## Project Outline:

### An Exploration of Gender Differences in Sharing Behavior on Scratch

#### Emilia Gan -- COM 521, Winter 2017

### Rationale

* Women are underrepresented in computer science and other computational fields.
* Previous research has suggested that boys and girls interact with the Scratch programming language and the online Scratch community in different ways.
* Scratch is the first experience with programming for many, if not most, children. The attitudes towards programming formed by children using the Scratch online programming platform thus likely affect future decisions such as whether to study computational fields or to pursue careers in tech.
* Our study will examine the differences in project sharing behavior between girls and boys using the Scratch platform. In particular, we will investigate whether there are differences in the sharing patterns of male versus female users and look into whether differences exist in the peer feedback received by projects based on the gender of the project creator.

### Initial Suppositions

Before beginning this project, I had a number of theories about how females might view sharing their projects relative to their male peers. Specifically, I thought: \* Females would be less likely to share projects \* Females would be more sensitive to negative feeback or absence of positive feedback

### Objectives

#### General Objective

To examine how peer recognition of shared projects affects future project sharing behavior by the same Scratcher.

#### Specific Objectives

* To determine whether there are differences in the pattern of Scratch project creation and sharing between male and female users of the Scratch online programming platform.
* To determine if there is a difference in the frequency of receiving peer feedback between projects created by female users and projects created by male users.
* To determine if receiving peer feedback (e.g. love-its and comments) affects the future sharing behavior of the project creator.

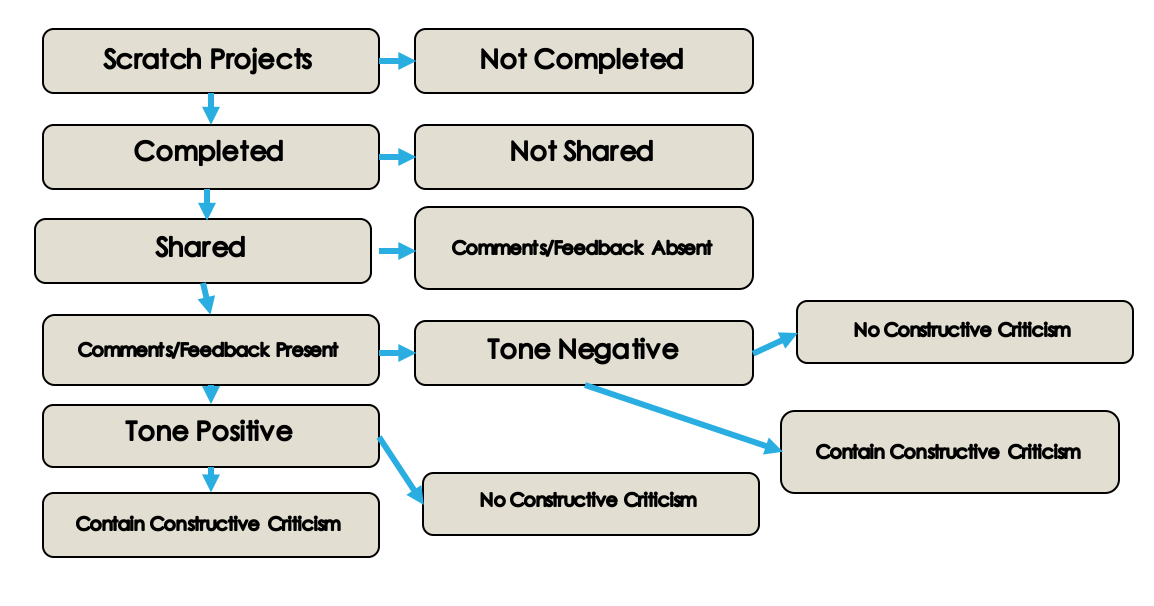
### Null Hypotheses

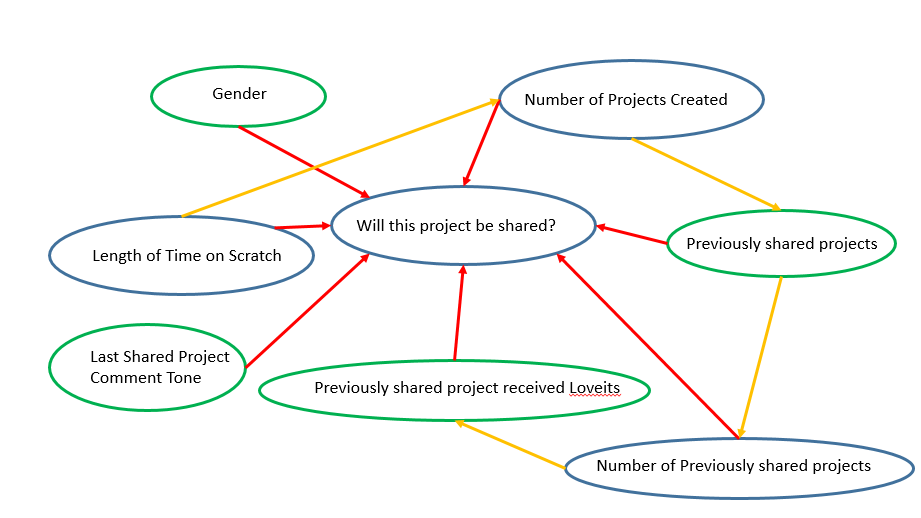
* Projects created by females are just as likely to be shared as projects created by males.
* Projects created by females are just as likely to receive peer recognition as projects created by males.
* Scratchers whose shared projects receive no peer recognition are just as likely to share future projects as are Scratchers whose earlier shared projects did receive peer recognition.
* Female Scratchers whose shared projects receive no peer recognition are just as likely to share future projects as are male Scratchers whose shared projects receive no peer recognition.

### Conceptual Diagrams

#### Conceptual Diagram 1: Project Outcomes

The first conceptual diagram depicts the possible outcomes for a newly created project.

 The second conceptual diagram depicts the factors that may contribute to the decision to share or not share a project.



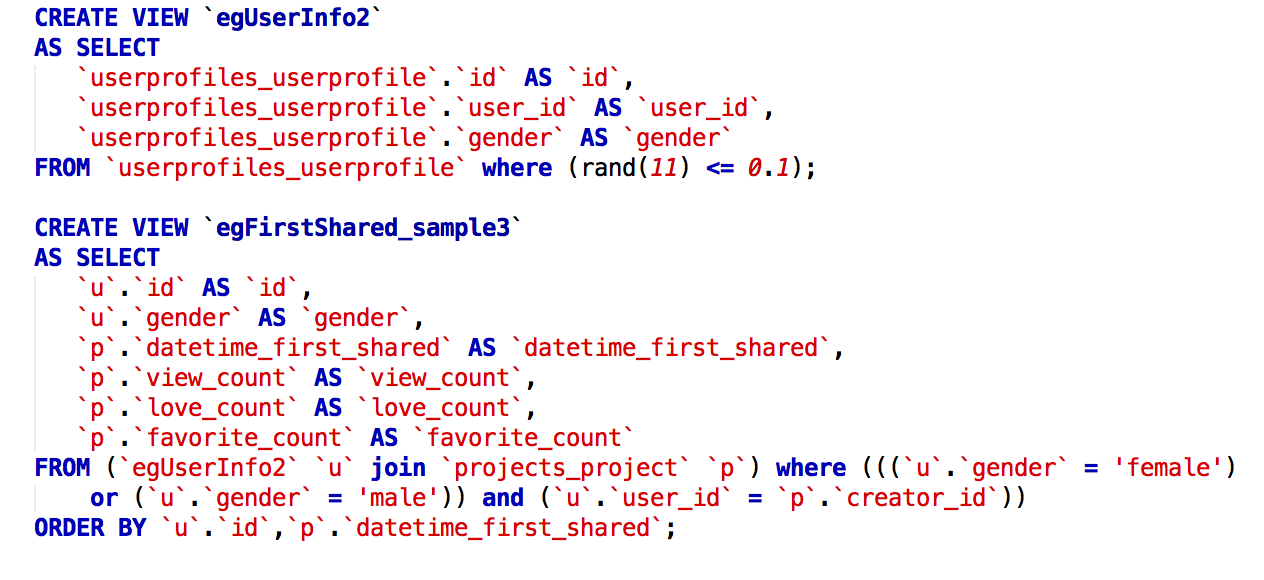
Conceptual Diagram for Factors Potentially Influencing Sharing

## Dummy Tables

### Taking a Sample from the Scratch Data Set

### Looking for Gender Differences in Project Sharing

The SQL code below was used to sample the Scratch Database. Setting the random seed to 11 and selecting for values < 0.1, ~871K users were randomly selected for the data set. These users created over 4.86 million projects.



SQL code used to generate the sample data set.

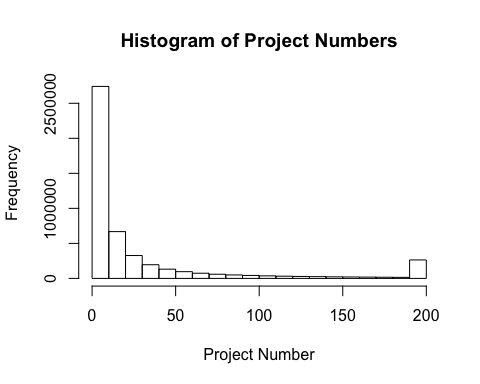
Before starting any analysis, an examination of the data gives the following information:

|  |  |  |
| --- | --- | --- |
|  | Projects | Percent |
| Female | 1855552 | 38.19 |
| Male | 3003650 | 61.81 |
| Total | 4859202 | 100.00 |
| Table 1: | Project Cou | nt by Gender of Creator |

This data set contains information on 4,859,202 Scratch projects. Approximately 40% of the projects were created by female Scratchers. Almost 600,000 unique Scratchers are represented in this dataset (236,606 female (39.6 %) and 360,428 male(60.4 %)).

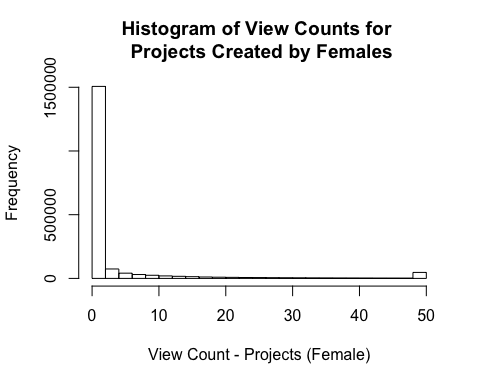
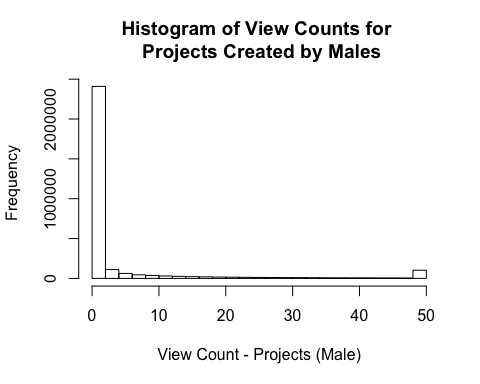
The following visualizations were created to help with the initial data exploration process.

Looking first at numbers of projects created (Project numbers 200 and higher are grouped together):

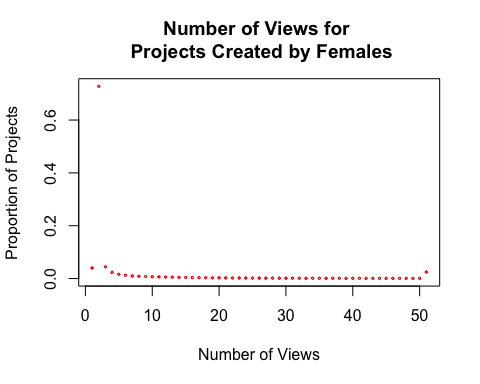
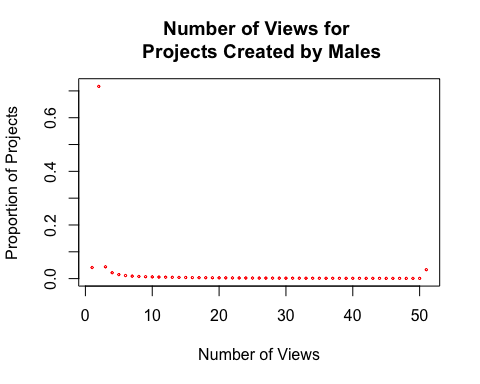


As this histogram shows, most people create a small number of projects, thus the bin containing 1st through 10th projects contains many more projects than any other bin. However, significant numbers of Scratchers create many more projects, with the most prolific Scratcher in this dataset creating close to 4000 projects.

Looking next at histograms of view counts for both the female and the male subsets in the data:

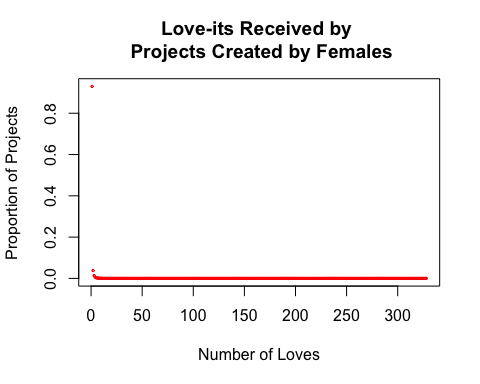
 

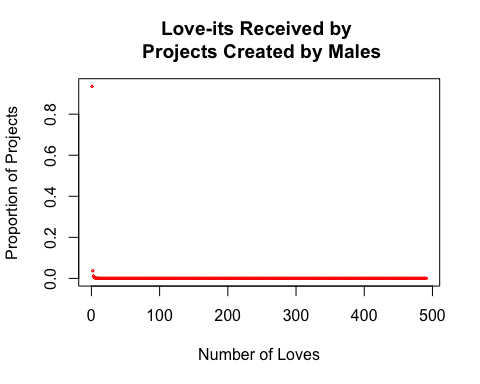
What this shows us is that most projects get very few views. The following plots present the same data, looking at proportions rather than raw counts:

  Proportion of Project Receiving a Given Number of Views

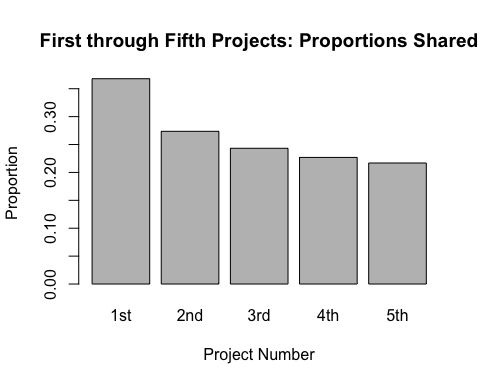
After looking at some features of the entire sample, the data set was processed to add an additional feature, "prev.projects." This feature makes data on how many projects had been created prior to any given project easy to access, allowing the creation of subsets such as "first.project", "second.project", etc.

While number of views is one measure of the interest generated by a project, it is not the only one available to us. A more active measure of positive interest in a given project is the number of love-its the project receives from other Scratchers. The following plots look at the distribution of love-its awarded to projects created by females and males.



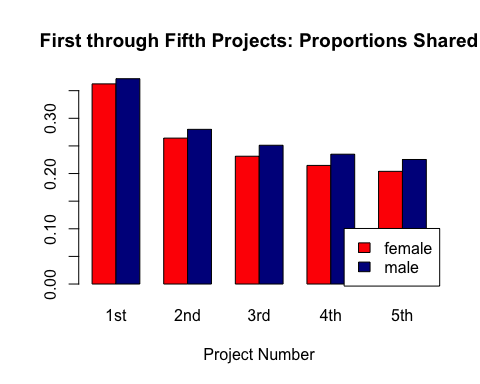
 Not surprisingly, most projects receive a small number of love-its. What is notable from these two plots, however, is that the maximum number of love-its received by a single project created by a male is much higher than the maximum received by a project created by female. In addition, the plot shown above is a scatter plot. Thus, it is evident that it is not just a simgle male-created project that got a larger number of love-its than the highest love-it-earning female-created project. In spite of these two observations, the Scratcher with the maximum cumulative total number of love-its received is female (30,119 versus 23,188). For both females and males, the Scratcher with the maximum number of projects is not the same individual as the Scratcher with the largest cumulative love-its total.

We can now look at sharing based on project number and run analyses to look at the effect that reactions to earlier projects has on sharing of later projects.

The following bar plot shows the proportions shared for each of 1st through 5th projects. 

### Looking for Gender Differences in Project Sharing

There appears to be a decreasing trend in proportion of projects shared as the number of projects created increases. Next, we'll look at the data when grouped according to the gender of the project creator. The following bar plot displays the proportions shared by females and males, for easy comparison:

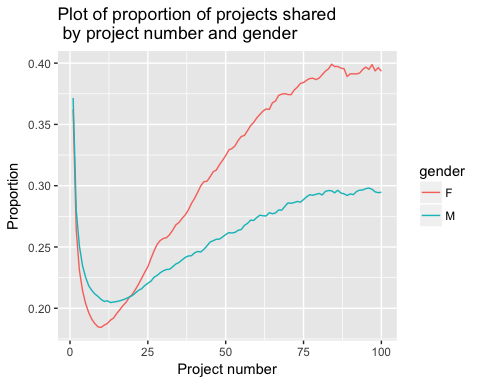


### What Are the Sharing Patterns for Projects?

Performing a test of equal or given proportions (prop.test) will test the null hypothesis that the proportions in each group are the same.

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: shared\_projects out of all\_projects  
## X-squared = 220.59, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.005275521 0.006883682  
## sample estimates:  
## prop 1 prop 2   
## 0.2632510 0.2571714

Test of Proportions: Proportion of Shared Projects out of All Projects Created, by Gender

Loking at the first five projects showed a decreasing trend with females consistently sharing a smaller proportion of their projects than men did. However, the proportions test over the whole dataset was not consistent with this pattern, meaning that at some point, females start sharing more projects relative to males. To get a better sense of what happens over time, we need to look at a longer trend.  As this plot demonstrates, although females start out less likely to share projects than males, if they go on to create more projects, they begin sharingi them at greater rates, and eventually overtake males for the proportion of projects shared. To see more precisely where the switchover occurs, looking at the data in tabular form is helpful:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| Female | 0.196 | 0.199 | 0.202 | 0.205 | 0.209 | 0.211 | 0.215 | 0.220 | 0.224 | 0.229 | 0.234 |
| Male | 0.205 | 0.206 | 0.207 | 0.208 | 0.209 | 0.210 | 0.212 | 0.215 | 0.216 | 0.219 | 0.220 |
| Looking a | t the ac | tual pro | portion | values, | we see t | he cross | over occ | urs at 2 | 0 projec | ts creat | ed. |

Looking at an even longer trend covering data on 1000 projects created, we see that females, once they overtake males, retain this lead for a considerable period. Males don't catch up again until roughly 400 projects have been created and, at that point, the difference seems likely to have minimal, if any, practical significance (though this has not formally verified):



### Effect of First Project Receiving Loveits on Second Project Sharing

When a new Scratcher creates his or her first project, a decision must be made whether ot not to share the project. Once a project is shared, other Scratchers can view the project and react to it. Possible feedback includes the number of views the project attracts, the number of Loveits the project receives, the number of times the project is identified as a "Favorite", and whether or not the project receives comments. As a first filter for whether a shared project received peer approval, we will look at how Loveits correlate with future project sharing.

First project shared and given Loveit(s):

Looking to see whether there is a difference in the proportion of first projects receiving Loveits based on project creator's gender:

|  |  |  |
| --- | --- | --- |
|  | Female | Male |
| Got LoveIts | 12265 | 21295 |
| Got NO LoveIts | 73453 | 112652 |

Table 4: LoveIts and First Shared Project

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: first\_loves out of firsts  
## X-squared = 101.91, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.01895930 -0.01283135  
## sample estimates:  
## prop 1 prop 2   
## 0.1430855 0.1589808

Test of Proportions: Proportion of First Projects Getting LoveIts by Gender

### Effect on Second Project

Will now use the data on first project sharing to see how receiving Loveits affects whether or not the second project is shared. First, we'll look at proportions of Scratchers sharing a second project after receiving Loveits for their first project, then at those sharing a second project after the first did NOT receive Loveits:

|  |  |  |
| --- | --- | --- |
|  | Female | Male |
| 1st Got LoveIts | 7868 | 14058 |
| 1st Got NO LoveIts | 37264 | 58185 |
| Total 2nd Shared Projects | 45132 | 72243 |
| Total 2nd Projects | 170866 | 258014 |

Table 5: Sharing of Second Projects

prop.test(second\_shared\_first\_loved, first\_proj\_got\_loves)

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: second\_shared\_first\_loved out of first\_proj\_got\_loves  
## X-squared = 11.876, df = 1, p-value = 0.0005687  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.029325671 -0.007983854  
## sample estimates:  
## prop 1 prop 2   
## 0.6415002 0.6601550

prop.test(second\_shared\_first\_not\_loved, first\_proj\_got\_no\_loves)

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: second\_shared\_first\_not\_loved out of first\_proj\_got\_no\_loves  
## X-squared = 14.976, df = 1, p-value = 0.0001089  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.013842026 -0.004527061  
## sample estimates:  
## prop 1 prop 2   
## 0.5073176 0.5165021

Now, we'll look at second project sharing by Scratchers overall:

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: second\_projects\_shared out of second\_projects\_all  
## X-squared = 129.99, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.01857951 -0.01313974  
## sample estimates:  
## prop 1 prop 2   
## 0.2641368 0.2799964

Now we will look at the proportion of shared second projects by users who got Loveits for their first project -- i.e. out of all second projects that are shared, how many are shared by people who received Loveits on their first projects?

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: first\_loved out of second\_projects\_shared  
## X-squared = 63.432, df = 1, p-value = 1.66e-15  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.02595798 -0.01573734  
## sample estimates:  
## prop 1 prop 2   
## 0.2422228 0.2630705

##   
## 2-sample test for equality of proportions with continuity  
## correction  
##   
## data: second\_projects\_shared out of not.first\_loved  
## X-squared = 185.75, df = 1, p-value < 2.2e-16  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.02294754 -0.01719082  
## sample estimates:  
## prop 1 prop 2   
## 0.2821914 0.3022606

## HYPOTHESIS TESTING

### HYPOTHESIS 1

* Projects created by females are just as likely to be shared as projects created by males.
* logit(p(share-gender))

##   
## Call:  
## glm(formula = shared ~ female + log1p(prev.projects) + female \*   
## log1p(prev.projects), family = binomial("logit"), data = info\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.8691 -0.7844 -0.7637 1.5763 1.7534   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.9572917 0.0023239 -411.94 <2e-16 \*\*\*  
## femaleTRUE -0.1799691 0.0037296 -48.26 <2e-16 \*\*\*  
## log1p(prev.projects) -0.0444043 0.0008308 -53.45 <2e-16 \*\*\*  
## femaleTRUE:log1p(prev.projects) 0.0911158 0.0013166 69.21 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5564039 on 4859201 degrees of freedom  
## Residual deviance: 5558860 on 4859198 degrees of freedom  
## AIC: 5558868  
##   
## Number of Fisher Scoring iterations: 4

### HYPOTHESIS 2

* Projects created by females are just as likely to receive peer recognition as projects created by males.
* logit(p(love>0-gender))

##   
## Call:  
## glm(formula = got.love ~ gender + log1p(prev.projects), family = binomial("logit"),   
## data = info\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.3309 -0.6872 -0.3193 0.5886 2.4502   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.764422 0.003601 -1045.43 <2e-16 \*\*\*  
## gendermale 0.189948 0.002541 74.75 <2e-16 \*\*\*  
## log1p(prev.projects) 1.174018 0.001100 1067.66 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 6284980 on 4859201 degrees of freedom  
## Residual deviance: 4158858 on 4859199 degrees of freedom  
## AIC: 4158864  
##   
## Number of Fisher Scoring iterations: 5

### HYPOTHESIS 3

* Scratchers whose shared projects receive no peer recognition are just as likely to share future projects as are Scratchers whose earlier shared projects did receive peer recognition.
* logit(p(share-recognition)

# Get SUM of previous loveits for projects

##   
## Call:  
## glm(formula = shared ~ log1p(prev.projects) + got.love, family = binomial("logit"),   
## data = info\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5043 -0.7851 -0.5783 0.9410 2.5727   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.9521721 0.0018818 -506.0 <2e-16 \*\*\*  
## log1p(prev.projects) -0.3862634 0.0009035 -427.5 <2e-16 \*\*\*  
## got.love 1.9618491 0.0029826 657.8 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5564039 on 4859201 degrees of freedom  
## Residual deviance: 5082637 on 4859199 degrees of freedom  
## AIC: 5082643  
##   
## Number of Fisher Scoring iterations: 4

### HYPOTHESIS 4

* Female Scratchers whose shared projects receive no peer recognition are just as likely to share future projects as are male Scratchers whose shared projects receive no peer recognition.
* logit(p(share-recognition-gender)

##   
## Call:  
## glm(formula = shared ~ female + log1p(prev.projects) + got.love,   
## family = binomial("logit"), data = info\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5340 -0.7850 -0.5833 0.9538 2.5516   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.9916971 0.0020876 -475.04 <2e-16 \*\*\*  
## femaleTRUE 0.0995143 0.0022470 44.29 <2e-16 \*\*\*  
## log1p(prev.projects) -0.3868950 0.0009039 -428.03 <2e-16 \*\*\*  
## got.love 1.9684075 0.0029890 658.54 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5564039 on 4859201 degrees of freedom  
## Residual deviance: 5080681 on 4859198 degrees of freedom  
## AIC: 5080689  
##   
## Number of Fisher Scoring iterations: 4

##   
## Call:  
## glm(formula = shared ~ female + log1p(prev.projects) + got.love +   
## female \* got.love, family = binomial("logit"), data = info\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5771 -0.7822 -0.5793 0.9719 2.5796   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.9440523 0.0022351 -422.37 < 2e-16 \*\*\*  
## femaleTRUE -0.0134086 0.0030264 -4.43 9.4e-06 \*\*\*  
## log1p(prev.projects) -0.3884367 0.0009049 -429.24 < 2e-16 \*\*\*  
## got.love 1.8754945 0.0034067 550.54 < 2e-16 \*\*\*  
## femaleTRUE:got.love 0.2545960 0.0045351 56.14 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 5564039 on 4859201 degrees of freedom  
## Residual deviance: 5077523 on 4859197 degrees of freedom  
## AIC: 5077533  
##   
## Number of Fisher Scoring iterations: 4