Diagnosis data and EDA in R

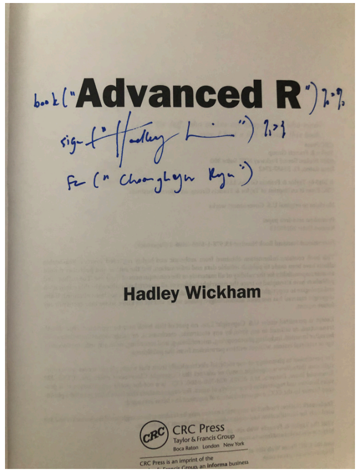
- 2021-03-10
- 유충현, Tidyverse Korea

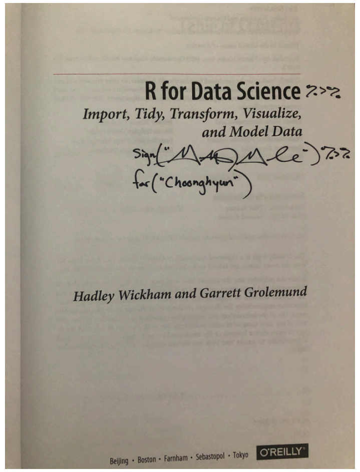
Tidyverse
Korea

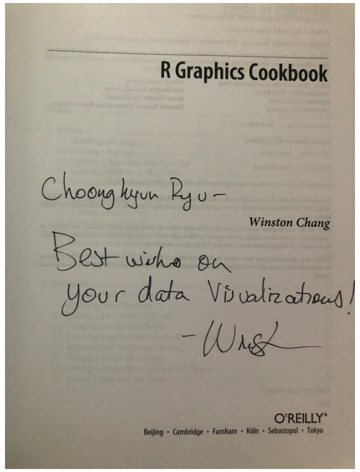
CONTENTS

Motivation – RStudio Conference 2018 (1/31 ~ 2/3)

- "tidyverse packages에 합류할 수 있는 패키지를 만들어 보자."
- "나의 Hexbin 스티커를 만들어 보자."







Submitted the CRAN (2018-04-27)

- 데이터 품질 진단, 탐색적 자료분석(EDA), 데이터 변환을 수행하는 패키지
- tidyverse packages와 협업하면, 기능이 배가됨

CRAN - Package dlookr

cran.r-project.org/web/packages/dlookr/index.html

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:	0.4.2
Depends:	R (≥ 3.2.0)
Imports:	cli, corrplot, dplyr (≥ 0.7.6), extrafont, ggplot2 (≥ 3.0.0), grid, gridExtra, hrbrthemes, kableExtra, knitr (≥ 1.22), magrittr, methods, mice, partykit, prettydoc, purrr, RcmdrMisc, rlang, rmarkdown, tibble, tidyr, tidyselect, utils
Suggests:	DBI, classInt, dbplyr, DMwR, forecast, ISLR, nycflights13, randomForest, rpart, RSQLite, stringr
Published:	2021-02-12
Author:	Choonghyun Ryu [aut, cre]
Maintainer:	Choonghyun Ryu <choonghyun.ryu at gmail.com>
BugReports:	https://github.com/choonghyunryu/dlookr/issues
License:	GPL-2 | file LICENSE
NeedsCompilation:	no
Materials:	NEWS
In views:	MissingData
CRAN checks:	dlookr results

Downloads:

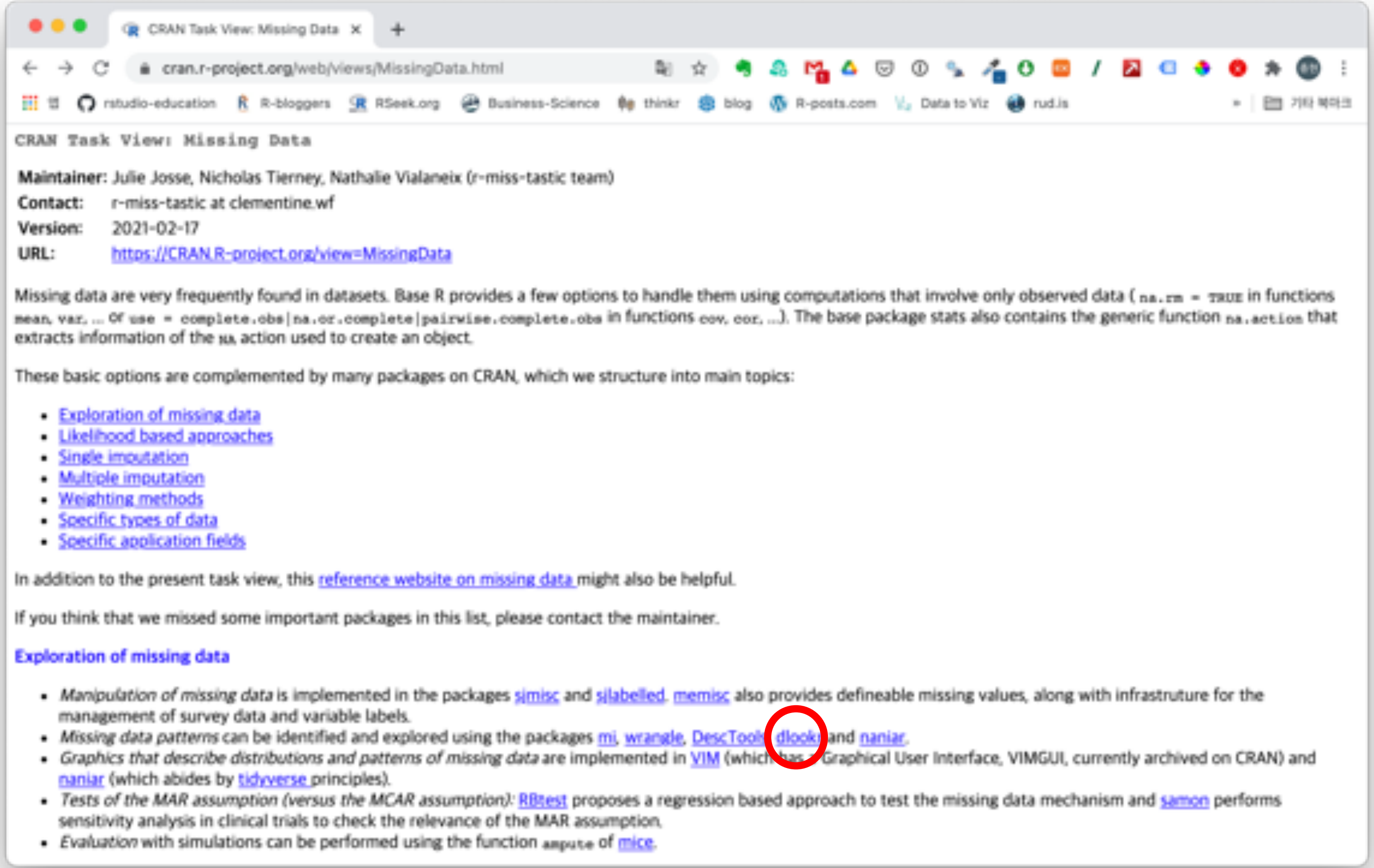
Reference manual:	dlookr.pdf
Vignettes:	Exploratory Data Analysis Data quality diagnosis Introduce dlookr Data Transformation
Package source:	dlookr_0.4.2.tar.gz
Windows binaries:	r-devel: dlookr_0.4.2.zip, r-release: dlookr_0.4.2.zip, r-oldrel: dlookr_0.4.2.zip
macOS binaries:	r-release: dlookr_0.4.2.tgz, r-oldrel: dlookr_0.4.2.tgz
Old sources:	dlookr archive

- CRAN : https://cran.r-project.org/web/packages/dlookr/

- GitHub : https://github.com/choonghyunryu/dlookr

몇 가지의 성과

- CRAN Task View에 등록됨 : Missing Data section
- 몇 편의 해외 논문에 소개/비교됨


CRAN Task View: Missing Data

cran.r-project.org/web/views/MissingData.html

CRAN Task View: Missing Data

Maintainer: Julie Josse, Nicholas Tierney, Nathalie Vialaneix (r-miss-tastic team)
Contact: r-miss-tastic at clementine.wf
Version: 2021-02-17
URL: https://CRAN.R-project.org/view=MissingData

Missing data are very frequently found in datasets. Base R provides a few options to handle them using computations that involve only observed data (`na.rm = TRUE` in functions `mean`, `var`, ... or `use = complete.obs|na.or.complete|pairwise.complete.obs` in functions `cov`, `cor`, ...). The base package stats also contains the generic function `na.action` that extracts information of the `NA` action used to create an object.

These basic options are complemented by many packages on CRAN, which we structure into main topics:

- Exploration of missing data
- Likelihood based approaches
- Single imputation
- Multiple imputation
- Weighting methods
- Specific types of data
- Specific application fields

In addition to the present task view, this reference website on missing data might also be helpful.

If you think that we missed some important packages in this list, please contact the maintainer.

Exploration of missing data

- *Manipulation of missing data* is implemented in the packages sjmisc and sjlabelled. memisc also provides defineable missing values, along with infrastruture for the management of survey data and variable labels.
- *Missing data patterns* can be identified and explored using the packages mi, wrangle, DescTools, dlookr and naniar.
- *Graphics that describe distributions and patterns of missing data* are implemented in VIM (which has a Graphical User Interface, VIMGUI, currently archived on CRAN) and naniar (which abides by tidyverse principles).
- *Tests of the MAR assumption (versus the MCAR assumption):* RBtest proposes a regression based approach to test the missing data mechanism and samon performs sensitivity analysis in clinical trials to check the relevance of the MAR assumption.
- *Evaluation* with simulations can be performed using the function `ampute` of mice.

The R Journal

JOSS
The Journal of Open Source Software

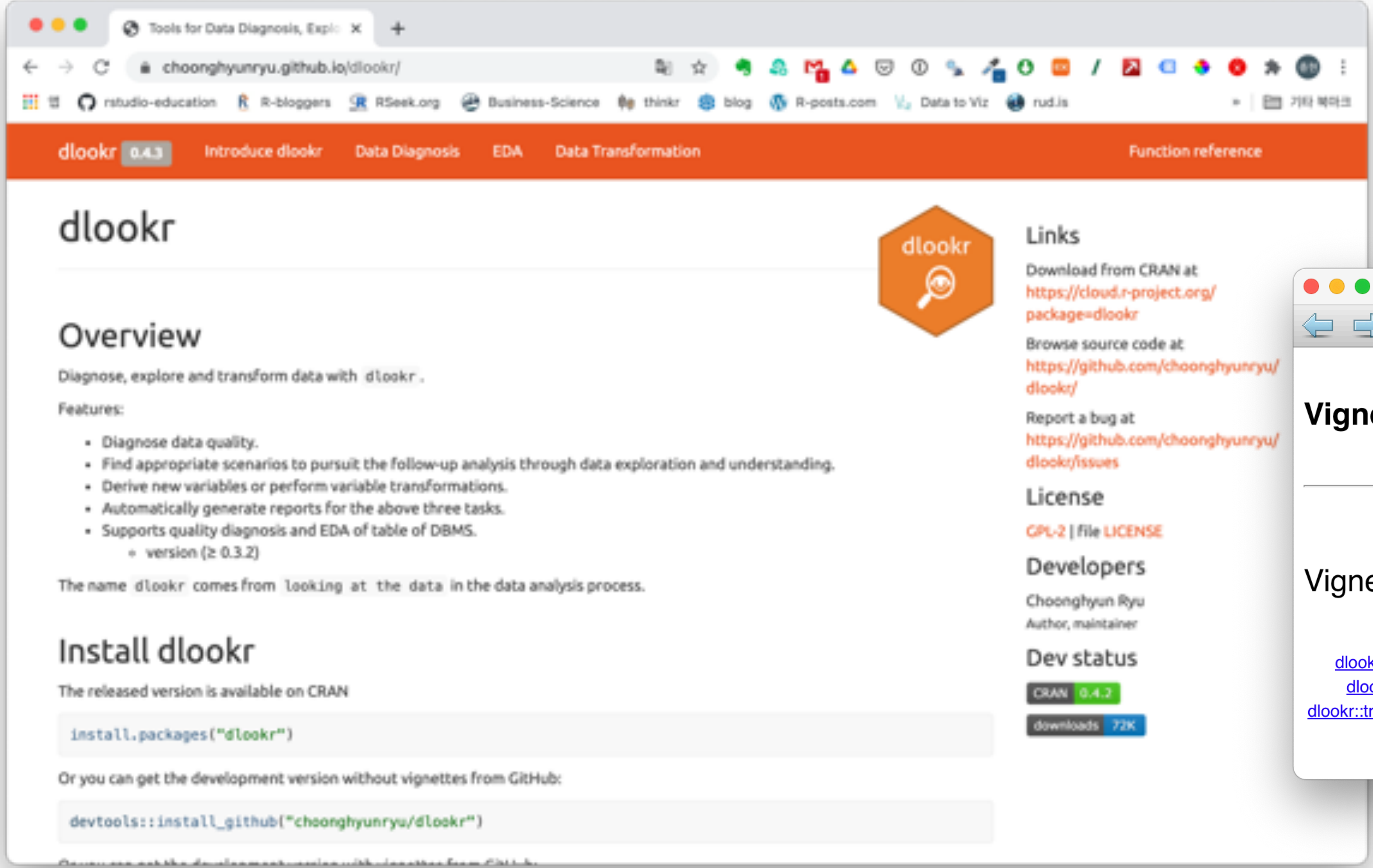
Journal of Statistical Software
MMMMMM YYYY, Volume VV, Issue II. doi: 10.18637/jss.v000.i00

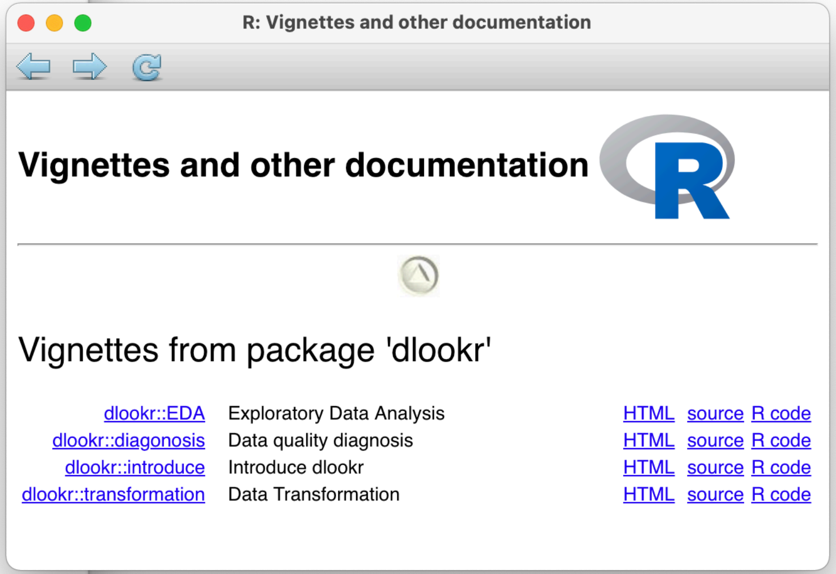
dlookr package 기능 ※ 이번에 다룰 내용

- 데이터 진단
  - 데이터 요약
  - 데이터 품질 진단
  - 범주형 품질 진단
  - 수치형 품질 진단
  - 이상치 진단
  - 결측치 진단
  - 데이터 진단 보고서
- EDA
  - 일변량 변수 탐색
  - 이변량 변수 탐색
  - 상관관계 파악
  - 정규성 검정
  - 인과관계 파악
  - EDA 보고서
- 데이터 변환
  - 특이 변수 추출
  - 특이값의 대체
  - 비닝
  - 최적 비닝
  - 데이터 변환
  - 데이터 변환 보고서
- 기타
  - 통계량 구하기
  - 프로그램 지원

dlookr package 기능 익히기

- https://choonghyunryu.github.io/dlookr/
- Vignettes

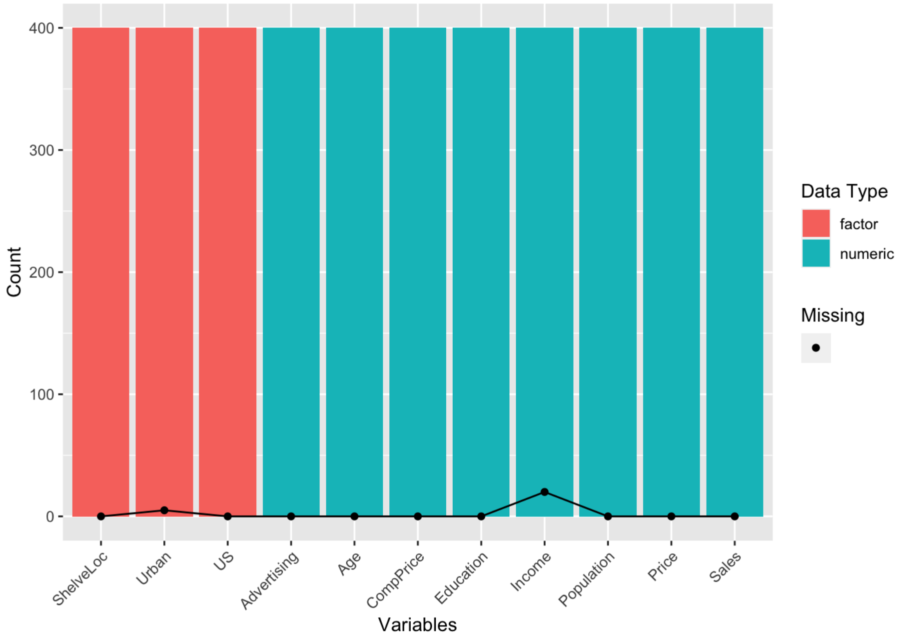




데이터 요약

```r
library(dlookr)
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20),
  "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5),
   "Urban"] <- NA

ov <- carseats %>%
  overview()

ov

summary(ov)
plot(ov)
```

```
> summary(ov)
── Data Scale ──────────────────────────────
● Number of observations              :     400
● Number of variables                 :      11
● Number of values                    :   4,400
● Size of located memory(bytes)       :  59,648

── Missing Data ────────────────────────────
● Number of completed observations :     375
● Number of observations with N/A  :      25
● Number of variables with N/A     :       2
● Number of N/A                    :      25

── Data Type ───────────────────────────────
● Number of numeric variables         :       8
● Number of integer variables         :       0
● Number of factors variables         :       3
● Number of character variables       :       0
● Number of other variables           :       0

── Individual variables ────────────────────
    Variables Data Type
1       Sales   numeric
2   CompPrice   numeric
3      Income   numeric
4 Advertising   numeric
5  Population   numeric
6       Price   numeric
7   ShelveLoc    factor
8         Age   numeric
9   Education   numeric
10      Urban    factor
11         US    factor
```



데이터 품질 진단

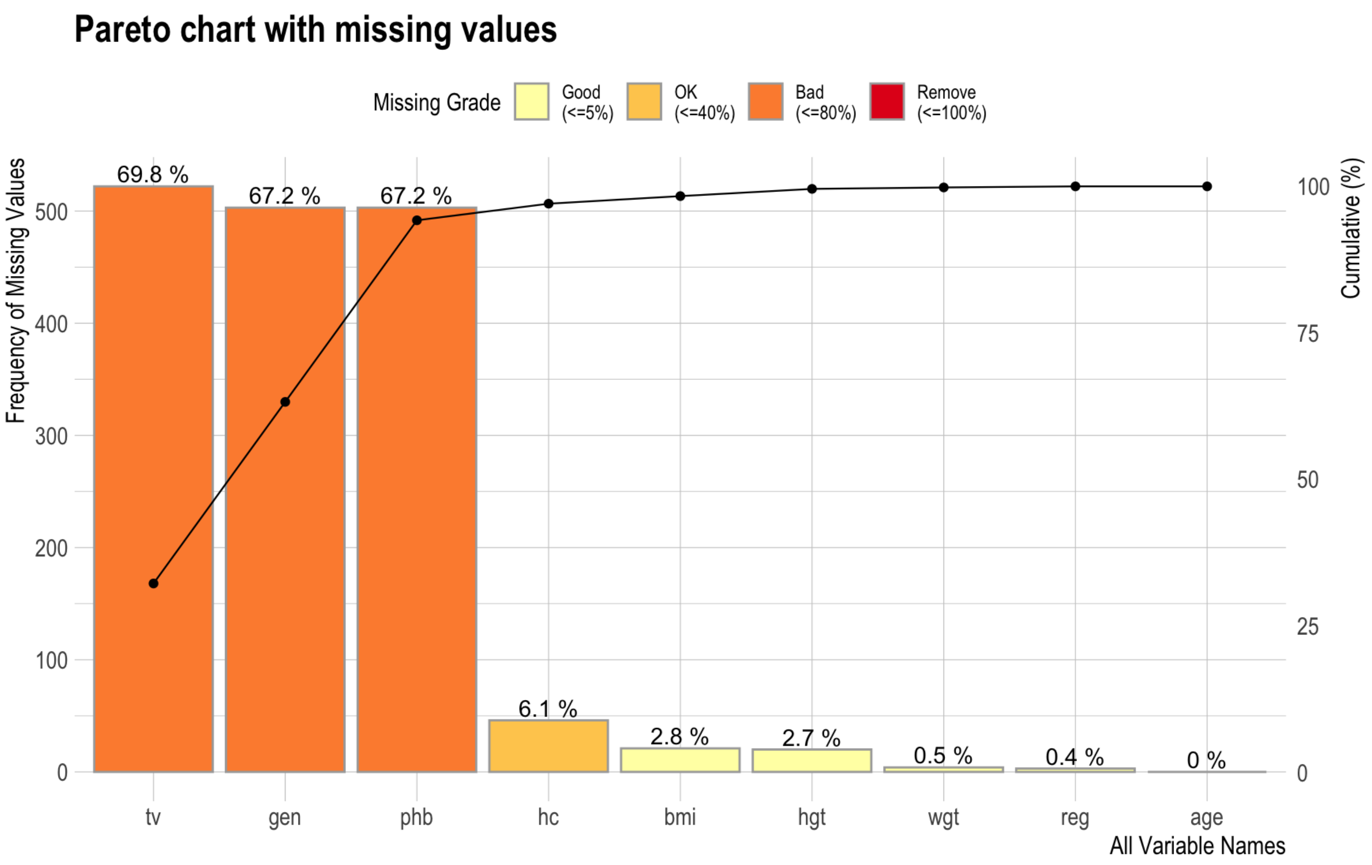
dlookr package   tidyverse packages

```
# 앞에서 5 변수의 진단
carseats %>%
  diagnose(1:5)

# 결측치가 있는 변수 추출
carseats %>%
  diagnose() %>%
  filter(missing_count > 0)

# 범주형 변수중 60%가 넘는 수준 추출
carseats %>%
  diagnose_category() %>%
  filter(ratio >= 60)

# 0을 포함하고 있는 수치형 변수 추출
carseats %>%
  diagnose_numeric() %>%
  filter(zero > 0)

# 이상치를 포함하고 있는 수치형 변수 추출
carseats %>%
  diagnose_outlier() %>%
  filter(outliers_ratio > 1)
```

```
# A tibble: 5 x 6
  variables   types   missing_count missing_percent unique_count unique_rate
  <chr>       <chr>           <int>           <dbl>        <int>       <dbl>
1 Sales       numeric             0               0          336       0.84
2 CompPrice   numeric             0               0           73       0.182
3 Income      numeric            20               5           99       0.248
4 Advertising numeric             0               0           28       0.07
5 Population  numeric             0               0          275       0.688

# A tibble: 2 x 6
  variables types   missing_count missing_percent unique_count unique_rate
  <chr>     <chr>           <int>           <dbl>        <int>       <dbl>
1 Income    numeric            20            5              99      0.248
2 Urban     factor              5            1.25            3      0.0075

 variables levels   N freq ratio rank
1     Urban    Yes 400  280  70.0    1
2        US    Yes 400  258  64.5    1

# A tibble: 2 x 10
  variables     min    Q1  mean median    Q3   max  zero minus outlier
  <chr>       <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <int> <int>   <int>
1 Sales           0  5.39  7.50   7.49  9.32  16.3     1     0       2
2 Advertising     0  0     6.64   5    12     29     144     0       0

 variables outliers_cnt outliers_ratio outliers_mean with_mean without_mean
1     Price            5           1.25         100.4   115.795     115.9899
```

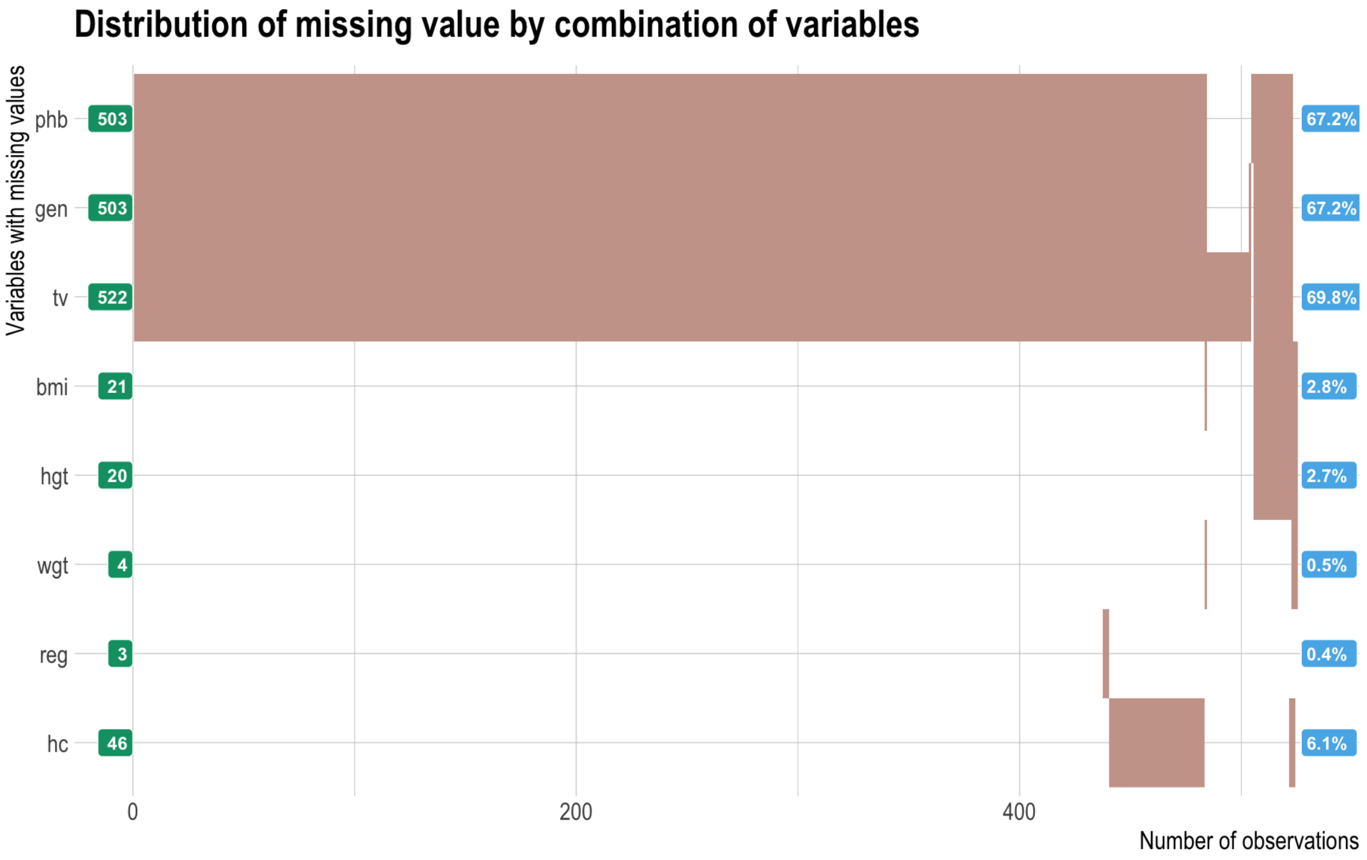
결측치 진단

```
# Visualize distribution of missing value by
pareto chart.

mice::boys %>%
  plot_na_pareto()
```



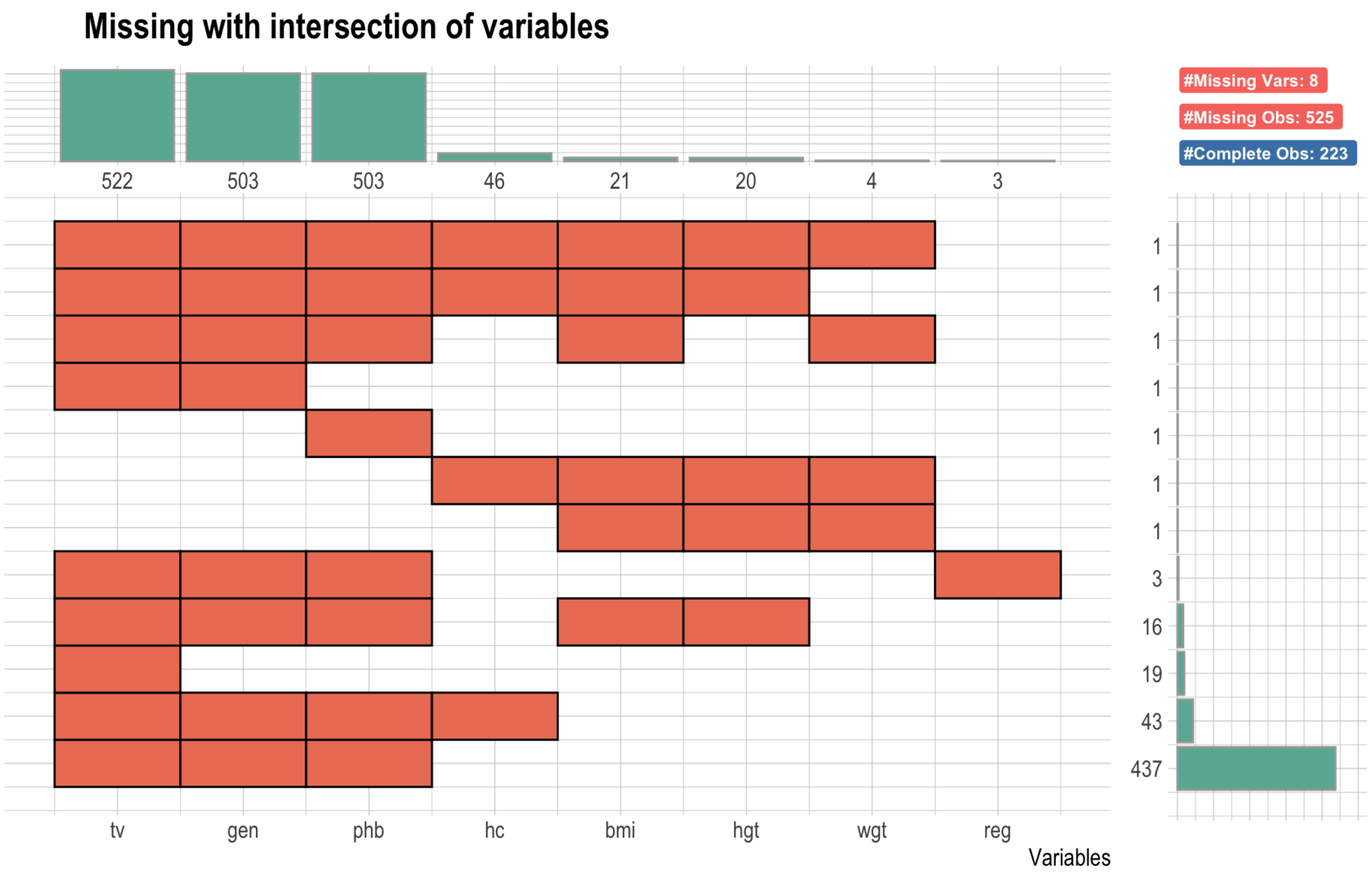
결측치 진단

```
# Visualize distribution of missing value by
hierarchy cluster.

mice::boys %>%
  plot_na_hclust()
```



결측치 진단

```
# Visualize distribution of missing value by
combination of variables.

mice::boys %>%
 plot_na_intersect()
```

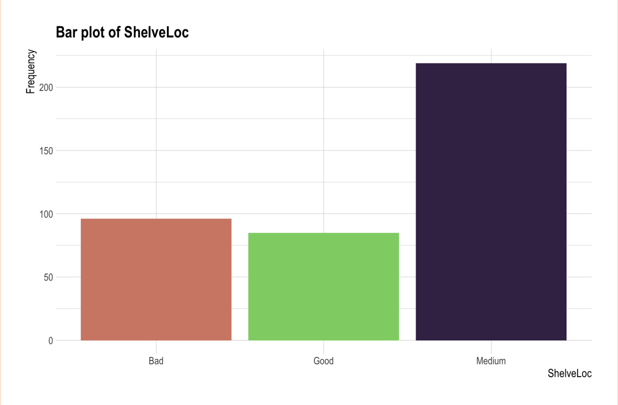


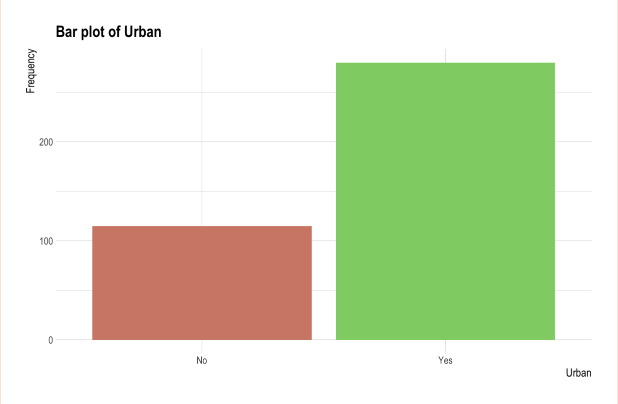
일변량 변수 탐색

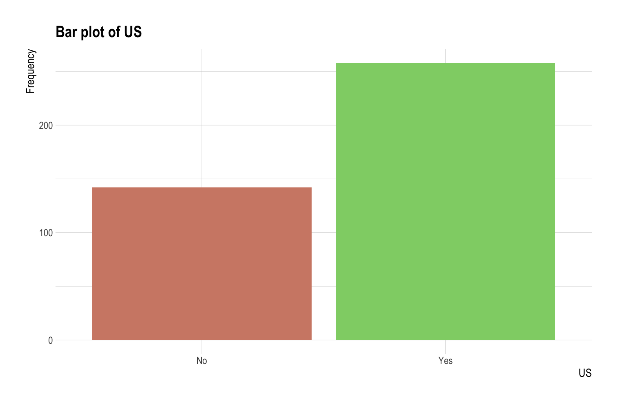
- categorical variables

```r
all_var <- carseats %>%
  univar_category()

all_var

summary(all_var)
plot(all_var)
```

```
> all_var
$ShelveLoc
  ShelveLoc   n   rate
1       Bad  96 0.2400
2      Good  85 0.2125
3    Medium 219 0.5475

$Urban
  Urban   n   rate
1    No 115 0.2875
2   Yes 280 0.7000
3  <NA>   5 0.0125

$US
   US   n  rate
1  No 142 0.355
2 Yes 258 0.645


> summary(all_var)
  variables statistic      p.value df
1 ShelveLoc  83.01500 9.408530e-19  2
2     Urban  68.92405 1.023293e-16  1
3        US  33.64000 6.631492e-09  1
```







일변량 변수 탐색

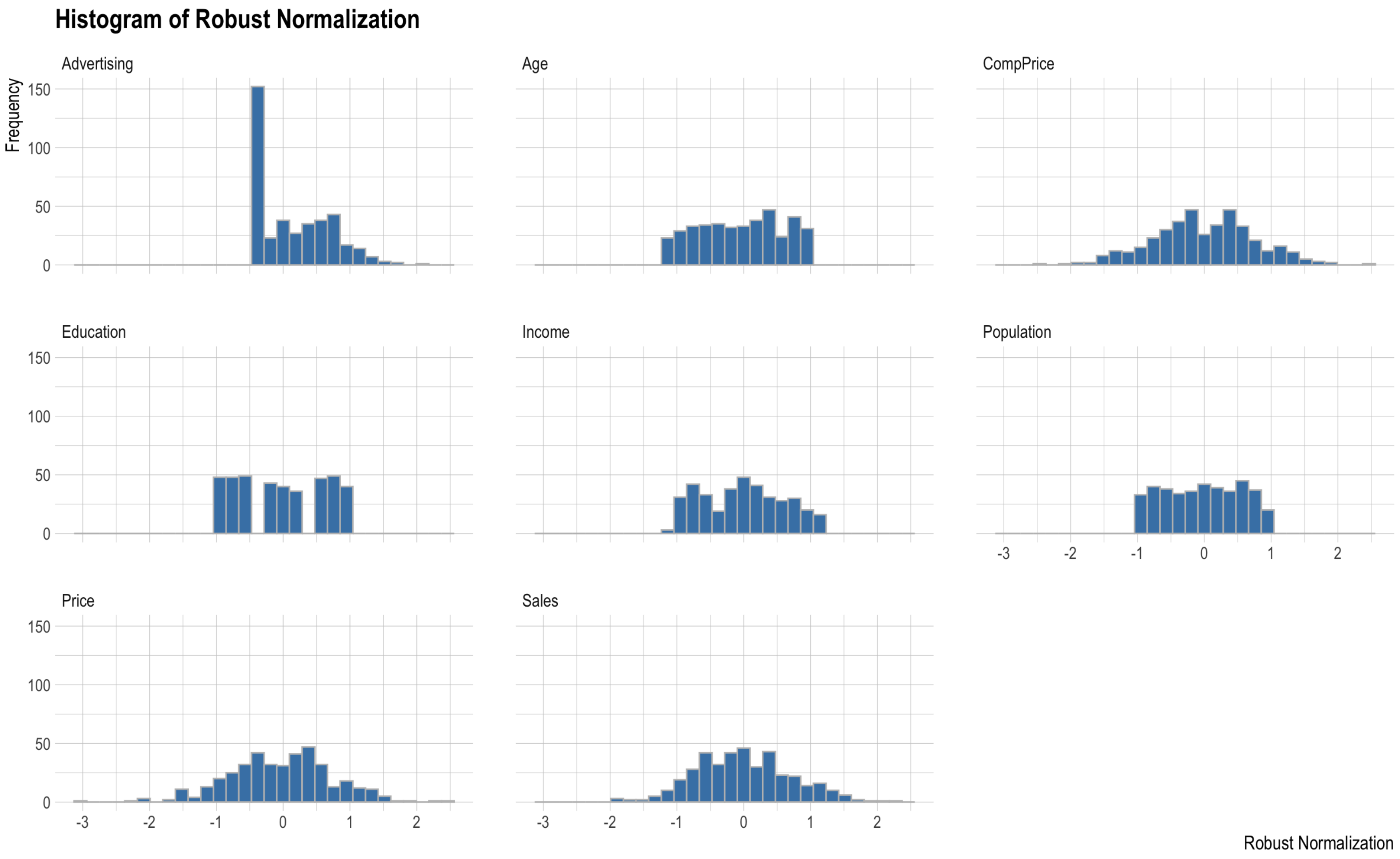
- numerical variables

```
all_var <- carseats %>%
  univar_numeric()

all_var

summary(all_var)
plot(all_var)
```

```
> all_var
$statistics
# A tibble: 8 x 10
  variable        n    na   mean     sd se_mean    IQR skewness kurtosis median
  <chr>       <int> <int>  <dbl>  <dbl>   <dbl>  <dbl>    <dbl>    <dbl>  <dbl>
1 Sales         400     0   7.50   2.82   0.141   3.93   0.186   -0.0809   7.49
2 CompPrice     400     0 125.    15.3    0.767  20     -0.0428   0.0417 125
3 Income        380    20  68.6   27.8    1.43   46      0.0442  -1.07     69
4 Advertising   400     0   6.64   6.65   0.333  12      0.640   -0.545     5
5 Population    400     0 265.   147.     7.37  260.    -0.0512  -1.20    272
6 Price         400     0 116.    23.7    1.18   31     -0.125    0.452   117
7 Age           400     0  53.3   16.2    0.810  26.2   -0.0772  -1.13     54.5
8 Education     400     0  13.9    2.62   0.131   4      0.0440  -1.30     14

> summary(all_var)
# A tibble: 8 x 8
  variable        mean    sd se_mean   IQR skewness kurtosis median
  <chr>          <dbl> <dbl>   <dbl> <dbl>    <dbl>    <dbl>  <dbl>
1 Sales        0.00161 0.719  0.0359     1   0.186   -0.0809      0
2 CompPrice   -0.00125 0.767  0.0383     1  -0.0428   0.0417      0
3 Income      -0.00795 0.604  0.0310     1   0.0442  -1.07        0
4 Advertising  0.136   0.554  0.0277     1   0.640   -0.545       0
5 Population  -0.0276  0.568  0.0284     1  -0.0512  -1.20        0
6 Price       -0.0389  0.764  0.0382     1  -0.125    0.452       0
7 Age         -0.0449  0.617  0.0309     1  -0.0772  -1.13        0
8 Education   -0.025   0.655  0.0328     1   0.0440  -1.30        0
```

일변량 변수 탐색

- numerical variables



이변량 변수 탐색

- categorical variables

```
all_var <- carseats %>%
 compare_category()

all_var

summary(all_var)
plot(all_var)
```

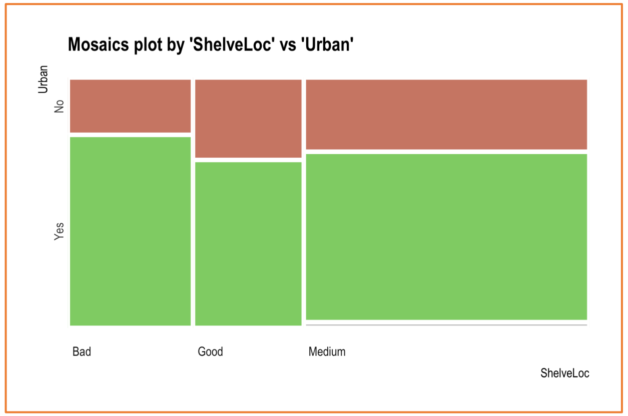
```
> all_var
$`ShelveLoc vs Urban`
# A tibble: 7 x 6
  ShelveLoc Urban     n   rate var1_rate var2_rate
  <fct>     <fct> <int>  <dbl>     <dbl>     <dbl>
1 Bad       No       22 0.055     0.229     0.191
2 Bad       Yes      74 0.185     0.771     0.264
3 Good      No       28 0.07      0.329     0.243
4 Good      Yes      57 0.142     0.671     0.204
5 Medium    No       65 0.162     0.297     0.565
6 Medium    Yes     149 0.372     0.680     0.532
7 Medium    NA        5 0.0125    0.0228    1

$`ShelveLoc vs US`
# A tibble: 6 x 6
  ShelveLoc US        n  rate var1_rate var2_rate
  <fct>     <fct> <int> <dbl>     <dbl>     <dbl>
1 Bad       No       34 0.085     0.354     0.239
2 Bad       Yes      62 0.155     0.646     0.240
3 Good      No       24 0.06      0.282     0.169
4 Good      Yes      61 0.152     0.718     0.236
5 Medium    No       84 0.21      0.384     0.592
6 Medium    Yes     135 0.338     0.616     0.523

$`Urban vs US`
# A tibble: 6 x 6
  Urban US        n   rate var1_rate var2_rate
  <fct> <fct> <int>  <dbl>     <dbl>     <dbl>
1 No    No       44 0.11      0.383     0.310
2 No    Yes      71 0.178     0.617     0.275
3 Yes   No       96 0.24      0.343     0.676
4 Yes   Yes     184 0.46      0.657     0.713
5 NA    No        2 0.005     0.4       0.0141
6 NA    Yes       3 0.0075    0.6       0.0116
```

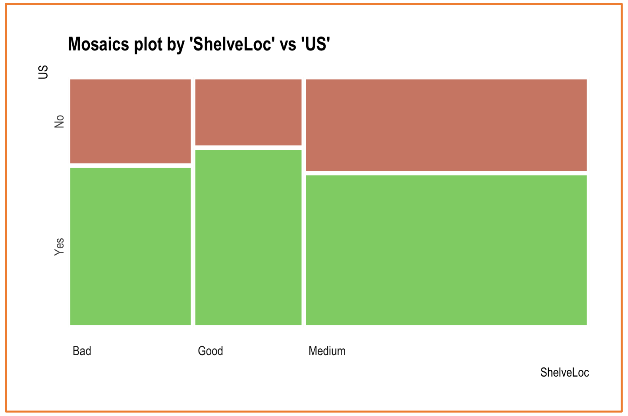
이변량 변수 탐색

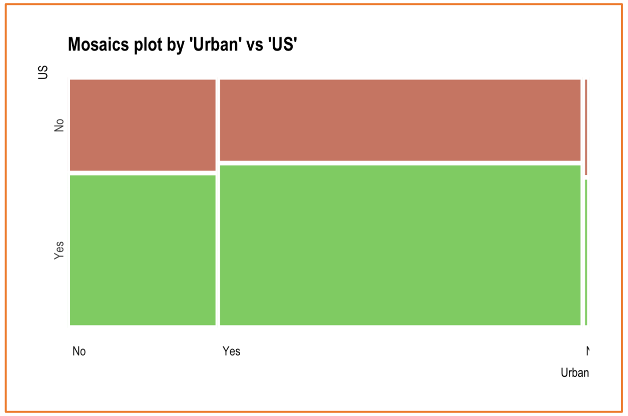
- categorical variables

```
> summary(all_var)
── Contingency tables ──── Number of table is 3 ─
$`ShelveLoc vs Urban`
         Urban
ShelveLoc  No Yes
   Bad     22  74
   Good    28  57
   Medium  65 149

$`ShelveLoc vs US`
         US
ShelveLoc  No Yes
   Bad     34  62
   Good    24  61
   Medium  84 135

$`Urban vs US`
     US
Urban  No Yes
  No   44  71
  Yes  96 184
```

```
── Relative contingency tables ──── Number of table is 3 ──
$`ShelveLoc vs Urban`
         Urban
ShelveLoc         No        Yes
   Bad    0.05569620 0.18734177
   Good   0.07088608 0.14430380
   Medium 0.16455696 0.37721519

$`ShelveLoc vs US`
         US
ShelveLoc     No    Yes
   Bad    0.0850 0.1550
   Good   0.0600 0.1525
   Medium 0.2100 0.3375

$`Urban vs US`
     US
Urban        No       Yes
  No  0.1113924 0.1797468
  Yes 0.2430380 0.4658228

── Chi-squared contingency table tests ─ Number of table is 3 ─
  variable_1 variable_2 statistic   p.value df
1  ShelveLoc      Urban  2.554420 0.2788141  2
2  ShelveLoc         US  2.739667 0.2541492  2
3      Urban         US  0.402651 0.5257233  1
```







이변량 변수 탐색

- numerical variables

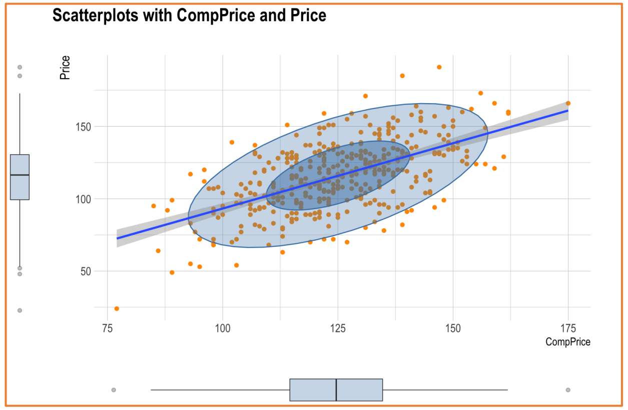
```
all_var <- carseats %>%
  compare_numeric()

all_var

summary(all_var)
plot(all_var)
```

```
> all_var
$correlation
# A tibble: 28 x 3
   var1      var2        coef_corr
   <chr>     <chr>           <dbl>
 1 Sales     CompPrice      0.0641
 2 Sales     Income         0.142
 3 Sales     Advertising    0.270
 4 Sales     Population     0.0505
 5 Sales     Price         -0.445
 6 Sales     Age           -0.232
 7 Sales     Education     -0.0520
 8 CompPrice Income        -0.0815
 9 CompPrice Advertising   -0.0242
10 CompPrice Population    -0.0947
# … with 18 more rows

$linear
# A tibble: 28 x 14
   var1      var2        r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC deviance df.residual  nobs
   <chr>     <chr>           <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
 1 Sales     CompPrice    0.00411       0.00160   2.82     1.64  2.01e- 1     1  -982. 1969. 1981.    3169.         398   400
 2 Sales     Income       0.0203        0.0177    2.82     7.83  5.41e- 3     1  -932. 1870. 1882.    3003.         378   380
 3 Sales     Advertising  0.0726        0.0703    2.72    31.2   4.38e- 8     1  -967. 1941. 1953.    2951.         398   400
 4 Sales     Population   0.00255       0.0000412 2.82     1.02  3.14e- 1     1  -982. 1970. 1982.    3174.         398   400
 5 Sales     Price        0.198         0.196     2.53    98.2   7.62e-21     1  -938. 1882. 1894.    2552.         398   400
 6 Sales     Age          0.0537        0.0514    2.75    22.6   2.79e- 6     1  -971. 1949. 1961.    3011.         398   400
 7 Sales     Education    0.00270       0.000194  2.82     1.08  3.00e- 1     1  -982. 1970. 1982.    3174.         398   400
 8 CompPrice Income       0.00664       0.00401  15.3      2.53  1.13e- 1     1 -1575. 3156. 3168.   88650.         378   380
 9 CompPrice Advertising  0.000586     -0.00193  15.3      0.233 6.29e- 1     1 -1659. 3324. 3336.   93769.         398   400
10 CompPrice Population   0.00897       0.00648  15.3      3.60  5.84e- 2     1 -1657. 3321. 3333.   92982.         398   400
# … with 18 more rows
```

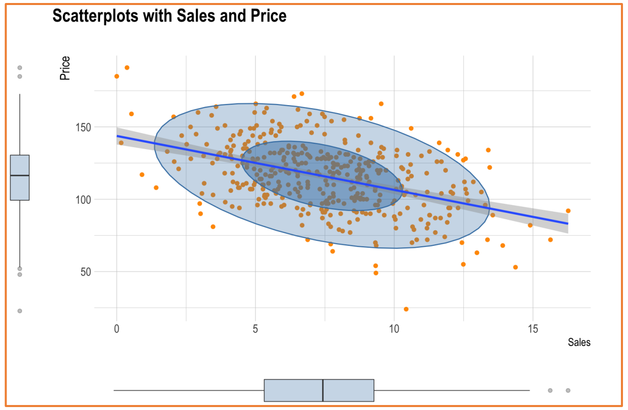
이변량 변수 탐색

- numerical variables

```
> summary(all_var)
── Correlation check : abs(r) > 0.3 ──────────────────── Number of pairs is 2/28 ──
# A tibble: 2 x 3
  var1      var2  coef_corr
  <chr>     <chr>     <dbl>
1 CompPrice Price     0.585
2 Sales     Price    -0.445
── R.squared check : R^2 > 0.1 ─────────────────────────── Number of pairs is 2/28 ──
# A tibble: 2 x 14
  var1      var2  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC deviance df.residual  nobs
  <chr>     <chr>     <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
1 CompPrice Price     0.342         0.340 12.5      207.  4.50e-38     1 -1575. 3157. 3169.   61732.         398   400
2 Sales     Price     0.198         0.196  2.53      98.2 7.62e-21     1  -938. 1882. 1894.    2552.         398   400
```



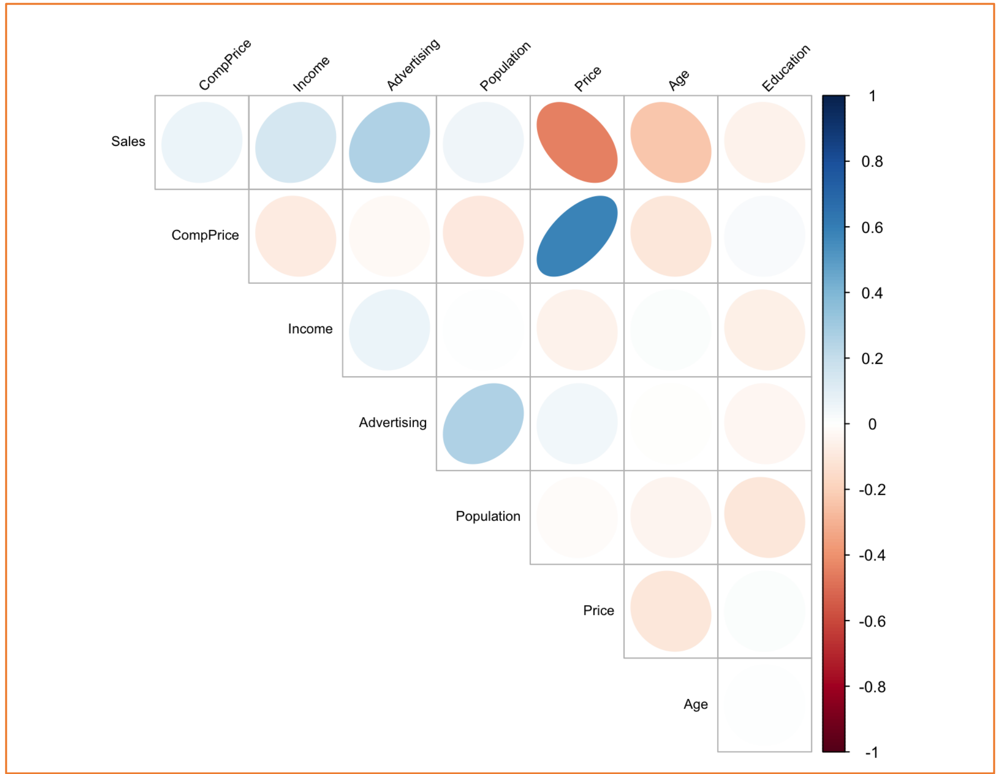
...



상관관계 파악

```
carseats %>%
  correlate() %>%
  filter(as.integer(var1) > as.integer(var2))

carseats %>%
  plot_correlate()
```

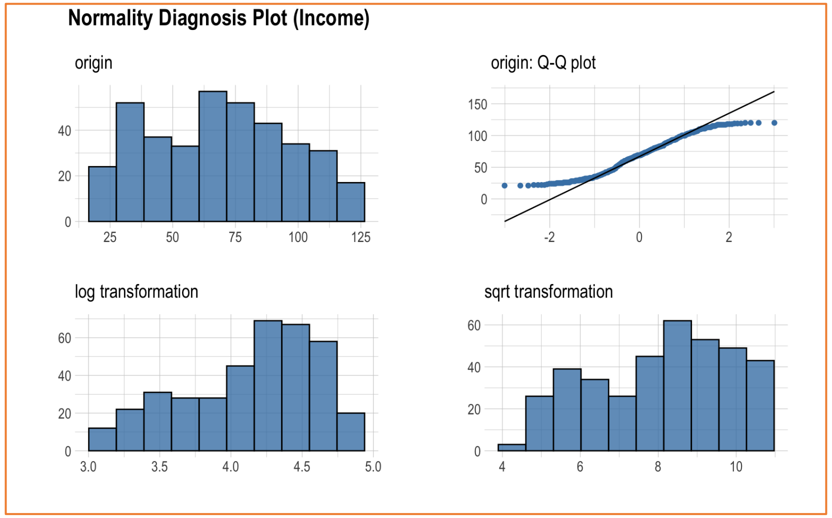
```
# A tibble: 28 x 3
   var1        var2      coef_corr
   <fct>       <fct>         <dbl>
 1 CompPrice   Sales        0.0641
 2 Income      Sales        0.142
 3 Advertising Sales        0.270
 4 Population  Sales        0.0505
 5 Price       Sales       -0.445
 6 Age         Sales       -0.232
 7 Education   Sales       -0.0520
 8 Income      CompPrice   -0.0815
 9 Advertising CompPrice   -0.0242
10 Population  CompPrice   -0.0947
# … with 18 more rows
```

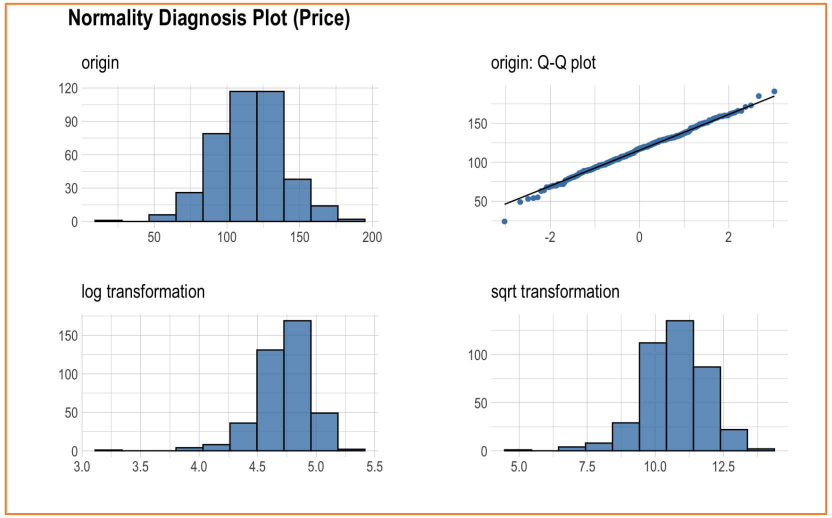


정규성 검정

```
carseats %>%
  normality()

carseats %>%
  plot_normality("Income", "Price")
```

```
# A tibble: 8 x 4
  vars          statistic  p_value sample
  <chr>             <dbl>    <dbl>  <dbl>
1 Sales             0.995 2.54e- 1    400
2 CompPrice         0.998 9.77e- 1    400
3 Income            0.962 2.38e- 8    400
4 Advertising       0.874 1.49e-17    400
5 Population        0.952 4.08e-10    400
6 Price             0.996 3.90e- 1    400
7 Age               0.957 1.86e- 9    400
8 Education         0.924 2.43e-13    400
```



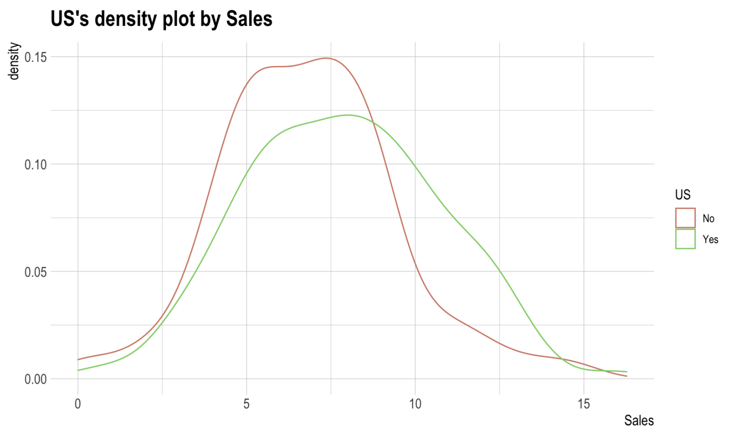


인과관계 파악

- target variable: category, indicator: numeric.

```r
categ <- carseats %>%
  target_by(US)


cat_num <- categ %>%
  relate(Sales)

cat_num

summary(cat_num)

plot(cat_num)
```

```
> cat_num
# A tibble: 3 x 27
  variable US        n    na  mean    sd se_mean   IQR skewness kurtosis   p00   p01   p05   p10   p20
  <chr>    <fct> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl>    <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 Sales    No      142     0  6.82  2.60   0.218  3.44   0.323     0.808  0    0.468  3.25  3.92  4.75
2 Sales    Yes     258     0  7.87  2.88   0.179  4.23   0.0760   -0.326  0.37 1.65   3.15  4.18  5.33
3 Sales    total   400     0  7.50  2.82   0.141  3.93   0.186    -0.0809 0    0.906  3.15  4.12  5.07
# … with 12 more variables: p25 <dbl>, p30 <dbl>, p40 <dbl>, p50 <dbl>, p60 <dbl>, p70 <dbl>, p75 <dbl>,
#   p80 <dbl>, p90 <dbl>, p95 <dbl>, p99 <dbl>, p100 <dbl>

> summary(cat_num)
   variable              US          n                 na           mean             sd
 Length:3           No   :1   Min.   :142.0   Min.   :0   Min.   :6.823   Min.   :2.603
 Class :character   Yes  :1   1st Qu.:200.0   1st Qu.:0   1st Qu.:7.160   1st Qu.:2.713
 Mode  :character   total:1   Median :258.0   Median :0   Median :7.496   Median :2.824
                              Mean   :266.7   Mean   :0   Mean   :7.395   Mean   :2.768
                              3rd Qu.:329.0   3rd Qu.:0   3rd Qu.:7.682   3rd Qu.:2.851
                              Max.   :400.0   Max.   :0   Max.   :7.867   Max.   :2.877
    se_mean             IQR             skewness
 Min.   :0.1412   Min.   :3.442   Min.   :0.07603
 1st Qu.:0.1602   1st Qu.:3.686   1st Qu.:0.13080
 Median :0.1791   Median :3.930   Median :0.18556
 Mean   :0.1796   Mean   :3.866   Mean   :0.19489
 3rd Qu.:0.1988   3rd Qu.:4.077   3rd Qu.:0.25432
 Max.   :0.2184   Max.   :4.225   Max.   :0.32308
                                              ...
p95              p99              p100
 Min.   :11.28   Min.   :13.64   Min.   :14.90
 1st Qu.:11.86   1st Qu.:13.78   1st Qu.:15.59
 Median :12.44   Median :13.91   Median :16.27
 Mean   :12.08   Mean   :13.86   Mean   :15.81
 3rd Qu.:12.49   3rd Qu.:13.97   3rd Qu.:16.27
 Max.   :12.54   Max.   :14.03   Max.   :16.27
```

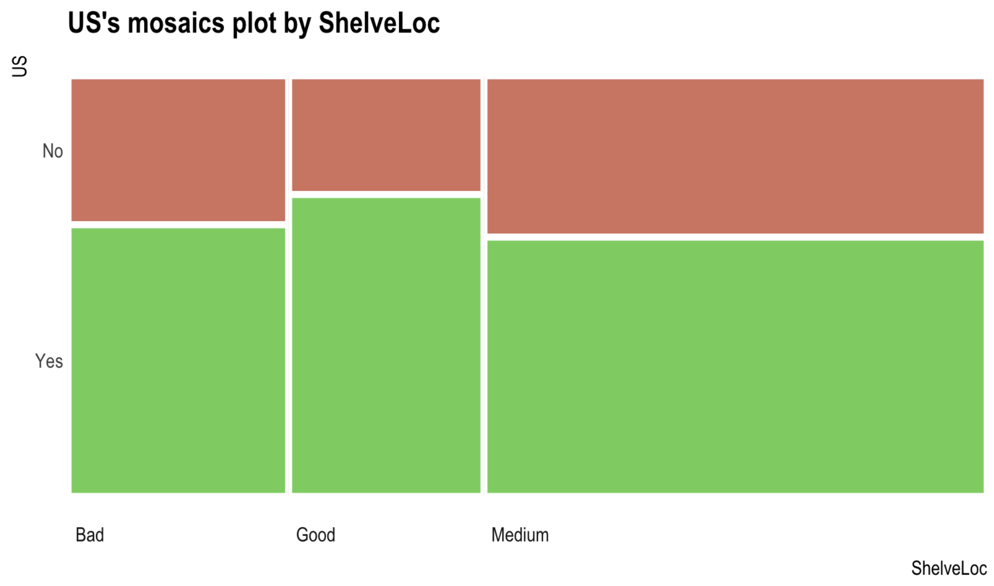


인과관계 파악

- target variable: category, indicator: category.

```r
categ <- carseats %>%
  target_by(US)


cat_cat <- categ %>%
  relate(ShelveLoc)

cat_cat

summary(cat_cat)

plot(cat_cat)
```

```
> cat_cat
     ShelveLoc
US     Bad Good Medium
  No    34   24     84
  Yes   62   61    135

> summary(cat_cat)
Call: xtabs(formula = formula_str, data = data, addNA = TRUE)
Number of cases in table: 400
Number of factors: 2
Test for independence of all factors:
              Chisq = 2.7397, df = 2, p-value = 0.2541
```

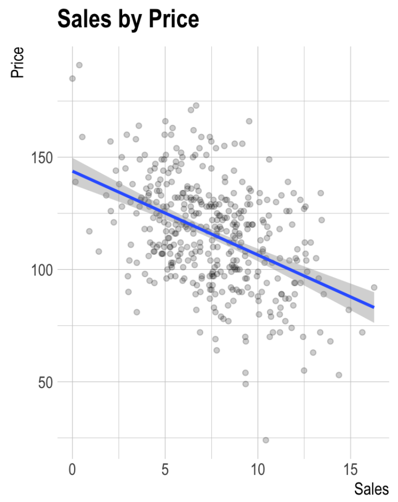


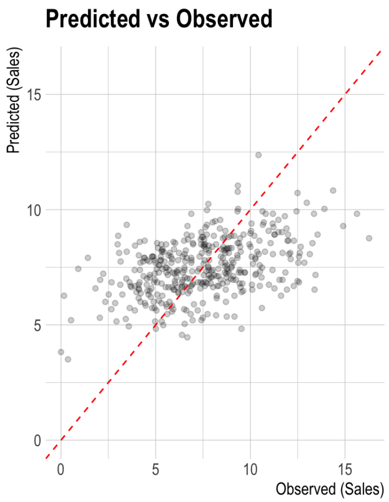
인과관계 파악

- target variable: numeric, indicator: numeric.

```
categ <- carseats %>%
  target_by(Sales)


num_num <- categ %>%
  relate(Price)

num_num

summary(num_num)

plot(num_num)
```

```
> num_num

Call:
lm(formula = formula_str, data = data)

Coefficients:
(Intercept)        Price
   13.64192     -0.05307

> summary(num_num)

Call:
lm(formula = formula_str, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5224 -1.8442 -0.1459  1.6503  7.5108

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.641915   0.632812  21.558   <2e-16 ***
Price       -0.053073   0.005354  -9.912   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.532 on 398 degrees of freedom
Multiple R-squared:  0.198, Adjusted R-squared:  0.196
F-statistic: 98.25 on 1 and 398 DF,  p-value: < 2.2e-16
```



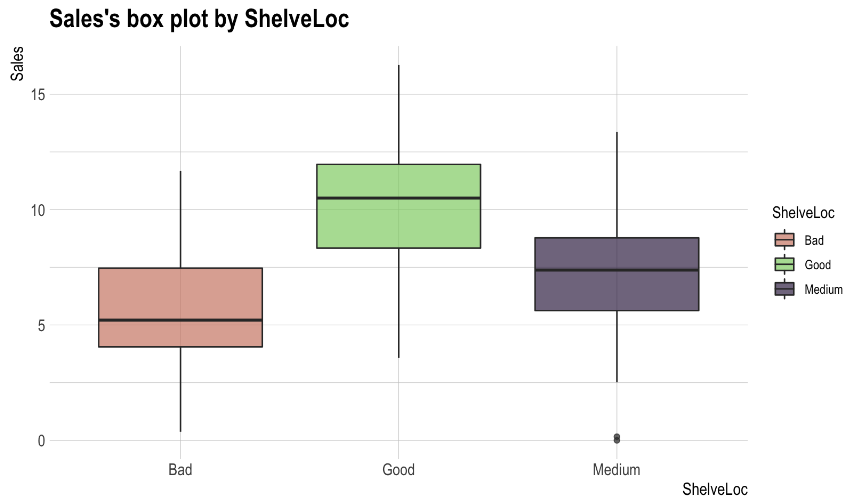


인과관계 파악

- target variable: numeric, indicator: category.

```r
categ <- carseats %>%
  target_by(Sales)


num_cat <- categ %>%
  relate(ShelveLoc)

num_cat

summary(num_cat)

plot(num_cat)
```

```
> num_cat
Analysis of Variance Table

Response: Sales
           Df Sum Sq Mean Sq F value    Pr(>F)
ShelveLoc   2 1009.5  504.77   92.23 < 2.2e-16 *
Residuals 397 2172.7    5.47
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*
> summary(num_cat)

Call:
lm(formula = formula(formula_str), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3066 -1.6282 -0.0416  1.5666  6.1471

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.5229     0.2388  23.131  < 2e-16 ***
ShelveLocGood     4.6911     0.3484  13.464  < 2e-16 ***
ShelveLocMedium   1.7837     0.2864   6.229  1.2e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.339 on 397 degrees of freedom
Multiple R-squared:  0.3172,	Adjusted R-squared:  0.3138
F-statistic: 92.23 on 2 and 397 DF,  p-value: < 2.2e-16
```



Automated Report - pdf

- 데이터 품질 진단, EDA, 데이터 변환 3종의 자동화 리포트 지원
- pdf, html의 2가지 리포트 포맷

REPORT SERIES WITH DLOOKR

Data Quality Diagnosis Report

Author:
dlookr package

Version:
0.3.15

January 21, 2021

Contents

ShelveLoc

1. Analysis of Variance

Table 4.8: Analysis of Variance Table : ShelveLoc

	Df	Sum Sq	Mean Sq	F value	Pr(> | F |)
ShelveLoc	2	1009.53	504.77	92.23	0
Residuals	397	2172.74	5.47	NA	NA

2. Simple Linear Model Information

Residual standard error: 2 on 397 degrees of freedom
Multiple R-squared: 0.31724, Adjusted R-squared: 0.3138
F-statistic: 92 on 2 and 397 DF, p-value: 0

Table 4.9: Simple Linear Model coefficients : ShelveLoc

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	5.52	0.24	23.13	0
ShelveLocGood	4.69	0.35	13.46	0
ShelveLocMedium	1.78	0.29	6.23	0

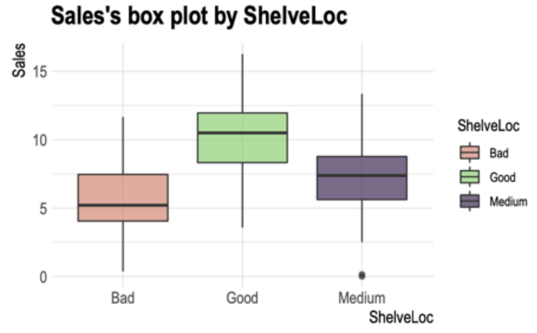


Figure 4.8: ShelveLoc

Automated Report - html

Data Quality Diagnosis Report

Report by dlookr package

2021-03-02

1.2.2 Diagnosis of numerical variables

General list of numerical diagnosis

variables	min	Q1	mean	median	Q3	max	zero	minus	outlier
Sales	0	5.39	7.50	7.49	9.32	16.27	1	0	2
CompPrice	77	115.00	124.97	125.00	135.00	175.00	0	0	2
Income	21	42.00	68.01	68.50	90.00	120.00	0	0	0
Advertising	0	0.00	6.64	5.00	12.00	29.00	144	0	0
Population	10	139.00	264.84	272.00	398.50	509.00	0	0	0
Price	24	100.00	115.80	117.00	131.00	191.00	0	0	5
Age	25	39.75	53.32	54.50	66.00	80.00	0	0	0
Education	10	12.00	13.90	14.00	16.00	18.00	0	0	0

1.2.3 List of numerical diagnosis (zero)

List of numerical diagnosis (zero)

variables	min	median	max	zero	zero ratio(%)
Advertising	0	5.00	29.00	144	36.00
Sales	0	7.49	16.27	1	0.25

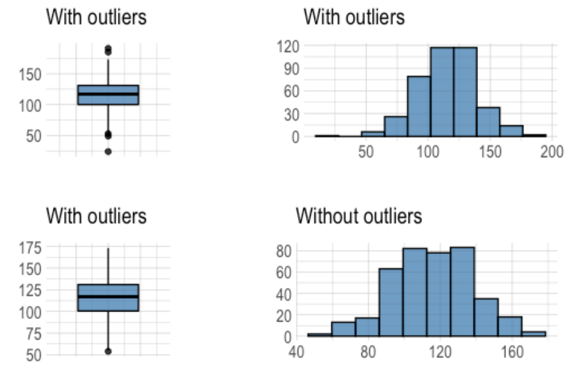
2.2 Detailed outliers diagnosis

variable : Price

Outliers information of Price

Measures	Values
Outliers count	5.00
Outliers ratio (%)	1.25
Mean of outliers	100.40
Mean with outliers	115.80
Mean without outliers	115.99

Outlier Diagnosis Plot (Price)



Supported DBMS table

- data.frame, tibble 뿐만 아니라, DBMS의 table에 포함된 데이터도 지원함 (모든 기능이 아닌 일부 기능 지원)

```
library(dplyr)

# Generate data for the example
carseats <- ISLR::Carseats
carseats[sample(seq(NROW(carseats)), 20), "Income"] <- NA
carseats[sample(seq(NROW(carseats)), 5), "Urban"] <- NA

# connect DBMS
con_sqlite <- DBI::dbConnect(RSQLite::SQLite(), ":memory:")

# copy carseats to the DBMS with a table named TB_CARSEATS
copy_to(con_sqlite, carseats, name = "TB_CARSEATS", overwrite = TRUE)

# describe from DBMS
con_sqlite %>%
  tbl("TB_CARSEATS") %>%
  describe(Sales, CompPrice, Income)
```

```
# A tibble: 3 x 26
  variable      n    na   mean    sd se_mean   IQR skewness kurtosis   p00    p01   p05   p10   p20   p25   p30   p40   p50   p60   p70
  <chr>     <int> <int>  <dbl> <dbl>   <dbl> <dbl>    <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 Sales       400     0   7.50  2.82   0.141  3.93   0.186   -0.0809     0  0.906  3.15  4.12  5.07  5.39  5.87  6.61  7.49  8.08  8.81
2 CompPrice   400     0 125.   15.3    0.767 20     -0.0428   0.0417    77 89.0   98    106   113.  115   117   121   125   130   133
3 Income      380    20  68.1  28.1    1.44  48      0.0800  -1.09      21 21.8   26.0   30    38    42    47.7  60.6  68.5  76.4  84
# … with 6 more variables: p75 <dbl>, p80 <dbl>, p90 <dbl>, p95 <dbl>, p99 <dbl>, p100 <dbl>
```

Collaborated tidyverse packages

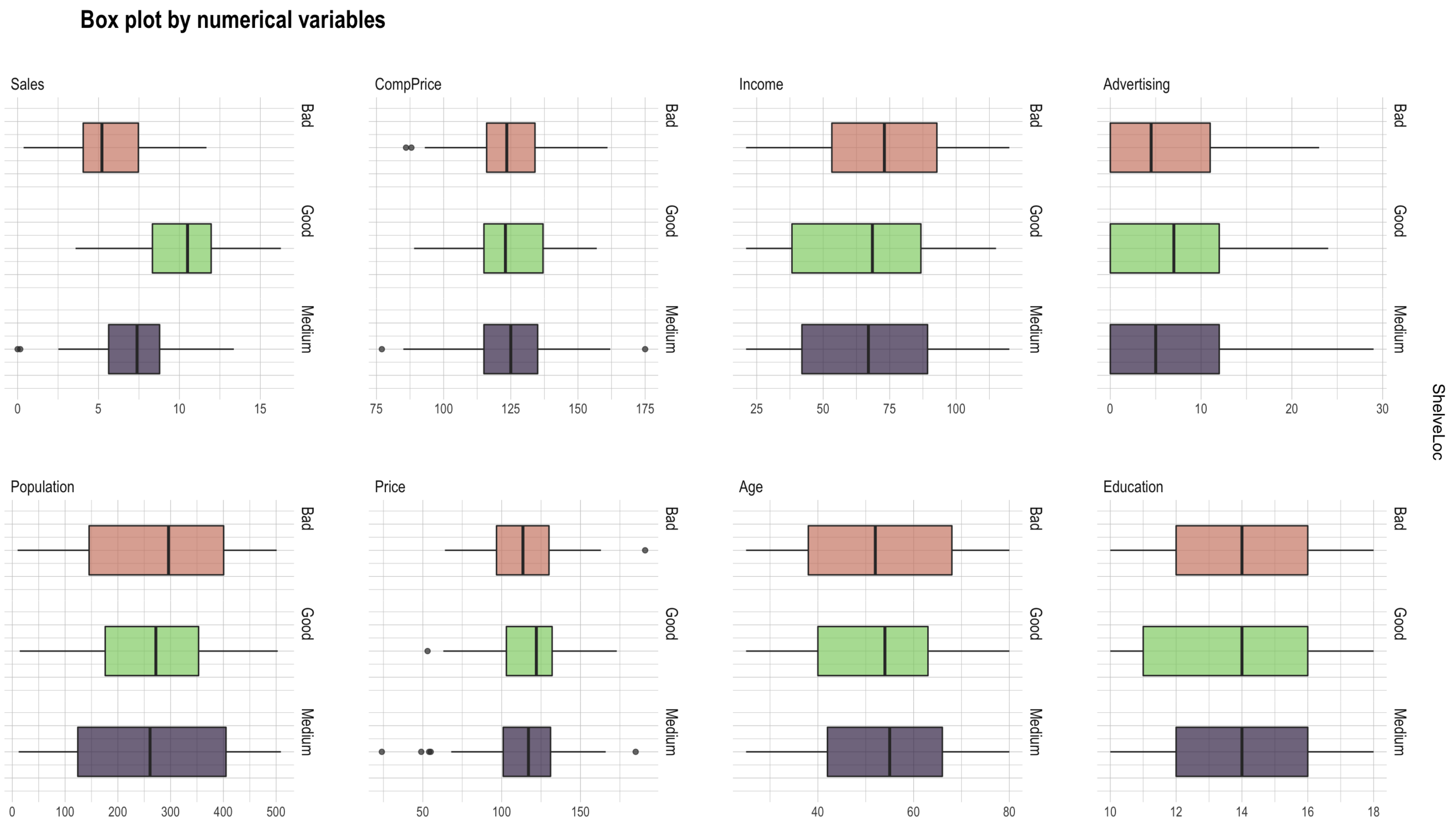
- select, mutate, filter, group_by 등 tidyverse packages의 함수와 혼용 가능

```
# select와 같은 기능 내재화
carseats %>%
  diagnose(Sales, Age)

# select 사용
carseats %>%
  select(Sales, Age) %>%
  diagnose()

# ShelveLoc, US별로 상관관계 파악
carseats %>%
  group_by(ShelveLoc, US) %>%
  correlate(Sales) %>%
  filter(abs(coef_corr) >= 0.5)

# 로그 변환 후 특정 그룹별로 정규성 검정
carseats %>%
  mutate(log_income = log(Income)) %>%
  group(ShelveLoc, US) %>%
  normality(log_income) %>%
  filter(p_value > 0.01)
```

```
# A tibble: 2 x 6
  variables types   missing_count missing_percent unique_count unique_rate
  <chr>     <chr>           <int>           <dbl>        <int>       <dbl>
1 Sales     numeric             0               0          336        0.84
2 Age       numeric             0               0           56        0.14

# A tibble: 2 x 6
  variables types   missing_count missing_percent unique_count unique_rate
  <chr>     <chr>           <int>           <dbl>        <int>       <dbl>
1 Sales     numeric             0               0          336        0.84
2 Age       numeric             0               0           56        0.14

# A tibble: 6 x 5
  ShelveLoc US    var1  var2  coef_corr
  <fct>     <fct> <fct> <fct>     <dbl>
1 Bad       No    Sales Price    -0.527
2 Bad       Yes   Sales Price    -0.583
3 Good      No    Sales Price    -0.811
4 Good      Yes   Sales Price    -0.603
5 Medium    No    Sales Price    -0.610
6 Medium    Yes   Sales Price    -0.538

# A tibble: 1 x 6
  variable   ShelveLoc US    statistic p_value sample
  <chr>      <fct>     <fct>     <dbl>   <dbl>  <dbl>
1 log_income Bad       No        0.945  0.0873     34
```

Collaborated tidyverse packages

```
# ShelveLoc별로 박스 플롯 시각화
carseats %>%
 group_by(ShelveLoc) %>%
  plot_box_numeric()
```



E. O. D