# Group Assignment 2 Submission

AUTHORS
Chris Chian (1759674)
Qianhui Li (1756140)
Frederick Arnhold (1756567)

## Part A. Cleaning and Analysing Data (60 points)

### 1. Introduction (10 pts)

- Dataset used: CEO_pay.
- Research Question: State the research question you are answering. Ensure it is answerable with the data you will use.

What CEO types emerge from the data when we perform clustering analysis based on CEO characteristics and/or firm characteristics? How can these CEO types inform our clients' decisions on how much to pay their CEOs? -

- Method: Is your approach descriptive, causal or predictive?
  -Our approach is descriptive. We are using K-Means clustering to segment CEO pays based on past data on a wide variety of firms and CEOs. We clustered based on both CEO and financial characteristics producing around four clusters in total.

We believe that our clients can use these clusters to examine and understand past trends in CEO pays, to either use it as a benchmark for future decisions on how to pay their CEOs or as a look at an ineffective status quo our clients may need to break away from.

### 2. Data Preparation & Analysis (50 pts)

**Instructions**

- Show and explain your data cleaning, wrangling, and variable creation.
- Present models, visualisations, and diagnostics.
- Ensure code is reproducible and commented.

Load Libraries:

```r
# load the libraries you need for your analysis here

library(tidyverse)
library(dplyr)
library(ggplot2)
library(recipes)
library(cluster)
library(factoextra)
library(broom)
# If you need to install a package run install.package('package_name') once
# then load it via library(package_name)
```

Load the initial dataset:

```
ceo_pay <-
  read_csv("data/ceo_pay.csv")
# load the initial dataset in this cell
```

## Data Cleaning

Perform any data cleaning steps in this section. Insert code chunks and comments / explanations as necessary.

```
Q1 <- quantile(ceo_pay$ceo_total_compensation, 0.25)
Q3 <- quantile(ceo_pay$ceo_total_compensation, 0.75)
IQR <- Q3 - Q1

lower_bound <- Q1 - (1.5 * IQR)
upper_bound <- Q3 + (1.5 * IQR)

# delete outliers
ceop_no_outliers <-
 ceo_pay[ceo_pay$ceo_total_compensation >= lower_bound & ceo_pay$ceo_total_compensatio
  drop_na() # take out all NAs
```

```
ceop_groups <-
  ceop_no_outliers |>
  drop_na() |>
  group_by(ceo_name, ceo_id) |>
  summarise( #to aggregate all variables to CEO-firm level like the pdf said
    mean_ceo_pay = mean(ceo_total_compensation),
    mean_ceo_age = mean(ceo_age),
    mean_ceo_tenure = mean(ceo_tenure),
    mean_ceo_roe = mean(comp_perf_roe),
    mean_ceo_mktval = mean(mktval)
  )
```

```
# Create the recipe
rec <- recipe(~ ., data = ceop_groups |>
                select(mean_ceo_pay, mean_ceo_roe)) |>
        step_normalize(all_numeric_predictors())

# Prepare and apply the recipe
fin_scaled <- prep(rec) |>
             bake(new_data = NULL)

fin_scaled <-
  fin_scaled |>
  drop_na() |>
 ungroup() |>
   select(mean_ceo_pay, mean_ceo_roe) # include only the numerical columns to make the
```

```
#Create the recipe
rec2 <- recipe(~ ., data = ceop_groups |>
                select(mean_ceo_age, mean_ceo_tenure)) |>
        step_normalize(all_numeric_predictors())

# Prepare and apply the recipe
char_scaled <- prep(rec2) |>
             bake(new_data = NULL)

#Do the same event as before but focus on CEO characteristics instead?
char_scaled <-
  char_scaled |>
  drop_na() |>
 ungroup() |>
   select(mean_ceo_age, mean_ceo_tenure)
```

## Data Analysis

Perform your data analysis in this section. Insert code chunks and comments / explanations as necessary.

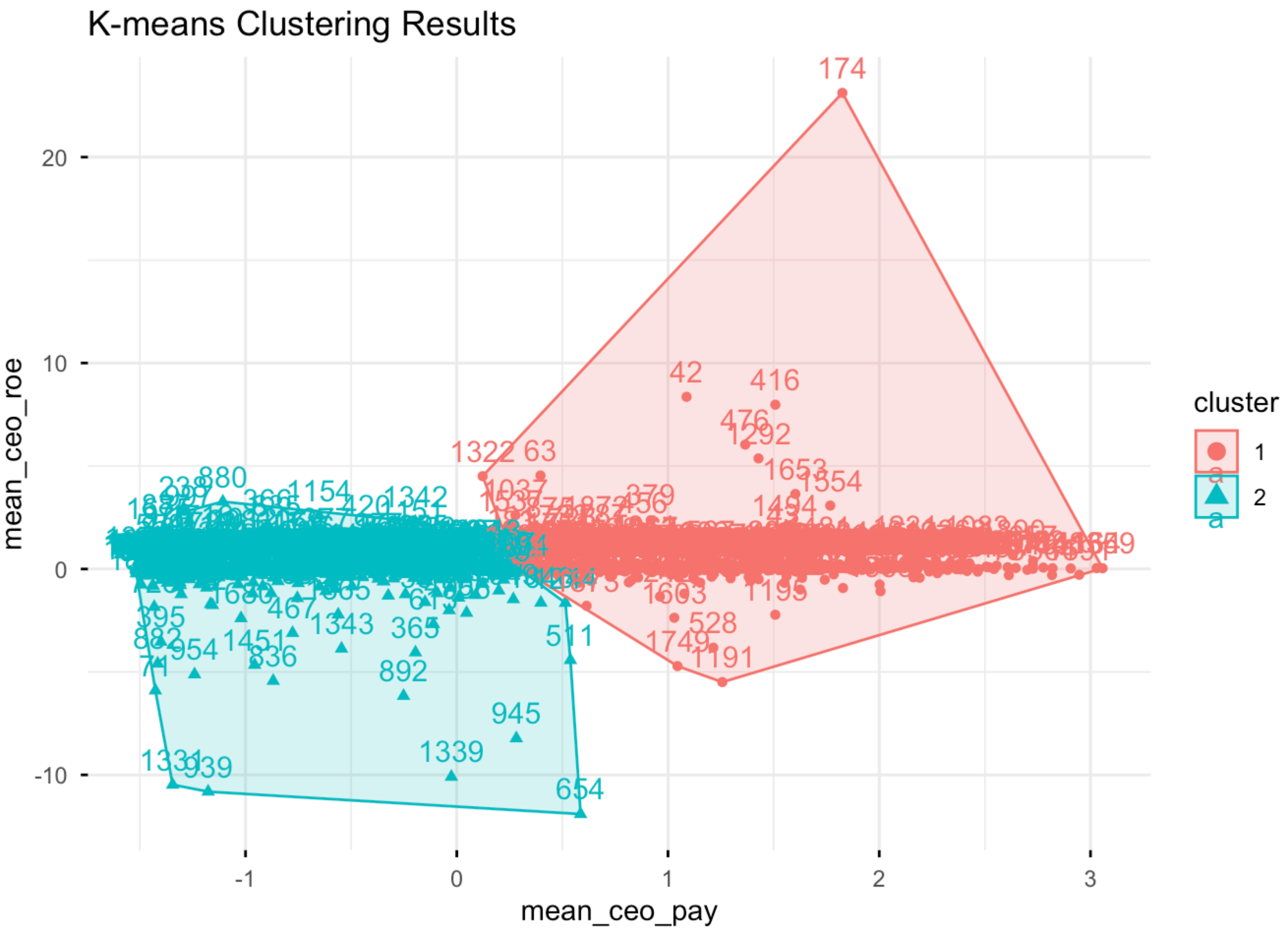
###FOR FINANCIAL

```
glimpse (fin_scaled) # check if data frame is scaled
```

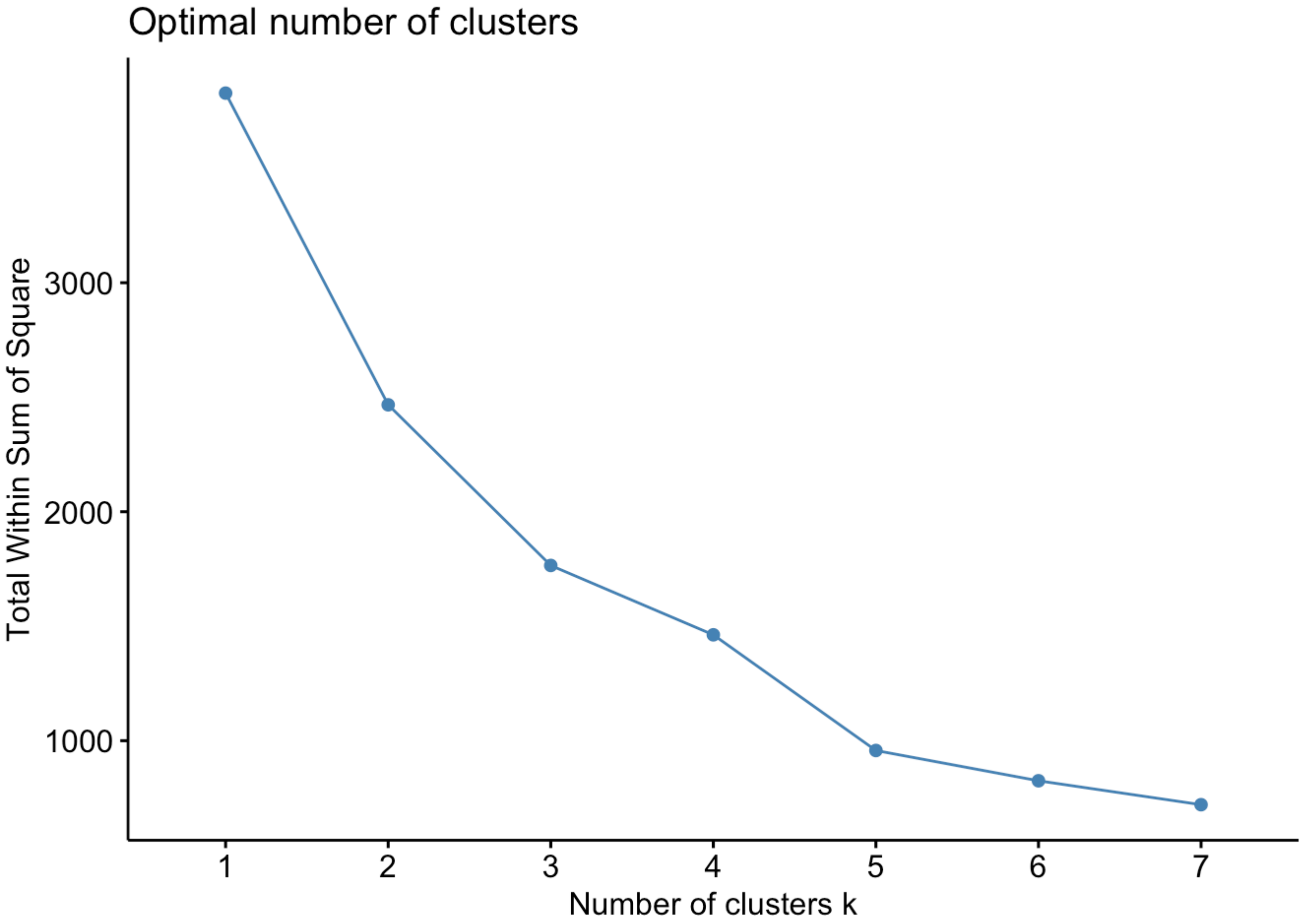
```
Rows: 1,915
Columns: 2
$ mean_ceo_pay <dbl> -1.48389521, -0.43283639, -0.42017880, -1.15643644, -0.30…
$ mean_ceo_roe <dbl> 0.14725801, 0.04499825, -0.05437027, -1.75540885, 0.09754…
```

```
# Now run K-means
set.seed(123)
kmeans_model <-
  kmeans(fin_scaled,
         centers = 2,
         nstart = 5) #generate K-means model based on financial characteristics
```

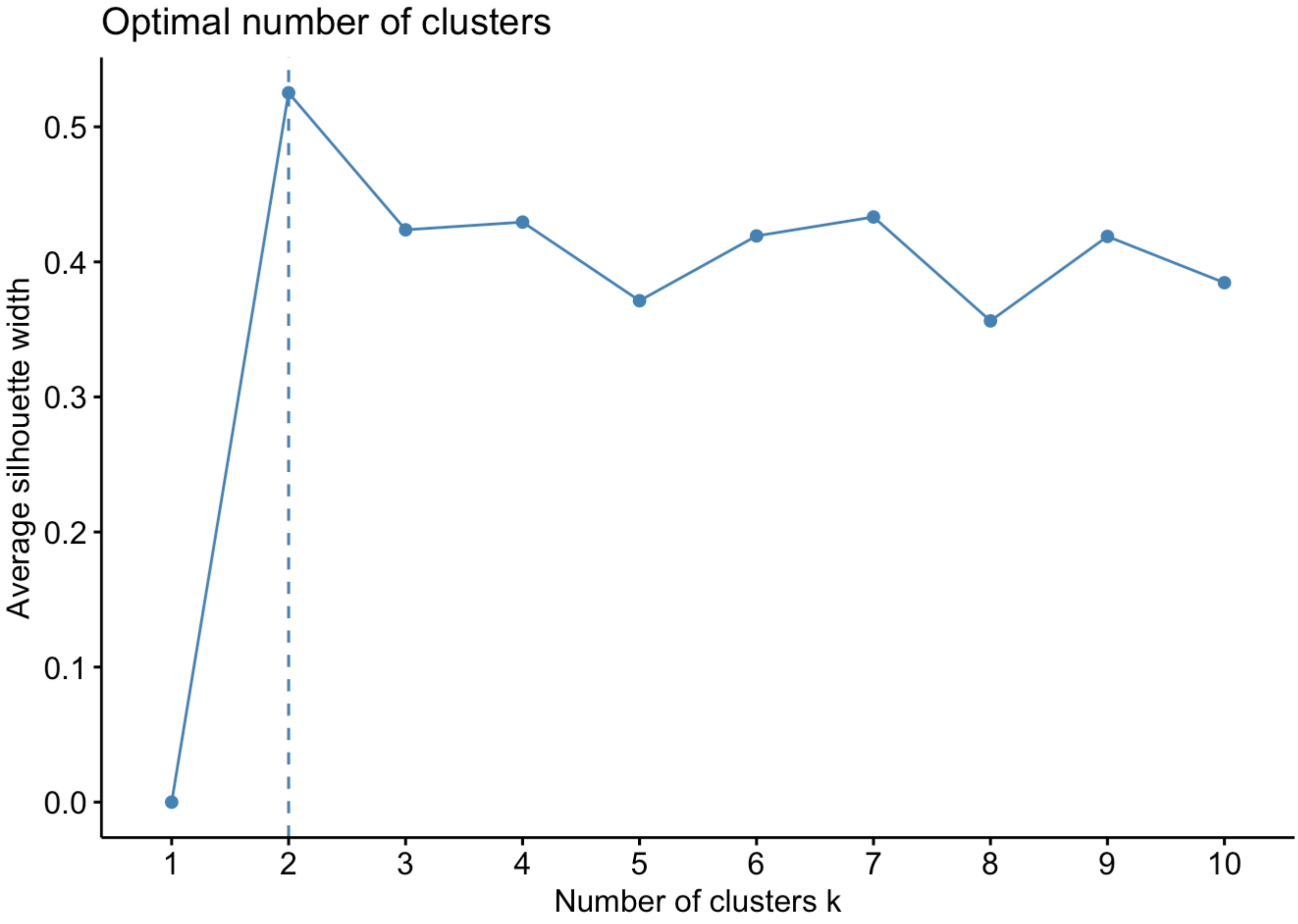
```
fin_cluster_plot <- fviz_cluster(
  kmeans_model,
  data = fin_scaled,
  ggtheme = theme_minimal(),
  main = "K-means Clustering Results") #plot the Financial clusters

ggsave(
  filename = "figs/fin_cluster_plot.png",
  plot = fin_cluster_plot,
  width = 8, height = 6, dpi = 300)

fin_cluster_plot
```



```
fviz_nbclust (
  fin_scaled,
  kmeans,
  method = "wss",
  k.max = 7,
  nstart = 10
) # use elbow method to check number of clusters
```



```
 fviz_nbclust(
  fin_scaled,
  kmeans,
  method = "silhouette"
) # use silhoutte method to cross check optimal number of clusters
```



```
fin_clustered <-
  ceop_groups |>
  select(mean_ceo_pay, mean_ceo_roe) # select only pay and ROE

fin_clustered <- augment(kmeans_model, fin_clustered)

fin_clustered |>
    group_by(.cluster) |>
    summarise(across(mean_ceo_pay:mean_ceo_roe, mean))
```

```
# A tibble: 2 × 3
  .cluster mean_ceo_pay mean_ceo_roe
  <fct>           <dbl>        <dbl>
1 1              13766.       0.314
2 2               4726.      -0.0568
```

```
#Two/ Three Clusters form

#Cluster 1 has higher pay and higher ROE
#Cluster 2 has lower ROE and lower pay
```

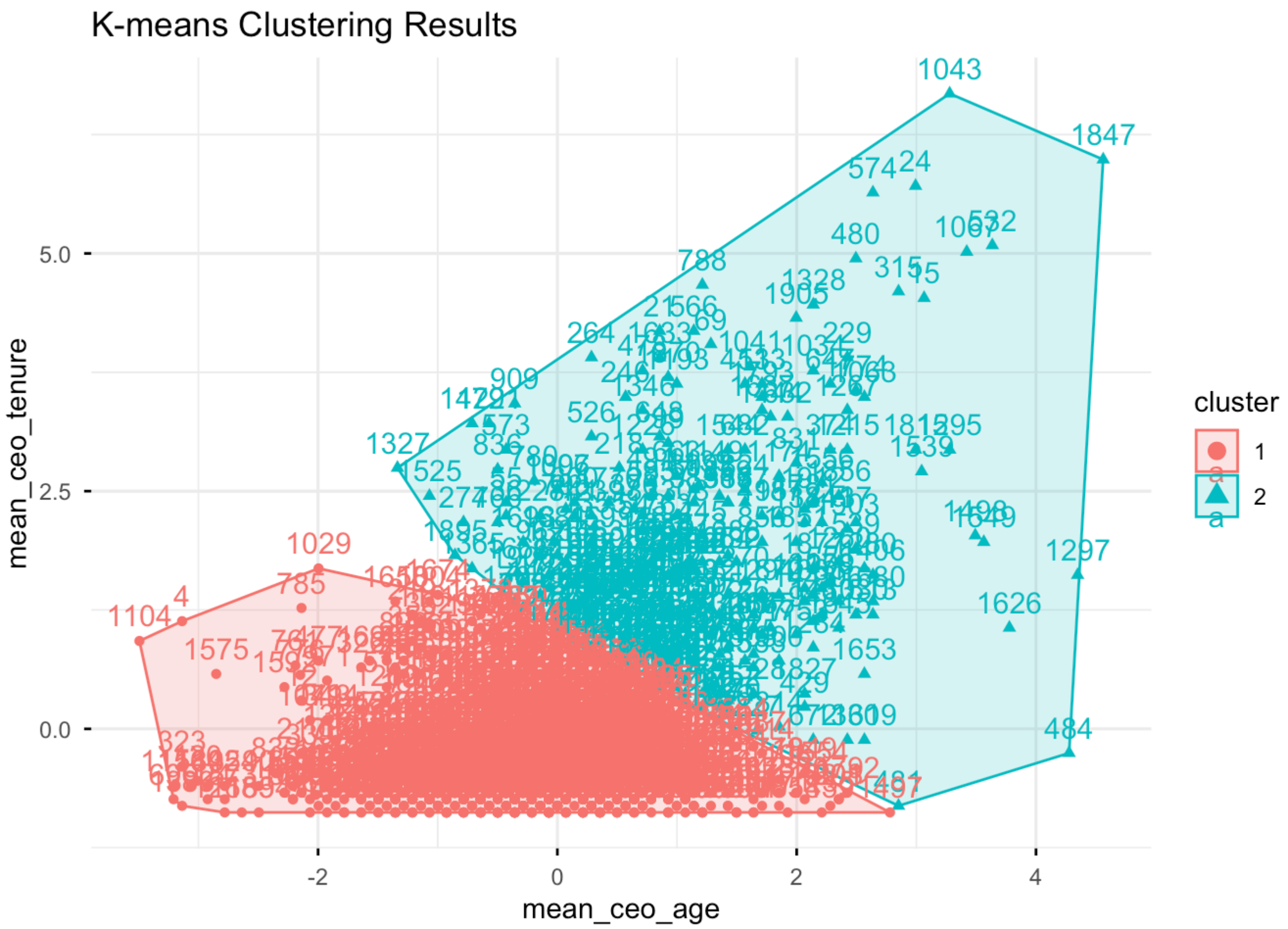
## FOR CEO CHARACTERISTICS

```
glimpse (char_scaled) # check if everything is scaled
```

```
Rows: 1,915
Columns: 2
$ mean_ceo_age    <dbl> -1.1711264334, -0.8554631831, -1.8533663614, -3.136384…
$ mean_ceo_tenure <dbl> 0.27057866, -0.53193824, -0.53193824, 1.13254126, 0.71…
```

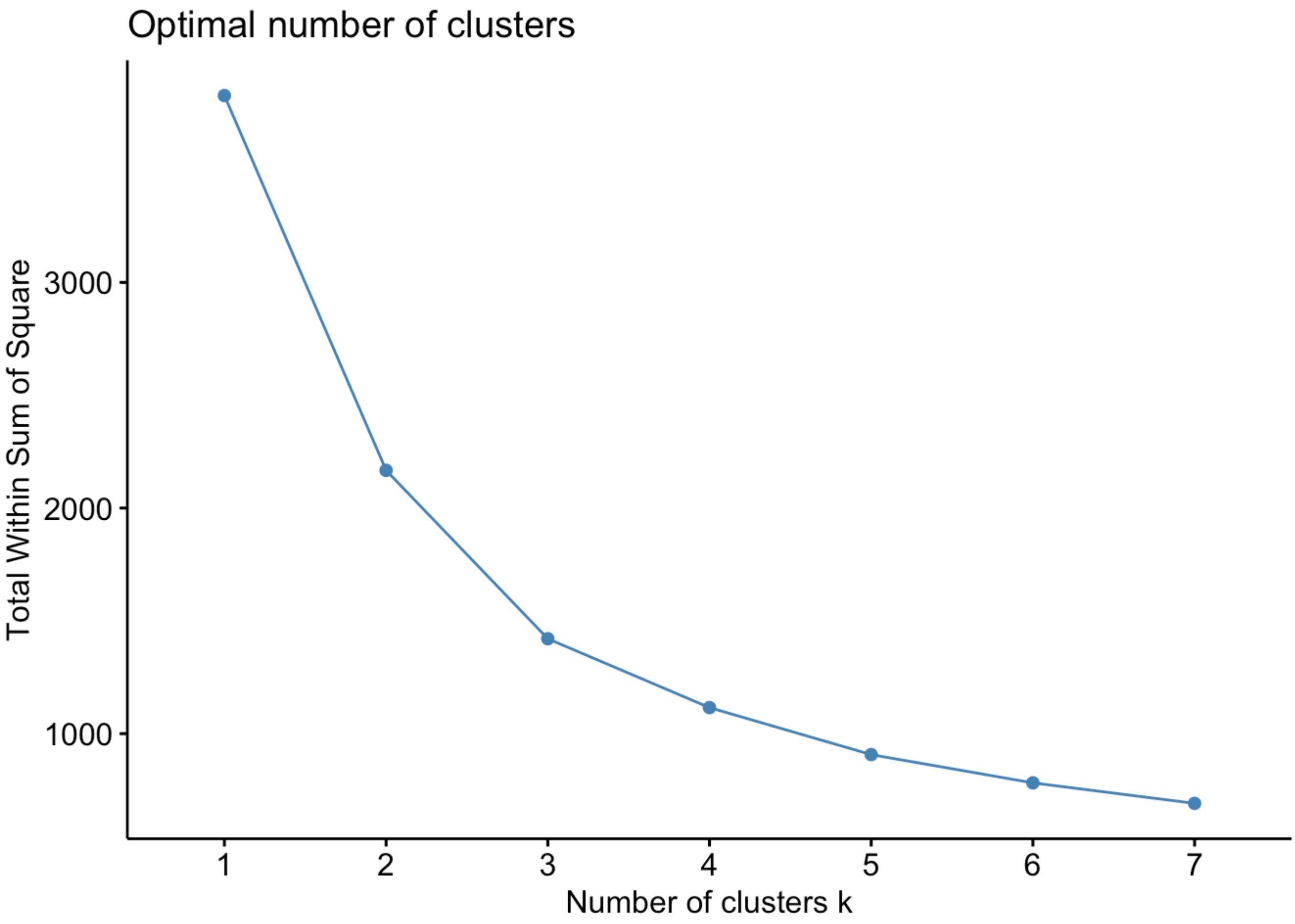
```
# Now run K-means
set.seed(1544)
kmeans_model <-
  kmeans(char_scaled,
         centers = 2,
         nstart = 6) # cluster again based on CEO characteristics
```

```
fviz_cluster(
  kmeans_model,
  data = char_scaled,
  ggtheme = theme_minimal(),
  main = "K-means Clustering Results"
) # graph CEO characteristics K-Means cluster model
```

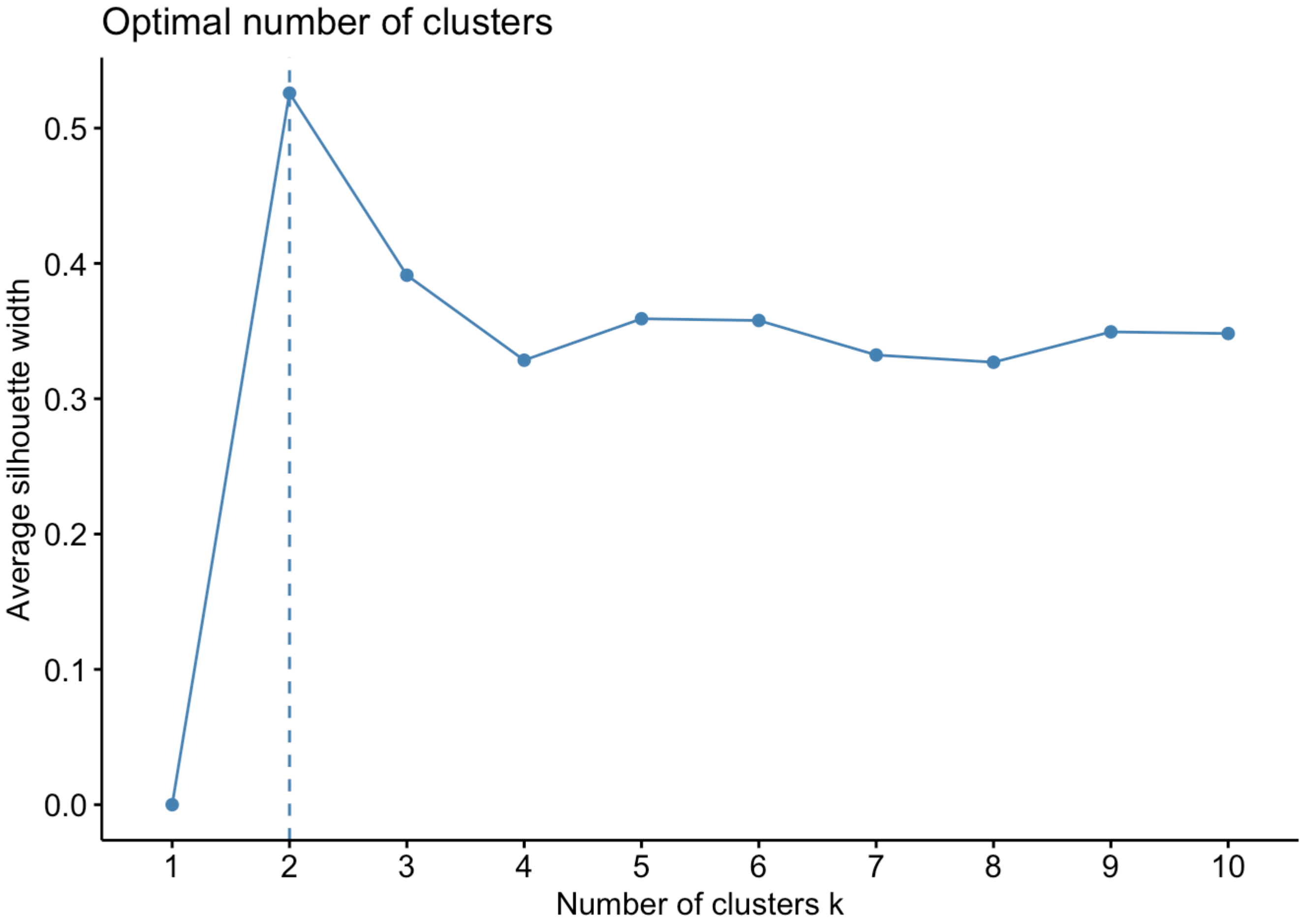


```
char_cluster_plot <- fviz_cluster(
  kmeans_model,
  data = char_scaled,
  ggtheme = theme_minimal(),
  main = "K-means Clustering Results") #save plot as char_cluster_plot

ggsave(
  filename = "figs/char_cluster_plot.png", #save plot as image
  plot = char_cluster_plot,
  width = 8, height = 6, dpi = 300)
```

```
fviz_nbclust (
  char_scaled,
  kmeans,
  method = "wss",
  k.max = 7,
  nstart = 10 #cross check optimal number of cluster
)
```



```
 fviz_nbclust(
  char_scaled,
  kmeans,
  method = "silhouette"
) #cross check optimal number of cluster
```



```
char_clustered <-
  ceop_groups |>
  select(mean_ceo_pay, mean_ceo_age, mean_ceo_tenure)

char_clustered <- augment(kmeans_model, char_clustered)

char_clustered |>
    group_by(.cluster) |>
    summarise(across(mean_ceo_pay:mean_ceo_tenure, mean))
```

```
# A tibble: 2 × 4
  .cluster mean_ceo_pay mean_ceo_age mean_ceo_tenure
  <fct>           <dbl>        <dbl>           <dbl>
1 1               7946.         55.9            3.76
2 2               6789.         65.3           18.7
```

```
#Cluster 1 higher pay but younger and shorter tenure
#Cluster 2 lower pay but older and longer tenure

#Clusters reveal that when paying CEOs, tenure and age should not matter
```

## Part B. Analytics Team Memo (20 points)

Instructions

**Length:** 600 words. **Exhibits:** Maximum 2 figures and/or tables.

Write this as a bullet-point working document for your analytics peers and manager.

Bullets must be written as full sentences. Use Writing Tips from “Write Like an Amazonian”. Only display code in memo if it is to define or explain. Do not produce any new analysis, table or figure in this section.

# Memo

PLACE EXHIBITS IN APPROPRIATE SECTION AS NEEDED

## 1. Research Question

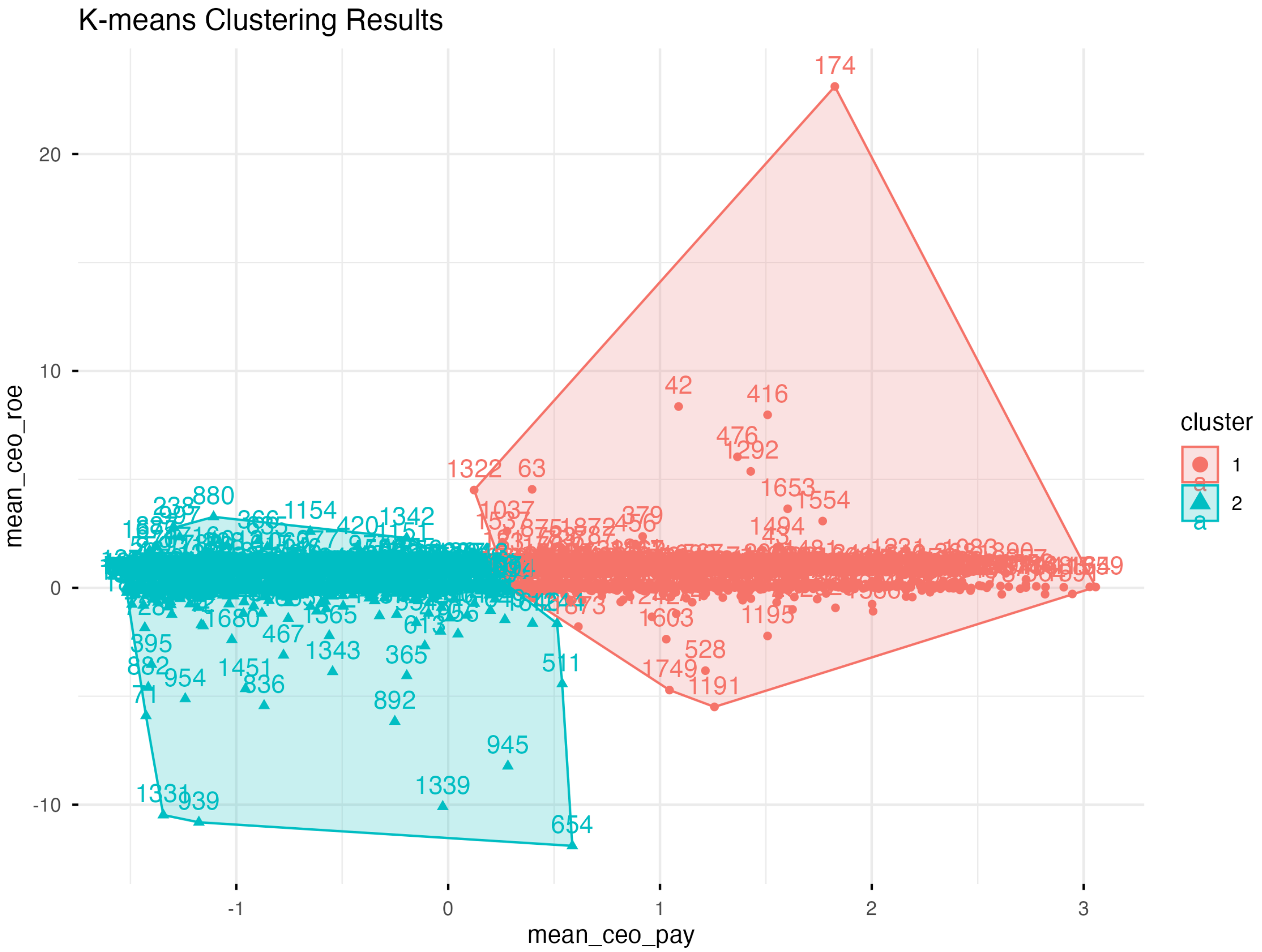
What CEO types emerge from the data when we perform clustering analysis based on CEO characteristics and/or firm characteristics? How can these CEO types inform our clients’ decisions on how much to pay their CEOs?

-

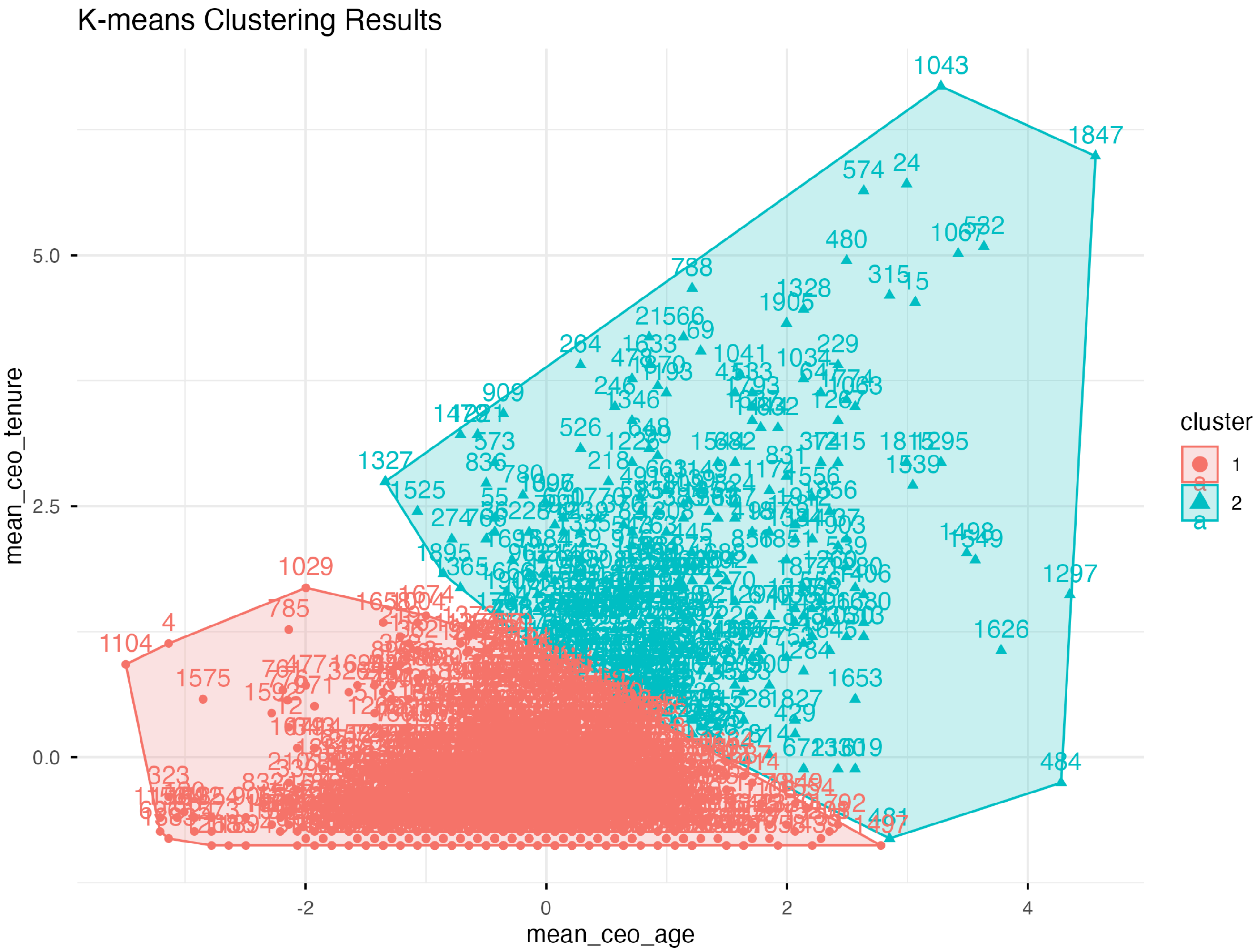
## 2. Data & Approach (5 pts)

- Dataset: The ceo_pay dataset includes executive-level and firm-level information which outline both performance and characteristics of both firms and CEOs. Key pieces of data in the set include CEO total compensation, age, tenure, ROE, profit, sales, and market value.
- Key variables we used in our approach were ceo_total_compensation, CEO_age, and CEO_tenure to evaluate CEO characteristics, while we used comp_perf_roe and profit_margin as our main financial performance indicators.
- We implemented descriptive analysis using K-means clustering to uncover groups of CEOs based on personal characteristics (CEO_age and CEO_tenure) as well as financial features (ceo_total_compensation, comp_perf_roe, profit_margin). Variables were cleaned, winsorised to remove outliers, normalised via the recipes package, and then clustered.
-
- Explain dataset used, key variables, and methods.
- Be concise but clear.

## 3. Preliminary Findings (6 pts)



- Two distinct financial-based clusters emerged when analysing CEO pay: high performance high pay, and low performance low pay.
- In the high performance, high pay cluster, CEOs lead firms with higher profitability recording a mean ROE of 31.4%, this is correlated with relatively higher pay, with a mean of 13766.25.
- In the low performance, low pay cluster, CEOs lead firms with weaker profitability, recording a mean ROE of -5.7% which is reflected in their lower levels of compensation.
- This clustering supports the idea that boards and compensation committees tie pay closely with financial performance rather than arbitrary characteristics. This is clearly reflected in the apparent correlation between pay and return on equity. Given that, of all clusters tested, ROE showed the clearest clustering, it can also be inferred that the use of shareholders’ money is viewed as the most important measure of performance for businesses.



- The second pair of clusters that emerged was based on CEO characteristics, namely age and tenure. - In this, the cluster of younger, less-tenured CEOs tended to earn more than older, longer-tenured CEOs. - This apparent pattern of younger CEOs earning greater compensation may stem from younger CEOs being able to more aggressively target long-term growth strategies (Han & Jo, 2024) because they have a long career ahead of them.

## 4. Assumptions & Limitations (3 pts)

Assumptions: - We assumed NAs and outliers are not useful for the analysis and so are excluded.

- We assumed that mean is the best way to aggregate year by year is to take the average.
- We assumed market valuation was insignificant in determining CEO pay.

Limitations: - Limitation in that the analysis is descriptive, not causal, clustering does not establish direction or significance of effects, only correlation.

- Omitted variables such as gender and industry since they were not numeric.

## 5. Open Questions for Feedback (3 pts)

- List up to 3 specific questions where you'd seek peer/manager input.
- What additional performance indicators would be optimal to improve segmentation accuracy of clusters?

- Would it be more optimal to cluster independently based on each industry given differing priorities between industries?
- What other data visualisation techniques would strengthen the interpretability of results?

## 6. Next Steps (3 pts)

- Outline what you would test or refine next if you were to keep working on the project.
- No more than 3 points.
- …

In future iterations we would like to: - integrate more data to evaluate CEO performance such as share price charts and performance ratings. -test the viability of companies that are using performance to inform compensation through causal analytics (e.g. assessing whether bonuses will drive performance). - analyse data on an industry-by-industry basis to tailor advice and feedback for managers of different industries.

# Part C. Corporate Brief (20 points)

**Instructions**

**Length:** 400 words. **Exhibit:** 1 table OR figure.

Write this as an executive brief to be read and understood by managers.

Bullets must be written as full sentences. Use Writing Tips from "Write Like an Amazonian". No visible code in the brief. Do not produce any new analysis, table or figure in this section.

INSERT YOUR FIGURE OR TABLE HERE

## Executive Summary (5 pts)

3–4 sentences summarising the research question, key findings and why it matters.

-We aimed to answer whether there were any meaningful segments (clusters) that could be formed from this data and how can these segments be used to inform future decisions of clients. -Our key findings were that, based on the K-means clustering algorithm, the key groups were: (1) high performing and highly paid, (2) low performing and lower paid, (3) highly paid younger CEOs, and (4) lowly paid older CEOs. -We believe that these clusters are instrumental in informing you of past patterns on CEO pay, specifically why and how they came to be, allowing you to make better decisions on how you want to pay your CEOs in the future.

## Key Insights (5 pts)

- The financial clustering cleanly separates CEOs into a high-performance/high-pay group and a low-performance/low-pay group, potentially indicating that compensation in the data is in line with performance.
- The characteristics cluster groups CEOs by age and tenure, with younger, shorter tenured CEOs garnering higher pay on average, possibly due to a mix of growing firms seeking bright young leaders as well as biases from performance indicators. -Generally, CEO compensation packages are developed based on metrics measuring returns and are positively correlated with firm

performance (Kweh et al., 2022) -In South Korea, young CEOs are more likely to take risks that lead to long-term firm performance compared to older CEOs who prefer stable financial performance (Han & Jo, 2024).

### Business Implications (5 pts)

- Tying pay to performance is both observable in the market and justifiable to shareholders, reducing risk of overpaying poor leadership and creating a quantifiable anchor to dictate compensation.
- The data suggests some reference to age and tenure to guide CEO compensation, however this risks misalignment with value creation, potentially weakening incentives and eroding investor confidence.
- Using the performance-based clusters as external reference points appears to be the logical route of justifying compensation to the board, as it supports disclosure narratives and improves benchmarking rigor.

### Recommended Actions (5 pts)

- Establish compensation bands anchored to objective performance metrics, using the ROE-pay clusters to inform the expansion into the use of further indicators to inform compensation.
- Calibrate a balance between fixed and variable pay so that bonuses paid move materially with performance to maintain the connection between pay and performance.
- Move away from demographic-based hiring and pay, instead stressing the performance-first nature of compensation to build rapport and strengthen governance.

## References

-Han, J., & Jo, S. J. (2024). How Much Does the CEO's Age Impact Corporate Performance Under a Changing Environment? Administrative Sciences, 14(11), 304. https://doi.org/10.3390/admsci14110304

-Kweh, Q. L., Tebourbi, I., Lo, H., & Huang, C. (2022). CEO compensation and firm performance: Evidence from financially constrained firms. Research in International Business and Finance, 61, 101671. 10.1016/j.ribaf.2022.101671

**Instructions:**

- List any sources, readings, or external material you cited.
- Use a consistent referencing style (APA).

## GenAI Use Declaration

**How to Reference GenAI Use**

If you used GenAI (e.g., ChatGPT, Copilot) for any part of your work, you must acknowledge it in **References** or a footnote. Examples:

- *Text generation*:
  "Some passages in the Analytics Team Memo were drafted with the assistance of ChatGPT (OpenAI,

> 2025), reviewed and edited by the authors.”
>
> - *Code assistance*:
>   “Code snippets for data cleaning were adapted from suggestions generated by ChatGPT (OpenAI, 2025).”
> - *Idea generation*:
>   “Research question ideas were brainstormed with assistance from ChatGPT (OpenAI, 2025).”
>
> Always make clear: **what was generated, and how you verified it.**

Please tick the boxes that apply by replacing the blank spaces between parentheses with an X.

- ☑ **No GenAI tools used.**
- ☑ **GenAI used for grammar or style editing** (e.g., spelling/grammar suggestions).
- ☑ **GenAI used for code assistance** (e.g., debugging, syntax help).

Note that all other forms of GenAI use in this group assignment 2 are **NOT** allowed.

**If GenAI was used:**
- I have **cited or referenced** it appropriately in my submission.
- I confirm that I have **reviewed, verified, and taken responsibility** for all content submitted.
- I understand that **copy-pasting unedited GenAI output is not acceptable**.
- Copy-pasting GenAI output, even with a citation, may lead to a finding of **poor academic performance or an academic misconduct case** against **each member of the group**.
- Any use of GenAI must be **substantially adapted, verified, and integrated** into our group’s own work.

**Declaration:**
By submitting this assignment for grading, we declare that this work represents our own group effort, and any use of GenAI tools has been transparently disclosed and appropriately referenced.