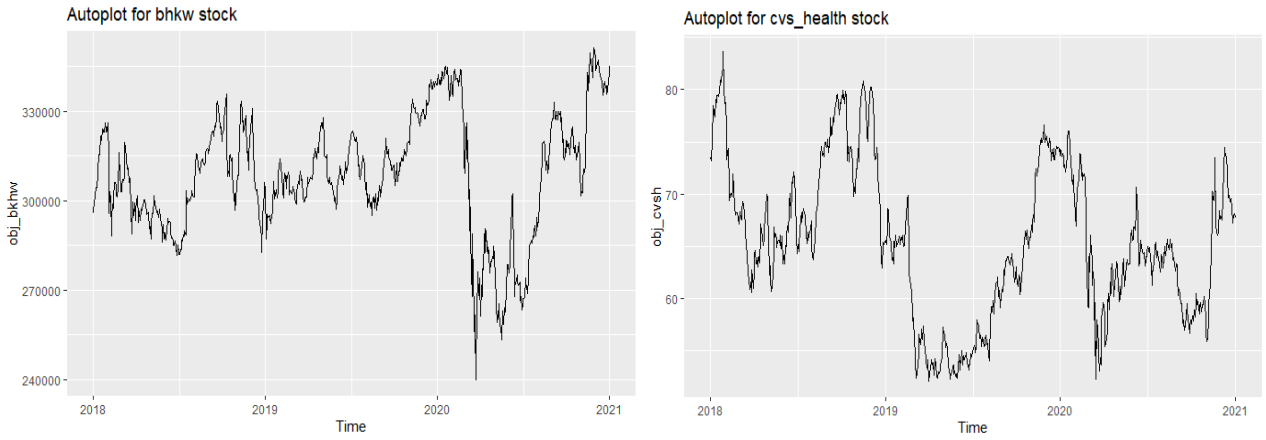
MSCI 718 PROJECT GROUP 16

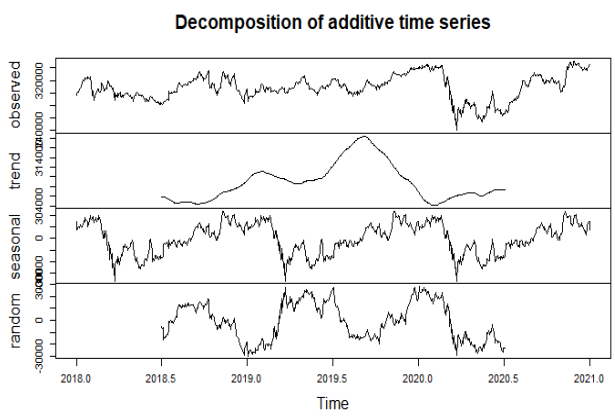
INTRODUCTION: A company's worth can be determined using its market capitalization (i.e., multiplying the company's stock price by the number of shares outstanding). The stock price is a relative and proportional value of a company's worth. Therefore, it only represents a percentage change in a company's market capitalization at any given point in time. This project will conduct a time series analysis and future stock price prediction for CVS health and Berkshire Hathaway (among the top 6 in the fortune 500 lists of companies). CVS Health is an American healthcare company that owns a pharmacy, retail pharmacy chain, health insurance provision, and many other brands. Berkshire Hathaway is an American multinational conglomerate holding company that owns multiple businesses and investments. Time-series analysis using the ARIMA forecasting technique is conducted based on the historical stock data for the last three years to check the effect and changes occurring due to the covid 19 pandemic on both the companies' stock prices and future market capitalization.

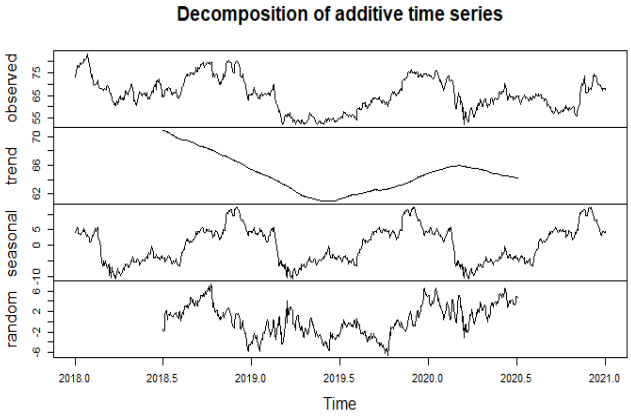
DATA & EDA: The project consists of two datasets stock price of CVS Health and Berkshire Hathaway, upon which the ARIMA modelling and hypothesis testing are conducted. The dataset used has the following characteristics:

1. The dataset consists of six features: Date, Open, Close, High, Low, Volume with values ranging from 03/01/2006 to 31/12/2020. Furthermore, the dataset has been pre-processed to remove NaN values and remove unnecessary features like open, high, low, and volume.
2. For ARIMA analysis, the final dataset has been prepared from 03/01/2018 to 31/12/2020, consisting of only the closing value for that stock, as shown in the figure below.
3. A separate **DASHBOARD** has been attached to check the live plot for the stocks with values.



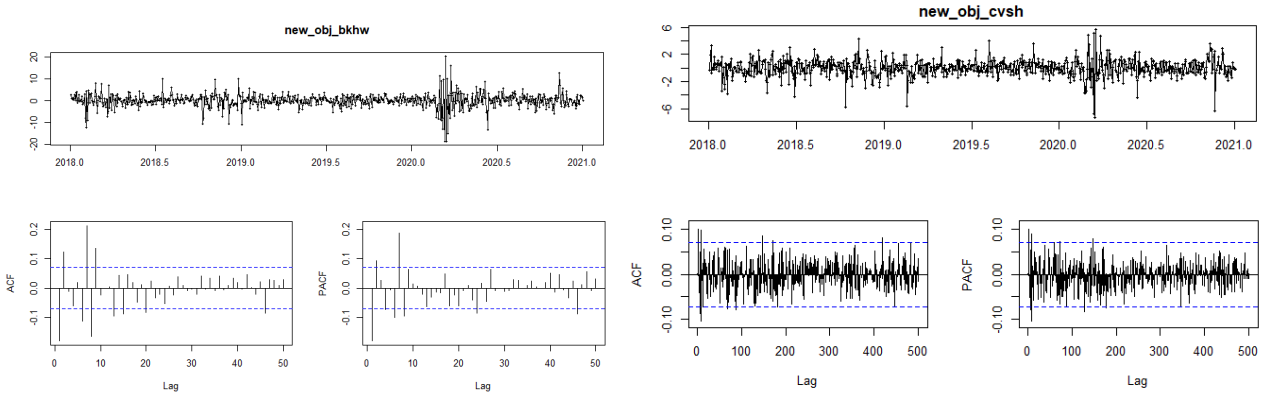
The plot above shows the movement of CVS health and Berkshire Hathaway stock price for three years. The range of CVS health stock varies between 55$ to 85$ and getting a peak value near January. The stock price of Berkshire Hathaway varies between 240k to 340k, with the stock performing at a stable rate between 280k and 320k except for the bottom-most point observed in January 2020. A detailed description of the stock data can be made using the separation of time components achieved by the decomposition plots. The graph below shows the decomposition plot that shows the additive decomposition of the time series plot in terms of seasonality, trend, and random components. The CVS health stock price shows a seasonal behaviour, in which the trend decreases initially and then increases. However, the stock price of Berkshire Hathaway shows a non-seasonal behaviour, in which the trend observes multiple peaks. It looks like, for trend, the price gradually decreases in the beginning and eventually increases at the end of the year for both companies.





PLANNING & ASSUMPTIONS: Assumption plays a significant role in determining the parameters of the model; for instance, ARIMA consists of three different parts: AR (Autoregressive part; p), I (Integration, degree of differencing; d) & MA (Moving average part, q) which will be decided based on the assumptions of the dataset. The following assumptions were tested before finalizing the hyperparameter for the ARIMA model.

STATIONARITY: ARIMA analysis requires data to be stationary (data with constant mean and variance). The autoplot for both the stocks shows a significant variation over time with changing mean, but no conclusive proof indicates stationarity. Therefore, the Augmented Dickey-Fuller test (ADF is conducted on the data). The null hypothesis of the ADF test cannot be rejected in the CVS health stock price as the p-value is significantly greater than 0.05 (Dickey-Fuller = -2.831, p-value = 0.2266). Thus, the data is not stationary. A similar result is obtained in the case of Berkshire Hathaway stock price; the null hypothesis of the ADF test cannot be rejected in the Berkshire Hathaway stock price as the p-value is significantly greater than 0.05 (Dickey-Fuller = -2.9251, p-value = 0.1867). Therefore, differencing is applied to the data to make it stationary and satisfy the assumption of ARIMA.



The figure shows the differenced data, and it can be seen that it has a constant mean and significantly less variation over time. The stationarity is confirmed using the Augmented Dickey-Fuller test (ADF). The null hypothesis of the ADF test can be rejected with a 95% confidence interval in the CVS health stock price as the p-value is significantly less than 0.05 (Dickey-Fuller = -8.31, p-value = 0.01). Thus, the data is stationary. A similar result is obtained in the case of Berkshire Hathaway stock price; the null hypothesis of the ADF test can be rejected with a 95% confidence interval in the Berkshire Hathaway stock price as the p-value is significantly less than 0.05 (Dickey-Fuller = -9.5, p-value = 0.01). Therefore, ARIMA with d = 1 must be applied to the modelling of both the dataset.

AUTOCORRELATION: Autocorrelation measures the extent of a linear relationship between lagged values of a time series. A correlogram is plotted using the ACF (), and PACF () function is used to check the autocorrelation between different lags in the time series. In the ARIMA model, the moving average part (q) is decided using ACF () plot by looking at the significant lag above 95% confidence interval. A similar approach is used for predicting the autoregressive part (p), which is decided using the PACF () plot. The key points observed from the above plots are:

- The ACF () plot for CVS health stock shows a seasonal behaviour due to the presence of a regular pattern and shape in the plot. However, the ACF () plot for Berkshire Hathaway stock price shows no signs of seasonality.
- The data for Berkshire Hathaway is not seasonal, ARIMA (p, d, q) will be used for the final modelling. The PACF () curve shows that the first eight lags are significant, which denotes that it is at least an ARIMA (8,1,0) model. The ACF () plot shows that at least the first five lags are significant (crossing the blue line indicating 95% confidence interval), meaning it is at least an ARIMA (8,1,5). Therefore, the final model must be selected based on residuals and AIC score.
- The time-series data for CVS health is seasonal, so ARIMA (p, d, q) x (P, D, Q) [m] will be used for final modelling. Based on ACF () and PACF () plots, we can check the seasonal patterns and get a lag value to capture the seasonal data. The final model selection for this case will be achieved using auto.arima function.

ARIMA MODEL: The model for the Berkshire Hathaway dataset involves fitting various ARIMA models between ARIMA (8, 1, 5) to ARIMA (10, 1, 10) to find the best fitting model based on the AICc score. The final model that provided the best results is ARIMA (9, 1, 9) that has 9 MA (moving average) coefficients and 9 AR (autoregressive) coefficients to form a linear time-dependent equation used in forecasting future values.

Model	Stock	AIC	AICc	BIC
ARIMA (9,1,9)	Berkshire Hathaway	3819.18	3820.21	3907.06
ARIMA (0,1,0) x (0,1,0) [251]	CVS Health	1971.63	1971.64	1975.85

The model for the CVS health dataset involves the use of auto.arima function to get the best model based on the AIC score. The results provide ARIMA (0,1,0) X (0,1,0) [251] as the best fitted model based on residuals and AIC score. The results have no moving average and autoregressive coefficient and have used a single differencing on the data, with the seasonal part of the model having 251 observations

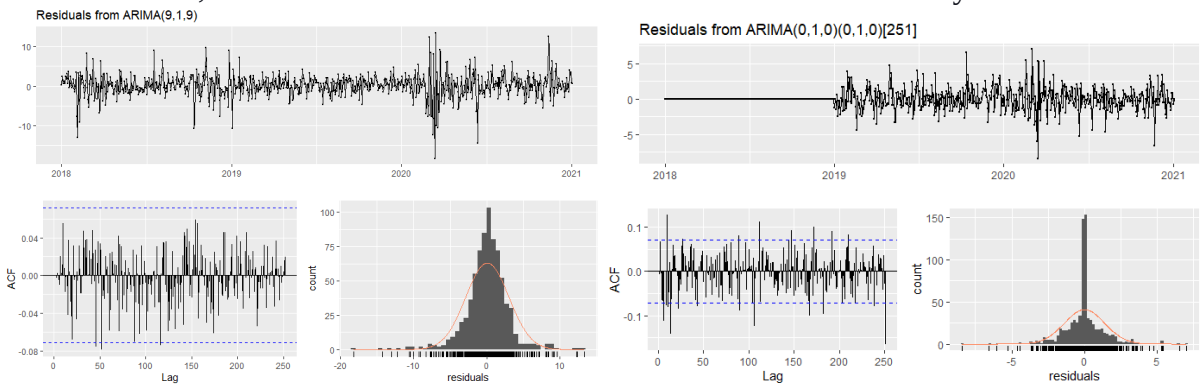
RESIDUALS DIAGNOSTICS: The residuals in a time series model are what is left over after fitting a model. In Arima time series modelling, residuals must be check for correlation (must be uncorrelated to ensure that forecasting is better and no information is left out), mean (must have mean zero or very close to zero to ensure that model is not biased), variance (must have a constant variance) and normal distribution.

The graph for the Berkshire Hathaway model (ARIMA (9, 1, 9)) shows that it is an appropriate fit for the data and forecasts the results unbiased. The above statement can be supported by the following evidence observed from residuals:

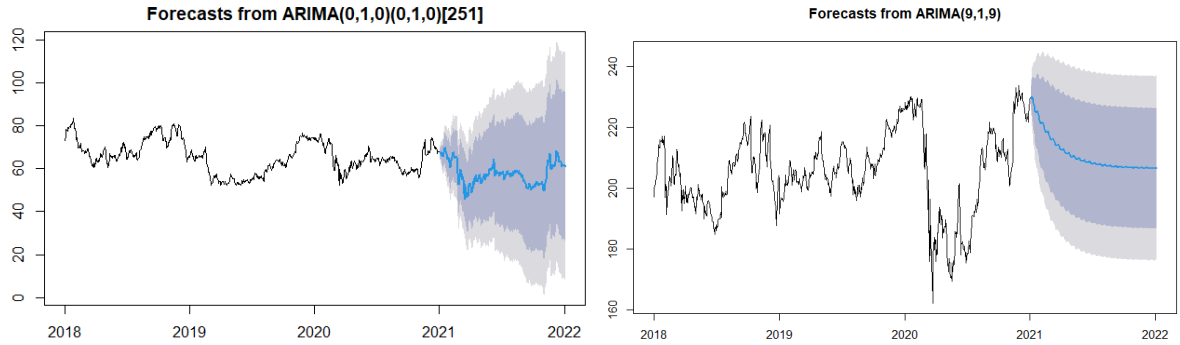
1. The mean of the residual is very close to zero (mean = 0.0920).
2. There is no significant correlation in the residual series based on the ACF () plot results.
3. The histogram suggests that the residuals are not entirely normal but very close to the normal distribution, except for the head part seems a little too long.
4. The Ljung-Box test suggests that the results are not significant as the p-value is very large (Q* = 109.28, p-value = 0.93). Thus, we can conclude that the residuals are not distinguishable from a white noise series, and hence no autocorrelation.

The graph for the CVS health model (ARIMA (0, 1, 0) X (0, 1, 0) [251]) shows that it is an appropriate fit for the data and forecasts the seasonal pattern of the stock. The above statement can be supported by the following evidence observed from the residuals:

1. The mean of the residual is very close to zero.
2. There are a few points which show a correlation among few data point in residuals. Since the data is seasonal, we continue to assume that there is no significant autocorrelation in the residual data.
3. The histogram suggests that the residuals are not entirely normal but very close to the normal distribution, except for the part at zero, which is very long.
4. The Ljung-Box test suggests that the results are significant as the p-value is small (Q* = 254.21, p-value = 2.98e-07). Thus, we can conclude that the residuals are distinguishable from a white noise series, and hence there is an autocorrelation due to the seasonality of data.



RESULTS: The results show the stock price prediction for Berkshire Hathaway () and CVS Health () for 2021, along with the 95% and 99% confidence interval. The highest point observed for CVS health is 69.64 $ around March 2021, and for Berkshire Hathaway, it is observed in January at 230.1$. The prediction for the first stock shows a seasonal behaviour based on the auto Arima results. The second stock keeps a steady decline in the first half then remains almost constant (significantly less variation).



CONCLUSION: The company's market value is decided based on outstanding shares of a company in that month or year and the company's share price. The outstanding shares of a company can fluctuate for several reasons, and that's why we will predict the market value of companies for the year 2021 based on the latest outstanding shares of both companies. The calculation shows CVS health market value for 2021 is 91.43732 billion (based on the stock's highest predicted value using the ARIMA model). Moreover, a similar result is observed in Berkshire Hathaway's market value for 2020 is 352.081458 million. The market value of CVS health is more elevated than Berkshire Heathway because of covid all the businesses slowed down, but the health sector experienced a boost in the last two years. Thus, the predicted market value of CVS health is higher compared to Berkshire Heathway for the coming year.

REFERENCES:

https://otexts.com/fpp2/arima.html

Introduction to Time Series Analysis and Forecasting in R | Udemy

https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/

https://plotly.com/r/candlestick-charts/

https://www.youtube.com/watch?v=hKDalfhDawA

https://www.youtube.com/watch?v=P-I0ljQpRCI

https://www.youtube.com/watch?v=qaZNDKFnX_Y

https://towardsdatascience.com/analyzing-stocks-using-r-550be7f5f20d

https://rpubs.com/kapage/523169

https://docs.scipy.org/doc/scipy/reference/stats.html

https://www.geeksforgeeks.org/time-series-analysis-using-arima-model-in-r-programming/

https://ademos.people.uic.edu/Chapter23.html

https://otexts.com/fpp2/stationarity.html

https://www.kaggle.com/thebasss/currency-exchange-rates

https://github.com/rkarwayun/MSCI-718-Project/blob/master/Project.Rmd

TEAM MEMBERS:

DEEP ASHISH JARIWALA--------(20909290)

URVI PATEL----------(20877623)

AYUSHI AHJOLIA ----------(20907061)

FARHAN ZAHID ---------- (2091079)