lab09:Halloween mini project

Yufei Zhang A16987415

2024-05-05

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

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

	chocolate	fruity	caramel	peanutyalmondy	nou…	crispedricewafer	h…	b..
	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
100 Grand	1	0	1	0	0	1	0	1
3 Musketeers	1	0	0	0	1	0	0	1
One dime	0	0	0	0	0	0	0	0
One quarter	0	0	0	0	0	0	0	0
Air Heads	0	1	0	0	0	0	0	0
Almond Joy	1	0	0	1	0	0	0	1

6 rows | 1-10 of 13 columns

Q1. How many different candy types are in this dataset?

```
nrow(candy)
```

```
## [1] 85
```

There are 85 types of candy.

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

```
sum(candy[,2])
```

```
## [1] 38
```

There are 38 fruity candy types.

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

```
## [1] 81.64291
```

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

```
candy["Sour Patch Kids", ]$winpercent
```

```
## [1] 59.864
```

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")
```

```
## Warning: package 'skimr' was built under R version 4.3.3
```

```
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	▇▁▁▁▂

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

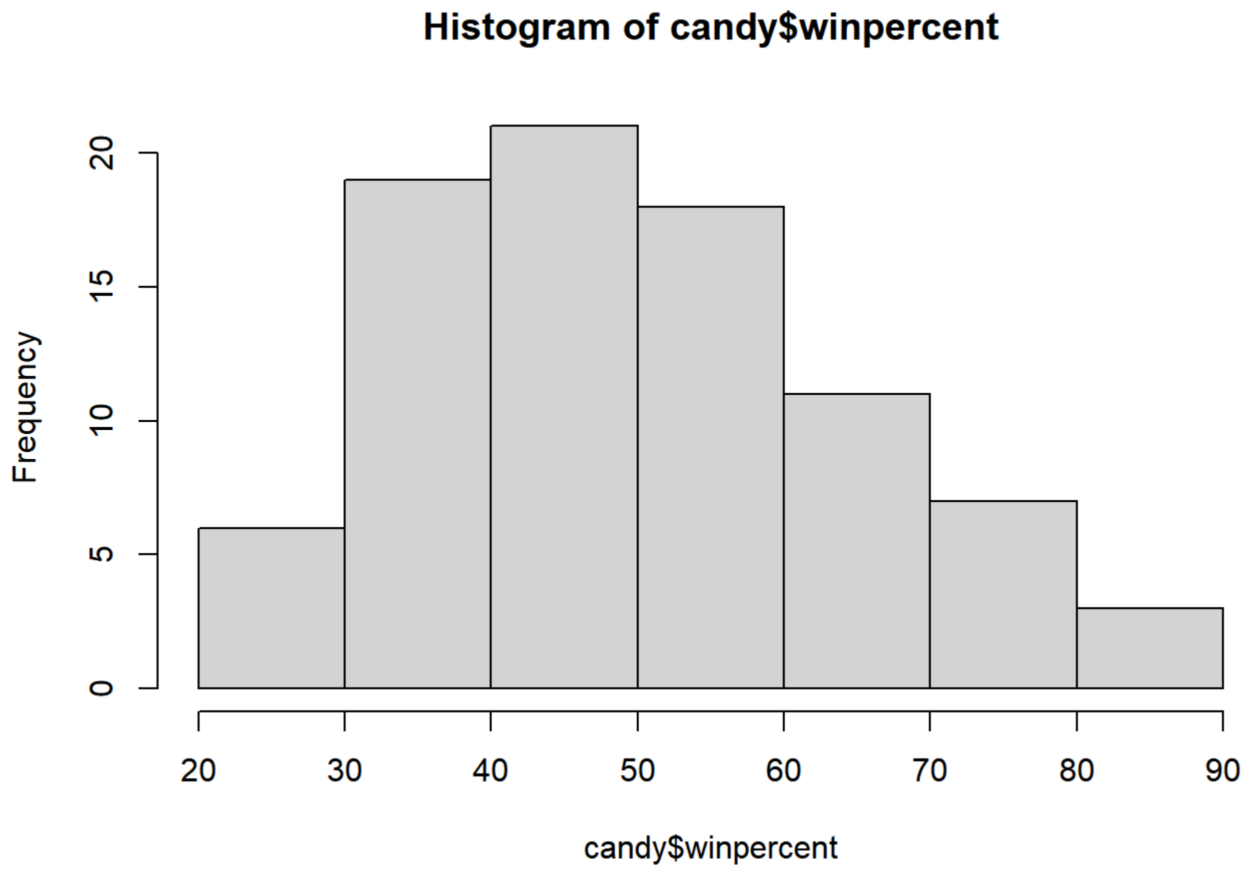
All percentages are continuous between 0 and 1 except winpercent looks to be on a different scale. It seems to be in % but not in decimal. All types column are either 0 or 1.

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

A zero means this candy type does not contain chocolate and a one means it contains chocolate.

Q8. Plot a histogram of winpercent values

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



Q9. Is the distribution of winpercent values symmetrical?

No, it is not symmetrical.

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

It is below 50%

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

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

chocolate candy is ranked higher.

Q12. Is this difference statistically significant?

Yes. Because the t value is 2.87 e-08 which is very very small.

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

```
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
```

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

	chocolate <int>	**fruity** <int>	**caramel** <int>	**peanutyalmondy** <int>	**nou…** <int>	**crispedricewafer** <int>	**h.** <i
Nik L Nip	0	1	0	0	0	0	
Boston Baked Beans	0	0	0	1	0	0	
Chiclets	0	1	0	0	0	0	
Super Bubble	0	1	0	0	0	0	
Jawbusters	0	1	0	0	0	0	

5 rows | 1-9 of 13 columns

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

	chocolate <int>	**fruity** <int>	**caramel** <int>	**peanutyalmondy** <int>	**nou…** <int>	**crispedricewafer** <int>	**h.** <i
Nik L Nip	0	1	0	0	0	0	
Boston Baked Beans	0	0	0	1	0	0	
Chiclets	0	1	0	0	0	0	
Super Bubble	0	1	0	0	0	0	

	chocolate	fruity	caramel	peanutyalmondy	nou…	crispedricewafer	h.
	<int>	<int>	<int>	<int>	<int>	<int>	<i
Jawbusters	0	1	0	0	0	0	

5 rows | 1-9 of 13 columns

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	nou…	crispedric
	<int>	<int>	<int>	<int>	<int>	
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	

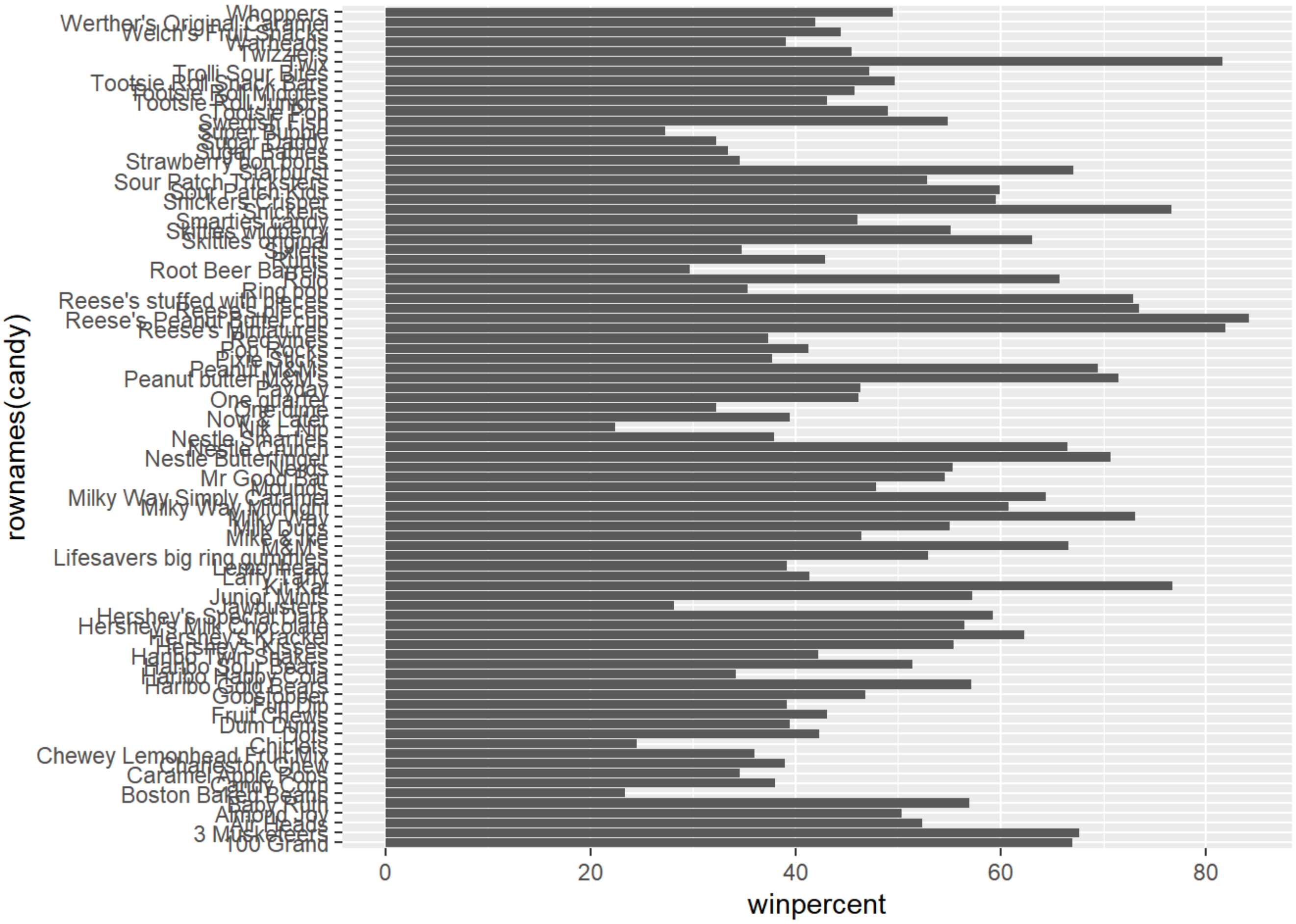
5 rows | 1-8 of 13 columns

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

```
library(ggplot2)
```

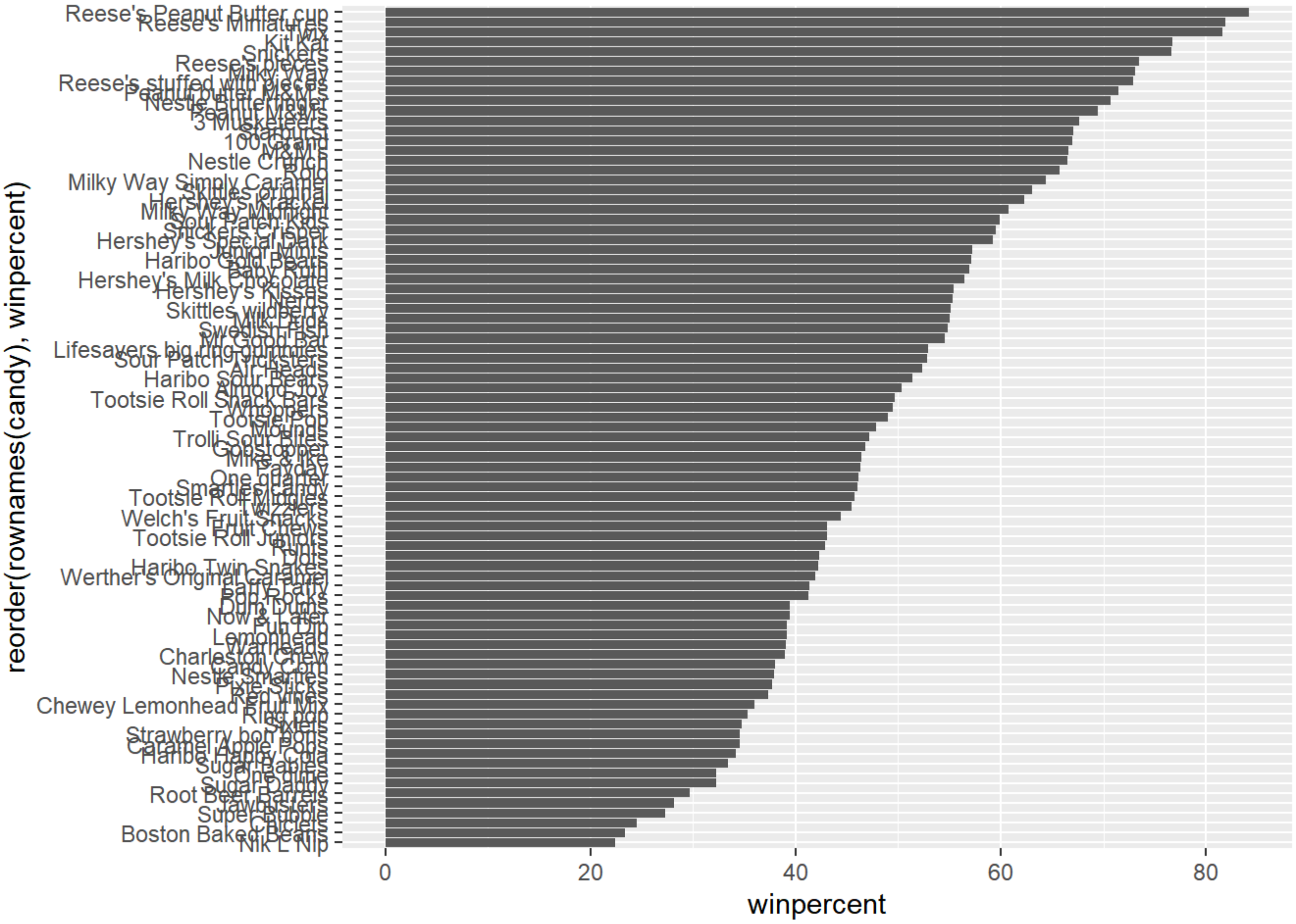
```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

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

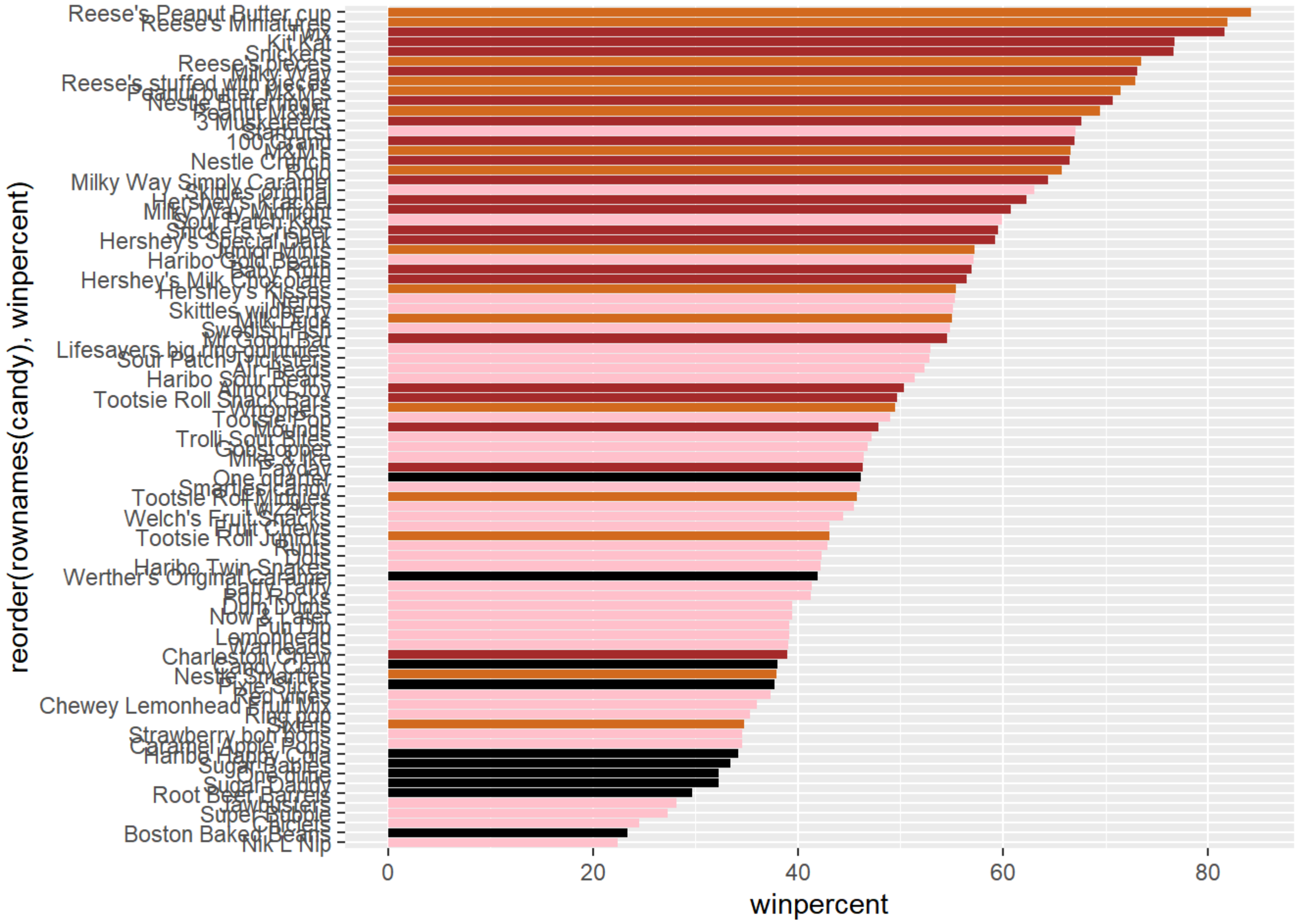


Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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



```
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)
```



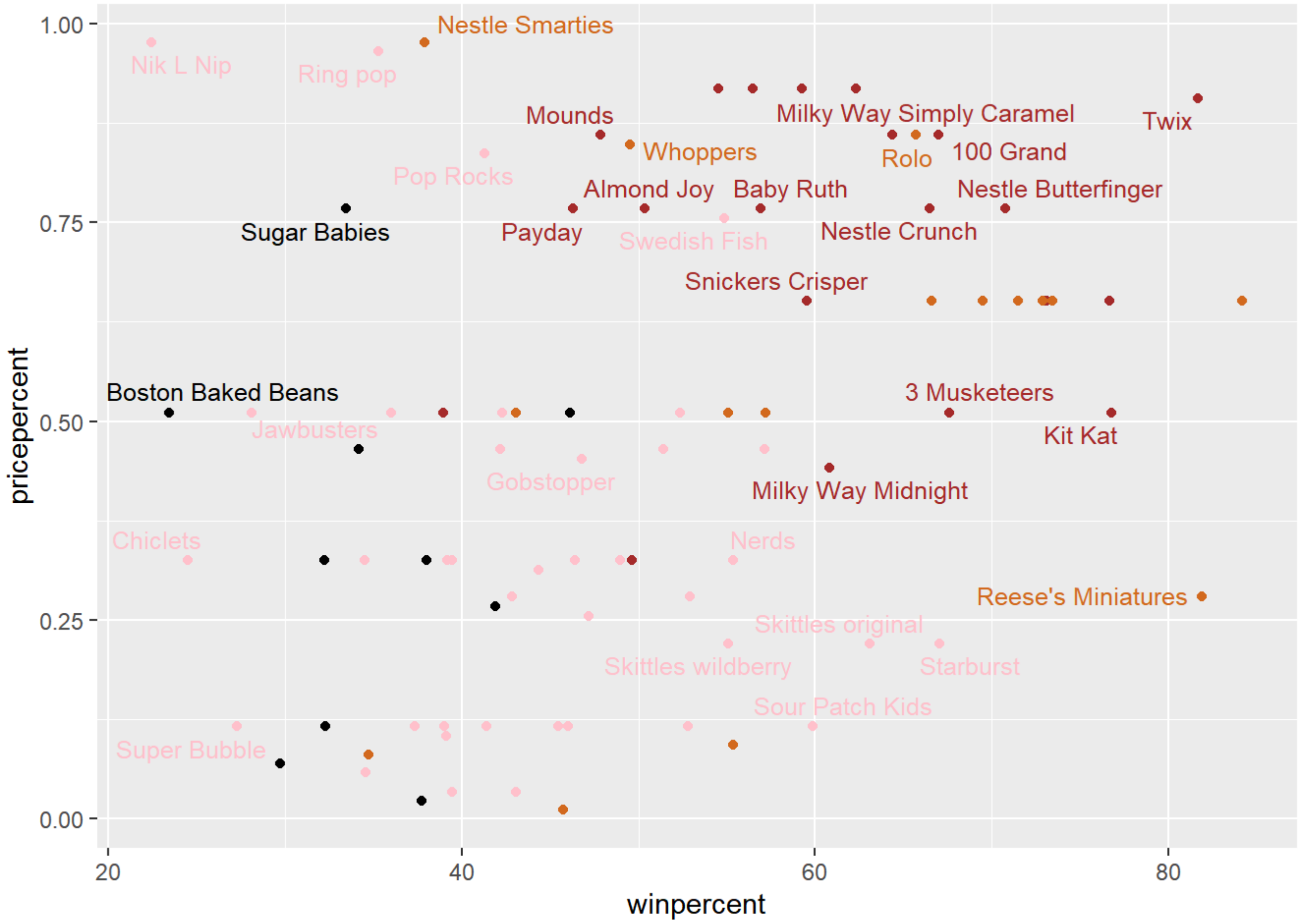
Now, for the first time, using this plot we can answer questions like: - Q17. What is the worst ranked chocolate candy?

Sixlets

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

```
library(ggrepel)

# 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 = 5)
```

```
## Warning: ggrepel: 53 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?

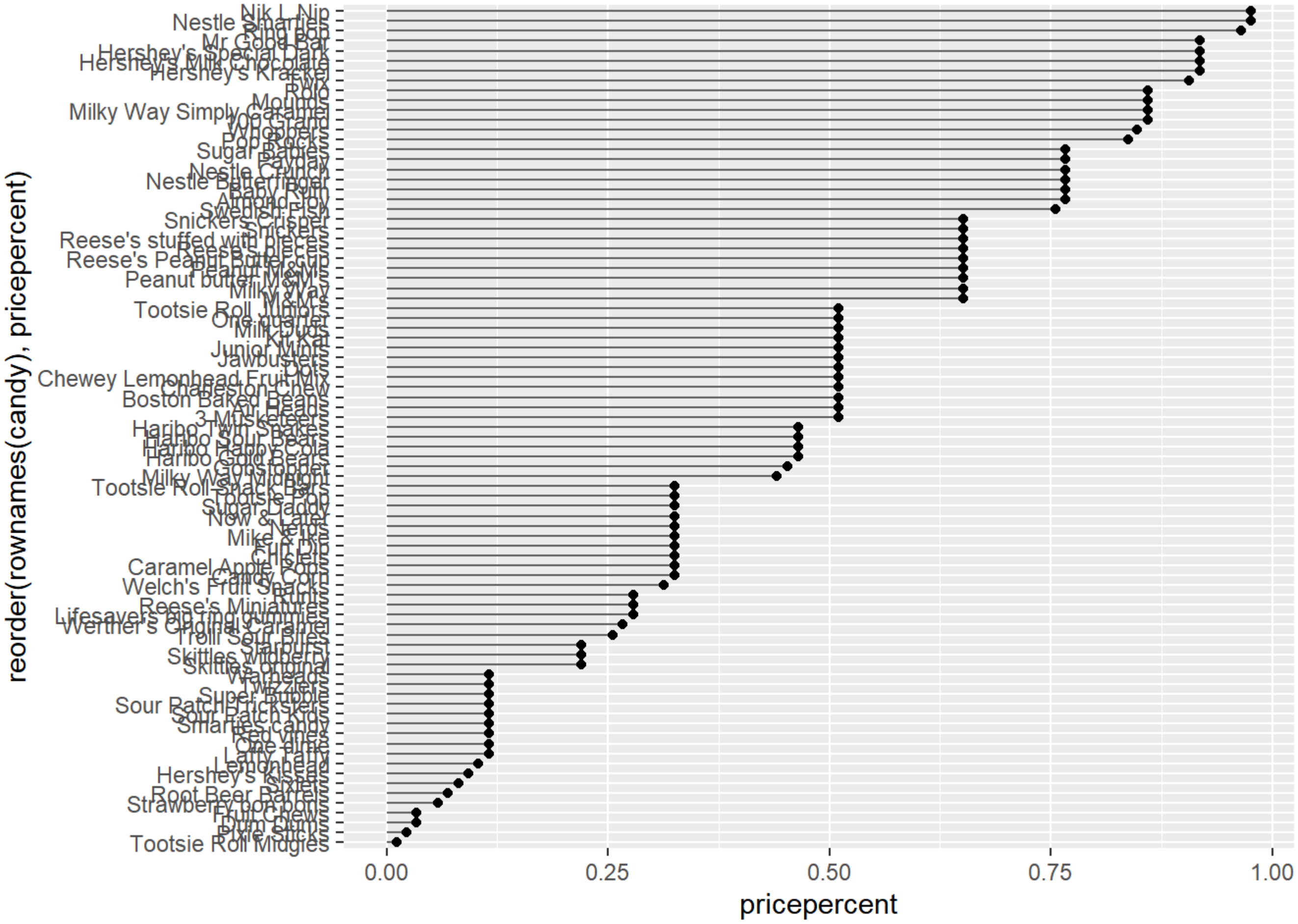
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

	pricepercent <dbl>	winpercent <dbl>
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

5 rows

Nik L Nip is lease popular among the 5.

```
# Make a lollipop chart of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```

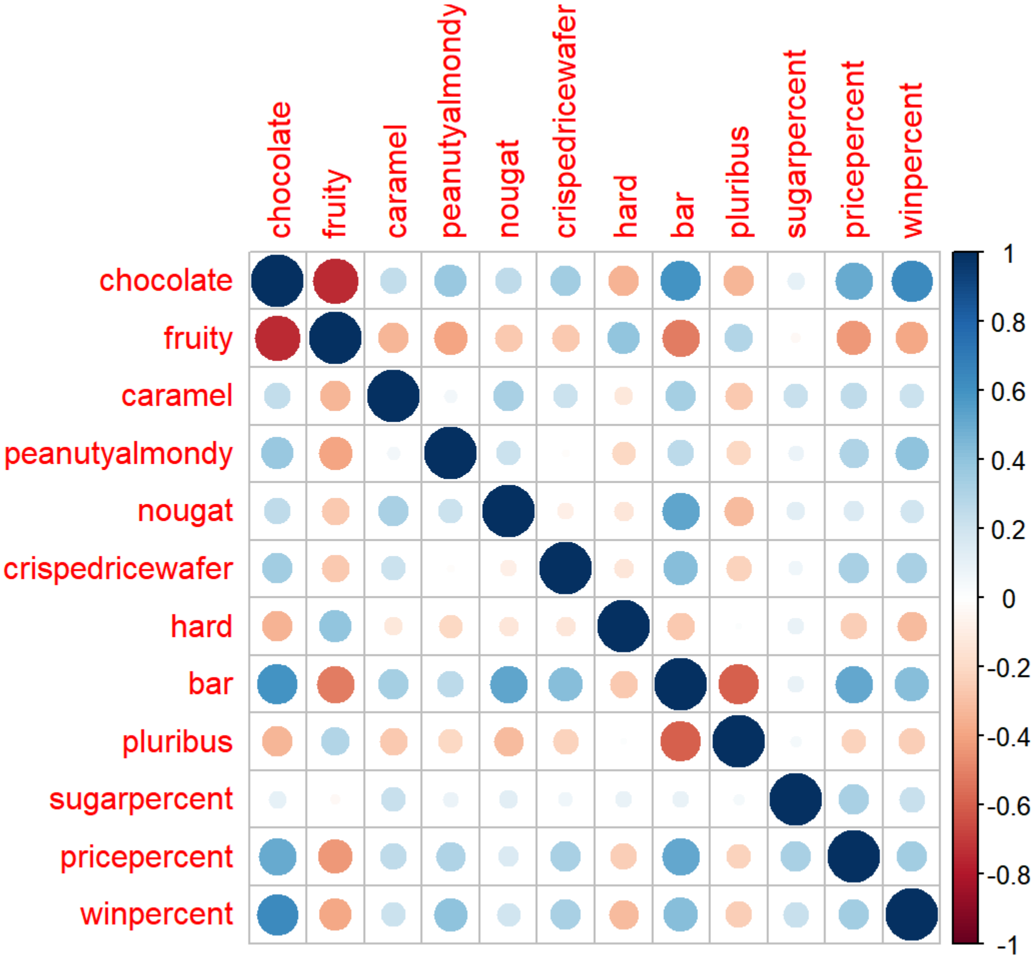


```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.92 loaded
```

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



Q22.

Examining this plot what two variables are anti-correlated (i.e. have minus values)? Fruity and chocolate,fruity and bar, pluribus and bar.

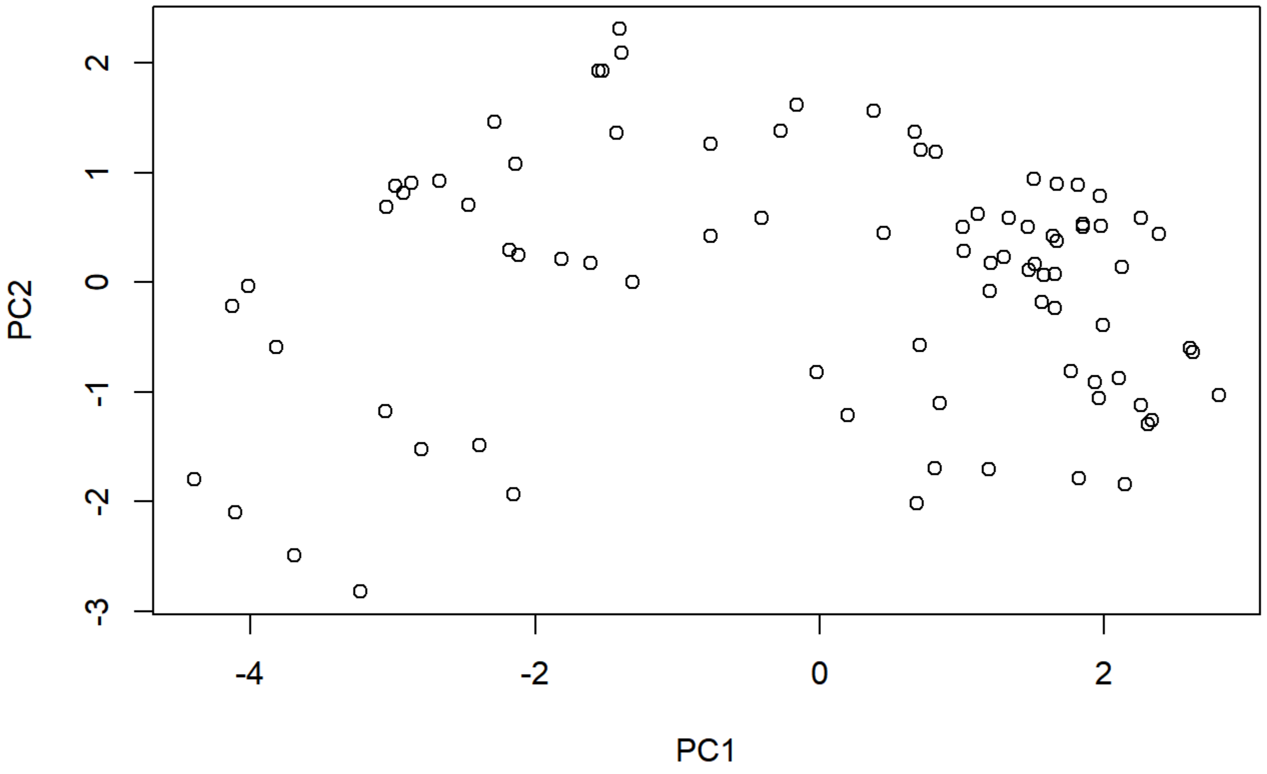
Q23. Similarly, what two variables are most positively correlated?

chocolate and winpercent, chocolate and bar.

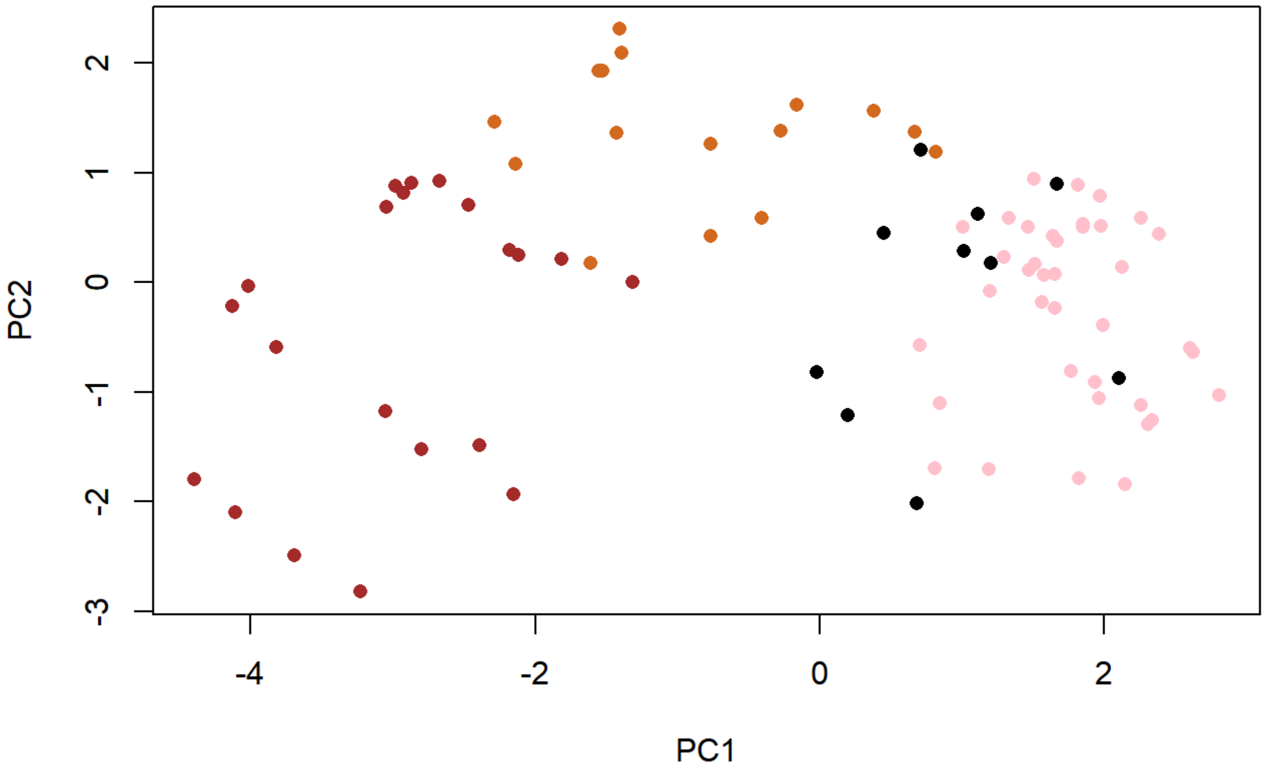
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
## 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
```

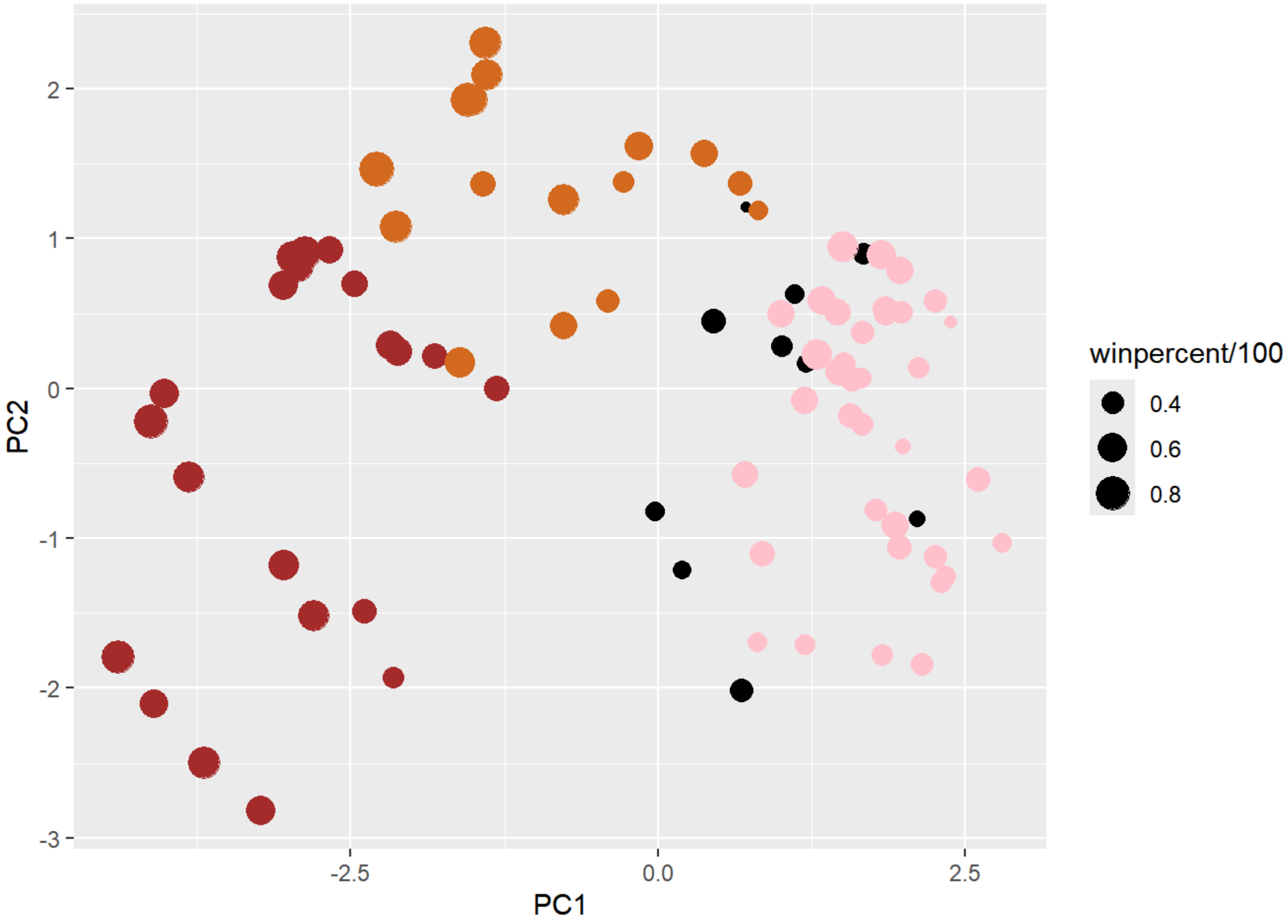
```
plot(pca$x[, 1:2])
```



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

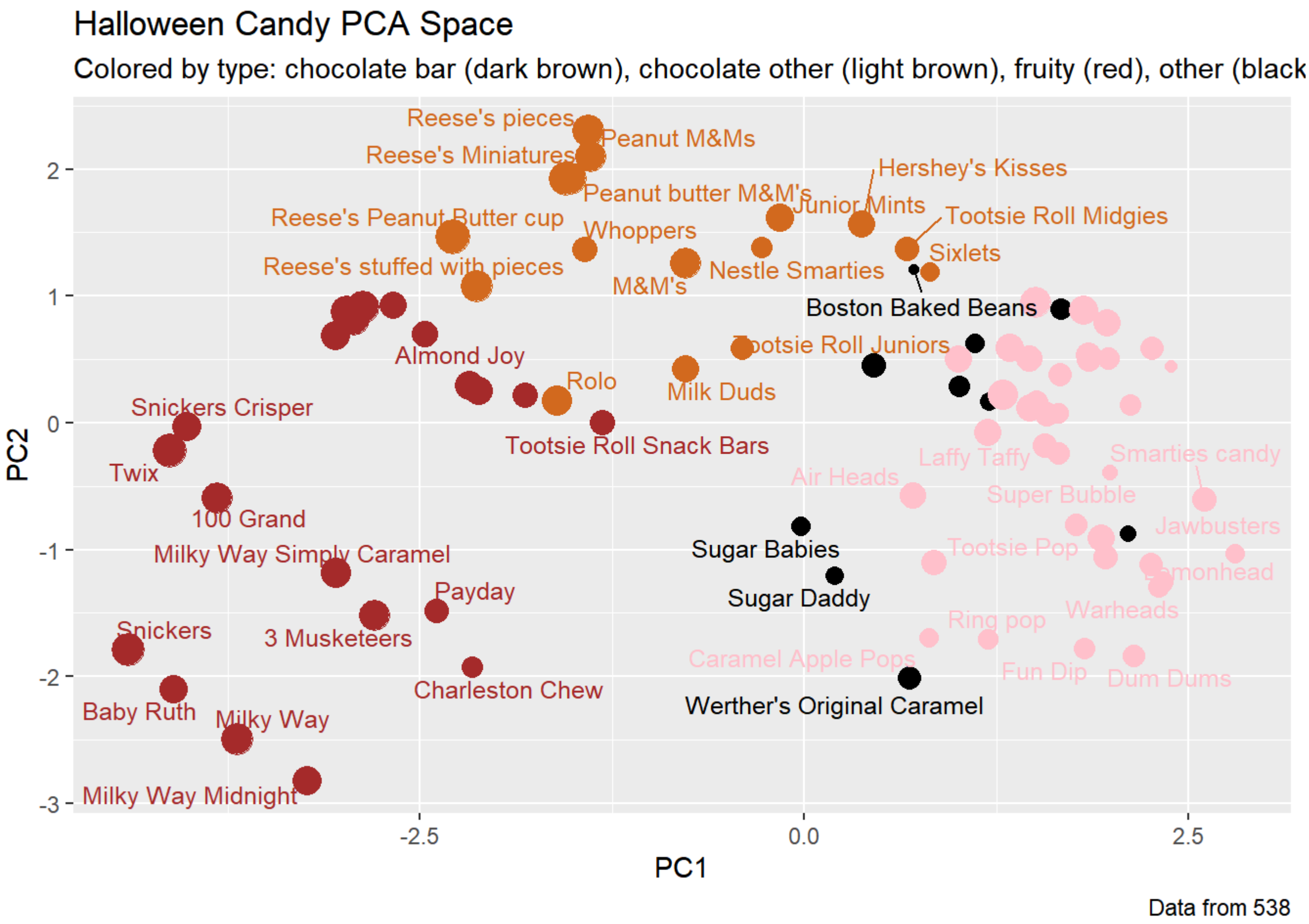


```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

p
```



```
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), fru
ity (red), other (black)",
       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

```
library(plotly)
```

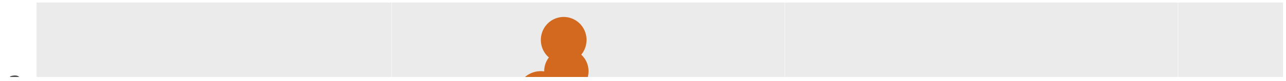
```
##
## Attaching package: 'plotly'
```

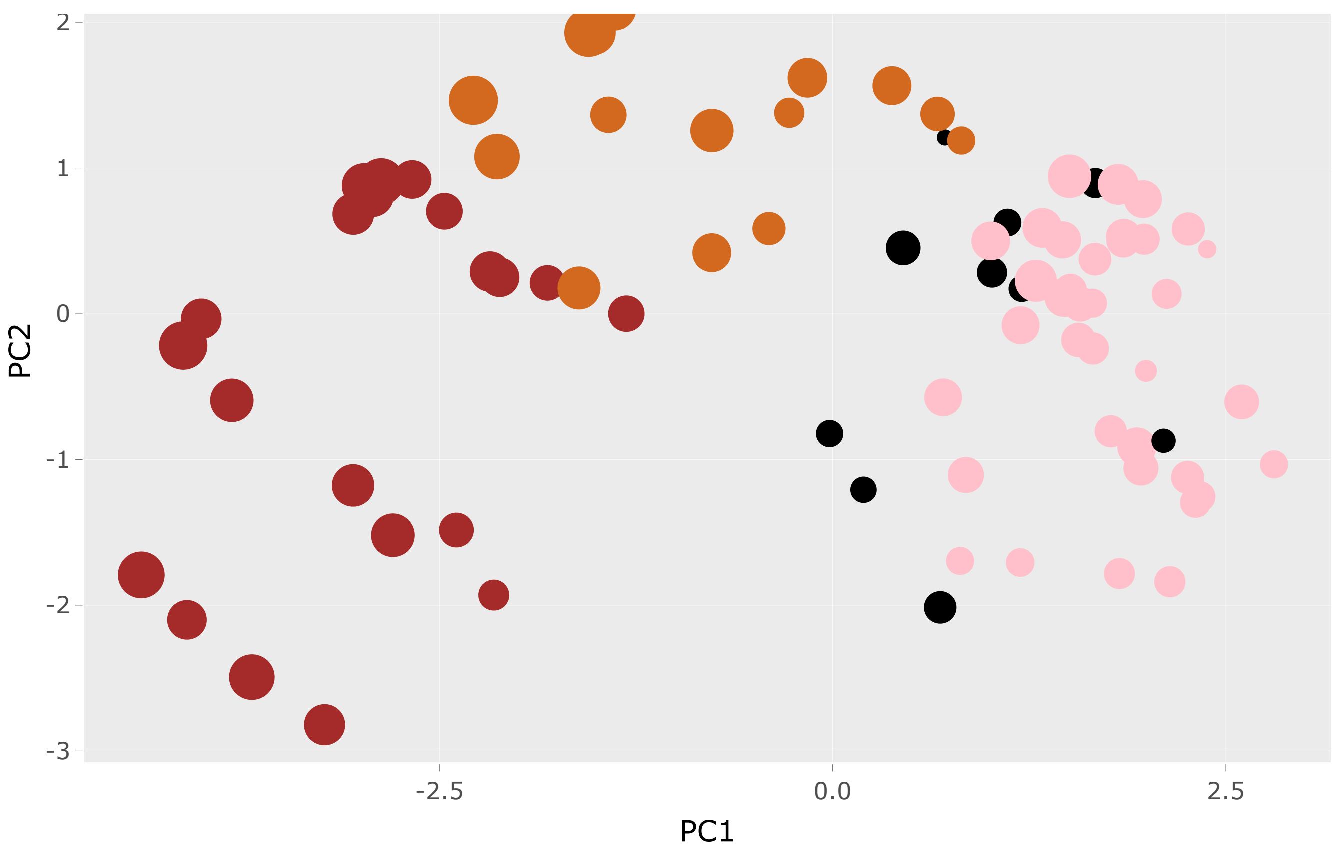
```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

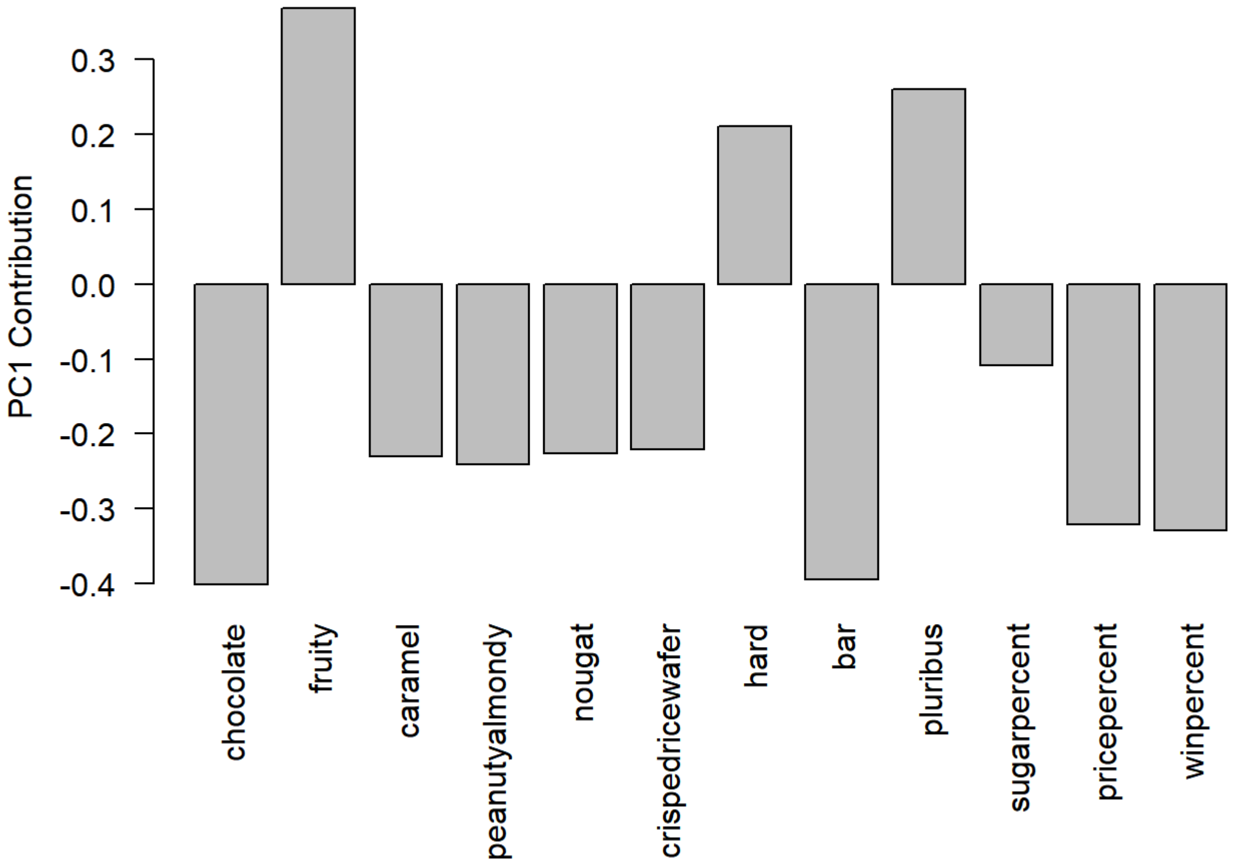
```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
ggplotly(p)
```





```
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? HINT. pluribus means the candy comes in a bag or box of multiple candies.

Fruity and pluribus. Yes since fruity candys usually comes in a bag or box of multiple candies.