Class09_candyproject

Eloise Simpson

Table of contents

Background

In today's mini-project we will analyze candy data with the exploratory graphics, basic statistics, correlation analysis and principal component analysis methods we have been learning thus far.

Data Import

The data comes as a CSV file from 538

```
head(read.csv("/Users/eloisesimpson/Documents/BIMM143/class09_candyproject/candy-data.csv"))
```

```
  competitorname chocolate fruity caramel peanutyalmondy nougat
1      100 Grand         1      0       1               0      0
2   3 Musketeers         1      0       0               0      1
3       One dime         0      0       0               0      0
4    One quarter         0      0       0               0      0
5      Air Heads         0      1       0               0      0
6     Almond Joy         1      0       0               1      0
```

```
  crispedricewafer hard bar pluribus sugarpercent pricepercent winpercent
1                1    0   1        0        0.732        0.860   66.97173
2                0    0   1        0        0.604        0.511   67.60294
3                0    0   0        0        0.011        0.116   32.26109
4                0    0   0        0        0.011        0.511   46.11650
5                0    0   0        0        0.906        0.511   52.34146
6                0    0   1        0        0.465        0.767   50.34755
```

```
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?

There are 85 rows in this dataset

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

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

```
[1] 38
```

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

```
[1] 81.64291
```

Q3. What is your favorite candy (other than Twix) in the dataset and what is it's winpercent value?

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

```
[1] 52.34146
```

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)
```

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	

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

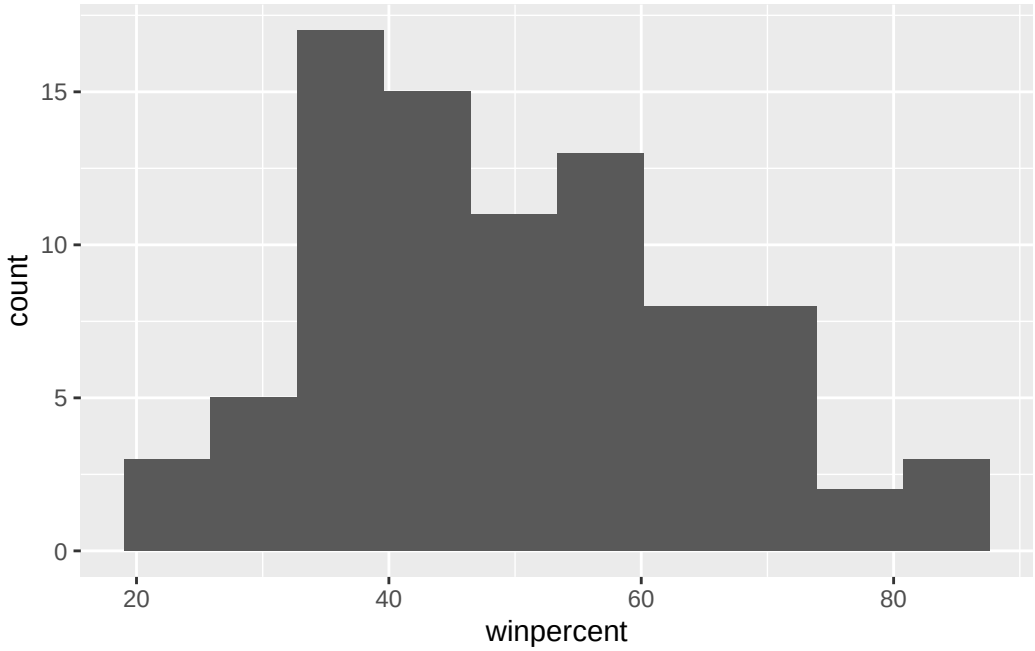
The winpercent column

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

A one means it is chocolate candy and a zero means it is not a chocolate candy

Q8. Plot a histogram of winpercent values using both base R an ggplot2.

```
library(ggplot2)
ggplot(candy, aes(x=winpercent)) +
  geom_histogram(bins=10)
```



Q9. Is the distribution of winpercent values symmetrical?

No

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

Below

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

1. Find all chocolate candy in dataset
2. Extract or find their winpercent values
3. calculate the mean of these values.
4. find all fruit candy
5. find their winpercent values
6. calculate their mean value

```
choc.candy <- candy[candy$chocolate == 1, ]
choc.win <- choc.candy$winpercent
mean(choc.win)
```

```
[1] 60.92153
```

```
fruity.candy <- candy[candy$fruity == 1,]
fruity.win <- fruity.candy$winpercent
mean(fruity.win)
```

```
[1] 44.11974
```

Chocolate is higher

Q12. Is this difference statistically significant?

```
t.test(choc.win, fruity.win)
```

```
    Welch Two Sample t-test

data:  choc.win and fruity.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
```

The p value is very small so yes it is.

Overall Candy Rankings

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

```
inds <- order(candy$winpercent)
head(candy[inds,],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

Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, Jawbusters

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

```
tail(candy[inds,],5)
```

	chocolate	fruity	caramel	peanutyalmondy	nougat
Snickers	1	0	1	1	1
Kit Kat	1	0	0	0	0
Twix	1	0	1	0	0
Reese's Miniatures	1	0	0	1	0
Reese's Peanut Butter cup	1	0	0	1	0

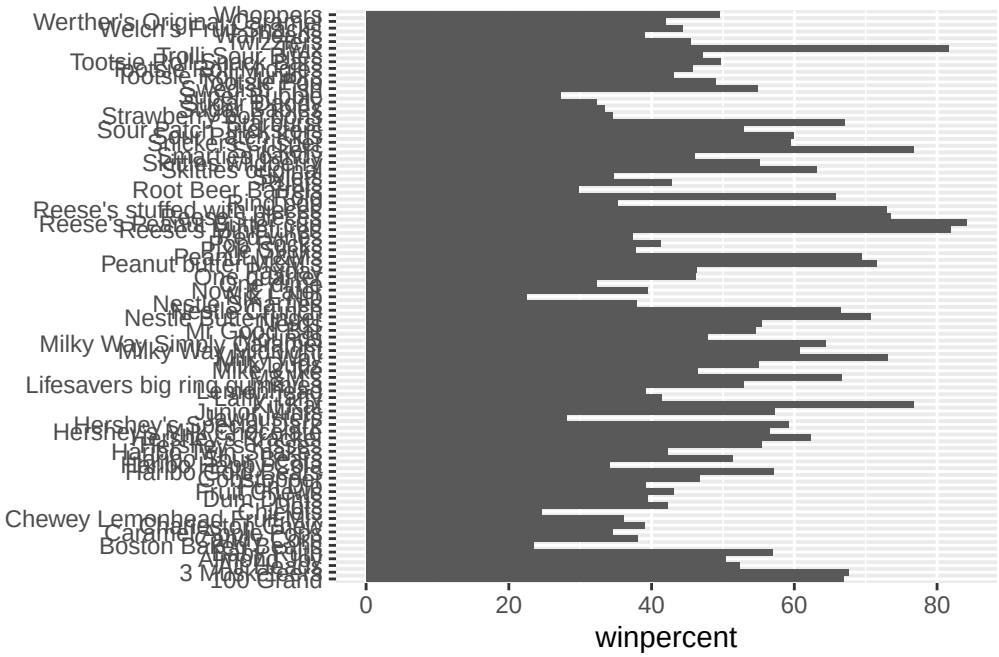
	crispedricewafer	hard	bar	pluribus	sugarpercent
Snickers	0	0	1	0	0.546
Kit Kat	1	0	1	0	0.313
Twix	1	0	1	0	0.546
Reese's Miniatures	0	0	0	0	0.034
Reese's Peanut Butter cup	0	0	0	0	0.720

	pricepercent	winpercent
Snickers	0.651	76.67378
Kit Kat	0.511	76.76860
Twix	0.906	81.64291
Reese's Miniatures	0.279	81.86626
Reese's Peanut Butter cup	0.651	84.18029

Reese's Peanut Butter cup, Reese's Miniatures, Twix, Kit Kat, Snickers.

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

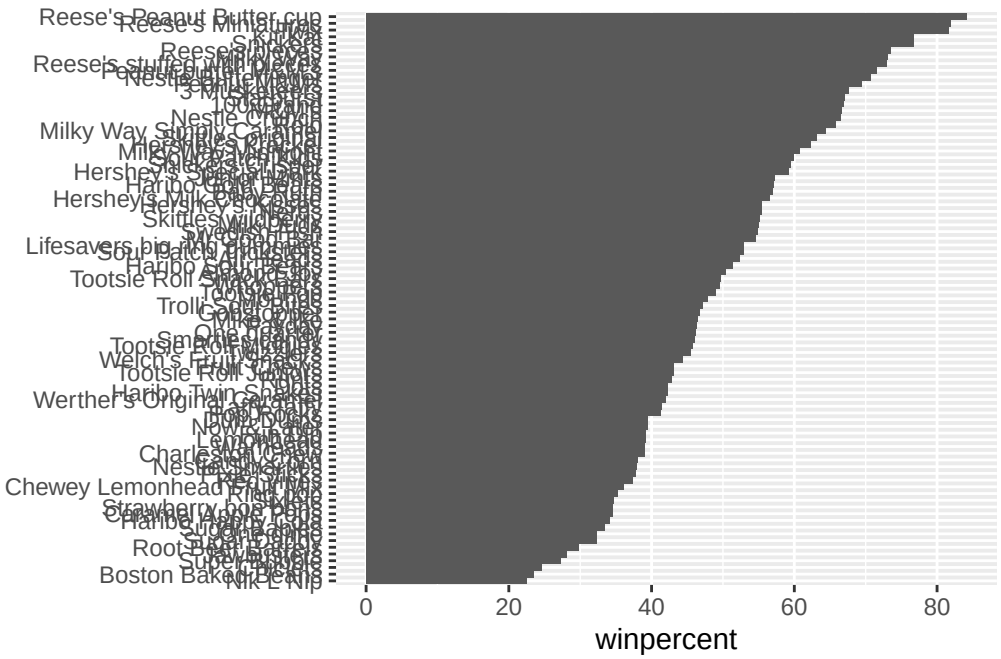
```
ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_col() +
  ylab("") ##to remove label
```




```
ggsave("barplot1.png", height = 10, width = 6)
```

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() +
  ylab("") ##to remove label
```




```
ggsave("barplot2.png", height = 10, width = 6)
```

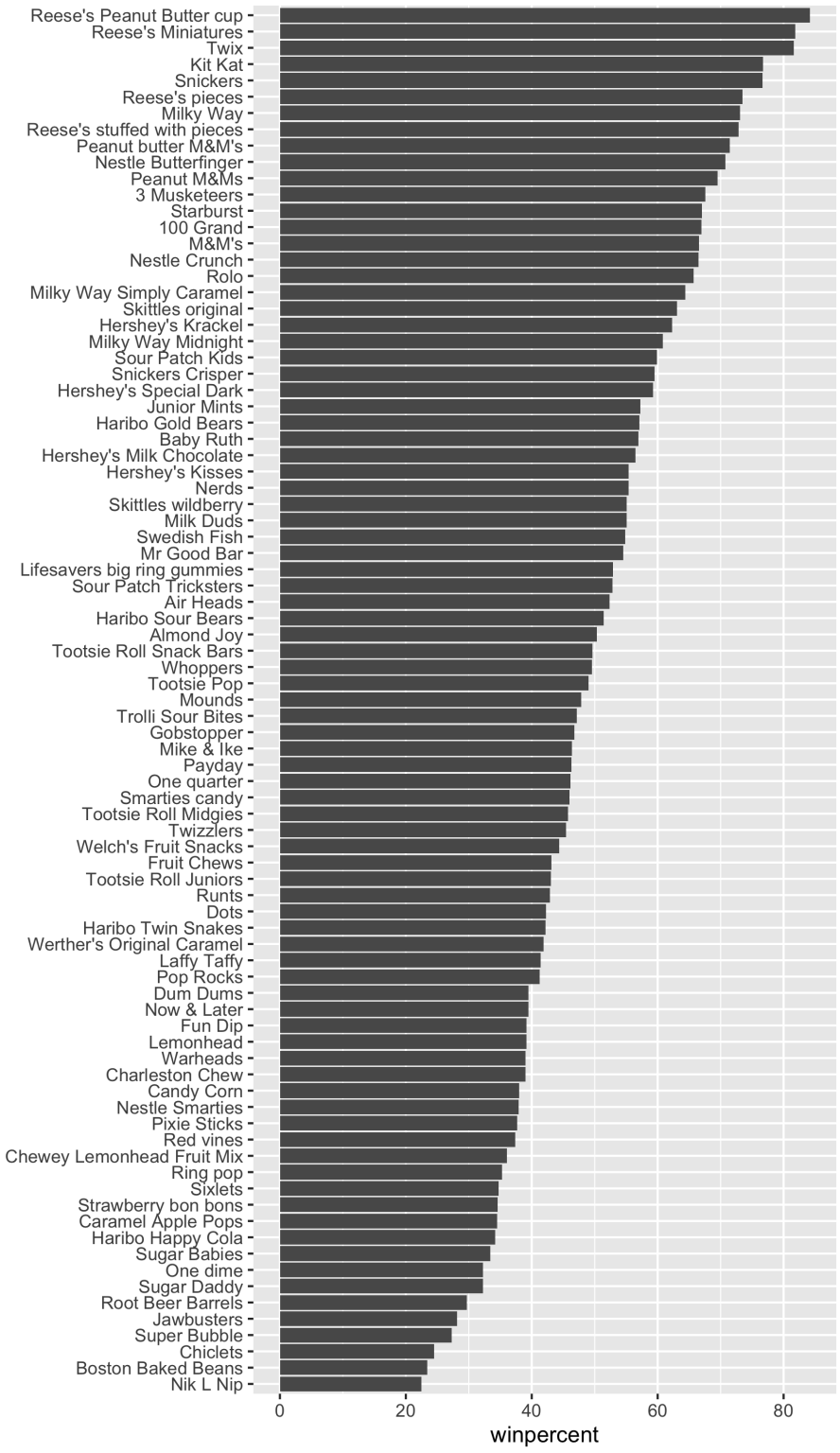
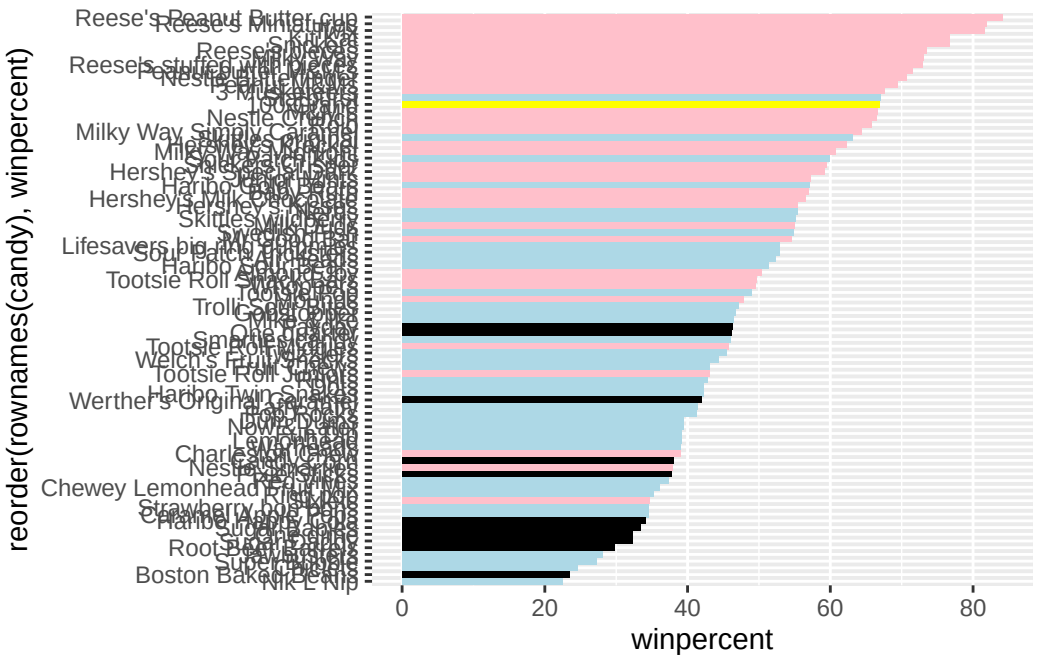


Figure 1: My second barplot for Q16

I want custom colors that I pick so we need to make this ourselves

```
my_cols <- rep("black", nrow(candy))
my_cols[candy$chocolate == 1] <- "pink"
my_cols[candy$bar] <- "yellow"
my_cols[candy$fruity == 1] <- "lightblue"
```

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




```
ggsave("barplot3.png", height = 10, width = 6)
```

Q17. What is the worst ranked chocolate candy?

Nik L Nip

Q18. What is the best ranked fruity candy?

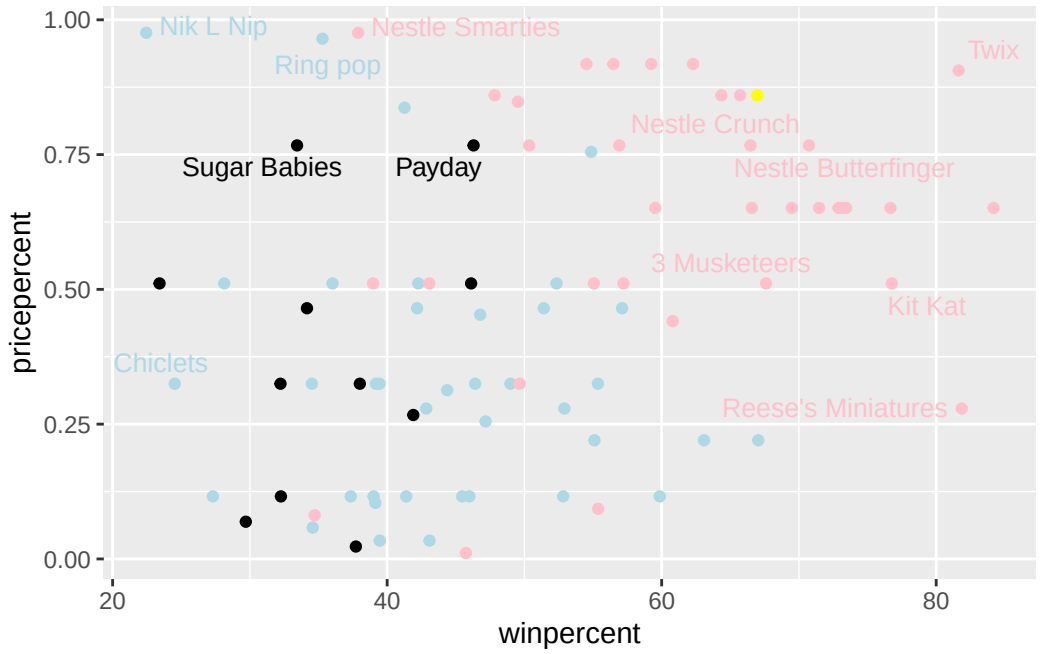
Reece's peanut butter cup

Taking a look at pricepercent

Make a plot of winpercent vs the pricepercent

```
library(ggrepel)
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.5, max.overlaps=5)
```

```
Warning: ggrepel: 73 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?

Reece's peanut butter cups

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

Nik l nip, chiclets, sugar babies

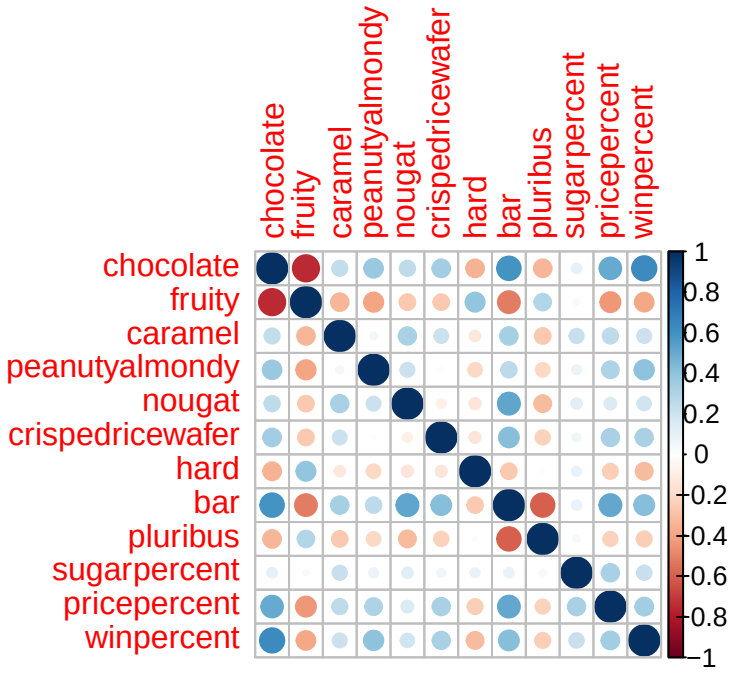
Exploring the correlation structure

Pearson correlation values range from -1 to +1

```
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)?

chocolate and winpercent

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

fruity and winpercent

Principal Component Analysis

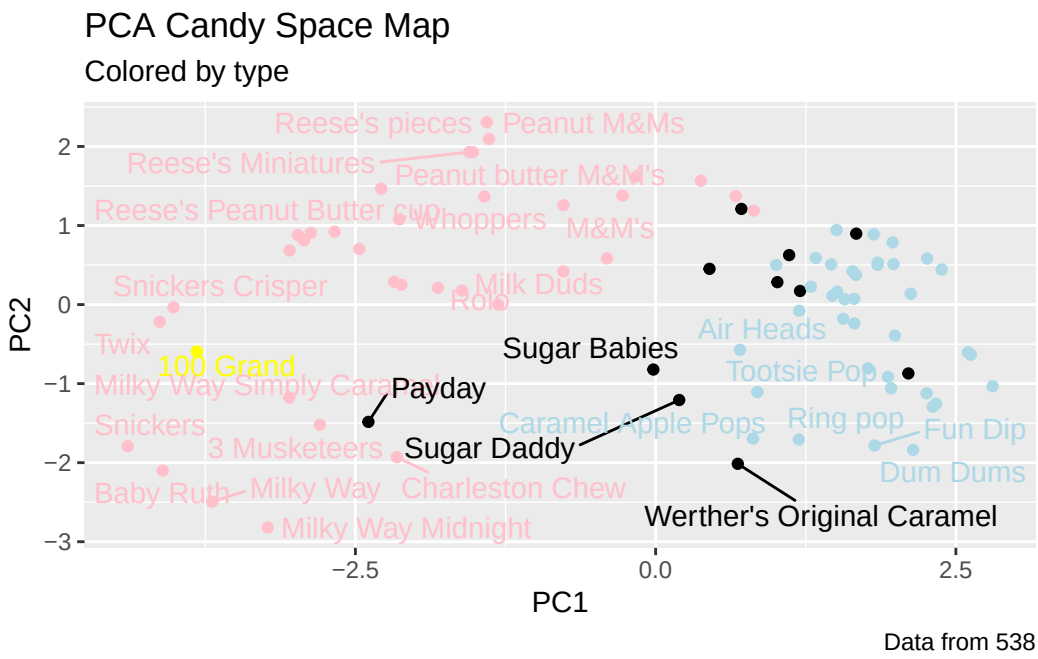
```
pca <- prcomp(candy, scale = T)
summary(pca)
```

```
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
```

The main results figure: the PCA score plot:

```
ggplot(pca$x) +
  aes(PC1, PC2, label=rownames(pca$x)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols) +
  labs(title="PCA Candy Space Map",
       subtitle = "Colored by type",
       caption = "Data from 538")
```

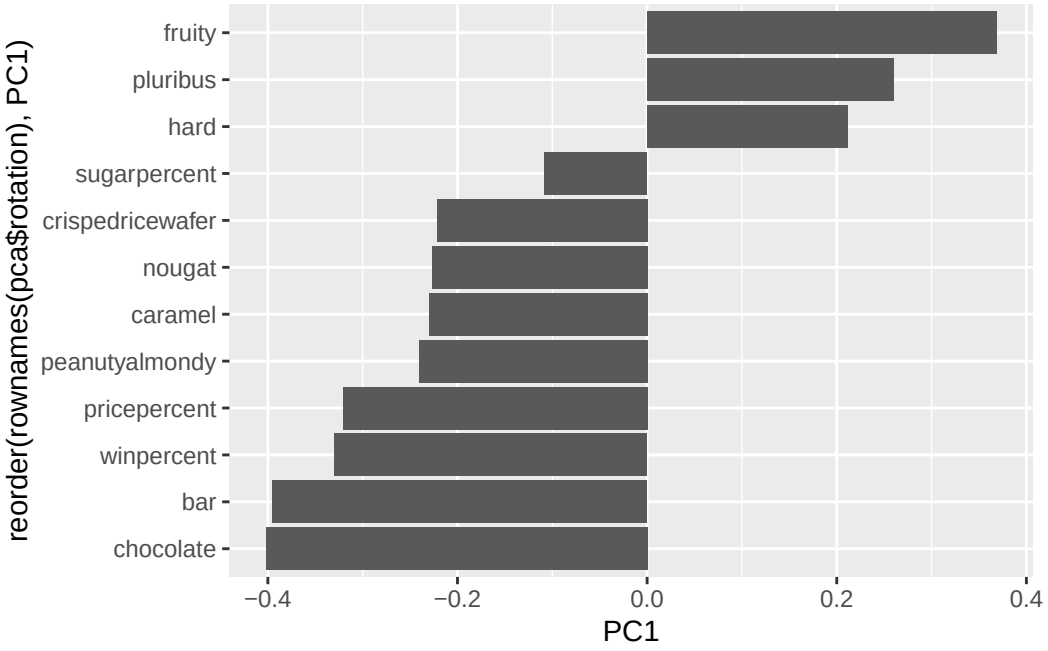
```
Warning: ggrepel: 56 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



The "loadings" plot for PC1

aes(winpercent, reorder (rownames(candy), winpercent))

```
ggplot(pca$rotation) +
  aes(PC1, reorder (rownames(pca$rotation), PC1)) +
  geom_col()
```



Q24. Complete the code to generate the loadings plot above. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you? Where did you see this relationship highlighted previously?

Winpercent, pricepercent, bar, chocolate, fruity. In the correlation plot.

Q25. Based on your exploratory analysis, correlation findings, and PCA results, what combination of characteristics appears to make a "winning" candy? How do these different analyses (visualization, correlation, PCA) support or complement each other in reaching this conclusion?

Having chocolate, being a bar, a good pricepercent. The visualisation helps to see whether the data supports a correlation between characteristics and popularity. The PCA displays which combinations are most powerful in driving people's preferences.