# What makes an Android App popular?

Presented by Xu Han, Data Science ScM

Oct 20, 2019

Github repo

# Introduction

- People nowadays spend more time with their phones. While this is a great opportunity for the market, it also becomes a challenge for developers to develop a popular app. The data was downloaded from Kaggle.
- In this project, the goal is to predict the number of installs from features of the app itself such as category, size, current version, rating, paid or free etc.
- Target variable: number of installs. '<10k', '<500k', '<5 million', '> 5 million'.
- This is a classification problem.

# Preprocessing: Original df

| | App | Category | Rating | Reviews | Size | Installs | Type | Price | Content Rating | Genres | Last Updated | Current Ver | Android Ver |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Photo Editor & Candy Camera & Grid & ScrapBook | ART_AND_DESIGN | 4.1 | 159 | 19 | 10,000+ | Free | 0 | Everyone | [Art & Design] | 7-Jan-18 | 1.0.0 | 4.0.3 and up |
| **1** | Coloring book moana | ART_AND_DESIGN | 3.9 | 967 | 14 | 500,000+ | Free | 0 | Everyone | [Art & Design, Pretend Play] | 15-Jan-18 | 2.0.0 | 4.0.3 and up |
| **2** | U Launcher Lite – FREE Live Cool Themes, Hide ... | ART_AND_DESIGN | 4.7 | 87510 | 8 | 5,000,000+ | Free | 0 | Everyone | [Art & Design] | 1-Aug-18 | 1.2.4 | 4.0.3 and up |
| **3** | Sketch - Draw & Paint | ART_AND_DESIGN | 4.5 | 215644 | 25 | 50,000,000+ | Free | 0 | Teen | [Art & Design] | 8-Jun-18 | Varies with device | 4.2 and up |

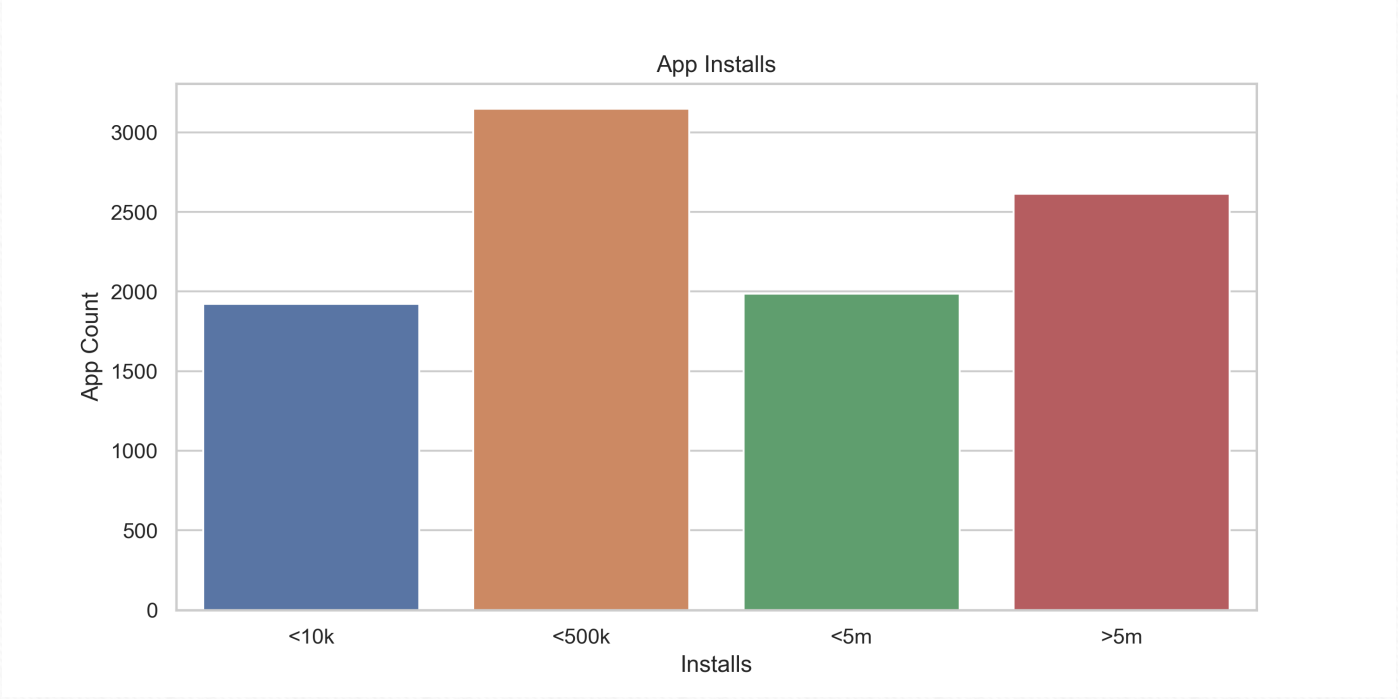
# Preprocessing: Data Cleaning

- Original df shape: (10841, 13); df_preprocessed shape: (9676, 68)
- Drop duplicates, adjust columns that has datatype 'Object', use ordinal encoder to preprocess target variable and reduce the number of groups to four
- Use Regular Expression to extract number from string and convert it back to int/float64.

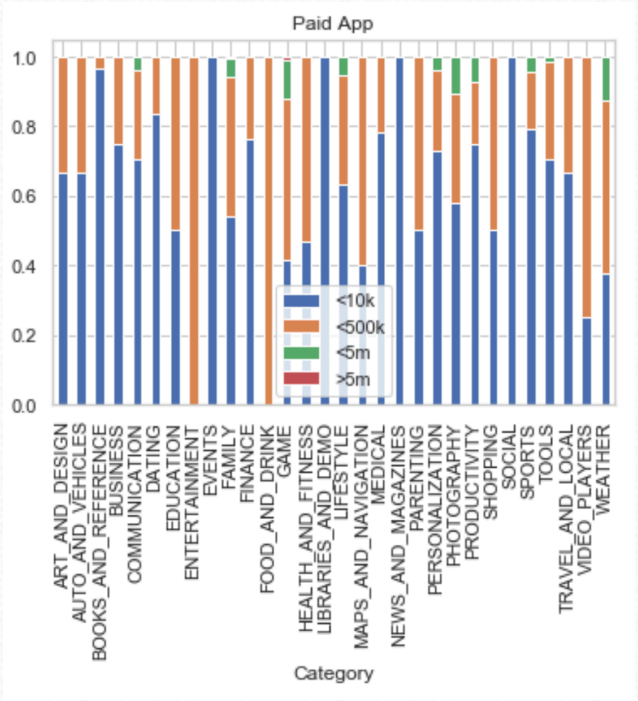
# Preprocessing: Missing values

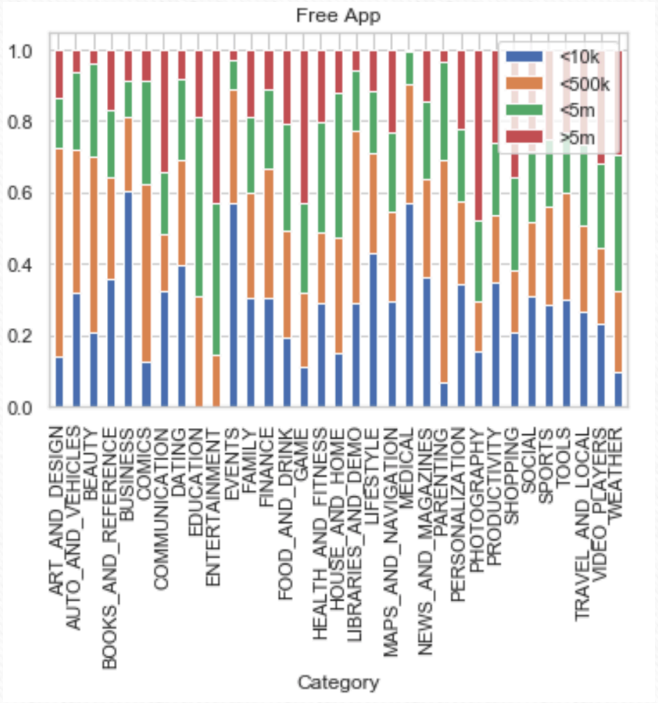
- Run MCAR test on df_preprocessed (test result p = 0)
- Train test split
- Iterative impute missing values for continuous features in test dataset
- Use minmax scaler, standard scaler on continuous features, onehot encoder, ordinal encoder on categorical features in test dataset
- For categorical features, make a new column for missing values
- Check percentage of missing data and drop two columns that has 91% data missing

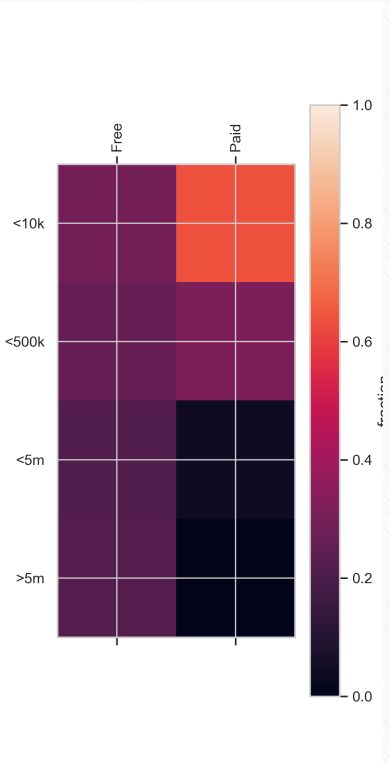
# EDA: Data Balance



# EDA: Free & entertainment apps have more downloads



# EDA: People care about popular paid app?