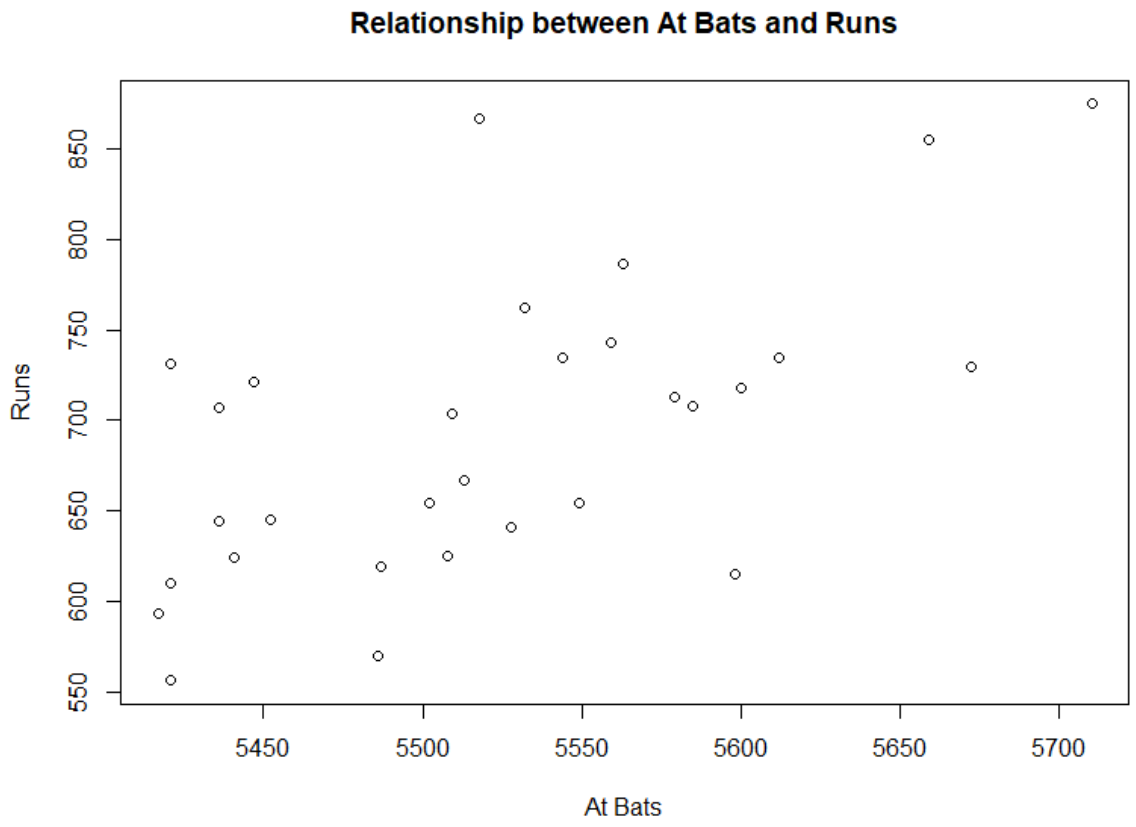
R Assignment 13

Code segments provided for each question, full R script for the assignment at the bottom

> 1: What type of plot would you use to display the relationship between runs and one of the other numerical variables? (Note that both variables are numeric!) Plot this relationship (using base R or ggplot2) using the variable at_bats as the x-variable.

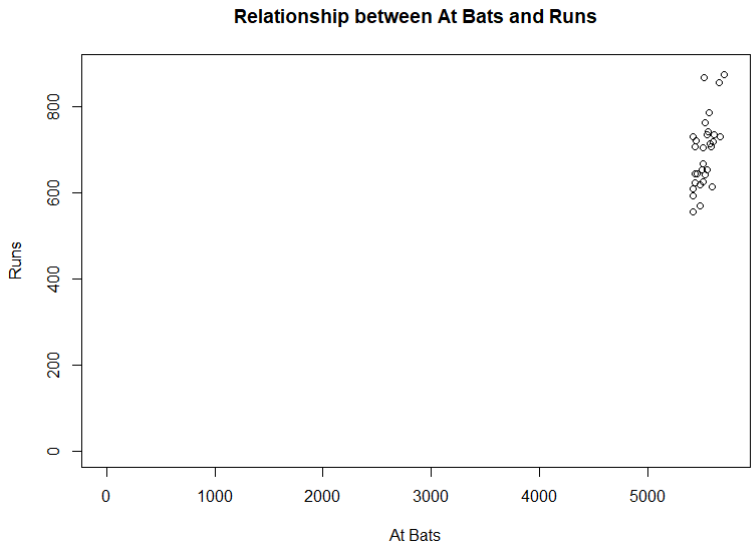
I would use a scatter plot to display the relationship between two numerical variables.

```
attach(mlb11)
plot(at_bats, runs, main = "Relationship between At Bats and Runs",
     xlab = "At Bats", ylab = "Runs")
```



> 2: Does the relationship in your plot from Question 1 look linear? If you knew a team's at_bats, would you be comfortable using a linear model to predict the number of runs?

Looking just at the "zoomed" area (axis do not start at 0), the relationship does not look linear. When the axis limit is set to 0, the relationship looks much more linear, so I would be comfortable using a linear model to predict the number of runs

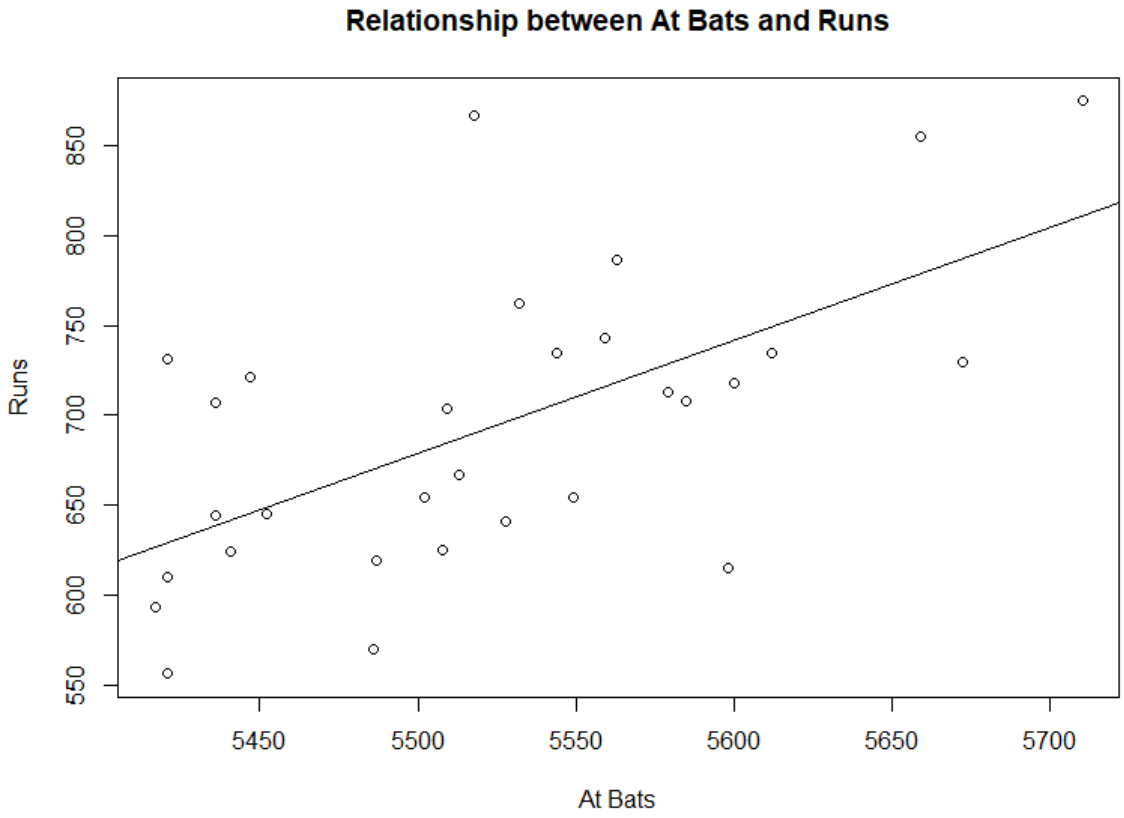


> 3: Using the cor() function, report the correlation between runs and at_bats. What does the correlation measure about the relationship between these two variables?

```
corrRunsAtBats = cor(runs, at_bats)
```

The correlation between runs and at_bats is 0.611. The correlation coefficient suggests a moderate positive relationship between `runs` and `at_bats`.

> 4: Fit a linear model to the above data, with runs as the outcome and at_bats as the predictor. Create a table in Word of the model fitting results (the point estimates, standard errors, t-statistics, and p-values only).



```
model1 <- lm(runs ~ at_bats, data = mlb11)
abline(lm(runs ~ at_bats))
summary(model1)
```

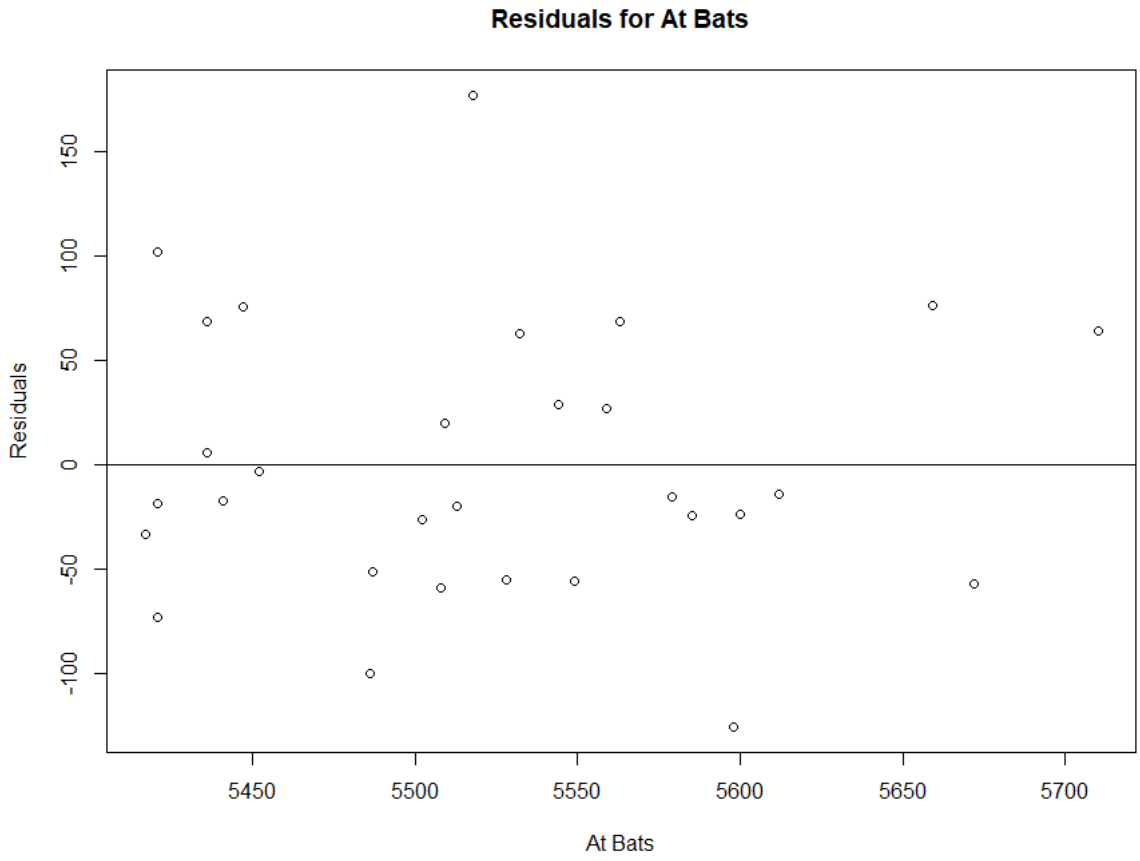
Coeff.	Point Estimate	Std. Error	t-statistics	p-value
(Intercept)	-2789.2429	853.6957	-3.267	0.002871
at_bats	0.6305	0.1545	4.080	0.000339

> 5: If a team manager saw the least squares regression line (and not the actual data), how many runs would he or she predict for a team with 5,579 at-bats? Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?

```
toPredict = 5579
predicted = predict(model1, data.frame(at_bats=c(toPredict)))
actual = mlb11$runs[mlb11$at_bats == toPredict]
residualPred = predicted - actual
```

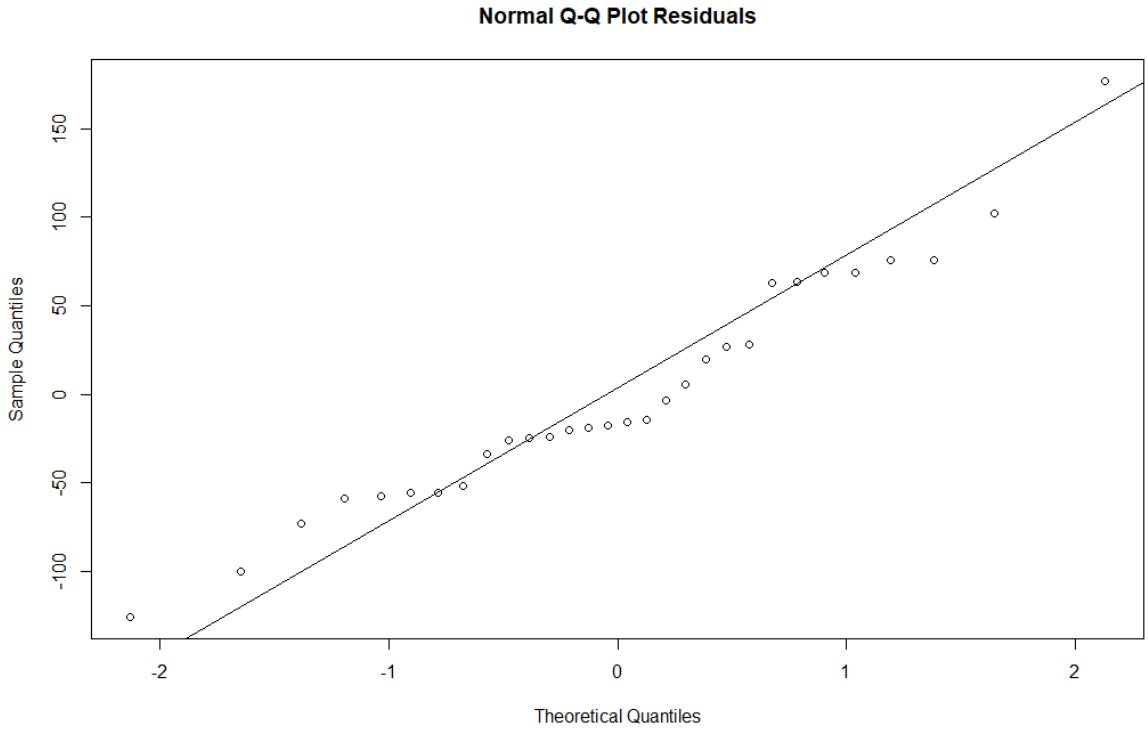
The team manager would predict 728.5955 runs, which is an overestimation by 15.59552, as the actual value is 713.

> 6: Share your plot of the residuals vs. the predictor here. Does this plot show a linear relationship between runs and at_bats?



The plot shows a linear relationship between `runs` and `at_bats` with both positive and negative outliers.

> 7: Share your normal probability plot here. Would you say the normality assumption has been met?



While the residuals do not follow a perfect normal distribution, it is close enough that you can approximate the residuals with a normal distribution, so the normality assumption is met.

Constant variability:

> 8: Based on the plot in Question 6, does the constant variability condition appear to be met?

Apart from one outlier (~5525 at bats), the constant variability condition appears to be met.

> 9: Based on your model-fitting results, does it appear that at_bats is a significant predictor of runs (using a significance level of 0.05)? Report the hypotheses for this test here, too.

H_0: The true slope of the linear regression model for at_bats and runs is 0.

H_A: The true slope of the linear regression model for at_bats and runs is not 0.

With a significance value of 0.05, we reject the null hypothesis that the true slope of the linear regression model is 0 (t-statistic of 4.080 and a p-value 0.000339) and conclude that `at_bats` is a significant predictor for runs.

> 10: Report the R2 value from your model fit. What does this value tell you about your model?

Multiple R-squared: 0.3729
Adjusted R-squared: 0.3505

The R^2 value of 0.3729 means, that 37.29 percent of the response variable (runs) is explained by the model-

Complete Script for R Assignment 13

```
load("C:/Users/Caro/Documents/R_631/mlb11.RData")
View(mlb11)

attach(mlb11)
plot(at_bats, runs, main = "Relationship between At Bats and Runs",
     xlab = "At Bats", ylab = "Runs")
# plot(at_bats, runs, main = "Relationship between At Bats and Runs", xlab = "At
Bats", ylab = "Runs", xlim = c(0, max(at_bats) + 10), ylim = c(0,  max(runs) +
10))

corrRunsAtBats = cor(runs, at_bats)

# fit linear model
model1 <- lm(runs ~ at_bats, data = mlb11)
abline(lm(runs ~ at_bats))
summary(model1)

# predict  new data
toPredict = 5579
predicted = predict(model1, data.frame(at_bats=c(toPredict)))
```

```r
actual = mlb11$runs[mlb11$at_bats == toPredict]
residualPred = predicted - actual

# plotting residuals
residualsModel = resid(model1)
plot(at_bats, residualsModel, ylab = "Residuals", xlab = "At Bats",
     main = "Residuals for At Bats")
abline(0, 0)

# normal probability
qqnorm(residualsModel, main = "Normal Q-Q Plot Residuals")
qqline(residualsModel)
```