Understanding dlookr package

실습을 중심으로

유충현

Updated: 2018/05/22

Overview

1. Motivation

2. dlookr exercises

Motivation

First Topic - rstudio::conf 2018

2018-01-31 2018-02-03에 걸쳐 San Diego에서 진행된 rstudio::conf 2018에 참석하는 행운을 얻었다. 컨퍼런스는 tidyverse package 와 Shiny package에 대한 세션들이 주를 이루었다. 2일 과정의 Data Science in tidyverse 교육 세션과 2일 과정의 컨퍼런스를 통해 tidyverse의 성장세를 느낄 수 있었다.

tidyverse package는 데이터를 조작하는 dplyr package와 데이터를 시각적으로 탐색하는 ggplot2 package가 주류를 이루고 있다. 자유도가 높은 Analytics에서 Modeling 과정을 제외한 데이터 입출력, 데이터 전처리, 시각화 과정에 어느 정도 표준으로 자리매김하고 있다.

rstudio::conf 2018 - 컨퍼런스 참석

그림: rstudio::conf 2018 참석

rstudio::conf 2018 - 유명 인사들을 만나다

그림: Hadley Wickham

rstudio::conf 2018 - 저자 사인북

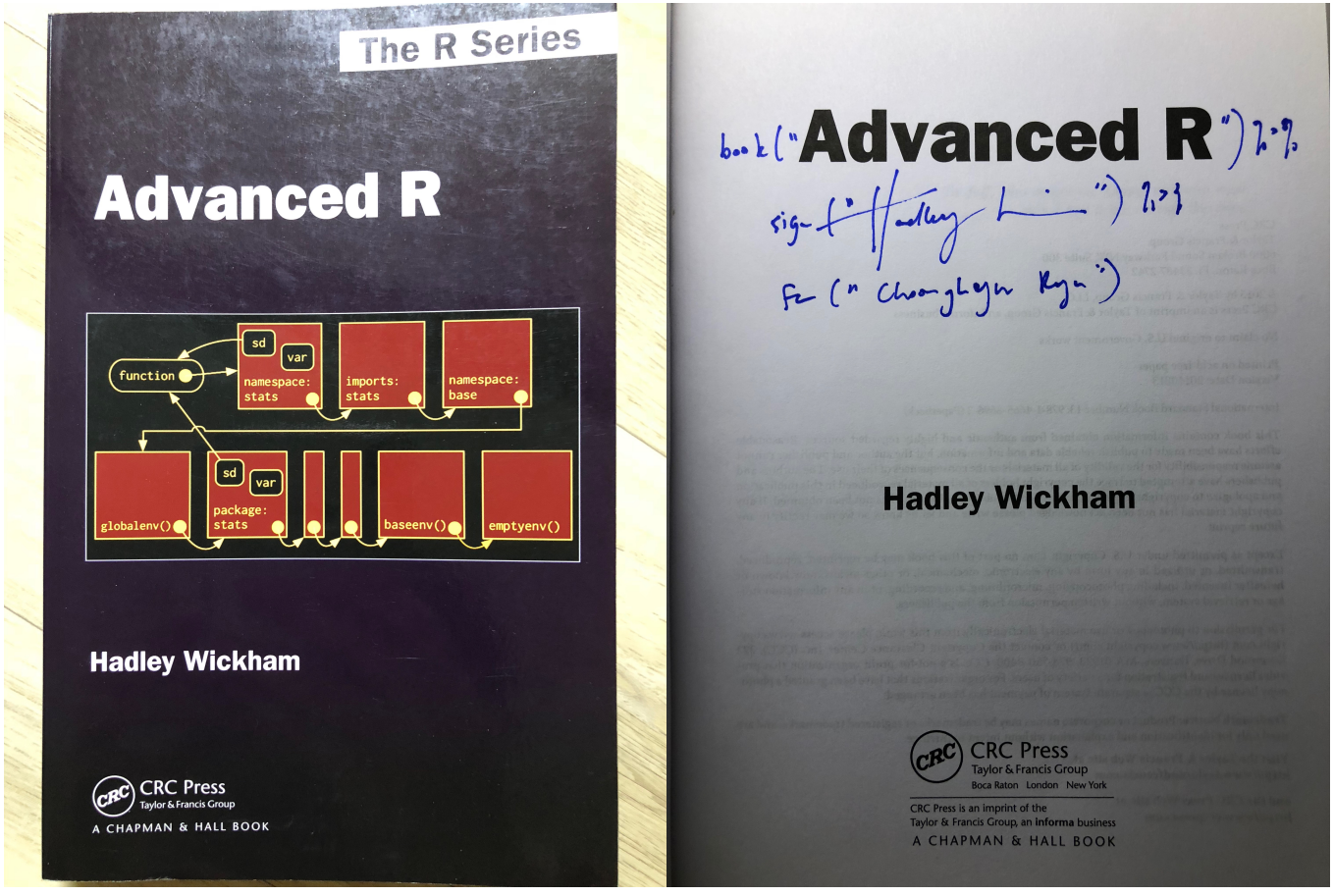

그림: Advanced R - Hadley Wickham

rstudio::conf 2018 - 저자 사인북

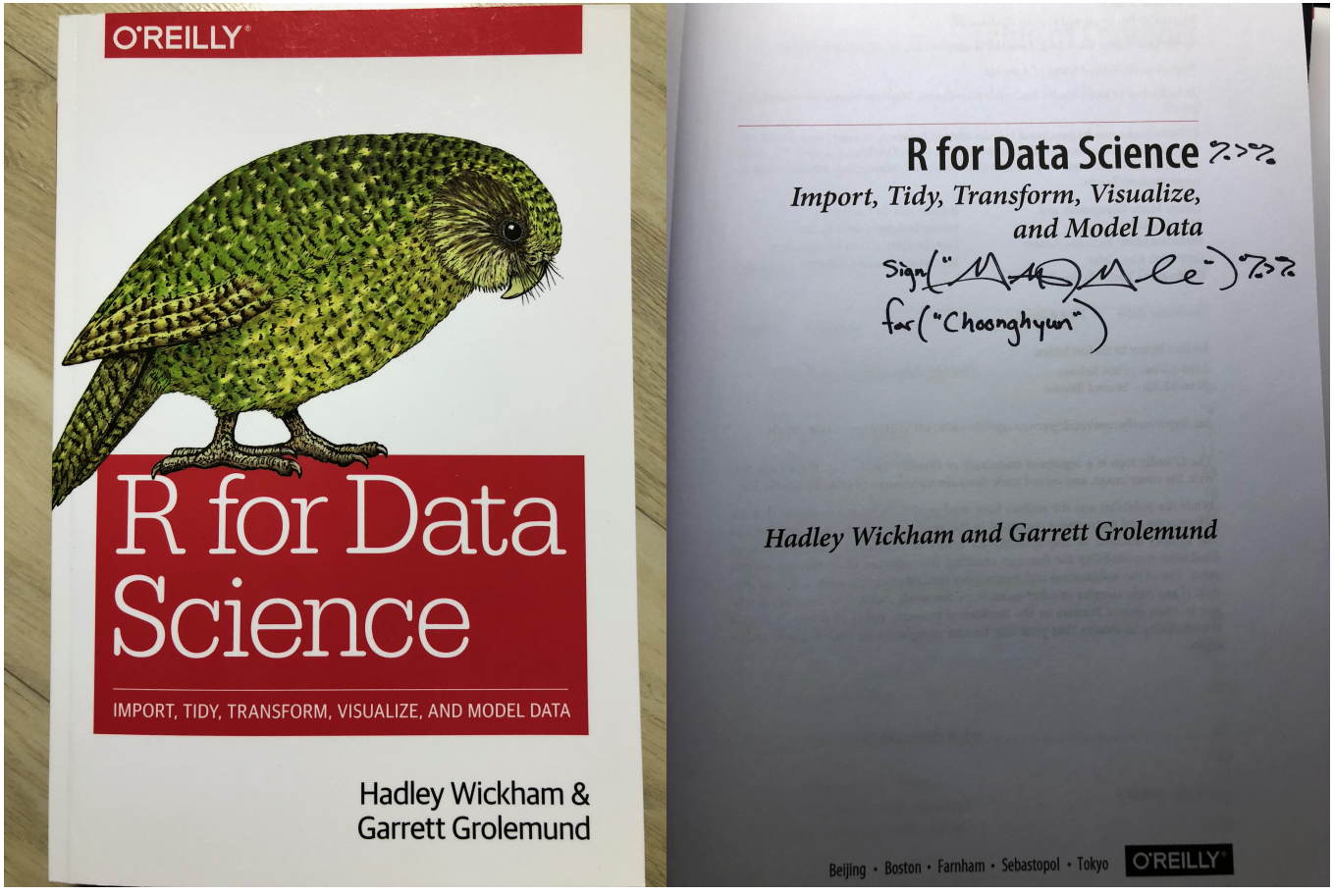

그림: R for Data Science - Garrett Grolemund

rstudio::conf 2018 - 저자 사인북

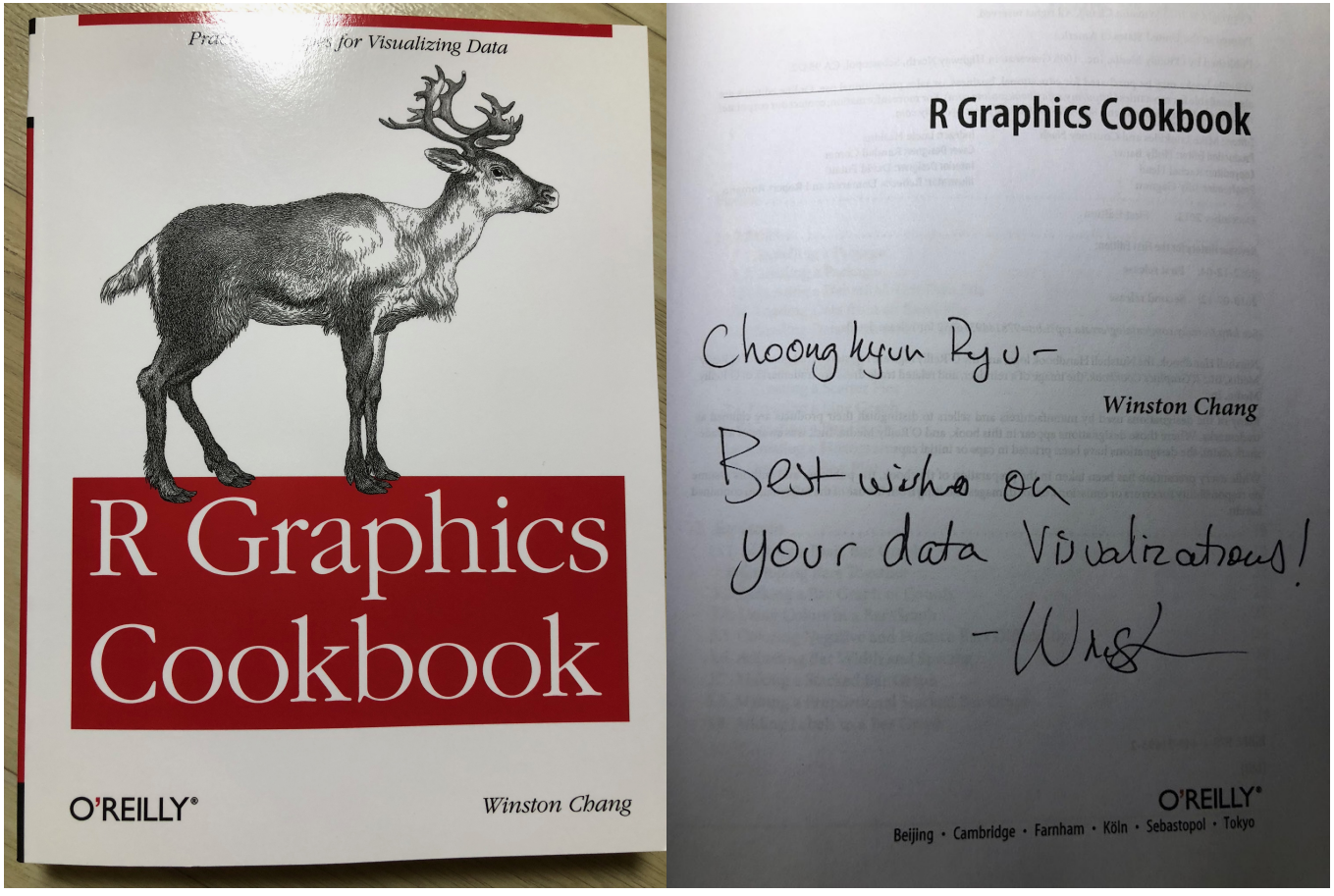

그림: R Graphics Cookbook - Windton Chang

rstudio::conf 2018 - Stickers

그림: RStudio family package's Stickers

Second Topic - dlookr package

rstudio::conf 2018를 다녀와서 "R User에게 도움을 줄 수 있는 packages를 만들어보자."는 Motivation이 생겼다. 이왕이면 tidyverse package와 궁합이 맞는 packages를 만들고 싶었다.

주말을 이용한 3개월의 기간동안 데이터 품질 진단, 탐색적 데이터 분석, 변수 변환을 지원하는 dlookr package를 만들어 배포할 수 있었다.

회사에서 사용하는 private package를 7종 개발하였지만, CRAN의 public package의 여러 제약 사항과, github의 버전 관리와 협업 체계는 너무나 낯설고 어려운 과정이었다.

dlookr package - bucket lists

"나도 유용한 패키지를 만들어서 누군가의 노트북에 스티커로 붙여 놓을 수 있었으면 좋겠다."

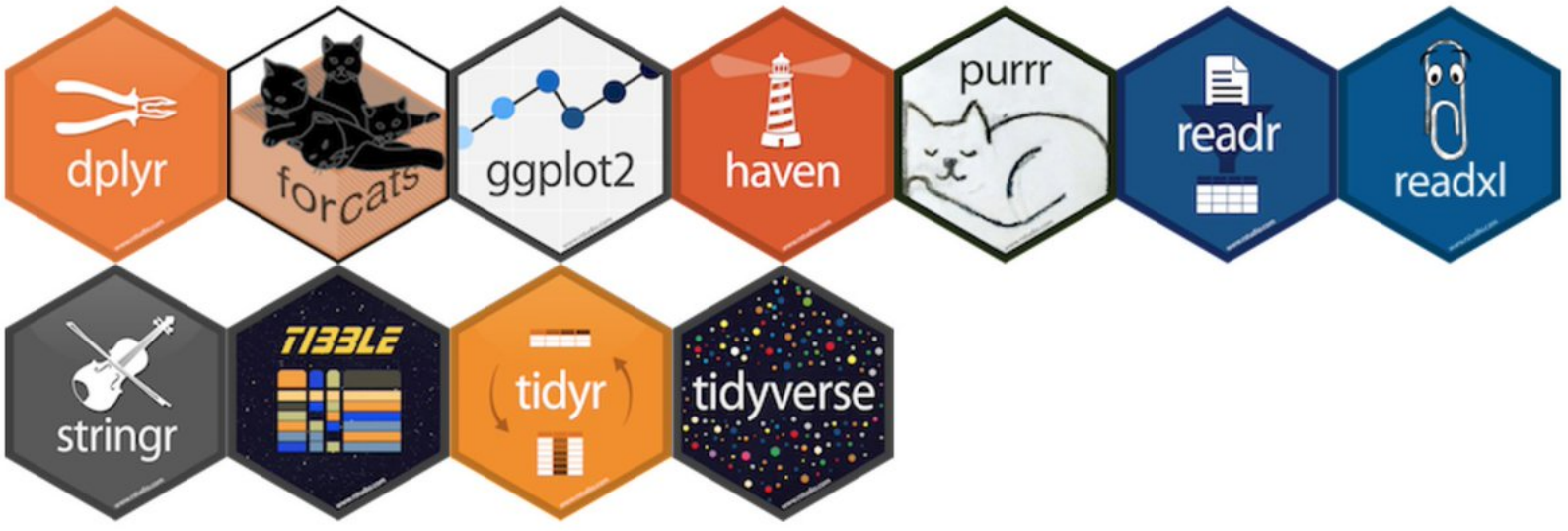

그림: My bucket lists - hexbin stickers

무엇을 만들것인가 ?!

"tidyverse package는 데이터 모델링을 위한 패키지라는 느낌보다는 데이터 조작과 시각화 및 함수형 프로그래밍을 위한 low levels 패키지라는 느낌을 지울 수 없다. 좀더 분석가에게 유용하게 활용될 수 있는 데이터 분석 툴을 만들어 보자."

자유도가 높은 데이터분석에서 표준적인 툴이 가능하겠는가?!!!

- dlookr - 0.3.0 개발:
  - 전처리, EDA, 변수변환 과정을 지원하자., `look data`.
- alookr - not yet:
  - 데이터 모델링 과정을 지원하자., `look analytics`.
- mlookr - not yet:
  - 모델 배포 및 관리를 지원하자., `look models`.

개발은 쉬운데, 배포가 어려워!!!

"패키지 개발은 그럭저럭 쉬웠는데, CRAN 배포와 github 관리가 어려움을 알았다."

- CRAN:
  - 도움말 한글 불가
    - private 패키지는 한글 도움말이 가능하나, CRAN은 불가
    - 부족한 영어 실력
  - 도움말 예제 `Runtime` 제약
    - 도움말 예제 수행 속도가 5초를 상회하는 것 불가
    - \dontrun{}이 아닌, \donttest{} 사용해야 함
  - `library(), require()` 불가
    - package::function() 형식으로 호출
    - NAMESPACE 파일에 importFrom(package,function)을 기술
  - `DESCRIPTION` 파일의 엄격한 심사
    - Title을 유사 기능의 패키지와 구별되도록 구체적으로 명기
    - Description의 대소문자 체크
  - 행운이 있었다
    - 심사자가 B.D. Ripley가 아닌 Swetlana Herbrandt

dlookr package - bucket lists

그림: My bucket lists - hexbin stickers 완료

dlookr exercises

about dlookr

dlookr: Tools for Data Diagnosis, Exploration, Transformation

A collection of tools that support data diagnosis, exploration, and transformation. Data diagnostics provides information and visualization of missing values and outliers and unique and negative values to help you understand the distribution and quality of your data. Data exploration provides information and visualization of the descriptive statistics of univariate variables, normality tests and outliers, correlation of two variables, and relationship between target variable and predictor. Data transformation supports binning for categorizing continuous variables, imputates missing values and outliers, resolving skewness. And it creates automated reports that support these three tasks.

Version:	0.3.0
Depends:	R (≥ 3.2.0)
Imports:	dplyr, magrittr, tidyr, ggplot2, RcmdrMisc, corrplot, rlang, purrr, tibble, tidyselect, classInt, moments, kableExtra, prettydoc, smbinning, xtable, knitr, rmarkdown, RColorBrewer, gridExtra, tinytex, methods, DMwR, mice, rpart, randomForest
Suggests:	ISLR, nycflights13, testthat
Published:	2018-04-27
Author:	Choonghyun Ryu [aut, cre]
Maintainer:	Choonghyun Ryu <choonghyun.ryu at gmail.com>
License:	GPL-2
NeedsCompilation:	no
CRAN checks:	dlookr results

Downloads:

Reference manual:	dlookr.pdf
Vignettes:	Exploratory Data Analysis Data quality diagnosis Data Transformation
Package source:	dlookr_0.3.0.tar.gz
Windows binaries:	r-devel: dlookr_0.3.0.zip, r-release: dlookr_0.3.0.zip, r-oldrel: dlookr_0.3.0.zip
OS X binaries:	r-release: dlookr_0.3.0.tgz, r-oldrel: not available

그림: CRAN - dlookr page

about dlookr

dlookr

CRAN 0.3.0

Overview

Diagnose, explore and transform data with `dlookr`.

Features:

- Diagnose data quality.
- Find appropriate scenarios to pursuit the follow-up analysis through data exploration and understanding
- Derive new variables or perform variable transformations.
- Automatically generate reports for the above three tasks.

The name `dlookr` comes from `looking at the data` in the data analysis process.

Install dlookr

The released version is available on CRAN

```
install.packages("dlookr")
```

Or you can get the development version without vignettes from GitHub:

```
devtools::install_github("choonghyunryu/dlookr")
```

그림: github - dlookr README.md

dlookr의 기능

"model fitting 이전 과정에서의 데이터진단/EDA/변수변환을 지원한다."

- 데이터진단
  - 수치변수/범주형 변수의 품질진단
  - 이상치 진단 및 시각화
- EDA
  - 기술통계량 계산 및 정규성 검정과 시각화
  - 상관계수 계산 및 상관관계 시각화
  - Target variable과 predictor와의 관계 규명 및 시각화
- 변수변환
  - 결측치와 이상치의 대체 및 표준화
  - binning 및 optimal binning
- 자동화된 보고서 작성
  - 데이터진단/EDA/변수변환 보고서 작성

skimr과의 비교

"이미 유사한 기능의 패키지가 있었다"

- skimr
  - 출시일 : 2017-12-21
  - 기능 : Compact and Flexible Summaries of Data
  - dplyr과의 궁합 - use pipe
  - support for inline spark graphs
- dlookr
  - 출시일 : 2018-04-27
  - 기능 : Tools for Data Diagnosis, Exploration, Transformation
  - dplyr과의 궁합 - use pipe
  - automated reporting

exercises: dlookr 설치

"dlookr 패키지를 설치하자"

```
> # CRAN으로부터의 설치
> install.packages("dlookr")
>
> # github으로부터의 설치
> devtools::install_github("choonghyunryu/dlookr")
>
> # github으로부터의 설치 (vignettes 포함, 권장함)
> install.packages(c("nycflights13", "ISLR"))
> devtools::install_github("choonghyunryu/dlookr",
+   build_vignettes = TRUE)
```

exercises: 어떻게 학습할 것인가?

"vignettes을 활용하자."

- vignettes - 영문으로 제공
  - browseVignettes(package = "dlookr")
- blog - vignettes을 한글로 제공
  - 데이터 품질진단
    - https://choonghyunryu.github.io/ko/2018/05/dlookr-데이터-품질-진단/
  - 탐색적 데이터분석
    - https://choonghyunryu.github.io/ko/2018/05/dlookr-탐색적-데이터-분석/
  - 데이터 변환
    - hhttps://choonghyunryu.github.io/ko/2018/05/dlookr-데이터-변환/

exercises: 데이터의 준비

"예제를 위한 데이터를 준비하자"

```
> # 2013년 NYC를 출발한 모든 항공편에 출발과 도착에 대한 정보
> library("nycflights13")
> data(flights)
>
> # 400개 매장의 아동용 카시트를 판매 시뮬레이션 데이터
> carseats <- ISLR::Carseats
>
> set.seed(123)
> carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
> set.seed(456)
> carseats[sample(seq(NROW(carseats)), 10), "Urban"] <- NA
```

exercises: 자동화된 리포트 1

"데이터의 품질을 진단해 보자"

```
> # latex이 설치된 경우 -  데이터 품질 진단
> flights %>%
+   diagnose_report()
>
> # latex이 설치되지 않은 경우 -  데이터 품질 진단 (웹버전)
> diagnose_report(flights, output_format = "html",
+   output_file = "Diagn.html")
```

exercises: 자동화된 리포트 2

"탐색적 데이터 분석을 수행하자"

```
> # 탐색적 데이터 분석 - target variable == numeric
> carseats %>%
+   eda_report(target = Sales, output_format = "html",
+     output_file = "EDA.html")
> # 탐색적 데이터 분석 - target variable == categorical
> eda_report(carseats, US, output_format = "html",
+   output_file = "EDA.html")
> # 탐색적 데이터 분석 - target variable is null
> eda_report(carseats, output_format = "html",
+   output_file = "EDA2.html")
```

exercises: 자동화된 리포트 3

"변수의 변환을 수행해 보자"

```
> carseats <- ISLR::Carseats
> carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
> carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA
>
> # 변수의 변환 - target variable is null
> transformation_report(carseats, output_format = "html")
>
> # 변수의 변환 - target variable is binary class
> transformation_report(carseats, US, output_format = "html",
+   output_file = "Transformation.html")
```

exercises: 데이터 품질 진단 실습

https://choonghyunryu.github.io/ko/2018/05/dlookr-데이터-품질-진단

- 학습 목표
  - diagnose()을 이용한 변수의 개괄적 진단
  - diagnose_numeric()을 이용한 수치형 변수의 상세 진단
  - diagnose_category()을 이용한 범주형 변수의 상세 진단
  - diagnose_outlier()를 이용한 이상치 진단
  - plot_outlier()를 이용한 이상치의 시각화
  - diagnose_report()를 이용한 진단 보고서 작성

exercises: 탐색적 데이터 분석 실습

https://choonghyunryu.github.io/ko/2018/05/dlookr-탐색적-데이터-분석

- 학습 목표
  - describe()을 이용한 기술통계량 계산
  - normality()을 이용한 수치형 변수의 정규성 검정
  - plot_normality()를 이용한 수치변수의 정규성 시각화
  - correlate()을 이용한 상관계수 계산
  - plot_correlate()를 이용한 상관행렬의 시각화
  - Target 변수에 기반한 EDA
    - 4가지 case에 대한 EDA의 이해
  - eda_report()를 이용한 EDA 보고서 작성

exercises: 데이터 변환 실습

https://choonghyunryu.github.io/ko/2018/05/dlookr-데이터-변환

- 학습 목표
  - imputate_na()을 이용한 결측치의 대체
  - imputate_outlier()을 이용한 이상치의 대체
  - transform()을 이용한 표준화
  - transform()을 이용한 치우친 데이터의 보정
  - binning()을 이용한 개별 변수의 Binning
  - binning_by()을 이용한 Optimal Binning
  - transformation_report()를 이용한 데이터변환 보고서 작성

THE END