CALIFORNIA STATE UNIVERSITY EAST BAY

Capstone Project

Impact of Covid-19 on E-Commerce Sales

Instructor : Professor Surendra Sarnikar
Presented By : Nidhi Patel (ke9695)

Table of Contents

Summary

The scope of the project is to analyze the impact of Covid-19 on the US e-commerce industry. The COVID-19 pandemic impacted businesses across all industries significantly and at the same time it totally reshaped consumers spending and purchasing behaviors.

As the novel coronavirus (COVID-19) spread across the United States in 2020, it has taken an ever-increasing toll on public health, as well as a number of other industries and sectors. From travel and tourism to finance and construction – almost every aspect of the U.S. economy has been affected by the global pandemic. One industry that has seen particularly noticeable changes over the past few months is E-commerce. As most states have issued stay-at-home orders to slow the spread of the disease, many Americans are now self-isolating while turning to technology for work, education, communication, and shopping.

In this project, we have analyzed the last 20 years of US quarterly total sales data and e-commerce sales data across various categories like general merchandise, food and beverages, clothing, building material, sports etc. Then we applied various time series analytics models to project e-commerce sales data for the year 2020 and identified the impact of covid 19 on e-commerce sales by comparing it with actual numbers. We also predicated e-commerce sales data for each quarter of year 2021 and identified how covid-19 has changed customer purchasing patterns and how it will impact the future growth of e-commerce sales. We also identified the impact of covid 19 on e-commerce sales across various categories and provided detailed insights.

Background

E-commerce sales have seen an upwards trend in the last 20 years as technology advances. Every year more and more businesses started increasing their online presence via various channels, which led to overall rise in online sales. However, after the novel coronavirus outbreak in March 2020 around the world, the e-commerce industry has seen dramatic change. As people have embraced social distancing as a way to slow the spread of the pandemic, there has naturally been a drop-off in brick-and-mortar shopping. That would seem to mean there would likely be an increase in online shopping as people turn to ecommerce to purchase the items they might have otherwise purchased in person. The pandemic has affected many businesses and industries but on the other side It has also served as an accelerant to many industries, pushing them years ahead of where their natural growth would have otherwise taken them.

E-commerce also benefited from lockdown orders in various countries across the world. As consumers faced stay-at-home orders and in-person shopping was replaced with online commerce for many consumers. This has changed customer's purchasing behaviors and had a significant impact in terms of various categories of products that consumers purchase online. This project provides insights on the exact impact of Covid-19 on e-commerce sales within the United States and it also provides findings on how consumers' shopping behaviors have changed across various categories of products.

Problem Statement

The goal of this project is to analyze the impact of Covid-19 on the US e-commerce industry. The COVID-19 pandemic has impacted consumer spending habits significantly across all categories and at the same time it totally reshaped consumers purchasing behaviors. The project will also identify the future impact on US e-commerce sales due to this behavior shift. For this, the US e-commerce sales data for the last 20 years has been acquired and analyzed through various analytics and prediction techniques.

Proposed Solution

- For this project, we used the past quarterly sales data of US E-Commerce Sales and retail sales from Q1 2000 to Q1 2020 to forecast future sales from Q2 of 2020 to Q4 of 2020. This dataset represents various sales numbers with defined timeline and the predictions that need to be done are for future quarters, the dataset can be well defined as time series dataset and because of that various Time Series analytics prediction models have been applied to provide the solution. We will be using the following methods for building the forecasting models.
    - Holt's winter model
    - ARIMA Model
    - Regression model with various types of Trend and Seasonality
    - Autoregressive Model
    - Facebook's Prophet model

- Once the prediction model building is done, the next step is to evaluate accuracy of all models and find out models with the highest accuracy that can be used to forecast future data. Those forecasted data then will be compared with 2020 actual sales data to evaluate the Impact of Covid 19 on e-commerce sales. Also, the prediction model will be used to forecast 2021 sales numbers to identify the shift in e-commerce sales due to changing purchasing habits of customers.
- This proposed solution has been divided into logical steps.
    1. Data Acquisition and Data Cleaning
    2. Data Exploration
    3. Data Partitioning
    4. Building Prediction Models
    5. Calculating Mode Accuracy
    6. Identify Best Models
    7. Generate Predictions

Step 1 - Data Acquisition and Data Cleaning

Data Acquisition

All the Dataset has been acquired from the US Census website, where the government publishes official US E-commerce sales and Retail sales numbers across various categories.

Data Source: - US census (https://www.census.gov/retail/index.html)

Below table represents fields of the dataset and some of the derived fields calculated from it.

Field Name	Data Type	Description
Quarter	String	1999Q1 to 2020Q1
Retail Sales, total	Numeric	Total Retail Sales in USD
Ecommerce sales, total	Numeric	Total E-Commerce Sales in USD
E Commerce as percent of total	Numeric	E-Commerce share in compare to Total Sales
Derived Fields		
QoQ% of Total Sales	Numeric	Quarterly Growth in Total Sales
QoQ% Ecommerce Sales	Numeric	Quarterly Growth in E-Commerce Sales
YoY% Total Sales	Numeric	Yearly Growth in Total Sales
YoY% Ecommerce Sales	Numeric	Yearly Growth in E-Commerce Sales

We also had a dataset of category-wise E-commerce and total retail sales, which will be used to evaluate impact of e-commerce sales across various categories in 2020.

Data Cleaning

- Some of the sales amount had unnecessary special character like various brackets, hashtag(#), asterisks (*) in it which needs to be cleaned and converted to proper number format.
- Some of the categorical data has some null values in it which needs to be removed.
- Below screenshots represents final dataset files after applying data cleaning.

Quarter	Retail Sales, Total	E-commerce Sales, Total	E-commerce as a Percent of Total	QoQ %, Total Sales	QoQ %, Ecommerce Sales	YoY%, Total Sales	YoY%, Ecommerce Sales
2020Q4	1,476,952	206,666	14	0.5	-1.2	6.9	32.1
2020Q3	1,469,769	209,251	14.2	12.1	-1.1	7	36.6
2020Q2	1,311,345	211,595	16.1	-3.8	31.9	-3.5	44.5
2020Q1	1,363,543	160,414	11.8	-1.3	2.6	2.1	14.8
2019Q4	1,381,250	156,581	11.3	0.5	2.2	3.9	16.6
2019Q3	1,374,212	153,274	11.2	1.1	4.7	3.9	17.3
2019Q2	1,359,250	146,394	10.8	1.8	4.8	3.2	13.8
2019Q1	1,335,812	139,713	10.5	0.5	4	2.5	11.8
2018Q4	1,329,085	134,291	10.1	0.5	2.8	2.7	11
2018Q3	1,322,811	130,625	9.9	0.4	1.6	4.6	13.2
2018Q2	1,317,610	128,616	9.8	1.1	2.9	5.3	14.2
2018Q1	1,302,741	124,936	9.6	0.7	3.2	4.4	15.5

1	Quarter	Category	Ecommerce Sales
2	2020Q1	Motor vehicle and parts dealers	9,241
3	2020Q1	Furniture and home furnishings stores	3,037
4	2020Q1	Electronics and appliance stores	13,262
5	2020Q1	Building mat. and garden equip. and supplies dealers	3,977
6	2020Q1	Food and beverage stores	3,526
7	2020Q1	Health and personal care stores	1,729
8	2020Q1	Gasoline stations	
9	2020Q1	Clothing and clothing access. stores	10,348
10	2020Q1	Sporting goods, hobby, musical instrument, and book stores	1,729
11	2020Q1	General merchandise stores	9,567
12	2020Q1	Miscellaneous store retailers	3,717

Year	Category	Total	Ecommerce
2020	Total Retail Trade	5,636,721	791,700
2020	Motor vehicle and parts dealers	1,250,394	44,638
2020	Furniture and home furnishings stores	111,528	16,397
2020	Electronics and appliance stores	83,173	72,038
2020	Building mat. and garden equip. and supplies dealers	381,892	25,763
2020	Food and beverage stores	853,125	24,502
2020	Health and personal care stores	364,761	9,554
2020	Gasoline stations	421,739	NA
2020	Clothing and clothing access. stores	196,539	59,152
2020	Sporting goods, hobby, musical instrument, and book stores	84,263	11,351
2020	General merchandise stores	731,561	65,444
2020	Miscellaneous store retailers	133,640	15,239
2019	Total Retail Trade	5,411,037	578,501
2019	Motor vehicle and parts dealers	1,239,767	42,684
2019	Furniture and home furnishings stores	120,517	2,159
2019	Electronics and appliance stores	90,536	1,822
2019	Building mat. and garden equip. and supplies dealers	372,432	2,165
2019	Food and beverage stores	773,647	8,303
2019	Health and personal care stores	342,569	945
2019	Gasoline stations	512,377	S
2019	Clothing and clothing access. stores	268,735	12,607
2019	Sporting goods, hobby, musical instrument, and book stores	80,757	3,197
2019	General merchandise stores	716,476	S
2019	Miscellaneous store retailers	131,560	5,224

Step 2- Data Exploration

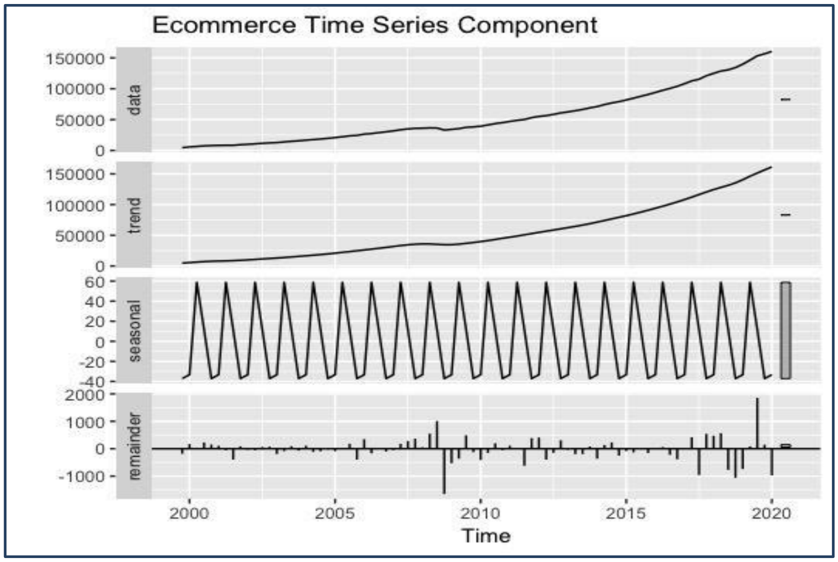
The main goal of this step is to understand various components of time series data. Any time series data consist of components like Trend, Seasonality, Noise etc. We will convert our data into time series format and use R libraries to understand its various components.

➢ **Create Time Series Data**

```
# Create Time Series DataSet For Adidas
Sales.ts <- ts(Sales_final.data$EcommerceSales_Total,
                start = c(1999, 4), end = c(2020, 1), freq = 4)
```

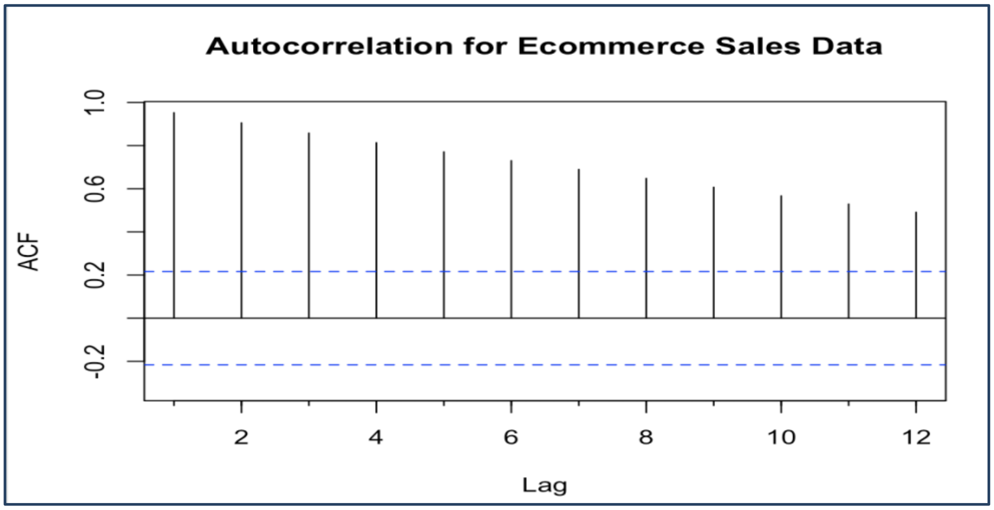
➢ **Time Series Decomposition Using stl () function**

- By using stl() command in R, we can decompose the quarterly US e-commerce sales time series data into three components namely trend, season and remainder to provide detailed analysis on time series components.
- Looking at the stl() chart, we see that both trend and seasonal components are present and they are strong. The seasonality pattern repeats itself over the years. The trend component is increasing from year 2000 to 2008 and there is little drop in 2008-2009 as there was an economic recession and then it gradually increases over the years.



➢ **Autocorrelation for E-Commerce Sales data:**

- By applying the autocorrelation on the historical data using the Acf() function present in R, we can visualize the time series components present in the data using the below chart.
- Looking at the autocorrelation chart, we can see a positive autocorrelation coefficient in lag 1 or first several periods (lag 1, lag 2, etc.) is substantially higher than the horizontal threshold (significantly greater than zero) which indicates there is strong upward trend component present in the data. There is no seasonality present in data.



Step 3- Data Partitioning

Partitioning the time series data is an important preliminary step to be considered before applying any forecasting method. The main scope of this step is to divide the time series data into training and validation datasets. We will develop forecasting models based on training data and test the models' performance using the validation data. The first 70 records (83.33%) of the Ecommerce data (1999Q4 through 2017Q1) will be partitioned as training data and rest 12 records (16.67%) of quarterly revenue (2017Q2 to 2020Q1) will be partitioned as validation data.

➢ **Training Dataset:-** Here is the training dataset that will be used for building various forecasting models.

```
> train.ts.ad
```

	Qtr1	Qtr2	Qtr3	Qtr4
1999				4476
2000	5691	6465	7419	7840
2001	8135	8336	8335	9314
2002	9904	10742	11543	12231
2003	12738	13773	14833	15588
2004	16697	17519	18506	19631
2005	20801	22233	23653	24364
2006	26417	27367	28842	30138
2007	31728	33524	34841	35784
2008	36017	36514	36292	33042
2009	34132	35282	37402	38110
2010	39289	41303	43474	45075
2011	47020	48815	50140	53123
2012	55144	56363	58449	60815
2013	62458	64383	66457	69005
2014	71108	74290	76943	79024
2015	81837	84704	87754	90687
2016	94057	97459	100519	103952
2017	108157			

➢ **Validation Dataset:-** Here is the validation dataset that will be used for evaluating the accuracy of various forecasting models.

```
> valid.ts.ad
```

	Qtr1	Qtr2	Qtr3	Qtr4
2017		112644	115419	121019
2018	124936	128616	130625	134291
2019	139713	146394	153274	156581
2020	160414			

Step 4- Prediction Models

The goal of this step is to apply various time series techniques to build prediction models.

Model 1:Holt's Winter Model

➢ **Objective:**

To develop the most optimal Holt-Winters Model with R's automated selection of error, trend, and seasonality options to forecast the quarterly ecommerce sales for validation period from 2017Q1 to 2020Q1.

➢ **Scope:**

- Holt-Winter model is used to forecast time series with trend and seasonality. On setting parameters of the model as 'ZZZ', ets() function from the forecast package evaluates all possible combinations of Error, Trend and Seasonality and gives the most optimal smoothing parameters.

- In model ZZZ, first Z represents additive or multiplicative error, second Z represents additive or multiplicative trend and third Z represents additive or multiplicative seasonality.
- The smoothing constants associated with Holt-Winters model are alpha(α), Beta(β) and Gamma(γ) each indicating exponential smoothing constants for level, trend and seasonality respectively, each having values between 0 to 1.

```
> hw.ZZZ.train.ad <- ets(train.ts.ad, model = "ZZZ")  #Model Received is AAN
> hw.ZZZ.train.ad
```

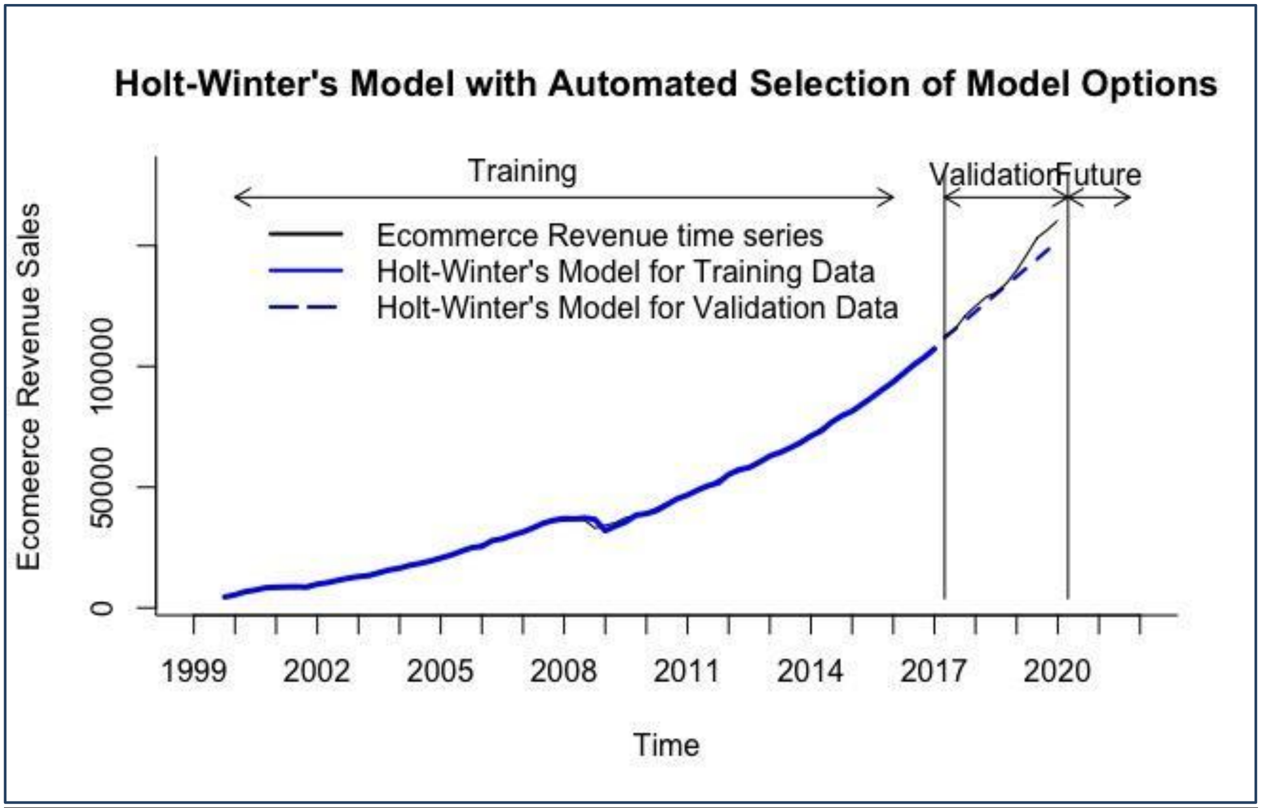
A summary of the Holt-Winter's (HW) model with the automated selection of the model options and automated selection of the smoothing parameters for the training period is shown below:

```
> hw.ZZZ.train.ad <- ets(train.ts.ad, model = "ZZZ")  #Model Received is AAN
> hw.ZZZ.train.ad
ETS(A,A,N)

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

  Smoothing parameters:
    alpha = 0.9999
    beta  = 0.3965

  Initial states:
    l = 3573.0633
    b = 944.0662

  sigma:  757.7752

     AIC     AICc      BIC
1231.530 1232.467 1242.772
```

➢ **Observations:**

- From summary, the optimal model for Adidas Training Set data comes out to be (A, A, N) indicating additive error, additive trend and none seasonality.
- The optimal value for exponential smoothing constant (alpha) is 0.9999 and smoothing constant for trend estimate (beta) is 0.3965. The alpha value of this model indicates that the model's level component tends to be more local. The smoothing constant for trend(beta) value is 0.3965 which indicates that the data captures global trend as the beta value is close to zero.

- **Holt’s Winter Model with Automated Selection for Training and Validation Data**



Above graph, showcases the original Ecommerce Training and Validation Data and how the Holt’s Winter forecast fits over the original data.

- **Accuracy measures on validation data:**

```
> round(accuracy(hw.ZZZ.train.pred.ad, valid.ts.ad), 3)
                   ME     RMSE      MAE   MPE  MAPE  MASE  ACF1 Theil's U
Training set   96.927  735.806  482.400 0.087 1.782 0.080 0.011        NA
Test set     3548.569 4793.188 3549.589 2.420 2.421 0.589 0.744     0.981
```

- Accuracy measures for HW model with additive error, additive trend and additive seasonality results into RMSE = 4793.188 and MAPE =2.421.

Model 2: Auto Arima Model

Autoregressive Integrated Moving Average (Arima Model) OR Box-Jenkins approach.

➢ **Objective:**

To develop an Auto ARIMA Model to forecast the quarterly US e-commerce sales for validation period from 2017Q1 to 2020Q1.

➢ **Scope:**

- ARIMA is a very powerful model for forecasting time series data. This model is capable of presenting any time series component – level (stationary), trend, and seasonality – or a combination of these components.
- Before implementing ARIMA, we need to detrend and de-seasonalize the time series and determine the order of parameter p (order of auto-regressive model) and order of parameter q (moving average for error lags) along with order of differencing (q) for both trend and seasonality. The data preparation and parameter tuning processes end up being really time consuming.
- Auto ARIMA makes this task simple for us as it eliminates steps like manually determining values of p, d and q and selects best values automatically by applying various combinations of values. An ARIMA model includes three parts:

1) p- Order of Auto-Regressive model for various order of lags – number of autocorrelation lags equal to p.
2) q - Order of Moving Average MA for error lags
3) d - Order of differencing to remove linear trend (d)
4) P - Order of Auto-regressive model for Seasonality
5) Q - Order of Moving Average MA for error lags

6) D - Order of Differencing to remove Linear Trend

7) M - for number of seasons in the data

➢ **Model Execution:**

A summary of the Auto Arima model for the training period is shown below.

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

Coefficients:
          ma1
      -0.5939
s.e.   0.1086

sigma^2 estimated as 565667:  log likelihood=-546.56
AIC=1097.11   AICc=1097.3   BIC=1101.55

Training set error measures:
                   ME     RMSE      MAE          MPE     MAPE       MASE        ACF1
Training set 89.41798 735.8156 479.2453 -0.005171975 1.721116 0.07952199 0.007930161
```

➢ **Observations:**

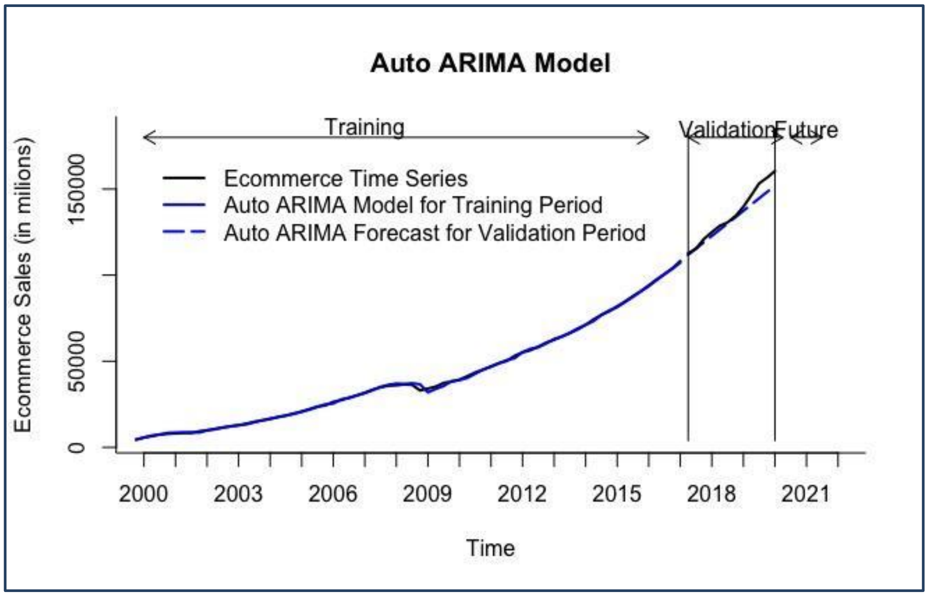
- This is a season Arima Model.
- The first three parameters generated by the auto.arima() function represent trend components. The series we received based upon auto arima model is (0,2,1) which means the following:

p = 0, order 0 autoregressive model *AR(1)*

d = 2, two differencing in trend

q = 1, 1 moving average for error lags

- **Auto ARIMA Model with Automated Selection for Training and Validation Data**



- Above graph, showcases the original Adidas Training and Validation Data and how the Auto ARIMA forecast fits over the original data.
- As seen from the graph above, the Auto ARIMA model forecast under predicts the validation data.

- **Accuracy measure on validation data:**

```
> # Accuracy on the validation dataset
> round(accuracy(train.auto.arima.ad.pred, valid.ts.ad), 3)
                   ME     RMSE      MAE    MPE  MAPE  MASE  ACF1 Theil's U
Training set   89.418  735.816  479.245 -0.005 1.721 0.080 0.008        NA
Test set     3471.176 4712.588 3476.176  2.365 2.370 0.577 0.743     0.964
```

Accuracy measures for the Auto ARIMA model results in RMSE = 4712.588 and MAPE = 2.370 on Validation data.

Model 3: 2 Level Model (Regression + MA Trailing for Residuals)

➢ **Objective:** To develop a 2 level Trailing Moving Average Model with a window of 4 to forecast the quarterly Ecommerce sales for validation period from 2017Q1 to 2020Q1.

➢ **Scope:**

- Trailing Moving Average is a predictive forecasting approach which uses historical data to make a forecast.
- The number of historical points used to make a forecast is specified in the window. In our case, we used a window of four historical points to make a forecast.
- Regression helps to model trend and seasonality and can be used to remove the trend (detrending) and seasonality (deseasonalizing) in historical data.
- We used a Regression model with Quadratic Trend and Seasonality for training data and MA-Trailing model for Residuals.

➢ **Model Execution:**

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

Residuals:
    Min      1Q  Median      3Q     Max 
-4288.2 -2657.0  -678.9  2142.6  6303.8 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8076.3527  1251.8331   6.452 1.69e-08 ***
trend        103.6394    71.9281   1.441    0.154    
I(trend^2)    17.6576     0.9819  17.983  < 2e-16 ***
season2     -169.3524  1013.3175  -0.167    0.868    
season3       22.1442  1013.1628   0.022    0.983    
season4     -261.0619   997.7195  -0.262    0.794    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2993 on 64 degrees of freedom
Multiple R-squared:  0.9898,	Adjusted R-squared:  0.989 
F-statistic:  1241 on 5 and 64 DF,  p-value: < 2.2e-16
```

➢ Model Equation

$$Y_t = 8076.3527 + 1\text{-}3.6394t + 17.65t^2 - 169.3524D^2 + 22.14D^3 - 261.06D^4$$

➢ Observations for Above Regression Model:

- The regression model with 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). All trend and season variables appear to be significant variables for the model. P value of trend is 0.1. P values for trend square and season 2 ,season 3 and season 4 are less than 0.1.
- Fstatistic's P-value for the model is pretty low which is 2.2 * 10 ^ -16 and Adj R Square is very high of value 0.989. As well the model has a high R-square of 0.9898. F statistics is pretty high, indicating it's a good model.
- The model appears to be statistically significant and there is trend present.
- **The combined two-level validation forecast for Ecommerce sales 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                  104277.2                    5592.260      109869.5
2                  107097.4                    6612.182      113709.6
3                  109478.2                    7632.105      117110.3
4                  112438.5                    8652.027      121090.6
5                  115003.8                    9671.949      124675.7
6                  117965.2                   10691.871      128657.1
7                  120487.3                   11711.794      132199.1
8                  123588.9                   12731.716      136320.6
9                  126295.4                   13751.638      140047.1
10                 129398.1                   14771.561      144169.7
11                 132061.4                   15791.483      147852.9
12                 135304.3                   16811.405      152115.7
```

➢ **Accuracy measures on validation data:**

```
> round(accuracy(reg.trend.seas.pred, valid.ts.ad), 3)
                    ME      RMSE       MAE    MPE   MAPE  MASE  ACF1 Theil's U
Training set      0.00  2861.556  2462.988 -2.211  9.428 0.409 0.916        NA
Test set      15877.53 16907.694 15877.527 11.408 11.408 2.635 0.762     3.622
> round(accuracy(ts.forecast.4, valid.ts.ad), 3)
                ME    RMSE      MAE   MPE  MAPE  ACF1 Theil's U
Test set 4675.694 5355.55 4675.694 3.319 3.319 0.725     1.122
```

Accuracy measures for 2 level (Regression + MA-Trailing) model results into RMSE = 728.271 and MAPE = 13.05.

Model 4: Regression Based Models:

➢ **Objective:**

To develop Regression based time series forecasting models to forecast the quarterly revenue of ecommerce sales from 2020Q2 to 2020Q4.

➢ **Scope:**

- Ecommerce's historical sales information 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.
- MLR will help generate a model which captures Trend, Seasonality and external information which will help us create a robust prediction model that can consider various parameters in predicting sales.

➢ **Execution of the models:**

MLR Sub Model 1: Regression model with linear trend:

Used to fit a global trend that applies to the entire time series of ecommerce historical sales data and will apply in the forecasting period.

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

Residuals:
   Min     1Q Median     3Q    Max
-10649  -4022  -1545   3981  20210

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7075.72    1728.77  -4.093 0.000115 ***
trend        1357.47      42.32  32.074  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7155 on 68 degrees of freedom
Multiple R-squared:  0.938,	Adjusted R-squared:  0.9371
F-statistic:  1029 on 1 and 68 DF,  p-value: < 2.2e-16
```

➢ Model Equation

$\boxed{Y_t = -7075.72 + 1357.47t}$

➢ **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.

- F-statistic's P-value for the model is very low which is 2.2 * 10 ^ -16 and Adj R Square is pretty decent which is 0.9371. As well the model has a relatively high R-square of 0.938. F statistics is pretty high, indicating it's a decent model.
- The model appears to be statistically significant and there is a linear trend present.
- Thus, this model can be used for Time series forecasting.

➢ **Accuracy measures on validation data:**

```
> round(accuracy(train.ad.lin.pred, valid.ts.ad), 3)
                   ME      RMSE       MAE    MPE   MAPE  MASE  ACF1 Theil's U
Training set     0.00  7051.663  5542.399 10.795 25.171 0.920 0.914        NA
Test set     38556.73 40021.997 38556.727 27.962 27.962 6.398 0.759     8.665
```

MLR Sub Model 2: Regression model with Exponential Trend:

Exponential trend implies multiplicative increase/decrease of the series over time; also means a percentage of growth from period to period (e.g., quarterly).

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

Residuals:
     Min       1Q   Median       3Q      Max 
-0.54163 -0.09002 -0.01475  0.09798  0.26694 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8.9079517  0.0365676  243.60   <2e-16 ***
trend       0.0401653  0.0008952   44.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1513 on 68 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1 
F-statistic: 3.146e+12 on 1 and 68 DF,  p-value: < 2.2e-16
```

➢ **Model Equation**

$\boxed{(log Y_t) = 8.91 + 0.04t}$

➢ **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.
- F-statistic's P-value for the model is very low which is 2.2 * 10 ^ -16 and Adj R Square is perfect one (1). As well the model has a perfect R-square of 1. F statistics is pretty high, indicating it's a decent model.
- The model appears to be statistically significant and there is an exponential trend present.
- Thus, this model can be used for Time series forecasting.

➢**Accuracy measures on validation data:**

```
> #Accuracy for Model2:  Regression Model with Exponential Trend:
> round(accuracy(train.ad.expo.pred, valid.ts.ad), 3)
                   ME     RMSE     MAE     MPE   MAPE MASE  ACF1 Theil's U
Training set      Inf      Inf     Inf     NaN    NaN  NaN    NA        NA
Test set     -25839.4 26797.16 25839.4 -18.763 18.763  NaN 0.711      5.82
```

MLR Sub Model 3: Regression model with Quadratic Trend

Fit order 2 polynomial Regression with y_t as output, and t and t^2 as predictors.

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

Residuals:
    Min      1Q  Median      3Q     Max 
-4188.0 -2591.3  -661.1  2244.5  6405.8 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7964.2882  1080.4029   7.372 3.26e-10 ***
trend        104.1322    70.2231   1.483    0.143    
I(trend^2)    17.6526     0.9585  18.417  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2927 on 67 degrees of freedom
Multiple R-squared:  0.9898,	Adjusted R-squared:  0.9895 
F-statistic:  3242 on 2 and 67 DF,  p-value: < 2.2e-16
```

- ### Model Equation

$$\boxed{Y_t = 7964.2882 + 104.1322t + 17.7t^2}$$

- ### <u>Observations for Sub-Model 3:</u>

- The regression model with Quadratic trend consists of two predictors which are Trend (t) and Trend Square(t^2). The Trend Square component appears to be a significant variable for this model with its p value below 0.01 and Trend appears to be a significant variable for the model, with its p value 0.1.
- F-statistic’s P-value for the model is very low which is 2.2 * 10 ^ -16 and Adj R Square is pretty high of value 0.9895. As well the model has a perfect R-square of 0.9895. F statistics is pretty high, indicating it’s a decent model.
- Thus, this model can be used for Time series forecasting.

➢ **Accuracy measures on validation data:**

```
> #Accuracy for Model 3:Regression mode with quadratic trend
> round(accuracy(train.ad.quad.pred, valid.ts.ad), 3)
                   ME      RMSE       MAE    MPE   MAPE  MASE  ACF1 Theil's U
Training set     0.00  2864.004  2462.589 -2.224  9.424 0.409 0.914        NA
Test set     15879.03 16911.292 15879.031 11.408 11.408 2.635 0.760     3.624
```

MLR Sub Model 4: Regression model with Seasonality:

Regression model with seasonality fits a series that falls into some seasonal pattern

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

Residuals:
   Min     1Q Median     3Q    Max 
-36605 -24564  -5741  19342  65861 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  42296.1     6871.0   6.156 4.97e-08 ***
season2      -2350.7     9859.0  -0.238    0.812    
season3       -801.9     9859.0  -0.081    0.935    
season4      -1618.4     9717.1  -0.167    0.868    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29150 on 66 degrees of freedom
Multiple R-squared:  0.0009664, Adjusted R-squared:  -0.04444 
F-statistic: 0.02128 on 3 and 66 DF,  p-value: 0.9957
```

➢ Model Equation

$$\boxed{Y_t = 42296.1 - 2350.7D^3 - 801.9D^3 - 1618.4D^4}$$

➢ **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.

- F-statistic's P-value for the model is pretty high which is 0.9957 and Adj R Square is pretty low of value-0.0444. As well the model has a low R-square of 0.00096. F statistics is pretty low, indicating it's not a good model.
- The model does not appear to be statistically significant and there is no seasonality present.
- Thus, this model cannot be used for Time series forecasting.

➢ <u>**Accuracy measures on validation data:**</u>

```
> #Accuracy for Model 4:Regression model with seasonality
> round(accuracy(train.ad.season.pred, valid.ts.ad), 3)
                  ME     RMSE      MAE     MPE    MAPE   MASE  ACF1 Theil's U
Training set     0.0 28306.17 23474.88 -91.939 121.504  3.895 0.946        NA
Test set     94223.8 95440.08 94223.80  69.252  69.252 15.635 0.768    20.879
```

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

Assuming that, the historical data has a linear trend, quadratic trend and quarterly seasonality, we can combine all the patterns in one forecasting model.

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

Residuals:
    Min      1Q  Median      3Q     Max 
-4288.2 -2657.0  -678.9  2142.6  6303.8 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8076.3527  1251.8331   6.452 1.69e-08 ***
trend        103.6394    71.9281   1.441    0.154    
I(trend^2)    17.6576     0.9819  17.983  < 2e-16 ***
season2     -169.3524  1013.3175  -0.167    0.868    
season3       22.1442  1013.1628   0.022    0.983    
season4     -261.0619   997.7195  -0.262    0.794    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2993 on 64 degrees of freedom
Multiple R-squared:  0.9898,	Adjusted R-squared:  0.989 
F-statistic:  1241 on 5 and 64 DF,  p-value: < 2.2e-16
```

➢ **Model Equation**

$$\boxed{Yt = 8076.35 + 103.6394t + 17.6576t^2 - 169.35D2 + 22.144D3 - 261.0619D4}$$

➢ **Observations for Sub Model 5:**

- The regression model with 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). All trend and season variables appear to be significant variables for the model. P value of trend is below 0.001 while p values for trend square and season 2, season 3 and season 4 are below 0.05.

- F-statistic's P-value for the model is pretty low which is 2.2 * 10 ^ -16 and Adj R Square is very high of 0.9169. As well the model has a high R-square of 0.924. F statistics is pretty high, indicating it's a good model.
- The model appears to be statistically significant and there is trend and seasonality present.
- Thus, this model can be used for Time series forecasting.

➢ <u>**Accuracy measures on validation data:**</u>

```
> #Accuracy for Model 5: Regression model with quadratic trend and seasonality.
> round(accuracy(train.ad.trend.season.pred, valid.ts.ad),3)
                   ME      RMSE       MAE    MPE   MAPE  MASE  ACF1 Theil's U
Training set     0.00  2861.556  2462.988 -2.211  9.428 0.409 0.916        NA
Test set     15877.53 16907.694 15877.527 11.408 11.408 2.635 0.762     3.622
```

Model 5: Facebook Prophet Model

➢ **Objective:** To develop model using Facebook's open-source Prophet library to forecast the quarterly Ecommerce sales for validation period from 2017Q1 to 2020Q1.

➢ **Scope:**

- Facebook's open source prophet model will consider all aspects of Time Series data like Trend, Seasonality (Daily, Monthly, Quarterly, Yearly) to provide a forecasting using historical data with better accuracy.
- Facebook's prophet model is specially good at scaling with large amount of data and still provide faster predictions.

➢ **Model Execution:**

```
#installing fb prophet packages
install.packages('prophet')
library(prophet)
install.packages('Rcpp')
install.packages('rlang')

#reading dataset file in data frame
df = read.csv("tsadjustedsales_copy_ Final_Facebook.csv",header = TRUE)
head(df)

#selecting date and e-commerce sales columns for prediction
final_df = df[,c('ds','y')]
head(final_df)

#applying fb prophet model
m=prophet(final_df)
future= make_future_dataframe(m,periods =6,freq = 'quarter')
head(future)
tail(future)

#Generating predictions with fb prophet
forcast = predict(m,future)
tail(forcast[c('ds','yhat','yhat_lower','yhat_upper')])
plot(m,forcast)
prophet_plot_components(m,forcast)
```

➢ **<u>Accuracy measures on validation data:</u>**

```
> result=rmse(actual,predicted)
> print(result)
[1] 13053.14
> result=mape(actual,predicted)
> print(result)
[1] 0.08329877
```

Step 5 - Comparing Accuracy of Prediction Models

- **Selection of best models:**

For easy comparison of the models' performance, we have listed the RMSE and MAPE values in the below table.

Model Name	RMSE	MAPE
Two level MA trailing model	5355.55	3.319
Holt's winter model	**4793.188**	**0.589**
Regression model with linear trend	40021.997	27.962
Regression model with Exponential trend	26797.6	18.763
Regression model with quadratic trend	16911.292	2.635
Regression model with Seasonality	95440.08	69.252
Regression model with quadratic trend and seasonality	16907.694	11.408
Two level Autoregressive model	11199.63	6.689
ARIMA Model	**4712.588**	**2.370**
Facebook Prophet	**13053.14**	**0.083**

Based on the RMSE and MAPE accuracy measures for the validation period, the first three best models are (in descending order): ARIMA model, Holt's winter model and Prophet model.

Forecasting Ecommerce Sales using the best models [2020Q2-2021Q4]

- In this section, we forecasted US E-commerce sales values for 2020Q2 to 2021Q4 using the best models developed in previous sections, to identify how US e-commerce sales would have been if Covid-19 pandemic would have not happened and then compared it with actual US E-Commerce sales numbers to evaluate the impact of Covid-19 on US E-commerce industry. [In US, country wide lockdowns have been announced in March 2020, start of 2020Q2]

Predicted E-commerce sales values (2020Q2 to 2021Q4) using ARIMA model

```
> # Apply forecast() function to make predictions for ts with
> # auto ARIMA model in validation set.
> Ecom.ARIMA.pred <- forecast(Ecom.ARIMA, h = 7, level = 0)
> Ecom.ARIMA.pred
        Point Forecast     Lo 0     Hi 0
2020 Q2       164673.6 164673.6 164673.6
2020 Q3       169043.2 169043.2 169043.2
2020 Q4       173441.2 173441.2 173441.2
2021 Q1       177846.5 177846.5 177846.5
2021 Q2       182253.7 182253.7 182253.7
2021 Q3       186661.4 186661.4 186661.4
2021 Q4       191069.3 191069.3 191069.3
> # Accuracy measure for entire data set
> round(accuracy(Ecom.ARIMA.pred$fitted, Sales.ts),3) # RMSE = , MAPE =
              ME    RMSE     MAE   MPE  MAPE   ACF1 Theil's U
Test set 145.554 902.742 615.654 0.059 1.632 -0.025     0.437
```

Predicted E-commerce sales values (2020Q2 to 2021Q4) using Holt's winter model

```
> hw.ZZZ.Ecom.pred <- forecast(hw.ZZZ.Ecom, h = 8, level = 0)
> hw.ZZZ.Ecom.pred
        Point Forecast     Lo 0     Hi 0
2020 Q2       164924.1 164924.1 164924.1
2020 Q3       169434.0 169434.0 169434.0
2020 Q4       173944.0 173944.0 173944.0
2021 Q1       178454.0 178454.0 178454.0
2021 Q2       182963.9 182963.9 182963.9
2021 Q3       187473.9 187473.9 187473.9
2021 Q4       191983.8 191983.8 191983.8
2022 Q1       196493.8 196493.8 196493.8
> # Accuracy measure for entire data set
> round(accuracy(hw.ZZZ.Ecom.pred$fitted, Sales.ts),3) # RMSE = , MAPE =
              ME    RMSE     MAE   MPE  MAPE  ACF1 Theil's U
Test set 147.333 927.756 634.411 0.162 1.998 0.108      0.53
```

Predicted E-commerce sales values (2020Q2 to 2021Q4) using Facebook's Prophet model

```
> forcast = predict(m,future)
> tail(forcast[c('ds','yhat','yhat_lower','yhat_upper')])
           ds     yhat yhat_lower yhat_upper
83 2020-04-01 162579.2   161621.5   163536.1
84 2020-07-01 166233.1   165130.5   167270.6
85 2020-10-01 169412.0   168296.1   170504.0
86 2021-01-01 173278.6   171990.3   174439.2
87 2021-04-01 177446.6   175978.7   178763.1
88 2021-07-01 181127.4   179413.3   182839.6
```

- In this below section, we forecasted US E-commerce sales values for 2021Q1 to 2022Q3 using the best models developed in previous sections, to identify how Covid-19 Pandemic will Impact US e-commerce sales in the future. For this we re-trained our best models by providing actual 2020 numbers, to generate forecasts of 2021Q1-2022Q3 which include covid effects also.

Predicted E-Commerce Sales values (2021Q1 - 2022Q3) using ARIMA model

- The ARIMA model has been re-trained to consider covid-19 effects, before generating below forecast, so that it can be used to evaluate future e-commerce growth impact.

```
> Ecom.ARIMA_new.pred
        Point Forecast     Lo 0     Hi 0
2021 Q1       249878.6 249878.6 249878.6
2021 Q2       286597.0 286597.0 286597.0
2021 Q3       286251.8 286251.8 286251.8
2021 Q4       296303.6 296303.6 296303.6
2022 Q1       335654.8 335654.8 335654.8
2022 Q2       362426.4 362426.4 362426.4
2022 Q3       366626.6 366626.6 366626.6
> # Accuracy measure for entire data set
> round(accuracy(Ecom.ARIMA_new.pred$fitted, Sales.ts),3) # RMSE = , MAPE =
              ME     RMSE      MAE   MPE  MAPE  ACF1 Theil's U
Test set 166.999 2075.661 1399.455 0.115 3.425 0.488     0.819
```

Predicted E-Commerce Sales values (2021Q1 - 2022Q3) using Holt's winter model

- The Holt's winter model has been re-trained to consider covid-19 effects, before generating below forecast, so that it can be used to evaluate future e-commerce growth impact

```
> hw.ZZZ.Ecom_new.pred
        Point Forecast     Lo 0     Hi 0
2021 Q1       214116.2 214116.2 214116.2
2021 Q2       221565.1 221565.1 221565.1
2021 Q3       229014.1 229014.1 229014.1
2021 Q4       236463.1 236463.1 236463.1
2022 Q1       243912.1 243912.1 243912.1
2022 Q2       251361.1 251361.1 251361.1
2022 Q3       258810.1 258810.1 258810.1
> # Accuracy measure for entire data set
> round(accuracy(hw.ZZZ.Ecom_new.pred$fitted, Sales_new.ts),3) # RMSE = , MAPE =
            ME     RMSE      MAE   MPE  MAPE   ACF1 Theil's U
Test set 532.7 5524.584 1484.871 0.408 2.082 -0.196     0.628
```

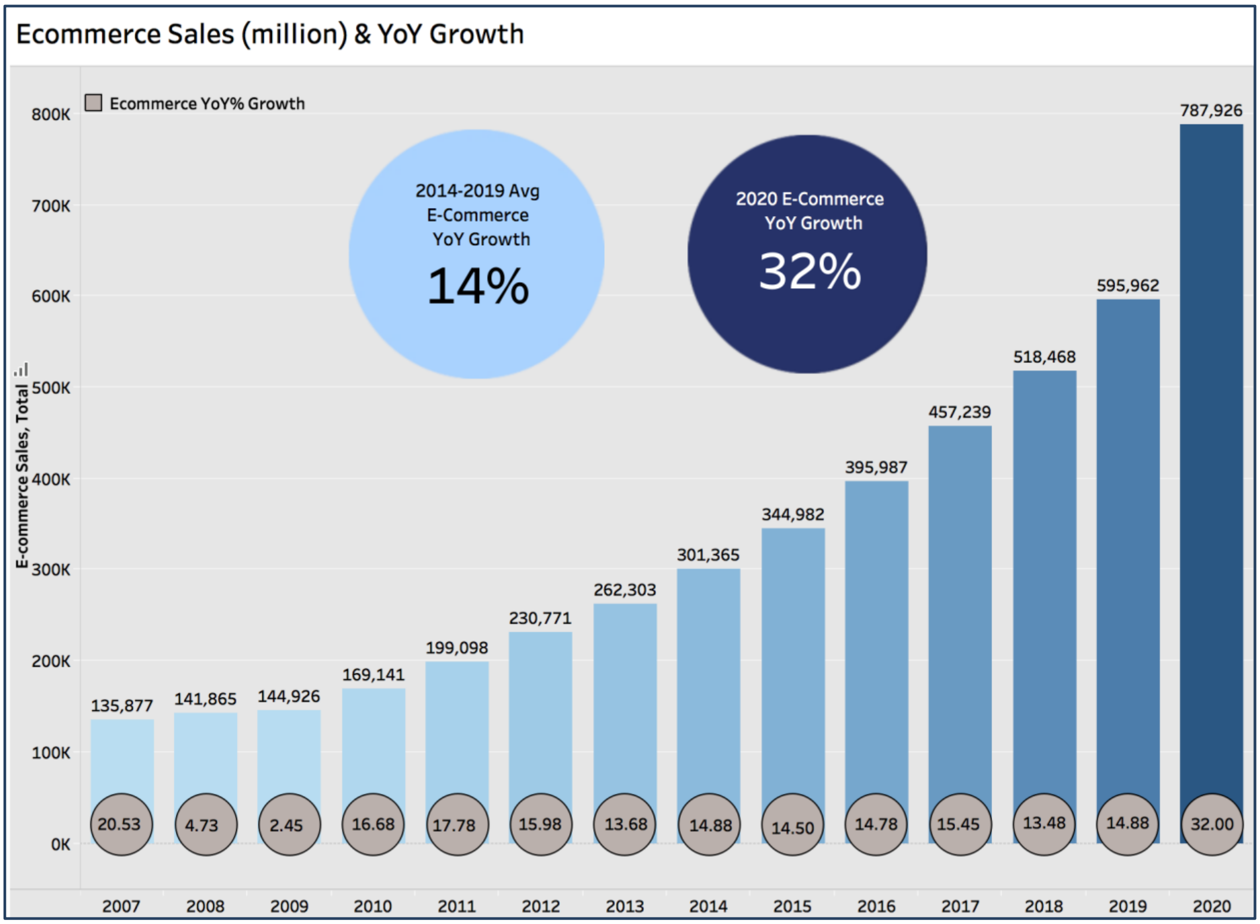
Predicted E-commerce sales values (2020Q2 to 2021Q4) using Facebook's Prophet model

```
> forcast = predict(m,future)
> tail(forcast[c('ds','yhat','yhat_lower','yhat_upper')])
           ds     yhat yhat_lower yhat_upper
89 2021-10-01 210512.9   203363.0   217071.4
90 2022-01-01 215162.9   207831.1   221840.3
91 2022-04-01 220223.2   212998.7   227766.7
92 2022-07-01 225578.3   218694.8   232760.4
93 2022-10-01 231126.5   224088.9   238794.5
94 2023-01-01 236828.2   229511.9   244501.0
```

Impact of Covid19 on US E-Commerce Sales [Insights]

US E-Commerce sales YoY Growth

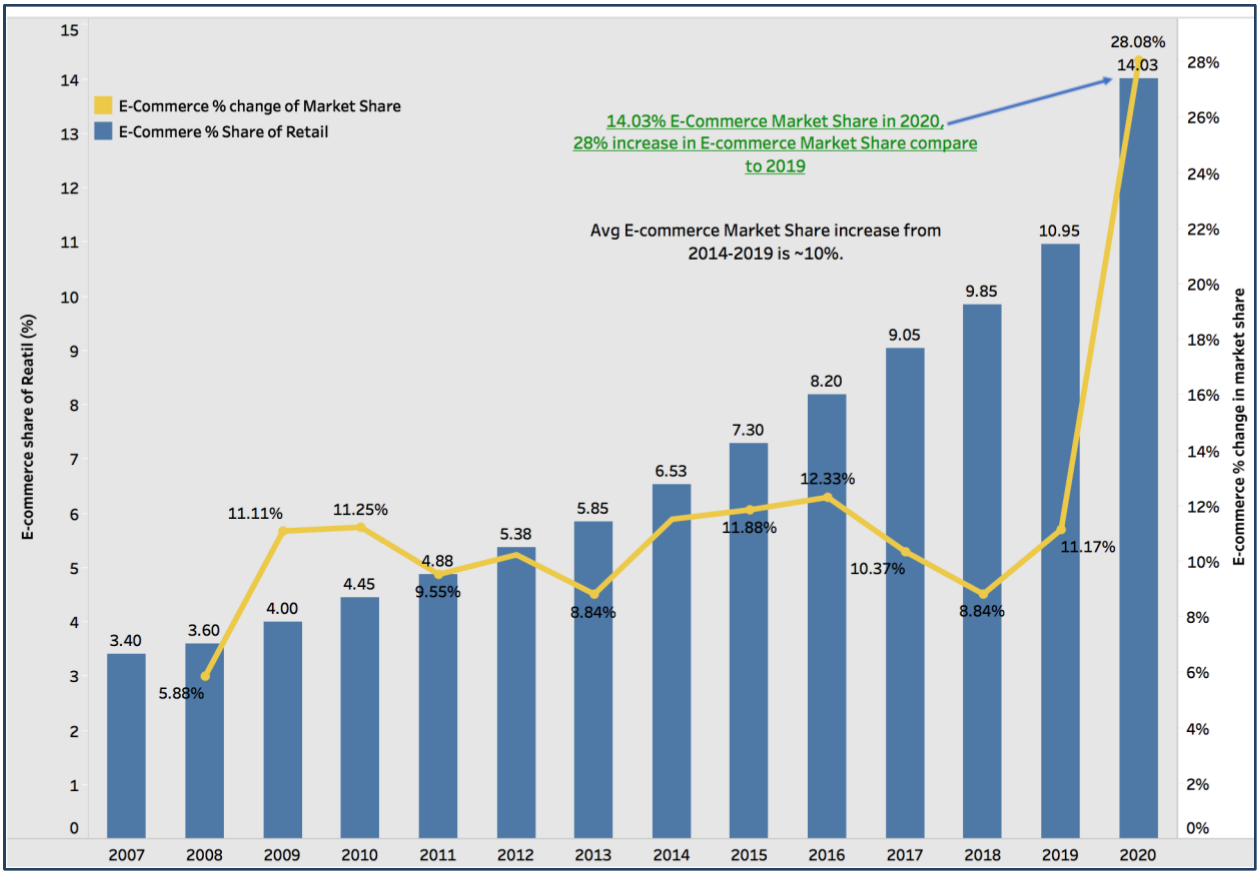
- In this section, we list down some of the interesting insights that has been discovered from the predictions that has been generated in previous section and also comparing those with actual e-commerce sales numbers, which outline the overall impact of Covid-19 on US E-Commerce Sales.



- From the above graph, we can see that US E-Commerce sales increased by a staggering 32% in Year 2020, which represents more than double YoY Growth from 2019, which is around 14.88%.

- ***In the previous 5 years before Covid from 2014-2019, avg US E-commerce sales growth was ~14%. Due to the impact of Covid in 2020, US E-commerce sales increased by more than double(~2.5x) to 32%.***

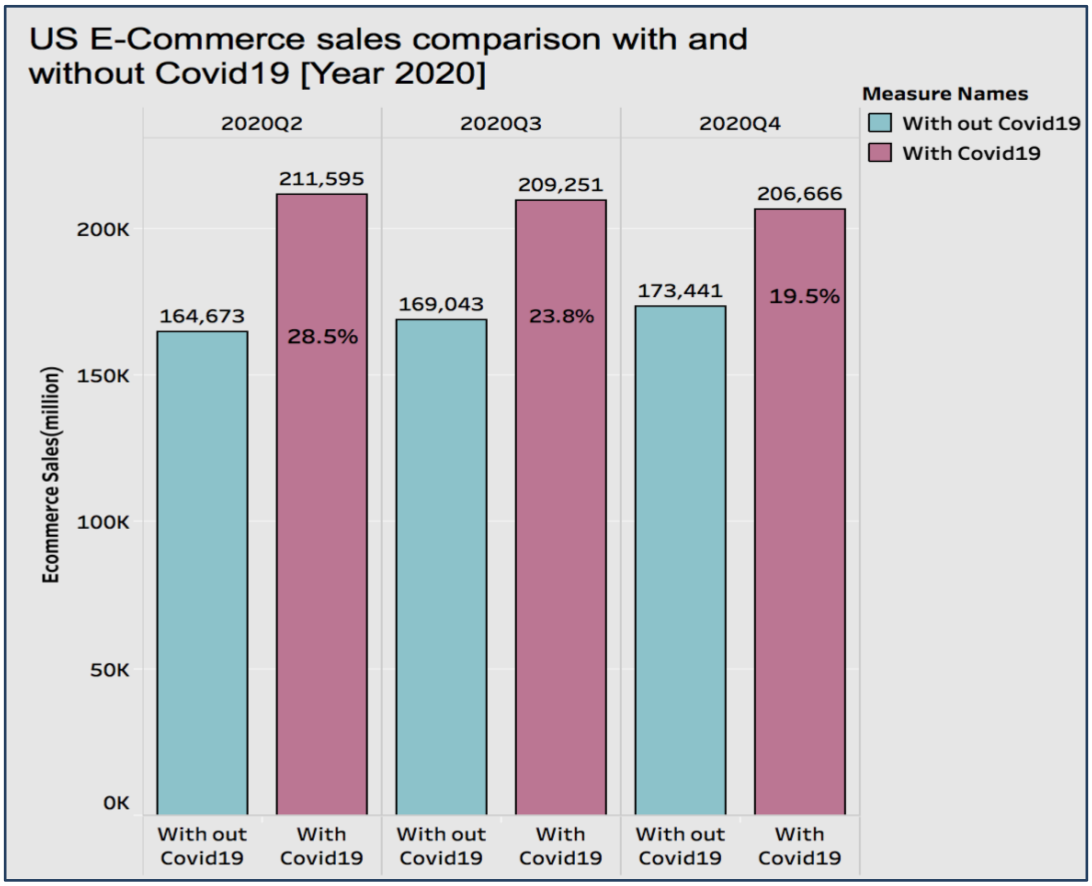
US E-Commerce Market Share YoY Growth



- From the above graph, we can see that US E-Commerce Market Share increased by 28% in Year 2020, which represents more than double YoY Growth in Market Share compared to 2019, which is around 11.7%.
- ***In the previous 5 years before Covid from 2014-2019, avg US E-commerce Market Share growth was ~10%. Due to the impact of Covid in 2020, US E-commerce Market sales increased by almost three times by 28%.***

US E-Commerce sales comparison with and without Covid [Year 2020]

- Below graph represents 2020 US E-Commerce sales comparison with and without Covid pandemic. The comparison has been done between 2020Q1 to 2020Q3 , since US shelter in place order started in March 2020 (start of Q2).
- We can see that due to pandemic E-Commerce Sales have seen 28.5%, 23.8% and 19.5% increase in 2020Q2, 2020Q3 and 2020Q4 respectively compared to before pandemic projected numbers.

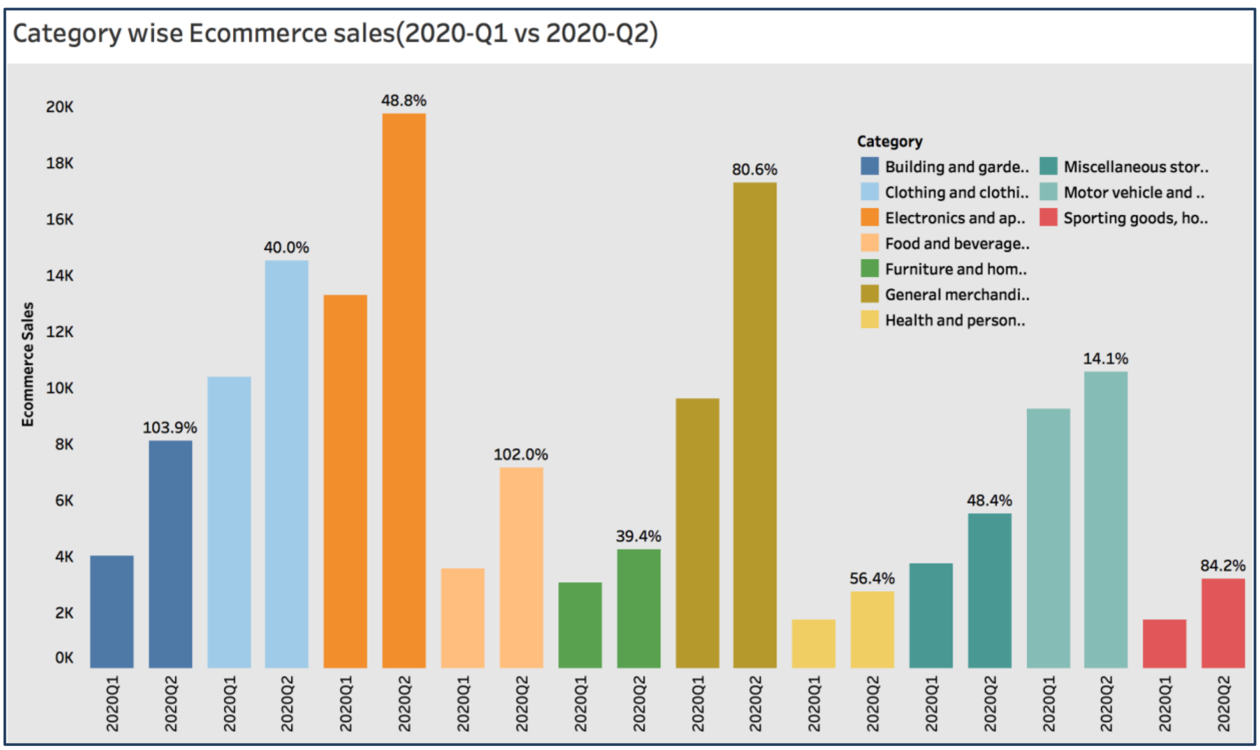


- US ecommerce saw an average quarterly Sales increase of ~44.5% throughout 2020 (compared to 2019).

Category wise US E-Commerce sales comparison [2020Q1 vs 2020Q2]

**[2020Q2 - shelter in place started]*

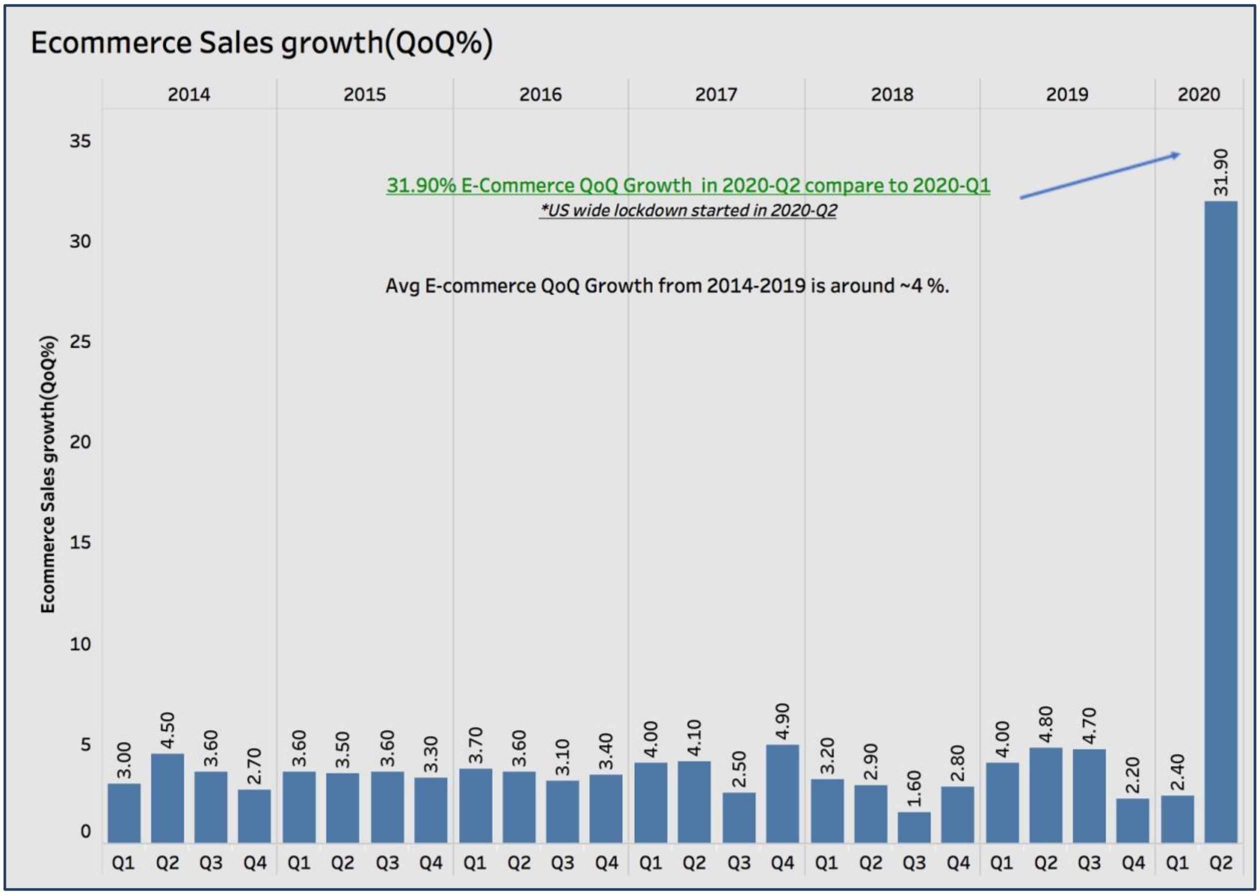
- Below graph provides category wise US E-Commerce sales comparison between 2021Q1 vs 2020Q2 to identify how covid has affected customer's purchasing habits and how customer spending has changed across various categories in shelter in place. The comparison has been done between 2020Q1 vs 2020Q2, since US shelter in place order started in March 2020 (start of Q2).
- Here, we can see that almost all categories have seen a rise in E-Commerce sales in 2020Q2 but there are certain categories that have seen sharp rise. **Food and Beverages** category **has seen increase of almost 102% in E-commerce sales in 2020Q2**, since more people started ordering food online due to shelter in place and dining restrictions.



- **General merchandise and groceries stores** like Walmart, Costco and others have seen almost **80% increase in E-commerce sales in 2020Q2**, since more people started avoiding going out for essential purchases due to safety concerns and started ordering those things online.
- Building Materials and Gardening categories has also seen a rise of about 100% in E-commerce sales, since most of the companies have announced work from home situations for the entire 2020, so people started spending more on building materials for home improvements and started doing in-home activities like gardening etc. for staying active.

US E-Commerce Sales QoQ Growth [Year 2020]

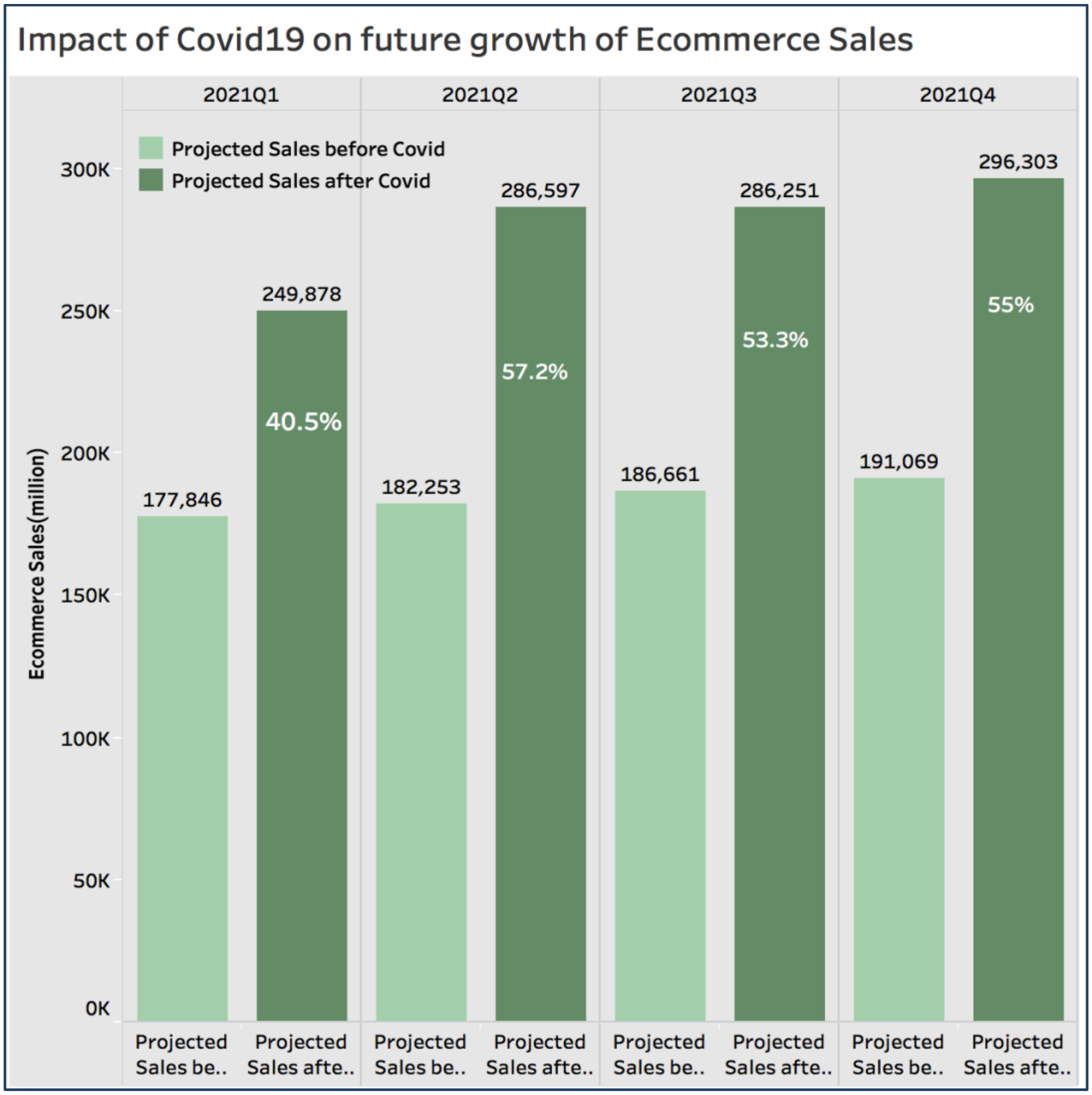
- Below graph provides US E-Commerce Sales QoQ growth from 2014Q1 to 2020Q2 to identify how covid has affected E-Commerce sales. The details has been provided till 2020Q2, since US shelter in place order started in March 2020 (start of Q2).



- Here, we can see that in 2020Q2 US E-Commerce sales have been increased by almost 31.90%. Average E-Commerce QoQ growth from 2014-2019 is around 4%.
- ***In the previous 5 years before Covid from 2014-2019, avg US E-commerce QoQ Sales growth was ~4%. Due to the impact of Covid in 2020, US E-commerce QoQ sales increased to 31.90%.***

Impact of Covid on future growth of US E-Commerce sales

- Below graph represents 2021 US E-Commerce sales comparison with and without Covid pandemic. The comparison has been done with Predicted E-Commerce Sales numbers for 2021 before and after Covid. This will help us evaluate the impact of Covid on future growth of E-Commerce sales in the US.



- Here, we can see that due to pandemic E-Commerce Sales in 2021 expected to be increased by avg 50% in all Quarters of 2021 compared to before pandemic projected numbers.

Conclusion

- The goal of the project is to evaluate the impact of Covid-19 on US E-Commerce sales, for this we built various E-Commerce sales prediction models with Time Series techniques by using the last 20 years of Sales data from the US Census.
- After evaluating accuracy of trained models on validation data set, we identified 3 models with highest accuracy as best forecasting models, which are Holt's-Winter, ARIMA and FB Prophet model. Based on the analysis and predictions, we outline some of the interesting impacts of Covid-19 pandemic on the E-Commerce industry.
- Consumers spent $787.926 billion online with U.S. retailers in 2020, up 32.0% from $595.962 billion in 2019. Online spending represented 28.08% of total retail sales last year, compared with 10.9% the year prior.
- Perhaps most striking, in 2020 E-commerce market share jumps to 14.03% of total US retail sales, representing almost 28% rise. In prior 5 years from 2014 to 2019, E- commerce market share increased averaged to only 10%.
- Changing consumer spending habits as a result of the coronavirus pandemic contributed to the spike in ecommerce sales across various categories last year. Food and Beverage, General merchandise and Home improvements and gardening are amongst the top categories who experienced close to 100% rise in online sales.
- The COVID-19 pandemic boosted U.S. online shopping by $193 billion. This figure represents the increase in online shopping during the months of March 2020, when the pandemic began in the U.S, through February 2021. During this time, U.S. consumers spent a total of $844 billion online. The pandemic itself produced a "rare step change in online spending, equivalent to a 20% boost," and noted this impact will continue even as the

pandemic comes to an end in the months to come as this has changed consumer's spending habit a lot.

- The pandemic has served as a great accelerant to E-commerce, pushing them 3 to 5 years ahead of where their natural growth would have otherwise taken them.

Bibliography

- https://www.census.gov/retail/index.html
- https://www.marketplacepulse.com/stats/us-ecommerce/us-e-commerce-growth-rate-55
- https://commonthreadco.com/blogs/coachs-corner/coronavirus-ecommerce?_ga=2.164983679.1592814449.1619211866-2071772300.1617151149#coronavirus-ecommerce-data
- https://www.statista.com/topics/6321/coronavirus-covid-19-impact-on-e-commerce-in-the-us/
- *Shmueli, G. and Lichtendahl Jr., K.C. Practical Time Series Forecasting with R, 2nd Edition, Axelrod Schnall Publishers, 2016. ISBN-13: 978-0-9978479-1-8.*

- ***Prof. Surendra Sarnikar***: *Thank you professor for your continuous support, guidance, feedback and invaluable knowledge that you shared during this capstone project.*