Derrick Robinson

DSC 530

Final Project

Exploratory Data Analysis on Injury Recovery Time

Statistical/Hypothetical Question:

The primary research question investigated in this study was: *Do taller athletes tend to have a longer recovery time?* This hypothesis was tested using the Injury Prediction Dataset from Kaggle, which contains various metrics related to athletes' injuries, including height, age, and recovery duration.

Outcome of Exploratory Data Analysis (EDA)

The EDA involved descriptive statistics, visualizations, and hypothesis testing to understand the relationship between height and recovery time. Histograms were used to analyze the distribution of key variables, while measures such as mean, median, mode, and spread provided insights into their central tendency and variability. Probability Mass Functions (PMFs) and Cumulative Distribution Functions (CDFs) helped in comparing different scenarios within the dataset.

Regression analysis was conducted to determine whether height significantly predicted recovery time. The results of the simple linear regression showed a weak correlation between height and recovery duration, as indicated by a low R-squared value and an insignificant p-value. This suggests that height alone is not a strong determinant of an athlete's recovery time. Additionally, a two-sample t-test comparing shorter and taller athletes failed to reveal a statistically significant difference in mean recovery time.

Missed Aspects During Analysis

One limitation of the analysis was the exclusion of other physiological and external factors that might impact recovery time, such as injury type, treatment methods, training regimens, and nutrition. These variables could have provided a more holistic view of what influences recovery duration. Additionally, potential outliers or missing values in the dataset may have skewed the analysis, which could have been addressed with further data preprocessing techniques.

Potentially Useful Variables

Several variables could have strengthened the analysis. For instance, injury severity and type would have provided better context for understanding differences in recovery time. Additionally, athlete weight and fitness level might also influence how quickly one recovers from injuries. Incorporating categorical variables such as sport type or position played might have helped identify trends specific to different athletic roles.

Assumption and Validity

A major assumption made was that height alone could serve as a significant predictor of recovery time. However, the analysis suggested that other factors likely play a larger role. Another assumption was that the dataset was representative of a broader athletic population, but without demographic details, this could not be verified. Lastly, it was assumed that the data was recorded accurately and consistently, yet inconsistencies in data collection could introduce bias.

Challenges and Areas of Uncertainty

One of the primary challenges faced during the analysis was understanding how to apply different statistical methods correctly, especially when testing for correlations and causation. Determining whether to use a linear regression model or a more complex nonlinear approach required careful consideration. Additionally, interpreting the results of statistical tests such as the t-test and Pearson's correlation required deeper statistical knowledge. Lastly, understanding potential confounding variables and how to control for them remained an area of difficulty.

Conclusion

While the initial hypothesis suggested that taller athletes may have longer recovery times, the analysis did not provide strong statistical evidence to support this claim. The weak correlation indicates that height alone is not a significant predictor of recovery time, and additional factors should be considered for a more comprehensive analysis. Future research incorporating more explanatory variables and a larger, more detailed dataset would likely yield more meaningful insights into the factors influencing injury recovery.