**Data Storytelling and Visualization Report**

**Dataset: Google Play Store**

**Name: Yashawant Prabhakar Parab**

**Department: Big Data and Business Analytics**

**Date: 13th September 2019**

Abstract

The aspiration of mining data and analysis of application in google play, the Google’s largest android app store, is to deliver in-depth insight of the data repository to application developer. This approach gives a brief perspective about the current market situation which would help in making appropriate decisions before creating the application and releasing it on the play store. To analyze the market, data such as description of app, Rating, price, installs, version, category, app version, and reviews are collected from app history and stored in a clean structure to further analyze each application. This logical study mainly focuses in measuring pair wise correlation between features thereby clustering the data set into groups. The significant level of pair-wise correlation is identified between certain features from dataset. The cluster analysis is done based on the recommendation system and statistical modeling method.

Table of Contents

[1. Introduction 4](#_Toc19370675)

[2. Motivation 4](#_Toc19370676)

[3. Problem Statement 5](#_Toc19370677)

[4. Literature Review 5](#_Toc19370678)

[5. Proposed Solution 6](#_Toc19370679)

[6. Contribution 6](#_Toc19370680)

[7.Method/Architecture 7](#_Toc19370681)

[7.1 Data cleaning and preparation 8](#_Toc19370682)

[7.2 Visualization techniques suitable for the chosen dataset 9](#_Toc19370683)

[7.3 Justification of the chosen visualization technique 9](#_Toc19370684)

[7.4 Marks and Channels 9](#_Toc19370685)

[7.6 Dashboard design 10](#_Toc19370686)

[8.Result 13](#_Toc19370687)

[9. Evaluation 18](#_Toc19370688)

[10. Discussion and Conclusions 18](#_Toc19370689)

[Appendix 18](#_Toc19370690)

[Bibliography 19](#_Toc19370691)

# Introduction

The information present in app store is categorized into three main categories: Application developer, Business and technical views and customer perspective. Customer’s perspective is considered in the form of Rating, user review, tags and content review. The business view contains application’s statistical and organizational information like category, price, downloads and size of each app. In the end the technical information helps in data mining. The technical information that is necessary to group similar application into same cluster is extracted.

The google play store is found to be largest app market in the world. Also, it has been perceived that downloads done using google play store is double as compared to that of the apple app store but makes only half the money compared to the apple app store. Our dataset is analyzed, and several Bar Charts are created to visualize the relationship between each attribute. Then features such as the number of ratings as well as subjectivity with installs are analyzed. Additionally, an analysis on the features App Name, Content Rating and Type were performed that resulted in some valid conclusion. Using those Bar Charts, the variance of the rating from 1 through 5 was observed. To further investigate, neural network for new game price predication was used. The explanations for all the analysis performed are described with graphical representation.

# Motivation

Smartphone app industry is booming in a significant way and this gives more rivalry among the developing community who create applications. This kind of threat which individual developer face in the form of a greater number of competitions may result in a chance to increase the application value in app store. But even if an app is launched and has reached standard rating together with strong number of downloads, these factors are not enough to retain the position of app forever because there will be many other apps which provide exact same features with better flexibility. Due to expansion and the competition among the market some attributes are chosen to analyze google play store data that resulted in our own defined success metric which would help the developer community to know their success rate and decide future obligatory features which they need to maintain in the next app to be released or version to be released. As per the google play store data some sentiment rating, reviews and price analysis were performed. And in addition to the analysis as per the developer’s point of view, neural net is applied on the price to predict the best fit price of application based on overall dataset.

# Problem Statement

The enlargement of smart phone over the year is driving the fast growth in development application of app store. Currently three global platforms are available; google has their own official google play store for Android Operating system, Microsoft has Microsoft windows phone store and Apple’s app store for iOS Users. But in present-day Google play store and Apple’s app store are the two major platforms who are ruling the development cycle and managing higher number of downloads, users and reviews. Increasing smartphone app development industry has resulted in an increased chance of failure on daily basis. The failure of application impacts the business revenue resulting in a loss in business. For example, Guitar-Tuna app that used to tune guitar, which is present on google play store in Music & Audio category has also several different competitive apps which are used to tune guitar. But over the period with the analysis the developers developed apps which has different instrumental tuning capability and gives basic musical overview to new users as well as the professional users community. The features of google play store were analyzed and it was concluded that developer must understand their app success rate before releasing the app. Our proposed success parameters would help the customer to perform a success rate analysis.

# [Literature Review](#_https://writingcenter.unc.edu/tips-)

In google play store dataset there are many attributes which helps to get better outcome regarding new predicated application. Extract as many features as possible from google play store to get better understanding of market review as well as the upcoming booming feed. To predict success matric of an app, revenue is the key factor which is not found in this dataset. Instead of revenue we used installation and user rating which gives glimpse whether the application is user friendly or is there any need of upgradation. For our analysis we picked Google’s play store and analyze some statistical and predictive modeling mechanism. But as per the problem statement consideration is it necessary to analyze the historical data of play store? Well, the answer Yes, that is because the average of 6410 mobile [apps released per day as per the statistics [1].](#_https://www.statista.com/statistics) The available apps on google play store is around 270000 in 2019 as per the [overview statistics from December 2009 to June 2019 [2].](#_https://www.statista.com/statistics_1)  The application released on daily based has very big competition to the developer to analyze and develop new app with the consideration of all the factor including the size, operating system, upcoming and existing smartphone user interface. Develop and release app without any change on same version gives the best output to the development community. Enormous amount of time, efforts and money are invested in the process of development of application.

# Proposed Solution

The proposed solution is based on a theoretical characterization of feature lifecycles in google play stores that would help app developers to identify trends and to find undiscovered requirements. To investigate and analyze the app features in current app store developers need historical app descriptions information. After analyzing the dataset, we proposed some machine learning technique and neural network mechanism which improve business wealth and developer time. Being aware of which are the intransitive structures in a given group may support developers in recognizing crucial (‘must-have’) requirements for their apps.

We started this research study with the aim achieving results to the following questions:

1. Number of apps available by category.
2. Installation overview according to the type of application.
3. Paid and Free apps on play store
4. Minimum Android version that is needed to run the apps.
5. Genre Split-up according to free and paid.

The most optimized method to get solutions to above mentioned problems is by using machine learning algorithms.

1. Predicted price of paid app
2. Correlation between Reviews and Rating using R.
3. Correlation between Reviews and Rating using Microsoft Excel.

# Contribution

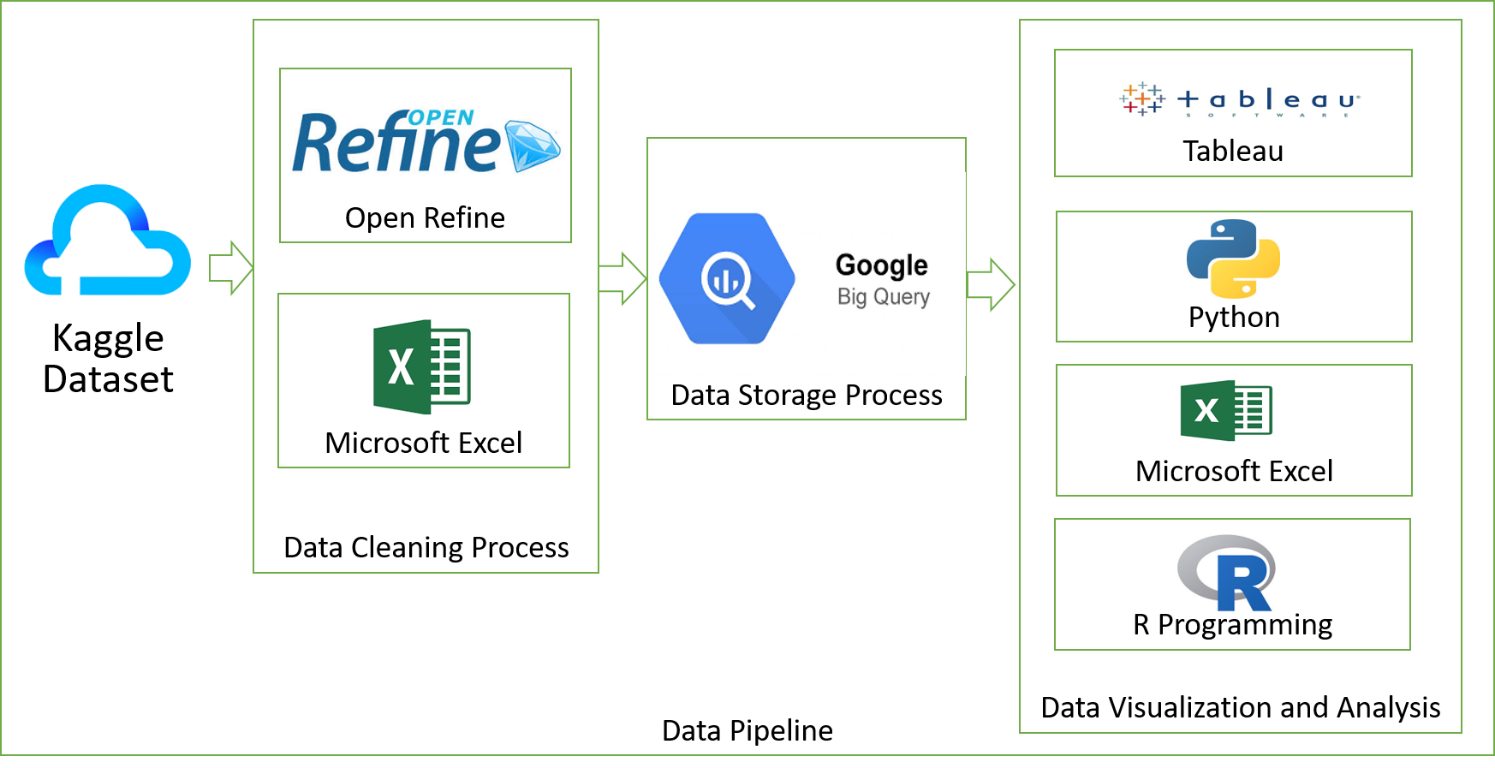
Previous research has examined the users reviews and rating which helps developer to know the market situation. This research has typically shown that consumers correlation between the application genres and their install rates. In contrast, the current research is the first attempt to study the expected cost of application using the machine learning technique. Used Rectified linear unit algorithm with 50 epochs to get better price prediction.

In contrast to what one might expect, getting ready for a future task is demonstrated to increase rather than decrease effort expenditure on current tasks because the resources for future task demands may inadvertently carry over to unrelated tasks in the present.

# 7.Method/Architecture

The Architecture of conducting this research is defined as per the users need as well as from developer’s point of view. The architecture embodies the components which can be active example: Downloads, rating, reviews and some are passive include category, price and compatible android version. We perform some query on google big query for visualization and data verification.

To perform operations on dataset we use some data cleaning, data preparation, and data visualization tools. The play store data pipeline shown below;



Google Data Pipeline

We used following tools for better productivity:

|  |  |
| --- | --- |
| Process | Tools |
| Data Cleaning | 1. OpenRefine 2. Microsoft Excel |
| Data Storage | [Google Big Query](#_https://console.cloud.google.com/bi) |
| Data Visualization and Dashboard | 1. Tableau 2. Microsoft Excel |
| Analysis | Python (Jupyter Notebook)  R Programming |

## 7.1 Data cleaning and preparation

The purpose of our project was to gather and analyze detailed information on apps in the Google Play Store in order to provide insights on app features and the current state of the Android app market. Google dataset has 32k data which include 11 different attributes which holds excessive amount of information with respective to the google play store app. We perform some cleaning process on data and delete some unused data. We used App, Category, Rating, Reviews, Size, Price, Genres, and Current version attributes as a key factor.

In addition, we embrace 2 attributes called as Android Version Name and Type of sale. Android Version has all the android version mapping with the defined name and Type of sale contain the sale type of the application which is Free or Paid.

Attributes and Description:

|  |  |
| --- | --- |
| Attributes | Description |
| App | Application name |
| Category | Category the app belongs to |
| Rating | Overall user rating of the app (as when scraped) |
| Reviews | Number of user reviews for the app (as when scraped) |
| Size | Size of the app (as when scraped) |
| Installs | Number of user downloads/installs for the app (as when scraped) |
| Price | Price of the app (as when scraped) |
| Content Rating | Age group the app is targeted at - Children / Mature 21+ / Adult |
| Genres | An app can belong to multiple genres (apart from its main category). For e.g. a musical family game will belong to Music, Game, Family Genres. |
| Last Updated | Date when the app was last updated on Play Store (as when scraped) |
| Current Ver | Current version of the app available on Play Store (as when scraped) |
| Android Ver | Min required Android version (as when scraped) |
| Type | Paid or Free |

## 7.2 [Visualization techniques](#_https://towardsdatascience.com/big-) suitable for the chosen dataset

Data visualization means taking information and placing it into a visual context, such as a map or graph. Data visualizations helps human brain to understand by visualizing which also makes it easier to detect patterns, trends, and outliers in groups of data.

[Following are the few techniques we used on this Dataset](#_https://www.searchenginejournal.com):

1. Pie Chart
2. Heat Map
3. Bar Graph
4. Scatter Plot

We perform complex algorithms which are much easier to understand in a visual format as opposed to lines and lines of text and numbers.

## 7.3 Justification of the chosen visualization technique

To find out which technique is appropriate for use cases we delve deep down to the big data visualization technique.

1. Pie Chart: Pie chart split into sectors illustrating with some numerical values, Labels and angle with same proportional to get value representation in definite method. We use pie chart for getting value distribution based on Type of app in a circular statistical graphic manner.
2. Heat map: To get a matrix representation with worm-to-color spectrum we use heat map. To get the visualization of paid app with respect to the genre we use heat map.
3. Bar Graph: Perform series operation on installation overview, split-up of free and paid apps using genre we use bar graph chart.
4. Scatter plot: instead of Line chart we analyze some feature using scatter plot. Scatter plot mainly used for finding out correlation between two attributes.

## 7.4 Marks and Channels

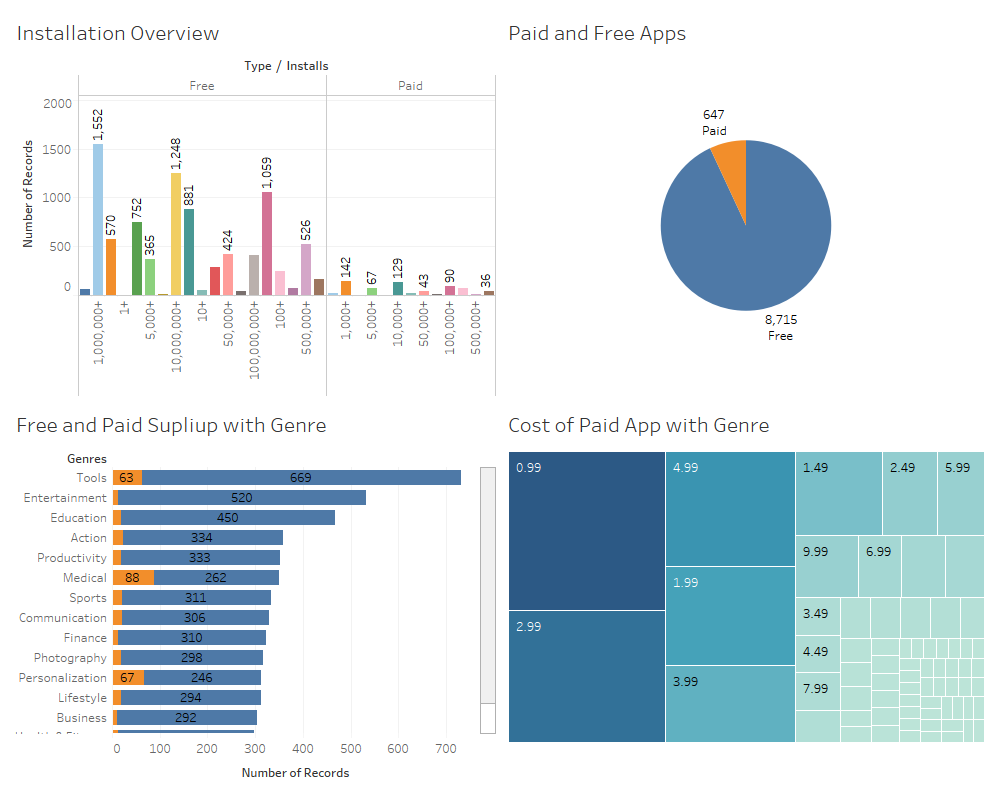
|  |  |
| --- | --- |
| Marks | Channels |
| Points  By using Microsoft Excel, we perform one relation between rating and reviews. In this we plot scatter plot and used points marks | Color  We used channels based on attribute appearance. We use silent color which suits with google official logo |
|  | Size  We used decent size of chart, Font and length which easily visible to human eye |

## 7.6 Dashboard design

Dashboard 1

Overall Dashboard

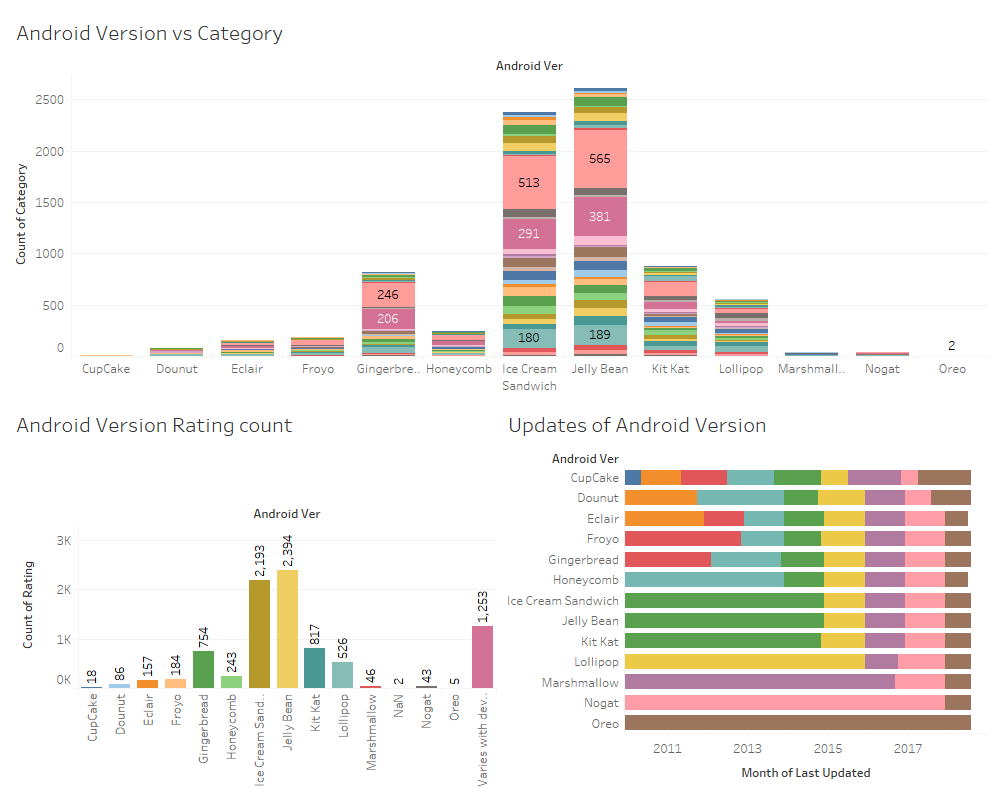
Dashboard Shows Overall preview of attributes which has Pair and free app distribution, Number of records that has been downloaded by users, Cost of application as per the genre, and Free and paid app split-up with respect to Genre.



Dashboard 2

Android Dashboard

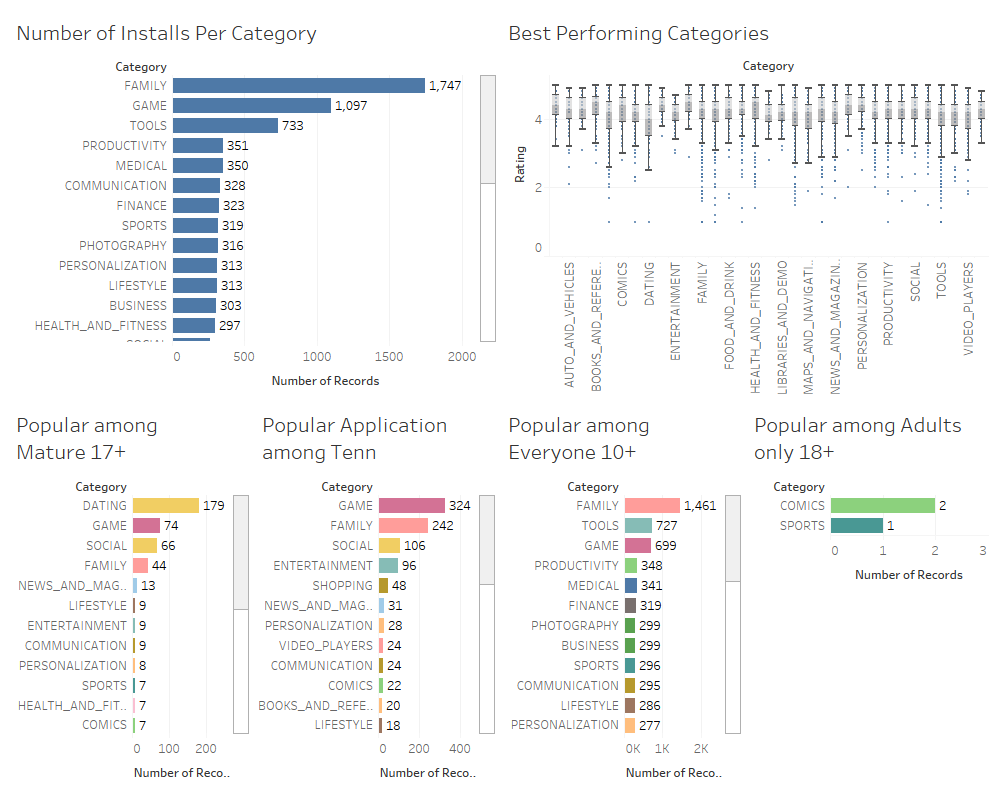
From 2008 to till date there are 14 operating system which google has introduced shown in below dashboard. With the update version attribute, we create android version histogram graph which indicates the year wise updates of version. There must be a correlation between rating and android version which shown in bar graph.



Dashboard 3

Application Installation Dashboard

Number of installs per category, best performing category with respect to installs rate and popularity of applications among mature, adults and teen separated by the bar and box plot.

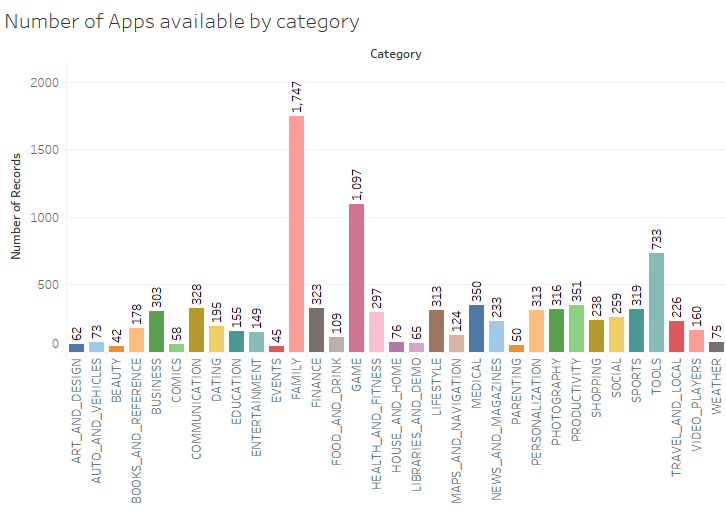


# 8.Result

**Exhibit: 1) Number of apps available by category.**

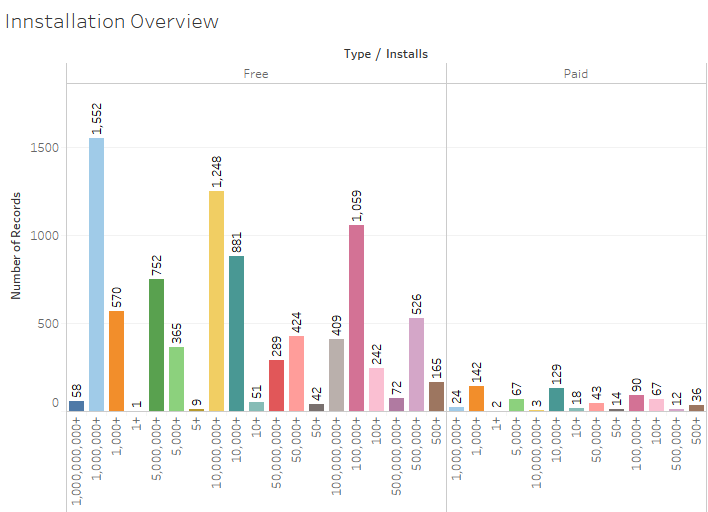
We analyze number of apps present in play store by their respective category using bar graph. As per the bar graph the family category has highest number of apps available in play store. Games including all types of genres and tool attributes present all types of utilities app like google translate, File manager, Share it etc.

Use of bar graph gives better visualization representation as per the category.



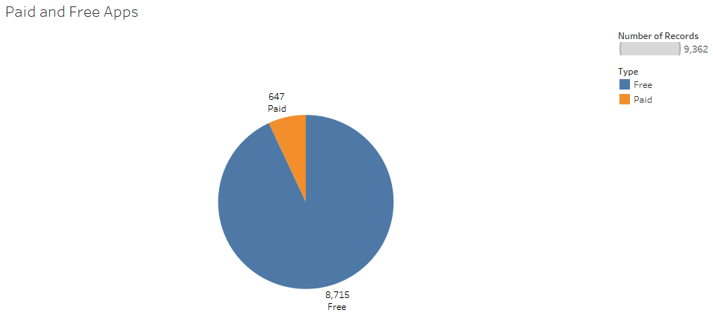
**Exhibit: 2) Installation overview according to the type of application.**

In this, sum of all the records of all installs are broken down by the type whether it is free or paid. The Google Play Store offers a wide range of applications, but most of these apps had been downloaded only by a small number of people. In our dataset, 93% of the apps are marked as free and about 7% of the apps are paid.



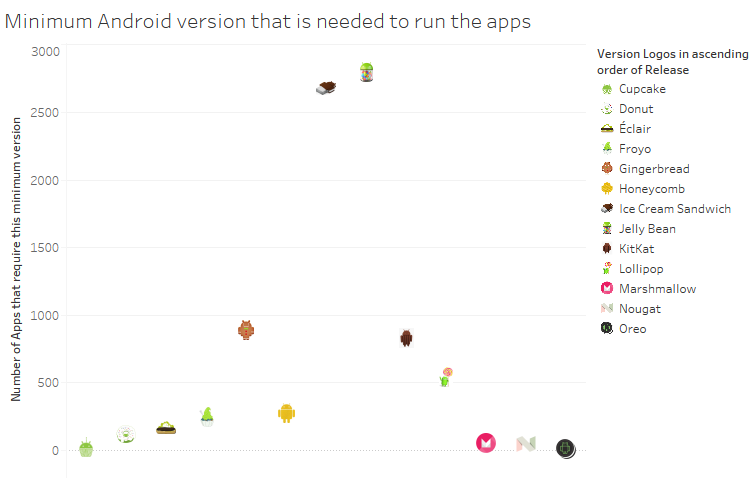
**Exhibit: 3) Paid and Free apps on play store.**

In our dataset more than 92% of free apps in our dataset were free in contrast to the paid apps. From the records less than 8% apps where most downloaded in paid type.



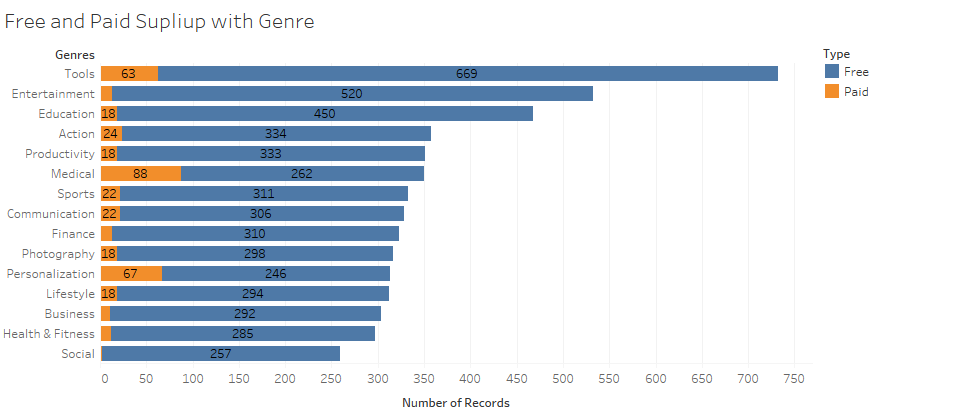
**Exhibit: 4) Minimum Android version that is needed to run the apps.**

From 2008 to till date there are [13 different version of operating system](#_https://en.wikipedia.org/wiki/Andro) available for different smartphone with respect to the specification of phone. Over the time operating system evolved their features and specification. We separate number of apps with minimum operation system version requirement. Most of the applications need Ice-creme sandwich and jellybean operating system to run app seamlessly.



**Exhibit: 5) Genre Split-up according to free and paid.**

As per the previous analysis we have split all the apps based on whether it is free or paid app. Now as an application developer it is important to know which genre is most famous and user friendly in which users would happily invest. We consider top 15 genre for this study. As per the analysis Personalization, medical, and tools genres has most paid users.



**Exhibit: 5) Predicted price of paid app**

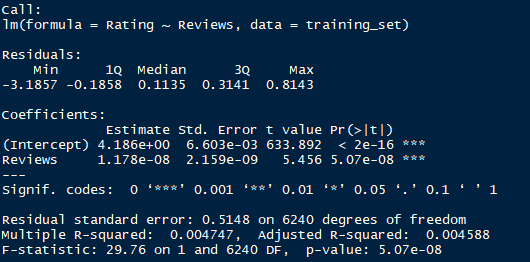
We use price attribute separately and [applied neural network to predict price of application](#_https://missinglink.ai/guides/neura).

Activation Function: We used ReLU (Rectified liner Unit) activation function which is responsible for transforming the summed weighted input from the price node into the activation of the output for that input.

For better prediction we run the neural net at 50 epochs and as per the rectified linear unit we come up with 4.6968 expected application cost.

**Exhibit: 6) Correlation between Reviews and Rating using R.**

To find the best fit between rating and reviews we used liner regression. The data was split into training and test data and a R inbuild function called lm was used. Lm function is defined for linear regression and the result shows all the residuals including Min, Max and median. Also, this function calculates alpha and beta value which helps to find perfect regression points on scatter plot. The regression has best fit as considering the p value which is less than 0.05. R squared and residual standard error shown below in R statistics.



**Exhibit: 7) Correlation between Reviews and Rating using Microsoft Excel.**

We calculate Lm (Liner regression) function in Excel by applying formula. After manipulating the data according to the test and train, alpha, beta and R² values were calculated.

Formula to estimate all equation

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Alpha, Beta and R squared Formula | | | | | | | | |
|  |  | | | | | |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  | |  | | --- | |  | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

# 9. [Evaluation](#_https://www.guidelines.kaowarsom.be)

This study is concerned with reviews: user-submitted applications evaluations that association an ordinal rating and text. The availability of reviews presents a task to researchers working on google play Store Review Mining and Analysis, because it distresses generalizability. For instance, the research with review mining and analysis has analyzed with the sentiments. As a way of exploring the issue empirically, we assess the level of representation an app review subset can provide.

# 10. Discussion and Conclusions

We show that by appropriate data reduction of user reviews to a subset of user requests, we can learn important results through correlation analysis. In contrast, we find a strong linear regression (correlation) between the Reviews and the ratings of apps. As per the R Statistics we came to know the p value is less than the expected which shows the regression has best fit. We build a neural network using ReLU for predicting and identify prevalent price in apps.

# Appendix



# Bibliography

<https://www.statista.com/statistics/276703/android-app-releases-worldwide/>

<https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/>

<https://writingcenter.unc.edu/tips-and-tools/literature-reviews/>

<https://en.wikipedia.org/wiki/Android_version_history>

<https://www.guidelines.kaowarsom.be/evaluating_scientific_papers>

<https://www.searchenginejournal.com/what-is-data-visualization-why-important-seo/288127/#close>

<https://towardsdatascience.com/big-data-information-visualization-techniques-f29150dea190>

<https://missinglink.ai/guides/neural-network-concepts/7-types-neural-network-activation-functions-right/>

<https://www.youtube.com/watch?v=vhrmmvAvOfQ>

<https://console.cloud.google.com/bigquery?project=datastory-252220&p=datastory-252220&page=project>

<https://en.wikipedia.org/wiki/Android_(operating_system)>