Университет ИТМО

# Практическая работа №3

по дисциплине «Визуализация и моделирование»

**Автор:** Логвинов Лев Анатольевич
**Поток:** 11.03.02
**Группа:** K3220
**Факультет:** ИКТ
**Преподаватель:** Чернышева А.В.

Санкт-Петербург, 2021 г.

Ссылка на датасет: https://www.kaggle.com/gregorut/videogamesales

Данный датасет содержит список видеоигр, у которых было продано более 100 тыс. копий. В датасете представлена информация о самих играх (название, платформа, год выпуска и т.п.), а также объемы их продаж на различных территориях (Северная Америка, Европа, Япония, остальной мир) и по всему миру в целом.

| Название столбца | Даннные, хранящиеся в столбце | Тип данных | Шкала |
|---|---|---|---|
| Rank | Рейтинг общих продаж | Целое число | Порядковая |
| Name | Название игры | Строка | Номинальная |
| Platform | Платформа выпуска игры | Строка | Номинальная |
| Year | Год выпуска игры | Целое число | Относительная |
| Genre | Жанр игры | Строка | Номинальная |
| Publisher | Издатель игры | Строка | Номинальная |
| NA_Sales | Продажи в Северной Америке (млн.) | Число с плавающей точкой | Относительная |
| EU_Sales | Продажи в Европе (млн.) | Число с плавающей точкой | Относительная |
| JP_Sales | Продажи в Японии (млн.) | Число с плавающей точкой | Относительная |
| Other_Sales | Продажи в остальном мире (млн.) | Число с плавающей точкой | Относительная |
| Global_Sales | Общий объем продаж по всему миру | Число с плавающей точкой | Относительная |

Столбцы датасета были проверены на наличие ошибочных данных. Были получены все уникальные значения столбцов Platform (содержащий данные о платформе, на которую была выпущена игра) и Genre (содержащий данные о жанре игры).

```
df_n["Platform"].unique()
```

```
array(['Wii', 'NES', 'GB', 'DS', 'X360', 'PS3', 'PS2', 'SNES', 'GBA',
       '3DS', 'PS4', 'N64', 'PS', 'XB', 'PC', '2600', 'PSP', 'XOne', 'GC',
       'WiiU', 'GEN', 'DC', 'PSV', 'SAT', 'SCD', 'WS', 'NG', 'TG16',
       '3DO', 'GG', 'PCFX'], dtype=object)
```

```
df_n["Genre"].unique()
```

```
array(['Sports', 'Platform', 'Racing', 'Role-Playing', 'Puzzle', 'Misc',
       'Shooter', 'Simulation', 'Action', 'Fighting', 'Adventure',
       'Strategy'], dtype=object)
```

Ячейки датасета были проверены на отсутствие данных (значение ячейки = NULL). Так как пустых ячеек оказалось немного (271 в столбце Year и 58 в столбце Publisher), учитвая, что в датасете около 16500 строк, было решено удалить строки, содержащие данные ячейки, так как это почти не повлияет на целостность датасета.

```
cols = list(df_n.columns)
df_na = {col: list(pd.isna(df_n[col])).count(True) for col in cols}
df_na
```

```
{'Rank': 0,
 'Name': 0,
 'Platform': 0,
 'Year': 271,
 'Genre': 0,
 'Publisher': 58,
 'NA_Sales': 0,
 'EU_Sales': 0,
 'JP_Sales': 0,
 'Other_Sales': 0,
 'Global_Sales': 0}
```

Были получены данные о столбце Year, представляющем информацию о дате выхода игры (максимальное, минимальное, среднее значения, а также медиана, мода и интерквартильный размах). В целома для анализа датасета важны все представленные значения, то есть выбросов нет.

```
print(df_n["Year"].max())
print(df_n["Year"].min())
```

```
2020.0
1980.0
```

```
year_stat = {"mean": df_n["Year"].mean(),
             "median": df_n["Year"].median(),
             "mode": df_n["Year"].mode().to_list(),
             "interquartile_range": df_n["Year"].quantile(0.75) - df_n["Year"].quantile(0.25),
             }
year_stat
```

```
{'mean': 2006.4055613528942,
 'median': 2007.0,
 'mode': [2009.0],
 'interquartile_range': 7.0}
```

Проблемы в данных:

1. Столбец Year - Пустые ячейки. Способ решения: удалить строки, содержащие пустые ячейки. Также можно заменить на среднее значение, моду или медиану, так как они не сильно разнятся, учитывая рассматриваемый диапазон.
2. Столбец Publisher - Пустые ячейки. Способ решения: удалить строки, содержащие пустые ячейки.
3. В столбцах, хранящих данные об объемах продаж (NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales), есть выбросы (подробнее в Jupiter), однако, я не считаю, что это является проблемой, которую нужно обработать, потому что данные потеряют свою целостность. Очевидно, что есть игры, которые выстрелили, из-за чего их продажи в несколько раз превышают продажи остальных игр и очевидно, что таких игр не много. Я считаю, что для дальнейшей визуализации эти данные важны.

Гипотеза: при визуализации после обработки данного датасета видимых для нас изменений не будет, так как была изменена незначительная часть датасета (были удалены <0,001% строк, также были оставлены очень большие значения - выбросы, ввиду их важности для целостности данных).