ADVANCED DATA ANALYSIS

First Assignment

Professor: I. Ntzoufras

16/12/2023
MSc in Statistics - AUEB
Grammenos Konstantinos
AM : f3612302
Assignment 23
Title: Four-Mile Run Dataset

Contents

List of Figures

List of Tables

Abstract

This report presents an analysis of the FourMileRun dataset which contains data for 19 runs. The name of the runner is Kevin and his aim is to see improve his effectiveness of his training. The data were collected by the runner of the four-mile course using a Garmin Forerunner 610 GPS watch. In the data there are several and various metrics like the run duration, the pace, numerous heart rate statistics and many others. Through exploratory analysis, visual methods and hypothesis testing we examined Kevin's improvement over these 19 runs and we analyze and found associations between some key variables. Finally we construct several regression models to assess the progress of his performance and to see which variables play a significant role on his improvement. We hope that our analysis and our findings will offer insights and guidance to the individual Kevin and will suggest modifications in order to help him optimize his health benefits and improve his exercise program.

Introduction

The file we analyze includes measurements for 19 observations by a Global Positioning System (GPS) watch worn by a runner of a four-mile course. More specifically, the variables contained in the dataset are displayed in the table below.

Variable Number	Variable's Name	Type	Meaning	Value
1	Run	Integer	Run number	From 1 to 19
2	Time	Character	Time of 4-mile run	In minutes and Seconds
3	Pace	Character	Avg time to run 1 mile during a run	In minutes and Seconds
4	Calories Burned	Integer	# of calories burned during the run	From 302 to 446
5	Training Effect	Numeric	Training-induced development of fitness and performance	From 1 to 5
6	Max HR	Integer	Maximum HR during run (BPM)	From 143 to 173
7	Avg HR	Integer	Average heart rate during the 4-mile run (BPM)	From 103 to 153
8	Avg Speed	Numeric	Average speed during run (miles per hour)	From 6.7 to 7.7
9	Max Speed	Numeric	Max speed during run (miles per hour)	From 7.8 to 9.7
10	HR Rest	Integer	Heart rate immediately after run (beats per minute)	From 117 to 152
11	HR Rest1	Integer	Heart rate 1 minute after run (beats per minute)	From 77 to 112
12	HR Rest2	Integer	Heart rate 2 minutes after run	From 72 to 108
13	HR Change1	Integer	Difference in heart rate from start of rest until one-minute later	From 21 to 63
14	HR Change2	Integer	Difference in HR from start of rest until 2-min later	From 29 to 67

Table 1: Variables of our Dataset

In this assignment we examined the relations between the variables that play a role in the progress of Kevin. Our goal was to do a descriptive analysis for all the variables individually, but also for the pair wise associations that we have. In addition to the descriptive analysis, we will check through visual methods his progress and by analyzing the data that we have we will construct a linear model to examine his improvement through the association of the training effect and the other variables.

Descriptive analysis and exploratory data analysis

With the use of R Studio, we did some statistical analysis and plenty of graphs for our data. The data analysis will be done at significance level **a=5%.** First of all we import our data in R and we put them into a data frame that is called FourMileRun. When we imported the data we saw an extra column with column name x which was full with missing values (NaN) so we deleted this column and we kept the rest Data Frame. Furthermore, we constructed some new variables. The first is one named as Training.Effect.Cat which includes the categories of the Training Effect depending on its value. For example, Minor (1.0-1.9), Maintaining (2.0-2.9), Improving (3.0-3.9), Highly Improving (4.0-4.9), and Overreaching (5.0).Secondly, we split the time column into 2 new columns, Time.Minutes which are the minutes of each individual run and Time.Seconds which are the seconds of the run. Also, we made 2 more columns as regards the Time, one names as Time.In.Minutes that is the total minutes of each run and Time.In.Seconds which are the seconds of each run. We did the same procedure and we made the corresponding columns for the Pace variable. The name of the corresponding columns are Pace.Minutes, Pace.Seconds, Pace.In.Minutes and Pace.In.Seconds. These new variables are either numeric or integers and that is why we convert the initial Time and Pace variables which were of character class.

We began our analysis by constructing some graphs. Firstly we did a bar plot so see the percentage of each training effect category in our data. As we observe, above 80% of his runs belong to the Improving Category which means that most of the time he improves his performance.

<u>Time in Minutes</u>: Although, we observe some fluctuations we can see a trend of decreasing run time over the runs, indicating an improvement in Kevin's running efficiency and speed.

<u>Training Effect:</u> generally is above 3.0, which suggests that most runs have a positive impact on Kevin's fitness, as this score belongs to the Improving Category. There is a significant drop in the run 15 as we will see also from the boxplot (TE in run 15 is an outlier), but after this the training effect returns to high scores indicating that Kevin was able to come back from this bad run. Generally, we observe some fluctuation but also an increasing trend in its score something that would suggest that Kevin's running sessions are becoming more effective and he constantly improving his fitness.

<u>Calories Burned:</u> we observe some variability, likely influenced by the intensity of HR during each run. There is no clear up or down trend which is expected because the calories that someone is burning during exercise can rely on various factors. The maximum calories burned were 446 and the average 372. The pace of Kevin seems not to show an upward or downward trend but there are some peaks and troughs. There are runs in which the pace significantly increases (indicating a slower run) which followed by runs where the pace decreases sharply (faster runs).

<u>Max and Avg HR</u>: Both the maximum and average heart rates show the same pattern over the runs, with the average heart rate being around 136 bpm and the maximum heart rate around 159 bpm. From the graph we can see a stable cardiovascular response of Kevin's heart during his runs.

HR.Change1 and HR.Change2: As regards now the plot of HR.Change1 and HR.Change2 we can tell that larger values on the y-axis indicate a greater drop in heart rate after running, something which is generally a sign of improving cardiovascular fitness. In the early runs we see some variability but after the 15 run we observe an upward trend in both lines, HR.Change1 and HR.Change2 which may suggesting that Kevin's heart rate recovery is improving.

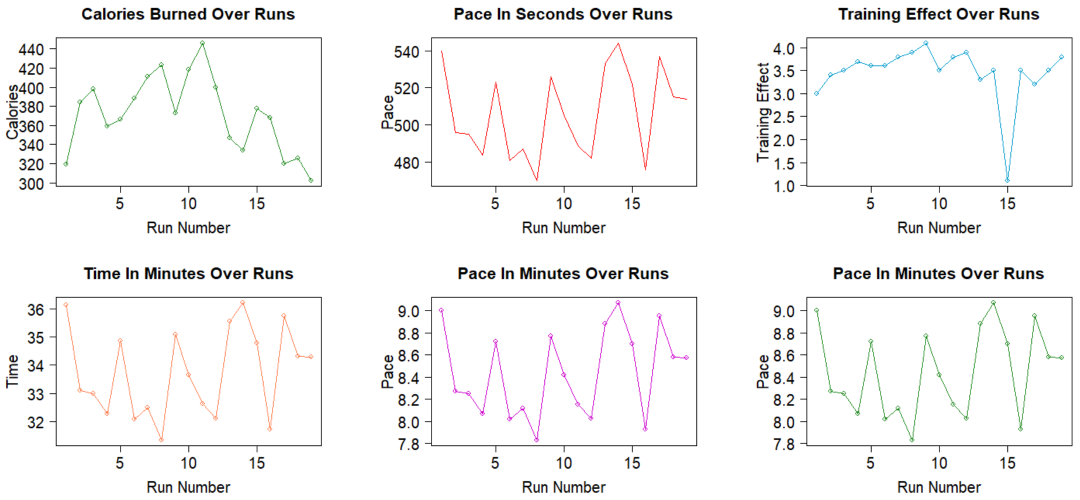


Figure 1: Plots of Numeric Variables

In the table below we can see more descriptive statistics about our variables. More specifically we present the mean, the standard deviation, the median, the trimmed mean, the mad, the min, the max, the range, the skewness, the kurtosis and the standard error for each of our quantitative variables.

- The average Time (in Minutes) to complete the 4 Mile Run is about 33.76 minutes with a standard deviation of about 1.59 minutes. These results show some variability in run times.

- The average pace is approximately 8.44 minutes per mile, with a standard deviation of 0.40 minutes. This shows a relatively consistent pace across runs, as the standard deviation is relatively small.

- The average Max HR is 159 beats per minute(bpm) while the average of the average HR is about 36 bpm. These values are very important in order to evaluate Kevin's cardiovascular fitness.

- Average speed is around 7.17 mph, with the maximum speed averaging at 8.72 mph.

- HR.Rest, HR.Rest1 and HR.Rest2 are cardiovascular metrics which show us how quickly Kevin's HR returns to a lower value, which is a sign of a good cardiovascular system. HR immediately after run averaged at 137 bpm and falls to 93 and 90 bpm 1 and 2 minutes after.

- HR Change1 and HR Change2 show the change in heart rate one and two minutes after the run, indicating the recovery speed HR. Average decrease in heart rate after one and two minutes post-run are 44 and 47 bpm, respectively.

	Mean	Sd	Median	Mad	Min	Max	Range
Calories Burned	371.58	39.74	373	40.03	302	446	144
Max HR	159	7.74	159	4.45	143	173	30
Avg HR	135.68	10.71	136	7.41	103	153	50
Avg Speed	7.17	0.32	7.10	0.44	6.70	7.70	1
Max Speed	8.72	0.59	8.80	0.74	7.80	9.70	1.90
HR Rest	137	10.52	140	7.41	117	152	35
HR Rest1	93.16	9.49	90	10.38	77	112	35
HR Rest 2	90.05	8.93	93	5.93	72	108	36
HR Change 1	43.84	10.20	46	7.41	21	63	42
HR Change 2	46.95	9.19	49	5.93	29	67	38
Time In Minutes	33.76	1.59	33.65	2.03	31.32	36.23	4.91
Pace In Minutes	8.44	0.40	8.42	0.52	7.83	9.07	1.24

Table 2: Descriptive Statistics of our variables

Furthermore, we did some QQ plots for our variables to see if they follow the normal distribution .Then, we use Shapiro-Wilks test in order to check for normality. Using Shapiro test and a significance level of α=5% we found that the p value of the test for the variables: Calories.Burned, Max.HR, Avg.Speed, Max.Speed, HR.Rest, HR.Rest1, HR.Rest2, HR.Change1, HR.Change2, Pace.In.Minutes and for Time.In.Minutes is above 0.05 so we do not reject Ho so we can assume normality for all these variables. We cannot assume normality for Avg.Hr and of course for Training.Effect.

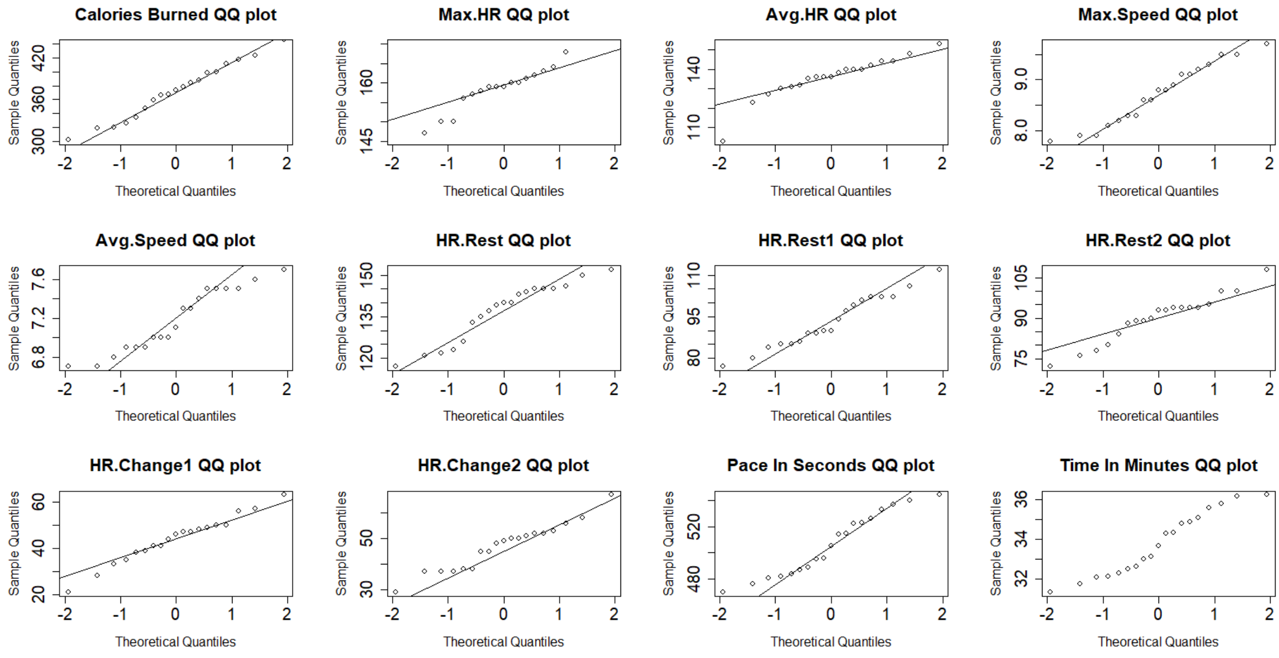


Figure 2:Normal QQ Plots for the numeric variables

We also made a test for outliers and we found that for the variable Training Effect we found only one outlier, for Max.HR 4 outliers, for variable Avg.HR 1, for HR.Rest2 2 outliers, for HR.Change1 1. We can observe them clearly in the boxplots below.

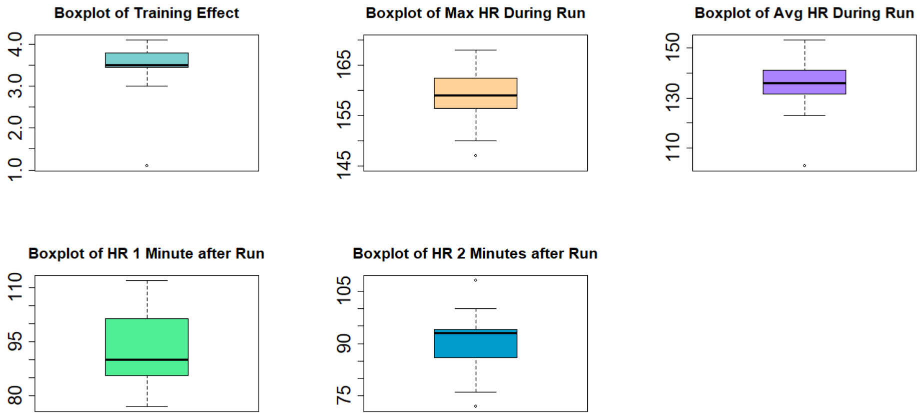


Figure 3: Boxplots of some important variables

We see an outlier in the Training Effect with a value of 1.1 (observation 15) that belongs to the Minor Training Effect category. We can assume that maybe Kevin was a bit tired this day or also the conditions on this specific day were not on his favour. For example, there could be some adverse weather conditions, like extreme heat or cold and may these very low or high temperatures made the run much more difficult to him. We observe that for the same observation (15), there are outliers in Avg.HR (103), in Max.HR (143) and in HR.Change1 (21), which might prove that our assumption about the adverse weather conditions is true. As we all know temperature plays a significant role and affects heart rate during any exercise. For example, in observation 15 Max.HR is 143 which is like 16 beats down from the mean (159) and the same is true for Avg.HR where it is 103 and the mean is 135 so it is 32 beats less per minute. Finally, although the low training effect value for Run 15 may seem to be an outlier, we can say that it is to be expected given bad weather conditions.

For the Max.HR as we said the one outlier is the one of the run 15 but there are 3 more. The second one is in the follow run of the run 15, that is the 16 run which indicates maybe some fatigue in those 2 runs. There are also 2 high values of Max HR which are outliers, the one is at the run 13 (172) and the other on the run 19 (the last run). For these 2 runs Kevin may pushed himself harder and maybe the low Max HR of the fellow runs (15,16) is a result of the high intensity 13 run. As regards the last run Kevin may wanted to give his best and try his limits and that’s why we saw this peak in his maximum HR.

As regards now the boxplot of HR 1 Minute after Run we observe a box ranges from almost 85 to 100 bpm indicating consistent recovery rates across the runs The median is about 90 bpm.

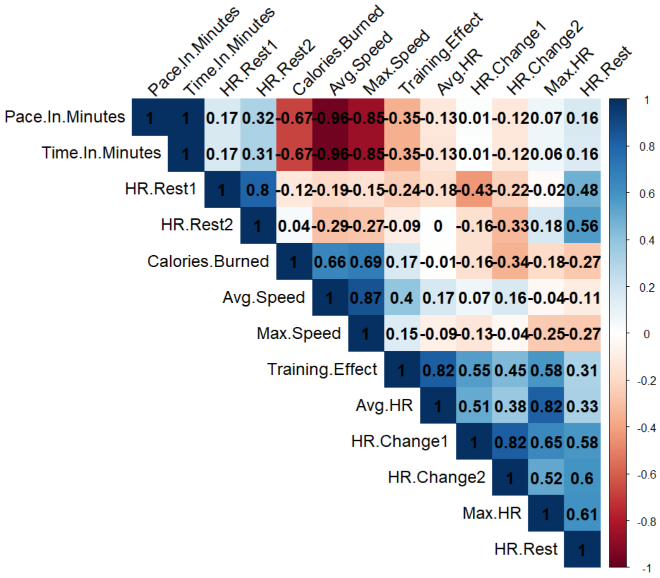
Last but not least, the Heart Rate 2 minutes after run shows very small variability as the 50% of the values range from somewhere above 85 to something lower that 95 bpm. We observe 2 outliers, one is the run 13 with HR after 2 minutes to be 108 bpm but as we have said before this was a high intensity run with an extremely high value of Max HR, so this outlier makes sense in a way. The 2nd outlier is that of the run 19 with value 76 where we also have a high intensity run with hug Max HR value. In this run Kevin’s heart rate dropped exceptionally quickly. The fact that both runs (13 and 19) were high intense runs but in the one the heart rate did not drop so much but in the other dropped significantly show us that Kevin’s heart has been adjusted to exercise and that Kevin throughout these 19 runs has improve a lot his cardiovascular fitness. The drop of the HR rate in the 19th run show us a good cardiovascular recovery that was achieved through Kevin’s training.

Because of the few observations that we have (19) we decided not to do any histograms and density plots.

Finally, we did split the column of Time.In.Minutes in 2 , the first and the second half in order to compare those 2 halves and check if there is statistically significant difference between them. After checking the normality assumption for each part we tested it using a paired t-test and we found that we do not reject the null hypothesis that there is no difference in mean running time between the two halves(p value=0.387). We found that the mean difference is -0.7133 which means that runs of the second half are faster by 0.7133 minutes but this is not statistically significant so we cannot really extract any important meaning from this analysis.

Pairwise comparisons

Since we cannot draw a conclusion only by investigating and analyzing each individual variable separately, we continued our analysis by examining the pair association between our variables. Firstly and most significantly, we created a Pearson correlation matrix, in order to investigate the relationship between our most important continuous variables. Each cell in the figure 4 shows the correlation between two variables, and the value inside the cell represents the correlation coefficient. We know that value 1 indicating a perfect positive correlation, the value -1 indicating a perfect negative correlation, and value 0 indicates no correlation.



	Pace.In.Minutes	Time.In.Minutes	HR.Rest1	HR.Rest2	Calories.Burned	Avg.Speed	Max.Speed	Training.Effect	Avg.HR	HR.Change1	HR.Change2	Max.HR	HR.Rest
Pace.In.Minutes	1	1	0.17	0.32	-0.67	-0.96	-0.85	-0.35	-0.13	0.01	-0.12	0.07	0.16
Time.In.Minutes		1	0.17	0.31	-0.67	-0.96	-0.85	-0.35	-0.13	0.01	-0.12	0.06	0.16
HR.Rest1			1	0.8	-0.12	-0.19	-0.15	-0.24	-0.18	-0.43	-0.22	-0.02	0.48
HR.Rest2				1	0.04	-0.29	-0.27	-0.09	0	-0.16	-0.33	0.18	0.56
Calories.Burned					1	0.66	0.69	0.17	-0.01	-0.16	-0.34	-0.18	-0.27
Avg.Speed						1	0.87	0.4	0.17	0.07	0.16	-0.04	-0.11
Max.Speed							1	0.15	-0.09	-0.13	-0.04	-0.25	-0.27
Training.Effect								1	0.82	0.55	0.45	0.58	0.31
Avg.HR									1	0.51	0.38	0.82	0.33
HR.Change1										1	0.82	0.65	0.58
HR.Change2											1	0.52	0.6
Max.HR												1	0.61
HR.Rest													1

Figure 4: Correlation Matrix

In the whole sample, we see that there is strong negative correlation between the Calories Burned during the run , the pace and of course the time in minutes. Faster runs lead to burning more calories.Also, as we would expected, there is very strong negative correlation between pace in minutes and time in minutes and of course there is very strong correlation between max and avg speed with pace and time. As Kevin's average speed increases, the time it takes to complete his runs decreases. In addition, we see strong positive correlation between the calories burned and avg and max speed, but generally there's no clear linear trend with heart rate or pace, indicating that calories burned are influenced by a combination of factors. Furthermore, we observe positive strong correlation between max.HR and HR.Change1 and also between HR.Change1 with the Training Effect. The training effect generally seems to have some association with heart rate metrics, and more particularly with the maximum and the average heart rate. This suggests that runs with higher intensity might contribute more to training effectiveness.The correlation between HR.Change2 and Max.HR, Training Effect and Avg.HR seems to be more moderate that this of the HR.Change1 and the same variables. In addition, we see very strong positive correlation between HR.Rest1 and HR.Rest2, moderate positive correlation between HR.Rest and HR.Rest1 and strong positive correlation between HR.Change1 and HR.Rest and the same is true for HR.Change2. This is absolutely logical and expected as HR.Change1 is HR.Rest- HR.Rest1 and HR.Change2 is HR.Rest- HR.Rest2. Strong positive correlation there is also between Max.HR and HR.Rest.Finally, we see that there is very strong positive correlation between the Max.HR and the avg.HR and also between Training Effect and avg.HR. The rest of the variables are weak or moderate correlated (weak linear dependence).

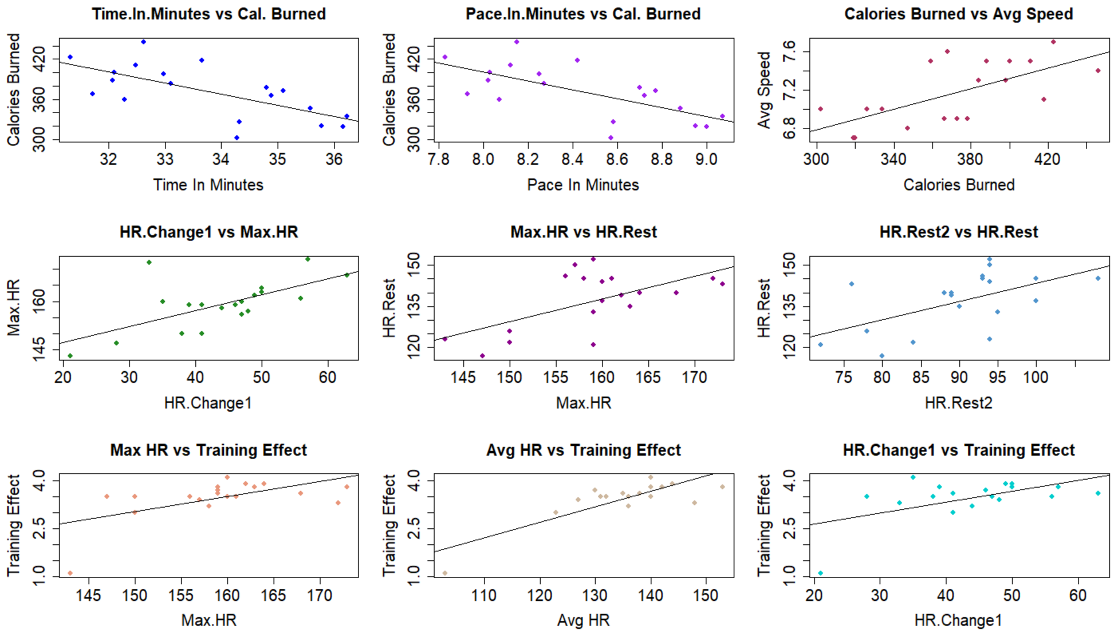


Figure 5: Scatter Plots of correlated variables

Finally,we conducted some analysis to examine the association between the categorical variable Training.Effect.Cat and several quantitative variables. However, a limitation of our analysis is the uneven distribution of observations across the different categories of Training.Effect.Cat. Out of the 19 runs, we observed only one run each for the 'Minor' and 'Highly Improving' categories, while the majority (17 runs) fell under the 'Improving' category.

Due to the limited number of observations in the 'Minor' and 'Highly Improving' categories, our ability to draw statistically meaningful conclusions for these specific categories is restricted. With only one observation in the 'Minor' and 'Highly Improving' categories, the variability within those categories cannot be estimated and computing reliable measures such as mean and standard deviation becomes impractical. Thus, we do not have information about the potential range of distribution of values for those categories.

Despite this limitation, we proceeded to create boxplots of some key quantitative variables, such as Average Heart Rate, Max Heart Rate, HR after the run, HR.Change1, and HR.Change2, by each category. The boxplots provide a visual representation of the distribution of these variables in each category, offering insights into the central tendency and variability of the data mostly within the Improving category.

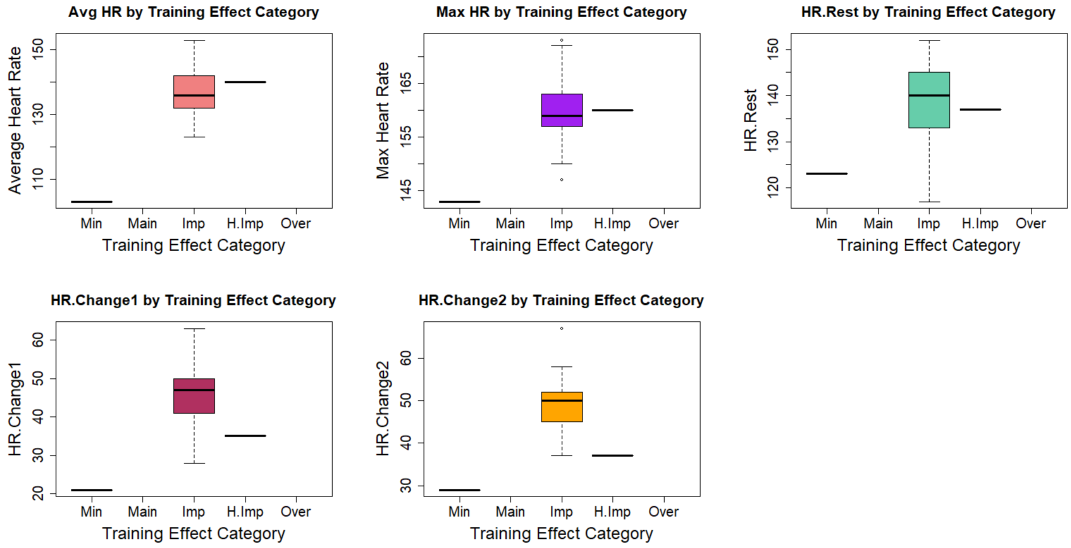


Figure 6: Boxplots by Training Category

Our interpretation of the boxplots will be focused on the 'Improving' category.

- In this category the boxplot of Avg.HR display a median Avg.HR close to 130 beats per minute. The box's range which represents the middle 50% of the data, is relatively narrow, suggesting that the Avg.HR values in this category are quite consistent. The whiskers extend from approximately 120 to something above150 beats per minute, and that is the range of Avg.HR. We do not have any outliers, which implies that all Avg.HR values fall within a typical range without extreme values. As we observe the boxplot of Avg.HR is similar to the ones we would do if we have the whole dataset, which is expected because 17 out of 19 observations lies in Improving Category. The difference is that now we observe the outlier of the run 15 as a straight line (because it is only 1 value) in the Minor Category.

- For the Max HR boxplot in the Improving Category we see an extremely narrow boxplot with the median around 160 bpm and the middle 50% of values to be somewhere above 155 and slightly lower that 165 bpm. Now we can see 2 outliers in the Improving Category while in the boxplot of Max HR of the whole dataset we observed 4. The one outlier is lower than 145 bpm and it is the run 16 which make senses to have low Max HR in this run because it followed by a bad run too (the run 15). We suppose that Kevin was too tired in these 2 runs or may the weather conditions were extreme and did not help him at all. The other outlier is a value of something above 170 bpm which is the last run with Max HR of 173 bpm. Kevin may wanted to increase the intensity of his workout significantly in his last run, pushing his heart rate higher than in previous runs.

- For the boxplot of HR.Rest boxplot we see that the median for the Improving category is something above 140 bpm, which is a moderate post-exercise heart rate and indicates that Kevin's runs are effectively improve his cardiovascular fitness. The interquartile range ranges from 130 to 145 bpm, indicating that there is some variability in his immediate post-run heart rate, but generally, it stays within a 15 bpm range, which we would say that is a relatively small range, suggests a level of consistency in his immediate HR after runs. We do not observe any outliers in this boxplot.

- Finally for the boxplot of the change in HR from 1 minute after the run and the HR immediate after run we observe an extremely narrow box suggesting a consistent

HR recovery . Also we do not observe any outliers.

- For the last boxplot of the change between HR from 2 minutes after the run and the HR immediate after run, we also see a smaller box ranges from 45 bpm to something above 15 bpm and we also see an outlier indicating that Kevin's heart rate dropped significantly more than the drop in other runs. This happened in the last run, the 19 run with HR.Change2 to be 67 bpm. For the last run we observed also a very high value of Max HR, so this fact combine with this high value of HR.Change2 proves us the high intensity that the last run may had.

Overall given the limitations of our dataset, it might be more appropriate for us to focus on descriptive statistics.

Predictive or Descriptive models

Since we studied and analyzed our variables both individually and pair wised and found which associations are significant, we constructed some regression models to assess the progress of Kevin. Our aim is to create linear models in order to examine the association between the training effect and the rest of variables. As you may understand, our response variable here is the training effect.

In order to build the appropriate model, we create a subset of the original data set, basically removing some variables that there was no meaning to be in the construction of the model (see appendix 3).

First of all, we fitted the full model that means the model with every variable of this subset data set and then we implemented the stepwise procedure, starting from this full model . After finding the best model from this technique we made some transformations to finally find that the best possible result is the one below.

We have the Training Effect as our response variable and Avg.HR, I(Avg.HR^2), the squared term of Average heart rate and the Time.In.Minutes are the explanatory ones. We thought about adding the quadratic effect when we saw that the relationship between Avg HR and Training Effect appears to be nonlinear. You can see that in figure 12.

Coefficients	Estimate	Std.Error	t value
Intercept	-23.87	4.38	-5.449
Avg.HR	0.4124	0.06346	6.498
I(Avg.HR^2)	-0.001421	0.0002458	-5.781
Time.In.Minutes	-0.06871	0.02981	-2.305
Residual Std. Error	0.19171 on 15 degrees of freedom		
Multiple R-squared	0.9182		
Adjusted R-squared	0.9019		

Table 3:Coefficients of the model

After finding that this is the best model by comparing the AIC of the best models that we had (see apendix), we checked and verified our model assumptions. First, we checked the assumption of the normality of the residuals using Shapiro Test (p-value = 0.3185) and we found that we do not reject the Null Hypothesis so the residuals may follow a normal distribution.We also did QQ Plots and Histograms for the residuals (see Figure 7). Then, we checked for Homoscedasticity of the residuals using the Levene test and we found that we do not reject Ho that the variances of the residuals are equal across quartiles of the fitted values(pvalue= 0.4155), so Homoscedasticity may holds (Figure 8). As we see in the figure 9 the residuals are scattered around the horizontal line at zero, which indicates that there's no systematic pattern to the residuals. When there is no pattern is a good sign and it indicates that the residuals are randomly distributed.The spread of the residuals appears to be fairly constant as we move from left to right along the fitted values. This visual inspection is consistent with the Levene test result.Furthemore, we check for the Independence of the residuals using the runs test and we discover that we do not reject Ho that the sequence was produced in a random manner (p-value = 0.627) so we have signs of independence of the residuals. Also, we did some plots to show the independence (see figure 8).

Also we saw from the Anova table below that all the covariates of the model are statistically significant.

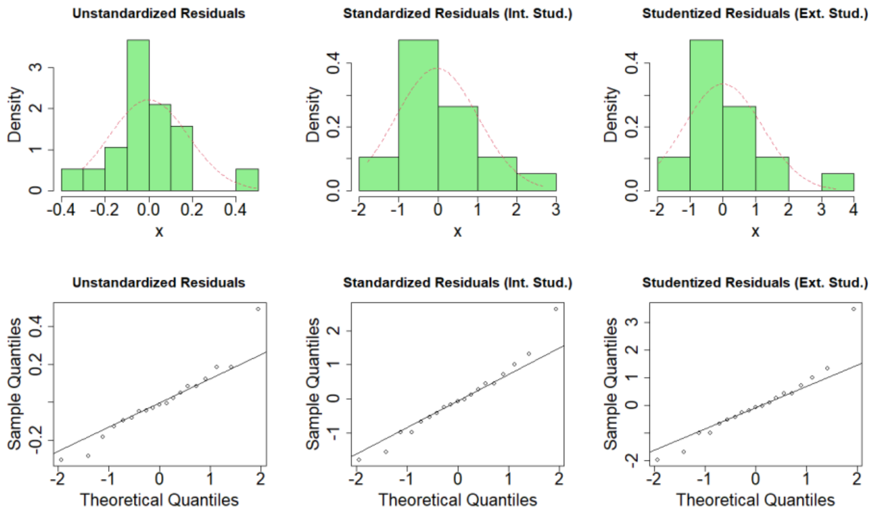


Figure 7: Histograms and QQ Plots of the Residuals

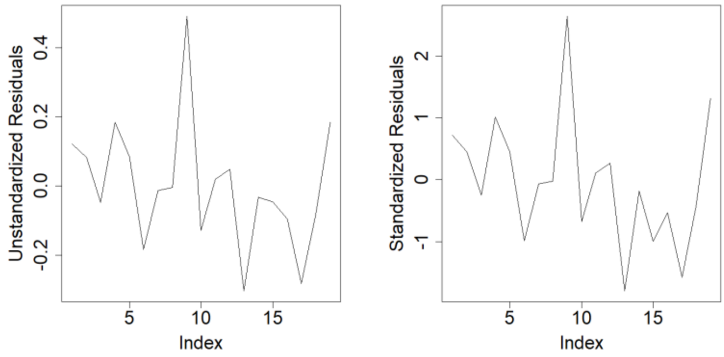


Figure 8: Independence Plots

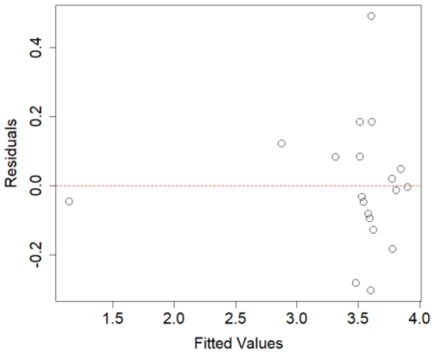


Figure 9: Residuals Vs Fitted Values

As we see in the figure 9, almost 95% of the residuals are in [-2,2] interval. Only 3 observations are outside of this interval. Also we do not see any structures in the plot which is a good sign. Furthermore we did an Anova test and we realized that every explanatory variable is statistically significant for the model.

So now, we proceed to further interpretation.

Our constant is β0 (intercept) represents the value of Training effect when the rest of the variables are zero. Yet, in our example there is no meaning for us to say that there is something like zero for our predictors. That's why we thought about centering the predictors of the model around their means. This process will not change the shape of the model or its fit but it will simply shifts the baseline to a more meaningful point of reference. We can see the coefficients of the new model in the table 5.

Coefficients	Estimate	Std.Error	t value
Intercept	3.6122	0.0525	68.783
Centered.Avg.HR	0.0026	0.0055	4.851
I(Centered.Avg.HR^2)	-0.0014	0.00024	-5.781
Centered.Time.In.Minutes	-0.06871	0.02981	-2.305
Residual Std. Error	0.19171 on 15 degrees of freedom		
Multiple R-squared	0.9182		
Adjusted R-squared	0.9019		

Table 4: Coefficients of the Centered Model

Now that we have the new model, the intepretations may have more meaning.

Intercept: Our constant is β0 (intercept) =3.61 which now represents the value of Training effect when the rest of the variables are at their reference level, that is when they are equal to their means. So when Avg.HR and Time.In.Minutes are at their average values the score of Training Effect will be 3.61 and this belongs to the Improving category.

Centered_Avg_HR: For each unit increase in the Avg.HR from its mean value that is for 1 beat per minute increase on the mean of Avg.HR, our response variable, the Training Effect will be increased by 0.0268 units, when of course all the other variables are constant. This increase you may think that is small but it is statistically significant.

I(Centered_Avg_HR^2): this represents the change in the effect of Avg.HR on Training.Effect for each unit increase in the squared term of Avg.HR from its mean value. As Avg.HR increases, the rate of increase in Training.Effect decreases by -0.0014208.

Centered_Time_In_Minutes: For each minute that Time.In.Minutes increases the there is a decrease in Training Effect and this is about 0.0687 units. This decrease is statistically significant. It is important to say that we again checked the assumption of the new model and we saw that the centerization of the predictors did not have any impact on them.

Finally, we tried to make some predictions using the best model which we describedabove. In the table 6 we can see the predictions of the value of Training.

Avg.HR	Time.In.Minutes	Predicted_Training_Effect
132	34.3	3.4569
140	32.5	3.7881
145	33.0	3.7911
150	33.5	3.7230
154	33.23	3.6635
151	33.45	3.7112
137	34.05	3.6250
128	34.18	3.2931
135	32.80	3.6591
142	33.30	3.7566
152	33.60	3.6828

Table 5: Prediction of Training Effect

Conclusions and Discussion

This comprehensive analysis of the Kevin's 19 run sessions has provided us various and numerous insights into his training progress and cardiovascular fitness. From the exploratory analysis in the first part we revealed man useful descriptive statistics and we showed the distribution of many of our variables during these 19 runs. After this, pairwise associations highlighted significant relationships, particularly between heart rate metrics and training effects but also between speed metrics and time metrics. Noticeably, the variable that plays the most significant role in the training effect is the Average Heart Rate during each run. Also the inclusion of the quadratic term of Avg HR showed us a non-linear relationship between the Training Effect and the Avg HR, indicating that the impact of heart rate on Training Effect may follows a more complex pattern. The final model also includes the Time.In.Minutes which implies that the duration of the run influence the training outcome. Max HR and other HR predictors were not significant predictors of Training Effect at least in this approach, which may show us that peak and post exercise HR may not give us much information about the Training Effect despite the high correlation that they have with our response variable (see figure 4). This model was selected because of its high Multiple R2 value and also because its small AIC (-2.28).
The findings from our analysis align with current exercise physiology research, which emphasizes the role of intensity in cardiovascular improvement. The positive relationship between average heart rate and training effect proved us one more time that moderate to high intensity training it is extremely significant in order to have a good training outcome. Though we believe that given more data and observations we may have found many more crucial insights about his training and we could help him more on his improvement.

References

[1] Ntzoufras I., Advanced data analysis with R, educational notes for MSc program Statistics AUEB

[2] Paul J. Laumakis, Rowan University, Analyzing Exercise Training Effect and Its Impact on Cardiorespiratory and Cardiovascular Fitness, Journal of Statistics Education Volume 22, Number 2 (2014)

[3] Benson, R. and Connelly, D. (2011), Heart Rate Training, Human Kinetics.

Appendix 1

Referring to descriptive analysis and exploratory data analysis

We made some Boxplots for the rest of our quantitative variables in which we do not observed any outliers as we see in the graphs.

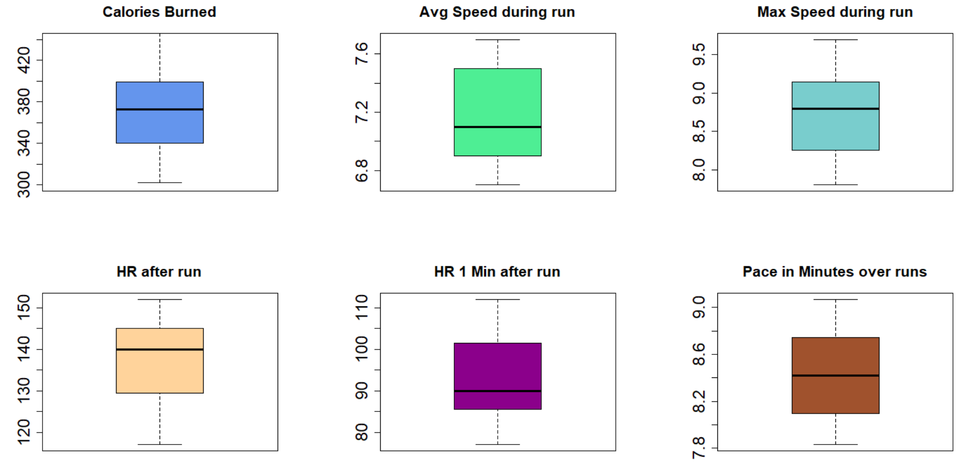


Figure 10: Boxplots of our quant variables

Boxplot of Calories Burned: We see that the 50% of the data are quite gathered and there is not so much variability, indicating that the calories burned in each run are fairly consistent. The median and the mean are almost the same as we can see also in the Table 2, something which may tell us that we have a symmetrical distribution. More specifically the mean is approximately 372 and the median 373.

Boxplot of Average Speed During Run: We observe that the average speed has a wider IQR which means more variability in Kevin's average speed from run to run. The median here is about 7.10 while the mean is 7.17. No outliers in the distribution of Avg Speed.

Boxplot of Max Speed During Run: We see that the max speed shows much less variability than the average speed. The median is centrally located (8.80) pretty close to the mean (8.72), which may show us a symmetric distribution. Also here we have no outliers, indicating consistent performance in reaching top speeds.

Boxplot of HR After Run: This is the boxplot of the heart rate immediately after Kevin's run. We observe a very wide IQR, which show us some fluctuation in Kevin's immediate post-run heart rate. These fluctuations are expected due to the differing intensities of each run. The median here is larger that the mean. Again we have no outliers.

Boxplot of HR 1 Minute After Run: This boxplot shows us the heart rate one minute after the run. The box of this boxplot is smaller (less variability) than the one from the HR immediate after the run which is logical and suggests that Kevin's heart rate tends to decreases and stabilize after one minute of recovery. The median here is a bit smaller than the mean and also we have no outliers.

Boxplot of Time in Minutes over Runs: This the boxplot of the Time (in minutes) that Kevin needed in order to complete each run. We do not see so much variability which is a good sign of Kevin's performance and may suggests that Kevin's run completion times are quite consistent. There are no outliers, and the median is slightly towards the lower end of the box, which could show us a general trend towards faster run times.

Appendix 2

Pairwise Associations

After this, we made some hypothesis testing between HR.Rest1 and HR.Rest2 to see the difference between the mean of the Heart rate one-minute after run and the mean of the Heart rate 2-minutes after run.We made a paired t test as we know that we have normality for both variables, and after checking the p-value(0.03243) we reject the Null Hypothesis that their true mean difference is equal to 0. That’s why we constructed an Error Bar.

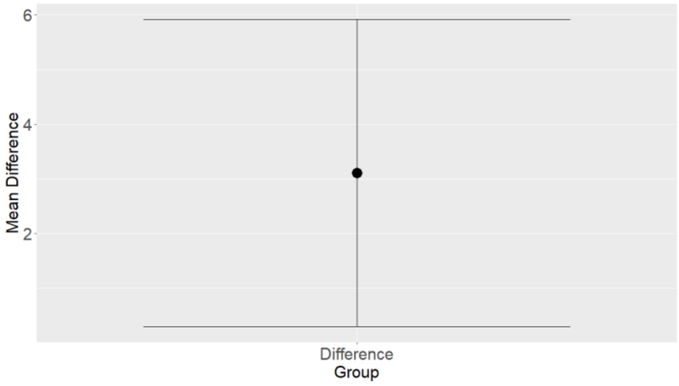


Figure 11: Error Bar Plot of HR.Rest1 and HR.Rest2

Actually their mean difference is statistically significant and it is about 3.10 bpm and we found that 95% of the true difference in means between "HR.Rest1" and "HR.Rest2" lies between 0.2907285 and 5.919797 beats per minute.

A significant drop in heart rate within the first two minutes post-exercise is a sign of good cardiovascular recovery. If Kevin's HR.Rest1 is significantly higher than HR.Rest2, it suggests that his HR is quickly returning to a resting state.Monitoring these changes over time can help Kevin improve his cardiovascular fitness.

We also did a Hypothesis testing between Avg.HR and HR.Rest to see the diffrence between the median of the Average Heart rate during run and the Hear Rate immediately after run. We do not have normality for Avg.HR and also we have a small dataset<50, so that’s why we did a Wilcoxon test. As checking the p value (0.7625) we found that we do not reject the Null hypothesis that there is no difference in the medians between the two paired groups.

Appendix 3

Predictive or Descriptive models

The subset of our initial data that we created, contained the variables Time.In.Minutes, Calories Burned, Training Effect, Max.HR, Avg.HR, Avg Speed, Max Speed, HR Rest., HR Rest1, HR Rest2, HR Change1, HR Change2.
So the full model contained all these variables with the response to be the Training Effect. Firstly, using the stepwise procedure with direction=both, we found the best model derived from this algorithm , which was this one lm(formula = Training.Effect ~ Max.HR + Avg.HR + HR.Rest + HR.Rest1 + Time.In.Minutes, data = subset_data). This model was a quite good one with Multiple R2=0.8194 which means that it could explain about 82% of the variability of the data. The AIC of this model was 16.7658401 with 7 df.

Coefficients	Estimate	Std.Error	t value
Intercept	2.7224	2.2741	1.197
Max HR	-0.0434	0.0224	-1.938
Avg HR	0.0609	0.0135	4.510
HR Rest	0.0270	0.0117	2.298
HR Rest1	-0.0167	0.0100	-1.664
Time In Minutes	-0.0819	0.0494	-1.665
Residual Standard Error	0.3146 on 13 df		
Multiple R-squared	0.8194		
Adjusted R-squared	0.75		

Table 6: Summary of the best model from stepwise procedure

From the summary of this model we see that the coefficient of Avg.HR is highly statistically significant and HR.Rest coefficient is also statistically significant but no so much as Avg HR. Coefficients of Max HR, Time.In.Minutes and HR.Rest1 are not statistically significant(p value > 0.05). From the Anova we saw that only the Max.HR and the Avg.HR are statistically significant (p value < 0.05) while the other variables not (p value > 0.05). After evaluating the performance of the model and the above results the first thing that we thought was to add the quadratic term of Avg HR. We thought about that after observing from the Figure 12 of the report, from the scatter plot of Avg HR with Training Effect that the association between those 2 does not seem to be linear. In the figure 12 the left plot is the simple scatter plot between Training Effect and Avg HR with a linear fit line. From this we can see a positive relationship between those. The right plot is a quadratic regression plot, which includes a curved fit line, indicating that the relationship between Avg HR and Training Effect as we see is better explained by a quadratic model. The very first increase in Avg HR leads to a rapid and big increase in Training Effect but as avg HR continue to increase the rate of increase slows down as we can see.

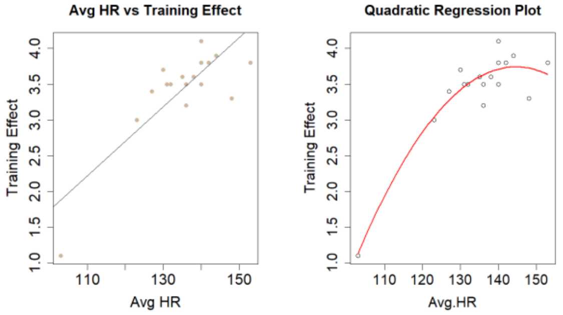


Figure 12: Avg HR plots with Training Effect (Simple and Quadratic)

Coefficients	Estimate	Std.Error	t value
Intercept	-22.45	6.449	-3.481
Max HR	-0.00131	0.01849	-0.071
Avg HR	0.3916	0.08274	4.732
I(Avg HR^2)	-0.0013	0.00033	-4.021
HR Rest	0.00609	0.01193	0.639
HR Rest1	-0.00441	0.00751	-0.587
Time In Minutes	-0.07346	0.03369	-2.181
Residual Standard Error	0.2137 on 12 df		
Multiple R-squared	0.9231		
Adjusted R-squared	0.8846		

Table 7: Coefficients of the stepwise model with quadratic term

So, adding the quadratic effect in the previous model we take the results in the table above. Now we see that the coefficients of Avg HR, the quadratic effect, the intercept and also the coefficient of Time.In.Minutes are statistically significant. Also we see an increase in Multiple R-squared which now has a value of 0.9231. The AIC of this model is 2.55.
From the Anova Analysis, we see that the Avg.HR has the highest F-value of 56.8837, indicating a very strong relationship with the training effect, and is highly significant. I(Avg.HR^2) is also highly statistically significant suggesting that the relationship between average heart rate and training effect is not strictly linear but includes a quadratic component. Time.In.Minutes shows a significant contribution (p = 0.0498374) with an F-value of 4.7552, indicating that the time to complete the runs has a quite significant impact on the training effect. All the others are statistically insignificant.
Finally, we tried to fit one more model which as it proved it was the best one. That was this one: lm(formula = Training.Effect ~ Avg.HR + I(Avg.HR^2) +Time.In.Minutes , data = subset_data) and we can see the summary of the coefficients of this model in the report in table 3. We compared the AIC of these 3 models and we end up choosing the last one model which had an AIC with value -2.28.