Case of study: **<u>“Tobacco data “</u>**

Alumni:

A01686401 Gonzalez Carreon Edgar
A01275180 Hernández Spínola Alexis

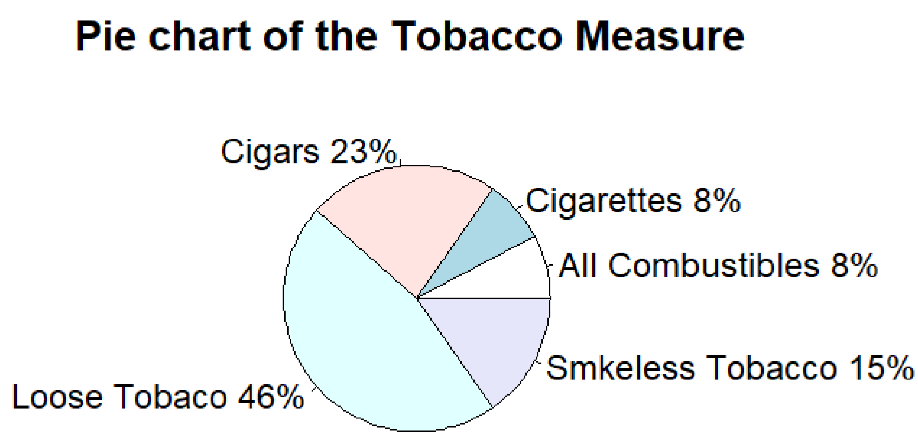
March 06, 2022

```
> Tobacco_Consumption <- read_excel("C:/Users/ediek/Desktop/DSCPy/Tobacco_Consumption.xlsx")
> tc <- Tobacco_Consumption
> summary(tc)
     Year      LocationAbbrev     LocationDesc        Population
 Min.   :2000   Length:273         Length:273         Min.   :209786736
 1st Qu.:2005   Class :character   Class :character   1st Qu.:222003984
 Median :2010   Mode  :character   Mode  :character   Median :235153929
 Mean   :2010                                         Mean   :234547860
 3rd Qu.:2015                                         3rd Qu.:247773709
 Max.   :2020                                         Max.   :256662010
    Topic             Measure           Submeasure        Data Value Unit
 Length:273         Length:273         Length:273         Length:273
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


    Domestic            Imports              Total
 Min.   :1.725e+06   Min.   :8.369e+03   Min.   :1.927e+06
 1st Qu.:3.754e+07   1st Qu.:1.439e+06   1st Qu.:4.036e+07
 Median :2.302e+09   Median :3.112e+08   Median :2.622e+09
 Mean   :5.244e+10   Mean   :3.069e+09   Mean   :5.550e+10
 3rd Qu.:1.014e+10   3rd Qu.:2.412e+09   3rd Qu.:1.335e+10
 Max.   :4.372e+11   Max.   :2.473e+10   Max.   :4.507e+11
 Domestic Per Capita Imports Per Capita Total Per Capita
 Min.   :  0.000    Min.   : 0.00     Min.   :  0.000
 1st Qu.:   0.099    1st Qu.: 0.00     1st Qu.:   0.101
```

```
 Median :  11.000    Median :  1.00     Median :  11.000
 Mean   : 226.650    Mean   : 13.32     Mean   : 240.012
 3rd Qu.:  44.000    3rd Qu.: 10.00     3rd Qu.:  54.000
 Max.   :2084.000    Max.   :128.00     Max.   :2148.000
> # data of last 20 years
> str(tc)
tibble [273 x 14] (S3: tbl_df/tbl/data.frame)
 $ Year              : num [1:273] 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
 $ LocationAbbrev    : chr [1:273] "US" "US" "US" "US" ...
 $ LocationDesc      : chr [1:273] "National" "National" "National" "National" ...
 $ Population        : num [1:273] 2.1e+08 2.1e+08 2.1e+08 2.1e+08 2.1e+08 ...
 $ Topic             : chr [1:273] "Noncombustible Tobacco" "Combustible Tobacco" "Combustible Tobacco" "Combustible Tobacco" ...
 $ Measure           : chr [1:273] "Smokeless Tobacco" "Cigarettes" "Cigars" "Loose Tobacco" ...
 $ Submeasure        : chr [1:273] "Chewing Tobacco" "Cigarette Removals" "Total Cigars" "Total Loose Tobacco" ...
 $ Data Value Unit   : chr [1:273] "Pounds" "Cigarettes" "Cigars" "Cigarette Equivalents" ...
 $ Domestic          : num [1:273] 4.55e+07 4.23e+11 5.61e+09 8.29e+09 1.68e+07 ...
 $ Imports           : num [1:273] 9.20e+04 1.23e+10 5.48e+08 7.03e+08 1.43e+06 ...
 $ Total             : num [1:273] 4.56e+07 4.36e+11 6.16e+09 8.99e+09 1.83e+07 ...
 $ Domestic Per Capita: num [1:273] 0.217 2018 27 40 0 ...
 $ Imports Per Capita : num [1:273] 0 59 3 3 0 0 0 2 0 2 ...
 $ Total Per Capita   : num [1:273] 0.217 2076 29 43 0 ...
> #8 numerical variables, the rest are categorical
> ncol(tc)
[1] 14
> nrow(tc)
[1] 273
> #6 categorical variables
```

```
> sum(!duplicated(tc))
[1] 273
> sum(duplicated(tc))
[1] 0
> sum(!complete.cases(tc))
[1] 0
> #clean data
> MT <- round(100* (table(tc$Measure)/length(tc$Measure)))
> MT

 All Combustibles       Cigarettes          Cigars     Loose Tobacco
            8              8              23             46
Smokeless Tobacco
           15
> label <- c("All Combustibles","Cigarettes","Cigars","Loose Tobaco","Smkeless Tobacco")
> label <- paste(label,MT)
> label <- paste(label,"%", sep = "")
```

> pie(table(tc$Measure),label, main = "Pie chart of the Tobacco Measure ")



As we can appreciate the first variable that it presented to us as categorical variable is call “Tobacco Measure”. We decided to plot with a pie graph, and as a result it shows us that most of the population is focused on Loose Tobacco, and a 38% is divided in tow categories: “Cigars” and “Smkeless” and just 16% consumes cigarettes and all combustibles.

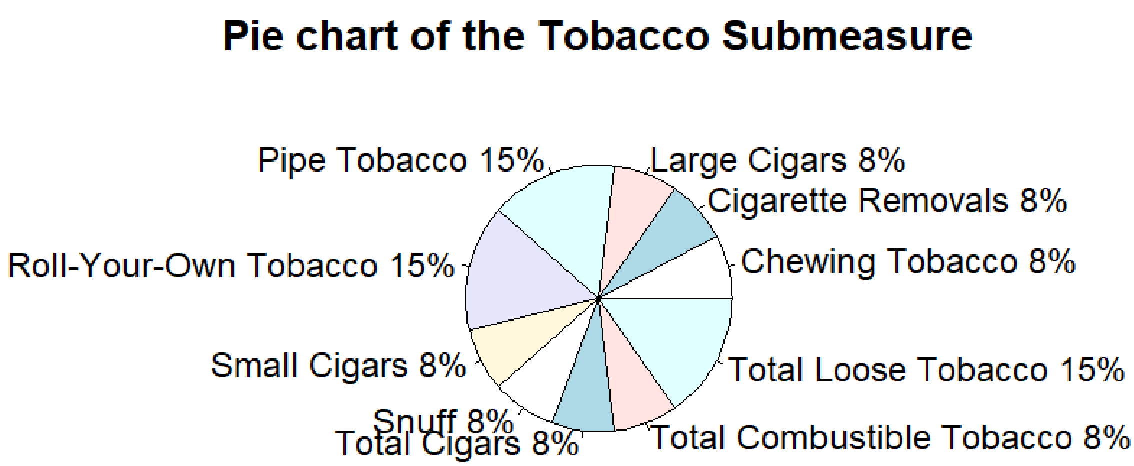
With this first view to the data, we can see that Loose tobacco & Cigars represent the 69% of our measure. One question we could make is why 69% is focused on those two groups? also is it diminishing the consumption of cigarettes or why it is just represented by the 8% of our measure?

> SM <- round(100*(table(tc$Submeasure)/ length(tc$Submeasure)))

> SM

```
   Chewing Tobacco      Cigarette Removals           Large Cigars
          8                   8                    8
     Pipe Tobacco    Roll-Your-Own Tobacco           Small Cigars
         15                  15                    8
            Snuff          Total Cigars Total Combustible Tobacco
          8                   8                    8
```

```
     Total Loose Tobacco
                   15
> label <- c("Chewing Tobacco","Cigarette Removals","Large Cigars","Pipe Tobacco","Roll-Your-Own Tobacco","Small Cigars","Snuff","Total Cigars","Total Combustible Tobacco","Total Loose Tobacco")

> label <- paste(label,SM)

> label <- paste(label,"%", sep = "")

> pie(table(tc$Submeasure),label, main = "Pie chart of the Tobacco Submeasure")
```



In the second categorical variable that we decided to analyze it represent a sub measure of the first variable that we analyze. This graph shows us a more detail about the consumption of tobacco of our measure.

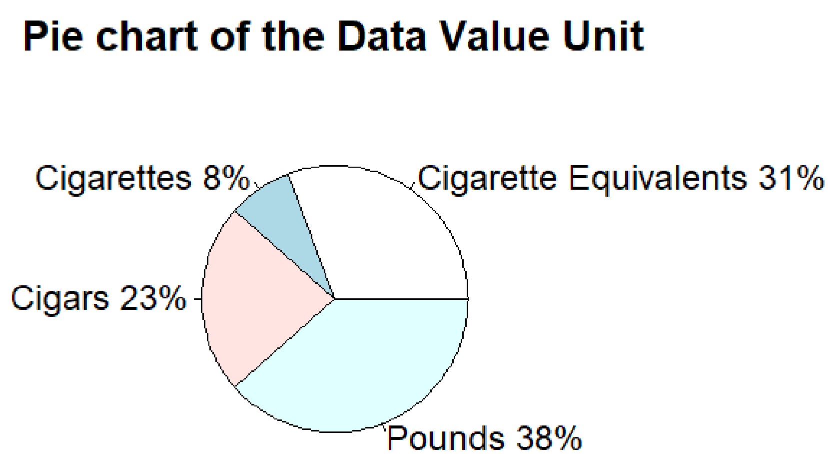
As we can see the 45% is focused on the type of "Pide Tobacco", "Roll-Your-own-Tobacco" and "Total loose Tobacco" and rest of the subcategories each one represents 8% of our data.

What we can interpretate from this graph is probably some preferences about how our measure prefers to consume the product. Maybe marketing should promote more on the three subcategories that the market prefers or maybe to stop some production line of the subcategories that represent 8% or less of the market.

As company we should start thinking why three subcategories possess almost the doble of percentage compared with the rest of the categories?

```
> DVU <- round(100*(table(tc$`Data Value Unit`)/ length(tc$`Data Value Unit`)))
```

```
> DVU

Cigarette Equivalents          Cigarettes            Cigars
               31                    8                 23
            Pounds
               38
> label <- c("Cigarette Equivalents","Cigarettes","Cigars","Pounds")
> label <- paste(label,DVU)
> label <- paste(label,"%",sep = "")
> pie(table(tc$`Data Value Unit`),label, main = "Pie chart of the Data Value Unit")
```



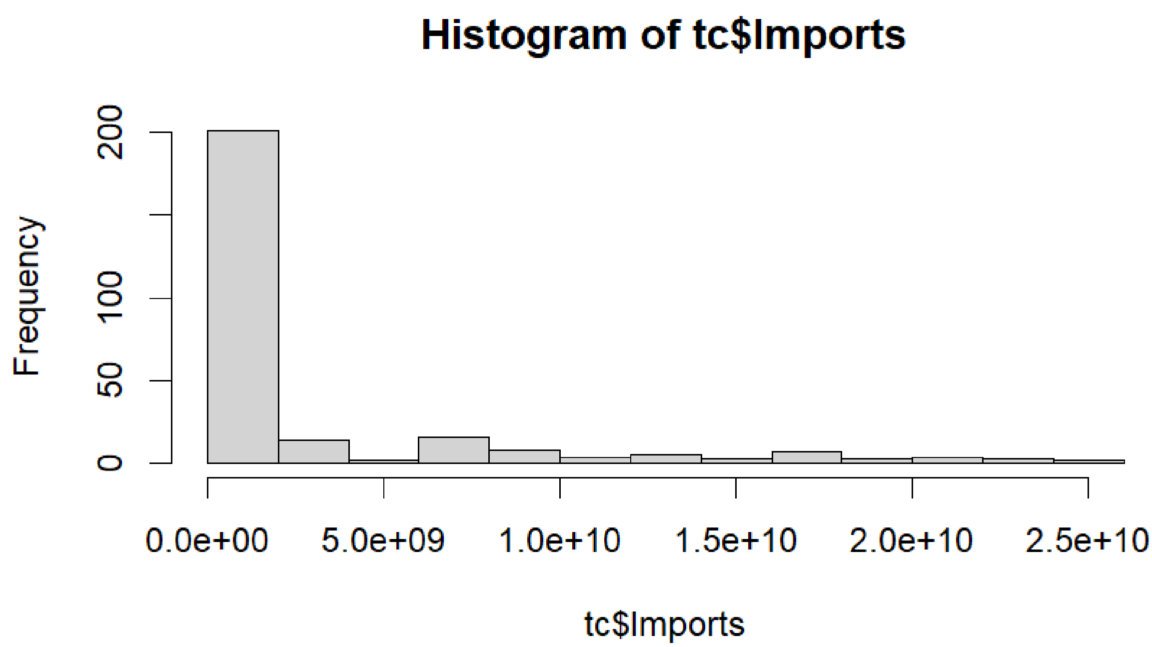
Finally with this third categorical variable we can appreciate four ways that probably the market prefers to consume their product. We can appreciate that 69% of the market is focused in “Pounds” & “Cigarette Equivalents” and we can appreciate that cigarettes represent the lowest size of the market and in the variable that we analyze before this one, cigarette is also represented in one small portion of the market.

What we can infer with this variable is that the market poses some clear preferences about how to consume the product and cigarettes are not popular as they used to be.

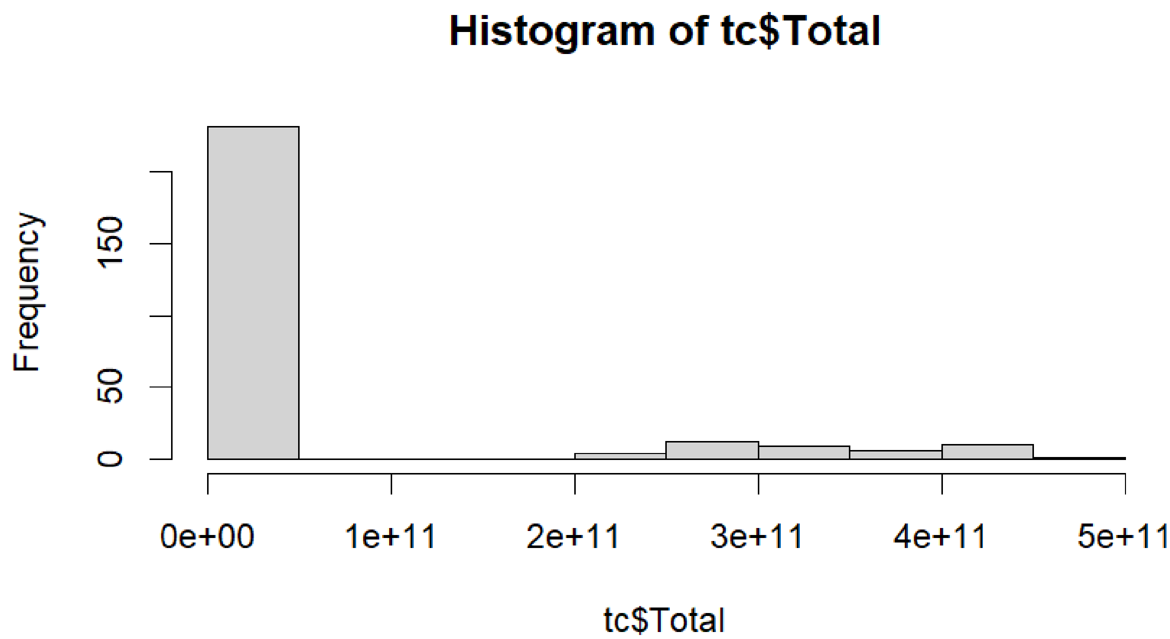
So if the main product of the company are cigarettes probably they should star seeing what the market is preferring to consume or we can appreciate as the tobacco market behavior in the US.

> #if i try to analyze some numerical variables with a box&plots or histogram tends to show lots of otuliers and some behavior like a gamma distribution, that's why is decided to analyze the numerical data with some time series, before we do that, we will show you two examples of what is mention before about the outliers.

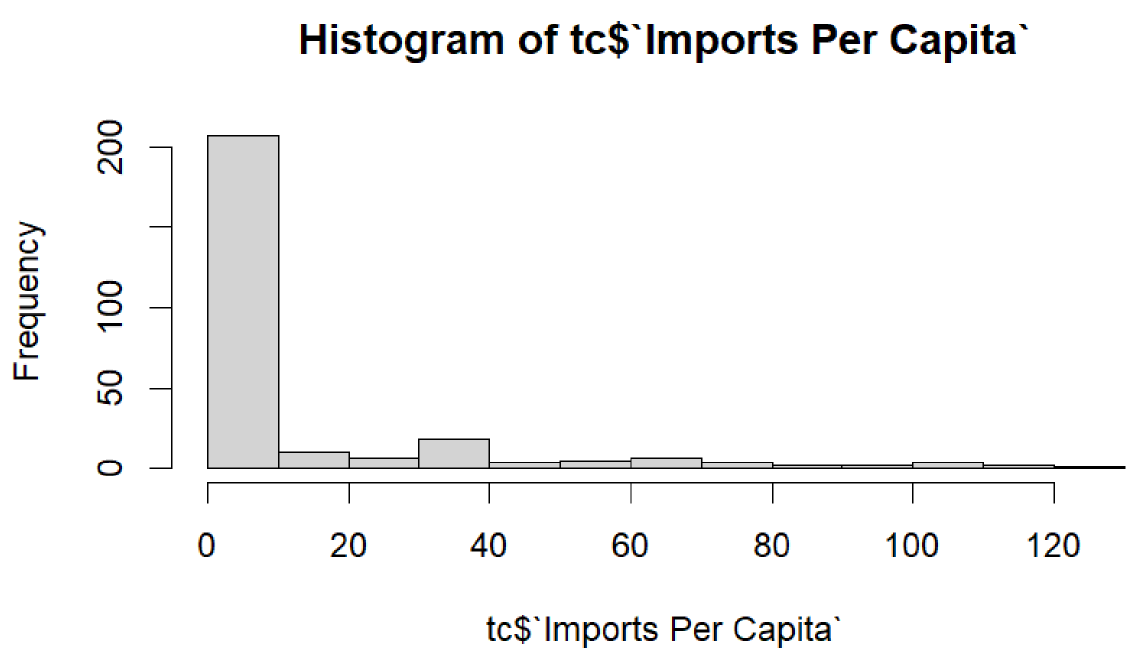
> hist(tc$Imports)



> hist(tc$Total)



> hist(tc$`Imports Per Capita`)



> # there are lots of outliers in the numerical variables. Let´s go and see some correlation

> install.packages("psych")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/

Installing package into ‘C:/Users/ediek/Documents/R/win-library/4.1’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.1/psych_2.1.9.zip'

Content type 'application/zip' length 4244452 bytes (4.0 MB)

downloaded 4.0 MB

package ‘psych’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\ediek\AppData\Local\Temp\Rtmpkp0hSa\downloaded_packages

> cor 1 <- cor(tc$Topic,tc$Submeasure)

Error: unexpected numeric constant in "cor 1"

> cor1 <- cor(tc$Year,tc$Total)

> cor1

[1] -0.07309212

> #years & total goes in different directions, or poses an inverse correlation

> cor1 <- cor(tc$Total,tc$`Total Per Capita`)

> cor1

[1] 0.9978109

> # Total & total per capita poses a strong positive correlation, if the total diminishes then the total per capita will diminish as well and it make sense because both comes from the same data with the same quantity the only difference is that per capita data is divided by population.

> cor2 <- cor(tc$Population,tc$`Total Per Capita`)

> cor2

[1] -0.09987042

> #makes sense that if the population diminish the total per capita increase because same amount of money with less people is equal to more money per habitant.

> cor3 <- cor(tc$Domestic,tc$Imports)

> cor3

[1] 0.8966838

> cor4 <- cor(tc$Domestic,tc$`Domestic Per Capita`)

> cor4

[1] 0.9976981

> #makes sense that each of the cor 3 possess a strong correlation.

Now we want to make a time series analysis. To achieve this goal first is needed to group the data by years because we only have years as a Date, not more specific time. Due to the problem of outliers, the best way to appreciate the data is with mean & median (because it is not sensitive to outliers) and see the trends, seasons, cicles and what happens

Data addressing the mean

> new_data <- tc %>% group_by(Year) %>% summarise(Year, mean_Dom = mean(Domestic), mean_Imp = mean(Imports), mean_Total = mean(Total)) %>% unique()

`summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.

> new_data

A tibble: 21 x 4

Groups: Year [21]

Year   mean_Dom   mean_Imp   mean_Total

<dbl>   <dbl>   <dbl>   <dbl>

1 2000 68335266449. 2184249582. 70519516031

2 2001 66265925336. 2619347357. 68885272693.

3 2002 63860561082. 3579422188. 67439983270.

4 2003 61269300051. 3932165895. 65201465946.

5 2004 61205829509. 3853052279. 65058881789.

6 2005 59948584075. 3330456706 63279040781.

7 2006 60305655741. 3025268508. 63330924249

8 2007 58114472773. 2719497527. 60833970300.

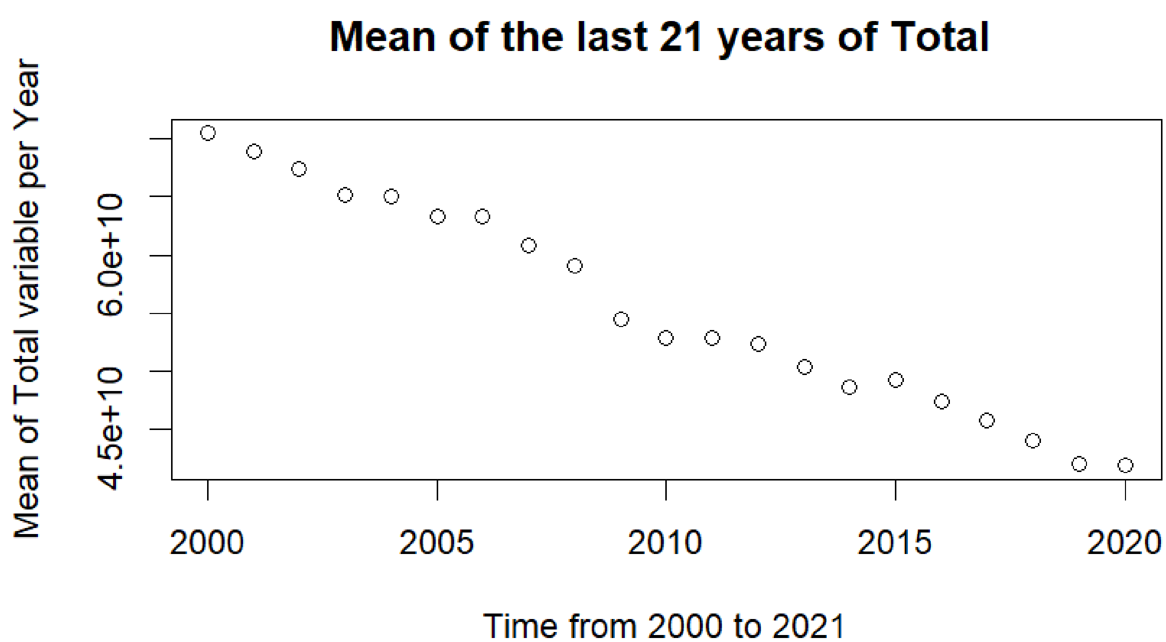
9 2008 56617916488. 2424823114. 59042758063.

10 2009 52327312584. 2196646723. 54523959307.

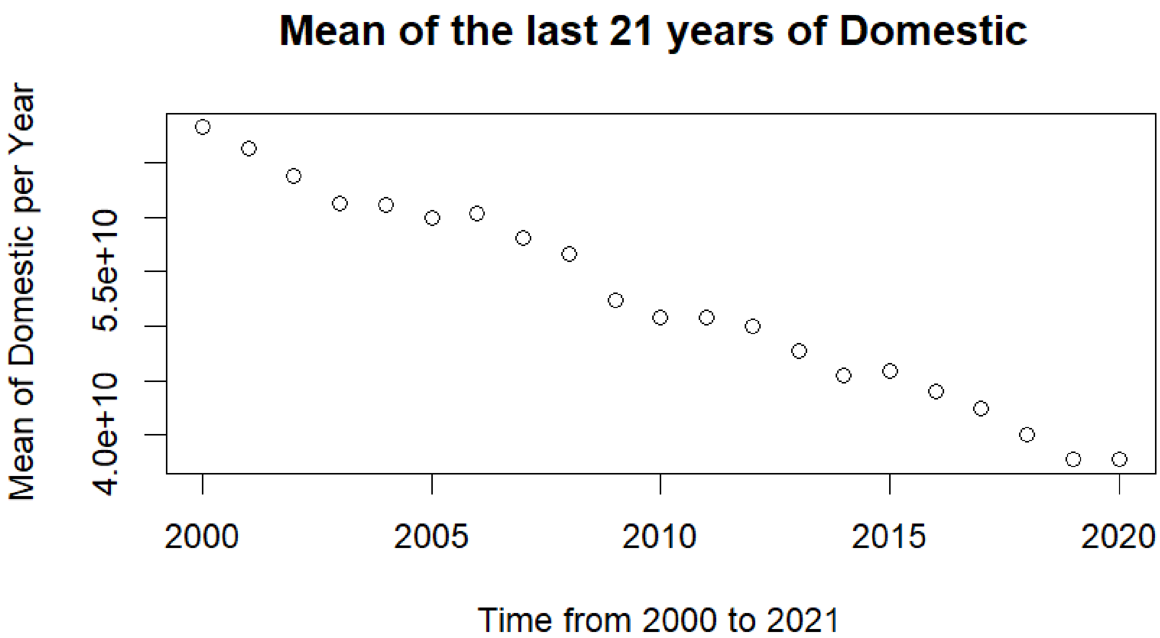
... with 11 more rows

#Lets plot some graphs

> plot(new_data$Year,new_data$mean_Total, ylab = "Mean of Total variable per Year", xlab = "Time from 2000 to 2021" , main = "Mean of the last 21 years of Total ")

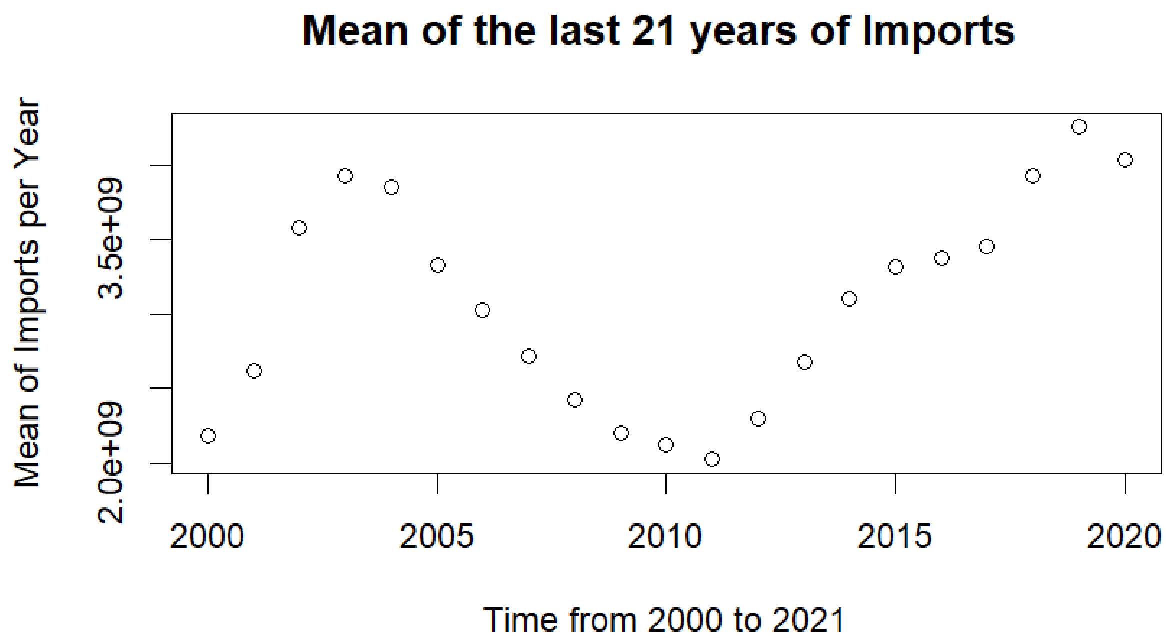


> plot(new_data$Year,new_data$mean_Dom, ylab = "Mean of Domestic per Year", xlab = "Time from 2000 to 2021" , main = "Mean of the last 21 years of Domestic ")



##Untile this moment the data is showing a decrement

> plot(new_data$Year,new_data$mean_Imp, ylab = "Mean of Imports per Year", xlab = "Time from 2000 to 2021" , main = "Mean of the last 21 years of Imports ")



In this plot we can appreciate a probable seasonality but before we try to do more analysis lets see what happen with the median

> new_data <- tc %>% group_by(Year) %>% summarise(Year, mead_Dom = median(Domestic), mead_Imp = median(Imports), mead_Total = median(Total)) %>% unique()

`summarise()` has grouped output by 'Year'. You can override using the `.groups` argument.

> new_data

A tibble: 21 x 4

Groups: Year [21]

```
   Year  mead_Dom  mead_Imp mead_Total
  <dbl>     <dbl>     <dbl>      <dbl>
 1  2000 2635167015 338489600 2999419077
 2  2001 2421344492 454477785 2914666831
 3  2002 2278010585 479059692 2757070277
 4  2003 2301972488 343673108 2474341488
 5  2004 2701646262 354785477 2917089262
 6  2005 3772041108 507187692 3968494108
 7  2006 4162220407 561063385 4434095407
```

8 2007 4658667130 816734031 5160825473

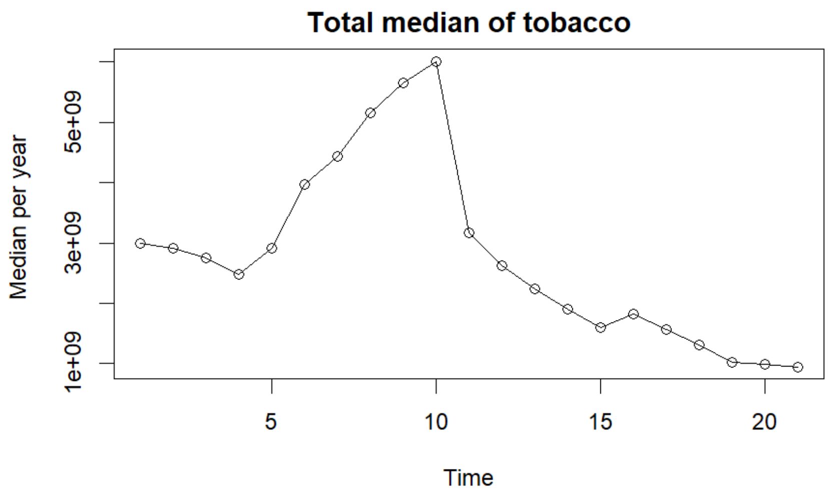
9 2008 4672736408 732498708 5657018408

10 2009 5279921723 314944000 6005708800

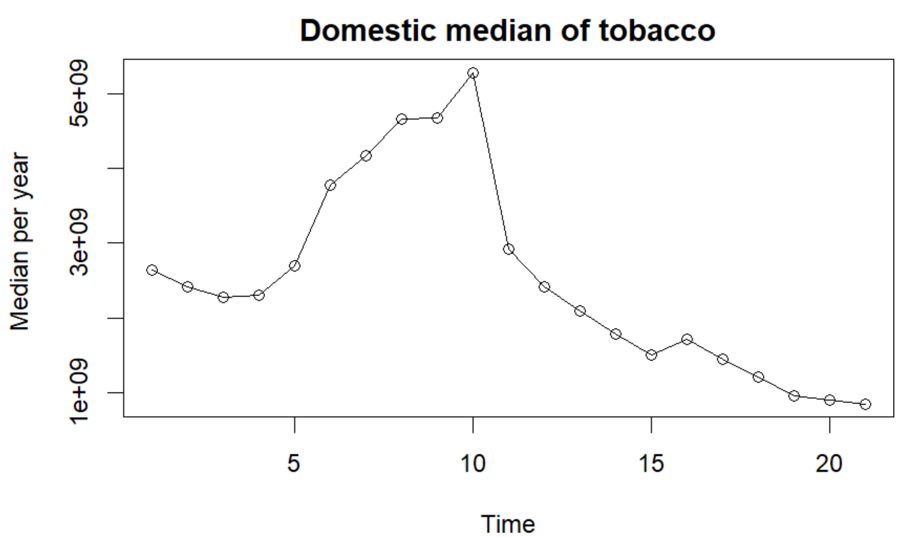
... with 11 more rows

##lets do some ploting

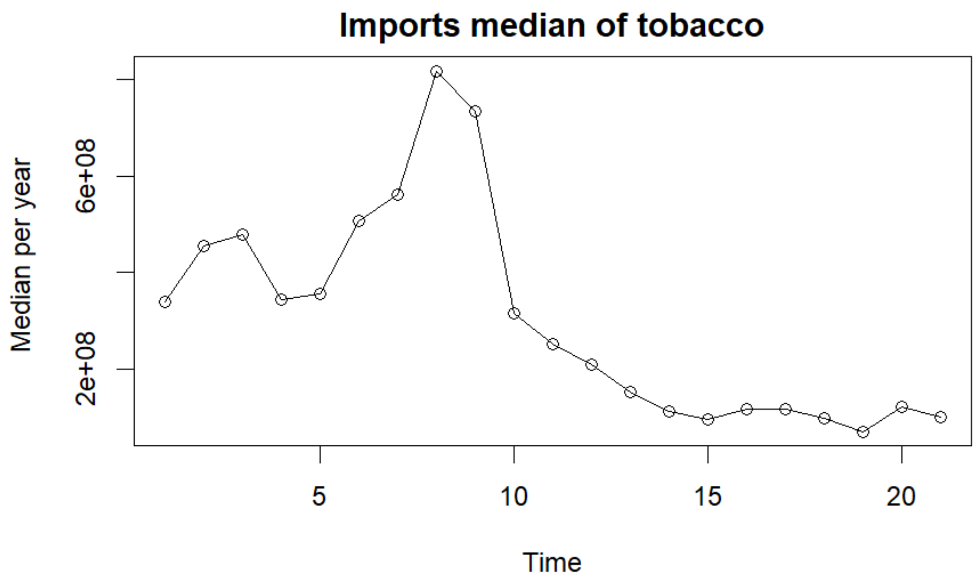
> plot(new_data$mead_Total, type = "o", ylab= "Median per year ", main = "Total median of tobacco", xlab = "Time")



> plot(new_data$mead_Dom, type = "o", ylab= "Median per year ", main = "Domestic median of tobacco", xlab = "Time")



> plot(new_data$mead_Imp, type = "o", ylab= "Median per year ", main = "Imports median of tobacco", xlab = "Time")



##With these plots we can infer two things. First, the outliers are clearly affecting the data as we expected to do, and secondly the three variables that are analyzed shows a similar patron.

we can appreciate that between 2008-2010 it extremely decreased imports & domestic and clearly is showed in the total variable. One reason it could be the 2008 crisis. The important insight from here is that after that it continues to decrease for the next 10 years, some change in the consumption habit of tobacco was affected.

Now lets do some time series analysis

nts <- ts(new_data)

> nts

Time Series:

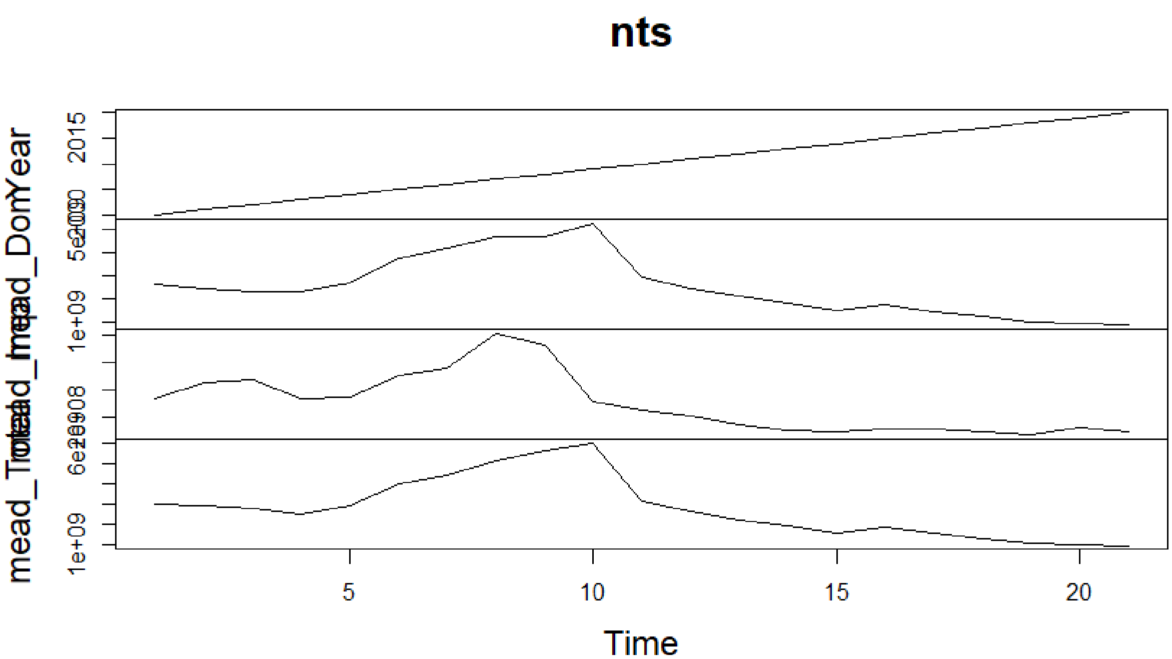
Start = 1

End = 21

Frequency = 1

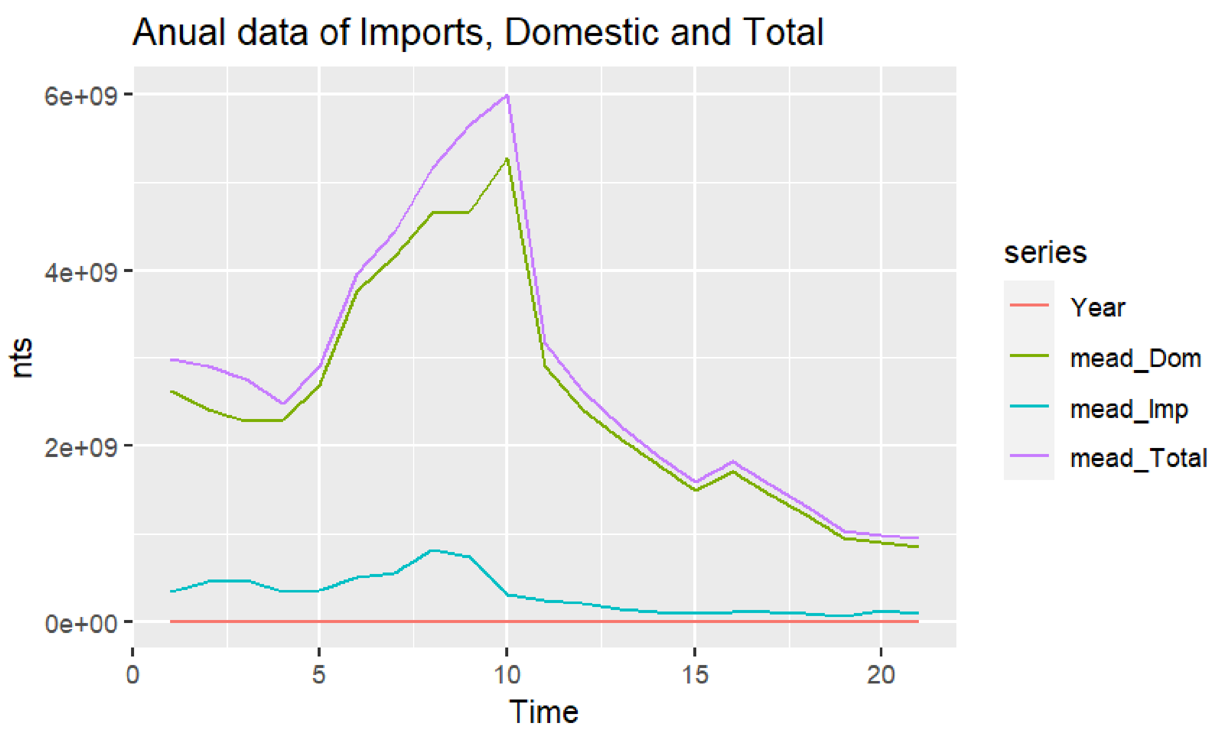
Year mead_Dom mead_Imp mead_Total

```
 1 2000 2635167015 338489600 2999419077
 2 2001 2421344492 454477785 2914666831
 3 2002 2278010585 479059692 2757070277
 4 2003 2301972488 343673108 2474341488
 5 2004 2701646262 354785477 2917089262
 6 2005 3772041108 507187692 3968494108
 7 2006 4162220407 561063385 4434095407
 8 2007 4658667130 816734031 5160825473
 9 2008 4672736408 732498708 5657018408
10 2009 5279921723 314944000 6005708800
11 2010 2915654892 252016246 3167671138
12 2011 2412067938 209522215 2621590154
13 2012 2087086277 152780308 2239866585
14 2013 1786230154 111391508 1897621662
15 2014 1497868308  95975385 1593843692
16 2015 1710605785 117874215 1828480000
17 2016 1446586092 117479385 1564065477
18 2017 1212400738  99499323 1311900062
19 2018  960193477  70862769 1031056246
20 2019  898257723 122838000  985198769
21 2020  849026462 101098000  948618339
# If we plot “nts”
> plot(nts)
```



According to the graph above, it exists lots of similarity between the three variables (Domestic, Import and the Total). These results relate to the covariances that we found earlier and allows to say that if one variable such as mead_total goes down is because the other two variable were affected first, at the end helps us to decide to work just with Total to predict what is going to happen to the demand for the next three years.

> autoplot(nts) + labs(title = "Anual data of Imports, Domestic and Total", ylab("Median of each year"))



lets observe some predictors with the original data and new data

#Original data

```
> train.tc <- head(tc$Total,219)
> test.tc <- tail(tc$Total,54)
> .8*length(tc$Total)
[1] 218.4
> ##Lets try SARIMA model
> sarima.tc <- auto.arima(train.tc)
> summary(sarima.tc)
Series: train.tc
ARIMA(0,0,0) with non-zero mean

Coefficients:
```

mean

57458566448

s.e. 667243532

sigma^2 = 1.642e+22: log likelihood = -5911.5

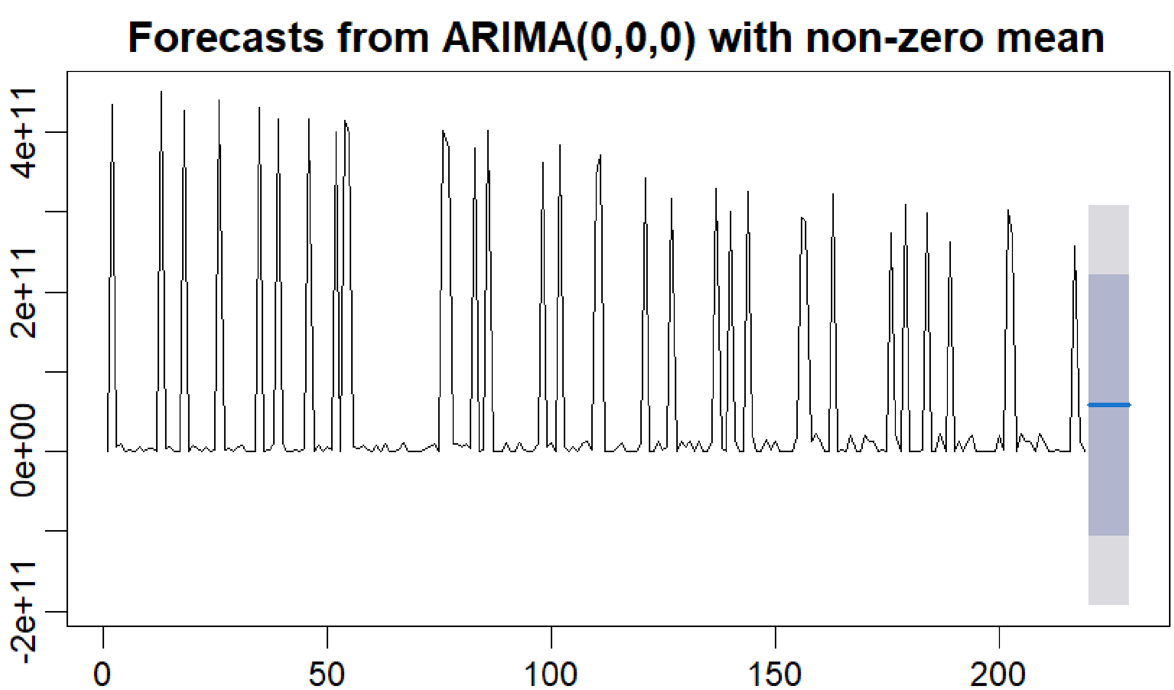
AIC=11827 AICc=11827.06 BIC=11833.78

Training set error measures:

	ME	RMSE	MAE	MPE	MAPE
Training set	-3.775441e-05	1.27862e+11	89964627672	-154360.1	154385.3

	MASE	ACF1
Training set	0.9386689	-0.01042226

> plot(forecast(sarima.tc))

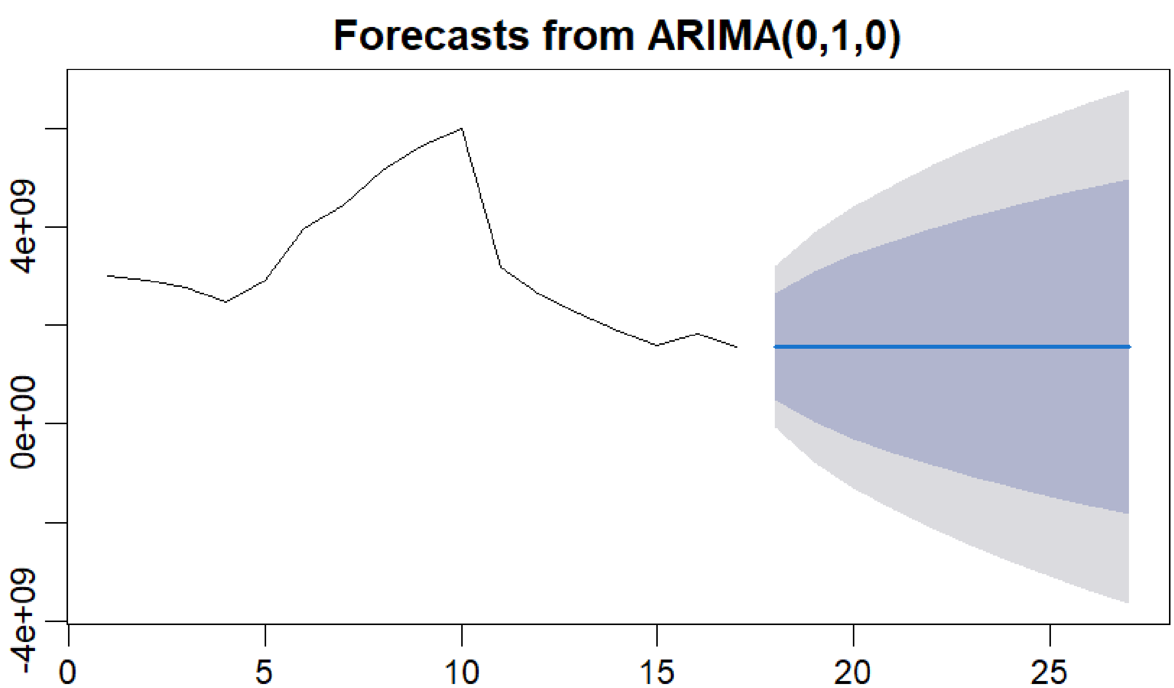


##New data ARIMA MODEL

train_tc_1 <- head(new_data$mead_Total,17)

> test_tc_1 <- tail(new_data$mead_Total,4)

```
> sarima_tc_1 <- auto.arima(train_tc_1)
> summary(sarima_tc_1)
Series: train_tc_1
ARIMA(0,1,0)

sigma^2 = 7.099e+17:  log likelihood = -351.53
AIC=705.07   AICc=705.35   BIC=705.84

Training set error measures:
                    ME      RMSE       MAE       MPE     MAPE      MASE
Training set -84256128 817417471 527668250 -6.338458 17.36943 0.9414913
                  ACF1
Training set 0.2245633
> plot(forecast(sarima_tc_1))
```




```
accuracy(forecast(sarima_tc_1,4),test_tc_1)
```

ME RMSE MAE MPE MAPE MASE

Training set -84256128 817417471 527668250 -6.338458 17.36943 0.9414913

Test set -494872123 515157277 494872123 -48.637862 48.63786 0.8829748

ACF1

Training set 0.2245633

Test set NA

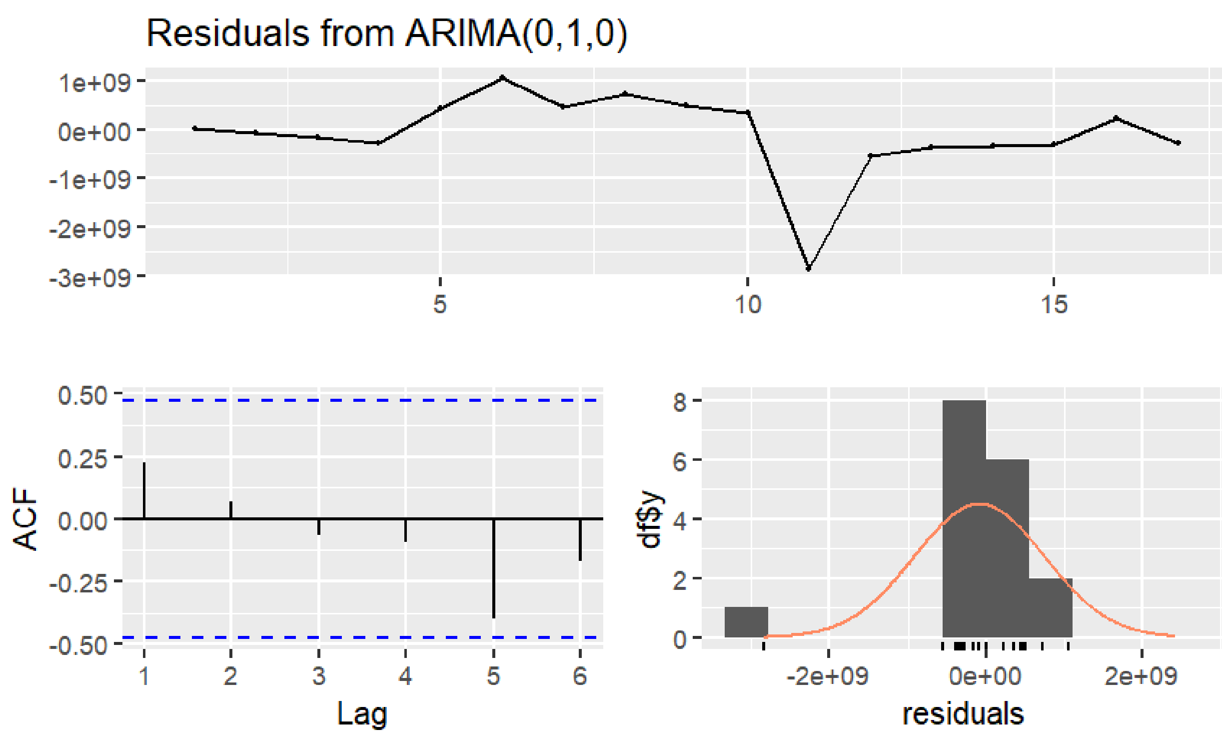
> Box.test(sarima_tc_1$residuals)

#As we can see in the forecasts that the ARIMA model provide us, the model does not adapt in a good way to our data, mainly because the model does not recognize the trend in our data.

Box-Pierce test

data: sarima_tc_1$residuals

X-squared = 0.85729, df = 1, p-value = 0.3545



> checkresiduals(sarima_tc_1)

Ljung-Box test

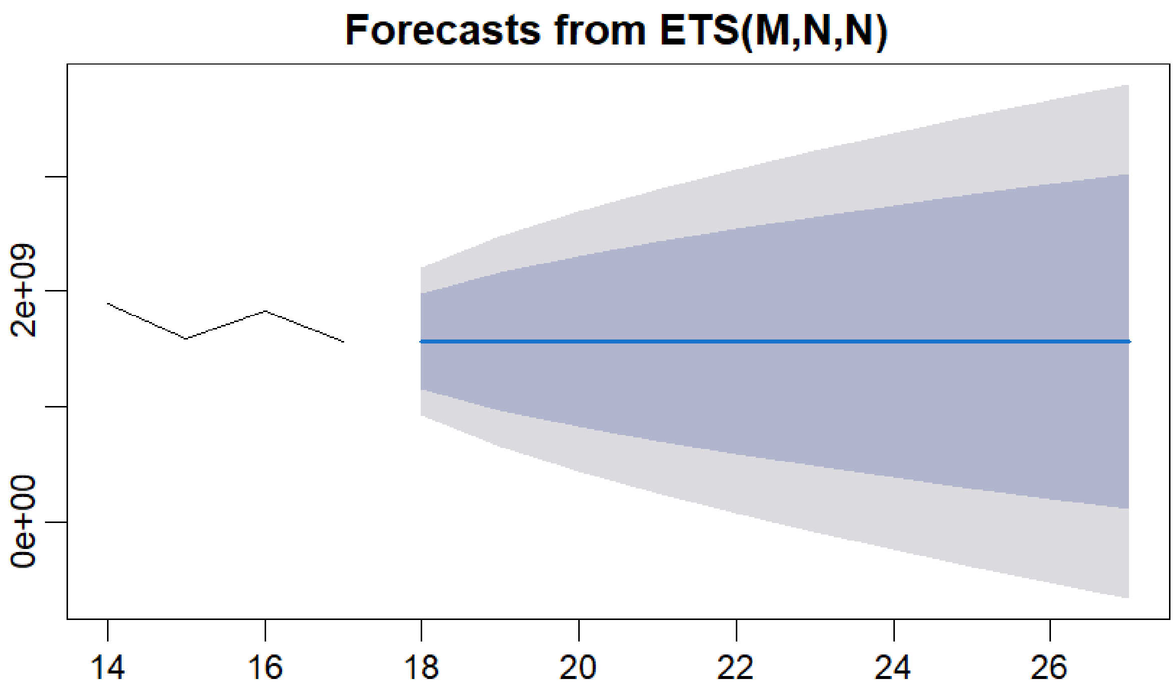
```
data:  Residuals from ARIMA(0,1,0)

Q* = 1.2233, df = 3, p-value = 0.7474

Model df: 0.   Total lags used: 3

##MODEL 2 NEW DATA

## lets try some ETS

>

> train_tc_ETS <- head(new_data$mead_Total,17)

> test_tc_ETS <- tail(new_data$mead_Total,4)

> ets.total <- ets(train_tc_ETS)

> summary(ets.total)

ETS(M,N,N)

Call:

 ets(y = train_tc_ETS)

  Smoothing parameters:

    alpha = 0.9999

  Initial states:

    l = 3928872911.6398

  sigma:  0.2084

     AIC     AICc      BIC

741.7011 743.5473 744.2008
```

Training set error measures:

	ME	RMSE	MAE	MPE	MAPE	MASE
Training set	-139118681	847950122	582208488	-8.168039	19.1877	1.038805

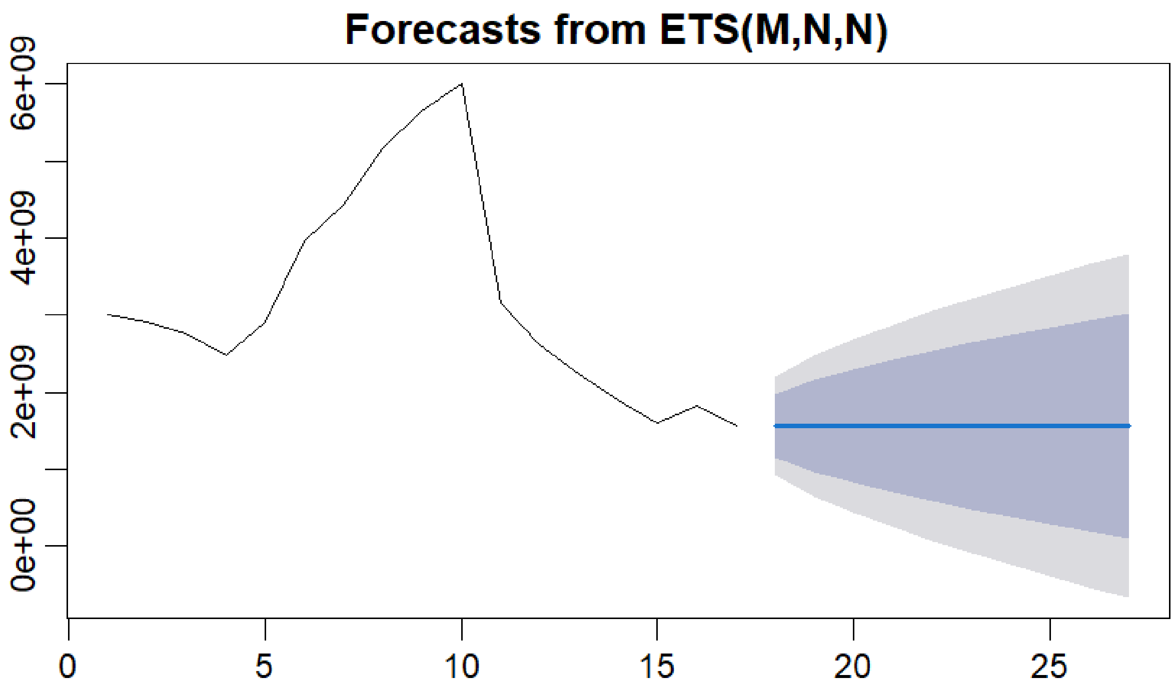
	ACF1
Training set	0.2124977

> plot(forecast(ets.total),4)



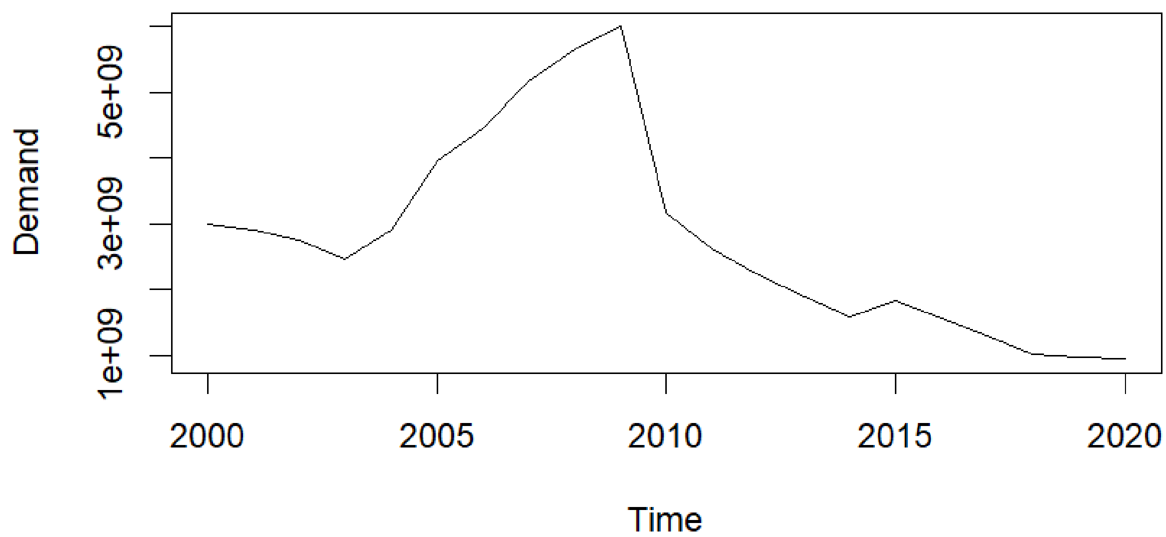
> plot(forecast(ets.total),29)

#In this case, the model also does not recognize the trend in the data, tha´s why it gives us forecasts that have a MAPE 19.18 which is still high.

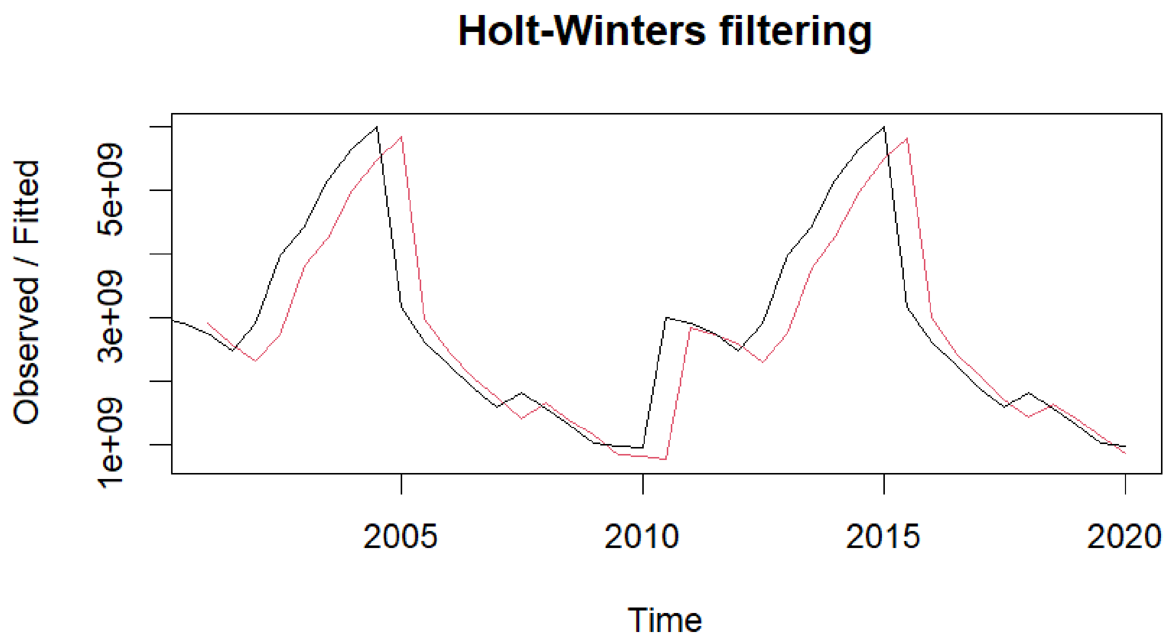



```
accuracy(forecast(ets.total,4), test_tc_ETS)
                    ME      RMSE       MAE       MPE     MAPE     MASE
Training set -139118681 847950122 582208488  -8.168039 19.18770 1.038805
Test set     -494898573 515182686 494898573 -48.640376 48.64038 0.883022
                  ACF1
Training set 0.2124977
Test set            NA
> Box.test(ets.total$residuals)

        Box-Pierce test

data:  ets.total$residuals
X-squared = 1.9764, df = 1, p-value = 0.1598
```

```
> Demand  <- ts(new_data$mead_Total,2000,2020, frequency = 2)
```

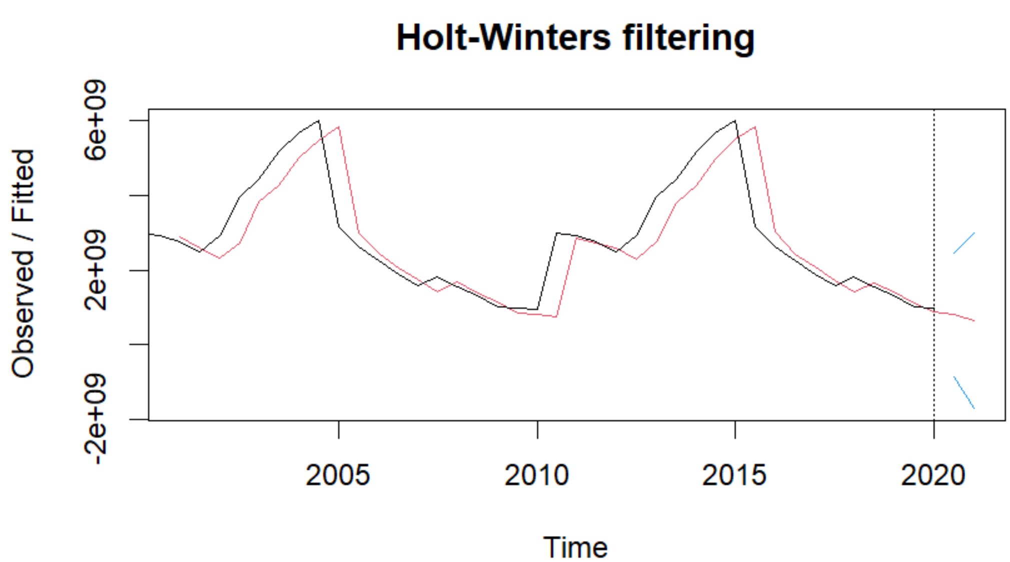



```
> hw <- HoltWinters(Demand, seasonal = "additive")
> plot(hw)
```



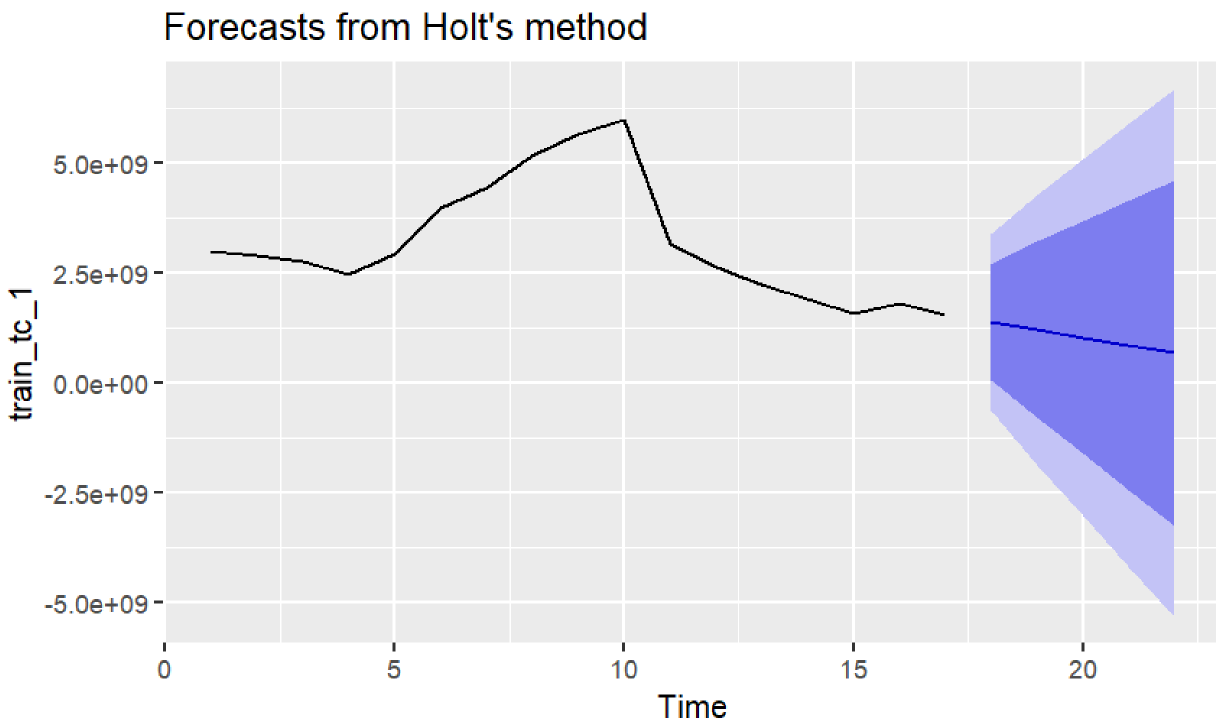
> ftc <- predict(hw,2,prediction.interval = TRUE)

> plot(hw,ftc)



##This plot and forecast were closed to be the correct but to create we had to gave a frequency of two, in other words, we said that we observe each data two times and that is not wrong so we decided to look for a more basic model to allows forecast in a proper manner

```
> Demand  <- ts(new_data$mead_Total,2000,2020, frequency = 1)

> train_tc_1

 [1] 2999419077 2914666831 2757070277 2474341488 2917089262 3968494108

 [7] 4434095407 5160825473 5657018408 6005708800 3167671138 2621590154

[13] 2239866585 1897621662 1593843692 1828480000 1564065477

> test_tc_1

[1] 1311900062 1031056246  985198769  948618339

> library(tidyverse)

-- Attaching packages ------------------------------------ tidyverse 1.3.1 --

v ggplot2 3.3.5     v purrr   0.3.4

v tibble  3.1.6     v dplyr   1.0.7

v tidyr   1.1.4     v stringr 1.4.0

v readr   2.1.1     v forcats 0.5.1

-- Conflicts --------------------------------------- tidyverse_conflicts() --

x dplyr::filter() masks stats::filter()

x dplyr::lag()    masks stats::lag()

> library(fpp2)

Registered S3 method overwritten by 'quantmod':

  method            from

  as.zoo.data.frame zoo

-- Attaching packages --------------------------------------------- fpp2 2.4 --

v forecast  8.16     v expsmooth 2.3

v fma       2.4


> holt_1 <- holt(train_tc_1, h = 5)

> autoplot(holt_1)
```



Like we can appreciate in this plot it seems that trend goes accordingly as we expected and in congruence with the tendency of the last three years

```
> accuracy(holt_1,test_tc_1)
                    ME      RMSE       MAE       MPE      MAPE      MASE
Training set -208294583 892997913 515854471 -7.794960 17.762411 0.9204126
Test set      -55900255 110901787  99357355 -4.899792  9.480887 0.1772782
                  ACF1
Training set 0.1123374
Test set            NA
> holt_1$model
Holt's method

Call:
 holt(y = train_tc_1, h = 5)
```

Smoothing parameters:

alpha = 0.9999

beta = 0.1612

Initial states:

l = 1931001462.1407

b = 395166734.0915

sigma: 1021182182

```
     AIC     AICc      BIC
758.9079 764.3624 763.0739
> accuracy(holt_1,test_tc_1)
                    ME      RMSE       MAE       MPE      MAPE      MASE
Training set -208294583 892997913 515854471 -7.794960 17.762411 0.9204126
Test set      -55900255 110901787  99357355 -4.899792  9.480887 0.1772782
                  ACF1
Training set 0.1123374
Test set            NA
```

#At the end we found that the Holt´s two parameter method is the best model to our data. Mainly because it considers trend in its parameters, which we think is the key feature in our data. The model accuracy is pretty good when we consider the lack of data, we used to train the model.

We had to use the median of tobacco sales because we had a dataset with a lot of outliers that were affecting the tobacco mean. Also, we could not use the data by period of the year because we didn´t have consistent data of all the periods of each year.

Our mean absolute percentage error in the test set is 9.48. With that in mind we can infer based on the data provided and the Holt´s two parameter-method that the sales of tobacco will continue decreasing probably because the constant and increasing popularity of electronic cigarettes in the US.

According to the CDC foundation, which is a non-profit organization that is part of the Centers for Disease Control and Prevention's critical health, said that from February 2020 to March 2021, e-cigarette sales increase 50%.

Our recommendation is that any business that sells cigarettes should start planning a change to tobacco-free cigarettes like e-cigarettes in order to increase their sales and avoid losing money.

Also some action plans that they can take care of meanwhile the renew the business model, cold be to decrease the amount of production of the sub products that the market are not consuming like: “Large cigars”, “” Chewing Tobacco” , “ Cigarette removals”, “Small cigars”, and “Snuff” .

That the production plan of the next three years should be decreasing each year compared to the actual production plan but don´t forget to study the market frequently. We are on a stochastic ecosystem so we should expect the unexpected anytime.

Other action that they can consider to do is to focus on what the clients most value about their product like: “Cigarette equivalents” & “Pounds”, so is the company is not giving this kind of service they should consider this way to do it.

One possibility that we can think about is to find other markets like Latin America where the tobacco consumption habits could be different and could fit with the actual business model of the company.

At the end, we still believe that adapt to new markets should be the must profitable answer.