Anomaly Detection

Brian Michira

9/10/2021

#Research Question To check whether the are anomalies in the given sales data.The objective of the task is fraud detection.

```
# Load libraries
suppressWarnings(
        suppressMessages(if
                         (!require(tidyverse, quietly=TRUE))
                 install.packages("tidyverse")))
library(tidyverse)
suppressWarnings(
        suppressMessages(if
                         (!require(anomalize, quietly=TRUE))
                 install.packages("anomalize")))
library(anomalize)
suppressWarnings(
        suppressMessages(if
                         (!require(tibbletime, quietly=TRUE))
                 install.packages("tibbletime")))
library(tibbletime)
suppressWarnings(
        suppressMessages(if
                         (!require(dplyr, quietly=TRUE))
                 install.packages("dplyr")))
library(dplyr)
```

```
#Load the data
df <-read.csv("http://bit.ly/CarreFourSalesDataset")
```

```
#checking the info
str(df)
```

```
## 'data.frame':    1000 obs. of  2 variables:
##  $ Date : chr  "1/5/2019" "3/8/2019" "3/3/2019" "1/27/2019" ...
##  $ Sales: num  549 80.2 340.5 489 634.4 ...
```

```
#changing the data type
df$Date <- as.Date(df$Date, format ="%m/%d/%Y")
df$Date <- sort(df$Date, decreasing = FALSE)
df <- as_tbl_time(df, index = Date)
df <- df %>%as_period("daily")
head(df)
```

```
## # A time tibble: 6 x 2
## # Index: Date
##   Date       Sales
##   <date>     <dbl>
## 1 2019-01-01  549.
## 2 2019-01-02  246.
## 3 2019-01-03  452.
## 4 2019-01-04  464.
## 5 2019-01-05  418.
## 6 2019-01-06  536.
```

```
#confirming if the data type has changed
str(df)
```

```
## tbl_time [89 x 2] (S3: tbl_time/tbl_df/tbl/data.frame)
##  $ Date : Date[1:89], format: "2019-01-01" "2019-01-02" ...
##  $ Sales: num [1:89] 549 246 452 464 418 ...
##  - attr(*, "index_quo")= language ~Date
##   ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##  - attr(*, "index_time_zone")= chr "UTC"
```

```
#Check the shape
dim(df)
```

```
## [1] 89  2
```

The data has 89 rows and 2 columns.

#Cleaning

```
#checking for missing values
sum(is.na(df))
```

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

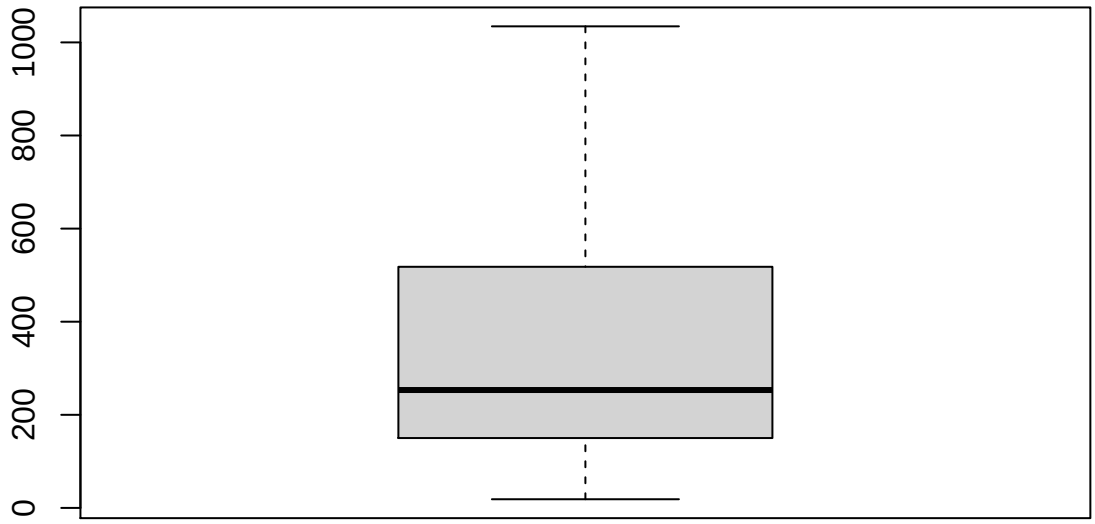
No missing values

```
#checking for missing values
duplicated_rows<-df[duplicated(df),]
duplicated_rows
```

```
## # A time tibble: 0 x 2
## # Index: Date
## # ... with 2 variables: Date <date>, Sales <dbl>
```

There are no duplicated rows.

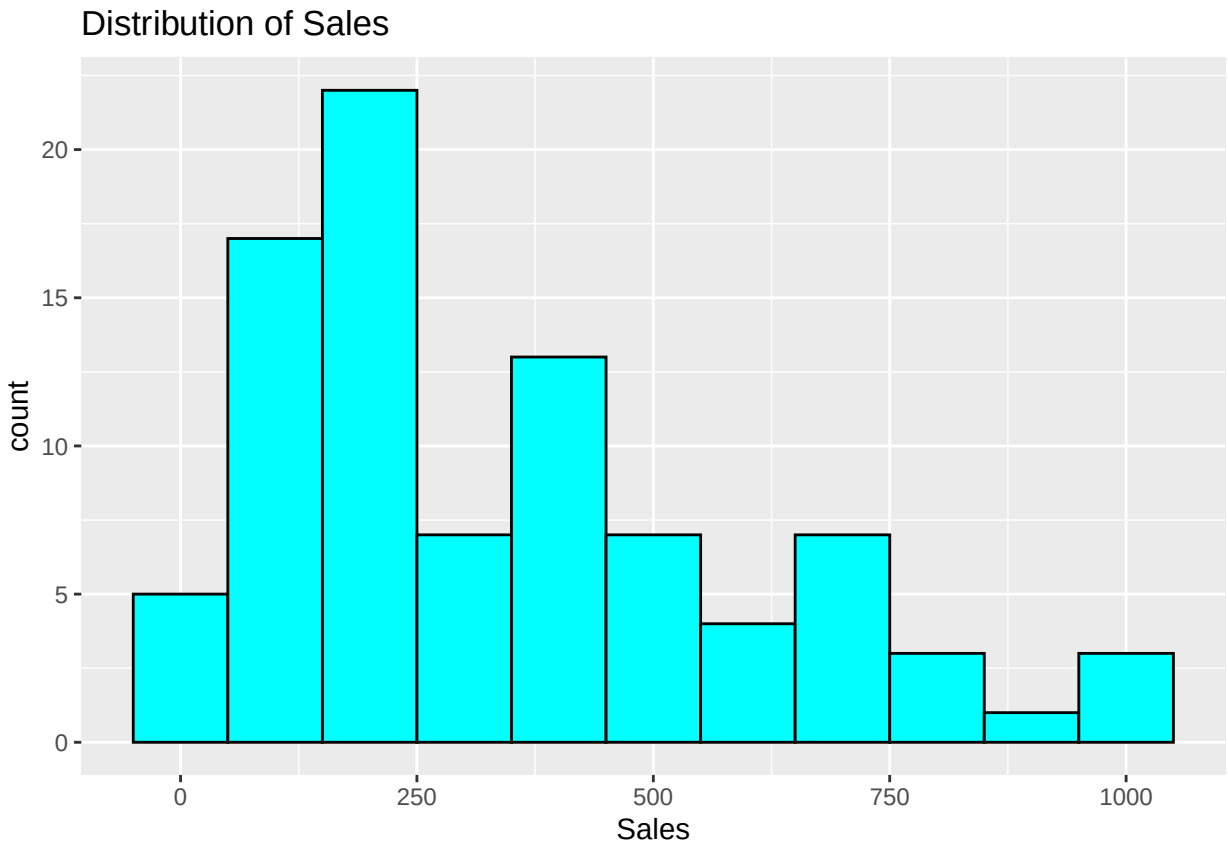
```
#checking for ouliers
boxplot(df$Sales)
```



There are no outliers

Univariate Analysis

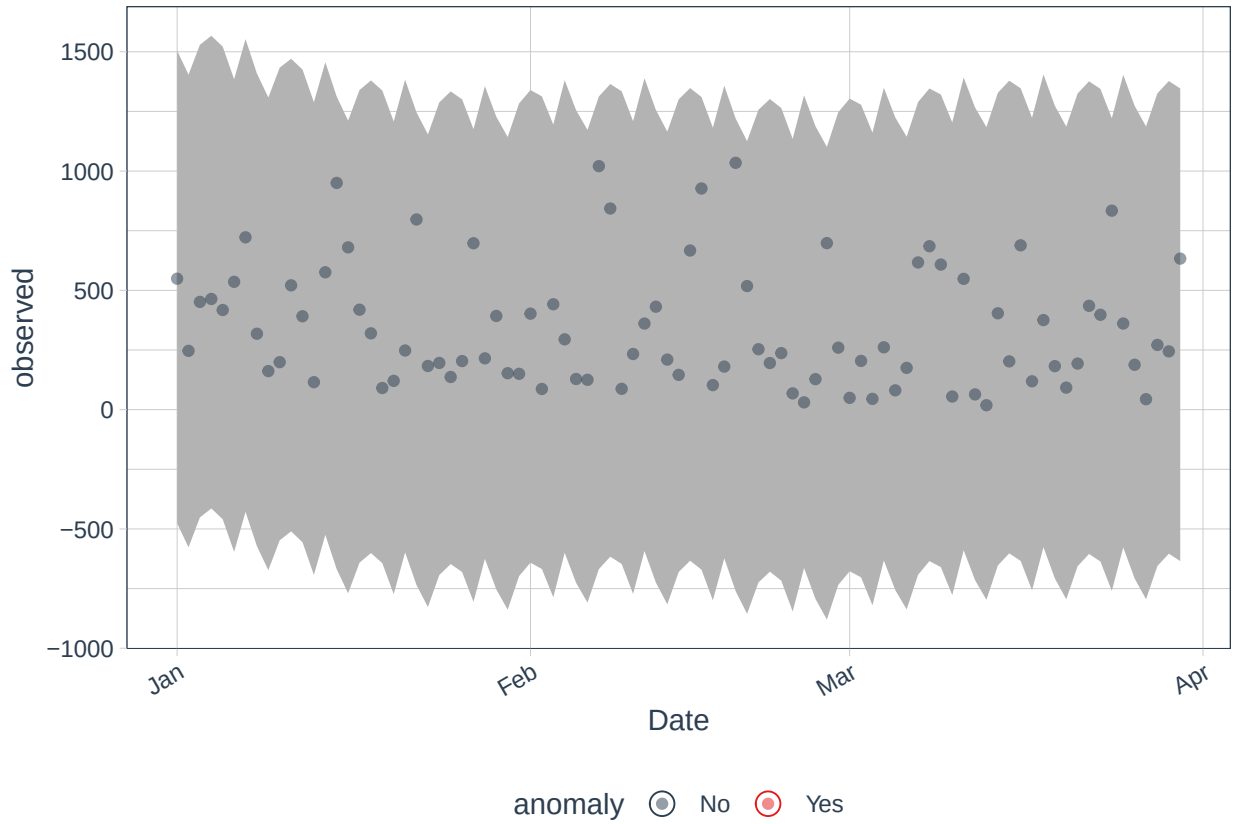
```
# plot the histogram of Unit.price
ggplot(df, aes(x = Sales)) +
  geom_histogram(fill = "cyan",
                 color = "black",
                 binwidth = 100) +
  labs(title="Distribution of Sales",
       x = "Sales")
```



Distibution of sales are positively skewed showing that most of the values are greater than the mean.

```
# Detecting our anomalies

df %>%
    time_decompose(Sales) %>%
    anomalize(remainder) %>%
    time_recompose() %>%
    plot_anomalies(time_recomposed = TRUE, ncol = 3, alpha_dots = 0.5)
```

```
## frequency = 7 days

## trend = 30 days

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

## Warning: `type_convert()` only converts columns of type 'character'.
## - `df` has no columns of type 'character'
```



There are no anomalies in our dataset.