**ALY 6070 DATA COMMUNICATION AND VISUALIZATION**

**EFFECTS OF SUCCESSFUL SCRATCH PROJECTS**

By

Amol Dasani, Shengyuan He, Kaushik Burra, Shuhang Li.

### **Table of Contents**

* Introduction
* Research Questions
* Data Source
* Data Cleaning
* Methods
* Analysis and results
* Conclusion

### **Introduction**

The scratch community is a digital platform where young minds concur to create, share and interact among themselves building multimedia projects using the Scratch programming environment. The primary goal of this community is to make it easier for researchers to study how young minds learn, create, communicate and interact in an informal digitized learning environment.

The scratch platform also enables the users to work in diverse fields and grow in large by sharing with others. The Scratch community works hard to build service community projects that drive the student's innovations and bring out their qualities and inspire each other’s work.

Scratch is an innovative way to let people advertise their projects too. And a good project can inspire the user's creativity and encourage more users to participate, thereby enhancing the importance of digital learning.

It is our goal in this database to utilize the tools we have learned to obtain the information we need. Our aim is to find out how the Scratch community works hard to build successful community projects by understanding the “**variables”** **followers, followed, favorites, downloads and viewers**, also dig deep in understanding their pattern in different countries. Furthermore, analyzing the correlation among “**factors” such as images, sounds, scripts, sprites and blocks** with the above variables in different countries and bring out the inference based on the data analysis.

### **Research Questions**

Some of our followed research questions are as follows (merging data and slicing by year) :

1. What is the trend between number of projects being followed and favorited over the year 2007 to 2012? (Plot 1)
2. Which country has projects being more followed,favored,downloaded and viewed and how does it influence the trend over years 2007-2012? (plot 2)
3. Based on the parameters images, sounds, scripts and sprites, how does the behaviour of variable(s) change for a project? (plot 3)
4. How do the derivativeness of parameters and variables change with respect to each country over the year 2007 to 2012? (plot 4)

### **Data Source**

The data set was provided to us by our instructor, Professor Alice Mello. The original source of the data set is derived from LLK (Lifelong Kindergarten) group at MIT labs.The variables used in our analysis are **Projects, Views, Favorites, Users, downloads, and Followers.**

### **Data Cleaning**

This dataset was provided in a clean format except for few missing values and unfiltered variables. We have filtered the top number of users based on the projects and established our analysis. Secondly, the part of regions (unfiltered) which belonged to countries are filtered into country itself. For example, the “mushroom islands” are filtered into Japan as they belong in Japan territory etc.

Null values:

* Radar chart (plot2): The null values have been excluded specifically for the countries alone in the radar chart.
* Heat map chart (plot 3): The null values of countries and factors including images, sounds, scripts, sprites and blocks have been excluded in heat map chart.

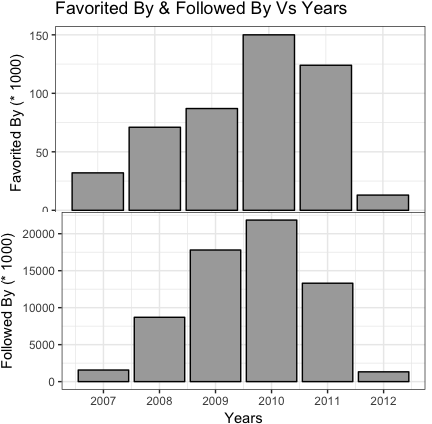
### **Methods**

We explored the dataset and brought an interesting theme for our paper. We based our findings on how the Scratch community diversified in expanding successful projects. We performed R programming on the dataset given to us, merged the datasets and plotted bar and radar graph to find out trends in successful projects based on variables followers, followed, views and downloads. Furthermore, we also built a correlation matrix on the factors responsible for establishing successful projects. Our analysis was mainly based on the fixed variable **“projects”** and sliced by **“year”.**

**Analysis and Results**

**Plot 1: Bar Graph**

We decided to make two charts to represent the relationship between the number of projects being followed and number of projects being favourited based on the year trend from 2007 to 2012.In this case, we believe that using bar charts is a great choice. It is very useful when presenting a series of data and displaying trends over time



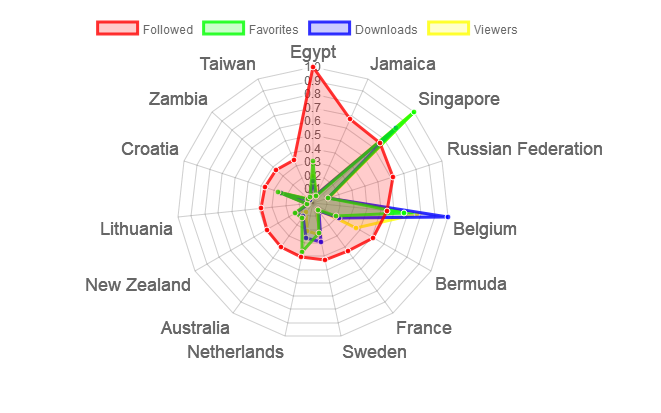
Look at these two vertical bar charts. They share the same horizontal axis which is based on a six-year period. (Although we only have one quarter of data in 2012, it’s not bad to think of it as a year-round reference).

In the upper one, the vertical axis shows the number of projects that are favourited categorized by year. We can easily see that the trend for the first four years has been steadily rising. In the year 2010, it peaks. And then there is a small drop in next year.

In the second chart, we can see how the number of projects being followed over time. The trend looks really like the first one, especially in the first four years. But it is important to note that there’s a big drop here in the next year. Therefore, we can see from here that these two data trends are not quite similar.

From the above analysis we infer that the scratch projects have become more and more successful over time. Kids have ingested the habit of working on these scratch projects collaboratively and utilized in grasping new concepts and learning over the years. After we analysed the trends of followers and favourites that the projects get in years, let’s see what it would look like to analyse the projects in various countries. In this case, is the bar chart still the best option? Can we visualize these data more intuitively?

**Plot 2: Radar chart**



The radar chart is built using polar coordinates which is visually effective and sound for the audience to look at the relationship between the projects with favorites, followed, viewers and downloads based on countries.

The radar chart changes or varies when we change the year from 2007 to 2012. We developed a radar graph to establish an average of the successful projects in top fifteen countries (excluding the null values of countries alone) having high number of viewers, downloads, favorites and being followed the most.

From the radar chart illustrated above we analyze that Egypt has maximum projects that are being followed. The rest countries have an almost similar ratio of projects being followed that are far less than the projects being followed from Egypt.

Furthermore, considering the variables favorites, downloads and viewers, we analyze that the trend being high in Singapore and Belgium, the rest having a similar ratio.

We also found an interesting analysis that the plots among downloads, favorites and viewers are overlapping on most of the countries, which explains their correlation too.

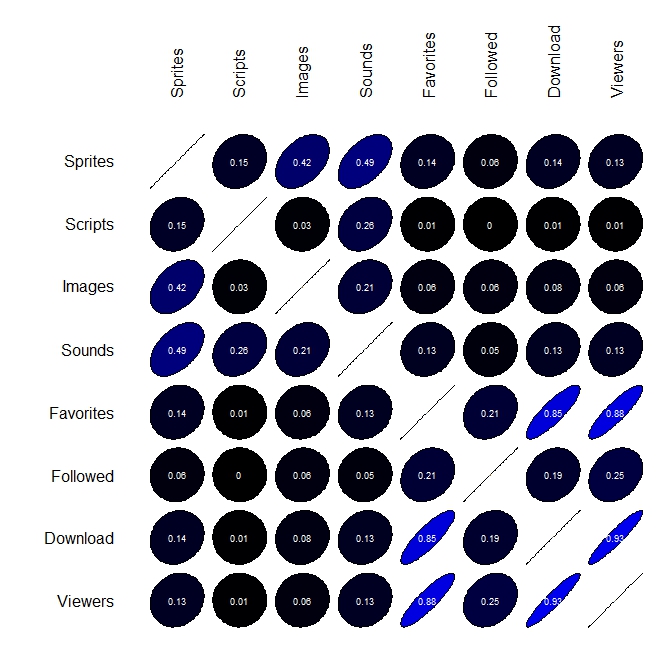
We concur from this analysis that even though Egypt has more number of projects being followed, it may not establish the fact that those projects are successful in ingesting digital learning to empower kids. Similarly, Singapore has the highest ratio of projects being favored and Belgium has the highest number of projects being downloaded, however it still does not attract the users to follow these projects than those in Egypt.

However, we also do find an interesting analysis, it shows that the variables downloads, views and favorites fairly overlap on top of one another for most of the countries, which explains the correlation. (**Further correlation matrix is charted in plot 3 to explain clearly**).

Thus, we can infer that, the projects that are being followed does not establish the fact that these projects are more successful, however the projects that have more number of downloads, views and favored more, may tend to be more successful, depending on the change in year.

.

**Plot 3: Correlation matrix**



To analyse this question, we intend to use correlation matrix to explore the correlation between variables to identify what factors could help a project become successful. What we can derive from the correlation matrix is the strength of the correlation R among the **“variables”** and “**factors”.** This is illustrated by the intensity and direction of the correlation coefficient between two variables is always between +1 and -1. If the value of r is close to -1 or 1, the correlation between two variables will be stronger and the image will be like a line. Conversely, if the value is close to 0, the image will be nearly circular.

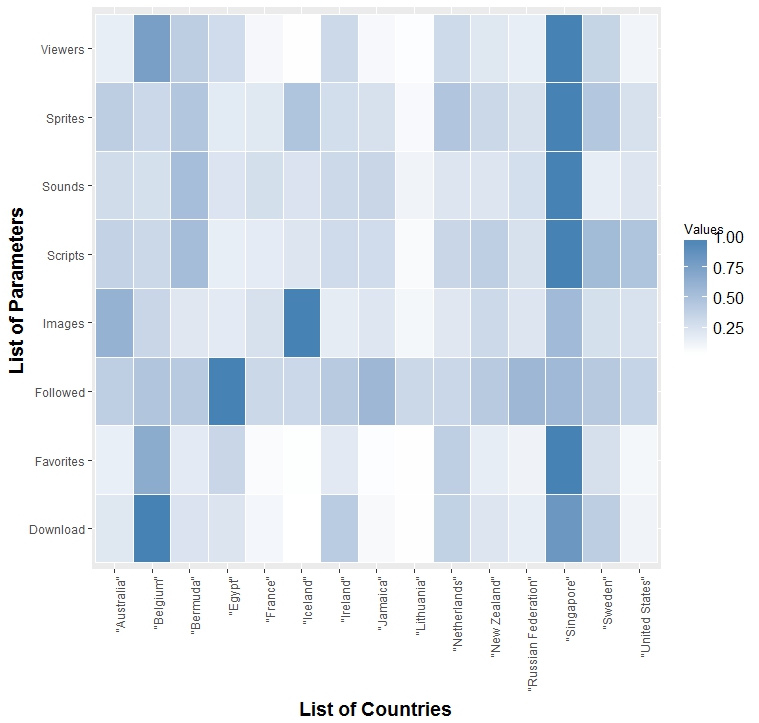
First, we found that sprites, scripts, images and sounds have a very low correlation with favourites, which means there is no necessary correlation between these variables. This also shows that whether a project can be favourited or not doesn’t depend on whether it has more contents, like sprites, scripts, images, and sounds. So even a project with less content can get more favourites.

In addition, we also discovered the contents of a project, such as sprites, scripts, images, and sounds, have a low correlation with followed, downloads, and views. Therefore, we infer that these items of the project cannot determine whether a project is more popular. There must be other factors that affect the popularity of a project.

Finally, favourites have a very low correlation with projects being followed, but a high correlation with the number of download and view. This shows that a project has higher downloads or views may also be more favourited.

We cannot find out in detail which exact and small factors determine the number of favourites and follow of a project. However, we can decide whether a project is more popular based on its number of downloads and views.

**Plot 4: Heat map chart**



The heat map shows the degree of parameter derivativeness in each country over the years 2007-2012. From the above plot we analyse the degree of each parameter for the top fifteen countries (countries being filtered based on the variables **“followed”** and **“favourited”).**

From the above graph we find that the country Singapore has a highest degree of parameters favourites, downloads, viewers, sounds and sprites. This explains that during the year 2007, Singapore has created higher successful projects to empower young minds in digital education.

Similarly, the trend changes once we change the year as the data changes over time. This way we can find the trend of countries having more successful projects over the span of five years from 2007-2012.

### Conclusion

We as a group have ingested the data sets, studied them, analysed and plotted various charts to establish our stats on how scratch projects can be successful with respect to the variables and parameters given in the data sets. Scratch community certainly works hard to create successful projects based on all these factors and empower young kid’s education in simple, informal, and creative aspects.

We demonstrated on how the variables downloads, views and favourites have a positive correlation and may be the reason for a project to be successful, but not just with the thought that projects being followed more become more successful.

We worked on analysing how each country shows its presence in providing successful projects through various means to make sure all the kids from various countries learn, grasp and create a digitized educational society in learning new things.

### References

1. Community service ideas for kids: Why giving back matters, by  [Marilyn Price-Mitchell,Ph.D.](https://www.rootsofaction.com/author/marilyn-price-mitchell/), <https://www.rootsofaction.com/community-service-ideas-for-youth/>.
2. Community service ideas for kids all ages, by Barbara Shelby, <http://www.kidactivities.net/post/community-service-ideas-for-kids.aspx>.
3. Keeping Pace with K-12 Digital Learning: An Annual Review of Policy and Practice. Eleventh Edition, by Watson John, “ <https://eric.ed.gov/?id=ED558147>”.
4. Value of digital learning,” <https://openlearning.mit.edu/value-digital-learning>”.

1. The impact of digital technology on learning : A summary for the education endowment foundation, by professor Steven Higgins, ZhiMin Xiao and Maria Katsipataki, School of Education, Durham University, <https://pdfs.semanticscholar.org/d26b/b59f2536107b57f242b8289b1eb6f51d8765.pdf>.
2. A study of the Effects of Digital Learning on learning Motivation and Learning Outcome, by Ming-Hung Lin, “ <http://www.ejmste.com/A-Study-of-the-Effects-of-Digital-Learning-on-Learning-Motivation-and-Learning-Outcome,69635,0,2.html>”.
3. A longitudinal dataset of five years of public activity in the Scratch online community,” <https://www.nature.com/articles/sdata20172>”.
4. Teachers report mixed impact of Digital Media, by Ian Quillen,” <https://www.edweek.org/ew/articles/2012/11/07/11techsurveys.h32.html>”.
5. Impact of students use of technology on their learning achievements in physiology courses at the university of Dammam, by Mohammed T. Al-Hariri, <https://www.sciencedirect.com/science/article/pii/S1658361216300683>.