HalloweenMiniProject

Laura Biggs

Load in candy data

```
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? There are 12 different candy types in this dataset, signified by different columns.

```
dim(candy)
```

```
[1] 85 12
```

Q2. How many fruity candy types are in the dataset? There are 38 candies that fall into the fruity category.

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

```
[1] 38
```

Using winpercent

Q3. What is your favorite candy in the dataset and what is it's winpercent value? I like sour gummy worms, particularly the Trolli Sour Bites. The Trolli candy has a winpercent value of 47.17, meaning that people choose this candy over others less than 50% of the time.

```
candy["Trolli Sour Bites", ]$winpercent
```

```
[1] 47.17323
```

Q4. What is the winpercent value for "Kit Kat"? Kit Kat's are popular with a winpercent of 76.7.

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

```
[1] 76.7686
```

Q5. What is the winpercent value for "Tootsie Roll Snack Bars"? Tootsie roll snack bars are less popular with a winpercent of 49.6

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

```
[1] 49.6535
```

Using the skimr package

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

```
Warning: package 'skimr' was built under R version 4.1.3
```

```
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	
winpercent	0	1	50.32	14.71	22.45	39.14	47.83	59.86	84.18	

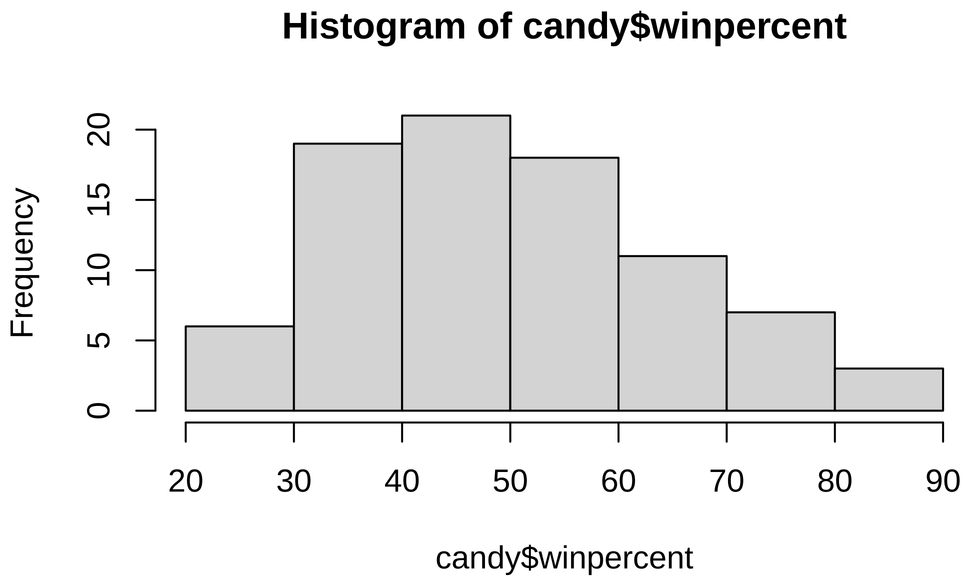
Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset? The winpercent variable is scaled differently relative to the other variables, and is scaled by 100 rather than 1.

Q7. What do you think a zero and one represent for the candy$chocolate column? A zero likely represents a candy that does not have chocolate, or otherwise answers false (=0) to the logical, while a one represents a candy that contains chocolate and answers true (=1) to the logical.

Plotting the data

Q8. Plot a histogram of winpercent values

```
hist(candy$winpercent)
```



Q9. Is the distribution of winpercent values symmetrical? The distribution isn't quite symmetrical or bell shaped, and is somewhat skewed as the frequency of observations cluster from 30%-50% rather than being evenly distributed around 50%.

Q10. Is the center of the distribution above or below 50%? The center of the distribution is below 50%, at 47.8%.

```
median(candy$winpercent)
```

```
[1] 47.82975
```

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

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

```
[1] 60.92153
```

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

```
[1] 44.11974
```

Q12. Is this difference statistically significant? This difference is statistically significant with a p value of 2.871e-08.

```r
t.test(chocolate, fruity)
```

```

	Welch Two Sample t-test

data:  chocolate and fruity
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 
```

Overall candy rankings

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

```r
library(dplyr)
```

```
Warning: package 'dplyr' was built under R version 4.1.3
```

```

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
```

Q14. What are the top 5 all time favorite candy types out of this set?

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

```
                           chocolate fruity caramel peanutyalmondy nougat
ReeseÃ•s Peanut Butter cup         1      0       0              1      0
ReeseÃ•s Miniatures                1      0       0              1      0
```

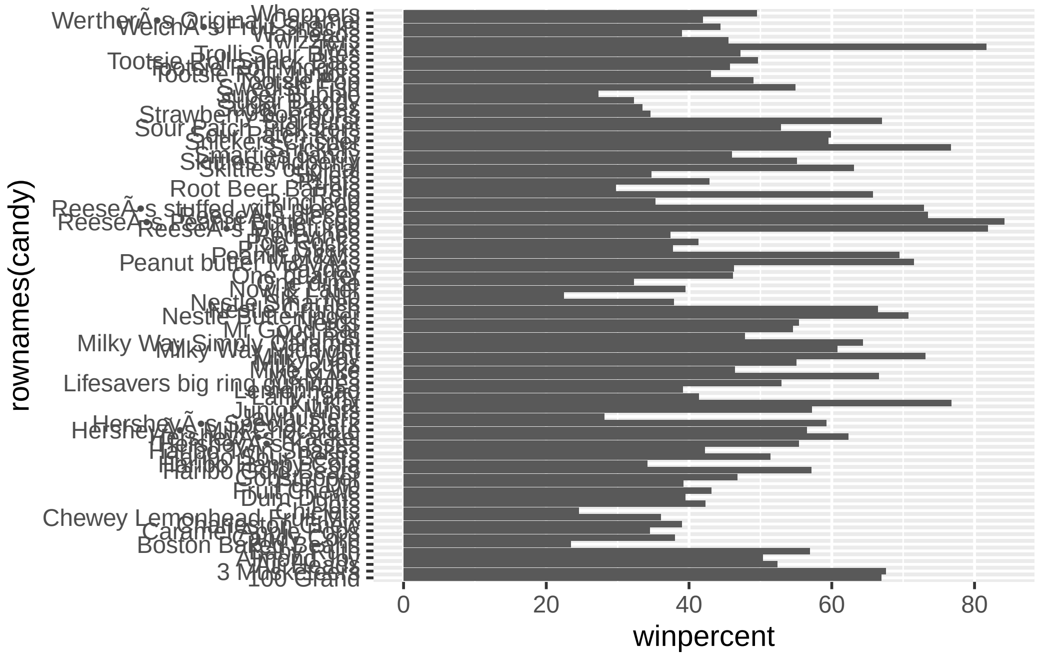
```
Twix                                1      0       1              0      0
Kit Kat                             1      0       0              0      0
Snickers                            1      0       1              1      1
                           crispedricewafer hard bar pluribus sugarpercent
ReeseÃ•s Peanut Butter cup               0    0   0        0        0.720
ReeseÃ•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
ReeseÃ•s Peanut Butter cup        0.651   84.18029
ReeseÃ•s Miniatures               0.279   81.86626
Twix                              0.906   81.64291
Kit Kat                           0.511   76.76860
Snickers                          0.651   76.67378
```

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

```
# with ggplot2
library(ggplot2)
```

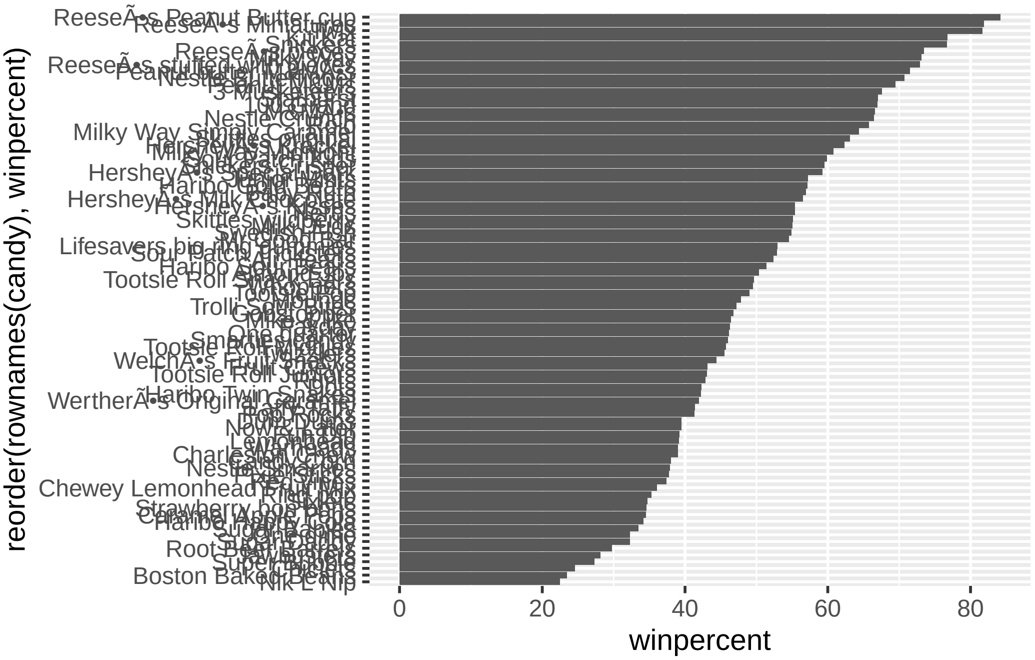
```
Warning: package 'ggplot2' was built under R version 4.1.3
```

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



Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

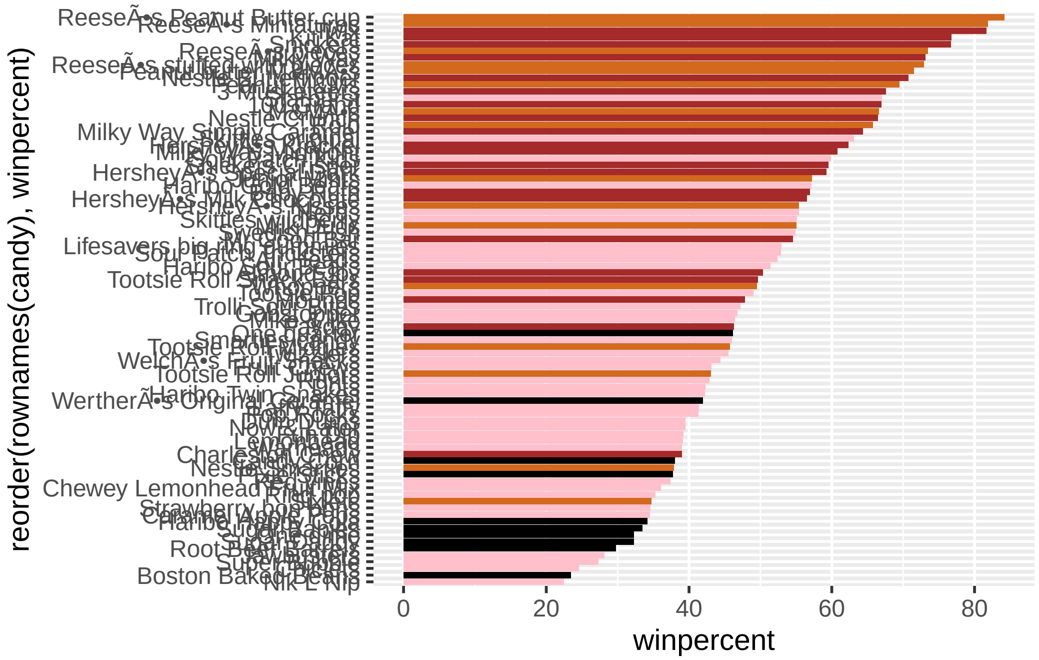
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent)) +
  geom_col()
```



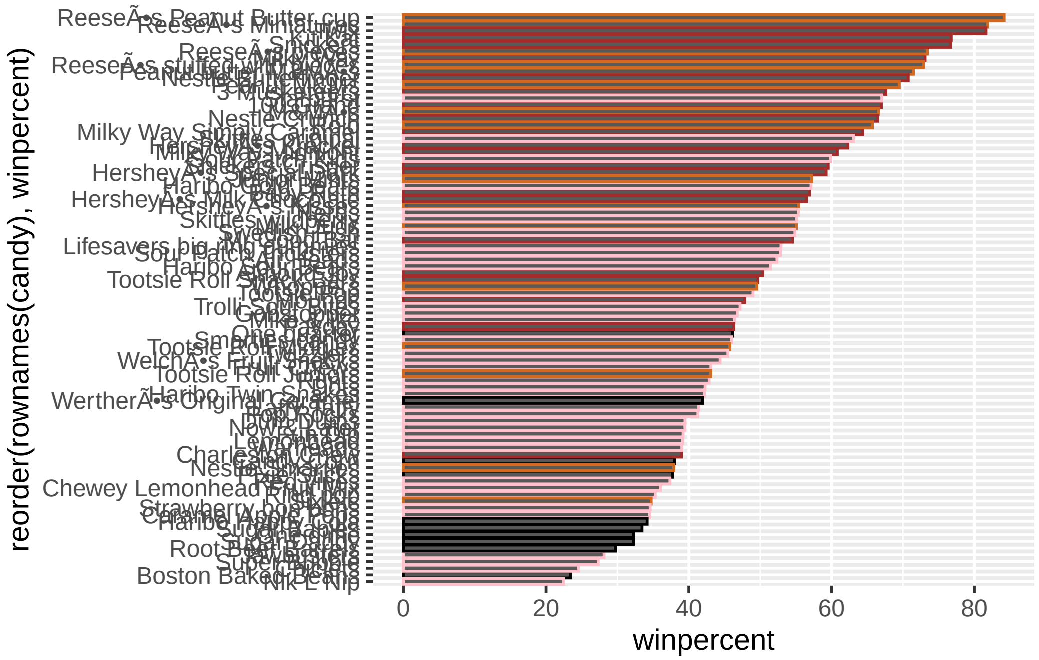
Add color to barplot

```
# Make color vectors
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"
```

```
#Add to barplot
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent)) +
  geom_col(fill=my_cols)
```



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



Q17. What is the worst ranked chocolate candy? Sixlets are the worst ranked chocolate candy.

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

Pricepercent

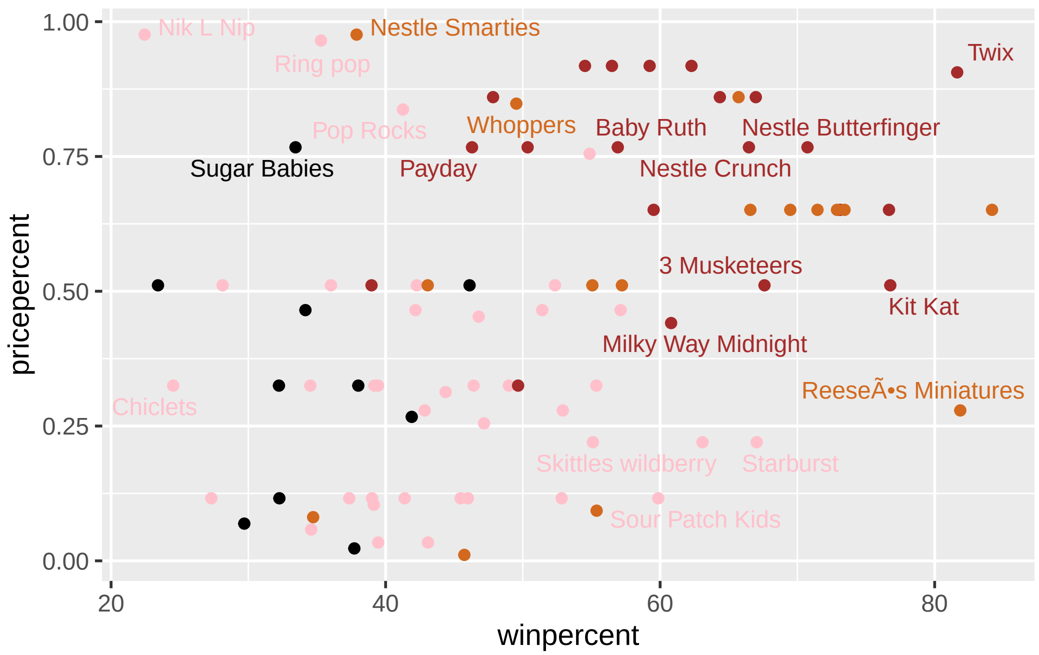
Plot pricepercent vs winpercent

```
library(ggrepel)
```

```
Warning: package 'ggrepel' was built under R version 4.1.3
```

```
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 5)
```

```
Warning: ggrepel: 66 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? Tootsie roll midgies are the highest ranked, 45.7%, for the lowest cost relative to other candies.

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

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

Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular? The 5 most expensive candies are Nik L Nip, Nestle Smarties, Ring pop, Hersheys Krackel, Hersheys Milk Chocolate. Nik L Nip 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
HersheyÃ•s Krackel                     0.918   62.28448
HersheyÃ•s Milk Chocolate              0.918   56.49050
```

Corelation Structure

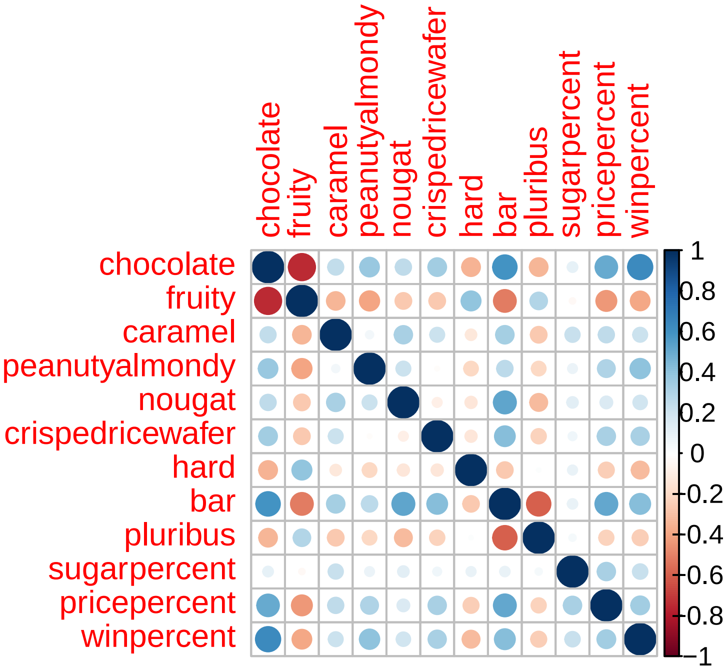
Using the corrplot package

```
library(corrplot)
```

```
Warning: package 'corrplot' was built under R version 4.1.3
```

```
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 fruity are anti-correlated.

Q23. Similarly, what two variables are most positively correlated? Chocolate and bar are most positively correlated.

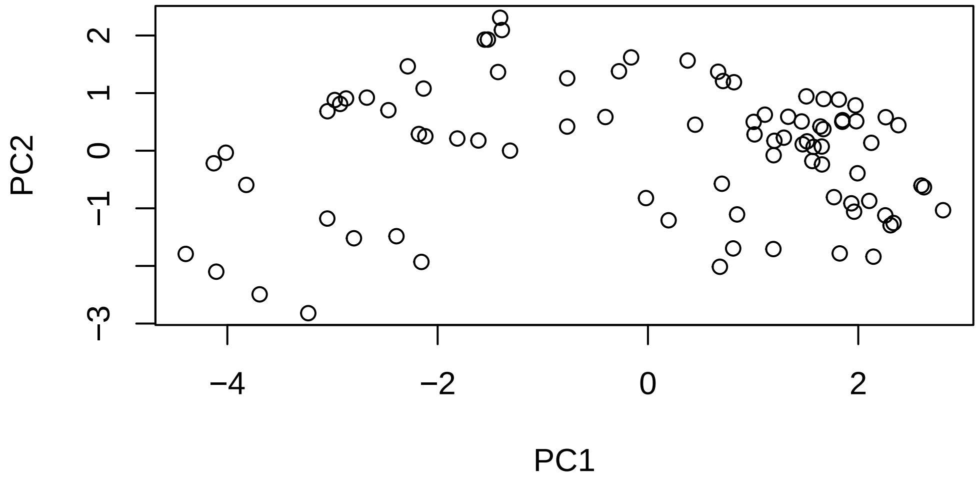
PCA

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

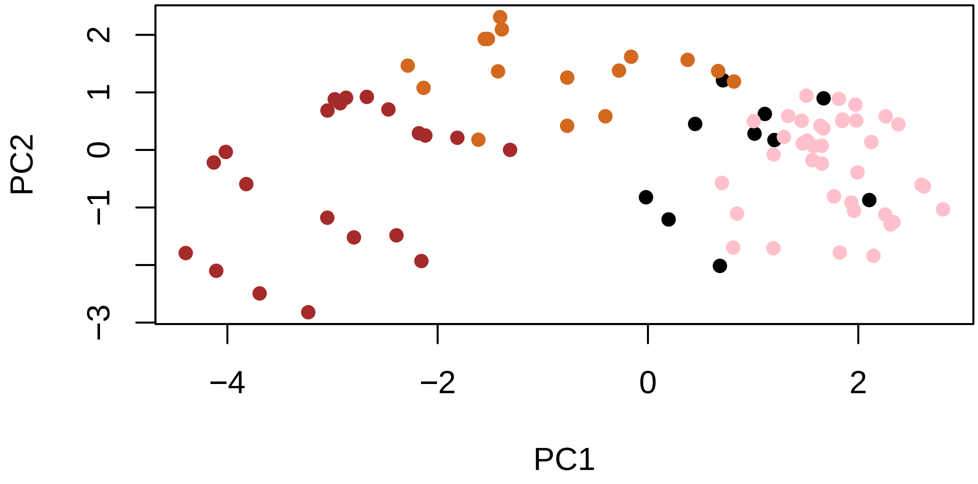
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 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 plot

```
# x plots scores, not rotation
plot(pca$x[,1:2])
```



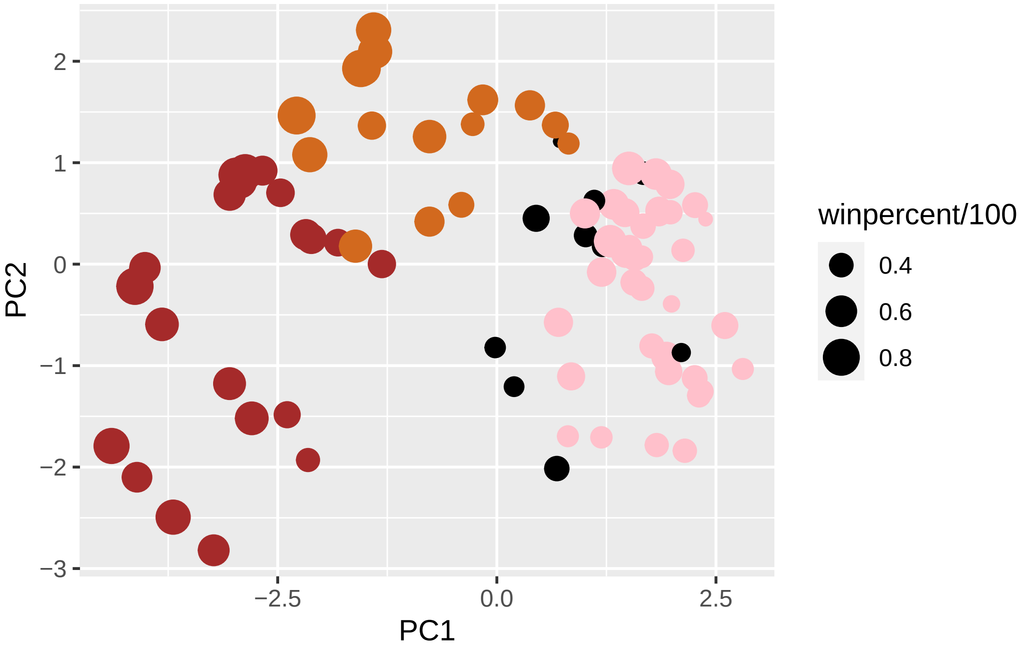
```
# add color
plot(pca$x[,1:2], col=my_cols, pch=16)
```



ggplot PCA

```
#Make a new df with PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
```

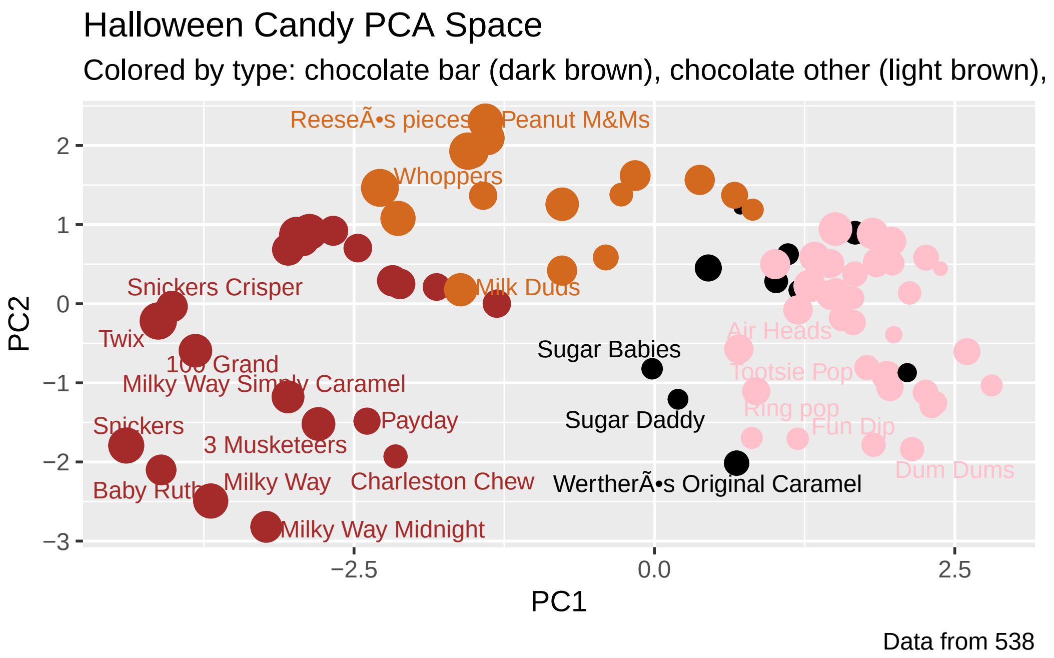
```
p <- ggplot(my_data) +
      aes(x=PC1, y=PC2,
          size=winpercent/100,
          text=rownames(my_data),
          label=rownames(my_data)) +
  geom_point(col=my_cols)

p
```



Using ggrepel

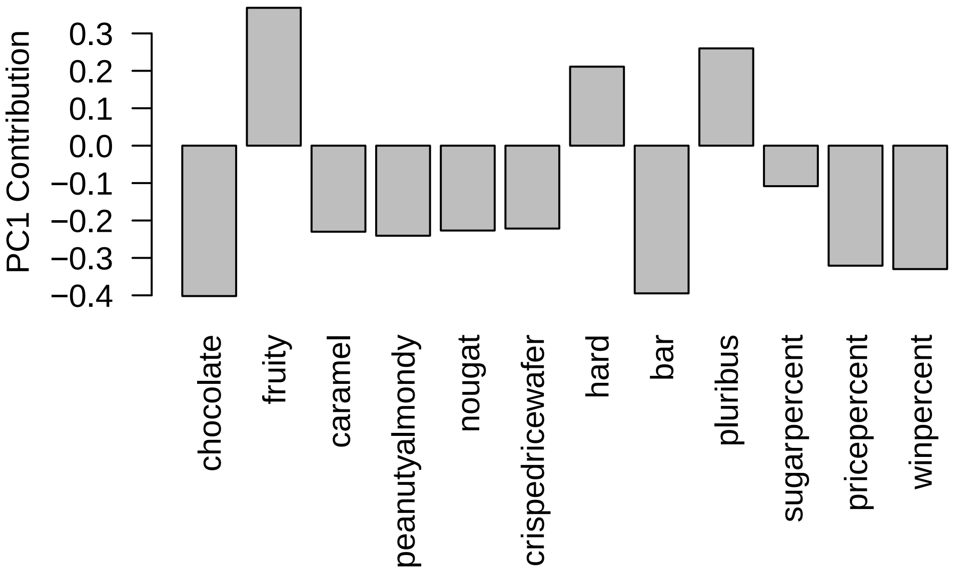
```
p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 7)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown
       caption="Data from 538")
```

```
Warning: ggrepel: 62 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



PCA by loading

```
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? PC1 resolves the variance between fruity candy (positive direction) and chocolate/bars (negative direction). Looking at the PCA plot we can see PC1 separates the fruity candy (pink) from the chocolate bars (dark brown) well. Pluribus also contributes to PC1 in the positive direction and is associated with the fruity candy type as there are many candies in a package as opposed to chocolate bars where this is only 1 candy inside.