CIS 3850 Final Project Analysis Thomas Harrington

I have tried to develop a while or for loop that will loop through the entire table and run a correlation on each column but I cannot get it to work. As a result I am brute forcing my way through the file to find the factors that seem to correlate to G3/. This is probably one of the least efficient ways to do this but I cannot get the loops to do what I want.

For the Student-Mat data:

I selected failures, desire for higher education = yes, going out with friends, reason school was chosen = reputation and age. These all had the lowest p-values and R seemed to think they have the greatest significance.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
failures1	-3.2374	0.6861	-4.719	3.36e-06	***
failures2	-5.0589	1.1298	-4.478	1.00e-05	***
failures3	-5.2193	1.1492	-4.542	7.53e-06	***
higheryes	3.51565	1.03081	3.411	0.000717	***
goout	-0.69676	0.23048	-3.023	0.002671	**
reasonreputation	1.2482	0.5722	2.181	0.0298	*
age	-0.477384	0.194650	-2.453	0.0146	*

This is somewhat disturbing or I am too new to statistical analysis to know how common this is. I can only find one positive significant correlation between factors and G3. I found multiple significant negative correlations.

Residuals:

Min	1Q	Median	3Q	Max
-11.1709	-1.9012	0.3386	2.7250	10.0119

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	12.2277	3.1590	3.871	0.000128	***
failures1	-2.8071	0.6708	-4.185	3.54e-05	***
failures2	-4.4885	1.0730	-4.183	3.57e-05	***
failures3	-4.9769	1.1167	-4.457	1.10e-05	***
internetyes	0.9587	0.5831	1.644	0.100987	
romanticyes	-0.9285	0.4659	-1.993	0.046961	*
famrel	0.1302	0.2423	0.537	0.591380	
freetime	0.3091	0.2269	1.362	0.173976	
goout	-0.4567	0.2023	-2.258	0.024516	*
reasonhome	0.5537	0.5405	1.025	0.306246	
reasonother	1.3274	0.7906	1.679	0.093988	.
reasonreputation	1.1237	0.5463	2.057	0.040385	*
age	-0.1328	0.1792	-0.741	0.459101	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.215 on 382 degrees of freedom
Multiple R-squared: 0.1795,	Adjusted R-squared: 0.1538
F-statistic: 6.966 on 12 and 382 DF, p-value: 1.861e-11

For this group the group residuals look good, the group p-value looks good but the r squared value is far too low. This group does not account for enough variance to be reliable.

For the final test I kept selected failures, romantic interest, going out with friends and the reason school was chosen.
The only positive correlation that is significant that I can find is the Reason the School was chosen and the choice was “Reputation”. The remainder of the correlations were negative. The information seems to predict students who are not going to perform better than students who will perform.
Residuals:
    Min   1Q Median   3Q   Max
-11.7557 -1.9495   0.3768   2.7289   9.8677

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   12.0667   0.7167 16.837 < 2e-16 ***
failures1   -3.0261   0.6471 -4.676 4.05e-06 ***
failures2   -4.6310   1.0681 -4.336 1.86e-05 ***
failures3   -5.0786   1.1011 -4.612 5.42e-06 ***
romanticyes   -0.9288   0.4586 -2.025   0.0435 *
goout   -0.3609   0.1934 -1.866   0.0629 .
reasonhome   0.5353   0.5375   0.996   0.3200
reasonother   1.3179   0.7926   1.663   0.0972 .
reasonreputation 1.1326   0.5456   2.076   0.0386 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.227 on 386 degrees of freedom
Multiple R-squared: 0.166,	Adjusted R-squared: 0.1487
F-statistic: 9.603 on 8 and 386 DF, p-value: 3.741e-12

For the Student-Por data:

Interestingly there are positive correlations that are significant with this data set. I have chosen school, sex, study time, failures, higher and internet due to their low p values.

Residuals:
    Min   1Q Median   3Q   Max

```
-12.1232  -1.3940  -0.0873   1.6503   7.5164

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.5778     0.4631  22.840  < 2e-16 ***
schoolMS     -1.2264     0.2370  -5.174 3.07e-07 ***
sexM         -0.6396     0.2266  -2.823 0.004904 **
studytime2    0.3963     0.2560   1.548 0.122117
studytime3    1.1641     0.3509   3.317 0.000961 ***
studytime4    1.0407     0.5022   2.072 0.038646 *
failures1    -2.9583     0.3587  -8.248 9.23e-16 ***
failures2    -2.6751     0.7053  -3.793 0.000163 ***
failures3    -3.0686     0.7499  -4.092 4.83e-05 ***
higheryes     1.8161     0.3730   4.869 1.42e-06 ***
internetyes   0.5594     0.2614   2.140 0.032714 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.708 on 638 degrees of freedom
Multiple R-squared:  0.308,        Adjusted R-squared:  0.2972
F-statistic:  28.4 on 10 and 638 DF,  p-value: < 2.2e-16
```

The residuals are not quite as tight as I would like. The min and max are not as even as I would like. The min seems to represent the fact that there is more negative correlation than positive in the model. The group p-value is good and the group R squared value accounts for 30% of variance, ok but not strong. I will remove the lowest correlations: Internet and sex and rerun the analysis.

```
Residuals:
    Min      1Q  Median      3Q     Max
-11.7465  -1.5538  -0.0149   1.6498   7.2535

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.6349     0.3925  27.094  < 2e-16 ***
schoolMS     -1.2684     0.2309  -5.494 5.67e-08 ***
studytime2    0.5449     0.2519   2.163   0.0309 *
studytime3    1.4125     0.3432   4.116 4.37e-05 ***
studytime4    1.1486     0.5040   2.279   0.0230 *
failures1    -2.9162     0.3612  -8.074 3.38e-15 ***
failures2    -2.8961     0.7076  -4.093 4.80e-05 ***
failures3    -3.2631     0.7534  -4.331 1.72e-05 ***
higheryes     1.8351     0.3758   4.884 1.32e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 2.729 on 640 degrees of freedom
Multiple R-squared: 0.2952, Adjusted R-squared: 0.2864
F-statistic: 33.5 on 8 and 640 DF, p-value: < 2.2e-16

The residuals still lean towards the negative. Out group p-value is still good but our R squared actually lowed by a percentage point or two.

This model shows some positive correlations: School Attended, Desire for Higher Education and Study Time. The model also shows some negative correlations: Number of Past Class Failures

---

Common correlations between Student-Mat and Student-Por data sets.

The correlation the two data sets had in common was a negative correlation between G3 and Classes failed. The higher the number of classes previously failed the greater the likelihood that the G3 grade will be low. The models did not share significant positive correlations.

The number of rows are different because some of the students have their rows concatenated with the matching row in the 2nd file.

The "by=" function is used to merge based on columns. The function will check to see that the columns in each dataset are present in the other and combine them. I do not know what happens if there are differing columns. I would imagine either the different column in added or an error is thrown.

I can get the files to merge but I cannot get R to do anything but summarize them. I keep getting object cannot be found errors for ds.merge2.obj.

The most consistent strongest predictor for both Math and Portuguese was "Previously Failed Courses". My recommendation would be to identify failing students as quickly as possible and provide them with the resources to catch their peers.