Statistical Learning Assessed Practical 1

Julian Bara-Mason | 201674483

2023-03-12

```
knitr::opts_chunk$set(echo=TRUE)
```

```
## [1] NA
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

Task 1 - Model 1

Use Population and GDP as inputs and medal count in 2012 Olympics as outputs for the linear regression model. Assess the model predictions using the same inputs against the medal count in 2016.

To predict the 2016 Olympics medal count using the 2012 data, the model used is as follows:

$$Medals_{won} = \beta_0 + \beta_1(GDP) + \beta_2(Population) + \epsilon, \epsilon \sim N(0, \sigma^2)$$

, *where* $Medals_{won}$ is the output, and $\beta_1(GDP)$ & $\beta_2(Population)$ are the inputs

Using a generalized linear regression model, the resulting model, Model 1, is:

$$Medals_{won} = 6.076 + 7.987 * 10^{-3}(GDP) + 5.642 * 10^{-9}(Population) + \epsilon$$

Interpreting coefficient values of model

In this model, the input, $\beta 1$ (GDP), appears to be a more significant predictor on the medals won than $\beta 2$ (Population).

The higher the absolute value of the t-value, the stronger the evidence against the null hypothesis of zero coefficient. $\beta 1$ (GDP) has a high t value of 10.33, indicating the coefficient for GDP is significantly different from zero. $\beta 2$ (Population) has a t value of 0.73, which is close to zero, indicating the coefficient for Population is not significantly different from zero.

The p-value for $\beta 1$ (GDP) is very small, $1.45e{-}15$, indicating that the GDP coefficient is highly significant in predicting the medals won. On the other hand, the p-value for $\beta 2$ (Population) has a relatively high p-value (> 0.05), indicating the Population coefficient is not significant in predicting medals won.

Confidence Intervals

Table 1. Estimates and their confidence intervals

X.	Confidence_Interval	Estimate_in_Interval
Intercept	3.14, 9.02	Yes
GDP	6.13e-3, 8.90e-3	Yes
Population	-8.85e-9, 1.93e-8	Yes

Table 1 above shows that the estimated values fall within their respective confidence interval, indicating a consistent model.

2016 Olympics Prediction with model

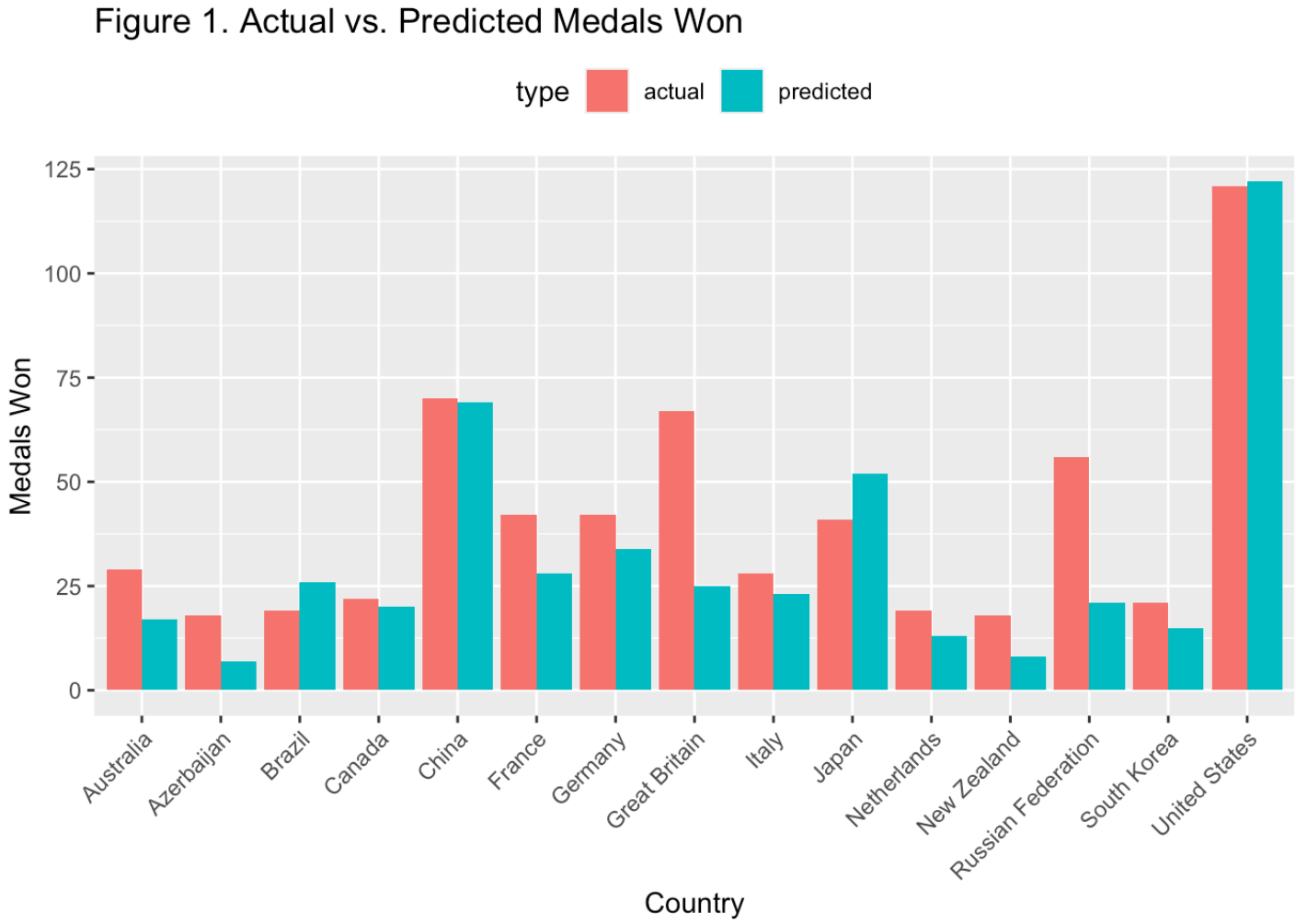


Figure 1 above is a bar chart of the top 15 Actual Medals Won vs Predicted Medals Won in the 2016 Olympics. The predicted medals won appear to be reasonably close to the actual medals won.

```
## The Root Mean Squared Error (RMSE) of Model 1 is  9.17
```

The model appears to be performing reasonably well with a RMSE that is smaller than the range of actual values (120) but relatively close to the mean (13.4).

Task 2 - Model 2

Repeat the task 1 for log-transformed inputs. Which model performs better. Justify your reasoning. Discuss potential benefits and reasons for using the transformation.

The resulting log-transformed model is:

$$Medals_{won} = -49.363 + 5.545(GDP) + 1.983(Population) + \epsilon$$

As with the earlier model, $\beta 1$ (GDP), appears to be a more significant predictor on the medals won than $\beta 2$ (Population).

Table 2. Log estimates and their confidence intervals

X.	Confidence_Interval	Estimate_in_Interval
Intercept	-97.61, -1.12	Yes
log(GDP)	2.52, 8.57	Yes
log(Population)	-1.53, 5.49	Yes

The estimated values fall within their respective confidence interval, indicating a consistent model.

Using the log-transformed model to predict the 2016 Olympics and comparing with previous model.

Table 3. Log-Transformed Model vs Original Model

Model	Residual_Deviance	AIC	RMSE
Log-Transformed	17,310.00	599.70	5.98e+17
Original	8,986.20	553.20	9.17

Table 3 shows the Residual Deviance, AIC, and RMSE scores of both models.

The residual deviance measures how well the model fits the data. The residual deviance of the original model is significantly lower than the residual deviance of the log-transformed model, indicating that it is a better fit.

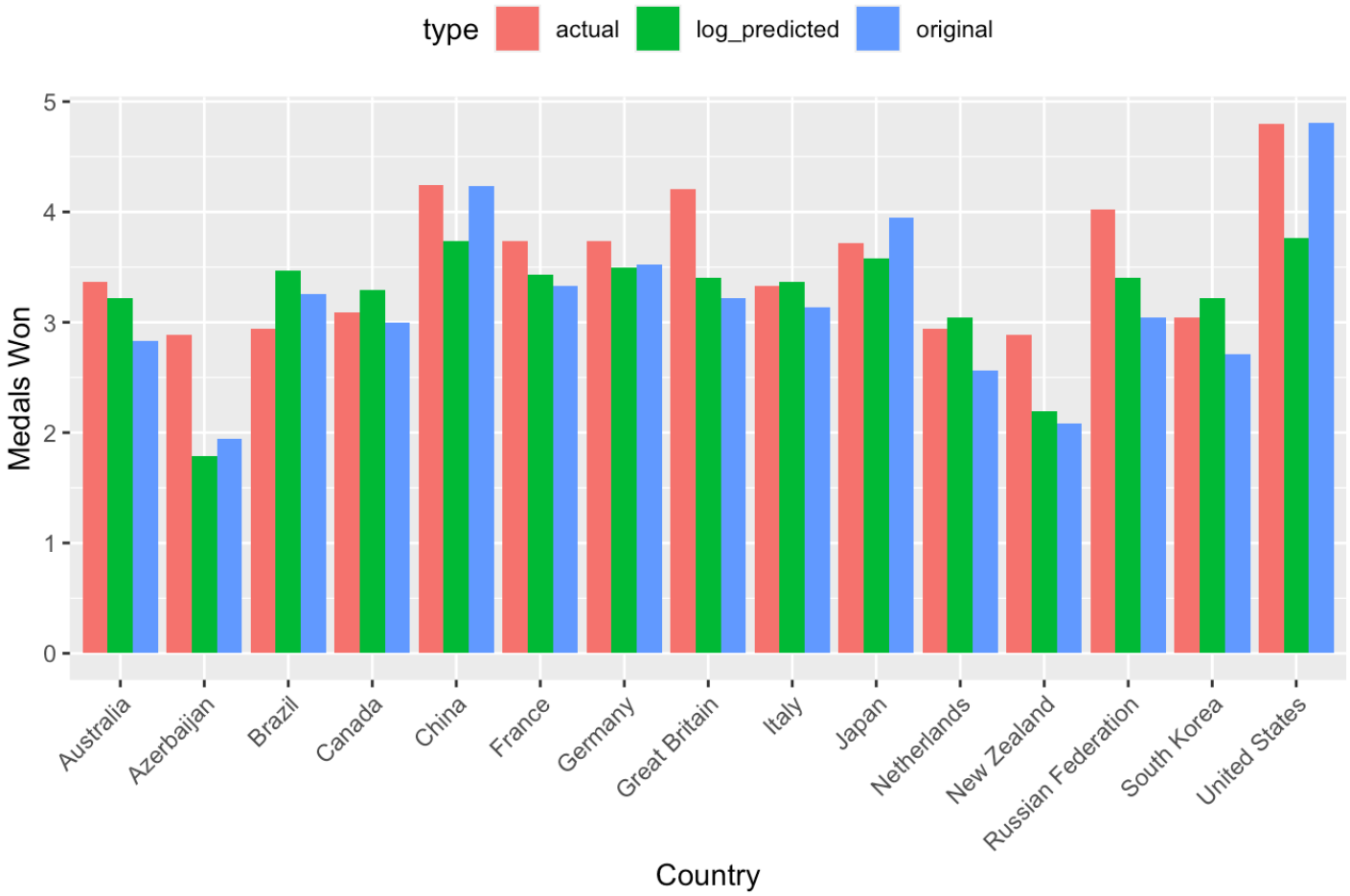
RMSE measures the average squared difference between the predicted values and the actual values - the accuracy of the predictions. The lower the RMSE, the better the model.

AIC, Akaike Information Criterion, measures the relative quality of the models based on their goodness of fit and complexity. The lower the AIC score, the better. The original model has a lower AIC value as per Table 2.

The original model, Model 1, seems to perform better as it has a lower residual deviance, AIC, and RMSE.

Figure 2 below is the top 15 medals won and the original model and log-transformed model predictions.

Figure 2. Actual vs. Predicted vs Log_Transformed Medals Won



Generally, log-transformations are useful for improving the performance and ease of interpreting linear regression models.

Log-transformations help improve performance by improving the model's accuracy by reducing the data's variability in events of highly skewed data. Also, non-linear relationships between the input and output variables can be transformed into linear relationships by log-transformation.

Task 3 - Model 3

Use K-means algorithm (kmeans function in R) to cluster the log-transformed inputs. Repeat the task 1 by training a linear regression model for each cluster (partitioning of the inputs) using the corresponding outputs. Combine the predictions suitably. Validate the optimal number of clusters. Justify your choice. Discuss potential benefits of the approach.

Figure 3. Elbow Plot

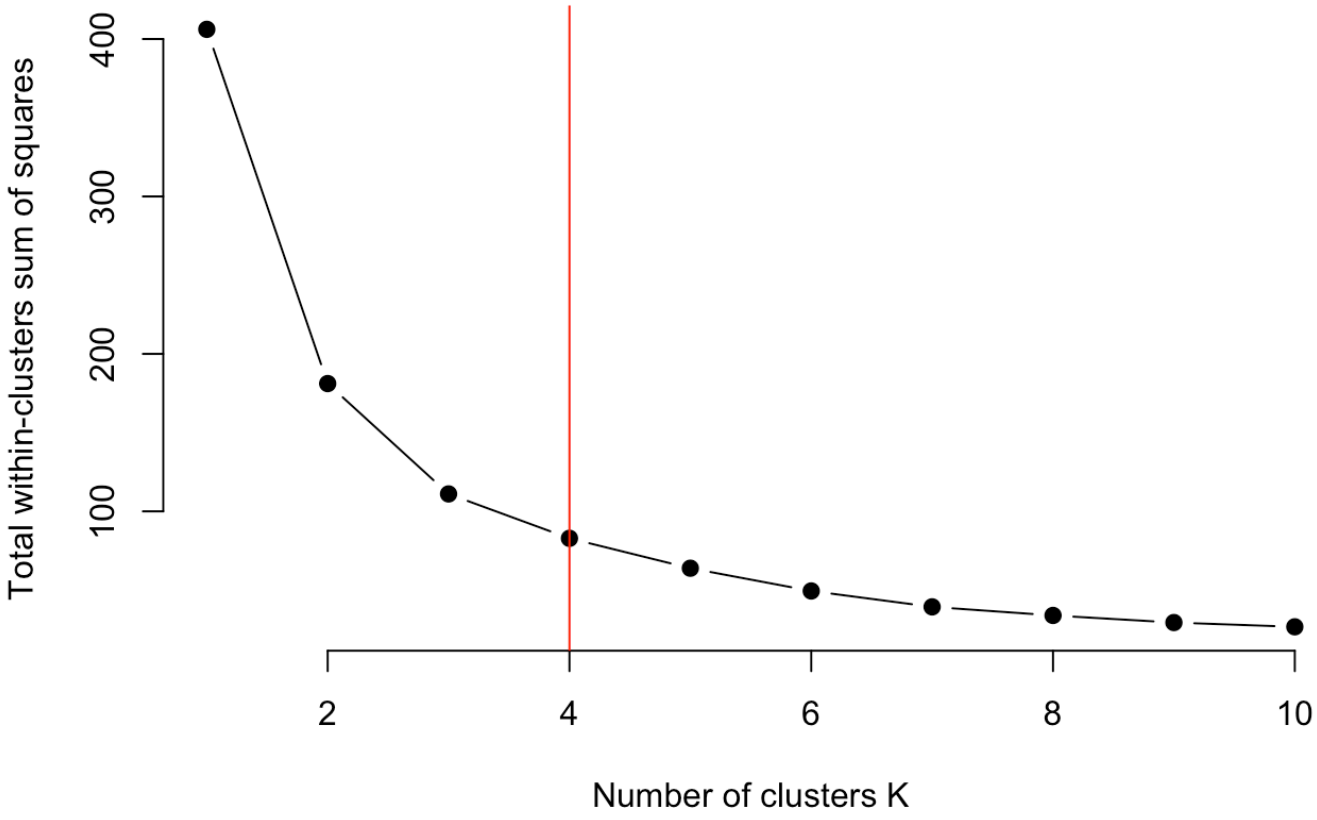


Figure 3 above is a plot of the Within-Cluster Sum of Squares (WSS) for different number of clusters to measure the sum of squared distance between the data points. The plot is a visual aid in identifying the optimal number of clusters that result in the least amount of variation within each cluster, while at the same time, minimizing the number of clusters used. It helps against overfitting or underfitting data.

From the figure, 4 appears to be the optimal number of clusters as there appears to be a relatively constant rate of decline of WSS from 4 clusters.

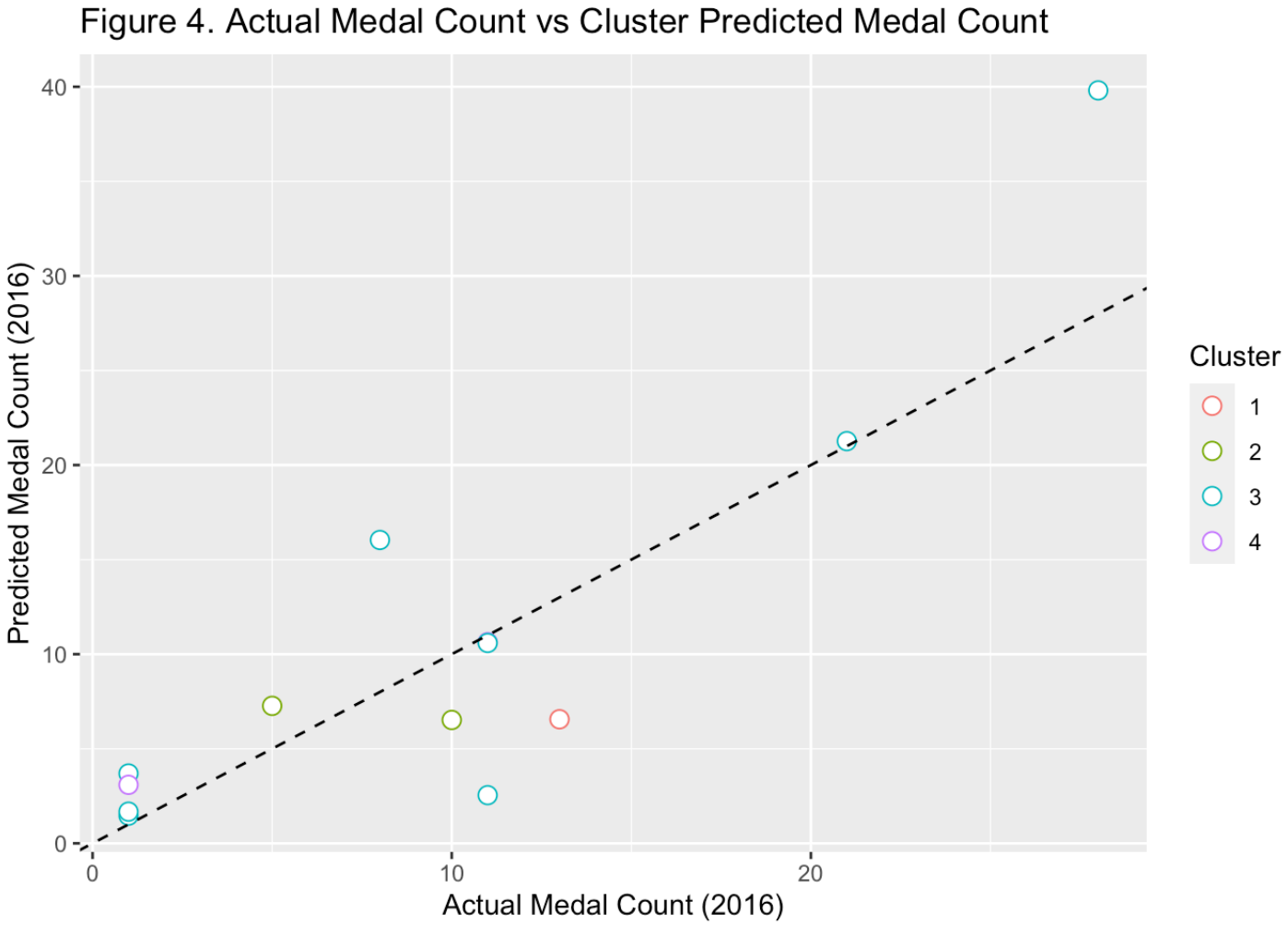


Figure 4 above is a scatterplot of the Actual Medal Count vs the Cluster Predictions.

The approach used here helps identify the different countries with similar output-input relationships and different patterns, thus helping to improve the accuracy of the predictions.

Task 4

Derive and compute the probability that a country wins at least one medal given the estimated model parameters. Use the UK as the country and the model from task 2.

```
## The probability that Great Britain wins at least one medal is 0.9654
```

Task 5

Perform model selection for the different models from tasks 1 to 3, respectively, and report your results. Which model would you choose to accurately predict the medal count? Justify your reasoning. Explain the model and comment on your findings.

Two tools were used to select the best model:

1. Cross-Validation

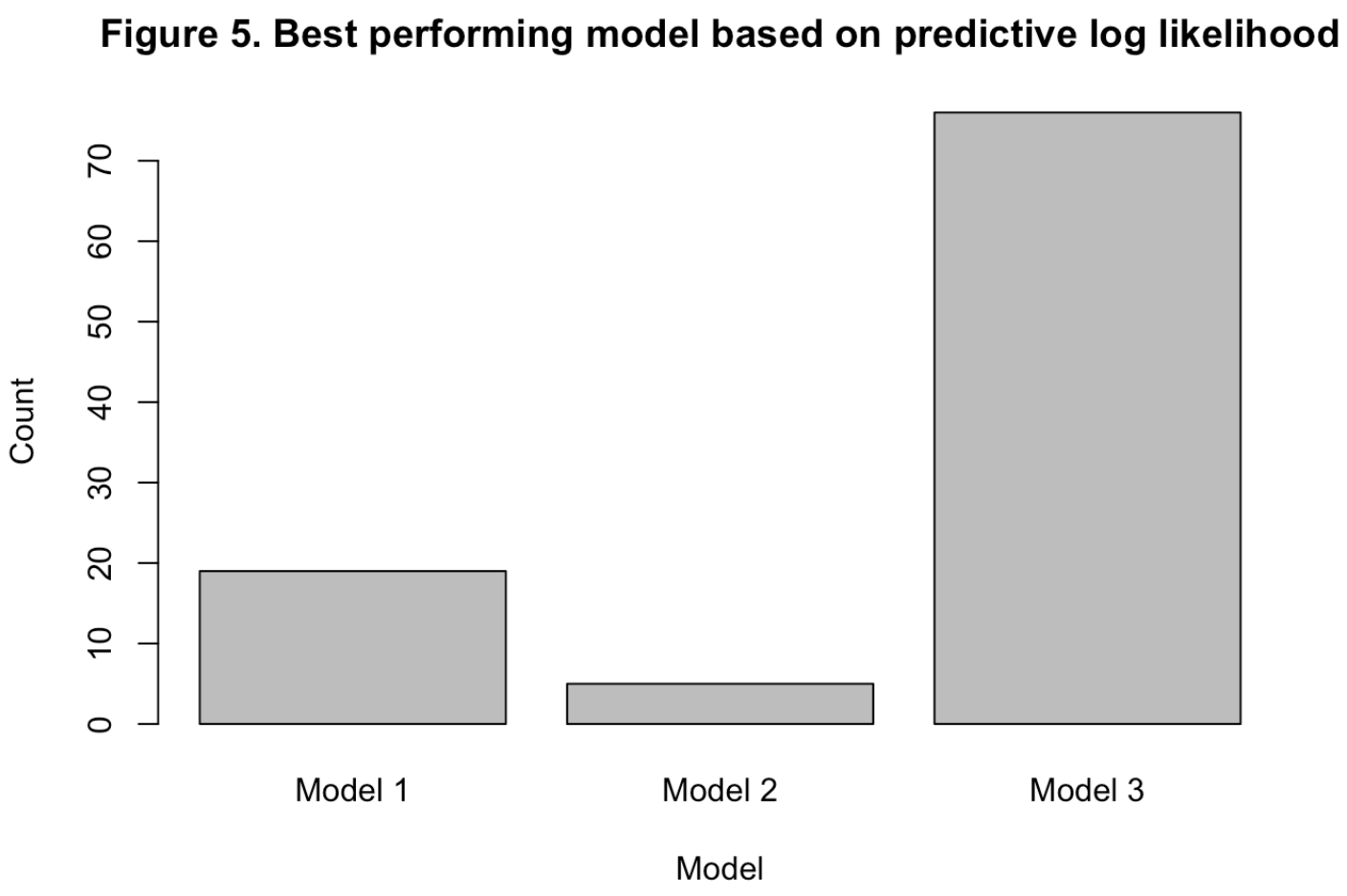
2. Root Mean Squared Error (RMSE)

Cross-Validation Cross-validation was used to evaluate the performance of each model on a test dataset. The performance of the model was measured by the predictive log-likelihood, as calculated by the cross-validation.

The data was split into a train set and a test set. Each model was trained on the train set and then evaluated with the test set. This process was repeated hundred times, and the average log-likelhood across both the train set and the test set was calculated.

As seen in Figure 5 below, Model 3 had the highest predictive log-likelihood the most times, making it the best model to use for predictions.

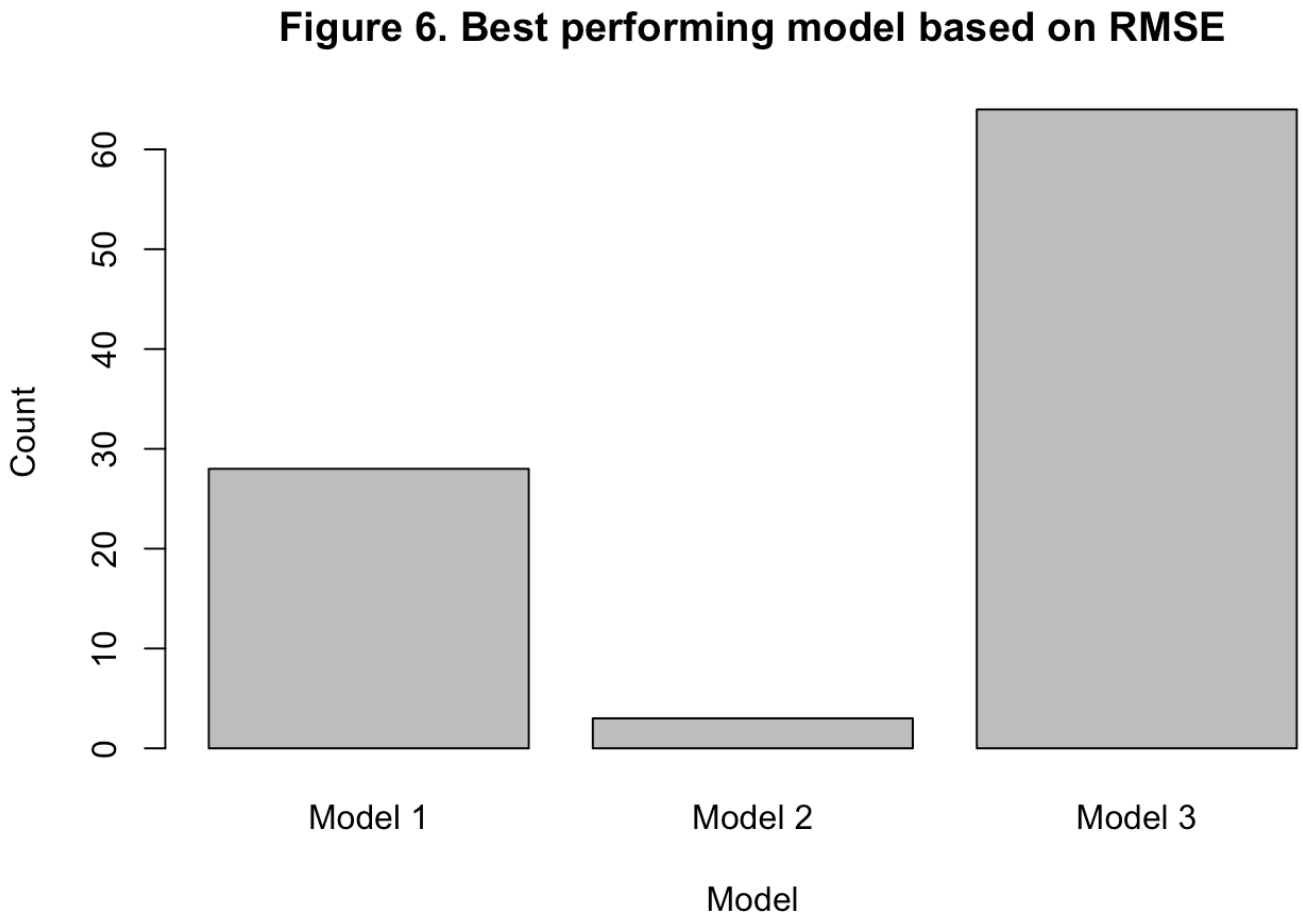
```
winner_ll_freq <- table(winner_ll)

barplot(winner_ll_freq, main="Figure 5. Best performing model based on predictive l
og likelihood", xlab="Model", ylab="Count")
```




```
#winner_mse_freq <- table(winner_mse)
#barplot(winner_mse_freq, main="Best performing model based on MSE", xlab="Model",
ylab="Count")
```

RMSE The best performing model based on RMSE is shown in Figure 6 below. Model 3 had the lowest RMSE the most times, indicating it is the best model to use.

```
winner_rmse_freq <- table(winner_rmse)
barplot(winner_rmse_freq, main="Figure 6. Best performing model based on RMSE", xla
b="Model", ylab="Count")
```



Overall, Model 3 is the best Model to use to predict future olympics because it was most often the model with the highest predictive log-likelihood and lowest RMSE.

Model 3 uses kmeans cluster to the countries into 4 Clusters using the log-transformed inputs. These clusters are then assigned to the data with the original, raw, non log-transformed inputs and then uses Model 1 to predict. This approach improves the model accuracy as seen in the results here.