DSC530 Final Project

David Suffolk

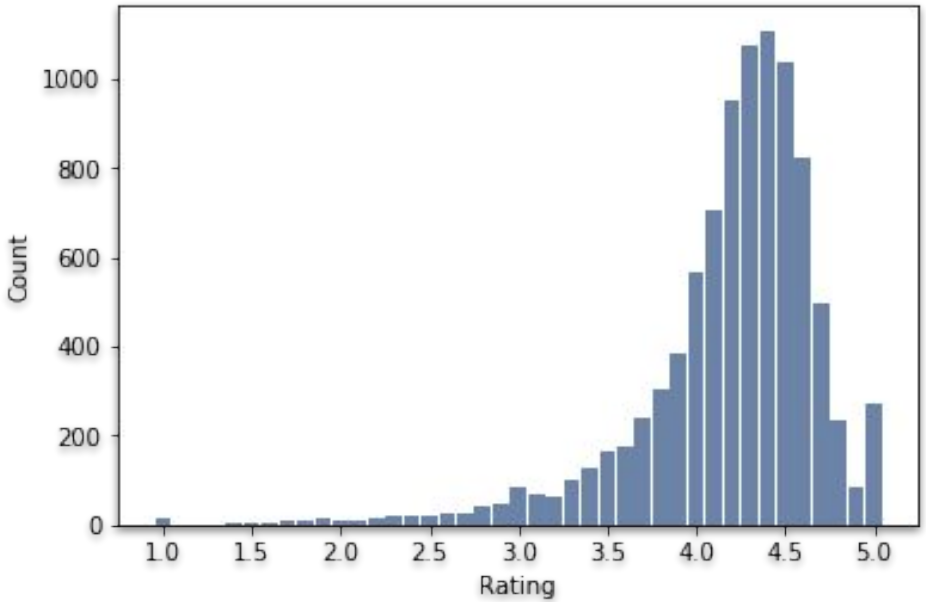
Statistical Question/Hypothesis

- If a customer pays for an app on their phone (compared to getting it for free), is that a signal of higher satisfaction with the product?
- The marketplace for apps is filled with free applications and many have the option to purchase a premium version while some are only available if purchased.
- If a customer is happy enough with the free app that they are willing to upgrade to a paid version, shouldn't paid apps have higher ratings in the app store?
- Since there are free options available and a customer chooses to purchase a different app instead, does this reflect higher satisfaction in the app through ratings and reviews?

Overview of Variables

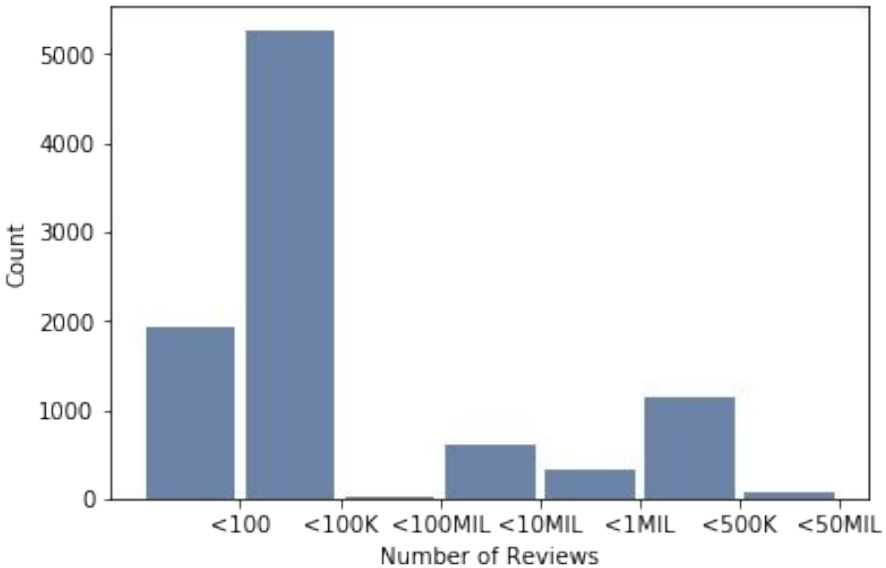
- The dataset contains information on apps within the Google Play Store
- The five variables that will be the focus of the project are the following:
  - Rating - The overall rating the app has received based on user reviews
  - Reviews - The number of reviews the app has received
  - Installs - The number of installs the app has received
  - Type - Binary variable defining if the app is Free or Paid
  - Content Rating - The recommended maturity level for the user accessing the app
- Rating, Reviews, and Installs all speak to the engagement between user and app and can give an overview of how a particular app is viewed by users.
- Content Rating is one method of determining what types of apps are installed, rated, and reviewed.

Rating



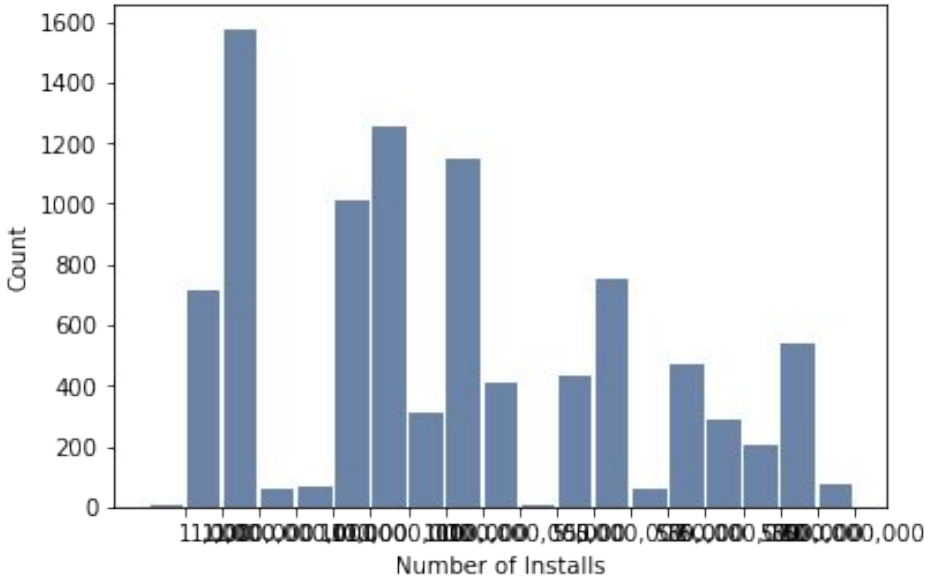
- Histogram shows a left-skewed distribution
- Mean: 4.192
- Mode: 4.4
- Median: 4.3
- Variance: 0.265
- Standard Deviation: 0.515

Reviews



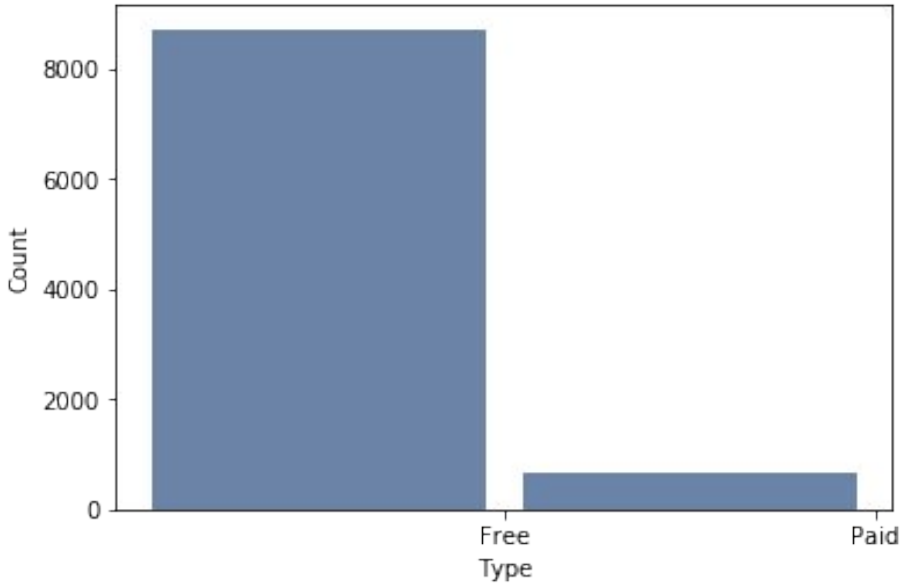
- Histogram shows that most apps receive less than 100K reviews
- Mean: 514,049
- Mode: 2
- Median: 5,930.5
- Variance: 9,885,000,896,407.732
- Standard Deviation: 3,144,042.127

Installs



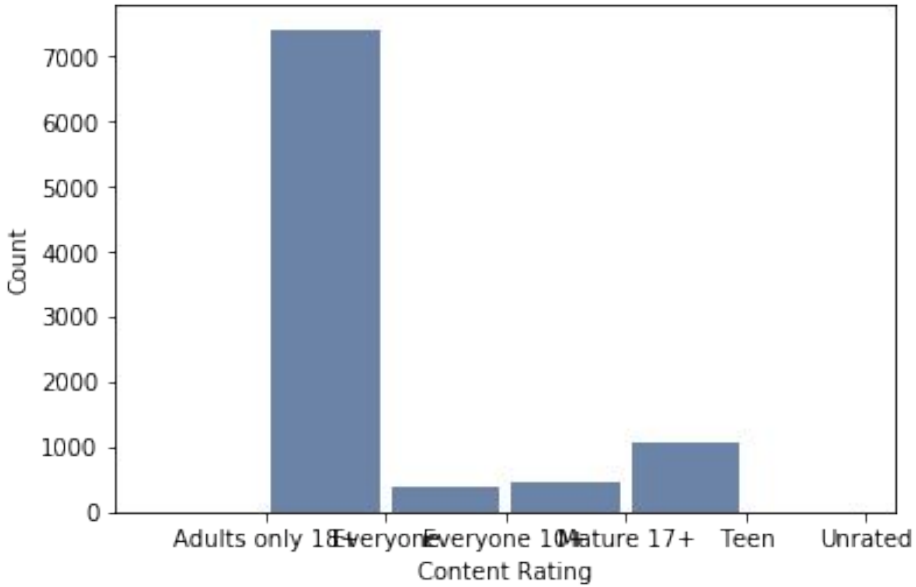
- The histogram shows that most apps have less than 1 million installs
- Mean: 17,897,443.726
- Mode: 1,000,000
- Median: 500,000
- Variance: 8,324,412,310,163,610
- Standard Deviation: 91,238,217.377

Type



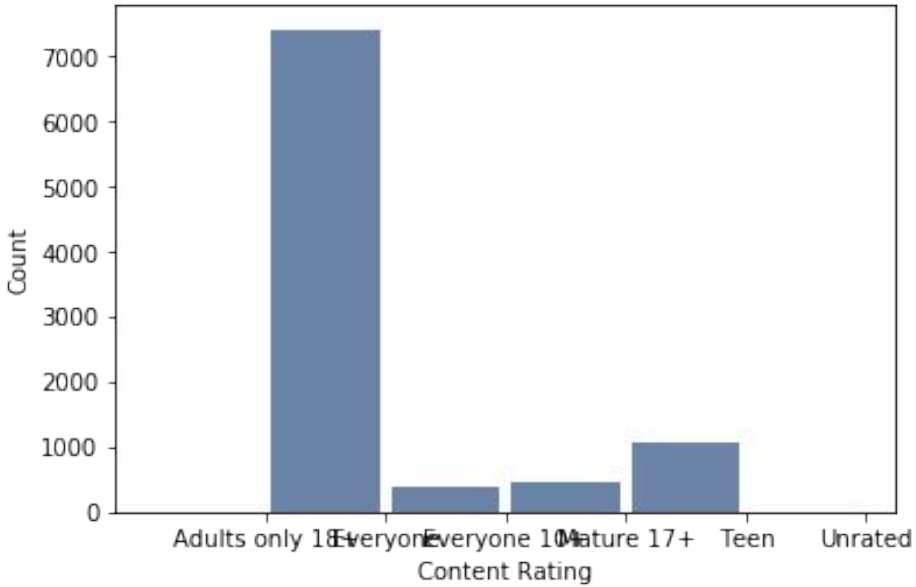
- Histogram shows that most of the apps in the data are Free
- For calculations, the values were made binary. Free = 0. Paid = 1.
- Mean: 0.069
- Mode: 0
- Median: 0
- Variance: 0.064
- Standard Deviation: 0.254

Content Rating



- The histogram shows that most of the apps have a Content Rating of Everyone
- For calculations, the following numeric values were assigned:
  - Unrated - 0
  - Everyone - 1
  - Everyone 10+ - 2
  - Teen - 3
  - Mature 17+ - 4
  - Adults only 18+ - 5

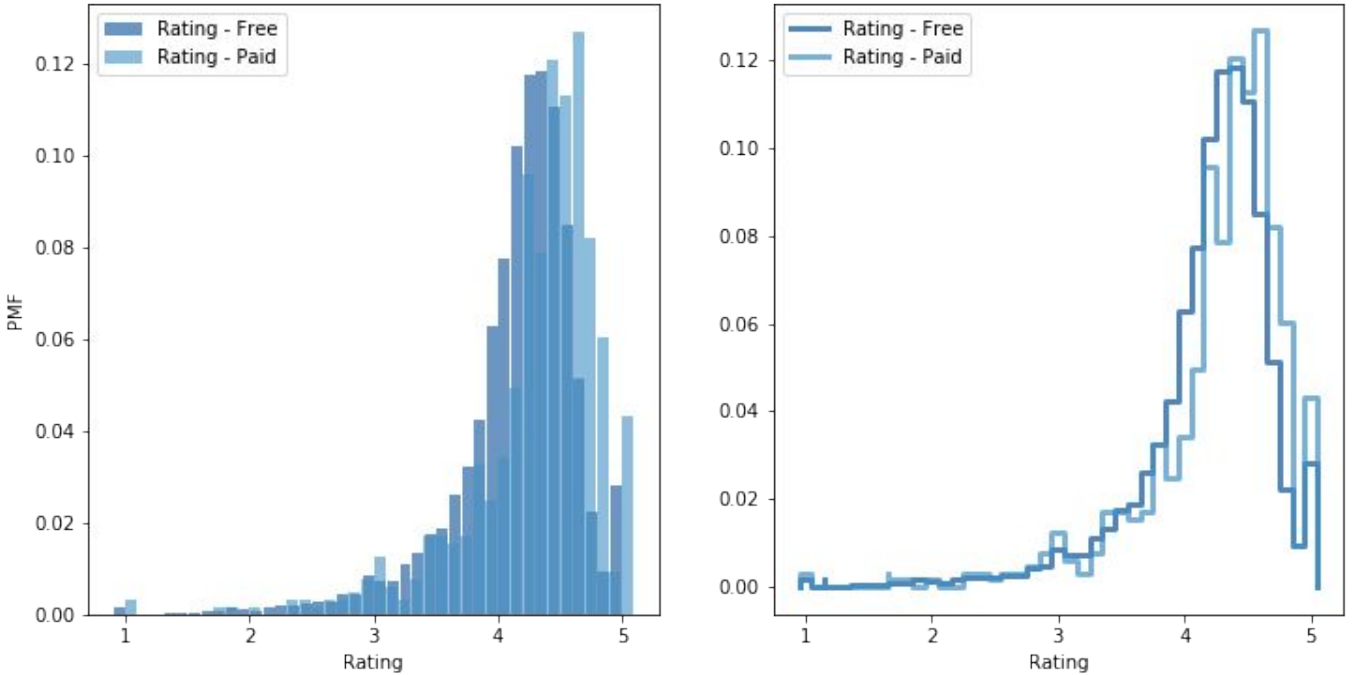
Content Rating



Unrated - 0, Everyone - 1, Everyone 10+ - 2, Teen - 3, Mature 17+ - 4, Adults only 18+ - 5

- Mean: 1.423
- Mode: 1 (Everyone)
- Median: 1 (Everyone)
- Variance: 0.775
- Standard Deviation: 0.88

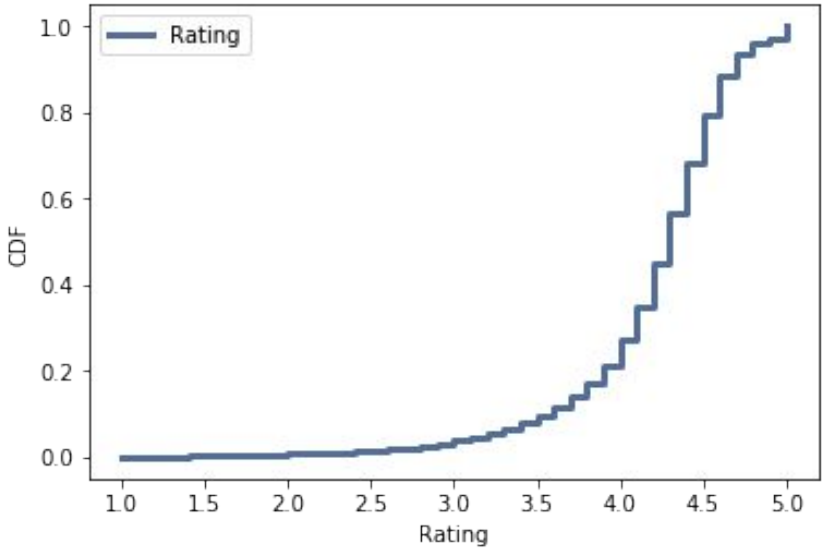
PMF


Rating - Free
Rating - Paid
PMF
0.12
0.10
0.08
0.06
0.04
0.02
0.00
1
2
3
4
5
Rating
Rating - Free
Rating - Paid
0.12
0.10
0.08
0.06
0.04
0.02
0.00
1
2
3
4
5
Rating

PMF

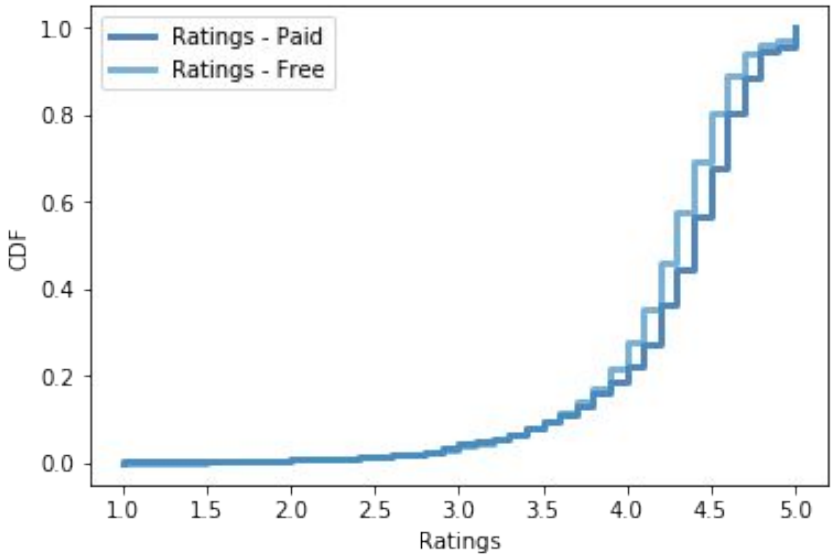
- For PMF, I looked into the ratings of Free apps against the ratings of Paid apps
- Based on the graphs on the previous slide, Paid apps do seem to receive a higher frequency of higher ratings.

CDF



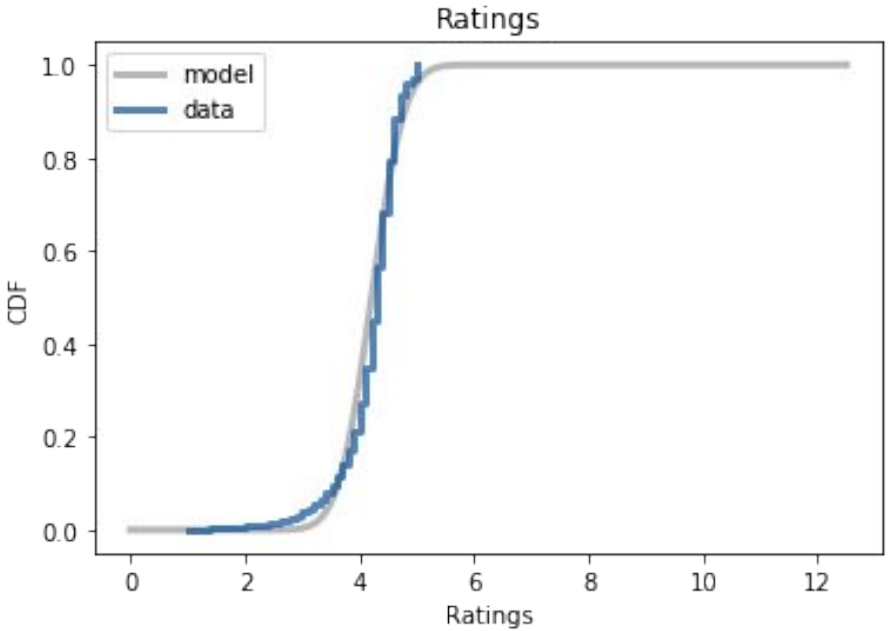
- The CDF for Ratings shows that a very low percentage of apps get a rating lower than 4. In fact, only about 20% of the apps have a rating that is below 4 out of 5.

CDF Comparison



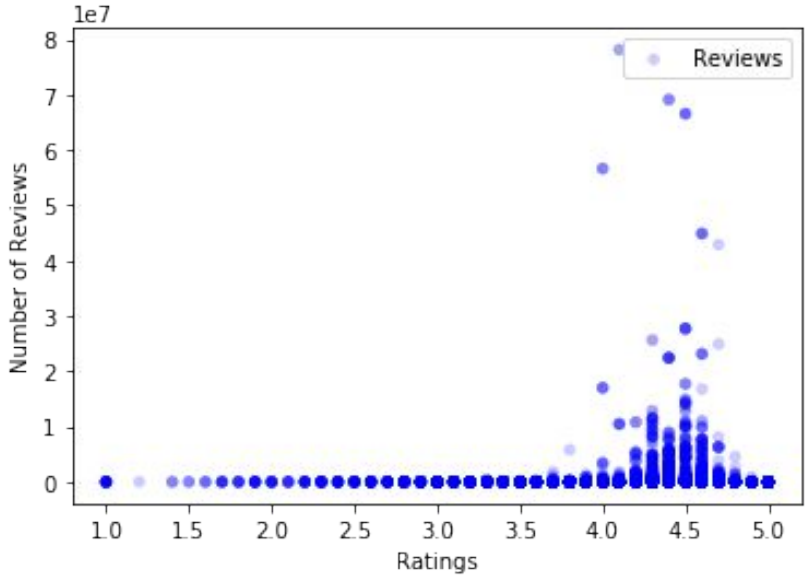
- When comparing the CDF of ratings for Paid and Free apps, we see that the line for Paid apps is slightly to the right of Free apps.
- We can see that there is a slightly higher chance of Paid apps receiving a higher rating than Free apps.

Analytical Distribution



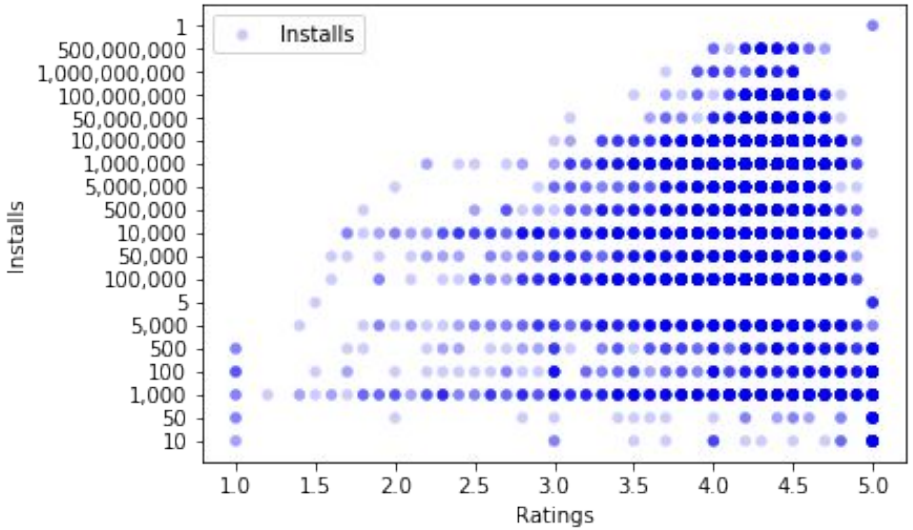
- Compare the CDF of Ratings to a normal model with the same variance and mean
- Mean: 4.208
- Variance: 0.202
- The data for the CDF of Ratings fits the model very closely indicating a normal distribution for a model with the same variance and mean.

Scatter Plot #1



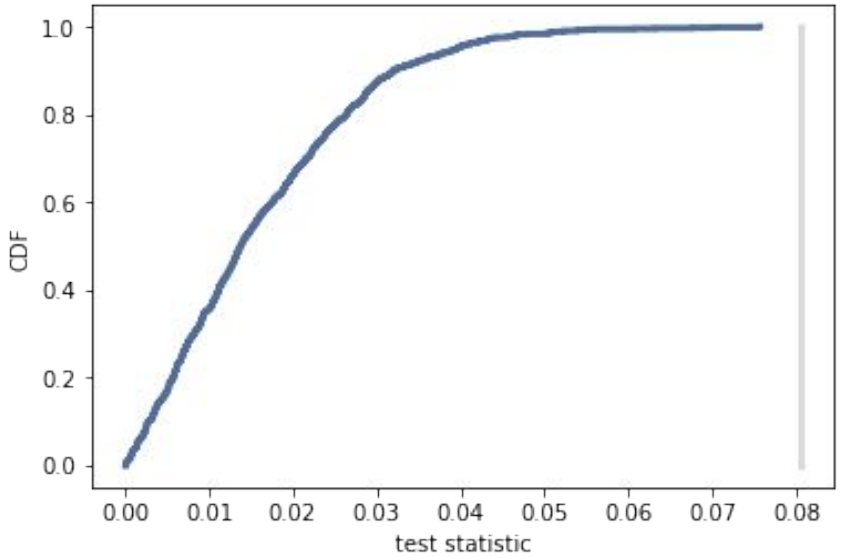
- The Scatter Plot shows the Rating against the Number of Reviews
- We can see that most reviews give a rating between 4 and 5
- Covariance: 110,368.519
  - Indicates a positive relationship between the two variables
- Pearson: 0.068
  - Close to zero indicating no relationship
- Spearman: 0.156
  - Higher than 0.05 so it is not statistically significant

Scatter Plot #2



- The Scatter Plot shows Ratings against the number of Installs
- We can see that as the installs increase, the variance in Ratings decreases
- Covariance: 2,413,799.731
  - Indicates a positive relationship between the two variables
- Pearson: 0.051
  - Close to zero indicating no relationship
- Spearman: 0.0695
  - Slightly higher than 0.05 so we should consider it statistically insignificant

Hypothesis Test



- I used Difference of Means with the Null Hypothesis that there is no difference between the ratings for Paid apps vs. Free apps.
- The result was 0 which indicates that the null hypothesis should be accepted indicating that there is no significant difference between the ratings for Paid and Free apps.

Regression Analysis

```
Optimization terminated successfully.
         Current function value: 0.250396
         Iterations 7
```

Logit Regression Results

Dep. Variable:	Type_Numeric	**No. Observations:**	9366
Model:	Logit	**Df Residuals:**	9364
Method:	MLE	**Df Model:**	1
Date:	Thu, 08 Aug 2019	**Pseudo R-squ.:**	0.003406
Time:	18:05:41	**Log-Likelihood:**	-2345.2
converged:	True	**LL-Null:**	-2353.2
		LLR p-value:	6.236e-05

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-4.0659	0.388	-10.490	0.000	-4.826	-3.306
Rating	0.3465	0.090	3.831	0.000	0.169	0.524

- Compared Rating as a predictor of Type (Free vs. Paid)
- The Coefficient is 0.3465 which means that there is a moderate positive relationship between the rating of a paid app.

Regression Analysis

```
Optimization terminated successfully.
         Current function value: 0.208602
         Iterations 15
```

Logit Regression Results

Dep. Variable:	Type_Numeric	**No. Observations:**	9366
Model:	Logit	**Df Residuals:**	9361
Method:	MLE	**Df Model:**	4
Date:	Thu, 08 Aug 2019	**Pseudo R-squ.:**	0.1697
Time:	18:09:26	**Log-Likelihood:**	-1953.8
converged:	True	**LL-Null:**	-2353.2
		LLR p-value:	1.317e-171

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.6718	0.339	-10.822	0.000	-4.337	-3.007
Rating	0.4861	0.077	6.283	0.000	0.334	0.638
Reviews	5.618e-06	4.64e-07	12.103	0.000	4.71e-06	6.53e-06
Installs_Numeric	-2.537e-06	2.07e-07	-12.271	0.000	-2.94e-06	-2.13e-06
Content_Rating_Numeric	-0.1114	0.057	-1.961	0.050	-0.223	-3.51e-05

- Compared all variables as a predictor of Type (Free vs. Paid)
- The Coefficient for Rating increases when controlled for Reviews, Installs, and Content Rating. The value is 0.4861.
- However, this relationship is still a moderately positive.