INTRODUCTION

The model described in this report was created to predict the total number of medals that will be won by each country in the 2016 Olympics held in Rio de Janeiro. The model is based on information regarding Games from 2000 to 2016.

The data set is composed of 15 variables. *TotYY* is the total number of medals won in a year by a country and it represents the response variable. The other variables are the explanatory variables and they give information, of all countries participating in Rio 2016 Olympics, about GDP in million US dollars and the population in different years; how many gold medals each country won in a certain year; the number of athletes representing each country in different years; and other information such as the altitude of each country's capital, if the country is a former/current communist state or a member of the former Soviet Union, or if it is a Muslim majority country or a one party state. Finally, the data set gives also information about whether a country has ever hosted the Olympics from 2000 to 2016.

The second part of this report is the exploratory analysis which provide information about the general relationship within variables. Then there will be an analysis of different models created using different distributions, in this case are Poisson and Negative Binomial. Finally, there will be discussion and conclusion about the best model predicting the number of medals won by a country in the 2016 Olympics.

EXPLORATORY ANALYSIS

The data must have the right format; therefore, they need to be checked before being used. Some variables have been modified as character. Furthermore, it has to be ascertained whether any transformations are needed, for example if it is appropriate to calculate the logarithm or the square root of the data. In this case since some variables have wide range of values, a log transformation has been applied to those variables. However, before doing that, a small amount (0.1) have been added to many values equal to zero of some variables, in order to not have NANs (Not A Number) after the log transformation.

The whole data set has been divided into a train set and a test set. The test set consists of all data regarding the Games of 2016. It represents the set on which the model created on the train data will be tested in order to check how well the model can predict future observations. The remaining part of the data set has been modified with the purpose to have all the observations together rather than divided by years.

Plots of the response variable against the explanatory variables have been made with the aim of examination of their relationships. In this phase only data from 2012 have been used.

In figure 1 it is possible to observe positive correlations between total number of medals won by a country and GDP, population, number of gold medals won and number of athletes representing each country. On the other hand, it is clear that there is no correlation between the total number of medals won by a country and the altitude of its capital, as it is shown in figure 2.

In figure 3 it is possible to notice that there is not a big difference between Soviet and No Soviet countries, and this is also the case for communist and no communist countries, whereas there is a substantial difference between Muslim and no Muslim countries, and this is also the case between countries with one party and countries with more than one party.

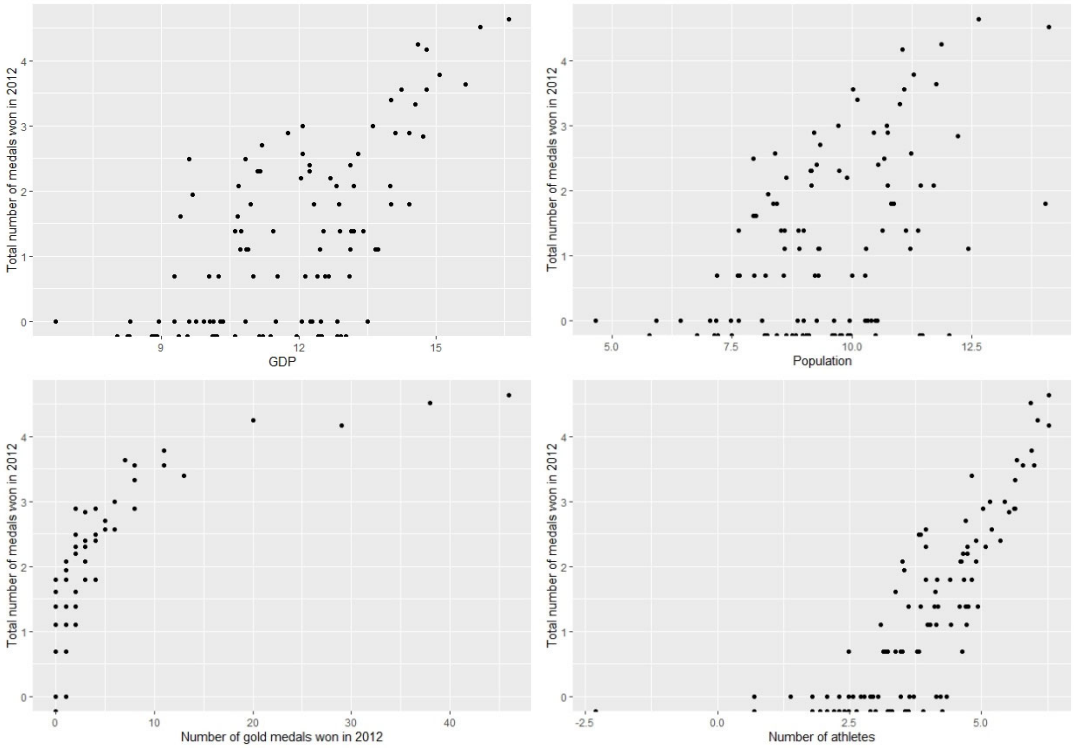

Figure 1: Total number of medals vs GDP, Population, number of gold medals and number of athletes

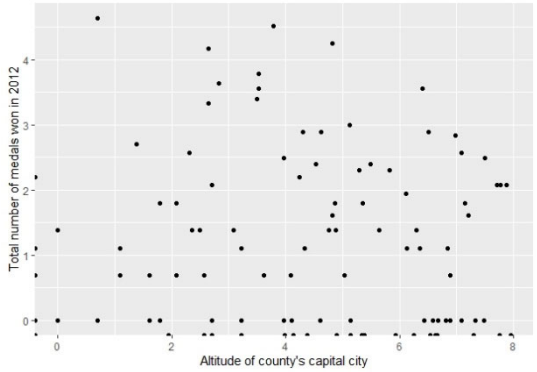

Figure 2: Total number of medals vs heights of each country's capital

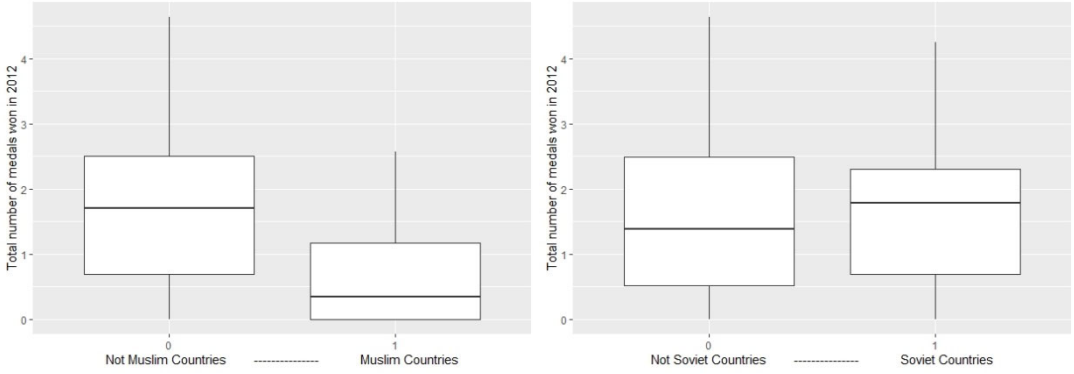

Figure 3: Total number of medals vs Muslim-No Muslim counties and Soviet-No Soviet countries

It is interesting to check how strong the correlations between the response variable and the explanatory variables are.

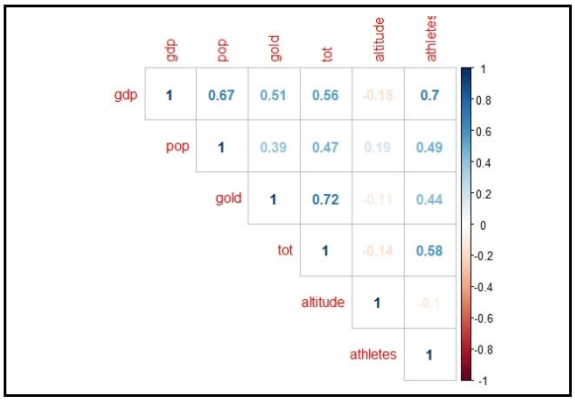

Figure 4: Correlation matrix

Figure 4 represents a matrix showing the correlations between all the numeric variables of the data set. In general, there is not a strong correlation between these variables: the highest is 0.72 and it is between *tot* and *gold*. Since *gold* is the number of gold medals won and *tot* is the total number of medals won, included the gold ones, these 2 variables are not independent because the higher is the number of gold medal won by a country, the higher will be the total number of medal won by that country. For this reason the variable *gold* will not be used in the models. All the other correlations are relatively weak, with the second highest, only related to the response variable *tot*, equal to 0.58 and it is versus the number of athletes.

ANALYSIS

In this phase models will be created using the train data which consist of all data from 2000 to 2012. The first model created is a very general one using glm function and Poisson distribution. The response is *tot* and all the explanatory variables have been used. The summary of this model, called model0, is the following:

```
Call:
glm(formula = tot ~ gdp + pop + athletes + host + soviet + comm +
    muslim + oneparty, family = poisson, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9665  -1.1843  -0.5463   0.6417   5.3828

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.58615    0.23480 -19.532  < 2e-16 ***
gdp          0.05671    0.02232   2.541 0.011060 *
pop          0.07100    0.02073   3.426 0.000613 ***
athletes     1.00732    0.03500  28.783  < 2e-16 ***
host         0.23242    0.06645   3.498 0.000469 ***
soviet       0.24819    0.07665   3.238 0.001204 **
comm         0.22466    0.07088   3.170 0.001526 **
muslim      -0.30838    0.09291  -3.319 0.000903 ***
oneparty     0.26418    0.09336   2.830 0.004659 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 8036.3  on 430  degrees of freedom
Residual deviance: 1051.3  on 422  degrees of freedom
AIC: 2191.3
```

```
Number of Fisher Scoring iterations: 5
```

This model has a very large Residual deviance (1051.3 on 422 degree of freedom) and a very large value of AIC (2191.3). For now, this is not really important because this first model is needed for variable selection process. Three techniques have been used: step, backward and forward and step using BIC criterion instead of AIC criterion of Mass library. The three technique show that the statistically significant explanatory variables are: *pop* (population), *athletes* (number of athletes representing the countries), *host* (whether the country has ever hosted Olympics or not) and *muslim* (whether the country is a Muslim majority or not). This is also confirmed from the summary of model0 since p-value is lower than 0.05 for these variables. Therefore, a second model called model 1 has been created using glm function and Poisson distribution on the afore mentioned explanatory variables. The summary of this second model is:

```
Call:
glm(formula = tot ~ pop + athletes + host + muslim, family = poisson,
    data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4685  -1.2146  -0.4815   0.6622   5.4793

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.06497    0.18622 -21.829  < 2e-16 ***
pop          0.14653    0.01311  11.180  < 2e-16 ***
athletes     1.09987    0.03049  36.077  < 2e-16 ***
host         0.05955    0.05383   1.106    0.269
muslim      -0.39763    0.09068  -4.385 1.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 8036.3  on 430  degrees of freedom
Residual deviance: 1142.7  on 426  degrees of freedom
AIC: 2274.7

Number of Fisher Scoring iterations: 5
```

CONFIDENCE INTERVAL MODEL 1		
	2.5%	97.5%
Intercept	0.01190491	0.02470568
Pop	1.12833011	1.18781893
Athletes	2.83056605	3.18988799
Host	0.95519272	1.17960819
muslim	0.56033933	0.56033933

This model still has very high values of residual deviance and AIC. It is possible to check if the residual deviance is led by outliers and also if a linear model is appropriate to explain the data.

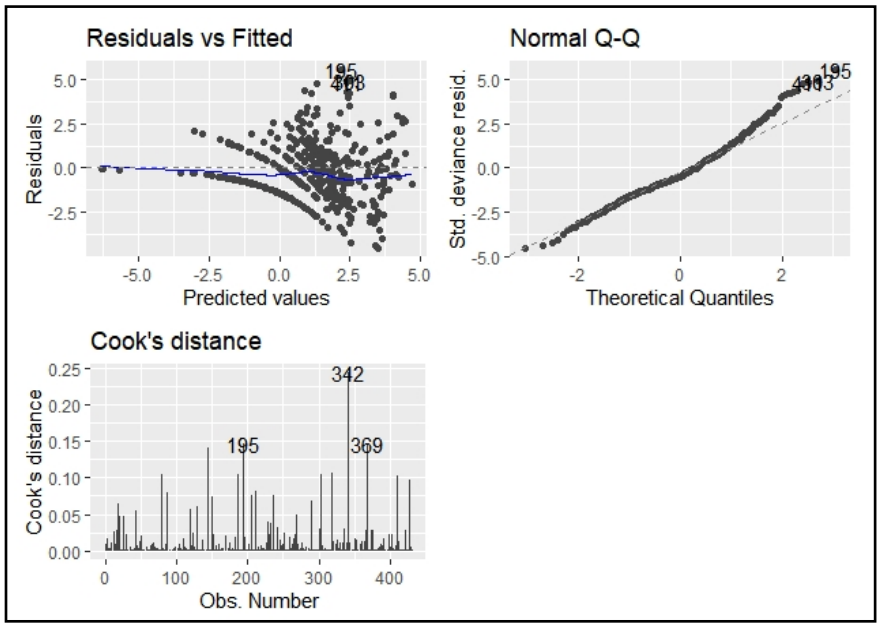

Figure 5: Outliers and linearity

Plot Residuals vs Fitted of figure 5 shows that the linearity might be questionable. The other two plots do not show any evidence of presence of outliers.

From the summary of model 1 it is possible to deduct that the factor *host* can be dropped, this is also confirmed from the confidence interval which contains 1. Thus, the next model, model 2, considers the variables *pop* and *athletes* and the factor *muslim*. This model does not seem to be better than model 1. Considering an interaction with the factor *muslim* (model 3) both the residual deviance and AIC slightly decrease.

In model 4 another distribution has been taken into account: The Negative Binomial. This model is based on a variable selection which indicates to use the variables *athletes*, *pop* and *comm*. Since the Poisson models has a very high residual deviance compared with the degree of freedom, probably due to overdispersion, a Negative Binomial can deal better in this kind of situations.

The summary of model 4 is the following:

```
Call:
glm.nb(formula = tot ~ athletes + pop + comm, data = train.data,
    init.theta = 3.722844049, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8327  -0.9209  -0.3940   0.4392   3.1714

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.63609    0.28587 -16.218  < 2e-16 ***
athletes     1.11204    0.04230  26.292  < 2e-16 ***
pop          0.11790    0.02827   4.170 3.05e-05 ***
comm         0.30552    0.08144   3.752 0.000176 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(3.7228) family taken to be 1)

    Null deviance: 2398.52  on 430  degrees of freedom
Residual deviance:  470.02  on 427  degrees of freedom
AIC: 1940.9

Number of Fisher Scoring iterations: 1
```

```
              Theta:  3.723
          Std. Err.:  0.528
 2 x log-likelihood:  -1930.863
```

Both the residual deviance and AIC have dramatically decreased, going from values of 1142.7 and 2274.7 to 470.02 and 1940.9, respectively. Model (model 5) has been created considering an interaction between *athletes* and *comm* and *pop* and *comm*. The summary of model 5 shows no improvement from the previous model.

```
Call:
glm.nb(formula = tot ~ athletes * comm + pop * comm, data = train.data,
    init.theta = 3.78457114, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8261  -0.9161  -0.3898   0.3716   3.2440

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.58174    0.75421  -6.075 1.24e-09 ***
athletes       1.32745    0.14133   9.392  < 2e-16 ***
comm           0.29320    0.51471   0.570    0.569
pop            0.01075    0.08861   0.121    0.903
athletes:comm -0.18811    0.11679  -1.611    0.107
comm:pop       0.09079    0.06661   1.363    0.173
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(3.7846) family taken to be 1)

    Null deviance: 2421.60  on 430  degrees of freedom
Residual deviance:  470.39  on 425  degrees of freedom
AIC: 1942.2

Number of Fisher Scoring iterations: 1


              Theta:  3.785
          Std. Err.:  0.541

 2 x log-likelihood:  -1928.162
```

However, as expected, the Negative Binomial is a better choice when dealing with overdispersion. In effect, the 95th percentile of the chi squared distribution with degrees of freedom 427 is 477, lower than 470.02: this indicates no evidence of lack of fit.

CONCLUSIONS

Predictions are made on model 3 because has the lowest values of residual deviance and AIC amongst the models using Poisson distribution, and model 4 which also has the lowest values of residual deviance and AIC amongst the model using Negative Binomial distribution. These predictions are made using test data set. To decide which of the two models, and so distributions better predicts the total number of medals won by a country, the root mean squared error (RMSQ) has been used. RMSE is the square root of the mean of the squared difference between the fitted and observed values and for the two models are:

Distribution	RMSQ
Poisson	8.073518
Negative Binomial	8.341499

From the values above seems that the better choice in this case would be model 3 with Poisson distribution.

FUTURE IMPROVEMENT

This analysis can be improved by adding more data to the data set about other Games, not only Olympics but also regional or local games, in order to have more information about athletes or any other factor which can affect their performance during Olympic games.

More detailed models can be created using for example also other distributions like Normal or Zero Inflated. A better variable selection can be used and also checking thoroughly any interactions with factors. As shown in the plot Residuals vs Fitted of figure 5, it would be equally interesting to crate nonlinear models.