UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

COURSE NAME: PROBABILITY AND STATISTICAL DATA ANALYSIS (SECI2143)

Title of PROJECT 2: The Sciences of Marathon Running

Submitted to: Dr.Azurah binti A Samah

Report Prepared by:

Group 9-The 3 WHISKERS

Section:9

1. Abraham Loh Tze Lung(A20EC0001)

2. Afifa Jumana(A20EC4009)

3. Henry Meshack Okinyi Odongo(A20EC0268)

Due Date: 3rd July,2021.

Table of Contents

Project 2 Topic: The Sciences of Marathon Running

1.0 Introduction/background

Marathon running (42.195km) has been a prevalent sport all across the nooks of world including Tokyo Marathon in Japan, Berlin Marathon in Germany, Boston Marathon and New York City Marathon in United States that attracts avid runners from various countries. There are two distinguishable types of runners and we named them recreational runners and competitive runners. The former do routine runs for a healthier body such as losing weight, improving blood circulation, sweating to remove unwanted toxics from body and strengthen their minds. The latter are well-trained athletes that follow professional training programs to break their own personal records.

In this project, the subjects we targeted on are male competitive marathon runners. There are various fallacy claims with unsupported proof all over the runner's forums and discussions such as "East Africans possessed running talents, they are born to be good runners", "The more miles you run, the better finishing times you will clocked in marathon running", and "The recent carbon-plated Nike running shoes will confirm a personal records in marathon running". Therefore, we are interested to carry out several inferential statistics tests to verify those claims from solid dataset retrieved from online sources to test what variables actually contribute to faster marathon running times.

In the later sections, we carried out 2-sample hypothesis test on the mean Berlin Marathon 2018 (Adidas shoes) results and mean Berlin Marathon 2019 (Nike shoes) results of male world class athletes. We also carried out correlation and regression analysis on the relationship between weekly training mileage of selected marathon runners and their marathon performances. The last test we carried out is chi-square test of independence to test whether running performances is independent or not independent of the geographical area of the runners reside in.

In this project, the test calculations and important scatter plot graphs were assisted by RStudio programming language. In the following section, we roughly described the dataset we retrieved from and more detailed can be found in appendix section.

2.0 Dataset

2-sample Hypothesis Test

The dataset are retrieved from official Berlin Marathon event website that consists of full results report.

https://www.bmw-berlin-marathon.com/en/impressions/statistics-and-history/results-archive/

Pre-processing:

Top 30 athletes in Berlin Marathon and their respective finishing times (in minutes) were selected from both years 2018 and 2019. (Refer to appendix for raw dataset and processing)

Linear Correlation and Regression Analysis

The dataset are from the Strava and Garmin Connect fitness tracking Apps that can view local runners running logs and their weekly mileage. The personal records can also be found from their profile.

Pre-processing:

30 random male runners were selected and both their average weekly running mileage and their marathon personal best records were noted and transferred to excel. (More details in appendix)

Chi-square Test of Independence

The dataset are retrieved from official IAAF(International Association of Athletics Federations) website containing full list of track and field events that can be searched via custom filter.

https://www.worldathletics.org/records/all-time-toplists/middlelong/5000-metres/outdoor/men/senior

Pre-processing:

15 random African male runners and 15 random non-African male runners under age 15 were selected and their 5,000m best records and nationality were observed and transferred to excel. (More details in appendix)

Why we choose the parameter/variables and possible outcome of the test based on our chosen variables:

For 2-sample hypothesis testing, we chose top 30 finishers in Berlin Marathon 2018 and Berlin Marathon 2019 and compared their overall finishing times. This is because in 2018, the

top athletes wore Adidas sponsored running shoes and in 2019, they wore carbon-plated Nike running shoes sponsored by Nike Company that deemed to aid faster running times.

For linear and regression test, we randomly selected 30 local competitive male runners from Strava fitness tracking apps to test whether weekly training mileage would affect marathon running times. Because there are other approaches that are as well believed such as low running mileage per week but was done at faster running speed or incorporating more strength training in the gym.

For chi-square test of independence, we chose 15 male competitive runners from East Africa and another 15 from non-African country under age 15. The performance of their 5,000m events were noted respectively and number of them being capable of less than 15 minutes were observed accordingly. This is to test the claim whether East African are genetically blessed with running talents or hard work that matters.

3.0 Data Analysis

3.1 2-samples Hypothesis Testing

Null Hypothesis, H_0: The population mean of the finish times by elite male runners in Berlin Marathon 2018 is the same as the population mean of the finish times by elite male runners in Berlin Marathon 2019.

Alternative Hypothesis, H_1: The population mean of the finish times by elite male runners in Berlin Marathon 2018 is not the same as the population mean of the finish times by elite male runners in Berlin Marathon 2019.

By using $\alpha=0.05$, $\alpha/2=0.025$ (For a two-tailed hypothesis test)

The formula of the test statistic calculation is given by:

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

The degree of freedom is

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}$$

```
> View(Hypothesis_Testing_2_sample)
> G2018 = Hypothesis_Testing_2_sample$G2018
> G2018
 [1] 121.65 126.38 126.80 128.27 128.77 129.30 129.93 131.12 132.95 133.15 133.72 134.95
[13] 135.00 135.07 135.32 135.62 136.43 136.47 136.65 137.48 137.80 139.10 139.23 139.75
[25] 139.88 139.90 140.30 140.87 141.02 141.23
> G2019 = Hypothesis_Testing_2_sample$G2019
> G2019
 [1] 121.68 122.80 123.60 126.75 128.12 128.43 128.47 128.58 128.93 129.65 130.43 130.95
[13] 131.13 131.23 131.83 132.48 132.68 133.32 133.37 133.63 133.65 133.80 133.85 133.88
[25] 133.95 134.03 134.08 134.28 134.45 134.52
> xbar1 = mean(G2018)
> xbar1
[1] 134.8037
> xbar2 = mean(G2019)
> xbar2
[1] 130.9517
>
> s1 = sd(G2018)
> s1
[1] 5.083211
> s2 = sd(G2019)
> s2
```

Figure 1: The results generated using RStudio including mean and standard deviation finishing time in Berlin Marathon 2018 and 2019

```
> s2
[1] 3.609539
> n1 = 30
> n2 = 30
>
> t0 = (xbar1-xbar2-0)/(sqrt((s1^2/n1)+(s2^2/n2)))
> t0
[1] 3.384168
>
> v = ((s1^2/n1)+(s2^2/n2))^2/((((s1^2/n1)^2)/(n1-1)) + (((s2^2/n2)^2)/(n2-1)))
> v
[1] 52.31699
>
> alpha = 0.05
> t.alpha = qt(alpha/2,floor(v))
> c(-t.alpha, t.alpha)
[1]  2.006647 -2.006647
>
```

Figure 2: The results generated using RStudio including test statistic, degree of freedom and critical regions at alpha=0.05

Conclusion from test:

Since the test statistic, T_0=3.3841 lies within the critical region $T_{crititcal}$>2.006647, we decide to reject H_0. We have sufficient evidence to conclude that the advent of carbon-plated Nike Vaporfly shoes have significant improvement on the overall marathon running times in Berlin Marathon course at alpha=0.05.

3.2 Correlation test

In our correlation test conducted, we calculated the coefficient of correlation, r using The Pearson's Product-Moment Correlation Coefficient given by the formula:

$$r = \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{[(\sum x^2) - (\sum x)^2/n][(\sum y^2) - (\sum y)^2/n]}}$$

Where r=sample correlation of coefficient

n=sample size (30 in our case)

x=value of the independent variable (miles per week in our case)

y=Value of dependent variable (Marathon finishing times in our case)

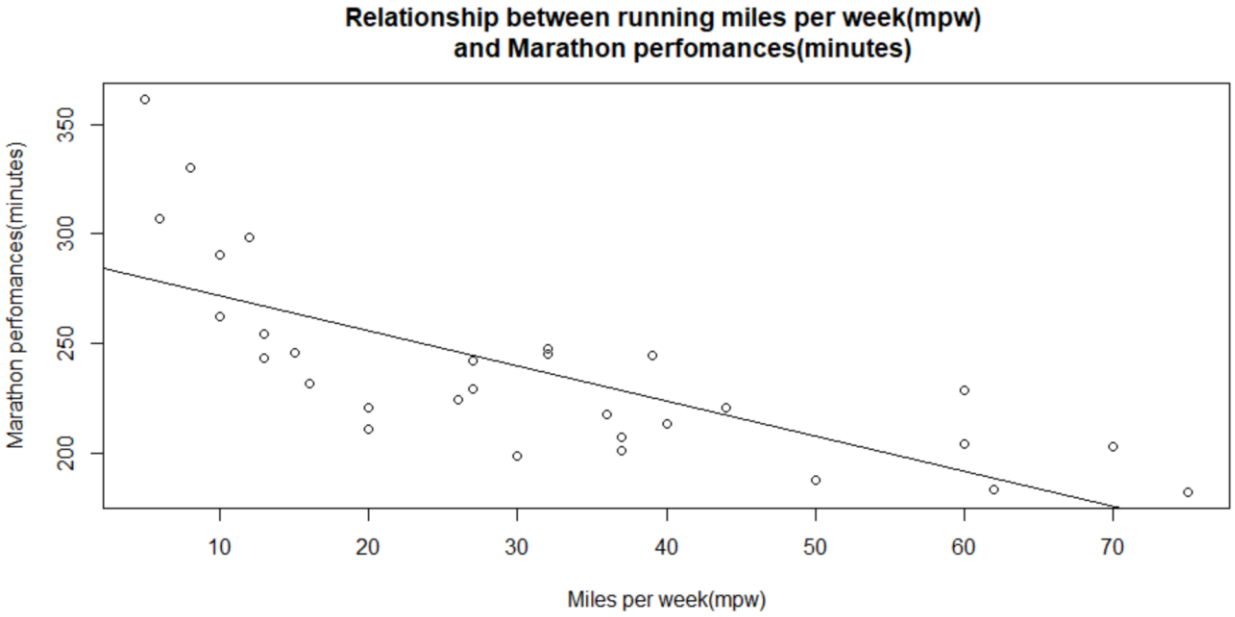

Figure 3: Scatter plot graph of Marathon performances in minutes against Miles per week in miles

```
        Pearson's product-moment correlation

data:  x and y
t = -5.7705, df = 28, p-value = 3.4e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8670689 -0.5129734
sample estimates:
       cor
-0.7370327
```

Figure 4: The results generated using RStudio including test statistic, degree of freedom, p-value, confidence interval and coefficient of correlation

Interpretation and discussion on what we can learn from the tests and graph

From Figure 4, the correlation coefficient is found to be -0.7370327 from RStudio. This tells us that there is a fairly strong correlation that the more miles a particular runners ran, the less time required to finish a marathon race (better performances). However, we have to accept the fact that there are other numerous factors come into play to the marathon finishing time.

Significant Test for Correlation

Null hypothesis, H_0: There is no correlation between weekly running mileage and marathon running performances

Alternative hypothesis, H_1: There exists a correlation between weekly running mileage and marathon running performances

By using α=0.05 and degree of freedom = 30-2=28

Test statistic is calculated from the following formula:

Test statistic:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

From Figure 4, the test statistic calculated was -5.7705. Also, the 95% confidence interval for the correlation coefficient is (-0.8670689, -0.5129734). From t-table, the critical regions are $t_{0.025,28}$<-2.048 or $t_{0.025,28}$>2.048.

Conclusion from test:

Since the test statistic calculated was -5.7705 lies within the critical region $t_{0.025,28}$<-2.048 at left tail, we had decided to reject the null hypothesis. At α=0.05, we can conclude that there is a correlation between weekly running mileage and the marathon finishing time.

3.3 Regression test

In regression test, we decide to further delve into how significant weekly running mileage could have an impact on marathon finishing time. This is because most of the runners hold a firm belief on weekly mileage rather than incorporating a speed training session that could have a risk of injury. There are also arguments that running mileage per week only work to a certain point.

Estimated Regression Model:

$Y = b_0 + b_1X$

Where,

Y = Estimated (or predicted) Y value

b_0 = Estimate of the regression intercept

b_1 = Estimate of the regression slope

X = Independent variable

We use n=30, x= running miles per week and y= best effort for marathon running finishing time in minutes and a regression model was generated by using RStudio.

```
Residuals:
    Min      1Q  Median      3Q     Max
-44.612 -20.872  -7.112  17.647  81.647

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 287.7404    10.1626   28.31  < 2e-16 ***
x            -1.5994     0.2772   -5.77  3.4e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.56 on 28 degrees of freedom
Multiple R-squared:  0.5432,    Adjusted R-squared:  0.5269
F-statistic:  33.3 on 1 and 28 DF,  p-value: 3.4e-06

> confint(model)
                2.5 %     97.5 %
(Intercept) 266.923372 308.557516
x            -2.167175  -1.031652
> sigma(model)*100/mean(y)
[1] 12.41853
```

Figure 5: The results generated using RStudio including important calculation of regression analysis

From the RStudio summary, we can get the formula for estimated regression model as:

Y= 287.7404 – 1.5994X

From the model, we again confirmed that the more miles a particular runner ran per week(X), the value of Y will decrease at a factor of 1.5994(Thus faster marathon finish times). However, this model is just a rough prediction. The one of the flaws of this model is that it assumes a non-runner or untrained athlete (0 mile ran per week) is capable of 287 minutes marathon or 4hours and 47 minutes which is very unlikely. Also, the improvement of marathon running is of diminishing returns, meaning the improvement is getting less obvious at the same dose of training mileage.

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{sum of square explaind by regression}}{\text{total sum of squares}}$$

From Figure 5, the value of R^2 is 0.5432.

Test Statistical of Regression

$H_0 = \beta_1 = 0$ (no linear relationship)

$H_1 = \beta_2 \neq 0$ (linear relationship)

Test statistic, $t = \frac{b_1 - \beta_1}{S_{b_1}}$

Degree of freedom = n-2

Where:

b_1= Sample regression slope coefficient

β_1= Hypothesized slope

S_{b1}=Estimator of the standard error of the slope

Conclusion from test:

Based on Figure 5, test statistic, t = 6.06. P-value is the significance level of the t-test. P-value is 3.4×10^{-6}. Since P-value is 0.0000034 is less than significance level 0.05, therefore we reject the null hypothesis. So, there is sufficient evidence that the miles per week ran affects the marathon running race performances.

3.4 Chi-Square test of independence

In this test, we selected randomly 15 male East African runners (including Ethiopia, Kenya, Somalia, Eritrea, Uganda, Morocco) and 15 male Non-African runners (Japanese, Chinese, European, American) high school students under 15 years old and the number of runners capable of finishing 5,000m in under 15 minutes was observed.

< 15 minutes / Participants	Yes	No	Total
African	10	5	15
Non-African	6	9	15
Total	16	14	30

Table 1: A two-way contingency table of our data

Null Hypothesis, H_0: The ability to run 5,000m under 15minutes at age 15 is independent of nationality of participant.

Alternative Hypothesis, H_1: The ability to run 5,000m under 15minutes at age 15 is not independent of nationality of participant.

```
        Pearson's Chi-squared test

data:  tbl
X-squared = 2.1429, df = 1, p-value = 0.1432

> alpha <- 0.05
> x2.alpha <- qchisq(alpha,df = 1,lower.tail = FALSE)
> print(x2.alpha)
[1] 3.841459
> output <- chisq.test(tbl,correct = FALSE)
> output$statistic
X-squared
 2.142857
> output$observed
```

Figure 6: The calculation results generated using RStudio including X-squared, degree of freedom and p-value

```
> output$observed

      A NA
NO    5  9
YES 10  6
> output$expected

     A NA
NO   7  7
YES 8  8
```

Key: A=African, NA=Non-African, Yes=capable, No= Not capable

Figure 7: Two-ways contingency tables generated using RStudio including observed frequency and expected frequency

Using alpha=0.05, and degree of freedom calculated from the formula (row-1)(column-1), we get degree of freedom equals to 1. The test statistic of chi-square test of independence is given by

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

The test statistic calculated by RStudio is 2.14285 whereas the critical region is >3.841459.

Conclusion from test:

Since the test statistic calculated is 2.14285 does not fall within the critical region $X^2_{critical}$>3.841459, we do not have sufficient evidence to reject the null hypothesis. Thus, we

conclude that the ability to run 5,000m under 15minutes at age 15 is independent of nationality of participant at alpha=0.05.

4.0 Conclusion

What have you learned from all activities done in Project 2 (choosing dataset, pre-processing and analysis process etc.)?

We have learned quite a lot of things from all the activities done in project 2, things such as having to select a group of data and be able to state results based on calculations performed on the dataset. We exposed to the practical way of carrying out various statistical tests on our favorite topic and applying the knowledge of inferential statistics. We also get to learn basics of R programming language that we deemed important in our future career.

What is your best/interesting findings from your results?

From our calculations and data analyses in 2-sample hypothesis testing, we have discovered that the recent invention of the Nike Vaporfly shoes that consist of carbon-plate had helped most elite runners in marathon running over the years. The carbon-plate shoes made the runners a smoother and bouncy running over the course. The imminent Tokyo Olympics 2020 even considering prohibiting it as it will give runners to gain unfair advantages. From correlation and regression test, we discovered that miles ran per week do indeed have considerable impact on overall marathon performances. However, we believed there are other factors that contributed to faster marathon running times and it will be highly individual, meaning not a single kind of medicine could be suitable to all the patients. In chi-square test of independence, the myth over the decades that running is being dominated by East African runner was proven wrong in our project. However, the limitations of this project are small sample size and chosen value of alpha. Therefore, further research are required to test those claims.

5.0 Appendix

- Sample of our original/raw dataset:
  (Source: https://www.bmw-berlin-marathon.com/en/impressions/statistics-and-history/results-archive/)
  Berlin Marathon 2019 results from top 1 to 30 finishers:

Place	First Name	Last Name	Nationality	#SN	Club	Sex	Net	Gross	Photo	Certificate
Search	Search	Search	Search	Search	Search	Search	Search	Search		
1	Kenenisa	Bekele	ETH	2	Äthiopien	M	02:01:41	02:01:41		
2	Birhanu	Legese	ETH	5	Äthiopien	M	02:02:48	02:02:48		
3	Sisay	Lemma	ETH	4	Äthiopien	M	02:03:36	02:03:36		
4	Jonathan	Korir	KEN	7	Kenia	M	02:06:45	02:06:45		
5	Felix	Kandie	KEN	6	Kenia	M	02:08:07	02:08:07		
6	Yohanes	Gebregergish	ERI	8	Eritrea	M	02:08:26	02:08:26		
7	Guojian	Dong	CHN	19	China	M	02:08:28	02:08:28		
8	Bethwel	Yegon	KEN	10		M	02:08:35	02:08:35		

- Processed dataset:

The marathon finish times were converted to minutes and transferred to excel as shown:

	A	B
1	G2018	G2019
2	121.65	121.68
3	126.38	122.8
4	126.8	123.6
5	128.27	126.75
6	128.77	128.12
7	129.3	128.43
8	129.93	128.47
9	131.12	128.58
10	132.95	128.93
11	133.15	129.65
12	133.72	130.43
13	134.95	130.95
14	135	131.13
15	135.07	131.23
16	135.32	131.83
17	135.62	132.48
18	136.43	132.68
19	136.47	133.32
20	136.65	133.37
21	137.48	133.63
22	137.8	133.65
23	139.1	133.8
24	139.23	133.85

Figure: The carbon-plated Nike running shoes (Nike Vaporfly Next Percent) that believed to have broken multiple marathon world records
Source: Google Search

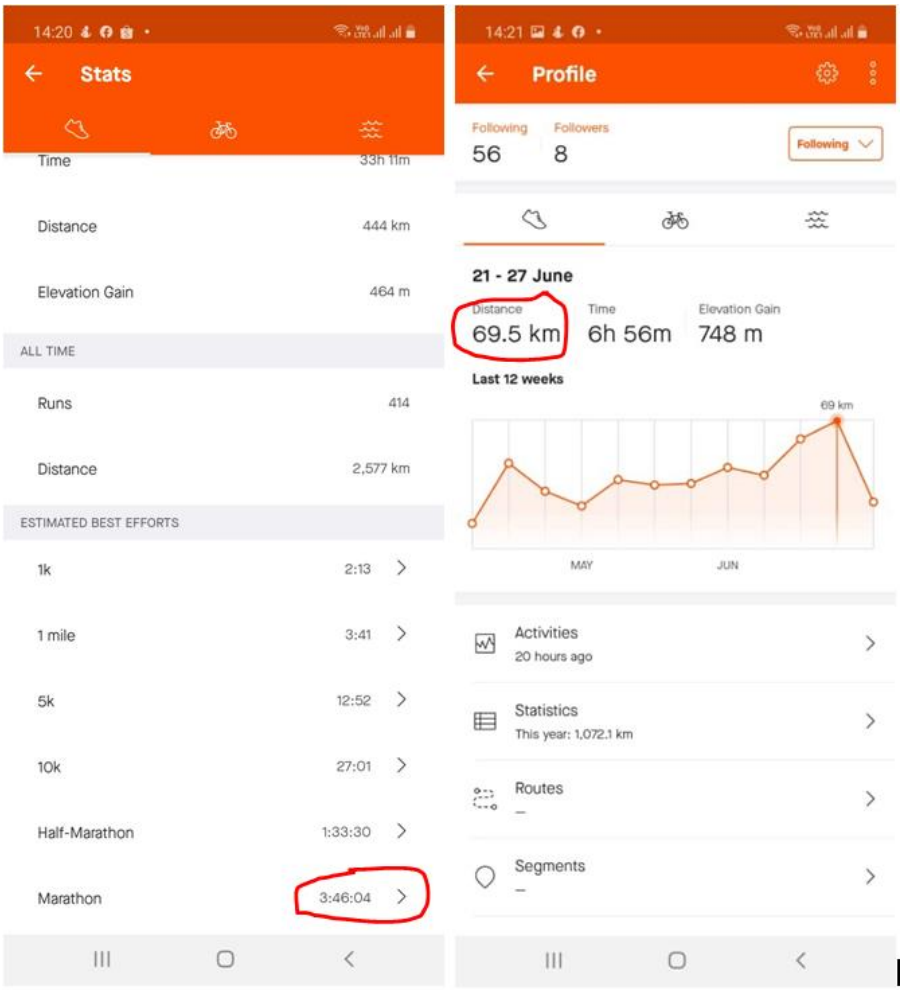

Figure: Strava fitness tracking apps that shows a sample of marathon best effort (left) in hour: minute: second and weekly running mileage (right) in kilometers

30 random local male runners were observed and the data were transferred into excel as shown:

mpw	performance
6	307.3
8	330.34
13	254.31
27	242.13
37	207.43
60	204.58
32	247.83
60	228.86
44	221.18
62	183.5
30	198.68
10	262.6
37	201.13
50	187.68
16	231.8
26	224.32
40	213.53
10	290.37
5	361.39
75	182.5
20	220.95
15	245.65
20	211.14

Figure: Excel containing weekly mileage (converted to miles) and marathon performance (converted to minutes)

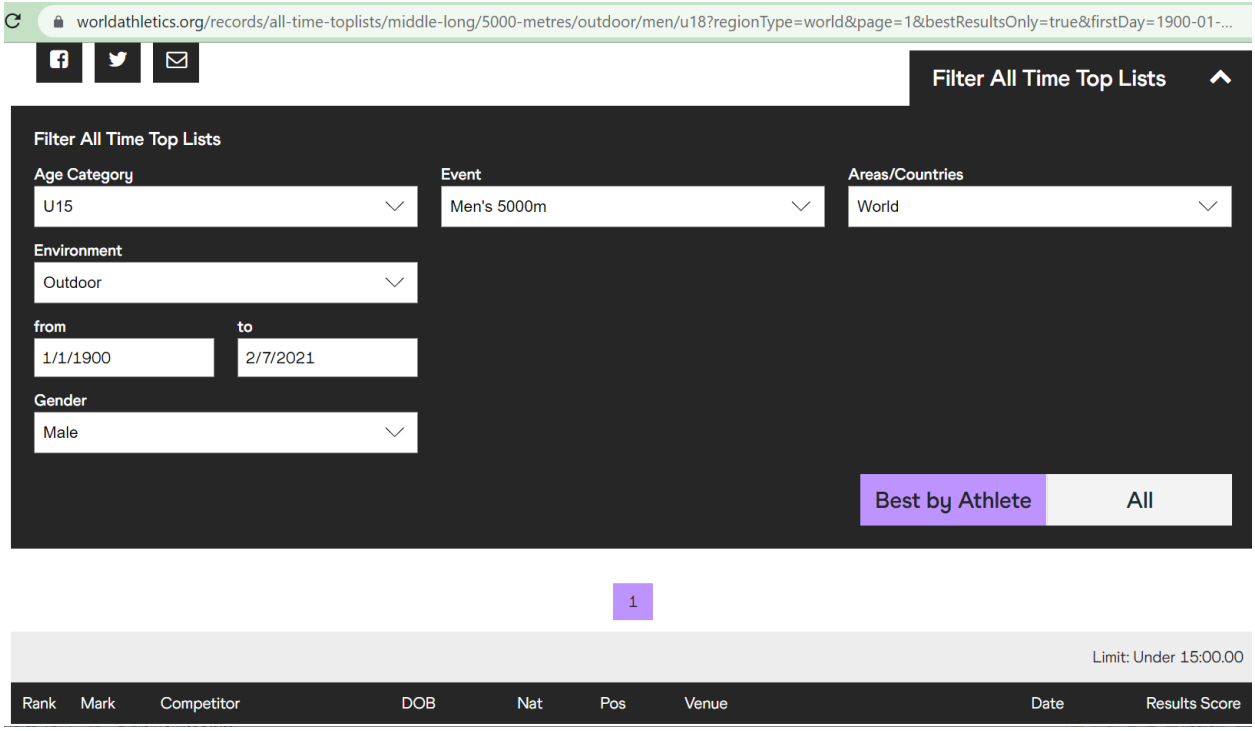

Figure: Raw data to collect male runners under 15 years old who capable to run 5000m under 15 minutes from all nations

81	13:49.9h	Solomon BUSIENEI	1984	KEN	2	Réduit (MRI)	18 AUG 2001	1034
82	13:50.07	Emmanuel KIPRONO	14 JUL 1995	KEN	2	Kericho (KEN)	05 JUL 2012	1033
83	13:50.27	Solomon MOLLA	20 JAN 1987	ETH	2	Argentan (FRA)	09 JUN 2003	1032
83	13:50.27	Hiroto YOSHIOKA	18 MAY 2004	JPN	3f18	NITTAIDAI Athletic Stadium, Yokohama (JPN)	15 NOV 2020	1032
85	13:51.58	Andrew LESUUDU	15 APR 1991	KEN	6	Tanger (MAR)	13 JUL 2008	1028
86	13:51.89	John Gathaiya MURUGU	01 JUL 1995	KEN	8f1	Hiroshima (JPN)	28 APR 2012	1027
87	13:52.02	Joseph MUIGAI	06 AUG 2004	KEN	3f9	Athletics Stadium, Konosu (JPN)	29 APR 2021	1026
88	13:52.22	Peter RONO	31 JUL 1967	KEN	1	Rieti (ITA)	31 AUG 1980	1026
89	13:52.37	Mike KIPRUTO	2001	KEN	4	Doha (QAT)	27 APR 2016	1025
90	13:52.43	Okubamichael FISSEHATSION	15 SEP 1997	ERI	3	Oordegem (BEL)	05 JUL 2014	1025
91	13:52.58	Titus Kwemoi MASAI	09 OCT 1989	KEN	3	Nairobi (KEN)	22 MAY 2004	1024
92	13:53.28	Bernard Kipkirui LANGAT	10 OCT 2001	KEN	10	Nairobi (KEN)	11 JUN 2018	1022
93	13:53.31	Rodgers KIBET	21 FEB 2005	UGA	8f2	Mandela National Stadium, Kampala (UGA)	08 MAY 2021	1022
94	13:54.29	Henry KIRUI	05 APR 1972	KEN	1	Laurentian University Stadium, Sudbury (CAN)	30 JUL 1988	1019
95	13:54.44	Andrew KWEMOI	15 JUN 2002	KEN		Mandela National Stadium, Kampala (UGA)	26 JUL 2019	1018
96	13:54.6h	Edward KIBET	2001	KEN	3	Eldoret (KEN)	18 MAY 2017	1018

Figure: Sample of the male athletes displaying their respective nations and their finishing time for 5,000m

	Participan	under15
2	NA	YES
3	NA	NO
4	NA	NO
5	NA	YES
6	NA	NO
7	NA	NO
8	NA	YES
9	NA	NO
10	NA	YES
11	NA	YES
12	NA	NO
13	NA	NO
14	NA	NO
15	NA	YES
16	NA	NO
17	A	YES
18	A	YES
19	A	NO
20	A	NO
21	A	YES
22	A	YES
23	A	YES
24	A	YES

Figure: 30 random runners were picked including 15 non-African(NA) and 15 African(A) and whether they are capable of sub 15 5,000m is observed.

6.0 References

- **https://www.worldathletics.org/records/all-time-toplists/middlelong/5000-metres/outdoor/men/senior**

- **https://www.bmw-berlin-marathon.com/en/impressions/statistics-and-history/results-archive/**

- **https://runrepeat.com/research-marathon-performance-across-nations**

- **https://www.runnersworld.com/races-places/a20813340/the-incredible-shrinking-marathoner/**

- **https://www.kaggle.com/rojour/boston-results?select=marathon_results_2015.csv**

- **https://towardsdatascience.com/analysing-elite-running-performance-with-historical-data-b9c6bdd9c5d8**