

# Московский Авиационный Институт (Национальный исследовательский университет)

# Факультет «Информационные технологии и прикладная математика»

# Кафедра 806 «Вычислительная математика и программирование»

Постановка задачи

Требуется сформировать/получить два набора данных соответствующие следующим критериям:

1) Один из датасетов должен представлять собой корпус документов. Язык, источник и тематика произвольна

2) Второй датасет должен содержать категориальные, количественные признаки. Для данного датасета определить предсказываемые признаки (для задачи регрессии и классификации). Если такого признака нет, спроектировать

По каждому датасету построить распределения признаков (в случае корпуса документов – построить распределение слов) и объяснить имеющуюся картину. Вычислить статистические характеристики признаков. Обнаружить и решить возможные проблемы с данными. Если решить данную проблему невозможно, объяснить почему.

Выполнение работы

Оба датасета были взяты с сайта Kaggle:

<https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps>

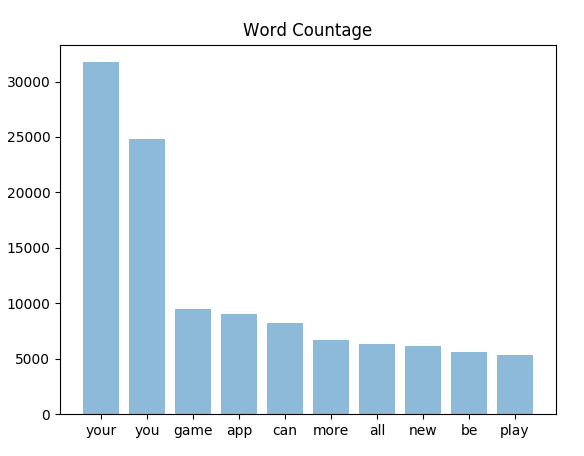
Оба датасета представляют собой наборы данных о приложениях из App Store.

# Текстовый датасет

Этот датасет включает в себя:

* ID – идентификатор приложения
* track\_name – название приложения
* size\_bytes – размер приложения в байтах
* app\_desc – описание приложения

Для анализа мы возьмем только поле *app\_desc*, которое содержит текст.



Не учитываются такие слова, как “the”, ”that”, ”and”… потому что они не несут никакой информации.

# Датасет с категориальными и количественными признаками

Этот датасет включает в себя:

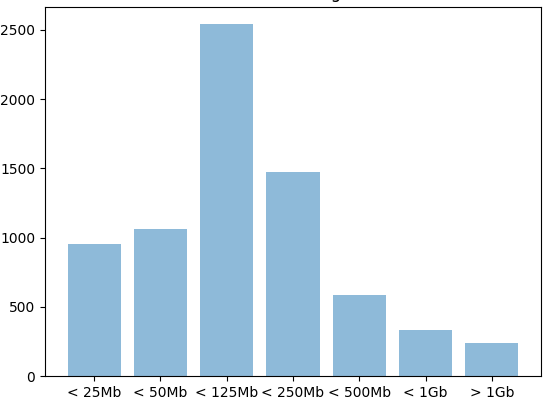
* ID – идентификатор приложения
* track\_name – название приложения
* size\_byte – размер приложения в байтах
* currency – валюта, в которой указана стоимость приложения
* price – стоимость приложения
* rating\_count\_tot – количество отзывов о приложении (для всех версий)
* rating\_count\_ver – количество отзывов о приложении (для текущей версии)
* user\_rating – средняя оценка пользователей (для всех версий)
* user\_rating\_ver – средняя оценка пользователей (для текущей версии)
* ver – код последней версии приложения
* cont\_rating – возрастные ограничения
* prime\_genre – жанр
* sup\_devices.num – количество поддерживаемых устройств
* ipadSc\_urls.num – количество скриншотов, показанных для превью
* lang.num – количество поддерживаемых языков
* vpp\_lic – лицензирование на основе устройств VPP

### ID – идентификатор приложения, track\_name – название приложения

Не представляют для нас никакого интереса

### size\_bytes



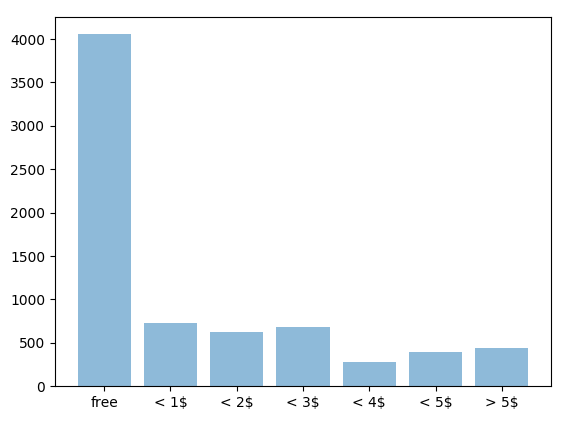


### currency

Во всех рядах атрибут равен “USD”, поэтому он не является информативным

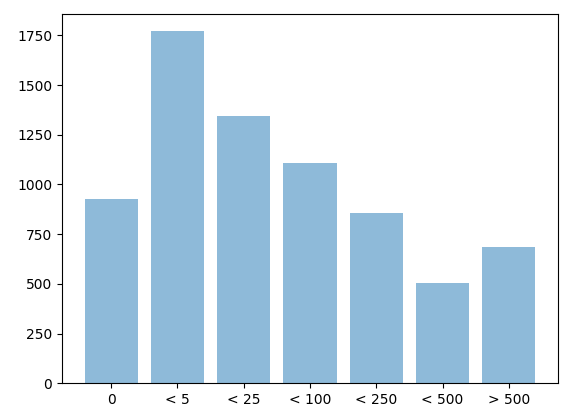
### price





### rating\_count\_tot (кол-во оценок по всем версиям)





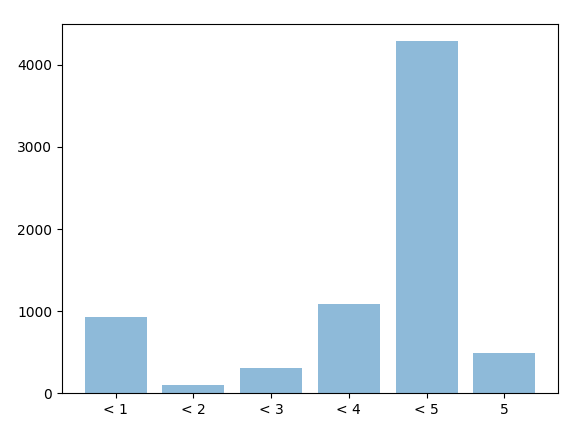
### rating\_count\_ver (кол-во оценок последней версии)





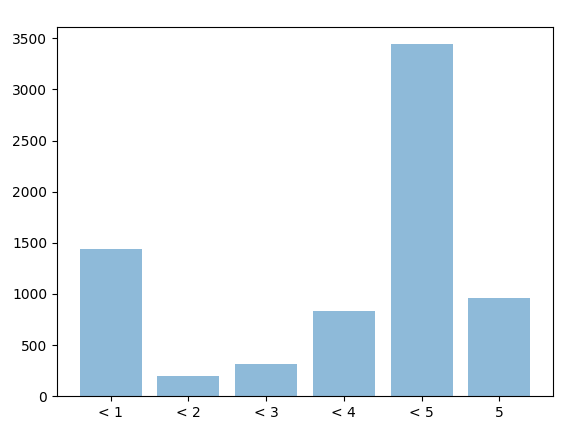
### User\_rating (рейтинг по всем версиям)





### User\_rating\_ver (рейтинг по последней версии)

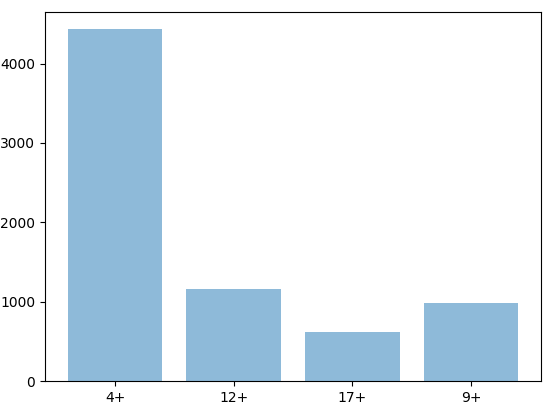




### ver

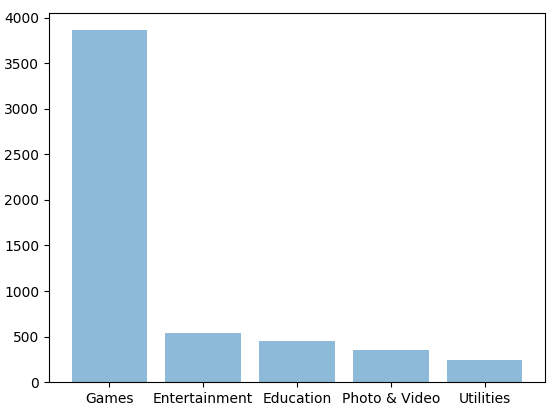
Не представляет для нас никакого инетереса

### cont\_rating (возрастные ограничения)



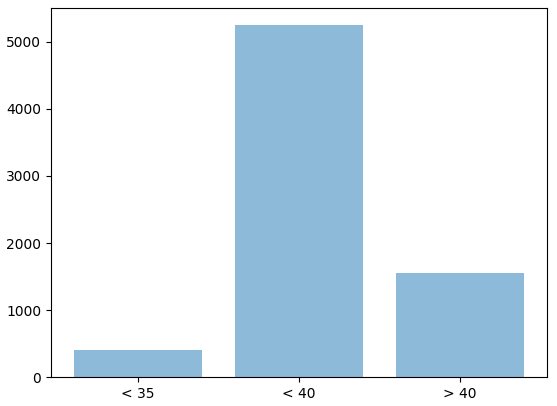
### prime\_genre

ниже приведено 5 самых популярных жанров



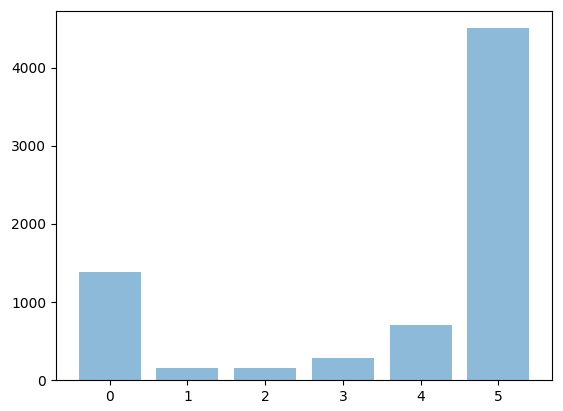
### sup\_devices.num (кол-во поддерживаемых устройств)





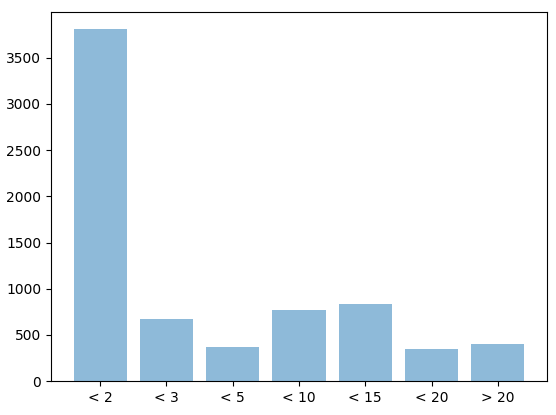
### ipadSc\_urls.num (кол-во скриншотов в маркете)





### lang.num (кол-во языков)





### vpp\_lic

Не представляет для нас интереса

Вывод

Работая над данной лабораторной работой, я нашел два датасета, построил для них распределение признаков и проанализировал их для построения модели обучения.