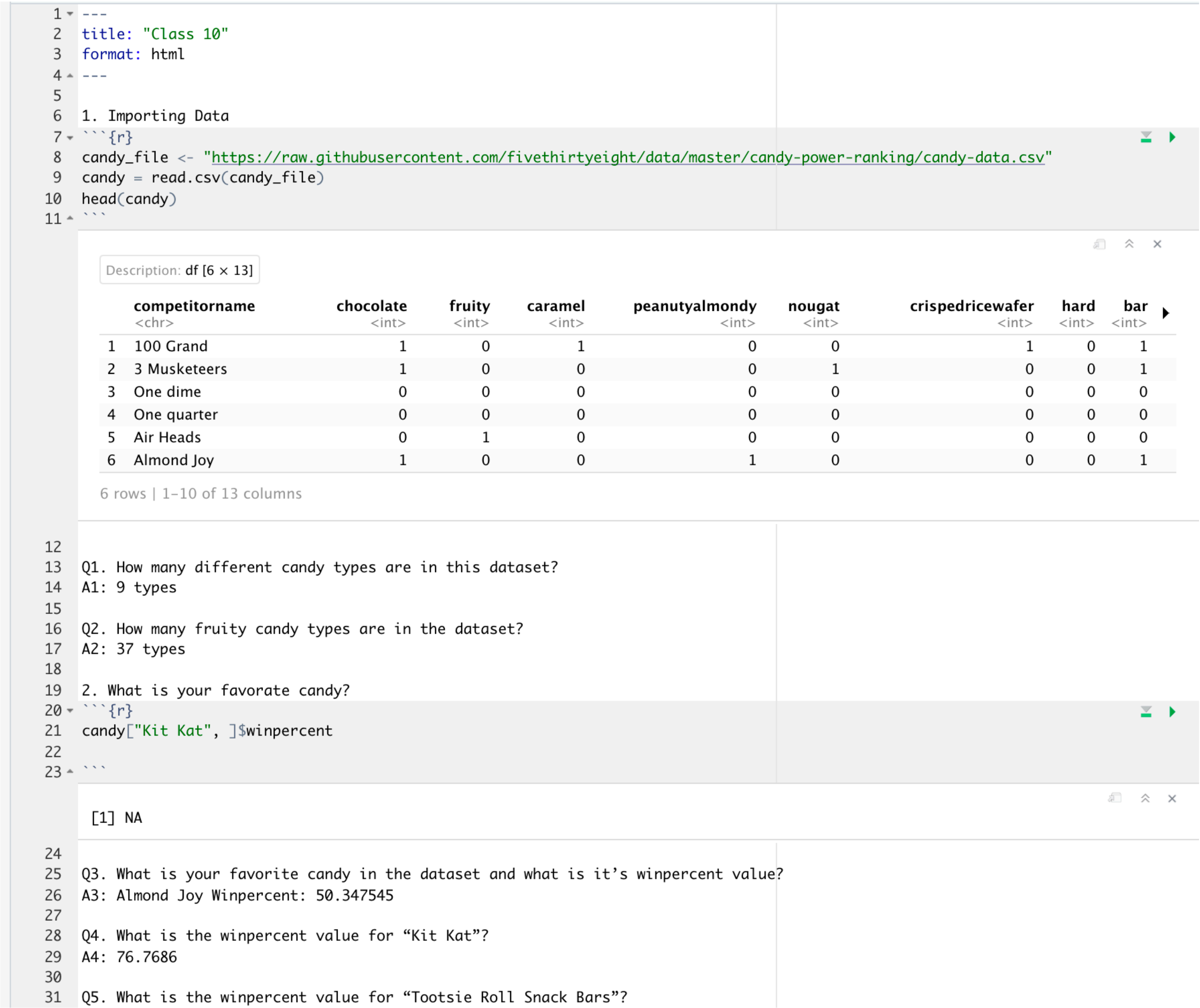


~~~
---
title: "Class 10"
format: html
---

1. Importing Data
```{r}
candy_file <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-ranking/candy-data.csv"
candy = read.csv(candy_file)
head(candy)
```
~~~

Description: df [6 × 13]

	competitorname <chr>	chocolate <int>	fruity <int>	caramel <int>	peanutyalmondy <int>	nougat <int>	crispedricewafer <int>	hard <int>	bar <int>
1	100 Grand	1	0	1	0	0	1	0	1
2	3 Musketeers	1	0	0	0	1	0	0	1
3	One dime	0	0	0	0	0	0	0	0
4	One quarter	0	0	0	0	0	0	0	0
5	Air Heads	0	1	0	0	0	0	0	0
6	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?
A1: 9 types

Q2. How many fruity candy types are in the dataset?
A2: 37 types

2. What is your favorate candy?
```{r}
candy["Kit Kat", ]$winpercent

```
~~~

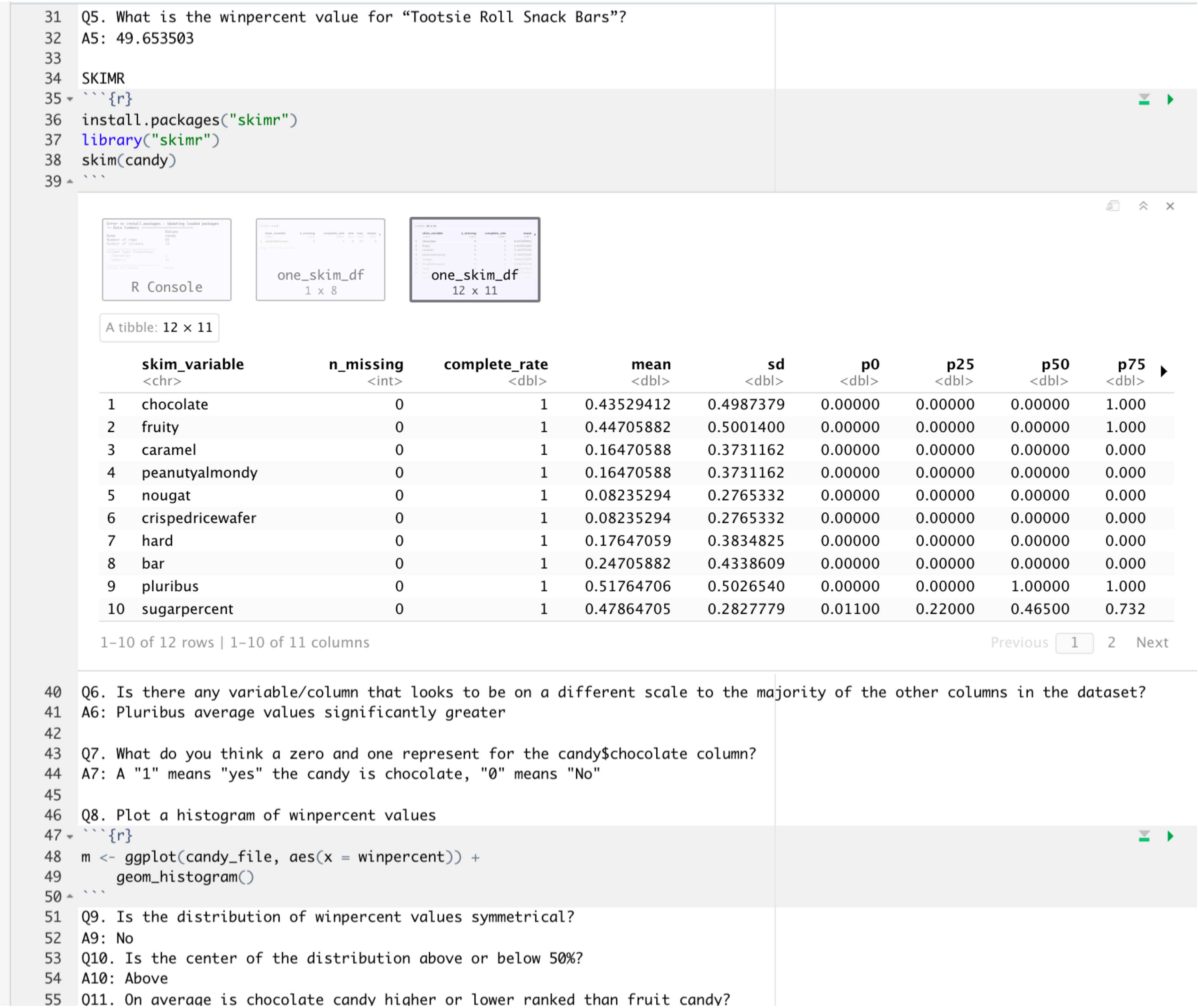
```
[1] NA
```

````

Q3. What is your favorite candy in the dataset and what is it’s winpercent value?
A3: Almond Joy Winpercent: 50.347545

Q4. What is the winpercent value for “Kit Kat”?
A4: 76.7686

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



Q5. What is the winpercent value for “Tootsie Roll Snack Bars”?
A5: 49.653503

SKIMR

~~~
```{r}
install.packages("skimr")
library("skimr")
skim(candy)
```
~~~

A tibble: 12 × 11

	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	mean <dbl>	sd <dbl>	p0 <dbl>	p25 <dbl>	p50 <dbl>	p75 <dbl>
1	chocolate	0	1	0.43529412	0.4987379	0.00000	0.00000	0.00000	1.000
2	fruity	0	1	0.44705882	0.5001400	0.00000	0.00000	0.00000	1.000
3	caramel	0	1	0.16470588	0.3731162	0.00000	0.00000	0.00000	0.000
4	peanutyalmondy	0	1	0.16470588	0.3731162	0.00000	0.00000	0.00000	0.000
5	nougat	0	1	0.08235294	0.2765332	0.00000	0.00000	0.00000	0.000
6	crispedricewafer	0	1	0.08235294	0.2765332	0.00000	0.00000	0.00000	0.000
7	hard	0	1	0.17647059	0.3834825	0.00000	0.00000	0.00000	0.000
8	bar	0	1	0.24705882	0.4338609	0.00000	0.00000	0.00000	0.000
9	pluribus	0	1	0.51764706	0.5026540	0.00000	0.00000	1.00000	1.000
10	sugarpercent	0	1	0.47864705	0.2827779	0.01100	0.22000	0.46500	0.732

1–10 of 12 rows | 1–10 of 11 columns

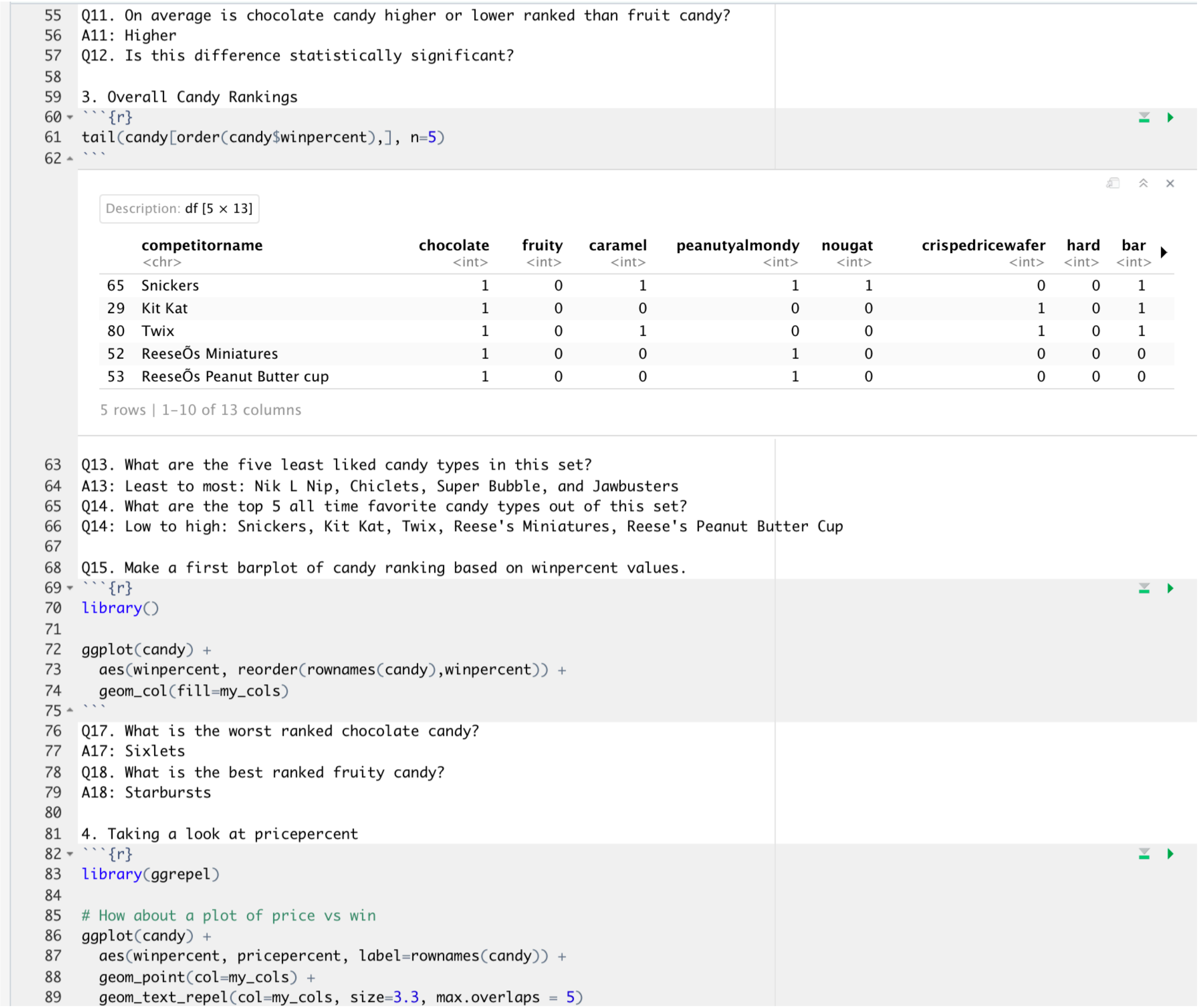
Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?
A6: Pluribus average values significantly greater

Q7. What do you think a zero and one represent for the candy$chocolate column?
A7: A "1" means "yes" the candy is chocolate, "0" means "No"

Q8. Plot a histogram of winpercent values

~~~
```{r}
m <- ggplot(candy_file, aes(x = winpercent)) +
    geom_histogram()
```
~~~

Q9. Is the distribution of winpercent values symmetrical?
A9: No
Q10. Is the center of the distribution above or below 50%?
A10: Above
Q11. On average is chocolate candy higher or lower ranked than fruit candy?



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

A11: Higher

Q12. Is this difference statistically significant?

3. Overall Candy Rankings

~~~
```{r}
tail(candy[order(candy$winpercent),], n=5)
```
~~~

Description: df [5 × 13]

	competitorname <chr>	chocolate <int>	fruity <int>	caramel <int>	peanutyalmondy <int>	nougat <int>	crispedricewafer <int>	hard <int>	bar <int>
65	Snickers	1	0	1	1	1	0	0	1
29	Kit Kat	1	0	0	0	0	1	0	1
80	Twix	1	0	1	0	0	1	0	1
52	ReeseÕs Miniatures	1	0	0	1	0	0	0	0
53	ReeseÕs Peanut Butter cup	1	0	0	1	0	0	0	0

5 rows | 1-10 of 13 columns

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

A13: Least to most: Nik L Nip, Chiclets, Super Bubble, and Jawbusters

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

Q14: Low to high: Snickers, Kit Kat, Twix, Reese's Miniatures, Reese's Peanut Butter Cup

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

~~~
```{r}
library()

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

Q17. What is the worst ranked chocolate candy?

A17: Sixlets

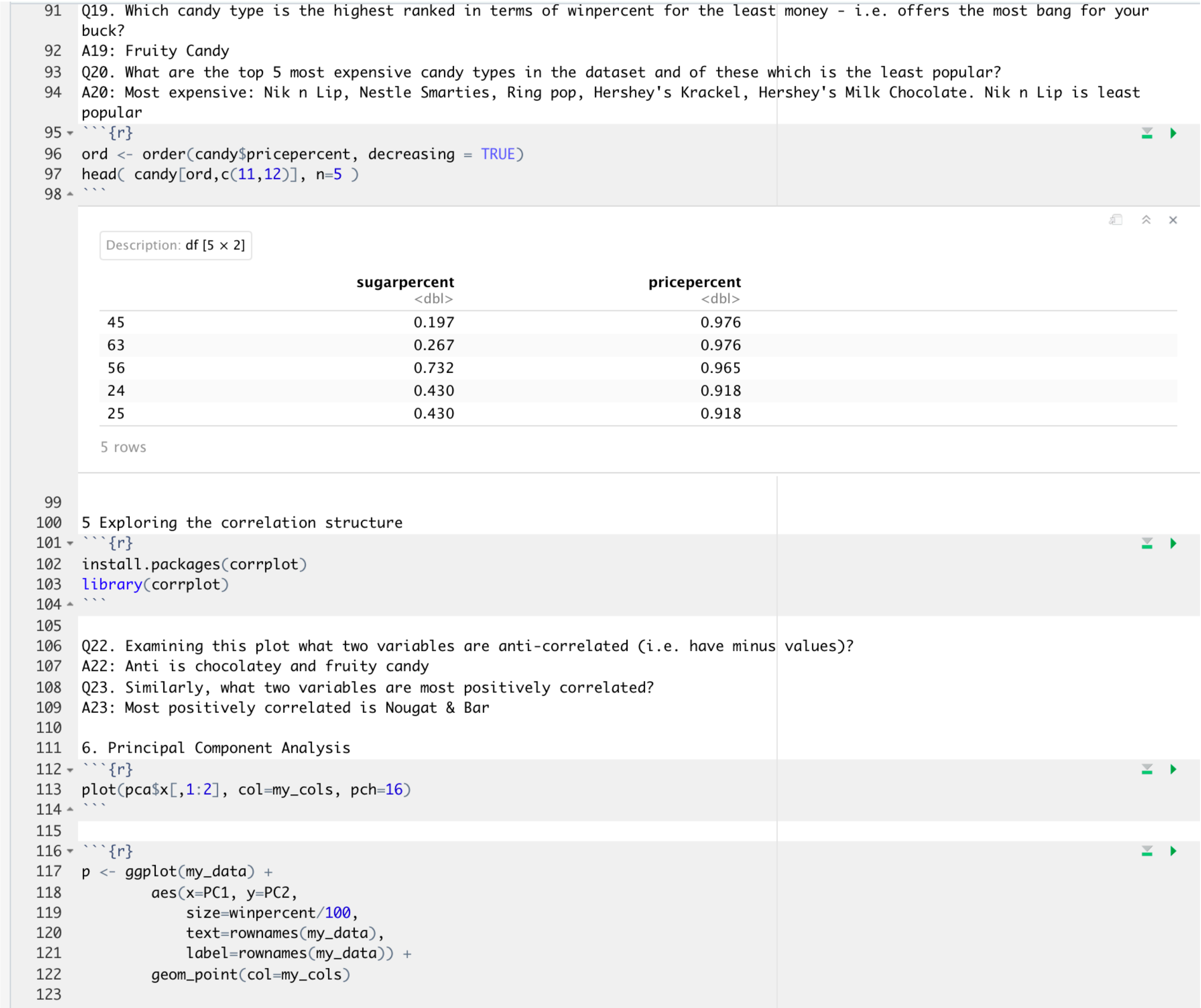
Q18. What is the best ranked fruity candy?

A18: Starbursts

4. Taking a look at pricepercent

~~~
```{r}
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)
~~~



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?

A19: Fruity Candy

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

A20: Most expensive: Nik n Lip, Nestle Smarties, Ring pop, Hershey's Krackel, Hershey's Milk Chocolate. Nik n Lip is least popular

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

Description: df [5 × 2]

	sugarpercent <dbl>	pricepercent <dbl>
45	0.197	0.976
63	0.267	0.976
56	0.732	0.965
24	0.430	0.918
25	0.430	0.918

5 rows

5 Exploring the correlation structure

~~~
```{r}
install.packages(corrplot)
library(corrplot)
```
~~~

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

A22: Anti is chocolatey and fruity candy

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

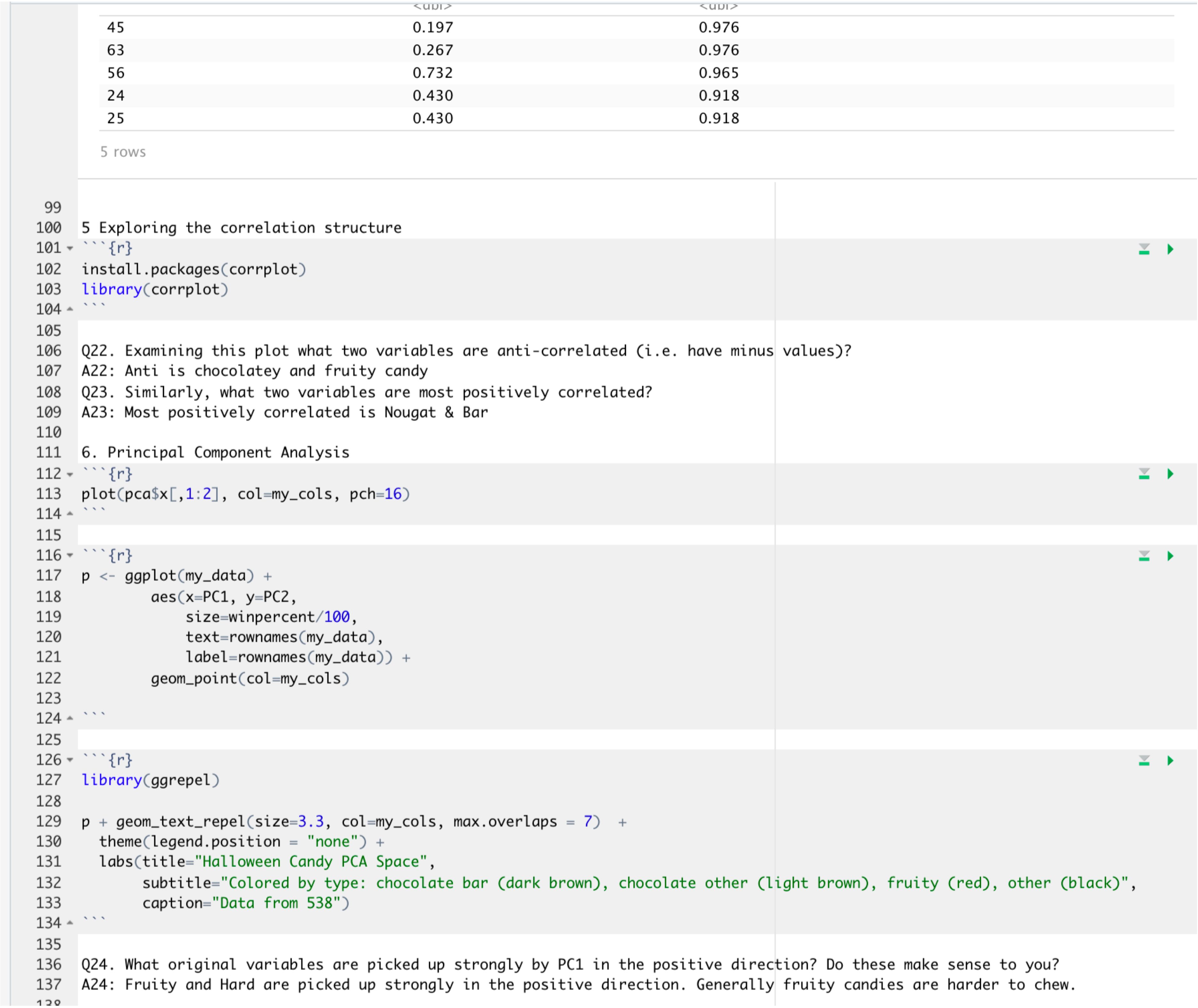
A23: Most positively correlated is Nougat & Bar

6. Principal Component Analysis

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

~~~
```{r}
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)

~~~



	<dbl>	<dbl>
45	0.197	0.976
63	0.267	0.976
56	0.732	0.965
24	0.430	0.918
25	0.430	0.918

5 rows

5 Exploring the correlation structure

~~~
```{r}
install.packages(corrplot)
library(corrplot)
```
~~~

Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)?
A22: Anti is chocolatey and fruity candy
Q23. Similarly, what two variables are most positively correlated?
A23: Most positively correlated is Nougat & Bar

6. Principal Component Analysis

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

~~~
```{r}
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)

```
~~~

~~~
```{r}
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 (red), other (black)",
       caption="Data from 538")
```
~~~

Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?
A24: Fruity and Hard are picked up strongly in the positive direction. Generally fruity candies are harder to chew.