INFLUENCE OF RATINGS AND REVIEWS ON ANDROID APP INSTALLATIONS FROM GOOGLE PLAY STORE

PROJECT GROUP – 4

BINDU PRIYA VODDHULA – A658M593

VINAY CHOWDARI MANDAVA – F233Z974

LOKESH MUPPALLA – J639H476

Instructor: **Justin Keeler**

BSAN 775 - 17396 - Introduction to Business Analytics

TR – 2:00 to 3:15

ABSTRACT

Software application is vital because specific software is required in almost every industry, in every business, and for each function. It became more important as time goes on. Mobile app distribution platform such as Google play store gets flooded with millions of new applications uploaded by developers every day. So in this project, we aim on analyzing Google play store that provides a particular app description and data such as reviews, ratings, price and number of downloads. The objective of this is to analyze the desire of the customer through the reviews provided in the feedback section and apps trend in the market to help the organization & developers. To this end, we provide an idea about app that managed to get maximum and minimum number of downloads and predicting the category of apps that is most likely to be downloaded in the coming years. Moreover, doing sentimental analysis on the apps that generated most positive and negative sentiments, sustainability of app in market on basis of previous data and current market situation. Furthermore, also analyzing the apps that has maximum downloads have they managed to get average rating so that concluding the co-relation between number of downloads and ratings received.

The google play store is one of the largest and most pop- ular Android app stores. It has an enormous amount of data that can be used to make an optimal model. We have used a raw data set of Google Play Store from the Kaggle website. This data set contains 13 different features that can be used for predicting whether an app will be successful or not using different features. This data set is scraped from the Google Play Store. This journal talks about different classifier models that we used for prediction purposes and finding which one gives the highest accuracy. This journal also gives de- tailed information on feature extraction and the complete Data visualization done on this data set. Our project code can be found at https://github.com/Rimshamaredia/CSCE-421-Project.

TABLE OF CONTENTS:

1. INTRODUCTION

Mobile applications are one of the fastest-growing segments of downloadable software application markets. Out of all of the markets we choose Google Play store due to its increasing popularity and recent fast growth [10]. One of the main reasons for this popularity is the fact that about 81% of the apps are free of cost [3]. The market has in- creased to over 845900 Apps and 226,500 unique sellers in April 2013 [2]. This rapid market has, in turn, led to over 500 million users downloading around 40 billion Apps all over the world [2]. Developers and users play key roles in determining the impact that market interactions have on future technology. However, the lack of a clear understanding of the inner working and dynamic of popular app markets impacts both the developers and users. In this article, we seek to shed light on the dynamics of the Google Play Store and how we can use different features from this data set for prediction purposes.

In this article, we will provide a longitudinal study of Google Play app metadata which will give unique information that is not available through the standard approach of capturing a single app snapshot. Using feature extraction from a longitudinal app analysis will be used to find whether an app will be successful or not. Our Analysis is divided into four phases: data extraction, data cleaning, data visualization, and applying different models, and it is depicted in figure 8. First, we collect the data from the Kaggle website. In the next step, we try to do data cleaning on the data set to reduce the error percentage. After the data set is ready, we try to analyze the data set using different plots and re- move the stuff not needed from the data set. The last step includes using different classification algorithms on the data set to see which one gives the highest percentage of accuracy. Finally, we narrate the analysis results to provide a clear vision of the relationship among the areas of interest.

We include a detailed discussion of the applicability and future research directions in the last section called Conclusion and future work.

2. PROBLEM STATEMENT

The Play Store apps data has enormous potential to drive app-making businesses to success. Android is expanding as an operating system and Mobile app industry is increasing in significantly and thus giving rise to more competitions to the ones that are creating applications. Due to the competition in the market and also expansion in order to help our developer understand what kinds of apps are likely to attract more users and what is the motivating factor for the people to download an app we analyze and research relevant data. For the app development industry where they can analyses the downloads and demand off app download in the industry.

We aim on providing doing sentimental analysis on the apps that generated most positive and negative sentiments and sustainability of app in market on basis of previous data and current market. The lack of thorough understanding of the dynamic and inner workings of well-known app markets has an impact on both the creators and the users. This article aims to clarify the Google Play Store's dynamics and demonstrate how various features from this data set can be used to make predictions. As an operating system, Android is growing and it has a market share of roughly 74%, which accurately reflects the vast number of people who use android. Our objective is to assist Android developers in understanding the driving force behind app downloads. Identifying the elements that influence a person's decision to download an app will also be helpful. For this reason, we want to investigate the relationships between category, reviews, pricing, and ratings.

3. OBJECTIVE

The aim of our analysis is to provide insights about android applications and their categories. We want to deep dive in data for the factors of influences on an application, to know why and how certain applications succeed and others. Also, what is required for an application to be considered as successfully topping the charts. This study's goal is to predict how user reviews, ratings, product size, and cost will impact programs installed. In this work, multiple linear regression is used. In sight of user reviews, ratings, size and price as well as user evaluations, the goal is to gather data on the number of installations done by users.

4. MOTIVATION

The rate of global mobile market expansion has greatly inspired us. Market expansion will present numerous chances and difficulties for business analyst students to increase productivity.

Mobile Industry is expected to develop at a compound annual growth rate of 22.9%, reaching a value of USD337.8 billion by 2027.

The market will be propelled in large part by the swiftly rising adoption of smartphones and tablets along with the high internet penetration in developing countries like China and India. (04.Market Data Forecast, 2022)

5. LITERATURE REVIEW

The Literature survey here outlines preceding researches on play store app analysis, the algorithms and graphs used by them. The writings we present here is the work of many pertinent papers explored by us so by collecting the combination of keywords and snowballing we have improvised our project. The literature review makes us contemplate and understand the earlier innovations related to the project.

- 1. Rimsha Maredia, Google Play Store Analysis Predict the popularity of an app on the Google Play Store using data.

In this study, the classification model was used to forecast how popular an app would be on the Google Play store.

- 2. Rashi Sharma's, analysis of the Google Play Store

In this study, the key factors influencing a user's decision to download an app were identified using an exploratory data analysis methodology.

- 3. The study on Play Store App Analysis from the International Research Journal of Engineering and Technology (IRJET)

In order to forecast which parameter will have the greatest impact on users' decisions to download the app, they implemented a linear regression model with Python software interfaces.

[1] In this paper, they proposed a completely unique and automatic framework IDEA, which aims to spot Emerging App issues effectively supported online review analysis. They evaluated IDEA on six popular apps from Google Play and Apple's App Store, employing the official app changelogs as their ground truth. Feedback from engineers and merchandise managers shows that 88.9% of them think that the identified issues can facilitate app development in practice. To make the topics comprehensible, IDEA labels each topic with the most relevant phrases and sentences based on an effective ranking scheme considering both semantic relevance and user sentiment.

[2] In this context, traditional recommendation techniques are introduced into APPs recommendation. However, different from traditional context, APPs recommendation is a very unique task since people use APPs for different reasons. In this paper, they analyzed user's usage and download behaviors supported a true Android Market data to hunt useful information which may benefit APPs recommendation task. APPs usage, Usage of APPs represents user's preferences and demands. Latent Models like matrix factorization can help to factorize user-item preference matrix into user preference vectors and item feature vectors. In their recommendation algorithm, they utilized user's usage history as user's preferences for APPs.

[3] In this paper, they used Sentiment analysis, Lexicon based sentiment analysis is a method that can be used for determining the sentiment polarization of a review or a comment in the App Store. There are two resources needed in lexicon based sentiment analysis for Indonesian language: machine translation and lexicon resource. In this study, they compared the performance of several different combinations of machine translations and lexicon resources in order to know the best resource combination that can be used in lexicon based sentiment analysis on App Review. The result shows that the combination of Google Translate and SentiWordNet can reach the highest overall accuracy by getting the score 0.72.

[4] They demonstrate an optimization-based aggregation method for ranking extortion and ranking misrepresentation recognition framework for versatile Apps. It is divided into three parts: 1) ranking based evidence, 2) rating based evidence and 3) review based evidence, by demonstrating Apps' ranking, rating and survey practices through measurable theories tests. They used here opinion analysis for finding how much a review is positive or negative. This review score is employed to reinforce the rating score of the user and therefore the emoticons within the reviews or comments. User has provided the rating, review & comments.

[5] In this paper, they represent a large-scale comparative study of cross-platform apps. They mine the characteristics of 80,000 app-pairs (160K apps in total) from a corpus of two .4 million apps collected from the Apple and Google Play app stores. They quantitatively compare their app store attributes, like stars, versions, and costs. They measure the aggregated user-perceived ratings and find many differences across the platforms. Further, they employ machine learning to classify 1.7 million textual user reviews obtained from 2,000 of the mined app-pairs. They also follow up with the developers to know the explanations behind identified differences. they contacted app developers to understand some of the major differences in app-pair attributes such as prices, update frequencies, AUR rates and top rated apps existing only on one platform.

[6] This talk presents results on analysis and testing of mobile apps and app stores, reviewing the work of the UCL App Analysis Group (UCLappA) on App Store Mining and Analysis. The talk also covers the work of the UCL CREST enter on Genetic Improvement, applicable to app improvement and optimization.

[7] In this paper, the google play store is one of the largest and most popular Android app stores. They have used a raw data set of Google Play Store from the Kaggle website. This data set contains 13 different features that can be used for predicting whether an app will be successful or not using different features. They conduct data modeling by using three models: Gaussian Naive Bayes Model, K- nearest neighbor model, and Decision Tree model. They also discovered how different algorithms work in different cases. They found that the Decision tree is easy to visualize and explain the model implementation and it also saves computational power.

6. DATA DESCRIPTION

The dataset contains details of Android applications present on Google Play. The dataset is extracted from Kaggle link. For analysis of the mentioned data, our business case is to locate the best Apps, which we measure by Review and Rating check. There are 13 parameters included that depict each application and an aggregate of 10841 applications. Following variables were initially included:

Table Header	Second Header
App	Name of the App
Category	Category of the app
Rating	Overall user rating of the app out of 5 on the Play Store
Reviews	Number of user reviews for the app
Size	Size of app
Installs	Number of user downloads/installs for the app
Type	Paid or Free
Price	Cost of the App
Content Rating	Age group the app is targeted at
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 Ver	Current version of the app available on Play Store
Android Ver	Minimum required Android Version

We found most popular category of apps on two basis - Number oF Installs and Number of reviews. Personalization wins in former criteria whereas Sports wins in later criteria. This data is good to implement machine learning models which was not a part of this project. It are often

considered as an improvement for future. A more can be done using Last updated variable where month can be separated and clubbed with a lot of other variable in order to insightful information.

7. METHODOLOGY

Our analysis approach is divided into three phases: data extraction, data cleansing, and visualization, and data modeling. In the first step, we collected the raw data from Kaggle. Then we did basic data cleaning and data visualization. After visualizing the data set, we removed some unnecessary features and made it ready for data modeling. Nest we conduct data modeling by using three models: Gaussian Naive Bayes Model, K-nearest neighbor model, and Decision Tree model.

- We have chosen IBM SPSS software for analyzing various relations between the variables in the dataset.

- To define the degree of association between the two variables, we used statistical technique correlation.

- To analyze the direct association between each independent variable and the dependent variable, we have used linear regression method.

- We employed the multiple regression method to examine the association between one dependent variable and several independent variables as a whole. To interpret their association with variations, we also derive a regression equation.

- Using Residuals and Fitted value, we describe the best fit of the independent variable.

Note :

We could also visualize this dataset in Tableau for better understanding of Android Apps between Category, Type, Reviews and Ratings which creates an impact on count of installations by users.

7.1 DATA CLEANING:

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

The raw data consisted of around 10,000 samples as shown below.

➔ **Frequencies**

Statistics

		Rating	Reviews	Size	Installs(K)	Price
N	Valid	9366	10244	9236	10840	10840
	Missing	1474	596	1604	0	0

Fig.1

After removing the duplicate, irrelevant and extreme values, the data set comprised to 7000 samples approximately as shown in the statistics.

Statistics

		Rating	Reviews_1	Size	Installs(K)	Price
N	Valid	7820	7820	7820	7820	7820
	Missing	0	0	0	0	0

Fig.2

To run a frequency distribution, click Analyze, Descriptive Statistics, then Frequencies. Then click on the variable name that you are checking and move it to the Variable box.

Statistics

		Rating	Reviews_1	Size	Installs(K)	Price
N	Valid	7820	7820	7820	7820	7820
	Missing	0	0	0	0	0
Mean		4.175	20.493371	14050.312	9417.79569	1.11449233
Std. Error of Mean		.0061	2.1112702	224.0324	650.789928	.195635111
Median		4.300	.113950	5632.000	100.000000	.000000000
Mode		4.4	.0001	1536.0	1000.00000	.000000000
Std. Deviation		.5421	186.701236	19811.3598	57549.8504	17.3001623
Variance		.294	34857.352	392489977	3.312E+9	299.296
Skewness		-1.759	20.978	2.057	13.083	22.309
Std. Error of Skewness		.028	.028	.028	.028	.028
Kurtosis		5.138	557.236	4.055	198.555	504.675
Std. Error of Kurtosis		.055	.055	.055	.055	.055
Range		4.0	6911.9315	99326.0	999999.999	400.000000
Minimum		1.0	.0001	2.0	.001000000	.000000000
Maximum		5.0	6911.9316	99328.0	1000000.00	400.000000
Sum		32648.4	160258.160	109873443	73647162.3	8715.33000

Fig.3

From the above Statistics, we observe that variables having the minimum and maximum values, mean and several observations with zero missing values.

7.4 HISTOGRAMS:

Frequency Distribution for Ratings

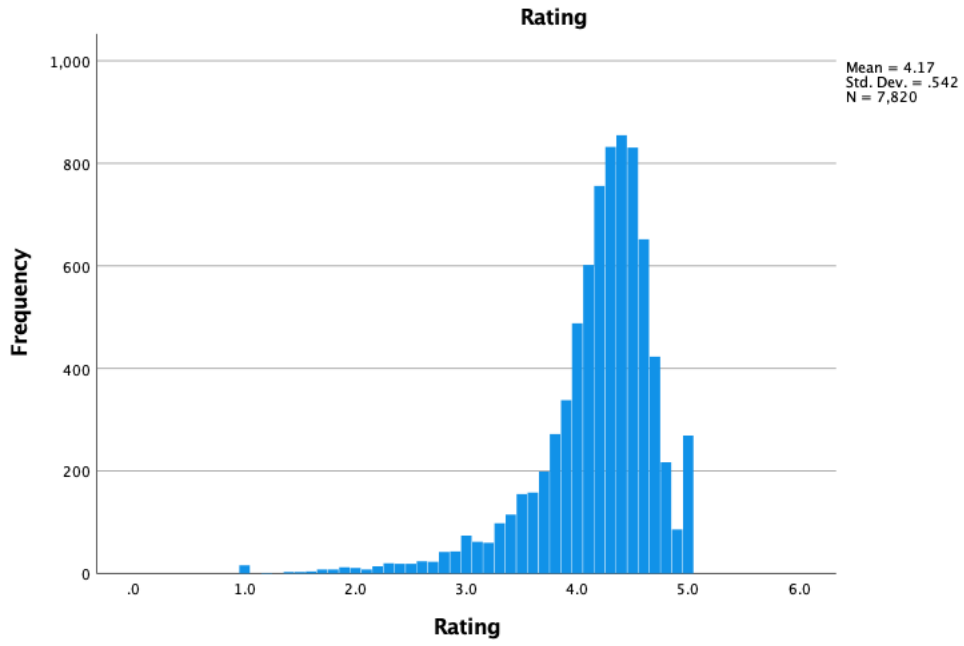

Fig.4

The Ratings are distributed by frequency; we get the histogram plot it tells that the data is normally distributed. The Rating data is very unique data to done the analysis.

Frequency Distribution for Reviews

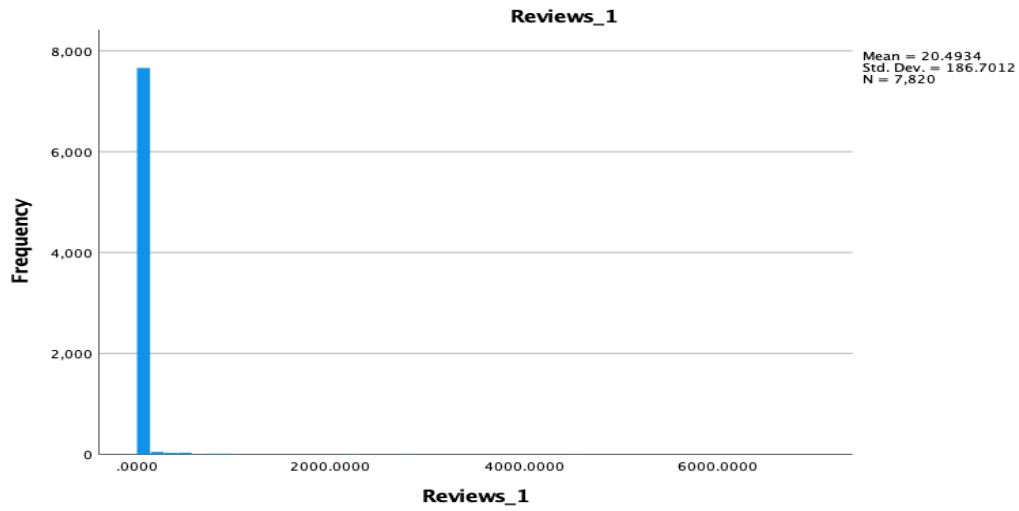

Fig.5

Here also Reviews distributed by frequency, but in this histogram the data is not normally distributed. Because in the data most of the values are zero's. So the data is quite difficult while doing the analysis.

Frequency Distribution for Size

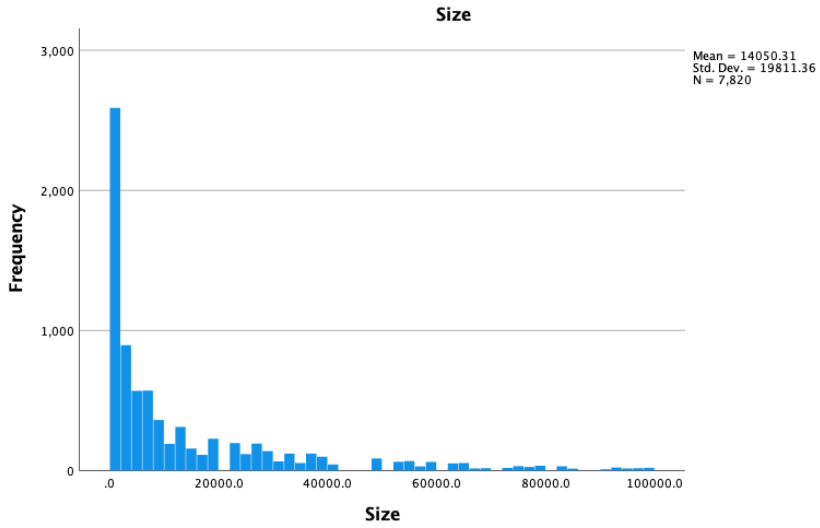

Fig.6

The Size is distributed by frequency; we get the histogram plot it tells that the data is not normally distributed. Whether Size data is skewed downward. In this data have more zero's. So we changed the data into different sizes.

Frequency Distribution for Installs

Installs are dispersed according to frequency here as well, although the data in this histogram is not evenly distributed. Because majority of the values in the data are zeros. Analyzing the data might be fairly challenging. Therefore, we changed the Installs data's quantities.

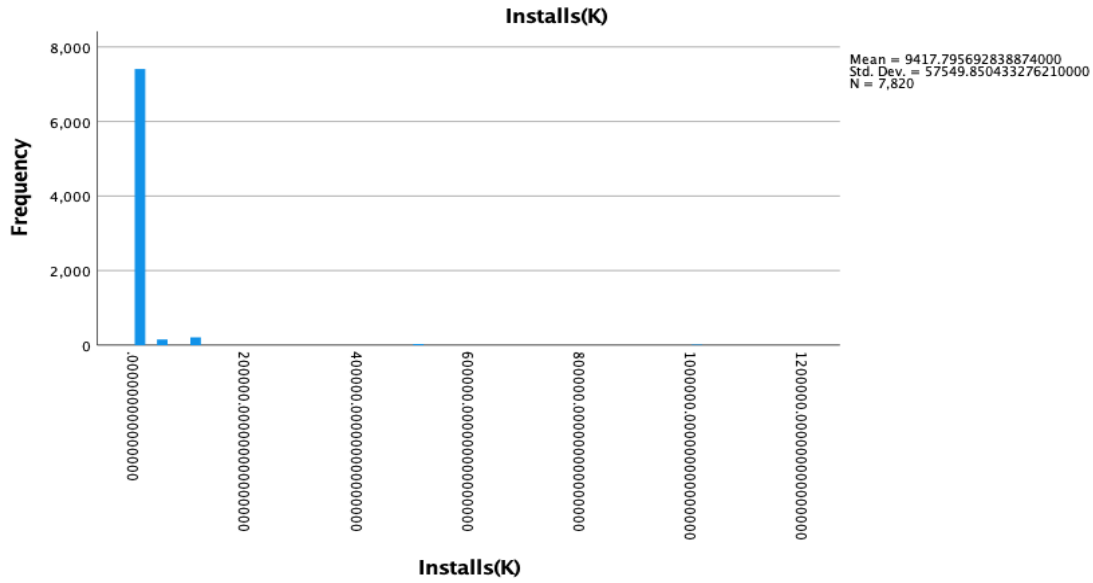

Fig.7

Frequency Distribution for Price

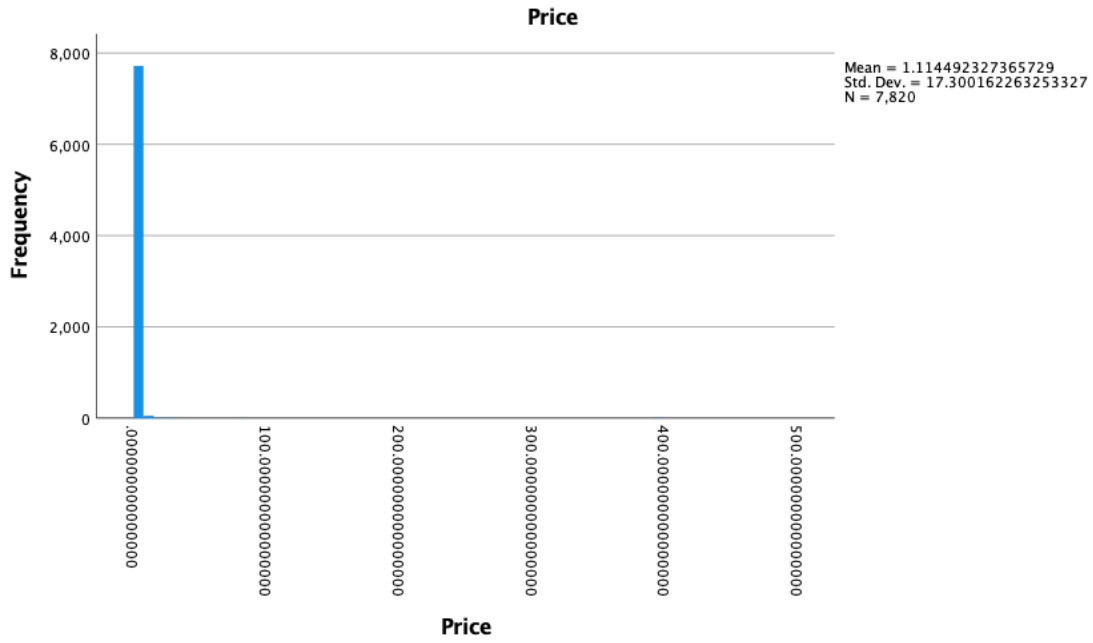

Fig.8

Although the data in this histogram is not normally distributed, the price is distributed here according to frequency. Only two categories—paid as 1 and free as 0—are present in this price data. The majority of consumers only utilize free apps.

7.5 DATA PRE-PROCESSING:

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labeling raw data into a form suitable for machine learning (ML) algorithms and then exploring and visualizing the data.

To obtain descriptive statistics for nominal variables, click Analyze, Descriptive Statistics, Frequencies. Move the nominal variables that you want to examine into the Variables box. Then click on the Statistics button. Check the following boxes:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
Rating	7820	4.0	1.0	5.0	4.175	.0061	.5421	.294
Reviews_1	7820	6911.9315	.0001	6911.9316	20.493371	2.1112702	186.701236	34857.352
Size	7820	99326.0	2.0	99328.0	14050.312	224.0324	19811.3598	392489977
Installs(K)	7820	999999.999	.001000000	1000000.00	9417.79569	650.789928	57549.8504	3.312E+9
Price	7820	400.000000	.000000000	400.000000	1.11449233	.195635111	17.3001623	299.296
Valid N (listwise)	7820							

Fig.9

7.6 VARIABLE SELECTION:

Variable selection means choosing among many variables which to include in a particular model, that is, to select appropriate variables from a complete list of variables by removing those that are irrelevant or redundant.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Price, Reviews_1, Rating, Size[b]	.	Enter

a. Dependent Variable: Installs(K)

b. All requested variables entered.

Fig.10

From the variable selection, The Target variable is Installs(k). The Independent variables are Ratings, Reviews_1, Size and Price.

7.5 MULTIPLE REGRESSION ANALYSIS:

Multiple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value. Each predictor value is weighed, the weights denoting their relative contribution to the overall prediction.

(1)Y=a+b1X1+b2X3+…+bnXn

Here Y is the dependent variable, and $X_1,\ldots,X_n$ are the *n* independent variables. In calculating the weights, a, $b_1,\ldots,b_n$, regression analysis ensures maximal prediction of the dependent variable from the set of independent variables. This is usually done by least squares estimation.

1. Click on Analyze\Regression\Linear.
2. Move your continuous dependent variable into the Dependent box.
3. Move your independent variables into the Independent box.
4. For Method make sure Enter is selected.

7.6 CORRELATION:

The bivariate Pearson Correlation produces a sample correlation coefficient, *r*, which measures the strength and direction of linear relationships between pairs of continuous variables. By extension, the Pearson Correlation evaluates whether there is statistical evidence for a linear relationship among the same pairs of variables in the population, represented by a population correlation coefficient, ρ (“rho”). The Pearson Correlation is a parametric measure.

Correlations

		Installs(K)	Rating	Reviews_1	Size	Price
Pearson Correlation	Installs(K)	1.000	.047	.530	.124	-.011
	Rating	.047	1.000	.052	.043	-.021
	Reviews_1	.530	.052	1.000	.093	-.007
	Size	.124	.043	.093	1.000	-.013
	Price	-.011	-.021	-.007	-.013	1.000
Sig. (1-tailed)	Installs(K)	.	<.001	.000	<.001	.176
	Rating	.000	.	.000	.000	.029
	Reviews_1	.000	.000	.	.000	.267
	Size	.000	.000	.000	.	.132
	Price	.176	.029	.267	.132	.
N	Installs(K)	7820	7820	7820	7820	7820
	Rating	7820	7820	7820	7820	7820
	Reviews_1	7820	7820	7820	7820	7820
	Size	7820	7820	7820	7820	7820
	Price	7820	7820	7820	7820	7820

Fig.11

7.7 MODEL SUMMARY:

The model summary table reports the strength of the relationship between the model and the dependent variable. R, the multiple correlation coefficient, is the linear correlation between the observed and model-predicted values of the dependent variable. Its large value indicates a strong relationship.

Model Summary[b]

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.535[a]	.287	.286	48618.5724	.287	785.144	4	7815	.000	1.675

a. Predictors: (Constant), Price, Reviews_1, Rating, Size

b. Dependent Variable: Installs(K)

Fig.12

According to the Model Summary, R-square is too low (0.287). Therefore, the analysis and the model do not fit well. We believe these incorrect results are the result of numerous data values of various sizes. The Durbin Watson test is 1.675 < 2. So it is positive correlation.

7.8 ANOVA :

Analysis of Variance, i.e. ANOVA in SPSS, is used for examining the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables. Essentially, ANOVA in SPSS is used as the test of means for two or more populations.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.424E+12	4	1.856E+12	785.144	.000[b]
	Residual	1.847E+13	7815	2.364E+9		
	Total	2.590E+13	7819			

a. Dependent Variable: Installs(K)

b. Predictors: (Constant), Price, Reviews_1, Rating, Size

Fig.13

- From the ANOVA results, the F-value is too high (785.144). It is calculated by dividing two mean squares. The higher the F-value in an ANOVA, the higher the variation between sample means relative to the variation within the samples.
- The higher the F-value, the lower the corresponding p-value.
- If the p-value is below a certain threshold, the value is 0.00 < 0.05. we can reject the null hypothesis of the ANOVA and conclude that there is a statistically significant difference between group means.

7.9 COEFFICIENTS:

Regression coefficients are estimates of the unknown population parameters and describe the relationship between a predictor variable and the response. In linear regression, coefficients are the values that multiply the predictor values.

The sign of each coefficient indicates the direction of the relationship between a predictor variable and the response variable.

- A positive sign indicates that as the predictor variable increases, the response variable also increases.
- A negative sign indicates that as the predictor variable increases, the response variable decreases.

The coefficient value represents the mean change in the response given a one unit change in the predictor.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-4123.407	4280.860		-.963	.335		
	Rating	1728.193	1016.596	.016	1.700	.089	.995	1.005
	Reviews_1	160.905	2.961	.522	54.336	.000	.989	1.011
	Size	.217	.028	.075	7.779	<.001	.990	1.010
	Price	-18.484	31.791	-.006	-.581	.561	.999	1.001

a. Dependent Variable: Installs(K)

Fig.14

Estimated regression equation for Installs:

The regression equation from above unstandardized coefficient data is

Installs = -4123.407 + 1728.193(Rating) + 160.905(Reviews_1) + 0.217(size) – 18.484(Price)

A variance inflation factor (VIF) is a measure of the amount of multi-collinearity in regression analysis. Multi-collinearity exists when there is a correlation between multiple independent variables in a multiple regression model. A VIF of 1 means that there is no correlation among the target predictors and the remaining predictor variables.

7.10 RESIDUAL STATISTICS:

In statistical models, a residual is the difference between the observed value and the mean value that the model predicts for that observation. Residual values are especially useful in regression and ANOVA procedures because they indicate the extent to which a model accounts for the variation in the observed data.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-5490.1392	1115649.25	9417.79569	30812.8049	7820
Residual	-626484.88	993080.750	.000000000	48606.1348	7820
Std. Predicted Value	-.484	35.902	.000	1.000	7820
Std. Residual	-12.886	20.426	.000	1.000	7820

a. Dependent Variable: Installs(K)

Fig.15

From the residual statistics, the residual and predicted values of the variables with different measurements as mean, standard deviation, maximum and minimum values.

Coefficient Correlations[a]

Model			Price	Reviews_1	Rating	Size
1	Correlations	Price	1.000	.005	.021	.011
		Reviews_1	.005	1.000	-.048	-.091
		Rating	.021	-.048	1.000	-.038
		Size	.011	-.091	-.038	1.000
	Covariances	Price	.001	4.588E-7	.001	9.964E-9
		Reviews_1	4.588E-7	8.769E-6	.000	-7.518E-9
		Rating	.001	.000	1.033	-1.077E-6
		Size	9.964E-9	-7.518E-9	-1.077E-6	7.782E-10

a. Dependent Variable: installs_k_k

Fig.16

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Rating	Reviews_1	Size	Price
1	1	2.486	1.000	.00	.00	.01	.06	.00
	2	1.003	1.575	.00	.00	.21	.00	.78
	3	.968	1.603	.00	.00	.77	.00	.22
	4	.534	2.157	.00	.00	.02	.93	.00
	5	.008	17.307	.99	.99	.00	.00	.00

a. Dependent Variable: installs_k_k

Fig.17

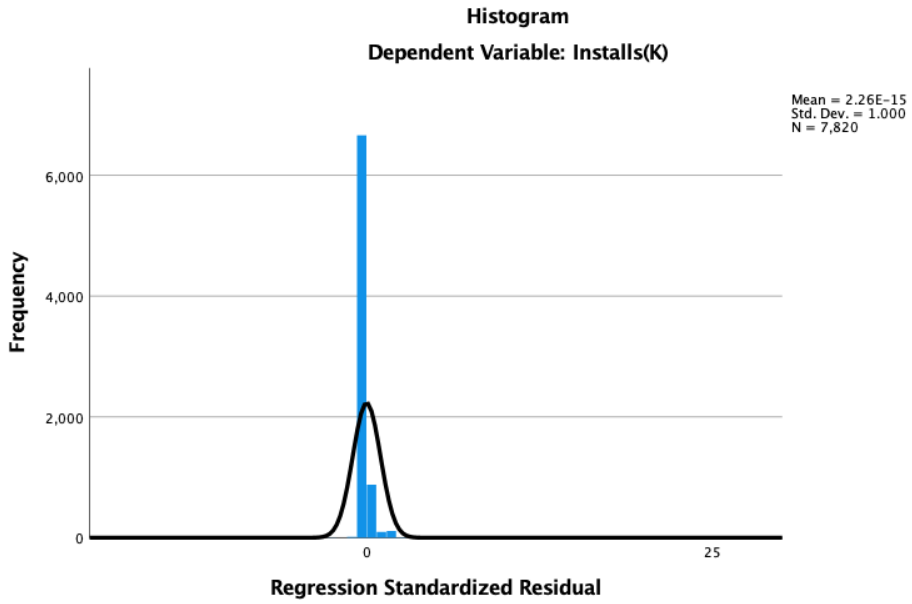

Fig.18

The above diagram is the Histogram of the regression standardized residual as dependent variable (Installs) distributed by frequency with mean = 2.26EE-15, Standard deviation = 1 and N = 7820. Here graph is not normally distributed.

7.11 HOMOSCEDASTICITY PLOT:

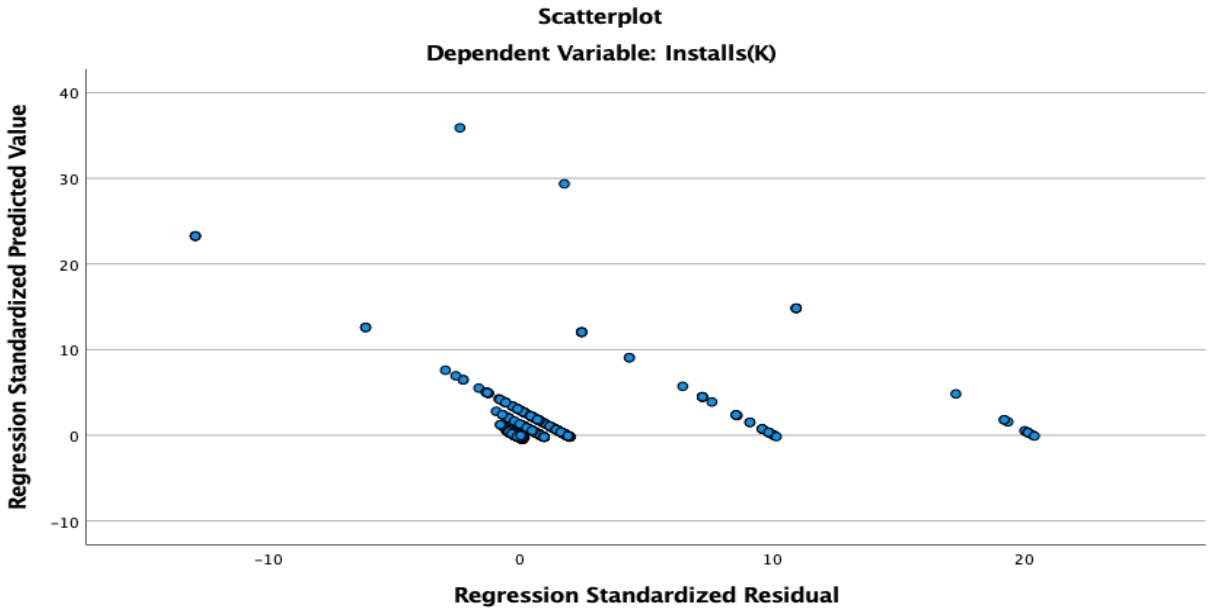

Fig.19

The above figure is Scatterplot of the regression standardized residual value v/s regression standardized predicted value is distributed by frequency with Installs as dependent variable. Residual scatter plots provide a visual examination of the assumption homoscedasticity between the predicted dependent variable scores and the errors of prediction. It shows there is not a good relation between the variables.

7.12 NORMAL P-P PLOT:

- A normal probability plot is the method by which the assumptions of normality and **homogeneity of variance** are tested in multiple regression. A normal probability plot is also commonly known as a P-P plot.
- The p-value is a probability that measures the evidence against the null hypothesis. A smaller p-value provides stronger evidence against the null hypothesis. Larger values for the Anderson-Darling statistic indicate that the data do not follow the normal distribution.
- Here the p-value is 7820 is high value, so the data is not normally distributed.

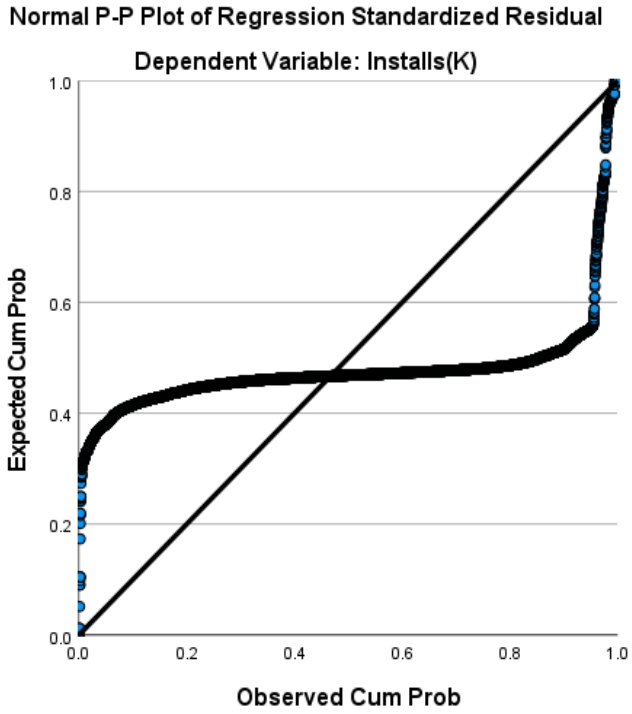

Fig.20

7.13 SCATTER PLOTS:

A scatterplot is a type of plot that we can use to display the relationship between two variables. It helps us visualize both the direction (positive or negative) and the strength (weak, moderate, strong) of the relationship between the two variables.

1. Click Graphs -> Legacy Dialogs -> Scatter/Dot
2. Select "Simple Scatter"
3. Click "Define"
4. Click "Reset" (recommended)
5. Select the predictor/independent variable and move it into the "X Axis" box
6. Select the criterion/dependent variable and move it into the "Y Axis" box
7. Select "Titles" to add a title (recommended)
8. Select "OK"

Plot between Installs and Rating

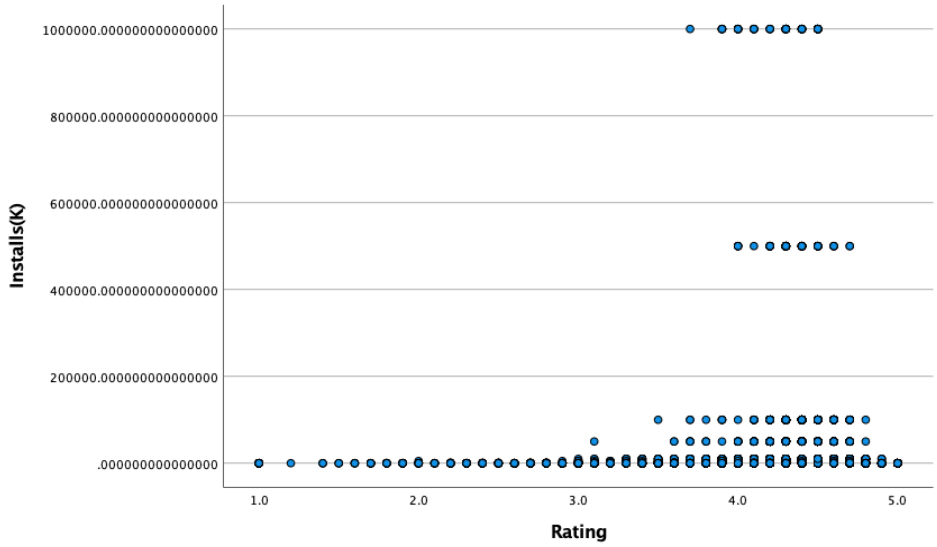

Fig.21

Plot between Installs and Reviews

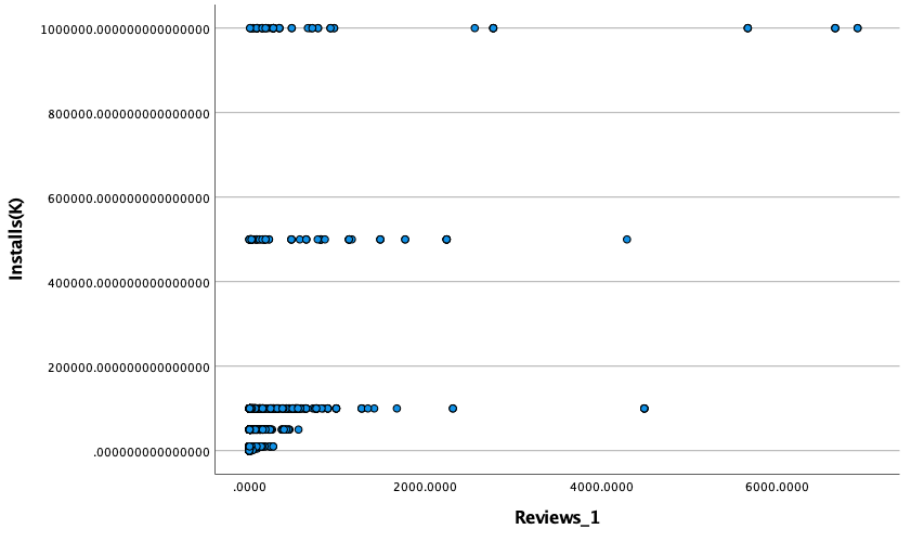

Fig.22

Plot between Installs and Size

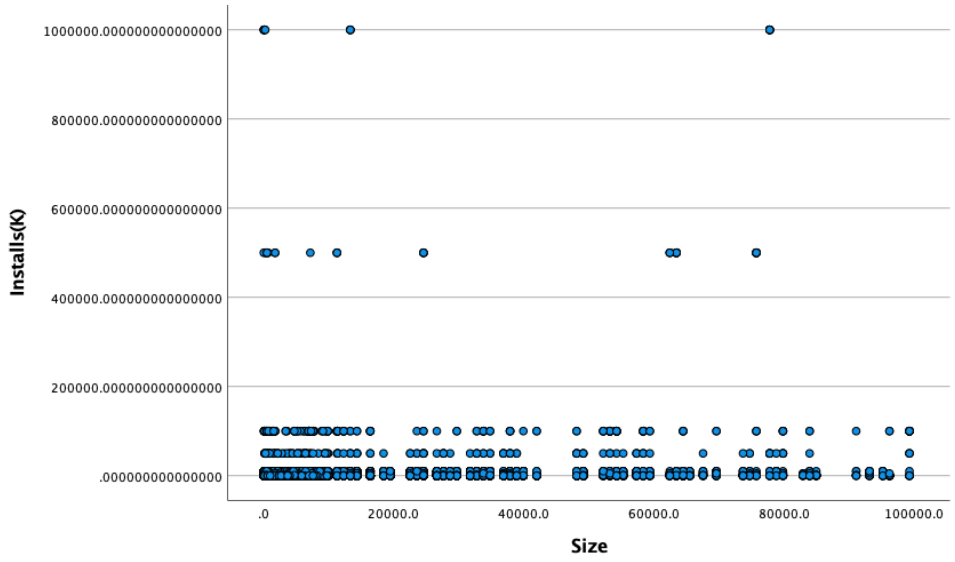

Fig.23

Plot between Installs and Price

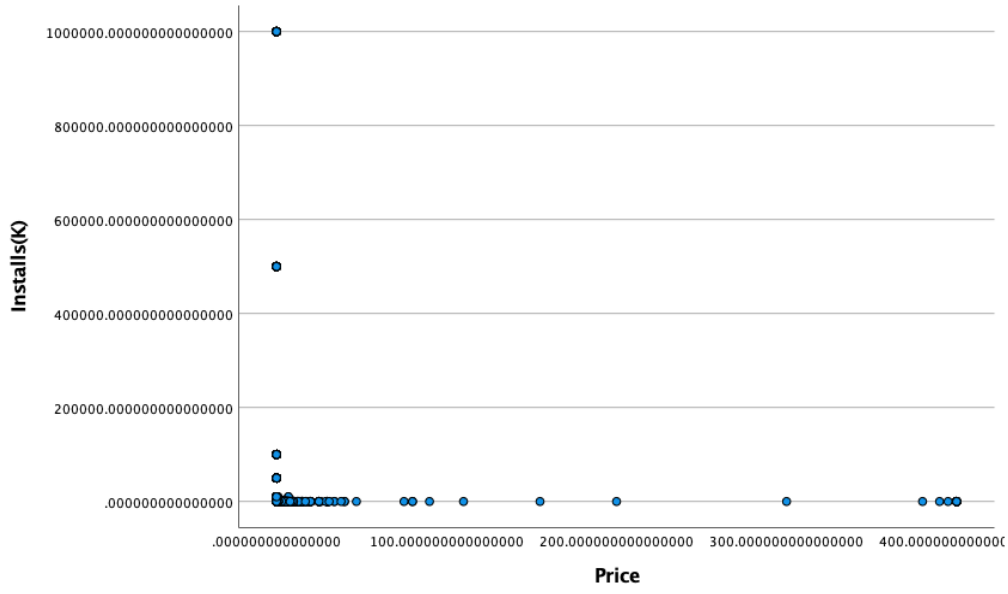

Fig.24

From the above all scatterplots of the Dependent variable v/s Independent variables it tells that the data is not even normally distributed and graphs does not show better relationship between the variables.

8. PROJECT OUTCOME

- From the Regression analysis, we clearly state that Reviews can reject the Null hypothesis.
- From the obtained Regression equation, we could interpret that the positive impact on installations is impacted by Ratings, Reviews and Size.
- Price is creating negative impact on installations.
- We Clearly states that Ratings can makes the more impact on the maker's decision while installing the apps from play store.
- Reviews and Size also shows the impact on maker's decision but less than the ratings and Price shows the negative impact on it.

9. LIMITATIONS

- Because of the huge amount of data in this dataset, data cleaning took more time than anticipated, and the resulting conclusions didn't accurately reflect whether the model was a good fit or not.
- From the scatterplots, we noticed that the target and independent variables didn't interact much and there was more variation between them.
- There isn't much difference in the results when we performed the regression analysis using different approaches, such as step-wise and backward.
- While performing the regression analysis, we noticed that the data in the file should be in simple sizes then it takes easy to do any analysis.

10. CONCLUSION

Thus the app development companies could decide what application should be developed and they can also see the prediction of their developed application. In this they also get to see the categorized reviews of all the application in one interface which will help them decide which app is liked by the users and which apps need to be developed more. The dataset contains immense possibilities to improve business values and have a positive impact. It is not limited to the problem taken into consideration for this project.

- My goal across the research was to examine the information and identify the key factors that influence people' decisions to download applications. After analyzing the data, I came to the conclusion that users prefer free apps more. Since most of the apps in the Play Store are roughly the same size, size doesn't really influence their decisions.
- Finally, in comparison to other characteristics, the Reviews parameter demonstrates the biggest influence on the target variable. From the outcomes, we deduced that.

11. FUTURE WORK

In future we could display live downloads and top applications of the play store. We could add a system that would create application on its own by using the data set and creating the best user interface by the highly rated apps. ANALYSIS USING EMOTICON", International Conference

Future work can also include:

1.Optimization of the pie-charts There are multiple domains in the same slice. The multiple domains could be separated and added to the same field to get a more detailed version of this pie chart.

2. Prediction of the number of reviews and installs by using the regression model.

3. Identifying the categories and stats of the most installed apps.

4. Exploring the correlation between the size of the app, the version of Android, etc. on the number of installs.

Many other interesting possibilities can be explored using this dataset.

- The dataset contains immense possibilities to improve business values and have a positive impact. It is not limited to the problem taken into consideration for this project. In this they can see the categorized reviews of all the application in an interface which will help them decide which app is most liked by users.
- In the future, we might show the most popular programs and live downloads from the Play Store. A system that would automatically create applications utilizing the data set and the best user interface from the highly rated apps might be included.

12. REFERENCES

- Kaggle.com.(2018). Google Play Store Apps.[online] https://www.kaggle.com/datasets/lava18/google-play-store-apps?select=googleplaystore.csv
- Amit Chile, Dr. P. R. Gundalwar.(2019). Analysis of Google Play Store Application.[online] https://www.ijraset.com/fileserve.php?FID=24134
- R. P. Rajeswari, K. Juliet, and Aradhana, "Text Classification for Student Data Set using NaiveBayes Classifier and KNN Classifier," Int. J. Comput. Trends Technol., vol. 43, no. 1, pp. 8–12, 2017. https://doi.org/10.14445/22312803/ijctt-v43p103
- Grover, S. 3 apps that failed (and what they teach us about app marketing). [online] https://blog.placeit.net/apps-fail-teach-us-appmarketing/
- "Mining and Analysis of Apps in Google Play," Pro- ceedings of the 9th International Conference on Web Information Systems and Technologies, 2013.
- Google play store: number of apps 2018(2018). [online] https://www.statista.com/ statistics/266210/number-of-available-applications-in- the-google-play-store/
- Amit Chile, Dr. P. R. Gundalwar.(2019). Anal- ysis of Google Play Store Application.[online] http://ijraset.com/fileserve.php?FID=24134 [Accessed 3 Mar. 2020]
- T. Denoeux, M. Skarstein-Bjanger, Induction of deci- sion trees from partially classified data, in: Proceedings of the 2000 IEEE International Conference on Systems, Man and Cybernetics (SMC'00), IEEE, Nashville, TN, 2000, pp. 2923–2928.
- Harman, M., Jia, Y., and Zhang, Y. (2012). App store mining and analysis: Msr for app stores. In 2012 9th IEEE Working Conference on Mining Software Repos- itories (MSR),pages 108–111.
- R. P. Rajeswari, K. Juliet, and Aradhana, "Text Classification for Student Data Set using Naive Bayes Classifier and KNN Classifier," Int. J. Com- put. Trends Technol., vol. 43, no. 1, pp. 8–12, 2017. https://doi.org/10.14445/22312803/ijctt-v43p103
- Jong, J. (2011). Predicting rating with sentiment anal- ysis. [online] http://cs229.stanford. edu/proj2011/Jong- PredictingRatingwithSentimentAnalysis.pdf.
- H. G. Schnack, M. Nieuwenhuis, N. E. van Haren, L. Abramovic, T. W. Scheewe, R. M. Brouwer, H. E. Hulshoff Pol, and R. S. Kahn, "Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophre- nia, bipolar disorder and healthy subjects," NeuroIm- age, vol. 84, pp. 299–306, jan 2014.