County Based Opportunites for Children

BQOM 2578 | Data Mining | Homework 6 Unsupervised Learning

Group 8: Anthony Pulleo, Hannah Shernisky, Theresa Wohlever

Sunday, November 23, 2025

Table of contents

DELETE ME | Instructions

Cluster Analysis is a very common tool and you may want to apply it to your final project dataset for this assignment if it has metric / continuous data. If your final project is primarily categorical data, please select a different dataset with continuous / metric data for this assignment.

Objective: Apply unsupervised learning to a dataset of your choice.

Steps:

Briefly explain / introduce your the dataset (even if you explained it in prior assignments) and what you hope to get from unsupervised learning analysis of this dataset. 1. Provide an Exploratory Analysis: summary statistics and scatterplots/histograms for example. 2. Explain your Cluster Model Selection - for example, Cluster size Provide a Cluster visualization and interpretation. Be sure to NOT JUST GIVE A TECHNICAL interpretation; how would you interpret / describe the different clusters and what insight do you get from them from your analysis? Submission format: From your QUARTO file, render either a PDF or MS Word Document (*.docx) document that covers the assignment and it should include the R code, the output and your text description, explanation, analysis, and interpretation. The first page must include an executive summary on the points of the assignment - without an executive summary, your submission will not be graded. Only one submission is required by the team. Do not exceed 30 pages.

HW6 Submission Link: https://canvas.pitt.edu/courses/324587/assignments/1871895

Executive Summary

1. With the intention of predicting the Child Opportunity Index (COI), a resource quality measure for healthy development of children our full dataset is aggregated from the following sources:
    - diversitydatakids.org
    - Census data
    - Urban influence codes, and
    - 2025 Country Health data
2. Using our project data we read in the full dataset, clean, wrangle, then prune. The pruned dataset is refined specifically for Random Forest (RF) and Artificial Neural Network (ANN). For the purpose of this homework, we omitted data exploration topics, covered in the final project. The COI score is continuous, therefore we will use regression models through the course of the homework.
3. For random forest, we identified an *optimal mtry at 16*. Running this model, we see that the average number of unhealthy days influences the dependent variable most across the trees. Further, an *OSR of 93, represents a 27% increase in model "accuracy"* over a regression tree.
4. For the neural network, we generated both a zero and one hidden layer model. OSR calculations for both models returned model performance in the *low 80%*. When compared against the Random forest, this is 10-11% less, therefore RF performed better than ANN.

Group 8: Anthony Pulleo, Hannah Shernisky, Theresa Wohlever

Data Preparation

The Child Opportunity Index (COI) measures and maps the quality of resources and conditions like these that matter for children's healthy development in the neighborhoods where they live.

Importing Data, Cleaning, & Wrangling

The data from diversitydatakids.org contains a series of indices, and does not provide the raw data that was used in the index calculation. The indices are normalized across different areas, like education or housing. We will pull from other datasets to see if factors calculated / assessed by those datasets influence the COI.

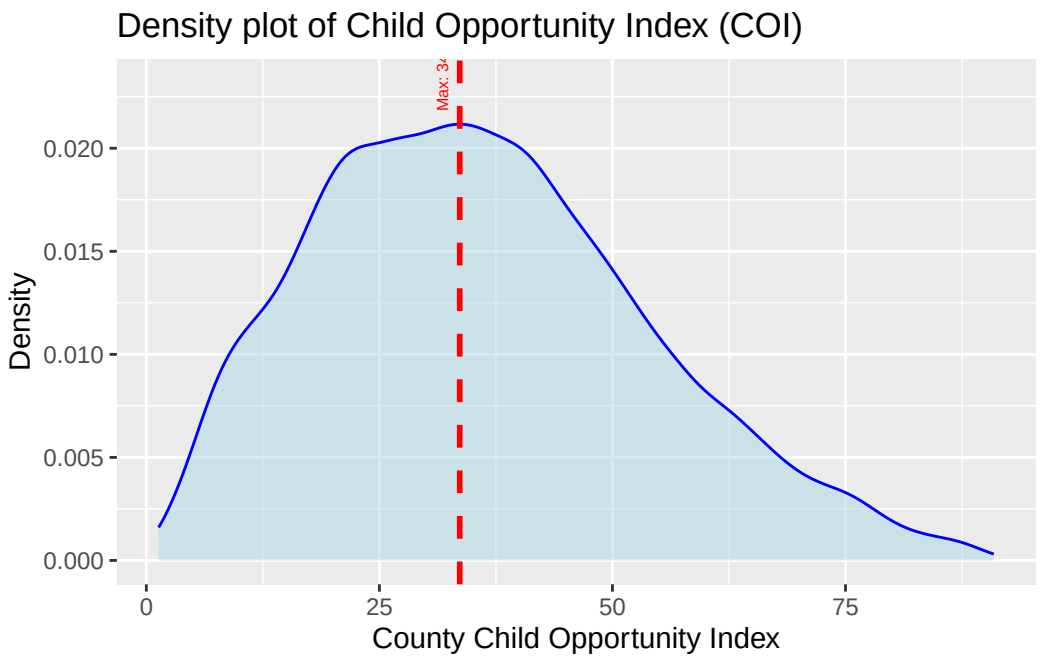
These datasets include the Urban Influence Codes from USDA, Census Data as part of the American Community Survey, and the 2025 County Health Rankings Data. Select variables will be appended via a left_join by County FIPS codes.

Data Exploration

Identify dimensions of interest

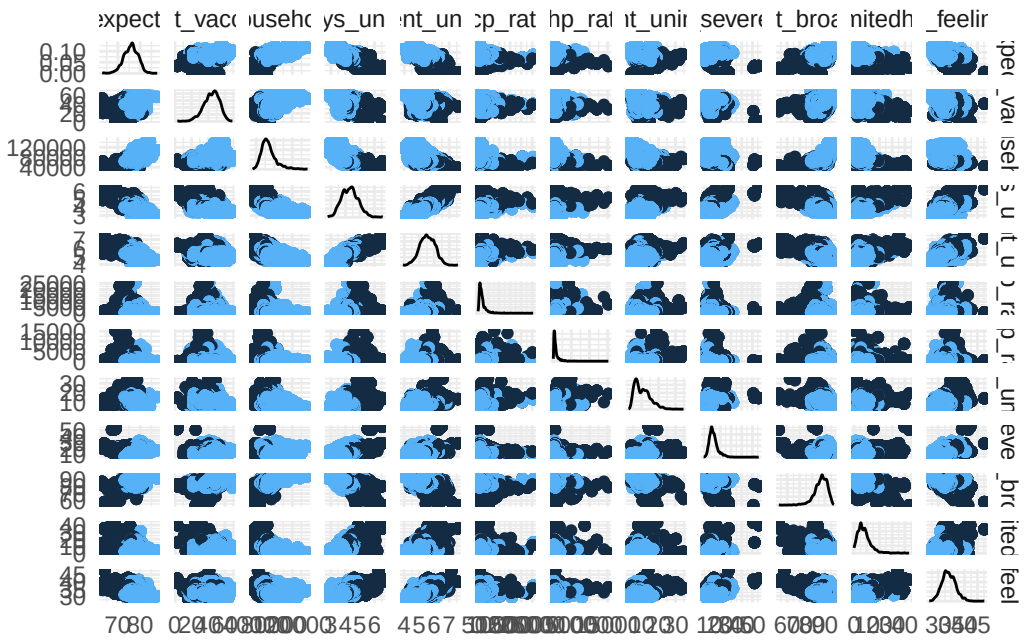
```
df$target <- df$r_COI_nat


##
## Visualize target values
##


# Histogram of target values
target_density <- density(df$target)

# Convert the density estimate to a function
dens_func <- approxfun(target_density$x, target_density$y)

# Use optimize() to find the maximum in a specified interval (choose based on your data)
result <- optimize(dens_func, interval = c(min(df$target), max(df$target)), maximum = TRUE)
local_max_x <- result$maximum     # The x value where local max occurs
local_max_y <- result$objective   # The max density value
```

```r
#  Create density plot with ggplot2 and add vertical line at max
df_density <- data.frame(x = df$target)
ggplot(df_density, aes(x = df$target)) +
  geom_density(fill = "lightblue", color = "blue", alpha = 0.5) +
  geom_vline(xintercept = local_max_x , color = "red", linetype = "dashed", size = 1) +
  annotate("text", x = local_max_x, y = local_max_y + 0.002,
    label = sprintf("Max: %.0f", local_max_x), color = "red", angle = 90, vjust = -1, size = 2) +
  labs(title = "Density plot of Child Opportunity Index (COI)", x = "County Child Opportunity Index", y = "Density")
```




```r
## Make TARGET binary
target_bin_cutoff <- local_max_x
df$target_bin <- ifelse(df$target < target_bin_cutoff, 0, 1)



scatter_plot_matrix <- ggpairs(
  df[ , colsOfInterest_scatter],
  aes(color = df$target_bin),
  upper = list(continuous = "points"),
  lower = list(continuous = "points"),
  diag  = list(continuous = "densityDiag")
) +
  theme_minimal()
```

```
print(scatter_plot_matrix)
```



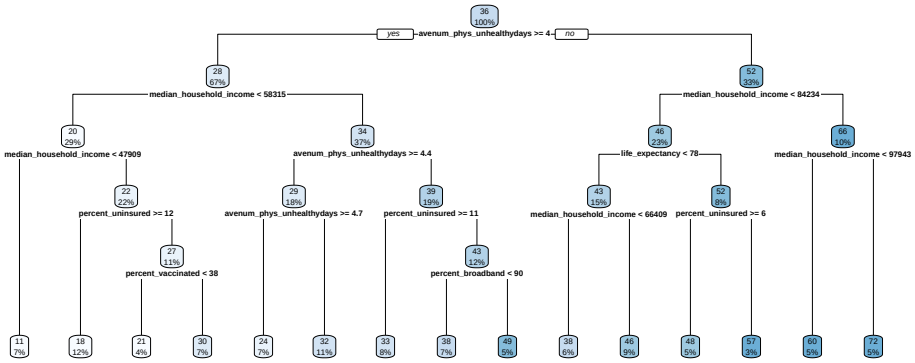
```
df_wTarget <- df


df$target_bin <- NULL
df$target <- NULL
```

Modeling

Split data into training and testing

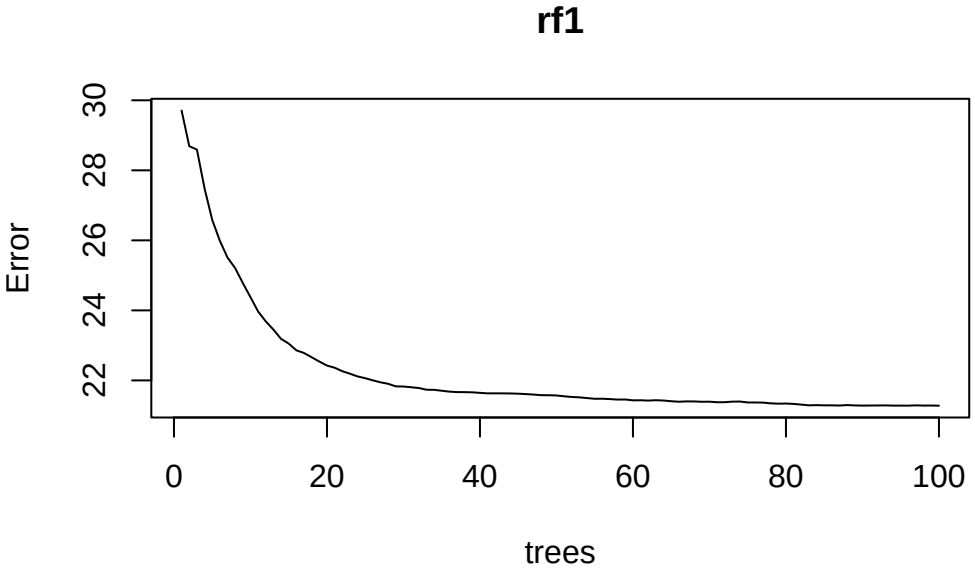
We will deviate slightly from the dataset we will use in our final project for the purpose of this homework. When doing the ANN, we identified a few variables that made the ANN return NaNs, and therefore inhibited our ability to successfully run these models. We will create df in the code below with variables that allow the ANN to work.

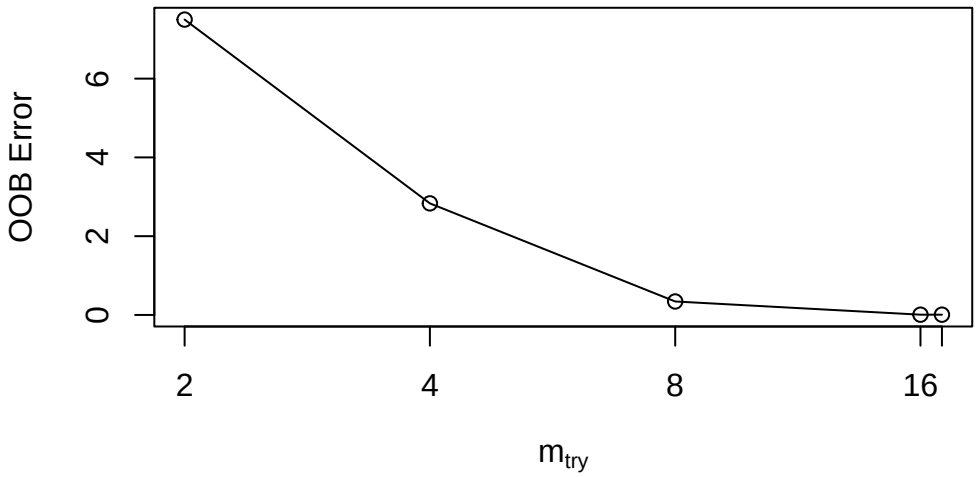
Regression Tree



Random Forest

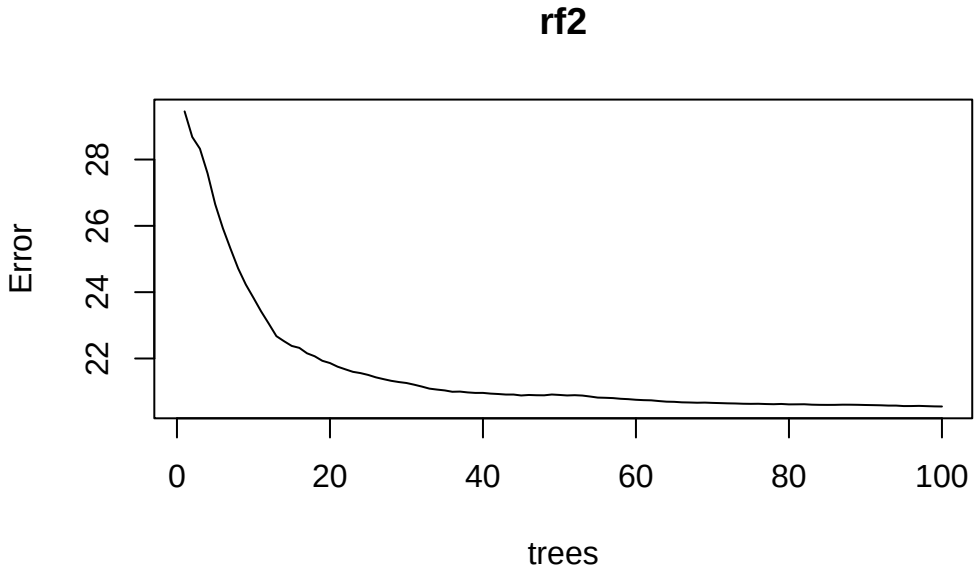
We will start by creating a random forest with similar parameters called for in class, and then we will identify the best mtry.

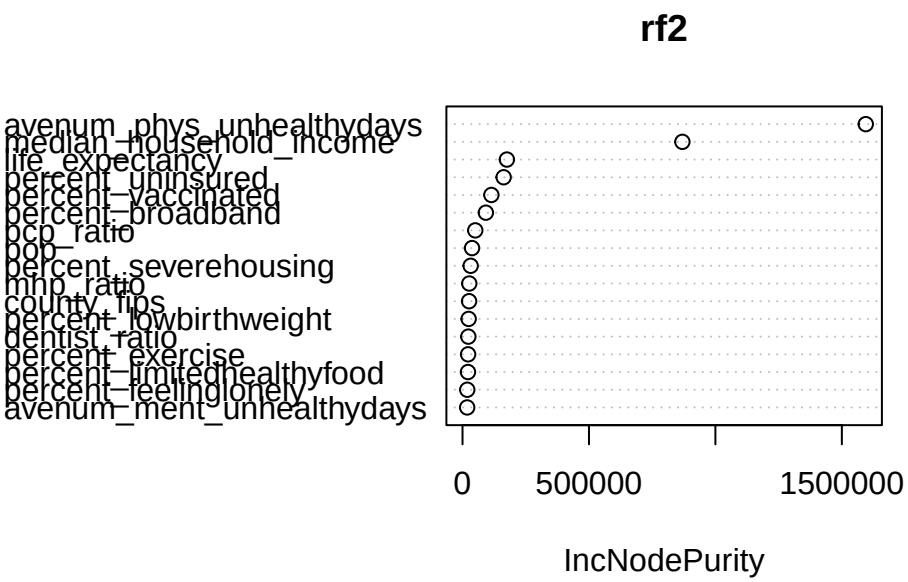




The lowest Out-of-Bag error from the dataset and model indicates that an optimal mtry is 16, because this represents the lowest error in the model. Despite the step of 2, we are fairly confident there are not further dips in the OOB Error curve, given the gentle descent of the curve.

We will then generate a "final" Random Forest, with the optimal mtry:





Average number of unhealthy days has the highest impact on inceasing node purity, followed by median household income, and life expectancy

Comparison to Regression Tree

Artificial Neural Network

Comparing with the RF performance:

The Random Forest Performs approximately 11% better than the neural network performed. This may be attributable to the fact that the Neural Network as defined above is not fully optimized (as more nodes and layers may have a positive benefit on the overall performance of the model).

Unsupervised Learning

```
# JAMBU ELBOW
iss <- function(k) {
  kmeans(df_cluster,
         centers = k,
         iter.max=100,
         nstart=100,
         algorithm="Lloyd" )$tot.withinss
}
```

```r
k.values <- 1:10
iss_values <- map_dbl(k.values, iss)
```

```
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
```

```
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
```

```
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
```

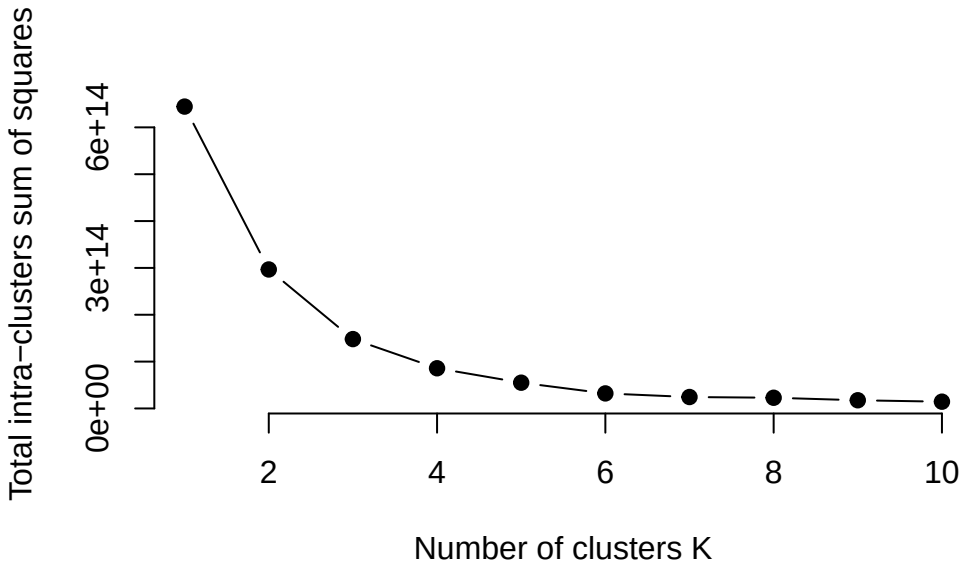
```
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
```

```
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
```

```
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
```

```
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
Warning: did not converge in 100 iterations
```

```r
plot(k.values, iss_values,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total intra-clusters sum of squares")
```




```r
total_ss <- sum((df_cluster- colMeans(df_cluster))^2)


iss <- function(k) {
```

```
  wss <- kmeans(df_cluster,
         centers = k,
         iter.max=100,
         nstart=100,
         algorithm="Lloyd")$tot.withinss
  return(1 - wss / total_ss)
}

k.values <- 1:10
perc_variance <- map_dbl(k.values, iss)
```

```
k_pop_asymptote_hunter <- asymptote(
  x = k.values,
  y = perc_variance,
  degree = "optim",
  upper.degree = 5,
  threshold = 0.95,  # Once y reaches 95% of asymptote
  proportional = TRUE,
  estimator = "glm",
  ci.level = NULL # We don't need confidence intervals - prevent ERROR: object 'y_lwr' not found
)
```

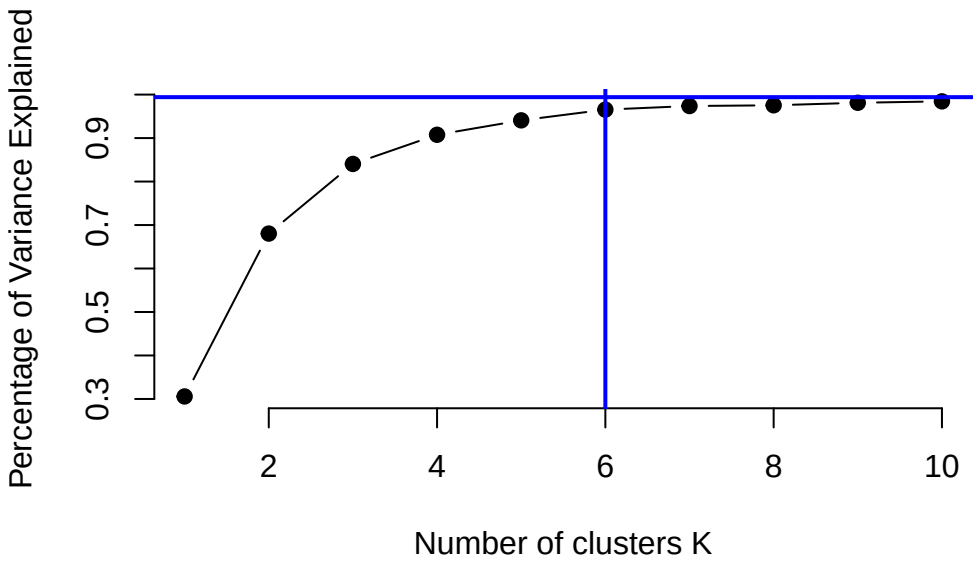
```
Estimated horizontal asymptote: ~0.9941373

Asymptote reached at x = 6
Asymptote threshold used: y = 0.9444304
Confidence level used:
```

```
k_pop_asymptote_hunter
```

```
$results
    x         y        ys
1   1 0.3057321 0.3057321
2   2 0.6803821 0.6803821
3   3 0.8404634 0.8404634
4   4 0.9076044 0.9076044
5   5 0.9407322 0.9407322
6   6 0.9654296 0.9654296
7   7 0.9737176 0.9737176
8   8 0.9753771 0.9753771
9   9 0.9811542 0.9811542
```

```
10 10 0.9844434 0.9844434

$h.asymptote
[1] 0.9941373

$min.n
[1] 6

$optimal.degree
[1] 5

$glm.results
[1] 1.315083
```

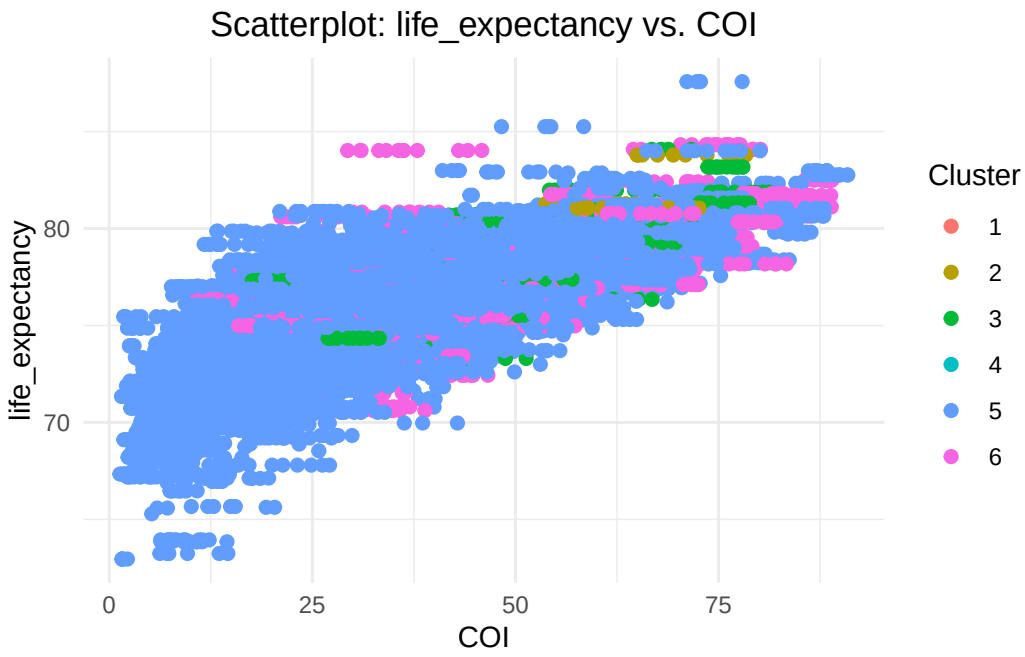
```r
plot(k.values, perc_variance,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Percentage of Variance Explained")
abline(h = k_pop_asymptote_hunter$h.asymptote, col = "blue", lwd = 2)
abline(v = k_pop_asymptote_hunter$min.n, col = "blue", lwd = 2)
```

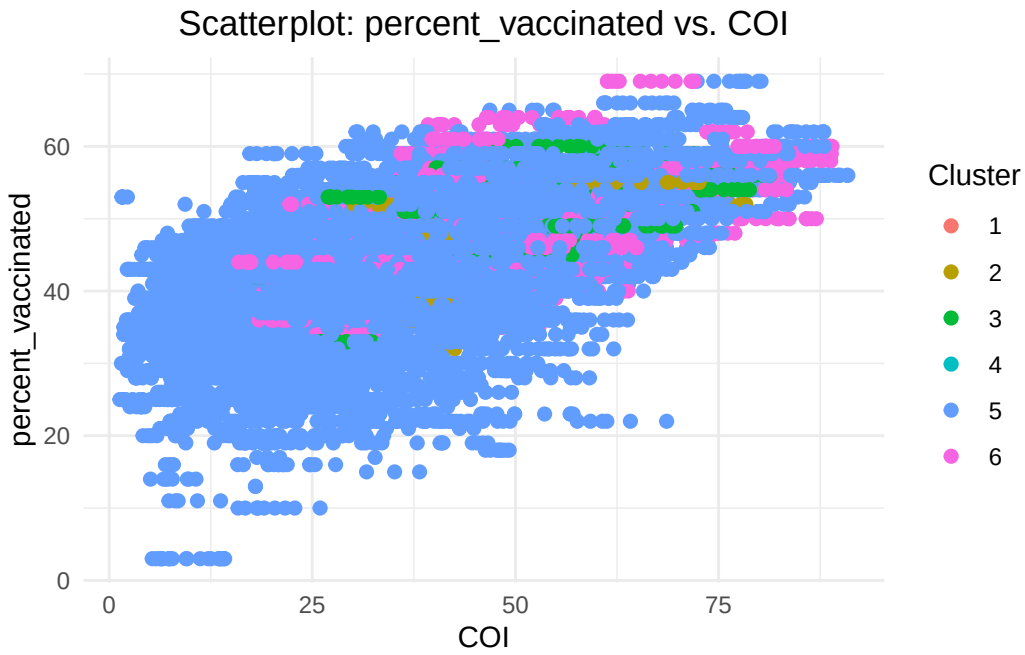


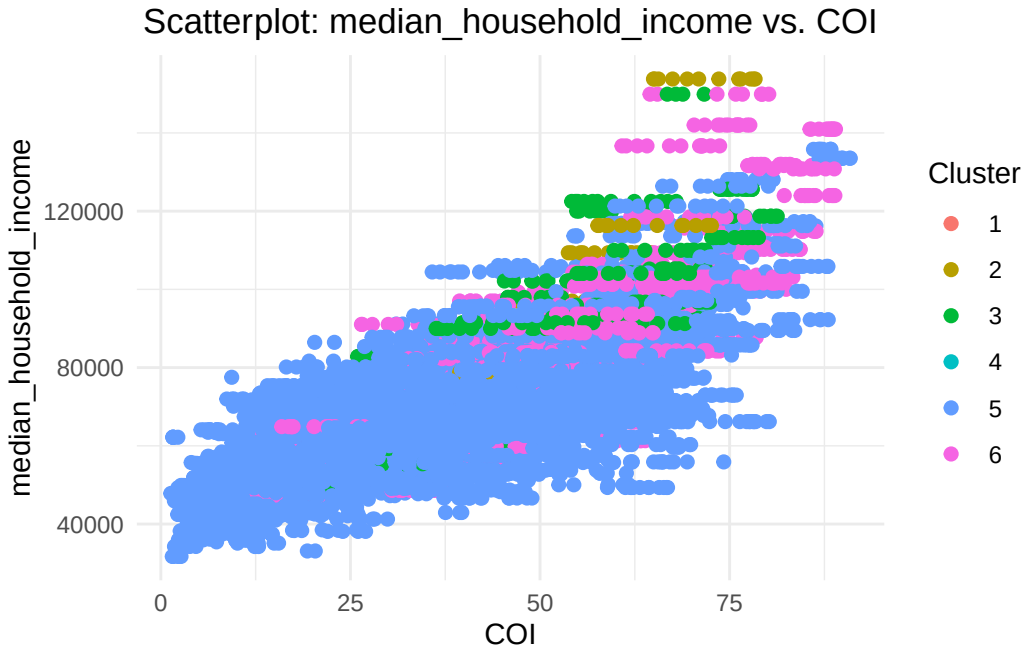

```r
k_pop_best_K <- k_pop_asymptote_hunter$min.n
```

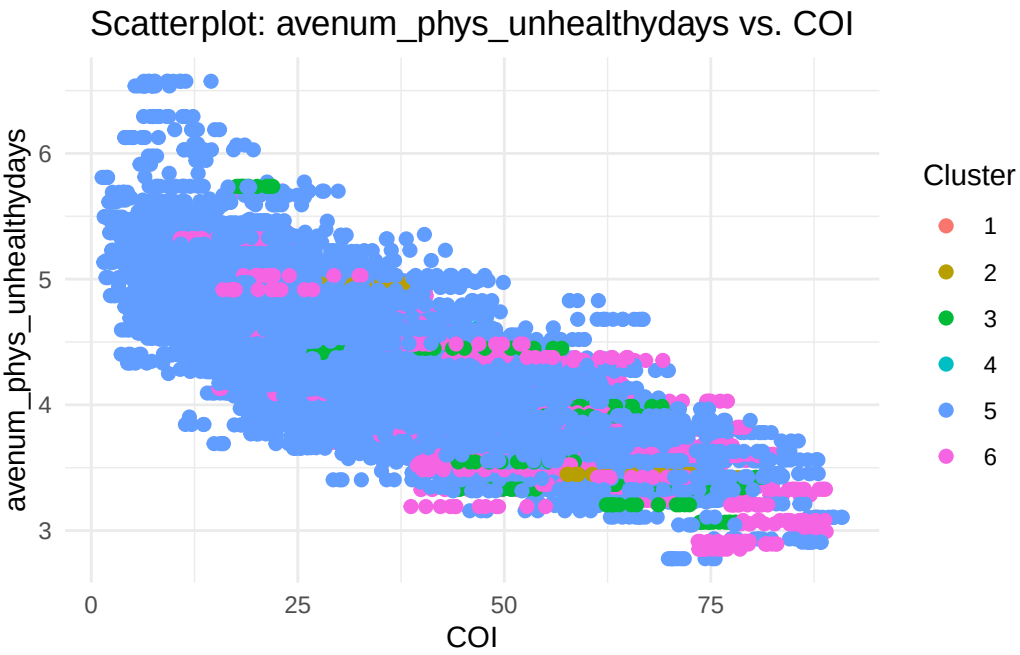
```r
set.seed(seed_this)
Kpop <- kmeans(df_cluster, k_pop_best_K, iter.max=100, nstart=50)
df$ClusterNumber <- Kpop$cluster
```

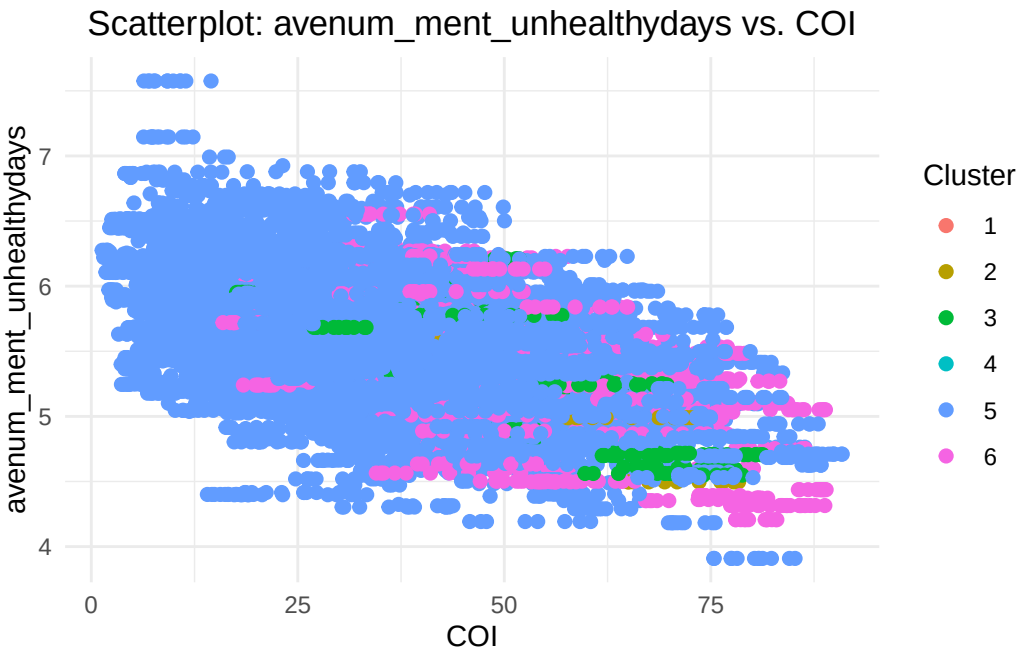
```
for(col in colsOfInterest_scatter) {
  # Use aes_string or .data pronoun for dynamic column selection
  p <- ggplot(df, aes(x=r_COI_nat, y=.data[[col]],
                      color=as.factor(ClusterNumber))) +
    geom_point(size=2) +
    labs(x = "COI", y = col, color = "Cluster") +
    ggtitle(paste("Scatterplot:", col, "vs. COI")) +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5))

  print(p)

}
```

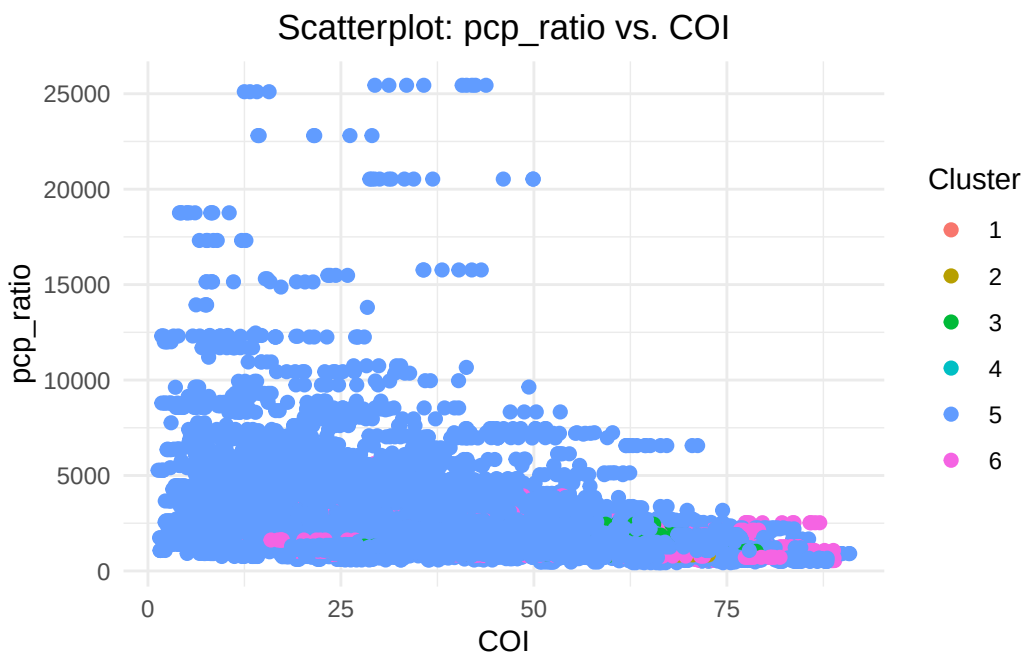


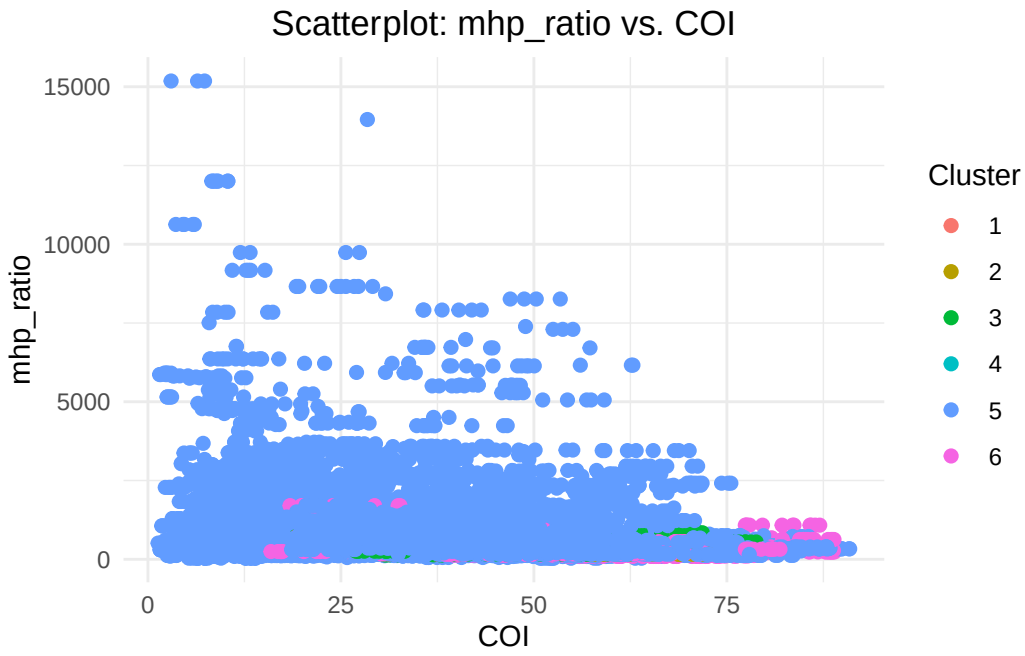

Scatterplot: percent_vaccinated vs. COI
percent_vaccinated
COI
0
20
40
60
0
25
50
75
Cluster
1
2
3
4
5
6

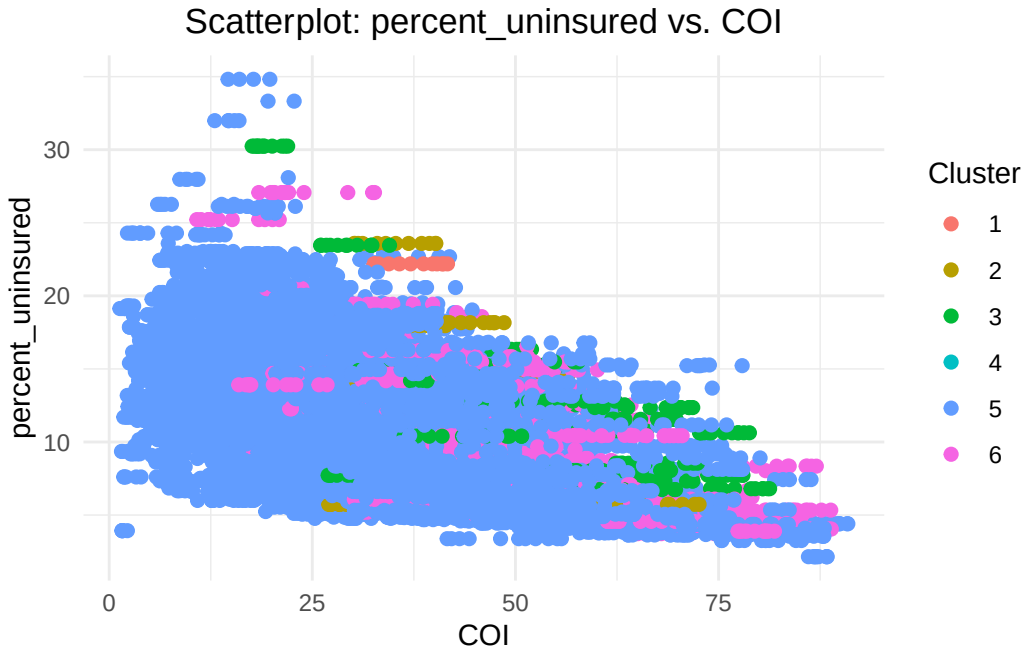

Scatterplot: median_household_income vs. COI
median_household_income
COI
40000
80000
120000
0
25
50
75
Cluster
1
2
3
4
5
6

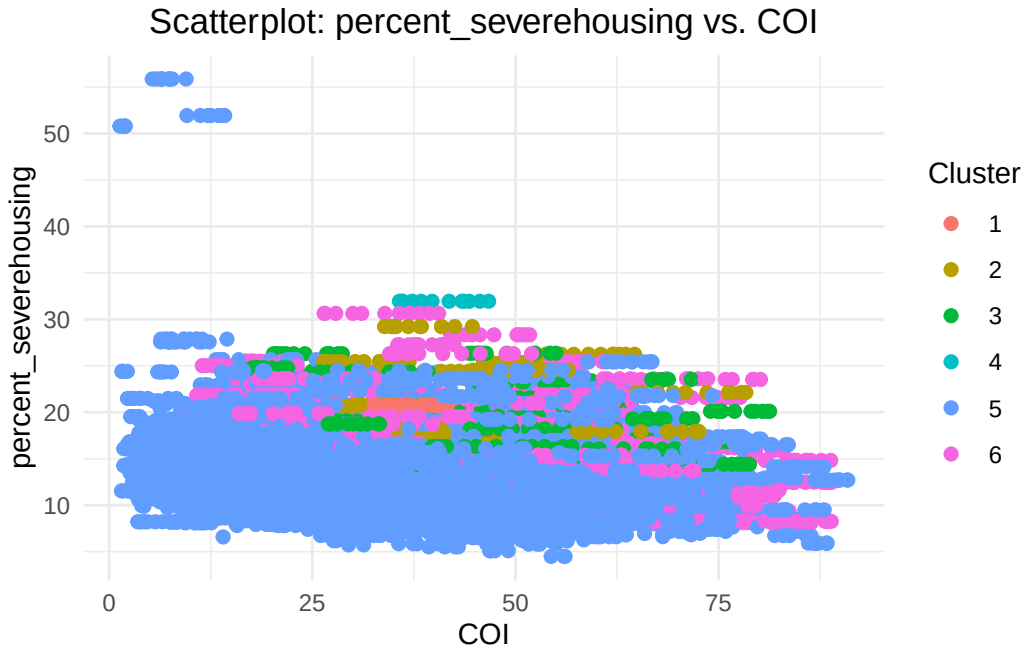

Scatterplot: avenum_phys_unhealthydays vs. COI
avenum_phys_unhealthydays
6
5
4
3
0
25
50
75
COI
Cluster
1
2
3
4
5
6

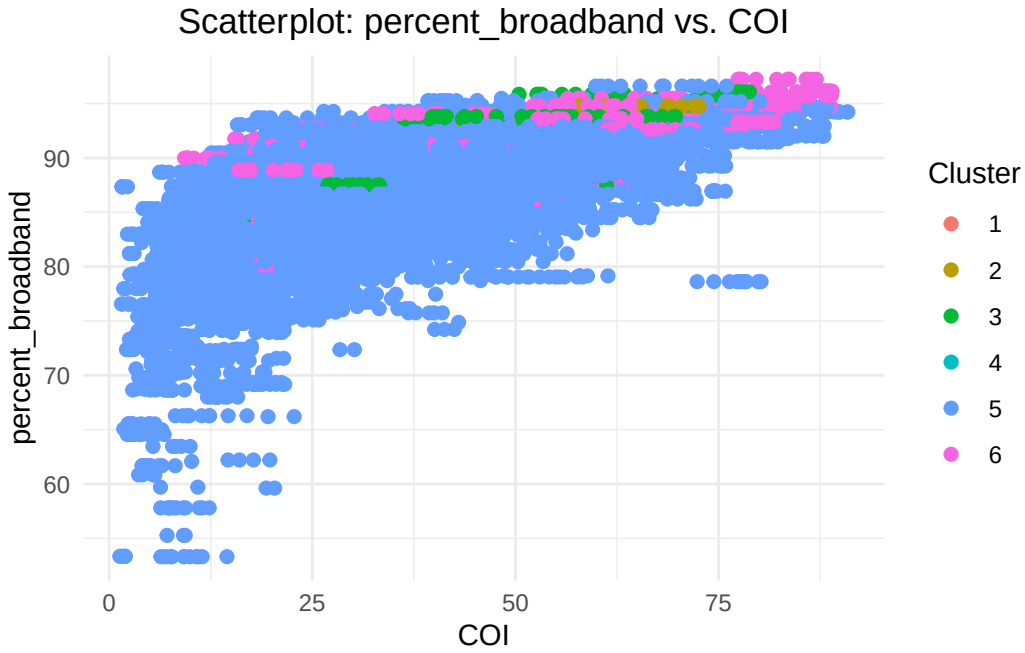

Scatterplot: avenum_ment_unhealthydays vs. COI
avenum_ment_unhealthydays
7
6
5
4
0
25
50
75
COI
Cluster
1
2
3
4
5
6

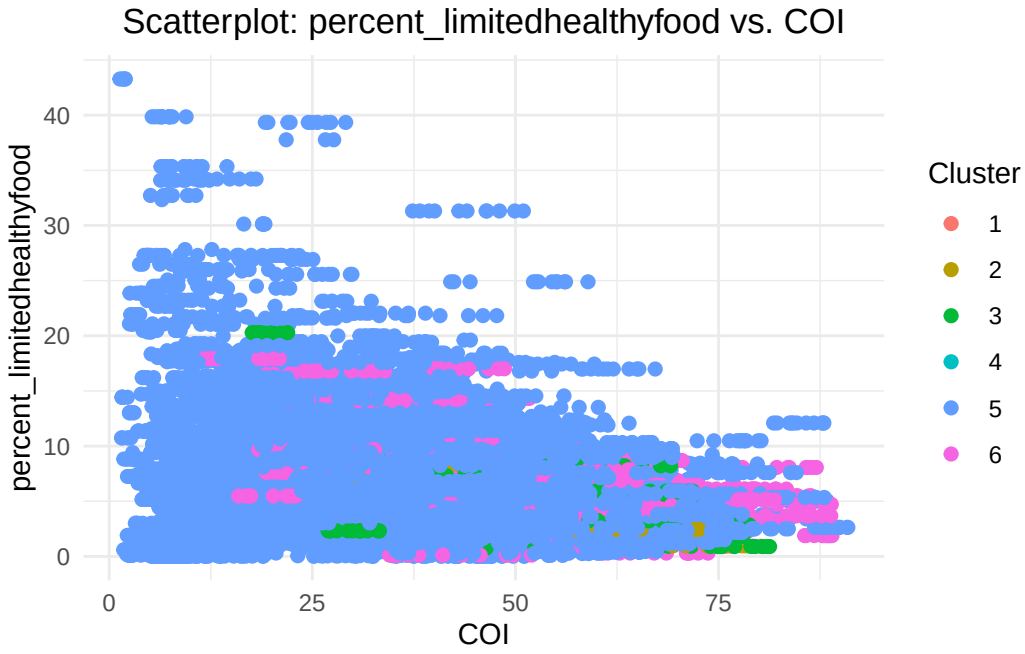

Scatterplot: pcp_ratio vs. COI
pcp_ratio
25000
20000
15000
10000
5000
0
0
25
50
75
COI
Cluster
1
2
3
4
5
6

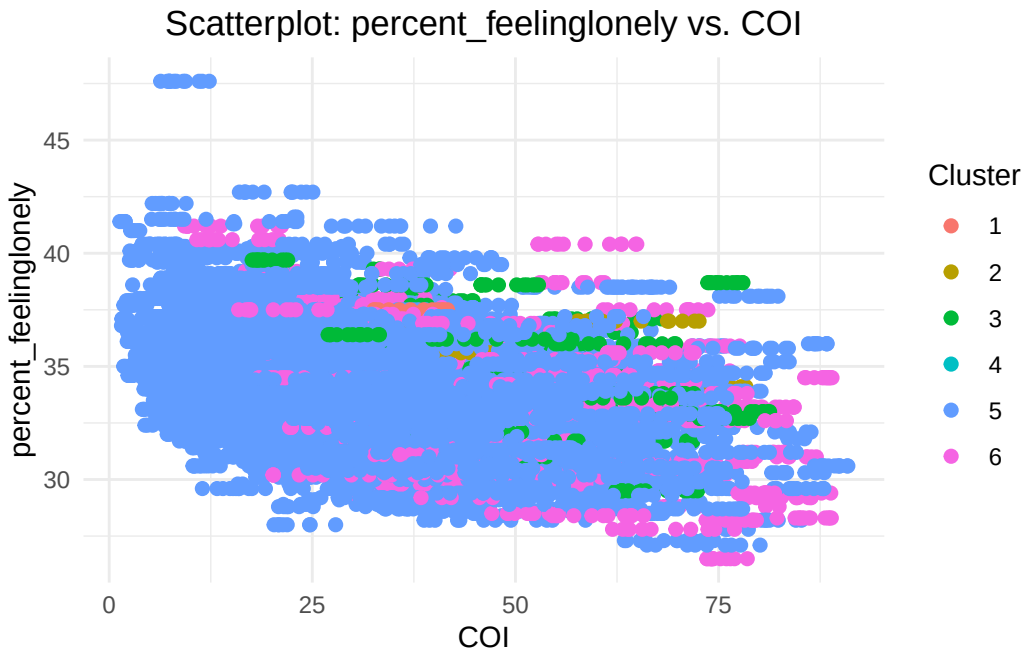

Scatterplot: mhp_ratio vs. COI
mhp_ratio
15000
10000
5000
0
0
25
50
75
COI
Cluster
1
2
3
4
5
6


Scatterplot: percent_uninsured vs. COI
percent_uninsured
30
20
10
0
25
50
75
COI
Cluster
1
2
3
4
5
6


Scatterplot: percent_severehousing vs. COI
percent_severehousing
50
40
30
20
10
0
25
50
75
COI
Cluster
1
2
3
4
5
6


Scatterplot: percent_broadband vs. COI
percent_broadband
90
80
70
60
0
25
50
75
COI
Cluster
1
2
3
4
5
6


Scatterplot: percent_limitedhealthyfood vs. COI
percent_limitedhealthyfood
40
30
20
10
0
0
25
50
75
COI
Cluster
1
2
3
4
5
6



References

https://www.diversitydatakids.org/research-library/child-opportunity-index-30-2023-county-data

https://www.diversitydatakids.org/research-library/research-brief/what-child-opportunity

https://www.ers.usda.gov/data-products/urban-influence-codes