Week 5 Halloween Mini Project

David Alvarez

```
candy_file <- "candy-data.txt"
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 'winpercent' value?

```
candy["Welch's Fruit Snacks", ]$winpercent
```

```
[1] 44.37552
```

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 Snacks Bars'

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

```
[1] 49.6535
```

Q6. Is there any variable/column that looks to be on a different scale?

```
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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	

```
# After loading the data summary using the function above, it seems like the 'winpercent'
```

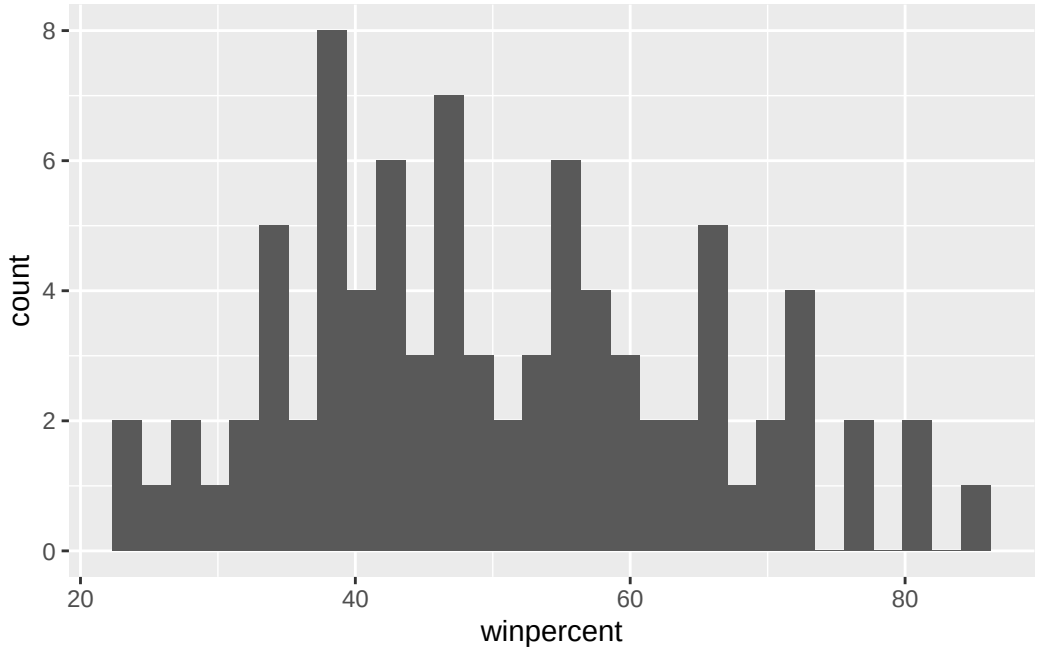
Q7. What do you think a zero and one represent for the ‘candy$chocolate’ column?

```
# I think the zero and one represent minimum and maximum values or probability in which th
```

Q8. Plot a histogram of winpercent values

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

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



Q9. Is the distribution of winpercent values symmetrical?

```
# The distribution is not symmetrical since it looks a bit skewed to the left, however it
```

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

```
# The center of distribution is just under 50 percent.  it looks to be around the range of
```

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

```
# Chocolate candy data
Choc.Inds <- as.logical(candy$chocolate)
Choc.win <- candy[Choc.Inds, "winpercent"]
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
```

```
# Fruit candy data
Fruit.Inds <- as.logical(candy$fruity)
Fruit.win <- candy[Fruit.Inds, "winpercent"]
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
```

```
# Comparing fruit and chocolate 'winpercent'

mean(Choc.Inds)
```

```
[1] 0.4352941
```

```
mean(Fruit.Inds)
```

```
[1] 0.4470588
```

Q12. Is the difference above statistically significant?

```
t.test(Choc.win, Fruit.win)
```

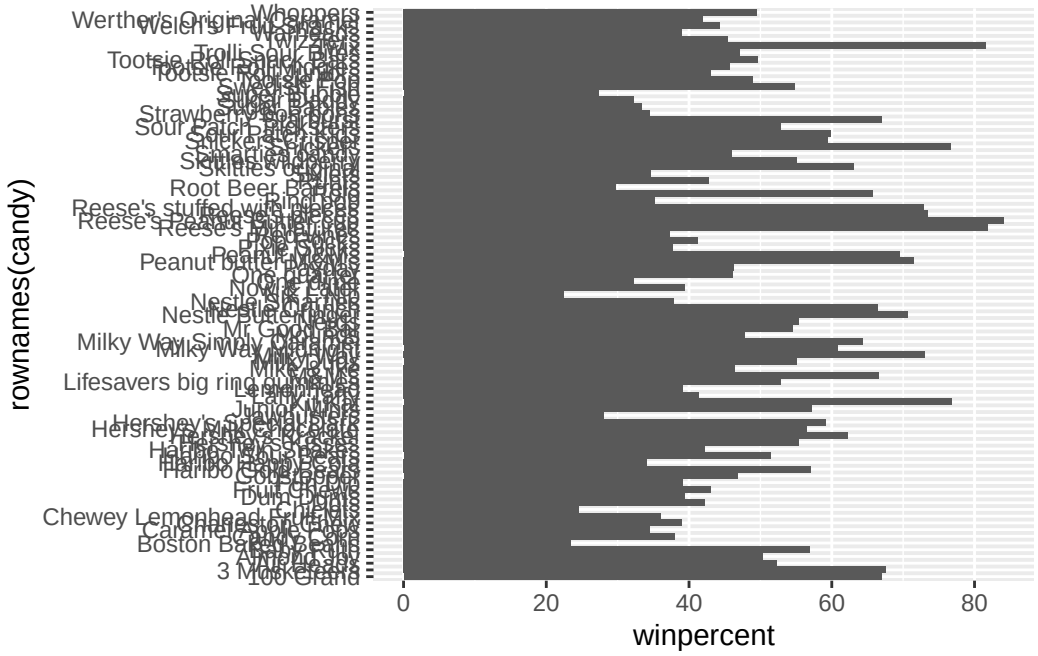
```
    Welch Two Sample t-test

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
```

```
# Values seem to be significant
```

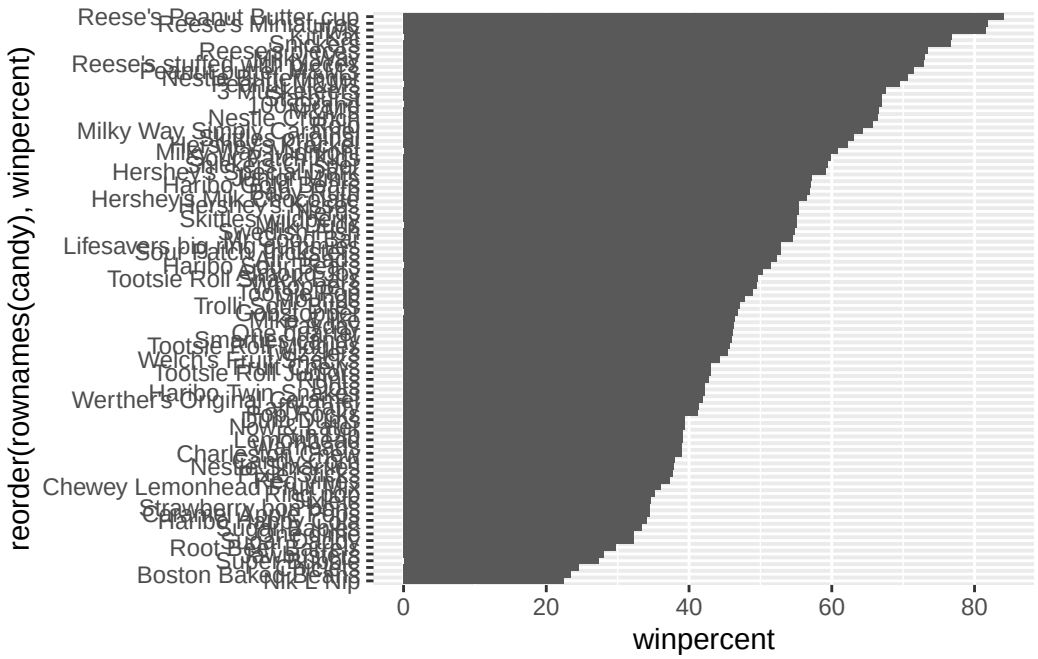
Q15. Make a barplot of candy ranking based on 'winpercent'

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



Q16. Sort the barplot

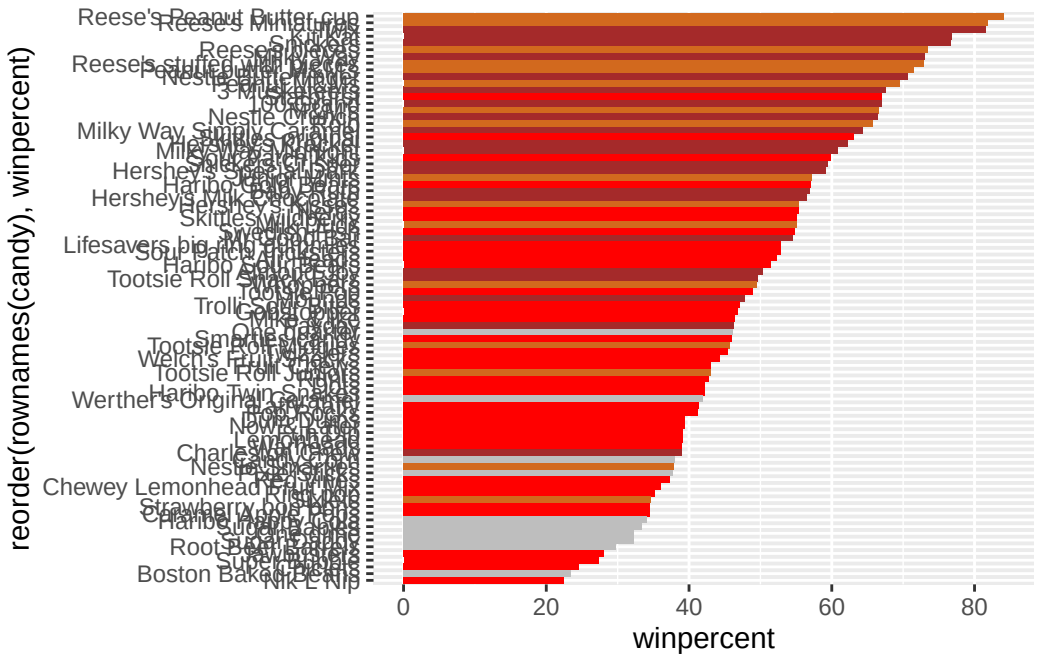
```
ggplot(candy) + aes(winpercent, reorder(rownames(candy),winpercent)) + geom_col()
```




```
mycols <- rep("gray", nrow(candy))
mycols[as.logical(candy$chocolate)] <- "chocolate"
mycols[as.logical(candy$fruity)] <- "red"
mycols[as.logical(candy$bar)] <- "brown"
mycols
```

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

```
# Improving the barplot
ggplot(candy) + aes(winpercent, reorder(rownames(candy),winpercent)) + geom_col(fill=mycol
```



Q17. What is the worst ranked chocolate candy?

```
# The worst ranked chocolate is Sixlets
```

Q18. What is the best ranked fruity candy?

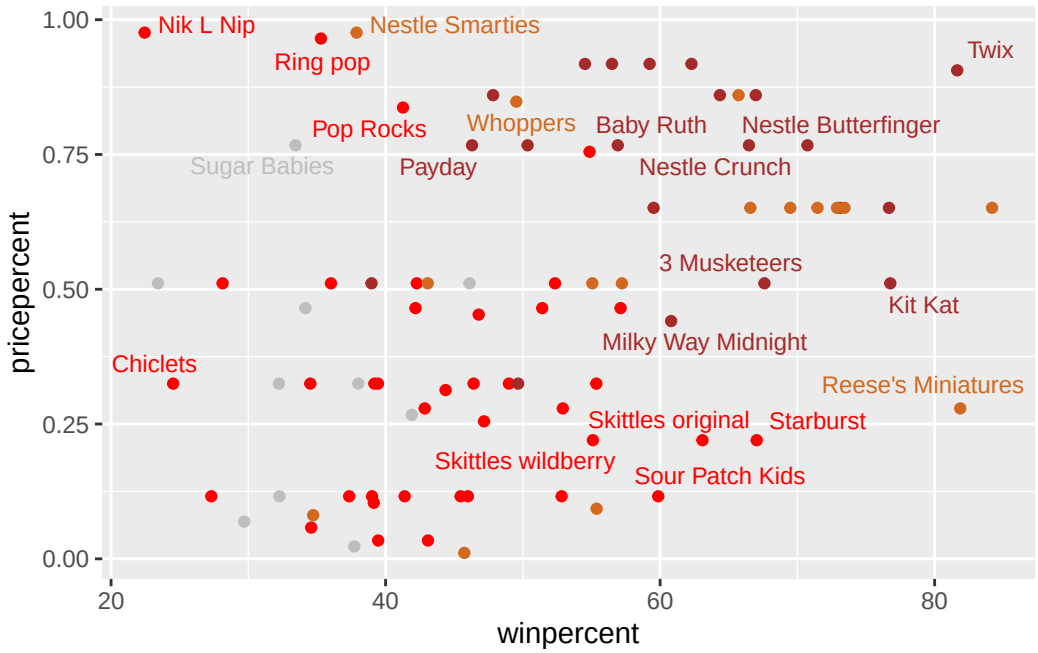
```
# The best ranked fruity candy is Starburst
```

Q19. Which candy type is the highest ranked in terms of 'winpercent' for the least money?

```
library(ggrepel)

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

```
Warning: ggrepel: 65 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




```
# After observing the plot, it seems as though Reeses Miniatures seem to have the most ban
```

Q20. What are the top 5 most expensive candy types in the dataset and 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
```

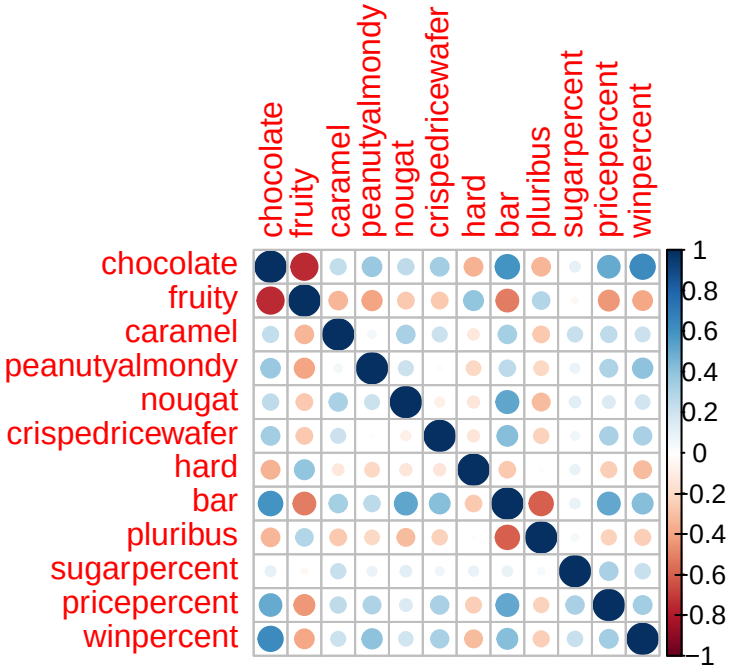
```
# The least popular of the candy presented was Nik L Nip
```

Q22. Examining the plot below, what two variables are anti-correlated?

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

```
## corrplot 0.90 loaded
cij <- cor(candy)
corrplot(cij)
```




```
# After examining this plot, the two variables that are anti-correlated are chocolate and
```

Q23. What two variables are most positively correlated?

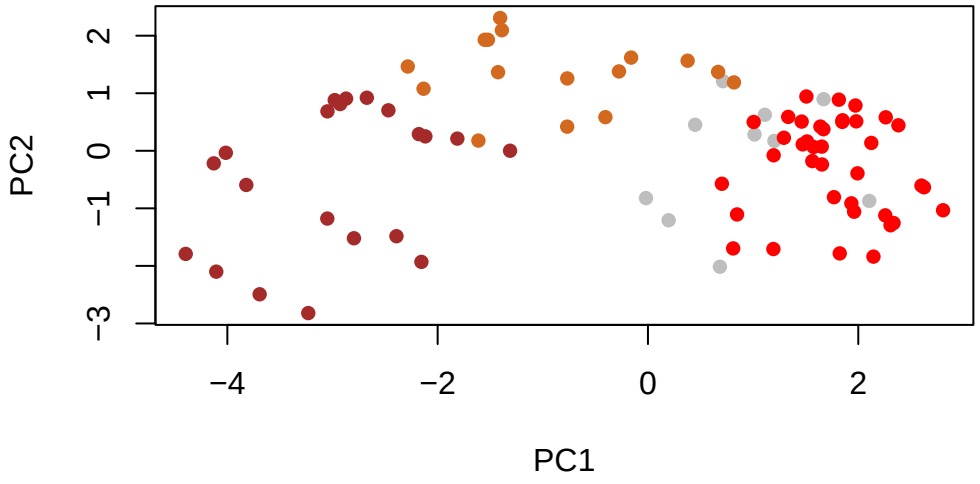
```
# Aside from the variables that are aligned with themselves, ie chocolate and chocolate, i
```

```
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
```

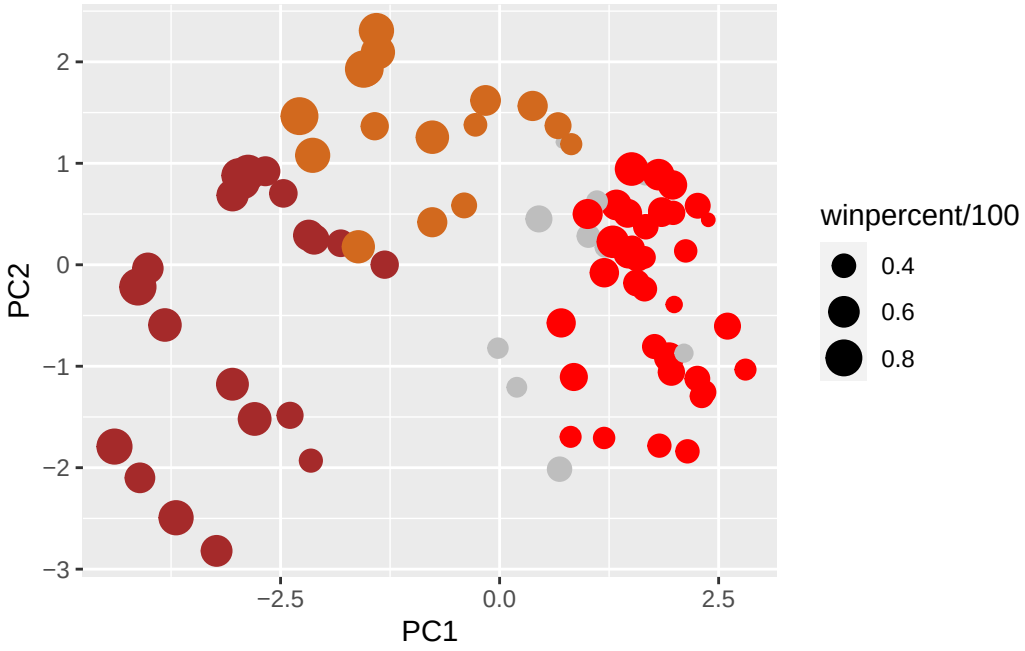
```
                            PC8     PC9    PC10    PC11    PC12
Standard deviation     0.74530 0.67824 0.62349 0.43974 0.39760
Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

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



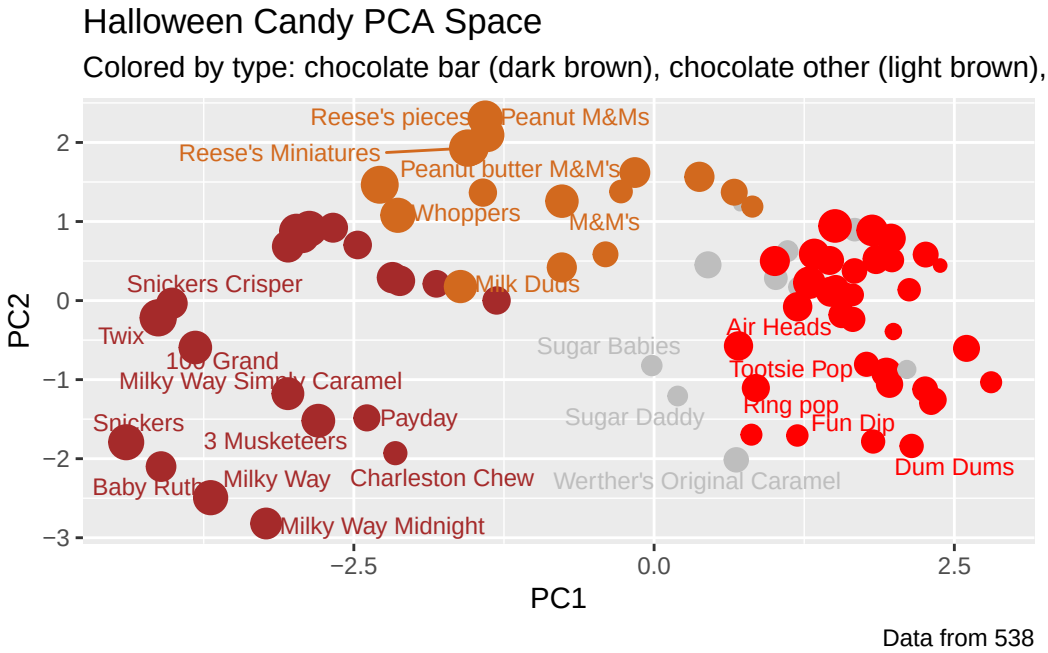

```r
# New data frame of PCA results
my_data <- cbind(candy, pca$x[,1:3])

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




```
library(ggrepel)
p + geom_text_repel(size=3.3, col=mycols, 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")
```

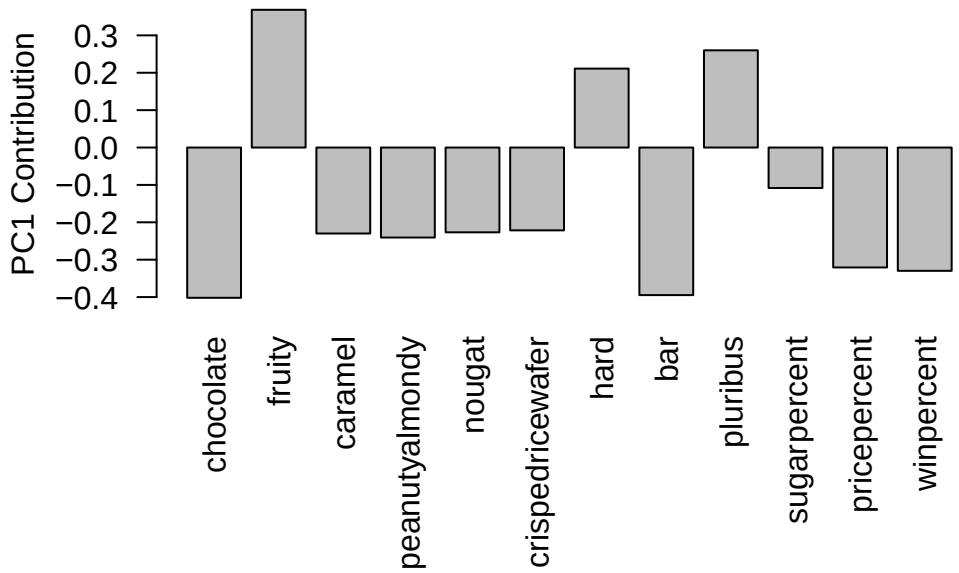
```
Warning: ggrepel: 59 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




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

Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense?

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




```
# The original variables picked up strongly by PC1 in the positive direction are 'fruity',
# These variables do seem to make sense to me since fruity type candies have these charact
```

```
library(tinytex)
```

```
Warning: package 'tinytex' was built under R version 4.3.2
```