UNIVERSITÀ DEGLI STUDI DI TRIESTE

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Bachelor's Degree - Industrial Engineering
A Case of Study - ITC Chain Italia S.R.L.

Data analysis for Anticipation of Future Demand

Submitted by:
Jose Agustin SPACCESI

Supervised by:
Prof. Marino NICOLICH

2017-2018

Acknowledgment

Without the support of my family during all these years, this would not be possible. I am thankful mom, dad and sister for being my main source of motivation, for impulse me every time my speed decreases and for being, even now in the distance, my company through this research.

Thank you to all my friends, who also directly or indirectly have supported me along the path and have helped me with this project.

And last, but not least, thank you to all the professors who built my knowledge until now, who aroused my sense of curiosity in these topics and who have helped me with its development. Specially to my tutor, Marino Nicolich, for his time and dedication.

Abstract

The goal of this thesis is the optimization of the production, according to the analysis of past and present data of sales collected throughout the company around the past 2 years.
This has been done to minimize unsealed products maximizing the use of materials and space in the warehouse, minimize deliver times and maximize profits.
In the initial phase of this project, we have encountered a company in full reorganization phase, looking in particular to industry 4.0 improvements. In order to accomplish that, use existing data to obtain information is very interesting for KITO and can make a big difference in its productive process and efficiency of the plant.

Keywords: Optimization; Forecast; Demand; Data; Analysis; Industry 4.0

Resumen

El objetivo de este trabajo final de grado (TFG) es la optimización de la producción según el análisis de datos de ventas pasados y presentes recolectados por la compañía en los últimos dos años.
Esto se ha hecho para minimizar productos no vendidos, maximizando el uso de los materiales y el espacio en el almacén, minimizando tiempos de entrega y maximizando los beneficios.
En la fase inicial del proyecto, nos encontramos una compañía en completo proceso de reorganización, buscando en particular mejoras relacionadas con la Industria 4.0. Por este motivo, el uso de datos para obtener información es muy interesante para KITO y puede hacer una gran diferencia en sus procesos productivos y en la eficiencia de la planta.

Palabras clave: Optimización; Forecast; Demanda; Datos; Análisis; Industria 4.0

Table of Contents

List of Figures

List of Tables

List of Equations

Introduction

KITO Chain Italia and Weissenfels

KITO CORPORATION is a global company born in Omori, Tokyo in 1932, leader in manufacturing of manual and electric chain hoists, with 22 different subsidiaries (including KITO Chain Italia S.R.L.) all around the world and 2,364 employees (March 31, 2017). In February 2016, KITO acquis Weissenfels Tech Chains S.R.L a metallurgical company located in Fusine (UD), Italy born in 1462. The main activity of the firm was the production of chains, accessories and master links made of iron and steel.

KITO
CHAIN ITALIA S.R.L.

Fig. 1: KITO CHAIN ITALIA S.R.L. logo

KITO's plan is continue the historical activity of Weissenfels, in order to make it one of the most important companies in its sector in Europe. [1][2][3]

Industry 4.0 and Data Science

Industry 4.0 refers to the "fourth industrial revolution" but, in contrast to the past three revolutions, there is not one single technology identified such us the key of the change. In despite of, the term Industry 4.0 describes a set of technological changes and organization processes based on innovation, communication and adaptability. [4]
One of the technology that Industry 4.0 refer is Data Science and Big data. First, it is necessary to clarify that Data science is the study field which includes several techniques from, statistics, data analysis, machine learning and their associated methods, to extract information and knowledge from data and Big data is a part of data science which treat huge (Petabyte 10^{15} byte) and complex (diverse types of data and formats, such as, audio, video, photography, etc.) data sets. [5]

As it can see in the following graphic,

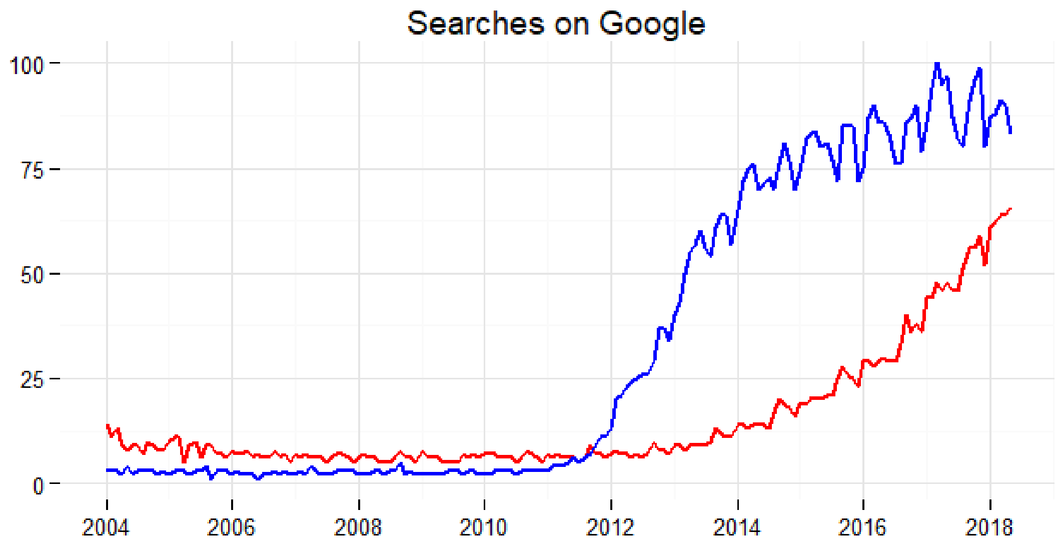


Fig. 2: Popularity of "Big data" and "Data science" on Google [6]

the interest in data is not new, but data science's popularity has been increasing in the last years. One reasons of that is that companies are increasing the use of sensors to capture data at all stages of a product's life, increasing sources of data and opportunities.

But the most of engineers does not have data science knowledge causing that most companies do not know how to capture, how to store or how to interpret them to improve their processes and products. Yet correct use of data can make industry more efficient, profitable and sustainable and many companies, such as KITO Chain Italia, are now trying to implement Industry 4.0 and Data Analysis on their productive processes. [7]

In the frame previously described and taking into account that disinformation is very expensive for a company, demand forecasting can help KITO to optimize production planning, take decisions or predict future capacity requirements.

Technics

Used programs

In this thesis it was used R and Excel to analyses the data. R was chosen because is a free software environment deeply extended between statistics and data analysts with many useful packages and tools to process, analyze and visualize data in a professional and complete way. Despite of it power, R is not simple enough for people without data science knowledge, that is why, the caption of data on the company was do it on Microsoft Excel as well as the presentation of result. Excel was chosen because is an extended and understandable program, easy to find in many computers and known by many people. [8][9]

Fig. 3: R logo. Fig. 4: Excel logo

On the other hand, the R code will be write on RStudio, an IDE (Integrated Development Environment) for R with many useful tools for plotting, history, debugging and workspace management. [10]

Fig. 5: RStudio logo

Forecasting

Forecasting is the process of predict or estimate a future event or trend based on present and past data. Certainly, as a prediction process, uncertainly is an important part on forecasting and most of the times is almost mandatory indicate the degree of uncertainty, as well as the accuracy attached to the analysis. Some examples of applications of forecasting are calculation of future weather conditions or financial trends. R has multiples tools and methods for forecast Time Series. [11]

Forecasting Techniques

In this part, we focus on forecasting techniques we used the case of study. That is: The Seasonal Naïve Approach and Holt Winters.

The Seasonal Naïve is an adaptation of Naive method for seasonal time series. In this method each prediction will be equal to the last observed value of the last season. For example, if you want to forecast the sales volume for next October, you would use the sales volume from the previous October. The following equation, where *"m"* is the seasonal period and *"k"* is the smallest integer greater than *"(h–1)/m"*, summarizes the method.

$$\hat{y}_{T+h|T} = y_{T+h-km} \qquad \textit{Eq. 1}$$

On the other hand, Holt winters or Triple Exponential Smoothing is an exponential smooth that include trend and seasonality. There are two different versions of this method, one for additive seasonality and one for multiplicative seasonality. Additive seasonality is when data experiment a constant variation on the same period every year while in Multiplicative seasonality the variation depend on previous values. For example, in Additive seasonality, sales could experiment an increase of 1000 pieces on April while on Multiplicative seasonality the increase is of 40%.

The component form for the additive method is:

$$\hat{y}_{t+h|T} = l_t + hb_t + s_{t+h-m(k+1)} \qquad \textit{Eq. 2}$$

$$l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \qquad \textit{Eq. 3}$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1} \qquad \textit{Eq. 4}$$

$$s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \qquad \textit{Eq. 5}$$

where *"m"* is the frequency of the seasonality and *"k"* is the integer part of *"(h−1)/m"*.

Furthermore, the smoothing parameter should be between $0 \leq \alpha \leq 1$, the trend parameter between $0 \leq \beta^* \leq 1$ and the seasonal parameter between $0 \leq \gamma \leq 1 - \alpha$.

The component form for the multiplicative method is:

$$\hat{y}_{t+h|T} = (l_t + hb_t)s_{t+h-m(k+1)} \quad \textit{Eq. 6}$$

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1}) \quad \textit{Eq. 7}$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1} \quad \textit{Eq. 8}$$

$$s_t = \gamma \frac{y_t}{l_{t-1}+b_{t-1}} + (1-\gamma)s_{t-m} \quad \textit{Eq. 9}$$

The parameters have the same definition as in the additive method. [14]

On R, Seasonal Naïve method is implemented with the *snaive()* function and Holt Winters with the *hw()* function with the `seasonal` argument equals to "additive" or "multiplicative". Both functions are from the **forecast** package. [forecast]

Current situation

KITO's midterm plan

According to KITO Mid-Term plan published on May 18, 2016, the goals for the next 5 years are return to a high margin business structure, growth through product portfolio expansion, evolve into a globally integrated enterprise and double the EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization) in order to become the most trusted anti-gravity equipment manufacturer in the global market.

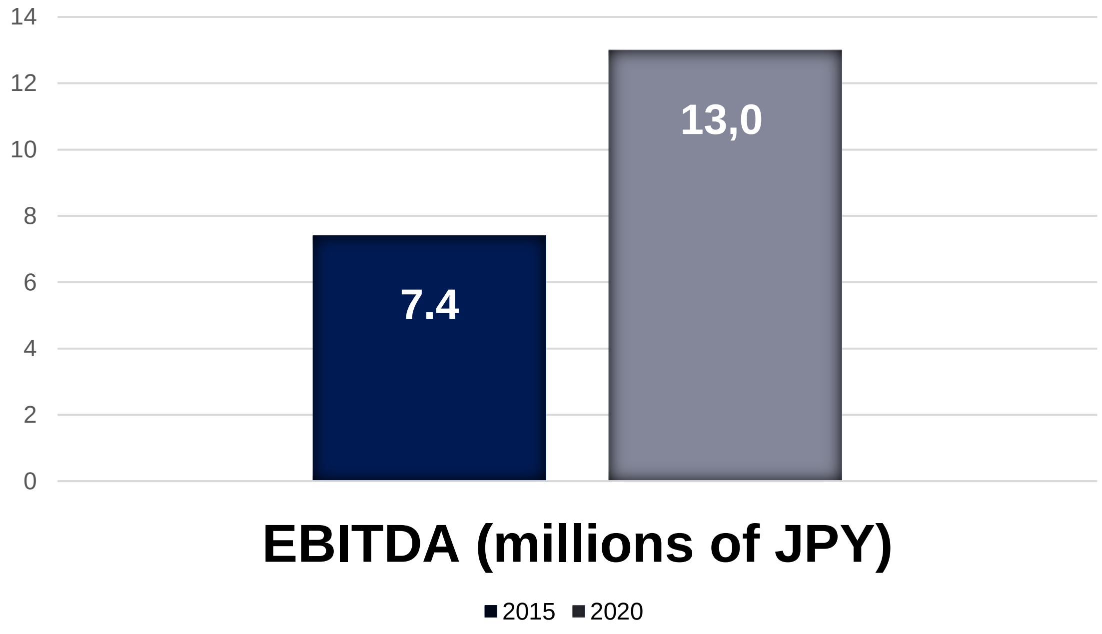


Fig. 6: Financial target of KITO CORP.

In order to accomplish their task and make KITO Chain Italia an important pillar for the group's European operations, KITO plan to raise the actual production of 287.680 pieces a year, to 1.000.000, four times the actual production, building a global production and supply system for chain-related items anticipating significant synergistic effects using the group's extensive sales network.

At the moment, KITO is embolden in a process of modernization, reorganizing the production, buying new machines and trying to approach Industry 4.0 goals. [1]

Actual distribution of demand

At the moment in KITO Chain Italia, the production is based on historical sales data and it is difficult to respond rapidly to demand variations in terms of stocks and level of service.

Critical issues and considerations

Independently of the machine where the part was mechanically transform, all parts must pass for a heat treatment process to reach the mechanical characteristic specification of the product. The heat treatment on KITO consists of two steps: quenching and tempering. During quenching metal is heated upper the critical temperature (around 724°C for steel) and cooling at a rapid rate producing martensite transformation and making ferrous alloys harder. Martensitic steel, while very hard, is too breakable to be useful for industrial applications, the second step, tempering, mitigate the problem heating steel below the lower critical temperature and cooling to provide some toughness by alleviating the internal tensions of the metal. [12]

KITO's furnaces are loaded with around 450 kg part batches. In order to plan the production, we have to take special attention to this point because depending on the dimensions and steel grade of the piece the time to heat and cold the pieces is different, so probably is difficult to put different parts on the furnace at the same time.

On the other hand, the space necessary to store finished parts can be considerate as unlimited.

Pieces can be different surface finish, but we will focus on the shape and material of the pieces that means that every piece with the same `CODICE_CATALOGO` essentially is the same pieces even if has a different `ARTICLE` because the second variable describe also other attributes that are not important to us.

The company does not have older data than two years ago because KITO's group bought the company two years ago. The analysis should to be more solid when we increment the quantity on data.

Data Analysis

First of all, to get a better understanding of the system it is important to know how the information is saved in the data set and what relationship there are between different entities on the reality we are analyzing. In order to get a better understanding of the situation is really useful to do an Entity-Relationship model.

The Entity-Relationship model (ER model) is a conceptual model of data that describe with high detail level how data will be logically and physically represented. The model components are Entity and Relationship. An Entity represent a concept of the real world with independent existence. In our model, the entities are the PRODUCTS and the CLIENTS. Relationship excites between Entities. In our model, clients BUY products. Both components can have attributes. Every Product has DIMENSIONS, DESCRIPTIONS, etc and every Client is located in a NAZIONE and has CLIENTE number. Every Entity has a key. A key is one or more obligatory attributes (the minimum necessary) that identify unequivocally the entity. The key of Products is ARTICLE and the key of Client is CLIENTE.

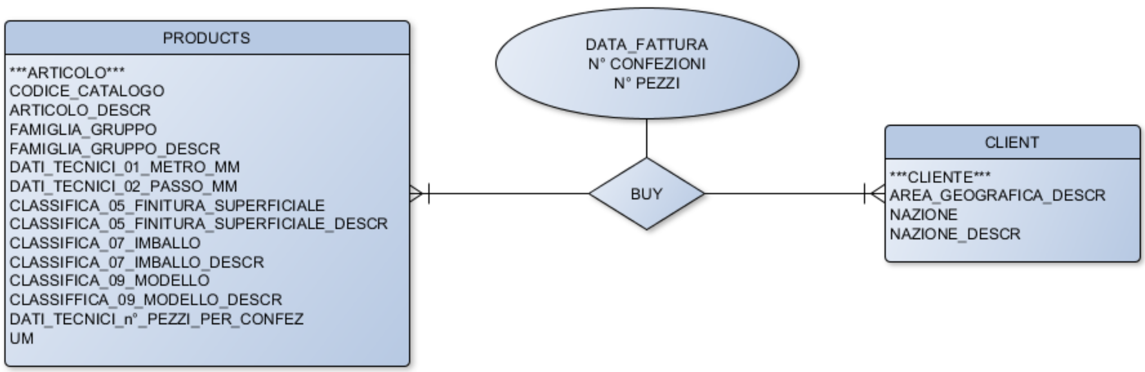


Fig. 7: ER model of data.

PRODUCT (<u>ARTICOLO</u>, CODICE_CATALOGO, ARTICOLO_DESCR, FAMIGLIA_GRUPPO, FAMIGLIA_GRUPPO_DESCR, ...)

CLIENT (<u>CLIENTE</u>, AREA_GEOGRAFICA_DESCR, NAZIONE, NAZIONE_DESCR)

BUY (<u>ARTICOLO</u>, <u>CLIENTE</u>, <u>DATA_FATTURA</u>, Nº CONFEZIONI, <u>Nº PEZZI</u>)

The data analysis will follow the next diagram,

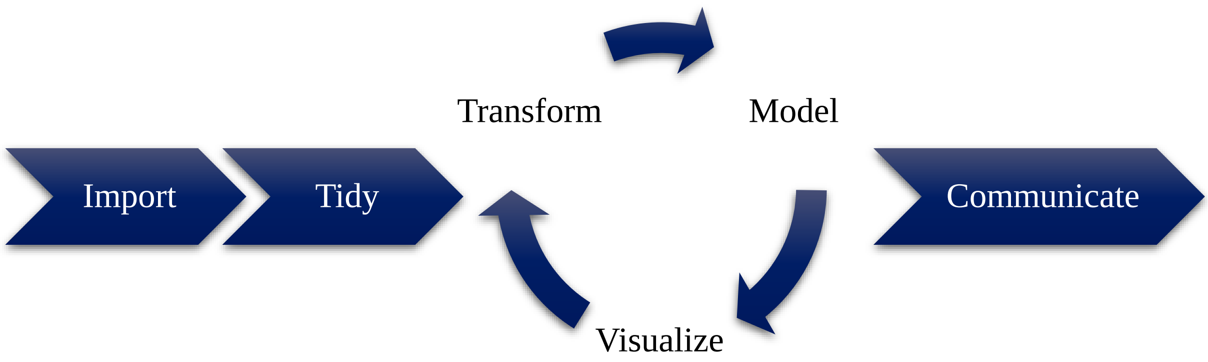


Fig. 8: Data analysis flow diagram.

Importing and Exploring

The data set used was proportionated by KITO and keep the data of sales of the last 2 years. The file has a size of around 765Mb and it was elaborated by KITO on Excel (.xlsx format).

Fig. 9: Excel file with the sales of the last 2 years.

To import the data sets on R, *readXL()* function was used because is good to import .xls or .xlsx files (Excel files) and unlikely than *read_excel()* or other most used functions, has the option of set all strings as factors and our data set is full of them. This function is from **RcmdrMisc** package. This package is not preinstalled in R or RStudio so it is necessary to download, install and load it before use it. [RcmdrMisc]

```
#Dowloading and installing the RcmdrMisc package
>     install.package(“RcmdrMisc”)

#Loading the RcmdrMisc package
>     library(RcmdrMisc)
```

Calling *readXL()* with `stringsAsFactors` argument sets as TRUE, in order to transform every character column into a factor column. Factors are variables in R which take on a limited number of different values.

```
#Importing Excel fil into as sales_original
>       sales_original       <-       readXL("VenditeKITO_2016_2018.xlsx",
stringsAsFactors = TRUE)
```

To get a first visualization of the data set, it was used a function from the **dplyr** package, *glimpse()* that print a compact summary of the internal structure of an R object, in this case a data frame. It also belongs to the **Tidyverse** core. It is also necessary to install this package. The syntax is the same one we use to install the last package. [dplyr]

```
> glimpse(sales_original)

Observations: 6,298
Variables: 24
$ AREA_GEOGRAFICA_DESCR                      <fct> Europa (in EU), Europa (out EU),
Europa...
$ NAZIONE                                 <fct> 18, 03, 18, 18, 43, 43, 43, 43, 43,
43,...
$ NAZIONE_DESCR                           <fct> HOLLAND, SLOVENIA, HOLLAND, HOLLAND,
JA...
$ CLIENTE                                       <fct> 739200, 574000, 886601, 886601,
514501,...
$ TIPO_CLIENTE..user..U...retailer..R.        <fct> R, R, R, R, R, R, R, R, R, R, R, R,
R, ...
$ ARTICOLO                        <fct> W3ANA30220A01, W5CGRA0100A03, W3ANF1022...
$ CODICE_CATALOGO                            <fct> D22E16, GSC10, F22, F22M16, SHC10,
SHC1...
$ ARTICOLO_DESCR                          <fct> COMPLESSIVO D22E16 GRIGIO GRADO 10,
GAN...
$ FAMIGLIA_GRUPPO                         <fct> B.12.06, B.12.05, B.12.09, B.12.09,
B.1...
$ FAMIGLIA_GRUPPO_DESCR                      <fct> ANELLONI <=26MM GR100, ACCESSORI
GR100,...
$ DATA_FATTURA                    <dbl> 20160519, 20160606, 20160616, 20160616,...
$ CLASSIFICA_05_FINITURA_SUPERFICIALE        <fct> 25, 25, 47, 47, 25, 25, 25, 25, 25,
25,...
$   CLASSIFICA_05_FINITURA_SUPERFICIALE_DESCR   <fct>   V.epox   grigioRAL9007,   V.epox
grigioRAL9...
```

```
$ CLASSIFICA_07_IMBALLO                          <fct> NA, 02, NA, NA, 06, 06, 06, 06, 06,
06,...
$ CLASSIFICA_07_IMBALLO_DESCR                    <fct> ** NON TROVATO **, Scatola, ** NON
TROV...
$ CLASSIFICA_09_MODELLO                          <fct> AA, CC, AA, AA, CE, CE, CE, CE, CE,
EC,...
$ CLASSIFICA_09_MODELLO_DESCR              <fct> Anellone, Gancio Clevis Grab, Anellone,...
$ DATI_TECNICI_01_DIAMETRO_MM                 <dbl> 23, 10, 23, 23, 10, 13, 16, 6, 8, 13,
2...
$ DATI_TECNICI_02_PASSO_MM                      <dbl> 160, 0, 270, 270, 0, 0, 0, 0, 0, 0,
0, ...
$ DATI_TECNICI_10_PEZZI_PER_CONFEZIONE        <dbl> 1, 10, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1,...
$ UM                                             <fct> N, N, N, N, N, N, N, N, N, N, N, N,
N, ...
$ N..CONFEZIONI                                 <dbl> 50, 1, 2, 2, 10, 10, 10, 10, 10, 10,
10...
$ PESO.TOT..Kg                                     <dbl> 156.00, 11.70, 5.12, 8.00, 11.20,
20.00...
$ N..PEZZI                                      <dbl> 50, 10, 2, 2, 10, 10, 10, 10, 10, 10,
1...
```

As can be seen, the data set has 6.298 rows, every row corresponds to an observation and every observation to a sale of the company. The type of data of each column it also appears on the internal structure summary. It is easy to see than the column `DATA_FATTURA` is double but it should be a date.

There are some mistakes we will have to take care in the structure, as `CLASSIFICA_07_IMBALLO_DESCR` where ** NON TROVATO ** value should be NA.

Moreover, the data set is composed by 24 columns, every column corresponds to a variable:

- `AREA_GEOGRAFICA_DESCR` is a **factor** that describe the geographical area where the `CLIENT` is located.
- `NAZIONE` is a **factor** describe, using a numeric code, the country where the `CLIENT` is located.
- `NAZIONE_DESCR` is a **factor** with the name of the country where the `CLIENT` is located.
- `CLIENTE` is a **factor** that, using a numeric code, shows unequivocally the client who buy the article.

- `TIPO_CLIENTE..user..U...retailer..R.` is a **factor** that describe if the `CLIENT` is a retailer (R) or an user (U). In this data set, all clients are retailers (R).
- `ARTICOLO` is a **factor** that, using a code, identify unequivocally the article sold to the client. This variable, unlikely than `CODICE_CATALOGO` define the color and package of the order.
- `CODICE_CATALOGO` is a **factor** that describe the catalog code of the piece.
- `ARTICOLO_DESCR` is a **factor** that describe the article sold to the client.
- `FAMIGLIA_GRUPPO` is a **factor** that group products by families using a brief description.
- `FAMIGLIA_GRUPPO_DESCR` is a **factor** that group products by families using a code.
- `DATA_FATTURA` is a double but it should be a **date**. The date when the sale was made is composed by eight characters without any separation. The first four are the year, the following two, the month and the last two, the day ("YYYYMMDD"). We are not sure if this date is not the date when the facture was made but we are going to work with it because is what the company provided us.
- `CLASSIFICA_05_FINITURA_SUPERFICIALE` is a **factor** that, using a numeric code, classify the surface finish of the article ordered by the client.
- `CLASSIFICA_05_FINITURA_SUPERFICIALE_DESCR` is a **factor** that describe the surface finish of the article ordered by the client.
- `CLASSIFICA_07_IMBALLO` is a **factor**, using a numeric code, classify the packing used to send the articles to the client.
- `CLASSIFICA_07_IMBALLO_DESCR` is a **factor** that describe the packing used to send the articles to the client.
- `CLASSIFICA_09_MODELLO` is a factor that classify pieces by models using two letters.
- `CLASSIFICA_09_MODELLO_DESCR` is a factor that classify pieces by models using a brief description.
- `DATI_TECNICI_01_DIAMETRO_MM` is a double with a characteristic measure of the piece.
- `DATI_TECNICI_02_PASSO_MM` is a double with a characteristic measure of the piece.
- `DATI_TECNICI_10_PEZZI_PER_CONFEZIONE` some of the pieces are sold in packs. This variable is a **double** with the number of pieces that are in the sold pack.
- `UM` we are not sure what this variable means but, in this data, set this column has always an "N" value.

- `N..CONFEZIONI` some of the products are sold in packs. This variable is a **double** that identify how many pieces are in the pack.
- `PESO` is a **double** that contain the weight of the package of pieces sent to the client.
- `N..PEZZI` is a **double** that contain the number of pieces of the `ARTICOLO` bought by the `CLIENTE` in `DATA_FATTURA`.

It is useful see how the data set looks like but print the data set on the console is not a good idea because we are talking about thousands of observations. To do a first visualization it was used function from **utils** package, *head()*, which as default shows the first six rows of the data set (with n arguments we can change the number of rows showed). It is not necessary install anything to use utils packages in R. [utils]

```
> head(sales_original)

  AREA_GEOGRAFICA_DESCR NAZIONE NAZIONE_DESCR CLIENTE
1         Europa (in EU)      18       HOLLAND  739200
2        Europa (out EU)      03      SLOVENIA  574000
3         Europa (in EU)      18       HOLLAND  886601
4         Europa (in EU)      18       HOLLAND  886601
5                   Asia      43         JAPAN  514501
6                   Asia      43         JAPAN  514501
  TIPO_CLIENTE..user..U...retailer..R.       ARTICOLO CODICE_CATALOGO
1                                    R W3ANA30220A01          D22E16
2                                    R W5CGRA0100A03           GSC10
3                                    R W3ANF10220001             F22
4                                    R W3ANF30220001          F22M16
5                                    R W4CSGS0100A03           SHC10
6                                    R W4CSGS0130A03           SHC13
                                                        ARTICOLO_DESCR
1                                    COMPLESSIVO D22E16 GRIGIO GRADO 10
2 GANCIO ACCORCIATORE GSC10 GRIGIO 10 PZ  10MM GRADO 10 TIPO A FORCELLA
3                       ANELLONE OVALE OFFSHORE F22 PORPORA     GRADO 8
4                       COMPLESSIVO OFFSHORE F22M16 PORPORA     GRADO 8
5      GANCIO SICUREZZA SHC10 GRIGIO 10MM      GRADO 10 TIPO A FORCELLA
6      GANCIO SICUREZZA SHC13 GRIGIO 13MM      GRADO 10 TIPO A FORCELLA
  FAMIGLIA_GRUPPO FAMIGLIA_GRUPPO_DESCR DATA_FATTURA
1         B.12.06 ANELLONI <=26MM GR100     20160519
2         B.12.05        ACCESSORI GR100     20160606
3         B.12.09      ANELLONI OFFSHORE     20160616
4         B.12.09      ANELLONI OFFSHORE     20160616
5         B.12.05        ACCESSORI GR100     20160623
6         B.12.05        ACCESSORI GR100     20160623
  CLASSIFICA_05_FINITURA_SUPERFICIALE
1                                  25
2                                  25
3                                  47
4                                  47
```

```
5                                  25
6                                  25
  CLASSIFICA_05_FINITURA_SUPERFICIALE_DESCR CLASSIFICA_07_IMBALLO
1                      V.epox grigioRAL9007                  <NA>
2                      V.epox grigioRAL9007                    02
3                      V.epox porporRAL3004                  <NA>
4                      V.epox porporRAL3004                  <NA>
5                      V.epox grigioRAL9007                    06
6                      V.epox grigioRAL9007                    06
  CLASSIFICA_07_IMBALLO_DESCR CLASSIFICA_09_MODELLO
1          ** NON TROVATO **                    AA
2                     Scatola                    CC
3          ** NON TROVATO **                    AA
4          ** NON TROVATO **                    AA
5             Sacchetto nylon                    CE
6             Sacchetto nylon                    CE
  CLASSIFICA_09_MODELLO_DESCR DATI_TECNICI_01_DIAMETRO_MM
1                    Anellone                          23
2           Gancio Clevis Grab                          10
3                    Anellone                          23
4                    Anellone                          23
5          Gancio Clevis Sling                          10
6          Gancio Clevis Sling                          13
  DATI_TECNICI_02_PASSO_MM DATI_TECNICI_10_PEZZI_PER_CONFEZIONE UM
1                      160                                    1  N
2                        0                                   10  N
3                      270                                    1  N
4                      270                                    1  N
5                        0                                    1  N
6                        0                                    1  N
  N..CONFEZIONI PESO.TOT..Kg N..PEZZI
1            50       156.00       50
2             1        11.70       10
3             2         5.12        2
4             2         8.00        2
5            10        11.20       10
6            10        20.00       10
```

Cleaning and Tidying

First of all, using *summary()* function, we display a summary of every variable, in order to identify possible fails in the data set. *summary()* is a **base** function used to produce summaries depending on the class of the argument. [base]

```
> summary(sales_original)
     AREA_GEOGRAFICA_DESCR     NAZIONE            NAZIONE_DESCR
 Europa (in EU) :2374       15      :1252   UNITED KINGDOM:1252
 Asia           :1190       50      : 897   U.S.A.        : 897
 Nord America   : 897       18      : 576   ITALIA        : 578
 Europa (out EU): 863       41      : 433   HOLLAND       : 576
 Italia         : 578       03      : 400   SOUTH KOREA   : 433
 Oceania        : 271       (Other):2162   SLOVENIA      : 400
 (Other)        : 125       NA's    : 578   (Other)       :2162
    CLIENTE      TIPO_CLIENTE..user..U...retailer..R.
 855800 :1252    R:6298
 651500 : 897
 739200 : 568
 652000 : 271
 633900 : 267
 270700 : 247
 (Other):2796
          ARTICOLO      CODICE_CATALOGO
 W3ANE10220A01:   81    WLK10   : 109
 W3ANE10260A01:   71    D22     :  96
 W5WWLK0100A06:   70    WLK8    :  94
 W3ANE10180A01:   69    WLK16   :  92
 W5WWLK0080A03:   67    SHC7-8  :  89
 W5WWLK0130A03:   67    (Other):5815
 (Other)      :5873     NA's    :   3
                                                   ARTICOLO_DESCR
 ANELLONE OVALE D22 GRIGIO GRADO 10                        :  81
 ANELLONE OVALE D26 GRIGIO GRADO 10                        :  71
 GIUNZIONE MECCANICA WLK10 GRIGIA 40 PZ  10MM GRADO 10:  70
 ANELLONE OVALE D18 GRIGIO GRADO 10                        :  69
 GIUNZIONE MECCANICA WLK13 GRIGIA 20 PZ  13MM GRADO 10:  67
 GIUNZIONE MECCANICA WLK16 GRIGIA 12 PZ  16MM GRADO 10:  67
 (Other)                                                   :5873
 FAMIGLIA_GRUPPO           FAMIGLIA_GRUPPO_DESCR   DATA_FATTURA
 B.12.01:1807     ACCESSORI GR <=80     :1807      Min.   :20160519
 B.12.02: 600     ACCESSORI GR100       :2594      1st Qu.:20170228
 B.12.05:2594     ANELLONI <= 26 MM     : 600      Median :20170907
 B.12.06:1141     ANELLONI <=26MM GR100:1141      Mean   :20171863
 B.12.09: 156     ANELLONI OFFSHORE     : 156      3rd Qu.:20180131
                                                   Max.   :20180531

 CLASSIFICA_05_FINITURA_SUPERFICIALE
 25      :2872
 10      :1985
 12      : 699
 06      : 248
 47      : 156
```

```
(Other): 141
NA's   : 197
CLASSIFICA_05_FINITURA_SUPERFICIALE_DESCR CLASSIFICA_07_IMBALLO
V.epox grigioRAL9007:2872                 02  :3227
V.epox rossoRAL3020 :1985                 06  :1161
V.epox bluRAL5002   : 699                 08  :  43
Zincatura galvanica : 248                 NA's:1867
** NON TROVATO **   : 197
V.epox porporRAL3004: 156
(Other)             : 141
   CLASSIFICA_07_IMBALLO_DESCR CLASSIFICA_09_MODELLO
** NON TROVATO **:1867         AA      :1897
Sacchetto nylon  :1161         WC      :1103
Sacco            :  43         CE      : 641
Scatola          :3227         CC      : 551
                               SB      : 443
                               EC      : 376
                               (Other):1287
     CLASSIFICA_09_MODELLO_DESCR DATI_TECNICI_01_DIAMETRO_MM
Anellone              :1897         Min.   : 0.00
Weisslock             :1103         1st Qu.:10.00
Gancio Clevis Sling: 641            Median :13.00
Gancio Clevis Grab : 551            Mean   :17.03
Set Ricambi           : 443         3rd Qu.:22.00
Gancio Sling          : 376         Max.   :70.00
(Other)               :1287
DATI_TECNICI_02_PASSO_MM DATI_TECNICI_10_PEZZI_PER_CONFEZIONE UM
Min.   :  0.00           Min.   :  0.000                      N:6298
1st Qu.:  0.00           1st Qu.:  1.000
Median :  0.00           Median :  4.000
Mean   : 61.86           Mean   :  8.642
3rd Qu.:135.00           3rd Qu.: 14.000
Max.   :480.00           Max.   :100.000

N..CONFEZIONI       PESO.TOT..Kg         N..PEZZI
Min.   :   1.00   Min.   :   0.035   Min.   :   0.00
1st Qu.:   2.00   1st Qu.:  12.400   1st Qu.:  10.00
Median :   5.00   Median :  33.000   Median :  20.00
Mean   :  23.91   Mean   : 102.415   Mean   :  85.68
3rd Qu.:  15.00   3rd Qu.:  87.500   3rd Qu.:  60.00
Max.   :1800.00   Max.   :4080.000   Max.   :6480.00
```

As we said before, DATA_FATTURA should be a Date and not a number. If it stays as a number, forecast analysis will not success and there can be possible errors during other analysis and visualizations. The following code will transform the column into a Date, the format attribute equals to "%Y%m%d" means that the year is in the first position and its compose by four numbers (%Y), the second that the two-following numbers describe the month (%m) and the last one that the day it is in the last position compose by two numbers(%d). On the other hand, there is no separation between them because in the data set there is not either.

```
#Transforming dates
> sales$DATA_FATTURA <- as.Date.factor(original_sales$DATA_FATTURA,
format = "%Y%m%d")
```

In the `CLASSIFICA_07_IMBALLO_DESCR` column, NA`s values are represented by a string: "** NON TROVATO **", same as `CLASSIFICA_05_FINITURA_SUPERFICIALE_DESCR`. In order to homogenize to data set, we changed that value for NA using the following code,

```
#Replacing "** NON TROVATO **" values
sales[sales$CLASSIFICA_07_IMBALLO_DESCR == "** NON TROVATO **",
"CLASSIFICA_07_IMBALLO_DESCR"] <- NA
sales[sales$CLASSIFICA_05_FINITURA_SUPERFICIALE_DESCR == "** NON
TROVATO **", "CLASSIFICA_05_FINITURA_SUPERFICIALE_DESCR"] <- NA
```

Many missing values introduce an error on our results because R does not know how to treat them, any operation that involves a NA value has a NA value as a result. It is necessary to check for missing values to make our predictions more accurately.

```
#Checking for missing values (NA) by column
missing_values <- sapply(sales, is.na)
colSums(missing_values)

                     AREA_GEOGRAFICA_DESCR
                                         0
                                   NAZIONE
                                       578
                             NAZIONE_DESCR
                                         0
                                  CLIENTE
                                         0
      TIPO_CLIENTE..user..U...retailer..R.
                                         0
                                 ARTICOLO
                                         0
                          CODICE_CATALOGO
                                         3
                           ARTICOLO_DESCR
                                         0
                          FAMIGLIA_GRUPPO
```

```
                                   0
              FAMIGLIA_GRUPPO_DESCR
                                   0
                        DATA_FATTURA
                                   0
     CLASSIFICA_05_FINITURA_SUPERFICIALE
                                 197
CLASSIFICA_05_FINITURA_SUPERFICIALE_DESCR
                                 197
               CLASSIFICA_07_IMBALLO
                                1867
          CLASSIFICA_07_IMBALLO_DESCR
                                1867
              CLASSIFICA_09_MODELLO
                                   0
         CLASSIFICA_09_MODELLO_DESCR
                                   0
         DATI_TECNICI_01_DIAMETRO_MM
                                   0
            DATI_TECNICI_02_PASSO_MM
                                   0
  DATI_TECNICI_10_PEZZI_PER_CONFEZIONE
                                   0
                                  UM
                                   0
                       N..CONFEZIONI
                                   0
                        PESO.TOT..Kg
                                   0
                            N..PEZZI
                                   0
```

According to results there are 578 NA's on NAZIONE, but this column and NAZIONE_DESCR, both have the same information. NA's values on NAZIONE represents "ITALIA" on NAZIONE_DESCR. There are 197 NAs on CLASSIFICA_05_FINITURA_SUPERFICIALE and 1867 on CLASSIFICA_07_IMBALLO but these columns have no major value for us.

There also NA`s values on CODICE_CATALOG that correspond with the following products:

- SET DI CONNESSIONE TARGHETTA/ACCESSORIO
- MAGLIA PIEGATA 13X60 MM NN MARC.LOTTO
- SUB LINK 40X170 MM NN MARC.L40

These three products are not in the catalog so is not possible to assign a catalog code, nevertheless these products are rarely ordered.

It is possible to see that some columns contain the same information. Such as, NAZIONE and NAZIONE_DESCR, FAMIGLIA_GRUPPO and FAMIGLIA_GRUPPO_DESCR, CLASSIFICA_05_FINITURA_SUPERFICIALE and

`CLASSIFICA_05_FINITURA_SUPERFICIALE_DESCR`, `CLASSIFICA_07_IMBALLO` and `CLASSIFICA_07_IMBALLO_DESCR`, `CLASSIFICA_09_MODELLO` and `CLASSIFICA_09_MODELLO_DESCR`. We will select only one of each of them. Furthermore, `UM` and `TIPO_CLIENTE..user..U...retailer..R.`, both have the same value, "N" and "R" respectively, all around the data set. In order to make the analysis more efficiency we will select only the columns with useful information using **dplyr** functions. [dplyr]

```
#Selecting columns
sales <- sales %>%
  select(CLIENTE, NAZIONE_DESCR, DATA_FATTURA, ARTICOLO,
CODICE_CATALOGO, FAMIGLIA_GRUPPO,
CLASSIFICA_05_FINITURA_SUPERFICIALE_DESCR,
CLASSIFICA_07_IMBALLO_DESCR, CLASSIFICA_09_MODELLO, N..PEZZI)
```

Finally, printing the structure again we check everything is ready for the analysis.

```
> glimpse(sales)
Observations: 6,298
Variables: 9
$ CLIENTE                                   <fct> 739200, 574000, 88...
$ NAZIONE_DESCR                             <fct> HOLLAND, SLOVENIA,...
$ DATA_FATTURA                              <date> 2016-05-19, 2016-...
$ ARTICOLO                                  <fct> W3ANA30220A01, W5C...
$ CODICE_CATALOGO                           <fct> D22E16, GSC10, F22...
$ FAMIGLIA_GRUPPO                           <fct> B.12.06, B.12.05, ...
$ CLASSIFICA_05_FINITURA_SUPERFICIALE_DESCR <fct> V.epox grigioRAL90...
$ CLASSIFICA_07_IMBALLO_DESCR               <fct> NA, Scatola, NA, N...
$ CLASSIFICA_09_MODELLO                     <fct> AA, CC, AA, AA, CE...
$ N..PEZZI                                  <dbl> 50, 10, 2, 2, 10, ...
```

Transforming

During the analysis we are going to be working with regular Time Series. A regular time series is a series of data points indexed taken at successive equally spaced points in time.

As we can imagine, our data is not regular. In order to make it regular we design the following function, called *ts_by_()*. In this function the **lubridate**, **dplyr** and **tidyr** packages were used. [lubridate][dplyr][tidyr]

```
#This function was created to treat with irregular time series for the thesis
"Data analysis for Anticipation of Future Demand".
#      "data" first column must be Dates and the second column, values.
#      "from" and "to" should be in "%Y-%m-%d" format.
#      "frequency" possible values are "month" and "week" to order the data set
by months or weeks respectively.
#The function will return a regular "ts" object.

ts_by_ <- function(data, from, to, frequency){
  require(lubridate)
  require(dplyr)
  require(tidyr)

  order  <-  as.data.frame(seq(from  =  as.Date(from),  to  =  as.Date(to),  by  =
frequency ))

  #By month
  if(frequency == "month"){
  #Months from, from to, to
    months <-  order %>%
      mutate(month_num = month(order[, 1])) %>%
      mutate(year_num = year(order[, 1])) %>%
      unite("month_year", month_num, year_num) %>%
      rename("Date" = 1) %>%
      select(Date, month_year)
  #Ordering data by month
    data_month <- data %>%
      mutate(month_num = month(data[, 1])) %>%
      mutate(year_num = year(data[, 1])) %>%
      group_by(month_num, year_num) %>%
      summarise(pieces = sum(N..PEZZI)) %>%
      unite("month_year", month_num, year_num) %>%
      select(month_year, pieces)
  #JOINING together
    data_ <- months %>%
      left_join(data_month, by = "month_year") %>%
      select(pieces)
  #Replacing NAs
    data_[is.na(data_)] <- 0
  #Creating a ts object
    data_by_month  <-  ts(data_,  start  =  c(year(from),  month(from)),  end  =
c(year(to), month(to)), frequency = 12)
```

```
  #Returning result
    return(data_by_month)
  }

  #By week
  if(frequency == "week"){
  #Weeks from, from to, to
    weeks <- order %>%
      mutate(week_num = week(order[, 1])) %>%
      mutate(year_num = year(order[, 1])) %>%
      unite("week_year", week_num, year_num) %>%
      rename("Date" = 1) %>%
      select(Date, week_year)
  #Ordering data by week
    data_week <- data %>%
      mutate(week_num = week(data[, 1])) %>%
      mutate(year_num = year(data[, 1])) %>%
      group_by(week_num, year_num) %>%
      summarise(pieces = sum(N..PEZZI)) %>%
      unite("week_year", week_num, year_num) %>%
      select(week_year, pieces)
  #JOINING together
    data_ <- weeks %>%
      left_join(data_week, by = "week_year") %>%
      select(pieces)
  #Replacing NAs
    data_[is.na(data_)] <- 0
  #Creating a ts object
    data_by_week   <-   ts(data_,   start   =   c(year(from),month(from)),   end   =
c(year(to),month(to)), frequency = 52)
  #Returning result
    return(data_by_week)
  }
}

save(ts_by_, file = "ts_by_.rda")
```

We can use the function anywhere just loading the `.rda` file called `"ts_by_.rda"` before, using the following syntax,

```
load("ts_by_.rda")
```

Instructions to use the function are in the first lines.

Analyzing

The following graphic shows the total number of pieces by month across time. Graphics in this analysis were made mostly by **ggplot2** package. [ggplot2]

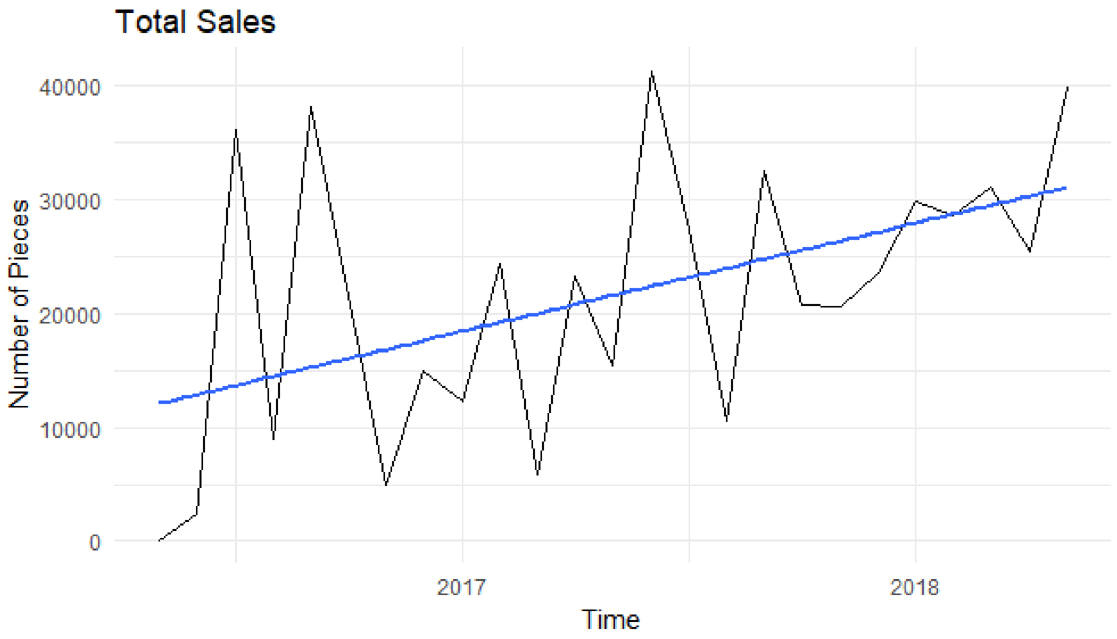


Fig. 10: Total sales

It is easy to see that sells trends to grow with the time on KITO.

The following graphic shows the distribution of KITO´s sales by country.

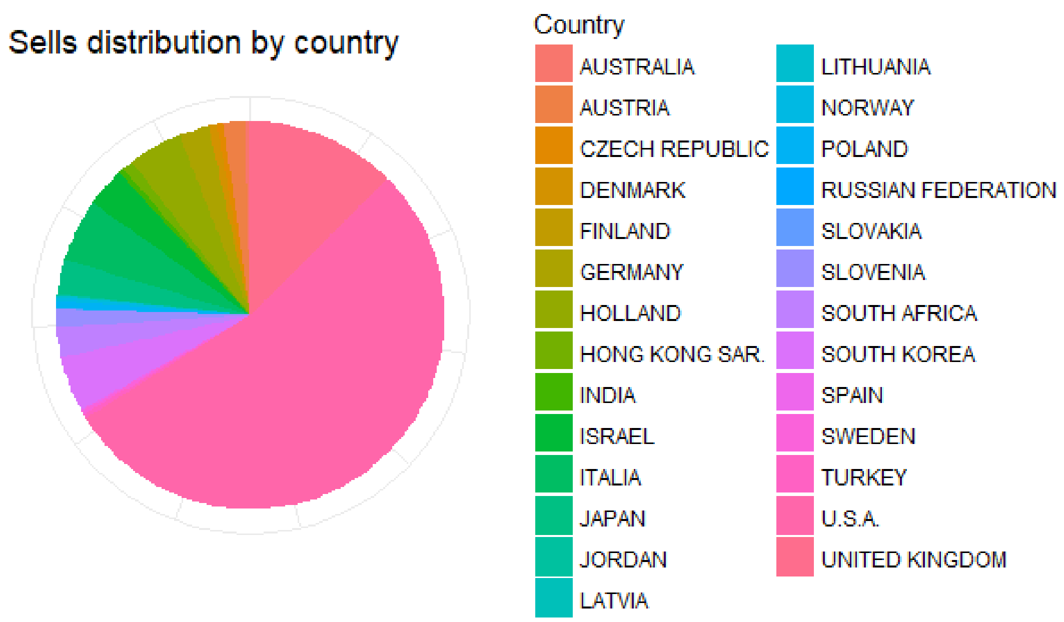


Fig. 11: Distribution of sales by country

As it can see, more than 50% of sales are done in U.S.A. (289.393 pieces). In the following table we can see the top 9 best buyers countries.

Table 1: Sells Distribution around the world.

COUNTRY	NUMBER OF PIECES	PERCENTAGE (%)
U.S.A.	289.393	53,63
UNITED KINGDOM	67.461	12,5
ITALIA	28.264	5,24
SOUTH KOREA	24.714	4,58
HOLLAND	24.327	4,51
ISRAEL	17.729	3,29
JAPAN	15.687	2,91
GERMANY	13.428	2,49
SOUTH AFRICA	13.416	2,49

KITO's plant produce more than 345 different articles. Make a complete analysis of all of them will consume a lot of time and is not efficient. In other to focalize our efforts in the important product we will select the best sell products. As follows,

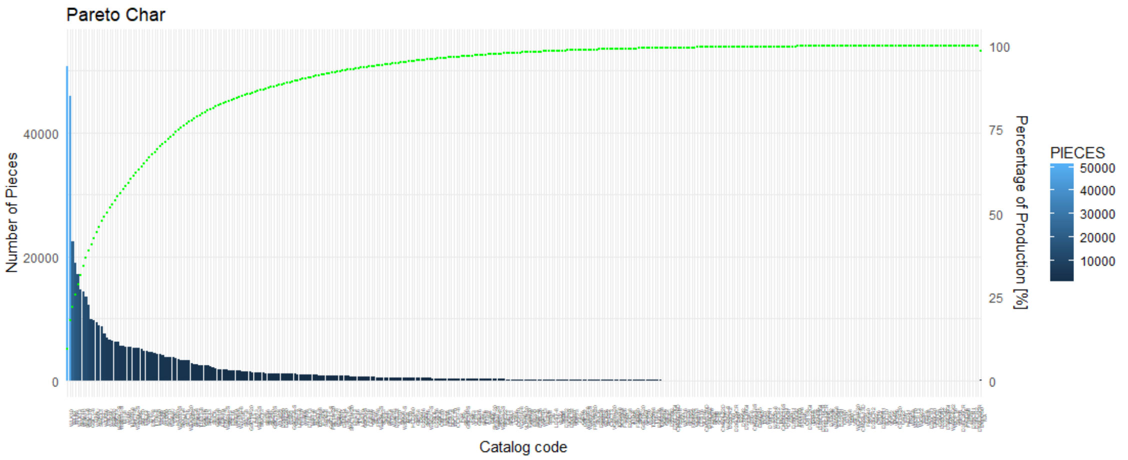


Fig. 12: Distribution of sales in Pareto Char

It is difficult to obtain information from this graphic and most of the products in the catalog are rarely sell.

The following graphic shows only the 25 most sell articles in the catalog.

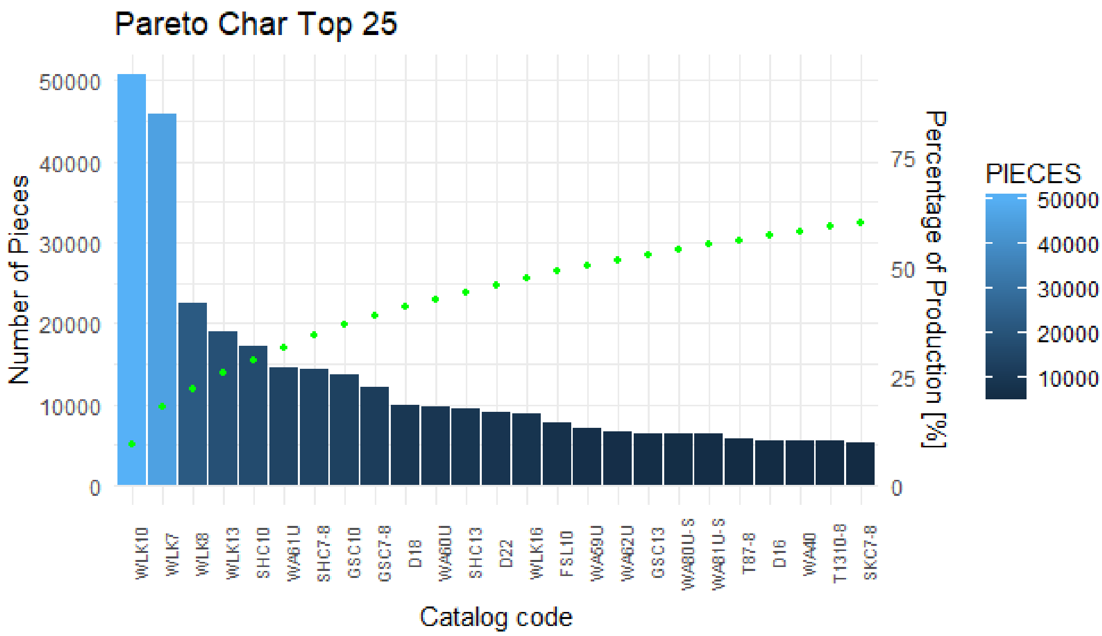
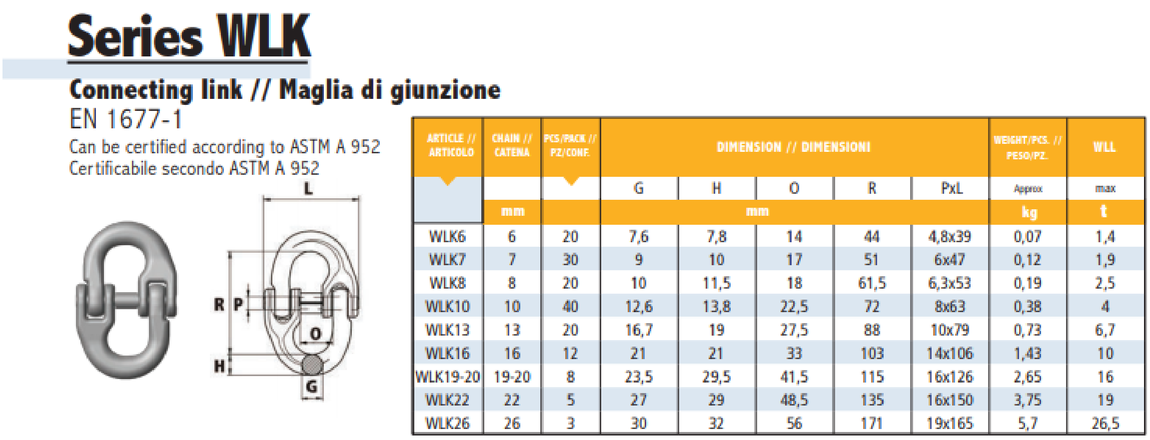


Fig. 13: Top 25 best sell products.

As we can see, the top 4 best sell articles are Connecting links from the “Series WLK”. These four products represent the 25,57% of the company sales and top 25 represent the 60.147%. In the following image we can see the Series WLK in the KITO’s catalog.



ARTICLE // ARTICOLO	CHAIN // CATENA	PCS/PACK // PZ/CONF.	DIMENSION // DIMENSIONI					WEIGHT/PCS. // PESO/PZ.	WLL
			G	H	O	R	PxL	Approx	max
	mm		mm					kg	t
WLK6	6	20	7,6	7,8	14	44	4,8x39	0,07	1,4
WLK7	7	30	9	10	17	51	6x47	0,12	1,9
WLK8	8	20	10	11,5	18	61,5	6,3x53	0,19	2,5
WLK10	10	40	12,6	13,8	22,5	72	8x63	0,38	4
WLK13	13	20	16,7	19	27,5	88	10x79	0,73	6,7
WLK16	16	12	21	21	33	103	14x106	1,43	10
WLK19-20	19-20	8	23,5	29,5	41,5	115	16x126	2,65	16
WLK22	22	5	27	29	48,5	135	16x150	3,75	19
WLK26	26	3	30	32	56	171	19x165	5,7	26,5

Fig. 14: Series WLK on KITOS’s catalog [13]

As we can see, in the catalog there is information about different aspect of every piece (the information change depending on the Serie), for more information about other Series see the KITO’s catalog. [13]

In order to forecast time series, it is important to study the seasonality.

Seasonality is when a time series data experiment regular, predictable and annual pattern. It is different from a cyclic pattern because in the second one the duration of the fluctuations is not fixed, can be four months or four years. For analyze the seasonality on our case, we use the *ggseasonplot()* function from with the polar argument equals to `TRUE`. This function is form **fpp2** package which is a useful for forecast analysis and it is not in R by default, so we need to install it as we did before with other packages. The following graphic was made with data of the total sales. [fpp2]

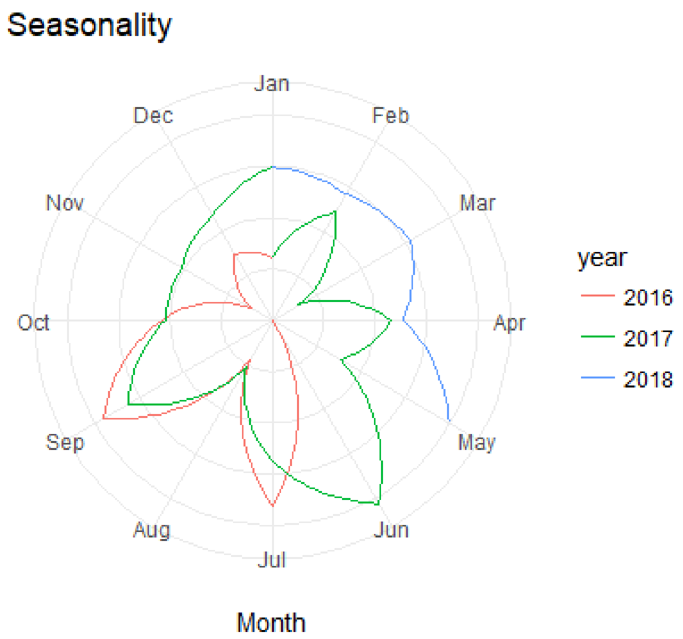


Fig. 15: Seasonality analysis of KITO sales.

As we can see in this graphic, September looks like a month of a high activity, same as June-July, while August is the opposite. Make sense because the month of August is vacations for most of the north hemispheric countries (main clients of KITO, as we saw on Fig. 11) and for KITO itself, so its logical to think that this are months of low activity in factories.

This trend is repeated in most of the articles, for example in the WLK10,

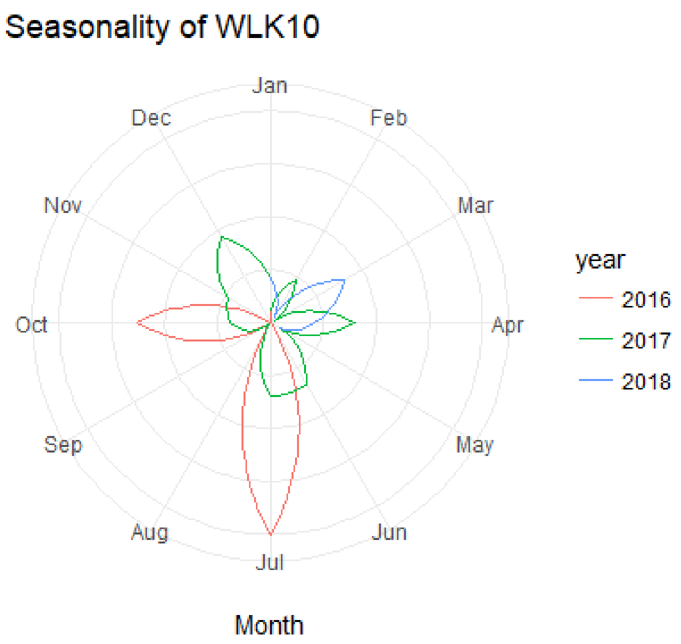


Fig. 16: Seasonality of WLK10.

The seasonality of this data set is not very strong and is too soon to say if it is additive or multiplicative seasonality but during the forecasting we will use Additive seasonality because it is easier and give less problems.

Forecasting of Demand

In this part we will try to predict the future demand of some of the best sell articles of KITO Chain Italia S.R.L. but have to evaluate the method to see which one has better results in our case. To check our result we will use the *accuracy()* function, from **forecast** package that calculate automatically the following parameters:

- ME: Mean Error
- RMSE: Root Mean Squared Error
- MAE: Mean Absolute Error
- MPE: Mean Percentage Error
- MAPE: Mean Absolute Percentage Error
- MASE: Mean Absolute Scaled Error
- ACF1: Autocorrelation of errors at lag 1.

As an agreement we will pick the method with the minimum Mean Absolute Percentage Error (MAPE). This function need as a parameter the training data set and the validation data set. [forecast]

As we said before we are going to evaluated two methods: Seasonal Naïve, and Holt Winter.

This forecast graphic shows that 90% of time the demand will be in the soft blue area while 80% of time it will be on the dark blue area. The blue line is the mean value while the red line is the real value of demand in every period. Graphics here were done by *plot()* function from **graphics** package. [graphics]

This forecast was done with Seasonal Naïve method and *snaive()* function.

```
#Total sales
total_sales <- select(sales, DATA_FATTURA, N..PEZZI)

#Regular ts by month
load("ts_by_.rda")
total_ts <- ts_by_(total_sales, from = "2016-05-01", to = "2018-05-30",
frequency = "month")

#Generating training data set
total_window <- window(total_ts, end = c(2018, 3))
```

```
#Using snaive()
total_sn <- snaive(total_window, h = 6)
```

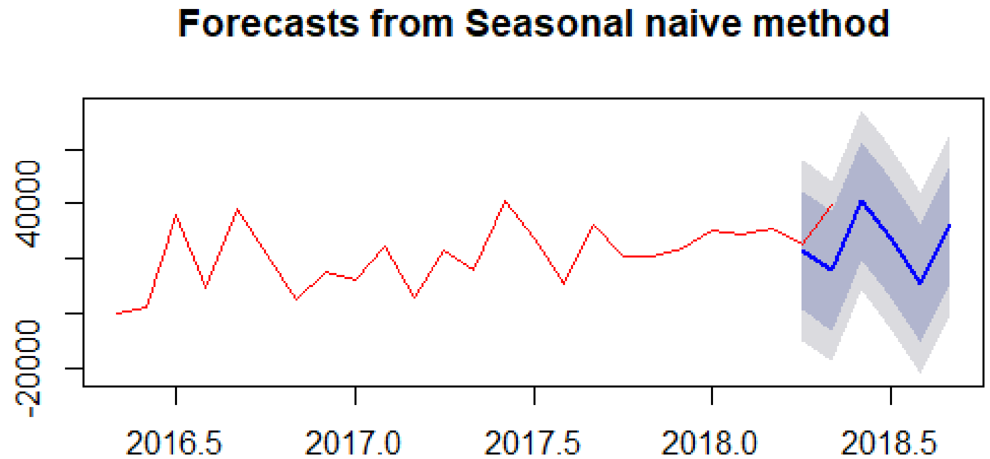


Fig. 17: Seasonal Naïve forecast of total sales.

```
> accuracy(total_sn, total_ts)
                   ME     RMSE      MAE      MPE     MAPE     MASE
Training set 10174.55 16880.61 12907.45 38.47314 48.02643 1.000000
Test set     13337.50 17342.93 13337.50 35.07810 35.07810 1.033318
                    ACF1 Theil's U
Training set -0.06882887        NA
Test set     -0.50000000  1.698164
```

Where `total_ts` is the complete data set and `total_sn` is the Season naïve forecast of `total_ts` form the second period of 2018.

```r
#Using Holt Winters
total_hw <- hw(total_window, h = 5)
```

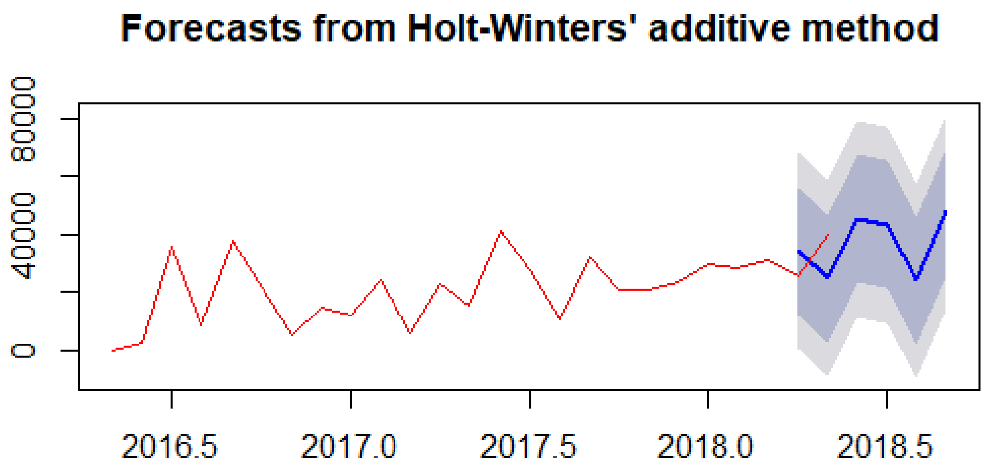


Fig. 18: Holt Winters forecast of total sales.

```
> accuracy(total_hw, total_ts)
                    ME      RMSE       MAE        MPE      MAPE
Training set -138.2215  9523.661  6038.141 771.502133 807.64246
Test set     3109.5064 12336.013 11937.679   1.544991  36.22784
                  MASE       ACF1 Theil's U
Training set 0.4678027 -0.0307952        NA
Test set     0.9248670 -0.5000000  1.046251
```

As an example, this is the Dynamic harmonic regression method that seams to be more accurately but is more complicated and less reproductible because has some parameters as K that needs to be estimated case by case and could work better in the future with more data. That are the reason why we do not include it on the options on this particular forecast analysis but can be used in future analysis, in particular if we use the *ts_by_()* function with `frequency` argument equals to "week", the frequency could be too high for other methods but not for Dynamic harmonic regression.

```r
#Using Dynamic harmonic regression
fit <- auto.arima(total_window, xreg = fourier(total_window, K = 5),
seasonal = FALSE)
total_dhr <- forecast(fit, xreg = fourier(total_window, K = 5), h = 5)
```

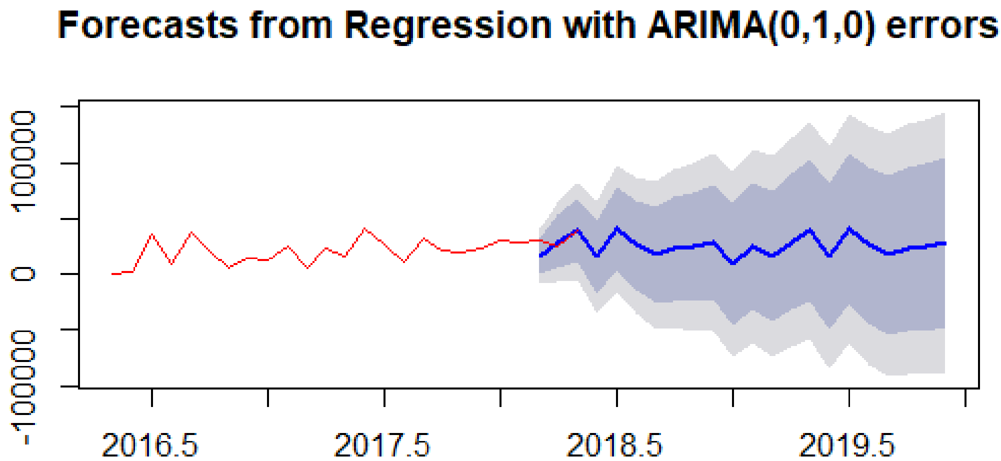


Fig. 19: Dynamic harmonic regression of total sales.

```
> accuracy(total_dhr, total_ts)
                   ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set  736.0506 9050.445 6469.979 -21.26198 52.46666 0.5540409 -0.5375037
Test set     3980.5774 8909.728 6581.196  11.63212 21.84906 0.5635647 -0.3258831
             Theil's U
Training set        NA
Test set     0.2200088
```

As we can see Holt Winters seems to be more accurately than Seasonal Naïve, so we will use this method to forecast the Top 5 products according to the Pareto Chart. To select the sales of every product function *filter()* from **dplyr** was used with `CODICE_CATALOGO == "......"` as a condition. The rest of the process is the same as with total sales.

1. WLK10

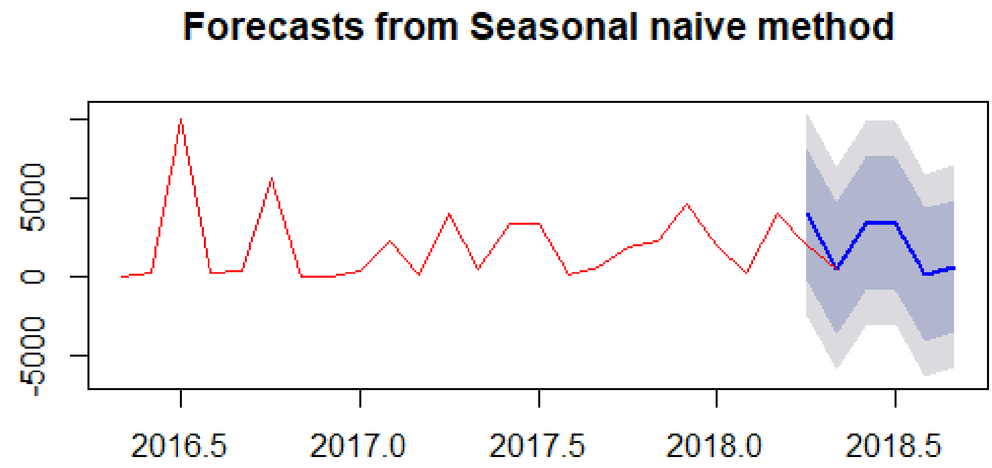


Fig. 20: Seasonal Naïve forecast of WLK10.

```
> accuracy(wlk10_sn, wlk10_ts)
                    ME     RMSE      MAE       MPE      MAPE      MASE
Training set  302.1818 3304.708 2684.727 -52.12372 162.27906 1.0000000
Test set    -1005.0000 1359.136 1005.000 -54.80769  54.80769 0.3743397
                   ACF1  Theil's U
Training set -0.1760162         NA
Test set     -0.5000000 0.05769231
```

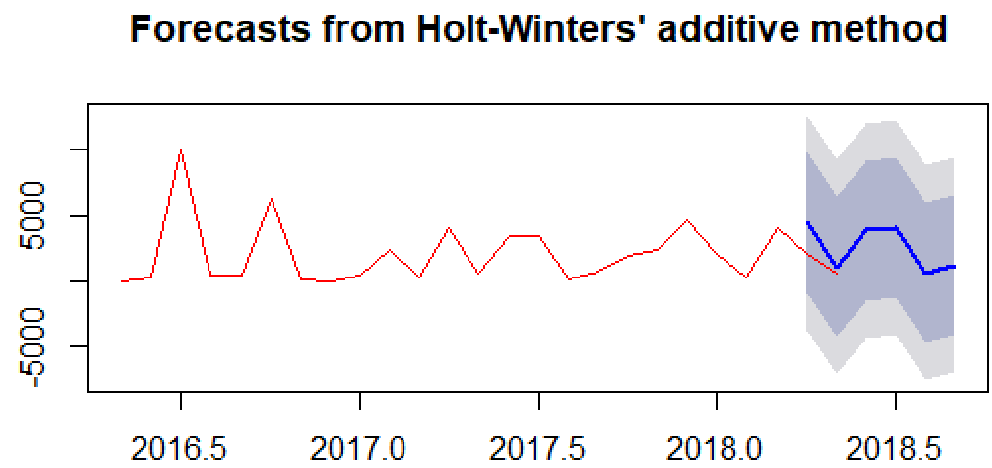


Fig. 21: Holt Winter forecast of WLK10.

```
> accuracy(wlk10_hw, wlk10_ts)
                    ME     RMSE      MAE       MPE     MAPE      MASE
Training set  -18.0404 2312.734 1345.792       NaN      Inf 0.5012768
Test set    -1456.7393 1721.468 1456.739 -108.9404 108.9404 0.5426023
                   ACF1 Theil's U
Training set -0.1641877        NA
Test set     -0.5000000 0.3458208
```

2. WLK7

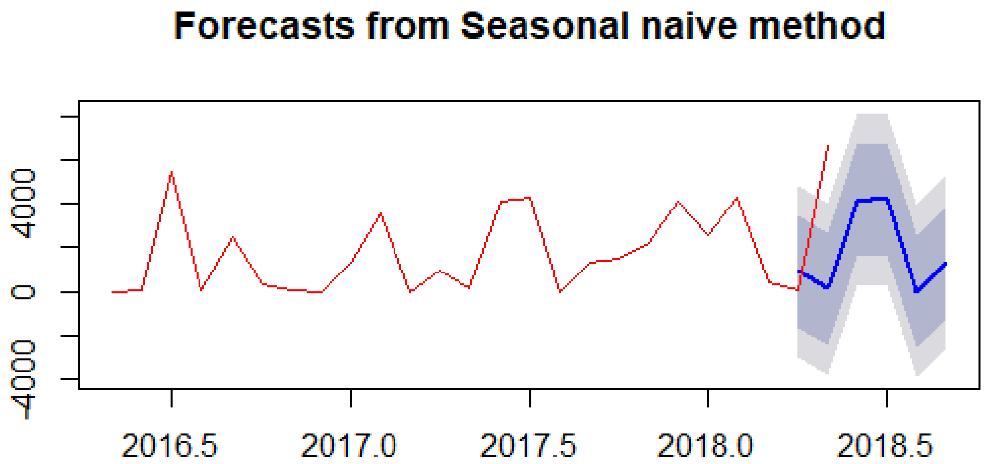


Fig. 22: Seasonal Naïve forecast of WLK7.

```
> accuracy(wlk7_sn, wlk7_ts)
                    ME     RMSE      MAE       MPE     MAPE     MASE
Training set 1044.727 1997.324 1489.273      -Inf      Inf 1.000000
Test set     2790.000 4626.035 3690.000 -1450.909 1549.091 2.477719
                    ACF1 Theil's U
Training set -0.01218963        NA
Test set     -0.50000000 0.9863014
```

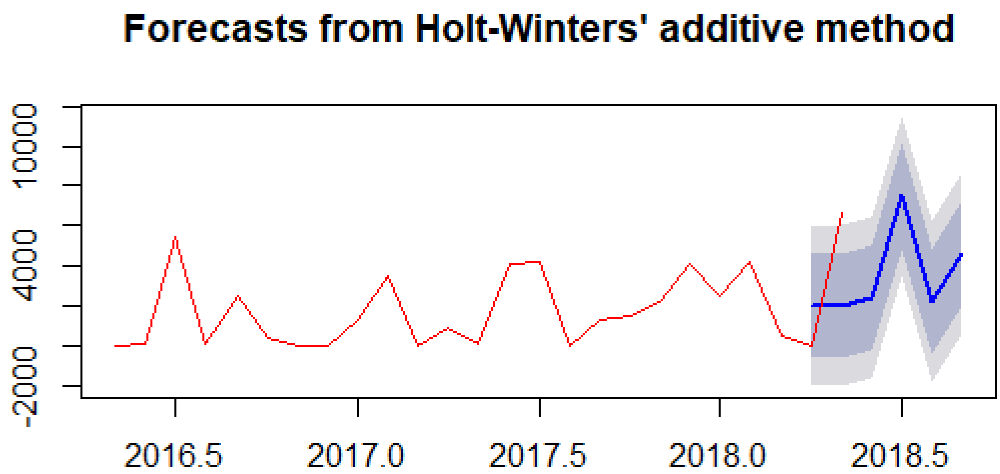


Fig. 23: Holt Winter forecast of WLK7.

```
> accuracy(wlk7_hw, wlk7_ts)
                     ME     RMSE       MAE       MPE     MAPE      MASE
Training set  -32.33401 1128.320  645.9212       NaN      Inf 0.4337159
Test set     1316.99985 3514.061 3257.9344 -3200.232 3269.549 2.1876009
                     ACF1 Theil's U
Training set  0.003059321        NA
Test set     -0.500000000  0.696337
```

3. WLK8

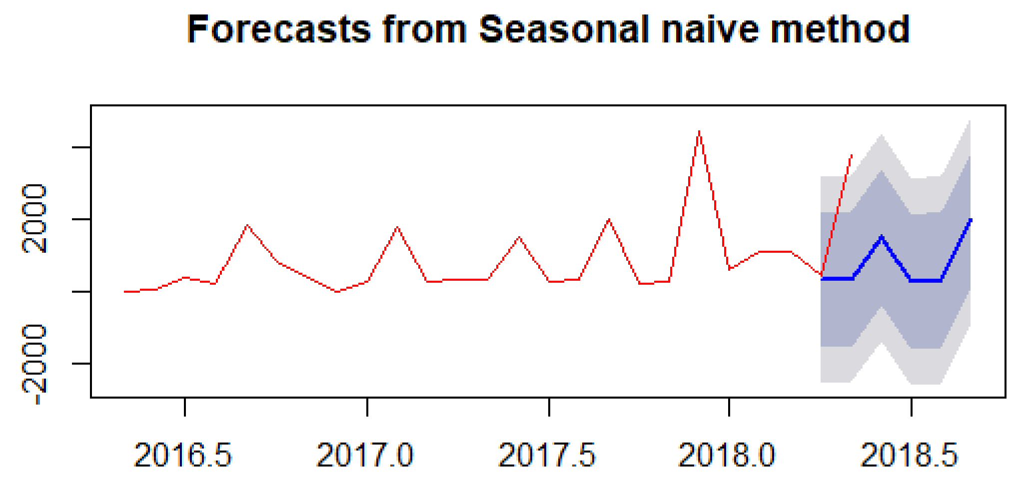


Fig. 24: Seasonal Naïve forecast of WLK8.

```
> accuracy(wlk8_sn, wlk8_ts)
                    ME    RMSE       MAE       MPE     MAPE     MASE
Training set  560.9091 1463.52  833.4545  3.291626 83.44406 1.000000
Test set     1801.0000 2436.49 1801.0000 62.896328 62.89633 2.160886
                  ACF1 Theil's U
Training set -0.117699        NA
Test set     -0.500000  1.036123
```

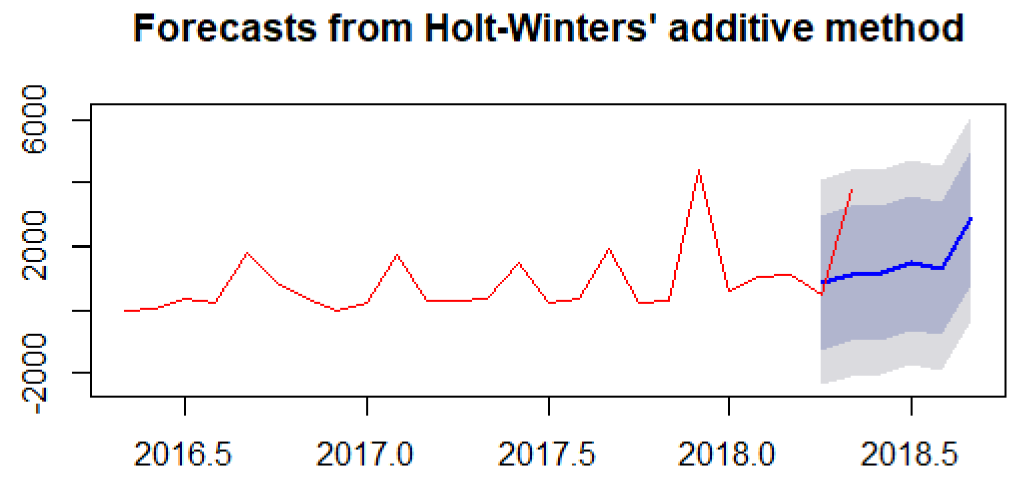


Fig. 25: Holt Winter forecast of WLK8.

```
> accuracy(wlk8_hw, wlk8_ts)
                     ME      RMSE      MAE       MPE     MAPE     MASE
Training set  -10.54596  911.6635  436.276      -Inf      Inf 0.523455
Test set     1124.67749 1887.5562 1515.905 -7.614892 77.43464 1.818822
                   ACF1 Theil's U
Training set -0.1200141        NA
Test set     -0.5000000 0.7948774
```

4. WLK13

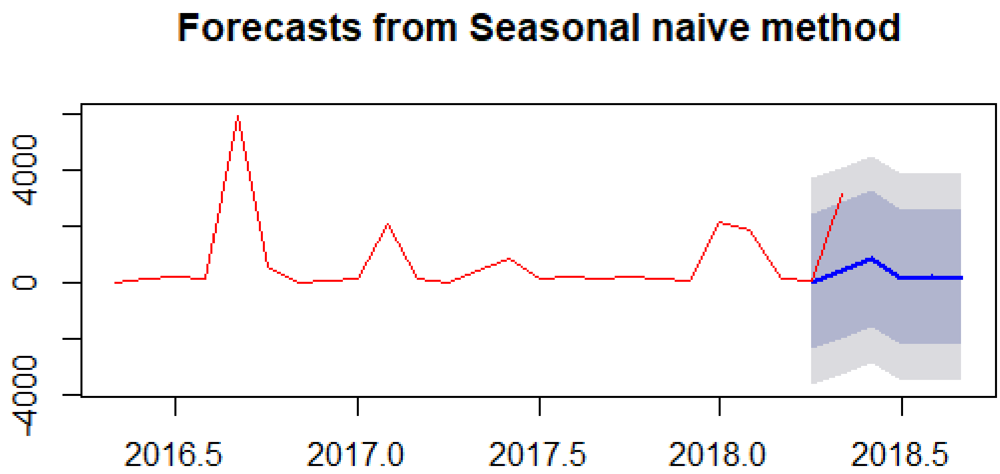


Fig. 26: Seasonal Naïve forecast of WLK13.

```
> accuracy(wlk13_sn, wlk13_ts)
                    ME     RMSE       MAE       MPE     MAPE     MASE
Training set -278.6364 1869.371  901.1818 -324.7443 403.1740 1.000000
Test set     1378.0000 1923.499 1378.0000   73.3121  73.3121 1.529103
                    ACF1 Theil's U
Training set  0.01204131        NA
Test set     -0.50000000 0.8831169
```

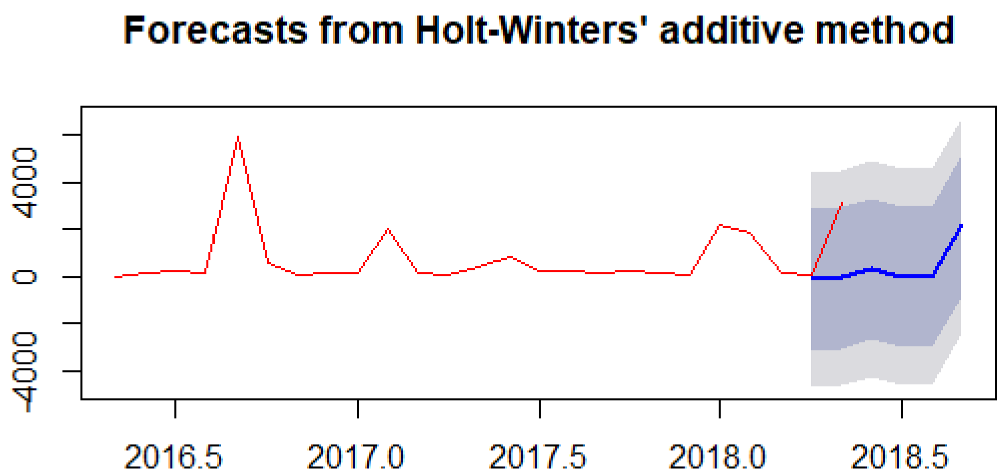


Fig. 27: Holt Winter forecast of WLK13.

```
> accuracy(wlk13_hw, wlk13_ts)
                      ME     RMSE       MAE      MPE     MAPE     MASE
Training set  -26.89992 1285.915  502.7342      Inf      Inf 0.557861
Test set     1715.93478 2279.556 1715.9348 230.6254 230.6254 1.904094
                    ACF1 Theil's U
Training set  0.02155194        NA
Test set     -0.50000000  1.044345
```

5. SHC10

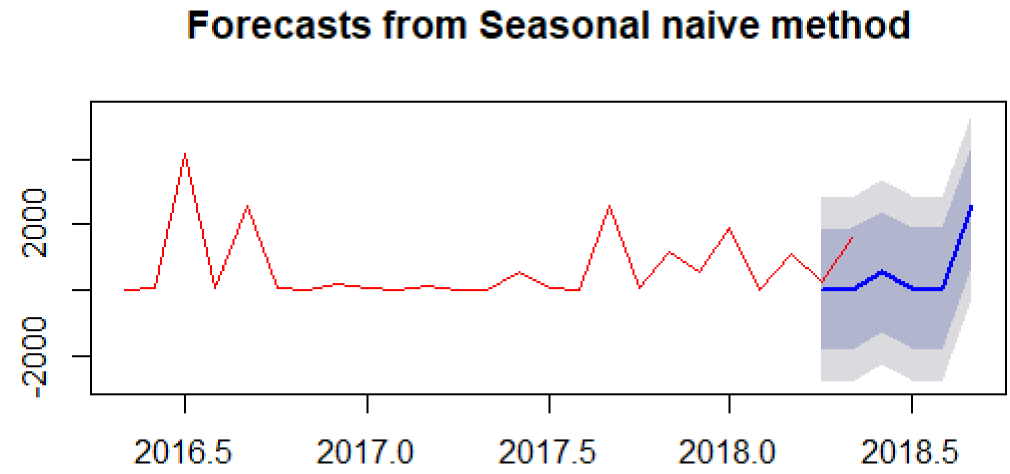


Fig. 28: Seasonal Naïve forecast of SHC10.

```
> accuracy(shc10_sn, shc10_ts)
                    ME     RMSE      MAE        MPE     MAPE     MASE
Training set  60.09091 1439.020 812.8182 -581.91765 696.87023 1.000000
Test set     894.00000 1102.974 894.0000   99.35897  99.35897 1.099877
                    ACF1 Theil's U
Training set -0.02451092        NA
Test set     -0.50000000   1.17378
```

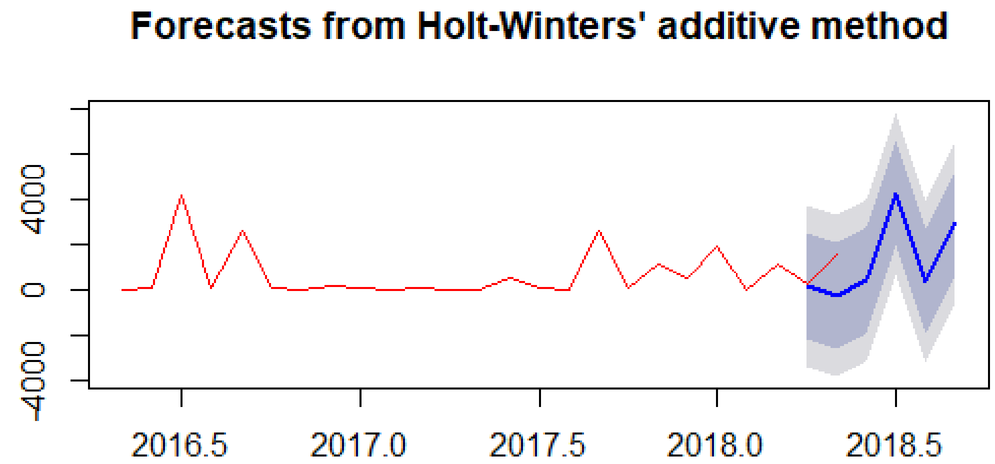


Fig. 29: Holt Winter forecast of SHC10.

```
> accuracy(shc10_hw, shc10_ts)
                    ME      RMSE      MAE      MPE     MAPE      MASE
Training set -20.92533  997.5222 438.4190      NaN      Inf 0.5393813
Test set     950.27915 1294.9849 950.2792 72.87481 72.87481 1.1691165
                     ACF1 Theil's U
Training set -0.009003745        NA
Test set     -0.500000000  1.394837
```

Results

According to the forecast analysis, we recommend producing products enough to cover the 80% area in the top 5 best selled products.

Table 2: Forecast result of the top 5 products and total sales.

	PRODUCT (CATALOG CODE)	JUNE (2018)	JULY (2018)	AUGUST (2018)	SEPTEMBER (2018)	METHOD	MAPE (%)
	TOTAL	62.934	49.006	32.168	54.121	SN	35,07
1	WLK10	7.655	7.711	4.375	4.975	SN	50,08
2	WLK7	6.679	6.819	2.559	3.869	SN	1549
3	WLK8	3.355	2.121	2.195	3.835	SN	62,89
4	WLK13	3.245	2.575	2.635	2.555	SN	73,31
5	SHC10	2.765	6.594	2.710	5.233	HW	72,87

This method can be reproducing to analyze every product produce in KITO's plant.
As we can see, the simpler method (Seasonal Naïve) looks better than Holt Winter to forecast demand on KITO's plant. This fact probably will change with time, across time KITO will register more sales that is data that will do our analysis better and because Holt Winter is essentially a more complicated method that can manage trends and seasonality probably would be better.

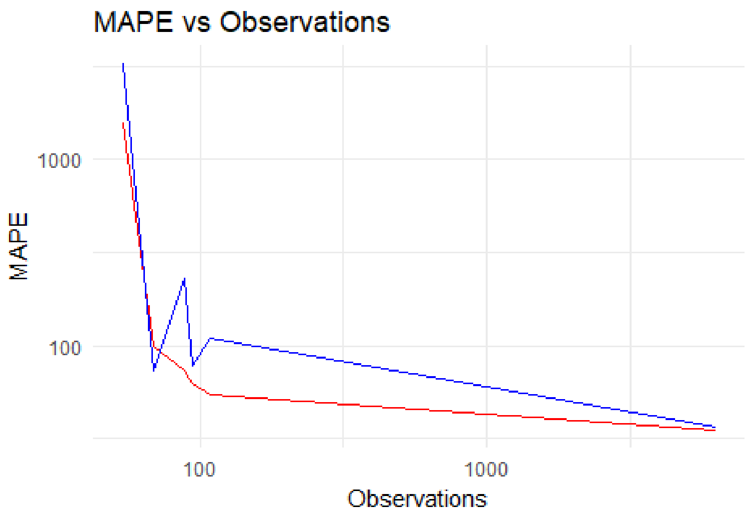


Fig. 30: MAPE vs Observations. S. Naïve and Holt Winter.

As we can see on Fig. 30 MAPE trends to decrease with the number of observations and Holt Winter has a higher decreasing pendent than Seasonal Naïve.

Some considerations. Dynamic harmonic regression is a particular case of Dynamic. Dynamic regression can be used to add more information to a model forecast. For example, if we are forecasting demand, price or advertising expenses could be valuable information to add. In Dynamic harmonic regression Fourier terms are used to handle multiple seasonality. The results of this method is in the Fig. 19. This can be modelized with the ARIMA function setting `seasonal` equals to `FALSE` and `xreg` equals to a Fourier series of the data. ARIMA is an acronym that stands for Auto-Regressive Integrated Moving Average. As we said this method can be better in the future.

Future opportunities with data on KITO

Data analysis is an amazing tool that can provide us a lot of useful information and as we said on the introduction disinformation is very expensive.

The new machines that KITO bought are provided by many sensors that give us thousands of dates that bring us the opportunity to improve the efficiency of the production process. For example, we can anticipate when some key components of the machines will break and replace them before, minimizing the time that the machine remains paralyzed and unforeseen when we have to make articles.

Data can help KITO to design new products using information of the best selled products, choosing products requirements knowing the client's needs and take marketing decisions using the information of sales by country.

As we said before there are other methods that with more density of data could be better for this analysis. In the future, KITO will have more data and new methods could be better for understand and predict the demand.

The implementation of new data analysis technics as real-time analysis, machine learning and deep learning are recommended in KITO's plant.

Real-time analytics is the use of data as soon as it enters in the system and can help to improve the response time. Machine learning is the use of statistical techniques to give computers the ability to "program itself" with data and can be use, for example to calibrate the machines automatically. Deep learning is a part of machine learning with algorithms inspired by the structure and function of the brain and can be use to recognize failure products in the production line automatically.

Conclusion

During this project we tried to reduce the disinformation on KITO's plant forecasting the future demand giving KITO the possibility to be ready and anticipate future events.

As we saw before, despite of the simplicity of Season Naïve is the method with the higher cost-benefit ratio and provide very useful information.

Season Naïve looks like the best way to forecast KITO's data for now but we also saw that the error of our forecast is big, so we must be carefully using this information.

Seasonal Naïve result shows a highest MAPE of 1549.09% and a minimum of 35,07% while Holt Winters a maximum of 3.269,55% and a minimum of 36,07%. The minimum error was founded on Total sales and the higher on WLK7 coinciding with the higher and lower number of observations (6.298 and 45 respectively). This together with the Fig. 30 makes us conclude than MAPE is decreasing with the number of observations, so we can expect better result in the future when the number of observation will be higher. Furthermore, the pendent of Holt winter method in the same graphic is higher than the other method, that means we can expect better result in the future with Holt Winter method.

Concluding, the method do not seems to be accurately to forecast the KITO Chain Italy demand right now because of the low amount of data they have because is a young company but the analysis of the result shows that the method could be use in the future for that end with good results.

References

[1] http://kito.com

[2] http://www.kitochainitalia.com/en

[3] https://www.kito.net/en/news/kito-acquisition-of-weissenfels-tech-chains/

[4] Industry 4.0 Analytical Study, European Parliament, POLICY DEPARTMENT A: ECONOMIC AND SCIENTIFIC POLICY.

[5] https://en.wikipedia.org/wiki/Data_science

[6] https://trends.google.com/trends/

[7] Smart Manufacturing Must Embrace Big Data by Andrew Kusiak, University of Iowa.

[8] https://www.r-project.org/

[9] https://products.office.com/en/excel

[10] https://www.rstudio.com/products/rstudio/

[11] https://en.wikipedia.org/wiki/Forecasting

[12] https://en.wikipedia.org/wiki/Heat_treating

[RcmdrMisc] https://cran.r-project.org/web/packages/RcmdrMisc/index.html

[dplyr] https://cran.r-project.org/web/packages/dplyr/

[utils] https://cran.r-project.org/web/packages/R.utils/index.html

[base] https://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html

[lubridate] https://cran.r-project.org/web/packages/lubridate/

[tidyr] https://cran.r-project.org/web/packages/tidyr/index.html

[ggplot2] https://cran.r-project.org/web/packages/ggplot2/

[13] http://www.kitochainitalia.com/en/products/download/KCI_catalogue_ed5_web.pdf

[fpp2] https://cran.r-project.org/web/packages/fpp2/index.html

[forecast] https://cran.r-project.org/web/packages/forecast/index.html

[graphics] http://stat.ethz.ch/R-manual/R-devel/library/graphics/html/00Index.html

[14] Book: Forecasting: Principles and Practice by Rob J Hyndman and George Athanasopoulos. https://www.otexts.org/fpp2