

class09_Halloween_mini_Project

Mina Wu (PID:A59013200)

Importing candy data

Today we will analyze some data from 538 about typical Halloween candy.

Read the csv file of the Halloween candy mini project

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

	chocolate	fruity	caramel	peanutyalmondy	nougat
100 Grand	1	0	1	0	0
3 Musketeers	1	0	0	0	1
One dime	0	0	0	0	0
One quarter	0	0	0	0	0
Air Heads	0	1	0	0	0
Almond Joy	1	0	0	1	0
Baby Ruth	1	0	1	1	1
Boston Baked Beans	0	0	0	1	0
Candy Corn	0	0	0	0	0
Caramel Apple Pops	0	1	1	0	0
Charleston Chew	1	0	0	0	1
Chewey Lemonhead Fruit Mix	0	1	0	0	0
Chiclets	0	1	0	0	0
Dots	0	1	0	0	0
Dum Dums	0	1	0	0	0
Fruit Chews	0	1	0	0	0
Fun Dip	0	1	0	0	0
Gobstopper	0	1	0	0	0
Haribo Gold Bears	0	1	0	0	0
Haribo Happy Cola	0	0	0	0	0
Haribo Sour Bears	0	1	0	0	0
Haribo Twin Snakes	0	1	0	0	0

Hershey's Kisses	1	0	0	0	0
Hershey's Krackel	1	0	0	0	0
Hershey's Milk Chocolate	1	0	0	0	0
Hershey's Special Dark	1	0	0	0	0
Jawbusters	0	1	0	0	0
Junior Mints	1	0	0	0	0
Kit Kat	1	0	0	0	0
Laffy Taffy	0	1	0	0	0
Lemonhead	0	1	0	0	0
Lifesavers big ring gummies	0	1	0	0	0
Peanut butter M&M's	1	0	0	1	0
M&M's	1	0	0	0	0
Mike & Ike	0	1	0	0	0
Milk Duds	1	0	1	0	0
Milky Way	1	0	1	0	1
Milky Way Midnight	1	0	1	0	1
Milky Way Simply Caramel	1	0	1	0	0
Mounds	1	0	0	0	0
Mr Good Bar	1	0	0	1	0
Nerds	0	1	0	0	0
Nestle Butterfinger	1	0	0	1	0
Nestle Crunch	1	0	0	0	0
Nik L Nip	0	1	0	0	0
Now & Later	0	1	0	0	0
Payday	0	0	0	1	1
Peanut M&Ms	1	0	0	1	0
Pixie Sticks	0	0	0	0	0
Pop Rocks	0	1	0	0	0
Red vines	0	1	0	0	0
Reese's Miniatures	1	0	0	1	0
Reese's Peanut Butter cup	1	0	0	1	0
Reese's pieces	1	0	0	1	0
Reese's stuffed with pieces	1	0	0	1	0
Ring pop	0	1	0	0	0
Rolo	1	0	1	0	0
Root Beer Barrels	0	0	0	0	0
Runts	0	1	0	0	0
Sixlets	1	0	0	0	0
Skittles original	0	1	0	0	0
Skittles wildberry	0	1	0	0	0
Nestle Smarties	1	0	0	0	0
Smarties candy	0	1	0	0	0
Snickers	1	0	1	1	1

Snickers Crisper	1	0	1	1	0	
Sour Patch Kids	0	1	0	0	0	
Sour Patch Tricksters	0	1	0	0	0	
Starburst	0	1	0	0	0	
Strawberry bon bons	0	1	0	0	0	
Sugar Babies	0	0	1	0	0	
Sugar Daddy	0	0	1	0	0	
Super Bubble	0	1	0	0	0	
Swedish Fish	0	1	0	0	0	
Tootsie Pop	1	1	0	0	0	
Tootsie Roll Juniors	1	0	0	0	0	
Tootsie Roll Midgies	1	0	0	0	0	
Tootsie Roll Snack Bars	1	0	0	0	0	
Trolli Sour Bites	0	1	0	0	0	
Twix	1	0	1	0	0	
Twizzlers	0	1	0	0	0	
Warheads	0	1	0	0	0	
Welch's Fruit Snacks	0	1	0	0	0	
Werther's Original Caramel	0	0	1	0	0	
Whoppers	1	0	0	0	0	
	crisped	rice	wafer	hard bar	pluribus sugarpercent	
100 Grand		1	0	1	0	0.732
3 Musketeers		0	0	1	0	0.604
One dime		0	0	0	0	0.011
One quarter		0	0	0	0	0.011
Air Heads		0	0	0	0	0.906
Almond Joy		0	0	1	0	0.465
Baby Ruth		0	0	1	0	0.604
Boston Baked Beans		0	0	0	1	0.313
Candy Corn		0	0	0	1	0.906
Caramel Apple Pops		0	0	0	0	0.604
Charleston Chew		0	0	1	0	0.604
Chewey Lemonhead Fruit Mix		0	0	0	1	0.732
Chiclets		0	0	0	1	0.046
Dots		0	0	0	1	0.732
Dum Dums		0	1	0	0	0.732
Fruit Chews		0	0	0	1	0.127
Fun Dip		0	1	0	0	0.732
Gobstopper		0	1	0	1	0.906
Haribo Gold Bears		0	0	0	1	0.465
Haribo Happy Cola		0	0	0	1	0.465
Haribo Sour Bears		0	0	0	1	0.465
Haribo Twin Snakes		0	0	0	1	0.465

Hershey's Kisses	0	0	0	1	0.127
Hershey's Krackel	1	0	1	0	0.430
Hershey's Milk Chocolate	0	0	1	0	0.430
Hershey's Special Dark	0	0	1	0	0.430
Jawbusters	0	1	0	1	0.093
Junior Mints	0	0	0	1	0.197
Kit Kat	1	0	1	0	0.313
Laffy Taffy	0	0	0	0	0.220
Lemonhead	0	1	0	0	0.046
Lifesavers big ring gummies	0	0	0	0	0.267
Peanut butter M&M's	0	0	0	1	0.825
M&M's	0	0	0	1	0.825
Mike & Ike	0	0	0	1	0.872
Milk Duds	0	0	0	1	0.302
Milky Way	0	0	1	0	0.604
Milky Way Midnight	0	0	1	0	0.732
Milky Way Simply Caramel	0	0	1	0	0.965
Mounds	0	0	1	0	0.313
Mr Good Bar	0	0	1	0	0.313
Nerds	0	1	0	1	0.848
Nestle Butterfinger	0	0	1	0	0.604
Nestle Crunch	1	0	1	0	0.313
Nik L Nip	0	0	0	1	0.197
Now & Later	0	0	0	1	0.220
Payday	0	0	1	0	0.465
Peanut M&Ms	0	0	0	1	0.593
Pixie Sticks	0	0	0	1	0.093
Pop Rocks	0	1	0	1	0.604
Red vines	0	0	0	1	0.581
Reese's Miniatures	0	0	0	0	0.034
Reese's Peanut Butter cup	0	0	0	0	0.720
Reese's pieces	0	0	0	1	0.406
Reese's stuffed with pieces	0	0	0	0	0.988
Ring pop	0	1	0	0	0.732
Rolo	0	0	0	1	0.860
Root Beer Barrels	0	1	0	1	0.732
Runts	0	1	0	1	0.872
Sixlets	0	0	0	1	0.220
Skittles original	0	0	0	1	0.941
Skittles wildberry	0	0	0	1	0.941
Nestle Smarties	0	0	0	1	0.267
Smarties candy	0	1	0	1	0.267
Snickers	0	0	1	0	0.546

Snickers Crisper	1	0	1	0	0.604
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
Strawberry bon bons	0	1	0	1	0.569
Sugar Babies	0	0	0	1	0.965
Sugar Daddy	0	0	0	0	0.418
Super Bubble	0	0	0	0	0.162
Swedish Fish	0	0	0	1	0.604
Tootsie Pop	0	1	0	0	0.604
Tootsie Roll Juniors	0	0	0	0	0.313
Tootsie Roll Midgies	0	0	0	1	0.174
Tootsie Roll Snack Bars	0	0	1	0	0.465
Trolli Sour Bites	0	0	0	1	0.313
Twix	1	0	1	0	0.546
Twizzlers	0	0	0	0	0.220
Warheads	0	1	0	0	0.093
Welch's Fruit Snacks	0	0	0	1	0.313
Werther's Original Caramel	0	1	0	0	0.186
Whoppers	1	0	0	1	0.872

	pricepercent	winpercent
100 Grand	0.860	66.97173
3 Musketeers	0.511	67.60294
One dime	0.116	32.26109
One quarter	0.511	46.11650
Air Heads	0.511	52.34146
Almond Joy	0.767	50.34755
Baby Ruth	0.767	56.91455
Boston Baked Beans	0.511	23.41782
Candy Corn	0.325	38.01096
Caramel Apple Pops	0.325	34.51768
Charleston Chew	0.511	38.97504
Chewey Lemonhead Fruit Mix	0.511	36.01763
Chiclets	0.325	24.52499
Dots	0.511	42.27208
Dum Dums	0.034	39.46056
Fruit Chews	0.034	43.08892
Fun Dip	0.325	39.18550
Gobstopper	0.453	46.78335
Haribo Gold Bears	0.465	57.11974
Haribo Happy Cola	0.465	34.15896
Haribo Sour Bears	0.465	51.41243
Haribo Twin Snakes	0.465	42.17877

Hershey's Kisses	0.093	55.37545
Hershey's Krackel	0.918	62.28448
Hershey's Milk Chocolate	0.918	56.49050
Hershey's Special Dark	0.918	59.23612
Jawbusters	0.511	28.12744
Junior Mints	0.511	57.21925
Kit Kat	0.511	76.76860
Laffy Taffy	0.116	41.38956
Lemonhead	0.104	39.14106
Lifesavers big ring gummies	0.279	52.91139
Peanut butter M&M's	0.651	71.46505
M&M's	0.651	66.57458
Mike & Ike	0.325	46.41172
Milk Duds	0.511	55.06407
Milky Way	0.651	73.09956
Milky Way Midnight	0.441	60.80070
Milky Way Simply Caramel	0.860	64.35334
Mounds	0.860	47.82975
Mr Good Bar	0.918	54.52645
Nerds	0.325	55.35405
Nestle Butterfinger	0.767	70.73564
Nestle Crunch	0.767	66.47068
Nik L Nip	0.976	22.44534
Now & Later	0.325	39.44680
Payday	0.767	46.29660
Peanut M&Ms	0.651	69.48379
Pixie Sticks	0.023	37.72234
Pop Rocks	0.837	41.26551
Red vines	0.116	37.34852
Reese's Miniatures	0.279	81.86626
Reese's Peanut Butter cup	0.651	84.18029
Reese's pieces	0.651	73.43499
Reese's stuffed with pieces	0.651	72.88790
Ring pop	0.965	35.29076
Rolo	0.860	65.71629
Root Beer Barrels	0.069	29.70369
Runts	0.279	42.84914
Sixlets	0.081	34.72200
Skittles original	0.220	63.08514
Skittles wildberry	0.220	55.10370
Nestle Smarties	0.976	37.88719
Smarties candy	0.116	45.99583
Snickers	0.651	76.67378

Snickers Crisper	0.651	59.52925
Sour Patch Kids	0.116	59.86400
Sour Patch Tricksters	0.116	52.82595
Starburst	0.220	67.03763
Strawberry bon bons	0.058	34.57899
Sugar Babies	0.767	33.43755
Sugar Daddy	0.325	32.23100
Super Bubble	0.116	27.30386
Swedish Fish	0.755	54.86111
Tootsie Pop	0.325	48.98265
Tootsie Roll Juniors	0.511	43.06890
Tootsie Roll Midgies	0.011	45.73675
Tootsie Roll Snack Bars	0.325	49.65350
Trolli Sour Bites	0.255	47.17323
Twix	0.906	81.64291
Twizzlers	0.116	45.46628
Warheads	0.116	39.01190
Welch's Fruit Snacks	0.313	44.37552
Werther's Original Caramel	0.267	41.90431
Whoppers	0.848	49.52411

Q1. How many different candy types are in this dataset? 85 Q2. How many fruity candy types are in the dataset? 38

```
type_candy <- nrow(candy)
type_candy
```

```
[1] 85
```

```
fruity<- sum(candy$fruity)
fruity
```

```
[1] 38
```

Q. How many variables/dimensions are there?

```
ncol(candy)
```

```
[1] 12
```

What is your favorite candy?

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

```
[1] 81.64291
```

My favorite candy is Peanut M&Ms.

```
favorite_candy <- candy["Peanut M&Ms",]$winpercent  
favorite_candy
```

```
[1] 69.48379
```

```
kitkat <- candy["Kit Kat",]$winpercent  
kitkat
```

```
[1] 76.7686
```

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

```
[1] 49.6535
```

Q3. What is your favorite candy in the dataset and what is its winpercent value? My favorite candy is peanut M&Ms and its winpercent value is 69.48379%.

Q4. What is the winpercent value for “Kit Kat”? 76.7686%.

Q5. What is the winpercent value for “Tootsie Roll Snack Bars”? 49.6535%

Folks like the skimr package for a first peek into a new dataset. Let's see what it does for this candy dataset.

```
library("skimr")  
#allow you to call skimr without library using the ::  
# skimr::(candy)  
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? Yes, the variable winpercent.

Q7. What do you think a zero and one represent for the candy\$chocolate column? 1 means the candy is chocolate (true). 0 means that the candy is not chocolate (false).

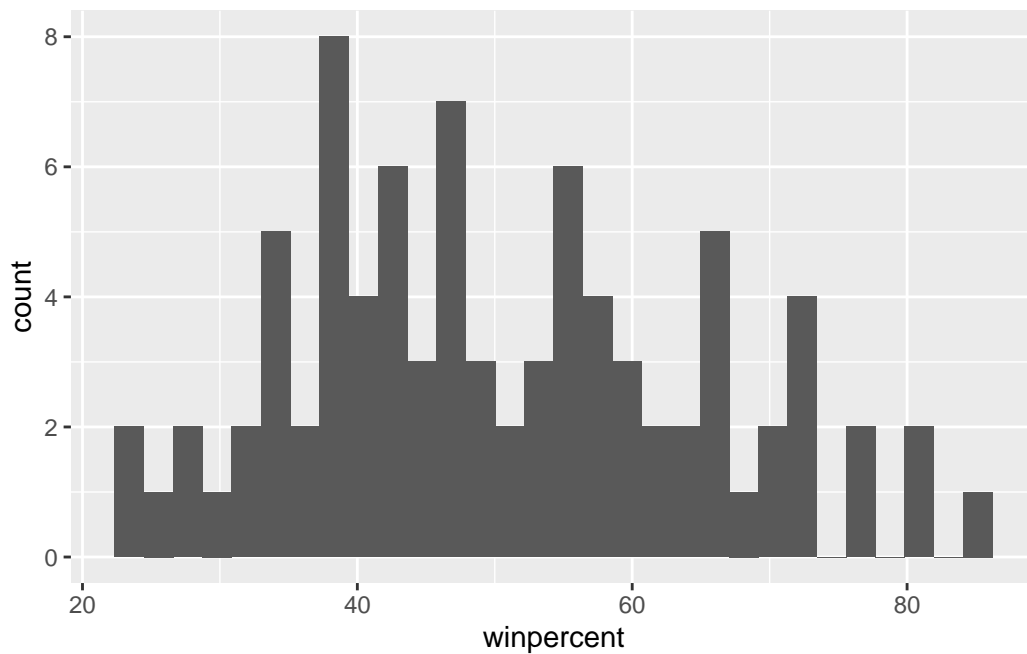
```
chocolate <- candy$chocolate
chocolate
```

```
[1] 1 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 0 0 0 1 1 0 1 1 1
[39] 1 1 1 0 1 1 0 0 0 1 0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 1 1
[77] 1 1 0 1 0 0 0 0 1
```

Q8. Plot a histogram of winpercent values.

```
library(ggplot2)

ggplot(candy, aes(winpercent))+
  geom_histogram(bins =30)
```



Q9. Is the distribution of winpercent values symmetrical? No, the distribution of winpercent values is not symmetrical. They are left-skewed.

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

-First find all chocolate candy(subset) -Get their winpercent values -Summarize these values into one metric

```
#Find mean for the chocolate candy
chocolate_ranked <- candy$winpercent[as.logical(candy$chocolate)]
mean_chocolate <- mean(chocolate_ranked)
mean_chocolate
```

```
[1] 60.92153
```

```
#Find mean for the fruity candy
fruity_ranked <- candy$winpercent[as.logical(candy$fruity)]
mean_fruity<- mean(fruity_ranked)
mean_fruity
```

```
[1] 44.11974
```

```
#t-test for chocolate and fruity
t.test(chocolate_ranked, fruity_ranked)
```

Welch Two Sample t-test

```
data: chocolate_ranked and fruity_ranked
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. On average is chocolate candy higher or lower ranked than fruit candy? Yes, chocolate is preferred with an mean of 60.92 compared to fruity with a mean of 44.12.

Q12. Is this difference statistically significant? Yes, it is significantly different, because the p-value is 2.87810^{-08} , which is less than <0.05 .

Overall Candy Rankings

load the dplyr package

```
library(dplyr)
```

Warning: package 'dplyr' was built under R version 4.0.5

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	peanut	almond	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

	crisped	rice	wafer	hard	bar	pluribus	sugar	percent	price	percent
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. What are the five least liked candy types in this set? Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters.

Use arrange with desc to sort the candy in descending order.

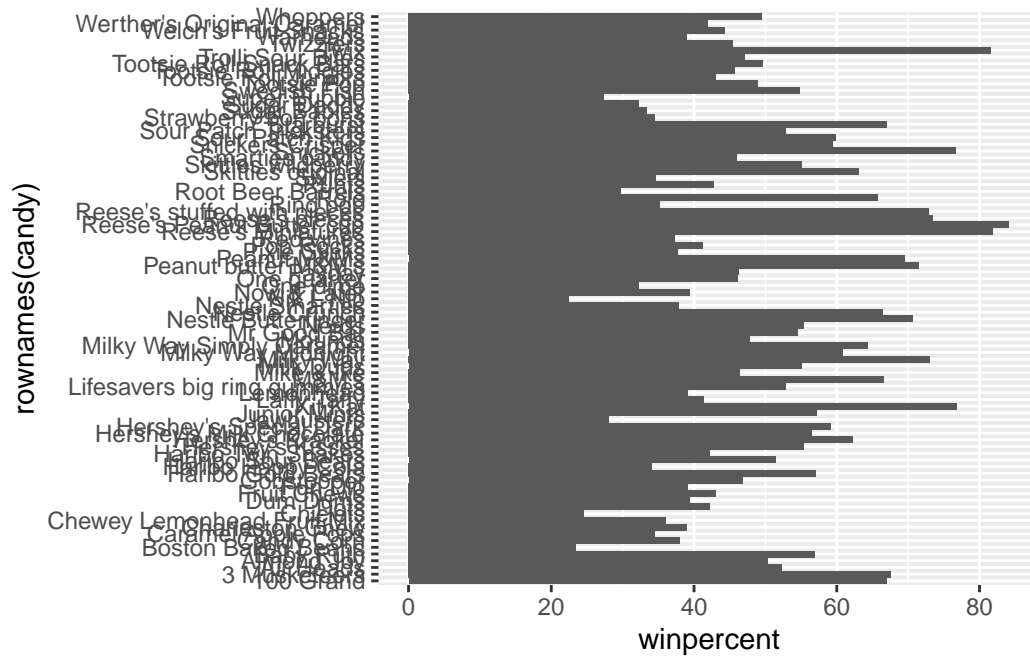
```
candy %>%  
  arrange(desc(winpercent)) %>%  
  head(5)
```

	chocolate	fruity	caramel	peanut	almond	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
	crisp	edricewafer	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. What are the top 5 all time favorite candy types out of this set? Reese's Peanut Butter cup, Reese's miniatures, Twix, Kit Kat, and Snickers

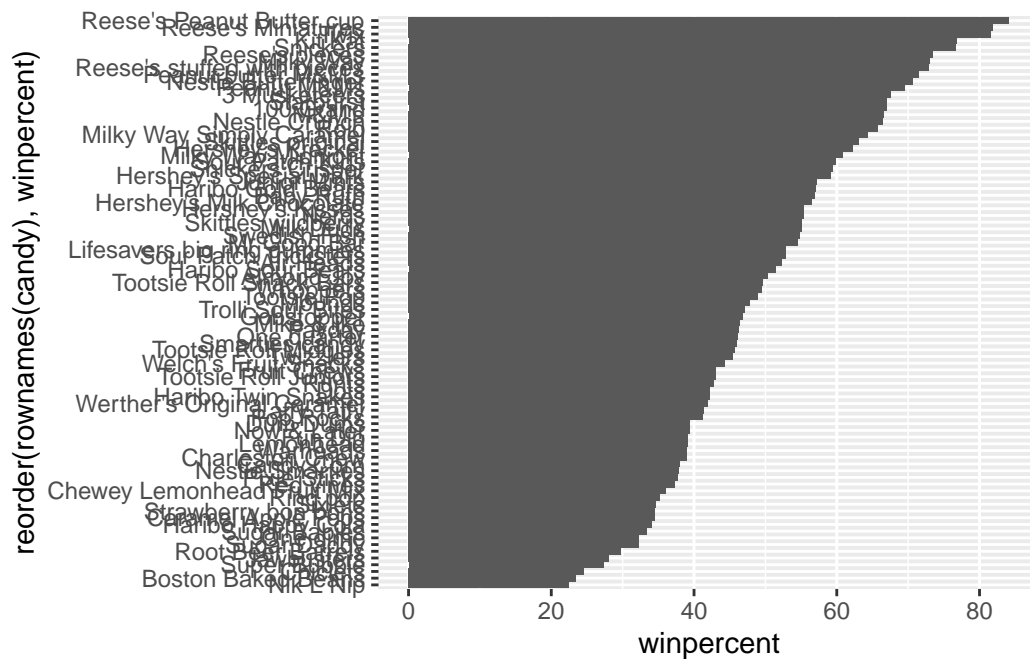
Q15. Make a first barplot of candy ranking based on winpercent values.

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



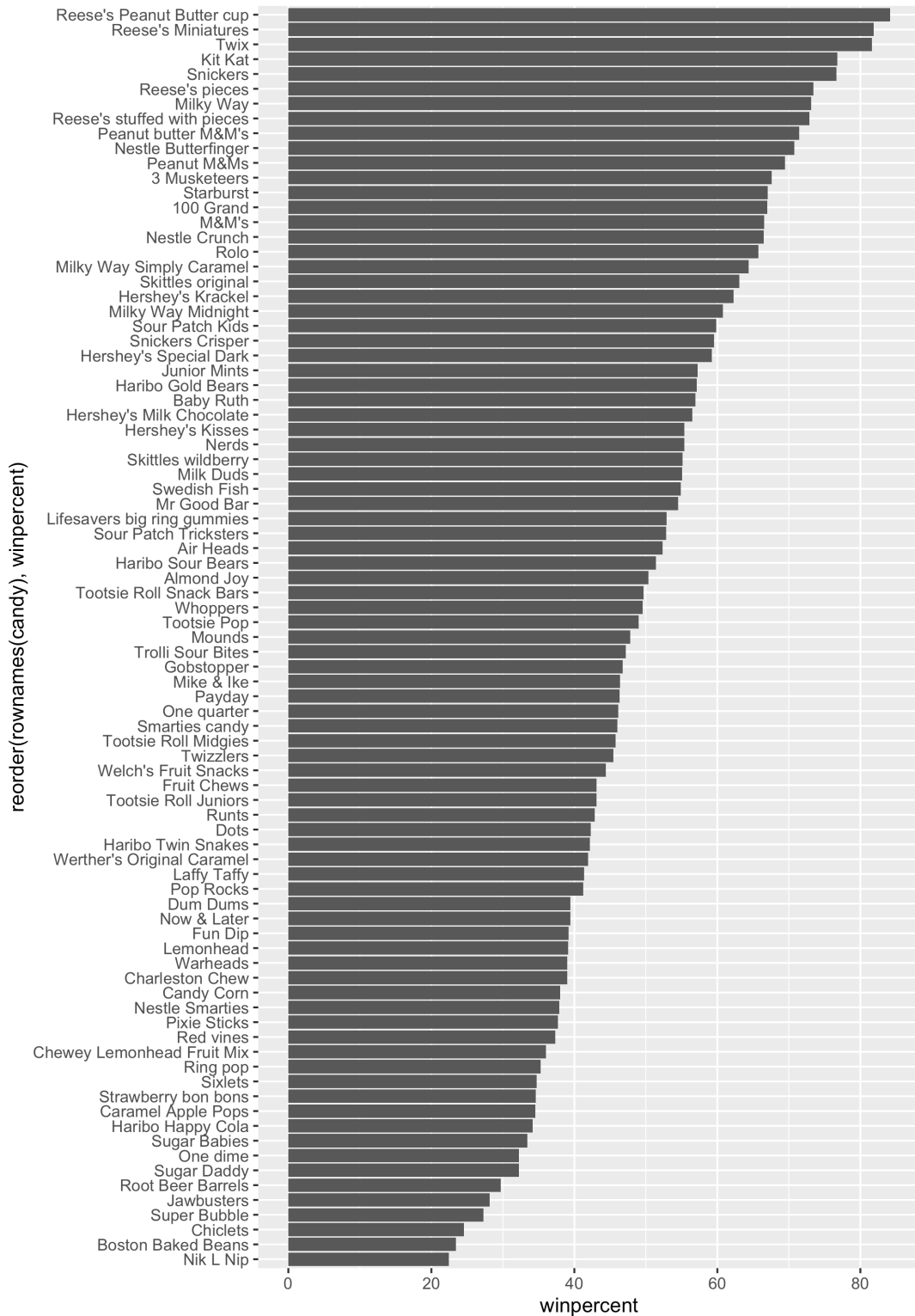
Q16. This is quite ugly, use the `reorder()` function to get the bars sorted by winpercent?

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



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

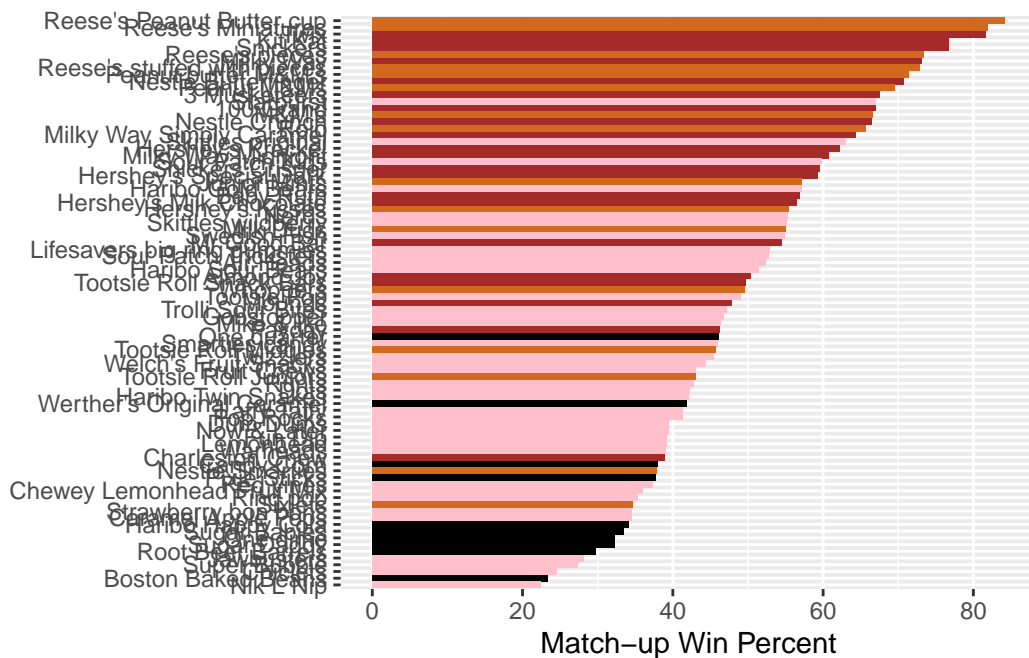
We can now insert any image using markdown syntax. This is ! followed by square brackets and then normal brackets.



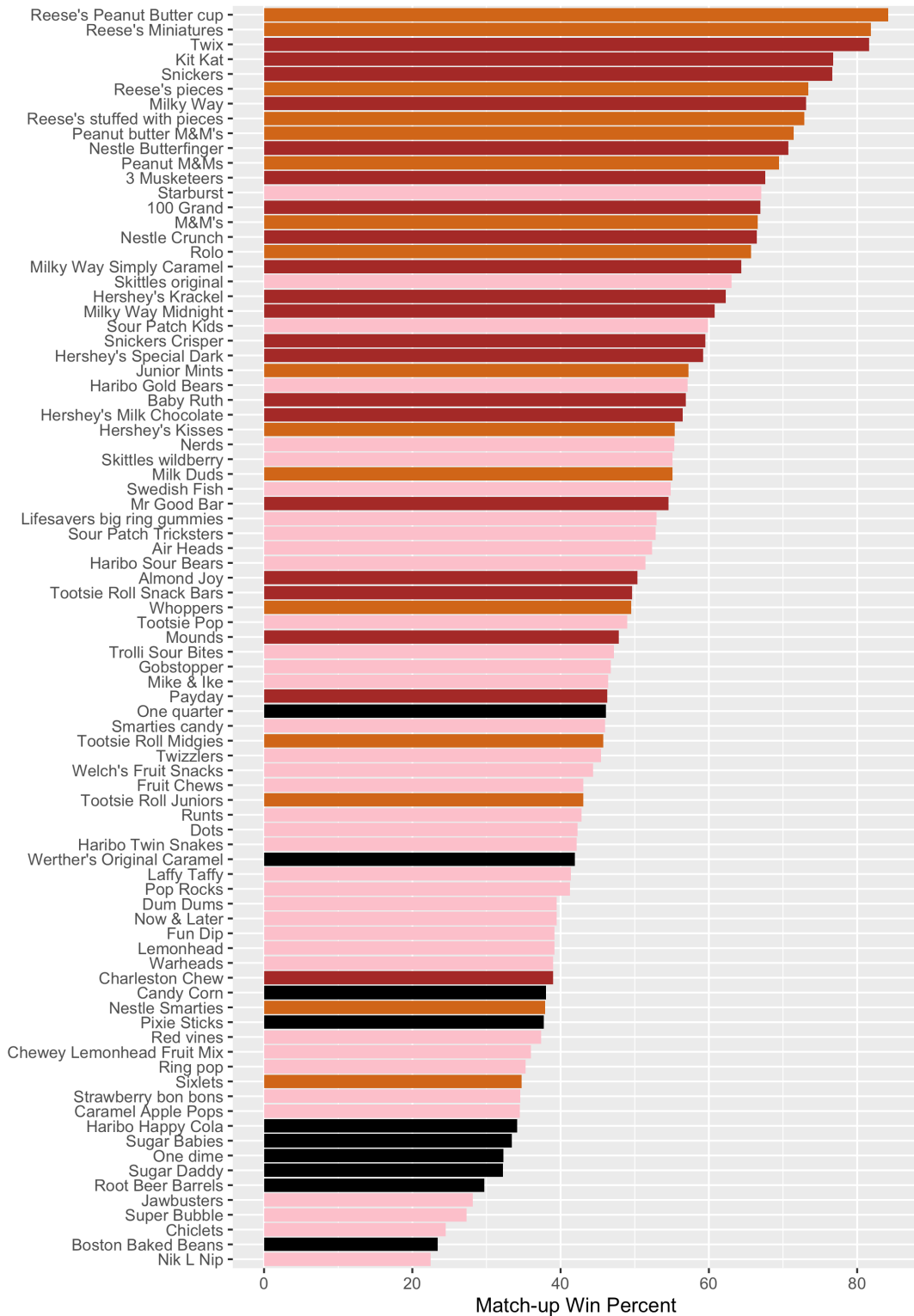
Making a vector of all black values (one for each candy). Then we overwrite chocolate (for chocolate candy), brown (for candy bars) and red (for fruity candy) values.

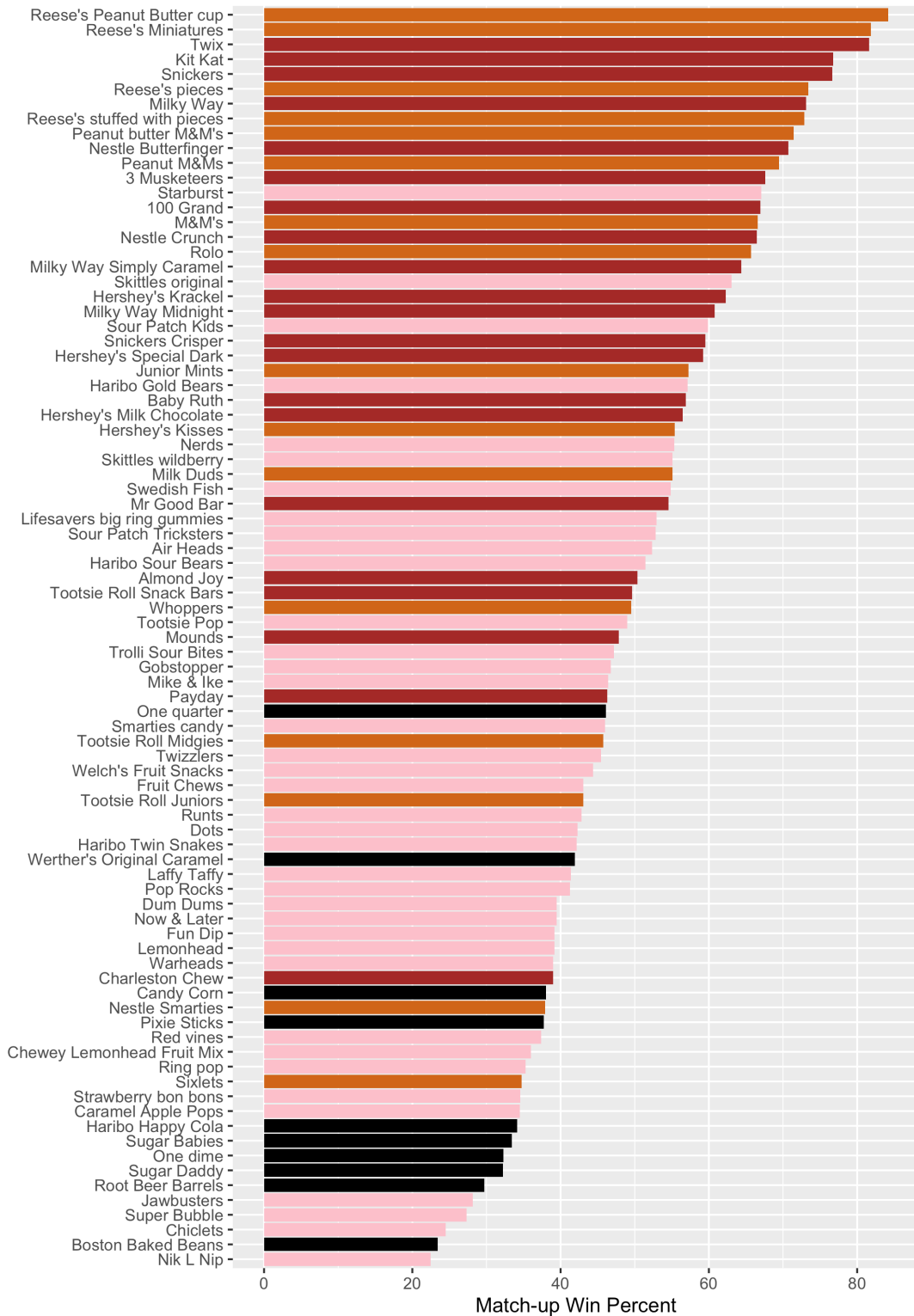
```
my_cols=rep("black", nrow(candy))
my_cols[as.logical(candy$chocolate)] = "chocolate"
my_cols[as.logical(candy$bar)] = "brown"
my_cols[as.logical(candy$fruity)] = "pink"

ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols) +
  labs(x="Match-up Win Percent", y= NULL)
```

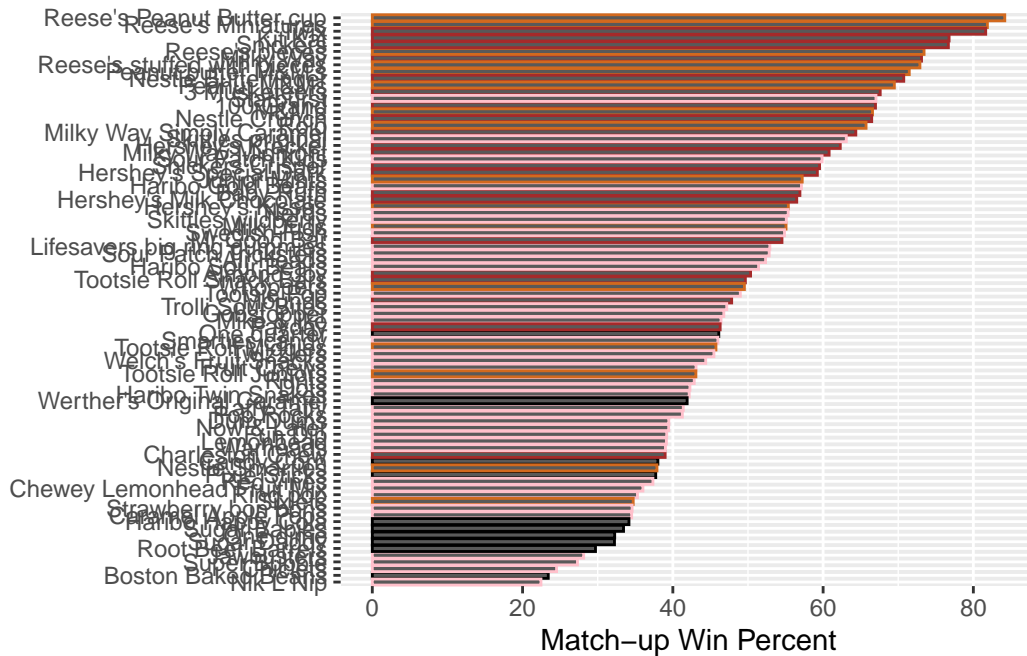


```
ggsave("barplot2.png", height =10, width=7)
```





```
ggplot(candy) +  
  aes(winpercent, reorder(rownames(candy), winpercent)) +  
  geom_col(col=my_cols) +  
  labs(x="Match-up Win Percent", y=NULL)
```



Q17. What is the worst ranked chocolate candy? Nik L nip

Q18. What is the best ranked fruity candy? Reeses Peanut Butte cup.

Taking a look at pricepercent

Install and load `ggrepel`.

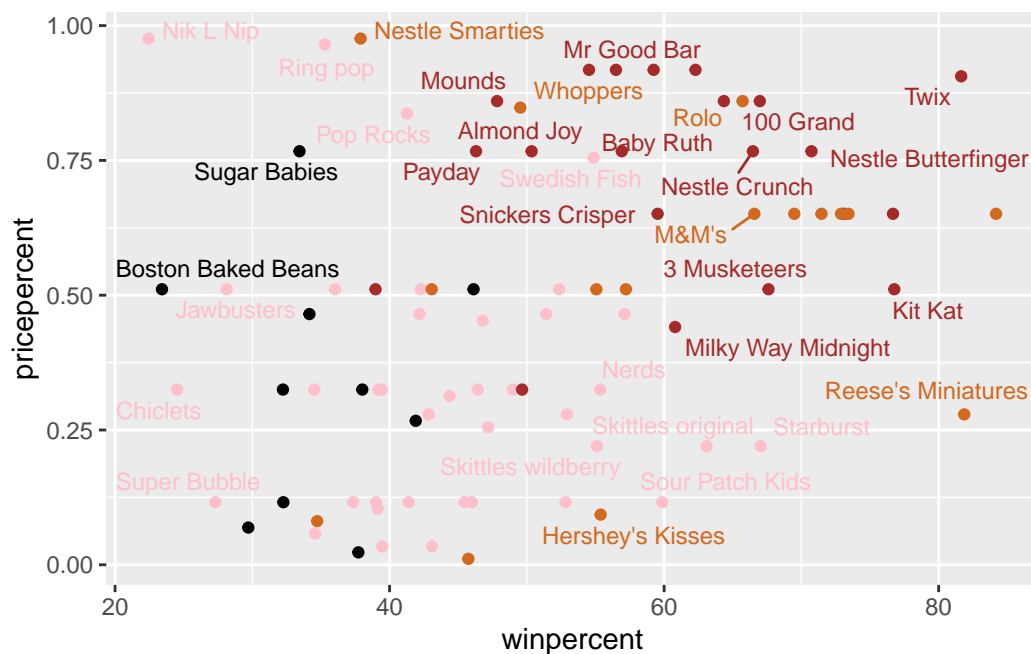
```
library (ggrepel)
```

Warning: package 'ggrepel' was built under R version 4.0.2

```
# How about a plot of price vs win
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
```

```
geom_point(col=my_cols) +  
geom_text_repel(col=my_cols, size=3.3, max.overlaps = 8)
```

Warning: ggrepel: 52 unlabeled data points (too many overlaps). Consider increasing max.overlaps



```
winpercent_to_pricepercent <- candy$winpercent/candy$pricepercent  
best_value <- candy[which.max(winpercent_to_pricepercent),]  
best_value
```

	chocolate	fruity	caramel	peanutyalmondy	nougat
Tootsie Roll Midgies	1	0	0	0	0
	crispedrice	wafer	hard bar	pluribus	sugarpercent
Tootsie Roll Midgies	0	0	0	1	0.174
	pricepercent	winpercent			
Tootsie Roll Midgies	0.011	45.73675			

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

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

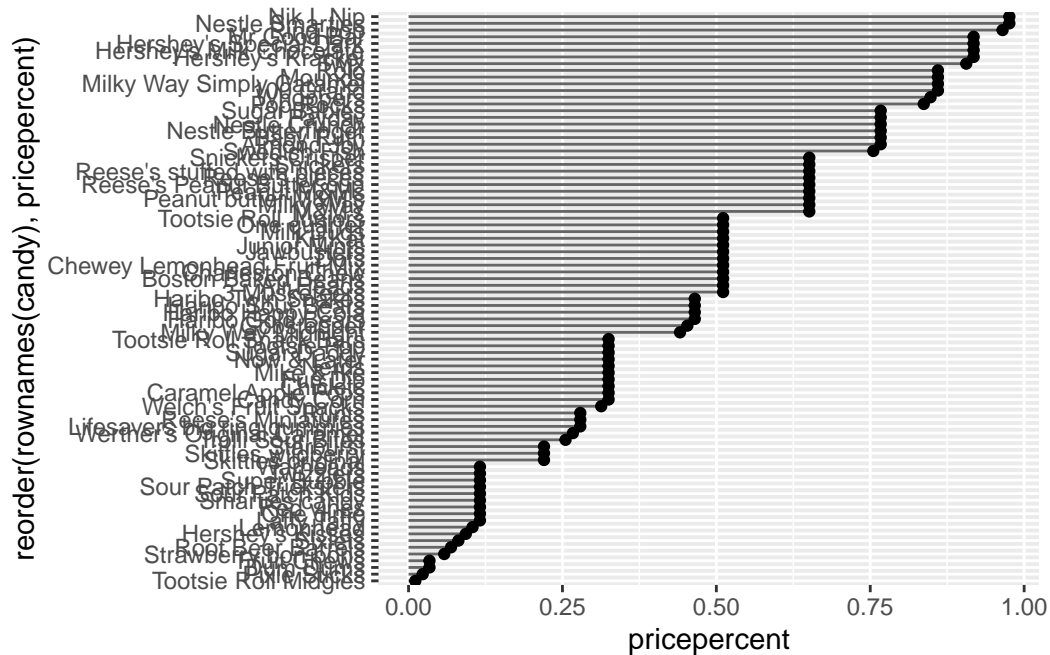
	pricepercent	winpercent
Nik L Nip	0.976	22.44534
Nestle Smarties	0.976	37.88719
Ring pop	0.965	35.29076
Hershey's Krackel	0.918	62.28448
Hershey's Milk Chocolate	0.918	56.49050

Q20. 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, Hershey's Krackel, and Hershey's Milk Chocolate. The least popular out of these 5 is the Nik L Nip.

Q21. Make a barplot again with `geom_col()` this time using `pricepercent` and then improve this step by step, first ordering the x-axis by value and finally making a so called "dot chat" or "lollipop" chart by swapping `geom_col()` for `geom_point()` + `geom_segment()`.

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



5 Exploring the correlation structure

We will calculate all Pearson correlation vlaues. load the corrrplot

```
library(corrplot)
```

corrplot 0.92 loaded

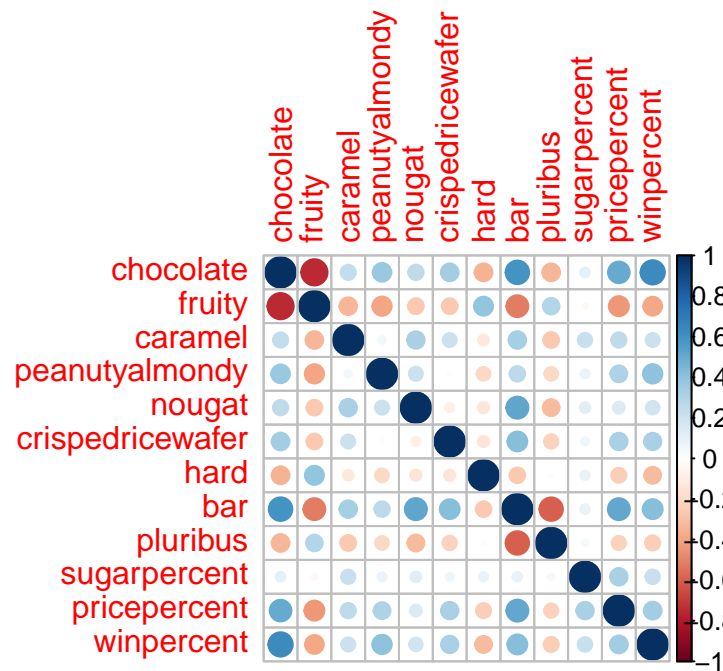
```
cij <- cor(candy)
head(cij)
```

	chocolate	fruity	caramel	peanutyalmondy	nougat
chocolate	1.0000000	-0.7417211	0.24987535	0.37782357	0.25489183
fruity	-0.7417211	1.0000000	-0.33548538	-0.39928014	-0.26936712
caramel	0.2498753	-0.3354854	1.00000000	0.05935614	0.32849280
peanutyalmondy	0.3778236	-0.3992801	0.05935614	1.00000000	0.21311310
nougat	0.2548918	-0.2693671	0.32849280	0.21311310	1.00000000
crispedricewafer	0.3412098	-0.2693671	0.21311310	-0.01764631	-0.08974359
	crispedricewafer	hard	bar	pluribus	sugarpercent
chocolate	0.34120978	-0.3441769	0.5974211	-0.3396752	0.10416906

fruity	-0.26936712	0.3906775	-0.5150656	0.2997252	-0.03439296
caramel	0.21311310	-0.1223551	0.3339600	-0.2695850	0.22193335
peanutyalmondy	-0.01764631	-0.2055566	0.2604196	-0.2061093	0.08788927
nougat	-0.08974359	-0.1386750	0.5229764	-0.3103388	0.12308135
crispedricewafer	1.00000000	-0.1386750	0.4237509	-0.2246934	0.06994969

	pricepercent	winpercent
chocolate	0.5046754	0.6365167
fruity	-0.4309685	-0.3809381
caramel	0.2543271	0.2134163
peanutyalmondy	0.3091532	0.4061922
nougat	0.1531964	0.1993753
crispedricewafer	0.3282654	0.3246797

```
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? winpercent and chocolate

Principal Component Analysis

The main function is called `prcomp()` we need to scale.

Let's apply PCA using the `prcomp()` function to our candy dataset

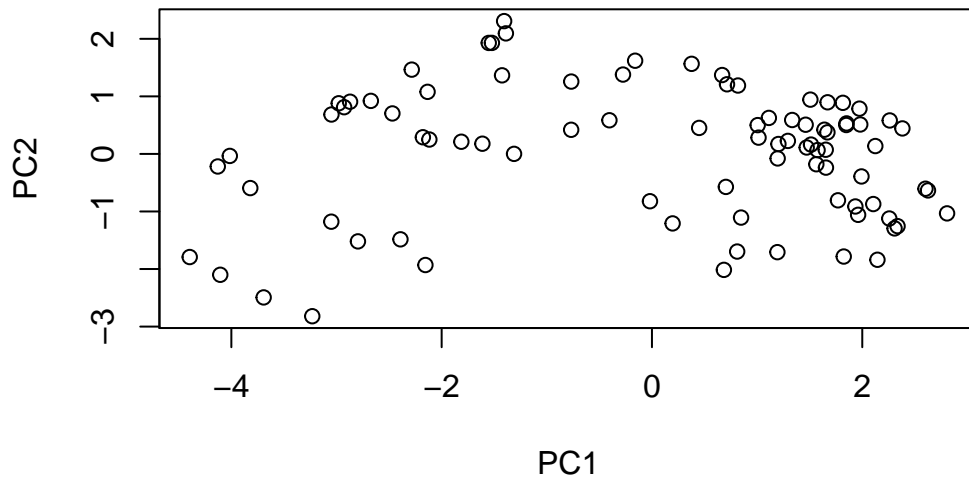
```
pca <- prcomp(candy, scale= TRUE)
summary(pca)
```

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

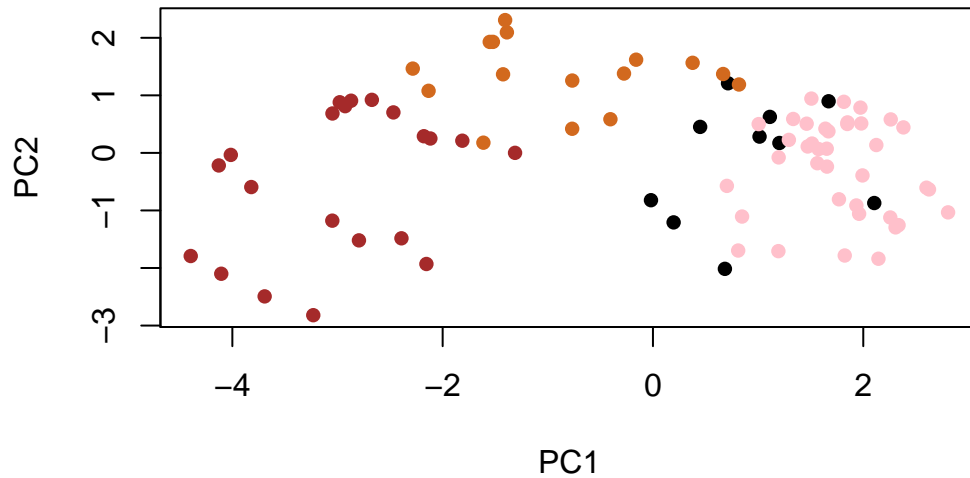
	PC8	PC9	PC10	PC11	PC12
Standard deviation	0.74530	0.67824	0.62349	0.43974	0.39760
Proportion of Variance	0.04629	0.03833	0.03239	0.01611	0.01317
Cumulative Proportion	0.89998	0.93832	0.97071	0.98683	1.00000

```
plot(pca$x[,1:2])
```



Change the plotting character and add some color

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

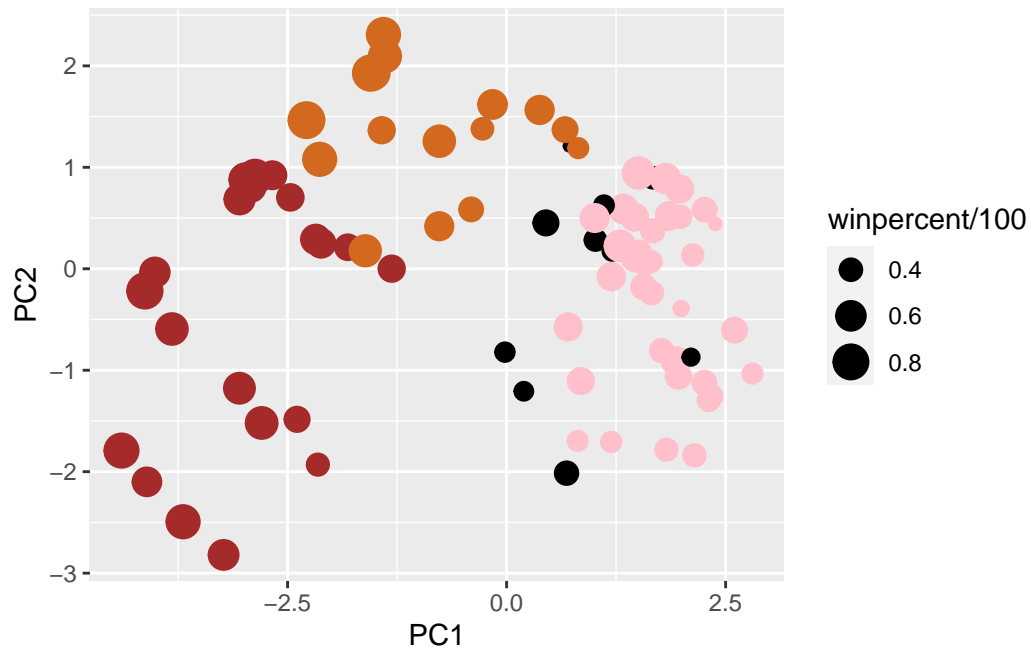


Make the plot with ggplot

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

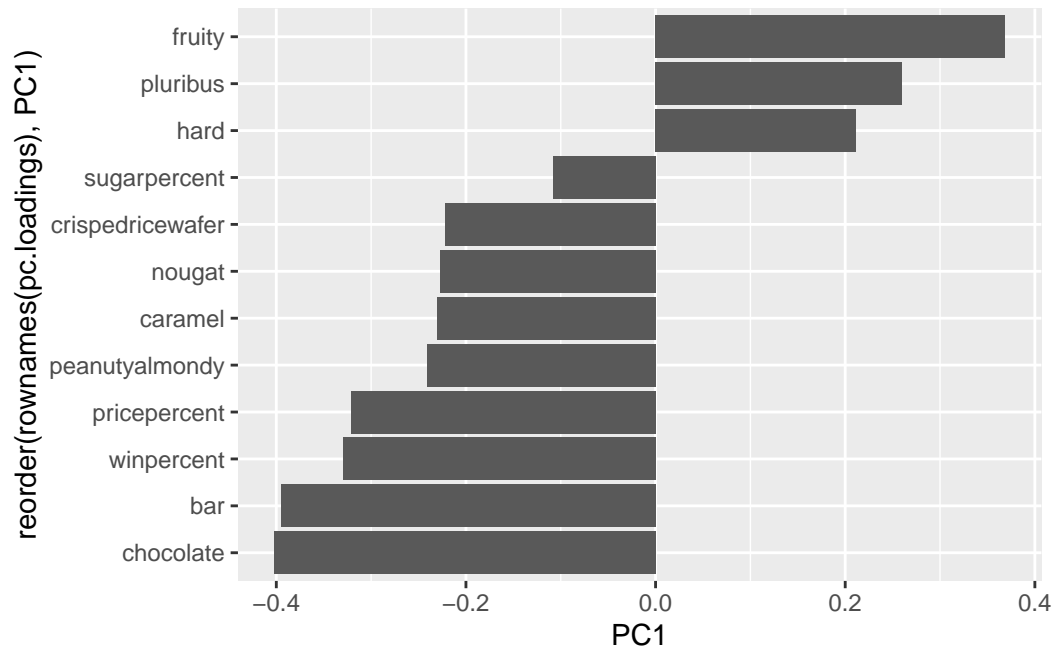
```
p <- ggplot(my_data) +
  aes(x=PC1, y=PC2,
      size=winpercent/100,
      text=rownames(my_data),
      label=rownames(my_data)) +
  geom_point(col=my_cols)
```

p



```
pc.loadings <- as.data.frame(pca$rotation)

ggplot(pc.loadings)+
  aes(PC1,reorder(rownames(pc.loadings), PC1))+
  geom_col()
```



Use the `ggrepel` package and the function `ggrepel::geom_text_repel()` to label up the plot with non overlapping candy names like.

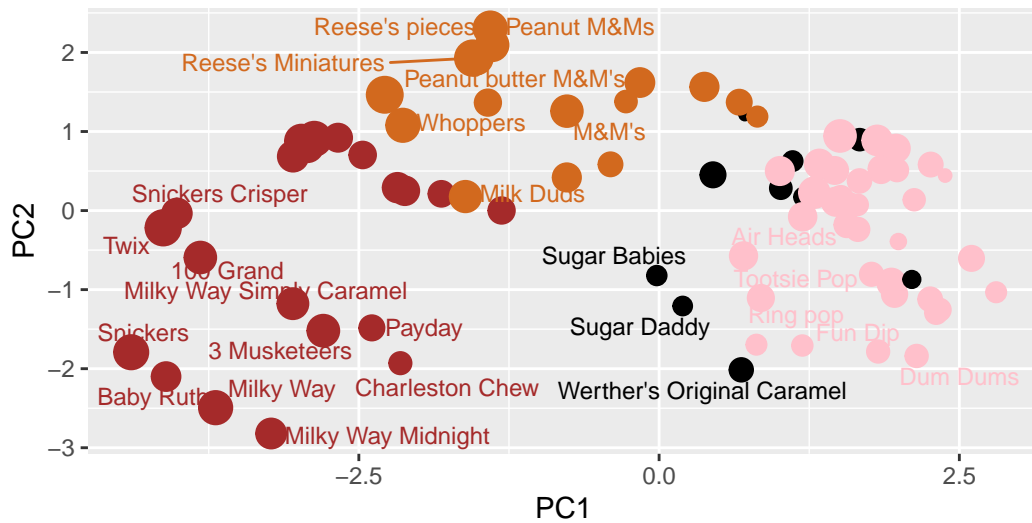
```
library(ggrepel)

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`: 59 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),



Data from 538

Pass the ggplot object `p` to `plotly` like so to generate an interactive plot that you can mouse over to see labels:

```
library(plotly)
```

Attaching package: 'plotly'

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

`last_plot`

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

`filter`

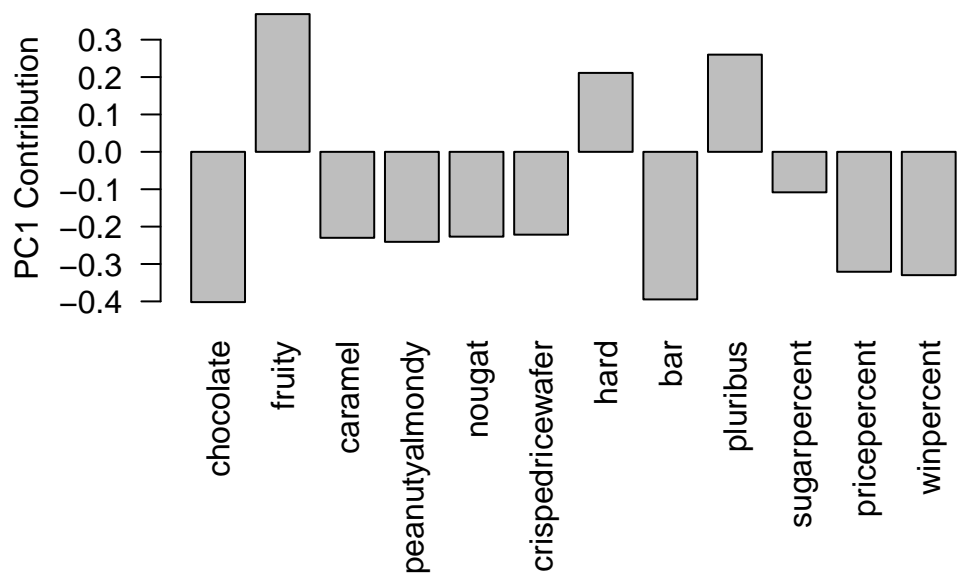
The following object is masked from 'package:graphics':

`layout`

```
#ggplotly(p)
```

Taking a quick look at PCA our loadings

```
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? Fruity, hard, and pluribus are picked up by PC1.