COMP 3400: Data Preparation Techniques Project

Analysis of popular movies from 1980 to 2019

Group Members:
- Liudmila Strelnikova 201819885
- David Chicas 201919354

Instructor's feedback:

- Distributions for variables you 'removed' are skewed. The method you used seems to have worked (std-like cutoff) but maybe a better approach is to use values above 95 percentile.

Changes we made in Iteration 2:

- We changed the way we delete the outliers in the columns "Gross" and "Budget" following the advice of instead deleting any values using the IQR method.

Cleaning outliers for **budget, gross, runtime, rating,** and **genre.**

The function replaces outliers higher than the maximum value with the maximum value
The replacement is done for the attributes specified in the variable columns
@df is the dataframe for which the replacements are performed
@columns is an array with the names of the attributes that need to be fixed

```
def replace_max_outliers_IQR(df, columns):
    for c in columns:
        q1=df[c].quantile(0.25)
        q3=df[c].quantile(0.75)
        iqr = q3-q1
        high_lim = q3 + 1.5*iqr
        df[c] = np.where(df[c] >= high_lim, high_lim, df[c])

m2_copy = m2.copy()
IQR_names = ['budget', 'gross', 'runtime', 'rating', 'genre']
replace_max_outliers_IQR(m2_copy, IQR_names)
```

Cleaning outliers for **score.**

The function replaces outliers lower than the minimum value with the minimum value
The replacement is done for the attributes specified in the variable columns
@df is the dataframe for which the replacements are performed
@columns is an array with the names of the attributes that need to be fixed

```
def replace_min_outliers_IQR(df, columns):
    for c in columns:
        q1=df[c].quantile(0.25)
        q3=df[c].quantile(0.75)
        iqr = q3-q1
        low_lim = q1 - 1.5*iqr
        df[c] = np.where(df[c] <= low_lim, low_lim, df[c])

replace_min_outliers_IQR(m2_copy, ['score'])
```

	name	rating	genre	year	score	director	country	budget	gross	company	runtime	date
0	The Shining	R	Drama	1980.0	8.4	Stanley Kubrick	United Kingdom	19000000.0	46998772.0	Warner Bros.	146.0	1980-06-13
1	The Blue Lagoon	R	Adventure	1980.0	5.8	Randal Kleiser	United States	4500000.0	58853106.0	Columbia Pictures	104.0	1980-07-02
3	Airplane!	PG	Comedy	1980.0	7.7	Jim Abrahams	United States	3500000.0	83453539.0	Paramount Pictures	88.0	1980-07-02
4	Caddyshack	R	Comedy	1980.0	7.3	Harold Ramis	United States	6000000.0	39846344.0	Orion Pictures	98.0	1980-07-25
5	Friday the 13th	R	Horror	1980.0	6.4	Sean S. Cunningham	United States	550000.0	39754601.0	Paramount Pictures	95.0	1980-05-09
...	...	...	...	...	...	...	...	...	...	...	...	...
7579	A Madea Family Funeral	PG-13	Comedy	2019.0	4.5	Tyler Perry	United States	20000000.0	74747725.0	The Tyler Perry Company	109.0	2019-03-01
7588	K-12	Not Rated	Fantasy	2019.0	6.5	Melanie Martinez	United States	5000000.0	359377.0	Atlantic Records	96.0	2019-09-05
7594	Unplanned	R	Biography	2019.0	5.8	Chuck Konzelman	United States	6000000.0	21354152.0	Unplanned Movie	109.0	2019-03-29
7604	Mine 9	Not Rated	Drama	2019.0	6.4	Eddie Mensore	United States	350000.0	226421.0	Emphatic Films	83.0	2020-05-19
7616	High on the Hog	R	Action	2019.0	3.5	Tony Wash	United States	1200000.0	45696.0	Hicktown Entertainment	85.0	2019-04-16