MGMT 6962 Assignment 7

Name: Yihui Yang **RIN**: 661979332

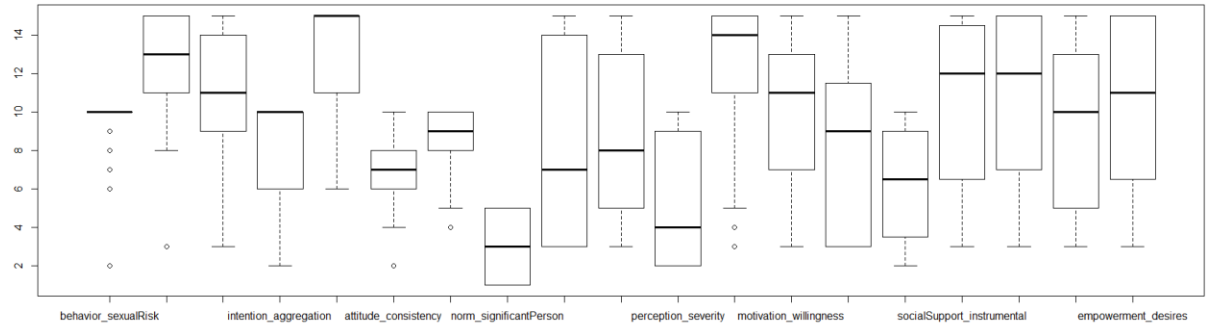
The data set I choose is Cervical Cancer Behavior Risk Data Set and Wine Quality Data Set.

1. Exploratory Data Analysis

First, take a look at cervical cancer behavior risk data. There are 19 variables in this data set.

	behavior_sexualRisk	behavior_eating	behavior_personalHygine	intention_aggregation	intention_commitment	attitude_consistency	attitude_spontaneity	norm_significantPerson	norm_fulfillment
1	10	13	12	4	7	9	10	1	8
2	10	11	11	10	14	7	7	5	5
3	10	15	3	2	14	8	10	1	4
4	10	11	10	10	15	7	7	1	5
5	8	11	7	8	10	7	8	1	5

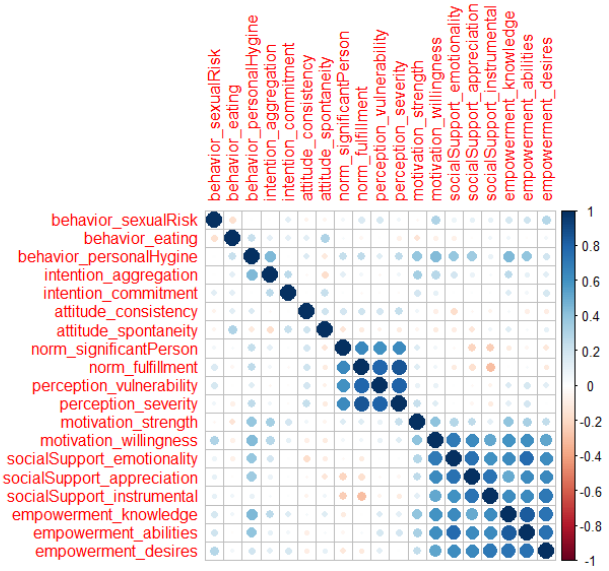
The first 18 columns are independent variables, which comes from 8 main variables including behavior, intention, attitude, norm, perception, motivation, social support and empowerment. The last columns is whether the this patient has cancer or not.



As the first 18 columns are numeric, I can plot the boxplot to check the distribution. Most of the 'behavior_eating' score is 10, but there are a few outliners.

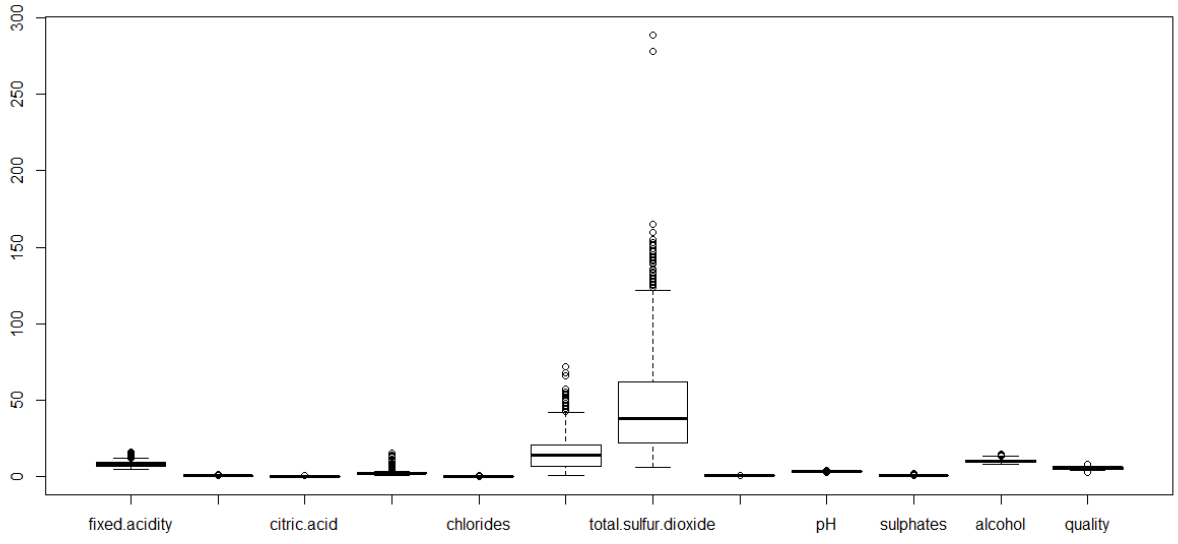
From the graph above, we can conclude that most data located in the range from 2 to 14.

I used the correlation maps to find the relationship within the independent variables.

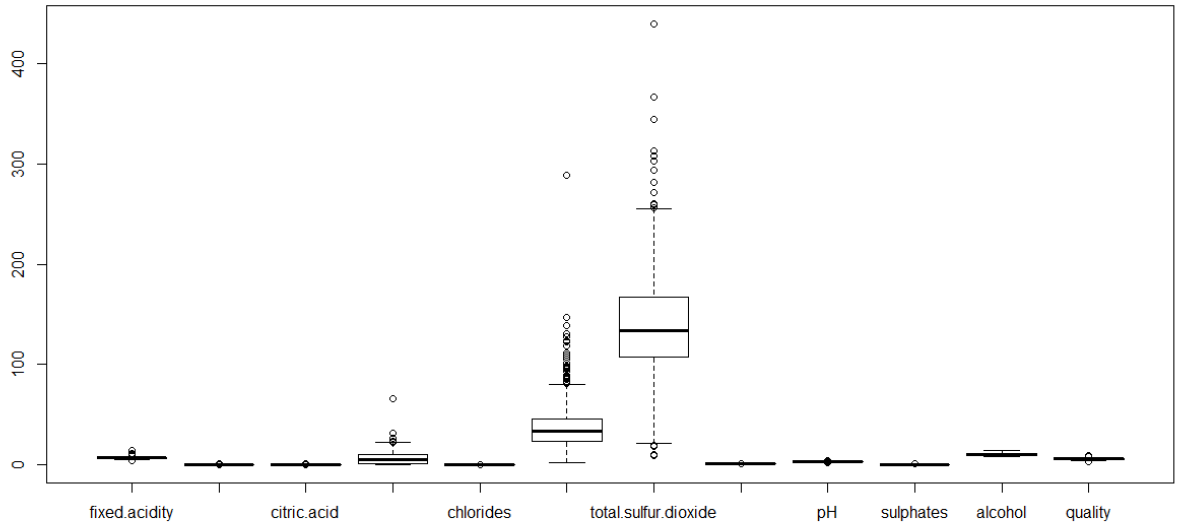


Behavior, intention and attitude demensions have no strong correlation with each other. Norm and perception have strong correlation with each other. Motivation, social support and empowerment have strong correlation with each other.

Then, take a look at the wine quality data. The wine quality data consist of red and white wine data.



red

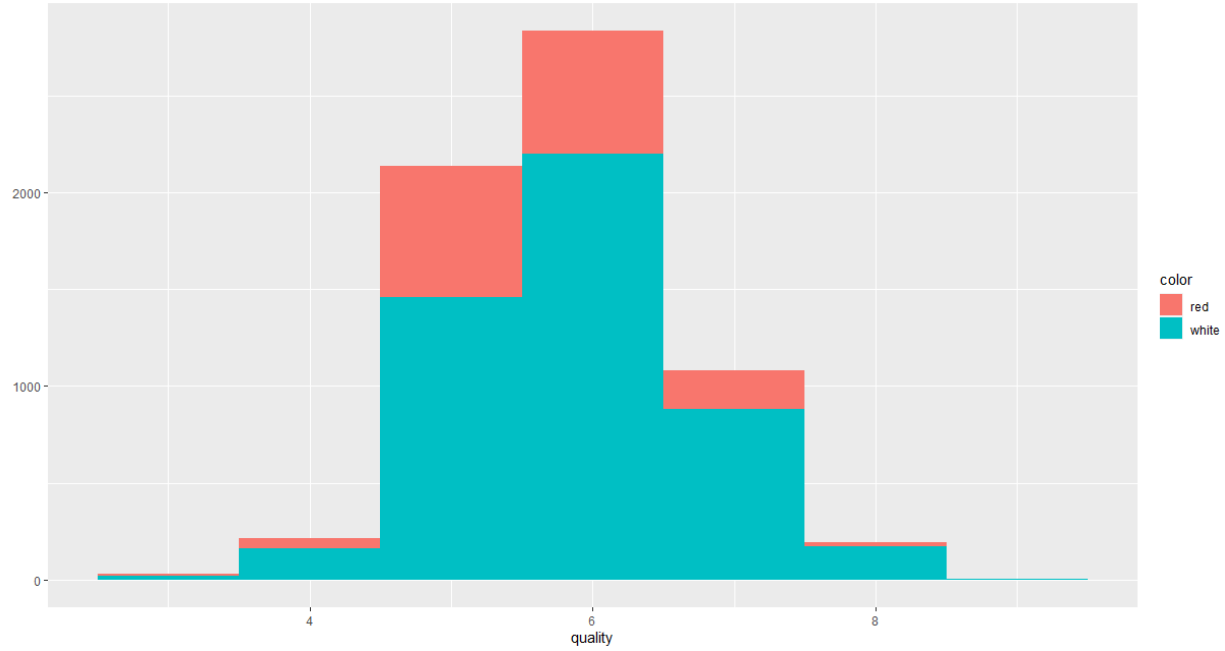


White

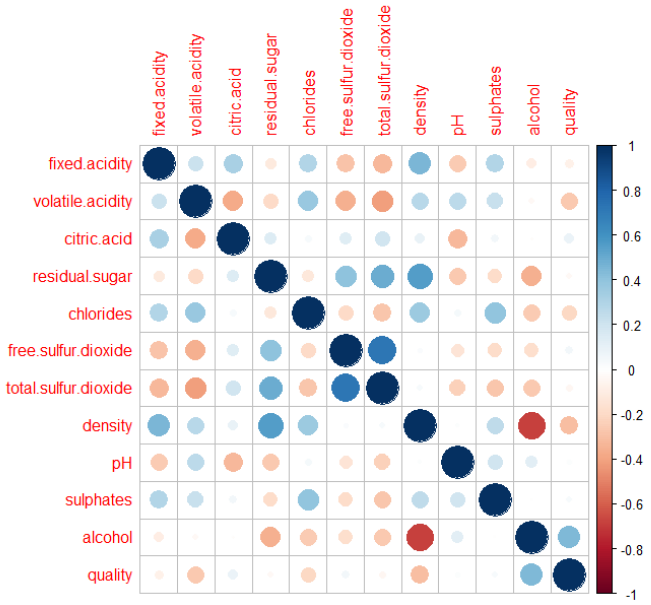
Both these data sets have same number of dimensions, but it seems that red and white wine have different scales. In the future models, I should conduct normalization method before modeling.

Also, red and white wine seem to have different characters. I will combine both the data set together and do some analysis after.

After combining them together, here is the distribution of quality.



Here is the correlation map.



Free sulfur dioxide and total sulfur dioxide have strong correlation with each other.

2. Model Development

For Cervical Cancer Behavior Risk Data Set, I conduct 3 models.

1) Logistic regression

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.523232	0.514573	2.960	0.00462	**
behavior_sexualRisk	-0.058853	0.035167	-1.674	0.10023	
behavior_eating	0.030551	0.017519	1.744	0.08709	.
behavior_personalHygine	-0.000195	0.018160	-0.011	0.99148	
intention_aggregation	-0.021454	0.016366	-1.311	0.19564	
intention_commitment	-0.018698	0.016177	-1.156	0.25302	
attitude_consistency	0.069021	0.025548	2.702	0.00929	**
attitude_spontaneity	-0.021806	0.027603	-0.790	0.43313	
norm_significantPerson	-0.031311	0.029431	-1.064	0.29230	
norm_fulfillment	0.000484	0.016664	0.029	0.97694	
perception_vulnerability	0.007709	0.016646	0.463	0.64522	
perception_severity	-0.056776	0.024031	-2.363	0.02192	*
motivation_strength	-0.018871	0.013494	-1.399	0.16789	
motivation_willingness	0.012804	0.014134	0.906	0.36913	
socialSupport_emotionality	0.010425	0.017797	0.586	0.56056	
socialSupport_appreciation	-0.057083	0.027500	-2.076	0.04288	*
socialSupport_instrumental	0.043057	0.016584	2.596	0.01222	*
empowerment_knowledge	-0.023004	0.020239	-1.137	0.26091	
empowerment_abilities	-0.020228	0.021194	-0.954	0.34428	
empowerment_desires	-0.023293	0.014938	-1.559	0.12498	

As we can see above, attitude_consistency, perception_serverity, and social support play a significant role in the model.

2) Random forest classification

```
Call:
 randomForest(formula = ca_cervix ~ ., data = df_sobar)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 11.11%
Confusion matrix:
   0  1 class.error
0 48  3  0.05882353
1  5 16  0.23809524
```

For the random forest model, I got 88.89% overall accurate rate, which is good. The OOB estimate of error rate it 11.11%

3) SVM

For SVM, here is the summary of the model:

```
svm(formula = ca_cervix ~ ., data = training_set, type = "C-classification", kernel = "linear")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Number of Support Vectors:  14

 ( 6 8 )


Number of Classes:  2

Levels:
 0 1
```

Here is the confusion matrix:

```
   y_pred
     0  1
  0 13  0
  1  1  4
```

Overall, I got 94.44% accurate rate, which is better than the result given by the random forest model.

For wine quality data set, I also conduct 3 models.

1) Random forest

```
        OOB estimate of  error rate: 29.4%
Confusion matrix:
  3  4    5    6   7  8 9 class.error
3 0  1   16   11   0  0 0   1.0000000
4 1 23  106   52   1  0 0   0.8743169
5 0  4 1339  417  14  0 0   0.2452086
6 0  3  322 1723 131  3 0   0.2103575
7 0  0   21  316 518  9 0   0.4004630
8 0  0    1   42  52 67 0   0.5864198
9 0  0    0    1   4  0 0   1.0000000
```

Overall accurate rate is 70.6%. Mistakes are very likely to happen when the quality is too high or too low.

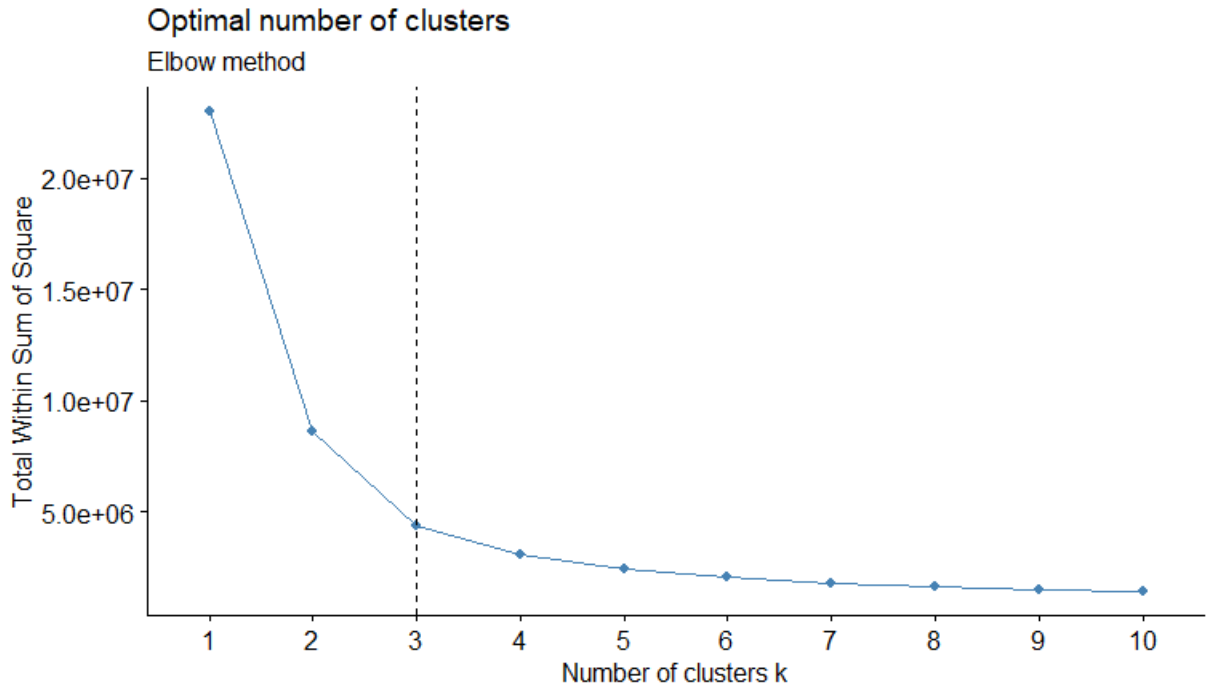
2) KNN

	prc_test_pred					
test_wine_target	4	5	6	7	8	Row Total
3	0	2	0	0	0	2
	0.009	3.488	1.028	0.362	0.022	
	0.000	1.000	0.000	0.000	0.000	0.002
	0.000	0.005	0.000	0.000	0.000	
	0.000	0.002	0.000	0.000	0.000	
4	2	18	13	0	0	33
	22.395	7.472	0.929	5.970	0.356	
	0.061	0.545	0.394	0.000	0.000	0.025
	0.333	0.048	0.019	0.000	0.000	
	0.002	0.014	0.010	0.000	0.000	
5	3	162	171	25	3	364
	1.034	30.447	1.399	25.342	0.217	
	0.008	0.445	0.470	0.069	0.008	0.280
	0.500	0.431	0.256	0.106	0.214	
	0.002	0.125	0.132	0.019	0.002	
6	1	153	368	125	7	654
	1.352	6.962	2.985	0.378	0.000	
	0.002	0.234	0.563	0.191	0.011	0.503
	0.167	0.407	0.551	0.532	0.500	
	0.001	0.118	0.283	0.096	0.005	
7	0	36	103	75	1	215
	0.993	11.058	0.517	33.514	0.749	
	0.000	0.167	0.479	0.349	0.005	0.166
	0.000	0.096	0.154	0.319	0.071	
	0.000	0.028	0.079	0.058	0.001	
8	0	5	13	10	3	31
	0.143	1.759	0.543	3.439	21.272	
	0.000	0.161	0.419	0.323	0.097	0.024
	0.000	0.013	0.019	0.043	0.214	
	0.000	0.004	0.010	0.008	0.002	
Column Total	6	376	668	235	14	1299
	0.005	0.289	0.514	0.181	0.011	

The overall accurate rate is 45.96%. It seems that random forest performs better than this one. In this model, since the number of quality ‘3’ is quality low, the prediction result will miss this result.

3) Kmeans

First, I use elbow function to find the k number. Before that, I subset the data by filtering out the quality list because I want to compare the model result with this column and hope to draw some conclusion from it.



From the chart above, k equals to 3.

df1$cluster	df1$quality 3	4	5	6	7	8	9	Row Total
1	11	54	782	850	208	49	0	1954
	0.433	1.850	30.042	0.010	41.833	1.410	1.504	
	0.006	0.028	0.400	0.435	0.106	0.025	0.000	0.301
	0.367	0.250	0.366	0.300	0.193	0.254	0.000	
	0.002	0.008	0.120	0.131	0.032	0.008	0.000	
2	14	82	579	693	225	25	0	1618
	5.705	14.792	4.071	0.249	7.111	11.068	1.245	
	0.009	0.051	0.358	0.428	0.139	0.015	0.000	0.249
	0.467	0.380	0.271	0.244	0.209	0.130	0.000	
	0.002	0.013	0.089	0.107	0.035	0.004	0.000	
3	5	80	777	1293	646	119	5	2925
	5.357	3.058	35.766	0.206	52.848	11.866	3.357	
	0.002	0.027	0.266	0.442	0.221	0.041	0.002	0.450
	0.167	0.370	0.363	0.456	0.599	0.617	1.000	
	0.001	0.012	0.120	0.199	0.099	0.018	0.001	
Column Total	30	216	2138	2836	1079	193	5	6497
	0.005	0.033	0.329	0.437	0.166	0.030	0.001	

On the left, it is the cluster predicted by kmeans. On the bottom, it's the quality of the wine.

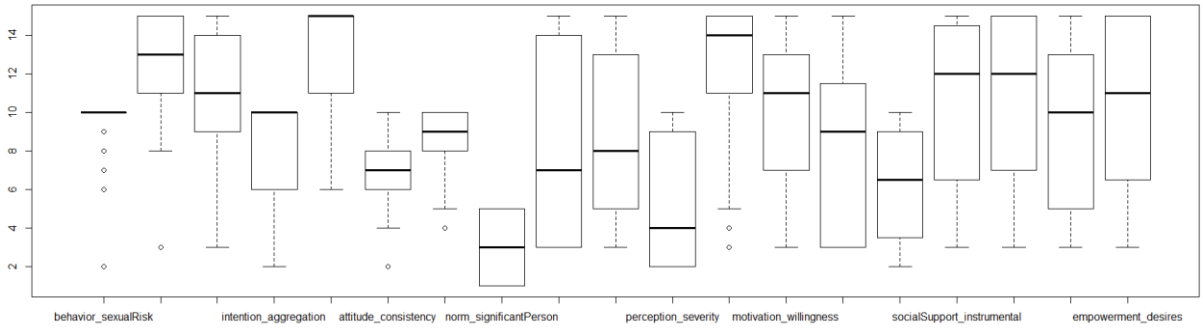
For group one, majority of the records are from low quality to middle level quality.

For group two, most of the records are middle quality.

For group three, most of the records are high quality wine.

3. Decisions (3%)

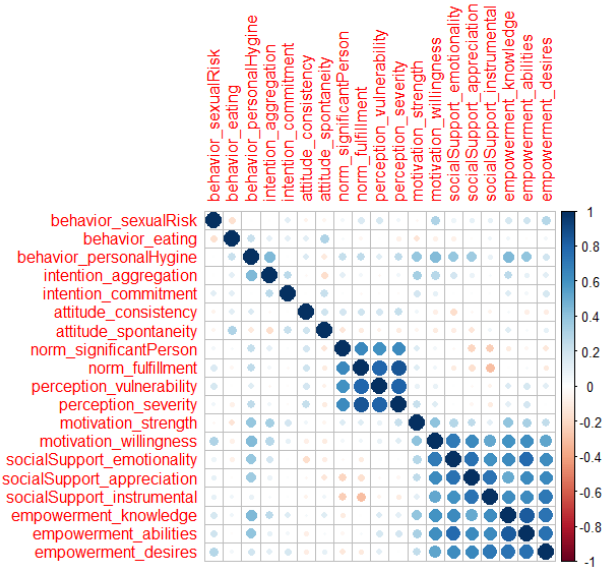
For Cervical Cancer Behavior Risk Data Set, all independent data are numeric and within similar scale – from 4 to 20. Therefore, I can easily conduct classification model without normalization the data.



After comparing the models, the KNN model with accurate rate more than 90% outperform the rest of them.

```
  y_pred
    0  1
0  13  0
1   1  4
```

For the wine data, I first combine two data sets together and plot correlation maps to see the relationship within the data.



The data within the same or similar categories have strong correlation.

I conduct several supervised models. At last, the random forest model beats the KNN by 70% accurate rate.

Meanwhile, I also conduct an unsupervised model and turned out to be interested.

	df1$quality							
df1$cluster	3	4	5	6	7	8	9	Row Total
1	11	54	782	850	208	49	0	1954
	0.433	1.850	30.042	0.010	41.833	1.410	1.504	
	0.006	0.028	0.400	0.435	0.106	0.025	0.000	0.301
	0.367	0.250	0.366	0.300	0.193	0.254	0.000	
	0.002	0.008	0.120	0.131	0.032	0.008	0.000	
2	14	82	579	693	225	25	0	1618
	5.705	14.792	4.071	0.249	7.111	11.068	1.245	
	0.009	0.051	0.358	0.428	0.139	0.015	0.000	0.249
	0.467	0.380	0.271	0.244	0.209	0.130	0.000	
	0.002	0.013	0.089	0.107	0.035	0.004	0.000	
3	5	80	777	1293	646	119	5	2925
	5.357	3.058	35.766	0.206	52.848	11.866	3.357	
	0.002	0.027	0.266	0.442	0.221	0.041	0.002	0.450
	0.167	0.370	0.363	0.456	0.599	0.617	1.000	
	0.001	0.012	0.120	0.199	0.099	0.018	0.001	
Column Total	30	216	2138	2836	1079	193	5	6497
	0.005	0.033	0.329	0.437	0.166	0.030	0.001	

The model results can clearly cluster the data into high quality, middle quality and low quality groups.