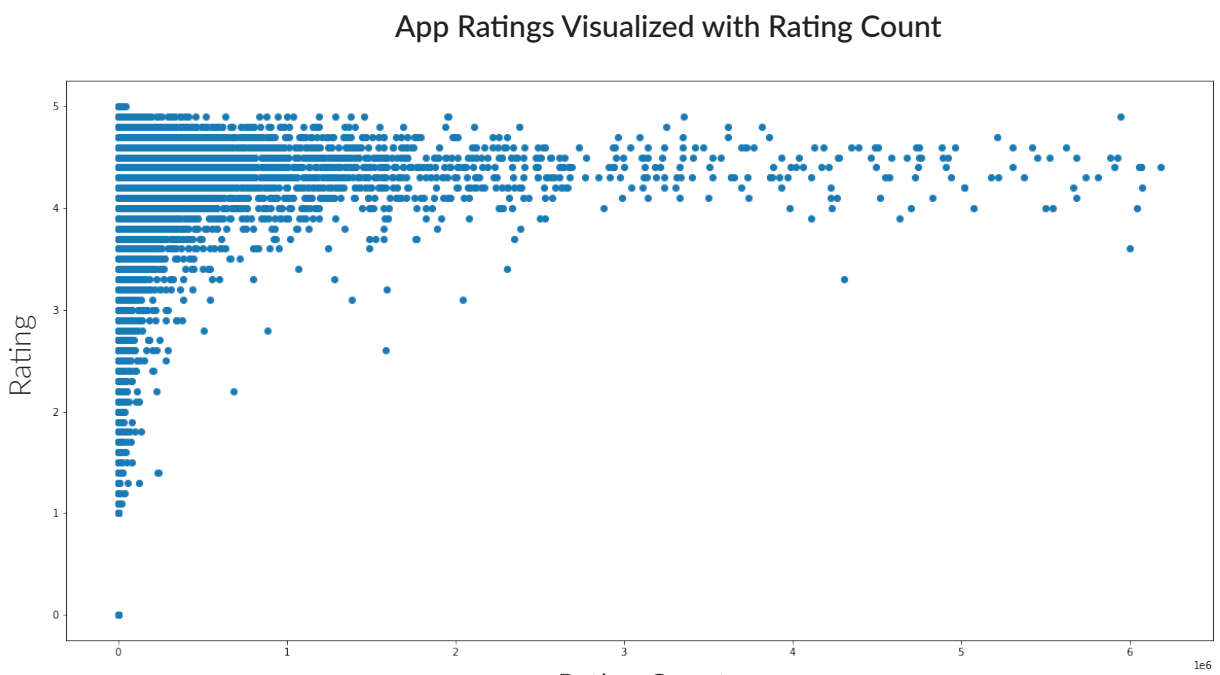
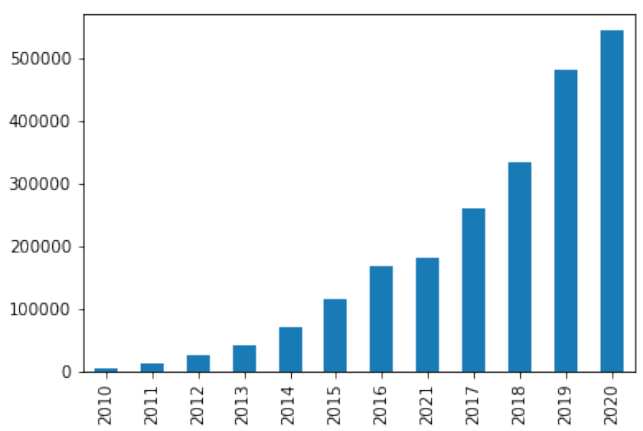
Google Play Store Data Analysis

Rauno Raa, Aleksander Kattel

Introduction

For our project, we decide to take a look at Google Play Store data. We gathered the data from the Kaggle website, from a public dataset, which include info of over 2 million apps on the Play Store. The dataset contains info about the apps' rating, installs, android version, updates, etc. Our objective was to get to know the data and visualize as well as train a machine learning model to predict app installs.





Data exploration

The dataset was very large and informative, so we had to make decisions about which parts to explore further. Something that stood out to us was the rating of the app and how it was correlated with the amount of ratings it had. We visualized the rating and rating count on a scatter plot and the results turned out quite interesting. We saw that the rating tends to even out with more ratings. Also, the highest rated apps were ones with fewer ratings and as a general rule the more ratings an app had, the greater the rating was. We also looked at application installs throughout the years, where we saw rapid growth but also maybe a sign of slowing down in 2020 as the growth was not that big compared to 2019. Another thing we explored was the ratio of paid and free apps, where the results were as expected: free apps took up 98% and paid apps 2% of apps on the Play Store.

Machine Learning Model

We decided to divide installs into size categories, because we thought that it is not the exact number that matters with the installs but rather the magnitude. Because the installs are divided into classes, we thought that good machine learning algorithms for predicting the install class would be Random Forest Classifier, Linear Support Vector Machine and Decision Tree Classifier. However Linear Support Vector Machine was too slow and Decision Tree Classifier's accuracy was about 54%. Random Forest Classifier got the "best" result of approximately 57% of accuracy. Thus, it seems that app installs aren't that much dependent on the features in the dataset, but the quality of the app.