**BAN 673-03**

**Time Series Analytics**

**Forecasting**

**CVS Health Stock Price**

Under the guidance of

**Professor Zinovy Radovilsky**

By

**Kishore Kumar Nakka**

**Table of Contents**

[**Summary**](#_cz26td5y2zs3) **5**

[**Introduction**](#_80wr7cz9ccpp) **6**

[**Main Chapter**](#_rkddbzaurb85) **7**

[Step 1: Define goal](#_h4i4tfp26zmu) 7

[Step 2: Get data](#_nos2qjyy6m79) 7

[Step 3: Explore & visualize series](#_sf1lrsvlszm1) 8

[Step 4: Data pre-processing](#_j1nok8eygbko) 10

[Step 5: Partition series](#_skyr9uprijow) 10

[Step 6: Apply forecasting Methods](#_y37c5uhiqkj1) 11

1. [Naive Forecasting](#_9yb9man0xm3b) 11
2. [Seasonal Naive Forecasting](#_ewkgmveaudbv) 11
3. [Moving Average](#_e5t1fgnp09rq) 11
4. [Advanced Exponential Smoothing (Holt-Winters Model)](#_c21vwwxhn3vn) 12
5. [Regression based models](#_tb089643q1lb) 13

[i. Regression model with linear trend](#_5v5zgeqj8vwy) 13

[ii. Regression model with quadratic trend](#_d47uubvex4th) 14

[iii. Regression model with seasonality](#_ot0jcobzomh9) 15

[iv. Regression model with linear trend and seasonality](#_7mnblo75ygu9) 16

[v. Regression model with quadratic trend and seasonality.](#_3mb01nanufhr) 17

[Comparing all these regression models](#_hmtdpu5ovuli) 18

1. [Auto - ARIMA](#_yst0v150wq7k) 18

[Step 7: Evaluate & compare performance](#_3glshqz6awxv) 20

[Step 8: Implement forecast system](#_ge95i2a0coi0) 20

[**Conclusion**](#_4re2ri89oj4k) **20**

[**Bibliography**](#_bqgyyb2qhhe6) **21**

[**Appendices**](#_zgxjpigw1m8c) **22**

# Summary

This is a project on Time Series Analysis and Forecasting to predict CVS Health[1] stock[3] using R Studio[2]. All the steps of the Time series forecasting methods were followed starting from data selection and exploring, visualizing the series, evaluating predictability, pre-processing of data which was not required, partitioning of time series, followed by generating some forecasting model, comparing the results of these models and then implementing the best model to forecast the future data, and then the conclusions.

The methods used for forecasting / model generation were:

1. Naive and Seasonal Naive
2. Moving Average - Trailing (with 6 different window widths as follows: 2, 4, 6, 8, 12)
3. Advanced Exponential Smoothing using Holt-Winters method
4. Regression models with (a) Linear Trend, (b) Quadratic Trend, (c) Seasonality, (d) Linear Trend and Seasonality, and (e) Quadratic Trend and Seasonality
5. Auto ARIMA.

From all the models that we developed, the best model we can use for prediction is the Regression model with Linear Trend and seasonality which has 10.261 as RMSE (Root Mean Square Error) and 11.134 as MAPE (Mean Absolute Percentage Error).

# Introduction

The CVS Health stock data for four years was taken. CVS Health Corporation provides health services and plans in the United States.The company was formerly known as CVS Caremark Corporation and changed its name to CVS Health Corporation in September 2014. The company was founded in 1963 and is headquartered in Woonsocket, Rhode Island.

The stock price was $1.59 in February of 1973. On December 1st of 2020, the stock price is $67.54.

Everyday’s closing price of CVS Health stock from January of 2016 through December of 2019 is taken from the website here: <https://finance.yahoo.com/quote/CVS/history>

In this project, I used the R Studio, a programming language to perform a time series analysis for CVS Health stock analysis. The aim is to find a good model that could be used to forecast the future values. I have used many modeling techniques for this project which will be discussed further in the paper.

# Main Chapter

This is a project on time series analysis and forecasting. The steps followed were: data selection & exploring, visualizing the series, evaluating predictability, pre-processing of data, partitioning of time series, followed by generating numerous forecasting model, comparing the results of these models and then implementing the best model(s) for forecasting of data for future, and then the conclusions.

## Step 1: Define goal

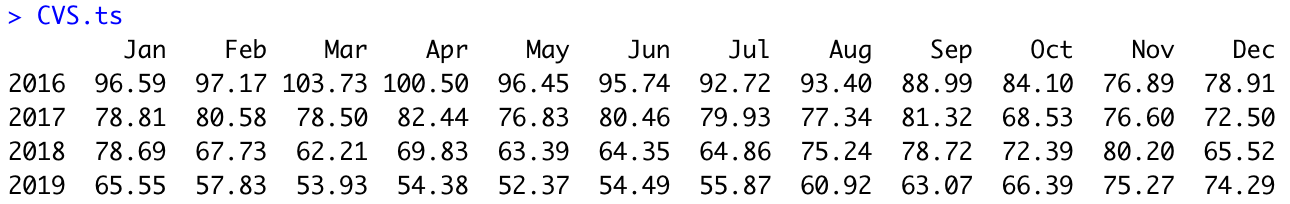
The goal of this project is to predict the CVS stock price. The resulting forecasts will be used to monitor CVS stock price. The forecasting models developed for this project were done via the R language.

## Step 2: Get data

Everyday’s closing price of CVS Health stock from January of 2016 through December of 2019 is taken from the website here: <https://finance.yahoo.com/quote/CVS/history>

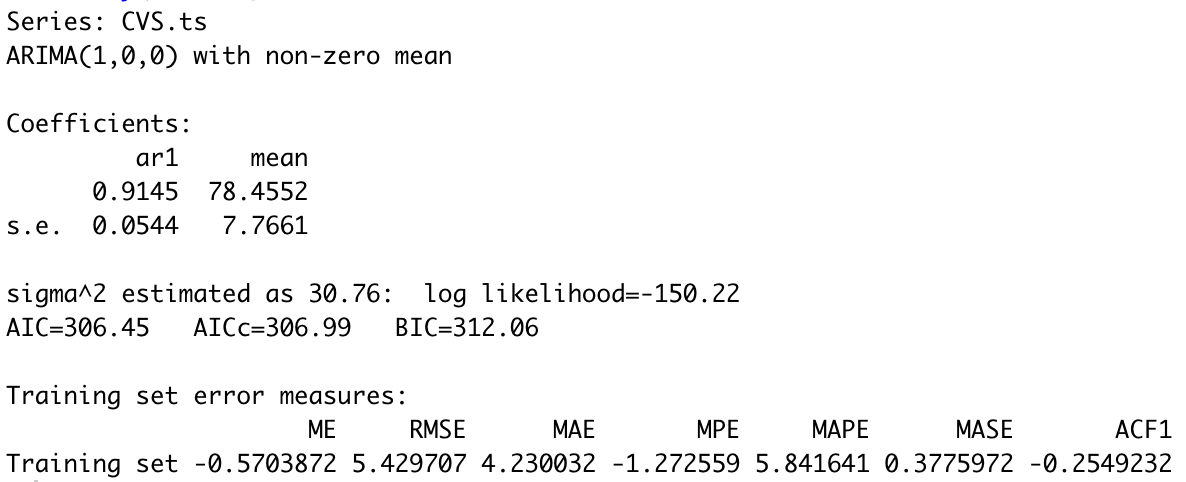
The data had other factors too, but the relevant column was taken, saved as a csv file (CVS.csv) used as input for this analysis and forecasting.

Here is how the data looks:



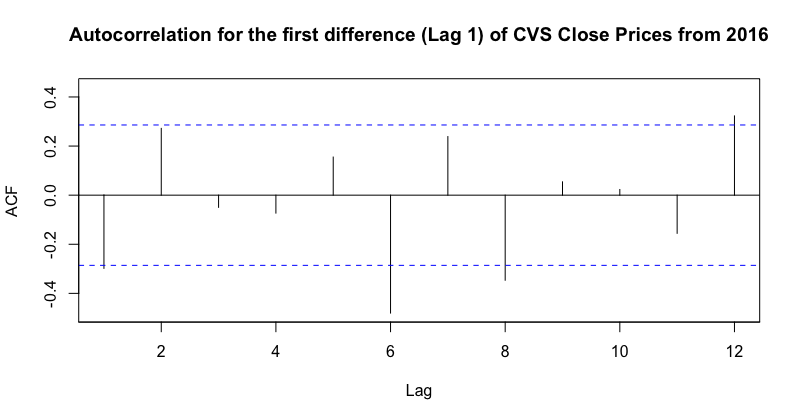
## Step 3: Explore & visualize series

Firstly, let us see if the data is predictable or not. From the summary, it is seen that the ar1 value is 0.91 that is kind of close to 1. Which basically means that it might not be a random walk and the future is predictable. But, the data we are taking is stock data. So, this is common for data like this.



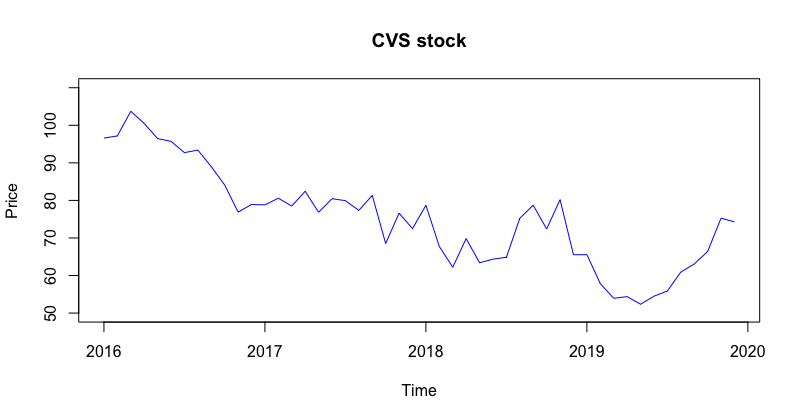
The model’s equation is: et = 78.45 + 0.91 et-1

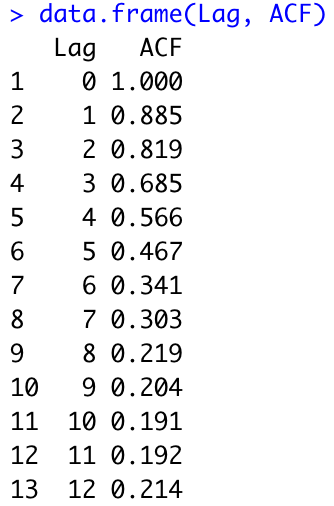
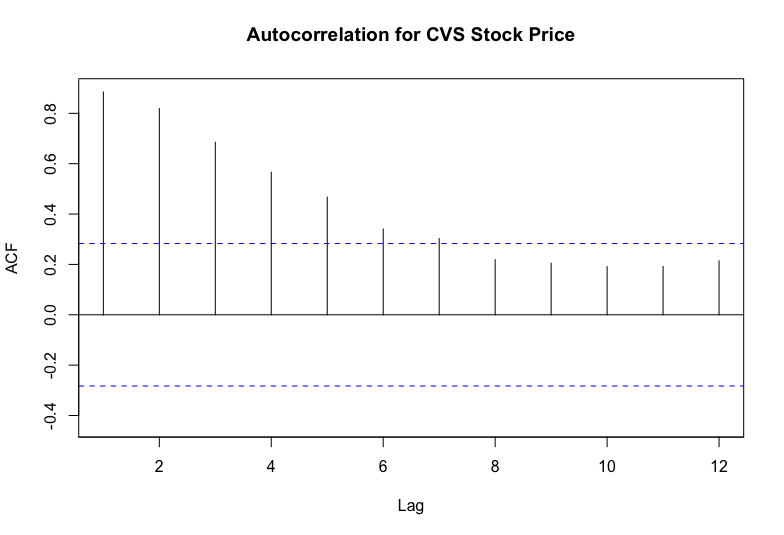
The coefficient of the ar1 (Yt-1) variable, 0.91, is below 1. Therefore, the data time series might be likely to be predictable and might not be a random walk. Let’s analyse further.



By the above chart, we can say that the data is not a random walk and the future is predictable. We see a strong negative autocorrelation at lag 6 saying that it has a half yearly seasonality.

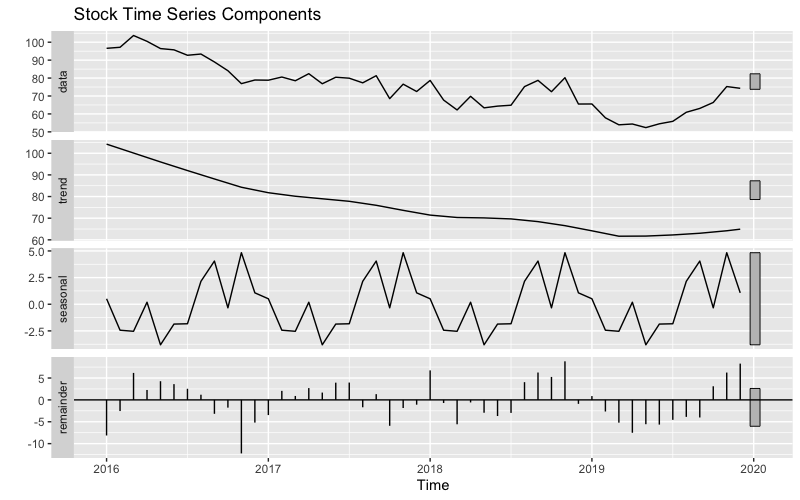
The time series data plotted below is the CVS stock data from 2016 to 2019 which appear to have a downward trend.



From the below auto-correlation chart and table, we see that the data is highly correlated, as the autocorrelation coefficients in all the lags are substantially higher than the horizontal threshold (significantly greater than zero). 

We can also say that the auto-correlation is very high with a lag of 1 and it decreases respectively for the further lags. From this, we can visualize, it has a strong trend relationship.

Here is a plot of the time-series components which shows the trend, seasonality and noise.



The above plot shows the data has a downward trend.

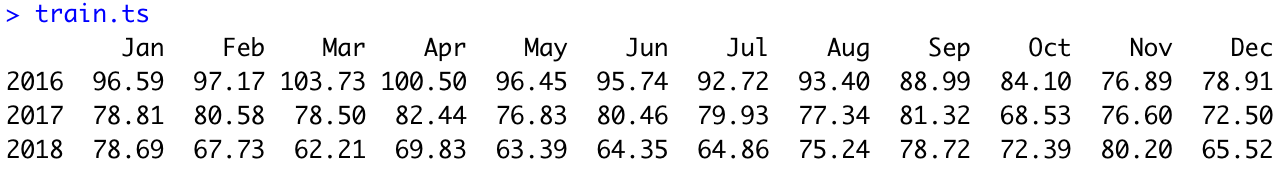
## Step 4: Data pre-processing

We need not do any pre-processing for this data.

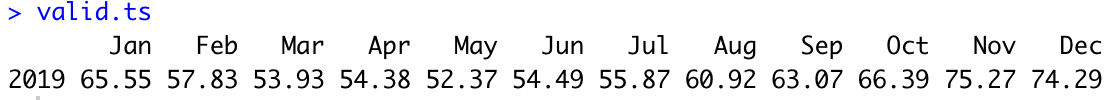
## Step 5: Partition series

The data was partitioned to train data and validation data.

Train data has 3 years of the data. From 2016 to 2018. It is shown as below.



Validation data has the year 2019. It is shown as below.



## Step 6: Apply forecasting Methods

### Naive Forecasting

### This shows the RMSE value of 8.685 and MAPE of 12.95

### Seasonal Naive Forecasting

### This shows the RMSE value of 11.048 and MAPE of 17.656.

### Moving Average

### Trailing Moving Averages were generated using rollmean() function with window widths of 2, 4, 5, 6, 8, and 12. The lowest values of MAPE and RMSE are for the window width of 2. The RMSE is 2.775 and MAPE is 3.021. This is the best till now.

### Advanced Exponential Smoothing (Holt-Winters Model)

### Holt-Winters Model was used with ets() function and model = “ZZZ” to get the optimum model selected by the system for error trend and seasonality. The model is as shown below with Additive error, no trend and no seasonality (A,N,N) . The alpha value is 0.62. It indicates the level component of the model. Here is the accuracy measure for the Holt-Winters Model. This is the accuracy of this model. It has RMSE of 11.713 and MAPE of 18.358.

### Regression based models

### Below are the summaries for the regression models developed later we will compare and get the better results.

#### i. Regression model with linear trend

### The Regression model with linear trend has an Adjusted R-squared value of 0.7 which says that the model accounts for 70% of the variations. Also, the p-value is less than 0.01. So, it is statistically significant. This regression model with linear trend has only one variable, period index (t). The equation for this is as follows: Y​t​ = 95.57 - 0.89 (t)

#### ii. Regression model with quadratic trend

### The Regression model with linear trend has an Adjusted R-squared value of approximately 0.796 which says that the model accounts for 79.6% of the variations. The p-value of this regression model indicates that it is statistically insignificant because it is <0.01 This regression model with linear trend has two independent variables, period index (t), and squared period index squared (t2​​). The equation for this is as follows: Y​t​ = 104.98 - 2.22 (t) - 0.03 (t2​)

### 

#### iii. Regression model with seasonality

### The Regression model with seasonality has an Adjusted R-squared value of -0.3. And the p-value is 0.98 (greater than 0.05) which states the model is statistically insignificant. This regression model with seasonality contains 11 seasonal variables season2 (D2), season3 (D3), and so on upto season12 (D12) The equation for this is as follows: Y​t​ = 84.697 - 2.87 (D​2)​ - 3.217 (D​3)​ - 0.44 (D​4) + ….. - 6.8 (D11) - 12.387 (D12)​

#### iv. Regression model with linear trend and seasonality

### The Regression model with seasonality has an Adjusted R-squared value of 0.63. And the p-value is 0.00012, which states the model is statistically significant. This regression model with seasonality contains 11 seasonal variables season2 (D2), season3 (D3), and so on upto season12 (D12) The equation for this is as follows: Y​t​ = 96.525 - 0.909 (t) - 1.96 (D​2)​ - 1.396 (D​3)​ - ….. - 2.377 (D12)​

#### v. Regression model with quadratic trend and seasonality.

### The Regression model with seasonality has an Adjusted R-squared value of 0.78. And the p-value is less than 0.01, which states the model is statistically significant. This regression model with seasonality contains 11 seasonal variables season2 (D2), season3 (D3), and so on upto season12 (D12) The equation for this is as follows: Y​t​ = 104.56 - 2.287 (t) - 0.037 (t2) - 1.587 (D​2)​ - 0.726 (D​3)​ - ….. - 2.377 (D12)​

#### Comparing all these regression models

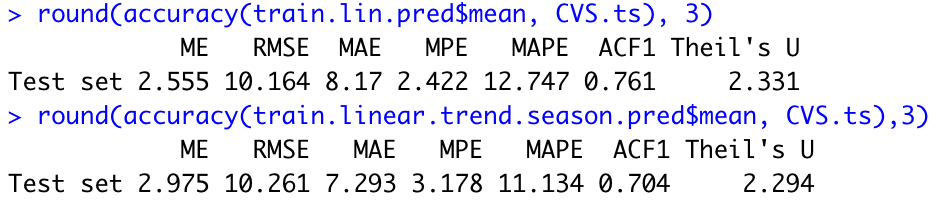
### From the above table, we can see that the RMSE and MAPE values of the regression model with linear trend and seasonality have the lowest values as 10.261 and 11.134. The next best is the regression model with linear trend with RMSE value of 10.164 and MAPE of 12.747.

### Auto - ARIMA

### Auto-ARIMA model is the model which is used to identify optimal ARIMA model and its perspective p, d, q parameters which indicate level, trend and seasonality. Here is the summary for the Auto-Arima model The equation for this can be given as below: Yt - Yt-1 = - 0.46 (Yt-1 - Yt-2) This the accuracy values of the auto - ARIMA. The RMSE is 11.740 and MAPE value is 18.513.

## Step 7: Evaluate & compare performance

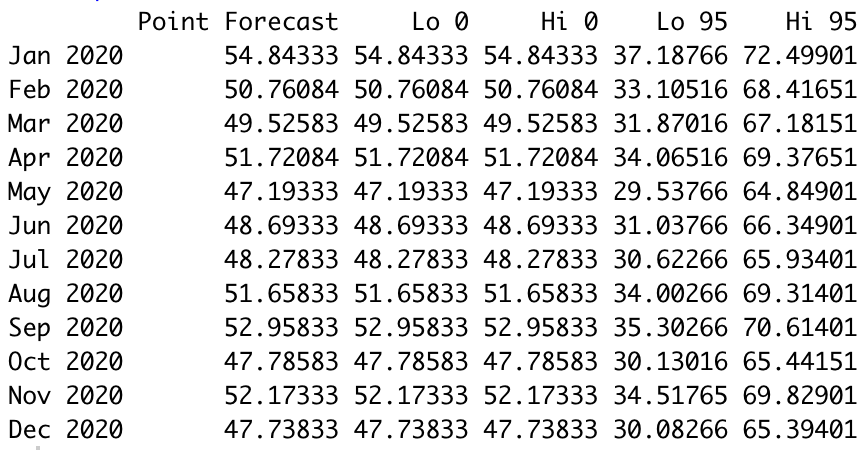
Below are the accuracy measures of the two best models. Regression model with linear trend, and Regression model with linear trend and seasonality



From these compared accuracies, we can say that both are really close in values. But, we can see that the Regression model with linear trend and seasonality has slightly lower values of the ME, MAE, ACF1 and Theil’s U. So, we choose the next best one as a regression model with linear trend and seasonality. It has the RMSE value of 10.261 and MAPE of 11.134.

## Step 8: Implement forecast system

We have used a Regression model with linear trend and seasonality to implement the forecast.



From the result above, we can see the prediction for the CVS Health of the year 2020.

# Conclusion

This project on time series analysis and forecasting on **CVS Health stocks** can be predicted best using the **regression model with linear trend and seasonality**. The model has the **RMSE value of 10.261 and MAPE of 11.134**.

Although the moving average and naive models have better RMSE and MAPE values, we do not get the right prediction values for this data. Not everything can be forecast reliably, if the factors that relate to what is being forecast are known and well understood and there is a significant amount of data that can be used, very reliable forecasts can often be obtained. If this is not the case or if the actual outcome is affected by the forecasts, the reliability of the forecasts can be significantly lower. And so, we had to choose the regression model with the linear trend model.

This project was a great learning opportunity to me because I have learnt a lot by the challenges I have faced in accomplishing the goals and overcoming the challenges and understanding the subject.

# Bibliography

[1] CVS Health is the name of the company that was referred in this document.   
More details can be seen here: <https://en.wikipedia.org/wiki/CVS_Health>

[2] R Studio is the programming language mainly used for statistical analysis.

More details are here: <https://en.wikipedia.org/wiki/RStudio>  
  
[3] Stock is the price of a share.   
More can be known here: <https://en.wikipedia.org/wiki/Stock>

# Appendices

1. PPT’s and other study material provided in the Time Series course by **Dr. Zinovy Radivosky** at California State University, East Bay.
2. Data from the wikipedia website to know about CVS.   
   Link given here: <https://en.wikipedia.org/wiki/CVS_Health>
3. Data used from Yahoo website to do the analysis.   
   Link provided here: <https://finance.yahoo.com/quote/CVS/>