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

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. 3.Clustering was conducted using **K-means** on a broad set of standardized variables reflecting life expectancy, vaccination rates, income, health outcomes, and more.Feature selection and standardization were appropriately attempted to mitigate scale issues. The number of clusters (K) was selected using a **variance explained/asymptote** (elbow-type) approach, suggesting an optimal $K \approx 6$.

3. Cluster analysis performed on the county-level Child Opportunity Index (COI) and health/socioeconomic indicators did **not yield clear, actionable or interpretable groupings**. Key technical and data issues hindered the identification of distinct clusters, substantially limiting potential insights for policy or intervention.

    - **Multiple warnings of non-convergence** in the K-means algorithm indicated instability and unreliable assignment of cluster labels.
    - **High overlap of clusters** in scatterplots across all variable pairs showed a failure of the algorithm to find meaningful group separation.
    - **No clear, distinguishable patterns** among clusters could be identified in terms of key variables such as COI, healthcare access, or economic measures.
    - **Underlying Reasons for Failure:**
        - Continued presence of non-informative or highly correlated features.
        - High dimensionality with relatively little structure.
        - The variety in magnitude/range across input variables, despite attempted standardization, likely contributed to the curse of dimensionality.
        - Insufficient signal or inherent structure in the data to naturally group counties into distinct types.

4. Cluster labels cannot be reliably connected to actionable geographic, socioeconomic, or health profiles.

5. Insights are Gained Despite Unsuccessful Clustering

    - **Current variables or selected data do not naturally segment counties into meaningful subgroups**, at least not in the space defined by K-means and these features.
    - **Clustering is sensitive to feature selection, scaling, and algorithm choice**: future efforts may require
        - More targeted variable selection, perhaps focusing on a smaller, more interpretable set.
        - Dimensionality reduction (e.g., PCA before clustering).

  - Trying alternative clustering algorithms (e.g., hierarchical, DBSCAN).
- **Domain knowledge is crucial**: Automated clustering alone does not guarantee segments that are useful for intervention or narrative.

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.

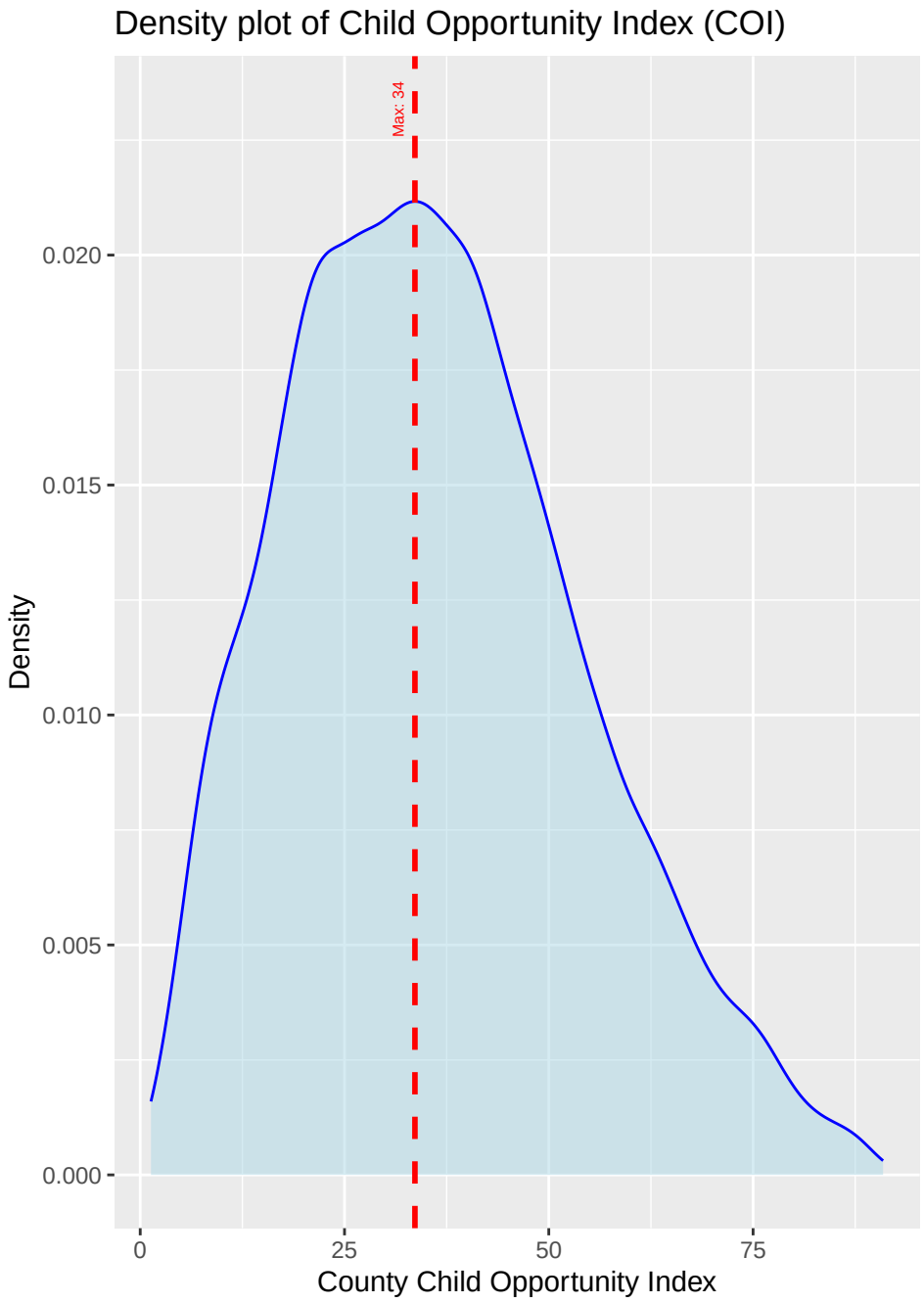
These datasets include the Urban Influence Codes from USDA, Census Data as part of the American Community Survey, and the 2025 County Health Rankings Data. Select variables will be appended via a left_join by County FIPS codes.

Data Exploration

Identify dimensions of interest

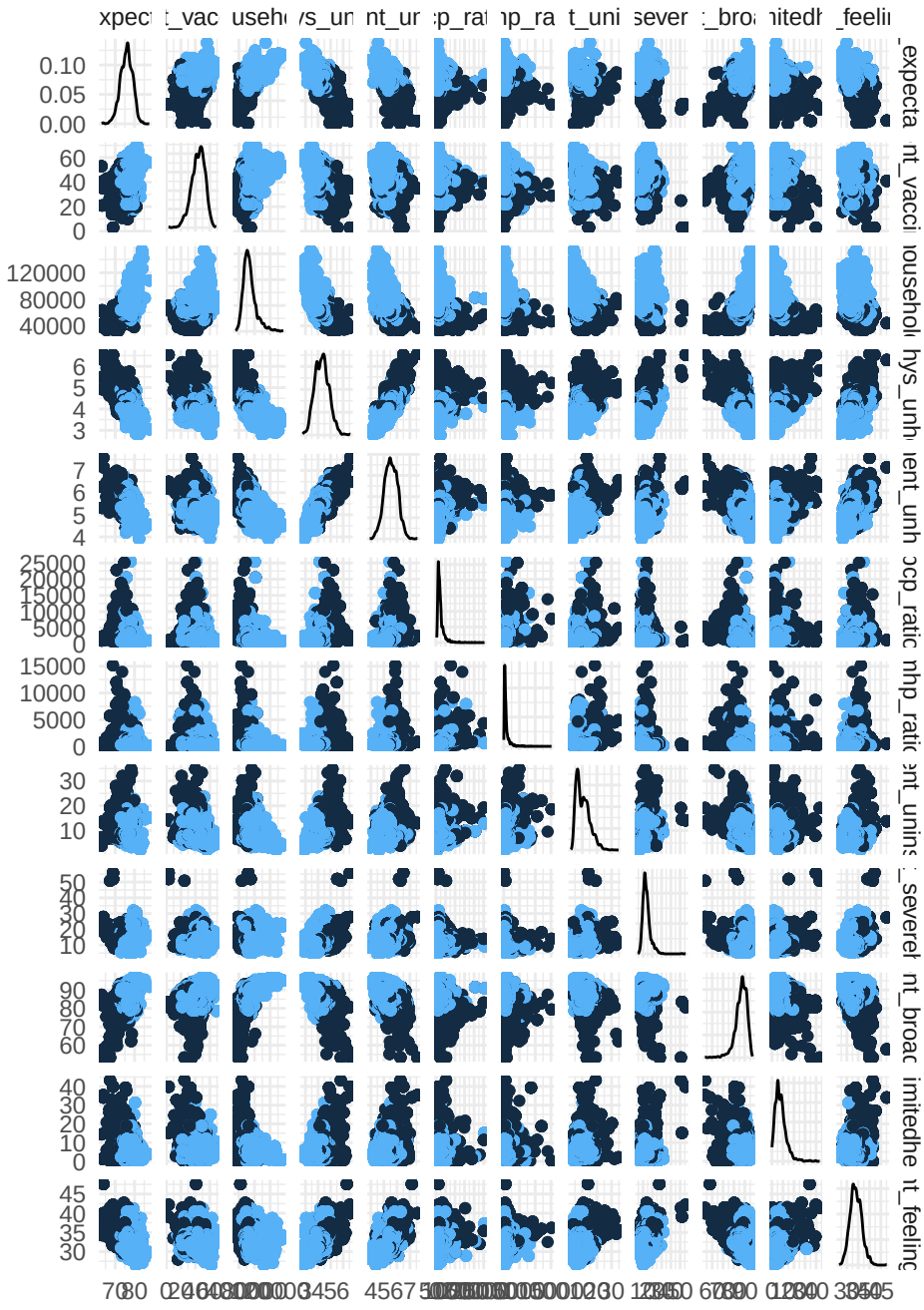
```
df$target <- df$r_COI_nat


##
## 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)
```

```r
# 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

#  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")
```




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

```r
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

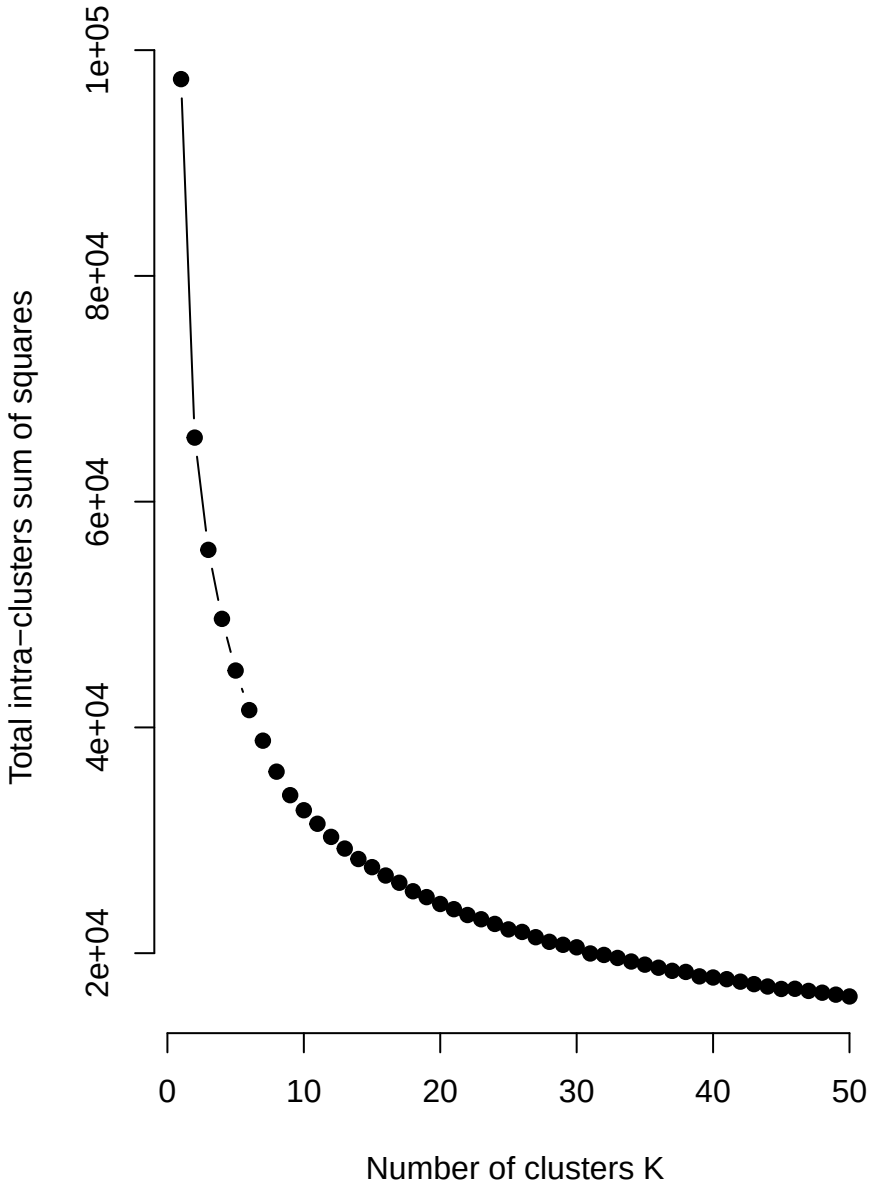
Split data into training and testing

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.

Unsupervised Learning

K-Means Clustering

```
WSS decreases at each step:


 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[25]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[37]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
[49]  TRUE
```




```r
k_asym <- function(x,y,kd,kud){
  asymptote(
  x = x,
  y = y,
  degree = kd, # "optim",
  upper.degree = kud,
  # threshold = 0.90,  # Once y reaches 90% of asymptote
  proportional = TRUE,
  estimator = "mean", # "glm",
```

```
  ci.level = NULL # We don't need confidence intervals - prevent ERROR: object 'y_lwr' not found
  )
}

k_pop_asymptote_hunter <- k_asym(k.values, perc_variance, kpop_degree, kpop_uppper.degree)
```

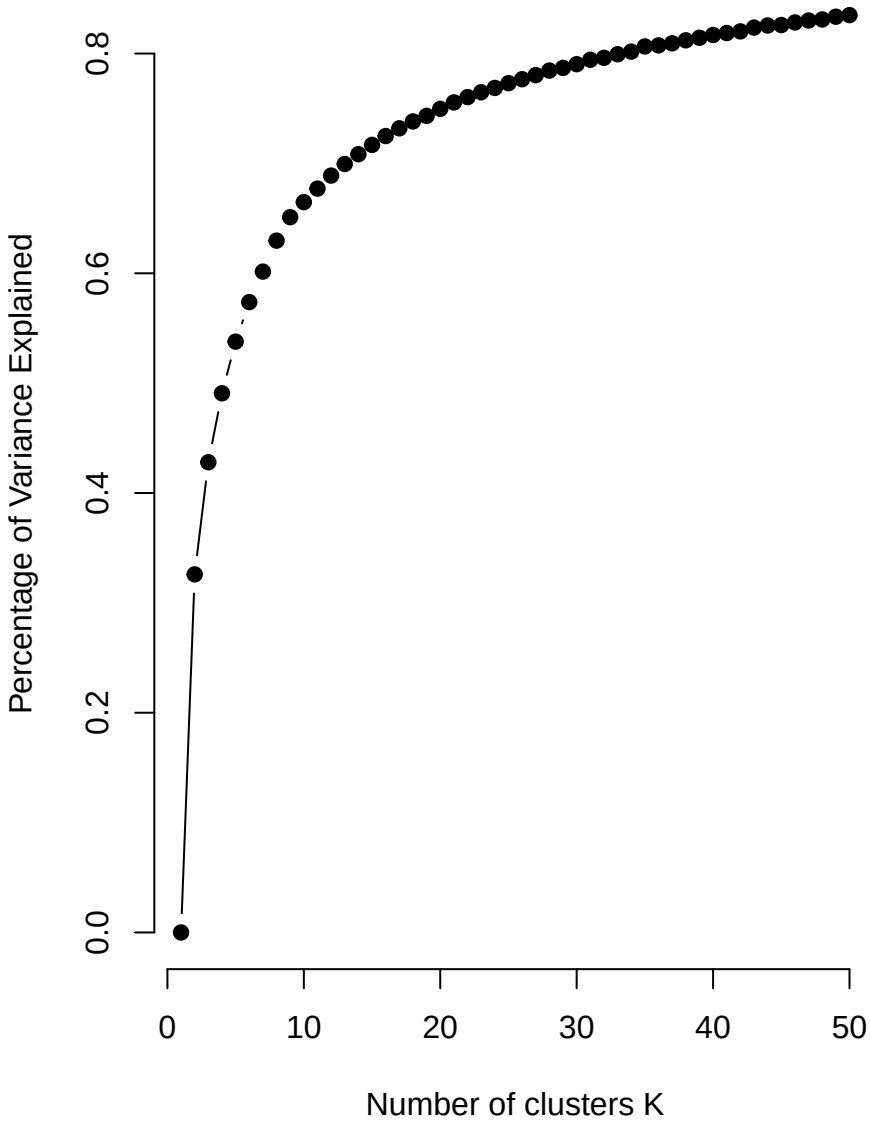
```
Estimated horizontal asymptote: ~0.9433625


Asymptote threshold used: y = 0.8961943
Confidence level used:
```

```
k_pop_asymptote_hunter
```

```
$results
    x             y            ys
1   1 -1.387779e-13 -2.916065e-06
2   2  3.259451e-01  3.259604e-01
3   3  4.280117e-01  4.279039e-01
4   4  4.907647e-01  4.924867e-01
5   5  5.377914e-01  5.383384e-01
6   6  5.737362e-01  5.735602e-01
7   7  6.015142e-01  6.017404e-01
8   8  6.297011e-01  6.278606e-01
9   9  6.510750e-01  6.521331e-01
10 10  6.647710e-01  6.648795e-01
11 11  6.770953e-01  6.772949e-01
12 12  6.889304e-01  6.888022e-01
13 13  6.994655e-01  6.992039e-01
14 14  7.083766e-01  7.085811e-01
15 15  7.168707e-01  7.170631e-01
16 16  7.249261e-01  7.247739e-01
17 17  7.319532e-01  7.318266e-01
18 18  7.382079e-01  7.390217e-01
19 19  7.432911e-01  7.441082e-01
20 20  7.496794e-01  7.494065e-01
21 21  7.555295e-01  7.570979e-01
22 22  7.603695e-01  7.602815e-01
23 23  7.647920e-01  7.646815e-01
24 24  7.687221e-01  7.689129e-01
25 25  7.730353e-01  7.729353e-01
26 26  7.766546e-01  7.767556e-01
```

```
27 27  7.803547e-01  7.803846e-01
28 28  7.845152e-01  7.838147e-01
29 29  7.869529e-01  7.899538e-01
30 30  7.903280e-01  7.904900e-01
31 31  7.943680e-01  7.934952e-01
32 32  7.961374e-01  7.963945e-01
33 33  7.993728e-01  7.990777e-01
34 34  8.017609e-01  8.022709e-01
35 35  8.064554e-01  8.048259e-01
36 36  8.074573e-01  8.070601e-01
37 37  8.093134e-01  8.096394e-01
38 38  8.120631e-01  8.120481e-01
39 39  8.143801e-01  8.143783e-01
40 40  8.169269e-01  8.166430e-01
41 41  8.186656e-01  8.188541e-01
42 42  8.201734e-01  8.220366e-01
43 43  8.236525e-01  8.231838e-01
44 44  8.255551e-01  8.260578e-01
45 45  8.259744e-01  8.259728e-01
46 46  8.283015e-01  8.282698e-01
47 47  8.300886e-01  8.301160e-01
48 48  8.310518e-01  8.310294e-01
49 49  8.335440e-01  8.335450e-01
50 50  8.349801e-01  8.349800e-01

$h.asymptote
[1] 0.9433625

$min.n
[1] NA

$optimal.degree
[1] 4
```

```r
plot(k.values, perc_variance,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Percentage of Variance Explained")
```



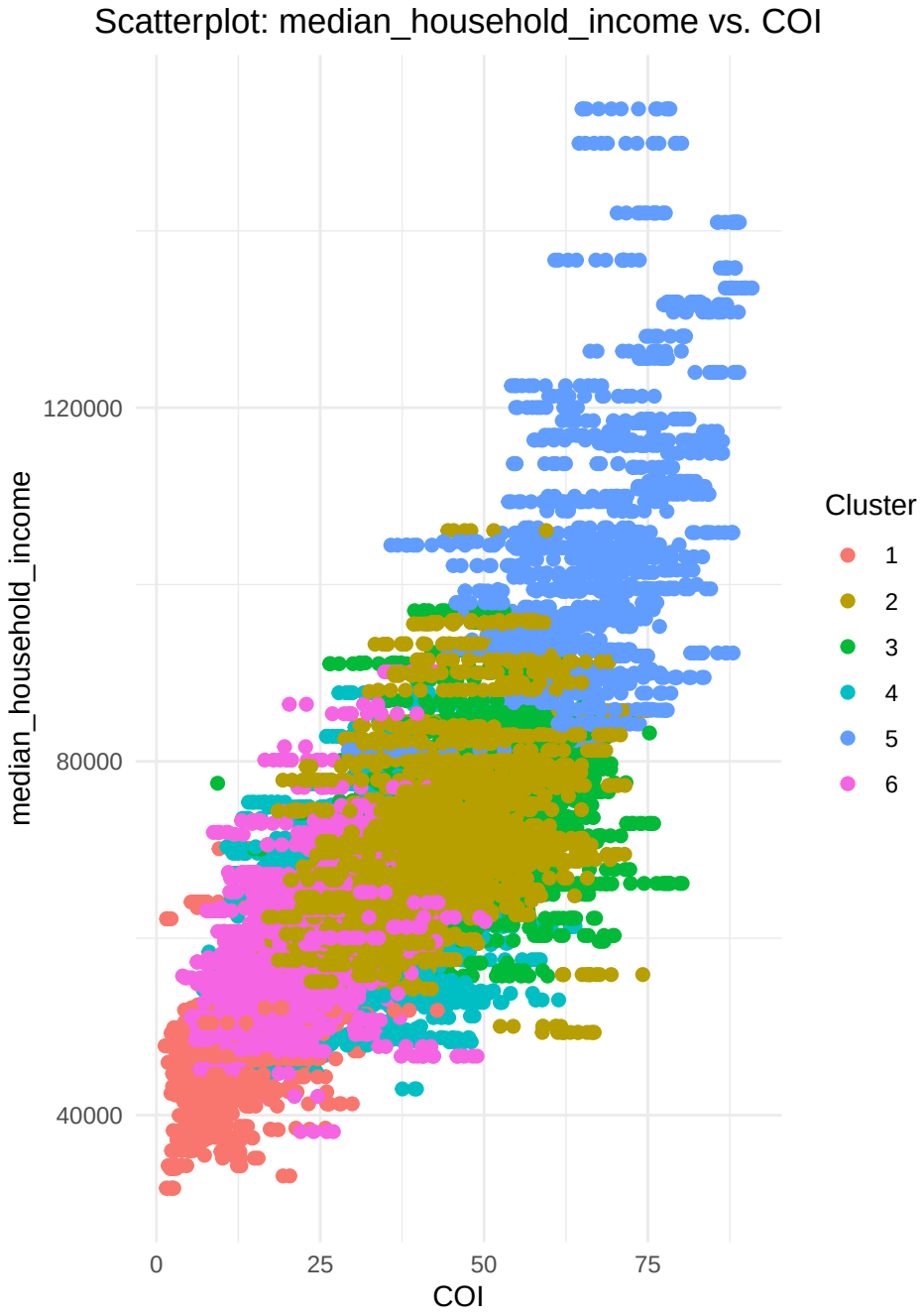

```
# points(k_pop_asymptote_hunter$min.n, k_pop_asymptote_hunter$h.asymptote, col = "blue", lwd = 2)

k_pop_best_K <- k_pop_asymptote_hunter$min.n
```

```
set.seed(seed_this)

## Kpop <- kmeans(df_cluster, k_pop_best_K, iter.max=100, nstart=50)
Kpop <- kmeans(df_cluster,6, 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)

}
```

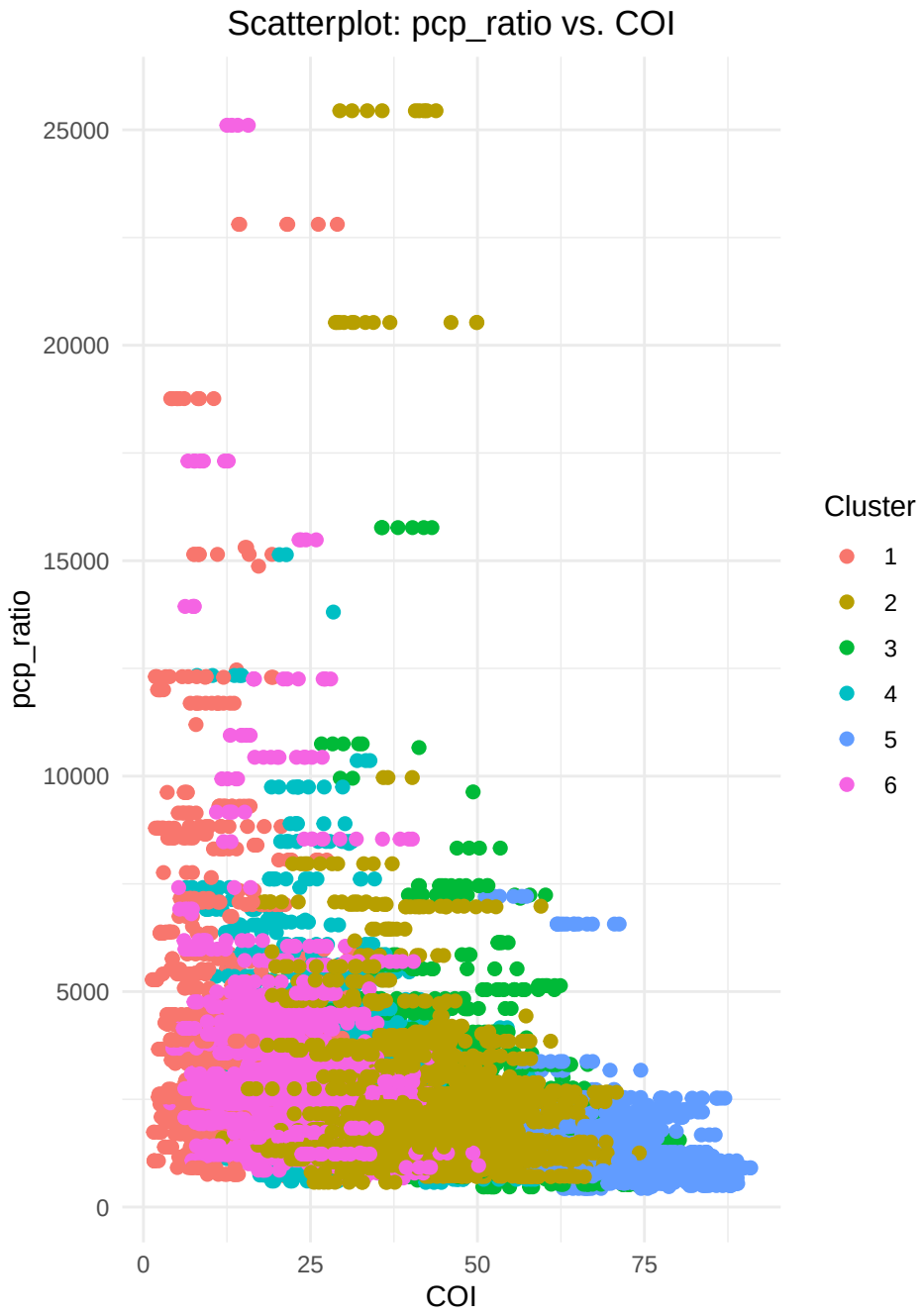

Scatterplot: life_expectancy vs. COI
life_expectancy
80
70
0
25
50
75
COI
Cluster
1
2
3
4
5
6

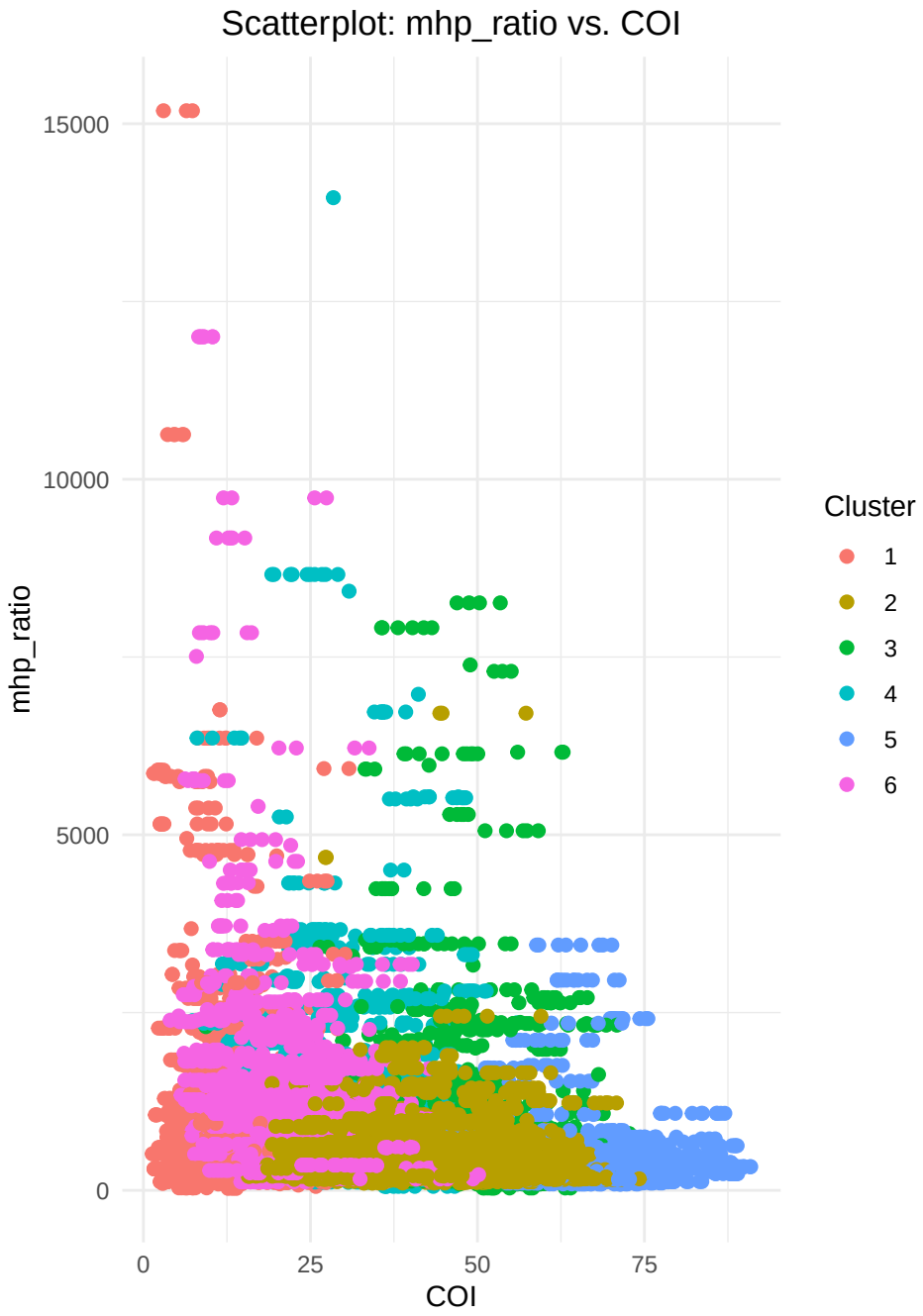

Scatterplot: percent_vaccinated vs. COI
percent_vaccinated
COI
0
20
40
60
0
25
50
75
Cluster
1
2
3
4
5
6

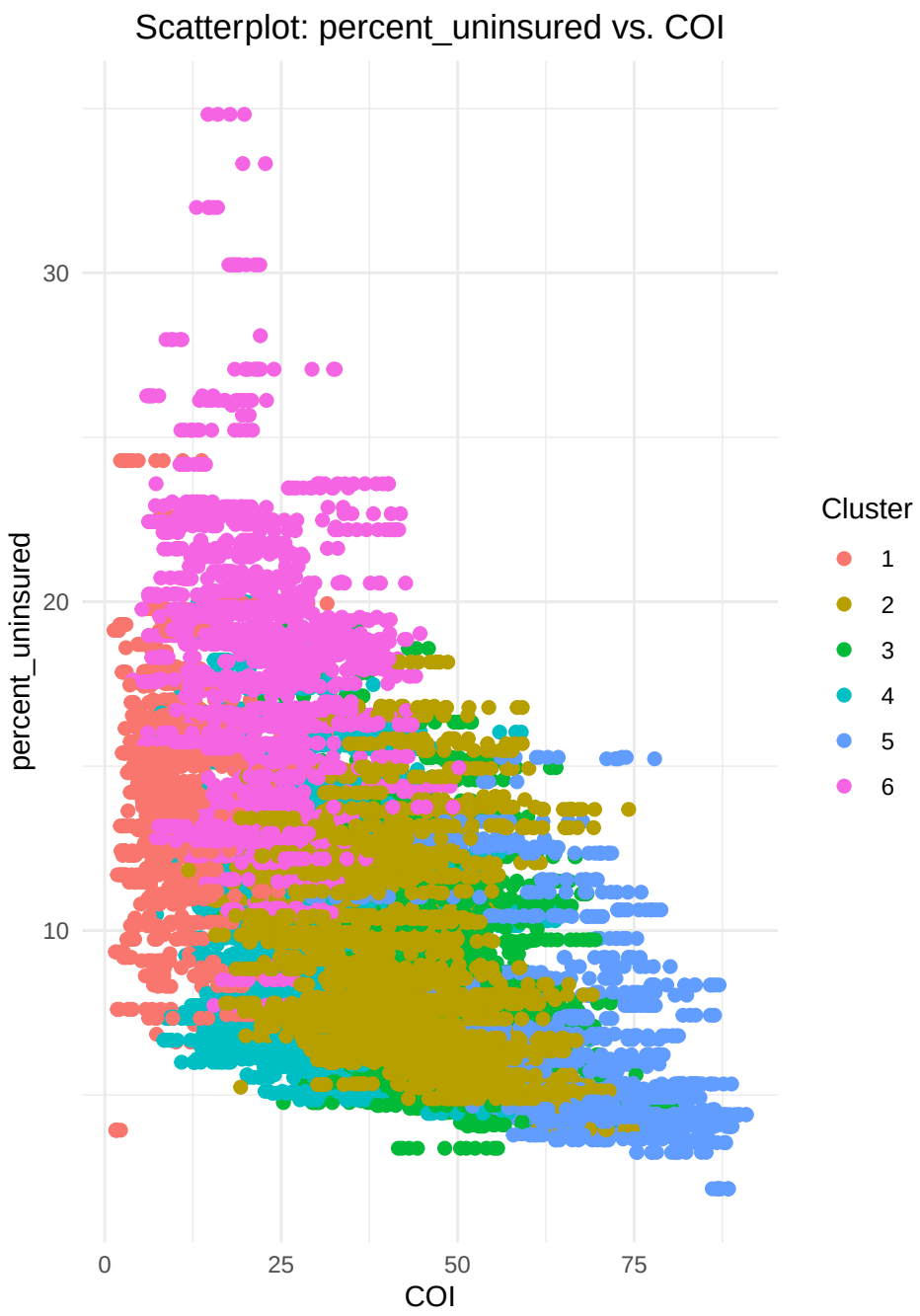

Scatterplot: median_household_income vs. COI
median_household_income
120000
80000
40000
0
25
50
75
COI
Cluster
1
2
3
4
5
6

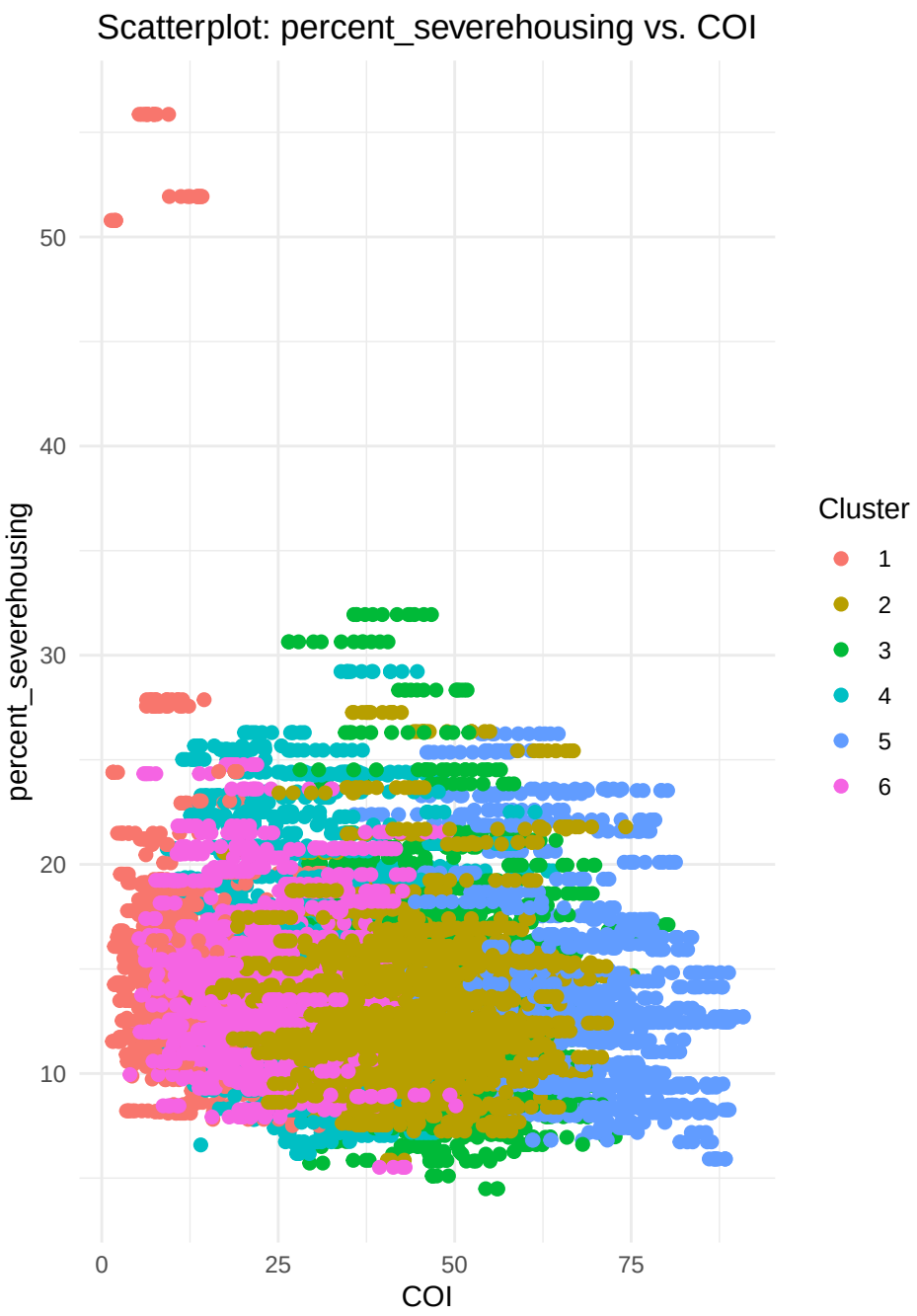

Scatterplot: avenum_phys_unhealthydays vs. COI
avenum_phys_unhealthydays
6
5
4
3
0
25
50
75
COI
Cluster
1
2
3
4
5
6

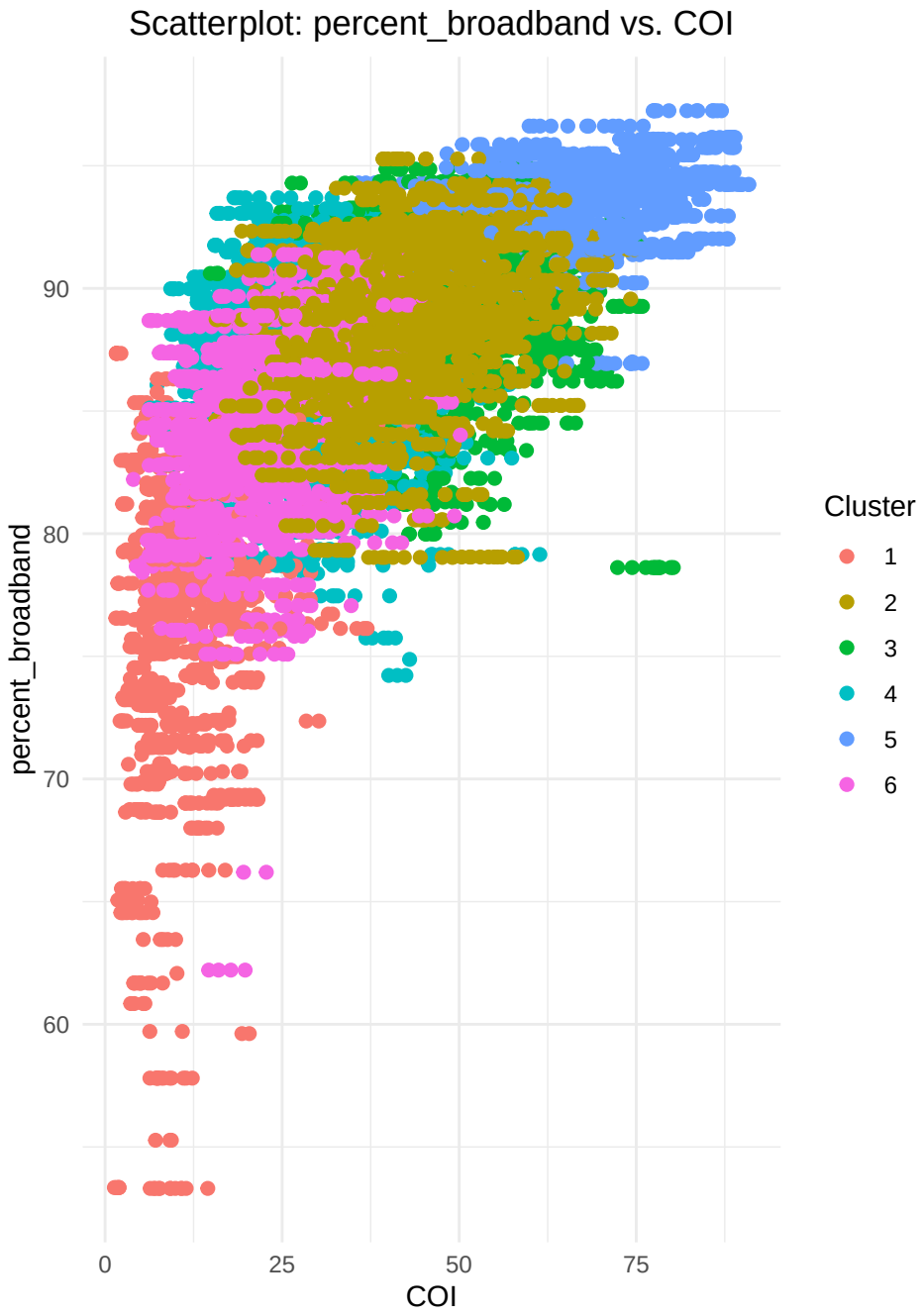

Scatterplot: avenum_ment_unhealthydays vs. COI
avenum_ment_unhealthydays
7
6
5
4
0
25
50
75
COI
Cluster
1
2
3
4
5
6


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
COI
0
10
20
30
40
0
25
50
75
Cluster
1
2
3
4
5
6

Initial K-Means Clustering Analysis

Scatter plots above clearly do nothing to define differentiated groupings nor help assess differences across groups.

Problem 1: Complete Convergence Failure

Clustering generated hundreds of "did not converge in 100 iterations" warnings, indicating the algorithm failed to stabilize. This renders all cluster assignments unreliable and non-reproducible. This is likely due to - Non-informative variables (county_fips, pop) included in distance calculations - No standardization applied despite variables ranging from 0-100 (percentages) to 0-25,000 (ratios) - Insufficient iterations (100) for 17-dimensional space - High dimensionality creates "curse of dimensionality" effects

Problem 2: Poor Cluster Separation

Scatter plots show heavily overlapping clusters across all variable pairs, indicating weak separation and low interpretability.This is likely due to - Variables on vastly different scales cause large-scale variables to dominate Euclidean distance calculations - High dimensionality (17 variables) weakens meaningful distance metrics - No cluster quality validation performed (silhouette width, Davies-Bouldin index)

Problem 3: Weak Interpretability

Current cluster descriptions lack actionable insights and quantifiable separation metrics.

K-Means Clustering Analysis

Interpretation and Insights from Cluster Analysis

The cluster analysis performed on the county-level Child Opportunity Index (COI) and health/socioeconomic indicators did **not yield clear, actionable or interpretable groupings**. Key technical and data issues hindered the identification of distinct clusters, substantially limiting potential insights for policy or intervention.

Description of Clusters and Attempted Interpretation

- Clustering was conducted using **K-means** on a broad set of standardized variables reflecting life expectancy, vaccination rates, income, health outcomes, and more.
- Feature selection and standardization were appropriately attempted to mitigate scale issues.
- The number of clusters (K) was selected using a **variance explained/asymptote** (elbow-type) approach, suggesting an optimal $K \approx 6$.

Key Observations from Cluster Output

- **Multiple warnings of non-convergence** in the K-means algorithm indicated instability and unreliable assignment of cluster labels.
- **High overlap of clusters** in scatterplots across all variable pairs showed a failure of the algorithm to find meaningful group separation.
- **No clear, distinguishable patterns** among clusters could be identified in terms of key variables such as COI, healthcare access, or economic measures.
- **Underlying Reasons for Failure:**
  - Continued presence of non-informative or highly correlated features.
  - High dimensionality with relatively little structure.
  - The variety in magnitude/range across input variables, despite attempted standardization, likely contributed to the curse of dimensionality.
  - Insufficient signal or inherent structure in the data to naturally group counties into distinct types.

What the Clusters Do NOT Show

- Clusters did **not map clearly to high/low COI, health, wealth, or resource divides**; cluster assignments are heavily overlapped and non-interpretable.
- Cluster labels cannot be reliably connected to actionable geographic, socioeconomic, or health profiles.

What Insights are Gained Despite Unsuccessful Clustering?

- **Current variables or selected data do not naturally segment counties into meaningful subgroups**, at least not in the space defined by K-means and these features.
- **Clustering is sensitive to feature selection, scaling, and algorithm choice**: future efforts may require
  - More targeted variable selection, perhaps focusing on a smaller, more interpretable set.
  - Dimensionality reduction (e.g., PCA before clustering).
  - Trying alternative clustering algorithms (e.g., hierarchical, DBSCAN).
- **Domain knowledge is crucial**: Automated clustering alone does not guarantee segments that are useful for intervention or narrative.

Recommendations for Future Clustering Efforts

- Conduct more thorough feature engineering and correlation analysis to reduce redundancy and noise.
- Consider using **dimensionality reduction** to find latent factors before re-applying clustering.
- Evaluate cluster solutions with **external validity metrics** (e.g., silhouette width).
- Triangulate unsupervised clustering with qualitative insights or geographic overlays to assist interpretation.

Conclusion

The attempted cluster analysis highlights important lessons regarding the **challenges of high-dimensional, mixed-data clustering in real-world policy data**. While the technical execution followed standard procedures, the lack of natural group structure in the data (and technical warnings) resulted in clusters that do **not provide actionable insights or interpretable segments** for understanding or addressing disparities in child opportunity or county health.

Compare

```
print(df_modelCompare)
```

```
                          Model OSR2_Phase3Report OSR2_this seed_this
1              Linear Regression             0.869        NA        NA
2                Regression Tree             0.757        NA        NA
3                  Random Forest             0.961        NA        NA
4  Artificial Neural Network ONE             0.822        NA      2511
5 Artificial Neural Network ZERO                NA        NA      2511
```

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