Assignment_09_SinghalSarika_

Sarika Singhal

Aug 07, 2021

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com/).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```r
setwd("~/Documents/repo/Week2/Week 2 - R/hello-world/week9")

#install.packages("readxl")
library("readxl")
#read_excel("week-7-housing.xlsx")

#the excel sheet below represent Crypto Current Market Cap Data

currencydata <- read_excel("allcurrenciesfinal12.18.17.xlsx")

#here is the structure of the data

summary(currencydata)
```

```
##  Currencyname            Date                          MarketCap
##  Length:535168      Min.   :2013-12-27 00:00:00   Min.   :0.0
00e+00
##  Class :character   1st Qu.:2015-09-27 00:00:00   1st Qu.:1.7
15e+04
##  Mode  :character   Median :2016-10-01 00:00:00   Median :1.0
81e+05
##                     Mean   :2016-07-14 05:40:24   Mean   :7.1
69e+07
##                     3rd Qu.:2017-06-15 00:00:00   3rd Qu.:9.7
01e+05
##                     Max.   :2017-11-24 00:00:00   Max.   :1.3
74e+11
##                     NA's   :13496                 NA's   :134
96
##      Close               Open               High
Low
##  Min.   :     0.0   Min.   :      0.0   Min.   :      0.0   M
in.   :     0.0
##  1st Qu.:     0.0   1st Qu.:      0.0   1st Qu.:      0.0   1
st Qu.:     0.0
##  Median :     0.0   Median :      0.0   Median :      0.0   M
edian :     0.0
##  Mean   :    88.5   Mean   :     90.1   Mean   :    102.3   M
ean   :    77.7
##  3rd Qu.:     0.1   3rd Qu.:      0.1   3rd Qu.:      0.1   3
rd Qu.:     0.1
##  Max.   :793273.0   Max.   :1013620.0   Max.   :1146320.0   M
ax.   :732467.0
##  NA's   :13496      NA's   :13496       NA's   :13496         N
A's   :13496
##      Volume
##  Min.   :0.000e+00
##  1st Qu.:2.200e+01
##  Median :3.160e+02
##  Mean   :2.111e+06
##  3rd Qu.:5.952e+03
##  Max.   :8.957e+09
##  NA's   :13496
```

```
#Data preparation and cleansing steps.

# 1. Familiarize yourself with the data set

file.info("allcurrenciesfinal12.18.17.xlsx")$size
```

```
## [1] 33921675
```

```
#File Size - 33921675 bytes

#an initial look at the data frame
str(currencydata)
```

```
## tibble [535,168 × 8] (S3: tbl_df/tbl/data.frame)
##  $ Currencyname: chr [1:535168] "0x" "0x" "0x" "0x" ...
##  $ Date        : POSIXct[1:535168], format: "2017-08-16" "201
7-08-17" ...
##  $ MarketCap   : num [1:535168] 6.70e+07 1.34e+08 1.23e+08 1.
77e+08 2.83e+08 ...
##  $ Close       : num [1:535168] 0.224 0.207 0.293 0.479 0.424
...
##  $ Open        : num [1:535168] 0.112 0.223 0.206 0.295 0.471
...
##  $ High        : num [1:535168] 0.28 0.239 0.35 0.544 0.475
...
##  $ Low         : num [1:535168] 0.104 0.207 0.206 0.284 0.403
...
##  $ Volume      : num [1:535168] 5232600 2752410 12793800 5267
7500 16016500 ...
```

```
#2 . Check for structural errors  - we’ll evaluate the data fram
e for structural errors. These include entry errors such as faul
ty data types, non-unique ID numbers, mislabeled variables, and
 string inconsistencies.
#If there are more structural pitfalls in your own dataset than
 the ones covered below, be sure to include additional steps in
 your data cleaning to address the idiosyncrasies.

#install.packages("dplyr")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
currencydata <- currencydata %>% rename(CryptoCurrencyname = Cur
rencyname)

#Examine if datatypes are faulty
typeof(currencydata$MarketCap)
```

```
## [1] "double"
```

```
#Non-unique ID numbers - In this dataset uniqueness is not a problem

#3 .Check for data irregularities, like invalid values and outliers.

summary(currencydata)
```

```
##  CryptoCurrencyname      Date                           MarketCap
##  Length:535168      Min.   :2013-12-27 00:00:00   Min.   :0.0
00e+00
##  Class :character   1st Qu.:2015-09-27 00:00:00   1st Qu.:1.7
15e+04
##  Mode  :character   Median :2016-10-01 00:00:00   Median :1.0
81e+05
##                     Mean   :2016-07-14 05:40:24   Mean   :7.1
69e+07
##                     3rd Qu.:2017-06-15 00:00:00   3rd Qu.:9.7
01e+05
##                     Max.   :2017-11-24 00:00:00   Max.   :1.3
74e+11
##                     NA's   :13496                 NA's   :134
96
##      Close              Open               High
Low
##  Min.   :     0.0   Min.   :      0.0   Min.   :      0.0   M
in.   :     0.0
##  1st Qu.:     0.0   1st Qu.:      0.0   1st Qu.:      0.0   1
st Qu.:     0.0
##  Median :     0.0   Median :      0.0   Median :      0.0   M
edian :     0.0
##  Mean   :    88.5   Mean   :     90.1   Mean   :    102.3   M
ean   :    77.7
##  3rd Qu.:     0.1   3rd Qu.:      0.1   3rd Qu.:      0.1   3
rd Qu.:     0.1
##  Max.   :793273.0   Max.   :1013620.0   Max.   :1146320.0   M
ax.   :732467.0
##  NA's   :13496      NA's   :13496       NA's   :13496        N
A's   :13496
##      Volume
##  Min.   :0.000e+00
##  1st Qu.:2.200e+01
##  Median :3.160e+02
##  Mean   :2.111e+06
##  3rd Qu.:5.952e+03
##  Max.   :8.957e+09
##  NA's   :13496
```

```
#Data look ok

#4: Decide how to deal with missing values

sum(is.na(currencydata))
```

```
## [1] 107968
```

```
#percent missing values per variable
apply(currencydata, 2, function(col)sum(is.na(col))/length(col))
```

```
## CryptoCurrencyname               Date          MarketCap
Close
##         0.02521825         0.02521825         0.02521825
0.02521825
##               Open               High                Low
Volume
##         0.02521825         0.02521825         0.02521825
0.02521825
```

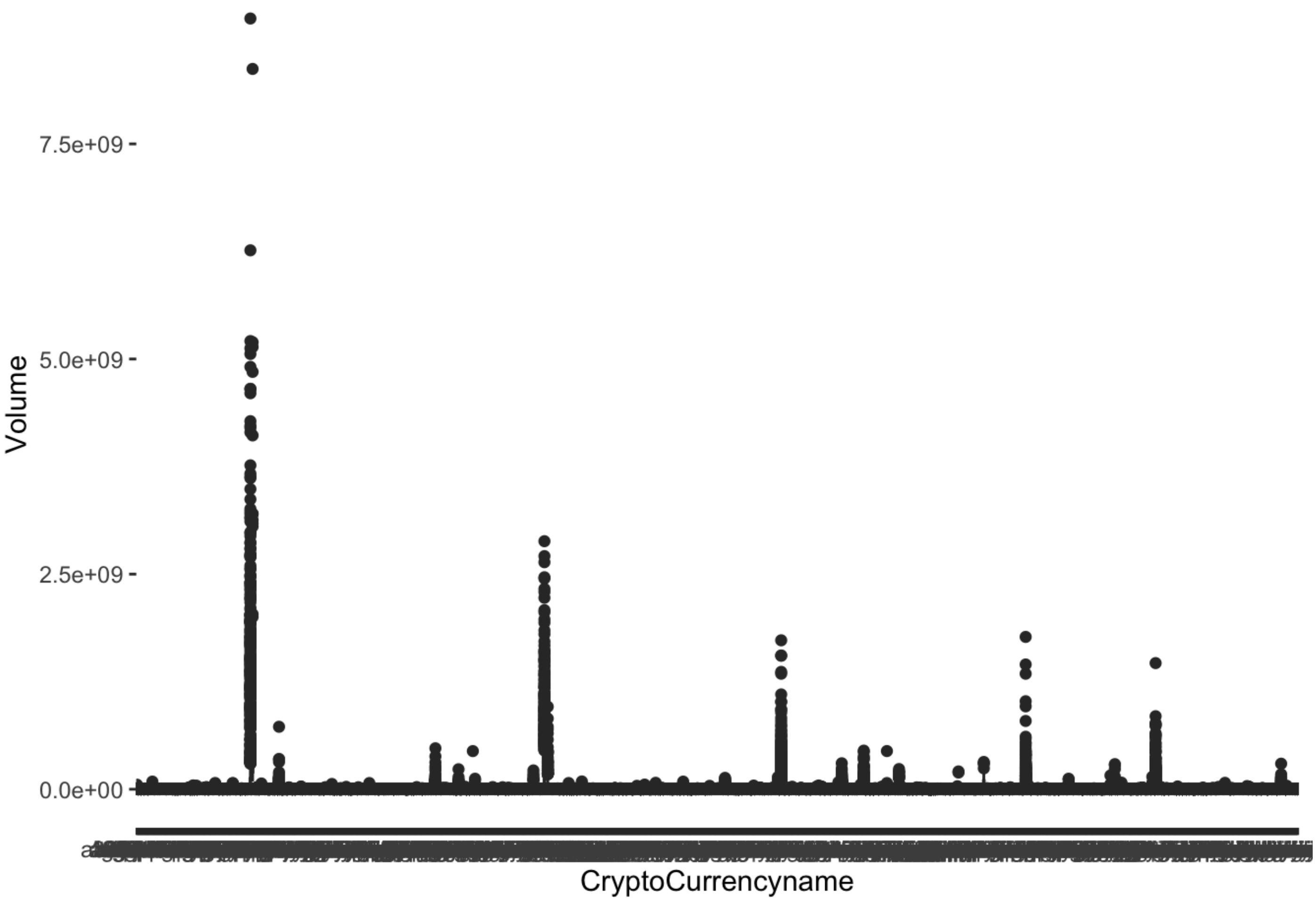
```
#identifying the rows with NAs
currencydata_NA <- rownames(currencydata)[apply(currencydata, 2,
anyNA)]

summary(currencydata_NA)
```

```
##    Length     Class      Mode
##    535168 character character
```

```
#removing all observations with NAs
currencydata_clean <- currencydata %>% na.omit()

#Clean Data Set
summary(currencydata_clean)
```

```
##  CryptoCurrencyname      Date                       MarketCap
##  Length:521672      Min.   :2013-12-27 00:00:00   Min.   :0.0
00e+00
##  Class :character   1st Qu.:2015-09-27 00:00:00   1st Qu.:1.7
15e+04
##  Mode  :character   Median :2016-10-01 00:00:00   Median :1.0
81e+05
##                     Mean   :2016-07-14 05:40:24   Mean   :7.1
69e+07
##                     3rd Qu.:2017-06-15 00:00:00   3rd Qu.:9.7
01e+05
##                     Max.   :2017-11-24 00:00:00   Max.   :1.3
74e+11
##      Close               Open                High
Low
##  Min.   :     0.0   Min.   :      0.0   Min.   :      0.0   M
in.   :     0.0
##  1st Qu.:     0.0   1st Qu.:      0.0   1st Qu.:      0.0   1
st Qu.:     0.0
##  Median :     0.0   Median :      0.0   Median :      0.0   M
edian :     0.0
##  Mean   :    88.5   Mean   :     90.1   Mean   :    102.3   M
ean   :    77.7
##  3rd Qu.:     0.1   3rd Qu.:      0.1   3rd Qu.:      0.1   3
rd Qu.:     0.1
##  Max.   :793273.0   Max.   :1013620.0   Max.   :1146320.0   M
ax.   :732467.0
##      Volume
##  Min.   :0.000e+00
##  1st Qu.:2.200e+01
##  Median :3.160e+02
##  Mean   :2.111e+06
##  3rd Qu.:5.952e+03
##  Max.   :8.957e+09
```

```r
#Discuss how you plan to uncover new information in the data tha
t is not self-evident.

#install.packages("ggplot2")
library(ggplot2)

ggplot(data = currencydata_clean, aes(x=CryptoCurrencyname,y=Vol
ume)) + geom_boxplot()
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

References

install.packages(“knitr”)