Animating Question

The primary question guiding this research is: How do the leading causes of death in New York City vary by race and ethnicity across different age groups?

Data Source

The data for this project was sourced from the NYC Leading Causes of Death Dataset, available through the NYC Open Data portal. This dataset provides comprehensive information on the leading causes of death, categorized by race, ethnicity, age group, and sex, offering a rich basis for detailed analysis.

New York City Department of Health and Mental Hygiene. (n.d.). New York City Leading Causes of Death. NYC Open Data. Retrieved April 25, 2024, from https://data.cityofnewyork.us/Health/New-York-City-Leading-Causes-of-Death/jb7j-dtam

Validation Strategies Employed

Cross-validation was employed as the primary validation strategy to ensure the robustness of the results. This method involves partitioning the data into multiple subsets, training the model on some subsets while validating it on the remaining ones. This process is repeated several times, and the results are averaged to provide a reliable measure of model performance.

Supervised Techniques Used

Two supervised learning techniques were employed which are logistic regression and random forest classifiers. Logistic regression is a statistical method that models the probability of a certain class or event, such as the leading cause of death, based on one or more predictors. Random forest is a joint learning method that builds multiple decision trees, merges their results to improve prediction accuracy, and controls over-fitting.

Logistic regression is a generalized linear model used for binary or multinomial classification. It estimates the parameters of a logistic model using maximum likelihood. Random forest is an ensemble method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. Each tree in a random forest is trained on a random subset of the data, and a random subset of features is considered for splitting at each node.

Logistic Regression

The logistic regression model achieved an accuracy of approximately 42.75%.

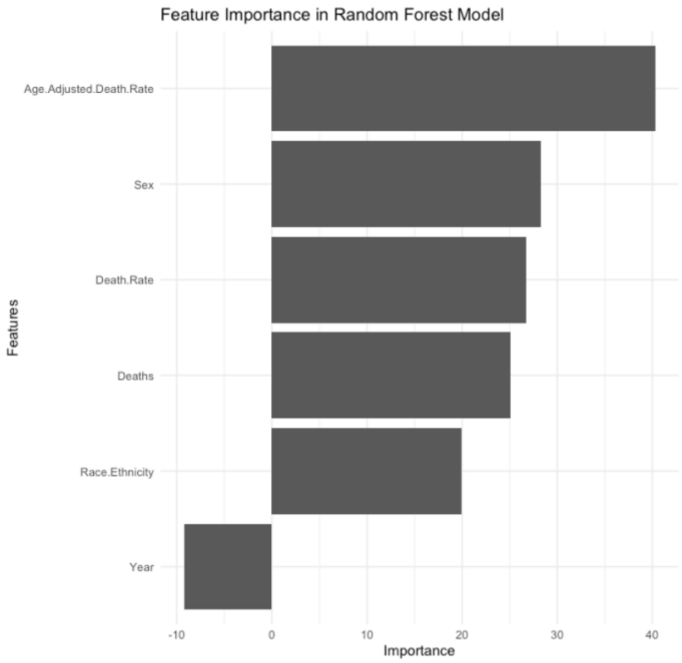
Random Forest

The random forest model initially achieved an accuracy of approximately 61.83%.

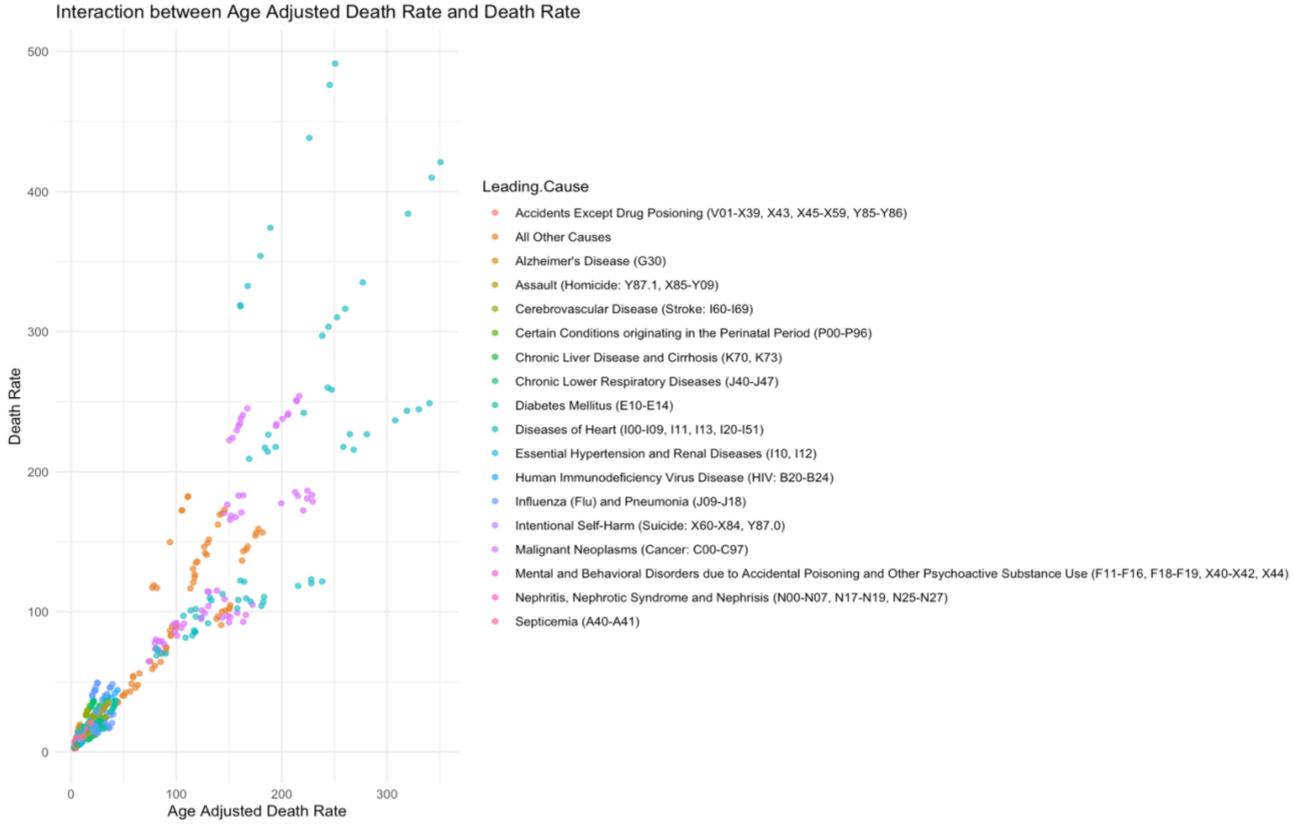
The results suggest that random forest classifiers are more effective than logistic regression for this dataset, providing higher accuracy in predicting the leading causes of death. This indicates that complex, non-linear relationships between variables are better captured by the ensemble method.

The model accuracies were validated using 10-fold cross-validation, a technique that helps mitigate overfitting by ensuring the model performs well on undetected data. This method divides the data into 10 subsets, trains the model on nine subsets, and validates it on the remaining one, repeating the process 10 times.

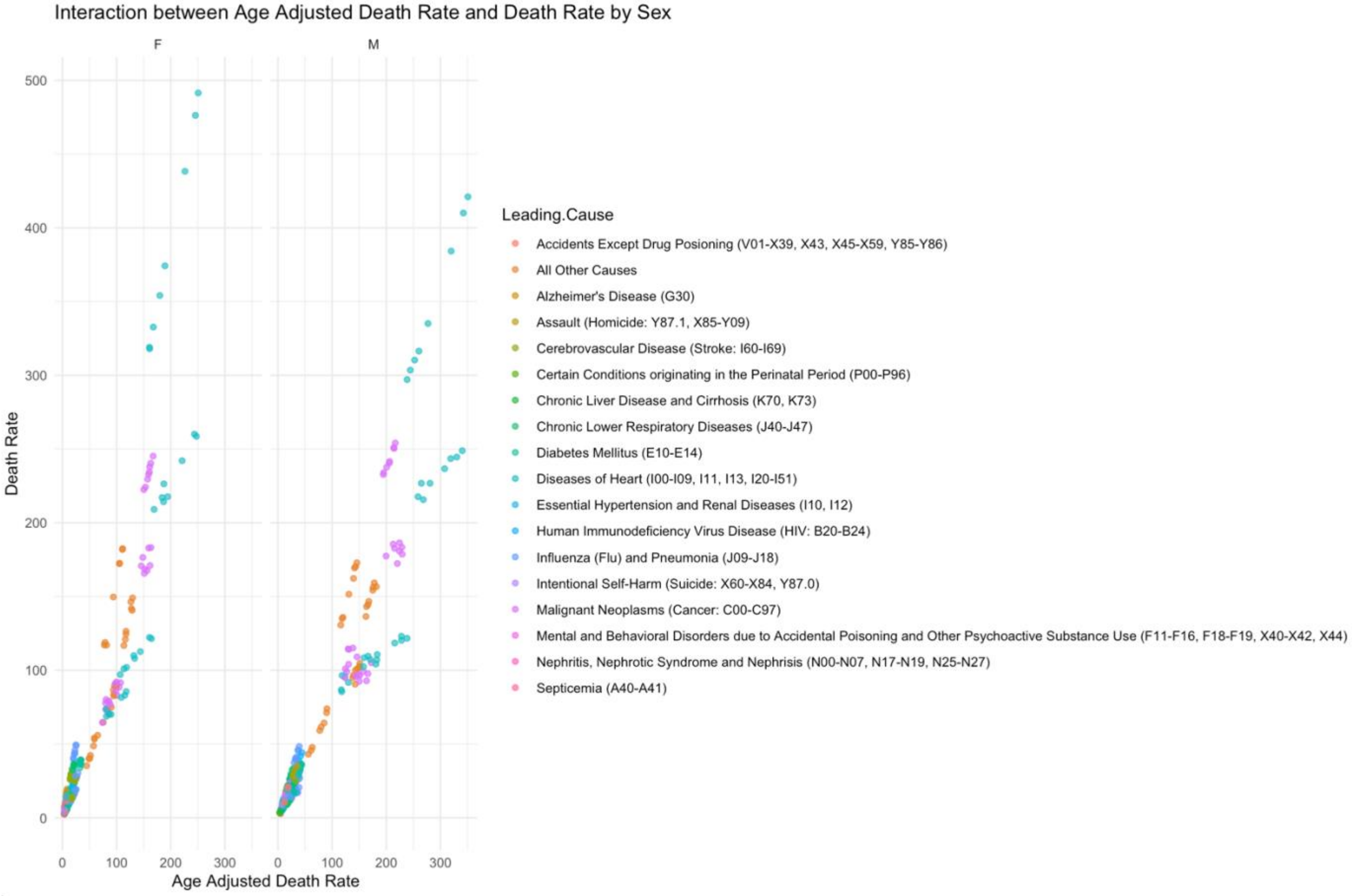
Feature Importance in Tuned Random Forest Model



Interaction between Age Adjusted Death Rate and Death Rate



Interaction between Age Adjusted Death Rate and Death Rate by Sex

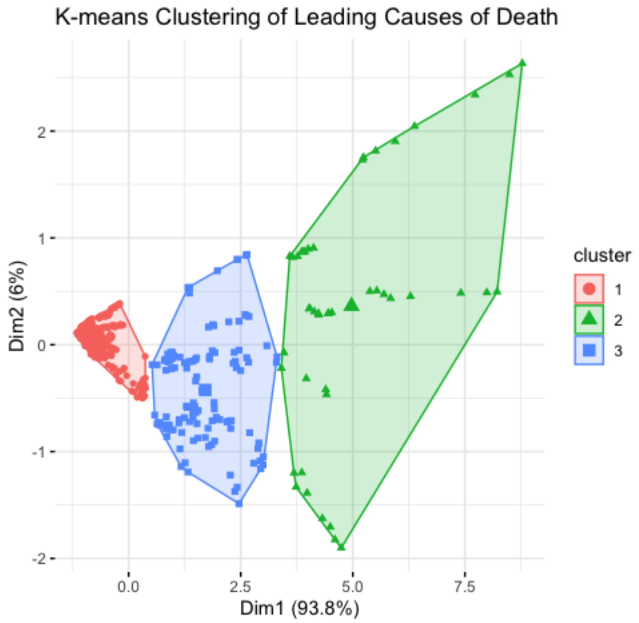


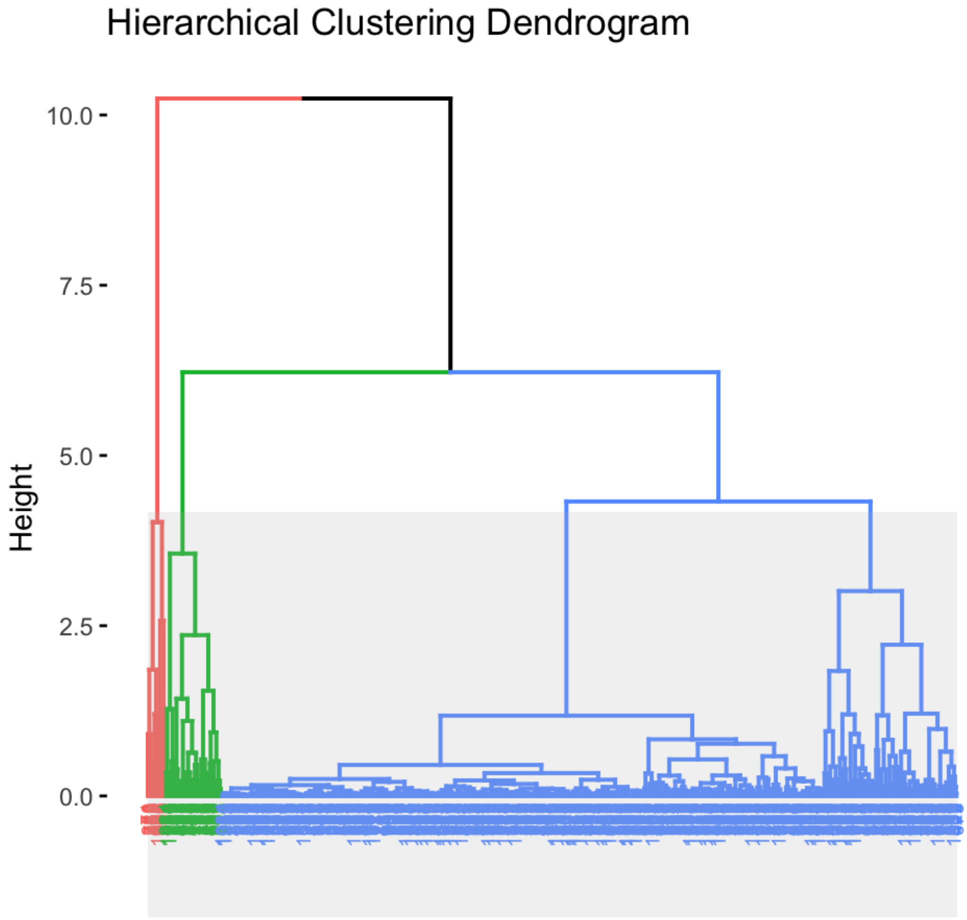
Unsupervised Techniques Used

Clustering techniques, such as K-means and hierarchical clustering, were employed to group similar causes of death and discover underlying patterns. K-means clustering partitions the data into a specified number of clusters, optimizing the assignment of points to clusters to minimize the variance within each cluster. Hierarchical clustering builds a tree-like structure (dendrogram) of the data points, where each node represents a cluster, and merges clusters based on their similarity.

K-means clustering initializes a set of cluster centroids and iteratively reassigns points to the nearest centroid and recalculates centroids until convergence. Hierarchical clustering computes a dissimilarity matrix and uses linkage methods (e.g., complete linkage) to iteratively merge the most similar clusters.

Results





The clustering analysis reveals distinct groups of leading causes of death that share similar characteristics, such as death rates and age-adjusted death rates. This helps in understanding commonalities and differences among various causes of death, which can inform targeted public health interventions. For example, identifying clusters that represent higher risks for specific demographic groups can help design focused health campaigns and allocate resources more effectively.

Cluster memberships were validated by examining the consistency of clusters formed by different methods (K-means and hierarchical clustering). The number of clusters was chosen based on the visual inspection of the dendrogram and the elbow method for K-means, ensuring that the selected number of clusters appropriately represents the data structure without overfitting or underfitting.

Conclusions

The study reveals significant differences in the leading causes of death across different racial and ethnic groups, influenced by factors such as sex and age-adjusted death rates. Random forest models proved to be more effective in capturing these patterns compared to logistic regression. K-means and hierarchical clustering, revealed distinct groups of causes of death with similar characteristics, such as death rates and age-adjusted death rates. This clustering highlighted significant differences across demographic groups, underscoring the need for targeted public health interventions. The study's results emphasize the importance of addressing specific health risks faced by different populations, informing more focused health policies and resource allocation.

The dataset may not capture all variables that influence death rates, such as socioeconomic factors, healthcare resources, lifestyle, and environmental influences. The supervised learning models, particularly logistic regression and random forest, have inherent assumptions and limitations. Logistic regression assuming a linear relationship between predictors and the log odds of the outcome. While clustering techniques effectively group similar causes of death, the interpretation of these clusters can be subjective, and the choice of the number of clusters and the linkage method in hierarchical clustering can influence the results.

Suggestions for Future Study

- Unsupervised Learning: Apply clustering techniques to identify patterns and subgroups within the dataset.
- Detailed Analysis: Perform a detailed analysis on specific clusters or outliers to understand underlying causes.
- Policy Recommendations: Develop specific policy recommendations based on the insights gained from the analysis.

Appendix

1. Software and Tools:
    a. R Programming Language: For data analysis, modeling, and visualization.
        i. R Libraries: Tidyverse, caret, randomForest, factoextra, and cluster.
2. Generative AI Technology:
    a. ChatGPT by OpenAI: Assisted in debugging R code.

3. R-Output:

```
library(tidyverse)
library(caret)
library(randomForest)
library(nnet)

data <- read.csv("/Users/kamala/Desktop/New_York_City_Leading_Causes_of_Death_20240425.csv")

# Data Preprocessing
data$Sex <- as.factor(data$Sex)
data$Race.Ethnicity <- as.factor(data$Race.Ethnicity)
data$Leading.Cause <- as.factor(data$Leading.Cause)

# Convert character variables to numeric
data$Deaths <- as.numeric(data$Deaths)
data$Death.Rate <- as.numeric(data$Death.Rate)
data$Age.Adjusted.Death.Rate <- as.numeric(data$Age.Adjusted.Death.Rate)

# Remove rows with missing values
data <- na.omit(data)

# Remove factor levels with no data in either training or testing set
factor_levels <- levels(data$Leading.Cause)
valid_levels <- factor_levels[sapply(factor_levels, function(x) sum(data$Leading.Cause == x)) > 1]
data <- data[data$Leading.Cause %in% valid_levels,]
data$Leading.Cause <- factor(data$Leading.Cause)  # Reset factor levels

# Split the data into training and testing sets
set.seed(123)
trainIndex <- createDataPartition(data$Leading.Cause, p = .8,
                                  list = FALSE,
                                  times = 1)
dataTrain <- data[trainIndex,]
dataTest <- data[-trainIndex,]

# Logistic Regression Model
log_model <- train(Leading.Cause ~ ., data = dataTrain, method = "multinom",
                   trControl = trainControl(method = "cv", number = 10))
```

```
log_pred <- predict(log_model, dataTest)
log_accuracy <- confusionMatrix(log_pred, dataTest$Leading.Cause)$overall['Accuracy']

# Random Forest Model
rf_model <- randomForest(Leading.Cause ~ ., data = dataTrain, importance = TRUE)
rf_pred <- predict(rf_model, dataTest)
rf_accuracy <- confusionMatrix(rf_pred, dataTest$Leading.Cause)$overall['Accuracy']

# Output the model accuracies
print(log_accuracy)
print(rf_accuracy)

# Plot Feature Importance
ggplot(importance_df, aes(x = reorder(Feature, Importance), y = Importance)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Feature Importance in Random Forest Model",
       x = "Features",
       y = "Importance")


# Interaction Plot between Age Adjusted Death Rate and Death Rate
ggplot(data, aes(x = Age.Adjusted.Death.Rate, y = Death.Rate, color = Leading.Cause)) +
  geom_point(alpha = 0.7) +
  theme_minimal() +
  labs(title = "Interaction between Age Adjusted Death Rate and Death Rate",
       x = "Age Adjusted Death Rate",
       y = "Death Rate")

# Interaction Plot by Sex
ggplot(data, aes(x = Age.Adjusted.Death.Rate, y = Death.Rate, color = Leading.Cause)) +
  geom_point(alpha = 0.7) +
  facet_wrap(~Sex) +
  theme_minimal() +
  labs(title = "Interaction between Age Adjusted Death Rate and Death Rate by Sex",
       x = "Age Adjusted Death Rate",
       y = "Death Rate")

# Select relevant features for clustering
clustering_data <- data %>% select(Age.Adjusted.Death.Rate, Death.Rate, Deaths)

# Normalize the data
clustering_data <- scale(clustering_data)

# K-means Clustering
set.seed(123)
```

```
kmeans_result <- kmeans(clustering_data, centers = 3, nstart = 25)

# Visualize K-means Clusters
fviz_cluster(kmeans_result, data = clustering_data, geom = "point",
             ellipse.type = "convex", ggtheme = theme_minimal(),
             main = "K-means Clustering of Leading Causes of Death")

# Hierarchical Clustering
dissimilarity_matrix <- dist(clustering_data)

# Hierarchical clustering using complete linkage
hc_result <- hclust(dissimilarity_matrix, method = "complete")

# Plot the dendrogram
fviz_dend(hc_result, k = 3,
          cex = 0.5,
          color_labels_by_k = TRUE,
          rect = TRUE,
          rect_fill = TRUE,
          main = "Hierarchical Clustering Dendrogram")

# Add cluster membership to the original data
data$KMeansCluster <- as.factor(kmeans_result$cluster)
data$HierarchicalCluster <- as.factor(cutree(hc_result, k = 3))

kmeans_clusters <- table(data$KMeansCluster)
hierarchical_clusters <- table(data$HierarchicalCluster)

kmeans_clusters
hierarchical_clusters
```