Data Scraping from Google Play Store and Visualization of its Content for Analytics

Rana M. Amir Latif
Department of Computer Science
COMSATS UNIVERSITY ISLAMABAD,
Sahiwal Campus
Sahiwal, Pakistan
ranaamir10611@gmail.com

Syed Umair Aslam Shah
Department of Computer Science
COMSATS University Islamabad,
Sahiwal Campus
Sahiwal, Pakistan
umairaslamsahiwal@gmail.com

Farah Ijaz
Department of Computer Science
COMSATS University Islamabad,
Sahiwal Campus
Sahiwal, Pakistan
farahijaz777@gmail.com

M. Talha Abdullah
Department of Computer Science
COMSATS UNIVERSITY ISLAMABAD,
Sahiwal Campus
Sahiwal, Pakistan
talha.ch.0334@gmail.com

Muhammad Farhan
Department of Computer Science
COMSATS UNIVERSITY ISLAMABAD,
Sahiwal Campus
Sahiwal, Pakistan
farhansajid@gmail.com

Abdul Karim
National College Of
Business Administration & Economics,
Rahim Yar Khan, Pakistan
abdullkarim@hotmail.com

***Abstract*— There are millions of applications uploaded by the developers on the daily basis. Without any check and balance millions of users download these applications. Theses duplicated applications damage the users trust on Google play store and can grab the confidential information of user. There is no more information provided by developers on the front end of the application that can define the legitimacy of the application. In this paper, by using a Google-play-scraper build a Google play store dataset with all categories of games. Scraping at least 550 applications of each category of games in free and respectively in paid applications by using Google play scraper, cumulatively scrape the 3600 paid applications and 10k free applications of all categories in games. The categories of these games' applications use respectively are Word, Trivia, Simulation, Sports, Strategy, Racing, Role_Playing, Puzzle, Music, Educational, Card, Casino, Casual, Board, Action, Adventure, and Arcade. On each application on Google play store, scrape maximum 70 attributes, but use four attributes for analysis in this paper that is Installs, Advertisements support, InApplicationPurcahses and Ratings. In this paper, visualizing the InAppPurchase rate of free and paid applications, Percentage of the advertisement support in free and paid applications, Ratings of free and paid application with histogram, Installs of free and paid application with a histogram with all categories of games application. To check the relationship in between attributes also, visualize them in CIRCOS. This visualization is more helpful for game developers in the development phases, also for the users of the game's application for the selection of the game that they want to play.**

Keywords*— *measurement and analysis; scraping; games applications; big data; google play store; circos visualization

I. INTRODUCTION

Google play store allows many third-party applications to be download by users. Personal information of the user is registered by third-party applications. These third-party applications can be installed by millions of users on their phones and tablets. Daily thousands of developers upload millions of applications. Millions of users download from play store but unfortunately, most of the content uploaded is unchecked [1]. Little information is available here on the Google play store. Lack of scalable tool for discovering and analyzing is the main reason. Specific third-party have the source code of the application. Inform of compressed binary packages submission of source code on Google play store is not even access by play store [2]. Crawling and scraping techniques are used to get information from the play store. Through crawling on content grow day by day and play daily basis get information. So, the content of Google play store and metadata will measure and analyze [3]. After millions of applications revivals get two types of applications on play store paid and free applications also have different categories list of game and home applications. Why some applications are more popular in users? What content of attributes should use? In these things' developers can take help from measurement and analysis. Applications performance and efficiency can be increased by developers. Analysis will not only useful for developers, but the user can also take advantage. The content of Google play store applications will characterize at a specific scale. This research will show paid and free applications relationship by measure and analyze attribute as the ratio of advertisements, number of downloads and price relationship with ranking [4].

The impact can be of two types positive or negative such things with additional visibility of applications how to change with a self-categorization choice will also discuss. This research work will also cover that how over time evolved, update, released and removed content of Google play store

978-1-5386-9509-8/19/$31.00 ©2019 IEEE

applications. All type of applications contains free application accounts in small percentage [5]. Millions of applications at play store are discarded by the recent policy of cloned of famous applications [6]. Clone removal effect will also check after analysis. Through play store dataset after crawling, similar applications can be easily detected based on applications description. Correlation between different application can be visualized by different techniques based on different attributes. In x pie3D scales value so NAs and zeros total 2*pi drops. Pie3D will calculate drawing sector and calls sequence. Empty slot at sector is also displayed and there is also a title. In the case of labels are supplied than in each sector label is placed [7]. In case of label supplied, with x number, label positions and sector color should match. The radius of pip can be shrunk and change the position in case of long labels. For detail graphical parameter arguments and plotting, generic function R objects are used. In R for data and functions, density objects and frames, many methods can used but simple and scatter plot default value is used [8]. Positions of objects and objects are visualized in a circular format known as CIRCOS. Data is layered richly, increase publication infographics quality and ideal tool with attractive looks are some properties which makes it more suitable [9].

II. Literature Review

Application obfuscation data is available in a small amount and frequently applications are plagiarized or repackaged. Software obfuscation challenges and use are done in a comprehensive analysis. Developers obfuscated 24.92% apps finds by authors after 1.7 million free applications analysis [10]. Then attitudes and experiences about obfuscation of 308 developers were surveyed to understand this obfuscation rate better. For their own app's developers applying obfuscation by self-report. Authors also found that when developers feel the risk of plagiarism, they do not fear theft of their own apps. For better understanding follow-up study is conducted by authors in which 70 participants failed from the vast majority to obfuscate a realistic sample app even while many mistakenly believed they had been successful. More work is needed for the obfuscation tool to make it more usable to educated developers and improve overall Android ecosystem health [11].

Consumer behaviors semantic analysis on social media for understanding opinion polarity has been increased on daily basis. For deep learning development little intention is required by user's reviews in China. Based on users of Google play store in China impact of deep learning checks in this paper. 196,651 reviews author collect by web mining technique from Google play store. For semantic analysis, different approaches are used by the author as deep learning and Long Short-Term Memory (LSTM) [12], Naïve Bayes (NB) [13] and support vector machine (SVM) [14]. Compared results after getting from these models. Deep learning model has more accuracy than models' other models mentioned in that paper. Firstly, results clear that deep learning model is perfect for non-average sampling data. Secondly, iSGoPaSD named semantic analysis dictionary is created. Thirdly, Google plays store prediction is improved by deep learning semantic analysis [15]. At Java source code, Dalvik bytecode and java level, evaluation strategies perform by authors in this research work. In C or C++, builders can compose code because it is enabled by Android, but these languages are cross-compiled to several binary architectures. Java-written components (C or C++) and native code component interact. "In this research work, authors perform the evaluation strategies at the java level, Dalvik bytecode and Java source code. Though, Android enables builders to compose code in C++ or C that is cross-compiled to several binary architectures. Moreover, the native code components and Java-written components (C or C++) interact.

Regulate the Dalvik Virtual Machine, as well as the java code entry, is the native code admission to all the Android APIs [16], so java unsound and rendering static analysis procedures for misleading. Moreover, malicious apps often use native code to launch kernel exploits or conceal their malicious capability in native code. This is the reason for implementing the security rules in sandboxing also, safety concerns that research in this paper has native code sandboxing. It is feasible without breaking the app functionality define the native code sandboxing [17] [18]. In this paper, with a broad assessment of the native code usage in 1.2 million Android applications. Firstly, authors perform in this paper the static examination to locate a lot of 446k application conceivably utilization native code, and after that dissected this set utilization of dynamic investigation. This assessment demonstrates that sandboxing native code and not utilizing a consent isn't in every case best, as applications' native code parts complete games that require Android authorizations [19].

In results, the analysis was so mature and trustworthy that shows the sandboxing in native code are more useful and feasible in practice. In all actuality, it changed into doable to routinely produce a native code sandboxing approach, which is gotten from examination, that limits numerous malignant practices in the meantime as in any case allowing a fitting execution of the lead saw sooner or later of dynamic assessment for 99.77% of the kind applications in dataset. The utilization of this framework to produce rules diminish the assault surface accessible to native code and, as an extra favorable position, it may furthermore empower additional dependable static examination of Java code."

The process of application is always thinking to be more scalable and effective. In this paper, authors realize that today's selection process of application is too slow and less capable to identify the threats in the working of the application. Authors analyze that now a day's applications in the market are just the repacking of the legitimate applications, the content, theme, and idea of these applications are loos like the same as legitimate application. These applications are constructed and spreading in the market by using the Android malware. These applications are often standalone and not supposed to relate each other. In this research paper author use technique which is called a MassVet which is used for the selection process in this paper on a large scale. The main idea of this paper to just to analyze and identify that application which is come into the market on just repacking basis also, that application which is already submitted into the market [20]. The MassVet technique is implemented on a stream processing engine and evaluating the 1.3 million applications from 33 applications market from all over the world. By using this technique, authors detect the application within 10 seconds with the very low false error rate.

Also, this technique performs the scanners in VirusTotal that include McAfee, NOD32. This scanner works in term of virus detection coverage, grab millions of malicious applications [21]."

III. Methodology

There are two major types of application on Google play store are free and paid. These types of applications also have several other categories of applications like games, movies, education and video etc. these all applications present on Google play store. In this paper, by using a Google-play-scraper build a Google play store dataset with all categories of games. Scraping at least 550 applications of each category of games in free and respectively in paid applications by using Google play scraper, cumulatively scrape the 3600 paid applications and 10k free applications of all categories in games. The categories of these games' applications use respectively are Word, Trivia, Simulation, Sports, Strategy, Racing, Role_Playing, Puzzle, Music, Educational, Card, Casino, Casual, Board, Action, Adventure, and Arcade. On each application on Google play store, scrape maximum 70 attributes, but use four attributes for analysis in this paper that is Installs, Advertisements support, InApplicationPurcahses and Ratings. which use in analysis is shown in "Fig. 1,". Analysis of dataset is performed in R-Studio and to check the relationship between attributes use CIRCOS as shown in "Fig. 2,".

	ratings	installs	offersIAP	adSupported
2	1114	100,000+	FALSE	TRUE
3	2735281	100,000,000+	TRUE	TRUE
4	249269	10,000,000+	TRUE	TRUE
5	1430466	50,000,000+	TRUE	TRUE
6	7167674	100,000,000+	TRUE	FALSE
7	8471286	100,000,000+	TRUE	TRUE
8	5428362	100,000,000+	TRUE	TRUE
9	519031	10,000,000+	TRUE	FALSE
10	2026003	50,000,000+	TRUE	TRUE
11	4716463	50,000,000+	TRUE	TRUE
12	292717	10,000,000+	TRUE	FALSE
13	7099	1,000,000+	TRUE	TRUE
14	4887	500,000+	TRUE	FALSE
15	4674	500,000+	TRUE	TRUE
16	10336264	100,000,000+	TRUE	FALSE
17	98471	10,000,000+	TRUE	TRUE
18	8393568	500,000,000+	TRUE	TRUE
19	245987	10,000,000+	TRUE	TRUE
20	2828	1,000,000+	TRUE	TRUE

Fig. 1. *Small Sample Screenshot of Dataset*

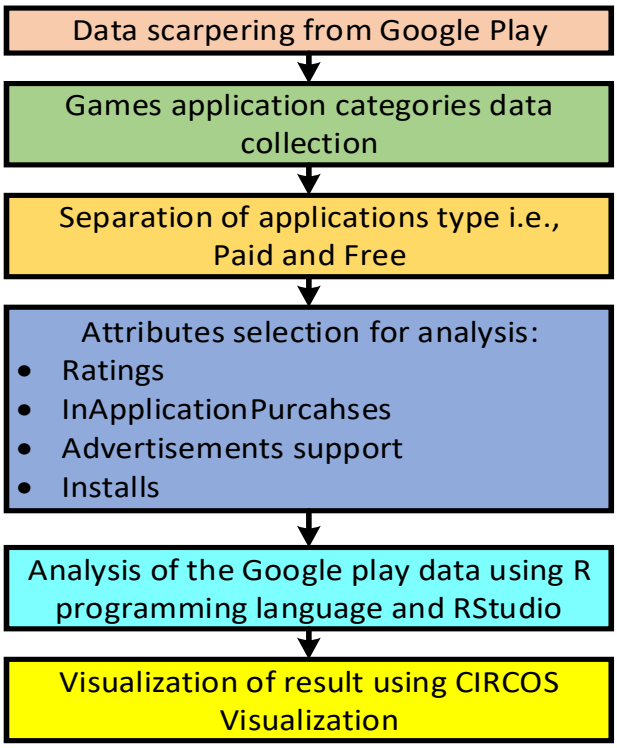

Fig. 2. *Methodology Diagram of Google Play Store Dataset Visualization*

IV. Results and Discussion

There are millions of applications uploaded by the developers on the daily basis. Without any check and balance, millions of users download these applications. Theses duplicated applications damage the users trust on Google play store and can grab the confidential information of the user. There is no more information provided by developers on the front end of the application that can define the legitimacy of the application. In this paper, visualizing the InAppPurchase rate of free and paid applications, Percentage of the advertisement support in free and paid applications, Ratings of free and paid application with histogram, Installs of free and paid application with a histogram with all categories of games application. To check the relationship in between attributes also, visualize them in CIRCOS. This visualization is more helpful for game developers in the development phases, also for the users of the game's application for the selection of the game that they want to play.

A. Free Games InApplicationPurchases (IAP)

In Google play store many games contain such items, products, credits etc. which get by performing actions which are not relevant to the game. For example, some games give you coins by just seeing ads, videos, by filling survey forms, buying through credit card etc. This thing knew as in application purchase. Free applications did not charge anything at the time of install that is why more than half offers IAP. Use the pie3D chart to visualize this attribute. The pie3D chart shows that 66% of free games offer IAP and 34% free games did not offer IAP as shown in "Fig. 3,".

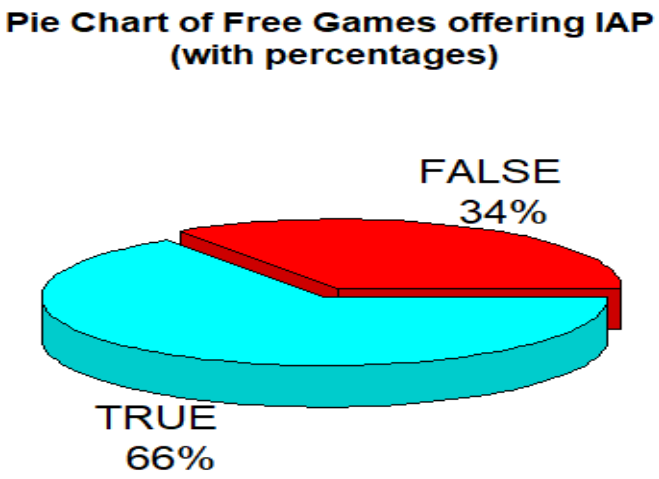

Fig. 3. *PieChart of Free Offering InApplicationPurcahes (IAP)*

B. Paid Games InApplicationPurchases (IAP)

Paid applications try to provide maximum comfort and satisfaction to its users. Paid applications charge money at the time of install that is why maximum did not offer IAP. Use the pie3D chart to visualize this attribute. The pie3D chart shows that just 17% paid games to offer IAP and just 83% paid games did not offer IAP as shown in "Fig. 4,".

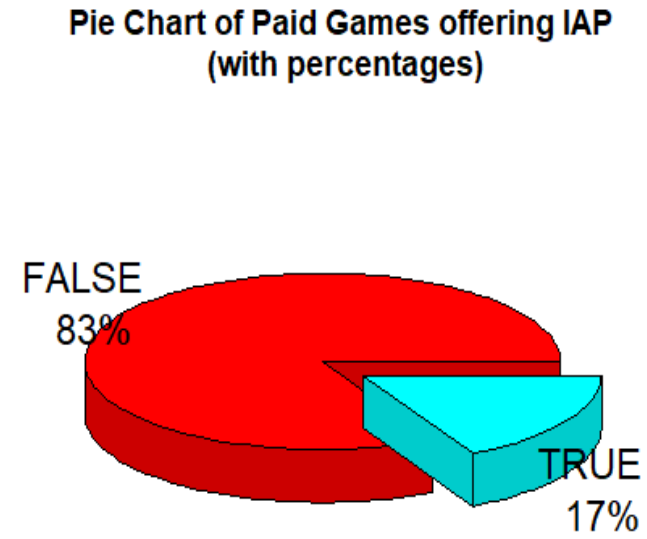

Fig. 4. *PieChart of Paid Offering InApplicationPurcahes (IAP)*

C. Free Games Advertisements

In Google play store many applications contain advertisements of companies, brands etc. Free applications did not charge anything at the time of install that is why majority offers advertisements. Use the pie3d chart to visualize this attribute. The pie3D chart shows that 87% of free games offer advertisements and just 13% free games did not offer advertisements as shown in "Fig. 5,".

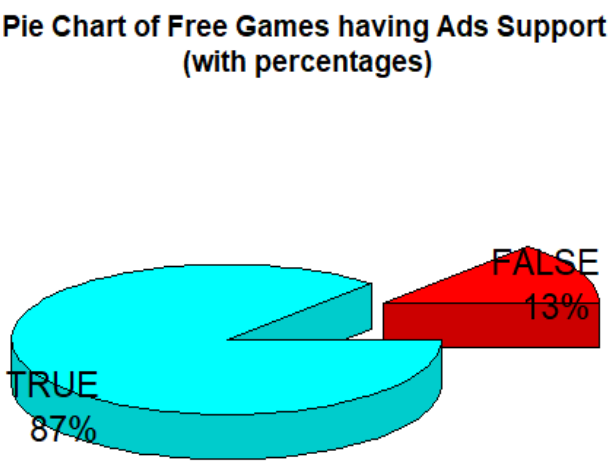

Fig. 5. *PieChart of Free Games having Advertisements Support*

D. Paid Games Advertisements

Paid applications charge at the time of install that is why majority did not offer advertisements. Paid applications try to provide maximum comfort and satisfaction to its users. Use the pie3D chart to visualize this attribute. The pie3D chart shows that just 14% paid games to offer advertisements and 86% paid games did not offer advertisements as shown in "Fig. 6,".

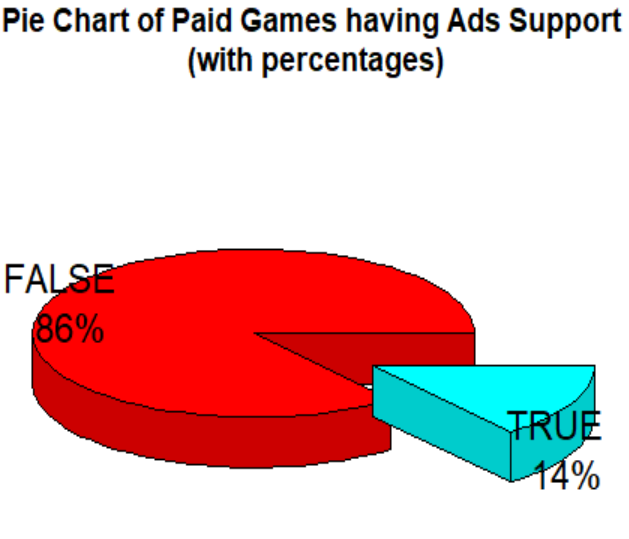

Fig. 6. *PieChart of Paid Games having Advertisements Support*

E. Free Games Rating Value

In Google play store users give ratings to the application according to there experience about that application. In Google play, store ratings can be given in the range from 0 to 5.0 represents very bad experience and 5 represents a very good experience. This attribute can be visualized in the histogram that how people ratings to free games. It can be clearly visualized in the histogram that most people give 4.5 ratings to the free game as shown in "Fig. 7,".

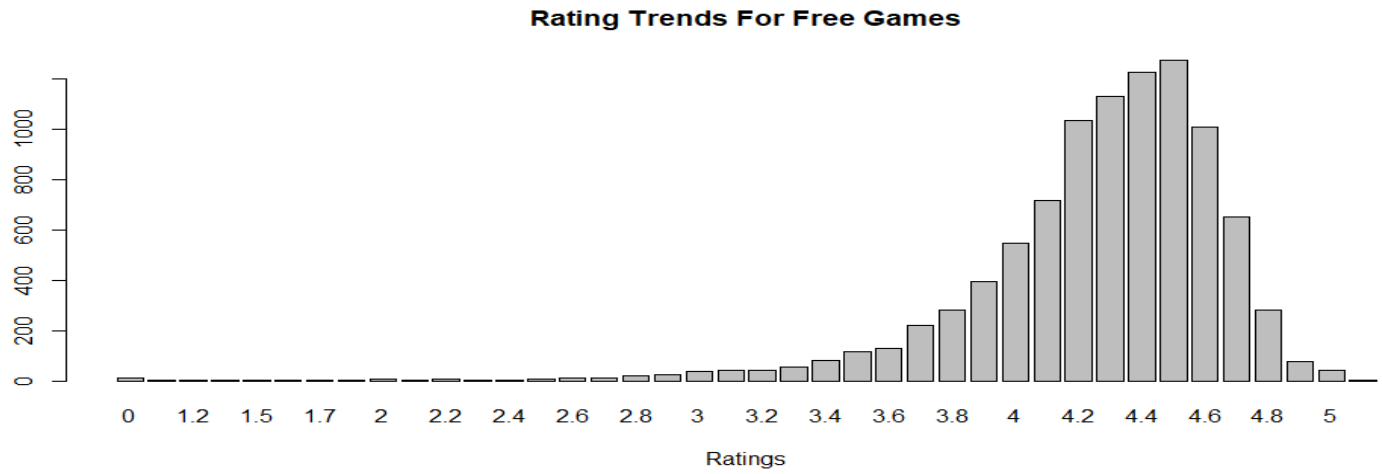

Fig. 7. *Rating Trends of Free Games*

F. All Categories of Free Games Ratings

In Google play store game applications contains categories as Arcade, Adventure, Action, Board, Casual, Casino, Card, Educational, Music, Puzzle, Role_Playing, Racing, Strategy, Sports, Simulation, Trivia, and Word. It can be clearly visualized in the histogram that most people give ratings to card category as shown in "Fig. 8".

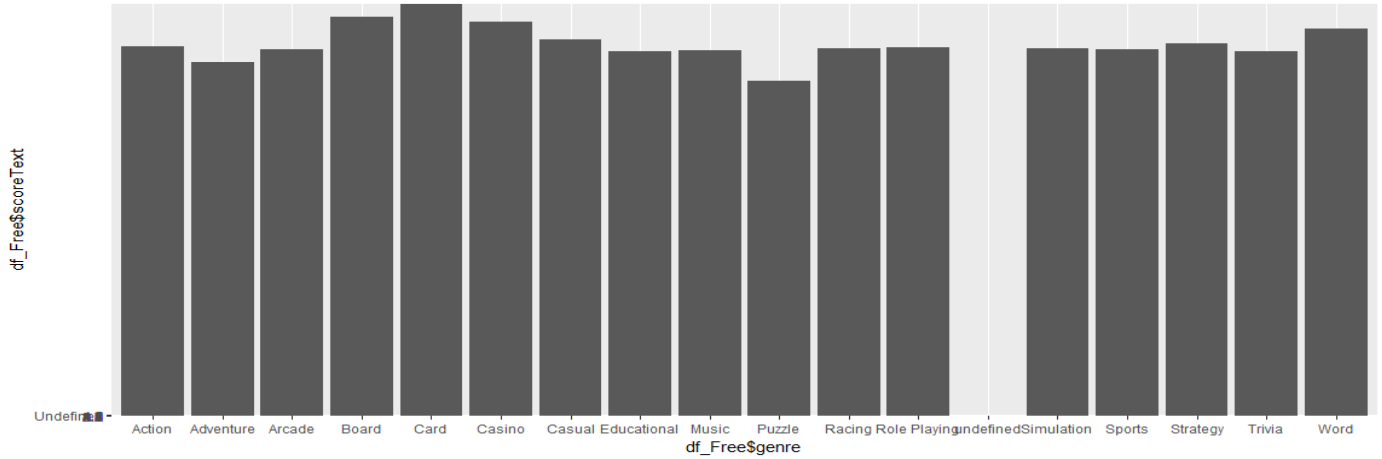

Fig. 8. *All Categories of Free Games Ratings*

G. All Categories of Paid Games Ratings

In Google play store ratings matter a lot in decision making for an application. The histogram shows that which category of paid games more ratings. It can be clearly visualized by a histogram that most people give ratings to the puzzle category as shown in "Fig. 9,".

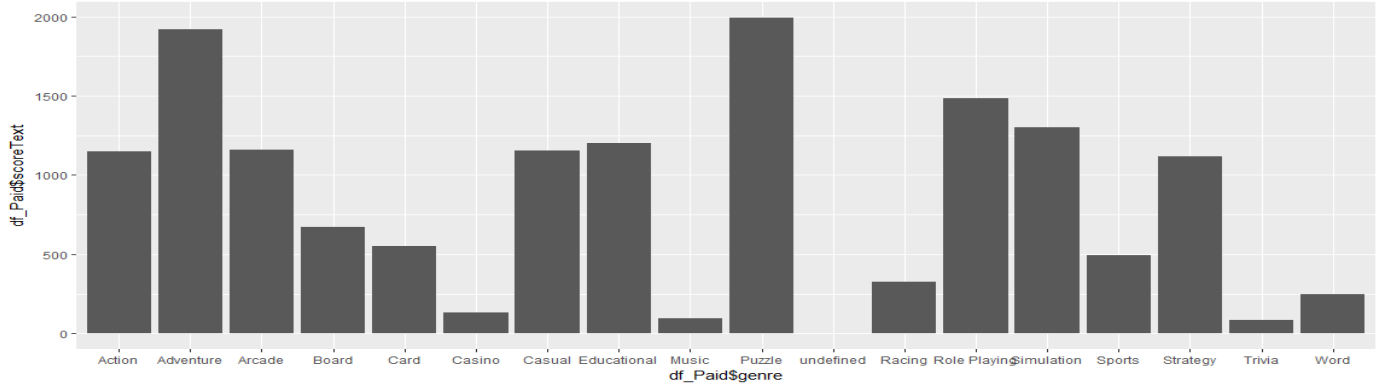

Fig. 9. *All Categories of Paid Games Ratings*

H. *All Categories of Free Games Installs*

In Google play, store number of installs play a vital role to make rough view about an application. Following 2 diagrams shows several installs of free games with respect to categories. As Arcade category have a greater number of installs with respect to other categories as shown in “Fig. 10,”.

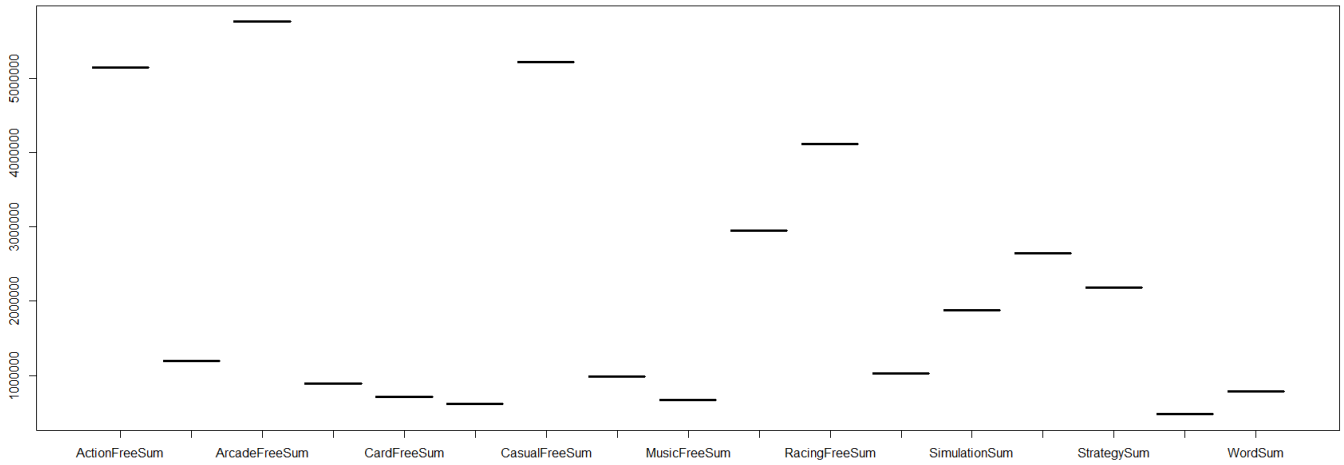

Fig. 10. *All Categories of Free Games Installs*

I. *All Categories of Paid Games Installs*

If in Google play store find category which installs is most, guide developers to try their skills in that category. Following 2 diagrams shows several installs of paid games with respect to categories. As Arcade category have a greater number of installs with respect to other categories as shown in “Fig. 11,”.

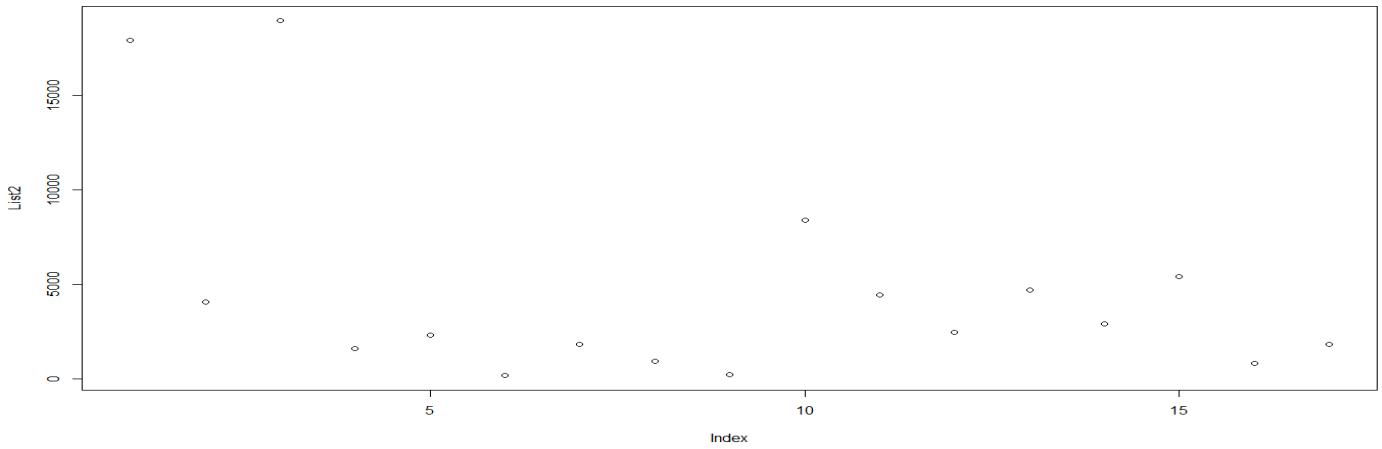

Fig. 11. *All Categories of Paid Games Installs*

J. *CIRCOS visualization for free and paid Games Applications*

CIRCOS is the tool in which multiple dependent and independent variables could be easily identified by the systematic point of view and best corporation of the layout. Ratings, InApplicationPurcahses, Advertisements support, and Installs are being used for the analysis and better visualization of the attributes data of free and paid games applications. The CIRCOS is the software package which is being used for the information and visualization of the information given to the package software. The circular layout of the data could be easily identified and the position of the objects. The circular region has more advantages due to the attractive responses. CIRCOS is the ideal creating for the publication quality and it also illustrates the high data into ink ratio. Which richly identifies the symmetries among all the values. The audience details and focuses the point of the figured requirement of the image. For the visualization of the genomic data and the genomics, migration could be identified by the system data. The multi-layered of the mathematical art. The data that describes the relationships of the multi-layered and annotations for the scales of CIRCOS as shown in “Fig. 12,”.

a) CIRCOS visulizatoin for Free Games Applications

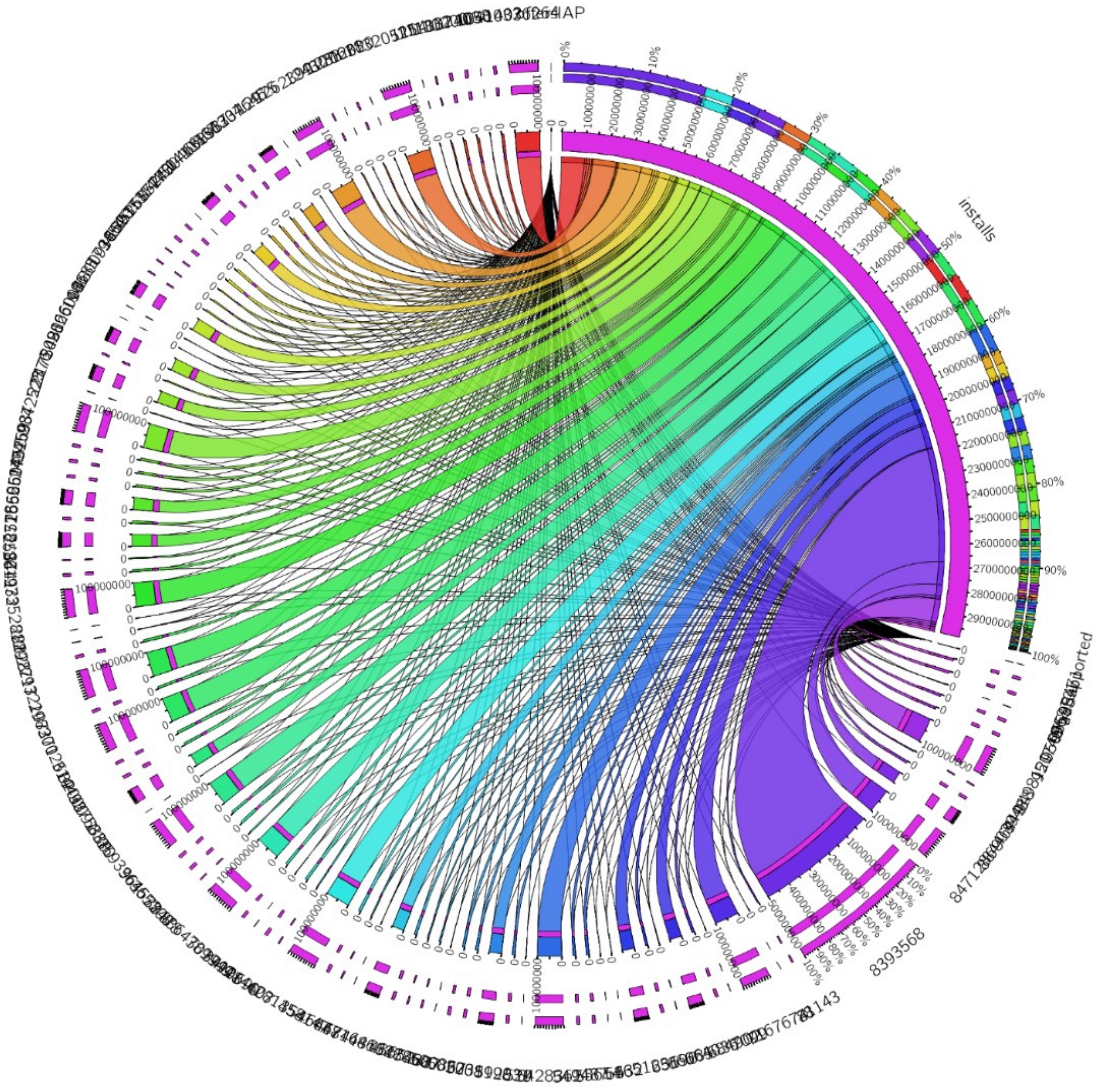

Fig. 12. *CIRCOS Visualization for Free Games with Four Attributes*

V. Conclusion and Future Work

There are two major types of application on Google paly store are free and paid. These types of applications also have several other categories of applications like games, movies, education and video etc. these all applications present on Google play store. In this paper, by using a Google-play-scraper build a Google play store dataset with all categories of games. Scraping at least 550 applications of each category of games in free and respectively in paid applications by using Google play scraper, cumulatively scrape the 3600 paid applications and 10k free applications of all categories in games. The categories of these games' applications use respectively are Word, Trivia, Simulation, Sports, Strategy, Racing, Role_Playing, Puzzle, Music, Educational, Card, Casino, Casual, Board, Action, Adventure and Arcade. On each application on Google play store, scrape maximum 70 attributes, but use four attributes for analysis in this paper that is Installs, Advertisements support, InApplicationPurcahses and Ratings. Analysis of dataset is performed in R-Studio by using different packages in R also to check the relationship in different attributes that are used in this research paper used CIRCOS. In CIRCOS also, check the different configuration that shows the statistical information about the different libraries like percentile, Components, and color. In future work scrape more attributes that will be helpful for more measurement and precise analysis of Google play application, the similarity factor in different applications can be checked, also make clusters of different attributes, check the relationship between different attributes with the help of that clusters.

References

[1] W. Yang, J. Li, Y. Zhang, Y. Li, J. Shu, and D. Gu, "APKLancet: tumor payload diagnosis and purification for Android applications," in Proceedings of the 9th ACM symposium on Information, computer and communications security, 2014, pp. 483-494: ACM.

[2] J. Crussell, C. Gibler, and H. Chen, "Attack of the clones: Detecting cloned applications on android markets," in European Symposium on Research in Computer Security, 2012, pp. 37-54: Springer.

[3] S. Bagnasco, D. Berzano, A. Guarise, S. Lusso, M. Masera, and S. Vallero, "Monitoring of IaaS and scientific applications on the Cloud

using the Elasticsearch ecosystem," in Journal of Physics: Conference Series, 2015, vol. 608, no. 1, p. 012016: IOP Publishing.

[4] W. Zhou, Y. Zhou, M. Grace, X. Jiang, and S. Zou, "Fast, scalable detection of piggybacked mobile applications," in Proceedings of the third ACM conference on Data and application security and privacy, 2013, pp. 185-196: ACM.

[5] S.-H. Seo, A. Gupta, A. M. Sallam, E. Bertino, K. J. J. o. N. Yim, and C. Applications, "Detecting mobile malware threats to homeland security through static analysis," vol. 38, pp. 43-53, 2014.

[6] H. Wang, H. Li, L. Li, Y. Guo, and G. Xu, "Why are Android apps removed from Google Play?: a large-scale empirical study," in Proceedings of the 15th International Conference on Mining Software Repositories, 2018, pp. 231-242: ACM.

[7] D. Bruns, "An Introduction to the Simplicity and Power of SAS/Graph," Tutorials of the Thirtieth Annual SAS, 2005.

[8] R. Suzuki and H. Shimodaira, "Pvclust: an R package for assessing the uncertainty in hierarchical clustering," Bioinformatics, vol. 22, no. 12, pp. 1540-1542, 2006.

[9] M. I. Krzywinski et al., "CIRCOS: an information aesthetic for comparative genomics," 2009.

[10] D. Wermke, N. Huaman, Y. Acar, B. Reaves, P. Traynor, and S. J. a. p. a. Fahl, "A Large Scale Investigation of Obfuscation Use in Google Play," 2018.

[11] A. Deo, S. K. Dash, G. Suarez-Tangil, V. Vovk, and L. Cavallaro, "Prescience: Probabilistic guidance on the retraining conundrum for malware detection," in Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security, 2016, pp. 71-82: ACM.

[12] S. Hochreiter and J. J. N. c. Schmidhuber, "Long short-term memory," vol. 9, no. 8, pp. 1735-1780, 1997.

[13] J. Chen, H. Huang, S. Tian, and Y. J. E. S. w. A. Qu, "Feature selection for text classification with Naïve Bayes," vol. 36, no. 3, pp. 5432-5435, 2009.

[14] V. Cherkassky and Y. J. N. n. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," vol. 17, no. 1, pp. 113-126, 2004.

[15] M.-Y. Day and Y.-D. Lin, "Deep Learning for Sentiment Analysis on Google Play Consumer Review," in Information Reuse and Integration (IRI), 2017 IEEE International Conference on, 2017, pp. 382-388: IEEE.

[16] D. J. T. r. Ehringer, "The dalvik virtual machine architecture," vol. 4, no. 8, 2010.

[17] B. Yee et al., "Native client: A sandbox for portable, untrusted x86 native code," in Security and Privacy, 2009 30th IEEE Symposium on, 2009, pp. 79-93: IEEE.

[18] V. Afonso et al., "Going native: Using a large-scale analysis of android apps to create a practical native-code sandboxing policy," in The Network and Distributed System Security Symposium, 2016, pp. 1-15.

[19] M. Spreitzenbarth, F. Freiling, F. Echtler, T. Schreck, and J. Hoffmann, "Mobile-sandbox: having a deeper look into android applications," in Proceedings of the 28th Annual ACM Symposium on Applied Computing, 2013, pp. 1808-1815: ACM.

[20] F. Zhang, H. Huang, S. Zhu, D. Wu, and P. Liu, "ViewDroid: Towards obfuscation-resilient mobile application repackaging detection," in Proceedings of the 2014 ACM conference on Security and privacy in wireless & mobile networks, 2014, pp. 25-36: ACM.

[21] K. Chen et al., "Finding Unknown Malice in 10 Seconds: Mass Vetting for New Threats at the Google-Play Scale," in USENIX Security Symposium, 2015, vol. 15.