

1. Для обработки выбрана база данных основных характеристик размещенных в Play Market приложений: <https://www.kaggle.com/lava18/google-play-store-apps/data>. В качестве дополнения, возможно, понадобится информация по Apple App Store: <https://www.kaggle.com/cmquib19/763k-ios-app-info>

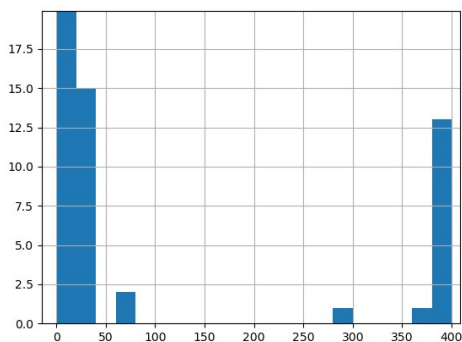
Цель – попытаться проанализировать рынок приложений. Какая успешность (количество скачиваний, стоимость) ждет приложение с заданными характеристиками? Как зависит стоимость от типа приложения, языка, возраста целевой аудитории? Есть ли какие-то особенности, объединяющие успешные приложения? Также интересно понаблюдать за динамикой спроса – возможно, временные зависимости как-то определяют успешность проекта.

2. Знакомство с первоначальными данными [10841 rows x 13 columns]

название колонки	unique() элементы	комментарий
App	9660	название само по себе не очень информативно, скорее всего можно убрать
Category	34	'ART_AND_DESIGN' 'AUTO_AND_VEHICLES' 'BEAUTY' 'BOOKS_AND_REFERENCE' 'BUSINESS' 'COMICS' 'COMMUNICATION' 'DATING' 'EDUCATION' 'ENTERTAINMENT' 'EVENTS' 'FINANCE' 'FOOD_AND_DRINK' 'HEALTH_AND_FITNESS' 'HOUSE_AND_HOME' 'LIBRARIES_AND_DEMO' 'LIFESTYLE' 'GAME' 'FAMILY' 'MEDICAL' 'SOCIAL' 'SHOPPING' 'PHOTOGRAPHY' 'SPORTS' 'TRAVEL_AND_LOCAL' 'TOOLS' 'PERSONALIZATION' 'PRODUCTIVITY' 'PARENTING' 'WEATHER' 'VIDEO_PLAYERS' 'NEWS_AND_MAGAZINES' 'MAPS_AND_NAVIGATION' '1.9'???
Rating	41	Десятичное число, бывает NAN
Reviews	6002	Количество отзывов, стоит учитывать в суперпозиции с рейтингом
Size	462	Размер приложения, иногда зависит от типа устройства
Installs	22	Не забыть посмотреть на долю тех, кто оставил отзыв
Type	4	Paid/Free; избыточно
Price	93	
Content Rating	7	Возрастной маркер
Genres	120	вроде бы совпадает с Category
Last Updated		Есть ли сезонность?
Current Ver	2833	Как давно на рынке, может ли много стоить молодое приложение?
Android Ver	34	важен ли этот параметр для определения стоимости?

Итак, осталось 7-8 параметров.

3. Последовательный анализ столбцов:



Гистограмма для Price

- a) **Category.** Удаляем строчку с значением '1.9' (10841->10840)
- b) **Rating.** Удаляем строки с NaN и с значениями > 5. (10840->9366)
- c) **Reviews.** Приводим к типу int.
- d) **Installs.** Приводим к типу int.
- e) **Price.** Приводим к типу float. Из интереса построим гистограмму.
- f) **Content Rating, Last Updated, Current Ver.** Пока пропускаем

4. Кардинальный шаг – исключим приложения, сопутствующие раскрученным проектам. Такие приложения бесплатны и имеют огромное количество скачиваний ($>10^6$). Примеры из жизни – Яндекс.Такси; Аэрофлот.Бонус; Додо.Пицца и т.д. (9366 ->6537)
5. Интересно исследовать динамику. Добавим похожий файл для Play Market, но уже за 2019 год <https://www.kaggle.com/gauthamp10/google-playstore-apps#Google-Playstore-32K.csv>
Как именно лучше объединить таблицы, включать или нет информацию по AppStore станет ясно, когда появится описание к следующим лабораторным.