ExploratoryAnalyses

MRB

October 27, 2017

Purpose

Document for data exploration for **Data Challenge 2017**!

Look at the data

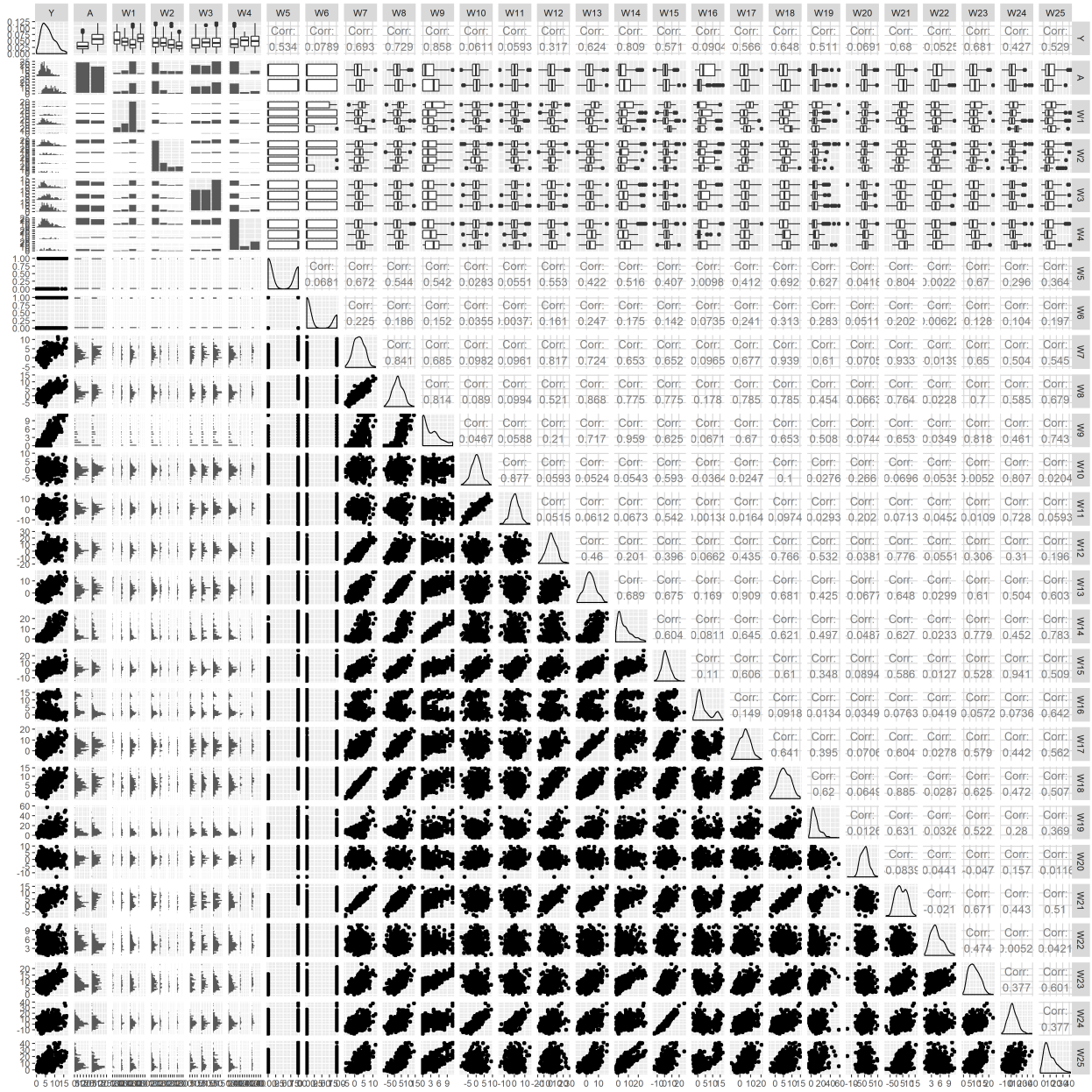
Table 1: **Overall Data Summary**

	Overall (N=400)
Y	
Mean (SD)	6.26 (3.47)
Q1, Q3	3.62, 8.43
Range	0.042 - 16.6
A	
0	215 (53.8%)
1	185 (46.2%)
W1	
A	42 (10.5%)
B	75 (18.8%)
C	263 (65.8%)
D	20 (5%)
W2	
A	252 (63%)
B	71 (17.8%)
C	37 (9.25%)
D	40 (10%)
W3	
A	115 (28.8%)
B	115 (28.8%)
C	170 (42.5%)
W4	
A	289 (72.2%)
B	34 (8.5%)
C	77 (19.2%)
W5	
0	232 (58%)
1	168 (42%)
W6	
0	279 (69.8%)
1	121 (30.2%)
W7	
Mean (SD)	2.29 (2.99)
Q1, Q3	0.028, 4.33
Range	-5.2 - 11.5

	Overall (N=400)
W8	
Mean (SD)	2.36 (3.68)
Q1, Q3	-0.22, 4.87
Range	-7.7 - 14.3
W9	
Mean (SD)	2.93 (3.14)
Q1, Q3	0, 5
Range	0 - 11
W10	
Mean (SD)	0.089 (3.03)
Q1, Q3	-1.9, 2.02
Range	-9.1 - 9.27
W11	
Mean (SD)	0.307 (4.47)
Q1, Q3	-2.6, 3.11
Range	-14 - 14.9
W12	
Mean (SD)	1.55 (7.67)
Q1, Q3	-3.4, 6.65
Range	-19 - 28.1
W13	
Mean (SD)	3.02 (4.67)
Q1, Q3	-0.152, 5.93
Range	-7.7 - 17.1
W14	
Mean (SD)	5.68 (6.6)
Q1, Q3	0.478, 9.04
Range	-2.7 - 27.4
W15	
Mean (SD)	2.01 (5.69)
Q1, Q3	-1.7, 5.29
Range	-14 - 27.2
W16	
Mean (SD)	3.8 (5)
Q1, Q3	0.297, 6.11
Range	-3.4 - 16.9
W17	
Mean (SD)	4.66 (5.14)
Q1, Q3	0.817, 8.17
Range	-7.9 - 20.3
W18	
Mean (SD)	5.54 (3.61)
Q1, Q3	2.92, 8.32
Range	-3 - 14.9
W19	
Mean (SD)	9.09 (9.26)
Q1, Q3	2.67, 12.6
Range	-5.2 - 57.6
W20	
Mean (SD)	-0.079 (3.11)
Q1, Q3	-2.4, 1.86
Range	-13 - 9.71

	Overall (N=400)
W21	
Mean (SD)	3.79 (4.31)
Q1, Q3	0.321, 7.3
Range	-8.4 - 15.6
W22	
Mean (SD)	4.73 (2.01)
Q1, Q3	3.21, 6.03
Range	0.631 - 10.9
W23	
Mean (SD)	8.5 (4.26)
Q1, Q3	5.14, 11.5
Range	-0.232 - 23.5
W24	
Mean (SD)	2.89 (8.34)
Q1, Q3	-2.9, 8.1
Range	-17 - 38
W25	
Mean (SD)	9.81 (9.21)
Q1, Q3	2.86, 15.4
Range	-4.9 - 41.5

Scatterplot Matrix without Transformation



Scatterplot Matrix with log transform on response variable.

Table 2: **Model Summaries: Y = W_i + A, for i = 1, 2, . . . , 25**

	estimate	std.error	p.value	adj.r.squared
(Intercept)	6.58	0.469	<0.001	0.324
W1 B	-1.4	0.55	0.009	.
W1 C	-2.5	0.475	<0.001	.
W1 D	0.736	0.777	0.344	.
A 1	3.4	0.287	<0.001	.
(Intercept)	4.59	0.25	<0.001	0.242
W2 B	0.247	0.406	0.542	.
W2 C	0.634	0.541	0.242	.

	estimate	std.error	p.value	adj.r.squared
W2 D	-0.342	0.521	0.512	.
A 1	3.46	0.312	<0.001	.
(Intercept)	4.39	0.312	<0.001	0.242
W3 B	0.327	0.398	0.412	.
W3 C	0.436	0.365	0.232	.
A 1	3.43	0.303	<0.001	.
(Intercept)	4.42	0.222	<0.001	0.257
W4 B	0.112	0.546	0.838	.
W4 C	1.2	0.383	0.002	.
A 1	3.46	0.302	<0.001	.
(Intercept)	3.66	0.198	<0.001	0.433
W5 1	3.14	0.271	<0.001	.
A 1	2.76	0.268	<0.001	.
(Intercept)	4.53	0.226	<0.001	0.245
W6 1	0.471	0.328	0.152	.
A 1	3.42	0.302	<0.001	.
(Intercept)	3.34	0.159	<0.001	0.625
W7	0.727	0.036	<0.001	.
A 1	2.7	0.216	<0.001	.
(Intercept)	3.45	0.141	<0.001	0.692
W8	0.636	0.026	<0.001	.
A 1	2.83	0.194	<0.001	.
(Intercept)	2.69	0.111	<0.001	0.832
W9	0.869	0.023	<0.001	.
A 1	2.22	0.146	<0.001	.
(Intercept)	4.67	0.206	<0.001	0.242
W10	0.046	0.05	0.354	.
A 1	3.42	0.303	<0.001	.
(Intercept)	4.65	0.206	<0.001	0.245
W11	0.05	0.034	0.142	.
A 1	3.44	0.302	<0.001	.
(Intercept)	4.52	0.196	<0.001	0.324
W12	0.13	0.019	<0.001	.
A 1	3.31	0.286	<0.001	.
(Intercept)	3.58	0.167	<0.001	0.57
W13	0.427	0.024	<0.001	.
A 1	2.99	0.229	<0.001	.
(Intercept)	2.91	0.126	<0.001	0.775
W14	0.39	0.013	<0.001	.
A 1	2.46	0.168	<0.001	.
(Intercept)	4.23	0.168	<0.001	0.509
W15	0.317	0.021	<0.001	.
A 1	3.01	0.245	<0.001	.
(Intercept)	4.21	0.274	<0.001	0.253
W16	0.082	0.033	0.012	.
A 1	3.75	0.326	<0.001	.
(Intercept)	3.2	0.193	<0.001	0.51
W17	0.351	0.024	<0.001	.
A 1	3.06	0.245	<0.001	.
(Intercept)	1.86	0.221	<0.001	0.575
W18	0.56	0.032	<0.001	.
A 1	2.8	0.229	<0.001	.

	estimate	std.error	p.value	adj.r.squared
(Intercept)	3.41	0.21	<0.001	0.429
W19	0.164	0.014	<0.001	.
A 1	2.91	0.267	<0.001	.
(Intercept)	4.67	0.206	<0.001	0.243
W20	-0.058	0.049	0.231	.
A 1	3.42	0.303	<0.001	.
(Intercept)	3.28	0.171	<0.001	0.581
W21	0.481	0.027	<0.001	.
A 1	2.48	0.231	<0.001	.
(Intercept)	4.31	0.439	<0.001	0.242
W22	0.07	0.076	0.361	.
A 1	3.49	0.308	<0.001	.
(Intercept)	0.538	0.245	0.029	0.635
W23	0.514	0.025	<0.001	.
A 1	2.92	0.211	<0.001	.
(Intercept)	4.32	0.188	<0.001	0.389
W24	0.16	0.016	<0.001	.
A 1	3.19	0.273	<0.001	.
(Intercept)	2.21	0.199	<0.001	0.597
W25	0.226	0.012	<0.001	.
A 1	3.95	0.222	<0.001	.

Playtime!

```
## Call:
## rpart(formula = formulize("Y", c(names(dc1)[4:28], "A")), data = dc1)
##   n= 400
##
##           CP nsplit rel error    xerror       xstd
## 1 0.54543153      0 1.0000000 1.0057679 0.06915336
## 2 0.09872680      1 0.4545685 0.4581356 0.03272062
## 3 0.05146478      2 0.3558417 0.3906314 0.02544897
## 4 0.04592446      3 0.3043769 0.3801340 0.02430105
## 5 0.02691989      4 0.2584524 0.3289846 0.02095169
## 6 0.01774951      5 0.2315325 0.2971572 0.02071810
## 7 0.01617286      6 0.2137830 0.2816212 0.01994626
## 8 0.01610594      7 0.1976102 0.2760416 0.01986078
## 9 0.01000000      8 0.1815042 0.2559134 0.01820837
##
## Variable importance
##  W9 W14  W8 W23 W25 W13   A W16 W21  W7 W12 W17 W18 W19
##  24  18  13  13  11  10   3   1   1   1   1   1   1   1
##
## Node number 1: 400 observations,    complexity param=0.5454315
##   mean=6.255473, MSE=11.98608
##   left son=2 (244 obs) right son=3 (156 obs)
##   Primary splits:
##       W9  < 3.5        to the left,  improve=0.5454315, (0 missing)
##       W14 < 5.763803   to the left,  improve=0.5181421, (0 missing)
##       W8  < 2.992449   to the left,  improve=0.4821869, (0 missing)
```

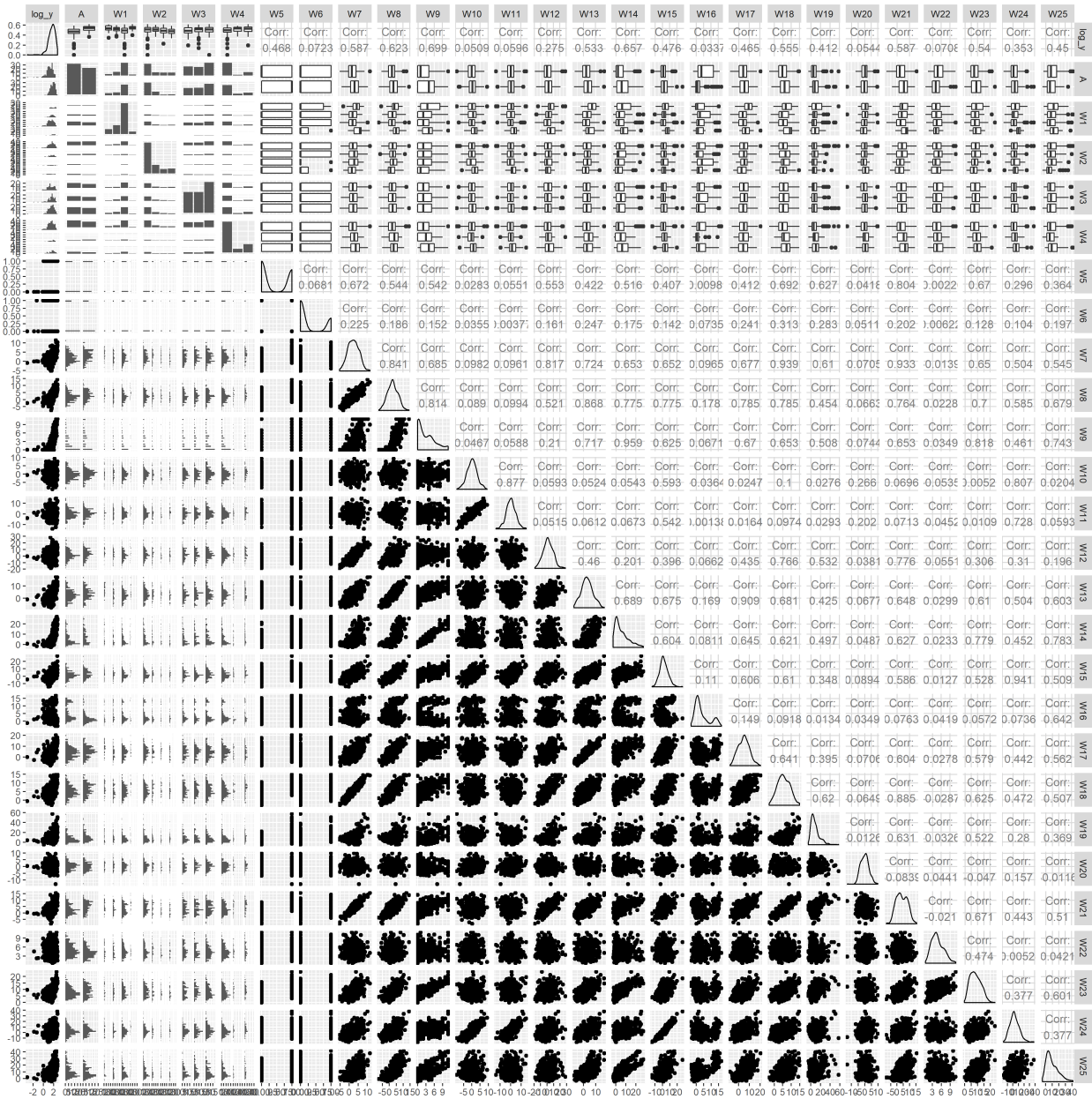
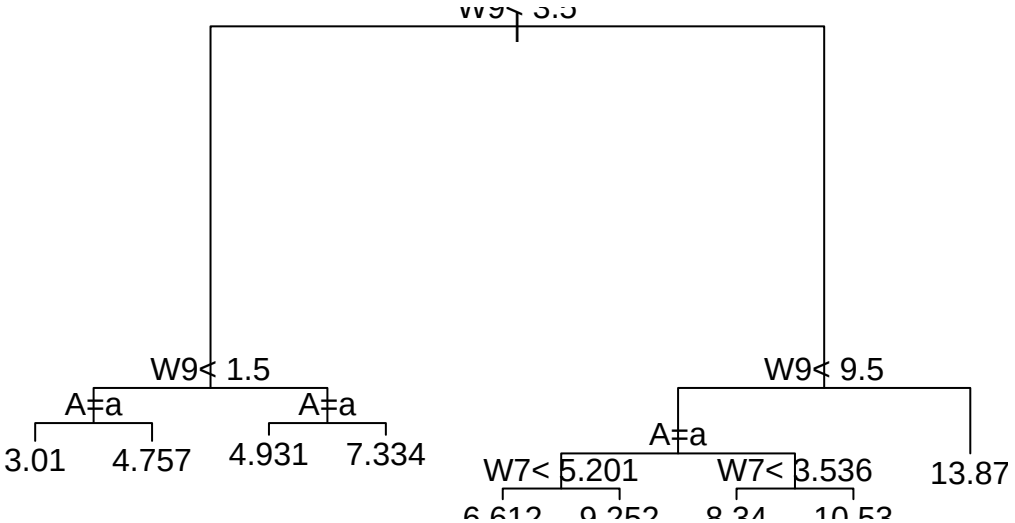


Figure 1: knit to html to see

```
##       W7  < 3.478499   to the left,  improve=0.4175468, (0 missing)
##       W21 < 4.90963    to the left,  improve=0.4038650, (0 missing)
##   Surrogate splits:
##       W14 < 5.262932   to the left,  agree=0.940, adj=0.846, (0 split)
##       W8  < 2.92829    to the left,  agree=0.872, adj=0.673, (0 split)
##       W23 < 9.745147   to the left,  agree=0.855, adj=0.628, (0 split)
##       W25 < 11.65137   to the left,  agree=0.805, adj=0.500, (0 split)
##       W13 < 3.613826   to the left,  agree=0.802, adj=0.494, (0 split)
##
## Node number 2: 244 observations,    complexity param=0.05146478
##   mean=4.211024, MSE=3.867181
##   left son=4 (184 obs) right son=5 (60 obs)
##   Primary splits:
##       W9  < 1.5        to the left,  improve=0.2614948, (0 missing)
##       A   splits as  LR, improve=0.2540243, (0 missing)
##       W7  < 1.591133   to the left,  improve=0.2526917, (0 missing)
##       W21 < 1.663075   to the left,  improve=0.2055047, (0 missing)
##       W8  < 2.017509   to the left,  improve=0.1808941, (0 missing)
##   Surrogate splits:
##       W14 < 2.53977    to the left,  agree=0.893, adj=0.567, (0 split)
##       W8  < 2.903125   to the left,  agree=0.836, adj=0.333, (0 split)
##       W25 < 13.25427   to the left,  agree=0.803, adj=0.200, (0 split)
##       W16 < 11.3988    to the left,  agree=0.799, adj=0.183, (0 split)
##       W13 < 4.309836   to the left,  agree=0.791, adj=0.150, (0 split)
##
## Node number 3: 156 observations,    complexity param=0.0987268
##   mean=9.4532, MSE=7.921833
##   left son=6 (135 obs) right son=7 (21 obs)
##   Primary splits:
##       W9  < 9.5        to the left,  improve=0.3830205, (0 missing)
##       W7  < 4.999441   to the left,  improve=0.3164421, (0 missing)
##       W14 < 13.03272   to the left,  improve=0.3133315, (0 missing)
##       A   splits as  LR, improve=0.2846859, (0 missing)
##       W21 < 4.90963    to the left,  improve=0.2515501, (0 missing)
##   Surrogate splits:
##       W14 < 20.21393   to the left,  agree=0.936, adj=0.524, (0 split)
##       W23 < 15.62551   to the left,  agree=0.910, adj=0.333, (0 split)
##       W11 < 10.56538   to the left,  agree=0.885, adj=0.143, (0 split)
##       W17 < 15.30834   to the left,  agree=0.885, adj=0.143, (0 split)
##       W13 < 13.02812   to the left,  agree=0.878, adj=0.095, (0 split)
##
## Node number 4: 184 observations,    complexity param=0.02691989
##   mean=3.636782, MSE=2.5425
##   left son=8 (118 obs) right son=9 (66 obs)
##   Primary splits:
##       A   splits as  LR, improve=0.27588740, (0 missing)
##       W7  < 1.39736    to the left,  improve=0.13836180, (0 missing)
##       W21 < 3.076161   to the left,  improve=0.10365300, (0 missing)
##       W12 < 1.487588   to the left,  improve=0.08995449, (0 missing)
##       W1  splits as  RRLR, improve=0.08627216, (0 missing)
##   Surrogate splits:
##       W16 < 0.7283739  to the right, agree=0.674, adj=0.091, (0 split)
##       W22 < 1.713776   to the right, agree=0.674, adj=0.091, (0 split)
##       W23 < 2.328696   to the right, agree=0.674, adj=0.091, (0 split)
```

```
##       W8  < -5.151453  to the right, agree=0.663, adj=0.061, (0 split)
##       W25 < -2.560438  to the right, agree=0.663, adj=0.061, (0 split)
##
## Node number 5: 60 observations,    complexity param=0.01774951
##   mean=5.972035, MSE=3.817129
##   left son=10 (34 obs) right son=11 (26 obs)
##   Primary splits:
##       A   splits as  LR, improve=0.3715656, (0 missing)
##       W7  < 1.590959   to the left,  improve=0.3063926, (0 missing)
##       W21 < 2.455516   to the left,  improve=0.2793826, (0 missing)
##       W12 < 0.9629807  to the left,  improve=0.2108504, (0 missing)
##       W8  < 1.879978   to the left,  improve=0.1933822, (0 missing)
##   Surrogate splits:
##       W25 < 4.990111   to the right, agree=0.800, adj=0.538, (0 split)
##       W16 < 1.168843   to the right, agree=0.783, adj=0.500, (0 split)
##       W17 < 5.369905   to the right, agree=0.733, adj=0.385, (0 split)
##       W22 < 3.75348    to the right, agree=0.683, adj=0.269, (0 split)
##       W2  splits as  LRLL, agree=0.667, adj=0.231, (0 split)
##
## Node number 6: 135 observations,    complexity param=0.04592446
##   mean=8.766184, MSE=5.145359
##   left son=12 (58 obs) right son=13 (77 obs)
##   Primary splits:
##       A   splits as  LR, improve=0.3169800, (0 missing)
##       W7  < 4.999441   to the left,  improve=0.2928541, (0 missing)
##       W21 < 4.90963    to the left,  improve=0.2803883, (0 missing)
##       W12 < 1.297934   to the left,  improve=0.2278584, (0 missing)
##       W9  < 5.5        to the left,  improve=0.2232378, (0 missing)
##   Surrogate splits:
##       W16 < 7.491763   to the right, agree=0.756, adj=0.431, (0 split)
##       W25 < 16.86896   to the right, agree=0.756, adj=0.431, (0 split)
##       W21 < 3.170922   to the left,  agree=0.681, adj=0.259, (0 split)
##       W19 < 4.571226   to the left,  agree=0.659, adj=0.207, (0 split)
##       W12 < 0.1056728  to the left,  agree=0.652, adj=0.190, (0 split)
##
## Node number 7: 21 observations
##   mean=13.86973, MSE=3.23064
##
## Node number 8: 118 observations
##   mean=3.010417, MSE=1.935495
##
## Node number 9: 66 observations
##   mean=4.756646, MSE=1.672211
##
## Node number 10: 34 observations
##   mean=4.930597, MSE=2.456896
##
## Node number 11: 26 observations
##   mean=7.333916, MSE=2.322863
##
## Node number 12: 58 observations,    complexity param=0.01617286
##   mean=7.294701, MSE=3.41116
##   left son=24 (43 obs) right son=25 (15 obs)
##   Primary splits:
```

```
##       W7  < 5.200705   to the left,  improve=0.3919170, (0 missing)
##       W14 < 15.69639   to the left,  improve=0.3720128, (0 missing)
##       W9  < 7.5        to the left,  improve=0.3720128, (0 missing)
##       W18 < 8.31034    to the left,  improve=0.3595877, (0 missing)
##       W19 < 12.06569   to the left,  improve=0.3563726, (0 missing)
##   Surrogate splits:
##       W18 < 9.142641   to the left,  agree=0.948, adj=0.800, (0 split)
##       W21 < 8.249895   to the left,  agree=0.948, adj=0.800, (0 split)
##       W19 < 15.61409   to the left,  agree=0.931, adj=0.733, (0 split)
##       W12 < 8.316907   to the left,  agree=0.897, adj=0.600, (0 split)
##       W9  < 7.5        to the left,  agree=0.862, adj=0.467, (0 split)
##
## Node number 13: 77 observations,    complexity param=0.01610594
##   mean=9.874574, MSE=3.592136
##   left son=26 (23 obs) right son=27 (54 obs)
##   Primary splits:
##       W7  < 3.536022   to the left,  improve=0.2791772, (0 missing)
##       W9  < 5.5        to the left,  improve=0.2493821, (0 missing)
##       W14 < 12.33164   to the left,  improve=0.2082537, (0 missing)
##       W21 < 4.786183   to the left,  improve=0.1811454, (0 missing)
##       W8  < 4.274089   to the left,  improve=0.1773352, (0 missing)
##   Surrogate splits:
##       W18 < 6.649654   to the left,  agree=0.896, adj=0.652, (0 split)
##       W8  < 3.479762   to the left,  agree=0.857, adj=0.522, (0 split)
##       W12 < -0.2520162 to the left,  agree=0.844, adj=0.478, (0 split)
##       W21 < 4.786183   to the left,  agree=0.844, adj=0.478, (0 split)
##       W17 < 5.275259   to the left,  agree=0.792, adj=0.304, (0 split)
##
## Node number 24: 43 observations
##   mean=6.611798, MSE=1.850068
##
## Node number 25: 15 observations
##   mean=9.252359, MSE=2.716976
##
## Node number 26: 23 observations
##   mean=8.340136, MSE=2.780899
##
## Node number 27: 54 observations
##   mean=10.52813, MSE=2.507683
```



From the scatterplot matrix there appears to be several variables that are very strongly correlated. Techniques to consider: lasso, pls, others?

Of the variables that are highly correlated, can select the variable(s) that appear to be most correlated with the response

Must not lose focus. The objective is to measure the treatment effect of A, not predict Y.