Northeastern University

ALY 6040 Module 1 Technique Practice (1)

Student’s name: Mohammad Hossein Movahedi

Assignment title: Module 1 Technique Practice

Course number and title: ALY6040 71368 Data Mining Applications SEC 09 Fall 2022 CPS [TOR-A-HY]

Term: 202315_A Fall 2022 CPS Quarter First Half

Instructor’s name: Hootan Kamran, PhD

Sep 21, 2022

Table of Contents

Introduction

Data analytics can sometimes be fun. Remember all the time you want to get ice cream from MacDonald’s, and it turns out the Icecream machine is broken. The dataset I will clean and explore in this assignment targets this widespread problem.

For this assignment, I’m going to do the following actions.

1. cleaning the dataset with EDA techniques
2. doing some initial expository analytics on the dataset
3. calculate some appropriate summary and statistics
4. explain the findings and their importance

now let’s dive into it.

Introducing the Dataset and Data Preparation

The McDonalds Ice Cream Machines Breaking has 11 variables, but not all of them are suitable for this project. I deleted some location data that wasn’t useful for this project and then corrected the remaining data types.

First, I load libraries and the data set.

```
print('Mohammad Hossein Movahedi')
print('Assignment 1 AlY 6040')
#loading needed libraries
install.packages("tidyverse")
install.packages("funModeling")
install.packages("Hmisc")
install.packages('magrittr')
install.packages('FSA')
install.packages('FSAdata')
install.packages('magrittr')
install.packages('dplyr')
install.packages('tidyr')
install.packages('plyr')
install.packages('tidyverse')
install.packages('outliers')
install.packages('ggplot2')
install.packages('lubridate')
install.packages('corrplot')
library(funModeling)
library(tidyverse)
library(Hmisc)
library(magrittr)
library(ggplot2)
library(corrplot)
library(dplyr)
#reading the dataset
data <- read.csv('mcdonalds_dataset.csv')
```

Then I looked at the structure of the dataset

```
#looking into the structure of the data
str(data)
```

the result is shown below

```
> str(data)
'data.frame': 16671 obs. of  11 variables:
 $ lat         : num  -74 -74 -74 -74 -74 ...
 $ lon         : num  40.7 40.7 40.7 40.7 40.7 ...
 $ alt         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ is_broken   : chr  "False" "False" "False" "False" ...
 $ is_active   : chr  "True" "True" "True" "True" ...
 $ dot         : chr  "working" "working" "working" "working" ...
 $ state       : chr  "NY" "NY" "NY" "NY" ...
 $ city        : chr  "New York" "New York" "New York" "New York" ...
 $ street      : chr  "114 Delancey St" "208 Varick St" "724 Broadway" "102 1st Ave" ...
 $ country     : chr  "USA" "USA" "USA" "USA" ...
 $ last_checked: chr  "Checked 142 minutes ago" "Checked 142 minutes ago" "Checked 142 minutes ago" "Checked 142 minutes ago" ...
```

As can be seen, the format of the variables isn’t correct. Therefore, I corrected them.

```
#correcting data types
data$is_broken <- as.factor(data$is_broken)
data$is_active <- as.factor(data$is_active)
data$dot<-as.factor(data$dot)
data$state <-as.factor(data$state)
data$city<-as.factor(data$city)
data$country <-as.factor(data$country)
```

And then, I deleted the unnecessary variables.

```
#Deleting unused columns
drops <- c("lat","lon","alt","street",'last_checked')
data <- data[ , !(names(data) %in% drops)]
```

Then I subset the USA because the State variable is only useful in the USA.

```
#sub-setting the data set by country only for USA
udata <- subset(data, country == "USA")
```

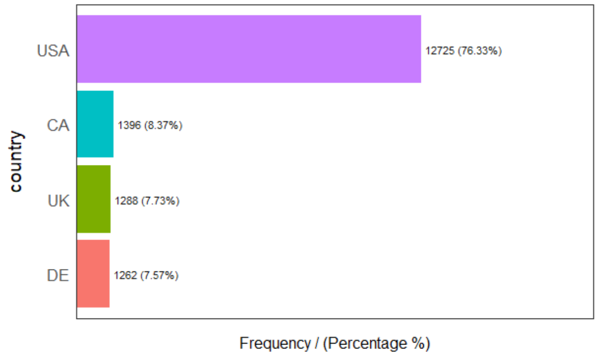
After this, I can now look at the summary and frequency of the data.

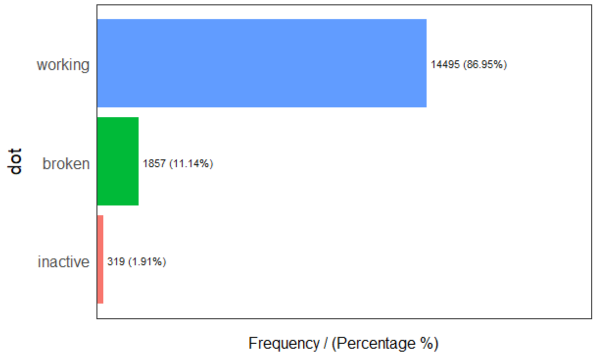
```
> summary(data)
 is_broken     is_active           dot            state               city      
 False:14814   False:  319   broken  : 1857          :3946   Houston    :  119  
 True : 1857   True :16352   inactive:  319   CA     :1158   Chicago    :  105  
                             working :14495   TX     :1056   San Antonio:   66  
                                              FL     : 819   Las Vegas  :   64  
                                              IL     : 626   Dallas     :   61  
                                              OH     : 585   Los Angeles:   60  
                                              (Other):8481   (Other)    :16196  
 country     
 CA : 1396  
 DE : 1262  
 UK : 1288  
 USA:12725  
```

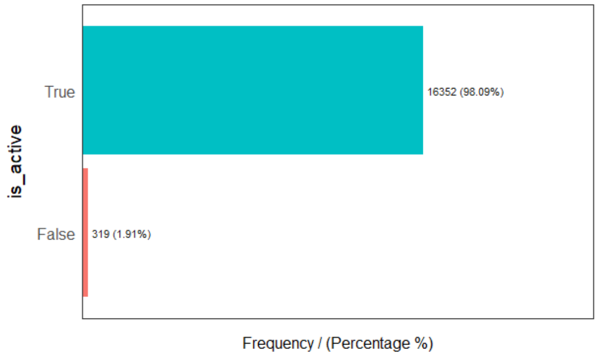
```
> freq(data[ , !(names(data)== "city")])
  is_broken frequency percentage cumulative_perc
1     False     14814      88.86           88.86
2      True      1857      11.14          100.00

  is_active frequency percentage cumulative_perc
1      True     16352      98.09           98.09
2     False       319       1.91          100.00

       dot frequency percentage cumulative_perc
1  working     14495      86.95           86.95
2   broken      1857      11.14           98.09
3 inactive       319       1.91          100.00

        state frequency percentage cumulative_perc
1                  3946      23.67           23.67
2          CA      1158       6.95           30.62
3          TX      1056       6.33           36.95
4          FL       819       4.91           41.86
5          IL       626       3.76           45.62
6          OH       585       3.51           49.13
7          NY       560       3.36           52.49
8          MI       501       3.01           55.50
9          PA       456       2.74           58.24
10         NC       440       2.64           60.88
11         GA       421       2.53           63.41
12         VA       380       2.28           65.69
13         IN       329       1.97           67.66
14         TN       308       1.85           69.51
15         MO       295       1.77           71.28
16         WI       281       1.69           72.97
17         MD       256       1.54           74.51
18         AZ       253       1.52           76.03
19         KY       243       1.46           77.49
20         NJ       242       1.45           78.94
21         WA       236       1.42           80.36
22         AL       230       1.38           81.74
23         MA       222       1.33           83.07
24         LA       217       1.30           84.37
25         SC       215       1.29           85.66
26         MN       209       1.25           86.91
27         CO       184       1.10           88.01
28         OK       182       1.09           89.10
29         AR       159       0.95           90.05
30         OR       148       0.89           90.94
31         IA       137       0.82           91.76
32         KS       136       0.82           92.58
33         MS       134       0.80           93.38
34         CT       130       0.78           94.16
35         NV       117       0.70           94.86
36         UT       102       0.61           95.47
37         WV        99       0.59           96.06
38         NM        88       0.53           96.59
39         NE        76       0.46           97.05
40         HI        71       0.43           97.48
41         ME        57       0.34           97.82
```

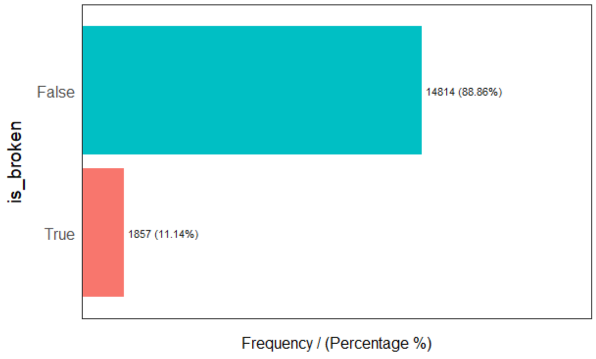
```
42         ID         56       0.34           98.16
43         NH         51       0.31           98.47
44         MT         45       0.27           98.74
45         DE         35       0.21           98.95
46         RI         29       0.17           99.12
47         SD         27       0.16           99.28
48         DC         25       0.15           99.43
49         WY         25       0.15           99.58
50         AK         24       0.14           99.72
51         VT         24       0.14           99.86
52         ND         23       0.14          100.00
53 Maharastra          3       0.02          100.00

  country frequency percentage cumulative_perc
1     USA     12725      76.33           76.33
2      CA      1396       8.37           84.70
3      UK      1288       7.73           92.43
4      DE      1262       7.57          100.00
```

The Following charts are also produced.









As can be seen, there is incorrect data in the State variable as Maharastra is not a US state. Also, there is a difference between is_active and is_broken, showing being broken is more common than being just inactive.

Before going to the next steps, I clean the State variable.

```
#deleting incorrect value in State
datan <- data %>%
  na_if("Maharastra")
```

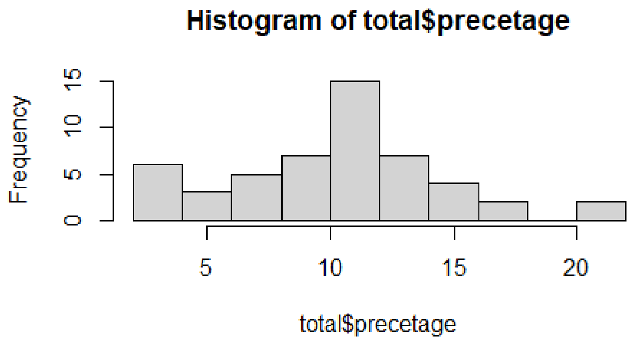
Now we can calculate some summary for the cleaned dataset.

```
#summerizing the data
cdata <-udata %>%
  group_by(state) %>%
  count(state)
cdata2 <-udata %>%
  group_by(state) %>%
  filter(is_broken == "True") %>%
  count(state)
total <- merge(cdata,cdata2,by="state")
total$precetage <-round(total$n.y/total$n.x,2)*100
total <-total[order(total$precetage, decreasing = TRUE), ]
head(total)
hist(total$precetage)
```

The result is shown below

```
head(total)
   state n.x n.y precetage
26    MS 134  28        21
1     AK  24   5        21
27    MT  45   8        18
38    OR 148  25        17
51    WY  25   4        16
11    GA 421  62        15
```

And the following histogram is created.



Explaining the findings and their importance

As can be seen, the percentage of broken machines in the US has a normal distribution with a mean of around 11 percent. The maximum percentage of broken machines is around 20 percent, and it belongs to Mississippi state. Also, despite popular belief, most Ice cream machines world wild are active. This raises the question then why the public thinks broken Ice cream machines are common in MacDonald's.

Q&A

- What did you do with the data in the context of exploration?

the data has been cleaned grouped and summarized in this assignment

- How many entries are in the dataset?

There are 12725 records from US which is used for analytics

- Was there missing data? Duplications? How clean was the data?

there where no missing data in the dataset

- Were there outliers or suspicious data?

there were 3 outliers in the data set and they have been replaced with NA

- What did you find? What intrigued you about the data? Why does that matter?

I found that the always broken Icecream machine is a myth

- What would your proposed next steps be? How do you plan to approach the cleansing of the data?

Some psychologist might find it interesting to look into why this pulp fiction exists

Bibliography

How to use R to display distributions of data and statistics. (2022). Influentialpoints.com. https://influentialpoints.com/Critiques/displaying_distributions_using_R.htm

How to Sort a DataFrame in R ? - GeeksforGeeks. (2021, May 27). GeeksforGeeks. https://www.geeksforgeeks.org/how-to-sort-a-dataframe-in-r/

Appendix

```
print('Mohammad Hossein Movahedi')
print('Assignment 1 AlY 6040')
#loading needed libraries
install.packages("tidyverse")
install.packages("funModeling")
install.packages("Hmisc")
install.packages('magrittr')
install.packages('FSA')
install.packages('FSAdata')
install.packages('magrittr')
install.packages('dplyr')
install.packages('tidyr')
install.packages('plyr')
install.packages('tidyverse')
install.packages('outliers')
install.packages('ggplot2')
install.packages('lubridate')
install.packages('corrplot')
library(funModeling)
library(tidyverse)
library(Hmisc)
library(magrittr)
library(ggplot2)
library(corrplot)
library(dplyr)
#reading the dataset
data <- read.csv('mcdonalds_dataset.csv')
summary(data)
status(data)
#looking into structure of the data
str(data)
#correcting data types
data$is_broken <- as.factor(data$is_broken)
```

```
data$is_active <- as.factor(data$is_active)
data$dot<-as.factor(data$dot)
data$state <-as.factor(data$state)
data$city<-as.factor(data$city)
data$country <-as.factor(data$country)
#Deleting unused columns
drops <- c("lat","lon","alt","street",'last_checked')
data <- data[ , !(names(data) %in% drops)]
#sub-setting the data set by country only for USA
udata <- subset(data, country == "USA")

# now we look at summery and freq
summary(data)
freq(data[ , !(names(data)== "city")])
#deleting incorrect value in State
datan <- data %>%
  na_if("Maharastra")
freq(datan)

#summerizing the data
cdata <-udata %>%
  group_by(state) %>%
  count(state)
cdata2 <-udata %>%
  group_by(state) %>%
  filter(is_broken == "True") %>%
  count(state)
total <- merge(cdata,cdata2,by="state")
total$precetage <-round(total$n.y/total$n.x,2)*100
total <-total[order(total$precetage, decreasing = TRUE), ]
head(total)
hist(total$precetage)
```