MACHINE LEARNING FOR BUSINESS HW2 –GROUP8

Nehal Taya

Derek Alexander Rice-Porter

Milan Gabriella Caggiano

OUTLINE

Visualization

Variance Inflation Factor(VIF)

Neural network – Binary Output

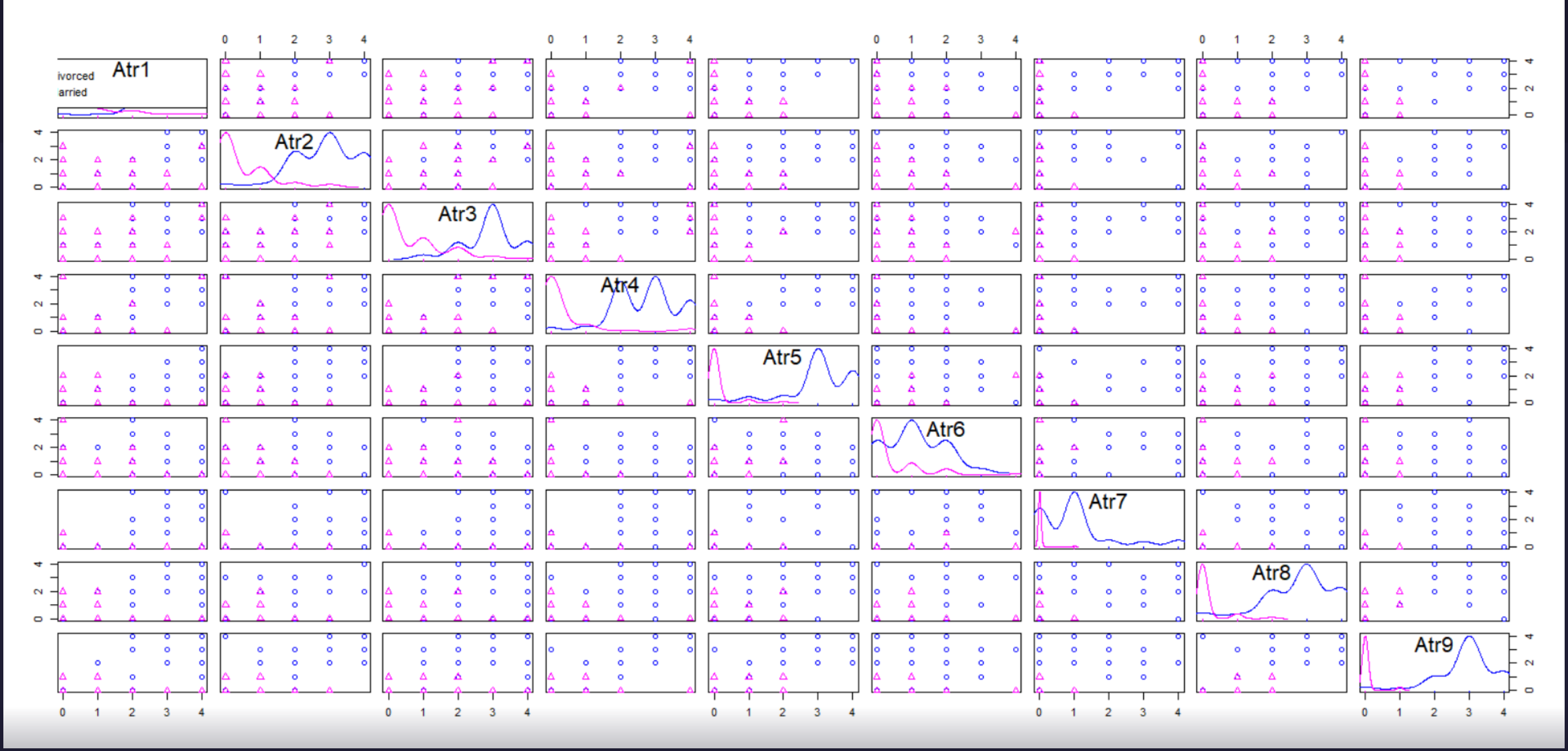
Lessons Learned

DIVORCE DATASET OVERVIEW

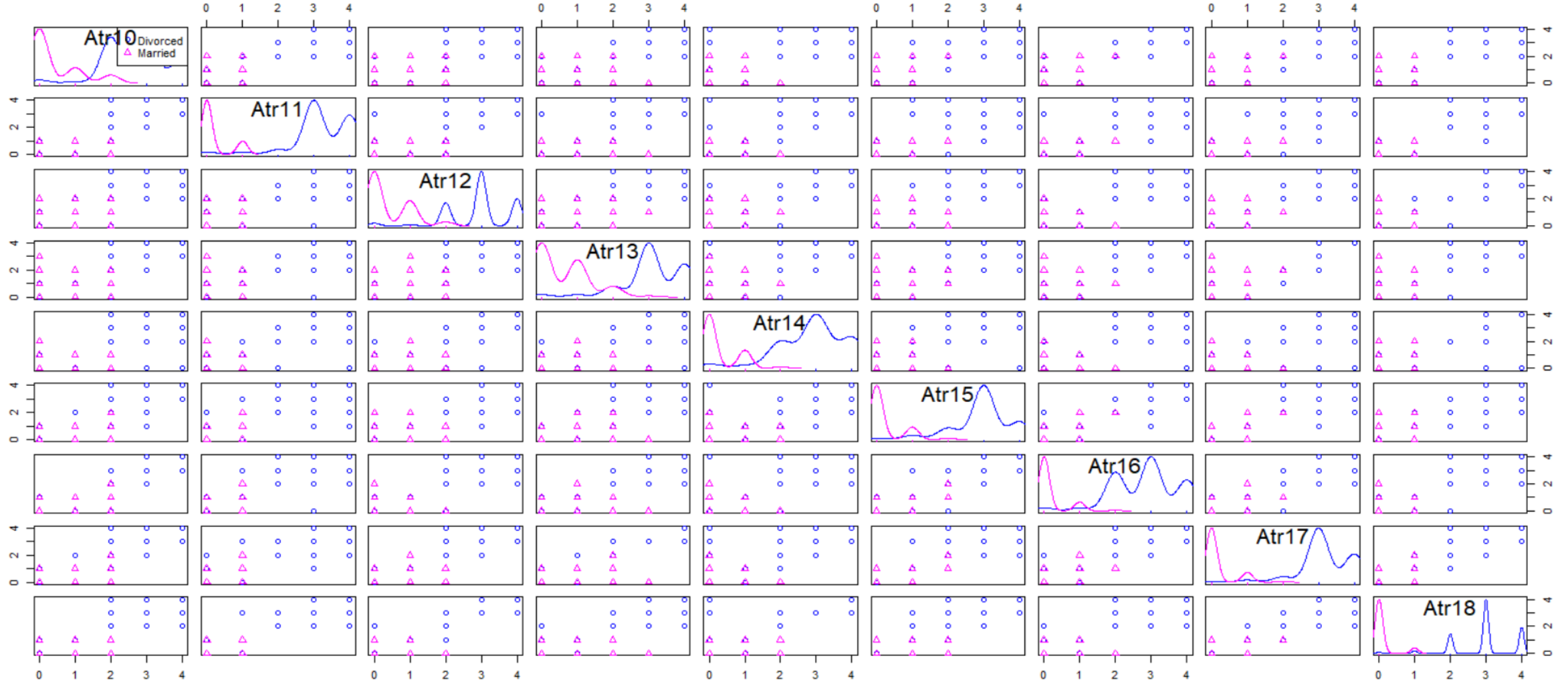
- 170 rows and 56 columns
- Dependent variable – Class
- 0 for married and 1 for divorced
- Independent Variables – 54 Attributes

VISUALIZATIONS

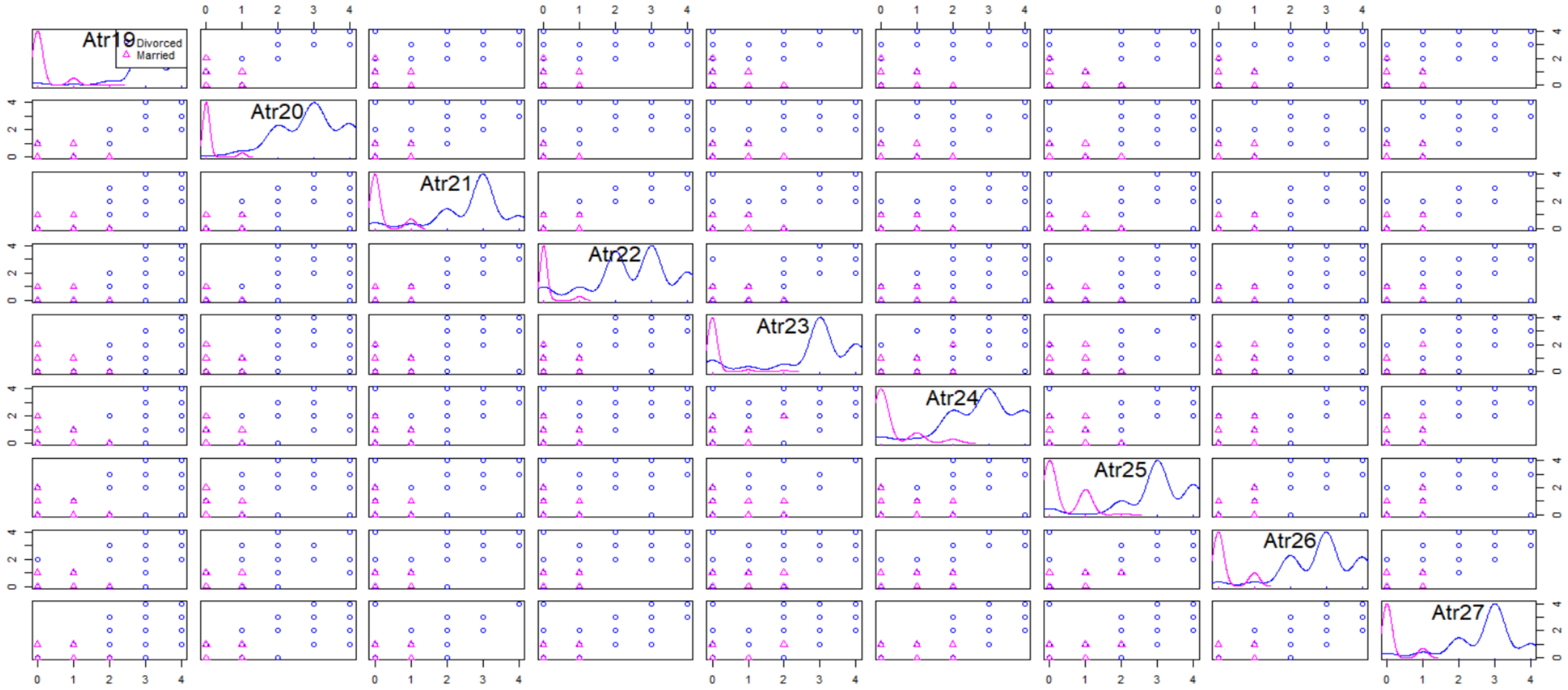
SCATTERPLOT MATRIX OF DIVORCE V/S ATTRIBUTES 1-9



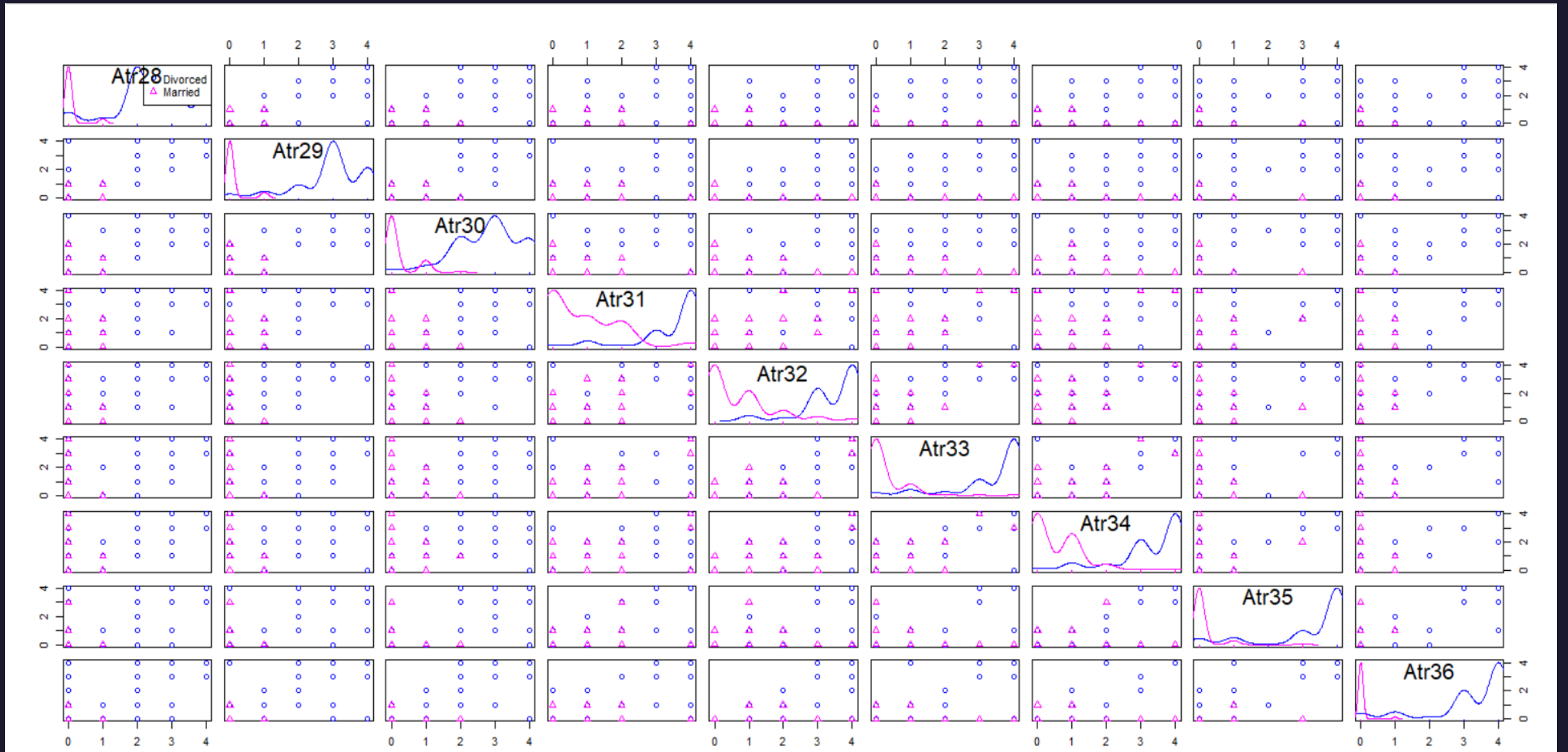
SCATTERPLOT MATRIX OF DIVORCE V/S ATTRIBUTES 10-18



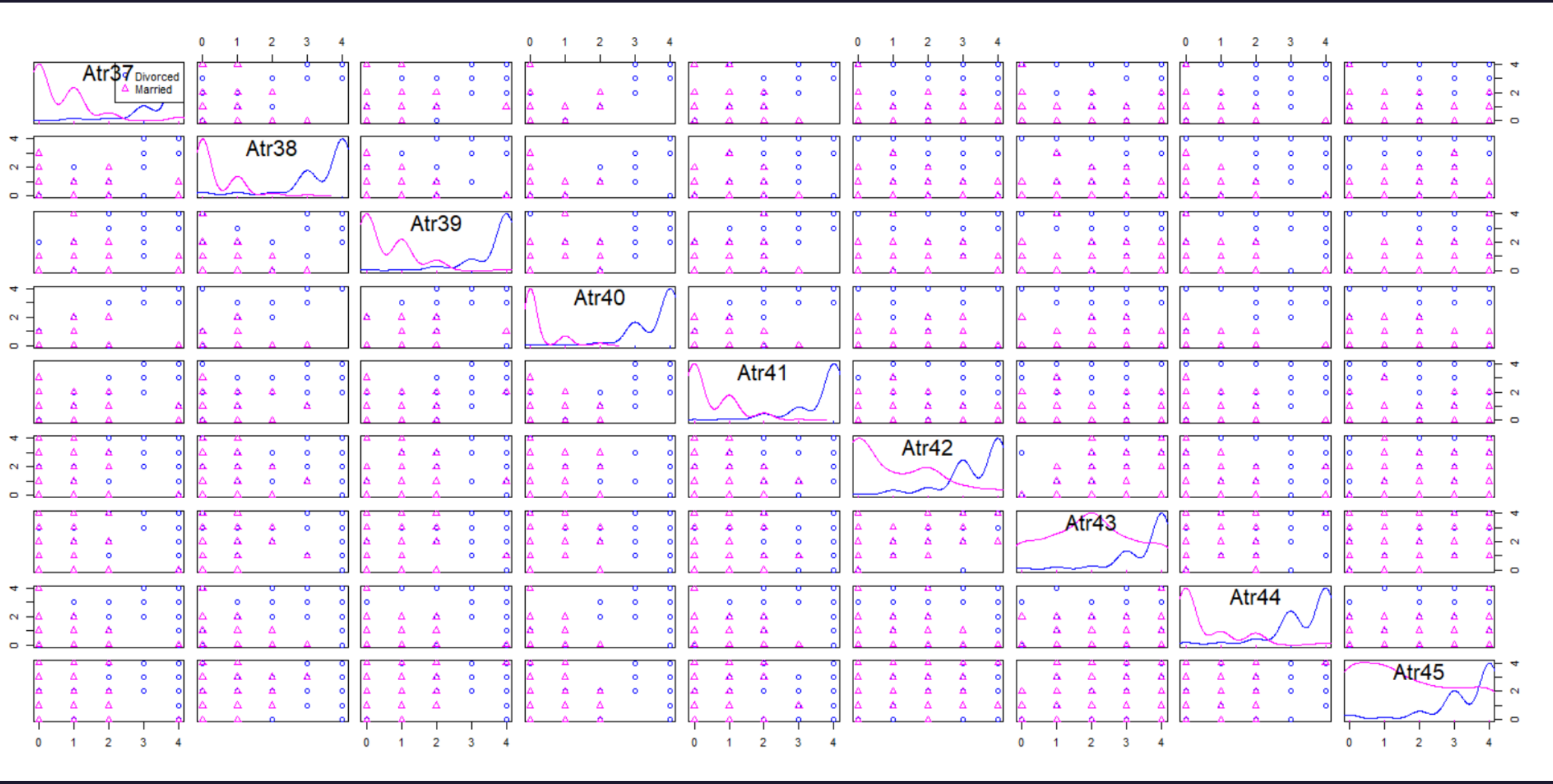
SCATTERPLOT MATRIX OF DIVORCE V/S ATTRIBUTES 19-27



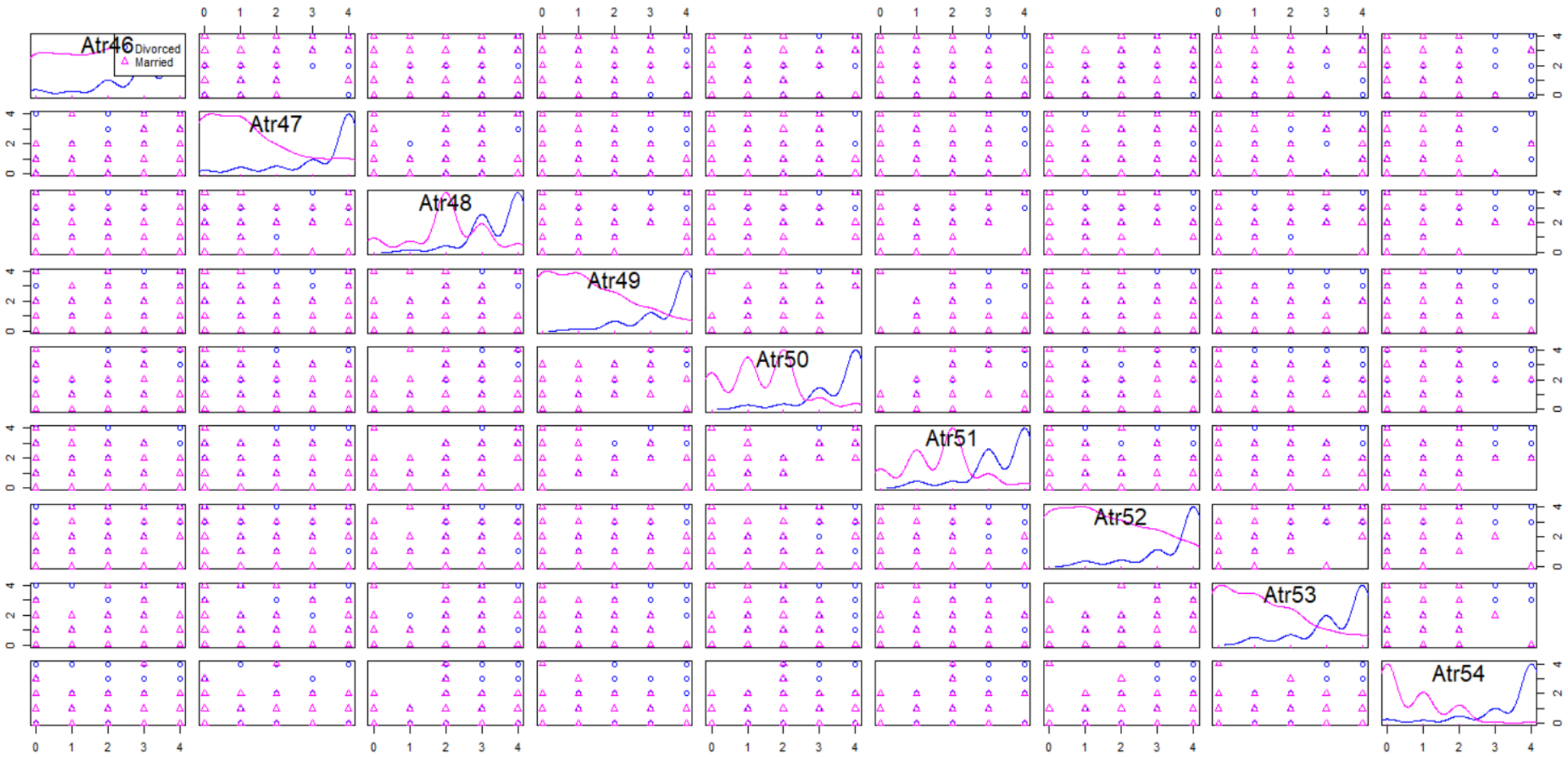
SCATTERPLOT MATRIX OF DIVORCE V/S ATTRIBUTES 28-36



SCATTERPLOT MATRIX OF DIVORCE V/S ATTRIBUTES 37-45



SCATTERPLOT MATRIX OF DIVORCE V/S ATTRIBUTES 46-54



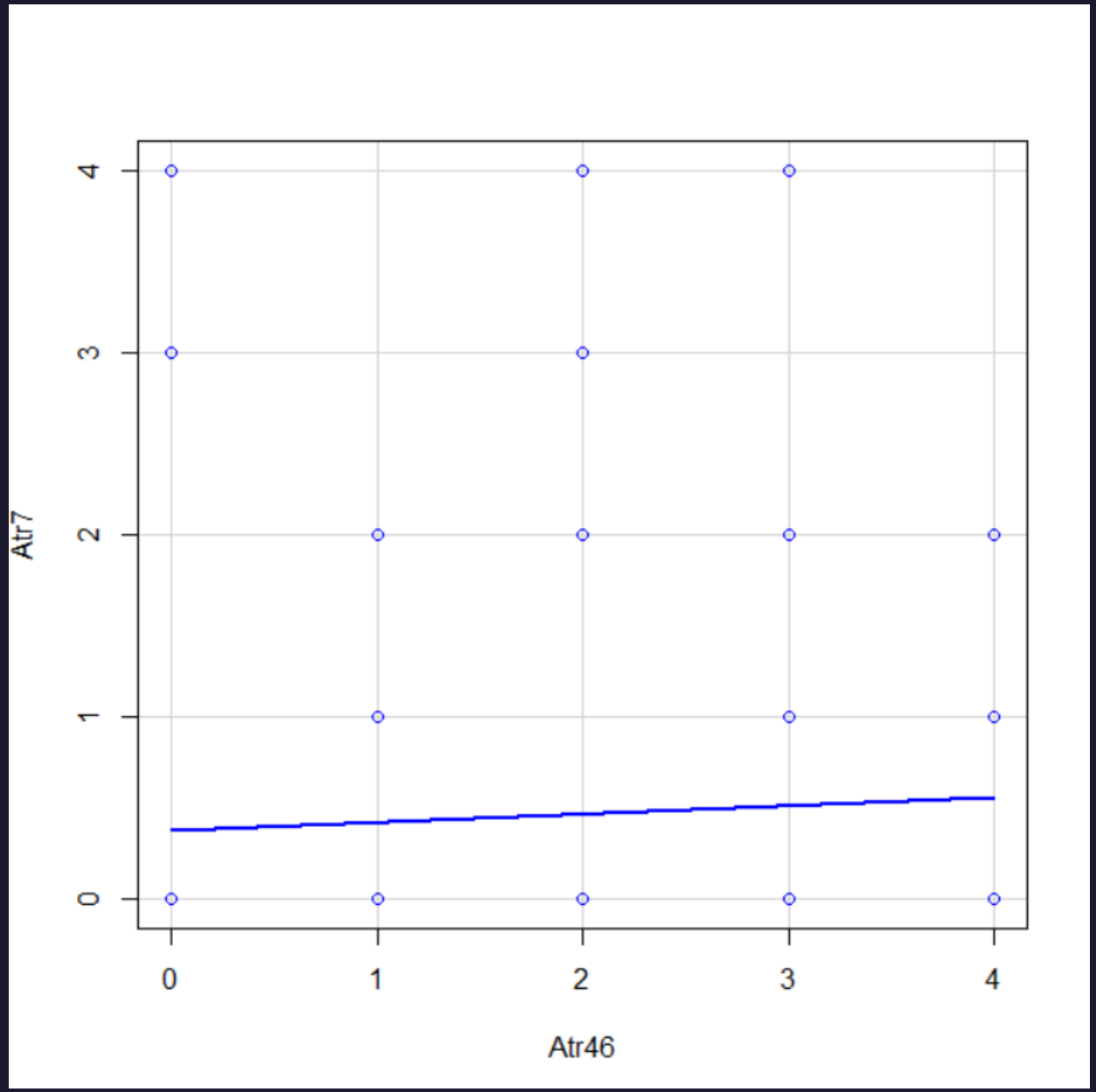
GRAPHS WITH NO RELATIONSHIP BETWEEN PAIR OF QUESTIONS

R CODE TO FIND WEAKEST 10 CORRELATION BETWEEN 54 ATTRIBUTES

- divorce <- read_excel("C:/Users/DELL/Desktop/divorce.xlsx")
- View(divorce)
- my_data <- divorce[, c(1:54)]
- res <- cor(my_data)
- sort(res,decreasing=F)[1:10]
- [1] 0.06984964 0.06984964 0.09482031 0.09482031 0.10284253 0.10284253 0.12775852 0.12775852
- [9] 0.14992964 0.14992964

GRAPHS WITH NO RELATIONSHIP BETWEEN PAIRS OF QUESTIONS

	Atr45	Atr46
Atr1	0.51016	0.400296
Atr2	0.489062	0.389519
Atr3	0.427409	0.308149
Atr4	0.446798	0.34024
Atr5	0.591656	0.470758
Atr6	0.09482	0.127759
Atr7	0.199548	0.06985

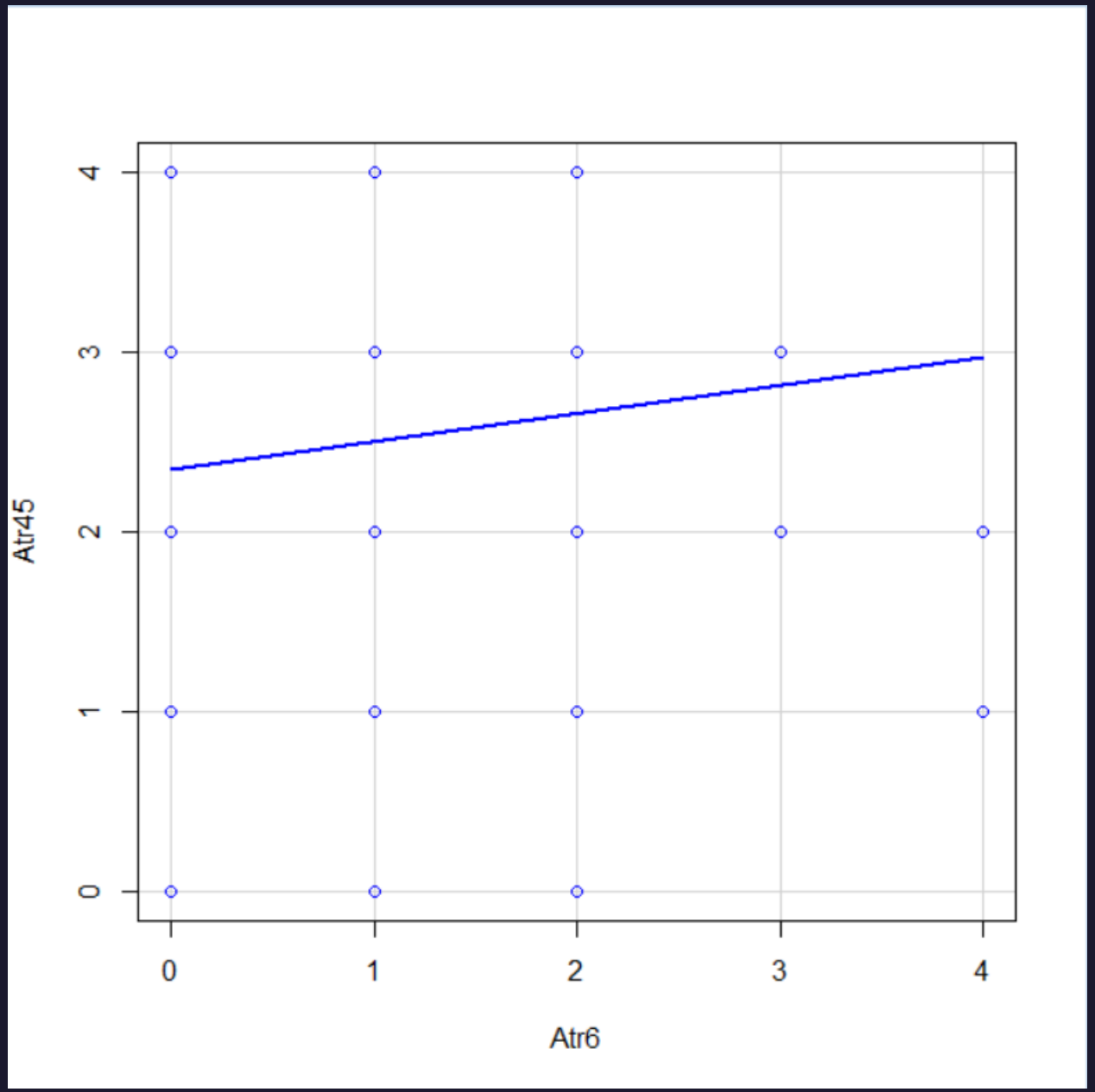


ATTRIBUTE 7 V/S 46

Attribute	List of Questions
Atr46	Even if I'm right in the discussion, I stay silent to hurt my spouse.
Atr7	We are like two strangers who share the same environment at home rather than family.

GRAPHS WITH NO RELATIONSHIP BETWEEN PAIRS OF QUESTIONS

	Atr6	Atr7
Atr1	0.28714	0.427989
Atr2	0.102843	0.417616
Atr45	0.09482	0.199548
Atr46	0.127759	0.06985
Atr47	0.212979	0.254225
Atr48	0.200673	0.31111

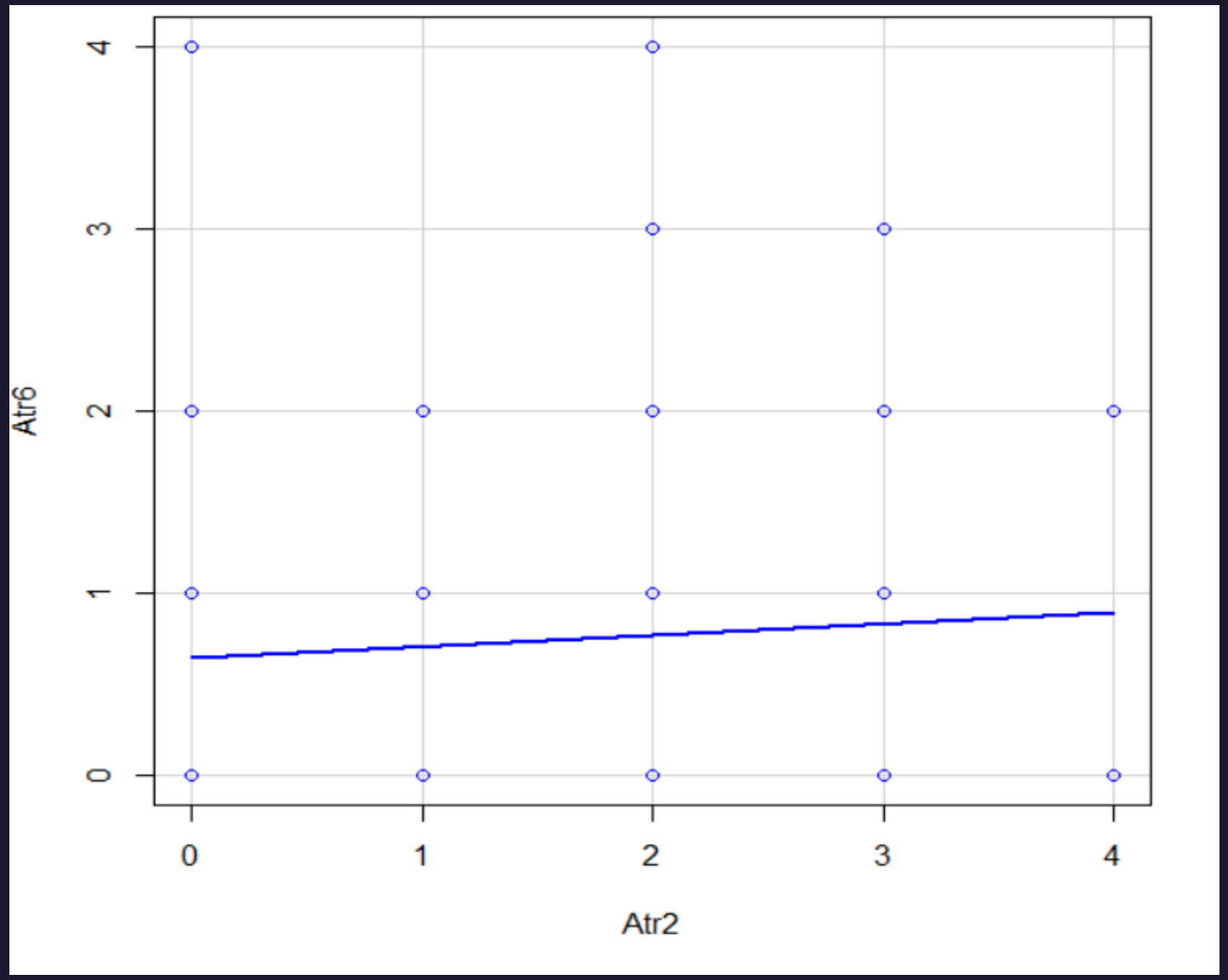


ATTRIBUTE 6 V/S 45

Attribute	List of Questions
Atr6	We don't have time at home as partners.
Atr45	I'd rather stay silent than discuss with my spouse.

GRAPHS WITH NO RELATIONSHIP BETWEEN PAIRS OF QUESTIONS

	Atr1	Atr2
Atr1	1	0.819066
Atr2	0.819066	1
Atr3	0.832508	0.805876
Atr4	0.825066	0.791313
Atr5	0.881272	0.81936
Atr6	0.28714	0.102843

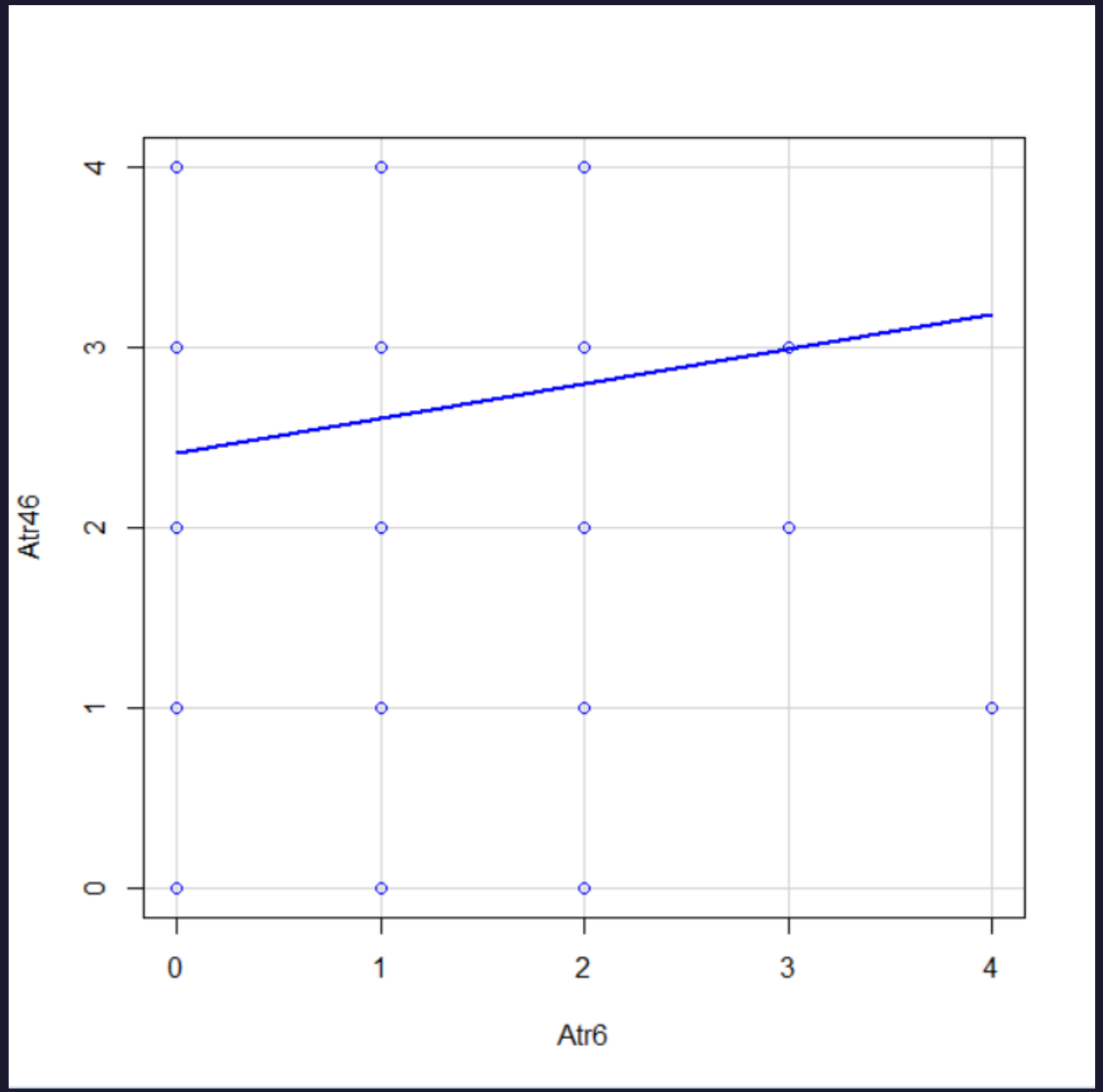


ATTRIBUTE 2 V/S 6

Attribute	List of Questions
Atr6	We don't have time at home as partners.
Atr2	I know we can ignore our differences, even if things get hard sometimes.

GRAPHS WITH NO RELATIONSHIP BETWEEN PAIRS OF QUESTIONS

	Atr5	Atr6
Atr43	0.613142	0.171599
Atr44	0.799453	0.339918
Atr45	0.591656	0.09482
Atr46	0.470758	0.127759

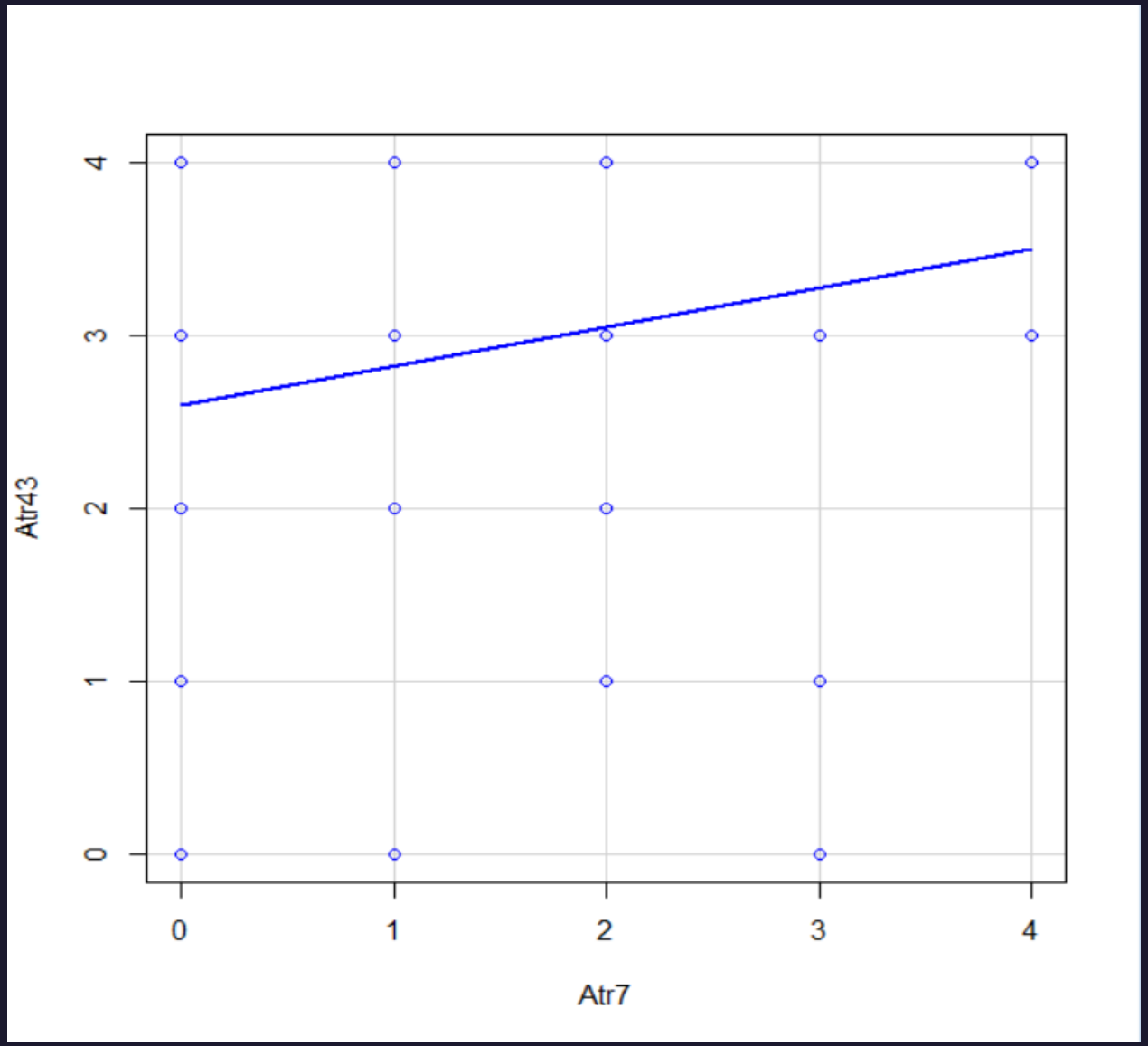


ATTRIBUTE 6 V/S 46

Attribute	List of Questions
Atr6	We don't have time at home as partners.
Atr46	Even if I'm right in the discussion, I stay silent to hurt my spouse.

GRAPHS WITH NO RELATIONSHIP BETWEEN PAIRS OF QUESTIONS

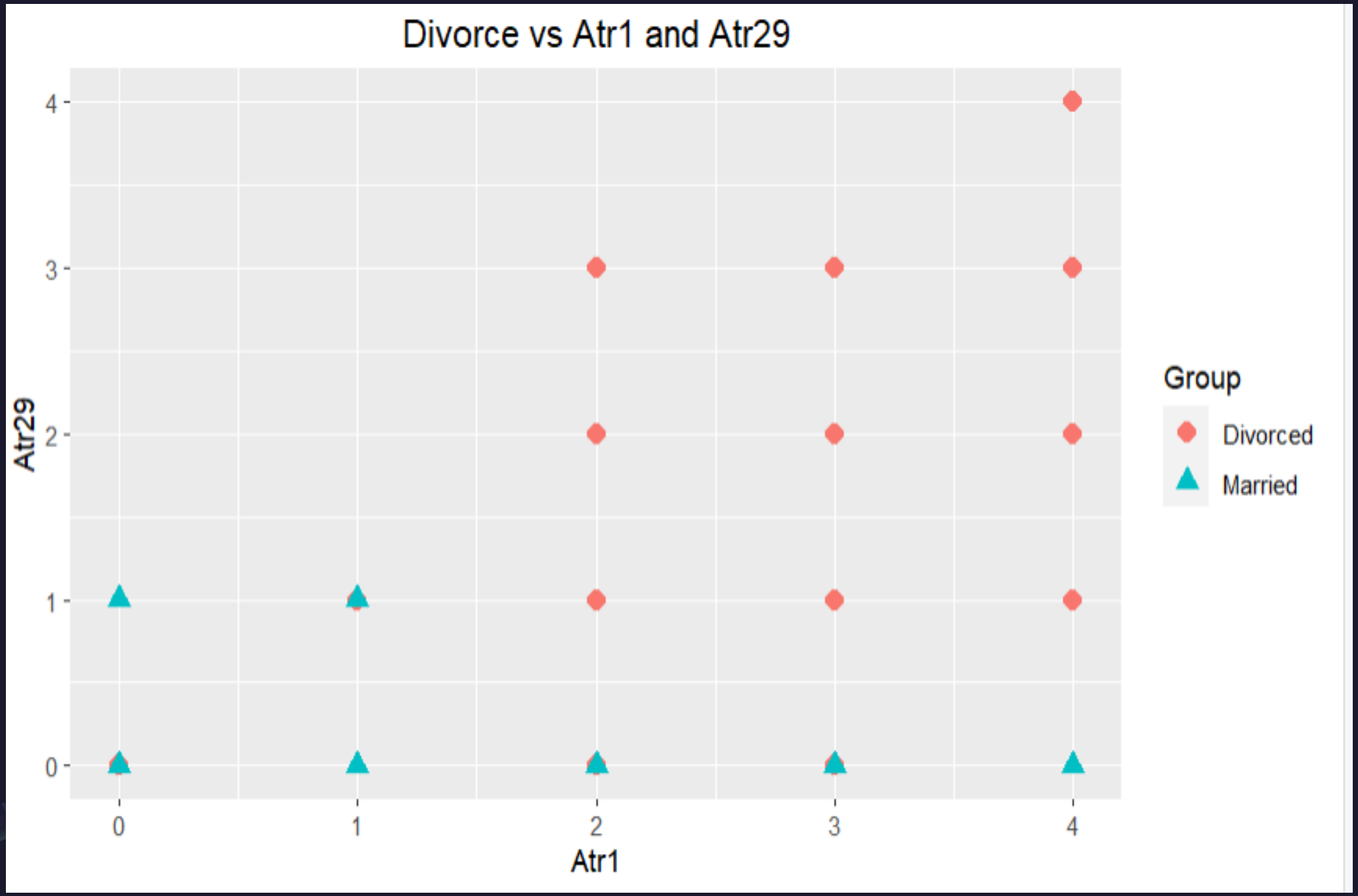
	Atr6	Atr7
Atr43	0.171599	0.14993
Atr44	0.339918	0.425874
Atr45	0.09482	0.199548
Atr46	0.127759	0.06985
Atr47	0.212979	0.254225
Atr48	0.200673	0.31111

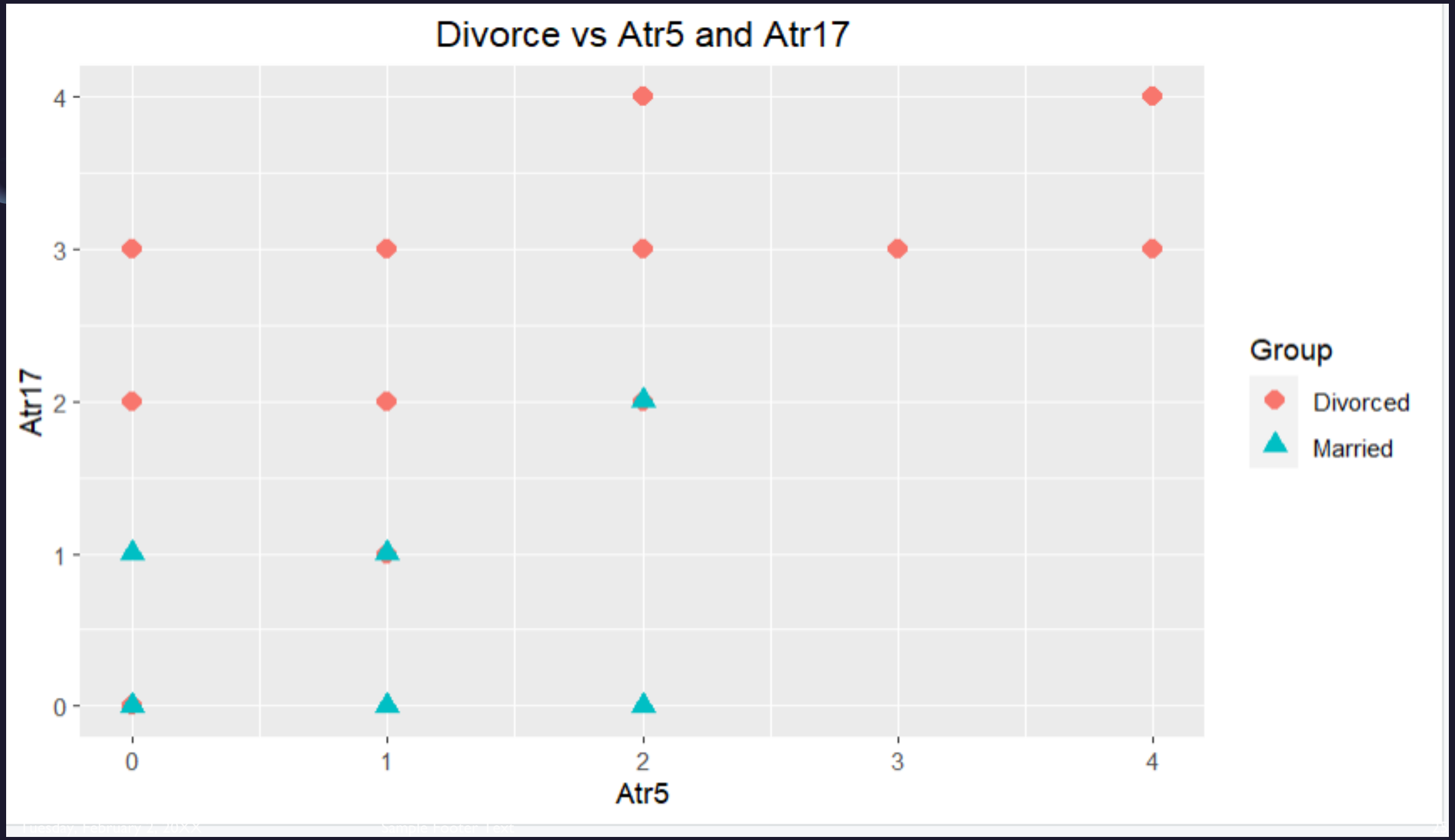


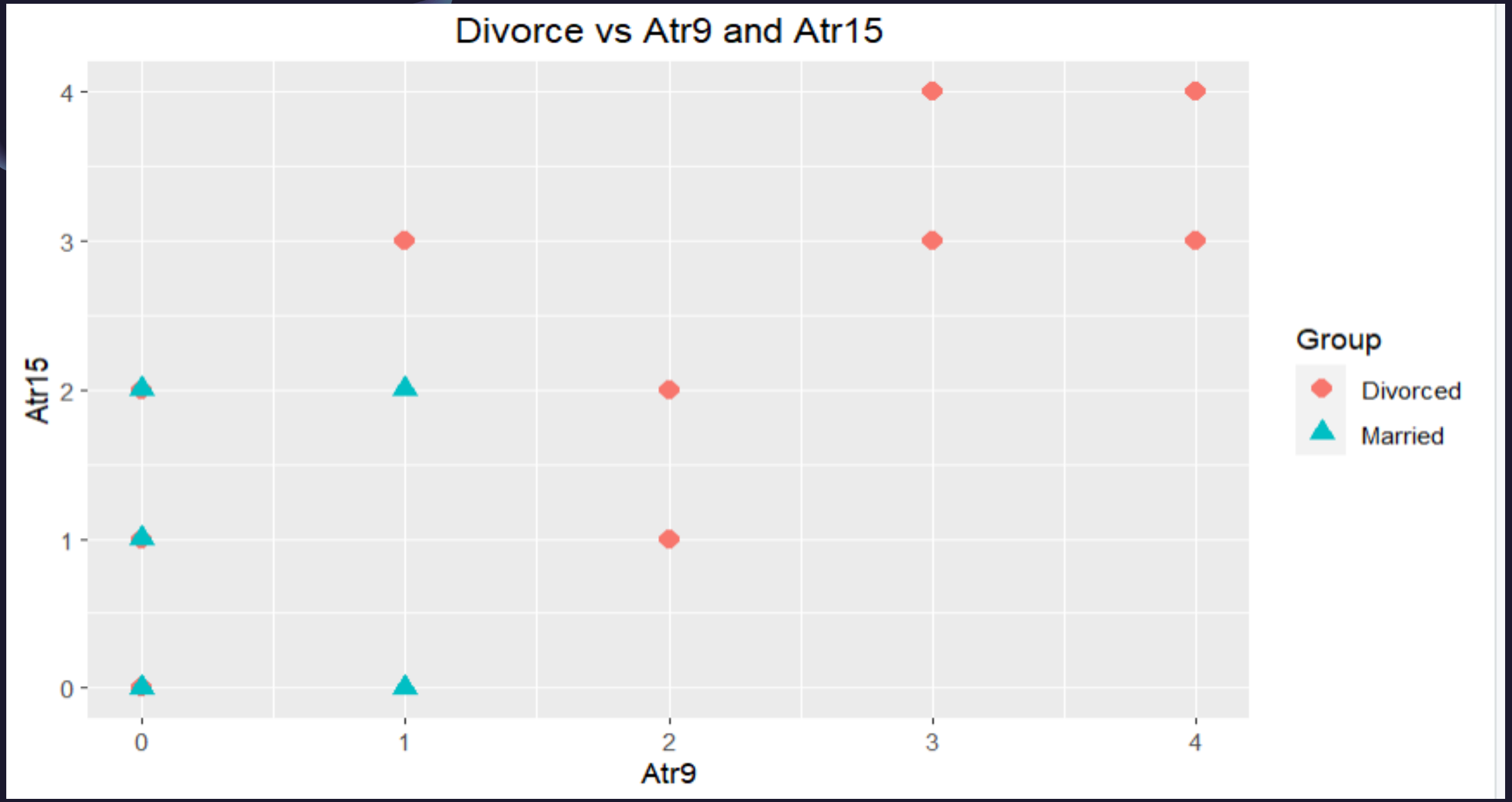
ATTRIBUTE 7 V/S 43

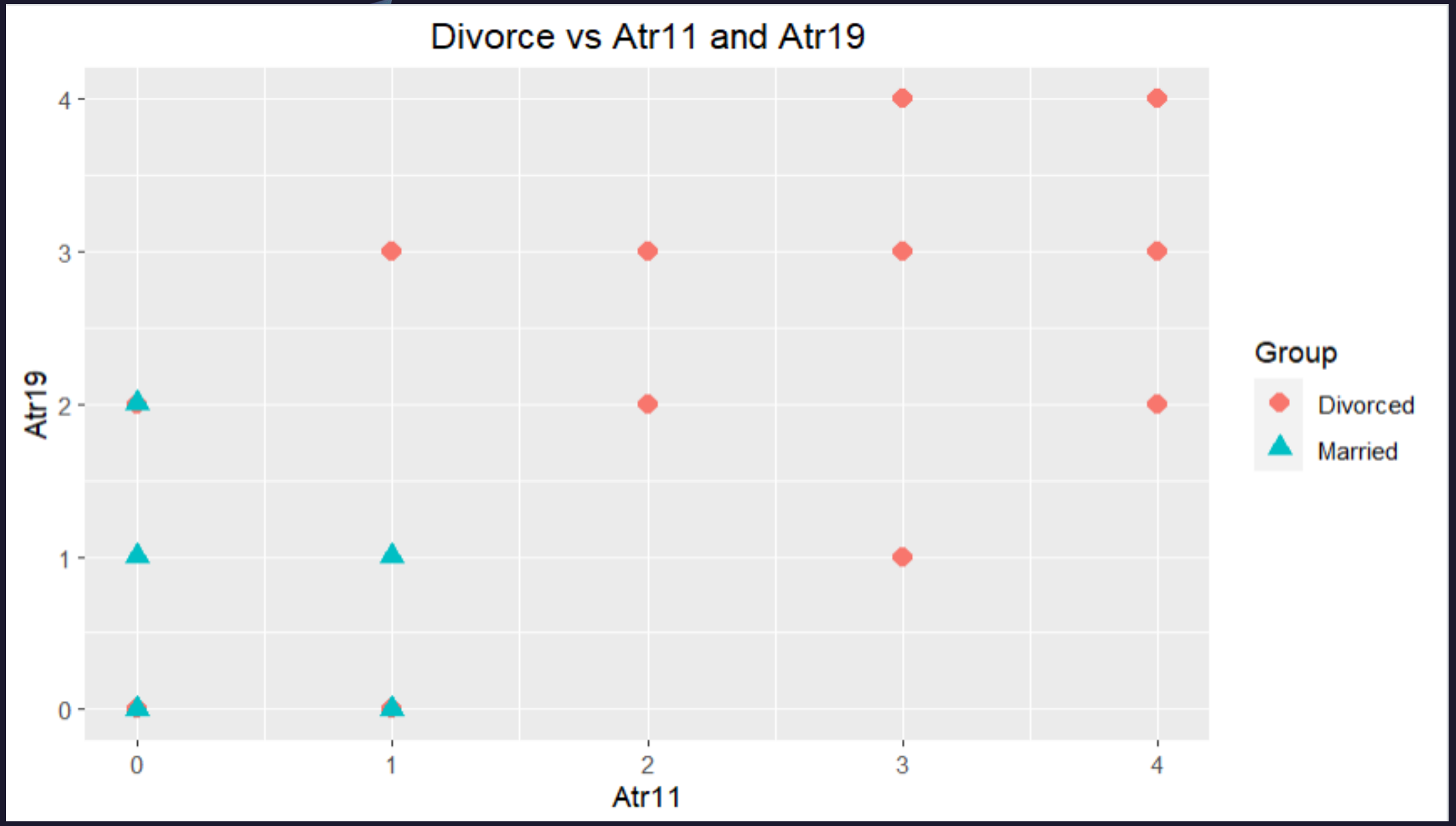
Attribute	List of Questions
Atr7	We are like two strangers who share the same environment at home rather than family
Atr43	I mostly stay silent to calm the environment a little bit.

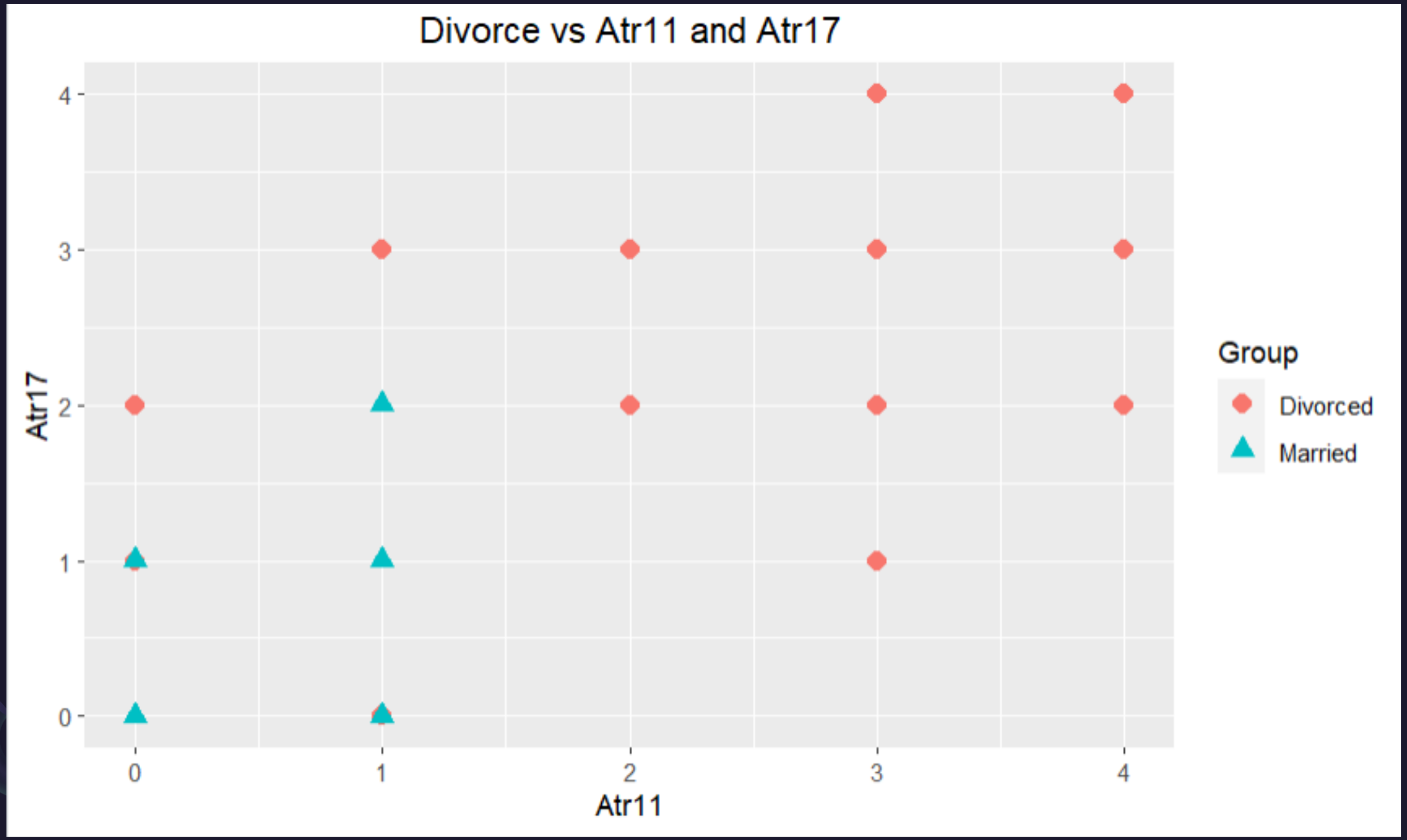
GRAPHS WITH PAIRS OF CONTINUOUS VARIABLES AND DIVORCE EVENTS


Divorce vs Atr1 and Atr29
Atr29
Atr1
0
1
2
3
4
Group
Divorced
Married


Divorce vs Atr5 and Atr17
Atr17
Atr5
0
1
2
3
4
Group
Divorced
Married


Divorce vs Atr9 and Atr15
Atr15
Atr9
0
1
2
3
4
Group
Divorced
Married


Divorce vs Atr11 and Atr19
Atr19
Atr11
0
1
2
3
4
Group
Divorced
Married


Divorce vs Atr11 and Atr17
Atr17
Atr11
0
1
2
3
4
Group
Divorced
Married

VARIANCE INFLATION FACTOR (VIF)

VIF ANALYSIS

```
> vif(RegModel.1)
     Atr1      Atr2      Atr3      Atr4      Atr5      Atr6      Atr7      Atr8      Atr9     Atr10     Atr11
14.450489 10.796914  9.692544 16.144843 30.967992  3.151901  5.158030 26.573779 40.222139 14.699374 36.709357
    Atr12     Atr13     Atr14     Atr15     Atr16     Atr17     Atr18     Atr19     Atr20     Atr21     Atr22
25.126364 14.733672 19.366610 26.705352 27.600872 52.861564 68.359229 43.431438 47.882343 35.714073 31.432679
    Atr23     Atr24     Atr25     Atr26     Atr27     Atr28     Atr29     Atr30     Atr31     Atr32     Atr33
31.683841 14.623710 31.060015 32.770170 25.172946 28.476411 51.338480 27.848034  9.829187 11.623947 30.098833
    Atr34     Atr35     Atr36     Atr37     Atr38     Atr39     Atr40     Atr41     Atr42     Atr43     Atr44
19.256379 37.294931 49.607464 14.392122 22.657649 20.180150 35.517048 18.587083  6.617568  5.339367 10.174080
    Atr45     Atr46     Atr47     Atr48     Atr49     Atr50     Atr51     Atr52     Atr53     Atr54
 5.781046  4.108797  6.708301  4.401816  8.109705 12.534393  9.182539  6.797882  8.545853 12.693577
```

VIF ANALYSIS

Atr1	Atr2	Atr3	Atr4	Atr5	Atr6	Atr7	Atr8	Atr9	Atr10	Atr11	
14.45049	10.79691	9.692544	16.14484	30.96799	3.151901	5.15803	26.57378	40.22214	14.69937	36.70936	
	Atr12	Atr13	Atr14	Atr15	Atr16	Atr17	Atr18	Atr19	Atr20	Atr21	Atr22
25.12636	14.73367	19.36661	26.70535	27.60087	52.86156	68.35923	43.43144	47.88234	35.71407	31.43268	
	Atr23	Atr24	Atr25	Atr26	Atr27	Atr28	Atr29	Atr30	Atr31	Atr32	Atr33
31.68384	14.62371	31.06002	32.77017	25.17295	28.47641	51.33848	27.84803	9.829187	11.62395	30.09883	
	Atr34	Atr35	Atr36	Atr37	Atr38	Atr39	Atr40	Atr41	Atr42	Atr43	Atr44
19.25638	37.29493	49.60746	14.39212	22.65765	20.18015	35.51705	18.58708	6.617568	5.339367	10.17408	
	Atr45	Atr46	Atr47	Atr48	Atr49	Atr50	Atr51	Atr52	Atr53	Atr54	
	5.781046	4.108797	6.708301	4.401816	8.109705	12.53439	9.182539	6.797882	8.545853	12.69358	

VIF ANALYSIS

Question	VIF
Atr6	3.151901
Atr46	4.108797
Atr48	4.401816
Atr7	5.15803
Atr43	5.339367
Atr45	5.781046
Atr42	6.617568
Atr47	6.708301
Atr52	6.797882
Atr49	8.109705
Atr53	8.545853
Atr51	9.182539
Atr3	9.692544
Atr31	9.829187
Atr44	10.17408
Atr2	10.796914
Atr32	11.623947
Atr50	12.534393
Atr54	12.693577
Atr37	14.392122
Atr1	14.450489
Atr24	14.62371
Atr10	14.699374
Atr13	14.733672
Atr4	16.144843
Atr41	18.587083
Atr34	19.256379
Atr14	19.36661
Atr39	20.18015
Atr38	22.657649
Atr12	25.126364
Atr27	25.172946
Atr8	26.573779
Atr15	26.705352
Atr16	27.600872
Atr30	27.848034
Atr28	28.476411
Atr33	30.098833
Atr5	30.967992
Atr25	31.060015
Atr22	31.432679
Atr23	31.683841
Atr26	32.77017
Atr40	35.517048
Atr21	35.714073
Atr11	36.709357
Atr35	37.294931
Atr9	40.222139
Atr19	43.431438
Atr20	47.882343
Atr36	49.607464
Atr29	51.33848
Atr17	52.861564
Atr18	68.359229

ATTRIBUTES WITH VIF<8

Attribute	VIF	List of Questions
Atr6	3.151901	We don't have time at home as partners.
Atr46	4.108797	Even if I'm right in the discussion, I stay silent to hurt my spouse.
Atr48	4.401816	I feel right in our discussions.
Atr7	5.15803	We are like two strangers who share the same environment at home rather than family.
Atr43	5.339367	I mostly stay silent to calm the environment a little bit.
Atr45	5.781046	I'd rather stay silent than discuss with my spouse.
Atr42	6.617568	When I argue with my spouse, ı only go out and I don't say a word.
Atr47	6.708301	When I discuss with my spouse, I stay silent because I am afraid of not being able to control my anger.
Atr52	6.797882	I wouldn't hesitate to tell my spouse about her/his inadequacy.

CORRELATION MATRIX BETWEEN 9 ATTRIBUTES WITH VIF<8

	Atr6	Atr7	Atr42	Atr43	Atr45	Atr46	Atr47	Atr48	Atr52
Atr6	1	0.42421224	0.227993	0.171599	0.09482	0.127759	0.212979	0.200673	0.205056
Atr7	0.42421224	1	0.333211	0.14993	0.199548	0.06985	0.254225	0.31111	0.243104
Atr42	0.22799335	0.33321101	1	0.719095	0.675723	0.548272	0.717083	0.618045	0.580543
Atr43	0.17159894	0.14992964	0.719095	1	0.775103	0.561868	0.69556	0.409197	0.51379
Atr45	0.09482031	0.19954849	0.675723	0.775103	1	0.592041	0.720692	0.413413	0.495944
Atr46	0.12775852	0.06984964	0.548272	0.561868	0.592041	1	0.664794	0.448773	0.55021
Atr47	0.21297863	0.25422502	0.717083	0.69556	0.720692	0.664794	1	0.602951	0.571223
Atr48	0.20067288	0.31110962	0.618045	0.409197	0.413413	0.448773	0.602951	1	0.513645
Atr52	0.20505634	0.24310427	0.580543	0.51379	0.495944	0.55021	0.571223	0.513645	1

PAIR OF QUESTIONS WITH HIGH CORRELATION

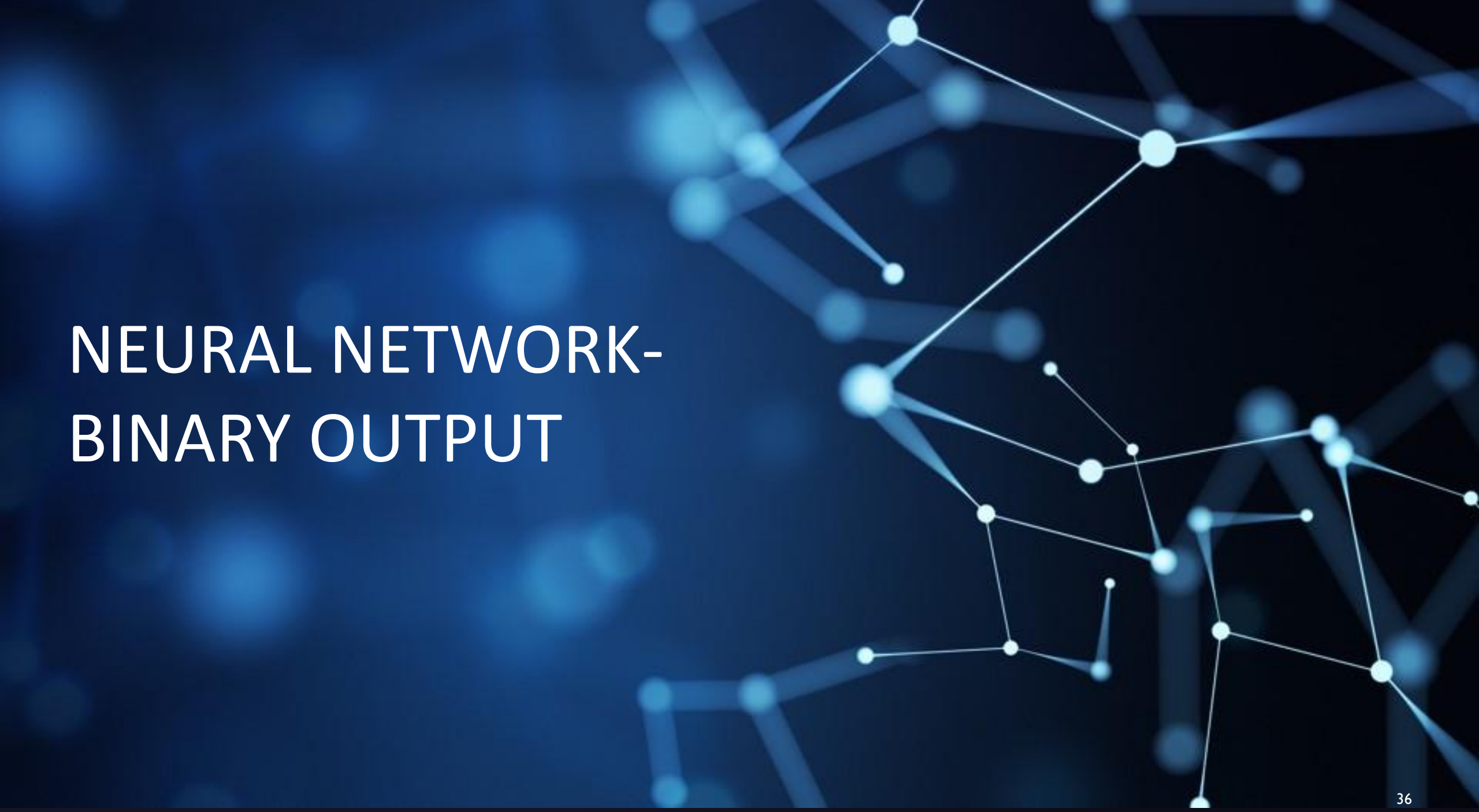
Attribute	VIF	List of Questions
Atr43	5.339367	I mostly stay silent to calm the environment a little bit.
Atr42	6.617568	When I argue with my spouse, ı only go out and I don't say a word.

Attribute	VIF	List of Questions
Atr42	6.617568	When I argue with my spouse, ı only go out and I don't say a word.
Atr47	6.708301	When I discuss with my spouse, I stay silent because I am afraid of not being able to control my anger.

Attribute	VIF	List of Questions
Atr43	5.339367	I mostly stay silent to calm the environment a little bit.
Atr45	5.781046	I'd rather stay silent than discuss with my spouse.

Attribute	VIF	List of Questions
Atr45	5.781046	I'd rather stay silent than discuss with my spouse.
Atr47	6.708301	When I discuss with my spouse, I stay silent because I am afraid of not being able to control my anger.

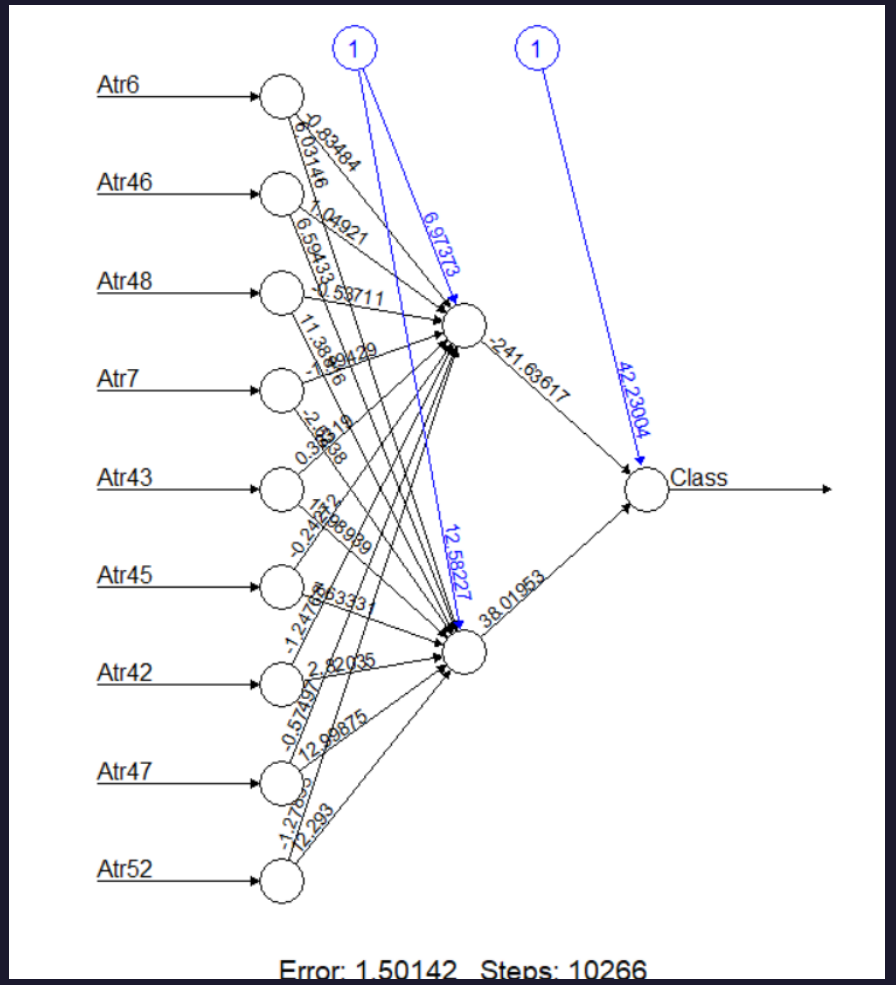
NEURAL NETWORK- BINARY OUTPUT



Neural network with Binary Output Class using the Nine Attributes (Atr) from the VIF analysis and all data points

NEURAL NETWORK OUTPUT

```
> divorcenet$result.matrix
                                  [,1]
error                        1.501419643
reached.threshold            0.009531608
steps                    10266.000000000
Intercept.to.llayhid1        6.973726235
Atr6.to.llayhid1            -0.834836409
Atr46.to.llayhid1            1.049207447
Atr48.to.llayhid1           -0.537107253
Atr7.to.llayhid1            -1.494291029
Atr43.to.llayhid1            0.333193909
Atr45.to.llayhid1           -0.242722284
Atr42.to.llayhid1           -1.247680313
Atr47.to.llayhid1           -0.574972541
Atr52.to.llayhid1           -1.278948215
Intercept.to.llayhid2       12.582270892
Atr6.to.llayhid2             6.031455236
Atr46.to.llayhid2            6.594330724
Atr48.to.llayhid2           11.388159878
Atr7.to.llayhid2            -2.623797629
Atr43.to.llayhid2           11.989394226
Atr45.to.llayhid2            7.633309423
Atr42.to.llayhid2            2.820353454
Atr47.to.llayhid2           12.998752473
Atr52.to.llayhid2           12.293002244
Intercept.to.Class          42.230043819
llayhid1.to.Class         -241.636165333
llayhid2.to.Class           38.019526865
>
```



- divorcenet <- neuralnet(Class ~ Atr6+ Atr46+Atr48+Atr7+Atr43+ Atr45+Atr42+Atr47+Atr52 , divorce, hidden=2, lifesign="minimal", linear.output=FALSE, threshold=0.01)
- divorcenet$result.matrix
- plot(divorcenet)

Neural Network 2 with Binary Output Class using the Nine

Attributes (Atr) from the VIF analysis and all Data Points using 50%

Training Data, Randomly selected

NEURAL NETWORK 2 WITH 50% TRAINING DATA RANDOMLY SELECTED

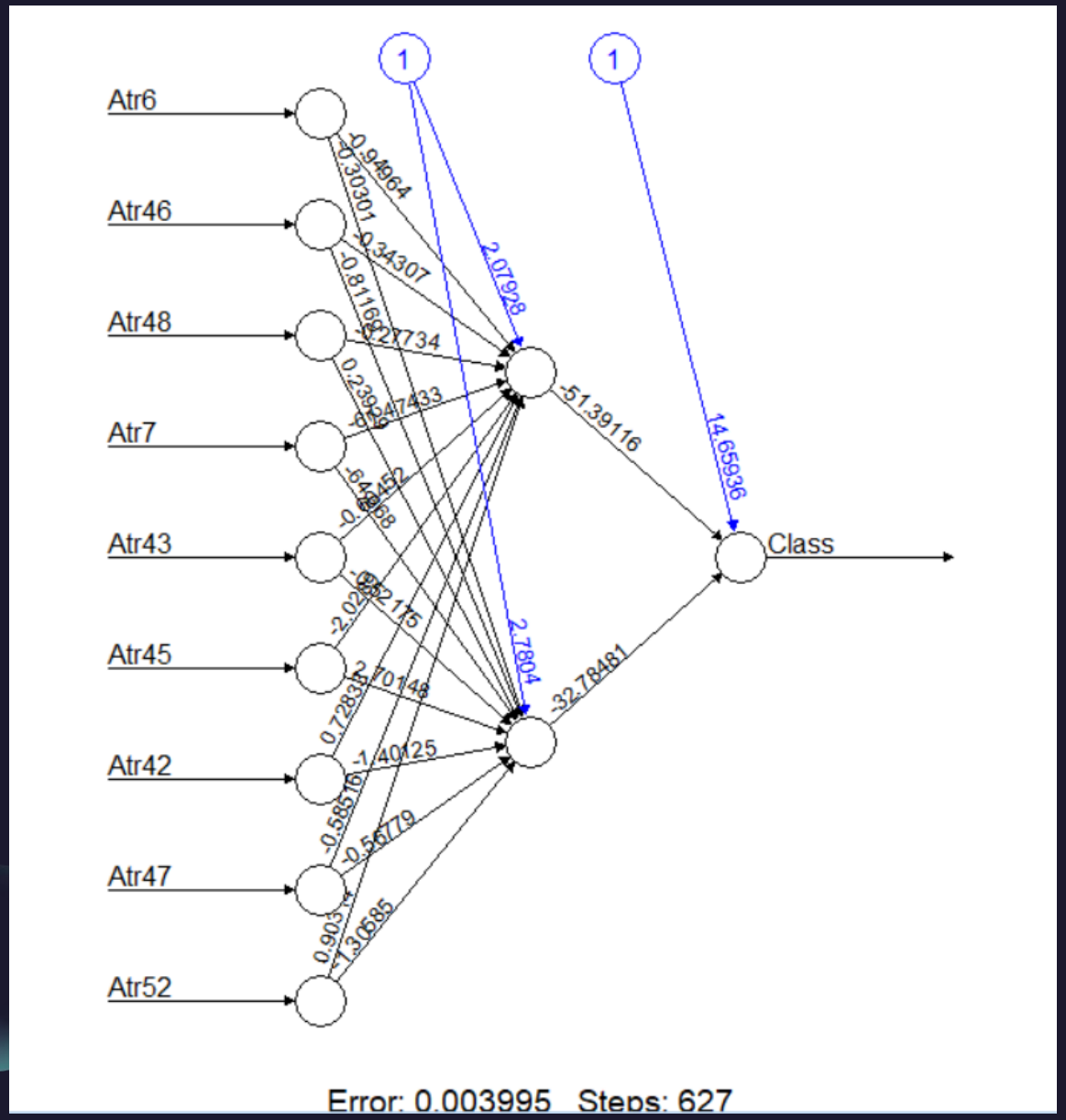
- divorce_index <- sample(nrow(divorce), 1/2 * nrow(divorce))
- divorce_train <- divorce[divorce_index,]
- divorce_test <- divorce[-divorce_index,]
- head(divorce_train)
- head(divorce_test)
- divorcenet <- neuralnet(Class ~ Atr6+ Atr46+Atr48+Atr7+Atr43+Atr45+Atr42+Atr47+Atr52, divorce_train, hidden=2 lifesign="minimal", linear.output=FALSE, threshold=0.01)
- plot(divorcenet)
- temp_test <- subset(divorce_test, select=c("Atr46","Atr48","Atr7","Atr43","Atr52","Atr6","Atr47","Atr45","Atr42"))
- head(temp_test)

50% TRAINING DATA

```
> head(divorce_train)
    Atr1 Atr2 Atr3 Atr4 Atr5 Atr6 Atr7 Atr8 Atr9 Atr10 Atr11 Atr12 Atr13 Atr14 Atr15 Atr16 Atr17 Atr18 Atr19 Atr20 Atr21 Atr22 Atr23 Atr24
53     4    3    2    3    4    1    0    3    2     3     4     3     4     3     2     3     4     3     4     3     2     3     4     3
117    0    0    0    0    0    2    0    0    0     0     0     0     1     0     0     1     1     0     0     0     0     0     0     0
47     3    3    3    2    3    1    1    3    3     2     3     3     3     3     3     2     3     3     3     3     3     2     3     3
148    0    0    1    0    0    0    0    0    0     0     0     0     0     0     0     0     0     1     0     0     0     0     2     2
106    0    0    0    0    0    0    0    0    0     0     0     1     1     1     1     0     0     0     0     0     1     1     0     0
46     3    3    2    3    3    1    1    3    2     3     3     3     3     3     2     3     3     3     3     3     2     3     3     3
    Atr25 Atr26 Atr27 Atr28 Atr29 Atr30 Atr31 Atr32 Atr33 Atr34 Atr35 Atr36 Atr37 Atr38 Atr39 Atr40 Atr41 Atr42 Atr43 Atr44 Atr45 Atr46
53      4     3     2     3     4     3     4     3     4     3     4     3     4     3     4     3     4     3     4     3     4     3
117     0     0     0     0     0     1     0     0     0     1     0     0     1     0     0     0     0     2     4     0     4     3
47      3     3     3     2     3     3     4     4     4     4     4     4     4     4     4     4     4     4     4     4     4     4
148     1     1     1     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     4     4     4     4
106     0     0     0     1     0     0     0     0     0     1     0     0     1     0     1     0     0     0     2     2     2     0
46      3     3     2     3     3     3     4     4     4     4     4     4     4     4     4     4     4     4     4     4     4     4
    Atr47 Atr48 Atr49 Atr50 Atr51 Atr52 Atr53 Atr54 Class    Group
53      4     3     4     4     3     4     3     4     1 Divorced
117     1     1     2     2     2     1     0     0     0  Married
47      4     4     4     4     4     4     4     4     1 Divorced
148     4     0     4     0     0     0     0     0     0  Married
106     0     2     2     2     2     1     1     0     0  Married
46      4     4     4     4     4     4     4     4     1 Divorced
```

DIVORCENET PLOT WITH

50% TRAINING DATA



50% RANDOM TESTING DATA

```
> head(temp_test)
   Atr46 Atr48 Atr7 Atr43 Atr52 Atr6 Atr47 Atr45 Atr42
1      2     3    0     1     3    0     1     3     1
3      3     3    2     2     2    3     2     2     3
7      3     3    4     3     2    3     2     3     3
8      0     2    0     3     1    1     1     0     4
10     2     2    0     3     4    2     0     2     2
11     4     4    0     4     4    0     4     4     4
> 
```

PREDICTION ON 50% TESTING DATA

- divorcenet.results <- compute(divorcenet, temp_test)
- results <- data.frame(actual=divorce_test$Class, prediction=divorcenet.results$net.result)
- results[1:20,]
- results$prediction <- round(results$prediction)
- results [1:20,]

PREDICTION OUTCOME FIRST 20 ROWS- ALL CORRECT PREDICTIONS

Index	Actual	Prediction
1	1	1
3	1	1
7	1	1
8	1	1
10	1	1
11	1	1
12	1	1
13	1	1
14	1	1
15	1	1
16	1	1
18	1	1
20	1	1
21	1	1
22	1	1
27	1	1
31	1	1
32	1	1
33	1	1
35	1	1

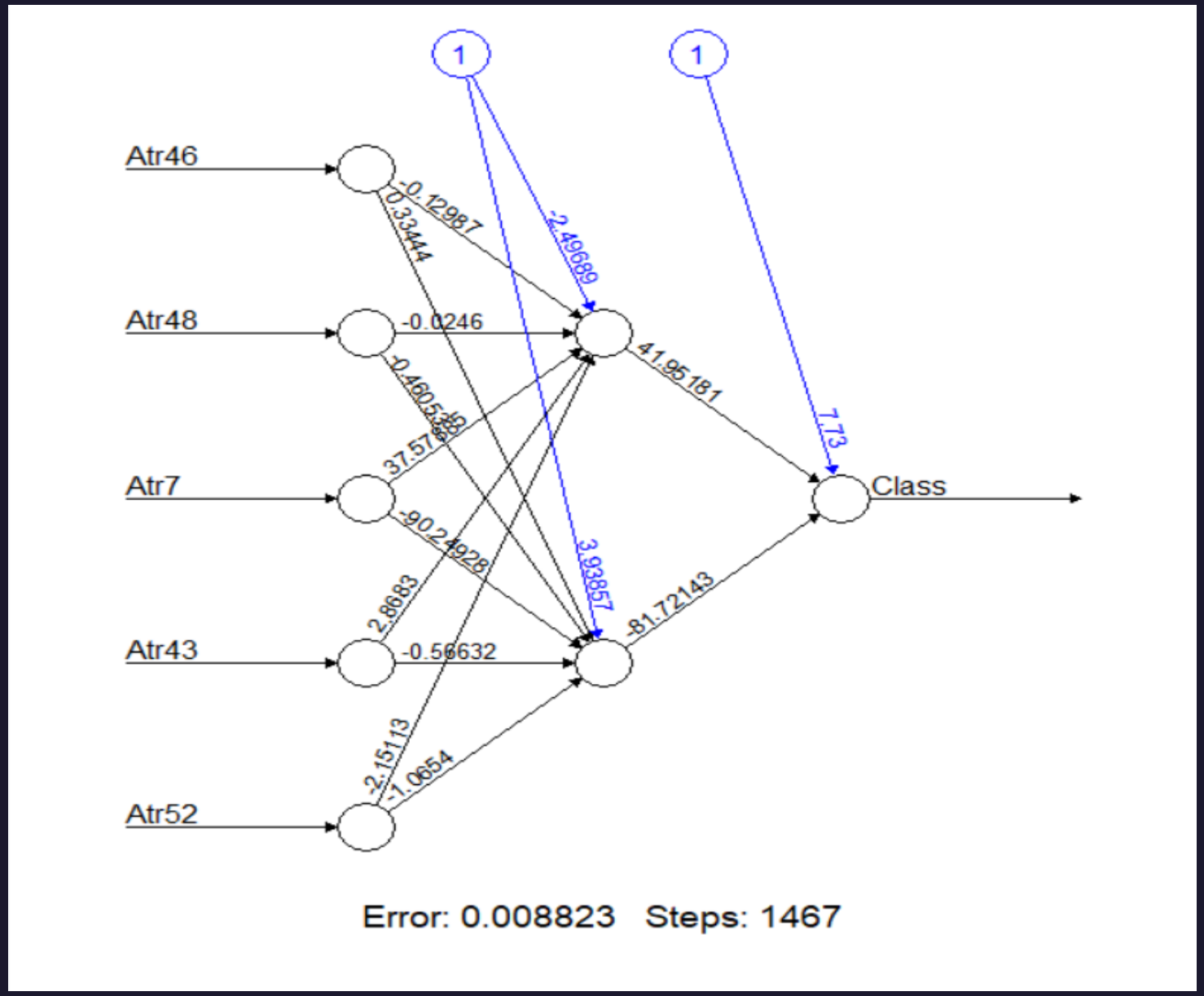
Third Neural Network with Binary Output Class using the Attributes

Atr7, Atr43, Atr46, At48, Atr52

NEURAL NETWORK 3

- divorce_index <- sample(nrow(divorce), 1/2 * nrow(divorce))
- divorce_train <- divorce[divorce_index,]
- divorce_test <- divorce[-divorce_index,]
- head(divorce_train)
- head(divorce_test)
- divorcenet <- neuralnet(Class ~ Atr46+Atr48+Atr7+Atr43+Atr52, divorce_train, hidden=2, lifesign="minimal", linear.output=FALSE, threshold=0.01)
- plot(divorcenet)
- temp_test <- subset(divorce_test, select = c("Atr46","Atr48","Atr7","Atr43","Atr52"))
- head(temp_test)

PLOT


1
1
Atr46
Atr48
Atr7
Atr43
Atr52
Class
-0.12987
0.33444
-2.49689
-0.0246
-0.46053
37.576
-90.24928
2.8683
-0.56632
-2.15113
-1.0654
3.93857
41.95181
-81.72143
7.73
Error: 0.008823 Steps: 1467

TESTING DATA

```
> head(temp_test)
   Atr46 Atr48 Atr7 Atr43 Atr52
1      2     3    0     1     3
2      2     3    0     3     4
6      2     1    0     2     1
7      3     3    4     3     2
9      1     1    1     2     1
12     4     4    0     4     4
>
```

PREDICTION ON 50% TESTING DATA

```
divorcenet.results <- compute(divorcenet,
temp_test)
```

```
results <-
data.frame(actual=divorce_test$Class,
prediction=divorcenet.results$net.result)
```

```
results[1:20,]
```

```
results$prediction <-
round(results$prediction)
```

```
results [1:20,]
```

PREDICTION OUTCOME FIRST 20 ROWS -2 INCORRECT PREDICTIONS

Index	Actual	Prediction
1	1	0
2	1	1
6	1	0
7	1	1
9	1	1
12	1	1
13	1	1
17	1	1
20	1	1
22	1	1
28	1	1
30	1	1
32	1	1
33	1	1
38	1	1
39	1	1
40	1	1
43	1	1
44	1	1
45	1	1

USING SEED FUNCTION TO GET SAME SAMPLE EVERYTIME

```
> set.seed(5)
> divorce_index <- sample(nrow(divorce), 1/2 * nrow(divorce))
> divorce_index
 [1]  66 107 121  41  71 147 131  38  58  53  76  16 154 128  27  74  85 122  51 113  92  13  54 146 138  42 116   8 109 104 123  70  91  99  72 152  80 143 160  60 125  90 127  97  15 118
[47] 100  81  20 114  11   9   4   3  28 156 137 130 115 167 163   2 159  21 164 119  17   6  61 170   7  24  29  40 106  94  62 112 141  78  48 140  63  77 149
> set.seed(5)
>
> divorce_index <- sample(nrow(divorce), 1/2 * nrow(divorce))
> divorce_index
 [1]  66 107 121  41  71 147 131  38  58  53  76  16 154 128  27  74  85 122  51 113  92  13  54 146 138  42 116   8 109 104 123  70  91  99  72 152  80 143 160  60 125  90 127  97  15 118
[47] 100  81  20 114  11   9   4   3  28 156 137 130 115 167 163   2 159  21 164 119  17   6  61 170   7  24  29  40 106  94  62 112 141  78  48 140  63  77 149
> set.seed(5)
>
> divorce_index <- sample(nrow(divorce), 1/2 * nrow(divorce))
> divorce_index
 [1]  66 107 121  41  71 147 131  38  58  53  76  16 154 128  27  74  85 122  51 113  92  13  54 146 138  42 116   8 109 104 123  70  91  99  72 152  80 143 160  60 125  90 127  97  15 118
[47] 100  81  20 114  11   9   4   3  28 156 137 130 115 167 163   2 159  21 164 119  17   6  61 170   7  24  29  40 106  94  62 112 141  78  48 140  63  77 149
>
```

NEURAL NETWORK 2 V/S NEURAL NETWORK 3 PREDICTION COMPARISION

Random Sample	Total Rows in Test Dataset	Number of Rows Correctly Predicted by NN2	Number of Rows Correctly Predicted by NN3	Neural Network 2 Correct Prediction %	Neural Network 3 Correct Prediction%
set.seed(1)	85	52	75	61.18%	88.24%
set.seed(2)	85	49	78	57.65%	91.76%
set.seed(3)	85	47	77	55.29%	90.59%
set.seed(4)	85	49	78	56.47%	91.76%
set.seed(5)	85	48	79	56.47%	92.94%
set.seed(6)	85	44	77	51.76%	90.59%
set.seed(7)	85	48	76	56.47%	89.41%

CONFUSION MATRIX

		Predictions	
		Married	Divorced
Actual Data	Married	8	36
	Divorced	1	40

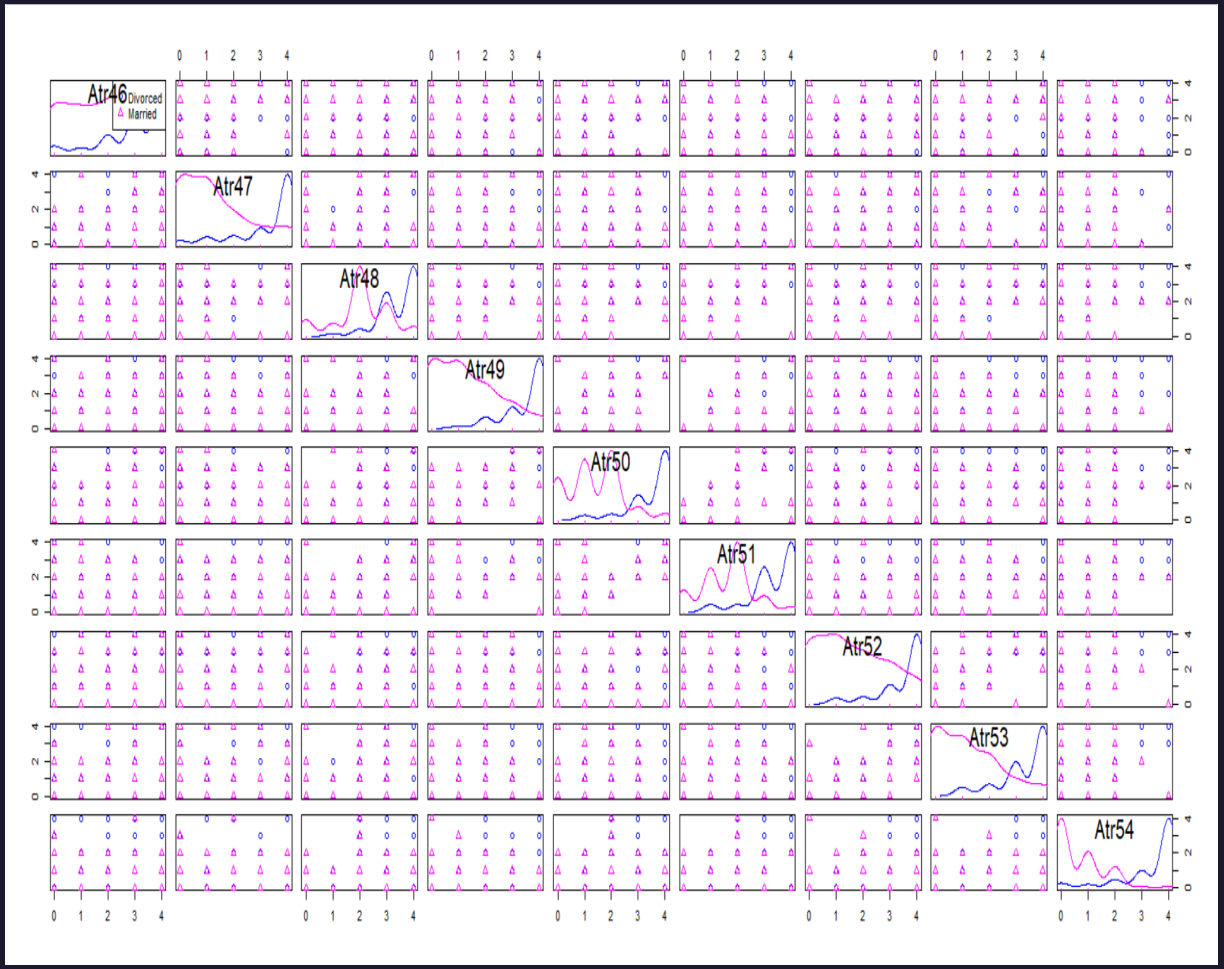
CONCLUSION

Accuracy of Neural Network 3 is way too high than Neural Network 2

There is a very high correlation between Attributes **Atr6,Atr47,Atr45,Atr42**

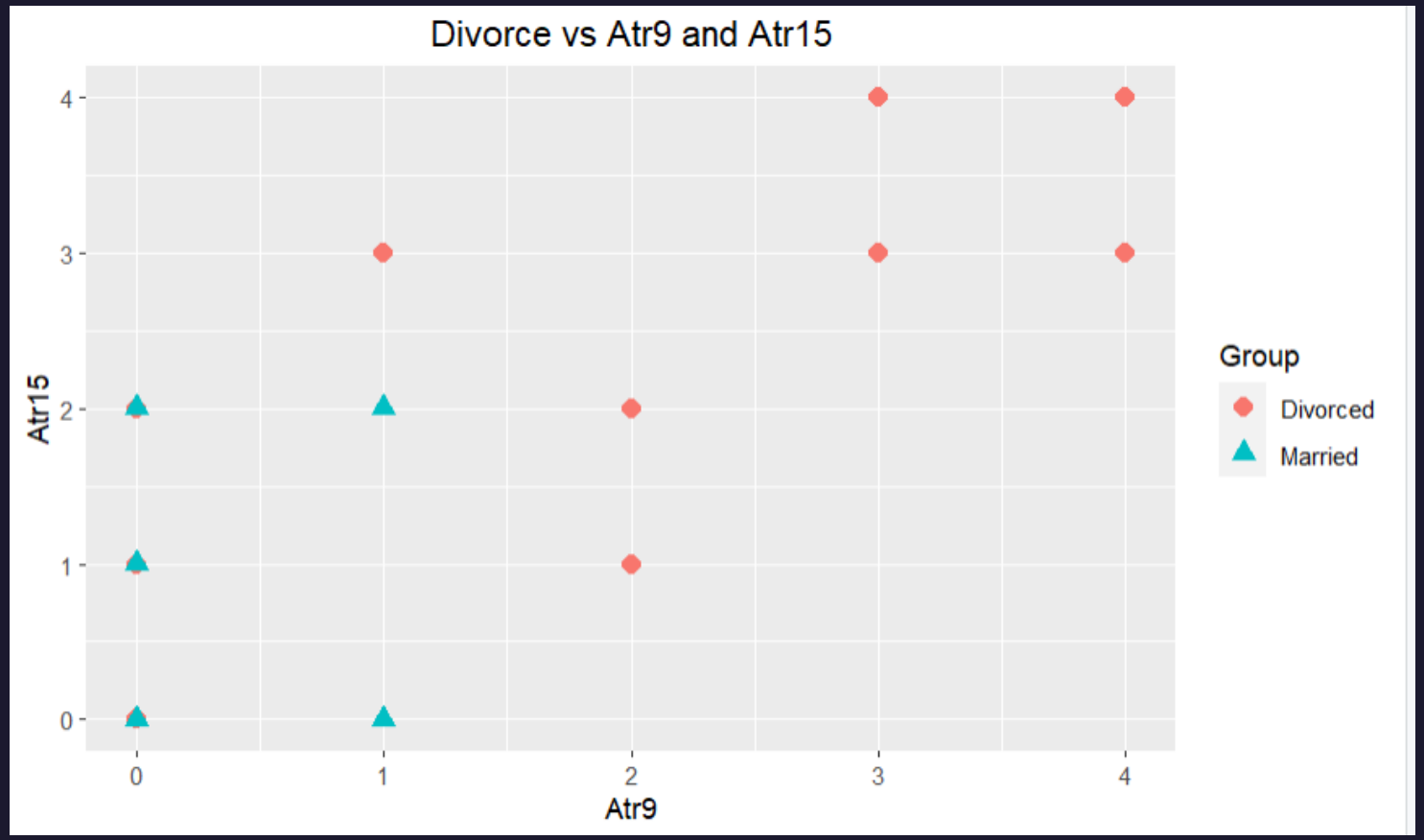
WHEN DO VISUALIZATIONS HELP ? WHEN THEY DO NOT?

WHEN VISUALIZATIONS DO NOT HELP



- When the dataset is huge and there are many independent variables
- It's very difficult to interpret relationship between X variables and divorce event in the scatterplot matrix
- Splitting scatterplot matrix into groups of 6 as there are 54 independent variables
- Difficult to plot relationship between for e.g. attribute 9 and attribute 10

WHEN VISUALIZATIONS HELPS



- Scatterplot helps in understanding the relationship between two attributes and plot divorce events
- Gives a clear picture compared to scatterplot matrix

THANK YOU

QUESTIONS?