Week 14 IP Part 4

Anomaly Detection

```
# Libraries
library(data.table)
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.0.5

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()

library(tibble)
library(tibbletime)

## Warning: package 'tibbletime' was built under R version 4.0.5

##
## Attaching package: 'tibbletime'

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

library(anomalize)

## Warning: package 'anomalize' was built under R version 4.0.5

## == 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 </>
```

```r
# Loading the data
df <- read_csv("http://bit.ly/CarreFourSalesDataset")
```

```
##
## -- Column specification --------------------------------------------------
------
## cols(
##   Date = col_character(),
##   Sales = col_double()
## )
```

```r
head(df)
```

```
## # A tibble: 6 x 2
##   Date      Sales
##   <chr>     <dbl>
## 1 1/5/2019  549.
## 2 3/8/2019   80.2
## 3 3/3/2019  341.
## 4 1/27/2019 489.
## 5 2/8/2019  634.
## 6 3/25/2019 628.
```

```r
# Information about the dataset
str(df)
```

```
## spec_tbl_df [1,000 x 2] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Date : chr [1:1000] "1/5/2019" "3/8/2019" "3/3/2019" "1/27/2019" ...
##  $ Sales: num [1:1000] 549 80.2 340.5 489 634.4 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Date = col_character(),
##   ..   Sales = col_double()
##   .. )
```

```r
# Change date column to datetime
df$Date <-as.Date(df$Date, format="%m/%d/%Y")
```

Data Cleaning

```r
# Column names
# Changing column names to lower case, and replacing spaces with underscores
colnames(df) = tolower(str_replace_all(colnames(df), c(' ' = '_')))

# Checking column names.
colnames(df)
```

```
## [1] "date"  "sales"
```

```r
# Null values
colSums(is.na(df))
```

```
##  date sales
##     0     0
```

- No null values were found

```
# Checking for duplicated records
sum(duplicated(df))
```

```
## [1] 0
```

- No duplicates were found

```
class(df)
```

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

```
df <- as.tibble(df)
```

```
## Warning: `as.tibble()` was deprecated in tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.
```

```
df <-as_tbl_time(df, index=date)
df <- dplyr::arrange(df,date)
```

```
# Detecting anomalies
df_anomalized <- df%>%
  as_period("daily")%>%
  time_decompose(sales)%>%
  anomalize(remainder)%>%
  time_recompose()
```

```
## frequency = 7 days
```

```
## trend = 30 days
```

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