PS1 - IO

Tate Mason

1 - Data Import and Descriptive Statistics

```
# Load necessary libraries
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(readr)
library(psych)
```

```
Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

    %+%, alpha
```

```
# Import the dataset
df <- read_delim("~/SchoolWork/Y2S1/IO/PS1/IRI.csv")
```

```
Rows: 63966 Columns: 18
-- Column specification --------------------------------------------------------
Delimiter: "\t"
chr  (5): market_name, state, brand, parent, segment
dbl (13): year, month, store_id, week_id, flavored, fortified, fiber, sugars...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(df)
```

```
Rows: 63,966
Columns: 18
$ year        <dbl> 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008~
$ month       <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
$ store_id    <dbl> 369, 108, 374, 221, 294, 117, 378, 310, 237, 321, 120, 275~
$ week_id     <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5~
$ market_name <chr> "BOSTON", "BOSTON", "BOSTON", "BOSTON", "BOSTON", "BOSTON"~
$ state       <chr> "Massachusetts", "Massachusetts", "Massachusetts", "Massac~
$ brand       <chr> "KELLOGGS RAISIN BRAN CRUNCH", "GENERAL MILLS GOLDEN GRAHA~
$ parent      <chr> "KELLOGGS", "GENERAL MILLS", "KELLOGGS", "GENERAL MILLS", ~
$ flavored    <dbl> 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1~
$ fortified   <dbl> 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0~
$ fiber       <dbl> 3.759398, 0.000000, 0.000000, 0.000000, 0.000000, 2.000000~
$ sugars      <dbl> 9.022556, 9.000000, 3.000000, 13.000000, 3.000000, 8.00000~
$ price       <dbl> 0.1680204, 0.2733062, 0.1723915, 0.2528735, 0.3389252, 0.2~
$ quantity    <dbl> 600.6000, 96.0000, 1889.2100, 165.1040, 60.0000, 224.0000,~
$ puma        <dbl> 2700, 2700, 2700, 2700, 2700, 2700, 2700, 2700, 2700, 2700~
$ sugar_price <dbl> 211.9174, 211.9174, 211.9174, 211.9174, 211.9174, 211.9174~
$ M           <dbl> 117679.10, 55841.92, 204192.90, 29956.03, 35946.88, 96805.~
$ segment     <chr> "Family", "Family", "Child", "Child", "Child", "Adult", "A~
```

```
describe(df$brand)
```

```
    vars     n  mean    sd median trimmed   mad min max range skew kurtosis
X1*    1 63966 19.96 11.28     20   19.95 14.83   1  39    38    0    -1.21
      se
X1* 0.04
```

```r
describe(df$parent)
```

```
    vars     n mean   sd median trimmed  mad min max range skew kurtosis se
X1*    1 63966 2.12 1.09      2    2.03 1.48   1   4     3 0.58    -0.97  0
```