ANA515 Assignment 2

Jiaxuan Ren

2023-06-18

```
# Global options that apply to every chunk.
knitr::opts_chunk$set(echo = TRUE, message=FALSE, tidy=TRUE, tidy.opts=list(width.cutoff=60))
```

Description of the data

The candy-power-ranking dataset measures "How often did a fun-sized candy of a given type win its matchups against the rest of the field?" and I hope to use this dataset to answer research question "What Halloween candy people most prefer?". The dataset was saved in a 'csv' file. If it is saved in a flat file, I would like to set a fixed width and delimit each attribute by a comma (,) or the tab character (\t) to make the data well-formatted and well-look. If it is saved as a binary file, it can be opened based on their file type (e.g., IDEs like Visual Studio, Eclipse, or Xcode are primarily used for programming and can handle binary files, especially for source code files and executables).

Read Data

```
# Load tidyverse library
library(tidyverse)

# Use read_csv function to read the data 'candy-data.csv'
mydata <- read_csv("candy-data.csv")

# Print out all columns of the data
glimpse(mydata)
```

```
## Rows: 85
## Columns: 13
## $ competitorname   <chr> "100 Grand", "3 Musketeers", "One dime", "One quarter~
## $ chocolate        <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ fruity           <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1,~
## $ caramel          <dbl> 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ peanutyalmondy   <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ nougat           <dbl> 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ crispedricewafer <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ hard             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1,~
## $ bar              <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ pluribus         <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1,~
## $ sugarpercent     <dbl> 0.732, 0.604, 0.011, 0.011, 0.906, 0.465, 0.604, 0.31~
## $ pricepercent     <dbl> 0.860, 0.511, 0.116, 0.511, 0.511, 0.767, 0.767, 0.51~
## $ winpercent       <dbl> 66.97173, 67.60294, 32.26109, 46.11650, 52.34146, 50.~
```

Clean the data

```r
# Rename the 'competitorname' column as 'candy_brand' and
# assigned the cleaned data into a new object called
# 'cleaned_data'.
cleaned_data <- mydata %>%
    rename(candy_brand = "competitorname")
glimpse(cleaned_data)
```

```
## Rows: 85
## Columns: 13
## $ candy_brand      <chr> "100 Grand", "3 Musketeers", "One dime", "One quarter~
## $ chocolate        <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ fruity           <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1,~
## $ caramel          <dbl> 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ peanutyalmondy   <dbl> 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ nougat           <dbl> 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ crispedricewafer <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ hard             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1,~
## $ bar              <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ pluribus         <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1,~
## $ sugarpercent     <dbl> 0.732, 0.604, 0.011, 0.011, 0.906, 0.465, 0.604, 0.31~
## $ pricepercent     <dbl> 0.860, 0.511, 0.116, 0.511, 0.511, 0.767, 0.767, 0.51~
## $ winpercent       <dbl> 66.97173, 67.60294, 32.26109, 46.11650, 52.34146, 50.~
```

```r
# Filter the data to only include top-10 'winpercent' candy
# and assigned the filtered data into a new object called
# 'filtered_data'.
filtered_data <- cleaned_data %>%
    top_n(10, winpercent)
glimpse(filtered_data)
```

```
## Rows: 10
## Columns: 13
## $ candy_brand      <chr> "Kit Kat", "Peanut butter M&M's", "Milky Way", "Nestl~
## $ chocolate        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
## $ fruity           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ caramel          <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 1, 1
## $ peanutyalmondy   <dbl> 0, 1, 0, 1, 1, 1, 1, 1, 1, 0
## $ nougat           <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 1, 0
## $ crispedricewafer <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 1
## $ hard             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
## $ bar              <dbl> 1, 0, 1, 1, 0, 0, 0, 0, 1, 1
## $ pluribus         <dbl> 0, 1, 0, 0, 0, 0, 1, 0, 0, 0
## $ sugarpercent     <dbl> 0.313, 0.825, 0.604, 0.604, 0.034, 0.720, 0.406, 0.98~
## $ pricepercent     <dbl> 0.511, 0.651, 0.651, 0.767, 0.279, 0.651, 0.651, 0.65~
## $ winpercent       <dbl> 76.76860, 71.46505, 73.09956, 70.73564, 81.86626, 84.~
```

Characteristics of the data

This dataframe has 85 rows and 13 columns. After cleaning and filtering, 10 rows and 13 columns are kept for further analysis. The names of the columns and a brief description of each are in the table below:

```r
# This makes a new data.frame called 'Column Description'
# with two columns, 'Columns' and 'Description'
library(knitr)  # import knitr library
```

```r
col_summary <- data.frame(Columns = c(colnames(filtered_data)),
    Description = c("the brand name of candy.", "Does it contain chocolate?",
        "Is it fruit flavored?", "Is there caramel in the candy?",
        "Does it contain peanuts, peanut butter or almonds?",
        "Does it contain nougat?", "Does it contain crisped rice, wafers, or a cookie component?",
        "Is it a hard candy?", "Is it a candy bar?", "Is it one of many candies in a bag or box?",
        "The percentile of sugar it falls under within the data set.",
        "The unit price percentile compared to the rest of the set.",
        "The overall win percentage according to 269,000 matchups."))

kable(col_summary, caption = "Column Description")
```

Table 1: Column Description

Columns	Description
candy_brand	the brand name of candy.
chocolate	Does it contain chocolate?
fruity	Is it fruit flavored?
caramel	Is there caramel in the candy?
peanutyalmondy	Does it contain peanuts, peanut butter or almonds?
nougat	Does it contain nougat?
crispedricewafer	Does it contain crisped rice, wafers, or a cookie component?
hard	Is it a hard candy?
bar	Is it a candy bar?
pluribus	Is it one of many candies in a bag or box?
sugarpercent	The percentile of sugar it falls under within the data set.
pricepercent	The unit price percentile compared to the rest of the set.
winpercent	The overall win percentage according to 269,000 matchups.

Summary statistics

The statistics were calculated on columns 'sugarpercent', 'pricepercent', and 'winpercent'. There is no missing values for all columns in the dataset.

```r
# Summary statistics on column 'sugarpercent'
sugar_sum <- summary(filtered_data["sugarpercent"])
print(sugar_sum)
```

```
##   sugarpercent
##  Min.   :0.0340
##  1st Qu.:0.4410
##  Median :0.5750
##  Mean   :0.5586
##  3rd Qu.:0.6910
##  Max.   :0.9880
```

```r
# Summary statistics on column 'pricepercent'
price_sum <- summary(filtered_data["pricepercent"])
print(price_sum)
```

```
##   pricepercent
##  Min.   :0.2790
##  1st Qu.:0.6510
##  Median :0.6510
##  Mean   :0.6369
##  3rd Qu.:0.6510
##  Max.   :0.9060
```

```
# Summary statistics on column 'winpercent'
win_sum <- summary(filtered_data["winpercent"])
print(win_sum)
```

```
##    winpercent
##  Min.   :70.74
##  1st Qu.:72.94
##  Median :75.05
##  Mean   :76.28
##  3rd Qu.:80.42
##  Max.   :84.18
```