Lab 10

Enter the names of the group members here: Austine Do, Graceanne Becker, Catherine Zhong

This assignment is due by the end of the lab. Only one student in the group submits a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

Let's load the appropriate packages for today:

```r
library(tidyverse)
library(ggcorrplot)
library(factoextra)
library(cluster)
```

In this lab, we will use a dataset that contains demographic information about 850 customers of a credit card company:

```r
Customer_Segmentation <- read_csv("https://raw.githubusercontent.com/laylaguyot/datasets/main//Customer_

# Take a quick look!
head(Customer_Segmentation)
```

```
## # A tibble: 6 x 7
##      Id   Age Education Years_Employed Income Card_Debt Other_Debt
##   <dbl> <dbl>     <dbl>          <dbl>  <dbl>     <dbl>      <dbl>
## 1     1    41         2              6     19     0.124      1.07
## 2     2    47         1             26    100     4.58       8.22
## 3     3    33         2             10     57     6.11       5.80
## 4     4    29         2              4     19     0.681      0.516
## 5     5    47         1             31    253     9.31       8.91
## 6     6    40         1             23     81     0.998      7.83
```

The `Education` variable is coded as: 1 = high school, 2 = some college, 3 = college degree, 4 = graduate degree. The `Income`, `Card_Debt`, and `Other_Debt` are reported in thousands of dollars.

Suppose that the company is about to advertise new types of credit cards. The goal of the lab is to identify different groups of customers based on the variables available in the dataset.

Question 1 (2 pts)

Which variable in the dataset above should we drop in this analysis and why? Overwrite `Customer_Segmentation` without that variable.

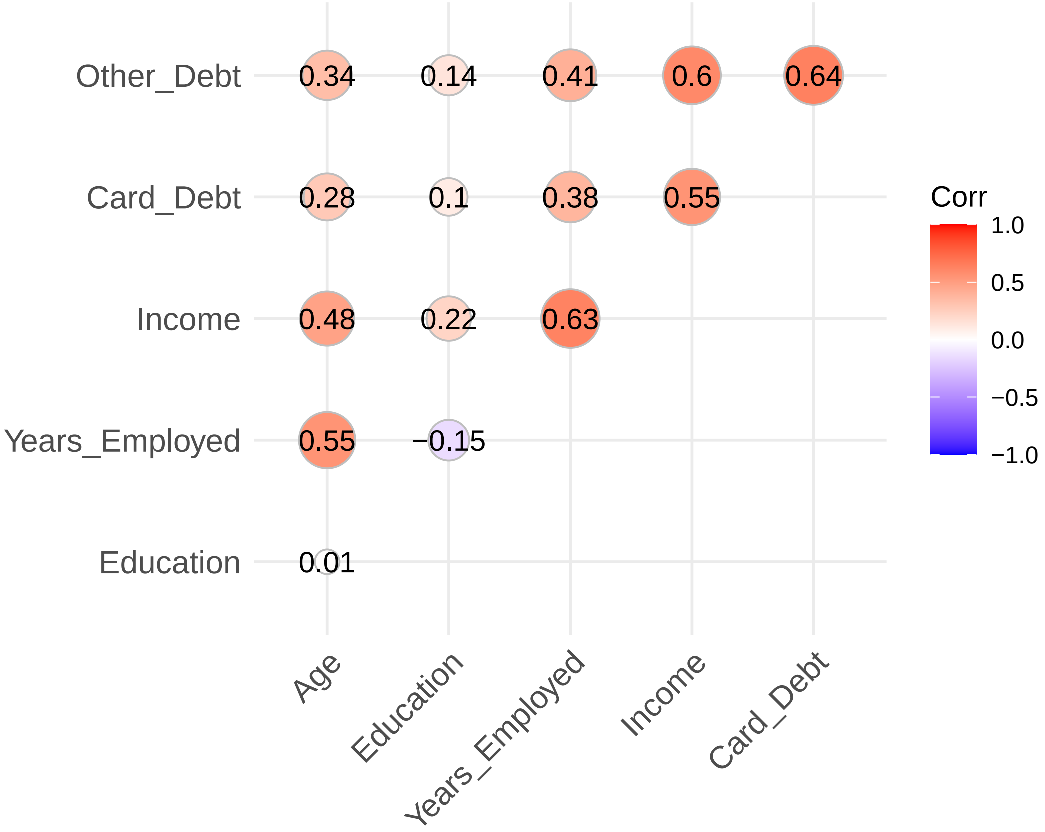
We should drop the `Id` variable because it is not a meaningful variable terms of useful information for each observation

```
# Selecting the variables that make sense
Customer_Segmentation <- Customer_Segmentation |>
    select(Age, Education, Years_Employed, Income, Card_Debt, Other_Debt)
```

Question 2 (3 pts)

Create a correlation matrix displaying the correlation coefficient for each pair of variables.

```
# Visualization of the correlation coefficient matrix for each pair of variables
ggcorrplot(cor(Customer_Segmentation),
           type = 'upper',
           lab = TRUE,
           method = 'circle')
```



Which pair of variables has the strongest positive correlation? Why do you think that makes sense?

`Other_Debt` and `Card_Debt` have the highest correlation and this makes sense since if you have card debt then probably also accumulated some other type of debt like student debt.

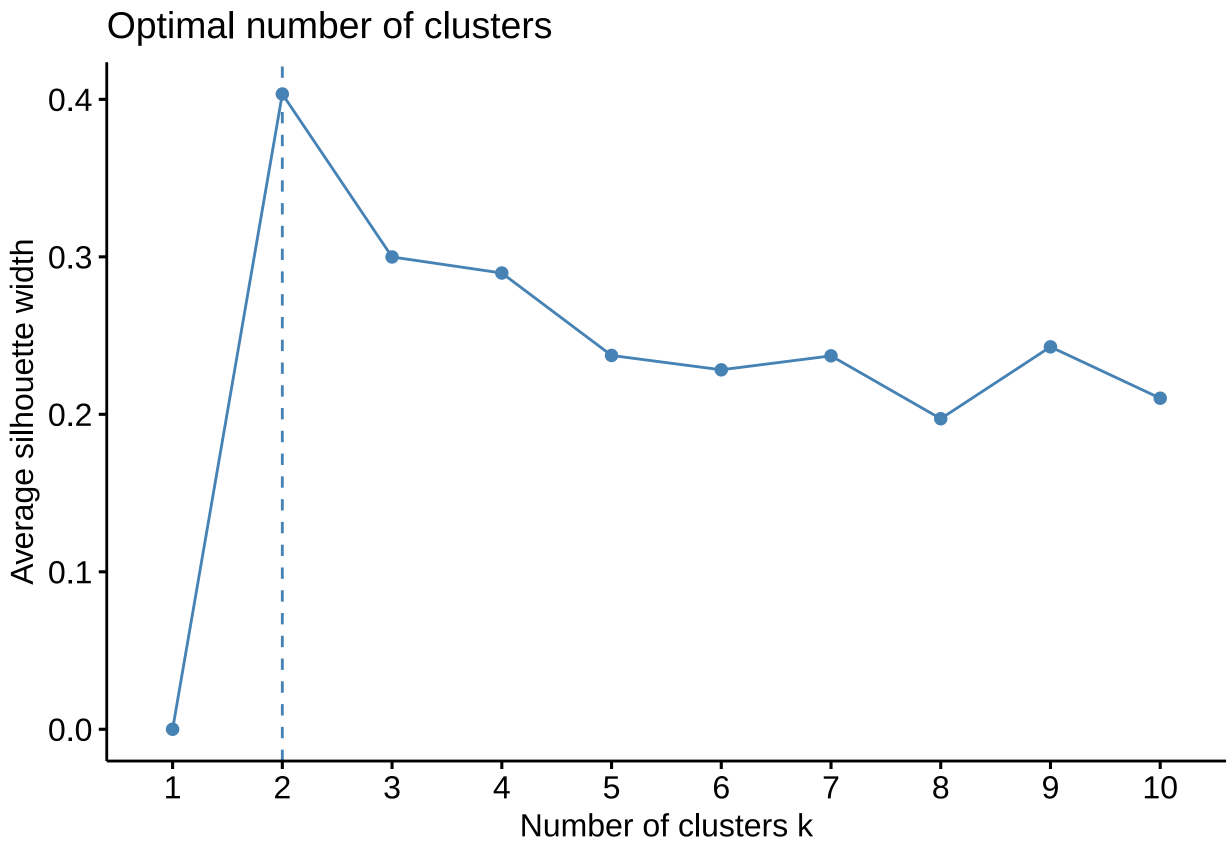
Which pair of variables has a negative correlation? Why do you think that makes sense?

`Years_Employed` and `Education` have a negative correlation and this makes sense because there are those who immediately enter the work force following high school.

Question 3 (8 pts)

We will try to identify groups of customers based on their characteristics in `Customer_Segmentation`. First, find how many clusters you should consider according to the average silhouette width. *Hint: use `fviz_nbclust()`.*

```
# Scaled the dataset and then found number of appropriate clusters
customer_seg_scaled <- Customer_Segmentation |>
    scale()
fviz_nbclust(customer_seg_scaled, kmeans, method = "silhouette")
```



We should consider 2 clusters

Second, use the appropriate number of clusters you found above and apply the clustering algorithm,`kmeans`, on `Customer_Segmentation` (remember to scale the variables first):

```
# For reproducible results
set.seed(322)
```

```r
# applying clustering algorithm
kmeans_results <- customer_seg_scaled |>
    kmeans(center = 2)

kmeans_results
```

```
## K-means clustering with 2 clusters of sizes 185, 665
##
## Cluster means:
##          Age  Education Years_Employed     Income  Card_Debt Other_Debt
## 1  0.9932068  0.2711557      1.1957107  1.2752423  1.0509821  1.1539959
## 2 -0.2763056 -0.0754343     -0.3326413 -0.3547667 -0.2923785 -0.3210365
##
## Clustering vector:
##   [1] 2 1 1 2 1 1 2 2 2 1 1 2 2 2 2 2 2 2 1 2 2 2 2 1 1 2 2 2 2 2 1 1 2 2 2 2 2
##  [38] 2 2 1 2 1 2 1 2 1 2 2 2 2 1 2 2 1 1 2 2 2 2 2 1 2 1 1 2 2 2 2 2 2 2 2 1 2
##  [75] 2 2 2 2 1 1 2 1 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [ reached getOption("max.print") -- omitted 750 entries ]
##
## Within cluster sum of squares by cluster:
## [1] 1743.650 1800.971
##  (between_SS / total_SS =  30.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Once you have the results of the clustering algorithm, add a variable to `Customer_Segmentation` to assign the appropriate cluster to each customer. Find the mean of each variable in `Customer_Segmentation` for each cluster.

```r
# Assigning clusters and finding means/centers
Customer_Segmentation |>
    mutate(cluster = as.factor(kmeans_results$cluster))
```

```
## # A tibble: 850 x 7
##      Age Education Years_Employed Income Card_Debt Other_Debt cluster
##    <dbl>     <dbl>          <dbl>  <dbl>     <dbl>      <dbl> <fct>
##  1    41         2              6     19     0.124      1.07  2
##  2    47         1             26    100     4.58       8.22  1
##  3    33         2             10     57     6.11       5.80  1
##  4    29         2              4     19     0.681      0.516 2
##  5    47         1             31    253     9.31       8.91  1
##  6    40         1             23     81     0.998      7.83  1
##  7    38         2              4     56     0.442      0.454 2
##  8    42         3              0     64     0.279      3.94  2
##  9    26         1              5     18     0.575      2.22  2
## 10    47         3             23    115     0.653      3.95  1
## # i 840 more rows
```

```r
kmeans_results$centers
```

```
##          Age  Education Years_Employed     Income  Card_Debt Other_Debt
## 1  0.9932068  0.2711557      1.1957107  1.2752423  1.0509821  1.1539959
## 2 -0.2763056 -0.0754343     -0.3326413 -0.3547667 -0.2923785 -0.3210365
```

How would you describe an average customer in each cluster?

The average customer in cluster 1 appear to be more educated, have more work experience, higher income, and higher overall debt while the average customer in cluster 2 is the complete opposite

---

Question 4 (4 pts)

The credit card company wants to use the clustering results for their marketing campaign advertising for new credit cards (this strategy is called market segmentation and aims at selling more products with less marketing expenses). Here is the advertisement for three new types of credit cards:

- *CreditStarter* card: "Start Building Your Credit! Our Introductory Credit Card offers a low limit and benefits designed for cautious spenders—begin your credit journey responsibly."
- *DebtConsolidator* card: "Simplify Your Finances! Transfer balances hassle-free with our Low-Interest Balance Transfer Card, designed to help you consolidate debt and save on interest payments."
- *ElitePlus* card: "Unlock Elite Benefits! Our Premium Card offers exclusive rewards and privileges tailored to your lifestyle. Enjoy higher credit limits and perks designed for your financial confidence."

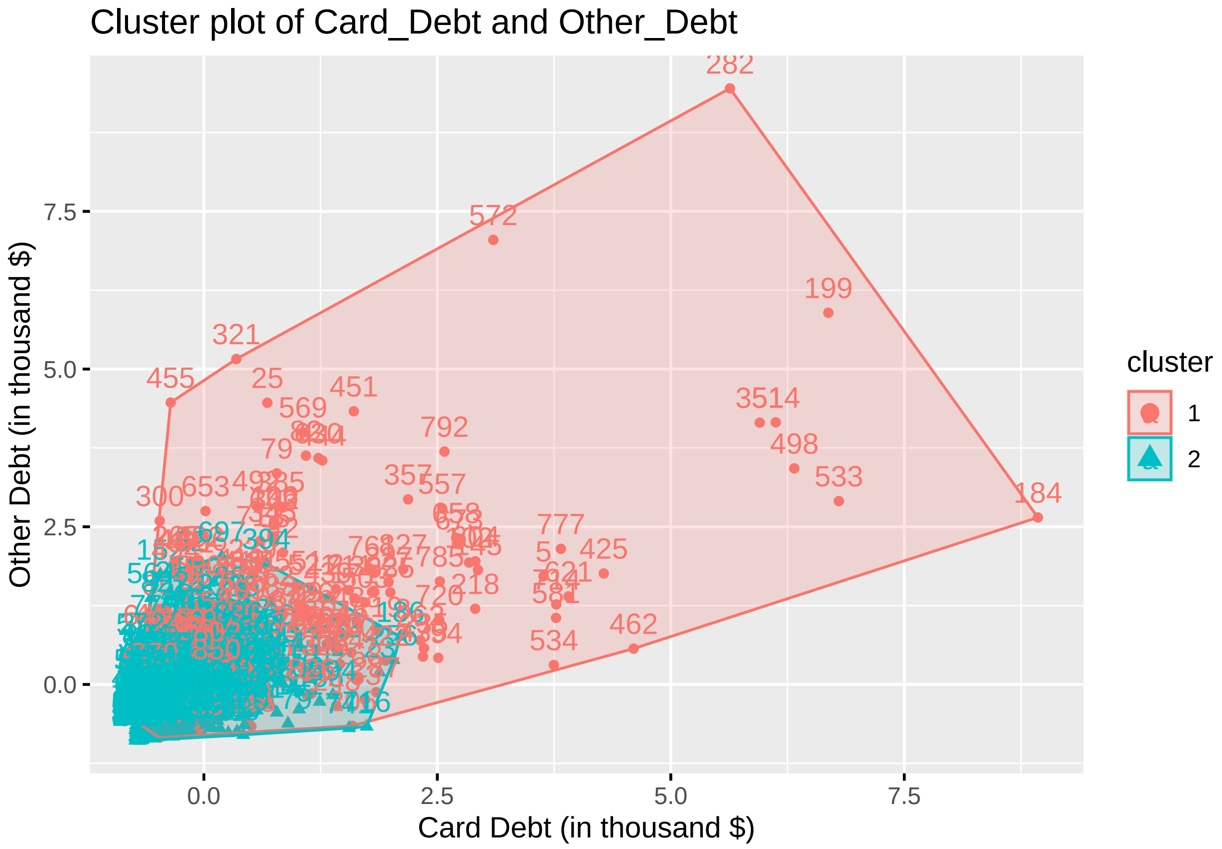
As a data scientist at the credit card company, which card would you recommend for the customers in each cluster? Why?

The credit card I would recommend to cluster 1 is ElitePlus since they have more work experience, more education, and higher income which may allow them to better use this product. I would recommend CreditStarted or DebtCosolidator to cluster 2 since they seem to be younger, less educated, and have much less debt than those in cluster 1, enabling those in cluster 2 to better leverage the benefits of the CreditStarter or DebtConsolidator products.

Question 5 (6 pts)

Let's visualize the clusters. First, pick any two variables in `Customer_Segmentation` and represent the relationship between these two variables, coloring by each cluster. How do the groups of customers differ based on these two variables?

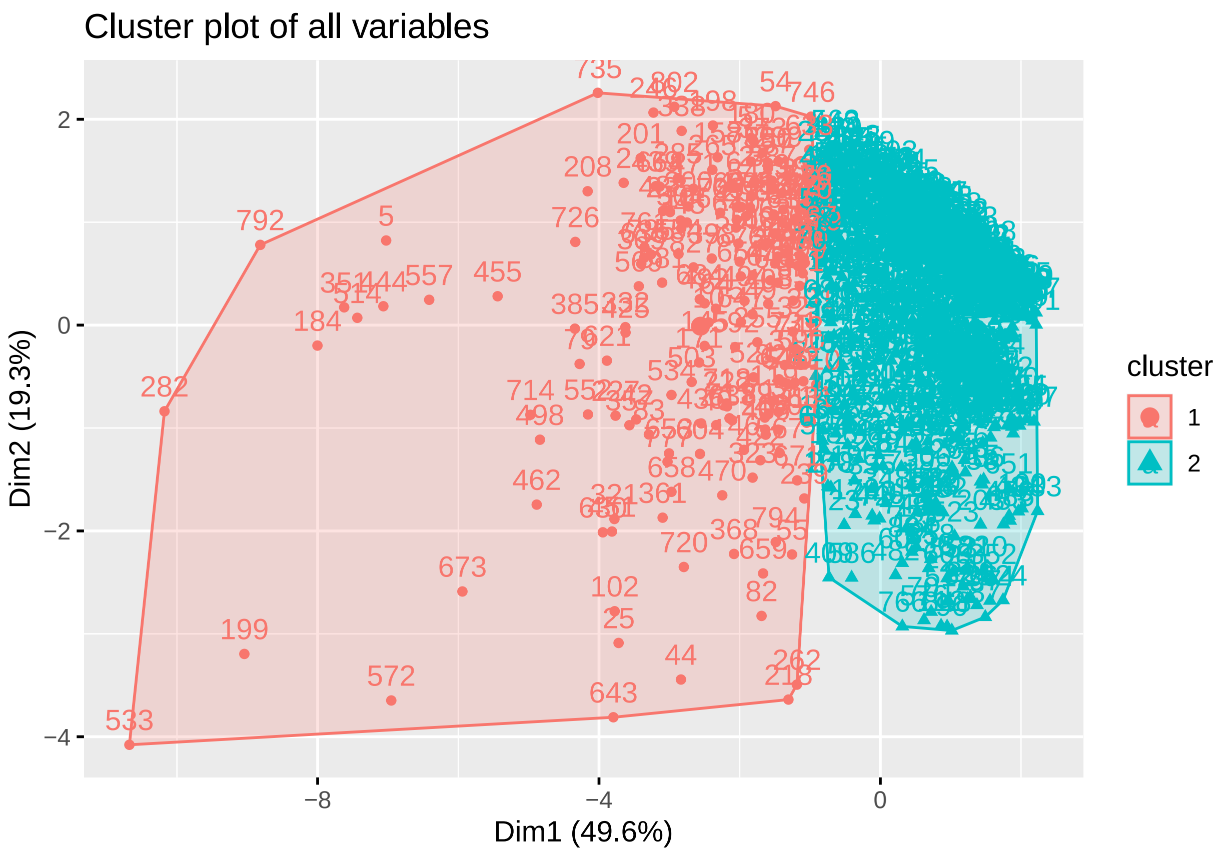
```r
# Visualization of `Card_Debt` and `Other_Debt` in the cluster plot
fviz_cluster(kmeans_results, data = Customer_Segmentation |> select(Card_Debt, Other_Debt)) +
    labs(title = 'Cluster plot of Card_Debt and Other_Debt',
         x = 'Card Debt (in thousand $)',
         y = 'Other Debt (in thousand $)')
```



It seems that variance in both types of debt across cluster 1 is much higher and more spread out while the variance of both types of debt across cluster 2 is much lower, indicating that those customers in cluster 2 have lower debt than cluster 1.

Second, represent the clusters using all variables available but in a 2-dimensional plots with the first two principal components. *Hint: this is really easy to do with `fviz_cluster()`.*

```r
# Visualization of all variables in the dataset in the cluster plot
fviz_cluster(kmeans_results, data = Customer_Segmentation) +
    labs(title = 'Cluster plot of all variables')
```



Can you see a better distinction between the two groups of customers?

Yes this plot better visualizes the distinction between the two groups of customers

How much variation is explained by the first two principal components?

In total the first two principle components account for 68.9% of the variation in the data

---

Question 6 (1 pt)

After exploring how customers naturally split into different groups, did the data match your expectations or not? If the data differed from your expectation, provide a possible explanation for why the data differed from what you expected.

The data matched our expectations as we assumed a higher income, higher education level, and more work experience would lead to higher Card and Other debt since those with higher income, more education, and more work experience would be able to take on more debt than those with lower income, less education, and less work experience.

Formatting: (1 pt)

Make sure the names of all group members are included at the beginning of the document.

Knit your file! You can knit into pdf directly or into html. Once it knits in html, click on `Open in Browser` at the top left of the window pops out. Print your html file into pdf from your browser.

Any issue? Ask other classmates or TA!

Finally, remember to select pages for each question when submitting your pdf to Gradescope and to identify your group members.