Analysis Report

Data Description

We are analyzing the online sales of one product in different countries, and we are measuring the Net Sales(USD).
In a matter of fact:

Net Sales (USD) = Gross sales (USD) + VAT/Tax - Returns (USD)
= Gross units sold*Sale price + VAT/Tax - Returns (USD) * Sale price
= Sale price * Net Units Sold + VAT/Tax

The game sales data covers two periods from 1 December 2016 to 31 January 2017 and from 1 December 2017 to 31 January 2018 respectively.

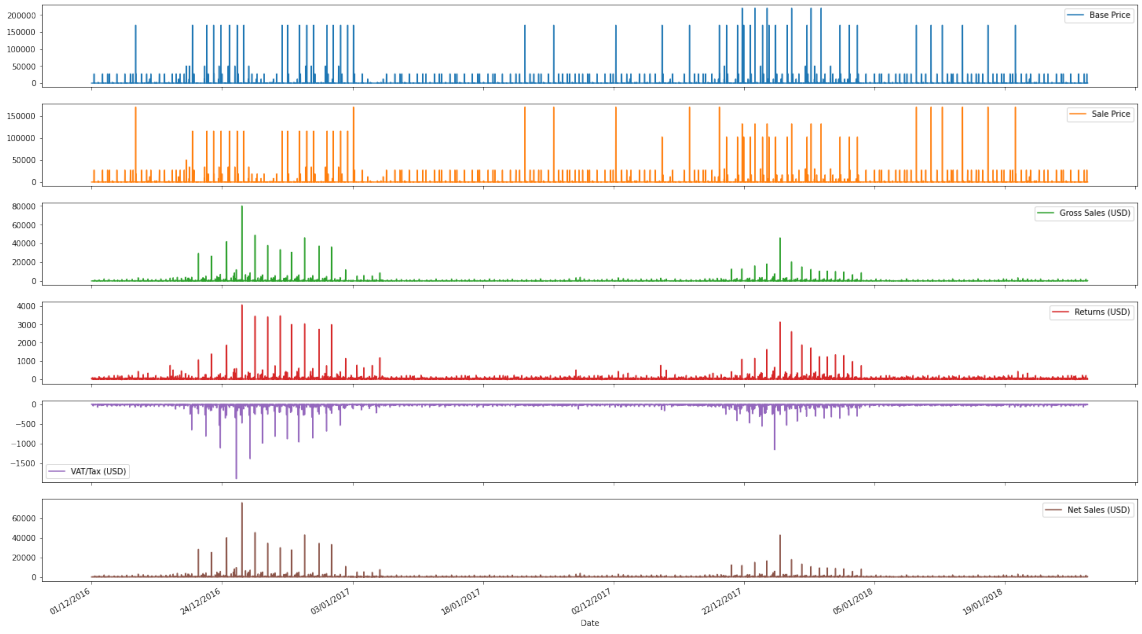
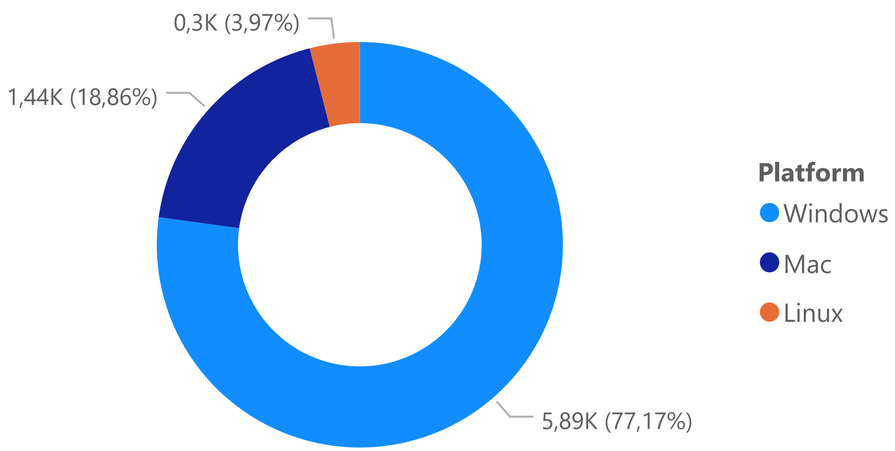


Figure 1: The distribution of Sales variables over days

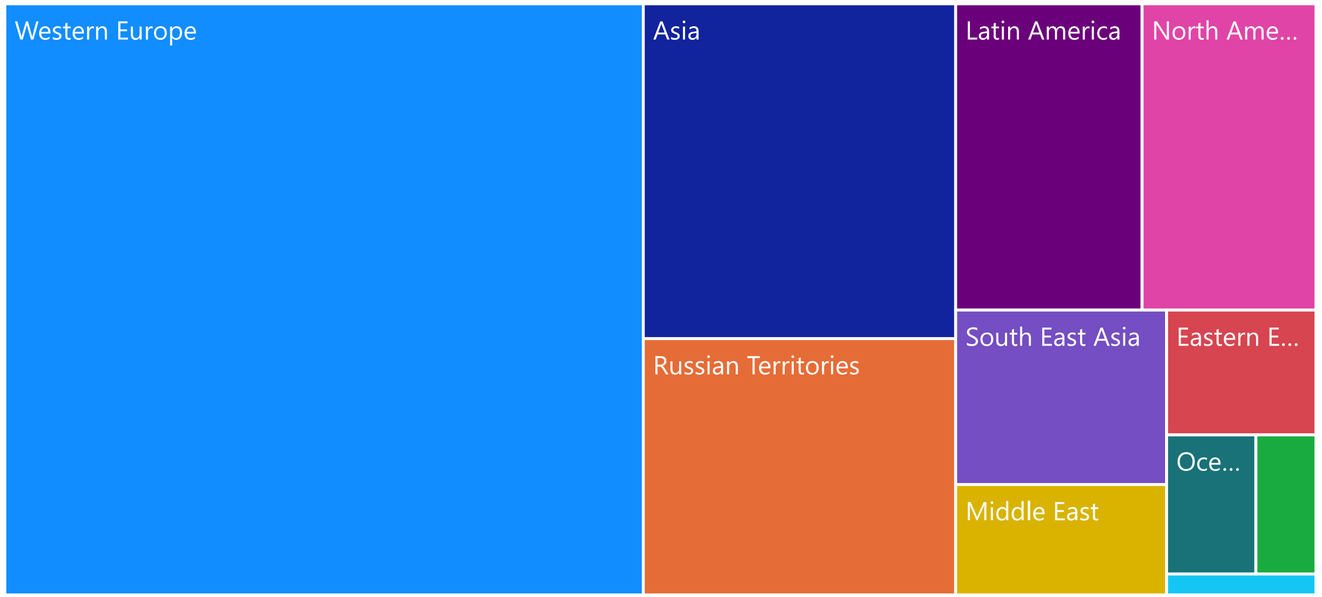
Exploratory Data Analysis

The Power BI dashboard illustrates the regions of activity as well as the Net Sales and Net Unit sold by region. We detect that the product is mostly sold is North America, however, the Net Sales (USD) is more generated in Western Europe.
Moreover, we plot the evolution of Gross Sales (USD) et Net Sales (USD) over year to determine how each value is trending over a period of time. If both lines increase together, this could indicate trouble with product quality because costs are also increasing, but it may also be an indication of a higher volume of discounts. We observe a similarity in the two time series. Therefore, we will look more closely to Net Sales (USD).
Our data rows are sale transactions and scattered per OS system or country. Consequently,it makes sense to aggregate our data by date and we sum up. This is to say, every row has a unique date.

Nombre de Platform par Platform



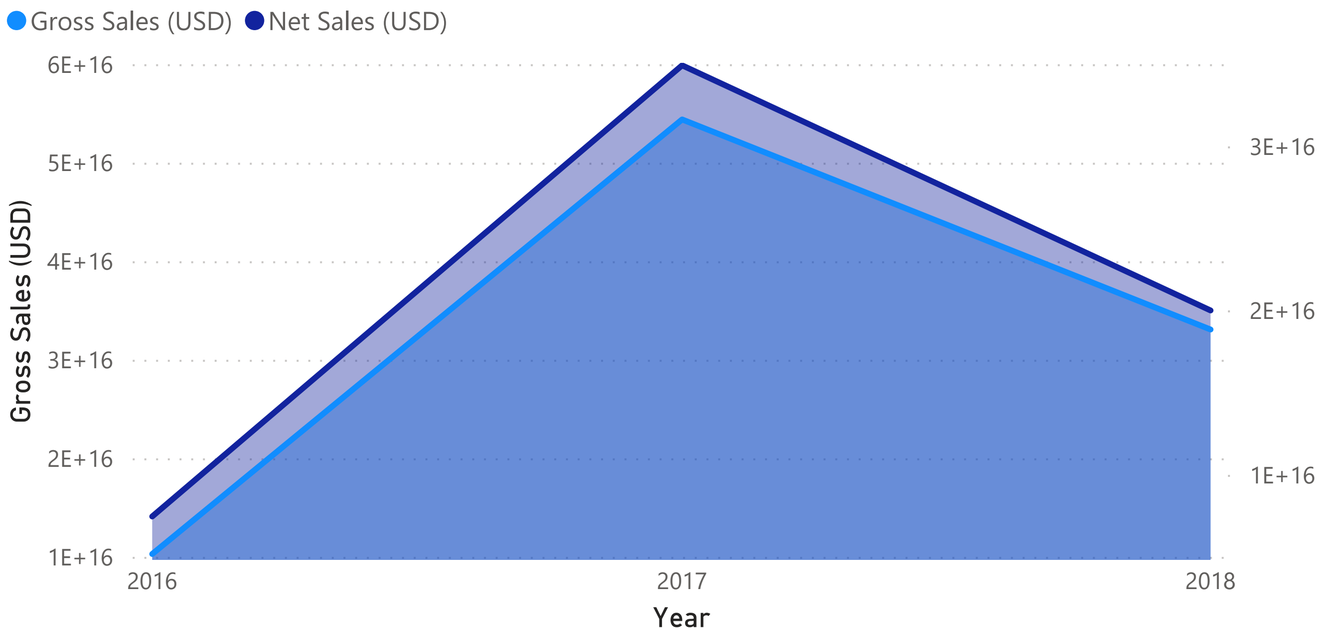
Net Sales (USD) par Region



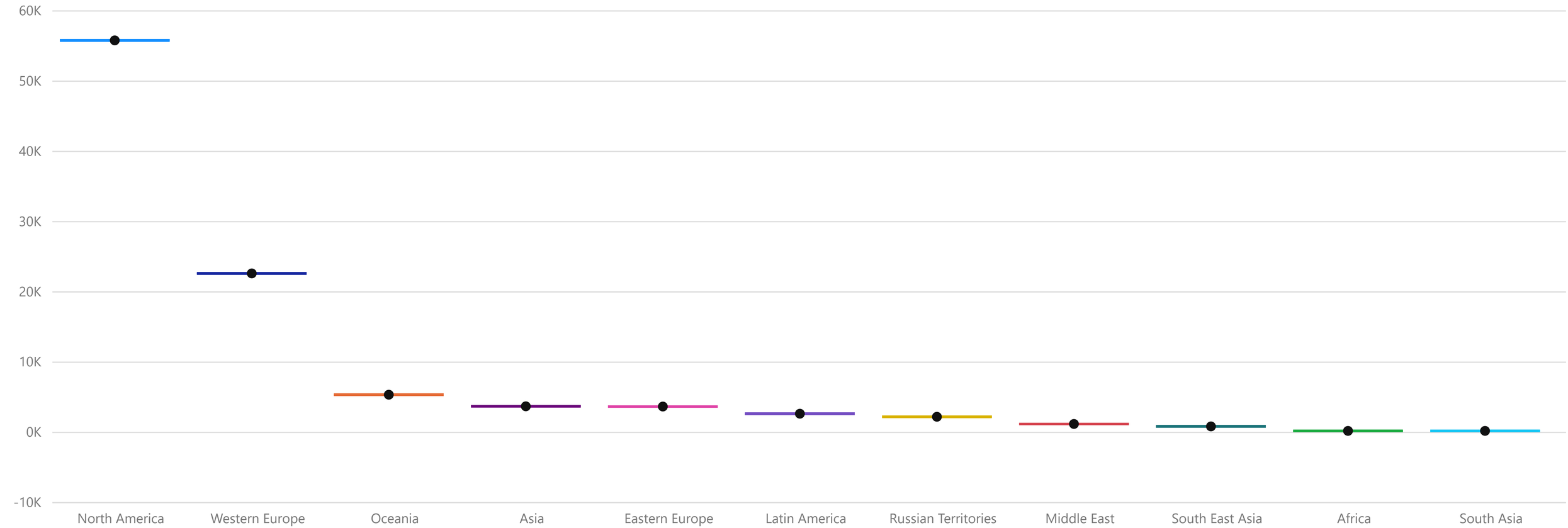
34M

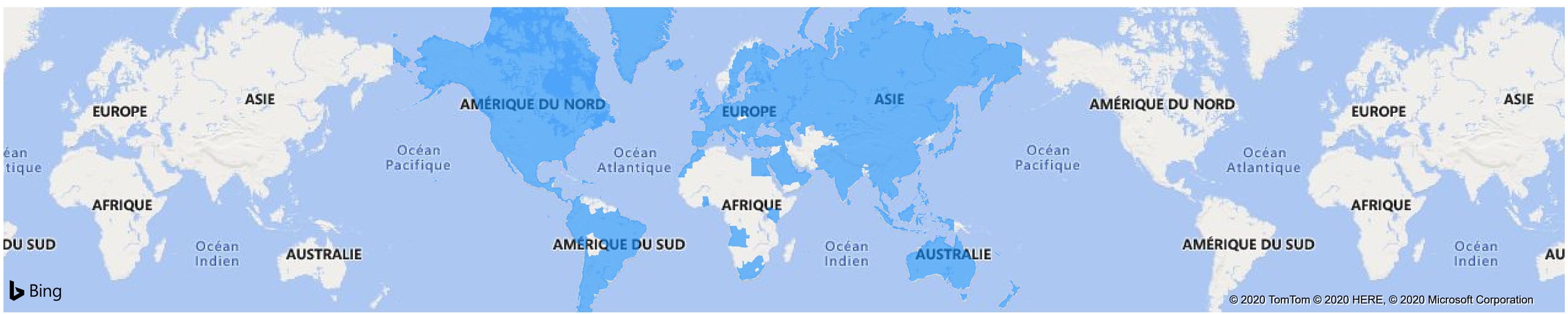
Sale Price

Gross Sales (USD) et Net Sales (USD) par Year



Net Units Sold par Region





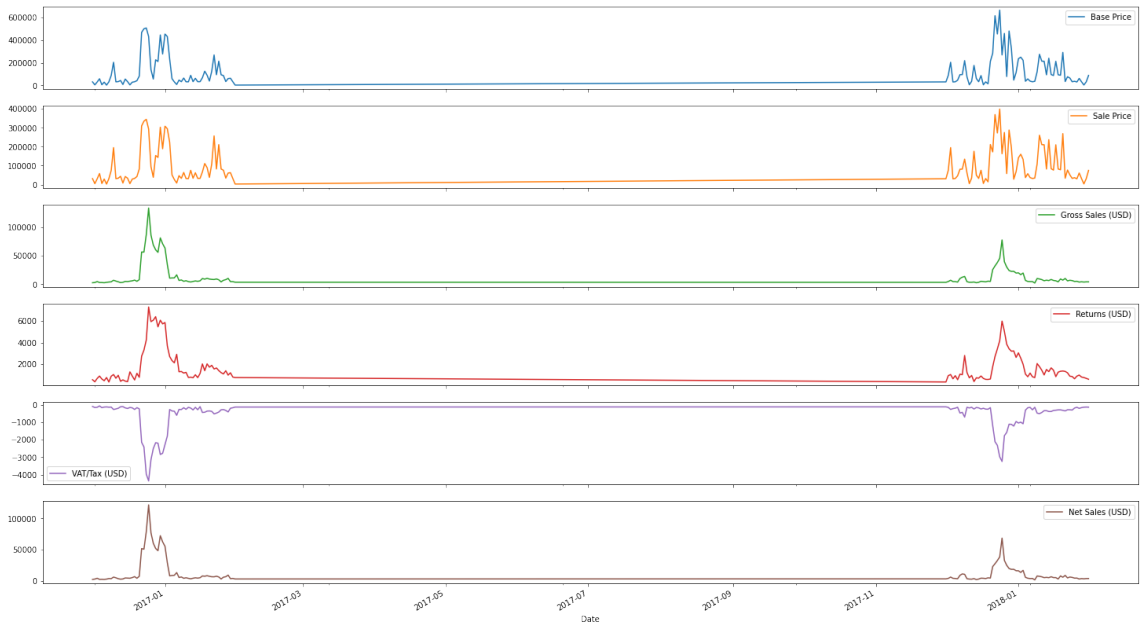


Figure 2: The distribution of our variables aggregated by Date

Univariate Data Analysis

The line charts for our numerical variables shows two periods with peaks around 25th December for Gross Sales, Net Sales and Returns. We can assume it is an e-commerce platform like Amazon selling in Christmas. To take a closer look, shall we see some basic statistics of our variables in Figure3. For instance, he variance (std) is high for the Sale Price and Net Sales and low for Net Unit Sold.

	Package	Product(ID#)	Gross Units Sold	Returns	Net Units Sold	\
count	7636.0	7636.0	7636.000000	7636.000000	7636.000000	
mean	-1.0	91203.0	14.604112	1.689890	12.914222	
std	0.0	0.0	111.726437	8.487162	103.923956	
min	-1.0	91203.0	0.000000	-2.000000	-45.000000	
25%	-1.0	91203.0	2.000000	0.000000	2.000000	
50%	-1.0	91203.0	2.000000	0.000000	2.000000	
75%	-1.0	91203.0	5.000000	2.000000	5.000000	
max	-1.0	91203.0	4679.000000	239.000000	4440.000000	

	Base Price	Sale Price	Gross Sales (USD)	Returns (USD)	\
count	7636.000000	7636.000000	7636.000000	7636.000000	
mean	2100.259810	1627.116960	248.323921	28.240493	
std	14833.274549	11133.476778	1865.259902	140.542828	
min	4.250000	2.550000	0.000000	-39.980000	
25%	22.990000	16.990000	25.028450	0.000000	
50%	24.990000	22.990000	48.800000	0.000000	
75%	78.000000	78.000000	89.950000	31.200000	
max	220000.000000	169999.000000	79984.489271	4074.562887	

	VAT/Tax (USD)	Net Sales (USD)
count	7636.000000	7636.000000
mean	-10.623679	209.459749
std	52.450715	1717.895432
min	-1902.319284	-764.550000
25%	-8.116400	21.210000
50%	0.000000	39.953400
75%	0.000000	74.970000
max	17.584700	75435.600000

Figure 3: Summary Statistics

In the fist period, as the Christmas season approaches the Sale Price variate from 22.99(USD) to 16.99(USD). As a result, the discount percentage is 26 %. In the second period, the sale price is 14.99 and the original price is 22.99 (the most frequent in the period prior to Christmas since the price is highly variable)so the discount is 35%. Looking attentively to the Base Price in Figure1 in the second period, we reveal a rise unlike the first period. If we extra fees like shipping charges and extra equipment or supply options to the Base Price we obtain the Sale Price. When setting the Sale Price we study the target customer and the country's price of market. That is why Sale Price is flexible but it is abnormal to find it inferior to the Base Price specially in the second period. Besides, to see the variability of Net Sales(USD) over the weeks we draw those boxplots in Figure4. Clearly, we detect a rise around the 4th week (last week of December 2016 and 2017) and seasonality in this time series.

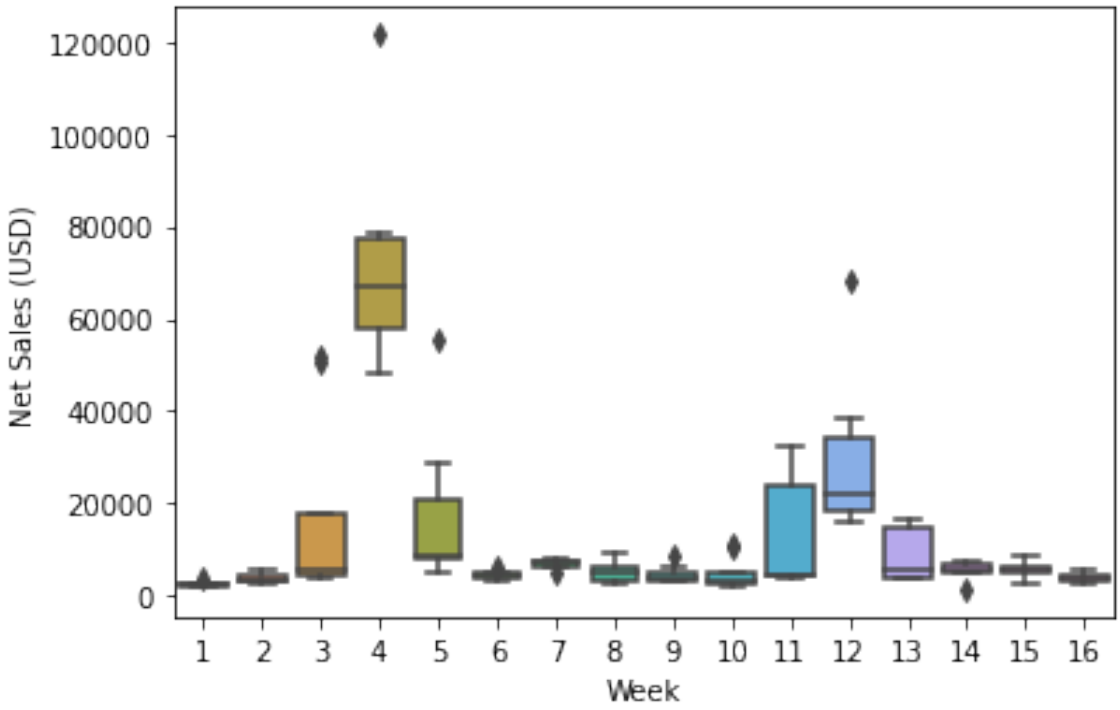


Figure 4: Boxplot of Net Sales(USD) over Weeks

Bivariate Data Analysis

In this section, we are studying the relationship between our variables. Thus, we plot them simultaneously in Figure4.

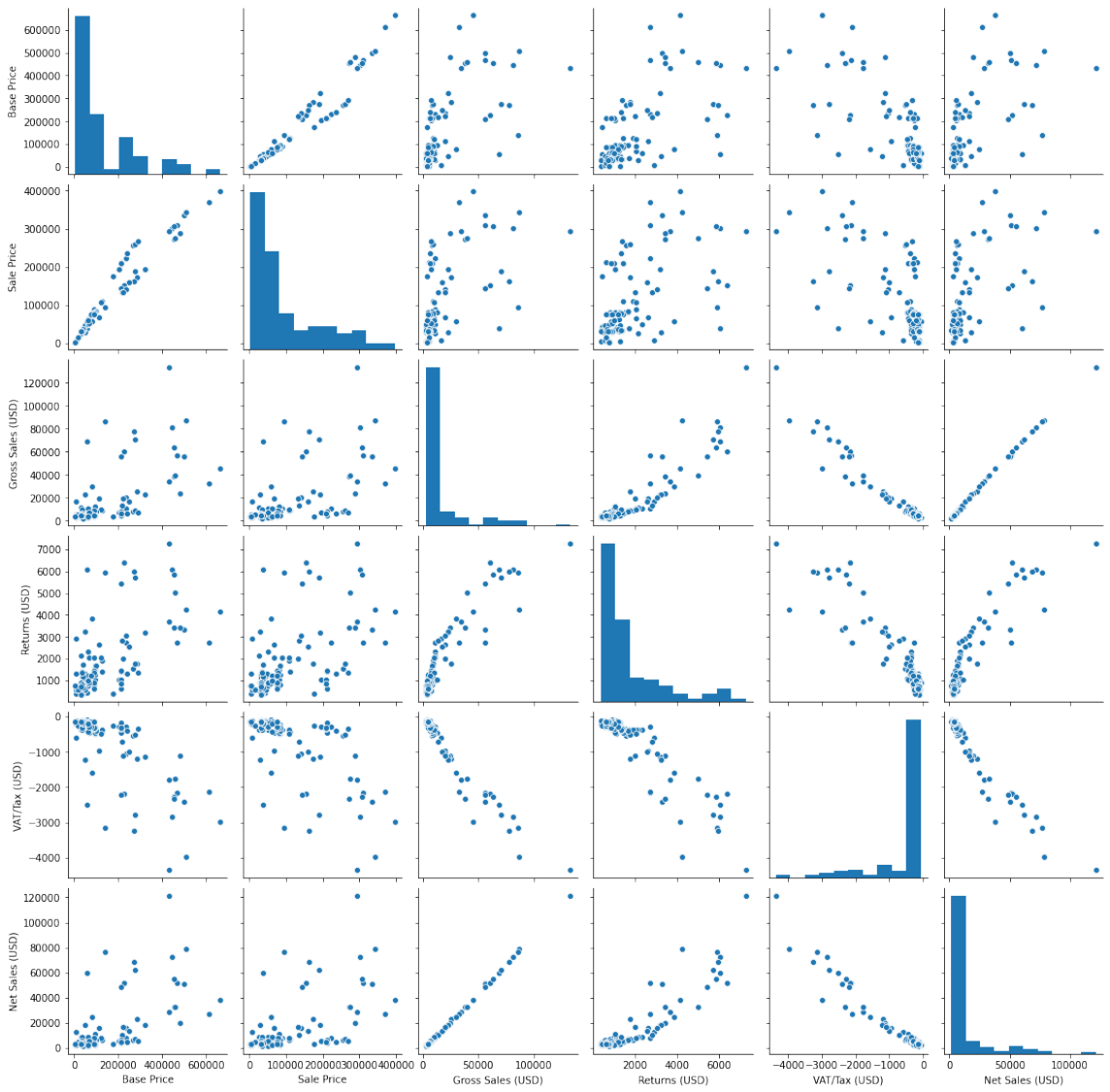


Figure 5: Bivariate Analysis

There is a linear relationship between Net Sales, Gross Sales, Returns and VAT/tax. We will focus on Net Sales which is basically a time series.

Time series Analysis

In this section, we will examine the Net Sales (USD) time series aggregated by date. Overall, We use the Box-Jenkins the methodology and we check:

1. The Stationary and Seasonal aspect: This includes stationarizing and Seasonal differencing.
2. Identification and finding the best fit of the model for the Autoregressif AR and Moving Average MA component.
3. Estimation for many models and selecting the best one.
4. Validation of the model after testing its parameters and residuals

After that, the time series is ready for forcasting and at this stage we can smooth and forecast the data with Simple Exponential Smoothing SES method.

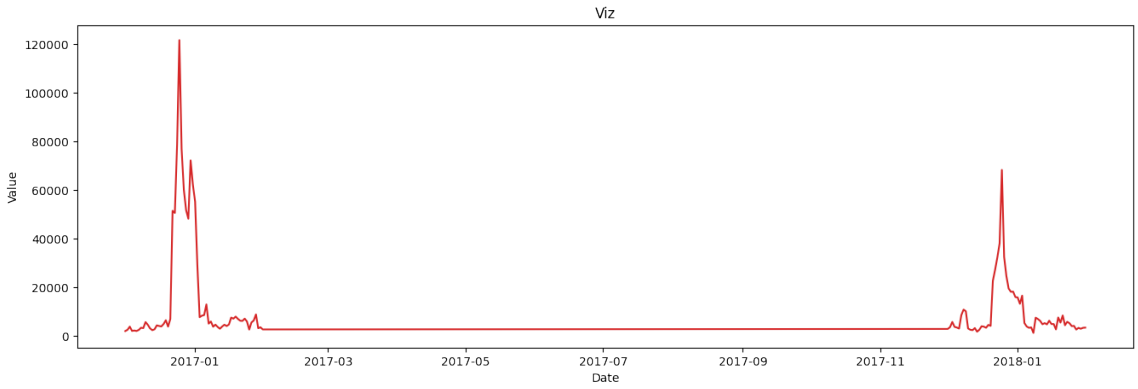


Figure 6: Net Sales (USD) aggregated by Date

Stationarity A stationary time series is one whose statistical properties such as mean, variance, and autocovariance are all constant and not a function of time.

$$E(X_t) = m \qquad \forall t \tag{1}$$

$$cov(X_t, X_{t+h}) = \gamma_h \qquad \forall t \tag{2}$$

To test the two equation, I calculate the mean of Net Sales (USD) for each day and the covariance .Then, I plot the evolution in Figure7 and Fifure8. The line is not constant which indicates that this time series is not stationary. The covariance is a also a function of time.

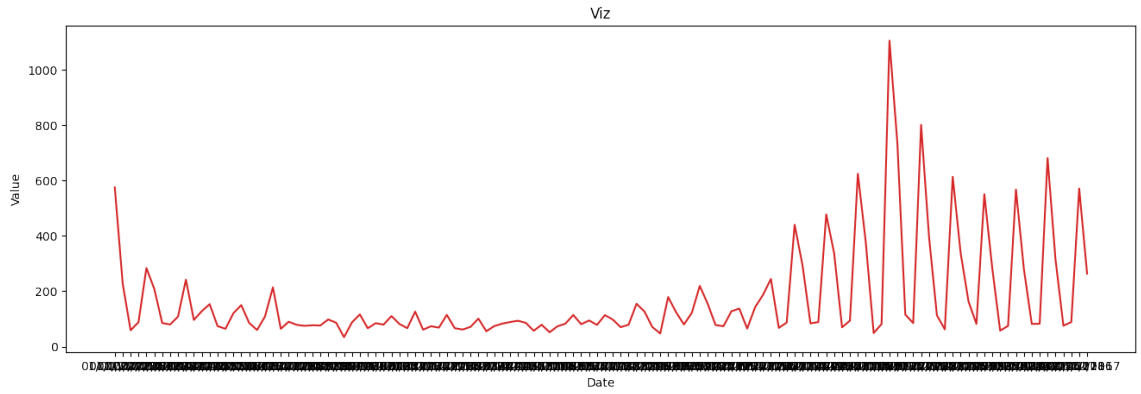


Figure 7: The mean of Net Sales(USD) per day over time

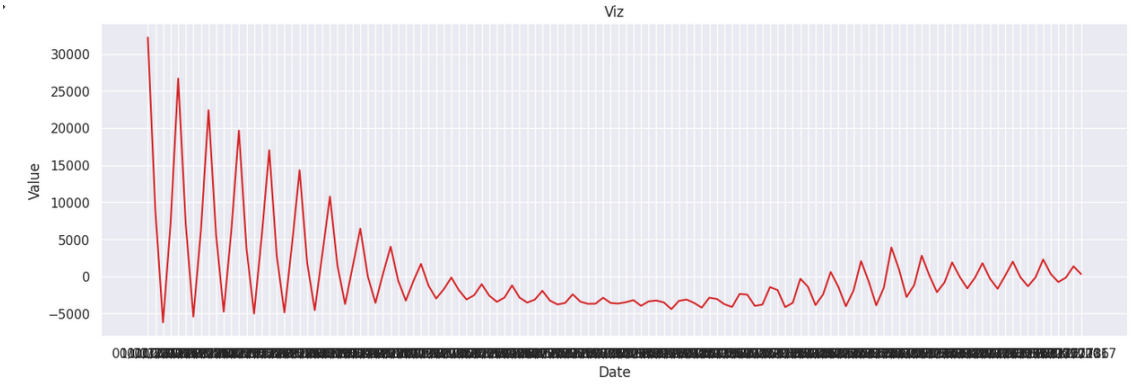


Figure 8: The covariance of Net Sales(USD) per day over time

It is not enough to verify empirically the stationarity, we need to test our hypothesis with Augmented Dickey Fuller ADF test and KPSS test. Our time series is not stationary since the p-value in ADF is higher than 0.05. However, it is trend stationary accordingly to KPSS test.

```
Results of Dickey-Fuller Test:
Null Hypothesis: Unit Root Present
Test Statistic < Critical Value => Reject Null
P-Value =< Alpha(.05) => Reject Null

Test Statistic                  -1.302951
p-value                          0.627821
#Lags Used                       5.000000
Number of Observations Used    118.000000
Critical Value 1%               -3.487022
Critical Value 5%               -2.886363
Critical Value 10%              -2.580009
dtype: float64
```

Figure 9: Augmented Dickey Fuller test on Net Sales (USD) time series

```
Results of KPSS Test:
Null Hypothesis: Data is Stationary/Trend Stationary
Test Statistic > Critical Value => Reject Null
P-Value =< Alpha(.05) => Reject Null

Test Statistic           0.565845
p-value                  0.026837
Lags Used               13.000000
Critical Value 10%       0.347000
Critical Value 5%        0.463000
Critical Value 2.5%      0.574000
Critical Value 1%        0.739000
dtype: float64
```

Figure 10: KPSS test on Net Sales (USD) time series

Let's get more insights from the Autocorrelation Function and Partial Autocorrelation PACF which can tell us about the Autoregressif AR and Moving average MA level. The MA models are stationary by default.As we can see, the ACF is oscillating between negative and positive values and it doesn't reach a plateau in zero. This tell us that there is seasonality in the time series.(see Figure 11)

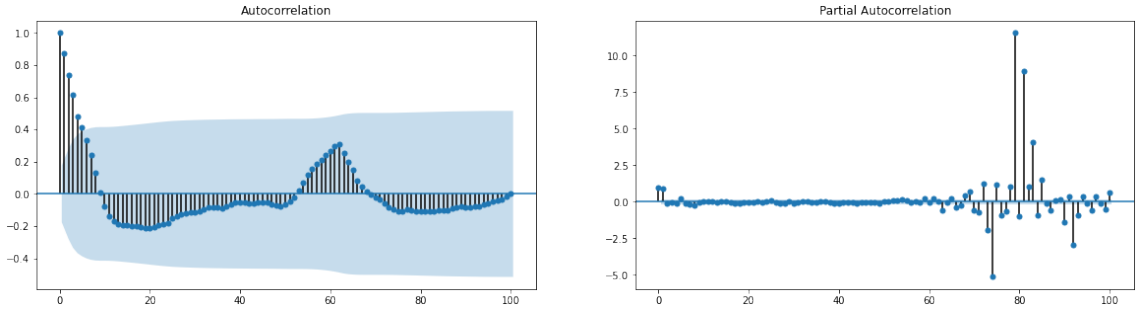


Figure 11: Net Sales (USD) Autocorrelation and Partial Autocorrelation Function

Differencing In order to make our time series stationary we need to differentiate it. Differencing is subtracting an observation from an observation at the previous time step. Differencing generates a time series of the changes between raw data points and helps us create a time series that is stationary. Normally, the correct amount of differencing is the lowest order of differencing that yields a time series which fluctuates around a well-defined mean value and whose autocorrelation function (ACF) plot decays fairly rapidly to zero. After each differencing operation, like we perform below, we can conduct an Augmented Dickey-Fuller (adf) and Kwiatkowski-Phillips-Schmidt-Shin (kpss) test to check for stationarity. Now, let's see the resulted time series

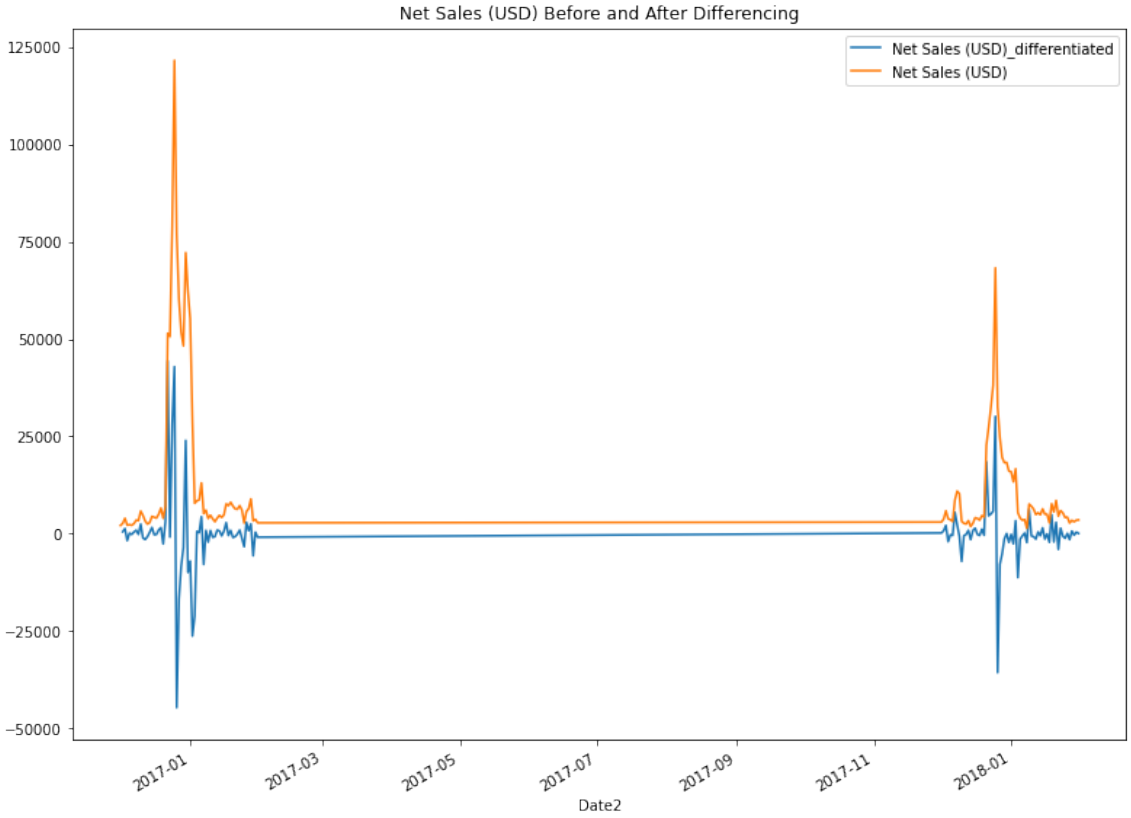


Figure 12: differentiated Net Sales (USD) and the original one

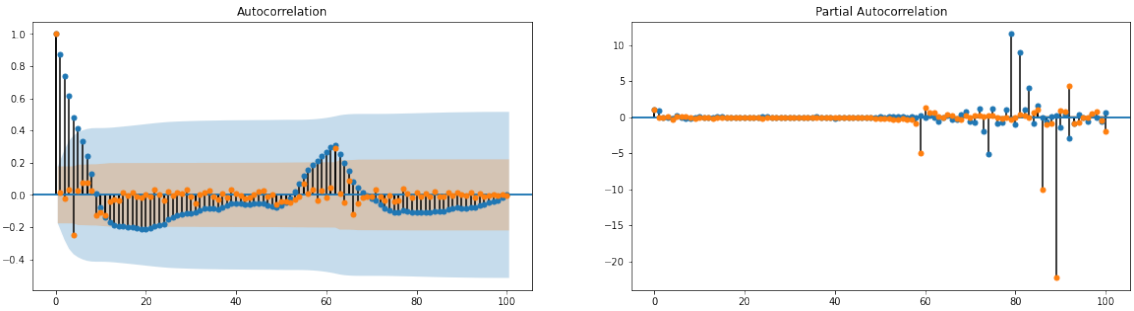


Figure 13: Net Sales (USD) Autocorrelation and Partial Autocorrelation Function for Net Sales (USD) and differentiated time series

Seasonality and Patterns We can think about our time series as composed of a combination of level, trend, seasonality, and noise.

Level: The average value in the series.
Trend: The increasing or decreasing value in the series.
Seasonality: The repeating short-term cycle in the series.
Noise: The random variation in the series.

To do this,there is two models Additive and. Multiplicative
y(t) = Level + Trend + Seasonality + Noise
y(t) = Level * Trend * Seasonality * Noise

In a multiplicative time series, the decomposed components multiply together to make the time series. In a multiplicative series there is increasing trend, the amplitude of seasonal activity increases and everything becomes more exaggerated. Multiplicative trend looks more like an exponential curve and multiplicative seasonality has waves that grow in amplitude of the course of time.
In an additive model we assume the components of the time series have an additive effect, that the amplitude of the seasonal effect is roughly the same, that the size of the residuals are mostly constant. Therefore, our model is additive.
Also, HPF filter method allow us to distinguish between components.(see Figure 16)

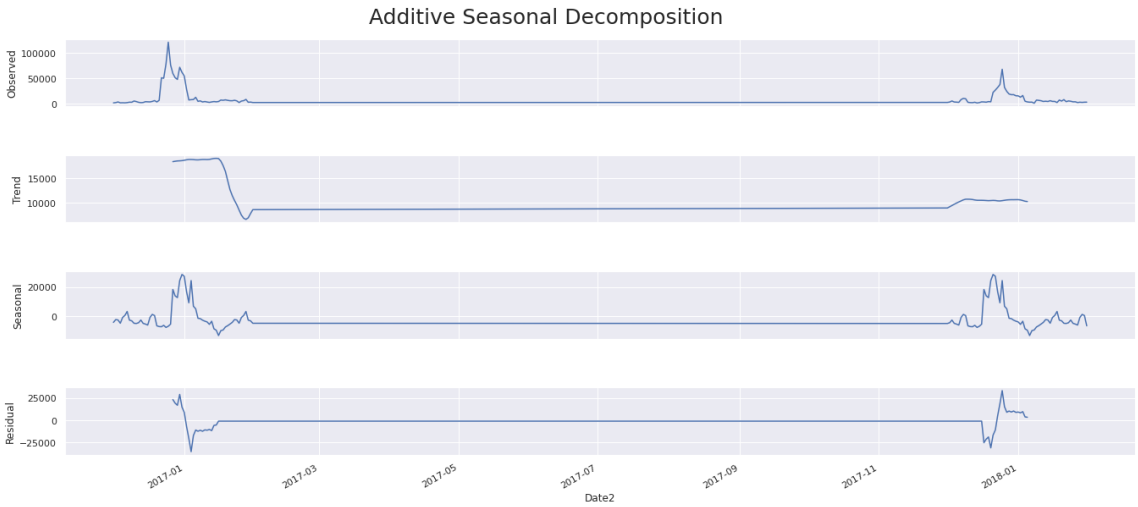


Figure 14: Additive decomposition

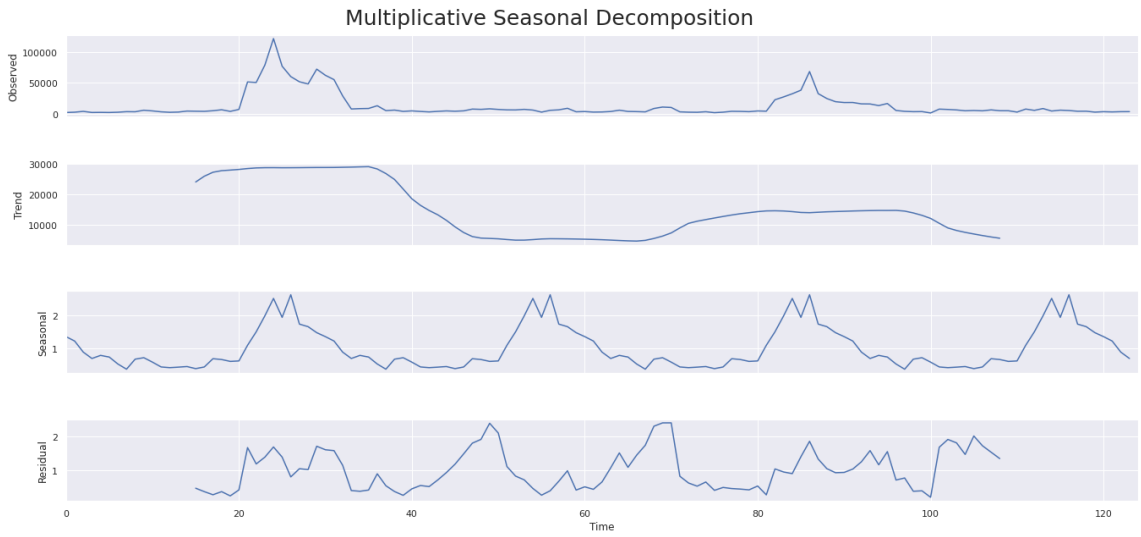


Figure 15: Multiplicative decomposition

Also, HPF filter method allow us to distinguish between components.

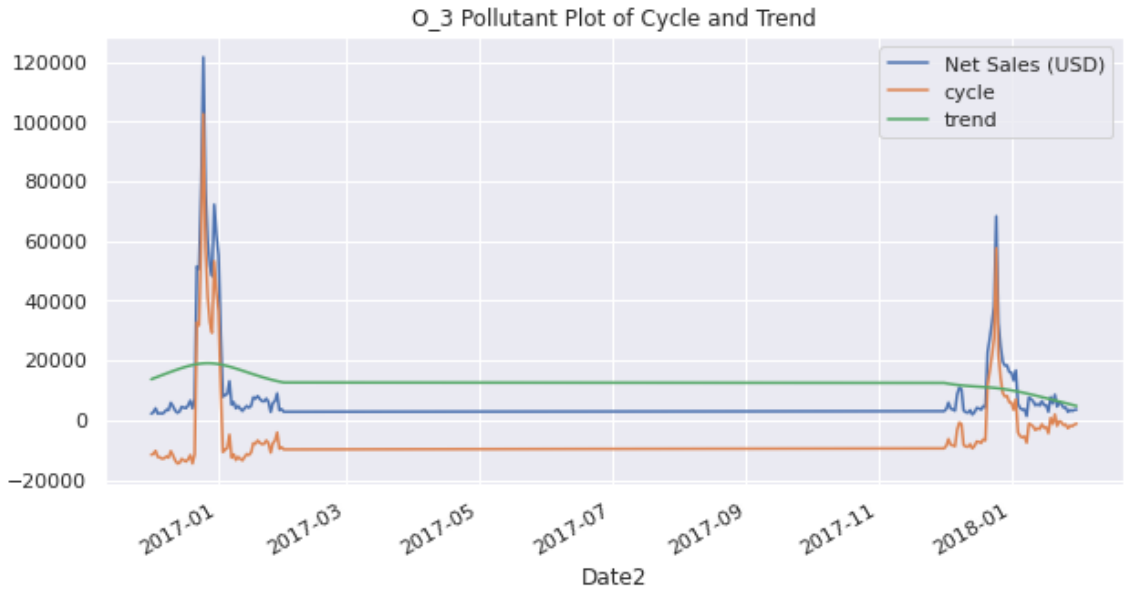


Figure 16: HPF filter decomposition

Difference between the two periods

```
[ ]  Recrutiment_period1['Net Sales (USD)'].describe()
```

```
count        62.000000
mean      16283.804401
std       25211.635139
min        2128.166400
25%        3728.759254
50%        5429.216781
75%        8349.200128
max      121579.466980
Name: Net Sales (USD), dtype: float64
```

```
[ ]  Recrutiment_period2['Net Sales (USD)'].describe()
```

```
count       61.000000
mean      9620.309584
std      11351.497772
min       1370.420925
25%       3545.211180
50%       4991.629248
75%      10279.297530
max      68288.562491
Name: Net Sales (USD), dtype: float64
```

Figure 17: Difference between the two periods

The second period has lower income (9620.3 USD as mean) than the first period (16283.8 USD)

Quantifying the similarity between the two periods For that, I calculated the cross-correlation 0.7283168644267012 and Distance Time Warping metric 75197.48319143995,

Business Performance

To define KPIs for Business Performance As we can see in Figure2 the VAT/Tax is lower in the second period than the first one.
A value-added tax (VAT) is a consumption tax placed on a product whenever value is added at each stage of the supply chain, from production to the point of sale. The amount of VAT that the user pays is on the cost of the product, less any of the costs of materials used in the product that have already been taxed. More than 160 countries around the world use value-added taxation, and it is most commonly found in the European Union. Moreover, European countries sold more in the first period. This can be a factor for the drop in sales in the second period.
Targeting ,Positioning and Segmenting: this a Marketing strategy is more effective is the first period where the European countries are targeted. Also, in the second period the Base Price is sometimes lower than the Sale Price which indicates non efficiency.
For THE 4 P's marketing: Price, Product, Place and Promotion, The price is basically the same but the variance of the Sale Price is higher is the second period.

Other Information

If I were in charge of this project in a real-world situation, I would ask the data of competitors in the market, and the Promotion done by this e-commerce platform in this particular occasion which is Christmas. May be our unique product is Christmas tree.

Figure 18: Christmas tree