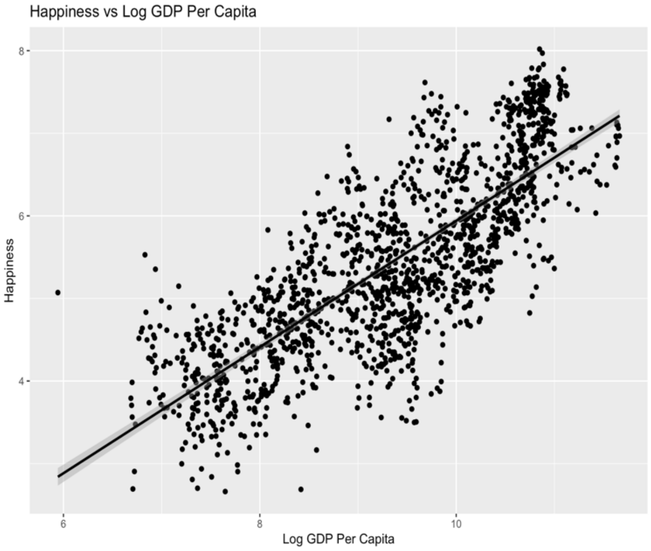
Regression and Time Series – Course Project Report

Kaggle name - Ashwath Ramsundar

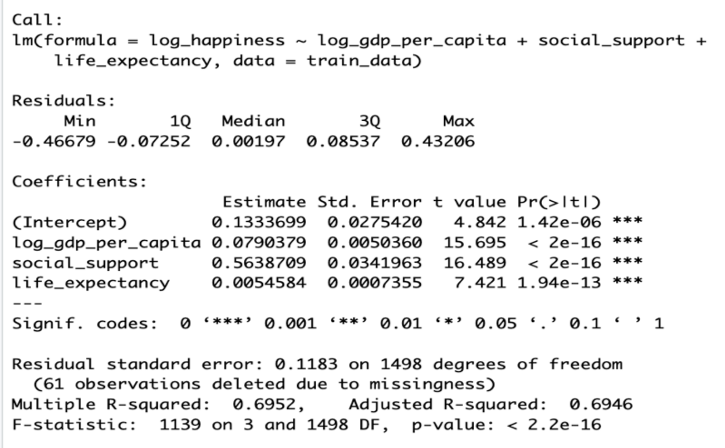
The 2018 World Happiness Report ranks 156 countries by happiness levels, and 117 countries by the happiness of their immigrants. (https://worldhappiness.report/ed/2018/)

I incorporated R program to add regression models to visualize the data and better understand it.

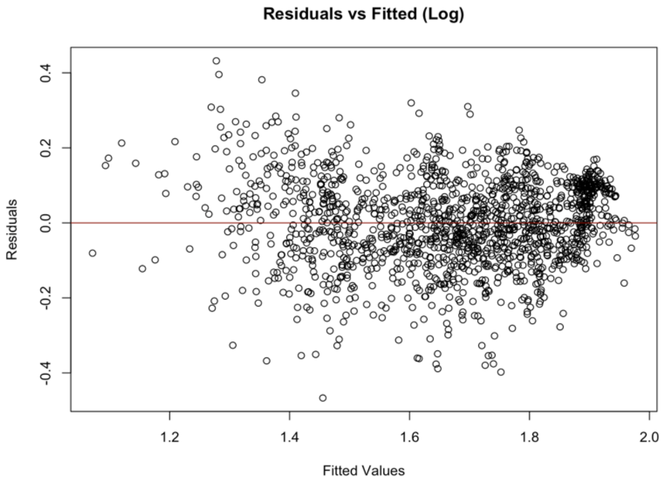


This is a ggplot that compares two parameters happiness and Log GDP per Capita. There is a clear positive linear relationship between Log GDPs Per Capita and Happiness. As Log GDP Per Capita increases, Happiness also tends to increase, suggesting that countries with higher GDP per capita have higher average happiness scores.

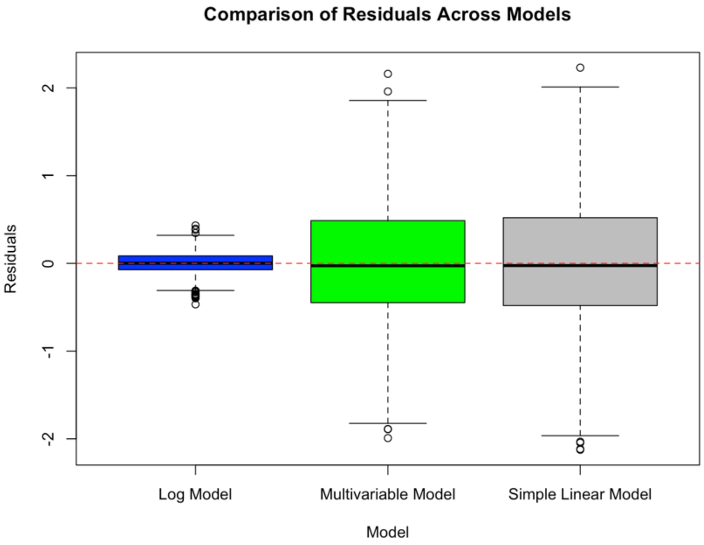
I used log transformation to better stabilize variance and note the heteroscedasticity. The drastic reduction in the residual error, which was at 0.6324 previously, and now at 0.1183 proves that it's a much tighter fit to the data. To further prove the previous, the F statistic is 1139, with a p-value < 2.2e-16 confirms that the fit is significant.



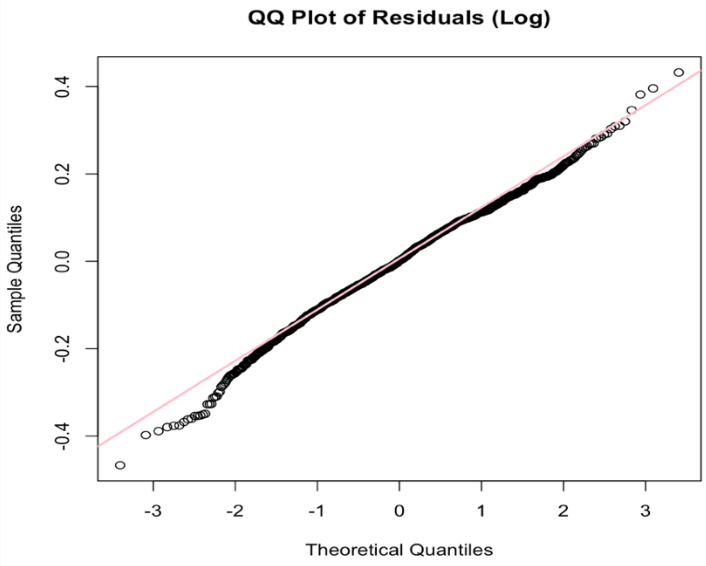
```
Call:
lm(formula = log_happiness ~ log_gdp_per_capita + social_support +
    life_expectancy, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.46679 -0.07252  0.00197  0.08537  0.43206 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        0.1333699  0.0275420   4.842 1.42e-06 ***
log_gdp_per_capita 0.0790379  0.0050360  15.695  < 2e-16 ***
social_support     0.5638709  0.0341963  16.489  < 2e-16 ***
life_expectancy    0.0054584  0.0007355   7.421 1.94e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1183 on 1498 degrees of freedom
  (61 observations deleted due to missingness)
Multiple R-squared:  0.6952,	Adjusted R-squared:  0.6946 
F-statistic:  1139 on 3 and 1498 DF,  p-value: < 2.2e-16
```



The plot suggests that the log-transformed model is a good fit. The residuals are randomly scattered, with no clear patterns, indicating that the key model assumptions—linearity, independence, and homoscedasticity—are being met. Although there are a few outliers, they don't seem to have a significant impact on the model's overall performance. This aligns with the earlier findings, confirming that the log-transformed model is well-fitted.



This boxplot shows the Log Model outperforms the Simple Linear and Multivariable Models, with a tighter spread of residuals and fewer extreme outliers. While all models have medians near zero, indicating unbiased predictions, the Log Model has the least variability, demonstrating its consistency and improved fit due to the log transformation.



The QQ plot shows that the residuals for the log-transformed model are roughly normally distributed. Most of the points follow the red line closely, indicating that the normality assumption is satisfied. While there are a few slight deviations at the ends, they are minor and unlikely to affect the model's overall performance. This reinforces the reliability of the log-transformed model.

Model	R^2	Residual Error	Comments
Simple Linear Model	0.6224	0.6901	Simple model with just one predictor
Multivariable Model	0.6883	0.6324	More predictors improved the model a bit, but the data can be better
Log Model	0.6946	0.1183	Best fit amongst the three. Stabilized variance

<u>Conclusion</u>:

I used a log-transformed linear regression model to predict happiness based on predictors such as log GDP per capita, social support, and life expectancy. This model was chosen to account for the non-linearity in the data and improve homoscedasticity. The R-squared of the model was 0.695, indicating that approximately 69.5% of the variation in happiness could be explained by the predictors. Residual diagnostics, including a residual vs. fitted plot and QQ plot, confirmed that the model assumptions of linearity, independence, and homoscedasticity were met