amazon

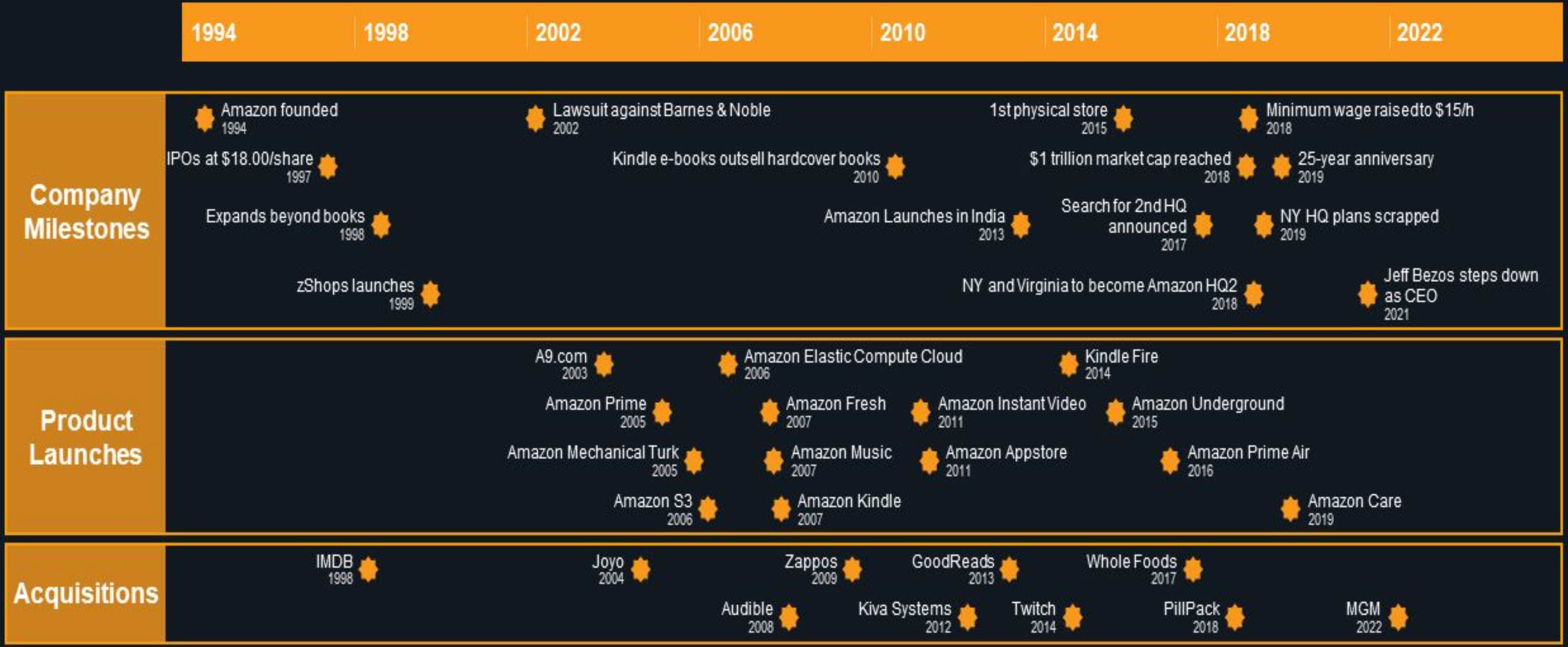
A Machine Learning Stock Forecasting Project

(R Programming Language)

Anabelle Capois Espinal

Tech Talent & Strategy, Data Science | February-May ‘23 Cohort | Instructor: Preston Menke amazon

All



1994 1998 2002 2006 2010 2014 2018 2022

Company Milestones

- Amazon founded 1994
- IPOs at $18.00/share 1997
- Expands beyond books 1998
- zShops launches 1999
- Lawsuit against Barnes & Noble 2002
- Kindle e-books outsell hardcover books 2010
- Amazon Launches in India 2013
- 1st physical store 2015
- Search for 2nd HQ announced 2017
- $1 trillion market cap reached 2018
- NY and Virginia to become Amazon HQ2 2018
- Minimum wage raised to $15/h 2018
- 25-year anniversary 2019
- NY HQ plans scrapped 2019
- Jeff Bezos steps down as CEO 2021

Product Launches

- A9.com 2003
- Amazon Prime 2005
- Amazon Mechanical Turk 2005
- Amazon S3 2006
- Amazon Elastic Compute Cloud 2006
- Amazon Fresh 2007
- Amazon Music 2007
- Amazon Kindle 2007
- Amazon Instant Video 2011
- Amazon Appstore 2011
- Kindle Fire 2014
- Amazon Underground 2015
- Amazon Prime Air 2016
- Amazon Care 2019

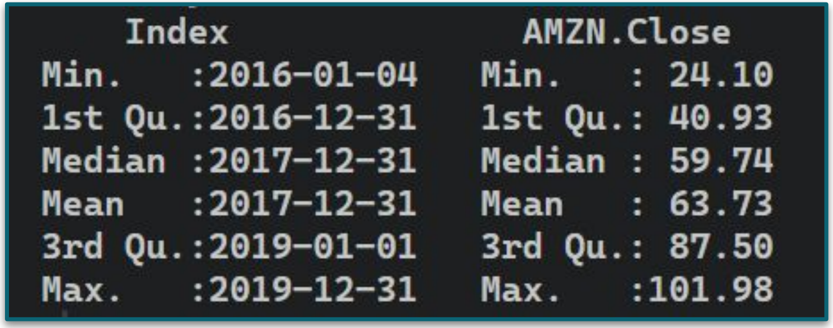
Acquisitions

- IMDB 1998
- Joyo 2004
- Audible 2008
- Zappos 2009
- Kiva Systems 2012
- GoodReads 2013
- Twitch 2014
- Whole Foods 2017
- PillPack 2018
- MGM 2022

Made with Office Timeline

About the Data

- Stock data from 01-01-2016 to 01-01-2020
- Only the closing stock price will be used
- Used to create a model that predicts 90 days (3 months) into 2020
- Supervised Machine Learning method is the **ARIMA model**, which will be used for forecasting stock prices
- Unsupervised method is **K-Means clustering**, will be used for competitors stock performance

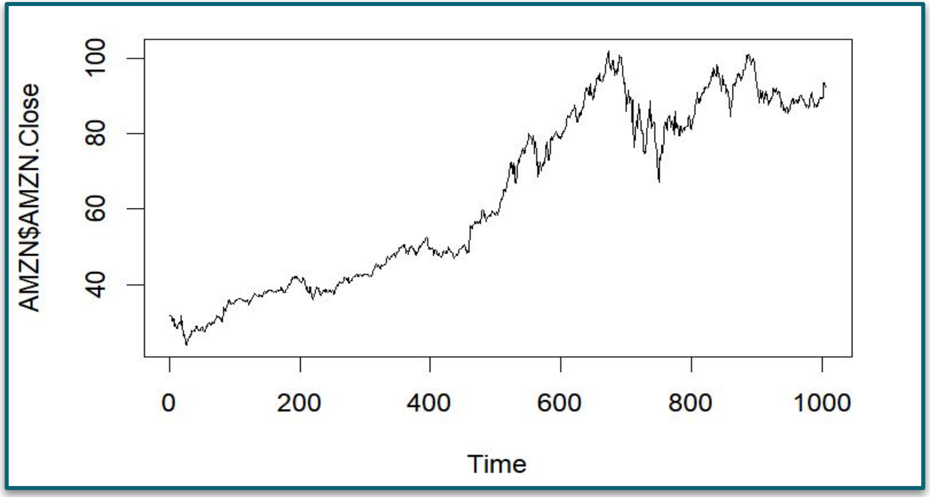


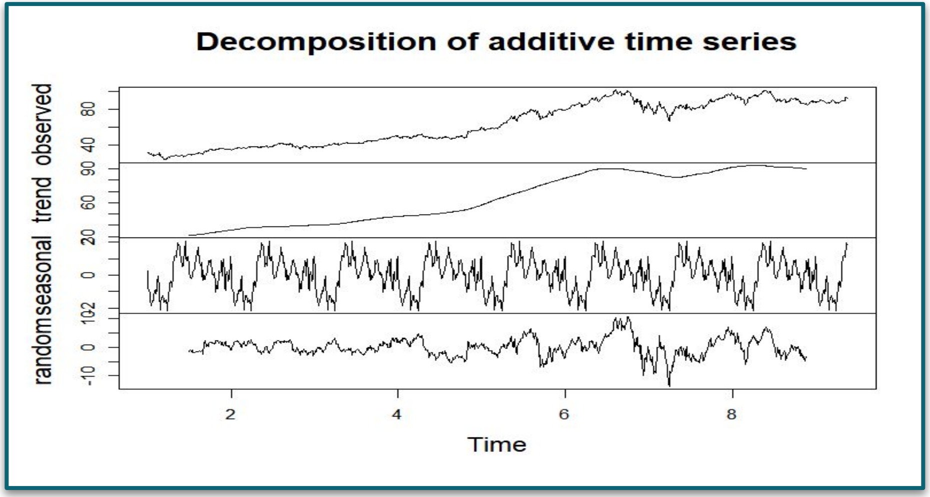
```
     Index             AMZN.Close
 Min.   :2016-01-04   Min.   : 24.10
 1st Qu.:2016-12-31   1st Qu.: 40.93
 Median :2017-12-31   Median : 59.74
 Mean   :2017-12-31   Mean   : 63.73
 3rd Qu.:2019-01-01   3rd Qu.: 87.50
 Max.   :2019-12-31   Max.   :101.98
```

Descriptive statistics for Amazon's closing stock price from 2016-01-01-01 to 2020-01-01

Amazon Stock Data 2016-2019

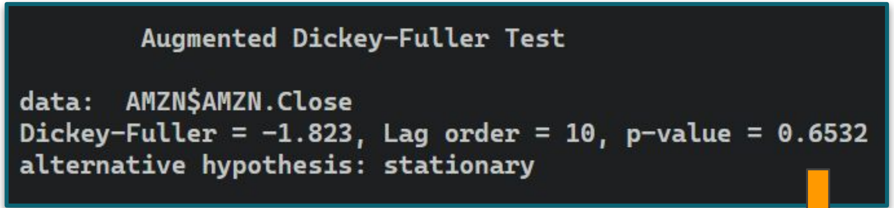
- Plotting the data reveals that the data is not stationary
- The time series graph can be broken down to see its individual components: random, seasonal, trend, and observed
- Time series models, like ARIMA, assume mean and variance are consistent
  - For better accuracy, the data gathered needs to be made stationary



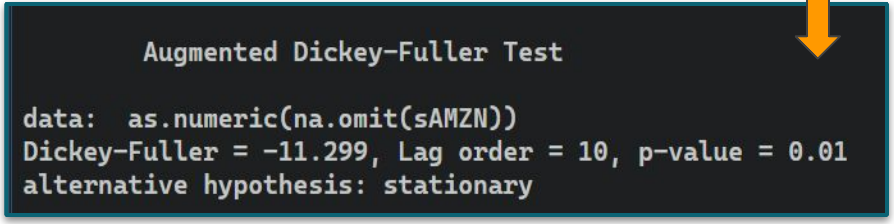


Stationary Data

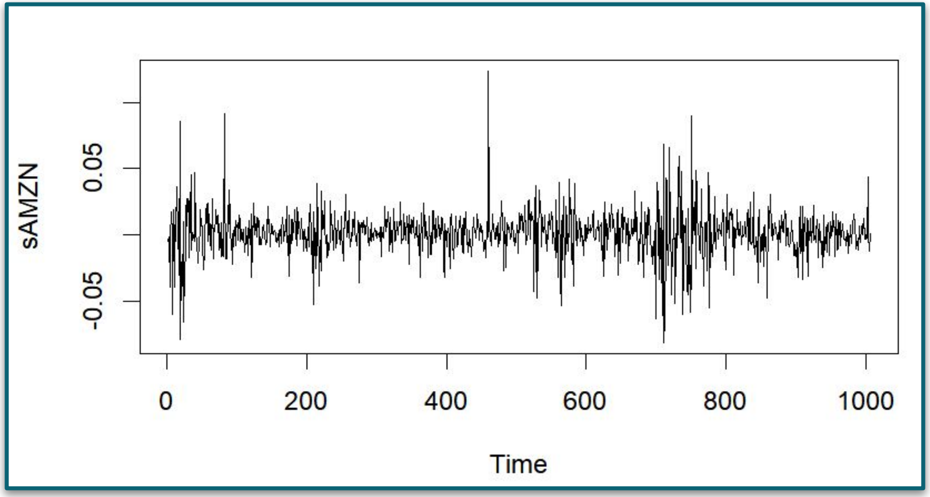
- Stationary = mean and variance do not vary across time
- Augmented Dickey-Fuller test confirms data is not stationary as is
  - P-value is much greater than 0.05
- Differencing can remove the effect of trend or seasonality, making the data stationary
  - P-value is below 0.05 and DF is a higher negative value
  - Plotting will show that the data is now stationary



```
        Augmented Dickey-Fuller Test

data:  AMZN$AMZN.Close
Dickey-Fuller = -1.823, Lag order = 10, p-value = 0.6532
alternative hypothesis: stationary
```

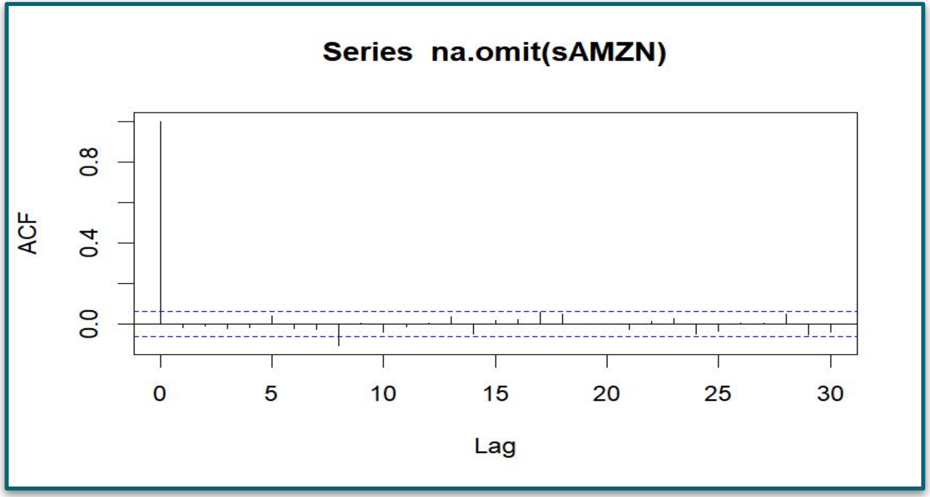


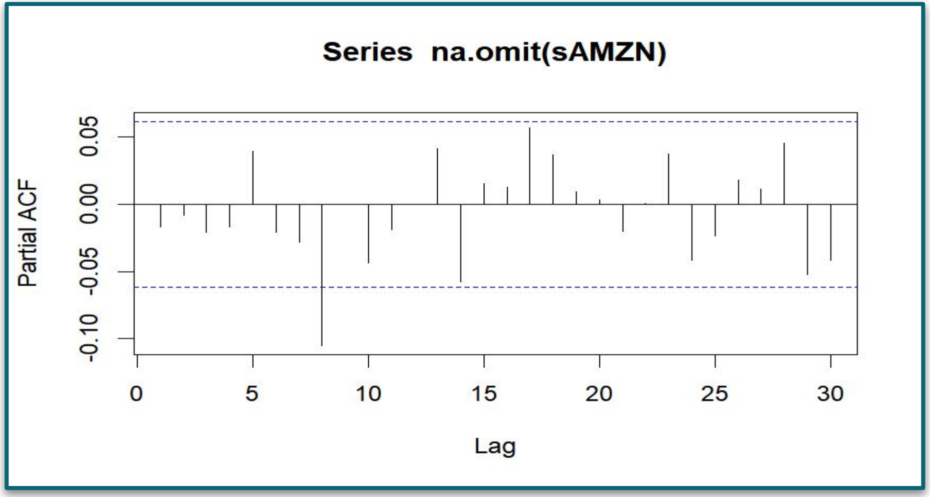
```
        Augmented Dickey-Fuller Test

data:  as.numeric(na.omit(sAMZN))
Dickey-Fuller = -11.299, Lag order = 10, p-value = 0.01
alternative hypothesis: stationary
```



Autocorrelation Analysis

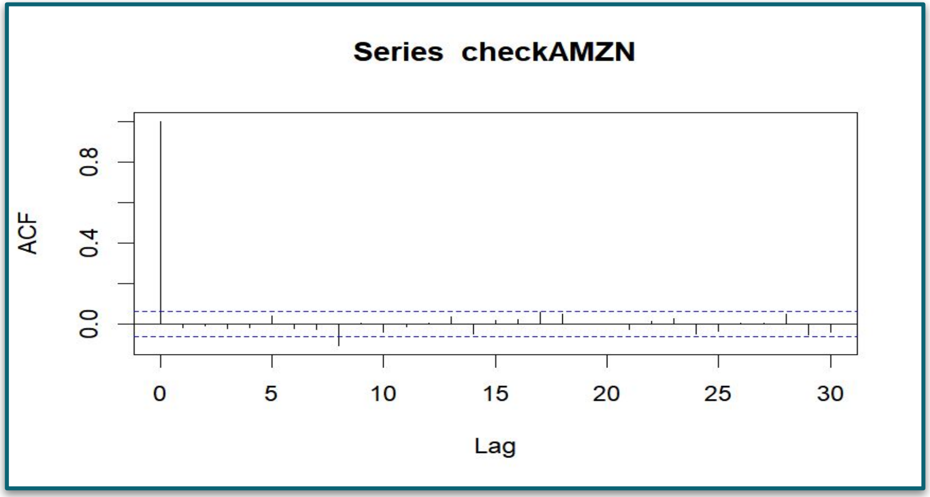
- Autocorrelation analysis to detect patterns and check for randomness
  - One significant non-zero autocorrelation = not random
  - Only one of the values is outside of the bounds for the PACF
    - White Noise = uncorrelated random variables with constant mean and variance
      - can’t obtain parameters from the ACF and PACF = likely use ARIMA (0,0,0)

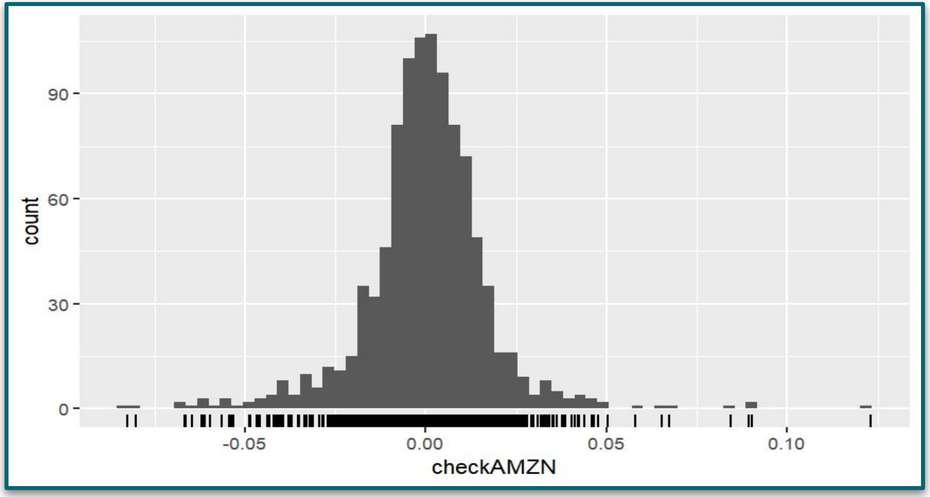


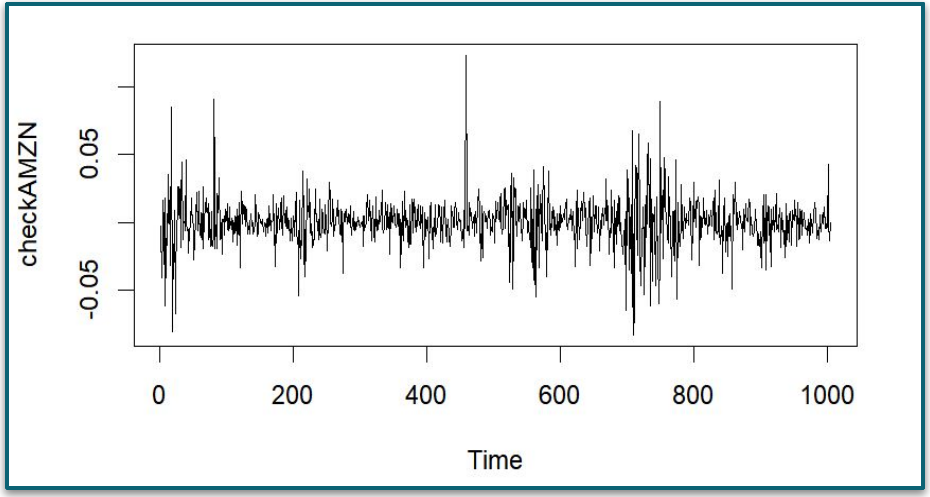


Diagnostic Check

- To confirm that ARIMA (0,0,0) is best fit, residual are:
  - Not correlated/are independent
  - Have zero mean
  - Normally distributed

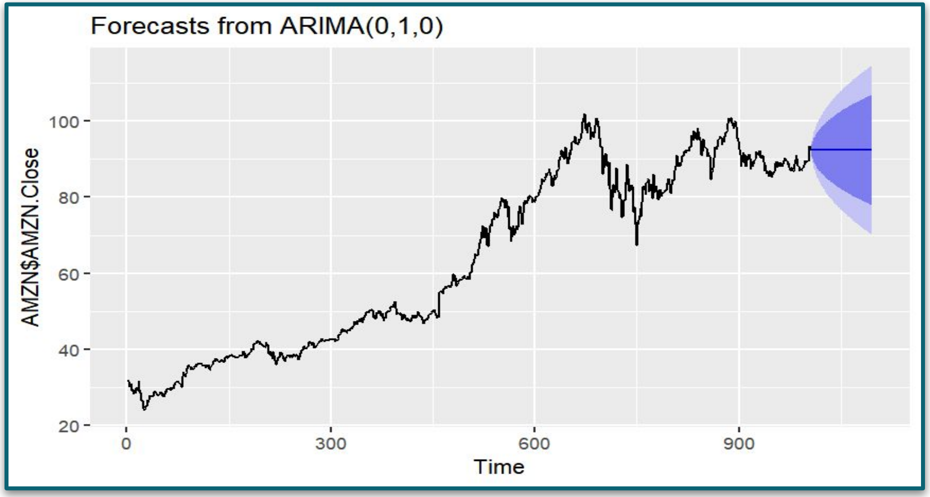


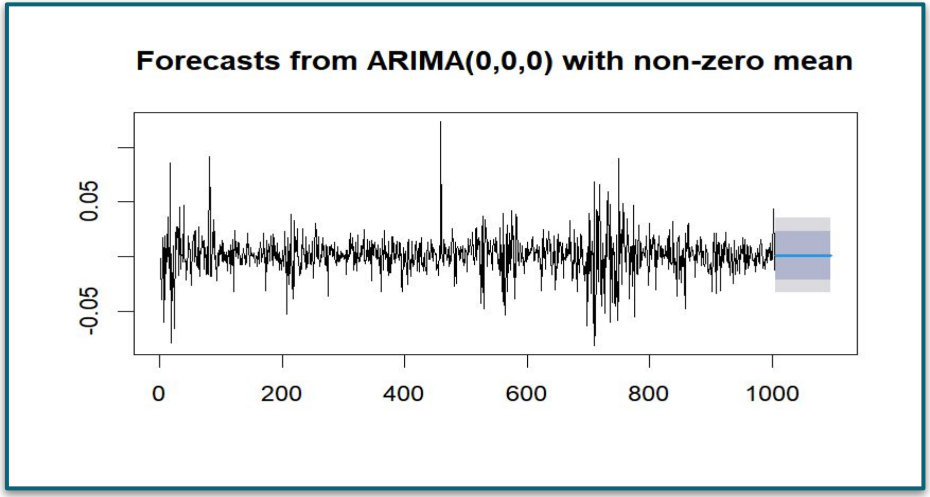




Forecasting 90 Days(3 Months)

- Forecasting with the original data, which is not stationary, can be observed in the left graph
- The stationary data forecasting can be observed on the right graph
- The lighter areas represent high and low at 95% confidence that the true population forecast is somewhere between these values
- The darker areas represent high and low at 80% confidence

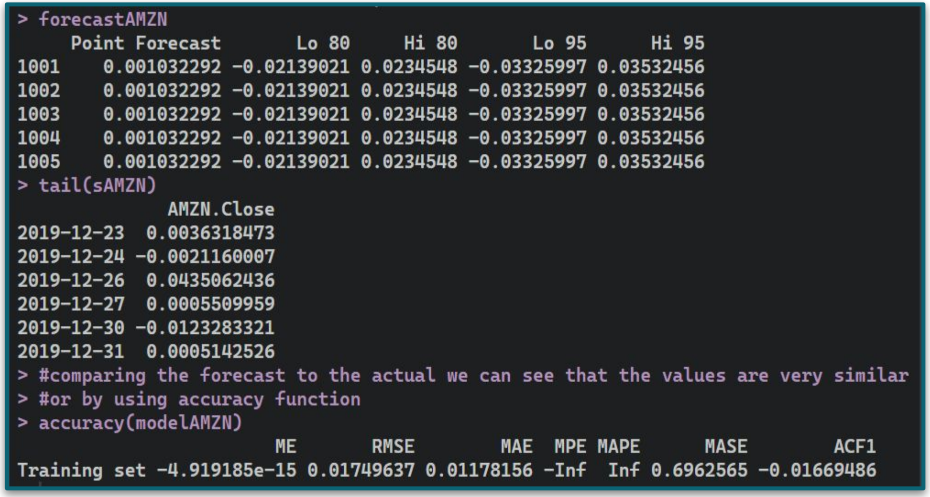




Checking for Accuracy

Model has high accuracy!

- Comparing predicted with actual data = values are similar
  - tail values can be compared to what the model would predict for these same values
- Looking at the MAPE value
  - Because there are 0 values, MAPE comes out to infinity, so we look at Mean Absolute Error
- Looking at the MAE value
  - Low value represents high accuracy of the model



```
> forecastAMZN
     Point Forecast       Lo 80     Hi 80       Lo 95      Hi 95
1001    0.001032292 -0.02139021 0.0234548 -0.03325997 0.03532456
1002    0.001032292 -0.02139021 0.0234548 -0.03325997 0.03532456
1003    0.001032292 -0.02139021 0.0234548 -0.03325997 0.03532456
1004    0.001032292 -0.02139021 0.0234548 -0.03325997 0.03532456
1005    0.001032292 -0.02139021 0.0234548 -0.03325997 0.03532456
> tail(sAMZN)
              AMZN.Close
2019-12-23  0.0036318473
2019-12-24 -0.0021160007
2019-12-26  0.0435062436
2019-12-27  0.0005509959
2019-12-30 -0.0123283321
2019-12-31  0.0005142526
> #comparing the forecast to the actual we can see that the values are very similar
> #or by using accuracy function
> accuracy(modelAMZN)
                        ME       RMSE        MAE  MPE MAPE      MASE        ACF1
Training set -4.919185e-15 0.01749637 0.01178156 -Inf  Inf 0.6962565 -0.01669486
```

The Competition

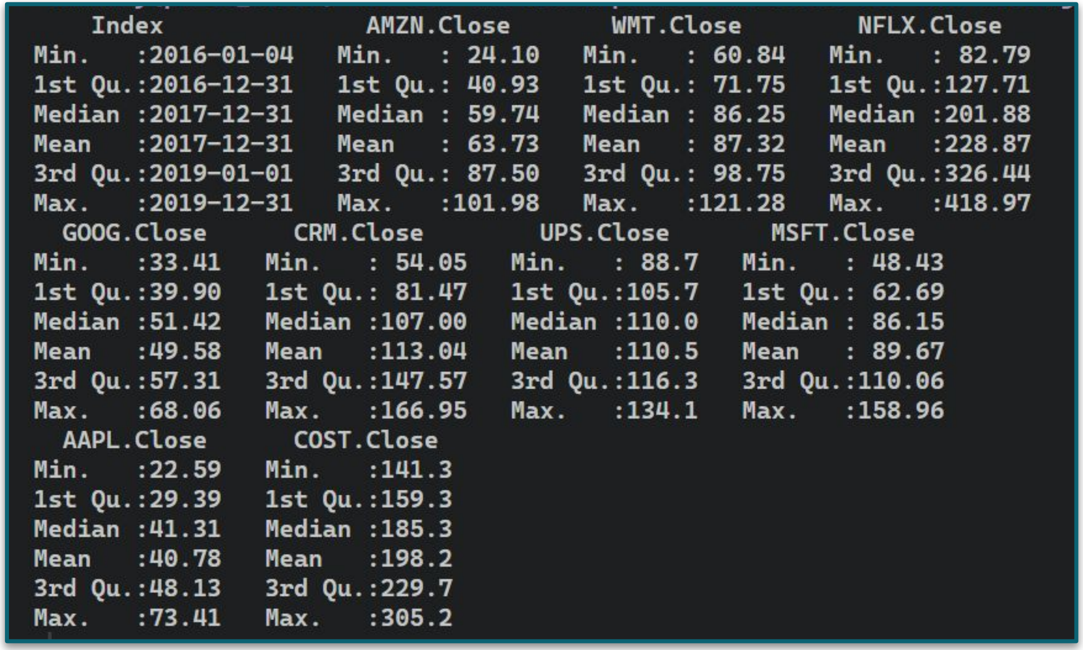
“Our current and potential competitors include:
(1) physical, e-commerce, and omnichannel **retailers**, publishers, vendors, distributors, manufacturers, and producers of the products we offer and sell to consumers and businesses;
(2) publishers, producers, and distributors of physical, digital, and interactive **media of all types** and all distribution channels;
(3) **web search engines**, comparison shopping websites, social networks, web portals, and other online and app-based means of discovering, using, or acquiring goods and services, either directly or in collaboration with other retailers;
(4) companies that provide e-commerce services, including website development and hosting, **omnichannel sales**, inventory, and supply chain management, advertising, fulfillment, customer service, and payment processing;
(5) companies that provide **fulfillment and logistics services** for themselves or for third parties, whether online or offline;
(6) companies that provide **information technology services or products**, including on-premises or cloud-based infrastructure and other services;
(7) companies that design, manufacture, market, or sell **consumer electronics**, telecommunication, and electronic devices; and
(8) companies that sell **grocery products** online and in physical stores. “

-Amazon.com, Inc. 2020 Form 10k

Chosen Companies

Walmart = WMT
Netflix = NFLX
Google = GOOG
Salesforce = CRM
UPS = UPS
Microsoft = MSFT
APPLE = APPL
Costco = COST

The Competition

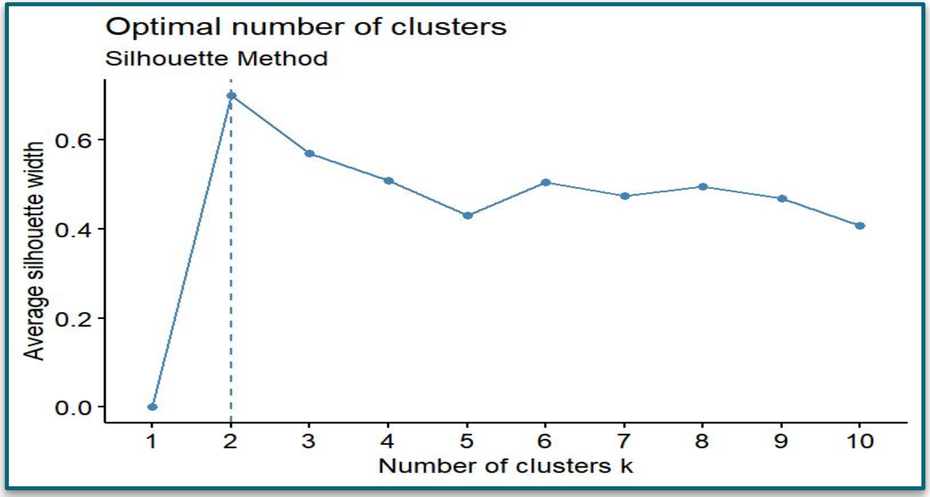


```
     Index                 AMZN.Close        WMT.Close         NFLX.Close
 Min.   :2016-01-04   Min.   : 24.10   Min.   : 60.84   Min.   : 82.79
 1st Qu.:2016-12-31   1st Qu.: 40.93   1st Qu.: 71.75   1st Qu.:127.71
 Median :2017-12-31   Median : 59.74   Median : 86.25   Median :201.88
 Mean   :2017-12-31   Mean   : 63.73   Mean   : 87.32   Mean   :228.87
 3rd Qu.:2019-01-01   3rd Qu.: 87.50   3rd Qu.: 98.75   3rd Qu.:326.44
 Max.   :2019-12-31   Max.   :101.98   Max.   :121.28   Max.   :418.97
   GOOG.Close        CRM.Close          UPS.Close         MSFT.Close
 Min.   :33.41   Min.   : 54.05   Min.   : 88.7   Min.   : 48.43
 1st Qu.:39.90   1st Qu.: 81.47   1st Qu.:105.7   1st Qu.: 62.69
 Median :51.42   Median :107.00   Median :110.0   Median : 86.15
 Mean   :49.58   Mean   :113.04   Mean   :110.5   Mean   : 89.67
 3rd Qu.:57.31   3rd Qu.:147.57   3rd Qu.:116.3   3rd Qu.:110.06
 Max.   :68.06   Max.   :166.95   Max.   :134.1   Max.   :158.96
   AAPL.Close        COST.Close
 Min.   :22.59   Min.   :141.3
 1st Qu.:29.39   1st Qu.:159.3
 Median :41.31   Median :185.3
 Mean   :40.78   Mean   :198.2
 3rd Qu.:48.13   3rd Qu.:229.7
 Max.   :73.41   Max.   :305.2
```

Descriptive statistics for each company's closing stock price
from 2016-01-01-01 to 2020-01-01

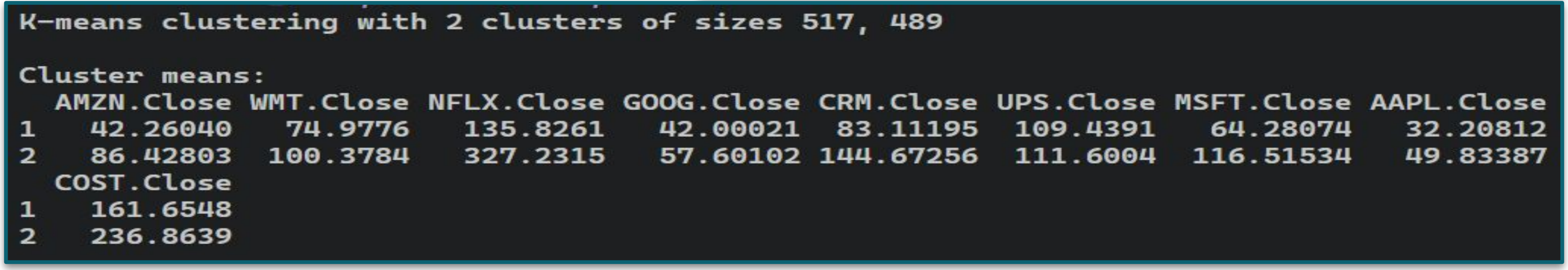
Clustering Stock Data

- The companies can be split into clusters based on stock performance
- Optimal number of clusters = 2, according the the Silhouette Method
- Being able to cluster in this way may be helpful to investors to begin to decide how to diversify portfolio



Optimal number of clusters

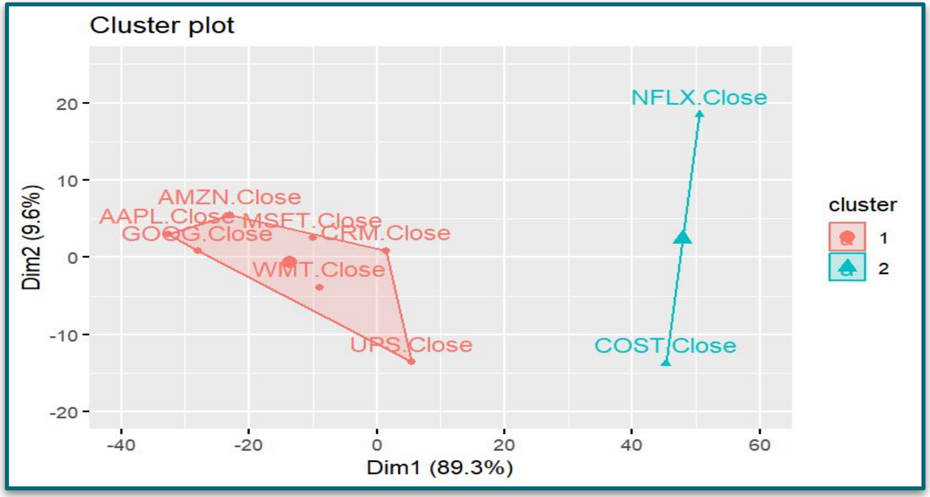
Silhouette Method

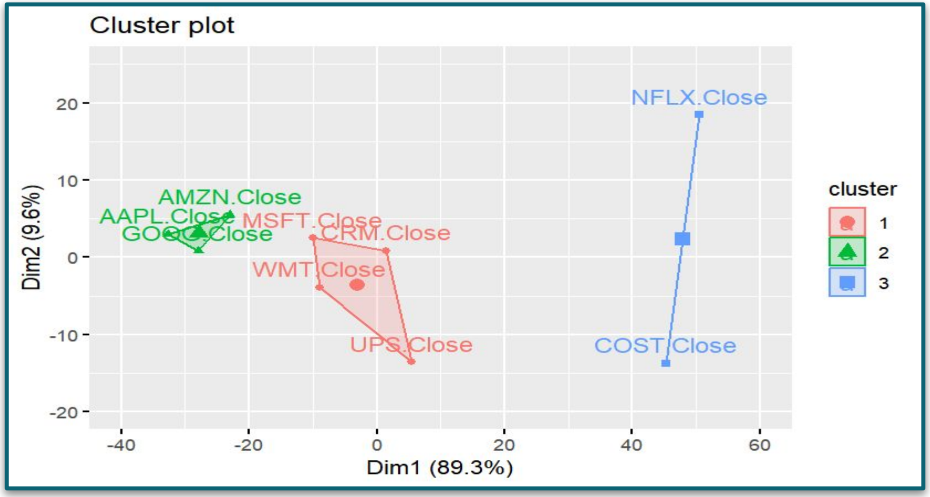


```
K-means clustering with 2 clusters of sizes 517, 489

Cluster means:
  AMZN.Close WMT.Close NFLX.Close GOOG.Close CRM.Close UPS.Close MSFT.Close AAPL.Close
1   42.26040   74.9776   135.8261   42.00021  83.11195  109.4391   64.28074   32.20812
2   86.42803  100.3784   327.2315   57.60102 144.67256  111.6004  116.51534   49.83387
  COST.Close
1   161.6548
2   236.8639
```

Top Competition

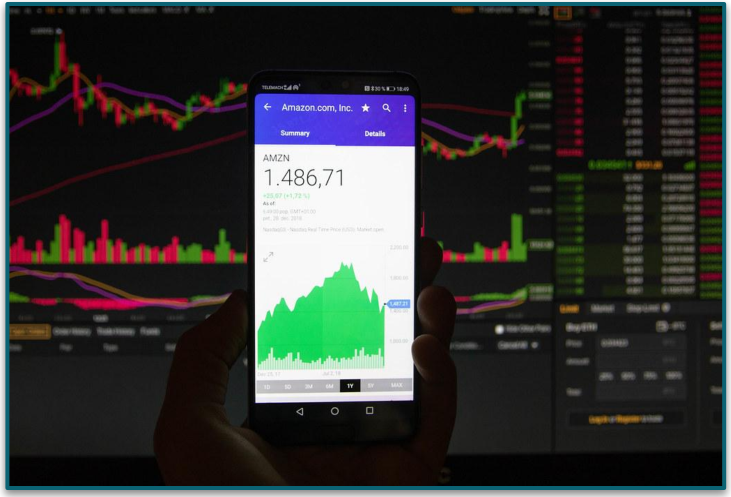
- Two clusters(optimal):
  - All companies but Netflix and Costco are comparable to AMZN stock
- Three clusters:
  - Only Google and Apple stocks are comparable to Amazon





Some Takeaways

- Seasonality and other factors can influence the performance of a stock
- Time series machine learning models may be less accurate when these factors are not taken into account
- ML models can provide high accuracy predictions when data is stationary
- K-means is a useful initial approach for investors who want to know how to best diversify their portfolio



Resources

- https://d18rn0p25nwr6d.cloudfront.net/CIK-0001018724/4d39f579-19d8-4119-b087-ee618abf82d6.pdf
- https://youtube.com/playlist?list=PLzAfHlPtM1I537hUVaqNDUBffDINzEmva
- https://medium.com/@aaronyen/https-medium-com-aaronyen-arimaproject-ab892486dc84
- https://bozliu.medium.com/financial-data-forecasting-using-r-7a55f2a1599
- https://towardsdatascience.com/interpreting-acf-and-pacf-plots-for-time-series-forecasting-af0d6db4061c
- https://www.geeksforgeeks.org/supervised-and-unsupervised-clustering-in-r-programming/
- https://www.youtube.com/watch?v=5mlth-yM2NE&t=337s
- https://towardsdatascience.com/machine-learning-for-stock-clustering-using-k-means-algorithm-126bc1ace4e1
- https://otexts.com/fpp2/index.html