This study aims to statistically assess the effects of training schedules and uncontrollable external factors on mental preparedness, as measured by trait emotional intelligence in 100-kilometer ultra-marathon runners. These factors include elevation gain, course surface type, average weekly mileage, and demographic characteristics. By examining these factors, the study seeks to offer a technical comprehension of the interplay between mental and physical readiness, which is essential for improving endurance sports performance. This work supports my own objective of utilizing data-driven techniques to maximize athletic training in sports like boxing and soccer, merging mental and physical fitness to create evidence-based performance gains. It also has practical importance for sports performance research.

Eric Samtleben's (2023) study, which looks at the relationship between emotional intelligence and performance in 100-kilometer ultra-marathon runners, is the source of the dataset. It includes assessments of emotional intelligence together with important demographic, racial, and physical training factors based on self-reported data from 288 individuals. The study was approved by the Trent University Research Ethics Board, and participants were sourced online using sites including Facebook, Reddit, Strava, and Raceroster.com. This extensive dataset provides insightful information about the physical and psychological aspects of ultra-marathon performance.

This dataset is especially helpful for investigating how mental preparation, as determined by emotional intelligence, is impacted by practice and outside, uncontrolled events. In my exploratory data analysis (EDA), I considered predictors like age, sex, pb_surface, pb_elev, pb100k_dec, and avg_km, but I also concentrated on important factors like teique_sf, the main measure of trait emotional intelligence. These factors allow us to evaluate the potential interactions between an athlete's mental readiness and the physical demands of ultra-marathon running. Clear definitions and ranges for every variable are provided by the comprehensive data dictionary and related documentation, guaranteeing that my analyses are based on a thorough comprehension of the underlying components. The dataset is extremely valuable for research attempting to create more comprehensive training and performance plans in endurance sports since it integrates psychological and physical metrics.

For my analysis, I started by importing the ultra-running dataset and eliminating any rows with missing values. Ten factors make up the dataset, which also includes assessments of emotional intelligence, physical performance metrics, and demographic data. In addition to predictors like age, sex, pb_surface (race surface type), pb_elev (elevation gain), pb100k_dec (personal best time in decimal hours), and avg_km (average kilometers ran each week), my main variables of interest are the trait emotional intelligence score (teique_sf).

I began by analyzing the distributions of the categorical variables, pb_surface and sex, to spot any possible outliers and have an idea of the distribution of these variables. A possible outlier among females is shown in the boxplot of teique_sf by sex (see Figure 1), which may be a result of data input errors or a real variance in the population. Likewise, Figure 2's boxplot of teique_sf by pb_surface raises the possibility that there is an outlier in the "trail" group with a lower emotional intelligence score.

Figure 1

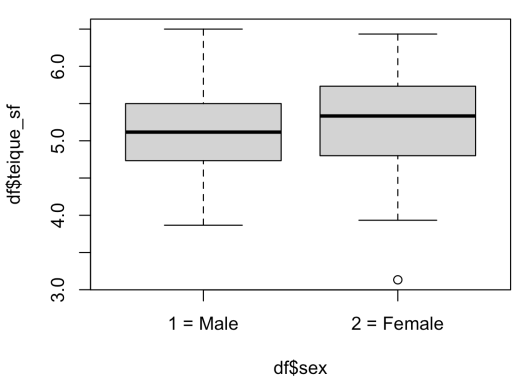
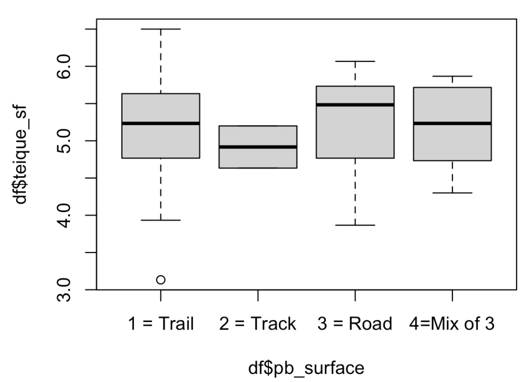


Figure 2



The frequency distributions of sex and pb_surface were also displayed using bar plots (Figures 3 and 4), demonstrating that the dataset is evenly represented in both categories. These illustrations assist us in determining if the final model's robust handling or additional transformation of the categorical components is necessary.

Figure 3

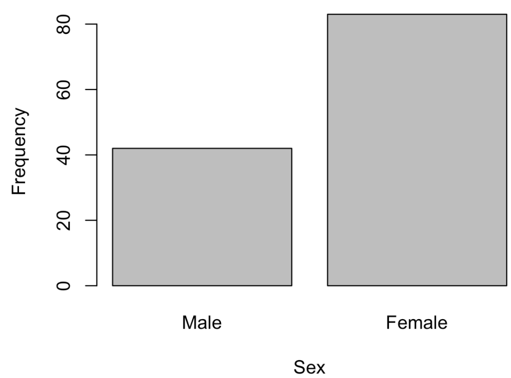
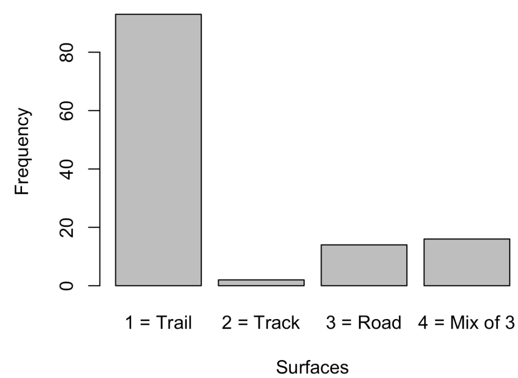


Figure 4



Then I looked at each numerical predictor separately. A possible upper outlier is indicated by age histograms and boxplots (Figures 5 and 6), which imply that some participants are significantly older than the bulk of the sample.

```
summary(df$age)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 19.0    33.0    38.0    40.1    45.0    87.0
```

Figure 5

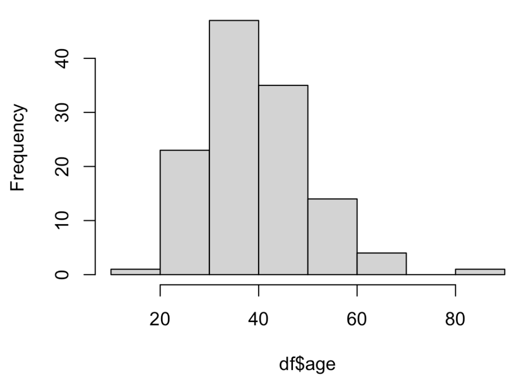
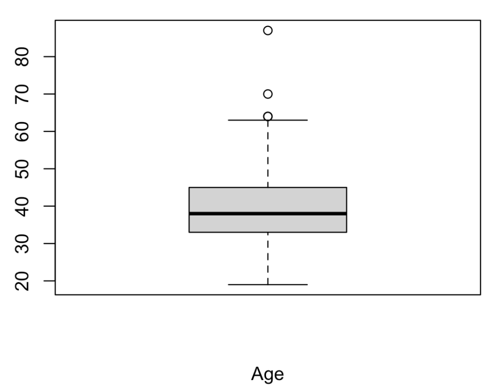


Figure 6



On the other hand, pb100k_dec seems to be regularly distributed, which is advantageous for regression modeling (Figures 7 and 8).

```
summary(df$pb100k_dec)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 6.50   12.16   14.38   14.69   17.00   23.50 
```

Figure 7

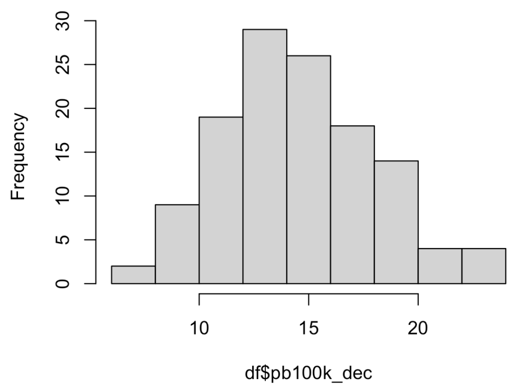
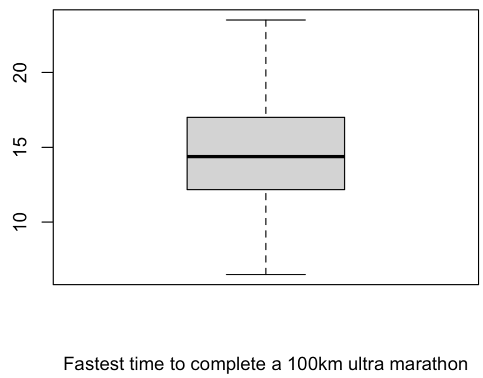


Figure 8



However, pb_elev is noticeably right-skewed (Figures 9 and 10), with several high values on the top end. This skewness implies that any model that uses pb_elev must take this non-normality into account.

```
summary(df$pb_elev)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0     800    2200    2523    3658    9000 
```

Figure 9

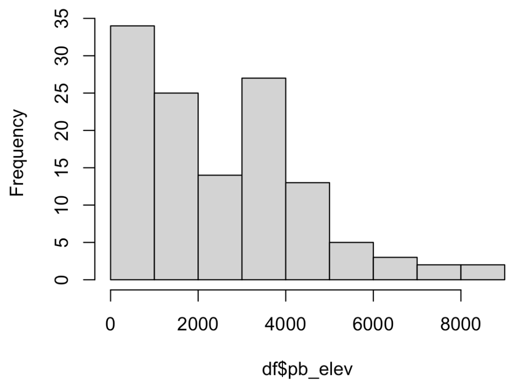


Figure 10

Although there are a few outliers at both ends, the average kilometers ran each week (avg_km) shows an approximately normal distribution (Figures 11 and 12).

```
> summary(df$avg_km)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    5.0    56.0    70.0    72.1    80.0   160.0
```

Figure 11

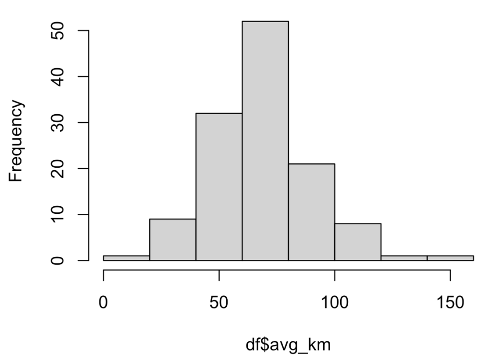
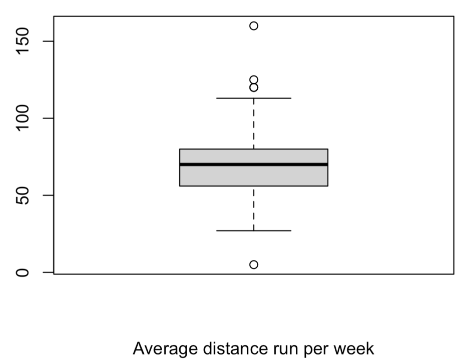
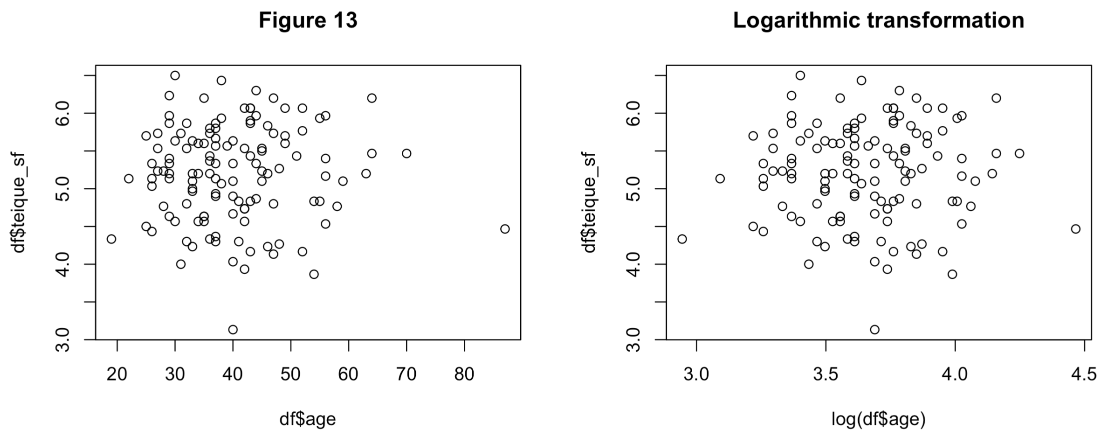


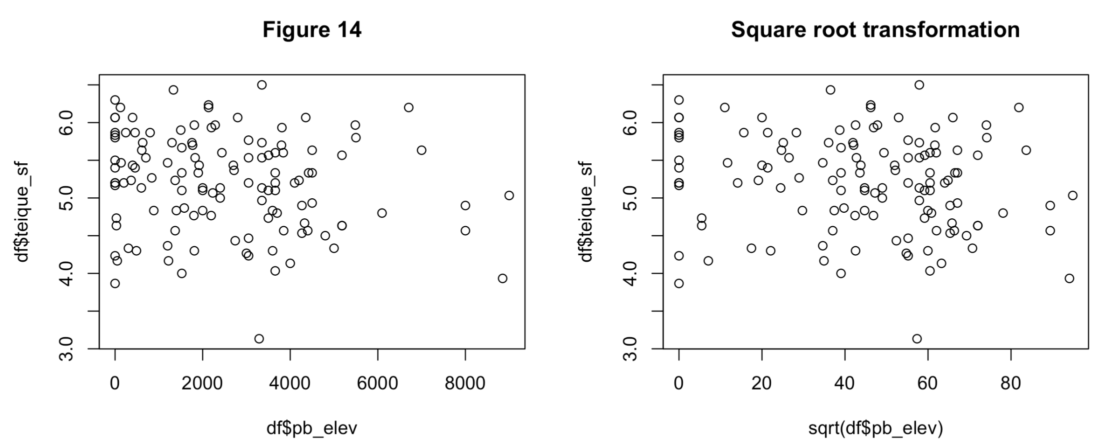
Figure 12



When comparing teique_sf to age in Figure 13, a single individual in their 80s jumps out as a possible outlier that may drag any fitted line, but there isn't a clear linear trend. The data span a wide range of elevation values with several extreme points on the high end, and no transformation significantly improved the linearity of the connection.



The data in Figure 14, which graphs teique_sf versus pb_elev spans a broad range of elevation values, with several extreme spots on the high end. Applying a square root transformation seems to provide a more balanced distribution and a little more linear scatter pattern, even if a square root transformation did not significantly increase the linearity of the connection. This implies that the square root of pb_elev could be better than the untransformed version, which might improve residual diagnostics and model fit in subsequent studies.



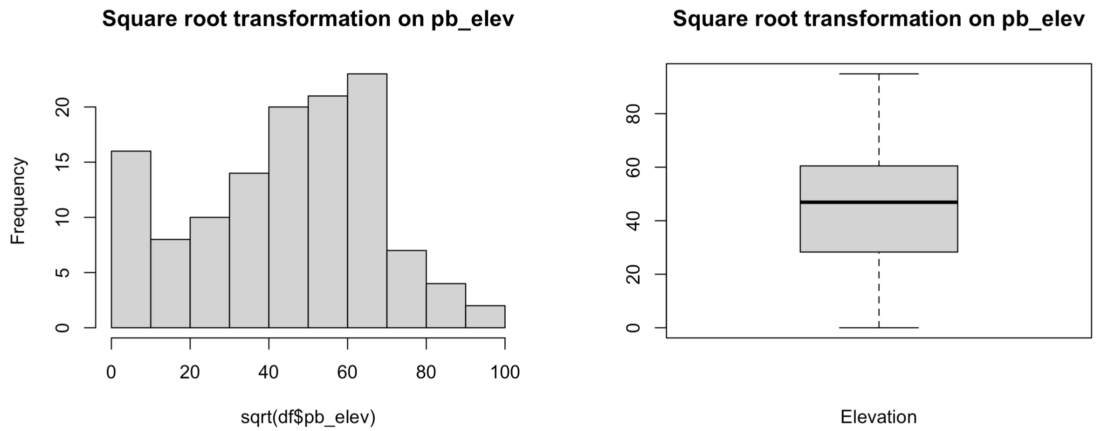
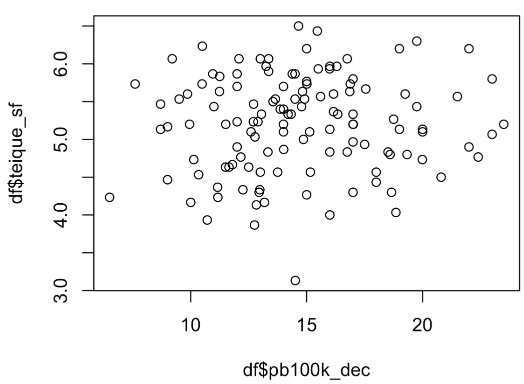
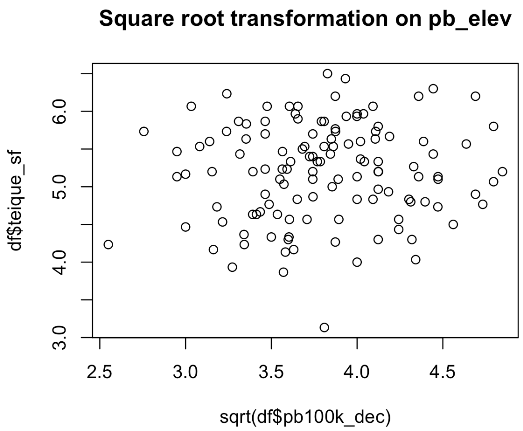


Figure 15 shows a more pronounced linear trend between teique_sf and pb100k_dec, suggesting that faster timings may be linked to higher emotional intelligence. This variable also seems to be relatively normally distributed based on previous histograms. I tried doing a square root and logarithmic transformation in pb100k_dec to have more linearity, but the results produced a nearly similar scatter plot, indicating that it could be simpler to leave the variable untransformed.

Figure 15





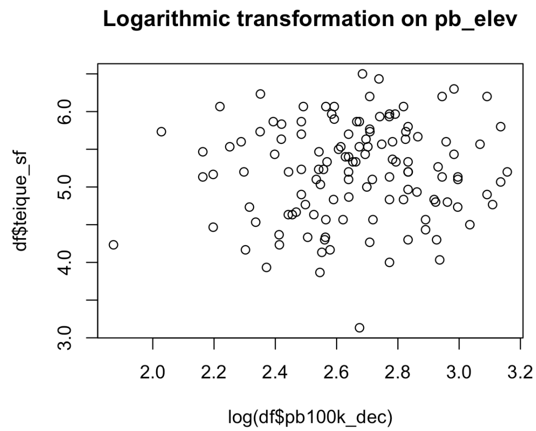
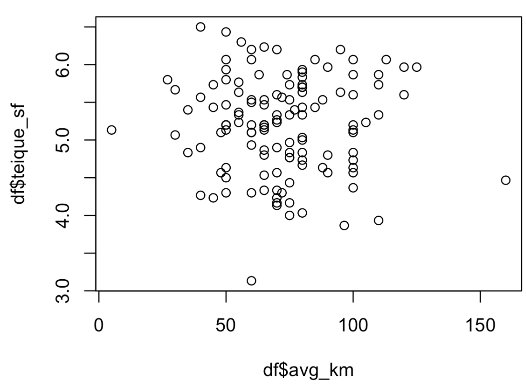
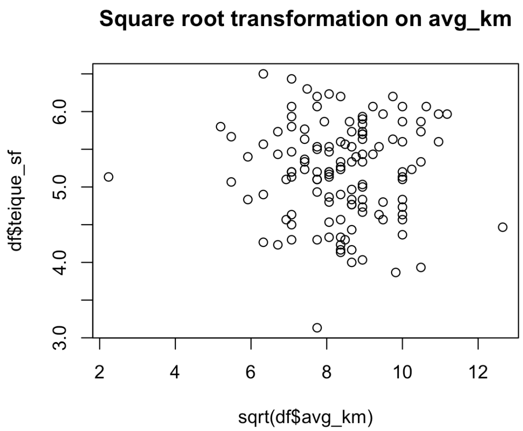
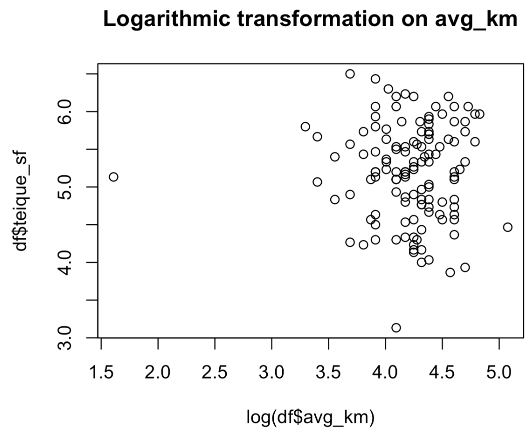


Figure 16, which displays Teique_sf vs. avg_km, shows a mostly linear trend with a few high-kilometrage outliers that might alter the slope. To obtain greater linearity as previously, I attempted a square root transformation, which just moved the data points to the right, and a logarithmic transformation, which tightened them even more while still shifting them to the right, indicating that neither transformation is helpful.

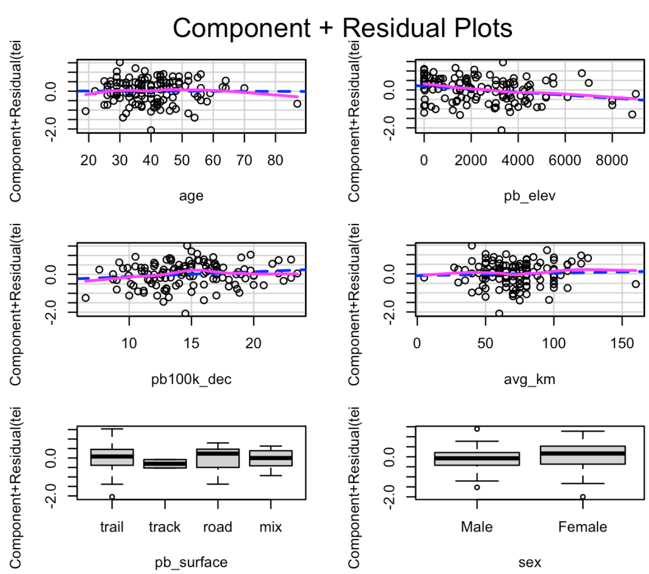
Figure 16



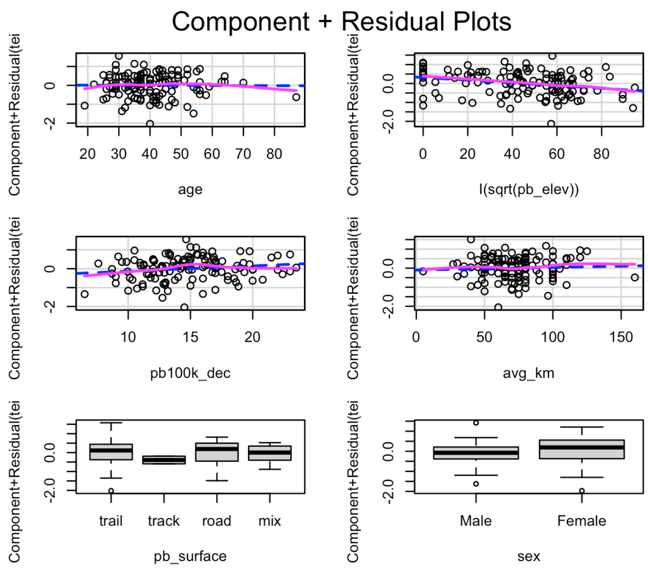




The partial regression graphs for the model utilizing the untransformed pb_elev are displayed in Figure 1. The age partial plot shows a largely horizontal band, but it is nevertheless impacted by one older person who slightly pulls the fitted line. Elevation gain has a generally linear relationship with teique_sf, though the residuals have a moderate dispersion around the fitted line. Despite a few high-kilometrage data in avg_km, pb100k_dec and avg_km preserve reasonably linear associations with teique_sf. Except for modest dispersion in the trail category, the boxplots for the categorical variables pb_surface and sex demonstrate how each category moves the response around the baseline level.



Following a square root transformation of pb_elev, the partial regression plots are shown in Figure 2. A little tighter clustering of residuals and a wider spread is shown in the partial plot for sqrt(pb_elev), indicating that this adjustment could enhance the model's capacity to represent the impact of elevation gain on teique_sf.



No matter if I used the original pb_elev or its square root transformation, it seems from the correlation matrices that none of the pairwise correlations between the numerical predictors have values greater than 0.4. This implies that the regression model's collinearity is not likely to be a significant issue. The overall pattern of low-to-moderate correlations suggests that the independent variables are largely separate, even though there is a moderate negative correlation (around –0.33) between pb100k_dec and avg_km. As a result, collinearity is most likely not an issue between the independent variables.

```
> #Correlation between independent variables
> #No correlation is high
> cor(df[, -c(2, 3, 5, 8,9, 10)])
                  age     pb_elev pb100k_dec      avg_km
age         1.0000000 -0.18024694 -0.1590007  0.20980204
pb_elev    -0.1802469  1.00000000  0.2593787  0.05380689
pb100k_dec -0.1590007  0.25937869  1.0000000 -0.32707374
avg_km      0.2098020  0.05380689 -0.3270737  1.00000000
> #Transformed pb_elev correlation between independent variables
> print(cor_transformed)
                    age sqrt_pb_elev pb100k_dec      avg_km
age           1.0000000  -0.18864963 -0.1590007  0.20980204
sqrt_pb_elev -0.1886496   1.00000000  0.2994052  0.04213896
pb100k_dec   -0.1590007   0.29940517  1.0000000 -0.32707374
avg_km        0.2098020   0.04213896 -0.3270737  1.00000000
```

Based on the patterns found in the exploratory study as well as domain knowledge, a few interactions now appear likely. For example, avg_km × sex may capture a possible variation in how weekly training volume effects trait across sexes, while age × sex may indicate if the effect of age on emotional intelligence varies between males and girls. Given that runners who contend with greater elevations could gain more, or less from increased training kilometers, sqrt_pb_elev × avg_km is another potential value. According to my exploratory investigation, the distributions seem acceptable following the necessary adjustments, and the predictors show primarily linear connections with the answer. A small number of outliers exist, but they don't appear to interfere with the general homoscedasticity or normalcy. As a result, at this point, no assumption is obviously broken, however diagnostics will be required for verification.

For the full analysis, I plan on using the previously studied predictors, as well as carefully chosen interaction terms to create a comprehensive multiple linear regression model, with teique_sf as the response variable. The model will next be improved using diagnostic tests to make sure the assumptions of linearity, homoscedasticity, and residual normality are satisfied. Additionally, models will be compared to identify the optimal mix of predictors and interactions. In the end, I want to understand how uncontrollable variables (like race surface and innate demographic traits) and controllable factors (like average mileage and training habits) interact to affect trait emotional intelligence. This information could help guide focused interventions to improve mental preparedness in endurance sports.

Works Cited:

Samtleben, E. "Ultra-running the Upcoming Sport of the Endurance World: Is Emotional Intelligence Associated with Performance?" *Journal of Multidisciplinary Research at Trent*, vol. 3, no. 1, 2021, pp. 138–154. http://creativecommons.org/licenses/by-nc-sa/4.0/.https://causeweb.org/tshs/ultra-running/

"Ultra-Running and Emotional Intelligence Data Dictionary." 2021, http://creativecommons.org/licenses/by-nc-sa/4.0/. https://causeweb.org/tshs/datasets/ultrarunning_data_dictionary.pdf

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.