Halloween Mini Project

Audrey Ting Zhu (A16898668)

2024-11-03

##Class 10: Halloween Mini-Project

```
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? ANS: There are 85 candy types.

```
nrow(candy)
```

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

Q2. How many fruity candy types are in the dataset? ANS: There are 38 fruity candy types.

```
sum(candy$fruity==1)
```

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

Q3. What is your favorite candy in the dataset and what is it’s winpercent value? ANS: My favorite candy is Air Heads. The winpercentage is 52.34146.

```
candy["Air Heads", ]$winpercent
```

```
## [1] 52.34146
```

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

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

```
## [1] 76.7686
```

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

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

```
## [1] 49.6535
```

```
library("skimr")
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	▇▁▁▁▂
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	▇▁▁▁▁

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

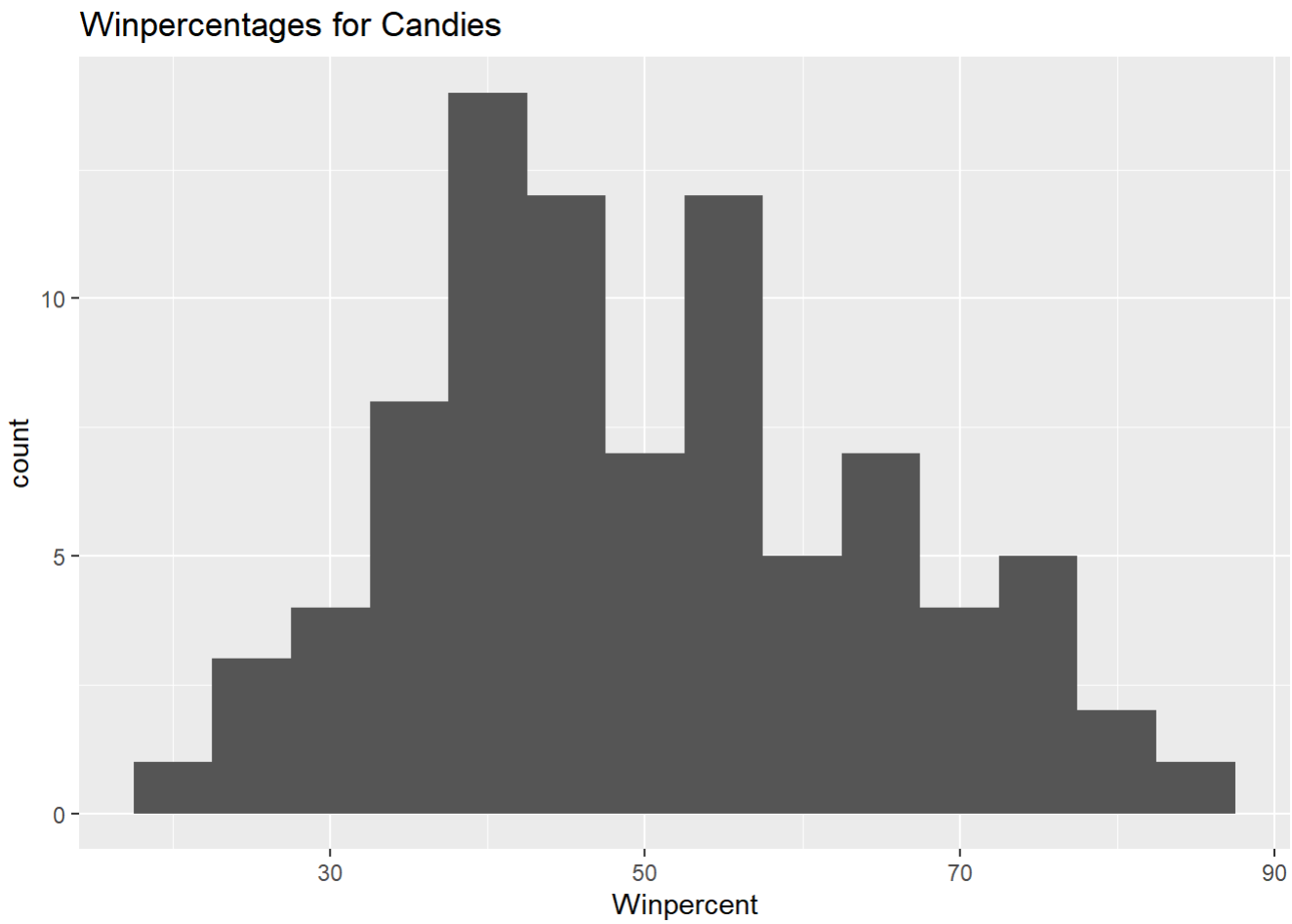
ANS: Winpercent is in a different range as it is a percentage score ranging up to 100. The other variables scale up to 1.

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

ANS: 0 means that there is no chocolate in the candy composition. 1 means that there is chocolate in the candy.

Q8. Plot a histogram of winpercent values

```
library(ggplot2)

ggplot(candy, aes(x=winpercent))+geom_histogram(binwidth=5)+labs(title="Winpercentages for Candies", x="Winpercent")
```



Q9. Is the distribution of winpercent values symmetrical?

ANS: The distribution of winpercent values are not 100 percent symmetrical, but it is roughly.

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

ANS: Depends if you look at mean or median. The mean is slightly above 50 but the median is 47.82, which is below 50%.

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

```
## [1] 50.31676
```

```
median(candy$winpercent)
```

```
## [1] 47.82975
```

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

ANS: Chocolate candy has an higher score.

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

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

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

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

Q12. Is this difference statistically significant?

```
t.test(candy$winpercent[as.logical(candy$chocolate)],candy$winpercent[as.logical(candy$fruit
y)])
```

```
##
##  Welch Two Sample t-test
##
## data:  candy$winpercent[as.logical(candy$chocolate)] and candy$winpercent[as.logical(candy$f
ruity)]
## 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
```

ANS: P-value is less than 0.05. It is pretty significant.

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

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

```
##                           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
##                           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
##                           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? ANS: The benefit of dplyr is that it has much simpler and readable syntax. However, you also need to download a package to make it work.

```
library("dplyr")
```

```
##
## 载入程序包：'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

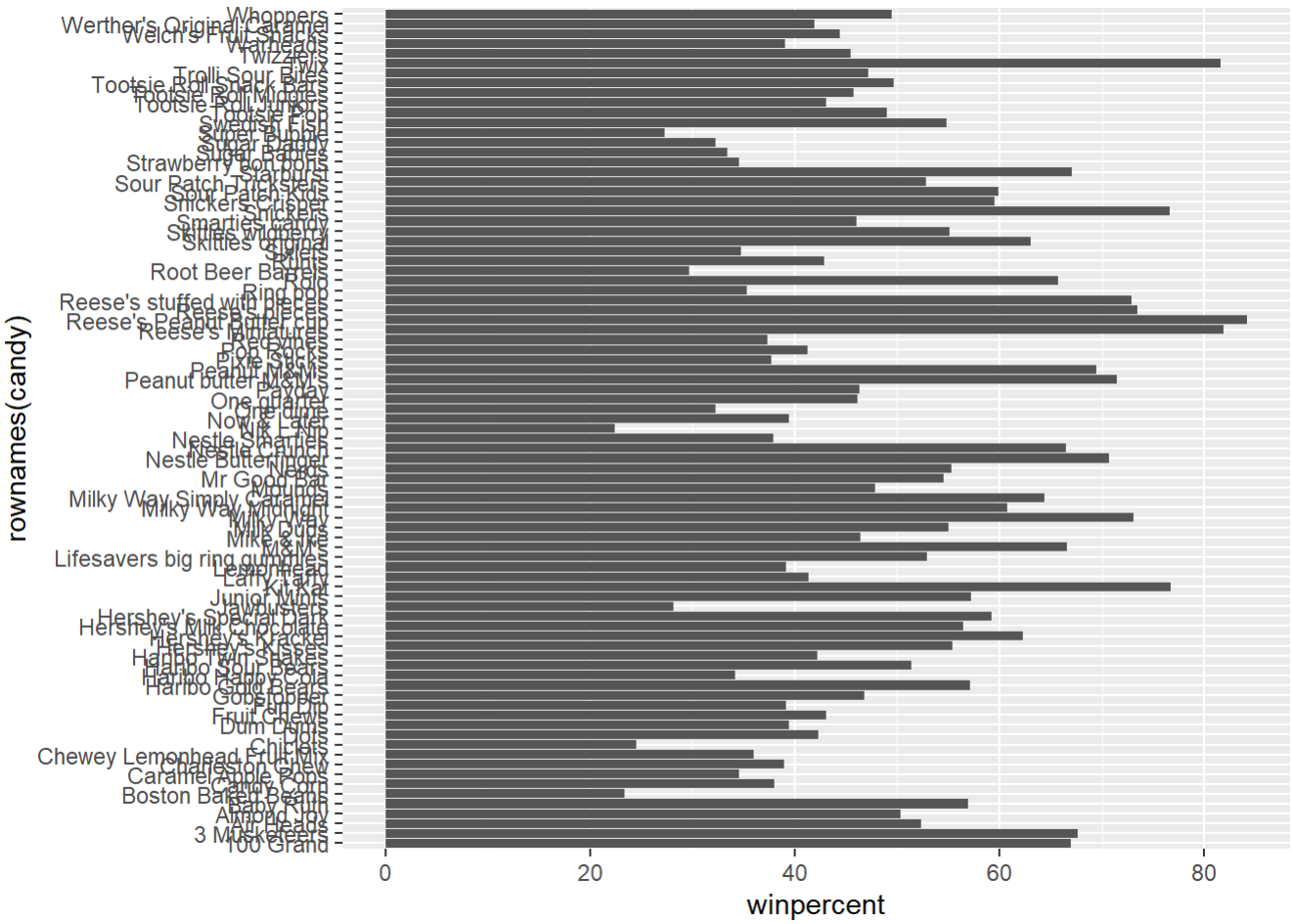
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

```
##                     chocolate fruity caramel 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
```

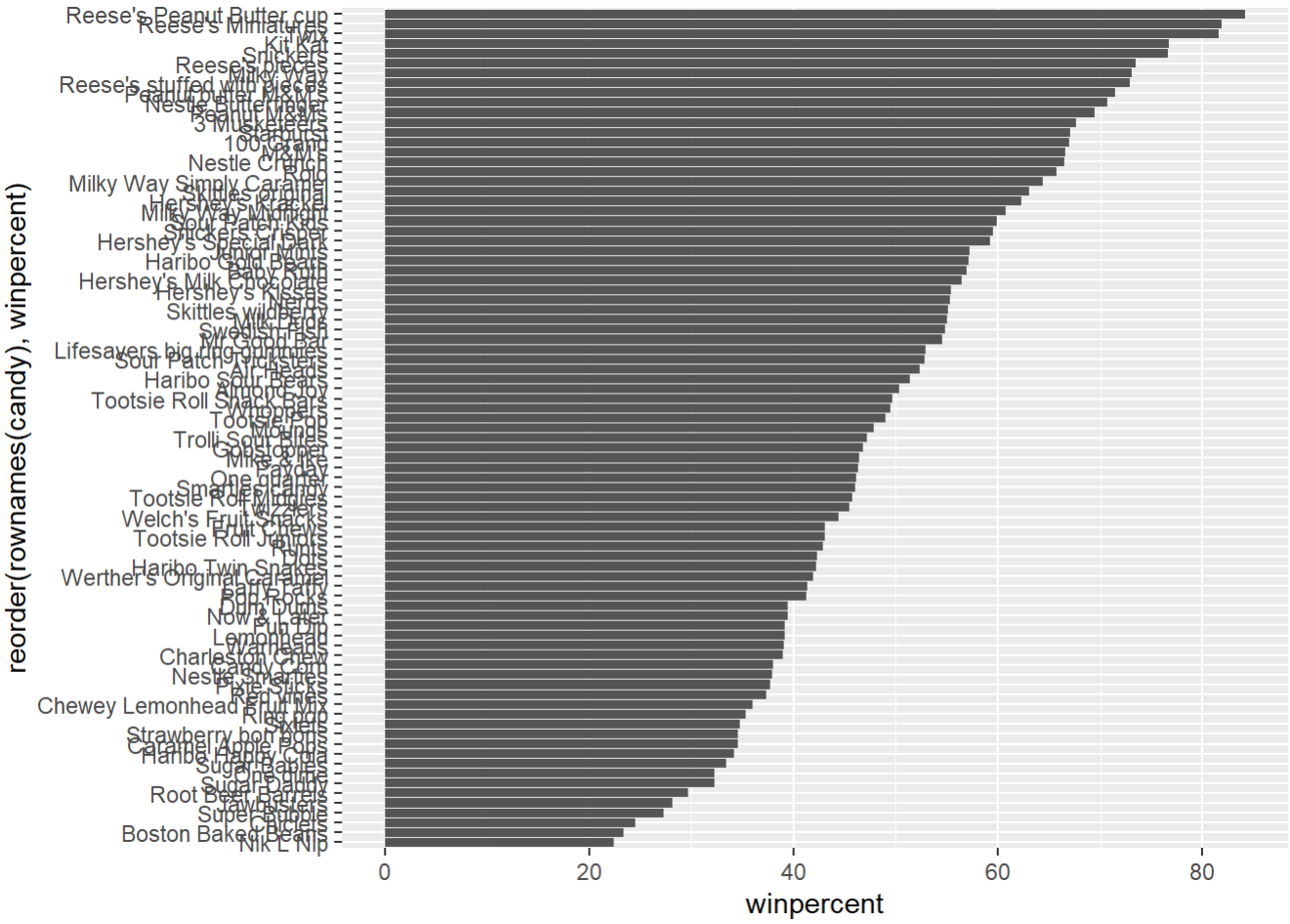
Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)

ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_bar(stat="identity")
```

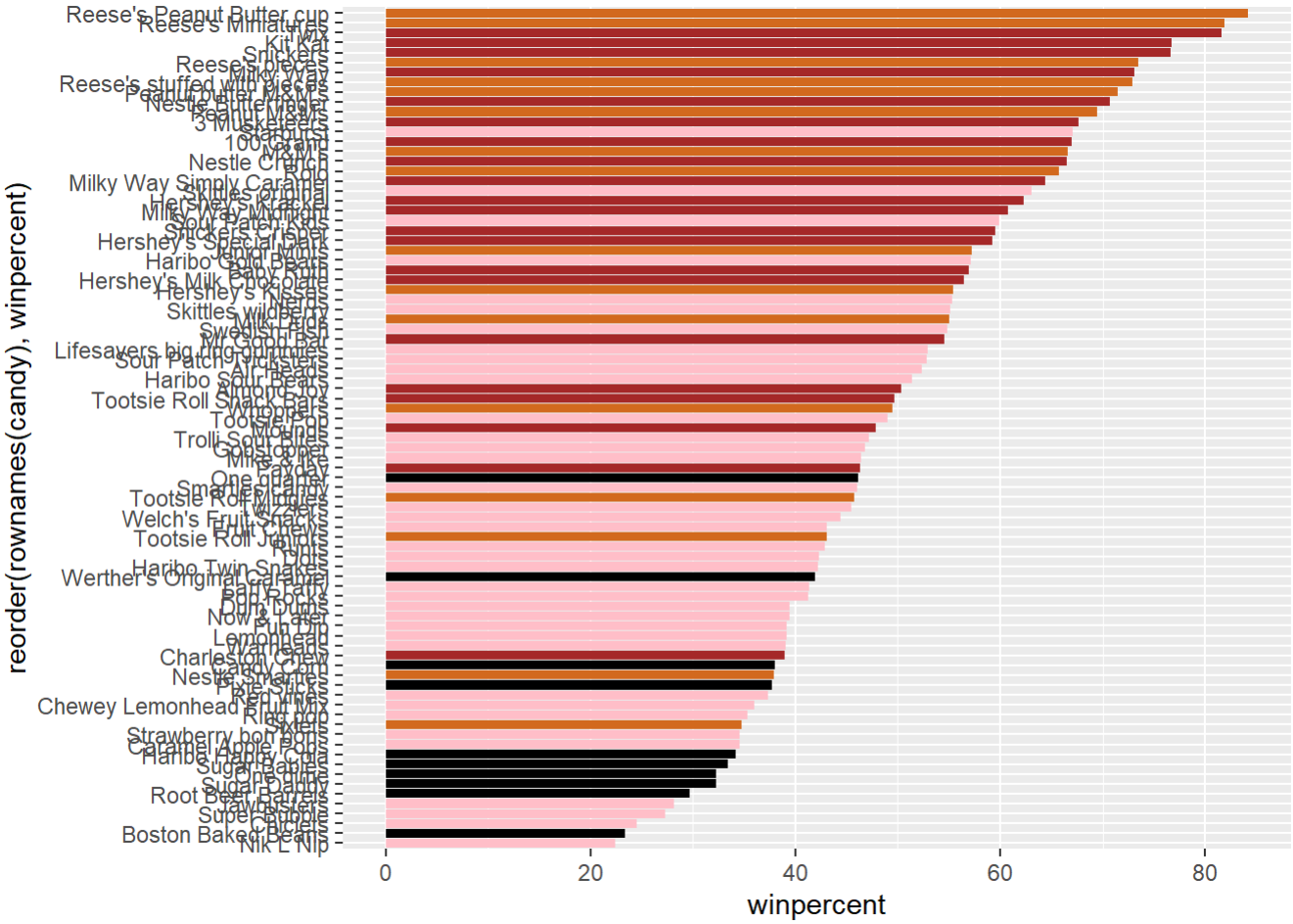


Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

```
library(ggplot2)

ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_bar(stat="identity")
```




```
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)
```

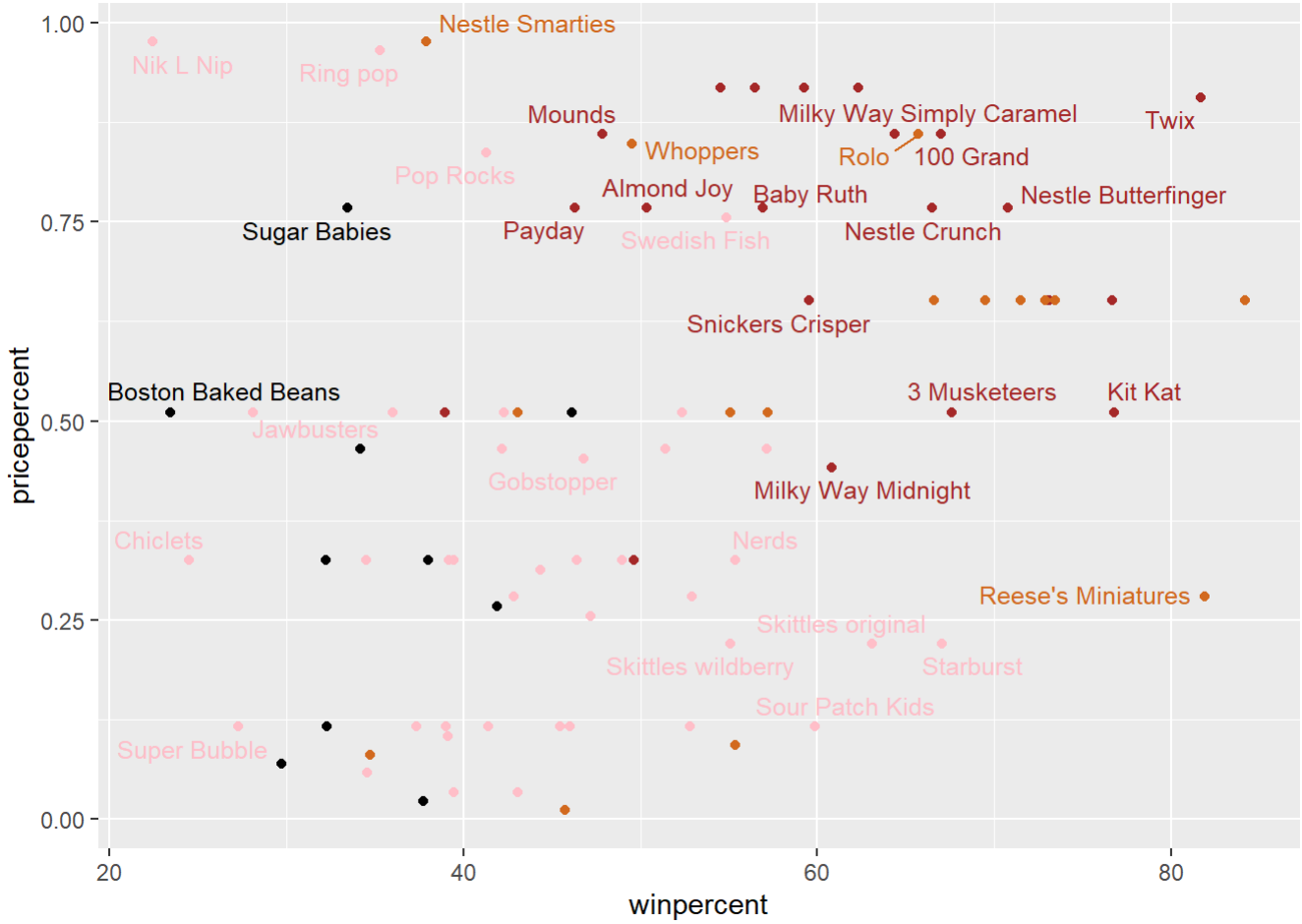


>Q17. What is the worst ranked chocolate candy? >ANS: Apparently people dont sixlets. >Q18. What is the best ranked fruity candy? >ANS: People like Starburst.

```
library("ggrepel")
```

```
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? ANS: Reeses Minatures

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

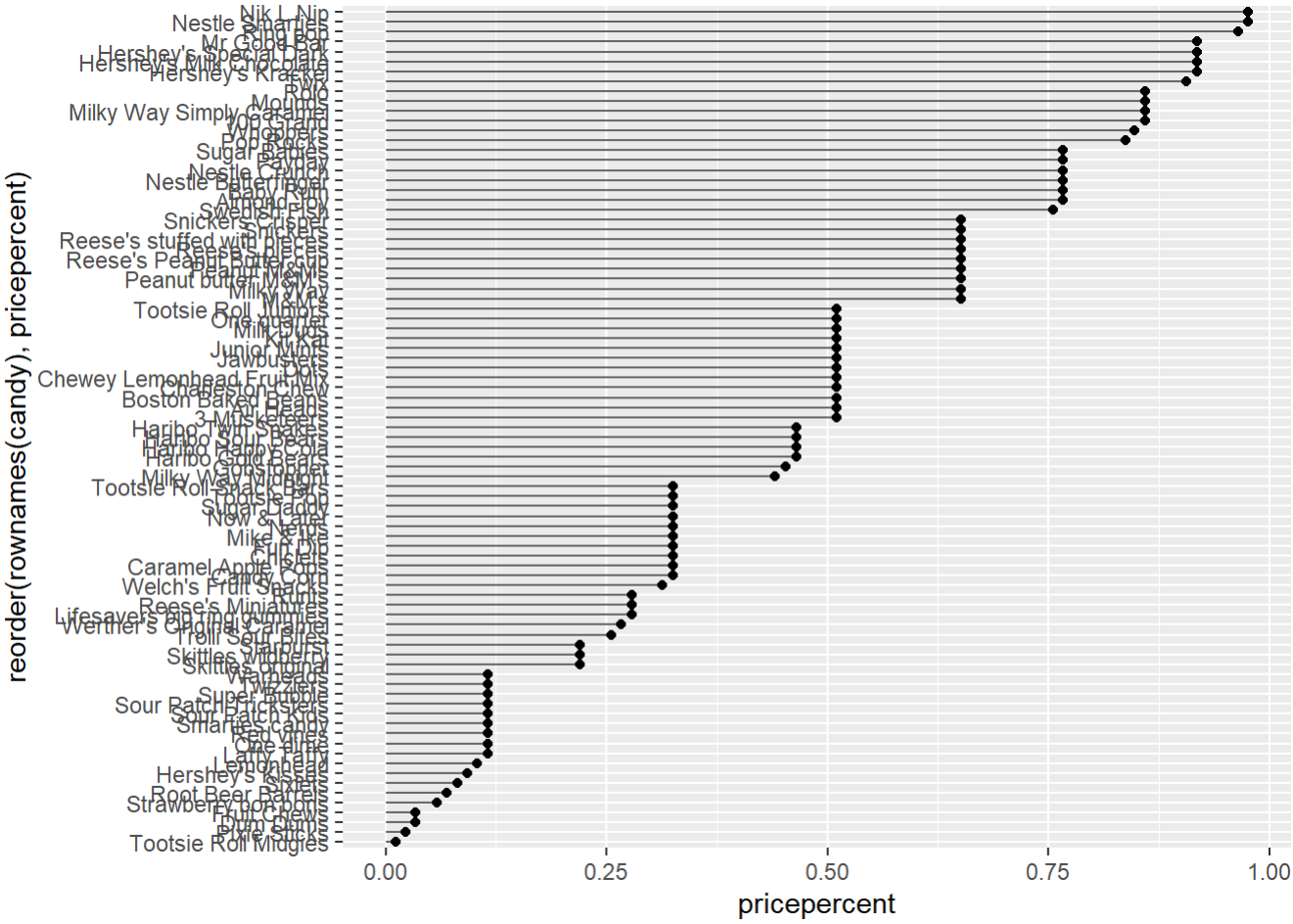
ANS: Nik L Lip manages to be expensive and unliked the most.

```
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
```

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().

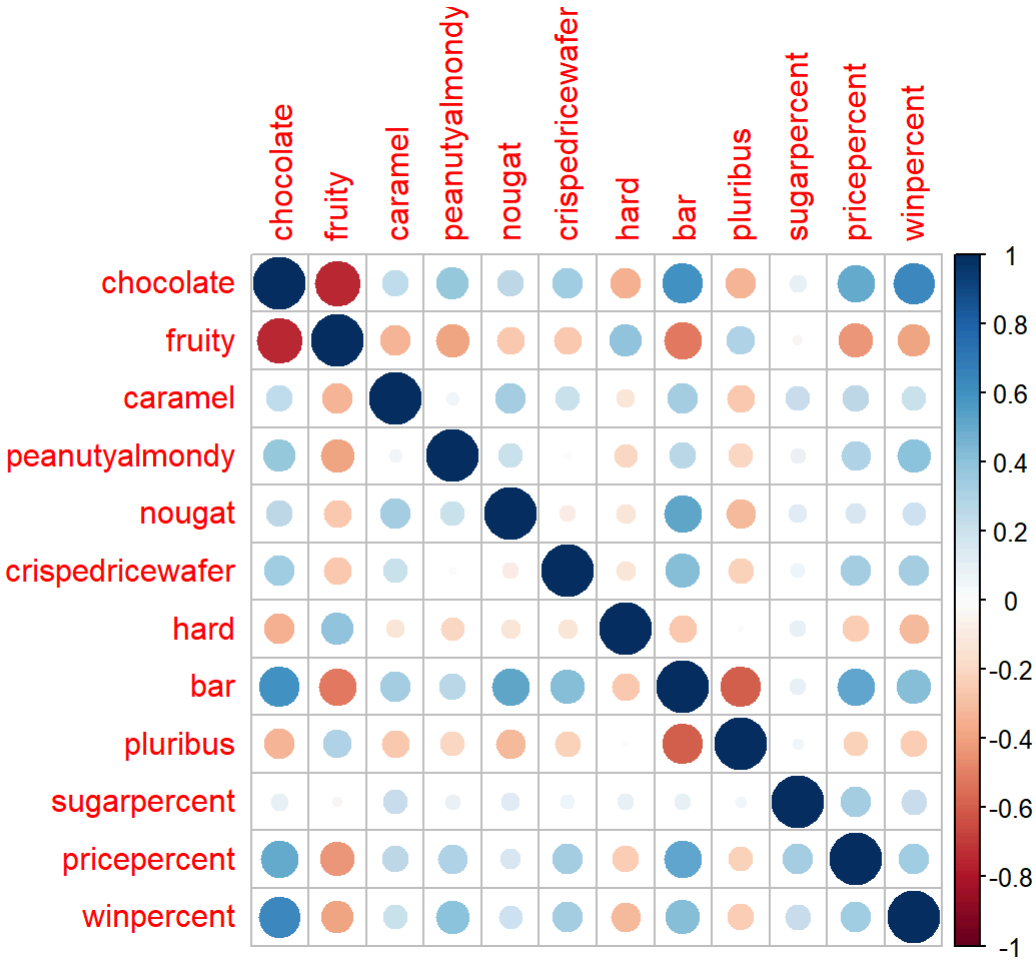
```
# Make a lollipop chart of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```




```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

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



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

ANS: Chocolate and fruity are anti-corellated, which is a shame because I like chocolate fruity candies.

Q23. Similarly, what two variables are most positively correlated? Ans: Chocolate and winpercent are correlated. Chocolate and bar are also correlated.

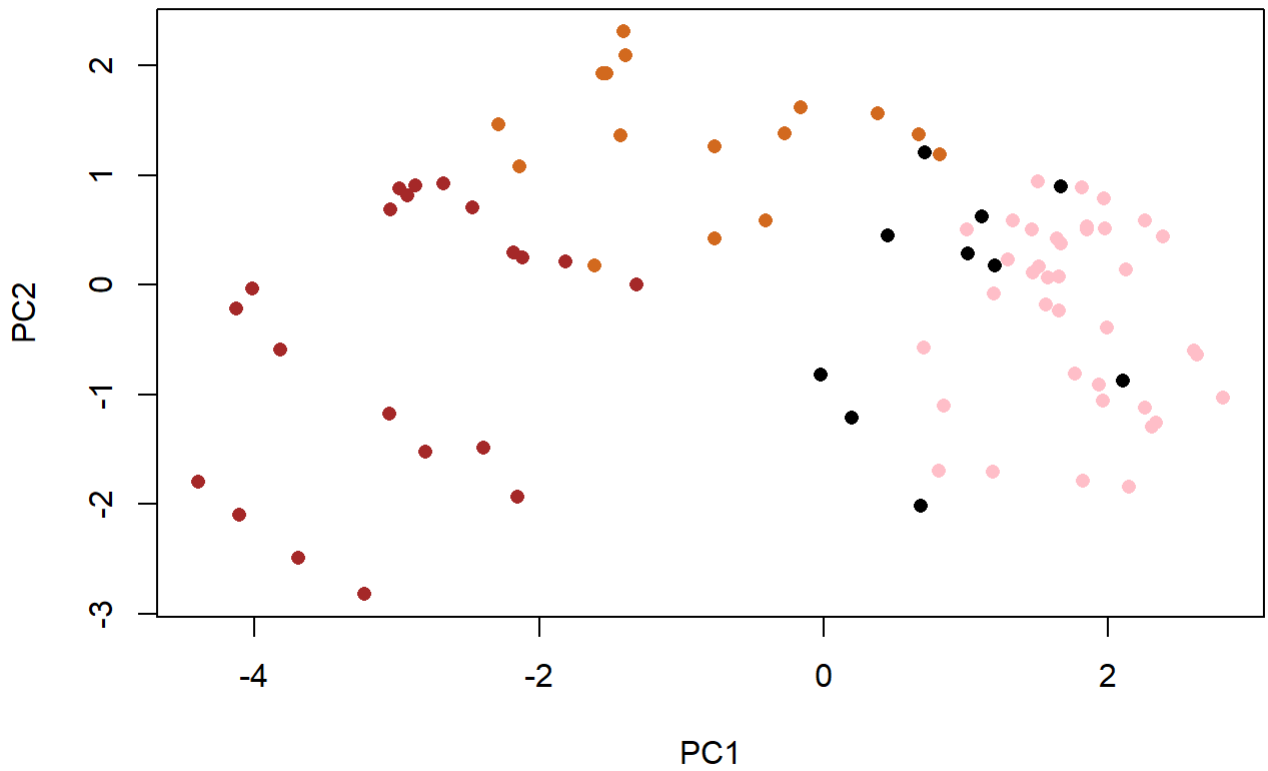
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
## 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
```

```
pca$rotation[,1]
```

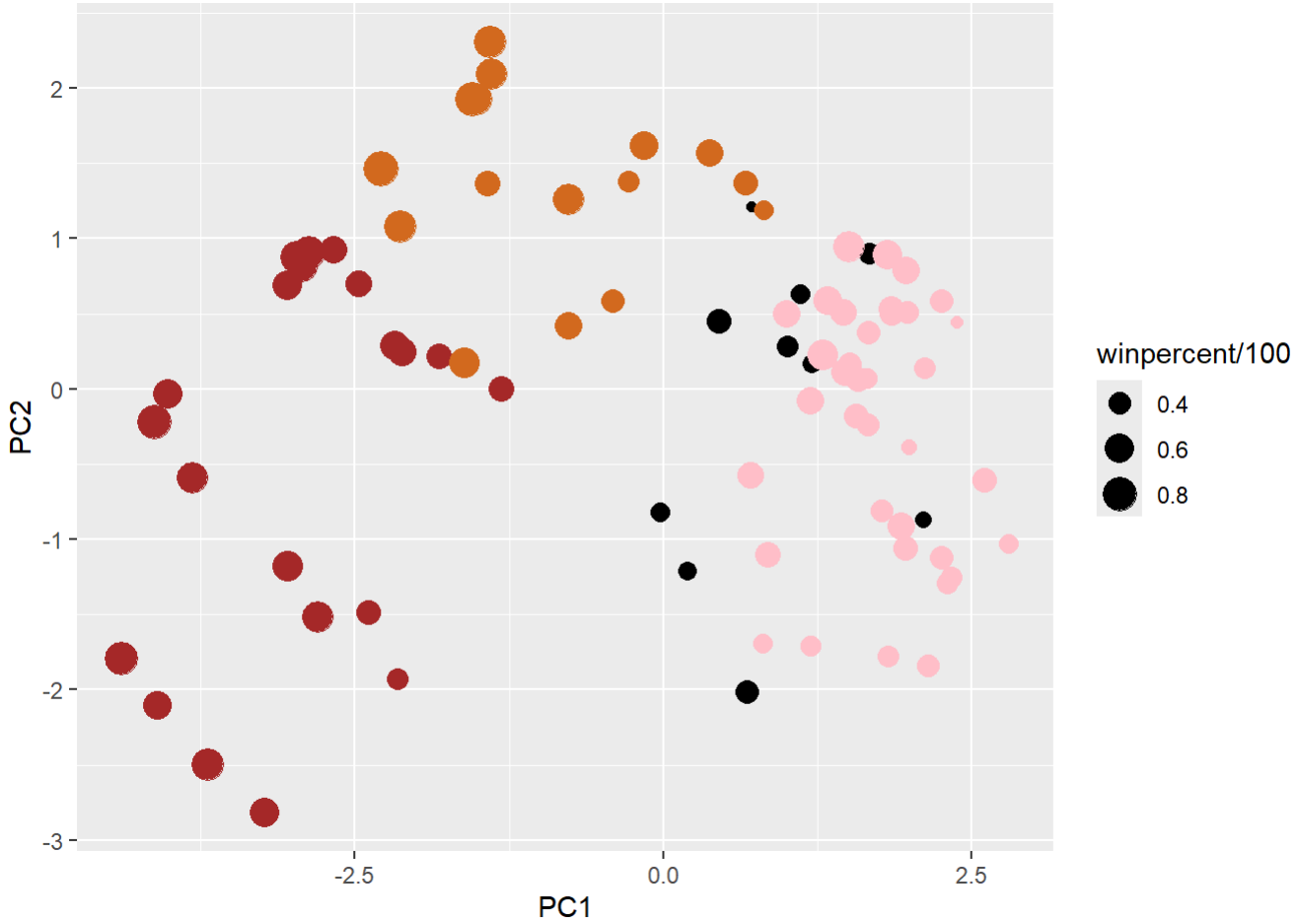
```
##        chocolate           fruity          caramel   peanutyalmondy
##       -0.4019466        0.3683883       -0.2299709       -0.2407155
##           nougat crispedricewafer             hard              bar
##       -0.2268102       -0.2215182        0.2111587       -0.3947433
##         pluribus     sugarpercent     pricepercent       winpercent
##        0.2600041       -0.1083088       -0.3207361       -0.3298035
```

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



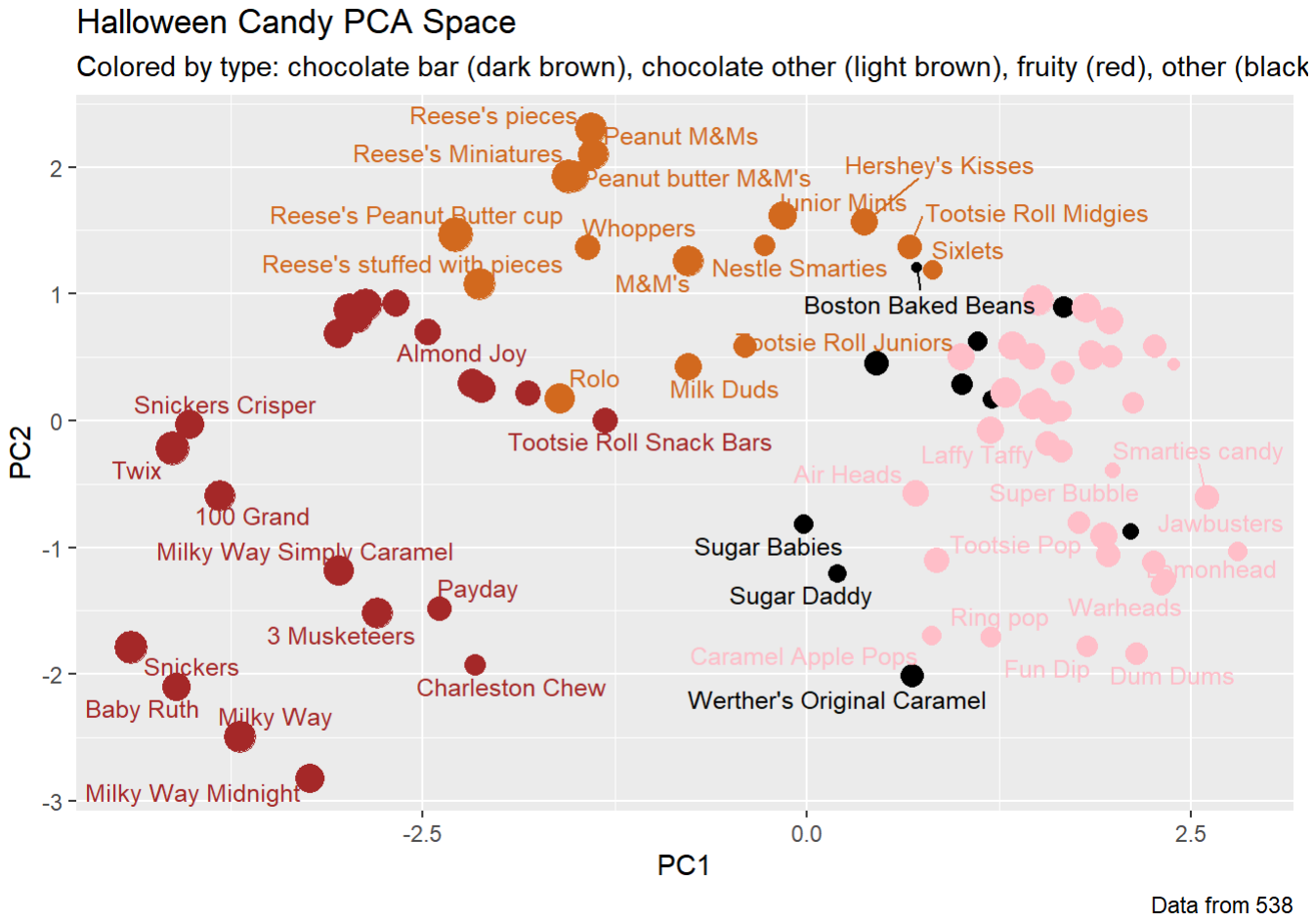

```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
```

```
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), fr
uity (red), other (black)",
       caption="Data from 538")
```

```
## Warning: ggrepel: 40 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
library(plotly)
```

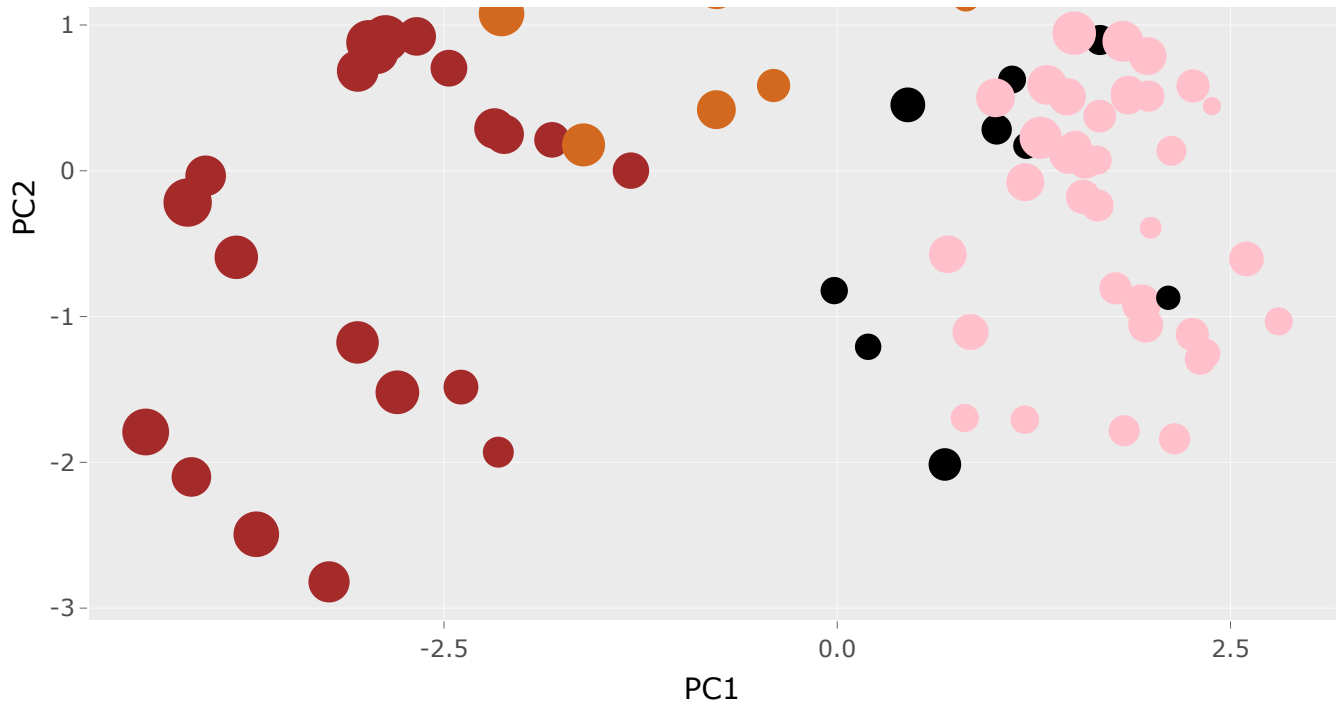
```
##
## 载入程序包：'plotly'
```

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

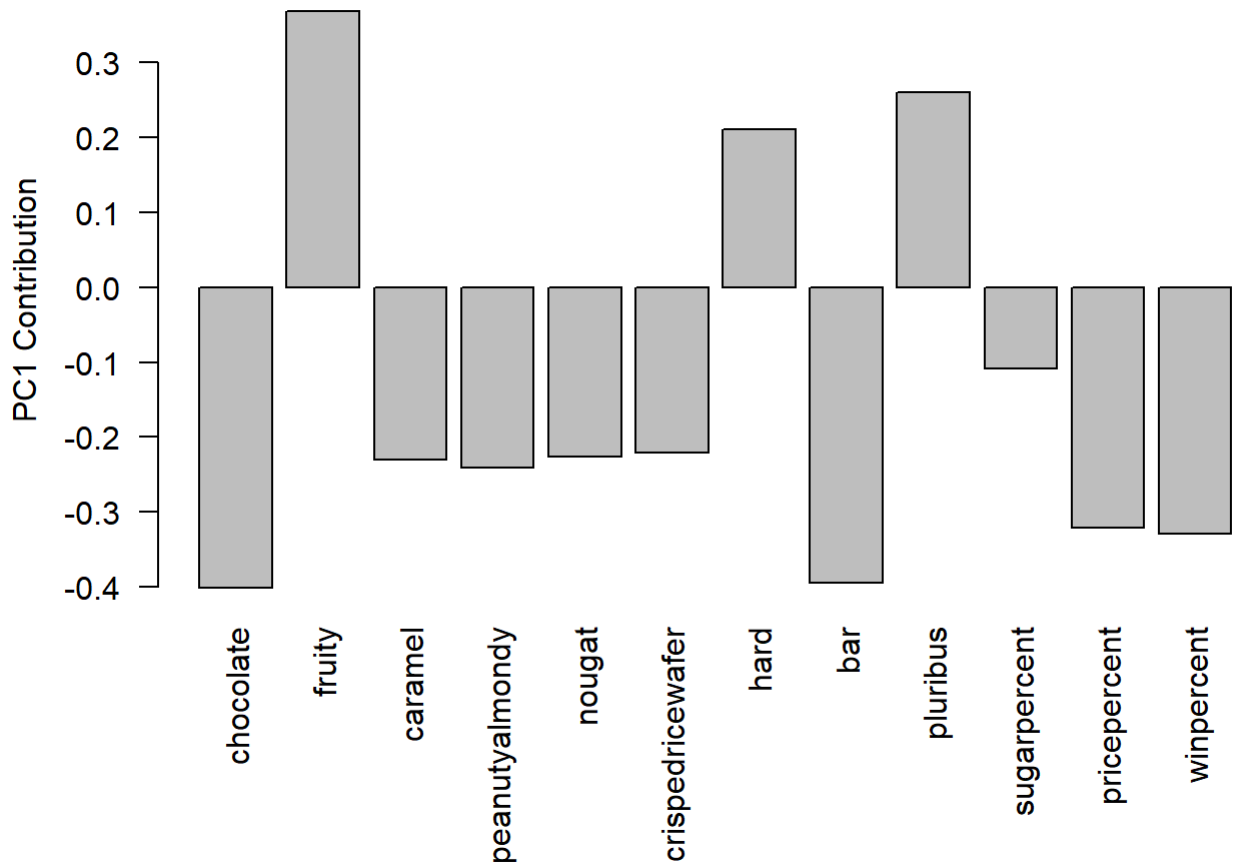
```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## 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? ANS:Fruity, hard and pluribus are heavily picked up by PC1. This makes sense. Afterall, many candies have the characteristics of being packed with multiple pieces in a bag, being fruity, and being hard overlap together.