A Time Series Analysis-Based Forecasting for the Amazon Revenue

A project report submitted in fulfilment of the requirement for the

BAN 673 – Time Series Analytics

Submitted by
Abhisha Burande (rq5588),
Anshika Sharma (fn5583),
Maitreyee Das (xf7587),
Priyanka Kushwaha (dv7467)

Master of Science in Business Analytics California State University, East Bay
Under the guidance of

Dr. Zinovy Radovilsky

College of Business & Economics

California State University, East Bay

CONTENTS

Abstract

This research study was designed to analyze the time-series based forecasting for the Amazon Revenue. At present, there is an incessant increase in demand for online shopping and Amazon is the biggest e-commerce business tycoon. So, a time series dataset of Amazon quarterly revenue from Q1-2005 to Q3-2020 was collected to fulfill the purpose of the study. The main objective of the study was to use Amazon's Quarterly revenue data to predict the dynamics of the same data in the future that is forecasting revenue from Q4-2020 to Q3-2021. Further, the time series dataset was partitioned into training and validation dataset and then former was used to build five different models and latter was used to validate the accuracy of the models. The five different models used were two-level (regression and trailing MA), Holt-winters', Regression based models, autoregressive and ARIMA models. Comparing RMSE and MAPE for these models, two best models (ARIMA and Holt-winters') with least errors were selected and fitted on the entire dataset for the forecasting. As per the analysis of the forecast, the model predicts an increase in revenue. Furthermore, exponential trend with an external variable (US-GDP) was used for multivariate analysis and future forecasting. Based on this forecasting, recommendations have been provided for making better strategic decisions.

1.Introduction

1.1. Company

Amazon.com, Inc. is a multinational company founded by Jeff Bezos in Seattle, Washington, on July 5,1994. The company specializes in e-commerce, cloud computing, digital streaming, and artificial intelligence. It is one of the top five companies in the information technology industry in the USA.

Company’s rapid growth triggered several acquisitions including Whole Foods Market for $13.4 billion. Bezos announced in 2018 that its Amazon Prime service (two-day delivery) crossed around 100 million subscribers worldwide. In November 2020, Amazon.com commenced delivery of prescription drugs opening a new competitive front against CVS and Walgreens. Amazon Web Services rents data storage and computing resources over the internet. In 2012, one percent of total internet traffic in North America traveled in and out of Amazon.com data centres which indicates the company's substantial online presence.

In 2018, Amazon.com was ranked 8th on the Fortune 500 rankings of the largest US corporations by total revenue. The company reported US$232.887 billion annual revenue with an increase of 30.9% compared to previous fiscal year. As of 2007, incessant expansion of the company has led to an increase in sales from 14.835 billion to 232.887 billion. In early February 2020, market capitalization was more than US$1 trillion after fourth quarter 2019 results. Revenue forecasts of such companies highly impact companies' growth and market outlook. Accurate revenue forecasts would help companies improve existing strategies and in introducing better strategies for maximizing profit.

1.2. Problem Statement:

Amazon planned to spend $4 billion (expected Quarter 2 profit) on COVID-19 related expenses. Company projected a loss of $1.5 billion in Quarter 2 of 2020; on the contrary, it generated total net sales of almost $96.15 billion exceeding its net sales of $69.98 billion in the same quarter of 2019. Earlier this year, even Jeff Bezos warned investors of a decrease in the company's revenue. This necessitates a better revenue projection for the company.

In this project, we have used a dataset of Amazon's quarterly revenue from Quarter 1 of 2005 to Quarter 3 of 2020. We will apply various time series models to the data set and finally select the most accurate one to forecast Amazon's future revenue from Q4 of 2020 to Q3 of 2021.

1.3. Proposed Solution:

We will train the following time series models on a training set for forecasting revenue on a validation set. Then, choose the best model to forecast future revenue using the entire data set.

1.3.1 Trailing Moving Average:

The revenue at time (t) will be calculated as the average of the raw historical observations at and before the time (t). This model only uses historical observations and thus it is viable for forecasting.

1.3.2 Holt-Winters' Model:

The Amazon data set demonstrates trend and seasonality. Hence, we will use Holt-Winters' exponential smoothing model as it incorporates both trend and seasonality variation.

1.3.3 Regression-based Models:

Regression based models can be used for fitting linear, exponential, and polynomial trends. Additive and multiplicative seasonality components can also be incorporated in the model. It is also useful in multi-period forecasting with an external variable.

1.3.4 Autoregressive Model:

A wide range of time series patterns can be handled remarkably using an autoregressive model. It predicts variable of interest based on past observations. The model is very useful when there is correlation between values in the historical data.

1.3.5 ARIMA Model:

This model predicts future points while considering correlation between historical data with a specific lag, differencing of the values to eliminate non-stationarity, and error lags.

1.4. Amazon Time Series Dataset:

We are considering a multivariate time series dataset as the dataset consists of Amazon's quarterly revenue from Q1-2005 to Q3-2020 and another time dependent variable U.S. GDP. Revenue not only depends on past values but also has some dependency on GDP. Therefore, we can use this dependency for forecasting Amazon's future revenue.

Variables in the dataset are:

- ➔ **Quarter:** Quarter 1 of 2005 to Quarter 3 of 2020
- ➔ **Revenue:** Quarterly revenue in US $Billion
- ➔ **US GDP:** Gross Domestic Product in US $Billion

2. Main Chapter:

2.1. Step 1: Define Goal

Our main goal is to forecast Amazon's quarterly revenue from Q4-2020 to Q3-2021. We identified time series components (level, trend, seasonality) in the data and extrapolated it to predict future revenue using various time series techniques. We worked on finding out the most parsimonious model with least number of parameters to incorporate the identified patterns in the historical data. Our major goal was to use the model with highest accuracy to predict future revenue, which would corroborate in making more informed decisions. Moreover, we utilized an external variable, US GDP, to demonstrate its impact on the main goal - forecasting revenue.

2.2. Step 2: Data Collection:

We collected Amazon quarterly revenue (Q1-2005 to Q3-2020) dataset from data.world. The dataset consists of 63 historical data observations.

Additionally, we extracted data for external variables, U.S. GDP, from Federal Reserve Economic Data. The external variable was used for multivariate time series analysis considering the high correlation with revenue.

2.3. Step 3: Explore and Visualize series:

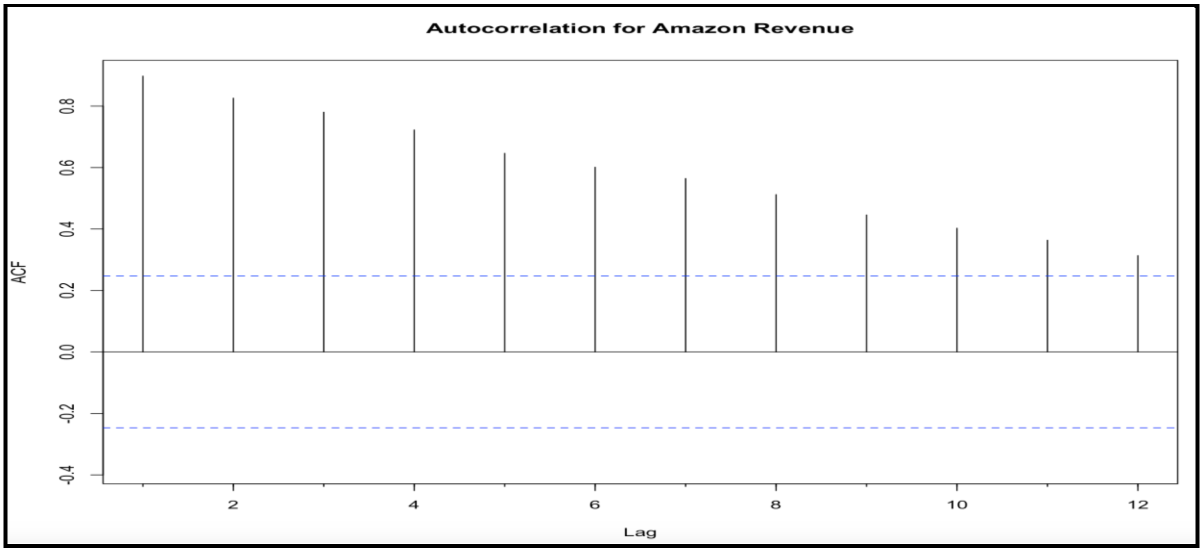
2.3.1 Create time series for the dataset:

We used data frame, Amazon_final.data, to create time series Amazon.ts using ts() function.

```
> Amazon.ts <- ts(Amazon_final.data$Revenue,
+                 start = c(2005, 1), end = c(2020, 3), freq = 4)
> Amazon.ts
      Qtr1  Qtr2  Qtr3  Qtr4
2005  1902  1753  1858  2977
2006  2279  2139  2307  3986
2007  3015  2886  3262  5672
2008  4135  4063  4264  6704
2009  4889  4651  5449  9520
2010  7131  6566  7560 12947
2011  9857  9913 10876 17431
2012 13185 12834 13806 21268
2013 16070 15704 17092 25586
2014 19741 19340 20579 29328
2015 22717 23185 25358 35746
2016 29128 30404 32714 43741
2017 35714 37955 43744 60453
2018 51042 52886 56576 72383
2019 59700 63404 69981 87437
2020 75452 88912 96145
```

2.3.2 Identify autocorrelation in the dataset:

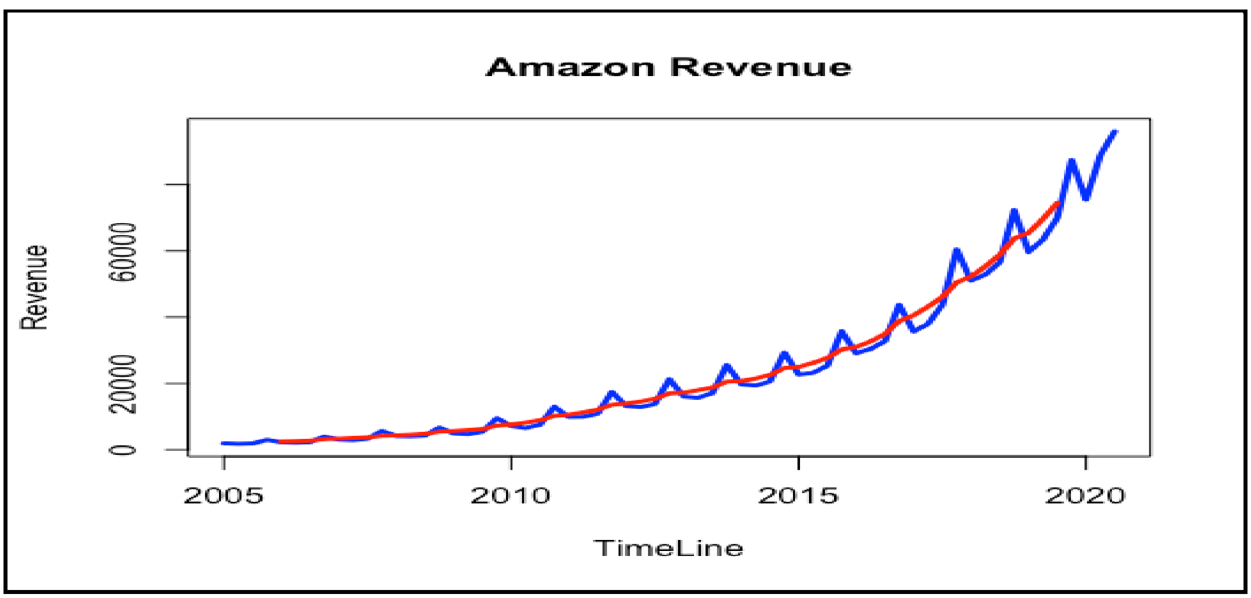
Using acf() function on the historical data with maximum 12 lags we created a correlogram to check whether observations are autocorrelated.



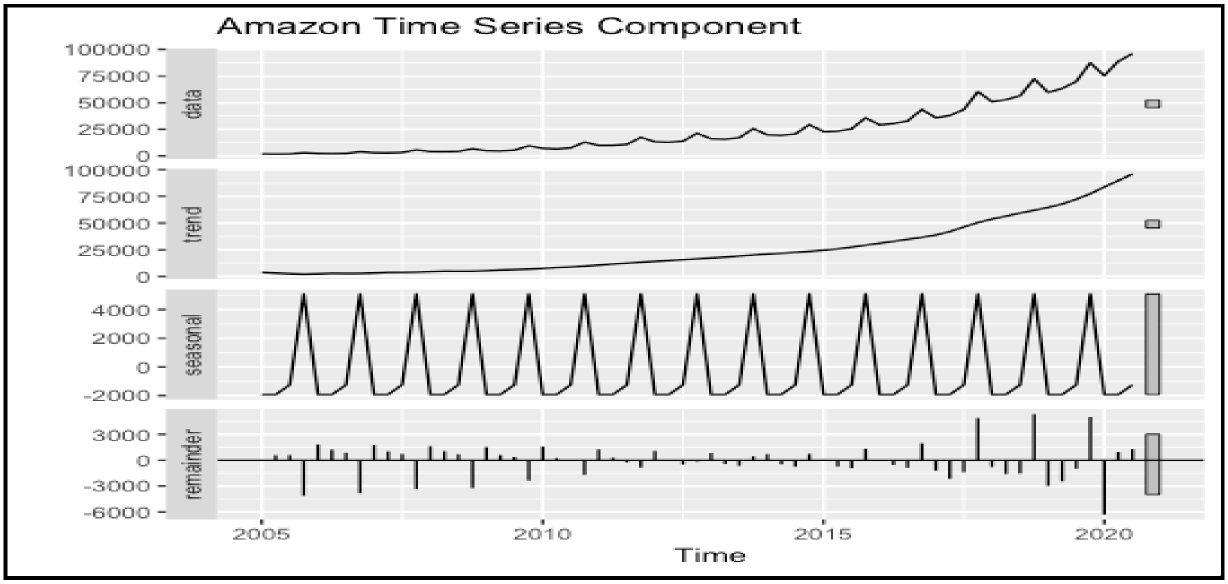
From the above graph, we can see that there is significant autocorrelation in all the lags. Lag 1 has significant autocorrelation which indicates trend in the data. Lag 12 is above the horizontal threshold which denotes possibility of seasonality in the data. Lag 4 and lag 8 are statistically significant which indicates high autocorrelation in quarterly data.

2.3.3 Visualize time series historical data:

Using plot() we visualized historical data to identify time series patterns in the data.



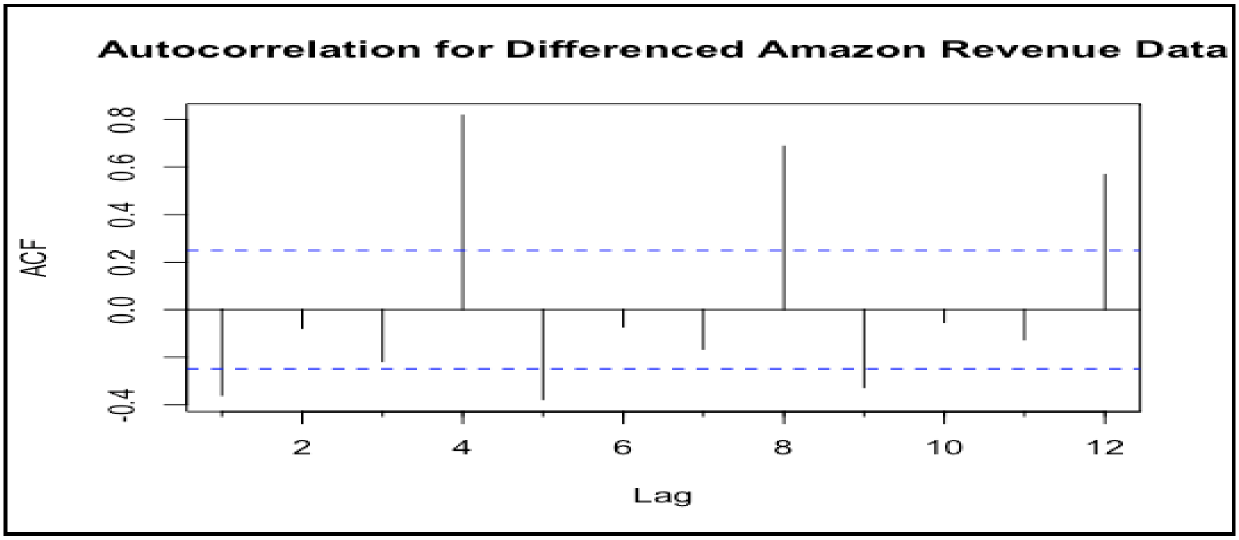
The above plot shows an upward exponential trend with seasonality. However, the plot does not display any outliers. Revenue seems to have an increasing trend over the years.



We used stl() function to decompose time series data into three components. The above chart consists of an original data, upward trend, and seasonality. It also displays level and noise components. Looking at the grey bar on the right side of the chart, we can conclude that the seasonal component has smaller variance (high bar) while the trend component has higher variance (small bar).

2.3.4 Check predictability of the differenced Amazon revenue data:

Using the first differencing (lag-1) of the difference data and Acf() function we created a correlogram with a maximum of 12 lags to check predictability of the dataset.



From the above correlogram, we can see that there is significant autocorrelation in differenced data. Lag 4 and lag 8 are statistically significant which indicates high autocorrelation in quarterly data. Lag 12 is significantly autocorrelated which denotes seasonality in the differenced data. There is significant negative autocorrelation in lag 1, lag 5, and lag 9 as the coefficients are above the horizontal threshold. Hence, using the first differencing we can conclude that Amazon revenue data is predictable.

2.4. Step 4: Pre-process data:

We extracted Amazon quarterly revenue dataset from data.world. The dataset had three attributes - Quarter, Revenue, and Net Income. However, we removed Net Income from the dataset because it was not a significant attribute in our time series analysis. As per our project scope we used Quarter and Revenue for most of the univariate forecasting models. Though, for multivariate forecasting we utilized US GDP dataset collected from Federal Reserve Economic Data. Not much preprocessing was required as both the datasets were of high quality - no missing values and no data value errors.

2.5. Step 5: Partition series:

The need for partitioning time series data is to check accuracy of the data which is excluded from model development. Time series data is divided into training and validation sets. Forecasting model is developed using the training set and validation set is used for validating the model performance. Partitioning eliminates the chances of overfitting thus preventing the model from performing poorly. We partitioned the training set from Q1-2005 to Q3-2016 and validation set from Q4-2016 to Q3-2020.

Following are the partitioned series:

- Training Set:

```
> train.ts.az
```

	Qtr1	Qtr2	Qtr3	Qtr4
2005	1902	1753	1858	2977
2006	2279	2139	2307	3986
2007	3015	2886	3262	5672
2008	4135	4063	4264	6704
2009	4889	4651	5449	9520
2010	7131	6566	7560	12947
2011	9857	9913	10876	17431
2012	13185	12834	13806	21268
2013	16070	15704	17092	25586
2014	19741	19340	20579	29328
2015	22717	23185	25358	35746
2016	29128	30404	32714	

- Validation set:

```
> valid.ts.az
```

	Qtr1	Qtr2	Qtr3	Qtr4
2016				43741
2017	35714	37955	43744	60453
2018	51042	52886	56576	72383
2019	59700	63404	69981	87437
2020	75452	88912	96145	

2.6. Step 6: Apply forecasting methods:

The dataset of Amazon was analyzed by employing different forecasting methods. In this study, we took five different models to predict the future data. These models are discussed in detail in further sections.

2.6.1 Model 1: Two-level Model (Regression + MA Trailing for Residuals)

In general, trailing MA should be used for forecasting in time series that lack seasonality and trend. But as we saw in the above sections that our dataset comprises trend as well as seasonality so we will be executing a two-level model (Regression + MA Trailing for Residuals).

Objective:

To develop a two-level Trailing Moving Average Model with a window of 4 to forecast the quarterly revenue of Amazon for the validation period from Q4-2016 to Q3-2020.

Scope:

- Trailing moving averages uses only current and historical observations to predict the future.
- This method is used in time series for making predictions. Before forecasting, we assume that the trend and seasonality components of the time series have already been removed or adjusted for.

Model Execution:

A regression model with quadratic trend and seasonality was created for the training data. The summary for this model is given below:

```
> reg.trend.seas <- tslm(train.ts.az ~ trend + I(trend^2) + season)
> summary(reg.trend.seas)

Call:
tslm(formula = train.ts.az ~ trend + I(trend^2) + season)

Residuals:
     Min       1Q   Median       3Q      Max
-2912.23  -571.58    95.21   729.38  3156.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1709.397    659.156   2.593   0.0131 *
trend        -68.256     56.764  -1.202   0.2361
I(trend^2)    15.328      1.147  13.367  < 2e-16 ***
season2     -703.089    526.285  -1.336   0.1889
season3     -412.001    526.832  -0.782   0.4387
season4     4207.599    539.328   7.802 1.25e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1289 on 41 degrees of freedom
Multiple R-squared:  0.9844,    Adjusted R-squared:  0.9825
F-statistic: 516.4 on 5 and 41 DF,  p-value: < 2.2e-16
```

Model Equation:

Model Equation: y_t = 1709.397 -68.256t + 15.328 t^2 – 703.089 D_2 -412.001 D_3 – 4207.599 D_4

Results and observations of this Regression Model:

- This model of regression with quadratic trend and seasonality consists of five predictors which are trend, trend square, seasonal dummy variables for quarter 2, quarter 3 and quarter 4 i.e**.,** D_2, D_3 and D_4. As per the summary we can see that the trend square variable and dummy variable for season 4 are significant and all other variables are insignificant.
- The F-statistics p-value is 2.2*10-16 which is below 0.05. The adjusted R-squared is very high i.e., 0.9825 along with the multiple R-squared which is 0.9844. The F-statistics is 516.4 which is very high So, overall, this model is a good model.

The 2016-2020 validation data Amazon Revenue forecast using this regression model is presented below (confidence interval is not used):

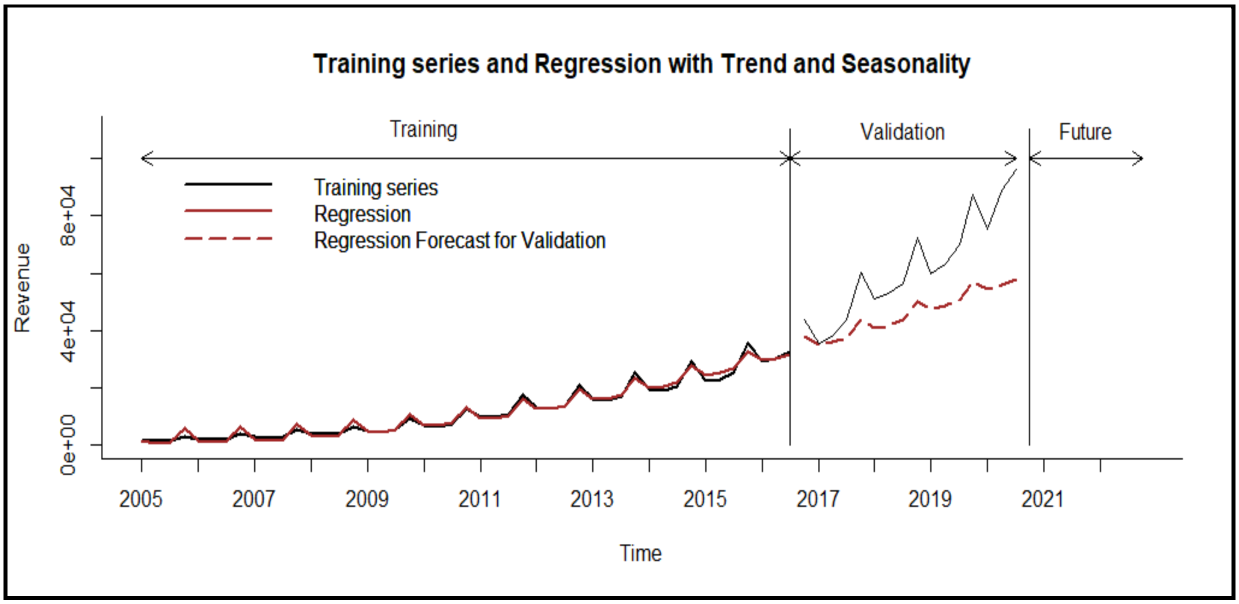
```
> reg.trend.seas.pred <- forecast(reg.trend.seas, h = 16, level = 0)
> reg.trend.seas.pred
        Point Forecast     Lo 0     Hi 0
2016 Q4       37957.02 37957.02 37957.02
2017 Q1       35168.00 35168.00 35168.00
2017 Q2       35914.16 35914.16 35914.16
2017 Q3       37685.14 37685.14 37685.14
2017 Q4       43815.30 43815.30 43815.30
2018 Q1       41148.91 41148.91 41148.91
2018 Q2       42017.69 42017.69 42017.69
2018 Q3       43911.30 43911.30 43911.30
2018 Q4       50164.08 50164.08 50164.08
2019 Q1       47620.32 47620.32 47620.32
2019 Q2       48611.72 48611.72 48611.72
2019 Q3       50627.96 50627.96 50627.96
2019 Q4       57003.36 57003.36 57003.36
2020 Q1       54582.23 54582.23 54582.23
2020 Q2       55696.26 55696.26 55696.26
2020 Q3       57835.12 57835.12 57835.12
```

The combined two-level validation forecast for Amazon Revenue is presented below.

(Two level = Regression Forecast + MA Trailing Residual Forecast):

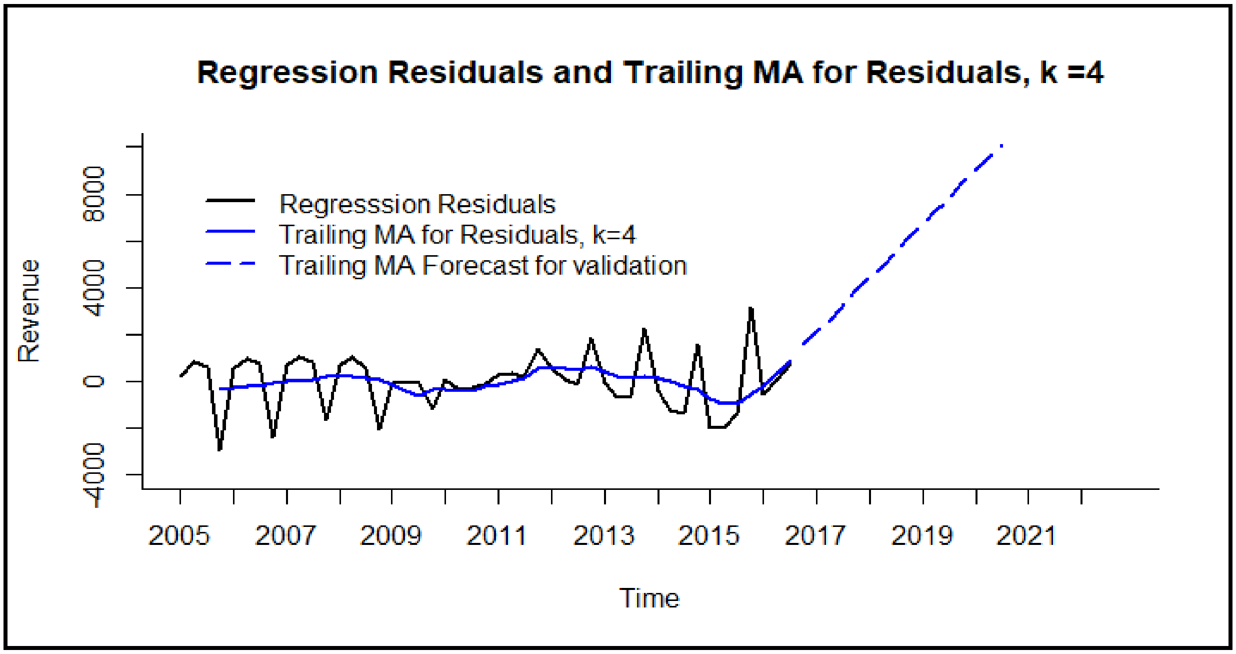
```
> total.reg.ma.pred
   reg.trend.seas.pred.mean ma.trailing.res_4.pred.mean ts.forecast.4
1                  37957.02                    1591.787      39548.80
2                  35168.00                    2128.689      37296.69
3                  35914.16                    2648.826      38562.98
4                  37685.14                    3176.546      40861.69
5                  43815.30                    3899.700      47715.00
6                  41148.91                    4436.602      45585.51
7                  42017.69                    4956.740      46974.43
8                  43911.30                    5484.460      49395.76
9                  50164.08                    6207.614      56371.69
10                 47620.32                    6744.515      54364.83
11                 48611.72                    7264.653      55876.37
12                 50627.96                    7792.373      58420.33
13                 57003.36                    8515.527      65518.89
14                 54582.23                    9052.429      63634.66
15                 55696.26                    9572.567      65268.83
16                 57835.12                   10100.286      67935.41
```

Plot of Training Series and regression with trend and seasonality:



In the above graph, a regression model is observed for the validation data set for good fit. Here, we can clearly see the lines representing the training series, regression and regression forecast for validation of 16 periods. As per the above graph, this regression forecast is under-predicting for validating because the original values are much higher than the predicted ones.

Plot of Regression Residuals and Trailing MA for Residuals, k = 4:



The above graph represents the regression residual with black line, Trailing MA for residuals with k=4 with blue line and trailing MA forecast for validation with blue dashed line. This graph shows that the trailing MA for residual of validation is increasing linearly in each quarter.

Accuracy measures on validation data:

```
> round(accuracy(reg.trend.seas.pred, valid.ts.az), 3) # RMSE=19192.227, MAPE=22.838
                  ME      RMSE       MAE    MPE   MAPE  MASE   ACF1
Training set     0.0  1203.615   915.618  2.094 14.568 0.329 -0.128
Test set     15985.4 19192.227 15985.402 22.838 22.838 5.752  0.605
             Theil's U
Training set        NA
Test set         1.624
> round(accuracy(ts.forecast.4, valid.ts.az), 3)# RMSE=13127.41, MAPE=14.622
               ME     RMSE      MAE    MPE   MAPE  ACF1 Theil's U
Test set 10137.07 13127.41 10410.91 13.868 14.622 0.486     1.101
```

As per the observations of accuracy measures, the RMSE and MAPE value for combined two level model of regression and trailing MA is 13127.41 and 14.622% respectively.

2.6.2 Model 2: Holt's Winter Model

Winter's (Holt-Winters) model is used for time series that contain level, trends (a slope) and seasonality (cyclical repeating pattern).

Objective:

Develop the most optimal Holt-Winters Model with R's automated selection of error, trend, and seasonality options to forecast the quarterly revenue of Amazon for the validation period from Q4-2016 to Q3-2020.

Scope:

- Holt-Winters model is used to forecast time series with level, trend, and seasonality. The setting of ets() function to 'ZZZ' from the forecast package evaluates all possible combinations of Error, a slope (trend) over time, and a cyclical repeating pattern (seasonality)and gives the most optimal smoothing parameters.
- Holt-Winters uses exponential smoothing to encode lots of values from the past and use them to predict level values for the present and future.

Model Execution:

Use ets() function with model value 'ZZZ' to fit Holts' winter model. Then use summary () to show Holt's winter model and its parameters.

```
> hw.ZZZ.train.az
ETS(M,A,M)

Call:
 ets(y = train.ts.az, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.6307
    beta  = 0.2164
    gamma = 0.3693

  Initial states:
    l = 1689.4215
    b = 246.1127
    s = 1.3167 0.8657 0.8562 0.9614

  sigma:  0.0564

     AIC     AICc      BIC
771.1881 776.0529 787.8394
```

A summary of the multiplicative Holt-Winters (HW) model with multiplicative error, additive trend, and multiplicative seasonality (model = "MAM") for the training period is shown above.

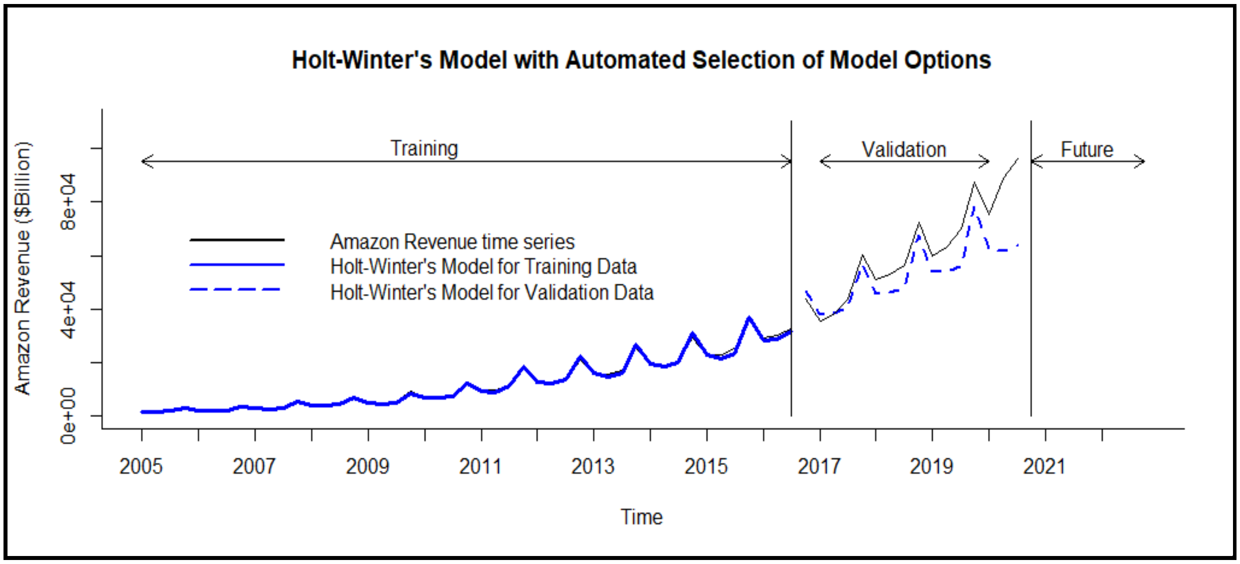
It can be seen from the model's summary that the optimal value for exponential smoothing constant (alpha) is 0.6307 the optimal smoothing constant for trend (beta) is 0.2164, and the optimal smoothing constant for seasonality estimate (gamma) is 0.3693.

The exponential smoothing constants (α, β, γ) are closer to 1 indicates that the model components tend to be more local.

Using this HW model, the point forecast in the validation period is presented below:

```
> hw.ZZZ.train.pred.az
        Point Forecast     Lo 0     Hi 0
2016 Q4       46777.04 46777.04 46777.04
2017 Q1       38020.85 38020.85 38020.85
2017 Q2       38518.60 38518.60 38518.60
2017 Q3       40314.41 40314.41 40314.41
2017 Q4       57115.26 57115.26 57115.26
2018 Q1       45990.67 45990.67 45990.67
2018 Q2       46197.26 46197.26 46197.26
2018 Q3       47976.08 47976.08 47976.08
2018 Q4       67519.21 67519.21 67519.21
2019 Q1       54013.16 54013.16 54013.16
2019 Q2       53928.58 53928.58 53928.58
2019 Q3       55692.22 55692.22 55692.22
2019 Q4       77999.54 77999.54 77999.54
2020 Q1       62096.52 62096.52 62096.52
2020 Q2       61720.47 61720.47 61720.47
2020 Q3       63470.71 63470.71 63470.71
```

Plot ts data with trend and seasonality data, and predictions for validation period.



Holt’s winter model with automated selection of Model options i.e model “MAM’ is shown above. From the graph, model prediction for validation data is under predicting.

Holts Winter Model Accuracy

```
> round(accuracy(hw.ZZZ.train.pred.az, valid.ts.az), 3) #RSME=12763.096 MAPE=13.166
                   ME      RMSE      MAE    MPE   MAPE  MASE  ACF1 Theil's U
Training set  118.419   744.407  520.890  0.839  4.255 0.187 0.113        NA
Test set     8635.902 12763.096 9374.214 11.306 13.166 3.373 0.627     0.983
```

Accuracy measures for Holt’s winter model (MAM) results in RMSE = 12763.096 and MAPE = 13.166% on Validation data.

2.6.3 Regression Model

Objective:

Develop the most optimal Regression Based time Series Forecasting Models to forecast the quarterly revenue of Amazon for the validation period from Q4-2016 to Q3-2020.

Scope:

Leverage Amazon’s historical Quarterly revenue information of past 15 years (2005-2020) to build Regression based time series models by fitting either of linear, exponential, quadratic trend or seasonality or a combination of these to forecast quarterly revenue on validation data.

Model Execution:

<u>MLR Sub Model 1: Regression model with linear trend:</u>

Objective is to fit a global trend that applies to the training time series of Amazon' historical revenue sales data and will apply in the validation period.

```
> train.az.lin <- tslm(train.ts.az ~ trend)
> summary(train.az.lin)

Call:
tslm(formula = train.ts.az ~ trend)

Residuals:
    Min      1Q  Median      3Q     Max 
-4520.9 -2764.1  -488.7  2436.2  9986.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3585.50     997.52  -3.594 0.000803 ***
trend         666.93      36.18  18.432  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3365 on 45 degrees of freedom
Multiple R-squared:  0.883,	Adjusted R-squared:  0.8804 
F-statistic: 339.7 on 1 and 45 DF,  p-value: < 2.2e-16
```

Model Equation:

$$yt = -3585.50 + 666.93\, t$$

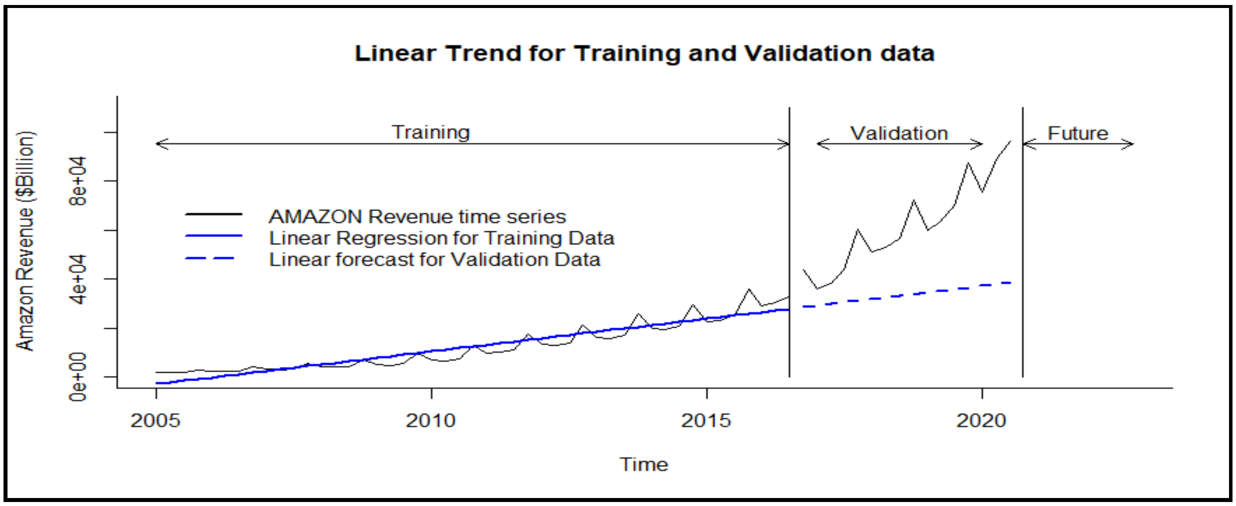
Observations for Sub Model 1:

The regression model with linear trend contains a single independent predictor which is Trend (t). Trend appears to be a significant variable for the model, with its p value below 0.001. The model's summary shows an extremely high R-squared of 0.883 and adj. R_squared of 0.8804, statistically significant F-statistic (p-value is substantially lower than 0.05), trend (t) is statistically significant (p-value <0.05). This regression model is statistically significant and a good fit for the historical data set, and thus can be used for forecast validation data.

The forecast result for validation dataset using Regression Model with Linear Trend is given below (confidence interval is not used. -

```
> train.az.lin.pred <- forecast(train.az.lin, h = nValid.az, level = 0)
> train.az.lin.pred
        Point Forecast     Lo 0     Hi 0
2016 Q4       28427.08 28427.08 28427.08
2017 Q1       29094.01 29094.01 29094.01
2017 Q2       29760.94 29760.94 29760.94
2017 Q3       30427.86 30427.86 30427.86
2017 Q4       31094.79 31094.79 31094.79
2018 Q1       31761.72 31761.72 31761.72
2018 Q2       32428.65 32428.65 32428.65
2018 Q3       33095.58 33095.58 33095.58
2018 Q4       33762.51 33762.51 33762.51
2019 Q1       34429.44 34429.44 34429.44
2019 Q2       35096.37 35096.37 35096.37
2019 Q3       35763.29 35763.29 35763.29
2019 Q4       36430.22 36430.22 36430.22
2020 Q1       37097.15 37097.15 37097.15
2020 Q2       37764.08 37764.08 37764.08
2020 Q3       38431.01 38431.01 38431.01
```

Plot of time series data with trend, and predictions for validation period



From the above graph, we can observe that the linear forecast for the revenue is significantly lower than the actual revenue in the validation period. Hence, the linear regression model is under predicting the revenue.

<u>MLR Sub Model 2: Regression model with Exponential trend:</u>

Objective is to fit an exponential trend that applies to the training time series dataset of Amazon' historical revenue sales figures and then will use the same model to forecast for validation period time series Dataset.

```
> #Model 2: Regression Model with Exponential Trend:
> train.az.expo <-tslm(train.ts.az ~ trend, lambda = 0)
> summary(train.az.expo)

Call:
tslm(formula = train.ts.az ~ trend, lambda = 0)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.22893 -0.14547 -0.05739  0.07964  0.43893 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.497812   0.056838  131.91   <2e-16 ***
trend       0.065331   0.002062   31.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1917 on 45 degrees of freedom
Multiple R-squared:  0.9571,	Adjusted R-squared:  0.9562 
F-statistic:  1004 on 1 and 45 DF,  p-value: < 2.2e-16
```

Model Equation:

$$\log (y_t) = 7.49 + 0.065\,t$$

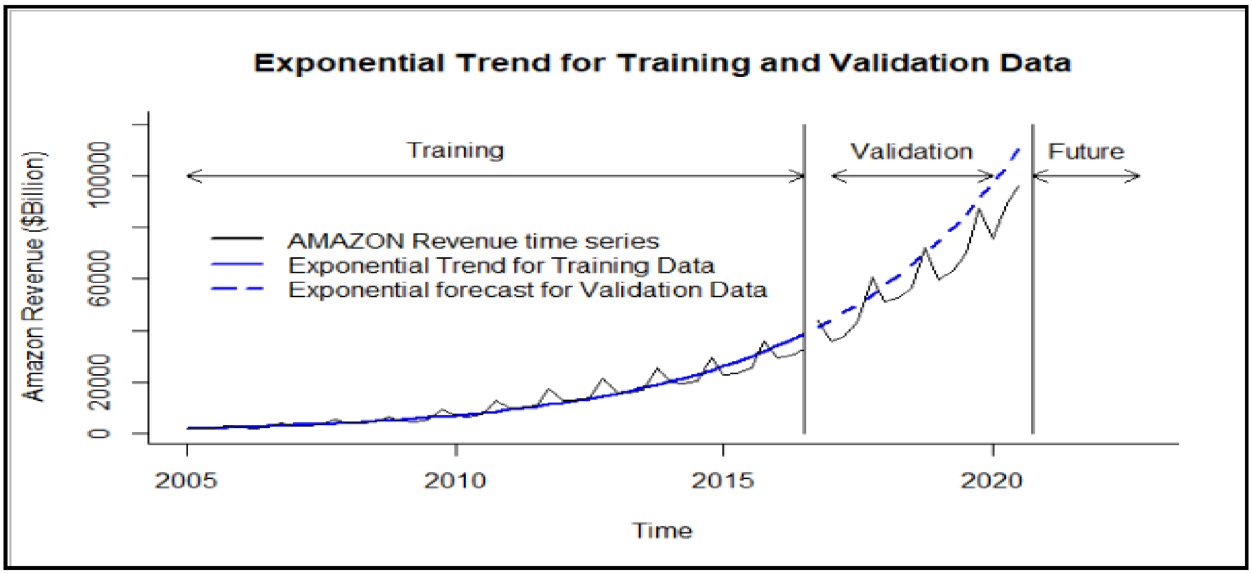
Observations for Sub Model 2:

The regression model with exponential trend contains a single independent predictor which is Trend (t). Trend appears to be a significant variable for the model, with its p value below 0. 001.The model's summary shows a very high R-squared of 0.957 and adj. R_squared of 0.9562, statistically significant F-statistic (p-value is substantially lower than 0.05), trend (t) is statistically significant (p-value <0.05).This regression model appears to be statistically significant and a good fit for the historical data set, and thus can be used for forecast .

The forecast result for validation dataset using Regression Model with Exponential Trend is given below (confidence interval is not used).-

```
> train.az.expo.pred <- forecast(train.az.expo, h = nValid.az, level = 0)
> train.az.expo.pred
        Point Forecast      Lo 0      Hi 0
2016 Q4       41510.18  41510.18  41510.18
2017 Q1       44312.63  44312.63  44312.63
2017 Q2       47304.27  47304.27  47304.27
2017 Q3       50497.88  50497.88  50497.88
2017 Q4       53907.11  53907.11  53907.11
2018 Q1       57546.49  57546.49  57546.49
2018 Q2       61431.58  61431.58  61431.58
2018 Q3       65578.96  65578.96  65578.96
2018 Q4       70006.34  70006.34  70006.34
2019 Q1       74732.62  74732.62  74732.62
2019 Q2       79777.98  79777.98  79777.98
2019 Q3       85163.97  85163.97  85163.97
2019 Q4       90913.57  90913.57  90913.57
2020 Q1       97051.35  97051.35  97051.35
2020 Q2      103603.50 103603.50 103603.50
2020 Q3      110598.00 110598.00 110598.00
```

Plot of time series data with exponential trend, and predictions for validation period:



The above graph showcases the exponential trend for training and validation data. The exponential forecast for the revenue is closer to the actual revenue for a few quarters in the validation period, For the next quarters the forecast is slightly overestimating the actual revenue.

MLR Sub Model 3: Regression model with Quadratic trend:

Objective is to fit order 2 polynomial Regression with y_t as output, and t and t^2 as predictors to the training time series dataset of Amazon' historical revenue sales figures and then use the same model to forecast revenue for validation period time series Dataset.

```
> train.az.quad <- tslm(train.ts.az~ trend + I(trend^2))
> summary(train.az.quad)

Call:
tslm(formula = train.ts.az ~ trend + I(trend^2))

Residuals:
    Min      1Q  Median      3Q     Max 
-3253.8 -1346.3  -480.5     0.6  6880.5 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2051.61    1081.71   1.897   0.0645 .  
trend         -23.33     103.95  -0.224   0.8235    
I(trend^2)     14.38       2.10   6.849 1.91e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2367 on 44 degrees of freedom
Multiple R-squared:  0.9434,	Adjusted R-squared:  0.9408 
F-statistic: 366.6 on 2 and 44 DF,  p-value: < 2.2e-16
```

Model Equation:

$$\mathbf{y_t = 2051.61 - 23.33\ t + 14.38\ t^2}$$

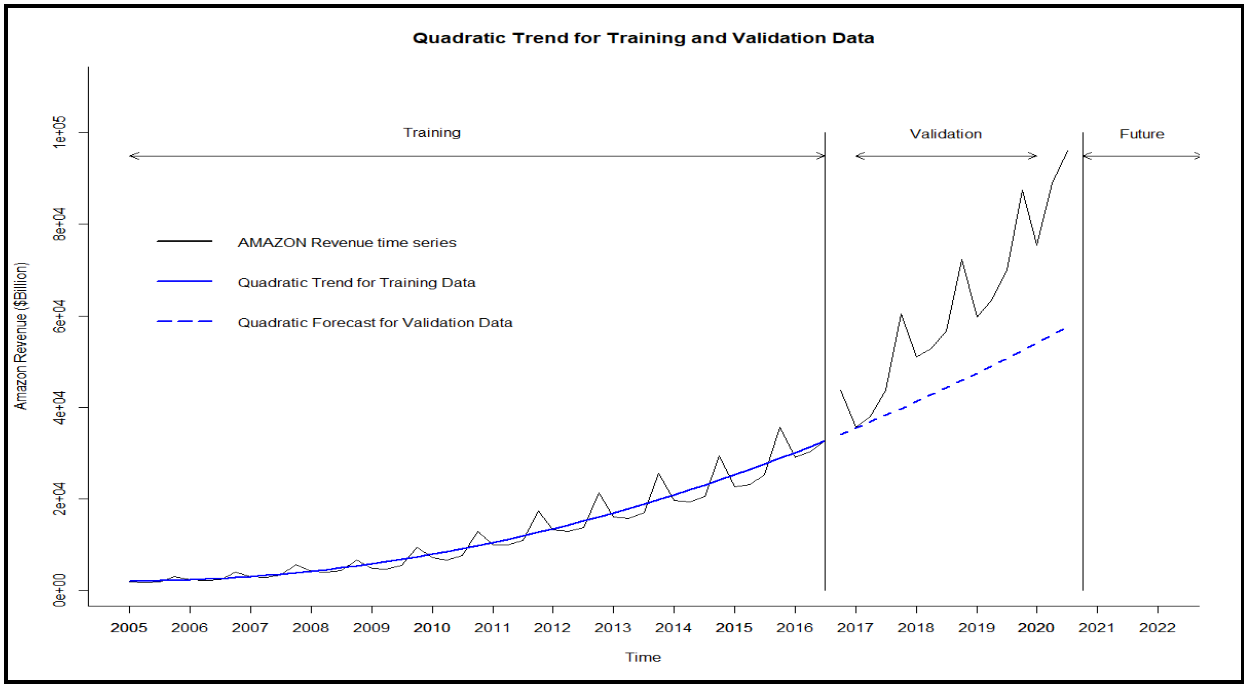
Observations for Sub Model 3:

The regression model with exponential trend contains two predictors which are Trend (t) and Trend Square(t2). Trend appears to be an Insignificant variable for the model, with its p value above 0.001 and Trend Square appears to be a significant variable for the model, with its p value below 0.1. The model's summary shows a very high R-squared of 0.943 and adj. R_squared of 0.9408, statistically significant F-statistic (p-value is substantially lower than 0.05). This regression model appears to be statistically significant and can be a good fit time series forecast .

The forecast result for validation dataset using Regression Model with Quadratic Trend is given below (confidence interval is not used).

```
> #Forecasting for Validation period:
> train.az.quad.pred <- forecast(train.az.quad, h = nValid.az, level = 0)
> train.az.quad.pred
        Point Forecast     Lo 0     Hi 0
2016 Q4       34064.20 34064.20 34064.20
2017 Q1       35435.76 35435.76 35435.76
2017 Q2       36836.09 36836.09 36836.09
2017 Q3       38265.18 38265.18 38265.18
2017 Q4       39723.03 39723.03 39723.03
2018 Q1       41209.65 41209.65 41209.65
2018 Q2       42725.02 42725.02 42725.02
2018 Q3       44269.15 44269.15 44269.15
2018 Q4       45842.05 45842.05 45842.05
2019 Q1       47443.70 47443.70 47443.70
2019 Q2       49074.12 49074.12 49074.12
2019 Q3       50733.29 50733.29 50733.29
2019 Q4       52421.23 52421.23 52421.23
2020 Q1       54137.93 54137.93 54137.93
2020 Q2       55883.39 55883.39 55883.39
2020 Q3       57657.61 57657.61 57657.61
```

Plot of time series data with quadratic trend, and predictions for validation period:



The above graph showcases the quadratic trend for training and validation data. We can observe that the regression forecast for the revenue is under predicting the actual revenue in the validation period.

MLR Sub Model 4: Regression model with Seasonality:

Objective is to fit a regression Model with Seasonality to the training time series dataset of Amazon' historical revenue figures and then use the same model to forecast revenue for validation period time series Dataset.

```
> train.az.season <- tslm(train.ts.az ~ season)
> summary(train.az.season)

Call:
tslm(formula = train.ts.az ~ season)

Residuals:
   Min     1Q Median     3Q    Max
-12584  -8195  -2614   6964  20620

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11170.75    2855.47   3.912 0.000321 ***
season2       -50.92    4038.24  -0.013 0.989998
season3       923.00    4038.24   0.229 0.820291
season4      4389.70    4129.00   1.063 0.293653
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9892 on 43 degrees of freedom
Multiple R-squared:  0.03415,	Adjusted R-squared:  -0.03323
F-statistic: 0.5068 on 3 and 43 DF,  p-value: 0.6796
```

Model Equation:

$$\mathbf{y_t = 11170.75 - 50.92\ D2 + 923\ D3 + 4389.70\ D4}$$

Observations for Sub Model 4:

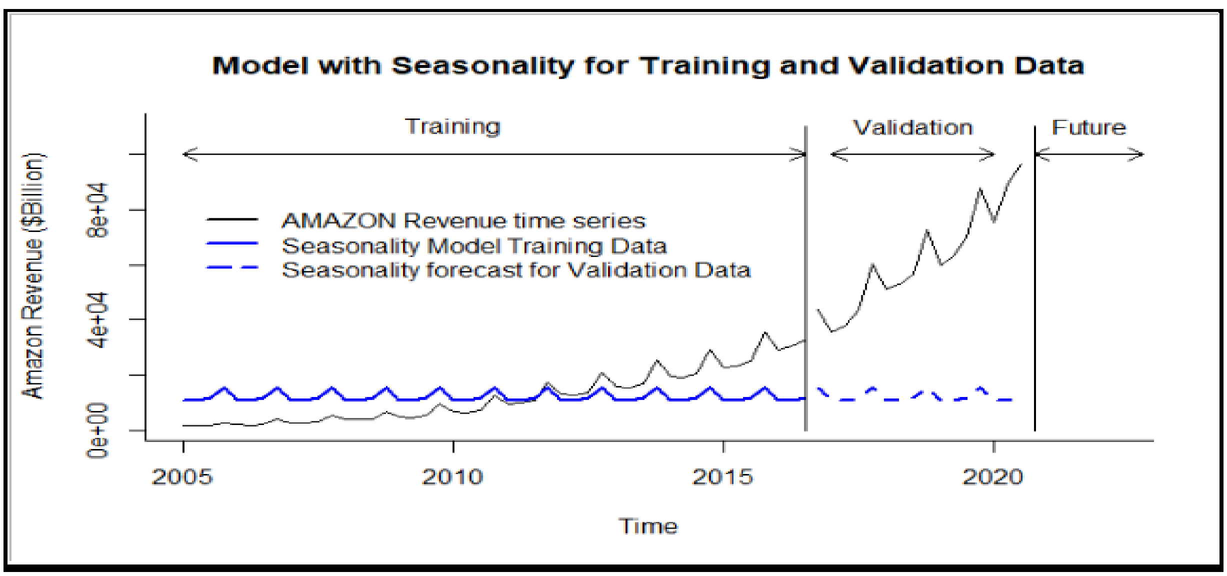
The regression model with Seasonality consists of 3 predictors which are seasonal dummy variables for Quarter 2(Season2), Quarter 3(Season 3) and Quarter 4 (Season 4). All season variables do not appear to be significant variables for the model, with their p values being very high and not statistically significant. The model's summary shows a very low R-squared of

0.03415 and adj. R_squared of -0.03323, pretty low F-statistic (p-value is substantially higher than 0.05). This regression model appears to be statistically Insignificant and cannot be a good fit time-series forecast .

The forecast result for validation dataset using Regression Model with Seasonality is given below (confidence interval is not used).

```
> train.az.season.pred
        Point Forecast    Lo 0     Hi 0
2016 Q4       15560.45 15560.45 15560.45
2017 Q1       11170.75 11170.75 11170.75
2017 Q2       11119.83 11119.83 11119.83
2017 Q3       12093.75 12093.75 12093.75
2017 Q4       15560.45 15560.45 15560.45
2018 Q1       11170.75 11170.75 11170.75
2018 Q2       11119.83 11119.83 11119.83
2018 Q3       12093.75 12093.75 12093.75
2018 Q4       15560.45 15560.45 15560.45
2019 Q1       11170.75 11170.75 11170.75
2019 Q2       11119.83 11119.83 11119.83
2019 Q3       12093.75 12093.75 12093.75
2019 Q4       15560.45 15560.45 15560.45
2020 Q1       11170.75 11170.75 11170.75
2020 Q2       11119.83 11119.83 11119.83
2020 Q3       12093.75 12093.75 12093.75
```

Plot of time series data with seasonality, and predictions for validation period:



The above graph showcases the Regression with Seasonality for training and validation data. We can observe that the regression forecast for the revenue is under predicting the actual revenue in the validation period seasonal forecast is significantly underperforming

MLR Sub Model 5: Regression model with Quadratic Trend and Seasonality:

Objective is to fit a regression Model with quadratic Trend and Seasonality to the training time series dataset of Amazon' historical revenue figures and then use the same model to forecast revenue for validation period time series Dataset.

```
Call:
tslm(formula = train.ts.az ~ trend + I(trend^2) + season)

Residuals:
     Min       1Q   Median       3Q      Max
-2912.23  -571.58    95.21   729.38  3156.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1709.397    659.156   2.593   0.0131 *
trend        -68.256     56.764  -1.202   0.2361
I(trend^2)    15.328      1.147  13.367  < 2e-16 ***
season2     -703.089    526.285  -1.336   0.1889
season3     -412.001    526.832  -0.782   0.4387
season4     4207.599    539.328   7.802 1.25e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1289 on 41 degrees of freedom
Multiple R-squared:  0.9844,    Adjusted R-squared:  0.9825
F-statistic: 516.4 on 5 and 41 DF,  p-value: < 2.2e-16
```

Model Equation:

$$\mathbf{y_t = 1709.397 - 68.256\ t + 15.328\ t^2 - 703.089\ D2 - 412.001\ D3 + 4207.599\ D4}$$

Observations for Sub Model 5:

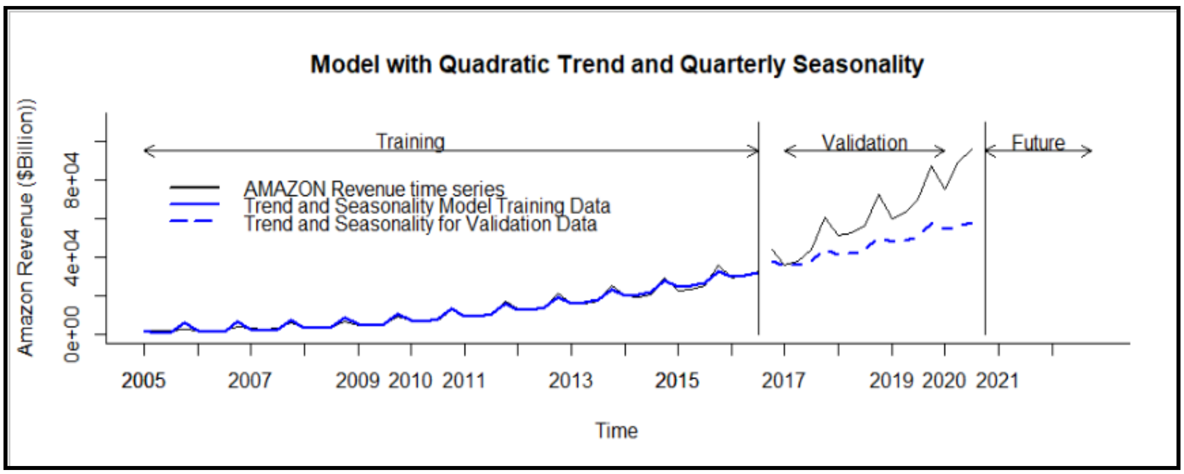
The regression model with Quadratic Trend and Seasonality consists of 5 predictors which are Trend, Trend square, seasonal dummy variables for Quarter 2(Season2), Quarter 3(Season 3) and Quarter 4 (Season 4). season variables Season4(Q4) appears to be only significant variables for the model, with their p values being very lower than .05. P value for Trend square has lower p value. The model's summary shows a very high R-squared of 0.9844 and adj. R_squared of 0.9825

, pretty high F-statistic (p-value is substantially lower than 0.05).This regression model appears to be statistically significant and can be a good fit time series forecast .

The forecast result for validation dataset using Regression Model with Quadratic Trend and Seasonality is given below (confidence interval is not used).

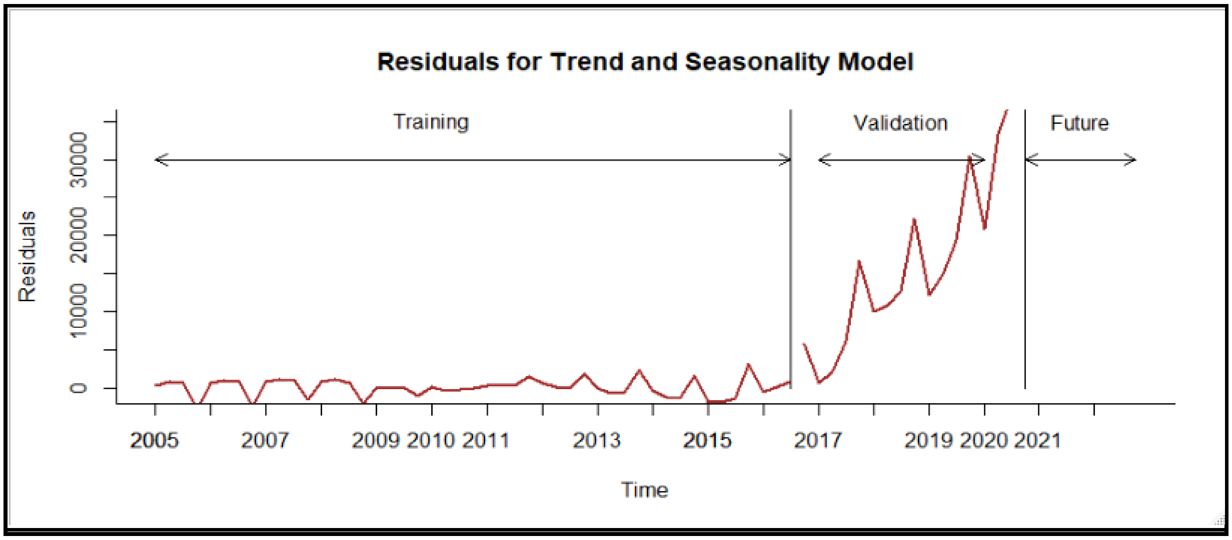
	Point Forecast	Lo 0	Hi 0
2016 Q4	37957.02	37957.02	37957.02
2017 Q1	35168.00	35168.00	35168.00
2017 Q2	35914.16	35914.16	35914.16
2017 Q3	37685.14	37685.14	37685.14
2017 Q4	43815.30	43815.30	43815.30
2018 Q1	41148.91	41148.91	41148.91
2018 Q2	42017.69	42017.69	42017.69
2018 Q3	43911.30	43911.30	43911.30
2018 Q4	50164.08	50164.08	50164.08
2019 Q1	47620.32	47620.32	47620.32
2019 Q2	48611.72	48611.72	48611.72
2019 Q3	50627.96	50627.96	50627.96
2019 Q4	57003.36	57003.36	57003.36
2020 Q1	54582.23	54582.23	54582.23
2020 Q2	55696.26	55696.26	55696.26
2020 Q3	57835.12	57835.12	57835.12

Plot of time series data with Quadratic Trend and seasonality, and predictions for validation period:



The above graph showcases the regression model with quadratic trend and seasonality for training and validation data. We can see that the forecast value is lower than the actual revenue in the validation period.

Plot of residuals of predictions with trend and seasonality:



The above graph showcases the residuals model for quadratic trend and seasonality model for training and validation data. The residual values in the validation period are positive, indicating that our forecast values are lower than the actual values. It means that our forecast is under predicting.

2.6.4 Model 4: Two level Model with Regression and AR Model for Residuals:

In two level models with Regression and AR, we are using a regression model with exponential trends with an autoregressive model with order 5 (AR).

Objective:

Develop a two-level model (Regression with exponential trend + AR (5) model for residuals) to forecast the quarterly revenue of Amazon for the validation period from Q4-2016 to Q3-2020.

Scope:

➢ In two level model we are using regression model with exponential trend and lambda equal to zero to generate forecast

- ➢ We will examine the forecast residual series for autocorrelation by utilizing time plot of forecast residual and ACF function plot
- ➢ If autocorrelation of residuals exists, we will fit AR model to forecast residual series

Model Execution:

The output for the regression model with exponential trend for the training period and forecast for the validation period are shown below.

```
> summary(train.expo.trend.season )

Call:
tslm(formula = train.ts.az ~ trend, lambda = 0)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22893 -0.14547 -0.05739  0.07964  0.43893

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.497812   0.056838  131.91   <2e-16 ***
trend       0.065331   0.002062   31.69   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1917 on 45 degrees of freedom
Multiple R-squared:  0.9571,    Adjusted R-squared:  0.9562
F-statistic:  1004 on 1 and 45 DF,  p-value: < 2.2e-16
```

This regression model with exponential trend contains 1 independent variables: trend index (t)

Model Equation:

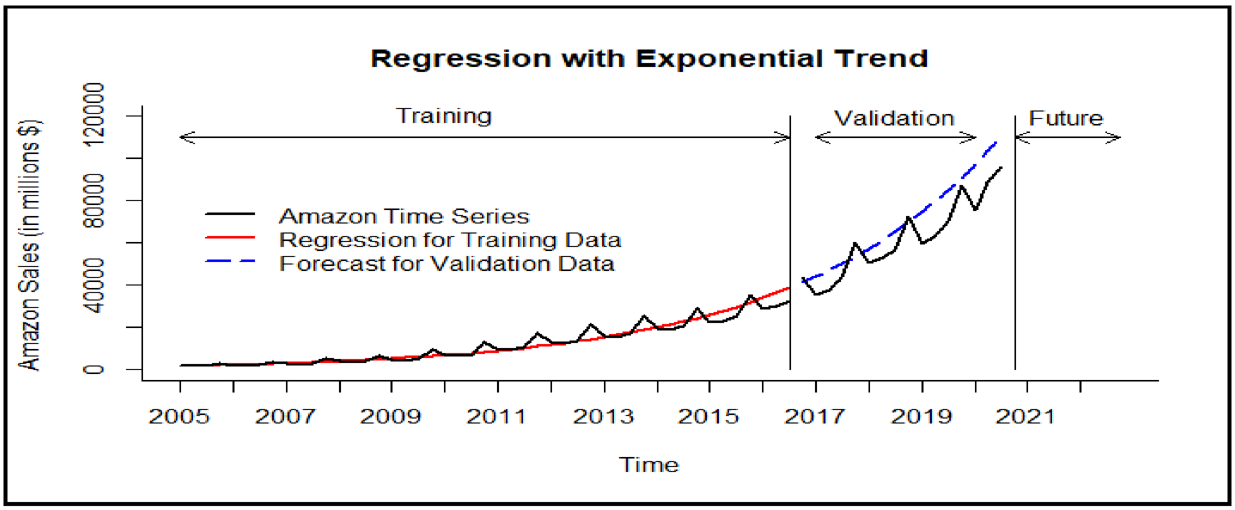
$$y_t = 7.497812+0.065331\,t$$

The model’s summary shows a very high R-squared of 0.9571 and adj. R_squared of 0.9652, statistically significant F-statistic (p-value is substantially lower than 0.05), trend (t) is statistically significant (p-value <0.05). This regression model is statistically significant and a good fit for the historical data set, and thus can be used for forecast validation data.

The validation forecast is shown below:

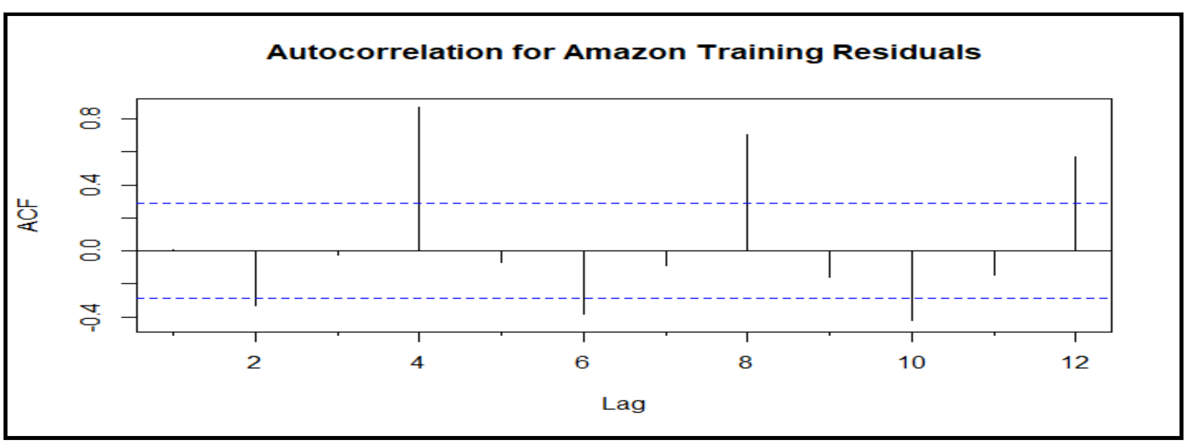
```
> train.expo.trend.season.pred
        Point Forecast     Lo 0      Hi 0
2016 Q4       41510.18  41510.18  41510.18
2017 Q1       44312.63  44312.63  44312.63
2017 Q2       47304.27  47304.27  47304.27
2017 Q3       50497.88  50497.88  50497.88
2017 Q4       53907.11  53907.11  53907.11
2018 Q1       57546.49  57546.49  57546.49
2018 Q2       61431.58  61431.58  61431.58
2018 Q3       65578.96  65578.96  65578.96
2018 Q4       70006.34  70006.34  70006.34
2019 Q1       74732.62  74732.62  74732.62
2019 Q2       79777.98  79777.98  79777.98
2019 Q3       85163.97  85163.97  85163.97
2019 Q4       90913.57  90913.57  90913.57
2020 Q1       97051.35  97051.35  97051.35
2020 Q2      103603.50 103603.50 103603.50
2020 Q3      110598.00 110598.00 110598.00
```

Plot ts data, Regression with exponential trend data, and predictions for validation period:



From the graph, it can be seen that the model is little over predicting the validation data

Acf() function to identify autocorrelation for the model residuals:



The chart shows strong significant autocorrelation of residuals for lags 4, lag 8, lag 12, as well as some negative autocorrelation can be found in lag-2, lag 6, and lag 10 which means that these autocorrelations (relationships) between residuals are not incorporated into the regression model. Thus, modeling this residual autocorrelation with an AR model and developing a two-level model may, overall, improve the forecast.

AR (5) model for training residuals is shown below:

The output of the AR (5) model for regression residuals is presented below. ARIMA (5, 0, 0) is an autoregressive (AR) model with order 5, no differencing, and no moving average model.

```
> summary(res.ar1)
Series: train.expo.trend.season.pred$residuals
ARIMA(5,0,0) with non-zero mean

Coefficients:
         ar1      ar2      ar3     ar4      ar5    mean
      0.7836  -0.0507  -0.0504  0.9258  -0.8382  0.0090
s.e.  0.0750   0.0250   0.0263  0.0247   0.0760  0.0246

sigma^2 estimated as 0.001662:  log likelihood=80.37
AIC=-146.75   AICc=-143.88   BIC=-133.8

Training set error measures:
                        ME       RMSE       MAE      MPE     MAPE      MASE      ACF1
Training set -0.0008643209 0.03807839 0.0286583 78.33351 96.63374 0.4570215 0.1694548
>
```

Model equation:

$$\boldsymbol{e_t = 0.0090+0.7836e_{t-1}-0.0507e_{(t-2)}-0.0504e_{(t-3)} +0.9258e_{(t-4)}-0.8382e_{(t-5)}}$$

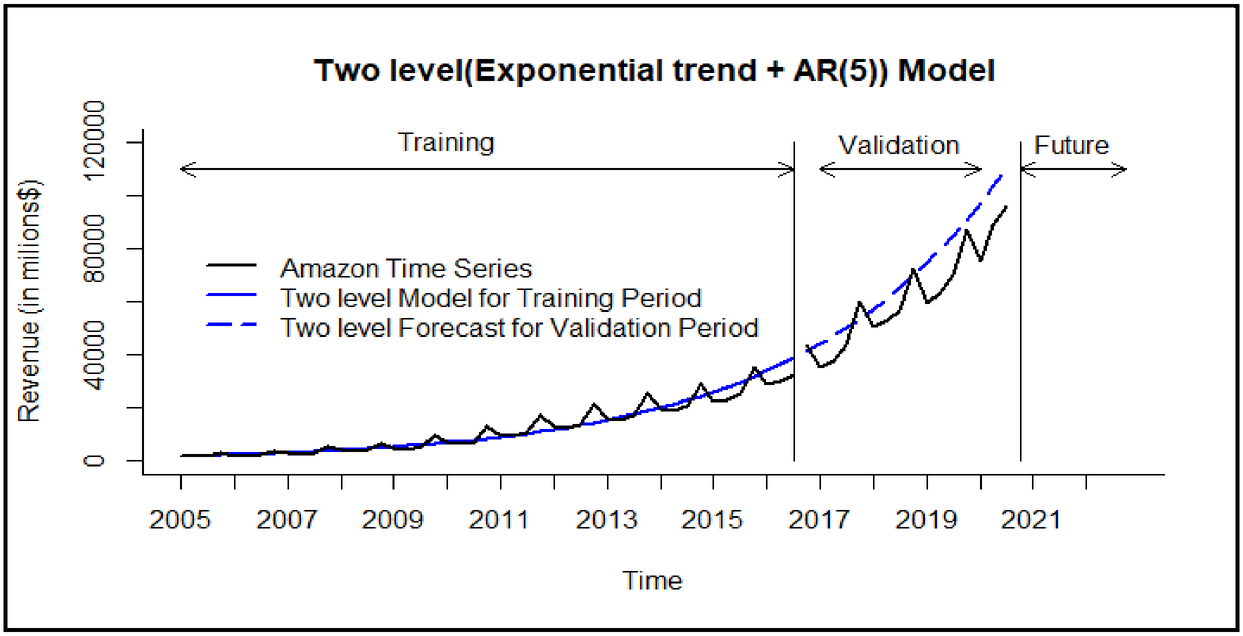
Forecast to make prediction of residuals in validation set:

```
> res.ar1.pred
        Point Forecast        Lo 0        Hi 0
2016 Q4     0.12663297  0.12663297  0.12663297
2017 Q1    -0.12107805 -0.12107805 -0.12107805
2017 Q2    -0.12516569 -0.12516569 -0.12516569
2017 Q3    -0.10476974 -0.10476974 -0.10476974
2017 Q4     0.19451130  0.19451130  0.19451130
2018 Q1    -0.05214145 -0.05214145 -0.05214145
2018 Q2    -0.05775747 -0.05775747 -0.05775747
2018 Q3    -0.04243188 -0.04243188 -0.04243188
2018 Q4     0.24227878  0.24227878  0.24227878
2019 Q1    -0.01434597 -0.01434597 -0.01434597
2019 Q2    -0.02908378 -0.02908378 -0.02908378
2019 Q3    -0.02307894 -0.02307894 -0.02307894
2019 Q4     0.24605392  0.24605392  0.24605392
2020 Q1    -0.01886373 -0.01886373 -0.01886373
2020 Q2    -0.03892626 -0.03892626 -0.03892626
2020 Q3    -0.03687110 -0.03687110 -0.03687110
>
```

Acf() function to identify autocorrelation for the model residuals:

As can be seen from the chart (correlogram), all autocorrelations of residuals of residuals created by AR (5) model are random. Thus, the AR (5) model for residuals has absorbed significant autocorrelation in all lags. Therefore, the AR (5) model for residuals can be combined with the regression model to improve the time series forecast.

Plot two-level modeling results, Regression + AR (5) for validation period:



From the graph, it can be seen that model prediction for validation data is little over predicting.

Two level Model (Regression with AR (5)) Accuracy:

```
> round(accuracy(valid.two.level.pred, valid.ts.az), 3) #RMSE=11414.8 MAPE 16.565
                ME    RMSE      MAE     MPE   MAPE  ACF1 Theil's U
Test set -8650.723 11414.8 10044.82 -14.163 16.565 0.111     1.056
>
```

Accuracy measures for two level Model (Regression with AR (5)) results in RMSE = 11414.8 and MAPE = 16.565% on Validation data.

3.6.5. Model 5: ARIMA Model

ARIMA is the abbreviation for Auto Regressive Integrated Moving Average. Auto Regressive (AR) terms refer to the lags of the differenced series, Moving Average (MA) terms refer to the lags of errors and I is the number of differences used to make the time series stationary.

Objective:

Develop an Auto ARIMA Model in order to forecast the quarterly revenue of Amazon for validation period Q4-2016 to Q3-2020.

Scope:

- ➢ ARIMA model can represent time series components like level, trend, and seasonality. This model is also capable of representing a combination of these components.
- ➢ In an ARIMA model we transform a time series into stationary one using differencing (to remove linear trend). (D) and (d) refers to the number of differencing transformations required by the time series to get stationary. If these values fail to revolve around a constant mean and variance then we find the second differencing using the values of the first differencing. We repeat this until we get a stationary series. The best way to determine whether the series is sufficiently differenced is to plot the differenced series and check to see if there is a constant mean and variance.

Model Execution:

Use auto.arima() function to fit ARIMA model. Then use summary () to show auto ARIMA model and its parameters.

```
Series: train.ts.ad
ARIMA(1,1,0)(2,1,0)[4]

Coefficients:
         ar1     sar1     sar2
      0.2839   0.3609   0.4816
s.e.  0.1512   0.1367   0.1413

sigma^2 estimated as 206747:  log likelihood=-317.49
AIC=642.98   AICc=644.06   BIC=649.93

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE
Training set 53.76597 414.1929 308.4874 0.2199496 3.335805
                 MASE        ACF1
Training set 0.111002 -0.03430737
```

Here, we get ARIMA (p,d,q)(P,D,Q) model for the historical data with level, trend, and seasonality components.

Non-seasonal Components:

- Autoregressive model with number of autocorrelation lags is 1 (p)
- Differencing order is 1(d) to remove linear trend
- Moving Average model of order is 0 (q) for error lags

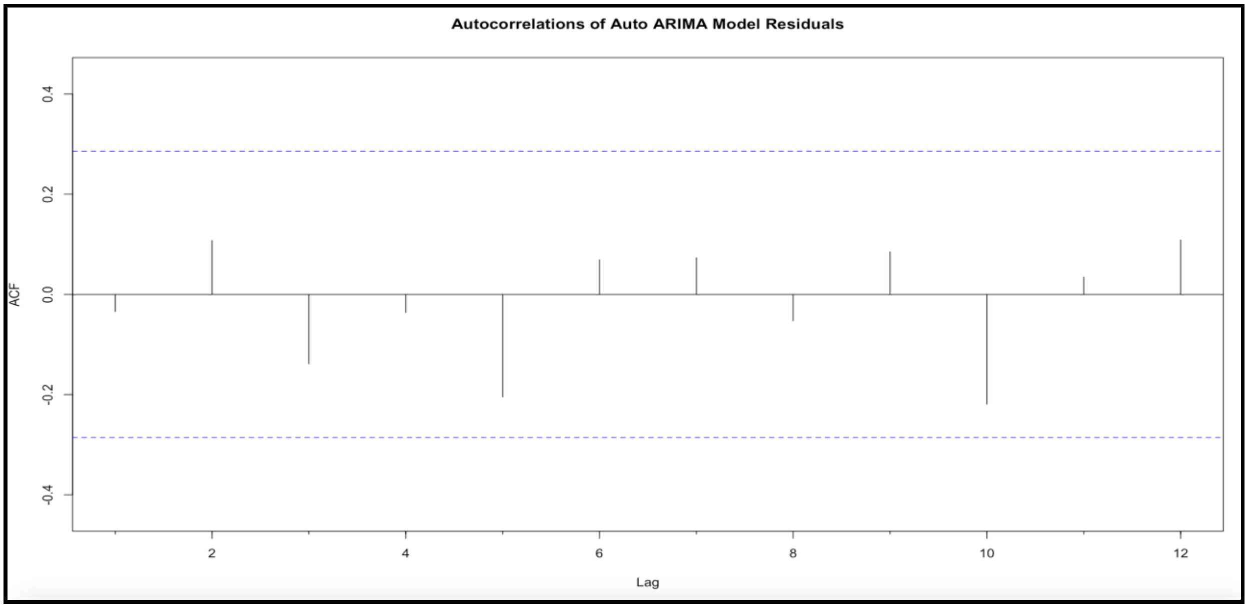
Seasonality components:

- Autoregressive model with number of autocorrelation lags is 2 (P)
- Differencing order is 1(D)
- order 0 moving average is (Q) for error lags

Model Equation:

Auto ARIMA Model Equation: $y_t - y_{t-1} = 0.284\ (y_{t-1} - y_{t-2}) + 0.361\ (y_{t-1} - y_{t-5}) + 0.482\ (y_{t-2} - y_{t-6})$

Apply Acf() to create autocorrelation chart:

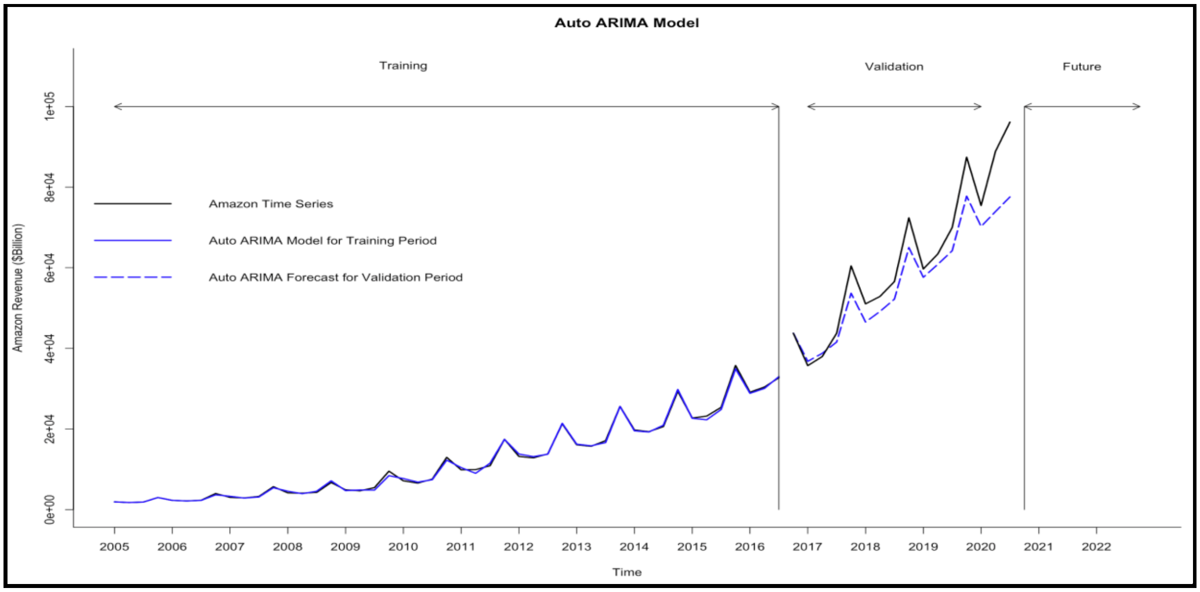


We can see only random noise in the chart as the model incorporates all the components – level, trend, and seasonality.

Apply forecast () function to make predictions for ts with auto ARIMA model in the validation set.

	Point Forecast	Lo 0	Hi 0
2016 Q4	43779.93	43779.93	43779.93
2017 Q1	36780.19	36780.19	36780.19
2017 Q2	38763.36	38763.36	38763.36
2017 Q3	41571.75	41571.75	41571.75
2017 Q4	53671.41	53671.41	53671.41
2018 Q1	46530.43	46530.43	46530.43
2018 Q2	49157.93	49157.93	49157.93
2018 Q3	52212.18	52212.18	52212.18
2018 Q4	65011.41	65011.41	65011.41
2019 Q1	57635.63	57635.63	57635.63
2019 Q2	60836.24	60836.24	60836.24
2019 Q3	64219.23	64219.23	64219.23
2019 Q4	77768.76	77768.76	77768.76
2020 Q1	70240.22	70240.22	70240.22
2020 Q2	73957.97	73957.97	73957.97
2020 Q3	77578.00	77578.00	77578.00

Plot ts data, trend and seasonality data, and predictions for validation period:



From the above graph, we can see auto ARIMA model fits well on historical training data. However, this model under is a bit under predicting the validation data.

Auto ARIMA model accuracy:

```
> # Accuracy on the validation dataset
> round(accuracy(train.auto.arima.az.pred, valid.ts.az), 3)# RMSE = 7471.543, MAPE = 7.813
                   ME     RMSE      MAE   MPE  MAPE  MASE   ACF1 Theil's U
Training set   53.767  414.193  308.487 0.220 3.336 0.111 -0.034        NA
Test set     5363.147 7471.543 5602.332 7.163 7.813 2.016  0.501     0.612
> 
```

Accuracy measures for Auto ARIMA model results into RMSE = 7471.543 and MAPE = 7.813% on Validation data.

2.7. Evaluate and compare performance

The accuracy performance of the above model for validation data are presented below:

Model 1- Two level (Regression + MA Trailing for Residuals)

```
> round(accuracy(reg.trend.seas.pred, valid.ts.az), 3) # RMSE=19192.227, MAPE=22.838
                 ME      RMSE       MAE    MPE   MAPE  MASE   ACF1 Theil's U
Training set    0.0  1203.615   915.618  2.094 14.568 0.329 -0.128        NA
Test set    15985.4 19192.227 15985.402 22.838 22.838 5.752  0.605     1.624
>
```

Model 2-Holt's Winter

```
> round(accuracy(hw.ZZZ.train.pred.az, valid.ts.az), 3) #RSME=12763.096 MAPE=13.166
                  ME      RMSE      MAE    MPE   MAPE  MASE  ACF1 Theil's U
Training set 118.419   744.407  520.890  0.839  4.255 0.187 0.113        NA
Test set    8635.902 12763.096 9374.214 11.306 13.166 3.373 0.627     0.983
>
```

Model 3- Regression

Regression model with linear trend

```
> round(accuracy(train.az.lin.pred, valid.ts.az), 3)  #RMSE: 32450.33 , MAPE: 43.003
                   ME     RMSE       MAE    MPE   MAPE   MASE  ACF1 Theil's U
Training set     0.00  3292.52  2754.279 10.494 43.992  0.991 0.359        NA
Test set     28791.27 32450.33 28791.269 43.003 43.003 10.360 0.639      2.82
>
```

Regression Model with Exponential Trend

```
> round(accuracy(train.az.expo.pred, valid.ts.az), 3) #RMSE: 11414.85 , MAPE: 16.565
                    ME     RMSE      MAE     MPE   MAPE MASE  ACF1 Theil's U
Training set       Inf      Inf      Inf     NaN    NaN  NaN    NA        NA
Test set     -8650.714 11414.85 10044.89 -14.163 16.565  NaN 0.111     1.056
>
```

Regression model with quadratic trend

```
> round(accuracy(train.az.quad.pred, valid.ts.az), 3)  #RMSE: 20293.347 , MAPE:24.104
                   ME      RMSE       MAE    MPE   MAPE  MASE   ACF1 Theil's U
Training set     0.00  2290.692  1631.545 -2.980 14.382 0.587 -0.214        NA
Test set     16862.72 20293.347 16862.724 24.104 24.104 6.068  0.504      1.74
>
```

Regression model with seasonality

```
> round(accuracy(train.az.season.pred, valid.ts.az), 3)   #RMSE:52755.052 , MAPE: 78.341
                   ME      RMSE       MAE      MPE    MAPE   MASE  ACF1 Theil's U
Training set     0.00  9461.359  8148.598 -108.417 141.632  2.932 0.914        NA
Test set     49734.12 52755.052 49734.116   78.341  78.341 17.896 0.734     4.798
>
```

Regression model with quadratic trend and seasonality

```
> round(accuracy(train.az.trend.season.pred, valid.ts.az),3) #RMSE:19192.227 #MAPE: 22.838
                  ME      RMSE       MAE    MPE   MAPE  MASE   ACF1 Theil's U
Training set     0.0  1203.615   915.618  2.094 14.568 0.329 -0.128        NA
Test set     15985.4 19192.227 15985.402 22.838 22.838 5.752  0.605     1.624
```

Model 4- Two level Model (expo trend + AR (5)) Model)

```
> round(accuracy(valid.two.level.pred, valid.ts.az), 3) #RMSE=11414.8 MAPE 16.565
                ME    RMSE      MAE     MPE   MAPE  ACF1 Theil's U
Test set -8650.723 11414.8 10044.82 -14.163 16.565 0.111     1.056
> 
```

Model 5- Auto ARIMA

```
> round(accuracy(train.auto.arima.az.pred, valid.ts.az), 3)# RMSE = 7471.543, MAPE = 7.813
                   ME     RMSE      MAE   MPE  MAPE  MASE   ACF1 Theil's U
Training set   53.767  414.193  308.487 0.220 3.336 0.111 -0.034        NA
Test set     5363.147 7471.543 5602.332 7.163 7.813 2.016  0.501     0.612
> 
```

Comparison of model based on MAPE and RMSE:

Model Name	**MAPE**	**RMSE**
Two level (Regression + MA Trailing for Residuals)	22.838	19192.227
Holt's Winter	**13.166**	**12763.096**
Regression model with linear trend	43.003	32450.33
Regression Model with Exponential Trend	16.565	11414.85
Regression model with quadratic trend	24.104	20293.347
Regression model with seasonality	78.341	52755.052
Regression model with quadratic trend and seasonality	22.838	19192.227
Two level Model (expo trend + AR (5)) Model)	16.565	11414.8

Auto ARIMA	**7.813**	**7471.543**

Based on the lowest values of MAPE and RMSE accuracy measures for the validation period, the two best models are (in descending order): ARIMA (MAPE and RMSE for the validation period forecast are 7.813% and 7471.543, respectively) and Holt's winter Model (MAPE and RMSE for the validation period forecast are 13.166% and 12763.096, respectively). These two models will be considered for forecasting Amazon revenue for the four quarters in Q4-2020 to Q3-2021 period.

2.8. Implementation of Two best model on entire data set

Applying best two models Auto Arima and Holt's winter on entire data. Summary, forecast plot and accuracy measure for these two models are shown below

2.8.1 Auto Arima

Summary of Auto Arima for entire data set

```
> summary(auto.arima)
Series: Amazon.ts
ARIMA(1,1,0)(2,1,0)[4]

Coefficients:
         ar1     sar1     sar2
      0.3284   0.2635   0.4295
s.e.  0.1286   0.1526   0.1569

sigma^2 estimated as 2712368:  log likelihood=-511.7
AIC=1031.4   AICc=1032.16   BIC=1039.64

Training set error measures:
                   ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set 183.2773 1538.811 815.1147 0.3700598 3.624398 0.1416731 0.03327963
>
```

Here, we get ARIMA (p,d,q)(P,D,Q) model for the historical data with level, trend, and seasonality components.

Non-seasonal Components:

- Autoregressive model with number of autocorrelation lags is 1 (p)
- Differencing order is 1(d) to remove linear trend

- Moving Average model of order is 0 (q) for error lags

Seasonality components:

- Autoregressive model with number of autocorrelation lags is 2 (P)
- Differencing order is 1(D)
- order 0 moving average is 0(Q) for error lags

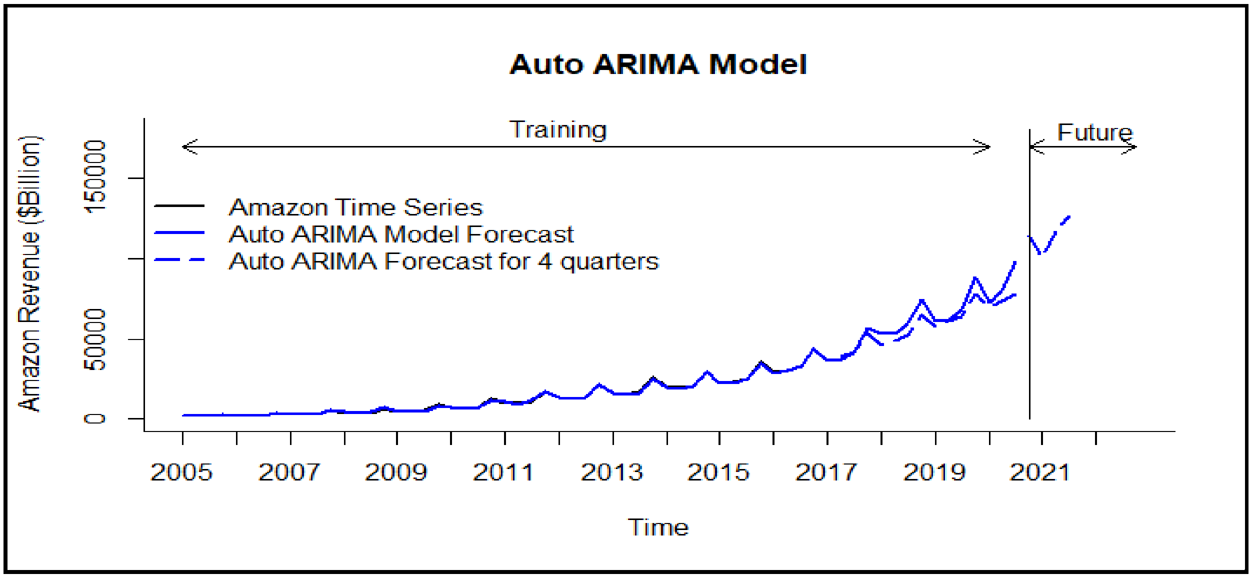
Model Equation:

$$y_t - y_{t-1} = -0.3284(y_{t-1} - y_{t-2}) + 0.2635(y_{t-1} - y_{t-5}) + 0.4295(y_{t-2} - y_{t-6})$$

Forecast for Q4-2020-Q3-2021:

```
> auto.arima.pred <- forecast(auto.arima, h = 4, level = c(85,95))
> auto.arima.pred
        Point Forecast     Lo 85    Hi 85     Lo 95    Hi 95
2020 Q4       113909.8 111539.00 116280.6 110681.89 117137.7
2021 Q1       100789.5  96847.41 104731.5  95422.25 106156.7
2021 Q2       117647.2 112438.12 122856.3 110554.87 124739.6
2021 Q3       126302.2 120032.53 132571.9 117765.86 134838.6
```

Plot ts data for training data and future forecast based on auto arima model:



The above graph shows the training data and future forecast. Our forecast using auto arima predicts exponential growth and increase in revenue for Amazon

2.8.2 Holt's Winter Model

Summary of Holt's Winter for entire data set

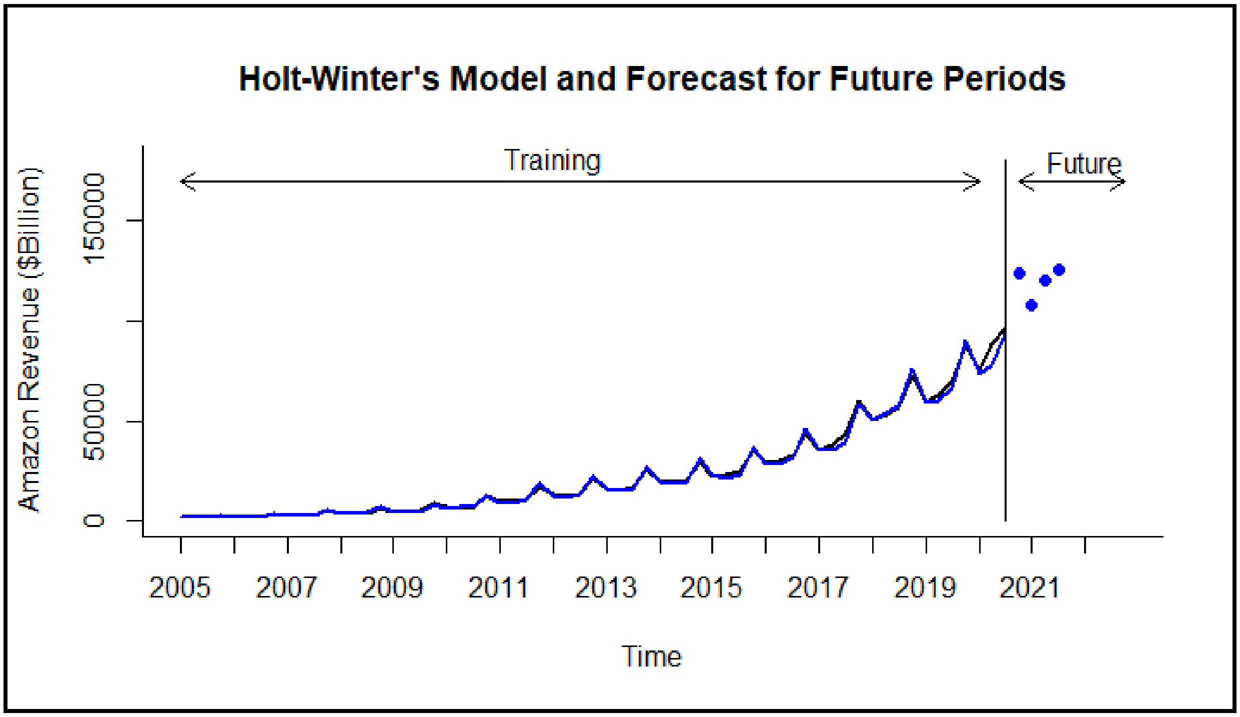
```
> HW.ZZZ # Model appears to be (M, A, M), with alpha =0.6077,Beta=0.2715,gamma =0.3923.
ETS(M,A,M)

Call:
 ets(y = Amazon.ts, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.6077
    beta  = 0.2715
    gamma = 0.3923

  Initial states:
    l = 1685.8558
    b = 222.199
    s = 1.311 0.8681 0.8576 0.9634

  sigma:  0.0564

     AIC     AICc      BIC
1110.879 1114.275 1130.167
>
```

A summary of the multiplicative Holt-Winters (HW) model with multiplicative error, additive trend, and multiplicative seasonality (model = "MAM") for the entire period is shown above. It can be seen from the model's summary that the optimal value for exponential smoothing constant (alpha) is 0.6077 the optimal smoothing constant for trend (beta) is 0.2715, and the optimal smoothing constant for seasonality estimate (gamma) is 0.3923.

Forecast for Q4-2020-Q3-2021:

```
> HW.ZZZ.pred
        Point Forecast     Lo 85    Hi 85     Lo 95    Hi 95
2020 Q4       124021.6 113951.11 134092.1 110310.32 137732.9
2021 Q1       108306.6  96924.59 119688.7  92809.66 123803.6
2021 Q2       120446.9 104381.49 136512.4  98573.36 142320.5
2021 Q3       125921.5 105220.21 146622.8  97736.10 154106.9
>
```

Plot ts data for training data and future forecast based on Holt's Winter model:



The above graph shows the training data and future forecast. Our forecast using Holt's Winter also predicts exponential growth and increase in revenue for Amazon

<u>Accuracy Performance of two best Model- Auto Arima and Holt's winter</u>

Auto Arima:

```
> round(accuracy(auto.arima.pred$fitted, Amazon.ts), 3)
              ME     RMSE     MAE  MPE  MAPE  ACF1 Theil's U
Test set 183.277 1538.811 815.115 0.37 3.624 0.033     0.203
>
```

Holts's winter

```
> round(accuracy(HW.ZZZ.pred$fitted, Amazon.ts), 3)
              ME   RMSE      MAE   MPE  MAPE  ACF1 Theil's U
Test set 337.882 1921.9 1056.206 0.972 4.203 0.255     0.216
>
```

Based on the lowest values of MAPE and RMSE accuracy measures for the entire period, Auto ARIMA (MAPE and RMSE for the entire period forecast are 3.624% and 1538.811, respectively) is best model compared to Holt’s winter Model (MAPE and RMSE for the entire period forecast are 4.203% and 1921.9, respectively).

2.9 Multivariate forecasting with External Variable

We will execute the regression model with exponential trends with US GDP. US GDP is highly correlated with revenue, so it is considered as external value to forecast Amazon revenue

Objective:

Develop a regression model with exponential trend with US GDP as external factor to forecast the quarterly revenue of Amazon

Scope:

- ➢ We are combining regression model with exponential trend with external variable US GDP
- ➢ For forecasting primary variable, first we must forecast external variable US GDP
- ➢ To Forecast US GDP, we have used regression model with trend and forecasted for Q4-2020 to Q3-2021
- ➢ New data frame is formed using the US GDP forecasted value. Regression model with exponential trend is applied on new data set for forecast of primary variable i.e Amazon Revenue

Model Execution:

The summary of regression model with exponential trend with external variable US GDP

```
> summary(az.expo.trend.external)

Call:
tslm(formula = Amazon.ts ~ trend + gdp.ts, lambda = 0)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27810 -0.11334 -0.04406  0.05537  0.41735 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.618e+00  5.302e-01  16.254   <2e-16 ***
trend        7.512e-02  6.099e-03  12.316   <2e-16 ***
gdp.ts      -8.692e-05  4.268e-05  -2.037   0.0461 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1726 on 60 degrees of freedom
Multiple R-squared:  0.9788,	Adjusted R-squared:  0.9781 
F-statistic:  1388 on 2 and 60 DF,  p-value: < 2.2e-16
```

This regression model with exponential trend and seasonality with external variable has contains 2 independent variables: trend index (t), and US GDP

Model equation:

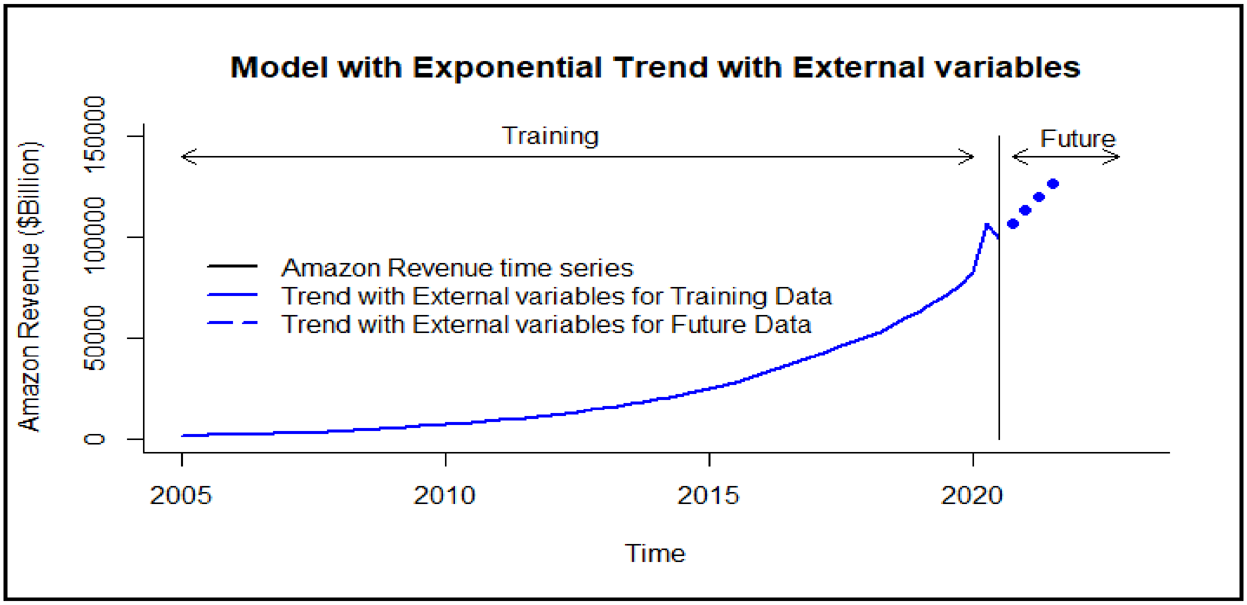
$$y_t = 8.618+0.07512\, t -0.00008692 US\ GDP$$

The model's summary shows a very high R-squared of 0.9788 and adj. R_squared of 0.9781, statistically significant F-statistic (p-value is substantially lower than 0.05), trend, and US GDP are statistically significant (p-value <0.05). This regression model is statistically significant and a good fit for the historical data set, and thus can be used for forecasting data.

Forecast output:

```
> az.expo.trend.external.pred
        Point Forecast     Lo 0     Hi 0
2020 Q4       107018.8 107018.8 107018.8
2021 Q1       113283.3 113283.3 113283.3
2021 Q2       119892.3 119892.3 119892.3
2021 Q3       126863.1 126863.1 126863.1
```

Plot ts data, Regression with exponential trend with external variable, and predictions for future period:



It is clear from graph that regression model with exponential trend with external variables predicts exponential increase in amazon revenue with trend

Regression Model with exponential trend with external variable Accuracy:

```
> round(accuracy(az.expo.trend.external.pred$fitted,Amazon.ts),3) #RMSE:4459.365 , MAPE:12.787
              ME     RMSE      MAE    MPE   MAPE  ACF1 Theil's U
Test set -97.855 4459.365 2764.734 -1.351 12.787 0.027     0.732
>
```

Accuracy measures for regression model exponential trend with external variable results in RMSE = 4459.365 and MAPE = 12.787% for entire data set

3. Conclusion:

The goal of the project was to forecast Amazon revenue for Q4-2020 to Q3-2021. Out of the five models, two best models – ARIMA and Holt Winters' Model have been used to predict future revenue forecasts on the entire dataset. ARIMA was the best model as it has the lowest MAPE and RMSE. Holt Winters' model was the second-best model for forecasting Amazon Revenue. In the extended project scope, exponential trend with external variables (U.S. GDP) can also be leveraged by the company for forecasting. As per CNBC news reports, amazon has projected Q4-2020 revenue in the range of $112.0 billion to $121.0 billion. According to the ARIMA model predictions made in this project, with 95% confidence level Amazon's projected revenue for Q4-2020 would range between $110 billion to $117 billion. As per the project findings approximately 25% revenue increase is projected as compared to Q4-2019. The massive shopping surge fueled by the COVID-19 pandemic might be the reason for Amazon's revenue increase. With this level of forecasting accuracy, the company can certainly use the models for future predictions and better strategic decisions.

4. Acknowledgments:

This project would not have been possible without the guidance of Dr. Zinovy Radovilsky, the instructor of this course (BAN673-Time Series Analytics). Additionally, thanks to our family and friends for encouraging us for new research. Lastly, thanks to our team members for introducing and finalizing this area of research.

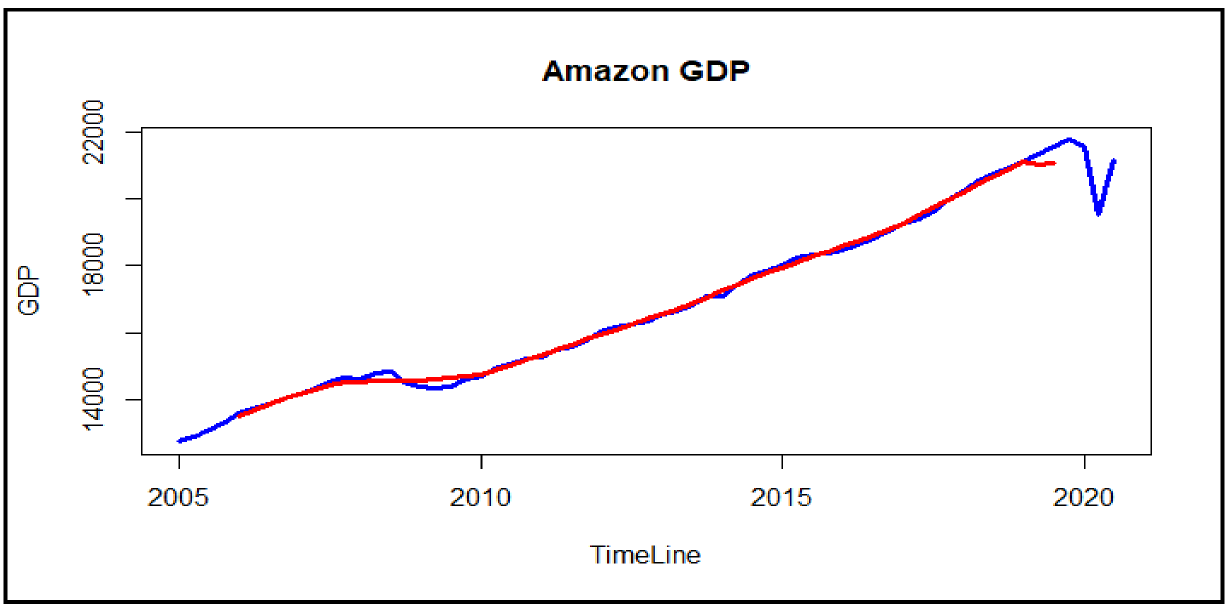
5. Bibliography

- **https://www.cnbc.com/2020/10/29/amazon-amzn-earnings-q32020.html#:~:text=Amazon%20said%20sales%20in%20the,38%25%20from%20a%20yer%20earlier.**

- https://otexts.com/fpp2/regression.html
- https://medium.com/opex-analytics/forecasting-in-times-of-disruption-9e7b2d9bd2e4
- https://www.researchgate.net/publication/327945077_Time_Series_Forecasting_using_ARIMA_Model_A_case_study_of_mining_face_drilling_rig
- https://www.researchgate.net/publication/330970319_Implementation_of_Exponential_Smoothing_for_Forecasting_Time_Series_Data
- Zinovy R. 2020. Time Series Analysis. 2020. Course Lecture Materials and Videos

6. Appendix

6.1 GDP plot



6.2 GDP Forecast

```
> trend.season.pred
        Point Forecast    Lo 0     Hi 0
2020 Q4       22093.09 22093.09 22093.09
2021 Q1       22302.84 22302.84 22302.84
2021 Q2       22514.73 22514.73 22514.73
2021 Q3       22728.77 22728.77 22728.77
```

6.3 Test Predictability Plot