National University of Singapore
School of Continuing & Lifelong Education (SCALE)

TBA2105 Web Mining
Tutorial/Lab 1

Learning Objectives

- Perform Exploratory Data Analysis
- Perform Data Wrangling using `dplyr`
- Perform Visualization using `ggplot2`

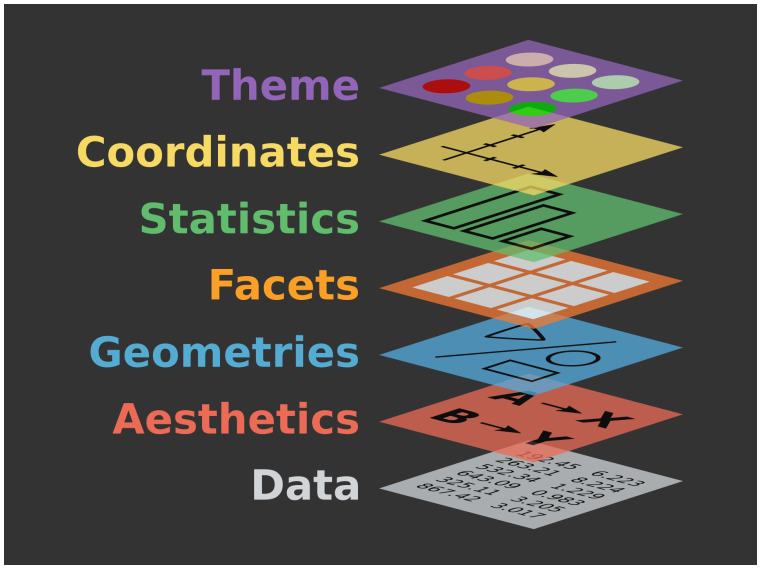
1. In this exercise, we will be using the `GooglePlayStore` (`googleplaystore.csv`) dataset which can be downloaded from LumiNUS/Autograder. This dataset consists of 10841 observations and 13 columns. The dataset was originally obtained from Kaggle (https://www.kaggle.com/lava18/google-play-store-apps) and the description of the columns can be found on the website.

   a) Install and load the `dplyr` and `ggplot2` R packages.

   b) Load the dataset into the `gplay` variable and remove away rows that contain NA values.

   c) Apply string manipulation to the `Installs` column and convert it to a proper numeric form. *Hint: read up on* `gsub()`

   d) ggplot2 is a charting tool for creating graphics based on the Grammar of Graphics. Each chart mainly consists of a combination of different types of layers (data, aesthetics, facets, statistics, coordinates, theme). Refer to the ggplot2 cheatsheet: https://raw.githubusercontent.com/rstudio/cheatsheets/master/data-visualization.pdf and generate a **scatterplot** of Reviews (y-axis) vs Rating (x-axis).



e) By looking at `gplay` dataframe, it can be observed that the `Reviews` column can be a small or large number. Thus, it makes sense to standardize the column by applying a log transformation. Regenerate the scatterplot from d) by applying a log transformation

f) Apart from the x-axis and y-axis, it is possible to add another dimension of visualization by coloring the scatterplots by its `Category`. Regenerate the scatterplot from e) to include the color dimension.

g) Try generating the same scatterplot but coloring using `Type`. What insights can you derive from this scatterplot?

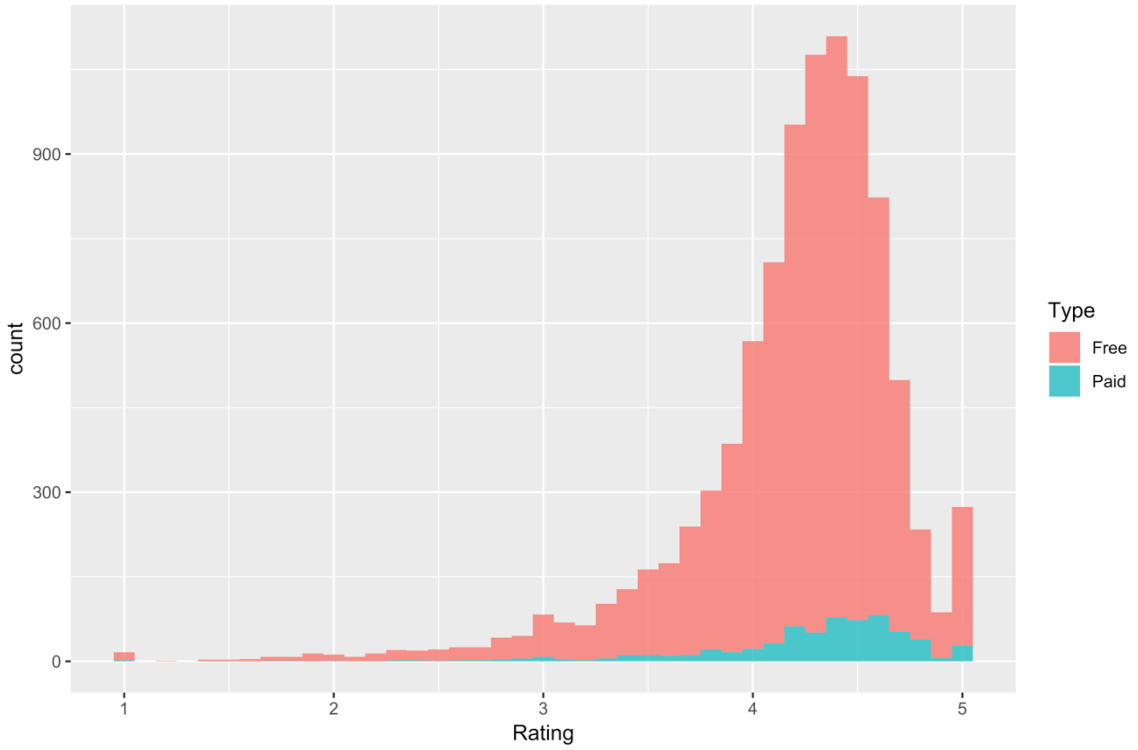
h) To perform various data manipulation, we could use the `dplyr` package. The idea behind `dplyr` is similar to `ggplot2` in that it also provides a grammar idea for doing data manipulation. Specifically, it provides a set of verbs for performing common tasks such as filtering, grouping, summarizing, and mutating data. See https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf. Try using the filter() verb to obtain a dataframe of Paid apps.

i) dplyr provides an easy way to perform exploratory data analysis using the `summarize()` function. We would also likely to want to use it together with the `group_by()` functions. The idea is similar to how we sometimes write SQL group-by statements. Use the `group_by()` and `summarize()` functions to obtain the mean rating (`meanRating`) and mean number of reviews (`meanReviews`). The output should be a tibble with 3 columns: `Category`, `meanRating`, and `meanReviews`.

j) We can further generate more visualization to better appreciate the data by category. Use ggplot to plot a **bar chart** results from i) (particularly `meanRating` vs `Category`), making the `meanRating` the y-axis and `Category` the x-axis. You should also color the bars with the `meanRating` value. To ensure proper display of the x-axis label, you should also rotate x-axis 90 degrees. Is there any insight we can derive from this bar chart?

k) To investigate whether there is a relationship between `Rating` and `Reviews`, use the `lm()` function to generate a regression model (`model1`) where `Rating` is the response and `Reviews` and `Installs` are the predictors.

l) Finally, we want to investigate based on the distribution of the apps to see whether there is a difference between the rating distribution of Free vs Paid apps. Try a histogram to obtain the distribution. You should use a bin width of 0.1 and alpha value of 80%). The histogram should look like this. What can we conclude from this chart?


900
600
300
0
count
1
2
3
4
5
Rating
Type
Free
Paid