Halloween Mini-Project

Jazz Zhang (A16149005)

```
candy_file <- read.csv("candy-data.csv")
candy = read.csv("candy-data.csv", row.names=1)

dim(candy)
```

```
[1] 85 12
```

Q1. 85 candy types

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

```
[1] 38
```

Q2. 38 fruity candy types

```
candy["Dum Dums",]$winpercent
```

```
[1] 39.46056
```

Q3. Dum Dums, 39.46%

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

```
[1] 76.7686
```

Q4. 76.77%

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

```
[1] 49.6535
```

Q5. 49.65%

```r
# install.packages("skimr")
library("skimr")
```

```
Warning: package 'skimr' was built under R version 4.3.1
```

```r
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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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. “winpercent” column

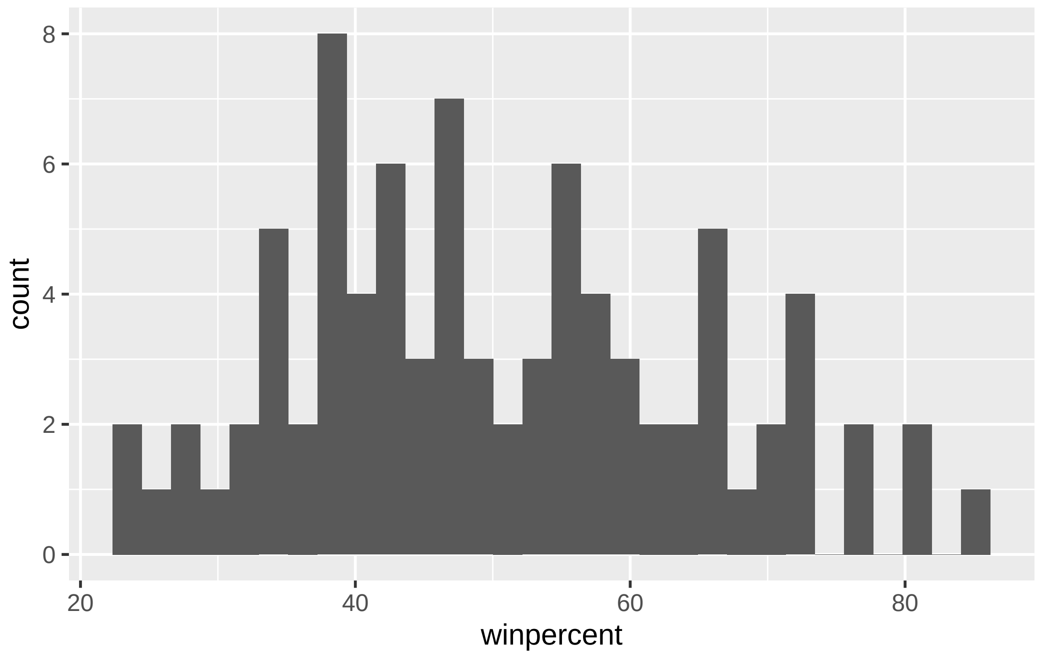
Q7. 0 and 1 represent the candy is either chocolate or not, repectively

```r
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.3.1
```

```r
# Q8.
ggplot(candy, aes(winpercent))+
  geom_histogram()
```

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



Q9. The distribution isn’t symmetrical

Q10. The ccenter is below 50

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

```

    Welch Two Sample t-test

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

Q11. Winpercent for chocolate is higher on average

Q12. The difference is statistically significant

```
library(dplyr)
```

```
Warning: package 'dplyr' was built under R version 4.3.1


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
```

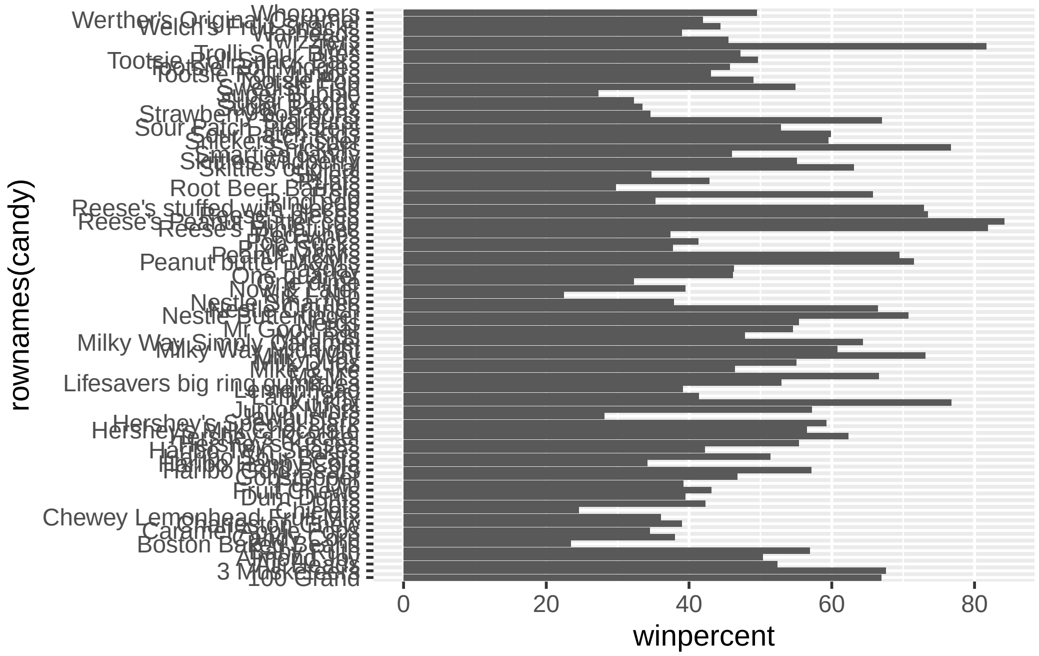
Q13. Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters

```
candy %>% arrange(desc(winpercent)) %>% head(5)
```

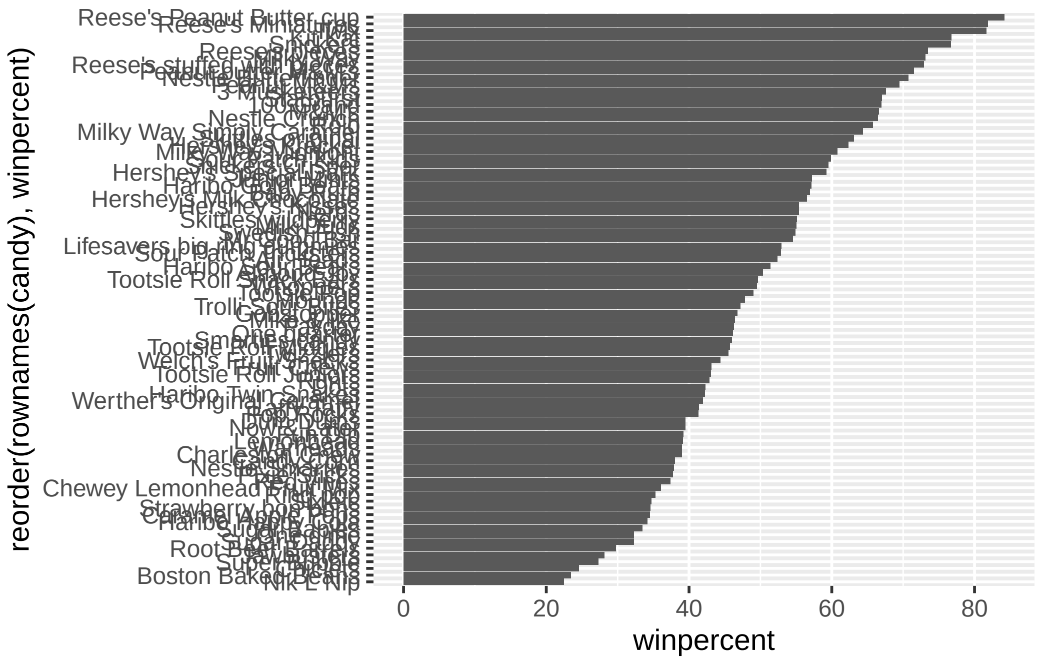
```
                          chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat                           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
```

Q14. Reese's Peanut Butter cup, Reese's Miniatures, Twix, Kit Kat, and Snickers

```
# Q15.
ggplot(candy) +
  aes(x=winpercent, y=rownames(candy)) +
  geom_bar(stat="identity")
```

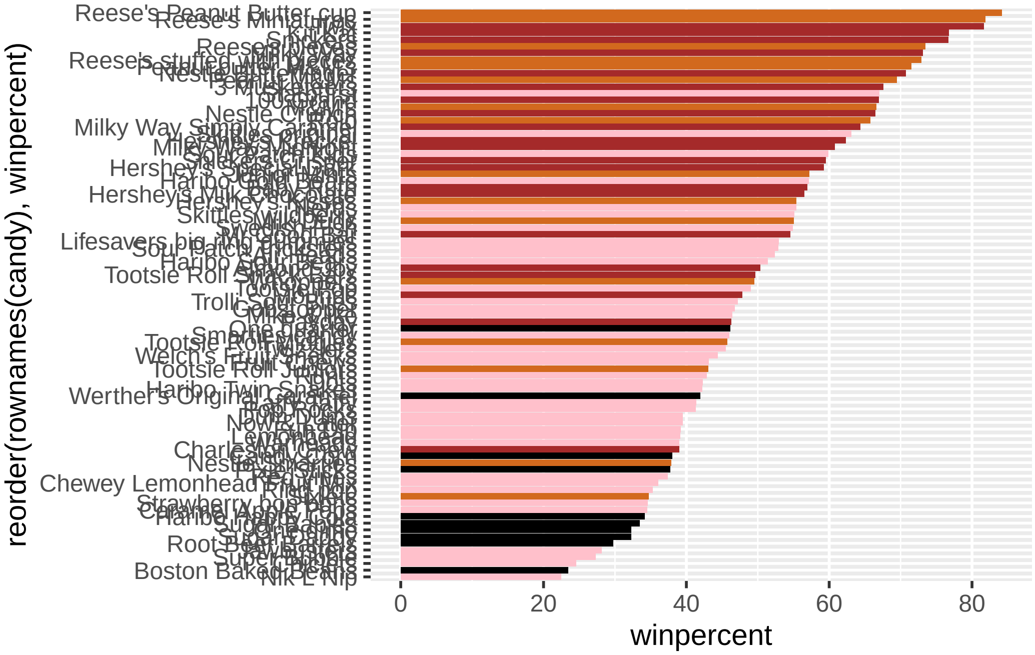


```
# Q16.
ggplot(candy) +
  aes(x=winpercent, y=reorder(rownames(candy),winpercent)) +
  geom_bar(stat="identity")
```



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



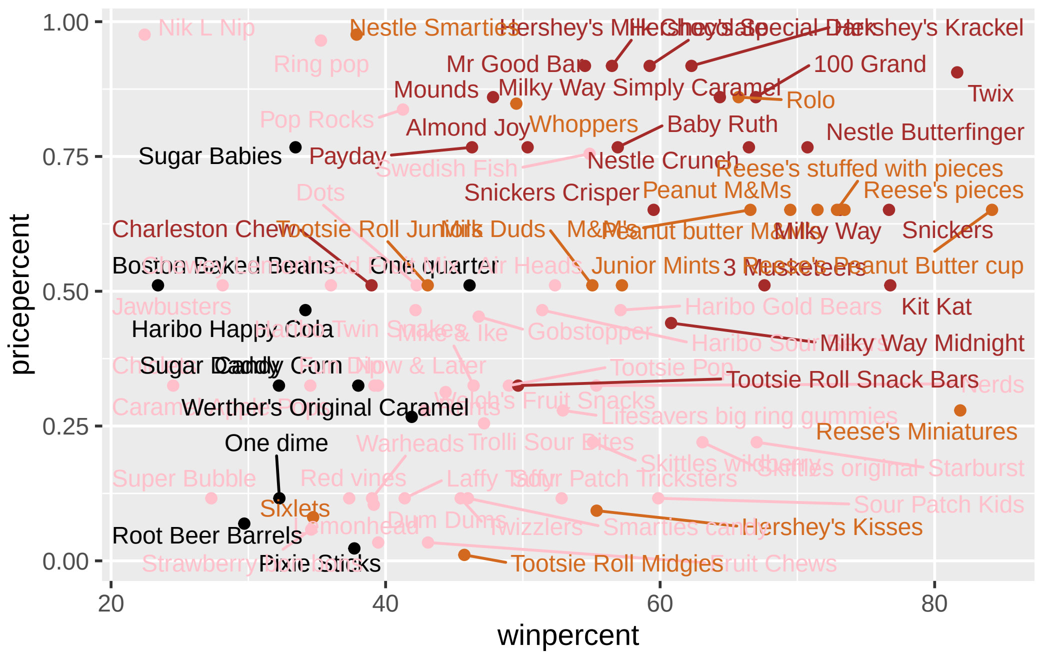
Q17. Sixlets

Q18. Starburst

```r
library(ggrepel)
```

```
Warning: package 'ggrepel' was built under R version 4.3.1
```

```r
ggplot(candy) +
  aes(x=winpercent, y=pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 50)
```



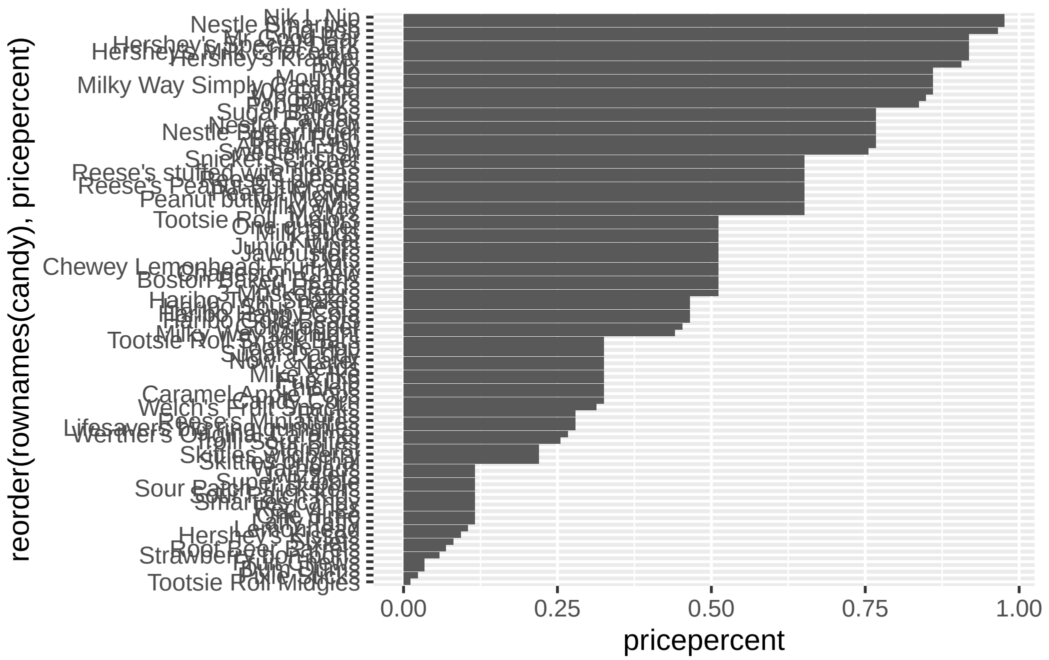
Q19. Reese's Miniatures

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

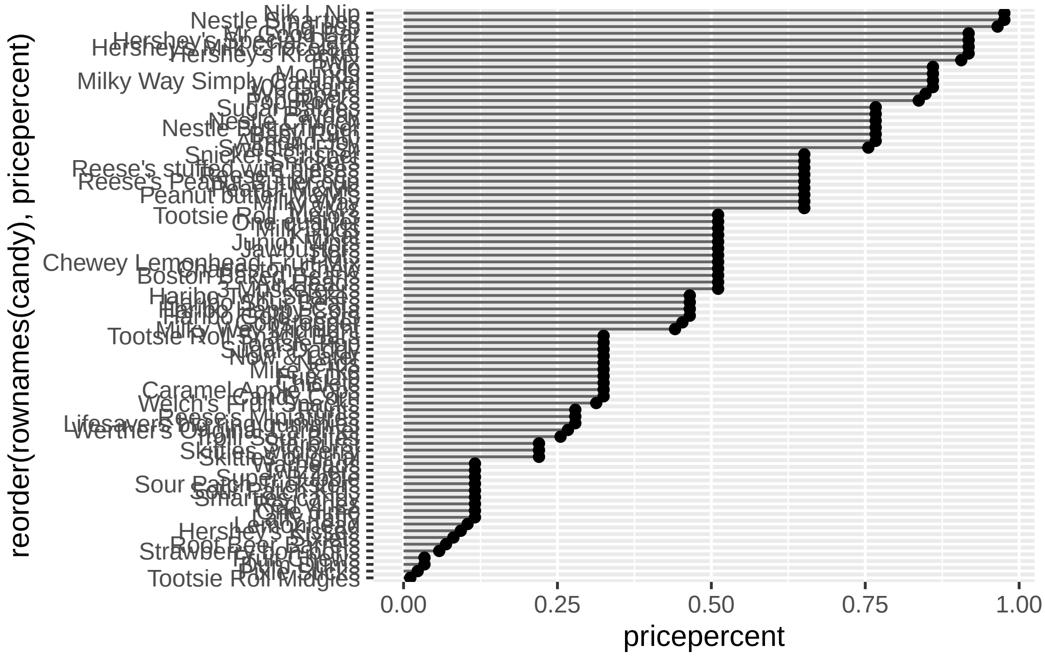
```
                         pricepercent winpercent
Nik L Nip                       0.976   22.44534
Nestle Smarties                 0.976   37.88719
Ring pop                        0.965   35.29076
Hershey's Krackel               0.918   62.28448
Hershey's Milk Chocolate        0.918   56.49050
```

Q20. Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, and Hershey's Milk Chocolate; Nik L Nip is the least popular

```
# Q21.
ggplot(candy) +
  aes(x=pricepercent, y=reorder(rownames(candy),pricepercent)) +
  geom_col()
```



```
# Q21. cont
ggplot(candy) +
  aes(x=pricepercent, y=reorder(rownames(candy),pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
  geom_point()
```



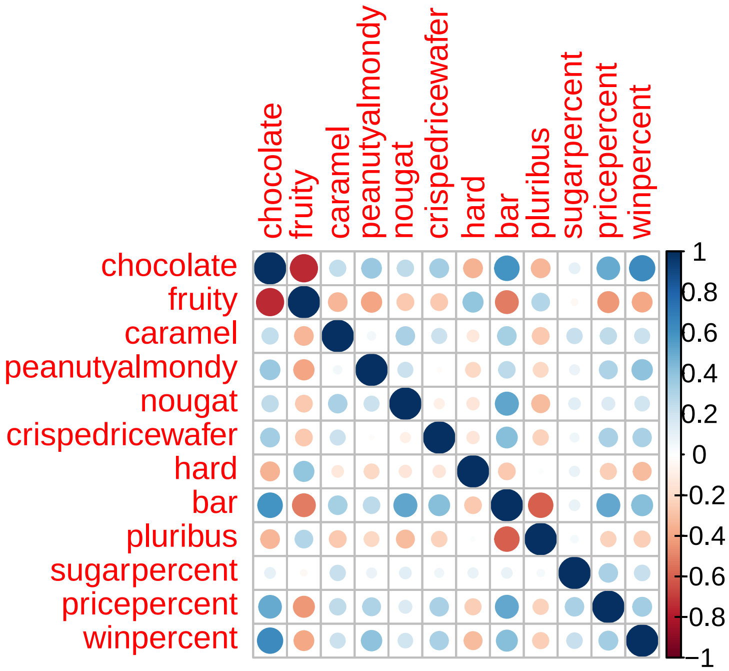
```
# install.packages("corrplot")

library(corrplot)
```

```
Warning: package 'corrplot' was built under R version 4.3.1
```

```
corrplot 0.92 loaded
```

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



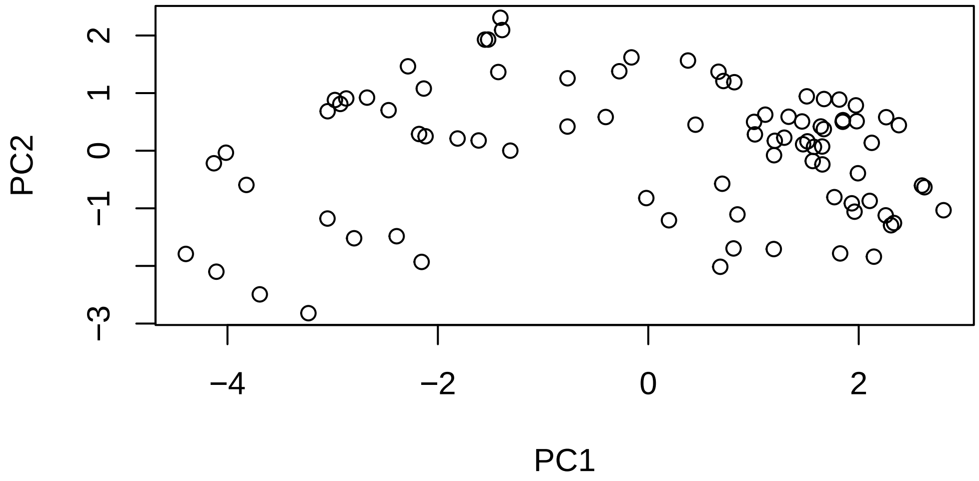
Q22. Chocolate and fruity

Q23. Chocolate and winpercent

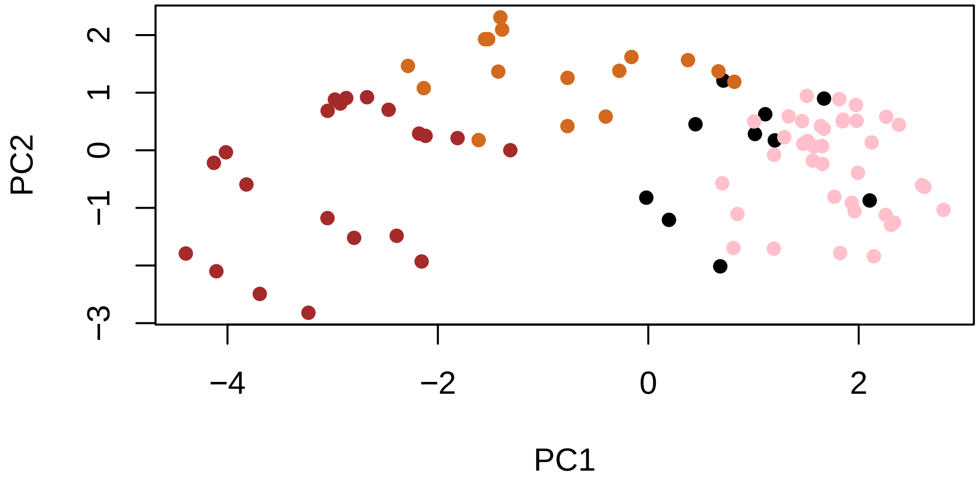
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
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])
```



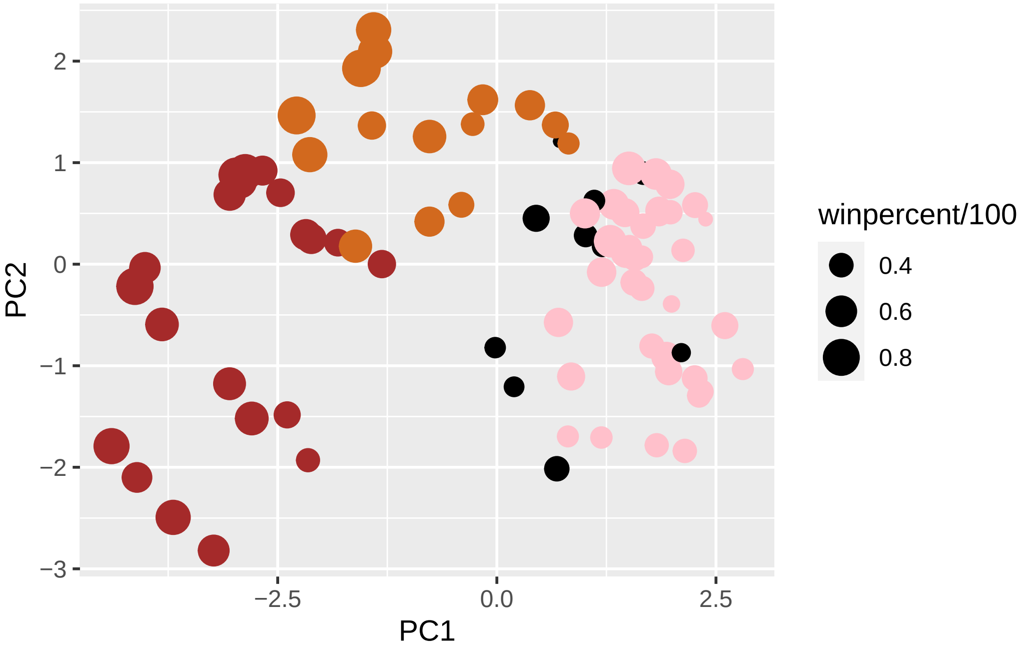
```
plot(pca$x[,1:2], col=my_cols, pch=16)
```



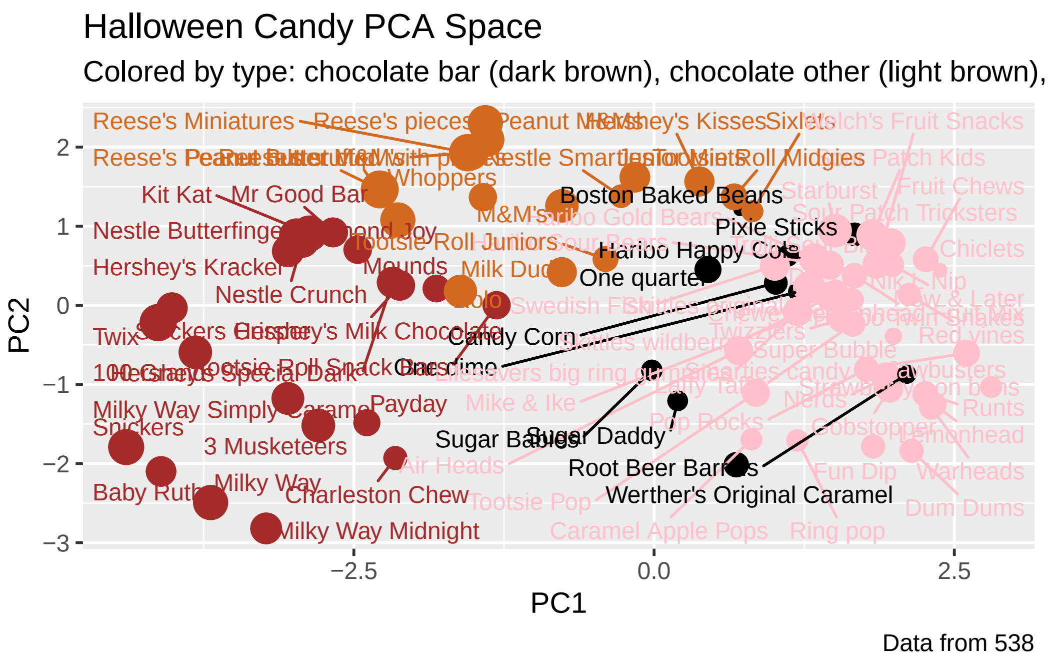
```r
my_data <- cbind(candy, pca$x[,1:3])

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
```



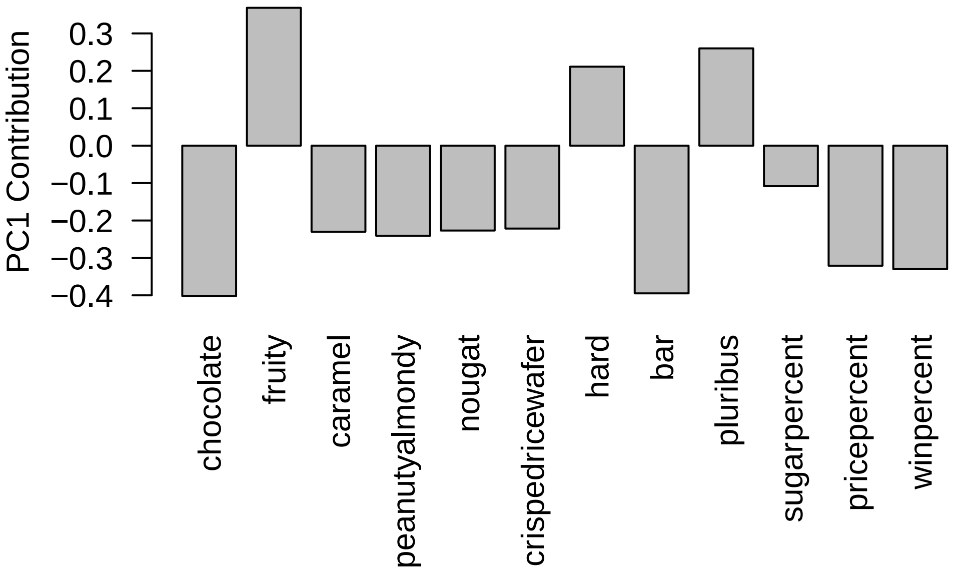
```r
p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 50)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown
       caption="Data from 538")
```



```
library(plotly)

ggplotly(p)
```

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



Q24. Fruity, pluribus, and hard.