Class 9: Halloween Mini Project

AUTHOR
Emily Hickey (PID: A15575724)

Importing candy data

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

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
```

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

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

```
[1] 38
```

What is your favorite candy?

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

```r
candy["Milky Way",]$winpercent
```

[1] 73.09956

The value for my favorite candy, Milky Way, is 73.09956

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

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

[1] 76.7686

The winpercent value for Kit Kat is 76.7686

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

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

[1] 49.6535

The winpercent value for Tootsie Roll Snack Bars is 49.6535

```r
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	▇▁▁▁▂

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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?

Yes, the winpercent value looks to be on a different scale to the majority of the other columns in the dataset. It was the only only that had values greater than 1.

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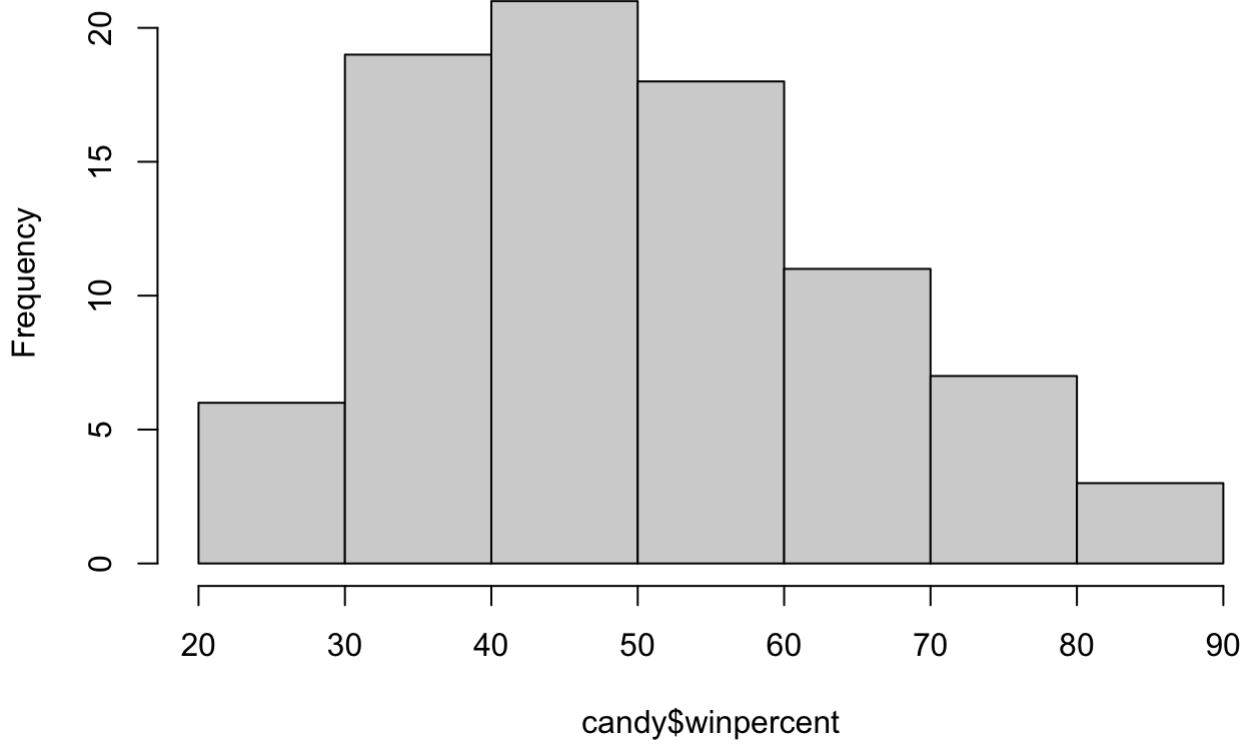
```
candy$chocolate
```

```
 [1] 1 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 0 0 0 1 1 0 1 1 1
[39] 1 1 1 0 1 1 0 0 0 1 0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1
[77] 1 1 0 1 0 0 0 0 1
```

The zeros and ones represent whether the candy contains chocolate or not. 1 means it does, 0 means it does not.

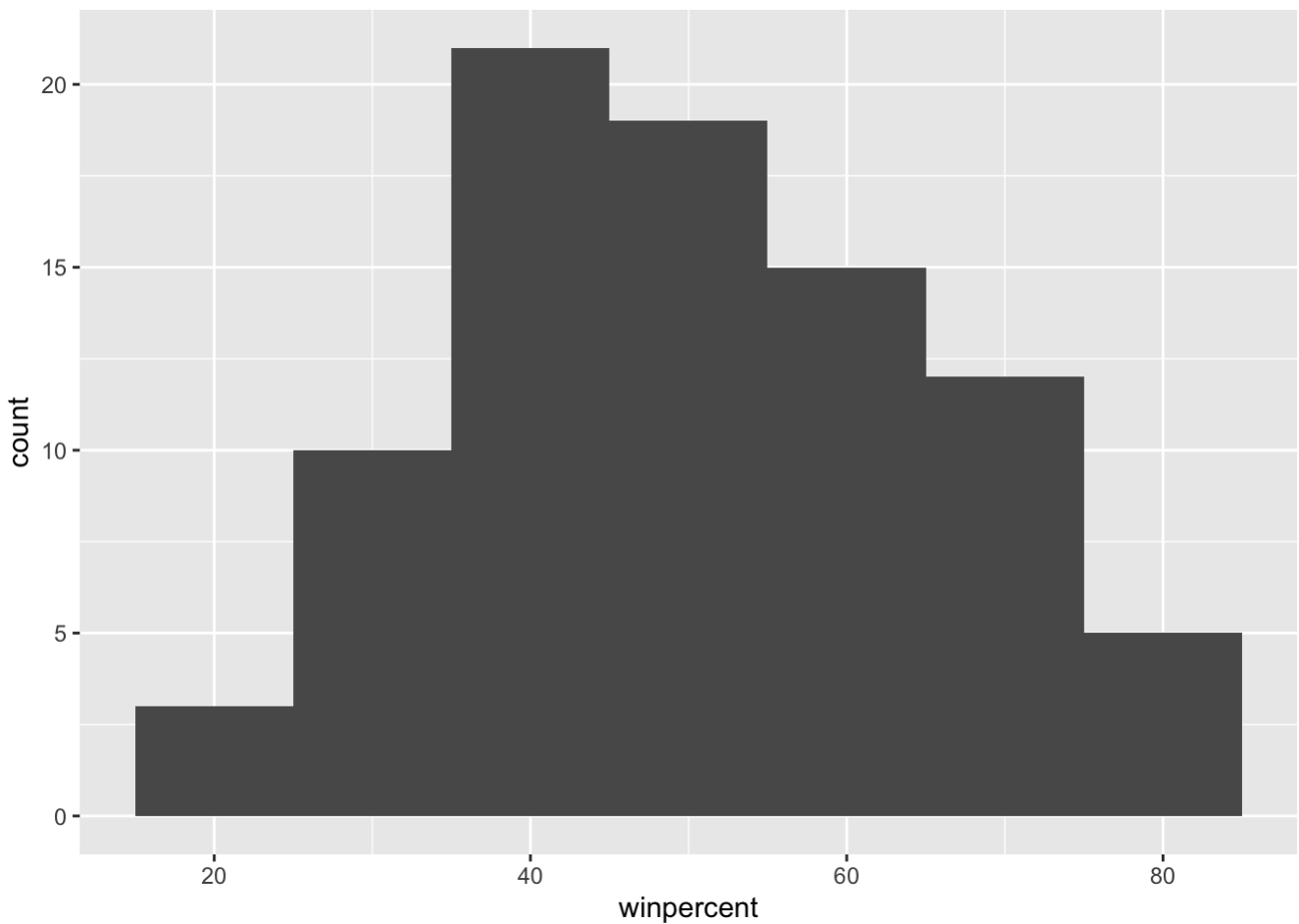> Q8. Plot a histogram of winpercent values

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

Histogram of candy$winpercent


```r
library(ggplot2)

ggplot(candy) +
  aes(winpercent) +
  geom_histogram(binwidth=10)
```

> Q9. Is the distribution of winpercent values symmetrical?

No.

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

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

```
[1] 50.31676
```

It's above 50%. It is 50.31676

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

```r
choc.inds <- as.logical(candy$chocolate)
choc.win <- candy[choc.inds,12]
choc.win
```

```
 [1] 66.97173 67.60294 50.34755 56.91455 38.97504 55.37545 62.28448 56.49050
 [9] 59.23612 57.21925 76.76860 71.46505 66.57458 55.06407 73.09956 60.80070
[17] 64.35334 47.82975 54.52645 70.73564 66.47068 69.48379 81.86626 84.18029
```

```
[25] 73.43499 72.88790 65.71629 34.72200 37.88719 76.67378 59.52925 48.98265
[33] 43.06890 45.73675 49.65350 81.64291 49.52411
```

```r
mean(choc.win)
```

```
[1] 60.92153
```

```r
fruit.inds <- as.logical(candy$fruity)
fruit.win <- candy[fruit.inds,12]
fruit.win
```

```
 [1] 52.34146 34.51768 36.01763 24.52499 42.27208 39.46056 43.08892 39.18550
 [9] 46.78335 57.11974 51.41243 42.17877 28.12744 41.38956 39.14106 52.91139
[17] 46.41172 55.35405 22.44534 39.44680 41.26551 37.34852 35.29076 42.84914
[25] 63.08514 55.10370 45.99583 59.86400 52.82595 67.03763 34.57899 27.30386
[33] 54.86111 48.98265 47.17323 45.46628 39.01190 44.37552
```

```r
mean(fruit.win)
```

```
[1] 44.11974
```

Chocolate is ranked higher than fruit candy.

> Q12. Is this difference statistically significant?

```r
t.test(choc.win,fruit.win)
```

```
	Welch Two Sample t-test

data:  choc.win and fruit.win
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, they are statistical different because the p-value is 2.871e-08, which is less than .05.

Overall Candy Rankings

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

```
library(dplyr)
```

```
Attaching package: 'dplyr'

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

    filter, lag

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

    intersect, setdiff, setequal, union
```

```
candy %>%
  arrange(winpercent) %>% head(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

Least liked: Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, Jawbusters

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

```
library(dplyr)
candy %>%
  arrange(winpercent) %>% tail(5)
```

	chocolate	fruity	caramel	peanutyalmondy	nougat
Snickers	1	0	1	1	1
Kit Kat	1	0	0	0	0
Twix	1	0	1	0	0
Reese's Miniatures	1	0	0	1	0

	crisped rice wafer	hard	bar	pluribus	sugarpercent
Reese's Peanut Butter cup	1	0	0	1	0

	crispedricewafer	hard	bar	pluribus	sugarpercent
Snickers	0	0	1	0	0.546
Kit Kat	1	0	1	0	0.313
Twix	1	0	1	0	0.546
Reese's Miniatures	0	0	0	0	0.034
Reese's Peanut Butter cup	0	0	0	0	0.720

	pricepercent	winpercent
Snickers	0.651	76.67378
Kit Kat	0.511	76.76860
Twix	0.906	81.64291
Reese's Miniatures	0.279	81.86626
Reese's Peanut Butter cup	0.651	84.18029

Most liked: Reese's Peanut Butter Cup, Reese's Miniatures, Twix, Kit Kat, Snickers

> Q15. Make a first barplot of candy ranking based on winpercent values.
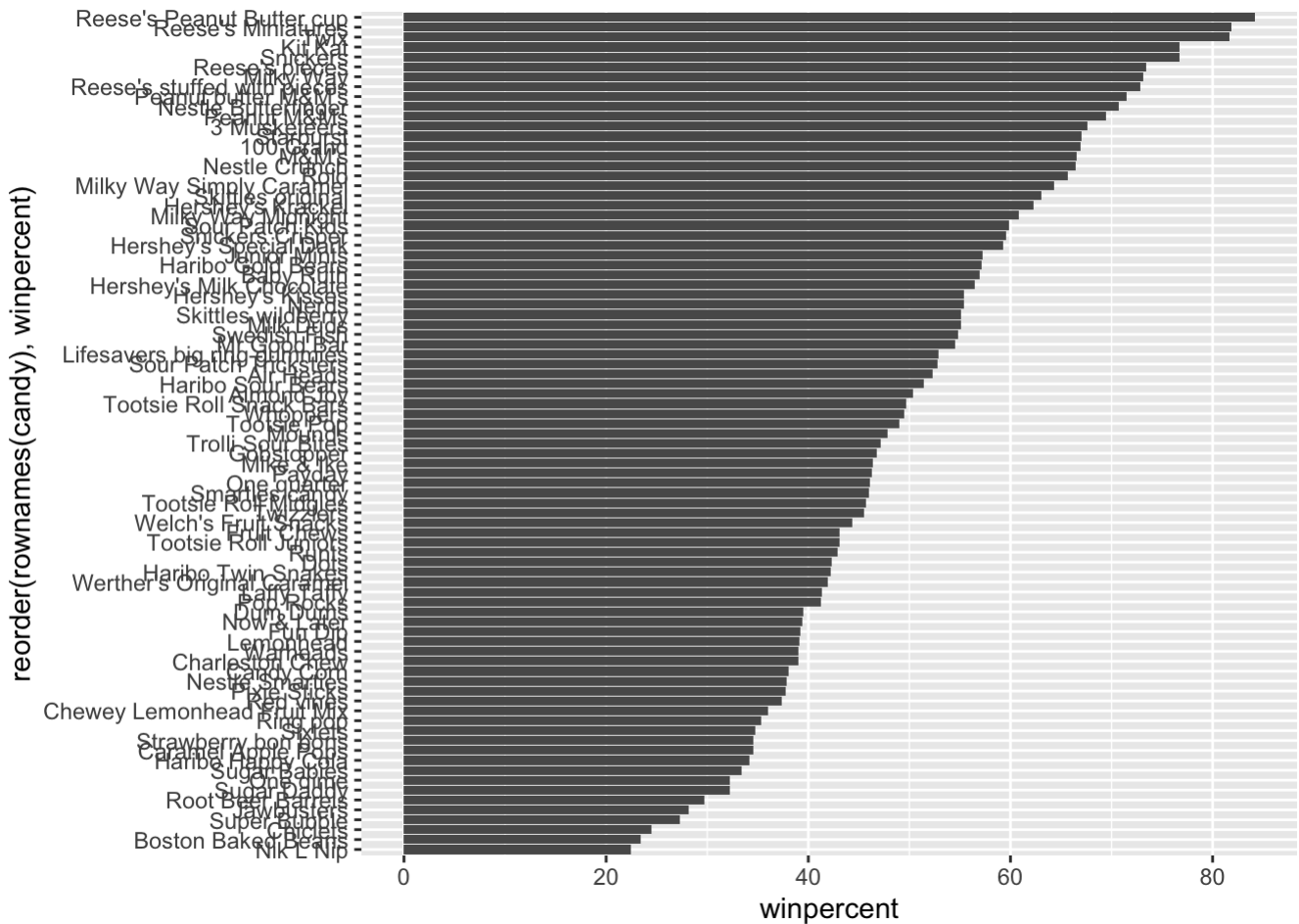
```
library(ggplot2)

ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_col()
```
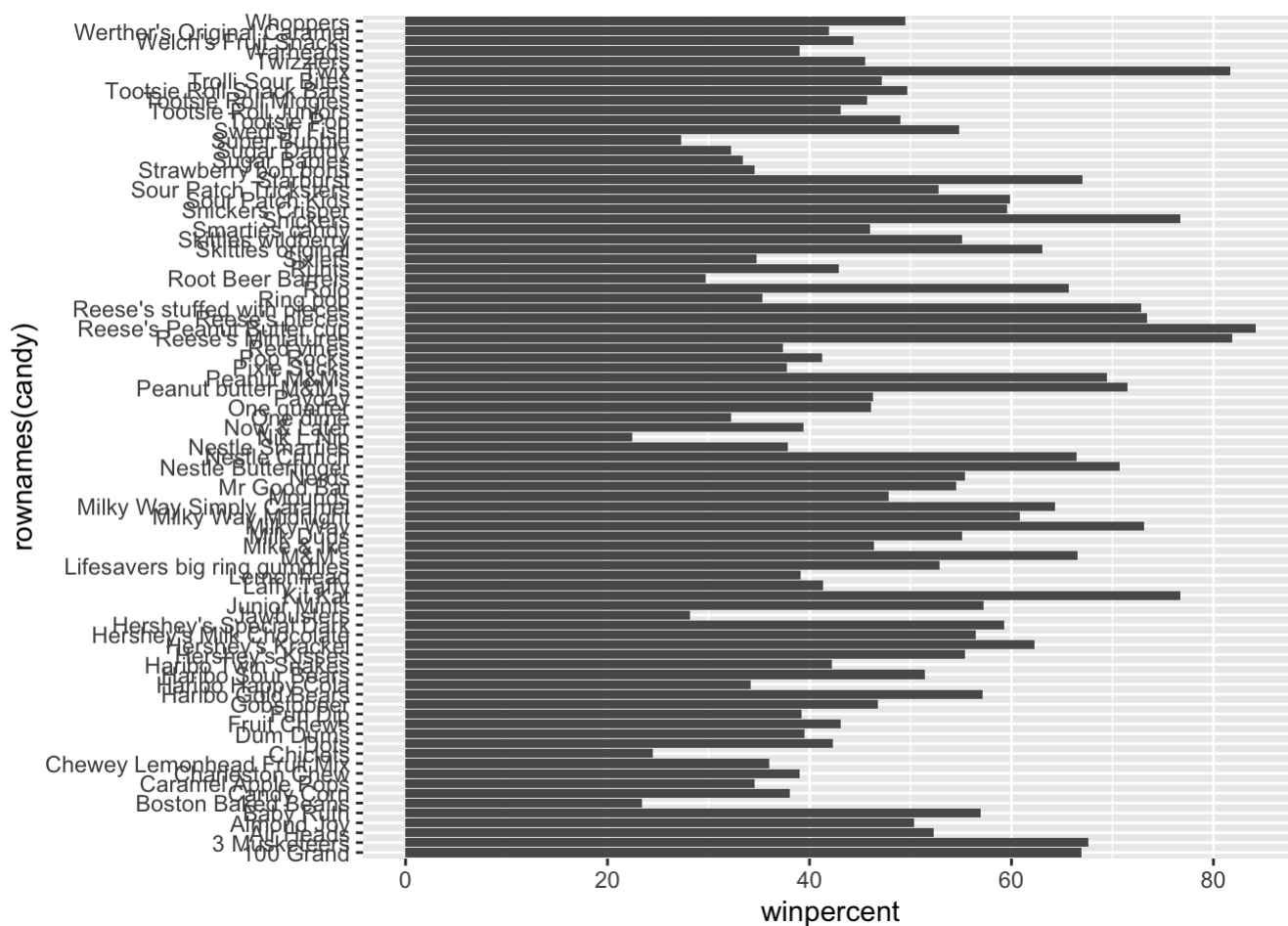
> Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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



```
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"
my_cols
```

```
 [1] "brown"     "brown"     "black"     "black"     "pink"      "brown"
 [7] "brown"     "black"     "black"     "pink"      "brown"     "pink"
[13] "pink"      "pink"      "pink"      "pink"      "pink"      "pink"
[19] "pink"      "black"     "pink"      "pink"      "chocolate" "brown"
[25] "brown"     "brown"     "pink"      "chocolate" "brown"     "pink"
[31] "pink"      "pink"      "chocolate" "chocolate" "pink"      "chocolate"
[37] "brown"     "brown"     "brown"     "brown"     "brown"     "pink"
[43] "brown"     "brown"     "pink"      "pink"      "brown"     "chocolate"
[49] "black"     "pink"      "pink"      "chocolate" "chocolate" "chocolate"
```

```
[55] "chocolate" "pink"      "chocolate" "black"      "pink"      "chocolate"
[61] "pink"      "pink"      "chocolate" "pink"      "brown"     "brown"
[67] "pink"      "pink"      "pink"      "pink"      "black"     "black"
[73] "pink"      "pink"      "pink"      "chocolate" "chocolate" "brown"
[79] "pink"      "brown"     "pink"      "pink"      "pink"      "black"
[85] "chocolate"
```

```r
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols)
```
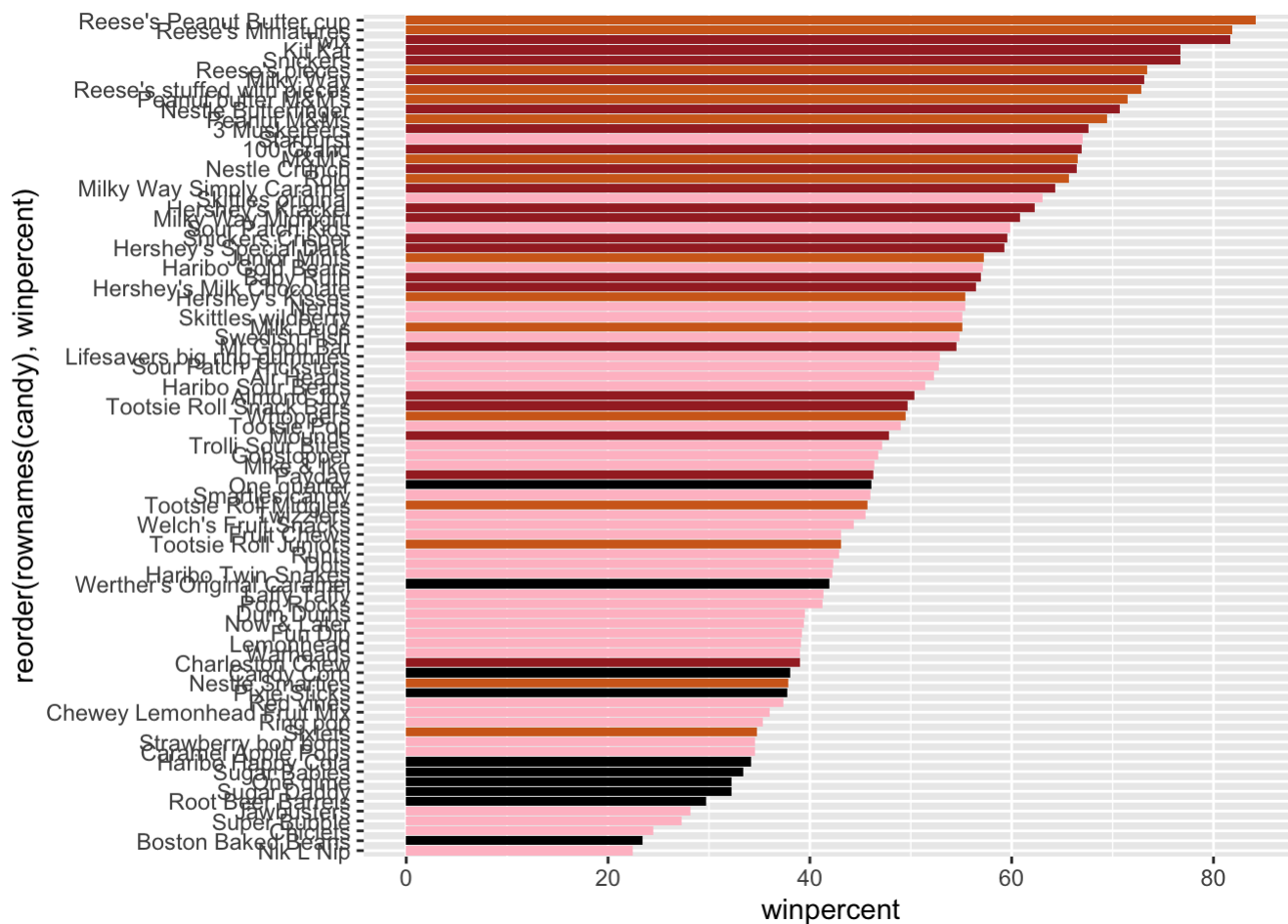

> Q17. What is the worst ranked chocolate candy?
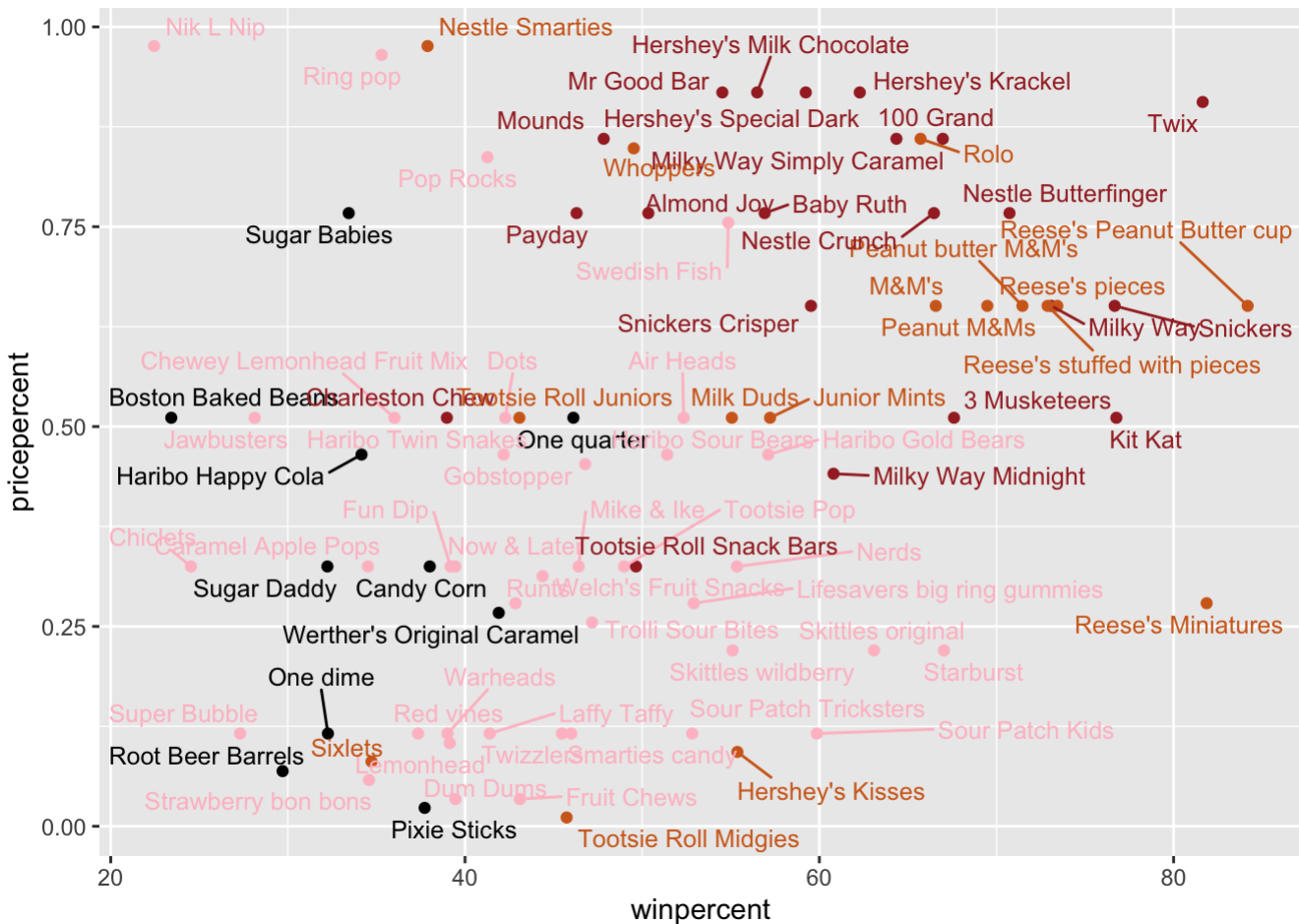
The worse ranked chocolate candy is Sixlets.

> Q18. What is the best ranked fruity candy?

The best ranked fruit candy is Starburst.

Taking a look at pricepercent

```
library(ggrepel)

ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps=23)
```


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?

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

```
                          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
```

Reese's Miniatures

> 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
Hershey's Krackel             0.918   62.28448
Hershey's Milk Chocolate      0.918   56.49050
```
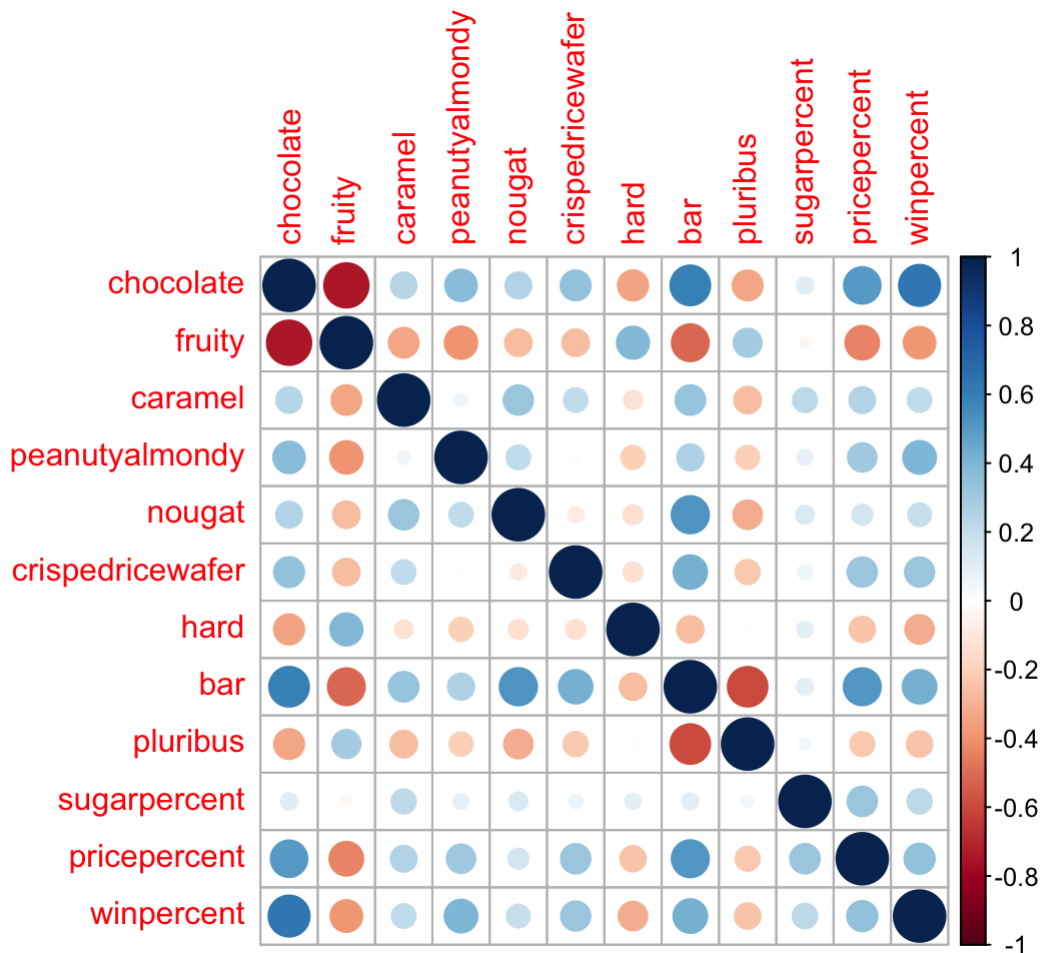
Top 5 Most Expensive: Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, Hershey's Milk Chocolate. Least popular: Nik L Nip

Exploring the correlation structure

```
library(corrplot)
```

corrplot 0.92 loaded

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

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

Chocolate and fruit are anti-correlated.

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

Winpercent and chocolate.

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
```
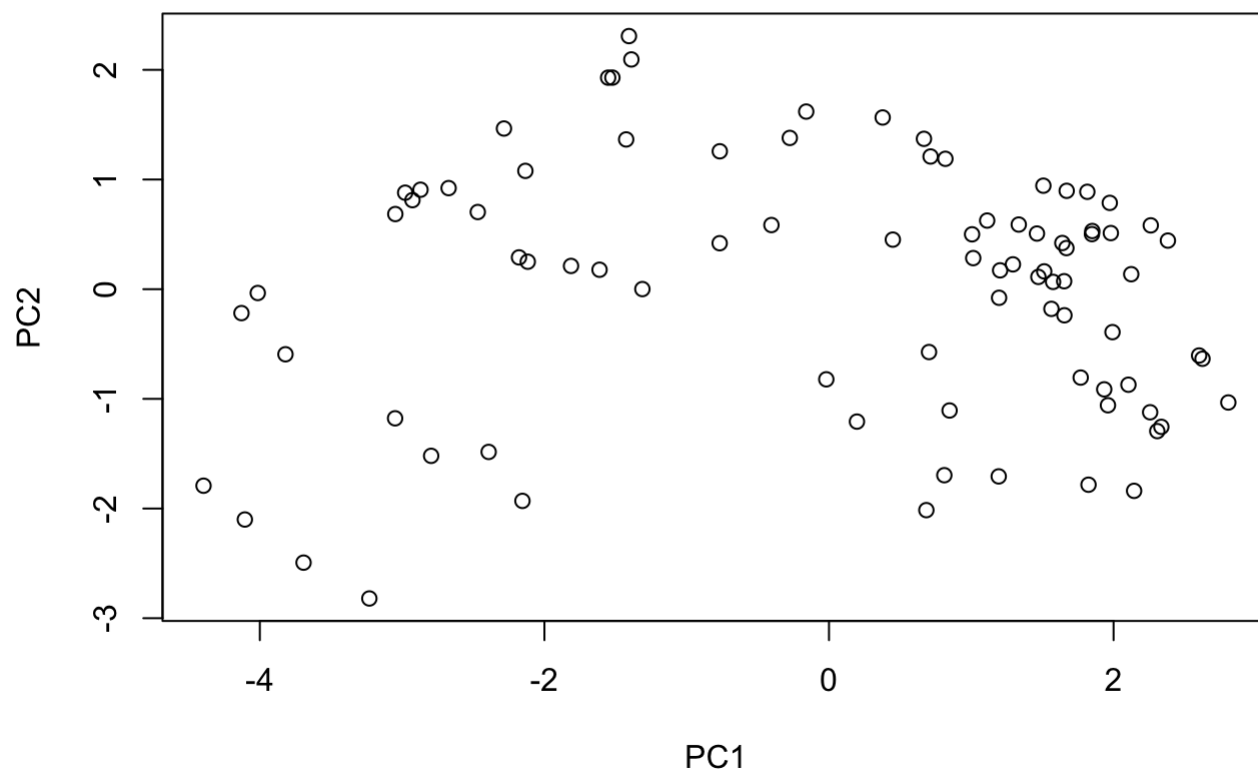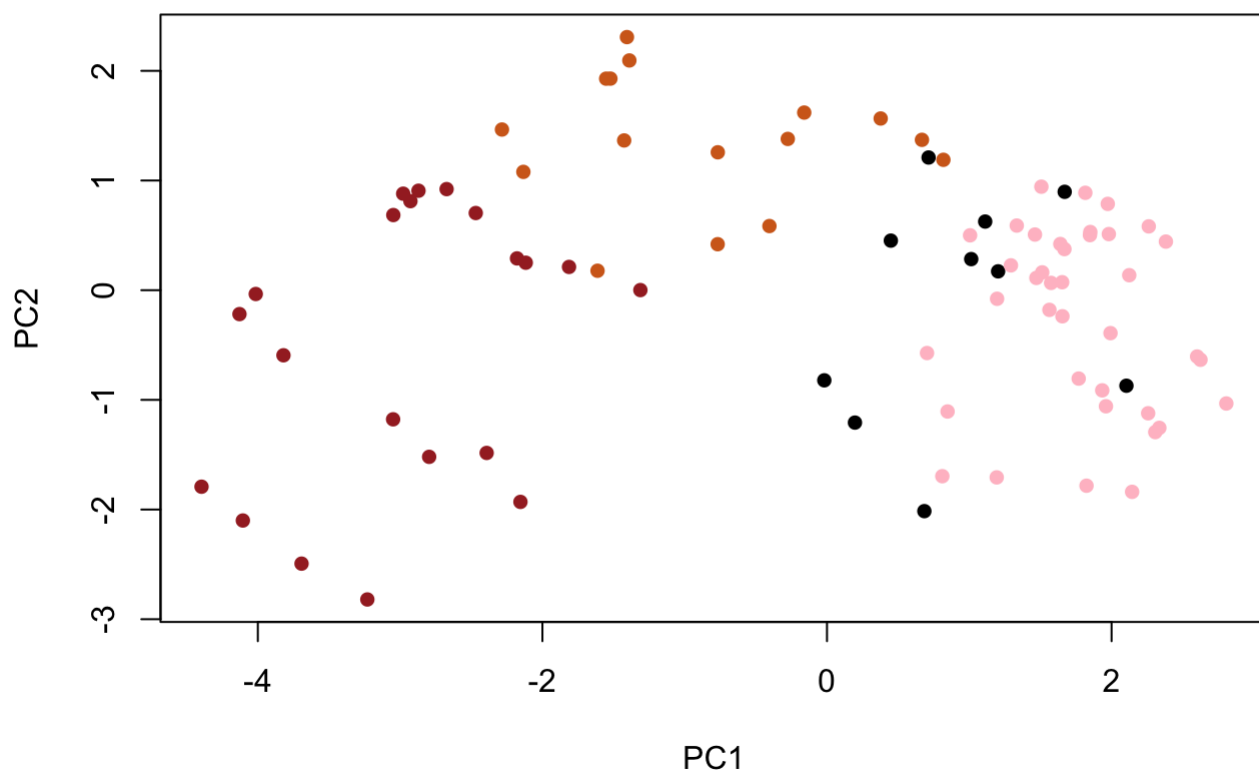
```
Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

```
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=my_cols)

p
```
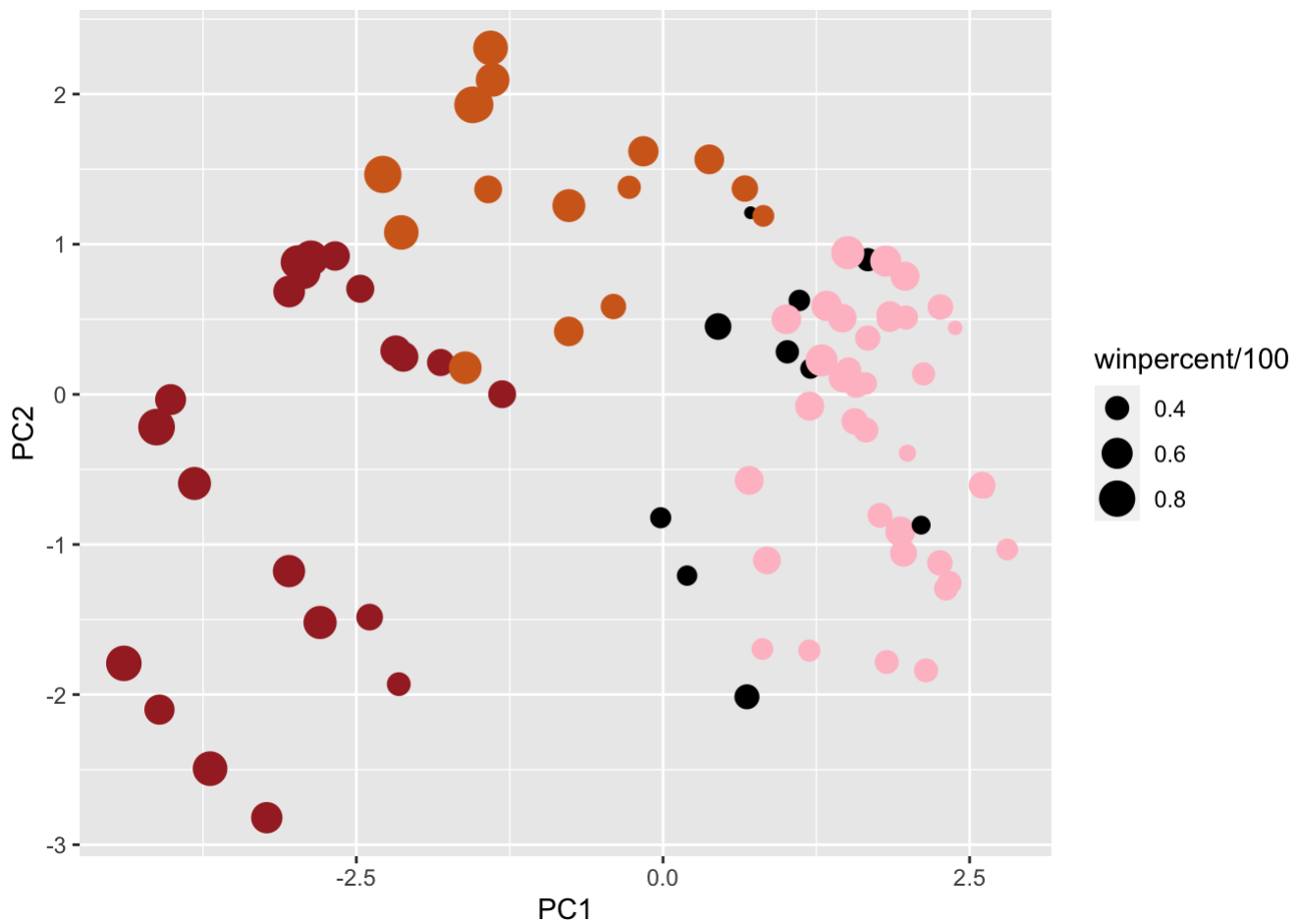
```
p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 100)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brow
       caption="Data from 538")
```

Halloween Candy PCA Space
Colored by type: chocolate bar (dark brown), chocolate other (light brown), fruity (red), other (blac
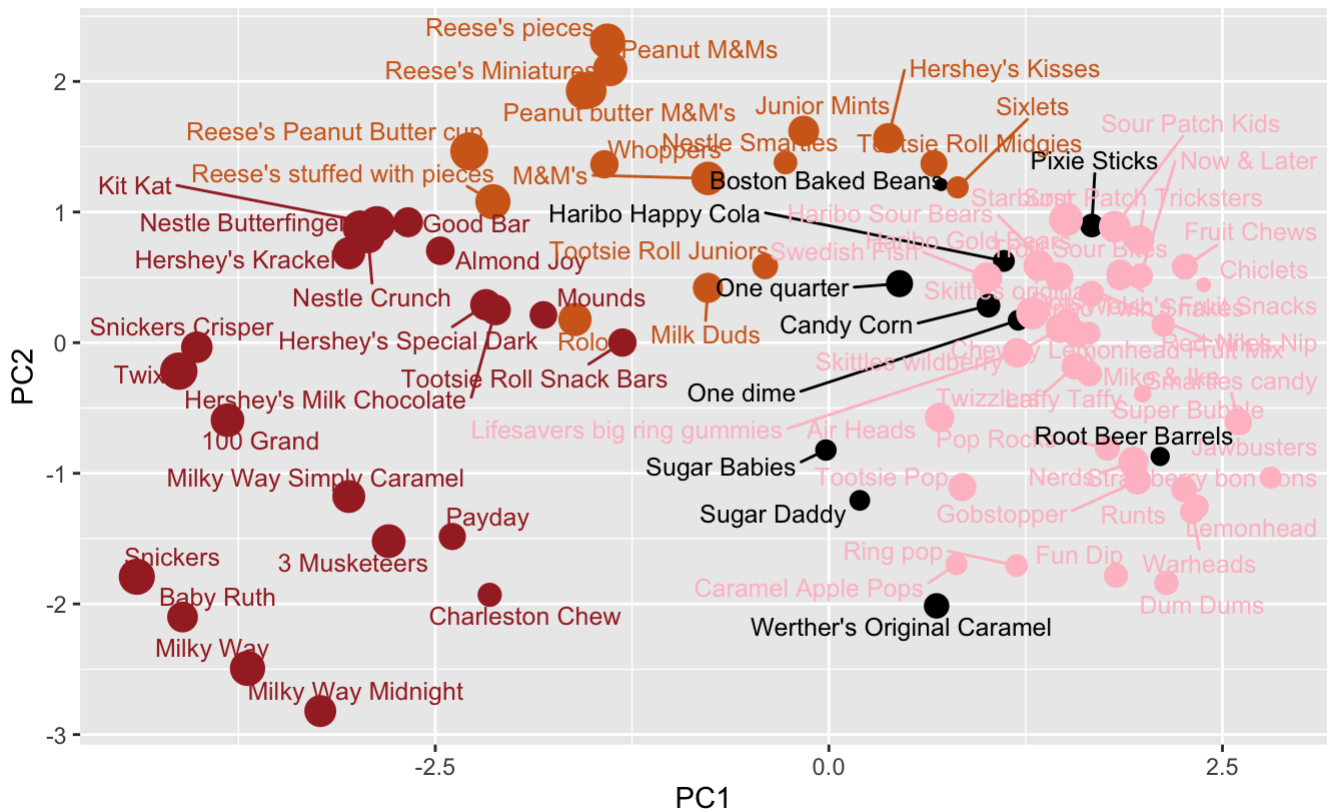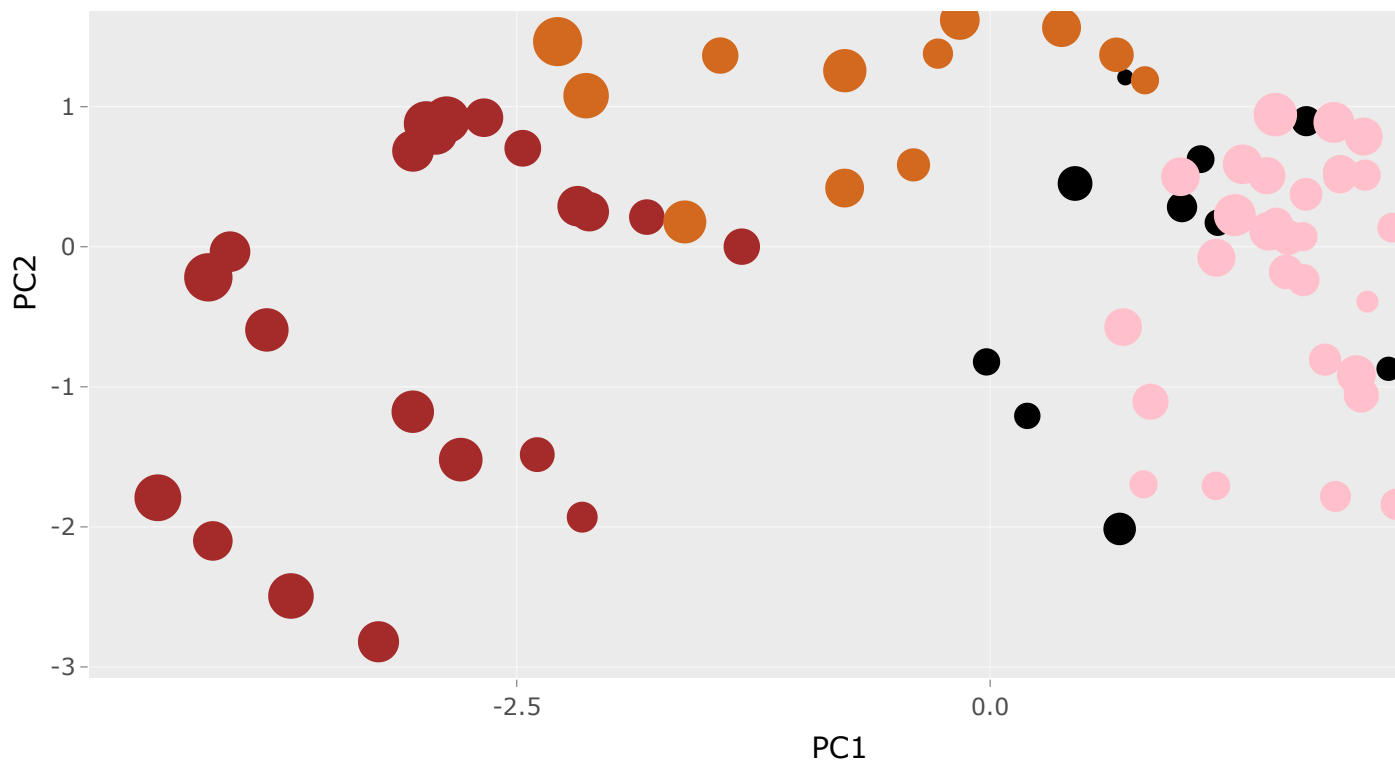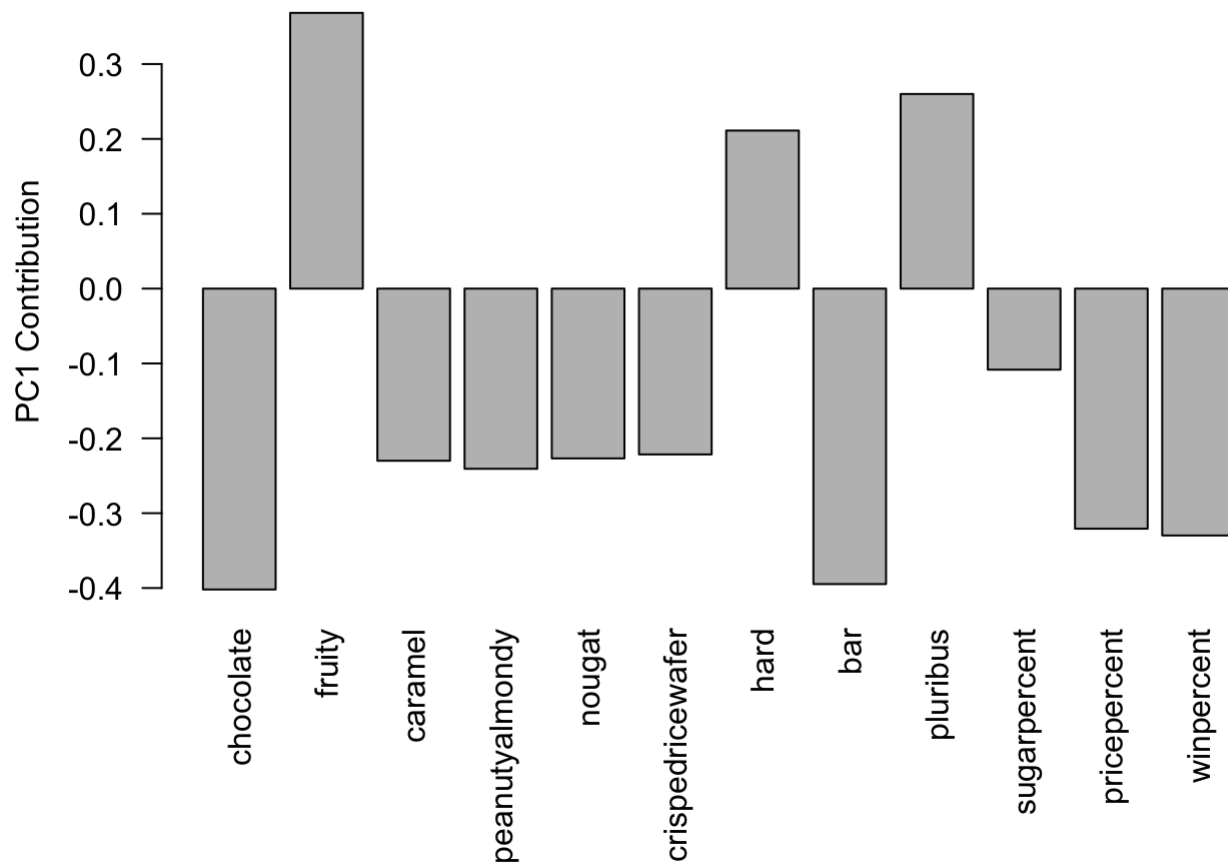

PC2 (vertical axis)
PC1 (horizontal axis)

Reese's pieces
Peanut M&Ms
Reese's Miniatures
Hershey's Kisses
Peanut butter M&M's   Junior Mints   Sixlets
Sour Patch Kids
Reese's Peanut Butter cup
Nestle Smarties   Tootsie Roll Midgies   Pixie Sticks   Now & Later
Whoppers   Starburst   Sour Patch Tricksters
Kit Kat   Reese's stuffed with pieces   M&M's
Boston Baked Beans
Fruit Chews
Nestle Butterfinger   Good Bar   Haribo Happy Cola   Haribo Sour Bears   Sour Bites
Hershey's Krackel   Tootsie Roll Juniors   Swedish Fish   Haribo Gold-Bears   Chiclets
Almond Joy
Nestle Crunch   Mounds   One quarter   Skittles original   Welch's Fruit Snacks
Snickers Crisper   Hershey's Special Dark   Rolo   Milk Duds   Candy Corn   Chewey Lemonhead Fruit Mix   Red Wicked Nip
Twix   Skittles wildberry   Mike & Ikes candy
Tootsie Roll Snack Bars   One dime   Twizzlers   Laffy Taffy   Super Bubble
Hershey's Milk Chocolate   Lifesavers big ring gummies   Air Heads   Pop Rocks   Root Beer Barrels   Jawbusters
100 Grand   Sugar Babies   Tootsie Pop   Nerds   Strawberry bon bons
Milky Way Simply Caramel   Sugar Daddy   Gobstopper   Runts   Lemonhead
Payday   Ring pop   Fun Dip   Warheads
Snickers   3 Musketeers   Caramel Apple Pops   Dum Dums
Baby Ruth   Werther's Original Caramel
Charleston Chew
Milky Way
Milky Way Midnight

Data from 538

```
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")
```

Fruity, hard, and pluribus are all picked up strongly by PC1 in the positive direction. This makes sense because fruity candy does typically come in hard form and in a bag or box of multiple candies. These traits follow the correlation structure we made earlier.