**Index**

Abstract & introduction ---------------------------------------------------------------------------- 3

Analysis of projects viewed, loved and downloaded ------------------------------------------- 4

Analysis of projects being loved ------------------------------------------------------------------ 7

Analysis of new Users and Projects -------------------------------------------------------------- 9

Analysis on behavior of most viewed projects ------------------------------------------------ 12

Analysis of relation between Lovers, Scripts, Sprites, and Downloads -------------------- 15

Conclusion ----------------------------------------------------------------------------------------- 17

Reference ------------------------------------------------------------------------------------------ 18

**Abstract**

This paper is written based on visualized statistical data from ALY6070 midterm group project. We write this paper to analyze how each element in the database influences each other, then we will make a comparison between our analysis result with other groups’ dataset.

**Introduction**

Scratch is a programming environment and an online community created at MIT Media lab where young people can learn, create new projects and communicate with other users. We have the dataset of public activity in the Scratch online community during the first five years from 2007 to 2012.

The dataset consisted of 42 variables after combing the datasets of all the groups. The dataset was enormous and had many aspects to it, so we cleaned the data using data interpreter of Tableau and also removed some outliers. We will concentrate on the following variables in this project report:

Viewers website, Downloaders website, lovers website, Scripts website, sprites website, project id, user id, is remix, saves, Images and sounds.

We will separate the main body of this paper into six parts, each part related to a particular requirement in the assignment introduction. Each of the six parts will contain the following ****contents:

Data source and variables used in analysis, methodology, graphs, and findings.

**Analysis:**

**1) Analysis of projects viewed, loved and downloaded**

**Data source used & variables used for analysis**

The focus of data analysis in this part is to explore the relationship between projects and viewed, projects loved and projects downloaded. Regarding variables, we use viewers as the independent variables, lovers, and downloaders as dependent variables to perform the analysis.

**Methods used in this part**

Scatter plots can show the distribution of data points on the Cartesian coordinate system in a regression analysis, and the scatterplot shows the approximate trend of the dependent variable as a function of independent variables. Especially, when we want to figure out the correlation between two variables, R and R^2 value could be a reliable factor ****for measuring the pattern of two variables which leads us to use scatter plot for this analysis.

**Findings**

The graphs below show the pattern between viewers, lovers, and downloaders within the dataset of group 1 to group 5:

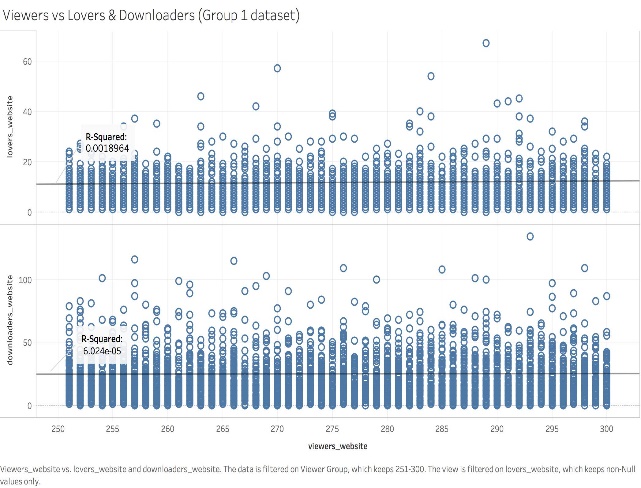
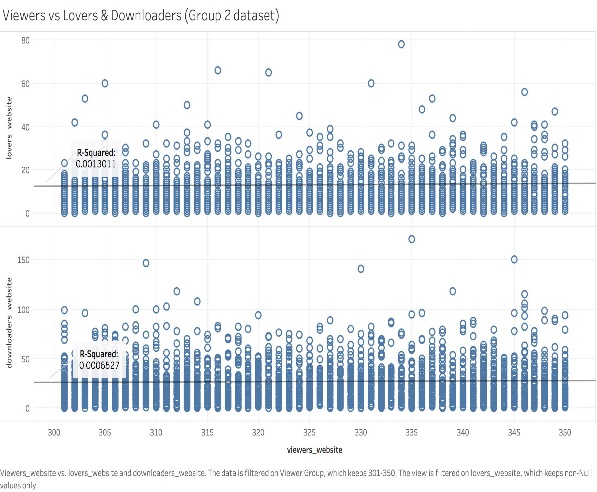
 

Figure 1. Scatter plot of group one Figure 2. Scatter plot of group two

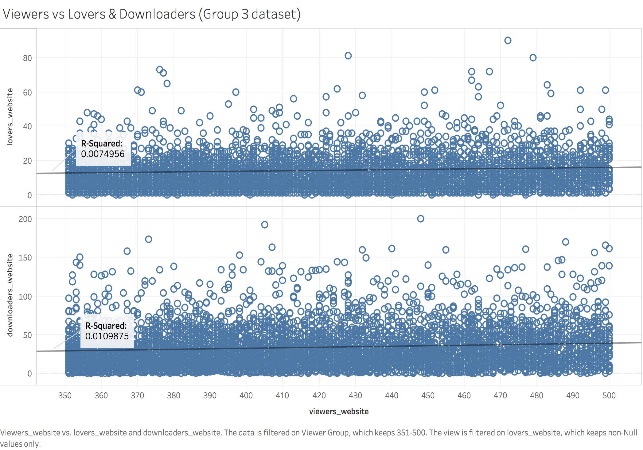
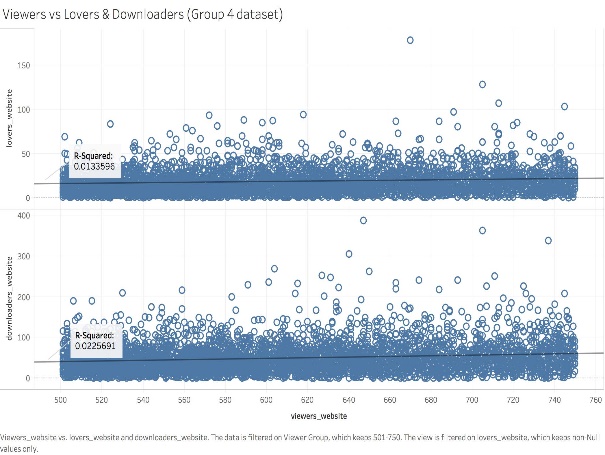
 

Figure 3. Scatter plot of group three Figure 4. Scatter plot of group four

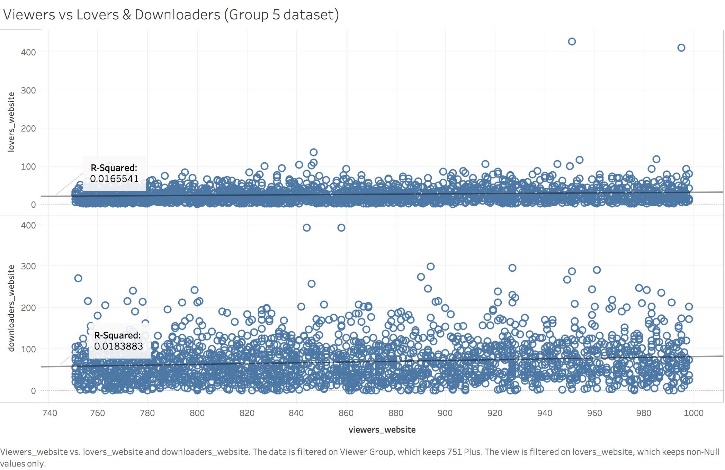
****

Figure 5. Scatter plot of group five

As we know, the correlation coefficient which is known as R-value is a good factor to describe the correlation between two variables and the more the R-value of the trend line is equal to 1 or close to 1, the more reliable the correlation is. Thus, we can get the correlation between viewers, lovers, and downloaders by seeing the R squared value shown in the graphs.

The value of R squared between viewers and lovers in the first graph is 0.0018964 which shows a weak positive correlation between lovers and viewers while the viewers range from 250-300; and the value of R squared between downloaders and viewers in the first graph is 6.024e-05 that indicates there seems no correlation between the number of downloaders and viewers ranging from 250-300. In the same way, we could get the R squared value of viewers vs lovers and viewers vs downloaders in the other four range are respectively 0.0013011 and 0.0006527; 0.0074956 and 0.0109875; 0.0133596 and 0.0225691; 0.0165541 and 0.0183883. Through the results, we get that in every range there ****is a weak positive correlation between viewers, lovers, and downloaders.

**2) Analysis of projects being loved**

**Data source used & variables used for analysis**

Through figure 1 to figure 5 we know that viewers have a weak positive correlation with lovers and downloaders, but we also notice that the correlation increases a little stronger with the number of viewers increasing. To figure out what leads the projects being loved, we chose the lovers, sprites, and scripts as variables to explore the outcome. Based on the data analysis we did in the midterm assignment, we notice that there are fewer data containing lovers with every group dataset, in order to make the results more reliable, we choose to combine five groups’ data together as a whole dataset as the data source.

**Methods used in this part**

Since we consider the sprites and scripts as the factor that affects if a project being loved, so we still need a plot for describing the correlation between every two variables where the scatter plot must be a great choice. Different from what we did in figuring out pattern between viewers, lovers, and downloaders, we choose to use the average value of lovers in every viewer range and the average value of the number of sprites and scripts to draw the scatter plot which is much clear and easy to understand.

**Findings**

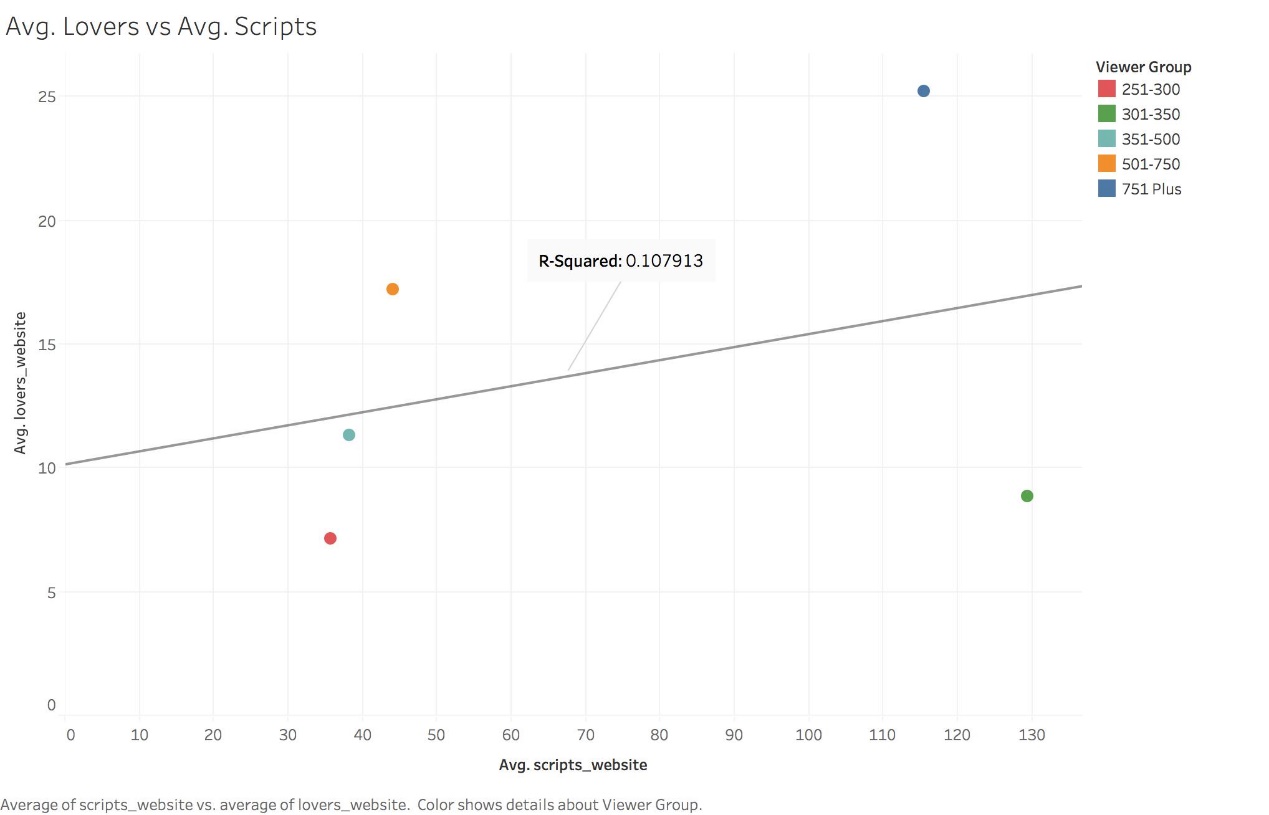
****The graphs below show the pattern between the average number of viewers, the average number of scripts and the average number of sprites:

Figure 6. Pattern between average lovers and average scripts

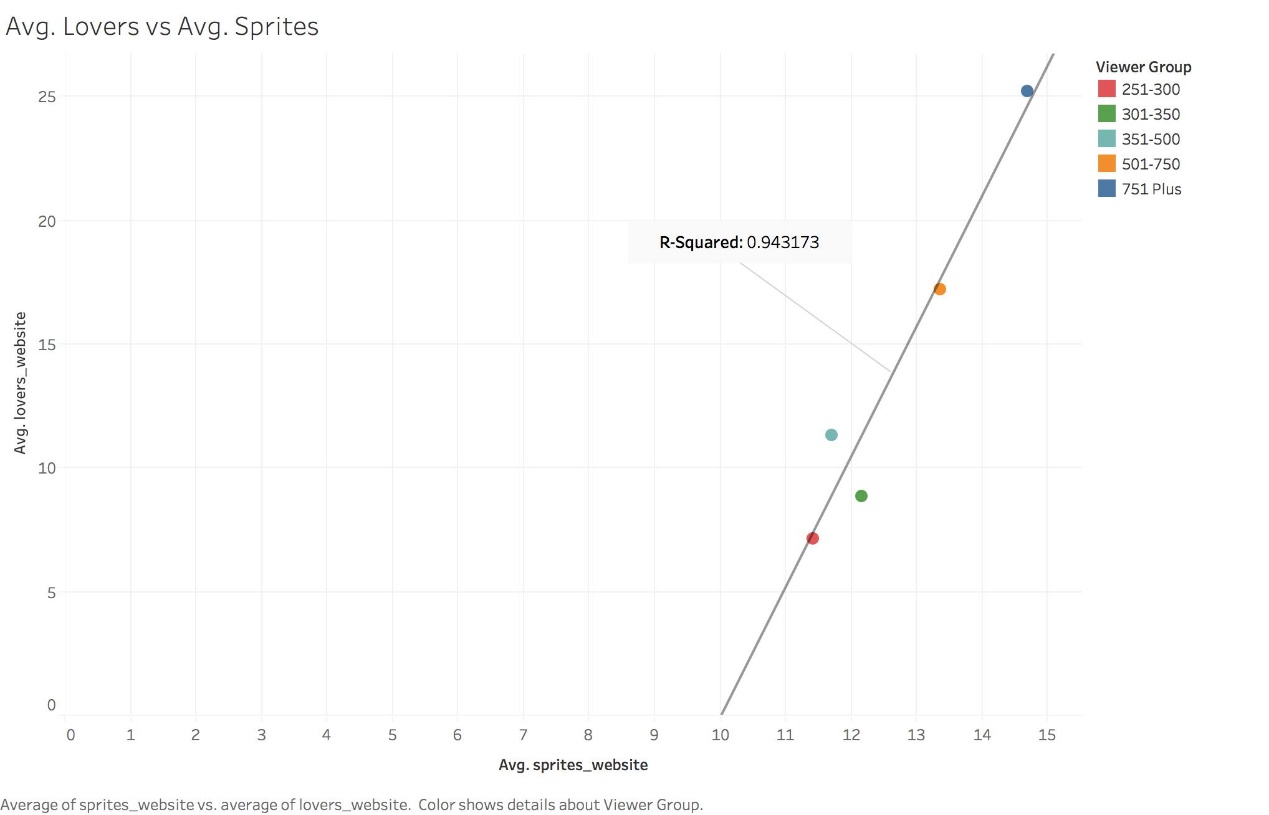


Figure 7. Pattern between average lovers and average sprites

From the graphs, we see that the number of projects being loved increases as the number ****of sprites and scripts increased. Moreover, we get that the value of R squared between viewers and scripts is 0.107913 which indicates a weaker positive correlation between those two variables while the value of R squared between viewers and sprites is 0.943137 that states a very strong positive correlation between viewers and sprites so that we can say both scripts and sprites will affect people on loving a project. In addition, the sprites are the main factor of influence compared with scripts.

**3) Analysis of new Users and Projects**

**Data source used & variables used for analysis**

We merged all five groups of datasets into one to determine the number of downloads, users, and projects.

**Methods used in this part**

We have merged our five dataset tables by appending rows from one table to another. It is possible as all the five data sets have the exact same variables and structures.

**Findings**

We can see the data of new projects and users from 2007 to 2012. There is a gradual increase in the number of projects and users from 2007 till 2011 but a sudden fall in 2012 as we have data only until Q1 of 2012.

****A screenshot of a cell phone

Description generated with very high confidence

Figure 8. New project V.S. users

The number of unique projects and users in 2007 was 1824 and 745 respectively. This number gradually increases till 2009 after which there is a significant hike in 2010 with 5911 new projects and 1679 new users. It increases in 2011 but falls to a lowest of 404 new users and 951 projects. As we have data only untill Q1 of 2012, it can be inferred that the number of new projects and users would be huge.

The number of new projects and users in 2007 is low as the Scratch community was new and not so popular. But we can see from the graph how it started becoming more popular over the years.

**Insight of Projects and Downloads**

A screenshot of a cell phone

Description generated with high confidence

figure 9. downloaded V.S. project remixed

**Findings**

The graph contains 6 horizontal stacked bars from 2007 to 2012 showing the number of downloads and whether the projects was a remix or an original.

A project can be classified as original or unoriginal depending if it’s an original project or an altered version. The primary goal of the Scratch project was to make it easy for users to download and modify other user’s projects. Hence, we can see there exists a large number of projects which were remixed.

**4) Analysis on behavior of most viewed projects**

**Data source used & variables used for analysis**

We merged all five groups of datasets into one to determine the specific behavior of most viewed projects.

**Methods used in this part**

We have merged our five dataset tables by appending rows from one table to another. It is possible as all the five data sets have the exact same variables and structures.

**Findings**

The dashboard consists of Images, Sounds and Saves plotted with 250 to 1000 views. We can see a few zig z-g graphs in the dashboard but after carefully examining it we can see for a project to be popular it is not necessary to have a large number of images, sounds or saves.

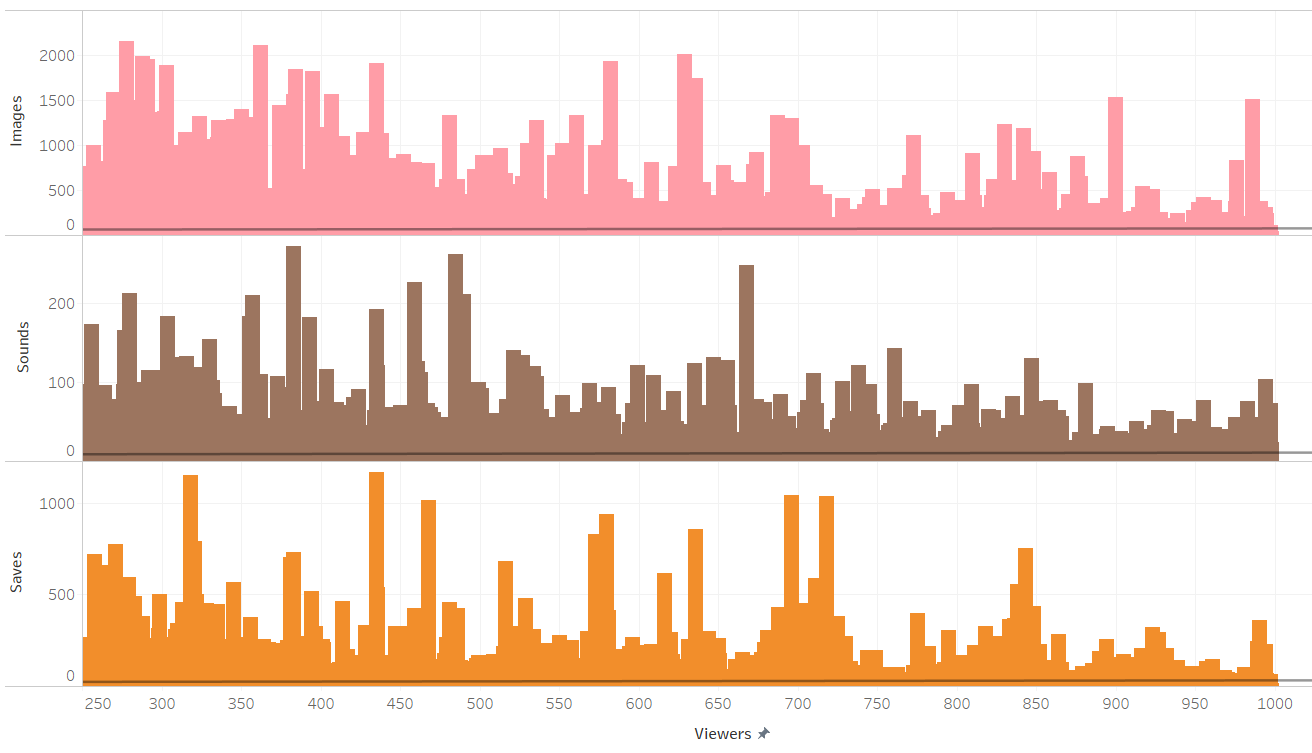
****

Figure 10. viewers V.S. images, sounds, and saves

From the graph consisting of Images plotted against views, we can see projects having views more than 900 views have very few images mostly around 500, while projects with views between 250 to 350 have images around 1500 on average. Hence a large number image does not necessarily make a project popular. It is possible that simple projects with small file size gain more popularity than complex projects with large files sizes.

We can see similar trends with sounds plotted against views that projects with a large number of views have less number of sound files. For saves plotted against views, we can see a similar trend but not as significant as images and sounds.

We can conclude that projects that are simple and small in media size usually gain ****more popularity and viewers.

**5) Analysis of relation between Lovers, Scripts, Sprites, and Downloads**

**Findings:**

The Dashboard consists of Sprites, Scripts and Lovers plotted with 250 to 1000 views. The shades of blue indicate the number of downloads of the project i.e. projects with fewer downloads would have a lighter shade of blue and projects with a high number of downloads would have darker shade of blue.

A view of a city

Description generated with very high confidence

Figure 11. views V.S. sprites

From Sprites plotted against Views, we can see the number of sprites does not make ****any significant impact on the number of views except for projects with more than 850 views which have less number of sprites on average.

From Scripts plotted against Views, we can see the number of scripts does affect the number of views.

We can see a positive correlation between Lovers and Views. Hence if a project is viewed more often, it gains more loves. Also, we can see darker shades of blue towards 800 to 1000 views indicating more number of downloads. Hence, we can conclude if a project is viewed more often it will be more likely to be loved and downloaded.

**6) Analysis on Similarities and Differences between Viewer Group**

**Data source used & variables used on analysis**

We merged all five viewer groups of data into one full set to examine the overall impact of website viewers on the average number of downloaders and lovers.

**Methods used in this part**

We used tableau’s Union function to combine our five dataset tables by appending rows from one table to another. This can only be done since all five data sets have exact same variables and structures. Next, we created 5 groups based on the “viewers\_website” with calculated field using an IF ELSE function, these five groups are shown in the graph below.

**Findings**

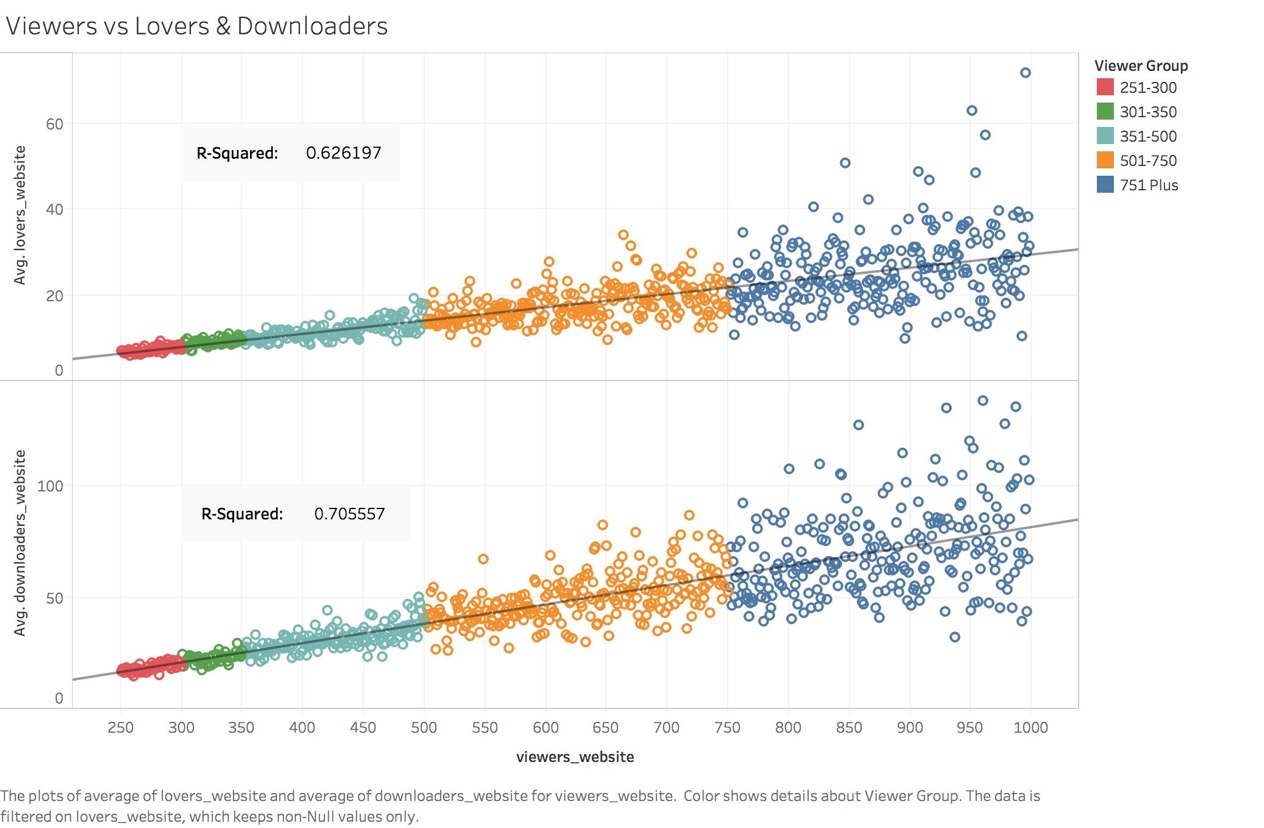


Figure 12 comparison with other groups

As we can see from the graph above, for both of these graphs, the average number of lovers and downloaders increase as the number viewer increases for each viewer group. However, there are some differences between viewer groups. As we can see from the graph above, the data in low-viewer groups such as a group in red, are more clustered. The data in high-viewer groups such as a group in blue, are more scattered. Therefore, this shows that the variation of the average number of downloaders and lovers in low-viewer groups is much smaller than that of in high-viewer groups. Thus, low-viewer groups tend to have more consistent results on the number of downloaders and lovers than that of the high-viewer group. We built a linear regression model for both data. The R-squared for both ****groups indicates that approximately 62.6% and 70.1% of the variability of the response variable around its mean is explained by the model respectively. These two R-squared are not as high probably due to the high variation within the high-viewers group, but still, is shows us viewer-group is still a good indicator which predicts the number of lovers and downloaders.

**Conclusion**

Some of the important conclusion of our analysis are:

* We can see there is a low correlation among viewers, lovers and downloaders which means the variation trend of datasetcould not influence others observably.
* There exists a positive correlation between lovers and viewers but the tendency is smooth as we can see mathematically R^2 in a scatter plot was below 0.005.
* The correlation between viewers and sprites is an exception with R squared around 0.9 which means both the variables are highly correlated.
* The popularity of scratch community substantially increased over the years from 2007 to 2012 with more number of projects and users created every year.
* The projects downloaded over the years 2007 to 2012 are mostly remixed reflecting the goal of scratch, to make it easy for users to download and modify other user’s project.
* Images, sounds and saves have a low correlation with views so it does not ****necessarily mean that projects with high number of media files would have a large number of views.
* Projects with highest views had relatively small number of media files indicating projecs with small file size would have more number of views.
* When compared with other groups database, we find there is actually a similar tendency, that group with the low-viewers database has a more density data correlation, where the high-viewers has a more discrete distribution.

**Reference**

1) Hill, B. M., & Monroy-Hernández, A. (2017, January 31). A longitudinal dataset of five years of public activity in the Scratch online community. Retrieved from <https://www.nature.com/articles/sdata20172#t1>

2) Scratch (programming language). (2018, February 17). Retrieved from <https://en.wikipedia.org/wiki/Scratch_(programming_language)>

3) Knaflic, Cole N. (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals. Hoboken, NJ. John Wiley & Sons. ISBN: 978-1-119-00225-3

4) Rumsey, Deborah. (n.d.). How to Interpret a Correlation Coefficient r. Retrieved from <http://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/>

5) Join Your Data. (n.d.). Retrieved from ****https://onlinehelp.tableau.com/current/pro/desktop/en-us/joining\_tables.html