Project in High Dimensional Data Analysis 2026

Part I

This part will introduce the fundamental properties, procedures, key concepts and basic rules of CA (Correspondence Analysis), MCA (Multiple Correspondence Analysis) and SVD (Singular Value Decomposition). The emphasis will be on how to use CA, MCA in practice, see the book "The Elements of Statistical Learning" by Trevor Hastie, Robert Tibshirani, Jerome Friedman.

Particular attention will therefore be paid to how CA, MCA and SVD can be useful in practice. Be pedagogical, act as if you were explaining to a non-initiated person.

Part II

In this part, you will apply linear regression, PCA (Principal Component Analysis) and SVD using a real dataset and R software (for PCA use the packages FactoMineR, factoextra) and R Markdown or Python (package Prince).

The file '**HDDAdataexam2026.csv**' on the drive contains data we collected together. The five columns correspond to the following variables.

- 'Phone' is the number of minutes an AIMS student spends watching the screen of his/her phone per day;
- 'SocialNetworks' is the number of minutes an AIMS student spends on social networks per day;
- 'Happiness' is the happiness score of an AIMS student on a scale of 10 (the most unhappy score is 1 and the happiest is 10);
- 'Walk' is the number of kilometers an AIMS student walks a day;
- 'InstagramRatio' is the Instagram follower/following ratio of an AIMS student, i.e. 'InstagramRatio' = 'Instagram followers' / ('People followed on Instagram' +1).

II-1

Export the dataset ‘HDDAdataexam2026.csv' to R and create a dataframe named ‘dataset’. Do a PCA of the above variables in the dataframe named ‘**dataset**’. Interpret the results.

II-2

Answer the questions below. Give details and interpretation of results for questions 2 to 10.

1) Construct the following variables and add them to the dataframe ‘**dataset**'

- '**LowInstagram**' is a dummy taking the value 1 if 'InstagramRatio' is less than 0.5, and zero otherwise;
- '**HighInstagram**' is a dummy taking the value 1 if 'InstagramRatio' is greater than 2, and zero otherwise;

2) We would like to investigate the smartphone addiction of AIMS students. Do a multiple regression with 'Phone' as dependent variable (Y) and the following independent (predictors) variables: 'SocialNetworks', 'Happiness', 'Walk', 'LowInstagram', and 'HighInstagram'.

3) Is the regression model in 2) useful in predicting 'Phone' time?.

Indication: if the P-value of the test of the anova table (of nullity of all coefficients except the constant) is greater than the critical level (usually 5%) then a linear model is not useful.

4) In the regression of 2), does 'SocialNetworks' significantly influence 'Phone'?

Indication: a predictor variable influences the response variable Y if the student test of nullity of its coefficient is not accepted.

5) According to the regression in 2), does 'Happiness' help reduce the time AIM students spend on their phone?

Indication: a predictor variable has a negative impact on the response variable Y if its coefficient is significantly negative.

6) Consider the regression output in 2) and two AIMS students A and B who provide exactly the same answers (120; 4; 15) for the variables 'SocialNetworks', 'Happiness' and 'Walk' respectively. Student A has an 'InstagramRatio' of 2.1 and Student B has a 'InstagramRatio' of 0.4. Give the predicted phone time for students A and B, respectively, denoted Phone(A) and Phone(B).

Indication: use ‘predit.lm’

7) Do a backward stepwise regression until you only have significant predictors left at the 5% significance level (it means delete one by one non-significant predictors beginning by the most non-significant).

8) Choose the model among those you obtained in 2) and 7) that is the most useful one according to the significance of the predictors variables and adjusted coefficient of determination to predict 'Phone' time.

9) Using the best model chosen in 8), give the predicted time an AIMS student spends watching his/her phone per day for a student who spends 30 minutes on social networks per day, has a happiness score of 8, walks 2 kilometers a day and has an Instagram ratio of 2.1.

10) Using the best model chosen in 8), give the predicted time an AIMS student spends watching his/her phone per day for a student who spends 30 minutes on social networks per day, has a happiness score of 8, walks 2 kilometers a day and has an Instagram ratio of 0.4.