Course: STQD 6444

Topic: Car Brand Toyota Sales Norway analysis using decomposition, smoothing model forecasting and ARIMA

Name: Syed Norman Daniel Bin Syed Mahadir(P125754)

1.0 Data background and source

This dataset is collected by Opplysningsrdet for Veitrafikken (OFV) is Norwegian road association has collected all types of car(Electric car,hybrid,etc) in between 2007 to 2017. This dataset contains 5 columns and 4378 rows but our main case focus is in toyota brand sales in Norway is about around 121 rows and 5 columns.This dataset contains year,month,the brand of the car, quantity of car sales and percent share in monthly total and also dataset is obtained from kaggle.com.

The variables are namely as below table:

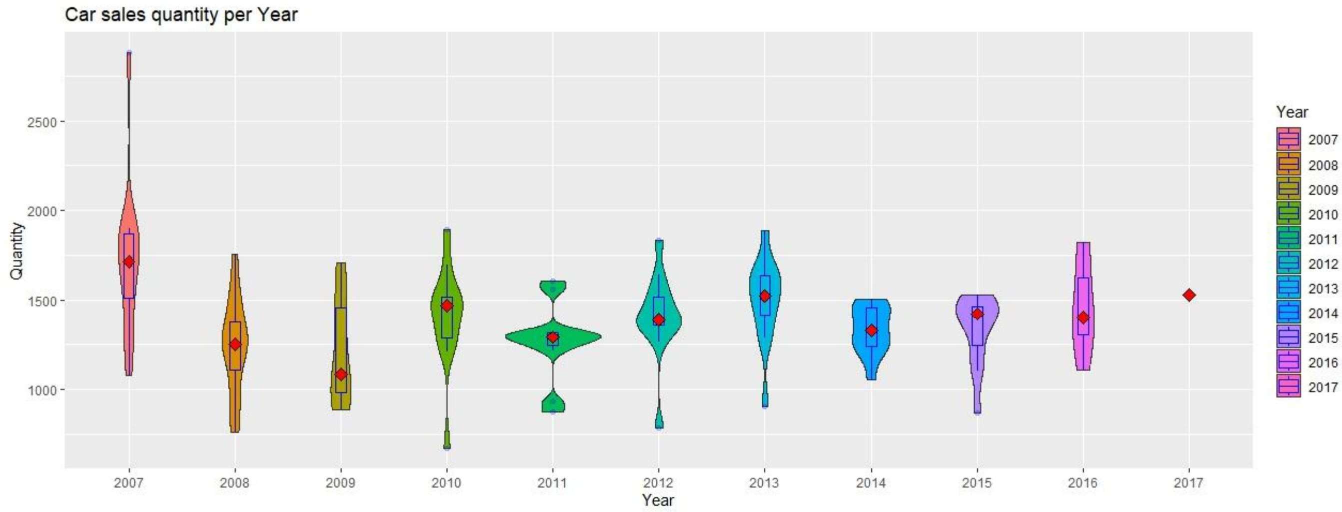
Variable Name	**Explanation**
Year	Year of sales Month
Month	Month of sales
Make	Car brand name (e.g. Volkswagen, Toyota, Tesla)
Quantity	Number of units sold
Pct	Percent share in monthly total

In this topic, there are one objectives that will be discussed:

1. To investigate the trend, seasonality, cyclic and random in car sales quantity through the year

2. To forecast the quantity of car sales of brand toyota

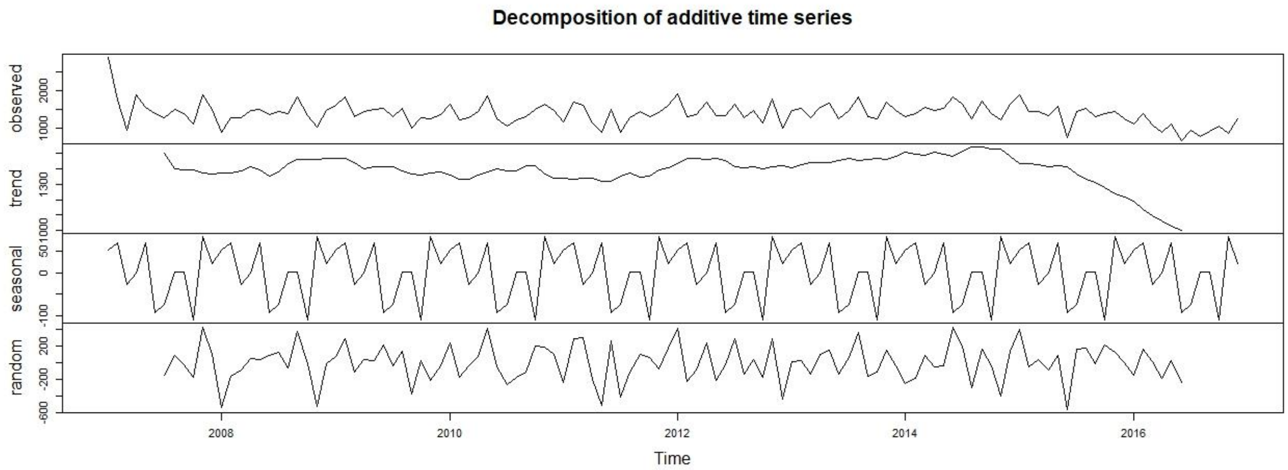
2.0 Description analysis of data



For the violin plot in quantity versus year, we can see that every year has different quantity sales of Toyota cars but for 2017 we cannot either boxplot r violin plot because the data for 2017 only contain jan only and else remaining not yet to be written,that's what why 2017, only shows one dot only. The red dots for every plot indicates that it is the median for every year and we can see that in 2007 has the highest median which means that it has a higher sales in that year and the lowest in 2009 at around 1100 . There are many main factors that in 2007 had good sales and after that year the sales cannot surpass the peak quantity sales, we can conclude that competition among other car brands as well as the appeal of the Honda brand is unable to appeal to the population in Norway for reasons that cannot challenge the sales in 2007.

In other parts of the wider spots, the wider spots indicate the frequent occurrence on that spot based on Y-axis value(Quantity). The good things in this dataset shows that there are no outliers which means that when we do a forecast there will be a good forecast because outliers can affect the statistical features of the data, making estimating the underlying patterns or trends difficult.By presenting a better and more consistent view of the data, eliminating outliers from the data can assist to enhance the accuracy and reliability of forecasting models.

3.0 Decomposition



Graph decomposition in this Car Sales brand Toyota, shows that from the graph time series that we observed its additive structure means that the amplitude of the seasonal component roughly remains the same over time and additive structure also can be used in forecast because it has a stationary time series. Below is the formula below is for additive structure:

Additive structure= Trend+Seasonal+Cyclical+Random

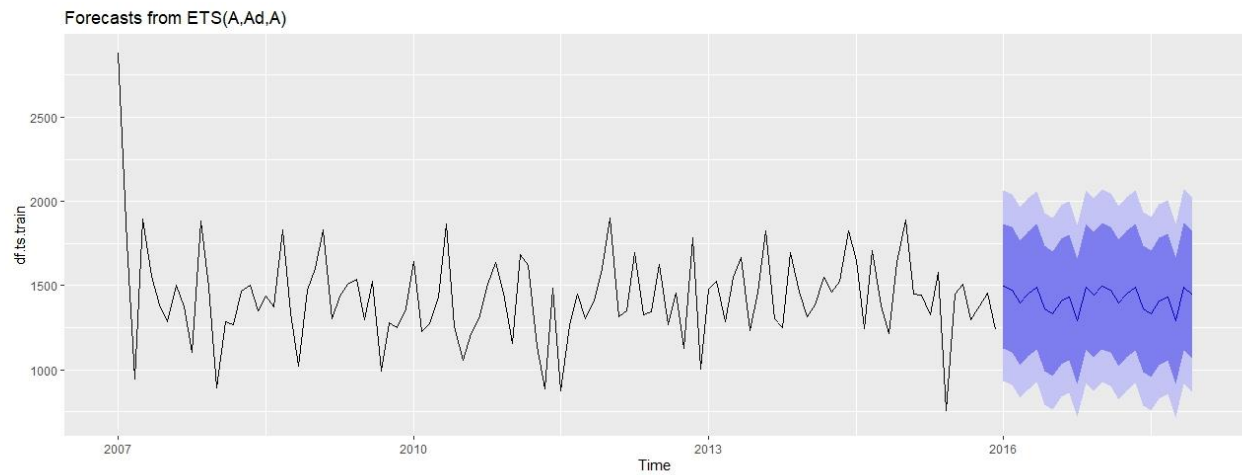
Next, the trend car sales quantity shows that it has a nearly flat trend between 2007 until 2015 means that it has statistically flat sales in those years and followed by another downward trend.Flat sales occur because Toyota needs to compete with other type cars which are more advanced(hybrid car,electric car) and also Norway applies high vehicle import duties and registration fees, making vehicles much more costly than in most other nations. Norway is essentially supporting EV sales at a level that other countries cannot afford by waiving these duties for electric cars.

The seasonal component of this graph of car quantity sales would catch the normal cyclical trends in car sales. Car purchases between the end and early of the year tend to be seen greater than middle of year due to maybe getting commission or bonus from the company and mostly likely buying a car in that particular month . Car sales, on the other hand, may be fewer during the middle months because people are less likely to have good financial money to purchase cars and mostly contribute to others things(Like:Travelling, Investment).

The unexpected and irregular graph changes in quantity sales that cannot be explained by either the trend or seasonal components would be represented by the random component of vehicle sales. This component may include any external variables that may have an effect on vehicle sales, such as economic changes, consumer behavior, or other factors.

4.0 Holt-Winters Seasonal smoothing

Since this data has seasonality and trend, we can use holt-winters seasonal smoothing method. The Holt-Winters Seasonal Method is used to generate forecasts using data with a pattern and seasonality. This technique can be done using either a "Additive" or a "Multiplicative" structure, depending on the data set. The Additive model works best when the seasonal pattern is consistent across the data collection. Seasonality and trend are present in this data; however, it is uncertain whether the seasonality is additive or multiplicative. To find the optimal fit model, we'll employ the Holt-Winters technique.



From the graph, we can see that there is a forecast using forecast() function and we identify this model as A,Ad,A as Addictive error,additive trend and additive seasonality since it has trend and seasonality.The forecast shows the 3 layers colors with different values.

```
> summary(ets.df.ts)
ETS(A,Ad,A)

Call:
 ets(y = df.ts.train, model = "AAA")

  Smoothing parameters:
    alpha = 0.041
    beta  = 1e-04
    gamma = 1e-04
    phi   = 0.9552

  Initial states:
    l = 1692.9651
    b = -11.2617
    s = 22.9123 70.7091 -132.869 8.0143 -16.7415 -90.3123
           -59.2865 70.101 25.9859 -25.1657 52.3197 74.3329

  sigma:  288.7496

     AIC     AICc      BIC
1746.934 1754.619 1795.212

Training set error measures:
                    ME     RMSE      MAE       MPE    MAPE      MASE        ACF1
Training set -6.550366 265.0515 193.5662 -3.872663 14.5058 0.6417446 -0.02047021
```

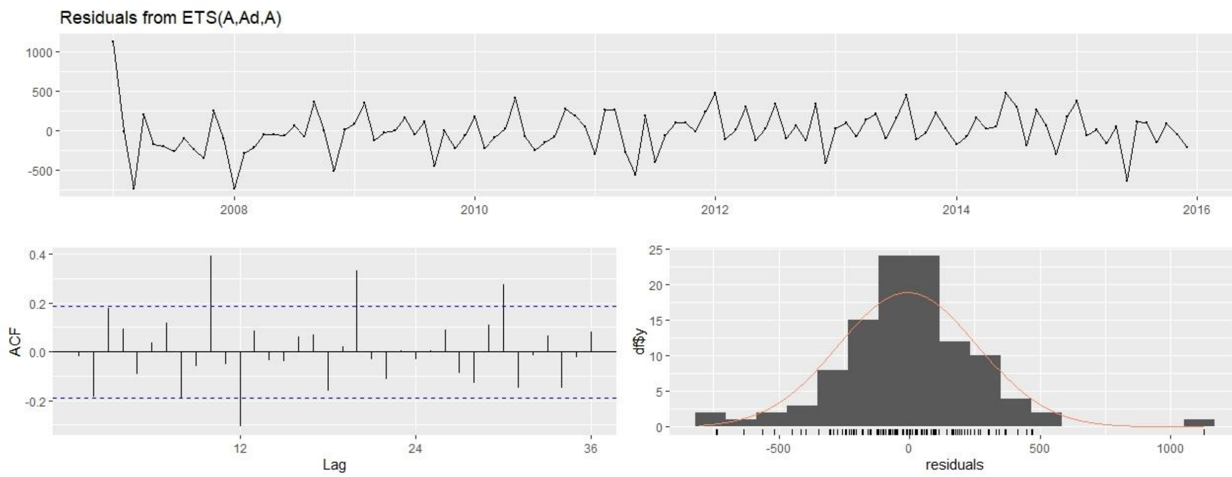
.After we plot the data, we need to find out alpha,beta and gamma value in our additive model.Below is the value for alpha, beta and gamma:

Alpha=0.041

Beta=0.0001

Gamma=0.0001

Next, we need to check out residuals to see if the residual grows larger over time or constant over time.



As we can see the residuals show the constant amplitude over time, meaning we no need change our ets to multiplicative models. Next we want to see predictive accuracy by using MAPE.

```
> #forecast 1 year
> df.ts.f<-forecast(ets.df.ts,h=24)
> accuracy(df.ts.f,df.ts.test)
                     ME     RMSE      MAE        MPE     MAPE
Training set  -6.550366 265.0515 193.5662  -3.872663 14.50580
Test set    -423.389647 460.0978 423.3896 -47.338226 47.33823
                  MASE        ACF1 Theil's U
Training set 0.6417446 -0.02047021        NA
Test set     1.4036955 -0.21747119   1.78765
>
```

The predictive accuracy is around 47.33%(according to MAPE).

5.0 ARIMA

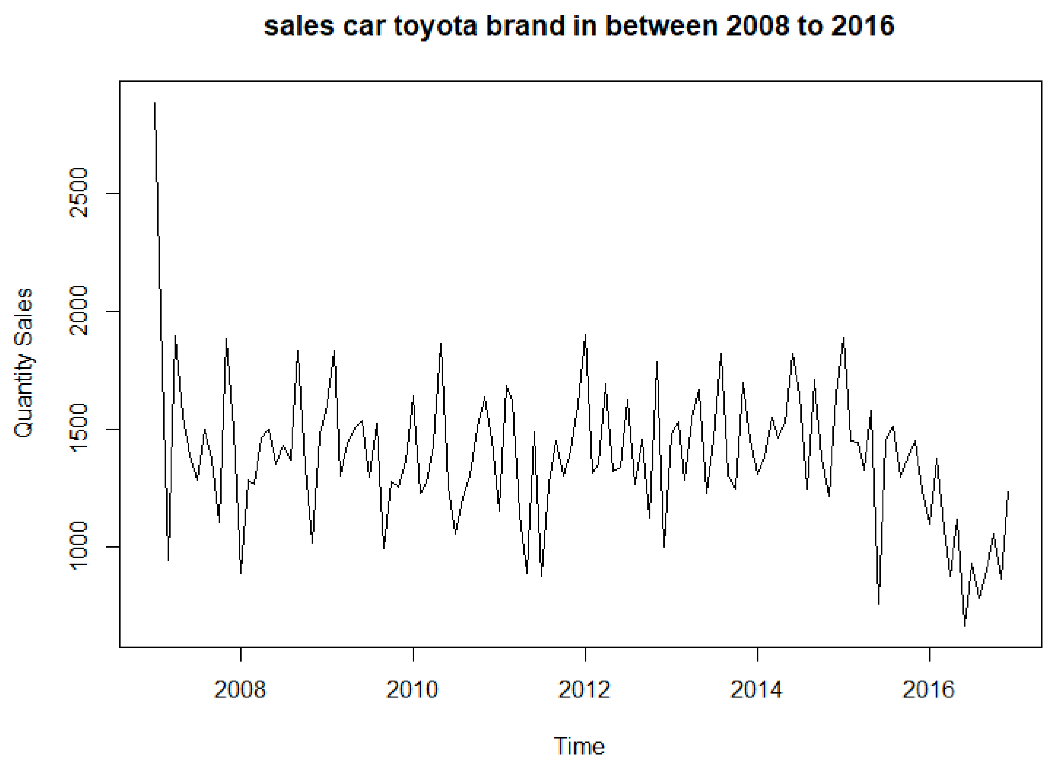


Figure above shows a visualization of time series in the Toyota Sales car between 2008 to 2016. As we can see for, beginning 2006 has a higher sales quantity car compare to another year and to be seen has statistically nearly flat sales in 2007 to 2015, and began to drop sales in early 2016.To forecast for this data set, we must check whether it is stationary data or not, we need to run the augmented dickey-fuller test.

```
> adf.test(df.ts)

        Augmented Dickey-Fuller Test

data:  df.ts
Dickey-Fuller = -2.9925, Lag order = 4, p-value = 0.1643
alternative hypothesis: stationary
```

The Augmented Dickey -fuller test shows p-value>0.05 which 0.05 is critical value , indicating that there is strong evidence to support H0.So it is a stationary state.

Next we need the best ARIMA model, we use the function auto.arima() to show the best ARIMA model. As we can see, ARIMA(0,1,1)(1,0,0)[12] is the best model . The best model can be concluded by looking at the smallest AIC.

```
> Model<-auto.arima(df.ts,ic="aic",trace = TRUE)

 ARIMA(2,1,2)(1,0,1)[12] with drift         : 1684.655
 ARIMA(0,1,0)            with drift         : 1741.795
 ARIMA(1,1,0)(1,0,0)[12] with drift         : 1714.925
 ARIMA(0,1,1)(0,0,1)[12] with drift         : 1682.785
 ARIMA(0,1,0)                               : 1739.972
 ARIMA(0,1,1)            with drift         : 1688.095
 ARIMA(0,1,1)(1,0,1)[12] with drift         : 1684.323
 ARIMA(0,1,1)(0,0,2)[12] with drift         : 1684.044
 ARIMA(0,1,1)(1,0,0)[12] with drift         : 1682.45
 ARIMA(0,1,1)(2,0,0)[12] with drift         : 1684.191
 ARIMA(0,1,1)(2,0,1)[12] with drift         : 1683.733
 ARIMA(0,1,0)(1,0,0)[12] with drift         : 1737.639
 ARIMA(1,1,1)(1,0,0)[12] with drift         : 1684.23
 ARIMA(0,1,2)(1,0,0)[12] with drift         : 1684.086
 ARIMA(1,1,2)(1,0,0)[12] with drift         : 1685.781
 ARIMA(0,1,1)(1,0,0)[12]                    : 1682.292
 ARIMA(0,1,1)                               : 1687.985
 ARIMA(0,1,1)(2,0,0)[12]                    : 1684.053
 ARIMA(0,1,1)(1,0,1)[12]                    : 1684.178
 ARIMA(0,1,1)(0,0,1)[12]                    : 1682.702
 ARIMA(0,1,1)(2,0,1)[12]                    : 1683.601
 ARIMA(0,1,0)(1,0,0)[12]                    : 1735.814
 ARIMA(1,1,1)(1,0,0)[12]                    : 1684.037
 ARIMA(0,1,2)(1,0,0)[12]                    : 1683.855
 ARIMA(1,1,0)(1,0,0)[12]                    : 1713.255
 ARIMA(1,1,2)(1,0,0)[12]                    : 1685.505

 Best model: ARIMA(0,1,1)(1,0,0)[12]
```

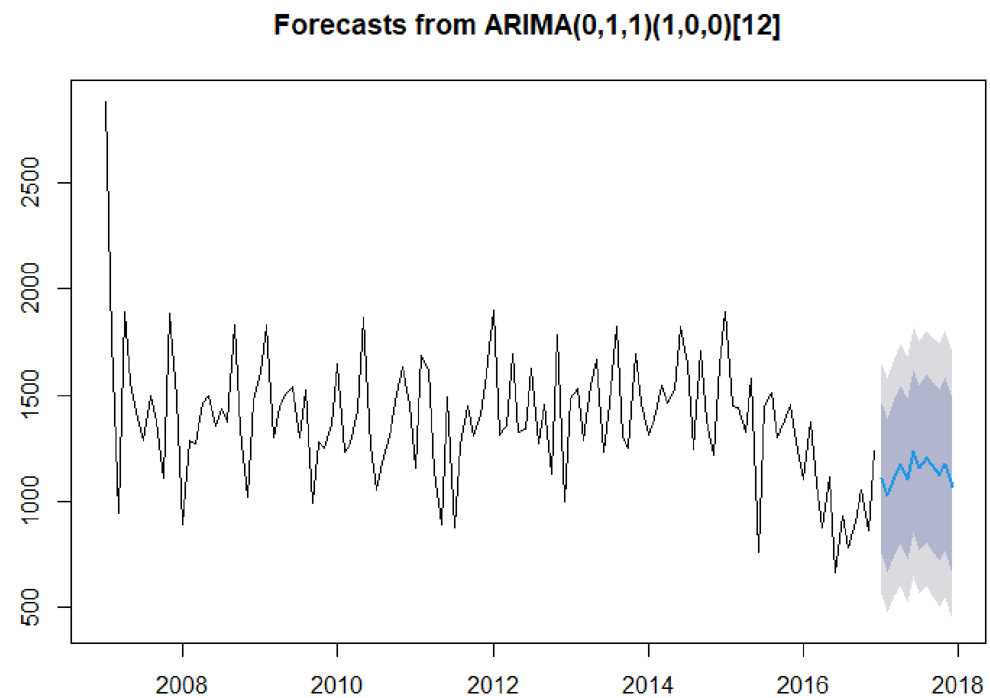
```
Call:
arima(x = df.ts, order = c(0, 1, 1), seasonal = c(1, 0, 0))

Coefficients:
          ma1     sar1
      -0.8175  -0.3016
s.e.   0.0568   0.1051

sigma^2 estimated as 75302:  log likelihood = -838.15,  aic = 1682.29

Training set error measures:
                    ME     RMSE      MAE       MPE     MAPE
Training set -54.25455 273.2673 209.0569 -7.417857 16.82162
                  MASE      ACF1
Training set 0.7239907 0.055285
```

Forecasting Arima



The figures above show forecasts based on ARIMA models of Norway quantity sales brand Toyota cars in between 2008 to 2018.The dataset only has until the end of month 2016 and we are doing some forecasting in one year about quantity sales. As we can see the actual sales in around 1000 to 1250 something and below the par average sales quantity car. Below is data for forecasting.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2017	1108.395	753.7275	1463.063	565.9778	1650.812
Feb 2017	1026.048	665.5226	1386.574	474.6717	1577.425
Mar 2017	1114.729	748.4388	1481.020	554.5363	1674.922
Apr 2017	1176.565	804.5988	1548.531	607.6919	1745.437
May 2017	1103.870	726.3146	1481.426	526.4485	1681.292
Jun 2017	1239.305	856.2405	1622.369	653.4585	1825.151
Jul 2017	1159.673	771.1783	1548.168	565.5216	1753.825
Aug 2017	1204.315	810.4649	1598.165	601.9732	1806.657
Sep 2017	1167.516	768.3817	1566.650	557.0930	1777.938
Oct 2017	1122.572	718.2234	1526.920	504.1742	1740.970
Nov 2017	1179.581	770.0843	1589.078	553.3098	1805.852
Dec 2017	1067.373	652.7915	1481.954	433.3256	1701.420

6.0 Conclusion

As a conclusion, we can conclude that:

- This dataset has a downward trend although in the middle of year has nearly flat sales quantity, has seasonality , cyclic and random.
- The actual forecast value is in the range between 1000 to 1250 in quantity sales .