Group 1: Bozzelli Giammarco (884962), Cavaliere Marco (887075), Isaac Lupi (885004), Parigi Michele (884948), Sibilia Beatrice (885016)

Group 1 - Mobile App Stores Analysis

Main idea: Study datasets of mobile applications of different App Stores' distributors (*Microsoft, Apple, Google*) and see whether we can find: common characteristics, users' preferences and correlations between different Apps' features

Data Loading and Cleaning:
We started with four different App Stores (*Microsoft Store, Applestore, Playstore* and *Steam*) but after a discussion we decided to discard the Steam dataset since it does not sell mobile apps, agreeing that the results of our projects would have been misleading if we were to use it.

We decided to load and clean the datasets one at a time and merge them together later since, by looking at them in a text editor, we realized they were really different and cleaning them all at once would have been counterproductive.

Notes on the cleaning process:
- Datasets had differente price currencies so we converted everything in USD ($)
- *Microsoft* dataset didn't have a '*size'* column so we added it and filled it with *zeros (0)* so that we could work without NaN values
- Some datasets had columns for *'reviews'* and *'installs'*, other had only *'reviews''* and other only *'Number of people rated'* so we decided to remove the *'installs'* column, and merge *'reviews'* and *'Number of people rated'* under a new *'Interactions'* column and use it as an indicator for popularity.

Q.1 Which App category do users prefer?

Since we wanted to find the category that users prefer, we had to give applications a more realistic rating, this because our original datasets had 5 stars rated apps with few interactions and we couldn't consider those apps on the same level of high rated apps with a lot of interactions. To give apps a "weighted rating", taking into consideration also the number of interactions for each app, we created the *"top_categories_weighted"* function in the *appstore module* that takes as input a Dataframe and optionally the store name that we want to study, and returns a dictionary with categories as keys and the weighted ratings as values.

We also created another function called *"graphic rating"* to create and print in a graphical way the dictionary coming from the *"top_categories_weighted"* function. We first analyzed each store differently using the above two functions, then we studied all data using the merged dataframe.

Results Q.1: All Stores have different preferred categories, in fact:

- Play Store --> Health & Fitness
- App Store --> Lifestyle
- Microsoft Store --> Education

But using the whole DataFrame we found that the highest rated category is **Health & Fitness**, which is also the preferred one in the Play Store.

Q.2: For which category are users willing to pay more?

In the first moment, this question was very tricky since we had to find out the right correlation between ‘*interactions’* (the indicator of the popularity) and ‘*price’* for each category. We proceeded in the following way:

For each category we computed the *price mean* taking into consideration all apps inside that category. Eventually we computed using the same method the *mean of interactions* per category

- **Problem:** We needed to find a correlation between mean of interactions and mean of price to find out which category was the most appealing one for the consumers.

-**Solution:** We solved this problem by multiplying the mean of interactions by the mean of price per app category, finding the mean revenues for each app, thus how *much people are willing to pay* on average for each app category.

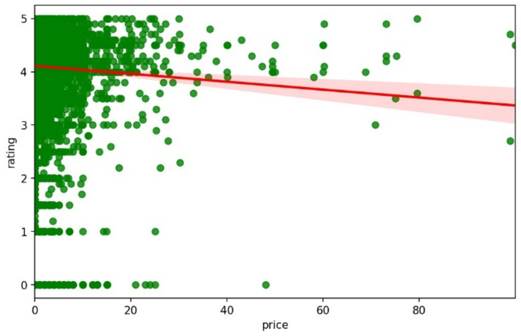
Results Q.2 : By representing graphically the data we found that applications in the **Photos & Videos** category are the most paid by the users.

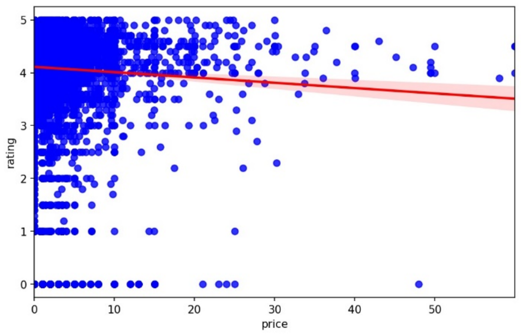
Q.3 Does users’ rating depend on price?

This question was interesting because we could make a prediction about it: the applications that cost more should be of a higher quality thus they should get higher ratings.
First of all, looking at the data set we noticed that there were just a few applications over the 100$ price, so we decided to take them away in order to have a more clear analysis.

We plotted a scatter plot with a regression line (**green**) with Seaborn, but looking at the graph, we noticed that there was nearly no correlation between ratings and price because the regression line was flat. At this point, we decided to try only with applications with a price lower than 60$ in order to have a more concentrated graph (**blue**) and here is the final result.

Results Q.3: As you can see, there is still no correlation between rating and price. This means that our prediction was wrong, in fact, even applications with lower prices get higher ratings.

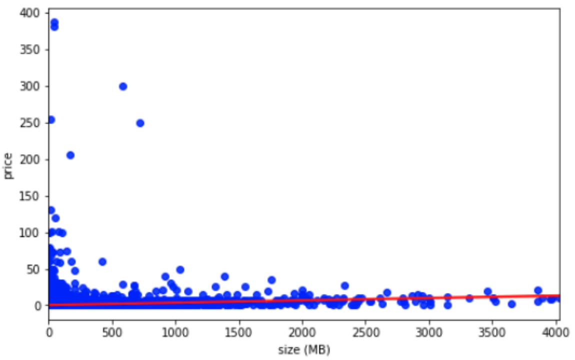


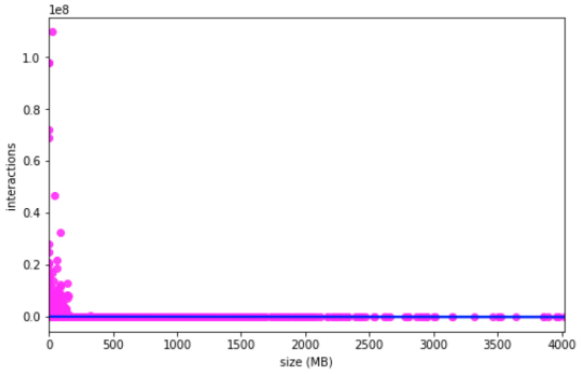


Q.4 Do ratings, popularity and prices depend on applications’ size?

The fourth question represents a correlation between four application characteristics: *price*, *user interactions* and *ratings*, and *size* expressed in Megabytes.

We decided to use the linear regression method, which, through a mathematical function, determines the relationship between two variables. To do this, we first created two functions. The first was to find the equation of the line, which specifies its intercept and slope. The second allowed us to find the linear correlation coefficient and the coefficient of determination, which represents the goodness of fit of the regression model found. We applied these formulas to the three separate cases, also representing the result with graphs using Seaborn.





Results Q.4: We can conclude that in all cases, we could not see a real correlation between the variables. We have determined this because the three linear correlation coefficients are very close to the value 0. Even from the graphs, in which the lines are represented almost horizontal, it can be understood that the values found are too small to determine evidence in the relationship

Q.5 Do most popular Apps have some characteristics in common?

All the possible common characteristics among apps we could find were: *size*, *category* type, *price* and *store* type. We decided to study each of them separately and plot the results.

First of all, to define which are the most popular apps we sorted by *'rating'* and *'interactions'* the merged data frame without all rows with *zero values* in the *'size (MB)'* column. Then we created a new dataframe with the first 50 apps of our sorted data frame.

To analyze this new dataframe we build four new functions :
1 - '*dictionary_top()'* that takes as input a *dataframe* and a *column name* and returns a dictionary with each value in the column as keys, and their frequency as values
2 - '*top_50_catplot()'* that takes as input a dictionary and plots a Bar Chart representing the data.
3 - '*top_50_sizeplot()'* that takes as input a DataFrame and uses the *name* and *size (MB)* columns to plot a graph representing the size in MB of each application
4 - *'top_50_priceplot()'* that takes as input a frequency dictionary and returns a graphical representation of it.

The outcome was unexpected: top apps didn't have many interactions, something that according to our assumptions means not being popular.
Later we saw that by ordering the starting dataframe sorted by '*interactions'* and '*rating'* results changed.

Results Q.5: This said, after looking at the results the latter method we used seemed more correct, thus we can say that applications in the top 50 have these characteristics in common:
- Category = Games
- Size <= 150 MB
- Price = Free
- Store = Playstore