

Mini Candy Project

Ebru Robinson

Table of contents

Winpercent vs Pricepercent	15
--------------------------------------	----

```
read.csv("candy-data.csv")
```

competitor- name	choco- late	fruity	caram- el	peanut- butter	yal- nougat	crisped wafer	rice- hard	bar	pluri-	sug- arper- cent	pri- ceper- cent	win- per- cent
100 Grand	1	0	1	0	0	1	0	1	0	0.732	0.860	66.97173
3 Musketeers	1	0	0	0	1	0	0	1	0	0.604	0.511	67.60294
One dime	0	0	0	0	0	0	0	0	0	0.011	0.116	32.26109
One quarter	0	0	0	0	0	0	0	0	0	0.011	0.511	46.11650
Air Heads	0	1	0	0	0	0	0	0	0	0.906	0.511	52.34146
Almond Joy	1	0	0	1	0	0	0	1	0	0.465	0.767	50.34755
Baby Ruth	1	0	1	1	1	0	0	1	0	0.604	0.767	56.91455
Boston	0	0	0	1	0	0	0	0	1	0.313	0.511	23.41782
Baked Beans												
Candy Corn	0	0	0	0	0	0	0	0	1	0.906	0.325	38.01096
Caramel	0	1	1	0	0	0	0	0	0	0.604	0.325	34.51768
Apple Pops												
Charleston Chew	1	0	0	0	1	0	0	1	0	0.604	0.511	38.97504
Chewy Lemonhead	0	1	0	0	0	0	0	0	1	0.732	0.511	36.01763
Fruit Mix												
Chiclets	0	1	0	0	0	0	0	0	1	0.046	0.325	24.52499
Dots	0	1	0	0	0	0	0	0	1	0.732	0.511	42.27208

competitor- name	choco- late	fruity	caram- chocolate	peanutyal- mondy	nougat	crispedrice- wafer	hardbar	pluribuscent	sug- arper- cent	pri- ceper- cent	win- per- cent	
Dum Dums	0	1	0	0	0	0	1	0	0	0.732	0.034	39.46056
Fruit Chews	0	1	0	0	0	0	0	0	1	0.127	0.034	43.08892
Fun Dip	0	1	0	0	0	0	1	0	0	0.732	0.325	39.18550
Gobstopper	0	1	0	0	0	0	1	0	1	0.906	0.453	46.78335
Haribo	0	1	0	0	0	0	0	0	1	0.465	0.465	57.11974
Gold Bears												
Haribo	0	0	0	0	0	0	0	0	1	0.465	0.465	34.15896
Happy Cola												
Haribo Sour Bears	0	1	0	0	0	0	0	0	1	0.465	0.465	51.41243
Haribo	0	1	0	0	0	0	0	0	1	0.465	0.465	42.17877
Twin Snakes												
Hershey's Kisses	1	0	0	0	0	0	0	0	1	0.127	0.093	55.37545
Hershey's Krackel	1	0	0	0	0	1	0	1	0	0.430	0.918	62.28448
Hershey's Milk Chocolate	1	0	0	0	0	0	0	1	0	0.430	0.918	56.49050
Hershey's Special Dark	1	0	0	0	0	0	0	1	0	0.430	0.918	59.23612
Jawbusters	0	1	0	0	0	0	1	0	1	0.093	0.511	28.12744
Junior	1	0	0	0	0	0	0	0	1	0.197	0.511	57.21925
Mints												
Kit Kat	1	0	0	0	0	1	0	1	0	0.313	0.511	76.76860
Laffy Taffy	0	1	0	0	0	0	0	0	0	0.220	0.116	41.38956
Lemonhead	0	1	0	0	0	0	1	0	0	0.046	0.104	39.14106
Lifesavers	0	1	0	0	0	0	0	0	0	0.267	0.279	52.91139
big ring gummies												
Peanut butter	1	0	0	1	0	0	0	0	1	0.825	0.651	71.46505
M&M's												
M&M's	1	0	0	0	0	0	0	0	1	0.825	0.651	66.57458
Mike & Ike	0	1	0	0	0	0	0	0	1	0.872	0.325	46.41172
Milk Duds	1	0	1	0	0	0	0	0	1	0.302	0.511	55.06407
Milky Way	1	0	1	0	1	0	0	1	0	0.604	0.651	73.09956

competitor- name	choco- late	fruity	caramel	peanutyal- mondy	nougat	crispedrice- wafer	hardbar	pluribuscent	sug- arper- cent	pri- ceper- cent	win- per- cent	
Milky Way	1	0	1	0	1	0	0	1	0	0.732	0.441	60.80070
Midnight												
Milky Way	1	0	1	0	0	0	0	1	0	0.965	0.860	64.35334
Simply												
Caramel												
Mounds	1	0	0	0	0	0	0	1	0	0.313	0.860	47.82975
Mr Good	1	0	0	1	0	0	0	1	0	0.313	0.918	54.52645
Bar												
Nerds	0	1	0	0	0	0	1	0	1	0.848	0.325	55.35405
Nestle	1	0	0	1	0	0	0	1	0	0.604	0.767	70.73564
Butterfinger												
Nestle	1	0	0	0	0	1	0	1	0	0.313	0.767	66.47068
Crunch												
Nik L Nip	0	1	0	0	0	0	0	0	1	0.197	0.976	22.44534
Now &	0	1	0	0	0	0	0	0	1	0.220	0.325	39.44680
Later												
Payday	0	0	0	1	1	0	0	1	0	0.465	0.767	46.29660
Peanut	1	0	0	1	0	0	0	0	1	0.593	0.651	69.48379
M&Ms												
Pixie Sticks	0	0	0	0	0	0	0	0	1	0.093	0.023	37.72234
Pop Rocks	0	1	0	0	0	0	1	0	1	0.604	0.837	41.26551
Red vines	0	1	0	0	0	0	0	0	1	0.581	0.116	37.34852
Reese's	1	0	0	1	0	0	0	0	0	0.034	0.279	81.86626
Miniatures												
Reese's	1	0	0	1	0	0	0	0	0	0.720	0.651	84.18029
Peanut												
Butter cup												
Reese's	1	0	0	1	0	0	0	0	1	0.406	0.651	73.43499
pieces												
Reese's	1	0	0	1	0	0	0	0	0	0.988	0.651	72.88790
stuffed with												
pieces												
Ring pop	0	1	0	0	0	0	1	0	0	0.732	0.965	35.29076
Rolo	1	0	1	0	0	0	0	0	1	0.860	0.860	65.71629
Root Beer	0	0	0	0	0	0	1	0	1	0.732	0.069	29.70369
Barrels												
Runts	0	1	0	0	0	0	1	0	1	0.872	0.279	42.84914
Sixlets	1	0	0	0	0	0	0	0	1	0.220	0.081	34.72200

competitor- name	choco- late	fruity	caram- chondy	peanutyal- nougat	crispedrice- wafer	hardbar	pluribuscent	sug- arper- cent	pri- ceper- cent	win- per- cent		
Skittles original	0	1	0	0	0	0	0	0	1	0.941	0.220	63.08514
Skittles wildberry	0	1	0	0	0	0	0	0	1	0.941	0.220	55.10370
Nestle Smarties	1	0	0	0	0	0	0	0	1	0.267	0.976	37.88719
Smarties candy	0	1	0	0	0	0	1	0	1	0.267	0.116	45.99583
Snickers	1	0	1	1	1	0	0	1	0	0.546	0.651	76.67378
Snickers Crisper	1	0	1	1	0	1	0	1	0	0.604	0.651	59.52925
Sour Patch Kids	0	1	0	0	0	0	0	0	1	0.069	0.116	59.86400
Sour Patch Tricksters	0	1	0	0	0	0	0	0	1	0.069	0.116	52.82595
Starburst	0	1	0	0	0	0	0	0	1	0.151	0.220	67.03763
Strawberry bon bons	0	1	0	0	0	0	1	0	1	0.569	0.058	34.57899
Sugar Babies	0	0	1	0	0	0	0	0	1	0.965	0.767	33.43755
Sugar Daddy	0	0	1	0	0	0	0	0	0	0.418	0.325	32.23100
Super Bubble	0	1	0	0	0	0	0	0	0	0.162	0.116	27.30386
Swedish Fish	0	1	0	0	0	0	0	0	1	0.604	0.755	54.86111
Tootsie Pop	1	1	0	0	0	0	1	0	0	0.604	0.325	48.98265
Tootsie Roll Juniors	1	0	0	0	0	0	0	0	0	0.313	0.511	43.06890
Tootsie Roll Midgies	1	0	0	0	0	0	0	0	1	0.174	0.011	45.73675
Tootsie Roll Snack Bars	1	0	0	0	0	0	0	1	0	0.465	0.325	49.65350
Trolli Sour Bites	0	1	0	0	0	0	0	0	1	0.313	0.255	47.17323
Twix	1	0	1	0	0	1	0	1	0	0.546	0.906	81.64291
Twizzlers	0	1	0	0	0	0	0	0	0	0.220	0.116	45.46628
Warheads	0	1	0	0	0	0	1	0	0	0.093	0.116	39.01190

competitor- name	choco- late	fruity	caramel	peanutyal- mondy	nougat	crispedrice- wafer	hardbar	pluribus	sug- arper- cent	pri- ceper- cent	win- per- cent	
Welch's Fruit Snacks	0	1	0	0	0	0	0	0	1	0.313	0.313	44.37552
Werther's Original Caramel Whoppers	0	0	1	0	0	0	1	0	0	0.186	0.267	41.90431
	1	0	0	0	0	1	0	0	1	0.872	0.848	49.52411

```
candy_file <- "candy-data.csv"
```

```
candy = read.csv(candy_file, row.names = 1)
head(candy)
```

	choco-		peanutyal-		crispedrice-				sug-	pri-	win-	
	late	fruity	caramel	mondy	nougat	wafer	hard	bar	pluribus	arper-	ceper-	per-
										cent	cent	cent
100	1	0	1	0	0	1	0	1	0	0.732	0.860	66.97173
Grand												
3	1	0	0	0	1	0	0	1	0	0.604	0.511	67.60294
Mus-												
ke-												
teers												
One	0	0	0	0	0	0	0	0	0	0.011	0.116	32.26109
dime												
One	0	0	0	0	0	0	0	0	0	0.011	0.511	46.11650
quar-												
ter												
Air	0	1	0	0	0	0	0	0	0	0.906	0.511	52.34146
Heads												
Al-	1	0	0	1	0	0	0	1	0	0.465	0.767	50.34755
mond												
Joy												

```
library(flextable)
flextable::flextable(head(candy))
```

chocolate	fruity	caramel	peanutyal-mondy	nougat	crispedrice-wafer	hard	bar	pluribus
1	0	1	0	0	1	0	1	0
1	0	0	0	1	0	0	1	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
1	0	0	1	0	0	0	1	0

```
library(dplyr)
candy |>nrow()
```

```
[1] 85
```

```
win <- candy$winpercent
win.mean <- mean(win)
round(win.mean)
```

```
[1] 50
```

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)
```

```
[1] 38
```

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

My favorite candy is Snickers.I can find its popularity as follows:

```
candy["Snickers", ]$winpercent
```

```
[1] 76.67378
```

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 Snack Bars”?

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

```
[1] 49.6535
```

```
library("skimr")
skim(candy) |> as.data.frame() |> kable()
```

skim_type	skim_vari- type	in_miss- ing	com- plete_rat- io	nu- meric.mean	nu- meric.sd	nu- meric.p0	nu- meric.p25	nu- meric.p50	nu- meric.p75	nu- meric.p100	nu- meric.hist
nu- meri- c	choco- late	0	1	0.435294	0.498737	0.000000	0.000000	0.000000	1.000	1.00000	
nu- meri- c	fruity	0	1	0.447058	0.500140	0.000000	0.000000	0.000000	1.000	1.00000	
nu- meri- c	caramel	0	1	0.164705	0.373116	0.000000	0.000000	0.000000	0.000	1.00000	
nu- meri- c	peanutyal- mond	0	1	0.164705	0.373116	0.000000	0.000000	0.000000	0.000	1.00000	
nu- meri- c	nougat	0	1	0.082352	0.276538	0.000000	0.000000	0.000000	0.000	1.00000	
nu- meri- c	crispedrice- wafer	0	1	0.082352	0.276538	0.000000	0.000000	0.000000	0.000	1.00000	
nu- meri- c	hard	0	1	0.176470	0.383482	0.000000	0.000000	0.000000	0.000	1.00000	
nu- meri- c	bar	0	1	0.247058	0.433860	0.000000	0.000000	0.000000	0.000	1.00000	
nu- meri- c	pluribus	0	1	0.517647	0.502650	0.000000	0.000000	1.00000	1.000	1.00000	

skim_variable	skim_type	n_missing	complete_rate	numeric.mean	numeric.sd	numeric.p0	numeric.p25	numeric.p50	numeric.p75	numeric.p100	numeric.hist
nu-mer	sugarpercent	0	1	0.478647	0.282777	0.0	0.11000	0.22000	0.46500	0.732	0.98800
nu-mer	pricepercent	0	1	0.468882	0.285730	0.0	0.11000	0.25500	0.46500	0.651	0.97600
nu-mer	winnerpercent	0	1	50.316763	33.714327	44.5349	49.14106	47.82975	59.864	84.18029	

```
skim(candy)
```

Table 5: 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	
crispedrice-wafer	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	

skim_vari- able	n_miss- ing	com- plete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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?

Yes. The variable winpercent is on a different scale than most others, because it ranges from about 22 to 84, whereas most variables are either 0 or 1. The variables sugarpercent and pricepercent are also continuous but remain within 0–1, so winpercent stands out the most.

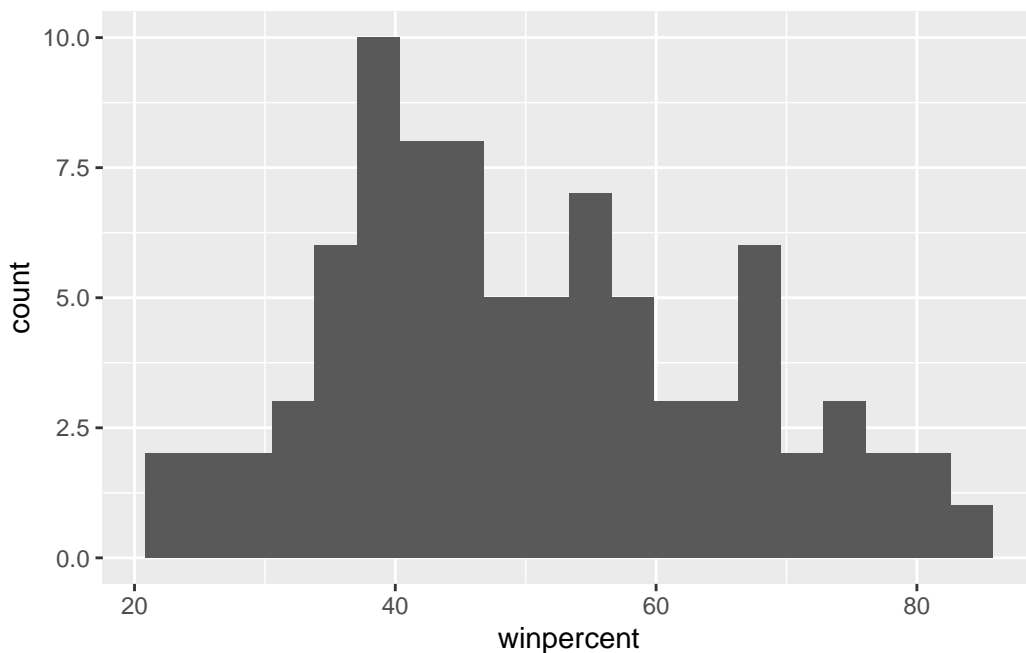
Q7. What do you think a zero and one represent for the candy\$chocolate column?

A zero indicates that the candy does not contain chocolate, and a one indicates that it does contain chocolate.

Q8. Plot a histogram of winpercent values

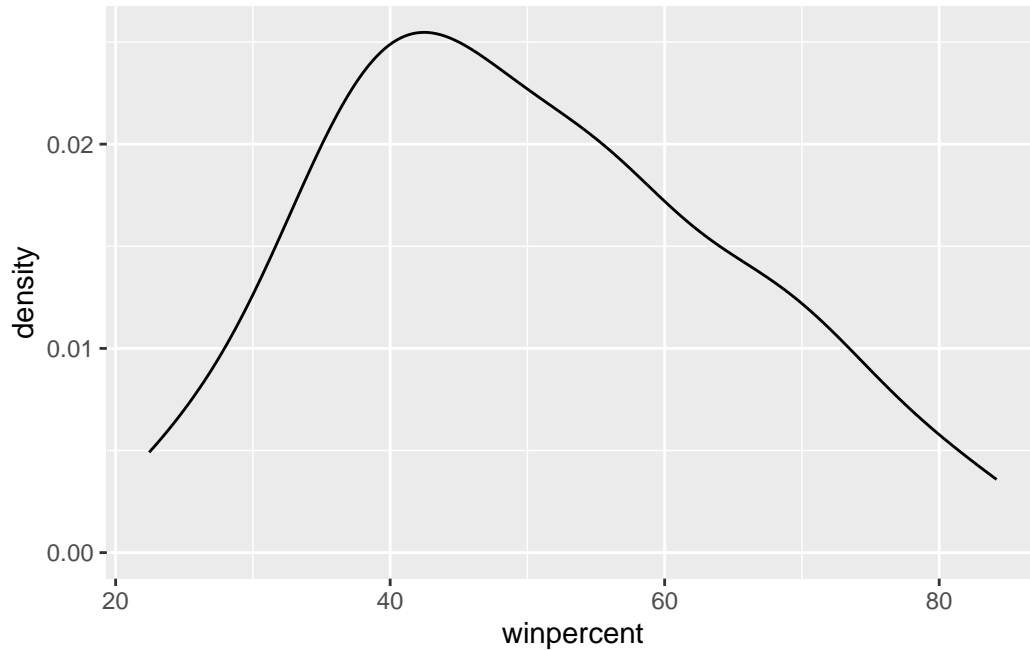
```
library(ggplot2)

ggplot(candy)+ aes(winpercent)+geom_histogram(bins=20)
```



Q9. Is the distribution of winpercent values symmetrical?

```
ggplot(candy)+ aes(winpercent)+geom_density()
```



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

Above

```
mean(candy$winpercent)
```

```
[1] 50.31676
```

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

```
# 1. Find all chocolate candies
choc.inds <- as.logical(candy$chocolate)
choc.candy <- candy[choc.inds, ]

# 2. Extract the winpercent values
choc.win <- choc.candy$winpercent

# 3. Find the mean winpercent for chocolate candies
```

```
choc.mean <- mean(choc.win)

# 4. Do the same for fruity candies
fruit.inds <- as.logical(candy$fruity)
fruit.candy <- candy[fruit.inds, ]

# 5. Extract winpercent values for fruity candies
fruit.win <- fruit.candy$winpercent

# 6. Find the mean winpercent for fruity candies
fruit.mean <- mean(fruit.win)

# 7. Compare the two means
choc.mean
```

```
[1] 60.92153
```

```
fruit.mean
```

```
[1] 44.11974
```

Q12. Is this difference statistically significant?

There is strong statistical evidence ($p < 0.001$) that chocolate candies are, on average, significantly more popular than fruity candies.

```
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
```

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

```
head(candy[order(candy$winpercent), ], n = 5)
```

	choco- late	fruity	caram- chondy	peanutyal- nougat	crispedrice- wafer	hard bar	pluribus	sug- arper- cent	pri- ceper- cent	win- per- cent
Nik L	0	1	0	0	0	0	1	0.197	0.976	22.44534
Nip										
Boston	0	0	0	1	0	0	1	0.313	0.511	23.41782
Baked										
Beans										
Chiclets	0	1	0	0	0	0	1	0.046	0.325	24.52499
Super	0	1	0	0	0	0	0	0.162	0.116	27.30386
Bubble										
Jaw- busters	0	1	0	0	0	1	1	0.093	0.511	28.12744

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

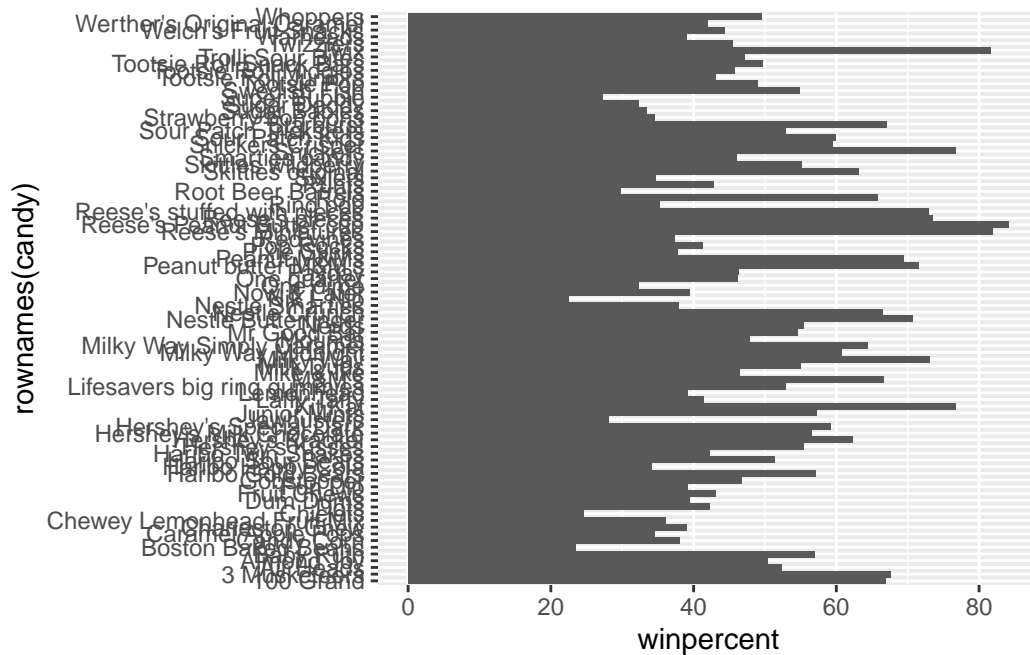
```
head(candy[order(-candy$winpercent), ], n = 5)
```

	choco- late	fruity	caram- chondy	peanutyal- nougat	crispedrice- wafer	hard bar	pluribus	sug- arper- cent	pri- ceper- cent	win- per- cent
Reese's	1	0	0	1	0	0	0	0.720	0.651	84.18029
Peanut										
Butter cup										
Reese's	1	0	0	1	0	0	0	0.034	0.279	81.86626
Miniatures										
Twix	1	0	1	0	0	1	0	0.546	0.906	81.64291
Kit Kat	1	0	0	0	0	1	0	0.313	0.511	76.76860
Snickers	1	0	1	1	1	0	0	0.546	0.651	76.67378

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

```
library(ggplot2)

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

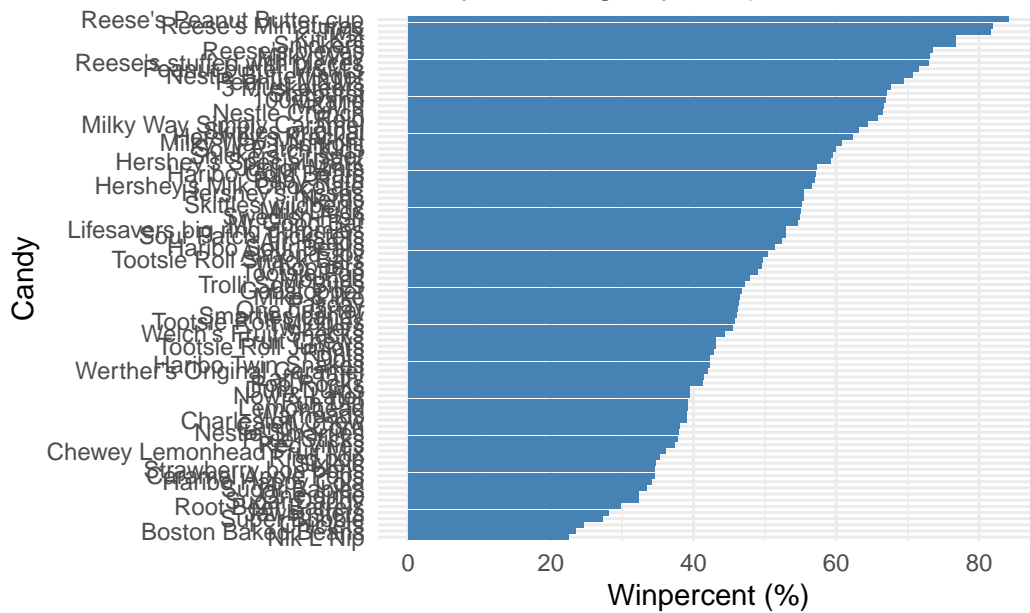


Q16. This is quite ugly, use the `reorder()` function to get the bars sorted by winpercent?

```
library(ggplot2)

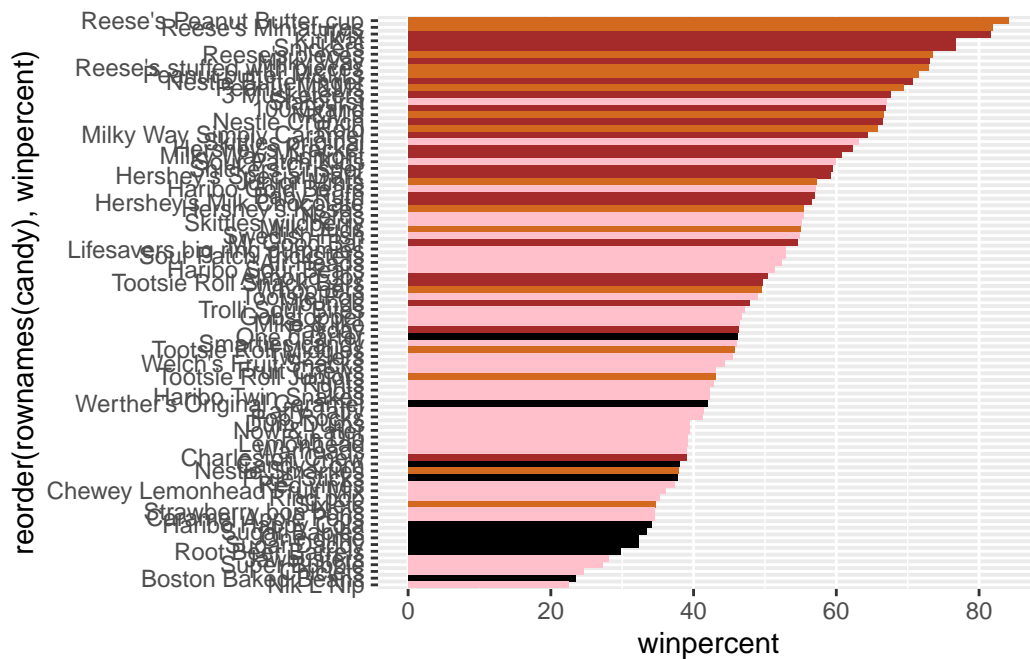
ggplot(candy) +
  aes(x = winpercent, y = reorder(rownames(candy), winpercent)) +
  geom_col(fill = "steelblue") +
  labs(
    title = "Candy Rankings by Winpercent",
    x = "Winpercent (%)",
    y = "Candy"
  ) +
  theme_minimal()
```

Candy Rankings by Winpercent



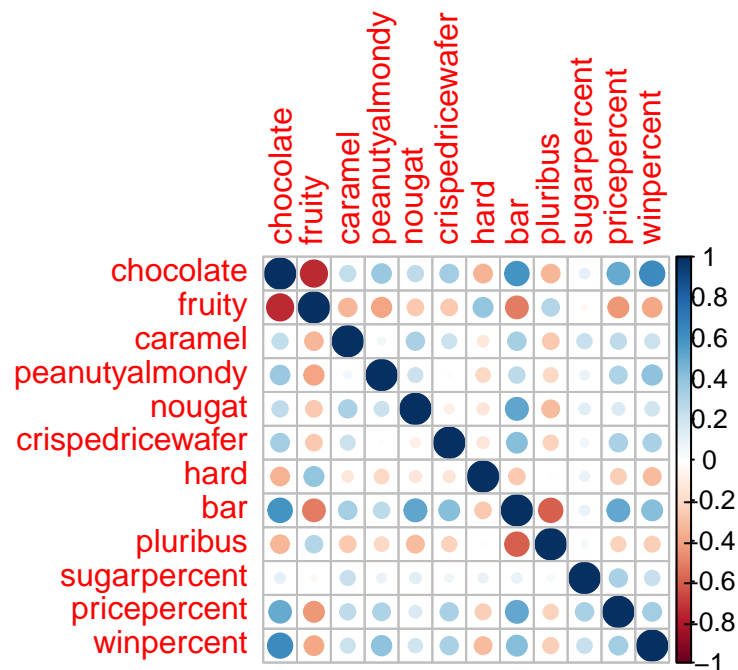
```
my_cols=rep("black", nrow(candy))
my_cols[as.logical(candy$chocolate)] = "chocolate"
my_cols[as.logical(candy$bar)] = "brown"
my_cols[as.logical(candy$fruity)] = "pink"

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



Winpercent vs Pricepercent

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

##PCA

The main function of this on R is `prcomp()` and we want to set `scale=TRUE` here:

```
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

Let's loom at the first main result figure- the "pc-plot" or PC1 vc PC2

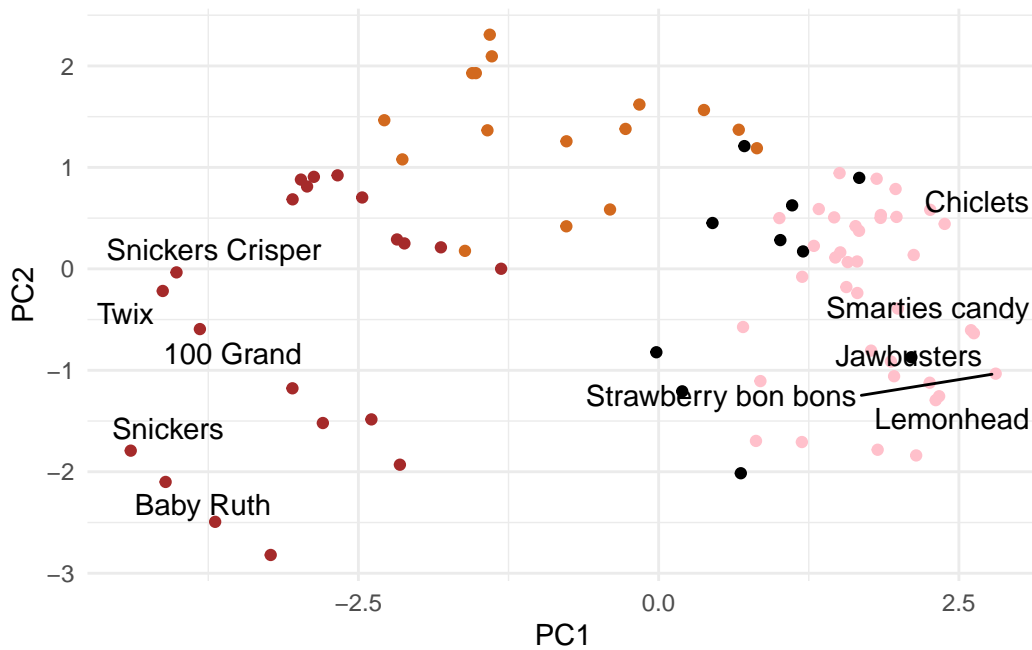
```
library(ggrepel)
# Find points to label
label_points <- as.data.frame(pca$x) %>%
  slice_max(PC1, n = 5) %>%
```

```

bind_rows(slice_min(as.data.frame(pca$x), PC1, n = 5))

ggplot(as.data.frame(pca$x), aes(PC1, PC2)) +
  geom_point(col = my_cols) +
  geom_text_repel(
    data = label_points,
    aes(label = rownames(label_points)),
    col = "black"
  ) +
  theme_minimal()

```



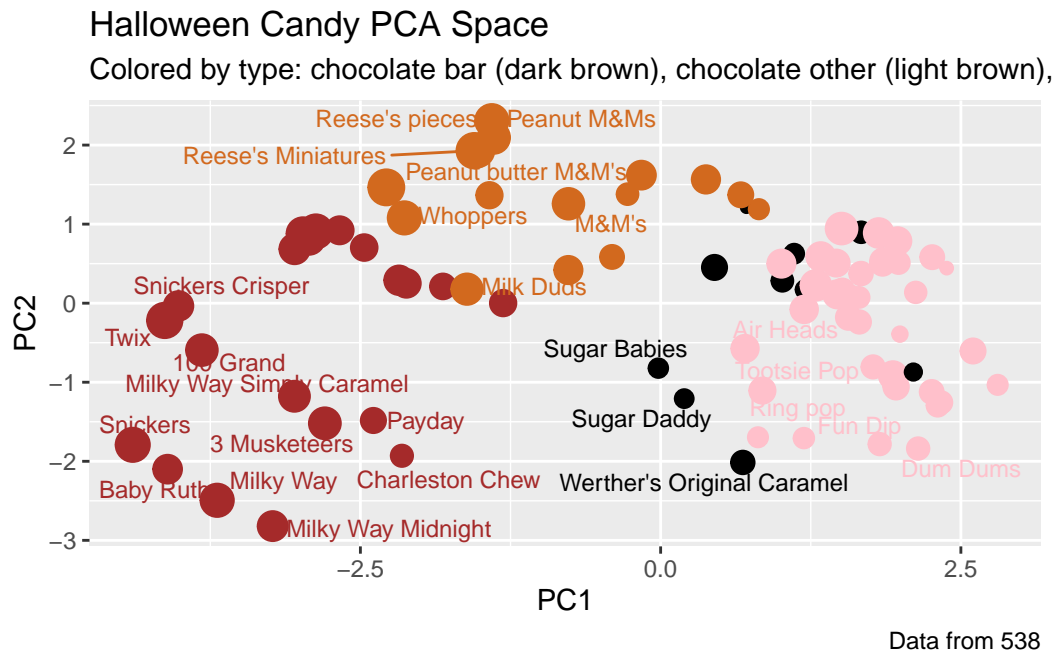
Don't forget about your variable "loadings" - how the original variables contribute to your new PC's

```

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 +
  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"
  )
)
```



```
280
281 {r}
282 #/ label: interactive-plot
283 #/ eval: knitr::is_html_output()
284 library(plotly)
285 plotly::ggplotly(p)
286
```

