Project Proposal
Sport Stat – Olympics Data

Audhi Aprilliant

audhiaprilliant.github.io

DESCRIPTION

- As a scientist in the Sport field, we are asked to do deep analysis which will be focused through data visualization
- The analysis will find out any trends in Olympics games such as the country that dominating in certain sports for 120 years and qualitative analysis to answer the question why this phenomenon happens
- Further, our journey steps to the side of statistics with regression problem to estimate the missing value in certain variables
- It's interesting to find out specific, rightful, and useful regression method to handle this problem

ASSUMPTIONS

The scopes of research are listed:

FIRST

Each rows is the unique person in difference time of period

SECOND

The chosen columns must have high correlation

QUESTIONS

To ensure the effectiveness of the research, the following question will be answered systematically:

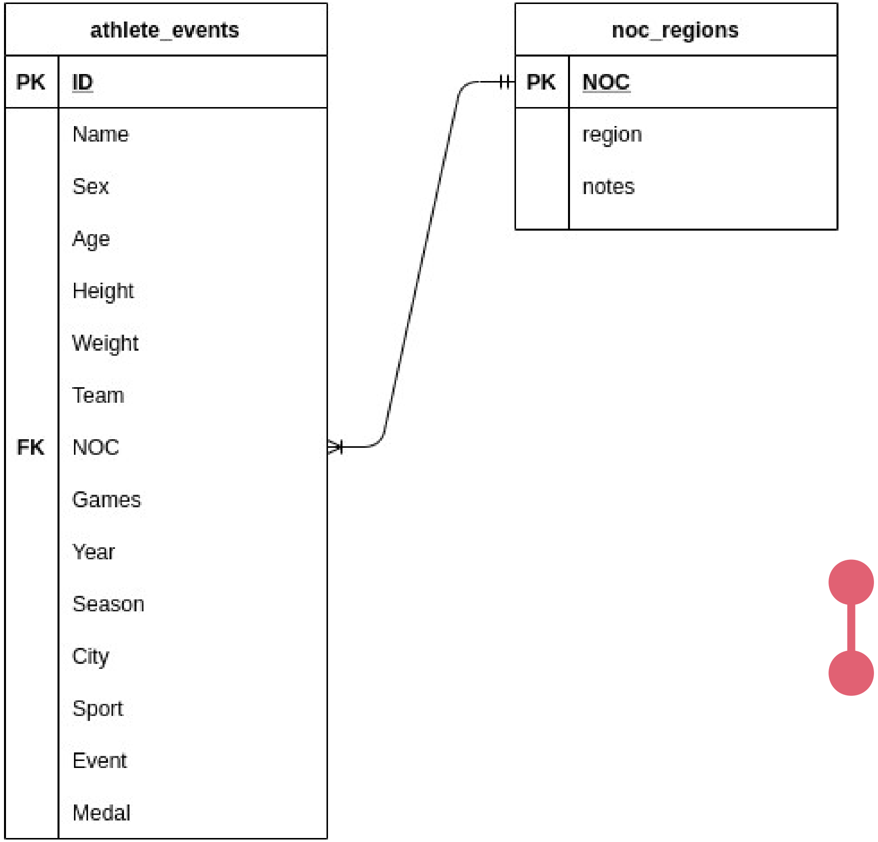
- How does data pre-processing will be developed?
- What kind of methods is more useful to handle and fill missing value in the certain columns?
- Are there unique trends visually in the data?

APPROACH

To ensure the effectiveness of the research, the following question will be answered systematically:

- Several columns will be dropped because it has no enough correlation with the main analysis
- To handle missing value, the linear regression and decision tree regressor will be compared
- Evaluation metrics use root mean square error, mean absolute error, and Pearson correlation
- Explanatory Data Analysis (EDA) is rightful method and mostly used to find out pattern in the whole data

ERD



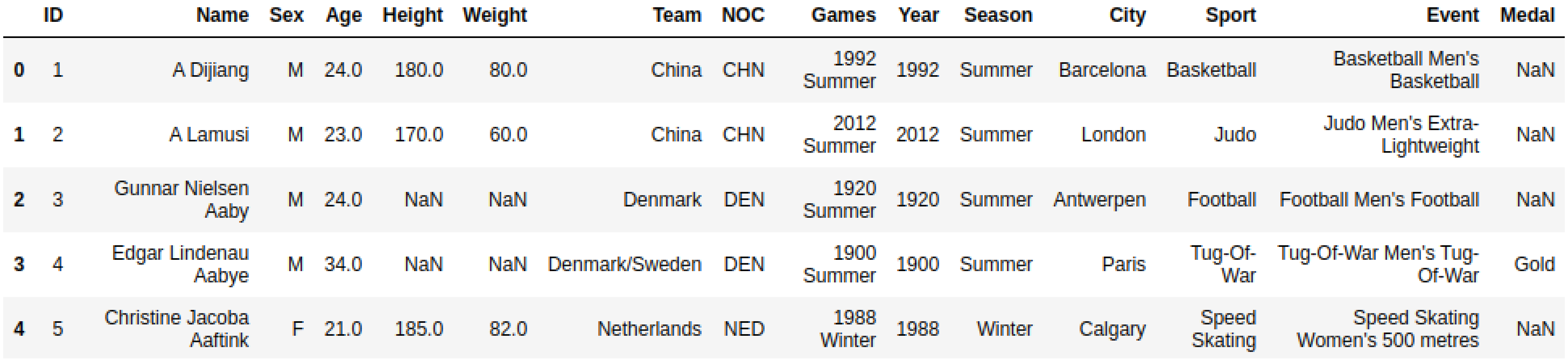
athlete_events	
PK	ID
	Name
	Sex
	Age
	Height
	Weight
	Team
FK	NOC
	Games
	Year
	Season
	City
	Sport
	Event
	Medal

noc_regions	
PK	NOC
	region
	notes

Two tables are included. But for the main analysis, we only need the athlete event table

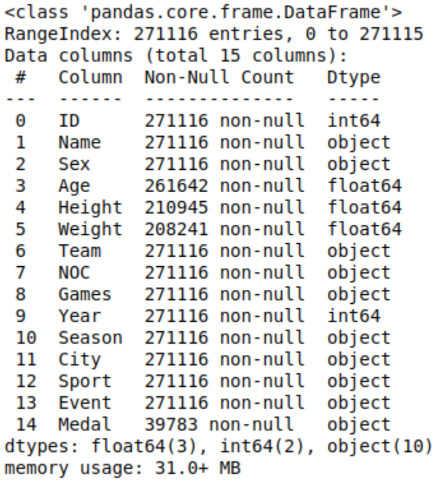
- Height
- Weight

EXPLORATION



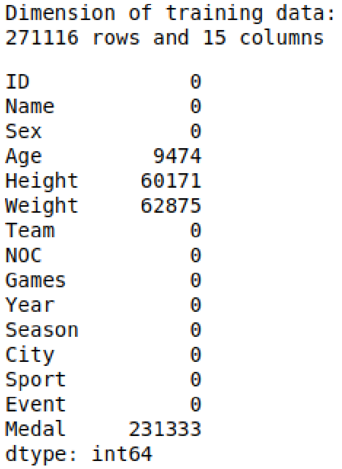
	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

Description of column type



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 15 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   ID      271116 non-null  int64
 1   Name    271116 non-null  object
 2   Sex     271116 non-null  object
 3   Age     261642 non-null  float64
 4   Height  210945 non-null  float64
 5   Weight  208241 non-null  float64
 6   Team    271116 non-null  object
 7   NOC     271116 non-null  object
 8   Games   271116 non-null  object
 9   Year    271116 non-null  int64
 10  Season  271116 non-null  object
 11  City    271116 non-null  object
 12  Sport   271116 non-null  object
 13  Event   271116 non-null  object
 14  Medal   39783 non-null   object
dtypes: float64(3), int64(2), object(10)
memory usage: 31.0+ MB
```

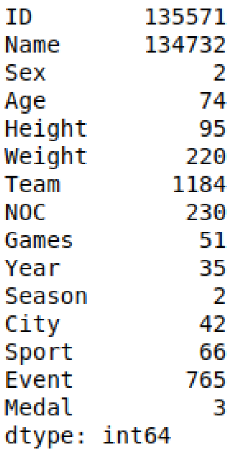
Missing value each columns



```
Dimension of training data:
271116 rows and 15 columns

ID             0
Name           0
Sex            0
Age         9474
Height     60171
Weight     62875
Team           0
NOC            0
Games          0
Year           0
Season         0
City           0
Sport          0
Event          0
Medal     231333
dtype: int64
```

Unique value each columns

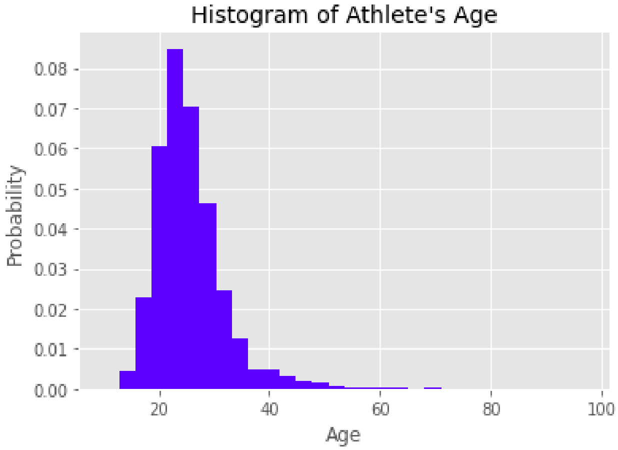


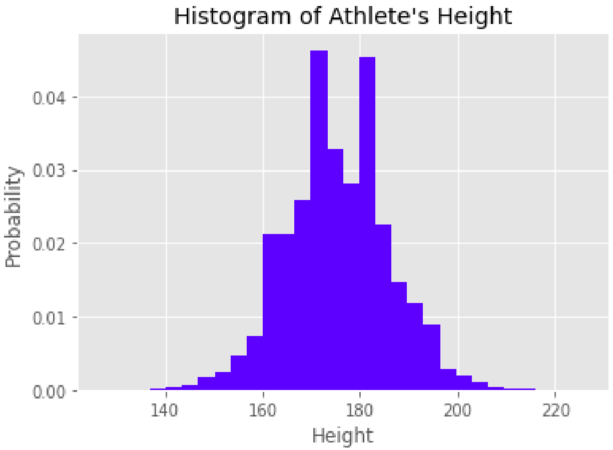
```
ID        135571
Name      134732
Sex            2
Age           74
Height        95
Weight       220
Team        1184
NOC          230
Games         51
Year          35
Season         2
City          42
Sport         66
Event        765
Medal          3
dtype: int64
```

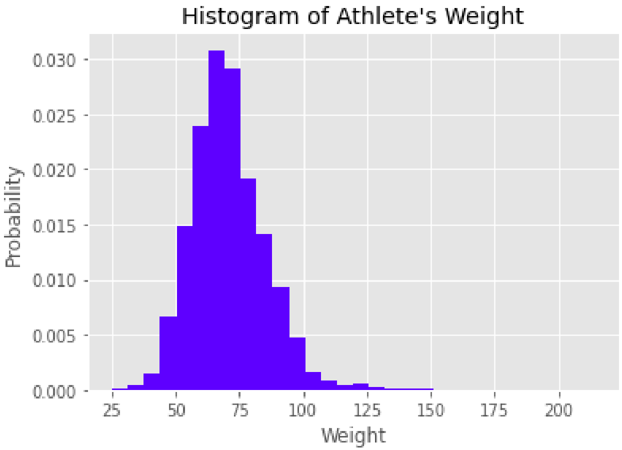
Findings

Three important variables for deep analysis need manipulation. These are athlete's age, weight, and height. So, it needs to find out the best method to fill those missing value properly

EXPLORATION





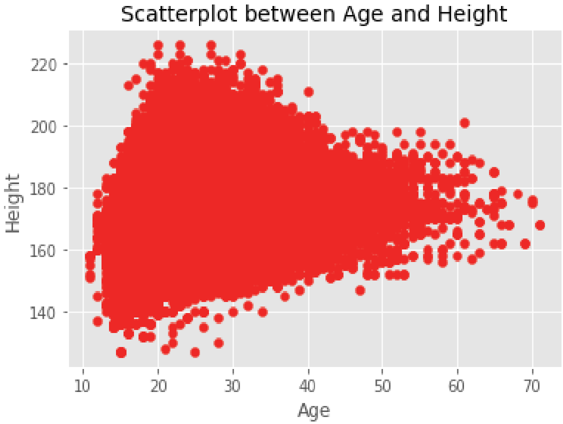


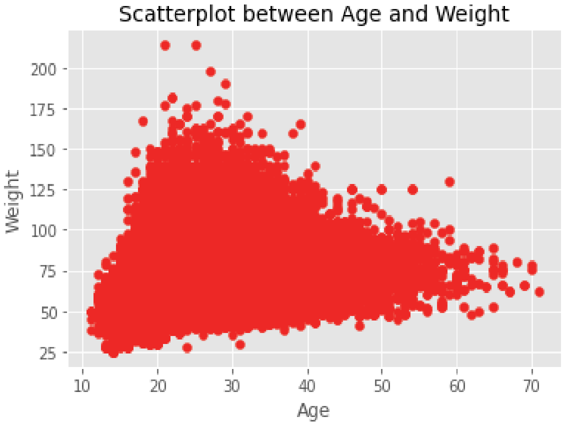
Findings

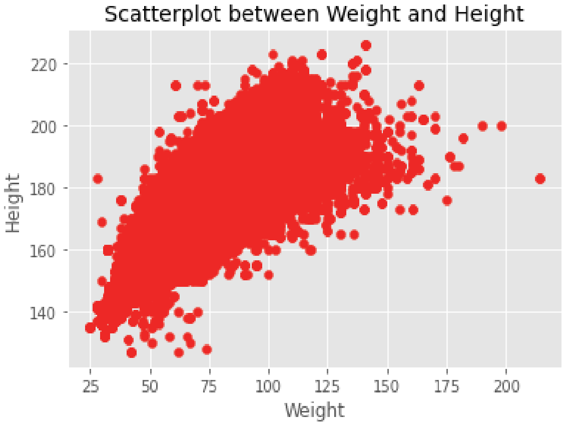
According to those histogram, athlete’s age and weight are right-skewed while the height is bell-shape, Normal distribution.

- The average of athlete’s age is about 97 yo. It’s unnatural. So, we need to do pre-processing
- The maximum of athlete’s weight is about 214 kg. This is why the histogram would be right-skewed

EXPLORATION


Scatterplot between Age and Height

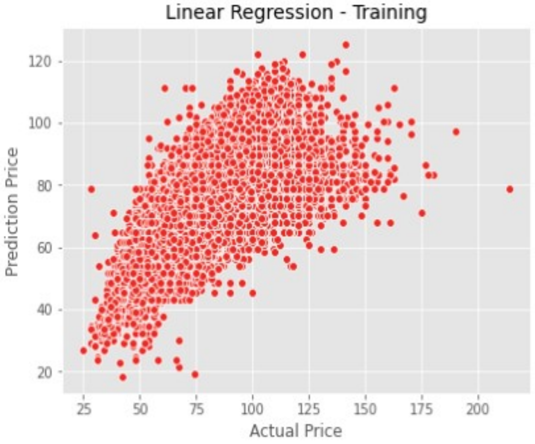

Scatterplot between Age and Weight

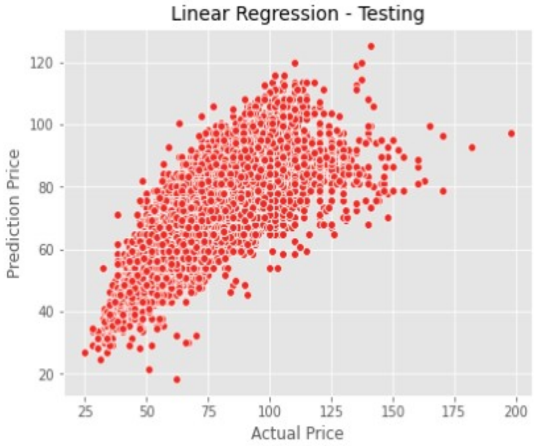

Scatterplot between Weight and Height

Findings

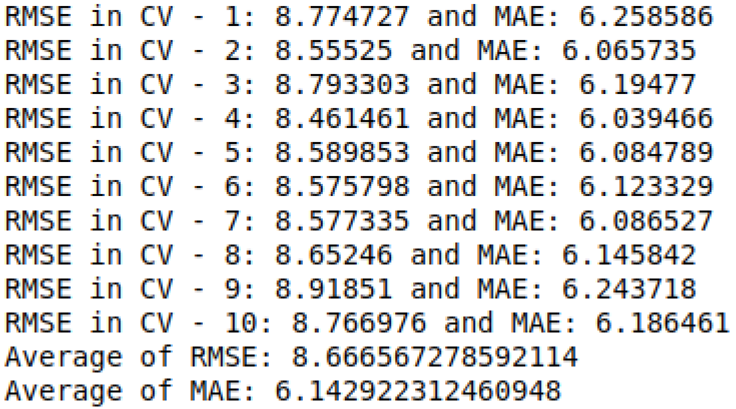
The correlation is the statistic indicating the relationship between two variables in the data. After exploring the numerical variables, the correlation between athlete’s weight and height is **high**.

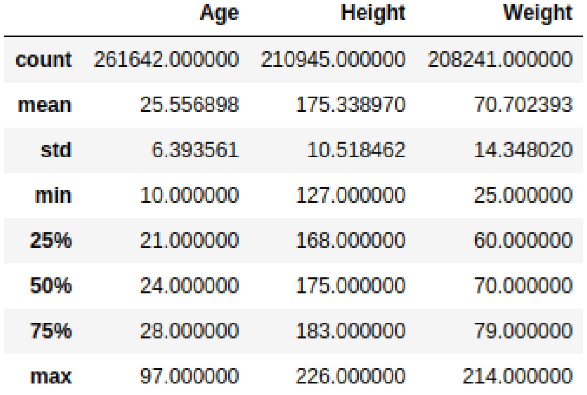
PRE-PROCESSING


Linear Regression - Training


Linear Regression - Testing

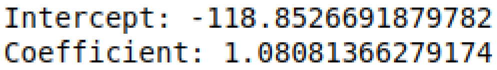
10 - Cross validation


```
RMSE in CV - 1: 8.774727 and MAE: 6.258586
RMSE in CV - 2: 8.55525 and MAE: 6.065735
RMSE in CV - 3: 8.793303 and MAE: 6.19477
RMSE in CV - 4: 8.461461 and MAE: 6.039466
RMSE in CV - 5: 8.589853 and MAE: 6.084789
RMSE in CV - 6: 8.575798 and MAE: 6.123329
RMSE in CV - 7: 8.577335 and MAE: 6.086527
RMSE in CV - 8: 8.65246 and MAE: 6.145842
RMSE in CV - 9: 8.91851 and MAE: 6.243718
RMSE in CV - 10: 8.766976 and MAE: 6.186461
Average of RMSE: 8.666567278592114
Average of MAE: 6.142922312460948
```


	Age	Height	Weight
count	261642.000000	210945.000000	208241.000000
mean	25.556898	175.338970	70.702393
std	6.393561	10.518462	14.348020
min	10.000000	127.000000	25.000000
25%	21.000000	168.000000	60.000000
50%	24.000000	175.000000	70.000000
75%	28.000000	183.000000	79.000000
max	97.000000	226.000000	214.000000

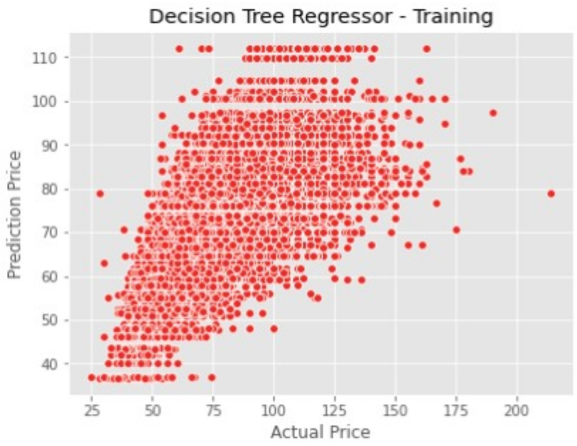
Findings

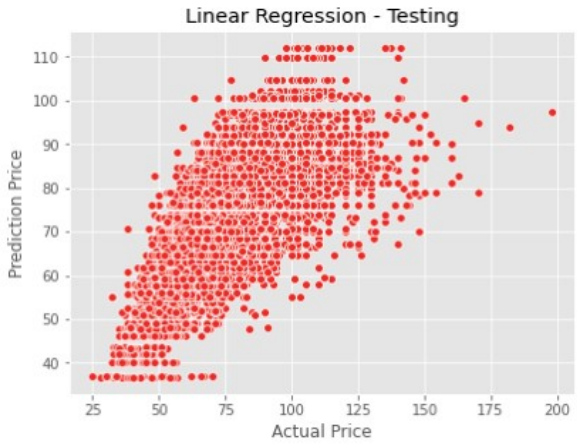
The RMSE of prediction is about 8.66 where it is comparable with the standard deviation of response variable. So, the linear regression model is quite good. The model equation is:

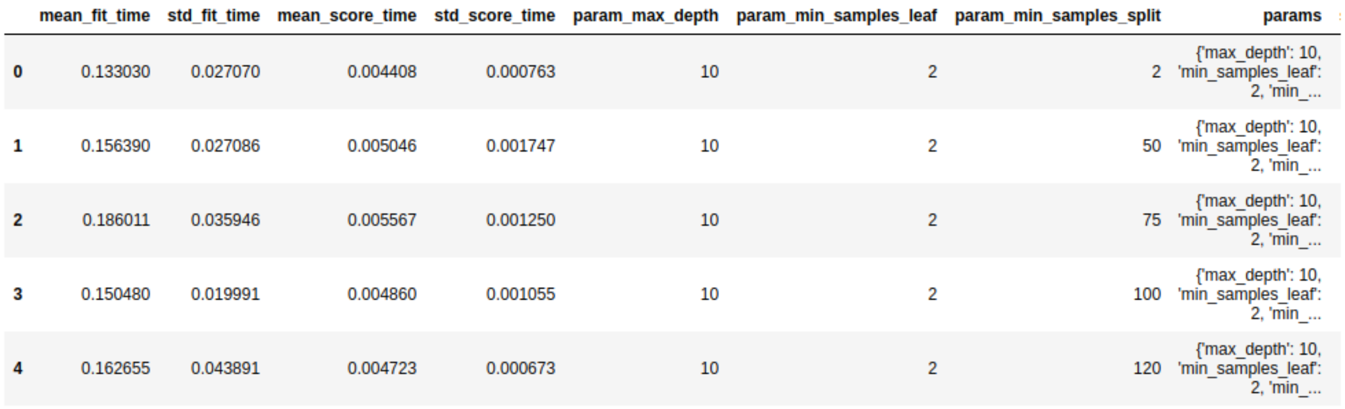

```
Intercept: -118.8526691879782
Coefficient: 1.08081366279174
```

Decision Tree Regressor ●

PRE-PROCESSING

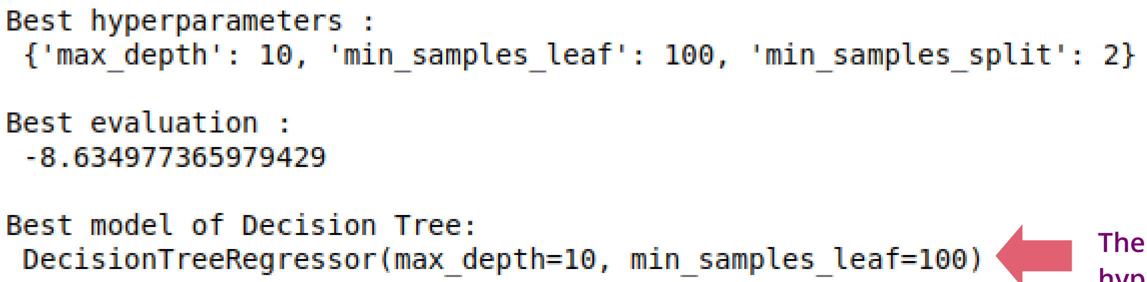






	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_max_depth	param_min_samples_leaf	param_min_samples_split	params
0	0.133030	0.027070	0.004408	0.000763	10	2	2	{'max_depth': 10, 'min_samples_leaf': 2, 'min_...
1	0.156390	0.027086	0.005046	0.001747	10	2	50	{'max_depth': 10, 'min_samples_leaf': 2, 'min_...
2	0.186011	0.035946	0.005567	0.001250	10	2	75	{'max_depth': 10, 'min_samples_leaf': 2, 'min_...
3	0.150480	0.019991	0.004860	0.001055	10	2	100	{'max_depth': 10, 'min_samples_leaf': 2, 'min_...
4	0.162655	0.043891	0.004723	0.000673	10	2	120	{'max_depth': 10, 'min_samples_leaf': 2, 'min_...

Grid-search to get optimum hyper parameters



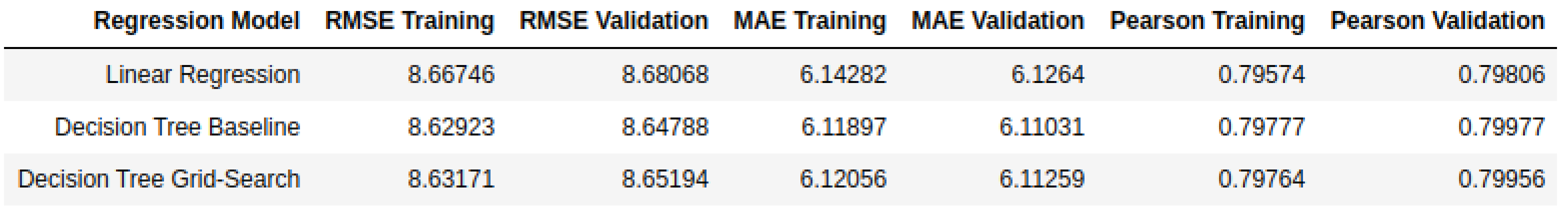
```
Best hyperparameters :
 {'max_depth': 10, 'min_samples_leaf': 100, 'min_samples_split': 2}

Best evaluation :
 -8.634977365979429

Best model of Decision Tree:
 DecisionTreeRegressor(max_depth=10, min_samples_leaf=100)
```

The chosen hyper parameters

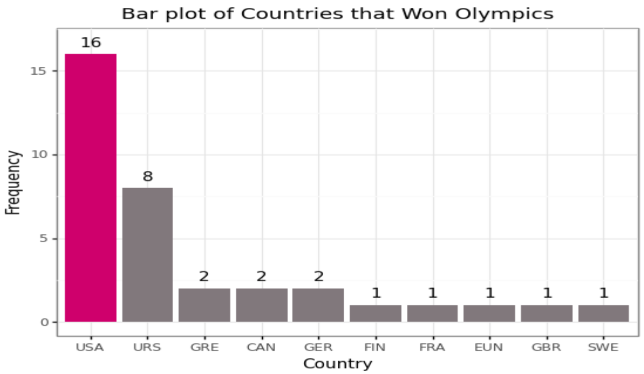
PRE-PROCESSING

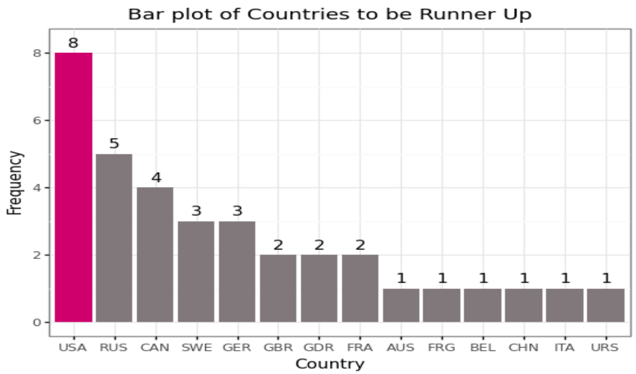


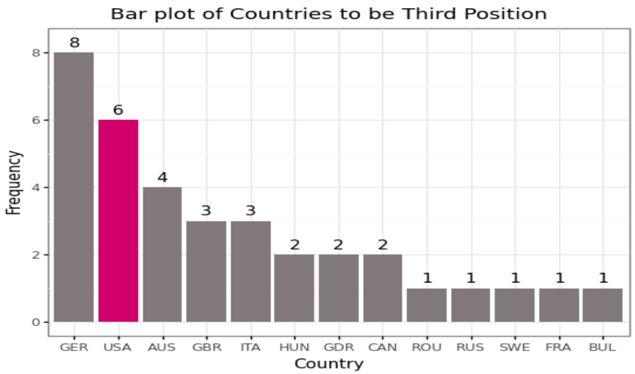
Regression Model	RMSE Training	RMSE Validation	MAE Training	MAE Validation	Pearson Training	Pearson Validation
Linear Regression	8.66746	8.68068	6.14282	6.1264	0.79574	0.79806
Decision Tree Baseline	8.62923	8.64788	6.11897	6.11031	0.79777	0.79977
Decision Tree Grid-Search	8.63171	8.65194	6.12056	6.11259	0.79764	0.79956

Linear regression is chosen because of its simplicity

DATA ANALYSIS


Bar plot of Countries that Won Olympics

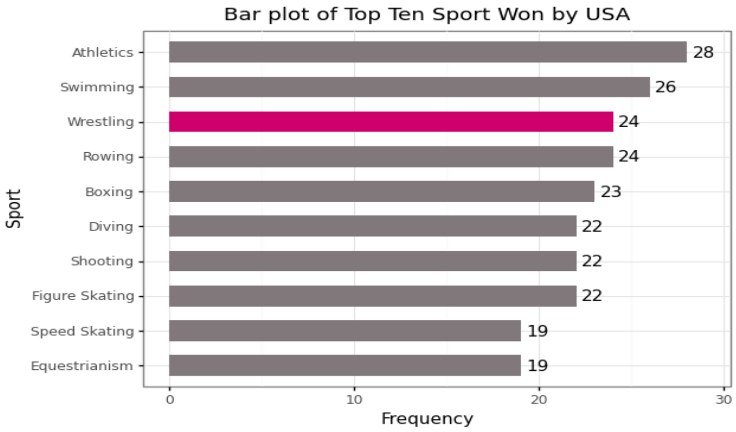

Bar plot of Countries to be Runner Up

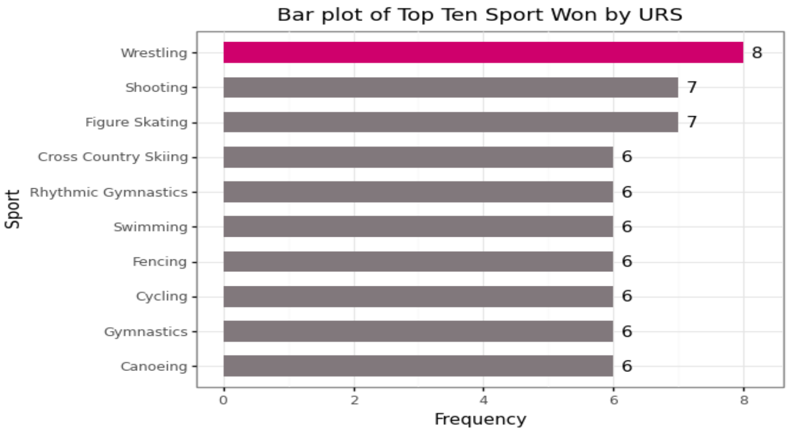

Bar plot of Countries to be Third Position

Findings

- For all Olympics event, United State of America (USA) have won the competition 16 times as general champion. Further, Uni Soviet has 8 times as general champion
- Despite not being 1st position, USA also active as runner up and 3rd position
- Uni Soviet is a rival of USA
- German and Canada are the other rival of USA with good potency

DATA ANALYSIS

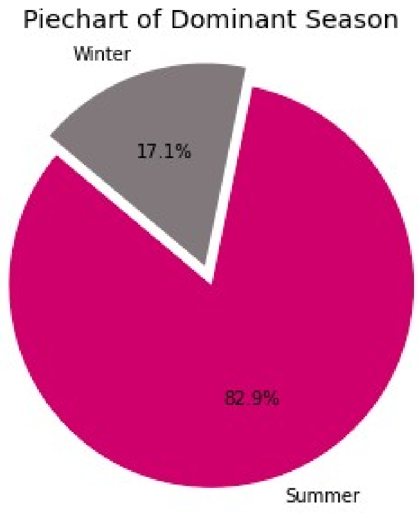


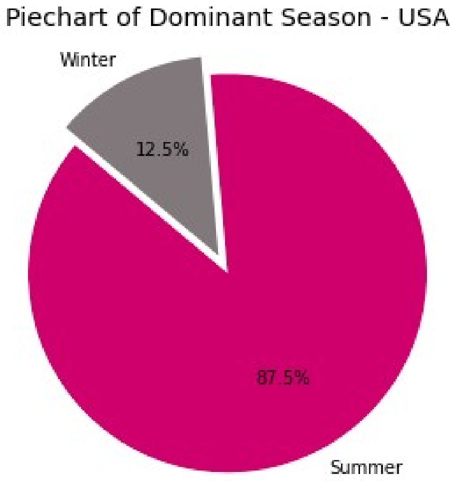


Findings

- As the rival of USA, Uni Soviet has the strongest sport with highest number of medals, that is wrestling
- The USA's sport with highest number of medals is athletics (28). It doesn't include in top ten sport won by the Uni Soviet
- Rowing, boxing, and diving can be optimized by USA in order to beat the real rival of Uni Soviet

DATA ANALYSIS

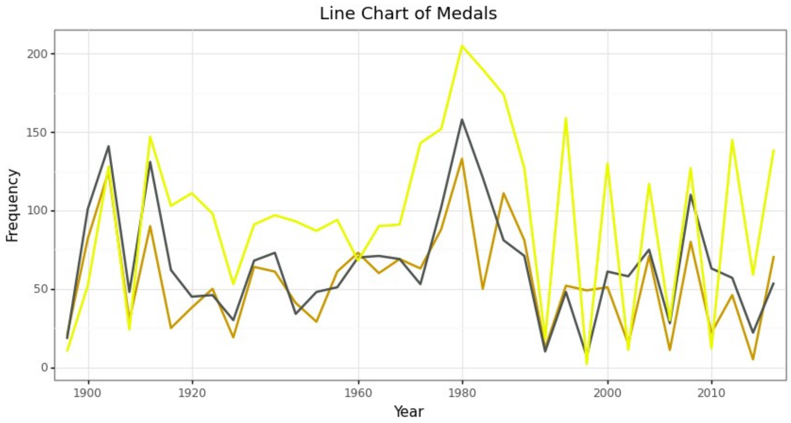




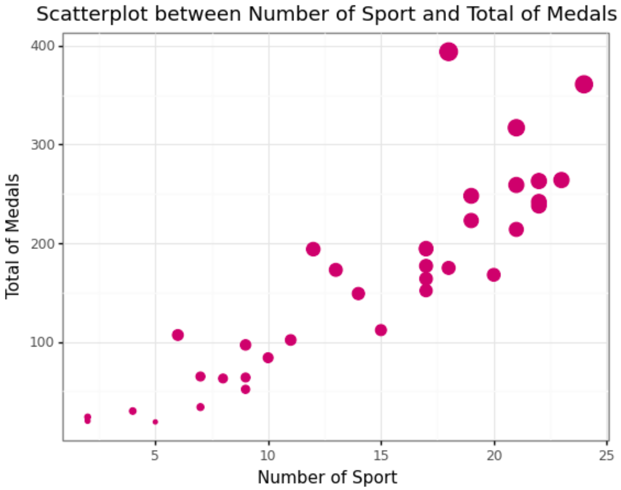
Season has not effect on the performance of USA in Olympics event

1986 - 2016

DATA ANALYSIS

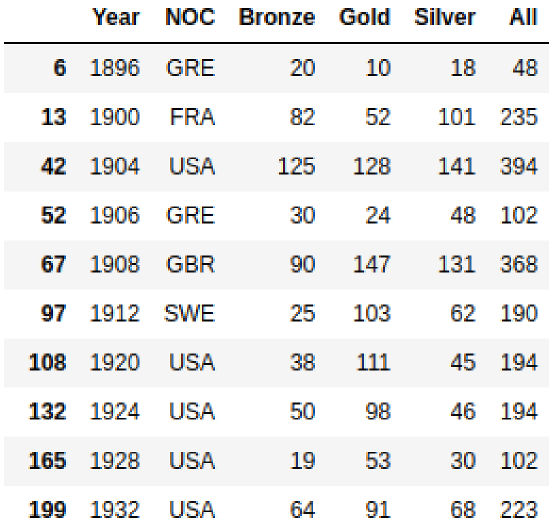


Line Chart of Medals



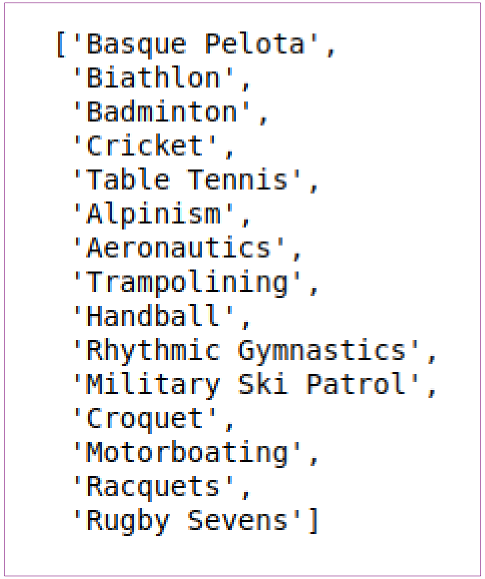
Scatterplot between Number of Sport and Total of Medals

Number of medal 1896 - 1932



	Year	NOC	Bronze	Gold	Silver	All
6	1896	GRE	20	10	18	48
13	1900	FRA	82	52	101	235
42	1904	USA	125	128	141	394
52	1906	GRE	30	24	48	102
67	1908	GBR	90	147	131	368
97	1912	SWE	25	103	62	190
108	1920	USA	38	111	45	194
132	1924	USA	50	98	46	194
165	1928	USA	19	53	30	102
199	1932	USA	64	91	68	223

The weakness of USA



```
['Basque Pelota',
 'Biathlon',
 'Badminton',
 'Cricket',
 'Table Tennis',
 'Alpinism',
 'Aeronautics',
 'Trampolining',
 'Handball',
 'Rhythmic Gymnastics',
 'Military Ski Patrol',
 'Croquet',
 'Motorboating',
 'Racquets',
 'Rugby Sevens']
```

Findings

- In 1980, it was a year with the highest number of medals to be contested
- Of course there is high positive correlation between number of sport with the total medals won by country (0.883)

CONCLUSION

- The USA dominates the Olympics event as the top three with highest medal in 1986 - 2016
- To defeat the USA in Olympics, other country must be discipline in sports training, especially athletics, swimming, and wrestling
- Other country are recommended to gain the medal from the list of 15 sport that had never been won by the USA in 1986 - 2016