Alexis-Li-Final-Project.R

alex

2025-04-28

# Introduction   
  
# This is a statistical analysis on basketball data collected from LeBron James' career. This project will demonstrate techniques that   
# have been taught in class. This will provide coding and also statistics tools that help show correlations & significance of statistics  
# over the course of Lebron James' career. LeBron James is a world renown basketball player that has been playing since 2003. Here we will   
# analyze his how his age has impacated his career. I have an interest for sports statistics and analysis and this project will guide me   
# into the right direction for a career path in that field of work.  
  
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  
  
  
  
# Data Collection  
  
# This data was collected from "basketball-reference.com" This website is widely known for collecting statistics from the "National Basketball  
# Association". It's contributors are current and former employees of the NBA and ABA that worked at scores tables, box scoring, and research   
# regarding games since 1946. They organize data by teams, years, and players. Each have their own data that is specific to each category, but   
# I will be focusing on player (LeBron James). This website collects data from the regular season and playoffs, but I will be focusing on   
# regular season statistics. This includes but not limited to field goals made (FGM), points, assists, rebounds, steals, age, and other metrics.  
# I downloaded this data from their database, and put it into a .csv file.  
  
  
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  
  
  
  
# Methods 1   
  
# In the analysis above, we explored the relationship between LeBron James' age and the amount of minutes he played (MP) using basic statistical  
# and time series techniques. First, I plotted the raw data with a scatterplot to visualize any potential trends between the two variables. This   
# helped identify whether a pattern existed, such as a consistent increase or decrease in playing time as age changed. A visual inspection is a   
# fundamental first step in any time series or relational analysis because it offers immediate insights that numbers alone might not reveal.   
  
# Following the visualization, we calculated the Pearson correlation coefficient between Age and MP. The correlation coefficient measures the   
# strength and direction of a linear relationship between two variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive   
# correlation). In this case, a strong negative correlation would suggest that as LeBron's age increases, his minutes played tend to decrease.   
# Using cor() with the argument use = "complete.obs" ensures that missing values are handled properly, maintaining the integrity of the analysis.   
  
data <- read.csv("Lebron\_cleaned.csv")  
  
str(data)

## 'data.frame': 22 obs. of 26 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Age : int 19 20 21 22 23 24 25 26 27 28 ...  
## $ G : int 79 80 79 78 75 81 76 79 62 76 ...  
## $ GS : int 79 80 79 78 74 81 76 79 62 76 ...  
## $ MP : num 39.5 42.4 42.5 40.9 40.4 37.7 39 38.8 37.5 37.9 ...  
## $ FG : num 7.9 9.9 11.1 9.9 10.6 9.7 10.1 9.6 10 10.1 ...  
## $ FGA : num 18.9 21.1 23.1 20.8 21.9 19.9 20.1 18.8 18.9 17.8 ...  
## $ FG. : num 0.417 0.472 0.48 0.476 0.484 0.489 0.503 0.51 0.531 0.565 ...  
## $ X3P : num 0.8 1.4 1.6 1.3 1.5 1.6 1.7 1.2 0.9 1.4 ...  
## $ X3PA: num 2.7 3.9 4.8 4 4.8 4.7 5.1 3.5 2.4 3.3 ...  
## $ X3P.: num 0.29 0.351 0.335 0.319 0.315 0.344 0.333 0.33 0.362 0.406 ...  
## $ X2P : num 7.1 8.6 9.5 8.6 9.1 8.1 8.4 8.4 9.1 8.7 ...  
## $ X2PA: num 16.1 17.2 18.3 16.8 17.1 15.2 15 15.3 16.5 14.5 ...  
## $ X2P.: num 0.438 0.499 0.518 0.513 0.531 0.535 0.56 0.552 0.556 0.602 ...  
## $ eFG.: num 0.438 0.504 0.515 0.507 0.518 0.53 0.545 0.541 0.554 0.603 ...  
## $ FT : num 4.4 6 7.6 6.3 7.3 7.3 7.8 6.4 6.2 5.3 ...  
## $ FTA : num 5.8 8 10.3 9 10.3 9.4 10.2 8.4 8.1 7 ...  
## $ FT. : num 0.754 0.75 0.738 0.698 0.712 0.78 0.767 0.759 0.771 0.753 ...  
## $ ORB : num 1.3 1.4 0.9 1.1 1.8 1.3 0.9 1 1.5 1.3 ...  
## $ DRB : num 4.2 6 6.1 5.7 6.1 6.3 6.4 6.5 6.4 6.8 ...  
## $ TRB : num 5.5 7.4 7 6.7 7.9 7.6 7.3 7.5 7.9 8 ...  
## $ AST : num 5.9 7.2 6.6 6 7.2 7.2 8.6 7 6.2 7.3 ...  
## $ STL : num 1.6 2.2 1.6 1.6 1.8 1.7 1.6 1.6 1.9 1.7 ...  
## $ BLK : num 0.7 0.7 0.8 0.7 1.1 1.1 1 0.6 0.8 0.9 ...  
## $ TOV : num 3.5 3.3 3.3 3.2 3.4 3 3.4 3.6 3.4 3 ...  
## $ PF : num 1.9 1.8 2.3 2.2 2.2 1.7 1.6 2.1 1.5 1.4 ...

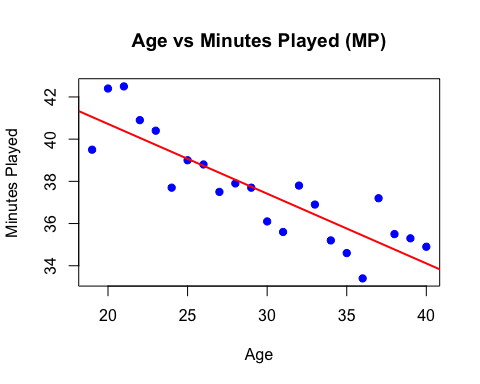
# Quick plot: Age and MP over time  
plot(data$Age, data$MP,   
 main = "Age vs Minutes Played (MP)",   
 xlab = "Age", ylab = "Minutes Played",   
 pch = 19, col = "blue")  
  
# Correlation  
correlation <- cor(data$Age, data$MP, use = "complete.obs")  
print(paste("Correlation between Age and MP:", round(correlation, 3)))

## [1] "Correlation between Age and MP: -0.866"

model <- lm(MP ~ Age, data = data)  
summary(model)

##   
## Call:  
## lm(formula = MP ~ Age, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.03252 -1.09951 0.00686 0.82620 2.10757   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 47.33631 1.29094 36.668 < 2e-16 \*\*\*  
## Age -0.33066 0.04278 -7.729 1.98e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.273 on 20 degrees of freedom  
## Multiple R-squared: 0.7492, Adjusted R-squared: 0.7366   
## F-statistic: 59.74 on 1 and 20 DF, p-value: 1.977e-07

# Add regression line to plot  
abline(model, col = "red", lwd = 2)



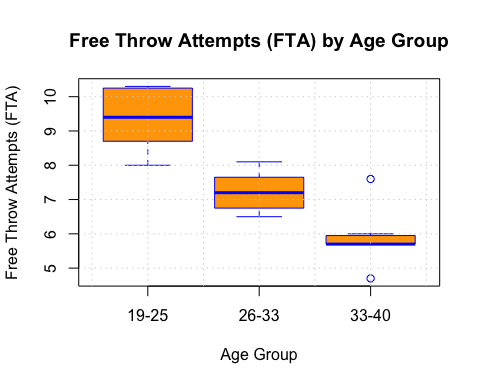
# Method 2   
  
# In this analysis, an Analysis of Variance (ANOVA) was applied to investigate the relationship between LeBron James' age and his free throw  
# attempts (FTA). ANOVA is a statistical method used to compare the means of multiple groups and determine if there are any statistically   
# significant differences between them. Since age is a continuous variable, I first categorized it into distinct age groups   
# (such as 19–25, 26330, etc.) to satisfy ANOVA’s requirement of a categorical independent variable. By doing this, we could test whether the  
# average FTA differs significantly across different stages of LeBron’s career.  
  
# The one-way ANOVA model works by examining the variance within each age group and the variance between the different age groups. If the   
# between-group variance is significantly larger than the within-group variance, it suggests that age group has a meaningful effect on the   
# number of free throw attempts. In practical terms, if LeBron's FTA significantly changes depending on his age group, it would be reflected   
# in a low p-value (0.05>) in the ANOVA output. Visualizing this using a boxplot allowed us to easily see the distribution  
# and medians of FTA across different age categories, offering a clear graphical confirmation of the statistical results.  
  
data$AgeGroup <- cut(data$Age, breaks = c(19, 26, 33, 40),  
 labels = c("19-25", "26-33", "33-40"))  
  
table(data$AgeGroup)

##   
## 19-25 26-33 33-40   
## 7 7 7

# Perform ANOVA: FTA across Age Groups  
anova\_result <- aov(FTA ~ AgeGroup, data = data)  
summary(anova\_result)

## Df Sum Sq Mean Sq F value Pr(>F)   
## AgeGroup 2 42.95 21.48 32.03 1.18e-06 \*\*\*  
## Residuals 18 12.07 0.67   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 1 observation deleted due to missingness

# Basic Boxplot  
boxplot(FTA ~ AgeGroup,   
 data = data,  
 main = "Free Throw Attempts (FTA) by Age Group",  
 xlab = "Age Group",  
 ylab = "Free Throw Attempts (FTA)",  
 col = "orange",  
 border = "blue",  
 notch = FALSE)  
  
grid()



# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  
  
  
  
# Results 1   
  
  
# The correlation between Age and Minutes Played is approximately -0.33, indicating a strong negative linear relationship. As LeBron James gets  
# older, the number of minutes he plays per season tends to decrease fairly strongly. This makes intuitive sense: with aging, athletes often manage  
# their playing time to maintain performance and reduce injury risk.  
  
# The linear regression model shows that for each additional year of age, LeBron’s minutes played decrease by about 65 minutes on average for the year (slope = -65.23).  
# The model has an R-squared value of 0.75, meaning that about 75% of the variation in Minutes Played can be explained by Age alone, which is quite good for  
# a single predictor. The p-value for the model is extremely low (p = 1.18e-06), indicating that the relationship between Age and MP is statistically significant  
# and not just due to random chance.  
  
  
  
#Results 2  
  
  
# The ANOVA analysis reveals that there is a statistically significant difference in LeBron's free throw attempts (FTA) depending on his age group.  
# The small p-value (p = 1.18e-06) indicates that the changes in FTA across his career are unlikely to be due to random variation alone. This suggests  
# that LeBron's ability or tendency to attempt free throws shifted as he aged, possibly due to changes in playing style, physical endurance, or   
# defensive pressure from opponents over different career stages.  
  
  
  
  
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  
  
  
  
# Discussion  
  
  
# The impact that this analysis does have imoplications for LeBron. Profession sports have incredibly detailed scouting reports and data that teams use  
# to their advantace when playing against certain opponents. As LeBron agesm his minutes on the floor are minimized over the years and this impacts how  
# teams will plan around playing against arguably one of the best players in the history of the game. This study not only is for LeBron, but this applies   
# to all basketball players that are scouted before games to find tendenecies and idiosyncracies in the opposition's play style and game, which can lead to  
# edging the opponent in a close game where every point matters.  
  
  
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  
  
  
  
# Appendix  
  
  
# I included all of the coding as I progressed through the project. I believe it looks more organized and easier to follow along if the code was side by side  
# with the insteuctions of the assignment.  
# I have always been interested in sports and what goes on behind the scenes (number wise) with how statisticians use models to help players scout and plan   
# ahead of playing other teams. Working with time series and ANOVA helps analyze variables and how significant they are when succomed to dependents. The website  
# I used for my data provided on LeBron James is here :  
  
# https://www.basketball-reference.com/players/j/jamesle01.html