Assignment\_10\_SinghalSarika\_ProjectMilestone3

Sarika Singhal

Aug 12, 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>.

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:

setwd("~/Documents/repo/Week2/Week 2 - R/hello-world/week9")  
  
#Introduction:  
  
## Initial appearance of Cryptocurrency/Bitcoin/altcoin happened in 2009.  
#Now we know it’s going to stay. Question being addressed is what the right percentage of crypto should be in one’s investment portfolio.  
  
  
#install.packages("readxl")  
library("readxl")  
  
#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. :0.000e+00   
## Class :character 1st Qu.:2015-09-27 00:00:00 1st Qu.:1.715e+04   
## Mode :character Median :2016-10-01 00:00:00 Median :1.081e+05   
## Mean :2016-07-14 05:40:24 Mean :7.169e+07   
## 3rd Qu.:2017-06-15 00:00:00 3rd Qu.:9.701e+05   
## Max. :2017-11-24 00:00:00 Max. :1.374e+11   
## NA's :13496 NA's :13496   
## Close Open High Low   
## Min. : 0.0 Min. : 0.0 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.0   
## Median : 0.0 Median : 0.0 Median : 0.0 Median : 0.0   
## Mean : 88.5 Mean : 90.1 Mean : 102.3 Mean : 77.7   
## 3rd Qu.: 0.1 3rd Qu.: 0.1 3rd Qu.: 0.1 3rd Qu.: 0.1   
## Max. :793273.0 Max. :1013620.0 Max. :1146320.0 Max. :732467.0   
## NA's :13496 NA's :13496 NA's :13496 NA'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

#The problem statement addressed.   
  
#This data set shows the MarketCap for various Crypto Currencies, date wise High low and Open and close.  
#This data can help us understand how each curries has changed in last few years and how the market cap has been for each of them.  
#That can help us anlayze how volatile each one of them is and answer few research question:  
#What are top cryptocurrency to be invested in and why?   
#What mix of cryptocurrency in portfolio is best? To determine risky vs safe  
#How much can cryptocurrency appreciate?  
  
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" "2017-08-17" ...  
## $ 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 52677500 16016500 ...

#How you addressed this problem statement  
#First data needs to be cleaned   
#Check for structural errors - we’ll evaluate the data frame for structural errors. These include entry errors such as faulty 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 = Currencyname)  
  
#Examine if datatypes are faulty  
typeof(currencydata$MarketCap)

## [1] "double"

#Non-unique ID numbers - In this dataset uniqueness is not a problem  
  
#Check for data irregularities, like invalid values and outliers.   
  
summary(currencydata)

## CryptoCurrencyname Date MarketCap   
## Length:535168 Min. :2013-12-27 00:00:00 Min. :0.000e+00   
## Class :character 1st Qu.:2015-09-27 00:00:00 1st Qu.:1.715e+04   
## Mode :character Median :2016-10-01 00:00:00 Median :1.081e+05   
## Mean :2016-07-14 05:40:24 Mean :7.169e+07   
## 3rd Qu.:2017-06-15 00:00:00 3rd Qu.:9.701e+05   
## Max. :2017-11-24 00:00:00 Max. :1.374e+11   
## NA's :13496 NA's :13496   
## Close Open High Low   
## Min. : 0.0 Min. : 0.0 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.0   
## Median : 0.0 Median : 0.0 Median : 0.0 Median : 0.0   
## Mean : 88.5 Mean : 90.1 Mean : 102.3 Mean : 77.7   
## 3rd Qu.: 0.1 3rd Qu.: 0.1 3rd Qu.: 0.1 3rd Qu.: 0.1   
## Max. :793273.0 Max. :1013620.0 Max. :1146320.0 Max. :732467.0   
## NA's :13496 NA's :13496 NA's :13496 NA'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  
  
#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. :0.000e+00   
## Class :character 1st Qu.:2015-09-27 00:00:00 1st Qu.:1.715e+04   
## Mode :character Median :2016-10-01 00:00:00 Median :1.081e+05   
## Mean :2016-07-14 05:40:24 Mean :7.169e+07   
## 3rd Qu.:2017-06-15 00:00:00 3rd Qu.:9.701e+05   
## Max. :2017-11-24 00:00:00 Max. :1.374e+11   
## Close Open High Low   
## Min. : 0.0 Min. : 0.0 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.0   
## Median : 0.0 Median : 0.0 Median : 0.0 Median : 0.0   
## Mean : 88.5 Mean : 90.1 Mean : 102.3 Mean : 77.7   
## 3rd Qu.: 0.1 3rd Qu.: 0.1 3rd Qu.: 0.1 3rd Qu.: 0.1   
## Max. :793273.0 Max. :1013620.0 Max. :1146320.0 Max. :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

head(currencydata\_clean)

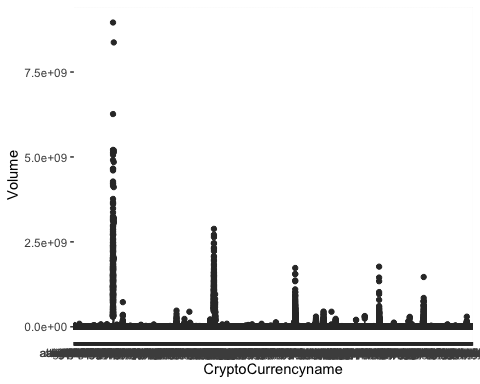
## # A tibble: 6 x 8  
## CryptoCurrencyna… Date MarketCap Close Open High Low Volume  
## <chr> <dttm> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0x 2017-08-16 00:00:00 67034800 0.224 0.112 0.280 0.104 5.23e6  
## 2 0x 2017-08-17 00:00:00 133813000 0.207 0.223 0.239 0.207 2.75e6  
## 3 0x 2017-08-18 00:00:00 123335000 0.293 0.206 0.350 0.206 1.28e7  
## 4 0x 2017-08-19 00:00:00 176747000 0.479 0.295 0.544 0.284 5.27e7  
## 5 0x 2017-08-20 00:00:00 282777000 0.424 0.471 0.475 0.403 1.60e7  
## 6 0x 2017-08-21 00:00:00 253364000 0.352 0.422 0.444 0.328 1.50e7

#max(currencydata\_clean$MarketCap)  
  
  
#install.packages("matrixStats")  
library(matrixStats)

##   
## Attaching package: 'matrixStats'

## The following object is masked from 'package:dplyr':  
##   
## count

#currencydata\_clean %>% rowwise() %>% mutate(Maximum\_price = max(c(currencydata\_clean$MarketCap)))  
  
##Summarize how you addressed this problem statement (the data used and the methodology employed, including a recommendation for a model that could be implemented).   
  
#What are different ways you could look at this data to answer the questions you want to answer?  
  
#There are various steps followed below to uncover the data   
  
#Discuss how you plan to uncover new information in the data that is not self-evident.  
  
#install.packages("ggplot2")  
library(ggplot2)  
  
ggplot(data = currencydata\_clean, aes(x=CryptoCurrencyname,y=Volume)) + geom\_boxplot()



##Summarize how you addressed this problem statement (the data used and the methodology employed, including a recommendation for a model that could be implemented).   
  
#What are different ways you could look at this data to answer the questions you want to answer?  
  
#There are various steps followed below to uncover the data   
  
### Call the functions on currencydata\_clean to examine the data frame  
dim(currencydata\_clean)

## [1] 521672 8

str(currencydata\_clean)

## tibble [521,672 × 8] (S3: tbl\_df/tbl/data.frame)  
## $ CryptoCurrencyname: chr [1:521672] "0x" "0x" "0x" "0x" ...  
## $ Date : POSIXct[1:521672], format: "2017-08-16" "2017-08-17" ...  
## $ MarketCap : num [1:521672] 6.70e+07 1.34e+08 1.23e+08 1.77e+08 2.83e+08 ...  
## $ Close : num [1:521672] 0.224 0.207 0.293 0.479 0.424 ...  
## $ Open : num [1:521672] 0.112 0.223 0.206 0.295 0.471 ...  
## $ High : num [1:521672] 0.28 0.239 0.35 0.544 0.475 ...  
## $ Low : num [1:521672] 0.104 0.207 0.206 0.284 0.403 ...  
## $ Volume : num [1:521672] 5232600 2752410 12793800 52677500 16016500 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:13496] 4376 7833 7834 7835 7836 7837 7838 7839 7840 7841 ...  
## ..- attr(\*, "names")= chr [1:13496] "4376" "7833" "7834" "7835" ...

summary(currencydata\_clean)

## CryptoCurrencyname Date MarketCap   
## Length:521672 Min. :2013-12-27 00:00:00 Min. :0.000e+00   
## Class :character 1st Qu.:2015-09-27 00:00:00 1st Qu.:1.715e+04   
## Mode :character Median :2016-10-01 00:00:00 Median :1.081e+05   
## Mean :2016-07-14 05:40:24 Mean :7.169e+07   
## 3rd Qu.:2017-06-15 00:00:00 3rd Qu.:9.701e+05   
## Max. :2017-11-24 00:00:00 Max. :1.374e+11   
## Close Open High Low   
## Min. : 0.0 Min. : 0.0 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.0 1st Qu.: 0.0   
## Median : 0.0 Median : 0.0 Median : 0.0 Median : 0.0   
## Mean : 88.5 Mean : 90.1 Mean : 102.3 Mean : 77.7   
## 3rd Qu.: 0.1 3rd Qu.: 0.1 3rd Qu.: 0.1 3rd Qu.: 0.1   
## Max. :793273.0 Max. :1013620.0 Max. :1146320.0 Max. :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

colnames(currencydata\_clean)

## [1] "CryptoCurrencyname" "Date" "MarketCap"   
## [4] "Close" "Open" "High"   
## [7] "Low" "Volume"

#Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.  
library("dplyr")  
step\_1\_df <- select(currencydata, -Volume)  
dim(step\_1\_df)

## [1] 535168 7

step\_1\_df\_novolume <- head(step\_1\_df, n = 10)  
#Arrange  
str(step\_1\_df\_novolume)

## tibble [10 × 7] (S3: tbl\_df/tbl/data.frame)  
## $ CryptoCurrencyname: chr [1:10] "0x" "0x" "0x" "0x" ...  
## $ Date : POSIXct[1:10], format: "2017-08-16" "2017-08-17" ...  
## $ MarketCap : num [1:10] 6.70e+07 1.34e+08 1.23e+08 1.77e+08 2.83e+08 ...  
## $ Close : num [1:10] 0.224 0.207 0.293 0.479 0.424 ...  
## $ Open : num [1:10] 0.112 0.223 0.206 0.295 0.471 ...  
## $ High : num [1:10] 0.28 0.239 0.35 0.544 0.475 ...  
## $ Low : num [1:10] 0.104 0.207 0.206 0.284 0.403 ...

##Model Suggestion  
  
## As per the analysis I am trying to do, Regression Model works best for this case as that will help  
## Forecast the future as per the problem statement  
## Analyze Market returns  
## Demand  
  
#How could you summarize your data to answer key questions?  
  
#filter\_df <- currencydata %>% filter(Currencyname == "dixasset" | Currencyname == "acoin" | Currencyname == "vector" | Currencyname == "ufocoin" )  
  
  
  
##Running a simple model on the data set to see Marketup predictions  
  
summary(glm(MarketCap ~ 1, data = currencydata))

##   
## Call:  
## glm(formula = MarketCap ~ 1, data = currencydata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -7.169e+07 -7.167e+07 -7.158e+07 -7.072e+07 1.374e+11   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 71687153 2159322 33.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 2.432386e+18)  
##   
## Null deviance: 1.2689e+24 on 521671 degrees of freedom  
## Residual deviance: 1.2689e+24 on 521671 degrees of freedom  
## (13496 observations deleted due to missingness)  
## AIC: 23565639  
##   
## Number of Fisher Scoring iterations: 2

## Intresting Insights : Which coins can help diversify your portfolio?  
##It is notable that almost all of the cryptocurrencies have become more correlated with each other across the board. In practice, when investing in the stock market, having a diversified portfolio reduces risk. However, having a diversified portfolio of the top 20 coins, does not have the same effect as in trading stocks.

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”)