Google Play Store Analysis With AppleStore Comparison

Table of Contents:

Title and author:

1. **Title**:

   Google Play Store Analysis With AppleStore Comparison

2. **Author**:

   - Name: Thai Quoc Hoang
   - Date of birth: January 09, 2000

- Email: quocthai9120@gmail.com (mailto:quocthai9120@gmail.com) / qthai912@uw.edu (mailto:qthai912@uw.edu)

Summary of research questions and Results:

1. Is there a strong linear correlation between the price of an application and the number of installs? Is there a strong linear correlation between rating and number of install? Is there a strong linear correlation between reviews and number of installs?

    - Result:
        - Correlations between "Installs" and another factor:
            - "Reviews" and "Installs": Even though the Spearman Correlation Coefficient gives a high result (about 0.96 with p_value near 0.0), because the graph of "Reviews" and "Installs" does not clearly show that relationship and the Spearman Correlation Coefficient is just a rank correlation, we could not give any conclusion about the correlation of "Reviews" and "Installs" at this time rather than they have a postive correlation.
            - "Rating" and "Installs": The Spearman Correlation Coefficient gives a low result (about 0.03 with p_value near 0.007) and the graph also shows they do not have strong correlation but they could have positive relationship to other.
            - "Size" and "Installs": The Spearman Correlation Coefficient gives a low result (about 0.2976 with p_value near 0.0) and the graph also shows they do not have strong correlation but they could have positive relationship to other.
            - "Price" and "Installs": The Spearman Correlation Coefficient gives a low result (about -0.2639 with p_value near 0.0) and the graph also shows they do not have strong correlation but they could have negative relationship to other.
        - Between other pair of factors:
            - "Price" and "Reviews": Even though between 'Price' and 'Reviews' there is a result of Pearson's correlation coefficient is about -0.01, the p-value of that result is about 0.37. Therefore, we cannot give a conclusion for this by the result above.
            - "Price" and "Rating": Between 'Price' and 'Rating' there is a result of Pearson's correlation coefficient is about -0.02 with p-value near 0.06 (slightly over 0.05), which shows that they do not have strong correlation. However, based on the result and the graph, we can recognize that 'Price' and 'Rating' may have a negative relationship.
            - "Price" and "Size": Even though between 'Price' and 'Size' there is a result of Pearson's correlation coefficient is about 0.01, the p-value of that result is about 0.1, which is higher than 0.05. Therefore, we cannot give a conclusion for this by the result above.
            - "Rating" and "Reviews": The Pearson Correlation Coefficient gives a low result (about 0.08 with p_value near 2.14^{-12}) and the graph also shows they do not have strong correlation but they could have positive relationship to other.
            - "Rating" and "Size": Even though between 'Rating' and 'Size' there is a result of Pearson's correlation coefficient is about -0.02, the p-value of that result is about 0.09, which is higher than 0.05. Therefore, we cannot give a conclusion for this by the result above.

    - "Reviews" and "Size": The Pearson Correlation Coefficient gives a low result (about 0.04 with p_value near 0.001) and the graph also shows they do not have strong correlation but they could have positive relationship to other.

2. If there is no strong correlation between these factors, are there any better relationships between these factors?
    - Result:
        - We could not able to find a relationship between "Price" and "Installs".
        - "Reviews" and "Installs" have a power relationship that log(Reviews) and log(Installs) are correlated.
        - "Rating" and "Installs" have an exponential relationship because when we group "Installs" by "Rating" (get the mean) then take log of that, there is a strong correlation between that value and "Rating" with a high result of Pearson Correlation Coefficient (about 0.7 with p_value near 6.3^{-7})
        - "Rating" and "Reviews" have an exponential relationship because when we group "Reviews" by "Rating" (get the mean) then take log of that, there is a strong correlation between that value and "Rating" with a high result of Pearson Correlation Coefficient (about 0.8 with p_value near 2.4^{-9})

3. Is the difference between the mean size of applications in Google Play Store dataset and size of applications in Appstore dataset is statistically significant?
    - Result: Observed difference between Applications mean size between Google Play Store and Apple Store, about -77.599 MB, is statistically significant.

4. What can we conclude about the mean size of applications on Google Play Store and Apple Store based on the dataset?
    - Result:
        - Google Play Store Applications mean size with 95% confidence: from about 17.4 MB to about 25.0 MB.
        - Apple Store Applications mean size with 95% confidence: from about 84.0 MB to about 113.9 MB.

Motivation and background:

- Nowadays, there is a huge number of smartphone applications. Moreover, there are also a huge number of applications that used to get a task done. From the publisher's perspective, so as to get more users for their applications, they need to understand the current state of the market that they publish the application to. From the users' perspective, in order to find the applications that really help them with their task, they also need to understand the market that they are finding application from.
- In the context of markets for smartphone applications, while Google Play Store and Apple Store are the two most popular markets, Google Play Store is undeniable currently much more dominating compared to Apple Store thanks to the Android Platform, according to comScore (https://www.comscore.com/Insights/Market-Rankings/comScore-Reports-January-2015-US-Smartphone-Subscriber-Market-Share). With that reason, it is extremely important to analyze Google Play Store market to understand several relationships and facts on that market.

- This project will analyze Google Play Store market to find the relationships between different numeric factors and find the factor that most affect another. Moreover, the project will also analyze the distribution of the factors to understand the behavior of the market.
- The statistical data about the relationships between different numerical factors on Google Play Store market is important and necessary for industries, developers, and publishers to compute and manage their products, predict the profits, clients, values, ... of the products, and to predict and estimates other products. Besides, the user can be based on the statistics to know what criteria they should look for while finding an application, which will improve their productivity.
- Moreover, it is necessary to understand the difference between the Android market and the IOS market so as to understand the state of the Android market. Therefore, I will also briefly compare the distribution of sizes of applications on the Google Play Store analysis with the Apple Store. With that results, people will be able to balance the products to be suitable with the market when creating applications, which will help them to plan for their products to serve a larger range of clients.

Datasets:

- Two datasets use for this project are a dataset of 10,000 Application on GooglePlay Store from Kaggle: Google Play Store App (https://www.kaggle.com/lava18/google-play-store-apps/home) and a dataset of 10,000 Applications on Apple Store from Kaggle: App Store Apple Data set (10K apps) (https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps/version/2).
- In this project, I am analyzing the "googleplaystore.csv" and "AppleStore.csv" to answer my research questions.
- The datasets can be downloaded by accessing the links above to download officially from the sources. I have also provides the csv files in the "zip" file with this notebook.

Google Play Store dataset:

According to the source of the Google Play Store dataset, the googleplaystore.csv file includes information about 10,000 applications with:

- **Name**: Application's name
- **Category**: The category of the application
- **Rating**: Users rating
- **Reviews**: Number of reviews
- **Size**: Application's size
- **Installs**: Number of installs
- **Type**: Free/Paid application
- **Price**: Price of application
- **Content Rating**: Age group the app is targeted at
- **Genres**: An app can belong to multiple genres (apart from its main category)
- **Last updated**: Application's last date of update
- **Current Version**: Current version of application
- **Android Version**: Minimum Android version required to run

Apple Store dataset:

According to the source of the Apple Store dataset, the AppleStore.csv file includes information about 10,000 applications with:

- **id**: ID of application
- **track_name**: Application's name
- **size_bytes**: Application's size (in Bytes)
- **currency**: Currency
- **price**: Price
- **rating_count_tot**: Total user rating counts
- **rating_count_ver**: user rating counts for current version
- **user_rating**: Average user rating value for all versions
- **user_rating_ver**: Average user rating value for current version
- **ver**: Latest version
- **cont_rating**: Age group the app is targeted at
- **prime_genre**: Primary Genre
- **sup_devices.num**: Number of supporting devices
- **ipadSc_urls.num**: Number of screenshots showed for display
- **lang.num**: Number of supported languages
- **vpp_lic**: Vpp Device Based Licensing Enabled

To get more information about the datasets, please go to the sources of the datasets by the links provided above!

Methodology:

Part 1: Data pre-processing:

- So as to analyze the 'googleplaystore.csv' dataset and 'AppleStore.csv', we have to pre-process the datasets so that we can get values from the datasets latter.
- For Google Play Store dataset (googleplaystore.csv), we need to remove rows that include NaN values, removes rows contain "Varies with device" in "Size" column, and remove unnecessary characters in "Size" ('k' and 'M'), "Price" ('$'), and "Installs" (',' and '+') columns. After removing unnecessary characters in these columns, we can read and store "Rating", "Reviews", "Size", "Installs", "Price" columns as DataFrames of type float.
- For Apple Store dataset (AppleStore.csv), we need to convert the size of applications in 'size_bytes' column to megabytes (MB). Then we can store 'size_bytes' column after converting to MB, 'price' column and 'user_rating' column as DataFrames of type float.

Part 2: Answer Research Questions:

Question 1:

We first let the null hypothesis to become: There are no correlation between any pairs of numerical factors in googleplaystore.csv dataset.

Then, we need to reject that null hypothesis by:

- Find linear correlation between 2 of the factors: "Size", "Price", "Installs", "Reviews", "Rating" in Google Play Store dataset by calculating the correlation coefficient of each pair of factors.
  - As values in "Installs" do not have a normal distribution, we need to use the Spearman correlation coefficient to find the correlation between "Installs" and another factor.
  - For other pair of factors, we can use the Pearson correlation coefficient to find the correlation between these pairs.
  - Besides of getting the results of the correlation coefficients, we need to get the p_value for these computations for reasoning.
- After calculating the correlation coefficient, we need to draw a graph of each pair that we have calculated to check the calculations above.
- Lastly, use the results above to give a conclusion.

Question 2:

As the range of each factor ("Size", "Price", "Installs", "Reviews", "Rating") in Google Play Store dataset is different, we may not able to find a strong correlation between each pair of factors in Question 1. Therefore, depends on the distribution of each pair of factors, they may have a relationship other than a linear correlation.

1. Find the relationship between 'Price' and 'Installs':

   - Because the price of most applications in the dataset has a range from 0 to 400 USD, while the number of installs gets up to 10^9, we need to shrink the range of the "Installs" column by getting its logarithm to make it less extreme.
   - We then let the null hypothesis to become: There are no correlation between these 2 factors and try to reject the null hypothesis.
     - After getting the logarithm of "Installs" column, we then find the Pearson correlation coefficient between 'Price' and 'Installs' column.
     - We need to get the p_value for these computations for our reasoning.
   - At this point, we can give a conclusion that whether "Installs" has an exponential relationship with respect to "Price" or not based on whether " $Price$ " has correlation with $a * log(" \ Installs \ ") + b$ or not.
   - Draw the scatterplot with a linear regression line using "Price" as x_axis and "Installs" as y_axis to check the computed result.

2. Find relationship between 'Review' and 'Installs':

   - In this case, because both "Review" and "Installs" have a large range, but the range of these factors are different, that is, while the range of "Review" is from 0 to more than 10000, the range of Installs runs through more than 10^9, we need to get logarithm of both columns.
   - We then let the null hypothesis to become: There are no correlation between these 2 factors and try to reject the null hypothesis.

    - Then find the Pearson correlation coefficient between 'Reviews' and 'Installs' column after getting logarithm of both of them.
    - Besides, we need to get the p_value for these computations for our reasoning.
  - Give a conclusion of the relationship of log("Reviews") and log("Installs") then reflect to the relationship of "Review" and "Installs".
  - We do need to draw a graph with scatterplot and linear regression line using "Reviews" as x_axis and "Installs" as y_axis to check the computed result.

3. Find the relationship between 'Rating' and other factors ('Installs', 'Reviews'):

   - Because for each value of rating there is a huge range of values of other factors. Therefore, one way to find a relationship is to group 'Rating' rows in the dataset with the mean of other rows that need to compare (get the mean of all values of "Installs" or "Reviews" that have the same value of "Rating")
   - After grouping the mean of each factor by each value of "Rating", we need to take the logarithm of these values ("Installs", "Reviews").
   - We then let the null hypothesis to become: There are no correlation between these 2 factors and try to reject the null hypothesis.
     - Find the Pearson correlation between "Rating" and each other factors after taking logarithm of them.
     - Besides, we need to get the p_value for these computations for our reasoning.
   - Give a conclusion about the exponential relationships between "Rating" and "Reviews"/ "Rating" and "Installs" based on the results above.
   - Plot the graph of 'Rating' (x_axis) and mean of that factor by 'Rating' column (y_axis) using scatterplot for real values and a line for that factor values predicted by 'Rating' using Linear Regression to check the accuracy of the result after calculated.

Question 3:

We set the hypothesis that our observed difference between mean size of applications in Google Play Store and Apple Store is not statistically significant as the null hypothesis. In order to reject that hypothesis, we need to:

- Define a function to get the confidence interval.
- Define a function to remove outlier.
- Remove outliers of "Size" column in both datasets to make the observation be more accurate.
- Calculate the difference between the mean size of applications in Google Play Store dataset and size of applications in Appstore dataset.
- Loop 10000 times. Each time randomly shuffle some values between the "Size" column in GooglePlayStore and Apple Store dataset then compute the mean of all applications' sizes in each dataset to add to the lists of mean for each dataset.
- Compute the 95% population distribution of the model.
- Plot a graph to show the result of chance to get the original result.

Question 4:

- Loops 10000 times. Each time get 100 random rows in each dataset to compute the mean size of that 100 rows then adds that mean value to the list of means for each dataset.
- Compute the 95% population distribution for the models of mean applications size on 2 datasets.
- Draw 2 histogram (Play Store and Apple Store) to show the model calculated above.
- As we have found the mean size of applications on Play Store and Apple Store based on the dataset, in order to helps the reader understand how the applications on both Play Store and Apple Store distributes with respect to the size of applications, we could draws 3 histograms (Apple Store, Play Store, and both Apple Store and Play Store) to show the distribution of applications' size (x_axis) by relative frequency (y_axis) next to each other (all of them need to be removed outliers). Label x_axis, y_axis as the two factors and set the title to help readability.

Results:

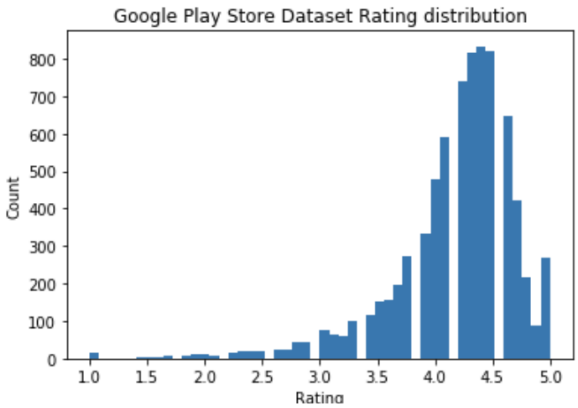
Question 1 result:

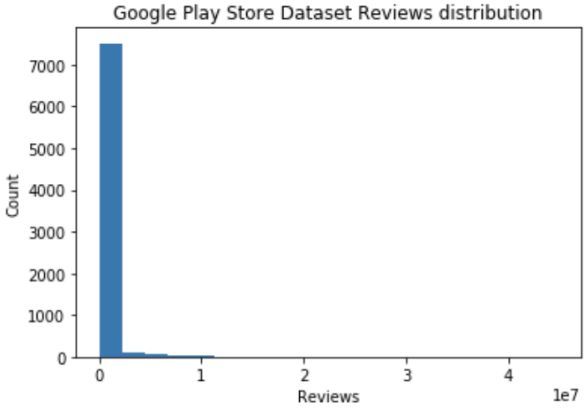
Is there a strong linear correlation between the price of an application and the number of installs? Is there a strong linear correlation between rating and number of install? Is there a strong linear correlation between reviews and number of installs?

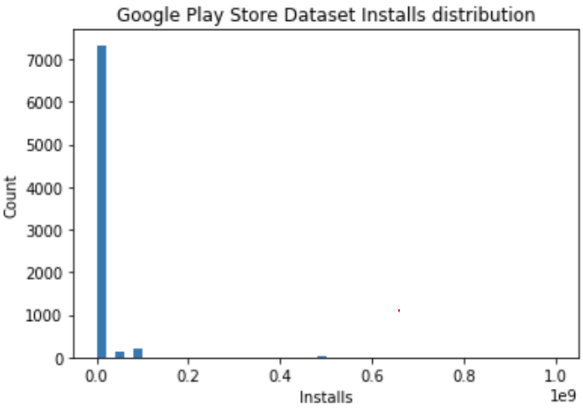
Google Play Store Dataset Distribution:

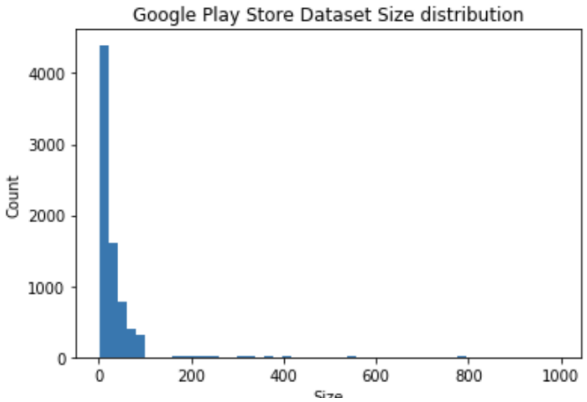
Conclusion:

- "Installs" is a categorical data that is not normally distributed.
- "Rating" is a categorical data that is ordinal and normally distributed.
- "Reviews" is a continuous data.
- "Size" is a continuous data.
- "Price" is a continuous data with majority elements are 0.00.









Discussions:

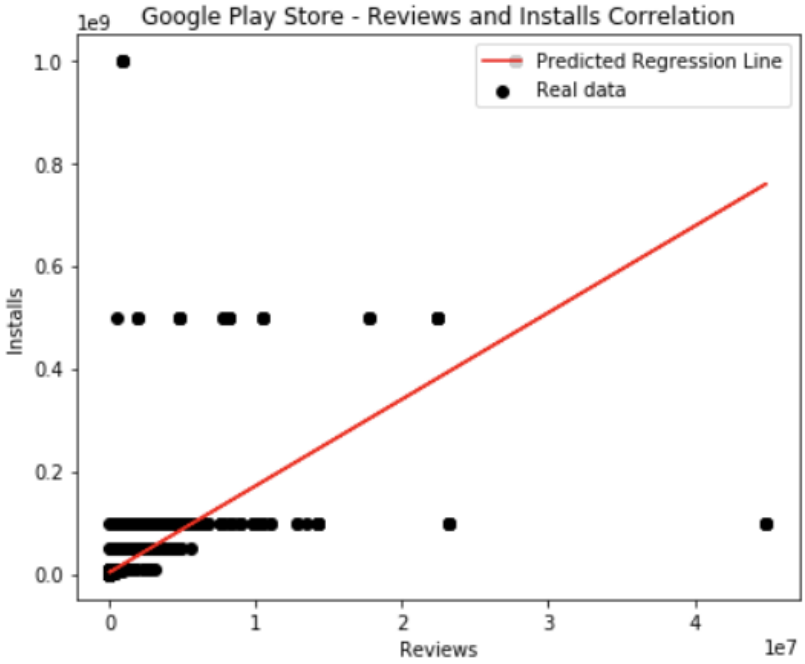
- We need to calculate correlations between "Installs" and another factor using Spearman's correlation coefficient.
- For other factors, we can calculate the correlations between them using Pearson's correlation coefficient. (Notice here, even though "Rating" is a categorical data, because it is ordinal and normally distributed, Pearson would be a accurate function to compute the correlation between that factor and the other.)

Google Play Store Correlation between "Installs" and another factor:

"Reviews" and "Installs":

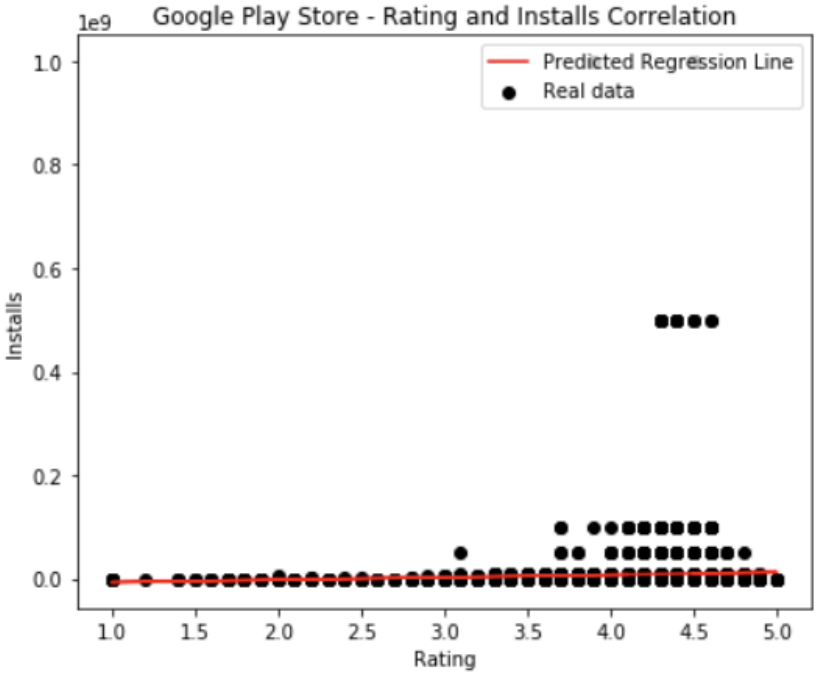
Even though the Spearman Correlation Coefficient gives a high result (about 0.96 with p_value near 0.0), because the graph of "Reviews" and "Installs" does not clearly show that relationship and the Spearman Correlation Coefficient is just a rank correlation, we could not give any conclusion about the correlation of "Reviews" and "Installs" at this time rather than they have a postive correlation.

On the other hand, the shape of the graph and the result of Spearman Correlation Coefficient do have the meanings that they do have a strong relationship, which encourages us to find out in Research Question 2.



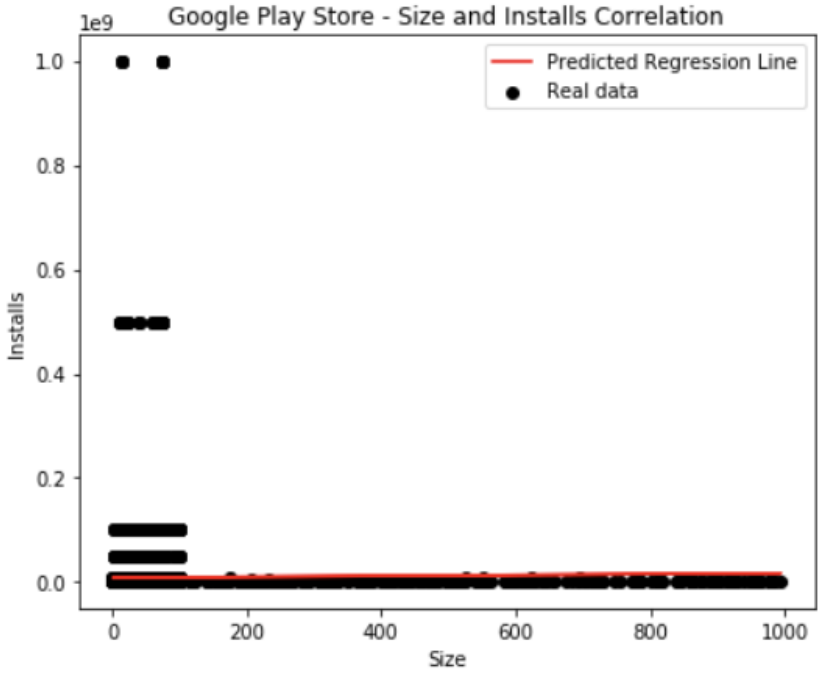
"Rating" and "Installs":

Between 'Rating' and 'Installs' there is a result of Spearman's correlation coefficient is about 0.03 with p-value near 0.007, which shows that they do not have strong correlation. However, based on the result and the graph, we can recognize that 'Rating' and 'Installs' may have a postive relationship.



"Size" and "Installs":

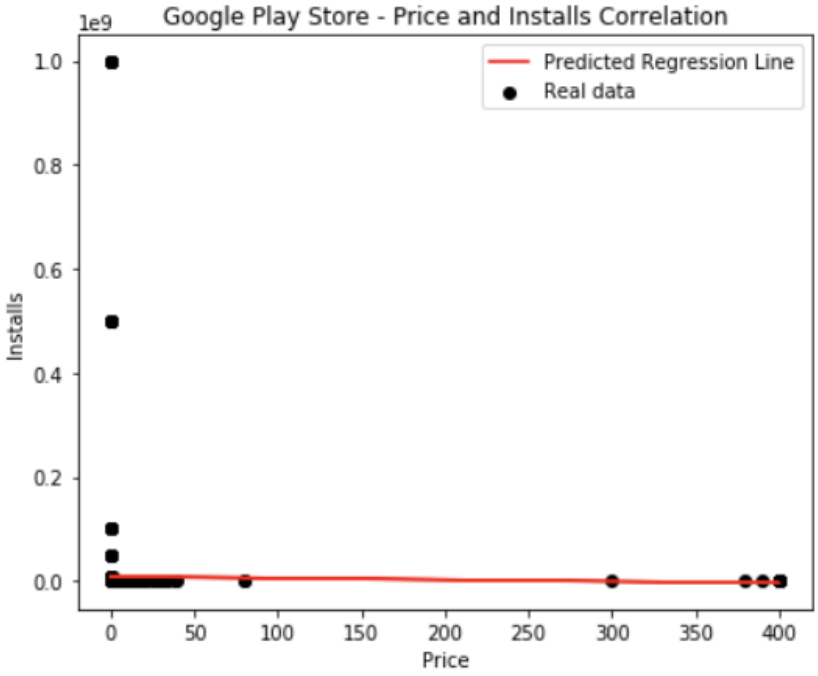
Between 'Size' and 'Installs' there is a result of Spearman's correlation coefficient is about 0.2976 with p-value near 9.94^{-158}, which shows that they do not have strong correlation. However, based on the result and the graph, we can recognize that 'Size' and 'Installs' may have a postive relationship.



"Price" and "Installs":

Between 'Price' and 'Installs' there is a result of Spearman's correlation coefficient is about -0.26

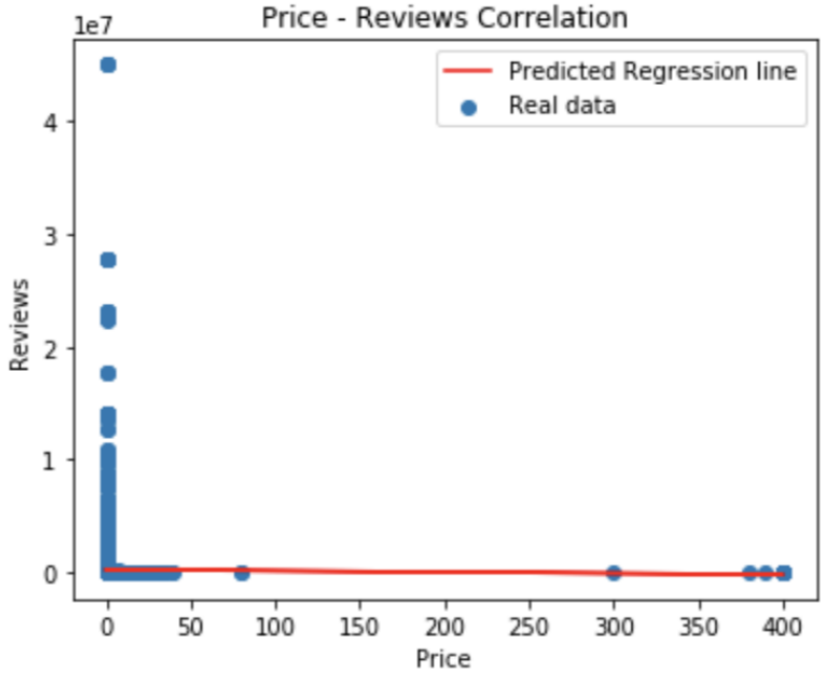
with p-value near 2.9^{-123}, which shows that they do not have strong correlation. However, based on the result and the graph, we can recognize that 'Price' and 'Installs' may have a negative relationship.



Correlation between other pair of factors:

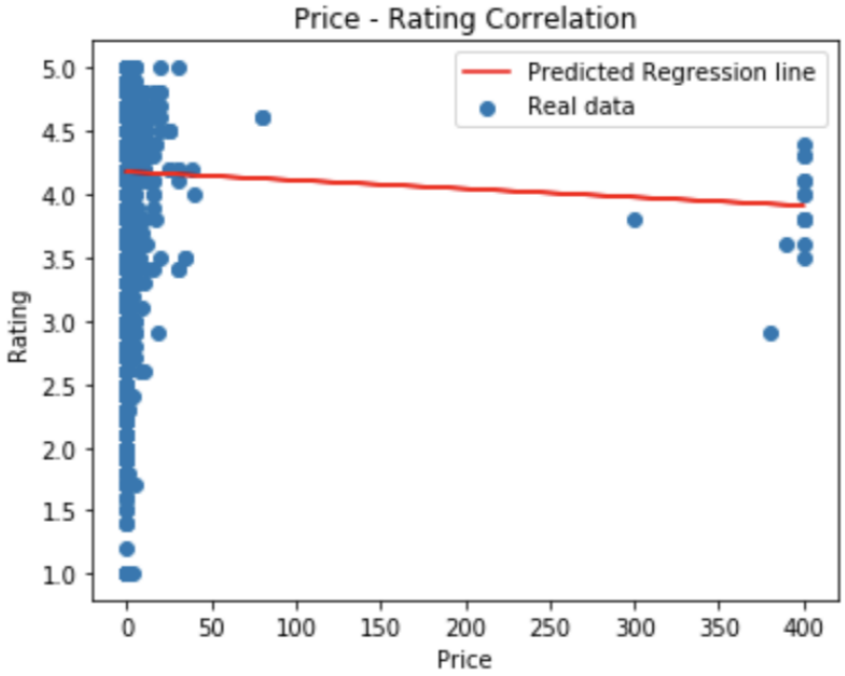
"Price" and "Reviews":

Even though between 'Price' and 'Reviews' there is a result of Pearson's correlation coefficient is about -0.01, the p-value of that result is about 0.37, which is much higher than 0.05. Therefore, as we cannot reject our null hypothesis with this high p-value, we cannot give a conclusion for this by the result above.



"Price" and "Rating":

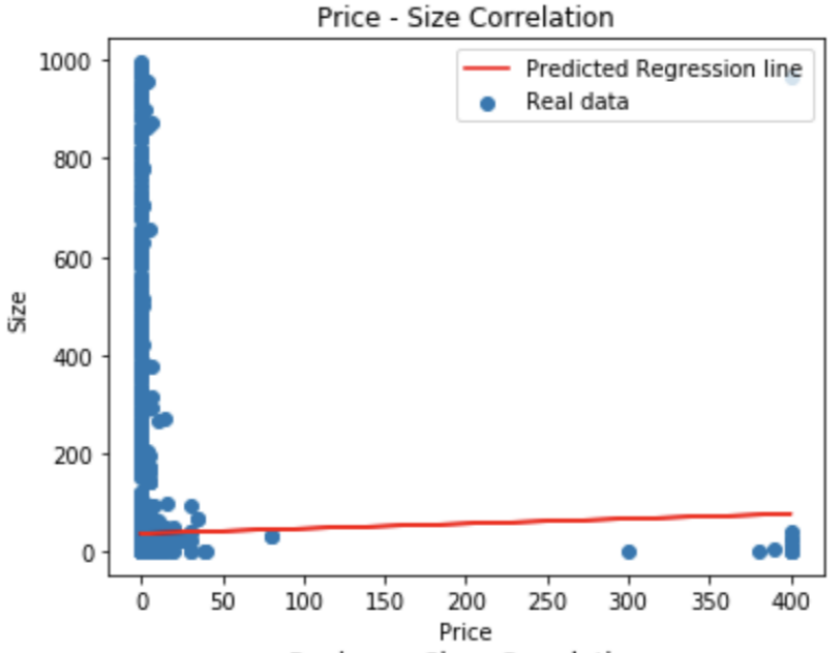
Between 'Price' and 'Rating' there is a result of Pearson's correlation coefficient is about -0.02 with p-value near 0.06 (slightly over 0.05), which shows that they do not have strong correlation. However, based on the result and the graph, we can recognize that 'Price' and 'Rating' may have a negative relationship.



"Price" and "Size":

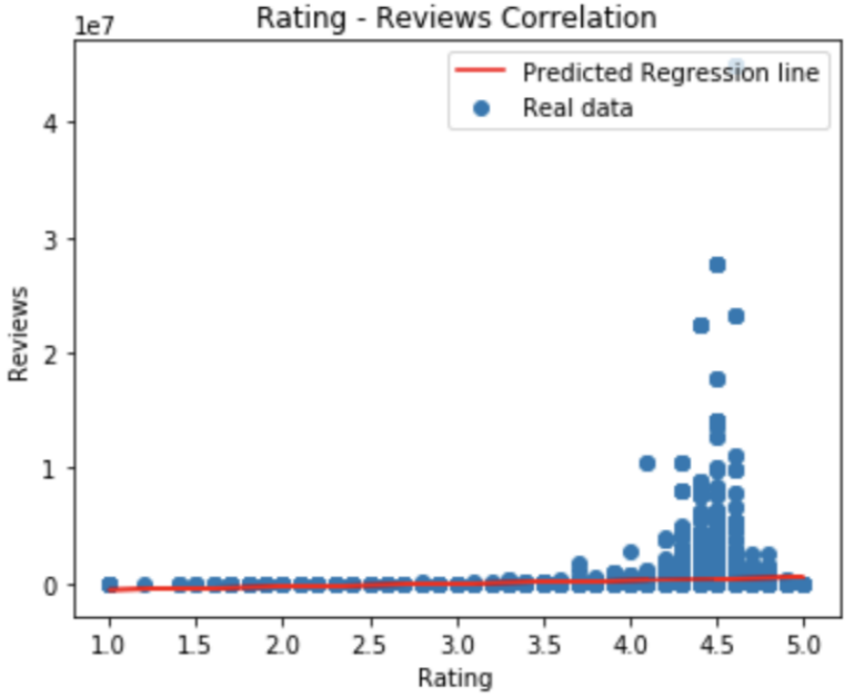
Even though between 'Price' and 'Size' there is a result of Pearson's correlation coefficient is about

0.01, the p-value of that result is about 0.1, which is higher than 0.05. Therefore, as we cannot reject our null hypothesis with this high p-value, we cannot give a conclusion for this by the result above.



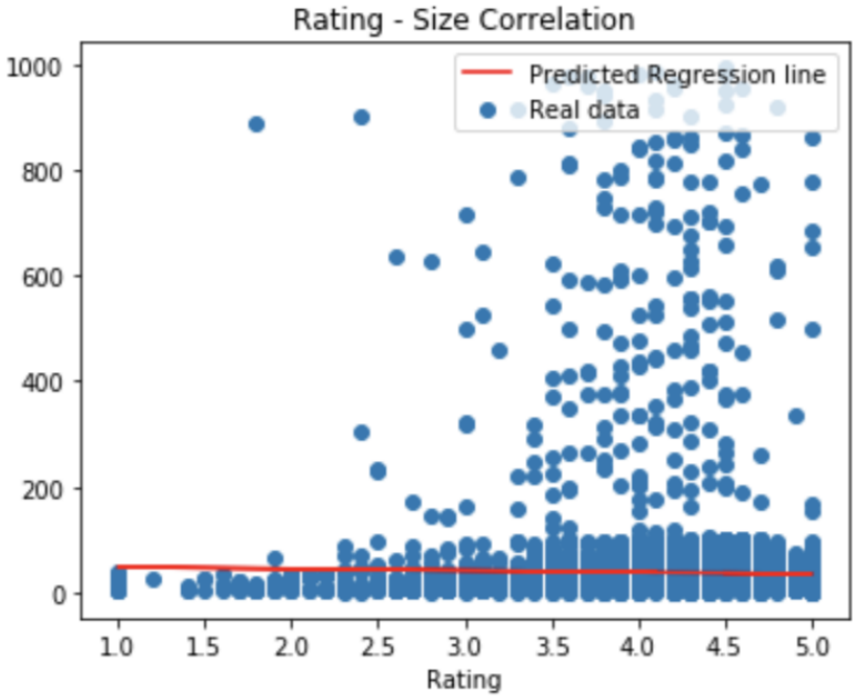
"Rating" and "Reviews":

Between 'Rating' and 'Reviews' there is a result of Pearson's correlation coefficient is about 0.08 with p-value near 2.15^{-12}, which shows that they do not have strong correlation. However, based on the result and the graph, we can recognize that 'Rating' and 'Reviews' may have a postive relationship.



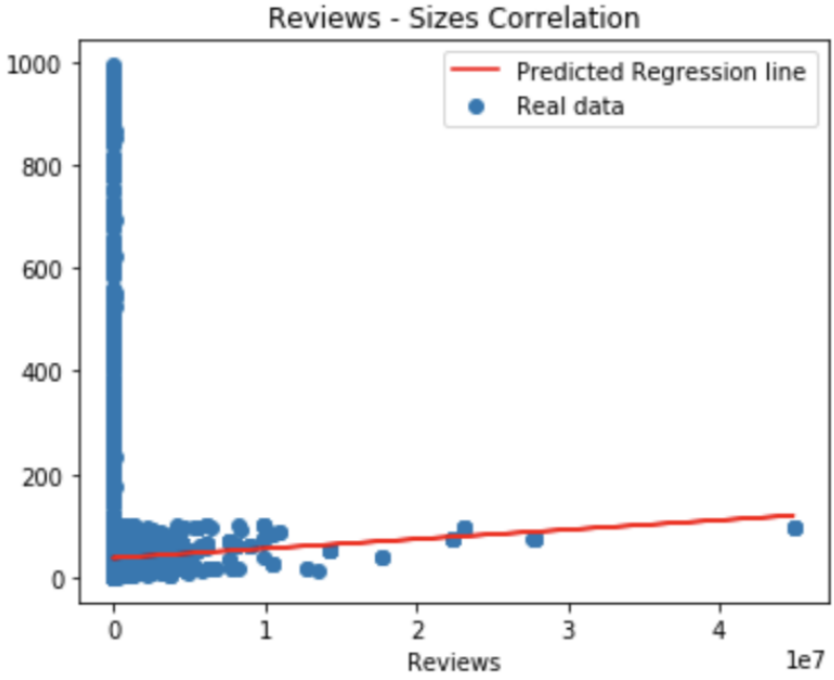
"Rating" and "Size":

Even though between 'Rating' and 'Size' there is a result of Pearson's correlation coefficient is about -0.02, the p-value of that result is about 0.09, which is higher than 0.05. Therefore, as we cannot reject our null hypothesis with this high p-value, we cannot give a conclusion for this by the result above.



"Reviews" and "Size":

Between 'Reviews' and 'Size' there is a result of Pearson's correlation coefficient is about 0.04 with p-value near 0.001, which shows that they do not have strong correlation. However, based on the result and the graph, we can recognize that 'Rating' and 'Reviews' may have a postive relationship.



Overall conclusion and Discussion:

- Most the factors that we have analyzed so far do not show any strong correlation. Up to now, we only able to notice whether each pair of factors have a positive or negative relationship or not. For serval pairs, we could not give any conclusion at this point.
- In this research question, we do not conclude that which factor affects another most because of the fact that these factors have different distributions and have different way to find the correlations.
- Notice that the "Price" factor has not been removed outliers. The reason is that because over 90% of the applications' price are $0.0, remove outliers will remove a lot of important data. However, we can easily understand that because most of the applications in our Google Play Store dataset are free, removing outlier could not improve that much.
- On the other hand, since several factors have dominant range, we can transform these factors for analytic and visualization purpose to find the relationship between the factors.

Question 2 result:

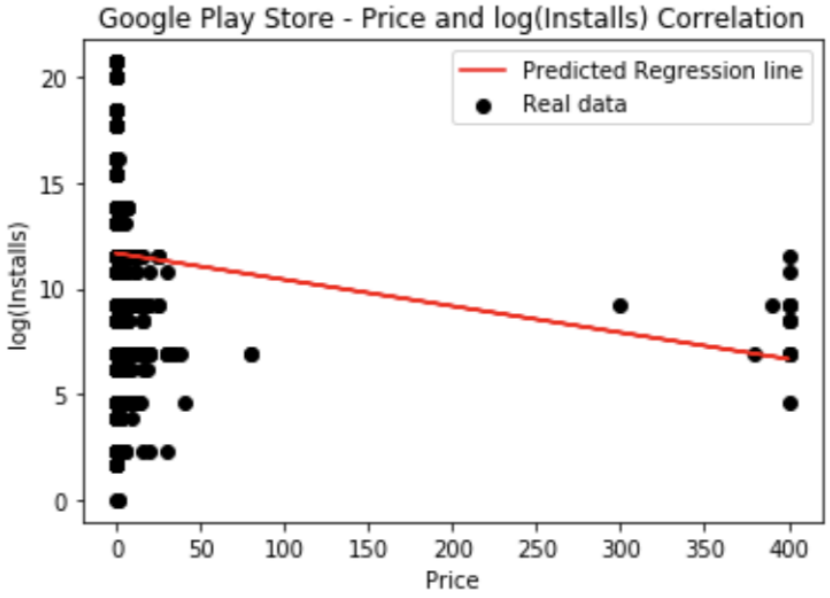
If there is no strong correlation between these factors, are there any better relationships between these factors?

Relationship between Price and Installs:

Conclusion:

As shown above that because most of the application in our Google Play Store dataset are free, we are currently not able to give a specific conclusion about the relationship bewteen these 2 factors rather than recognizing that they have a slightly negative relationship.

From our current analysis, we understand that even by taking logarithm of Install could not have result a correlation further from 0 (in this case the Pearson Correlation Coefficient is about -0.06). Therefore, we can be more confidence about our analysis in Research Question 1 that because most of the price are free, it is hard to find a relationship between Price and other factor if we keep using the current approach.



Discussion:

To understand about the relationship of "Price", it is better to do another analysis that focus especially on different kinds of applications ("Free" and "Paid").

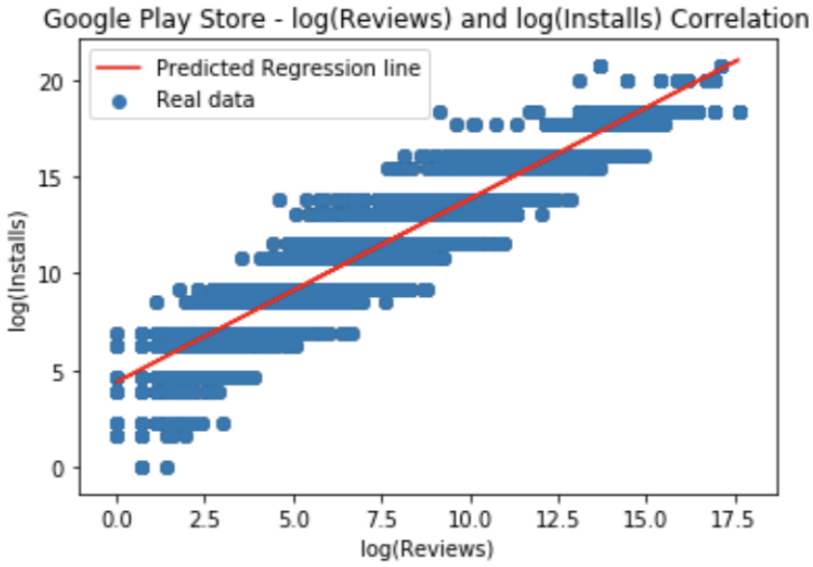
Relationship between Reviews and Installs:

Conclusion:

By taking logarithm of both Reviews and Installs in Google Play Store dataset, we could able to see that the pair log(Reviews) and log(Installs) have a Pearson's correlation coefficient of about 0.95 with p-value near 0.0, which shows that they have a really strong positive correlation.

Plotting the graph also makes we more understand that relationship.

Therefore, we can conclude that Reviews and Installs have a power relationship that log(Reviews) and log(Installs) are strong correlated.



Discussion:

Reflecting back to the Research Question 1, because we have found the strong correlation of log(Reviews) and log(Installs).

Here, we have found the strong correlation of log(Reviews) and log(Installs) with the regression line that has the form: $log(Installs) = a * log(Reviews) + b$, where a and b are constants, which means when we cancel the logarithm in both side, we will get $Installs = Reviews^a * B$, where a and B are constants.

Therefore, we can understand the reason for the high result for Spearmans correlation coefficient and the shape of the graph in Research Question 1.

From this analysis, we understand the fact about Google Play Store that Reviews and Installs do have strong relationship. Therefore, if the publishers want to gain more Installs, Reviews is a important factor that they want to focus on.

Relationship between Rating and Installs, between Rating and Reviews:

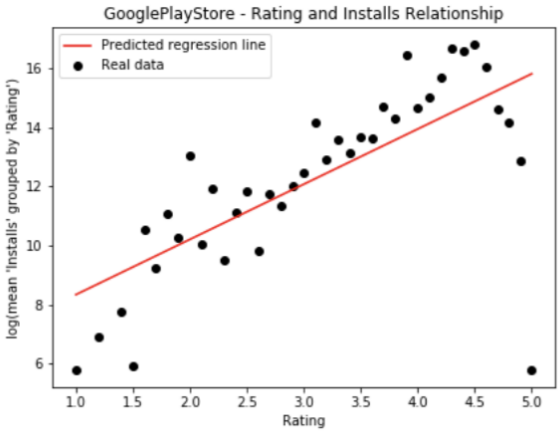
Conclusion:

When we group "Installs" by "Rating" (get the mean of all values of "Installs" that have the same value of "Rating") then take log of that, there is a strong correlation between that value and "Rating" with a high result of Pearson Correlation Coefficient (about 0.7 with p_value near 6.3^{-7})

Doing exactly similar, we also able to conclude that Rating and Reviews have a strongly positive exponential relationship since the Pearson's correlation coefficient between Rating and Reviews is about 0.79 with p-value is about 2.445^{-9}.

Plotting the graphs makes we understand more about the exponential relationships of both pairs.

Therefore, between 2 pairs: Rating and Installs, and Rating and Reviews all have strong postive exponential relationships.



Discussion:

- From this analysis, we understand the fact about Google Play Store that Rating and Installs, and Rating and Reviews do have strong relationship.
- Because we have found the strong correlation with regression line with the form "Rating" = log("mean Installs")$*a + b$, where a and b are constants, which means:

  "Mean Installs" = $10^{("Rating"-b)/a}$. From this, we could understand that, a little change in Rating could have a strong effect Installs.

- Similarly, from the strong correlation with the regression line with the form "Rating" = log("mean Reviews")$*a + b$, where a and b are constants, which means:

  "Mean Reviews" = $10^{("Rating"-b)/a}$, we could also understand that a little change in Rating could have a strong effect on Reviews.

=> Therefore, if the publishers want to gain more Installs or Reviews, Rating is a really important factor that they want to focus on.

Overall Discussion:

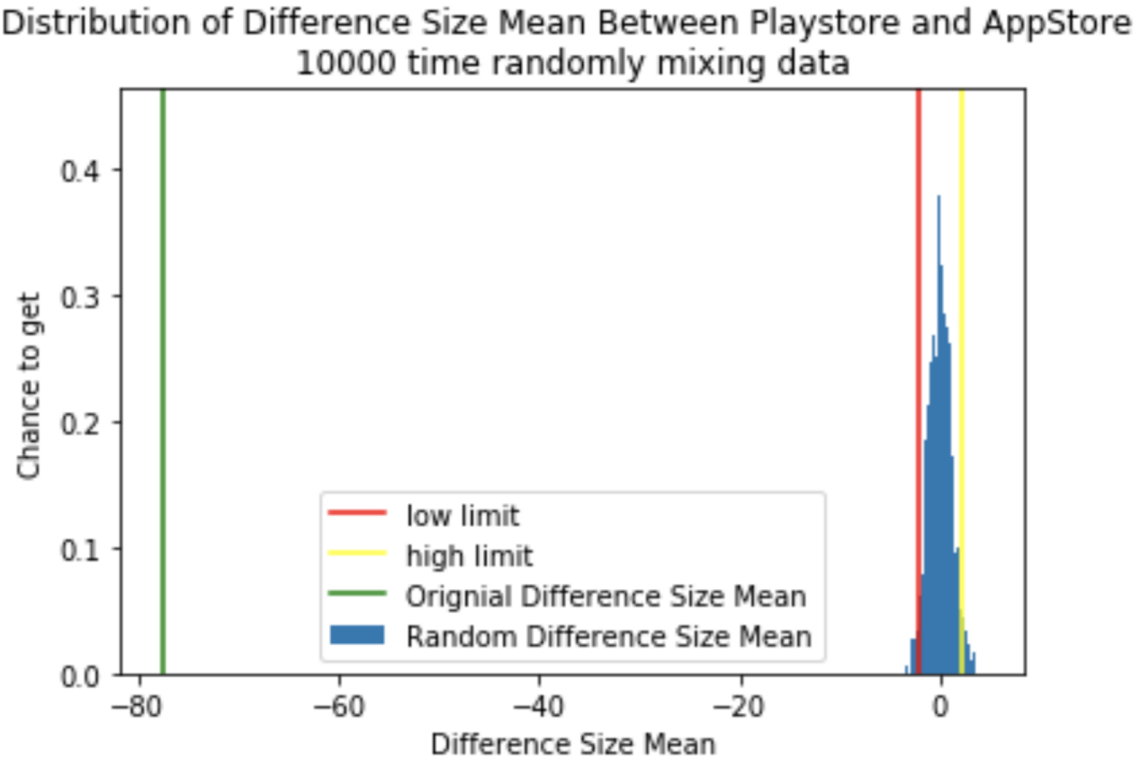
- Here, we do not give a conclusion that which factor affects "Installs" most because we have to use different method to find the relationships between these factors.
- In the future, if we can find a better approach for this problem, we will be able to give a better comparison.

Question 3 result:

Is the difference between the mean size of applications in Google Play Store dataset and size of applications in Appstore dataset is statistically significant?

Conclusion:

By doing hypothesis testing, we can give a conclusion that the observed difference between Applications mean size between Google Play Store and Apple Store, about -77.599 MB, is statistically significant.



Discussion:

From the conclusion above, we understand that most of applications on Google Play Store could be target to serve a more specific purpose compare to Apple Store. Besides, the result helps us strengthen our thought that Google Play Store targets a wider range of devices while Apple Store only targets IOS devices. So applications on that market would have smaller size so as to reach as many devices as possible.

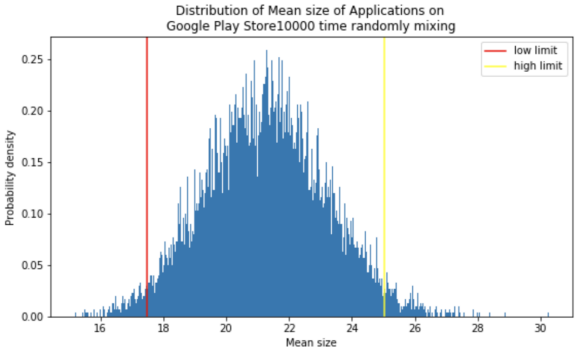
Question 4 result:

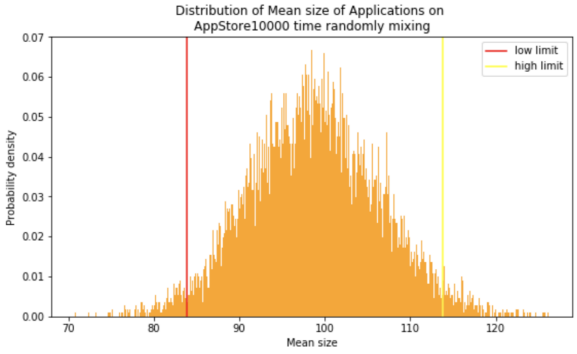
What can we conclude about the mean size of applications on Google Play Store and Apple Store based on the dataset?

Conclusion:

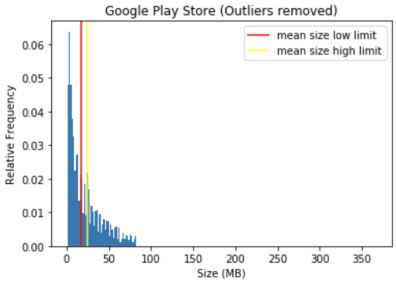
By computation and graphing, we are 95% confidence that:

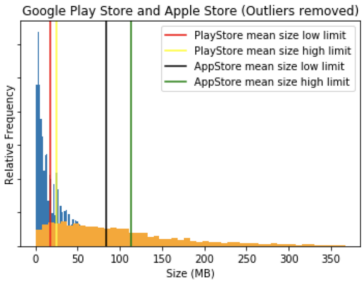
- Google Play Store Applications mean size is from about 17.4 MB to about 25.0 MB.
- Apple Store Applications mean size is from about 84.0 MB to about 113.9 MB.

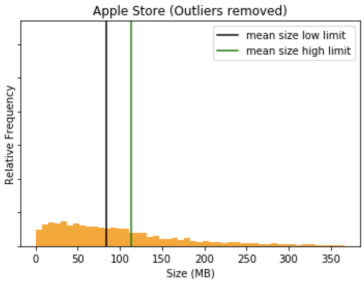




Besides, the distribution of our dataset shows that the most applications on Google Play Store has much smaller size compared with applications on Apple Store.







Discussion:

Understand the mean size of applications on Google Play Store and Apple Store, we understand the trend of applications on the both markets. Therefore, we are much more confidence with our argument above that applications on Google Play Store are normally focus on separated task rather than combine many tasks in one application. Besides, having a low mean size means that applications on Google Play Store can supports weaker devices, which means that the range of targeted audience of the publishers on Google Play Store market could probably bigger than that of publishers on Apple Store.

Reproducing results:

Here (CODE_Google_Play_Store_Analysis_With_AppleStore_Comparison.ipynb) is the source code for the Project for the purpose of reproducing results.

Work plan evaluation:

Part 1: Pre-processing the datasets:

Time estimated: (1 Hour)

PRE. Import Library

1. Import necessarily libraries: pandas, numpy, matplotlib, scipy, and sklearn.

I. Google Play Store dataset (googleplaystore.csv)

1. Use pandas to read the 'googleplaystore.csv' dataset file and store as a DataFrame.
2. Pre-processing the 'googleplaystore.csv' dataset file:
    - Remove rows that include NaN values.
    - 'Size':
        - Remove rows that contains "Varies with device" in 'Size' column.
        - Remove unecessary characters ('k', 'M') in the "Size" column.
    - 'Price':
        - Remove the $ sign in the "Price" column.
    - 'Installs':
        - Remove uneccessary characters (',' and '+') in the 'Installs' column.
3. Store 'Rating', 'Reviews', 'Sizes', 'Installs', 'Price' columns to pandas DataFrame as type float.

II. Apple Store dataset (AppleStore.csv)

1. Use pandas to read the 'ApplseStore.csv' dataset file and store as a DataFrame.
2. Pre-processing the 'AppleStore.csv' dataset file:
    - Remove rows that include NaN values.
    - Convert the size of applications in 'size_bytes' column to MB by dividing that size to 1024^2.
3. Store 'size_bytes' columns after converted to MB and 'user_rating' columns to pandas DataFrame as type float.

Evaluation: For this part, I have to spend about 2 hours because I have to deeply examine different factors in the datasets to deal with (NaN values, special cases in each column,...).

Part 2: Answer Research Questions:

Time estimated: (12 Hours - 36 Hours)

Question 1:

- Plot the graphs to test continuity of each column.
- Use scipy library to compute the Spearman correlation coefficient so as to find the correlation between "Installs" and another factor.
- For other pair of factors, use scipy library to compute the Pearson correlation coefficient.
- Use linear regression from sklearn library to fit the linear regression line for each graph and use matplotlib.pyplot to draw these graphs.

Evaluation: While working on this part, I have met an unexpected results that Reviews and Installs have a high result of Spearman's correlation Coefficient. Therefore, I have to spend more time to read the graph to understand the real reason there is that they seems to have a ranked relationship that even though applications with high number of reviews are reasonably to have high number of Installs, they are not linear correlated. Besides, I have followed correctly as my work plan.

Question 2:

- When finding the relationship between "Price" and "Installs", use np.log to get the log of "Installs". Then use scipy library to compute the Pearson correlation coefficient and draw a graph with a linear regression line to check it.

Evaluation: While working on this part and looking at the result, I was considered that whether should I remove the outliers for the "Price". However, as I understand the fact about the distribution of Price in the dataset, I have resolved my concern about whether removing outlier could make my computation better.

- When finding the relationship between "Reviews" and "Installs", use np.log to get the log of both "Reviews" and "Installs". Then use scipy library to compute the Pearson correlation coefficient and draw a graph with a linear regression line to check it.

Evaluation: The plan for this part is accurate. Therefore, I did not need to do any additional work.

- When finding the relationship between "Rating" and other factors ("Reviews", "Installs"), use DataFrame.groupby to group values of "Reviews" and "Installs" with respect to "Rating". Use "mean()" function to get the mean of each value in "Reviews" and "Installs". Then use np.log to get the log of "Reviews" and "Installs" and use scipy library to compute the Pearson correlation coefficient and draw a graph with a linear regression line to check it.

Evaluation: The plan for this part is accurate. Therefore, I did not need to do any additional work.

Question 3:

- Define a function to remove outlier.
- Define a function to get the confidence interval.
- Remove outliers of the size column.
- Calculate the difference between the mean size of applications in Google Play Store dataset and size of applications in Appstore dataset.
- Loop 10000 times. Each time randomly shuffle some values between the "Size" column in GooglePlayStore and Apple Store dataset then compute the mean of all applications' sizes in each dataset to add to the lists of mean for each dataset.
- Compute the 95% population distribution of the model.
- Plot a graph to show the result of chance to get the original result.

Evaluation: When working on this problem, I have to concern about how to define the function to get the confidence interval and I have to concern about how to define the function to get the remove outliers, which made me do several additional works.

Question 4:

- Loop 10000 times. Each time get 100 random rows in each dataset to compute the mean size of that 100 rows then adds that mean value to the list of means for each dataset.
- Compute the 95% population distribution for the models of mean applications size on 2 datasets.
- Draw 2 histogram (Play Store and Apple Store) to show the model calculated above.

- Draw 3 histograms (Apple Store, Play Store, and both Apple Store and Play Store) to show the distribution of applications' size (x_axis) by relative frequency (y_axis) next to each other (all of them need to be removed outliers). Label x_axis, y_axis as the two factors and set the title to help readability.

Evaluation: the plan for this part is accurate.

Part 3: Testing and Improving the algorithm in the analysis:

Time estimated: (24 Hours - 72 Hours)

- For the data pre-processing part: Check the values of each column by try printing the values out to find all special cases.
- For question 1: Compare the correlation coefficients with the graphs to test the calculation.
- For question 2: Compare the correlation coefficients with the graphs to test the calculation. Besides, try to find out another stronger relationship between different factors.
- For question 3: Test the remove outlier function, Test for the function to get the confidence interval.
- For question 4: Test for lengths of both lists of mean has length 10000 for further computations (Confidence interval).

Evaluation: I was not able to give a conclusion for all analysis in this project (relationships between Price and other factors). However, as I noticed that this should be separate to a new analysis that focus on 2 types of Price (Free and Paid), I have not showed the work for that in this project. Otherwise, the plan for this part is accurate.

Part 4: Gives a conclusion about the datasets

Time Estimated: (2 Hours)

Testing:

Here (CODE_Google_Play_Store_Analysis_With_AppleStore_Comparison.ipynb) (Testing part) is the source code for the Project for the purpose of testing results.

Live Presentation:

Project Presentation Slides (Presentation.pptx)

Collaboration:

I have done this project alone.