Class 11 Candy

David Ma

In today's class we will examine 558 candy data and see if this helps us gain some more feeling for how PCA and other methods work.

```
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?

```
nrow(candy)
```

```
[1] 85
```

Q2. How many fruity candy types are in this dataset?

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

```
[1] 38
```

Q. What are these fruity candies?

```
candy[ candy$fruity == 1, ]
```

	chocolate	fruity	caramel	peanutyalmondy	nougat
Air Heads	0	1	0	0	0
Caramel Apple Pops	0	1	1	0	0
Chewey Lemonhead Fruit Mix	0	1	0	0	0
Chiclets	0	1	0	0	0
Dots	0	1	0	0	0
Dum Dums	0	1	0	0	0
Fruit Chews	0	1	0	0	0
Fun Dip	0	1	0	0	0
Gobstopper	0	1	0	0	0
Haribo Gold Bears	0	1	0	0	0
Haribo Sour Bears	0	1	0	0	0
Haribo Twin Snakes	0	1	0	0	0
Jawbusters	0	1	0	0	0
Laffy Taffy	0	1	0	0	0
Lemonhead	0	1	0	0	0
Lifesavers big ring gummies	0	1	0	0	0
Mike & Ike	0	1	0	0	0
Nerds	0	1	0	0	0
Nik L Nip	0	1	0	0	0
Now & Later	0	1	0	0	0
Pop Rocks	0	1	0	0	0
Red vines	0	1	0	0	0
Ring pop	0	1	0	0	0
Runts	0	1	0	0	0
Skittles original	0	1	0	0	0
Skittles wildberry	0	1	0	0	0
Smarties candy	0	1	0	0	0
Sour Patch Kids	0	1	0	0	0
Sour Patch Tricksters	0	1	0	0	0
Starburst	0	1	0	0	0
Strawberry bon bons	0	1	0	0	0
Super Bubble	0	1	0	0	0
Swedish Fish	0	1	0	0	0
Tootsie Pop	1	1	0	0	0
Trolli Sour Bites	0	1	0	0	0

Twizzlers	0	1	0	0	0
Warheads	0	1	0	0	0
Welch's Fruit Snacks	0	1	0	0	0

	crisped	ricewafer	hard	bar	pluribus	sugarpercent
Air Heads	0	0	0	0		0.906
Caramel Apple Pops	0	0	0	0		0.604
Chewey Lemonhead Fruit Mix	0	0	0	1		0.732
Chiclets	0	0	0	1		0.046
Dots	0	0	0	1		0.732
Dum Dums	0	1	0	0		0.732
Fruit Chews	0	0	0	1		0.127
Fun Dip	0	1	0	0		0.732
Gobstopper	0	1	0	1		0.906
Haribo Gold Bears	0	0	0	1		0.465
Haribo Sour Bears	0	0	0	1		0.465
Haribo Twin Snakes	0	0	0	1		0.465
Jawbusters	0	1	0	1		0.093
Laffy Taffy	0	0	0	0		0.220
Lemonhead	0	1	0	0		0.046
Lifesavers big ring gummies	0	0	0	0		0.267
Mike & Ike	0	0	0	1		0.872
Nerds	0	1	0	1		0.848
Nik L Nip	0	0	0	1		0.197
Now & Later	0	0	0	1		0.220
Pop Rocks	0	1	0	1		0.604
Red vines	0	0	0	1		0.581
Ring pop	0	1	0	0		0.732
Runts	0	1	0	1		0.872
Skittles original	0	0	0	1		0.941
Skittles wildberry	0	0	0	1		0.941
Smarties candy	0	1	0	1		0.267
Sour Patch Kids	0	0	0	1		0.069
Sour Patch Tricksters	0	0	0	1		0.069
Starburst	0	0	0	1		0.151
Strawberry bon bons	0	1	0	1		0.569
Super Bubble	0	0	0	0		0.162
Swedish Fish	0	0	0	1		0.604
Tootsie Pop	0	1	0	0		0.604
Trolli Sour Bites	0	0	0	1		0.313
Twizzlers	0	0	0	0		0.220
Warheads	0	1	0	0		0.093
Welch's Fruit Snacks	0	0	0	1		0.313

	pricepercent	winpercent

Air Heads	0.511	52.34146
Caramel Apple Pops	0.325	34.51768
Chewey Lemonhead Fruit Mix	0.511	36.01763
Chiclets	0.325	24.52499
Dots	0.511	42.27208
Dum Dums	0.034	39.46056
Fruit Chews	0.034	43.08892
Fun Dip	0.325	39.18550
Gobstopper	0.453	46.78335
Haribo Gold Bears	0.465	57.11974
Haribo Sour Bears	0.465	51.41243
Haribo Twin Snakes	0.465	42.17877
Jawbusters	0.511	28.12744
Laffy Taffy	0.116	41.38956
Lemonhead	0.104	39.14106
Lifesavers big ring gummies	0.279	52.91139
Mike & Ike	0.325	46.41172
Nerds	0.325	55.35405
Nik L Nip	0.976	22.44534
Now & Later	0.325	39.44680
Pop Rocks	0.837	41.26551
Red vines	0.116	37.34852
Ring pop	0.965	35.29076
Runts	0.279	42.84914
Skittles original	0.220	63.08514
Skittles wildberry	0.220	55.10370
Smarties candy	0.116	45.99583
Sour Patch Kids	0.116	59.86400
Sour Patch Tricksters	0.116	52.82595
Starburst	0.220	67.03763
Strawberry bon bons	0.058	34.57899
Super Bubble	0.116	27.30386
Swedish Fish	0.755	54.86111
Tootsie Pop	0.325	48.98265
Trolli Sour Bites	0.255	47.17323
Twizzlers	0.116	45.46628
Warheads	0.116	39.01190
Welch's Fruit Snacks	0.313	44.37552

How often does my favorite candy win?

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

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

```
[1] 76.67378
```

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
```

```
skimr::skim(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	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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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` column is on a 0:100 scale as compared to others with a 0:1 scale.

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

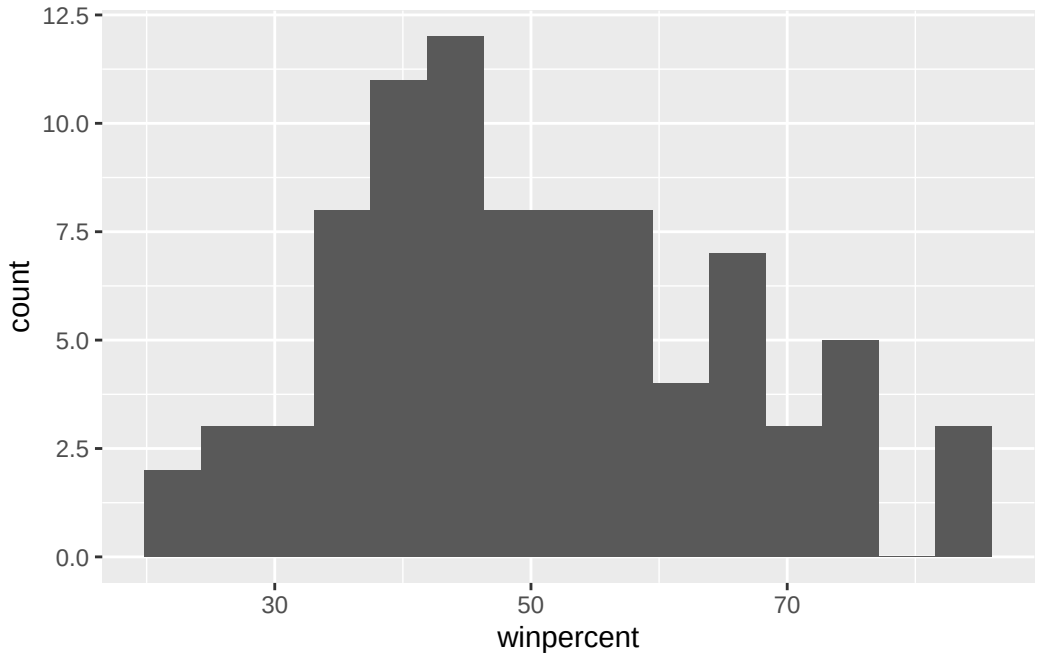
A zero would represent that the candy does not contain chocolate while a one would represent that the candy contains chocolate.

Q8. Plot a histogram of winpercent values

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.2.3
```

```
ggplot(candy, aes(x = winpercent)) +
  geom_histogram(bins = 15)
```



Q9. Is the distribution of winpercent values symmetrical?

No

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

Below 50% with a mean

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

```
[1] 50.31676
```

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

To answer this question I will need to:

- "Subset" the candy dataset to just chocolate candy,
- get their winpercent values,
- and then calculate the mean of these.

Then do the same for fruity candy and compare.

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

```
print(chocolate_avg)
```

```
[1] 60.92153
```

```
print(fruity_avg)
```

```
[1] 44.11974
```

On average, chocolate candy is ranked higher than fruity candy.

Q12. Is this difference statistically significant?

```
t.test((candy$winpercent[as.logical(candy$chocolate)]), candy$winpercent[as.logical(candy$
```

```
    Welch Two Sample t-test

data:  (candy$winpercent[as.logical(candy$chocolate)]) and candy$winpercent[as.logical(candy$
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
```

Overall Candy Rankings

There is a base R function called sort() for sorting vectors of input.

```
x <- c(5,2,10)

# sort(x, decreasing = TRUE)
sort(x)
```

```
[1]  2  5 10
```

The buddy function to `sort()` that is often more useful is called `order()`. It returns the "indices" of the input that would result in it being sorted.

```
order(x)
```

```
[1] 2 1 3
```

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

```
library(tidyverse)
```

```
Warning: package 'tibble' was built under R version 4.2.3
```

```
Warning: package 'dplyr' was built under R version 4.2.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v lubridate 1.9.2     v tibble    3.2.1
v purrr     1.0.1     v tidyr     1.3.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
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
```

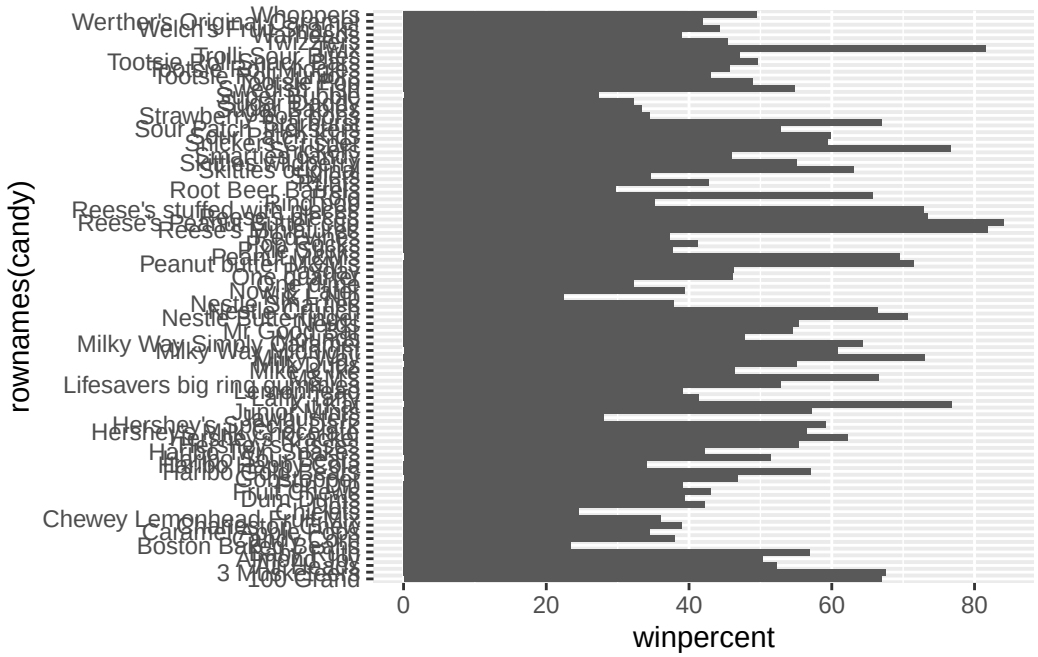
Q14. What are the top 5 all time favorite candy types out of this set?

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

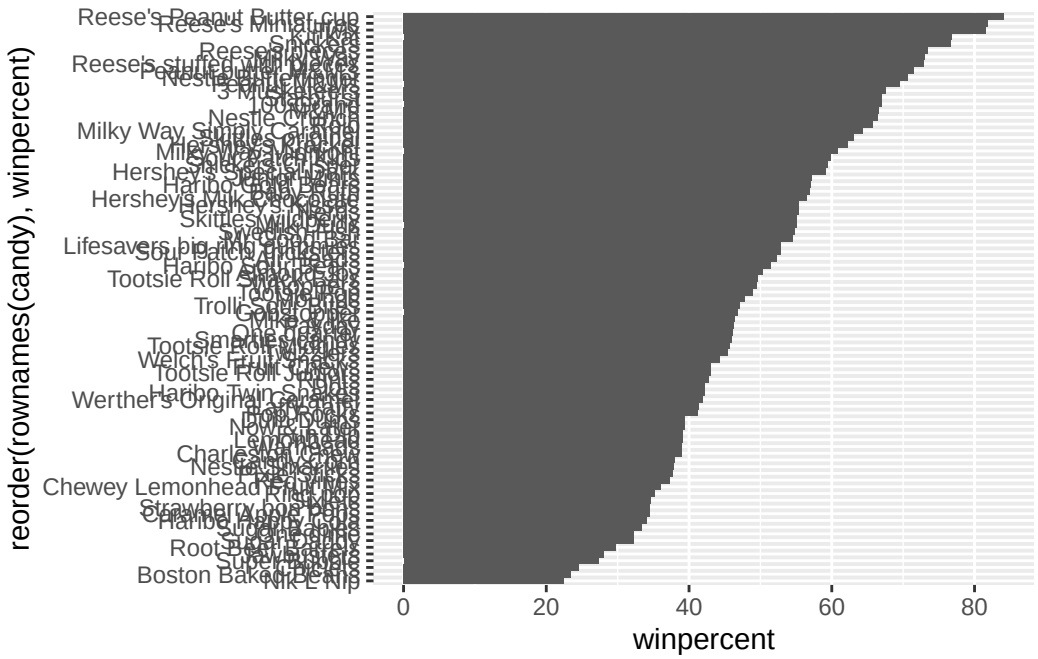
Q15. Make a barplot of candy ranking using `winpercent` values.

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



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

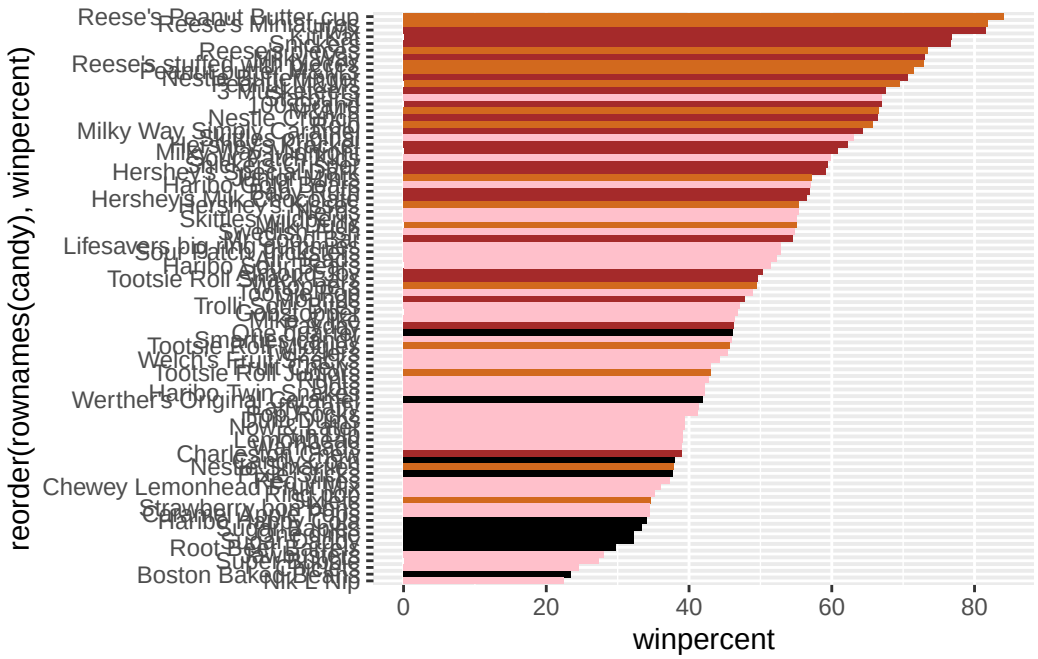
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col()
```



Adding colors to these bars

```
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)
```



Q17. What is the worst ranked chocolate candy?

Sixlets

Q18. What is the best ranked fruity candy?

Starburst

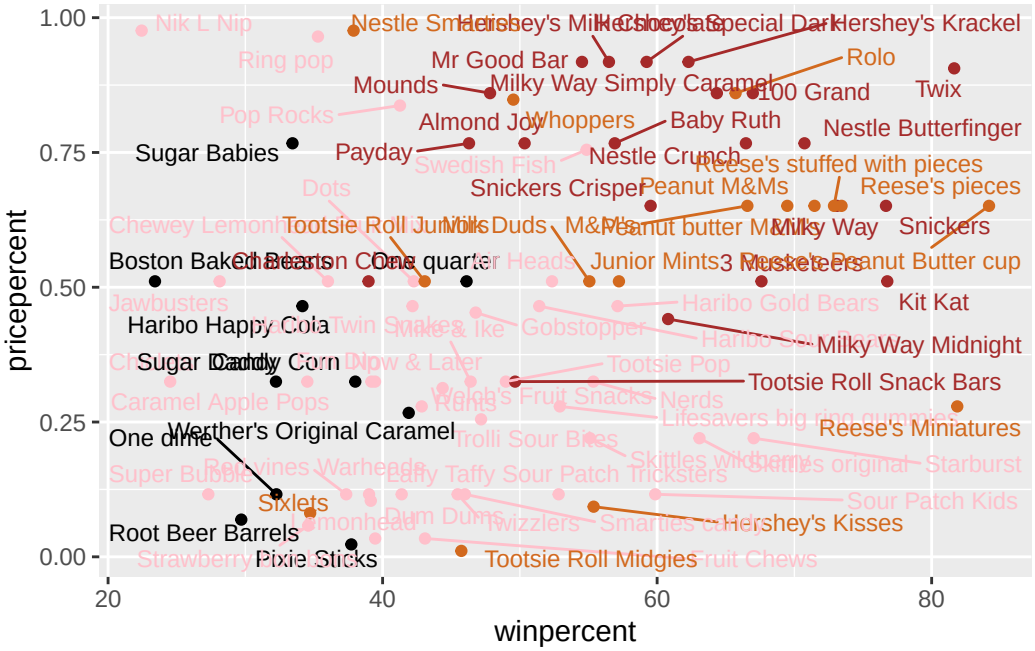
Taking a look at pricepoint

Q. What is the best candy for the least money?

```
library(ggrepel)
```

```
Warning: package 'ggrepel' was built under R version 4.2.3
```

```
# How about a plot of price vs win
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  # ggrepel helps deal with overlapping labels
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 30)
```



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?

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
```

Nik L Nip

Exploring the correlation structure

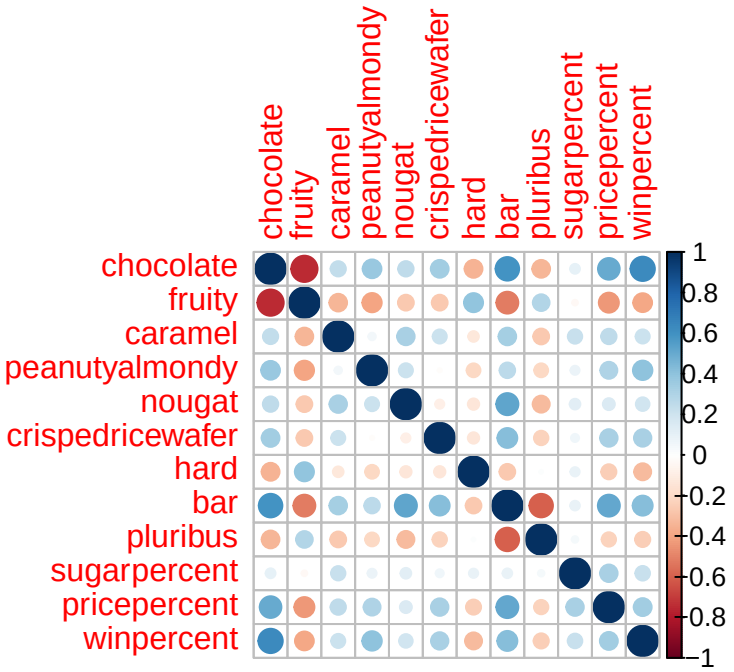
Pearson correlation goes between -‘ and +1 with zero indicating no correlation. Values close to one indicate being very highly correlated.

```
library(corrplot)
```

```
Warning: package 'corrplot' was built under R version 4.2.3
```

```
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 ani-correlated.

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

Chocolate and winpercent or chocolate and bar.

Principal Component Analysis

The base r function for PCA is called `prcomp()` and we can set "scale=TRUE/FALSE".

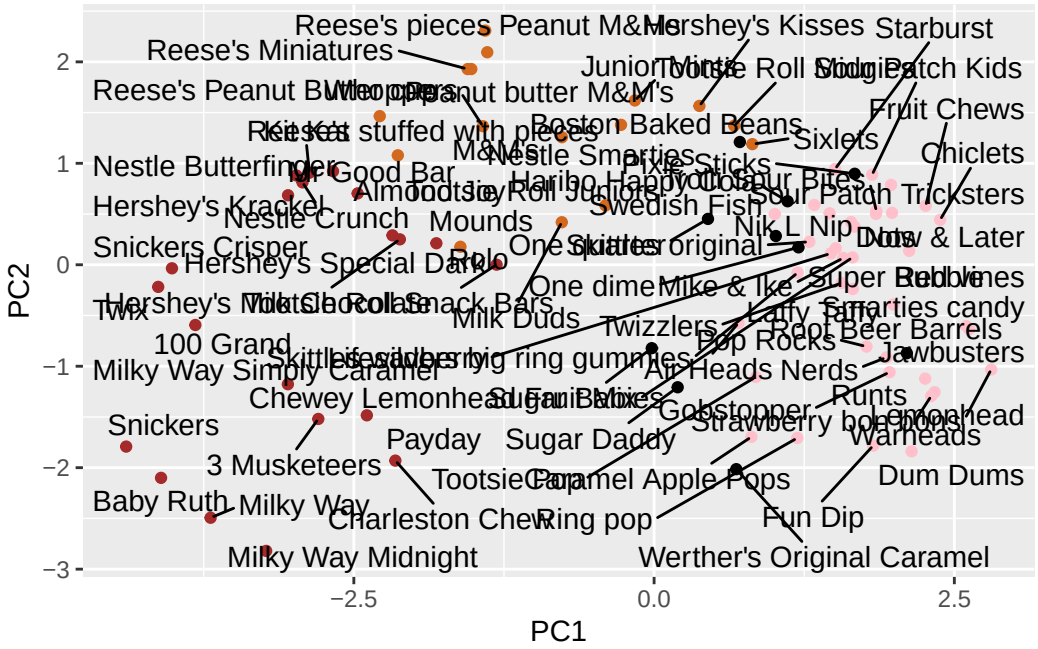
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
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
```

The main result of PCA - i.e. the new PC plot (projection of candy on our new PC axis) is contained in `pca$x`

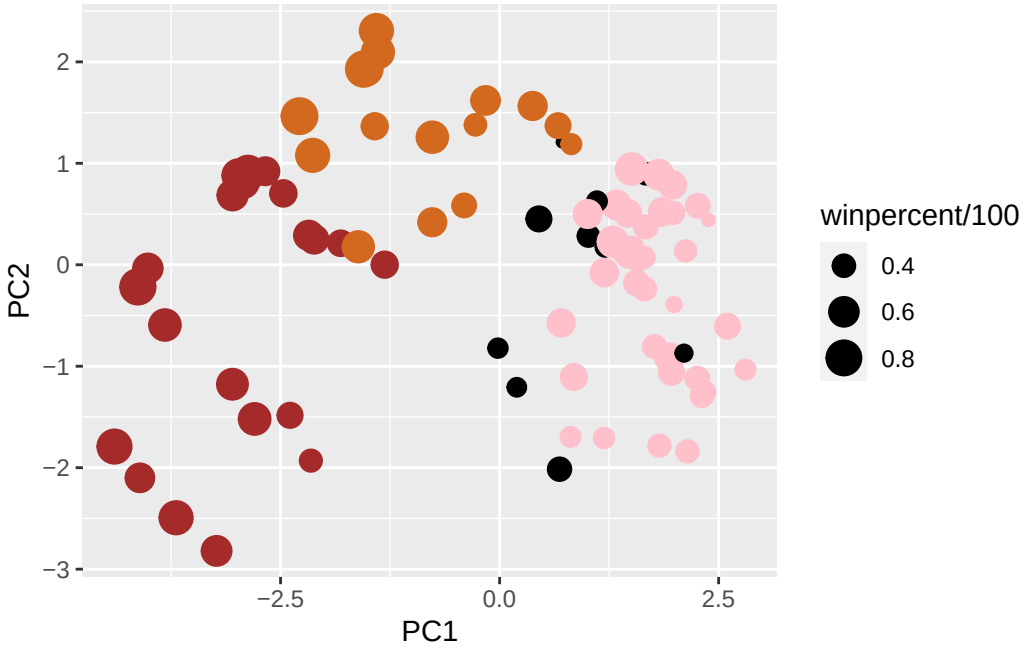
```
pc <- as.data.frame(pca$x)

ggplot(pc) +
  aes(PC1, PC2, label = rownames(pc)) +
  geom_point(col = my_cols) +
  geom_text_repel(max.overlaps  = 30)
```

```
Warning: ggrepel: 5 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



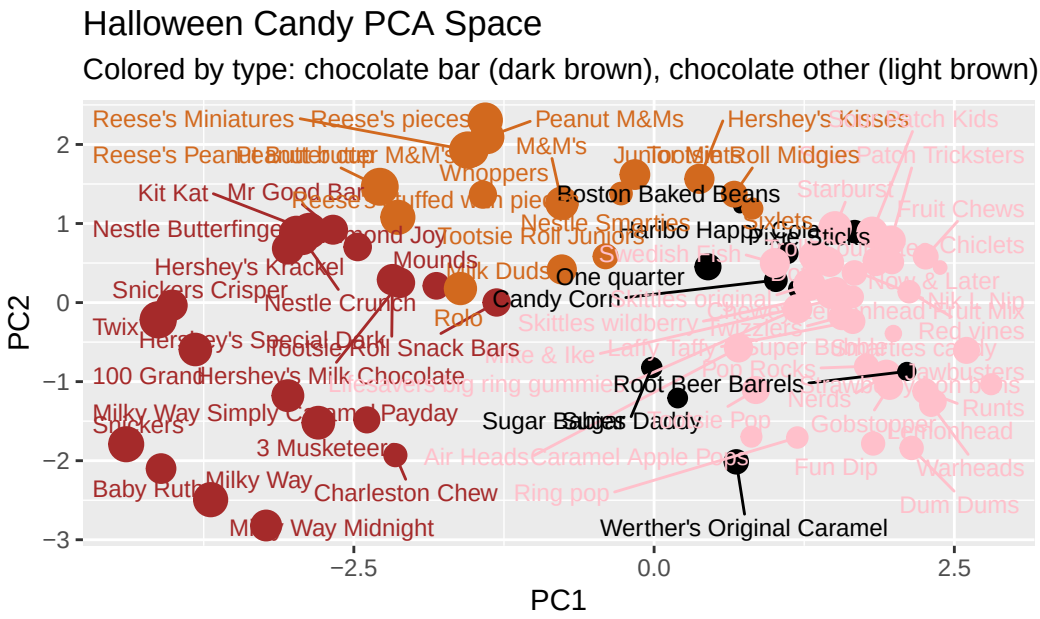

```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
```

```
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 = 30)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown
       caption="Data from 538")
```

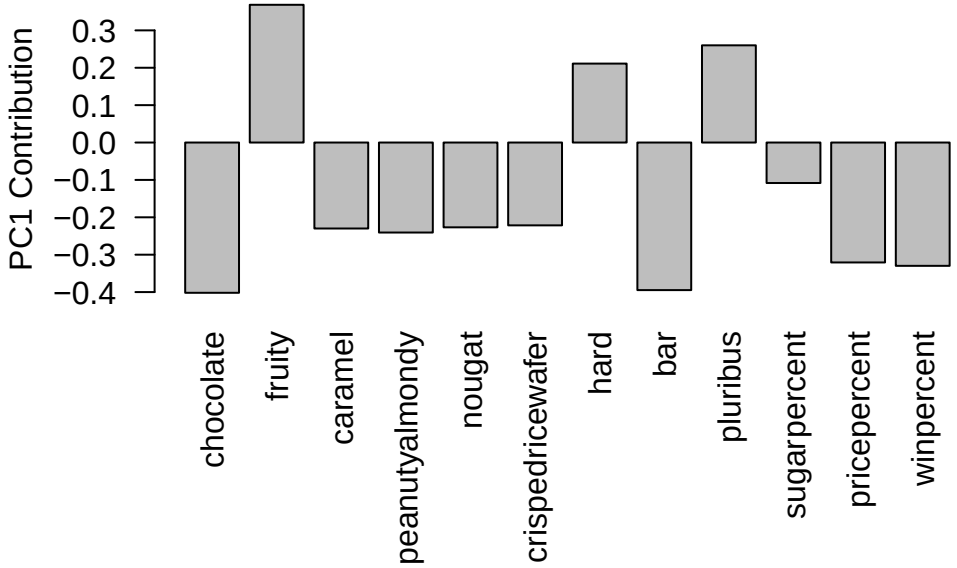
```
Warning: ggrepel: 5 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Data from 538

```
#library(plotly)
#ggplotly(p)
```

```
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. These make sense these variables are strongly correlated, as seen in our correlation structure.