Homework 3

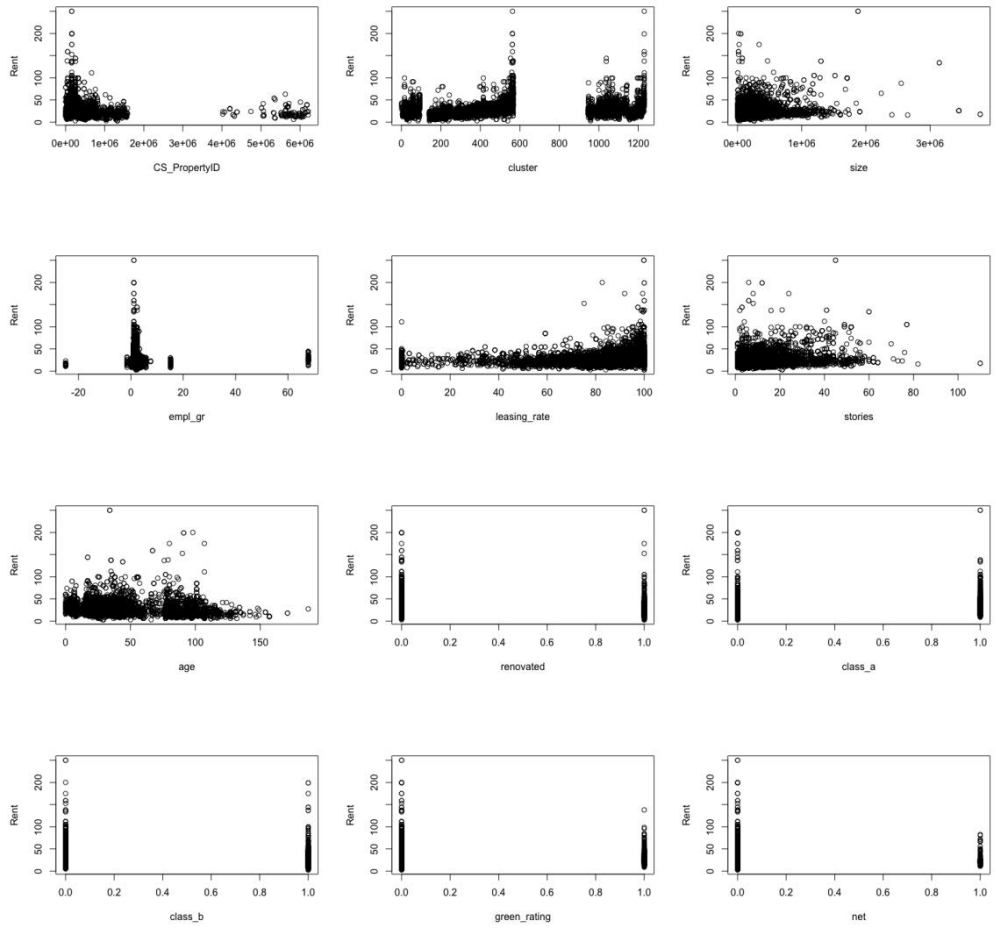
Xingya Wang

1. Predictive model building

1.1 To build the best predictive model possible for price

How could we best predict **Rent**- which represents price in this data set? To be able to answer this question, there are several different methods we could use, including KNN, trees, linear regression and so on. And I would like to predict the price under the linear model, using **step** methods.

First, see how the variables work in predicting **Rent** using all the variables except **LEED** and **Energystar**. Since the variable **green_rating** contains all the buildings with **LEED** or **Energystar** together, which also means "green certificated" mentioned in the question, we do not have to calculate the same content variables twice. The pictures are in Figure 1.



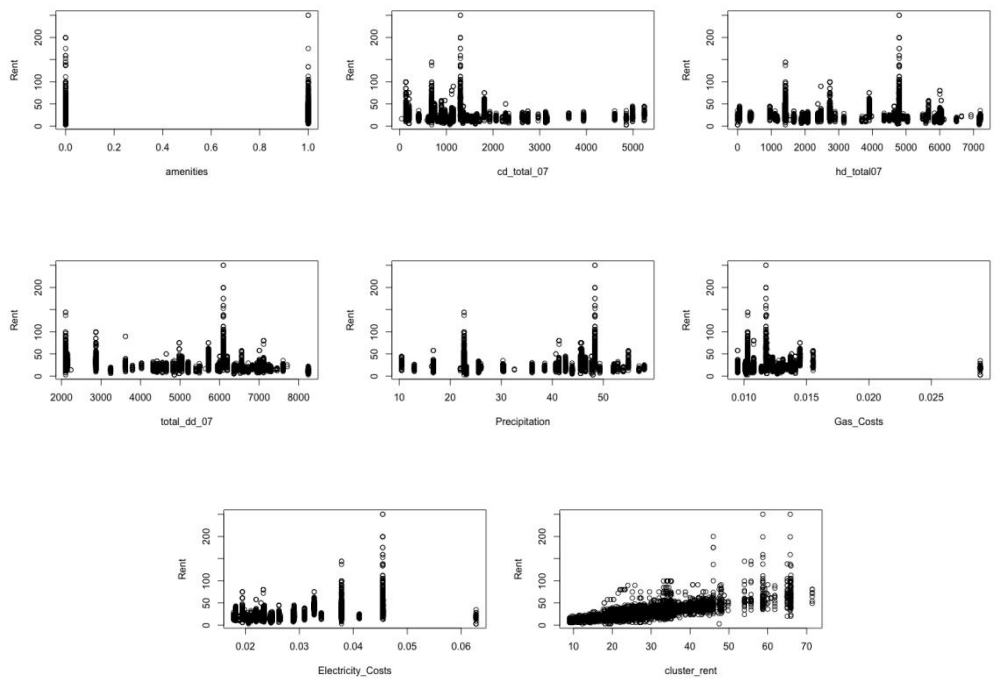


Figure 1- single variable versus **Rent** plots

From the figure above, it's hard to say if each off them does not relate to **Rent** at all, thus, I could not delete each of them unreasonably.

Second, setting a simple linear model and then using this linear regression to calculate a new **step** model under squared basics. **Getcall** the model and then name it as **lm_best**.

Finally, see how it works, or to say see if it improves the prediction of price. I've concluded the **summary** results in Table 1 below.

Models	Residuals					R^2 values	
	Min	1Q	Median	3Q	Max	R-squared	Adjusted R-squared
lm1	-53.765	-3.581	-0.530	2.483	173.892	0.6125	0.6115
lm_best	-55.527	-3.530	-0.640	2.726	148.465	0.6548	0.6514

Table 1- summary results of two models

As a conclusion, some values in residuals do not vary too much, but the max residual has decreased a lot in **step** model. To be more specific, we could just compare the R^2 values(note $0 \leqq R^2 \leqq 1$, and R^2 is greater means the model works better), and we have a kind of significant improve here.

Thus, the **lm_best** model, is my best predictive model possible for price.

1.2 to use this model to quantify the average change in rental income per square foot (whether in absolute or percentage terms) associated with green certification, h.a.e.f.

How could we estimate how much will the rental income per square change associated with green certification? We need to find all the variables which combine with green certification, and then calculate the change in **Rent**, assume other variables are at same value, and only change green certification from 0 (means don't have) to 1 (have this certification).

Green certification has the same meaning as **green_rating**, where **green_rating**=1 means the building has green certification and doesn't have if **green_rating**=0.

The estimate slopes related to green certification in the best model are listed below in Table 2.

Variable	Estimate slope
green_rating	1.105
green_rating*amenities	-2.006
green_rating*age	0.03513

Table 2- Coefficient result related to **green_rating**

According to the coefficient result of **lm_best**, while we holding other values fixed, if the **green_rating** level increase from 0 to 1, the total improvement in the price per square will be the formula:

1.105-2.006*amenities+0.03513*age

To get a specific number, we use the mean age and amenities here, which are 47.24 and 0.5266, then the per square change in rental income is 1.105-1.056+1.660=1.709 in dollars per square foot per calendar year.

2. What causes what

2.1 Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime?

According to the listening material, the main problem here is that it's hard to determine "what causes what". It would be reasonable to hire more polices in a city due to the large quantity of crimes, thus could not be the useful data in predicting "would more police decrease the number of crimes".

2.2 How were the researchers from UPenn able to isolate this effect?

To isolate this effect, the researchers find the terrorism alert system which could cause a increase in police and do not relate to crime. While it is under high-alert, their will be more police, and researchers could find how the crimes being affected under this situation.

From the table, it is easy to find that under high-alert, crime would decrease. And if we combine it with ridership, it shows that high-alert also contributes to crime decrease and the absolute value of slope number does not decrease too much, and also R^2 value has increased. Thus, it confirms the police increase will decrease the number of crimes, holding ridership the same. And for ridership, it shows greater importance in predicting daily crime than high-alert. It represents if there's more ridership, it would be higher daily crimes, holding all else fixed.

2.3 Why did they have to control for Metro ridership? What was that trying to capture?

Since there exists the possibility that the number of crimes decrease due to less victims on the street during high-alert times, so we have to control for Metro ridership to see how ridership changes.

And it trying to capture if ridership really be influenced by high-alert and then to predict if the ridership is almost the same, what will happen to crime numbers with or without high-alert. All these methods are trying to leave "police" as the only variable in predicting "crime".

2.4 Can you describe the model being estimated here? What is the conclusion?

The model here is trying to estimate how first police district area and other districts at high-alert days combine with the daily ridership would influence the daily number of crime.

As a conclusion, though it's true that more ridership leads to more crimes,however, at high-alert days, the first police district area would decrease more crime than other districts, holding ridership fixed. And it confirms that more police could help in diminishing crime.

3. Clustering and PCA

3.1 Which dimensionality reduction technique makes more sense to you for this data?

How could we find the perfect dimensionality reduction method which could also distinguish reds from white? To be able to compare the two methods (clustering and PCA), we need to do these two methods first, then to see the plots to find our final choice.

3.1.1 K-means

There are several clustering methods including K-means, H-cluster and so on. Here I'd like to use K-means clustering to see how the different K value works on this data.

First, we need to do some prepare work for the data, such as central and scale the data, then we just run different K value cluster, and see how it translate the data set. Here is a table contains K-means with n cluster and between_SS / total_SS value(note: we want this number to be as great as it could be to best represent the data).

N cluster	between_SS / total_SS	N cluster	between_SS / total_SS
2	21.4%	6	49.3%
3	36.2%	7	52.4%
4	43.0%	8	54.2%
5	46.7%	9	55.8%

Table 3- K-means with n cluster and its between_SS / total_SS values

It's obviously that while K-means cluster = 9, it could represent the data better. However, if we plot, the pictures , it seems that while cluster=2 in K-means method, this method works better in distinguishing reds from whites. Figure 2 shows how it works.

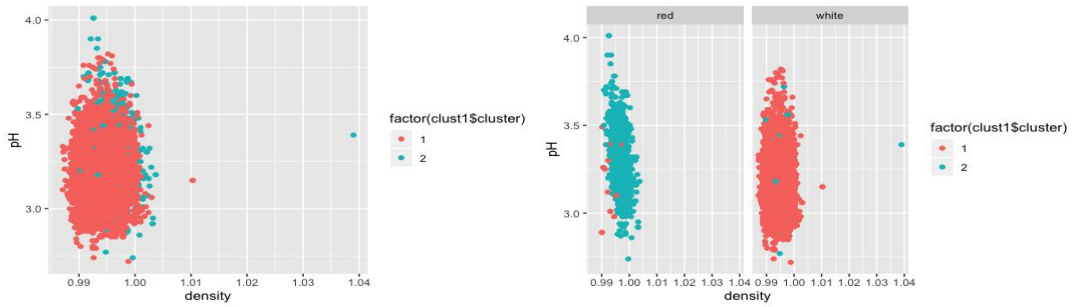


Figure 2- pH versus density plot with and without facet in wine color

Comparing these two pictures above, the points really mixed up in the first picture, and after showing under different wine colors, it nearly perfect concentrate in

different cluster numbers. We could run this plot using all the 11 chemical elements contains in the data, and the results are almost the same.

However, given the problem that while with 9 cluster K-means could contains more original data, we need to see how this works in distinguishing reds from whites. Let us focus on Figure 3 to see what happens.

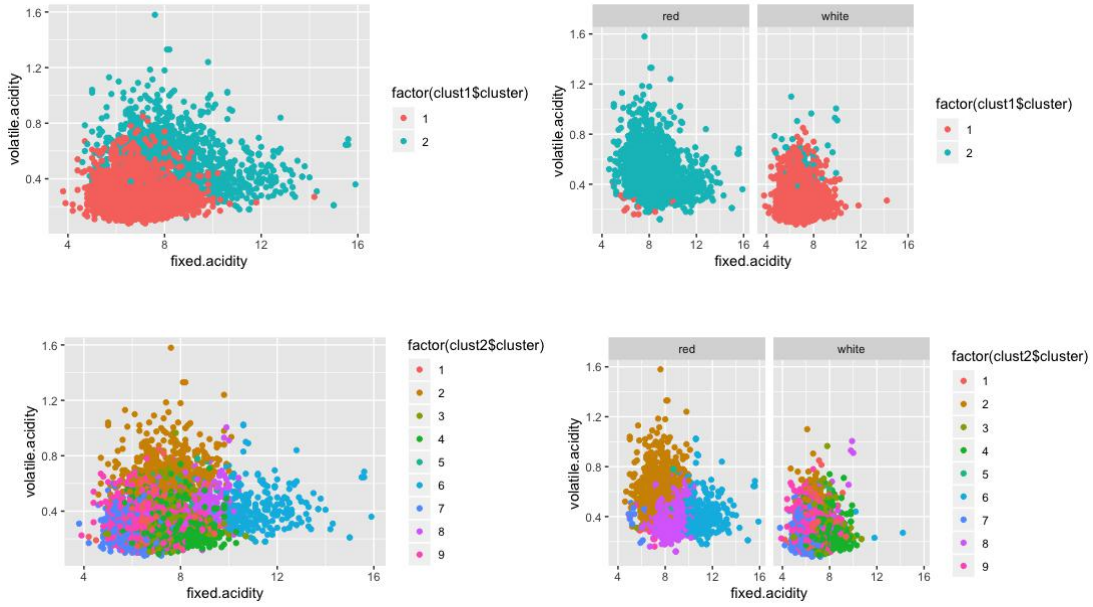


Figure 3- fixed versus volatile acidity K-means plots under 2 and 9 clusters

It is really obvious that we could not see what happens in the mess points, let alone distinguish reds from whites.

As a conclusion, we could distinguish reds and rights only use the 11 chemical properties, however, we have to give up the higher accuracy to let cluster = 2 in the K-means clustering.

3.1.2 PCA

Now we could try if PCA works better on this data set. Run simple codes and some of the importance of components are list in Table 4. (Only list the first 6 column in the 11 PCs)

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.7407	1.5792	1.2475	0.98517	0.84845	0.77930
Proportion of Variance	0.2754	0.2267	0.1415	0.08823	0.06544	0.05521
Cumulative Proportion	0.2754	0.5021	0.6436	0.73187	0.79732	0.85253

Table 4-Tmportance of components PCA with different numbers

It shows that higher principle component has lower PVE. Then, we just use the simple PCA to plot pictures. See the points plot on fixed and volatile acidity plot in

Figure 4.

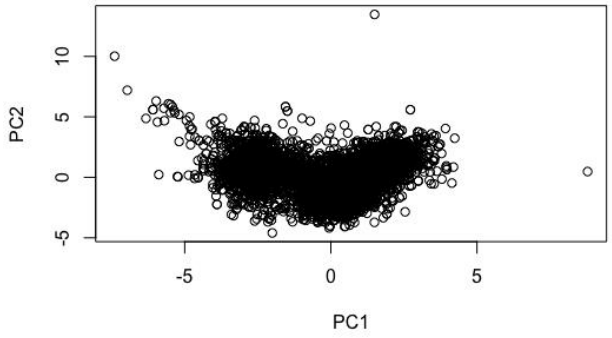


Figure 4-PC1 versus PC2 plot

It is important to notice here that I'm using the same variables which could not do well in cluster=9, and now we using different color to separate different wine colors in Figure 5.

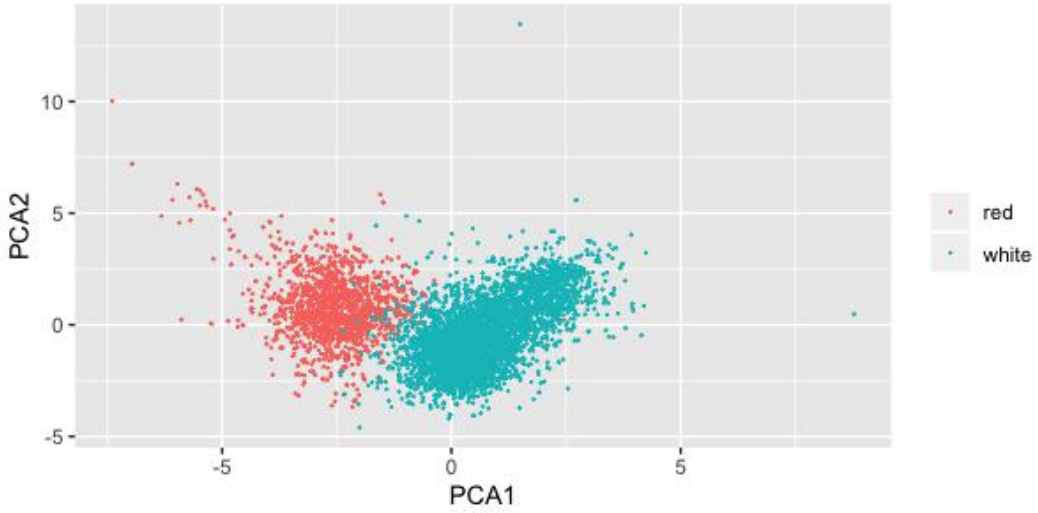


Figure 5-PC1 versus PC2 plot with wine colors

Since the absolute value of numbers are really small here, I use small point size in this picture to see what really plots here. From this picture, we can see that we could almost distinguish reds from whites: there are only several points mixed with each other here.

As a conclusion, PCA could do well in demensonality reduction and could help us in find reds and whites.

Using the "unsupervised' information, it seems that PCA works better in both accuracy and distinguishing wine colors.

3.2 Does this technique also seem capable of sorting the higher from the lower quality wines?

As a guess, I think this technique also seems capable to distinguish the different quality wines. How to prove this? Again, we run the same code using quality as the facet variable.

Since "quality" varies from 1 to 10 in the statement, however, there is no specific high quality and low quality standard level in the statement. Thus, to be more reasonably, I use the mean value here, which is 5.818, to distinguish higher quality from lower quality. (That is, when quality >5.818, it is high quality, and is "true" in the figure below.)

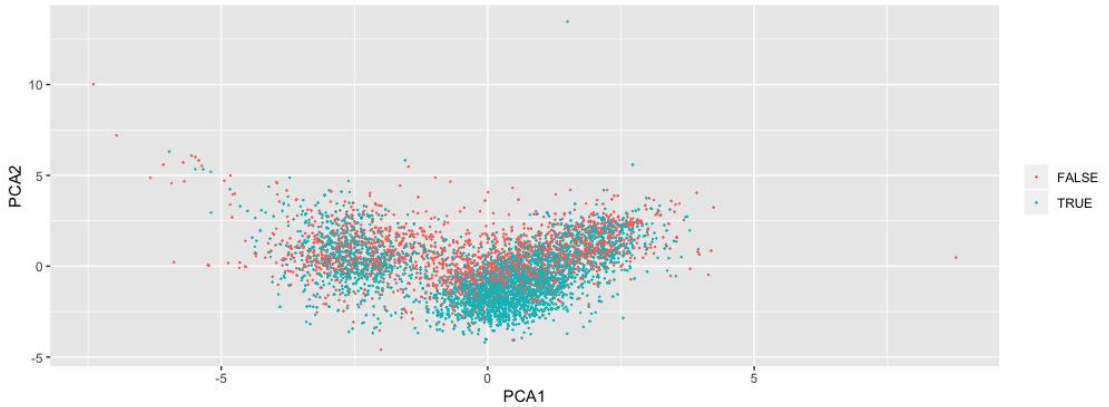


Figure 6-PCA plot with different wine quality

It is not as what I expected before, the picture really messed up, and could not tell the difference between different quality wines.

As a conclusion, it is not the same as my own point view, which means we couldn't sort the higher from the lower quality wines.

4. Market segmentation

To analyze the data and give your client some insight as to how they might position their brand to maximally appeal to each market segment.

We do not need to predict a single aimed value, instead, we should find the co-relationship between the variables to help my client giving them some insight as to how they might position their brand to maximally appeal to each market segment.

First, we need to know what we want to get from the data set here. As a summary of the data set, there are 36 different categories with integer numbers for total 7882 number of users. And we try to catch up the relationship of individuals in reading habits, that is: to find which the one who is interest in x will interest in other categories. It will help the brand to position correctly, to make related ads which could appeal the readers, then to gain customers.

After trying several models, PCA seems the best in working on this data set, and lets see how it plots figures below.

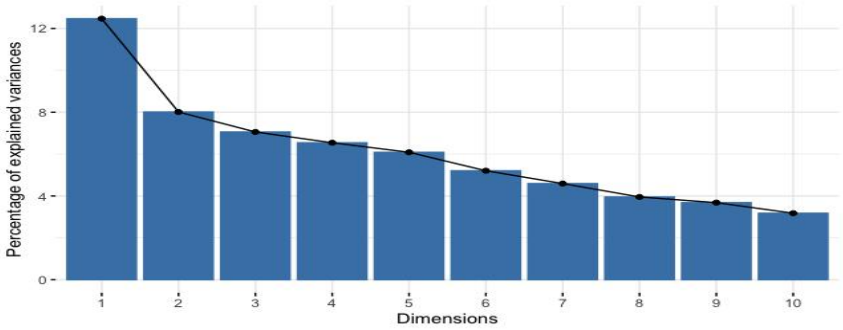
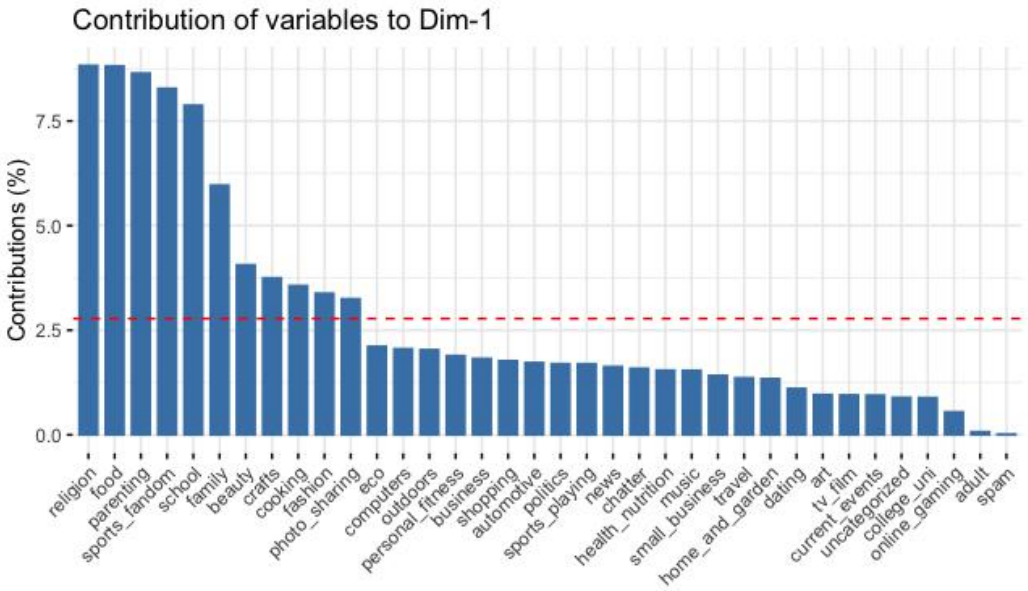


Figure 7- Proportion of var versus Dimension plot



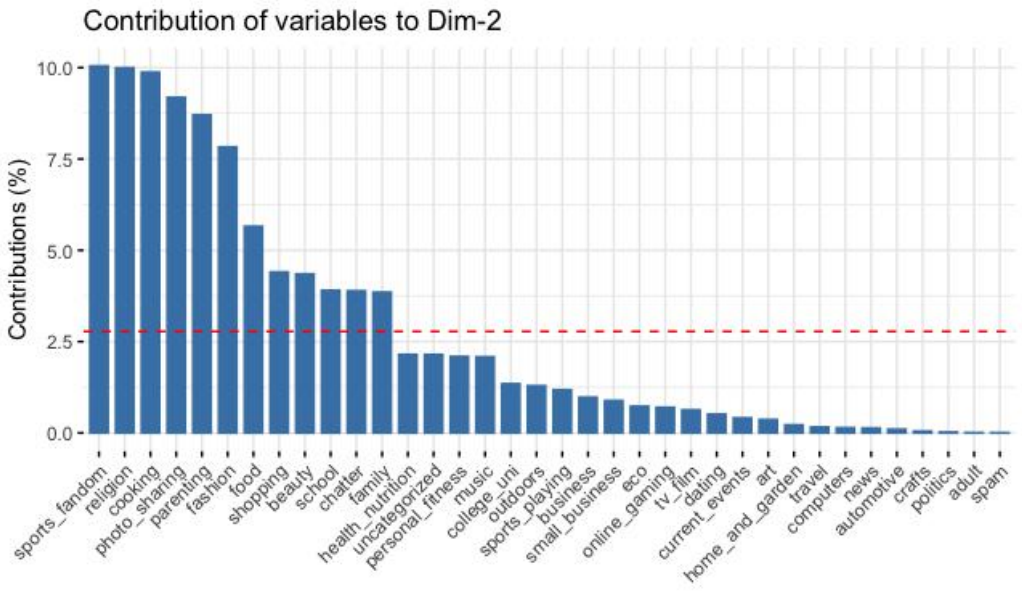


Figure 8-Contributions of each factor in Dim1&2

The factor "spam" and "adult" do not contribute too much in all the 36 variables, thus we do not have to worry about the "not accuracy" and "not allow" problems.

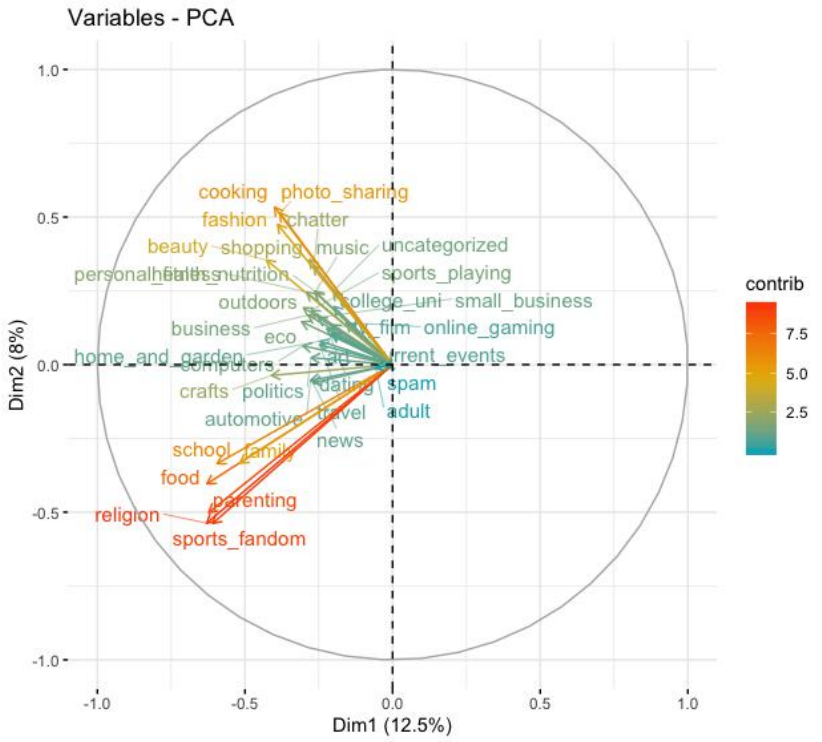


Figure 9- the plot of 36 variable names with different contributions

From Figure 9, the variables which are highly correlated with each other are in the same/similar colors and also changing in the same direction, which could be translate into the table below(without "spam" and "adult" and "uncategorized", since they do not value to our client). Since we are brand orientation, we only need to focus on the

	Categories		Categories
Family concern	Sports fandom	Business	Business
	Religion		Eco
	Parenting		Small business
	Food	Students	Online gaming
	School		College uni
	Family		Computers
Female Channel	Cooking	Art	Art
	Fashion		Crafts
	Photo sharing		Film
	Beauty		Home and garden
	Chatter	Going out	Automotive
	Shopping		Travel
Health	Personal fitness		Dating
	Music	News	Politics
	Health nutrition		News
	Outdoors		Current events
	Sports playing		

Table 5- Summary of categories

The above shows a brief summary of the categories, and it might not be really accurate, since it's hard to distinguish the variables between each other.

What else is that "family concern" and "female channel" contributes more to the Dimensions 1&2 above, and there is also more accuracy in their close relationship, which could help our client in their brand position.

As a conclusion, the company could find their aimed customers by posting ads on related websites above, and it seems the companies which focused on family and females could find their aimed consumer easily than others.