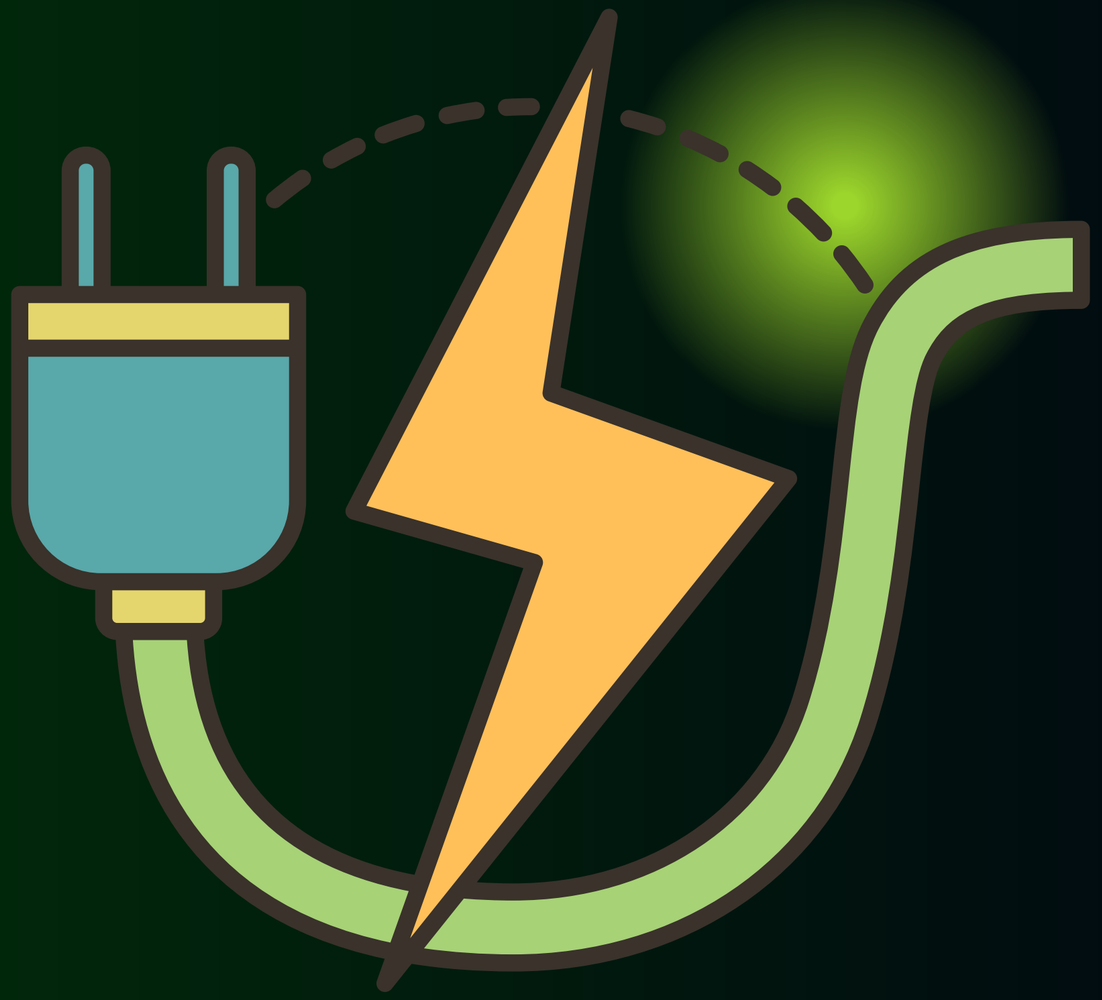
NARRATED PRESENTATION:

https://video.kent.edu/media/BA64060%20Final%20Project%20Presentation%20Group%203/1_yxsltuiw

BA64060 - GROUP 3
FINAL PROJECT

UNCOVERING PATTERNS IN U.S. POWER PLANT GENERATION & FUEL USE



MARVELLE HORTON
ALEXIS MCCARTNEY
KRISTEN DURKIN

WHAT WE WANTED TO UNDERSTAND

01 What operational patterns appear when we cluster U.S. power plants using the 2023 year-to-date generation and fuel data?

02 What characteristics define each cluster, and what do these differences suggest about plant behavior, performance and efficiency?

03 What groups of power plants naturally form on how much fuel they use and how much electricity they generate?

04 Are there meaningful differences in efficiency between these groups?

HOW WE APPROACHED THE ANALYSIS

Raw Data → Clean & prepare → Engineer Key Metric → Scale Variables → Test Cluster Options → Final Cluster & Insights

01

```
```{r}
df <- read.csv("Group3Data.csv") #read csv file
View(df) #Visual Check
```
```

02

```
```{r}
colSums(is.na(df)) #Count N/A values in each column
```
```

```
```{r}
num_cols <- ncol(df)   #total columns
ytd <- df[, (num_cols-5):num_cols] #grab last 6 columns (YTD data)
head(ytd)
colSums(is.na(ytd)) #double check for missing values
```
```

03

```
# I compute total within-cluster sum of squares (WSS) for k = 1 to 10.
# The "elbow" is the point where adding more clusters stops giving
# big decreases in WSS.
wss <- sapply(1:10, function(k) {
  kmeans(train, centers = k, nstart = 20)$tot.withinss
})

# Plot WSS vs k so I can visually inspect the elbow.
plot(
  1:10, wss, type = "b",
  xlab = "Number of Clusters (k)",
  ylab = "Total Within-Cluster Sum of Squares (WSS)",
  main = "Elbow Method for Choosing k"
)
```

04

```
# Create a new engineered feature:
# efficiency_mmbtu_per_mwh = total_fuel_mmbtu / net_gen_mwh
# This measures how many MMBtu of fuel are used per MWh produced.
df_2023 <- df_2023 %>%
  mutate(efficiency_mmbtu_per_mwh = total_fuel_mmbtu / net_gen_mwh)

# Quick sanity checks
head(df_2023)
str(df_2023)
summary(df_2023$efficiency_mmbtu_per_mwh)
```

05

```
# I run k-means on the training data using k = 3 clusters and
# multiple random starts (nstart = 25) for stability.
set.seed(123)
km_final <- kmeans(train, centers = k_opt, nstart = 25)

# Cluster sizes show how many units fall into each group.
km_final$size

# Cluster centers summarize the average standardized profile for each cluster.
km_final$centers
```

WHAT THE DATA REVEALED

CLUSTER 1

Low Intensity + Electricity Generation:

- LARGEST group
- Low fuel use
- Low electricity output
- Often backup or rarely operating units

CLUSTER 2

Moderate Intensity + Fuel Use:

- Medium-sized group
- Balanced electricity production
- Stable performance, but not the biggest contributor to electricity generation

CLUSTER 3

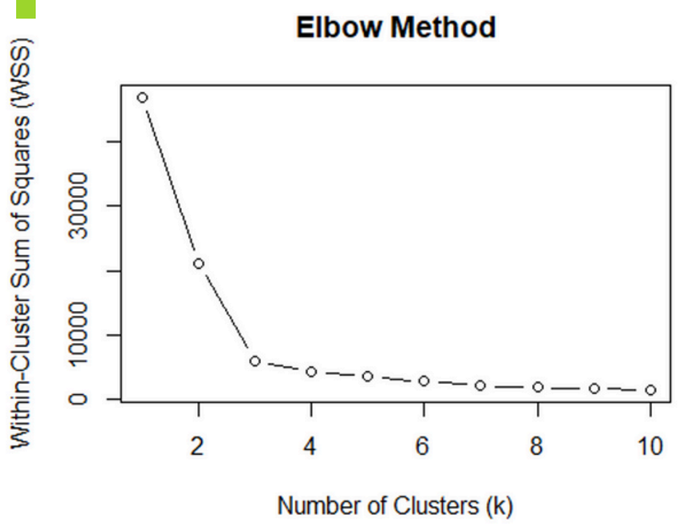
High Intensity, Fuel Use + Efficiency

- Smallest group by count
- Generates the most electricity
- Uses fuel the most efficiently
- Significant representation of U.S. power generation

```
219  cluster_summary <- train_original %>%
220    group_by(cluster) %>%
221    summarise(
222      n_units = n(),
223      avg_total_fuel_mmbtu = mean(total_fuel_mmbtu, na.rm = TRUE),
224      avg_net_gen_mwh = mean(net_gen_mwh, na.rm = TRUE),
225      avg_efficiency_mmbtu_mwh = mean(efficiency_mmbtu_per_mwh, na.rm =
226      .groups = "drop"
227    )
228
229  cluster_summary
```

HOW WE CHOSE THE NUMBER OF CLUSTERS

01



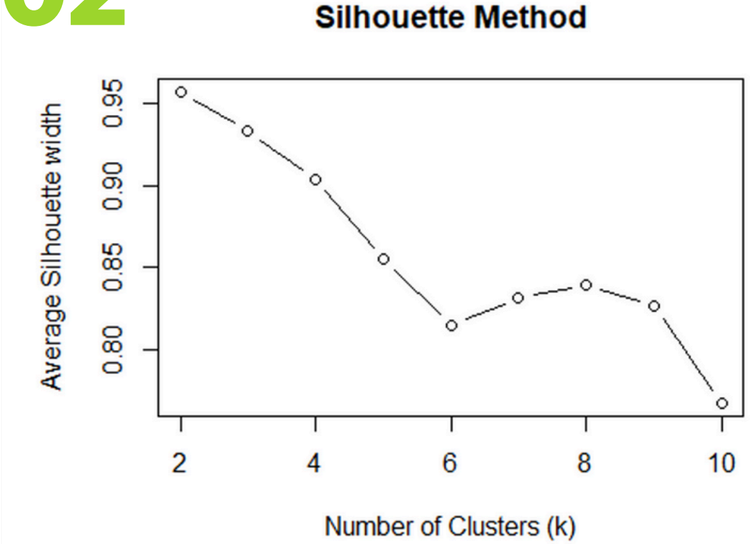
#Look where elbow - this shows when adding more would not be benefical

Results:

Cluster 1 = very high WSS with everything lumped together. k = 2 = largest change. k = 3 = “elbow” point.

Using the elbow method,a clear bend at k = 3 can be seen. While the WSS decreases sharply from k = 1 to k = 3, the reduction beyond k = 3 is minimal. This indicates that 3 clusters provide an ideal balance.

02



#Higher values mean more defined clusters

Confirmation k = 3

Based on the elbow method, it’s observed to have a clear bend at k = 3, indicating diminishing returns beyond this point. Although the silhouette method peaked at k = 2, the value for k = 3 remained high, suggesting well separated clusters.

For this reason we selected k = 3 as the optimal number of clusters.

WHAT THESE CLUSTERS TELL US

CLUSTER 1

Low Intensity + Electricity Generation:

- Make up the majority of U.S. Units
- Low fuel use + electricity output
- Likely, backup units, niche operations, or peaker plants
- Don’t contribute to the nation’s grid significantly

CLUSTER 2

Moderate Intensity + Fuel Use:

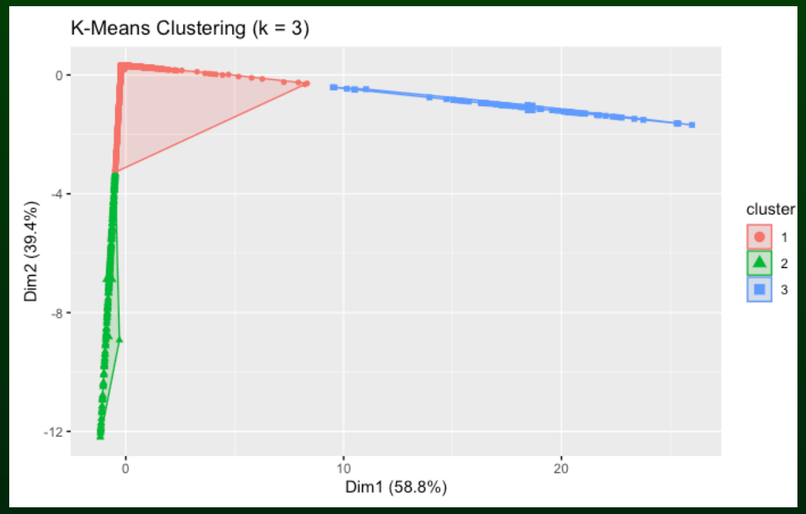
- Produce a stable, mid-level amount of electricity
- Operate consistently but not at a huge scale
- Regional plants that meet predictable demand

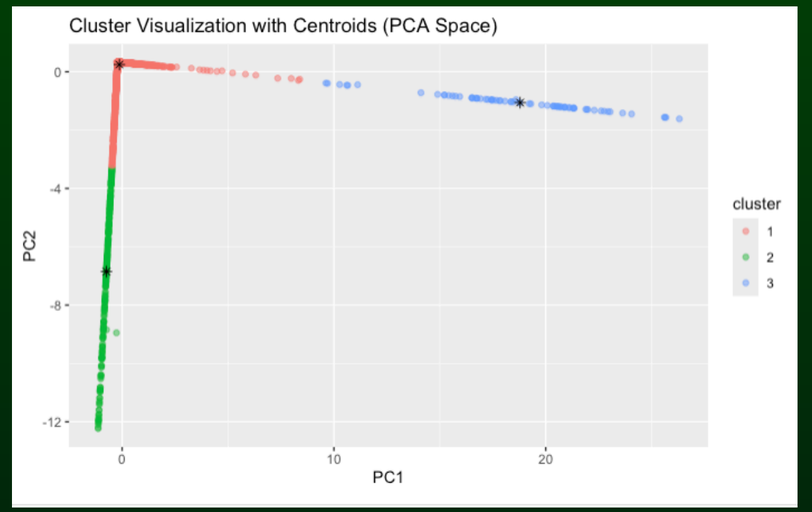
CLUSTER 3

High Intensity, Fuel Use + Efficiency

- Smallest group by count
- Generates the most electricity in the dataset
- Highest operational efficiency (best fuel to electricity conversion)
- These are the “workout” plants running the U.S. power grid

Clusters & Centroids of Clusters





Based off of the cluster formations, one can assume that cluster 3 is working as a base load power plant. That is shown by its linear power consumption and output sitting at high levels.

```r
#Centroid points

centers_unscaled <- sweep(km_final$centers, 2, attr(ytd_scaled, "scaled:scale"), "*")
centers_unscaled <- sweep(centers_unscaled, 2, attr(ytd_scaled, "scaled:center"), "+")
centers_unscaled


km_final$centers


library(ggplot2)

# PCA on the training data
pca <- prcomp(train, scale = FALSE)
pca_scores <- as.data.frame(pca$x[,1:2])  # first two PCs
pca_scores$cluster <- as.factor(km_final$cluster)

# centroid points
centroids <- aggregate(pca_scores[,1:2], list(Cluster = pca_scores$cluster), mean)

# point plot
ggplot(pca_scores, aes(PC1, PC2, color = cluster)) +
  geom_point(alpha = 0.5) +
  geom_point(data = centroids, aes(PC1, PC2),
             color = "black", size = 2, shape = 8) +   # centroid markers
  labs(title = "Cluster Visualization with Centroids (PCA Space)")
```

WHY THIS SEGMENTATION MATTERS

IT SHOWS WHERE THE U.S. GETS ITS ELECTRICITY

Even thoigh thousands of plants exist, only a small handful produce most of the power. This can be a good use case for investors, grid operators and policymakers to understand where reliable electricity comes from.

IT HIGHLIGHTS EFFICIENCY GAPS ACROSS THE SYSTEM

We are now able to see which plants operate efficiently (Cluster 3), waste fuel (Cluster 1) and sit comfortably in the middle (Cluster 2)

PROVIDES A CLEARER PICTURE FOR ENERGY PLANNING + INVESTMENT

If a small group of plants carries the grid, those are the ones that are most critical to maintain, impacted by outages and most likely to justify mondernization investment.

SIMPLIFIES A LARGE DATASET INTO AN UNDERSTANDABLE STORY

Our clustering methods turned this large dataset into low producers, mid-range producers and high output, efficient producers. This kind of insight is what leadership stakeholders can use to make decisions.

THANK YOU!