Class 10: Halloween Mini Project

```r
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? 85 Q2. How many fruity candy types are in the dataset? 38

```r
dim(candy)
```

```
[1] 85 12
```

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

```
[1] 38
```

> Q3. What is your favorite candy in the dataset and what is it's winpercent value? Reese's peanut butter cup, 84.18% Q4. What is the winpercent value for "Kit Kat"? 76.77% Q5. What is the winpercent value for "Tootsie Roll Snack Bars"? 49.65

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

```
[1] 84.18029
```

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

```
[1] 76.7686
```

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

```
[1] 49.6535
```

```r
library("skimr")
skim(candy)
```

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	▇▇▇▇▆
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? The winpercent variable is on a range from 1-100, whereas the values for all of the other variables are less than 1 Q7. What do you think a zero and one represent for the candy$chocolate column? I think a zero means that the candy does not have chocolate, and a one means that the candy does have chocolate.

> Q Find fruity candy with a win percnt above 50%

```r
library(dplyr)
```

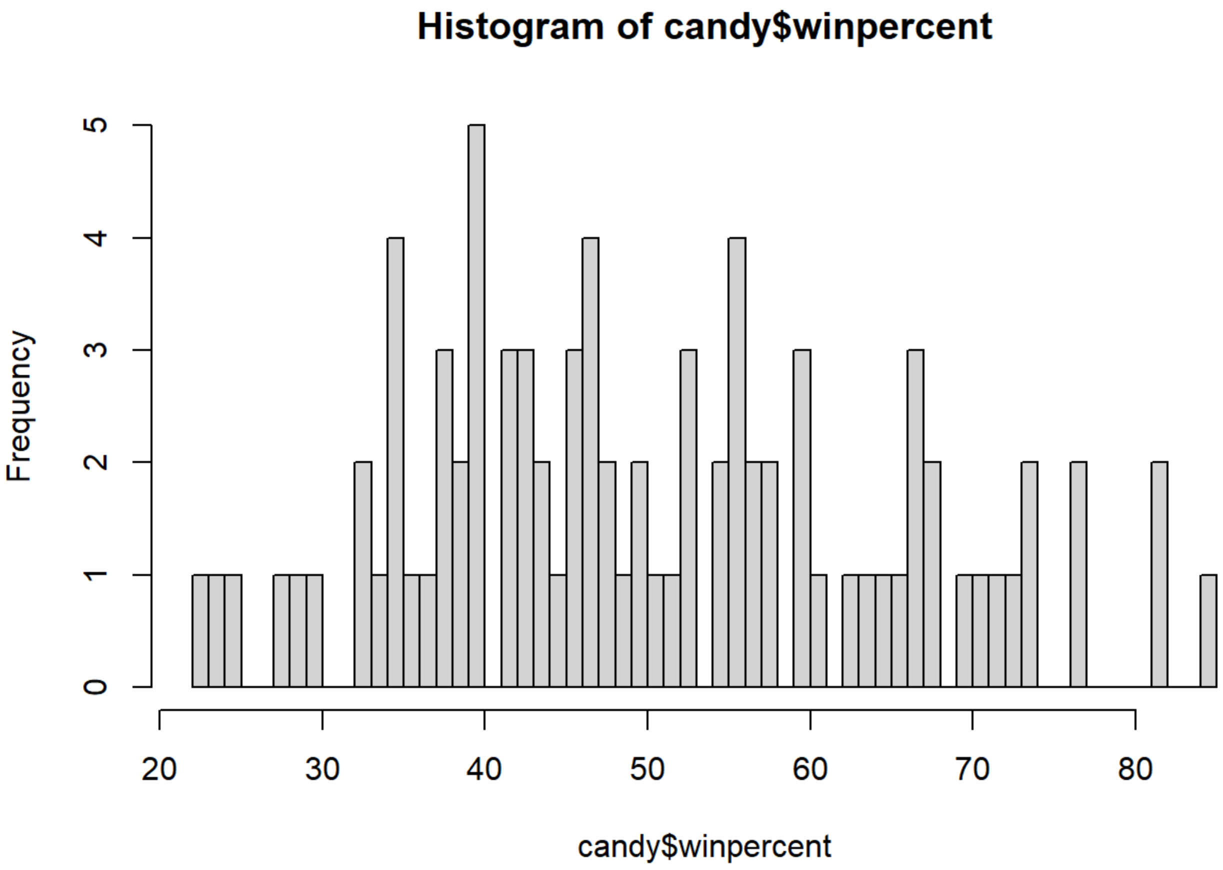
```

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
```

```r
candy |>
  filter(fruity == 1) |>
  filter(winpercent > 50)
```

```
                            chocolate fruity caramel peanutyalmondy nougat
Air Heads                           0      1       0              0      0
Haribo Gold Bears                   0      1       0              0      0
Haribo Sour Bears                   0      1       0              0      0
Lifesavers big ring gummies         0      1       0              0      0
Nerds                               0      1       0              0      0
Skittles original                   0      1       0              0      0
Skittles wildberry                  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
Swedish Fish                        0      1       0              0      0
                            crispedricewafer hard bar pluribus sugarpercent
Air Heads                                  0    0   0        0        0.906
Haribo Gold Bears                          0    0   0        1        0.465
Haribo Sour Bears                          0    0   0        1        0.465
Lifesavers big ring gummies                0    0   0        0        0.267
Nerds                                      0    1   0        1        0.848
Skittles original                          0    0   0        1        0.941
Skittles wildberry                         0    0   0        1        0.941
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
Swedish Fish                               0    0   0        1        0.604
                            pricepercent winpercent
Air Heads                          0.511   52.34146
Haribo Gold Bears                  0.465   57.11974
Haribo Sour Bears                  0.465   51.41243
Lifesavers big ring gummies        0.279   52.91139
Nerds                              0.325   55.35405
Skittles original                  0.220   63.08514
Skittles wildberry                 0.220   55.10370
```

```
Sour Patch Kids                0.116   59.86400
Sour Patch Tricksters          0.116   52.82595
Starburst                      0.220   67.03763
Swedish Fish                   0.755   54.86111
```

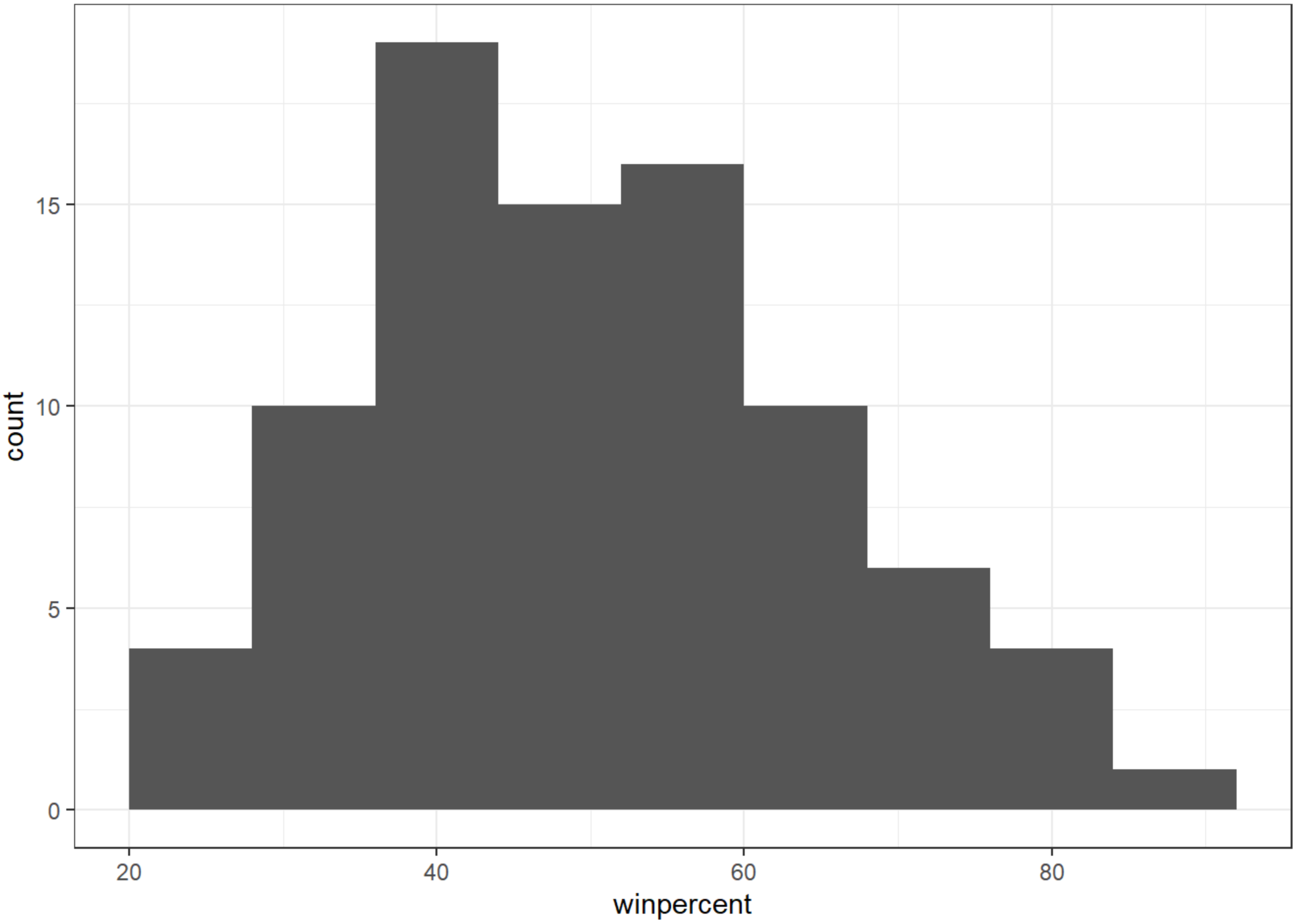
```r
hist(candy$winpercent, breaks = 50)
```



```r
summary(candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.45   39.14   47.83   50.32   59.86   84.18 
```

```r
library(ggplot2)
ggplot(candy) +
  aes(winpercent) +
  geom_histogram(binwidth = 8) +
  theme_bw()
```



```
chocolate <- candy |>
  filter(chocolate == 1)

fruity <- candy |>
  filter(fruity == 1)

mean(fruity$winpercent) > mean(chocolate$winpercent)
```

```
[1] FALSE
```

```
t.test(chocolate$winpercent, fruity$winpercent)
```

```

	Welch Two Sample t-test

data:  chocolate$winpercent and fruity$winpercent
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 
```

> Q8. Plot a histogram of winpercent values Q9. Is the distribution of winpercent values symmetrical? No Q10. Is the center of the distribution above or below 50%? The median is below 50%, but the mean is slightly above 50%. Q11. On average is chocolate candy higher or lower ranked than fruit candy? Chocolate candy is higher ranked than fruity candy Q12. Is this difference statistically significant? Since the p value is very small, the difference is statistically significant.

```r
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
```

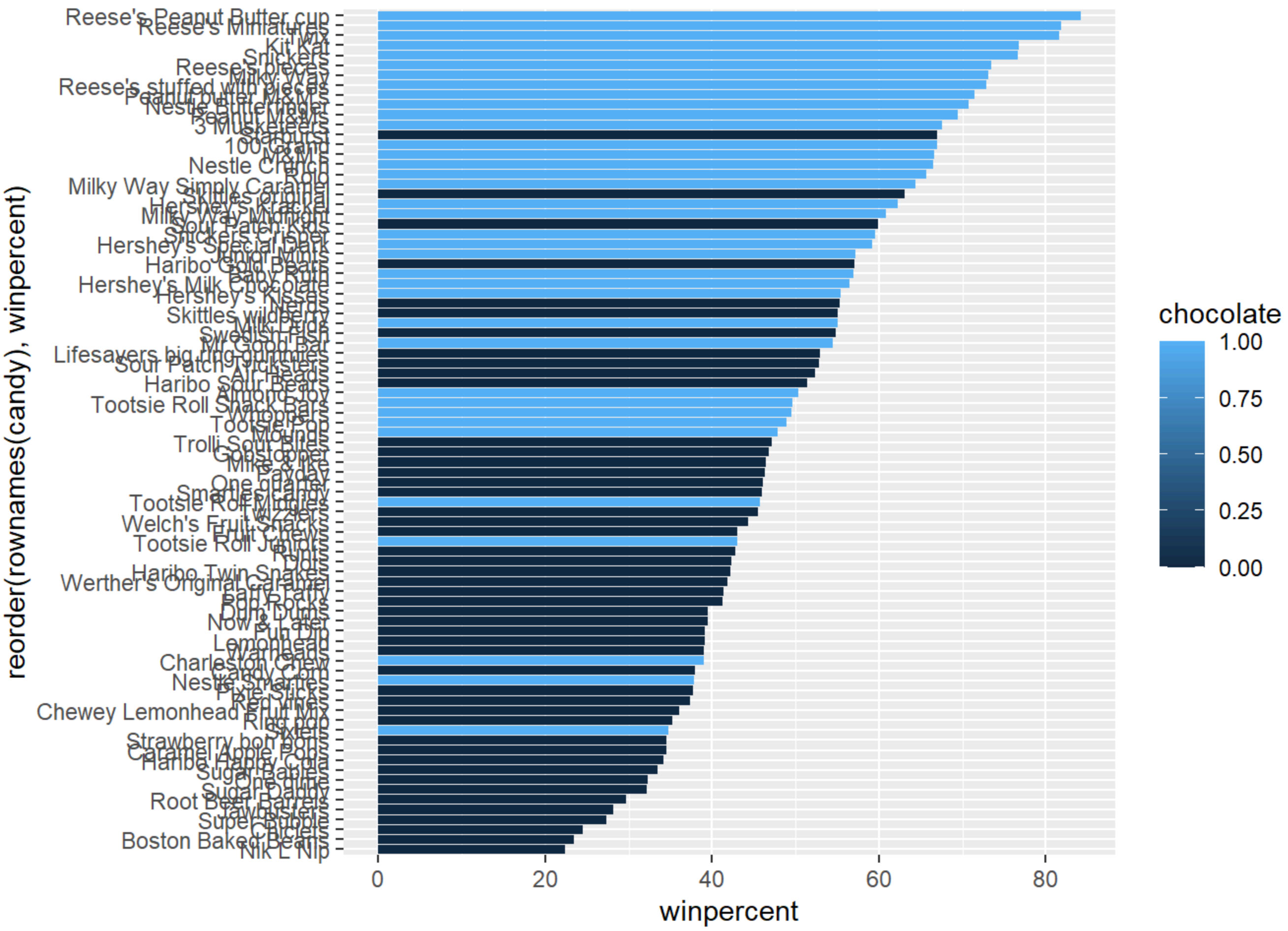
```r
candy %>% arrange(winpercent) %>% tail(5)
```

```
                          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
```

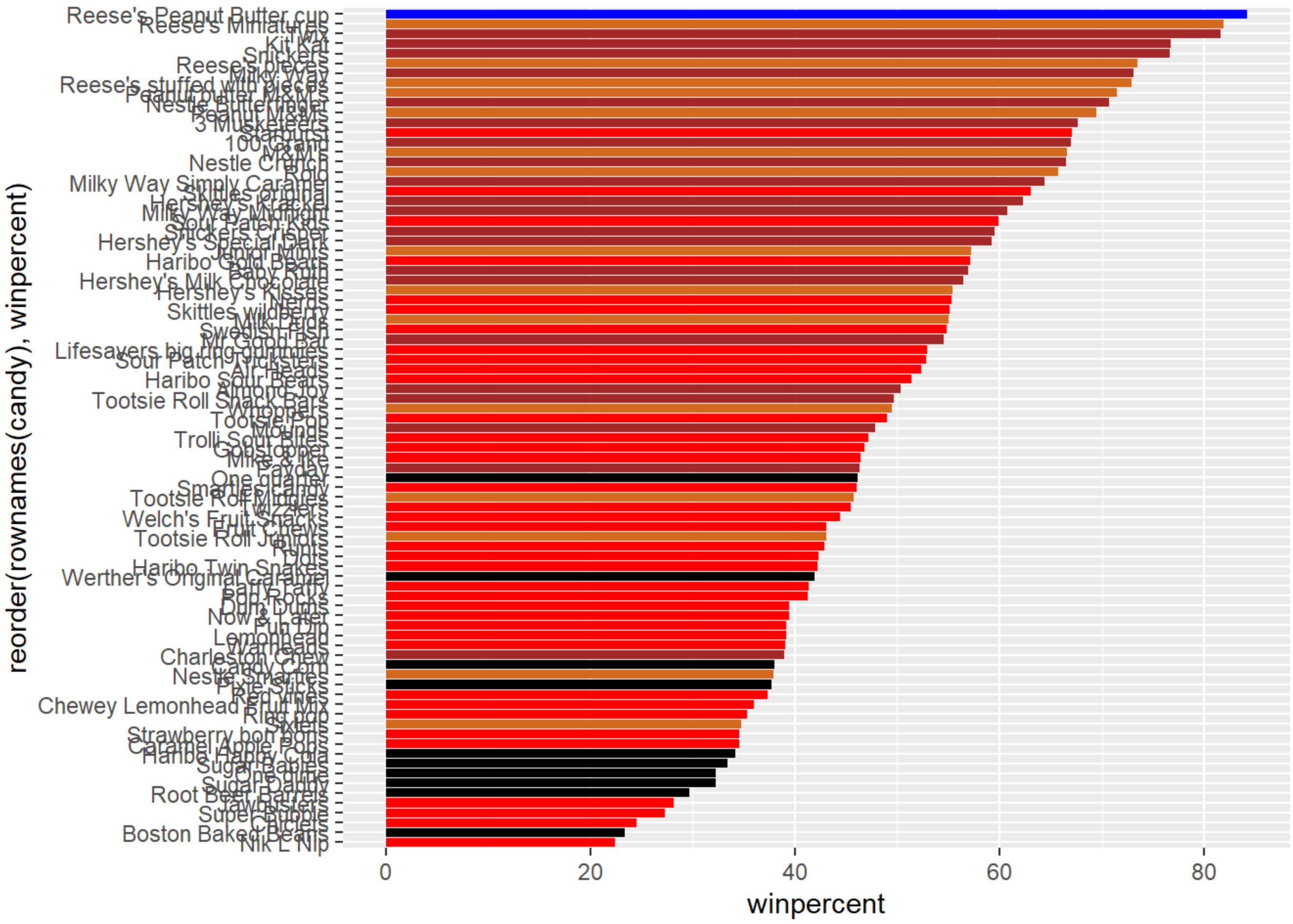
```
Reese's Miniatures                0.279    81.86626
Reese's Peanut Butter cup         0.651    84.18029
```

> Q13. What are the five least liked candy types in this set? Nik l nip, boston baked beans, chiclets, super bubble, jawbreakers Q14. What are the top 5 all time favorite candy types out of this set? Snickers, kit kat, twix, reese's miniature, reese's peanut butter cup

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



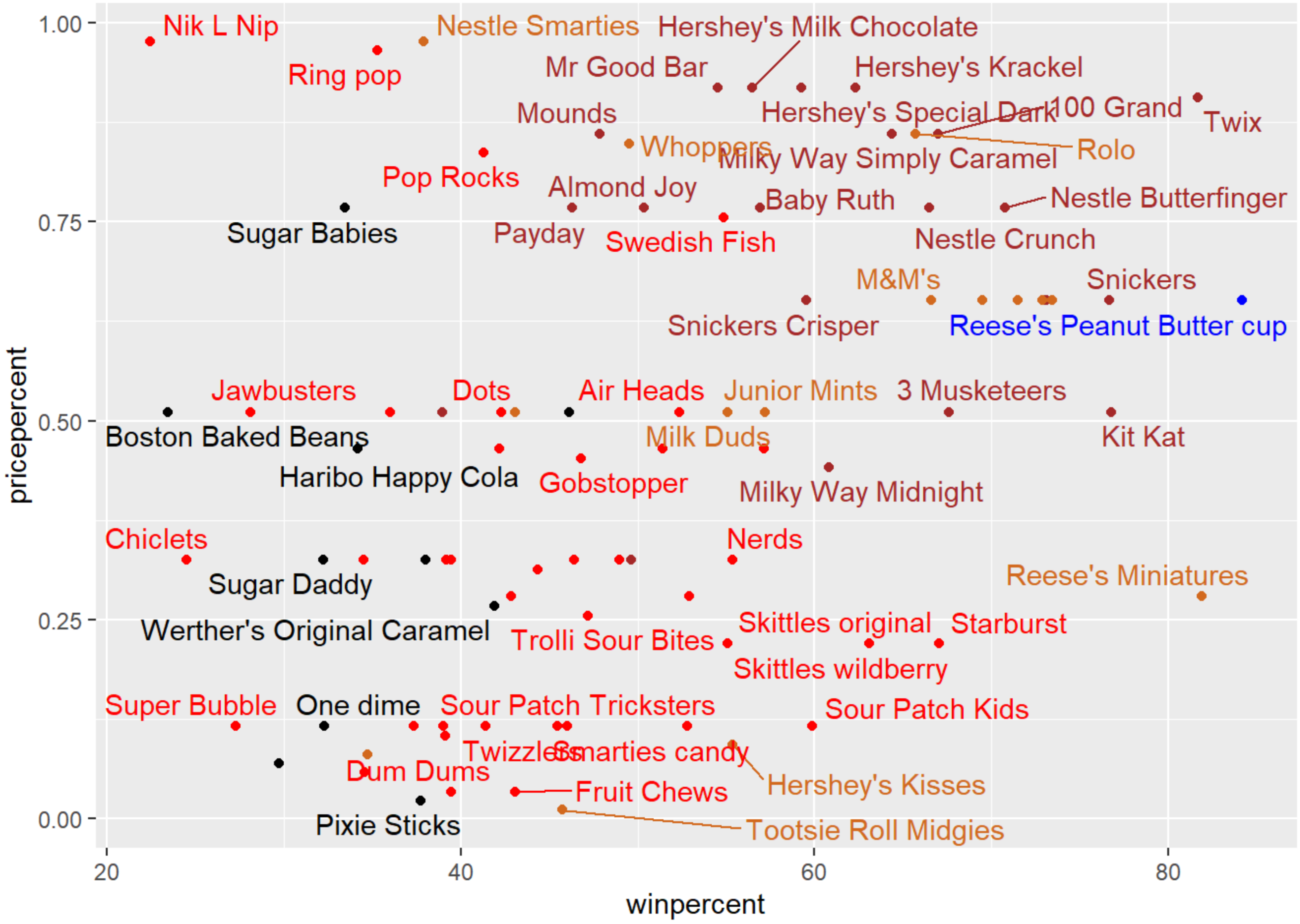
```
mycols <- rep("black", nrow(candy))
mycols[as.logical(candy$chocolate)] <- "chocolate"
mycols[as.logical(candy$bar)] <- "brown"
mycols[as.logical(candy$fruity)] <- "red"
mycols[rownames(candy) == "Reese's Peanut Butter cup"] <- "blue"
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent)) +
  geom_col(fill = mycols)
```



> Q17. What is the worst ranked chocolate candy? Sixlets Q18. What is the best ranked fruity candy? Starburst

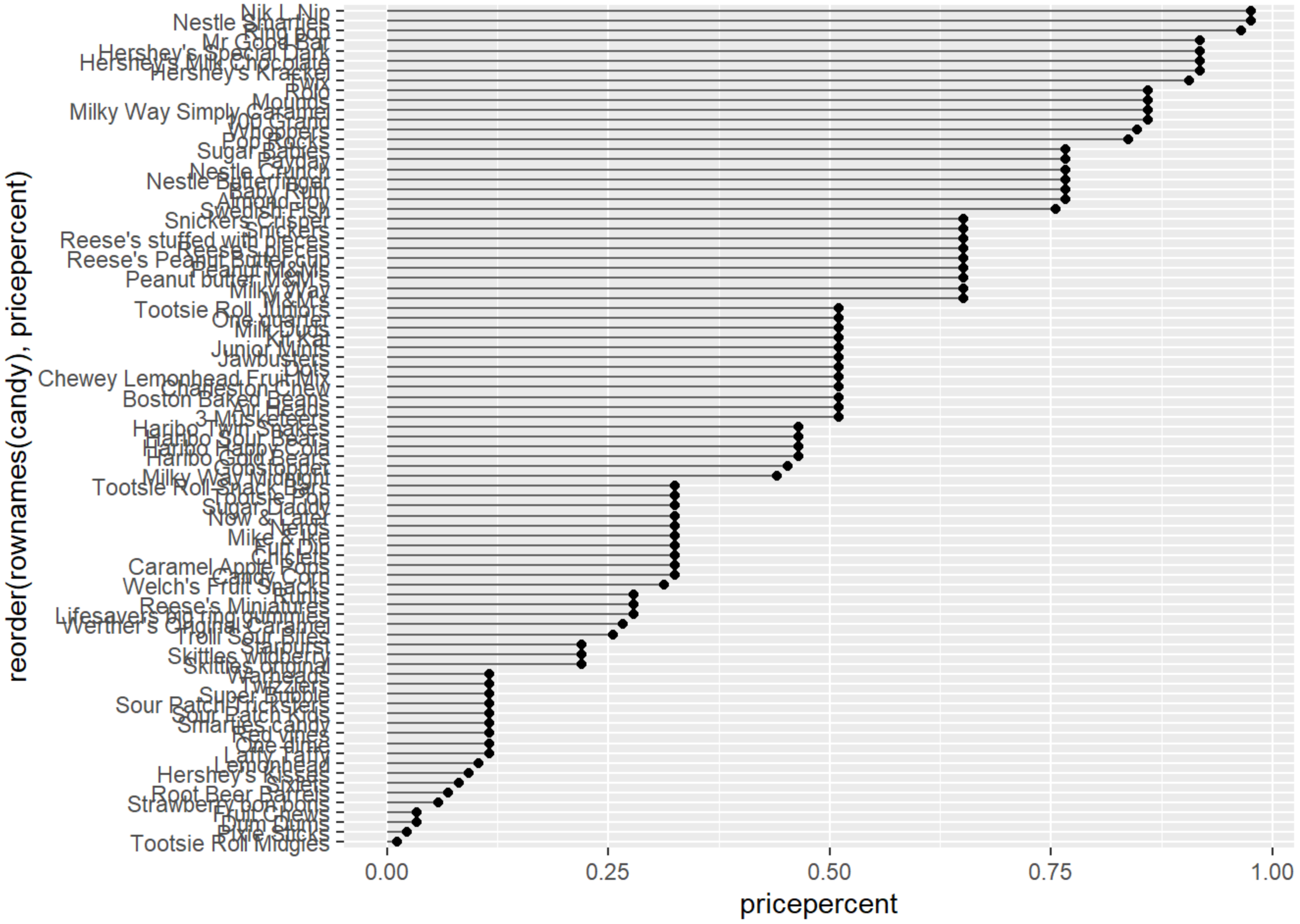
```
library(ggrepel)
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col = mycols) +
  geom_text_repel(col = mycols, max.overlaps = 10)
```

```
Warning: ggrepel: 29 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



> 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? Nik L Nip, Nestle Smarties, Ring Pop, Mr Good Bar, and Hershey's Krackel. Nik L Nip is the least popular.

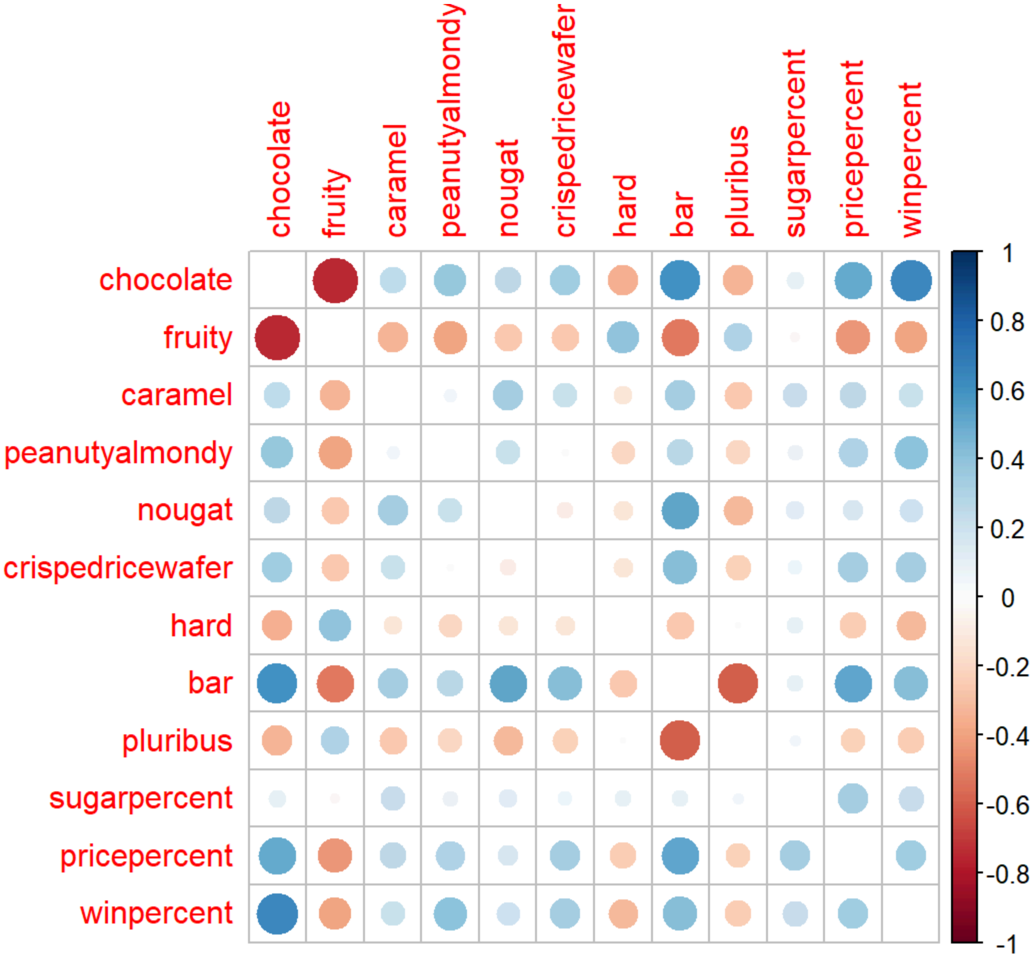
```r
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```



```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
cij <- cor(candy)
corrplot(cij, diag = F)
```

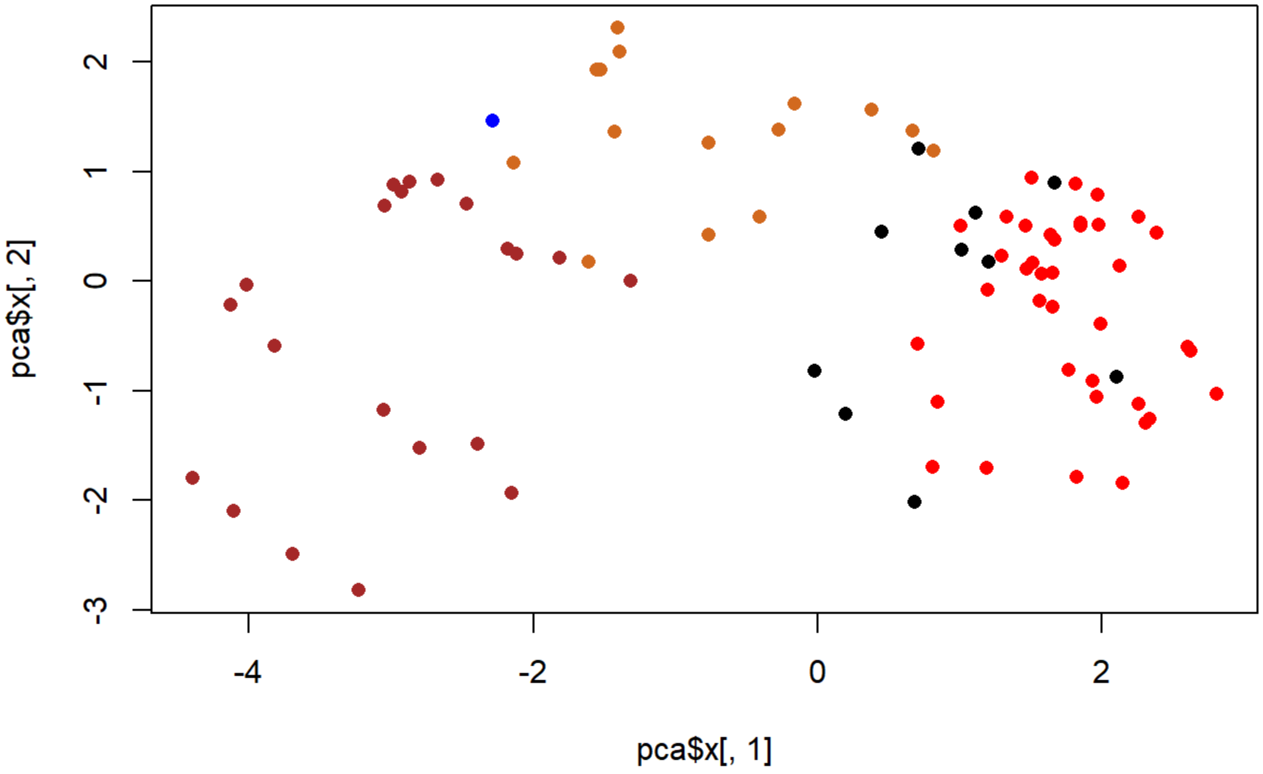


> Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)? Chocolate and fruity are the most anti-correlated, as they have biggest and darkest red circle at their intersection. Q23. Similarly, what two variables are most positively correlated? Chocolate and bar, as well as chocolate and win percent, seem to be the most positively correlated, as they have the biggest and darkest blue circles at their intersections.

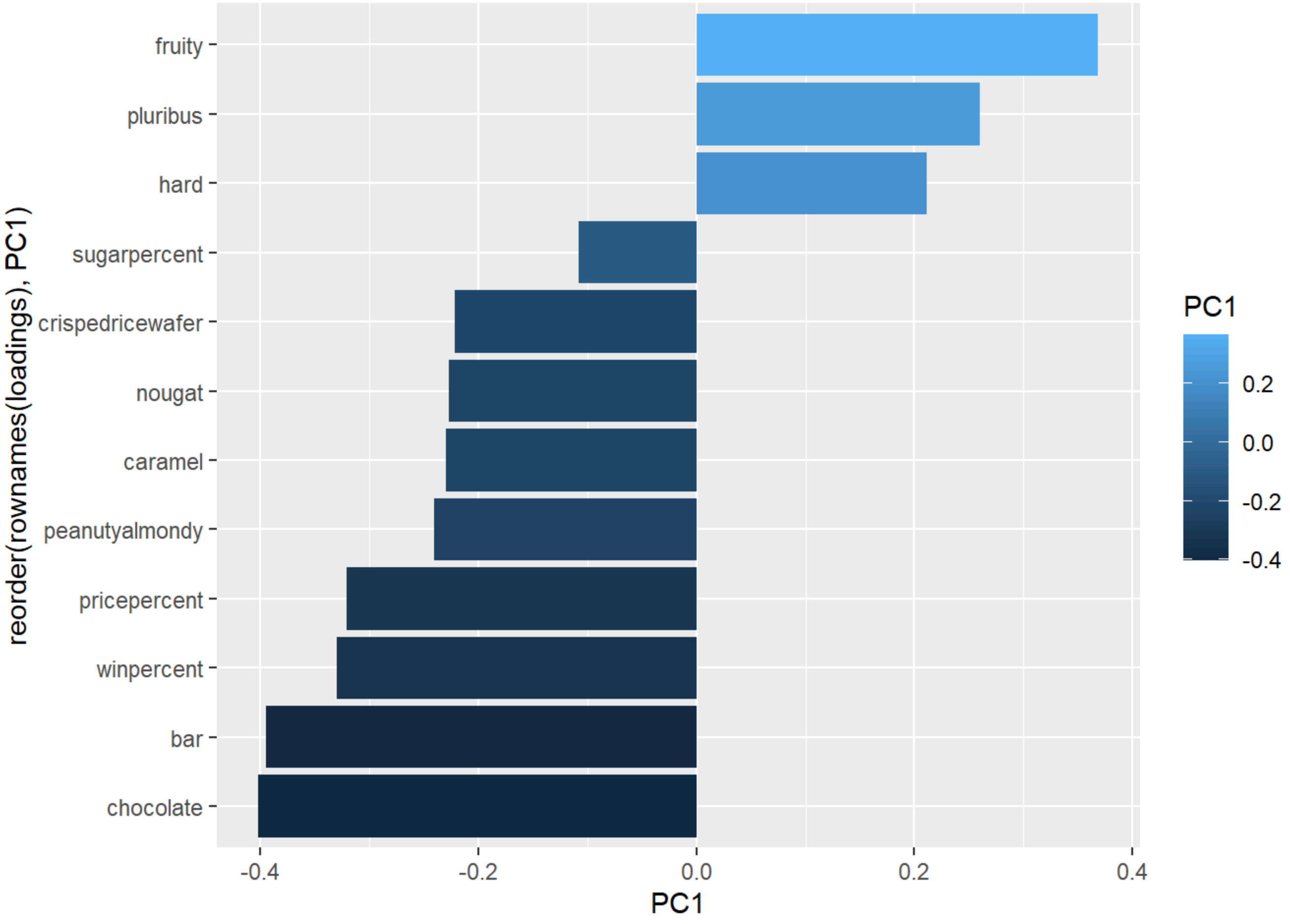
```r
pca <- prcomp(candy, scale = TRUE)
summary(pca)
```

```
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
```

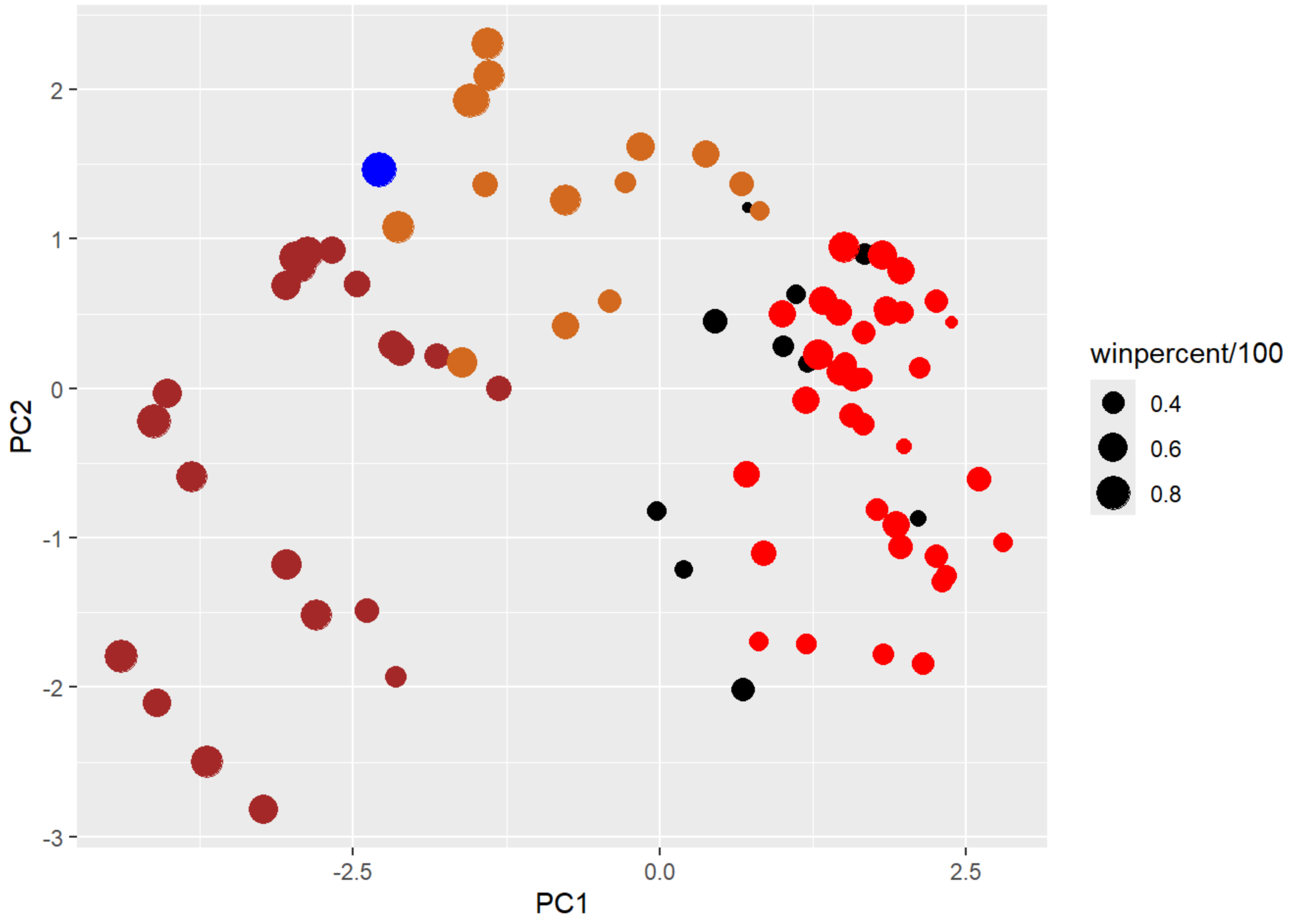
```r
plot(pca$x[,1], pca$x[,2], col = mycols, pch = 16)
```



```
loadings <- as.data.frame(pca$rotation)
ggplot(loadings) +
  aes(PC1, reorder(rownames(loadings), PC1), fill = PC1) +
  geom_col()
```

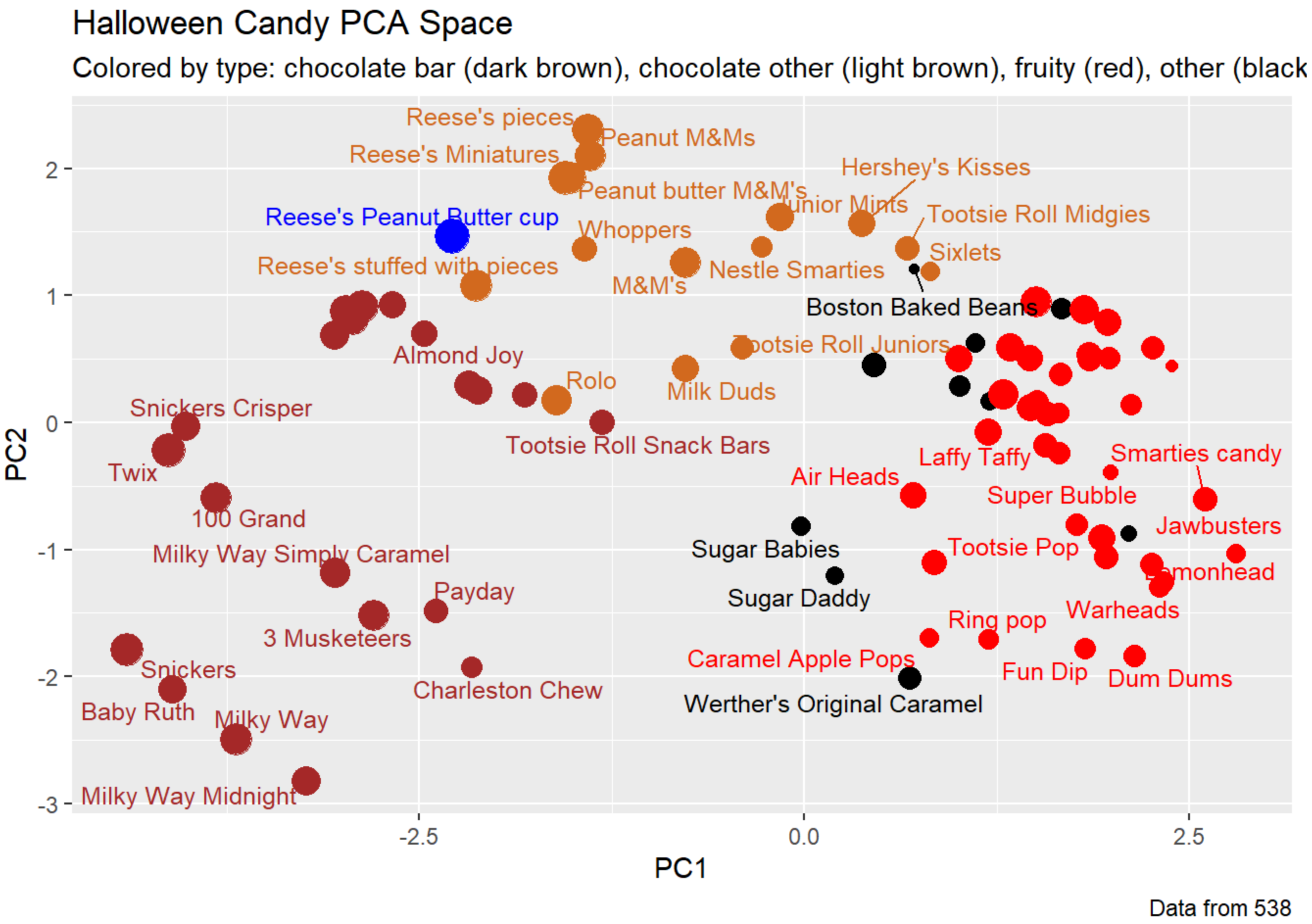


```r
my_data <- cbind(candy, pca$x[,1:3])
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=mycols)
p
```



```r
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), frui
       caption="Data from 538")
```

```
Warning: ggrepel: 40 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```


Halloween Candy PCA Space

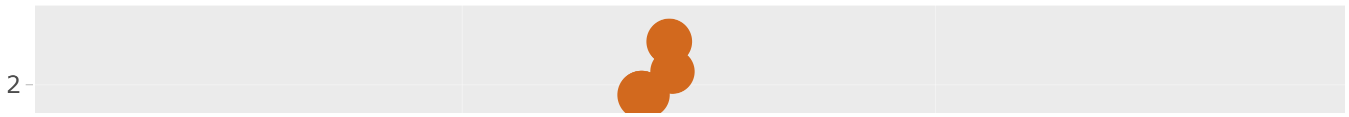
Colored by type: chocolate bar (dark brown), chocolate other (light brown), fruity (red), other (black

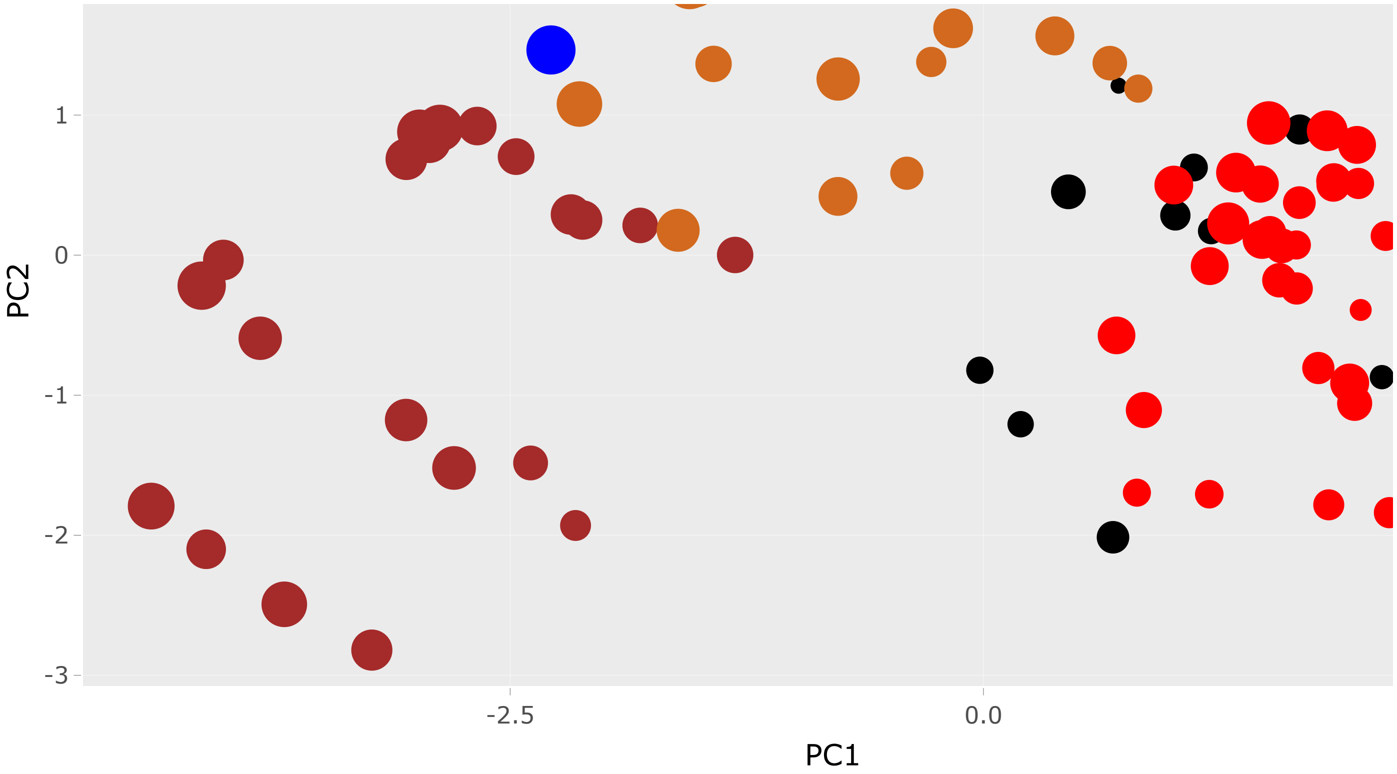
Data from 538

```r
library(plotly)
```

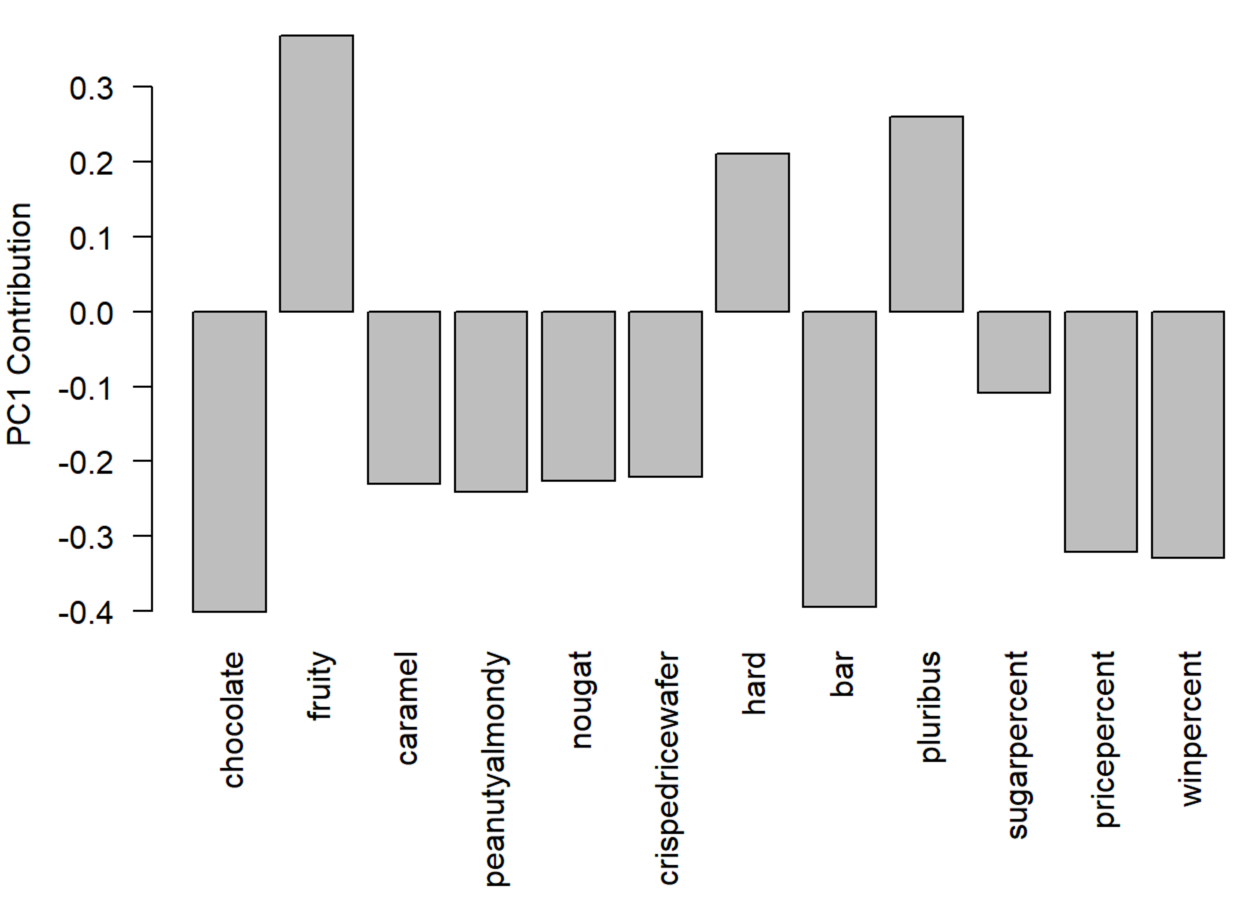
```
Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

    last_plot

The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
```

```r
ggplotly(p)
```





```r
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? The variables that are picked up strongly in the positive direction by PC1 are fruity, hard, and pluribus. This makes sense to me, because these three variables were correlated with each other and not very highly correlated with any other variables.