# Context

My habit has been to utilize one or two functions in a package without investigating other functionality. In this series I’m testing the idea of breaking that habit.

Each post will include how I was using a package integrated with a case-study that illustrates newly discovered functions.

# DataExplorer {package}

You can tell by the name of my blog that {DataExplorer} is perfectly suited for this series on R packages.



Exploratory Data Analysis (EDA) is the initial and an important phase of data analysis/predictive modeling. During this process, analysts/modelers will have a first look of the data, and thus generate relevant hypotheses and decide next steps. However, the EDA process could be a hassle at times. This R package aims to automate most of data handling and visualization, so that users could focus on studying the data and extracting insights.

Just about every time I’m working with new data, I’m loading {DataExplorer} from my library of R packages. However, I’m typically only using the plot\_missing() function.

While researching the package I was excited to discover functionality that has become core to my EDA process. In today’s case-study we will go over:

The function I use often: plot\_missing()

Newly discovered functions from {DataExplorer}

How {DataExplorer} provides insights that expedite EDA

* 1. Case-Study Setup

Let’s get started by loading our packages and importing a bit of data.

# Load Packages

# Core Packages library(tidyverse) library(tidyquant) library(recipes) library(rsample) library(knitr)

# Data Cleaning library(janitor)

# EDA

library(skimr) library(DataExplorer)

# ggplot2 Helpers library(scales) theme\_set(theme\_tq())

# Import Data

For our case-study we are using data from the Tidy Tuesday Project archive.

Each record represents bags of coffee that were assessed and “professionally rated on a 0-100 scale.” Each row has a score that originated from assessing X number of bags of coffee beans.

Out of the many features in the data set, there are 10 numeric metrics that when summed make up the coffee rating score (total\_cup\_points).

tuesdata <- tidytuesdayR::tt\_load(2020, week = 28)

##

## Downloading file 1 of 1: `coffee\_ratings.csv` coffee\_ratings\_tbl <- tuesdata$coffee\_ratings

# coffee\_ratings\_tbl <- read\_csv("static/01\_data/coffee\_ratings.csv")

# coffee\_ratings\_tbl <- read\_csv("../../static/01\_data/coffee\_ratings.csv")

# Data Caveats

If you have all 10 metrics then you don’t need a model to predict total\_cup\_points.

That said, this post is about preprocessing data in preparation for analysis and/or predictive modeling. I chose these data for the case-study because of the many characteristics and features present that will help illustrate the benefits of {DataExplorer}.

To illustrate the benefits, we assume total\_cup\_points is our target (dependent variable) and that all others are potential predictors (independent variables).

Let’s get to work!

# Preprocessing Pipeline

As usual, let’s setup our preprocessing data pipeline so that we can add to it as we gain insights. Read This Post to learn more about my approach to preprocessing data. coffee\_ratings\_preprocessed\_tbl <- coffee\_ratings\_tbl

# Case-Study Objectives

* + 1. Rapidly assess data.
    2. Gains insights that help preprocess data.

Let’s see how {DataExplorer} can expedite the process.

As usual, let’s take a glimpse() of our data to see how we should proceed.

coffee\_ratings\_preprocessed\_tbl %>% glimpse()

## Rows: 1,339

## Columns: 43

## $ total\_cup\_points 90.58, 89.92, 89.75, 89.00, 88.83, 88.83, 88.75…

## $ species "Arabica", "Arabica", "Arabica", "Arabica", "Ar…

## $ owner "metad plc", "metad plc", "grounds for health a… ## $ country\_of\_origin "Ethiopia", "Ethiopia", "Guatemala", "Ethiopia"… ## $ farm\_name "metad plc", "metad plc", "san marcos barrancas… ## $ lot\_number NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…

## $ mill "metad plc", "metad plc", NA, "wolensu", "metad… ## $ ico\_number "2014/2015", "2014/2015", NA, NA, "2014/2015", …

## $ company "metad agricultural developmet plc", "metad agr… ## $ altitude "1950-2200", "1950-2200", "1600 - 1800 m", "180…

## $ region "guji-hambela", "guji-hambela", NA, "oromia", "… ## $ producer "METAD PLC", "METAD PLC", NA, "Yidnekachew Dabe… ## $ number\_of\_bags 300, 300, 5, 320, 300, 100, 100, 300, 300, 50, …

## $ bag\_weight "60 kg", "60 kg", "1", "60 kg", "60 kg", "30 kg… ## $ in\_country\_partner "METAD Agricultural Development plc", "METAD Ag… ## $ harvest\_year "2014", "2014", NA, "2014", "2014", "2013", "20…

## $ grading\_date "April 4th, 2015", "April 4th, 2015", "May 31st… ## $ owner\_1 "metad plc", "metad plc", "Grounds for Health A…

## $ variety NA, "Other", "Bourbon", NA, "Other", NA, "Other… ## $ processing\_method "Washed / Wet", "Washed / Wet", NA, "Natural / …

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## | $ | aroma | 8.67, | 8.75, | 8.42, | 8.17, | 8.25, | | 8.58, | 8.42, | 8.25,… |
| ## | $ | flavor | 8.83, | 8.67, | 8.50, | 8.58, | 8.50, | | 8.42, | 8.50, | 8.33,… |
| ## | $ | aftertaste | 8.67, | 8.50, | 8.42, | 8.42, | 8.25, | | 8.42, | 8.33, | 8.50,… |
| ## | $ | acidity | 8.75, | 8.58, | 8.42, | 8.42, | 8.50, | | 8.50, | 8.50, | 8.42,… |
| ## | $ | body | 8.50, | 8.42, | 8.33, | 8.50, | 8.42, | | 8.25, | 8.25, | 8.33,… |
| ## | $ | balance | 8.42, | 8.42, | 8.42, | 8.25, | 8.33, | | 8.33, | 8.25, | 8.50,… |
| ## | $ | uniformity | 10.00, 10.00, 10.00, 10.00, | | | | | 10.00, 10.00, 10.00… | | | |
| ## | $ | clean\_cup | 10, 10, 10, 10, 10, 10, 10, | | | | | 10, 10, 10, 10, 10,… | | | |
| ## | $ | sweetness | 10.00, 10.00, 10.00, 10.00, | | | | | 10.00, 10.00, 10.00… | | | |
| ## | $ | cupper\_points | 8.75, | 8.58, | 9.25, | 8.67, | 8.58, | | 8.33, | 8.50, | 9.00,… |
| ## | $ | moisture | 0.12, | 0.12, | 0.00, | 0.11, | 0.12, | | 0.11, | 0.11, | 0.03,… |
| ## | $ | category\_one\_defects | 0, 0, | 0, 0, | 0, 0, | 0, 0, | 0, 0, | | 0, 0, | 0, 0, | 0, 0,… |
| ## | $ | quakers | 0, 0, | 0, 0, | 0, 0, | 0, 0, | 0, 0, | | 0, 0, | 0, 0, | 0, 0,… |

## $ color "Green", "Green", NA, "Green", "Green", "Bluish… ## $ category\_two\_defects 0, 1, 0, 2, 2, 1, 0, 0, 0, 4, 1, 0, 0, 2, 2, 0,… ## $ expiration "April 3rd, 2016", "April 3rd, 2016", "May 31st… ## $ certification\_body "METAD Agricultural Development plc", "METAD Ag… ## $ certification\_address "309fcf77415a3661ae83e027f7e5f05dad786e44", "30… ## $ certification\_contact "19fef5a731de2db57d16da10287413f5f99bc2dd", "19… ## $ unit\_of\_measurement "m", "m", "m", "m", "m", "m", "m", "m", "m", "m…

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## | $ | altitude\_low\_meters | 1950.0, | 1950.0, | 1600.0, | 1800.0, | 1950.0, | NA, | NA,… |
| ## | $ | altitude\_high\_meters | 2200.0, | 2200.0, | 1800.0, | 2200.0, | 2200.0, | NA, | NA,… |
| ## | $ | altitude\_mean\_meters | 2075.0, | 2075.0, | 1700.0, | 2000.0, | 2075.0, | NA, | NA,… |

## Wow, 43 columns!

Many of these are obviously unnecessary and so let’s get to work reducing these down to something more meaningful.

We can begin by removing a few columns and so lets add that step to our preprocessing.

coffee\_ratings\_preprocessed\_tbl <- coffee\_ratings\_tbl %>% # remove columns

select(-contains("certification"), -in\_country\_partner)

# Exploratory Data Analysis (EDA)

Integrating {DataExplorer} into our EDA process creates a work-flow that quickly assesses:

* + 1. Summary statistics: skimr::skim()
    2. Missing data: plot\_missing()
    3. Categorical data: plot\_bar()
    4. Numerical data: plot\_historgram

Once the data is assessed, we can decide on steps that might be added to a preprocessing data pipeline.

# Summary Statistics

skimr::skim() gives us everything we need to quickly derive insights.

coffee\_ratings\_preprocessed\_tbl %>% skimr::skim()

Table 1: Data summary

Name Piped data

Number of rows 1339

Number of columns 39

Column type frequency: character 20

numeric 19

Group variables None

## Variable type: character

**skim\_variable n\_missing complete\_rate min max empty n\_unique whitespace**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| species | 0 | 1.00 | 7 | 7 | 0 | 2 | 0 |
| owner | 7 | 0.99 | 3 | 50 | 0 | 315 | 0 |
| country\_of\_origin | 1 | 1.00 | 4 | 28 | 0 | 36 | 0 |
| farm\_name | 359 | 0.73 | 1 | 73 | 0 | 571 | 0 |
| lot\_number | 1063 | 0.21 | 1 | 71 | 0 | 227 | 0 |
| mill | 315 | 0.76 | 1 | 77 | 0 | 460 | 0 |
| ico\_number | 151 | 0.89 | 1 | 40 | 0 | 847 | 0 |
| company | 209 | 0.84 | 3 | 73 | 0 | 281 | 0 |
| altitude | 226 | 0.83 | 1 | 41 | 0 | 396 | 0 |
| region | 59 | 0.96 | 2 | 76 | 0 | 356 | 0 |
| producer | 231 | 0.83 | 1 | 100 | 0 | 691 | 0 |
| bag\_weight | 0 | 1.00 | 1 | 8 | 0 | 56 | 0 |
| harvest\_year | 47 | 0.96 | 3 | 24 | 0 | 46 | 0 |
| grading\_date | 0 | 1.00 | 13 | 20 | 0 | 567 | 0 |
| owner\_1 | 7 | 0.99 | 3 | 50 | 0 | 319 | 0 |
| variety | 226 | 0.83 | 4 | 21 | 0 | 29 | 0 |
| processing\_method | 170 | 0.87 | 5 | 25 | 0 | 5 | 0 |
| color | 218 | 0.84 | 4 | 12 | 0 | 4 | 0 |
| expiration | 0 | 1.00 | 13 | 20 | 0 | 566 | 0 |
| unit\_of\_measurement | 0 | 1.00 | 1 | 2 | 0 | 2 | 0 |

## Variable type: numeric

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **skim\_variable** | **n\_missing** | **complete\_rate** | **mean** | **sd** | **p0** | **p25** | **p50** | **p75** | **p100 hist** |
| total\_cup\_points | 0 | 1.00 | 82.09 | 3.50 | 0 | 81.08 | 82.50 | 83.67 | 90.58 ▁▁▁▁▇ |
| number\_of\_bags | 0 | 1.00 | 154.18 | 129.99 | 0 | 14.00 | 175.00 | 275.00 | 1062.00 ▇▇▁▁▁ |
| aroma | 0 | 1.00 | 7.57 | 0.38 | 0 | 7.42 | 7.58 | 7.75 | 8.75 ▁▁▁▁▇ |
| flavor | 0 | 1.00 | 7.52 | 0.40 | 0 | 7.33 | 7.58 | 7.75 | 8.83 ▁▁▁▁▇ |
| aftertaste | 0 | 1.00 | 7.40 | 0.40 | 0 | 7.25 | 7.42 | 7.58 | 8.67 ▁▁▁▁▇ |
| acidity | 0 | 1.00 | 7.54 | 0.38 | 0 | 7.33 | 7.58 | 7.75 | 8.75 ▁▁▁▁▇ |
| body | 0 | 1.00 | 7.52 | 0.37 | 0 | 7.33 | 7.50 | 7.67 | 8.58 ▁▁▁▁▇ |
| balance | 0 | 1.00 | 7.52 | 0.41 | 0 | 7.33 | 7.50 | 7.75 | 8.75 ▁▁▁▁▇ |
| uniformity | 0 | 1.00 | 9.83 | 0.55 | 0 | 10.00 | 10.00 | 10.00 | 10.00 ▁▁▁▁▇ |
| clean\_cup | 0 | 1.00 | 9.84 | 0.76 | 0 | 10.00 | 10.00 | 10.00 | 10.00 ▁▁▁▁▇ |
| sweetness | 0 | 1.00 | 9.86 | 0.62 | 0 | 10.00 | 10.00 | 10.00 | 10.00 ▁▁▁▁▇ |
| cupper\_points | 0 | 1.00 | 7.50 | 0.47 | 0 | 7.25 | 7.50 | 7.75 | 10.00 ▁▁▁▇▁ |
| moisture | 0 | 1.00 | 0.09 | 0.05 | 0 | 0.09 | 0.11 | 0.12 | 0.28 ▃▇▅▁▁ |
| category\_one\_defects | 0 | 1.00 | 0.48 | 2.55 | 0 | 0.00 | 0.00 | 0.00 | 63.00 ▇▁▁▁▁ |
| quakers | 1 | 1.00 | 0.17 | 0.83 | 0 | 0.00 | 0.00 | 0.00 | 11.00 ▇▁▁▁▁ |
| category\_two\_defects | 0 | 1.00 | 3.56 | 5.31 | 0 | 0.00 | 2.00 | 4.00 | 55.00 ▇▁▁▁▁ |
| altitude\_low\_meters | 230 | 0.83 1750.71 8669.44 | | | 1 1100.00 1310.64 1600.00 190164.00 ▇▁▁▁▁ | | | | |
| altitude\_high\_meters | 230 | 0.83 1799.35 8668.81 | | | 1 1100.00 1350.00 1650.00 190164.00 ▇▁▁▁▁ | | | | |
| altitude\_mean\_meters | 230 | 0.83 1775.03 8668.63 | | | 1 1100.00 1310.64 1600.00 190164.00 ▇▁▁▁▁ | | | | |

The skim() function gives an incredible amount of detail to help guide data preprocessing.

## New Insights

Breakout by data-type: 20 categorical and 19 numeric features Substantial missing values within some features

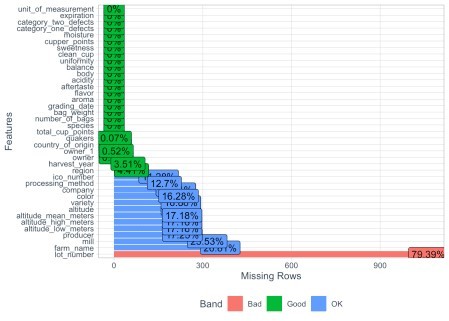
Many features with skewed distributions

Large number of features that appear unnecessary Categorical features with large number of unique values

# Missing Data

The visualization provided by plot\_missing() helps identify columns that may need attention.

coffee\_ratings\_preprocessed\_tbl %>% plot\_missing(ggtheme = theme\_tq())



This visual allows rapid assessment of features that may need to be dropped or have their values estimated via imputation.

## New Insights

Most features have complete data.

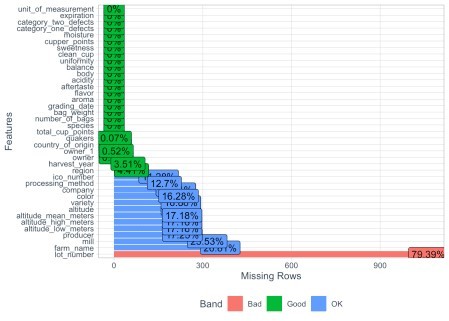
Many features (if kept) need imputation (estimate and replace missing data).

# Categorical Data

Equipped with plot\_bar() we can rapidly assess categorical features by looking at the frequency of each value.

coffee\_ratings\_preprocessed\_tbl %>%

plot\_bar(ggtheme = theme\_tq(), ncol = 2, nrow = 4)



I’m definitely impressed with this function and it is now part of my EDA toolbox 

## New Insights

Arabica dominates the species feature (we can remove)

Features exist with many categories but few values (we can lump into ‘other’) We can engineer a continent feature from country\_of\_orgin

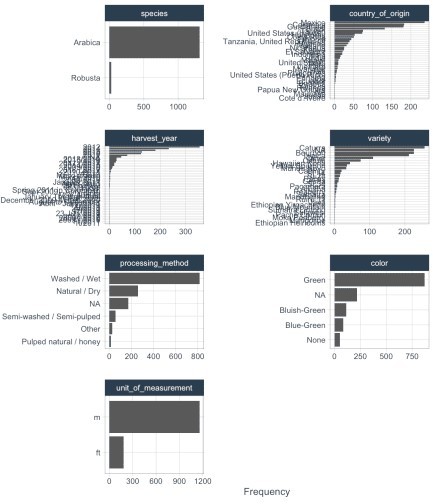
Cleaning and standardization is needed for harvest\_year Unit of measurement can be dropped

Better picture of where imputation is needed

# Numerical Data

Onward to assessing our numerical features using plot\_histogram(). coffee\_ratings\_preprocessed\_tbl %>%

plot\_histogram(ggtheme = theme\_tq(), nrow = 5, ncol = 4)



This is another function that swiftly made its way into my EDA toolbox 

## New Insights

Many features look normally distributed

Skewed features may need transformations (depending on modeling approach) We can probably keep mean altitude and drop the low and high versions Quakers (unripened beans) should probably be categorical

# Plot Altitude

Let’s test our assumption about dropping the low and high altitude features.

coffee\_ratings\_preprocessed\_tbl %>%

# select columns and pivot data select(contains("altitude\_")) %>% pivot\_longer(1:3) %>%

# plot data

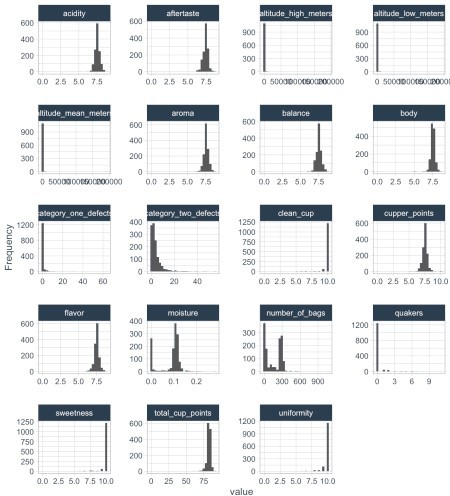
ggplot(aes(name, value, color = name)) + geom\_violin() +

geom\_jitter(alpha = 0.05) +

# formatting

scale\_y\_log10(label = scales::comma\_format()) + theme(legend.position = "none") +

labs(x = "", y = "Meters")



Looks good.

The variation between low and high isn’t substantial and so we would probably keep altitude\_mean\_meters and drop the others.

# Plot Quakers vs. Score

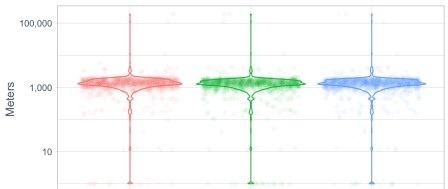
Let’s quickly double check quakers to see if it’s better to encode as a factor (categorical variable).

coffee\_ratings\_preprocessed\_tbl %>%

# select columns and plot data select(quakers, total\_cup\_points) %>%

ggplot(aes(as.factor(quakers), total\_cup\_points)) + geom\_violin() +

geom\_jitter(alpha = 0.2) + ylim(0, 100)





It doesn’t look like quakers explains much of the variation within total\_cup\_points. If kept, it would be worth updating to a categorical variable.

# Wrap Up

Rapidly assessing data is critical for speeding up analysis.

After researching {DataExplorer} I am convinced it is worth adding to the data practitioners toolbox  Three functions allowed us to quickly assess our data and gain insights:

DataExplorer::plot\_missing() DataExplorer::plot\_bar() DataExplorer::plot\_histogram()

These insights could then be used in the next step of cleaning and preprocessing these data for analysis and/or predictive modeling.