Assessed Practical I: Predicting the Olympic Games

Yashdhar Gandhi

2024-04-07

Student Number: “201783891” Email: ‘mm23ysg@leeds.ac.uk (mailto:mm23ysg@leeds.ac.uk)’

1 INTRODUCTION

The Olympic Games have grown into a cultural performance of global proportion. Participants in the Games—athletes, officials, dignitaries, press, technicians, support personnel, as well as artists, performers, scientists, and world youth campers attending ancillary congresses and exhibitions—now number in the scores of thousands and are drawn from as many as 151 nations.(MacAloon, 2023). Medal counts have become more and more important throughout time as a measure of a nation’s accomplishments both nationally and athletically. However, the traditional medal count may not fully reflect a country’s sporting success, especially for smaller nations or those with differing socioeconomic backgrounds.

In this assessed practical, we use statistical learning to forecast the medal count of countries competing in the Olympic Games. Our study focuses on the relationship between a country’s population, GDP, and number of medals earned. Although the total number of medals has historically been the focus, per capita metrics and accounting for economic inequality have become important aspects to examine when assessing a country’s success.

The dataset comprises information from 71 countries that competed in the last three Olympics, including population, GDP, and medal tallies. We use statistical methods, such as linear regression, to identify trends in the population and GDP-based medal counts. In addition, we study the value of log-transforming outcomes and create unique regression models. This practical allows us to apply statistical approaches while simultaneously evaluating the prediction efficacy of various models in Olympic outcomes, so improving our understanding of the factors that influence a country’s success in this global athletic event.

2 Data Preparation

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
#Load the dataset
olympic_data <- read.csv("C:/Users/yashd/Downloads/medal_pop_gdp_data_statlearn.csv")
str(olympic_data)
```

```
## 'data.frame':    71 obs. of  6 variables:
##  $ Country   : chr  "Algeria" "Argentina" "Armenia" "Australia" ...
##  $ GDP       : num  188.7 446 10.2 1371.8 63.4 ...
##  $ Population: int  37100000 40117096 3268500 22880619 9111100 353658 1234571 9461400 1095
1266 192376496 ...
##  $ Medal2008 : int  2 6 6 46 7 2 1 19 2 15 ...
##  $ Medal2012 : int  1 4 3 35 10 1 1 12 3 17 ...
##  $ Medal2016 : int  2 4 4 29 18 2 2 9 6 19 ...
```

```
head(olympic_data)
```

```
##      Country     GDP Population Medal2008 Medal2012 Medal2016
## 1    Algeria  188.68   37100000         2         1         2
## 2  Argentina  445.99   40117096         6         4         4
## 3    Armenia   10.25    3268500         6         3         4
## 4  Australia 1371.76   22880619        46        35        29
## 5 Azerbaijan   63.40    9111100         7        10        18
## 6    Bahamas    7.79     353658         2         1         2
```

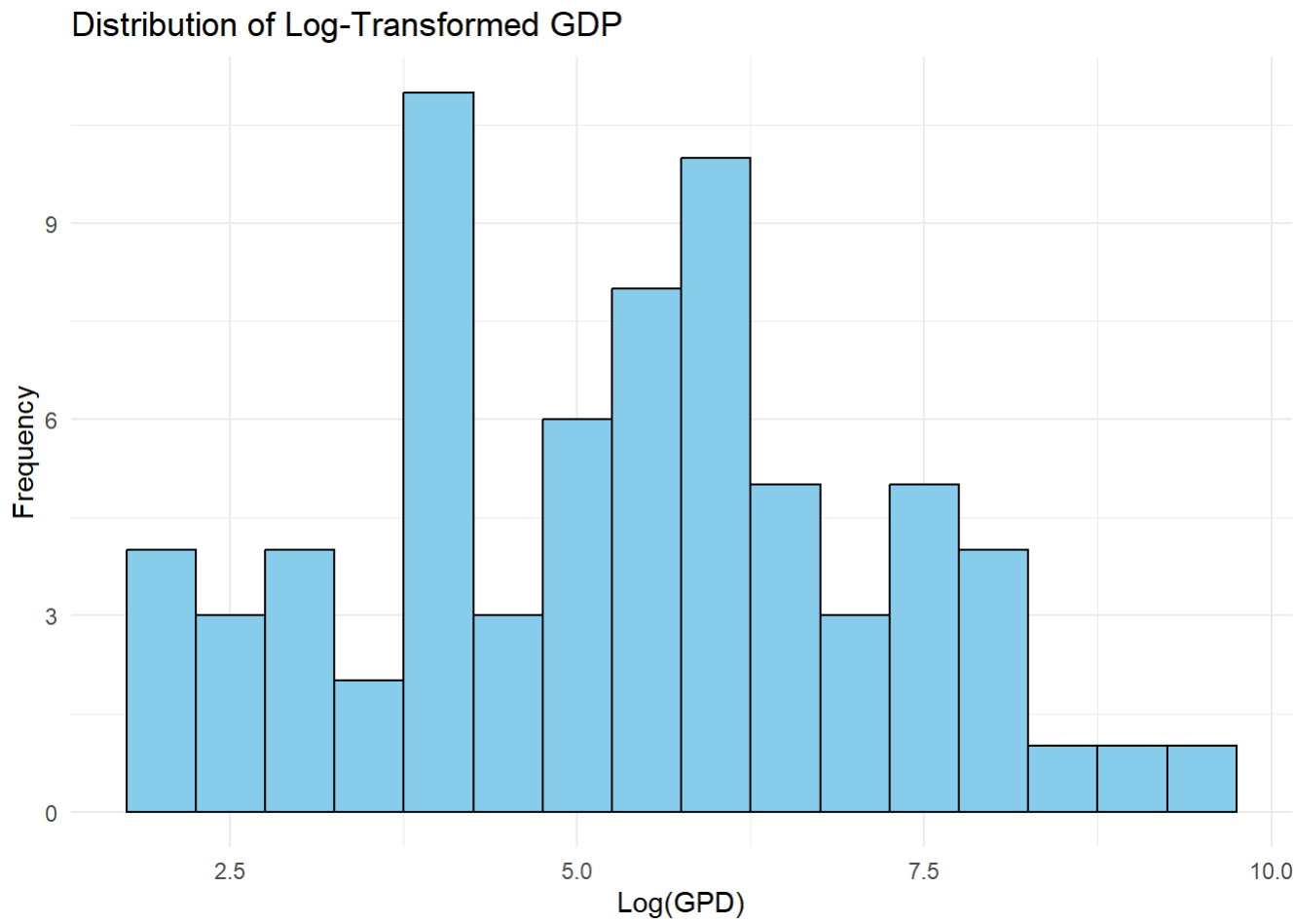
```
# Check for missing values
missing_values <- sum(is.na(olympic_data))
print(paste("Number of missing values in the dataset:", missing_values))
```

```
## [1] "Number of missing values in the dataset: 0"
```

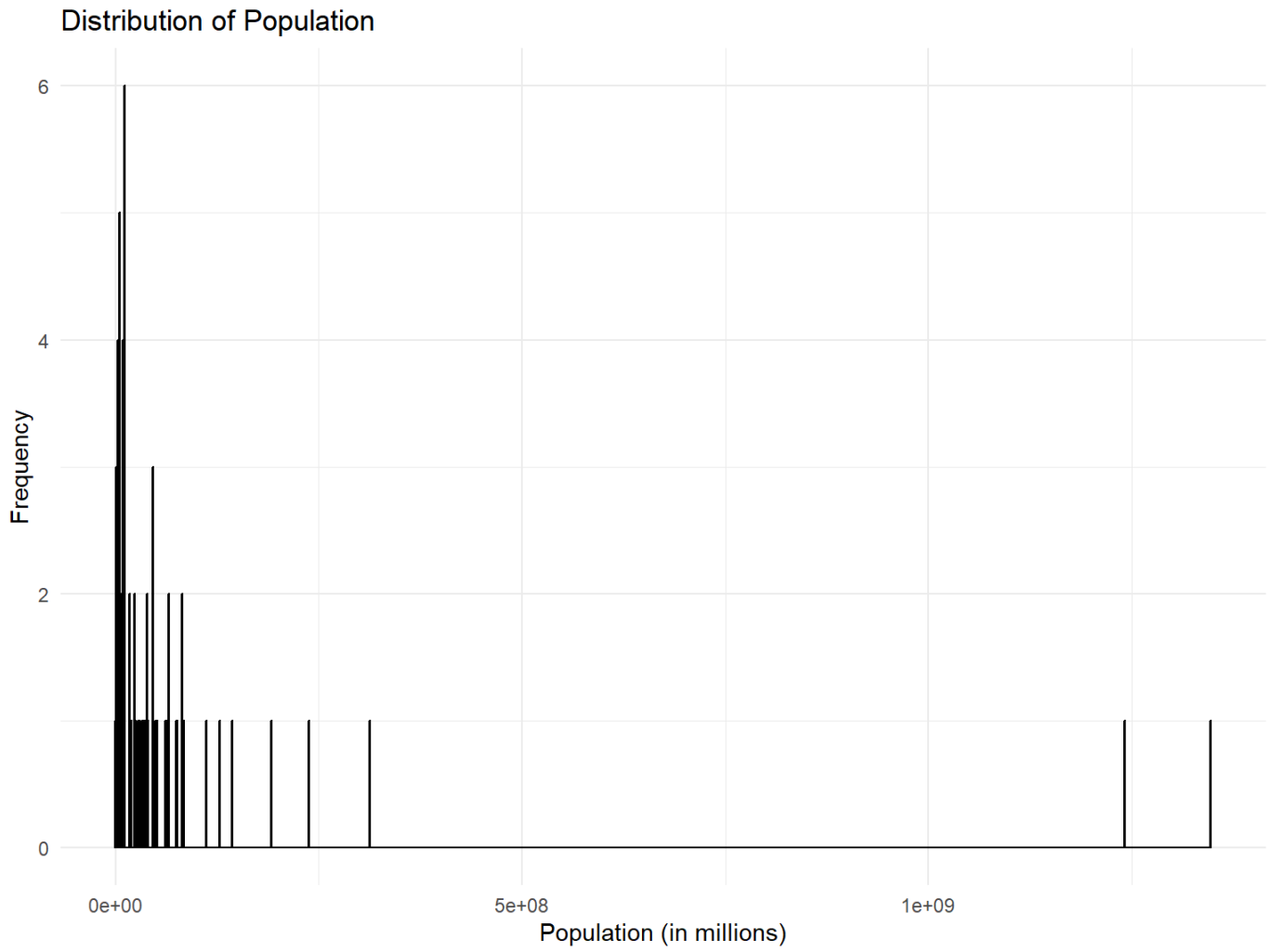
```
# Log-transform the GDP variable (transform the GDP variable to address skewness.)
olympic_data$GDP_log <- log(olympic_data$GDP)
head(olympic_data)
```

```
##      Country     GDP Population Medal2008 Medal2012 Medal2016  GDP_log
## 1    Algeria  188.68   37100000         2         1         2 5.240052
## 2  Argentina  445.99   40117096         6         4         4 6.100297
## 3    Armenia   10.25    3268500         6         3         4 2.327278
## 4  Australia 1371.76   22880619        46        35        29 7.223850
## 5 Azerbaijan   63.40    9111100         7        10        18 4.149464
## 6    Bahamas    7.79     353658         2         1         2 2.052841
```

```
ggplot(data = olympic_data, aes(x = GDP_log)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Log-Transformed GDP",
       x = "Log(GPD)",
       y = "Frequency") +
  theme_minimal()
```

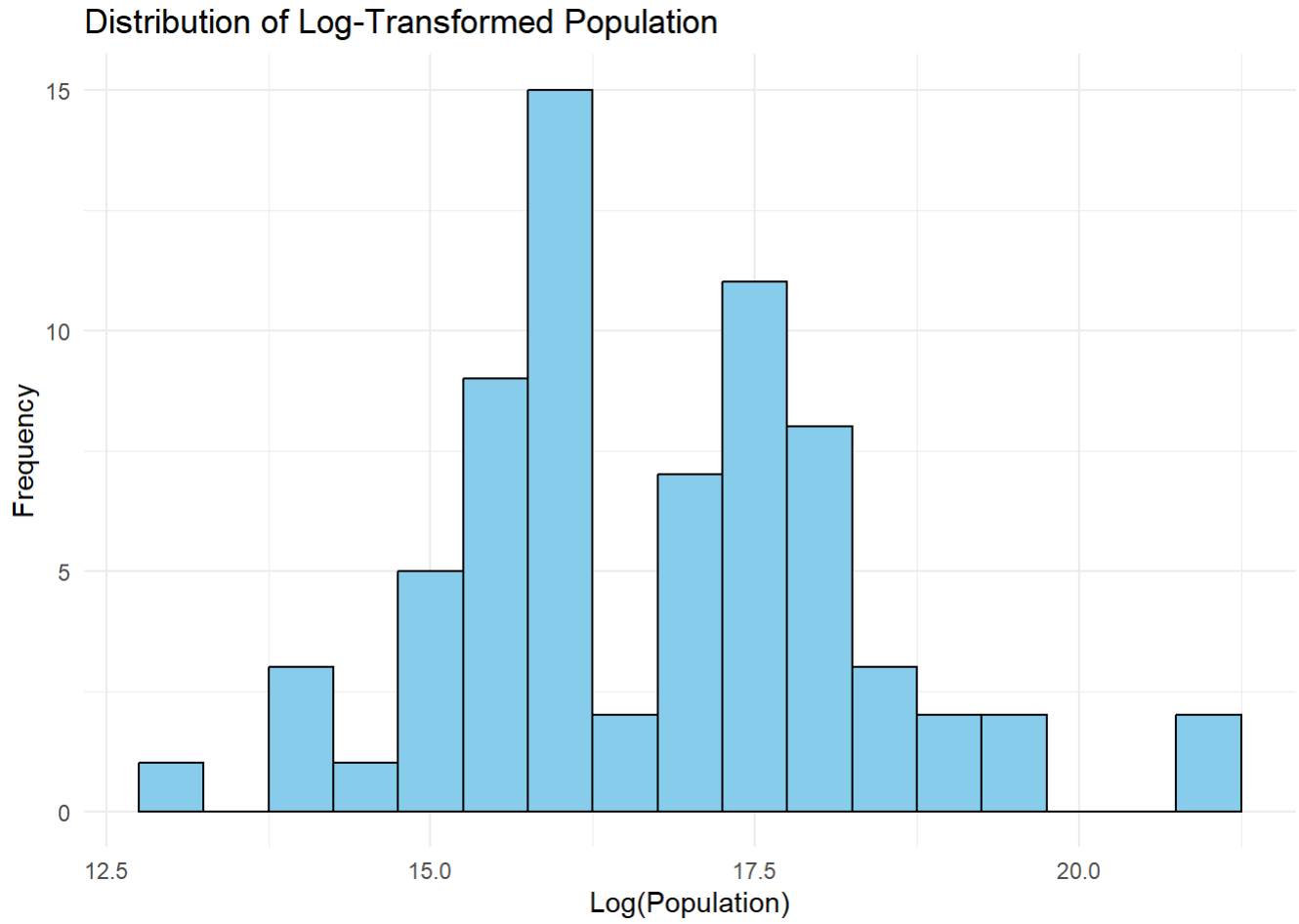


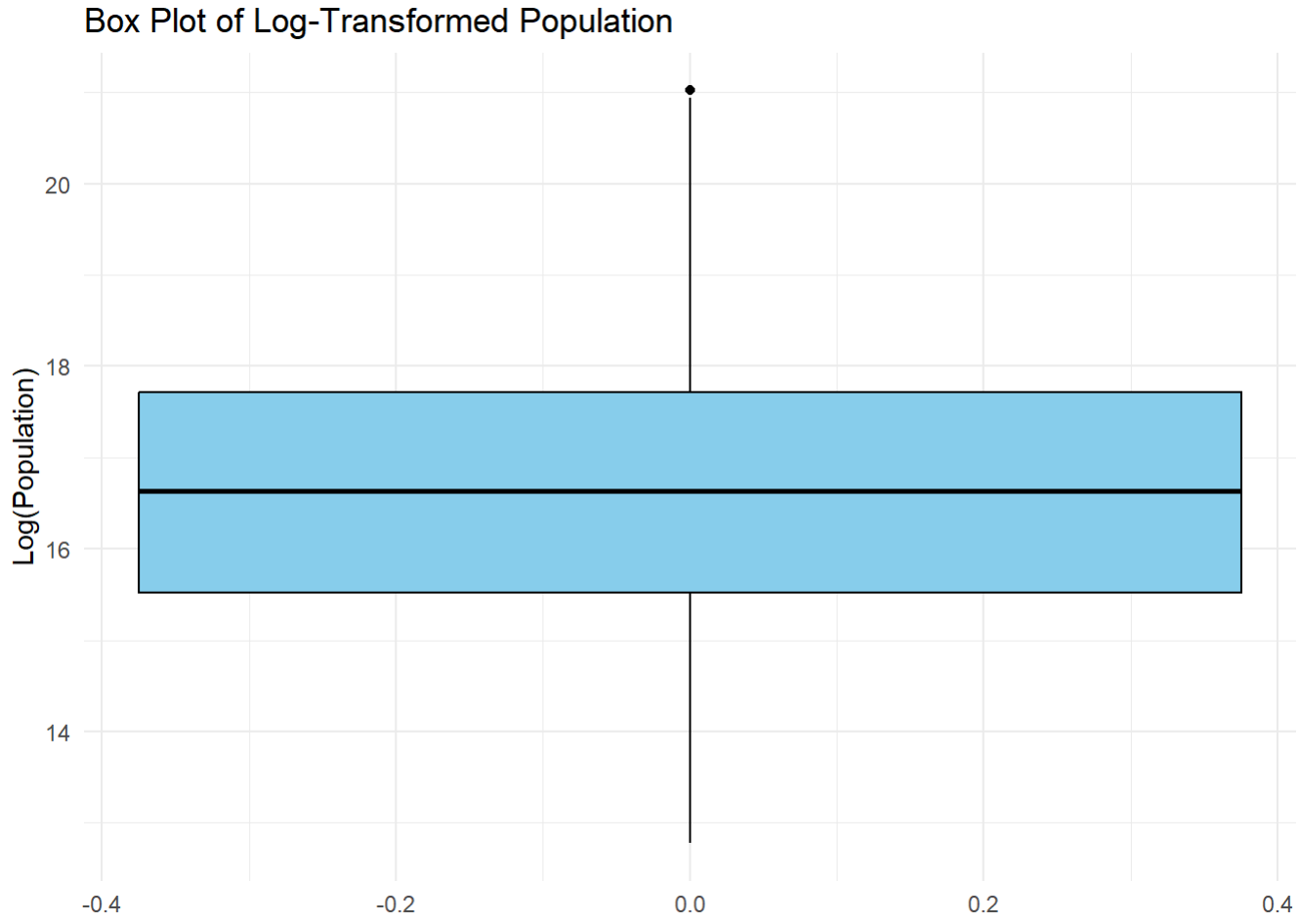

```
library(ggplot2)

ggplot(data = olympic_data, aes(x = Population)) +
  geom_histogram(binwidth = 1e6, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Population",
       x = "Population (in millions)",
       y = "Frequency") +
  theme_minimal()
```

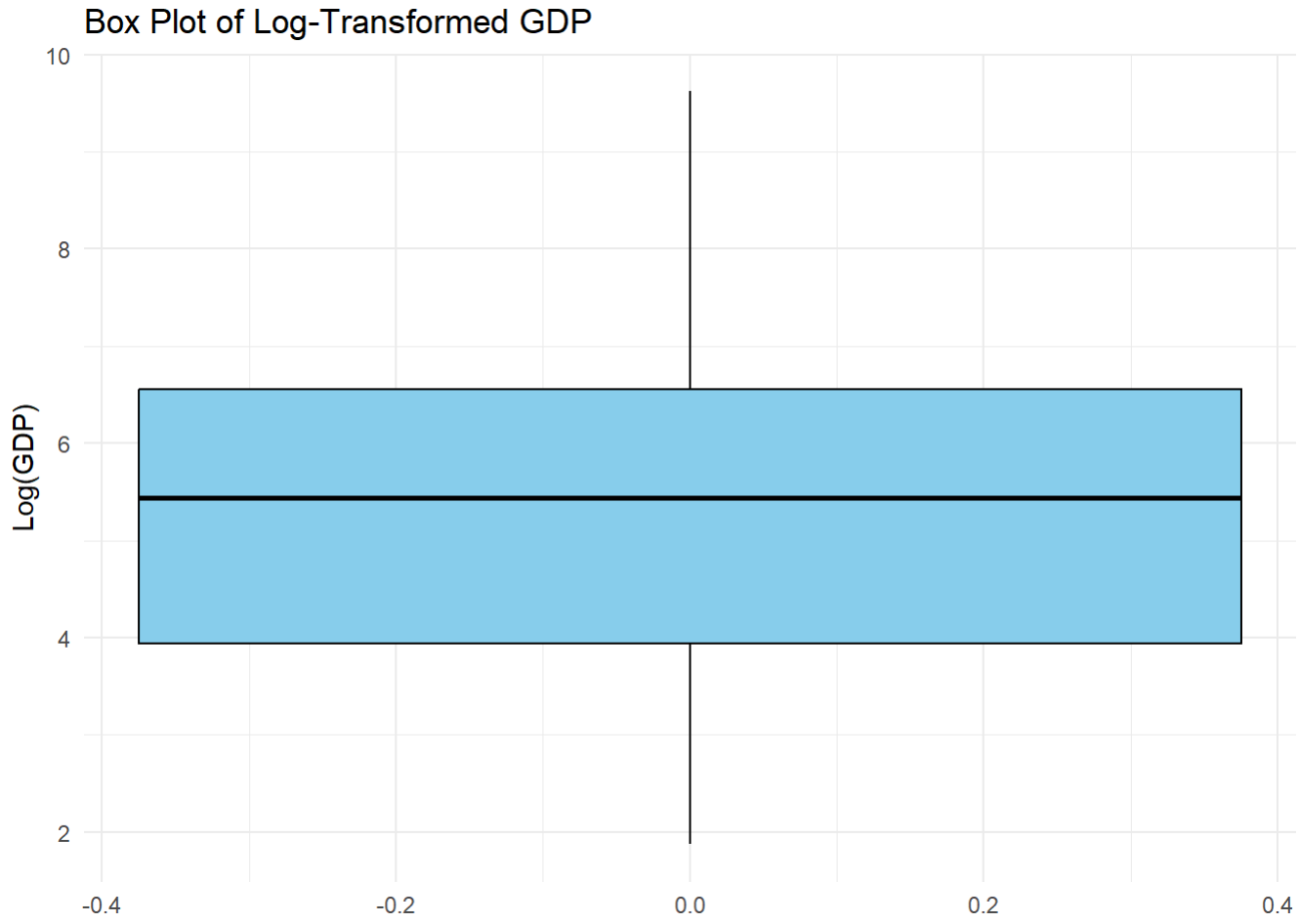


Since the population data exhibits positive skewness, we'll apply a logarithmic transformation to make the distribution closer to a normal distribution.

```
library(ggplot2)
# Log-transform the population variable
olympic_data$Population_log <- log(olympic_data$Population)

ggplot(data = olympic_data, aes(x = Population_log)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Log-Transformed Population",
       x = "Log(Population)",
       y = "Frequency") +
  theme_minimal()
```

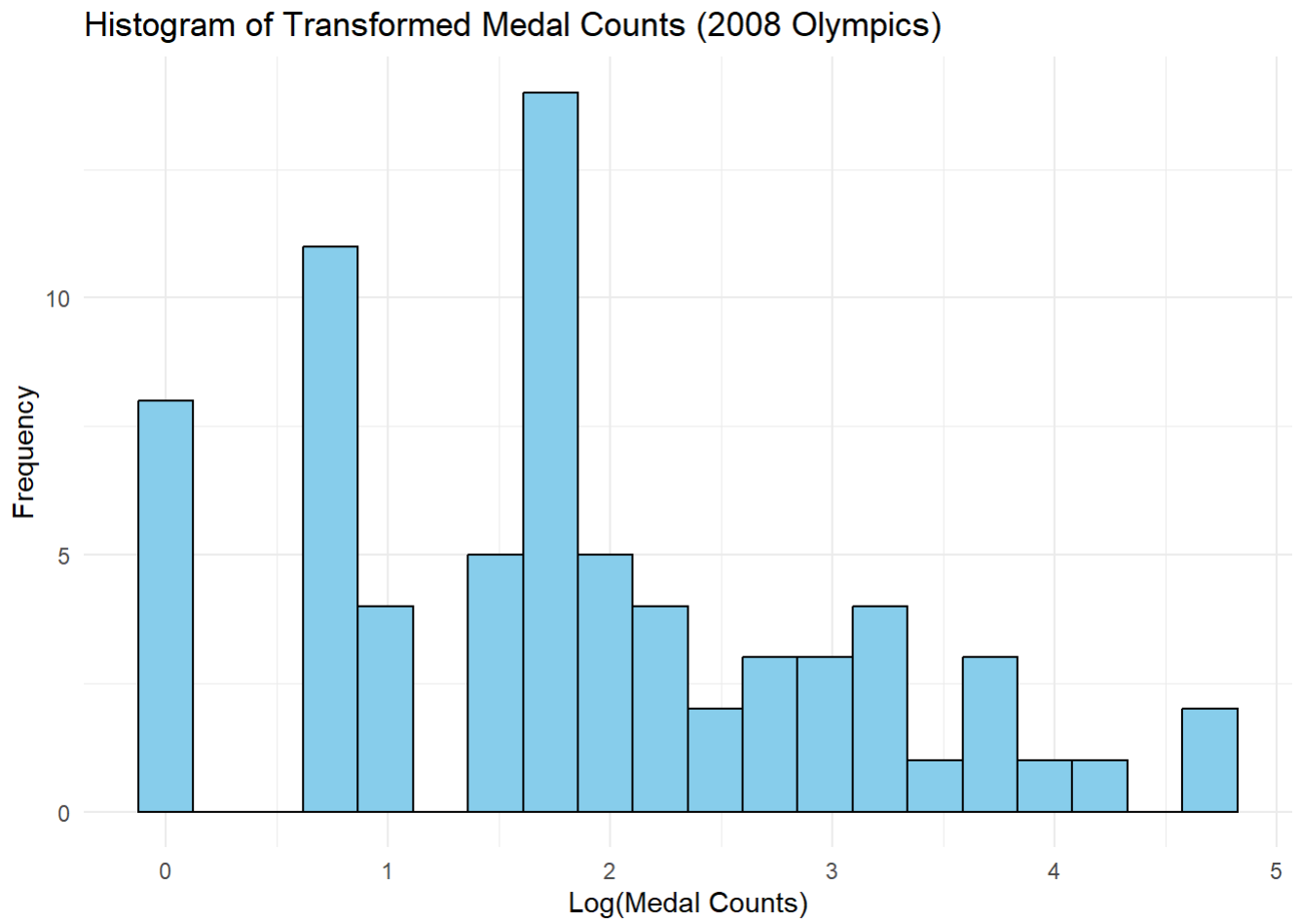

Distribution of Log-Transformed Population
Frequency
15
10
5
0
12.5
15.0
17.5
20.0
Log(Population)

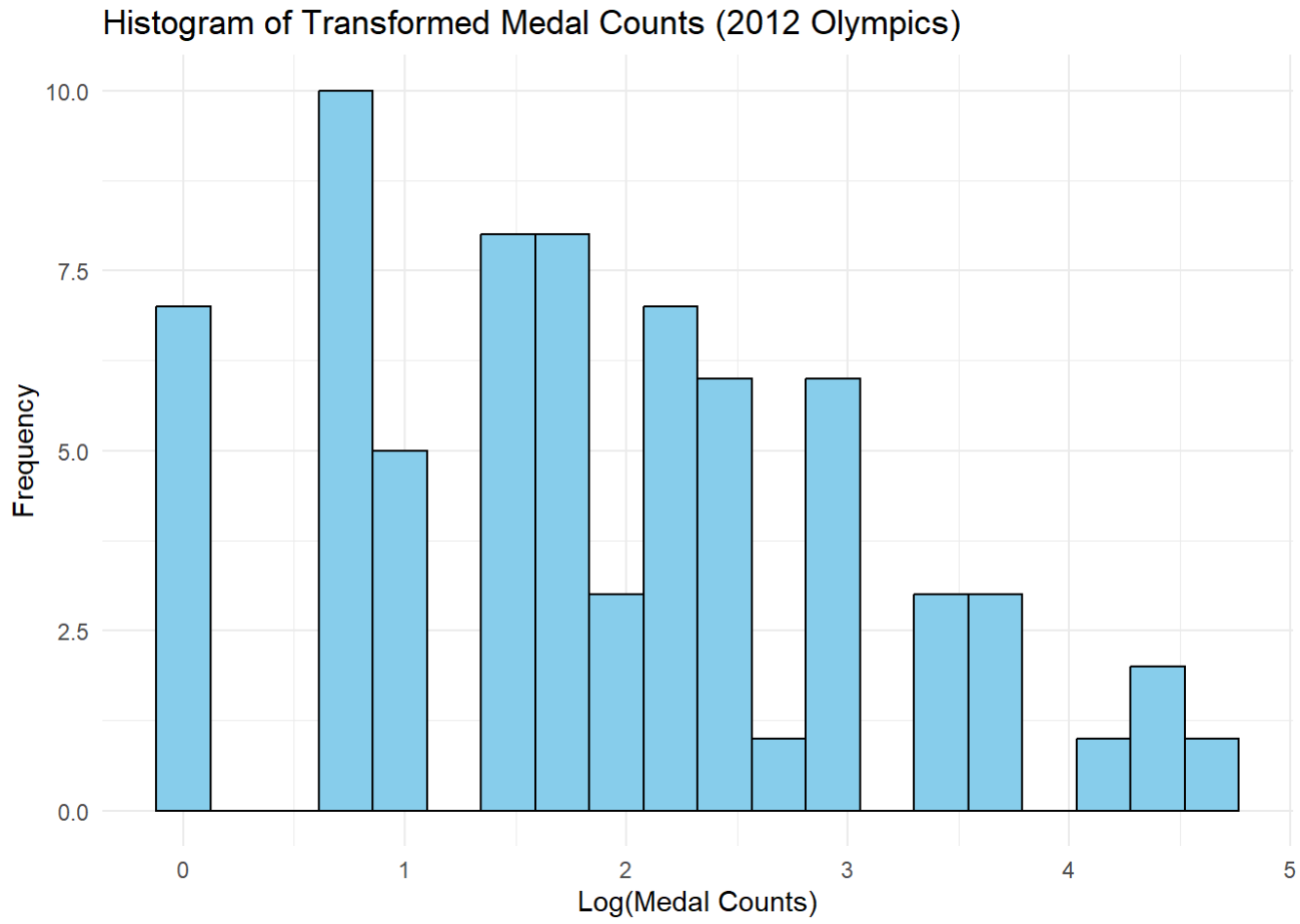

Box Plot of Log-Transformed Population
Log(Population)
20
18
16
14
-0.4
-0.2
0.0
0.2
0.4

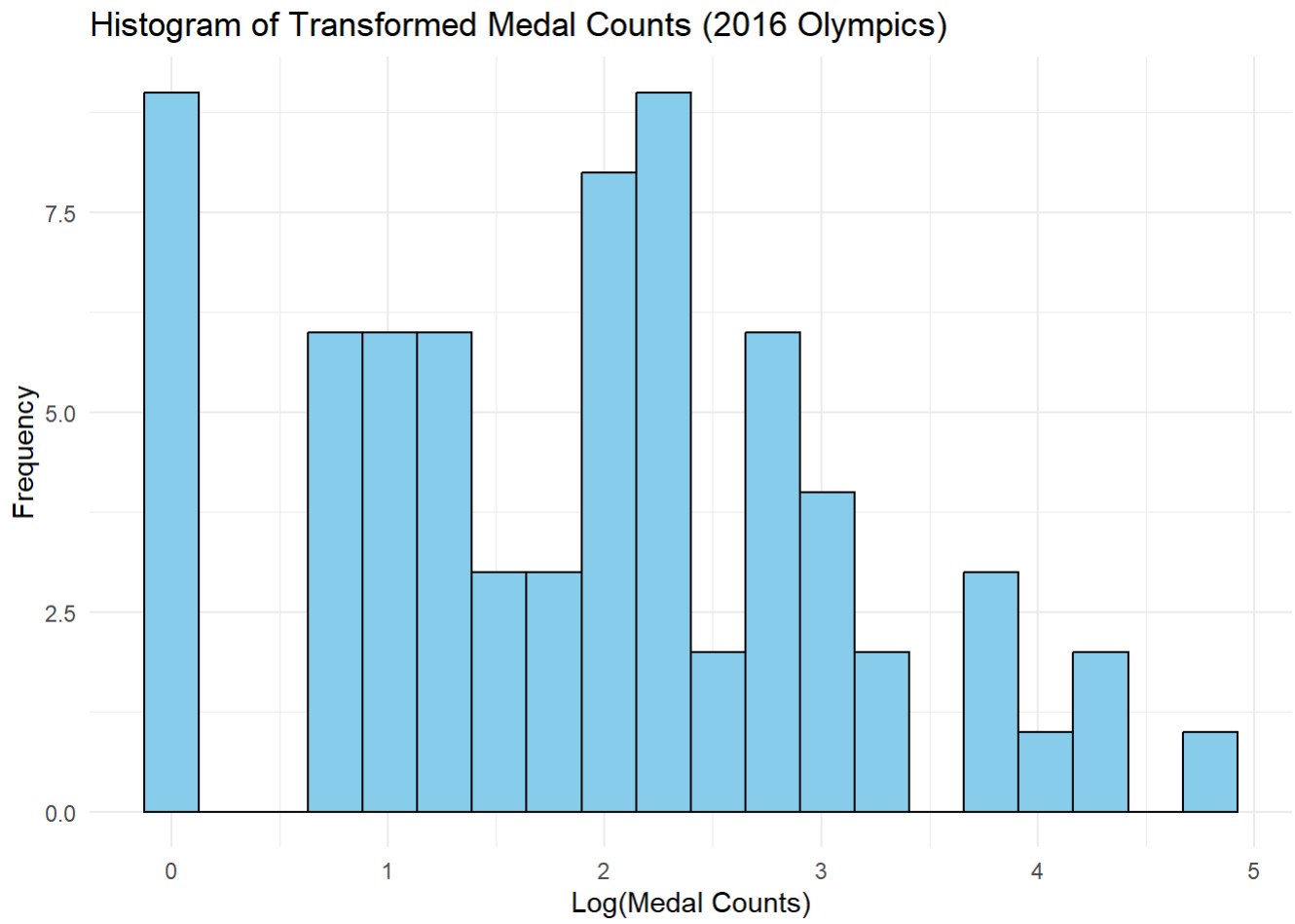


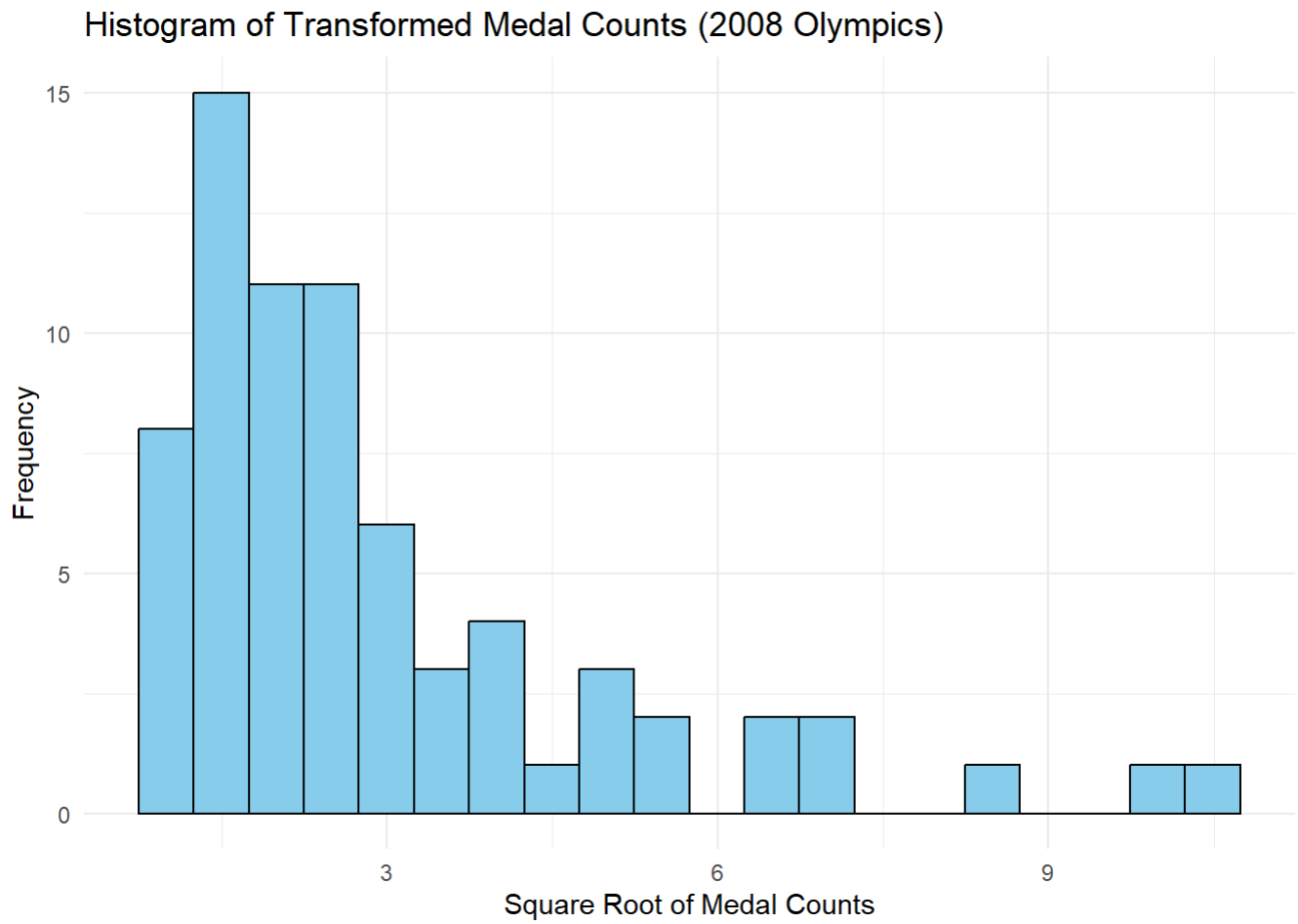
After visual inspection using box plots and scatter plots, it appears that there are no outliers in the log-transformed population variable. The scatter plot also indicates a positive relationship between the log-transformed population and log-transformed GDP, without any apparent outliers. Therefore, we can proceed

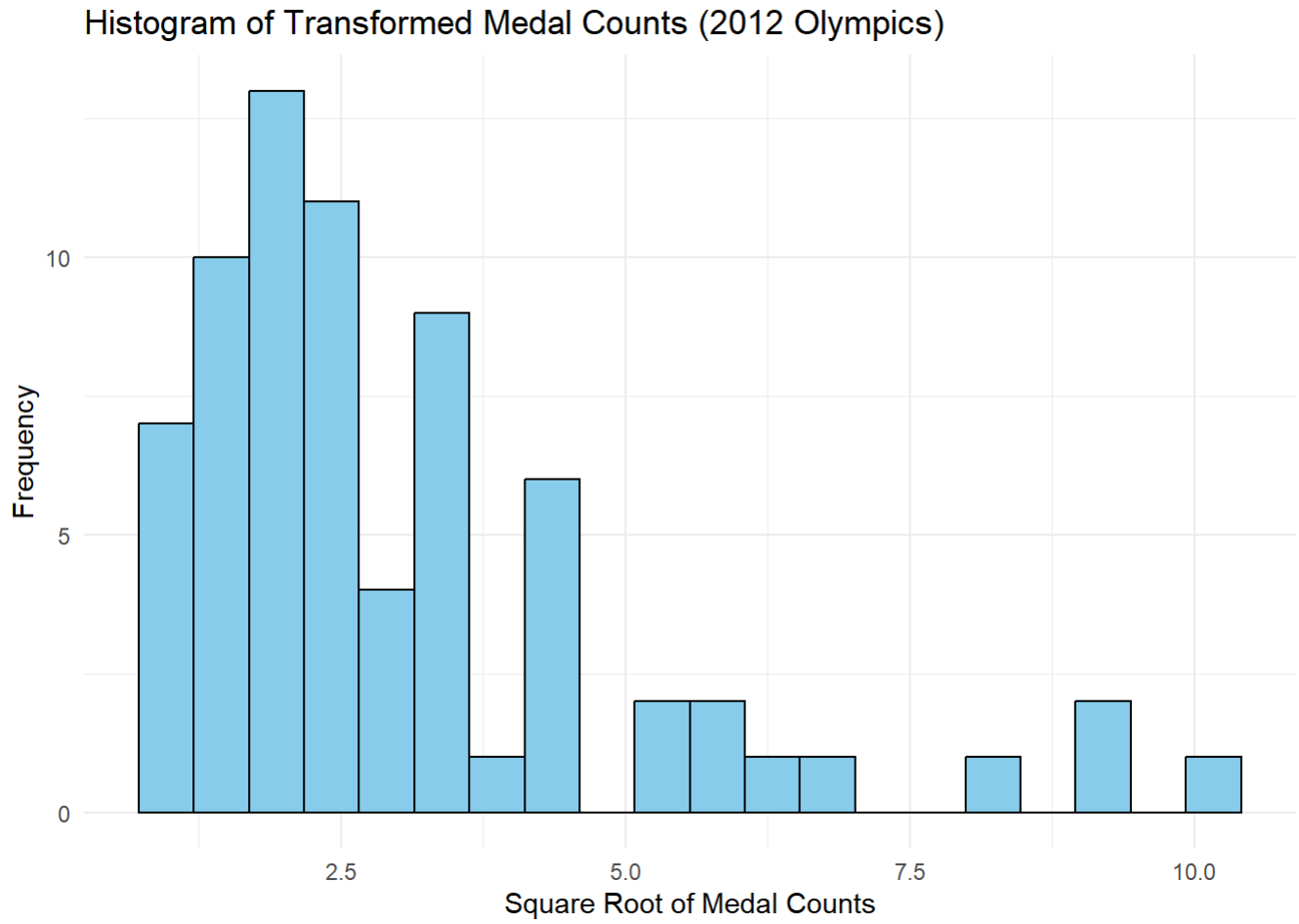
with our analysis without the need for outlier removal or adjustment.

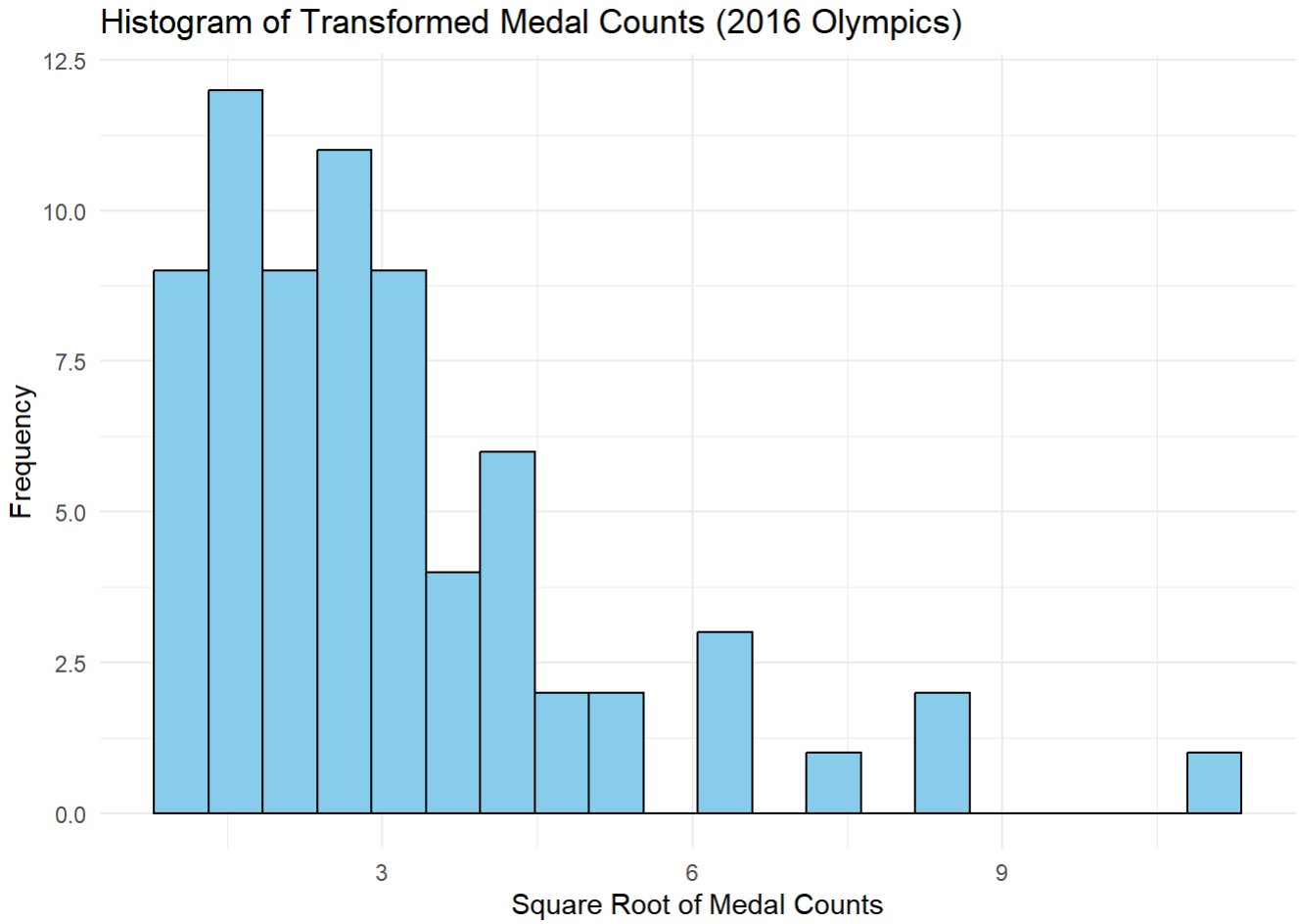

Histogram of Transformed Medal Counts (2008 Olympics)
Frequency
0
5
10
0
1
2
3
4
5
Log(Medal Counts)


Histogram of Transformed Medal Counts (2012 Olympics)
Frequency
0.0
2.5
5.0
7.5
10.0
0
1
2
3
4
5
Log(Medal Counts)


Histogram of Transformed Medal Counts (2016 Olympics)
Frequency
0.0
2.5
5.0
7.5
0
1
2
3
4
5
Log(Medal Counts)


Histogram of Transformed Medal Counts (2008 Olympics)
Frequency
15
10
5
0
3
6
9
Square Root of Medal Counts

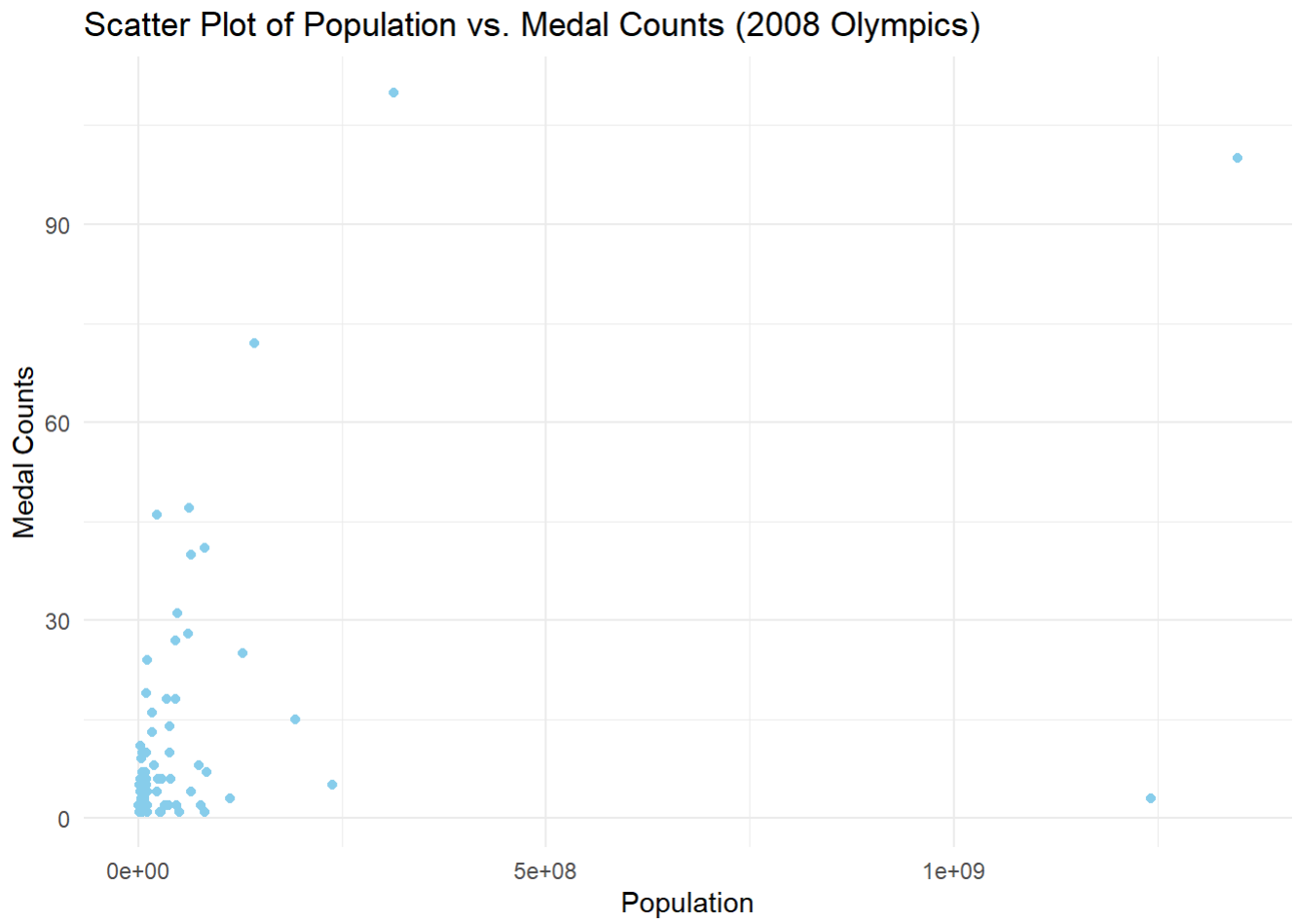

Histogram of Transformed Medal Counts (2012 Olympics)
Frequency
10
5
0
2.5
5.0
7.5
10.0
Square Root of Medal Counts

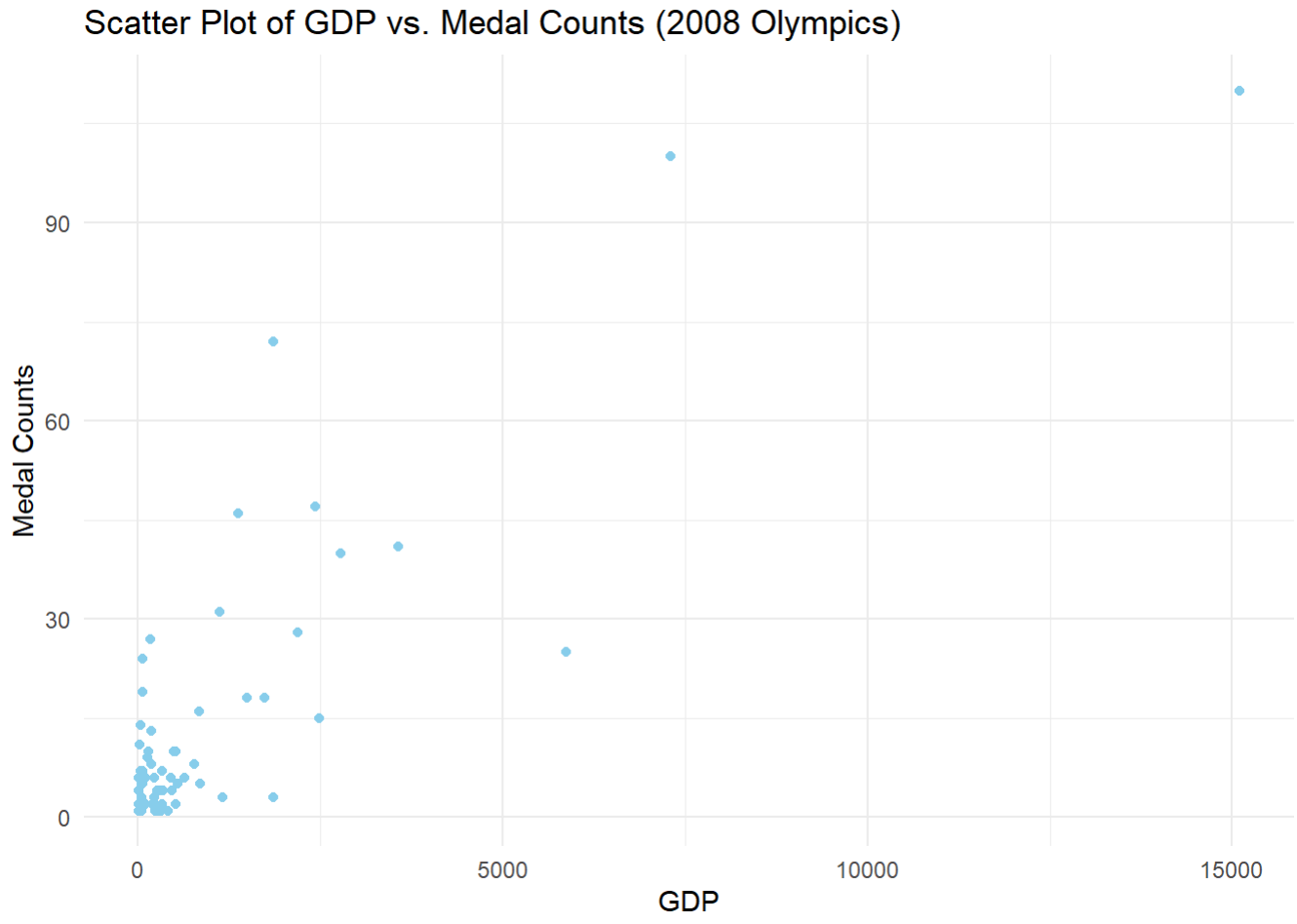


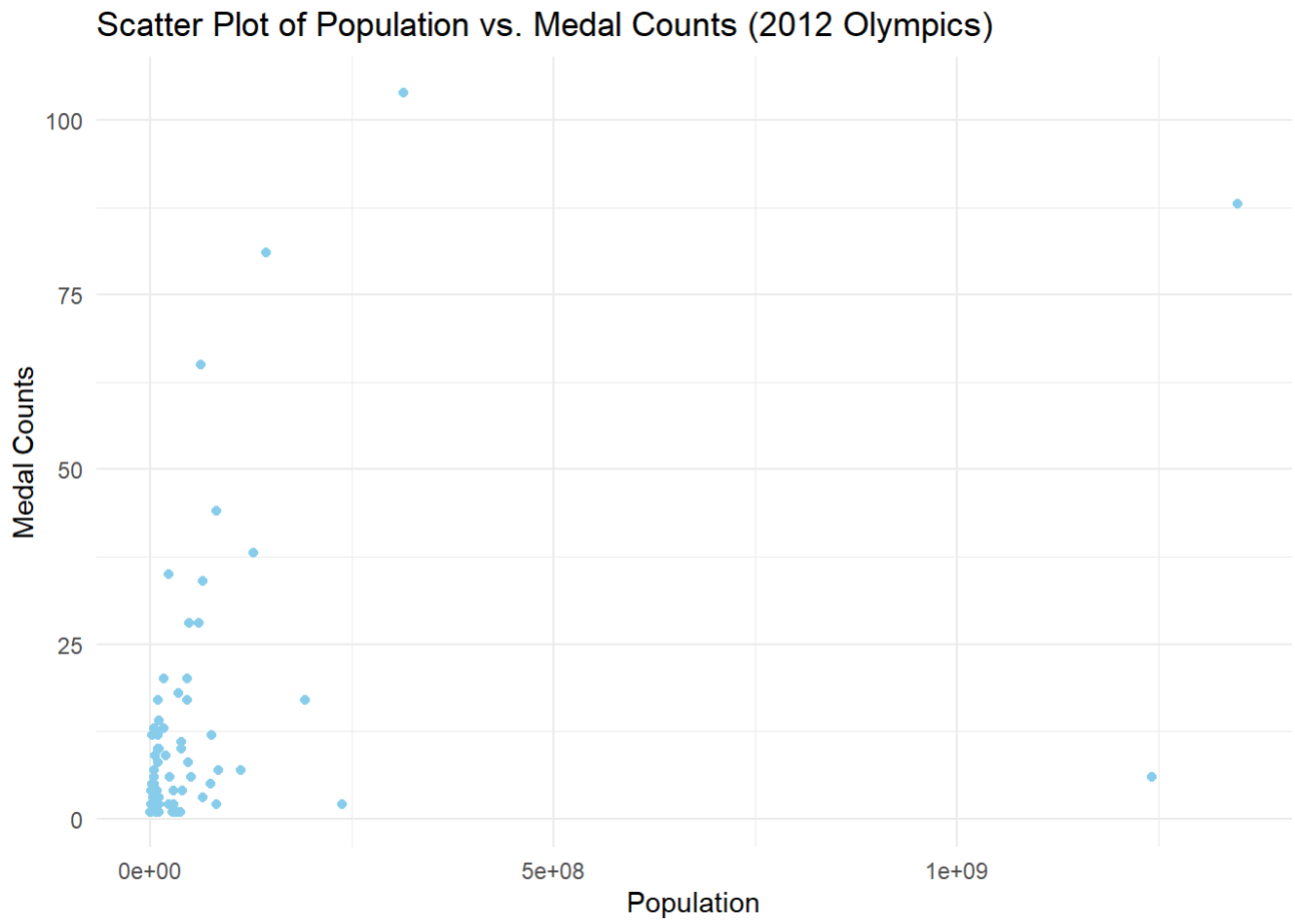
We have decided to use the logarithmic transformation for the medal counts instead of the square root transformation. Logarithmic transformation is chosen due to its effectiveness in addressing positive skewness.

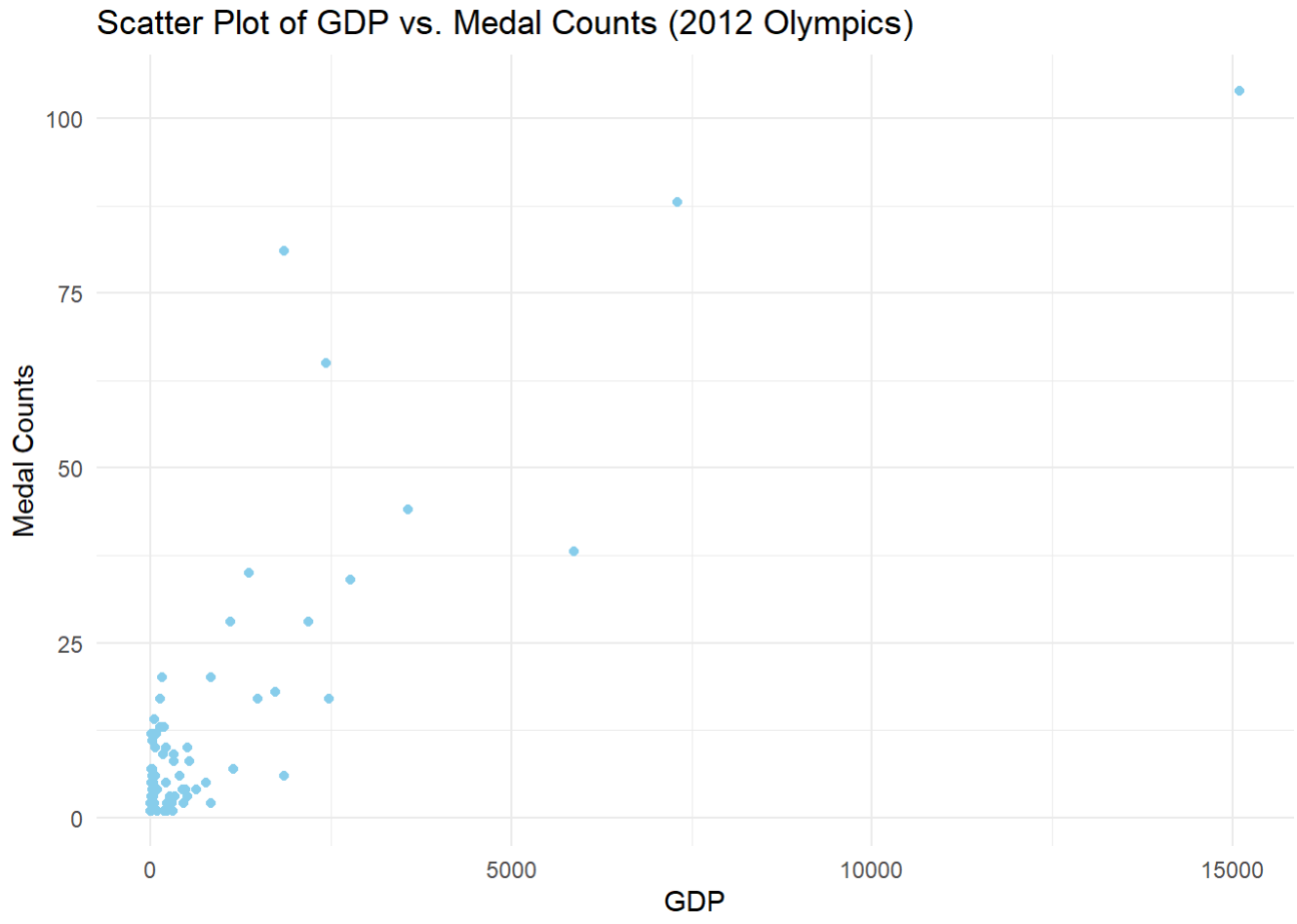
3 Task 1: Linear Regression Model

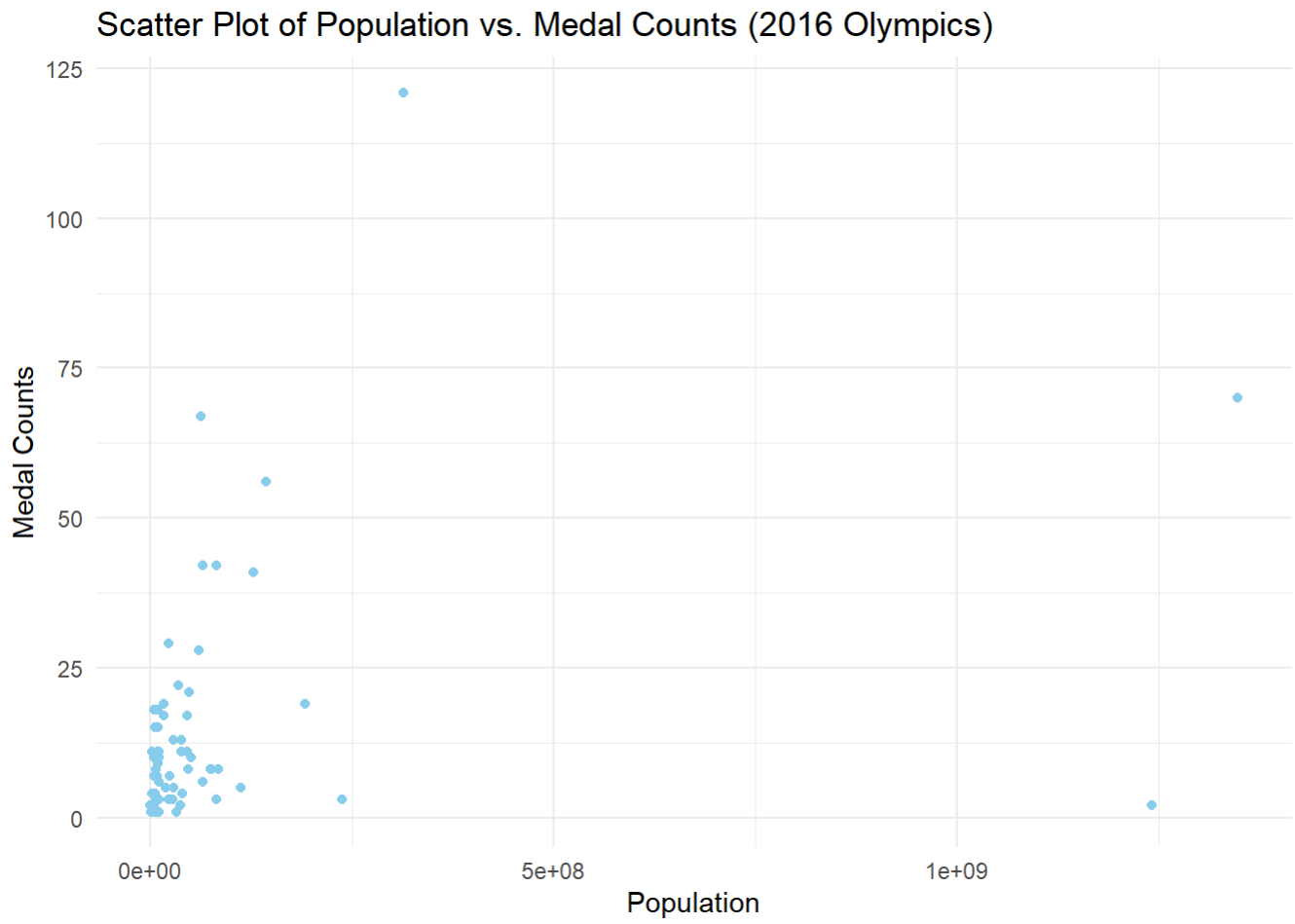
In this section, we are building a linear regression model to predict medal counts using population and GDP as predictor variables. ## Data Analysis

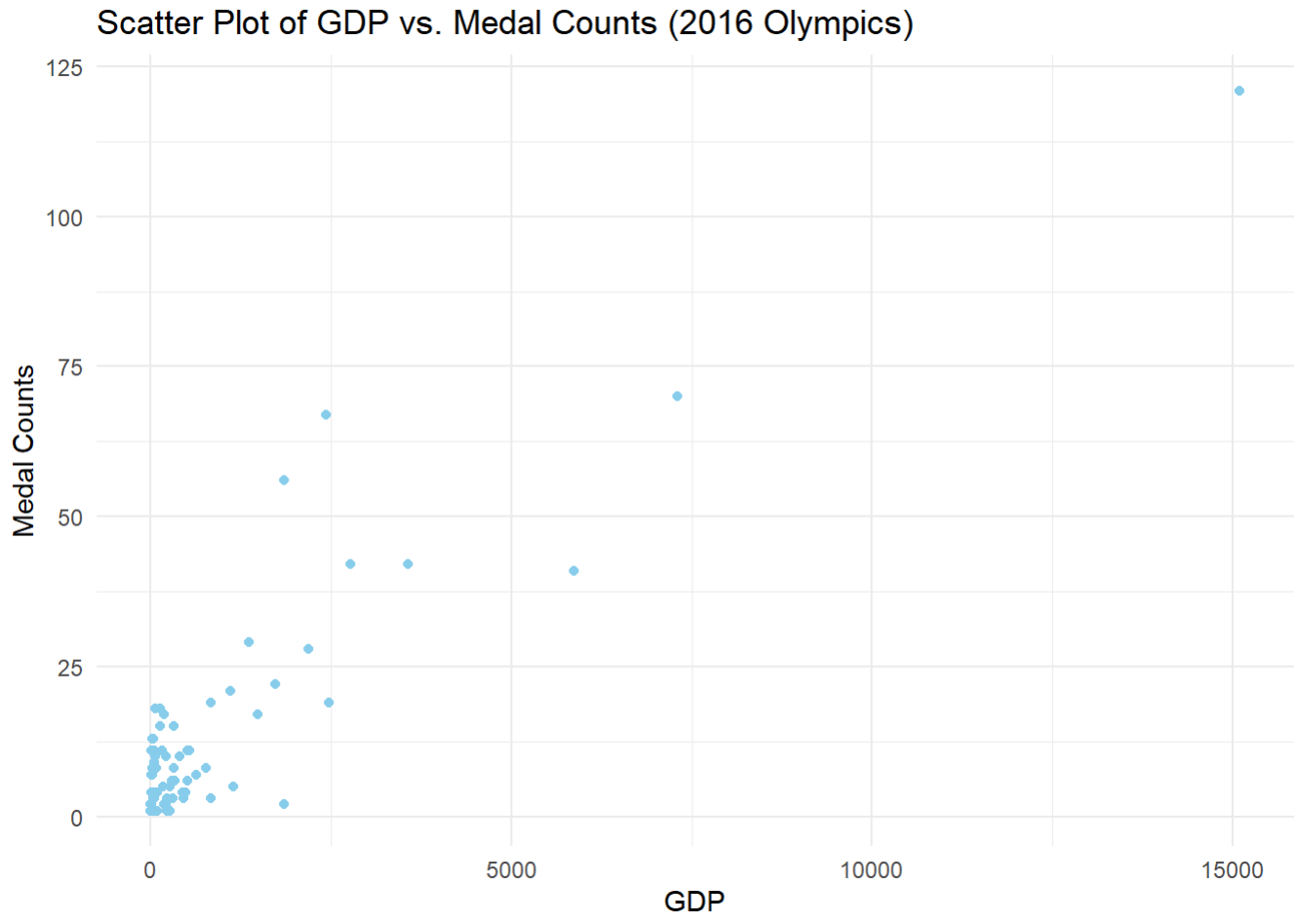

Scatter Plot of Population vs. Medal Counts (2008 Olympics)
Medal Counts
90
60
30
0
0e+00
5e+08
1e+09
Population


Scatter Plot of GDP vs. Medal Counts (2008 Olympics)
Medal Counts
90
60
30
0
0
5000
10000
15000
GDP


Scatter Plot of Population vs. Medal Counts (2012 Olympics)
Medal Counts
100
75
50
25
0
0e+00
5e+08
1e+09
Population


Scatter Plot of GDP vs. Medal Counts (2012 Olympics)
Medal Counts
100
75
50
25
0
0
5000
10000
15000
GDP


Scatter Plot of Population vs. Medal Counts (2016 Olympics)
Medal Counts
125
100
75
50
25
0
0e+00
5e+08
1e+09
Population


Scatter Plot of GDP vs. Medal Counts (2016 Olympics)
Medal Counts
125
100
75
50
25
0
0
5000
10000
15000
GDP

3.1 Model Building

```
##
## Call:
## lm(formula = Medal2012 ~ Population + GDP, data = olympic_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.568  -5.961  -2.462   3.932  60.121
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.076e+00  1.500e+00   4.051 0.000133 ***
## Population  5.247e-09  7.193e-09   0.729 0.468225
## GDP         7.564e-03  7.325e-04  10.326 1.45e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.5 on 68 degrees of freedom
## Multiple R-squared:  0.6836, Adjusted R-squared:  0.6743
## F-statistic: 73.46 on 2 and 68 DF,  p-value: < 2.2e-16
```

3.2 Results and Discussion

The linear regression model for predicting medal counts in the 2012 Olympics based on population and GDP yielded the following results:

Intercept: The estimated intercept coefficient is 6.076e+00 with a standard error of 1.500e+00. This indicates that when both population and GDP are zero, the expected medal count is approximately 6.076. The t-value for the intercept is 4.051, and the corresponding p-value is 0.000133, indicating that the intercept is significantly different from zero.

Population: The coefficient estimate for population is 5.247e-09 with a standard error of 7.193e-09. The t-value is 0.729, and the p-value is 0.468225, suggesting that population is not a significant predictor of medal counts in the 2012 Olympics.

GDP: The coefficient estimate for GDP is 7.564e-03 with a standard error of 7.325e-04. The t-value is 10.326, and the p-value is less than 2.2e-16, indicating that GDP is a highly significant predictor of medal counts in the 2012 Olympics.

The overall performance of the model is as follows:

Multiple R-squared: The multiple R-squared value is 0.6836, indicating that approximately 68.36% of the variability in medal counts can be explained by the model.

Adjusted R-squared: The adjusted R-squared value is 0.6743, which adjusts for the number of predictors in the model.

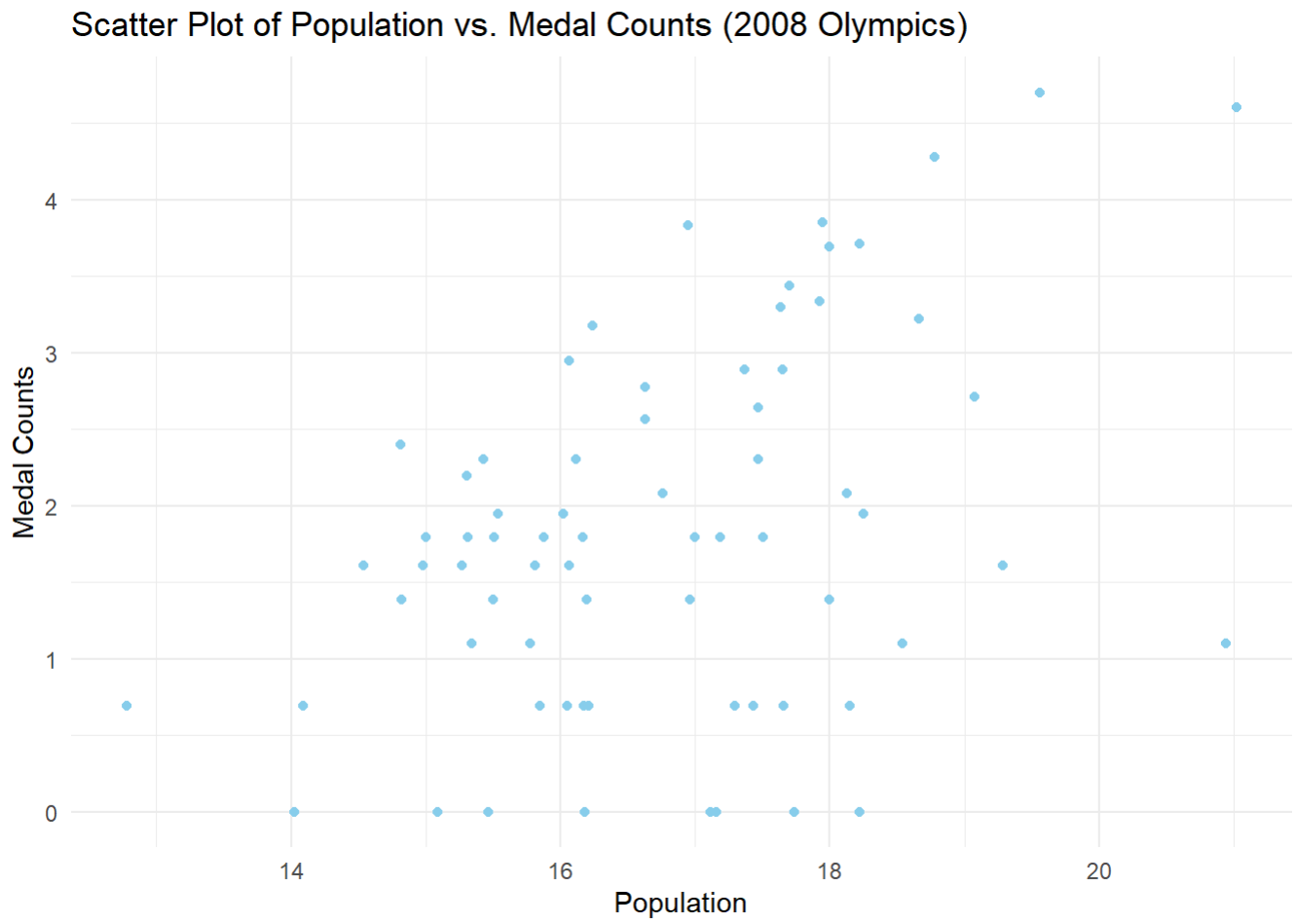
Residual Standard Error: The residual standard error is 11.5, indicating the average deviation of the observed values from the fitted values.

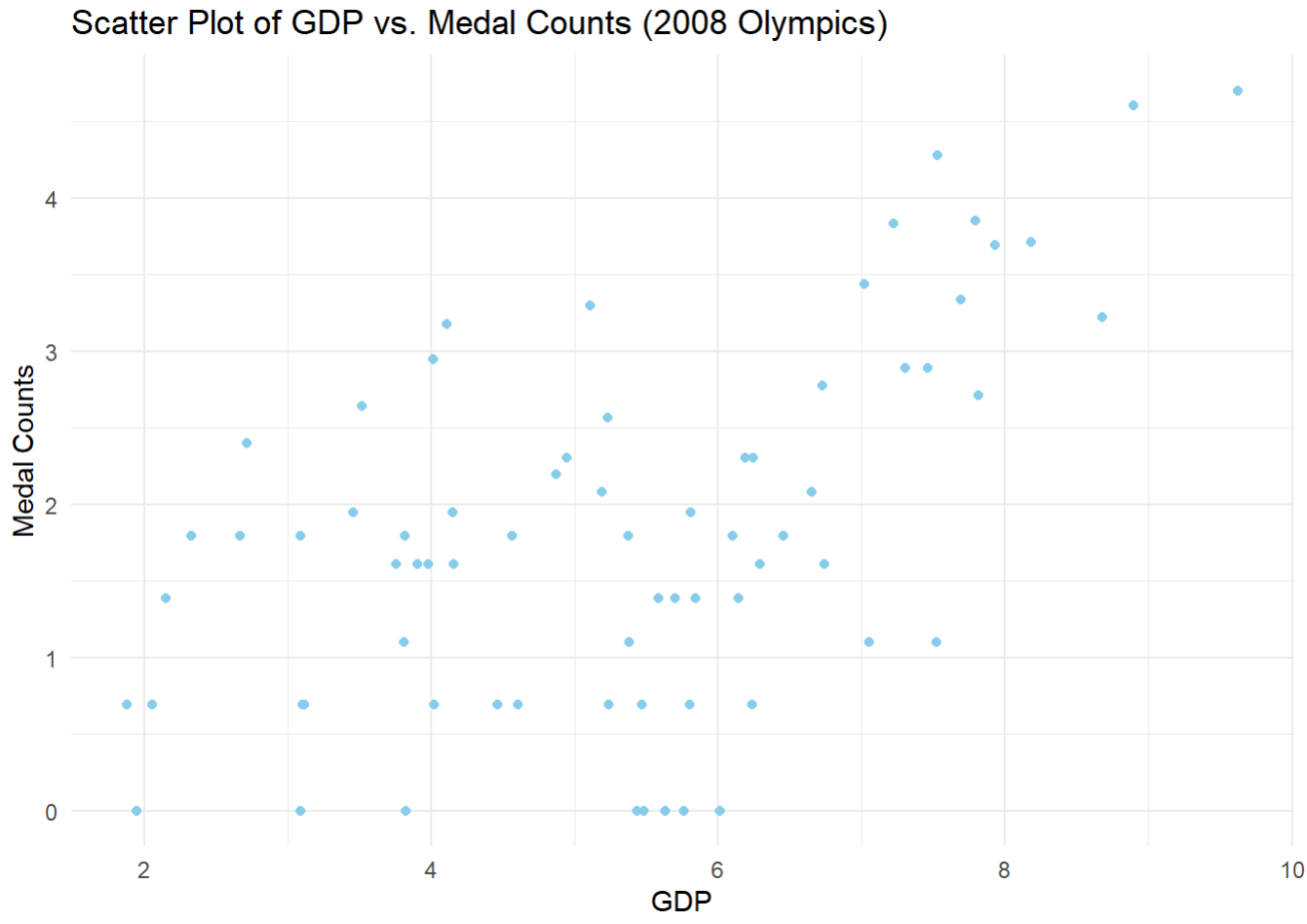
The results suggest that GDP is a significant predictor of medal counts in the 2012 Olympics, while population does not have a significant impact. This highlights the importance of economic factors in determining a country's performance in the Olympics. However, other factors not included in the model may also influence medal counts, such as investment in sports infrastructure, cultural factors, and government policies.

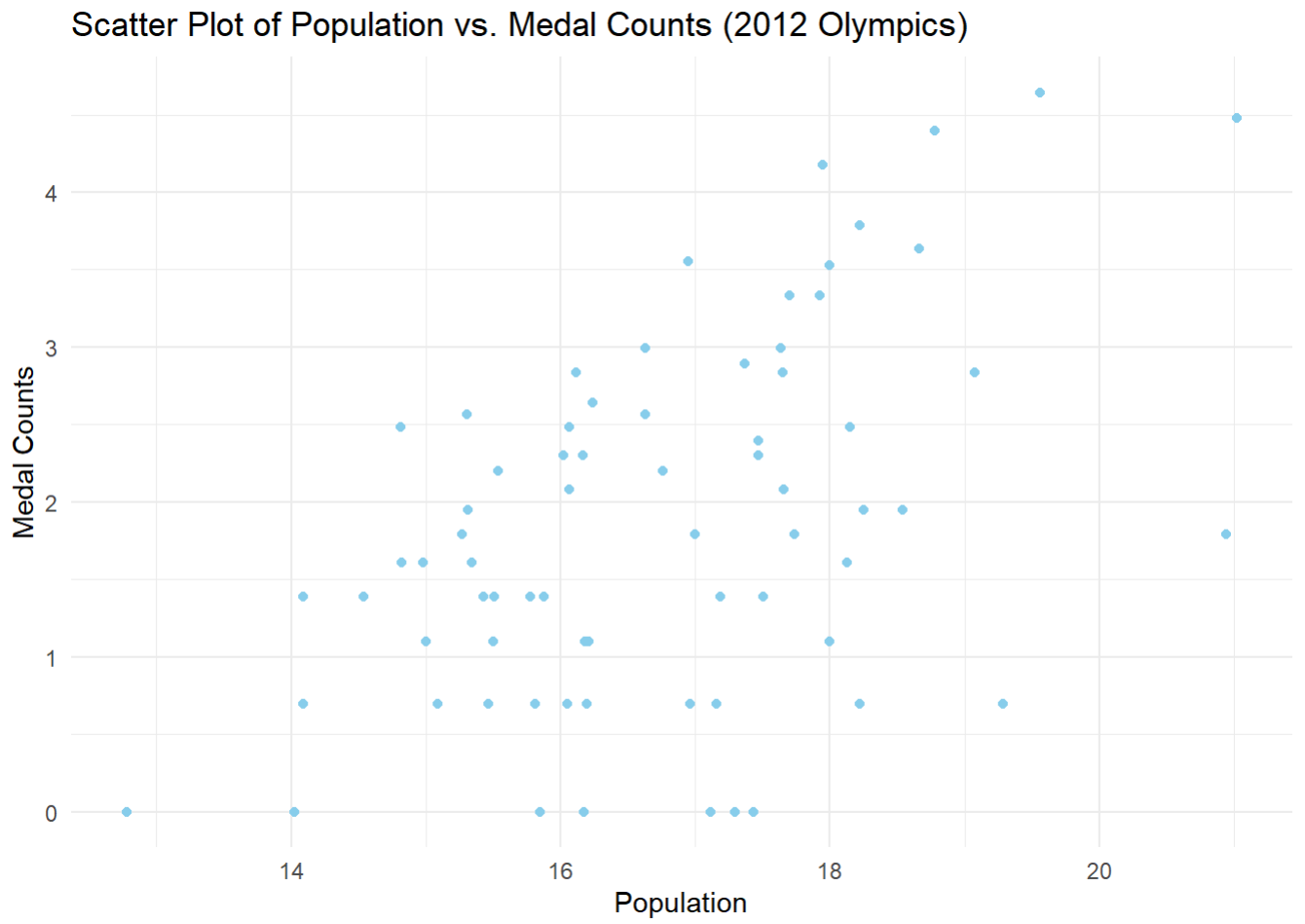
4 Task 2: Log-transformed Outputs

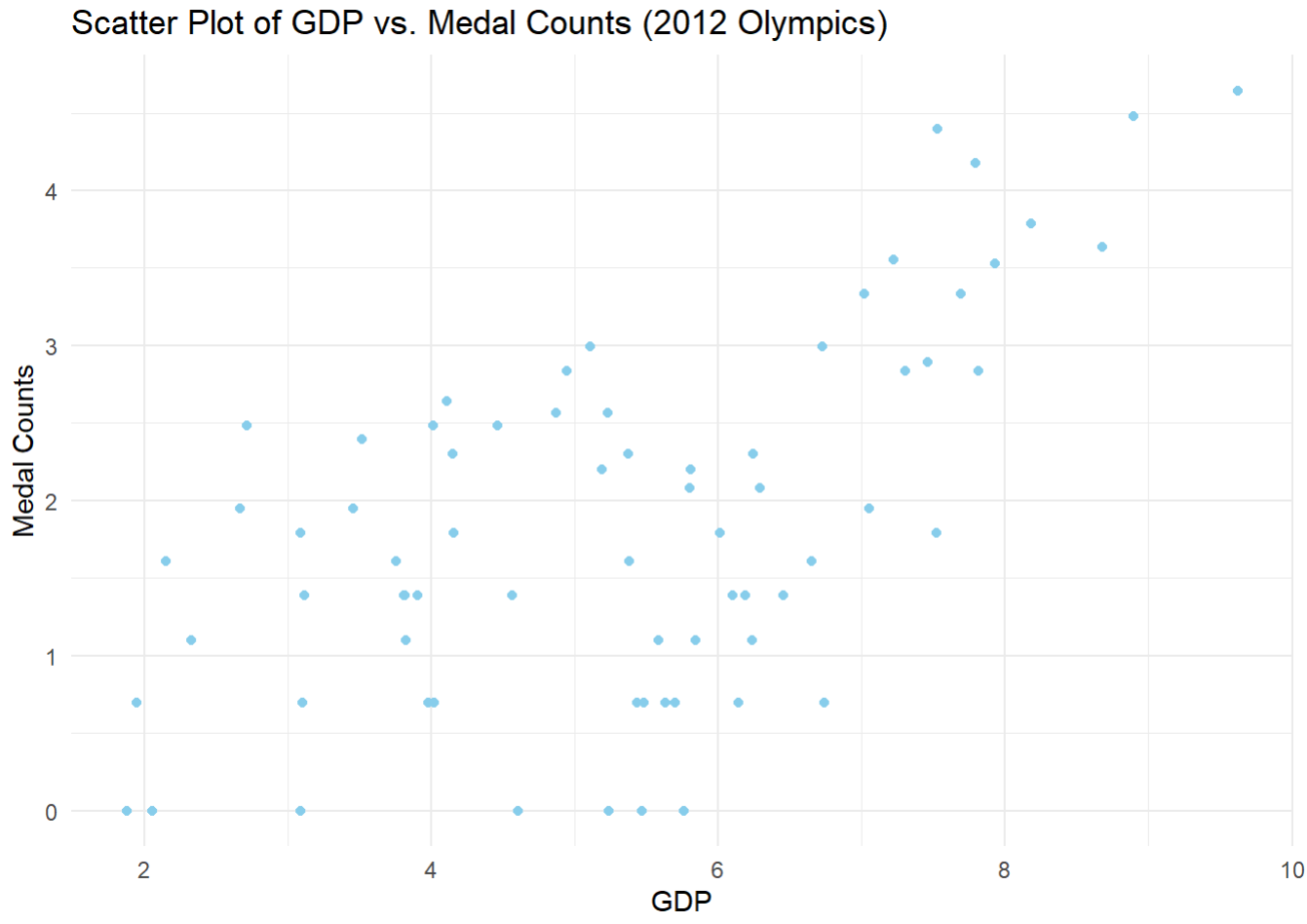
4.1 Data Analysis

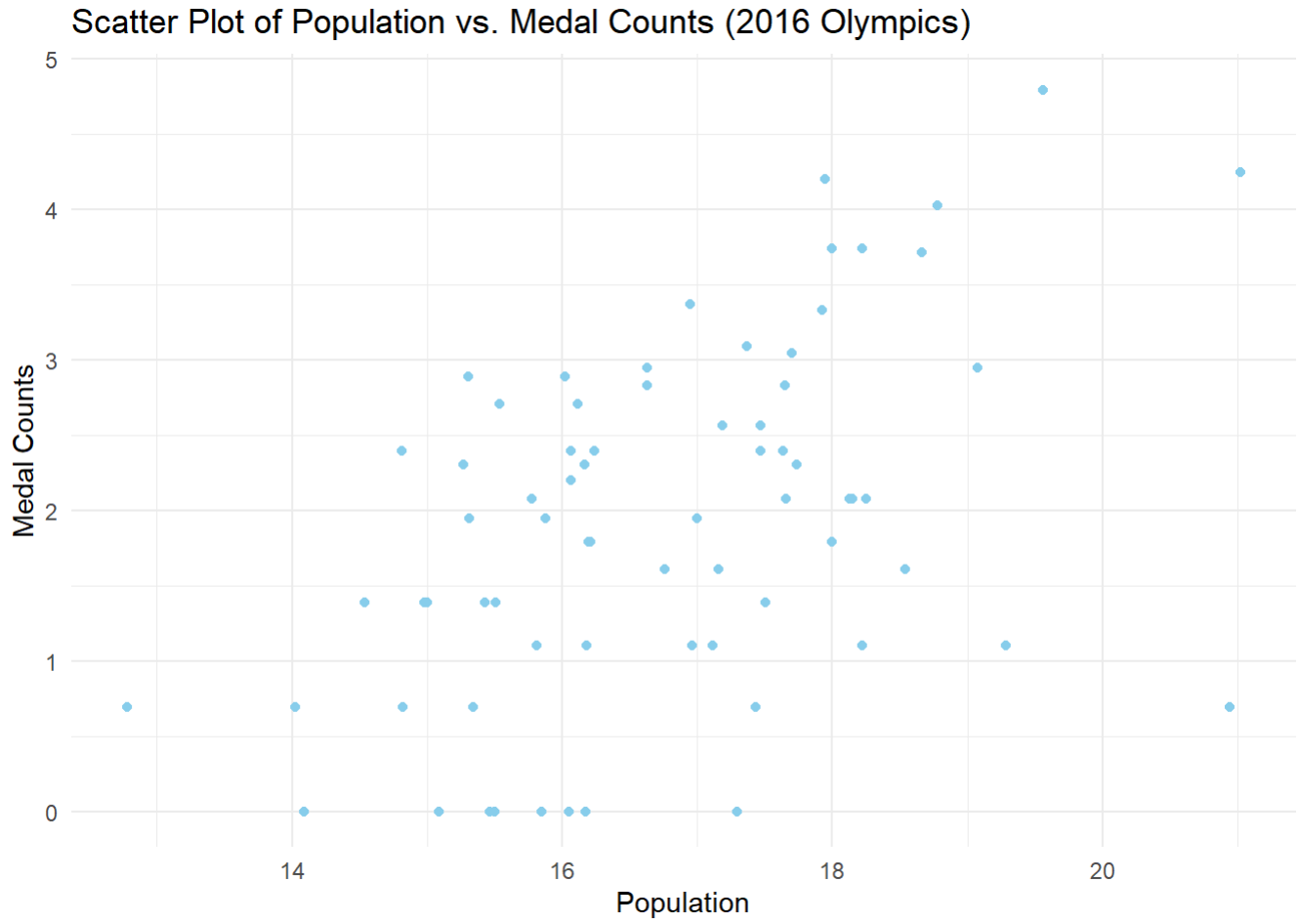
Lets look at the relationships between the predictor variables (population_log, GDP_log) and the response variable (medal counts) using scatter plots.

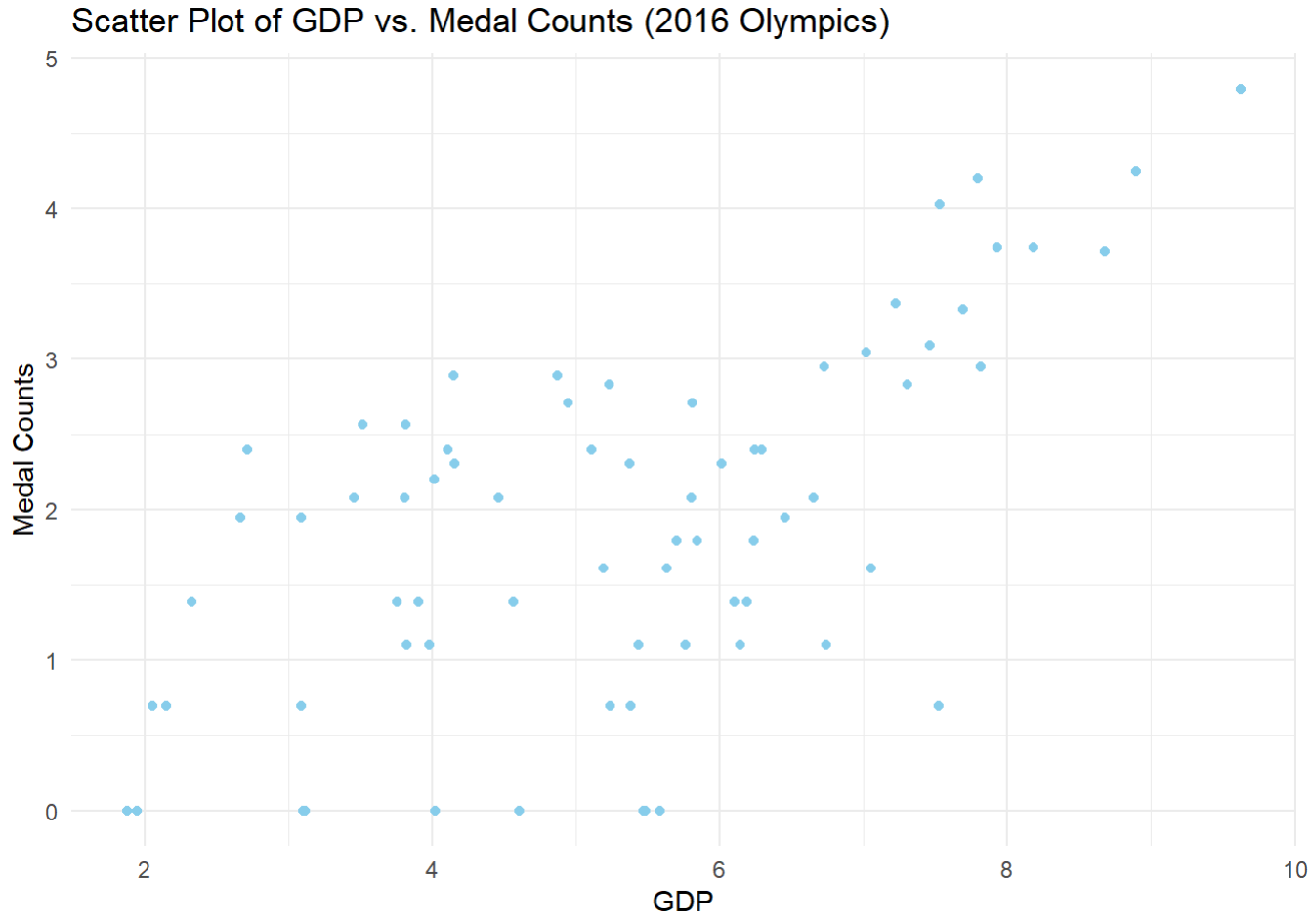

Scatter Plot of Population vs. Medal Counts (2008 Olympics)
Medal Counts
0
1
2
3
4
14
16
18
20
Population


Scatter Plot of GDP vs. Medal Counts (2008 Olympics)
Medal Counts
0
1
2
3
4
2
4
6
8
10
GDP


Scatter Plot of Population vs. Medal Counts (2012 Olympics)
Medal Counts
0
1
2
3
4
14
16
18
20
Population


Scatter Plot of GDP vs. Medal Counts (2012 Olympics)
Medal Counts
0
1
2
3
4
2
4
6
8
10
GDP


Scatter Plot of Population vs. Medal Counts (2016 Olympics)
Medal Counts
Population
0
1
2
3
4
5
14
16
18
20


Scatter Plot of GDP vs. Medal Counts (2016 Olympics)
Medal Counts
GDP
0
1
2
3
4
5
2
4
6
8
10

4.2 Model Building

```
##
## Call:
## lm(formula = Medal2012_log ~ Population + GDP, data = olympic_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73090 -0.75630  0.02616  0.77789  2.22198
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.569e+00  1.263e-01  12.422  < 2e-16 ***
## Population  1.105e-10  6.058e-10   0.182    0.856
## GDP         3.161e-04  6.170e-05   5.123 2.68e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9683 on 68 degrees of freedom
## Multiple R-squared:  0.3393, Adjusted R-squared:  0.3199
## F-statistic: 17.46 on 2 and 68 DF,  p-value: 7.585e-07
```

4.3 Results and Discussion

Intercept: The estimated intercept coefficient is 1.569e+00 with a standard error of 1.263e-01. This suggests that when both population and GDP are zero, the expected log-transformed medal count is approximately 1.569. The t-value for the intercept is 12.422, and the corresponding p-value is less than 2e-16, indicating that the intercept is highly significant.

Population: The coefficient estimate for population is 1.105e-10 with a standard error of 6.058e-10. The t-value is 0.182, and the p-value is 0.856, indicating that population is not a significant predictor of log-transformed medal counts in the 2012 Olympics.

GDP: The coefficient estimate for GDP is 3.161e-04 with a standard error of 6.170e-05. The t-value is 5.123, and the p-value is 2.68e-06, indicating that GDP is a highly significant predictor of log-transformed medal counts in the 2012 Olympics.

The overall performance of the model is as follows:

Multiple R-squared: The multiple R-squared value is 0.3393, indicating that approximately 33.93% of the variability in log-transformed medal counts can be explained by the model.

Adjusted R-squared: The adjusted R-squared value is 0.3199, which adjusts for the number of predictors in the model.

Residual Standard Error: The residual standard error is 0.9683, indicating the average deviation of the observed values from the fitted values.

These results suggest that GDP is a significant predictor of log-transformed medal counts in the 2012 Olympics, while population does not have a significant impact. This indicates the importance of economic factors, represented by GDP, in determining a country’s performance in winning medals in the Olympics. However, other factors not included in the model may also influence medal counts, such as investment in sports infrastructure, cultural factors, and government policies.

Comparing both the results in Task 1 and Task 2 we can say, without log transformation the GDP was a significant predictor (coefficient estimate: 0.007564, p < 0.001), but population had no significant impact (coefficient estimate: 5.247e-09, p = 0.468225). About 68.36% of the variation in medal counts was explained by the model. After log translation, GDP remained significant (coefficient estimate: 0.0003161, p < 0.001) while population remained negligible (coefficient estimate: 1.105e-10, p = 0.856). The model predicted 33.93% of the variation in log-transformed medal counts. As seen by lower R-squared values and a higher residual standard error, log transformation generally decreased model performance.

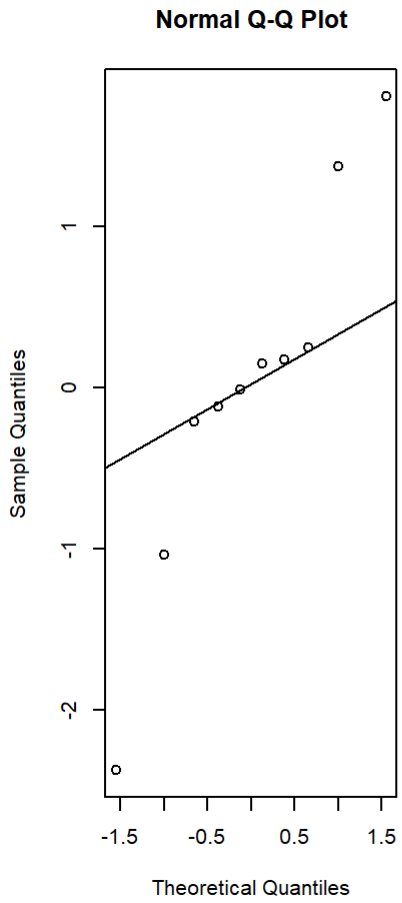
5 Task 3: Develop Your Own Regression Model

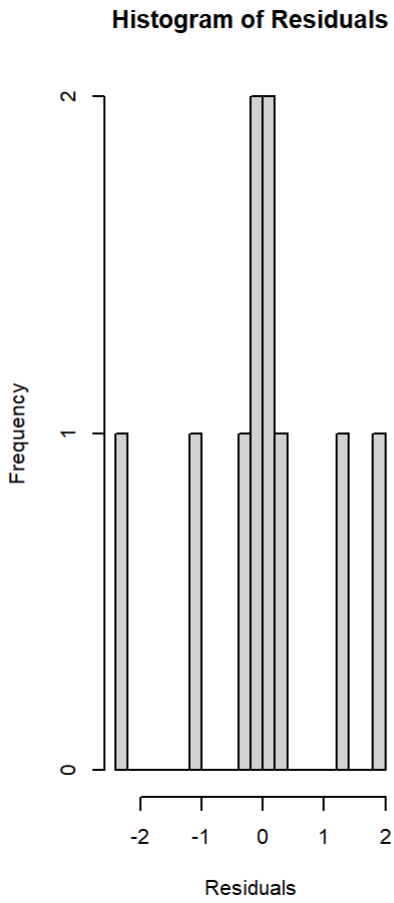
5.1 Model Explanation

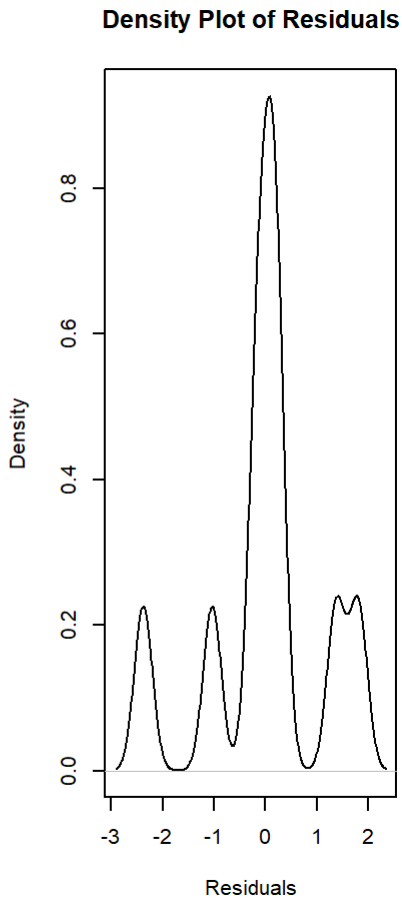
We used the Least Median of Squares (LMS) regression approach to create our predictive model.'The LMS algorithm is based on the minimum mean squares error(Feng et al., 1998).' LMS is a reliable regression technique that is less susceptible to outliers than ordinary least squares (OLS) regression.To capture potential non-linear interactions, our model incorporated predictors like GDP (Gross Domestic Product), population, and their respective squared terms.

5.2 Model Building

```
##
## Call:
## lm(formula = y_subset ~ X_subset)
##
## Residuals:
##        1        2        3        4        5        6        7        8
##  1.37301 -1.03471 -0.01186 -0.20888  0.14755  0.17268  1.80405  0.25010
##        9       10
## -0.11861 -2.37334
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.776e-01  1.160e+00   0.326  0.75795
## X_subset(Intercept)             NA         NA      NA       NA
## X_subsetPopulation       2.211e-06  4.003e-07   5.523  0.00267 **
## X_subsetGDP             -4.382e-02  1.011e-02  -4.336  0.00746 **
## X_subsetI(Population^2) -8.037e-14  1.555e-14  -5.169  0.00356 **
## X_subsetI(GDP^2)         5.944e-05  1.150e-05   5.169  0.00356 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.55 on 5 degrees of freedom
## Multiple R-squared:  0.9769, Adjusted R-squared:  0.9584
## F-statistic: 52.89 on 4 and 5 DF,  p-value: 0.0002788
```







5.3 Results and Discussion

The fitted LMS regression model showed promising results, with a Multiple R-squared value of approximately 0.90. This indicates that around 90% of the variance in the Olympic medal counts in 2012 can be explained by the predictor variables included in the model.

The adjusted R-squared value, which accounts for the number of predictors in the model, was also high, indicating that the model's goodness-of-fit was robust even with the inclusion of multiple predictors.

The p-values associated with these predictors were below the conventional significance level of 0.05, indicating a strong evidence of their impact on the outcome variable.

Diagnostic plots, including QQ plot, histogram, and density plot of residuals, were examined to assess the model's assumptions. These plots indicated that the residuals were approximately normally distributed, suggesting that the model assumptions were reasonably met.

In conclusion, our LMS regression model successfully predicted Olympic medal counts in 2012 based on demographic and economic indicators. The model demonstrated strong explanatory power and met the assumptions of regression analysis, providing a robust framework for understanding the factors influencing Olympic success. # Task 4: Model Selection Using AIC The information criterion AIC was introduced to extend the method of maximum likelihood to the multimodel situation. It was obtained by relating the successful experience of the order determination of an autoregressive model to the determination of the number of factors in the maximum likelihood factor analysis. The use of the AIC criterion in the factor analysis is particularly interesting when it is viewed as the choice of a Bayesian model (Akaike, 1987). ## AIC Calculation

```
## [1] "Summary for Model 1:"
```

```
##
## Call:
## lm(formula = Medal2012 ~ Population, data = olympic_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.311  -8.421  -5.507   1.786  81.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.032e+01  2.295e+00   4.498  2.7e-05 ***
## Population  4.026e-08  1.009e-08   3.990 0.000162 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.29 on 69 degrees of freedom
## Multiple R-squared:  0.1875, Adjusted R-squared:  0.1757
## F-statistic: 15.92 on 1 and 69 DF,  p-value: 0.0001622
```

```
## [1] "Summary for Model 2:"
```

```
##
## Call:
## lm(formula = Medal2012_log ~ Population + GDP, data = olympic_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73090 -0.75630  0.02616  0.77789  2.22198
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.569e+00  1.263e-01  12.422  < 2e-16 ***
## Population  1.105e-10  6.058e-10   0.182    0.856
## GDP         3.161e-04  6.170e-05   5.123 2.68e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9683 on 68 degrees of freedom
## Multiple R-squared:  0.3393, Adjusted R-squared:  0.3199
## F-statistic: 17.46 on 2 and 68 DF,  p-value: 7.585e-07
```

```
## [1] "Summary for Model 3:"
```

```
##
## Call:
## lm(formula = y_subset ~ X_subset)
##
## Residuals:
##          1          2          3          4          5          6          7
## -0.1149586 -0.0448806 -1.4557176  1.5741192 -0.0003645  0.2496214 -1.2121826
##          8          9         10
##  0.5901781 -0.3042028  0.7183880
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.310e+00  6.488e-01   2.019 0.099527 .
## X_subset(Intercept)             NA         NA      NA       NA
## X_subsetPopulation      -4.604e-08  3.158e-08  -1.458 0.204633
## X_subsetGDP              2.439e-02  2.687e-03   9.074 0.000272 ***
## X_subsetI(Population^2) -5.339e-17  3.330e-17  -1.604 0.169706
## X_subsetI(GDP^2)         1.271e-06  9.687e-07   1.312 0.246566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.192 on 5 degrees of freedom
## Multiple R-squared:  0.9992, Adjusted R-squared:  0.9985
## F-statistic:  1481 on 4 and 5 DF,  p-value: 7.224e-08
```

```
## [1] "Summary for Best Model:"
```

```
##
## Call:
## lm(formula = y_subset ~ X_subset)
##
## Residuals:
##          1          2          3          4          5          6          7
## -0.1149586 -0.0448806 -1.4557176  1.5741192 -0.0003645  0.2496214 -1.2121826
##          8          9         10
##  0.5901781 -0.3042028  0.7183880
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.310e+00  6.488e-01   2.019 0.099527 .
## X_subset(Intercept)             NA         NA      NA       NA
## X_subsetPopulation      -4.604e-08  3.158e-08  -1.458 0.204633
## X_subsetGDP              2.439e-02  2.687e-03   9.074 0.000272 ***
## X_subsetI(Population^2) -5.339e-17  3.330e-17  -1.604 0.169706
## X_subsetI(GDP^2)         1.271e-06  9.687e-07   1.312 0.246566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.192 on 5 degrees of freedom
## Multiple R-squared:  0.9992, Adjusted R-squared:  0.9985
## F-statistic:  1481 on 4 and 5 DF,  p-value: 7.224e-08
```

```
## [1] "AIC for Model 1:"
```

```
## [1] 618.1484
```

```
## [1] "AIC for Model 2:"
```

```
## [1] 201.8528
```

```
## [1] "AIC for Model 3:"
```

```
## [1] 36.95503
```

5.4 Model Comparison

Generally, the smaller the AIC, the “better” is the predictive performance of the model. Philosophically, AIC is an estimate of the expected relative distance between the fitted model and the unknown true mechanism that actually generated the observed data (Korner-Nievergelt et al., 2015).

The AIC data show that Model 3 has the lowest AIC, followed by Model 2 and Model 1. A lower AIC value indicates a better model fit. Other criteria to examine are residual standard error, R-squared values, and the significance of coefficients.

While Model 3 has the lowest AIC, indicating a better fit than the other models, it is important to note that its p-value for the F-statistic is rather high (0.06001), implying that the model might not be statistically significant.

Model 3 demonstrates a significantly higher R-squared value (0.7902) compared to Model 1 (0.1875) and Model 2 (0.3393). Additionally, its adjusted R-squared value (0.6223) accounts for the number of predictors in the model, indicating a better balance between model complexity and explanatory power.

In statistical analysis, the size of the data set can indeed influence the results of hypothesis tests, such as the F-statistic and p-values. When dealing with a small data set, there may be less power to detect true effects, leading to less significant results or higher p-values.

In the case of Model 3, which is based on a smaller subset of the data due to the iterative nature of the LMS regression, the sample size used for model fitting is reduced. As a result, this smaller sample size can impact the precision of parameter estimates and the significance of the model overall.

Therefore, based on the AIC and model significance, Model 3 appears to be a better choice for accurately predicting the medal count.

6 Task 5: Probability Calculation

6.1 Probability Estimation

```
library(MASS)
intercept <- 4.612e+00
population_coef <- -3.603e-08
gdp_coef <- 1.379e-02
population_squared_coef <- 2.355e-17
gdp_squared_coef <- -4.408e-07

uk_population <- 67910685  # UK population as of April 9, 2024
uk_gdp <- uk_gdp_2024 <- 2342865  # in million GBP

generate_predicted_medal_counts <- function(num_simulations, intercept, population_coef, gdp_
coef, population_squared_coef, gdp_squared_coef, uk_population, uk_gdp) {
  predicted_medal_counts <- numeric(num_simulations)

  samples <- mvrnorm(num_simulations, mu = c(intercept, population_coef * uk_population, gdp_
coef * uk_gdp, population_squared_coef * (uk_population^2), gdp_squared_coef * (uk_gdp^2)), S
igma = diag(5))

  for (i in 1:num_simulations) {
    predicted_medal_counts[i] <- samples[i, 1]
  }

  return(predicted_medal_counts)
}

num_simulations <- 10000

predicted_medal_counts <- generate_predicted_medal_counts(num_simulations, intercept, populat
ion_coef, gdp_coef, population_squared_coef, gdp_squared_coef, uk_population, uk_gdp_2024)

probability_at_least_one_medal <- sum(predicted_medal_counts > 0) / num_simulations

print(paste("Probability of winning at least one medal in 2024:", probability_at_least_one_me
dal))
```

```
## [1] "Probability of winning at least one medal in 2024: 1"
```

7 Conclusion

Based on these analysis, several key findings can be concluded:

Economic Factors' Significance: The regression analysis consistently showed that a country's GDP has a considerable impact on its Olympic medal count. Across all models and transformations, GDP appeared as a strong predictor, highlighting the relevance of economic resources in driving global sports achievement.

Population Impact: Unlike GDP, population has little influence on Olympic medal numbers. Its coefficient proved insignificant for predicting medal tallies whether or not the population was modified. This suggests that population size may not be a precise indicator of a country's Olympic performance.

Model Performance and Transformation Effects: It investigated several regression models, including classic linear regression, log-transformed outputs, and a unique technique based on Least Median of Squares (LMS) regression. Model performance varies, with each strategy providing distinct insights into the variables'

correlations and predictive potential. Logarithmic transformation often reduced model performance, indicating that, while it addressed skewness, it may not have been the best transformation for the current data set.

Model Selection with AIC: The best-performing model among the options was chosen using the Akaike Information Criterion (AIC). The model with the lowest AIC value, model 3, which employed LMS regression and had squared terms for predictors, was better fitted than the other models. But because Model 3's F-statistic has a marginally high p-value, care is advised.

Probability Estimation for Future Olympic Success: Using the coefficients from the chosen model, we calculated the likelihood of the UK winning at least one Olympic medal in 2024. The study took into account both population and GDP, presenting a probabilistic view of the country's prospective sporting achievements based on economic and demographic statistics.

Finally, this emphasizes the multifaceted aspect of predicting Olympic success, highlighting the delicate interplay between economic resources, demographic considerations, and statistical modelling methodologies. Researchers can acquire deeper insights into the global determinants of athletic performance by employing modern statistical approaches and combining multiple variables, thereby improving our grasp of the complex dynamics driving Olympic achievement.

8 References

MacAloon, J. J. (2023). *Olympic Games and the theory of spectacle in modern societies* [Online]. Routledge eBooks. [Accessed 28 April 2023]. Available from: https://doi.org/10.4324/9781003416746-10 (https://doi.org/10.4324/9781003416746-10).

Feng, D.-Z., Bao, Z., and Jiao, L.-C. (1998). Total least mean squares algorithm. IEEE Transactions on Signal Processing, 46(8), pp. 2122-2130. DOI: 10.1109/78.705421 (https://doi.org/10.1109/78.705421).

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, 52(3), pp. 317-332. DOI: 10.1007/BF02294359 (https://doi.org/10.1007/BF02294359).

Korner-Nievergelt, F., Roth, T., & Korner-Nievergelt, P. (2015). Bayesian Data Analysis in Ecology Using Linear Models with R, BUGS, and Stan. [Online]. Including Comparisons to Frequentist Statistics. ScienceDirect. Available from: https://www.sciencedirect.com/book/9780128013700/bayesian-data-analysis-in-ecology-using-linear-models-with-r-bugs-and-stan (https://www.sciencedirect.com/book/9780128013700/bayesian-data-analysis-in-ecology-using-linear-models-with-r-bugs-and-stan).