The Dataset

We elected to use the Mobile App Store Dataset (https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps). The data was extracted using the Itunes Search API. The data was originally generated to explore details that allow an app to make it to trending in the app store. This dataset was selected because we thought it would be easier to work with what we thought were easily identifiable relationships.

Below is a list and description of fields used in our analysis:

1. user_rating - the average user rating value the app receives across all versions (0-5)
2. user_rating_ver - the average user rating value for the current version of the app (0-5)
3. prime_genre - the primary genre of the application (ex. Education, Entertainment, Social Networking)
4. Lang.num - the number of languages supported by the app
5. Price - the cost of the app

Analysis 1:

Average App Ratings and App Genres

Hypothesis

We hypothesized that there could be a relationship between the genre of an app and the average rating of the app. If this was to be true, then different apps would have different expected average ratings based on their genre.

Finding

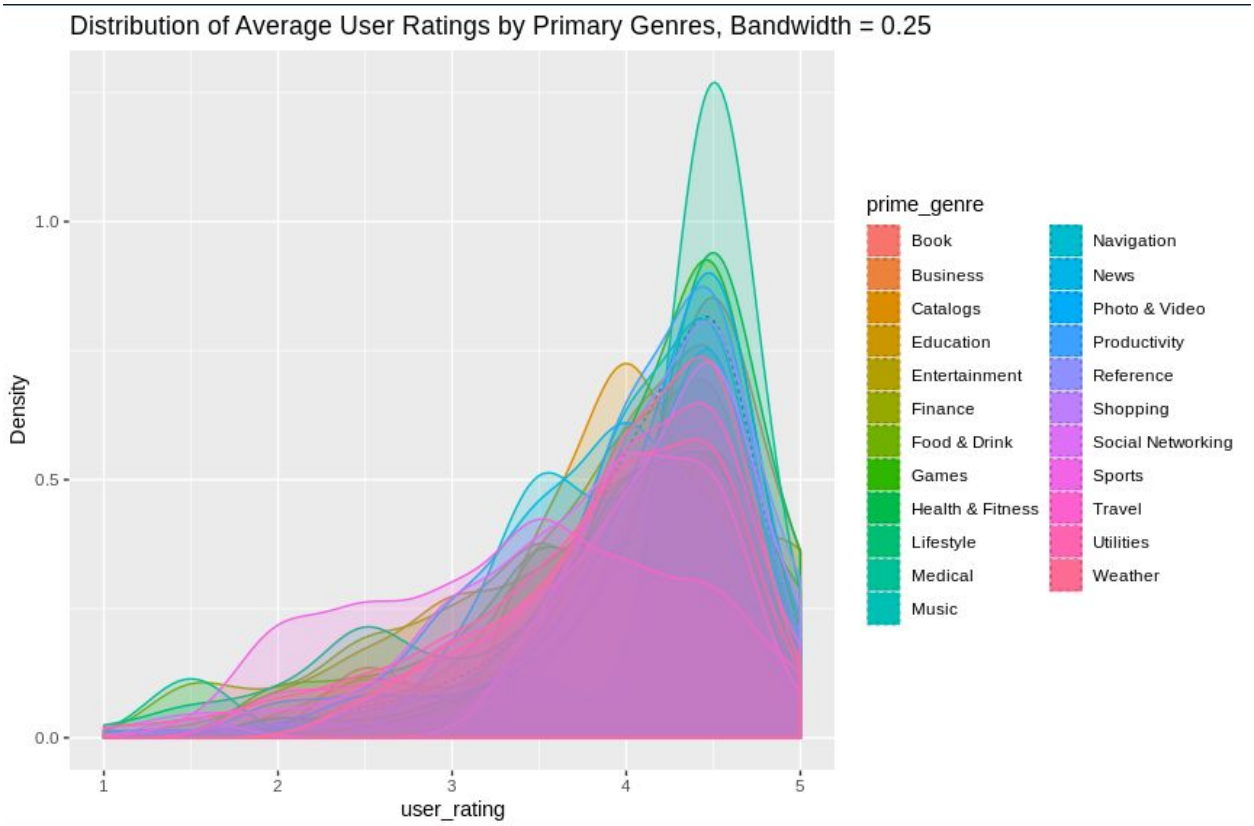
There is statistically significant difference between the expected average rating of apps with different genres. It might be of interest for quality assurance teams to compare the ratings of their apps to other apps of the same genre, rather than those of the overall ecosystem.

Investigation

Since there are only about 7000 apps in the dataset, versus the millions of apps offered in the Appstore, whatever differences between mean average ratings of different genres could result from innate randomness from the sample, rather than from any actual differences between the 24 genres. Therefore, even though there is some difference between the mean average ratings

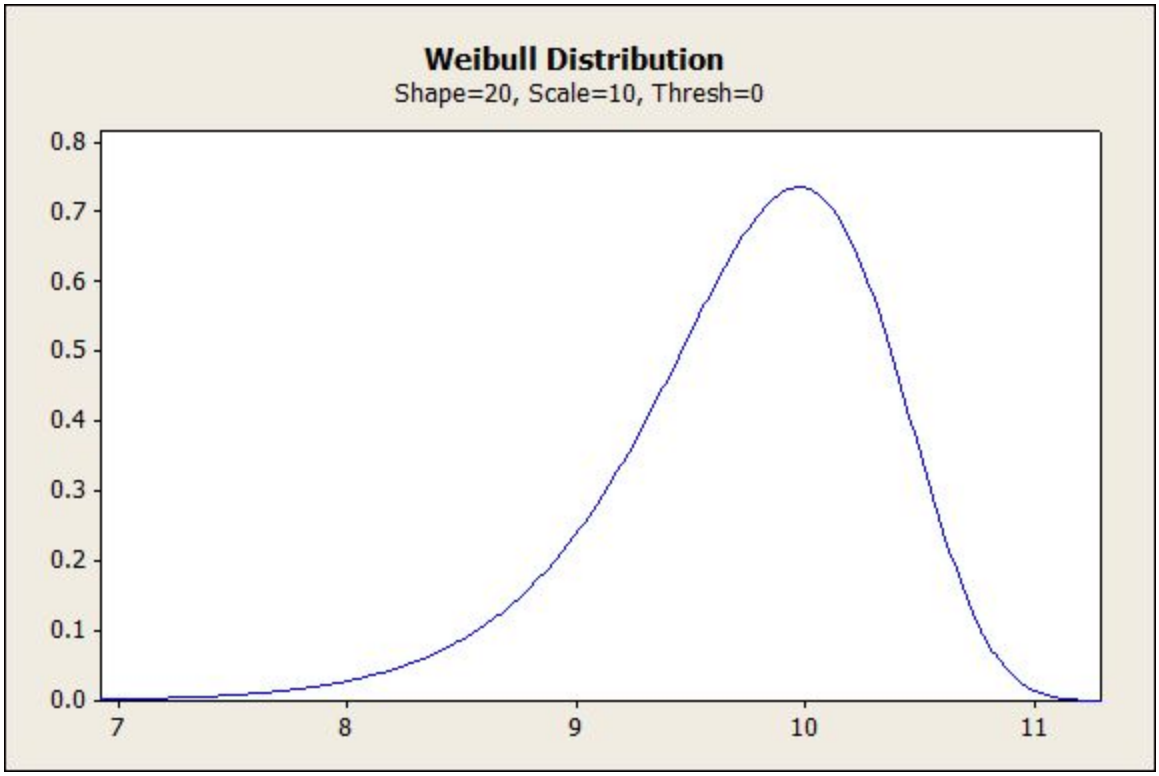
in the dataset (see Appendix A), it requires a statistical test for us to be confident about the differences despite the randomness in the sampling procedure.

Because the average rating of an app is a linear function of its cumulative rating, the distributions of average ratings should be subject to Central Limit Theorem and thus conform to a bell curve pattern. Looking at the summary statistics, we can also find that there is little difference between variances of each group, with the largest variance not exceeding two times the smallest variance. With a rule of thumb (https://data.library.virginia.edu/a-rule-of-thumb-for-unequal-variances/), the data suggests the use of one factor Analysis of Variance F test to investigate the differences between groups, as long as the groups are normally distributed.



Using the N > 30 rule for Central Limit Theorem, we filtered out the apps with fewer than 30 ratings, and investigated the distributions of the remainder. As we can see from the density plotting above, the average ratings tend to center around 4.3 stars, with visible negative skew compared to a bell shaped normal distribution. As the visualization of distributions seems to contradict the Central Limit Theorem, we carried out shapiro-wilk normality tests on each group to test our assumption. The results (p values included with histograms in Appendix B) allow us to be highly confident that the genres do not possess normally distributed average app ratings. Nevertheless, as seen in the density plot above, the average ratings are distributed in an identical, albeit not gaussian, manner. We make the reasonable assumption that they belong to

the same Weibull distribution family, an example of which is included below.



With this assumption in mind, we carried out an Kruskal-Wallis test against the null hypothesis that all the genres have the same expected average app ratings. We obtained a statistically significant result against the null hypothesis (chi-squared = 450.57, df = 22, p-value < 2.2e-16).

We were thus able to conclude that there was significant difference between expected average app ratings in apps of various genres.

Limitations and Future Work

Due to constraints on time and resources, we omitted some proposed procedures that could further reinforce our assumptions, such as carrying out a chi squared tests for algorithmically fitted weibull distributions, on top of visually analyzing kernel density estimations with an arbitrarily set bandwidth. We could also attempt to identify confounding variables that might cause differences between genres, such as complexity of UIs and number of features, which can give product owners a competitive edge against other apps of the same genre.

Analysis 2:

of languages vs. Rating/# Of Ratings

Hypothesis

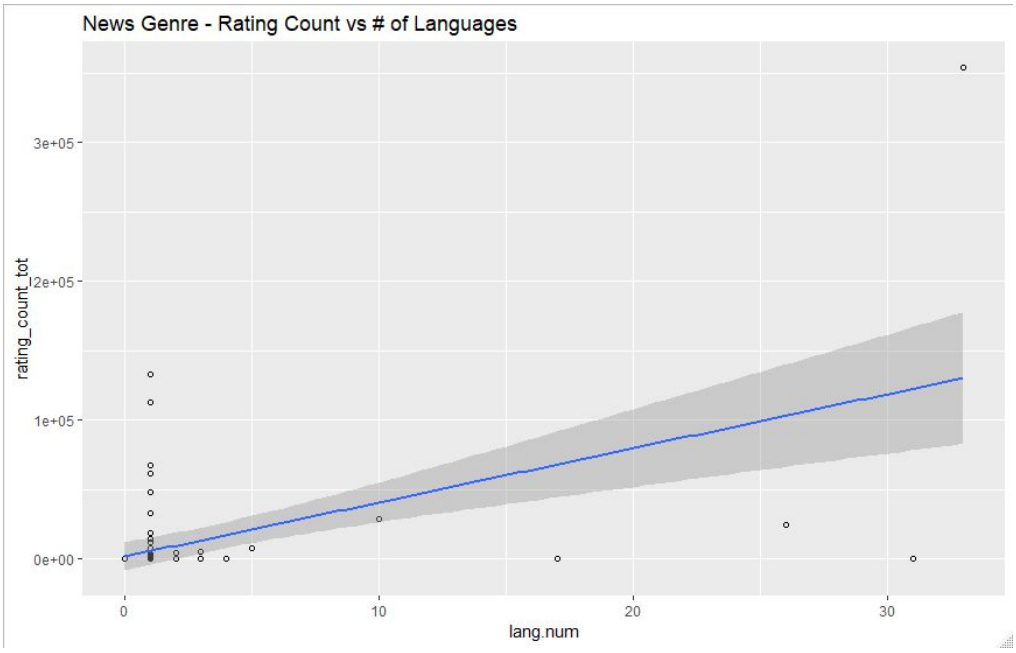
We hypothesized that there would be a correlation between the number of languages supported by an app and the number of downloads an app received. For this to be true the correlation coefficient between the two variables would need to be above .6.

Investigation

Due to a lack of explicit download data in the dataset, we chose to use rating and the number of ratings as a proxy for download numbers. The expected correlation between the number of languages and was not found. This may be due to the inexact download calculations made from the rating data. User that are request more language support would increase the number of reviews on low language apps. Also, users that are happy with the language support may not leave a review because they are only content with the with the app, not ecstatic about it. This is backed by the data showing that the most common ratings are 1 star and 5 star, which shows ratings are more common when the user has a strong opinion. Also, the rating data is rounded to the nearest .5. That means that the rating data is segmented which could be making it more difficult to show any correlation.

Due to these flaws, the overall analysis failed to show much. The main takeaway was the fact that a simple correlation between the number of languages and the ratings was minimal, which shows that more analysis would need to be done. It also implies that adding a random language to an app will not guarantee any increase in rating value or count.

However, once we grouped by genre a few results started to appear. Genres with high usage of text tended to have more/higher ratings. Notable exceptions to this general trend are Catalogs and News, however, this could be due to the local nature of those apps. For example, a French person would not be likely to download a Cleveland Plain dealer app. Another interesting thing to note is the News genre being highly correlated to rating count, but completely uncorrelated to average rating. This could be caused by a high variance in quality of news apps. Also, plotting the news genre separately shows that there is one outlier with a higher number of languages supported and avery high rating which creates the high correlation. This is another drawback of our relatively small dataset.



Future Work

Our conclusion would be more robust if we had the actual download data, instead of the implied data. Also, we could hold more variables constant and use a difference-in-differences evaluation to remove some of the inconsistencies in rating change found in the genre vs. rating analysis. Another interesting analysis would be identifying the top languages to support. The obvious answer is the countries in which the app store is the most popular. However, that initial thought does not consider the proportion of the population that are multilingual. An analysis on this front would provide information on what language to focus on when developing an app.

Appendix A:

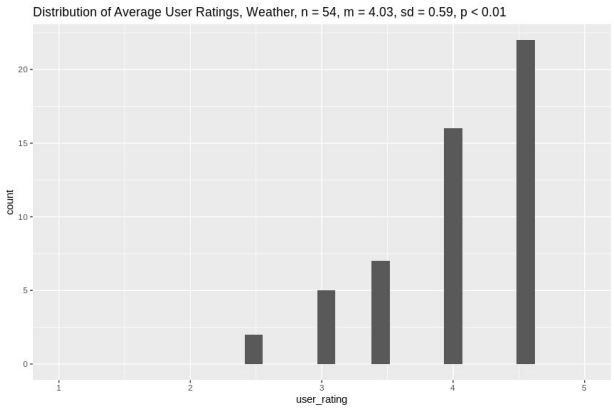
Included in this appendix is the summary statistics of average ratings by different genres. Decimals are rounded to four significant digits. Apps with fewer than 30 ratings are excluded.

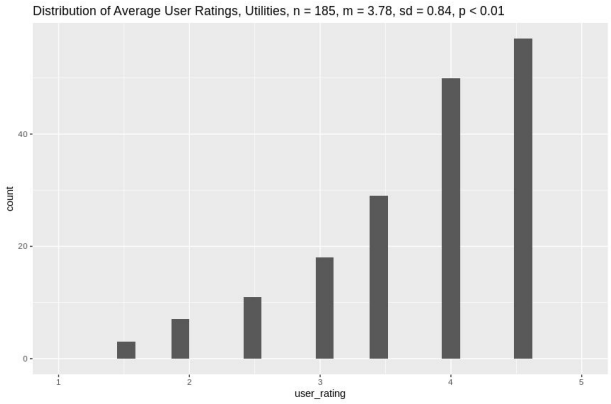
app genre	# of samples	mean average rating	standard deviation
Book	37	4.230	0.6831
Business	48	4.052	0.6782
Catalogs	5	4.200	0.5701
Education	288	3.998	0.6027

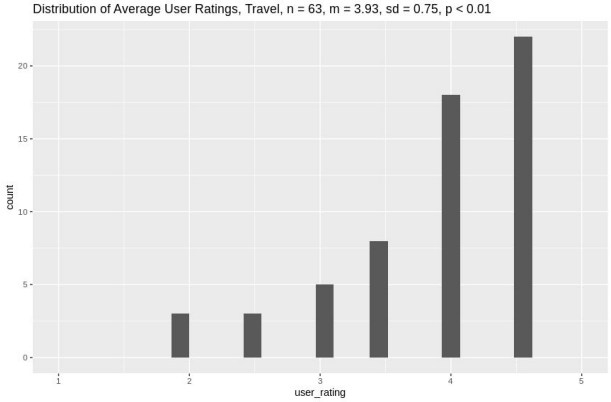
Entertainment	408	3.701	0.8123
Finance	50	3.580	0.9333
Food & Drink	36	3.917	0.8410
Games	2963	4.229	0.4858
Health & Fitness	139	4.288	0.6167
Lifestyle	92	3.712	0.9499
Medical	14	4.250	0.8492
Music	118	4.114	0.5524
Navigation	25	4.000	0.5204
News	53	3.755	0.6695
Photo & Video	287	4.132	0.7021
Productivity	152	4.178	0.5502
Reference	45	4.244	0.5995
Shopping	93	4.102	0.6366
Social Networking	122	3.725	0.7876
Sports	85	3.424	0.8981
Travel	63	3.929	0.7506
Utilities	185	3.784	0.8352
Weather	54	4.028	0.5941

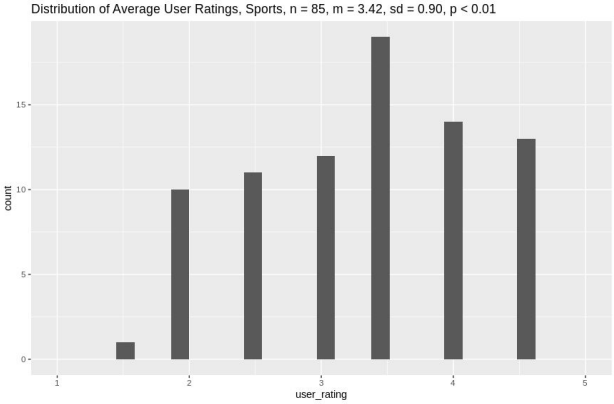
Appendix B:

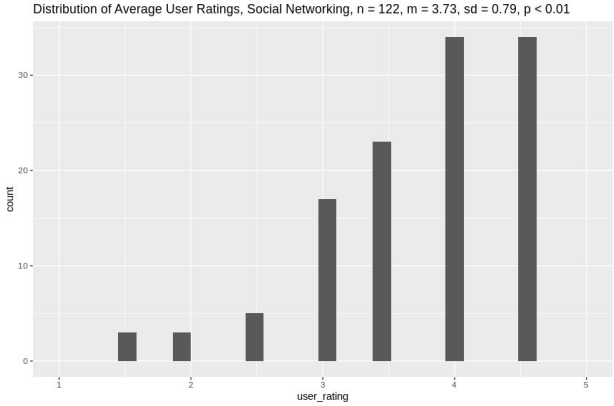
Included in this appendix is distributions of average app ratings by different genres, shown with histograms, and descriptions of the p values of the shapiro-wilk normality test (the lower it is, the less confident we are about the normality of the distribution).

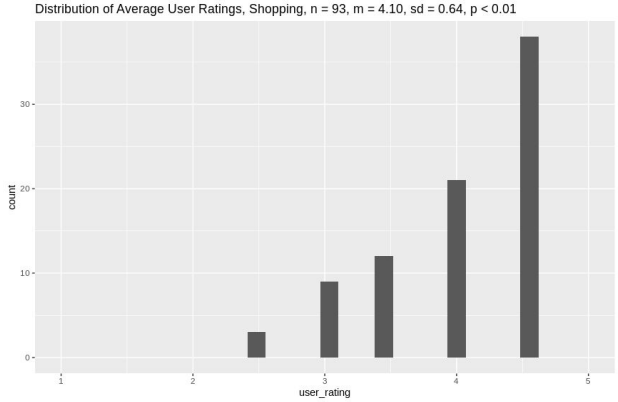

Distribution of Average User Ratings, Weather, n = 54, m = 4.03, sd = 0.59, p < 0.01
count
user_rating

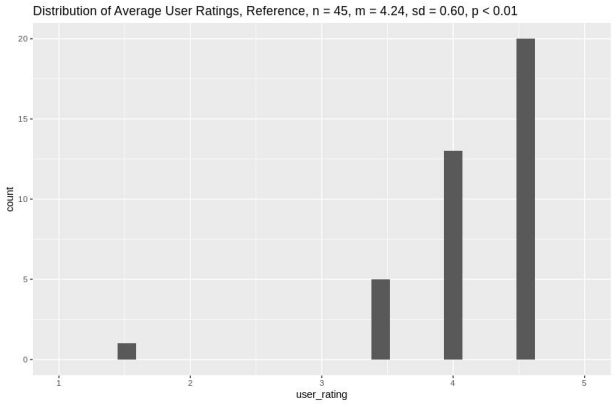

Distribution of Average User Ratings, Utilities, n = 185, m = 3.78, sd = 0.84, p < 0.01
count
user_rating

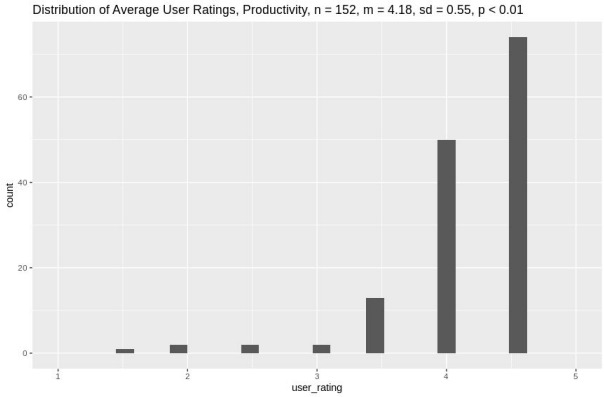

Distribution of Average User Ratings, Travel, n = 63, m = 3.93, sd = 0.75, p < 0.01
count
user_rating

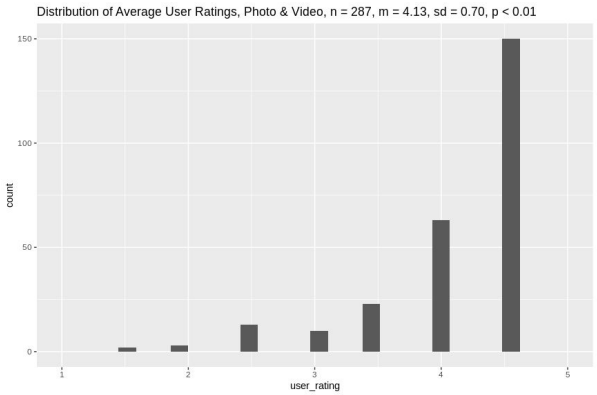

Distribution of Average User Ratings, Sports, n = 85, m = 3.42, sd = 0.90, p < 0.01
count
user_rating

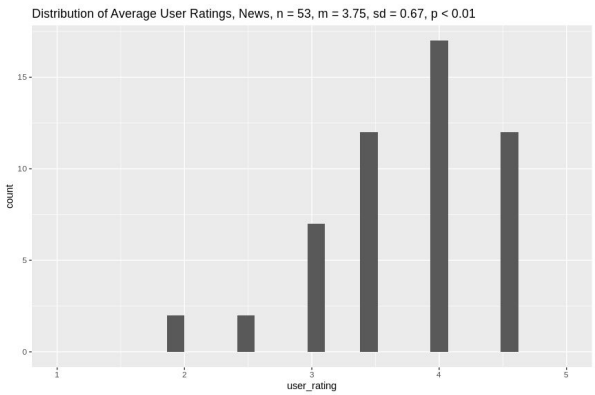

Distribution of Average User Ratings, Social Networking, n = 122, m = 3.73, sd = 0.79, p < 0.01
count
user_rating

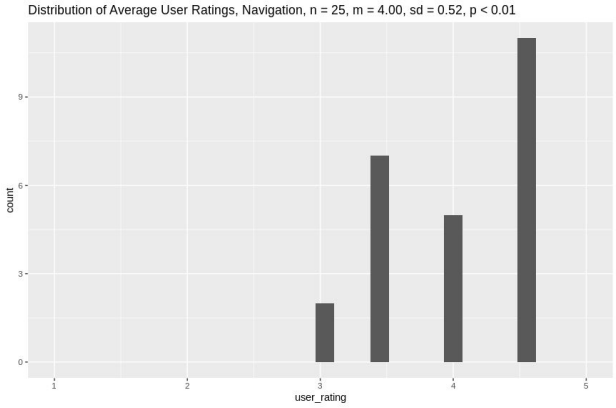

Distribution of Average User Ratings, Shopping, n = 93, m = 4.10, sd = 0.64, p < 0.01
count
user_rating

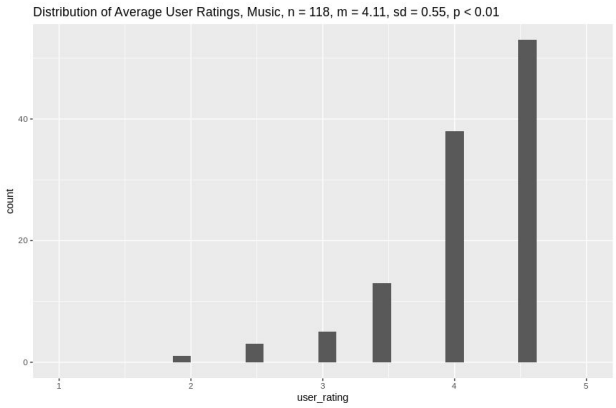

Distribution of Average User Ratings, Reference, n = 45, m = 4.24, sd = 0.60, p < 0.01
count
user_rating

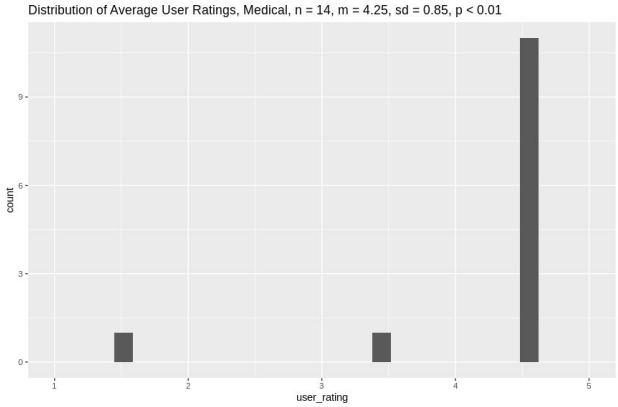

Distribution of Average User Ratings, Productivity, n = 152, m = 4.18, sd = 0.55, p < 0.01
count
user_rating

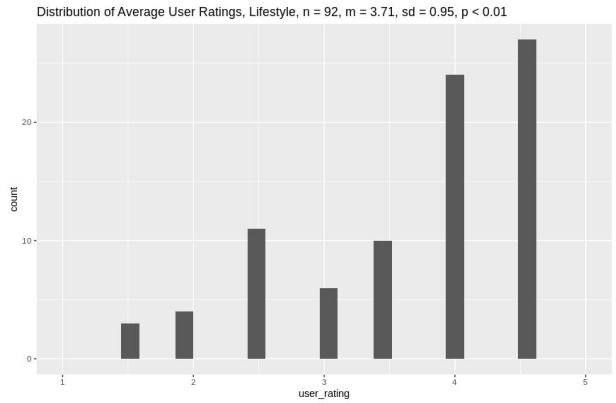

Distribution of Average User Ratings, Photo & Video, n = 287, m = 4.13, sd = 0.70, p < 0.01
count
user_rating

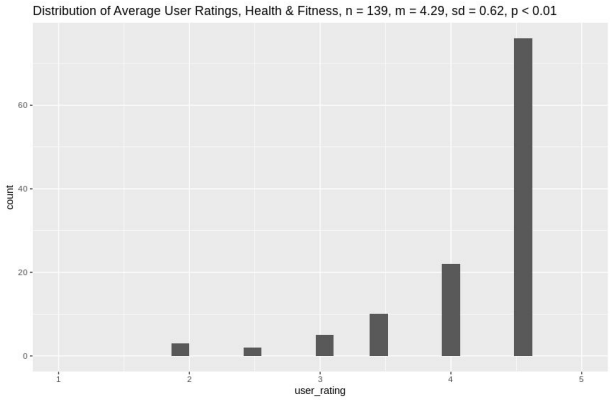

Distribution of Average User Ratings, News, n = 53, m = 3.75, sd = 0.67, p < 0.01
count
user_rating

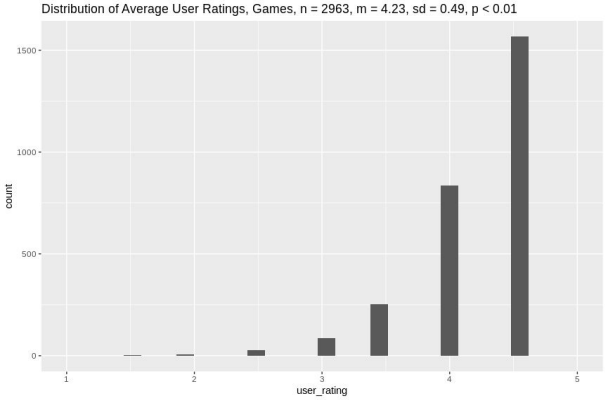

Distribution of Average User Ratings, Navigation, n = 25, m = 4.00, sd = 0.52, p < 0.01
count
user_rating

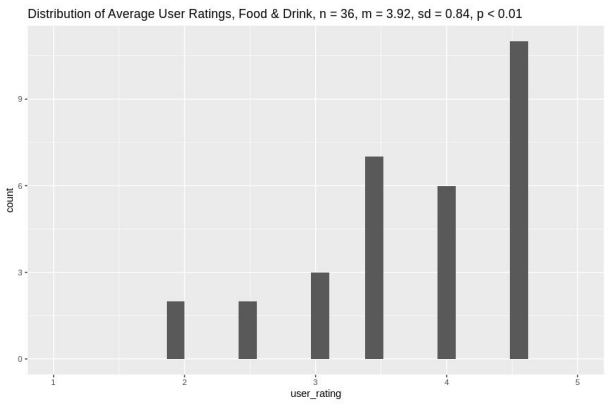

Distribution of Average User Ratings, Music, n = 118, m = 4.11, sd = 0.55, p < 0.01
count
user_rating

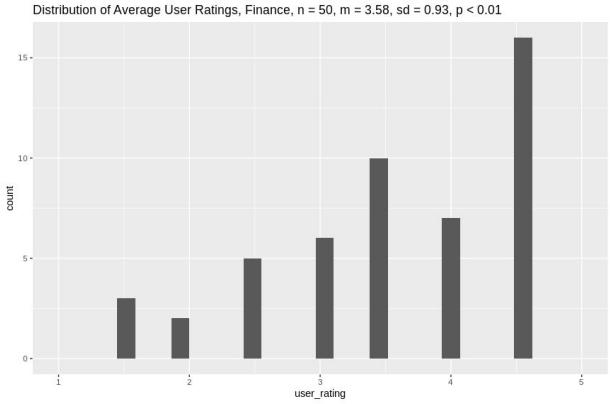

Distribution of Average User Ratings, Medical, n = 14, m = 4.25, sd = 0.85, p < 0.01
count
user_rating

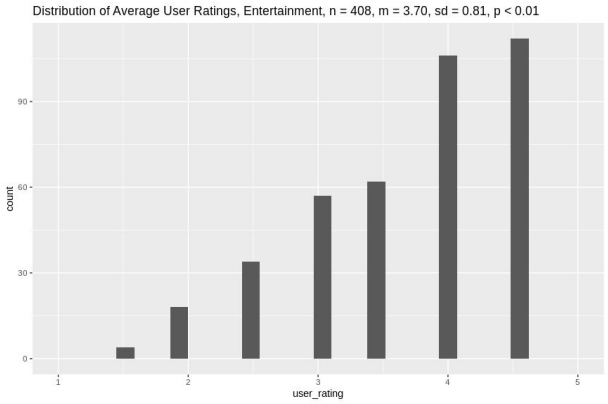

Distribution of Average User Ratings, Lifestyle, n = 92, m = 3.71, sd = 0.95, p < 0.01
count
user_rating

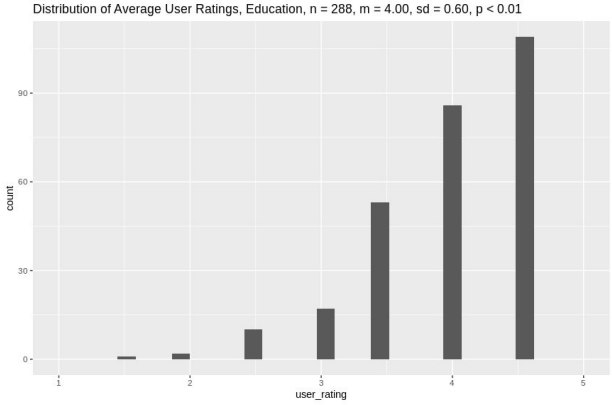

Distribution of Average User Ratings, Health & Fitness, n = 139, m = 4.29, sd = 0.62, p < 0.01
count
user_rating

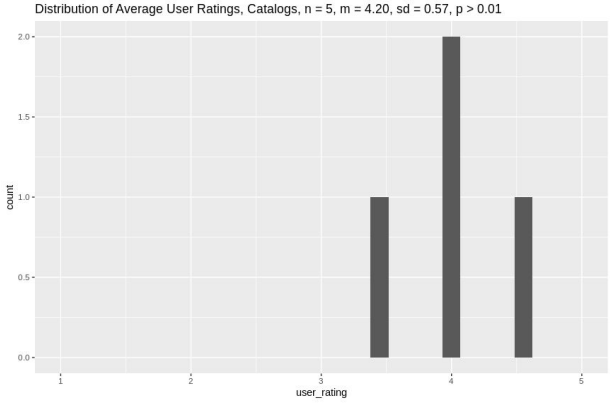

Distribution of Average User Ratings, Games, n = 2963, m = 4.23, sd = 0.49, p < 0.01
count
user_rating

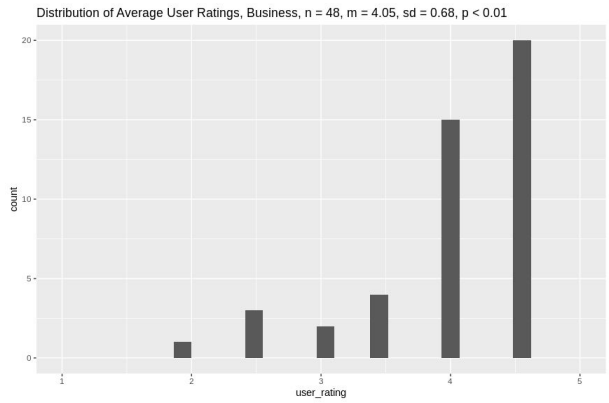

Distribution of Average User Ratings, Food & Drink, n = 36, m = 3.92, sd = 0.84, p < 0.01
count
user_rating

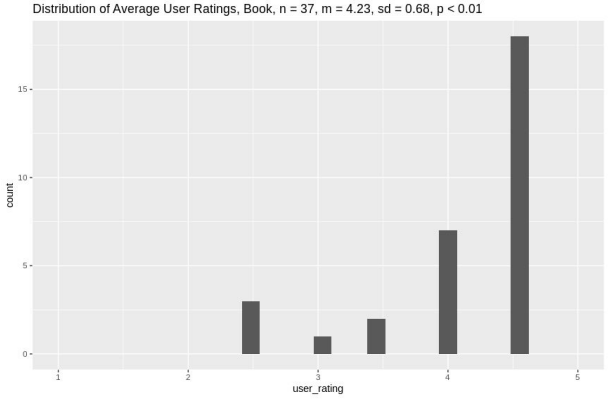

Distribution of Average User Ratings, Finance, n = 50, m = 3.58, sd = 0.93, p < 0.01
count
user_rating


Distribution of Average User Ratings, Entertainment, n = 408, m = 3.70, sd = 0.81, p < 0.01
count
user_rating


Distribution of Average User Ratings, Education, n = 288, m = 4.00, sd = 0.60, p < 0.01
count
user_rating


Distribution of Average User Ratings, Catalogs, n = 5, m = 4.20, sd = 0.57, p > 0.01
count
user_rating


Distribution of Average User Ratings, Business, n = 48, m = 4.05, sd = 0.68, p < 0.01
count
user_rating


Distribution of Average User Ratings, Book, n = 37, m = 4.23, sd = 0.68, p < 0.01
count
user_rating

Appendix C:

User Rating vs # of Languages Correlation: 0.170976276639192
Rating Count vs # of Languages Correlation: 0.137674716281491
User Rating this Version vs # of Languages Correlation: 0.175579616478395
Rating Count this Version vs # of Languages Correlation: 0.0132869345818919

Rating Count vs. # of Languages Correlation by Genre

Navigation	0.676335715
Medical	0.5742613964
News	0.5081977446
Reference	0.5070333027
Productivity	0.4906248899
Weather	0.404398071
Travel	0.2996641188
Book	0.2732762781
Utilities	0.2708604891
Business	0.2364874274
Social Networking	0.2326408041
Shopping	0.2045298604
Sports	0.1829764349
Lifestyle	0.1671909332
Photo & Video	0.1580176579
Entertainment	0.1151171455
Education	0.1140069356
Games	0.110467293
Health & Fitness	0.1062750588
Finance	0.0842137884
Music	0.06659335761
Food & Drink	-0.02033548743
Catalogs	-0.06123519183

Rating Count Current Version vs. # of Languages Correlation by Genre

Navigation	0.6787695789
News	0.4206289096
Book	0.3803521593
Medical	0.244502883
Travel	0.1867000602
Health & Fitness	0.1808849407
Social Networking	0.1666918568
Music	0.1386102454
Weather	0.1112611053

Productivity	0.1054815492
Reference	0.08462503709
Entertainment	0.07228110266
Business	0.04700236581
Photo & Video	0.04646080201
Finance	0.01662626643
Games	0.01057570243
Utilities	0.003301468864
Education	-0.0239099793
Shopping	-0.03362834009
Lifestyle	-0.03777511039
Food & Drink	-0.03938949106
Sports	-0.054728764
Catalogs	-0.0982444649

User Rating vs. # of Languages Correlation by Genre

Navigation	0.4032762836
Medical	0.3837935632
Catalogs	0.3247383979
Social Networking	0.3054161645
Utilities	0.2566672272
Health & Fitness	0.2309451273
Book	0.2220704216
Weather	0.2049288048
Reference	0.2009078676
Entertainment	0.1973657509
Lifestyle	0.1963945932
Education	0.191353167
Finance	0.1890738124
Games	0.1630538087
Business	0.1609487419
Music	0.152220198
Photo & Video	0.1419738303
Travel	0.1263537405
Food & Drink	0.0962527744
Shopping	0.08308233228
Productivity	0.06933983758
Sports	0.04128294726
News	-0.0401673562

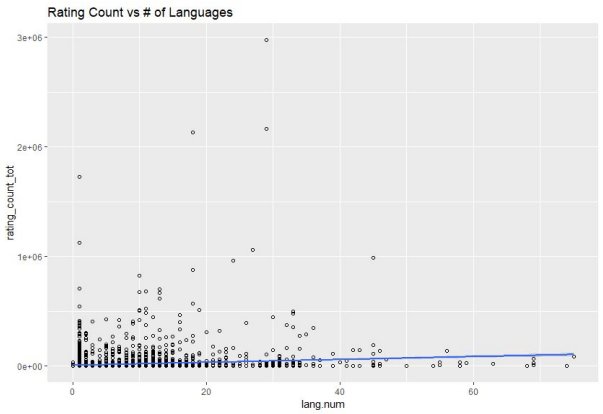
User Rating Current Version vs. # of Languages Correlation by Genre

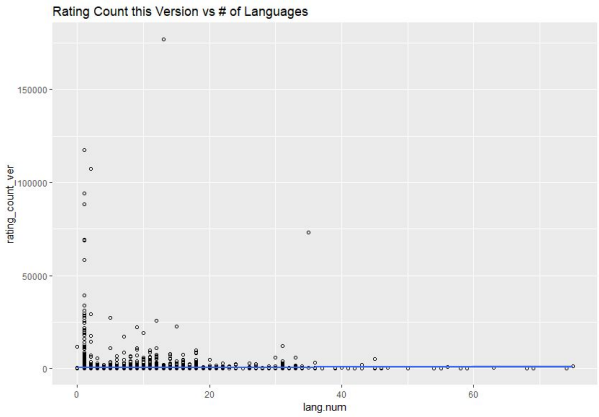
Catalogs	0.3892184869
Navigation	0.3682115137

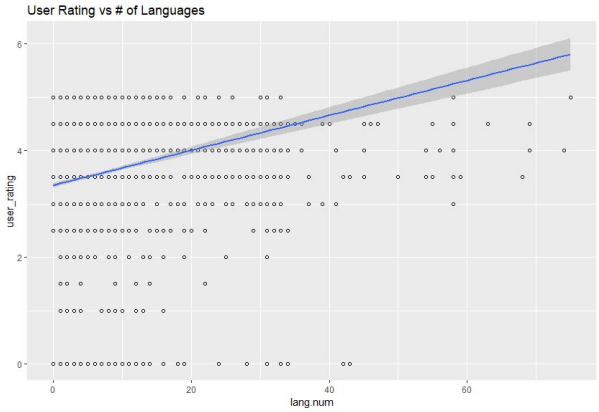
Medical	0.3214514652
Reference	0.3141150686
Utilities	0.3044357642
Social Networking	0.3029276502
Book	0.2768865231
Education	0.250554015
Music	0.2443176853
Lifestyle	0.2365887708
Health & Fitness	0.2297305703
Weather	0.2253550298
Entertainment	0.2118875033
Finance	0.2088767924
Shopping	0.197999057
Productivity	0.1721988501
Games	0.153187458
Business	0.1205436642
Travel	0.1131315341
Photo & Video	0.08378491561
Food & Drink	0.061705545
Sports	0.05098378591
News	0.03837862818

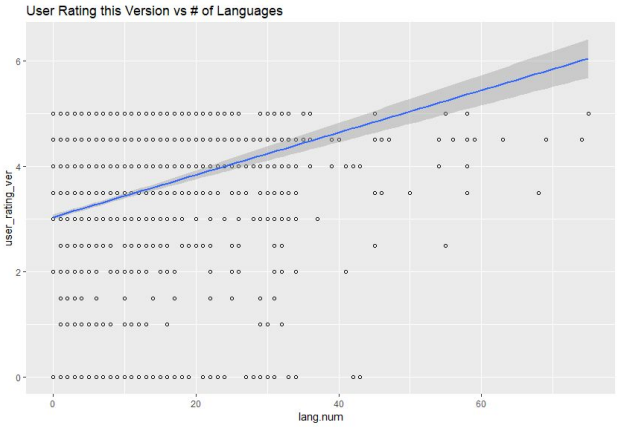
Appendix D:

Included in this appendix are graphs of rating count and user ratings vs the number of languages supported by the apps.


Rating Count vs # of Languages
rating_count_tot
3e+06
2e+06
1e+06
0e+00
0
20
40
60
lang.num


Rating Count this Version vs # of Languages
rating_count_ver
150000
100000
50000
0
0
20
40
60
lang.num


User Rating vs # of Languages
user_rating
6
4
2
0
0
20
40
60
lang.num


User Rating this Version vs # of Languages
user_rating_ver
6
4
2
0
0
20
40
60
lang.num