STA130 – Winter 2024

Week 7 Problem Set

N. Moon and J. Speagle

Instructions

How do I hand in my solutions and how do I check my work

You will submit your solutions (.Rmd and .pdf) on MarkUs at the following link: https://markus4.teach.cs.toronto.edu/2024-01/courses/1 Submissions are due at 5pm on Thursdays; see Quercus page for the specific deadline for each problem set.

Usually when you do an assignment, you don't find out whether your answers are correct until *after* the deadline, when you get your grade back. However, using MarkUs, you can submit your work before the deadline and run tests to check your solutions!

Note: Some parts of some questions may not be covered by tests in MarkUs, but you're still responsible for reviewing the posted solutions and make sure you understand them. Some of these parts will also be graded in some weeks.

Some of the optional questions have tests in MarkUs, but you won't be penalized for not completing these (or failing the tests for these parts) when we grade your work after the submission deadline. The tests for these parts are provided for your guidance.

What to do if a test fails on MarkUs

- Take a deep breath! Your work won't really be graded until the deadline, so start early to make sure you have lots of time to resolve issues before the deadline.
- Read the message to get hints about what the problem is. For example "variable X not present" means that you may have a typo in the name of the variable we're looking for - re-read the question carefully and make sure you're following the instructions.
- Search on Piazza to see if other classmates have encountered a similar error (and if not, consider posting a screenshot of the error message).
- Come to TA or instructor office hours with your issue.

IMPORTANT - SETUP

```r
# Instructions for how students should define tests
STUDENT_NUMBER <- 1010057028;  # replace 0 by your real student number
```

Question 1

In this question, you will work with the social media anxiety data from the Week 5 problem set (optional question).

There have been many questions regarding whether or not usage of social media increases anxiety levels. For example, does regular viewing of Instagram or WeChat posts create an unattainable sense of life success and satisfaction? Does procrastinating by watching YouTube or TikTok videos or reading Reddit threads contribute unnecessary stress from deadline pressure?

To try and answer some of these questions, a study was conducted to examine the relationship between social media usage and student anxiety. Students were asked to categorize their social media usage as "High" if it exceeded more than 2 hours per day, and then student anxiety levels were scored through using series of questions. Higher scores on the questionnaire suggest higher student anxiety.

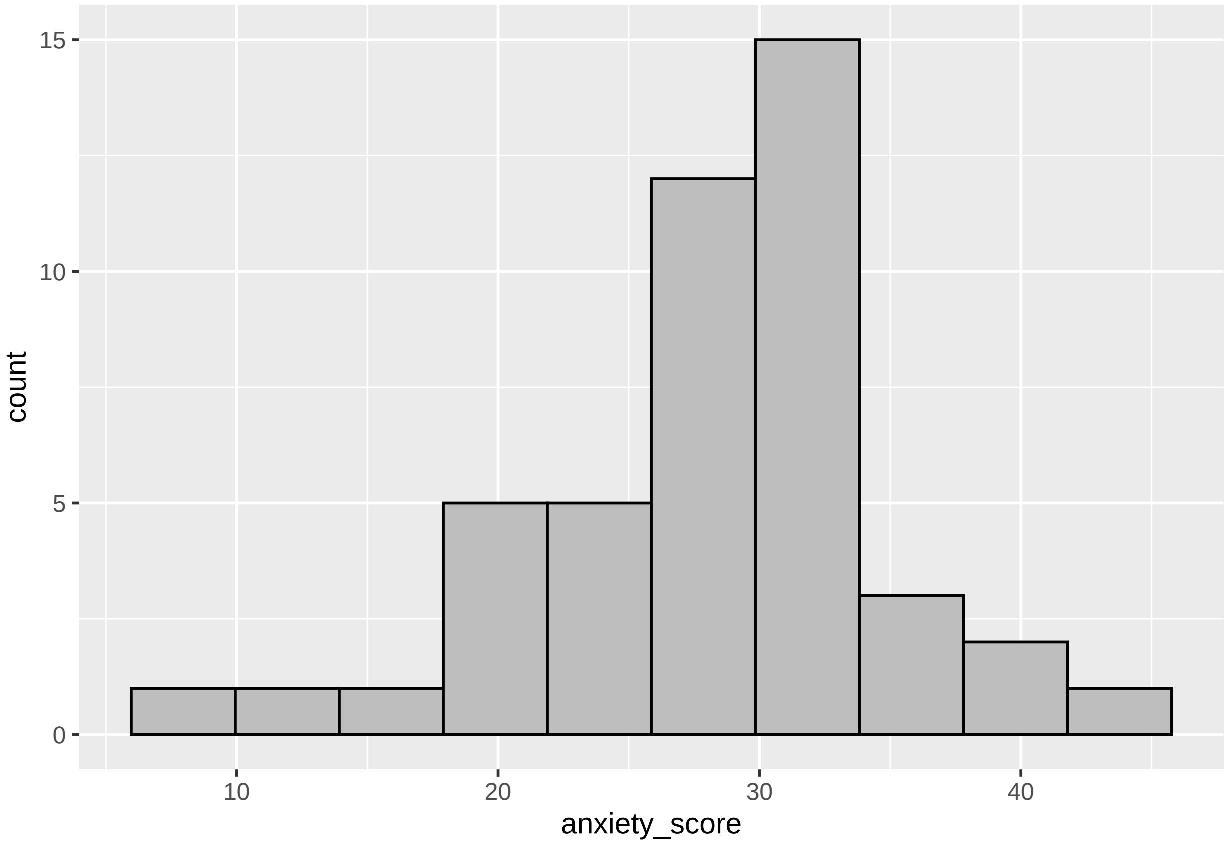
```
# define our initial dataset
social_media_usage <- c(rep("Low", 30), rep("High", 16));
anxiety_score <- c(24.64, 39.29, 16.32, 32.83, 28.02,
                   33.31, 20.60, 21.13, 26.69, 28.90,
                   26.43, 24.23, 7.10,  32.86, 21.06,
                   28.89, 28.71, 31.73, 30.02, 21.96,
                   25.49, 38.81, 27.85, 30.29, 30.72,
                   21.43, 22.24, 11.12, 30.86, 19.92,
                   33.57, 34.09, 27.63, 31.26,
                   35.91, 26.68, 29.49, 35.32,
                   26.24, 32.34, 31.34, 33.53,
                   27.62, 42.91, 30.20, 32.54)
anxiety_data <- tibble(social_media_usage, anxiety_score)

# preview our data
glimpse(anxiety_data)
```

```
## Rows: 46
## Columns: 2
## $ social_media_usage <chr> "Low", "Low", "Low", "Low", "Low", "Low", "Low", "L~
## $ anxiety_score      <dbl> 24.64, 39.29, 16.32, 32.83, 28.02, 33.31, 20.60, 21~
```

(a) Create a visualization to examine the distribution of anxiety scores. Describe the distribution, taking care to comment on the centre, shape and spread of the distribution, with particular attention to the number of clusters you believe there might be.

```
# Create your summary and visualization(s) here
anxiety_data %>% ggplot(aes(x=anxiety_score)) + geom_histogram(color = "black",
                  fill = "gray",
                  bins = 10)
```



```
data <- anxiety_data %>% select(anxiety_score) %>% summarize(min = min(anxiety_score), mean = mean(anxi

# Replace NULL by your answers below
Q1a_min <- data$min
Q1a_mean <- data$mean
Q1a_median <- data$median
Q1a_max <- data$max
```

(b) Use R to run the k-means algorithm with k=2 to estimate two clusters of people based on their anxiety scores. Save the results of kmeans in an R object called `clustering`.

```
set.seed(STUDENT_NUMBER) # DO NOT CHANGE THIS LINE

# run k-means on the data
clustering <- kmeans(anxiety_data$anxiety_score, 2)


Q1b_clustering_object_type <- class(clustering) # DO NOT CHANGE THIS LINE
```

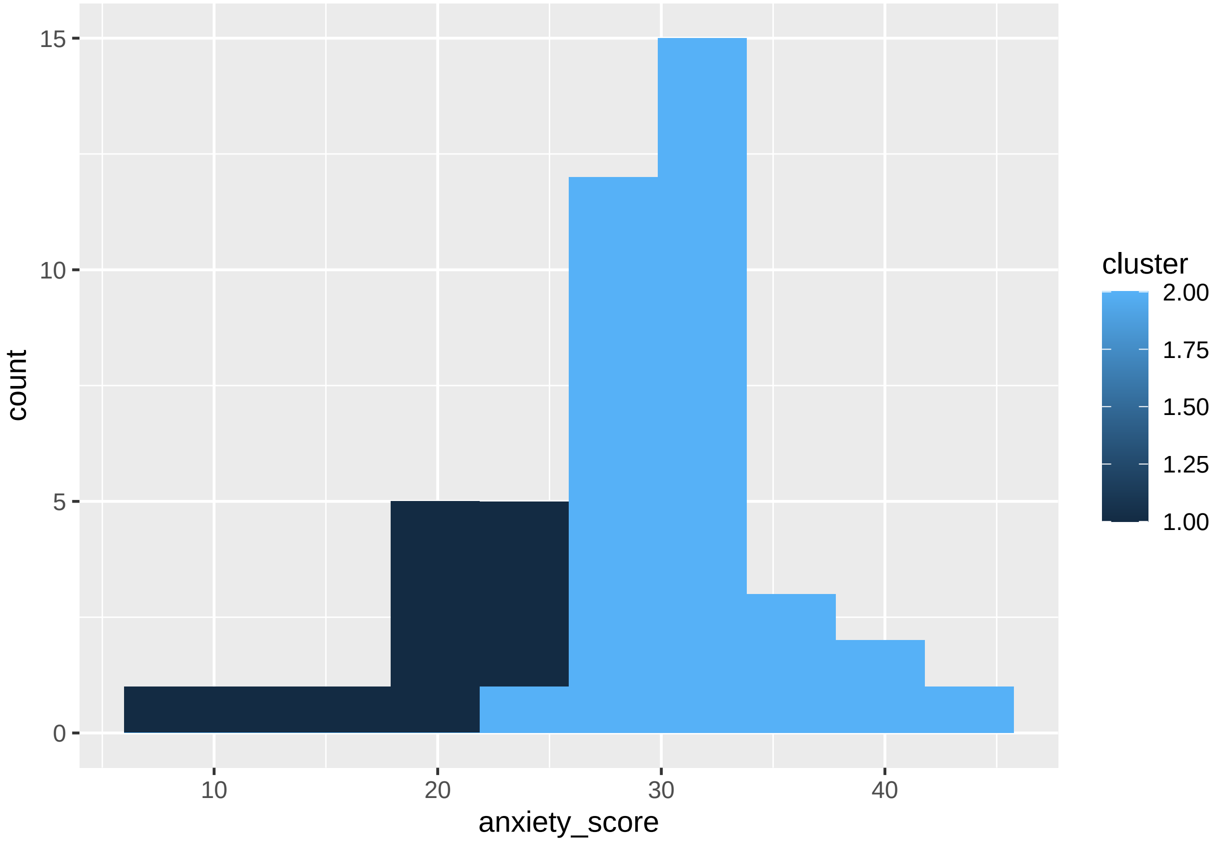
(c) Add a new column called `cluster` to the `anxiety_data` tibble by using `mutate(cluster = clustering$cluster)`.

```
# Add cluster labels to our original anxiety_data tibble
anxiety_data <- anxiety_data %>% mutate(cluster = clustering$cluster)
```

```
Q1c_anxiety_data_with_cluster_variable <- anxiety_data; # DO NOT CHANGE THIS LINE OF CODE
```

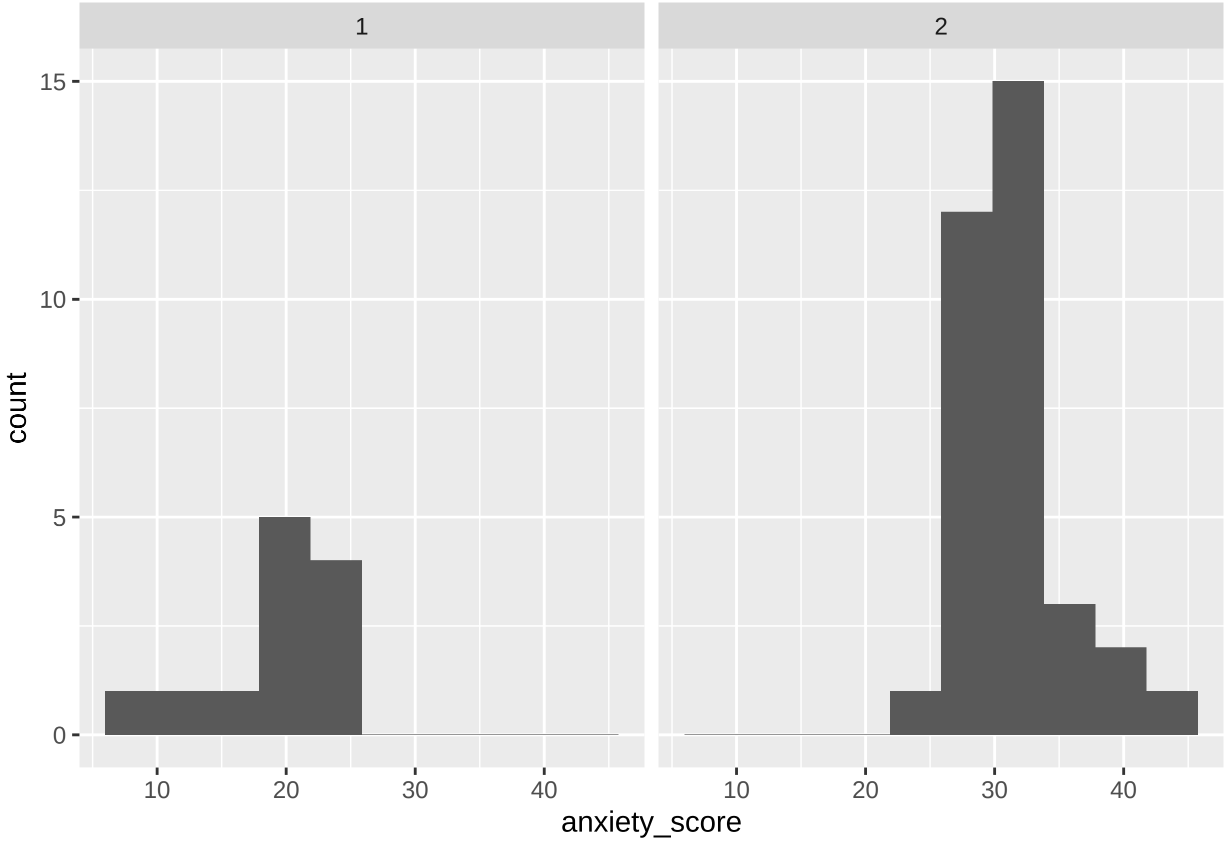
(d) Produce one histogram to visualize the two clusters you obtained above (in a single histogram). Hint: Look at the example on p43 of this week's slides.

```
# Create your visualization here
anxiety_data %>% ggplot(aes(x=anxiety_score, group=cluster, fill=cluster)) + geom_histogram(bins = 10)
```



(e) Produce two histogram to visualize the two clusters you obtained above, using facet_wrap(). Hint: Look at the example on p42 of this week's slides.

```
# Create your visualizations here
anxiety_data %>% ggplot(aes(x=anxiety_score)) + geom_histogram(bins = 10) + facet_wrap(~cluster)
```



(f) Based on your visualizations in the two parts above, comment on the clusters you obtained. Do the clusters appear to represent two distinct groups of anxiety scores? Replace NULL by the number of the cluster which has higher anxiety scores.

```
Q1f_cluster_number_high_anxiety <- 2 # Replace NULL by 1 or 2
Q1f_cluster_number_high_anxiety
```

```
## [1] 2
```

Question 2

In the anxiety data from Question 1, in addition to anxiety scores each person is also characterized as having a “Low” (less than 2 hours/day) or “High” (more than 2 hours/day) level of social media usage. In HW5, you conducted a 2-sample hypothesis test to investigate whether the median anxiety was different for “Low” and “High” social media users, but it’s not clear why 2 hours was used as the cutoff. In this question, you’ll compare the “High” vs “Low” grouping to the groupings we got from clustering the data in the previous question.

(a) You might be wondering if people who spend more time on social media have higher anxiety than those who spend less time on social media. Here, you’ll calculate the agreement between the “High” and “Low” levels of social media usage and the clusters we obtained in the previous question. Add a new variable called “categories_agree” to the `anxiety_data` tibble which takes the value TRUE if the individual is in the high social media usage AND high anxiety cluster or if they are in the low social media usage group and low anxiety group

```
anxiety_data <- anxiety_data %>%
  mutate(categories_agree = as.logical((social_media_usage == "High") ^ (cluster == 2)))  # Replace FAL
# calculate the values of this variable




Q2a_anxiety_data_with_category_agreement <- anxiety_data # DO NOT CHANGE THIS LINE
```

(b) Below is a summary value (Q2b_value) and summary table (Q2b_summary_table). After you’ve correctly completed the part above, interpret the meaning of Q2b_value and Q2b_summary_table (specifically the prop_correct column)

```
Q2b_value <- sum(anxiety_data$categories_agree) / nrow(anxiety_data)
Q2b_value
```

```
## [1] 0.6086957
```

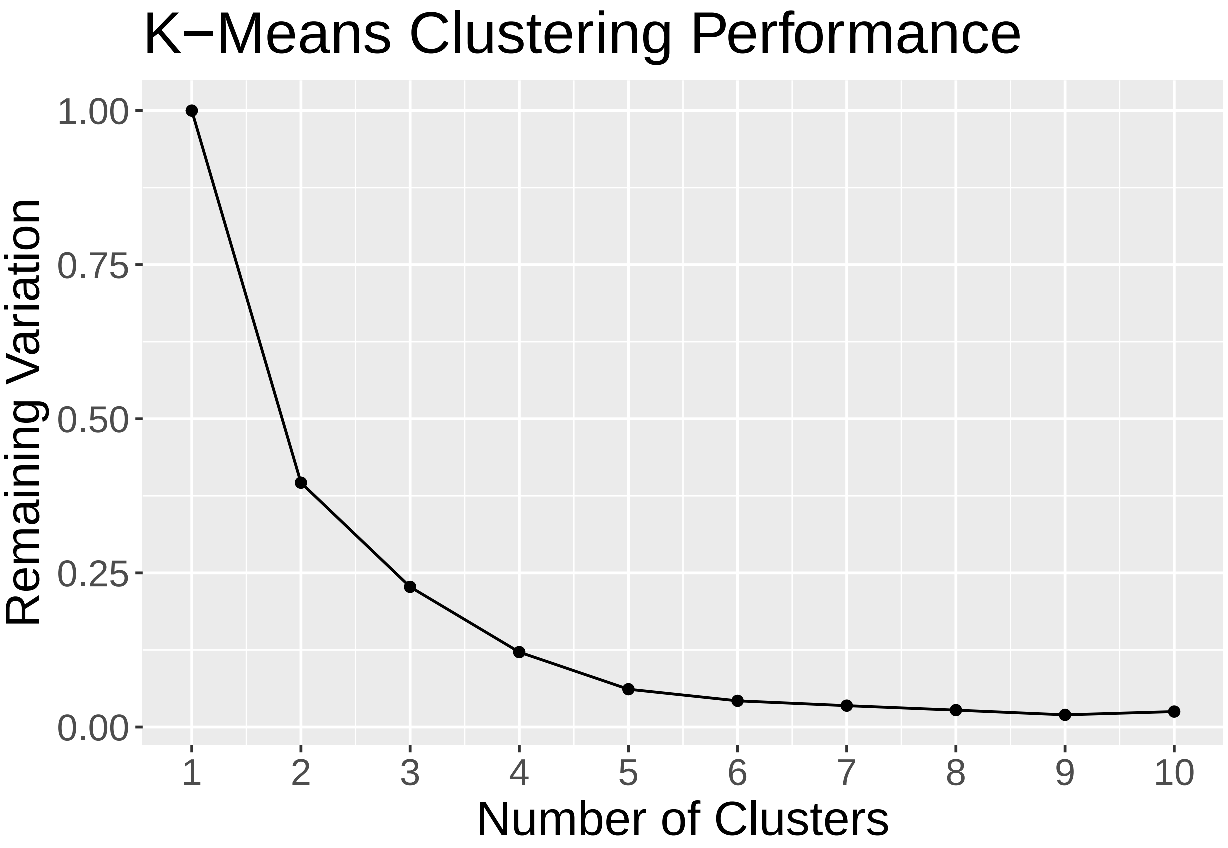
```
Q2b_summary_table <- anxiety_data %>% group_by(social_media_usage) %>%
  summarise(n=n(),
  prop_correct = sum(categories_agree) / n)
Q2b_summary_table
```

```
## # A tibble: 2 x 3
##   social_media_usage     n prop_correct
##   <chr>              <int>        <dbl>
## 1 High                  16          1
## 2 Low                   30          0.4
```

(b) The R code below shows the performance of k-means clustering with various values of k. Based on this plot, what might be a good choice for the number of clusters for anxiety score? Explain your answer in a few sentences.

```
set.seed(130) # DO NOT CHANGE THIS LINE
explained_ss <- rep(NA, 10)
for(k in 1:10){
  # run k-means on the data
  clustering <- kmeans(anxiety_data$anxiety_score, k)
  explained_ss[k] <- clustering$betweenss / clustering$totss
}
```

```
# Plot evolution of metric as a function of k
ggplot() +
  aes(x=1:10, y=1-explained_ss) +
  geom_line() +
  geom_point() +
  labs(x="Number of Clusters",
       y="Remaining Variation",
       title="K-Means Clustering Performance") +
  theme(text=element_text(size=18)) +
  scale_x_continuous(breaks=1:10) +
  scale_x_continuous(breaks=1:10)
```


K−Means Clustering Performance
Remaining Variation
Number of Clusters