Anomaly Detection

Joshua Kibuye

2022-04-04

Carrefour Marketing Analysis (Anomaly Detection)

1. Defining the Question

a) Specifying the Data Analytic Question.

What are most relevant marketing strategies that will result in the highest no. sales at Carrefour Kenya.

b) Defining the Metric for Success

To identify any anomalies in the dataset ## c) Understanding the context You are a Data analyst at Carrefour Kenya and are currently undertaking a project that will inform the marketing department on the most relevant marketing strategies that will result in the highest no. of sales (total price including tax). Your project has been divided into four parts where you'll explore a recent marketing dataset by performing various unsupervised learning techniques and later providing recommendations based on your insights.

d) Recording the Experimental Design

- Data cleaning
- Performing extensive exploratory data analysis where applicable.
- Detecting anomalies in our data.

```
#Loadig packages
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tibbletime)
```

```
## Warning: package 'tibbletime' was built under R version 4.1.3
```

```
##
## Attaching package: 'tibbletime'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(anomalize)
```

```
## Warning: package 'anomalize' was built under R version 4.1.3
```

```
## == Use anomalize to improve your Forecasts by 50%! ==============================
## Business Science offers a 1-hour course - Lab #18: Time Series Anomaly Detection!
## </> Learn more at: https://university.business-science.io/p/learning-labs-pro </>
```

```
library(timetk)
```

```
## Warning: package 'timetk' was built under R version 4.1.3
```

2. Data Understanding

```
#Loading the dataset
df <- read.csv('http://bit.ly/CarreFourSalesDataset')
```

```
#Looking at the top of our dataset
head(df)
```

```
##        Date    Sales
## 1  1/5/2019 548.9715
## 2  3/8/2019  80.2200
## 3  3/3/2019 340.5255
## 4 1/27/2019 489.0480
## 5  2/8/2019 634.3785
## 6 3/25/2019 627.6165
```

```
#Looking at the bottom of the dataset
tail(df)
```

```
##           Date     Sales
## 995  2/18/2019   63.9975
## 996  1/29/2019   42.3675
## 997   3/2/2019 1022.4900
## 998   2/9/2019   33.4320
## 999  2/22/2019   69.1110
## 1000 2/18/2019  649.2990
```

```
#Getting information of the dataset
glimpse(df)
```

```
## Rows: 1,000
## Columns: 2
## $ Date  <chr> "1/5/2019", "3/8/2019", "3/3/2019", "1/27/2019", "2/8/2019", "3/~
## $ Sales <dbl> 548.9715, 80.2200, 340.5255, 489.0480, 634.3785, 627.6165, 433.6~
```

2. Processing the data for Anomaly Detection

```
#Changing the date column to date time
df$Date <- as.Date(df$Date, format = "%m/%d/%Y")
```

The date is now in the right format.

```
df$Date <- as.POSIXct(df$Date)
```

```
# Convert df to a tibble
df <- as_tibble(df)
class(df)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

The dataframe has been converted into a tibble.

3. Time Series Decomposition

```
#Performing the time decomposition
df %>%
  time_decompose(Sales, method = "stl", frequency = "auto", trend = "auto", message = TRUE)  %>%
  anomalize(remainder, method = "gesd", alpha = 0.05, max_anoms = 0.2) %>%
  plot_anomaly_decomposition()
```

```
## Converting from tbl_df to tbl_time.
## Auto-index message: index = Date
```

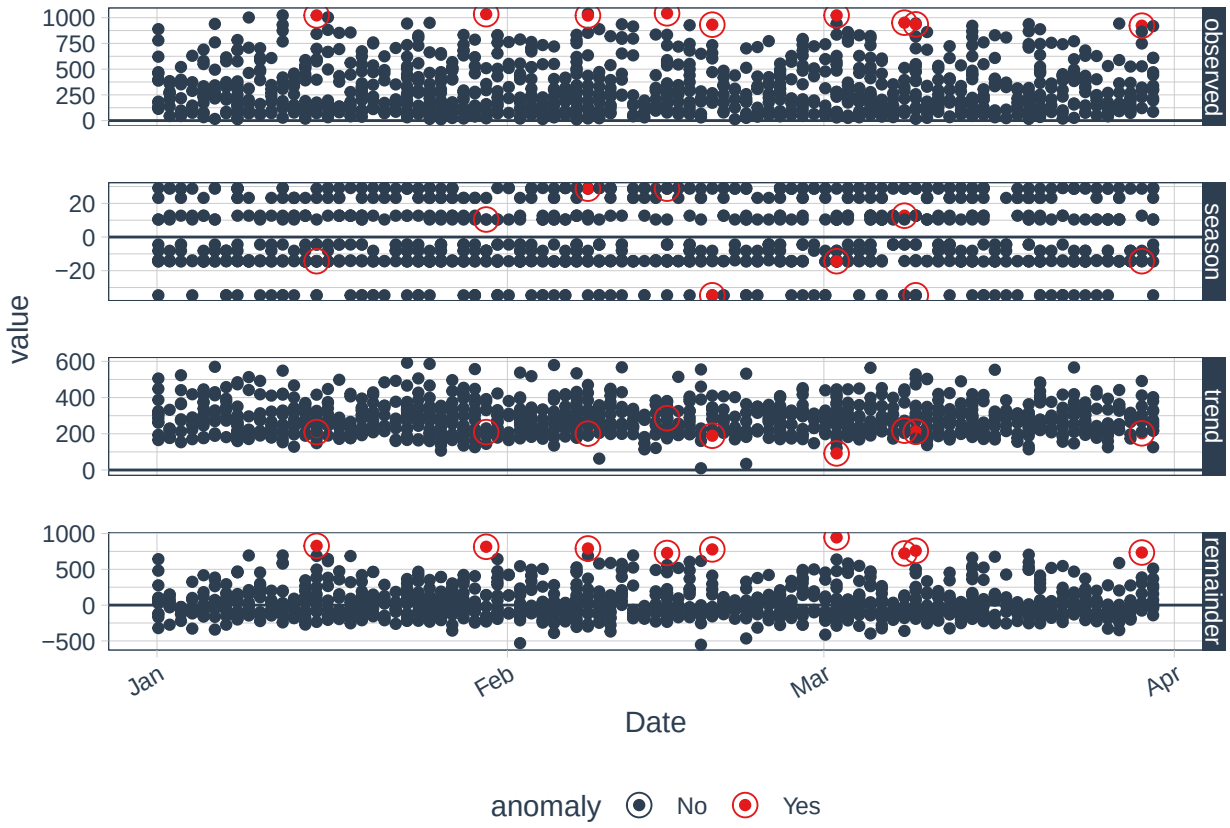
```
## Note: Index not ordered. tibbletime assumes index is in ascending order. Results may not be as desir
```

```
## frequency = 12 seconds

## Note: Index not ordered. tibbletime assumes index is in ascending order. Results may not be as desire

## trend = 12 seconds

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```



4. Detecting Anomalies

```
#Anomaly detection
df %>%
  time_decompose(Sales, method = 'stl', frequency = 'auto', trend = 'auto') %>%
  anomalize(remainder, method = 'gesd', alpha = 0.1, max_anoms = 0.1) %>%
  time_recompose() %>%
  plot_anomalies(time_recomposed = TRUE, ncol = 3, alpha_dots = 0.5)
```

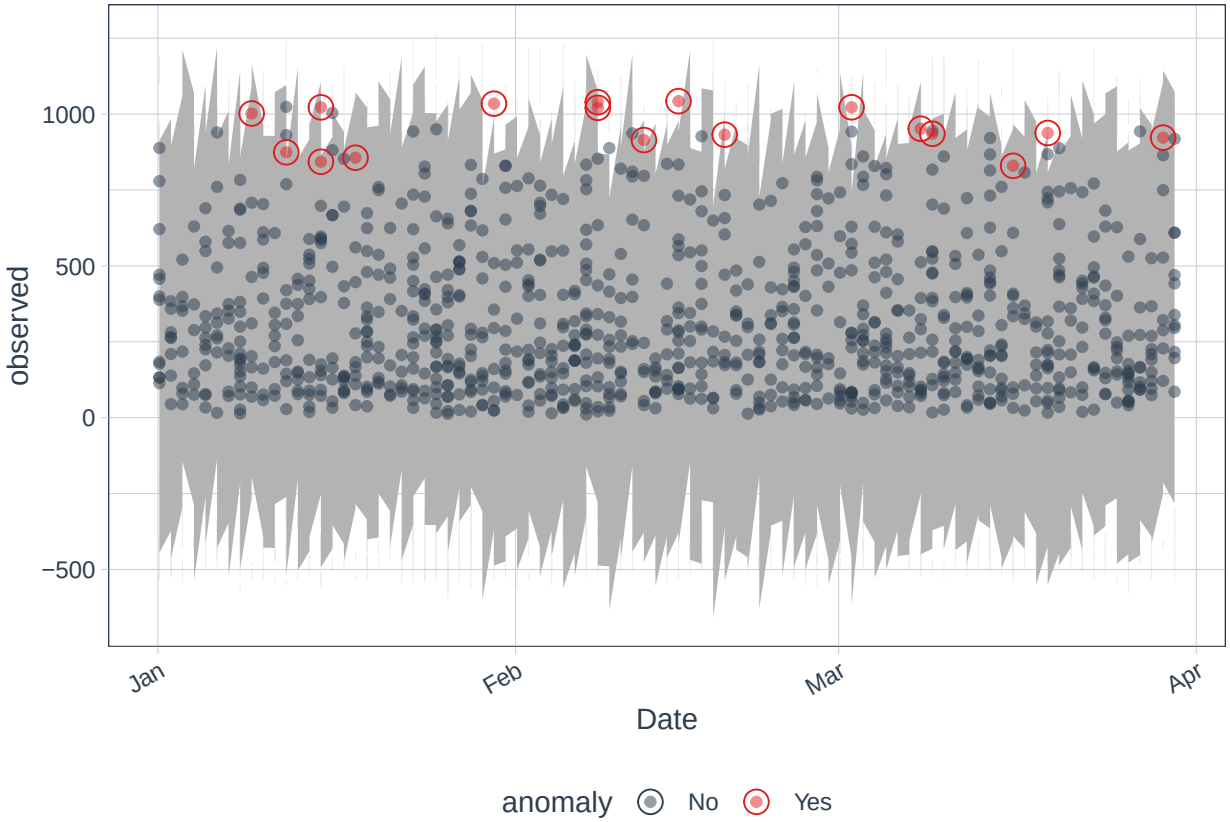
```
## Converting from tbl_df to tbl_time.
## Auto-index message: index = Date

## Note: Index not ordered. tibbletime assumes index is in ascending order. Results may not be as desire

## frequency = 12 seconds
```

```
## Note: Index not ordered. tibbletime assumes index is in ascending order. Results may not be as desire

## trend = 12 seconds
```



5. Conclusion There are 16 anomalies in the months of January to April.