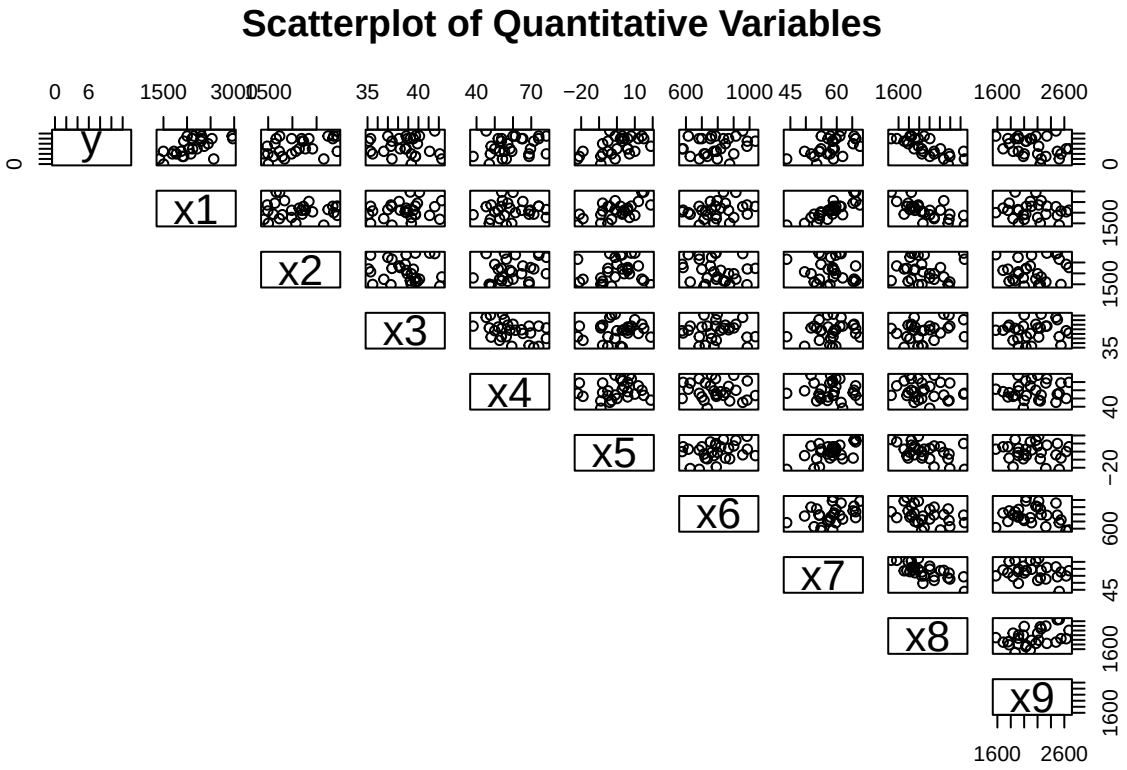
Guided Question Set 6 Solutions

1)

```
library(tidyverse)
Data<-read.table("nfl.txt", header=TRUE)
```

The scatterplot matrix is displayed below

```
pairs(Data, lower.panel = NULL, main="Scatterplot of Quantitative Variables")
```



The pairwise correlations are shown below

```
round(cor(Data),3)
```

```
##          y     x1     x2     x3     x4     x5     x6     x7     x8     x9
## y    1.000  0.593  0.483 -0.081  0.258  0.513  0.224  0.545 -0.738 -0.304
```

```
## x1  0.593  1.000 -0.037  0.212  0.070  0.600  0.253  0.837 -0.659 -0.111
## x2  0.483 -0.037  1.000 -0.069  0.302  0.135 -0.193 -0.197 -0.051  0.146
## x3 -0.081  0.212 -0.069  1.000 -0.413  0.115 -0.003  0.163  0.290  0.088
## x4  0.258  0.070  0.302 -0.413  1.000  0.149 -0.128 -0.101 -0.164  0.059
## x5  0.513  0.600  0.135  0.115  0.149  1.000  0.259  0.610 -0.470 -0.090
## x6  0.224  0.253 -0.193 -0.003 -0.128  0.259  1.000  0.367 -0.352 -0.173
## x7  0.545  0.837 -0.197  0.163 -0.101  0.610  0.367  1.000 -0.685 -0.203
## x8 -0.738 -0.659 -0.051  0.290 -0.164 -0.470 -0.352 -0.685  1.000  0.417
## x9 -0.304 -0.111  0.146  0.088  0.059 -0.090 -0.173 -0.203  0.417  1.000
```

a)

We note that predictors x_1, x_2, x_5, x_7, x_8 have moderate to high correlations with the number of wins. These predictors are rushing yards, passing yards, turnover differential, percent of plays that are rushes, and the opponent's rushing yards for the season. The last predictor is the only one that is negatively associated with the number of wins.

The predictors x_3, x_4, x_6, x_9 do not have a strong linear relationship with number of wins. These predictors are punting average, field goal percentage, penalty yards, and opponents' passing yards for the season.

b)

Notice that x_1, x_5, x_7, x_8 have moderately high correlations with each other. We noted earlier that these predictors are have some correlation with the number of wins.

c)

There are a couple of ways to frame an answer here:

- I would consider using x_1, x_2, x_5, x_7, x_8 in a MLR as these predictors have high correlation with the number of wins.
- I would consider using x_2 as it has a high correlation with the number of wins, and a subset of x_1, x_5, x_7, x_8 as these are correlated with the number of wins but are also correlated among themselves.

2)

```
result<-lm(y~x2+x7+x8, data=Data)
summary(result)
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = Data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0370 -0.7129 -0.2043  1.1101  3.7049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.808372   7.900859  -0.229 0.820899
## x2           0.003598   0.000695   5.177 2.66e-05 ***
## x7           0.193960   0.088233   2.198 0.037815 *
## x8          -0.004816   0.001277  -3.771 0.000938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on 24 degrees of freedom
## Multiple R-squared:  0.7863, Adjusted R-squared:  0.7596
## F-statistic: 29.44 on 3 and 24 DF,  p-value: 3.273e-08
```

The regression equation is $\hat{y} = -1.8084 + 0.0036x_2 + 0.1940x_7 - 0.0048x_8$.

3)

The predicted number of wins increases by 0.1940 for a percentage point increase in the percent of plays that are runs, when the passing yards and number of yards given up to opponents are held constant.

4)

```
newdata<-data.frame(x2=2000,x7=48,x8=2350)
predict(result,newdata,interval="prediction")
```

```
##        fit        lwr      upr
## 1 3.381448 -0.5163727 7.279268
```

The predicted number of wins for such a team is 3.3814. The 95% prediction interval is (-0.5164, 7.2793).

5)

$H_0 : \beta_2 = \beta_7 = \beta_8 = 0$, H_a : at least one of the coefficients in H_0 is not zero.

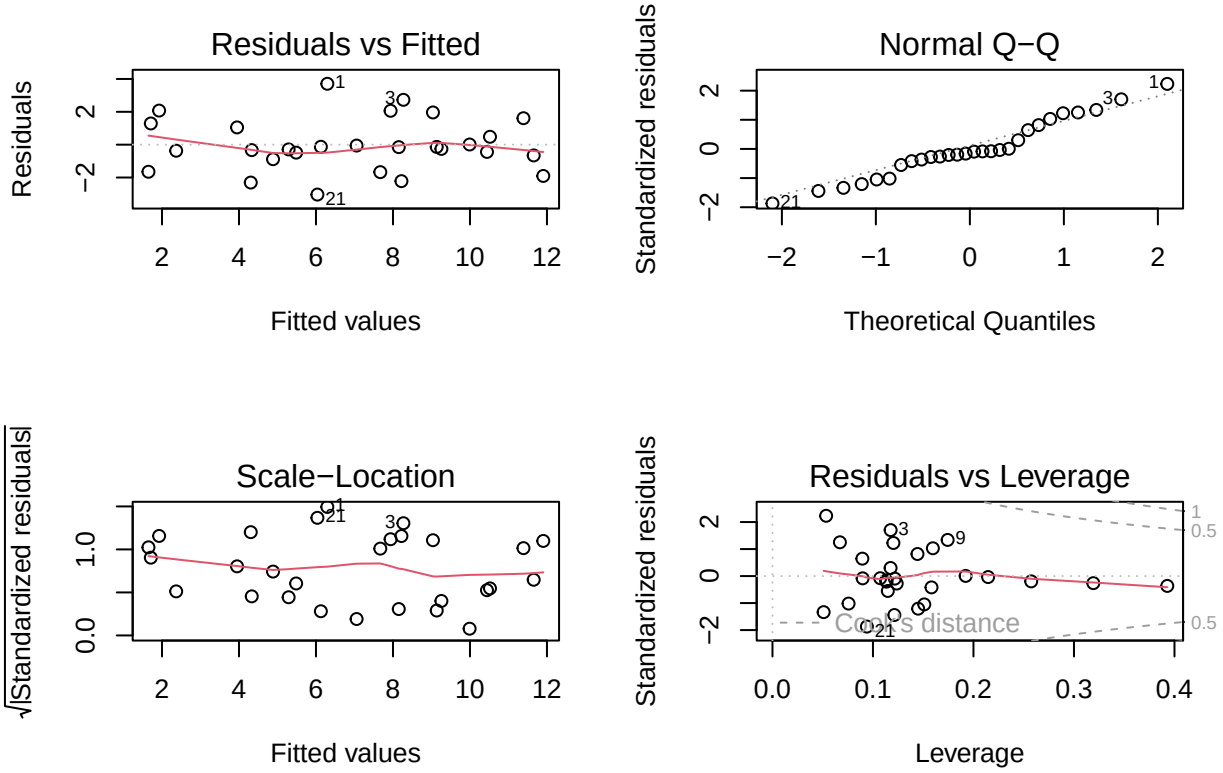
The ANOVA F statistic is 29.44 with p-value close to 0. The critical value is `qf(0.95,3,24)` which is 3.0088. Since the p-value is less than 0.05 (or since the F statistic is greater than 3.0088), we reject the null hypothesis. The data support the claim that our model with the three predictors is useful in predicting the number of wins.

6)

The t statistic is 2.198. Since the p-value is less than 0.05, we reject the null hypothesis. The predictor percent of plays that are rushes is useful in predicting the number of wins, when we already have pass yards and opponent's rush yards already in the model. The critical value is `qt(0.975,24)` which is 2.0639.

7)

```
par(mfrow=c(2,2))
plot(result)
```



The regression assumptions appear to be met. From the residual plot, we note the residuals are evenly scattered around 0 at random, with a constant vertical variance. From the QQ plot, the normality assumption is reasonably met as the residuals fall close to their theoretical values under normality.

8)

```
result2<-lm(y~x2+x7+x8+x1, data=Data)
summary(result2)
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8 + x1, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7456 -0.6801 -0.1941  1.1033  3.7580
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.8791718  8.1955007  -0.107  0.91550
## x2           0.0035214  0.0007191   4.897 6.02e-05 ***
## x7           0.1437590  0.1280424   1.123  0.27313
## x8          -0.0046994  0.0013131  -3.579  0.00159 **
## x1           0.0009045  0.0016489   0.549  0.58862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.732 on 23 degrees of freedom
## Multiple R-squared:  0.7891, Adjusted R-squared:  0.7524
## F-statistic: 21.51 on 4 and 23 DF,  p-value: 1.702e-07
```

The t test for the coefficient of x_1 is insignificant. We can remove this predictor and leave the others in the model. OR This predictor is insignificant in the presence of the other predictors. Disagree with the classmate.

To address the classmate's statement, we need to fit a simple linear regression with x_1, the team's rushing yards, as the only predictor. The MLR model is not meant to address the classmate's statement.

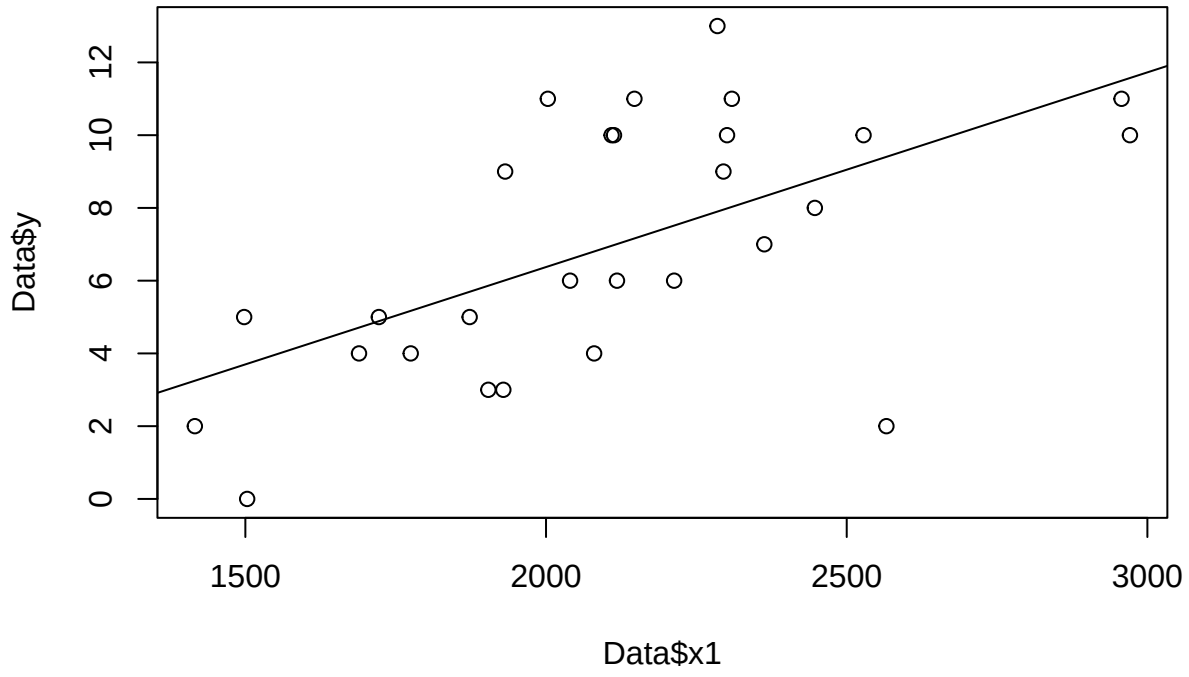
If you tried to fit a SLR with x_1 as the only predictor and create a scatterplot of the number of wins with x_1, you will see that x_1 is a significant predictor and is linearly related to the number of wins.

```
result3<-lm(y~x1, data=Data)
summary(result3)
```

```
##
## Call:
## lm(formula = y ~ x1, data = Data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -7.4037 -1.3665 -0.6416  2.2539  5.1002
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.330015   3.053809  -1.418 0.168093
## x1           0.005352   0.001424   3.758 0.000877 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.855 on 26 degrees of freedom
## Multiple R-squared:  0.3519, Adjusted R-squared:  0.327
## F-statistic: 14.12 on 1 and 26 DF,  p-value: 0.000877
```

```
plot(Data$y~Data$x1)
abline(result3)
```




```
cor(cbind(Data$x1,Data$x2,Data$x7,Data$x8))
```

```
##             [,1]        [,2]       [,3]        [,4]
## [1,]  1.00000000 -0.03674736  0.8372827 -0.65854627
## [2,] -0.03674736  1.00000000 -0.1969154 -0.05104783
## [3,]  0.83728269 -0.19691540  1.0000000 -0.68504573
```

```
## [4,] -0.65854627 -0.05104783 -0.6850457  1.00000000
```

Notice on its own x_1 is linearly related to the response. It is not needed in a model with x_2, x_7, x_8 as it doesn't improve the predictions significantly as it is highly correlated with a number of other predictors. So x_1 doesn't provide much additional insight to the prediction when the others are already in the model.