General description of the Click-Balling app and tips

Click-Balling (CB) efficiently investigates time series and panels especially in their first stage of eyeballing. This no-code web app is actually an end-to-end research toolbox i.e., it automatically builds an academic skeleton document including abstracts, tables, figures, references list, JEL, and Keywords. It is intended for empirical researchers who explore their database and anyone who tracks data over time and writes a paper, accordingly. These capabilities enable researchers to focus on their contribution rather than technicalities. CB meets at least 90 percent of data exploration needs and it saves valuable time, as compared to popular statistical programs currently available in the market. Yet, CB is not a replacement of current statistical software tools but a complement, especially in giving quick insights regarding the data set at hand. The app automatically detects various file structures, merges data from both local files and the web, and handles missing values and outliers using sophisticated statistical procedures. CB includes side by side a Control panel and an Output panel. Every click in the Control panel is immediately shows up in the Output panel - no coding is needed. In the Control panel you can change or select parameters such as data type conversion, frequency conversion, or sub sample selection while in the Output panel the results of statistical procedures are presented. Another advantage of CB is the accessibility of users to various updated statistical packages in the very same Output panel so that they can implement alternative statistical procedures in one click. For instance, it goes over the above parameters and estimates linear relations (OLS and OLS Interactions) between the selected series so that all possible model results are presented (up to 240 models). As most researchers do not examine all possible models, this one-click principle saves a lot of time in all cases and sometimes turns research failure into success. In addition, CB presents statistical tips to the user in cases some tests are used inappropriately e.g., run OLS regression when data are non-stationary. Finally, saving outputs from the output panel locally, enables to build a skeleton of an article namely, tables and figures in Word/Latex/PowerPoint etc. Thus, what is left to the researcher to compose a complete article is writing down descriptions of these outputs.
For efficiency and clarity, the CB app is divided into Powerful Exploratory Data Analysis (hereafter, PEDA) and Advanced CB packages. The former is for those who are interested in data wrangling such as reading & merging data, imputing missing data, outliers detecting, and creating quick figures and summary tables. The latter Empirical Research Tool Kit (hereafter, ERTK), which includes all PEDA procedures, is for researchers who use analytic and thorough examinations of the data in order to publish in professional journals.
The rest of this document describes both the parameters and statistical procedures of the two CB apps. 'ERTK only' in square brackets refers to procedures that are not included in PEDA but appear in the ERTK app, only.

(A) Control Panel (left side). You can select parameters as follows:

• **Series Picking: Select one or more series**. Categorical variables are colored in blue and symbolized by trending bars while binary series are green colored with a filled arrow icons. All others considered numeric variables. You can organize the series by their first appearance (Original), by increasing order (A:Z), by decreasing order (Z:A), or by their

suffix (suffix). The latter is useful for panel data. These 4 sorting options are titled 'Sort Type'.
Once a categorical or binary variable is selected a box above the current one is opened so that it is possible to filter the sample data by categorical/binary variables.

• **Conversion Type: Converts the data type of all selected numerical series**. Examines the original series (Original), natural without trend (deTrnd), cubic spline smoothed series (smooth), indexation of the first observation to 100 (index), de-seasoned series (seas), BoxCox transformation including negative numbers (BoxCox+), and rolling window sample (Roll). Once you choose the Roll option you can select the rolling conversion type (mean, median, max, min, sum, standard deviation (sd), CoV (Coefficient of Variation) = mean/sd), and the rolling window size (using a slider). See the nearby question mark for exact **conversion** definitions.

• **Transformation Type: Transforms the data of all selected numerical series**. Examines the original series (Original), change rates (ratio), Log Differences (diffLog), simple differences (diff), natural logarithm (log), normalization (scale), absolute values (abs), squared values (sqr), root squared values (sqrt), and 1 over the series values (1/x). See the attached question mark for exact **transformation** definitions.

• **Frequency of data: Converts data frequency**. Changes the frequency of data from higher frequency to a lower one (daily (day)/ Half weekly - every 3 days (Hweek)/ Half month - every 15 days (Hmonth)/ monthly / quarterly / Half year (Hyear)/ yearly/ 2 years (2y)/ 3 years (3y)/ 5 years (5y)/ 10 years (10y)). The app automatically detects the basic frequency of the data and allows to convert the frequency according to the mean, median, sum, first, or last observation of the higher frequency series (using the 'Frequency Method' bar below the 'Frequency of data' one).

• **Sample Period: Changes the sample period using the slider**. Use the slider to change either the sample start, the sample end, or both (left click the mouse on the slider and once you see a two sides arrow, move it). The latter is useful for comparing two sub periods of the same size.

• **Save Output.. (green button)**. Once you see results in the Output panel (figures/tables) you can save them locally for further usage as PNG files. Thus, these Output panel components can be incorporated in a document for publication. In most cases there are three output parts [ERTK only] e.g., regression results in a table, regression residuals in a graph, and residuals tests in a table. Regression residuals are also incorporated in the current data set as x_fitted and x_resid where, x is a regression type e.g., ols. You can download each one of them, separately by clicking the Ok button (let you browser accepting multiple downloads). For a full screenshot press the <PrtScn> button and paste the result into any painter for further usages.

• **Statistical Tips [ERTK only]**. The app detects misuse of statistical procedures, such as the F-test when the data are not normally distributed. Four data issues are exam-

ined: (1) 'Independence' of observations (serial correlations) by Ljung-Box statistics. (2) 'Normality' by Robust Jarque-Bera test. (3) 'Stationarity' by Augmented Dickey Fuller (ADF) and Phillips Perron (PP) unit root tests. (4) 'Seasonality' by 'QS', 'Kruskall Wallis', and 'Friedman' tests (at least two years of data are required and sometimes the above procedures cannot detect seasonality. In such cases, NA is shown). The background colors (4 different colors for each one of the above misuses) are on whenever the null (H0) is rejected by a P-value of less than 0.05, and, if that rejection is contrary to the assumptions of the statistical procedure, for example, non-stationarity in OLS regressions. For further details and comprehensive univariate tests see the menu: Exploration/Series Characteristics.

(B) Output panel (right side). Statistic results are organized, as follows:

(1) **Utility**

• **Import File (New)**. Reads user's data from Excel, CSV, Text, Stata, SAS, and SPSS files, while overwriting the current data (if exists). If data are in Excel or CSV, the app automatically detects where the dates column is located, what is the basic frequency e.g., daily, what are the variable types e.g., categorical, and whether the uploaded data can be further processed or not. For example, whether dates are missing, invalid or the variables are non-numeric. If the data are structured as a panel, the app converts the data into flat file format i.e., original series common prefix for each variable and specific suffix for each identity. For instance, for two variables: population (pop) and gdp of three IDs: UK, Japan, and USA the app will create 6 series: pop.UK, pop.Japan, pop.USA, gdp.UK, gdp.JAPAN, and gdp.USA. Then you can easily re-compose a panel of your own.

• **Merge File**. Reads a new data set either flat file or a panel and merge both files by the dates column according the following options (A is the current data set and B is the new (merged) one):
A - Merges the new data by the current dates, only.
B - Merges the new data by its own dates, only.
A and B - Merges the new data with the existing one only if dates of both A and B are the same.
A or B - Merges the new data with the existing one by either A dates or B dates.
Suffix - Merges the new data set with the existing one only if both series names share the same suffix. This is useful for panels in which you want to merge part of a large data set according the names of the existing data set (A). See an example in the use case video at `https://www.clickballing.com/`. It is highly recommended that high frequency data e.g., 'daily' will uploaded first (A) and lower frequency data will be merged (B). Finally, if you mix panel and flat files, upload the panel first (A) and merge the flat file next (B). Notice that you can download Web data directly to the current data set in the 'Utility/Download Web Data' menu [RTK only].

• **Save Current State**. Brings you back to the current state including parameters of the Control Panel. If your data consists of several (merged) files, using this option will save you a lot of time. The next time of using CB, you can use the "Load First/Demo Data"

button on the landing page to get started from the last state without uploading and merging the files again.

• **Show Demo**. Presents visual/verbal explanations on the app using the Home Page of Click-Balling app: `https://www.clickballing.com/`. In particular, it demonstrates a use case of research - The relationships between Credit Default Swaps (CDS) and the COVID-19 data. Note that some of the browsers cannot display that clip. In such a case you can download the clip to your local computer and display it from there.

• **Create Seasonal Variable**. To the selected variables, you can also add a new binary or categorical variable representing seasonality. Binary variables consist of two unique values such as 0/1 while categorical variables have limited number of unique values but more than two. The available seasonal variables are: day of the month (Day - 1:31), day of the week (Wday - 1:7), week of the year (Week - 1:53), month of the year (Month - 1:12), quarter (Qtr - 1:4), and year (Year e.g., 2017:2020). Each seasonal variable can be categorical i.e., all categorical values in a single series (select 'Categorical') or each value is saved in a separate series such that the number of new series will be the number of seasonal values (select 'Binary'). In such a case the new series is binary that gets the value 1 for the season and 0 otherwise. For instance, selecting 'Qtr' and 'Categorical' will create one variable named Qtr with categorical values: 1, 2, 3, 4 while selecting 'Qtr' and 'Binary' will create 4 series named 'Qtr1', 'Qtr2', 'Qtr3', 'Qtr4' with values of 0 and 1, accordingly. This option works on the original variables only. For more information see the attached question mark.

• **Create Binary/Categorical Variable**. Adds a Binary or categorical variable by double clicking on the most right hand side column. Then give the new series a name and save it (the default is 'Dum1'). For your convenience the day of the week, the day of the month, week number, month number and the year are presented, too. Additionally, you can change the selected data frequency (in the Control Panel) to a lower frequency e.g., from daily to monthly, and set your Binary/Categorical variable as before to get respective Binary/Categorical at the original higher frequency. For example, if your basic frequency is daily and you select a quarterly frequency (Qtr) then, by replacing the zero of 2021Q3 to 1 or any other number all days within that quarter turns to 1 once you keep changes <Ctrl>+<Entr> and save. Then you can see the new Binary/Categorical Variable by going to other menus such as, 'Single graph' or 'Data Table'.

• **Create Lead/Lag Variable**. Here you can add up to 25 periods of leads or lags to the selected variables. This option works on the original variables only, and not on the converted or transformed data types or frequencies. The new lag/lead series are symbolized by the prefix 'Lag'? / 'Lead'?, respectively where ? can be: 1:6, 10, 12, 15, 20, and 25. For example, if the selected variables are "oil" and "gold" and you want to add lag of ten periods to these two variables, the new series names will be Lag10oil and Lag10gold.

• **Create Numeric Variable**. Adds a numerical variable based on other selected variables. Here you can enter an expression in which a:z or A:Z represent the variable name that is shown above the input box. For example, (a + b∧2)/c [which is equal to $\frac{a+b^2}{c}$] is an appro-

priate expression. The series can be named by entering a new valid name in the upper input box. If that box remains empty the app will give the new variable the name 'X1'. As long as the expression appears OK, the results will appear on the right side of the Output Panel. By using this option, you can efficiently and quickly create your own variables by trial and error.

• **Create Panel**. Once your uploaded file includes panel data you can re-compose any panel you want as the series in CB app are kept as in flat file format i.e. with a common prefix and different suffix (see an example in the 'Import File (New)' menu above). It needs at least 2 variables with common Prefix and 3 IDs with different Suffix. This option can save you a lot of time if your panel consists of missing values, outliers, and is unbalanced. By carefully examining each series you can decide whether to exclude it, complete missing values, correct outliers, or even convert or transform it (using various parameters in the Control Panel). Then you can easily build your own panel and save it for further usage.

• **Download Market Data [ERTK only]**. There are 7 data types to choose from (Stock, Stock Index, ETF, Currency, Commodity, Cripto, Macro). After selecting one data type e.g., Stock, a menu of 7 choices is presented below the data types menu (Adjusted Close, Close, High, Low, Open, Volume, All). Choose one or more variables ('All' means all other variables). Then, choose one or more series of the specific data type e.g., Currencies, and First Date and Last Date inputs. Each time you make a selection the selected variables are shown in the upper line under the data types name e.g., 'Close High' under 'StockIndex'. 'Macro' is an exception in which the menu to choose from contains variable number of items depending on the keywords you type in the Free Search input box. Also, selected variables are numbered as Macro1, Macro2 etc. by their symbols. You can shrink the 'Macro' items list by selecting from the 'Country' menu. To reset a selection go back to any data types menu, select a variable e.g., 'High' and deselect it. You will see that your selection is disappeared from the upper line under the data types name. Once you select all required variables and series (selecting companies/countries is mandatory requirement) press the green 'Download' button. After a few seconds a basic stats table will show up. Then you can either download the data as an CSV/Excel/PDF/PRINT, merge it with the current data set, or replace the current data set with the new data using the upper green button. You can change the names of the selected series/variables or any cell in the basic stats table so that your new variable names will be understandable. Once you see a success message ('Done'), you will see the new variables at the series picker (in Control Panel).

• **Download Data by API URL [ERTK only]**. Here you can download time series or panel data from the internet using API. The URL address should be either XML, or HTML, or JSON, or CSV type and must contain one column of date. First, copy & paste a URL string from the relevant web site, that refers to data world contents e.g., exchange rates (usually more than one series) and select date ranges. Then, select series by their codes from the series picker e.g., EUR/USD exchange rate and a merge type (AND, OR, or Replace). Once you see a table with basic statistics of the downloaded data and a description of the codes, select the required series from the series picker. You can edit the series codes (most LHS of the table) by double clicking on any cell. Finally, by pressing the green button 'Go' you will get a 'Done' message that the merged/replaced data is ready.

To start over, clear the URL string, deselect all series, and wait a second. You can also download the table as CSV/EXCEL/PDF for further usage or print them. Notice, however, that sometimes the Web data tables use Java Script codes so, they cannot be downloaded using this menu.

• **Edit All Original Variables**. Displays in a table with sophisticated search engine all the original variables. By double clicking on a cell, you can edit a particular cell within the table. You can also change variable names by clicking on the name cells (to the right of 'Edit»' cell - the 3rd line from the table top). Deleting entire row or column is done by clearing the row or column numbers (most left column and top row, respectively). finally, by clicking on the right or left wide buttons on the table's top, you may save your changes, or get back to the original data set.

• **Build Literature Survey [ERTK only]**. Builds for you an abstracts list based on your keywords. Here you should select some keywords (spaces between them) to start a search (Scholar Search). Then remove non-required rows (Delete Rows) and finally save the abstracts (Save Abstracts) into your device. The saved file name should include the characters: 'Abstract' so that you will be able to run a report (in the below menu 'Utility/Build Article/Presentation'). You can also save or print the abstracts as-is using the buttons (up left): CSV, Excel, PDF, Print. The abstracts are based on Google Search (GS) therefore, try to use many keywords in order to make it accurate. Limit the searching period (mm-yyyy) to the latest years and choose searching parameters as follows:
- No. Pages: One/Two/Three. Each GS's page contains 18 abstracts.
- Publication Type: All Types/Journals/Journals & Books. The former option contains also HTML documents, works in progress etc. Use the latter options if you want to narrow down the search.
- References Style: MLA/APA/Chicago/Harvard/Vancouver. These options are the same as in GS.
The output includes: (1) Article name, (2) Author names and publication year (in parenthesis), (3) Journal name and its ranking (based on SCImago Journal rank - SJR), (4) Number of citations, (5) Full abstract. In some cases the app cannot recognize the journal so the GS's (short and incomplete) abstract is presented. Finally, the output file (AbsRef_[SearchKey].zip where, SearchKey is your keywords) contains two zipped files: The above abstracts list (Abstracts_[SearchKey].txt) and the references list (references.bib). Unzip the two files into the folder in which your other tables and figures are. Of course, you can save the Abstract list as csv/xlsx/pdf, or print is.

• **Build Article/Presentation [ERTK only]**. Builds an academic skeleton document including: (1) 'Title' page (Title, Author names, date). (2) 'Abstract' page with keywords and Journal of Economic Literature (JEL) keywords. These two items are automatically derived from the above menu (Build Literature Survey) so an abstract output file together with a bibTex file (References.bib) must be included in the 'Select Input Files' menu once you want to include literature survey and references in the skeleton file. (3) 'Introduction' page. (4) 'Literature Survey' page - Once an abstract output and bibTex files are be included (see the above menu), all abstracts are presented with the title name and their hyperlink to the

references list, which is located at the end of the skeleton document. (5) 'Model/Hypotheses' page. (6) 'Data description and results' page. This page includes two sub sections: Tables and figures. Once you include tables and/or figures they will appear within these two sub sections along with their automatic numbering. (7) 'Discussion of the results' page. (8) 'Robustness checks' page. (9) 'Summary' page. (10) 'References' page. Here all references from the included file 'References.bib' that appear in the 'abstract output' file will be depicted one after another. The references contain hyperlinks to the original studies so that you can read a study easily by clicking on its hyperlink (and read it all once you have an access to the respective journal).
The output file, which includes all of selected items, can be: word document (.docx), Tex-PDF - LaTex file for PDF (.tex), html file (.html), Microsoft PowerPoint (.pptx), TexBeamer - Latex file for Beamer presentation in PDF (.tex), pdf - Truly PDF file, ioslides - html file for ioslides presentation (.html), and slidy - html file for slidy presentation (.html). Once you see the skeleton document details (CB_yyyy-mm-dd.*) you can download the file and add your contribution using that skeleton file as a starting point. Except for the Word (.docx), Powerpoint (.pptx), and PDF (.pdf) files all other files are zipped since they include two files: (1) BuildFile.tex for Latex based files or BuildFile.md for HTML based file, (2) References.bib. Unzip the two files into the folder in which your other tables and figures are.

• **Preferences**. This option lets you choose the default of:
(1) Number of decimal points: 0..10
(2) Date format: dd/mm/YYYY, dd/mm/YY, YYYY-mm-dd, YY-mm-dd, dd.mm.YYYY, dd.mm.YY

(2) Visualization

• **Single Graph**. Shows an interactive graph with Zoom-In capabilities (hovering and brushing). You can select from various graph types (line, point, bar, stacked bar, area), point patterns (dot, triangle, square...), line patterns (solid, dashed,..., mix - gives each series different pattern), and line thickness (1..4, mix).

• **Data Table**. presents the selected variables in an Excel-like table. Here you can see the data as is and save (CSV, EXCEL, PDF) or print it.

• **Multiple Graphs**. Creates an interactive graph with Zoom-In capabilities (hovering and brushing) like the Single Graph above, but each series is presented separately (up to 18 series in three vertical sub panels). Furthermore, Zooming in by brushing affects all selected series, simultaneously.

• **Graph Gallery**. Creates various graph types including: Pie, Donut, Line (in one or or two Y axes), Dot, Combined bars/lines/dots, Box plot (in one or or two Y axes), Violin, Area, Contour, Bubble, and HeatMap. Some graphs can be drawn vertically or horizontally and by variable or by date. Additionally, customized graph titles can be added (Add Titles) such as: Caption, X axis, Y axis, and Secondary Y axis with various font size and decorations. The latter HTML tags are added in the inbox texts as follows: Bold - add

before the bold text and after, Italic - add before the italic text and after, Superscript - add ^{before the super script text and} after, and Subscript - add _{before the sub script text and} after. Finally, break a line with
.

• **Graph-Table**. Combines an interactive bar graph (upper panel) with a linked table (lower panel) of the selected variables. By clicking anywhere in the interactive bar graph, you can see the selected variables on the clicked date along with its surrounding data, i.e. 1 to 5 dates before and after the clicked date.

• **Histogram**. Shows histograms of each one of the selected series. You can choose the histogram type as follows: (1) 'count' - the values of a series are shown on the X-axis while number of occurrences per each bin appear on the Y-axis, (2) 'probability' - the same as (1) except that the probability (0..1) is shown per each bin. (3) 'density' - a line graph of the density function of the series. (4) 'overlay' - all series figures are overlaid (partially overlapped). 'stack' - all series figures are stacked one on each other. Histogram types (1)-(3) are sub-plotted while (4)-(5) are presented in one plot. You can also control the maximum number of bins (auto, 5, 10, 15, 20, 25) and per each series one or more basic statistical parameters (mean, standard deviation, skewness, and kurtosis) can be presented. For every sub-plot, the series name and the number of observations are depicted.

• **Q-Q Plot**. Plots a scatter diagram of the selected variables against the theoretical probability distributions of Normal, ChiSquared, Uniform, Studen-t, Exponential, Log-Normal, and Caushy. The diagrams can be plotted with an envelope (confidence levels of 0.9-0.995) and up to 4 extreme observations that are emphasized by the red color. For a description of the above probability distributions see `https://en.wikipedia.org/wiki/List_of_probability_distributions`

• **Pairwise Correlation (N>1)**. Presents pairwise correlation coefficients by Pearson, Kendall, or Spearman of at least 2 variables, as follows:
All - Correlation coefficients are presented with their significance levels (***, **, *, and ■ denote 0.001, 0.01, 0.05, and 0.1 significance level, respectively), self distribution, and scatter diagram.
One - When you select one variable from the selected variables, you will be able to see its correlation with all other variables.
Extreme - Shows only the most (10..30) extreme coefficient values either positive or negative, extreme coefficient values that are greater than 0.25..0.9, smaller than -0.25..-0.9, or greater, in absolute values, than 0.25..0.9.

• **Cross & Auto Correlation (N=2)**. Depicts the Cross Correlation, Auto Correlation and partial Correlation functions of two variables (N=2). You can control the number of lags/leads by using the slider (up to 40 periods lags/leads). In addition, it is possible to draw a scatter plot of the 2 selected variables with points marking, LOESS smoothing of the bi-plot center, and extreme observations.

• **3D Plot (N=3)**. Presents an interactive 3D graph with 3 series only (N=3). These 3

series may include Binary/Categorical variables and you can move, enlarge, and rotate the 3D graph by Right mouse button, Left mouse button, and Scroll mouse wheel, respectively.

• **Principal Component**. shows a bi-plot of the two largest principal components with the relative contribution of the selected series to the total variability (N>1). For explanation on the method see `https://www.statisticshowto.com/principal-component-analysis-2/#PCA`.

• **Clusters (N=1)**. Depicts cluster map using several distance methods (Euclidean, Manhattan, Minkowski, maximum, Sokal-Michener, and Gower) and various clustering methods (K-means, PAM, H-cluster, and HK-means). The cluster can be determined manually (1..15) or automatically (optimally) by the app. Each number in the map denotes a date. For explanation on the methods see `https://www.statisticshowto.com/?s=cluster`.

(3) Exploration

• **Series Characteristics**. Shows important characteristics of the selected variables with their significance levels (in test results):
Serial correlation - Number of observations, Lag coefficient of one period of the series, and the Ljung-Box Statistics for serial correlation.
Optimal ARIMA parameters - Based on Akaike Information Criterion (AIC) regarding the optimal lags number, presents the Auto-Regressive (AR), the Integration level (I), and Moving Average (MA) of each series.
Normality tests - The tests include Jarque-Bera, robust Jarque-Bera dealing with outliers, Symmetry for the existence of skewness, and Fat tails for the existence of kurtosis.
Stationarity (unit root) tests - Including Augmented Dickey Fuller (ADF), Phillips Perron (PP), Kwiatkowski–Phillips–Schmidt–Shin (KPSS) assuming Level Stationary, and KPSS assuming Trend Stationary.
Seasonality - Detects seasonality using the P-Values of QS, Kruskall Wallis, and Friedman tests. Notice that at least two years of data are required and any data frequency is accepted (including daily. See the R package 'seastests' for details).

• **Statistics by Categorical/Binary**. Presents basic statistics (mean, median, max, min, standard deviation (sd), sum, number of observation (Obs.)) broken down by **one** Binary or categorical variable. This can be useful in situations where the selected (numerical) variables behave differently over Binary or categorical values.

• **Distribution Classification**. presents the probability (in percents) of the selected series to be derived from a known distribution. In particular, it classifies the distribution of a model-family using machine learning. Since it is difficult to exactly predict the correct model family, consider this function as somewhat experimental (See the R function check_distribution of the package 'Performance'). The distributions list contains the following known distributions: 'bernoulli', 'beta', 'beta-binomial', 'binomial', 'chi', 'exponential', 'F', 'gamma', 'lognormal', 'normal', 'negative binomial', 'negative binomial (zero-inflated)', 'pareto', 'poisson', 'poisson (zero-inflated)', 'uniform', 'tweedie, and 'weibull'. For a descrip-

tion of the above probability distributions see `https://en.wikipedia.org/wiki/List_of_probability_distributions`

• **Build My Stats Table**. Builds a summary table of statistics that can be downloaded and included in any research document, later. The list of possible statistics includes: mean, median, max, min, sd, Obs. (number of observations) ,NAs (number of NAs), pos (number of positive numbers), sum, skew (skewness),kurt (kurtosis), Q10 (the 10^{th} percentile), Q25 (the 25^{th} percentile), Q75 (the 75^{th} percentile), Q90 (the 90^{th} percentile), AC1, AC2, AC3, AC4, AC5 (Auto Correlation Coefficient of one to five periods), JB (P-Value of Jarque-Bera test for Normality. H0: the series is normal distributed), Zeros (number of Zeros), Dup. (number of duplicated numbers), CoV (Coefficient of Variation: mean/sd), mode (most common value), NonLinear (Nonlinearity statistic based on Terasvirta's test of a time series. See help in '?'), entropy (a measures of forecastability of a time series based on Shannon spectral entropy. See help in '?'), stability (variance of non-overlapping windows' means), crossing (number of times a time series crosses the median), hurst (Hurst's coefficient that measures the long-term memory of a time series. See help in '?'), ADF (P-Value of the Augmented Dickey-Fuller unit root test with linear trend and automatic lags order. H0: the series is non-stationary), and PP (P-Value of the Phillips-Perron unit root test with linear trend and automatic lags order H0: the series is non-stationary). The summary table layout can be either vertical (default) i.e., statistics above the table and variables on the left side, or horizontal.

• **Long-Short Smoother (N=1)**. Presents the selected (N=1) series and its Long and Short term smoothing period counterparts in a Zoom-In graph. The smoothed series can be calculated by simple moving average (MA), Exponential moving average (EMA), Hull's moving average method (HMA), LOESS method, Daniell's Kernel, Dirichlet's Kernel, and HP filter. By moving the sliders of 'Smoothing Period 1' and 'Smoothing Period 2' one can find out when a short-term smoothing period crosses the long-term smoothing period from below/above. The crossing points are represented by red (down crossing) and blue (up crossing) dashed horizontal lines.
The button **Best ROR** goes over all slider options of Long and Short Term smoothers with 2-3 gaps between consecutive numbers depending on the sample size (No gap - up to 50 rows, 2 gaps - 50 to 100 rows, and 3 gaps - over 100 rows). Then, it finds the best accumulated ROR (%, Total Rate of Return) and changes the two sliders values, accordingly. If Fees (%) is not zero the app takes it into consideration. For example, if there are 5 crosses (open and close positions) and Fees = 0.5% then: Net ROR = Gross ROR - 2*5*0.005. Also, if for instance Stop Loss = 20% then the long (short) position, when the short smoother - the red line crosses the long smoother - the blue line from below (above), is terminated premature. Such a case occurs when $\frac{price_t}{1.2} < price_0$ for the long position and when $price_t * 1.2 > price_0$ for the short position. The Net ROR is calculated accordingly i.e., $ROR = 100 * (\frac{price_t}{price_0} - 1)$ for the long position and $ROR = 100 * (\frac{price_0}{price_t} - 1)$ for the short position (assuming no margins) In addition, you can allow or disallow short selling.
At the bottom of the lower panel a footnote depicts the Buy & Hold benchmark that is buying at the sample period start and selling at the period end. The figure and the lower

panel are useful in technical analysis.

• **Simple Seasonality (N=1)**. shows the seasonality impact on a selected (one) series in two tables: the upper table shows basic statistics of the series broken down by seasonal values e.g., days of the week (1..7) while the lower table presents P-Values of pairwise equality tests by Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Misses, and Kuiper. Note that this option is relevant for days of the week, weeks, months, and quarters, only.

• **Pairwise partial Correlation (N>1)**. Shows pairwise partial correlation coefficients by Pearson, Kendall, or Spearman. The upper matrix shows partial correlation coefficients while the lower matrix depicts Semi partial correlations. ***, **, and * denote 0.01, 0.05, and 0.1 significance level, respectively. For explanation on the methods see `https://www.statisticshowto.com/partial-correlation/`.

• **Predictive Power Score - PPS (N>1)**. The PPS is an asymmetric, data-type-agnostic score that can detect linear or non-linear relationships between two variables (columns). The score ranges from 0 (no predictive power) to 1 (perfect predictive power) and it can be used as an alternative to the correlation coefficients (matrix). The lower triangle of the output matrix presents relations from the x-axis (predictor) to the y-axis (target) while the upper triangle presents relations of opposite direction i.e., from the y-axis (target) to the x-axis (predictor). For more details, see Wetschoreck (2020) `https://towardsdatascience.com/ripcorrelation-introducing-the-predictive-power-score-3d90808b9598` and the R package: 'ppsr'.

• **Pairwise Granger's Causality Tests**. Presents bi-directional pairwise P-Values of the Granger's causality test. The test prior (H0) is that (1) variable X does not Granger cause variable Y and (2) Y does not Granger cause X. Accordingly, the upper matrix presents P-values of (1) namely, a variable in row does not Granger cause a variable in column while the lower matrix presents P-values of (2) namely, column does not Granger cause row. In addition, you can choose the maximum number of lowest P-values to show (4..25) or all pairs that are significant (0.01..0.15 significance levels). The app presents also series that uniquely causes the other but not the opposite. For instance, X Granger causes Y but Y does not Granger causes X. It is particularly useful if there are many factors exhibiting a linear trend, which means that the variables may frequently be the cause as well as causing each other. For a practical example see the help on Granger causality.

• **ANOVA**. Presents the effects of up to three binary/categorical variables on the dependent (numeric) variable in an one-two-three ways ANOVA. F-test, P-Values, Eta-Squared, and Partial Eta-Squared are presented. For explanation on the method see `https://www.statisticshowto.com/?s=ANOVA`.

• **Pairwise Equality Test**. Depicts P-Values of pairwise empirical equality tests. The equality tests include both parametric (Mean, Median, Std) and non-parametric tests particularly: Welch t-Test, Mann-Whitney, Kruskal-Wallis, F-Test, Ansari, Mood, Fligner, Kolmogorov-Smirnov (K-S), and Anderson-Darling (A-D). Most tests can be: two-sided,

greater or smaller. For more information see the attached (to the right) question mark.

• **Meta Data**. Presents first and last dates together with the basic frequency of the data as detected by the CB app. Shows also basic statistics of all original variables including: Series name, type (numeric/Binary/categorical), mean, max, min, number of observations (*#Obs.*), number of missing values (*#NAs*), number of zeros (*#Zeros*), and number of non-unique observations (*#Duplicates*). This option should be the first step of the eyeballing process after uploading any data set.

(4) Imputation

• **NAs Map**. Graphical representation of the percentage of missing data in the selected series and their distribution. See the Help menu for further details.

• **Univariate Imputation**. Three methodologies (Linear, Cubic Spline, and Stine) are optional for identifying and completing missing data in each of the selected series, separately. Clicking on the wide green button below the figure will add imputed series with the same name as the selected ones though with the prefix 'Imp'. For further details on the imputation methods, see the Help/Imputation menu.

• **Multivariate Imputation (N>1)**. Five multivariate methodologies (Cross mean, Copy mean, KNN - K Nearest Neighbors, IRMI - Iterative Robust Model-based Imputation, CART - Classification And Regression Trees) are optional for identifying and completing missing data (N>1). See the Help/Imputation menu for further details.

• **Mean combination (N>1)**. A simple mean combination of all of the eight methodologies above. All series of interest should be Numeric and in their original form, without being converted or transformed, and have a maximal sample period.

(5) Outlier

• **Univariate (N=1)**. Five methodologies for identifying outliers in a single series (N=1) including $\mu \pm K \times \sigma$, Percentiles, Inter Quarter Range (IQR), Hampel (1971) Method, and ARIMA. Basic statistics of each method, positive outliers, negative outliers, and all outliers, are provided in the lower panel. The red button, located below the statistics table (lower panel), shows the observations before and after any outlier (up to +/- 3 dates). For further details see the Help/Outlier menu.

• **Univariate Outlier Summary (N=1)**. This table shows the outlying dates sorted in decreasing order based on the amount of univariate outliers found.

• **Multivariate (N>1)**. There are four methodologies for identifying outliers in several series (N>1) including Cook's Distance, Mahalanobis Distance, Local Outlier factor (LOF), and OGK's Distance. The red button, located below the graph, shows the observations

before and after any outlier (up to +/- 3 periods). For further details see the Help menu.

• **Multivariate Outliers Summary (N>1)**. The table shows the outlying dates sorted in decreasing order based on the amount of multivariate outliers.
Note that in all of the above sub menus it is possible to zoom in on the outliers either by brushing within the interactive graph or by clicking on the red button located at the bottom of the graph.

• **Pairwise correlation (N>1)**. Like in the **Visualization** menu **Pairwise correlation/All** with an additional panel that shows the same figure without outliers that are found by the Robust Mahalanobis distance (see the Help/Outlier menu on this method).

• **Granger's Causality (N>1)**. Like in the **Exploration/Pairwise Granger's Causality Tests/All** with an additional lower panel that shows the same table without outliers that are found by the Robust Mahalanobis distance. For further details see the Help/Outlier menu.

• **Box Plot**. A box plot is shown in the upper panel for each of the selected series, while the lower panel shows the number of pairwise common outliers on a particular date.

(6) **Regression** [ERTK Only]

• **Model Selection**. This selection procedure selects the best regressors in an OLS regression using the Stepwise procedure. Once you select a dependent variable, the best regressors (up to 15) are presented in a table that includes: coefficients, Std. errors, t-values, P-values and goodness of fit measures (Adj. R^2, Akaike Information Criteria (AIC), Durbin-Watson (D.W.), Maximal VIF (a measure of Multicollinearity) among the variables). Then, it is easy to change parameters in the Control Panel (frequency, Transformation, Conversion, sample period) and find out how this affects the best regressors list.

• **Models Comparisons**. This option enables a systematic pre-check of linear relations between the dependent (selected) variable and all other exogenous variables. The check is done by transforming the series (deTrend, smooth, seas, All), converting them (ratio, diffLog, diff, Log, scale, All), changing both the data frequencies (day, week, month, qtr, year, All) and frequency method (mean, median, All). In case the basic frequency is lower e.g., month the frequency options are limited to that frequency and lower e.g., month, qtr, year, and All. There are two regression types: simple OLS and OLS with interactions and for each type a constant is automatically added to the equation which is shown above the model results table. Each entry of that table presents an OLS (or OLS-Interaction) regression model results as follow: Model number (Model), Transformation type (Trans.), Conversion type (Conv.), frequency (Freq.), Frequency Method (Freq.M), Number of observations (*#Obs.*), Adj. R^2, Most Important Independent Variable (MIIV) with its significance level (***, **, and * denote 0.01, 0.05, and 0.1 significance level, respectively.), the Durbin-Watson statistic for serial correlation of one lag (D.W.), the Akaike Information Criterion (AIC), the Bayesian (Schwarz) information criterion (BIC), the Breusch-Godfrey P-val. of test for serial correla-

tion of order up to 5 lags (BG.Pval), and the Breusch-Pagan test against heteroskedasticity (BP.Pval). For the two latter tests, figures lower than 0.05 might indicate a rejection of the null i.e., no serial correlation and homogeneity. All in all, 240 linear models are examined for each one of the two regression types (OLS and OLS-Interaction). Once you see the results table you can select the promising models for further investigation and fine tuning.

• **OLS**. There are four panels that present: (1) The coefficients of exogenous variables with their stepwise ranking and goodness of fit measures, (2) Residuals versus fitted values including outliers (by Cook's distance), and (3) Some residual tests for: (a) Serial correlation, (b) Non linearity and (c) Heteroscedasticity, and (4) The representing formula of the OLS regression. That formula can also be printed or saved (in various formats including Latex) as is, in *Italic*, with the actual coefficient values, or wrapped.

• **Rolling OLS**. Checks the coefficients' stability either by running the OLS with a fix moving window or by an expanding window. The representing formula of the basic OLS regression can also be printed or saved as in the OLS menu.

• **OLS Interactions** - Automatically adds the interactions between the selected exogenous variables (up to six variables). The representing formula can also be printed or saved as in the OLS menu.

• **Robust OLS**. Presents OLS coefficients that are robust to outliers using the methods of (Phi): Huber, BiSquare, and Hampel and compared with the standard OLS. It also depicts basic statistics of the residuals of the three robust methods alongside those of the OLS. Additionally, four dates of the lowest weights by the three robust methods are presented. The actual coefficient values of the representing formula are by the Huber method (1).

• **ARIMAX**. Shows a coefficients table and a residuals graph of the optimal ARIMA(P, I, Q) model with external regressors where P, I, Q are the number of ARs (P), Integration level (I), and number of MAs (Q), respectively. You can choose 'auto' for the auto.arima function (from the R package 'forecast'). Alternatively, you can select either P (up to 5 lags), or I (up to 2 lags), or Q (up to 5 lags) manually. The representing formula of the ARIMAX regression can also be printed or saved as in the OLS menu.

• **GLM (Logit / Probit)**. Depicts a coefficients table and a residuals graph of the LOGIT or PROBIT model in which the dependent variable is binary (two values only, such as 0/1). The represented formula of the LOGIT or PROBIT model can also be printed or saved as in the OLS menu.

• **Order logit / Order probit**. Presents a threshold coefficients table (unique to every state), a common coefficients table (to all states), and a residuals graph of the Order LOGIT/PROBIT model, in which the dependent variable is categorical. The representing formula of the Order LOGIT or Order PROBIT model can also be printed or saved as in the OLS menu.

• **TOBIT**. Exhibits the OLS and Tobit coefficients side by side (Tobit is useful in case of censored data) assuming the cut off value is zero. The representing formula of the OLS/Tobit model can also be printed or saved as in the OLS menu.

• **GARCH**. Depicts a coefficients table and a residuals graph of the selected GARCH model. Enables different types of models including sGARCH, eGARCH, iGARCH, gjr-GARCH, apARCH, csGARCH, various AR and MA lags, several distributional assumptions (Normal, sNormal, Student, sStudent, Ged, sGed, Nig), and external regressors in the conditional variance equation.

• **Markov Switching Model (MSM)**. Exhibits a coefficients table of the various regimes (2 to 4 are available) including switched external regressors, AR(1), and Standard Deviation. It also presents a transition matrix probabilities in a table and both smoothed probabilities and conditional means, by a regime.

• **Quantile Regression**. Shows a coefficients table by the number of Quantile's tau i.e., $n \in (2, 3, 4, 5, 10, 20)$ where $\tau = 1/n$. The representing formula of the basic OLS model can also be printed or saved as in the OLS menu.

(7) **Forecasts** [ERTK Only]

• **Univariate Methods**. Presents a graph of In-sample dependent, fitted, and residuals and Out-of-Sample point estimates with upper and lower Confidence Intervals (available CI are 0.8 to 0.99 and the training set from 60% to 99%). The forecasting methods include: time series LS including trend and seasonality (TSLM), Exponential smoothing state space (ETS), Holt-Winter, auto ARIMA, Trend and Seasonal components (TBAT), and Forward neural networks with a single hidden layer and lagged inputs (NNET). In addition, a forecast accuracy measures of the error (defined as Out-of-Sample actual observations - Mean/Upper/Lower estimates) are presented in a table. All procedures are gathered from the R package: forecast.

• **Multivariate (shrinkage) Methods**. Exhibits Out-of-Sample point estimates and forecast errors using various methods: Principal Components Regression (PCR), Partial Least Squares (PLS), k nearest neighbor regression (KNN), Ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, and Least Angle Regression (LAR). Additionally, a forecast accuracy measures of the error (defined as Out-of-Sample actual observations - point estimates) are presented in a table. All procedures are gathered from the R package: caret.

(8) **Panel/System** [ERTK Only]

• **Panel Builder**. Here you can easily construct your own panel by selecting a variable e.g., the GDP and an ID e.g., country. Once you select them both and click the 'Save Selection & Build' button, the series picker will automatically be updated and you will be able to go on to the following panel menus. This is much easier than selecting many series

one by one especially, when there are many variables or IDs. Notice that panel menus are available only if there are at least 2 variables and 3 IDs. If this is not the case you will see the message: » Not a Panel data ».

• **Panel Statistics**. Presents basic statistics (mean, max, min, sd, Obs., NAs, skew. kurt, Zeros, Dup.) of the panel by a selected variable either by its ID e.g., country, or by date, or by both. The minimum number of variables to construct a panel, is 2 and that of IDs is 3.

• **Panel Test**. Advises on the appropriate panel model by presenting several test results including: OLS Vs. Fixed Effect (FE), Hausman test (Random Effect (RE) Vs. FE), Pesaran's Cross-Sectional Dependence (CD), Lagrange Multiplier (LM) for Time Effects, and Goldfeld-Quandt for Heteroskedasticity. Also, presents panel information such as total number of observations and whether the panel is balanced?.

• **Panel Regression**. Depicts panel regression results by various models (Pooling, FE, RE, RE of Hausman & Taylor, and 2 steps GMM). Under FE and RE models both ID and date specific coefficients are presented too. There are also options to run the model with robust Standard Error (SE), or trim observations such that the panel is balanced. Finally, forecast accuracy measures are presented depending on a training set (from 50% to 99%) and number of 'step ahead' periods (from 1 to 30).

• **VAR (Vector Auto-Regression)**. Presents VAR-X results depending on the selected lags (up to 5 lags), model type (none, constant, trend, or both), and the exogenous variables (without lags). It also depicts residual graphs of the VAR models and multivariate residual tests in a table (Optimal number of lags, serial correlation test, ARCH test, and normality test).

• **Cointegration Checks**. The upper panel shows the Johansen Cointegration Test's Critical Values and Vector Error Correction (VEC) Statistics depending on the selected lags (from 2 to 6) and type (constant or trend). Two stars attached to the figures in the 5% column reflects the optimal rank. The lower panel presents cointegration relations among the selected variables (and a constant or trend) in which the eigenvectors are normalized to the first variable - first line).

• **VEC (Vector Error Correction)**. Depicts an Impulse Response Function (IRF) results or Forecast Error Variance Decomposition (FEVD) in a graph depending on the selected lags (up to 5 lags), model type (none, constant, trend, or both), number of steps ahead (1:20), confidence interval (none, 0.9, 0.95, 0.99), and whether the shocks are cumulative or not. Since this VEC procedure implements the Cholesky decomposition one can change the order of the variables (using Sort Type in the Control Panel) to test whether the output is sensitive to the order of variables.

• **Structural VAR**. Presents coefficients of the Structural VAR (SVAR) model depending on the selected lags (up to 5 lags), model type (none, constant, trend, or both), and the

SVAR model & constraints. Available models are matrix A or B, matrix AB, or Blanchard and Quah matrix type. The SVAR constraints are determined by clicking on the two upper (side by side) matrices, A and B, that are toggable between 0 and empty cell and where 0 means no constraints while empty cell reflects a constraint (equality) of the two respective variables in line and column. In SVAR model 'A or B' the coefficients are affected by matrix A or matrix B independently, in SVAR model 'AB' both matrices affect the coefficients (do not set empty cells in the same place in both matrices, A and B), and in Blanchard and Quah model the upper panel presents a contemporaneous impact coefficients matrix while the lower panel shows identified long-run impact coefficients matrix.

• **Structural VEC**. Shows an Impulse Response Function (IRF) results or Forecast Error Variance Decomposition (FEVD) in a graph depending on the selected lags (up to 5 lags), model type (none, constant, trend, or both), number of steps ahead (1:20), confidence interval (none, 0.9, 0.95, 0.99), and whether the shocks are cumulative or not. Similar to the SVAR, constraints are determined by clicking on the two upper (side by side) matrices, A and B, that are toggable between 0 and empty cell and where 0 means no constraints while empty cell reflects a constraint (equality) of the two respective variables in line and column. In model 'A' or model 'B' the coefficients are affected by matrix A or matrix B independently while in model 'AB' both matrices affect the graph (do not set empty cells in the same place in both matrices, A and B).

• **System - Two Equations**. Presents results of a two-equation system including coefficients of the two equations, goodness of fit, and 6 panel plots of the: two independent variables, the exogenous variables, and the residuals of both equations. The available models are: OLS (two independent regressions), Weighted LS (WLS), Seemingly Unrelated Regression (SUR), 2 Stage LS (2SLS), 2 Stage Weighted LS (W2SLS), and 3 Stage LS (3SLS).

(9) **Utility** [ERTK Only]

The aim of the utility procedures is mainly: (1) downloading historic micro and macro data directly to the current data set and (2) to allow the publishing of the outputs namely, creating tables and graphs on the fly.

• **Download My Institution's Data**. This menu is the same as the last one except two things: (1) it is relevant for in-house data rather than free Web data. (2) You must enter your Institution's Username and Password. Please consult your IT experts regarding the full URL addresses of the in-house data tables.

• **Build a Report**. This item assembles all figures and tables you select for building a report. First, you must select the files to be included in the report. Then, select an output file format (word, pdf, Tex, html, power point, slidy). The selected input files that have the appropriate extension (eps, tex, jpeg, png, bmp, svg, wmf, html, csv) will be included in the report; each output in a separate page. The report, named "CB_YYYY-MM-DD" (default) or any other name you choose, will be saved once you click the green button (Download Report...). Then, you can add explanations etc. turning the output report into a complete empirical paper.

• **Publish Figure**. After uploading any PNG file particularly, a file that was created by the 'Figure' button (located in the upper panel of the Control panel), you can customize the figure by adding a stylized caption and/or footer using the 'Set parameters' menu. The stylized figure may include: Alignment, font size, font family (including bold and italic), color, spaces between lines, grid (border) lines (including line type and width). Finally, save the figure as DOCX, PDF, HTML, or PNG file.

• **Publish Table**. Once you select this option the last table is ready for editing before publishing. You can stylize the table's body, headers, caption, and footer using the 'Set parameters' menu and save the table as PPTX, DOCX, PDF, HTML, or PNG file. In addition, you can edit each cell in the table body. The stylized table include: Theme, alignment, font size, font family (including bold and italic), color, spaces between lines, column width, number of decimal points, border lines (including line type and width). Moreover, one can stylize part of rows or columns by checking/unchecking specific rows/columns.

• **Preferences**. Here you can change the date format and the default number of decimal points (you can always change them later using the respective menu in the Control Panel).

(10) **Help** Includes help files particularly for the statistical procedures used.

(11) **The Ten Commandments**

Here are the CB's **ten commandments** of what to do and what to avoid:

1. One column in your data set should be a valid date in any format (e.g., 'dd/mm/yyyy' or 'yy-mm-dd'). A column of this type must be present in any time series data, and if the app does not detect one, it will generate an error message.
2. Select short (**English only**) names for your variables. Particularly, do not use spaces, or other non alphabetic letters (except _) within the variables names as the app converts such letters into dots. If your uploaded file is organized as a panel the variable names should be short as possible and with alphabetic characters only. This is required because during the uploading process CB turns any panel data into flat file format, using series name prefix as the common variable while the series name suffix is the idiosyncratic ID. For instance, if the uploaded panel contains an ID columns e.g., country (Afghanistan, Albania,..), a time columns e.g., date (2020-1-1, 2020-1-2....), and a numeric (data) column of temperatures then, the series names' prefix of the new series will be: hot.Afghanistan, hot.Albania etc.
3. Use the question marks (?) nearby menus for the methods used and operational tips. See also the respective Help files. Note that for the time being some of the procedures do not have help files.
4. Use the 'State' option (in the Control's upper panel) after merging files. Next time you will be able to uploading the last state at the point they were left last time.
5. In almost all graphs you can Zoom-In for further data insights (by mouse brushing of the investigated area) and you can download them by clicking on the wide green button (in the Control Panel) for further usage particularly, building a report.
6. Do not change parameters (mainly on the left panel but also below graphs or tables) too often. Wait between two consequent changes especially for time consuming procedures (sym-

bolized in the menus by TT - (Takes Time)). If the slider or other inputs start bouncing go temporarily to another frequency e.g., from Day to Week. If it still bouncing go temporarily to the 'Meta Data' sub menu under the 'Exploration' main menu.
If the filter (Categorical/Binary series) begin to bounce, temporarily deselect them, and then slow down the pace of the filters' changes. Additionally, if you change the filter values, the slider is also affected; therefore, you should return the slider to the previous position or any other state you prefer.
7. Do not mix data frequencies e.g., daily with weekly data or data types e.g., change rate (ratio) with logs in your input files.
8. Do not select many variables in multivariate procedures (including regressions) if you have relatively few observations. Notice that CB in its current version is not intended for big data analysis.
9. Do not forget to save your figures by downloading them using the wide green button (in the bottom of Control Panel) and tables by using the menu above them.
10. Do not leave CB in case of temporarily no response. It can be the reason of time consuming procedures (TT) or bug in the beta version (in such a case please send me an email with the case details). Try to refresh the web page by (1) Changing one of the parameters in the Control panel e.g., 'Frequency method' in order to see the expected change in the Output Panel, (2) Changing the menu selection to another one e.g., 'Meta Data', or (3) Re-upload the data set using the 'New' menu at the Control's upper panel. Since the latter may take more time, only use it if the others two do not help.