Университет ИТМО

Практическая работа №3

по дисциплине «Визуализация и моделирование»

Автор: Хоанг Минь Тханг
Поток: ВИМ1.1
Группа: К33212
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

1. Описание

Данные о видеоиграх, выпущенных с 1980 по 2016.

Датасет:

	Name	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	Critic_Count	User_Score	User_Count	Developer	Rating
0	Wii Sports	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	51.0	8	322.0	Nintendo	E
1	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN	NaN	NaN	NaN	NaN
2	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	73.0	8.3	709.0	Nintendo	E
3	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	73.0	8	192.0	Nintendo	E
4	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	NaN	NaN	NaN	NaN

Таблица с описанием данных в столбцах:

Название	Описание	Формат	Шкала	Проблема	Решение
Name	название видеоигры	str	номинальная	не нужно использовать при построении модели	удалить столбец
Platform	предназначенная платформа	str	номинальная	-	-
Year_of_Release	год выпуска	float	интервальная	тип данных: float	перевод в целое число
Genre	жанр видеоигры	str	номинальная	-	-
Publisher	издатель видеоигры	str	номинальная	-	-
NA_Sales	количество проданных единиц в Северной Америке (в миллионах)	float	относительная	-	-
EU_Sales	количество проданных единиц в Европе (в миллионах)	float	номинальная	-	-
JP_Sales	количество проданных единиц в Японии (в миллионах)	float	номинальная	-	-
Other_Sales	количество проданных единиц в других регионах (в миллионах)	float	номинальная	-	-
Global_Sales	количество проданных единиц в мире (в миллионах)	float	номинальная	-	-
Critic_Score	оценка критиков	float	относительная	-	-
Critic_Count	количество критиков	float	относительная	не нужно использовать при построении модели	удалить столбец
User_Score	оценка пользователей	str	относительная	тип даннх: str	перевод в число
User_Count	количество пользователей	float	относительная	не нужно использовать при построении модели	удалить столбец
Developer	разработающая компания	str	номинальная	не нужно использовать при построении модели	удалить столбец
Rating	рейтинг	str	номинальная	-	-

2. Предобработка данных

2.1. Удаление нерелевантных столбцов

```
df.drop('Name', inplace=True, axis=1)
df.drop('Critic_Count', inplace=True, axis=1)
df.drop('User_Count', inplace=True, axis=1)
df.drop('Developer', inplace=True, axis=1)
```

```
df.head()
```

	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	User_Score	Rating
0	Wii	2006.0	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	8	E
1	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN	NaN
2	Wii	2008.0	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	8.3	E
3	Wii	2009.0	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	8	E
4	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	NaN

2.2. Столбец Year_of_Release

Нужно переводить тип данных в целое число, потому что это год выпуска.

```
df["Year_of_Release"] = df["Year_of_Release"].astype('Int64')
df.head()
```

	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	User_Score	Rating
0	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	8	E
1	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	NaN	NaN	NaN
2	Wii	2008	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	8.3	E
3	Wii	2009	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	8	E
4	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	NaN	NaN	NaN

2.3. Столбец User_Score

```
df['User_Score'].unique()
```

```
array(['8', nan, '8.3', '8.5', '6.6', '8.4', '8.6', '7.7', '6.3', '7.4',
       '8.2', '9', '7.9', '8.1', '8.7', '7.1', '3.4', '5.3', '4.8', '3.2',
       '8.9', '6.4', '7.8', '7.5', '2.6', '7.2', '9.2', '7', '7.3', '4.3',
       '7.6', '5.7', '5', '9.1', '6.5', '8.8', '6.9', '9.4', '6.8', '6.1',
       '6.7', '5.4', '4', '4.9', '4.5', '9.3', '6.2', '4.2', '6', 'tbd',
       '3.7', '4.1', '5.8', '5.6', '5.5', '4.4', '4.6', '5.9', '3.9',
       '3.1', '2.9', '5.2', '3.3', '4.7', '5.1', '3.5', '2.5', '1.9', '3',
       '2.7', '2.2', '2', '9.5', '2.1', '3.6', '2.8', '1.8', '3.8', '0',
       '1.6', '9.6', '2.4', '1.7', '1.1', '0.3', '1.5', '0.7', '1.2',
       '2.3', '0.5', '1.3', '0.2', '0.6', '1.4', '0.9', '1', '9.7'],
      dtype=object)
```

Видим, что в столбце есть странное значение 'tbd'. Заменять это значение на NaN и переводить тип данных в float.

```
df['User_Score'] = df['User_Score'].replace("tbd", np.nan)
```

```
df['User_Score'].unique()
```

```
array(['8', nan, '8.3', '8.5', '6.6', '8.4', '8.6', '7.7', '6.3', '7.4',
       '8.2', '9', '7.9', '8.1', '8.7', '7.1', '3.4', '5.3', '4.8', '3.2',
       '8.9', '6.4', '7.8', '7.5', '2.6', '7.2', '9.2', '7', '7.3', '4.3',
       '7.6', '5.7', '5', '9.1', '6.5', '8.8', '6.9', '9.4', '6.8', '6.1',
       '6.7', '5.4', '4', '4.9', '4.5', '9.3', '6.2', '4.2', '6', '3.7',
       '4.1', '5.8', '5.6', '5.5', '4.4', '4.6', '5.9', '3.9', '3.1',
       '2.9', '5.2', '3.3', '4.7', '5.1', '3.5', '2.5', '1.9', '3', '2.7',
       '2.2', '2', '9.5', '2.1', '3.6', '2.8', '1.8', '3.8', '0', '1.6',
       '9.6', '2.4', '1.7', '1.1', '0.3', '1.5', '0.7', '1.2', '2.3',
       '0.5', '1.3', '0.2', '0.6', '1.4', '0.9', '1', '9.7'], dtype=object)
```

```
df['User_Score'] = df['User_Score'].astype(float)
```

.

2.4. Обработка пустых ячеек

Выясним, есть ли в датафрейме пустые значения.

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

```
{'Critic_Score': 8582,
 'EU_Sales': 0,
 'Genre': 2,
 'Global_Sales': 0,
 'JP_Sales': 0,
 'NA_Sales': 0,
 'Other_Sales': 0,
 'Platform': 0,
 'Publisher': 54,
 'Rating': 6769,
 'User_Score': 9129,
 'Year_of_Release': 269}
```

В реальности, у большого количества видеоигр нет определенного рейтинга, поэтому пустые значения в столбце Rating можно не обработать.

В столбцах Genre, Year_of_Release, Publisher количество пустых значений не существенное по сравнению с количеством строк в датасете, поэтому можно удалить соответствующие строки.

```
df.dropna(subset = ["Genre"], inplace=True)
df.dropna(subset = ["Year_of_Release"], inplace=True)
df.dropna(subset = ["Publisher"], inplace=True)
```

Для столбцов Critic_Score и User_Score, заменить все NaN на медиану.

```
df['User_Score'].replace(np.NaN, df['User_Score'].median(), inplace=True)
df['Critic_Score'].replace(np.NaN, df['Critic_Score'].median(), inplace=True)
```

```
df.head()
```

	Platform	Year_of_Release	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales	Critic_Score	User_Score	Rating
0	Wii	2006	Sports	Nintendo	41.36	28.96	3.77	8.45	82.53	76.0	8.0	E
1	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24	71.0	7.5	NaN
2	Wii	2008	Racing	Nintendo	15.68	12.76	3.79	3.29	35.52	82.0	8.3	E
3	Wii	2009	Sports	Nintendo	15.61	10.93	3.28	2.95	32.77	80.0	8.0	E
4	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37	71.0	7.5	NaN

3. Гипотезы

3.1. Nintendo - издатель самых продаваемых видеоигр.

3.2. Видеоигры жанров Action, Shooter, Fighting в основном ориентированы для взрослых геймеров.

3.3. Мировая продажа отражает оценку пользователей.

3.4. Новые видеоигры более популярные чем старые.

3.5. Количество ежегодных выпусков видеоигр растет экспоненциально.