Reflection for Final Portfolio

Diahmin Hawkins dlh2166@columbia.edu

12/9/2024

"The more reflective you are, the more effective you are," is a great quote by Hall and Simeral. This semester, I enrolled in a course Practical Data Analysis with Dr. Alice Paul where my analttical and problem solving were put to the test. Throughout the semester, I have become a better statistician and more aware of the things I need to work on and things I flourish in. In this reflection, we will discuss my first two projects "Impact of Environmental Conditions on Marathon Runners'Performance Based on Gender and Age" and "".

Project 1

Project 1 consisted of an exploratory data analysis that explored the impactful of environmental conditions on marathon runners' performance based upon their gender and age. This was one of my favorite projects because I actually talked with participants that participated in these marathons at the *Naragansett Run Club* to obtain a personal connection and understanding of their perspective. Even though it was my favorite project, many changes were brought to attention and adjusted accordingly. In my `Effects of Age on Marathon Performance in Men and Women Box Plot`, I removed the plot because it has similar information as the previous plot in the study. I changed my `Effects of Marathon Performance Stratified by Race, Gender, and Age Group Boxplot` into line trending plots for diverse plotting.

In Aim 2, I was able to explore the impact of environmental conditions on marathon performance, and whetjer the impact differs across age and gender. In this Aim 2 analysis approach, I explored a linear regression model to investigate the relationship between runtime s and other key variables that consisted of weather parameters, social deemographics, and marathon races. Previously, I did not include the marathon races as one of my predictor variables. So, in efforts to see the impact of races on runtimes, this was implemented in the model. In my analysis, I noticed age was non-linear and demonstrated a somehwayt U-shpaped. Due to this non-linearity, I created a model with that did not include age to see the difference in impact and measured the AIC and BIC. The AIC and BIC were lower in the original model, which demonstrated that age has an impact on the runtimes for the marathon runners. Lastly, my model indicated high VIF due to similarities and correlations in the weather parameters.

Last, but not least, Aim 3 asked us to identify the weather parameters (WBGT, Flag conditions, temperature,etc) that have the largest impact on marathon performance. In this analysis portion, I needed to change the factors severity of the flag conditions for better interpretation and readability.

From this project, I learned more about multicolinearity and its importance. Multicollinearity is a key concept in regression analysis that occurs when two or more independent variables are highly correlated, leading to redundancy in the information they provide to the model. This can inflate the variance of the coefficient estimates, making it harder to determine the individual effect of each variable on the dependent variable. In my analysis, I noticed high VIF values among the weather parameters, which highlighted the presence of multicollinearity. Addressing this is critical because it ensures that the predictors in the model contribute unique information, making the results more interpretable and reliable. Ignoring multicollinearity can lead to misleading conclusions or overly sensitive models that react to small changes in the data. By identifying and addressing multicollinearity, I learned how to refine my models and produce results that are more stable and trustworthy for decision-making.