Uncovering patterns in the 2019 KBC Dublin marathon

Daniel Rodriguez Ribera
School of Computing
National College of Ireland
Dublin, Ireland
x18142907@student.ncirl.ie

I. INTRODUCTION & SPECIFICATION OF THE HYPOTHESIS

The Dublin marathon is a sports event that takes place every year in Dublin city centre, on the last Sunday of October. There were 17,732 runners in this year's edition. The runners are classified in categories according to gender and age even though they all run at the same time.

Their overall position in the event is recorded regardless of their category. At the same time, runners are ranked on the position achieved within their category and obtain prize and recognition according to that.

This research will attempt to single out the factors that contribute to the performance of participants in the past 2019 KBC Dublin marathon.

Age categories are part of this study as well as a variable that accounts for whether participants are members of a running club or not. More detail regarding data will be provided later sections of this paper.

As a frequent runner in races of shorter distance and being a volunteer in this year's marathon himself the inspiration for this research springs from the author's own desire to run a marathon in the future.

Every runner looks up to the people who attempt to complete this challenge and the analysis presented in this paper will look into the key performance aspects that assure success in marathons.

This paper will try to answer the following research questions:

- RQ1: What is the best age category to run the Dublin Marathon?
- RQ2: Does joining a running club affect performance?
- RQ3: What race segment is key for a runner's success in the Dublin marathon?

By studying these three questions, the aim is not to predict but to give insights as to how runners succeeded in achieving the best overall position they could in the past edition of the Dublin marathon.

II. METHODOLOGY & BACKGROUND OF THE DATASET

The methodology used to tackle this research is Knowledge Discovery in Databases proposed in [1].

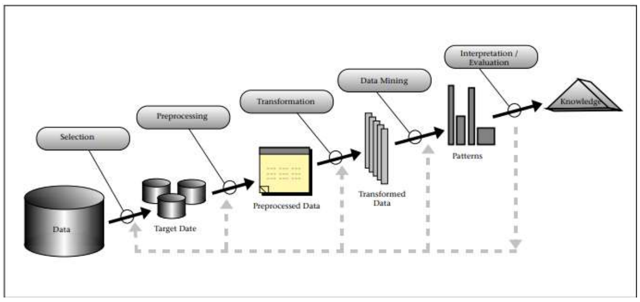


Fig. 1. An Overview of the Steps That Compose the KDD Process [1, Fig 1.]

1) *Data selection.*

This paper will study the data provided by the KBC Dublin Marathon 2019 website [2] that states attributes such as the timing achieved by every participant in a set of four sections that made up the 42.195 KM course.

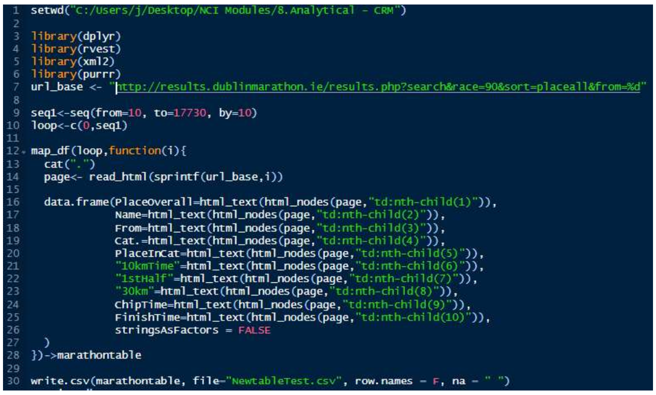
Other websites offer information about the Dublin marathon but the official website is the golden source.

Also this official website [3] has data available from past editions of the marathon but to avoid using time series, all data used in this regression has been collected from the year 2019. In every variable studied here, each observation represents one runner.

2) *Data pre-processing. Tasks:*

a) Web scraping

The data is not available in CSV or any Excel format therefore the table had to be obtained using a web scraping technique. The author's tool of choice for this task was RStudio. The link to the precise web page scraped is in [4]. "Fig.2" shows the code used for the web scraping technique.



```
setwd("C:/Users/j/Desktop/NCI Modules/8.Analytical - CRM")

library(dplyr)
library(rvest)
library(xml2)
library(purrr)
url_base <- "http://results.dublinmarathon.ie/results.php?search&race=90&sort=placeall&from=%d"

seq1<-seq(from=10, to=17730, by=10)
loop<-c(0,seq1)

map_df(loop,function(i){
  cat(".")
  page<- read_html(sprintf(url_base,i))

  data.frame(PlaceOverall=html_text(html_nodes(page,"td:nth-child(1)")),
             Name=html_text(html_nodes(page,"td:nth-child(2)")),
             From=html_text(html_nodes(page,"td:nth-child(3)")),
             Cat.=html_text(html_nodes(page,"td:nth-child(4)")),
             PlaceInCat=html_text(html_nodes(page,"td:nth-child(5)")),
             "10kmTime"=html_text(html_nodes(page,"td:nth-child(6)")),
             "1stHalf"=html_text(html_nodes(page,"td:nth-child(7)")),
             "30km"=html_text(html_nodes(page,"td:nth-child(8)")),
             ChipTime=html_text(html_nodes(page,"td:nth-child(9)")),
             FinishTime=html_text(html_nodes(page,"td:nth-child(10)")),
             stringsAsFactors = FALSE
  )
})->marathontable

write.csv(marathontable, file="NewtableTest.csv", row.names = F, na = " ")
```

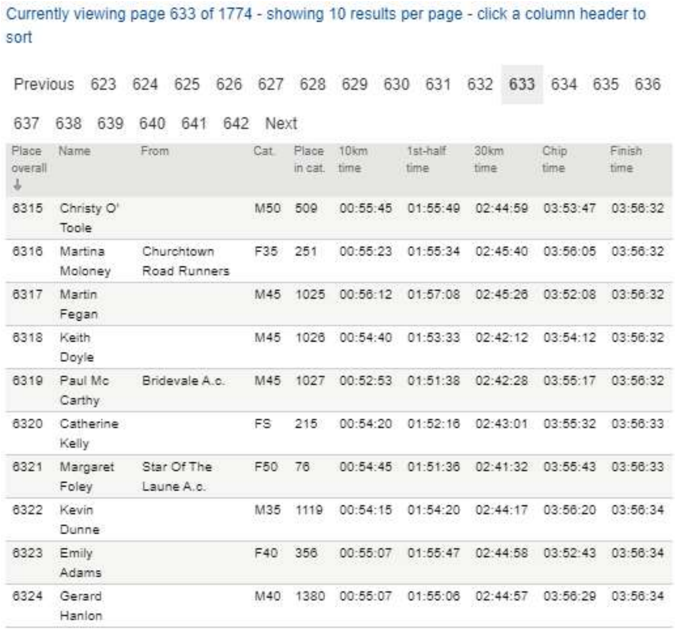
Fig. 2. RStudio web scraping code.

Table I. below shows the names and meaning of the variables of the table scraped.

TABLE I. ORIGINAL VARIABLES

Variable	Description
Place Overall	Place achieved by the runner according to Finish Time.
Name	Name and surname of the runner
From	Name of the club in which the runner is registered. This value is blank if the runner is not registered in any club.
Cat.	• "FS" and "MS" refer to Female and Male senior. These are runners aged 19-34. • The rest of categories describe Gender plus age in within intervals of 5 years, e.g. a 37 year-old man will be classified as "M35". A 42 year-old woman will be under the "F40" category. • "FU19" and "MU19"stand for Female and Male Under 19 years old.
Place in Cat.	Place within runners of same category by Finish Time.
10Km time	Time in minutes from the start till the 10Km mark.
1st-Half time	Time in minutes from the start till Half-Marathon mark (21.0975 Km).
30Km time	Time from the start until 30km mark.
Chip Time	Time it took the runner since he crossed the starting line till he crossed the finishing line.
Finish Time	Time it took the runner since the official time the race started until he/she crossed the finishing line. It's different from Chip Time since not all participants cross the starting line at the same moment.

"Fig.3" reveals a snip of the table scraped. It is a 10x10 table: 10 rows and equal number of columns that replicates itself 1,774 times in order to capture every runner's performance.



Currently viewing page 633 of 1774 - showing 10 results per page - click a column header to sort

Previous 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 Next

Place overall ↓	Name	From	Cat.	Place in cat.	10km time	1st-half time	30km time	Chip time	Finish time
6315	Christy O' Toole		M50	509	00:55:45	01:55:49	02:44:59	03:53:47	03:56:32
6316	Martina Moloney	Churchtown Road Runners	F35	251	00:55:23	01:55:34	02:45:40	03:56:05	03:56:32
6317	Martin Fegan		M45	1025	00:56:12	01:57:08	02:45:26	03:52:08	03:56:32
6318	Keith Doyle		M45	1026	00:54:40	01:53:33	02:42:12	03:54:12	03:56:32
6319	Paul Mc Carthy	Bridevale A.c.	M45	1027	00:52:53	01:51:38	02:42:28	03:55:17	03:56:32
6320	Catherine Kelly		FS	215	00:54:20	01:52:16	02:43:01	03:55:32	03:56:33
6321	Margaret Foley	Star Of The Laune A.c.	F50	76	00:54:45	01:51:36	02:41:32	03:55:43	03:56:33
6322	Kevin Dunne		M35	1119	00:54:15	01:54:20	02:44:17	03:56:20	03:56:34
6323	Emily Adams		F40	356	00:55:07	01:55:47	02:44:58	03:52:43	03:56:34
6324	Gerard Hanlon		M40	1380	00:55:07	01:55:06	02:44:57	03:56:29	03:56:34

Fig. 3. Example of the table scraped that spans 1,774 pages.

b) Dealing with missing values (NAs)

There were only 378 missing values in the original dataset out of 177,320 cells. This entails that runners that started the race but did not finish it. Also, instances where the chip used to record the runner's time malfunctioned.

The ratio of NAs is around 0.2% so they were removed from the dataset as such a small percentage of observations is highly unlikely to change any results.

3) Data transformation

a) Gender component is removed

There were 23 original variables that combined the participant's gender and age as explained in Table I.

This study is not concerned with the gender attribute and was removed. Consequently 12 categories remained:

- Under19 -> "Age18"
- Seniors -> "Age19-34": they will be used as baseline for comparison against other groups.
- People aged 35-39 -> "Age35

The rest of categories go on on the same way until "Age80" the oldest category with only four participants.

b) Dummy encoding

Dummy encoding was performed on these twelve new categories of the attribute age.

k-1 = 11 new binary variables were created. The baseline for this was the Senior group (age19-34). This means that if a participant belongs to the senior category then all the dummy variables are zero [5, p.510].

c) New variables to reflect the running Pace

The race is split in four sections of slightly different distance as seen in "Fig.4". The way to overcome this issue so that they can all be comparable is by "normalizing" them.

The way in which this was achieved was to create new variables that accounted for the average Pace per kilometre at which the runner had ran the section measure in minutes.

These new variables are named "MinPace1", "MinPace2", "MinPace3" and "MinPace4".

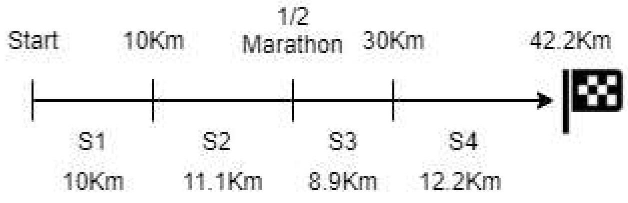


Fig. 4. Sections in which the marathon is divided.

4) Data mining.

5) Interpretation of results and knowledge discovery.

This will be addressed in the final section of this paper.

III. RESEARCH AND INVESTIGATION INTO SUITABLE TECHNIQUES

This research considered the following techniques for implementation:

A. Principal Components Analysis (PCA)

K. Tsiptsis and A. Chorianopoulos [6, p.66] consider Principal Components Analysis as "a statistical technique used to reduce the data of the original input fields. It derives a limited number of compound measures that can efficiently substitute for the original inputs while retaining most of their information."

PCA therefore is used for large datasets with many variables that need to be slimmed down in order to perform the analysis. This is done by studying the linear correlation among variables and only keeping those that give us more information.

Principal Components Analysis is an unsupervised technique suitable for numeric continuous fields that does not handle well categorical data [6, pp.67].

The statement above is the reason why although PCA was considered, it finally was deemed not appropriate to be part of the present study. In the present dataset, there are two important categorical variables: "age category" and "gender" therefore a technique that is not suitable for this type of data would not be the best choice.

B. Clustering

There are several clustering techniques (k-means being the most popular) but they are all essentially unsupervised machine learning techniques. This means that they are not suitable for analysis where there is a target variable.

K-means, TwoStep and Kohonen networks are different clustering algorithms and such they produce best results with numeric input fields. Although they can handle categorical inputs, it is widely advised to avoid using clustering with this type of data [6, pp.83].

The present analysis is focused on a clear target variable as "Place overall" in the race. This along with the fact that it is not advisable to use clustering with categorical variables.

Clustering analyses the data input creating groups of observations (clusters) where data points are more alike within the same group.

In brief, clustering has thus been ruled out as a technique to perform this analysis for the below reasons:

- Clustering does not handle well categorical data. Two of the attributes in the dataset" "age category" and "gender" are according to experience two strong candidate variable to explain the variance of the result in the race.
- Not a suitable technique when there is a target variable. In this case, the variable "Place Overall" in the race is the dominant variable that we are trying to get a better sense of.

C. Classification - Decision trees

Classification is a technique used to analyse datasets in order to predict categorical (discrete, unordered) class labels.

Classification is a two steps process: in the first step, a classification model is built based on previous data. In the second step, the accuracy of the model is tested and if the results are acceptable, then used to classify new data. [7, p.327].

This technique is frequently used in risk analysis where financial institutions want to predict and assign levels of risk to their investments or liabilities.

One of the most used algorithms for classification tasks is the decision tree algorithm in any of its variations: ID3, C4.5 or CART.

That being said, predictions fall out of the scope of this project and that's why classification was not used for the analysis presented here.

However, while studying the dataset of the 2019 Dublin marathon it was perceived the following two scenarios where an analysis based on classification could be employed on this dataset:

- In-race drop-out rate. There are a number of observations that are recorded on the first 10K, half-marathon point of 30K mark but do not have a chip time or finish time recorded. This refers to runners that started the marathon but did not make it until the end of the race. The proportion of such runners a little over 1% [8].
- No-shows. Official registration figures go up to 22,500 runners but a little less than 18,000 actually started the run on that day. This means there is a 20% no-show rate [8].

The above two ratios make up for a perfect scenario where a classification algorithm could be deployed in order to predict runners that either will not show on race day or will drop out in the middle of it.

Future work should could look into the above in order to perform a classification analysis and study the causes of these issues. The no-show ratio would be specially interesting since the popularity of the Dublin marathon has grown in recent years to the point where the entry of participants of the 2020 edition will be based on a lottery [9].

In any case, richer data will be needed in order to conduct the outlined analysis.

D. Support Vector Machines (SVM)

"Support Vector Machines" is a supervised machine learning algorithm which can be used for both classification or regression purposes. However, it is mostly applied to classification problems.

A SVM attempts to achieve maximum separation (margin) between classes. The difference with other techniques is that SVM can work with non-linear data.

"SVM is an algorithm that uses nonlinear mapping to transform the original training data into a higher dimension. Within this dimension, it searches for the linear optimal separating hyperplane (i.e., a "decision boundary" separating the tuples of one class from another). With an appropriate nonlinear mapping to a sufficiently high dimension, data form two classes can always be separated by a hyperplane" [7, pp.410].

However, SVM are suitable for binary classification problems and when used in regression, it is used in a logistic linear regression where the outcome is binary as well.

The way to overcome the mentioned shortcoming is by performing dummy encoding on the several categories. This workaround is simple to implement but it is not optimal if there is no dominant category [10].

The data used for this research assumes runners aged “19-24” as the dominant category so a priori SVMs could be a candidate technique to be implemented.

In spite of that, SVMs are said to be a black box method: very hard to understand and also suffer from the curse of dimensionality [10].

Although SVMs could be used in this analysis, ultimately I have favoured the next technique for its simplicity and effectiveness.

E. Linear Regression

This technique was ultimately chosen due to its simplicity and power to describe relationships. A closer look to this technique will be taken in the next section.

IV. Implementation of the chosen technique & The technique(s) employed

A. Theory behind linear regression

The objective of this linear regression is to find a model that correctly describes the relationship between the dependent variable (Y_i) from one or more independent variables (x_i) and their related parameters, b_1. Regression also accounts for some error associated with that model (ε_i).

This model differs from that of a correlation only in that it uses an *unstandardized* measure of the relationship (b_1) and consequently we include a parameter, b_0, that tells us the value of the outcome when the predictors are zero [5, p.371].

Linear regression will be called simple linear regression if there is only one independent variable and multiple linear regression if there are two or more. In any case, the outcome will be “equation (1)” as follows:

$$Y_i= (b_0 + b_1x_{1i} + b_2x_{2i} +\ldots+ b_{k-1}x_{k-1} + b_kx_k)+\varepsilon_i \quad (1)$$

where K is the number of independent variables.

In this research, implementations of both simple and multiple linear regressions will be performed in following sections of this paper.

B. Dependent variable: object of study

This section goes into detail about the independent variables included in the implementation according to past work, common sense and the author’s choice.

In the same way, the reasoning to rule out other independent variable is presented.

It’s important to remark once again that the dependent variable throughout this research will be “Place Overall in the 2019 Dublin Marathon” (Yi).

This variable is understandably highly correlated to the variable “Finish Time”, being its Pearson correlation value equal to (0.953). Pearson's correlation measure “p” can range from -1 to 1. [11]

- p = -1 indicates a perfect negative linear relationship between variables.
- p = 0 indicates no linear relationship between variables.
- p = 1 indicates a perfect positive linear relationship between variables.

Given the concepts above, explaining the relationship between “Finish Time” and “Place Overall” does not have any statistical value.

Instead, this research will focus on modelling the relationship of dependent variable “Place Overall” according to the following independent variables:

1) Independent variables considered:

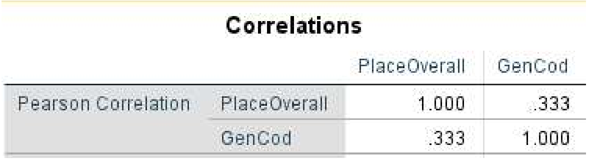
a) “Age”

b) “ClubOrNot”

c) Pace(min), kept across different sections of the race: “MinPace1”, “MinPace2”, “MinPace3”and “MinPace4”

2) Variables excluded:

a) “Gender”. The level of correlation between “Gender” and “Place Overall” is not worrisome, only 0.333. This means that Gender is a valid Independent Variable to include in the model. However, the author has deemed more appropiate to base the implementation on other variables and perhaps gender could be object of future study.



Correlations

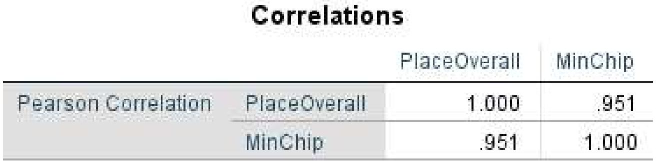
		PlaceOverall	GenCod
Pearson Correlation	PlaceOverall	1.000	.333
	GenCod	.333	1.000

Fig. 5. Pearson correlation coefficient “Place Overall” vs. “Gender”.

b) “Chip time”. As outlined before, this is the personal time achieved by every runner since they cross the starting line until they cross the finishing one.

“Place Overall” ranks runners according to “Finish time”but “Chip Time”just differs from the first in just a matter of few seconds or minutes depending on how long took the runner to cross the starting line since the official start time.

Consequently, “Chip Time” and the dependent variable are highly correlated as we see below.



Correlations

		PlaceOverall	MinChip
Pearson Correlation	PlaceOverall	1.000	.951
	MinChip	.951	1.000

Fig. 6. Pearson correlation coefficient “Place Overall” vs. “Chip Time(Min)”.

C. *Tools: IBM SPSS Statistics vs. RStudio*

The tool used to implement the linear regression technique was IBM SPSS Statistics. The next option for the implementation would have been the open source RStudio software that was already used to perform the web scraping.

Familiarity with SPSS regarding previous work along with a more interactive and user friendly interface [12] tipped the scales in SPSS's favour.

D. *Method of entering predictors into the model*

Having chosen the predictors, the order in which they are entered into the model needs to be specified. This research is made up of two regressions: one multiple and one one simple.

- *Multiple linear regression*

Hierarchical regression method for input is applied here. The predictors are entered into the model in order of importance based on past research [5, p.398]. The order of input is as follows:

1) "*Age*". However, this has been dummy encoded resulting in 12 new binary categories. All dummy variables must be entered into the model at once [5, p.509].

2) "*ClubOrNot*". This will represent the second block included in our model.

- *Simple linear regression*

This implementation has been deemed necessary to establish which section out the four is most important in the marathon.

The unit of measure used for this has been "MinPace1" to "MinPace4". Using a multiple regression in this case was not possible due to the high collinearity that this produced.

If a top runner runs at a pace of 3min per Km in section 1, it is very unlikely that that pace drops to 8min per Km in section 4 unless due to illness or extreme fatigue.

Pace across sections 1 to 4 varies, of course, for every runner studied. Nevertheless, these measures are correlated beyond doubt on account of the reason given in the previous paragraph.

Instead of a multiple linear regression, four simple linear regressions were performed to assess the suitability of "MinPace1" to "MinPace4" to explain success in the Dublin marathon event.

V. The Findings, in Terms of Both Quantitative Results and Their Qualitative Interpretation

This section will present the results of the implementation described before in relation to the research questions posed at the beginning.

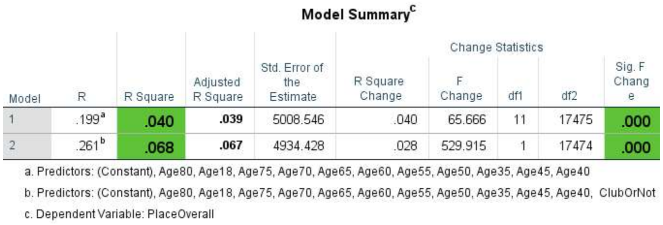
A. *RQ1: What is the best age category to run the Dublin Marathon?*

The original 23 runner categories were stripped of the gender component, as it is not a part of this analysis. After this 12 age categories remained and dummy encoding was performed thus leaving k-1 = 11 binary variables.

One of the categories acts as a baseline against which all other groups will be compared: this is the "19-34" age bracket.

1) Using age categories does explain success in the marathon.

- We can see in "Fig.7" that the p-value is 0.00 which confirms that this model using age is significantly better that having no model at all.
- R-Square is set at 0.04 i.e. 4% of the variance in the change of "Place Overall" in the Dublin marathon is explained by the age category at which the participants belonged.



Model Summary[c]

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.199[a]	.040	.039	5008.546	.040	65.666	11	17475	.000
2	.261[b]	.068	.067	4934.428	.028	529.915	1	17474	.000

a. Predictors: (Constant), Age80, Age18, Age75, Age70, Age65, Age60, Age55, Age50, Age35, Age45, Age40

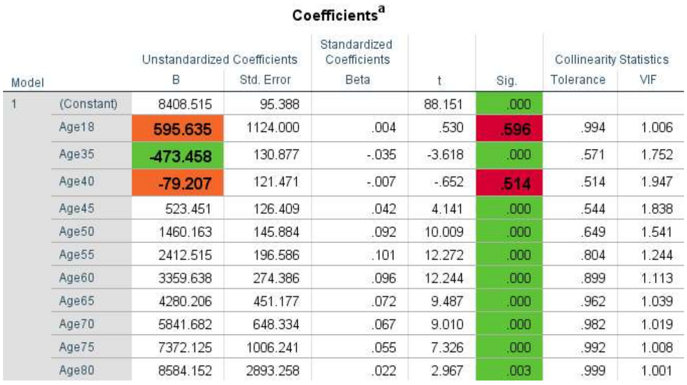
b. Predictors: (Constant), Age80, Age18, Age75, Age70, Age65, Age60, Age55, Age50, Age35, Age45, Age40, ClubOrNot

c. Dependent Variable: PlaceOverall

Fig. 7. Model summary for implementation needed in RQ1 and RQ2 .

2) Comparison of age categories vs. "19-34" base group.

- Category "Age35" that groups runners aged 35-39 has a negative coefficient -473.45. This means that runners in this category are likelier to achieve a better place in the race than those aged "19-34".
- The p-value associated to the t-test for the "Age35" is significant (0.00). This means that being in the "Age35" does improve the chances of qualify better than runners.
- "Age40" also produces a negative coefficient but in this case the p-value associated is higher than 0.05 therefore the test is not significant.
- The above means there is no evidence to suggest that runners in "Age40" have a different performance from those aged 19-34. This is a hopeful result for those worried that getting older decreases their chances of performing well.
- Similarly, "Age18" has a positive coefficient but a non-significant p-value (as it is higher than 0.05). This means that there is no evidence to suggest that runners in "Age18" perform differently from those of 19-34.



Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	8408.515	95.388		88.151	.000		
	Age18	595.635	1124.000	.004	.530	.596	.994	1.006
	Age35	-473.458	130.877	-.035	-3.618	.000	.571	1.752
	Age40	-79.207	121.471	-.007	-.652	.514	.514	1.947
	Age45	523.451	126.409	.042	4.141	.000	.544	1.838
	Age50	1460.163	145.884	.092	10.009	.000	.649	1.541
	Age55	2412.515	196.586	.101	12.272	.000	.804	1.244
	Age60	3359.638	274.386	.096	12.244	.000	.899	1.113
	Age65	4280.206	451.177	.072	9.487	.000	.962	1.039
	Age70	5841.682	648.334	.067	9.010	.000	.982	1.019
	Age75	7372.125	1006.241	.055	7.326	.000	.992	1.008
	Age80	8584.152	2893.258	.022	2.967	.003	.999	1.001

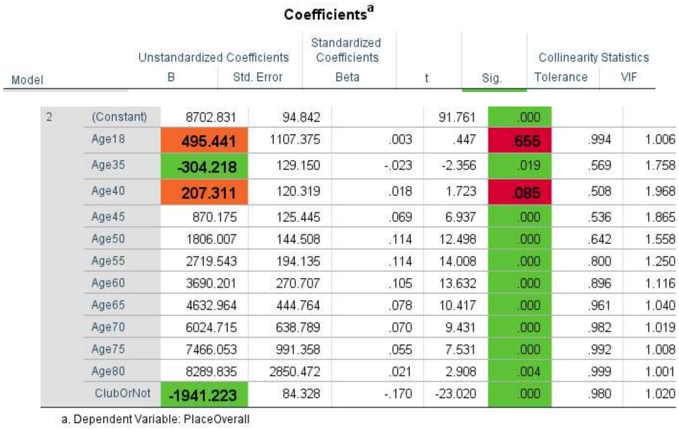
Fig. 8. B-values obtained in model 1 and their related p-values.

B. *RQ2: Should you join a running club?*

"Fig.7" above shows as well the results of the second multiple linear regression used to research this question.

- The p-value is significant so the model does explain the place of the runner in the race and the R-Square value jumps up to 0.068 in this second model.
- "Fig. 9" shows the B-coefficients and p-values of this second model. In general terms, the "ClubOrNot" variable does not alter the findings stated after model 1 regarding age groups and performance.
- As we can see t-tests for "Age18" and "Age40" are not significant which suggests that being a runner of this group does not imply a different performance from that of our baseline group: "19-34".

We can conclude that there is evidence to suggest that joining a running club does improve runner performance in the Dublin city marathon.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
2	(Constant)	8702.831	94.842		91.761	.000		
	Age18	495.441	1107.375	.003	.447	.655	.994	1.006
	Age35	-304.218	129.150	-.023	-2.356	.019	.569	1.758
	Age40	207.311	120.319	.018	1.723	.085	.508	1.968
	Age45	870.175	125.445	.069	6.937	.000	.536	1.865
	Age50	1806.007	144.508	.114	12.498	.000	.642	1.558
	Age55	2719.543	194.135	.114	14.008	.000	.800	1.250
	Age60	3690.201	270.707	.105	13.632	.000	.896	1.116
	Age65	4632.964	444.764	.078	10.417	.000	.961	1.040
	Age70	6024.715	638.789	.070	9.431	.000	.982	1.019
	Age75	7466.053	991.358	.055	7.531	.000	.992	1.008
	Age80	8289.835	2850.472	.021	2.908	.004	.999	1.001
	ClubOrNot	-1941.223	84.328	-.170	-23.020	.000	.980	1.020

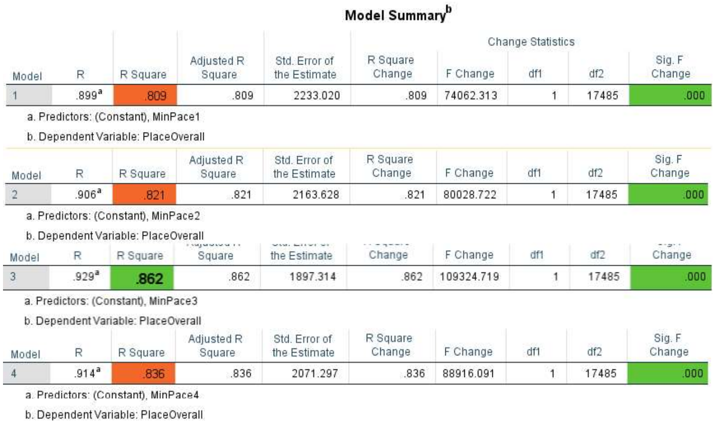
a. Dependent Variable: PlaceOverall

Fig. 9. B-values obtained in model 2 and their related p-values.

C. *RQ3: What race segment is key for a runner's performance in the Dublin marathon?*

Using the pace per minute at which the section is ran, four simple linear regressions were performed in order to find out which section of the race carried the most importance.

- The results in "Fig.10" show that it is actually the pace in section 3 (from the Half-marathon until the 30Km mark) is the one that accounts better for the final performance in the race.
- R-Square in model 3 is the highest. "MinPace3" is therefore the variable that explains best success in the marathon out of the four sections.



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.899[a]	.809	.809	2233.020	.809	74062.313	1	17485	.000

a. Predictors: (Constant), MinPace1

b. Dependent Variable: PlaceOverall

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
2	.906[a]	.821	.821	2163.628	.821	80028.722	1	17485	.000

a. Predictors: (Constant), MinPace2

b. Dependent Variable: PlaceOverall

Model	R	R Square	Square	the Estimate	Change	F Change	df1	df2	Change
3	.929[a]	.862	.862	1897.314	.862	109324.719	1	17485	.000

a. Predictors: (Constant), MinPace3

b. Dependent Variable: PlaceOverall

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
4	.914[a]	.836	.836	2071.297	.836	88916.091	1	17485	.000

a. Predictors: (Constant), MinPace4

b. Dependent Variable: PlaceOverall

Fig. 10. Summary models of four single linear regressions performed solely with the variable Pace per section in Min.

Marathoners sometimes have an in-race shock known as "hitting the wall" in runner's speak. This usually occurs within the last few kilometres of the race due to improper nutrition or lack of training.

Popular belief in line with this fact thinks that Section 4 is the most important in order to obtain a better result in the race but this simple linear regression has disproved this.

The pace at which section 3 is ran carries more importance than section 4.

VI. References

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining To Knowledge Discovery in Databases," *Artificial Intelligence Magazine, vol.* 17, no.3, pp.37–51,1996.

[2] [Online]. Available: http://kbcdublinmarathon.ie/ [Accessed on: Nov. 20, 2019].

[3] [Online]. Available: http://kbcdublinmarathon.ie/results-certificates/ [Accessed on: Dec. 1, 2019].

[4] [Online]. Available:http://results.dublinmarathon.ie/results.php?search&race=90&sort=placeall&from=0 [Accessed on: Nov. 20, 2019].

[5] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 5th ed, London:Sage Publications Ltd, 2018.

[6] K. Tsiptsis and A. Chorianopoulos, *Data Mining Techniques In CRM: Inside Customer Segmentation*, London: Wiley, 2009.

[7] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts And Techniques*, 3rd ed., London:Morgan Kaufmann, 2012.

[8] [Online]. Available: https://medium.com/running-with-data/kbc-dublin-marathon-analysis-8650782cf703 [Accessed on: Dec. 3, 2019].

[9] [Online]. Available: https://www.runireland.com/lottery-entry-dublin-marathon/ [Accessed on: Dec. 2, 2019].

[10] M. Tova-Izquierdo, "NCI Advanced Data Mining Lecture Notes: Support Vector Machines", 2018. [Online]. Available: https://moodle2018.ncirl.ie/course/view.php?id=1613 [Accessed on: Nov. 15, 2019].

[11] [Online].Available: http://onlinestatbook.com/2/describing_bivariate_data/pearson.html [Accessed on: Dec. 1, 2019].

[12] [Online]. Available: https://www.educba.com/r-vs-spss/ [Accessed on: Dec. 1, 2019].