EDA

Malathi

Feb 22, 2018

About EDA

R Programming package DataEXplorer

In Data Science, 80% of time is spent on preparing the data. The package DataExplorerin R does basic EDA with just one function create_re

Installation and Loading

Let us begin our EDA by loading the library: Install if the package doesn’t exist

Dataset

The dataset that we will be using for this analysis is Chocolate Bar Ratings. Loading input dataset into our R session for EDA:

```
choco = read.csv('D:/dataset/choco/flavors_of_cacao.csv', header = T, stringsAsFactors = F)
```

Data Cleaning

Some reformatting of data types are required before proceeding.For example, Cocoa.Percent is supposed to be a numeric value but read as a character due to the presence of % symbol, so needs to be fixed.

```
choco$Cocoa.Percent <- as.numeric(gsub('%','',choco$Cocoa.Percent))
choco$Review.Date = as.character(choco$Review.Date)
```

Variables

Get the names and data types of the variables.

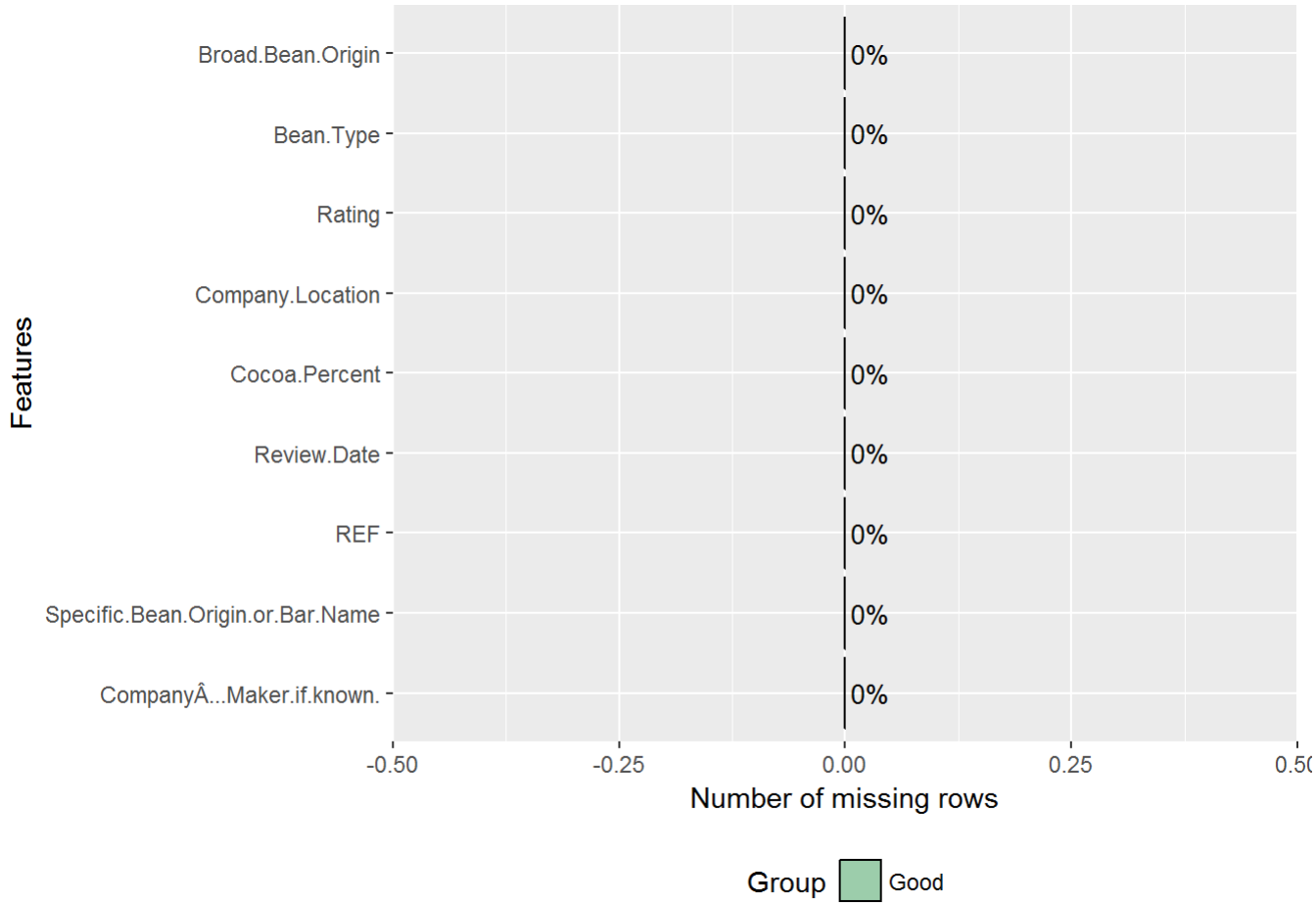
```
plot_str(choco)
```

We've got some Continuous variables and some Categorical variables.

Search for Missing Values

It's very important to see if the input data given for Analysis has got Missing values before diving deep into the analysis.

```
plot_missing(choco)
```

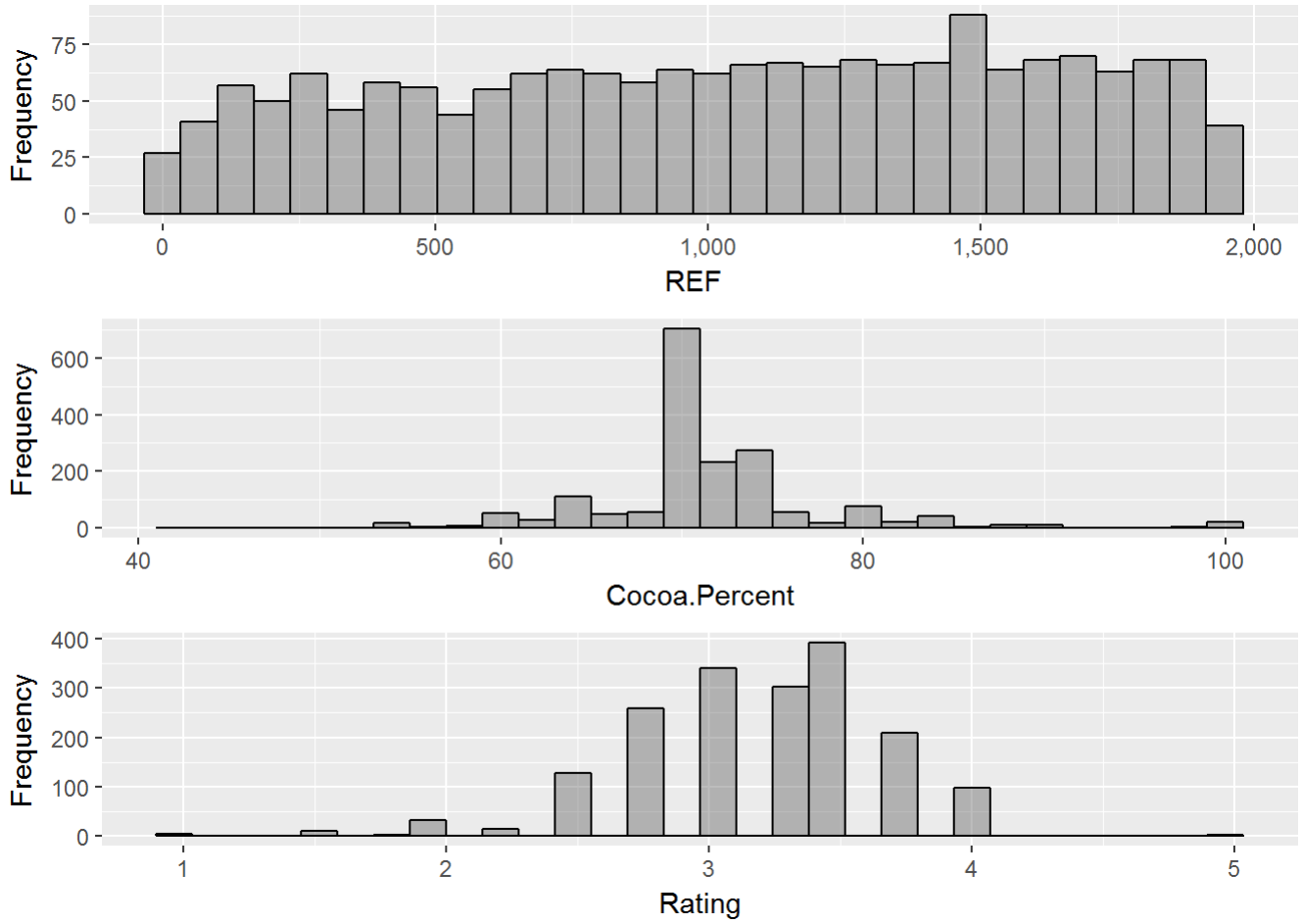


There's no missing value in this dataset.

Continuous Variables

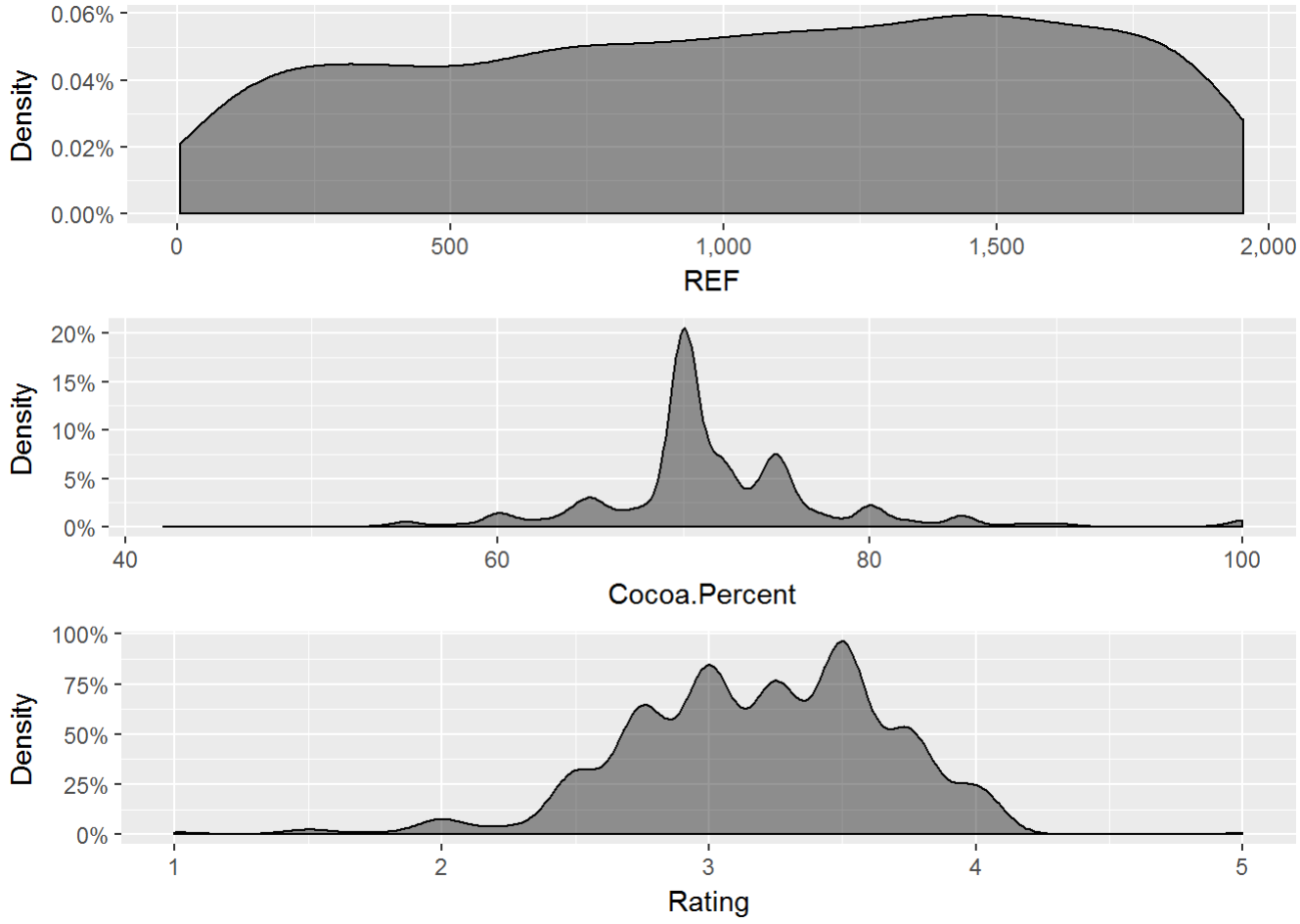
Histogram is analyst's best friend to analyse/represent Continuous Variables.

```
plot_histogram(choco)
```



DataExplorer has got a function for the density plot.

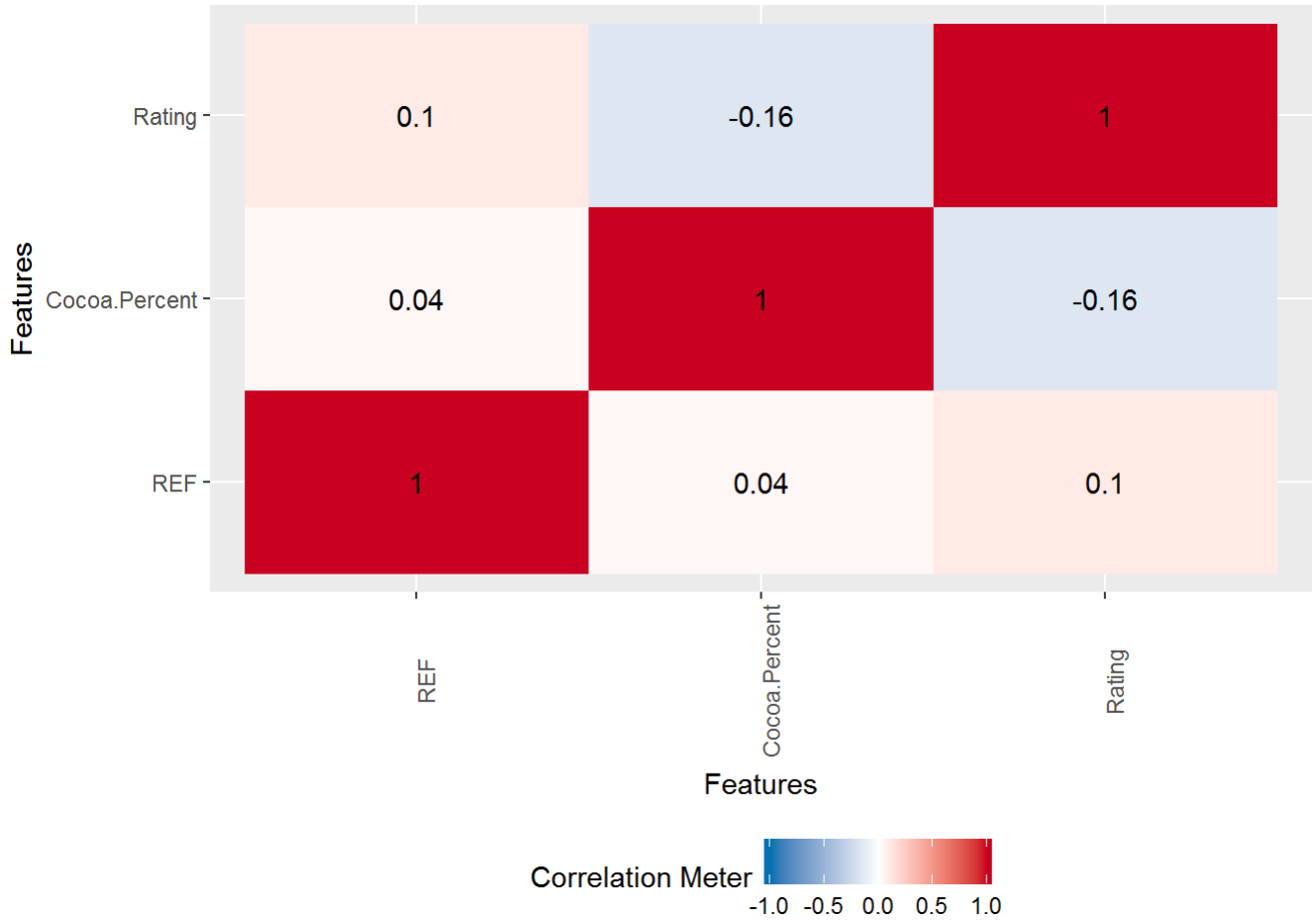
```
plot_density(choco)
```



Multivariate Analysis

That marks the end of univariate analysis and the beginning of bivariate/multivariate analysis, starting with Correlation analysis.

```
plot_correlation(choco, type = 'continuous','Review.Date')
```



##Gives this plot:

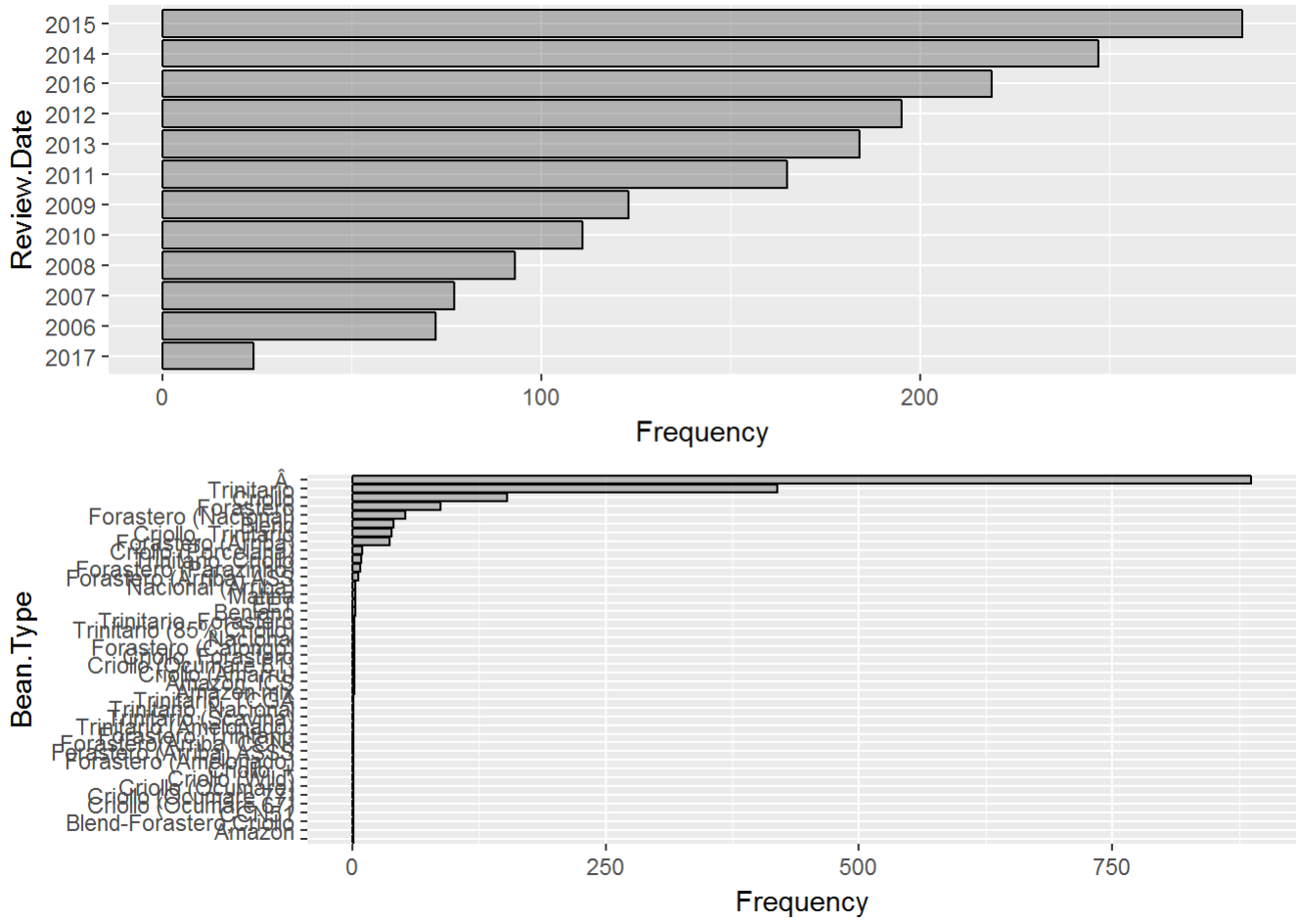
Similar to the correlation plot, DataExplorer has got functions to plot boxplot and scatterplot with similar syntax as above.

Categorical Variables???-???Barplots

So far we've seen the kind of EDA plots that DataExplorer lets us plot for Continuous variables and now let us see how we can do similar exercise for categorical variables. Unexpectedly, this becomes one very simple function

```
plot_bar(choco)
```

```
## 4 columns ignored with more than 50 categories.
## CompanyÂ...Maker.if.known.: 416 categories
## Specific.Bean.Origin.or.Bar.Name: 1039 categories
## Company.Location: 60 categories
## Broad.Bean.Origin: 101 categories
```



And finally, if you have got only a couple of minutes (just like in the maggi noodles ad, 2 mins!) just keep it simple to use create_report() that gives a very nice presentable/shareable rendered markdown in html. Hope this article helps you perform simple and fast EDA and generate shareable report with typical EDA elements. To learn more about Exploratory Data Analysis in R, check out this DataCamp Course.

References

https://cran.r-project.org/web/packages/DataExplorer/index.html (https://cran.r-project.org/web/packages/DataExplorer/index.html)

```
##
##
## processing file: report.rmd
```

```
##
  |
  |                                                                      |   0%
  |
  |.....                                                                 |   7%
##   ordinary text without R code
##
##
  |
  |.........                                                             |  14%
## label: global_options (with options)
## List of 1
##  $ include: logi FALSE
##
##
  |
  |...............                                                       |  21%
##    inline R code fragments
##
##
  |
  |....................                                                  |  29%
## label: data_structure
##
  |
  |.........................                                             |  36%
##   ordinary text without R code
##
##
  |
  |...............................                                       |  43%
## label: plot_data_structure
##
  |
  |.................................                                     |  50%
##   ordinary text without R code
##
##
  |
  |.....................................                                 |  57%
## label: missing_values
```

```
##
  |
  |.........................................                             |  64%
##   ordinary text without R code
##
##
  |
  |.............................................                         |  71%
## label: histogram
```

```
##
  |
  |.................................................              |  79%
##   ordinary text without R code
##
##
  |
  |.....................................................          |  86%
## label: bar
```

```
##
  |
  |.........................................................      |  93%
##   ordinary text without R code
##
##
  |
  |.............................................................| 100%
## label: correlation
```

```
## output file: c:/Users/DELL/Desktop/report.knit.md
```

```
## "D:/Program Files/RStudio/bin/pandoc/pandoc" +RTS -K512m -RTS "c:/Users/DELL/Desktop/report.u
tf8.md" --to html --from markdown+autolink_bare_uris+ascii_identifiers+tex_math_single_backslash
--output pandoc15407a5a4653.html --smart --email-obfuscation none --self-contained --standalone
--section-divs --table-of-contents --toc-depth 6 --template "d:\PROGRA~2\R\R-34~1.2\library\RMAR
KD~1\rmd\h\DEFAUL~1.HTM" --no-highlight --variable highlightjs=1 --variable "theme:cerulean" --i
nclude-in-header "C:\Users\DELL\AppData\Local\Temp\Rtmpm8ZaCV\rmarkdown-str15403fe17d45.html" --
mathjax --variable "mathjax-url:https://mathjax.rstudio.com/latest/MathJax.js?config=TeX-AMS-MML
_HTMLorMML"
```

```
##
## Output created: report.html
```

```
##
##
## Report is generated at "c:/Users/DELL/Desktop/report.html".
```