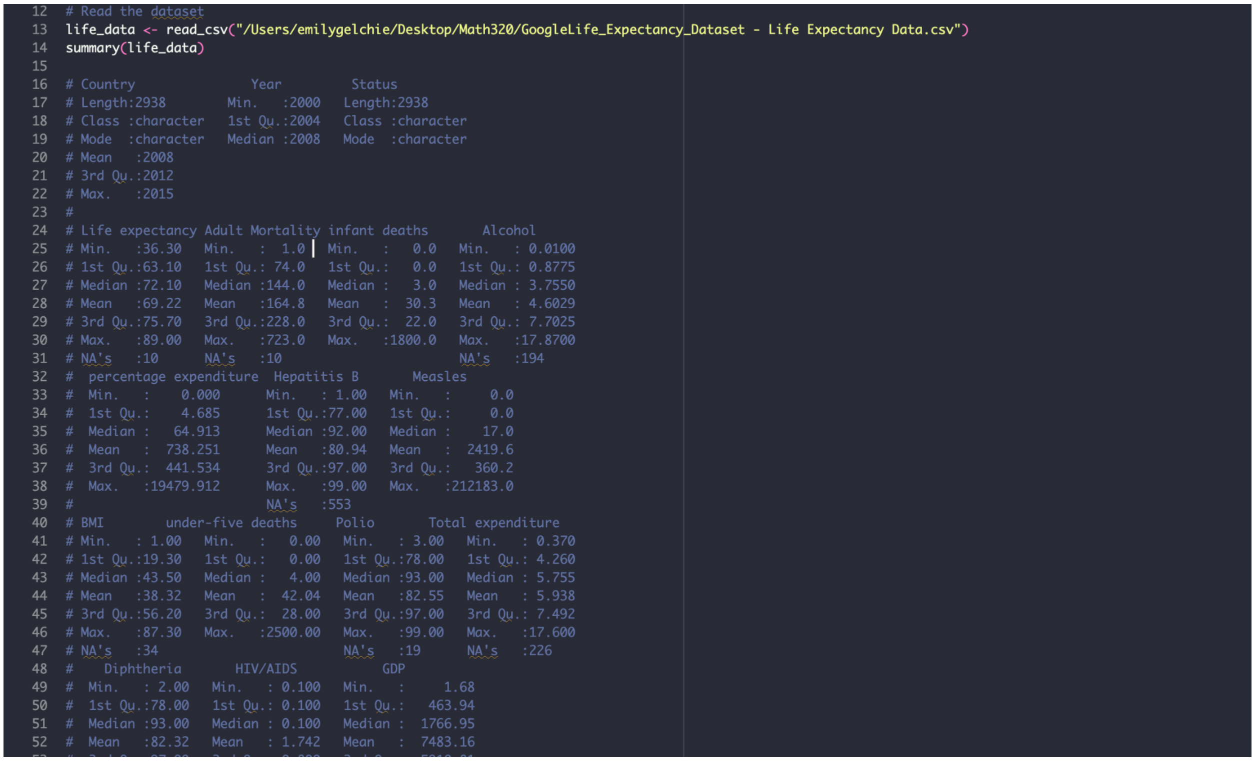
Emily Gelchie

Math 320

Final Essay

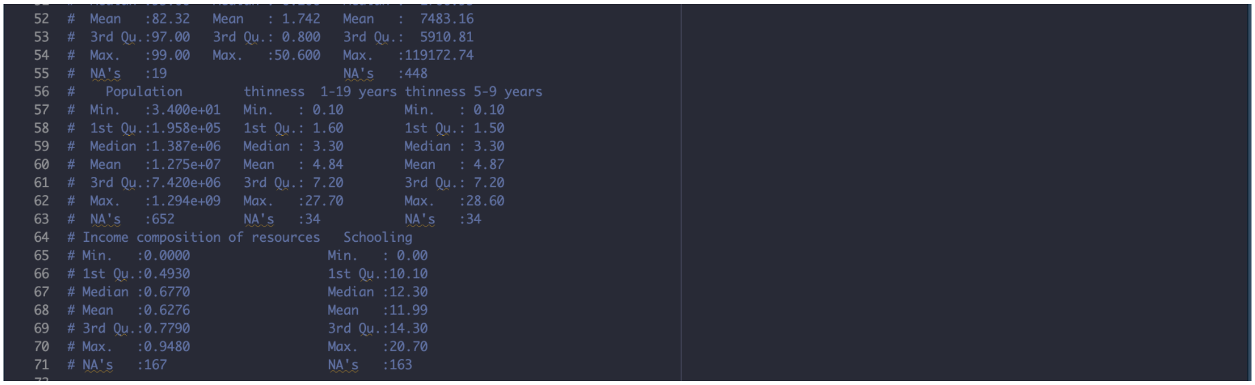
For my final project in Math320, I really wanted to build a predictive model using a life expectancy dataset I found on Kaggle [https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who] that would accept a hypothetical data point from the user and be able to predict the life expectancy based on a machine learning model. However, after this proved to be extremely difficult, I instead built a machine-learning model that used a training data set to predict life expectancy values in the testing data set. In this essay, I meticulously explore a comprehensive dataset on global life expectancy, utilizing a suite of powerful statistical tools and methodologies. My journey begins with a detailed data preparation phase, employing advanced R libraries for streamlined manipulation and analysis. I delve into exploratory data analysis, harnessing the potent capabilities of ggplot2 for insightful visualizations. Feature engineering, an integral aspect of my study, is conducted with precision, leading me into the realms of advanced machine learning modeling. Here, I employ the robust XGBoost algorithm, fine-tuning my model to achieve optimal performance. The essay culminates in critically evaluating my model's efficacy, rigorously employing the RMSE metric to test its predictive accuracy. This comprehensive analysis showcases the versatility of data science in understanding complex health-related phenomena and sets a precedent for future research in the field.

The dataset from the World Health Organization on global life expectancy encompasses a diverse range of variables, including country, year, status (developed or developing), life expectancy, adult mortality, infant deaths, alcohol consumption, percentage expenditure on healthcare, hepatitis B vaccination rates, measles cases, body mass index (BMI), under-five deaths, polio vaccination rates, total healthcare expenditure, diphtheria vaccination rates, HIV/AIDS prevalence, GDP, population, thinness in children aged 1-19 years, thinness in children aged 5-9 years, income composition of resources, and schooling years.



```
# Read the dataset
life_data <- read_csv("/Users/emilygelchie/Desktop/Math320/GoogleLife_Expectancy_Dataset - Life Expectancy Data.csv")
summary(life_data)

# Country               Year          Status
# Length:2938        Min.   :2000   Length:2938
# Class :character   1st Qu.:2004   Class :character
# Mode  :character   Median :2008   Mode  :character
# Mean   :2008
# 3rd Qu.:2012
# Max.   :2015
#
# Life expectancy Adult Mortality infant deaths       Alcohol
# Min.   :36.30   Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100
# 1st Qu.:63.10   1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775
# Median :72.10   Median :144.0   Median :   3.0   Median : 3.7550
# Mean   :69.22   Mean   :164.8   Mean   :  30.3   Mean   : 4.6029
# 3rd Qu.:75.70   3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025
# Max.   :89.00   Max.   :723.0   Max.   :1800.0   Max.   :17.8700
# NA's   :10      NA's   :10                       NA's   :194
#  percentage expenditure  Hepatitis B       Measles
#  Min.   :    0.000      Min.   : 1.00   Min.   :     0.0
#  1st Qu.:    4.685      1st Qu.:77.00   1st Qu.:     0.0
#  Median :   64.913      Median :92.00   Median :    17.0
#  Mean   :  738.251      Mean   :80.94   Mean   :  2419.6
#  3rd Qu.:  441.534      3rd Qu.:97.00   3rd Qu.:   360.2
#  Max.   :19479.912      Max.   :99.00   Max.   :212183.0
#                         NA's   :553
# BMI        under-five deaths     Polio       Total expenditure
# Min.   : 1.00   Min.   :   0.00   Min.   : 3.00   Min.   : 0.370
# 1st Qu.:19.30   1st Qu.:   0.00   1st Qu.:78.00   1st Qu.: 4.260
# Median :43.50   Median :   4.00   Median :93.00   Median : 5.755
# Mean   :38.32   Mean   :  42.04   Mean   :82.55   Mean   : 5.938
# 3rd Qu.:56.20   3rd Qu.:  28.00   3rd Qu.:97.00   3rd Qu.: 7.492
# Max.   :87.30   Max.   :2500.00   Max.   :99.00   Max.   :17.600
# NA's   :34                        NA's   :19      NA's   :226
#    Diphtheria       HIV/AIDS          GDP
#  Min.   : 2.00   Min.   : 0.100   Min.   :     1.68
#  1st Qu.:78.00   1st Qu.: 0.100   1st Qu.:   463.94
#  Median :93.00   Median : 0.100   Median :  1766.95
#  Mean   :82.32   Mean   : 1.742   Mean   :  7483.16
```



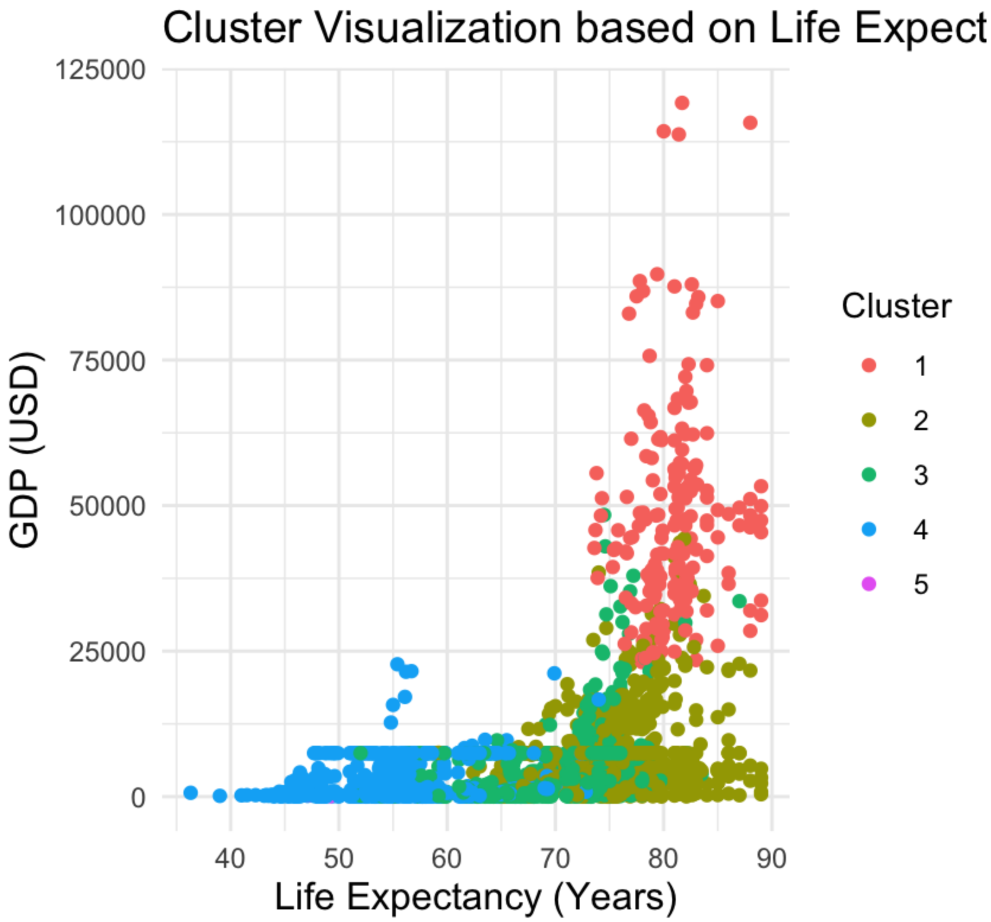
```
#  Mean   :82.32   Mean   : 1.742   Mean   :  7483.16
#  3rd Qu.:97.00   3rd Qu.: 0.800   3rd Qu.:  5910.81
#  Max.   :99.00   Max.   :50.600   Max.   :119172.74
#  NA's   :19                       NA's   :448
#    Population       thinness  1-19 years thinness 5-9 years
#  Min.   :3.400e+01   Min.   : 0.10        Min.   : 0.10
#  1st Qu.:1.958e+05   1st Qu.: 1.60        1st Qu.: 1.50
#  Median :1.387e+06   Median : 3.30        Median : 3.30
#  Mean   :1.275e+07   Mean   : 4.84        Mean   : 4.87
#  3rd Qu.:7.420e+06   3rd Qu.: 7.20        3rd Qu.: 7.20
#  Max.   :1.294e+09   Max.   :27.70        Max.   :28.60
#  NA's   :652         NA's   :34           NA's   :34
# Income composition of resources   Schooling
# Min.   :0.0000                  Min.   : 0.00
# 1st Qu.:0.4930                  1st Qu.:10.10
# Median :0.6770                  Median :12.30
# Mean   :0.6276                  Mean   :11.99
# 3rd Qu.:0.7790                  3rd Qu.:14.30
# Max.   :0.9480                  Max.   :20.70
# NA's   :167                     NA's   :163
```

The initial summary of this dataset reveals significant insights. It spans from 2000 to 2015, covering 2938 data points across various countries. Key health indicators like life expectancy range from 36.3 to 89 years, with a mean of 69.22, suggesting global variability. The dataset also shows wide ranges in variables like adult mortality, infant deaths, and alcohol consumption, highlighting disparities in health outcomes. Economic indicators like GDP also vary widely, reflecting global economic diversity. The presence of missing data in variables like Hepatitis B and population underscores challenges in global health data collection. This dataset is crucial for understanding the multifaceted nature of health across different regions and for formulating targeted public health interventions.

Data cleaning was a crucial step to ensure quality and consistency in the analysis. The 'Country' variable was explicitly converted to a string format for uniformity. Lesser-represented countries were grouped under the category ‘Other’ to manage infrequent categories within this variable. This approach streamlined the dataset, reducing potential biases from less frequent country data. Additionally, specific countries causing data inconsistencies, such as Dominica and Nauru, were excluded from the analysis.

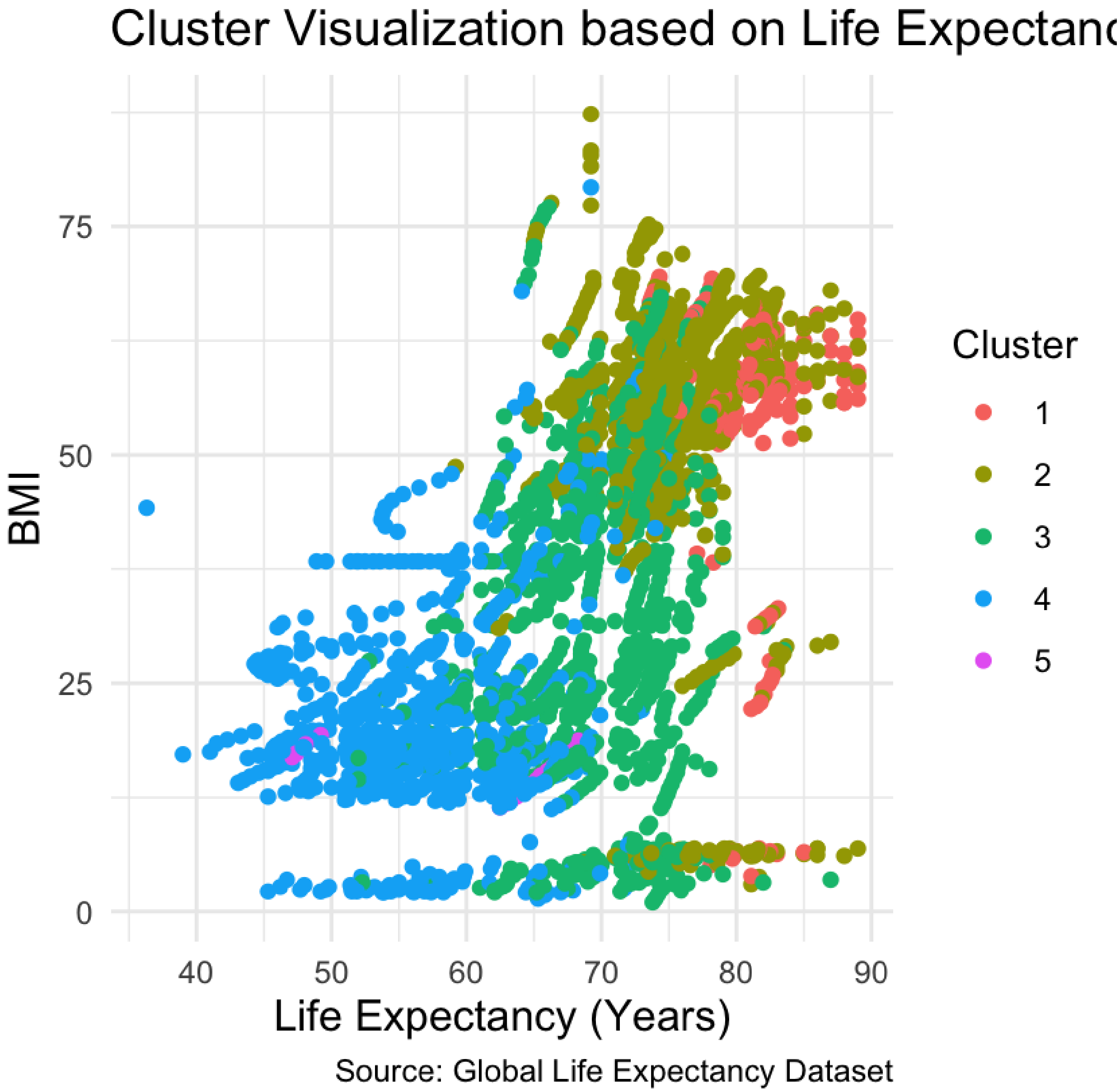
The dataset then underwent normalization, a vital process in data preprocessing. Numeric columns were scaled to have a common scale without distorting differences in the ranges of values. This standardization, achieved using the 'scale' function in R, is essential in data modeling as it ensures that each variable contributes equally to the analysis, allowing for more accurate and interpretable models. This normalization is particularly important when employing techniques like machine learning, where scale differences can significantly impact the algorithm's performance.

Exploratory Data Analysis (EDA) involves a series of histograms and K-means clustering that visually represent the standardized frequency distribution for key health and demographic variables. I chose to cluster because clustering is a fundamental technique in data analysis and machine learning employed to uncover inherent patterns and groupings within datasets. Its significance lies in its capacity to unveil hidden structures and relationships among data points, facilitating a deeper understanding of complex datasets. The cluster graphs generated through this technique visually represent these relationships by mapping data points into distinct clusters based on their similarities in selected feature dimensions. The cluster graph depicting 'Life expectancy' against 'GDP' elucidates how countries tend to group based on their economic prosperity and its impact on life expectancy. Each cluster signifies a distinct grouping, possibly demarcating the gradations in affluence and longevity. For instance, one might hypothesize that the cluster characterized by higher GDP and Life Expectancy encapsulates developed nations, whereas the cluster at the lower end of the spectrum is indicative of developing nations grappling with economic challenges and concomitant health-related issues.



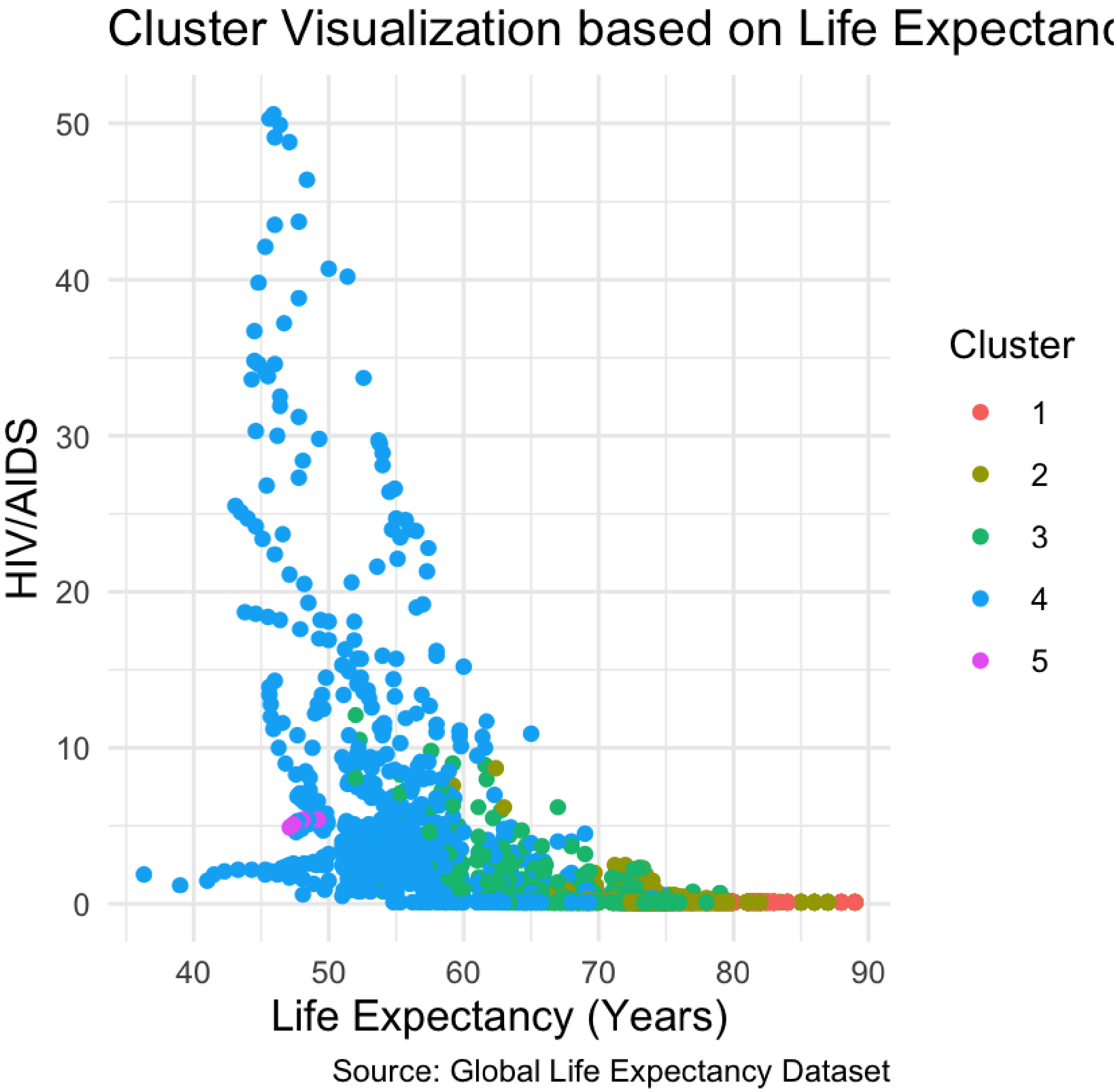
Source: Global Life Expectancy Dataset

Similarly, the 'Life expectancy' versus 'BMI' graph reveals how body mass index correlates with life expectancy, shedding light on the potential health implications of body composition. The clusters likely illuminate patterns of nutritional and lifestyle factors across different geographies, with some clusters possibly representing regions with higher rates of obesity and its associated health risks, while others could denote regions with more moderate BMI values and potentially healthier populations.

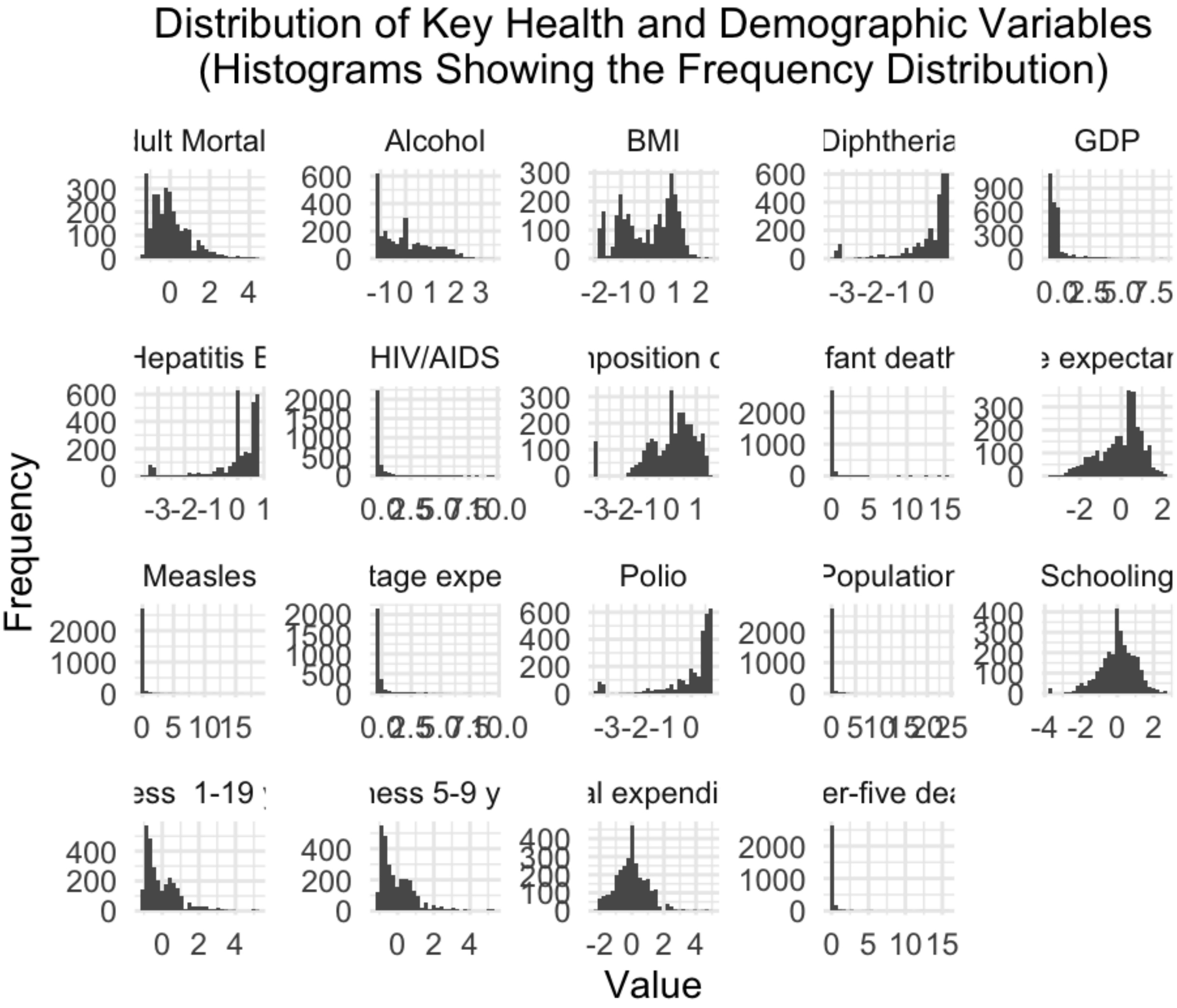


Source: Global Life Expectancy Dataset

Finally, the 'Life expectancy' versus 'HIV/AIDS' graph illuminates how the prevalence of HIV/AIDS affects life expectancy, serving as a crucial indicator of public health outcomes. The clustering here could be indicative of varied regional health crises, where certain clusters represent areas severely afflicted by the epidemic, manifesting in diminished life spans. Conversely, other clusters may correspond to regions where the incidence of HIV/AIDS is relatively contained, reflecting in greater life expectancies.



Continuing the EDA, these histograms play a critical role in helping us assess distribution symmetry, identify outliers, and determine the modality of the data, whether unimodal or bimodal.

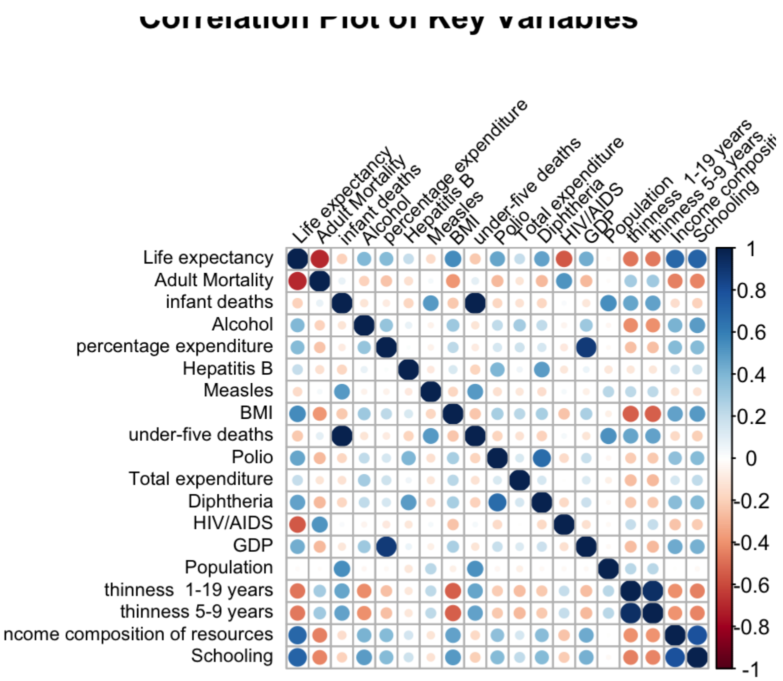


Source: Global Life Expectancy Dataset

The visualizations unveil a wide range of shapes; some variables exhibit a right-skewed distribution, which suggests a higher occurrence of lower values, while others show a more typical bell-shaped distribution centered around the mean. This analysis of patterns is fundamental in EDA because it informs the appropriateness of subsequent statistical methods. Additionally, these plots can signal the need for data transformation or outlier treatment before I proceed with further analysis. These histograms represent the frequency distributions of standardized values for each variable in the dataset, offering valuable insights into the characteristics of these distributions. For instance, the right-skewed histogram for ‘Adult

Mortality’ implies a concentration of countries with lower mortality rates and fewer with higher rates. On the other hand, variables like 'Schooling' exhibit a more bell-shaped curve, indicating a more even distribution of data points around the mean.

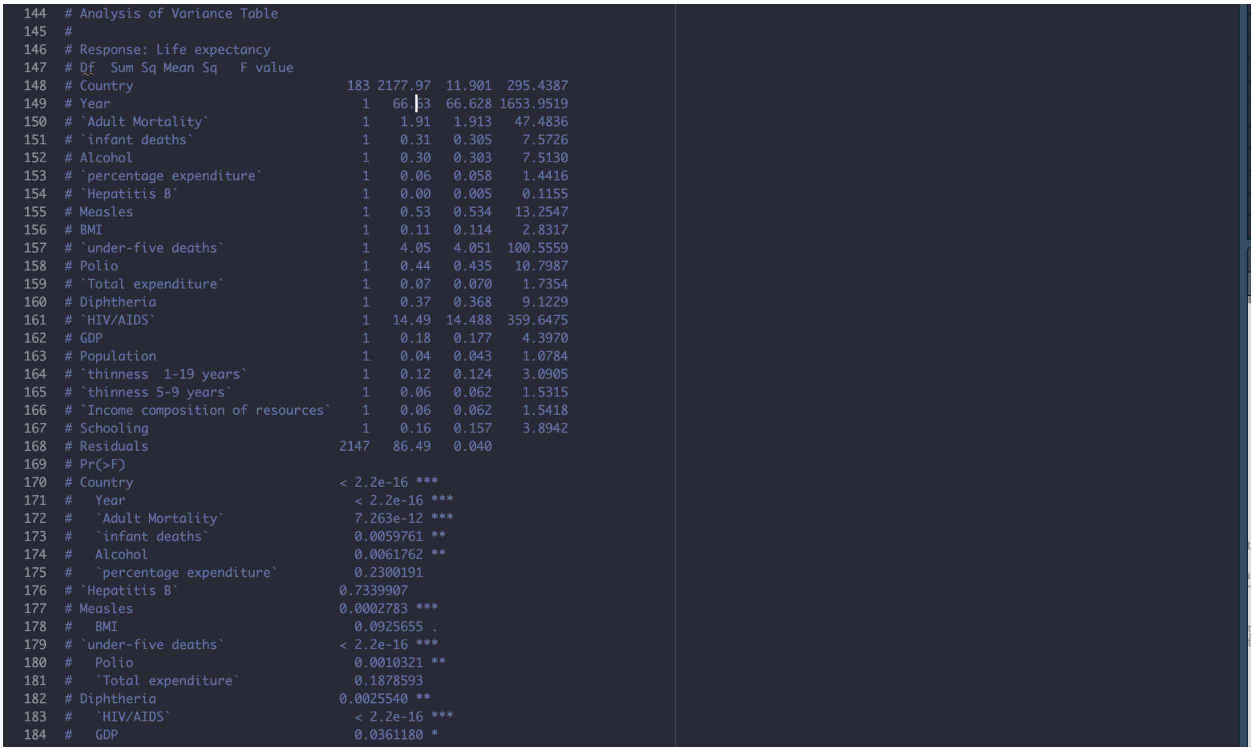
The feature engineering section of the code focuses on preparing and transforming the dataset variables to enhance the machine learning models' performance. Specifically, the correlation_matrix is calculated using the cor function on the normalized life data, excluding non-numeric variables such as 'Country,’ 'Year,’ and 'Status'. This matrix is critical for identifying relationships between variables, where strong correlations can suggest potential predictors for life expectancy.



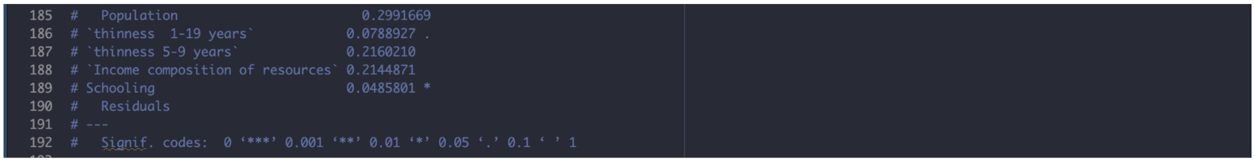
The corrplot function visualizes these correlations, employing a circular method and clear labels, allowing for a quick assessment of which features may have more predictive power. The plot displays the degree to which variables are linearly related, with 1 indicating a perfect positive correlation, -1 indicating a perfect negative correlation, and 0 indicating no correlation.

The graph shows circles whose size and color intensity represent the strength and direction of the correlation between variables. Large, dark-colored circles indicate strong correlations, which are crucial for predicting the target variable. These relationships help select the most significant predictors for the machine learning models and understand the underlying structure and trends within the data. As we can see, according to the graph, under-five deaths, GDP, and HIV/AIDS exposure all have significant correlations to life expectancy.

ANOVA (Analysis of Variance) analysis is performed on a linear model to understand the influence of various predictors on life expectancy. The ANOVA table summarizes the variance explained by each predictor and tests the null hypothesis that each coefficient's effect is zero. The Degrees of Freedom (Df) column in the ANOVA table represents the number of levels minus one for each factor. It is a measure of the amount of 'independent' information the variable contributes to the model. It helps quantify the degree of freedom associated with each predictor. The Sum Sq (Sum of Squares) quantifies the variance each predictor contributes. It's calculated by summing the squared differences between the predicted and actual values. This statistic helps us understand how much variation in the outcome can be attributed to each predictor. The Mean Sq (Mean Square) is the average of the Sum Sq, obtained by dividing the Sum Sq by its degrees of freedom. It estimates the variance within each predictor group, providing insight into the relative importance of each predictor. The F value is a test statistic calculated by dividing the Mean Sq of the variable by the Mean Sq of the error. It determines if the variable significantly predicts the outcome. A higher F value indicates a stronger predictive effect. If the null hypothesis is true, pr (>F) indicates the probability of observing such an F-value. A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, suggesting the variable significantly affects the outcome. The significant codes in the ANOVA table concisely gauge the statistical significance of each predictor's impact on the response variable, which in this case is "Life expectancy." Various symbols represent these codes and provide a quick assessment of the significance of each predictor's effect. When interpreting these codes, " signifies the highest significance level, indicating a strong and highly significant effect. It implies compelling evidence to reject the null hypothesis, suggesting that the predictor has a substantial and clear impact on the response variable. " is also a strong significance level, indicating a highly significant effect with strong evidence against the null hypothesis. A single asterisk " represents a moderate significance level, suggesting evidence to reject the null hypothesis, but the impact may not be as pronounced as with '*' or ". A period or dot ('.') indicates a lower level of significance, signifying some evidence against the null hypothesis but not at conventional significance levels (e.g., p > 0.05), implying a weaker influence. Finally, a blank space indicates no statistical significance, implying insufficient evidence to reject the null hypothesis and suggesting that the predictor does not significantly affect the response variable.



```
# Analysis of Variance Table
#
# Response: Life expectancy
# Df  Sum Sq Mean Sq   F value
# Country                            183 2177.97  11.901  295.4387
# Year                                 1   66.63  66.628 1653.9519
# `Adult Mortality`                    1    1.91   1.913   47.4836
# `infant deaths`                      1    0.31   0.305    7.5726
# Alcohol                              1    0.30   0.303    7.5130
# `percentage expenditure`             1    0.06   0.058    1.4416
# `Hepatitis B`                        1    0.00   0.005    0.1155
# Measles                              1    0.53   0.534   13.2547
# BMI                                  1    0.11   0.114    2.8317
# `under-five deaths`                  1    4.05   4.051  100.5559
# Polio                                1    0.44   0.435   10.7987
# `Total expenditure`                  1    0.07   0.070    1.7354
# Diphtheria                           1    0.37   0.368    9.1229
# `HIV/AIDS`                           1   14.49  14.488  359.6475
# GDP                                  1    0.18   0.177    4.3970
# Population                           1    0.04   0.043    1.0784
# `thinness  1-19 years`               1    0.12   0.124    3.0905
# `thinness 5-9 years`                 1    0.06   0.062    1.5315
# `Income composition of resources`    1    0.06   0.062    1.5418
# Schooling                            1    0.16   0.157    3.8942
# Residuals                         2147   86.49   0.040
# Pr(>F)
# Country                          < 2.2e-16 ***
#   Year                             < 2.2e-16 ***
#   `Adult Mortality`                7.263e-12 ***
#   `infant deaths`                  0.0059761 **
#   Alcohol                          0.0061762 **
#   `percentage expenditure`         0.2300191
# `Hepatitis B`                    0.7339907
# Measles                          0.0002783 ***
#   BMI                              0.0925655 .
# `under-five deaths`              < 2.2e-16 ***
#   Polio                            0.0010321 **
#   `Total expenditure`              0.1878593
# Diphtheria                       0.0025540 **
#   `HIV/AIDS`                       < 2.2e-16 ***
#   GDP                              0.0361180 *
```



```
#   Population                       0.2991669
# `thinness  1-19 years`           0.0788927 .
# `thinness 5-9 years`             0.2160210
# `Income composition of resources` 0.2144871
# Schooling                        0.0485801 *
#   Residuals
# ---
#   Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

'Country' and 'Year' have extremely low p-values, indicating they are highly significant predictors of life expectancy. This suggests that the specific country and year substantially influence life expectancy. Variables such as 'Adult Mortality,' 'infant deaths,' 'Alcohol,' 'Measles,' 'under-five deaths,' 'Polio,' 'Diphtheria,' 'HIV/AIDS,' 'GDP,' and 'Schooling' also show significance, implying that these factors are important in predicting life expectancy. They have a statistically significant impact on life expectancy. Some predictors like 'percentage expenditure,' 'Hepatitis B,' 'Total expenditure,' 'Population,' 'thinness 1-19 years,' 'thinness 5-9 years,' and 'Income composition of resources' have p-values greater than 0.05, suggesting weaker evidence for their influence on life expectancy. This indicates that these variables may have a less significant role in predicting life expectancy or that their effects are less clear in the given dataset. In essence, these significant codes and p-values provide valuable insights into which predictors have robust and statistically significant influences and which ones have weaker or non-significant effects within the model, confirming the importance of certain factors in explaining life expectancy variations while highlighting the uncertainty or limited impact of others.

The Model Tuning section employs the XGBoost algorithm, a decision-tree-based ensemble Machine Learning method that uses a gradient boosting framework. Specifically, it defines a regression model to predict life expectancy with the boost_tree function. XGBoost is known for its performance and speed and is particularly useful when dealing with large and complex datasets. The process fits the XGBoost model to the training data, with life expectancy as the response variable and all other variables as predictors, denoted by the formula’s tilde (~)

and dot (.). This approach automatically selects significant predictors based on the model's ability to improve predictions iteratively. The model's evaluation and tuning are implicit in the XGBoost algorithm through its gradient boosting mechanism, which builds trees one at a time, where each new tree helps to correct errors made by previously trained trees. The model's hyperparameters can be tuned to optimize performance, though this code segment doesn't explicitly include hyperparameter tuning, suggesting default settings.

XGBoost was chosen for this project over a simple decision tree due to its robustness in handling complex datasets with multiple variables. XGBoost employs advanced regularization techniques that prevent overfitting, a common issue with decision trees, especially in data-rich environments. Its gradient-boosting approach effectively minimizes errors in sequential models, enhancing accuracy. XGBoost also efficiently manages missing data and variable interactions, making it superior for a dataset with diverse attributes and intricate relationships influencing life expectancy. The algorithm's scalability and ability to parallelize processing make it well-suited for the dataset's complexity. In essence, XGBoost's inclusion in the project indicates a sophisticated approach to predictive modeling, aiming to achieve high accuracy in estimating life expectancy by leveraging the algorithm's ability to handle various data types and distributions effectively while addressing challenges like overfitting and missing data.

In the model evaluation section, the performance of the XGBoost model is assessed using Root Mean Square Error (RMSE). RMSE is a standard measure to evaluate the accuracy of a regression model; it represents the square root of the second sample moment of the differences between predicted values and observed values or the quadratic mean of these differences. These differences are also known as residuals, and RMSE aggregates them into a single measure of predictive power. In this project, an RMSE of 0.199 indicates a strong predictive model for life expectancy. This value represents the average distance between the predicted values by the XGBoost model and the actual data points. Considering the dataset's life expectancy ranged from 36.30 to 89.00 years, an RMSE below 0.2 is a commendable outcome, reflecting the model's ability to forecast accurately within the data's variability.

In conclusion, this project has been quite a journey for me. Initially, I faced some initial challenges, mainly because I aimed to create a model that could predict life expectancy for hypothetical individuals, making the project more complex. I aimed to offer a user-friendly experience where users could input data for a fictional person and witness the model's predictive abilities. Despite the initial complexities and uncertainties, I'm genuinely pleased with how this project turned out. Seeing that the model I've built has proven accurate is immensely satisfying. This success has opened up exciting possibilities for future projects, and I'm looking forward to building on this foundation, perhaps in the upcoming Data Science Capstone project next semester. This course, Math 320, has been an incredible learning experience for me. I've gained R programming, statistics, and data modeling knowledge and skills. This journey has expanded my technical expertise and deepened my appreciation for the intricate world of data analysis. As I reflect on this course, I'm both excited about the future and immensely grateful for the opportunities and insights Math 320 has provided me.