Exploratory Data Analysis Report

Athletes will dedicate several years of training into their performance at the Olympics. The goal is to push the limits of the human body in athletic excellence. We will be focusing on track events that involve strictly running, excluding the marathon. Some of these races are finished in under 11 seconds and others last over half an hour. How can we best predict an Olympic track athlete’s finishing time in these races? What is the relationship between an athlete’s finishing time, adjusted for distance, at the Olympics and their sex, age, weight, height, year of competition, and the athlete’s nationality, their country’s total GDP, and population? Our goal is to model an athlete’s race time with athlete-level characteristics and their team’s country-level characteristics.

While investigating relationships between these variables, we kept in mind that we were headed toward a hierarchical model and so, in an organizational strategy, we analyzed the variables by their levels. First, we will explore variables at level 1, which is an athlete’s race, then level 2, the country the athlete is from, with cross-level relationships throughout.

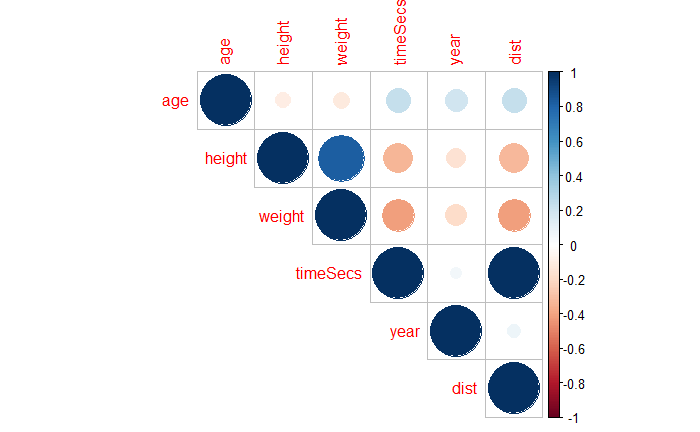
A level 1 unit is athlete’s race that is defined by the athlete, with characteristics such as age, height, weight, and sex, and defined by the race, with characteristics such a distance and year. Looking at the correlation plot in *Figure 1* and corresponding table of values of level one variables and the response in *Table 1*, there are some obvious pairs of variables with strong relationships. Height and weight are highly positively correlated (0.824). Distance of the race and the finishing time are nearly perfectly correlated (0.998). Then, there are some more interesting, albeit weaker, relationships. Both weight (-0.41) and height (-0.34) are negatively correlated with finishing time. Athletes that are taller or weigh more, presumably weight due to muscle, tend to have faster times. However, when we look at graphs of time versus height and time versus weight grouped by distance, we can see there are different slopes depending on the distance of the race seen in *Figure 5* and *Figure 6*. The effect of weight seems to matter less as the distance increases, while the effect of height peaks at the middle tier distance and matters less on especially short races or especially long races. Now, age is positively correlated with finishing, meaning the older an athlete the slower the time tends to result. Yet, when this is broken down by distance in *Figure 7*, there are positive slope for 800-meter races or longer and negative slopes for 400-meter races and shorter. Seemingly, older athletes perform better in longer distance races. Another nuance to age is that there is a positive correlation between year and age (0.1943), suggesting that athletes are competing longer into their careers. Remarkably in *Figure 12* year and finishing time has a very low (positive) correlation, but the plots of time versus year grouped by distance does show a downward trend in every distance and *Table 4* shows the coefficient for year in the regression line predicting time (seconds) from year and adjusting for distance is very statistically significant (t-value = -5.5).

On average, males tend to have lower finishing times than females. Maybe a lesser known difference between males and female athletes, across Olympic history there have been almost triple the amount of male medaling athletes compared to female (413 males; 172 females) seen in *Figure 4*. Also as *Figure 3* displays, there are less competitors overall as the distance increases and males’ finishing times appear to vary more as distance increases while females’ finishing times variation seems to stay about the same. In short, at level 1 there seems to be graphical and numerical indicators to include interaction with age and distance, weight and distance, height and distance, age and year, and sex and distance.

In our coming model we will nest athletes within country. From *Table* 2, there are 45 countries with athletes that medaled in Olympic track history with the events we have included. The USA has 3.3 times more medals than the next country, Great Britain, with 177 compared to 54. In fact, the USA makes up 30.26% of observations in this data set. Over half (22/45) of countries included have less than 5 medals. There are very large differences in country variation in finishing times seen in *Figure 10*. Overall, the variances of finishing times grouped by distance are approximately equal using *Figure 8.*  Country level characteristics include GDP total and population. The correlation plot and correlation coefficients are in *Figure 9* and *Table 3.* The obvious relationship is that GDP and population are positively correlated (0.654). We also thought population and finishing time would be negatively correlated because then there are larger pools of athletes to choose from. However, population and finishing time are only loosely negatively correlated (-0.069). When looking at log(population) versus finishing time this number moves to -0.137. When the 4 Chinese athletes are removed, this correlation becomes even stronger (-0.240), but, even though China has a population over 1 billion people, there is no sound reason to remove these athletes. As suspected, GDP and finishing time are negatively correlated (-0.205). We originally thought wealthier countries might have athletes with faster times because those countries can send more athletes to the Olympics and give them better training.

Looking ahead, we plan to fit a hierarchical linear model that considers country as a random effect and distance as a fixed effect as our null model. From there, we will continue to add fixed effects, the interactions mentioned above, and then fit models with random slopes. Our goal remains the same: to best model the finish times of Olympic track athletes with athlete race level information and country level information.

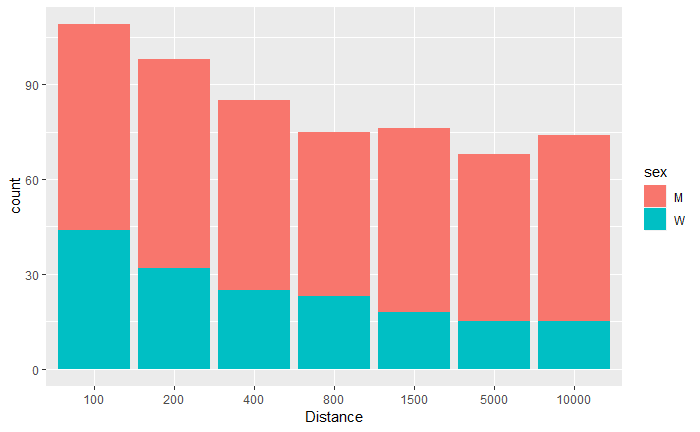
Appendix



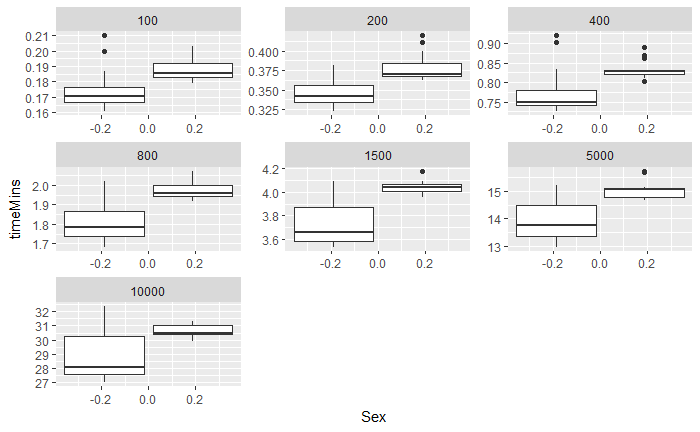
*Figure 1: Correlation plot of level 1 quantitative variables and response*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | age | height | weight | timeSecs | year | dist |
| age | 1 | -0.090 | -0.114 | 0.238 | 0.194 | 0.238 |
| height |  | 1 | 0.824 | -0.337 | -0.153 | -0.327 |
| weight |  |  | 1 | -0.414 | -0.187 | -0.411 |
| timeSecs |  |  |  | 1 | 0.055 | 0.998 |
| year |  |  |  |  | 1 | 0.070 |
| dist |  |  |  |  |  | 1 |

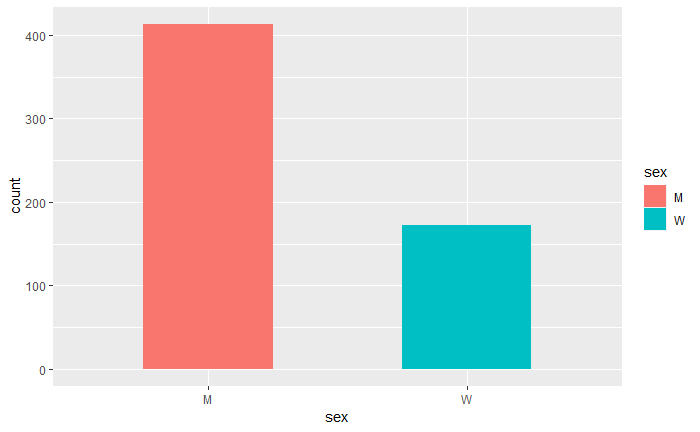
*Table 1: Correlation Coefficients of level 1 quantitative variables and response*



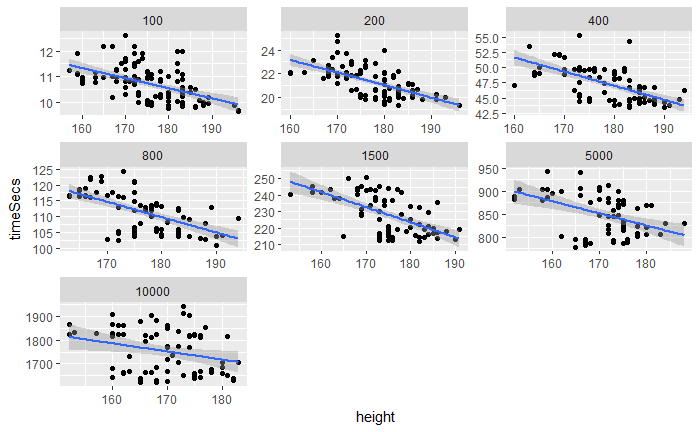
*Figure 2: Barplot of Athlete Counts by Distance and Segmented by Sex*



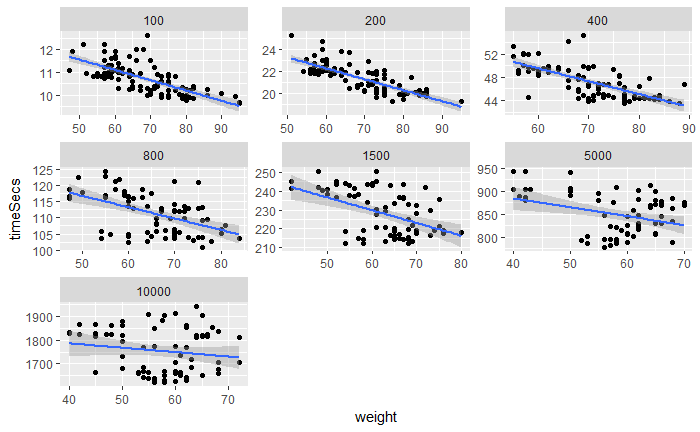
*Figure 3: Boxplots of Finishing Time (seconds) by Sex (Male: right, Female: left) grouped by Race Distance*



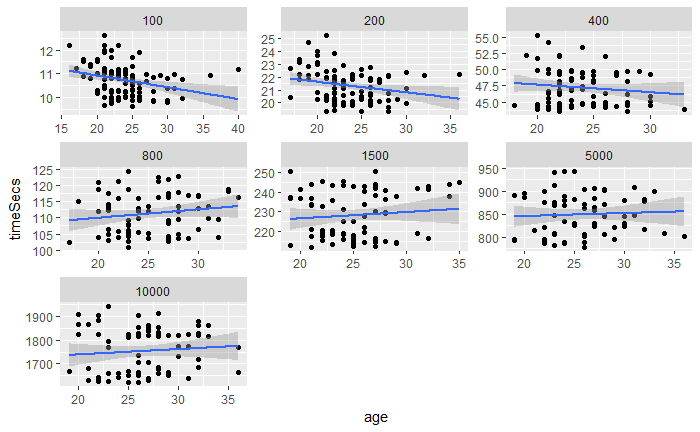
*Figure 4: Barplot of Medaled Olympic Athletes by Sex*



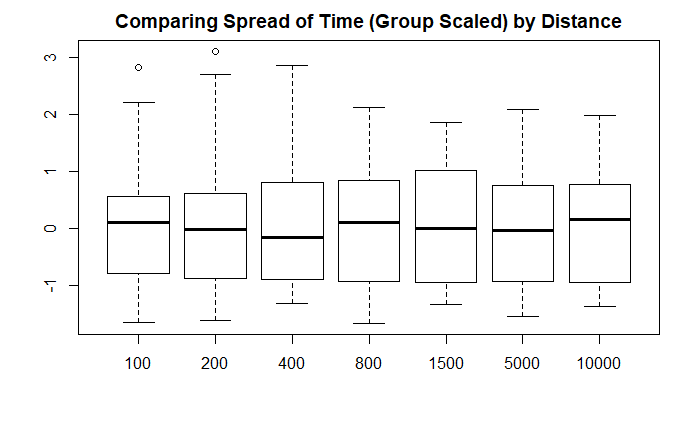
*Figure 5: Scatterplot of Finishing Time (seconds) vs. Height grouped by Race Distance*



*Figure 6: Scatterplot of Finishing Time (seconds) vs. Weight grouped by Race Distance*



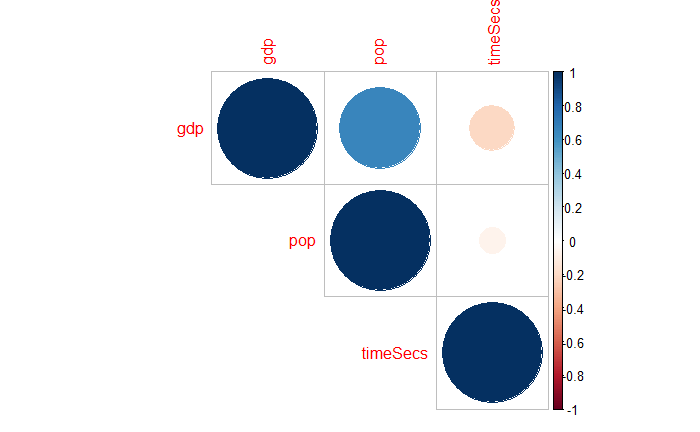
*Figure 7: Scatterplot of Finishing Time (seconds) vs. Age grouped by Race Distance*



*Figure 8: Boxplots of (Distance Group Scaled) Time by Distance*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **BAR BDI BLR BRA BRN CIV COL DOM ERI LUX MEX QAT UGA AUT HUN IRL MOZ NOR PAN TUR UKR ESP CHN NAM** | | | | | | | | | | | |
| **1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 4 4** | | | | | | | | | | | |
| **TUN BEL POR CUB RUS TTO ROU POL SWE CAN ITA MAR NZL FRA AUS FIN ETH KEN JAM GBR USA** | | | | | | | | | | |  |
| **4 5 6 7 7 9 10 11 11 12 12 12 12 16 29 33 34 39 45 54 177** |  |  |  |  |  |  |  |  |  |  |  |

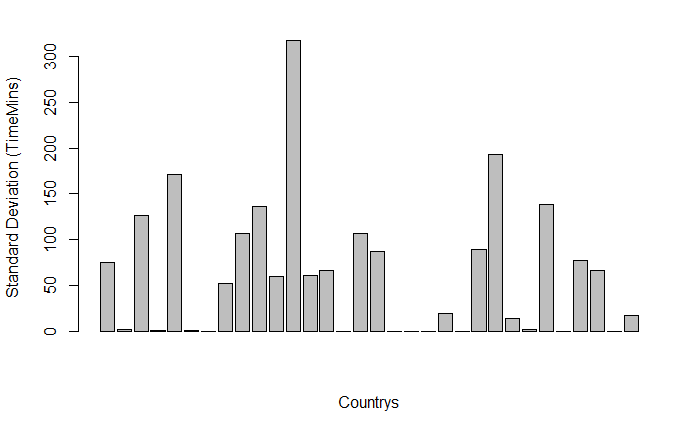
*Table 2: Medaled Athletes by Country*



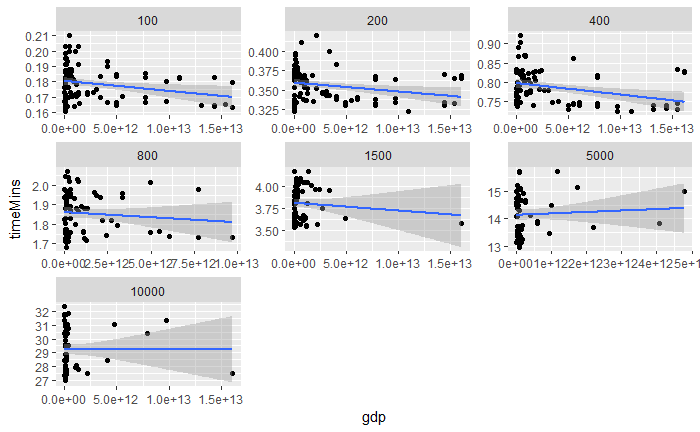
*Figure 9: Correlation Plot of Level 2 Quantitative Variables and Response*

|  |  |  |  |
| --- | --- | --- | --- |
|  | gdp | pop | timeSecs |
| gdp | 1 | 0.653682 | -0.2048 |
| pop |  | 1 | -0.06902 |
| timeSecs |  |  | 1 |

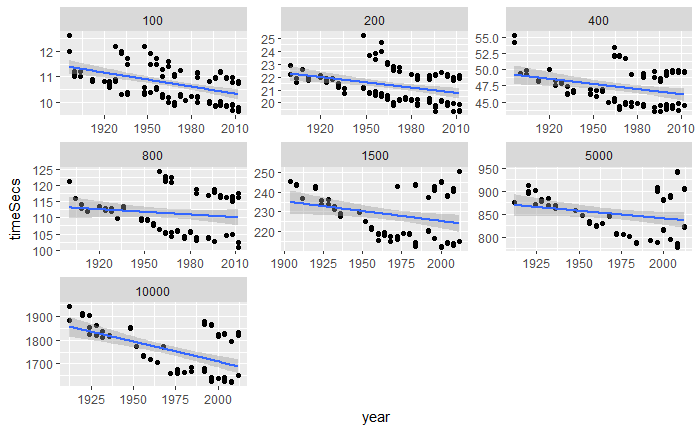
*Table 3: Correlation Coefficients of Level 2 Quantitative Variables and Response*



*Figure 10: Barplot of Standard Deviation of Finishing Times by Country*



*Figure 11: Scatterplot of GDP versus Finishing Time grouped by Event Distance*



*Figure 12: Scatterplot of Finishing Time (seconds) vs. Age grouped by Race Distance*

|  |
| --- |
|  |
| lm(formula = track$timeSecs ~ track$year + track$dist) |
|  |
| Residuals: |
| Min 1Q Median 3Q Max |
| -114.288 -16.199 2.434 12.669 180.290 |
|  |
| Coefficients: |
| Estimate Std. Error t value Pr(>|t|) |
| (Intercept) 512.325875 96.613845 5.303 1.62e-07 \*\*\* |
| track$year -0.271066 0.049081 -5.523 5.03e-08 \*\*\* |
| track$dist 0.176746 0.000479 369.021 < 2e-16 \*\*\* |
| --- |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |

*Table 4: Multiple Linear Regression of Finishing Time (seconds) vs. Year, Distance*