R Notebook

ANOMALY DETECTION

1. DEFINING THE QUESTION

a) Specifying the Question

Identifying anomalies in the dataset.

b) Defining the Metrics of Success

Identifying anomalies in the dataset which is frraud etection.

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

1. Defining the question, the metric for success, the context and the experimental design.
2. Reading and exploring the dataset.
3. Identififying anomalies in the dataset which is fraud detection.

e) Relevance of the data

The data used will inform the marketing department on the most relevant marketing strategies that will result in the highest number of sales and total price including tax. The dataset link:http://bit.ly/CarreFourSalesDataset

2. DATA ANALYSIS

a) Checking the Data

```
# Loading libraries
library(data.table)
library(tidyverse)
```

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

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- 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(anomalize)
```

```
## == 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 </>
```

```
# Importing the data
df <- fread('http://bit.ly/CarreFourSalesDataset')
df
```

```
##            Date     Sales
##          <char>     <num>
##    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
##   ---
##  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
```

3. ANOMALY DETECTION

```
library(data.table)
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(mvtnorm)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(PRROC)
```

```
summary(df)
```

```
##      Date               Sales
##  Length:1000        Min.   :  10.68
##  Class :character   1st Qu.: 124.42
##  Mode  :character   Median : 253.85
##                     Mean   : 322.97
##                     3rd Qu.: 471.35
##                     Max.   :1042.65
```

```
skew <- sum(as.numeric(df$Class))/nrow(df)
sprintf('Percentage of fraudulent transactions in the data set %f', skew*100)
```

```
## [1] "Percentage of fraudulent transactions in the data set 0.000000"
```

There are no frauduent transactions in the dataset.