# Final_Report_Logistic

## Introduction

There is widespread research looking into how happy a country is based on certain socioeconomic and individual-level factors (***find research***). Furthermore, there is also wide-ranging research looking into various mental health disorders and their impact on quality of life (***find research***). There is a general consensus that the better a countries' socioeconomic status, the higher they score on the happiness index (***find research***). On the other hand, mental health disorders are suggested to be caused by a plethora of factors, such as, genetics, environmental circumstances, financial circumstances, etc. (***find research***). The aim of the current research paper is to gain a better understanding of these two schools of research together.

The prevalence of depression is on the rise across the globe (Bell & Blanchflower, 2019). Furthermore, research has indicated that this may have to do with one's socioeconomic status, where individuals with lower socioeconomic statuses are more likely to experience depressive symptoms (Graham & Pinto, 2019). Contrarily, research has also shown that happiness interventions have been shown to reduce depression amongst individuals (D'raven et a., 2014). Therefore, there is an established correlation between depression and happiness. The current research paper would like to extend these findings and try to establish a predictive inference of depression based on happiness metrics.

On the other hand, there could be a reverse relationship between these factors as well. To better explain this, it is important to understand the concept of Quality of Life (QoL), which refers to the overall well-being of an individual or population (Teoli & Bhardwaj, 2023). This includes having good personal health (mental, physical, and spiritual), strong relationships, good education and social status, healthy work environment, good income, sense of safety/security, freedom to make choices, and healthy physical environment. Based on this definition, the current study could look into how/to what extent mental health disorders and the other QoL metrics mentioned could predict QoL.

For the same, the data used for this research was obtained from the World Happiness Report for information about happiness metrics [link1] and from University of Oxford for information about mental health disorders [link 2].

The aim of this research paper was to investigate the following avenues with respect to world happiness and mental health:

1. To what extent do happiness metrics (log GDP per capita, social support, life expectancy, freedom, perception of corruption, generosity) predict mental health disorders, specifically, depression?
2. To what extent do mental health disorders and happiness metrics (schizophrenia, bipolar disorder, eating disorders, anxiety disorders, depression, alcohol use disorder, drug abuse disorders, log GDP per capita, social support, freedom to make choices) predict quality of life?

## Method

*Materials and Data Merging*

As mentioned earlier, two datasets were used for this study: World Happiness Report (2022) and Mental Health Disorders Data (2022). These datasets were merged on common countries and years, and then a larger dataset with 1462 observations was created. It contained 18 variables, namely, country, year, life ladder, log GDP per capita, social support, life expectancy at birth, freedom, perception of corruption, generosity, positive affect, negative affect, Alcohol Use Disorders, Drug Use Disorders, Schizophrenia, Bipolar Disorders, Anxiety Disorders, Eating Disorders, and Depression.

**CHECK BELOW PARAGRAPH. PLEASE ADD OR SUBTRACT INFO IF NEED BE.**

The happiness metrics were a range of factual data collected from census surveys (e.g., GDP per capita, life expectancy at birth) and self-reported variables collected from the population of each country in a given year (e.g., life ladder, freedom to make choices, social support, etc). The final variables of this dataset were also self-reported measures of positive and negative emotions which were culminated and calculated into the variables positive affect and negative affect (required to analyze happiness reports for countries). For this dataset, countries were considered to have an acceptable sample size and were included in the dataset if their sample ranged from 1000-3000 individuals. The mental disorders dataset's variables indicated the percentage of the population that had the stated mental disorder in the given country and year per observation.

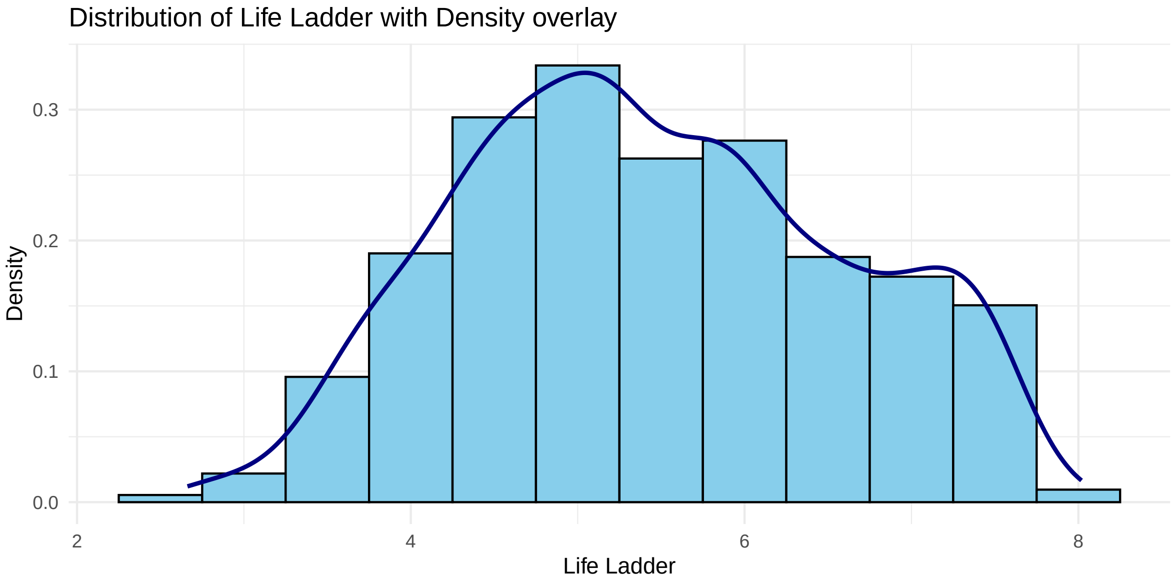
*Data Cleaning and Manipulation*

The merged dataset had multiple observations per country, each observation representing a different year. Though combining information by country was an option, this would involve averaging observations across years per country and removing information about time. However, in relation to happiness metrics and mental health, it is important to maintain information regarding time as research suggests that environmental changes across the years can have an impact on these factors (***FIND RESEARCH***). Therefore, observations were not grouped by country by averaging all information across different years.

However, for ease of interpretation of data visualization, countries were split into four different categories: low (<= $1,135 per capita), lower-middle ($1,136-4,465 per capita), upper-middle ($4,466-13,845 per capita) and high income (> $13,845 per capita) countries. This was done as per the latest income group classification put forth by the World Bank (World Bank, 2023). Converting the variable Log GDP per capita to a regular format, countries were then categorized into said income categories.

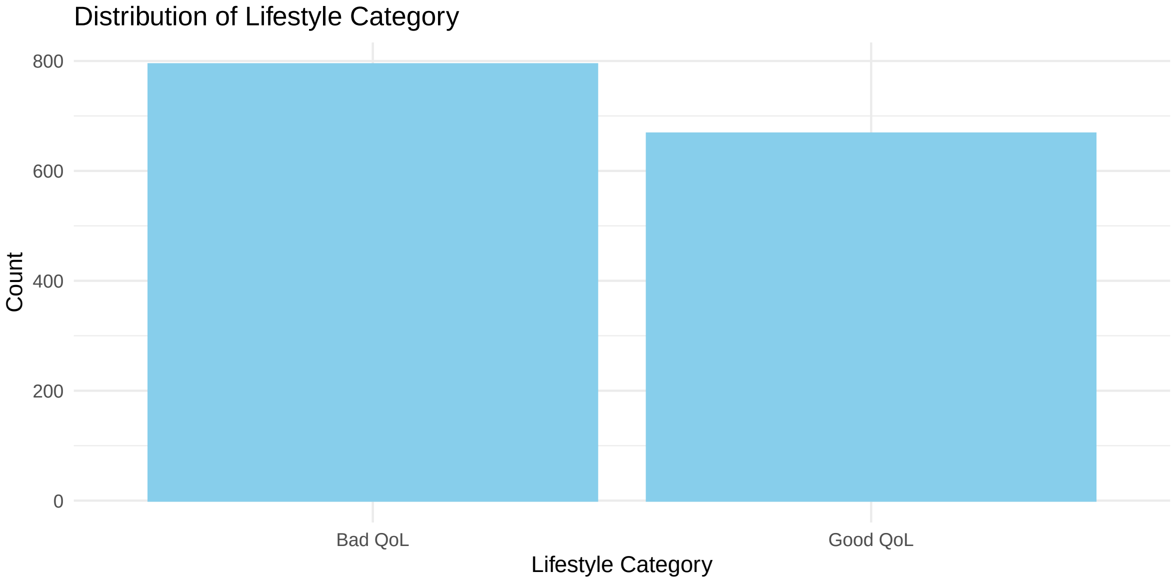
The last transformation done was in relation to the second research question. Since this question aimed to predict one's lifestyle category based on mental health and happiness metrics, a variable for "lifestyle category" needed to be created. This was done with the help of the variable Life Ladder in the World Happiness dataset. The original variable is called the Cantril ladder, which asks respondents to rate their current lives on a scale of 0 (worst possible life) to 10 (best possible life). In the World Happiness dataset, this variable was an average of all the results collected from citizens of the specific country in the given year (i.e., a float, continuous variable). ***Figure __*** visually depicts the distribution of the original variable life ladder in the dataset.

Figure __. Histrogram of the distribution of life ladder across countries.



To simplify this variable and create an overarching binary variable indicating "Lifestyle Category", i.e., Good Quality of Life (Good QoL) vs. Bad Quality of Life (Bad QoL), the variable life ladder was transformed. This was done using the midpoint of the Life Ladder scale (i.e., 5.5) and then recoding a Lifestyle Category variable to Bad QoL if the observation was below the midpoint and Good QoL if it was above. ***Figure* __**shows the distribution of this newly created binary variable, Lifestyle Category.

Figure __. Barplot for distribution of Lifestyle Category.



*Exploratory Data Analysis*

### FIRST RESEARCH QUESTION INFO GOES HERE

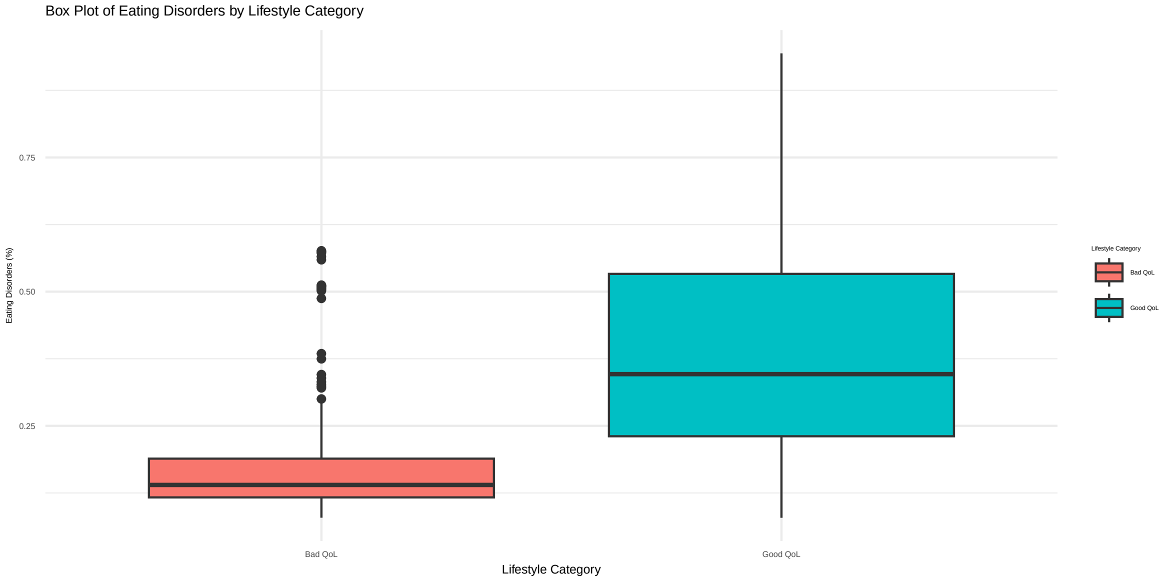
For the second research question, it was important to gain some insights into the variable descriptive statistics based on the binary variable, i.e., statistics for the Bad QoL category vs. the Good QoL category. *Table* __ provides this information.

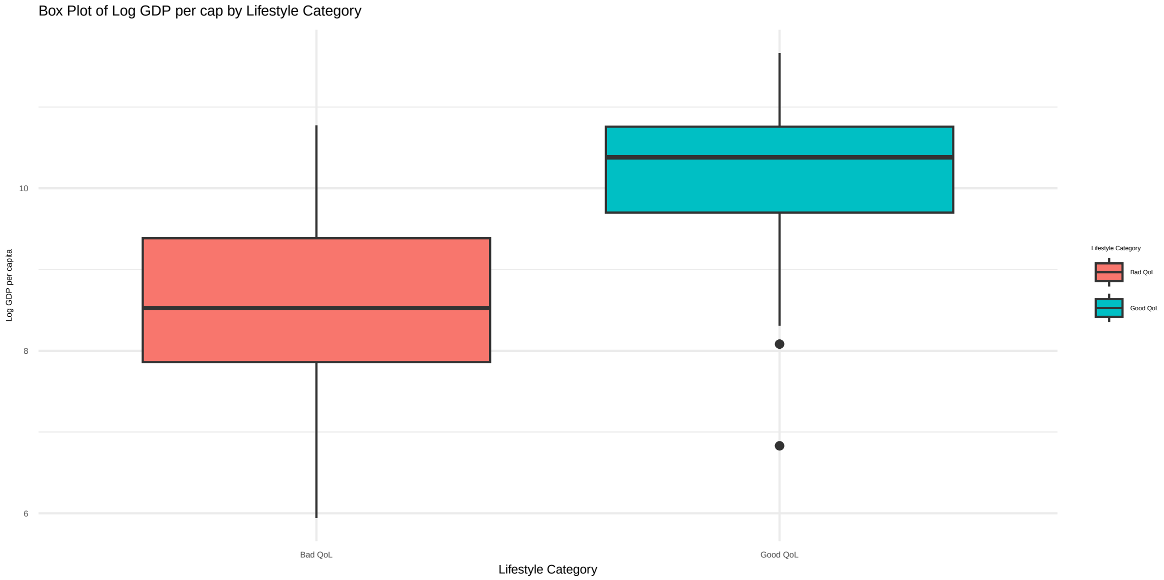
Table __. Descriptive Statistics for variables related to research question 2, factored based on the two levels of Lifestyle Category (Bad QoL vs. Good QoL).

```
cat(latex_table)
```

| Lifestyle Category | Schizophrenia (%)_mean | Schizophrenia (%)_sd | Schizophrenia (%)_median | Schizo |
|---|---|---|---|---|
| Bad QoL | 0.1920485 | 0.0318930 | 0.1875378 | |
| Good QoL | 0.2360202 | 0.0465964 | 0.2123363 | |

The boxplots below help visualize two of these statistics, one mental health disorder and one happiness metric (Figure __ and Figure __).





## Assessing Multicollinearity

We fit a basic regression model, and then we calculate the Variance Inflation Factor (VIF) values for the predictors in the model to assess multicollinearity. Multicollinearity refers to a situation in which two or more predictor variables in a multiple regression model are highly correlated. If these variables are highly correlated, it can be difficult to disentangle the separate effects of the predictors on the response variable.

```
# Fit a basic model
model <- glm(`Lifestyle Category` ~ `Depression (%)` + `Schizophrenia (%)` + `Bipolar diso

# Calculate VIF
vif_values <- vif(model)

# Print VIF values
vif_values
```

```
        `Depression (%)`            `Schizophrenia (%)`
                1.422584                       1.786717
  `Bipolar disorder (%)`         `Eating disorders (%)`
                2.836509                       3.272848
 `Anxiety disorders (%)`           `Log GDP per capita`
                2.843300                       2.661674
        `Social support` `Freedom to make life choices`
                1.382023                       1.458170
       `Positive affect`              `Negative affect`
                1.761800                       1.418236
`Drug use disorders (%)`
                1.671129
```

The VIF values for all predictors in the model were found to be low, indicating that there is no multicollinearity. This suggests that each predictor in the model has a unique contribution to the prediction of the response variable and that the effects of the predictors can be estimated separately.

## Results

### FIRST RESEARCH QUESTION REGRESSION GOES HERE.

To recap the components of the second research question, the response variable chosen was Lifestyle Category with two levels: Good QoL and Bad QoL. The predictor variables chosen for the model to predict Lifestyle Category can be categorized into mental health disorders (Schizophrenia, Depression, Anixety, Alcohol abuse, Drug use, Bipolar Disorder, and Eatiny Disorder) and happiness metrics (log GDP per capita, social support, and freedom to make choices). Furthermore, research has suggested that a lack of social support is most likely to increase the risk of substance abuse (Cherry, 2023; Eddie et al., 2019; Horigian et al., 2020). Therefore, it is likely that there is an interaction between the predictors social support and alcohol abuse and/or drug use. For the same, two interaction terms were included in this logistic regression model (i.e., Alcohol use * Social Support and Drug Use * Social Support) to be in line with current psychological and addictive behaviors literature.

### Logistic Regression Model Output Summary

| Variable | Estimate | Std. Error | z value | Pr(> |
|---|---|---|---|---|
| Intercept | -52.35236 | 67.33324 | -0.778 | 0.4369 |
| Depression (%) | -1.09615 | 0.26180 | -4.187 | 2.83e-05 *** |
| Schizophrenia (%) | 10.07029 | 5.55143 | 1.814 | 0.0697 . |
| Bipolar disorder (%) | 2.54807 | 1.39493 | 1.827 | 0.0678 . |
| Eating disorders (%) | 10.03718 | 2.10474 | 4.769 | 1.85e-06 *** |
| Anxiety disorders (%) | -0.15576 | 0.16797 | -0.927 | 0.3538 |
| Log GDP per capita | 0.46577 | 0.24318 | 1.915 | 0.0554 . |
| Social support | 7.42898 | 6.12312 | 1.213 | 0.2250 |
| Freedom to make life choices | 7.00062 | 1.08158 | 6.473 | 9.63e-11 *** |
| Drug use disorders (%) | -3.03971 | 4.47484 | -0.679 | 0.4970 |
| Year | 0.01618 | 0.03335 | 0.485 | 0.6276 |
| Alcohol use disorders (%) | 2.24420 | 1.75640 | 1.278 | 0.2013 |
| Social support: Alcohol use disorders (%) | -2.02845 | 2.00845 | -1.010 | 0.3125 |
| Social support: Drug use disorders (%) | 4.78372 | 5.33878 | 0.896 | 0.3702 |

The logistic regression analysis uncovered meaningful connections between predictor variables and the Lifestyle Category. Exponentiating the coefficients provides a lens to interpret the effects on odds. Notably, individuals with higher levels of Depression (%) experience a substantial 33.4% decrease in the odds ratio of belonging to specific lifestyle categories, underscoring the influential role of mental health in shaping lifestyle choices. Moreover, a 0.0001 unit increase in Freedom to make life choices remarkably boost the odds ratio by 10.9%, highlighting the pivotal role of autonomy in determining one's lifestyle. Eating disorders (%) also exhibit a noteworthy positive association with lifestyle choices, where every 0.0001 unit increase in eating disorders percent in a population corresponds to a 22.86% increase in the odds ratio of having a Good Quality of Life, indicating a significant impact of eating habits on lifestyle preferences. Although variables like Schizophrenia (%), Bipolar disorder (%), and Log GDP per capita show potential trends, they did not attain conventional significance levels. These findings contribute valuable insights into the intricate interplay of mental health, freedom, and eating habits in shaping lifestyle choices.

**The method used for this question was also a prediction based model, as the aim of this part of the study was to predict QoL based on the predictor variables chosen. The results are as follows:**

```
confusionMatrix(data = as.factor(predictions), reference = as.factor(testing_set$`Lifestyl
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 206  20
         1  22 175

               Accuracy : 0.9007
                 95% CI : (0.8682, 0.9275)
    No Information Rate : 0.539
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8004

 Mcnemar's Test P-Value : 0.8774

            Sensitivity : 0.8974
            Specificity : 0.9035
         Pos Pred Value : 0.8883
         Neg Pred Value : 0.9115
             Prevalence : 0.4610
         Detection Rate : 0.4137
   Detection Prevalence : 0.4657
      Balanced Accuracy : 0.9005

       'Positive' Class : 1
```

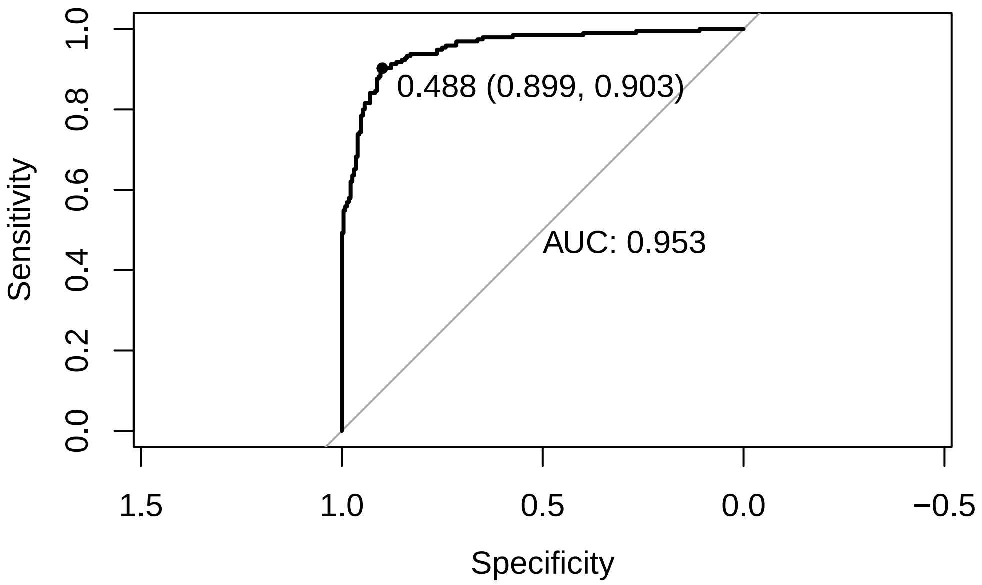
The confusion matrix and associated statistics provide an evaluation of the logistic regression model's performance. The model was assessed using a default threshold of 0.5, where the positive class corresponds to "Good Quality of Life" (QoL). The accuracy of the model is 90.07%, indicating the proportion of correctly classified instances among all predictions. The Kappa statistic, which considers agreement by chance, is 0.8004, suggesting substantial agreement beyond chance. Sensitivity (True Positive Rate) is 89.74%, indicating the model's ability to correctly identify individuals with a Good QoL, while specificity (True Negative Rate) is 90.35%, reflecting the model's proficiency in identifying those without a Good QoL. The positive predictive value (PPV) is 88.83%, representing the probability of actually having a Good QoL given a positive prediction, and the negative predictive value (NPV) is 91.15%, representing the probability of not having a Good QoL given a negative prediction. Overall, the balanced accuracy is high at 90.05%, indicating a well-performing model for distinguishing between the two classes.

```
# Predict probabilities using the logistic regression model on the testing set
probabilities <- predict(logistic_model, newdata = testing_set, type = "prob")

# Extract probabilities of the positive class
positive_probabilities <- probabilities[, "Good QoL"]

# Plot the ROC curve using the probabilities and the actual values from the testing set
roc_obj <- roc(testing_set$`Lifestyle OneHotEnc`, positive_probabilities)
```

```
Setting levels: control = 0, case = 1
```

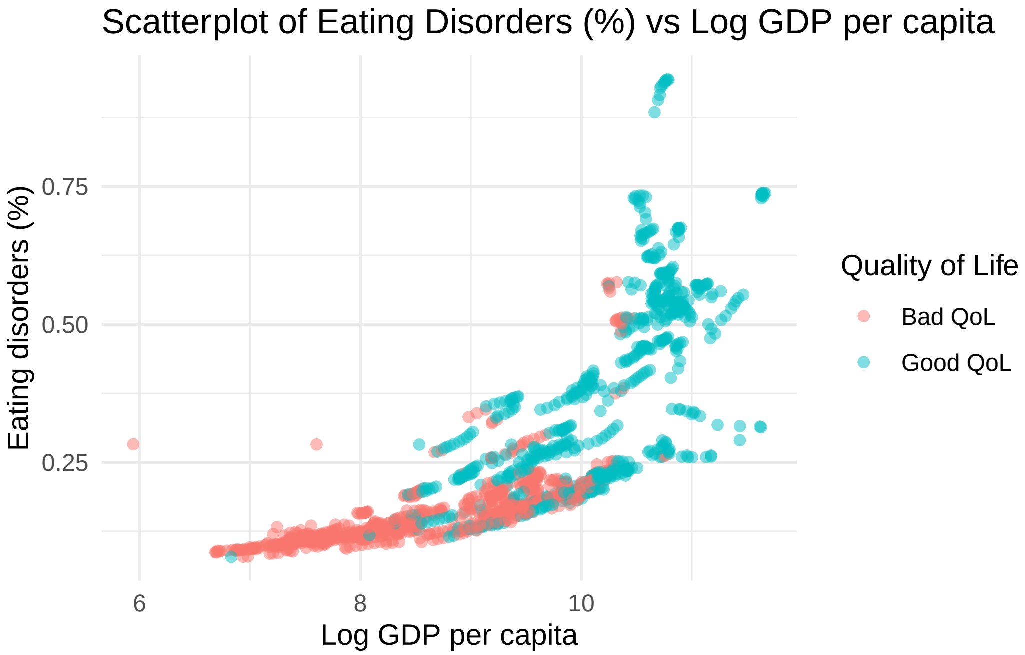
```
Setting direction: controls < cases
```

```
plot(roc_obj, print.thres = "best", print.auc = TRUE)
```



The ROC curve analysis further enhances the evaluation of the logistic regression model's discriminatory power. The Area Under the Curve (AUC) is 0.953, indicating a high level of accuracy in distinguishing between the two classes. The optimal threshold identified by the ROC curve analysis is 0.488, representing a balanced point that maximizes both sensitivity and specificity. At this threshold, the model achieves a sensitivity of 90.3%, denoting its ability to correctly identify individuals with a "Good Quality of Life" (QoL), and a specificity of 89.9%, signifying its proficiency in correctly classifying those without a "Good QoL." This balanced threshold contributes to the model's ability to maintain high performance across both classes. The associated confusion matrix at the optimal threshold is as follows:

| Actual/Predicted | Bad QoL | Good QoL |
|---|---|---|
| Bad QoL | 205 | 23 |
| Good QoL | 19 | 176 |



## References

Bell, D. N., & Blanchflower, D. G. (2019). The well-being of the overemployed and the underemployed and the rise in depression in the UK. *Journal of Economic Behavior & Organization*, *161*, 180-196.

Cherry, K. Mse. (2023, March 3). *A social support system is imperative for Health and well-being.* Verywell Mind. https://www.verywellmind.com/social-support-for-psychological-health-4119970#:~:text=1%20Poor%20social%20support%20has,Alcohol%20use

D'raven, L. T. L., Moliver, N., & Thompson, D. (2015). Happiness intervention decreases pain and depression, boosts happiness among primary care patients. *Primary health care research & development*, *16*(2), 114-126.

Eddie, D., Hoffman, L., Vilsaint, C., Abry, A., Bergman, B., Hoeppner, B., ... & Kelly, J. F. (2019). Lived experience in new models of care for substance use disorder: a systematic review of peer recovery support services and recovery coaching. *Frontiers in psychology*, *10*, 1052.

Graham, C., & Pinto, S. (2019). Using Well-Being Metrics to Assess Social Well-Being and Ill-Being: Lessons from Rising Mortality Rates in the United States. *The Economics of Happiness: How the Easterlin Paradox Transformed Our Understanding of Well-Being and Progress*, 319-353.

Horigian, V. E., Schmidt, R. D., & Feaster, D. J. (2021). Loneliness, mental health, and substance use among US young adults during COVID-19. *Journal of psychoactive drugs*, *53*(1), 1-9.

Teoli D, Bhardwaj A. Quality Of Life. [Updated 2023 Mar 27]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK536962/#

*The world by income and region.* WDI - The World by Income and Region. (2023). https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html