**Table of Content**

1. Introduction

2. Scope of the Analysis

3. Existing System

1. Drawbacks or limitations of existing system

4. Source of dataset

5. ETL process

6. Analysis on dataset (for each analysis)

1. Introduction
2. General Description
3. Specific Requirements, functions and formulas
4. Analysis results
5. Visualization

7. List of Analysis with results

8. Future scope

9. References

10. Bibliography

**INTRODUCTION TO DATA MANAGEMENT**

**PROJECT REPORT**

(Project Semester August-December 2018)

**GOOGLE PLAY STORE APPS**

Submitted by

Satya Kireet

Registration No: 11602210

Programme: B. Tech-CSE

Section: KEM39

Course Code: INT 217

Under the Guidance of

**Mrs. Savleen Kaur**

**Discipline of CSE/IT**

**Lovely School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Satya Kireet bearing Registration no 11602210 has completed INT217 project titled, **“GOOGLE PLAY STORE APPS.”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science & Engineering**

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I hereby declare that the project work entitled “GOOGLE PLAY STORE APPS” submitted to the LOVELY PROFESSIONAL UNIVERSITY, Phagwara is a record of an original work done by me under the guidance of Mrs. Savleen Kaur. I Satya Kireet hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Satya Kireet

Registration No: 11602210

Date: -11-2018

**ACKNOWLEDGEMENT**

I would like to express my special thanks of gratitude to my teacher Mrs. Savleen Kaur who gave me the golden opportunity to do this wonderful project on the topic “GOOGLE PLAY STORE APPS” which also helped me in doing a lot of Research and i came to know about so many new things I am really thankful to them.Secondly is would also like to thank my parents and friends who helped me a lot in finalizing this project within the limited time frame.

**INTRODUCTION**

**INTRODUCTION TO DATA SCIENCE**

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, like data mining.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.

The term "data science" has appeared in various contexts over the past thirty years but did not become an established term until recently. In an early usage, it was used as a substitute for computer science by Peter Naur in 1960. Naur later introduced the term "data logy". In 1974, Naur published Concise Survey of Computer Methods, which freely used the term data science in its survey of the contemporary data processing methods that are used in a wide range of applications.

In 1996, members of the International Federation of Classification Societies (IFCS) met in Kobe for their biennial conference. Here, for the first time, the term data science is included in the title of the conference ("Data Science, classification, and related methods"),after the term was introduced in a roundtable discussion by Chikio Hayashi.

In November 1997, C.F. Jeff Wu gave the inaugural lecture entitled "Statistics = Data Science?" for his appointment to the H. C. Carver Professorship at the University of Michigan. In this lecture, he initiated the modern, non-computer science, usage of the term "data science" and advocated that statistics be renamed data science and statisticians data scientists. Later, he presented his lecture entitled "Statistics = Data Science?" as the first of his 1998 P.C. Mahalanobis Memorial Lectures. These lectures honor Prasanta Chandra Mahalanobis, an Indian scientist and statistician and founder of the Indian Statistical Institute.

In 2001, William S. Cleveland introduced data science as an independent discipline, extending the field of statistics to incorporate "advances in computing with data" in his article "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics," which was published in Volume 69, No. 1, of the April 2001 edition of the International Statistical Review / Revue Internationale de Statistique. In his report, Cleveland establishes six technical areas which he believed to encompass the field of data science: multidisciplinary investigations, models and methods for data, computing with data, pedagogy, tool evaluation, and theory.

**INTRODUCTION TO EXCEL**

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications. It has been a very widely applied spreadsheet for these platforms, especially since version 5 in 1993, and it has replaced Lotus 1-2-3 as the industry standard for spreadsheets. Excel forms part of Microsoft Office.

Microsoft Excel has the basic features of all spreadsheets, using a grid of cells arranged in numbered rows and letter-named columns to organize data manipulations like arithmetic operations. It has a battery of supplied functions to answer statistical, engineering and financial needs. In addition, it can display data as line graphs, histograms and charts, and with a very limited three-dimensional graphical display. It allows sectioning of data to view its dependencies on various factors for different perspectives (using pivot tables and the scenario manager). It has a programming aspect, Visual Basic for Applications, allowing the user to employ a wide variety of numerical methods, for example, for solving differential equations of mathematical physics, and then reporting the results back to the spreadsheet.

In a more elaborate realization, an Excel application can automatically poll external databases and measuring instruments using an update schedule, analyze the results, make a Word report or PowerPoint slide show, and e-mail these presentations on a regular basis to a list of participants. Excel was not designed to be used as a database.

Microsoft allows for a number of optional command-line switches to control the manner in which Excel starts.

**INTRODUCTION TO THE PROJECT**

**Google Play** (previously **Android Market**) is a digital distribution service operated and developed by Google. It serves as the official app store for the Android operating system, allowing users to browse and download applications developed with the Android software development kit (SDK) and published through Google. Google Play also serves as a digital media store, offering music, magazines, books, movies, and television programs. It previously offered Google hardware devices for purchase until the introduction of a separate online hardware retailer, Google Store, on March 11, 2015.

Applications are available through Google Play either free of charge or at a cost. They can be downloaded directly on an Android device through the Play Store mobile app or by deploying the application to a device from the Google Play website. Applications exploiting hardware capabilities of a device can be targeted to users of devices with specific hardware components, such as a motion sensor (for motion-dependent games) or a front-facing camera (for online video calling). The Google Play store had over 82 billion app downloads in 2016 and has reached over 3.5 million apps published in 2017.[[4]](https://en.wikipedia.org/wiki/Google_Play#cite_note-3.5_million_apps-4) It has been the subject of multiple issues concerning security, in which malicious software has been approved and uploaded to the store and downloaded by users, with varying degrees of severity.

**Context**

While many public datasets (on Kaggle and the like) provide Apple App Store data, there are not many counterpart datasets available for Google Play Store apps anywhere on the web. On digging deeper, I found out that iTunes App Store page deploys a nicely indexed appendix-like structure to allow for simple and easy web scraping. On the other hand, Google Play Store uses sophisticated modern-day techniques (like dynamic page load) using J Query making scraping more challenging.

**Content**

Each app (row) has values for category, rating, size, and more.

**Acknowledgements**

This information is scraped from the Google Play Store. This app information would not be available without it.

**Inspiration**

The Play Store apps data has enormous potential to drive app-making businesses to success. Actionable insights can be drawn for developers to work on and capture the Android market!

**Content of google play store raw data**

Googleplaystore.csv

* App: Application name
* Category: Category the app belongs to
* Rating: Overall user rating of the app
* Reviews: Number of user reviews for the app
* Size: Size of the app
* Installs: Number of user downloads/installs for the app
* Type: Paid or free
* Price: Price of the app
* 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)
* Last Updated: Date when the app was last updated on Play Store
* Current Version: Current Version of the app available on Play Store
* Android Version: Minimum required android version

**SCOPE OF THE ANALYSIS**

The Play Store data taken from the source contained all the information that is required for the project scope of analysis. But the data contained lot of duplicate values and many blank places at most of the places. All the data was analyzed and the duplicate values were removed and the blank places were filled with the appropriate data that suits the work. Also, replacements were done at some points where the data was inadequate.

“Ctrl + G” and “Ctrl + F” are the operations used to analyze data in a keen perspective. There is a new feature in excel tools i.e. **KUTOOLS** downloaded from google. It is very helpful for the analysis of the data.

The Play Store app data is analyzed and came up with the following:

* Users try to download good apps with high user rating and with reasonable price. Good apps are recognized by good rating count.
* Memory is a constraint for some of the mobile users, so best among the apps(by rating) and memory size is used for analysis.
* Users will get to know the list of applications that are working bad, average, good and excellent. App developers can also analyze the applications and based on the user reviews they can rectify the mistakes.
* If the average, maximum and minimum price is known then it has a great value for the new developers to conduct feasibility study.
* If the minimum rating and its details are known, then the app developers will try to improve its quality.

**EXISTING SYSTEM**

**DRAWBACKS OR LIMITATIONS OF EXISTING SYSTEM**

The existing data doesn’t provide the following techniques that are required for the analysis

* FILTERING: The existing system does not allow to select the required data by using selection criteria’s.
* GROUPING: The existing does not have the capabilities of pivot table to group the data and report.
* MATHEMATICAL CALCULATIONS: The existing system does not have logical conditioning and mathematical calculations such as average, maximum, minimum.
* SORTING: Sorting allows the apps to be sorted by category or cost.

**SOURCE OF DATASET**

The dataset is taken from Kaggle. Kaggle is a community of data scientists and data enthusiasts. This platform enables you to learn from and mentor each other on your personal, academic, and professional data science journeys.

Kaggle is an online community of data scientists and machine learners, owned by Google, Inc. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. Kaggle got its start by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and short form AI education. On 8 March 2017, Google announced that they were acquiring Kaggle.

The community spans 194 countries. It is the largest and most diverse data community in the world, ranging from those just starting out to many of the world's best known researchers.

Kaggle competitions regularly attract over a thousand teams and individuals. Kaggle's community has thousands of public datasets

With millions of apps around nowadays, the following data set has become very key to getting top trending apps in Google app store. This data set contains more than 7000 Play Store app details. The data was extracted from the Google Play Store website. R and Linux web scraping tools were used for this study.

**ETL PROCESS**

In computing, extract, transform, load (ETL) is a process in database usage to prepare data for analysis, especially in data warehousing. The ETL process became a popular concept in the 1970s. Data extraction involves extracting data from homogeneous or heterogeneous sources, while data transformation processes data by transforming them into a proper storage format/structure for the purposes of querying and analysis; finally, data loading describes the insertion of data into the final target database such as an operational data store, a data mart, or a data warehouse. A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format so that application developers can build applications and end users can make decisions.

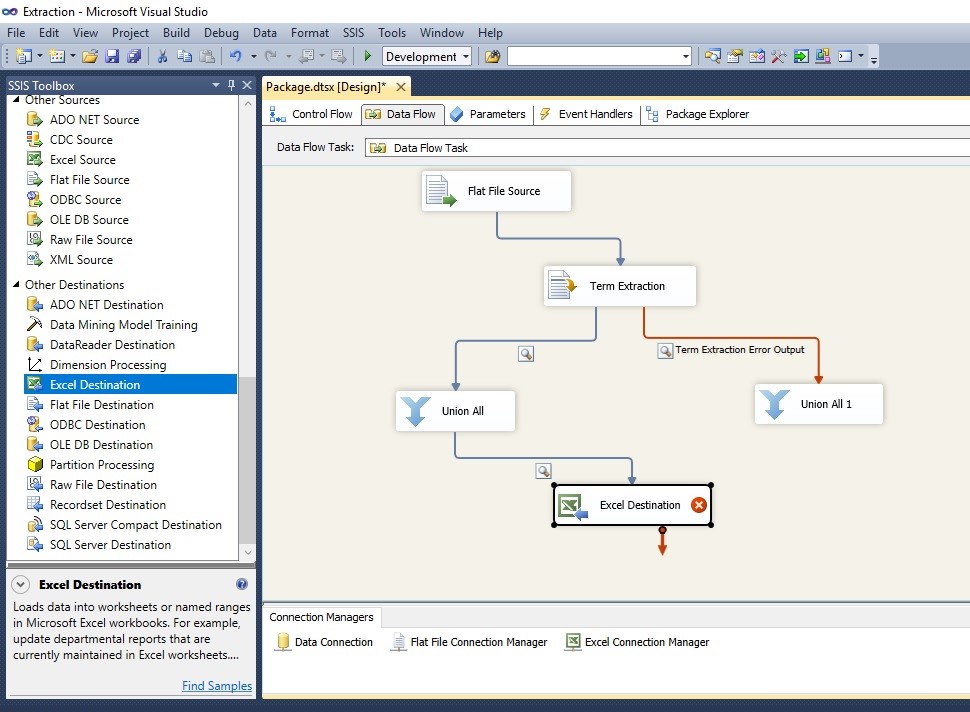
Since the data extraction takes time, it is common to execute the three phases in parallel. While the data is being extracted, another transformation process executes while processing the data already received and prepares it for loading while the data loading begins without waiting for the completion of the previous phases.

ETL systems commonly integrate data from multiple applications (systems), typically developed and supported by different vendors or hosted on separate computer hardware. The separate systems containing the original data are frequently managed and operated by different employees. For example, a cost accounting system may combine data from payroll, sales, and purchasing.

**Extract**

The first part of an ETL process involves extracting the data from the source system(s). In many cases, this represents the most important aspect of ETL, since extracting data correctly sets the stage for the success of subsequent processes. Most data-warehousing projects combine data from different source systems. Each separate system may also use a different data organization and/or format. Common data-source formats include relational databases, XML, JSON and flat files, but may also include non-relational database structures such as Information Management System (IMS) or other data structures such as Virtual Storage Access Method (VSAM) or Indexed Sequential Access Method (ISAM), or even formats fetched from outside sources by means such as web spidering or screen-scraping. The streaming of the extracted data source and loading on-the-fly to the destination database is another way of performing ETL when no intermediate data storage is required. In general, the extraction phase aims to convert the data into a single format appropriate for transformation processing.

An intrinsic part of the extraction involves data validation to confirm whether the data pulled from the sources has the correct/expected values in a given domain (such as a pattern/default or list of values). If the data fails the validation rules it is rejected entirely or in part. The rejected data is ideally reported back to the source system for further analysis to identify and to rectify the incorrect records.



**Transform**

In the data transformation stage, a series of rules or functions are applied to the extracted data in order to prepare it for loading into the end target. Some data does not require any transformation at all; such data is known as "direct move" or "pass through" data.

An important function of transformation is the cleaning of data, which aims to pass only "proper" data to the target. The challenge when different systems interact is in the relevant systems' interfacing and communicating. Character sets that may be available in one system may not be so in others.

In other cases, one or more of the following transformation types may be required to meet the business and technical needs of the server or data warehouse:

Selecting only certain columns to load: (or selecting null columns not to load). For example, if the source data has three columns (aka "attributes"), roll\_no, age, and salary, then the selection may take only roll\_no and salary. Or, the selection mechanism may ignore all those records where salary is not present (salary = null).

Translating coded values: (e.g., if the source system codes male as "1" and female as "2", but the warehouse codes male as "M" and female as "F")

Encoding free-form values: (e.g., mapping "Male" to "M")

Deriving a new calculated value: (e.g., sale\_amount = qty \* unit\_price)

Sorting or ordering the data based on a list of columns to improve search performance

Joining data from multiple sources (e.g., lookup, merge) and deduplicating the data

Aggregating (for example, rollup — summarizing multiple rows of data — total sales for each store, and for each region, etc.)

Generating surrogate-key values

Transposing or pivoting (turning multiple columns into multiple rows or vice versa)

Splitting a column into multiple columns (e.g., converting a comma-separated list, specified as a string in one column, into individual values in different columns)

Disaggregating repeating columns

Looking up and validating the relevant data from tables or referential files

Applying any form of data validation; failed validation may result in a full rejection of the data, partial rejection, or no rejection at all, and thus none, some, or all of the data is handed over to the next step depending on the rule design and exception handling; many of the above transformations may result in exceptions, e.g., when a code translation parses an unknown code in the extracted data

**LOAD**

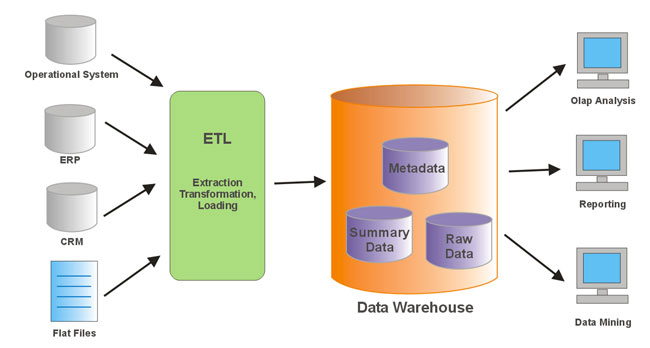
The load phase loads the data into the end target, which may be a simple delimited flat file or a data warehouse. Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information; updating extracted data is frequently done on a daily, weekly, or monthly basis. Other data warehouses (or even other parts of the same data warehouse) may add new data in a historical form at regular intervals—for example, hourly. To understand this, consider a data warehouse that is required to maintain sales records of the last year. This data warehouse overwrites any data older than a year with newer data. However, the entry of data for any one year window is made in a historical manner. The timing and scope to replace or append are strategic design choices dependent on the time available and the business needs. More complex systems can maintain a history and audit trail of all changes to the data loaded in the data warehouse.

As the load phase interacts with a database, the constraints defined in the database schema — as well as in triggers activated upon data load — apply (for example, uniqueness, referential integrity, mandatory fields), which also contribute to the overall data quality performance of the ETL process.

For example, a financial institution might have information on a customer in several departments and each department might have that customer's information listed in a different way. The membership department might list the customer by name, whereas the accounting department might list the customer by number. ETL can bundle all of these data elements and consolidate them into a uniform presentation, such as for storing in a database or data warehouse.

Another way that companies use ETL is to move information to another application permanently. For instance, the new application might use another database vendor and most likely a very different database schema. ETL can be used to transform the data into a format suitable for the new application to use.

An example would be an Expense and Cost Recovery System (ECRS) such as used by accountancies, consultancies, and legal firms. The data usually ends up in the time and billing system, although some businesses may also utilize the raw data for employee productivity reports to Human Resources (personnel dept.) or equipment usage reports to Facilities Management.



In our scenario, dataset is Mobileappstore.csv, so during the ETL process the data is extracted from this dataset, transformed to eliminate irrelevant data mentioned in the scope of analysis section and loaded into the excel where the required data resides. From this analysis reporting can be done.

**ANALYSIS ON DATASET**

**ANALYSIS 1**

**INTRODUCTION**

This analysis helps in getting the best app out of the large corpse of apps of same category on the basis of the reviews. One can easily sort out the best app based on the user reviews Users try to download good apps with the content on the reviews. Good apps are recognized by good reviews.

**GENERAL DESCRIPTION**

This analysis is to help play store users in figuring out good applications based on reviews. A good application is the one whose rating is greater than some value and have minimum number of user reviews. This analysis also includes costing factor where in applications with less than specified amount can be listed. Besides these two factors, one can choose the category as well.

**SPECIFIC REQUIREMENTS**

It is required to show the top reviews apps based on the category. This helps the users to identify the app which one to be downloaded. This can be well represented using line charts varying based on the category of apps.

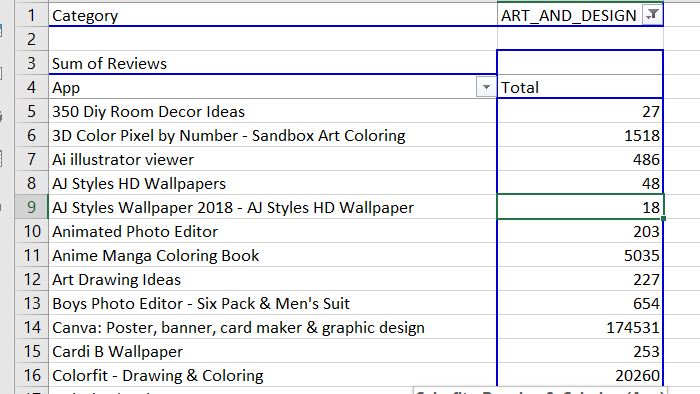
**FUNCTIONS AND FORMULAES**

In this analysis pivot table is used to find the top reviews of all the applications in each category.

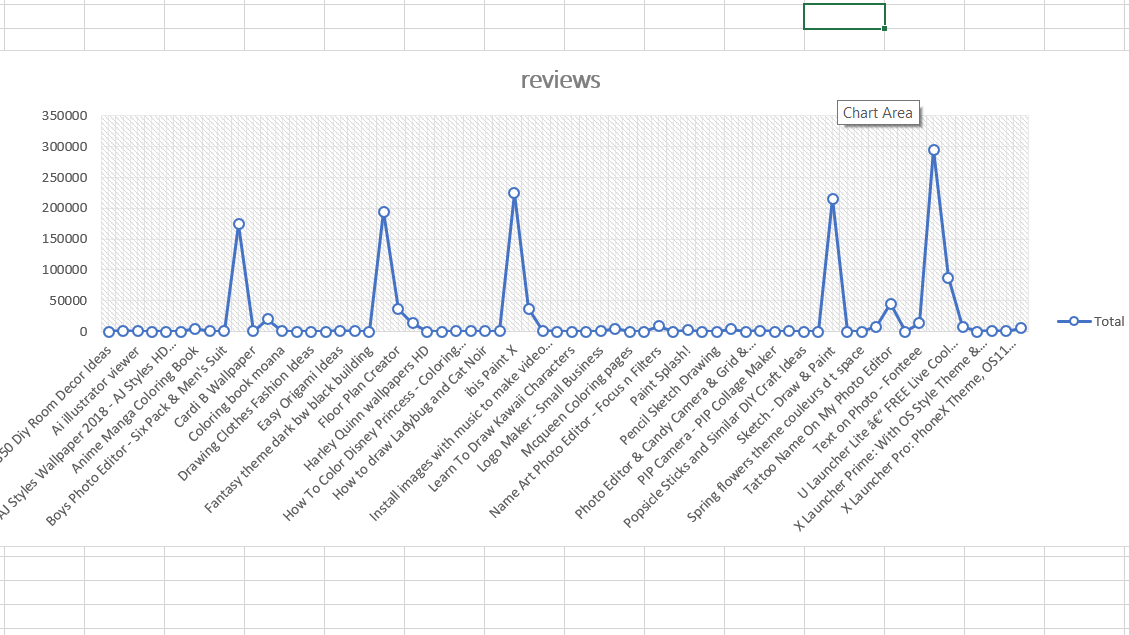
“App” is taken as the rows and “reviews” is taken in the values section and “category” is taken in the filter section.

Visualization is made through line graph.

**ANALYSIS RESULTS**



**VISUALIZATION**



**ANALYSIS 2**

**INTRODUCTION**

This analysis shows us the total number of apps that are present in the play store based on the content (i.e.” Everyone”,” Teen”,”18+” etc.). This restricts the users based on their age from downloading the app from the store. It also tells about in which category the application is based on and in which content rating types , also consists of the type of the application.

**GENERAL DESCRIPTION**

This analysis is to help play store users in figuring out applications based on content. Besides these this restricts children under the age group from viewing the content in these apps. And also the type whether it is paid or free and it is represented in bar graph.

**SPECIFIC REQUIREMENTS**

To show content rating of apps. This helps the users to identify the app which one to be downloaded. This can be well represented using bar graph varying based on the category of apps.

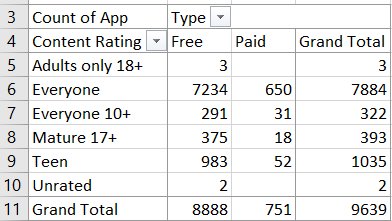
**FUNCTIONS AND FORMULAES**

In this analysis pivot table is used to find the type of content of all the applications in each category.

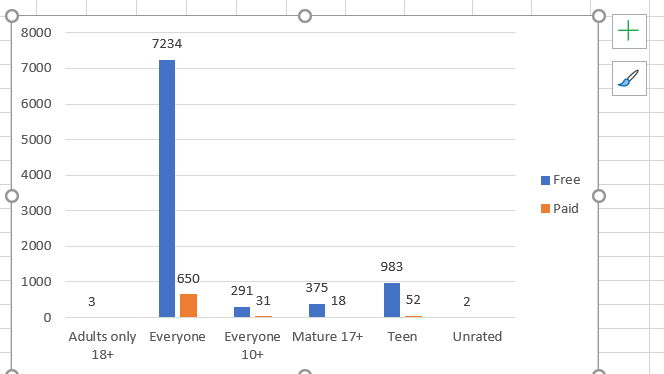
“Content Rating” is taken as the rows and “Count of Apps” is taken in the values section. Visualization is

made through bar graph.

**ANALYSIS RESULTS**



**VISUALIZATION**



**ANALYSIS 3**

**INTRODUCTION**

This analysis arranges the apps by assigning ranks to each one of them. Ranks are assigned based on downloads that are done all over the world.

**GENERAL DESCRIPTION**

This analysis is to helps in figuring out applications on the basis of their ranks. Besides these this distinguishes the apps based on their downloads (i.e. “100000+” , ”500000+” etc.).

**SPECIFIC REQUIREMENTS**

To show ranking of apps based on downloads. This helps to analyze between the topmost downloaded and the ones that are least downloaded. This can be well represented using line charts varying based on the rank.

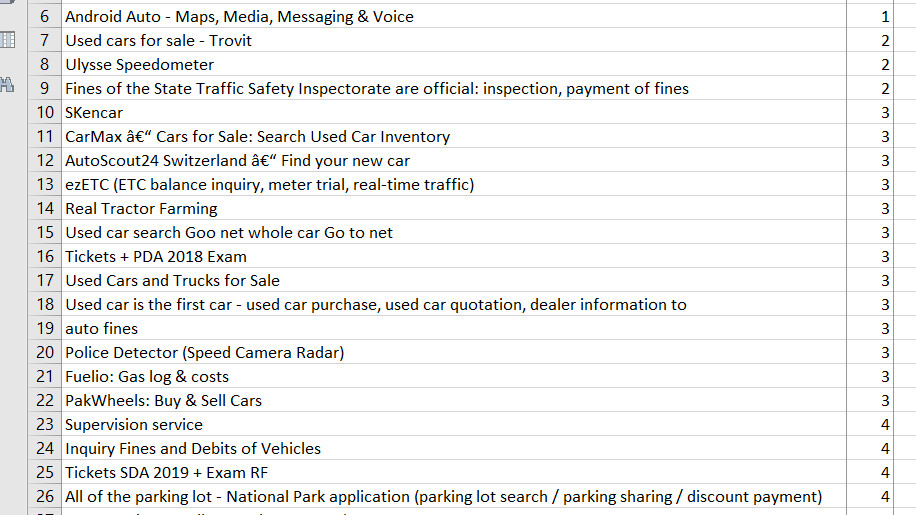
**FUNCTIONS AND FORMULAES**

In this analysis pivot table is used to distinguish between ranking of the apps.

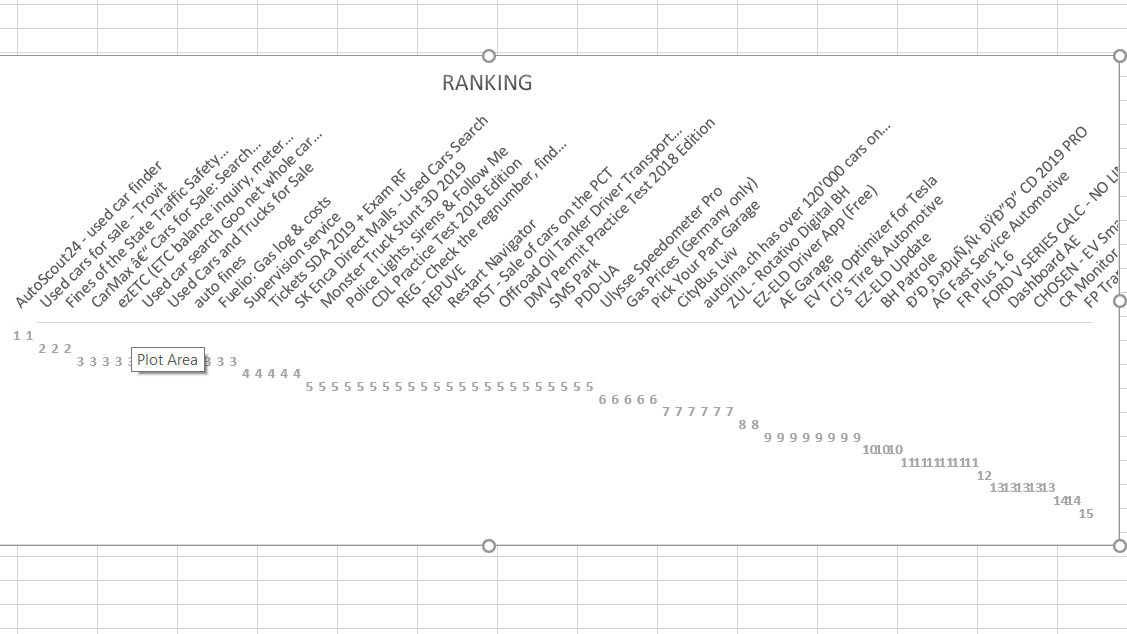
“App” is taken in the rows and “Rank” is taken in the values section. Visualization is

made through line chart. And the applications that are in the category of all the types in the data are ranked according to the installs of that application by the user.

**ANALYSIS RESULTS**



**VISUALIZATION**



**ANALYSIS 4**

**INTRODUCTION**

This analysis tells about the rating of the application accordingly the number of that application. It gives us the rating and as well as the number applications based on the category and it filters are used in the category column.

**GENERAL DESCRIPTION**

This analysis is to helps in figuring out applications based on their categories. The applications based on the category and tells about the count wise applications and followed by the rating of that application.

**SPECIFIC REQUIREMENTS**

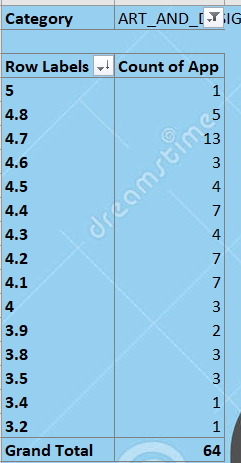
To show the applications based on the category and the applications rating are also taken in the parallel section and filters are applied on the category and this is the grouping of the data and is shown by applying the filters.

**FUNCTIONS AND FORMULAES**

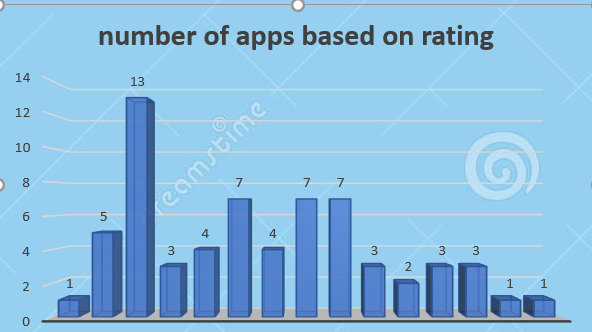
In this analysis pivot table is used to describes between installs and rating in each category of the application.

RATINGS are taken in rows and count of APPS is taken in the values section and is filtered by the category. Visualization is made through column chart.

**ANALYSIS RESULTS**



**VISUALIZATION**



**ANALYSIS 5**

**INTRODUCTION**

This analysis tells about the last updated applications based on the category of all the types that are mentioned on the given raw data.

**GENERAL DESCRIPTION**

This analysis helps in finding out the last updated applications based on the category and year and month wise. It tells about the applications that are last updated and are sorted in a newest to oldest.

**SPECIFIC REQUIREMENTS**

It is required to show the applications which are having latest updated according to the year-month wise and it is filtered accordingly by the sorting (newest to the oldest).

In the beginning the last updated column was not grouped by the year-month wise then later I converted it to year-month wise display by grouping the last updated column to the year-month wise option.

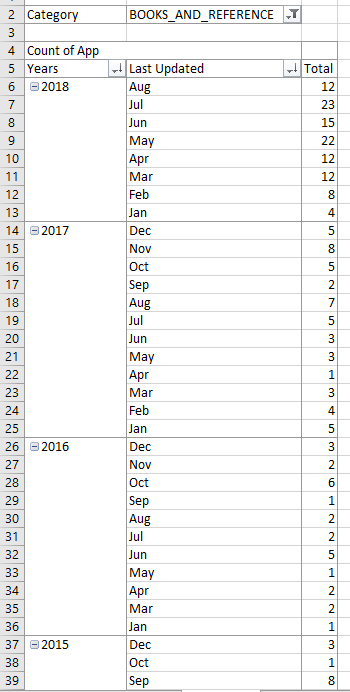
**FUNCTIONS AND FORMULAS**

In this analysis pivot table is used to the count of applications based on the last updated column.

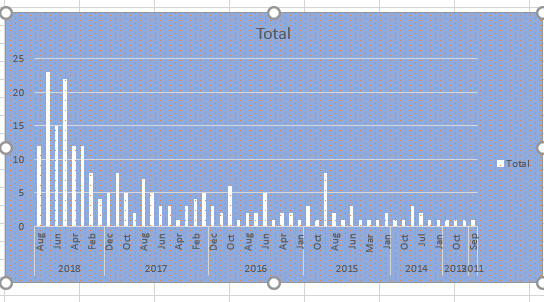
“LAST UPDATED”,”YEAR” is taken in the rows and “count of APPLICATIONS” is taken in the values section and “CATEGORY” is taken in the filter section.

Visualization is made through column chart. And the count of the applications that are in the category of all the types in the data are listed according to the last updated column that is grouped by the year and month wise.

**ANALYSIS RESULTS**



**VISUALIZATION**



**FUTURE SCOPE**

Data analytics is a process through which data is cleaned, analyzed and modelled using tools. This data is then used to derive insights. The insights are then used for business related decision-making purposes. There are many techniques that data analysts use in different fields of work. In the world of business, Data analytics is used for making strategies to get the desired business results. Today, data analytics has become a big career option in India. As a result, big data analytics courses are in huge demand.

* As more data comes in we can categorize the apps by language. This feature can allow app developers to know which languages to use for better reach and for users, the apps available in the language they want.
* Comparison between latest version and old version ratings. This feature will help the mobile app developers about the new features added and their acceptance levels in the market against the older version.

**BIBLIOGRAPHY**

Microsoft Excel 2010 Data Analysis And Business Modelling By Wayne L. Winson

Microsoft Excel 2010 Step by Step By Curtis Frye

<https://nycdatascience.com/blog/student-works/web-scraping/analysis-of-apps-in-the-google-play-store/>

<https://www.kaggle.com/datasets>

<https://www.youtube.com/results?search_query=dashboard+on+google+play+store+in+excel>