Abstract:

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].

Methods:

Merging Data: 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.

Dataset Information: 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.

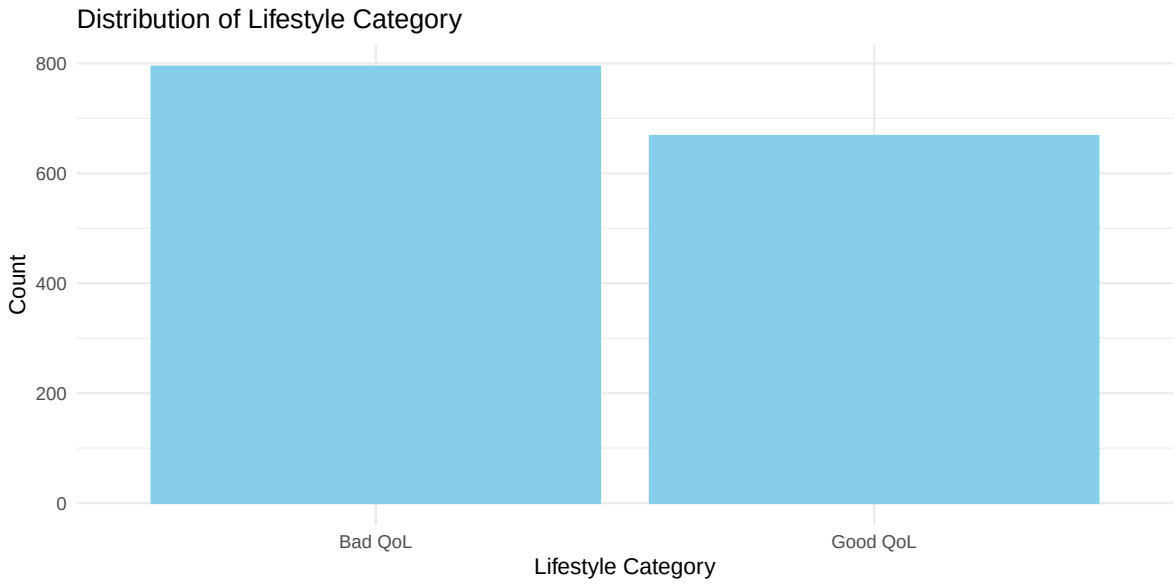
Data Cleaning and Manipulation: The merged dataset had multiple observations per country, each observation representing a different year. Though combining information by country was an option, this would involve averaging observations across years per country and removing information about time. However, 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. Instead, we then include a categorical variable that represents the continent the country resides in to collapse the countries into different regions (continents).

Another transformation was done 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.



EDA:

RQ1 Specific EDA and Steps for Model:

We first identify this is a **prediction problem** and proceed with examining the question as such. Since we are predicting depression % for those in the region (continents), we use logistic regression to build a model and select variables based on *priori* selection. We focus on predictor variables: `Log GDP per Capita`, `Social Support`, `Healthy life expectancy at birth`, `Freedom to make like choice`, `Genorisity`, `Perceptions of corruption`, `Positive affect`, `Negative affect`, `Continent`, and `Year` for our outcome variable `Depression %`. These variables were selected after careful examination of the dataset as a whole and incorporating predictor variables that reflect both economic and social conditions of people in their continent and year.

After building the initial benchmark model with these variables, we asses the significance of the variables using p-values, check the multicollinearity using `VIF()`, and check RMSE to see if our model will be predict well. To improve our model, we first remove variables to see if there is an impact; in the case, `Social Support`. After seeing no improvement, we examine the `Social Support` variable again and break it up into categories where a continent with high social support (above .75) would be considered "High" and those under would be "Low." We then make an interaction term with `Generosity` and `Social support`. After assessing the model again, we see that `Social Support` and the interaction between `Generosity` and `Social support` becomes significant so we keep this current model.

We then evaluate the influential points to see if there are high leverage points that affect the model. After looking at the plots, we remove observatons 221, which had high leverage. After removing these points, we see both `Genorisity` and `Freedom to make life choices` become more significant so we keep them.

After finalizing our model, we make test data to predict on our model to validate our model.

RQ2 Specific EDA and Steps:

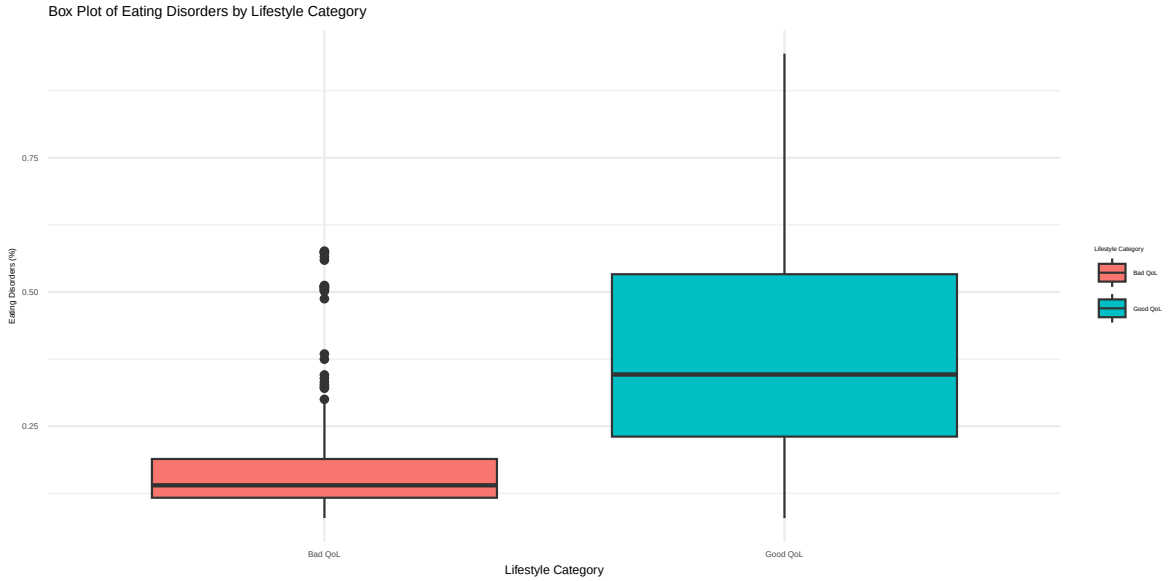
For the second research question, it was important to gain some insights into the variable descriptive statistics based on the binary variable, i.e., statistics for the Bad QoL category vs. the Good QoL category. *Table* _ provides this information.

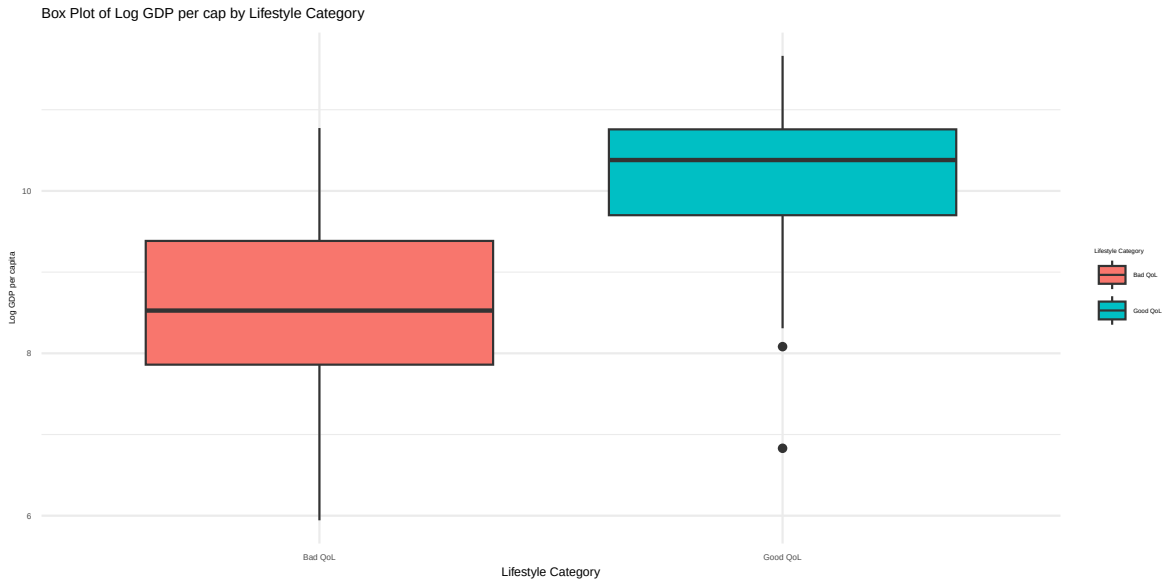
Table _. Descriptive Statistics for variables related to research question 2, factored based on the two levels of Lifestyle Category (Bad QoL vs. Good QoL).

```
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 _).





Results:

General EDA results:

RQ1 Results:

```
tab_model(model6, show.se = TRUE)
```

	Depression (%)			
Predictors	Estimates	std. Error	CI	p
(Intercept)	3.50	0.65	2.23 – 4.77	**<0.001**
Continent [Americas]	-0.69	0.07	-0.82 – -0.56	**<0.001**
Continent [Asia]	-0.57	0.06	-0.70 – -0.45	**<0.001**
Continent [Europe]	-0.49	0.07	-0.64 – -0.35	**<0.001**
Continent [Oceania]	0.03	0.15	-0.26 – 0.32	0.860
Year [2006]	-0.42	0.58	-1.55 – 0.72	0.470
Year [2007]	-0.39	0.58	-1.52 – 0.74	0.497
Year [2008]	-0.31	0.58	-1.44 – 0.83	0.597
Year [2009]	-0.41	0.58	-1.54 – 0.72	0.476
Year [2010]	-0.41	0.58	-1.54 – 0.72	0.478
Year [2011]	-0.40	0.58	-1.53 – 0.73	0.489
Year [2012]	-0.42	0.58	-1.55 – 0.71	0.468
Year [2013]	-0.43	0.58	-1.56 – 0.70	0.456
Year [2014]	-0.48	0.58	-1.61 – 0.65	0.405
Year [2015]	-0.46	0.58	-1.59 – 0.68	0.430
Year [2016]	-0.45	0.58	-1.58 – 0.68	0.434
Year [2017]	-0.47	0.58	-1.60 – 0.66	0.413
Log GDP per capita	0.26	0.03	0.20 – 0.31	**<0.001**
Healthy life expectancy at birth	-0.02	0.00	-0.03 – -0.01	**<0.001**
Freedom to make life choices	0.33	0.16	0.01 – 0.65	**0.041**
Generosity	0.27	0.12	0.03 – 0.52	**0.028**
Social support category [Low]	0.13	0.05	0.03 – 0.23	**0.014**
Perceptions of corruption	-0.59	0.11	-0.80 – -0.38	**<0.001**
Positive affect	-0.47	0.26	-0.98 – 0.03	0.068
Negative affect	0.80	0.23	0.34 – 1.25	**0.001**

Generosity × Social support category [Low]	-0.19	0.31	-0.80 – 0.41	0.530
Observations	1317			
R^2 / R^2 adjusted	0.249 / 0.235			

RQ1 Findings:

1. **Log GDP per capita (0.25):** For each one-unit increase in log GDP per capita, depression tends to rise by 0.25 percentage points.
2. **Healthy life expectancy at birth (-0.02):** A one-unit increase in healthy life expectancy is associated with a decrease of 0.02 percentage points in depression.
3. **Freedom to make life choices (0.35):** When people experience a one-unit increase in the freedom to make life choices, depression tends to increase by 0.35 percentage points.
4. **Perceptions of corruption (-0.63):** A one-unit increase in perceptions of corruption is associated with a decrease of 0.63 percentage points in depression.
5. **Generosity (0.25):** An increase in generosity by one unit corresponds to a rise of 0.25 percentage points in depression.
6. **Positive affect (-0.58):** An increase in positive affect is linked to a decrease of 0.58 percentage points in depression when social support is constant.
7. **Social support category (Low) (0.24):** Compared to the "High" category, being in the "Low" social support category is associated with an increase of 0.24 percentage points in depression.
8. **Negative affect (0.84):** An increase in negative affect by one unit is linked to an increase of 0.84 percentage points in depression.
9. **Interaction: Generosity * Social support category (Low) (1.26):** The interaction effect indicates that the combination of low social support and higher generosity is associated with an increase of 1.26 percentage points in depression.

RQ1 Predictions:

1. In a country categorized as part of the Americas continent, during the year 2015, with a log GDP per capita of 9.5, a social support categorized "High", a healthy life expectancy at birth of 70 years, freedom to make life choices of 0.7, generosity of 0.1, and perceptions of corruption at 0.05, the estimated depression percentage is predicted to be around 3.37%.

2. In a country situated in the Asia continent, during the year 2014, with a log GDP per capita of 10.0, a social support categorized as "Low", a healthy life expectancy at birth of 75 years, freedom to make life choices of 0.8, generosity of 0.2, and perceptions of corruption at 0.02, the estimated depression percentage is predicted to be approximately 4.08%.

RQ1 Plot:

RQ2 Results:

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.

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.

Model Results

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

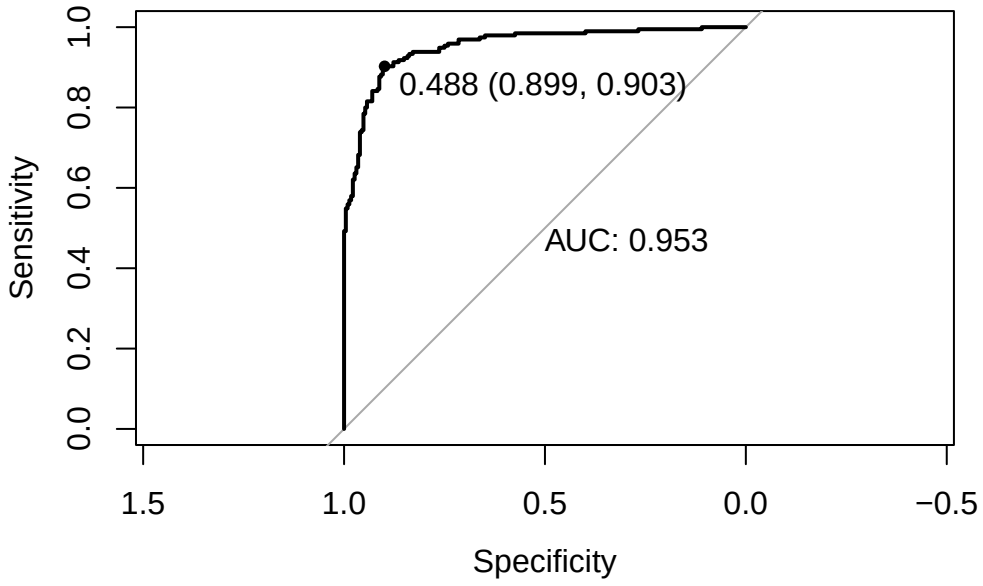
Variable	Estimate	Std. Error	z value	Pr(>
Intercept	-52.35236	67.33324	-0.778	0.4369
Depression (%)	-1.09615	0.26180	-4.187	2.83e-05 ***
Schizophrenia (%)	10.07029	5.55143	1.814	0.0697 .

Variable	Estimate	Std. Error	z value	Pr(>
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 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.

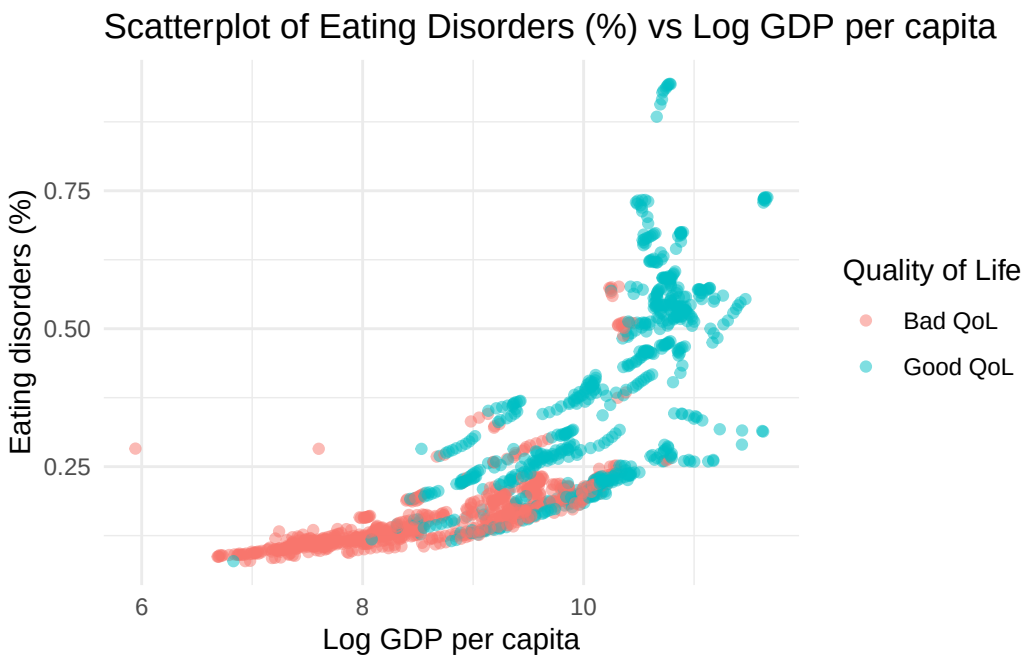


The ROC curve analysis further enhances the evaluation of the logistic regression model's discriminatory power. The Area Under the Curve (AUC) is 0.953, indicating a high level of accuracy in distinguishing between the two classes. The optimal threshold identified by the ROC curve analysis is 0.488, 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

Actual/Predicted	Bad QoL	Good QoL
Good QoL	19	176

RQ2 Result Plot



Conclusion:

Your report will be an 8-10 page self-contained document describing your analysis. It should be written as a professional document that can be understood by someone with limited statistics background (e.g., a client). **You are also required to submit a single QMD file that includes your code for the EDA and analysis.** The report should be organized as follows:

- **Abstract**: A few sentences describing the purpose of the analysis, the data, and key results
- **Introduction**: Provide more background on the data and research questions. Be sure to cite the data and background information appropriately (APA style is fine). Why are these questions worth exploring?
- **Methods**: Describe the process you used to conduct analysis. This includes EDA and any relevant data cleaning information (e.g., did you exclude missing values? If so, how many? Did you collapse categories for any variables?). Then describe the models you fit, and how you planned to assess the model, including influential points,

multicollinearity, and diagnostics. The organization of this section may depend on your particular dataset/analysis, but you may want to break it into subsections such as "Data," "Models," and "Model assessment." Note that you **do not** present any results in this section. This section reflects your statistical analysis plan. For example, you will state how you went about EDA but you will not present findings of the EDA.

- **Results**: Here you should present results for all aspects of the analysis. The structure of this section should mirror the structure of the methods section. For example, you can start with a few key EDA results (e.g., a table of descriptive statistics), then present model results, then address assessment. This is the section where you will primarily refer to tables and figures. You should have at least 1 figure for each research question that illustrates a key result of the analysis (not a diagnostic plot).
- **Conclusion**: Describe the key takeaways from your analysis, limitations, and future work that can be done to advance knowledge in this area.

A few things to keep in mind:

- You should never refer to actual variable names in the text, tables, or figures. For example, if a variable for height is called "ht___cm," you should always say "height," and the first time you mention it you should state that it is measured in cm. In plots and tables, it should say "height (cm)"
- The report should be produced in Quarto and rendered to PDF. All tables and figures should use appropriate labels.
- Someone should be able to read the abstract and look at the tables and figures and have a pretty good idea of 1) the goals of your analysis, and 2) the key results.
- I recommend using colorblind-friendly color palettes in your figures. It can be even better to differentiate with line types or symbols instead of relying on color.
- Keep you audience in mind! A non-statistician should be able to read your report and have a good idea of what you did, even if they may not understand all of the technical details.
- You can have an appendix if tables or figures are too large to fit into the main text. For example, if you have several predictors, you may want to put a table of model results in the appendix.