals. That data snippet can be seen to the right and is continued until day 60.

	Time	Temperature
0	56.00	24.185226
1	56.25	20.541786
2	56.50	19.705120
3	56.75	20.088444
4	57.00	20.789677

The following is an example of where using too many data points will result in the Runge's Phenomenon

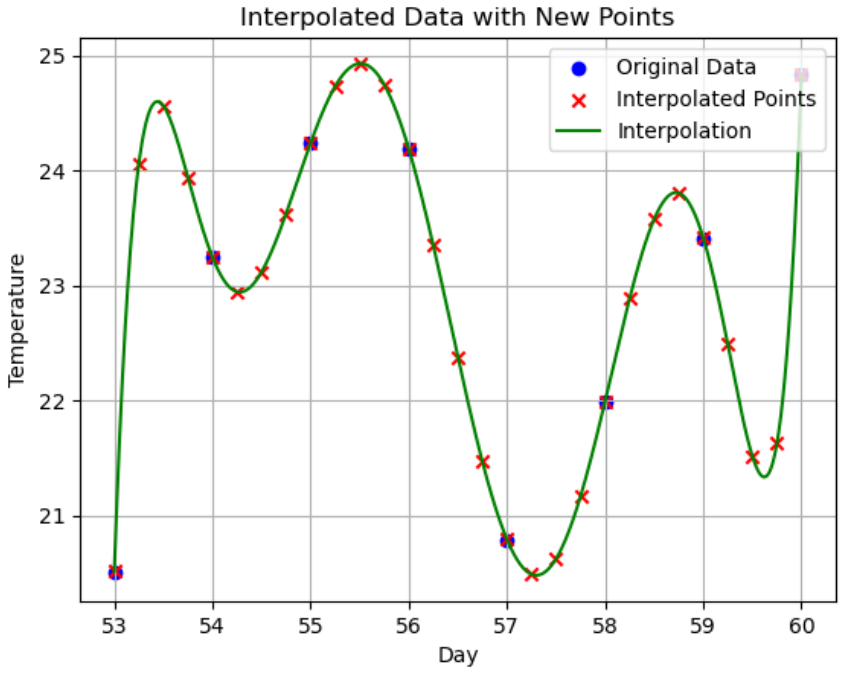
Graph 3: Runge's Phenomenon for Lagrange Interpolation



The polynomial has become very inaccurate and does not reflect the data well. You can see that the graph has a polynomial which goes up to 1*10^45 which is the result of having a polynomial to the degree of 47.

Graph 2 only had four values for the polynomial each day. However the following graph is going to plot all of those points so in order to have a more telling graph, using a polynomial which has 1000 different x values is going to make a smooth graph. The following graph plots the interpolated points that occur at 6am, 6pm and midnight as well as the original midday data points.

Graph 4: Interpolated Data with New Points



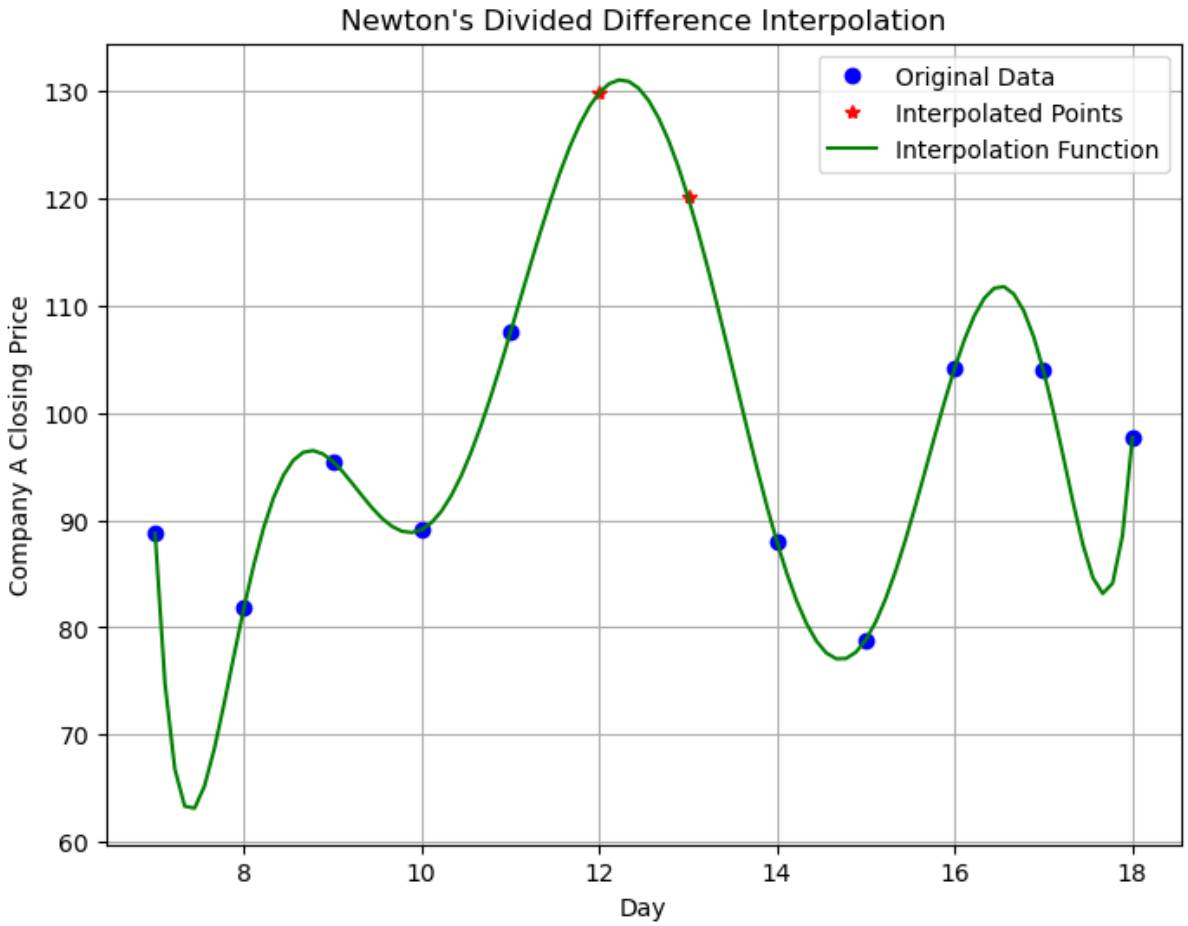
Problem 1.2: Stock Market Analysis
Use historical stock price data to interpolate missing values and analyze stock performance trends over time.

The stock market data includes dates and the closing prices of three different companies. There is no additional information apart from the closing price on the particular day. The missing data is seen with the weekend values missing as the market is closed. As a result Newton's Divided Difference is going to be the preferred method due to the uneven intervals of times being recorded. Obviously this data is not going to be that applicable to the real world as finding values on the weekend do not add a lot of value because the market is shut.

In order to analyse the performance of a stock the Augmented Dickey Fuller Test and the rolling mean & standard deviation can be used to evaluate how each of these different companies stocks have performed.

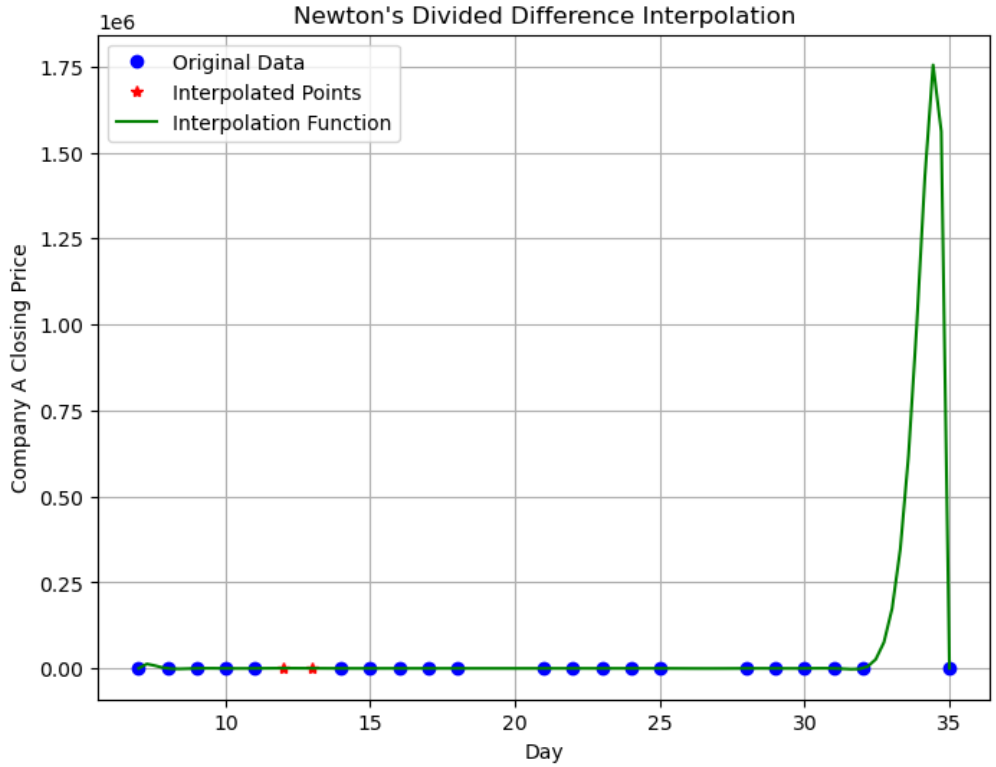
There is going to be a day column added which takes 2023-01-02 as day 0 and each day past that will add one value. This allows for easier repetition of the function for NDD and Lagrange's interpolation technique.

Graph 5: Newton's Divided Difference Interpolation for Company A



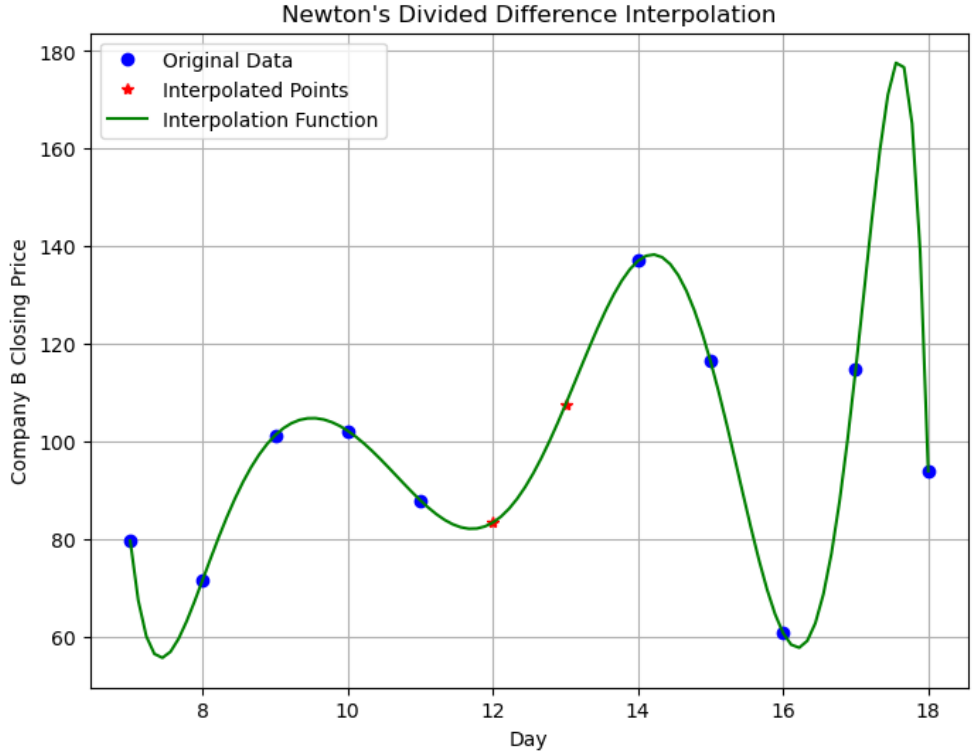
The graph above uses NDD to plot values for days 12 and 13 which fall on the weekend. You can see some of the limitations of NDD in this graph. In order to fit a function through every point it can be difficult to ensure accuracy in predictions. The weekend values (interpolated datapoints) are far above the recorded original data. It is likely that this data would not strong represent what the value of the stocks of Company A were on this weekend. This is when evaluation techniques could be used to ensure that the interpolation has some sense of accuracy.

Graph 6: Runge's Phenomenon for Newton's Divided Difference



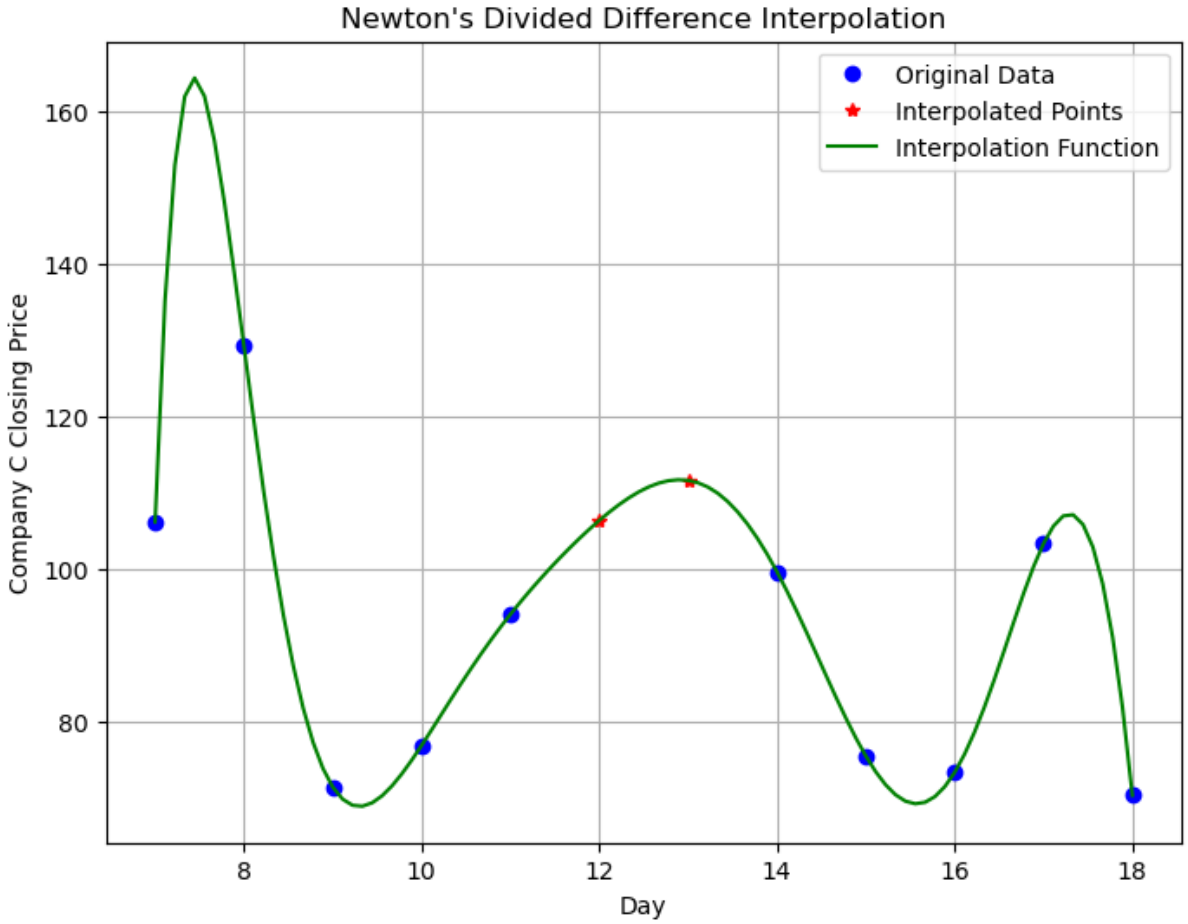
When you increase the number of data points to above 10 it results in the Runge's Phenomenon. I have included about 20 data points to show the extent of its effect. You can see that the graph goes through the other points but because of the magnitude of the spike between days 33 and 35 the model instantly becomes negligible because that behaviour would never occur.

Graph 7: Newton's Divided Difference Interpolation for Company B



Looking at Company B over the same time period you can see that the data that is used for days 12 and 13 seem to be a little bit more of what would be expected. They follow the trend that between day 11 and day 14 there is an increase in trend. However the spike that occurs between day 17 and 18 is most likely not what would have happened during the opening hours of the stock market.

Graph 8: Newton's Divided Difference Interpolation for Company C



Company C's interpolation shows less variability. However there is another spike that occurs between days 7 and days 8. This seems to be the beginning of the Runge Phenomenon because there would be no particular reason for the data to follow that movement. It is clear that the interpolation methods are good but they do have some constraints and not all data from the interpolation should be taken without further analysis of its accuracy.

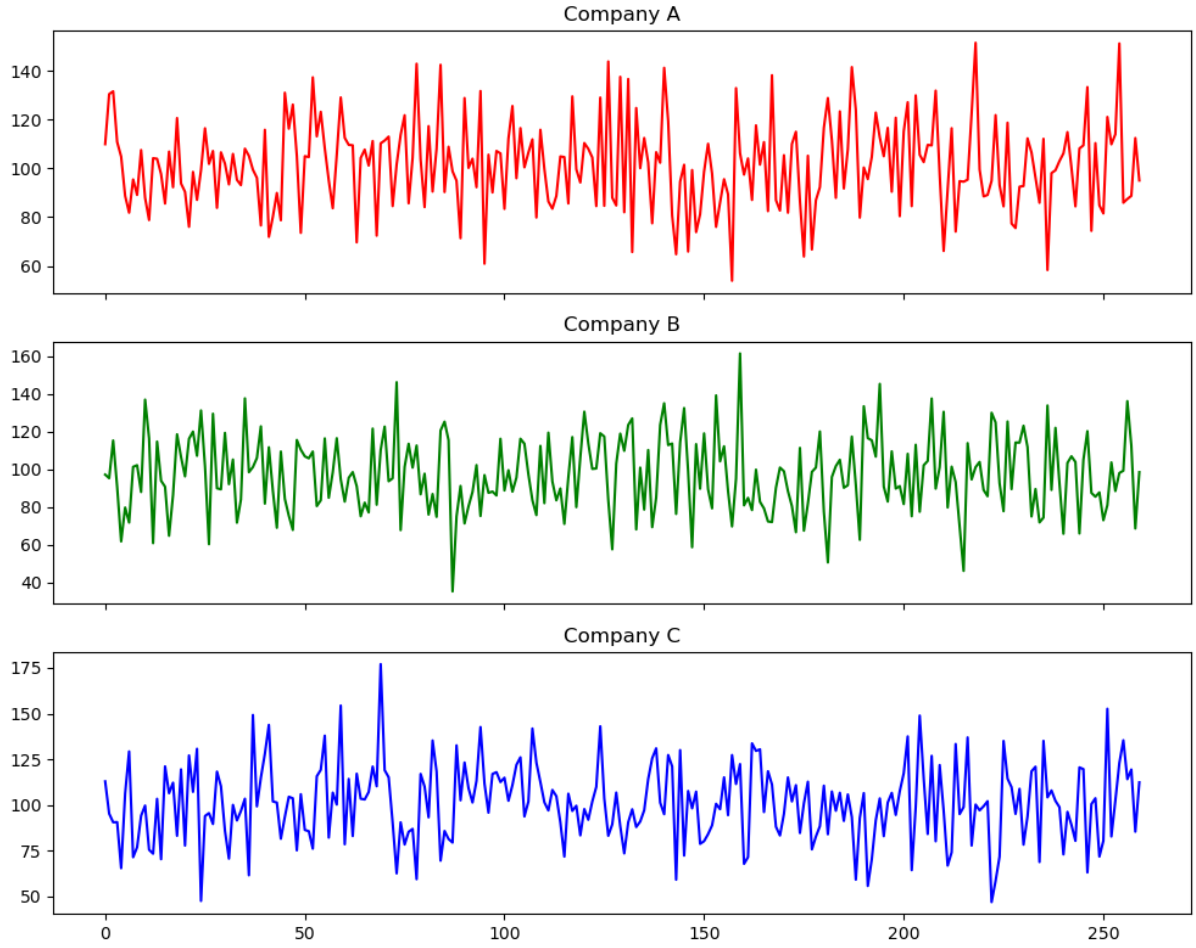
For the analysis of these companies, a subset is not required and the entire data can be looked at. The data takes place for about 300 days and the weekend values are skipped in the time series data. I have put all three companies changing in closing price into a subplot so you can compare trend, seasonality and determine if there is any cycles in the data.

The following methods are going to be used to evaluate the performance of the stock over the past 300 days.

- Augmented Dickey Fuller Test
- Rolling Mean & Standard Deviation

<u>Augmented Dickey Fuller Test</u>
Checks for the seasonality likelihood with a certain threshold for each level of confidence (90%,95% and 99%). Can help determine if there is seasonality or cycles.
<u>Rolling Mean & Standard Deviation</u>
Will look at the change in average closing price and how much it varies. This can help determine if there are any trends.

Graph 9: Company's Closing Price Comparison



Visualising the different company's closing price can usually help in the understand of components such as trends, cycles and seasonality.

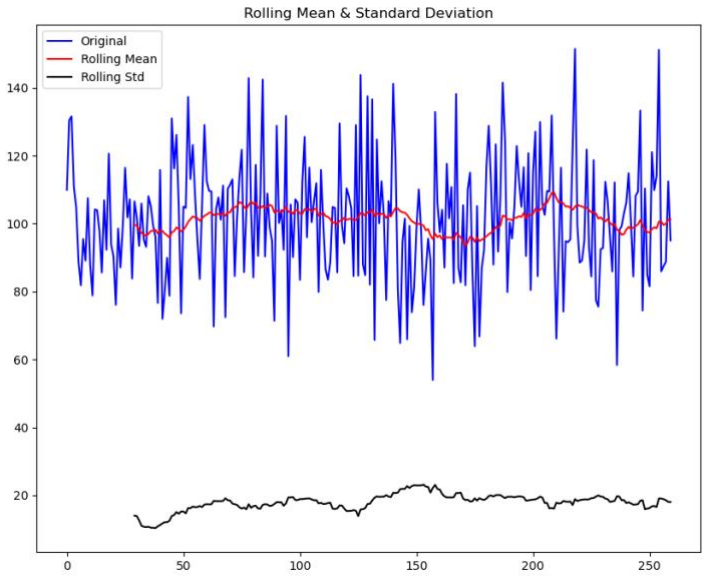
All three company's do not look as though they show any of the components stated above. You can see that the data follows a random and relatively constant variance in closing price. This would suggest that these company's have not displayed any strong signs of growth or any particular changes in price during certain time periods. When looking to see if the company's have similarities in their growth it does not look like they follow any similar pattern that is shared.

The ADF for each company is well above the threshold for having a 99% confidence that the graphs are stationary.

Stationarity

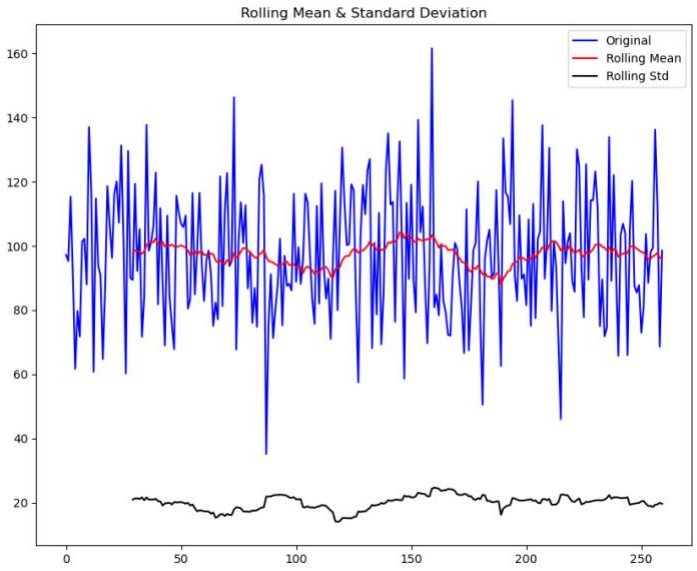
There is no relationship between the previous values and the current values. Meaning that time does not have a relationship with the values of the time series. This results in a constant mean and standard deviation

```
Company A - ADF Test
ADF Statistic:  -17.328377407052628
P-Value:  5.428714493560247e-30
Critical Value:
        1%: -3.46
        5%: -2.87
        10%: -2.57

Company B - ADF Test
ADF Statistic:  -12.478506195251528
P-Value:  3.1292222639171183e-23
Critical Value:
        1%: -3.46
        5%: -2.87
        10%: -2.57

Company C - ADF Test
ADF Statistic:  -6.168399934893085
P-Value:  6.899858545712441e-08
Critical Value:
        1%: -3.46
        5%: -2.87
        10%: -2.57
```



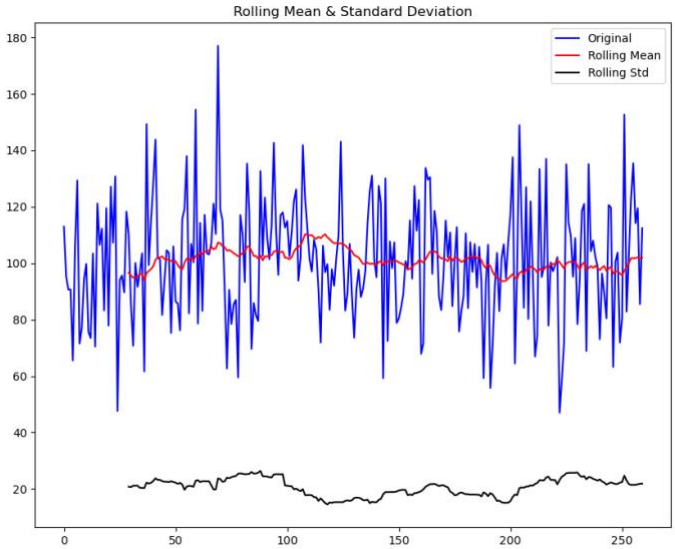
Company A Rolling Mean & Standard Deviation

The rolling mean has increases and decreases but it remains around the average of about $100 as the closing price. This graph shows that there is no trend as there is no overall trend in value of closing price, no seasonality because the variation of data is constant and no cycles because of the alteration around a constant mean.



Company B Rolling Mean & Standard Deviation

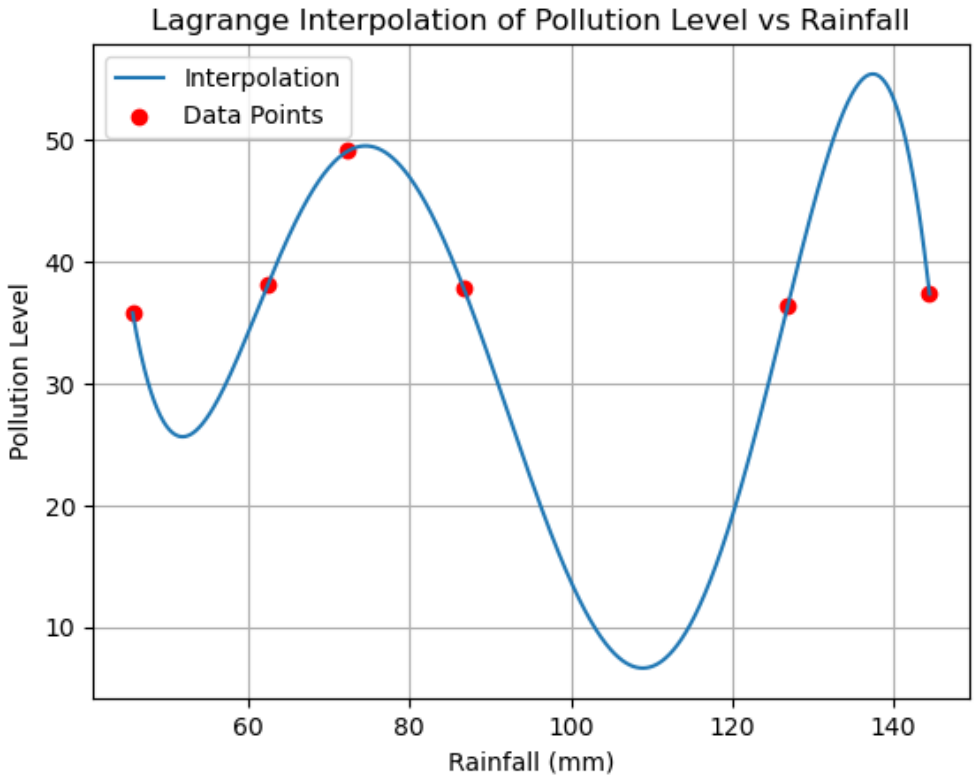
The rolling mean has increases and decreases but it remains around the average of about $100 as the closing price. This graph shows that there is no trend as there is no overall trend in value of closing price, no seasonality because the variation of data is constant and no cycles because of the alteration around a constant mean.



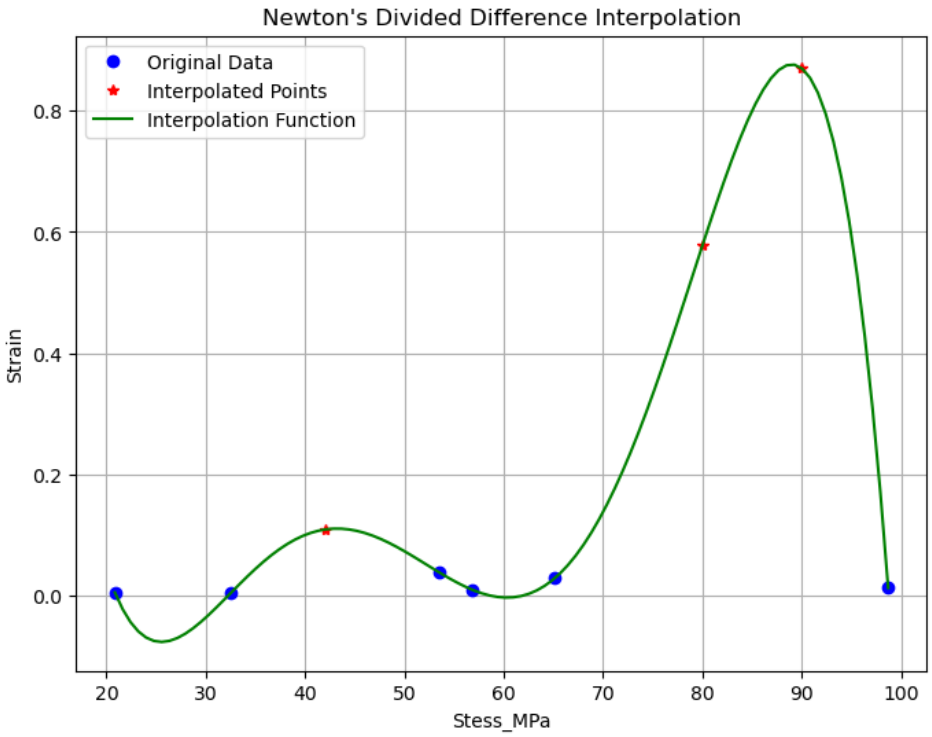
Company C Rolling Mean & Standard Deviation

The rolling mean has increases and decreases but it remains around the average of about $100 as the closing price. This graph shows that there is no trend as there is no overall trend in value of closing price, no seasonality because the variation of data is constant and no cycles because of the alteration around a constant mean.

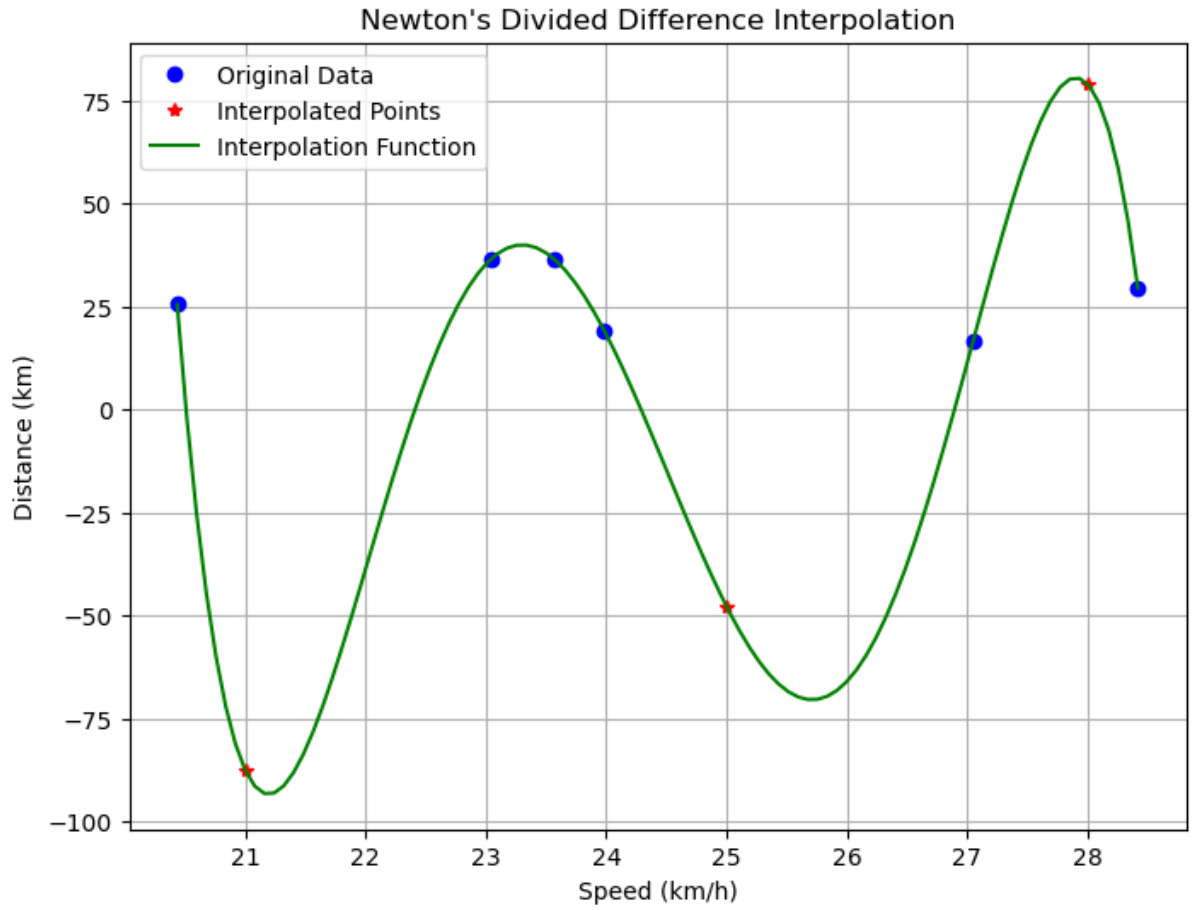
Problem 1.3: Environmental Studies
Interpolate data points for a geographic region based on sparse environmental measurements (e.g., pollution levels, rainfall).



Problem 1.4: Engineering Applications
Given experimental data points from materials testing (e.g., stress-strain values), interpolate missing values to model material behavior.



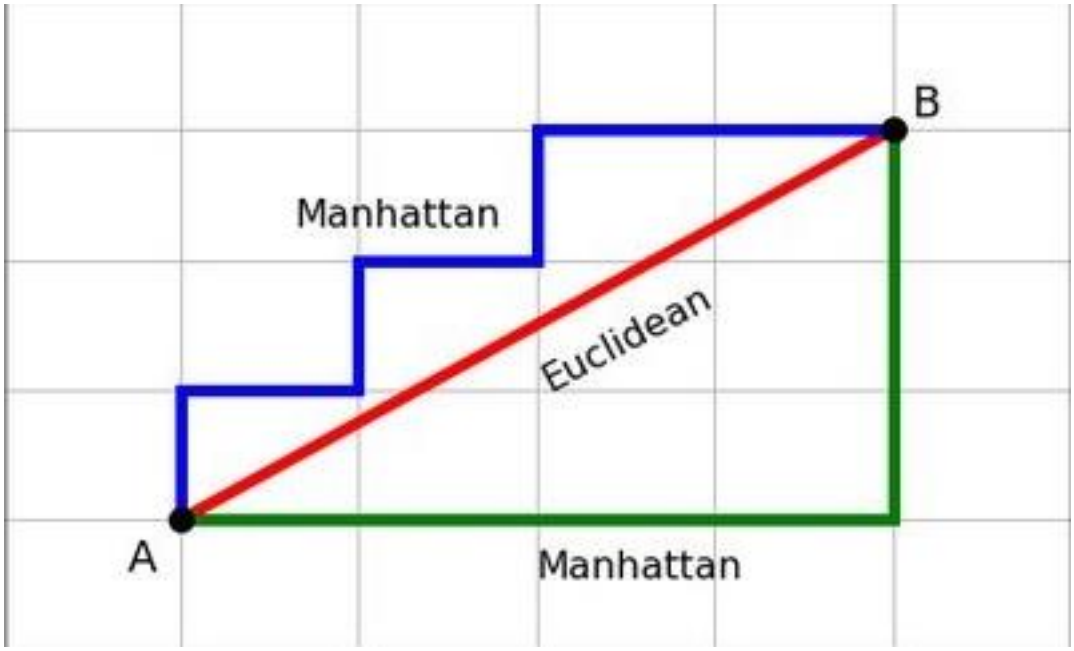
Problem 1.5: Sports Performance Analysis
Use recorded performance metrics (e.g., speed, distance) from sports events to interpolate missing values for a comprehensive performance analysis.



Problem 2: Chebyshev Distance

1. Calculate the Chebyshev Distance between two sets of points. (10 points)
2. Discuss the significance of Chebyshev Distance in the context of approximation and interpolation. How does it compare to other distance metrics like Euclidean distance? (10 points)

When comparing data points, you can find the distance between the based off their position from one another. The most popular forms of distance are Euclidian Distance and Manhattan Distance. As seen in the image below.



https://www.quora.com/What-are-the-differences-between-Manhattan-Distance-and-Chebyshev-Distance

Euclidian Distance is the most common distance metric as it creates a straight line between two points. This straight line can be found using the Pythagorean theorem:

$$a^2 + b^2 = c^2.$$

Manhattan Distance is commonly used in city block distances as it sums the differences of cartesian co-ordinates. Viewing a city as a grid you would be able to go down particular streets to minimise the time and distance needed to get from point a to point b. The equation is used to find the Manhattan Distance.

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

https://www.shiksha.com/online-courses/articles/all-about-manhattan-distance/

The Chebyshev distance takes a bit of a different approach to the other distance methods mentioned. The Chebyshev distance is about finding the maximum absolute difference between two points. This makes it particularly useful where movement can happen in any direction. It can commonly be seen in game development and grid based pathfinding.

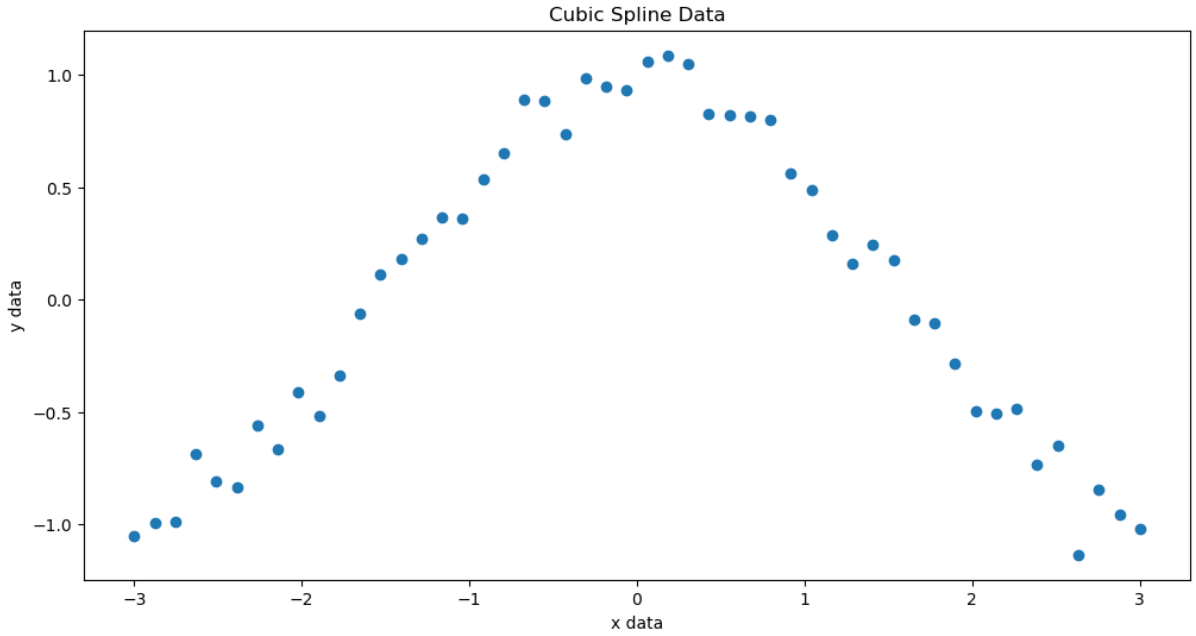
Chebyshev distance plays a vital role when interpolating polynomials to fit the existing data. In particular, it minimises the effects of the Runge's Phenomenon which takes place when a polynomial needs to fit a larger number of nodes. This usually results in the polynomial having a high magnitude of oscillation at a particular point of the polynomial to properly fit the data. However this makes the interpolation method not that effective.

Chebyshev nodes are distributed in a non-uniform manner with them concentrating more towards the end points of the interval. Choosing these points when interpolating is going to reduce the numerical instability and reduce the rapid oscillations when dealing with high degree polynomials.

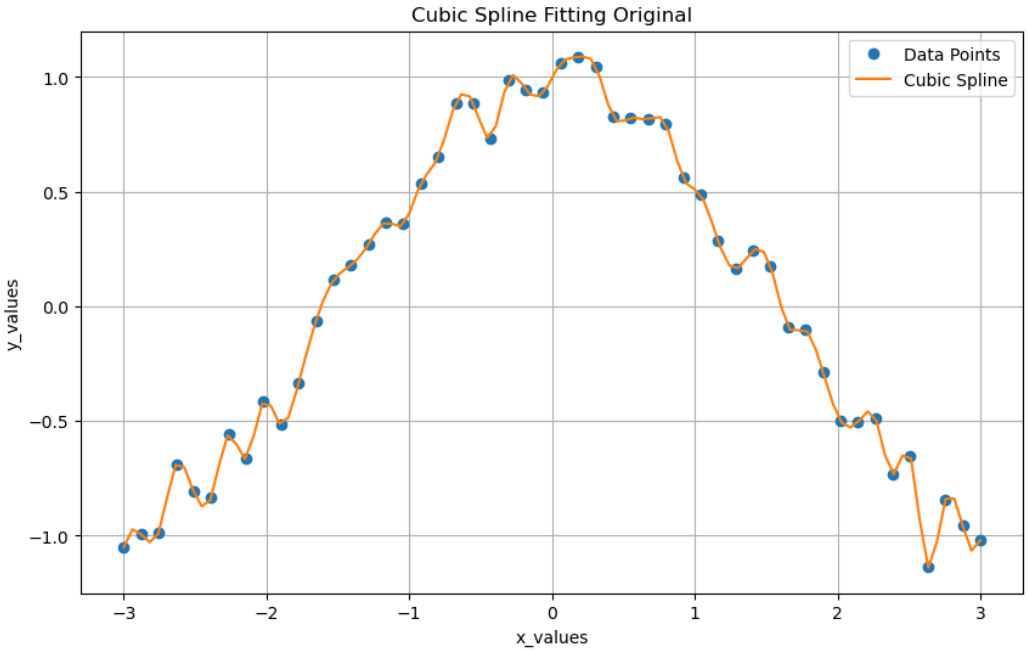
For the given dataset:
X: [-0.63635007 -0.39151551 -0.13610996 0.22370579 -0.4157107 -0.08786003 -0.60065244 0.18482914 0.2150897 -0.86989681]
Y: [-0.63319098 0.04951286 -0.41754172 -0.72101228 -0.26727631 0.57035192 0.02846888 -0.90709917 -0.65895175 0.89777107]
the following is the Chebyshev distance: 1.7676678885361075
This means that the maximum difference between x and y values is 1.77 which was found from looking at the difference between the 10th coordinate point.

Problem 3: Cubic Splines

1. Implement Cubic Splines to fit a smooth curve through a set of data points.
2. Evaluate the smoothness of the curve by adjusting the spline parameters. Discuss the trade-offs between curve smoothness and fitting accuracy.



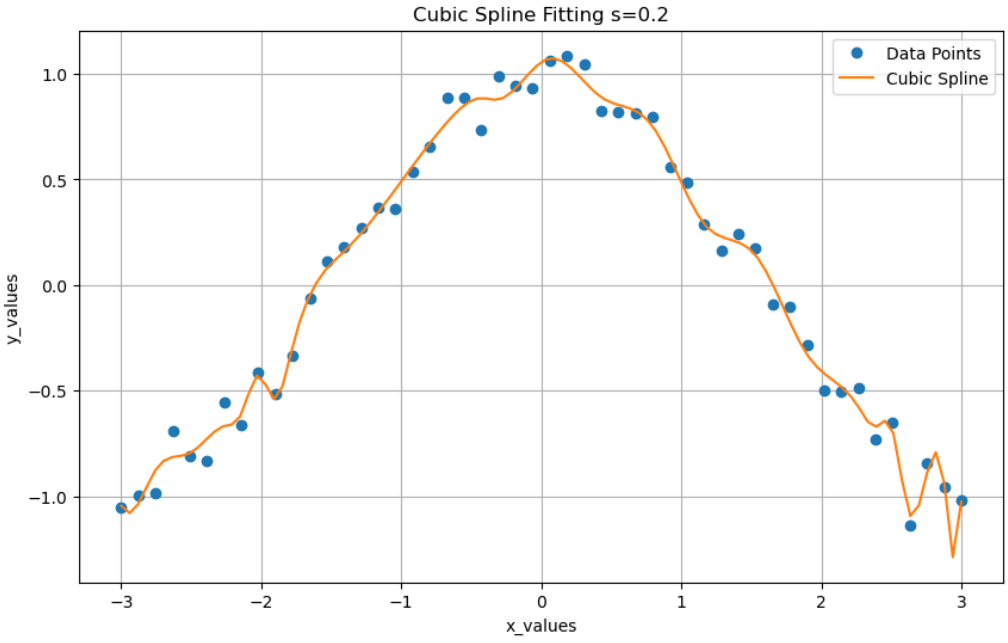
Above is the scatterplot for the data that is going to have the cubic spline interpolation technique used on it. You can observe that the data does follow an overall trend. It appears it is an upside-down quadratic with noise in the values. Using cubic splines you can create interpolation.



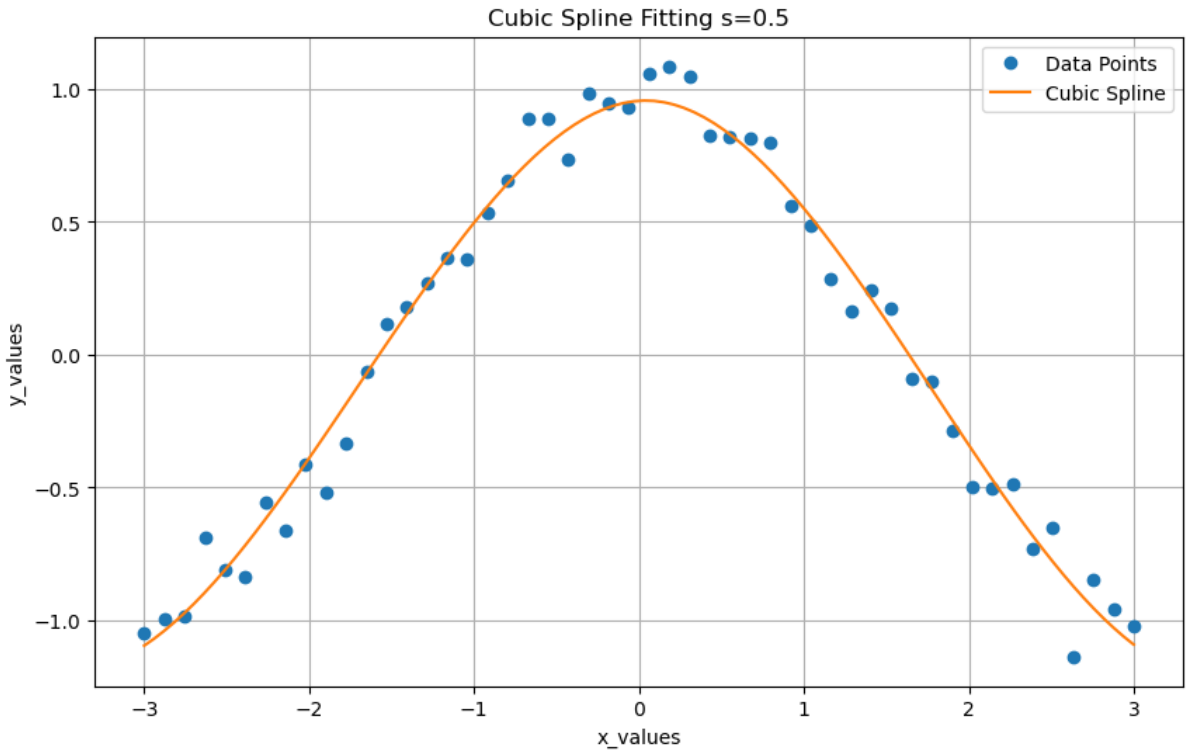
This is the base cubic spline interpolation, it matches the points very closely. This was an issue that was found using the other interpolation techniques, that they would usually have values that did not look like they would make logical sense. Not only that but the cubic spline is able to take place across all of the data which would not be possible using Lagrange or NDD due to the complexity of the polynomial. However, sticking with a cubic allows for fitted data but this would be a very large piecewise function that takes more complex mathematics to solve for.

Changing the smoothness of the Cubic Spline

The first component that is going to be changed is the s value for the cubic spline. The s parameter is the smoothing effect for the spline. As you increase this the smoothness of the spline increases. This is better for understanding the overall trend of the movement of the data that takes place. However it might miss out on vital movements that take place within the data. These could be outliers which represent a particular movement but using a higher smoothing value it would not be included in the interpolation of the data.



You can observe that the spline does not go through each of the original data points when changing the s value to a slightly higher one. This will mean that it will not follow the noise of the data as much. This seems to be a good value of s as it mitigates the extent of the noise and tries to capture the overall movement. Using the original cubic spline likely reflects overfitting as the trend of the data should not follow so closely if you are trying to generalise the interpolation, which is more likely the case.



The value of s at 0.5 further generalises the data. This does seem to create a very clear visualisation of the overall trend between x and y values. This is a good s value if you do not want to risk overfitting, however because of the trend of the data following a pretty clear shape this is not too much of an issue to worry about. If the data was more random with larger fluctuations then having a higher s value would become more important.

Conclusion

After exploring the different techniques that are used for interpolating and approximating across a variety of fields it has proven how valuable these techniques can be. Lagrange interpolation is a straightforward method for smaller datasets to get an approximation of a polynomial that goes through all the data. Like Newton's Divided Difference however it uses a recursive approach that works best for uneven data points. Both suffer from Runge's Phenomenon which limits both to best use case being with smaller ranges and data sets.

Chebyshev distance has a distinct approach to measuring the distance between points. This can help reduce the Runge Phenomenon in interpolation. It is used in enhancing stability and accuracy with high degree polynomials.

Cubic splines are a versatile method that is used for creating smooth interpolations through data points. It can manipulate different parameters to change the type of smoothing that exists.

Overall, the choice of the interpolation technique will depend on the understanding of the dataset and the desired outcomes. It is important to be aware of all of these techniques and to understand their strengths and weaknesses when applying them to real world scenarios.