Final Report

Analyzing Key Factors and Predicting App Success on the Google Play Store

Table of Contents

1. Introduction and Background

The advent of mobile technology has revolutionized the way individuals engage with digital content, and mobile applications have become integral to our daily lives. The Google Play Store stands as one of the largest platforms for distributing and accessing mobile applications, offering a diverse array of apps catering to various needs and interests. The Google Play Store Apps dataset, sourced from Kaggle, provides a comprehensive collection of information on a myriad of mobile applications available on the Google Play Store. This dataset encompasses a wide range of details, including app categories, user ratings, the number of installations, content ratings, pricing, and more. Understanding the patterns and trends within this dataset offers valuable insights into the dynamics of the mobile app ecosystem.
To achieve success on the Google Play Store, it is crucial to understand the factors that contribute to app growth and visibility. This includes features such as app quality, user ratings and reviews, the number of installations, and the frequency of app updates. Businesses aiming to succeed in the Google Play Store must carefully consider each of these factors and implement a holistic strategy to improve their app's visibility, downloads, and user engagement.

In our analysis, we aim to drill down on which factors are the most crucial for success on the Google Play Store.

1.1 Research Question and Motivation

Our research aims to identify the relationship between application rating and the number of downloads. Furthermore, we aim to identify the most critical factors regarding app success on the Google Play Store. We believe that conducting this research will allow businesses and individual entrepreneurs to have greater insight into the dynamics of the Google Play Store.

1.	What is the relationship between application rating and the number of installations? Is it a negative or positive relationship?
2.	Given that we have defined success, can we identify which factors majorly influence the success of the app?
3.	Given that we have defined success, can we predict the success of an app based on attributes such as reviews, size, installs, and price?
4.	Can we segment categories of apps based on user age ratings of the apps?

2. Dataset Description and Variable Introduction

For our data, we have retrieved a dataset from kaggle.com.

Table 2.1: Dataset Description

Variable	Description
app_name	The name of the mobile application.
category	The category or genre to which the app belongs (e.g., "Games," "Productivity," "Travel").
rating	The user rating of the app.
reviews	The number of user reviews for the app.
size	The size of the app in terms of storage.
installs	The number of times the app has been installed.
type	Whether the app is "Free" or "Paid."
price	The price of the app if it's a paid app.
content_rating	The content rating for the app (e.g., "Everyone," "Teen," "Mature 17+").
genres	Additional categorization of the app based on its features.
last_updated	The date when the app was last updated.
current_ver	The current version of the app.
android_ver	The minimum Android version required to run the app.

3. Data Summary

3.1 Data Summary Statistics

Our dataset consists of 12 predictors, out of which 5 are quantitative predictors and 7 are categorical. There are 7724 observations.

Table 3.1: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Max
Rating	7,724	4.174	0.545	1.000	5.000
Reviews	7,724	294,862.700	1,863,815.000	1	44,893,888
Size	7,724	22.967	23.450	0.008	100.000
Installs	7,724	8,423,109.000	50,154,239.000	1	1,000,000,000
Price	7,724	1.128	17.407	0.000	400.000

3.2 Correlation

Table 3.2: Correlation Matrix

	Rating	Reviews	Size	Installs	Price
Rating	1	0.080	0.084	0.053	-0.021
Reviews	0.080	1	0.240	0.626	-0.010
Size	0.084	0.240	1	0.163	-0.026
Installs	0.053	0.626	0.163	1	-0.011
Price	-0.021	-0.010	-0.026	-0.011	1

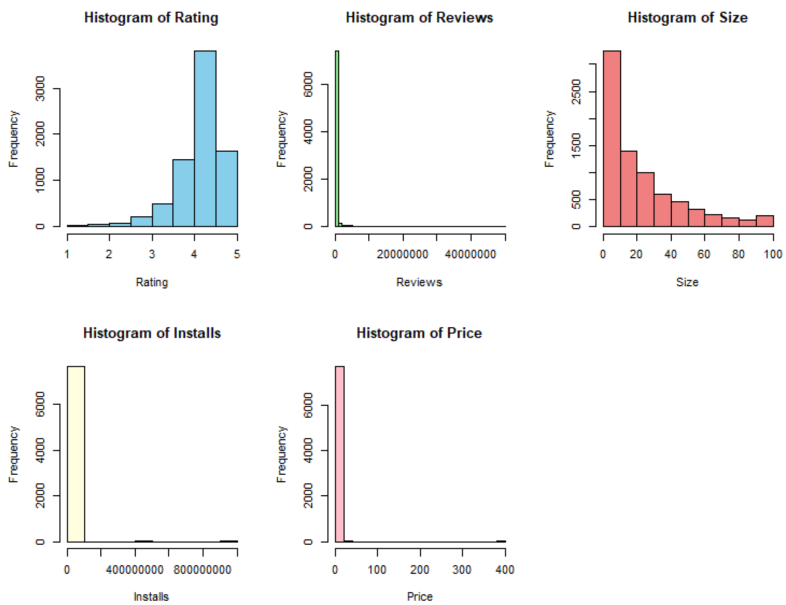
The most significant correlations in the matrix shown above are the correlation between the number of reviews and the number of installs (0.626), the number of reviews and the size of the app (0.240), and the number of installs and the size of the app (0.163). From this it can be inferred, that the greater the number of reviews and the greater the size of the app, the higher the number of app installs will be.

Some weaker correlations also exist, as can be seen in the matrix, such as the correlation between rating and the number of installs (0.053). The negative correlations in this dataset are also quite weak such as the correlation between price and number of installs (-0.011).

3.3 Histogram of Variables

Histograms in Graph 3.1 visualize the distribution of various attributes within the dataset. The histograms for the variables in our dataset are shown above. The 'Installs' histogram reveals the frequency of app installations, demonstrating a skewed distribution with a majority of apps falling in the lower installation count range. Only a small percentage of mobile applications on the Google Play Store have a billion downloads and so the histogram is skewed.

Graph 3.1: Histogram of Variables

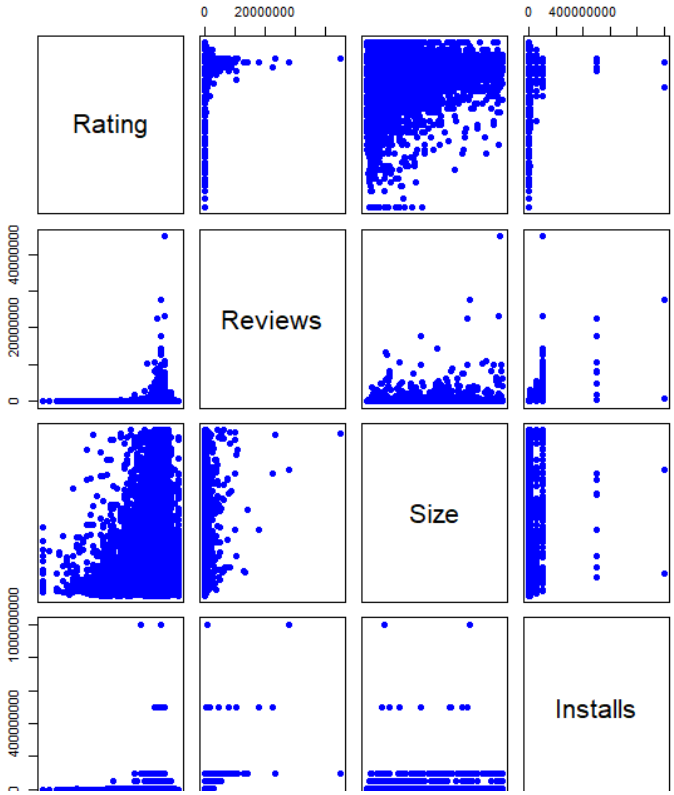


The 'Rating' histogram illustrates the distribution of app ratings, showcasing a generally positively skewed pattern with a concentration around high ratings. The 'Reviews' histogram indicates that most apps have relatively low review counts, but a few outliers have significantly higher review numbers. The 'Size' histogram exhibits a varied distribution of app sizes, with a concentration around smaller sizes. Finally, the 'Price' histogram demonstrates that the majority of apps are free, with a few high-priced apps. These visual representations offer insights into the underlying patterns and characteristics of the data, aiding in a comprehensive understanding of the distribution of key variables within the dataset.

3.4 Scatterplot of Variables

The scatterplot in Graph 3.2 shows that there is a positive correlation between ratings and reviews. This means that installations with more reviews tend to have higher ratings. However, there is a weak negative correlation between rating and size. This means that mobile apps that are larger tend to have lower ratings. In regards to ratings and number of installations, there is a weak positive correlation.

Graph 3.2: Scatterplot of Variables



There is a weak negative correlation between rating and price. This means that mobile apps that are more expensive tend to have lower ratings. There is a positive correlation between reviews and size. There also seems to be a positive correlation between reviews and installs. This reiterates the idea that apps with more reviews tend to have more installs. There is also a weak negative correlation between reviews and price. This means that mobile apps that are more expensive tend to have fewer reviews. Furthermore, there is a positive correlation between size and installs and between size and price. This means that mobile apps that are larger tend to be more expensive and are likely to have more reviews. Finally, there is a weak positive correlation between installs and price.

4. Data Mining Method Description

4.1 Analysis Approach

Correlation analysis, regression analysis, random forest, and clustering are all statistical methods that can be used to analyze factors contributing to app success.

Correlation analysis can be used to measure the strength and direction of the relationship between two variables. For example, we can use correlation analysis to measure the correlation between app rating and the number of installs.

Regression analysis can be used to model the relationship between one or more independent variables (e.g., app category, content rating, size, price, installs) and a dependent variable (e.g., app rating). This can be used to predict the value of the dependent variable based on the values of the independent variables. For example, we can use regression analysis to build a model to predict the number of installs for a new app based on its features.

Random Forest is an ensemble learning algorithm that combines multiple decision trees. It introduces randomness in data sampling and feature selection to enhance robustness and prevent overfitting. Widely used for classification and regression tasks, it provides high accuracy and features important insights.

Clustering can be used to segment users into different groups based on their app preferences. For example, we can use clustering to segment users into groups that are more likely to install games, productivity apps, or social media apps.

4.2 Data Cleaning

In our dataset, the size of the app is either a megabyte or kilobyte. We have converted all the observations to megabytes. We have also converted columns like installs, price, reviews, and size to numeric by removing the non-numeric characters. We have removed the duplicated rows if any and also removed the rows with null values.

4.3 Implementation of Method

4.3.1 Simple Linear Regression Model

Simple linear regression is a statistical modeling technique used to examine the relationship between a dependent variable and one or more independent variables. In essence, it aims to find the best-fitting linear equation that predicts the dependent variable based on the values of the independent variables.

Analysis of the Relationship Between Application Rating and Number of Installations

In the context of our analysis, a linear regression model serves as a foundational tool for exploring the relationship between the Application Rating(Dependent Variable) and the Number of Installations(Independent Variable) in the Google Play Store dataset.
The model equation is as follows:

- $\text{Rating} = \beta 0 + \beta 1 \times \text{Number of Installations}$

Table 4.1: Simple Linear Regression Results

	Dependent variable:
	Rating
Installs	0.000***
	(0.000)
Constant	4.169***
	(0.006)
Observations	7,724
R2	0.003
Adjusted R2	0.003
Residual Std. Error	0.544 (df = 7722)
F Statistic	21.502*** (df = 1; 7722)
Note:	*p<0.1; **p<0.05; ***p<0.01

Coefficients:

Intercept Interpretation:

The intercept, estimated at 4.17, represents the predicted application rating when the number of installations is zero. However, since installations cannot be negative, the intercept's practical interpretation is limited.

Installs Coefficient Interpretation:

The coefficient for "Installs" is approximately 0.0000000005722, indicating the estimated change in the application rating for a one-unit increase in the number of installations. The positive coefficient and low p-value (0.00000359) suggest that, on average, as the number of installations increases, the application rating also increases.

Statistical Significance:

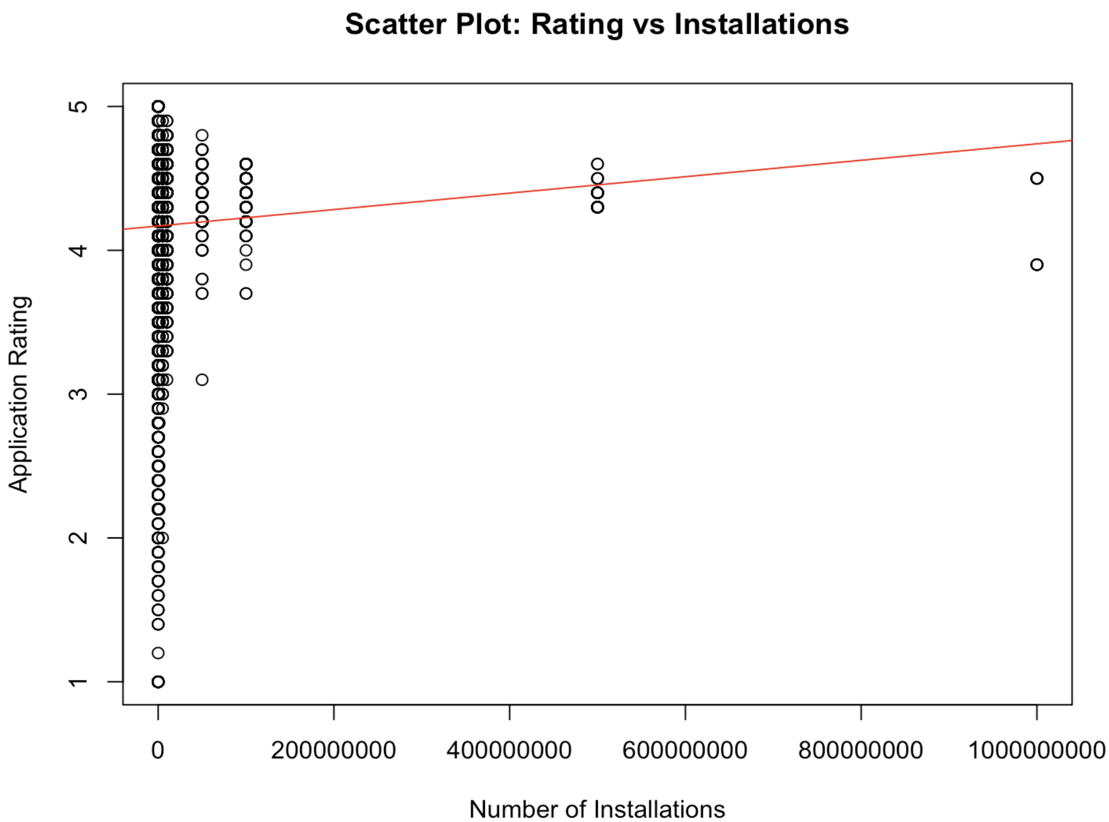
Both the intercept and Installs coefficient have very low p-values (<<0.001), indicating their statistical significance. This suggests that both intercept and install coefficients are unlikely to be zero, supporting their role in explaining the variability in application ratings.

R-squared Interpretation:

The R-squared value is 0.002777, representing the proportion of variance in application ratings explained by the number of installations. In practical terms, this indicates that only a very small percentage (0.28%) of the variability in application ratings is explained by the number of installations. The model's ability to predict application ratings based on the number of installations is limited.

The linear regression analysis of the relationship between application rating and the number of installations reveals a statistically significant but very weak positive relationship. The coefficient for the "Installs" variable is positive (0.0000000005722), indicating that, on average, as the number of installations increases, the application rating also increases. However, the small coefficient value and low R-squared values (0.002777) suggest that the model explains only a minimal proportion of the variability in application ratings based on the number of installations. Overall, while a positive relationship is observed, its practical significance is limited, and other factors likely influence application ratings to a greater extent.

Graph 4.1: Scatter Plot: Rating vs Installations



The scatter plot shown in Graph 4.1 illustrates the relationship between application ratings and the corresponding number of installations in the Google Play Store dataset. The scatter plot reveals an upward-sloping trend, indicating a positive correlation between application ratings and the number of installations. As the application ratings increase, there is a tendency for higher numbers of installations. The straight and upward orientation of the abline (regression line) further emphasizes the positive correlation. This line represents the best-fitting linear model based on the data, suggesting a consistent increase in installations with higher ratings. The positive slope of the regression line aligns with the intuition that well-rated applications tend to attract more installations. Users may be more inclined to download and install applications with higher ratings, contributing to this observed trend. While the trend

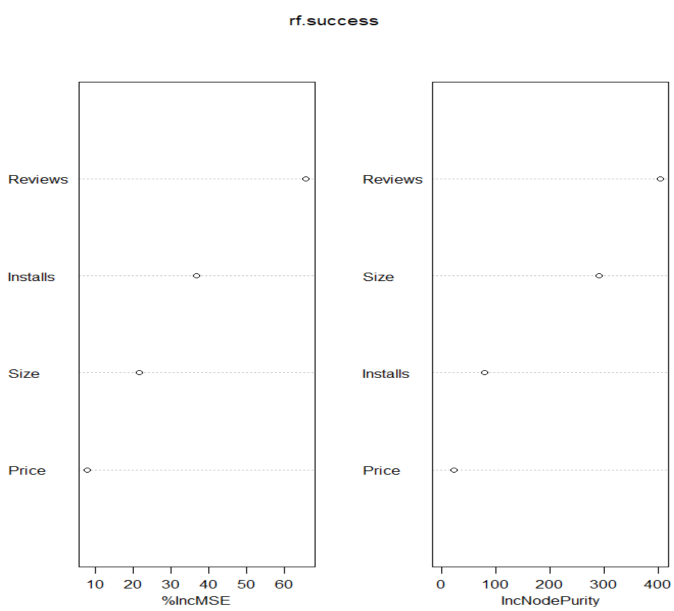
is positive, it's essential to note the dispersion of data points around the regression line. The spread indicates some variability, and the strength of the correlation may be influenced by other factors not considered in the linear model. Outliers can have a notable impact on statistical analyses, including linear regression models. In this context, they represent instances where the number of installations does not conform to the expected trend based on the application rating.
In summary, the scatter plot, coupled with the upward-sloping regression line, provides visual evidence of a positive relationship between application ratings and installations. The straight and upward orientation of the line suggests a consistent and linear growth in installations with increasing ratings.

4.3.2 Random Forest

In the case of this prediction analysis, application success was defined as having a rating greater than the mean rating in the dataset. We used supervised learning methods such as Random Forest as they can capture complex relationships, reduce overfitting, and assess variable importance. This model provides us with insights into the factors that influence mobile application success on the Google Play Store. The model implements evaluation metrics such as MSE, confusion matrix, and accuracy rates on training and testing sets. These metrics were able to provide the results that indicate that the number of reviews has a big influence in predicting success while suggesting that app size and app rating are comparatively less influential factors. The above is illustrated in the below figure.

In this model, we defined success as mobile applications having a user rating greater than or equal to the mean rating in the dataset (4.174). As indicated by Graph 4.2, the number of reviews is the biggest factor in predicting success.

Graph 4.2: Random Forest



4.3.3 Multiple Linear Regression

The goal of multiple linear regression is to model the relationship between a dependent variable (outcome variable) and two or more independent variables (predictors). We build the model by using the lm function in R. The rating of an app is often considered an important metric because it serves as a quick and accessible measure of user satisfaction, so we selected Rating as a dependent variable, and Reviews, Size, Installs, and Price as independent variables, to model the relationship between them.

Table 4.2: Multiple Linear Regression Results

	Dependent variable:
	Success
Installs	**0.000**
	(0.000)
Reviews	**0.00000*****
	(0.000)
Size	**0.002*****
	(0.0002)
Price	**-0.001*****
	(0.0003)
Constant	**0.584*****
	(0.008)
Observations	**7,724**
R2	**0.019**
Adjusted R2	**0.018**
Residual Std. Error	**0.479 (df = 7719)**
F Statistic	**37.250*** (df = 4; 7719)**
Note:	***p<0.1; **p<0.05; ***p<0.01**

Table 4.2 shows the summary output of a linear regression model fitted to the data using the lm function in R.

We find that the p-value for Installs is the largest one, so we use the backward selection method to delete one variable with the largest p-value. The goal of backward selection is to iteratively eliminate predictor variables from the model until a satisfactory or optimal model is achieved. It starts with a model that includes all the predictor variables and then systematically removes the least significant variables.

Table 4.3: Multiple Linear Regression Results.

	Dependent variable:
	Success
Reviews	**0.00000***** **(0.000)**
Size	**0.002***** **(0.0002)**
Price	**-0.001***** **(0.0003)**
Constant	**0.585***** **(0.008)**
Observations	**7,724**
R2	**0.019**
Adjusted R2	**0.018**
Residual Std. Error	**0.479 (df = 7720)**
F Statistic	**49.034*** (df = 3; 7720)**
Note:	***p<0.1; **p<0.05; ***p<0.01**

If we consider that if p-value < 0.05 for a certain variable then it is significant, then: Reviews, Size and Price have statistically significant effects on the dependent variable.

For the positive coefficient of Reviews (P-value=0.0000000000000251), we can interpret that the number of Reviews will have a positive (Coefficient=0.000000023003) and significant influence on the Rating. An increase in Reviews is associated with an increase in the Rating.

For the positive coefficient of Size (P-value=0.0000000000107965), we can interpret that the Size will have a positive (Coefficient=0.001630102897) and significant influence on the Success of the app. An increase in Size is associated with an increase in Success.

For the negative coefficient of Price (P-value=0.00469), we can interpret that the Price will have a negative (Coefficient=-0.000885913072) and significant influence on the Success of the app. An increase in Price is associated with a decrease in Success.

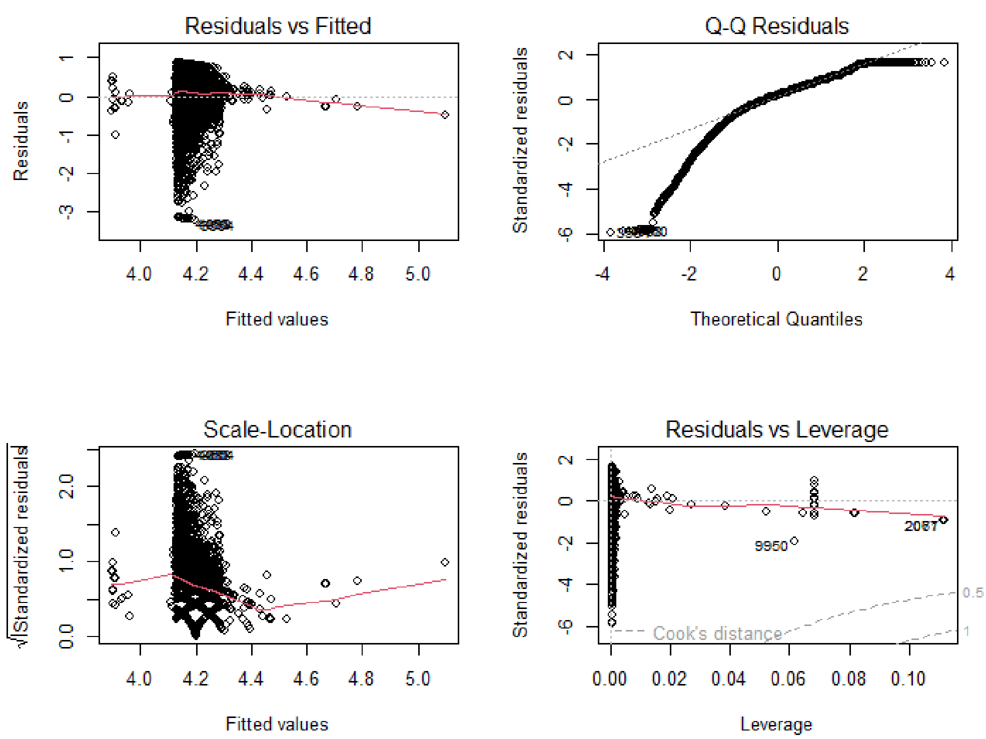
The adjusted R-squared value is 0.018, suggesting that the model explains a small proportion of the variance in the dependent variable. The F-statistic is significant, indicating that the overall model is statistically significant.

After building the Multiple linear regression model, we can answer our research question that we can predict Success based on attributes such as Review and Size. A good review will naturally improve the success of the app, and a large game size often means more content, high-quality graphics and better audio quality. With those features, it will give the customers a better game experience. In a competitive market, users may have expectations about the value they receive for the price paid. Higher-priced apps might face stronger competition in terms of user expectations, which could result in lower ratings.

Multicollinearity problem

Multicollinearity is a phenomenon that occurs when two or more independent variables in a regression model are highly correlated with each other. This can lead to several issues and challenges in the context of regression analysis. We also checked the Multicollinearity problem by using VIF, we found all the VIF is less than 5. The VIF of Reviews is 1.700146, Size is 1.062242, Installs is 1.645513, and Price is 1.000735. So we did not find multicollinearity problems.

Graph 4.3: Four diagnostic plots



Residual vs Fitted:

The point on the graph is randomly scattered around the horizontal axis with no clear pattern. This indicates that the relationship between the independent variable(s) and the dependent variable is approximately linear. The red line which is any data point that falls directly on is the estimated regression line.

Q-Q Residuals:

This plot shows the data points closely following the straight line at a 45% angle upwards. When the theoretical quantiles range from -1 to 2, the standardized residuals fall on the straight line, when the theoretical quantiles are bigger than 2, the standardized residuals start to spread out from the straight line. If it's a straight line, we consider the residuals to follow a normal distribution.

Scale-Location:

The square root of the standardized residuals is randomly scattered around the red line, there is no clear pattern among the residuals. The red line is approximately horizontal. The points on the "Scale-Location" plot are evenly spread across the horizontal axis with no clear pattern. This suggests that the variance of residuals is approximately constant across the range of fitted values.

Residuals vs Leverage:

The spread of standardized residuals decreases as leverage increases. This plot helps to measure the influence of the plots. The vertical position of the points represents the standardized residuals (residuals divided by their standard deviation). Observations with large, standardized residuals are plotted higher on the y-axis. Points with high Cook's Distance values are typically considered influential. Look for points that are both high in leverage and have large, standardized residuals.

Prediction of Success

To forecast the success of the apps, we employed the validation set approach and utilized a random forest algorithm. We randomly allocated 50% of our dataset for training the model and the remaining 50% for testing its performance. Table 4.4 presents the Mean Squared Error (MSE) obtained through both the validation set approach and the random forest technique. Notably, there is only a slight difference between the two approaches. Given its greater robustness and flexibility in managing complex data relationships, we opt for the random forest algorithm as the superior predictor for app success.

Table 4.4: Mean Squared Error

Method Type	MSE
Validation Set Approach	0.2278
Random Forest	0.2121

4.3.4 Clustering

Hierarchical clustering is a technique used in cluster analysis to group similar objects into clusters based on their similarity. The process creates a tree-like structure, known as a dendrogram, representing the hierarchy of clusters. Objects that share greater similarity are grouped at lower levels of the tree, forming branches that merge as we move up the hierarchy.

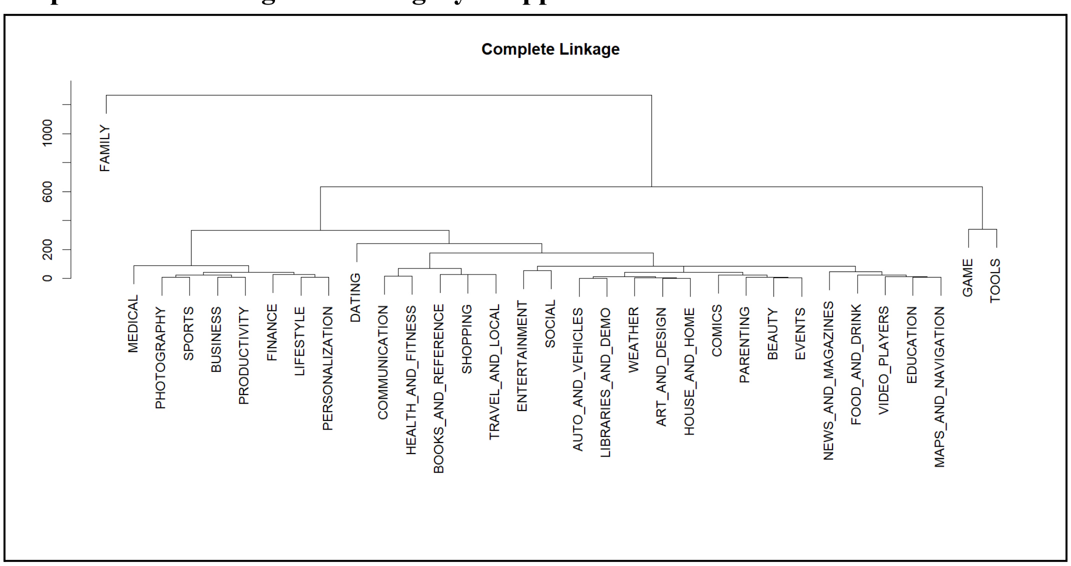
We utilized hierarchical clustering to segment app categories based on age ratings, which are divided into Everyone, Teen, Everyone 10+, Mature 17+, Adults only 18+ and Unrated as shown in Table 4.5. Our results in Graph 4.4 show that Family category apps are at the top of the hierarchy, while gaming and tools are second. Additionally, we observed that entertainment and social apps fall under one segment. This analysis provides valuable insights for app developers to make informed decisions.

Table 4.5: Contingency Table

	Adults only 18	Everyone	Everyone 10	Mature 17	Teen	Unrated
ART_AND_DESIGN	0	54	1	0	3	0
AUTO_AND_VEHICLES	0	61	1	0	1	0
BEAUTY	0	34	0	1	2	0
BOOKS_AND_REFERENCE	0	130	5	2	7	0
BUSINESS	0	242	0	0	4	0
COMICS	1	23	2	6	17	0

Given that Family category apps dominate the hierarchy, developers may want to focus on creating or promoting family-friendly apps. Since entertainment and social apps are in the same segment, developers might consider creating apps that blend elements of both, offering users a comprehensive social and entertainment experience. Cross-promoting apps within the same segment is another strategy to explore. For instance, users interested in a shopping app might be intrigued by another app from the same category.

Graph 4.4: Clustering of the category of apps



Understanding the hierarchy aids developers in identifying commonalities and preferences within specific segments. This knowledge can guide feature enhancements or updates to better align with user expectations.

5. Conclusion and Implications

5.1 Practical Implications

Our research questions can have several practical implications for both, individual app developers and large businesses.

For app developers, understanding user preferences and needs can help them create more user-friendly, appealing, and useful apps. They can also use this information to optimize their app listings for the Google Play Store, which can improve their app's visibility and chances of being discovered by potential users. Additionally, predicting app churn can help developers implement strategies to retain users, reduce uninstalls, and improve user engagement.

For businesses, insights into the Google Play Store can help them understand market trends and assess their competition. This information can be used to make informed business decisions about app development, marketing, and other areas.

Overall, examining the factors that contribute to app success on the Google Play Store can provide valuable insights for both app developers and businesses, which can help them to improve their apps and business strategies, and increase their chances of success.

5.2 Conclusion

In conclusion, our analysis delved into the factors contributing to the success of the Google Play Store Apps, utilizing data mining techniques and statistical models. We addressed key research questions related to the relationship between application rating and the number of installations, the identification of critical success factors, and the prediction of app success based on various attributes.

Our analysis revealed a statistically significant but weak positive relationship between application rating and the number of installations. While the trend suggests that higher-rated apps tend to have more installations, the practical significance was limited, indicating that other factors likely play a more substantial role in app success.

Our model for predicting app success highlighted the significant influence of the number of reviews, emphasizing its importance in the app ecosystem. Additionally, multiple linear regression explored the relationship between app success and attributes such as reviews, size, installs, and price. The results indicated that both reviews and app size had statistically significant effects on app success, suggesting that an increase in these factors was associated with higher app ratings.

Hierarchical clustering of app categories based on age ratings provided developers with valuable insights for strategic decision-making. Understanding the hierarchy of app categories can guide developers in creating family-friendly apps or exploring cross-promotion strategies within specific segments.

The practical implications of our findings extend to app developers and businesses, offering guidance on creating user-friendly apps, optimizing app listings, and making informed business decisions. The analysis provides a nuanced understanding of the dynamics of the Google Play Store, aiding stakeholders in enhancing their apps and strategies for greater success.

In the ever-evolving landscape of mobile applications, continuous analysis and adaptation to user preferences and market trends remain crucial for sustained success on platforms like the Google Play Store. Our study contributes to this ongoing dialogue, providing valuable insights for those navigating the dynamic and competitive app ecosystem.

6. References

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

Google Play Store Apps. (2019, February 3). Kaggle. https://www.kaggle.com/datasets/lava18/google-play-store-apps/data