Homework 8 Solution

Disclaimer

This homework solution is for the sole benefit of students taking Stat 220 from Prof. Bastola during Fall term 2022. Dissemination of this solution to people who are not registered for this course is not permitted and will be considered grounds for Academic Dishonesty for the all individuals involved in the giving and receiving of the solution.

Assignment prompt

Problem 1: Random classifiers

Suppose you have n observations and you would like to construct a classifier to predict positive outcomes. We know that n_N are actually negative cases and n_P are actually positive cases.

Suppose we are a lazy data scientist and don't actually build a model with predictors to classify our cases. Instead, we use a fair coin to classify negative and positive cases! This "coin classifier" will result in an even split between *expected* positive and negative predictions:

$$\hat{n}_N = \hat{n}_P = \frac{n}{2}$$

Using the same logic, half of the actual negatives will be positives and positives:

$$TN = FP = \frac{n_N}{2}$$

and half of the actual positives will be positives and negatives:

$$FN = TP = \frac{n_P}{2}$$

confusion matrix	truth: negative	truth positive	total
predict negative	$TN = \frac{n_N}{2}$	$FN = \frac{n_P}{2}$	$\hat{n}_N = \frac{n}{2}$
predict positive	$FP = \frac{n_N}{2}$	$TP = \frac{n_P}{2}$	$\hat{n}_P = \frac{n}{2}$
total	n_N	n_P	n

a.

Use the "coin classifier" confusion matrix above to compute accuracy, precision, sensitivity, and specificity for this fair coin prediction method.

Answer: With a fair coin flip predicting positives, the accuracy will be 50%.

$$accuracy = \frac{TN + TP}{n} = \frac{0.5n_N + 0.5n_P}{n} = 0.5\frac{n_N + n_P}{n} = 0.5$$

Precision will be equal to the rate of positives in the sample since 50% of the positives are correctly predicted and 50% of the overall sample is predicted to be positives:

$$precision = \frac{TP}{\hat{n}_P} = \frac{0.5n_P}{0.5n} = \frac{n_P}{n}$$

Sensitivity will be 50% since that is the rate at which we correctly predict positives:

$$Sensitivity = \frac{TP}{n_P} = \frac{0.5n_P}{n_P} = 0.5$$

Specificity will also be 50% since that is the rate at which we correctly predict negatives:

$$Specificity = \frac{TN}{n_N} = \frac{0.5n_N}{n_N} = 0.5$$

b.

Consider another classification method that doesn't actually build a model with predictors to classify our cases. Instead, we now predict all cases to be positive! (e.g. a biased coin that lands "positive" with probability 1)

Construct the confusion matrix for this classifier and compute accuracy, precision, sensitivity, and specificity.

Answer: Here all cases are positive, so $TN = 0$ (no negatives predicted) and $FN = 0$ (no negatives predicted). All negative cases are predicted positive, so $FP = n_N$ and all positive cases are predicted positive, so $TP = n_P$.

confusion matrix	truth negative	truth positive	total
predict negative	$TN = 0$	$FN = 0$	$\hat{n}_N = 0$
predict positive	$FP = n_N$	$TP = n_P$	$\hat{n}_P = n$
total	n_N	n_P	n

With a probability p predicting positives, the accuracy will be the rate of positives in the sample:

$$accuracy = \frac{TN + TP}{n} = \frac{0 + n_P}{n} = \frac{n_P}{n}$$

Precision will be equal to the rate of positives in the sample:

$$precision = \frac{TP}{\hat{n}_P} = \frac{n_P}{n}$$

Sensitivity will be 100% since that is the rate at which we correctly predict positives:

$$Sensitivity = \frac{TP}{n_P} = \frac{n_P}{n_P} = 1$$

Specificity will be 0% since that is the rate at which we correctly predict negatives:

$$Specificity = \frac{TN}{n_N} = \frac{0}{n_N} = 0$$

c.

Under what conditions in part (b) will accuracy of this classifier be highest? Under what conditions in part (b) will accuracy of this classifier be lowest?

Answer: Accuracy is the rate of positive cases in the data, $\frac{n_P}{n}$. So when we have a lot of positive cases in the data, predicting all cases to be positive will yield high accuracy. But when we have few positive cases the accuracy will be low.

d.

Describe the trade-off in part (b) in specificity vs sensitivity. What would these values be if instead of classifying every case as positive, you classified all cases as negative?

Answer: We have sensitivity of 1 because we are finding all true positive cases when predicting all positives. But we have specificity of 0 because no true negatives are correctly predicted. If we reverse the classification and predict all negative cases, the opposite will occur with specificy of 1 and sensitivity of 0.

Problem 2: Spam using k-nn

This example looks at a data set of about 4600 emails that are classified as spam or not spam, along with over 50 variables measuring different characteristic of the email. Details about these variables are found on the Spambase example on the machine learning data archive. The dataset linked to below is a slightly cleaned up version of this data. The only extra column in the data is `rgroup` which is a randomly assigned grouping variable (groups 0 through 99) which we will eliminate from the data.

Read the data in using the commands below to create a response `class` variable that contains the factor levels `spam` and `nonspam` with `spam` the first level.

```
> # tsv = tab separated values!
> spam <- read_delim("https://raw.githubusercontent.com/deepbas/statdatasets/main/spam.txt",
+         delim="\t")
>
> # some clean up
> spam <- spam %>%
+   mutate(class = fct_recode(
+     spam,
+     spam = "spam" ,
+     nonspam = "non-spam"), # rename levels because caret doesn't like "non-spam"
+     class = fct_relevel(class, "spam") # make "spam" the first level (our "positive")
+     ) %>%
+   select(-rgroup, -spam) # don't need random group variable and  spam variable
> levels(spam$class)  # verify "spam" is level 1
[1] "spam"    "nonspam"
```

What proportion of cases are spam in this data set? The variable `class` is the true classification of an email.

answer:

```
> spam %>%
+   count(class) %>%
+   mutate(n/sum(n))
# A tibble: 2 x 3
  class       n `n/sum(n)`
  <fct>   <int>      <dbl>
1 spam     1813      0.394
2 nonspam  2788      0.606
```

b.

We want to fit a k-nn classifier for `spam` using the 57 quantitative predictors (columns 1-57) from the spam data to an 80%/20% training and test set split. Create the training and test sets to do this.

Use the following seed and the `sample` command generate training data (this *should* mean your train/test split should be the same as the solution key!):

```
> set.seed(757302859)  # set a seed
```

answer:

```
> set.seed(757302859)  # set a seed
>
> spam_split <- initial_split(spam, prop = 0.80)
> # Create training data
> spam_train <- spam_split %>%
+                     training()
> # Create testing data
> spam_test <- spam_split %>%
+                     testing()
```

c.

Make a recipe for fitting k nearest-neighbor algorithm to the training data by inputting the formula and preprocessing steps.

```
> spam_recipe <- recipe(class ~ ., data = spam_train) %>%
+   step_scale(all_predictors()) %>%
+   step_center(all_predictors()) %>%
+   prep()
```

d.

Specify the nearest neighbor model to fit a classification model using 10 neighbors.

```
> spam_knn_spec <- nearest_neighbor(mode = "classification",
+                                   engine = "kknn",
+                                   weight_func = "rectangular",
+                                   neighbors = 10)
```

e.

Define the workflow object feeding in the recipe and model specification. Then, fit the model to the training data. *answer:*

```
> spam_workflow <- workflow() %>%
+   add_recipe(spam_recipe) %>%
+   add_model(spam_knn_spec)
>
> spam_fit <- fit(spam_workflow, data = spam_train)
```

f.

Evaluate the model on the test datasetand predict `class` for the test data set.

```
> test_features <- spam_test  %>% select(-class)
> spam_pred <- predict(spam_fit, test_features, type = "raw")
> spam_results <- spam_test %>%
+   select(class) %>%
+   bind_cols(predicted = spam_pred)
```

g.

Compute the accuracy, sensitivity, specificity and precision for predicting test set responses. Which rate(s) would you like to be high as a user of this spam filter?

Answer:

As someone who rarely looks in my spam folder, I want to make sure that there are very few "false positives", meaning very few non-spam emails that get dumped in my spam folder. So I want high specificity and high precision in the filter.

```
> custom_metrics <- metric_set(accuracy, sens, spec, ppv)
> metrics <- custom_metrics(spam_results,
+                truth = class,
+                estimate = predicted)
> metrics <- metrics %>% select(-.estimator)
```

h.

Use the `tidymodels` package to do 10-fold cross validation as follows:

- use the 80% training data split from part b.
- tune your knn spam classifier based on accuracy
- consider neighborhood sizes ranging from size 1 to 31

After running `train` to produce the 10-fold CV results (this could take ~1 minute), use the `results` from your `train` object to get the training set cross-validated estimates of the accuracy, precision, sensitivity and specificity of your final ("best") classifier.

And use the following seed before running your `train` command:

```
> set.seed(30498492)
```

answer

Run CV for knn:

```
> spam_recipe_new <- recipe(class ~ ., data = spam_train) %>%
+   step_scale(all_predictors()) %>%
+   step_center(all_predictors())%>%
+   prep()
```

```
> knn_spec <- nearest_neighbor(
+   weight_func = "rectangular",
+   neighbors = tune()
+ ) %>%
+   set_engine("kknn") %>%
+   set_mode("classification")
```

```
> spam_vfold <- vfold_cv(spam_train, v = 10, strata = class)
```

```
> k_vals <- tibble(neighbors = seq(from = 1, to = 30, by = 1))
```

```
> knn_fit <- workflow() %>%
+   add_recipe(spam_recipe_new) %>%
+   add_model(knn_spec) %>%
+   tune_grid(
+     resamples = spam_vfold,
+     grid = k_vals,
+     metrics = custom_metrics
+     )
```

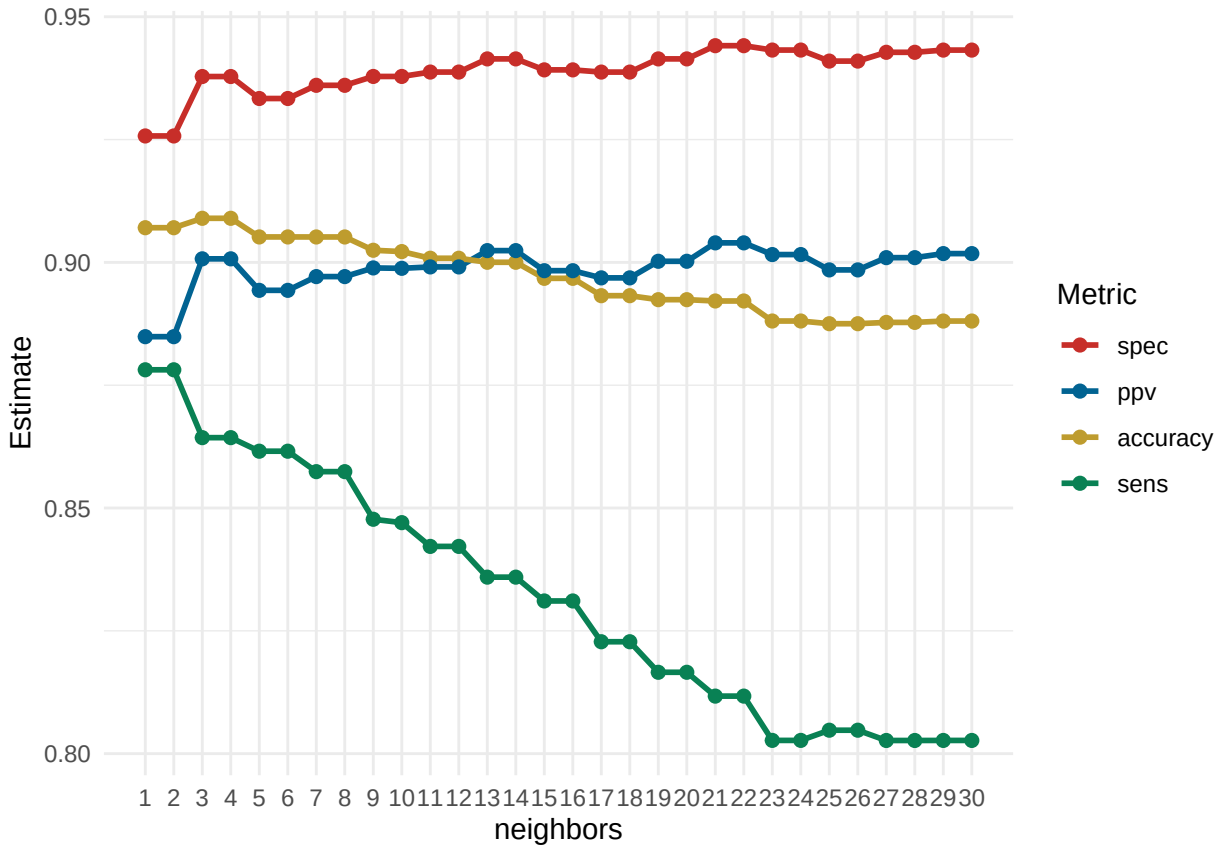
i.

Use the `results` object from your CV fit from part (i) to plot the accuracy, precision, sensitivity and specificity for the values of k that you tuned over. Comment on which neighborhood sizes are "best" in terms of these four metrics (your answer may depend on the metric).

answer

Low values of k seem best for high precision, accuracy and sensitivity while higher values yield higher specificity. Since "non-spam" is the majority of cases (~2/3 of cases are non-spam), the larger the neighborhood the more likely it is that we see majority "non-spam" which means higher specificity (correctly predicting non-spam cases).

```
> cv_metrics <- collect_metrics(knn_fit)
> final.results <- cv_metrics %>%  mutate(.metric = as.factor(.metric)) %>%
+   select(neighbors, .metric, mean)
>
> final.results %>%
+   ggplot(aes(x = neighbors, y = mean, color = forcats::fct_reorder2(.metric, neighbors, mean))) +
+   geom_line(size = 1) +
+   geom_point(size = 2) +
+   theme_minimal() +
+   scale_color_wsj() +
+   scale_x_continuous(breaks = seq(1,30)) +
+   theme(panel.grid.minor.x = element_blank()) +
+   labs(color='Metric', y = "Estimate")
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
This warning is displayed once every 8 hours.
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
generated.
```



Problem 3: Incoming student characteristic

We will look at a "classic" college data set of a random sample of colleges and universities. To simplify our look at this data, we will filter to only look at MN, MA, and CA schools

```
> colleges <- read_csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/Colleges.csv")
> names(colleges)
 [1] "...1"        "State"       "College"     "SATM"        "SATV"
 [6] "AppsReceive" "AppsAccept"  "HStop10"     "HStop25"     "FullTime"
[11] "Tuition"     "RoomBoard"   "Books"       "Ratio"       "Donate"
[16] "Expend"      "GradRate"    "Type"        "AvgSalary"   "NumFaculty"
> colleges2 <- colleges %>%
+   filter(State %in% c("MN","MA","CA"))
> colleges2 %>% count(State)
# A tibble: 3 x 2
  State     n
  <chr> <int>
1 CA       21
2 MA       19
3 MN       11
```

We will also just focus on student body characteristics (incoming class averages) for SAT and the HS variables (which are are the proportion of the incoming class that is in the top 10% or 25% of their HS class). Here we select just these characteristics and college name and state.

```
> colleges2 <- colleges2 %>% select(1,2,3,4,5,8,9)
> colleges2
# A tibble: 51 x 7
    ...1 State College                            SATM  SATV HStop10 HStop25
   <dbl> <chr> <chr>                             <dbl> <dbl>   <dbl>   <dbl>
 1     9 CA    California Institute of Technolo   750   660      98     100
 2    10 CA    California Lutheran University     495   436      23      52
 3    11 CA    California Polytechnic-San Luis    547   455      47      73
 4    12 CA    Chapman University                 501   456      23      48
 5    13 CA    Claremont McKenna College          670   600      71      93
 6    14 CA    Harvey Mudd College                740   630      95     100
 7    15 CA    Pitzer College                     590   560      37      73
 8    16 CA    Pomona College                     700   640      80      98
 9    17 CA    Scripps College                    590   560      60      83
10    18 CA    Occidental College                 570   510      52      81
# ... with 41 more rows
```

Let's cluster schools by their incoming class characteristics.

(a)

Why should we standardize the variables of interest before using a clustering method?

Answer:

The size and variability of the variables differs. If we didn't standardize then any method that uses Euclidean distance will favor SAT measures at the expense of the HS measures.

(b)

Standardize the four quantitative clustering variables (`SATM`, `SATV`, `HStop10`, `HStop25`).

answer:

```
> standardize <- function(x) {
+   (x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)
+ }
>
> colleges2_std <- colleges2 %>%
+   select(4:7) %>%
+   mutate_all(standardize)
```

(c)

Fit a k-means cluster algorithm to standardized predictors from (b) using k=5 group centers and 20 different starting locations. Extract the total within cluster and total sums of squares. Set a seed to make your results reproducible.

answer total within cluster SS is about 21 and the tota SS is about 200.

```
> set.seed(2957402)
> km5 <- kmeans(colleges2_std, centers = 5, nstart = 20)
> km5$totss   # total SS
[1] 200
> km5$tot.withinss   # total within SS
[1] 21.38091
```

(d)

Suppose we run `kmeans` with k=3. Will the within cluster variation go up or down compared to k=5? Explain your answer.

answer

Within cluster variation will increase because we are forcing more distant points to be included in the same cluster. We can verify that here:

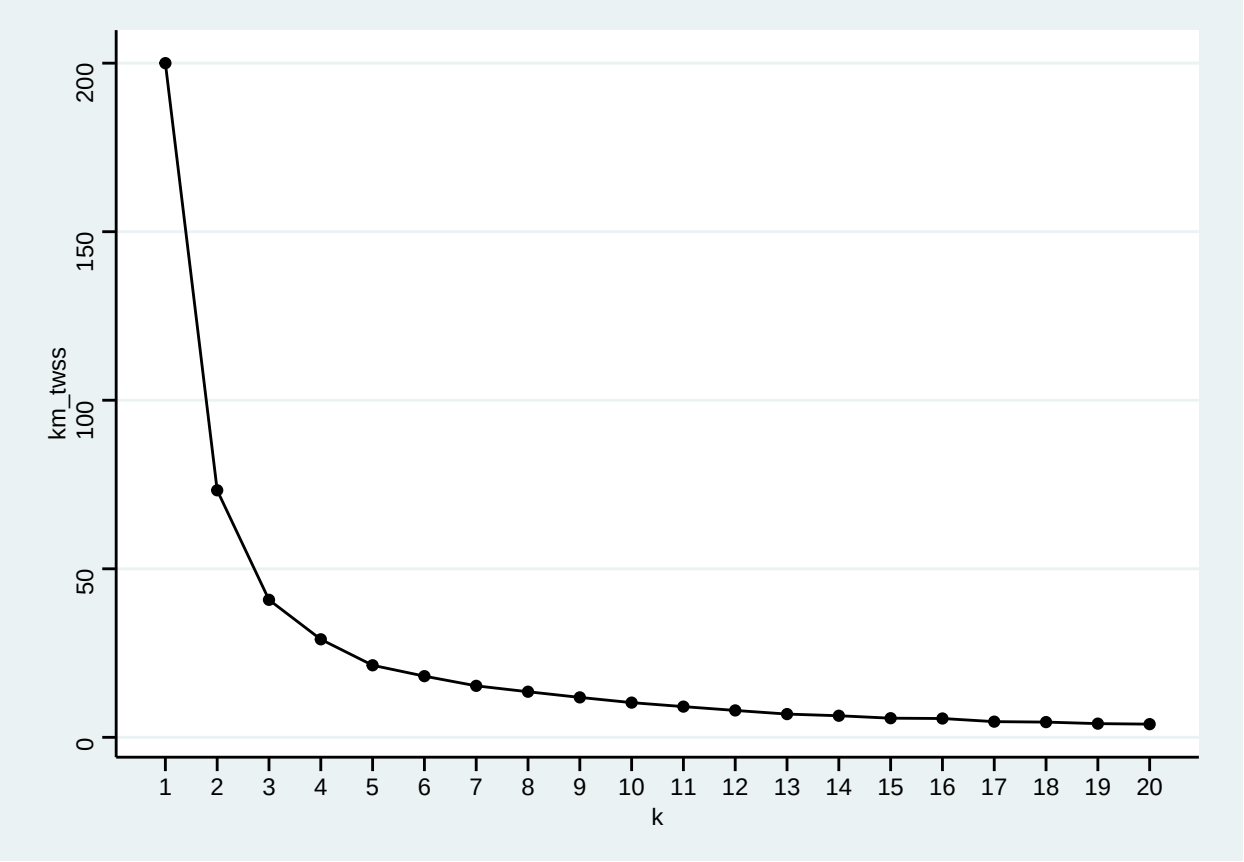
```
> set.seed(2957402)
> km3 <- kmeans(colleges2_std, centers = 3, nstart = 20)
> km3$totss   # total SS
[1] 200
> km3$tot.withinss   # total within SS
[1] 40.79375
```

(e)

Use a `map` family of command to fit 20 different kmean algorithms from $K = 1$ to $K = 20$ clusters. Plot the total within cluster SS against the number of clusters. Which choice of K looks best for grouping these schools?

answer:

```
> km_twss <- purrr::map_df(1:20, .f = function(x) {
+   twss <- kmeans(colleges2_std, centers=x, nstart=20)$tot.withinss
+   tibble(k = x, km_twss = twss)
+ })
> km_twss %>% head()
# A tibble: 6 x 2
      k km_twss
  <int>   <dbl>
1     1   200
2     2    73.3
3     3    40.8
4     4    29.1
5     5    21.4
6     6    18.2
> ggplot(km_twss, aes(x = k, y = km_twss)) +
+   geom_point() +
+   geom_line() +
+   scale_x_continuous(breaks = 1:20)
```



$K = 3$ is where we see the "elbow" in this plot. Higher cluster values will decrease the SS but not by enough to make it worth while (just adds complexity).

(f)

Fit a k-means clustering with $K = 3$, then add the cluster ID to the `college2` data frame. Make sure to make this ID a character vector.

answer:

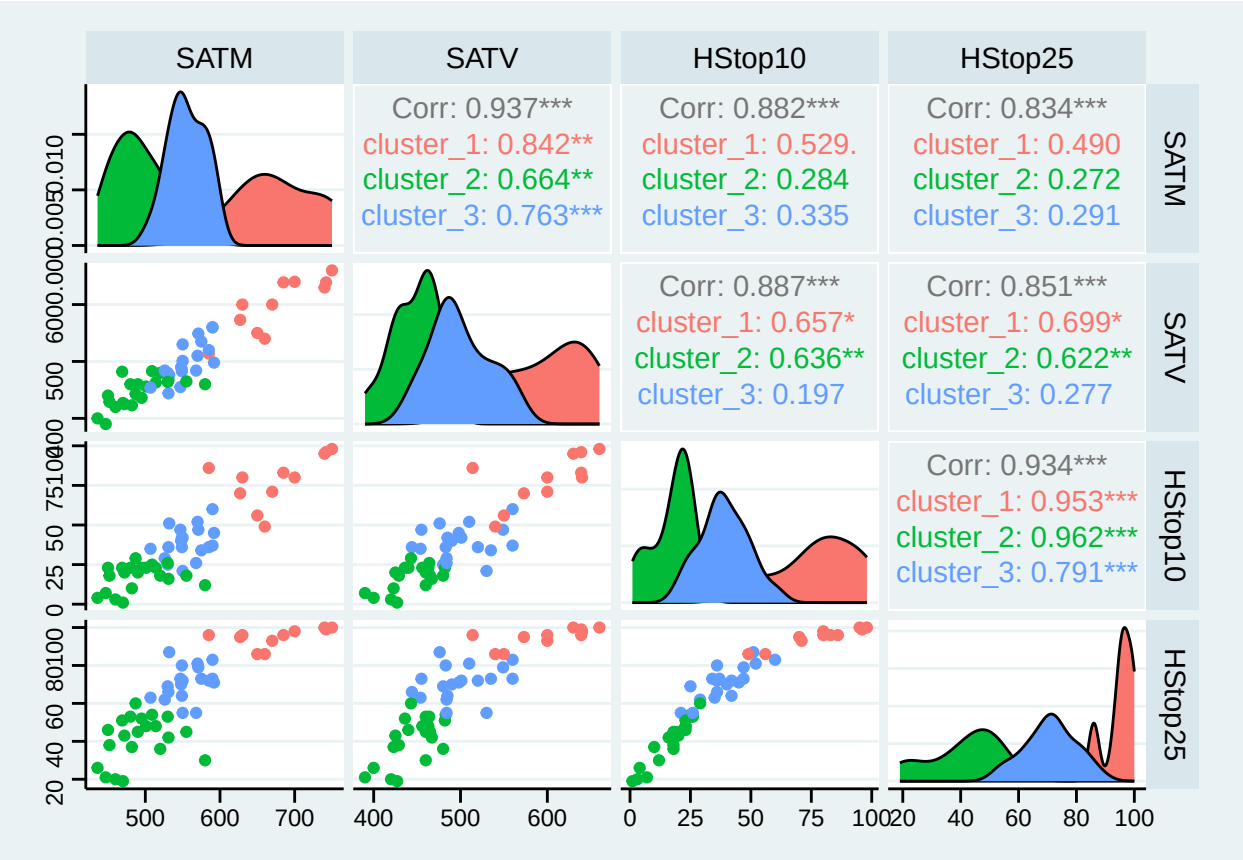
I already fit this model in part (d) and the output was named `km3`. Here we add cluster assignments called `cluster_km3` to the full data set

```
> colleges2 <- colleges2 %>%
+   mutate(cluster_km3 = stringr::str_c("cluster_", km3$cluster))
> glimpse(colleges2)
Rows: 51
Columns: 8
$ ...1        <dbl> 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23,~
$ State       <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA"~
$ College     <chr> "California Institute of Technolo", "California Lutheran U~
$ SATM        <dbl> 750, 495, 547, 501, 670, 740, 590, 700, 590, 570, 532, 472~
$ SATV        <dbl> 660, 436, 455, 456, 600, 630, 560, 640, 560, 510, 476, 425~
$ HStop10     <dbl> 98, 23, 47, 23, 71, 95, 37, 80, 60, 52, 51, 20, 86, 36, 23~
$ HStop25     <dbl> 100, 52, 73, 48, 93, 100, 73, 98, 83, 81, 87, 43, 96, 80, ~
$ cluster_km3 <chr> "cluster_1", "cluster_2", "cluster_3", "cluster_2", "clust~
```

(g)

Clustering methods don't just determine clusters based on individual variable values, but how these variables combine will all other variables (e.g. clusters in R^4 space for this problem). We can also look at a 2-d view of the data using a scatterplot matrix. We will use the ggplot "add on" package called `GGally`

```
> library(GGally)    # install if needed
> colleges2 %>%
+   ggpairs(aes(color = cluster_km3),
+           columns=c("SATM", "SATV", "HStop10", "HStop25"))
```



Use the graphical EDA above as a starting point to describe the k-means cluster assignments. (e.g. how to the cluster groupings relate SAT scores, type of college, etc).

Answer:

The k-means groups correspond to clusters of schools with "high" (red), "mid" (blue) and "low" (green) achieving students based on their incoming test and high school measures. Note that the coloring/cluster ID of your groups may be different from my groups (because the algorithm is random) but the basic overall pattern should be observed in "low", "middle" and "high" clusters.