**Olympics Games Analysis**

**Vineet Kapoor|11910076|ISB CBA Practicum|23rd September, 2018**

**Executive Summary:**

* There exist a strong correlation between Olympics performance with Nominal GDP, Number of internet users, country is host or not.
* Gender Gap index and Population of a country is correlated to some extent with the count of total medals.
* Countries like US bag more medals in Swimming due to height of their players.
* Nominal GDP is responsible for good performance of Russia , Australia, Germany, Great Britain and France in Olympics.
* Number of internet users also has some relation with Total medals tally.
* The host country has good performance in Olympics.
* China’s population and it’s sports culture, sports schools, sports trainees are responsible for it’s good performance.

**Business Problem:**

* Which countries dominate in Olympics and why?
* To Analyse different countries on the basis of Age, Height and Weight. of players, who won the medals for all seasons
* To Cluster different countries on the basis of Age, Height and Weight. of players, who won the medals for Summer and Winter seasons.
* To Cluster different countries on the basis of GDP, population and total medals.
* Identify the factors that are driving India behind other countries.
* Predicting the rank or performance of top 20 countries in next Olympics.
* Predicting the total medals , gold, silver , bronze medals of top 20 countries in next Olympics in Tokyo -2020.

**Introduction:**

* Olympics event is held every four years, where different teams participate across the world. In this project, clustering and different regression techniques have been used to solve out important questions related to the performance of players of best performing countries and it is compared with the India’s performance in Olympics.

**Motivation For Study:**

* The prediction of total medals from regression and other models can be used for building a predictive model . The K- fold cross validation and comparison of RMSE can improve the accuracy of the model.
* The K – means clustering of the GDP , population and number of medals will help to find out the relation of GDP, population with the performance of countries in Olympics.
* The K – means clustering of the Age, Height and weight for Summer and winter seasons can help to find the comparison between genes of Olympic athletes in top countries.
* Regression analysis will provide the further insights.

**Methodology & Tools:**

The Data has been taken from

Analytic Approach:

1. **The data has been filtered out for only those players , who have won medals in the original given dataset. The data is cleaned and all the missing values has been imputed. An additional data using GDP , population and total number of medals was also used for clustering and analysis is performed.**
2. **The Dataset for regression has been taken from Wikipedia, economic freedom index site, World Bank, Gender gap index.**
3. **Data was collected for top 20 countries in 2016 summer Olympics and India also.**
4. **Two datasets are used for analysis. The dataset used for clustering was cleaned. Missing values were imputed by their mean values. Dataset used for Regression analysis is created using the information from the Wikipedia and sites. Total number of medals for top 20 countries in Olympics 2020 are predicted**
5. **The training and validation data was created using sampling and K- fold cross validation. The whole modelling exercise is done on python using sklearn package.**
6. **K- Means Clustering has been performed on the age, weight and height variables for summer and winter seasons. and on whole data**
7. **K- Means Clustering has been performed on the age, weight and height variables**
8. **Methods used for performing regression models are**

* **Ordinary Least Squares**
* **Random Forest.**
* **Gradient Boosting Regressor .**
* **K Nearest Neighbours regressor.**
* **Optimization of Hyperparameters is also performed using Grid Search CV method.**

**Models:**

* OLS- ordinary least squares Regression Model
* GBM – Gradient Boosting Regressor model
* KNN Regressor model
* Random Forest model
* Optimization of Hyperparameters of KNN regressor model using Grid Search CV method
* Xgboost regressor model.
* K- Means clustering method, PCA.
* K fold Cross validation model.

**Message:**

**Means:**

* Visualization Tools - Tableau, Python, R.
* Modelling Tools – OLS model, K-means model, Lasso Regression model, KNN regressor, GBM, XGBregressor, Random forest packages from sklearn are used.
* R studio - K-means clustering is performed in R studio.
* Python – Regression analysis using different machine learning models is performed in Python. Optimization is also performed in python.
* Libraries used – in R: ggplot2, forecast, zoo, cluster, factoextra, magrittr, dplyr,
* In Python: numpy, pandas, matplotlib, seaborn, StandardScaler, statsmodels.formula.api.

**Connecting CBA Term – 1 courses:**

* K- means Clustering, PCA,
* Multiple Regression , Evaluation of Models and
* optimizing the Hyperparameters using Grid search method.

**Data collection & cleaning:**

**Analytic Approach:**

* Check for correlation between different variables with total medals and individual – gold, silver, bronze medals.
* Check for factors responsible for predicting total count of medals.
* To forecast the number of Gold medals for countries -
* To build a Multiple Regression Model using different techniques and evaluating the models using RMSE, LRMSE.
* To build clusters of observations using GDP, population, total medals data using K-means algorithm.
* Analyse the top 20 athletes, top 10 male and female athletes.
* Analyse mean height of players for all countries , who won at least one medal.

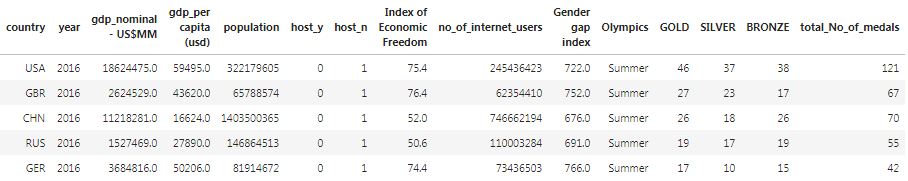
**Approach of Solving the prediction of medals in 2020 Olympics- Tokyo.**

1. Firstly, I collected the data from Wikipedia and various sites. Dataset has been created from scratch.

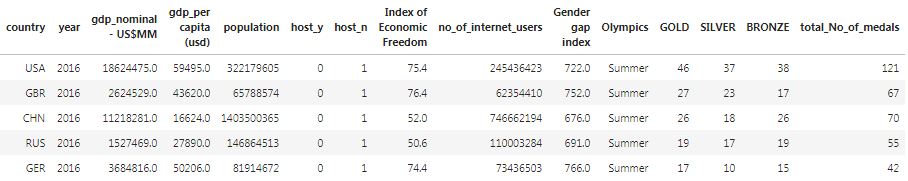
* **Conditions - Only summer olympics has been used.**
* **Only information from 2000 till date has been used for prediction.**
* **Only top 20 ranked countries from 2016 summer olympics have been used.**
* **India is also included as part of analysis.**
* **Variables have been used selectively after careful research of Olympics using google.**

**Objective – Predicting the count of total medals for these 21 countries for 2020 olympics in Tokyo. Count of Gold , silver and Bronze medals is also predicted using different machine learning models. The count of medals will be rounded off to greater integer for ranking purpose.**

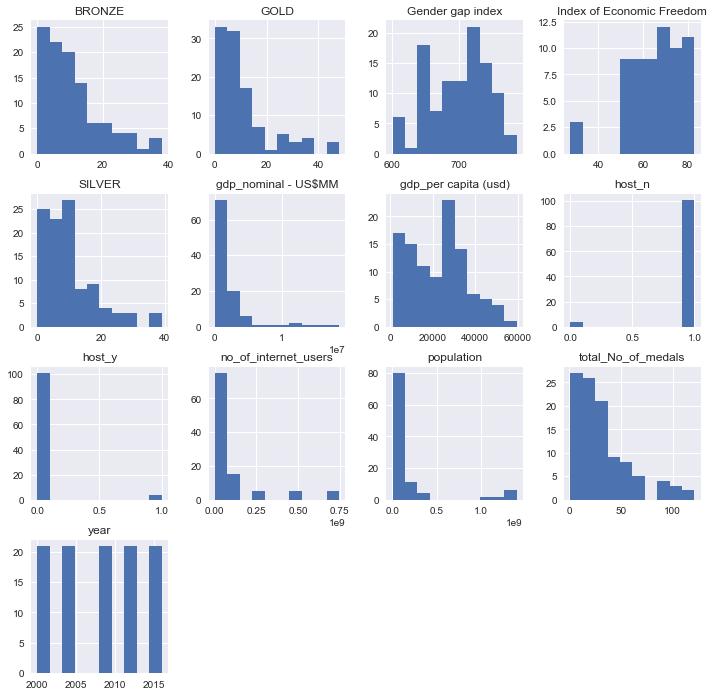
1. Variables that are used in regression are: Nominal GDP in US$, GDP per capita, Population, Gender Gap Index, No of internet users, Economic freedom index, Country is a host or not, Total medals.



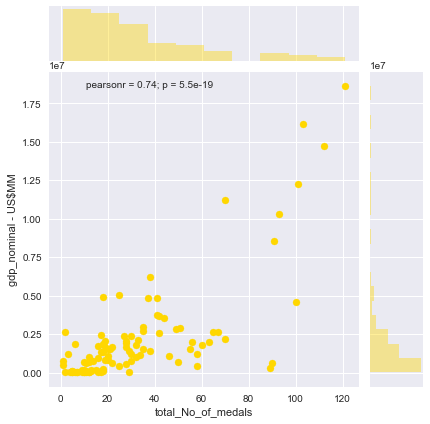
1. Then, the data is consolidated and cleaned. Missing values in GDP per capita and Nominal GDP in US ($) are imputed by the median and only few variables are selected for further analysis. Variables given below were used for further analysis.



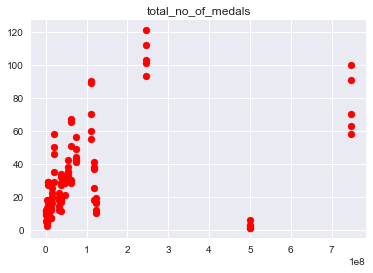
1. Descriptive statistics, Scatter plot of features and target variable has been plotted.
2. Histogram of individual variables



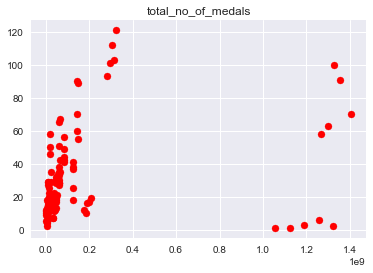
Scatter plot between Total medals and Nominal Medals.



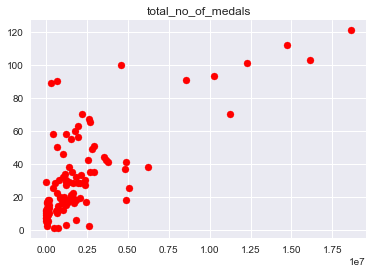
Scatter plot between internet users and number of models



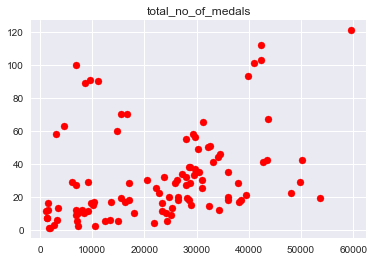
Scatter plot between population and number of models



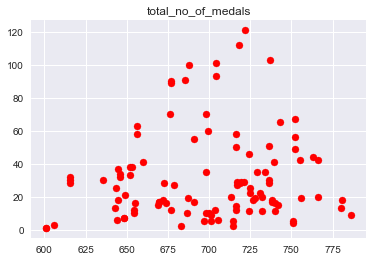
Scatter plot between gdp\_nominal - US$MM and number of models



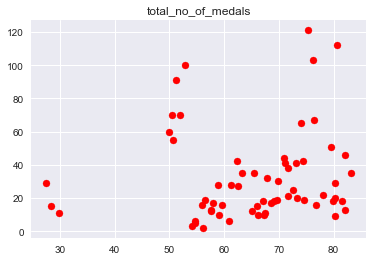
Scatter plot between gdp\_per capita (usd) and number of models



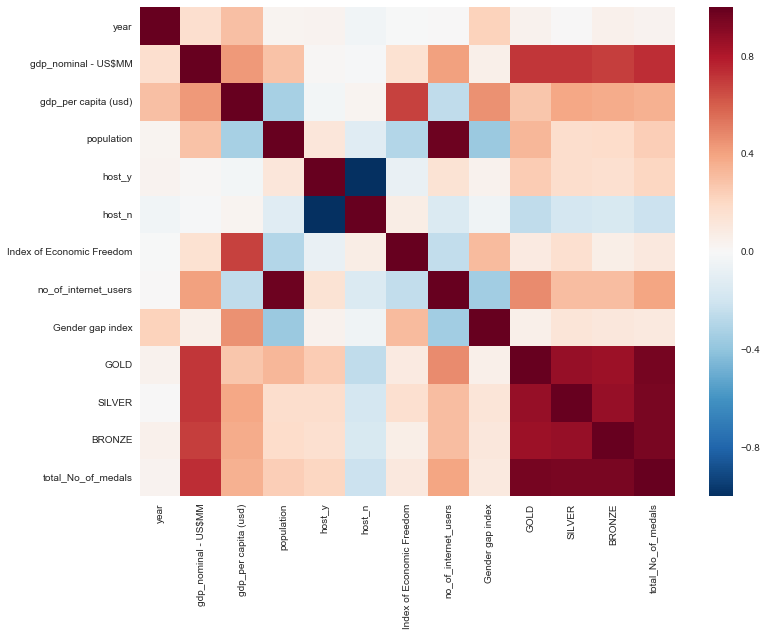
Scatter plot betweenGender gap index and number of models



Scatter plot betweenIndex of Economic Freedom



1. Correlation matrix between each variable.



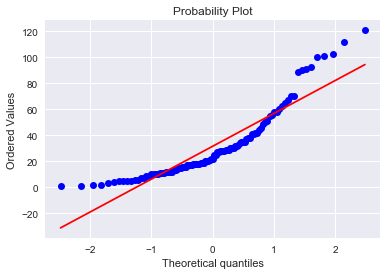
CORRELATION of variables with total medals

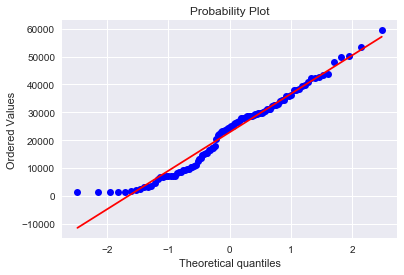
A screenshot of a cell phone

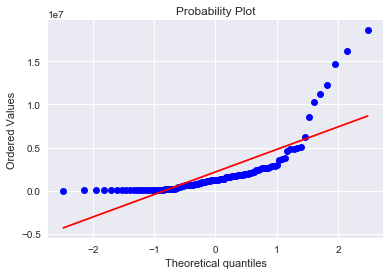
Description generated with very high confidence

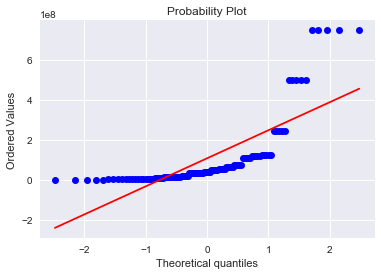
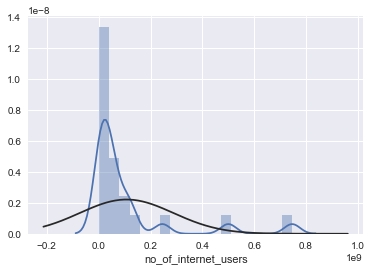
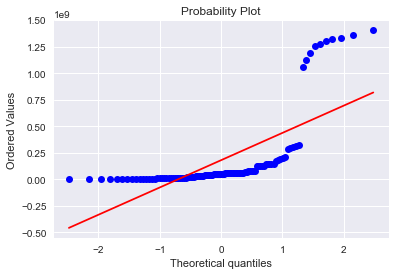
1. histogram and normal probability plot of variables

A close up of a white wall

Description generated with high confidenceA screenshot of a cell phone

Description generated with high confidenceA close up of a white wall

Description generated with high confidenceA close up of a white background

Description generated with high confidence

1. LinearRegression package is imported from sklearn to perform Linear regression. Dependent variables – no\_of\_internet\_users','population','gdp\_nominal - US$MM','gdp\_per capita (usd)

Independent variables – total no of medals.

1. **OLS** - ordinary least squares method in Linear regression is used for regression analysis. Package **statsmodels.formula.api** is used.

1. Backward elimination approach to make optimal model. All the variables which are influential for predicting the target variable are taken and then , the variable which has p-value greater than 0.05 is removed from next model.

A screenshot of a cell phone

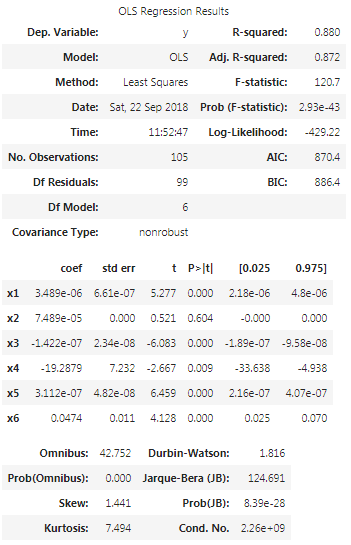
Description generated with very high confidence

1. Second model

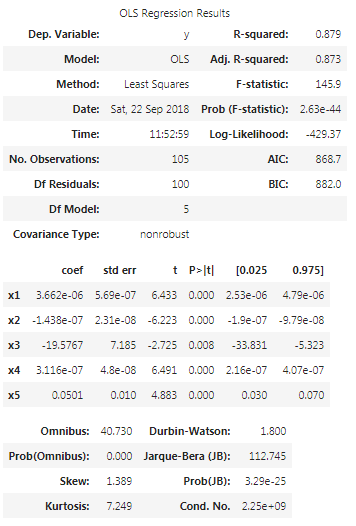
A screenshot of a cell phone

Description generated with very high confidence

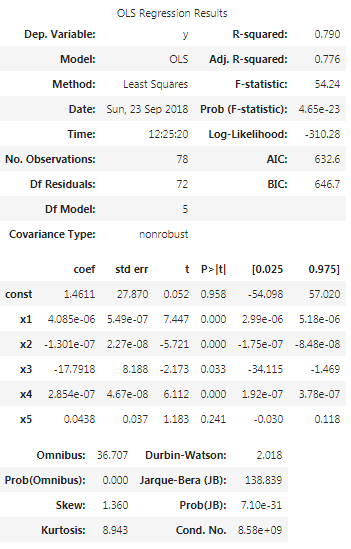
1. Third model



1. Fourth model

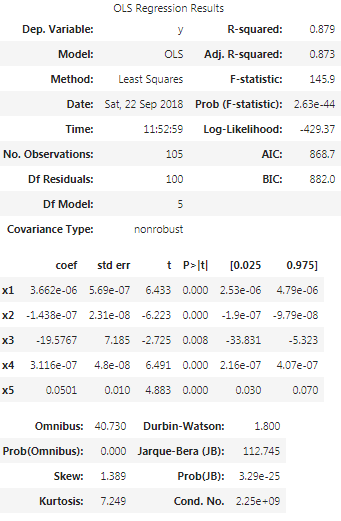


1. Model with constant term



The model has constant term. It shows that intercept is not significant and it is also making other variables insignificant. The intercept can be removed but it is kept to maintain the slope of the line and not forcefully fit it to pass through origin.

Model without intercept term.



As , Adj R-square is 0.873.

1. Split data into train and test set. Test data - 25% of the total data.

Training data - 75% of the total data.

The constant term is added into the model as sklearn comes without constant term.

1. The model is trained on training data and then, it is tested on the testing data. The model is used for prediction on testing data and the values are predicted. The values are :
2. array([54.70423253, 26.22290275, 89.57814271, 3.03950788, 32.75880007,
3. 35.55879045, 14.87948244, 15.3428213 , 11.08542232, 34.36031291,
4. 28.53409227, 23.58545376, 22.9815572 , 25.98829601, 25.21146098,
5. 25.08946282, 54.44232911, 35.93428501, 33.55573691, 25.74742028,
6. 37.59996345, 48.69003561, 32.46063478, 14.32226344, 24.38153615,
7. 26.50533086, 34.68780539])

These values are then compared with the actual values in test data.

array([38, 20, 70, 3, 28, 35, 10, 2, 16, 60, 90, 11, 46, 17, 21, 35, 65,

70, 17, 28, 51, 37, 55, 17, 12, 10, 42], dtype=int64)

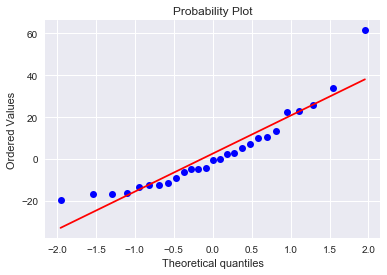
1. Scatter plot between predicted and actual y values.

A picture containing sky, outdoor

Description generated with high confidence

1. Regression Diagnostics are used to verify the model.

QQ Plot of residuals.



Plot of residuals and fitted values - Heterosckadasticity can be shown from this.

A picture containing sky

Description generated with very high confidence

**cook's distance**

A close up of a white wall

Description generated with high confidence

A screenshot of a cell phone

Description generated with high confidence

A screenshot of a cell phone

Description generated with high confidenceA screenshot of a cell phone

Description generated with very high confidence25.

A screenshot of a cell phone

Description generated with very high confidence

1. Root mean squared error - 18.390982933772655
2. Mean squared error, absolute mean squared error,MAPE

mean\_squared\_error(y\_test, y\_pred) - > 338.16267765895765

mean\_absolute\_error(y\_test, y\_pred) -> 13.58217196028744

y = np.mean(np.abs((y\_test - y\_pred) / y\_test)) \* 100 -> 67.33740754292982

1. Line plot

A close up of a device

Description generated with high confidence

1. Scaling using min max scaler function
2. RMSE using it and without it.
3. GBR – gradient Boosting Regressor
4. GradientBoostingRegressor(alpha=0.9, criterion='friedman\_mse', init=None,
5. learning\_rate=0.1, loss='ls', max\_depth=3, max\_features=None,
6. max\_leaf\_nodes=None, min\_impurity\_split=1e-07,
7. min\_samples\_leaf=1, min\_samples\_split=2,
8. min\_weight\_fraction\_leaf=0.0, n\_estimators=100,
9. presort='auto', random\_state=None, subsample=1.0, verbose=0,
10. warm\_start=False)

Mean Average Error - 11.553307096107396

rmse\_GB - 15.870378513251822

Predicted values - array([30.37322921, 23.20038308, 93.35850247, -0.32727455, 28.38373415,

36.81914088, 6.95268583, 12.04646516, 9.58618783, 36.36280138,

38.12632426, 20.01109248, 40.01432684, 19.53626986, 31.84835735,

23.44301863, 39.89373027, 42.54087747, 25.34346499, 38.51444573,

36.81914088, 29.06502374, 36.36280138, 13.24408643, 23.63657343,

18.2986027 , 40.60078111])

A picture containing sky

Description generated with high confidence

1. Random Forest Regressor

mean\_squared\_error(y\_test, pred) - > 380.0618518518518

mae -> 14.381481481481481

rmse\_RF -> 19.495175091592582

A close up of a device

Description generated with high confidence

1. KNeighborsRegressor

mean\_squared\_error(y\_test, pred) -> 100.38888888888889

mae -> 7.444444444444445

##residuals - fitted values plot

A picture containing sky

Description generated with very high confidence

A screenshot of a cell phone

Description generated with high confidence

A close up of a device

Description generated with high confidence

1. XGBOOST Regressor
2. XGBRegressor(base\_score=0.5, booster='gbtree', colsample\_bylevel=1,
3. colsample\_bytree=1, gamma=0, learning\_rate=0.1, max\_delta\_step=0,
4. max\_depth=3, min\_child\_weight=1, missing=None, n\_estimators=100,
5. n\_jobs=1, nthread=None, objective='reg:linear', random\_state=0,
6. reg\_alpha=0, reg\_lambda=1, scale\_pos\_weight=1, seed=None,
7. silent=True, subsample=1)

array([10.101154 , 9.565032 , 24.341003 , 1.1759995, 9.648361 ,

12.906151 , 2.3163729, 3.250154 , 3.929835 , 18.69073 ,

19.951208 , 7.0283065, 9.843973 , 6.037368 , 9.004144 ,

8.219137 , 12.1687975, 18.357227 , 7.7765718, 13.613796 ,

13.114792 , 11.923305 , 17.799566 , 3.0716302, 11.236645 ,

7.364332 , 11.404296 ], dtype=float32)

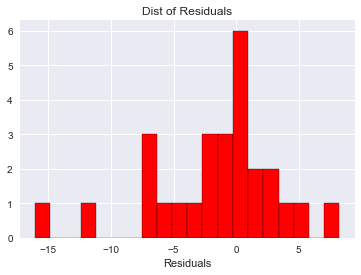
Plot of residuals vs fitted values

A picture containing sky, outdoor, tree

Description generated with high confidence

A close up of a device

Description generated with high confidence



Mae - > 3.6826813176826194

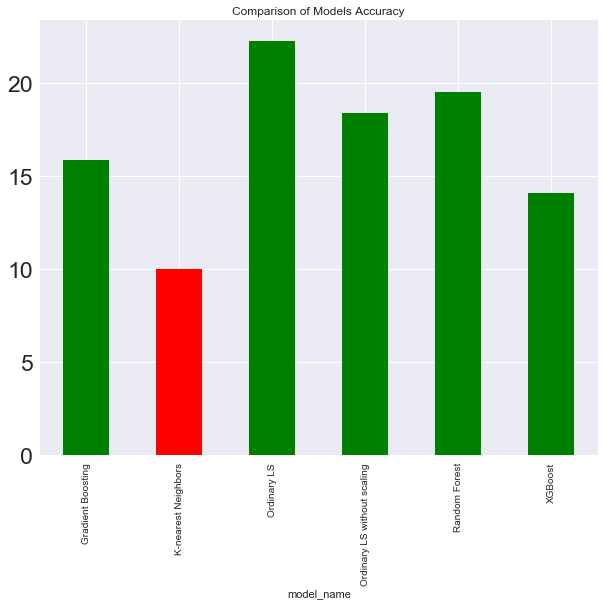
rmse\_xgb -> 5.307733160990386

Features importance plot using XG boost model.

A screenshot of a cell phone

Description generated with very high confidence

1. Comparison of Models Accuracy – RMSE of all the different models has been taken.



1. Applying Grid search method #Optimization techniques. The method has been applied on KNN Regressor model. The parameters are chosen and are used for further analysis.

Best Parameters -> {'algorithm': 'auto', 'leaf\_size': 20, 'n\_jobs': -1, 'n\_neighbors': 2, 'p': 1}

best\_accuracy -> 0.6721860613751406

1. K- fold CV

Here, 10 folds have been used to improve the score of the model.

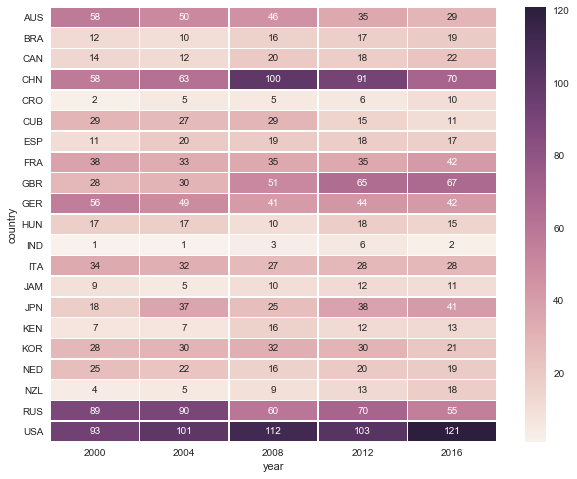
efficiency = cross\_val\_score(estimator = c, X= X\_train, y = y\_train, cv = 10)

efficiency - array([ 0.90308735, 0.40965325, 0.9327203 , 0.569522 , 0.9600856 ,

-0.12679426, 0.9458242 , 0.91388119, 0.4723171 , 0.71271263])

**efficiency.mean()** -> 0.6693009346625299

1. Heatmap of medals and country with year



1. Prediction on 2020 independent variables data - The test data - test\_data\_2020\_olympics is used for prediction of the medals in 2020 Olympics. KNN regressor model is used for prediction as it has lowest RMSE value. Missing value in variable gdp\_nominal - US$MM is imputed with median value.

Prediction of the total medals using same model.

array([112. , 48.5, 95.5, 65. , 43. , 21.5, 51. , 31. , 35.5,

39.5, 19.5, 16.5, 17.5, 29. , 31. , 11.5, 7. , 21. ,

15.5, 20. , 4. ])

Prediction of the gold medals using same model.

array([ 7. , 2.5, 43. , 0. , 11. , 11. , 1.5, 0.5, 1.5, 22. , 22. ,

4. , 16.5, 4.5, 13. , 12.5, 18. , 22. , 5. , 9. , 18. , 7. ,

22. , 5.5, 2.5, 1.5, 11. ])

array([32.5, 16. , 26.5, 18. , 15. , 7.5, 17. , 10. , 11.5, 13.5, 6.5,

3.5, 5. , 11. , 10. , 4. , 2. , 5. , 5.5, 4. , 1.5])

Prediction of the bronze medals using same model.

array([33.5, 14.5, 26. , 25. , 14. , 7. , 15. , 8. , 12. , 13.5, 6. ,

5. , 7.5, 9.5, 8. , 2.5, 2. , 9. , 5. , 13. , 2.5])

Rank of countries

1. Total medals

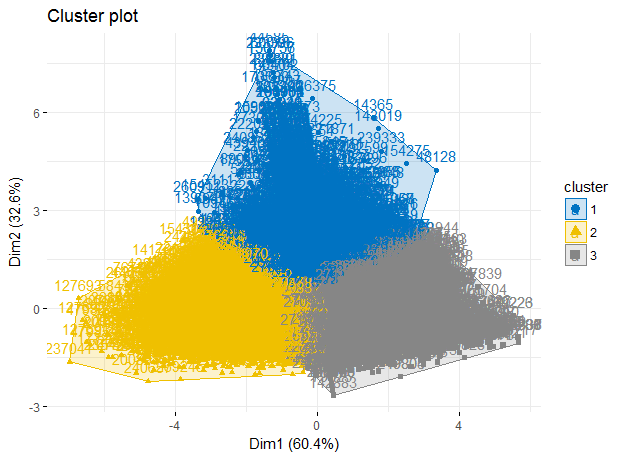
**Analysis using K-means Clustering:**

K- means clustering has been done using different set of variables.The primary dataset given has been used for this. The data was cleaned and then, only the players who have won medals are use in clustering.

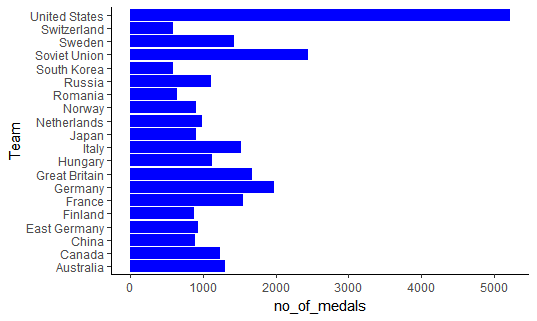
Variables use for clustering - Age, Height, Weight.

1. This visual shows the loadings of Age, Height, Weight on the 3 new components(dimensions). PCA is performed in this. Height, Weight has high loadings on component 1. Age has high loading on Component 2 and Height, Weight has high loadings on Comp 3.
2. > c
3. Loadings:
4. Comp.1 Comp.2 Comp.3
5. Age -0.143 0.989
6. Height -0.698 -0.134 -0.704
7. Weight -0.702 0.709

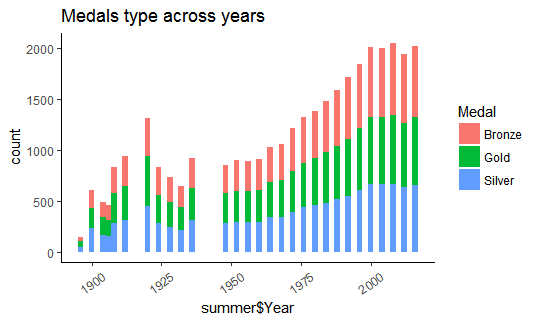
2. The three clusters have been plotted on 2- D chart, showing the observations on two best principal components.



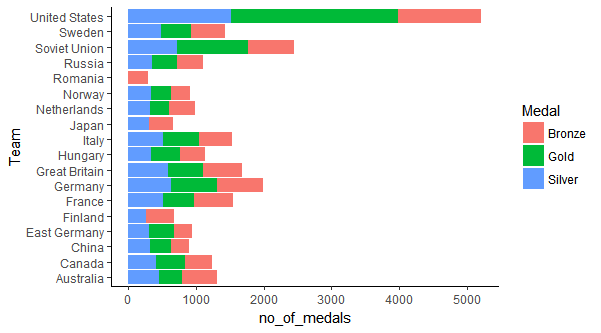
3. This shows US has won most medals. This is the chart of top 20 countries.



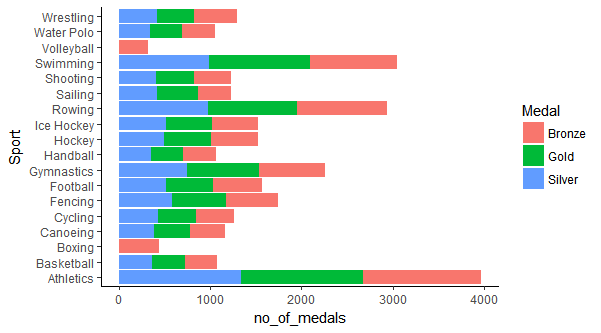
4. Count of medals is increasing from 1900 till 2016. The medals count is almost constant after 2000.



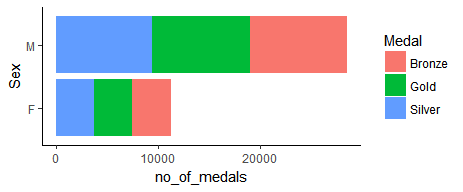
5. Chart shows bar plot of medals by country and medals also used as colours. Count of Gold medals is more than silver, bronze for USA.



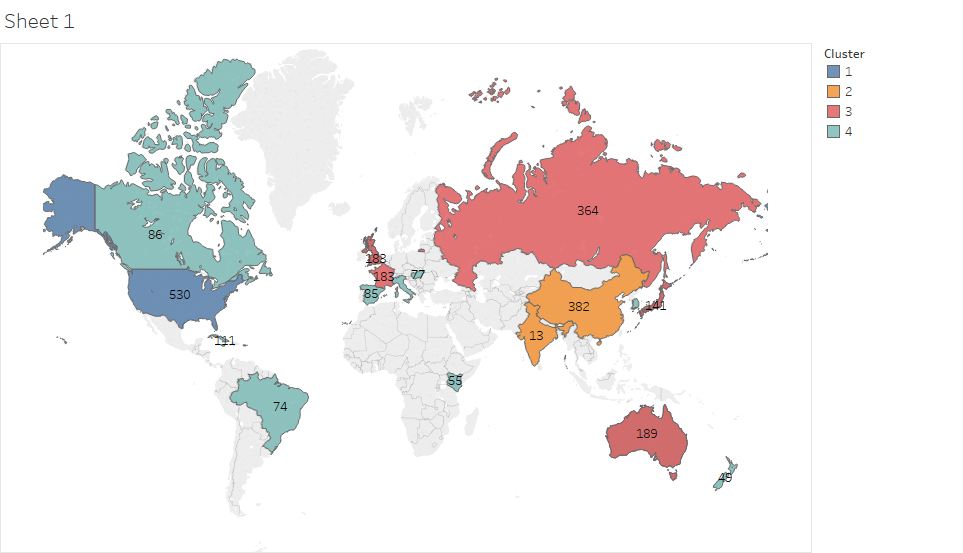
6. Athletics, Rowing and Swimming are the top 3 sports. These sports are being part of Olympics since its start. This chart shows the top 18 sports as per the count of medals in Olympics.



7. This chart shows that Males are more dominant in getting medals than females since the start of olympics till now.



8.



This chart shows the top 20 players according to the count of medals till now in the history of Olympics. Mostly the players are form US, Soviet Union, Japan, Germany. Michael Phelps is on the top because of the medals he won in swimming.

A close up of text on a black background

Description generated with very high confidence

Top 10 Male athletes according to the number of medals, mostly from USA.

A screenshot of a cell phone

Description generated with very high confidence

Top 10 female athletes according to the number of medals, mostly from Germany, US and soviet

A screenshot of a social media post with text and a blue sky

Description generated with high confidence

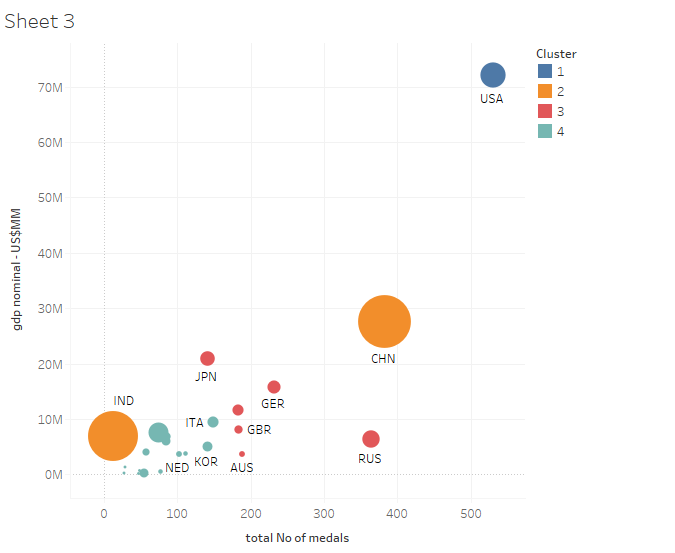
The K means clustering is performed on GDP, population and Total medals variables. The dataset is the same that is used for regression analysis. Only years after – 2000 are considered and only top 20 ranked countries in 2016 summer olympics are taken Optimum no of clusters are 4 and most of the records are in cluster 4.

1. First cluster shows that it has highest number of bronze medals.Countries have highest GDP per capita
2. Second cluster has least Average GDP per capita, least number of silver and gold medals. Average population is very high and it consist of high populated countries.
3. 3rd cluster has highest number of silver medals. Number of gold medals are also on the higher side. This cluster consist of US, china and top performing nations, teams
4. 4th cluster shows that avg no of internet users are very less in this cluster. Most of the records are in this cluster but the no of medals are lesser as relative to other cluster.This cluster does not consist of top performing nations.

A screenshot of a cell phone

Description generated with very high confidence

This chart is scatterplot, it shows relation between Nominal GDP and total count of medals for the same cluster as discussed just above. Cluster 1 contain only USA, Cluster 2 contain only China and India. Cluster 3 contain Russia, Germany , japan, Australia, GBR, Countries ranked from 10 to 20 in last summer Olympics come in 4th cluster. It clearly shows that both variables has some correlation with each other. USA is acting like an outlier in this chart.



The K- means clustering has been done for Winter and Summer olympics both . The optimum no of clusters are 3.

1. First cluster shows the average age is 35 and weight is 71. This cluster consist of experienced players and mature players who have won medals.
2. Second cluster shows the average age is 25 and weight is 88. This cluster consist of mixed players and it shows this the most suited combination of height, weight and age , which will secure more medals. Their height is also better relative to other clusters.
3. 3rd cluster shows the average age is 23 and weight is 66. This cluster consist of younger players and might contain specific sports events , in which weight is important for players who have won medals. Their weight is very less as compared to other clusters

A screenshot of a cell phone

Description generated with very high confidence

The treemap chart shows the genetic characteristics of the players in India who have won medals,

In hockey, weight is very less, and age is on higher or lower side. Some players have tall and have high weight.

In Badminton, weight is on the lowers side and age is also less. Youngsters can get more medals. Perfect age is 22 yrs.

In wrestling, age is 29 or 26 yrs, it proves that experienced persons are better in this sport and can get more medals.

Colour shows 3 different clusters . Most of players come in category of 3rd cluster who have won medals, Most of the hockey players come in 1st and 2nd cluster.

A screenshot of a cell phone

Description generated with very high confidence

Different clusters divided by two gender categories.

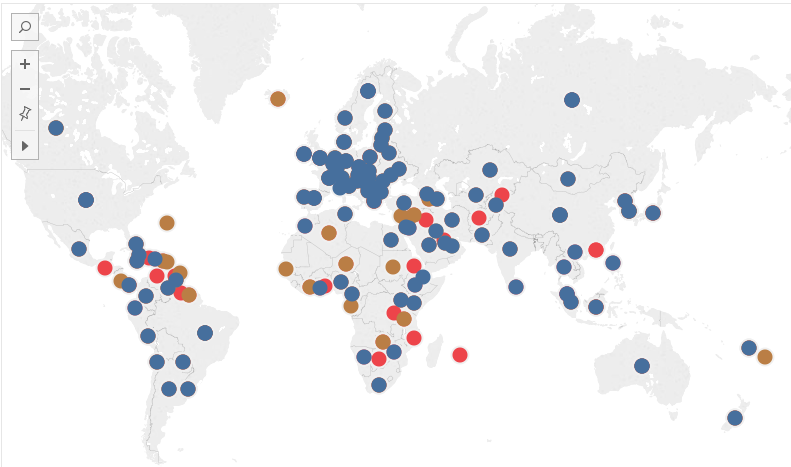
Most females come in 3rd cluster, it shows that the females who won medals in olympics should have less age.

A screenshot of a cell phone

Description generated with very high confidenceA screenshot of a cell phone

Description generated with high confidence

For summer olympics, this geo map chart shows the cities from where, the players belong and colour shows different clusters. Players from cluster 1 are spread across world. Players from cluster 2 are from Africa, middle east and asia regions and carribean countries.

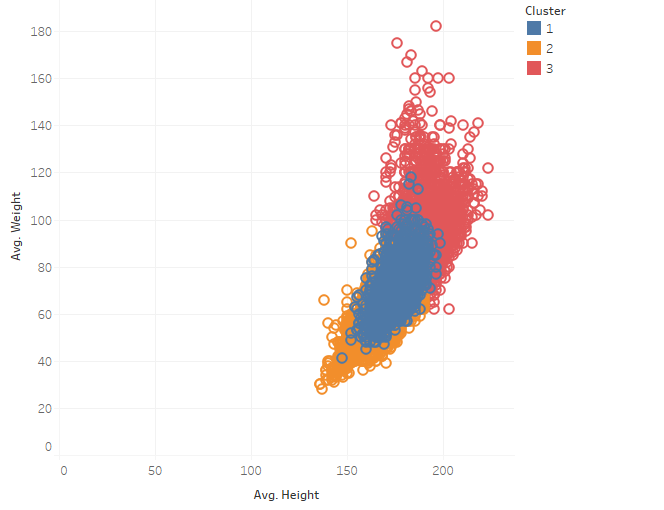


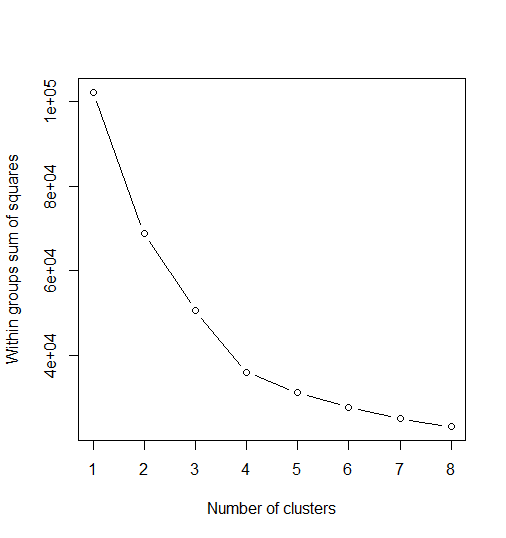
This Bar plot of all sports consist of players who have won medals, only from China country. China has huge medals in Diving, Gymnastics, Shooting, Skating, Table tennis, swimming, weightlifting and volleyball. These sports will be popular in sports schools in china and most of the players are from cluster 3 in gymnastics, diving, skating, weightlifting, badminton. This shows that all these sports need young players. This can be a lesson for Indian athletes to improve in these sports because India has lot of young sportspersons.

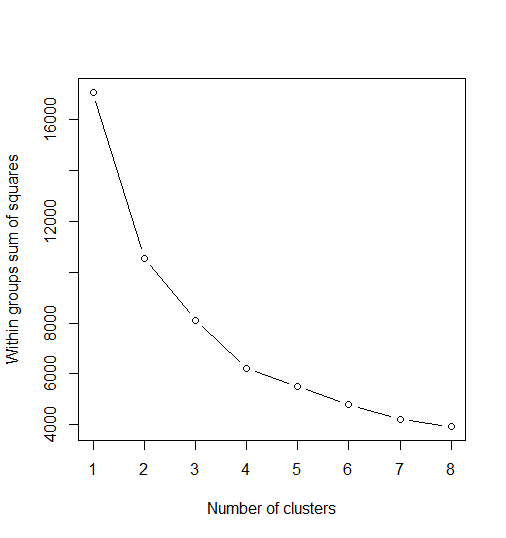
A picture containing object

Description generated with high confidence

This is the scatter plot between avg weight and avg height, The clusters are shown by colours and it can be seen clearly from here that players in cluster 3 has weight, height on higher side and cluster 2 are having less average weight and height. Players in cluster 1 comes in middle







The K- means clustering has been done for Summer olympics. The optimum no of cluster are 3.

1. First cluster shows the average age is 35 and weight is 73. This cluster consist of experienced players and mature players who have won medals.
2. Second cluster shows the average age is 23 and weight is 64. This cluster consist of young players and it proves that lesser age and weight is the factor in getting more medals. This cluster has highest number of players.
3. 3rd cluster shows the average age is 25 and weight is 90. This cluster consist of mixed players and might contain specific sports events , in which weight is important for players who have won medals. Their Average height is also highest , might consist of swimming players.

A screenshot of a cell phone

Description generated with very high confidence

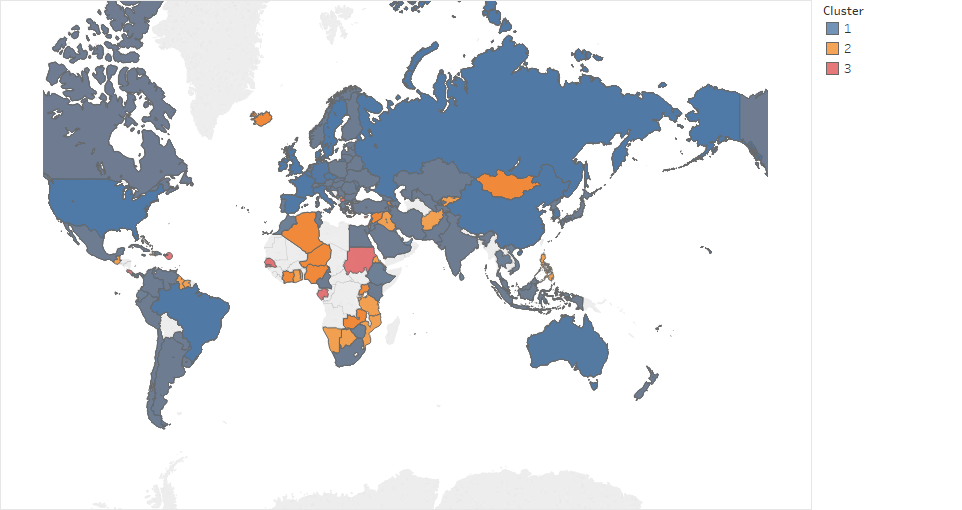
The K- means clustering has been done for Winter olympics. The optimum no of cluster are 3.

1. First cluster shows the average age is 31 and weight is 85. This cluster consist of experienced players and mature players who have won medals.
2. Second cluster shows the average age is 24 and weight is 75. This cluster consist of mixed players and it shows this the most suited combination of height, weight and age , which will secure more medals.
3. 3rd cluster shows the average age is 25 and weight is 60. This cluster consist of mixed players and might contain specific sports events , in which weight is important for players who have won medals. Their weight is very less as compared to other clusters

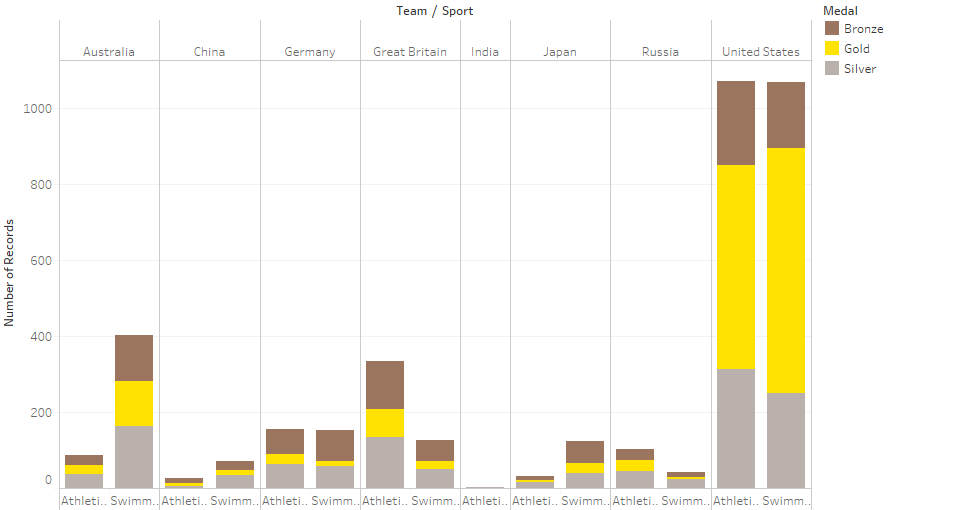
A screenshot of a cell phone

Description generated with very high confidence

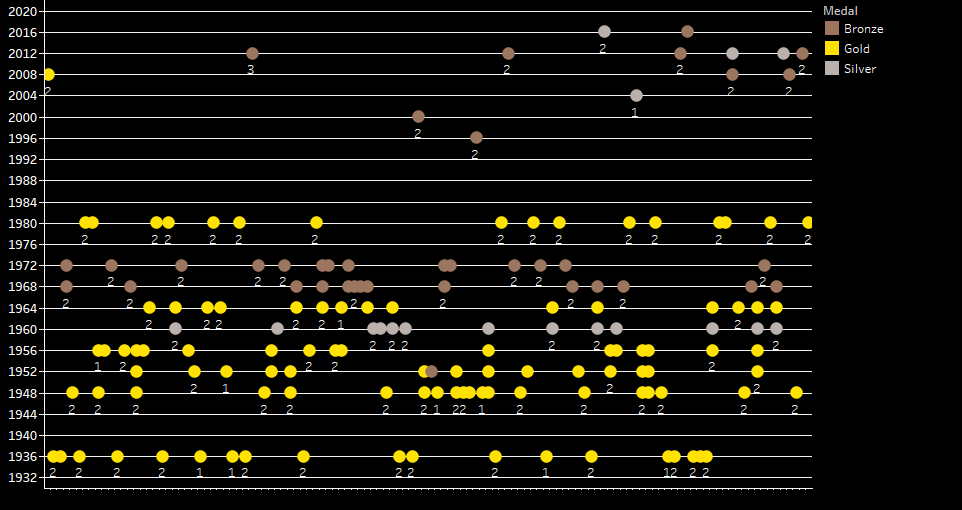
This geo- map chart shows different countries coloured by clusters of age, height and weight variables. It shows that cluster 2 – which has highest no of records and youngest players are mostly from Arab nations, African countries. Players in Cluster 1 are mostly from US, and big economy countries.



This chart shows the comparison only for Athletics and Swimming events in summer olympics cluster. This shows that USA and Aussies have highest number of medals in these 2 sports and it shows that the players genetic conditions are similar in these two countries. Great Britain and Germany also has good number of players in this.



This chart compares the medals tally for summer olympics only for Indian team, it shows that Hockey was primary sport of India in earlier days but now, trend is changing to wrestling, badminton, boxing and shooting. Most of the players were younger or older in terms of age as most of the players are from cluster 1 and 2. In 2012 only, one player from cluster 3 won the bronze medal.



This visual shows the medals in each sport in olympics for best performing countries including India. It shows that India has got all 3 types of medals only in Hockey, shooting. In wrestling, badminton , players and sports is rising in India. Badminton and Hockey are the only sports which India is at par with other nations.



Modelling, evaluation:

1. Naïve Bayes model was selected for text classification and sentiment classification of tweets.
2. Maximum entropy model is chosen sentiment analysis.
3. Confusion matrix is used for evaluation.

Business Recommendation:

* Forecasting model for medals can be made using Neural Network.
* The OPTIMIZATION TECHNIQUES can be used for optimizing the model.

Assumptions, Limitations:

1. The data was collected for only 5 years from 2000 to 2016.
2. The data used for regression was small and the sample contains only 20 countries data.
3. Not able to collect data related to Sports centres, sports funding for countries.
4. The test data for year 2020 was not available for all fields. So, data for 2018 year is considered.
5. The features in dataset are not normal.
6. Assumptions of Multiple Regression are used in modelling

References:

* <https://en.wikipedia.org/wiki/Olympic_Games>
* <https://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users>
* <https://data.worldbank.org/indicator/ny.gdp.pcap.pp.cd>
* <https://www.heritage.org/index/ranking>