World Happiness Project Report

Hannah Puisto

12/06/2023

Introduction

In this project, the second project in the HarvardX Data Science Capstone course, our goal is to choose and use a publicly available dataset to solve the problem of our choice by applying machine learning techniques that go beyond standard linear regression. The dataset chosen is about global happiness scores based on societal factors from the World Happiness Report published annually. This project studies the happiness of countries by implementing both a simple sum and a generalized linear model.

Our selection of models comes from the nature of the data itself. The description of the data files from the source provide insight that summing the key factors gives a good estimate, so we choose to explore this. Since the key factor variables are not values from a preset list and are a wide range of numbers, using a generalized linear model would be best to give us more flexibility as opposed to using a K-Nearest Neighbors or Classification model.

The data for this project was imported from Kaggle, posted by the Sustainable Development Solutions Network based on the raw data from the official World Happiness Report. Only data from 2015 to 2019 is available on Kaggle, however, the data for 2020-2023 available directly from the World Happiness Report website is either formatted differently or is not available. The 2022 file doesn't exists on its own, and the Life Expectancy variable in the 2020, 2021, and 2023 file is the actual age instead of being a rated number like in the 2015-2019 files. Therefore, we only use data from 2015 to 2019 for consistency in this report.

The happiness scores are based on answers from the Gallup World Poll to the main life evaluation question asked in the poll, known as the Cantril ladder, which asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. There are six key factors that contribute to the happiness scores: economic production (GDP), social support, life expectancy, freedom of choice, absence of government corruption, and generosity. Each of these key factors contribute to higher evaluations in each country than they are in *Dystopia*.

Dystopia is a hypothetical country with values equal to the world's lowest national averages for each of the other six factors. Dystopia is an imaginary country that has the world's least-happy people. The purpose in establishing Dystopia is to have a benchmark against which all countries can be favorably compared (no country performs more poorly than Dystopia) in terms of each of the six key variables.

The dystopian residual is a constant value, currently set at 1.85. Additionally, there are other residuals, or unexplained components, differing for each country, that reflect the extent to which the six key variables either over- or under-explain happiness scores. We choose these other residuals out of our analysis, keeping in mind that our predicted happiness scores won't be completely accurate because of it.

Given that there are residuals in the calculations of the scores, the six key factors do not directly impact the total score reported for each country, but they do explain why some countries rank higher than others. Simply summing up the six factors will not yield the overall score, but it will give a general idea of where a country might place in the rankings.

We will first take a look at the data in the *Data Summary* subsection before jumping into our analysis. In our exploration of the data, we will also see how each of the key six factors correlate to the scores and if they should be included in our analysis or removed.

In the *Methods and Analysis* section, we will present two models performed in two different ways. The first model is a simple summation of the six key factors to predict happiness scores. The second model is a generalized linear model. In both models, we perform our analysis using all six key factors and using only five. In the *Results* section, we will compare our models and then summarize our findings in the *Conclusion* section.

Data Summary

The raw data from the World Happiness Report contains the following for each year:

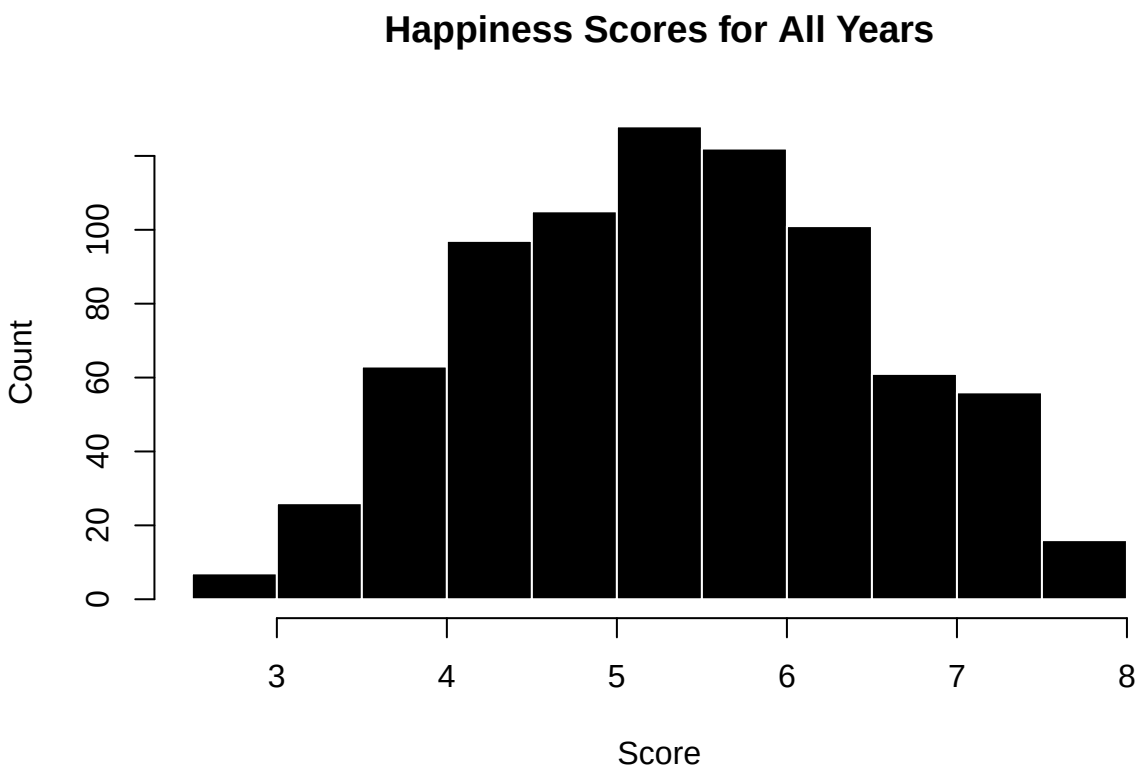
- 2015: 158 countries with 12 variables
- 2016: 157 countries with 13 variables
- 2017: 155 countries with 12 variables
- 2018: 156 countries with 9 variables
- 2019: 156 countries with 9 variables

We first modify the variable names so the same field has the same name in each file, add the year, and then remove variables that are not in all 5 data files. Next, we combine all 5 files into one large dataset consisting of 782 rows of data for the following 10 columns:

```
names(full_data)
```

```
##  [1] "country"        "rank"           "score"          "gdp_per_capita"
##  [5] "social_support" "life_exp"       "freedom"        "gov_trust"
##  [9] "generosity"     "year"
```

The data appears to be normally distributed when looking at the large dataset, as does each year separately when exploratory analysis was performed. The following histogram shows the distribution of happiness scores for all five years of data.



The lower the happiness score, the lower the values of each factor. We've been supplied with six key factors, but we will need to see if each of them actually has a significant impact on the happiness score. The following scatter plots explore the correlation between factors.

```
# Plot gdp_per_capita vs score
ggplot(data = full_data, aes(x = score, y = gdp_per_capita)) +
  geom_point(color='black') +
  geom_smooth(method = "lm", se = TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
# Plot social_support vs score
ggplot(data = full_data, aes(x = score, y = social_support)) +
  geom_point(color='black') +
  geom_smooth(method = "lm", se = TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
# Plot life_exp vs score
ggplot(data = full_data, aes(x = score, y = life_exp)) +
  geom_point(color='black') +
  geom_smooth(method = "lm", se = TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
# Plot freedom vs score
ggplot(data = full_data, aes(x = score, y = freedom)) +
  geom_point(color='black') +
  geom_smooth(method = "lm", se = TRUE)
```

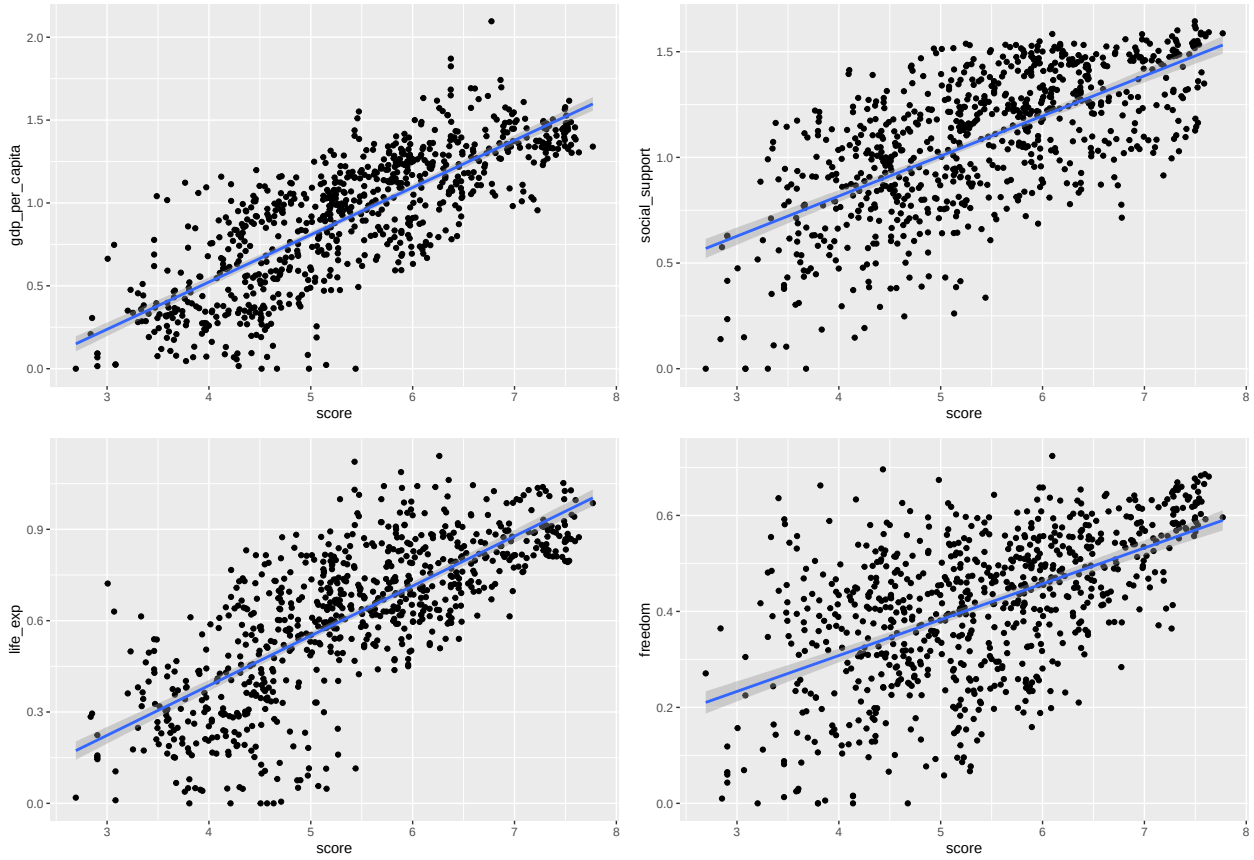
```
## `geom_smooth()` using formula = 'y ~ x'
```

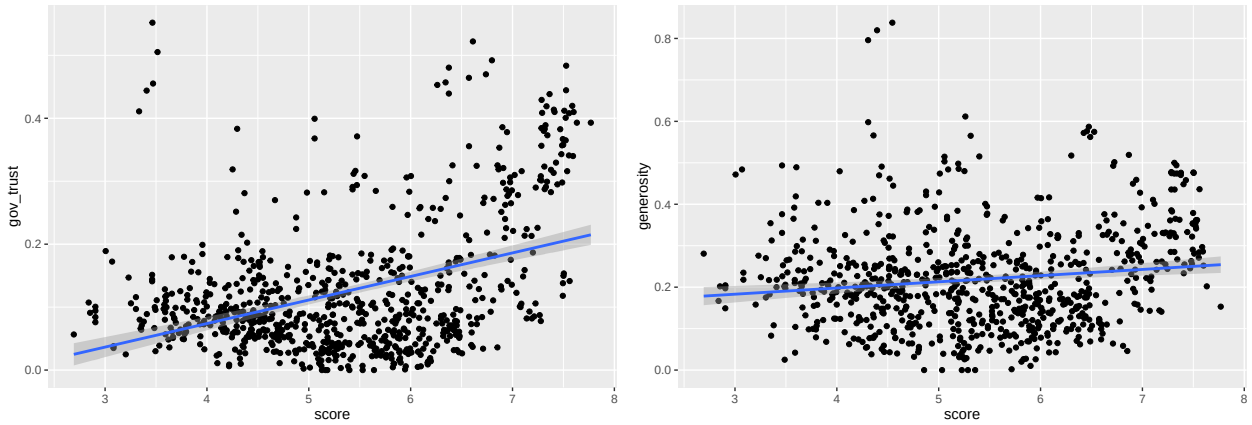
```
# Plot gov_trust vs score
ggplot(data = full_data, aes(x = score, y = gov_trust)) +
  geom_point(color='black') +
  geom_smooth(method = "lm", se = TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
# Plot generosity vs score
ggplot(data = full_data, aes(x = score, y = generosity)) +
  geom_point(color='black') +
  geom_smooth(method = "lm", se = TRUE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```





`Generocity` appears to have little correlation with happiness score, as can be seen by the horizontal best-fit line, so removing it from our model may be beneficial. We will perform our model with and without `Generosity` in the *Methods and Analysis* section. GDP per capita, social support, and life expectancy seem to have the largest correlation with happiness score, but all five other factors appear to have a positive correlation to the score.

We split our large dataset into two subsets: a training dataset that we will use create our models, and a testing dataset that we will use to validate our models. Given the smaller size of the dataset, even with multiple years in it, we do not want to split the data in a way that we don't have enough records in one or the other. Since there are 5 years' worth of data in our large file, we chose to split the data 80% and 20%, training to testing respectively, so the validation dataset has a similar number of rows as a single year's dataset would have.

```
# Create training and testing datasets using p=0.80
train_index <- createDataPartition(full_data$score, times=1, p=0.80, list=FALSE)
train <- full_data[train_index,]
test <- full_data[-train_index,]
```

Methods and Analysis

Model 1a: Summing the Key Factors

The first model we will use is very simple and obvious: adding all of the key factors together and including the 1.85 dystopian residual constant. As discussed in the *Introduction*, our predictions using this method won't be the same as the happiness score due to the other residuals, but it should give us a close estimate. We also use the Root Mean Square Error (RMSE) as our indicator for success. As we are performing a simple sum and not training a model, we will use the full dataset in our sum.

```
# Predict score by sum method
sumvar_model <- full_data %>%
  mutate(pred_score = gdp_per_capita + social_support +
         life_exp + freedom + gov_trust + generosity + 1.85,
         RMSE = RMSE(score, pred_score))

# Calculate the RMSE
sum_rmse = RMSE(sumvar_model$score, sumvar_model$pred_score)
```

The RMSE for summing the key factors and the dystopian constant is 0.5805411.

Model 1b: Summing the Key Factors without Generosity

Next, we apply the same first model except this time we remove `Generosity`.

```
# Predict score by sum method
sumvar_nogen_model <- full_data %>% mutate(pred_score = gdp_per_capita + social_support +
                                                        life_exp + freedom + gov_trust + 1.85,
         RMSE = RMSE(score, pred_score))

# Calculate the RMSE
sum_nogen_rmse = RMSE(sumvar_nogen_model$score, sumvar_nogen_model$pred_score)
```

The RMSE for summing the key factors (without `Generosity`) and the dystopian constant is 0.6820853.

Model 2a: Use the Generalized Linear Model (GLM)

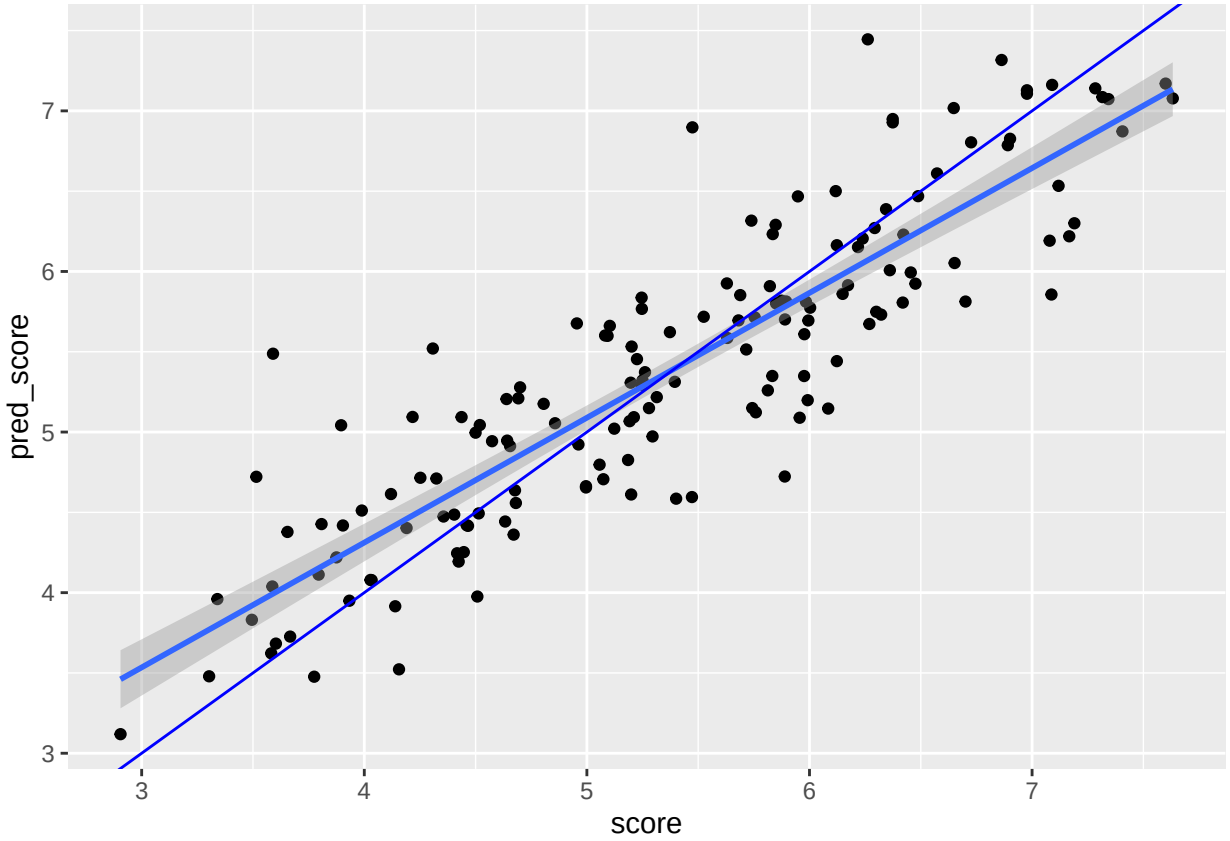
For our second model, we will use our split datasets created in the *Data Summary* section to fit a generalized linear model (GLM). Once we fit the data, we apply predicted results to our validation `test` dataset.

```
# Predict score using GLM
data_fit <- glm(score ~ gdp_per_capita + social_support + life_exp + freedom + gov_trust + generosity,
                data = train)

# Add predicted scores to test data frame
results <- test %>%
  mutate(pred_score = as.numeric(predict.glm(data_fit, newdata=test)),
         RMSE = RMSE(score, pred_score))

# Calculate the RMSE
glm_rmse = RMSE(results$score, results$pred_score)
```

By plotting the predicted scores to the actual scores, we can see how close our predictions are.

```
# Plot predicted scores vs actual scores with x=y line
ggplot(data = results, aes(score, pred_score)) +
  geom_point(color='black') +
  geom_smooth(method = "lm", se = TRUE) +
  geom_abline(color='blue')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The RMSE for applying the glm to the key factors is 0.5110445. From our model, we can calculate the coefficients to determine the equation for our model.

```
##    (Intercept) gdp_per_capita social_support       life_exp        freedom
##      2.1834052      1.1758644      0.6496119      0.8811847      1.5524362
##      gov_trust     generosity
##      0.9959357      0.5291949
```

$$\hat{y} = 2.183 + 1.176x_{GDP} + 0.65x_S + 0.881x_L + 1.552x_F + 0.996x_T + 0.529x_G \quad (1)$$

For the above equation, predicted score = $\hat{y}$, GDP per capita score = x_{GDP}, Social Support score = x_S, Life Expectancy score = x_L, Freedom score = x_F, Trust score = x_T, Generosity score = x_G.

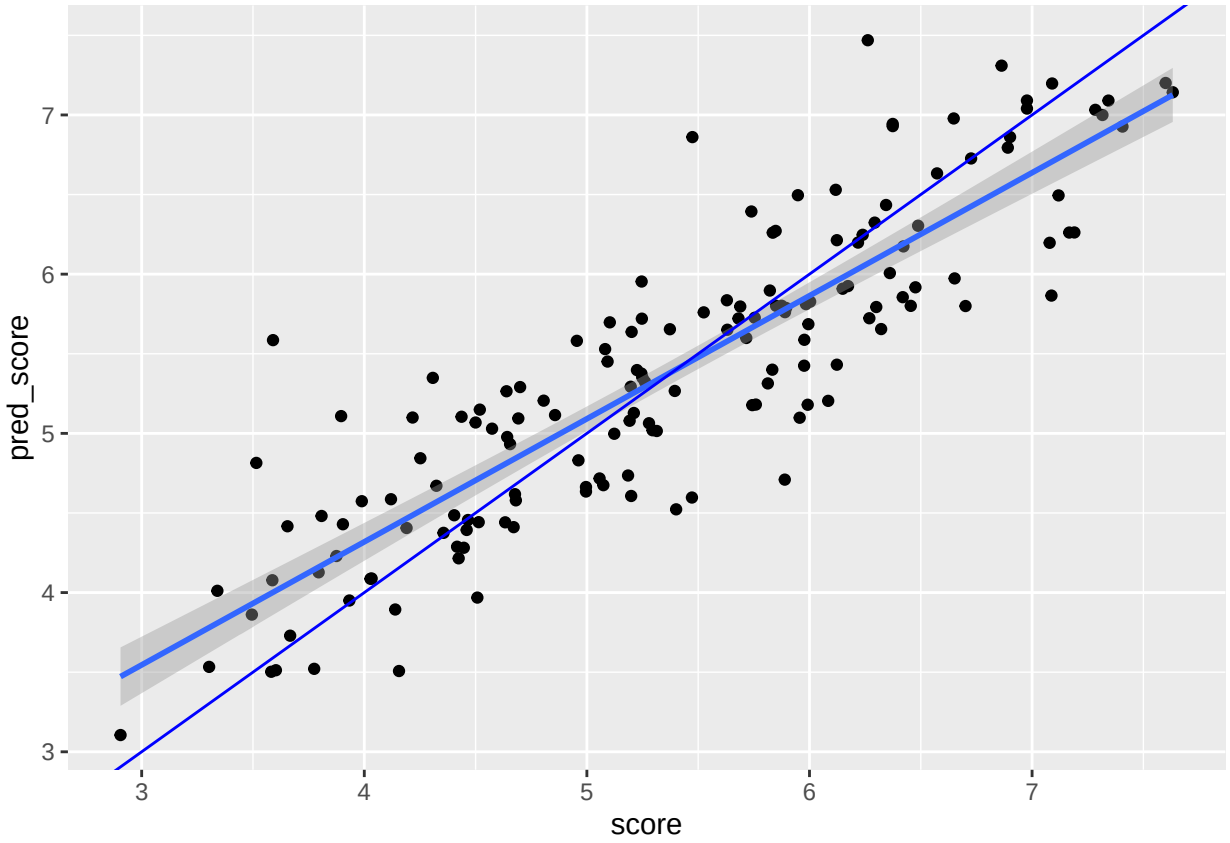
Model 2b: Use the GLM without Generosity

Now, we apply our second model again but remove `Generosity`.

```
# Predict score using GLM without generosity
data_fit_nogen <- glm(score ~ gdp_per_capita + social_support + life_exp + freedom + gov_trust,
                      data = train)

# Add predicted scores to test data frame
results_nogen <- test %>%
  mutate(pred_score = as.numeric(predict.glm(data_fit_nogen, newdata=test)),
```

```
        RMSE = RMSE(score, pred_score))

# Calculate the RMSE
glm_nogen_rmse = RMSE(results_nogen$score, results_nogen$pred_score)

# Plot predicted scores vs actual scores with x=y line
ggplot(data = results_nogen, aes(score, pred_score)) +
  geom_point(color='black') +
  geom_smooth(method = "lm", se = TRUE) +
  geom_abline(color='blue')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The RMSE for applying the glm to the key factors (without `Generosity`) is 0.519053. The calculated coefficients are then used to determine the new equation for our model.

```
##    (Intercept) gdp_per_capita social_support       life_exp        freedom
##      2.2745790      1.1526516      0.6224814      0.8921216      1.6678991
##      gov_trust
##      1.1630488
```

$$\hat{y} = 2.275 + 1.153x_{GDP} + 0.622x_S + 0.892x_L + 1.668x_F + 1.163x_T \tag{2}$$

.

Results

The four models discussed in the *Methods and Analysis* section above returned the following RMSEs:

- Sum of Factors Model: 0.5805411
- Sum of Factors Model Without Generosity: 0.6820853
- GLM Model: 0.5110445
- GLM Model Without Generosity: 0.519053

We can clearly see that the GLM model, including `Generosity` is the best model for our data. As a final step to this report, we apply our best model to the full dataset.

```
# Predict score using GLM
full_fit <- glm(score ~ gdp_per_capita + social_support +
         life_exp + freedom + gov_trust + generosity,
                data = full_data)

# Add predicted scores to full data frame
results_full <- full_data %>%
  mutate(pred_score = as.numeric(predict.glm(full_fit, newdata=full_data)),
         RMSE = RMSE(score, pred_score))

# Calculate the RMSE
glm_full_rmse = RMSE(results_full$score, results_full$pred_score)
```

The final RMSE after applying our best model to the full dataset containing 5 years' worth of data is 0.5474916.

Conclusion

In this report, we explored the correlation between six key factors in worldwide happiness using data from the World Happiness Report. Though we identified one key variable appeared to have little to no correlation to the overall happiness scores, removing it from our model did not reduce the error. After performing our analysis, we noticed that using the GLM model and including `Generosity` was the best model. When embarking on this analysis, I expected that the first model, summing all six key factors and the dystopian residual, would yield the best predictions. I was surprised to find that the GLM model performed the best.

The dataset for the chosen topic was relatively small, which limited the model selection possibilities and could have caused over-training to occur in our model. The missing residuals caused a whole in the data that could have given more insight into the validity of each model. If the data available for all years contained the same variables and included the residual values, a more accurate fit could be calculated for our predictions. In the future, additional training models could be applied to the historical data to predict upcoming years' results.

References

- The source for the data files is Kaggle: https://www.kaggle.com/datasets/unsdsn/world-happiness/
- The World Happiness Report website : https://worldhappiness.report/data/)
- Frequently Asked Questions for the World Happiness Report: https://worldhappiness.report/faq/
- Appendix 1 to the World Happiness Report: https://happiness-report.s3.amazonaws.com/2021/Appendix1WHR2021C2.pdf