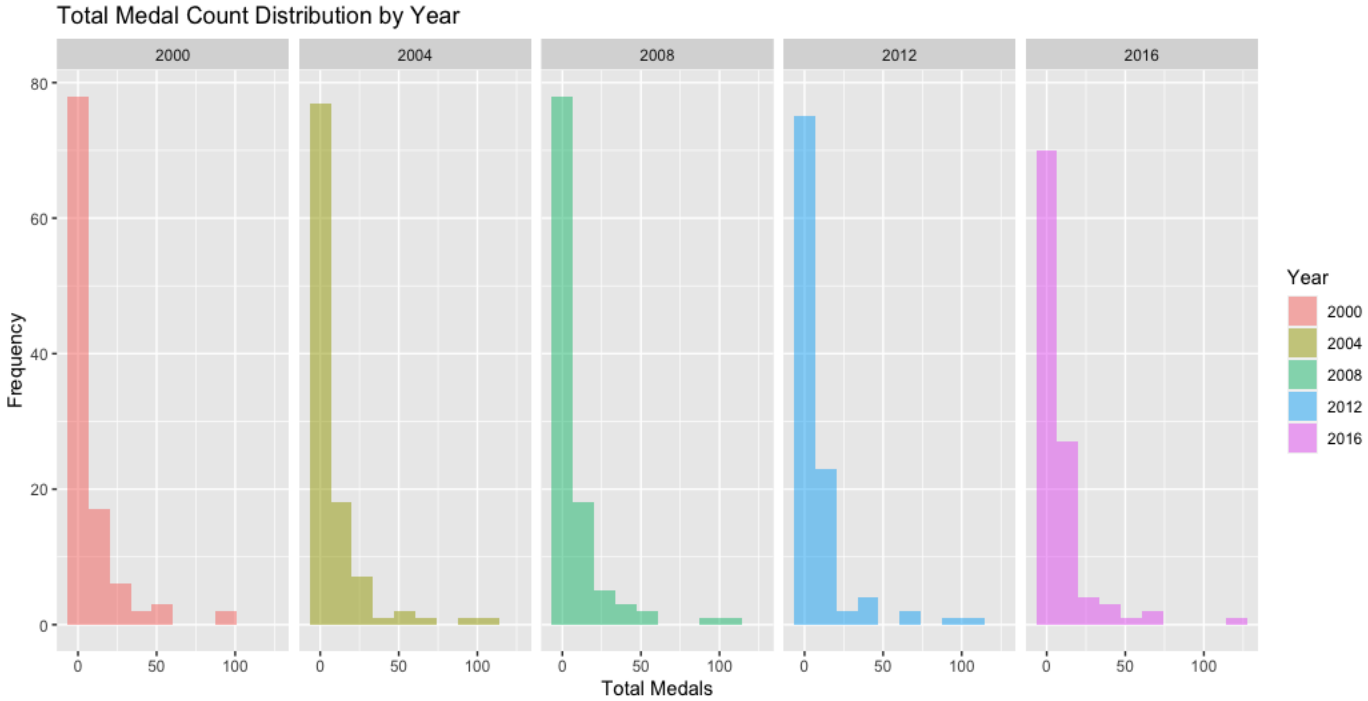
Predicting Tokyo Olympic Medal Counts

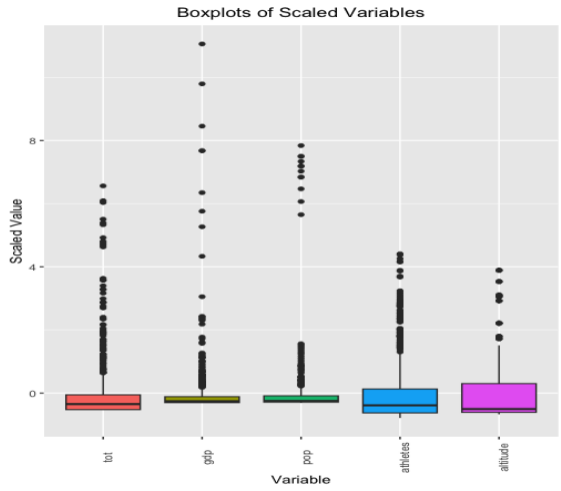
Introduction

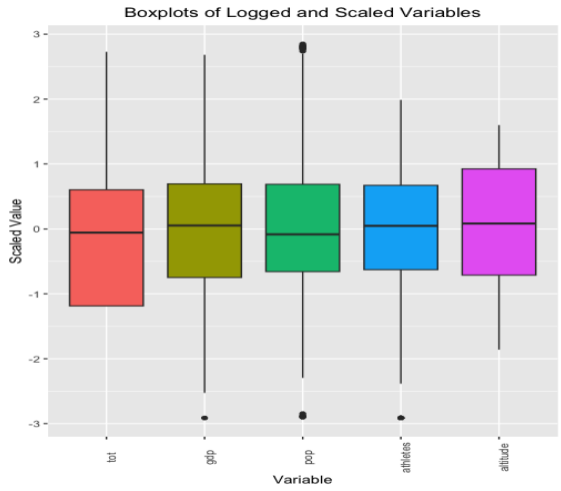
This report aims to develop models for predicting the number of medals won by each country at the Tokyo 2020 Olympics, using available pre-game information. Our goal is to estimate the total number of medals, including gold, silver, and bronze. As a result, because gold medals are inherently part of the total medal variable, we will not include them in our predictions. This prevents possible problems with collinearity between these variables. A new variable, "prev_tot," was generated by applying a lag operation on the "tot" variable for each country. By including the previous total medal count as a predictor in a predictive model, we can capture the historical performance and trends of countries in terms of winning medals.

We find little change when analysing the distribution of total medal counts over the five-year period. The histograms provided support this observation. Typically, the majority of nations take home 0 to 2 medals. The number of nations winning more medals after this point sharply declines. For example, approximately 22% of countries did not win any medals in 2012. The presence of a high proportion of zeroes suggests a potential excess of zeros in the data, indicating that a significant number of countries may have a low probability of winning any medals. Although, there is a very small number of countries included in our file who at times have not entered any athletes.



The scaled boxplot summaries provide a clearer visualization of the data distribution and outliers. Upon examination, we observe that the distributions remain largely consistent across different Olympic Games editions. Furthermore, the boxplots reveal that medal counts deviate from the typical distribution and qualify as outliers. However, applying a logarithmic transformation to the total medal counts yields a more symmetrical distribution, enhancing the data's overall symmetry and balance.





Handling Data

When the data are first examined, a number of problems show up that might present difficulties when modelling the data. These problems include missing values indicated by "#NA" and incorrect data types where some columns are set to character rather than integer.

1. Character variables: It is clear from looking at the dataset that GDP00, GDP16, and GDP20 are character variables. It is essential to change them from character to integer format in order to understand the distribution of these variables and perform thorough summaries.
2. Missing Values: NAs in the data; once we transform these variables from character to integer, we cansee that some NAs have been forced as a result.

```
> print(missing_counts)
    country country.code       gdp00       gdp04       gdp08       gdp12       gdp16       gdp20       pop00
          0            0           1           0           0           0           2           5           0
      pop04       pop08       pop12       pop16       pop20       pop21      soviet       comm      muslim
          0           0           0           0           1           1           0           0           0
   oneparty      gold00      gold04      gold08      gold12      gold16      gold20      tot00       tot04
          0           0           0           0           0           0           0           0           0
      tot08       tot12       tot16       tot20    totgold00   totgold04   totgold08   totgold12   totgold16
          0           0           0           0           0           0           0           0           0
 avmedals00  avmedals04  avmedals08  avmedals12  avmedals16  avmedals20    altitude  athletes00  athletes04
          0           0           0           0           0           0           0           0           0
 athletes08  athletes12  athletes16  athletes20        host
          0           0           0           0           0
```

GDP values and the missing population values were imputed using this algorithm. missForest is an imputation algorithm that utilizes random forests to estimate missing values based on the relationships and patterns observed in the available data. Presented below are examples of imputed values, along with the actual recorded GDP values for the corresponding years, encompassing different time periods:

country	gdp00	gdp04	gdp08	gdp12	gdp16	gdp20
Afghanistan	4894	5285	10191	20537	19469	20116

country	gdp00	gdp04	gdp08	gdp12	gdp16	gdp20
Cuba	30565	38203	60806	73141	88339	107352
Syrian Arab Republic	19326	25087	52631	73672	68588	74577

3. Data Format: The data was initially in a wide format, but we transformed it to a long format for better analysis. This change allows us to examine specific years and countries. We will train models using data from 2000-2012 and validate them with the 2016 data. The performance will be tested using the 2020 data.

••Linear regression can predict negative values because it models a continuous relationship between the predictor variables and the response variable, allowing for the estimation of values below zero.

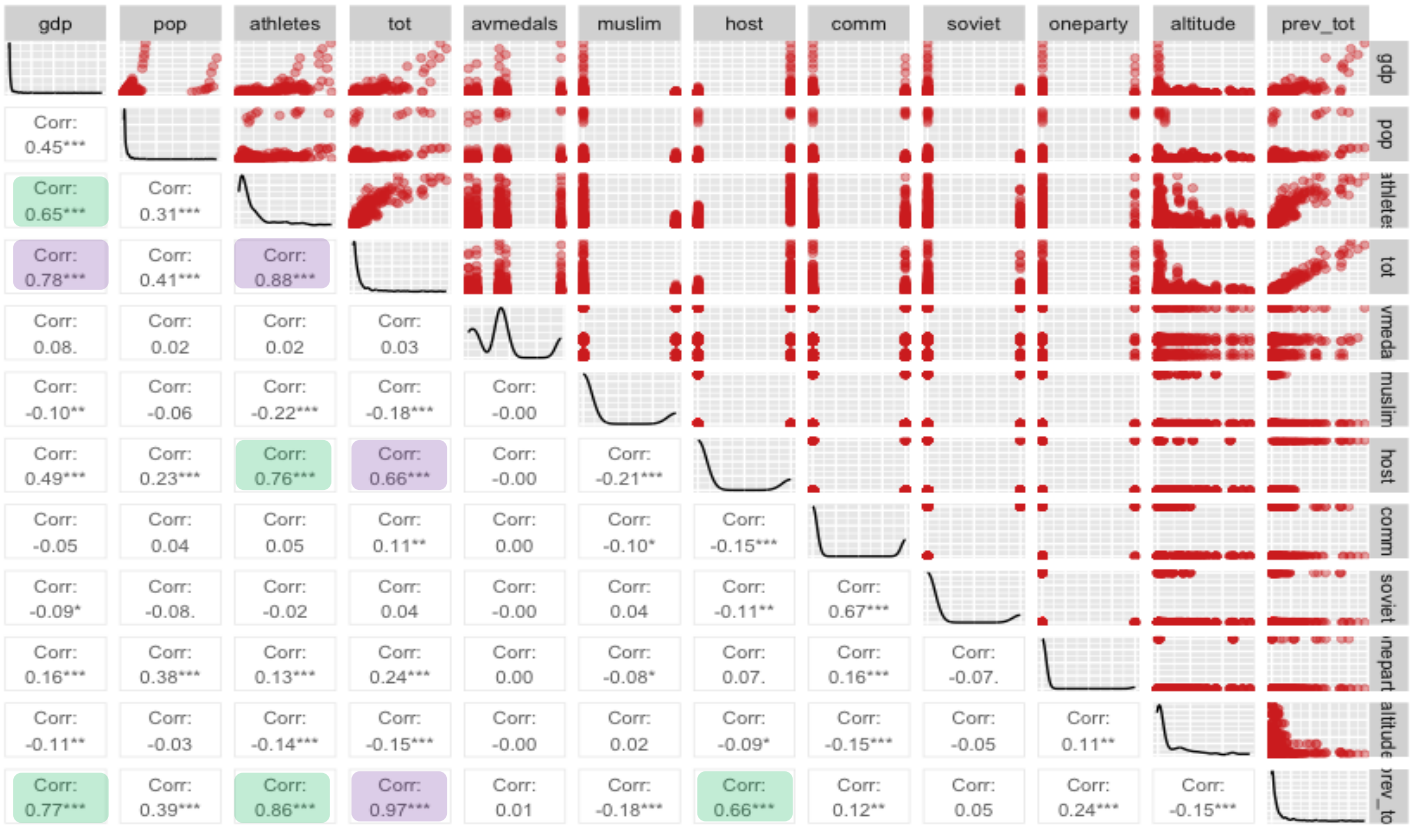
Explanatory Data Analysis

Summary Statistics

Total Medals 2000-2012 Summary Statistics						
Min	1st Qu.	Median	Mean	3rd Qu.	Max	Variance
0.000	0.000	3.000	8.667	8.000	112.000	283.6102

Given that the data involves count data, we can examine the variance and mean to determine whether the distribution resembles a Poisson distribution by looking at the total number of medals awarded to all nations from 2000 to 2012. In contrast to the strict Poisson requirement that the mean and variance be equal, the variance is found to be greater than the mean. This implies that there may be over-dispersion in the data. This should be considered when modelling the data, and we should consider other models to address excess zeros and reduce over-dispersion, such as zero-inflated models or hurdle models.

Pairs Plot

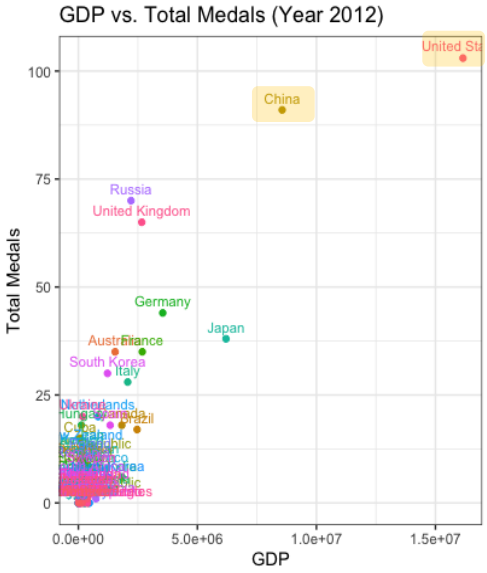


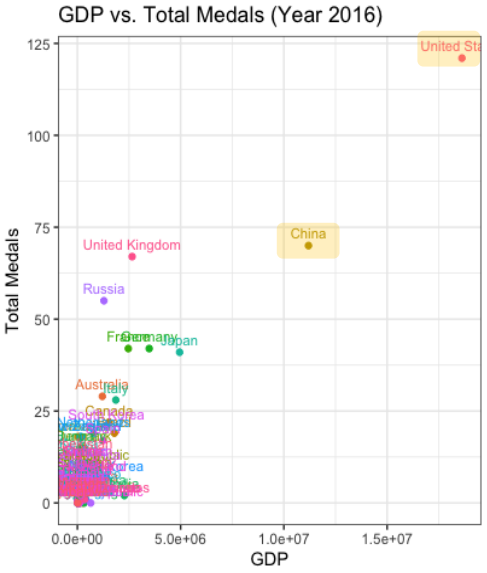
The pairs plot reveals the strongest correlations with the total medal count. Specifically, we observe high correlations between the total medals count and variables such as GDP (0.78), the number of athletes representing the country (0.88), the country's previous year's total number of medals won (0.97), and whether the country has hosted or will host the games (0.66). However, it is important to note that due to the transformation of the data into the long format, the observations are grouped by country, which affects their independence. Additionally, the presence of multicollinearity is indicated by the green boxes observed in the plot.

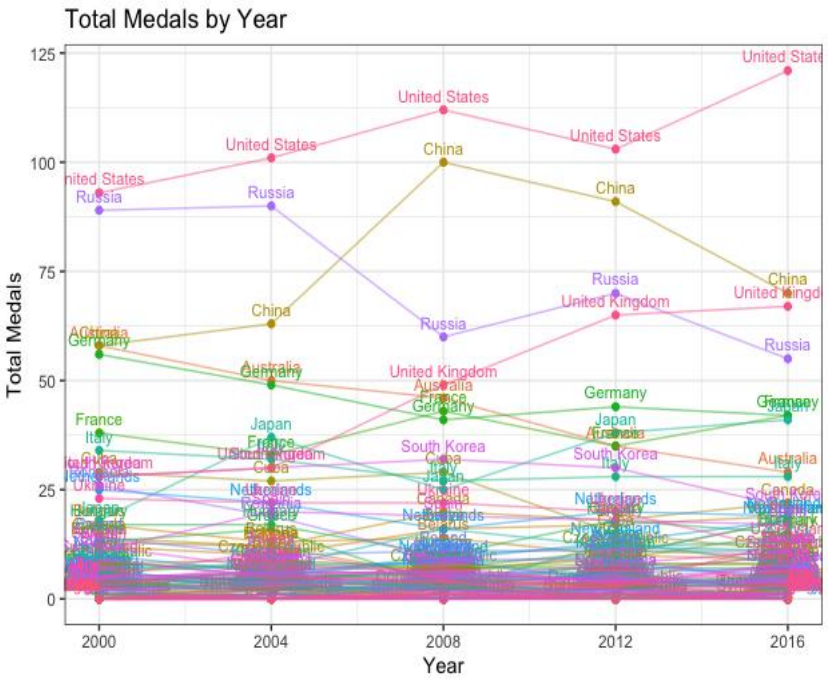
••Linear regression can predict negative values because it models a continuous relationship between the predictor variables and the response variable, allowing for the estimation of values below zero.

GDP vs Total Medals

There are some GDP outliers, as can be seen by looking at the plots for the years 2012 and 2016. Due to their high GDPs and medal counts, China and the US stand out particularly. The data distribution suggests a weak linear relationship between these variables.





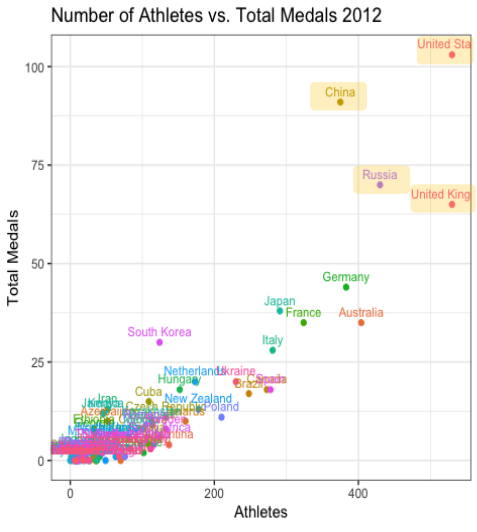


Upon closer examination of the outliers in the plot, it becomes evident that these countries consistently stand out as top performers over the years. This observation suggests a persistent trend where countries with higher GDP tend to excel in terms of medal counts. While there may be fluctuations in their performance over time, these countries consistently maintain a significantly higher count than the average.

This finding highlights a consistent association between higher GDP and better performance in terms of total medal counts in the Olympics. It suggests that countries with stronger economic resources are more likely to invest in sports programs, infrastructure, and athlete development, resulting in a sustained competitive advantage in the Olympic Games.

Athletes' vs Total Medals

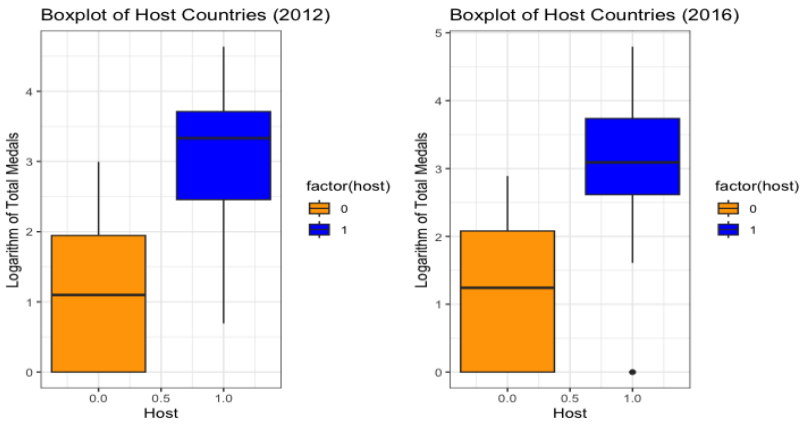
According to the analysis, there is a strong positive correlation between the number of athletes who represent a nation and the overall medal total they are likely to win. Notably, a few nations stand out as outliers in this relationship, including the US, the UK, Russia, and China. These countries consistently bring some of the most athletes to the games, earning some of the highest medal totals in both 2012 and 2016. By looking at the previous plot we can confirm our analysis.

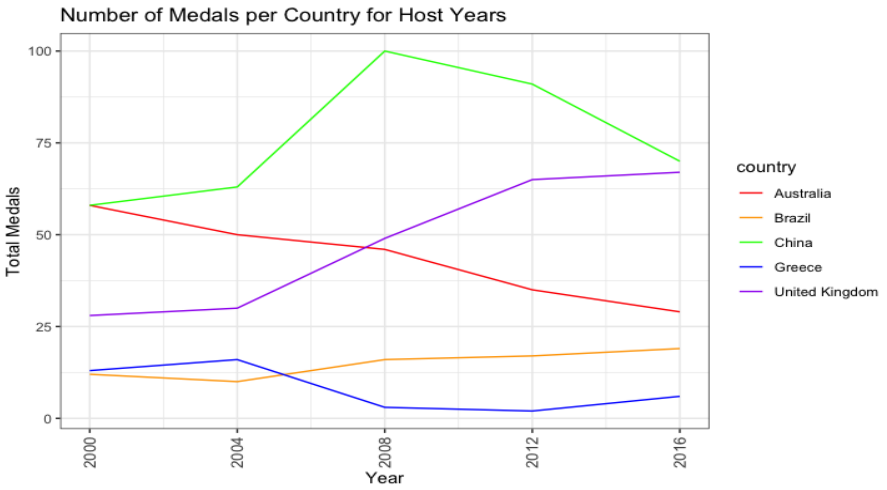


••Linear regression can predict negative values because it models a continuous relationship between the predictor variables and the response variable, allowing for the estimation of values below zero.

Host vs Total Medals

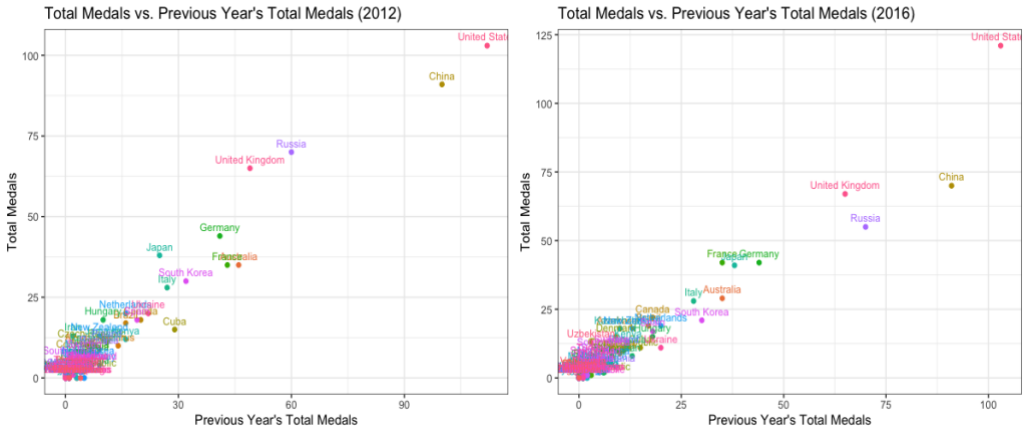
It seems that countries that have hosted the Olympic Games are more likely to win a larger number of medals overall. This finding is consistent with the high degree of multicollinearity among the variables. Additionally, there is a clear correlation between having a higher GDP and being a host country.





When we examine recent Olympic host nations like Australia (2000), Greece (2004), China (2008), the United Kingdom (2012), and Brazil (2016), we notice a recurring pattern. With the possible exception of Australia for gold medals, each nation saw a peak in the number of medals won during the years they served as hosts. With improved infrastructure, training, and national motivation, hosting the Games seems to be associated with higher medal achievements.

Previous Total Medals vs Total Medals



We can consider a country's past success or failure in winning medals by looking at its overall medal total. This variable includes significant elements like a nation's overall sporting prowess, sports investment, development initiatives, and strategic planning. It offers a starting point for comprehending a nation's potential to win medals.

The plot between previous total medals and overall total, which shows a clear linear relationship, suggests that countries with higher prior medal counts tend to have higher current medal counts. This connection implies that, at least in part, past success can predict future success.

••Linear regression can predict negative values because it models a continuous relationship between the predictor variables and the response variable, allowing for the estimation of values below zero.

Model Development

The data was divided into three parts: a training set consisting of observations from 2000 to 2012, a validation set with data from 2016, and a test set using observations from 2020.

Four regression techniques were investigated in the analysis:

1. Normal Linear Model
2. Poisson Model
3. Zero-Inflated Model (Poisson)

These models were evaluated and compared based on several performance metrics, including the root mean squared error (RMSE), the Akaike Information Criterion (AIC), the Mean Absolute Error (MAE), and the examination of residuals.

- The **RMSE** is a measurement of the typical discrepancy between the predicted values and the actual values that shows how accurate the model is on average. The performance of the model improves with decreasing RMSE.
- The **AIC** is a statistical measure used to assess the relative quality of different models. It considers the model's goodness of fit, and the number of parameters used. A lower AIC value indicates a better-fitting model.
- Another metric that measures the average discrepancy between predicted and actual values is the **MAE.** Lower MAE values indicate better performance. It gives a measure of the model's average prediction error.

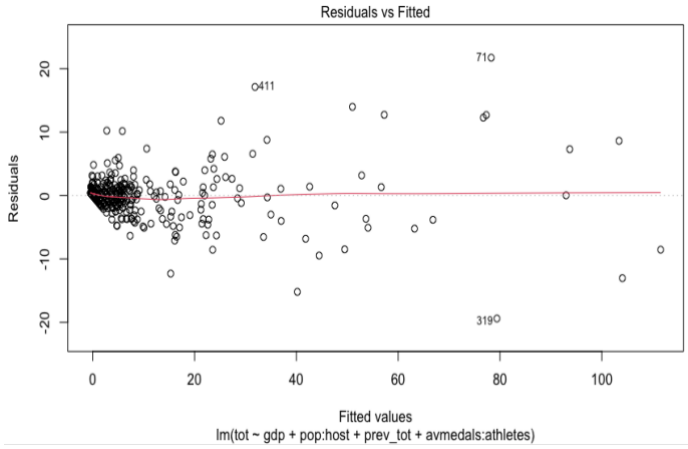
Normal Linear Model

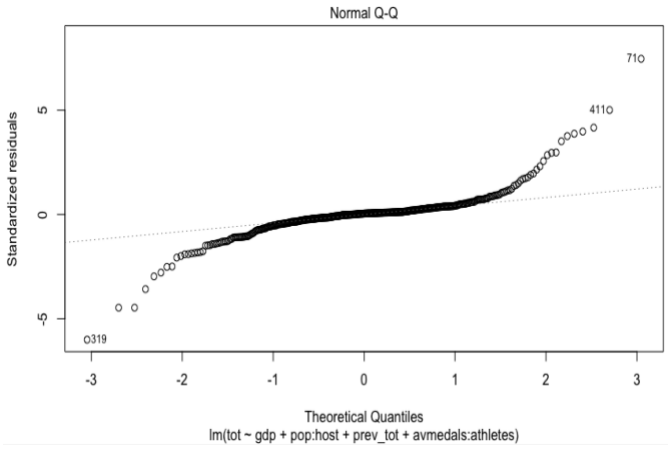
We begin with a linear model, and test a combination of variables to identify their significance and therefore which should be maintained within the model based on their p-values, as well as some step AIC testing, ultimately the best linear model appears to be:

$$\boldsymbol{Total\ Medals} = \boldsymbol{\beta_{GDP}} + \boldsymbol{\beta_{pop}} : \boldsymbol{\beta_{host}} + \boldsymbol{\beta_{prev\_tot}} + \boldsymbol{\beta_{athletes}} : \boldsymbol{\beta_{avmedals}}$$

The data seem to fit the linear regression model well. All of the predictors have significant coefficients ($p < 0.05$), which indicates their impact on the response variable, according to the model summary. The model's performance is evaluated using various metrics. The Adjusted R-squared value of 0.9581 suggests that approximately 95.81% of the variance in the response variable is explained by the model, indicating a strong relationship between the predictors and the number of medals won. The F-statistic is highly significant, indicating that the overall model is statistically significant.

The model's goodness of fit was further evaluated using residual analysis. The residual plots show that there is no apparent trend or systematic departures from the mean, and that the residuals are roughly normally distributed. The RSME, MAE and AIC were calculated for this model and will be compared with the others later in the report to conclude the model comparison.





••Linear regression can predict negative values because it models a continuous relationship between the predictor variables and the response variable, allowing for the estimation of values below zero.

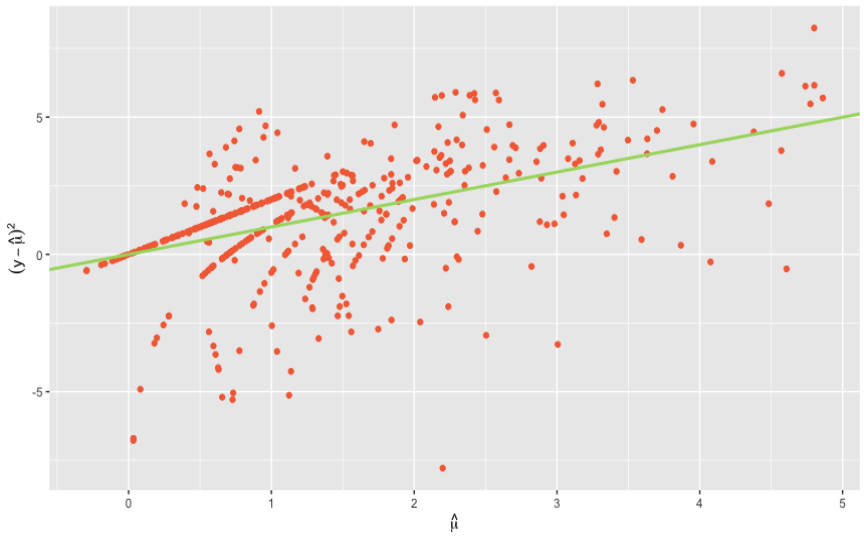
Poisson Model

A Poisson model was taken into account for the data while considering different variable options. It's essential to keep in mind that the Poisson distribution assumes that the variance and mean are equal. We anticipate possible over-dispersion in the Poisson model in this situation because the variance is greater than the mean. The Poisson model might not be the best option for the data, as evidenced by this departure from the assumption of equal variance and mean. Based on the lowest scoring AIC and RMSE the best performing Poisson model was:

$$\boldsymbol{Total\ Medals} = \boldsymbol{\beta}_{log(GDP)} * \boldsymbol{\beta}_{log(pop)} + \boldsymbol{\beta}_{prev\_tot} + \boldsymbol{\beta}_{athletes} + \boldsymbol{\beta}_{host} + \boldsymbol{\beta}_{comm}$$

The Poisson model summary illustrates clear evidence of over-dispersion in the data. This is demonstrated by the fact that the residual deviance, which is 1256.2, is significantly higher than the degrees of freedom, which are 424. The over-dispersion in the data is indicated by the significant difference between the residual deviance and degrees of freedom, which suggests that the model does not capture the variability present in the data adequately. The calculated value of the dispersion parameter is 3.03, further supporting the over-dispersion. This value, which is significantly higher than the typical over-dispersion parameter of 1, offers more proof of the data's excessive variability.

The over-dispersion that has been observed suggests that the Poisson model might not be the best option for modelling the data. To obtain more precise estimates and reliable predictions, alternative models that can consider over-dispersion, such as the negative binomial model or zero-inflated model.



Additionally, the dispersion plot shows that the data is widely dispersed both above and below the line of equality for mean and variance, confirming that the data is over-dispersed. As a result, we may conclude that this model does not provide a good fit for the data.

Also we can see that this model is severely underfitting zero counts. There are a total of 119 observations in the training dataset where the response variable has a value less than 1, indicating the presence of 0s in the response variable but our model is only able to predict about 40.

Zero- Inflated Model (Poisson)

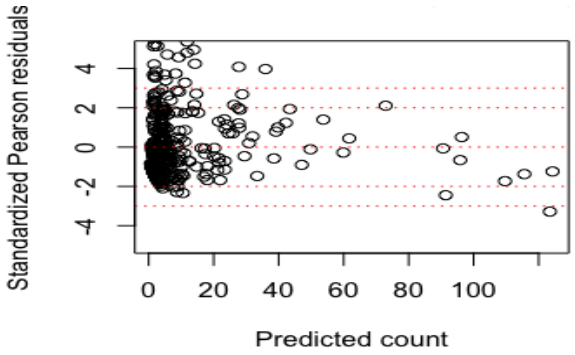
Finally, a zero-inflated model was applied to the data using the zeroinfl() function. Both the count model (Poisson with log link) and zero-inflation model (binomial with logit link) were estimated. The best performing model from this category appeared to be:

$$\boldsymbol{Total\ Medals} = \boldsymbol{\beta}_{comm} + \boldsymbol{\beta}_{log(pop)} * \boldsymbol{\beta}_{log(GDP)} + \boldsymbol{\beta}_{host} + \boldsymbol{\beta}_{prev\_tot} + \boldsymbol{\beta}_{athletes} \mid \boldsymbol{\beta}_{athletes} + \boldsymbol{\beta}_{prev\_tot}$$

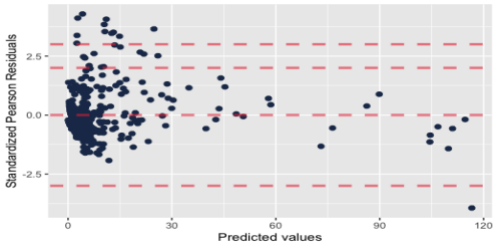
The coefficients in the count model show the relationship between the predictors and the expected count of non-zero values. The statistical significance of each predictor variable (p 0.001) indicates that they all significantly impact the expected counts. The zero-inflated model appears to fit the data more closely than the null model and the Poisson regression model, according to the likelihood ratio test (p 0.001). Although the overdispersion parameter is reduced to 1.81, indicating some improvement in handling overdispersion, it is important to keep in mind that the data still show signs of overdispersion which can be seen from the plots below.

••Linear regression can predict negative values because it models a continuous relationship between the predictor variables and the response variable, allowing for the estimation of values below zero.

Poisson Model



Zero-Inflated Model



Model Comparison

The four models will now be compared using a variety of comparative metrics to evaluate each one's performance and determine which model will provide the highest level of predictive accuracy. These metrics serve as indicators of how well the models fit the data, either by quantifying the amount of variance explained by the model or by evaluating the alignment between the predicted and observed values.

Model	RMSE	MAE	AIC	BIC
Normal Linear	3.42	1.93	2301.72	2326.13
Poisson	5.98	3.22	2394.20	2426.74
Zero-Inflated (Poisson)	5.22	2.81	2111.00	2155.75

It is clear from the comparison of the three models that each one has advantages and disadvantages. Due to its lower RMSE and MAE values, the Normal Linear Model exhibits greater accuracy in predicting the total number of medals won. This suggests that, in comparison to the Poisson and Zero-Inflated (Poisson) models, it offers estimates that are more accurate. On the other hand, because it has the lowest AIC and BIC values, the Zero-Inflated (Poisson) Model performs best in terms of model fit. This suggests that it achieves a better balance between model complexity and goodness of fit. The Poisson Model still offers a reasonable fit to the data, despite its lack of accuracy.

Considering both accuracy and model fit criteria, the Normal Linear Model emerges as the most favourable choice among the three models. Now, let's explore how well these models perform when it comes to predicting the 2016 total medals as validation and forecasting the 2020 total medals as a test.

Prediction

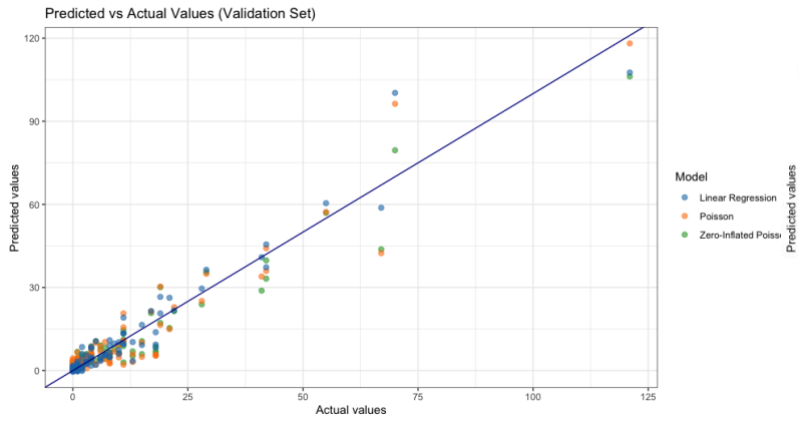
Based on the RMSE and MAE values, lower values indicate better predictive accuracy. Therefore, in terms of both validation and test performance, the Normal Linear Model outperforms the Poisson and Zero-Inflated (Poisson) models. It offers more precise predictions for both the validation and test datasets because it has the lowest RMSE and MAE values. However, the Zero-Inflated (Poisson) Model has the lowest AIC value when considering the AIC, which assesses the trade-off between model complexity and goodness of fit. This implies that, in comparison to the other two models, it offers a better balance between model fit and complexity.

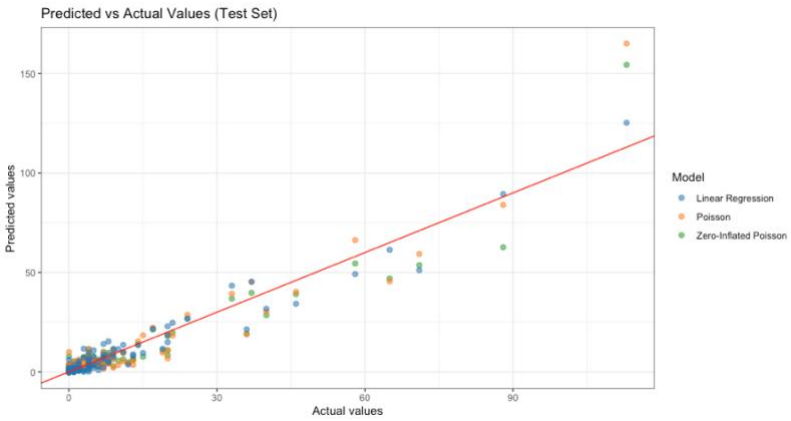
Model	Validation RMSE	Test RMSE	Validation MAE	Test MAE	AIC
Normal Linear	4.37	4.35	2.37	*2.77*	*2301.72*
Poisson	5.34	6.90	3.42	*3.75*	*2384.93*
Zero-Inflated (Poisson)	4.63	6.45	2.99	*3.45*	*2111.00*

Indeed, the Normal Linear Model appears to outperform the Poisson Model and the Zero-Inflated (Poisson) Model in terms of predictive accuracy, as indicated by the lower RMSE values on the test set. Additionally, the plots of the validation predictions and test predictions for all three models show similar results. There is also

••Linear regression can predict negative values because it models a continuous relationship between the predictor variables and the response variable, allowing for the estimation of values below zero.

underfitting present, it means that the models may not capture the full complexity of the data and fail to provide accurate predictions. This can be observed if the predicted values consistently deviate from the actual values in the plots.





We can go on to isolate some of the more extreme values from the top performing countries to see how this looks on an individual basis. Again, we can see that the Linear and the Zero-Inflated model are much closer to the actual observed values.

Country	2016 Actual	Linear Prediction	Poisson Prediction	Zero-Inflated Poisson Prediction
United Kingdom	67	58	42	44
China	70	89	84	63
Russia	55	57	57	60
United States	121	107	118	106
France	42	37	36	33

It is clear from comparing the 2016 Olympic results to the predictions made by the three models that the linear model offers the most accurate estimates of the observed values for the majority of the top-performing countries. The number of medals awarded to countries like the United Kingdom, China, and France is frequently underestimated by the Poisson and Zero-Inflated Poisson models. The fact that all three models exhibit a general tendency to underestimate the number of medals won by these nations, indicating some degree of underfitting in capturing the true medal-winning potential, is nevertheless significant. But when it comes to predictive accuracy, the Linear Model consistently outperforms the competition, demonstrating a closer fit with the observed data.

Country	2020 Actual	Linear Prediction	Poisson Prediction	Zero-Inflated Poisson Prediction
United Kingdom	65	61	46	47
China	88	89	64	63
Russia	71	51	61	54
United States	113	125	146	154
Japan	58	50	59	54
Mauritius	0	-0.1**	1	0

The performance of the three models is in line with the results from the validation set when looking at the predictions for the test set for the 2020 Olympics. Comparing the Linear Model to the Poisson and Zero-Inflated Poisson models, the Linear Model continues to show better predictive accuracy. For the majority of the countries, the predictions from the linear model closely match the actual observed values, accurately capturing the medal counts with only a few minor deviations.

••Linear regression can predict negative values because it models a continuous relationship between the predictor variables and the response variable, allowing for the estimation of values below zero.

In contrast, the predictions of the Poisson and Zero-Inflated Poisson models show a mixture of underestimation and overestimation. The Poisson Model tends to overestimate the counts for China and the United States while underestimating the counts for the United Kingdom, Russia, and Japan. Similar to this, the Zero-Inflated Poisson Model overestimates medal counts for China and the United States while underestimating them for the United Kingdom and Russia.

Conclusions & Considerations

Limitation of the Approach

1. The linearity assumption in the linear regression model is one of our method's limitations. Although this model was successful in our analysis, it depends on the predictor variables and response variables having a linear relationship. The relationship between these variables may, however, be more complicated and nonlinear. We could investigate more sophisticated modelling methods, like polynomial regression or regression with interaction terms, to overcome this limitation.
2. The use of a large sample size spanning several years to train the model could be another potential drawback. This gives a thorough overview of medal counts over time, but it might also cause problems with generalisation. We can more precisely and individually evaluate the model's performance by making predictions for just one year at a time. This strategy can also assist in addressing the worry that the observations might not be independent of one another because results from one year may have an impact on results from later years.
3. Every model used in the analysis contains unique data assumptions. While the Zero-Inflated Poisson Model assumes excess zeros follow a particular distribution, the Poisson Model assumes the mean and variance of the response variable to be equal. These presumptions might not always hold true in real-world scenarios, and if they are broken, the models' performance could be affected.
4. The limited amount of predictor variables used in our study's models is another drawback. Even though we have included some important elements, there might be additional significant variables or interactions that could enhance the models' capacity for prediction. Exploring additional predictor variables, such as the percentage of athletes with prior Olympic experience or data on each nation's investment in sports, could offer insightful information and possibly improve the models' accuracy.

Further Work for Model Improvement:

1. Ridge or Lasso regression are two examples of regularisation techniques that can be used to reduce overfitting and enhance generalisation to the data. Regularisation can handle multicollinearity and reduce model complexity, resulting in predictions that are more reliable and precise.
2. A more accurate assessment of the models' performance and assurance of their generalizability to new data could be obtained by carrying out a more rigorous cross-validation procedure, such as k-fold cross-validation. This would decrease the possibility of overfitting and validate the models' reliability.
3. Evaluation of the data in its original wide format can offer a more in-depth and granular perspective instead of aggregating the data by country. Keeping the variables separate and avoiding grouping them by country would entail doing this. The models can accurately represent the distinctive qualities and variations within each country's performance by treating each individual observation separately. This method enables a more nuanced analysis and may find additional insights and patterns that were possibly missed when the data was aggregated.

••Linear regression can predict negative values because it models a continuous relationship between the predictor variables and the response variable, allowing for the estimation of values below zero.