Are athletes good students?

Jack Lin

First glance of dataset

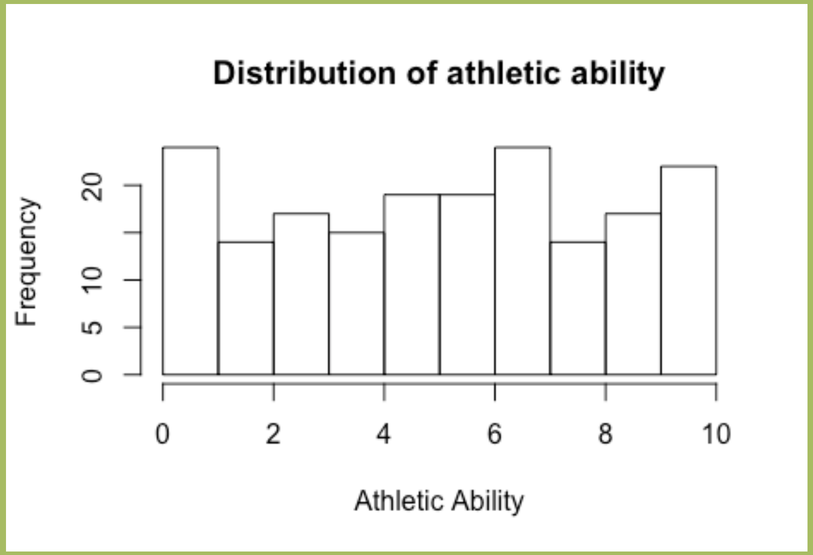


Fig 1. Balance distribution of athletic ability

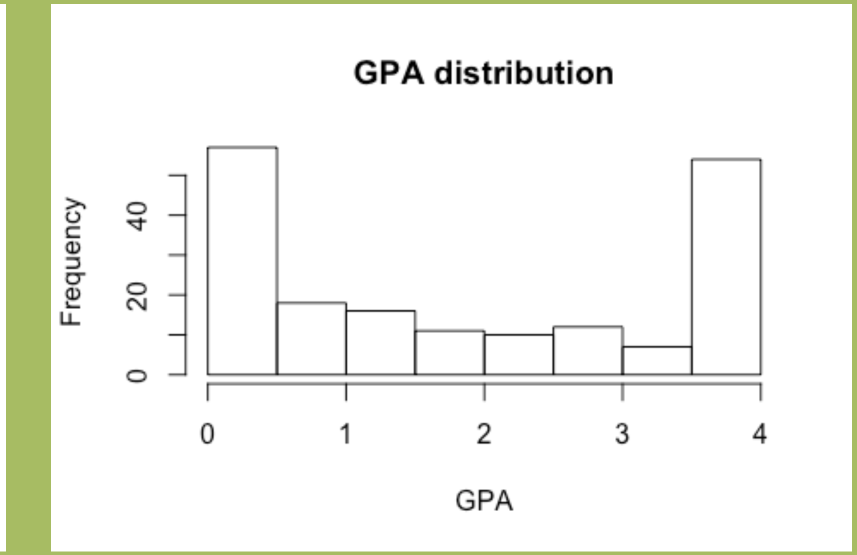


Fig 2. Bipolar GPA distribution

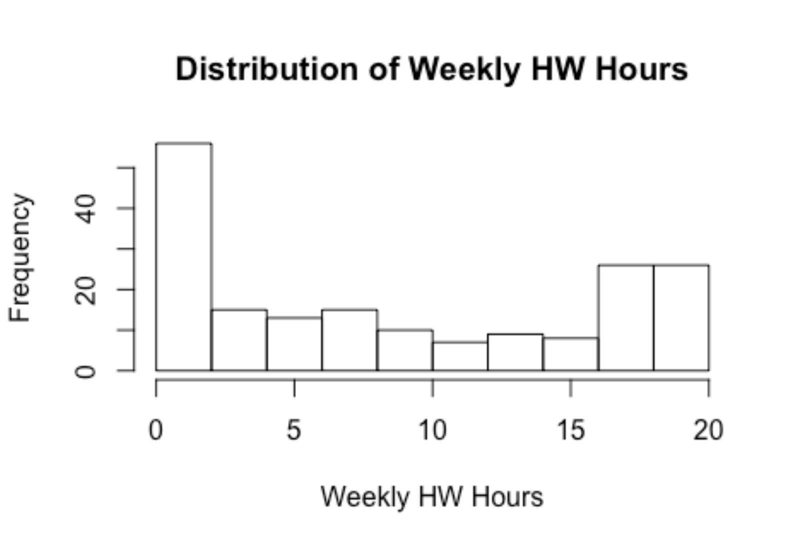


Fig 3. A considerable number of lazy students…

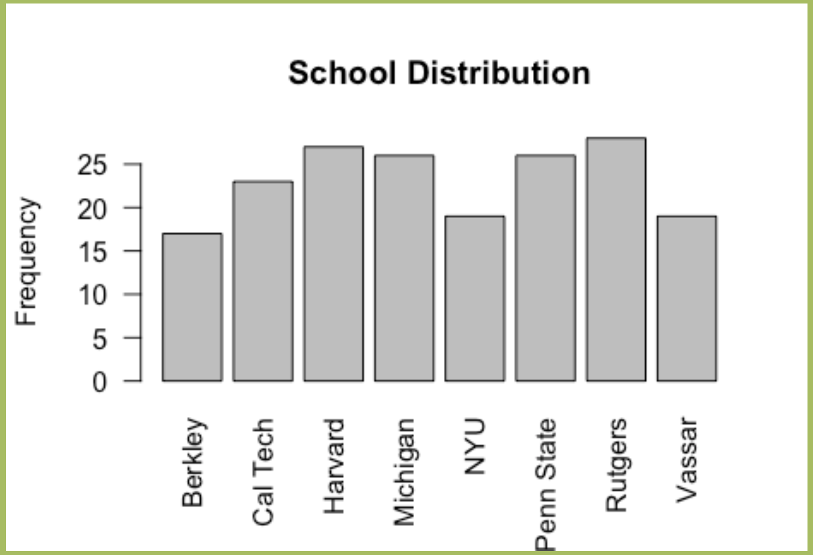


Fig 4. Balanced representation among schools

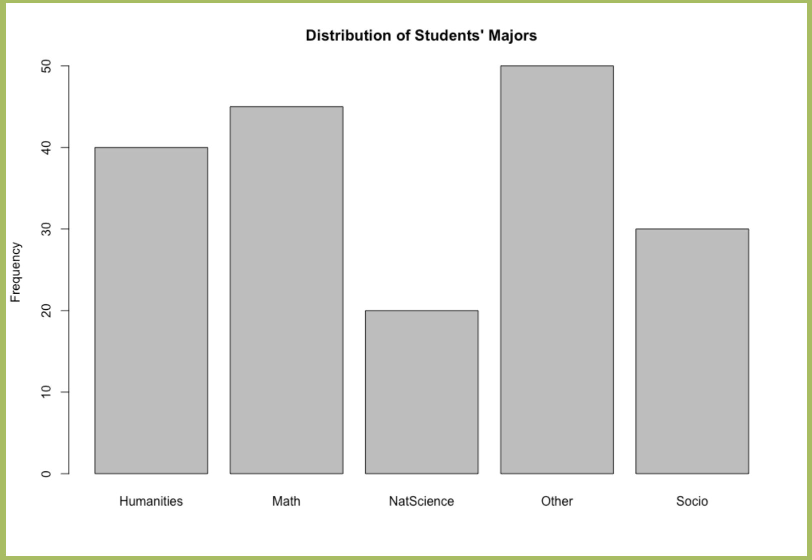


Fig 5. Natural science majors are a bit underrepresented

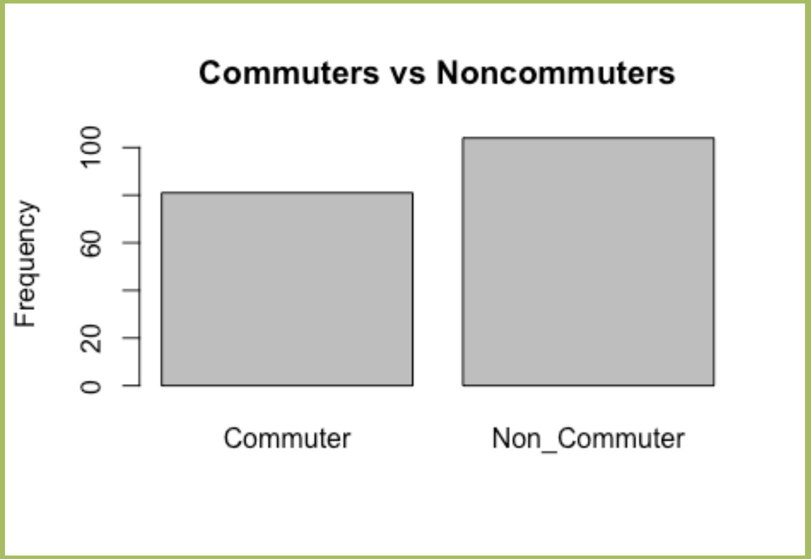


Fig 6. Commuters vs Non-commuters

No noticeable bias is observed in the raw dataset other than GPA and a large number of lazy students who don’t spend time doing homework

R code:

```
> hist(Athlete_GPA_Final$Athletic_Ability, main = "Distribution of athletic ability", xlab="Athletic Ability”)
> hist(Athlete_GPA_Final$GPA, main = "GPA distribution", xlab=“GPA")
> hist(Athlete_GPA_Final$Hours_Spent_On_Homework_Per_Week, main = "Distribution of Weekly HW Hours", xlab="Weekly HW Hours”)
> barplot(table(Athlete_GPA_Final$School), main = "School Distribution", ylab = "Frequency", las = 2)
> barplot(table(Athlete_GPA_Final$Major), main = "Distribution of Students' Majors", ylab = "Frequency", names.arg = c(“Humanities","Math","NatScience","Other","Socio"))
> barplot(table(Athlete_GPA_Final$Commuter_Status), main = "Commuters vs Noncommuters", ylab = "Frequency")
```

First glance of dataset (continued)

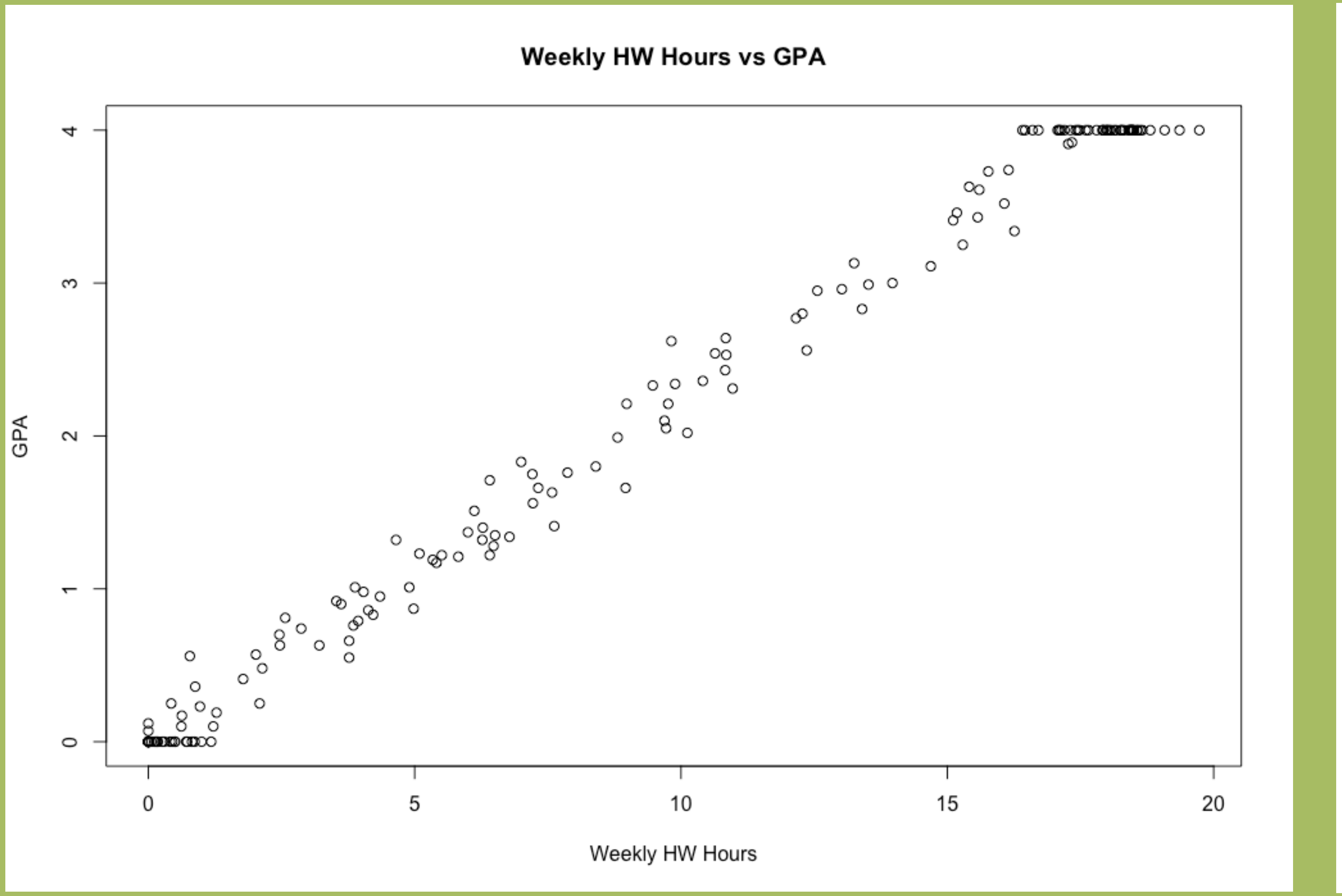


Fig 7. Weekly HW Hours vs GPA

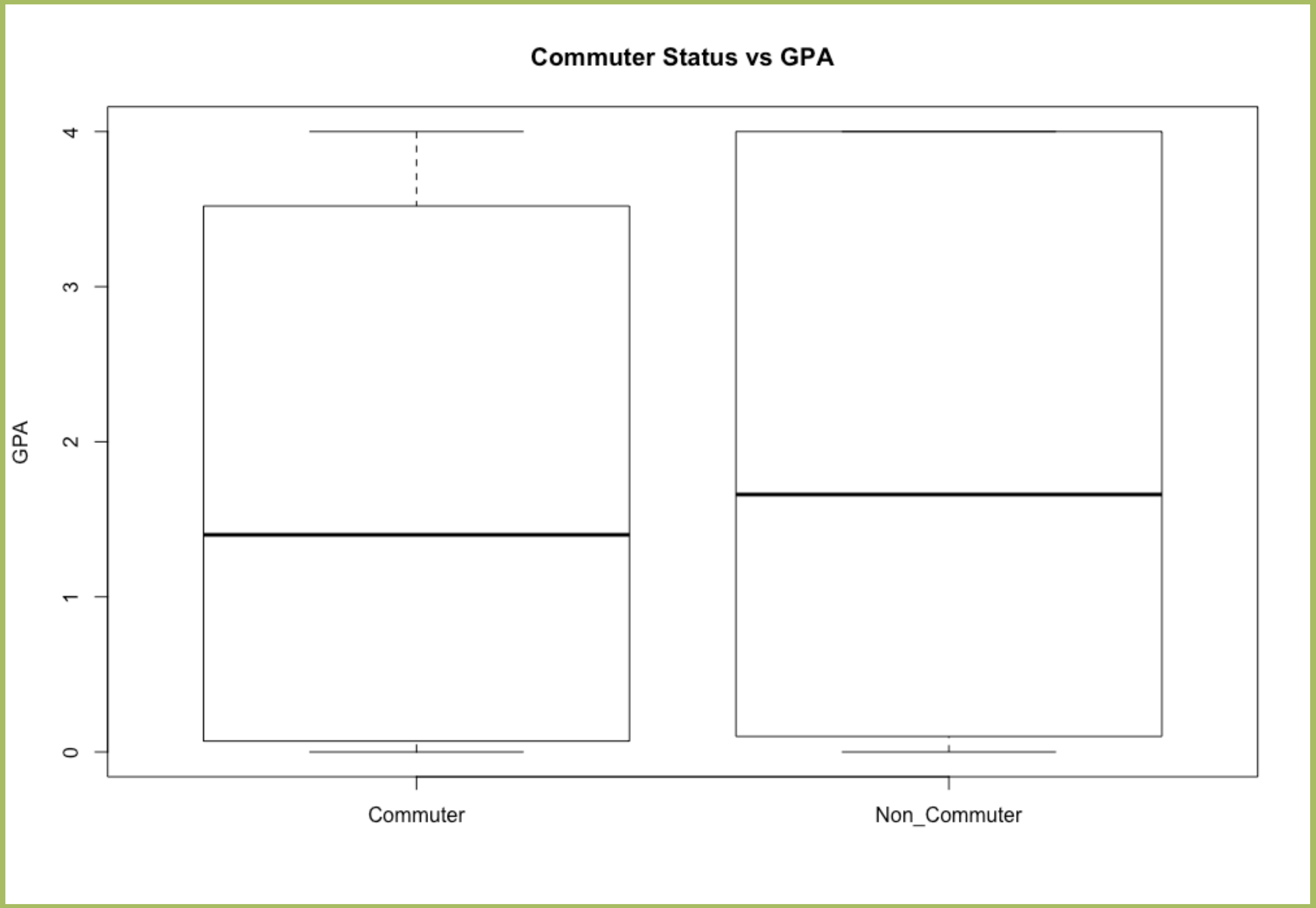


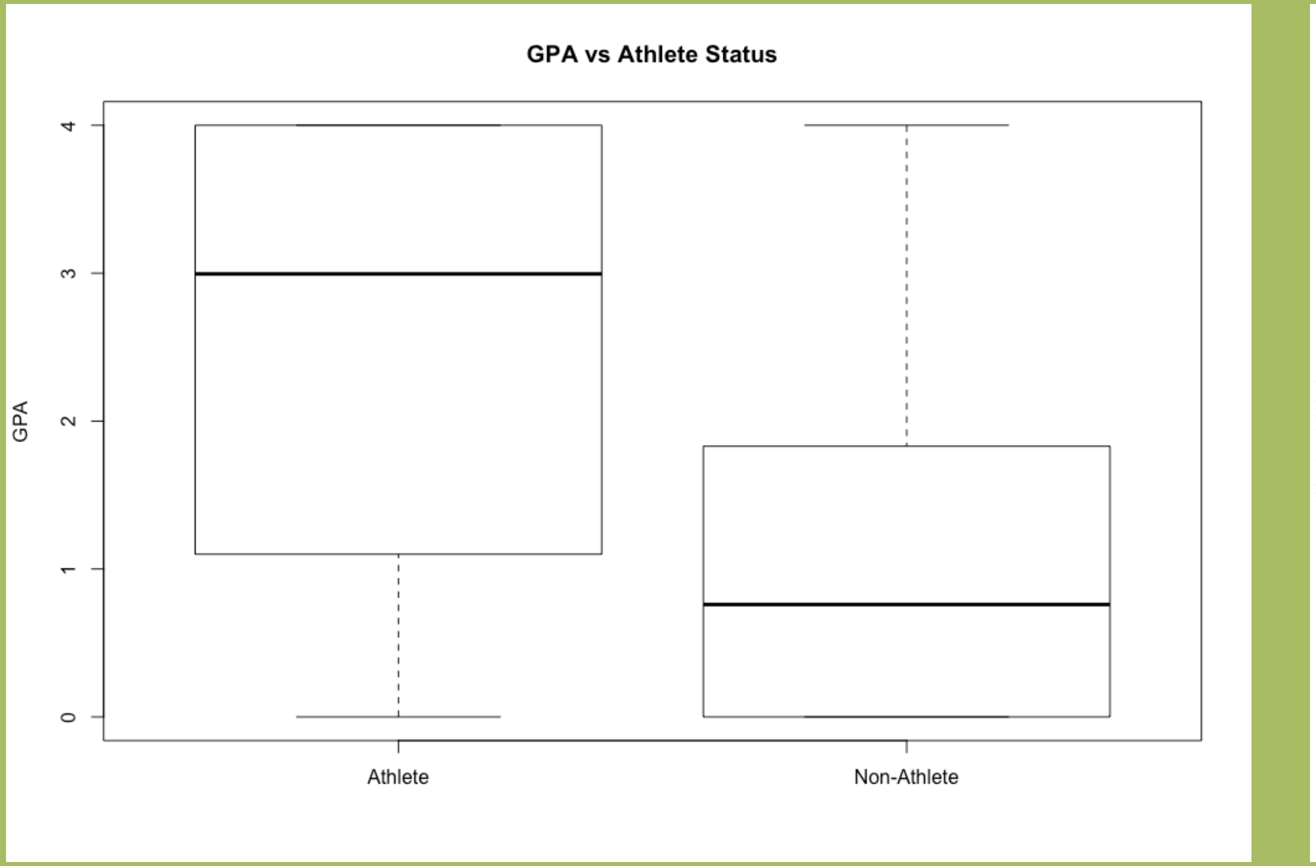
Fig 8. Commuter Status vs GPA

Fig. 7 demonstrates the positive correlation between the time spent on doing homework and GPA. Fig. 8 shows that commuter status may not affect GPA by much.

R code:

```
> plot(Athlete_GPA_Final$GPA~Athlete_GPA_Final$Hours_Spent_On_Homework_Per_Week, main = "Weekly HW Hours vs GPA", xlab = "Weekly HW Hours", ylab = “GPA”)
> boxplot(Athlete_GPA_Final$GPA~Athlete_GPA_Final$Commuter_Status, main = "Commuter Status vs GPA", ylab = "GPA")
```

Preprocessing



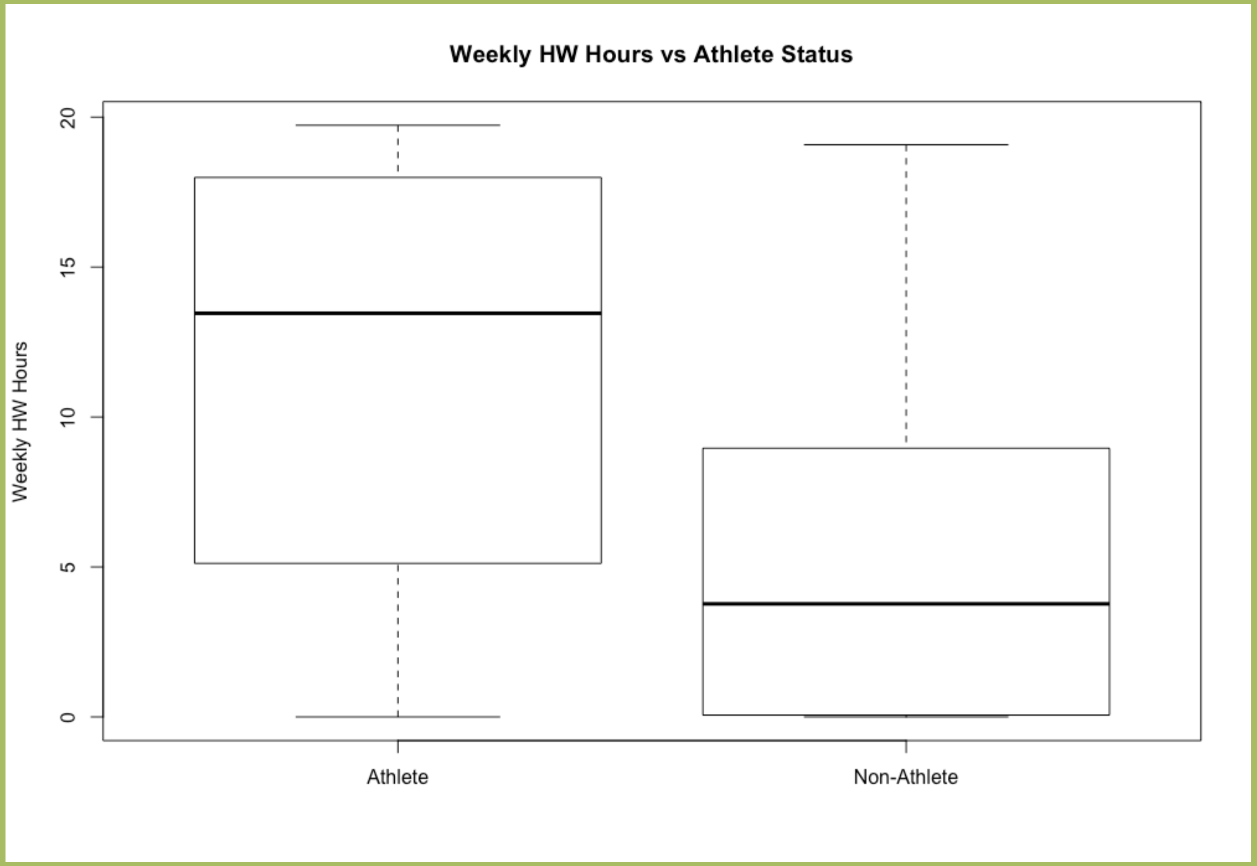


Fig 9. Student athletes' mean GPA is more than 2.0 higher than that of non-athlete students

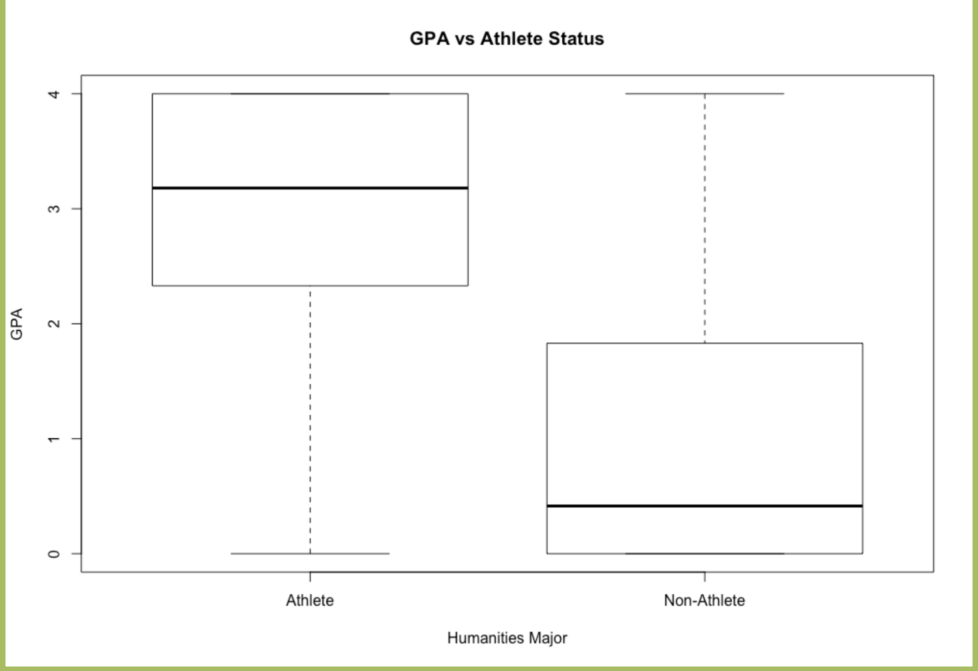
Fig 7. Student athletes spend about nine hours more weekly on homework than non-athlete students do

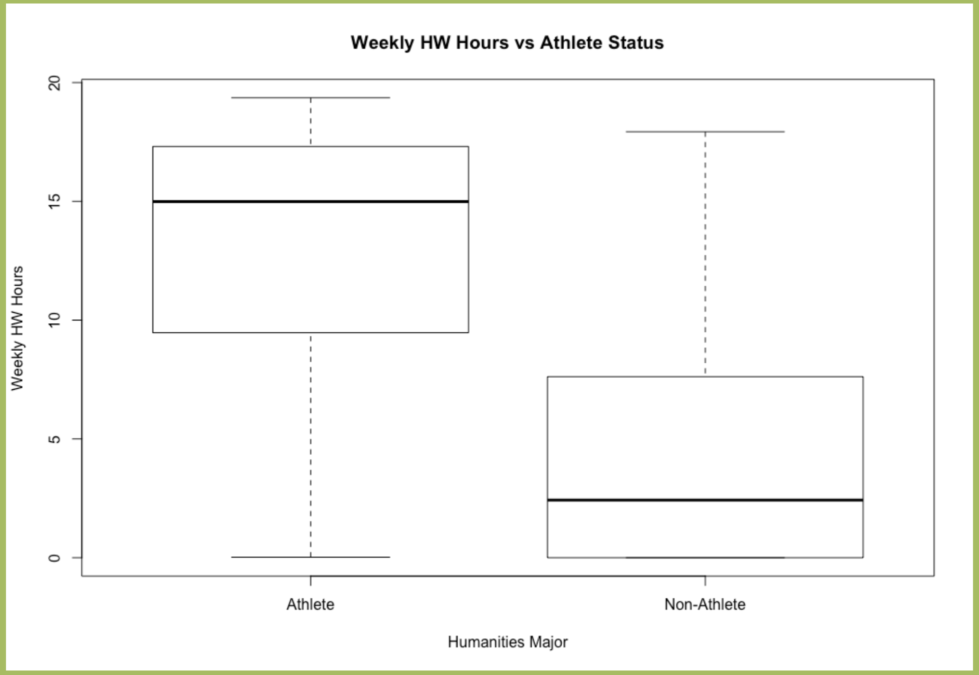
A 7th column named "Athlete" is added to differentiate student athletes and non-athletes in our dataset. The box plot demonstrates a significant difference of mean GPA between student athletes and non-athletes.

```
R code:
> Athlete_GPA_Final[,7] <- ""
> colnames(Athlete_GPA_Final)[7] <- “Athlete"
> Athlete_GPA_Final[Athlete_GPA_Final$Athletic_Ability >= 5, 7] = "Athlete"
> Athlete_GPA_Final[Athlete_GPA_Final$Athletic_Ability < 5, 7] = “Non-Athlete"
> boxplot(Athlete_GPA_Final$GPA~Athlete_GPA_Final$Athlete, main = "GPA vs Athlete Status", ylab = “GPA")
> boxplot(Athlete_GPA_Final$Hours_Spent_On_Homework_Per_Week~Athlete_GPA_Final$Athlete, main = "Weekly HW Hours vs Athlete Status", ylab = "Weekly HW Hours")
```

Different Majors

Humanities

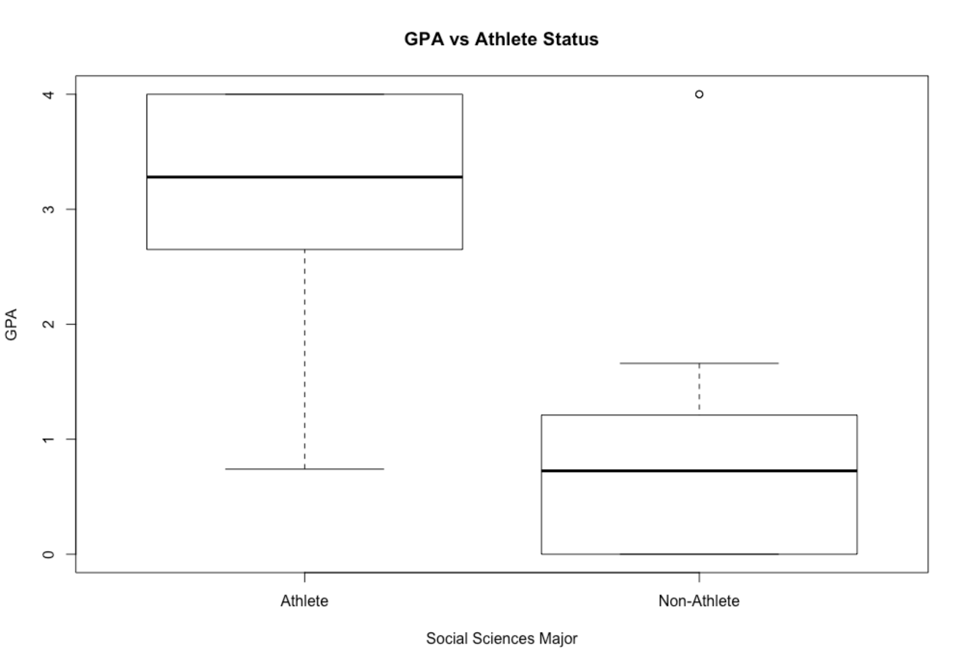


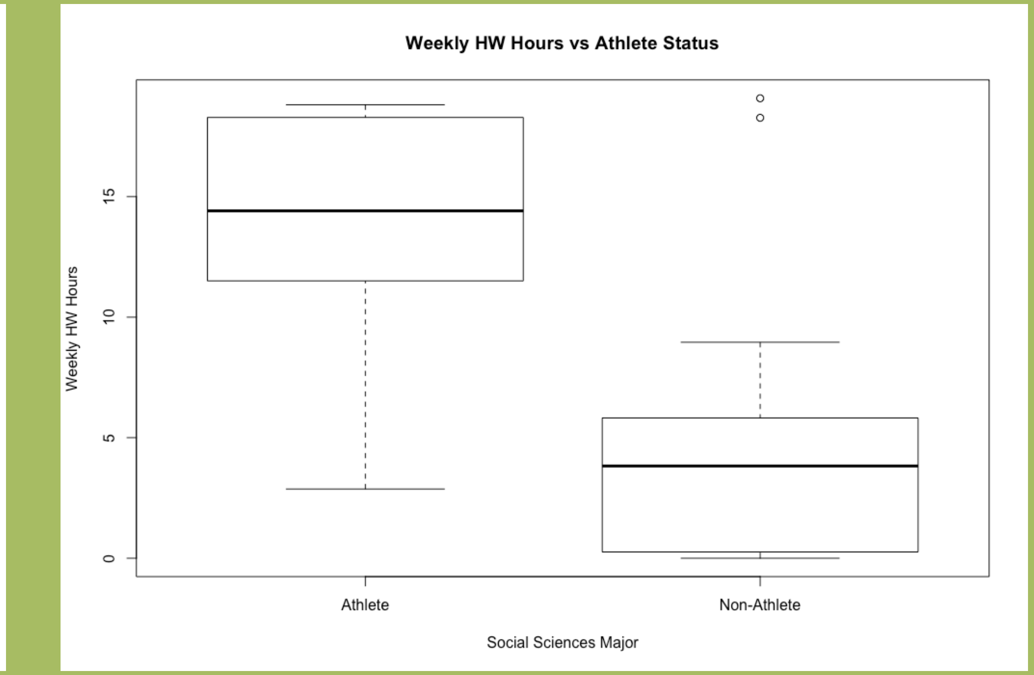


R code:

```
> humanities.data <- subset(Athlete_GPA_Final, Athlete_GPA_Final$Major == “Humanities")
> boxplot(humanities.data$GPA~humanities.data$Athlete, main = "GPA vs Athlete Status", xlab = "Humanities Major", ylab = "GPA")
> boxplot(humanities.data$Hours_Spent_On_Homework_Per_Week~humanities.data$Athlete, main = "Weekly HW Hours vs Athlete Status", xlab = "Humanities Major", ylab = "Weekly HW Hours")
```

Social Sciences



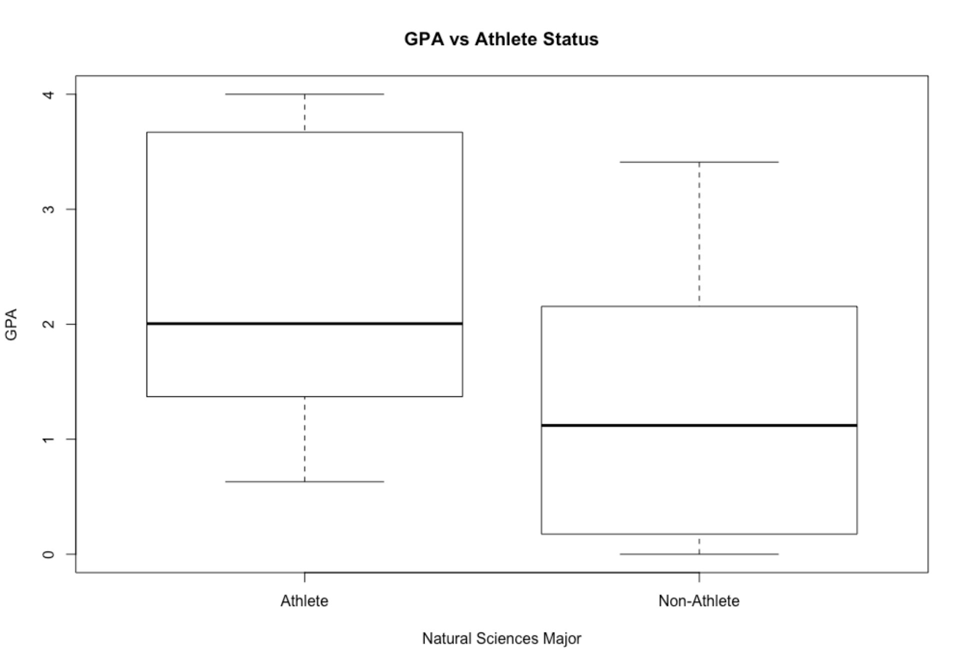


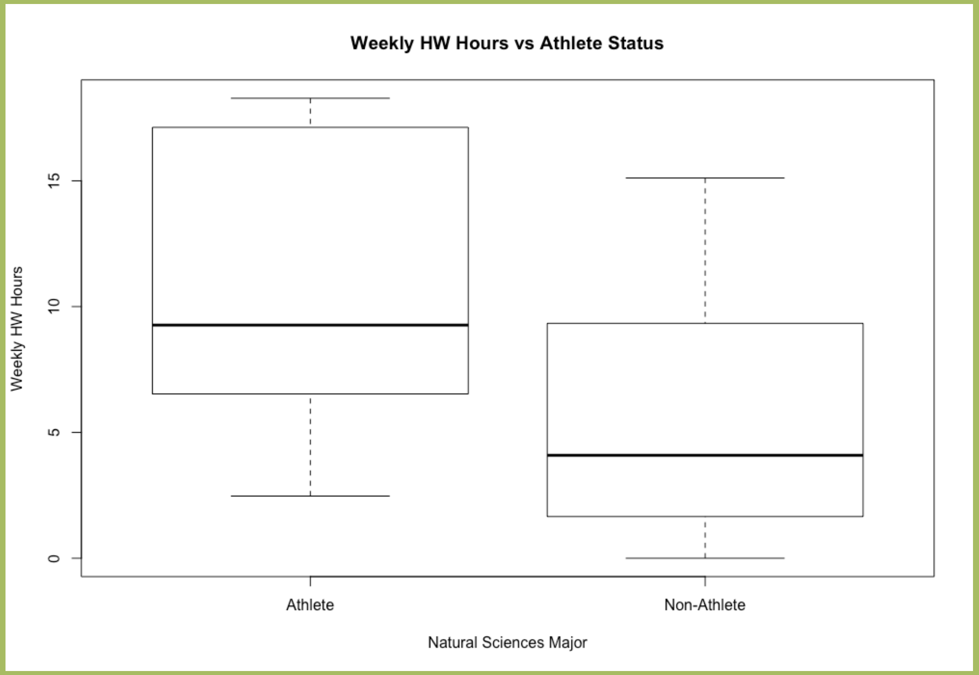
R code:

```
> social.sciences.data <-subset(Athlete_GPA_Final, Athlete_GPA_Final$Major == "Social Sciences")
> boxplot(social.sciences.data$GPA~social.sciences.data$Athlete, main = "GPA vs Athlete Status", xlab = "Social Sciences Major", ylab = "GPA")
> boxplot(social.sciences.data$Hours_Spent_On_Homework_Per_Week~social.sciences.data$Athlete, main = "Weekly HW Hours vs Athlete Status", xlab = "Social Sciences Major", ylab = "Weekly HW Hours")
```

Different Majors (Continued)

Natural Sciences

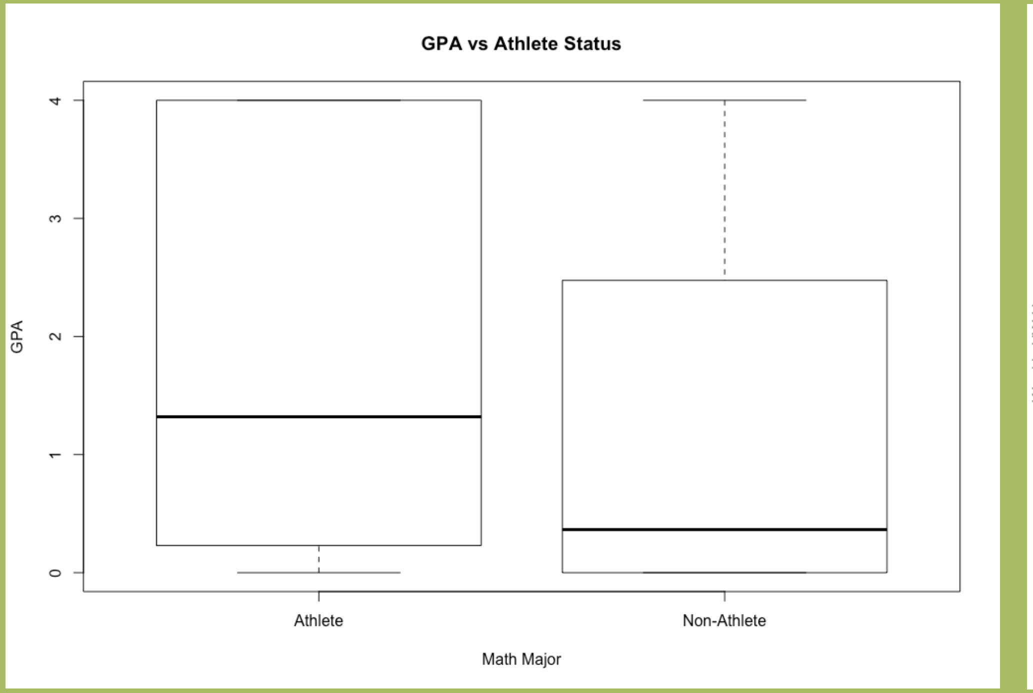


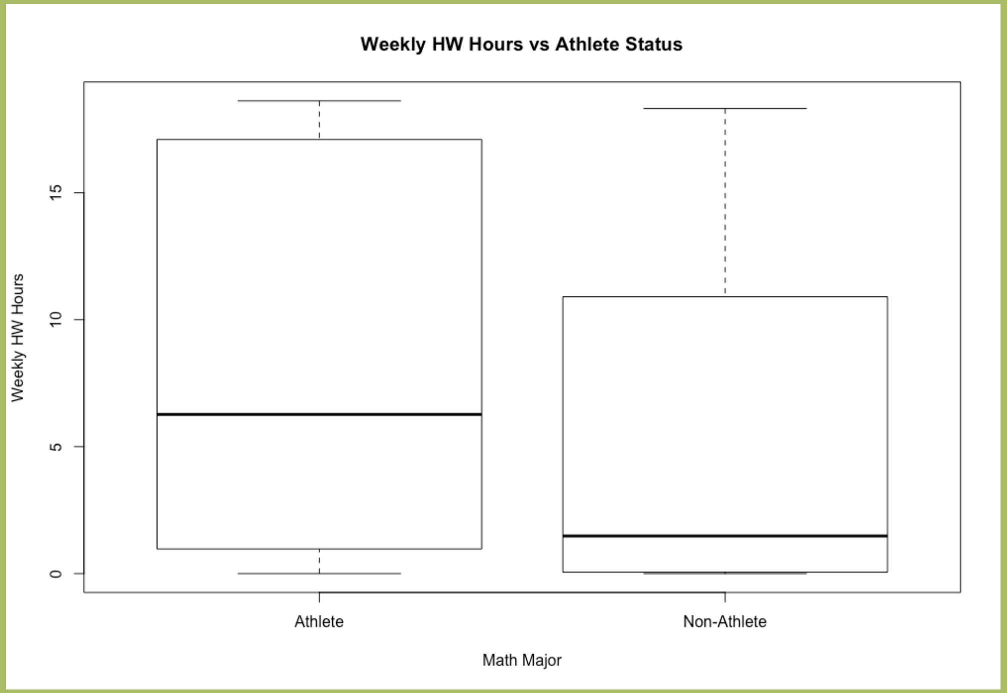


R code:

```
> natural.sciences.data <-subset(Athlete_GPA_Final, Athlete_GPA_Final$Major == "Natural Sciences")
> boxplot(natural.sciences.data$GPA~natural.sciences.data$Athlete, main = "GPA vs Athlete Status", xlab = "Natural Sciences Major", ylab = "GPA")
> boxplot(natural.sciences.data$Hours_Spent_On_Homework_Per_Week~natural.sciences.data$Athlete, main = "Weekly HW Hours vs Athlete Status", xlab = "Natural Sciences Major", ylab = "Weekly HW Hours")
```

Math



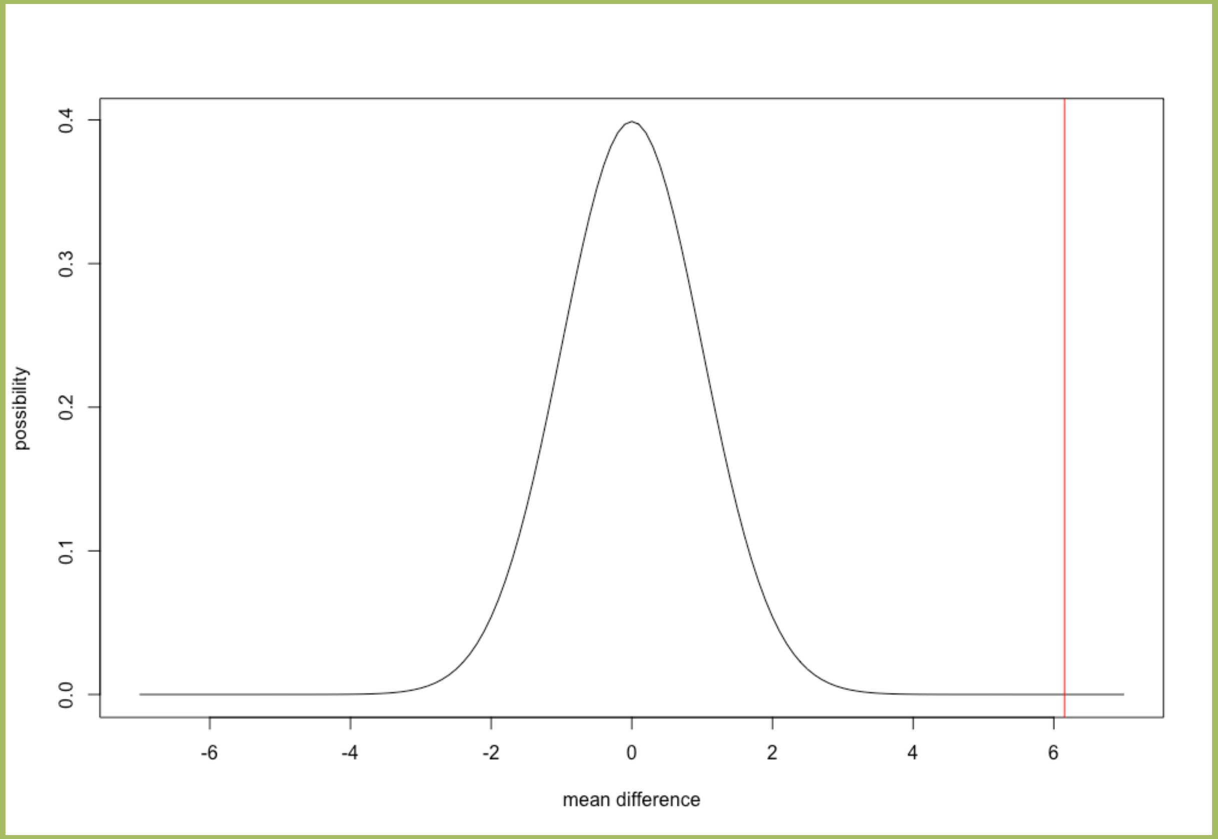


R code:

```
> math.major.data <- subset(Athlete_GPA_Final, Athlete_GPA_Final$Major == "Mathematics")
> boxplot(math.major.data$GPA~math.major.data$Athlete, main = "GPA vs Athlete Status", xlab = "Math Major", ylab = "GPA")
> boxplot(math.major.data$Hours_Spent_On_Homework_Per_Week~math.major.data$Athlete, main = "Weekly HW Hours vs Athlete Status", xlab = "Math Major", ylab = "Weekly HW Hours")
```

Z-test of mean GPAs

Now let's calculate the Z-score and P-value between average GPAs of student athletes and non-athletes



Z = 6.2515

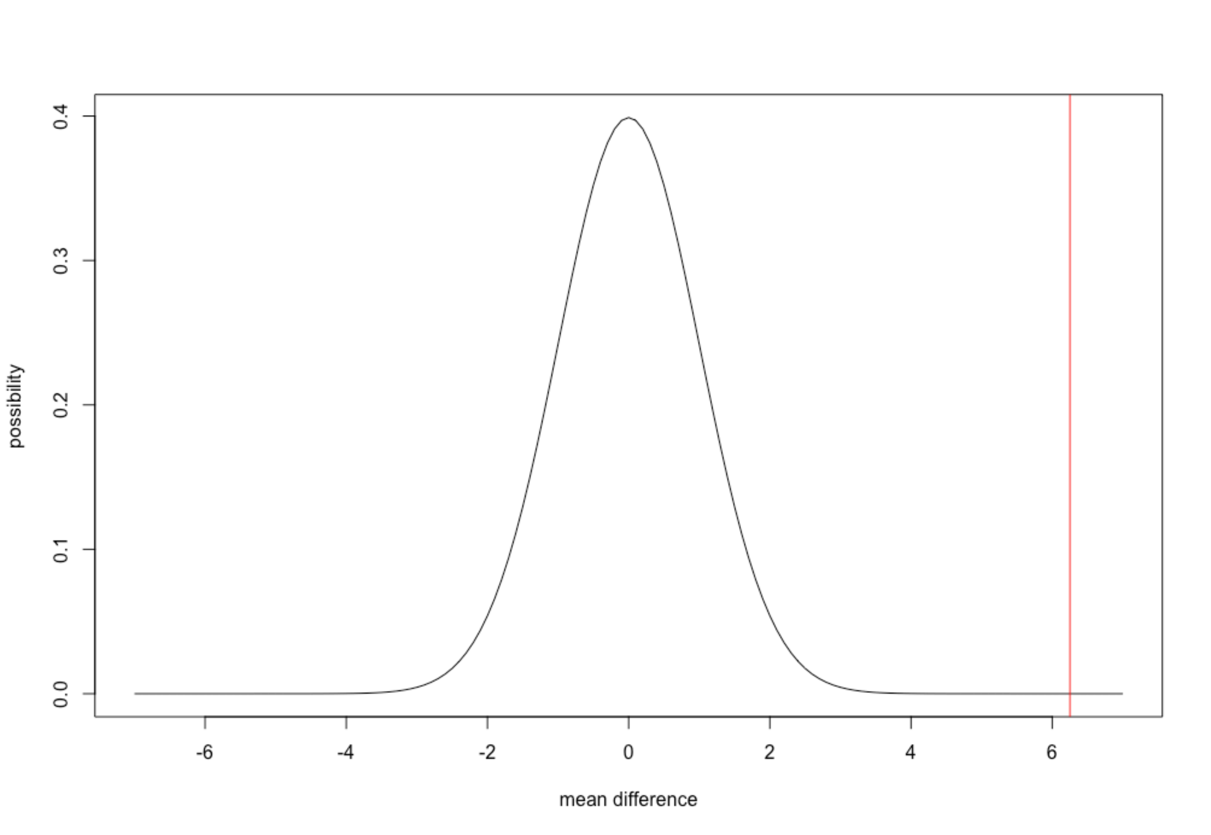
P =3.83923e-10

R code:

```
> athlete.data <- subset(Athlete_GPA_Final, Athlete_GPA_Final$Athlete == "Athlete")
> nonathlete.data <- subset(Athlete_GPA_Final, Athlete_GPA_Final$Athlete == “Non-Athlete")
> athlete.gpa <- athlete.data$GPA
> nonathlete.gpa <- nonathlete.data$GPA
> mean.athlete.gpa <- mean(athlete.gpa)
> mean.nonathlete.gpa <- mean(nonathlete.gpa)
> sd.athlete.gpa <- sd(athlete.gpa)
> sd.nonathlete.gpa <- sd(nonathlete.gpa)
> len_athlete.gpa <- length(athlete.gpa)
> len_nonathlete.gpa <- length(nonathlete.gpa)
> sd.gpa <- sqrt(sd.athlete.gpa^2/len_athlete.gpa + sd.nonathlete.gpa^2/len_nonathlete.gpa)
> zeta.gpa <- (mean.athlete.gpa - mean.nonathlete.gpa)/sd.gpa
> zeta.gpa
> plot(x=seq(from = -7, to= 7, by=0.1),y=dnorm(seq(from = -7, to= 7,  by=0.1),mean=0),type='l',xlab = 'mean difference',  ylab='possibility')
> abline(v=zeta.gpa, col=‘red’)
> p <- 1 - pnorm(zeta.gpa)
> p
```

More Z-test!

Because of the high correlation between GPA and hours spent on homework, we might as well examine the Z-score and P-value between **mean HW hours** by student athletes and non-athletes!



Z = 6.2515

P = 2.032643e-10

R code:

```
> athlete.hw <- athlete.data$Hours_Spent_On_Homework_Per_Week
> nonathlete.hw <- nonathlete.data$Hours_Spent_On_Homework_Per_Week
> mean.athlete.hw <- mean(athlete.hw)
> mean.nonathlete.hw <- mean(nonathlete.hw)
> sd.athlete.hw <- sd(athlete.hw)
> sd.nonathlete.hw <- sd(nonathlete.hw)
> len_athlete.hw <- length(athlete.hw)
> len_nonathlete.hw <- length(nonathlete.hw)
> sd.hw <- sqrt(sd.athlete.hw^2/len_athlete.hw + sd.nonathlete.hw^2/len_nonathlete.hw)
> zeta.hw <- (mean.athlete.hw - mean.nonathlete.hw)/sd.hw
> zeta.hw
> plot(x=seq(from = -7, to= 7, by=0.1),y=dnorm(seq(from = -7, to= 7,  by=0.1),mean=0),type='l',xlab = 'mean difference',  ylab='possibility')
> abline(v=zeta.hw, col='red')
> p <- 1 - pnorm(zeta.hw)
> p
```

Some schools have skewed GPAs!

Q. Should we discard statistical outliers in which the entire student body is student athlete/non-athlete?

A. No, the raw dataset covers schools from East Coast, Midwest, and West Coast and has a balanced distribution between public and private colleges.

Conclusion

a) Fig. 7 shows an obvious correlation between GPA and hours spent on HW.
b) Commuter status is nonsignificant to GPA
c) Athletes own a higher mean GPA than non-athletes across all majors in the dataset
d) Statistical significance is observed in both GPA vs athlete status (P =3.83923e-10) and hours spent on HW vs athlete status (P = 2.032643e-10)