class9

```
candy_file <- "~/Downloads/candy-data.csv"  # Update this path if needed
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?

```
colnames(candy)
```

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

```
num_candy_types <- length(unique(candy$CandyType))
num_candy_types
```

```
[1] 0
```

Q2. How many fruity candy types are in the dataset? 1

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

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

```
[1] 81.64291
```

Q4. What is the winpercent value for "Kit Kat"?

Q5. What is the winpercent value for "Tootsie Roll Snack Bars"?

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

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

Q8. Plot a histogram of winpercent values

Q9. Is the distribution of winpercent values symmetrical?

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

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

Q12. Is this difference statistically significant? Yes Q13. What are the five least liked candy types in this set?

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

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

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

Q17. What is the worst ranked chocolate candy?

- Q18. What is the best ranked fruity candy?

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?

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

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

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

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