Halloween project

Jaquish (PID: A59010386)

10/29/2021

```
#read.csv("candy-data.csv")
candy_file <- "candy-data.csv"
```

```
candy = read.csv("candy-data.csv", 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
```

```
row.names(candy)
```

```
##  [1] "100 Grand"                 "3 Musketeers"
##  [3] "One dime"                  "One quarter"
##  [5] "Air Heads"                 "Almond Joy"
##  [7] "Baby Ruth"                 "Boston Baked Beans"
##  [9] "Candy Corn"                "Caramel Apple Pops"
## [11] "Charleston Chew"           "Chewey Lemonhead Fruit Mix"
## [13] "Chiclets"                  "Dots"
## [15] "Dum Dums"                  "Fruit Chews"
## [17] "Fun Dip"                   "Gobstopper"
## [19] "Haribo Gold Bears"         "Haribo Happy Cola"
## [21] "Haribo Sour Bears"         "Haribo Twin Snakes"
## [23] "HersheyÕs Kisses"          "HersheyÕs Krackel"
## [25] "HersheyÕs Milk Chocolate"  "HersheyÕs Special Dark"
## [27] "Jawbusters"                "Junior Mints"
## [29] "Kit Kat"                   "Laffy Taffy"
## [31] "Lemonhead"                 "Lifesavers big ring gummies"
## [33] "Peanut butter M&MÕs"       "M&MÕs"
```

```
## [35] "Mike & Ike"                  "Milk Duds"
## [37] "Milky Way"                   "Milky Way Midnight"
## [39] "Milky Way Simply Caramel"    "Mounds"
## [41] "Mr Good Bar"                 "Nerds"
## [43] "Nestle Butterfinger"         "Nestle Crunch"
## [45] "Nik L Nip"                   "Now & Later"
## [47] "Payday"                      "Peanut M&Ms"
## [49] "Pixie Sticks"                "Pop Rocks"
## [51] "Red vines"                   "ReeseÕs Miniatures"
## [53] "ReeseÕs Peanut Butter cup"   "ReeseÕs pieces"
## [55] "ReeseÕs stuffed with pieces" "Ring pop"
## [57] "Rolo"                        "Root Beer Barrels"
## [59] "Runts"                       "Sixlets"
## [61] "Skittles original"           "Skittles wildberry"
## [63] "Nestle Smarties"             "Smarties candy"
## [65] "Snickers"                    "Snickers Crisper"
## [67] "Sour Patch Kids"             "Sour Patch Tricksters"
## [69] "Starburst"                   "Strawberry bon bons"
## [71] "Sugar Babies"                "Sugar Daddy"
## [73] "Super Bubble"                "Swedish Fish"
## [75] "Tootsie Pop"                 "Tootsie Roll Juniors"
## [77] "Tootsie Roll Midgies"        "Tootsie Roll Snack Bars"
## [79] "Trolli Sour Bites"           "Twix"
## [81] "Twizzlers"                   "Warheads"
## [83] "WelchÕs Fruit Snacks"        "WertherÕs Original Caramel"
## [85] "Whoppers"
```

```
nrow(candy)
```

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

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

85 different types of candy.

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

38 fruity candies

```
colnames(candy)
```

```
##  [1] "chocolate"        "fruity"           "caramel"          "peanutyalmondy"
##  [5] "nougat"           "crispedricewafer" "hard"             "bar"
##  [9] "pluribus"         "sugarpercent"     "pricepercent"     "winpercent"
```

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

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

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

My favorite is Butterfingers but they are not on the list so I'll go with Nerds at only 55%.

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

```
## [1] 55.35405
```

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

76%

```
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)
```

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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
winpercent	0	1	50.32	14.71	22.45	39.14	47.83	59.86	84.18	

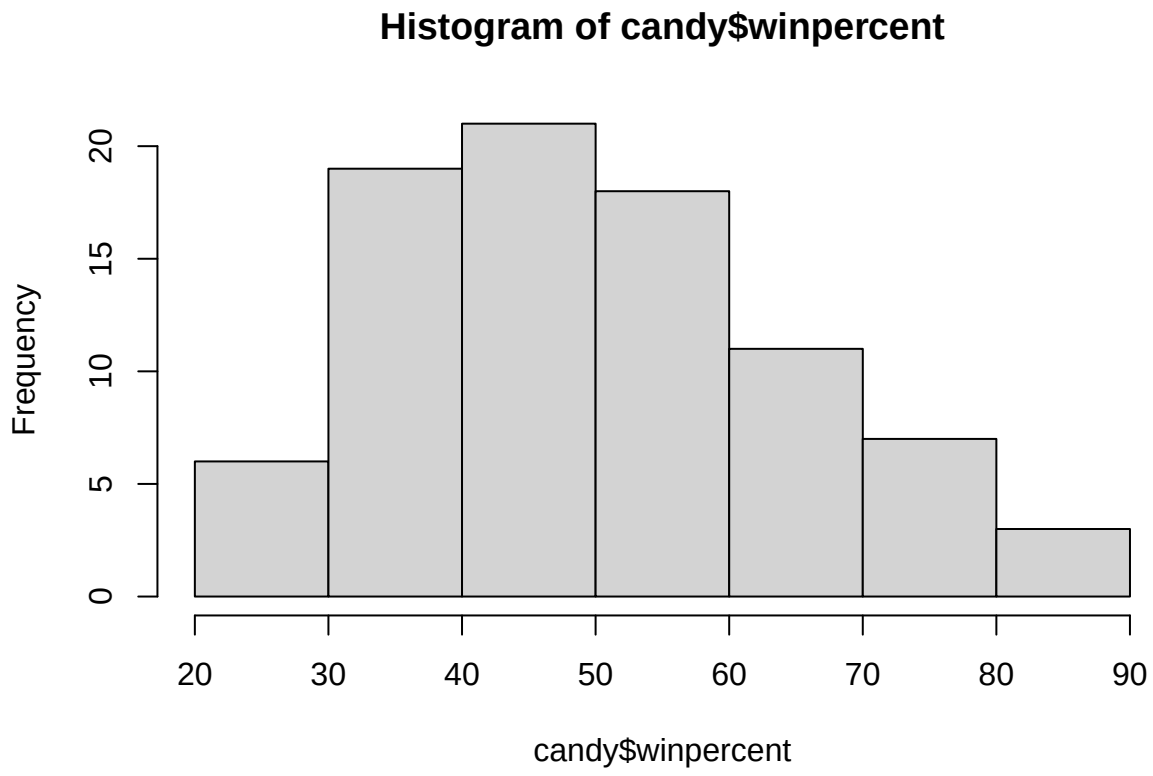
Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

The winpercent is on a different scale.

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

A 1 means that it was chosen over another candy while the zero means it lost against another candy.

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



Q9. Is the distribution of winpercent values symmetrical?

No 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?

The average of all winpercent values for chocolate is ranked higher than that of fruity.

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

```
## [1] 60.92153
```

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

```
## [1] 44.11974
```

Q12. Is this difference statistically significant?

```
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
```

Q12. Is this difference statistically significant?

It is signigicant based on p-value.

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

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

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

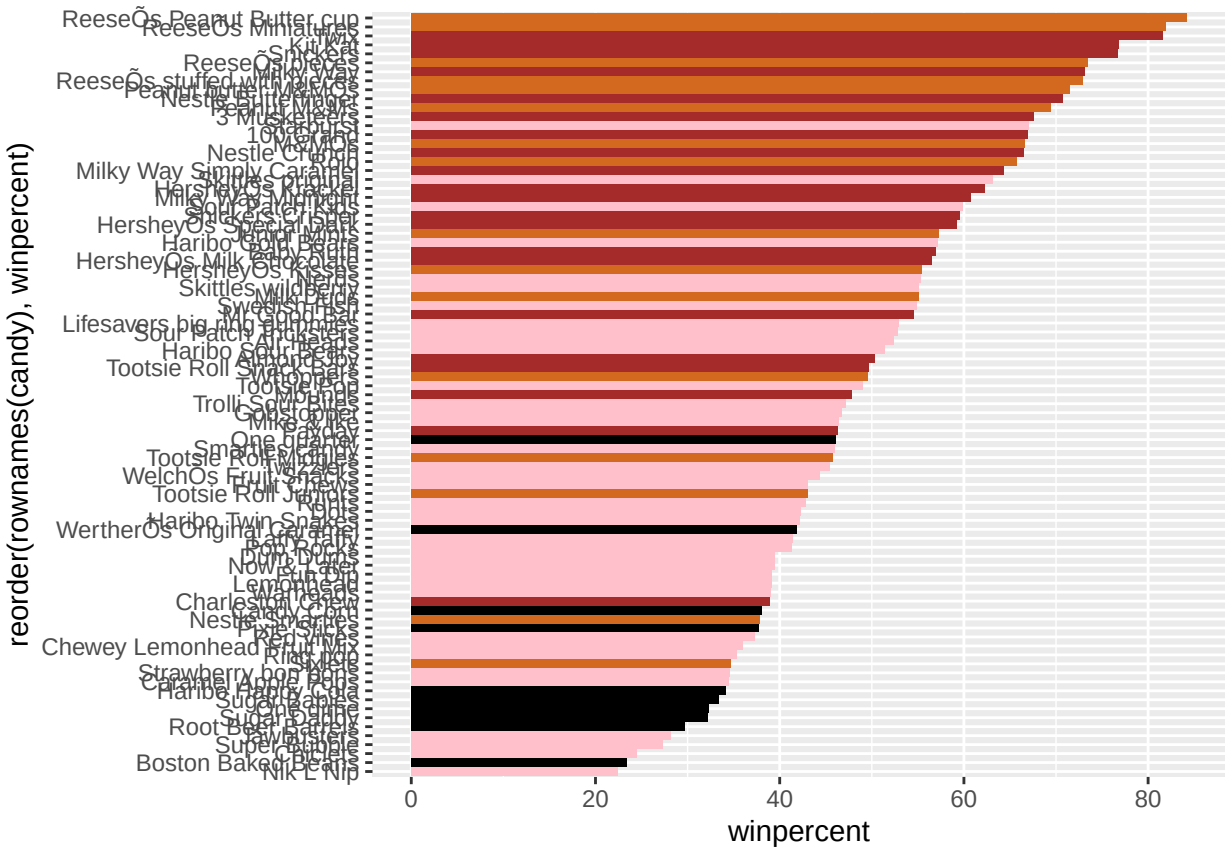
```
## [1] "ReeseÕs Peanut Butter cup" "ReeseÕs Miniatures"
## [3] "Twix"                      "Kit Kat"
## [5] "Snickers"
```

```
#library(dplyr)

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

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

```
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"
```

```
library(ggplot2)

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



> Q17. What is the worst ranked chocolate candy?

Nik L Lip

Q18. What is the best ranked fruity candy?

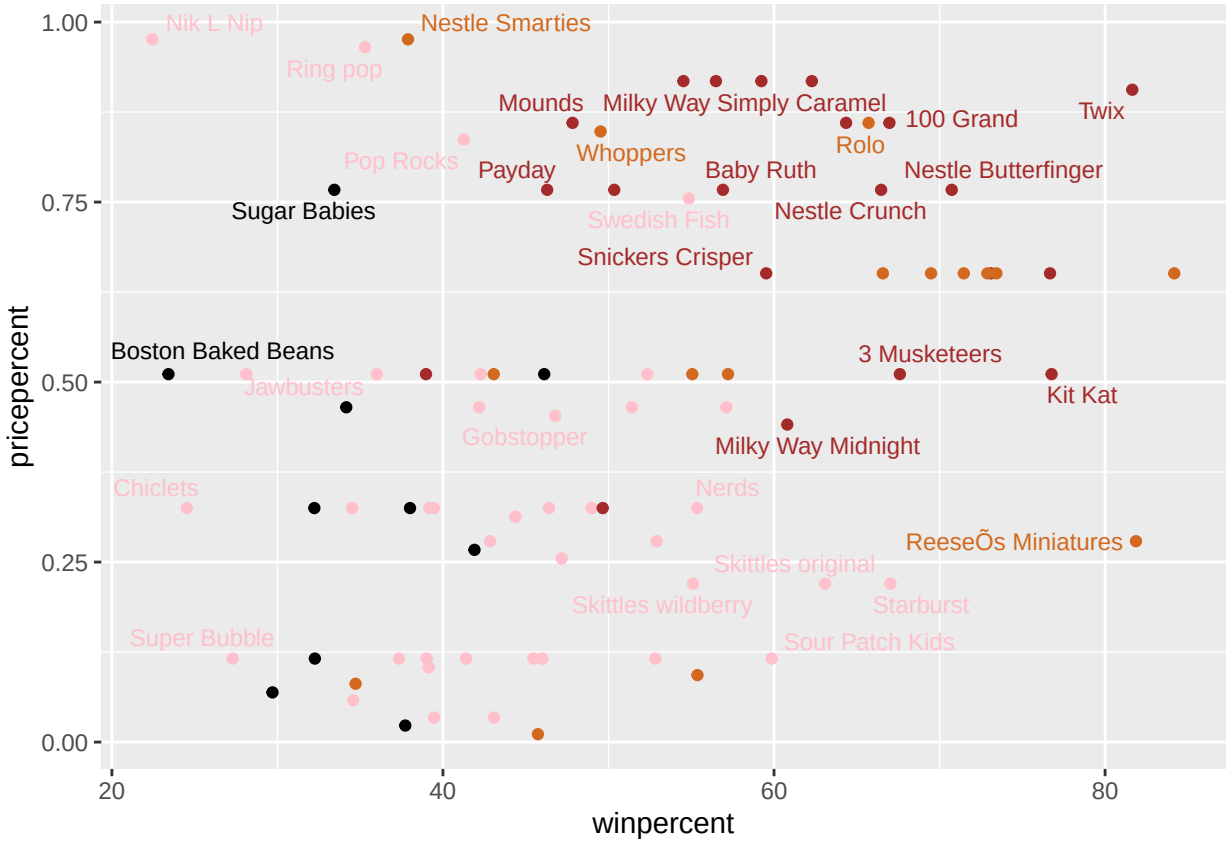
Starburst

```
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)
```

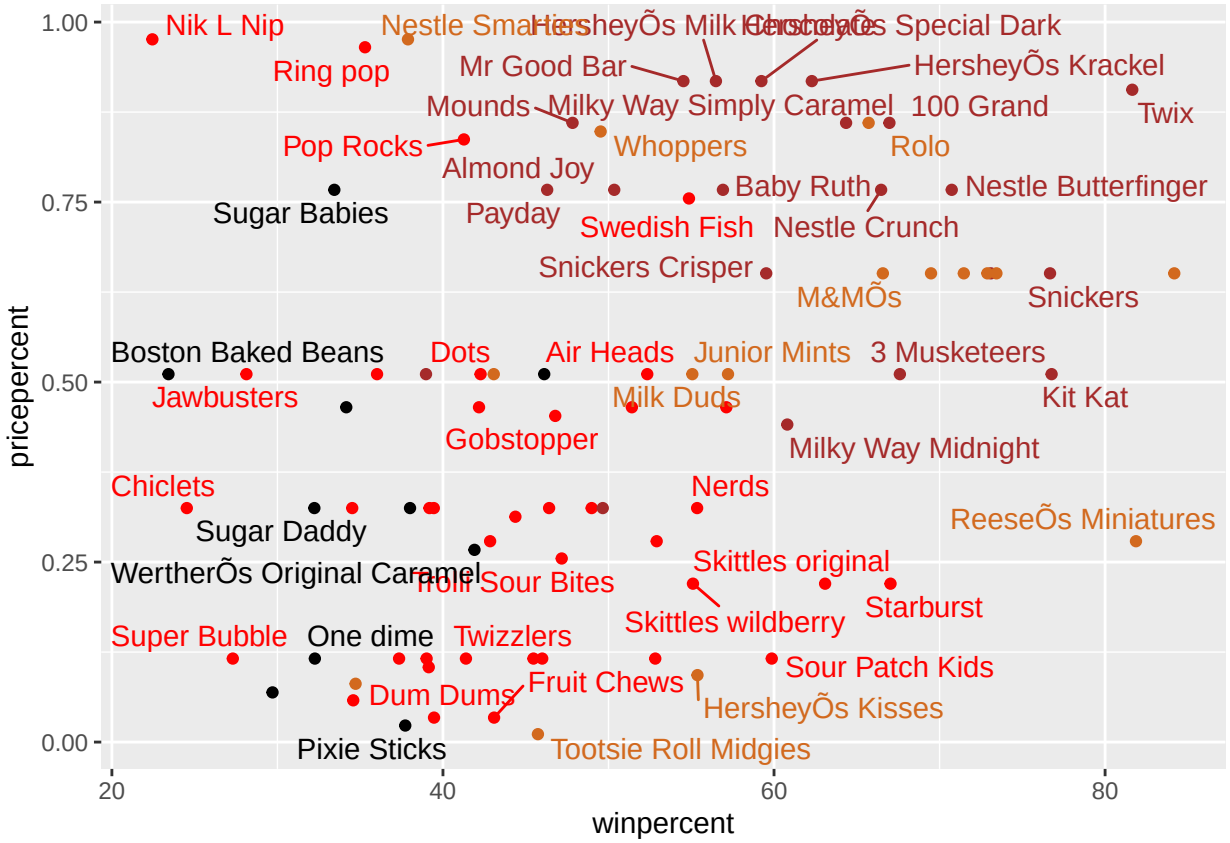
```
## Warning: ggrepel: 54 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



change my fruity color to red

```
my_cols[as.logical(candy$fruity)] <- "red"
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols)
```

```
## Warning: ggrepel: 33 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
gsub("Ö", "'", rownames(candy))
```

```
##  [1] "100 Grand"                 "3 Musketeers"
##  [3] "One dime"                  "One quarter"
##  [5] "Air Heads"                 "Almond Joy"
##  [7] "Baby Ruth"                 "Boston Baked Beans"
##  [9] "Candy Corn"                "Caramel Apple Pops"
## [11] "Charleston Chew"           "Chewey Lemonhead Fruit Mix"
## [13] "Chiclets"                  "Dots"
## [15] "Dum Dums"                  "Fruit Chews"
## [17] "Fun Dip"                   "Gobstopper"
## [19] "Haribo Gold Bears"         "Haribo Happy Cola"
## [21] "Haribo Sour Bears"         "Haribo Twin Snakes"
## [23] "HersheyÕs Kisses"          "HersheyÕs Krackel"
## [25] "HersheyÕs Milk Chocolate"  "HersheyÕs Special Dark"
## [27] "Jawbusters"                "Junior Mints"
## [29] "Kit Kat"                   "Laffy Taffy"
## [31] "Lemonhead"                 "Lifesavers big ring gummies"
## [33] "Peanut butter M&MÕs"       "M&MÕs"
## [35] "Mike & Ike"                "Milk Duds"
## [37] "Milky Way"                 "Milky Way Midnight"
## [39] "Milky Way Simply Caramel"  "Mounds"
## [41] "Mr Good Bar"               "Nerds"
## [43] "Nestle Butterfinger"       "Nestle Crunch"
## [45] "Nik L Nip"                 "Now & Later"
```

```
## [47] "Payday"                       "Peanut M&Ms"
## [49] "Pixie Sticks"                 "Pop Rocks"
## [51] "Red vines"                    "ReeseÕs Miniatures"
## [53] "ReeseÕs Peanut Butter cup"    "ReeseÕs pieces"
## [55] "ReeseÕs stuffed with pieces" "Ring pop"
## [57] "Rolo"                         "Root Beer Barrels"
## [59] "Runts"                        "Sixlets"
## [61] "Skittles original"            "Skittles wildberry"
## [63] "Nestle Smarties"              "Smarties candy"
## [65] "Snickers"                     "Snickers Crisper"
## [67] "Sour Patch Kids"              "Sour Patch Tricksters"
## [69] "Starburst"                    "Strawberry bon bons"
## [71] "Sugar Babies"                 "Sugar Daddy"
## [73] "Super Bubble"                 "Swedish Fish"
## [75] "Tootsie Pop"                  "Tootsie Roll Juniors"
## [77] "Tootsie Roll Midgies"         "Tootsie Roll Snack Bars"
## [79] "Trolli Sour Bites"            "Twix"
## [81] "Twizzlers"                    "Warheads"
## [83] "WelchÕs Fruit Snacks"         "WertherÕs Original Caramel"
## [85] "Whoppers"
```

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?

From this list it looks like Reeses Miniatures is the best bang for buck with high popularity.

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

```
##                            pricepercent winpercent
## ReeseÕs Peanut Butter cup        0.651   84.18029
## ReeseÕs Miniatures               0.279   81.86626
## Twix                             0.906   81.64291
## Kit Kat                          0.511   76.76860
## Snickers                         0.651   76.67378
```

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

Nik L Nips are most expensive and least popular.

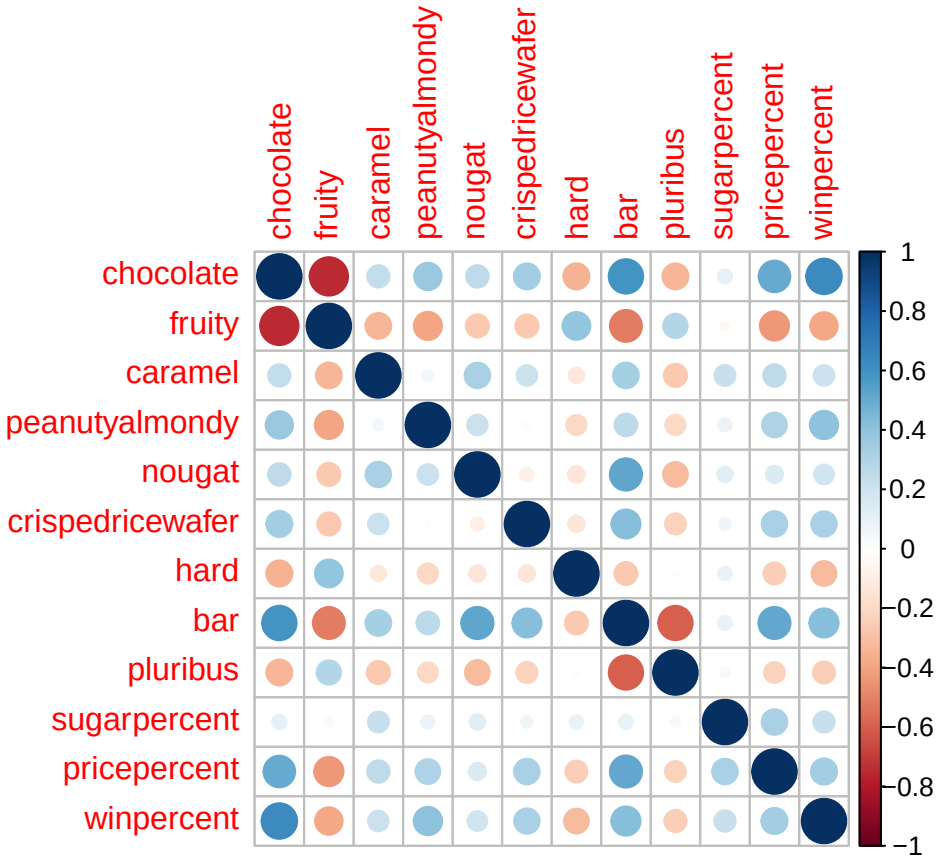
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

```
##                          pricepercent winpercent
## Nik L Nip                       0.976   22.44534
## Nestle Smarties                 0.976   37.88719
## Ring pop                        0.965   35.29076
## HersheyÕs Krackel               0.918   62.28448
## HersheyÕs Milk Chocolate        0.918   56.49050
```

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

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



Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)?

Fruity and chocolate are not correlated. Bar and pluribus as well.

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

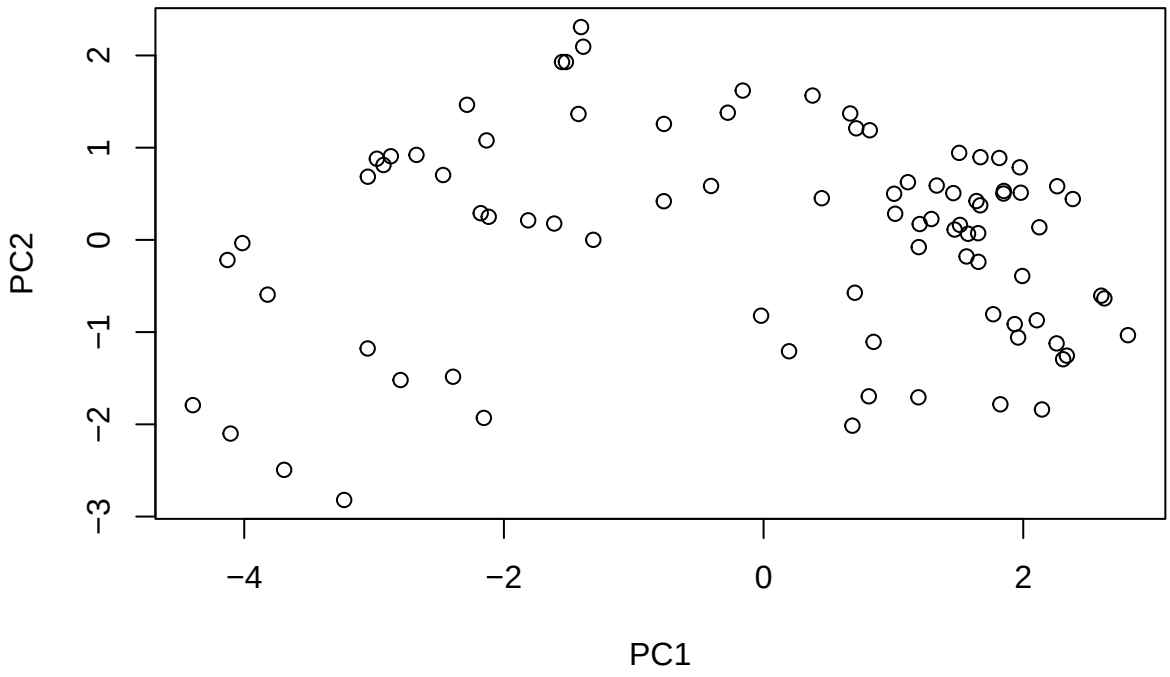
Chocolate and bar. Chocolate and winpercent. Chocolate and pricepercent.

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

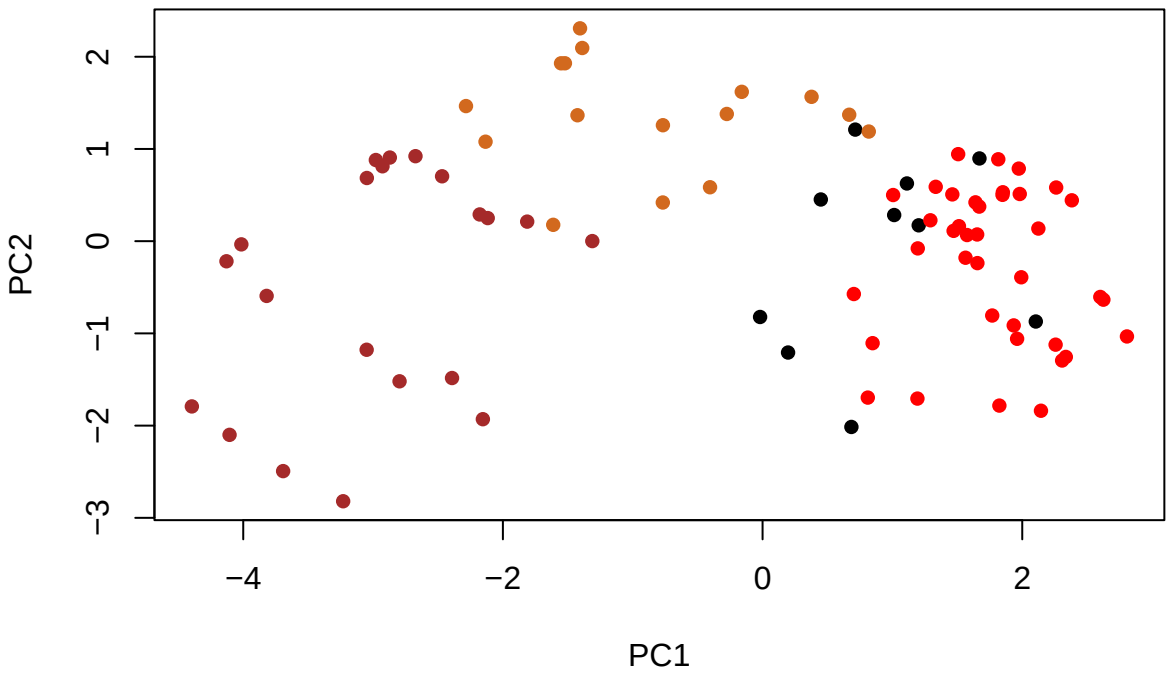
```
## Importance of components:
##                           PC1    PC2    PC3     PC4    PC5     PC6     PC7
## Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
## Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
```

```
## 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])
```

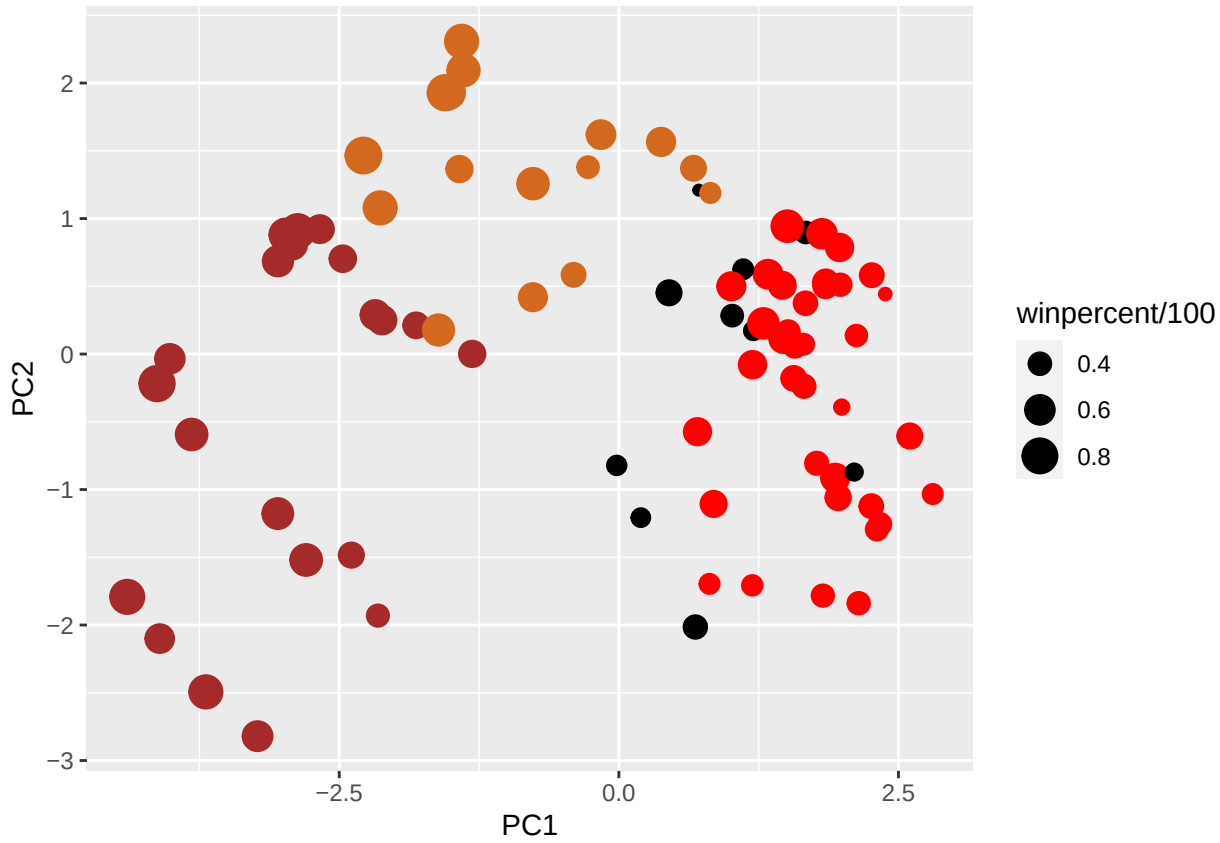



```
plot(pca$x[,1:2], col=my_cols, pch=16)
```



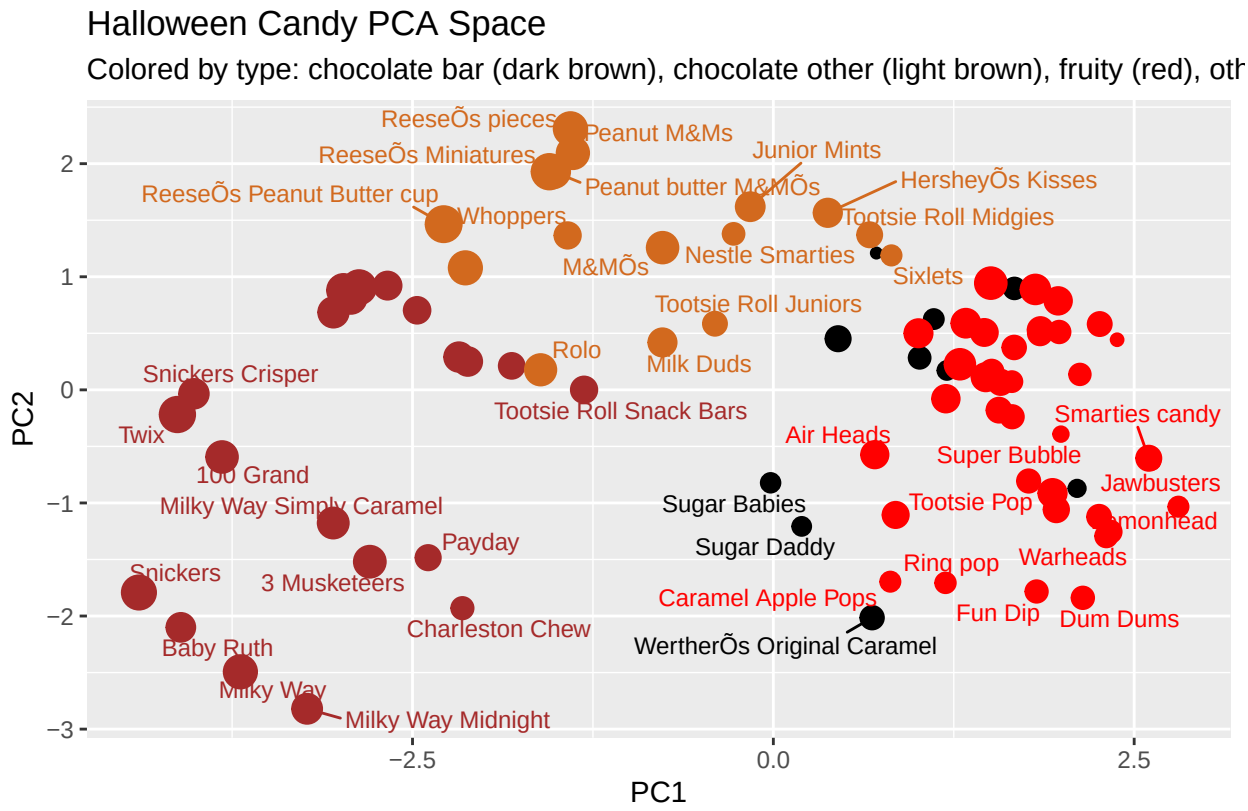

```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
```

```
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
```

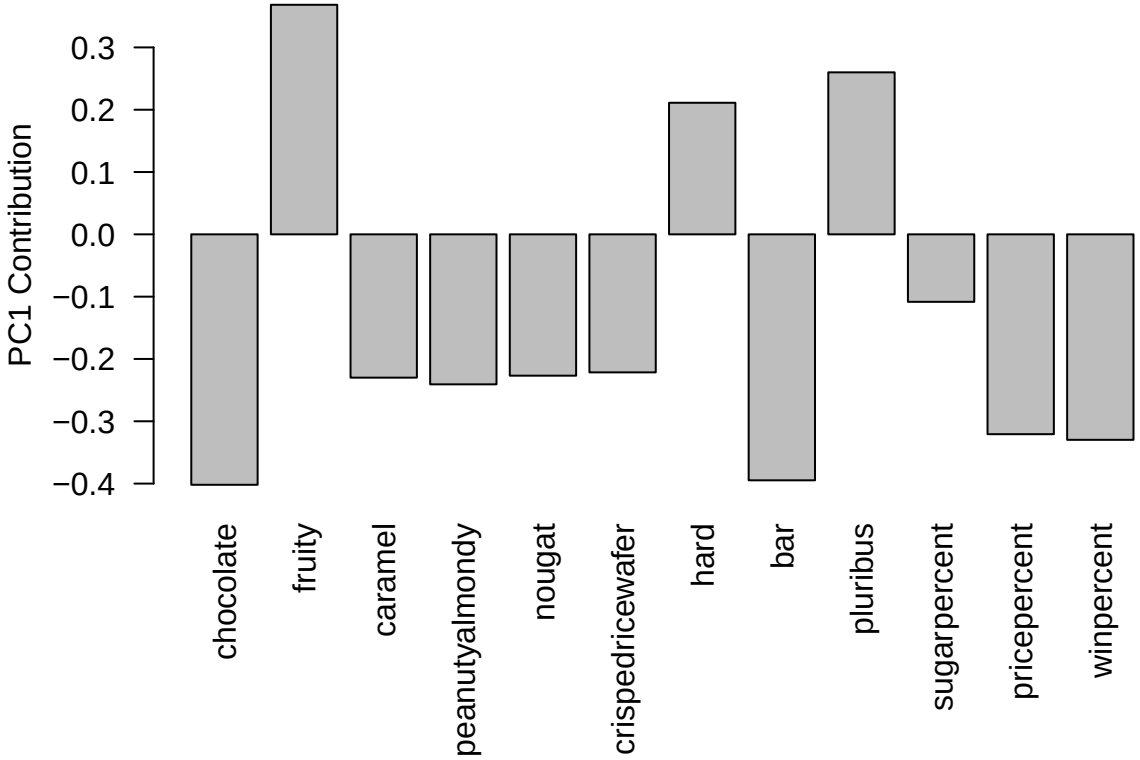



```
#library(ggrepel)

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), fruity (re
       caption="Data from 538")
```

```
## Warning: ggrepel: 44 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")
```



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

The fruity, hard and pluribus. This makes sense because most fruity candy tends to come in bags and are hard.