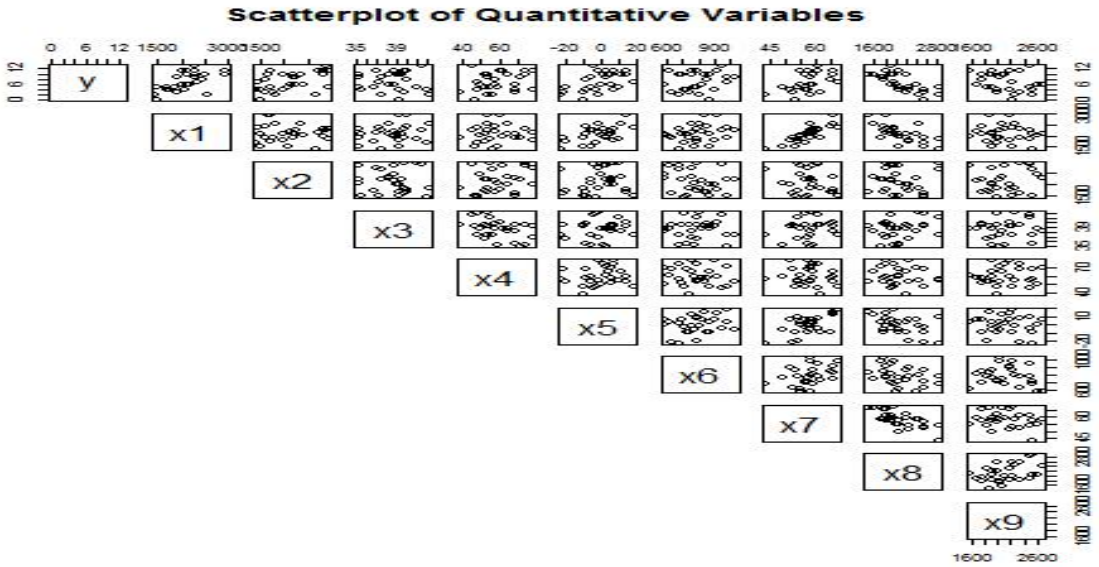
Stat 6021: Guided Question Set 4 Solutions

1. (a) The scatterplot matrix is displayed below:



The correlation matrix is shown below:

	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

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) 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.

2. $\hat{wins} = -1.8084 + 0.0036x_2 + 0.1940x_7 - 0.0048x_8$.

```
Call:
lm(formula = y ~ x2 + x7 + x8)

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
```

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

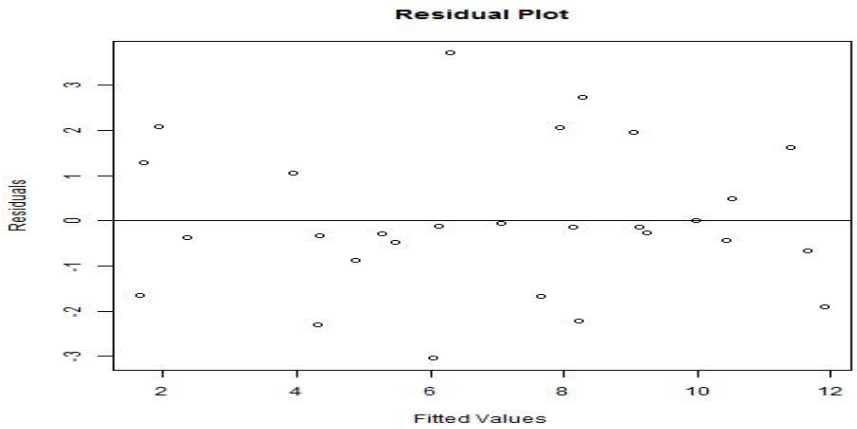
```
> newdata<-data.frame(x2=2000,x7=48,x8=2350)
> predict.lm(result,newdata,interval="prediction")
       fit        lwr      upr
1 3.381448 -0.5163727 7.279268
```

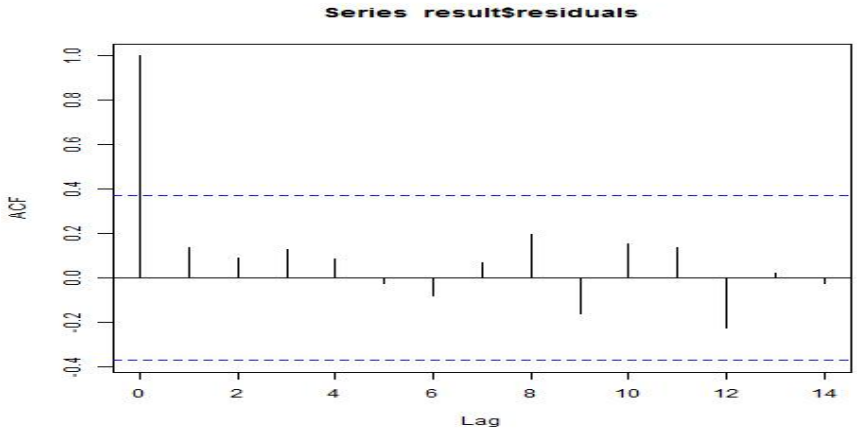
5. $H_0 : \beta_2 = \beta_7 = \beta_8 = 0, H_a$: at least one of the coefficients in H_0 is not zero.

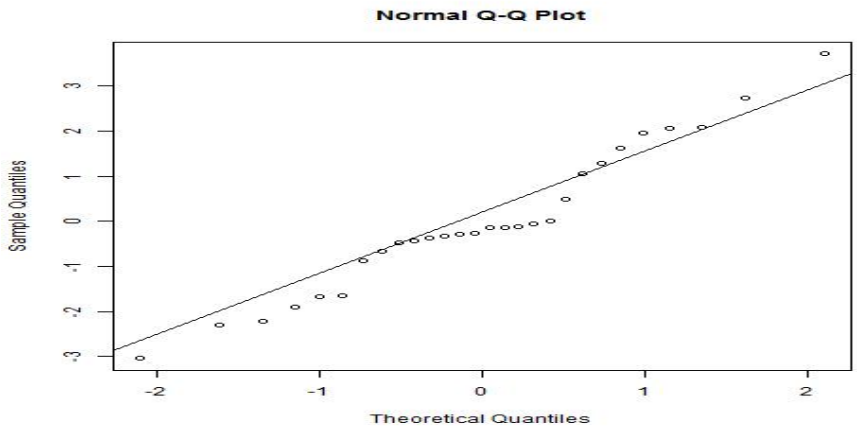
The ANOVA F statistic is 29.44 with p-value close to 0. The critical value is `qf(0.95,3,24)` which is 3.0088. Since the p-value is less than 0.05 (or since the F statistic is greater than 3.0088), we reject the null hypothesis. The data supports the claim that our model with the three predictors is 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)` 2.0639.

7. 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 variabce. The ACF plot shows the residuals are uncorrelated. From the QQ plot, the normality assumption is reasonably met as the residuals fall close to their theoretical values under normality.







8. 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.