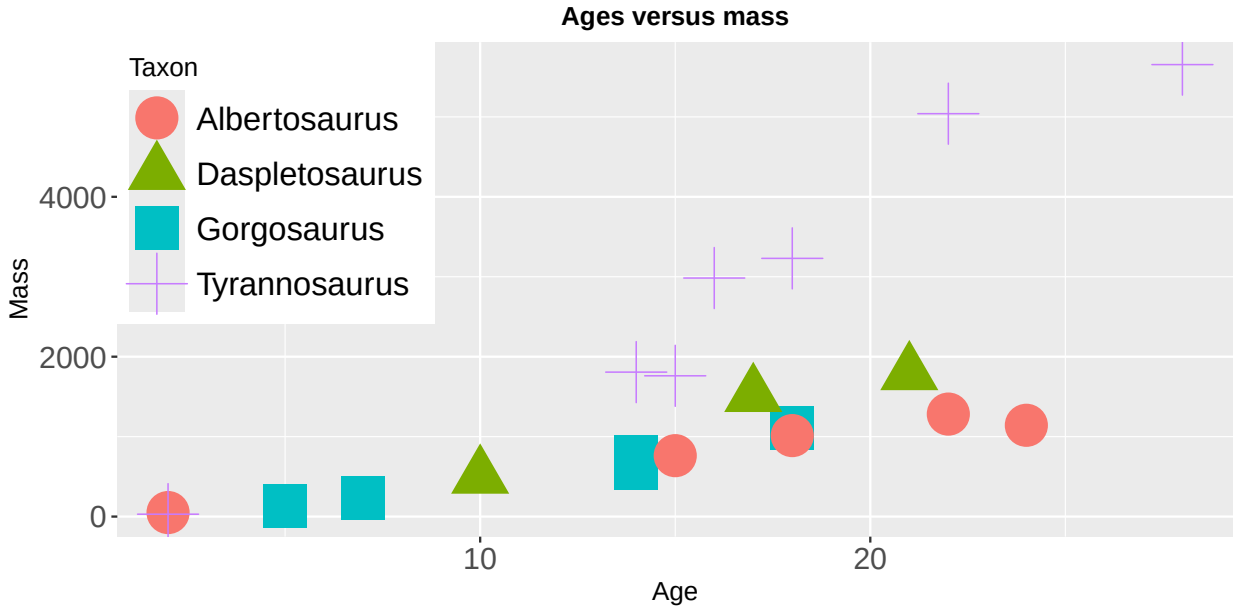
Problem 1: (a) Write a loop to compute the mean and standard deviation of the mass for each taxon.

```
data_curve <- read.csv('Growth curves data.csv')
means=numeric(4)
stds=numeric(4)
taxon_names=unique(data_curve['Taxon'][1:20,])
for (i in 1:length(taxon_names)) {
         means[i]=mean(data_curve$Mass[data_curve$Taxon==taxon_names[i]])
         stds[i]=sd(data_curve$Mass[data_curve$Taxon==taxon_names[i]])
}
tab=matrix(c(means,stds), ncol=2)
colnames(tab)=c('Means','SD')
rownames(tab)=c('Gorgosaurus','Albertosaurus','Daspletosaurus','Tyrannosaurus')
tab=as.table(tab)
tab
```

```
##                    Means        SD
## Gorgosaurus     563.0000  397.2430
## Albertosaurus   849.8600  486.2409
## Daspletosaurus 1268.3333  682.6466
## Tyrannosaurus  2929.4143 1958.0373
```

(b) Make a plot of age versus mass that includes all observations but a different plotting symbol (i.e., the pch option in plot) or color for each taxon.

```
library(ggplot2)
ggplot(data_curve, aes(x=Age, y=Mass, colour=Taxon, shape=Taxon))+
xlab("Age")+
ylab("Mass")+
ggtitle("Ages versus mass")+
geom_point(size=9, alpha=1)+
theme(axis.title.x=element_text(colour="Black", size=12),
axis.title.y=element_text(colour="Black", size=12),
axis.text.x=element_text(size=14),
axis.text.y=element_text(size=14),
legend.title=element_text(size=12),
legend.text=element_text(size=15),
legend.position=c(0,1),
legend.justification=c(0,1),
plot.title=element_text(size=12,face = "bold", hjust=.5),
)
```




```
linear_model=lm(log(Mass)~log(Age),data=data_curve)
summary(linear_model)
```

```
##
## Call:
## lm(formula = log(Mass) ~ log(Age), data = data_curve)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8279 -0.3692 -0.1350  0.4475  0.8437
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2946     0.4206   5.456 3.50e-05 ***
## log(Age)      1.7538     0.1598  10.977 2.09e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5199 on 18 degrees of freedom
## Multiple R-squared:   0.87,  Adjusted R-squared:  0.8628
## F-statistic: 120.5 on 1 and 18 DF,  p-value: 2.09e-09
```

Mass is a dependent variable, and age is a predictor. The residuals are the difference between the actual and predicted values. We want the median to be centered around zero so that residuals are somewhat symmetrical and the model predicts correctly at both the high and low ends of the dataset. Based on the residuals part of the summary, our model's prediction is quite good. Based on the coefficients part, both the intercept and log(age) are statistically significant because their p-values are sufficiently small. Our model is $y = 2.2946 + 1.7538x$, where x represents log(age). Every one unit increase in log(age) corresponds to 1.7538-unit increase in log(mass). The standard error of the coefficient is an estimate of the standard deviation of the coefficient. The standard error is used to create confidence intervals. Here, 95% confidence interval around slope is $(1.7538 \pm 1.96 \times 0.1598) = (1.44, 2.07)$. We see that RSE is 0.52. Considering mass data (the biggest value is 5654), having all our predictions be off by an average of 0.52 won't affect predictions significantly. The R-squared value shows what percentage of the variation within our dependent variable all the predictors are

explaining. It is 87% for the model, which is pretty good. F-statistic is very large, and our p-value is so small that it indicates there is a positive and statistically significant relationship between log age and log mass.