October 27, 2023 Class 08 Halloween Candy Project

Savannah Bogus A69027475

Importing Candy Data

```
read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-rankin
```

	competitorname	chocolate	fruity	caramel	peanutyalmondy	nougat
1	100 Grand	1	0	1	0	0
2	3 Musketeers	1	0	0	0	1
3	One dime	0	0	0	0	0
4	One quarter	0	0	0	0	0
5	Air Heads	0	1	0	0	0
6	Almond Joy	1	0	0	1	0
7	Baby Ruth	1	0	1	1	1
8	Boston Baked Beans	0	0	0	1	0
9	Candy Corn	0	0	0	0	0
10	Caramel Apple Pops	0	1	1	0	0
11	Charleston Chew	1	0	0	0	1
12	Chewey Lemonhead Fruit Mix	0	1	0	0	0
13	Chiclets	0	1	0	0	0
14	Dots	0	1	0	0	0
15	Dum Dums	0	1	0	0	0
16	Fruit Chews	0	1	0	0	0
17	Fun Dip	0	1	0	0	0
18	Gobstopper	0	1	0	0	0
19	Haribo Gold Bears	0	1	0	0	0
20	Haribo Happy Cola	0	0	0	0	0
21	Haribo Sour Bears	0	1	0	0	0
22	Haribo Twin Snakes	0	1	0	0	0
23	Hershey's Kisses	1	0	0	0	0

24	Hershey's Krackel	1	0	0	0	0
25	Hershey's Milk Chocolate	1	0	0	0	0
26	Hershey's Special Dark	1	0	0	0	0
27	Jawbusters	0	1	0	0	0
28	Junior Mints	1	0	0	0	0
29	Kit Kat	1	0	0	0	0
30	Laffy Taffy	0	1	0	0	0
31	Lemonhead	0	1	0	0	0
32	Lifesavers big ring gummies	0	1	0	0	0
33	Peanut butter M&M's	1	0	0	1	0
34	M&M's	1	0	0	0	0
35	Mike & Ike	0	1	0	0	0
36	Milk Duds	1	0	1	0	0
37	Milky Way	1	0	1	0	1
38	Milky Way Midnight	1	0	1	0	1
39	Milky Way Simply Caramel	1	0	1	0	0
40	Mounds	1	0	0	0	0
41	Mr Good Bar	1	0	0	1	0
42	Nerds	0	1	0	0	0
43	Nestle Butterfinger	1	0	0	1	0
44	Nestle Crunch	1	0	0	0	0
45	Nik L Nip	0	1	0	0	0
46	Now & Later	0	1	0	0	0
47	Payday	0	0	0	1	1
48	Peanut M&Ms	1	0	0	1	0
49	Pixie Sticks	0	0	0	0	0
50	Pop Rocks	0	1	0	0	0
51	Red vines	0	1	0	0	0
52	Reese's Miniatures	1	0	0	1	0
53	Reese's Peanut Butter cup	1	0	0	1	0
54	Reese's pieces	1	0	0	1	0
55	Reese's stuffed with pieces	1	0	0	1	0
56	Ring pop	0	1	0	0	0
57	Rolo	1	0	1	0	0
58	Root Beer Barrels	0	0	0	0	0
59	Runts	0	1	0	0	0
60	Sixlets	1	0	0	0	0
61	Skittles original	0	1	0	0	0
62	Skittles wildberry	0	1	0	0	0
63	Nestle Smarties	1	0	0	0	0
64	Smarties candy	0	1	0	0	0
65	Snickers	1	0	1	1	1
66	Snickers Crisper	1	0	1	1	0

67	Sour Patch Kids	0	1	0	0	0	
68	Sour Patch Tricksters	0	1	0	0	0	
69	Starburst	0	1	0	0	0	
70	Strawberry bon bons	0	1	0	0	0	
71	Sugar Babies	0	0	1	0	0	
72	Sugar Daddy	0	0	1	0	0	
73	Super Bubble	0	1	0	0	0	
74	Swedish Fish	0	1	0	0	0	
75	Tootsie Pop	1	1	0	0	0	
76	Tootsie Roll Juniors	1	0	0	0	0	
77	Tootsie Roll Midgies	1	0	0	0	0	
78	Tootsie Roll Snack Bars	1	0	0	0	0	
79	Trolli Sour Bites	0	1	0	0	0	
80	Twix	1	0	1	0	0	
81	Twizzlers	0	1	0	0	0	
82	Warheads	0	1	0	0	0	
83	Welch's Fruit Snacks	0	1	0	0	0	
84	Werther's Original Caramel	0	0	1	0	0	
85	Whoppers	1	0	0	0	0	

	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent
1	1	0	1	0	0.732	0.860	66.97173
2	0	0	1	0	0.604	0.511	67.60294
3	0	0	0	0	0.011	0.116	32.26109
4	0	0	0	0	0.011	0.511	46.11650
5	0	0	0	0	0.906	0.511	52.34146
6	0	0	1	0	0.465	0.767	50.34755
7	0	0	1	0	0.604	0.767	56.91455
8	0	0	0	1	0.313	0.511	23.41782
9	0	0	0	1	0.906	0.325	38.01096
10	0	0	0	0	0.604	0.325	34.51768
11	0	0	1	0	0.604	0.511	38.97504
12	0	0	0	1	0.732	0.511	36.01763
13	0	0	0	1	0.046	0.325	24.52499
14	0	0	0	1	0.732	0.511	42.27208
15	0	1	0	0	0.732	0.034	39.46056
16	0	0	0	1	0.127	0.034	43.08892
17	0	1	0	0	0.732	0.325	39.18550
18	0	1	0	1	0.906	0.453	46.78335
19	0	0	0	1	0.465	0.465	57.11974
20	0	0	0	1	0.465	0.465	34.15896
21	0	0	0	1	0.465	0.465	51.41243
22	0	0	0	1	0.465	0.465	42.17877
23	0	0	0	1	0.127	0.093	55.37545

24	1	0	1	0	0.430	0.918	62.28448
25	0	0	1	0	0.430	0.918	56.49050
26	0	0	1	0	0.430	0.918	59.23612
27	0	1	0	1	0.093	0.511	28.12744
28	0	0	0	1	0.197	0.511	57.21925
29	1	0	1	0	0.313	0.511	76.76860
30	0	0	0	0	0.220	0.116	41.38956
31	0	1	0	0	0.046	0.104	39.14106
32	0	0	0	0	0.267	0.279	52.91139
33	0	0	0	1	0.825	0.651	71.46505
34	0	0	0	1	0.825	0.651	66.57458
35	0	0	0	1	0.872	0.325	46.41172
36	0	0	0	1	0.302	0.511	55.06407
37	0	0	1	0	0.604	0.651	73.09956
38	0	0	1	0	0.732	0.441	60.80070
39	0	0	1	0	0.965	0.860	64.35334
40	0	0	1	0	0.313	0.860	47.82975
41	0	0	1	0	0.313	0.918	54.52645
42	0	1	0	1	0.848	0.325	55.35405
43	0	0	1	0	0.604	0.767	70.73564
44	1	0	1	0	0.313	0.767	66.47068
45	0	0	0	1	0.197	0.976	22.44534
46	0	0	0	1	0.220	0.325	39.44680
47	0	0	1	0	0.465	0.767	46.29660
48	0	0	0	1	0.593	0.651	69.48379
49	0	0	0	1	0.093	0.023	37.72234
50	0	1	0	1	0.604	0.837	41.26551
51	0	0	0	1	0.581	0.116	37.34852
52	0	0	0	0	0.034	0.279	81.86626
53	0	0	0	0	0.720	0.651	84.18029
54	0	0	0	1	0.406	0.651	73.43499
55	0	0	0	0	0.988	0.651	72.88790
56	0	1	0	0	0.732	0.965	35.29076
57	0	0	0	1	0.860	0.860	65.71629
58	0	1	0	1	0.732	0.069	29.70369
59	0	1	0	1	0.872	0.279	42.84914
60	0	0	0	1	0.220	0.081	34.72200
61	0	0	0	1	0.941	0.220	63.08514
62	0	0	0	1	0.941	0.220	55.10370
63	0	0	0	1	0.267	0.976	37.88719
64	0	1	0	1	0.267	0.116	45.99583
65	0	0	1	0	0.546	0.651	76.67378
66	1	0	1	0	0.604	0.651	59.52925

67	0	0	0	1	0.069	0.116	59.86400
68	0	0	0	1	0.069	0.116	52.82595
69	0	0	0	1	0.151	0.220	67.03763
70	0	1	0	1	0.569	0.058	34.57899
71	0	0	0	1	0.965	0.767	33.43755
72	0	0	0	0	0.418	0.325	32.23100
73	0	0	0	0	0.162	0.116	27.30386
74	0	0	0	1	0.604	0.755	54.86111
75	0	1	0	0	0.604	0.325	48.98265
76	0	0	0	0	0.313	0.511	43.06890
77	0	0	0	1	0.174	0.011	45.73675
78	0	0	1	0	0.465	0.325	49.65350
79	0	0	0	1	0.313	0.255	47.17323
80	1	0	1	0	0.546	0.906	81.64291
81	0	0	0	0	0.220	0.116	45.46628
82	0	1	0	0	0.093	0.116	39.01190
83	0	0	0	1	0.313	0.313	44.37552
84	0	1	0	0	0.186	0.267	41.90431
85	1	0	0	1	0.872	0.848	49.52411

```
candy_file<-"candy_data.csv"
candy=read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-
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
```

```
dim(candy)
```

```
[1] 85 12
```

There are 85 candies in this dataset.

```
table(candy$fruity)
```

```
 0  1
47 38
```

38 candies have a 1 in fruity, so 38 candies are fruity.

Favorite candy

Q3

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

```
[1] 76.67378
```

The win percent value is 76.67%, which is wayyyy too low in my opinion for snickers. I cannot BELIEVE Twix is higher. Snickers is like elevated Twix.

Q4 and Q5

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

```
[1] 76.7686
```

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

```
[1] 49.6535
```

Kit Kat=76.76 (HIGHER THAN SNICKERS? REALLY?) Tootsie Roll Snack Bars=49.65

```
library("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	
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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
winpercent	0	1	50.32	14.71	22.45	39.14	47.83	59.86	84.18	

Q6

win percent is on a different scale to the other columns, I believe.

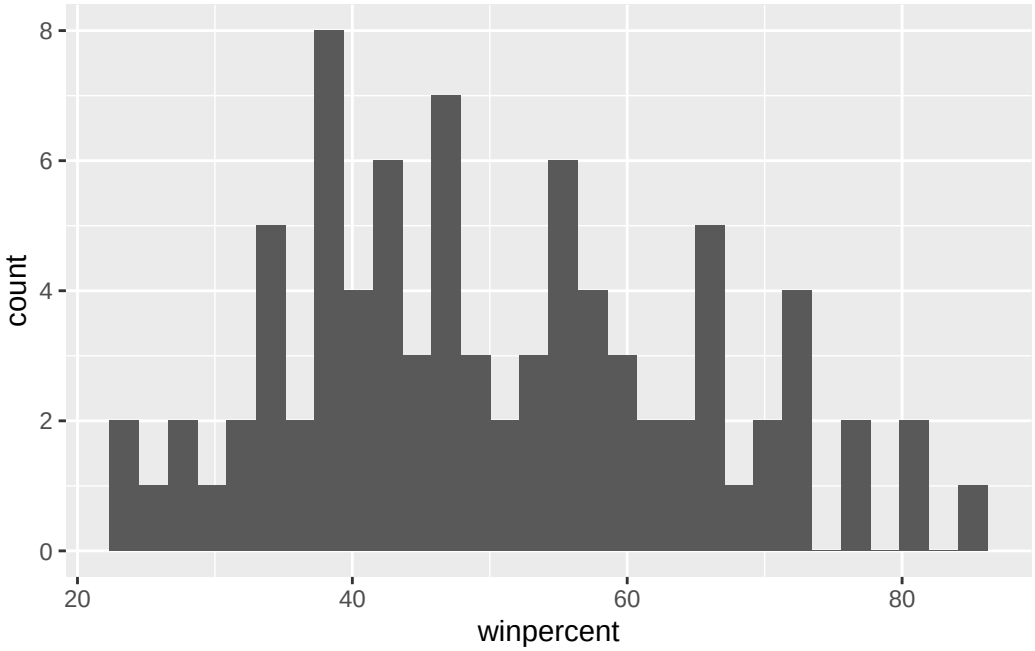
Q7

0=not chocolatey, 1=chocolatey. ie. Hershey's would have a 1 and Starbursts would have a 0.

Q8

```
library(ggplot2)
ggplot(candy)+
       aes(winpercent)+
    geom_histogram()
```

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



Q9

The distribution is not symmetrical.

Q10

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

```
[1] 50.31676
```

The center of the distribution is just above 50%.

Q11

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

```
[1] 60.92153
```

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

```
[1] 44.11974
```

The mean win percent for chocolate candy is higher than the mean win percent for fruity candy.

Q11

```
choc<-as.logical(candy$chocolate)
fruity<-as.logical(candy$fruity)

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

```
    Welch Two Sample t-test

data:  candy$winpercent[as.logical(candy$chocolate)] and candy$winpercent[as.logical(candy$fr
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, the difference is statistically significant

Overall Candy Rankings

```
order(candy$winpercent)
```

```
 [1] 45  8 13 73 27 58 72  3 71 20 10 70 60 56 12 51 49 63  9 11 82 31 17 46 15
[26] 50 30 84 22 14 59 76 16 83 81 77 64  4 47 35 18 79 40 75 85 78  6 21  5 68
[51] 32 41 74 36 62 42 23 25  7 19 28 26 66 67 38 24 61 39 57 44 34  1 69  2 48
[76] 43 33 55 37 54 65 29 80 52 53
```

```
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
```

```
tail(candy[order(candy$winpercent),],n=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
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
```

Q13

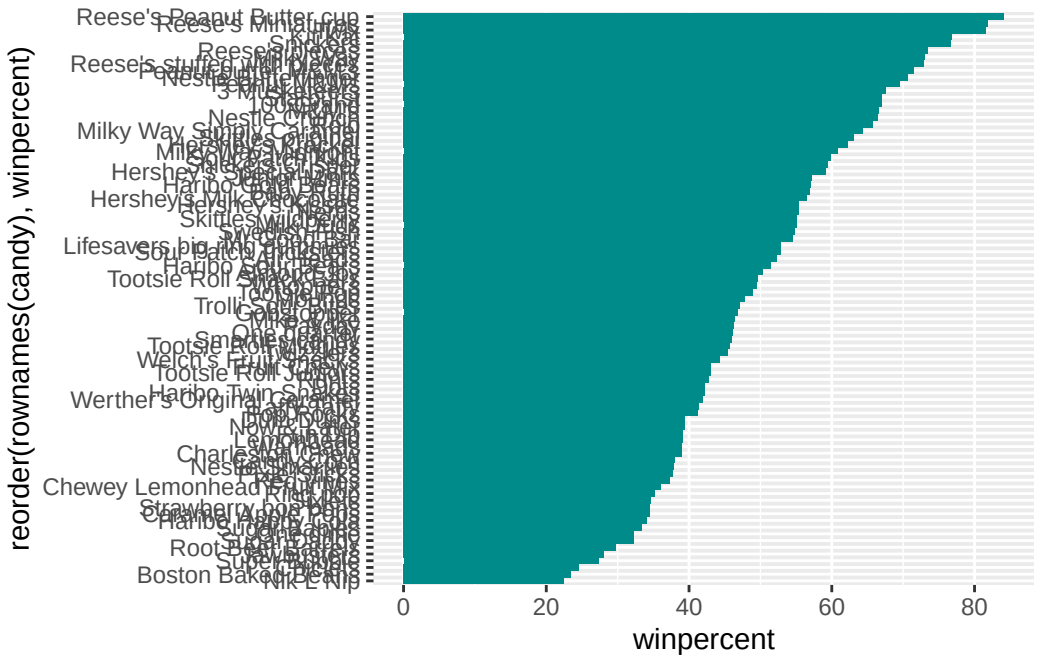
Least liked 5 candies are Nik L Nip (what on earth is that), boston baked beans (tell me this isn't actually baked beans and that it's a candy), chiclets, super bubble, and jawbusters.

Q14

Top 5 candies are Reese's Peanut Butter cup, Reese's Miniatures (1, the miniatures ARE worse. 2, this is the same candy), Twix, Kit Kat, Snickers.

Q15

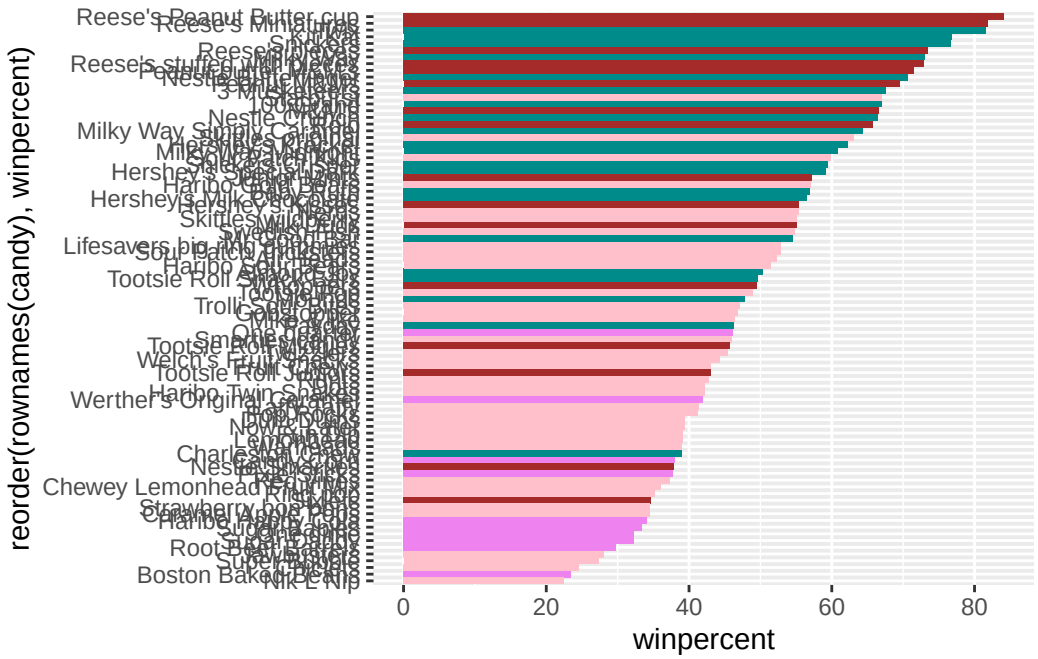
```
ggplot(candy)+
  aes(winpercent, reorder(rownames(candy),winpercent))+
  geom_bar(stat="identity",fill="darkcyan")
```



Trying again with new colors that you prefer. (Except we're still using my faves)

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

ggplot(candy)+
  aes(winpercent, reorder(rownames(candy),winpercent))+
  geom_bar(stat="identity",fill=my_colors)
```



I'm watching you make this graph in class right now (I'm on Q23 in real time) and it looks like I did mine differently. But, I guess it still works.

Q17 and Q18

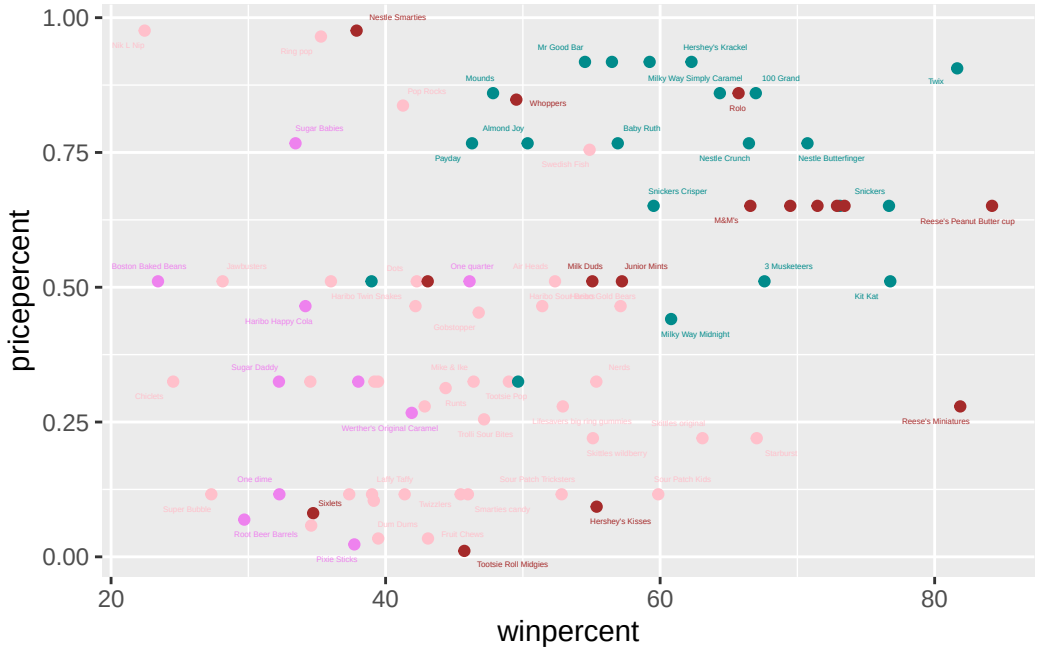
Worst ranked chocolate candy is sixlets. The best ranked fruity candy is starbursts.

Taking a look at pricepercent

```
library(ggrepel)

ggplot(candy)+
  aes(winpercent,pricepercent,label=rownames(candy))+
  geom_point(col=my_colors)+
  geom_text_repel(col=my_colors,size=1,max.overlaps=5)
```

```
Warning: ggrepel: 20 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Q19

```
order(candy$pricepercent)
```

```
 [1] 77 49 15 16 70 58 60 23 31  3 30 51 64 67 68 73 81 82 61 62 69 79 84 32 52
[26] 59 83  9 10 13 17 35 42 46 72 75 78 38 18 19 20 21 22  2  4  5  8 11 12 14
[51] 27 28 29 36 76 33 34 37 48 53 54 55 65 66 74  6  7 43 44 47 71 50 85  1 39
[76] 40 57 80 24 25 26 41 56 45 63
```

```
cheap<-order(candy$pricepercent,decreasing=TRUE)
head(candy[cheap,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
```

```
tail(candy[cheap,c(11,12)],n=5)
```

```
                    pricepercent winpercent
Strawberry bon bons        0.058   34.57899
Dum Dums                   0.034   39.46056
Fruit Chews                0.034   43.08892
Pixie Sticks               0.023   37.72234
Tootsie Roll Midgies       0.011   45.73675
```

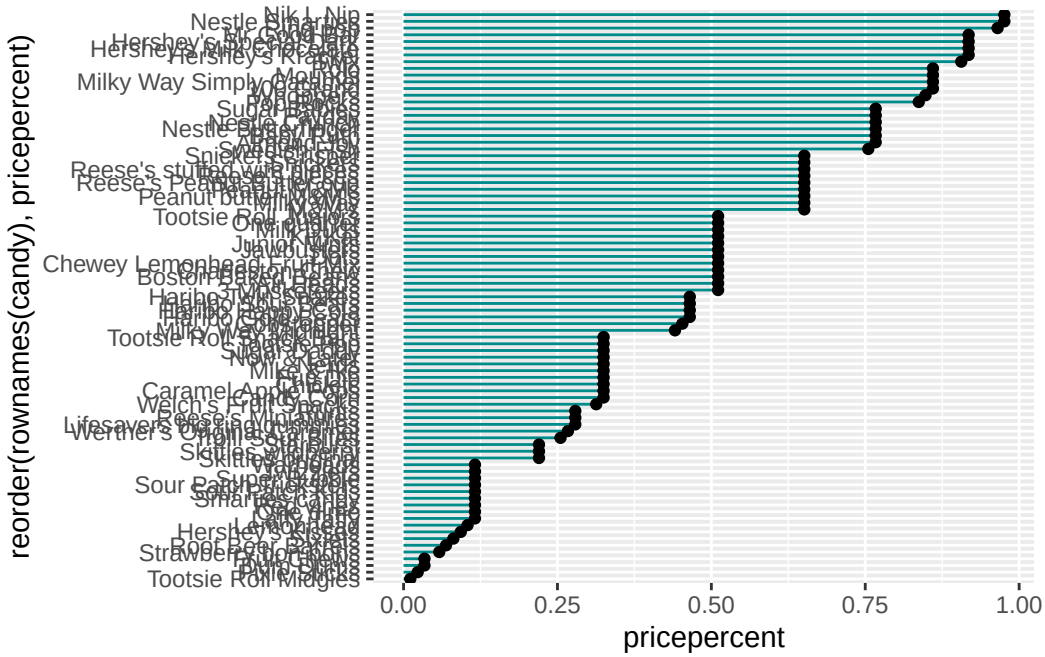
Reese's miniatures (not as good as fullsized) is ranked very highly and is very cheap, by eye, but out of the 5 cheapest candies, Tootsie roll midgies (what are these) are the best and cheapest.

Q20

Why is Nik L Nip expensive but also garbage. That's the least popular. Ring pops, Smarties, Pop Rocks, and Sugar Babies are also very very pricey, and apparently not all that great. (Except Ring pops are fantastic, or so they should be ranked)

Q21

```
ggplot(candy)+
  aes(pricepercent,reorder(rownames(candy),pricepercent))+
  geom_segment(aes(xend=0,yend=reorder(rownames(candy),pricepercent)),col="darkcyan")+
  geom_point()
```

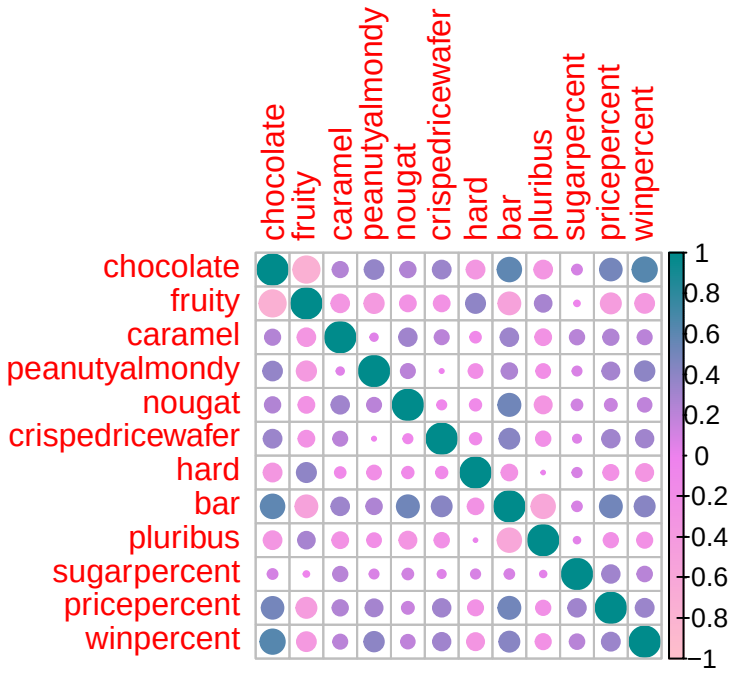


Exploring correlation structure

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

```
cij<-cor(candy)
corrplot(cij,col=colorRampPalette(c("pink","violet","darkcyan"))(200))
```



I spent a good 7 minutes figuring out how to implement my color scheme. I had never heard of the `colorRampPalette()` function.

Q22

Chocolate and fruity and pluribus and bars are negatively correlated. It's weird because I LOVE fruity chocolate (chocolate covered strawberries?? cmon) but I guess it's not very good in candy.

Q23

Winpercent and chocolate are positively correlated :(as well as chocolate and bar.

Principal Component Analysis

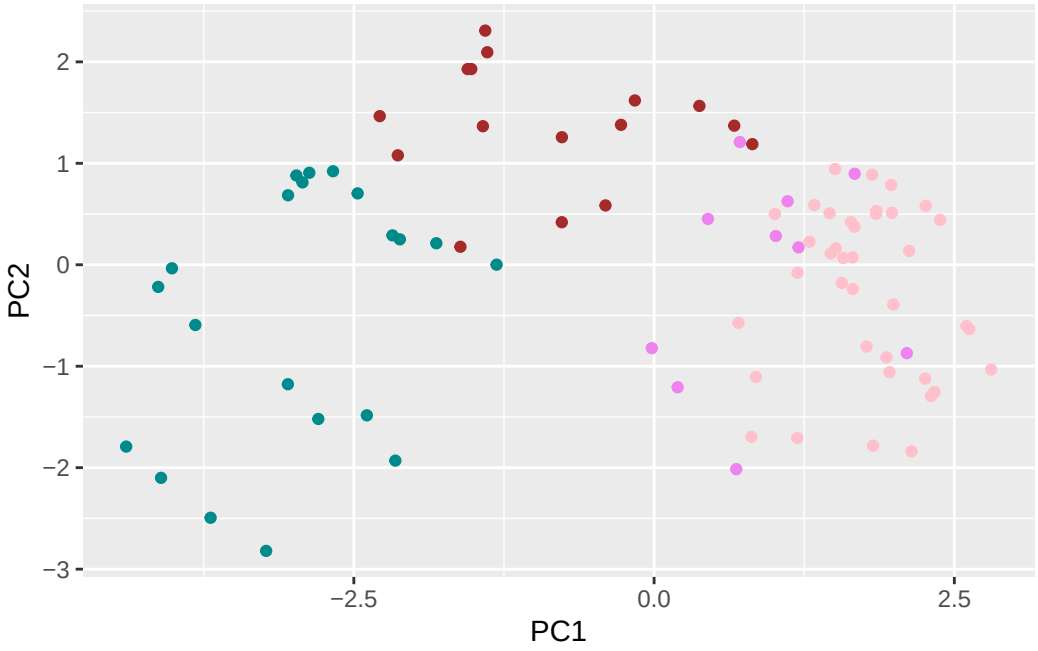
```
pca<-prcomp(candy,scale=TRUE)
summary(pca)
```

```
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
```

```
pca$rotation[,1]
```

```
     chocolate           fruity          caramel   peanutyalmondy
    -0.4019466        0.3683883       -0.2299709       -0.2407155
        nougat crispedricewafer             hard              bar
    -0.2268102       -0.2215182        0.2111587       -0.3947433
      pluribus     sugarpercent     pricepercent       winpercent
     0.2600041       -0.1083088       -0.3207361       -0.3298035
```

```
df_candy<-as.data.frame(pca$x)
ggplot(df_candy)+
  aes(PC1,PC2)+
  geom_point(col=my_colors)
```

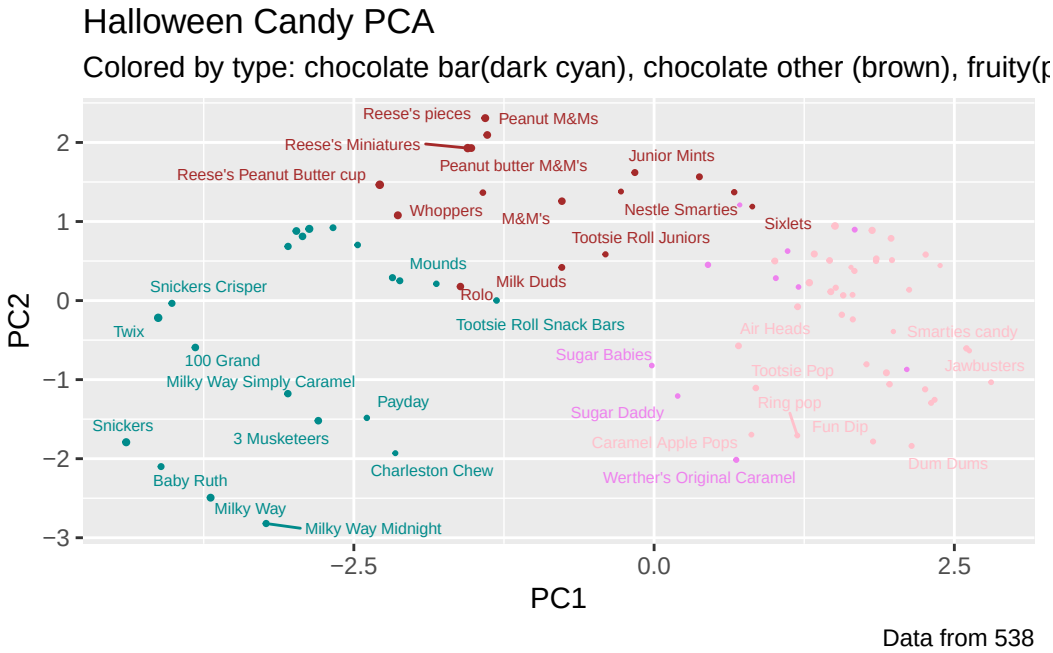


I realize now that you wanted us to use a base R plot, but I started working in ggplot right away because usually you say "use base R OR use this package" so I wanted to start using the package right away. Please don't take off points.

Now, we're going to use ggrepel.

```
ggplot(df_candy)+
  aes(PC1,PC2)+
  geom_point(col=my_colors,size=candy$winpercent/100)+
  geom_text_repel(col=my_colors, size=2, max.overlaps=7,label=rownames(candy))+
  theme(legend.position="none")+
  labs(title="Halloween Candy PCA",
       subtitle="Colored by type: chocolate bar(dark cyan), chocolate other (brown), fruit
       caption="Data from 538")
```

```
Warning: ggrepel: 48 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

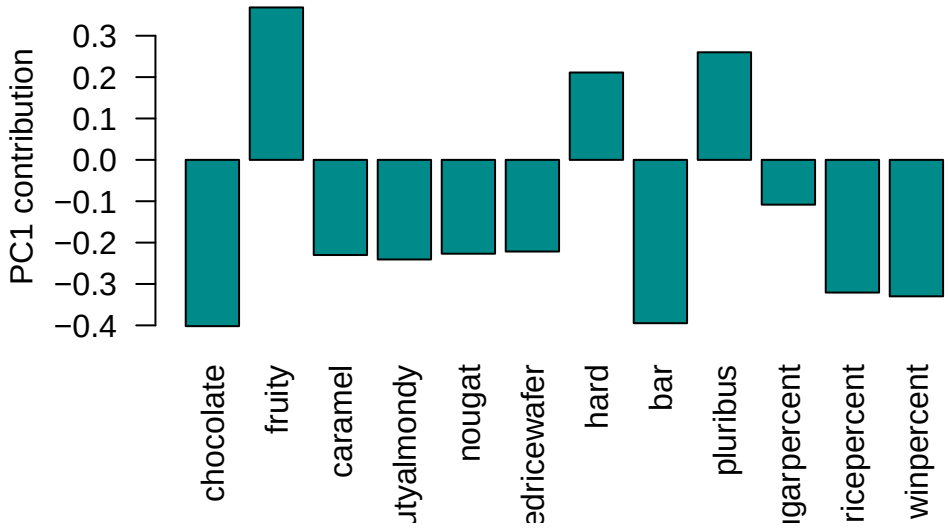


The interactive labeling seems like a great idea. I'm going to try that. However, when I try to render it, it doesn't work because I'm rendering a PDF. I'll just put my code in as the backtick code chunk and not as code to be read.

‘ library(plotly)

ggplotly(ggplot(df_candy)+ aes(PC1,PC2)+ geom_point(col=my_colors,label=rownames(candy))+ geom_text_repel(col=my_colors, size=2, max.overlaps=7,label=rownames(candy))+ theme(legend.position="none")+ labs(title="Halloween Candy PCA", subtitle="Colored by type: chocolate bar(dark cyan), chocolate other (brown), fruity(pink), other(black)", caption="Data from 538")) ‘ Now, we're going to look at the PCA loadings.

```
barplot(pca$rotation[,1],las=2,ylab="PC1 contribution",col="darkcyan")
```



Q24

If you're chocolate, you're probably a bar, and if you're fruity, you're probably a hard pluribus.