2020

Statistical Analysis on Decathlon Data

MAT022 COURSWORK

MISBAH SOHAIL

C1950696

Submitted on: 10 January 2019

Abstract:

Statistical analysis is performed on a given data set related to athletes' performances in Decathlon athletics events from year 1996 to 2006. The first part of the report is based on the descriptive statistics explaining the basic variations in data. The second part comprises of inferential statistics which describes the correlation between events and the corresponding points achieved, the reduction in the dimensions using Principal Component Analysis and compares the performances of most winning nationalities.

Table of Contents

Introduction

Decathlon is a ten day athletics event which consists of ten different events: Long-jump, Shot-put, High-jump, Discus, Javelin, Pole-vault, 100 metres race, 400 metres race, 110 meters hurdles and 1500m race.

The data set comprises of 24 variables, three of them which are related to an athlete's profile are categorical, while the rest of them are numerical. The categorical variables are as follows:

- DecathleteName- Decathlete's name
- Nationality- Decathlete's nationality
- yearEvent- Year of performance

The numerical variables are related to the performance of the ten events, distance/time of each event and the points associated with it. There is one variable for the total points awarded to each player.

The sample of data starting from year 1996 till 2006 was taken into consideration for the whole analysis.

The report starts with the analysis of the descriptive properties of the variable. Frequency of categorical variables were discussed while for the numerical variables, mean and the spread of distribution was mentioned.

The next part of the report explains inferential statistics. One of the main goals of the analysis was to find the similarities and differences in performances of different nationalities in the above mentioned time frame, especially the most frequently winning nationalities, and provide some inferences about the whole data-set. Different tests and methodologies were used in this part. Spearman test was performed to check for any correlation within events and the respective points given to the players based on the performance. Ten events were defined in four significant dimensions using principal component analysis and based on the result, inferences were made to compare nationalities performances.

For inferential analysis, the sample data was merged into rows, with each row representing a single nationality. This was done by taking the median of the total points awarded to the players from each participating nationality.

The results of this report are similar to the findings mentioned in studies in the background section.

Background

A lot of research has been done on data pertaining to athletics with the main goal of reducing dimensions by applying factor analysis.

In Park 2011, Multivariate Statistical Analysis was performed on the data of results of Decathlon Performances from year 1988 to 2008. Principal component analysis was performed on the basis of individual performances and it was concluded that inter-discipline correlations exists between the performances in all ten events which are: 'sprinting abilities', 'throwing capabilities', 'jumping' and 'endurance'. (Reference [1] attached)

A similar study was published by Dziadek 2018 on the data of best Polish decathlon competitors from the year 1985 to 2015. Runs, throws and jumps were the three factors taken into consideration for the analysis of this study. (Reference [2] attached)

Descriptive Analysis

Players from 107 different nationalities participated in Decathlon games from 1986 to 2006. The three nationalities with the highest number of records are United States of America (USA), Germany and France with USA at the top with 823 records followed by Germany (504 records) and France (303 records).

Less than ten percent of the participating nationalities had more than a 100 records and there were nine nationalities with just a single record in all the years.

Although the USA had the most number of records, it was observed that USA's players topped just a total of 4 times in the span of 10 years while Czech Republic's(CZE) players, Roman Sebrle and Tomas Dvorak, remained the leading competitors interchangeably from year 2000 to 2005.

Since the winners of the Decathlon were either from the USA or CZE, the inferential analysis of this report is based on comparing the athletes' performances from these two nationalities.

Statistics

		Totalpoints	P1500	Pjt	Ppv	Plj	P100m	Psp	Phj	P400m	P110h	Pdt
N	Valid	4120	4120	4120	4120	4120	4120	4120	4120	4120	4120	4120
	Missing	0	0	0	0	0	0	0	0	0	0	0
Mean		7336.49	649.70	642.35	741.65	807.44	811.06	672.93	743.38	788.15	826.35	653.54
Median		7235.00	657.00	638.00	731.00	804.00	810.00	672.00	740.00	789.00	828.00	651.00
Std. Deviation		415.599	89.663	89.490	105.873	76.152	61.469	76.194	76.556	64.513	66.427	86.084
Variance		172722.130	8039.526	8008.383	11209.140	5799.202	3778.490	5805.452	5860.826	4161.942	4412.548	7410.390
Skewness		.966	-.557	.146	.003	.223	.095	.079	.190	-.082	-.084	.150
Std. Error of Skewness		.038	.038	.038	.038	.038	.038	.038	.038	.038	.038	.038
Kurtosis		.509	.684	-.031	.170	.105	-.054	-.199	.038	-.020	.066	.048
Std. Error of Kurtosis		.076	.076	.076	.076	.076	.076	.076	.076	.076	.076	.076
Range		2226	767	628	831	520	426	512	645	461	518	716
Percentiles	25	7008.00	595.00	582.00	673.00	755.00	769.00	619.25	687.00	745.25	782.00	593.00
	50	7235.00	657.00	638.00	731.00	804.00	810.00	672.00	740.00	789.00	828.00	651.00
	75	7592.00	712.00	703.00	819.00	857.00	852.00	725.75	794.00	832.00	870.00	709.00

Table 1: Descriptive analysis of total and each event points from 1996 to 2006

Table 1 describes the basic statistical properties of total points and each events' points for performances from year 1996 to 2006. It can be observed from the table above that for most of the variables, the value of skew-ness is greater than 0.5 but less than 1, and the mean and median values are not the same, thus it can be stated that the data point's variables are moderately skewed. (Reference [5] attached)

Below is a histogram of frequency of Total Points, which have the largest absolute skew-ness value of 0.966 among all other variables.

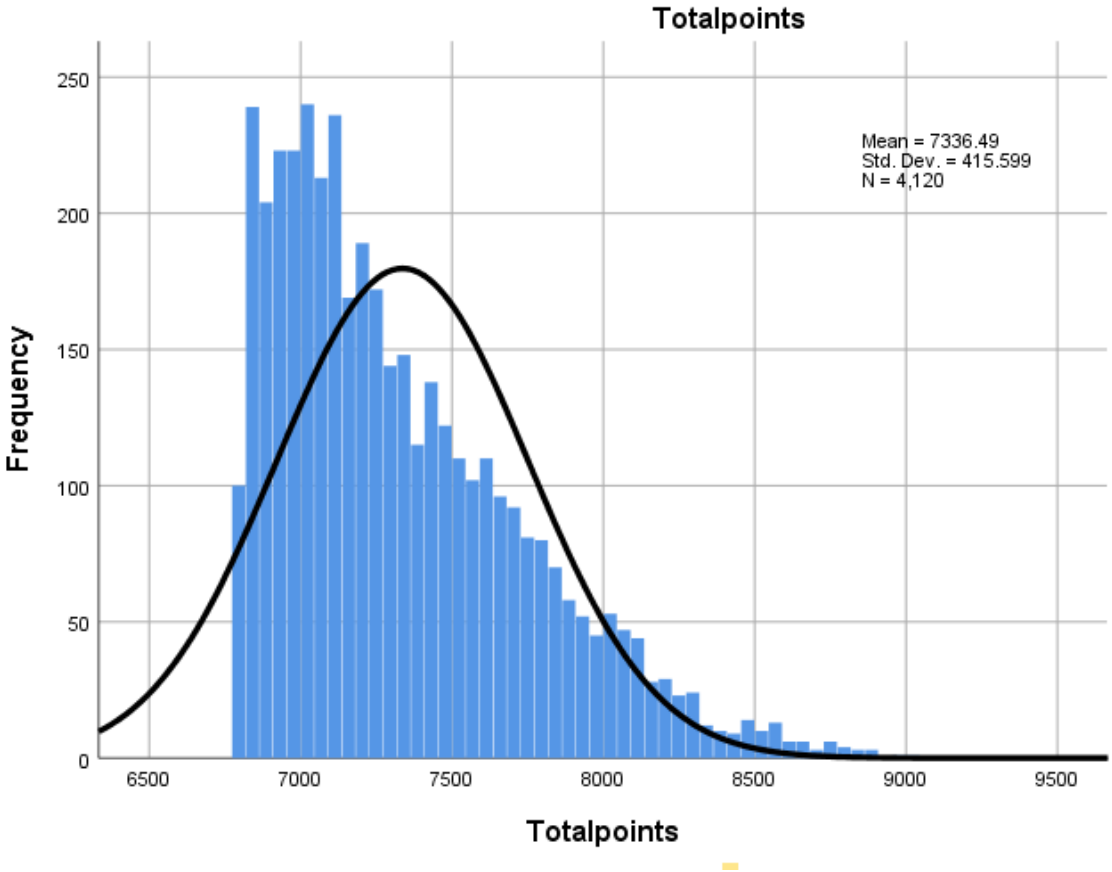

Figure 1: Histogram based on the frequency of Total-points

It is evident in Figure 1 that the data is not symmetrical around the mean. It is more clustered towards the left hence concluding that it is positively skewed.

Inferential Analysis

The section is divided into three parts: the first part explaining the correlation between the variables, the second part emphasizing on applying Principal Component Analysis on the data set, and the third part is based on comparing both nationalities' performances projected on the new dimensions and confirming if there is enough evidence to support the finding which are made after applying Principal Component Analysis.

Linear Regression

Since the data was found to be not-normal, in Descriptive Analysis, non-parametric test, Spearman Correlation test was applied to analyze any correlation between all events and the points achieved. It was observed that the points awarded for all event against their measured metrics displayed a perfectly linear relationship.

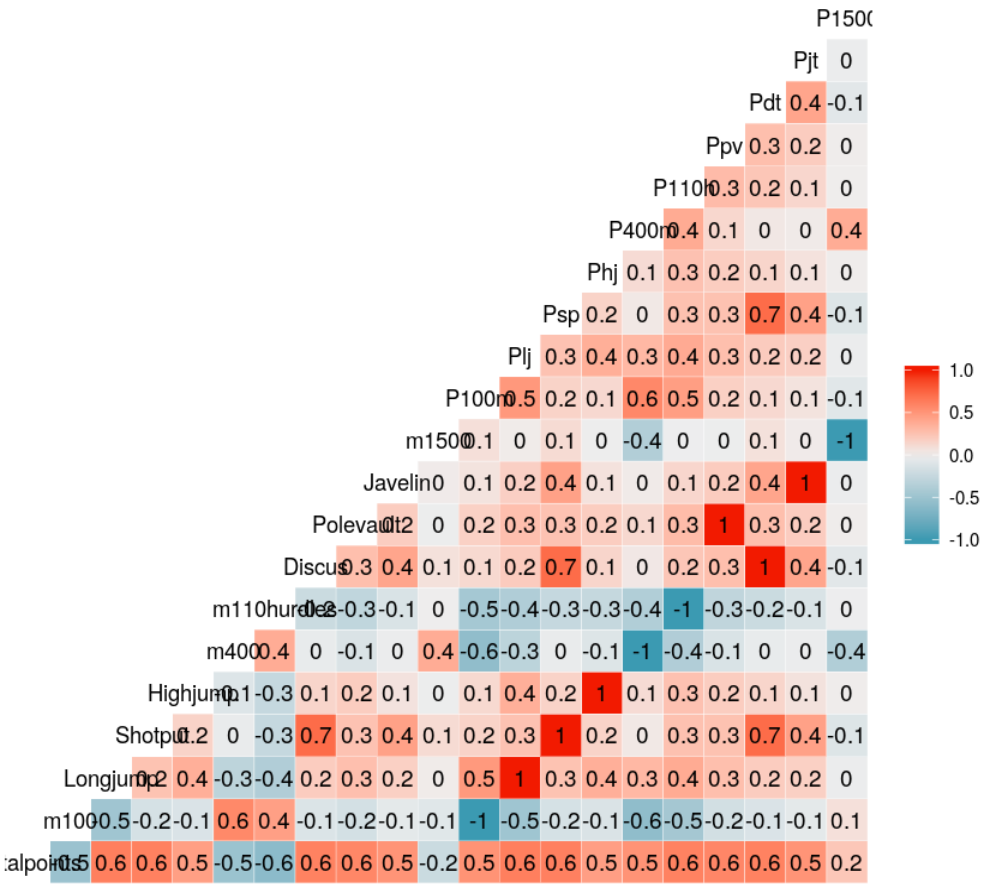

Figure 2: Correlation coefficients all variables (measured metric and the points rewarded for all 10 events)

The correlation coefficients were examined based on the standard alpha value of 0.05.

The results are illustrated in Figure 2. Positive correlation of value +1 was observed between points awarded and the distance metric based game. Similarly, negative correlation of value -1 was observed for time metric based games.

Hence, since it is evident that the measured metric and the points awarded are linearly related to each other, only the points awarded for events are used for further analysis.

As discussed in the section of descriptive analysis, most of the nationalities had less than 100 players participating in between 1996 to 2006. Inferential statistics were performed based on each nationality. The players from the same nationality were grouped together in a single row by taking the median of each scores column. The median scores are used to overcome the presence of outliers that would affect the mean score (see Table 2).

Nationality	Total_Points	P100	Long Jump	Putting the shot	High Jump	P400m	110m with hurdles	pole vault	Discus	Javelin	P1500
<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
ALG	7113	810	866	575.0	687	820.0	808	645.0	600.0	583.0	678
ARG	7335	830	827	656.5	731	822.5	821	745.5	645.0	656.5	647
AUS	7280	814	821	670.0	740	815.0	788	731.0	671.0	684.0	646
AUT	7282	809	811	691.5	723	791.0	824	702.0	647.5	680.5	652
BAH	7205	814	802	700.0	723	766.0	835	673.0	695.0	627.0	591
BAR	7766	929	952	614.0	859	869.0	969	482.0	535.0	723.0	730

Table 2: Top 5 rows of table displaying nationalities with points of all events

Principal Component Analysis

Principal Component Analysis is a feature reduction technique which generates new dimensions based on the linear correlations of original variables. These new dimensions are uncorrelated to one another and explain most of the variance in the original data. (Reference [3] and [4] attached)

Applying the technique to the data resulted in ten new principal dimensions (Dim). The summary in Table 3 tells that Dim.1 explains 20.74 percent of the total variance, Dim.2 explaining 19.6 giving a cumulative of 44 percent variance. The top 4 Dimensions maintains a cumulative variance of a significant 71.4 % which will be used for further analysis.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10
Variance	2.474	1.960	1.487	1.222	0.727	0.699	0.629	0.417	0.215	0.170
% of var	24.744	19.601	14.875	12.222	7.272	6.987	6.286	4.167	2.149	1.698
Cumulative % of var.	24.744	44.345	59.219	71.442	78.713	85.700	91.986	96.153	98.302	100.000

Table 3: Summary of cumulative variances as defined in Dimensions 1-10.

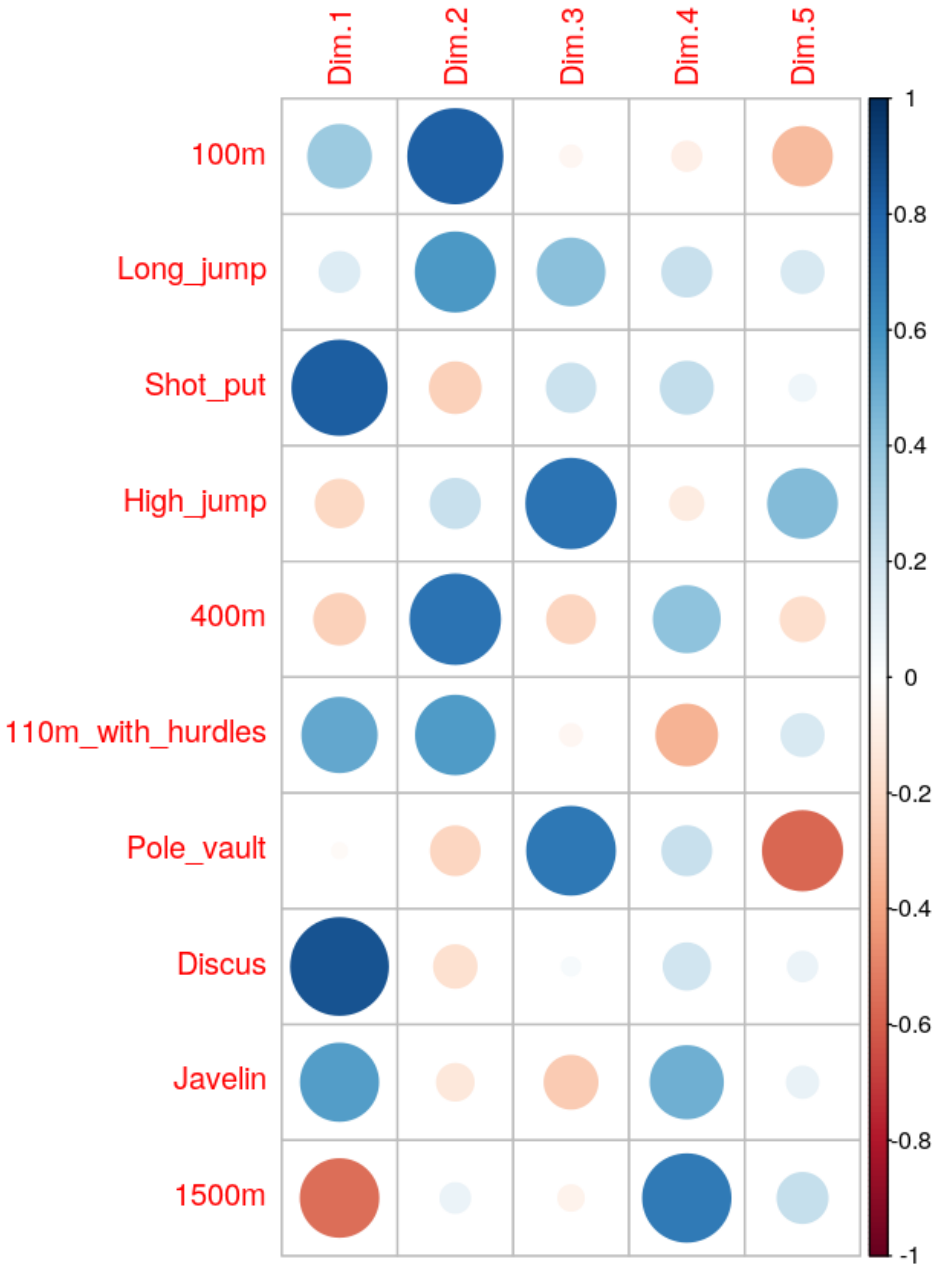

Figure 3 summarizes first 5 dimensions correlation values against all events.

As demonstrated by the plot Figure 3, based on the correlations of events, no particular category of events (events requiring similar capabilities) is correlated to Dim.5. Additionally it explains only 7.3% of the variance in data. Therefore, only the first four dimensions are considered further.

The Table 3 below shows a detailed description of first four dimensions with their correlation values.

```
decathlon.pca.desc <- dimdesc(decathlon.pca, axes = c(1:5), proba = 0.05)
$quanti
```

	correlation	p.value
Discus	0.8285145	3.367295e-28
Shot-put	0.7529845	8.449801e-21
110m with hurdles	0.5751878	9.181886e-11
Javelin	0.5242397	6.820720e-09
100m	0.4933741	6.687300e-08
High-jump	-0.2330752	1.569186e-02
P1500	-0.5294931	4.519027e-09

Table 4.1: Correlation table for Dim.1

	correlation	p.value
400m	0.7656377	7.672029e-22
100m	0.7268416	7.813436e-19
Long-jump	0.5191580	1.009027e-08
110m with hurdles	0.4637017	4.902566e-07
1500m	0.2040604	3.500834e-02
Discus	-0.2503728	9.294615e-03
Pole-vault	-0.3137102	1.000433e-03
Shot-put	-0.3457759	2.643760e-04

Table 4.2: Correlation table for Dim.2

	correlation	p.value
High-jump	0.7857789	1.212707e-23
Pole-vault	0.6173847	1.416170e-12
Long-jump	0.4480580	1.302408e-06
Shot-put	0.2044792	3.462793e-02
400m	-0.2291395	1.759290e-02
Javelin	-0.3756556	6.685052e-05

Table 4.3: Correlation table for Dim.3

	correlation	p.value
1500m	0.6756087	1.432528e-15
400m	0.4276785	4.333350e-06
Pole-vault	0.3982144	2.158335e-05
Javelin	0.3582723	1.512230e-04
Putting the shot	0.3176115	8.574051e-04
Discus	0.2308913	1.672350e-02
Long-jump	0.1904828	4.938417e-02
110m with hurdles	-0.3051374	1.394092e-03

Table 4.4: Correlation table for Dim.4

Table 4: Significant Correlations of first 4 Dimensions with events

Alpha is chosen to be 0.05. Since the p-values of all the variables are less than 0.05, the correlation coefficients of all the variables mentioned in the table are statistically significant. However, the top two events for first three dimensions as well as the first event in the Table4.4 (as highlighted) contributed the most towards each dimension. Higher values indicate a greater contribution towards a dimension.

Throwing events like Discus and Shot-put majorly contribute towards the variability in Dim.1 (See Table 4.1). Dim.2 (see Table 4.2) is largely being influenced by the running events (100m and 400m race). Dim.3 (see Table 4.3) depends on High-jump and Pole-vault which can be categorized as jumping events. Only P1500, which is largely based on endurance, has a high impact on Dim.4 (see Table 4.4)

Below are the loading plots which shows projections of all events points on Dim.1 & Dim.2 and Dim.3 & Dim.4 respectively. Besides describing how a variable influences a dimension, the length and the direction of a vector determines the strength and sign of correlation respectively. The angle between two vectors shows the strength of correlation between them.

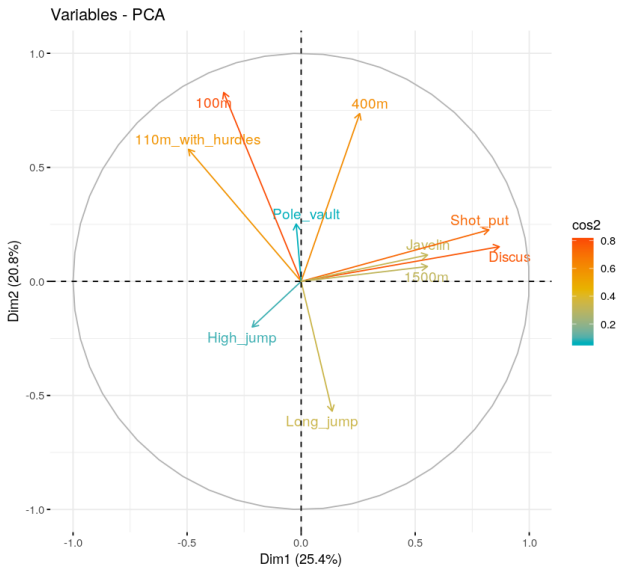

Figure 4: Loading plot of all events on Dim.1 and Dim.2 (Dim.1 on horizontal-axis, Dim.2 on vertical-axis)

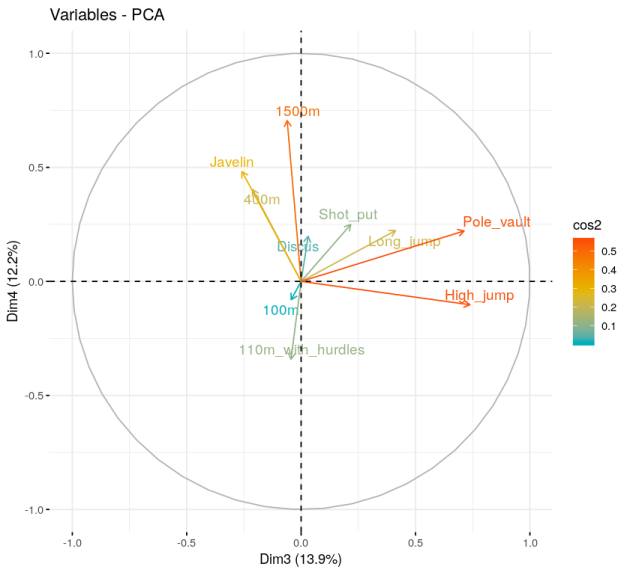

Figure 5: Loading plot of all events on Dim.3 and Dim.4 (Dim.3 on horizontal-axis, Dim.4 on vertical-axis)

Discus and Shot-put labeled vectors (See Figure 4) are the longest vectors pointing towards the right, depicting a strong positive correlation with Dim.1. Similarly, 400m and 100m races have a strong positive correlation with Dim.2. Justifiably, the small angle between Discus and Shot-put explains that the events are linked because they are both related to throwing. Vectors which are normal to one another imply that there is no significant relationship between them. Throwing events and running events like (100m) seem to have no meaningful relationship to one another.

Pole-vault and High-jump vectors illustrate a high correlation with Dim.3 while 1500m race is largely positively correlated with Dim.4 as can be seen in Figure 5.

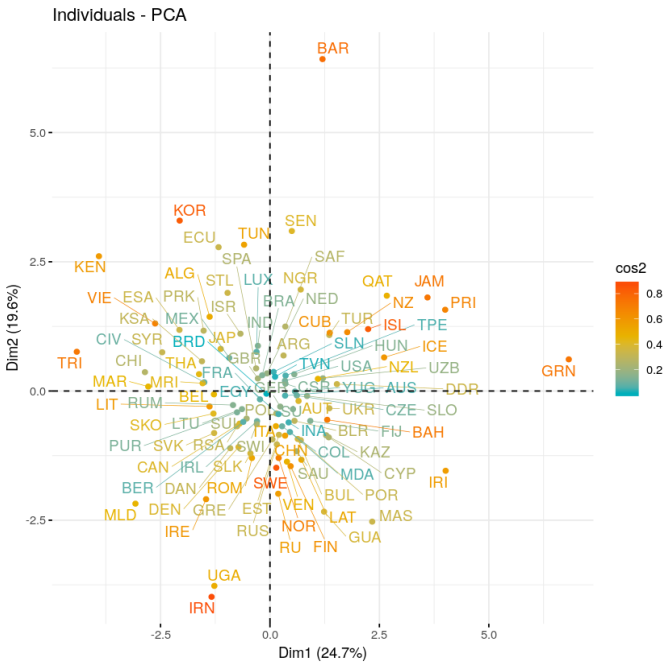

Figure 6: Nationalities position with respect to Dim.1 and Dim.2 (Dim.1 on horizontal-axis, Dim.2 on vertical-axis)

It can be observed that Barbados (BAR) and Grenada (GRN) have outlying positions with respect to Dim.1 and Dim.2 (See Figure 6). It can be implied that participants from BAR performed the best at running related events as compared to other nationalities. Similarly, GRN athletes did the best in throwing games comparatively to other countries. However, considering BAR and GRN are in the same geographical region, it can be assumed that athletes from the West Indian region have a higher tendency to perform in these events especially those demonstrating speed and throwing capabilities. Having said that, this cannot be defined as a concrete conclusion due to the fact that each of these nationalities had only 1 athlete playing over the span of 10 years.

Comparing Groups

The performance of the most winning nationalities (USA and Czech Republic) as mentioned in the descriptive analysis section will discussed in this section, based on the results of Principal Component Analysis and further performing statistical tests to confirm the findings.

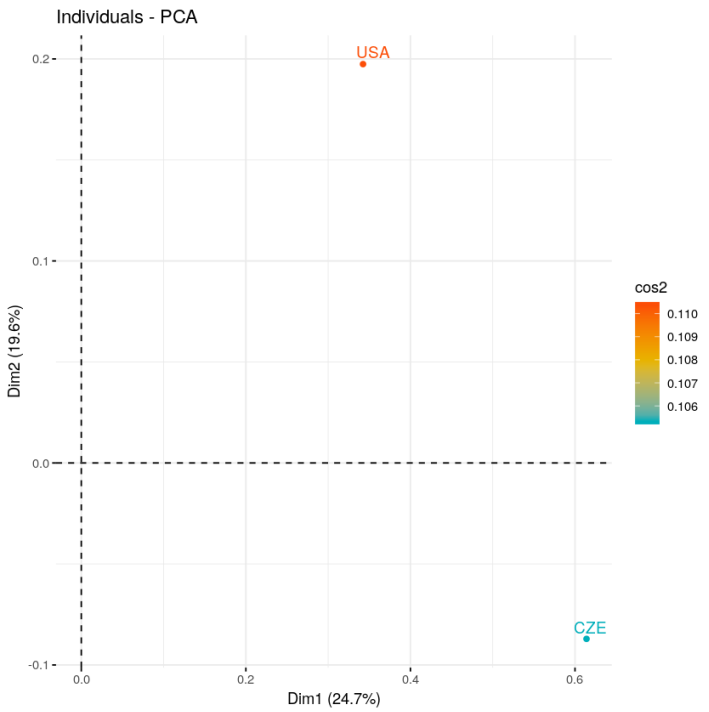

Figure 7: USA and CZE's position with respect to Dim.1 and Dim.2 (Dim1 on horizontal-axis, Dim.2 on vertical-axis)

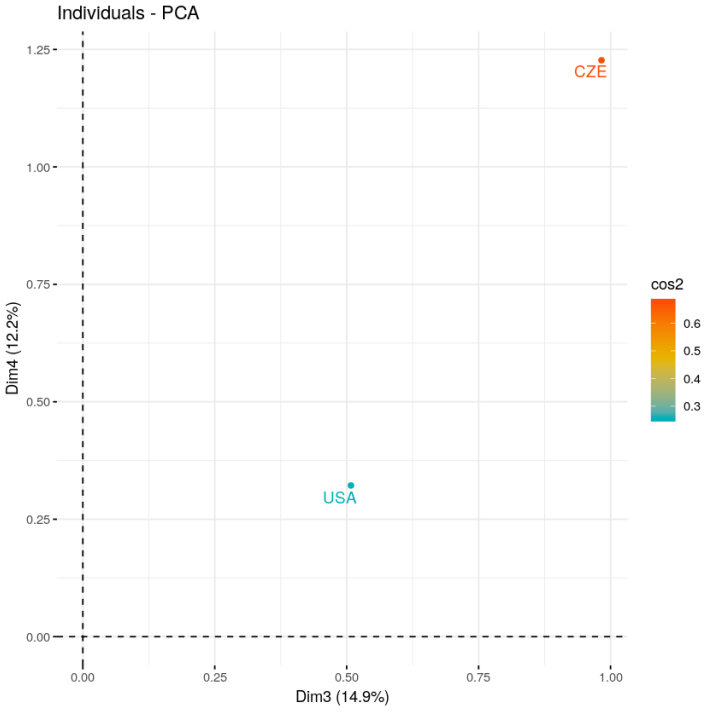

Figure 8: USA and CZE's position with respect to Dim.3 and Dim.4(Dim.3 on horizontal-axis, Dim.4 on vertical-axis)

Figure 7 shows that CZE's athletes performed better in events contributing towards Dim.1 which particularly define throwing capabilities while USA's players are more likely to perform better than CZE and other nationalities with their running strength (100m and 400m race) which majorly define Dim.2.

Similarly, as shown in Figure 8, CZE's players jumping and endurance capabilities appear better than the players from USA.

Further analysis was done to confirm if the above statements were actually being reflected in the whole sample data-set by comparing the points earned by players of USA and CZE in Shot-put, Discus, 100m race, 400m race, High-jump, Pole-vault and 1500m race.

Total points of all performances of both groups was used to check for normality of 2 groups.

Below are the results:

```
        Shapiro-Wilk normality test

data:  dataset1$Totalpoints[dataset1$Nationality == "USA"]
W = 0.93728, p-value < 2.2e-16

data:  dataset1$Totalpoints[dataset1$Nationality == "CZE"]
W = 0.92061, p-value = 1.867e-06
```

The test was conducted with an alpha value of 0.05. For both the groups, the null hypothesis is rejected as the p-values are less than 0.05 and it can be stated that the distributions are non-parametric for both groups.

Mann-Whitney-Wilcoxon Test, with an additional argument of less/greater, was used to compare the points earned in different events for two groups. The results are as follows:

Dim1(Throwing) Hypothesis: CZE>USA	Dim2(Running) Hypothesis: CZE<USA	Dim3(Jumping) Hypothesis: CZE>USA	Dim4(Endurance) Hypothesis: CZE>USA
Shot-put by Nationality W = 104614, p-value = 6.556e-05	100m by Nationality W = 75216, p-value = 0.00728	High-jump by Nationality W = 95898, p-value = 0.02459	1500 by Nationality W = 113601, p-value = 4.728e-09
Discus by Nationality W = 87286, p-value = 0.4483	400m by Nationality W = 84914, p-value = 0.3537	Pole-vault by Nationality W = 99956, p-value = 0.002299	

Table 5: Results of Wilxon Test for comparing USA and CZE's performances in seven different events

The alpha value was set at 0.05 (see Table 5 for results). According to the results discussed, the following can be concluded:

1. For throwing events, CZE's athlete's performance in Shot-put is significantly better than the players of the USA. But for Discus, the null hypothesis is being retained (p-value= 0.4483)
2. For running events, CZE's athlete's performance in 100m race is significantly less good than the players of the USA. But for 400m race, the null hypothesis is being retained with the (p-value= 0.3537)
3. CZE's athletes perform better in all jumping events (High-jump and Pole-vault) than the players of the USA. The null hypothesis is rejected in both the cases (p-value= 0.02459, p-value= 0.002299 respectively)

4. For the 1500m race, the null hypothesis is rejected (p-value= 4.728e-09) and it can be stated CZE's athletes perform better in 1500m race than the players of the USA.

Conclusion

The following conclusions can be inferred from the analysis:

1. The top performing nationalities are found to be USA and CZE
2. Event performance metrics and the point achieved are linearly correlated to one another
3. Seven of the Decathlon's events (Shot-put, Discuss, 100m race, 400m race, High-jump, Pole-vault, 1500m race) can be defined in 4 significant dimensions which describes 71 percent of the variability of all 10 events. In particular, the first dimension describing about throwing related events, the second one for running, and the third one was jumping and fourth dimension explains endurance.
4. CZE players perform better than the USA in Shot-put, High-jump, Pole-vault and 1500m race.
5. USA players perform better in 100m race.

References

1. Park, J. and Zatsiorsky, V.M., 2011. Multivariate statistical analysis of decathlon performance results in olympic athletes (1988-2008). World Academy of Science, Engineering and Technology, 5(5), pp.985-988.

2. Dziadek, B., Iskra, J. and Przednowek, K., 2018. Principal Component Analysis in the Study of Structure of the Best Polish Decathlon Competitors from the Period between 1985–2015. Central European Journal of Sport Sciences and Medicine, 23(3), pp.77-87.

3. Jolliffe, I.T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), p.20150202.

4. Sthda.com. (2017). PCA - Principal Component Analysis Essentials - Articles - STHDA. [online] Available at: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/[Accessed 10 Jan. 2020]

5. Gooddata.com. (2019). Normality Testing - Skewness and Kurtosis - Documentation. [online] Available at: https://help.gooddata.com/doc/en/reporting-and-dashboards/maql-analytical-query-language/maql-expression-reference/aggregation-functions/statistical-functions/predictive-statistical-use-cases/normality-testing-skewness-and-kurtosis?fbclid=IwAR2sckmMjQa2s8fuhX0prU2DX-w5383486FZdtelh97bH2wbx990ne2recU [Accessed 10 Jan. 2020]