

halloween-mini-project

```
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

Q3. What is your favorite candy in the dataset and what is its winpercent value?

```
candy["Reese's Peanut Butter cup", ]$winpercent
```

[1] 84.18029

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

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

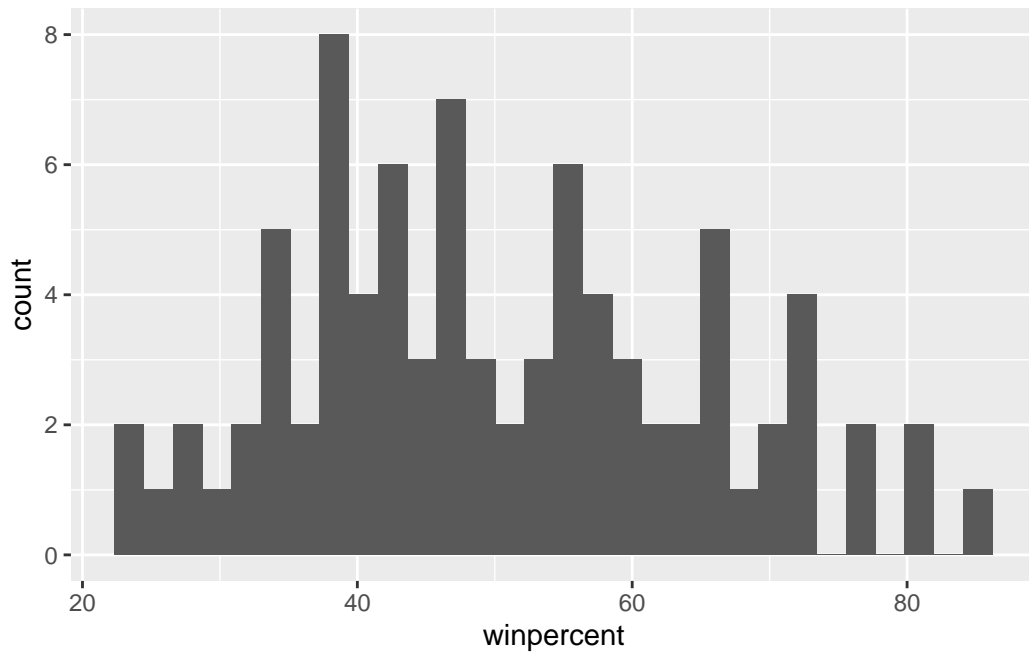
Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset? Yes, winpercent is on a different scale

Q7. What do you think a zero and one represent for the candy\$chocolate column? If it is a chocolate-y candy or not- 0= not, 1= yes

Q8. Plot a histogram of winpercent values

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

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



Q9. Is the distribution of winpercent values symmetrical?

No

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

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

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

```
[1] 60.05188
```

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

```
[1] 60.92153
```

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

```
[1] 44.11974
```

Higher!

Q12. Is this difference statistically significant?

```
choc<-candy$winpercent[as.logical(candy$chocolate)]
fruit<-candy$winpercent[as.logical(candy$fruity)]
t.test(choc,fruit)
```

Welch Two Sample t-test

```
data:  choc and fruit
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!

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

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

	chocolate	fruity	caramel	peanut	almond	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

	crisped	rice	wafer	hard	bar	pluribus	sugar	percent
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

	price	percent	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

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

	chocolate	fruity	caramel	peanut	almond	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

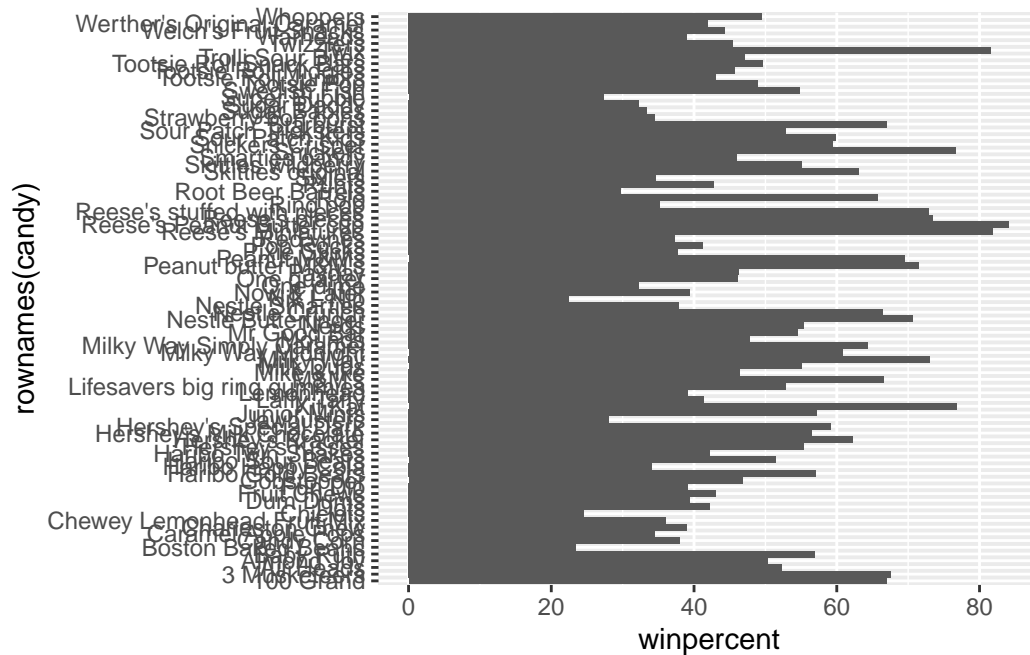
	crisped	rice	wafer	hard	bar	pluribus	sugar	percent	price	percent
Nik L Nip		0	0	0		1		0.197		0.976
Boston Baked Beans		0	0	0		1		0.313		0.511
Chiclets		0	0	0		1		0.046		0.325
Super Bubble		0	0	0		0		0.162		0.116
Jawbusters		0	1	0		1		0.093		0.511

	winpercent
Nik L Nip	22.44534
Boston Baked Beans	23.41782
Chiclets	24.52499
Super Bubble	27.30386
Jawbusters	28.12744

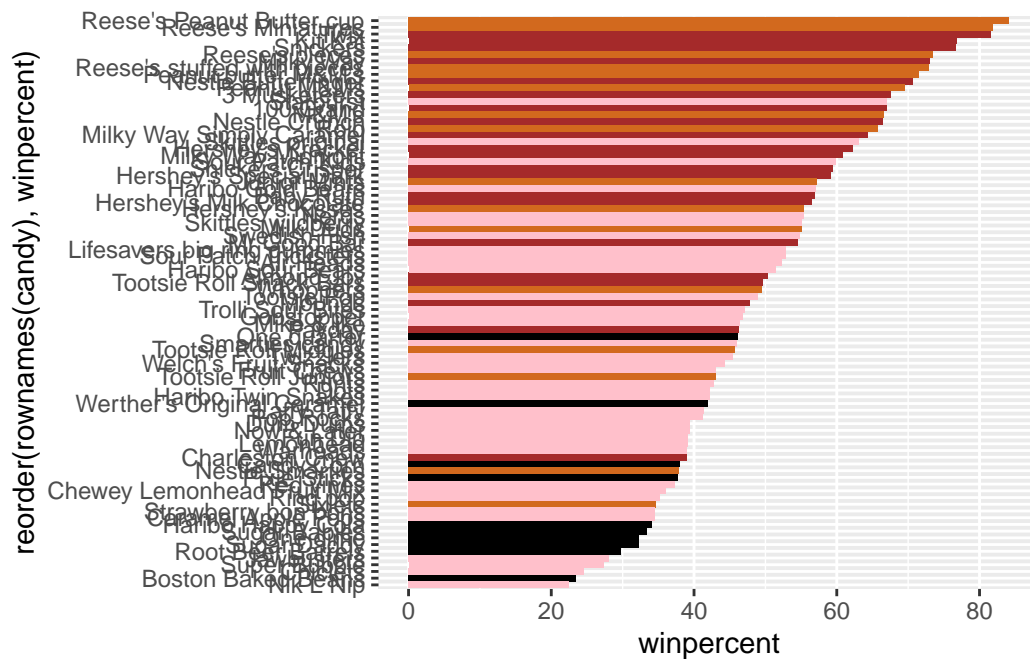
Nik L Nip, Boston Baked Beans, Super Bubble, Jawbusters Q14. What are the top 5 all time favorite candy types out of this set? Reese's Peanut Butter cup, Reese's miniatures, Twix, Kit Kat, Snickers

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

```
ggplot(candy)+
  aes(x=winpercent, y=rownames(candy)) +
  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"
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols)
```

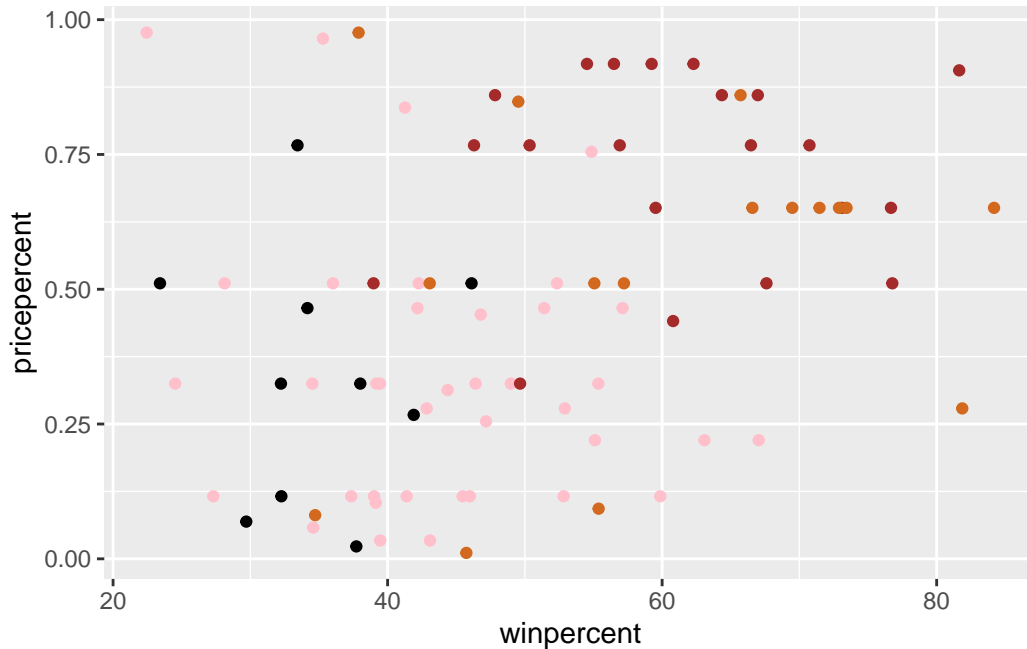


Now, for the first time, using this plot we can answer questions like: - Q17. What is the worst ranked chocolate candy? Boston Baked Beans

- Q18. What is the best ranked fruity candy? Starburst

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

# How about a plot of price vs win
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) #+
```



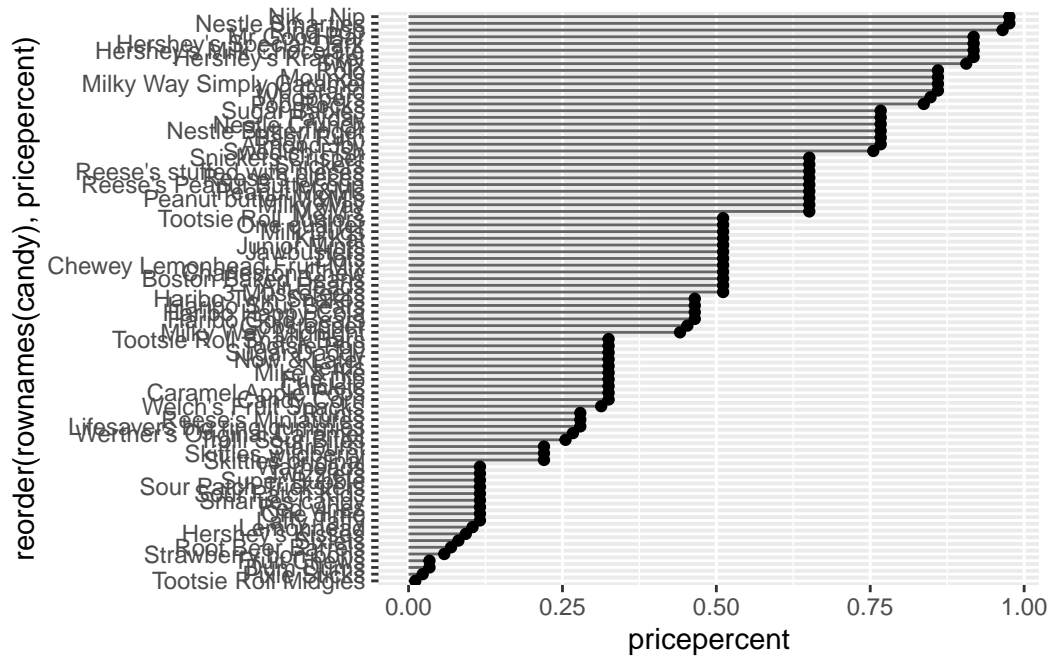
```
#geom_text_repel(col=my_cols, size=3.3, max.overlaps = 10)
```

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? Tootsie Roll Midgies

Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular? Nik L Nip (least popular), Nestle Smarties, Ring Pop, Mr. Good Bar, Hershey's Milk Chocolate

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()`.

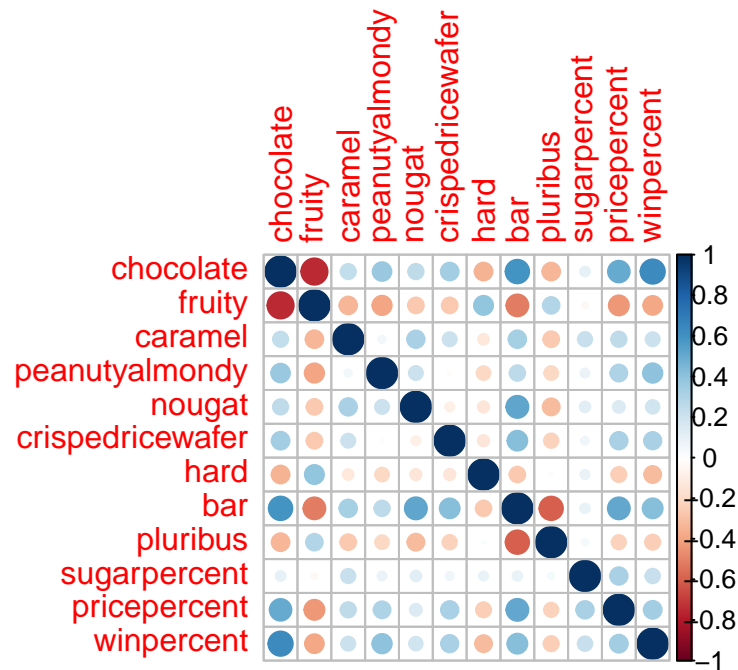
```
ggplot(candy, aes(pricepercent, reorder(rownames(candy), winpercent))) +  
  geom_col()
```

```
#install.packages("corrplot")
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 pricepercent Q23. Similarly, what two variables are most positively correlated?

```
cor(candy$chocolate,candy$winpercent)
```

```
[1] 0.6365167
```

```
cor(candy$chocolate,candy$bar)
```

```
[1] 0.5974211
```

winpercent and chocolate

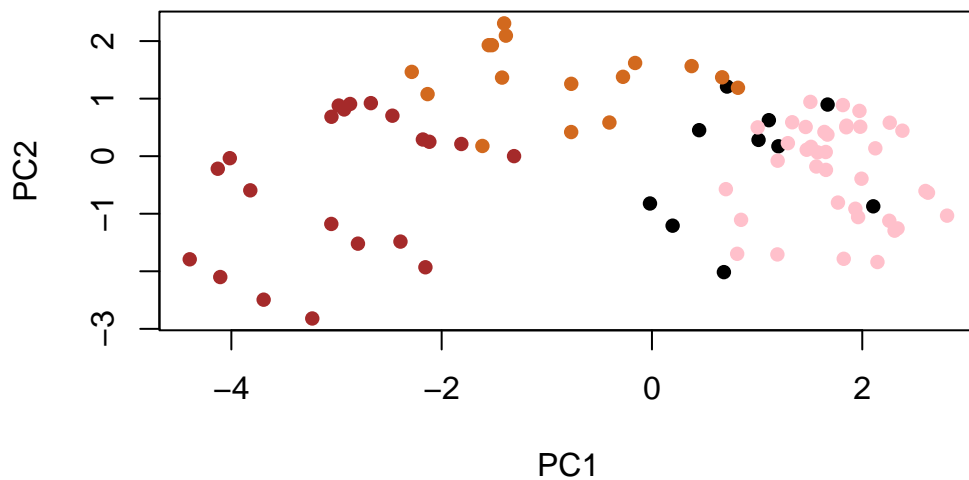
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

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

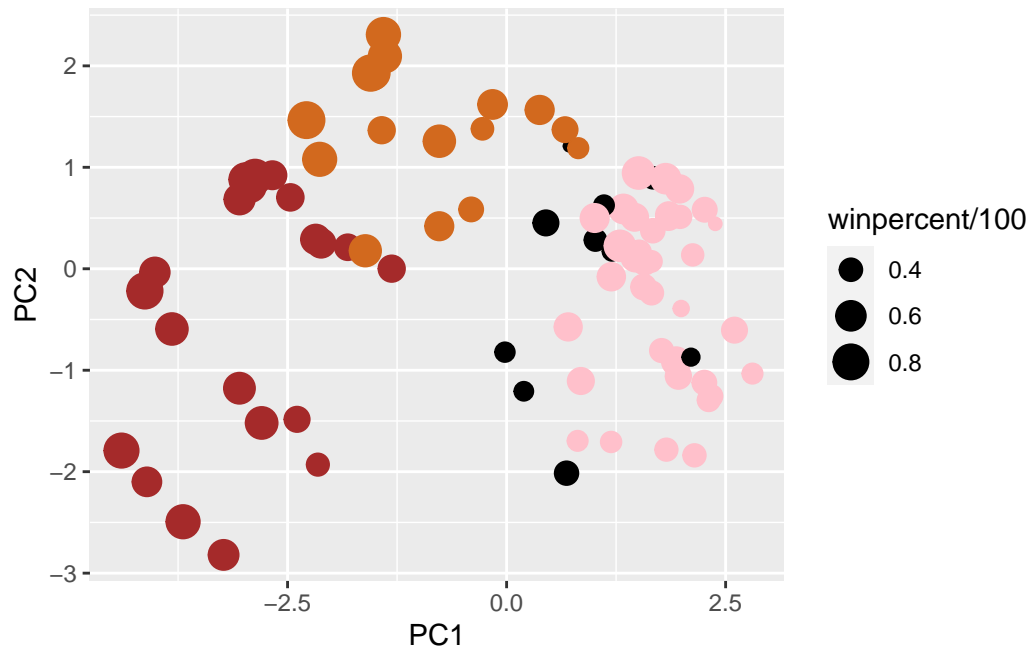
Proportion of Variance	0.3601	0.1079	0.1025	0.09636	0.0755	0.05593	0.05539
Cumulative Proportion	0.3601	0.4680	0.5705	0.66688	0.7424	0.79830	0.85369
	PC8	PC9	PC10	PC11	PC12		
Standard deviation	0.74530	0.67824	0.62349	0.43974	0.39760		
Proportion of Variance	0.04629	0.03833	0.03239	0.01611	0.01317		
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)
```

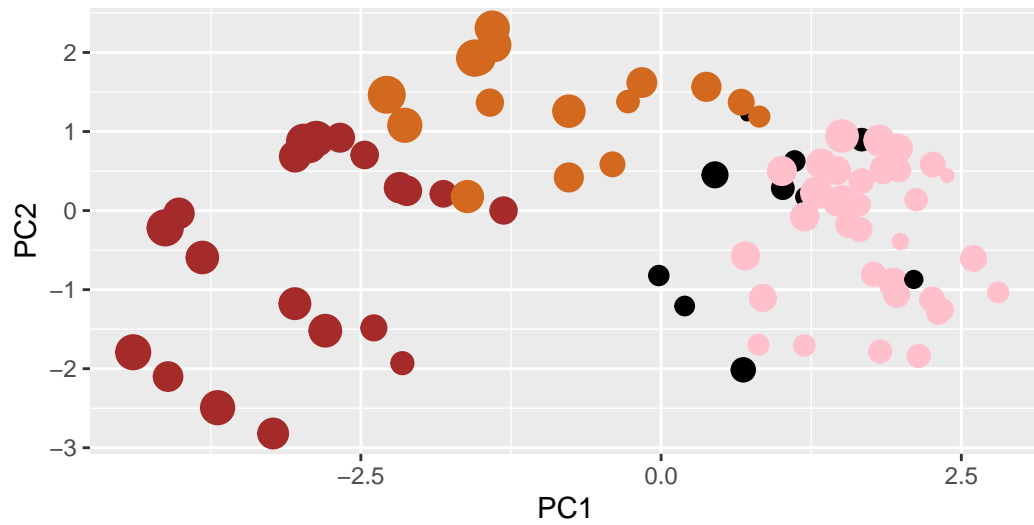
p



```
p + #geom_text_repel(size=3.3, col=my_cols, max.overlaps = 7) +
  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")
```

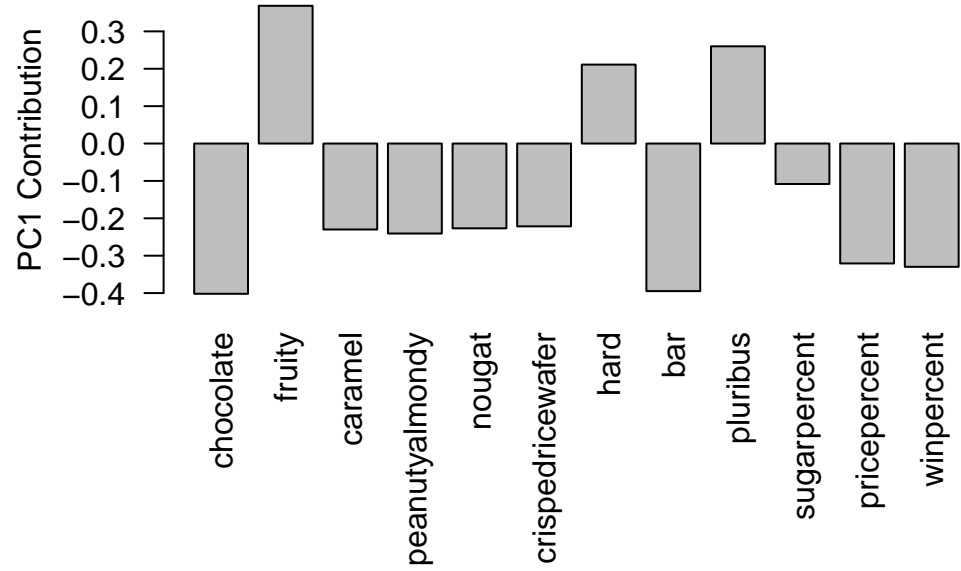
Halloween Candy PCA Space

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



Data from 538

```
#install.packages("plotly")
#library(plotly)
#ggplotly(p)
par(mar=c(8,4,2,2))
barplot(pca$rotation[,1], las=2, ylab="PC1 Contribution")
```



Q24 Pluribus and fruity