FINAL PROJECT REPORT

Introduction to Statistical Machine Learning

Student information:

Name	Rheina Trudy Williams
Student ID	2021380101

Northwestern Polytechnical University

Spring 2024

1. Introduction

Statistical machine learning develops models that learn from data and make predictions using statistical techniques. It merges machine learning techniques' computational efficiency and adaptability with statistical inferences and modeling capabilities.

Statistics forms the foundation for scientific analysis and provides a framework for the organization and interpretation of data. With statistics, researchers can identify patterns, trends, and relationships within datasets. This analytical capability in statistics is valuable for large and complex datasets, where summarizing and concluding data becomes critical. Machine learning, on the other hand, utilizes computational algorithms to enable computers to learn from data. Machine learning aims to develop adaptive models that can make data-driven predictions and improve their performance over time. Furthermore, these models generalize from specific examples to broader cases.

2. Goals

a. Preprocess data.
b. Utilize R language to analyze data.
c. Apply machine learning techniques to the data.
d. Evaluate model performance.
e. Interpret findings.

3. Tools and data

Rstudio
R 4.3.3
2018 County Health Rankings Data.xls

4. Theory overview

Several machine learning techniques can be used to analyze data:

a. Linear regression

A statistical technique for rank predictions, a linear modeling approach to quantify the relationship between a dependent variable and a set of independent variables. The principle of linear regression is to estimate the dependent variable's conditional probability distribution, given the independent variables' specific values.

b. Logistic regression

It is a method employed for classification tasks. The objective is to predict the likelihood of a data point belonging to a specific class. It uses statistical analysis to model the relationship between a binary dependent variable and one or more independent variables. This approach allows logistic regression to estimate the probability of an instance falling into a particular class.

c. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (K-NN) can handle various data types. It makes predictions for new data points by finding similar existing data points (neighbors) and basing the prediction on those neighbors' characteristics. This allows K-NN

to adapt to different patterns in the data without strict assumptions about the data's underlying structure.

d. Decision tree

A decision tree mimics a decision-making process like a flowchart. It starts with a question about the data, and then splits it into branches based on the answer. Each branch leads to another question or a final answer. This structure allows the tree to learn and predict by asking a series of questions about the data.

e. K-means clustering

K-means clustering is a technique for grouping similar data points. It begins by randomly selecting a set of K central points within the data space, called centroids. Each data point is then assigned to the closest centroid based on distance. Once all points are assigned, the centroids are recalculated to represent the center of their respective assigned points. This process of assigning points and recalculating centroids iterates until the centroids no longer significantly change, indicating that a stable grouping of the data has been achieved. This method assumes the number of clusters (K) is predetermined, and the goal is to assign each data point to one of these K groups.

f. Hierarchical clustering

Hierarchical clustering is a method for grouping data points into a hierarchy of clusters. It connects data points that are close together (based on a distance measure) into small groups. These small groups can then be further connected into larger clusters, forming a hierarchy of groups that reflect varying levels of similarity within the data.

5. Project content

In this project, I mainly analyze the ‘Outcomes & Factors Rankings’ data sheet using six different methods mentioned in the previous section.

1. Load data

The data is loaded from 2018 County Health Rankings Data.xls, renamed as health_data, focusing on Outcomes & Factors Rankings data sheet.

2. Data cleaning

The data was then renamed for easier accessibility.

- FIPS (Federal Information Processing Standard) -> FIPS

- State -> State
- County -> County
- # of Ranked Counties -> NumRankedCountries
- Health Outcomes Rank -> HealthOutcomesRank
- Health Outcomes Quartile -> HealthOutcomesQuartile
- Health Factors Rank -> HealthFactorsRank
- Health Factors Quartile -> HealthFactorsQuartile

Because NAs are introduced when factoring the data, the missing values are then erased from the data.

```
> health_data$HealthOutcomesRank <- as.numeric(health_data$HealthOutcomesRank)
Warning message:
NAs introduced by coercion
> health_data$HealthFactorsRank <- as.numeric(health_data$HealthFactorsRank)
Warning message:
NAs introduced by coercion
```

```
> # Remove rows with missing values in specific columns
> complete_rows <- complete.cases(health_data[, c("HealthOutcomesRank", "HealthFactorsRank")])
> health_data <- health_data[complete_rows, ]
> sum(is.na(health_data))
[1] 0
```

After ensuring there are no missing values, the data proceed to modelling and plotting.

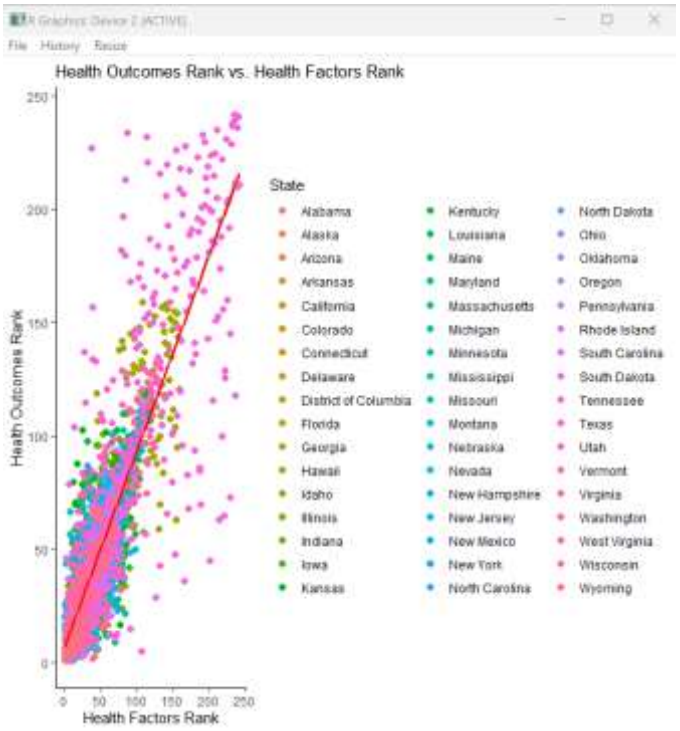
3. Modeling
   a. Linear regression

```
> lm_model <- lm(HealthOutcomesRank ~ HealthFactorsRank, data = health_data)
> summary(lm_model)

Call:
lm(formula = HealthOutcomesRank ~ HealthFactorsRank, data = health_data)

Residuals:
     Min       1Q   Median       3Q      Max
-135.283   -8.911   -1.978    7.930  187.685

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.451469   0.573055   11.26   <2e-16 ***
HealthFactorsRank 0.864833   0.009052   95.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.89 on 3076 degrees of freedom
Multiple R-squared:  0.7479,    Adjusted R-squared:  0.7479
F-statistic:  9127 on 1 and 3076 DF,  p-value: < 2.2e-16
```

In linear regression, the model is focused on analyzing Health Outcomes Rank and Health Factor Rank.

Based on the model summary, there is a strong negative relationship between the data. The coefficient for Health Factors Rank is positive (0.864833), indicating negative relationship because higher Health Factors Rank (better health factors) corresponds to a predicted decrease in Health Outcomes Rank (better health outcomes). The p-value for Health Factors Rank is less than 2.2e-16, indicating a high level of statistical significance. The R-squared value = 0.7479 means that the model explains 74.79% of the variance in Health Outcomes Rank after accounting for the model complexity.

- Plotting linear regression between Health Outcomes Rank and Health Factors Rank:



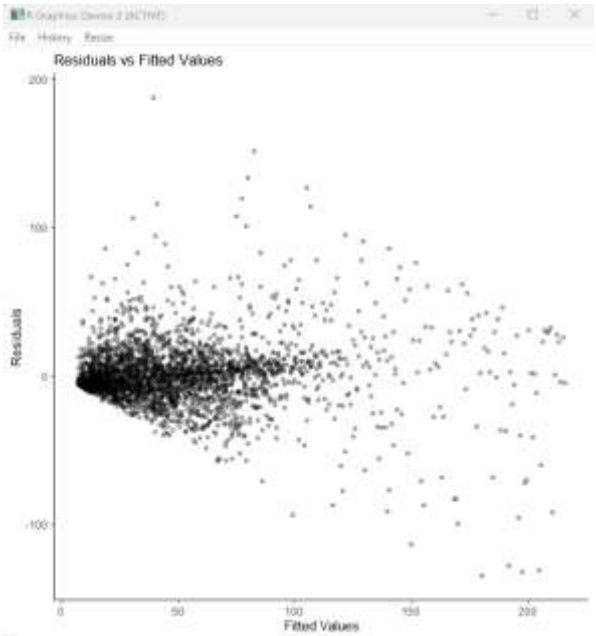
The data is then visualized through a scatter plot. There are outliers in the data, representing states with unique circumstances that influence their health outcomes rank despite their health factors rank. However, the data exhibits a huge scatter, indicating that there is variability in health outcomes even among states with similar health factor ranks.

After figuring the general relationships in the data, predictions between the Factor Rank and Outcomes Rank are tested.

```
> # Sample predictions for new data: rank
> new_data_lmpredict <- data.frame(HealthFactorsRank = c(20, 45, 70))
> predicted_outcomes <- predict(lm_model, new_data_lmpredict)
> predicted_outcomes
       1        2        3
23.74813 45.36896 66.98979
```

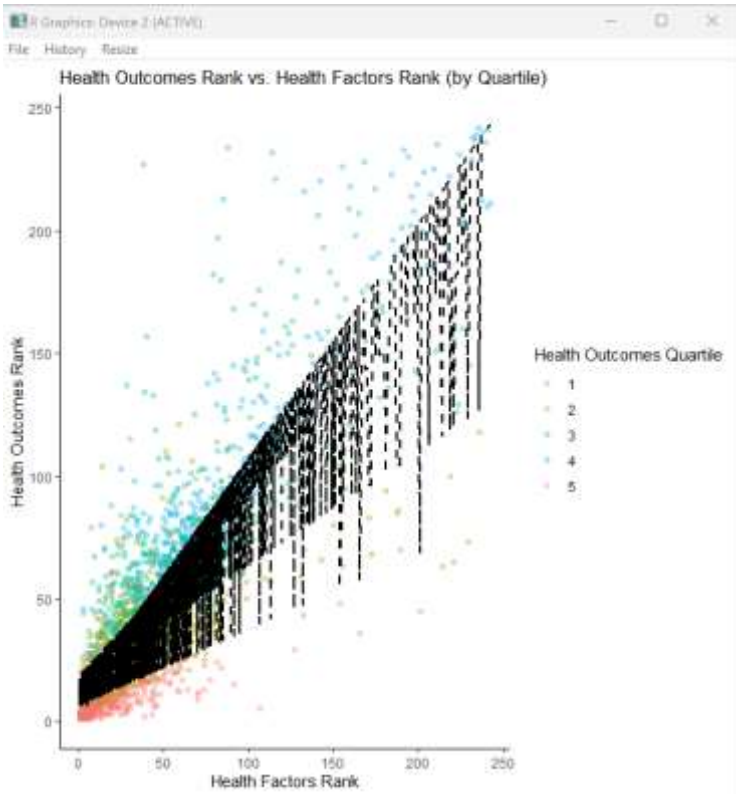
- Health Factors Rank of 20: The predicted Health Outcomes Rank is around 23.75. This suggests that for states/counties with a Health Factors Rank of 20, the model predicts a Health Outcomes Rank closer to the average.
- Health Factors Rank of 45: The predicted Health Outcomes Rank is around 45.37. This corresponds to a state/county with a mid-range Health Factors Rank, and the model predicts a Health Outcomes Rank near that value as well.
- Health Factors Rank of 70: The predicted Health Outcomes Rank is around 66.99. For states/counties with a high Health Factors Rank, the model predicts a lower Health Outcomes Rank.

• Plotting residuals and fitted values in the linear regression model:



In this plot, I tested the data for linear regression for residuals and fitted values comparison. It can be seen that the vertical spread of the residuals (y-axis) is not constant across the range of fitted values (x-axis). This means the variance of the errors might not be constant across the data.

- Plot of Health Outcomes Rank and Health Factors Rank by quartile:



The four colored lines represent the predicted health outcomes for each Health Outcomes Quartile (Quartile 1-4). The lines are roughly parallel, suggesting a negative association between Health Factors Rank and Health Outcomes Rank across all quartiles.

The dashed lines are predicted values. The data points are scattered around the predicted lines, indicating variability in the relationship between Health Factors Rank and Health Outcomes Rank. There might be other factors that affects the changes in rank, therefore this project also tests the data in classifications.

b. Logistic regression (multinomial)

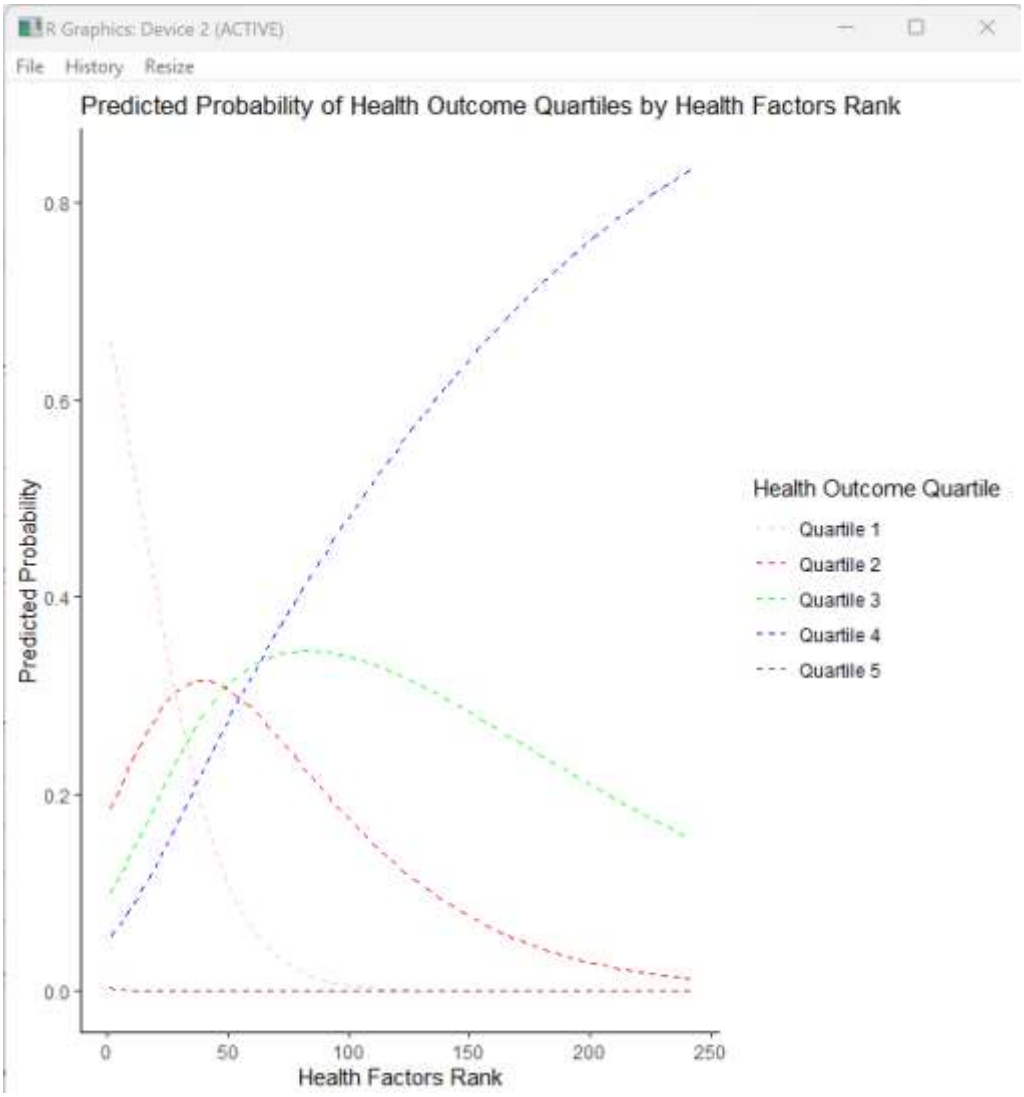
Multinomial logistic regression to find classification in data with more than two variables that are the quartiles. In this case, I set quartile 2-4, with 1 as the baseline and 5 as data with zero values.

```
> multinom_model <- multinom(HealthOutcomesQuartile ~ HealthFactorsRank, data = filtered_data)
# weights:  15 (8 variable)
initial  value 4953.849894
iter  10 value 3822.842368
iter  20 value 3717.559280
iter  30 value 3715.779850
final  value 3715.673745
converged
> summary(multinom_model)
Call:
multinom(formula = HealthOutcomesQuartile ~ HealthFactorsRank,
    data = filtered_data)

Coefficients:
  (Intercept) HealthFactorsRank
2   -1.323687        0.04709715
3   -1.962778        0.06012576
4   -2.556026        0.06953321
5   -4.948212       -0.18623524

Std. Errors:
  (Intercept) HealthFactorsRank
2  0.09052920       0.002832652
3  0.09824994       0.002858657
4  0.10679175       0.002896191
5  1.51559513       0.227254670

Residual Deviance: 7431.347
AIC: 7447.347
```

The model aims to estimate coefficients for each level of HealthOutcomesQuartile compared to the intercept. The coefficients for HealthFactorsRank are all negative, suggesting that as Health Factors Rank increases, the estimated coefficients become more negative. Although the interpretation of coefficients is limited to this, the negative signs suggest that a higher Health Factors Rank is associated with a decrease in the log odds of being in higher Health Outcomes Quartiles. This aligns with the findings from the linear regression model.

- Plot of predicted probability of Health Outcome Quartiles based on the Health Factors Rank:
  In this plot, I showed all the quartile, including quartile 5 (zero values quartile) and quartile 1 (the base line).
  Quartile 1: As Health Factors Rank increases, the probability of Health Outcome Quartiles decreases.
  Quartile 2: As Health Factors Rank increases, the probability of Health Outcome Quartiles decreases. However, there is an increase nearing Health Factors Rank 50. This emphasizes the trend where the

distribution shifts towards Quartile 2 (signifying average health outcomes) as Health Factors Rank approaches 50.
Quartile 3: As Health Factors Rank increases, the probability of Health Outcome Quartiles decreases. However, there is an increase nearing Health Factors Rank 50-100. This shows a trend of Health Outcome Quartiles for the range 50-100 for Quartile 3.
Quartile 4: As Health Factors Rank increases, the probability of Health Outcome Quartiles increases.



c. K-Nearest Neighbors (KNN)

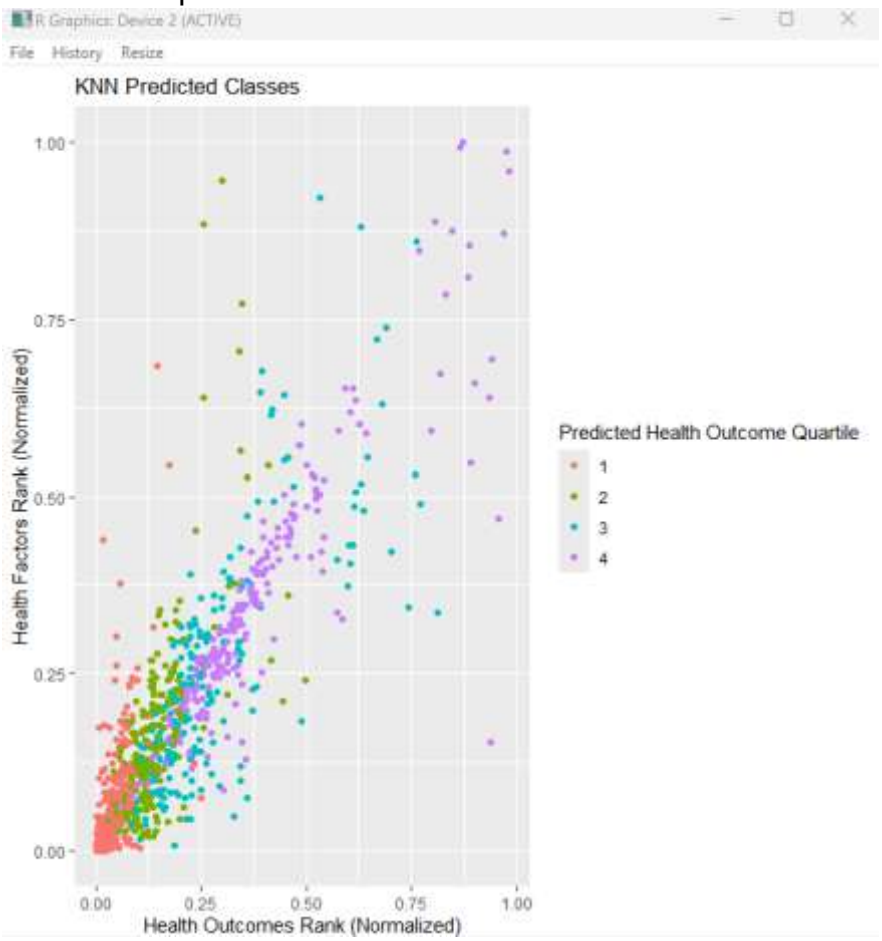
From the model, each quartile represents a range of health outcomes, with quartile 1 typically indicating the most favorable outcomes and quartile 3 indicating the least favorable. Quartile 1 has the highest count of predictions (279), followed by quartile 4 (237), while quartiles 2 and 3 have lower counts (205 and 203, respectively).

Quartile 1 having the highest count of predictions suggests that there is a significant proportion of counties where the health factors correspond to better health outcomes.

```
knn_pred      1   2   3   4   5  NR Quartile
  1         187  53  22  16   1   0        0
  2          40  87  50  28   0   0        0
  3           4  56  86  57   0   0        0
  4           1  29  66 141   0   0        0
  5           0   0   0   0   0   0        0
  NR          0   0   0   0   0   0        0
  Quartile    0   0   0   0   0   0        0
> summary(knn_pred)
       1        2        3        4        5       NR Quartile
     279      205      203      237        0        0        0
```

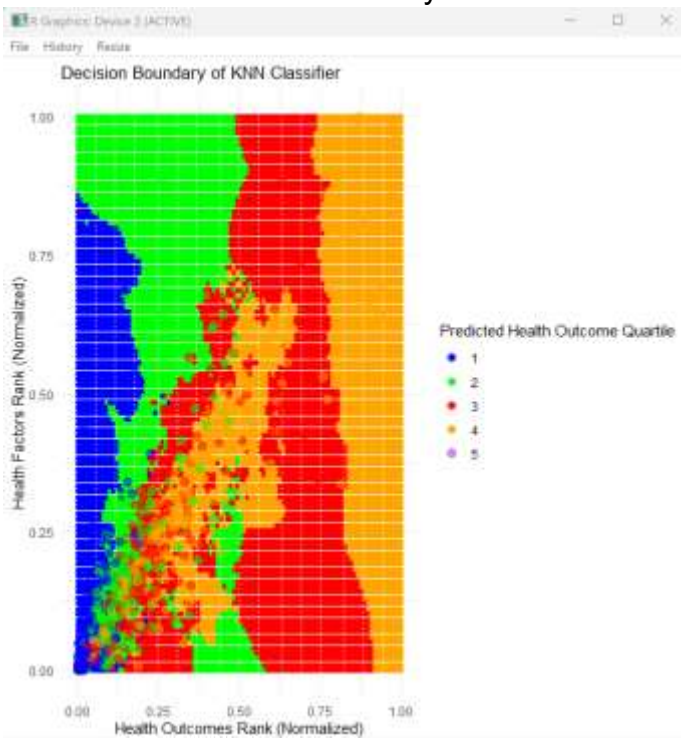
Above the summary is a contingency matrix comparing the predicted health outcome quartiles (rows) generated by the KNN model with the actual health outcome quartiles (columns) for the unseen test data. The diagonal elements of the table represent the counts where the predicted quartile matches the actual quartile. This table shows the correctness of the model as the value of 187 indicates that 187 data points were correctly classified as quartile 1. Some misclassifications can be found in off-diagonal elements. For example, in the first row, the column corresponding to quartile 2 (value of 53) indicates that 53 data points that belonged to quartile 2 were misclassified as quartile 1 by the KNN model.

- Plot of KNN predicted classes:



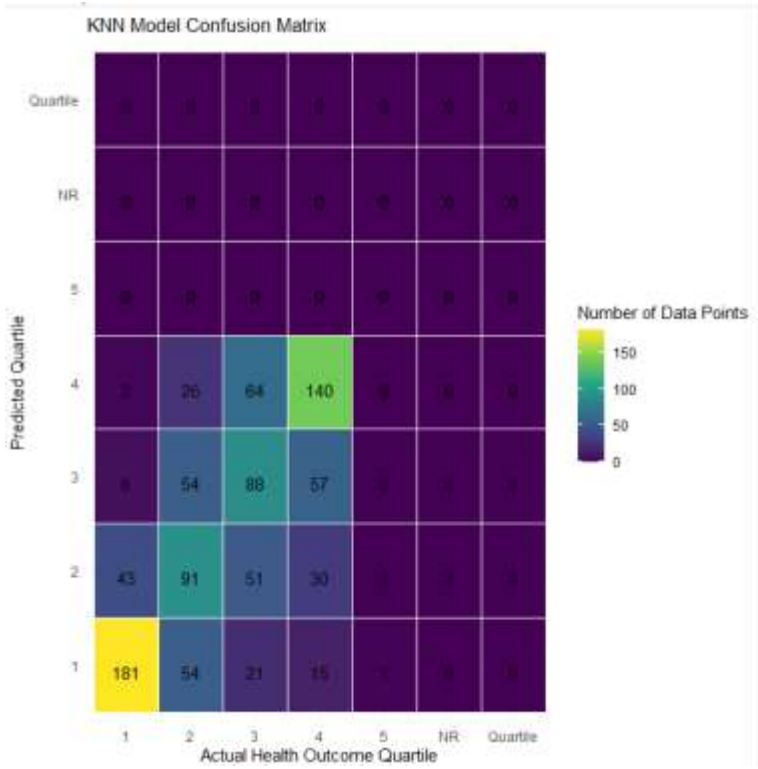
There seems to be a partial separation between the points represented in Quartiles 1 to 4. Quartile 2 appears separated from Quartiles 1 and 3, while Quartile 4 forms a distinct cluster in the upper right area. There is also overlap between the clusters, between Quartiles 1, 2, and 3. This indicates that some states/counties in these quartiles share similar characteristics in the underlying health factors data, showing lower to average health outcomes.

- Plot of KNN decision boundary



The decision boundary separates the data space into regions associated with predicted Health Outcomes Quartiles (1 to 4). The decision boundaries between Quartiles 1, 2, and 3 are less distinct compared to the boundary between Quartile 4 and the rest. This shows that states/counties in Quartiles 1, 2, and 3 might be harder to differentiate based on health factors alone. There are chances of some overlap in the health factor characteristics.

- Plot of KNN confusion matrix in heat map:



- Quartile 2: The model performs best for Quartile 2 (green) with a high chance of correct classifications (279) and relatively low misclassifications.
- Quartile 1, 3, and 4: The model struggles more with these quartiles. There are several misclassifications for Quartile 1 (89), Quartile 3 (102), and Quartile 4 (87).

The confusion matrix reveals that the KNN model is uneven across health outcome quartiles. Although it is good at classifying Quartile 2, it has difficulty distinguishing between Quartile 1, 3, and 4, showing limitations in using health factors alone to predict these quartiles.

d. Decision tree

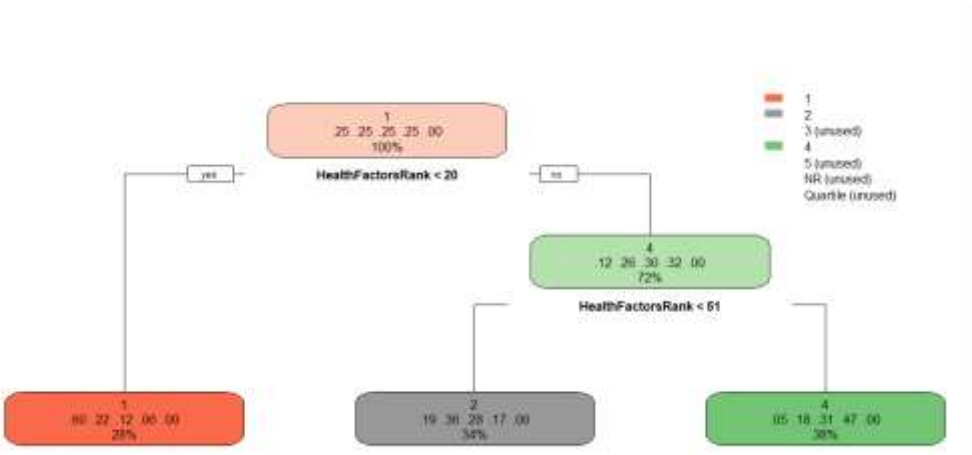
```
> tree_model <- rpart(HealthOutcomesQuartile ~ HealthFactorsRank, data = health_data, method = "clas
s")
> print(tree_model)
n= 3078

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 3078 2305 1 (0.25 0.25 0.25 0.25 0.00032)
  2) HealthFactorsRank< 19.5 866  348 1 (0.6 0.22 0.12 0.062 0.0012) *
  3) HealthFactorsRank>=19.5 2212 1495 4 (0.12 0.26 0.3 0.32 0)
    6) HealthFactorsRank< 50.5 1051  673 2 (0.19 0.36 0.28 0.17 0) *
    7) HealthFactorsRank>=50.5 1161  621 4 (0.048 0.18 0.31 0.47 0) *
```

```
> summary(tree_model)
Call:
rpart(formula = HealthOutcomesQuartile ~ HealthFactorsRank, data = health_data,
    method = "class")
  n= 3078

          CP nsplit rel error    xerror       xstd
1 0.20043384      0 1.0000000 1.0355748 0.01004292
2 0.08720174      1 0.7995662 0.8004338 0.01179436
3 0.01000000      2 0.7123644 0.7284165 0.01198475

Variable importance
HealthFactorsRank
              100

Node number 1: 3078 observations,    complexity param=0.2004338
  predicted class=1  expected loss=0.7488629  P(node) =1
    class counts:   773   772   761   771     1
   probabilities: 0.251 0.251 0.247 0.250 0.000
  left son=2 (866 obs) right son=3 (2212 obs)
  Primary splits:
      HealthFactorsRank < 19.5 to the left,  improve=208.7872, (0 missing)

Node number 2: 866 observations
  predicted class=1  expected loss=0.4018476  P(node) =0.2813515
    class counts:   518   190   103    54     1
   probabilities: 0.598 0.219 0.119 0.062 0.001

Node number 3: 2212 observations,    complexity param=0.08720174
  predicted class=4  expected loss=0.675859  P(node) =0.7186485
    class counts:   255   582   658   717     0
   probabilities: 0.115 0.263 0.297 0.324 0.000
  left son=6 (1051 obs) right son=7 (1161 obs)
  Primary splits:
      HealthFactorsRank < 50.5 to the left,  improve=78.65508, (0 missing)

Node number 6: 1051 observations
  predicted class=2  expected loss=0.6403425  P(node) =0.3414555
    class counts:   199   378   297   177     0
   probabilities: 0.189 0.360 0.283 0.168 0.000

Node number 7: 1161 observations
  predicted class=4  expected loss=0.5348837  P(node) =0.377193
    class counts:    56   204   361   540     0
   probabilities: 0.048 0.176 0.311 0.465 0.000
```

The decision tree confirms a negative relationship between Health Factors Rank and Health Outcomes Rank, aligns from the previous models. The tree splits the data based on Health Factors Rank at two key thresholds:

- First split: Health Factors Rank < 19.5 separates observations into two branches. Those below 19.5 are more likely to have worse health outcomes (Quartile 1).
- Second split (for observations with Health Factors Rank >= 19.5): Health Factors Rank < 50.5 further separates them. Those below 50.5 are more likely to have intermediate health outcomes (Quartile 2 or 3).

- Plot of decision tree:



The plot shows the distribution of Health Factors Rank (x-axis) within each Health Outcomes Quartile (1 to 4). For lower Health Factors Ranks, Quartile 4 has the highest proportion of observations. This suggests that states/counties with worse health factors are more likely to have worse health outcomes (Quartile 4). As Health Factors Rank increases the proportion of observations in Quartile 4 gradually decreases. There is some overlap in the distribution of Health Factors Rank across Quartiles 1 to 3 This indicates that even within a health outcome quartile, there is a variation in health factors.

e. K-means clustering

```
> kmeans_model <- kmeans(health_data_norm, centers = 4)
> health_data$kmeans_cluster <- as.factor(kmeans_model$cluster)
> table(health_data$kmeans_cluster)

   1    2    3    4
1099 1397  118  464
> summary(kmeans_model)
             Length Class  Mode
cluster      3078   -none- numeric
centers         8   -none- numeric
totss           1   -none- numeric
withinss        4   -none- numeric
tot.withinss    1   -none- numeric
betweenss       1   -none- numeric
size            4   -none- numeric
iter            1   -none- numeric
ifault          1   -none- numeric
```

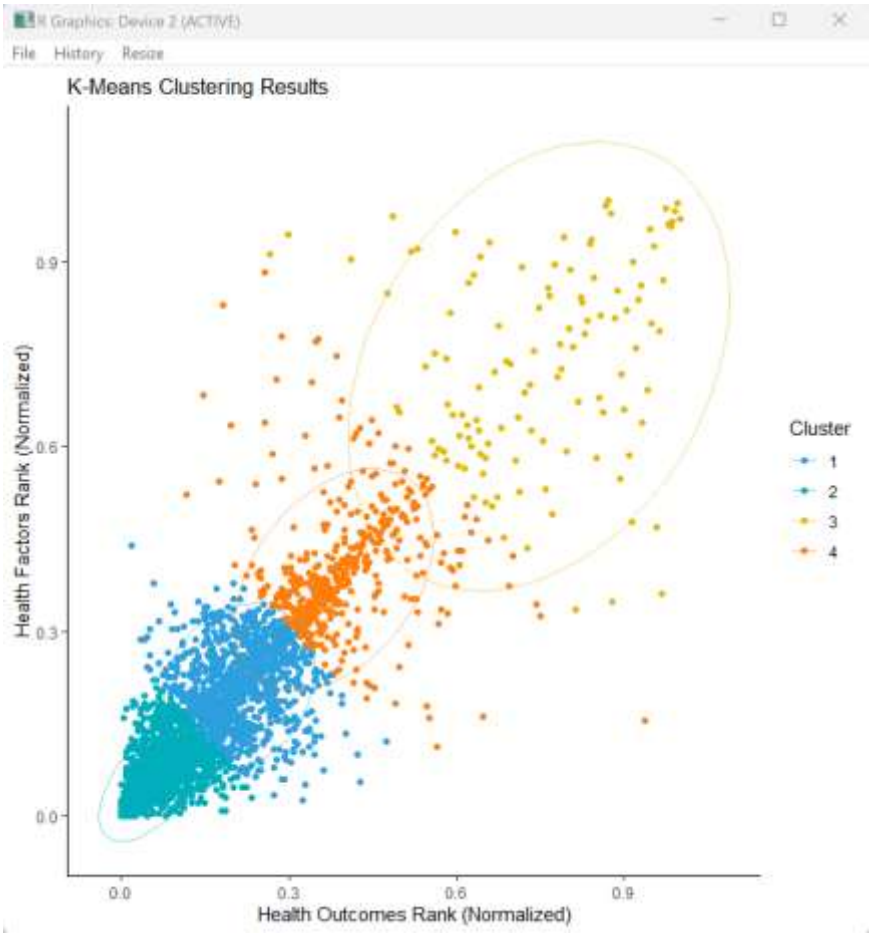
The table displays the number of observations (states/counties) assigned to each cluster (1, 2, 3, and 4) by the K-means algorithm.

- Plot of K-means clustering:

  Based on the plot, there are 4 clusters with cluster 4 as the widest cluster, followed by cluster 3, cluster 2, and cluster 1. The earlier clusters are more concentrated than the later clusters. This shows that there is more variation in distribution for the later cluster.

  This plot suggests that the data can be clustered into 4 groups. The first group has low Health Outcomes Rank and low Health Factors Rank, groups two and three have average Health Outcomes Rank and average Health Factors Rank, and group 4 with high Health Outcomes Rank has varieties in Health Outcomes Factors. Majority of the counties/regions are in the low-

average rank while the counties/regions that are in the higher outcomes rank has different factors rank.



f. Hierarchical clustering

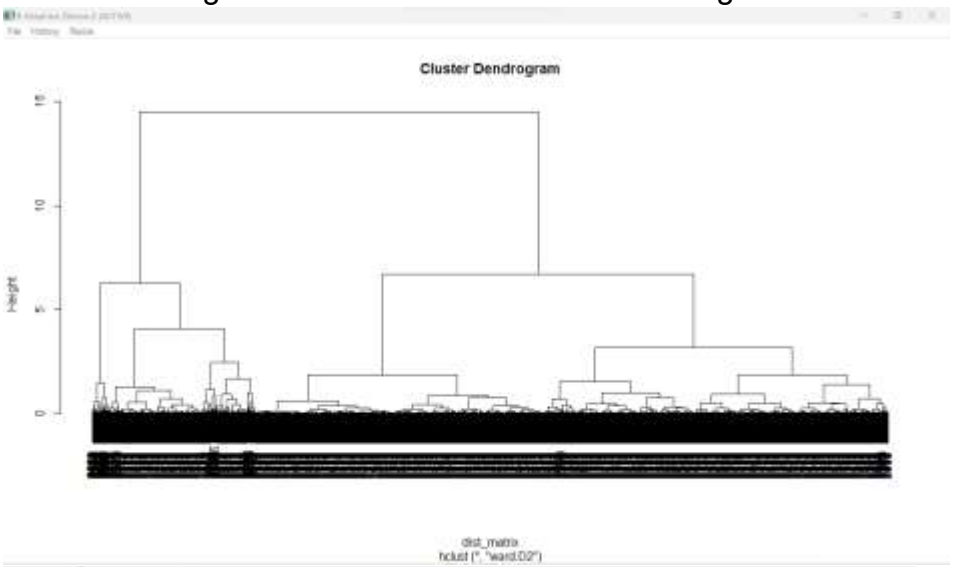
```
> summary(hclust_model)
            Length Class  Mode
merge       6154   -none- numeric
height      3077   -none- numeric
order       3078   -none- numeric
labels         0   -none- NULL
method         1   -none- character
call           3   -none- call
dist.method    1   -none- character
```

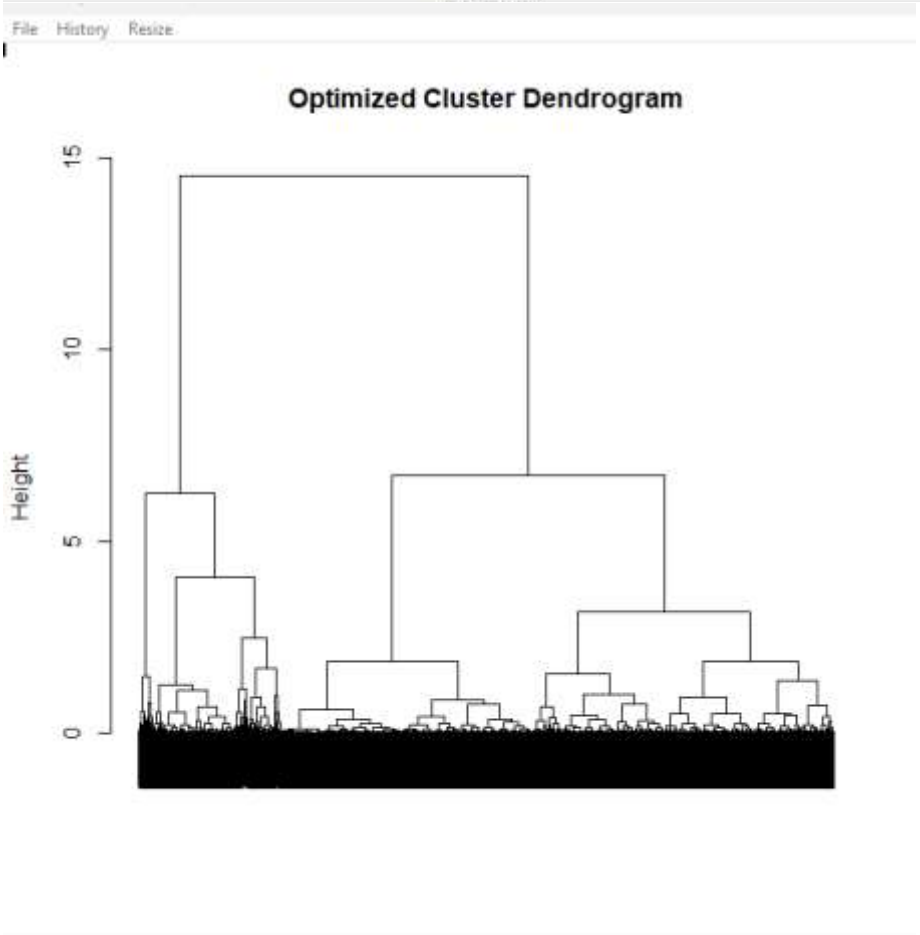
```
> print(hclust_model)

Call:
hclust(d = dist_matrix, method = "ward.D2")

Cluster method   : ward.D2
Distance         : euclidean
Number of objects: 3078
```

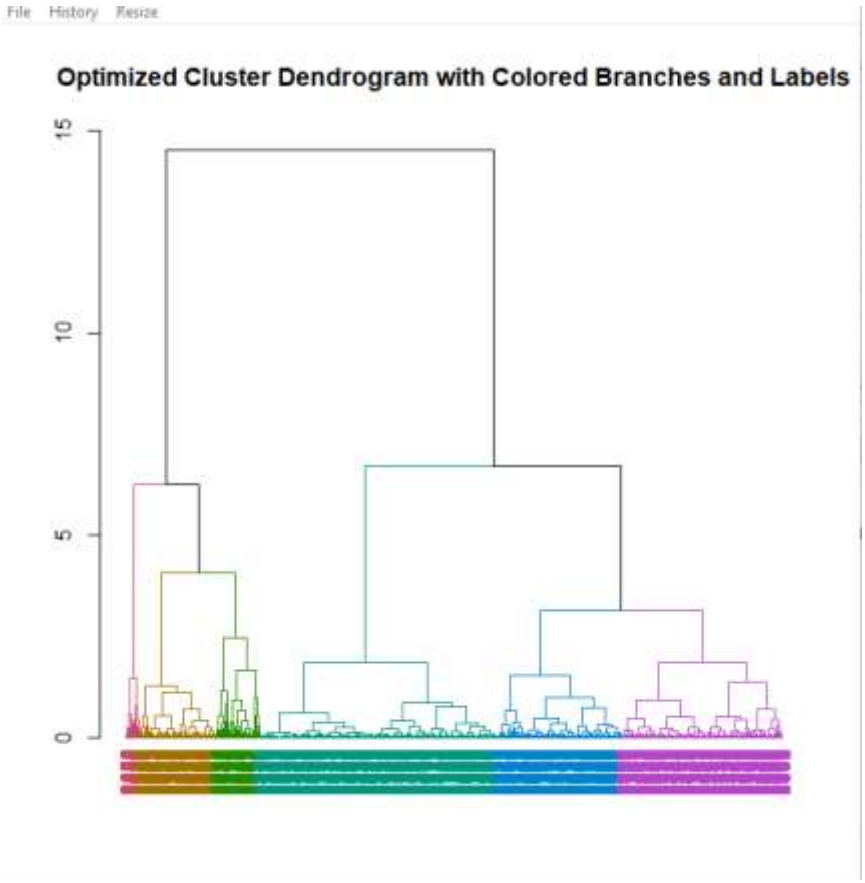
The model was built using Ward's minimum variance method (ward.D2) for clustering, minimizing the total within-cluster variance when forming clusters by recursively merging clusters that minimize the increase in total within-cluster variance. The distance metric used in this

model is Euclidean distance, measuring the straight-line distance between two data points in a multidimensional space.

- Plot of dendrogram based on hierarchical clustering:







For hierarchical clustering, I visualized two kinds of dendrograms:

- Original dendrogram

In the plot, multiple clusters exist because of the large data. It can be seen there are 4 big clusters with cluster 3 and cluster 4 owning the most instances.

- Optimized dendrogram

In the optimized dendrogram, I tried to compact the dendrogram into 60% of the original size. Although the data still cannot be observed specifically, clearer clusters can be seen. I did rectangle grouping based on the data spread and colored them for readability. From this dendrogram, 6 huge clusters are grouped with their similarity based on the heights. It also shows that the red group and light green group have the most similar features.

4. Results
   - Both linear regression and logistic regression analyses indicate a negative relationship between Health Outcomes Rank and Health Factors Rank. Higher Health Factors Rank corresponds to lower Health Outcomes Rank, suggesting that counties with better health factors tend to have better health outcomes. This relationship is consistent across different modeling approaches, showing the correct prediction.
   - The models, particularly KNN, show varying performance across health outcome quartiles. While Quartile 1 predictions are most accurate,

Quartiles 2, 3, and 4 present challenges, indicating limitations in predicting health outcomes solely based on health factors.

- Decision tree analysis reveals clear thresholds in the Health Factors Rank that separate counties into different Health Outcomes Quartiles. Similarly, K-means and hierarchical clustering identify distinct clusters based on Health Outcomes and Factors Rank. These methods give analysis into underlying patterns and groupings within the data, enabling a deeper understanding of county-level health disparities and similarities.
- The hierarchical clustering analysis reveals the presence of distinct clusters within the dataset. By visualizing the dendrogram, we can observe the grouping of counties based on similarities in health outcomes and factors.

6. Conclusion

In conclusion, this project aimed to employ statistical machine-learning techniques to analyze county-level health data and figure out the relationship between health outcomes and factors. The data are observed through analysis using linear regression, logistic regression, K-Nearest Neighbors (KNN), decision tree, K-means clustering, and hierarchical clustering. The results consistently showed a negative relationship between Health Outcomes Rank and Health Factors Rank, indicating that counties with better health factors tend to have better health outcomes. While this relationship was reliable in different modeling approaches, varying performance was observed across different health outcome quartiles, highlighting the complexity of predicting health outcomes solely based on health factors.

During the project, I encountered some difficulties to adjust the correct model for the data. At first, I want'sed to do polynomial regression, however, the result from the linear regression shows that clustering is the best method for this data. Othe difficulties that I encountered were syntax errors and plotting design. I fix them by browsing through the R stack forum and the lecture slides. Finally, I finish this project successfully.