ASSIGNMENT

Report in the table here below Name, Surname and ID Number for each member of your group.

Name	Surname	ID Number

Context and description of the dataset

Data regarding the use of Instagram between students at some universities have been collected in order to understand how to improve the performance, in term of views of Stories feature.

The dataset collects 243 observations regarding the following variables, updated on the 31st of October 2025.

- *sex*: sex of the owner of the profile, M for male, F for female
- *language*: language of the profile
- *siblings*: number of siblings of the owner of the profile, 0 for only child
- *account_num*: number of Instagram account the respondent can access to
- *story_views*: number of views of the last story published
- day_time_min: average number of minutes the profile owner spends daily in the Instagram app
- *num_follower*: number of followers of the profile
- *num_post*: number of posts published
- *effectiveness*: level of agreement (Strongly disagree / Disagree / Agree / Strongly agree) with the following statement "Instagram is an effective way to get in touch with other people"
- *attractiveness*: self-evaluation of the attractiveness of the profile (from 1 very bad to 10 very good)
- *private_d*: dummy variable with level equal to 1 for profile belonging to students attending private universities, 0 otherwise.

The aim of the assignment is to implement a statistical analysis and develop a report following the instructions reported here below.

Instructions

In the assignment folder you can find

- this file **30457_Assignment_Text_2025_2026.pdf**
- the dataset **30457_Assignment_Data_2025_2026.csv**

Include in the report also the scripts with the *Rstudio* commands used to implement the analysis and **send the final document in .doc or .docx format to** the following email address: **valerio.lange@unibocconi.it**

The assignment must be delivered **by 11:59pm on 21st of December 2025**. Assignments delivered later will not be considered. For any question related to the assignment please refer to the following email address: **valerio.lange@unibocconi.it**

A. Descriptive statistics of the sample

A.1. Report the main information for each variable (e.g.: appropriate graphs, frequency distribution table). Comment on results, taking into account central tendency measures but also non-central tendency measures and measures of dispersion.

A.2. If needed and based on the results obtained in the previous point, identify and exclude outliers from the next steps of the analysis.

B. Confidence intervals

B.1. Compute the 95% confidence interval for the number of views. Also comment on it, checking if there is any difference between only child and others.

B.2. Compute the 99% confidence interval for the number of followers. Also comment on it, checking if there is any difference between men and women.

B.3. Compute the 90% confidence interval for the proportion of accounts in Italian and comment on it.

C. Hypothesis testing

C.1. Is the number of followers significantly different between men and women? Comment on the results using different levels of the significance level *α* and compare your conclusion with point B.2.

C.2. Is the number of views significantly different between students in private and public universities? Comment on the results using the *p*-value.

C.3. Is the average time spent daily in the Instagram significantly different between people speaking English and people speaking Italian? Use α=0.01 and comment on the results.

D. Linear Regression

D.1. There is now interest to explore the relationship between the number of views and the number of followers. Run a linear regression model with the number of views as dependent variable and the number of followers as independent variable. Comment on the output results and evaluate the goodness of fit of the model. Finally check with appropriate techniques whether the simple linear regression model meets the linearity, normality and homoscedasticity assumptions and comment on the results.

D.2. Then, run a linear regression model including as independent variables the variables regarding sex, number of accounts, number of posts and the average time spent daily in the app to the model in the previous point. Comment on the output results and evaluate the goodness of fit of the model.

E. Prediction

E.1. By using the model in D.2, predict the expected number of views for the female median account.

E.2. Determine a 95% confidence interval for the prediction in the previous point.

F. Logistic Regression

F.1. Is it possible to predict if the account belongs to a student attending a private university looking at the number of followers, the number of posts and the number of views? Run a logistic regression model and comment on the output results.