Grup member's name:

1. (4 marks) Define the questions of interest (that is, statements of study objectives).   
These questions involve one or more variables. E.g. “Is there an association between gender and whether   
they smoke or not?" is a question involving two categorical variables. “What is the sample proportion of   
people who use Samsung phones?” is a question involving a single variable. There should be 5 questions of   
interest since you have to perform 3 univariate analyses and 2 bivariate analyses (regression, chi-square).   
Ideally, the questions are all related to a common topic (e.g., about smoking) so that the entire project   
appears connected and has a theme. The intention of these questions is so that you can find answers to   
these question through data analysis and address them in your powerpoint file (to be handed in). It takes   
time to find a suitable dataset so make sure you do not work on this at the last minute.   
2. (8 mark) Define the population of interest, sampling frame, sampling design, sample size   
You must be specific. E.g., “students” is not specific enough for a definition of a population. Use the format   
as taught in class: “The population is made up of all .....” is a good definition.   
3. (8 marks) Define the variables you use in your project and provide the source of your dataset.   
Identify 6 variables, 3 of which are quantitative and 3 of which are categorical. When defining a quantitative   
variable that has a unit, include it in your definition. E.g., Weekly wage (in $). Try to select variables that   
have a good chance of being strongly related to the question you are trying to answer. E.g., in regression,   
for every one-unit increase in the explanatory variable, the predicted value of the response variable   
changes by a fixed number of units. With that in mind, you should select two quantitative variables that   
have a good chance of being linearly related. Students should be aware that part of the project is to perform   
correlation and regression analysis, as well as a chi-square analysis. Much thought should be put into   
choosing suitable variables. Students have to identify which two variables are used for   
correlation/regression analysis, and which two variables are chosen for chi-square analysis. (Marks are   
given for clarity of description, identifying type (categorical/quantitative), and choice of variables   
suitable for regression analysis and chi-square test).

**Questions of interest**

* Which are the 5 most expensive apps in the play store dataset?
* How is the distribution of Rating? Can it be considered Normal distribution?
* How is the distribution of prices? Can it be considered Normal distribution?
* Can any of the numerical variables describe the behaviour of the price tendency?
* Can we consider Categories and Installs are strongly correlated?

**Objectives:**

**Population of interest:** All apps are available on Google Play Store.

**Sampling frame:**  All apps are available to download on Google Play Store in the zone where the data was collected (unknown)

**Sampling design:** Cluster sampling: Due to Play Store only show certain apps depends on the user´s location, all the data collection by web scrapping are based on only one region.

**Sample size:** 10841 records.

Define the variables you use in your project and provide the source of your dataset.

**Categorical:**

* **Genres:** An app can belong to multiple genres (apart from its main category). For e.g., a musical family game will belong to .
* **Category:** Category the app belongs to
* **Type:** Paid or Free
* **Content Rating:** Age group the app is targeted at - Children / Mature 21+ / Adult .
* **Installs:** Number of user downloads/installs for the app (as when scraped)

**Numerical:**

* **Price:** Price of the app (as when scraped)
* **Size:** Size of the app (as when scraped)
* **Rating:** Overall user rating of the app (as when scraped)

Source of the Data Set:

<https://www.kaggle.com/datasets/lava18/google-play-store-apps>