# MATH 308 Assignment 7

A solar flare is a sudden flash of light observed on the disk or edge of the Sun that emits a huge amount of energy up to 6 × 1025 joules. X-rays and ultraviolet radiation emitted by flares can affect the Earth's ionosphere and disrupt long-range radio communications. Radio wave radiation at decimeter wavelengths can directly interfere with the operation of radar and instruments and equipment that use these wavelengths. In this report, we study the characteristics of active regions on the sun where flares are not observed by considering features such as activity changes and evolution. The dataset can be found on http://archive.ics.uci.edu/ml/datasets/solar+flare (http://archive.ics.uci.edu/ml/datasets/solar+flare).

Variables in the dataset:1. class: Code for class (modified Zurich class) (A,B,C,D,E,F,H) 2. largestSpotSize: Code for largest spot size (X,R,S,A,H,K) 3. spotDistri: Code for spot distribution (X,O,I,C) 4. Activity: Activity (1 = reduced, 2 = unchanged) 5. Evolution: Evolution (1 = decay, 2 = no growth, 3 = growth) 6. pre24hActive: Previous 24 hour flare activity code (1 = nothing as big as an M1, 2 = one M1, 3 = more activity than one M1) 7. histComplex: Historically-complex (1 = Yes, 2 = No) 8. beHistComplex: Did region become historically complex on this pass across the sun's disk (1 = yes, 2 = no) 9. area: Area (1 = small, 2 = large) 10. areaLargestSpot: Area of the largest spot (1 = <=5, 2 = >5) And another 3 column for predicted variables.

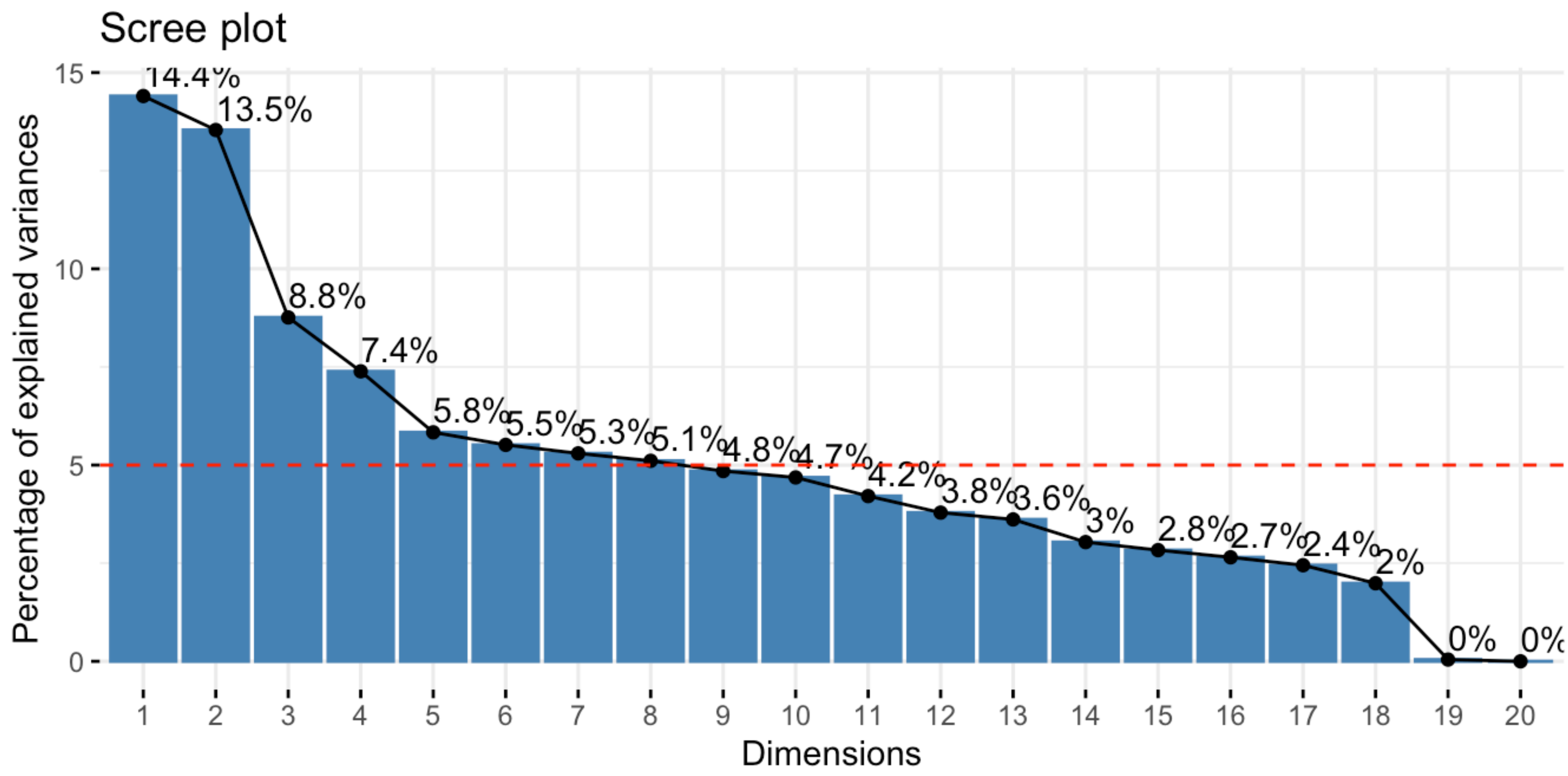
Since we are focusing on the solar-free regions, we only take the first 8 variables and remove the remaining variables. We proceed with the multiple correspondence analysis.

```
res.mca <- MCA(flare, graph = FALSE)
eig.val <- get_eigenvalue(res.mca)
head(eig.val)
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1  0.3600569        14.402276                    14.40228
## Dim.2  0.3384220        13.536880                    27.93916
## Dim.3  0.2190422         8.761687                    36.70084
## Dim.4  0.1847038         7.388151                    44.08899
## Dim.5  0.1458637         5.834548                    49.92354
## Dim.6  0.1378716         5.514862                    55.43841
```

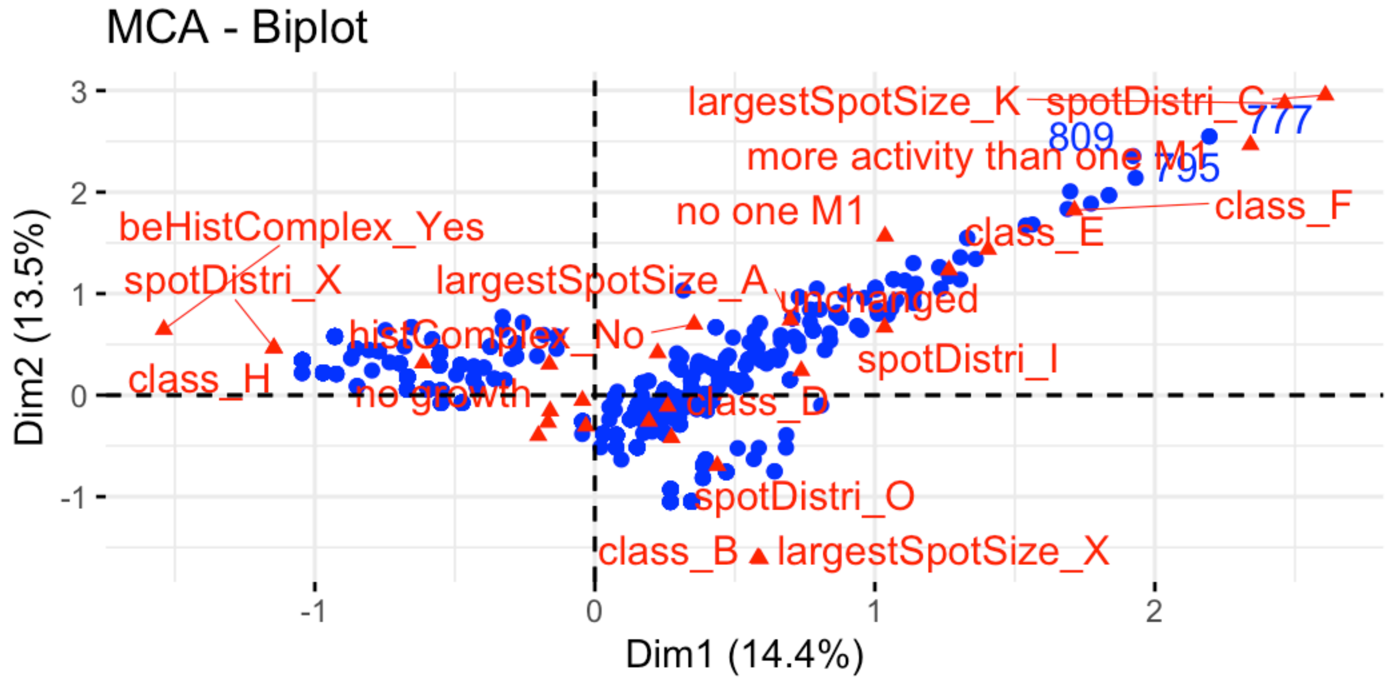
We can better see the percentages of inertia explained by each MCA dimensions with a scree plot.

```
# our data has 865 rows, 31 colomns
fviz_screeplot(res.mca, addlabels = TRUE,ncp=30)+
  geom_hline(yintercept=100/20,linetype=2,color="red")
```



We find that the first two dimensions retain 27.94% of the total inertia in the data. The elbow rule indicates that it is sufficient to represent the data with the first two or three dimensions.

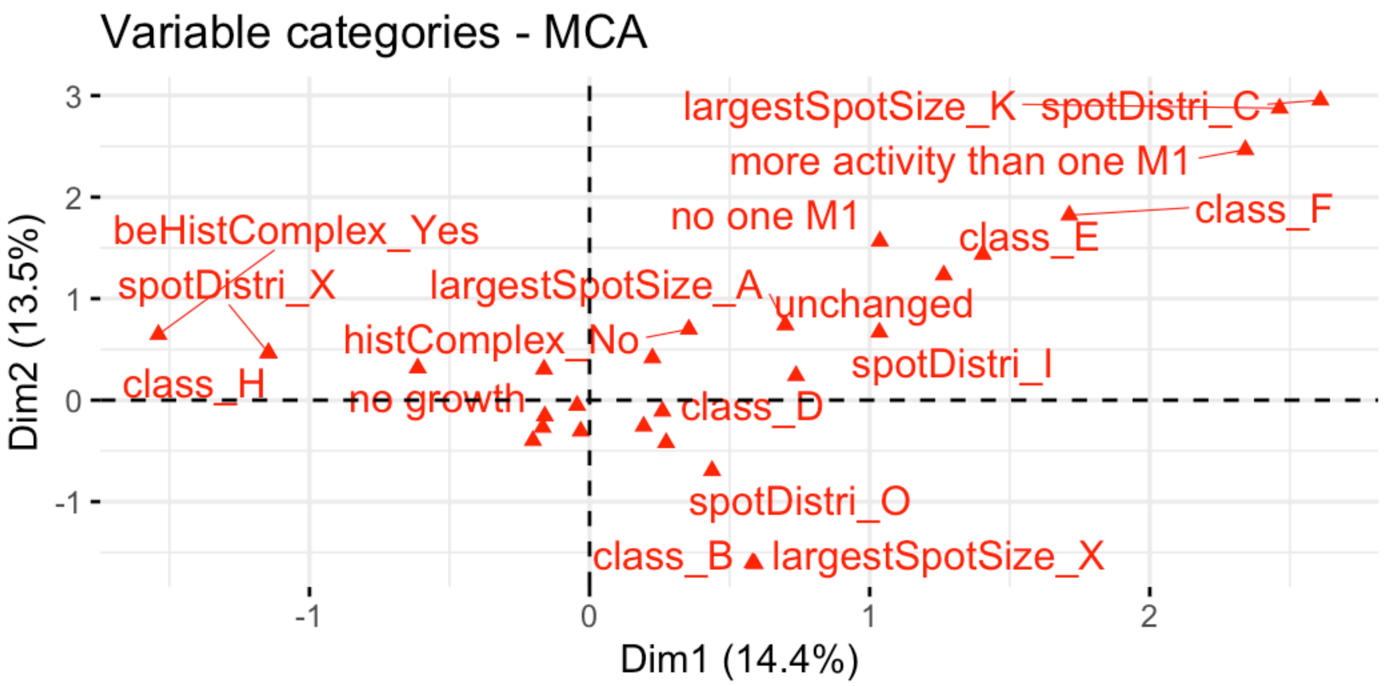
```
fviz_mca_biplot(res.mca, repel = TRUE)
```



Here the blue points indicate the solar regions, and the columns represent the variables. The biplot contains individuals and variables. Firstly, we can start with the analysis with the varibles.

Correlation between variables and principal dimensions

```
fviz_mca_var(res.mca, repel = TRUE,  ggtheme = theme_minimal())
```

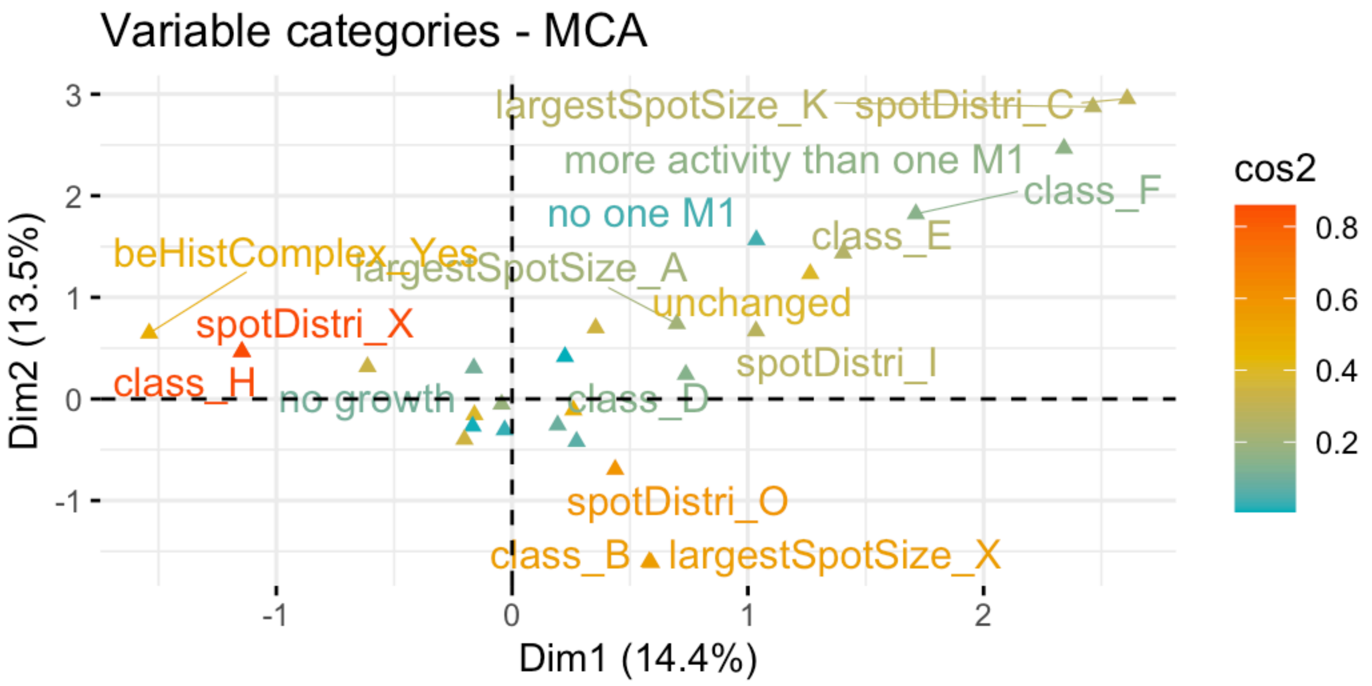


This plot visualizes the correlation between each variable and the MCA dimensions. Notice that the spot distribution C, largest spot size K, and more activity than one M1 are the most correlated with both the first two dimensions. Also, the varibles with similar profile are grouped together and with negative correlation are positioned in opposed quadrant. Quality of representation of variable categories

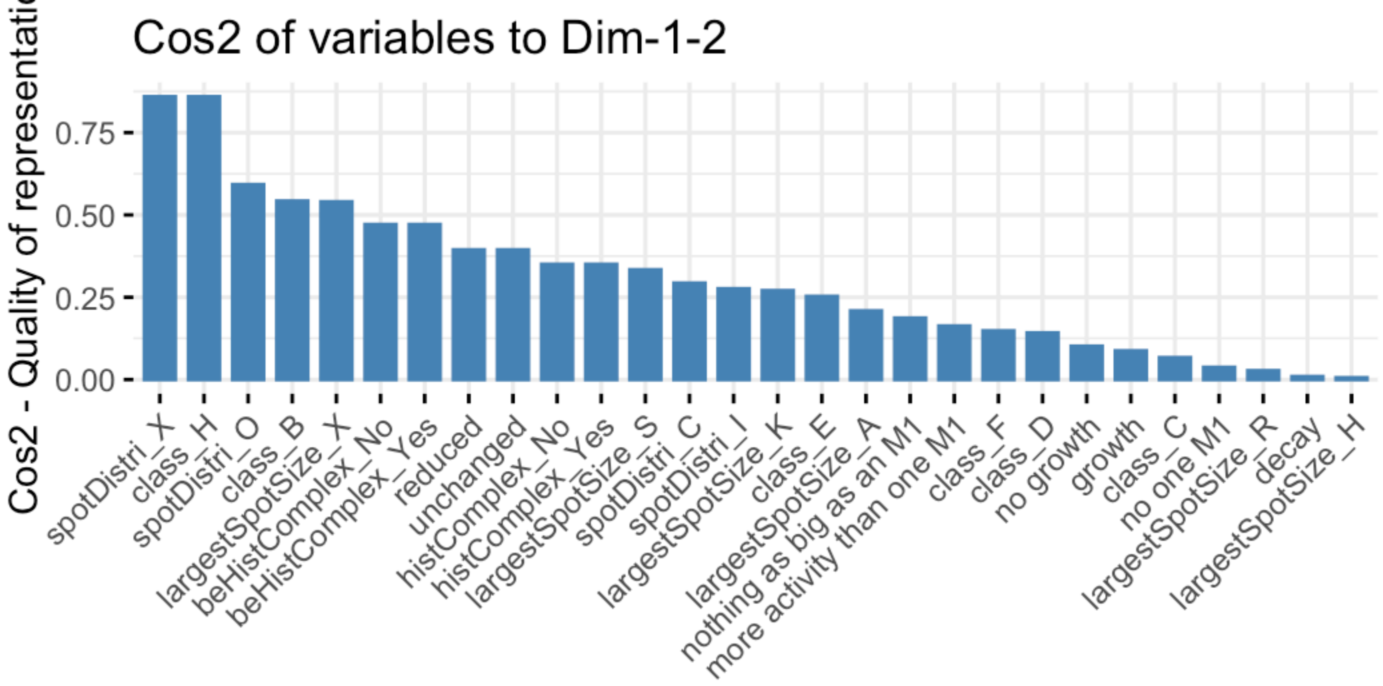
```
head(res.mca$var$cos2, 4)
```

```
##                Dim 1      Dim 2      Dim 3       Dim 4       Dim 5
## class_B 0.06326083 0.47905139 0.38923291 0.035918949 0.003551000
## class_C 0.01994251 0.04676014 0.27167033 0.314421252 0.037710739
## class_D 0.12807901 0.01370967 0.18399838 0.152252077 0.043332569
## class_E 0.12352366 0.12916619 0.02192278 0.002046771 0.006932972
```

```
fviz_mca_var(res.mca, col.var = "cos2",gradient.cols = c("#00AFBB", "#E7B800", "#FC4E
07"), repel = TRUE, ggtheme = theme_minimal())
```

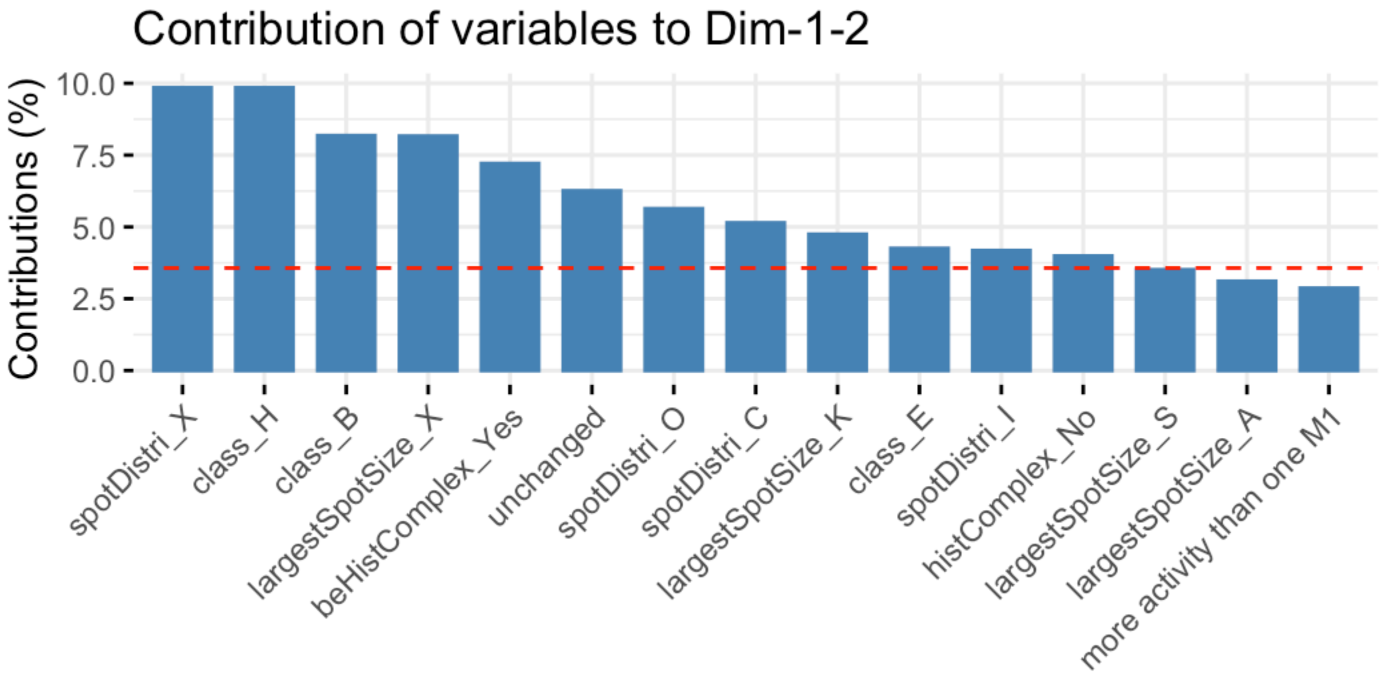


```
fviz_cos2(res.mca, choice = "var", axes = 1:2)
```

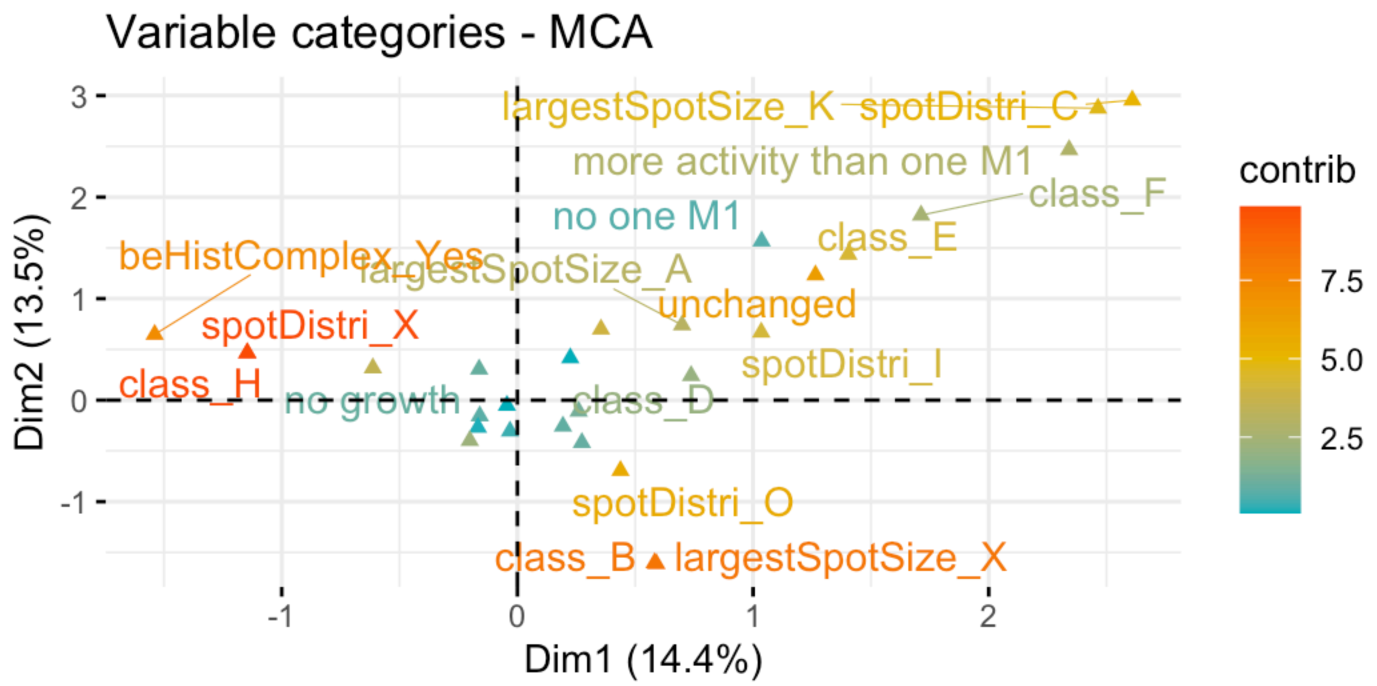

Cos2 of variables to Dim-1-2

Contribution of variable categories to the dimensions

```
# Dim1 and Dim2 are considered together
fviz_contrib(res.mca, choice = "var", axes = 1:2, top = 15)
```


Contribution of variables to Dim-1-2

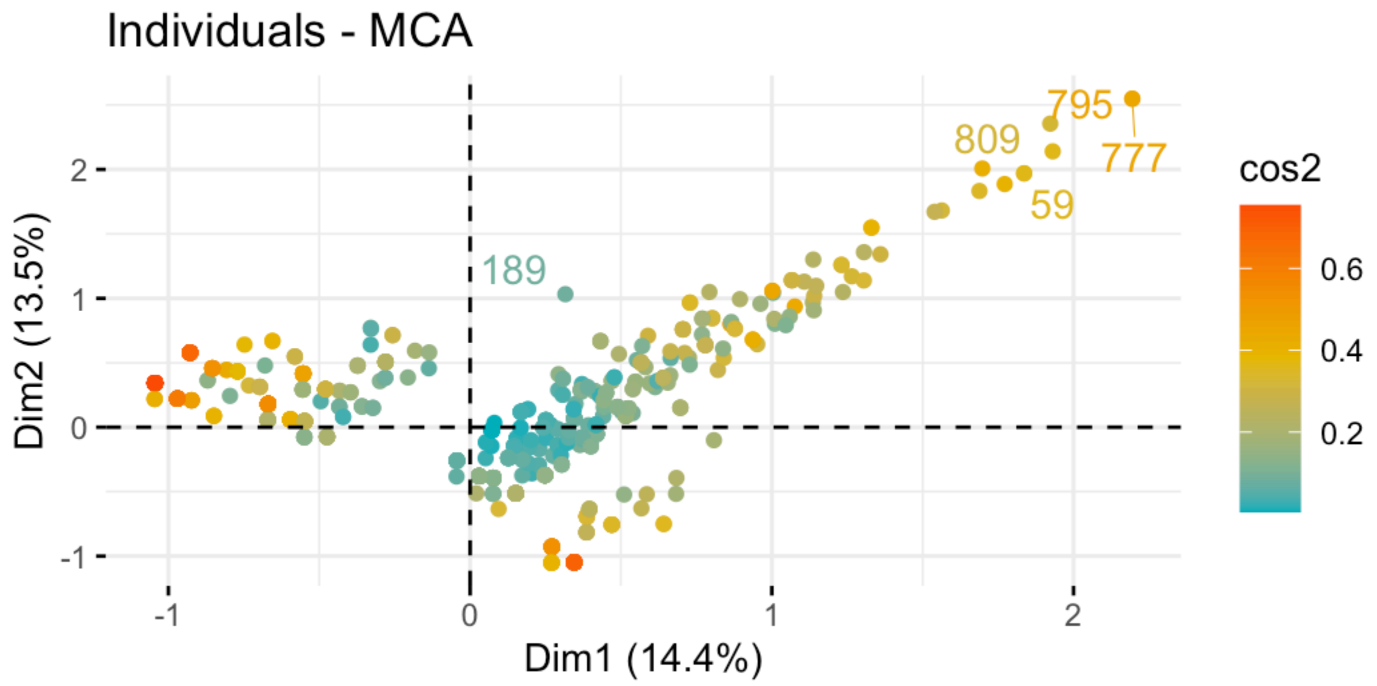
```
fviz_mca_var(res.mca, col.var = "contrib",gradient.cols = c("#00AFBB", "#E7B800", "#F
C4E07"), repel = TRUE, ggtheme = theme_minimal() )
```



In the second contribution plot, we only take consideration in the first 15 varibles for the given Dimension 1 and 2. It can be seen that the spotDistri_X and class_H occupy about 20% of contribution.The red line in the plot refer to the expected average value. Moreover, the contribution of the first 12 varibles are above the reference line, which we can consider them are important to the dimension1 and 2. For this Variable categories – MCA plot, it is easy to see that spotDistri_X and class_H have a important contribution to Dimension1 and class_F, largestSpotSize_A have contribution to the nagetive pole to Dimension1.

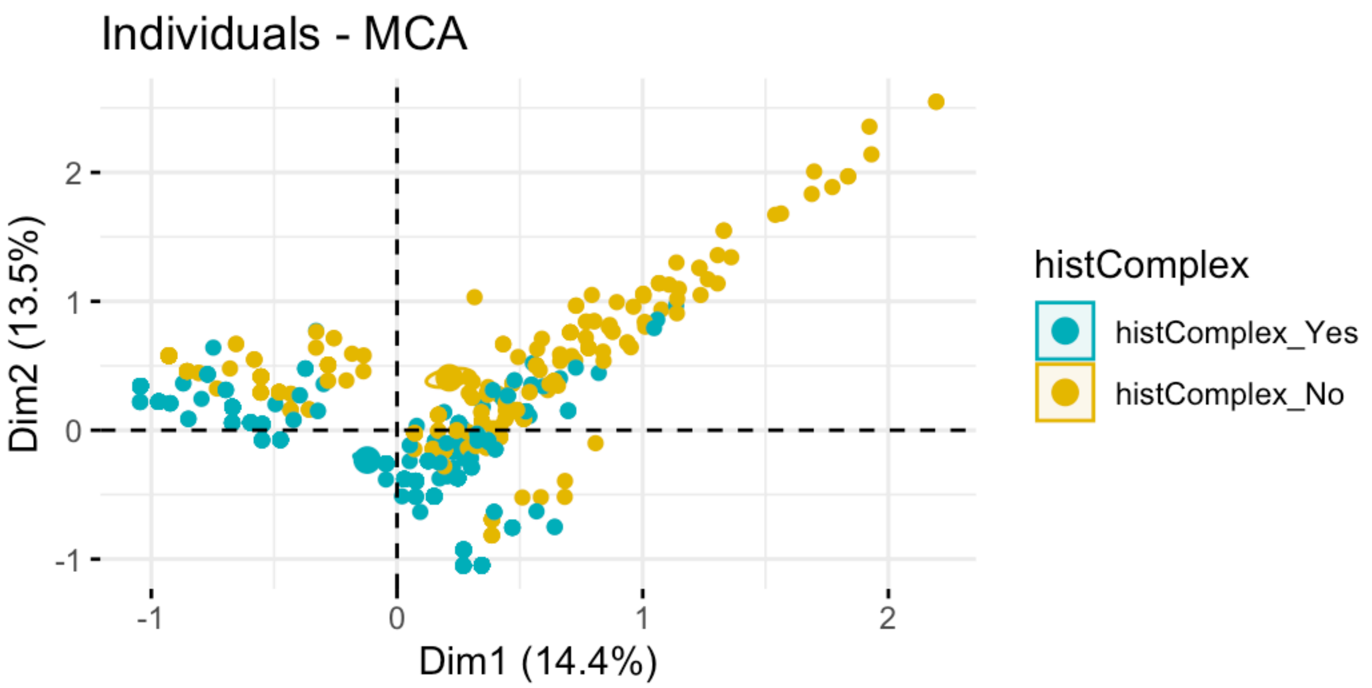
Secondly, this is our analysis for the Individuals. Individual’s Result

```
fviz_mca_ind(res.mca, col.ind = "cos2",gradient.cols = c("#00AFBB", "#E7B800", "#FC4E
07"), repel = TRUE,  ggtheme = theme_minimal())
```



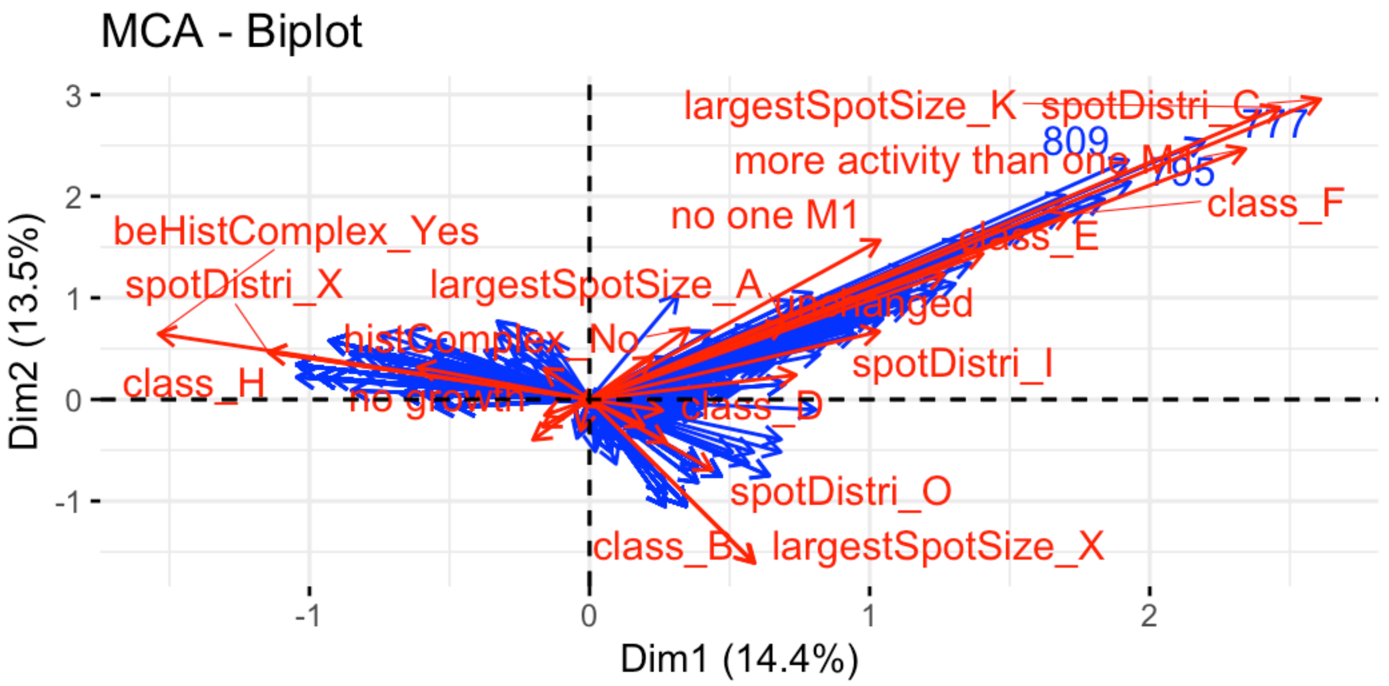
We can also color each region based on whether it is historically complex. We find that historically complex regions distribute uniformly on the first two dimensions, and the historically non-complex regions are more likely to find on the first quadrant, which implies regions positive in both the first two dimensions tend to be historically non-complex.

```
fviz_mca_ind(res.mca,label = "none",  habillage = "histComplex",  palette = c("#00AFB
B", "#E7B800"), addEllipses = TRUE, ellipse.type = "confidence", ggtheme = theme_mini
mal())
```



In this plot, we use levels of the variable ‘histComplex’ to color individuals by their groups. Asymmetric Biplot

```
fviz_mca_biplot(res.mca, map ="colgreen", arrow = c(T, T), repel = TRUE)
```



Using the Asymmetric Biplot can help us better interpret the distance between col points and row points, which means that the col information is presented in row space. In this plot, we scaled the cols and rows to have the variance equal to the square roots of eigenvalus.