Final Exam

Jacob Aylward

2025-06-27

```
library(readr)
data<- read_csv("C:/Users/jacob/Downloads/out_eia923__fuel_receipts_costs-2.csv")
```

```
## Rows: 696244 Columns: 36
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr  (25): plant_name_eia, utility_name_eia, state, contract_type_code, en
er...
## dbl   (9): plant_id_eia, plant_id_pudl, utility_id_eia, utility_id_pudl, f
ue...
## dttm  (2): report_date, contract_expiration_date
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this m
essage.
```

```
View(data)
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.
0.0 ──
## ✓ dplyr     1.1.4     ✓ purrr     1.0.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://g
oo.gl/ve3WBa
```

```
Data <- na.omit(data)
str(Data)
```

```
## tibble [696,244 × 36] (S3: tbl_df/tbl/data.frame)
##  $ report_date                            : POSIXct[1:696244], format: "20
08-01-01" "2008-01-01" ...
##  $ plant_id_eia                           : num [1:696244] 3 3 3 7 7 7 7 8
8 8 ...
##  $ plant_id_pudl                          : num [1:696244] 32 32 32 207 20
7 207 207 231 231 231 ...
##  $ plant_name_eia                         : chr [1:696244] "Barry" "Barry"
"Barry" "Gadsden" ...
##  $ utility_id_eia                         : num [1:696244] 195 195 195 195
195 195 195 195 195 195 ...
##  $ utility_id_pudl                        : num [1:696244] 18 18 18 18 18
18 18 18 18 18 ...
##  $ utility_name_eia                       : chr [1:696244] "Alabama Power
Co" "Alabama Power Co" "Alabama Power Co" "Alabama Power Co" ...
##  $ state                                  : chr [1:696244] "AL" "AL" "AL"
"AL" ...
##  $ contract_type_code                     : chr [1:696244] "C" "C" "C" "C"
...
##  $ contract_expiration_date               : POSIXct[1:696244], format: "20
08-04-01" "2008-04-01" ...
##  $ energy_source_code                     : chr [1:696244] "BIT" "BIT" "NG
" "BIT" ...
##  $ fuel_type_code_pudl                    : chr [1:696244] "coal" "coal" "
gas" "coal" ...
##  $ fuel_group_code                        : chr [1:696244] "coal" "coal" "
natural_gas" "coal" ...
##  $ supplier_name                          : chr [1:696244] "interocean coa
l" "interocean coal" "bay gas pipeline" "alabama coal" ...
##  $ fuel_received_units                    : num [1:696244] 259412 52241 27
83619 25397 764 ...
##  $ fuel_mmbtu_per_unit                    : num [1:696244] 23.1 22.8 1.04
24.61 24.45 ...
##  $ fuel_cost_per_mmbtu                    : chr [1:696244] "2.134999990463
257" "2.1150000095367 43" "8.630999565124512" "2.7760000228881836" ...
##  $ bulk_agg_fuel_cost_per_mmbtu           : chr [1:696244] "null" "null" "
8.603500366210938" "null" ...
##  $ fuel_consumed_mmbtu                    : num [1:696244] 5992418 1191095
2892180 625020 18677 ...
##  $ total_fuel_cost                        : chr [1:696244] "12793811" "251
9165.5" "24962406" "1735056" ...
##  $ fuel_cost_per_mmbtu_source             : chr [1:696244] "original" "ori
ginal" "original" "original" ...
##  $ sulfur_content_pct                     : num [1:696244] 0.49 0.48 0 1.6
9 0.84 ...
##  $ ash_content_pct                        : num [1:696244] 5.4 5.7 0 14.7
15.5 ...
##  $ mercury_content_ppm                    : chr [1:696244] "null" "null" "
null" "null" ...
##  $ primary_transportation_mode_code       : chr [1:696244] "RV" "RV" "PL"
```

```
"TR" ...
##  $ secondary_transportation_mode_code     : chr [1:696244] "null" "null" "
null" "null" ...
##  $ natural_gas_transport_code             : chr [1:696244] "firm" "firm" "
firm" "firm" ...
##  $ natural_gas_delivery_contract_type_code: chr [1:696244] "null" "null" "
null" "null" ...
##  $ moisture_content_pct                   : chr [1:696244] "null" "null" "
null" "null" ...
##  $ chlorine_content_ppm                   : chr [1:696244] "null" "null" "
null" "null" ...
##  $ data_maturity                          : chr [1:696244] "final" "final"
"final" "final" ...
##  $ mine_id_msha                           : chr [1:696244] "null" "null" "
null" "null" ...
##  $ mine_name                              : chr [1:696244] "mina pribbenow
" "mina pribbenow" "null" "alabama coal" ...
##  $ mine_state                             : chr [1:696244] "COL" "COL" "nu
ll" "AL" ...
##  $ coalmine_county_id_fips                : chr [1:696244] "null" "null" "
null" "01007" ...
##  $ mine_type_code                         : chr [1:696244] "SU" "SU" "null
" "SU" ...
```

```
Data_New <- Data[, !(names(Data) %in%c("mercury_content_ppm","natural_gas_del
ivery_contract_type_code","moisture_content_pct","chlorine_content_ppm","mine
_id_msha"))]
str(Data_New)
```

```
## tibble [696,244 × 31] (S3: tbl_df/tbl/data.frame)
##  $ report_date                      : POSIXct[1:696244], format: "2008-01
-01" "2008-01-01" ...
##  $ plant_id_eia                     : num [1:696244] 3 3 3 7 7 7 7 8 8 8
...
##  $ plant_id_pudl                    : num [1:696244] 32 32 32 207 207 207
207 231 231 231 ...
##  $ plant_name_eia                   : chr [1:696244] "Barry" "Barry" "Bar
ry" "Gadsden" ...
##  $ utility_id_eia                   : num [1:696244] 195 195 195 195 195
195 195 195 195 195 ...
##  $ utility_id_pudl                  : num [1:696244] 18 18 18 18 18 18 18
18 18 18 ...
##  $ utility_name_eia                 : chr [1:696244] "Alabama Power Co" "
Alabama Power Co" "Alabama Power Co" "Alabama Power Co" ...
##  $ state                            : chr [1:696244] "AL" "AL" "AL" "AL"
...
##  $ contract_type_code               : chr [1:696244] "C" "C" "C" "C" ...
##  $ contract_expiration_date         : POSIXct[1:696244], format: "2008-04
-01" "2008-04-01" ...
##  $ energy_source_code               : chr [1:696244] "BIT" "BIT" "NG" "BI
```

```
T" ...
##  $ fuel_type_code_pudl               : chr [1:696244] "coal" "coal" "gas"
"coal" ...
##  $ fuel_group_code                   : chr [1:696244] "coal" "coal" "natur
al_gas" "coal" ...
##  $ supplier_name                     : chr [1:696244] "interocean coal" "i
nterocean coal" "bay gas pipeline" "alabama coal" ...
##  $ fuel_received_units               : num [1:696244] 259412 52241 2783619
25397 764 ...
##  $ fuel_mmbtu_per_unit               : num [1:696244] 23.1 22.8 1.04 24.61
24.45 ...
##  $ fuel_cost_per_mmbtu               : chr [1:696244] "2.134999990463257"
"2.115000009536743" "8.630999565124512" "2.7760000228881836" ...
##  $ bulk_agg_fuel_cost_per_mmbtu      : chr [1:696244] "null" "null" "8.603
500366210938" "null" ...
##  $ fuel_consumed_mmbtu               : num [1:696244] 5992418 1191095 2892
180 625020 18677 ...
##  $ total_fuel_cost                   : chr [1:696244] "12793811" "2519165.
5" "24962406" "1735056" ...
##  $ fuel_cost_per_mmbtu_source        : chr [1:696244] "original" "original
" "original" "original" ...
##  $ sulfur_content_pct                : num [1:696244] 0.49 0.48 0 1.69 0.8
4 ...
##  $ ash_content_pct                   : num [1:696244] 5.4 5.7 0 14.7 15.5
...
##  $ primary_transportation_mode_code  : chr [1:696244] "RV" "RV" "PL" "TR"
...
##  $ secondary_transportation_mode_code: chr [1:696244] "null" "null" "null"
"null" ...
##  $ natural_gas_transport_code        : chr [1:696244] "firm" "firm" "firm"
"firm" ...
##  $ data_maturity                     : chr [1:696244] "final" "final" "fin
al" "final" ...
##  $ mine_name                         : chr [1:696244] "mina pribbenow" "mi
na pribbenow" "null" "alabama coal" ...
##  $ mine_state                        : chr [1:696244] "COL" "COL" "null" "
AL" ...
##  $ coalmine_county_id_fips           : chr [1:696244] "null" "null" "null"
"01007" ...
##  $ mine_type_code                    : chr [1:696244] "SU" "SU" "null" "SU
" ...
```

```
sample_data <- round(0.02 * nrow(Data_New))
sample. <-sample(1:nrow(Data_New), size = sample_data, replace =FALSE)
new_sample <- Data_New[sample.,]
str(new_sample)
```

```
## tibble [13,925 × 31] (S3: tbl_df/tbl/data.frame)
##  $ report_date                       : POSIXct[1:13925], format: "2023-02-
01" "2019-02-01" ...
```

```
##  $ plant_id_eia                      : num [1:13925] 2406 3456 3954 55927
3457 ...
##  $ plant_id_pudl                     : num [1:13925] 1159 416 426 291 327
...
##  $ plant_name_eia                    : chr [1:13925] "Linden" "Newman" "Mt
Storm" "Jasper" ...
##  $ utility_id_eia                    : num [1:13925] 65389 5701 19876 1753
9 55937 ...
##  $ utility_id_pudl                   : num [1:13925] 14245 103 349 292 113
...
##  $ utility_name_eia                  : chr [1:13925] "Linden Combined Cycl
e" "El Paso Electric Co" "Dominion Virginia Power" "South Carolina Electric&G
as Co" ...
##  $ state                             : chr [1:13925] "NJ" "TX" "WV" "SC" .
..
##  $ contract_type_code                : chr [1:13925] "C" "S" "C" "C" ...
##  $ contract_expiration_date          : POSIXct[1:13925], format: "2024-03-
01" "1970-01-01" ...
##  $ energy_source_code                : chr [1:13925] "NG" "NG" "BIT" "NG"
...
##  $ fuel_type_code_pudl               : chr [1:13925] "gas" "gas" "coal" "g
as" ...
##  $ fuel_group_code                   : chr [1:13925] "natural_gas" "natura
l_gas" "coal" "natural_gas" ...
##  $ supplier_name                     : chr [1:13925] "pseg energy" "conoco
phillips" "tri-star mining inc" "bp" ...
##  $ fuel_received_units               : num [1:13925] 738955 32211 21169 19
262 218713 ...
##  $ fuel_mmbtu_per_unit               : num [1:13925] 1.04 1.05 23.54 1.04
1.03 ...
##  $ fuel_cost_per_mmbtu               : chr [1:13925] "4.8846001625061035"
"2.257999897003174" "2.3429999351501465" "3.562999963760376" ...
##  $ bulk_agg_fuel_cost_per_mmbtu      : chr [1:13925] "4.8846001625061035"
"2.698899984359741" "2.268199920654297" "null" ...
##  $ fuel_consumed_mmbtu               : num [1:13925] 772208 33822 498339 1
9994 226149 ...
##  $ total_fuel_cost                   : chr [1:13925] "3771927" "76369.0546
875" "1167609.25" "71238.4609375" ...
##  $ fuel_cost_per_mmbtu_source        : chr [1:13925] "eiaapi" "original" "
original" "original" ...
##  $ sulfur_content_pct                : num [1:13925] 0 0 1.67 0 0 ...
##  $ ash_content_pct                   : num [1:13925] 0 0 18.8 0 0 ...
##  $ primary_transportation_mode_code  : chr [1:13925] "PL" "PL" "TR" "PL" .
..
##  $ secondary_transportation_mode_code: chr [1:13925] "null" "null" "null"
"null" ...
##  $ natural_gas_transport_code        : chr [1:13925] "interruptible" "inte
rruptible" "null" "firm" ...
##  $ data_maturity                     : chr [1:13925] "final" "final" "fina
l" "final" ...
```

```
##  $ mine_name                          : chr [1:13925] "null" "null" "job 3"
"null" ...
##  $ mine_state                         : chr [1:13925] "null" "null" "MD" "n
ull" ...
##  $ coalmine_county_id_fips            : chr [1:13925] "null" "null" "null"
"null" ...
##  $ mine_type_code                     : chr [1:13925] "null" "null" "S" "nu
ll" ...
```

From the 2% sample of data what is the most common fuel type used?

```
table(new_sample$fuel_group_code)
```

```
##
##           coal    natural_gas      other_gas      petroleum petroleum_coke
##           4670           7981             27           1180             67
```

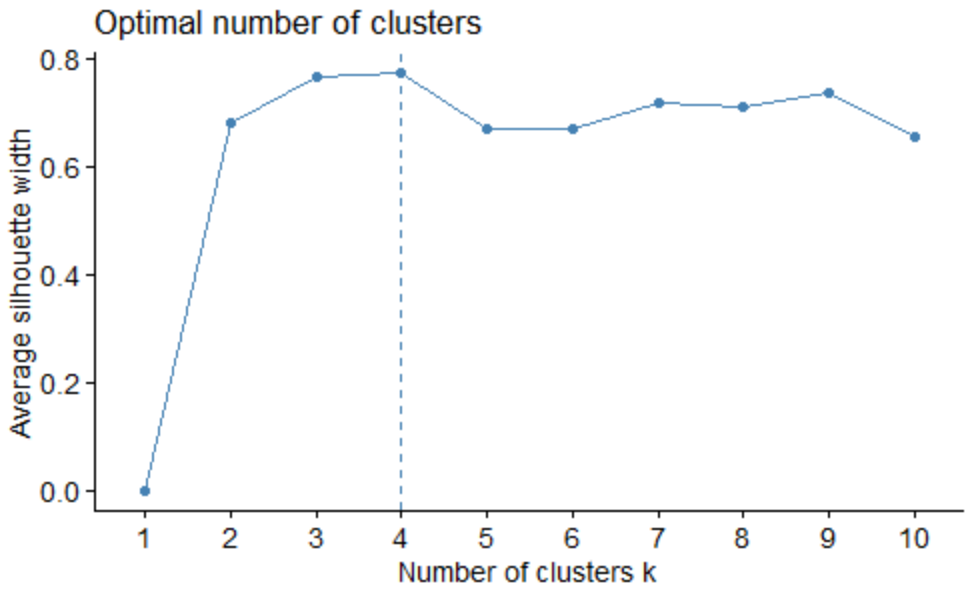
Does the average amount of fuel recieved units for each type support natural gas being the most commonly used? If not what are some assumptions that can be made?

```
aggregate(fuel_received_units ~ fuel_group_code, data = new_sample, FUN = mea
n, na.rm = TRUE)
```

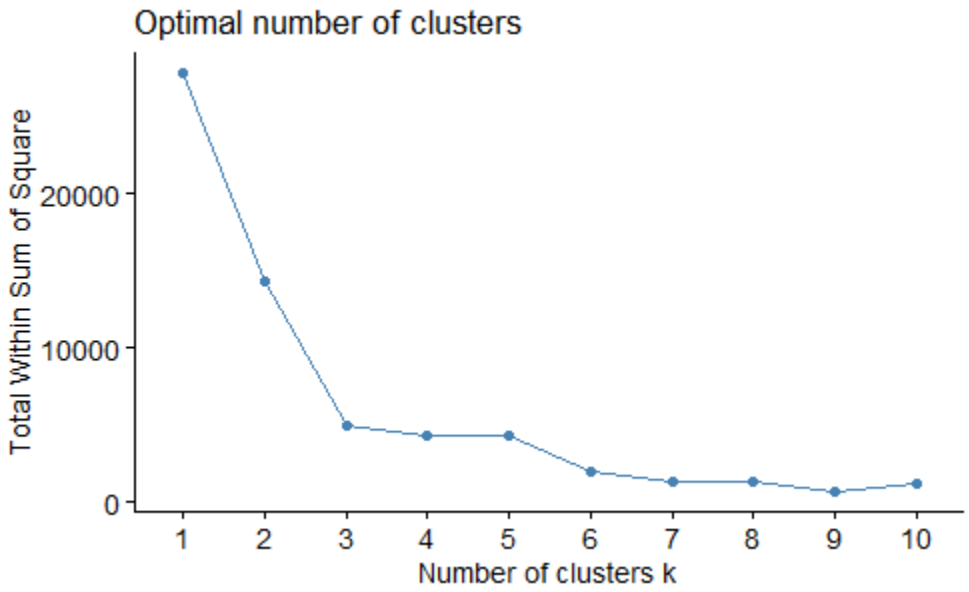
```
##   fuel_group_code fuel_received_units
## 1            coal           48467.56
## 2     natural_gas          425119.99
## 3       other_gas         1735262.89
## 4       petroleum            4996.10
## 5  petroleum_coke           26207.91
```

What are the clusters like for fuel_received_units and fuel_mmbtu_per_unit and what can this tell us based on the units received and fuel in each unit? Finding clusters based on fuel_received_units and fuel_mmbtu_per_unit (units received and fuel in each unit) The number of clusters is determined based on k-means

```
set.seed(1738)
SD <- new_sample[,c(15,16)]
SD <- scale(SD)
fviz_nbclust(SD, kmeans, method = "silhouette")
```



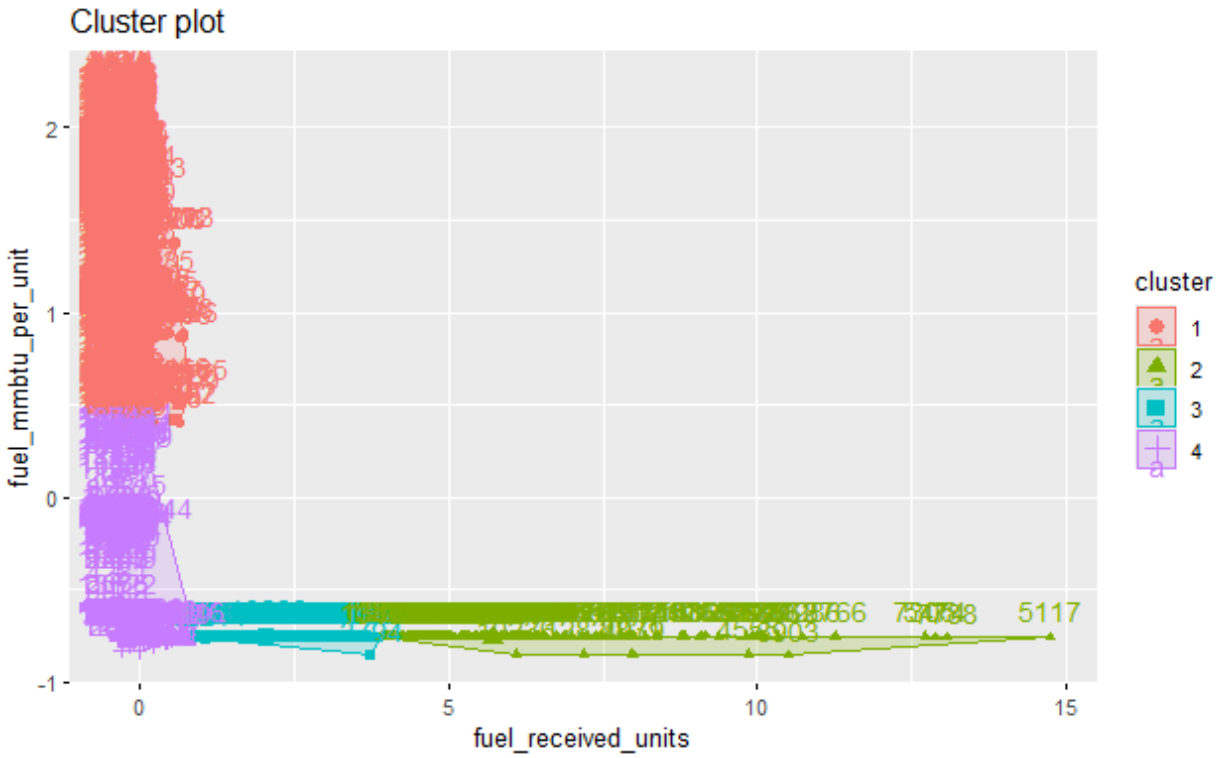

```r
fviz_nbclust(SD, kmeans, method = "wss")
```




```r
k4 <- kmeans(SD, centers = 4, nstart = 25)
centers <- k4$centers
k4$size
```

```
## [1]  4681  226 883 8135
```

```r
fviz_cluster(k4, data = SD)
```



From these clusters we find:

Cluster 1 represents fuel orders received that have the most fuel per unit based on the units received indicating that the most value may be gained from orders within this cluster.

Cluster 2 represents a large variety of fuel units ordered but with a common variable of being below positive value in terms of the fuel received in each unit. Specifically, fuel received units of 10 or more have very little value gained in terms of fuel per unit indicating that the overall cost paid for the units is not likely to be worth the value.

Cluster 3 represents fuel units received with the worst value of fuel per unit based on the units received. By having a limited amount of fuel per unit with minimal units being ordered there is not much value gained per unit.

Cluster 4 represents a mixed value of fuel per unit with small amounts of units received. For any orders that have a positive value for fuel received per unit likely gains value due to small amount of units ordered but any orders with a negative value for fuel do not gain much value.

What this data shows:

“Other Gases” likely has many units within Cluster 2 due to being the most units received on average but being the lowest form of fuel used. Due to Cluster 2 having many negative values of fuel per unit it likely means more units are needed per order to meet demand leading to having the highest average

Petroleum may have many orders in Cluster 1 due to being the lowest units received on average but being the 3rd most used fuel form. Meaning the orders that are created have positive value of fuel per unit leading to less units being needed.

Natural Gas likely has many orders in Cluster 4 and a good portion of Cluster 1 to support being the most used fuel form but also having the 2nd highest units received on average. Meaning more value of fuel per unit is gained in some orders compared to others.