Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

Институт информационных технологий и анализа данных

наименование института

ОТЧЕТ  
к лабораторной работе №3 по дисциплине

Интеллектуальные системы и технологии

«Анализ данных»

Наименование темы

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Выполнил студент группы: | ИСМб-19-1 |  |  |  | Вовиков Д.Е. |
|  | Шифр группы |  | подпись |  | Фамилия ИО |
| Проверил |  |  |  |  | Массель А.Г. |
|  | должность |  | подпись |  | Фамилия ИО |
| Содержание отчета на: | |  | 9 | стр. | |

Иркутск 2022 г

Содержание

[1 Введение 3](#_Toc120031940)

[1.1 Цель лабораторной 3](#_Toc120031941)

[1.2 Задачи 3](#_Toc120031942)

[1.3 Описание Video Game Sales 3](#_Toc120031943)

[2 Таблица и гипотезы, которые можно вывести 4](#_Toc120031944)

[2.1 Использованные данные: 4](#_Toc120031945)

[2.2 Гипотезы: 4](#_Toc120031946)

[3 Программная реализация 5](#_Toc120031947)

[3.1 Чистка данных 5](#_Toc120031948)

[3.2 Обработка данных 5](#_Toc120031949)

[3.2.1 Генерация цветовой схемы 5](#_Toc120031950)

[3.2.2 Сортировка объектов 5](#_Toc120031951)

[3.3 Построение графиков 5](#_Toc120031952)

[3.3.1 График жанров 5](#_Toc120031953)

[3.3.2 График издательств 5](#_Toc120031954)

[3.3.3 График платформ 6](#_Toc120031955)

[3.3.4 График годов издания 6](#_Toc120031956)

[4 Графики и выводы по ним 7](#_Toc120031957)

1 Введение

1.1 Цель лабораторной

Разобрать пример исследования и классификации данных с использованием некоторых библиотек на Python

1.2 Задачи

* выбрать интересующий нас набор данных (DataSet). Разнообразные наборы Dataset'ы можно скачать с [1]. DataSet обычно представляет собой файл с таблицей в формате JSON или CSV.
* Провести анализ данных и сделать выводы.
* \*Разбить датасет на 2 части (20/80) применить методы машинного обучения для прогнозирования.

1.3 Описание Video Game Sales

Этот набор данных содержит список видеоигр, продажи которых превышают 100 000 копий. Он был сгенерирован на основе данных vgchartz.com.

Поля включают

Rank – Рейтинг общих продаж

Name – Название игры

Platform — платформа выпуска игр (например, ПК, PS4 и т. д.)

Year – Год выхода игры.

Genre – Жанр игры.

Publisher – Издатель игры

NA\_Sales — продажи в Северной Америке (в миллионах)

EU\_Sales — Продажи в Европе (в миллионах)

JP\_Sales — продажи в Японии (в миллионах)

Other\_Sales — Продажи в остальном мире (в миллионах)

Global\_Sales — общий объем продаж по всему миру.

2 Таблица и гипотезы, которые можно вывести

2.1 Использованные данные:

Platform — платформа выпуска игр (например, ПК, PS4 и т. д.)

Year – Год выхода игры.

Genre – Жанр игры.

Publisher – Издатель игры

2.2 Гипотезы:

1. Самая распространенная платформа
2. Самый распространенный год
3. Самый распространенный жанр
4. Самый распространенный издатель игр

3 Программная реализация

3.1 Чистка данных

df = df.replace({None: 'unknown'})

df = df.replace({np.NaN: 'unknown'})

3.2 Обработка данных

3.2.1 Генерация цветовой схемы

PLOT\_LABEL\_FONT\_SIZE = 25

def getColors(n):

COLORS = []

cm = plt.cm.get\_cmap('hsv', n)

for i in np.arange(n):

COLORS.append(cm(i))

return COLORS

3.2.2 Сортировка объектов

def dict\_sort(my\_dict):

keys = []

values = []

my\_dict = sorted(my\_dict.items(), key=lambda x:x[1], reverse=True)

for k, v in my\_dict:

keys.append(k)

values.append(v)

return (keys,values)

3.3 Построение графиков

3.3.1 График жанров

breed\_count = pd.value\_counts(df['Genre'].values, sort=True)

breed\_count\_keys, breed\_count\_values = dict\_sort(dict(breed\_count))

TOP\_breed = len(breed\_count\_keys)

plt.figure(figsize=(20,10))

plt.title('Жанры', fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.bar(np.arange(TOP\_breed), breed\_count\_values, color=getColors(TOP\_breed)) #метод построения гистаграмм

plt.xticks(np.arange(TOP\_breed), breed\_count\_keys, rotation=90, fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.yticks(fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.ylabel('Количество игр в жанре', fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.show()

3.3.2 График издательств

breed\_count = pd.value\_counts(df['Publisher'].values, sort=True)

breed\_count\_keys, breed\_count\_values = dict\_sort(dict(breed\_count))

print(range(len(breed\_count\_keys)))

del breed\_count\_values[50:579]

del breed\_count\_keys[50:579]

TOP\_breed = len(breed\_count\_keys)

plt.figure(figsize=(20,10))

plt.title('Издатель', fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.bar(np.arange(TOP\_breed), breed\_count\_values, color=getColors(TOP\_breed)) #метод построения гистаграмм

plt.xticks(np.arange(TOP\_breed), breed\_count\_keys, rotation=90, fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.yticks(fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.ylabel('Количество игр в издании', fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.show()

3.3.3 График платформ

breed\_count = pd.value\_counts(df['Platform'].values, sort=True)

breed\_count\_keys, breed\_count\_values = dict\_sort(dict(breed\_count))

TOP\_breed = len(breed\_count\_keys)

plt.figure(figsize=(20,10))

plt.title('Платформы', fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.bar(np.arange(TOP\_breed), breed\_count\_values, color=getColors(TOP\_breed)) #метод построения гистаграмм

plt.xticks(np.arange(TOP\_breed), breed\_count\_keys, rotation=90, fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.yticks(fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.ylabel('Количество игр на платформе', fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.show()

3.3.4 График годов издания

breed\_count = pd.value\_counts(df['Year'].values, sort=True)

breed\_count\_keys, breed\_count\_values = dict\_sort(dict(breed\_count))

del breed\_count\_values[20]

del breed\_count\_keys[20]

TOP\_breed = len(breed\_count\_keys)

plt.figure(figsize=(20,10))

plt.title('Год издания', fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.bar(np.arange(TOP\_breed), breed\_count\_values, color=getColors(TOP\_breed)) #метод построения гистаграмм

plt.xticks(np.arange(TOP\_breed), breed\_count\_keys, rotation=90, fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.yticks(fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.ylabel('Количество игр изданных в год', fontsize=PLOT\_LABEL\_FONT\_SIZE)

plt.show()

4 Графики и выводы по ним

Самые распространенный жанр это Action.

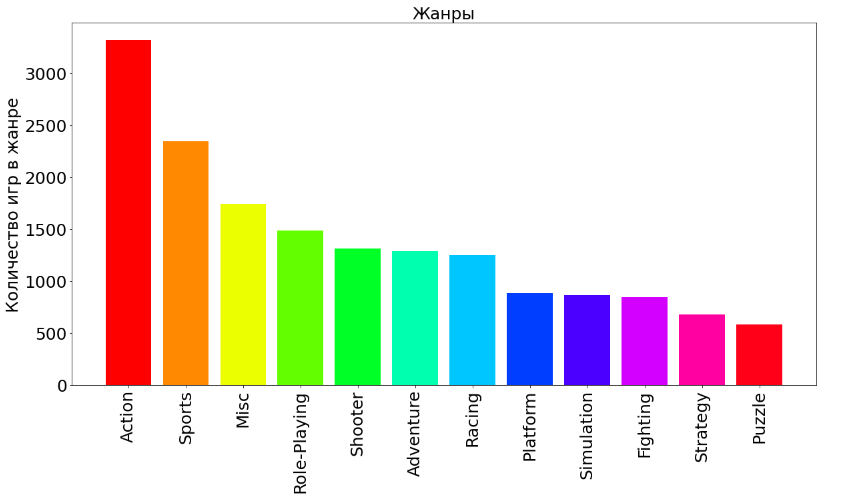


Рисунок 1 – Количество игр в жанрах

Самый распространённый издатель игр – это Electronic Arts.

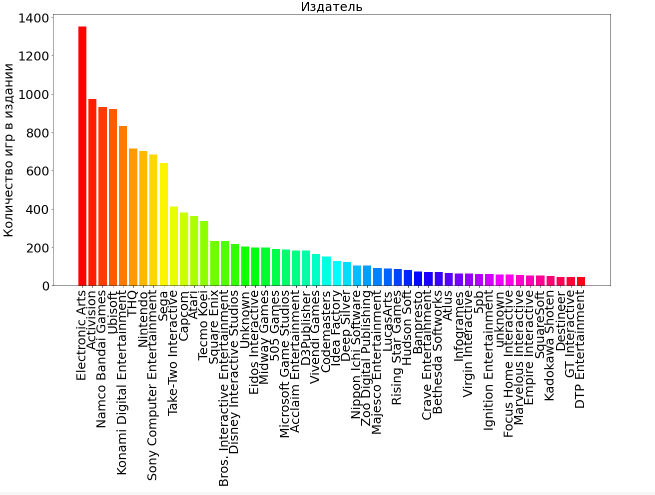


Рисунок 2 – Количество игр в издании

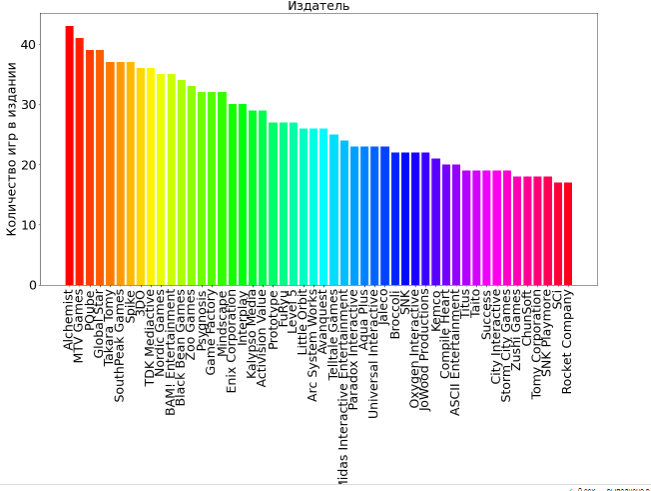


Рисунок 3 – Количество игр в издании

Самые распространенные платформы – это DS и PS2.

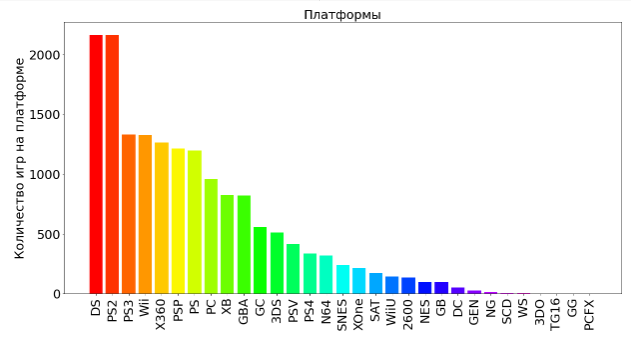


Рисунок 4 – Количество игр на платформе

Самый распространённый год для издания игра – это 2009 и 2008.

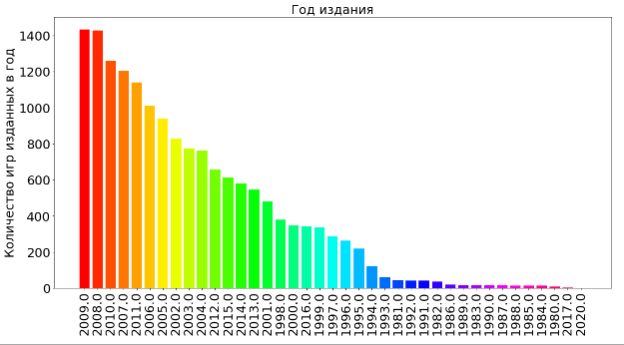


Рисунок 5 – Количество игр изданный в год