Assignment #7

Syamala Srinivasan

Introduction:

The factor analysis of liquor preferences conducted by Jean Stoetzel illustrates how factor analysis can be utilized to analyze variables that are difficult to quantitate such as preference. The sample population of this study includes a cross-section of the French population in 1956. These individuals were asked which type of liquor they liked best and to rank them from best to least. These liquors included; "Armagnac, Calvados, Cognac, Kirsh, Marc, Mirabelle, Rum, Whiskey, and Liqueurs.

Factor analysis is utilized to group these liquors together based on correlations that can be categorized into factors. In Stoetzel's paper the factors are stated to be strength of the liquor, price and local preference.

Results:

Task 1:

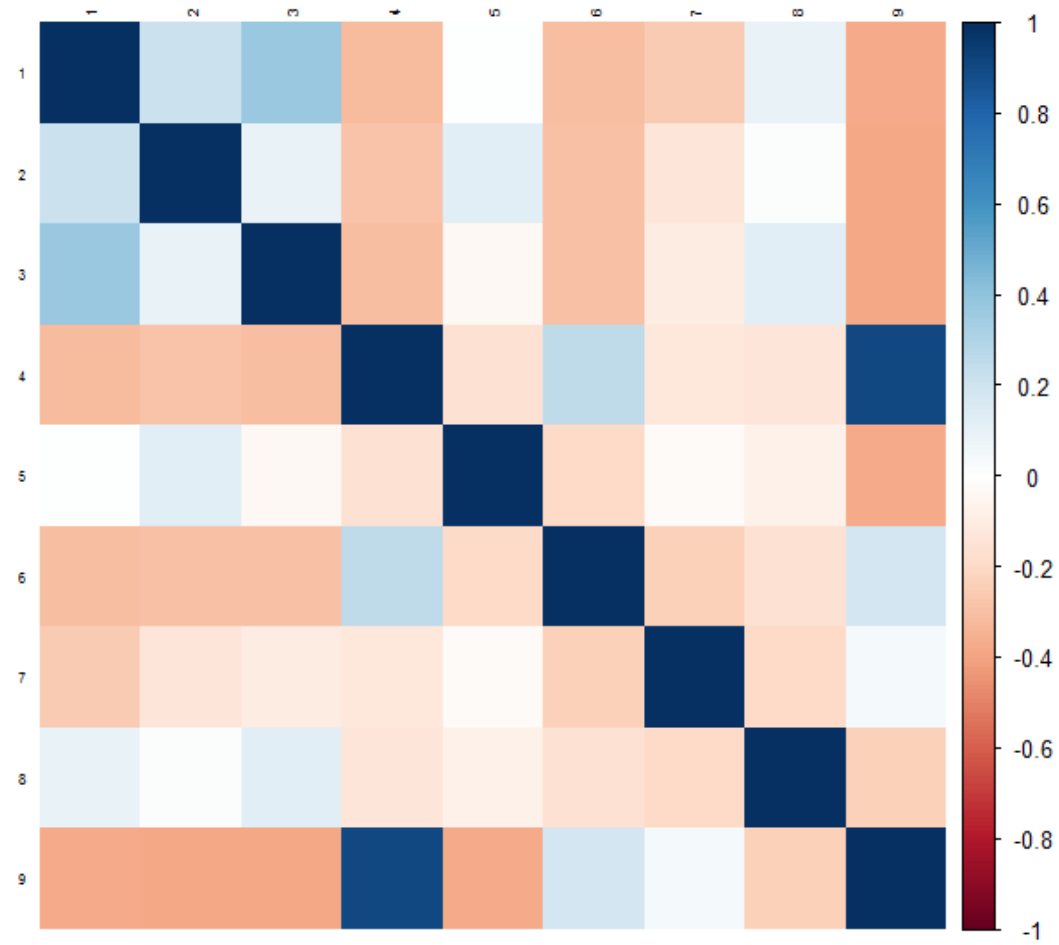
To begin the analysis the correlation matrix must be loaded into R. This is initially created by created a vector with the correlation values.

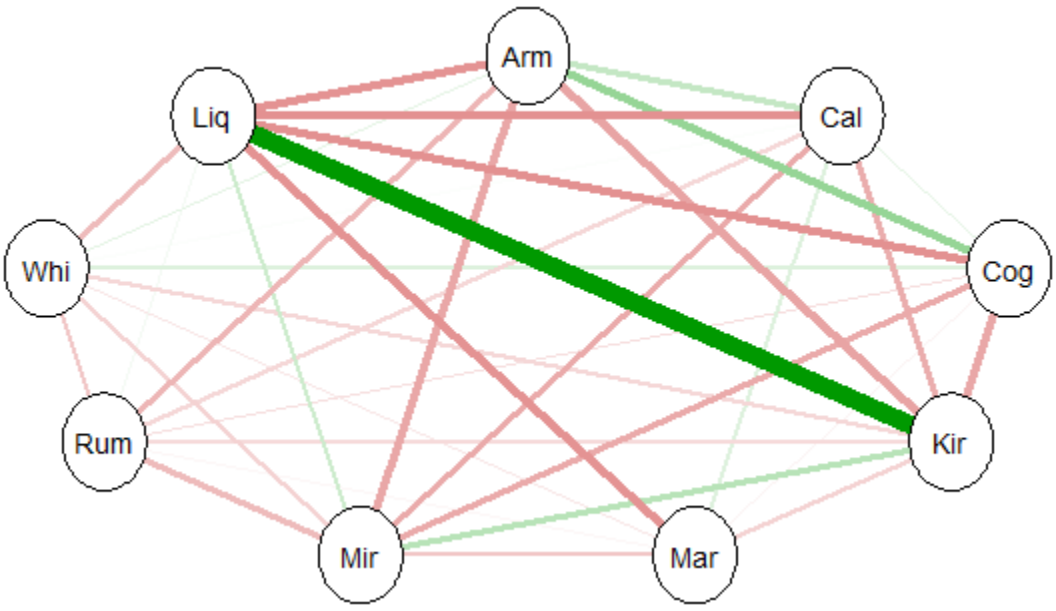
1.000	0.210	0.370	-0.32	0.000	-0.31	-0.26	0.090	-0.38
0.210	1.000	0.090	-0.29	0.120	-0.30	-0.14	0.010	-0.39
0.370	0.090	1.000	-0.31	-0.04	-0.30	-0.11	0.120	-0.39
-0.32	-0.29	-0.31	1.00	-0.16	0.25	-0.13	-0.14	0.900
0.00	0.120	-0.04	-0.16	1.000	-0.20	-0.03	-0.08	-0.38
-0.31	-0.30	-0.30	0.25	-0.20	1.000	-0.24	-0.16	0.180
-0.26	-0.14	-0.11	-0.13	-0.03	-0.24	1.000	-0.20	0.040
0.090	0.010	0.120	-0.14	-0.08	-0.16	-0.20	1.000	-0.24
-0.38	-0.39	-0.39	0.900	-0.38	0.180	0.040	-0.24	1.000

This matrix once transformed in R Studio to include the headers of the various types of liquors is outputted to look like:

	Arm	Cal	Cog	Kir	Mar	Mir	Rum	Whi	Liq
Arm	1.00	0.21	0.37	-0.32	0.00	-0.31	-0.26	0.09	-0.38
Cal	0.21	1.00	0.09	-0.29	0.12	-0.30	-0.14	0.01	-0.39
Cog	0.37	0.09	1.00	-0.31	-0.04	-0.30	-0.11	0.12	-0.39
Kir	-0.32	-0.29	-0.31	1.00	-0.16	0.25	-0.13	-0.14	0.90
Mar	0.00	0.12	-0.04	-0.16	1.00	-0.20	-0.03	-0.08	-0.38
Mir	-0.31	-0.30	-0.30	0.25	-0.20	1.00	-0.24	-0.16	0.18
Rum	-0.26	-0.14	-0.11	-0.13	-0.03	-0.24	1.00	-0.20	0.04
Whi	0.09	0.01	0.12	-0.14	-0.08	-0.16	-0.20	1.00	-0.24
Liq	-0.38	-0.39	-0.39	0.90	-0.38	0.18	0.04	-0.24	1.00

The correlation matrix can also be viewed as a plot.





Utilizing the correlation matrix and the corresponding plots, there is one area of high positive correlation that is immediately noticeable which is the correlation between Kirsh and Liquers. There is also a relatively high negative correlation between Liquers and any of Armagnac, Calvados, Cognac, and Marc.

Task 2:

In Stoetzel's paper it is estimated to be a three factor model. The factors in Stoetzel's paper are

- Factor 1: Strong to sweet
- Factor 2: Expensiveness of the liquer
- Factor 3: Local or national preference

Running the matrix in R provides us with the same, three factors.

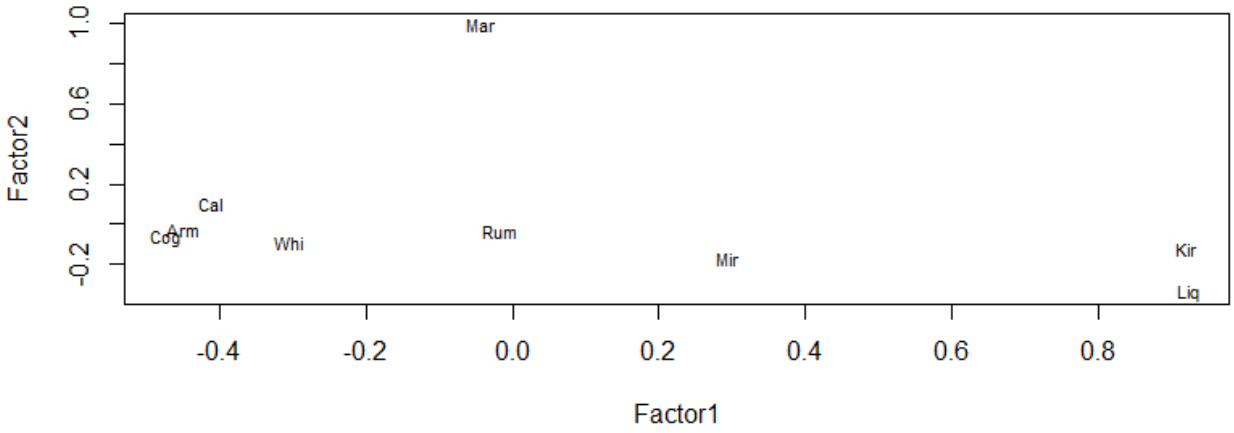
For factor 1 there are two variables with very high loadings and a few medium loadings. We could interpret this to mean that there are two very sweet wines while the rest are somewhere between strong and sweet.

For factor 2 there is a very high positive loading for Marc while every other liquer is in a low to medium range. This could indicate that Marc is significantly more expensive than other liquers.

Factor 3 has a very high negative loading for Mitrabelle while the others are very low which could indicate that Mitrabelle is a local favorite.

```
Loadings:
      Factor1 Factor2 Factor3
 [1,] -0.450           0.193
 [2,] -0.411   0.100   0.172
 [3,] -0.473           0.183
 [4,]  0.921  -0.121
 [5,]          0.996
 [6,]  0.293  -0.169  -0.938
 [7,]                  0.256
 [8,] -0.305
 [9,]  0.923  -0.344   0.158
```

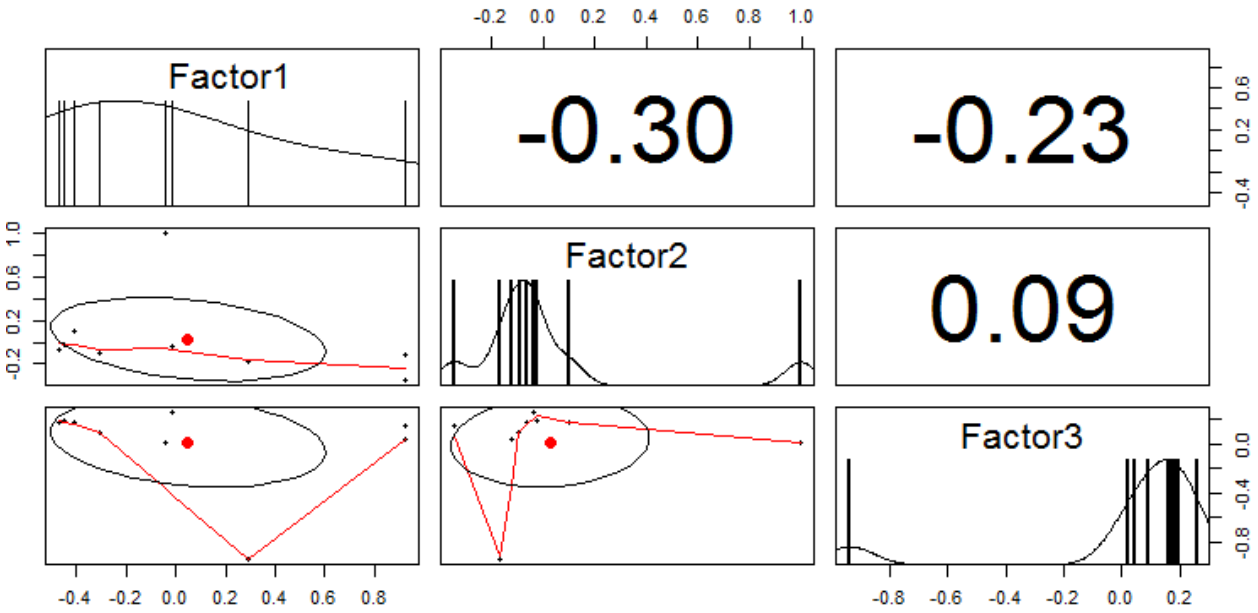
This balance between sweet/strong, expensive/cheap and preference can also be illustrated in a plot. Below we can compare factor 1 and 2. We can make some assumptions from this graph such as Kirsh and Liquers being seen as strong and cheap.



Of course, we can also view all the factors compared to each other, not just factors 1 and 2. This is illustrated below in the scatterplot matrix.

In the below scatterplot matrix there are a few relationships that stand out. For factor 2 (price) we can see there is a strong negative relationship between price and how strong a liquor is; as well as with preference, in that price is affected by how much of a local preference there is.

We can also see that there is a medium negative relationship between factors 1 and 2 and factors 1 and 3 while there is a small positive relationship between factors 2 and 3.



Utilizing the sums of squares of the loading values for each factor we can determine how much variance is explained by each factor. This is illustrated below using the SS loading.

Also illustrated below is the cumulative variance for the factors. Utilizing the below, we can see that 53% of the variation in the model with all three factors is explained by the three factors.

```
                Factor1 Factor2 Factor3
SS loadings       2.477   1.179   1.082
Proportion Var    0.275   0.131   0.120
Cumulative Var    0.275   0.406   0.527
```

The values obtained are slightly different than those in the paper as indicated in the below table.

Item	Factor 1	Factor 1 (Paper)	Factor 2	Factor 2 (Paper)	Factor 3	Factor 3 (Paper)
Armagnac	-0.450	-0.60	-0.22	-0.17	0.193	0.14
Calvados	-0.411	-0.52	0.100	-0.03	0.172	0.42
Cognac	-0.473	-0.52	-0.06	-0.03	0.183	0.42
Kirsch	0.921	0.50	-0.121	-0.06	0.043	-0.10
Marc	-0.04	-0.29	0.996	0.66	0.019	-0.39
Mirabelle	0.293	0.46	-0.169	-0.24	0.938	-0.19
Rum	-0.01	0.17	-0.036	0.74	0.256	0.97
Whisky	-0.305	-0.29	-0.094	-0.08	0.091	0.09
Liqueurs	0.923	0.64	-0.344	0.02	0.158	0.16

There are many reasons why two factor analyses performed on the same data would result in different outcomes. One reason could be that a different rotation style was utilized by each analysis.

To test that three factors is a sufficient number of factors we will have a null hypothesis that there is no difference if we change from a three factor model to a four factor model.

```
Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 1820.72 on 12 degrees of freedom.
The p-value is 0
```

```
Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 968.7 on 6 degrees of freedom.
The p-value is 5.26e-206
```

Looking at the above outputs we can reject our null hypothesis. We can do this because the chi square statistic is smaller.

Task 3:

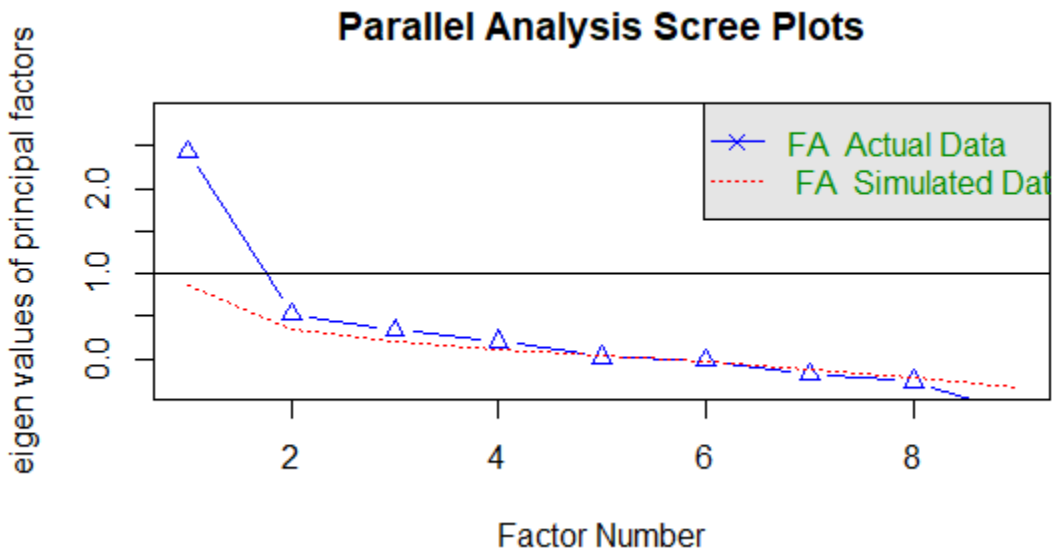
Looking at the below table we can see that there is a trade off between the significance of the chi square statistic and the percent of variation that is explained. The more factors added, the less significant the chi square value is however more of the variation is explained by the model. This would indicate that perhaps 3 is a sufficient number of factors due to it being a good middle ground for both.

Number of Factors	Chi Square Statistic Value	Percent of variation explained	MAE
1	2972.05	27%	.0862

2	2448.36	39%	.0705
3	1820.72	53%	.0486
4	968.7	60%	.0325
5	674.73	70%	.0200
6	N/A	N/A	.0200

Another method to determine the numbers of factors to utilize is parallel analysis which outputs a recommendation of four factors.

```
Parallel analysis suggests that the number of factors =  4
```



In the above graph it is illustrated that the simulated data reaches the actual data after 4 factors. According to the above graph the max number of factors to be considered would be 5.

Task 4:

The last model created utilized the rotation method of varimax, for the next model the rotation method promax will be used instead which is oblique.

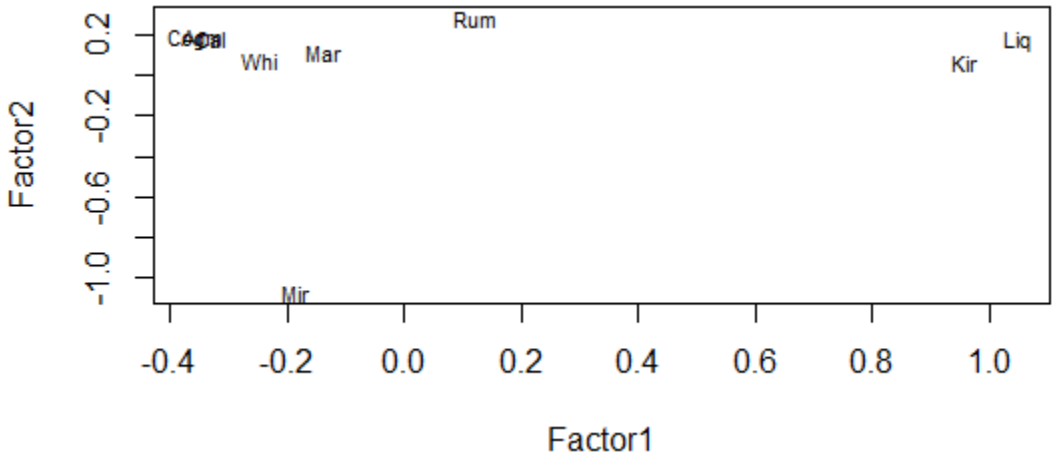
When this model is created also with three factors we can see that the promax model is very different from the original variamx model and that more of the variation of the model is explained for the second however, the MAE for the second model is higher at .1103 indicating that this model is not as good.

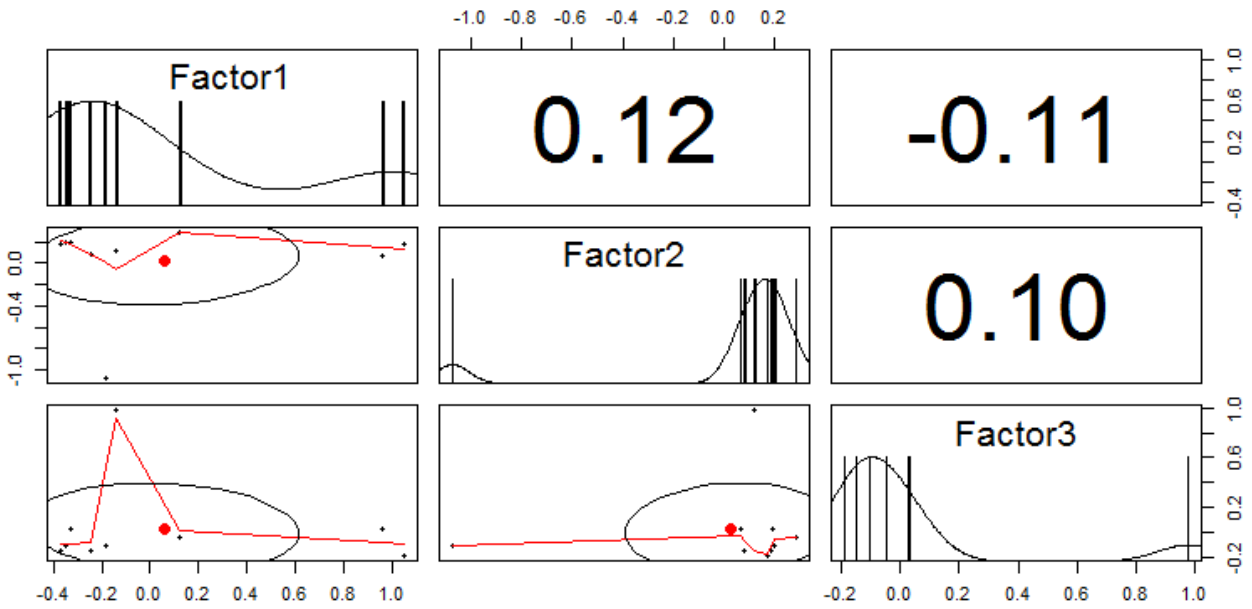
```
Loadings:
      Factor1 Factor2 Factor3
 [1,] -0.347   0.202  -0.102
 [2,] -0.331   0.193
 [3,] -0.371   0.186  -0.146
 [4,]  0.961
 [5,] -0.139   0.123   0.977
 [6,] -0.186  -1.075  -0.101
 [7,]  0.123   0.287
 [8,] -0.248          -0.146
 [9,]  1.047   0.175  -0.186

                Factor1 Factor2 Factor3
SS loadings      2.518   1.407   1.057
Proportion Var   0.280   0.156   0.117
Cumulative Var   0.280   0.436   0.554
```

When comparing the factors in the second model it is clear that there are defined relationships for each factor combination. This model has better interpretability than the varimax rotation.

An interesting point to highlight is that most liquors fall into factor one with only two favoring factor 2 and one in factor 3. There are also much higher correlations in this model than the original varimax model.

Using the plot of factor 1 and factor 2 again as a comparison, it is noticeable that there are much tighter groupings than before. Kirsh and liquers have a relationship with being very strong and cheap while Mirabelle is alone being very expensive and sweet. The remainders can be considered cheap and weak.





Task 5:

To assist in determining how many factors should be considered in the model, the mean absolute errors will be considered.

Number of Factors	MAE
1	.0861
2	.0816
3	.1103
4	.1472
5	.1239
6	.1284

For each factor added to the model the error experienced is increased.

The mean absolute errors of the second model are significantly higher than the original model indicating that the first model may be the better fit between the two.

Conclusions:

Factor analysis is crucial to understanding trends in data that may not easily produce an original hypothesis. The creation of factors helps to consolidate variables into interpretable trends. One of the hardest parts of factor analysis is determining how many factors to have in your model, this can be determined in many ways and there really is no wrong answer.