BaseTech

May 10, 2022

# 2019 SALES ANALYSIS

PROJET 8 : COMMUNIQUEZ VOS RESULTATS

PRESENTED TO

**OpenClassrooms**

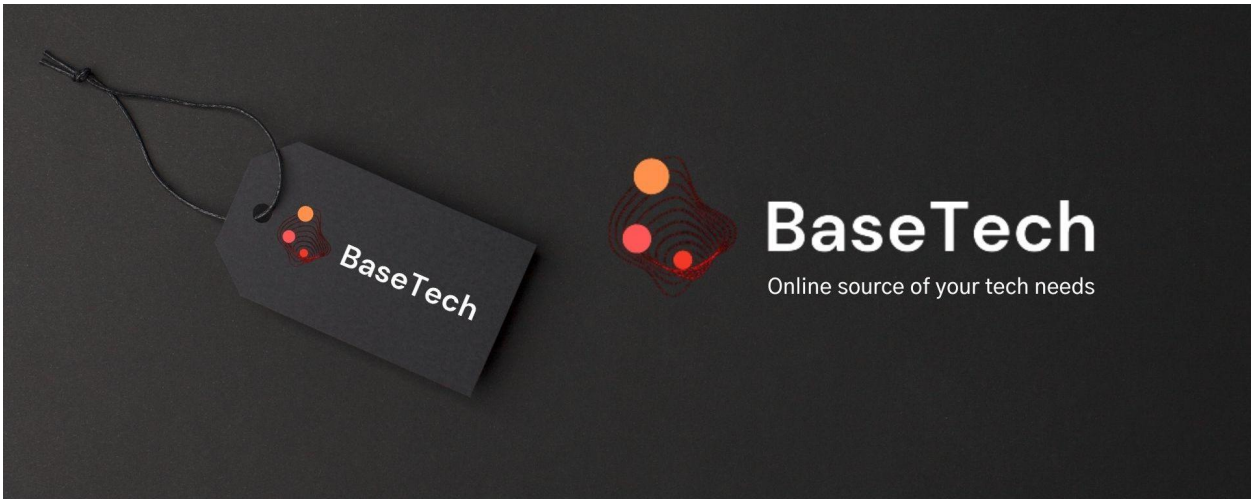
PRESENTED BY

**Katrina JUMADIAO**

# TABLE OF CONTENTS

# 2019
# SALES ANALYSIS
# BaseTech

## Introduction




BaseTech is a recent start-up American electronics company in the e-commerce industry. Launched in 2018, BaseTech is located in San Francisco, California. Although not yet diverse enough in terms of product offerings, the revenues BaseTech generated in 2019 showed growth and positive outcome overall.

This report will analyze BaseTech's sales revenues in 2019 to evaluate its performance and plot a course for future development. The data used in this analysis is from the stored and collected records from online purchases made on BaseTech's website throughout the year. The sales data is a compilation of 12 months' sales revenue, consisting of 186 850 entries, with 6 different variables identifying each order information.

An important mention is that BaseTech did not have official records on their first year in 2018, hence the lack of data to provide better sales comparison and forecasting for the upcoming year.

There are three main sections in this report:

- Processing and data manipulation for handling errors and standardization of information. Once the cleaning process has been completed and validated its accuracy and reliability, it will then be used for sales analytics moving forward.

- Detailed report of findings from the analysis conducted using a programming language and other external sources. Python is predominantly used to explore and respond to business queries with 12 months' worth of sales data.

- Present different methods of estimating future revenues, the model chosen to forecast BaseTech's sales for an upcoming period, and the challenges faced during the predictive analytics.

# 1. DATA PROCESSING

The file directory "Sales_Data" contains 12 CSV files representing each month's generated sales. The "sales_2019" data frame is a concatenation of all the records from the file directory, summing up 186 850 entries and 6 columns which can be classified into 2 main groups:

| Order information | Product information |
|---|---|
| Order ID | Product name |
| Date of purchase | Price per unit |
| Shipping address | Quantity of unit(s) sold |

The table below is an overview of the evolution of the dataset in the process:

| Step | Description | Number of entry | Number of variable |
|---|---|---|---|
| 0 | Raw data | 186 850 | 6 |
| 1 | Rename columns | 186 850 | 6 |
| 2 | Missing values | 186 305 | 6 |
| 3 | Duplicates | 185 687 | 6 |
| 4 | Outliers | 185 686 | 6 |
| 5 | Data types | 185 686 | 6 |
| 6 | Data enhancement | 185 686 | 15 |

## 1.1. Data cleaning

Data cleansing is one of the most critical processes in data analysis. This process may involve data enhancement and manipulation to ensure that it is free of irrelevant, incorrect, incomplete, and corrupted information.

### 1.1.1. Rename columns

Renaming columns in a data frame is essential to facilitate data manipulation. Words separated by underscores are necessary to enhance readability. When manipulating data with Python, it is easier to reference back to a variable without always using brackets and quotation marks. Almost all columns retained their initial names but are now written in lowercase with an underscore separating the words.

| Original columns | New columns |
|---|---|
| Order ID | order_id |
| Product | product |
| Quantity Ordered | order_qty |
| Price Each | unit_price |
| Order Date | order_date |

| | |
|---|---|
| Purchase Address | purchase_address |

## 1.1.2. Missing values

Missing values result in data due to errors during data collection or the absence of a response from respondents. 2 options are available when handling missing values: removing the observations from the data set and/or imputing new values based on other observations.
Once filtered out missing inputs, findings show 3 270 missing entries (545 rows x 6 columns), accounting for 1.75% of the data. Since these records do not have any information, inputting missing values is impossible. Thus, dropping those values is the only valid option.

## 1.1.3. Duplicates

Duplicates are copies in a dataset with more than one occurrence. This typically happens when a dataset is collected from multiple sources. In some other cases, duplicates arise when inputs are submitted more than once in a survey or errors during data entry.
After removing missing values, duplicated entries went from 1,162 to 618, accounting for merely 0.33% of the data to be removed. The data frame will have 185 687 entries in total at this stage.

## 1.1.4. Outliers

When a data point falls far outside the norm, it is considered an outlier. Outliers are detected when some data points are remarkably distinct from the other values.
They may skew an analysis too far in a certain direction. Outliers are misleading because they are the same type as other values. Outliers may result in biased results during the analysis, but removing them should have a valid reason.

This entry below does not have any valuable information to consider preserving in the dataset.

| order_id | product | order_qty | unit_price | order_date | purchase_address |
|---|---|---|---|---|---|
| Order ID | Product | Quantity Ordered | Price Each | Order Date | Purchase Address |

## 1.1.5. Data types

All data types were considered "objects" (strings) when loading the dataset. Hence, "order_qty" and "unit_price" should be regarded as numerical: integers (whole number) and float (decimal points), respectively, whereas "order_date" should be specified as date-time.

| Variable | Original data type | Converted data type |
|---|---|---|
| order_qty | object | int64 |
| unit_price | object | float64 |
| order_date | object | datetime64[ns] |

# 1.2. Data enhancement

Typically, a sales analysis consists primarily of quantitative data, such as key performance indicators (KPIs) and charts. In this case, adding new variables is necessary to perform structured analysis and carry out valuable data insights. The data will be enhanced by adding multiple columns that can be segmented into 4 main types: date, address, category, and sales.

## 1.2.1. Date

To identify customer behavior or statistics in a certain period, adding variables such as quarter, month, date, day of the week, and hour is needed. These mentioned columns were extracted from the variable "order_date."

## 1.2.2. Address

To determine sales generated from a precise geographic location, having variables such as city and state is necessary. This is done by extracting city and state from the address in the data.
One thing to note is the importance of including either the state or the zip code of a city. The United States has multiple cities with the same name but entirely different states. One example found in the data is Portland. When grouping the data into cities, Portland shows 2 counts of state. And once checked, Portland belongs to 2 different states, one from Maine and the other from Oregon. Adding a suffix name generally suffices to distinguish them better.

| | | | purchase_address |
|---|---|---|---|
| city | state | zipcode | |
| Atlanta | GA | 30301 | 12334 |
| Austin | TX | 73301 | 8609 |
| Boston | MA | 02215 | 15706 |
| Dallas | TX | 75001 | 12321 |
| Los Angeles | CA | 90001 | 21450 |
| New York City | NY | 10001 | 18807 |
| Portland | ME | 04101 | 2301 |
| | OR | 97035 | 8723 |
| San Francisco | CA | 94016 | 28324 |
| Seattle | WA | 98101 | 12212 |

## 1.2.3. Product

Grouping products based on category can help identify valuable perceptions of customers, emerging trends, and information about competitors and their marketing activities. It is conducted to position products and promotes them more effectively. In other words, product category analysis provides in-depth insights to help draw conclusions that will help set up a brand's key issues.

The objective of a product category analysis is to:

- Generate overall awareness about a particular category
- Drive higher growth for the specific category

In this case, a new variable "category" has been created to help implement strategies like the queries above.

## 1.2.4. Sales

Adding the "sales" column is one of the most important variables for this analysis. This is obtained by multiplying the "unit_price" of a product by the "order_qty."

# 2. DATA SUMMARY

This segment shows a brief overview of BaseTech's product offering components and online transactions made in 2019.

## 2.1. Price

BaseTech offers 17 distinct prices for its products

- Prices ranging from $2.99 up to $1 700
- 25% of the price offer falls under $14.95 and below
- 50% of the price offer falls under $150 and below
- 75% of the price offer falls under $400 and below

## 2.2. Order ID

- Customer with the most spending in a single purchase is order_id = 181069 (2 products for $3 779.99)
- Customer with the most diverse product purchase is order_id = 160873 (5 products for $1 476.94)

## 2.3. Product

BaseTech currently sells 19 products from 8 different categories:

| Appliance | TV | Monitor | Laptop | Smartphone | Headphone | Charging cable | Battery |
|---|---|---|---|---|---|---|---|
| LG washing machine | Flatscreen TV | 34” ultrawide monitor | MacBook Pro laptop | iPhone | Apple Airpods headphone | Lightning charging cable | AAA batteries (4-pack) |
| LG dryer | | 27” 4K gaming monitor | ThinkPad laptop | Google phone | Bose SoundSport headphone | USB-C charging cable | AA batteries (4-pack) |
| | | 27” FHD monitor | | Other smartphone | Wired headphone | | |
| | | 20” monitor | | | | | |

## 2.4. Order quantity

There are 185 686 sold products in total

- The average quantity per product ordered is 1
- The minimum quantity bought for the same product is 1
- The maximum quantity bought for the same product is 9 (worth $26.91 in total)

## 2.5. Transactions

There are 178 437 online purchases made in total

- The lowest basket amount sums up to $2.99
- The highest basket amount sums up to $3 779.99
- $11.95 and below represents 25% of the total price paid
- $14.95 and below represents 50% of the total price paid
- $150 and below represents 75% of the total price paid

## 2.6. Location

BaseTech reached 10 cities from 8 different U.S. states, as seen on the map in *Figure 1* and the table in *Figure 2* below.


*Figure 1. Basetech's consumers are located in 10 different cities in the U.S.*

| city | state |
|---|---|
| Atlanta | GA |
| Austin | TX |
| Boston | MA |
| Dallas | TX |
| Los Angeles | CA |
| New York City | NY |
| Portland | ME |
| | OR |
| San Francisco | CA |
| Seattle | WA |

*Figure 2. List of cities*

# 3. SALES ANALYSIS

*The tech company has experienced fluctuation in 2019.*

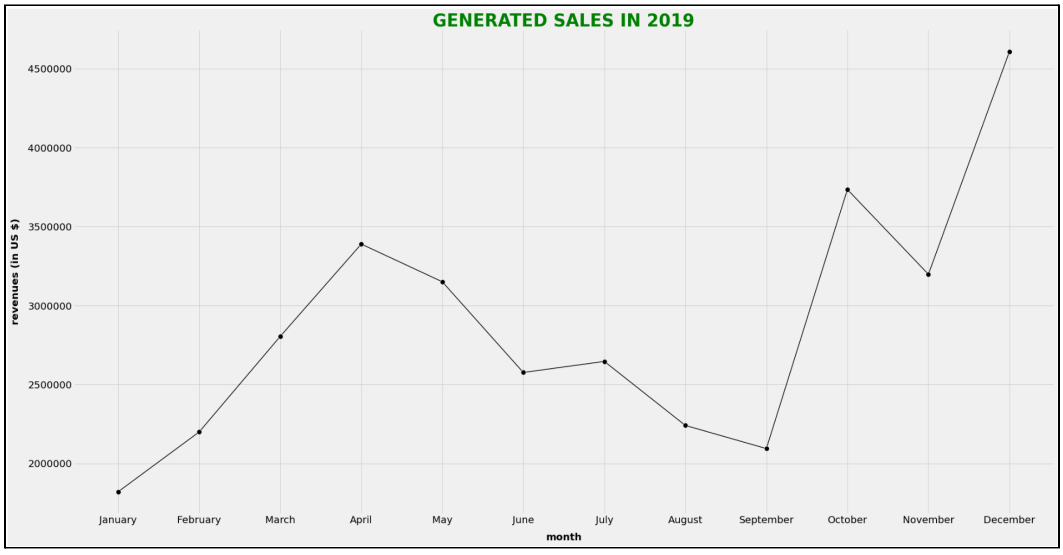



*Figure 3. BaseTech's sales revenues in 2019*

**KEY STAT:**

1. During the first quarter, sales made up 19.79% of total annual revenues
2. BaseTech shows an improvement of $2 298 320 million in the second quarter

3. BaseTech experienced a continuous drop in sales from $9 116 114 down to $6 982 010 million throughout the third quarter.
4. The tech company recovered erratically during the holiday sales, with a growth rate of 65.30%

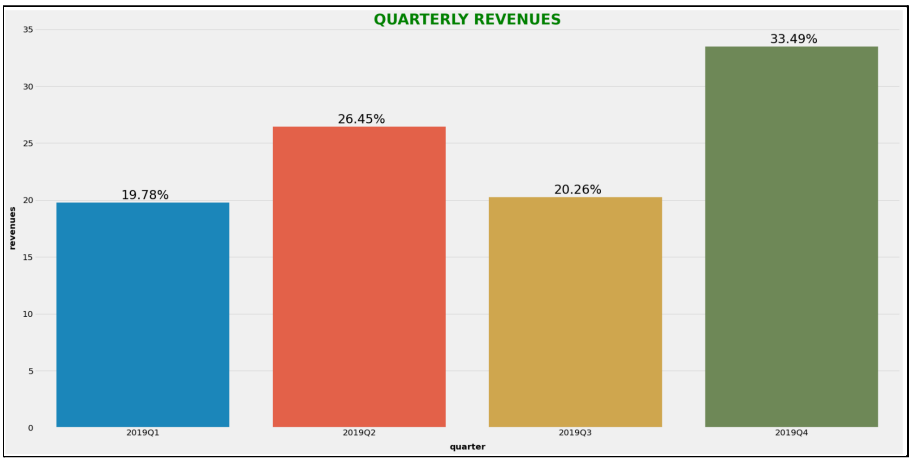



***Figure 4. BaseTech's quarterly sales revenues in 2019***

# 3.1. Quarterly sales performance

BaseTech's lack of data information may hinder the ability to compare various aspects of its sales cycle to help identify and provide in-depth insights into its business. Nonetheless, a 12-month worth of data still supplies a preview of how the tech company is performing over the year and would undoubtedly issue valuable findings beneficial to BaseTech in the decision-making process.

## 3.1.1. Q1 shows a steady rise for BaseTech

BaseTech indicates a continuous growth in sales during its first quarter, as shown in *Figure 3*, showing that it receives more online orders as days go by, even though its total revenues were the lowest (19.79%) in 2019.

Factors such as attractive deals during the seasonal sales may have encouraged the increase in demand for BaseTech, such as:

- Traditional period for post-Christmas discounting in January
- Valentine's Day and Superbowl LIII in February
- St. Patrick's Day in March

## 3.1.2. Q2 exhibits growing demand and good business health

BaseTech generated sales grew from $6 817 794 million to $9 116 114 million in the 2nd quarter (cf *Figure 4*), which is an increase of 33.71% ($2 298 320 million). Growth in demand may indicate an effective marketing strategy, such as social media presence and advertising improvements. Another element that could drive its boost in revenues is customer satisfaction, resulting in positive customer feedback through good reviews, which may have motivated new consumers to purchase from them. Increasing sales could also be due to good shipping and handling and better performance in the company's services, such as reliable customer service.

### 3.1.3. Q3 BaseTech faces a continuous decline in demand and revenues slippage

Following a gradual increase from the first quarter to the second quarter, sales dropped continuously during the third quarter (*cf Figure 4*). Unit sales substantially declined by $2 134 103 million, a considerable decrease of 23.41% from the previous quarter. Several reasons may have caused the drop in BaseTech's sales performance.

**E-Commerce Summer Slump**

Summer implies lower sales for most businesses. Gross sales can drop by 30% in July from high generating sales months like December. Shoppers spend less time on their devices while on vacation during the summer and make fewer purchases overall. And this is primarily true for e-commerce like BaseTech, where electronic devices are not as needed during the warmer months. Low sales in summer appear to be global in scope, then eventually recover back in September.

**Inventory shortage**

BaseTech may have experienced stock shortages. Supply shortages and improper stock management generally generate stock-outs. Another stock-out scenario is when BaseTech has stock left in their warehouse but not listed on their online shop. BaseTech can potentially lose about half of the intended purchases when a customer experiences a stock-out.

**Customer relation**

A consistent fall in sales for about 5 months can translate into BaseTech's lack of customer acquisition and retention. Factors that may have caused the drop in sales are lack of online presence, poor customer online purchase experience, unsatisfactory after-sale and customer service, and lack of engagement and loyalty.

**Online presence**

BaseTech may have a weak online presence and not enough traffic to their website. Not enough use of SEO may result in their website's poor position on search results pages.

**Online experience**

The quality of their website may also play a role. User experience and the interface are crucial, especially when running a solely online business like BaseTech. Not providing enough details on product descriptions and images or poor web navigation can easily deter consumers from buying, affecting customer acquisition and retention. Indeed, the first navigation or purchase experience is the most critical part of a customer's journey.
Another element to consider is payment gateways. BaseTech may haven't been able to deliver an easy checkout experience or lack online payment options. As for mobile shopping, BaseTech's website may not be optimized for mobile users.

There are also abandoned carts to take into account for. Based on recent online shopping statistics, additional costs like shipping, taxes, and fees result in 49% of cart abandonment. These costs dissuade shoppers from pursuing their intended purchase.

**After-sales**

Poor customer service results in clients' discontentment and negative experience, leading to decreased customer loyalty and increased churn rate. Some critical elements of poor customer service that BaseTech may have neglected over time are:

- Vague policies in returns, refunds, and guarantees
- Unreachable customer support
- Slow response time or getting put on hold for an extended period
- Directing customers to the website
- Overlooking buyer feedback
- Deliver inaccurate information

**Customer engagement and loyalty**

Another potential cause of the constant decrease in demand is the absence of repetitive purchases from existing customers. BaseTech may have overlooked the importance of building a connection with its existing clients, such as insufficient email marketing (newsletters), no rewards program, and no personalized recommendations.

### 3.1.4. Q4 BaseTech recovered from its loss

After a significant loss in sales from the last quarter, BaseTech exhibits remarkable improvement in its performance. Sales rocketed by $4 558 938 million, a sharp increase of 65.30%. The fourth quarter for BaseTech, accounting for 33.49% of the total sales, has been the most relieving and most successful period of the year.

The most significant online sales events occur before Halloween until December, when retailers often promote huge deals on their products. Consumers are most willing to spend, especially on electronic products, during Black Friday and Cyber Monday, where sales are the most attractive. So holiday shopping makes these the peak months for online retail, which must have been the biggest driver of BaseTech's push in sales.

***SUMMARY**: 2019 was a rather turbulent year for BaseTech, with a few fluctuations, and it must re-evaluate several areas to improve its performance. BaseTech displays an upward trend in sales revenues overall, starting from $1 821 413 million in January to $4 608 295 million by the end of the year. It generated a total of $34 465 537 million in 12 months.*

## 3.2. Key dates

### 3.2.1. Best performing month

The chart below in *Figure 5* shows that December had the highest sales, representing 13.37% of annual revenues, followed by October, which generated $3 734 777 million, a significant increase of 78.32% from the previous month. As mentioned in the Q4 sales analysis previously, this is due to the hike in consumer spending during the holiday seasons. As the third-best month, April is also one of the holiday periods where consumers expect promotions and are willing to spend their money on Easter sales, with sales revenue merely 1.01% lesser than in October.

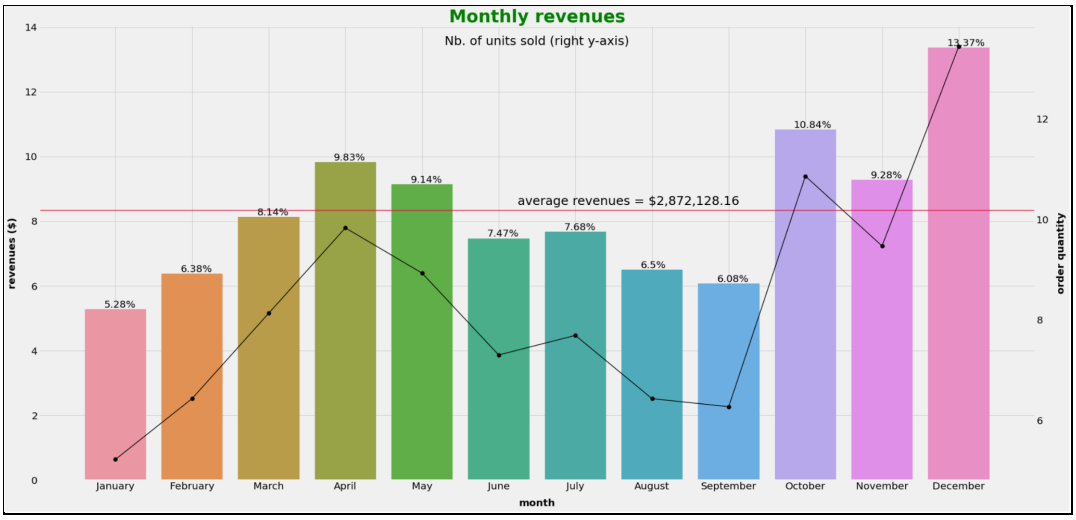



***Figure 5. BaseTech's monthly sales performance and number of units sold***

## 3.2.2. Monthly purchases

The number of generated revenues is reflected by the number of units sold, with December running up in the first place, October showing up in the second place, and April finishing in the third place.

### *January and mid-year slumps*

The months following the busy, strong periods (April and the last quarter of the year) are when brands expect a significant drop in sales. After numerous spendings during the holiday season, the following months are one of the lowest purchase months for an average U.S. consumer. The first and the third quarter usually bring low retail sales numbers, most commonly in business types like electronics.

## 3.2.3 Peak times

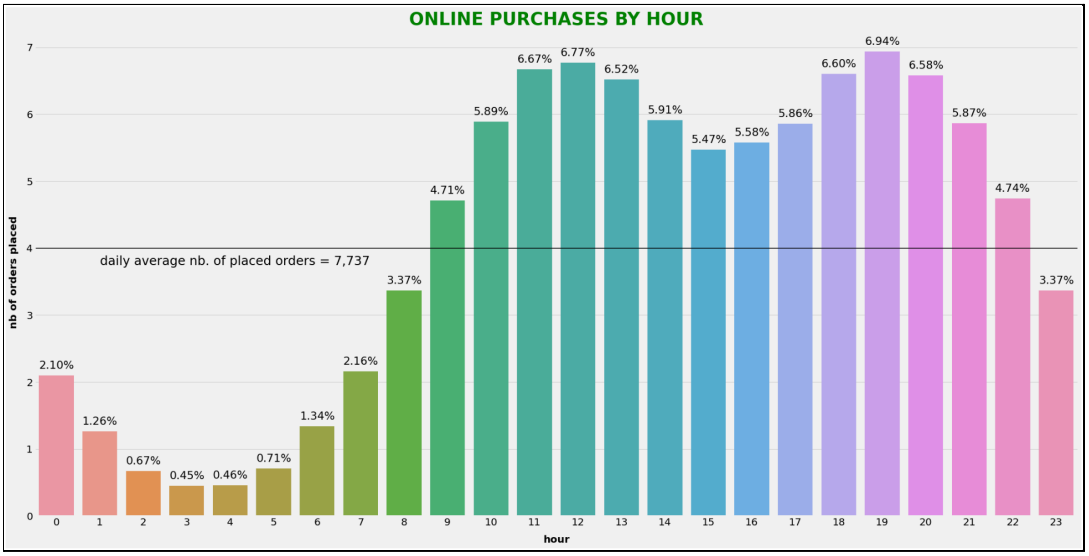



***Figure 6. Number of online purchases on BaseTech's website***

**PEAK ONLINE SHOPPING HOURS**

Based on *Figure 6* above, Tuesday has the highest traffic day of the week for online shopping, and the peak time of day for BaseTech occurs between 11 am and 1 pm, with a second peak at 6 pm and 8 pm.

There is a slight drop between 2 pm – 5 pm before reaching the second peak hours. Then there is a significant drop-off starting at 10 pm. Online sales reflect buyers' daily habits, with sales declining overnight before building up again at 6 am and increasing throughout the remainder of the day.

**WEEKLY ROUTINE FOR ONLINE PURCHASE**

There doesn't seem to be a particular day for online shopping. Tuesday is a little busier than the rest of the days, but there is no significant difference. The rise of mobile consumers buying online can explain the constant sales traffic throughout the week. Mobile shopping allows buyers to quickly purchase goods in just a few clicks, anywhere and anytime, as long as there's service.

**MONTHLY PATTERN FOR ONLINE PURCHASE**

Pay dates somewhat influence the monthly sales routine for BaseTech. Usually, peak days for sales come at the beginning and end of each month around payday. Shoppers tend to buy online at the beginning of the month than at the end. This type of routine can be justified by consumers who are likely to prioritize monthly bills at the end of the month than acquire new things.

***SUMMARY**: For BaseTech, knowing the busiest times can help improve service and performance and build an optimized basket abandonment strategy. These are findings from BaseTech's sales analysis, and it's different for other types of business. For example, restaurants may have their most substantial peak periods around Valentine's Day and much less around Black Friday and Cyber Monday, where most online electronic shops generally have the highest sales.*

## 3.3. Geographic distribution

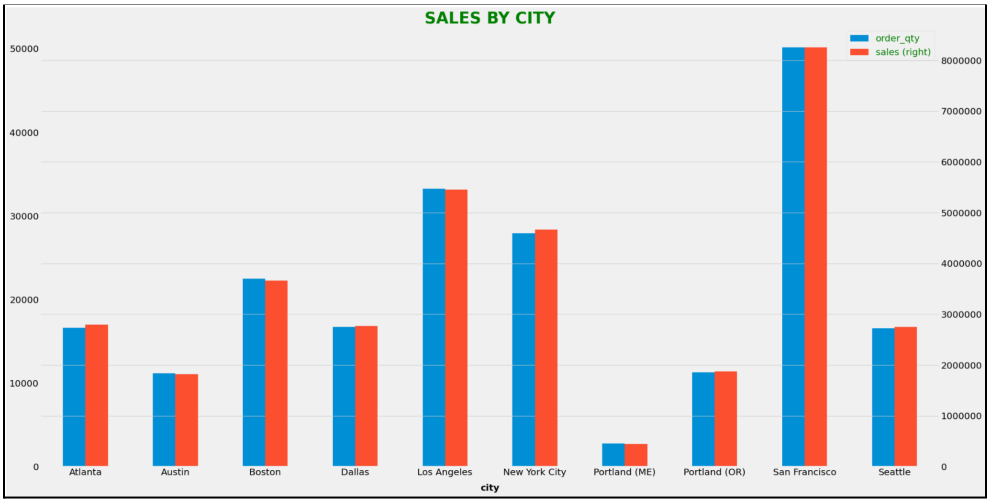



***Figure 7. Number of BaseTech's sales revenues per city***

BaseTech is located in San Francisco, California; it's no wonder that the highest sales come from San Francisco itself (*cf Figure 7*). With $8 254 743 million of sales generated, accounting for 23.95% of BaseTech's sales. This may be because BaseTech offers click and collect (drive-thru), so consumers from San Francisco have easier access to picking up their orders than Los Angeles and the rest of the city. Second in place comes Los Angeles, also located in California, representing 15.81%, $2 806 439 million less than San Francisco. This may translate to the fact that shipping delay is the lowest in the Californian state.

New York City, in third place, is known as one if not the busiest city in the world. Unsurprisingly, New York tends to purchase electronic goods to provide them with the necessary tools to accompany them in their active lifestyle. Especially in such an advanced and modern city like New York, people are mostly equipped with electronic devices as part of their daily routine.
For instance, it is ubiquitous to see New Yorkers rushing in the streets with their headphones on while talking to someone on the phone while carrying their laptops inside their briefcases. This has prompted New York to become a high-stakes testing base for metropolitan deliveries.

Next up are Boston, Massachusetts (10.62%), Atlanta, Georgia (8.11%), Dallas, Texas (8.02%), and Seattle, Washington (7.96%), respectively. While in the top-bottom comes Portland, Oregon (5.43%), Austin, Texas (5.27%), and Portland, Maine (1.30%).

***SUMMARY**: It's unsurprising that BaseTech customers are from well-known metropolitan areas. Consumers from big cities spend more than suburbian shoppers, while buyers from the rural areas are the ones who typically expend the least.*

# 3.4. Product ranking

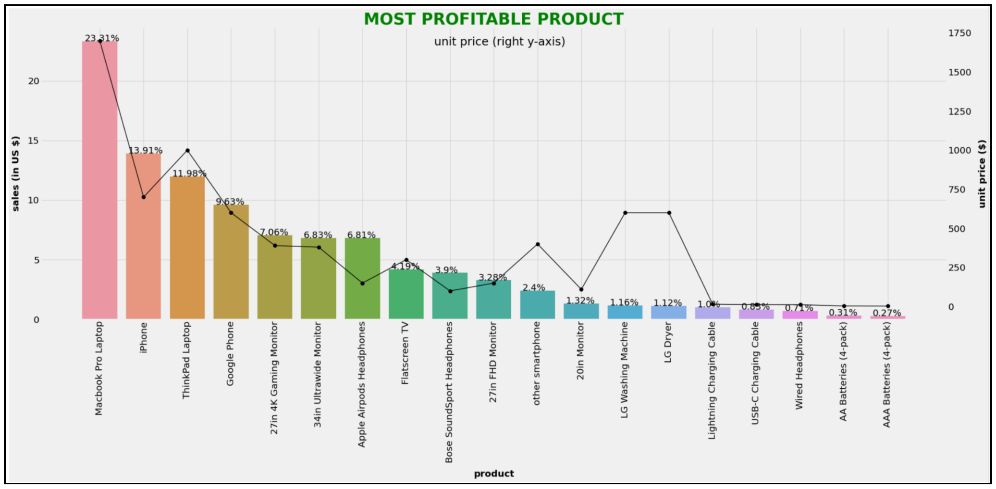



***Figure 8. BaseTech's sales revenues and price per product***

## 3.4.1 Most profitable product

According to the *Figure 8* above, the top two high sales generating units from BaseTech are Apple products, representing 37.22% of total annual sales: the MacBook Pro laptop hitting the first spot at 23.31% and the iPhone at 13.91% as the first spot second most profitable product. That leaves 62.78% for 17 other products BaseTech sells.

On the other hand, smaller electronic products such as batteries, charging cables, and wired headphones sit at the bottom of the ranking, accounting for merely 3.12% of the total annual sales.

Their rankings are primarily because of how much they cost: MacBook Pro has the highest selling price on BaseTech's listings, selling at $1 700, and iPhone being the most expensive item in the smartphone category at $700. The bottom 5 products on the list are the cheapest items available at BaseTech, costing roughly $15 or less.

While the statement is true for these products, it does not apply to home appliances. For instance, both the LG washing machine and LG dryer cost $600, the same price as the Google phone, yet they both ranked at the bottom of the list while the Google phone is part of the top 5. This means less demand for home appliances, which is very common in the electronic category.

Indeed, computers and smartphones are often considered business tools, media consumption, online business, education, and gaming. Not only are they an essential part of people's daily routine, but the fast-paced digital world with yearly releases keeps people buying and updating their devices to enjoy the new features and technology.

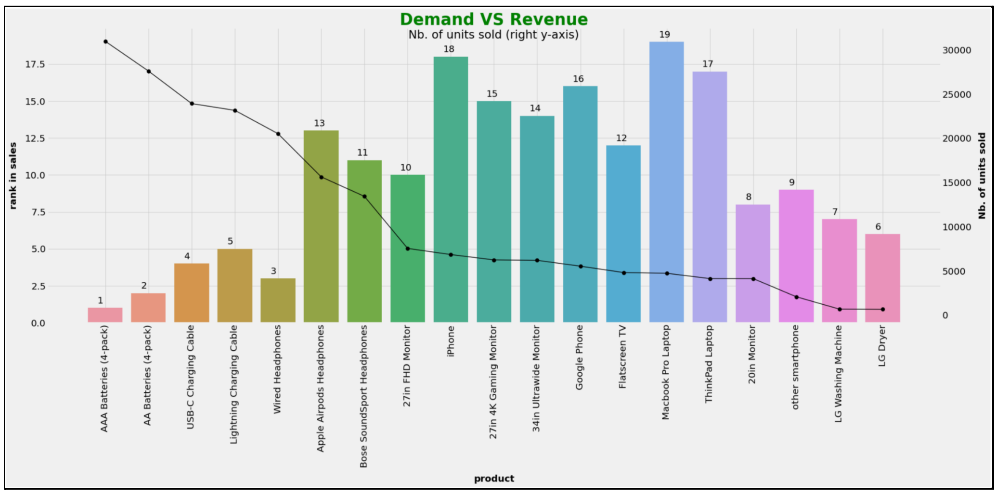



***Figure 9. BaseTech's sales revenues and number of units sold per product***

## 3.4.2 Highest selling product

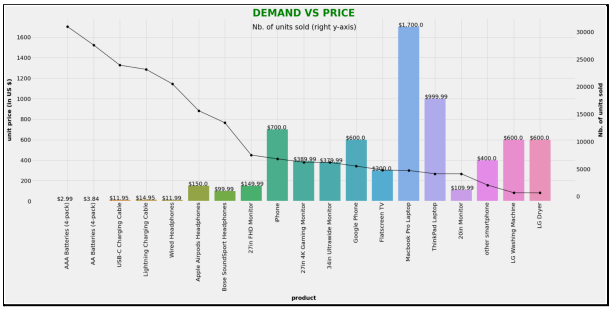



***Figure 10. Number of units sold and price per product***

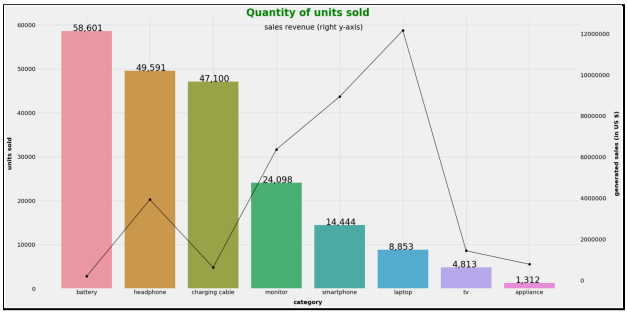



***Figure 11. Number of units sold and generated sales per product***

**Batteries, a household must-have**

Sold at the lowest price for less than $4, the highest-selling products are AAA Batteries (4-pack), which amounts to 14.84% of total sold products (*cf. Figure 10 and 11*), followed by AA Batteries (4-pack) at 13.22%. Cheapest items at BaseTech, these products are easy to store and deemed necessary, whether at home, school, or other buildings. These are the kind of products usually bought as complimentary items related to what consumers intended to buy in the first place. Although in this case, batteries are purchased as it is in bulk. Smaller electronic items with shorter life spans are often bought in bulk for several reasons:

- Having basics in stock means consumers don't have to worry about running out of those products for a while
- Consumers can save money because retailers usually offer discounted prices when buying in bulk

**From stored energy to transmitting energy**
Next to energy-storing batteries comes energy distributing charging cords. USB-C and Lightning charging cables, ranking third and fourth respectively, are items that usually get replaced frequently.
Any charger cord often wears out and loses effectiveness over time.
USB-C charging cable sells slightly higher than Lightning cable by 762 sold units, a minimal difference of 0.32%. This is because more devices use USB-C than Lightning cables. USB-C is much more universal than the proprietary Lightning cable exclusive to Apple's iPhones, AirPods, other computer accessories like Magic Keyboard, Magic Trackpad, Magic Mouse, and basic/older models of iPads. Logically, USB-C would sell better, even for some Apple consumers; Apple started implementing USB-C charging ports on their latest MacBook and iPads.

**Isolated yet connected**
Noise-canceling earbuds, higher sound quality wired headphones, it's portable… Many people can get through a whole car ride, a workday, a workout, or a meal without paying attention to anyone. As the course in technology continues to push towards uncovering new ways to isolate ourselves from the external world, it is also an instrument that re-connects people to the world. It's no wonder that people have such a solid connection to their devices, and headphones play a massive part in that.

No matter the price point, from $11.99 wired headphones to a staggering $150 AirPods, headphones are selling better than any smartphones, laptops, and monitors combined. Wired headphones in fifth place, representing 9.83% of total sold units, followed by over 15 thousand shipped AirPods, then Bose's SoundSport headphones in seventh place.

**Productivity devices**
Smartphones, monitors, and laptops are essential in everyday life. These are the primary devices people use daily, from personal to professional.

Even though these products are part of people's daily routine, they tend to sell less than smaller electronics. This is because they have a longer lifespan than batteries and charging cables. In addition to that, numerous people don't buy these products on their own, as opposed to headphones, because more often than not, many employers provide computers, phones, and monitors already. In addition, businesses usually have specific bigger suppliers that supply them with the tools they need, unlike smaller online shops like BaseTech.

Another reason to consider is that, as many people like to be updated on the latest technology, several people still don't see that as crucial. Once they have acquired the device they need, they hold on to that and stand by it for several years. On the other hand, one thing to notice is that Apple products are on the higher side of sold units than its counterpart. For instance, the iPhone sells better than the other smartphones BaseTech offers, and the MacBook Pro is a bit ahead of the ThinkPad laptop.

**Appliances**
Home appliances ended up at the bottom of most sold products, although they generated higher sales revenues than AAA batteries, primarily thanks to their higher price. In 2019, BaseTech only sold 666 units of LG washing machines and 20 units less for LG dryers. That's 13.37% less than the highest selling product, a massive difference of 29 674 units sold.

## 3.4.3. Frequently bought together

Apple's iPhone and Lightning charging cable with 521 transactions are the products most often bought together, then the Google phone and USB-C charging cable with 506 orders.

```
count = Counter()

for row in order_combo["group"]:
    row_list = row.split(", ")
    count.update(Counter(combinations(row_list, 2)))

for key, value in count.most_common(10):
    print(f"{key, value}")
```

```
(('iPhone', 'Lightning Charging Cable'), 521)
(('USB-C Charging Cable', 'Google Phone'), 506)
(('Google Phone', 'USB-C Charging Cable'), 491)
(('Lightning Charging Cable', 'iPhone'), 490)
(('iPhone', 'Wired Headphones'), 233)
(('Wired Headphones', 'iPhone'), 229)
(('Wired Headphones', 'Google Phone'), 229)
(('other smartphone', 'USB-C Charging Cable'), 196)
(('iPhone', 'Apple Airpods Headphones'), 195)
(('Google Phone', 'Wired Headphones'), 193)
```

***Figure 12. List of products frequently bought together and purchase frequency***

# 4. SALES FORECASTING

Demands and trend forecasting are no longer luxury commodities but a necessity. Many forecasting methods have been built to address managerial forecasting problems' increasing diversity and intricacy.

The selection of forecasting technique relies on multiple segments, such as the objective, availability of information, degree of precision, period, cost the business is willing to spend, and the availability of the time needed to deliver.

**Basic types of forecasting**
There are three general kinds of forecasting – qualitative approaches, time series analysis, and causal models.

- The qualitative technique uses qualitative information (expert opinion, for example) and details about particular occasions that may or may not consider the past.
- On the other hand, the second depends entirely on historical data and focuses on pattern changes.
- The last one is powerful enough to take particular circumstances properly and employs highly developed and specific information about associations between system components. Historical data is vital to causal models like time series and projection techniques.

This report will specifically cover time series analysis. These are statistical practices used when several years of data information are available and the trends are apparent and somewhat stable.

***** Although, in this case, BaseTech was only able to provide a 12-month worth of sales data.***

**Time series models**
A time series is a set of observed ordered values at subsequent points in time—in this context, BaseTech's monthly sales revenues help to determine and demonstrate:

- Regularity and variation in the series, also known as "seasonalities."
- Cyclical or repeating patterns
- Trends

Multiple models are used to forecast time-series data, like classical time series, supervised, and deep-learning-based models. This report will cover the classical time series model.

**Chosen method**
A distinct group of methods and techniques is fitted for forecasting value according to time in machine learning. Among other methods, the most common ones are ARIMA, SARIMA, VAR, and Holt-Winters.

This report will revolve around the ARIMA model, which stands for Autoregressive Integrated Moving Average. These parameters are labeled p, d, and q, respectively.

**p: Autoregressive (AR)**
The order of AR term is characterized by p. The autoregressive part refers to the association between the variable with its own lagged values; if p=2, the variable depends on the past two lagged values.
The Partial Auto-Correlation Function (PACF) plot can denote the order of AR.

**d: Integrated (I)**
The integrated part is fundamental when the data is non-stationary. It refers to the order of differencing, represented by d.

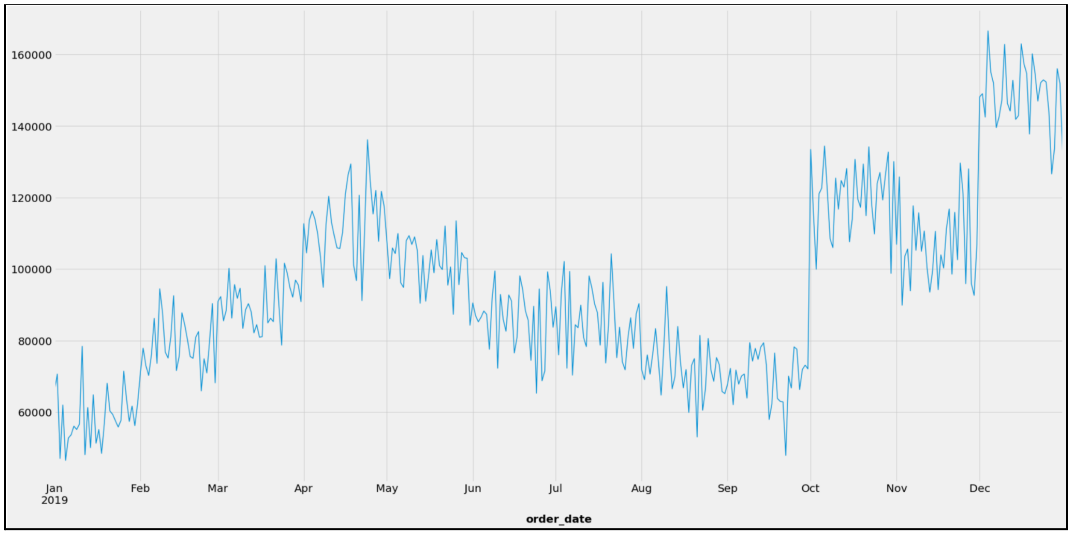
**q: Moving Average (MA)**
Moving Average order shows the dependence of the present value on the lagged error terms; its order is marked q.
The Auto-Correlation Function (ACF) plot can indicate the order of AR.

# 4.1. Data preparation

## 4.1.1. Plotting

Overall, there is a general upward trend present in the data (*cf. Figure 13*). Statsmodel can be performed to decompose this time series to understand the data's behavior better.

***Figure 13. BaseTech's sales revenues in 2019***



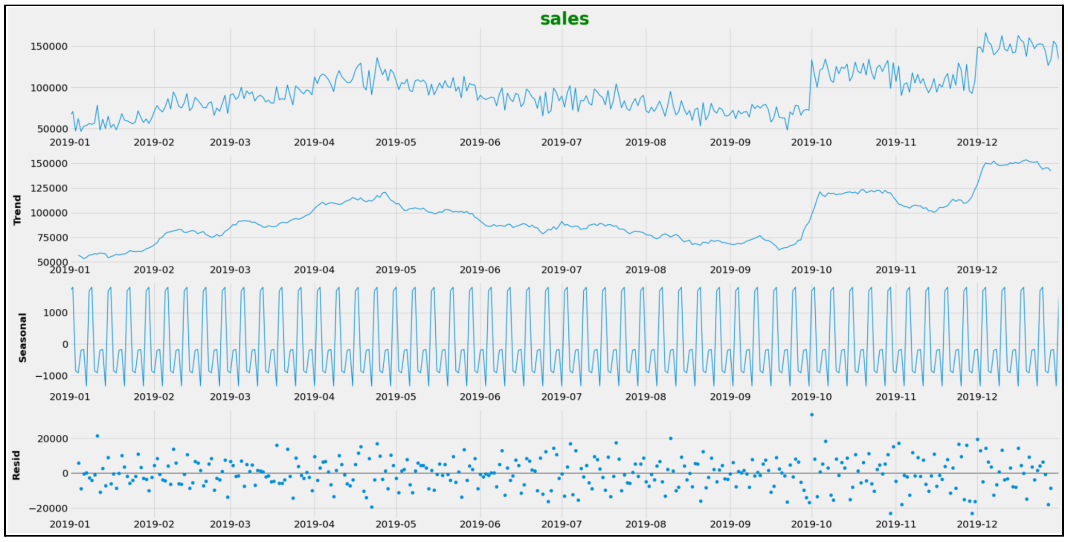

***** This analysis extensively relies on the statsmodel library, written in Python.***

## 4.1.2. Decomposition

Time series decomposition breaks down the observed time series data's trend, seasonality, and residual or noise components.

The time series exhibits a general upward trend (*cf. Figure 13 and Figure 14*). The seasonal component looks unpleasant, as shown in *Figure 14* below; the series can be assumed non-seasonal. Before applying any statistical model to a time series, ensuring the stationarity of the data is necessary.

***Figure 14. Decomposition of time-series featuring the trend, seasonality, and residuals***




## 4.1.3. Stationarity

The idea of stationarity revolves around the concept of consistency. If the mean of a time series increases over time, it's not stationary.

**Plotting rolling statistics**

Plotting rolling mean and variance is a good way to examine the series visually. If the plot shows an upward or downward trend with varying mean and variance, then the data is considered non-stationary. The rolling statistics plot below displays a non-stationary time series. An ADF test will verify this assumption.

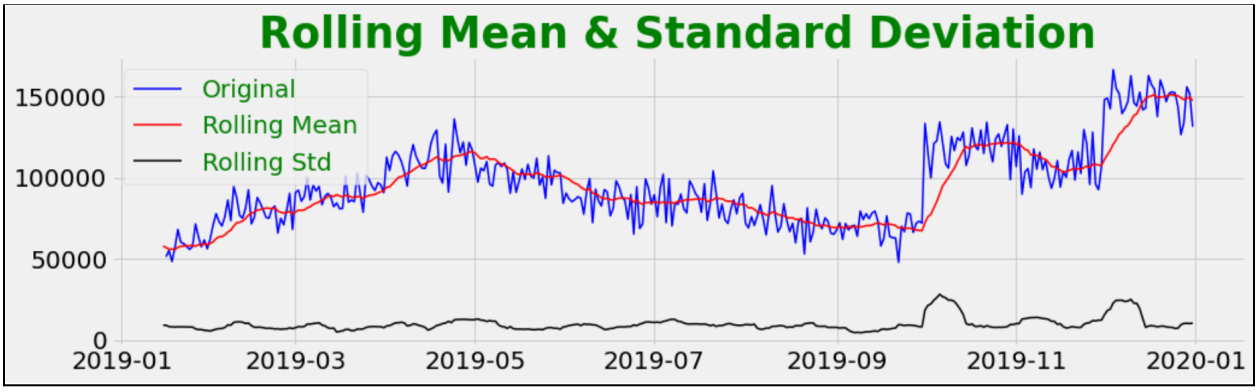



***Figure 15. 15-day moving average and rolling standard deviation***

**Augmented Dickey-Fuller Test**

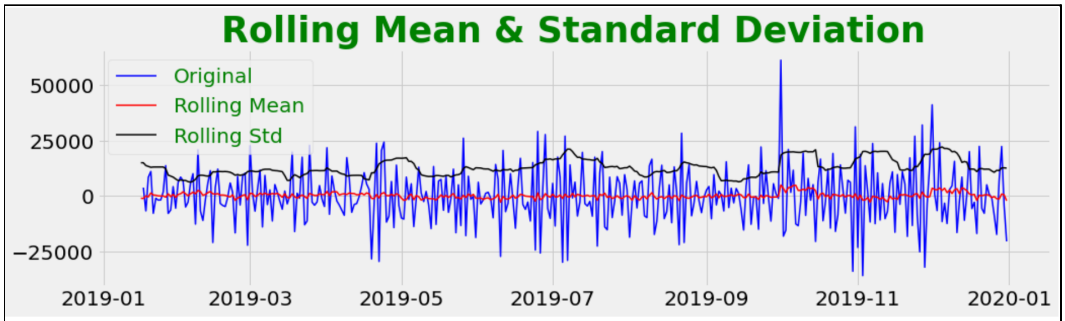
This statistical test evaluates if a time series is stationary. The ADF test (*cf. Figure 16*) resulted in a p-value greater than the threshold (0.05), thus confirming assumptions from the rolling statistics plot. Differencing is necessary to stationarize the data.

```
Results of Dickey-Fuller test :
------------------------------------------------
Test Statistic                  -1.616561
P-value                          0.474532
# Lags Used                      4.000000
Number of Observations Used    360.000000
dtype: float64


Is this data stationary ?
        Critical value 1%: -3.448645946352023 - The data is not stationary with 99% confidence
        Critical value 5%: -2.869602139060357 - The data is not stationary with 95% confidence
        Critical value 10%: -2.5710650077160495 - The data is not stationary with 90% confidence


CONCLUSION : time series is non-stationary, accept H0
```

***Figure 16. Augmented Dickey-Fuller test results with a p-value of 0.47, confirming a non-stationary time series***

## 4.1.4. Differencing

The method used for stationarizing the data is differencing. Differencing helps stabilize the mean by subtracting the previous value from the current value. Differencing is done by using the ***.shift()*** method in Python. Another method to get the order of differencing is by using the function "***ndiffs***." It estimates the number of differences needed to make a time series stationary. The data plotted below in *Figure 17* reach stationarity with one order of differencing.




```
Results of Dickey-Fuller test :
------------------------------------------------
Test Statistic               -1.404486e+01
P-value                       3.254470e-26
# Lags Used                   3.000000e+00
Number of Observations Used   3.600000e+02
dtype: float64


Is this data stationary ?
        Critical value 1%: -3.448645946352023 - The data is  stationary with 99% confidence
        Critical value 5%: -2.869602139060357 - The data is  stationary with 95% confidence
        Critical value 10%: -2.5710650077160495 - The data is  stationary with 90% confidence


CONCLUSION : time series is stationary, reject H0
```

***Figure 17. Rolling statistics and ADF test results showing a stationarized time series after 1st differencing***

## 4.1.5. ACF and PACF plots

Autocorrelation and Partial Autocorrelation plots find the required number of AR (p) and MA (q) terms.

- **Autocorrelation (ACF) plot**
  The ACF chart determines the optimal required number of MA terms to eliminate autocorrelation.
- **Partial autocorrelation (PACF) plot**
  PACF defines the optimal number of terms in the AR model and represents the pure correlation between the series and its lag. PACF plots.

*Before and after differencing*

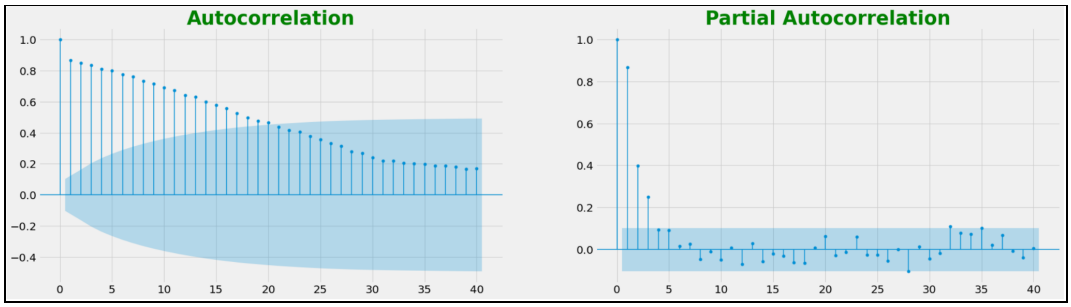



***Figure 18. Non-stationary time series correlation plot before differencing***

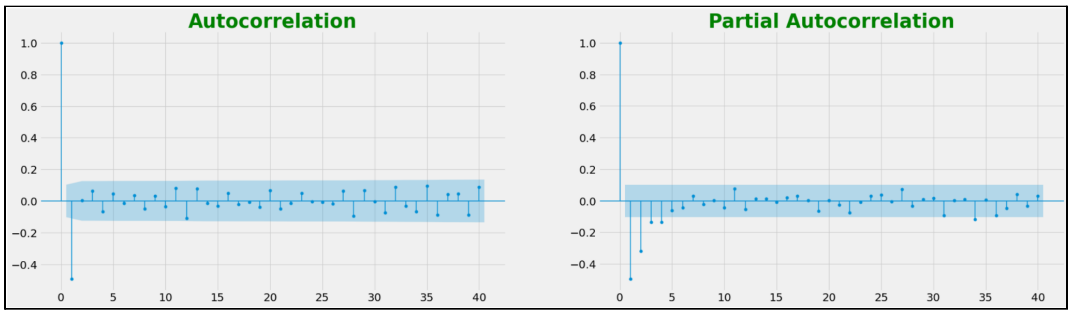



***Figure 19. Stationarized time series correlation plot after 1st differencing***

The ACF chart from the original series shows non-stationary data (*cf. Figure 18*), portrayed by the slow linear decline in the spikes. First-order differencing transformed the ACF plot with a single negative spike at lag 1 (*cf. Figure 19*).

The horizontal blue lines define the significance thresholds, and the vertical lines characterize the ACF and PACF values. Only the vertical lines exceeding the thresholds are significant. The chart in Figure 19 indicates an AR(4) and MA(1) model. But note that it's always better to choose a simplified model.

***** It's feasible to test various combinations of ARMA models. A diagnostic can be done to evaluate their performances. The model with the best (lowest) AIC and BIC scores will be chosen.***

### 4.1.6. Cross-validation

Cross-validation is a statistical approach utilized to evaluate the accuracy of the model. This is done by creating a train-test validation split.

The starting date of the time series is from January 2019 to December 2019. The training data will be from January to mid-December, while the testing data will contain the rest of the dates left (the last 15 days of December).

## 4.2. Building the model

There are several different models to consider oftentimes when building time series models. The Akaike Information Criterion (AIC) evaluates the quality of each model to find the most optimal: the lower the AIC value, the better. In this series, the model with the lowest AIC, as shown in *Figure 20* below, is ARIMA(0,1,1). Plotting the residuals helps to ensure that there are no patterns (constant mean and variance).

```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=7490.599, Time=0.22 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=7613.523, Time=0.02 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=7537.344, Time=0.05 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=7488.662, Time=0.16 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=7611.682, Time=0.02 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=7492.447, Time=0.05 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : AIC=7490.643, Time=0.19 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=7488.945, Time=0.26 sec
 ARIMA(0,1,1)(0,0,0)[0]             : AIC=7488.870, Time=0.03 sec

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

***Figure 20. Configuring the most optimal ARIMA order with the lowest AIC using auto-arima***

## 4.2.1. White noise

White noise is variations in data that any regression model cannot explain.
To read the above figure from top left to bottom right:

- **Top left**: the residual errors appear to fluctuate close to the mean of 0.
- **Top right**: the density plot indicates a gaussian distribution with a mean of around 0
- **Bottom left**: scatter points should align with the linear line. Any significant deviations would indicate a skewed distribution.
- **Bottom right**: the ACF chart displays no autocorrelation between the residual errors. Overall, the residual errors seem to be a good fit, with a mean near 0

**The Ljung-Box test hypothesis**
One uses of the Ljung-Box test is to determine if the errors are iid–independent, identically distributed (i.e., white noise). Ljung-Box is a test of lack of fit: a model is a good fit if the autocorrelations of the residuals are small.

According to the test (*cf. table on the right*), the p-value for lags from 5 to 30 are insignificant (greater than the 0.05 threshold), so the null hypothesis is accepted. Therefore, the time series is white noise and can be used for forecasting.

| lb_pvalue |
|---|
| 0.763736 |
| 0.947845 |
| 0.995226 |
| 0.998946 |
| 0.998731 |
| 0.998328 |

## 4.2.2. Model validation

No matter how good the model looks, it still needs evaluation. Validating a model assesses the model's accuracy of the prediction by comparing the predicted values to the actual values. The model used here has only 365 data points: 335 observations went into training, while the last 35 values were used for testing to try and predict the last 15 days in the dataset. The model resulted in a decent MAPE score of 5.64%. But will it result in accurate forecasting results?

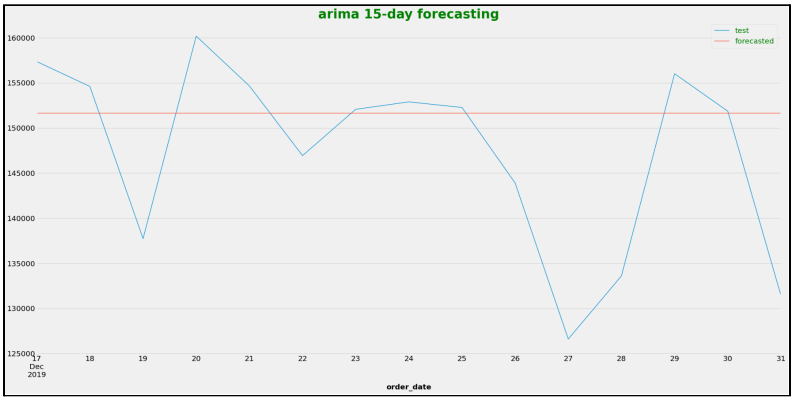



***Figure 21. 15-day prediction: Original values vs Forecasted values***

The score achieved may be acceptable, but the result of the prediction did not catch any data fluctuation (*cf. Figure 21*). This will not predict the future for the next month's sales as no seasonality was captured. This shows that a lack of data will not provide any clue of the future direction except for averaging all the observations—and that would only result in a straight line when visualizing. This type of forecast is undependable because it doesn't deliver a sure estimate of the future. Foretelling the future using this model would still give the same unreliable outcome.

## 4.3. Future prediction (forecasting the next 15 days)

Even though the series is worth one year of data, it is still considered insufficient to render a reliable forecast. The results gave some indications for the trend but none for seasonality. The only way to add seasonal components to this data is with business knowledge and tweaking the models manually, like data augmentation.

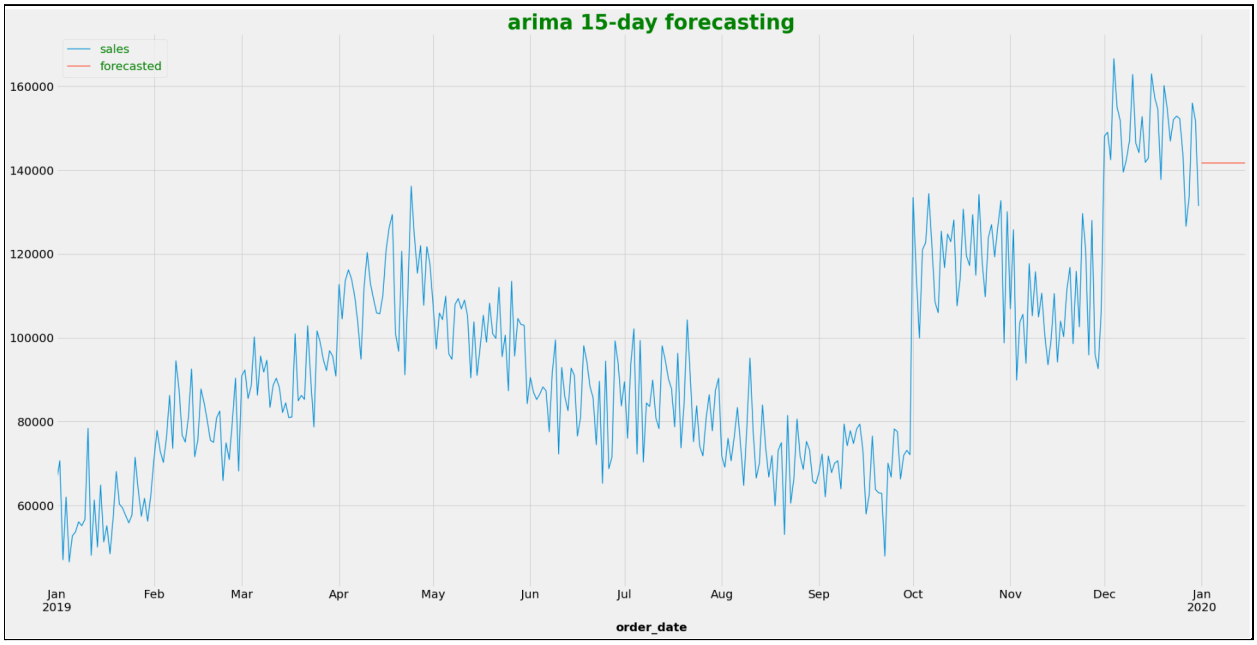



***Figure 22. 15-day forecasting results (January 1 to January 15, 2020)***

# THE TAKEAWAY:

This project aims to analyze BaseTech's sales performance in 2019 and use a forecasting technique to predict their sales revenues for an upcoming period.

In 2019, BaseTech experienced a somewhat turbulent year with ups and downs but managed to cope with the trend, resulting in a good year overall. The first four months were an excellent start for BaseTech, with the demands kept growing. Unfortunately, they experienced a continuous decline in sales in the next 5 months without a clear indication of the cause. Is it due to an inventory shortage? Are they losing clients because of poor customer service, unsatisfactory online shopping experience, and lack of client retention strategy? Weren't they investing enough in marketing to expand their online presence and customer acquisition?

I have implemented a time series forecasting technique, mainly the ARIMA method from the classical time series model. Forecasting will help BaseTech in its decision-making process, implement new strategies, and improve its sales inventory management for the upcoming weeks or months.

However, I didn't fully achieve my plan. Even though I have chosen a suitable approach for predicting BaseTech's future sales, it didn't result in my envisioned outcome. The lack of data deprived me of any further analysis to answer the questions like the ones listed above. It also prevented me from making successful and more accurate predictions.

Nonetheless, with all these statistical techniques developed over the years, there may be another better method of forecasting small data. The Data Analyst training provided me with the tools to collect, aggregate, process, explore, analyze, and train data for more practical uses like machine learning. It gave me the knowledge to understand how data works, but it also opened the doors for me in more in-depth advanced maths, which I hadn't had a chance to learn before.

Realizing how much more there is than the classical time series model for forecasting, so much more than PCA to interpret multicollinearity between features, and how many other machine learning techniques, probabilities, and hypotheses there are, I would like to pursue this domain to understand data better than just simple univariate and bivariate analysis. Data Scientist training will undoubtedly help me acquire the skills I am looking for.

# Questions?
# Contact me.

Katrina JUMADIAO
**Data Analyst**, *OpenClassrooms*
konz.katrina@gmail.com

BaseTech