Anomaly Detection

Dennis Kiarie

9/11/2021

```
library(tinytex)
```

1. Define the Question

1.1 Research Question

Our Research seeks to check whether there are any anomalies in the given sales dataset.

1.2 Metric of Success

To check whether there are any anomalies in the given sales dataset. The objective of this task being fraud detection ***1.3 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.

1.4 Experimental Design

1. Loading Data into RStudio.
2. Checking the Data.
3. Tidying the Data.
4. Conducting Exploratory Data Analysis i.e Univariate, Bivariate and Multivariate Analysis.
5. Anomaly Detection.
6. Implement the Solution
7. Challenge the Solution
8. Follow up Questions

1.5 Data Relevance

The data provided its appropriate for our analysis. The dataset for this analysis can be found in this link:[http://bit.ly/CarreFourSalesDataset]

2.Data Preparation

```
## Importing libraries
#---
#
library(pacman)
library(data.table)
pacman :: p_load(pacman,ggbiplot,plyr, dplyr,scales, readr, grid,factoextra, GGally,DataExplorer, ggplo
```

```
## Installing package into 'C:/Users/Denoo/OneDrive/Documents/R/win-library/4.1'
## (as 'lib' is unspecified)
```

```
## Warning: package 'FSelecto' is not available for this version of R
##
## A version of this package for your version of R might be available elsewhere,
## see the ideas at
## https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages
```

```
## Warning: unable to access index for repository http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contril
##   cannot open URL 'http://www.stats.ox.ac.uk/pub/RWin/bin/windows/contrib/4.1/PACKAGES'
```

```
## Warning: 'BiocManager' not available.  Could not check Bioconductor.
##
## Please use `install.packages('BiocManager')` and then retry.
```

```
## Warning in p_install(package, character.only = TRUE, ...):
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'FSelecto'
```

```
## Warning in pacman::p_load(pacman, ggbiplot, plyr, dplyr, scales, readr, : Failed to install/load:
## FSelecto
```

```
theme_set(theme_classic())
options(warn = -1)
```

```
## Loading the data from a csv file
#---
#
df <- fread('http://bit.ly/CarreFourSalesDataset')
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
##   ---
##  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
```

```
##preview the first five records
#---
#
head(df, n=5)
```

```
##         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
```

##preview the last 6 records of the dataset #—

```
tail(df)
```

```
##         Date     Sales
## 1: 2/18/2019   63.9975
## 2: 1/29/2019   42.3675
## 3:  3/2/2019 1022.4900
## 4:  2/9/2019   33.4320
## 5: 2/22/2019   69.1110
## 6: 2/18/2019  649.2990
```

2. Anomaly Detection

```
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"
```

Conclusion

There were no fradulent transactions in the data.