Adam Gincel

Professor Li

MA 331

PCB Study Report

Polychlorinated biphenyls, or PCBs, are collections of synthetic compounds, also known as congeners. They are very toxic to young children and the environment. Despite no longer being produced in the United States, they are still present in many places, particularly marine wildlife. As such, the EPA measures levels of PCBs in fish as a means of determining PCB presence. There are many different PCBs present in fish, so to compare them all more simply the Toxic Equivalent Score, or TEQ, was invented. Estimating TEQ efficiently based on a few measurements greatly speeds along the process of measuring how safe, or unsafe, fish may be to consume. Doing this successfully would also greatly reduce costs.

Graphs and numerical summaries described the variables and compared them to one another. After that, a multiple regression was used to see how many measurements were needed to accurately predict the total PCB content of a given fish. Outliers were later accounted for, making the measurements even more accurate. Later in the assignment the log of each measurement was used to further improve predictive accuracy using different datasets, which decreased standard error even further.

Using PCBs 52, 118, 138, and 180 and removing two outliers, I got an R-squared value of 0.9941, denoting a very highly accurate multiple regression prediction. Later on I was asked to use the TEQPCB, TEQDIOXIN, and TEQFURAN rows to predict total TEQ content, which resulted in an R-squared value of 1, meaning these three rows were able to perfectly accurately

predict the Toxic Equivalent Scores of marine wildlife. Through this and other values, we were able to successfully predict information with less information than given, which could potentially allow for accurate prediction with less comprehensive measurement, saving both time and money in the short- and long-run.

Adam Gincel

MA331

17 December 2016

I pledge my honor that I have abided by the Stevens Honor System

Statistics Final Project 11.42 – 11.52

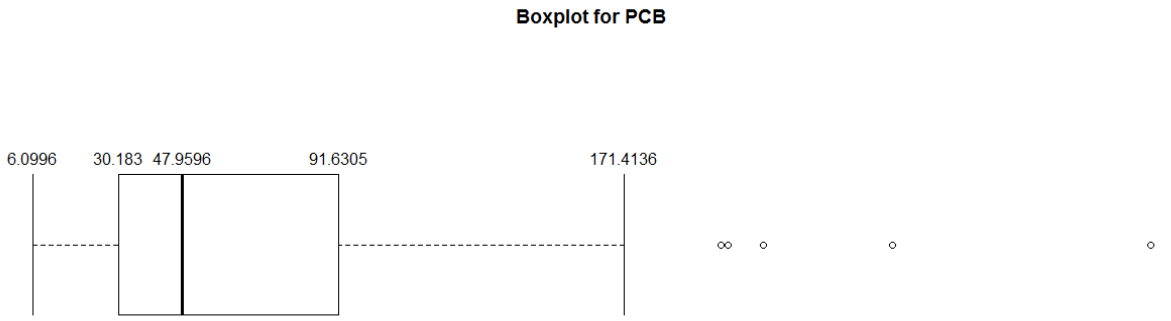
Chapter 11:

42. Consider the following variables: PCB (the total amount of PCB) and four congeners: PCB52, PCB118, PCB138, and PCB180.

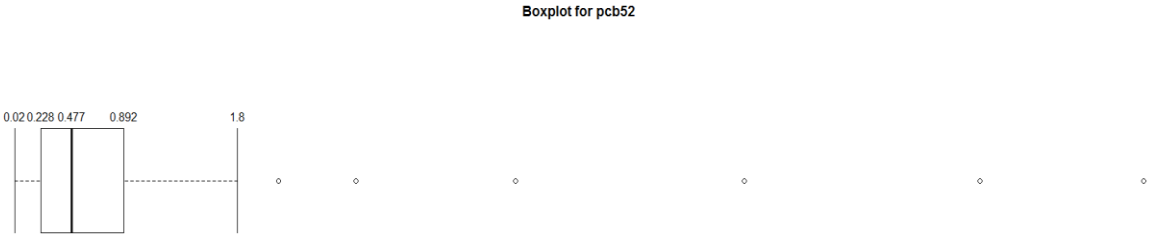
1. Using numerical and graphical summaries, describe the distribution of each of these variables.

	PCB	**PCB52**	**PCB118**	**PCB138**	**PCB180**
Min	6.0996	0.02	0.236	.64	0.395
Q1	30.138	0.228	1.49	3.18	1.24
Median	47.9596	0.477	2.42	4.92	2.69
Q3	91.6305	0.892	3.89	8.65	4.49
Maximum (not outlier)	171.4136	1.8	6.9	13.1	9.33
Mean	68.46737	0.9580435	3.256319	6.826812	4.15842
Standard Deviation	59.3906	1.59828	3.019118	5.862651	4.986439
Variance	3527.243	2.5545	9.115071	34.37068	24.86458

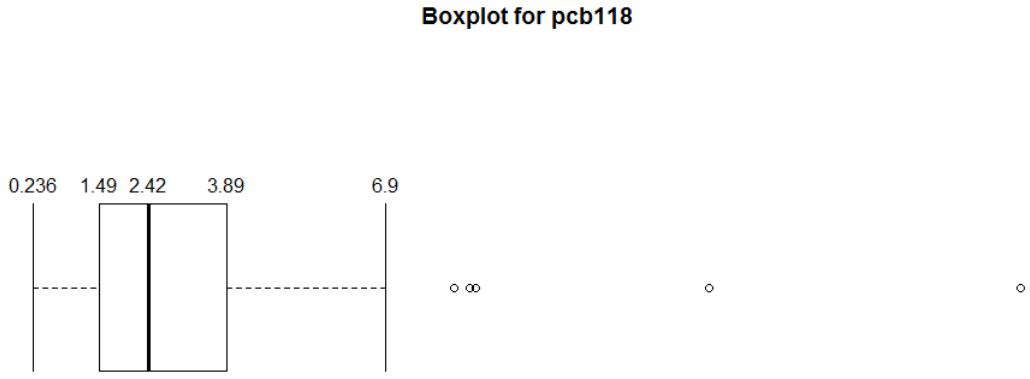
Boxplots:



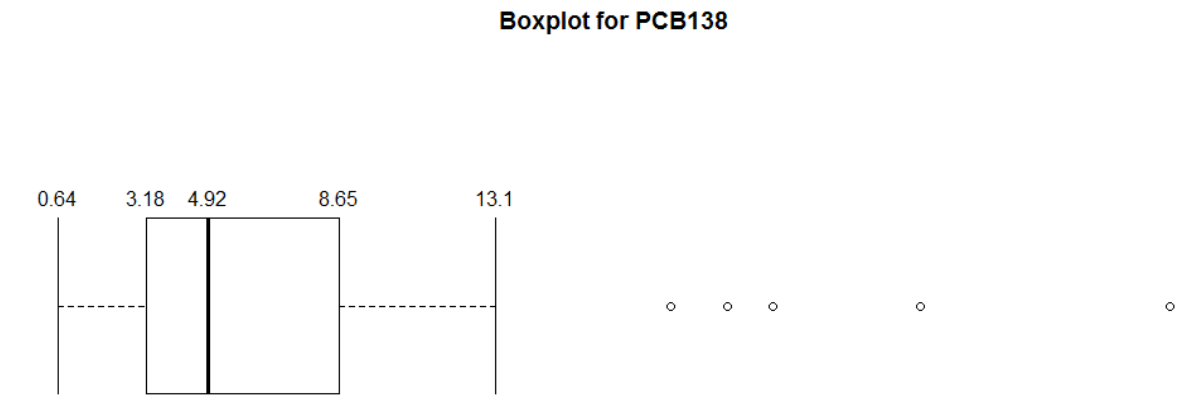
PCB is right skewed. The variance for the dataset is 3527, which is enormous.



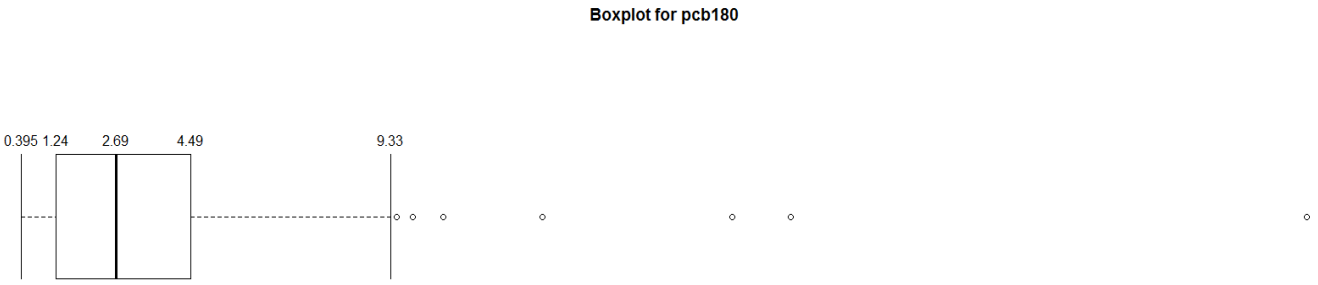
PCB52 is incredibly right skewed, with many spaced out outliers. This suggests a cluster of small numbers with a few very large values throughout.



PCB118 is right skewed but fairly normal in its distribution.



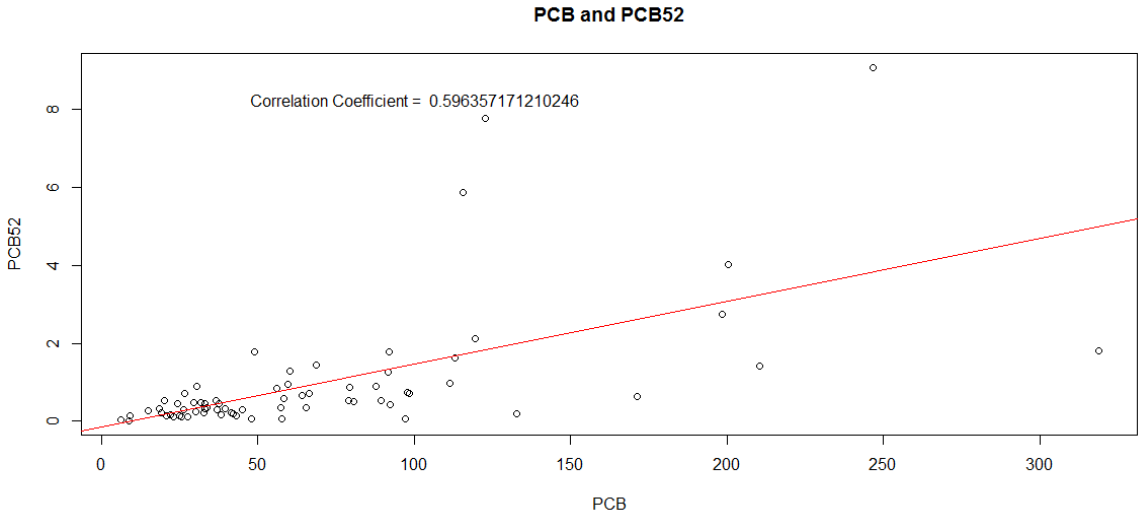
PCB138 is right skewed as well, with outliers that are not very severe.

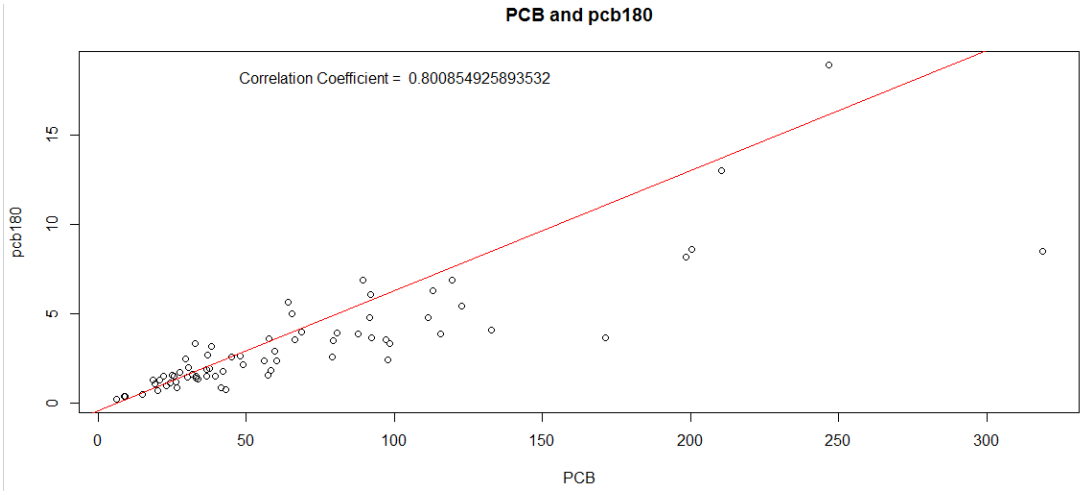


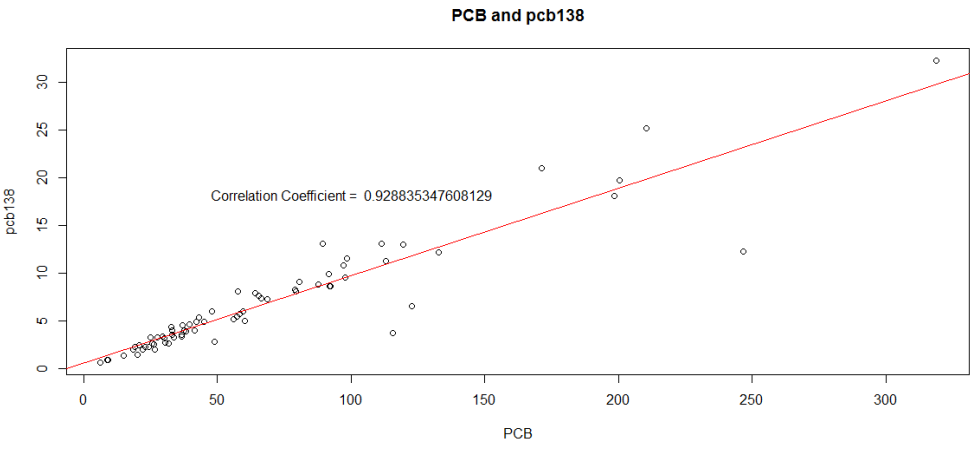
PCB180 is very right skewed, with one particularly large outlier.

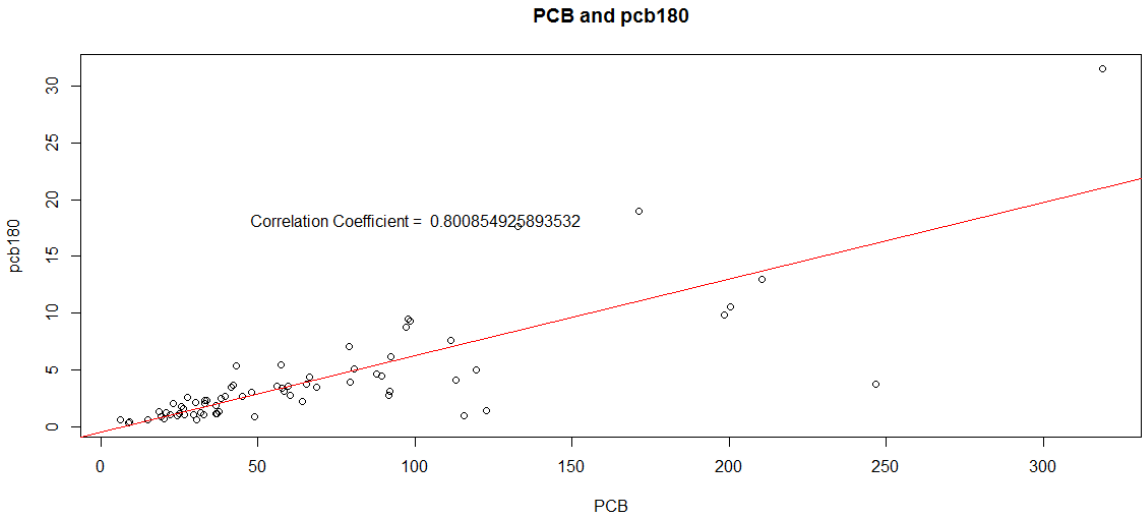
2. Using numerical and graphical summaries, describe the relationship between each pair of variables.

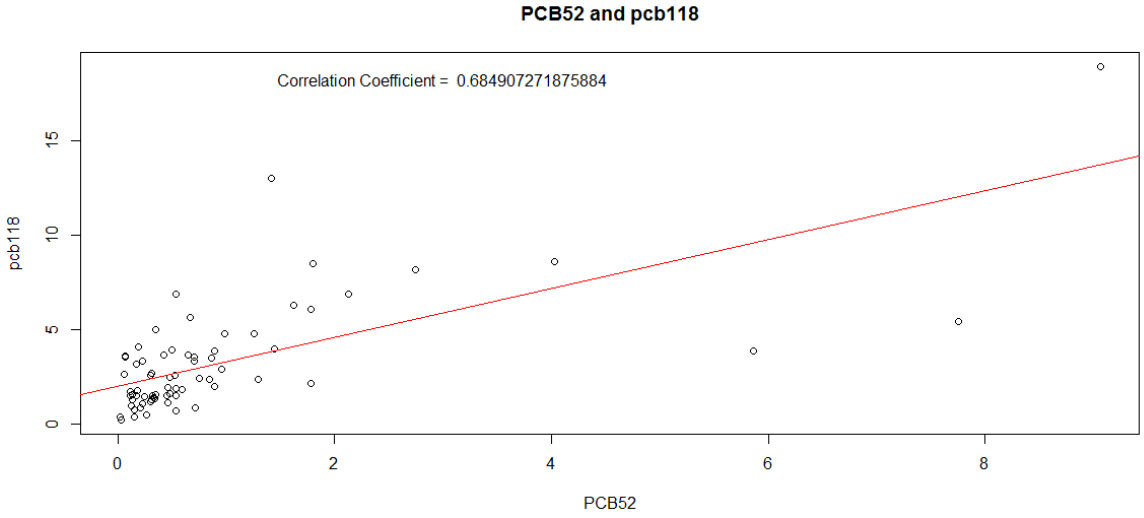
Comparative Scatterplots:

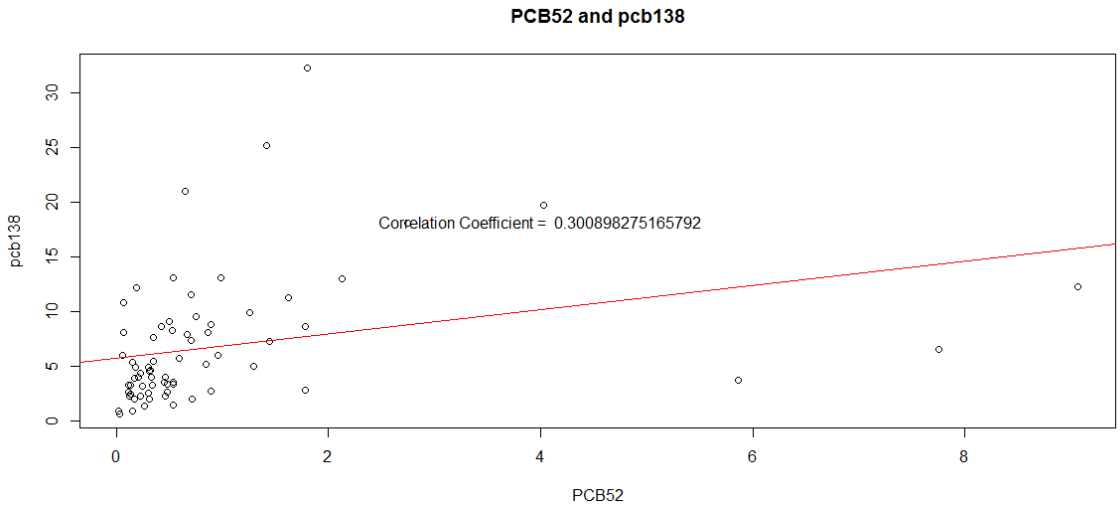


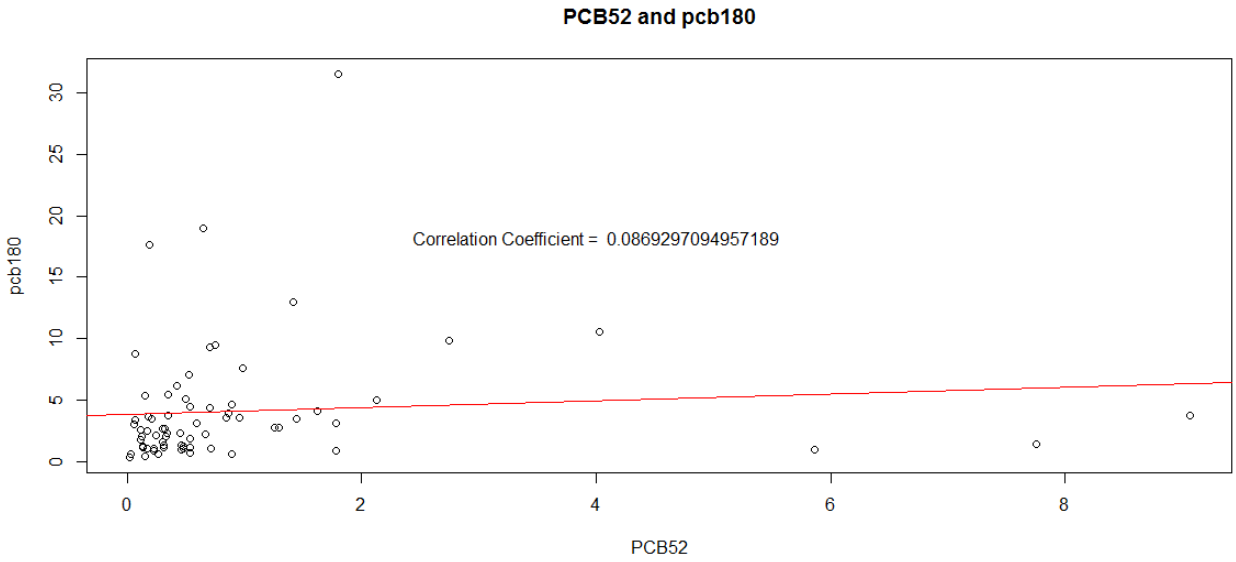




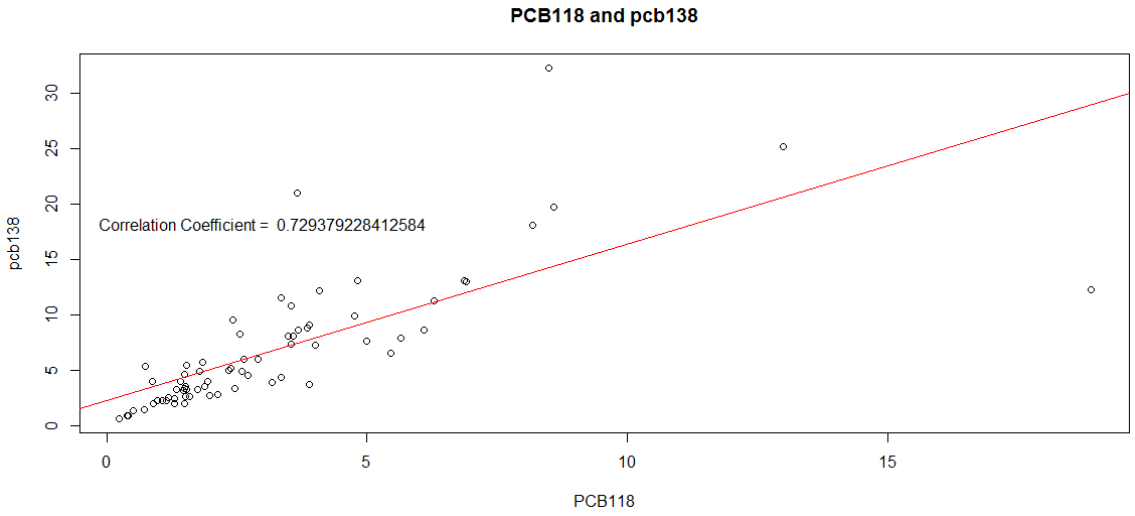

PCB and pcb180
Correlation Coefficient = 0.800854925893532
pcb180
PCB
0
5
10
15
20
25
30
0
50
100
150
200
250
300

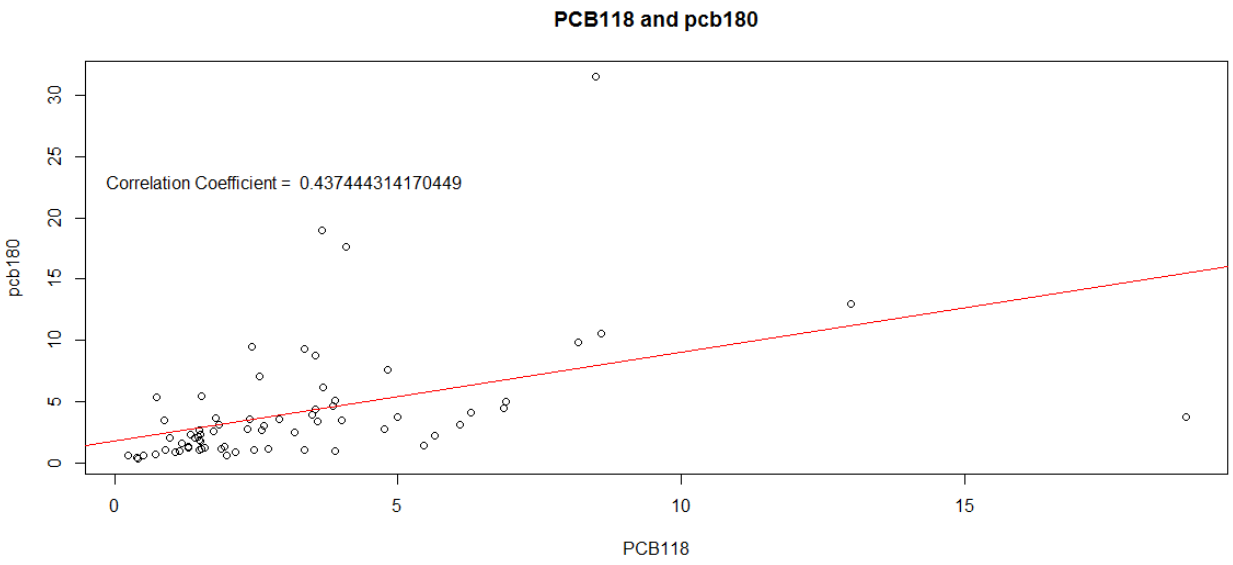

PCB52 and pcb118
Correlation Coefficient = 0.684907271875884
pcb118
PCB52
0
5
10
15
0
2
4
6
8

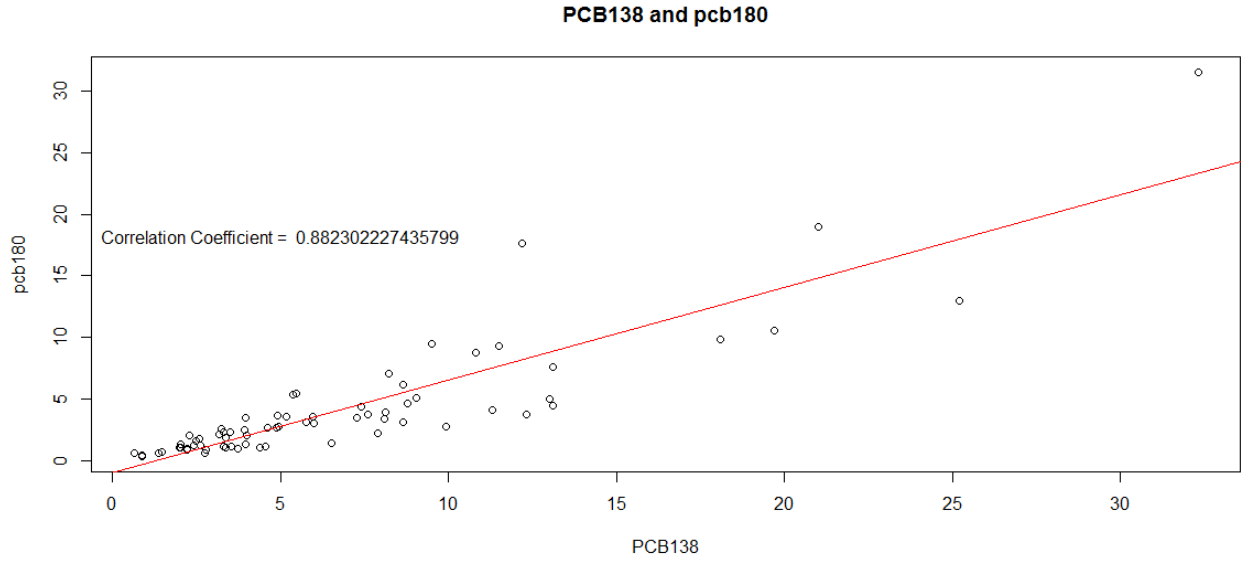

PCB52 and pcb138
Correlation Coefficient = 0.300898275165792
pcb138
PCB52
0
5
10
15
20
25
30
0
2
4
6
8



At 8% correlation, PCB52 and PCB180 are not at all correlated to each other.







43. Use the four congeners PCB52, PCB118, PCB138, and PCB180 in multiple regression to predict PCB.

1. Write the statistical model for this analysis. Include all assumptions.

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$, i from 1 to 69. ε_i are independent N(0, sigma) random normally distributed variables.

2. Run the regression and summarize the results.

```
> summary(fm)

Call:
lm(formula = pcb ~ pcb52 + pcb118 + pcb138 + pcb180)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0864  -2.4554   0.0278   2.7726  22.5487

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9369     1.2293   0.762    0.449
pcb52        11.8727     0.7290  16.287  < 2e-16 ***
pcb118        3.7611     0.6424   5.855 1.79e-07 ***
pcb138        3.8842     0.4978   7.803 7.19e-11 ***
pcb180        4.1823     0.4318   9.687 3.64e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.382 on 64 degrees of freedom
Multiple R-squared:  0.9891,    Adjusted R-squared:  0.9885
F-statistic:  1456 on 4 and 64 DF,  p-value: < 2.2e-16
```

The coefficients are 0.9369, 11.87, 3.76, 3.88, and 4.1. Most of these are significantly different from 0, except for 0.9369.

3. Examine the residuals. Do they appear to be approximately Normal? When you plot them versus each of the explanatory variables, are any patterns evident?

```
> res
          1           2           3           4           5           6
 1.277227199 -1.344066255 -0.943159230  2.811216840 -0.469721237 -1.562541821
          7           8           9          10          11          12
 4.780579947 -3.852704838  0.964095799  1.085172383  0.897188064  2.579050872
         13          14          15          16          17          18
-3.462115492 -1.541703972  0.027812561 -7.394577585 -0.638354046 -5.731690897
         19          20          21          22          23          24
 0.006181009  3.574209373  1.680032527  1.347611295 -0.843545607 -2.455448440
         25          26          27          28          29          30
-0.453215596 -2.190284709  4.302991465  0.741042524  3.839787739  4.370864406
         31          32          33          34          35          36
 3.357573258 -4.821365085 -4.828862396  0.667832089  2.338384014  0.356585555
         37          38          39          40          41          42
 7.953912908  2.389520186 -3.158377333  0.839461515 11.840513719  0.914297734
         43          44          45          46          47          48
 7.509891183 12.732683493 -0.283854210  7.927519126 -2.106000207  0.292992411
         49          50          51          52          53          54
-2.400210728 -22.086433486  8.249594717  6.620221587 -7.330680537 -1.413467501
         55          56          57          58          59          60
-6.049523233 -4.375379996  2.772641810 -11.979195316 -6.680819912  0.294339736
         61          62          63          64          65          66
-5.363775948 -1.952310376 -1.433718591 -3.633086432 22.548701115 -13.415763340
         67          68          69
-8.614067208  3.614435328  7.303856070
```

```
> stem(res)

  The decimal point is 1 digit(s) to the right of the |

  -2 | 2
  -1 |
  -1 | 32
  -0 | 977766555
  -0 | 444332222222111111000
   0 | 000001111111112223333444444
   0 | 5778888
   1 | 23
   1 |
   2 | 3
```

The patterns are roughly normal, as per the stem graph, though there appear to be two outliers, which are -22 and 23. No real patterns emerge.

44. The examination of the residuals in part (c) of the previous exercise suggests that there may be two outliers, one with a high residual and one with a low residual.

1. Because of safety issues, we are more concerned about underestimating PCB in a specimen than about overestimating. Give the specimen number for each of the two suspected outliers. Which one corresponds to an overestimate of PCB?

The two specimens are #50 and #65. As specimen 65 is greatly higher residual than most other values, it corresponds to an overestimate of PCB.

2. Rerun the analysis with the two suspected outliers deleted, summarize these results, and compare them with those you obtained in the previous exercise.

```
> summary(fm)

Call:
lm(formula = dsfixed[, "pcb"] ~ dsfixed[, "pcb52"] + dsfixed[,
    "pcb118"] + dsfixed[, "pcb138"] + dsfixed[, "pcb180"])

Residuals:
     Min       1Q   Median       3Q      Max
-12.2421  -2.1762  -0.1378   1.7036  14.2051

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.6277     0.8858   1.838   0.0709 .
dsfixed[, "pcb52"]   14.4420     0.6960  20.751  < 2e-16 ***
dsfixed[, "pcb118"]   2.5996     0.5164   5.034 4.40e-06 ***
dsfixed[, "pcb138"]   4.0541     0.3752  10.805 6.89e-16 ***
dsfixed[, "pcb180"]   4.1086     0.3175  12.942  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.555 on 62 degrees of freedom
Multiple R-squared:  0.9941,    Adjusted R-squared:  0.9938
F-statistic:  2629 on 4 and 62 DF,  p-value: < 2.2e-16
```

```
> stem(res)

  The decimal point is 1 digit(s) to the right of the |

  -1 | 2
  -0 | 987655
  -0 | 44433333322221111111100000000
   0 | 00001111111122222233344
   0 | 5778899
   1 | 4
```

By removing the two outliers, we see the same multiple regression go from R-squared = 0.98 to R-squared = 0.99, making the model even better at predicting the value of the PCB dataset. The stems now make an almost perfect normal distribution.

45. Run a regression to predict PCB using the variables PCB52, PCB118, and PCB138. Note that this is similar to the analysis you did in Exercise 11.43, with the change that PCB180 is not included as an explanatory variable.

1. Summarize the results.

```
> summary(fm)

Call:
lm(formula = pcb ~ pcb52 + pcb118 + pcb138)

Residuals:
     Min       1Q   Median       3Q      Max
-29.6219  -3.3502   0.8791   3.3785  29.5217

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.0184     1.8895  -0.539    0.592
pcb52        12.6442     1.1291  11.198   <2e-16 ***
pcb118        0.3131     0.8333   0.376    0.708
pcb138        8.2546     0.3279  25.177   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.945 on 65 degrees of freedom
Multiple R-squared:  0.9732,    Adjusted R-squared:  0.972
F-statistic: 786.7 on 3 and 65 DF,  p-value: < 2.2e-16
```

2. In this analysis, the regression coefficient for PCB118 is not statistically significant. Give the estimate of the coefficient and the associated P-value.

The estimate of the coefficient of PCB118 is 0.3131, which is significantly close to 0. The associated P value is seen in the table as well, which is 0.708.

3. Find the estimate of the coefficient for PCB118 and the associated P-value for the model analyzed in Exercise 11.43.

In exercise 11.43, the coefficient for PCB118 was 3.76, and its associated P-value was 1.79*e^-7, which is significant at the 5%, or even 1% level.

4. Using the results in parts (b) and (c), write a short paragraph explaining how the inclusion of other variables in a multiple regression can have an effect on the estimate of a particular coefficient and the results of the associated significance test.

By removing variables from a multiple regression, it is possible to lose a piece of the bigger picture. We see a drop in the significance of other variables, that may be relevant to the whole dataset, but not a subset of it. As such, it is important to include as much data as possible for a more accurate multiple regression, coefficient, and significance test.

46. Dioxins and furans are other classes of chemicals that can cause undesirable health effects similar to those caused by PCB. Three types of chemicals are combined using toxic equivalent scores (TEQs), which attempt to measure the health effects on a common scale. The PCB data file contains TEQs for PCB dioxins, and furans. The variables are called TEQPCB, TEQDIOXIN, and TEQFURAN. The data file also includes the total TEQ, desired to be the sum of these three variables.

1. Consider using a multiple regression to predict TEQ using the three components TEQPCB, TEQDIOXIN, and TEQFURAN as explanatory variables. Write the multiple regression model in the form TEQ = $\beta 0 + \beta 1$TEQPCB + $\beta 2$TEQDIOXIN + $\beta 3$TEQFURAN + ϵ. Give numerical values for the parameters $\beta 0$, $\beta 1$, $\beta 2$, and $\beta 3$.

Model is: TEQ = 3.426*e^-6 + (1)TEQPCB + (1)TEQDIOXIN + (1)TEQFURAN + ϵ

$\beta 0$ almost equals 0, $\beta 1$, $\beta 2$, and $\beta 3$ all equal 1.

2. The multiple regression model assumes that the ϵ's are Normal with mean zero and standard deviation σ. What is the numerical value of σ?

σ = SE/sqrt(n) = 7.95e-06/sqrt(69) approximately = 0. As such, the value of σ is 0.

3. Use software to run this regression and summarize the results.

```
> summary(fm)

Call:
lm(formula = ds[, "teq"] ~ ds[, "teqpcb"] + ds[, "teqdioxin"] +
    ds[, "teqfuran"])

Residuals:
       Min         1Q     Median         3Q        Max
-5.638e-06 -2.844e-06 -1.680e-06 -1.130e-06  3.714e-05

Coefficients:
                    Estimate Std. Error   t value Pr(>|t|)
(Intercept)        3.426e-07  1.917e-06 1.790e-01    0.859
ds[, "teqpcb"]     1.000e+00  8.239e-07 1.214e+06   <2e-16 ***
ds[, "teqdioxin"] 1.000e+00  1.761e-06 5.677e+05   <2e-16 ***
ds[, "teqfuran"]  1.000e+00  5.664e-06 1.766e+05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.95e-06 on 65 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 9.581e+11 on 3 and 65 DF,  p-value: < 2.2e-16
```

```
> stem(res)

  The decimal point is 5 digit(s) to the left of the |

  -0 | 655555
  -0 | 44444333333333332222222222222222211111111111111111111111
   0 |
   0 |
   1 |
   1 | 568
   2 |
   2 | 69
   3 |
   3 | 7
```

The R-squared value for this approximation is literally 1. As such, these values are able to perfectly approximate TEQ. The distribution of the residuals is a right-skewed normal distribution.

47. The information summarized in TEQ is used to assess and manage risks from these chemicals. For example, the World Health Organization (WHO) has established the tolerable daily intake (TD1) as 1 to 4 TEQs per kilogram of body weight per day. Therefore, it would be very useful to have a procedure for estimating TEQ using just a few variables that can be measured cheaply. Use the for PCB congeners PCB52, PCB118, PCB138, and PCB180 in a multiple regression to predict TEQ. give a description of the model and assumptions, summarize the results, examine the residuals and write a summary of what you have found.

The model is: $TEQ_i = \beta_0 + \beta_1 PCB52 + \beta_2 PCB118 + \beta_3 PCB138 + \beta_4 PCB180 + \varepsilon_i$, i from 1 to 69. ε_i are independent N(0, sigma) random normally distributed variables.

```
> summary(fm)

Call:
lm(formula = teq ~ pcb52 + pcb118 + pcb138 + pcb180)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6655 -0.6000 -0.1814  0.5162  2.7025

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.059965   0.184450   5.747 2.73e-07 ***
pcb52       -0.097277   0.109383  -0.889  0.37716
pcb118       0.306184   0.096388   3.177  0.00229 **
pcb138       0.105786   0.074697   1.416  0.16156
pcb180      -0.003905   0.064784  -0.060  0.95212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9576 on 64 degrees of freedom
Multiple R-squared:  0.6769,    Adjusted R-squared:  0.6568
F-statistic: 33.53 on 4 and 64 DF,  p-value: 4.489e-15
```

The coefficients are -0.097, 0.306, 0.106, and -0.0039, making the equation:

TEQ = 1.06 – 0.097PCB52 + 0.306PCB118 +0.106PCB138 - 0.0039PCB180.

```
> stem(res)

  The decimal point is at the |

  -1 | 775
  -1 | 3100
  -0 | 98777777666666555555
  -0 | 44433322221111
   0 | 000133333
   0 | 5566677888
   1 | 2444
   1 | 5
   2 | 03
   2 | 67
```

As per this stem plot of the residuals, they seem to be slightly right skewed. No clear patterns emerge.

48. Because distributions of variables such as PCB, the PCB congeners, and TEQ tend to be skewed researchers frequently analyze the logarithms of the measured variables. Create a data set that has the logs of each of the variables in the PCB data file. Note that zero is a possible value for PCB126. Most software packages will eliminate these cases when you request a log transformation.

1. If you do not do anything about the 16 zero values of PCB126, what does your software do with these cases? Is there an error message of some kind?

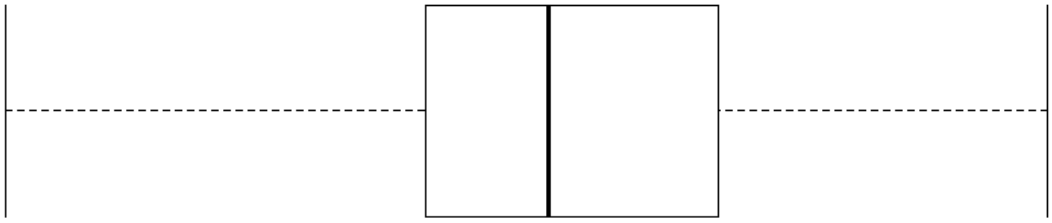
R sets all of the zeroes in PCB126 to negative infinity; as this is the value of log(0). There is no error message upon setting this, though it could cause problems when using the data in equations and graphs.

2. If you attempt to run a regression to predict the log of PCB using the log on PCB126 and the log of PCB52, are the cases with the zero values of PCB126 eliminated? Do you think that this is a good way to handle this situation?

These cases are not eliminated. As such, doing this throws an error with the software. I believe this is reasonable, as arbitrarily removing problematic datapoints without notifying the user could lead to confusion in certain cases. Notifying the user that something is wrong with the dataset, and leaving the solution up to them, is a much safer solution.

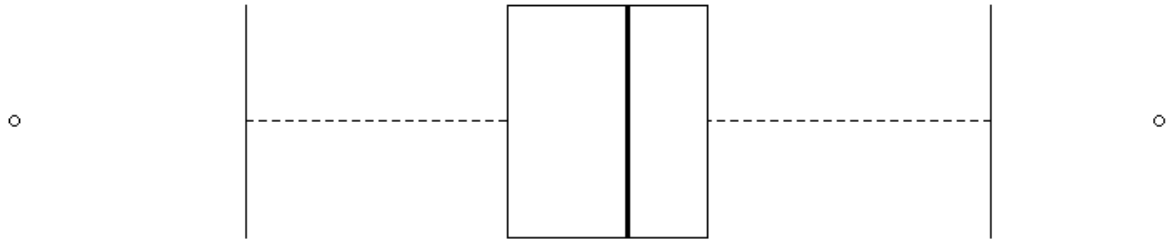
3. The smallest nonzero value of PCB126 is 0.0052. One common practice when taking logarithms of measured values is to replace the zeros by one-half of the smallest observed value. Create a logarithm data set using this procedure; that is, replace the 16 zero values of PCB126 by 0.0026 before taking logarithms. Use numerical and graphical summaries to describe the distributions of the log variables.

Boxplot: PCB



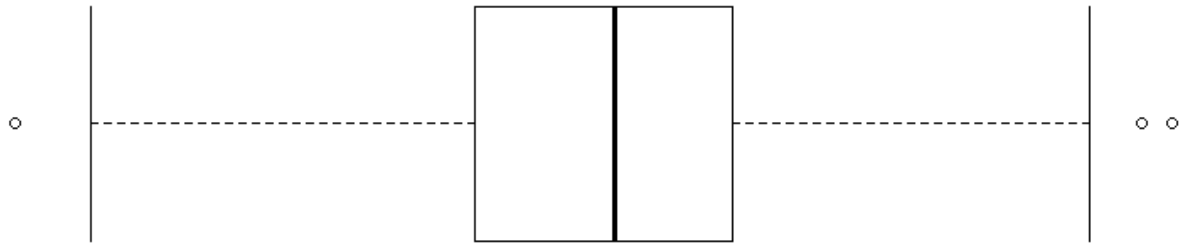
```
print(summary(dslog["pcb"]))
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.808   3.407   3.870   3.917   4.518   5.764
```

Boxplot: PCB28



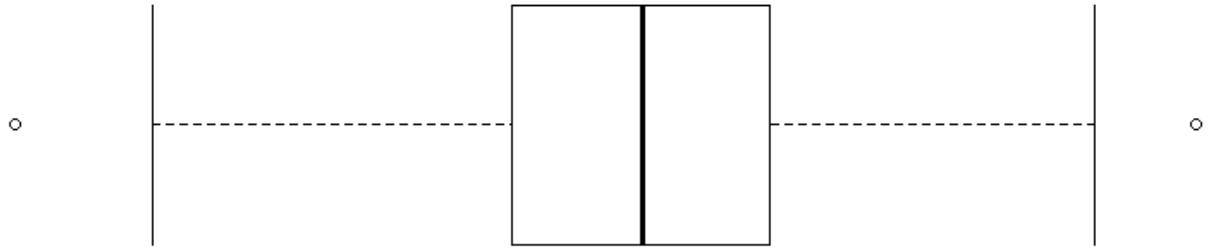
```
> print(summary(dslog[,"pcb28"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-5.1160 -2.0710 -1.3390 -1.3340 -0.8393  1.9360
```

Boxplot: PCB52



```
> print(summary(dslog[,"pcb52"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.9120 -1.4780 -0.7402 -0.7722 -0.1143  2.2040
```

Boxplot: PCB118

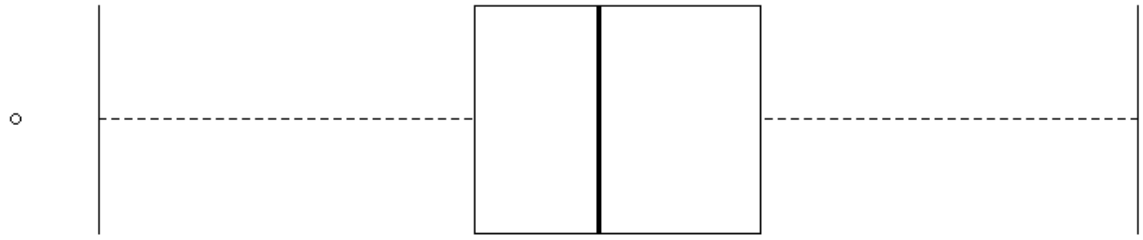


```
> print(summary(dslog[,"pcb118"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.4440  0.3988  0.8838  0.8559  1.3580  2.9390
```

Boxplot: PCB126

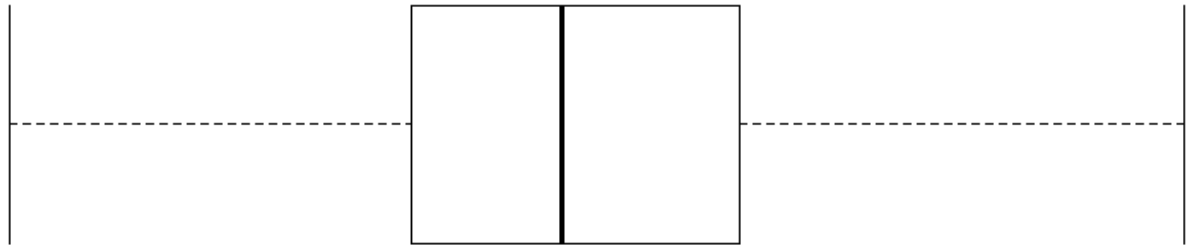
```
> print(summary(dslog[,"pcb126"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -5.952  -5.221  -4.906  -4.846  -4.220  -3.451
```

Boxplot: PCB138



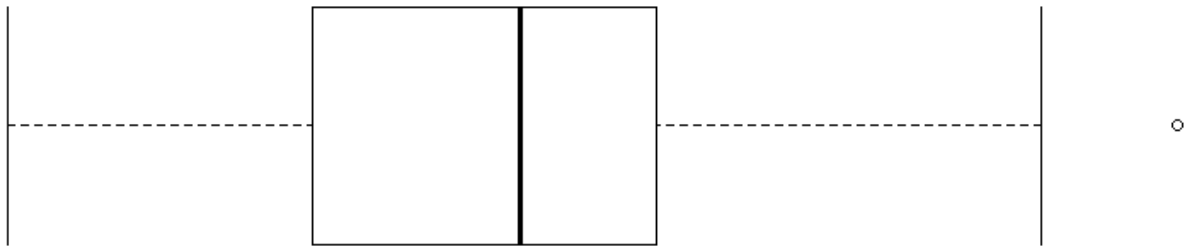
```
> print(summary(dslog[,"pcb138"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.4463  1.1570  1.5930  1.6140  2.1580  3.4750
```

Boxplot: PCB153



```
> print(summary(dslog[,"pcb153"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.1508  1.1940  1.6940  1.7030  2.2890  3.7730
```

Boxplot: PCB180

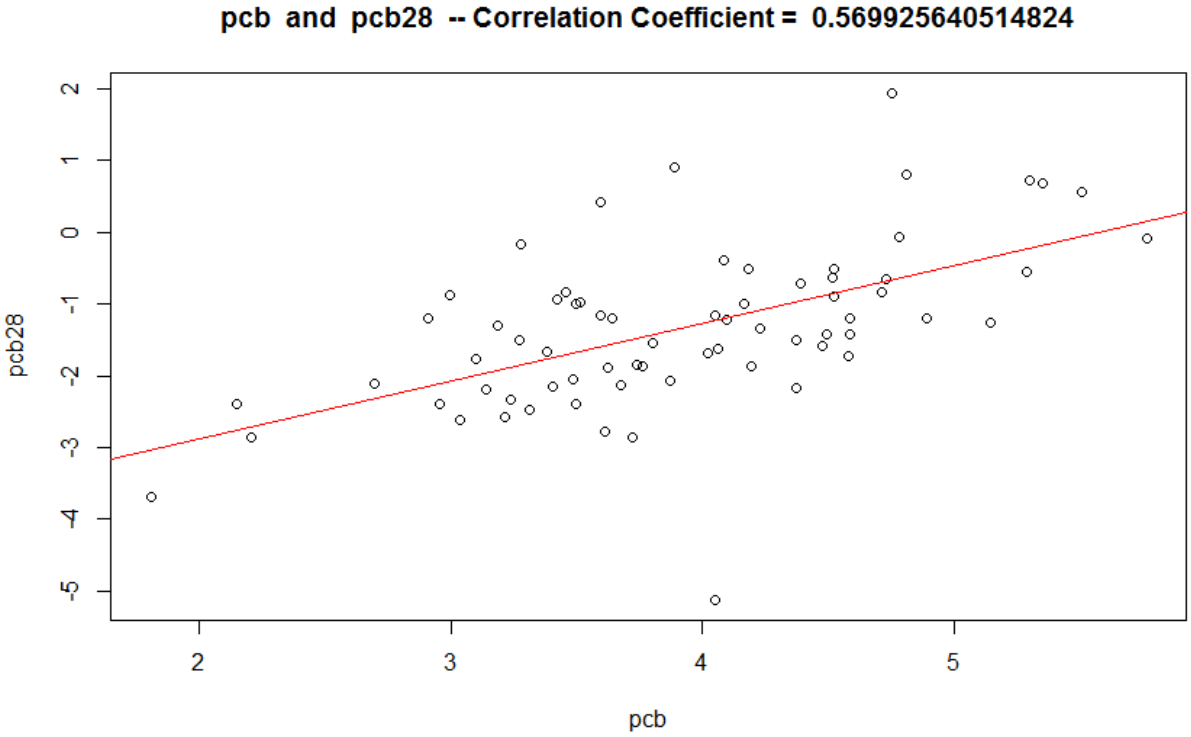


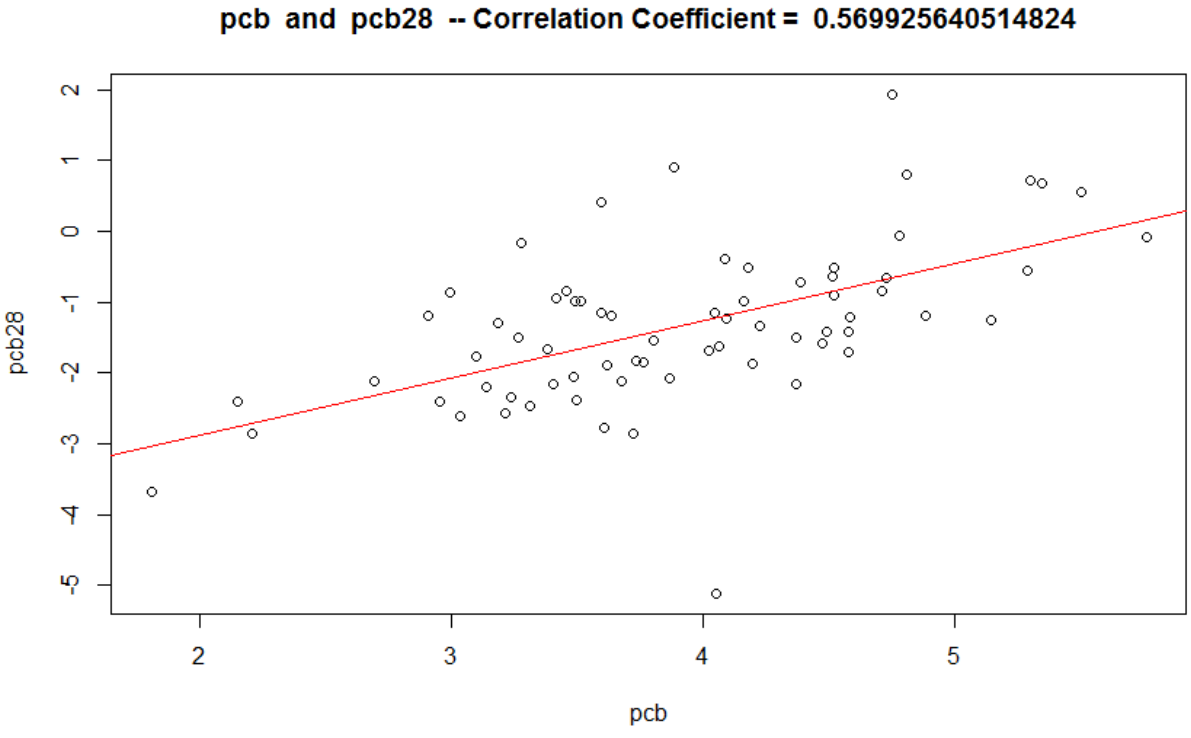
```
> print(summary(dslog[,"pcb180"]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.9289  0.2151  0.9895  0.9752  1.5020  3.4500
```

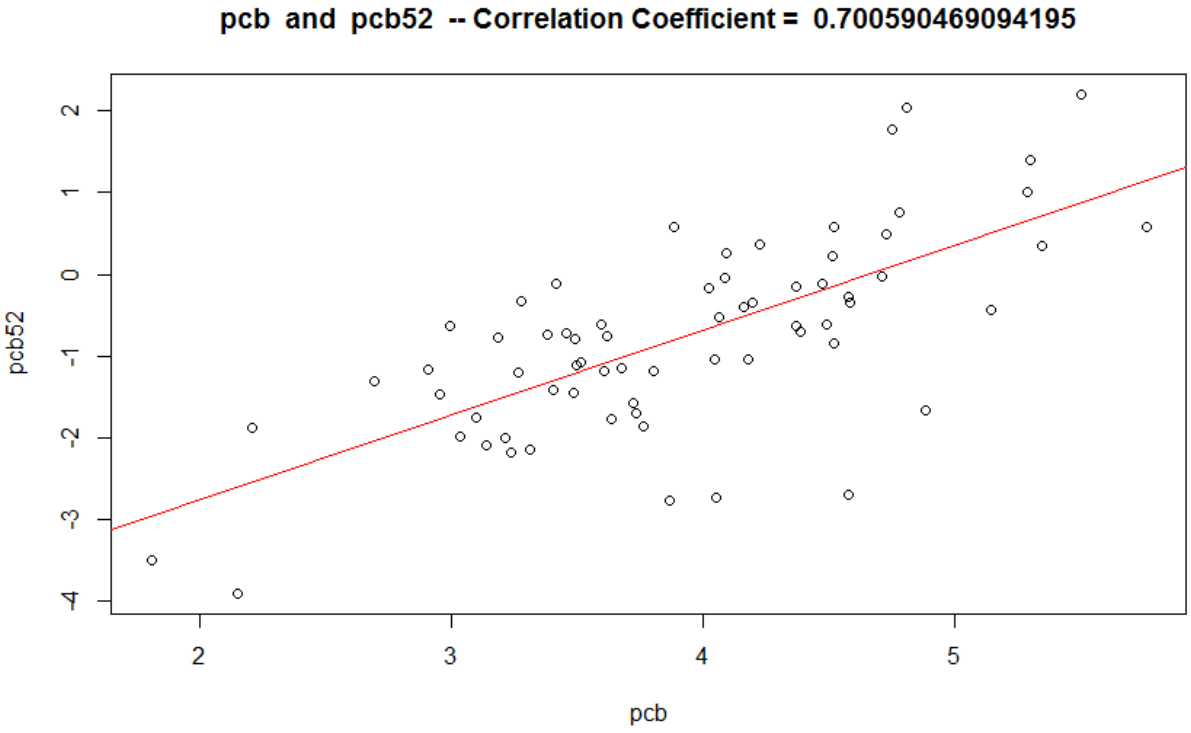
All distributions appear to be approximately normal.

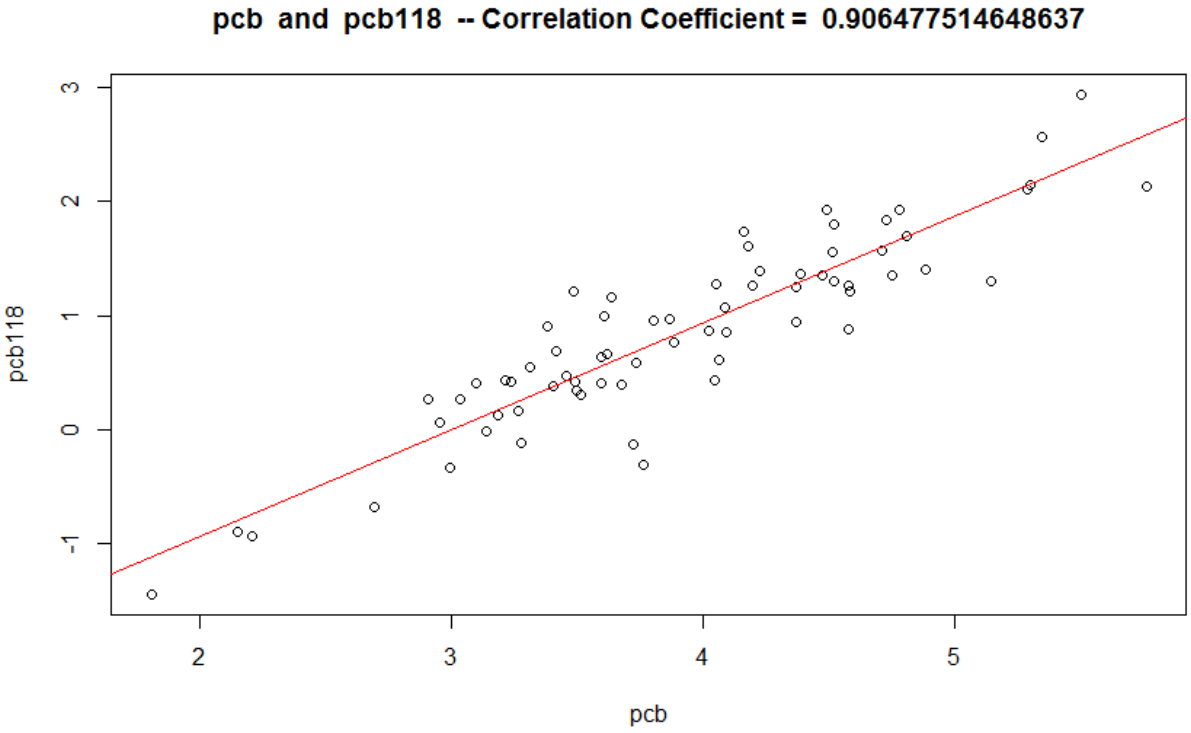
49. Refer to the previous exercise.

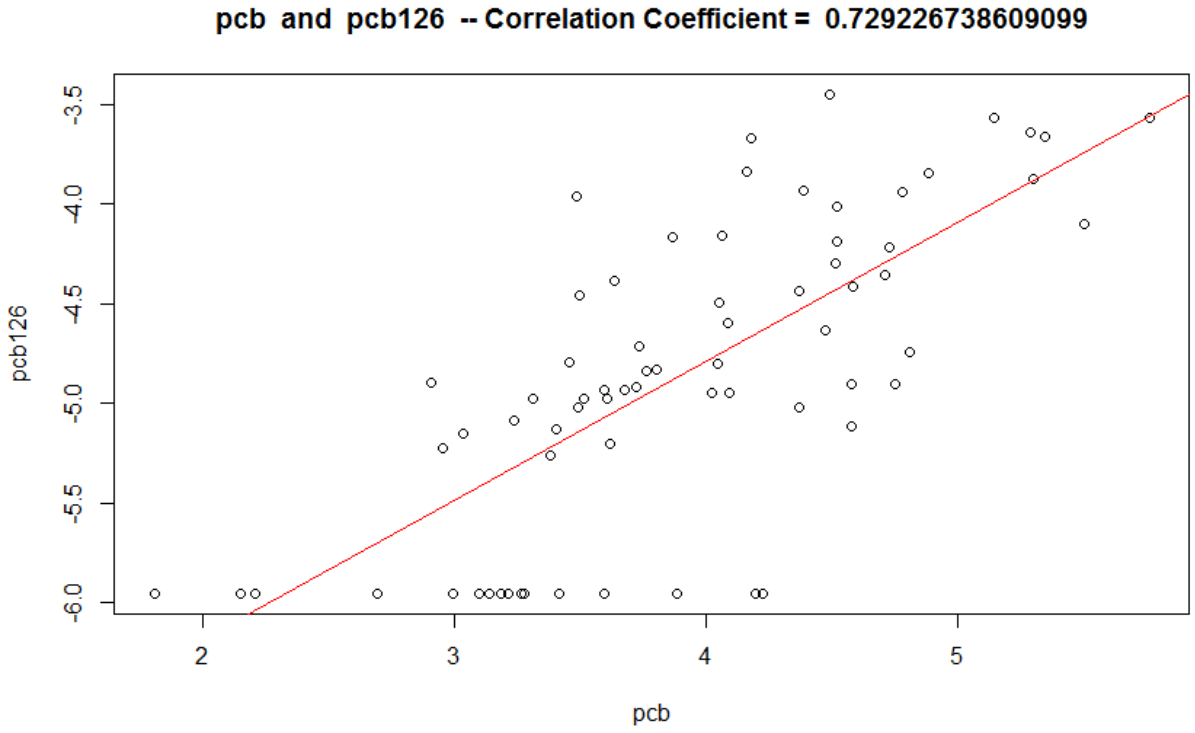
1. Use the numerical and graphical summaries to describe the relationships between each pair of log variables.

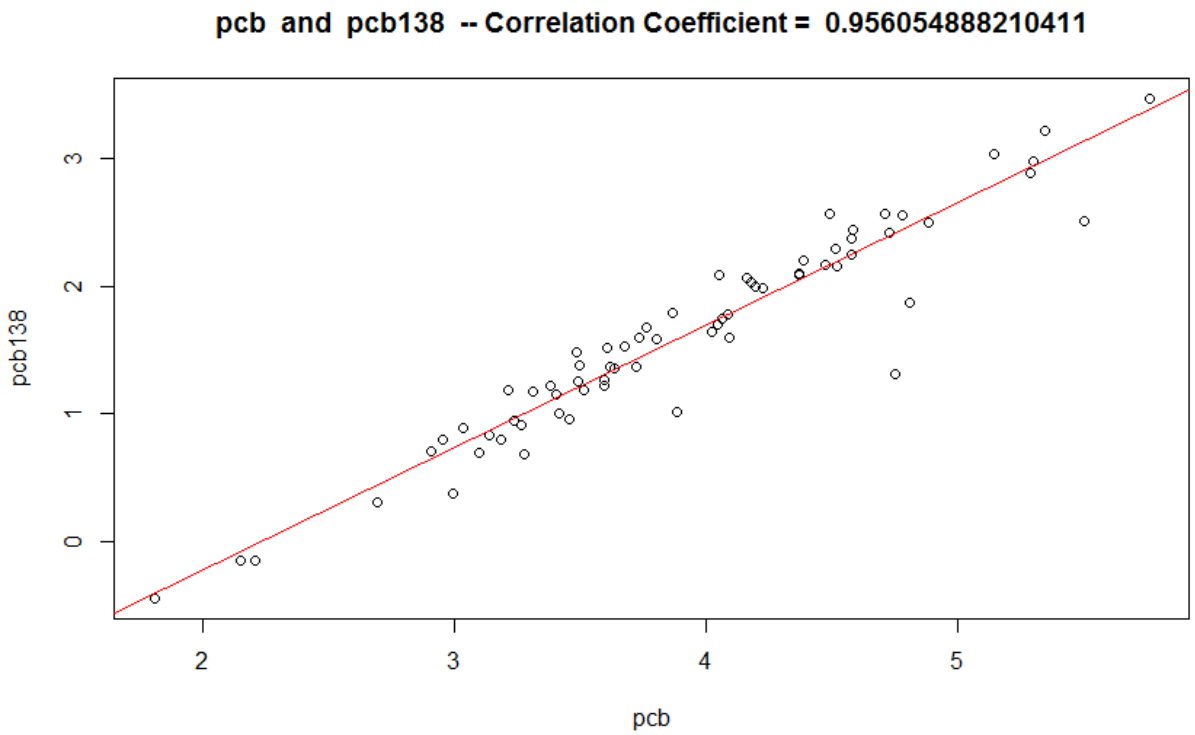

pcb and pcb28 -- Correlation Coefficient = 0.569925640514824
pcb28
pcb
2
3
4
5
-5
-4
-3
-2
-1
0
1
2

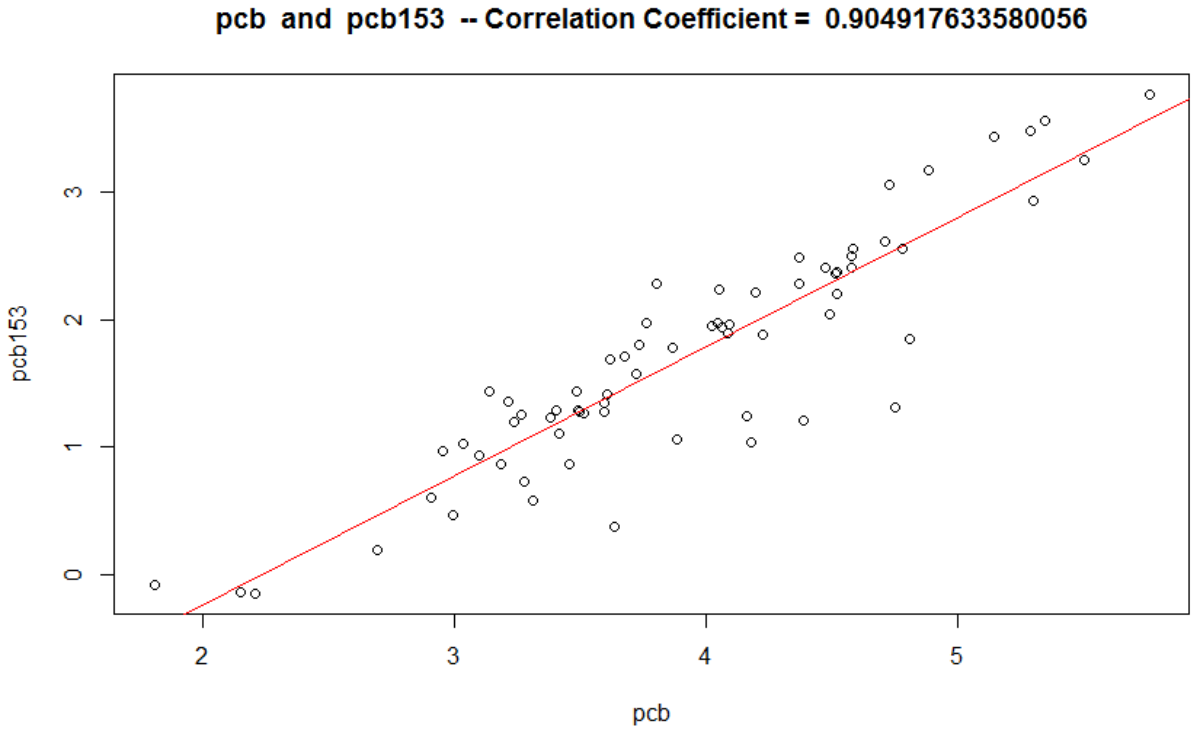

pcb and pcb28 -- Correlation Coefficient = 0.569925640514824
pcb28
pcb
2
3
4
5
-5
-4
-3
-2
-1
0
1
2

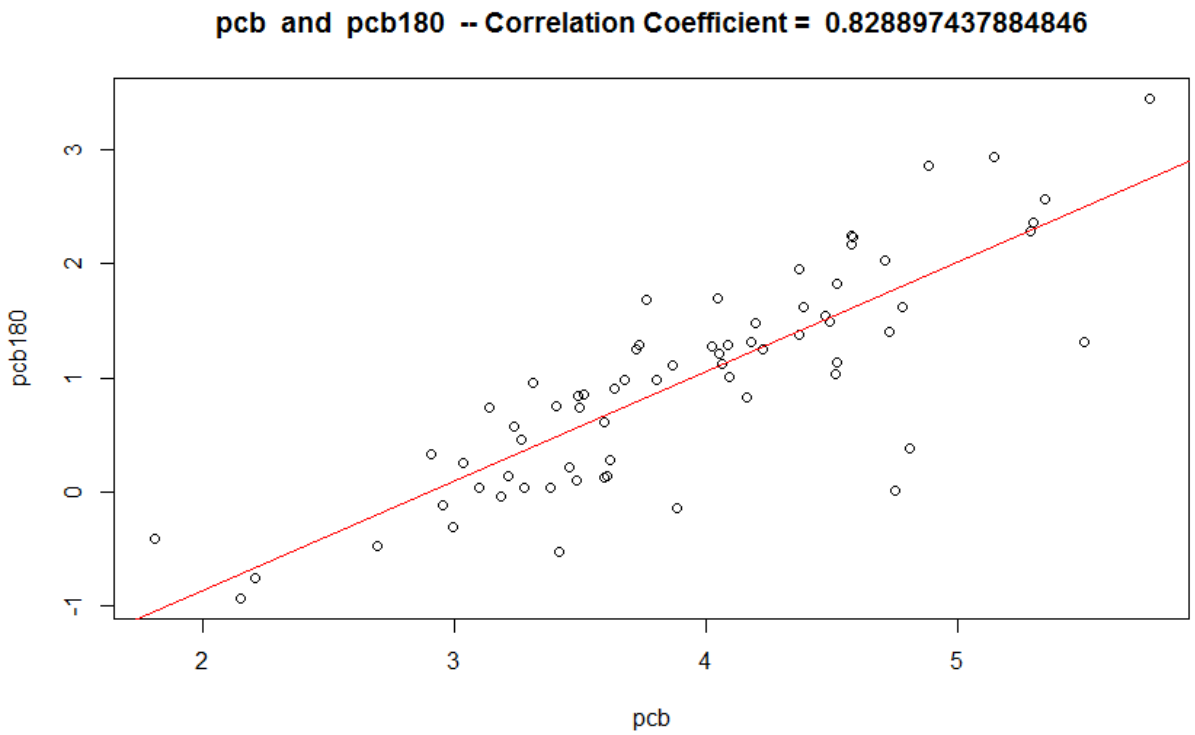

pcb and pcb52 -- Correlation Coefficient = 0.700590469094195
pcb52
pcb
2
3
4
5
-4
-3
-2
-1
0
1
2

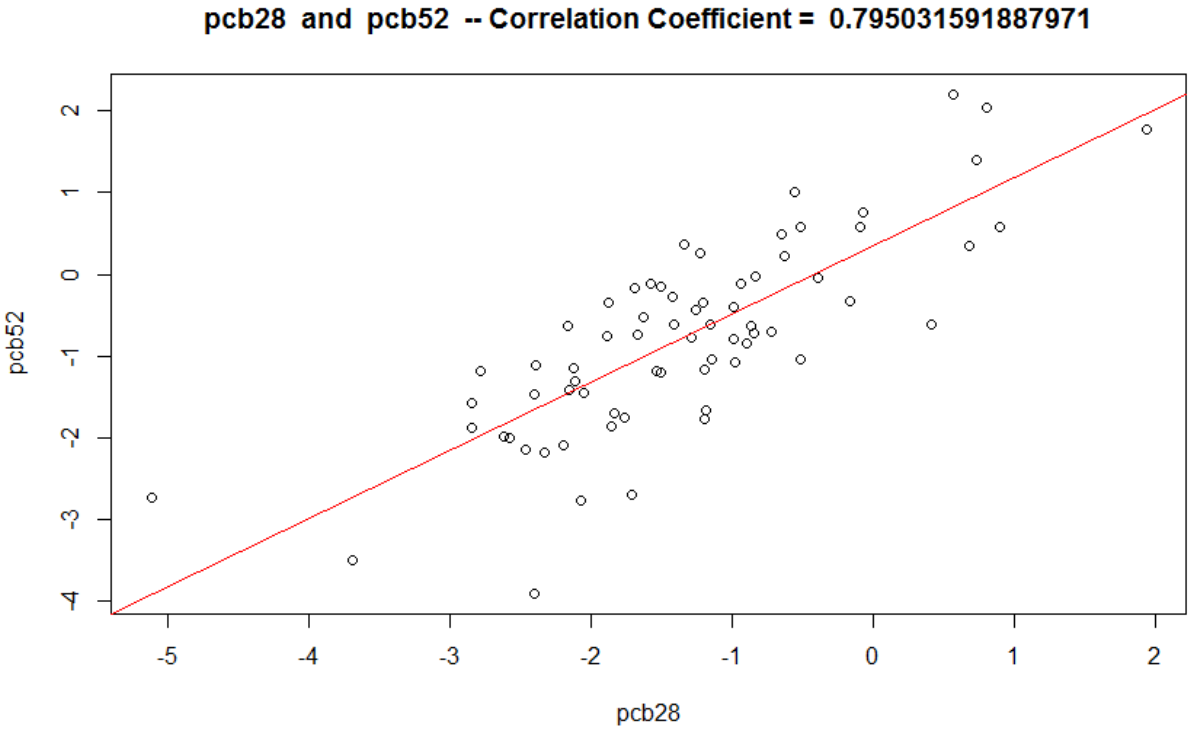

pcb and pcb118 -- Correlation Coefficient = 0.906477514648637
pcb118
pcb
2
3
4
5
-1
0
1
2
3

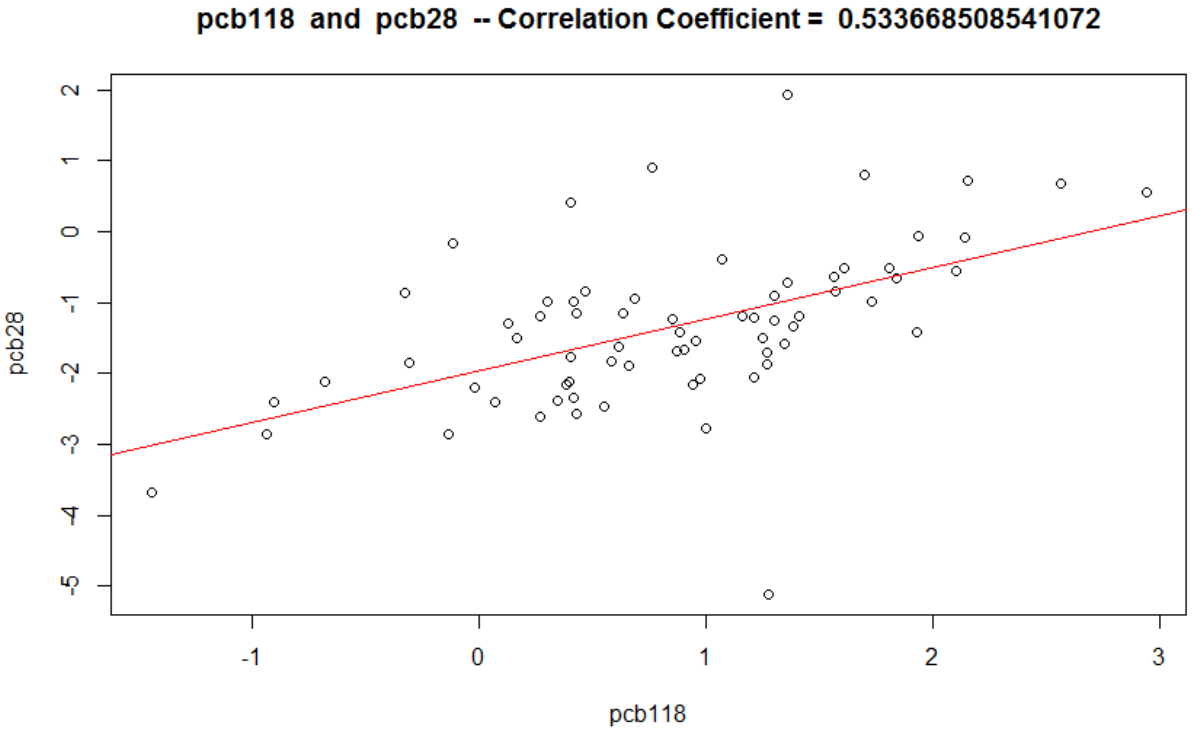

pcb and pcb126 -- Correlation Coefficient = 0.729226738609099
pcb126
-3.5
-4.0
-4.5
-5.0
-5.5
-6.0
2
3
4
5
pcb

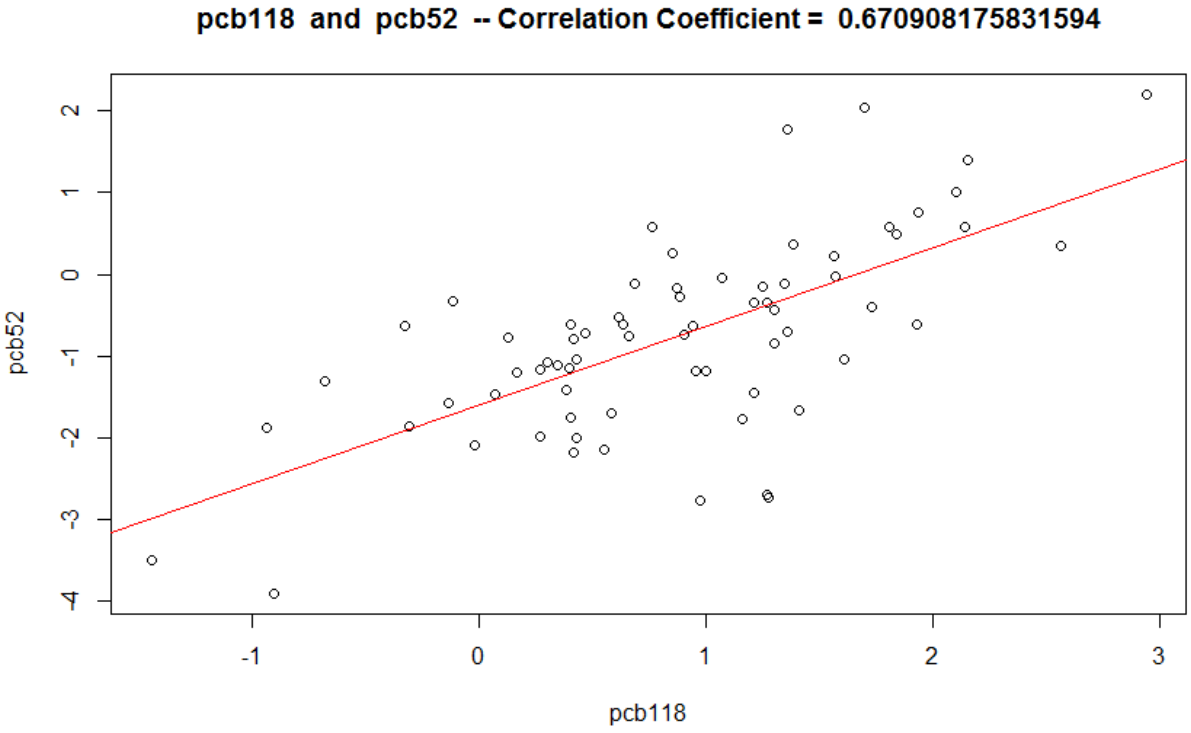

pcb and pcb138 -- Correlation Coefficient = 0.956054888210411
pcb138
3
2
1
0
2
3
4
5
pcb

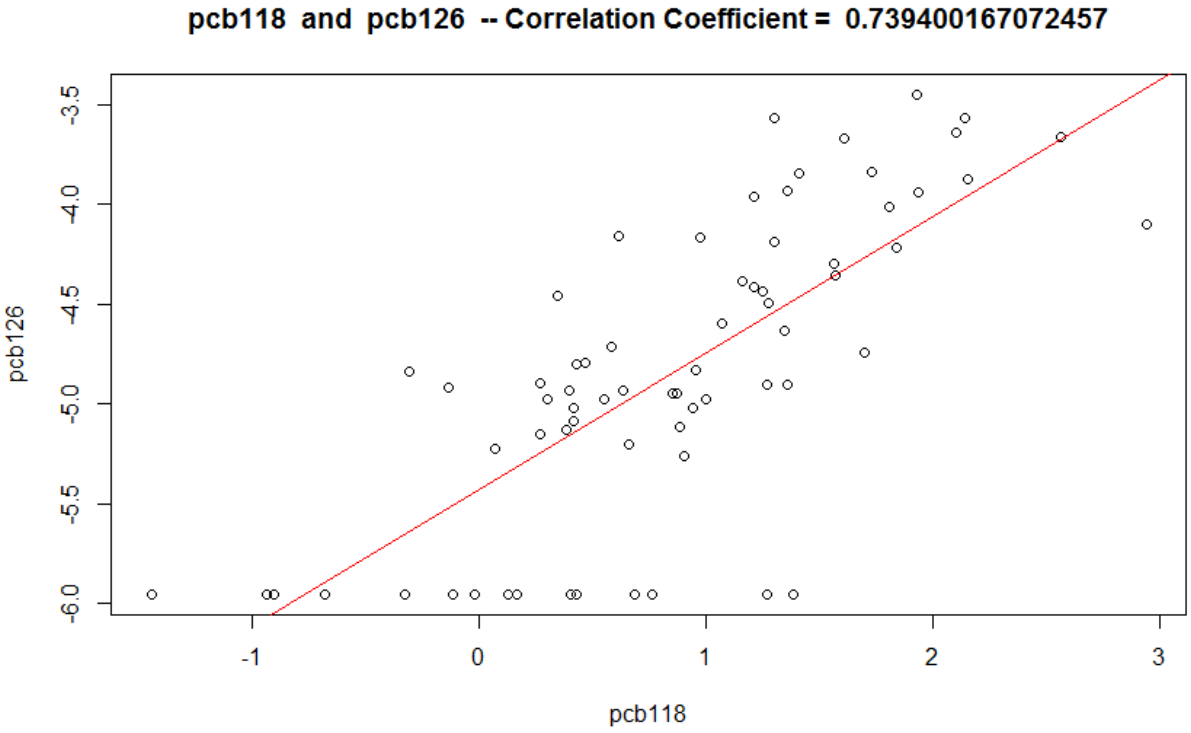

pcb and pcb153 -- Correlation Coefficient = 0.904917633580056
pcb153
pcb
0
1
2
3
2
3
4
5

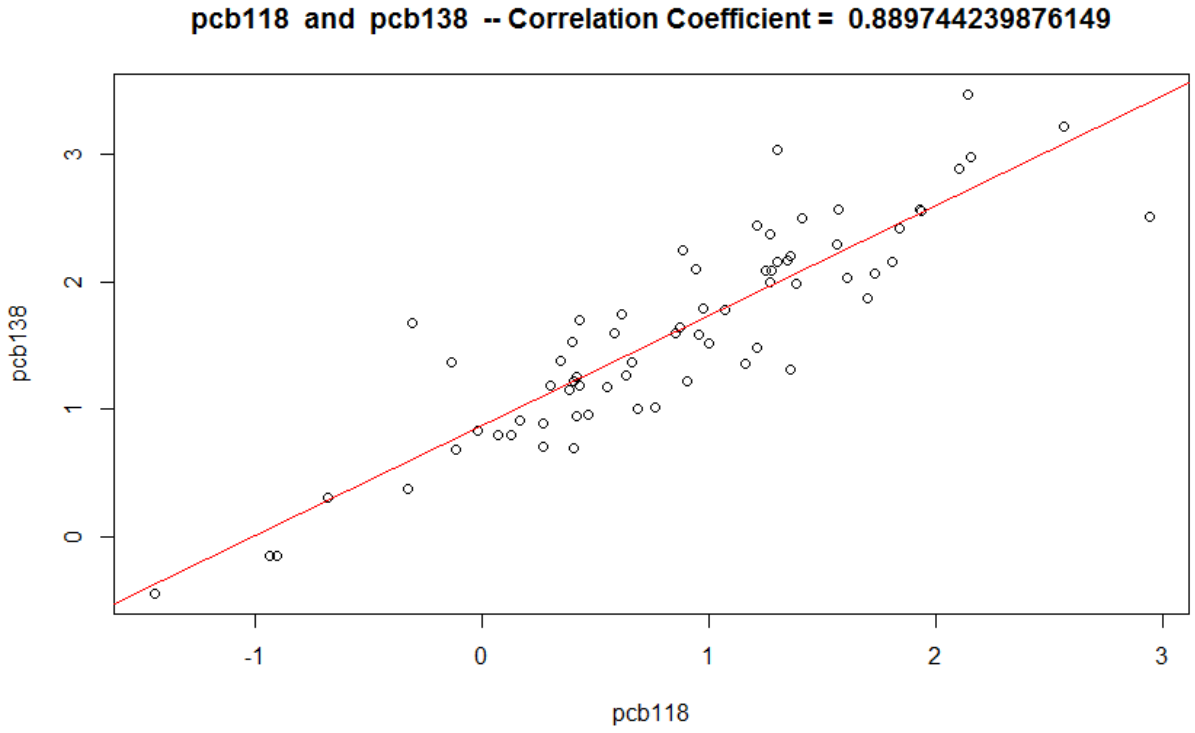

pcb and pcb180 -- Correlation Coefficient = 0.828897437884846
pcb180
pcb
-1
0
1
2
3
2
3
4
5

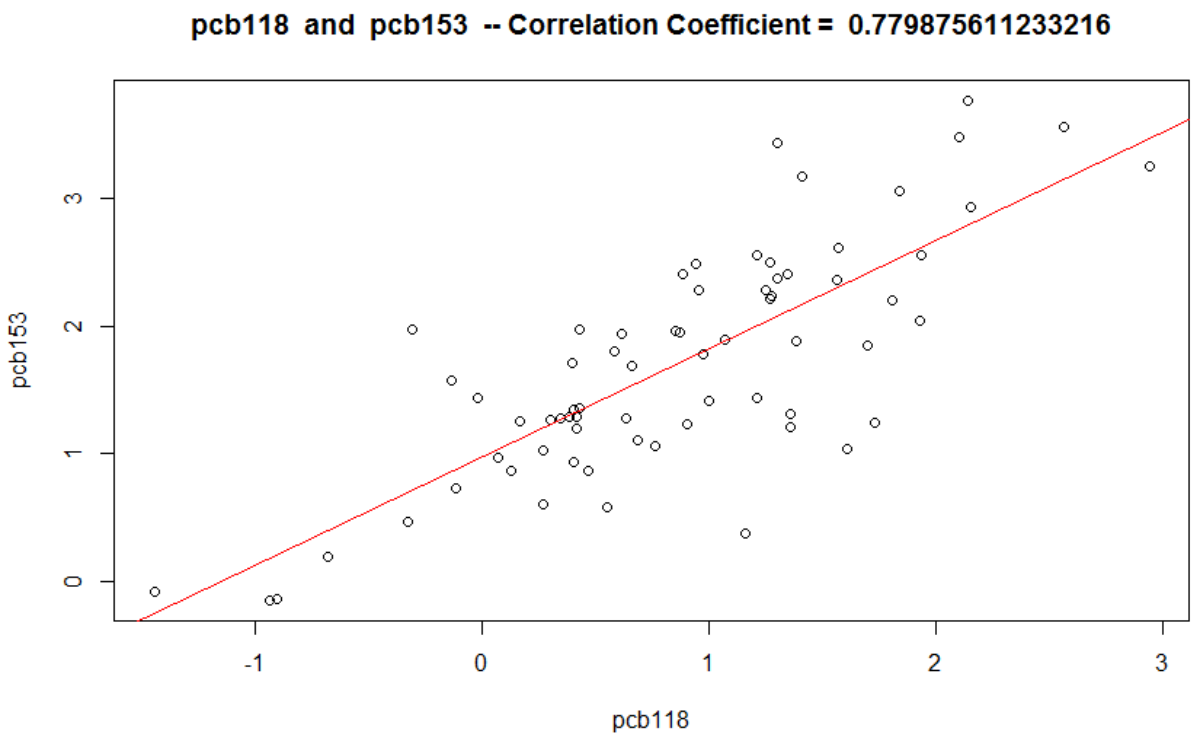

pcb28 and pcb52 -- Correlation Coefficient = 0.795031591887971
pcb52
pcb28
-5
-4
-3
-2
-1
0
1
2

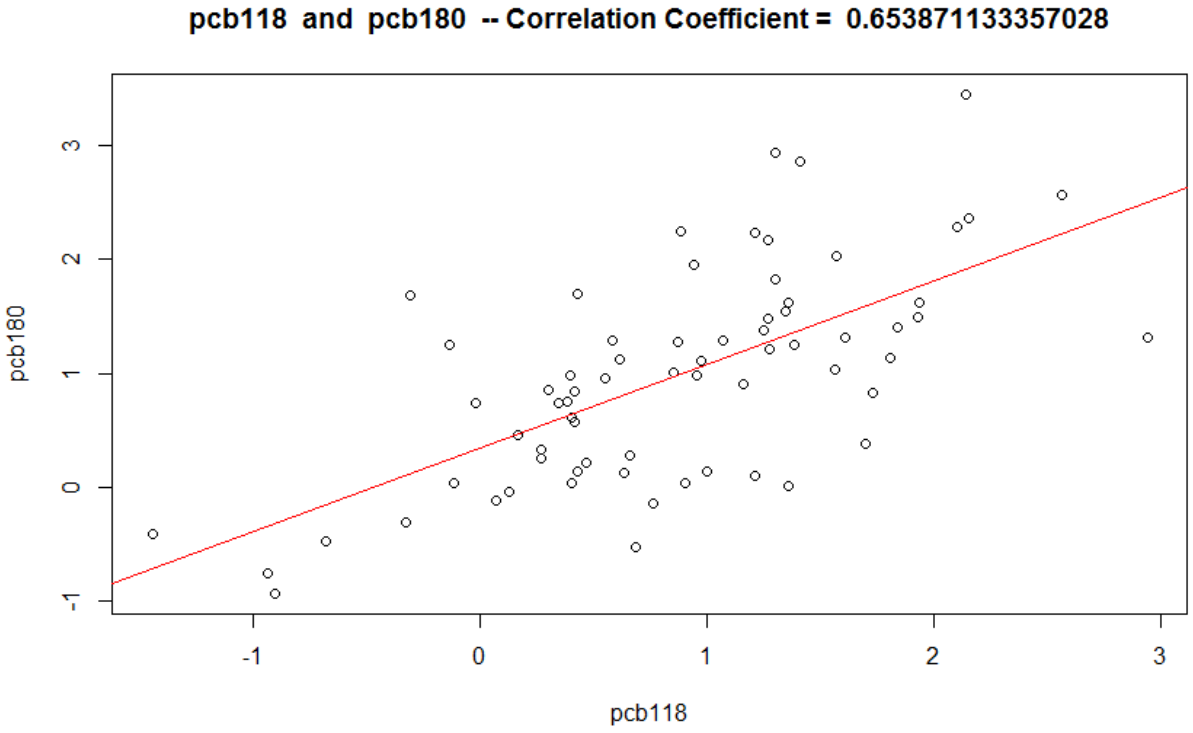

pcb118 and pcb28 -- Correlation Coefficient = 0.533668508541072
pcb28
pcb118
-5
-4
-3
-2
-1
0
1
2
3

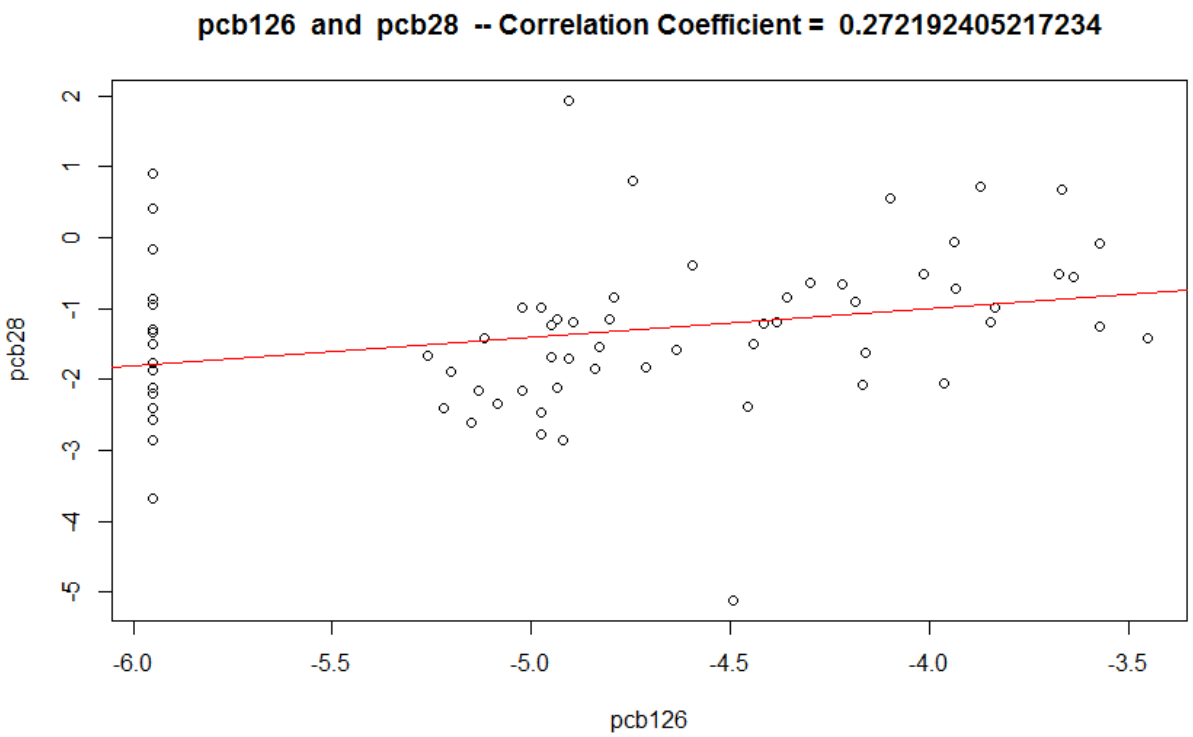

pcb118 and pcb52 -- Correlation Coefficient = 0.670908175831594
pcb52
pcb118
2
1
0
-1
-2
-3
-4
-1
0
1
2
3

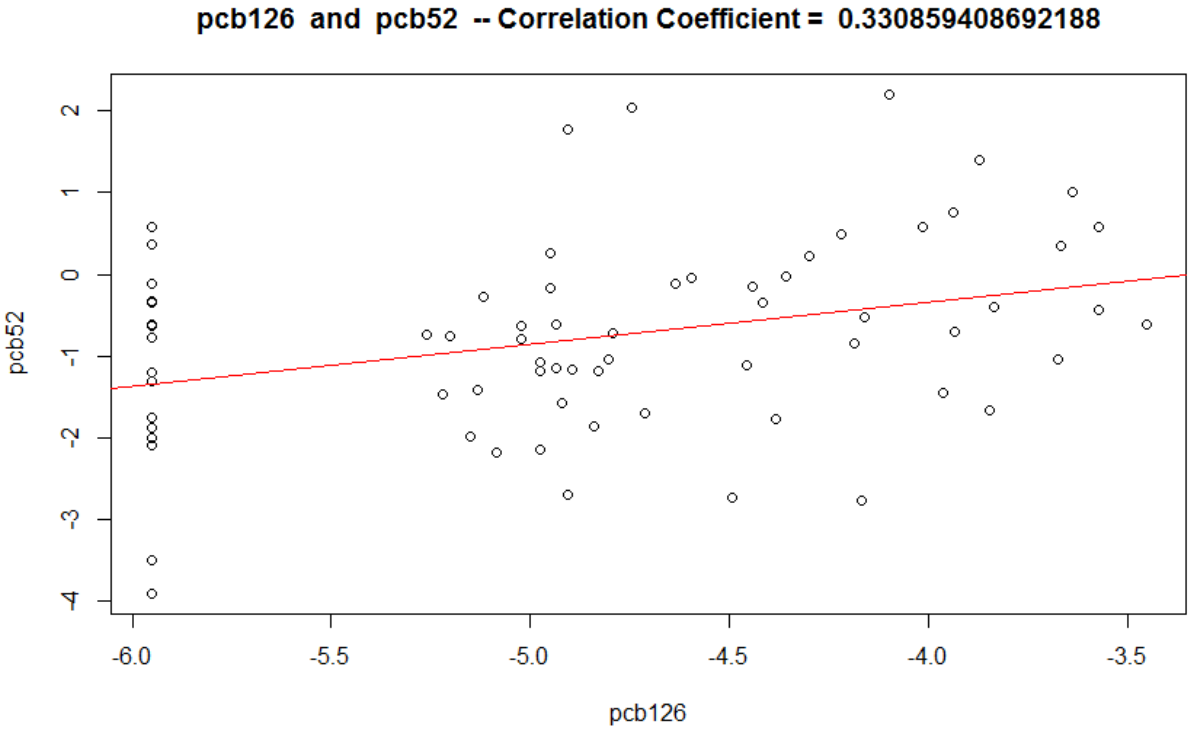

pcb118 and pcb126 -- Correlation Coefficient = 0.739400167072457
pcb126
pcb118
-3.5
-4.0
-4.5
-5.0
-5.5
-6.0
-1
0
1
2
3

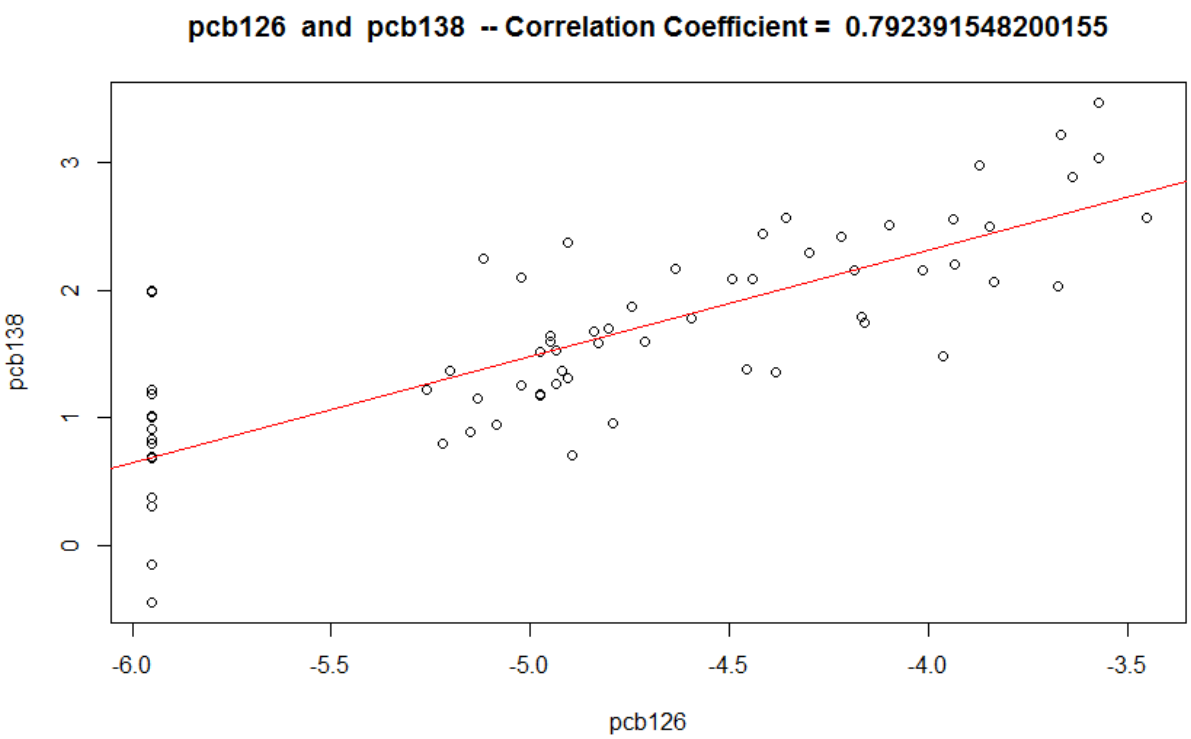

pcb118 and pcb138 -- Correlation Coefficient = 0.889744239876149
pcb138
pcb118
-1
0
1
2
3

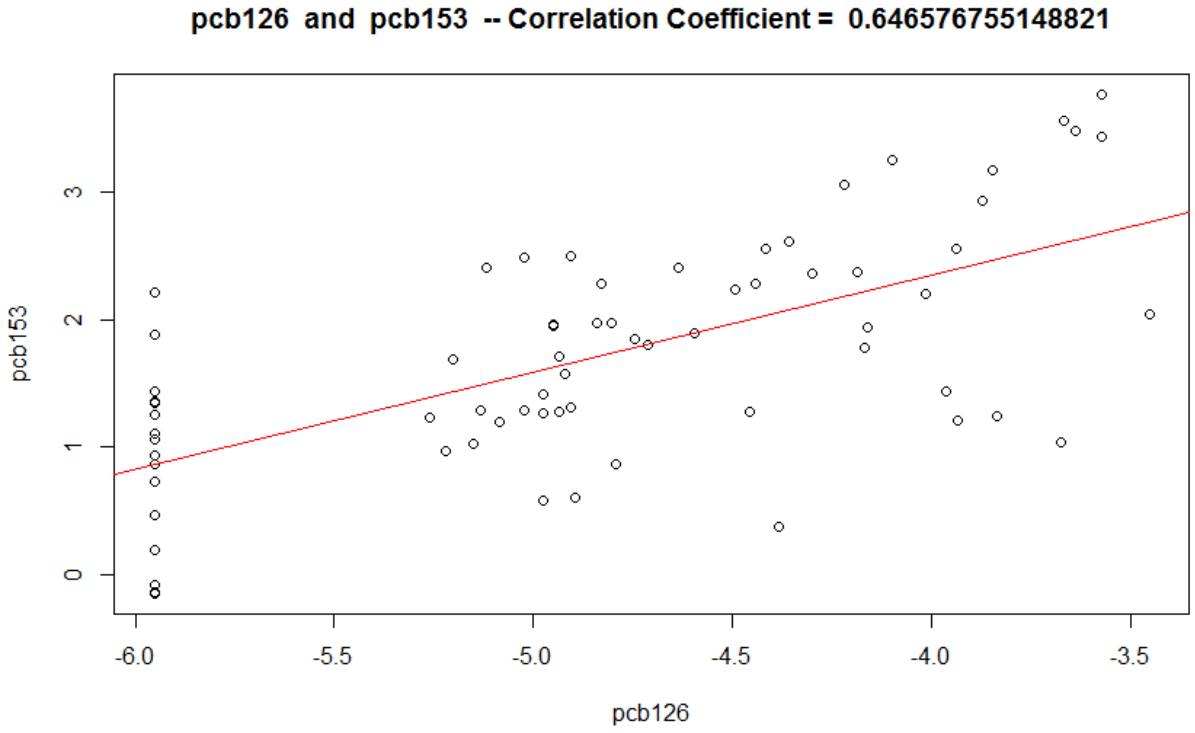

pcb118 and pcb153 -- Correlation Coefficient = 0.779875611233216
pcb153
pcb118
-1
0
1
2
3

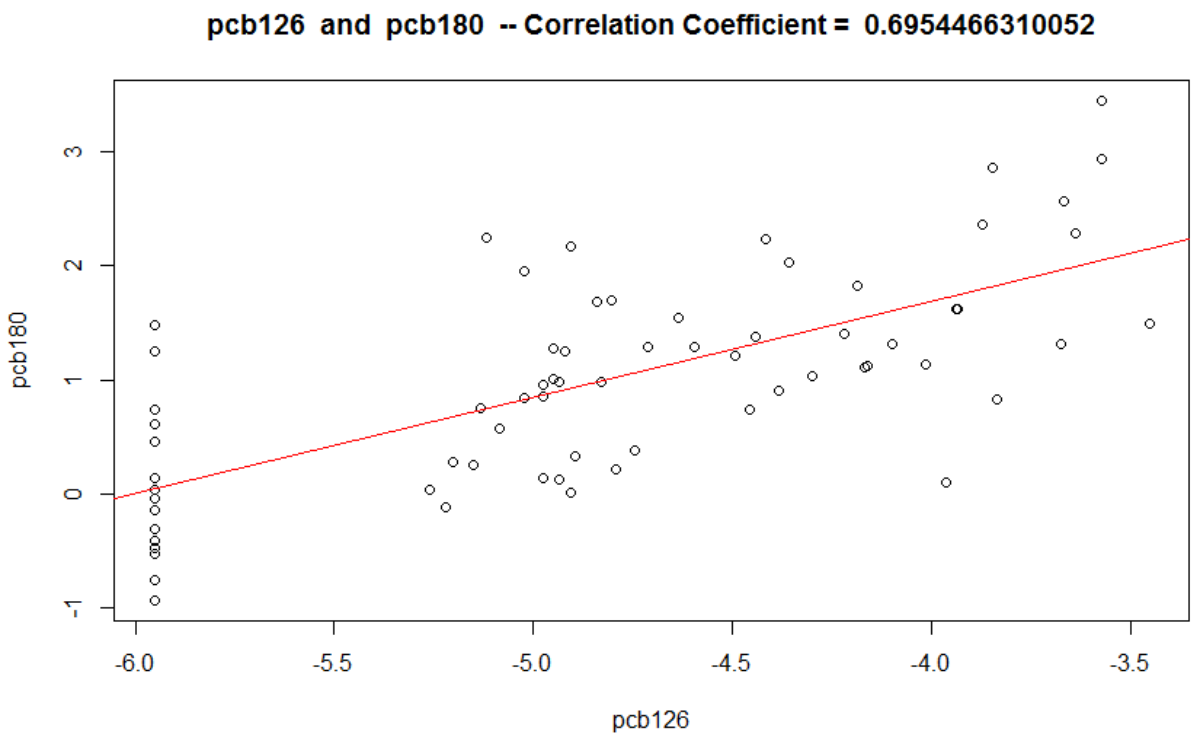

pcb118 and pcb180 -- Correlation Coefficient = 0.653871133357028
pcb180
pcb118
-1
0
1
2
3

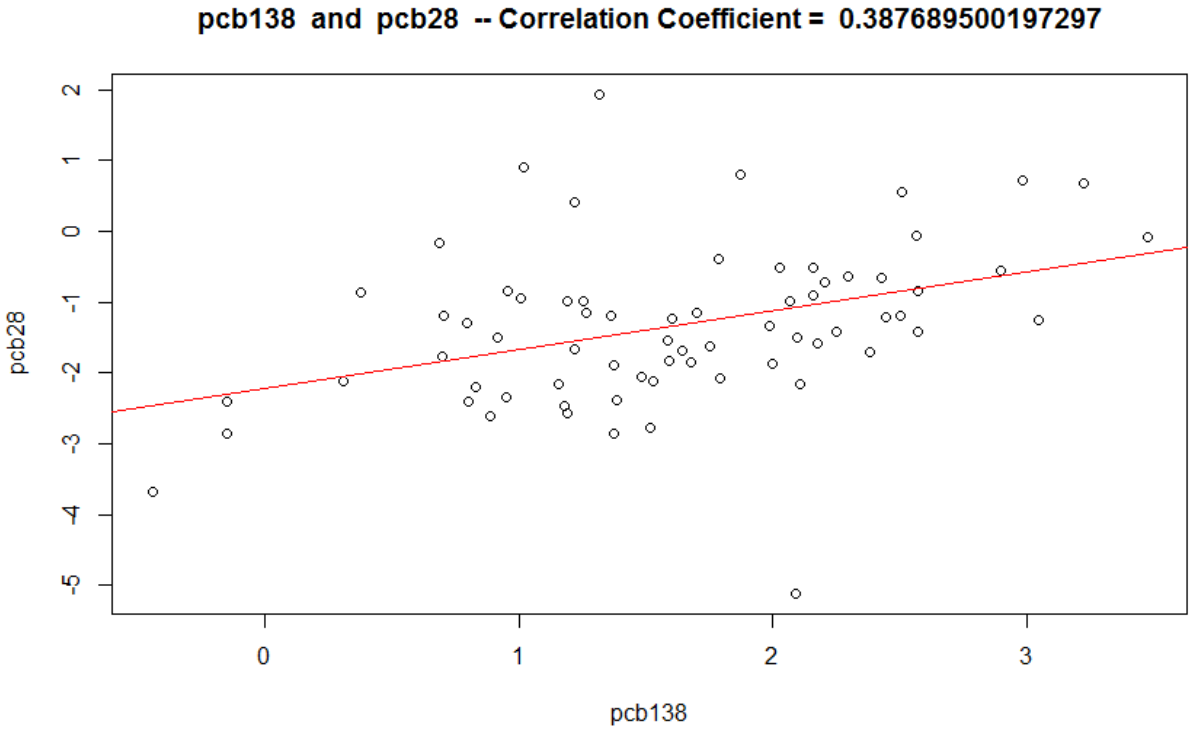

pcb126 and pcb28 -- Correlation Coefficient = 0.272192405217234
pcb28
pcb126
-6.0
-5.5
-5.0
-4.5
-4.0
-3.5
2
1
0
-1
-2
-3
-4
-5

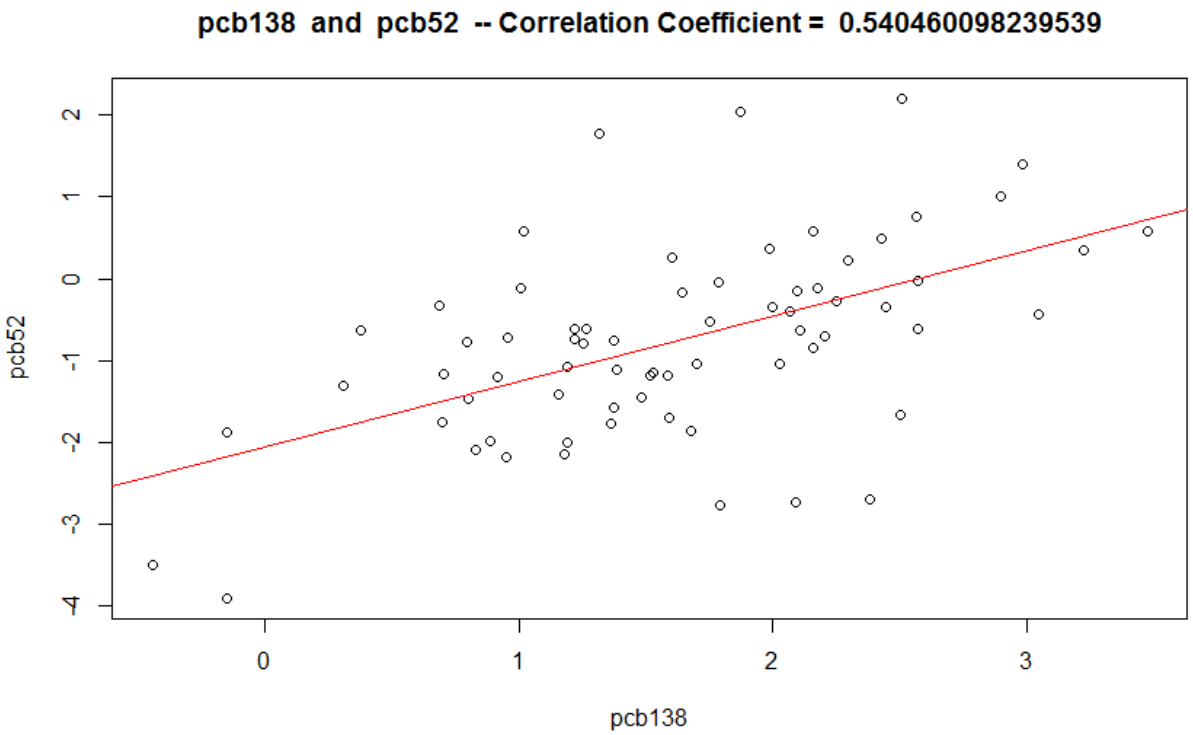

pcb126 and pcb52 -- Correlation Coefficient = 0.330859408692188
pcb52
pcb126
2
1
0
-1
-2
-3
-4
-6.0
-5.5
-5.0
-4.5
-4.0
-3.5

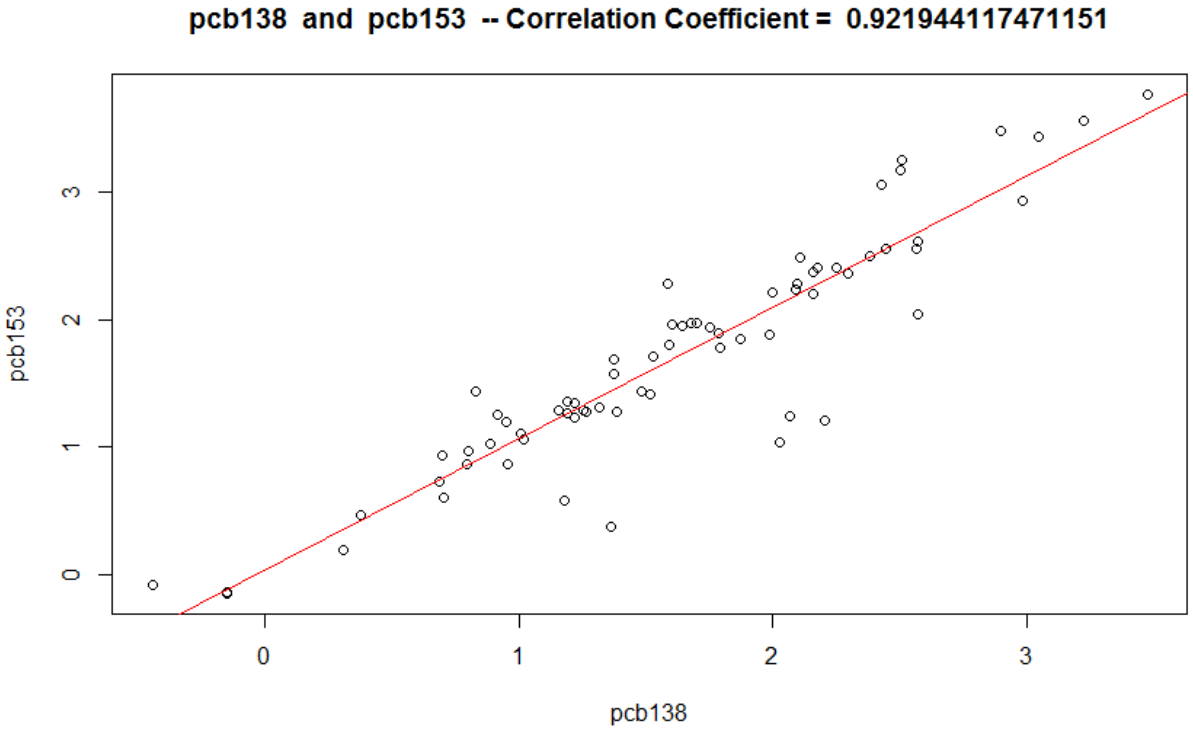

pcb126 and pcb138 -- Correlation Coefficient = 0.792391548200155
pcb138
pcb126
3
2
1
0
-6.0
-5.5
-5.0
-4.5
-4.0
-3.5

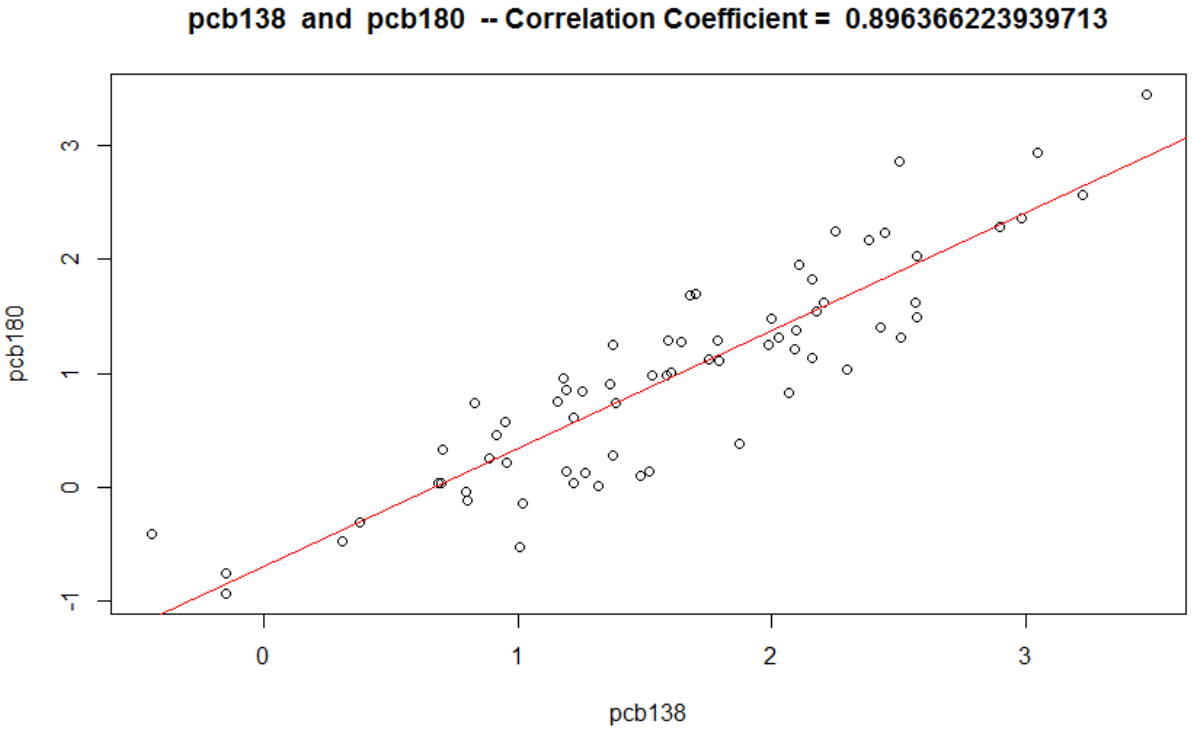

pcb126 and pcb153 -- Correlation Coefficient = 0.646576755148821
pcb153
pcb126
0
1
2
3
-6.0
-5.5
-5.0
-4.5
-4.0
-3.5

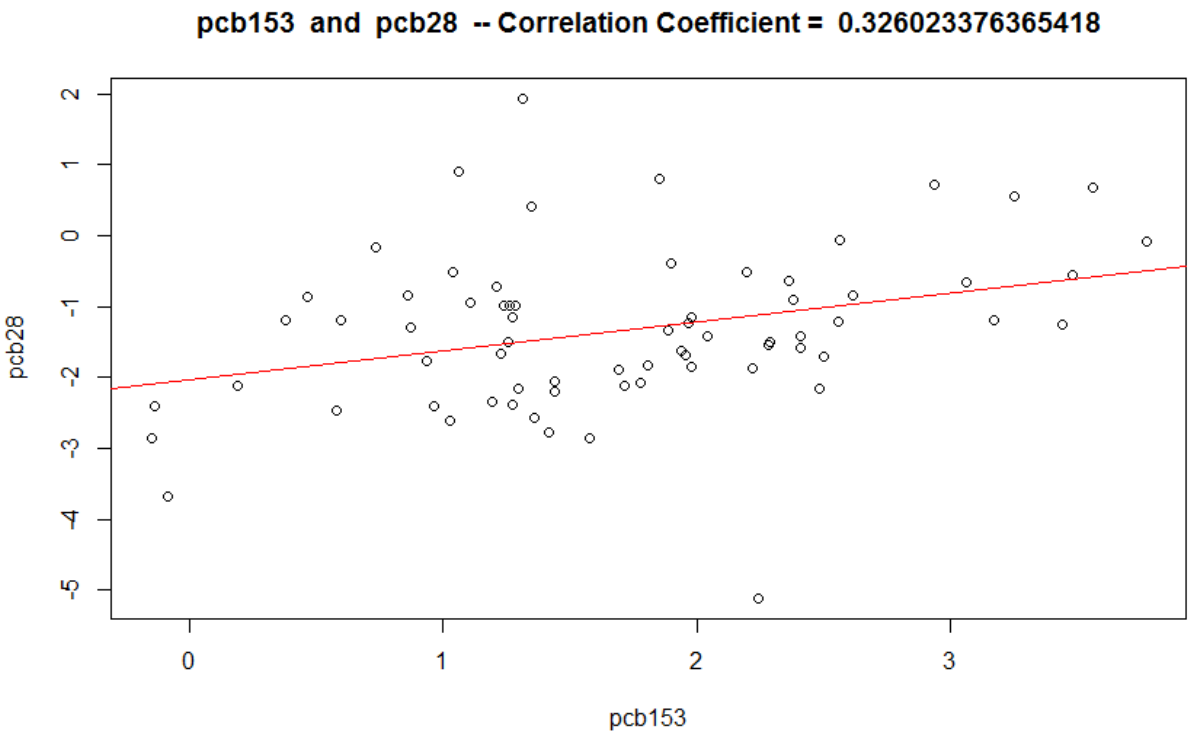

pcb126 and pcb180 -- Correlation Coefficient = 0.6954466310052
pcb180
pcb126
-1
0
1
2
3
-6.0
-5.5
-5.0
-4.5
-4.0
-3.5

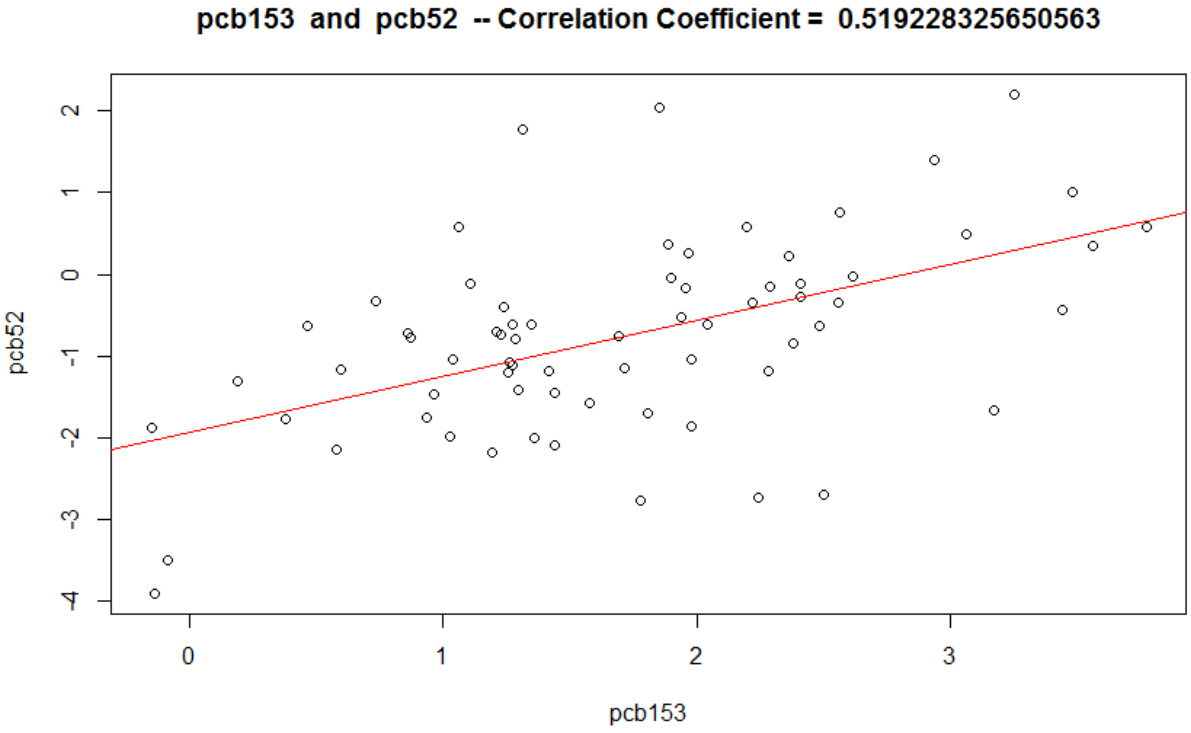

pcb138 and pcb28 -- Correlation Coefficient = 0.387689500197297
pcb28
pcb138
2
1
0
-1
-2
-3
-4
-5
0
1
2
3

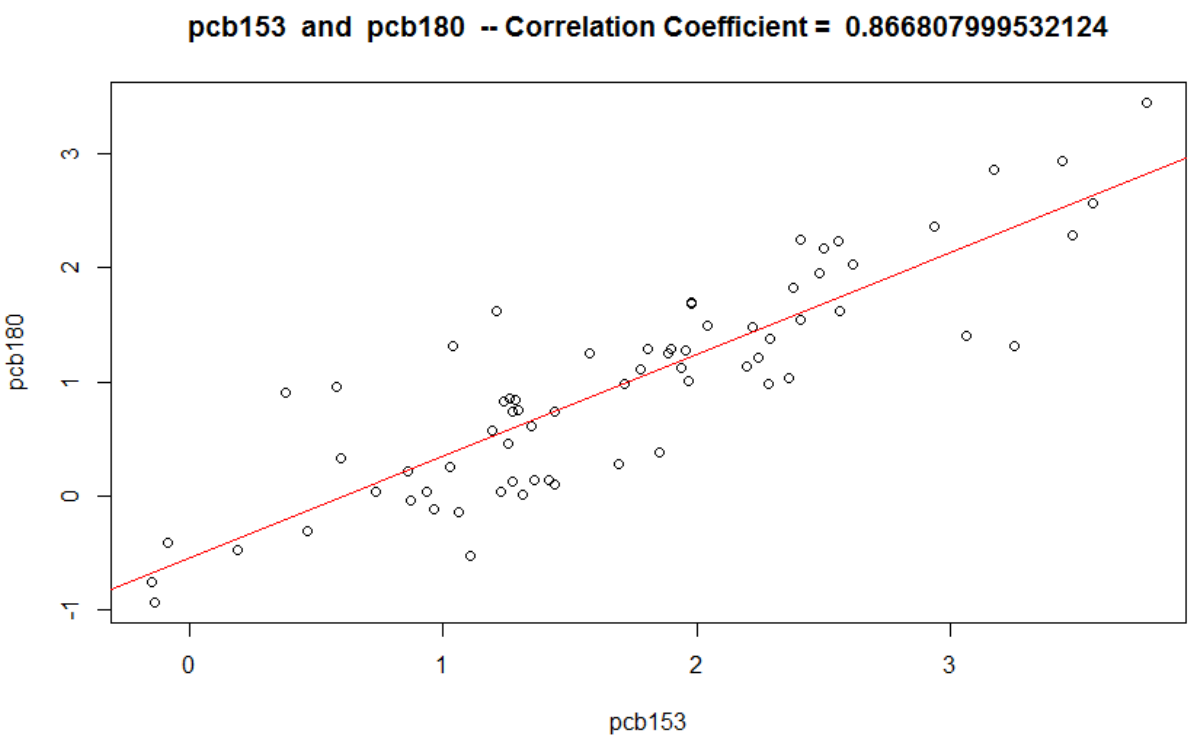

pcb138 and pcb52 -- Correlation Coefficient = 0.540460098239539
pcb52
pcb138
2
1
0
-1
-2
-3
-4
0
1
2
3

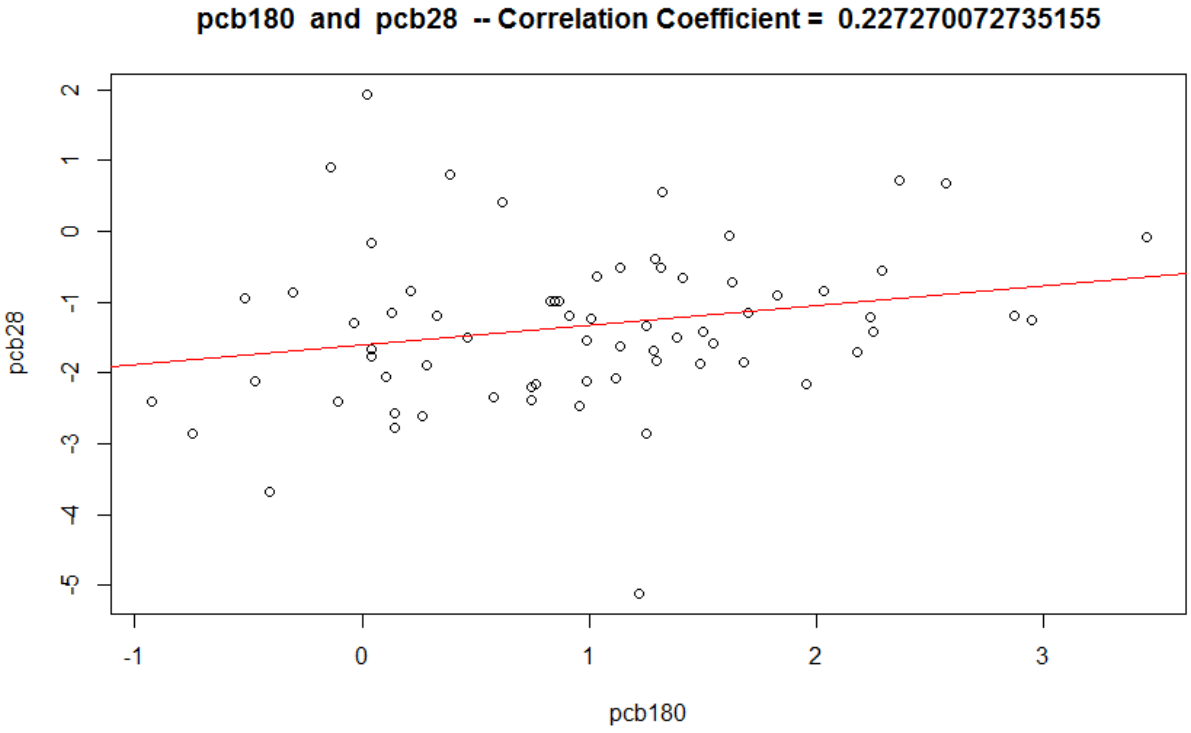

pcb138 and pcb153 -- Correlation Coefficient = 0.921944117471151
pcb153
0
1
2
3
pcb138
0
1
2
3

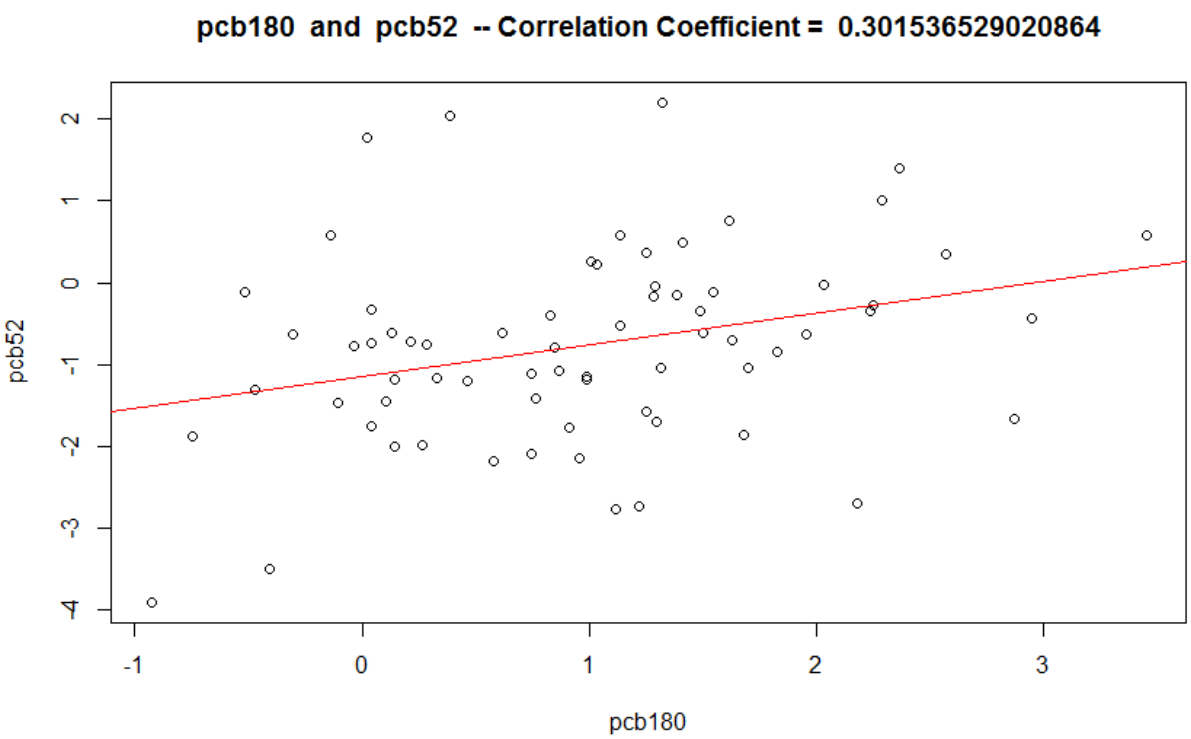

pcb138 and pcb180 -- Correlation Coefficient = 0.896366223939713
pcb180
pcb138
-1
0
1
2
3


pcb153 and pcb28 -- Correlation Coefficient = 0.326023376365418
pcb28
pcb153
-5
-4
-3
-2
-1
0
1
2
3


pcb153 and pcb52 -- Correlation Coefficient = 0.519228325650563
pcb52
pcb153
2
1
0
-1
-2
-3
-4
0
1
2
3


pcb153 and pcb180 -- Correlation Coefficient = 0.866807999532124
pcb180
pcb153
3
2
1
0
-1
0
1
2
3


pcb180 and pcb28 -- Correlation Coefficient = 0.227270072735155
pcb28
pcb180
2
1
0
-1
-2
-3
-4
-5
-1
0
1
2
3


pcb180 and pcb52 -- Correlation Coefficient = 0.301536529020864
pcb52
pcb180
2
1
0
-1
-2
-3
-4
-1
0
1
2
3

2. Compare these summaries with the summaries that you produced in Exercise 11.42 for the measured variables.

The correlations found for the logarithmic comparisons are higher than the correlations found for the measured variables.

50. Use the log data set that you created in Exercise 11.48 to find a good multiple regression model for predicting the log of PCB. Use only log PCB variables for this analysis. Write a report summarizing your results.

```
> summary(fm)

Call:
lm(formula = dslog[, "pcb"] ~ dslog[, "pcb52"] + dslog[, "pcb118"] +
    dslog[, "pcb138"] + dslog[, "pcb180"])

Residuals:
     Min       1Q   Median       3Q      Max
-0.30885 -0.08719 -0.02035  0.04706  0.60278

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.02587    0.09818  30.819  < 2e-16 ***
dslog[, "pcb52"]   0.18415    0.02118   8.693 1.94e-12 ***
dslog[, "pcb118"]  0.18690    0.06950   2.689  0.00912 **
dslog[, "pcb138"]  0.40934    0.11958   3.423  0.00108 **
dslog[, "pcb180"]  0.21816    0.06363   3.428  0.00107 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.148 on 64 degrees of freedom
Multiple R-squared:  0.968,     Adjusted R-squared:  0.966
F-statistic: 483.3 on 4 and 64 DF,  p-value: < 2.2e-16


> stem(res)

  The decimal point is 1 digit(s) to the left of the |

  -3 | 1
  -2 | 1
  -1 | 866554432211000
  -0 | 998777655444433332221111 10
   0 | 002222345666778
   1 | 378
   2 | 013468
   3 | 6
   4 |
   5 |
   6 | 0
```

By taking the multiple regression of several permutations of different PCB congeners, I determined that using LOGpcb52, LOGpcb118, LOGpcb138, and LOGpcb180 yields an R-squared value of 0.968, among the highest of all tested permutations. As this was the same combination used earlier, I decided to stick with it.

As seen in the stem graph of the residuals, the distribution of the residuals is a slightly right-skewed normal distribution.

51. Use the log data set that you created in Exercise 11.48 to find a good multiple regression model for predicting the log of TEQ. Use only log PCB variables for this analysis. Write a report summarizing your results and comparing them with the results you obtained in the previous exercise.

```
> summary(fm)

Call:
lm(formula = dslog[, "teq"] ~ dslog[, "pcb28"] + dslog[, "pcb126"])

Residuals:
     Min       1Q   Median       3Q      Max
-0.55087 -0.17821  0.02989  0.13940  0.96205

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.96348    0.22747  17.424  < 2e-16 ***
dslog[, "pcb28"]   0.10770    0.03246   3.318  0.00148 **
dslog[, "pcb126"]  0.62222    0.04801  12.960  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2916 on 66 degrees of freedom
Multiple R-squared:  0.7681,    Adjusted R-squared:  0.7611
F-statistic: 109.3 on 2 and 66 DF,  p-value: < 2.2e-16
```

```
> stem(res)

  The decimal point is 1 digit(s) to the left of the |

  -4 | 5225555
  -2 | 996643300
  -0 | 98554499773220
   0 | 000333668880001111224445 5
   2 | 012556857
   4 | 1658
   6 |
   8 | 6
```

Using previously selected rows yielded a fairly low R-squared value of < 0.55. By using the rows previously unused (pcb28 and pcb126), I got an R-squared value of 0.7681.

As before, the stem of the residuals is a slightly right-skewed normal distribution.

52. Use the results of your analysis of the log PCB data in Exercise 11.50 to write an explanation of how regression coefficients, Standard errors of regression coefficients, and tests of significance for explanatory variables can change depending on what other explanatory variables are included in the multiple regression analysis.

By cherry-picking datasets with minimal outliers (as seen when we removed two outliers earlier in this assignment), and datasets with distributions similar to the target dataset (when we picked 52, 118, 138, 180, while excluding 28 and 126), we get more accurate multiple regression predictions. These reduce the standard error, and increase the significance of the results we get when predicting datasets.

The most ideal set to use in multiple regression will, when averaged out, yield a similar distribution with a similar range to the target dataset. Conversely, the least ideal set would have a very different distribution and/or many outliers compared to the range of the target set.

These assumptions are fairly intuitive; similar datasets can be extrapolated to predict things in aggregate.

R-CODE FOLLOWS:

```
Adam Gincel

MA331

R Code for Final Project

ds = file.csv("pcb.csv")

#11.42

boxplot(ds[,"pcb"], main="Boxplot for pcb", horizontal=T, axes=F, staplewex=1)

text(x=boxplot.stats(ds[,"pcb"])$stats, labels=boxplot.stats(ds[,"pcb"])$stats,
y=1.25)

mean(ds[,"pcb"])

sd(ds[,"pcb"])

var(ds[,"pcb"])



boxplot(ds[,"pcb52"], main="Boxplot for pcb52", horizontal=T, axes=F, staplewex=1)

text(x=boxplot.stats(ds[,"pcb52"])$stats, labels=boxplot.stats(ds[,"pcb52"])$stats,
y=1.25)

mean(ds[,"pcb52"])

sd(ds[,"pcb52"])

var(ds[,"pcb52"])



boxplot(ds[,"pcb118"], main="Boxplot for pcb118", horizontal=T, axes=F, staplewex=1)
```

```
text(x=boxplot.stats(ds[,"pcb118"])$stats, labels=boxplot.stats(ds[,"pcb118"])$stats,
y=1.25)

print(mean(ds[,"pcb118"]))

print(sd(ds[,"pcb118"]))

print(var(ds[,"pcb118"]))


boxplot(ds[,"pcb138"], main="Boxplot for PCB138", horizontal=T, axes=F, staplewex=1)

text(x=boxplot.stats(ds[,"pcb138"])$stats, labels=boxplot.stats(ds[,"pcb138"])$stats,
y=1.25)

mean(ds[,"pcb138"])

sd(ds[,"pcb138"])

var(ds[,"pcb138"])


boxplot(ds[,"pcb180"], main="Boxplot for pcb180", horizontal=T, axes=F, staplewex=1)

text(x=boxplot.stats(ds[,"pcb180"])$stats, labels=boxplot.stats(ds[,"pcb180"])$stats,
y=1.25)

print(mean(ds[,"pcb180"]))

print(sd(ds[,"pcb180"]))

print(var(ds[,"pcb180"]))


plot(pcb, pcb52, main="PCB and PCB52", xlab="PCB", ylab="PCB52")

abline(lm(pcb52~pcb), col="red")

corrCoeff <- cor(pcb,pcb52)

text(labels = paste("Correlation Coefficient = ", corrCoeff), x=100.25, y=8.25)
```

```
# ^ repeat for all values


#11.43

fm <- lm(pcb ~ pcb52 + pcb118 + pcb138 + pcb180)

fm

summary(fm)


res <- residuals(fm)

res

stem(res)


#11.45

fm <- lm(pcb ~ pcb52 + pcb118 + pcb138)

fm

summary(fm)


res <- residuals(fm)

res

stem(res)


#11.46

fm <- lm(ds[,"teq"] ~ ds[,"teqpcb"] + ds[,"teqdioxin"] + ds[,"teqfuran"])
```

```
fm

summary(fm)



res <- residuals(fm)

res

stem(res)



#11.47

fm <- lm(teq ~ pcb52 + pcb118 + pcb138 + pcb180)

fm

summary(fm)



res <- residuals(fm)

res

stem(res)



#11.48

boxplot(dslog[,"pcb118"], main="Boxplot: PCB118", horizontal=T, axes=F, staplewex=1)

print(summary(dslog[,"pcb118"]))

#repeat for each pcb value



#11.49

for (i in pcblist) {
```

```
  for(j in pcblist) {
    if(i < j)
    {
      corrCoeff <- cor(dslog[,i],dslog[,j])
      plot(dslog[,i], dslog[,j], main=paste(i, " and ", j, " -- Correlation
Coefficient = ", corrCoeff), xlab=i, ylab=j)
      abline(lm(dslog[,j]~dslog[,i]), col="red")
    }
  }
}

#11.50
datasets <- c("pcb28", "pcb52", "pcb118", "pcb126", "pcb138", "pcb153", "pcb180")
permutations <- combn(datasets, 2) # use this to determine best multiple regression
for (i in 1:21)
{
  print(paste(permutations[1,i], permutations[2,i]))
  print(summary(lm(dslog[,"pcb"] ~ dslog[,permutations[1,i]] +
dslog[,permutations[2,i]])))
}
```