Halloween Mini Project

Angela Abraham

2025-10-30

Importing candy data

```
candy_file<-"https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-ranking/candy-dat
candy<-read.csv(candy_file, row.names=1)
head(candy)
```

```
##              chocolate fruity caramel peanutyalmondy nougat crispedricewafer
## 100 Grand            1      0       1              0      0                1
## 3 Musketeers         1      0       0              0      1                0
## One dime             0      0       0              0      0                0
## One quarter          0      0       0              0      0                0
## Air Heads            0      1       0              0      0                0
## Almond Joy           1      0       0              1      0                0
##              hard bar pluribus sugarpercent pricepercent winpercent
## 100 Grand       0   1        0        0.732        0.860   66.97173
## 3 Musketeers    0   1        0        0.604        0.511   67.60294
## One dime        0   0        0        0.011        0.116   32.26109
## One quarter     0   0        0        0.011        0.511   46.11650
## Air Heads       0   0        0        0.906        0.511   52.34146
## Almond Joy      0   1        0        0.465        0.767   50.34755
```

> Q1. How many different candy types are in this dataset?

```
nrow(candy)
```

```
## [1] 85
```

#85 candy types

> Q2. How many fruity candy types are in the dataset?

```
sum(candy$fruity)
```

```
## [1] 38
```

#38 candy types in the dataset

What is your favorite candy?

```
candy["Twix",]$winpercent
```

```
## [1] 81.64291
```

Q3. What is your favorite candy in the dataset and what is it’s winpercent value?

```
rownames(candy)
```

```
##  [1] "100 Grand"                   "3 Musketeers"
##  [3] "One dime"                    "One quarter"
##  [5] "Air Heads"                   "Almond Joy"
##  [7] "Baby Ruth"                   "Boston Baked Beans"
##  [9] "Candy Corn"                  "Caramel Apple Pops"
## [11] "Charleston Chew"             "Chewey Lemonhead Fruit Mix"
## [13] "Chiclets"                    "Dots"
## [15] "Dum Dums"                    "Fruit Chews"
## [17] "Fun Dip"                     "Gobstopper"
## [19] "Haribo Gold Bears"           "Haribo Happy Cola"
## [21] "Haribo Sour Bears"           "Haribo Twin Snakes"
## [23] "Hershey’s Kisses"            "Hershey’s Krackel"
## [25] "Hershey’s Milk Chocolate"    "Hershey’s Special Dark"
## [27] "Jawbusters"                  "Junior Mints"
## [29] "Kit Kat"                     "Laffy Taffy"
## [31] "Lemonhead"                   "Lifesavers big ring gummies"
## [33] "Peanut butter M&M’s"         "M&M’s"
## [35] "Mike & Ike"                  "Milk Duds"
## [37] "Milky Way"                   "Milky Way Midnight"
## [39] "Milky Way Simply Caramel"    "Mounds"
## [41] "Mr Good Bar"                 "Nerds"
## [43] "Nestle Butterfinger"         "Nestle Crunch"
## [45] "Nik L Nip"                   "Now & Later"
## [47] "Payday"                      "Peanut M&Ms"
## [49] "Pixie Sticks"                "Pop Rocks"
## [51] "Red vines"                   "Reese’s Miniatures"
## [53] "Reese’s Peanut Butter cup"   "Reese’s pieces"
## [55] "Reese’s stuffed with pieces" "Ring pop"
## [57] "Rolo"                        "Root Beer Barrels"
## [59] "Runts"                       "Sixlets"
## [61] "Skittles original"           "Skittles wildberry"
## [63] "Nestle Smarties"             "Smarties candy"
## [65] "Snickers"                    "Snickers Crisper"
## [67] "Sour Patch Kids"             "Sour Patch Tricksters"
## [69] "Starburst"                   "Strawberry bon bons"
## [71] "Sugar Babies"                "Sugar Daddy"
## [73] "Super Bubble"                "Swedish Fish"
## [75] "Tootsie Pop"                 "Tootsie Roll Juniors"
## [77] "Tootsie Roll Midgies"        "Tootsie Roll Snack Bars"
## [79] "Trolli Sour Bites"           "Twix"
## [81] "Twizzlers"                   "Warheads"
## [83] "Welch’s Fruit Snacks"        "Werther’s Original Caramel"
## [85] "Whoppers"
```

```
candy["Almond Joy", "winpercent"]
```

```
## [1] 50.34755
```

#Almond Joy; winpercent value: 50.35

> Q4. What is the winpercent value for "Kit Kat"?

```
candy["Kit Kat", "winpercent"]
```

```
## [1] 76.7686
```

#winpercent value: 76.77

> Q5. What is the winpercent value for "Tootsie Roll Snack Bars"?

```
candy["Tootsie Roll Snack Bars", "winpercent"]
```

```
## [1] 49.6535
```

#Winpercent value: 49.65

```
#install.packages("skimr)
library("skimr")
```

```
## Warning: package 'skimr' was built under R version 4.5.2
```

```
skim_without_charts(candy)
```

Table 1: Data summary

Name	candy
Number of rows	85
Number of columns	12
Column type frequency:	
numeric	12
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
chocolate	0	1	0.44	0.50	0.00	0.00	0.00	1.00	1.00
fruity	0	1	0.45	0.50	0.00	0.00	0.00	1.00	1.00
caramel	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
peanutyalmondy	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00
nougat	0	1	0.08	0.28	0.00	0.00	0.00	0.00	1.00
crispedricewafer	0	1	0.08	0.28	0.00	0.00	0.00	0.00	1.00
hard	0	1	0.18	0.38	0.00	0.00	0.00	0.00	1.00
bar	0	1	0.25	0.43	0.00	0.00	0.00	0.00	1.00
pluribus	0	1	0.52	0.50	0.00	0.00	1.00	1.00	1.00
sugarpercent	0	1	0.48	0.28	0.01	0.22	0.47	0.73	0.99
pricepercent	0	1	0.47	0.29	0.01	0.26	0.47	0.65	0.98
winpercent	0	1	50.32	14.71	22.45	39.14	47.83	59.86	84.18

> Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

#The last column, winpercent, does not have a number in between 0 and 1.

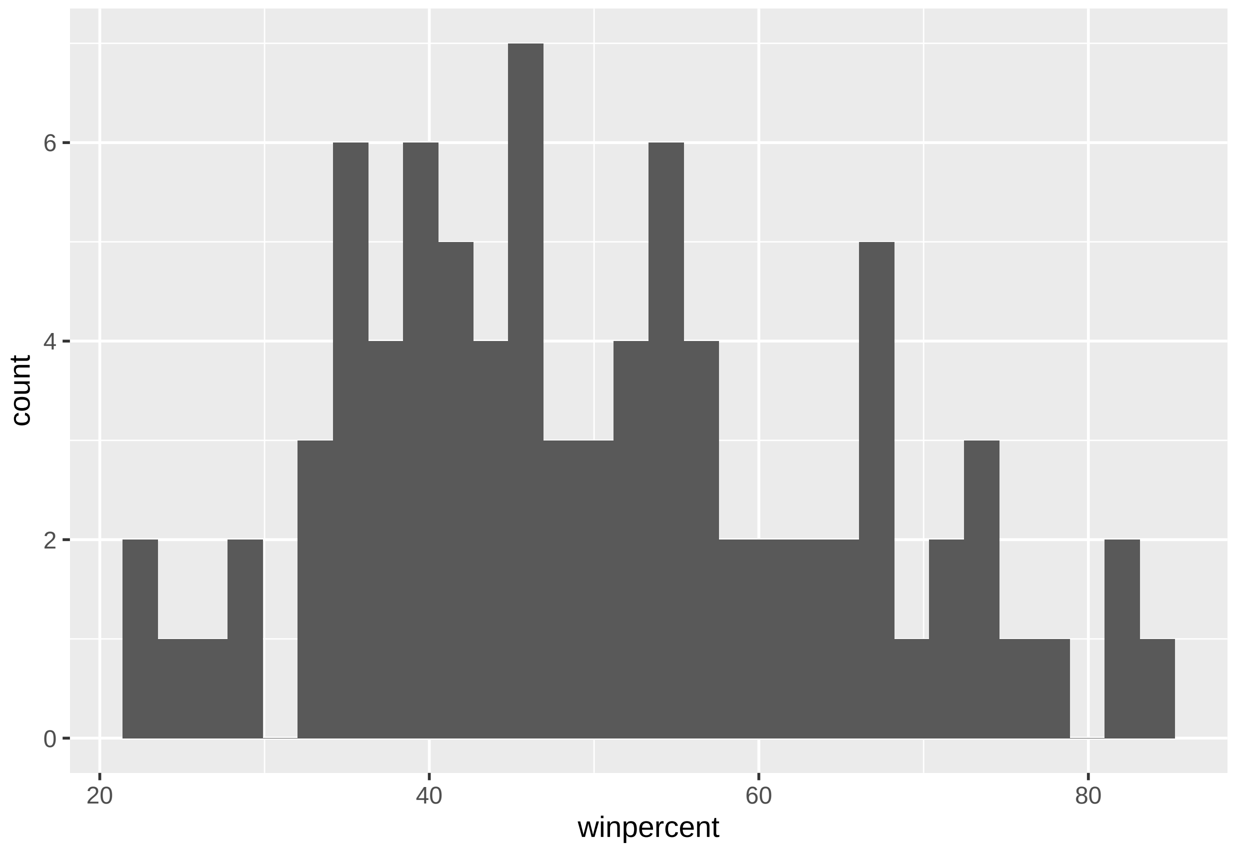
> Q7. What do you think a zero and one represent for the candy$chocolate column?

#Zero is the candy does not contain chocolate and one is that it does contain chocolate.

> Q8. Plot a histogram of winpercent values

```r
library(ggplot2)
ggplot(candy, aes(x=winpercent))+geom_histogram()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value ‘binwidth‘.
```



Q9. Is the distribution of winpercent values symmetrical?

#Distribution is not symmetrical, seems to be skewed right

Q10. Is the center of the distribution above or below 50%?

```
mean(candy$winpercent)
```

```
## [1] 50.31676
```

```
median(candy$winpercent)
```

```
## [1] 47.82975
```

#Center is below 50% of the distribution.

Q11. On average, is chocolate candy higher or lower ranked than fruit candy?

```
chocolate<-candy$winpercent[as.logical(candy$chocolate)]
fruity<-candy$winpercent[as.logical(candy$fruity)]
mean(chocolate)
```

```
## [1] 60.92153
```

```r
mean(fruity)
```

```
## [1] 44.11974
```

#Chocolate is ranked higher than fruit candy with a mean average of 60.9% which is higher than the fruity candy (mean average of 44.1%).

> Q12. Is this difference statistically significant?

```r
t.test(chocolate, fruity)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  chocolate and fruity
## t = 6.2582, df = 68.882, p-value = 2.871e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11.44563 22.15795
## sample estimates:
## mean of x mean of y 
##  60.92153  44.11974
```

#Yes, it is statistically significant. P-value is 2.871e-08 which is smaller than 0.05.

Overall Candy Rankings

> Q13. What are the five least liked candy types in this set?

```r
head(candy[order(candy$winpercent),], n=5)
```

```
##                    chocolate fruity caramel peanutyalmondy nougat
## Nik L Nip                  0      1       0              0      0
## Boston Baked Beans         0      0       0              1      0
## Chiclets                   0      1       0              0      0
## Super Bubble               0      1       0              0      0
## Jawbusters                 0      1       0              0      0
##                    crispedricewafer hard bar pluribus sugarpercent pricepercent
## Nik L Nip                         0    0   0        1        0.197        0.976
## Boston Baked Beans                0    0   0        1        0.313        0.511
## Chiclets                          0    0   0        1        0.046        0.325
## Super Bubble                      0    0   0        0        0.162        0.116
## Jawbusters                        0    1   0        1        0.093        0.511
##                    winpercent
## Nik L Nip            22.44534
## Boston Baked Beans   23.41782
## Chiclets             24.52499
## Super Bubble         27.30386
## Jawbusters           28.12744
```

#Five least liked candy types are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbreakers.

> Q14. What are the top 5 all time favorite candy types out of this set?

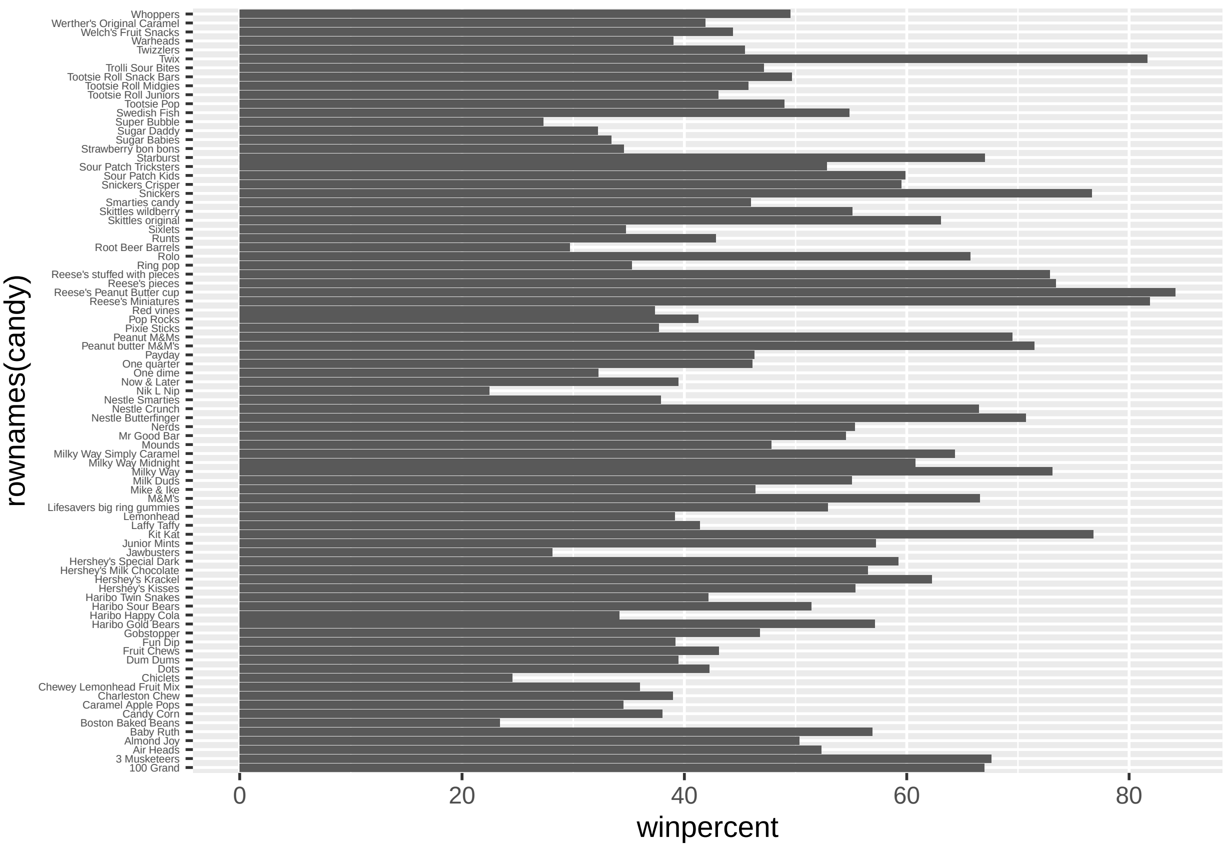
```
head(candy[order(candy$winpercent, decreasing=TRUE),], n=5)
```

```
##                           chocolate fruity caramel peanutyalmondy nougat
## Reese’s Peanut Butter cup         1      0       0              1      0
## Reese’s Miniatures                1      0       0              1      0
## Twix                              1      0       1              0      0
## Kit Kat                           1      0       0              0      0
## Snickers                          1      0       1              1      1
##                           crispedricewafer hard bar pluribus sugarpercent
## Reese’s Peanut Butter cup                0    0   0        0        0.720
## Reese’s Miniatures                       0    0   0        0        0.034
## Twix                                     1    0   1        0        0.546
## Kit Kat                                  1    0   1        0        0.313
## Snickers                                 0    0   1        0        0.546
##                           pricepercent winpercent
## Reese’s Peanut Butter cup        0.651   84.18029
## Reese’s Miniatures               0.279   81.86626
## Twix                             0.906   81.64291
## Kit Kat                          0.511   76.76860
## Snickers                         0.651   76.67378
```

#Top 5 all time favorite candy types in the set are Reese’s Peanut Butter Cups, Reese’s miniatures, Twix, Kit Kat, and Snickers.

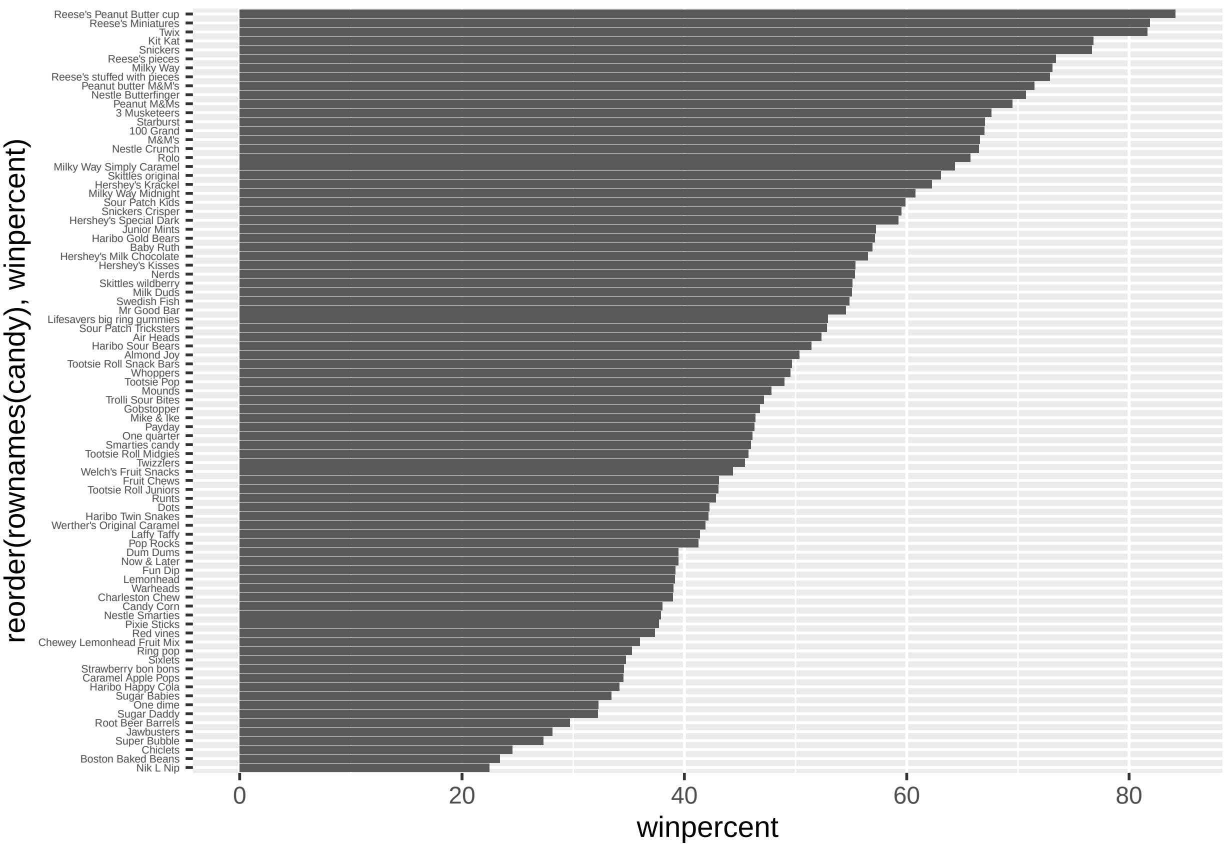
> Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)
ggplot(candy) + aes(winpercent, rownames(candy)) + geom_col() + theme(axis.text.y=element_text(size=4))
```



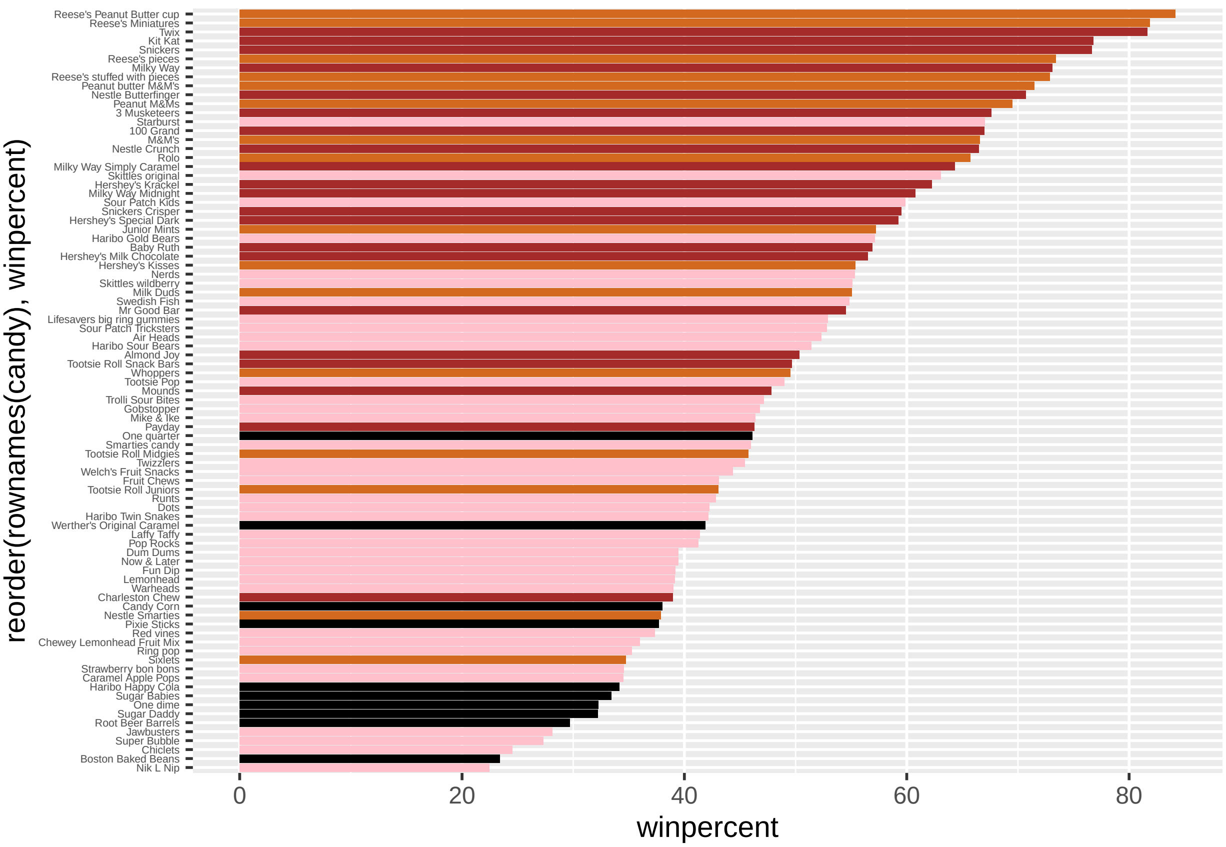
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent)) + geom_col() + theme(axis.text.y=element_text(s
```



```
my_cols=rep("black", nrow(candy))
my_cols[as.logical(candy$chocolate)] = "chocolate"
my_cols[as.logical(candy$bar)] = "brown"
my_cols[as.logical(candy$fruity)] = "pink"
```

```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols) + theme(axis.text.y=element_text(size=4))
```



Q17. What is the worst ranked chocolate candy?

#Worst ranked chocolate candy is Sixlets.

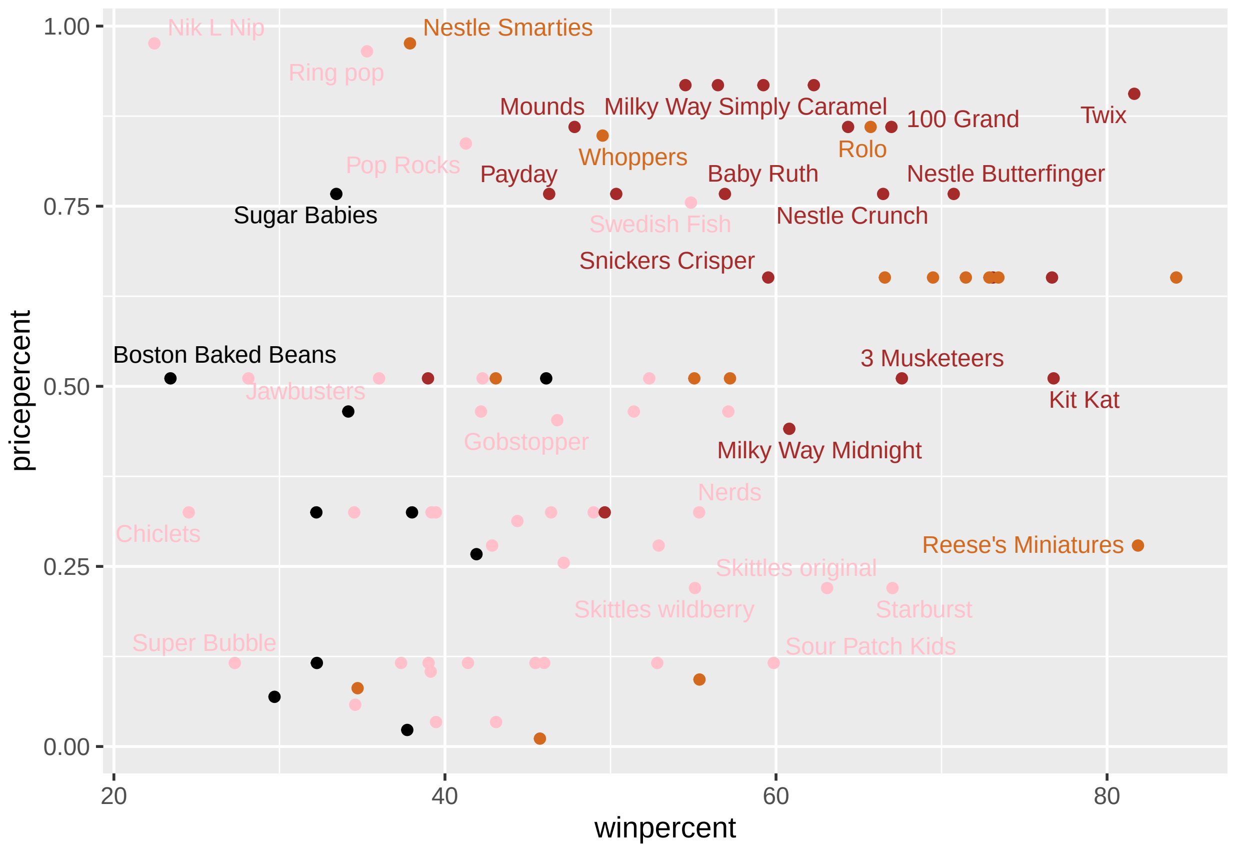
Q18. What is the best ranked fruity candy?

#Best ranked fruity candy is Starburst.

Taking a look at pricepercent

```
library(ggrepel)
ggplot(candy) + aes(winpercent, pricepercent, label=rownames(candy)) + geom_point(col=my_cols) + geom_te
```

```
## Warning: ggrepel: 54 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Q19. Which candy type is the highest ranked in terms of winpercent for the least money -i.e. offers the most bang for your buck?

```
cheap_candy<-candy[candy$pricepercent<0.3,]
cheap_candy[which.max(cheap_candy$winpercent),]
```

```
##                    chocolate fruity caramel peanutyalmondy nougat
## Reese's Miniatures         1      0       0              1      0
##                    crispedricewafer hard bar pluribus sugarpercent pricepercent
## Reese's Miniatures                0    0   0        0        0.034        0.279
##                    winpercent
## Reese's Miniatures   81.86626
```

#Reese's Miniatures

Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular?

```
expensive <-head(candy[order(candy$pricepercent, decreasing=TRUE),],n=5)
expensive
```

```
##                          chocolate fruity caramel peanutyalmondy nougat
## Nik L Nip                        0      1       0              0      0
## Nestle Smarties                  1      0       0              0      0
## Ring pop                         0      1       0              0      0
## Hershey’s Krackel                1      0       0              0      0
## Hershey’s Milk Chocolate         1      0       0              0      0
##                          crispedricewafer hard bar pluribus sugarpercent
## Nik L Nip                               0    0   0        1        0.197
## Nestle Smarties                         0    0   0        1        0.267
## Ring pop                                0    1   0        0        0.732
## Hershey’s Krackel                       1    0   1        0        0.430
## Hershey’s Milk Chocolate                0    0   1        0        0.430
##                          pricepercent winpercent
## Nik L Nip                       0.976   22.44534
## Nestle Smarties                 0.976   37.88719
## Ring pop                        0.965   35.29076
## Hershey’s Krackel               0.918   62.28448
## Hershey’s Milk Chocolate        0.918   56.49050
```

```
expensive[which.min(expensive$winpercent),]
```

```
##             chocolate fruity caramel peanutyalmondy nougat crispedricewafer hard
## Nik L Nip           0      1       0              0      0                0    0
##             bar pluribus sugarpercent pricepercent winpercent
## Nik L Nip     0        1        0.197        0.976   22.44534
```

#or

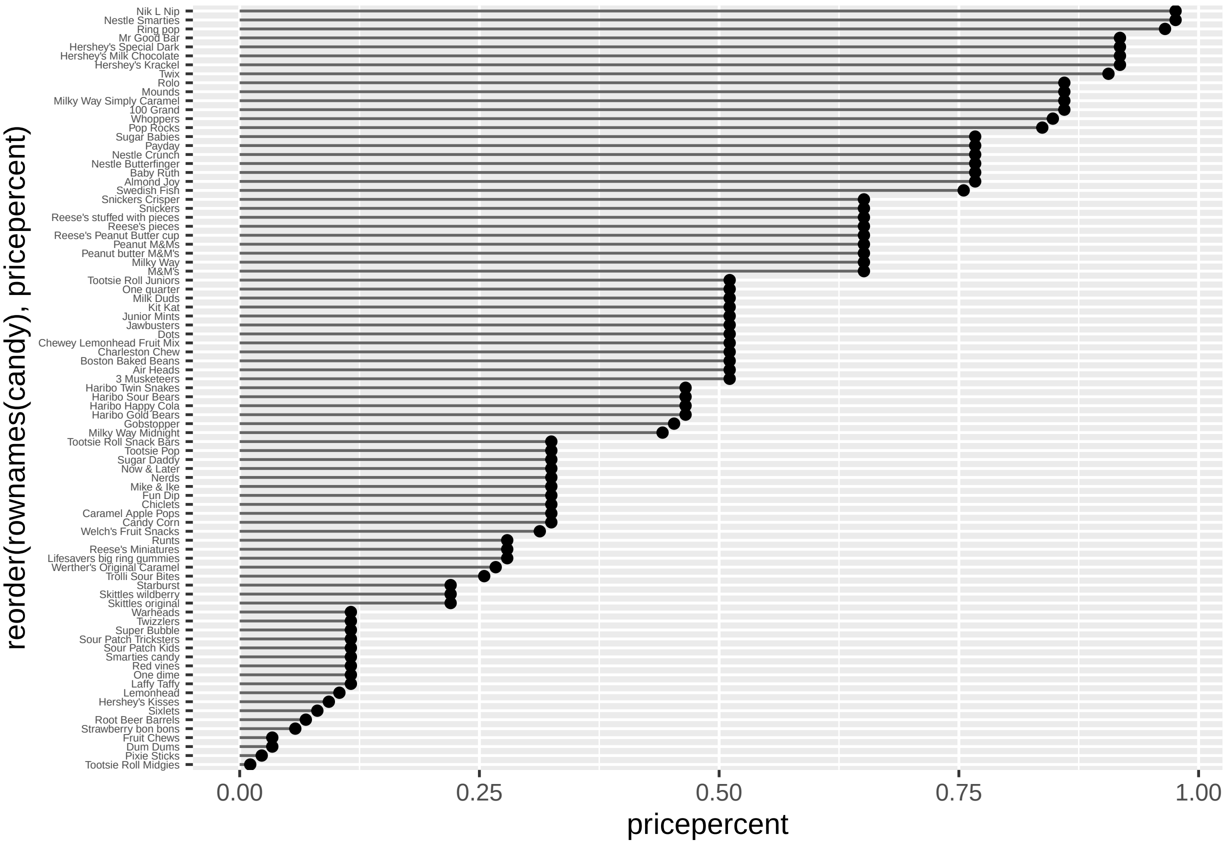
```
ord<-order(candy$pricepercent, decreasing=TRUE)
head(candy[ord, c(11,12)],n=5)
```

```
##                          pricepercent winpercent
## Nik L Nip                       0.976   22.44534
## Nestle Smarties                 0.976   37.88719
## Ring pop                        0.965   35.29076
## Hershey’s Krackel               0.918   62.28448
## Hershey’s Milk Chocolate        0.918   56.49050
```

#The top five most expensive candy types in the dataset are Nik L Nip, Nestle Smarties, Ring pop, Hersheys Krackel, and Hersheys Milk Chocolate. The least popular out of them all is Nik L Nip.

> Q21. Make a barplot again with geom_col() this time using pricepercent and then improve this step by step, first ordering the x-axis by value and finally making a so called “dot chat” or “lollipop” chart by swapping geom_col() for geom_point() + geom_segment().

```
ggplot(candy) + aes(pricepercent, reorder(rownames(candy), pricepercent))+ geom_segment(aes(yend=reorde
```

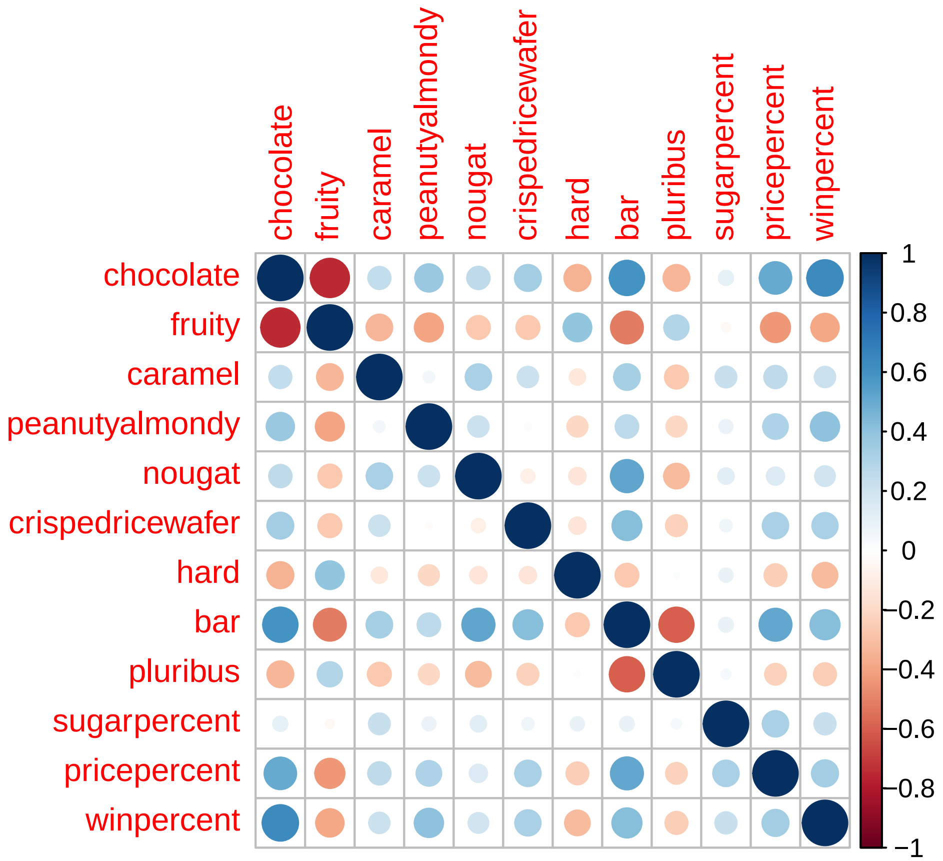


Exploring the correlation structure

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
cij<-cor(candy)
corrplot(cij) + theme(axis.text.y=element_text(size=6))
```



```
## NULL
```

Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)?

#Chocolate and fruity are anti-correlated.

Q23. Similarly, what two variables are more positively correlated?

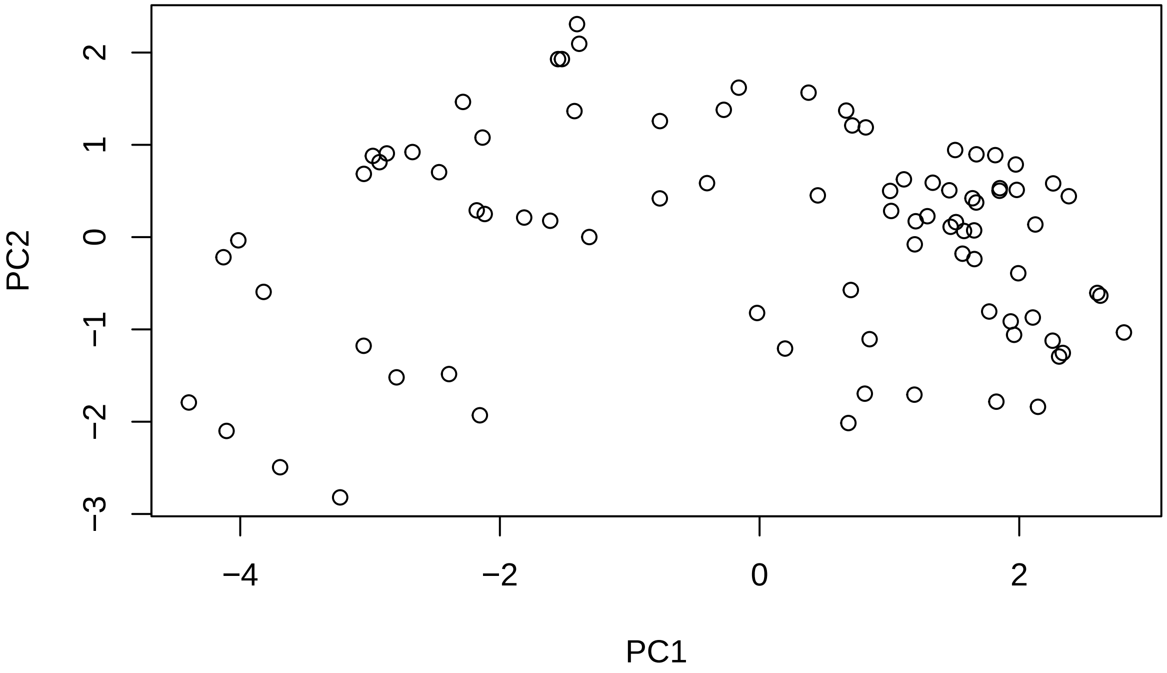
#Chocolate and winpercent , chocolate and bar are most positively correlated.

Principal Component Analysis

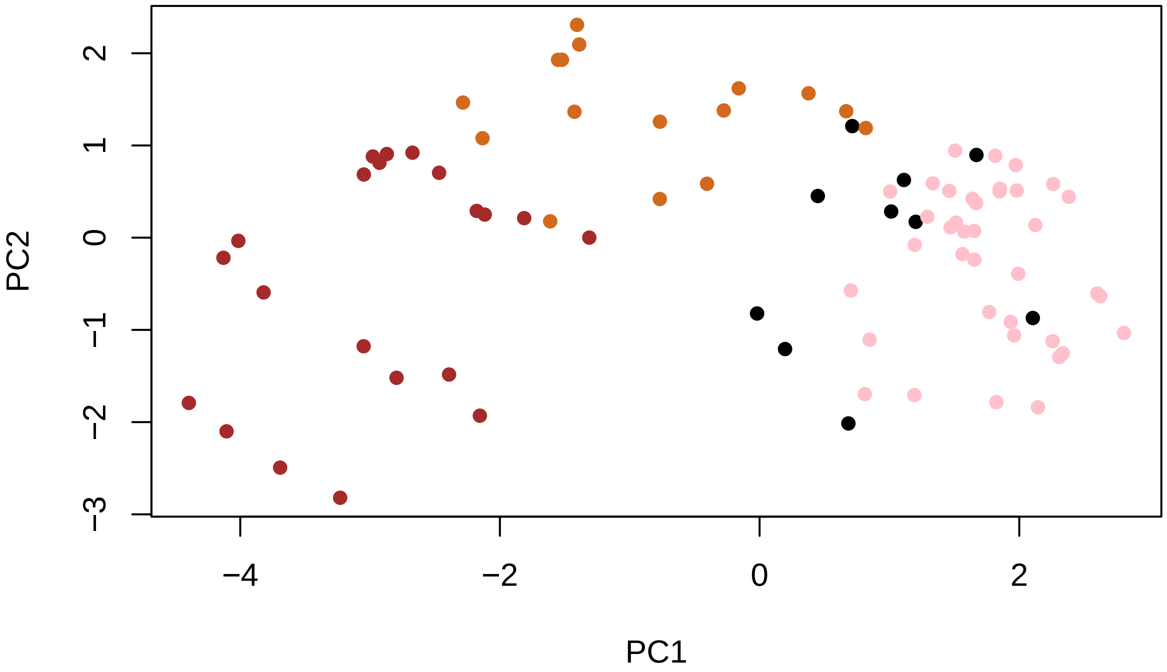
```
pca<-prcomp(candy, scale=TRUE)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4    PC5     PC6     PC7
## Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
## Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
## Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 0.7424 0.79830 0.85369
##                            PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.74530 0.67824 0.62349 0.43974 0.39760
## Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
## Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

```
plot(pca$x[,1:2])
```

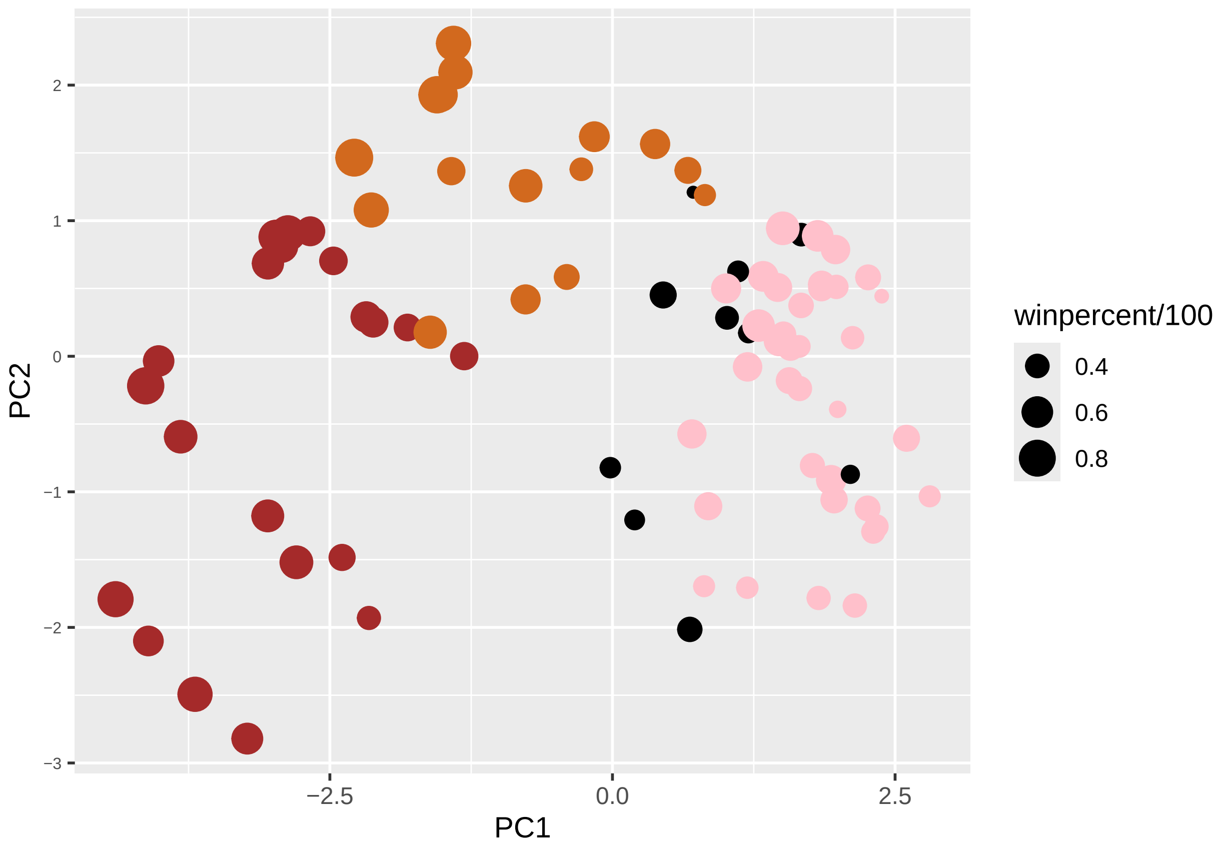


```
plot(pca$x[,1:2],col=my_cols,pch=16)
```



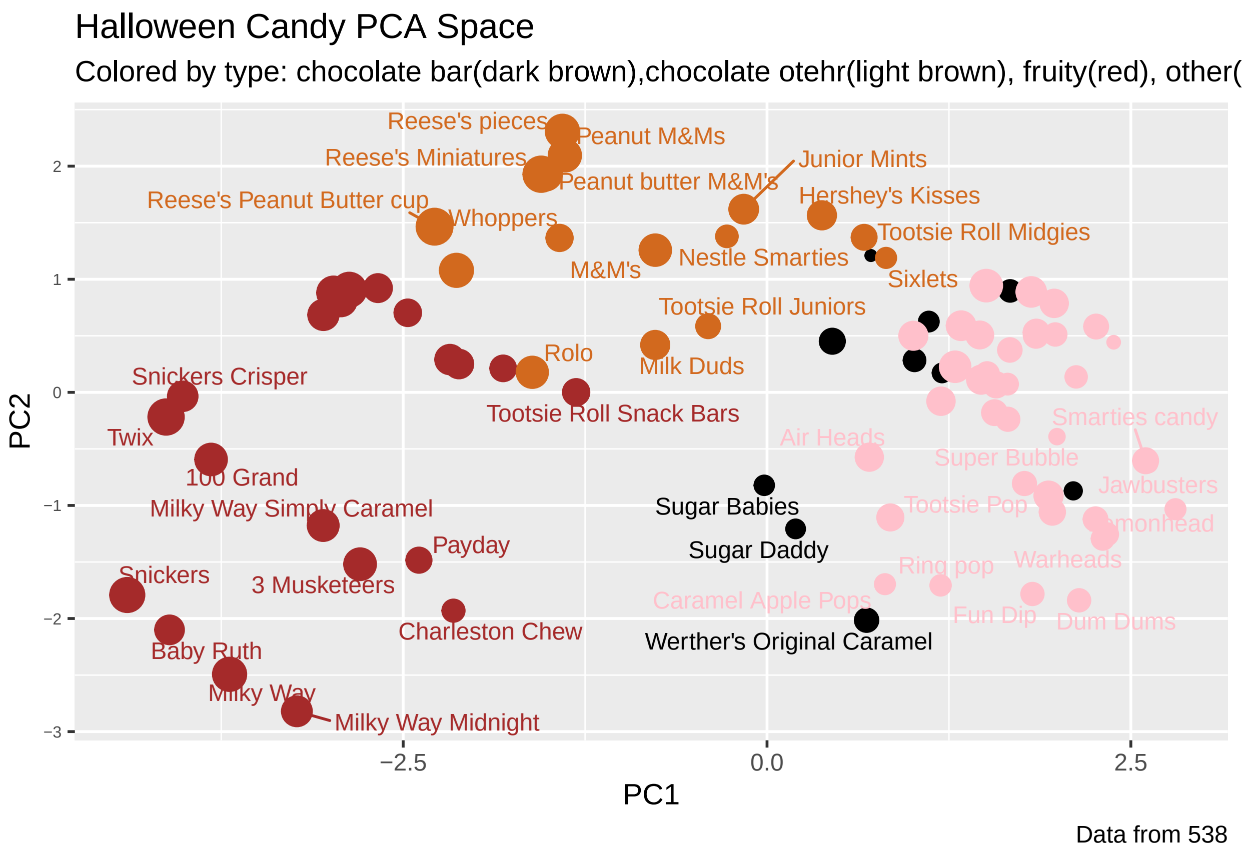
```
my_data<-cbind(candy,pca$x[,1:3])
```

```
p<-ggplot(my_data)+
  aes(x=PC1,y=PC2, size=winpercent/100,text=rownames(my_data), label=rownames(my_data)) + geom_point(col
p
```



```
library(ggrepel)

p+geom_text_repel(size=3.3, col=my_cols, max.overlaps=7) + theme(legend.position="none") + labs(title="
```

```
## Warning: ggrepel: 44 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Halloween Candy PCA Space

Colored by type: chocolate bar(dark brown),chocolate otehr(light brown), fruity(red), other(

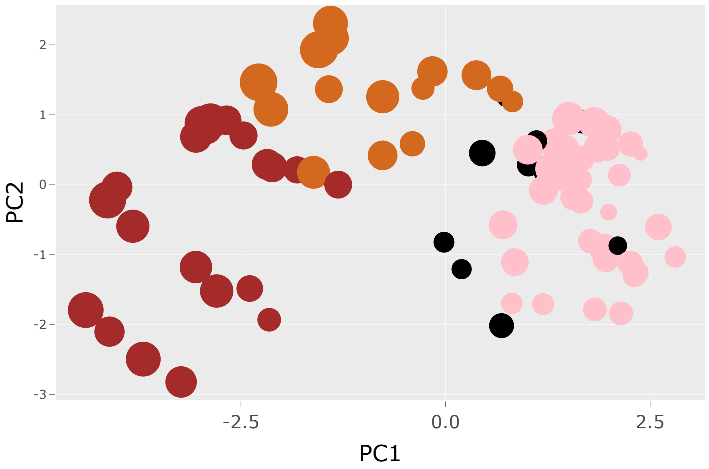
Data from 538

```
#install.packages("plotly)
library(plotly)
```

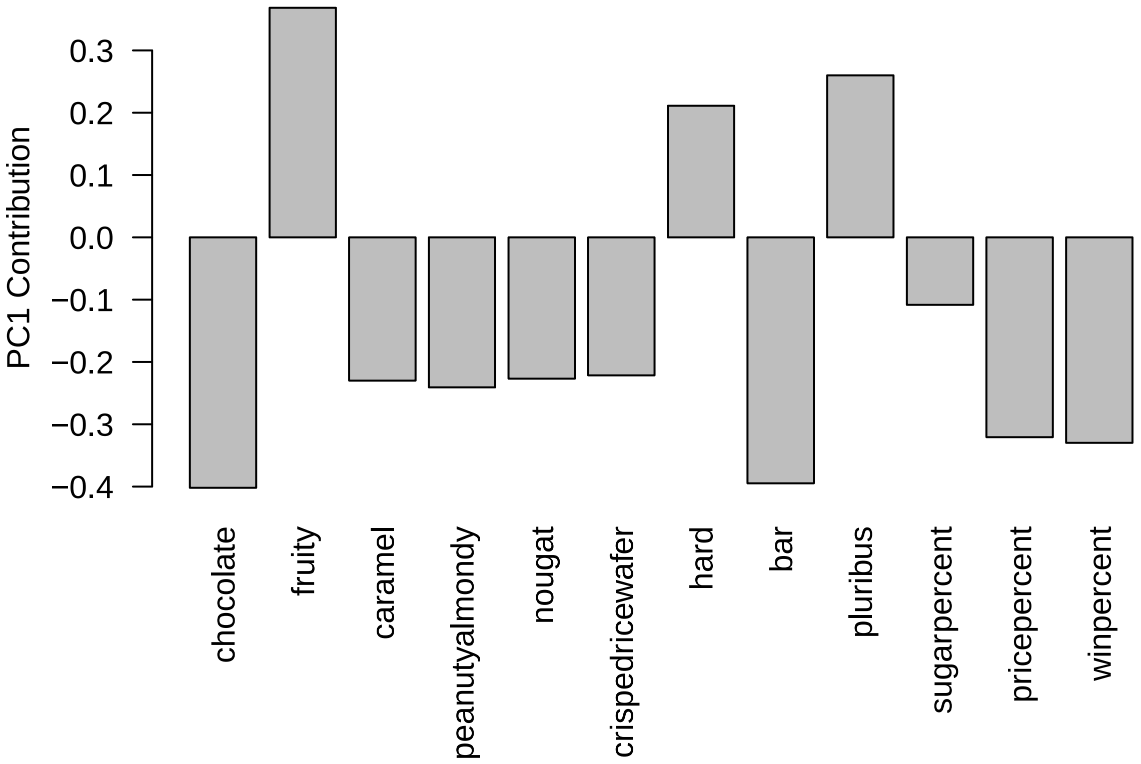
```
## Warning: package 'plotly' was built under R version 4.5.2

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
ggplotly(p)
```

```
## file:///C:/Users/Angela/AppData/Local/Temp/Rtmp82dYlm/file4ffc21bf1d63/widget4ffc29bd460e.html scree
```



```r
par(mar=c(8,4,2,2))
barplot(pca$rotation[,1], las=2,ylab="PC1 Contribution")
```



Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

#Fruity, hard, and pluribus are picked up strongly by PC1 in the positive direction. This makes sense as this captures the spectrum between fruity and chocolate. The high PC1 is the fruity, hard candies the come in bulk, while the low PC1 is chocolate bars that are one item and is more expensive.