Halloween_Mini_Project

Jocelyn Olvera

10/29/2021

###Class 10: Halloween Mini-Project

##Importing candy data

```
candy_file <- "candy-data.csv"

candy = read.csv("candy-data.csv", 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? ##Q2. How many fruity candy types are in the dataset? ##The functions dim(), nrow(), table() and sum() may be useful for answering the first 2 questions.

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

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

```
library("skimr")
skim(candy)
```

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	hist
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	
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	

Q1:

```
nrow(candy)
```

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

Q2:

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

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

```
sum(candy$chocolate)
```

```
## [1] 37
```

```
View(candy)
```

```
candy["Baby Ruth", ]$winpercent
```

```
## [1] 56.91455
```

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

```
candy["Baby Ruth",]$winpercent
```

```
## [1] 56.91455
```

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

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

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

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

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

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

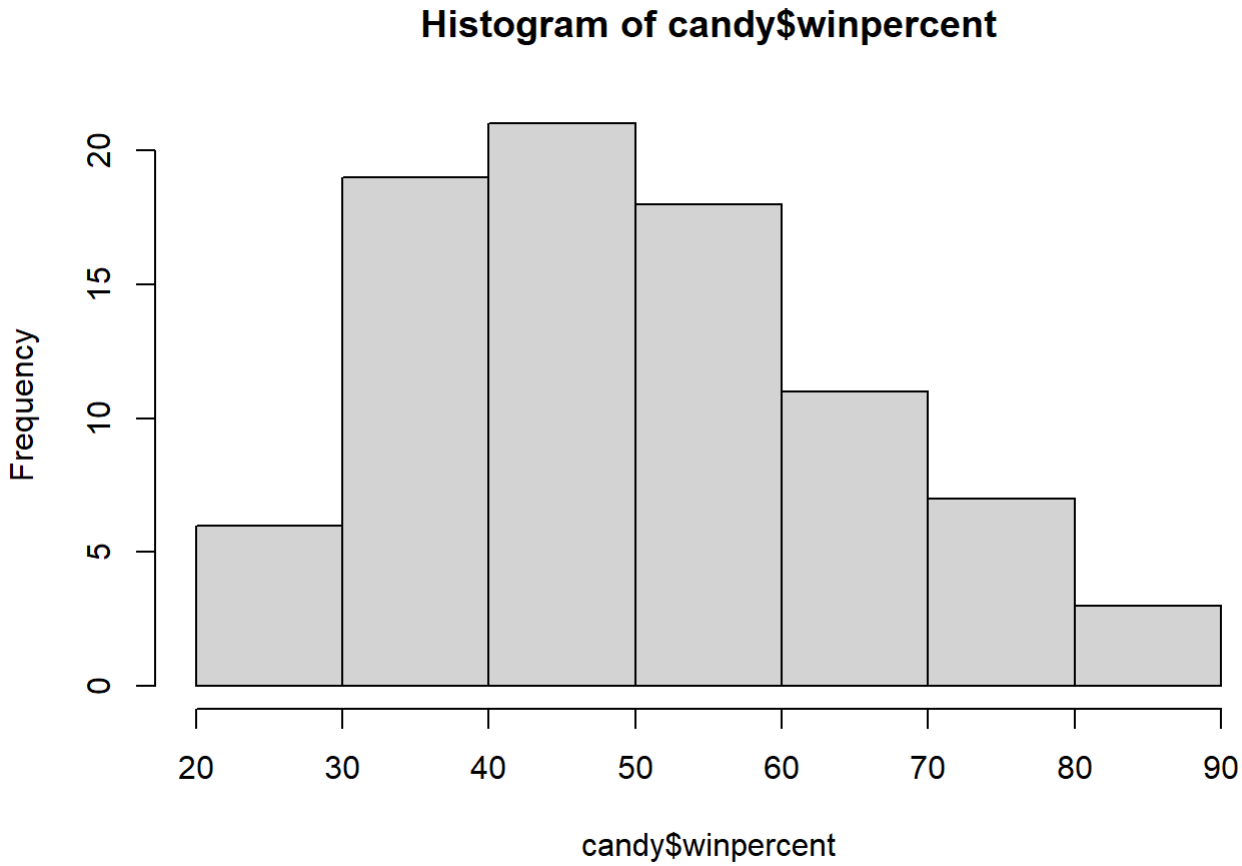
Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

yes,

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

##Q8. Plot a histogram of winpercent values

```
hist(candy$winpercent)
```



##Q9. Is the distribution of winpercent values symmetrical? ##No

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

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

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

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

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

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

##Q12. Is this difference statistically significant?

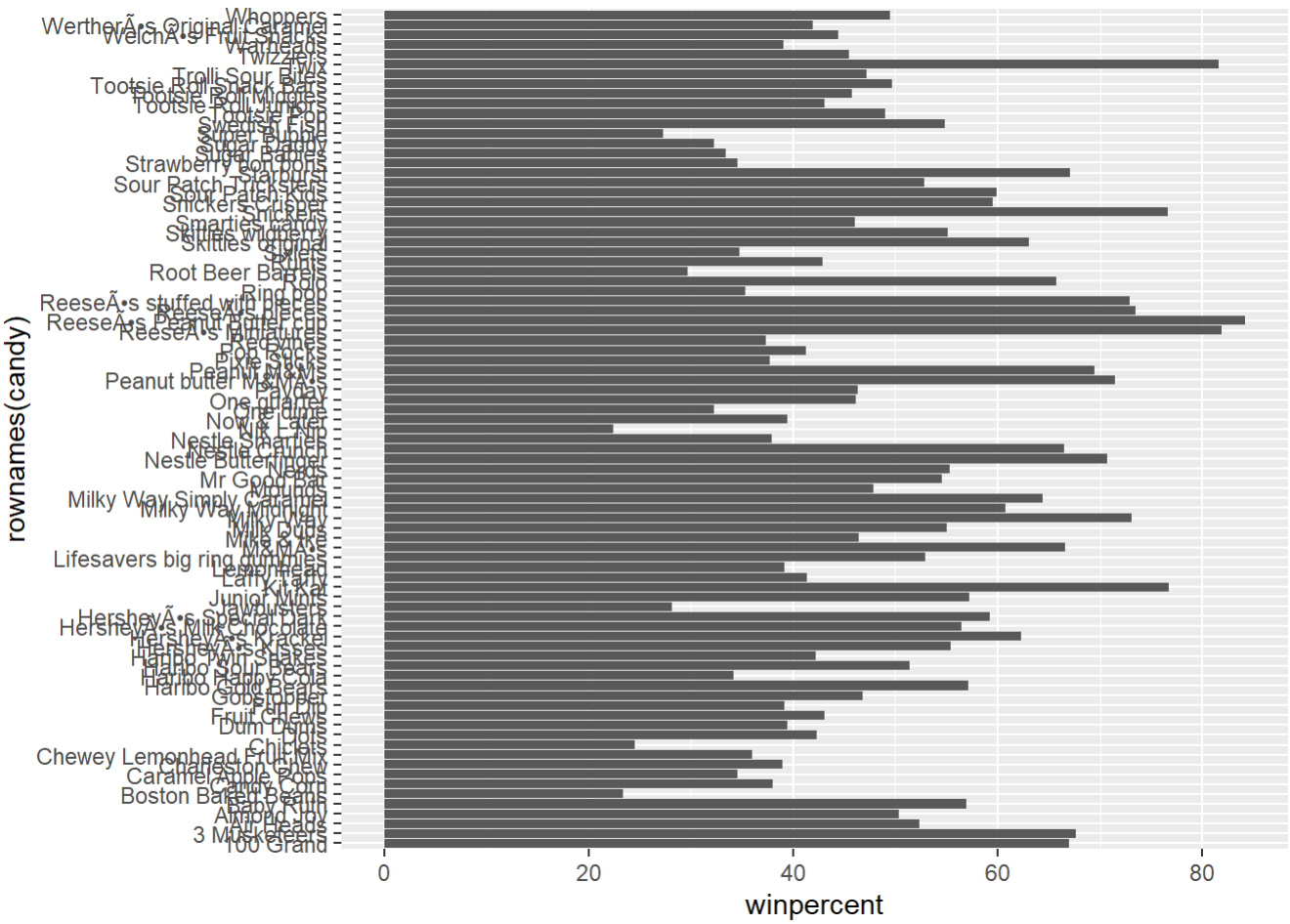
```
t.test(chocolate, fruity)
```

```
##
##  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
```

#3. Overall Candy Rankings ##Q13. What are the five least liked candy types in this set? ##Q14. What are the top 5 all time favorite candy types out of this set?

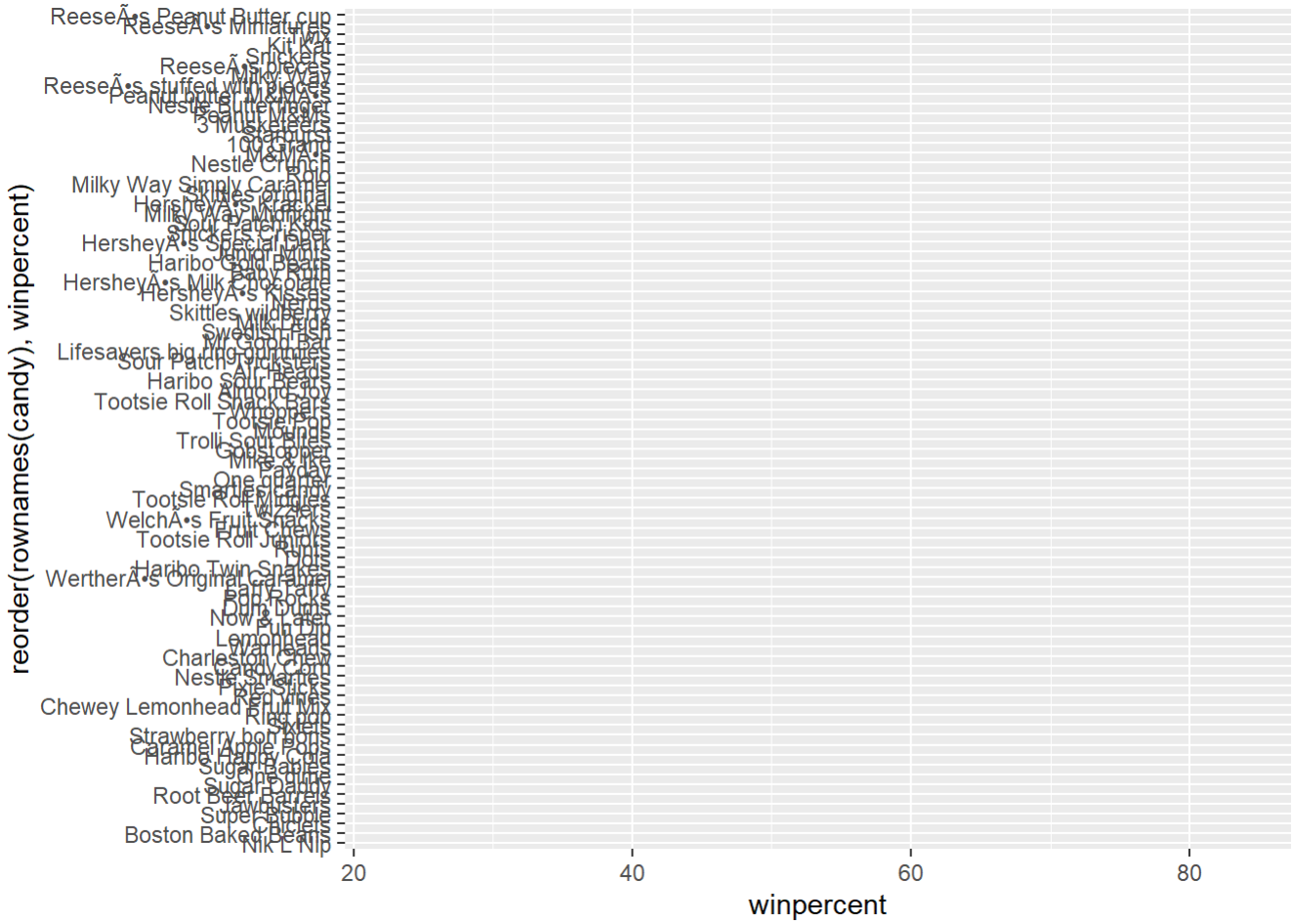
##Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)
ggplot(candy) +
  aes(winpercent, rownames(candy))+
  geom_col()
```



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

```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent))
```

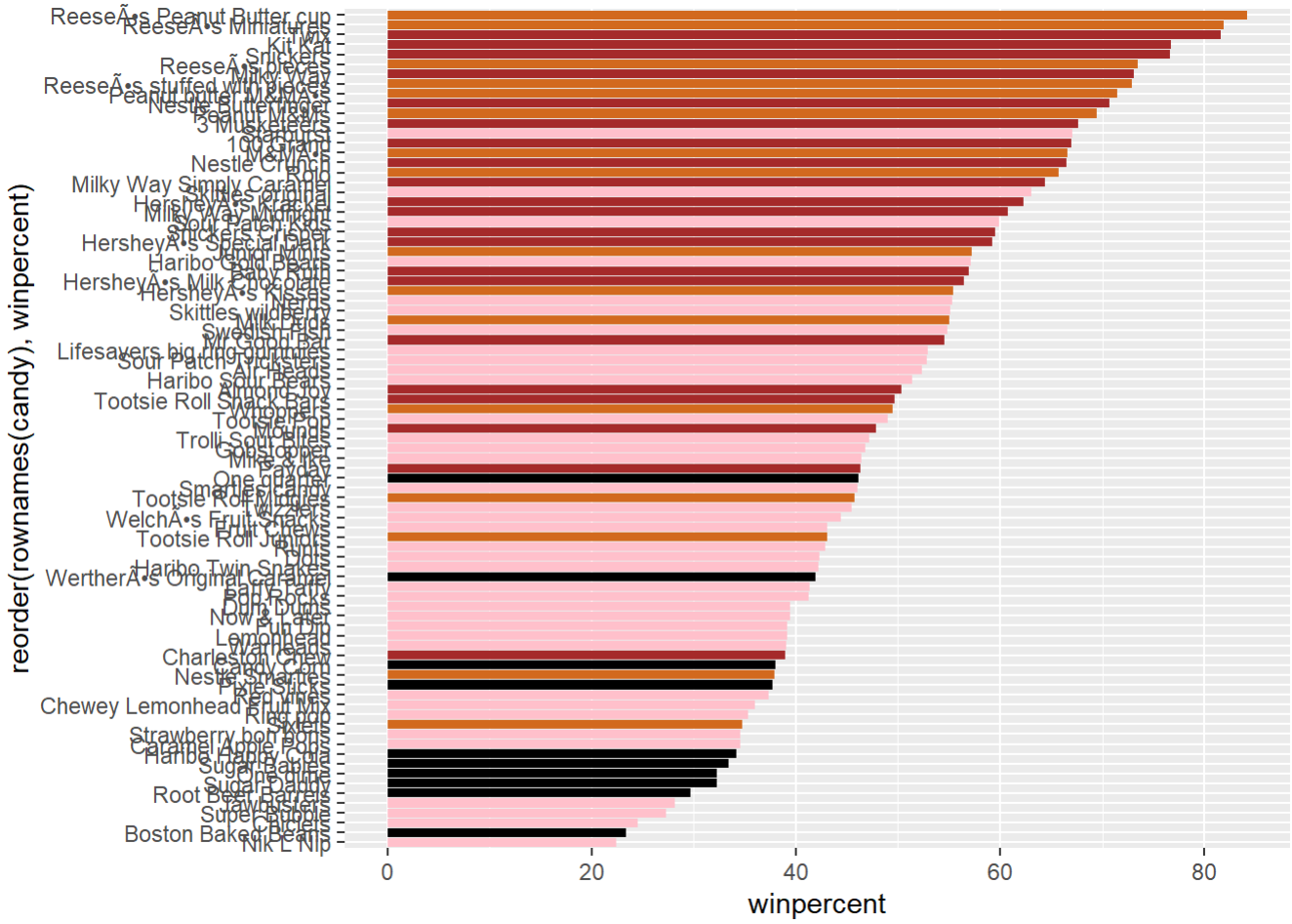



```
  geom_col()
```

```
## geom_col: width = NULL, na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_stack
```

```
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)
```



Finding out what these functions stand for. “rep” repeats. We make a vector here where it repeats the color “black”

```
my_cols=rep("black", nrow(candy))
my_cols
```

```
##  [1] "black" "black" "black" "black" "black" "black" "black" "black" "black"
## [10] "black" "black" "black" "black" "black" "black" "black" "black" "black"
## [19] "black" "black" "black" "black" "black" "black" "black" "black" "black"
## [28] "black" "black" "black" "black" "black" "black" "black" "black" "black"
## [37] "black" "black" "black" "black" "black" "black" "black" "black" "black"
## [46] "black" "black" "black" "black" "black" "black" "black" "black" "black"
## [55] "black" "black" "black" "black" "black" "black" "black" "black" "black"
## [64] "black" "black" "black" "black" "black" "black" "black" "black" "black"
## [73] "black" "black" "black" "black" "black" "black" "black" "black" "black"
## [82] "black" "black" "black" "black"
```

```
my_cols[as.logical(candy$chocolate)]="chocolate"
my_cols
```

```
##  [1] "chocolate" "chocolate" "black"     "black"     "black"     "chocolate"
##  [7] "chocolate" "black"     "black"     "black"     "chocolate" "black"
## [13] "black"     "black"     "black"     "black"     "black"     "black"
## [19] "black"     "black"     "black"     "black"     "chocolate" "chocolate"
## [25] "chocolate" "chocolate" "black"     "chocolate" "chocolate" "black"
## [31] "black"     "black"     "chocolate" "chocolate" "black"     "chocolate"
## [37] "chocolate" "chocolate" "chocolate" "chocolate" "chocolate" "black"
## [43] "chocolate" "chocolate" "black"     "black"     "black"     "chocolate"
## [49] "black"     "black"     "black"     "chocolate" "chocolate" "chocolate"
## [55] "chocolate" "black"     "chocolate" "black"     "black"     "chocolate"
## [61] "black"     "black"     "chocolate" "black"     "chocolate" "chocolate"
## [67] "black"     "black"     "black"     "black"     "black"     "black"
## [73] "black"     "black"     "chocolate" "chocolate" "chocolate" "chocolate"
## [79] "black"     "chocolate" "black"     "black"     "black"     "black"
## [85] "chocolate"
```

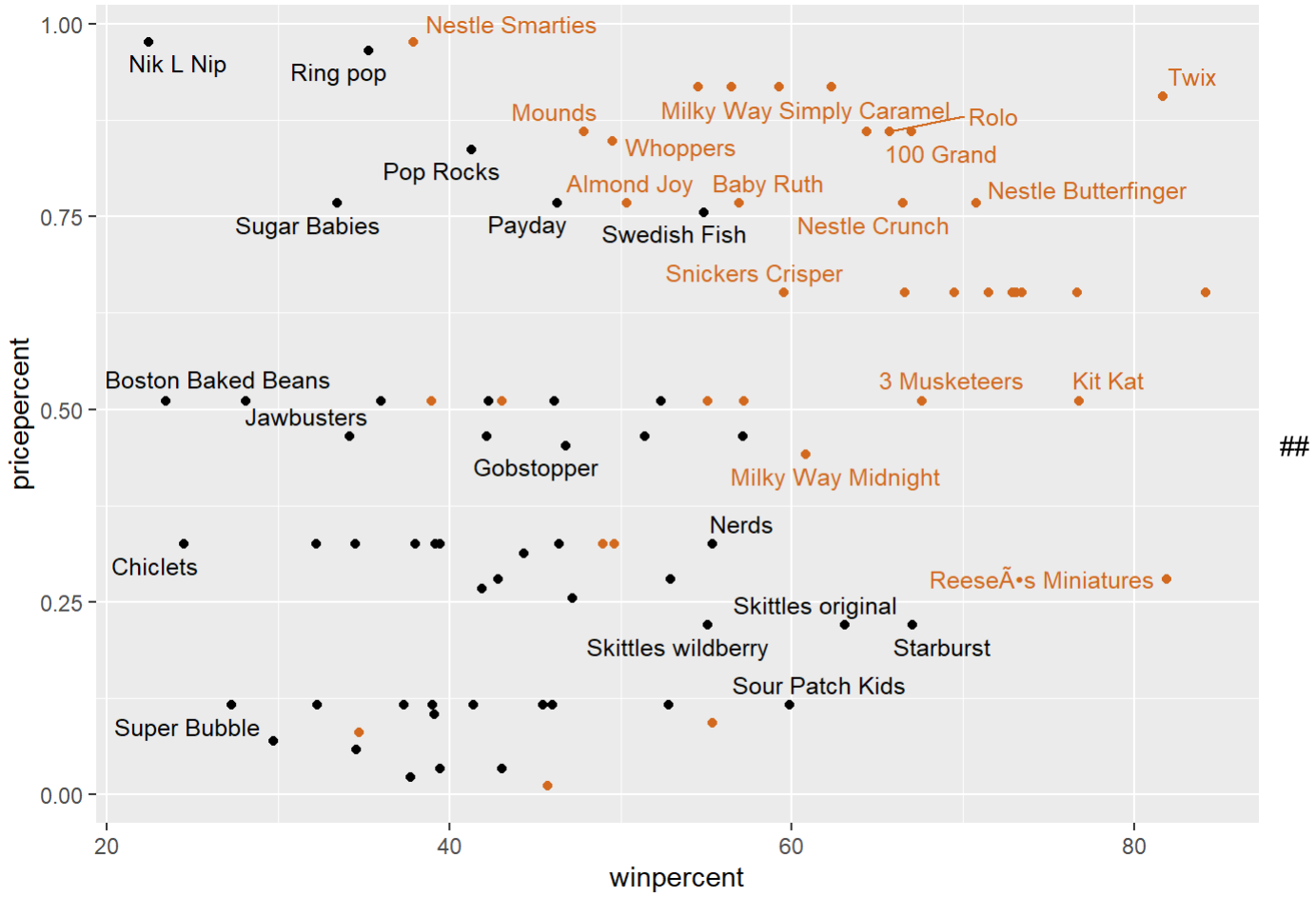
Q17. What is the worst ranked chocolate candy?

Q18. What is the best ranked fruity candy?

#4. Taking a look at pricepercent

```
library(ggrepel)

# How about a plot of price vs win
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 5)
```

```
## Warning: ggrepel: 53 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



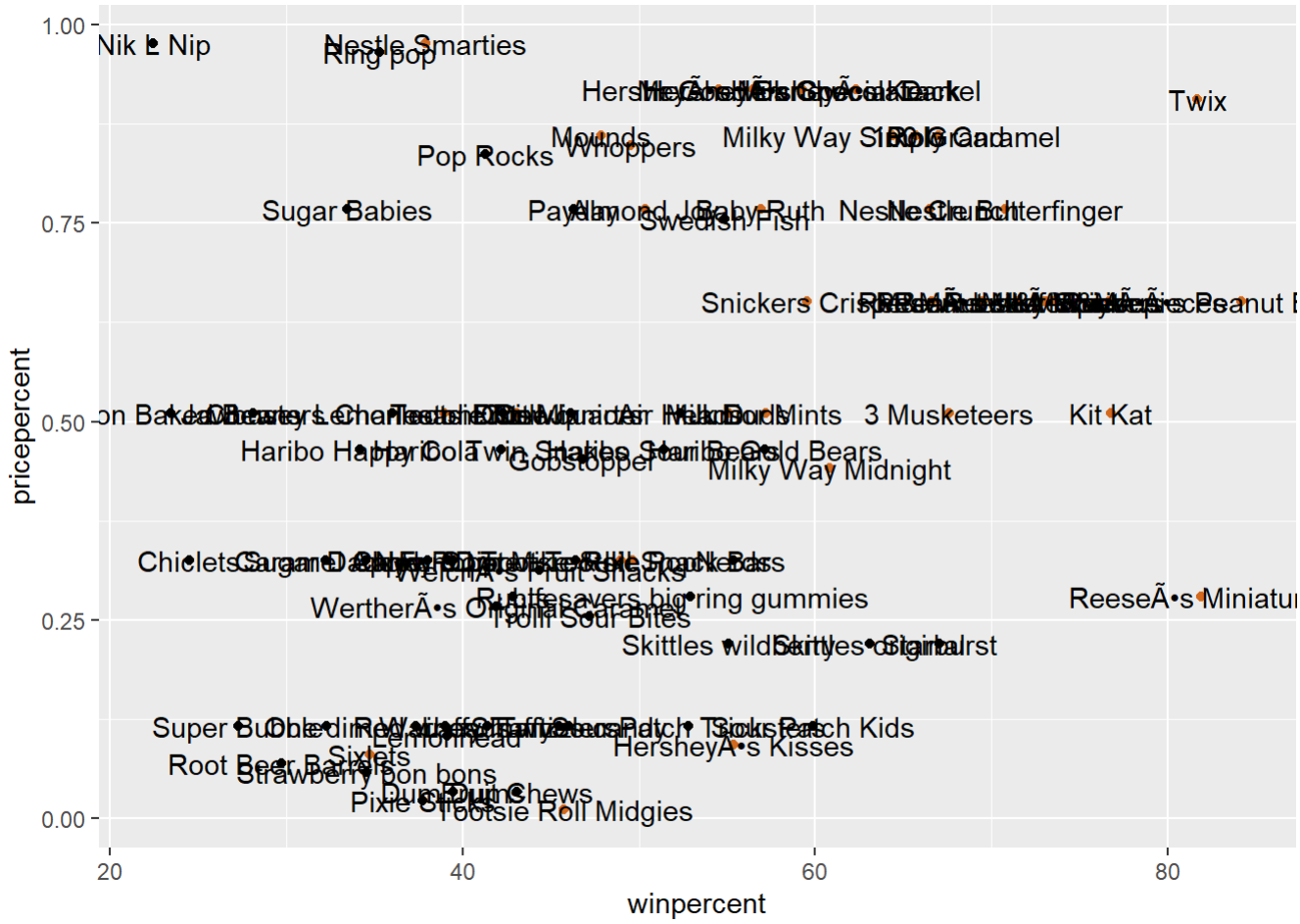
Q19. Which candy type is the highest ranked in terms of winpercent for the least money - i.e. offers the most bang for your buck?

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

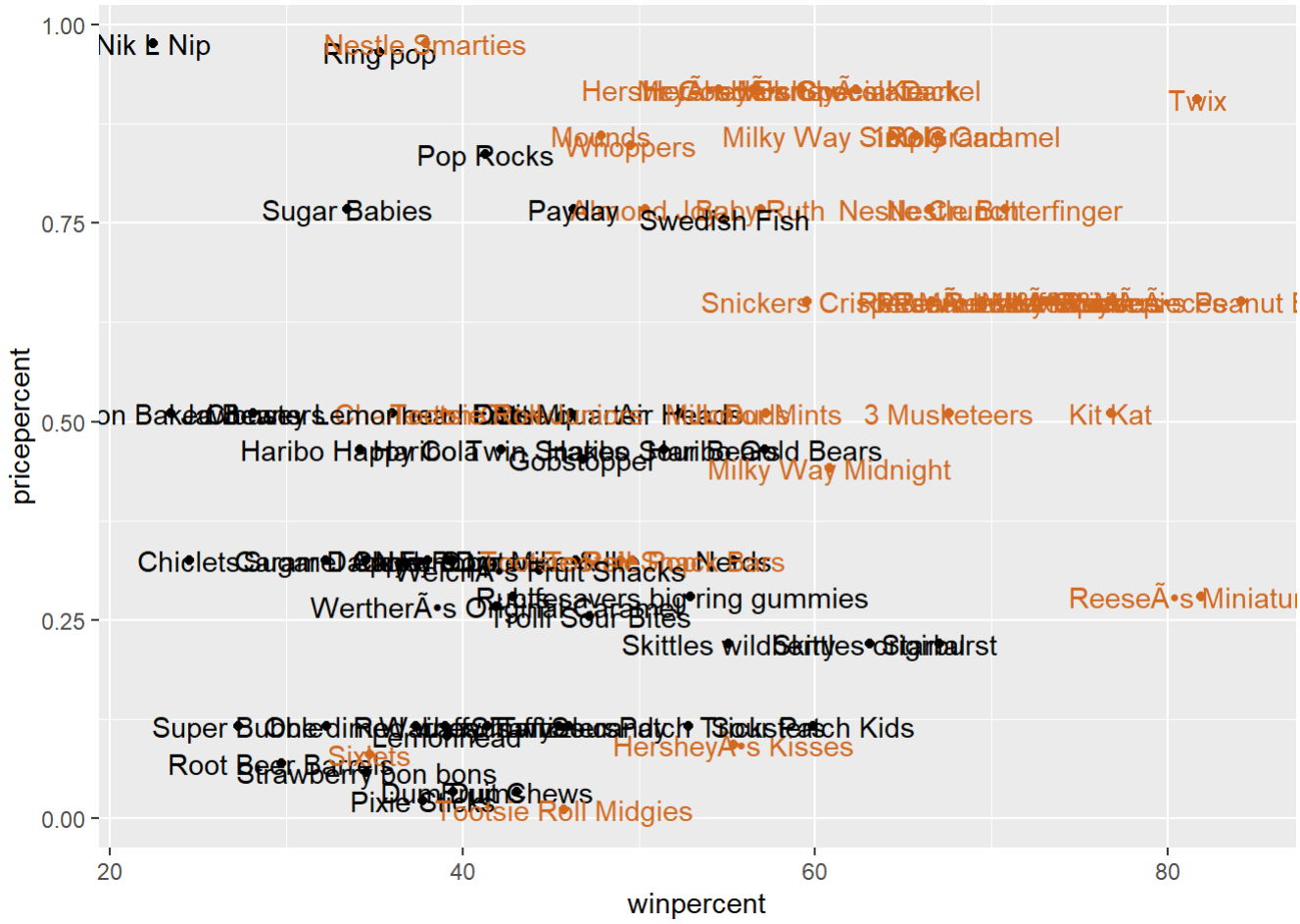
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

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

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




```
###my_cols[as.logical(candy$)]
  ggplot(candy) +
  aes(winpercent,pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text(col=my_cols)
```



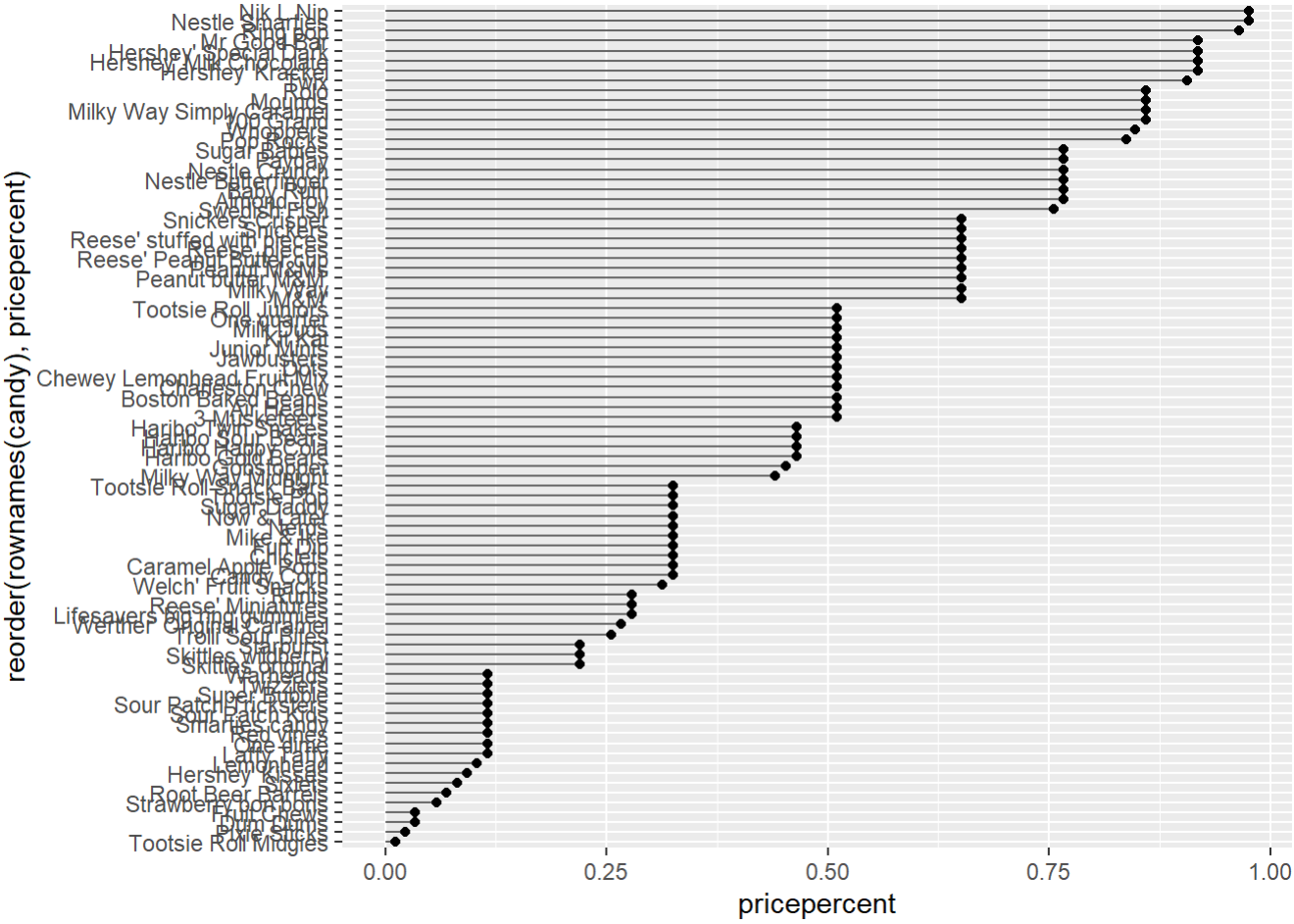

```
rownames(candy)
```

```
##  [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] "HersheyÃ•s Kisses"            "HersheyÃ•s Krackel"
## [25] "HersheyÃ•s Milk Chocolate"    "HersheyÃ•s Special Dark"
## [27] "Jawbusters"                   "Junior Mints"
## [29] "Kit Kat"                      "Laffy Taffy"
## [31] "Lemonhead"                    "Lifesavers big ring gummies"
## [33] "Peanut butter M&MÃ•s"         "M&MÃ•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"                    "ReeseÃ•s Miniatures"
## [53] "ReeseÃ•s Peanut Butter cup"   "ReeseÃ•s pieces"
## [55] "ReeseÃ•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] "WelchÃ•s Fruit Snacks"        "WertherÃ•s Original Caramel"
## [85] "Whoppers"
```

```
rownames(candy) <-gsub("Ã•s", "'",rownames(candy))
rownames(candy)
```

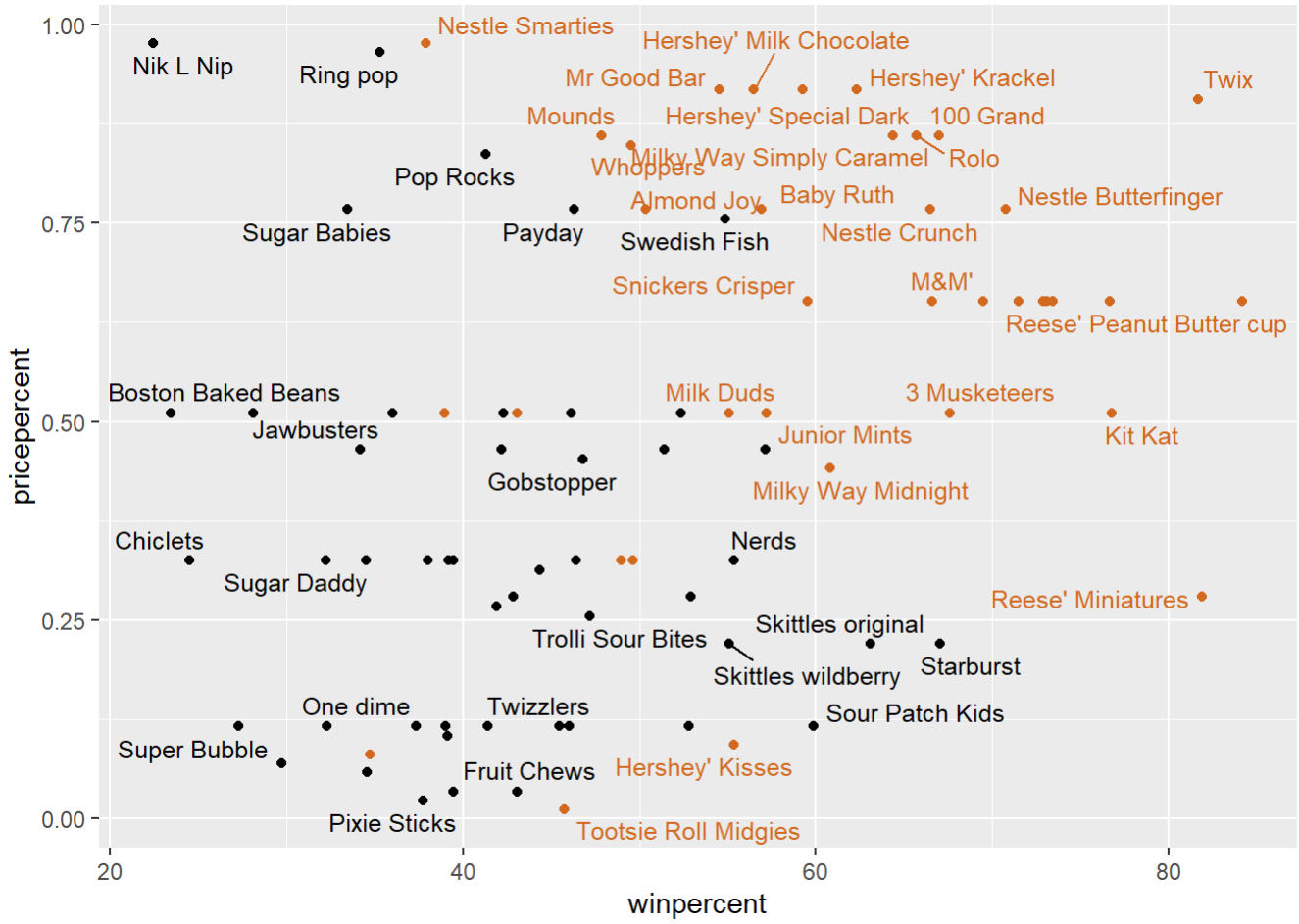
```
##  [1] "100 Grand"                   "3 Musketeers"
##  [3] "One dime"                    "One quarter"
##  [5] "Air Heads"                   "Almond Joy"
##  [7] "Baby Ruth"                   "Boston Baked Beans"
##  [9] "Candy Corn"                  "Caramel Apple Pops"
## [11] "Charleston Chew"             "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' Kisses"             "Hershey' Krackel"
## [25] "Hershey' Milk Chocolate"     "Hershey' Special Dark"
## [27] "Jawbusters"                  "Junior Mints"
## [29] "Kit Kat"                     "Laffy Taffy"
## [31] "Lemonhead"                   "Lifesavers big ring gummies"
## [33] "Peanut butter M&M'"          "M&M'"
## [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' Miniatures"
## [53] "Reese' Peanut Butter cup"    "Reese' pieces"
## [55] "Reese' 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' Fruit Snacks"         "Werther' Original Caramel"
## [85] "Whoppers"
```

```
# Make a lollipop chart of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```




```
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 7)
```

```
## Warning: ggrepel: 37 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

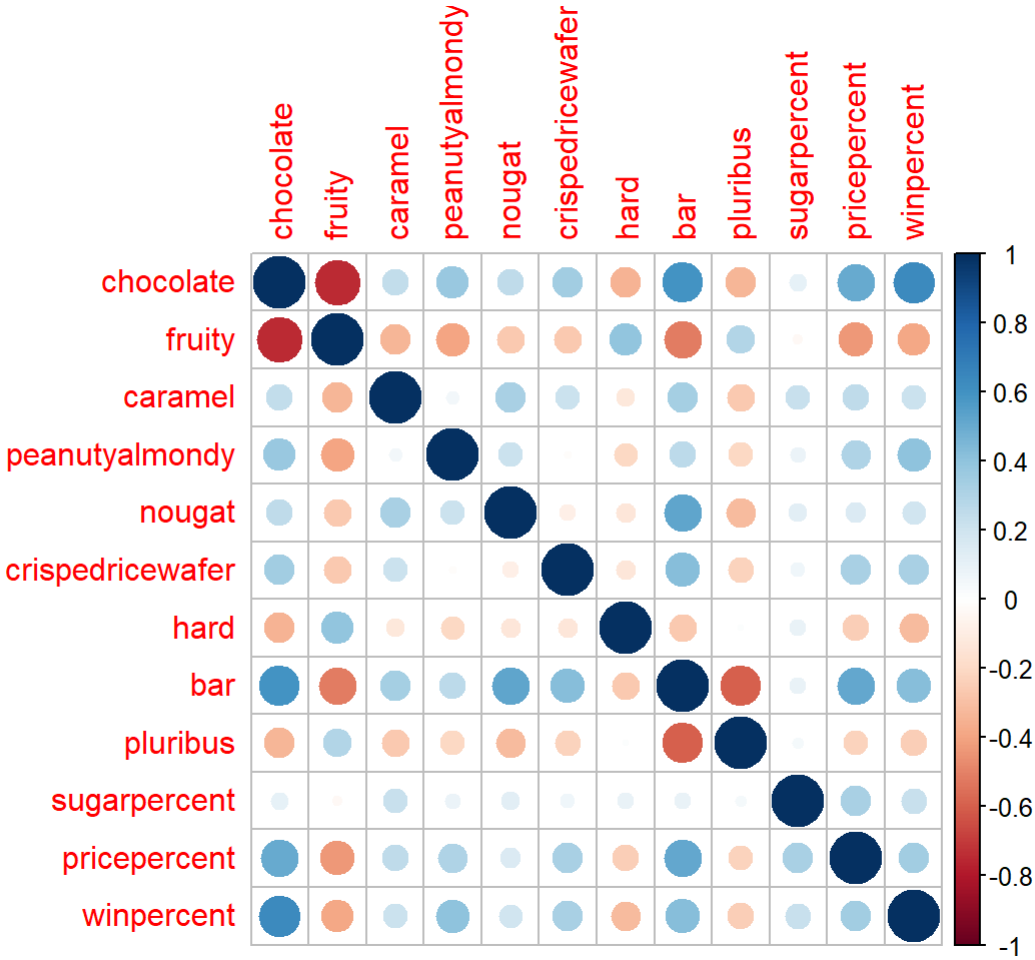


#Correlation analysis

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
cij <- cor(candy)
corrplot(cij)
```



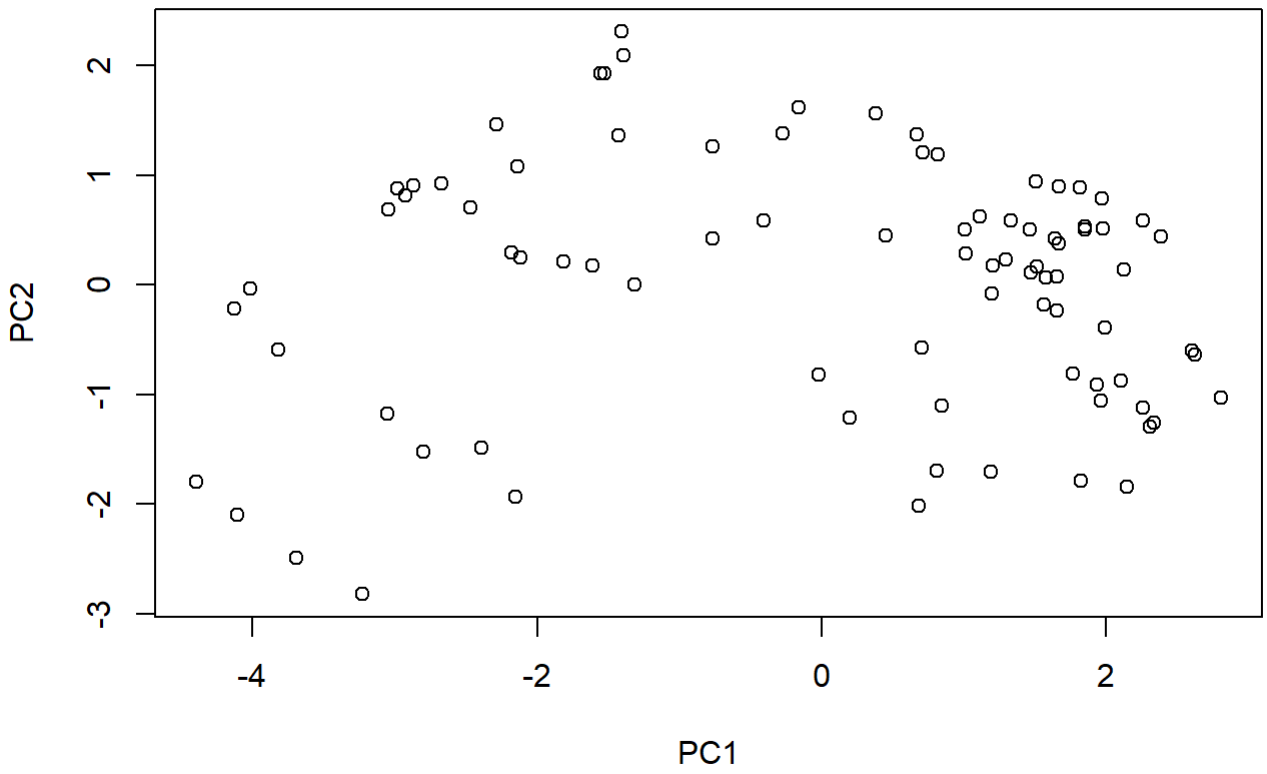
6.

Principal Component Analysis

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

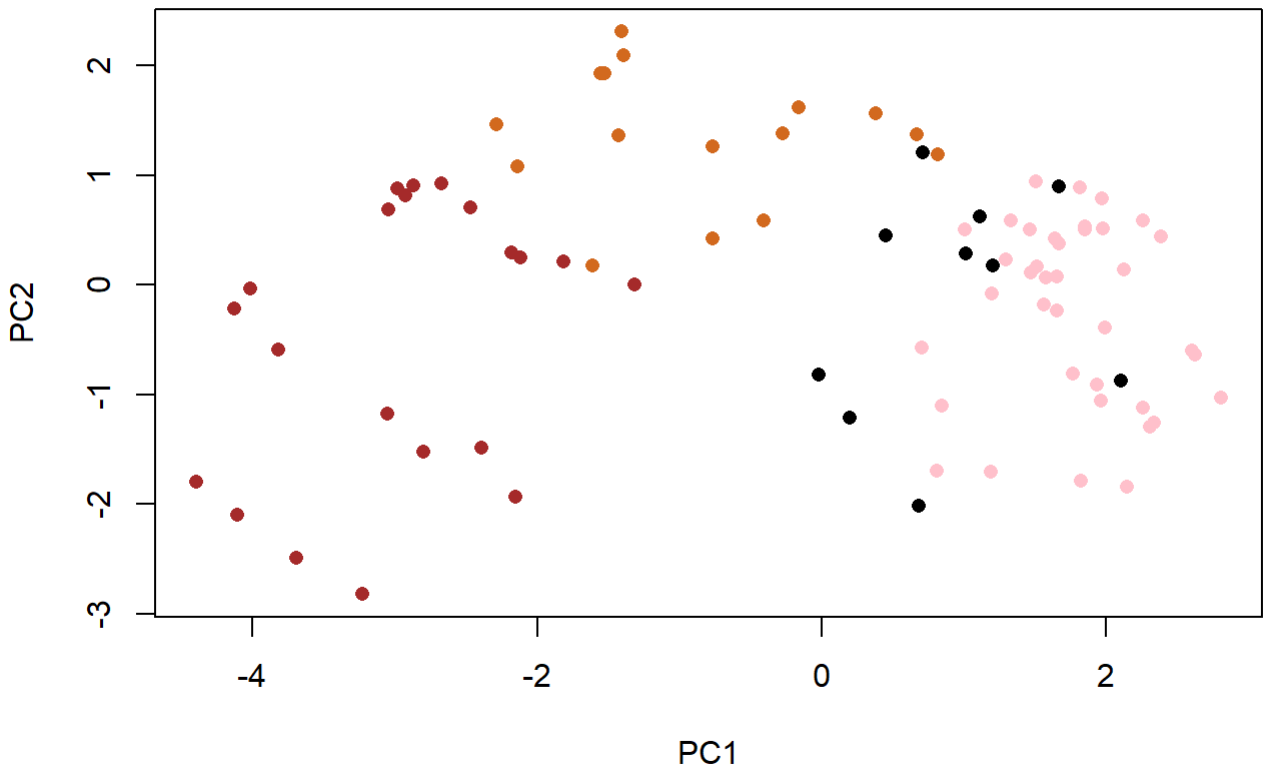
```
## Importance of components:
##                           PC1    PC2    PC3     PC4    PC5     PC6     PC7
## Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
## Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
## Cumulative Proportion  0.3601 0.4680 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])
```




```
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"
```

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

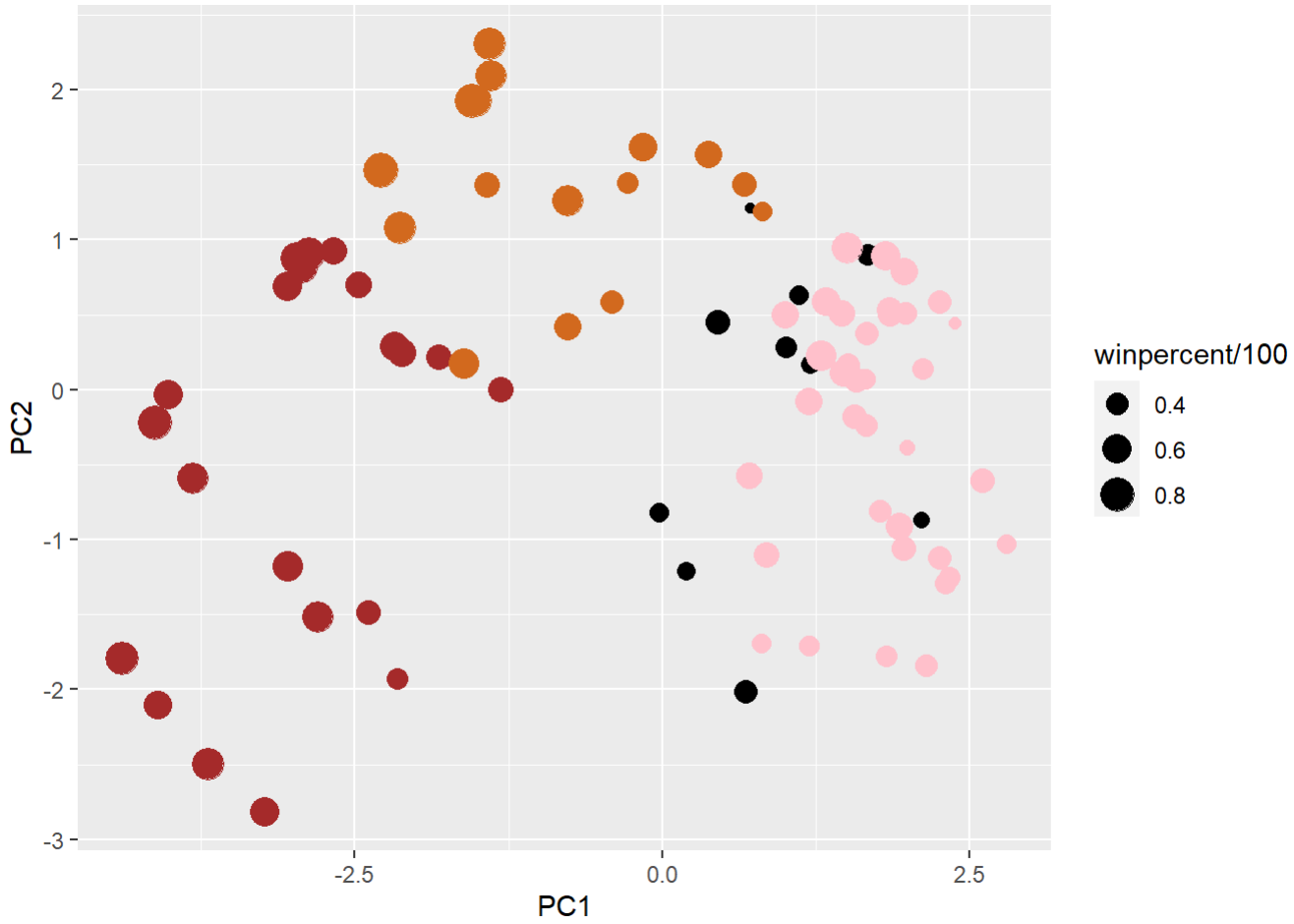


#

Make a new data-frame with our PCA results and candy data

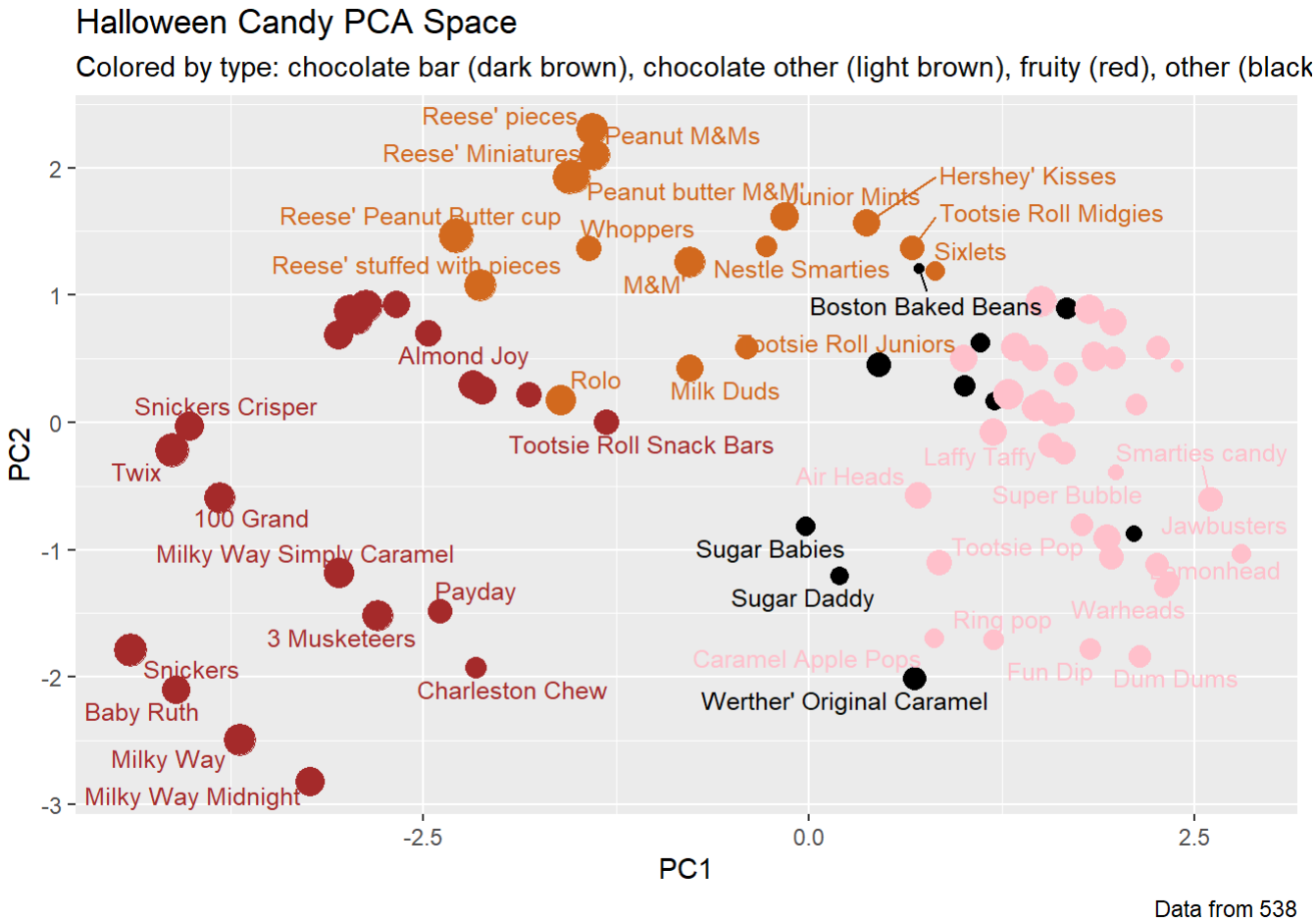
```
my_data <- cbind(candy, pca$x[,1:3])
```

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

p
```




```
library(ggrepel)

p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 7)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown), fru
ity (red), other (black)",
       caption="Data from 538")
```

```
## Warning: ggrepel: 40 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
library(plotly)
```

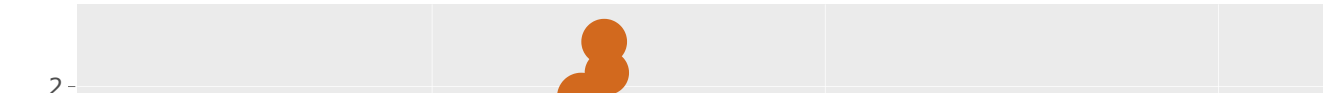
```
##
## Attaching package: 'plotly'
```

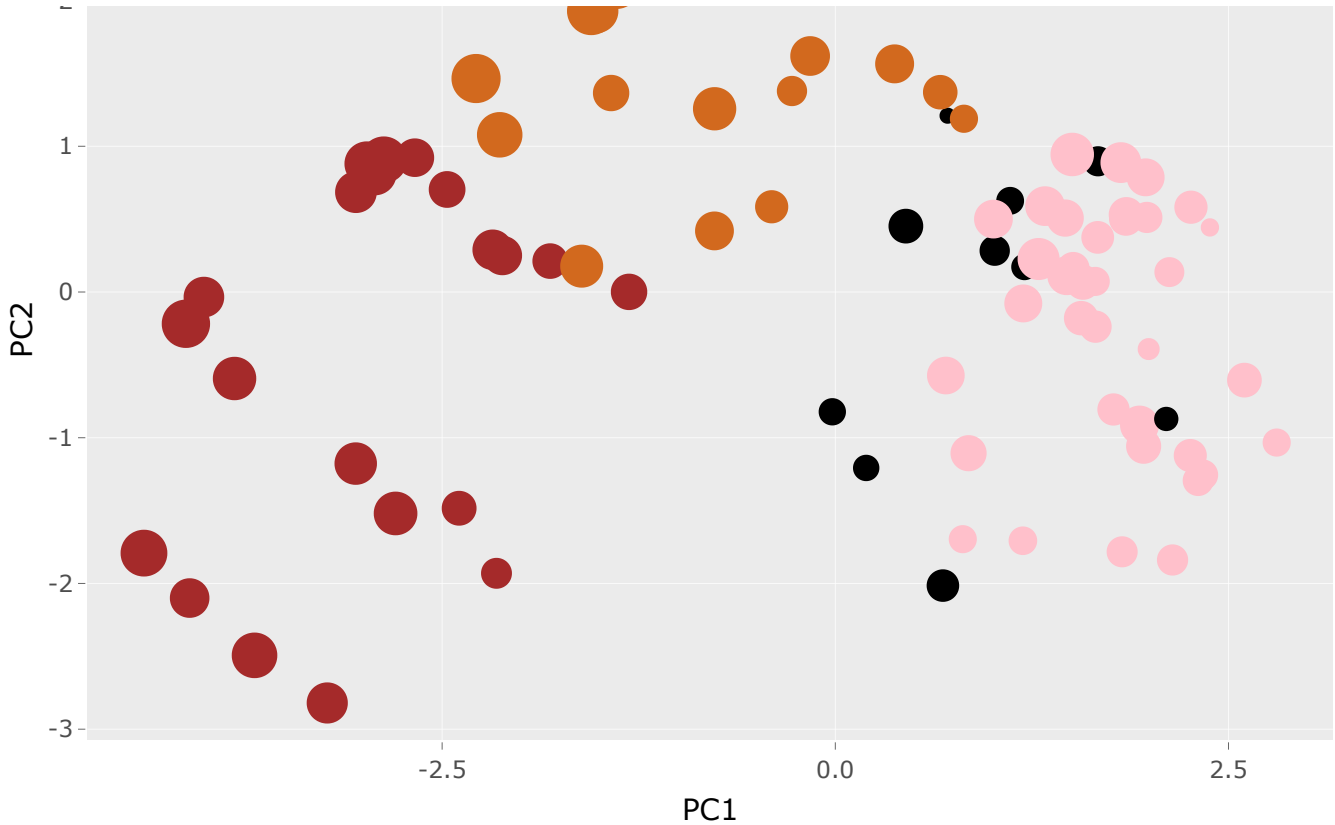
```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

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

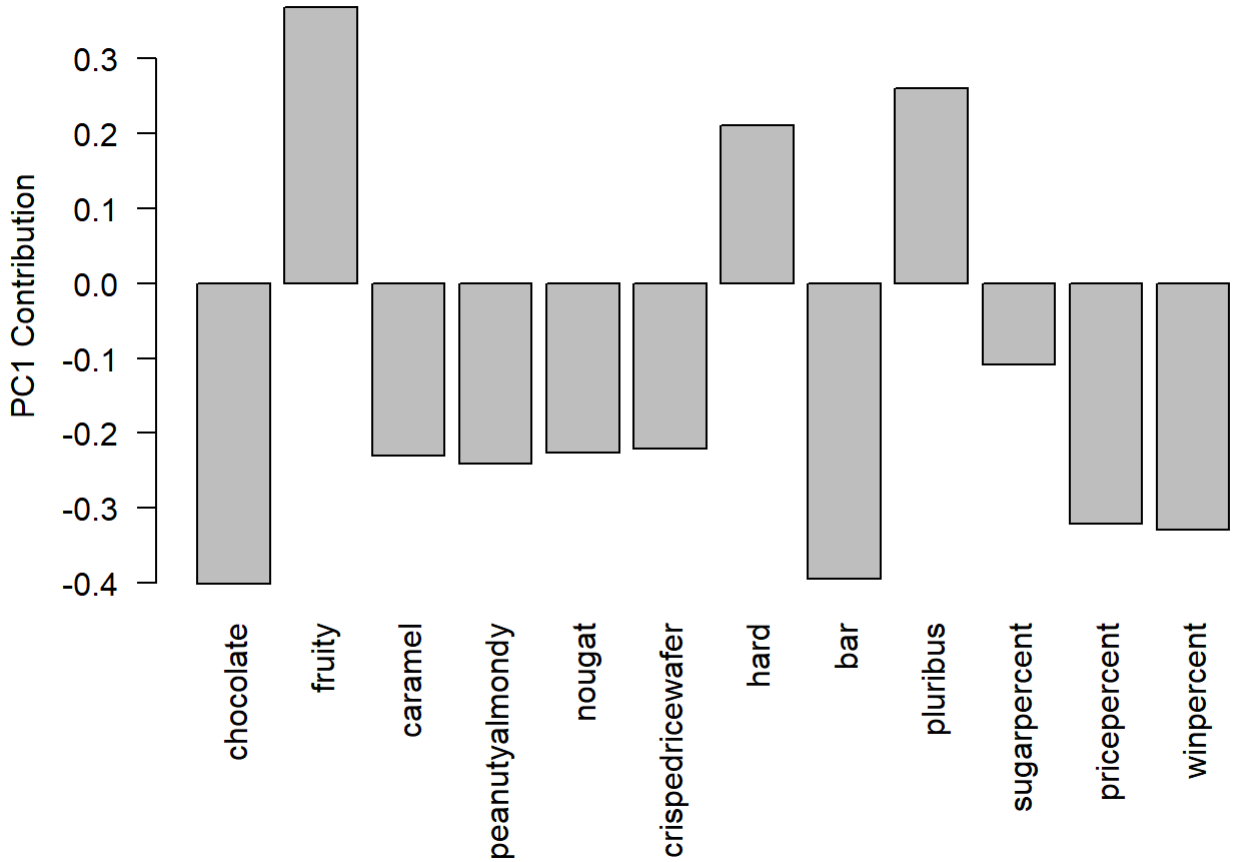
```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
ggplotly(p)
```






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



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

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