Rebecca Lee
STA 2260.04
013919802

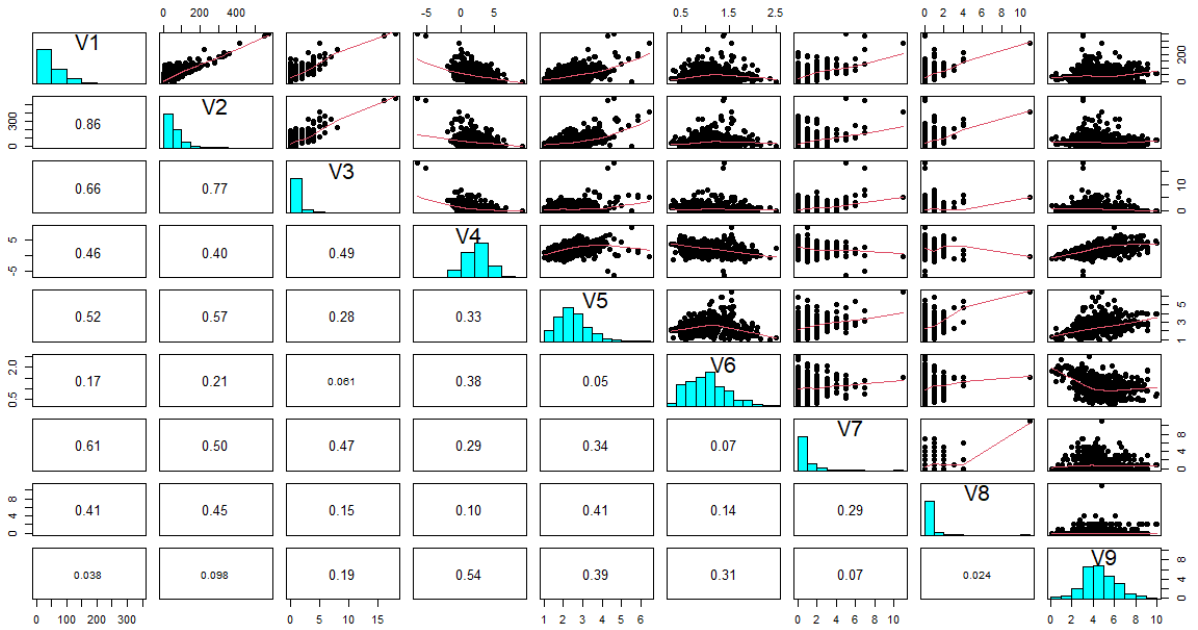
STA2260.04 Project: QSAR Aquatic Toxicity

DATA SET

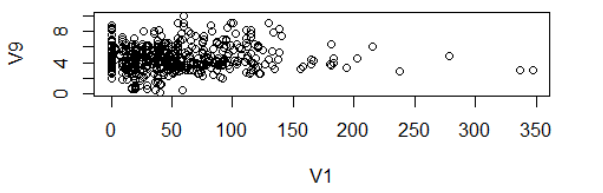
Variable Name	Description	Range
V1	TPSA (Tot); molecular properties	0.00 - 347.32
V2	SAacc; molecular properties	0.00 - 571.95
V3	H-050; atom-centered fragments	0.0000 - 18.0000
V4	MLOGP; molecular properties	-6.446 - 9.148
V5	RDCHI; connectivity indices	1.000 - 6.439
V6	GATS1p; 2D autocorrelations	0.281 - 2.500
V7	nN; constitutional indices	0.000 - 11.000
V8	C-040; atom-centered fragments	0.0000 - 11.0000
V9	LC50; OUTPUT VARIABLE; concentration of chemical that causes 50% of a specific fish in 48 hours	0.122 - 10.047

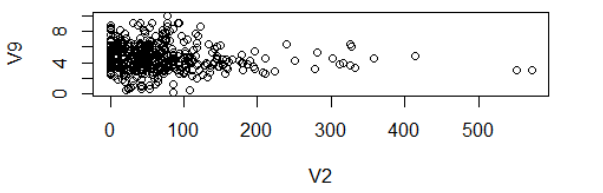
We will use LC50 (V9) as the output variable.
Unfortunately, the amount of information on the data was sparse, and didn't include further descriptions, range, or units. The ranges shown were calculated using R's summary() function.

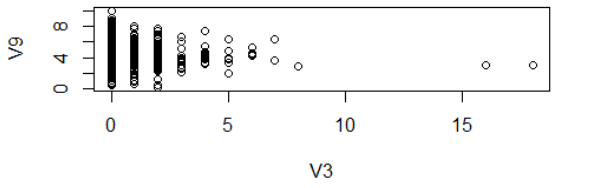
SCATTERPLOT ANALYSIS, DISCUSSION

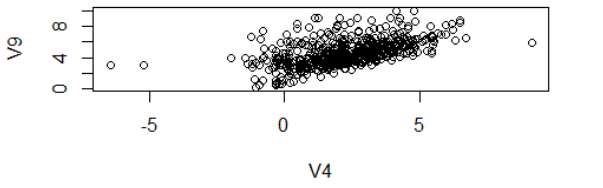


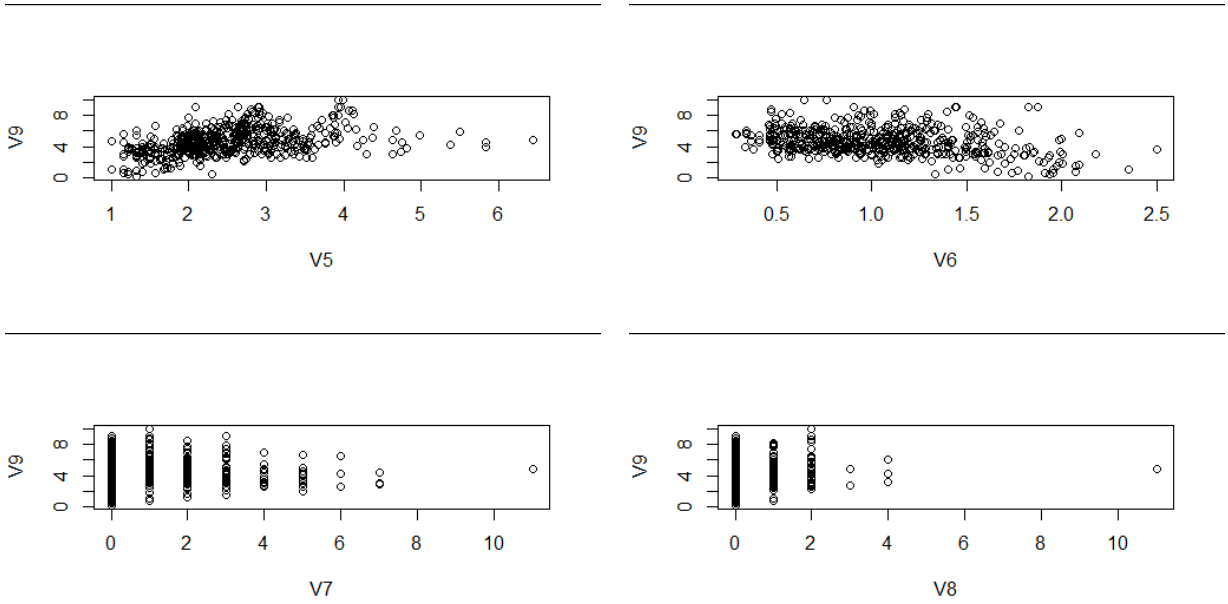
Our output variable, V9, is the last variable, so we will look at the bottom row.
The strongest correlation appears to be V4, MLOGP, at 0.54. This is still a somewhat weak correlation.











The scatterplots for V3, V7, and V8 are discrete, with a visible positive skew. We will not use these variables for M1.
The scatterplots for V1 and V2 have a positive skew, and don't appear to show any patterns. We will not use these variables either. The scatterplot for V5 also has a positive skew, but shows some semblance of a linear relationship. We may investigate this further.
The scatterplots for V4 and V6 most visibly show linear relationships. We will look into both of these variables.

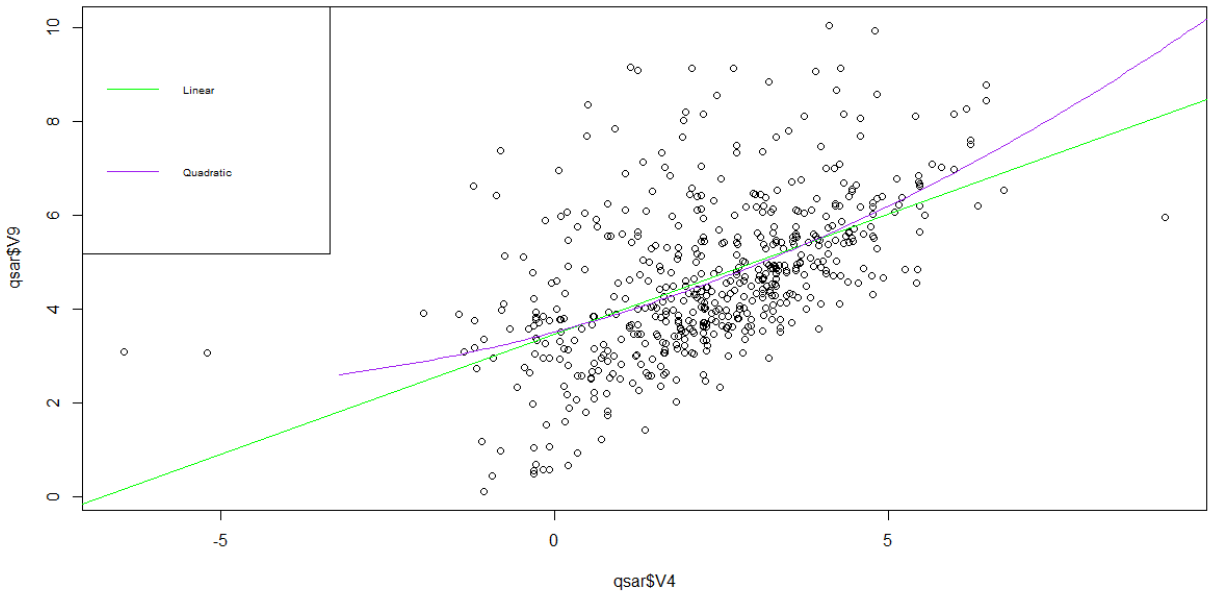
M1

Variable	Linear AIC	Quadratic AIC
V4	1926.827	1920.820
V5	2020.952	2976.263
V6	2066.500	2049.264

V4, MLOGP, has the lowest AIC of all the variables measured. We will compare its linear and quadratic graphs, shown in the table below.

Line Type	R^2	Adjusted R^2	AIC Value
Linear	0.2868	0.2855	1926.827
Quadratic	0.2972	0.2946	1920.820

The quadratic line is better suited for V4 because it has both a higher R^2 and a lower AIC than the linear model.



The equation of the green linear line is 3.47399 + 0.51197[V4]
The equation of the purple quadratic line is 3.51673 + 0.31867[V4] + 0.03087$[V4]^2$.

M2

I created 3 models; M2.1 using the 5 variables with the highest correlation values from the scatterplot analysis section (V2, V3, V5, V6, V7); M2.2 using the 3 variables with the highest correlation values (V3, V5, V6); and M2.3 using scatterplots that had similar shapes (V1, V2, V3, V5).

Model	R^2	Adjusted R^2	AIC Value
M2.1	1.296	0.4033	1843.467
M2.2	1.325	0.3726	1864.803
M2.3	1.32	0.3799	1862.419

M2.2, the 3 variables with highest correlation (V3, V5, V6), has the highest R^2.
M2.3, the scatterplots with similar shapes (V1, V2, V3, V5), has the lowest AIC value. We will use this model.
The formula for M2.3, from the R printout, is
lm(formula = V9 ~ V1 + I(V1^2) + V2 + I(V2^2) + V5 + I(V5^2) + V3, data = qsar)

MODEL COMPARISON

Model	R^2	Adjusted R^2	AIC
M1	0.2972	0.2946	1920.820
M2	1.32	0.3799	1862.419

M2 is technically the better model of the two because it has a higher R^2, adjusted R^2, and lower AIC. This may be due to grouping linked variables, since they were chosen on the basis of shape. Picking multiple graphs with a similar shape would show better results than only one. However, this opens up the possibility that the variables are linked by a hidden/lurking variable (like shark attacks and amount of ice cream sold).