candy_rankings538

4/13/2018

ชื่อ : นาย รัตนวิช สินบวรพิชัย

รหัสนักศึกษา: 60070084

ชื่อ : นาย ศุภกิตติ์ เธียรธัญญกิจ

รหัสนักศึกษา : 60070098

เวลาเรียน : อังคาร 13.30 - 16.30

โจทย์ปัญหา:

1ช็อคโกแลตเป็นส่วนประกอบที่ถูกใช้ในขนมมากที่สุดจริงหรือไม่?

2.ขนมชนิดใดมีส่วนประกอบของน้ำตาลมากที่สุด?

ชุดข้อมูลที่ใช้ fivethirtyeight : candy_rankings

1 Load libraries

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------------
--- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------------------ ti
dyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fivethirtyeight)
```

2 Look at the data (candy_rankings)

```
# print out the dataframe to see what it looks like
candy_rankings
```

```
## # A tibble: 85 x 13
##    competitorname chocolate fruity caramel peanutyalmondy nougat
##    <chr>          <lgl>     <lgl>  <lgl>   <lgl>          <lgl>
##  1 100 Grand      TRUE      FALSE  TRUE    FALSE          FALSE
##  2 3 Musketeers   TRUE      FALSE  FALSE   FALSE          TRUE
##  3 One dime       FALSE     FALSE  FALSE   FALSE          FALSE
##  4 One quarter    FALSE     FALSE  FALSE   FALSE          FALSE
##  5 Air Heads      FALSE     TRUE   FALSE   FALSE          FALSE
##  6 Almond Joy     TRUE      FALSE  FALSE   TRUE           FALSE
##  7 Baby Ruth      TRUE      FALSE  TRUE    TRUE           TRUE
##  8 Boston Baked ~ FALSE     FALSE  FALSE   TRUE           FALSE
##  9 Candy Corn     FALSE     FALSE  FALSE   FALSE          FALSE
## 10 Caramel Apple~ FALSE     TRUE   TRUE    FALSE          FALSE
## # ... with 75 more rows, and 7 more variables: crispedricewafer <lgl>,
## #   hard <lgl>, bar <lgl>, pluribus <lgl>, sugarpercent <dbl>,
## #   pricepercent <dbl>, winpercent <dbl>
```

3 Looking for answers

3.1 หาว่าในขนมทั้งหมดช็อคโกแลตถูกใช้เป็นส่วนประกอบในขนมกี่ชนิด

- **เลือกคอลัมน์ที่เป็นวัตถุดิบทั้งหมด**
- **นับจำนวน** TRUE **ใน คอลัมน์นั้นๆ**

```
ingredient <- candy_rankings %>%
              # select ingredients column
              select(chocolate:crispedricewafer)

#convert dataframe to vector
d1 <- c(ingredient)

#set v1 to all "chocolate" values
v1 <- d1[[1]]
#set v2 to all "fruity" values
v2 <- d1[[2]]
#set v3 to all "caramel" values
v3 <- d1[[3]]
#set v4 to all "peanutyalmondy" values
v4 <- d1[[4]]
#set v5 to all "nougat" values
v5 <- d1[[5]]
#set v6 to all "crispedricewafer" values
v6 <- d1[[6]]

#create vector of name column
ingredient_name <- c("chocolate", "fruity", "caramel", "peanutyalmondy", "nougat", "crispedricewafer")

#create vector of amount column by condition: length of "TRUE" value
amount <- c(
            length(v1[v1 == TRUE]),
            length(v2[v2 == TRUE]),
            length(v3[v3 == TRUE]),
            length(v4[v4 == TRUE]),
            length(v5[v5 == TRUE]),
            length(v6[v6 == TRUE])
            )

#combind 2 vector
combindmt <- cbind(ingredient_name, amount)

#convert vector to dataframe
sumdf <- as.data.frame(combindmt)

#sort dataframe adapt from https://www.statmethods.net/management/sorting.html
attach(sumdf)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     amount, ingredient_name
```

```
sortedsumdf <- sumdf[order(-amount),]

sortedsumdf
```

```
##    ingredient_name amount
## 2           fruity     38
## 1        chocolate     37
## 3          caramel     14
## 4   peanutyalmondy     14
## 5           nougat      7
## 6 crispedricewafer      7
```

chocolate is in the 2nd place with the total consumption of 37 candies

3.2 หา 20 อันดับขนมที่มีน้ำตาลมากที่สุก และเรียงลำดับ

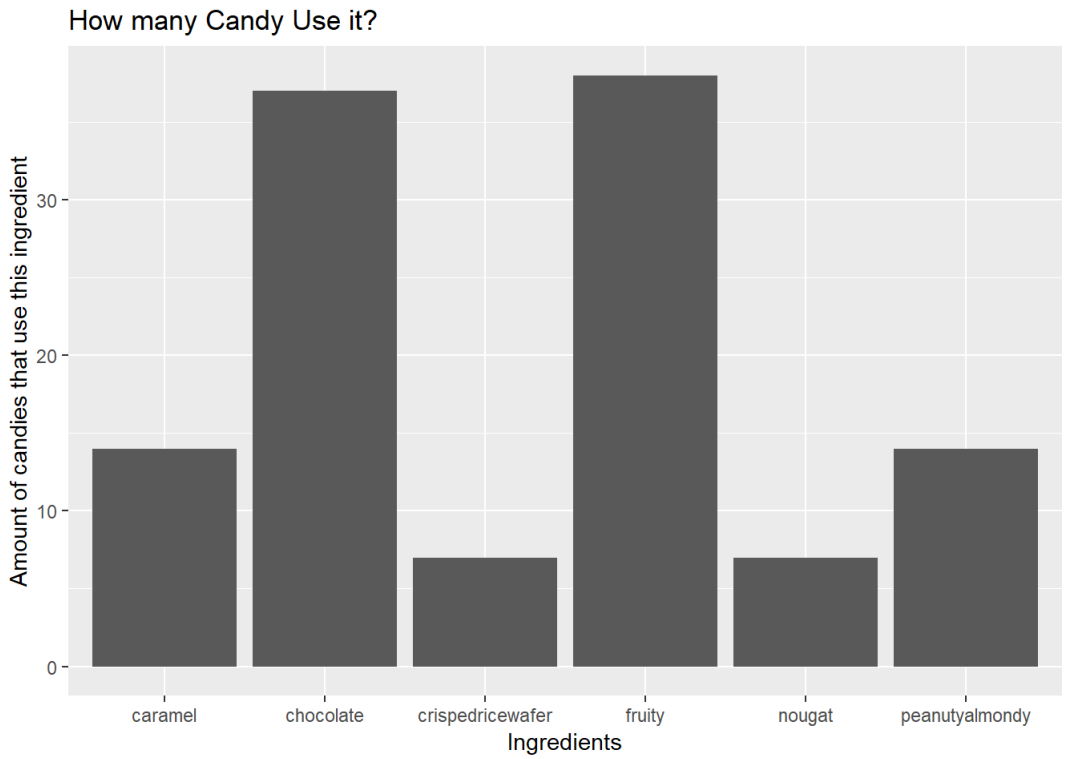
```
#select name and sugarpercent column
sugar <- candy_rankings %>%
         select(competitorname, sugarpercent)

#sort dataframe adapt from https://www.statmethods.net/management/sorting.html and select top 20
attach(sugar)

sortedsugar20 <- sugar[order(-sugarpercent),] %>%
              head(20)

sortedsugar20
```

```
## # A tibble: 20 x 2
##    competitorname              sugarpercent
##    <chr>                              <dbl>
##  1 Reese's stuffed with pieces        0.988
##  2 Milky Way Simply Caramel           0.965
##  3 Sugar Babies                       0.965
##  4 Skittles original                  0.941
##  5 Skittles wildberry                 0.941
##  6 Air Heads                          0.906
##  7 Candy Corn                         0.906
##  8 Gobstopper                         0.906
##  9 Mike & Ike                         0.872
## 10 Runts                              0.872
## 11 Whoppers                           0.872
## 12 Rolo                               0.860
## 13 Nerds                              0.848
## 14 Peanut butter M&M's                0.825
## 15 M&M's                              0.825
## 16 100 Grand                          0.732
## 17 Chewey Lemonhead Fruit Mix         0.732
## 18 Dots                               0.732
## 19 Dum Dums                           0.732
## 20 Fun Dip                            0.732
```

4 Visualize data

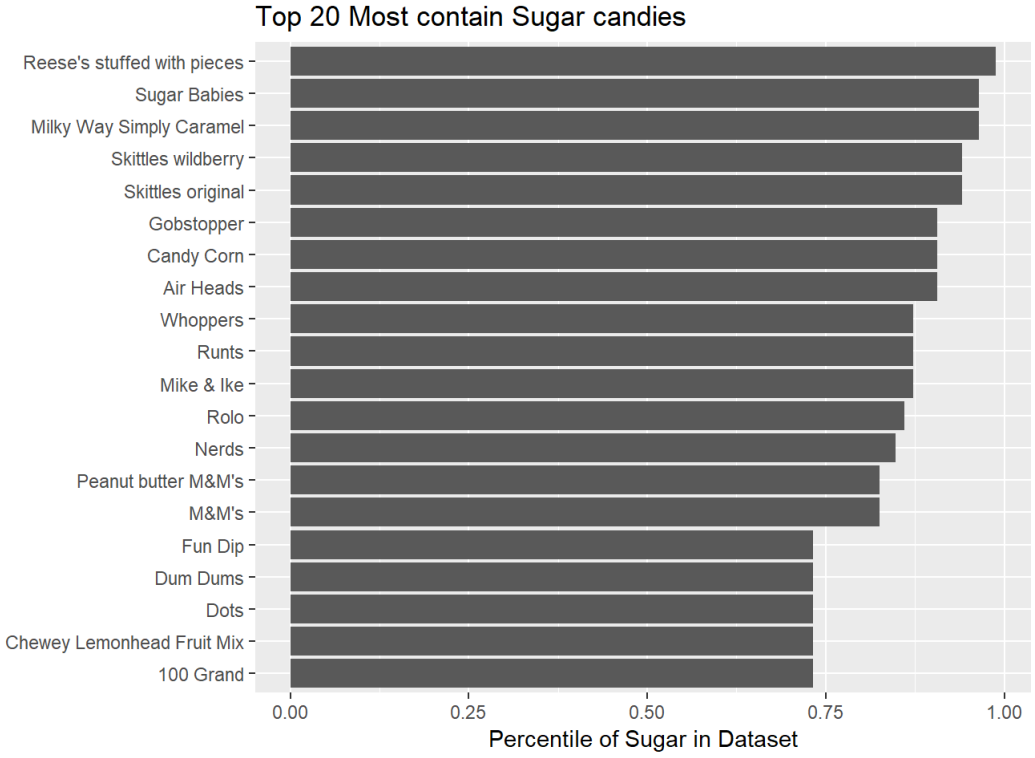
- **กราฟที่แสดงว่าวัตถุดิบแต่ละชนิดถูกใช้ในขนมกี่ชนิดบ้าง**

```
p1 <- ggplot(sortedsumdf, aes(ingredient_name, as.numeric(as.character(amount)))) +
        # add geometry of the plot (histogram)
        geom_col() +
        # add labels
        labs(x="Ingredients", y="Amount of candies that use this ingredient", title="How many Candy Use it?")

p1
```



- กราฟ 20 อันดับขนมที่มีส่วนประกอบของน้ำตาลมากที่สุด

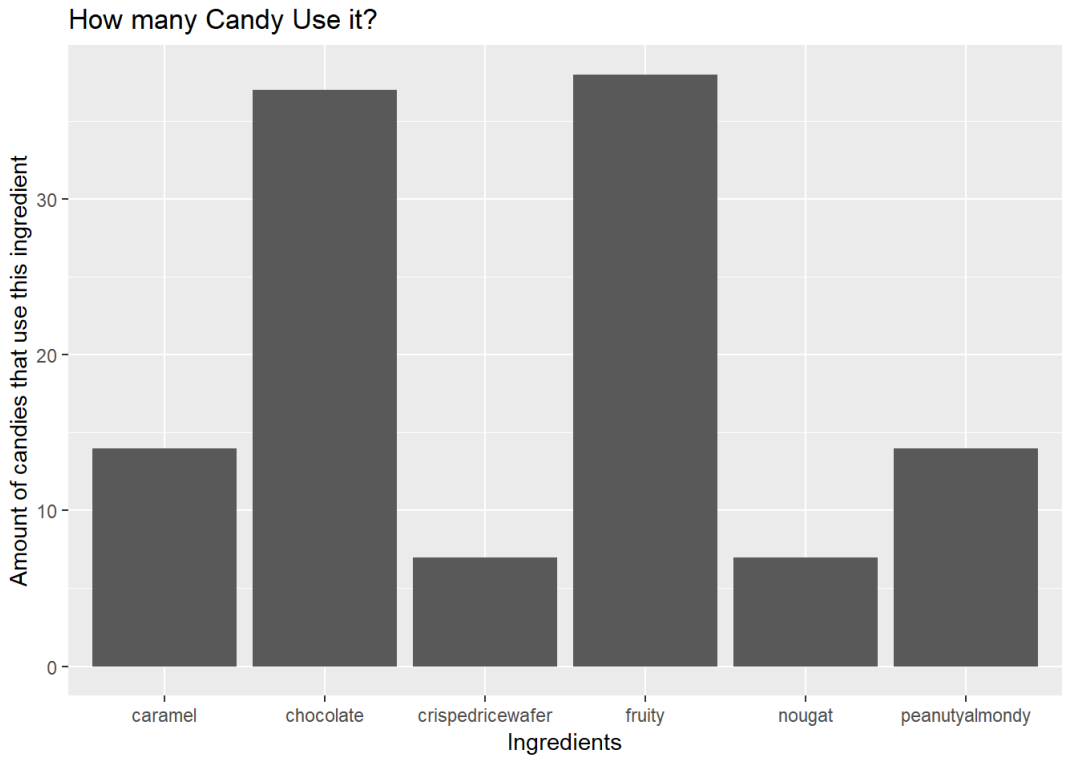
```
p2 <- ggplot(data=sortedsugar20, aes(x=reorder(competitorname, sugarpercent),y=sugarpercent)) +
        # add geometry of the plot (histogram)
        geom_col() +
        # flip coordinate
        coord_flip() +
        # add labels
        labs(x="", y="Percentile of Sugar in Dataset", title="Top 20 Most contain Sugar candies")

p2
```



สรุปผล

1. จากขนมทุกชนิดในผลสำรวจ ช็อคโกแลตถูกใช้งานเป็นอันดับ 2 รองมาจากกลิ่นผลไม้ และอันดับที่ 3 คือคาราเมล

```
p1
```



2. ขนมที่มีอัตราส่วนของน้ำตาลเป็นส่วนประกอบมากที่สุดคือ Reese’s stuffed with pieces ซึ่งเป็นขนมช็อคโกแลตที่ประกอบด้วยเนยถั่วอยู่ในPercentile ที่ 98.8 ของกลุ่มข้อมูลนี้ รองลงมาคือ Milky Way Simply Caramel ซึ่งเป็นขนมคาราเมลแบบแท่ง และ Sugar Babies ซึ่งเป็นขนมคาราเมลอัดเม็ด อยู่ที่ Percentile ที่ 96.5

```
p2
```