Class9

Sabrina Koldinger (A16368238)

Halloween Mini-Project

Here we analyze a candy dataset from the 538 website.

```
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?

```
sum(nrow(candy))
```

```
[1] 85
```

There are 85 different types.

Q2. How many fruity candy types are in the dataset?

```
sum(candy[,"fruity"])
```

```
[1] 38
```

There are 37 fruity types.

Favorite Candy

Q3. What is your favorite candy in the dataset and what is it's winpercent value?

```
candy["Swedish Fish", "winpercent"]
```

```
[1] 54.86111
```

Swedish Fish have a winpercent of 54.9%

Q4. What is the winpercent value for "Kit Kat"?

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

```
[1] 76.7686
```

Kit Kat win percent = 76.8%.

Q5. What is the winpercent value for "Tootsie Roll Snack Bars"?

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

```
[1] 49.6535
```

Tootsie roll win percent = 49.6%.

What is the least liked candy?

```
inds=order(candy[,"winpercent"])
head(candy[inds,])
```

```
                  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
Root Beer Barrels         0      0       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
Root Beer Barrels                0    1   0        1        0.732        0.069
                  winpercent
Nik L Nip           22.44534
Boston Baked Beans  23.41782
Chiclets            24.52499
Super Bubble        27.30386
Jawbusters          28.12744
Root Beer Barrels   29.70369
```

The least liked candy is Nik L Nip.

```
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	
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	

```
skimr::skim(candy)
```

Table 3: 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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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?

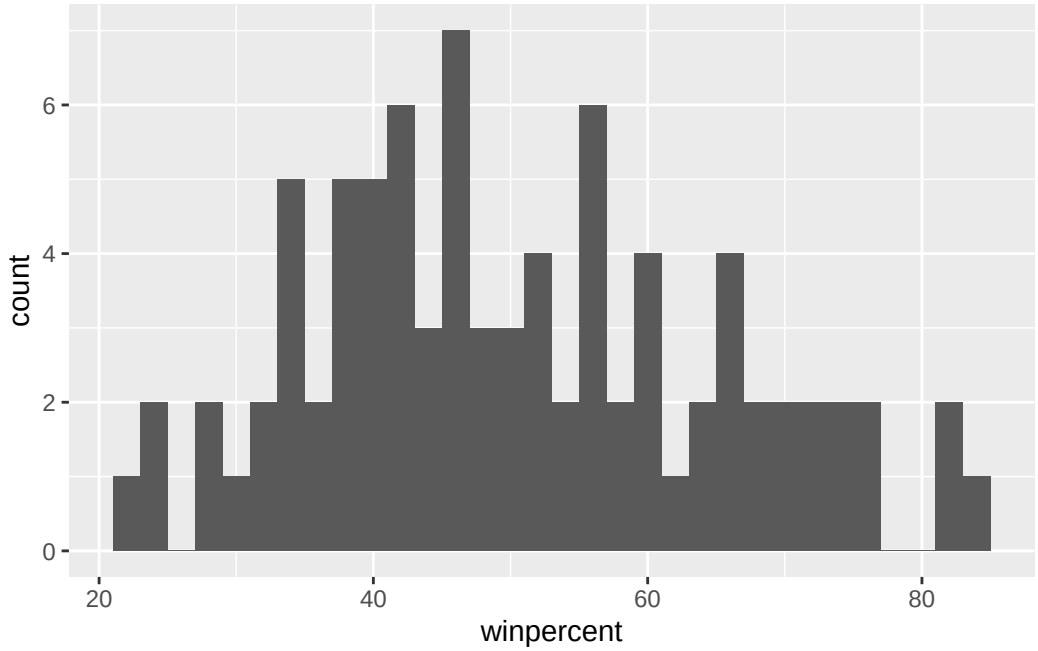
The win percent is on a different scale.

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

It represents whether the candy is chocolate or not.

Q8. Plot a histogram of winpercent values

```
library(ggplot2)
ggplot(candy)+aes(x=winpercent,)+ geom_histogram(binwidth=2)
```



Q9. Is the distribution of winpercent values symmetrical?

They are not symmetrical.

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

The center is below 50%

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

```
TFchocolate=as.logical(candy[,"chocolate"])
candy[TFchocolate,]$winpercent
```

```
 [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
```

```
mean(candy[TFchocolate,]$winpercent)
```

```
[1] 60.92153
```

```
TFfruity=as.logical(candy[,"fruity"])
mean(candy[TFfruity,]$winpercent)
```

```
[1] 44.11974
```

Chocolate candy on average is ranked higher.

Q12. Is this difference statistically significant?

```
t.test(candy[TFchocolate,]$winpercent, candy[TFfruity,]$winpercent)
```

```

    Welch Two Sample t-test

data:  candy[TFchocolate, ]$winpercent and candy[TFfruity, ]$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
```

The p-value is small, so it is statistically signficant.

Ranking

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

```
inds=order(candy[,"winpercent"])
head(candy[inds,])
```

```
                  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
Root Beer Barrels         0      0       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
Root Beer Barrels                0    1   0        1        0.732        0.069
                  winpercent
Nik L Nip           22.44534
Boston Baked Beans  23.41782
Chiclets            24.52499
Super Bubble        27.30386
Jawbusters          28.12744
Root Beer Barrels   29.70369
```

The least liked candy types are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters.

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

```
inds=rev(order(candy[,"winpercent"]))
head(candy[inds,])
```

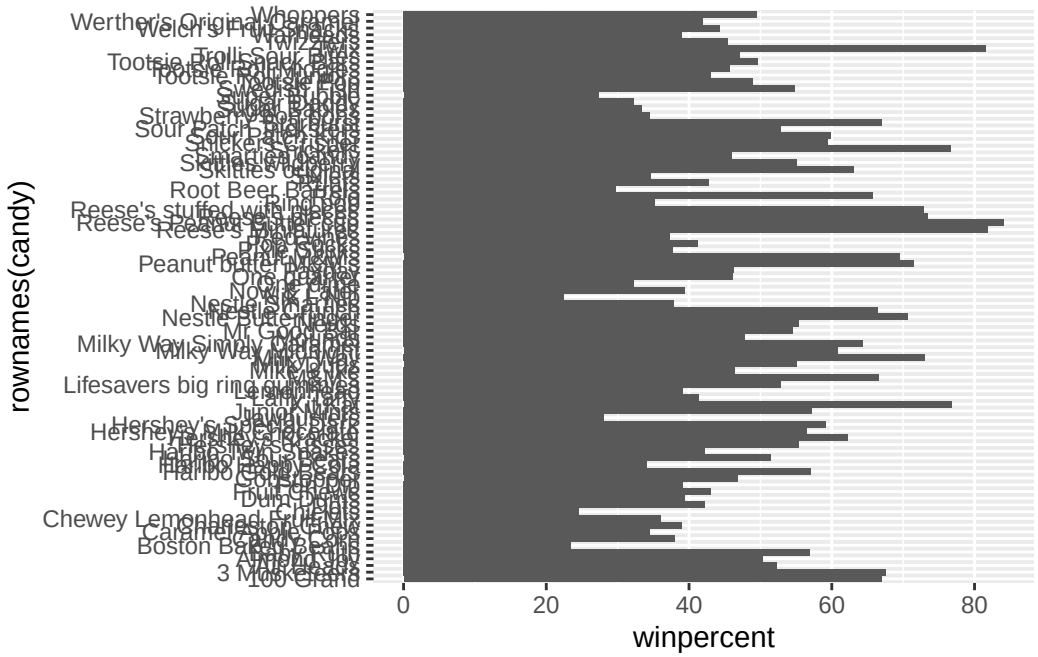
```
                        chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
```

```
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
Reese's pieces                    1      0       0              1      0
                          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
Reese's pieces                            0    0   0        1        0.406
                          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
Reese's pieces                   0.651   73.43499
```

The mosted liked Reese's peanut butter cup, Reese's Miniatures, Twix, Kit Kat, and snickers.

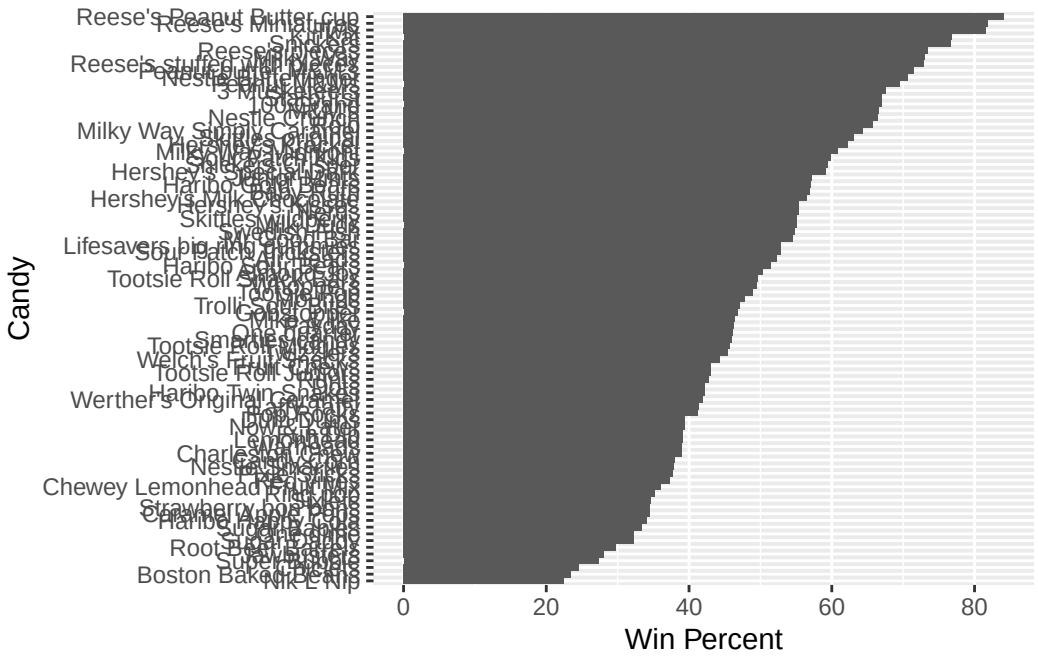
Q15. Make a first barplot of candy ranking based on winpercent values.

```
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() +
  labs(x="Win Percent", y="Candy")
```




```
ggsave('barplot1.png', width=7, height=10)
```

You can insert any image using the markdown syntax

Add some color to our ggplot.

```
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_co
```

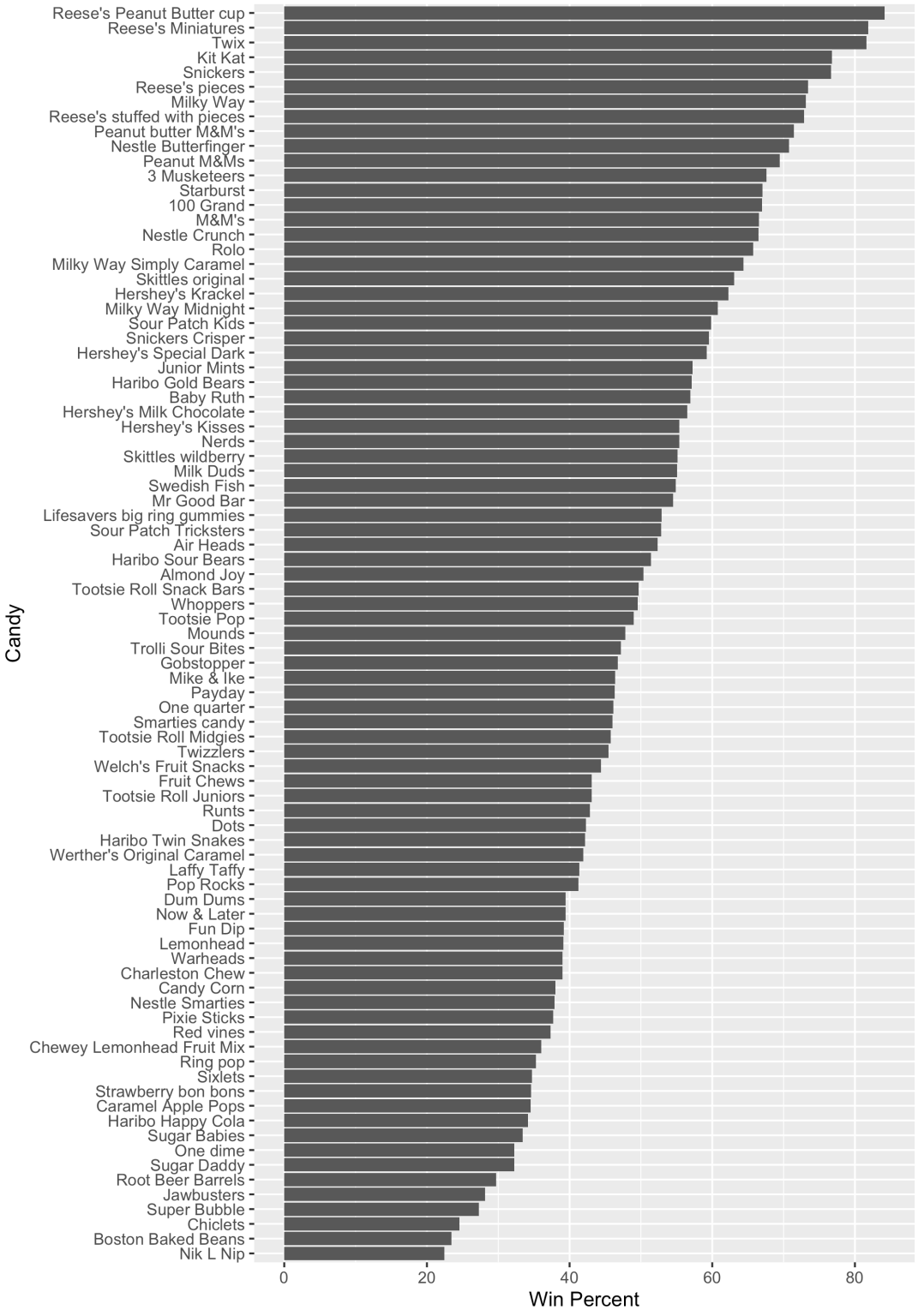
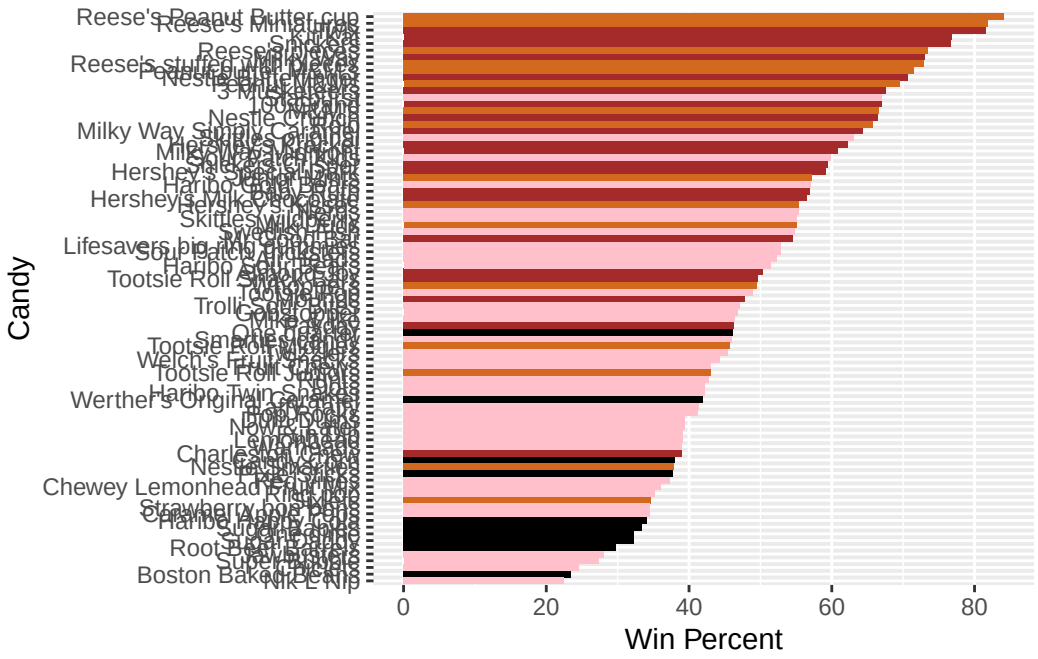


Figure 1: An example of photo insertion



Q17. What is the worst ranked chocolate candy?

Sixlet is ranked the worst.

Q18. What is the best ranked fruity candy?

Starburst is the best.

Pricepercent

If we want to see what is a good candy to buy in terks of minpercent and pricepercent we can plot these two vairables and then see the best candy for the least amount of money.

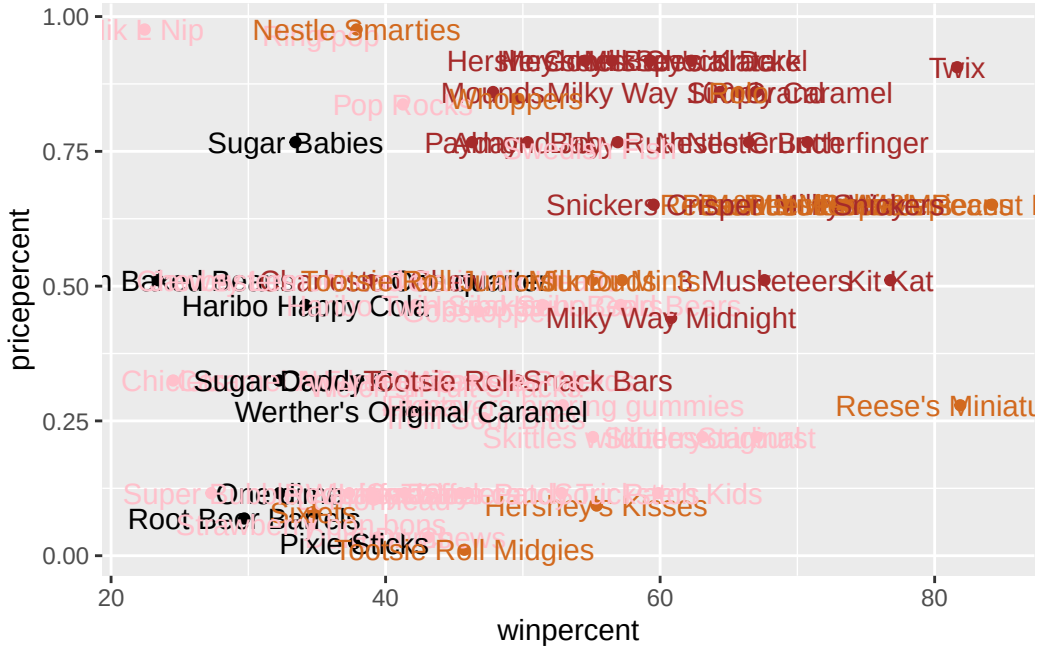
```
candy$pricepercent
```

```
 [1] 0.860 0.511 0.116 0.511 0.511 0.767 0.767 0.511 0.325 0.325 0.511 0.511
[13] 0.325 0.511 0.034 0.034 0.325 0.453 0.465 0.465 0.465 0.465 0.093 0.918
[25] 0.918 0.918 0.511 0.511 0.511 0.116 0.104 0.279 0.651 0.651 0.325 0.511
[37] 0.651 0.441 0.860 0.860 0.918 0.325 0.767 0.767 0.976 0.325 0.767 0.651
```

```
[49] 0.023 0.837 0.116 0.279 0.651 0.651 0.651 0.965 0.860 0.069 0.279 0.081
[61] 0.220 0.220 0.976 0.116 0.651 0.651 0.116 0.116 0.220 0.058 0.767 0.325
[73] 0.116 0.755 0.325 0.511 0.011 0.325 0.255 0.906 0.116 0.116 0.313 0.267
[85] 0.848
```

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?

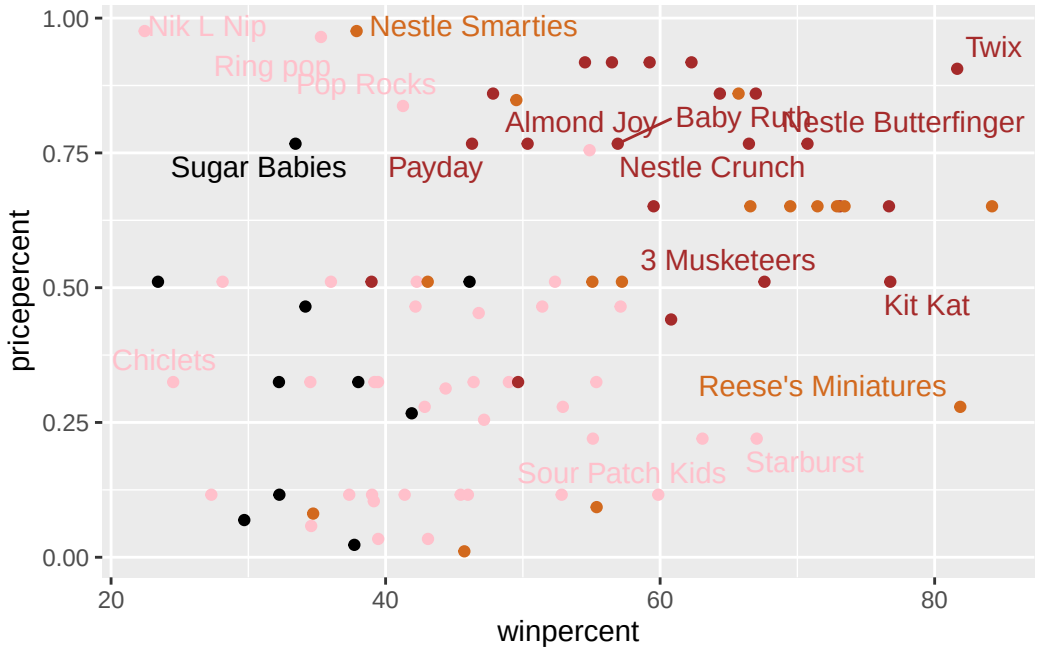
```
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) + geom_text(col=my_cols)
```



To avoid the overplotting of all these labels we can use an add on package ggrepl.

```
library(ggrepel
        )
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) + geom_text_repel(col=my_cols, max.overlaps =7)
```

```
Warning: ggrepel: 68 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Reese's miniatures is best ranked for the least amount.

Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular?

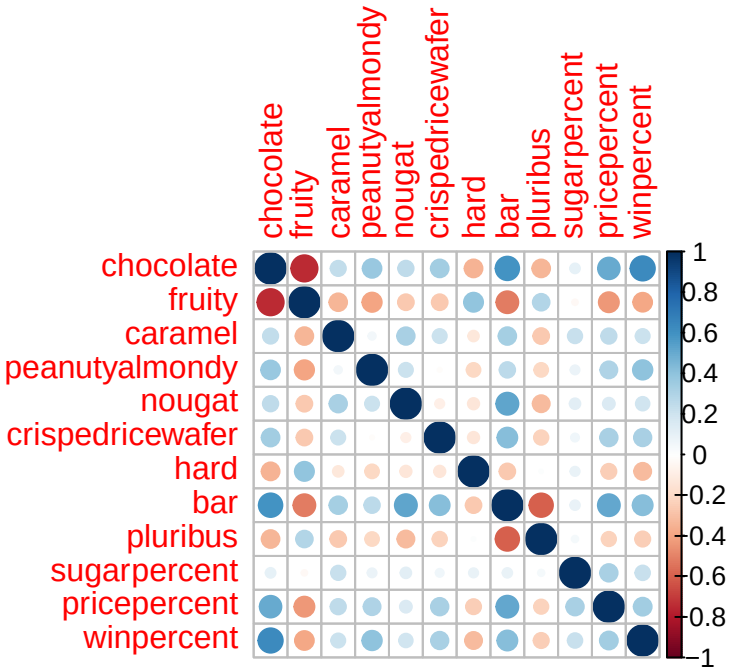
The most expensive is Nik L Nip, Ring Pop, Nestle Smarties, Mr. Good Bar, and Hershey's Milk Chocolate. The least popular is Nik L Nip.

Correlation

```
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 fruity

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

Chocolate and bar or chocolate and winpercent

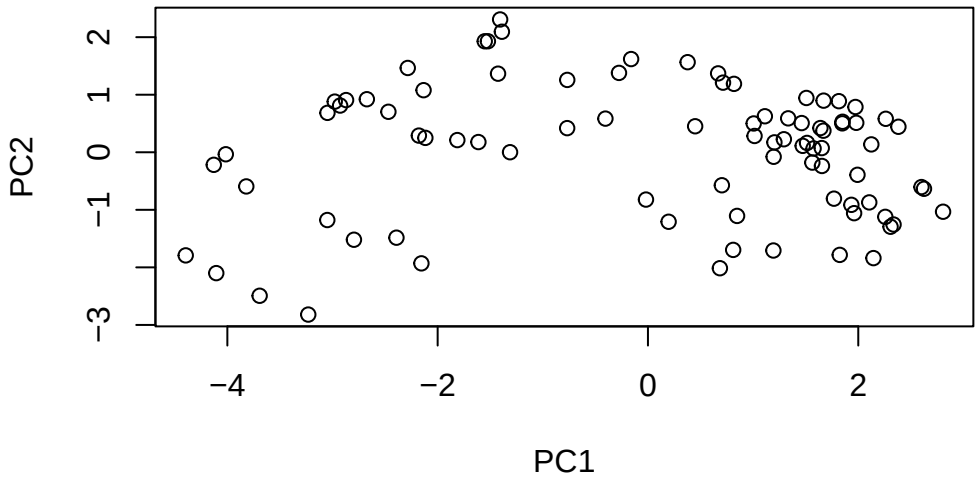
PCA

The main function for this is `precom()` and here we know we need to scale our data with `scale=TRUE` argument.

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

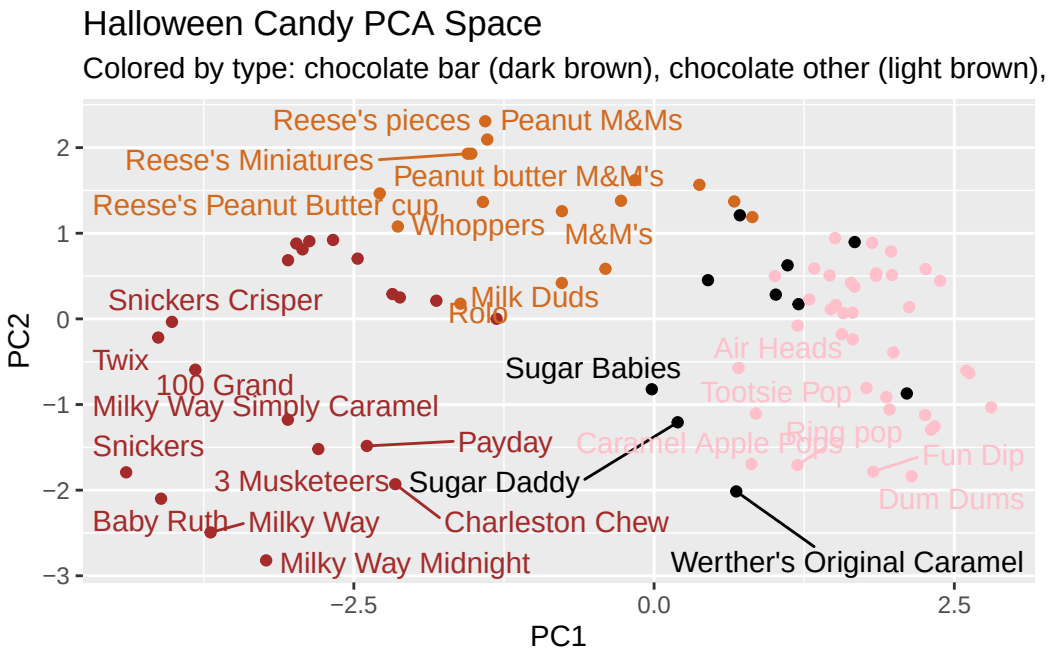
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688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

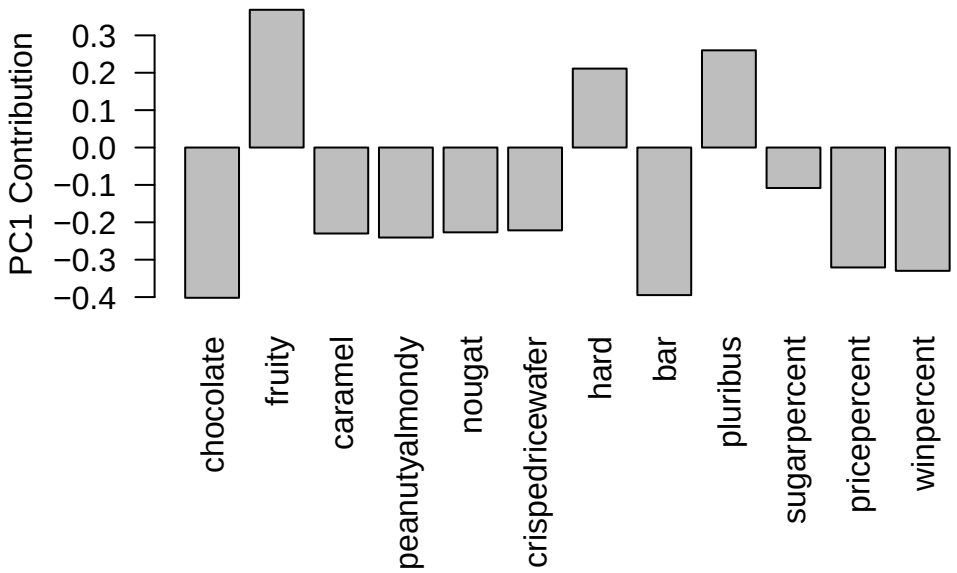



```r
library(ggrepel)
my_data <- cbind(candy, pca$x[,1:3])
ggplot(my_data) +
        aes(x=PC1, y=PC2,label=rownames(my_data)) +
        geom_point(col=my_cols) + geom_text_repel(col=my_cols)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown
```

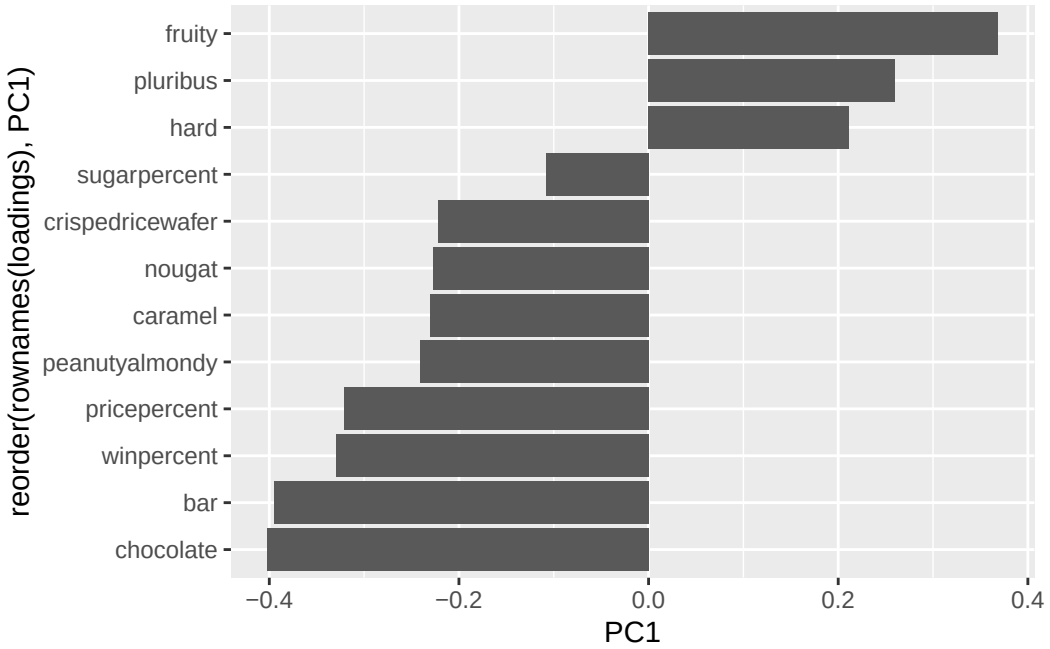
```
Warning: ggrepel: 56 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




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




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



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

The variables are fruity, pluribus, and hard are in the positive direction. These do make sense since fruity candy tends to have those two characteristics.