**Essential Information:**

* The project consists of applying econometric analyses based on a real-world dataset described below, using statistical software (Stata).
* Stata ver.15

# Dataset:

The panel dataset contains various metrics for apps in the top free and top paid charts of the Google Play app store for nine months (May 2016 to January 2017) at a monthly level.

Below is the description of the variables in the dataset.

* *log\_users*: the number of new users of the app per month.
* *updates*: the number of new updates for the app per month.
* *rating*: the star rating (between zero and five) an app received by users until the given month.
* *size*: the app file size, in megabytes, in the given month. It shows the complexity and

sophistication of the app.

* *log\_price*: natural logarithm of the app price in the given month. App price is the price that users must pay before downloading the *paid* apps (*price* is the non-logged version).
* *chart*: indicates if the app is free or paid.
* *rank*: the app rank in the top chart in the given month.
* *category*: app category.
* *life\_cysle*: an indicator of the app age; precisely, from 1st to 4th stage, indicating inception to maturity of an app life cycle.
* *app\_id* is the unique identifier of the app.
* *period*: the time-period identifier (at a monthly level) of the panel data.

Note: The log transformation applied for price variable (*log\_price*) is Ln(x+1), rather than Ln(x), to avoid losing observations with price=0; hence, if the price is zero, the log-transformed version will be zero as well— Ln(0+1)=0. For simplicity, you can interpret the effect size (if needed) as Ln(x).

Other variables (*calendar year, calendar month, and store)* are also included. The App name is dropped from the dataset, but the unique identifier (app\_id) is included, as explained above.

Overall, the objective is to estimate the effect of the number of new updates, app rating, app file size, and being free or paid on the app users (while controlling for some factors, as explained below).

# Content and Structure: Introduction

* Inspect your data and make sure there is no data entry mistake in the value of variables.

For example, the price of paid apps should not be zero, or the size of an app cannot be negative. If you detect some data entry mistake, drop those observations.

* Build the needed variables (as explained below); for example, natural logarithm

transformation of variables (if needed) or creating new categorical variables required for analysis.

* Provide a brief explanation for the methodology, such as data, the corrected and

cleansed sample, the definition of dependent, independent, and control variables, the objective of the analyses, and the baseline model.

# Descriptive Analysis

* Provide a two-way table for summary statistics of your model’s numeric variables for the whole sample, Game, and non-Game apps (altogether). Tip: you need to build a variable to distinguish between Game and non-Game apps.
* Provide the correlation matrix of the numeric variables of your model. Beefily discuss

the results.

* Apply a statistical test and evaluate if there is any significant difference (at 0.05 significance level) between the app categories regarding the number of users (logged). Can you qualitatively support the test results with a graph?

# Exploratory Analysis

* Inspect the data graphically, such as visual summary statistics across subsamples (such as categories, free vs paid, Game vs non-games, etc.), checking the distribution/skewness of *main* variables (i.e., dependent and independent variable), pre- checking the relationship between the dependent and independent variables, the longitudinal trend of the dependent variable (across subsamples), etc. The details and types of graphs are your decision—the objective is to provide a concise yet informative inspection of the data before running the regression. You may pick up a few of the above-mentioned list of potential graphs (or other graphs), which describe various aspects of the data efficiently.

# Main Regression Analysis:

* Conduct an OLS regression to estimate the effect of *the number of new updates*, *app rating, app file size (logged), and being free or paid on the number of new users (logged)* in the whole sample, while controlling for app price (logged), the, app category, and time-period. Carefully interpret and discuss the results (e.g., R-squared, the statistical significance of coefficients, the effect size). This will be the *baseline* model.
* Modify the baseline model to evaluate the differential effect of the number of new

updates for high-quality vs low-quality apps. For running your model, you need to build a variable to distinguish high vs low-quality apps. Define low-quality apps as those with a rating below four stars, and high-quality apps as those with four stars and above. Based on the results, discuss the statistical significance and effect size of the difference. You may use graphical illustration to enhance your discussion. Based on your experience or understanding of the mobile app context, can you provide some conceptual explanation for the results?

* Modify the baseline model to evaluate the differential effect of app rating across app

life cycles. You may use graphical illustration to enhance your discussion.

# Diagnostics and Robustness Analysis:

* Apply diagnostic analyses on the baseline model to check the potential heteroskedasticity and apply appropriate remedy if needed. Briefly compare the new results with the original results of the baseline.
* Applying statistical tests and graphical analyses, evaluate the possibility of having

outliers or influential observations in your baseline model.

* Run the baseline model with app fixed effects and robust standard errors. Briefly compare the new results with the original results of the baseline. Explain the effect of which variable seems to be biased in the baseline model. Explain how the fixed effect model can mitigate the endogeneity problem in your baseline model. Why are some variables dropped from the model with app fixed effects?
* If you want to improve your baseline model, what additional variables do you suggest

extracting from the Google Play app store (<https://play.google.com/store/apps>) to be incorporated in your model? Explain how the new variable(s) improve(s) the baseline model.

o Check out here for an example of publicly available app metrics: <https://play.google.com/store/apps/details?id=com.calm.android>(Calm app page) or any other app in Google Play app store; the available metrics are similar across apps.

# Appendix

* Copy the programming codes in the appendix in Word format. Do not copy the codes as a screenshot. Alternatively, you can upload the Stata do-file along with your report

# Format:

* The project file should be in Microsoft word format in Times New Roman 12-point font double spaced. The length of the project should be no more than 3500 words (excluding tables, graphs, and appendix). You should report the word count along with your name and student number at the beginning of your project. According to the university policy, exceeding the word count limit is subject to a 10-point penalty.

# Guideline and Tips:

* Apply the analyses required as explained *orderly*, section by section (from Introduction to Diagnostics and Robustness Analysis).
* The report —the writing, explanations, tables, and graphs— should be **clear and informative as a self-sufficient and stand-alone document** for readers who do not have access to this Final Project Description.
* In the **introduction**, concisely explain the aim of the empirical report, sample and data, and the definition of all final variables incorporated in your regression models. Some of this information (such as data and variables definition) has been already provided, but you need to give a concise summary of them in your report.
* All **tables and graphs** should be numbered and titled (and with captions if an additional explanation is required) and referred to in the report accordingly. The label of the variables in tables and graphs should be informative.
* Graphs should be visually clear (axis title, colour, legend, axis scale, etc.). You may use an image format for your graphs. Do not populate the report with lots of graphs; be selective and use the most informative ones for your purpose.
* Tables should be exported from the statistical software to a proper and readable Word format. You can report various models in one or two tables (each model in one column). Yet, you need to clearly number your models and refer to them in the discussions accordingly. Moreover, you don’t need to report the coefficients of the time-periods (and app categories) in your tables. Still, you should clearly indicate whether they are included in the models (see the Sample Report). All other coefficients should be reported in the tables.
* In **the regression tables**, standard errors should be reported below each coefficient (in the parenthesis), the significance level of the coefficient should be determined by asterisks. The R-squared and number of observations for each model should be reported (see the Sample Report).
* The **programming codes** used for preparing the tables, graphs, and regressions should be provided in a clear, easy to trace, and readable format in the appendix or as a separate do-file.
* You don’t need to cite any reference, but use a proper citation style and provide the reference list in the appendix if you intend to do so.
* Overall, the quality (i.e., clarity, rigour, precision, and depth) of the project is more important than the length.