Анализ данных по фильмам Disney

Петр Камнев, БПИ201

Датасет

Исследование проводилось над датасетом с Kaggle в формате .csv, описывающим статистику фильмов Disney

Изначально таблица содержала 432 строки и 32 столбца.

Ссылка: https://www.kaggle.com/therealsampat/disney-movies-dataset

	Unnamed: 0	title	Production company	Release date	Running time	Country	Language	Running time (int)	Bud (fl
0	0	Academy Award Review of	Walt Disney Productions	['May 19, 1937']	41 minutes (74 minutes 1966 release)	United States	English	41.0	
1	1	Snow White and the Seven Dwarfs	Walt Disney Productions	['December 21, 1937 (Carthay Circle Theatre ,...	83 minutes	United States	English	83.0	14900(
2	2	Pinocchio	Walt Disney Productions	['February 7, 1940 (Center Theatre)', 'Febru...	88 minutes	United States	English	88.0	26000(
3	3	Fantasia	Walt Disney Productions	['November 13, 1940']	126 minutes	United States	English	126.0	22800(
4	4	The Reluctant Dragon	Walt Disney Productions	['June 20, 1941']	74 minutes	United States	English	74.0	6000(

5 rows × 32 columns

Преобразованный датасет

title - название фильма (String)

Country - страна производства (String, 4 missing)

Language - язык (String, 2 missing)

Running time - длительность (Int, 10 missing)

Budget - бюджет (Float, USD, 159 missing)

Box office - сборы в кино (Float, USD, 77 missing)

Release date - дата выхода (datatime, 4 missing)

imdb - рейтинг imdb (Float, 10 missing)

	title	Country	Language	Running time	Budget	Box office	Release date	imdb
0	Academy Award Review of	United States	English	41.0	NaN	NaN	1937-05-19	7.2
1	Snow White and the Seven Dwarfs	United States	English	83.0	1490000.0	418000000.0	1937-12-21	7.6
2	Pinocchio	United States	English	88.0	2600000.0	164000000.0	1940-02-07	7.4
3	Fantasia	United States	English	126.0	2280000.0	83300000.0	1940-11-13	7.8
4	The Reluctant Dragon	United States	English	74.0	600000.0	960000.0	1941-06-20	6.9
...	...	...	...	...	...	...	...	...
95	King of the Grizzlies	['United States', 'Canada']	English	93.0	NaN	NaN	1970-02-11	5.7
96	The Boatniks	United States	English	100.0	NaN	18607492.0	1970-07-01	5.5
97	The Wild Country	United States	English	100.0	NaN	4000000.0	1970-12-15	6.4
98	The Aristocats	United States	English	79.0	4000000.0	191000000.0	1970-12-11	7.1
99	The Barefoot Executive	United States	English	96.0	NaN	NaN	1971-03-17	6.0

Транспонирование среза

```
data_raw.iloc[:5, :5].T #по заданию нужно сделать срез и транспонировать его
```

	0	1	2	3	4
Unnamed: 0	0	1	2	3	4
title	Academy Award Review of	Snow White and the Seven Dwarfs	Pinocchio	Fantasia	The Reluctant Dragon
Production company	Walt Disney Productions	Walt Disney Productions	Walt Disney Productions	Walt Disney Productions	Walt Disney Productions
Release date	['May 19, 1937']	['December 21, 1937 (Carthay Circle Theatre ,...	['February 7, 1940 (Center Theatre)', 'Febru...	['November 13, 1940']	['June 20, 1941']
Running time	41 minutes (74 minutes 1966 release)	83 minutes	88 minutes	126 minutes	74 minutes

Расчёт нового столбца с данными

Payback – окупаемость фильма

In [68]:
```
arr = []
for i, budget in enumerate(data['Budget']):
    box = data['Box office'].iloc[i]
    if budget != budget or box != box or budget == 0 or box/budget > 500000:
        arr.append(np.nan)
    else:
        arr.append(box/budget)
data['Payback'] = arr
data
```

Out[68]:

	title	Country	Language	Running time	Budget	Box office	Release date	imdb	Payback
289	Roving Mars	United States	English	40.0	1000000.0	11000000.0	2006-01-27	7.2	11.0
272	Sacred Planet	Other	English	40.0	NaN	1108356.0	2004-04-22	6.0	NaN
0	Academy Award Review of	United States	English	41.0	NaN	NaN	1937-05-19	7.2	NaN

Превращение интервального столбца в категориальный

Разделяем картины на 20 и 21 век

In [69]:
```
arr = []
for i, date in enumerate(data['Release date']):
    if date != date:
        arr.append('Unknown')
    else:
        arr.append(str(date)[:2] + '00\'s')
data['Century'] = arr
data
```

Out[69]:

	title	Country	Language	Running time	Budget	Box office	Release date	imdb	Payback	Century
289	Roving Mars	United States	English	40.0	1000000.0	11000000.0	2006-01-27	7.2	11.0	2000's
272	Sacred Planet	Other	English	40.0	NaN	1108356.0	2004-04-22	6.0	NaN	2000's
0	Academy Award Review of	United States	English	41.0	NaN	NaN	1937-05-19	7.2	NaN	1900's
7	Saludos Amigos	United States	['English', 'Portuguese', 'Spanish']	42.0	NaN	1135000.0	1942-08-24	6.1	NaN	1900's

Многоуровневая сортировка

Сортируем по продолжительности и бюджету

```
data = data.sort_values(by=['Running time', 'Budget']) #есть задание отсортировать по колонкам
data.head(20)
```

продолжительность бюджет

15	The Adventures of Ichabod and Mr. Toad	United States	English	68.0	NaN	1625000.0	1949-10-05	7.0
22	The Living Desert	United States	English	69.0	300000.0	2600000.0	1953-11-10	7.5
171	DuckTales the Movie: Treasure of the Lost Lamp	United States	English	69.0	20000000.0	18100000.0	1990-08-03	6.9
348	Winnie the Pooh	United States	English	69.0	30000000.0	50100000.0	2011-04-06	7.2
47	Jungle Cat	United States	English	69.0	NaN	2300000.0	1960-08-10	7.4
6	Bambi	United States	English	70.0	858000.0	267400000.0	1942-08-09	7.3
32	Secrets of Life	United States	English	70.0	NaN	NaN	1956-11-06	7.8
9	The Three Caballeros	United States	['English', 'Spanish', 'Portuguese']	71.0	NaN	3355000.0	1944-12-21	6.4

Цель и гипотезы

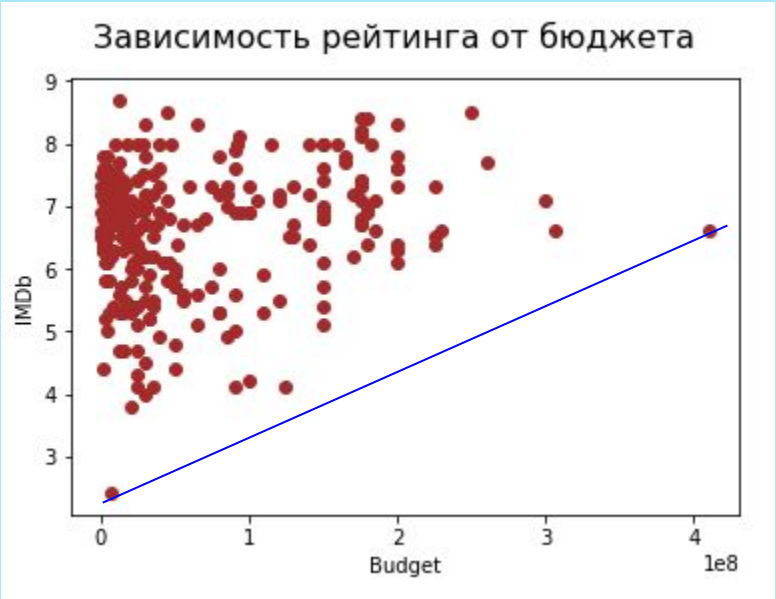
Цель исследования: выявить зависимости в данных о фильмах Disney

Гипотезы:

- Чем выше бюджет фильма, тем выше оценка IMDb
- Окупаемость коррелирует с оценкой IMDb
- Фильмы 20 века оценены выше, чем фильмы 21 века
- Продолжительность фильма зависит от бюджета
- Фильмы, снятые исключительно в США оценены выше
- У IMDb распределение, близкое к нормальному

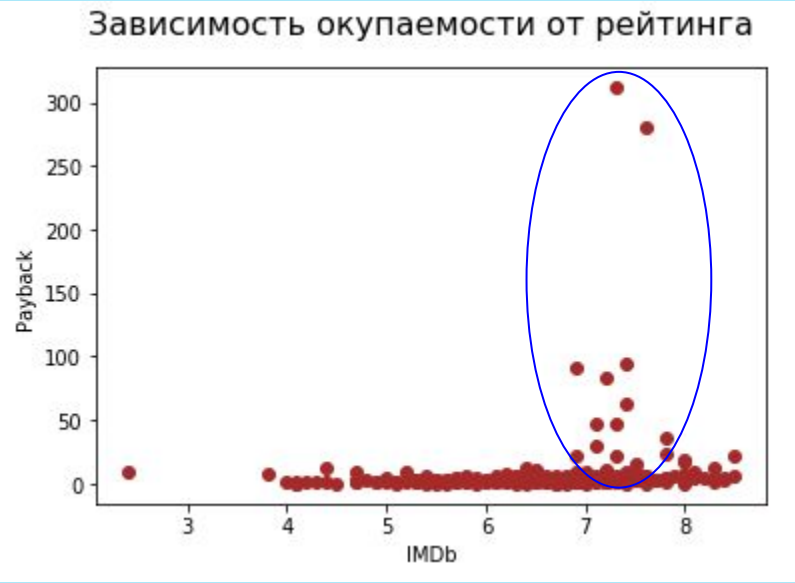
Графики

Видим, что при повышении бюджета снижается минимальный рейтинг



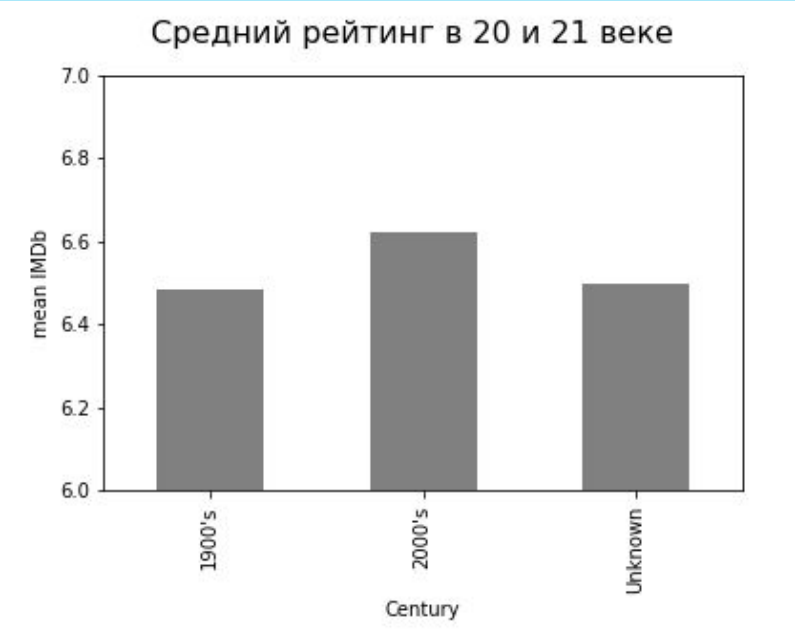
Графики

Явной зависимости не видно, но высокая окупаемость (> 50) только у фильмов с высоким рейтингом



Графики

В 21 веке средний рейтинг выше



Графики

При увеличении длительности фильма явно растет его бюджет

Графики

Распределение рейтингов IMDb похоже на нормальное

Сводная таблица

Эта таблица построена до обработки исходного датасета

Самый высокий средний рейтинг США смогли получить в коллаборации с другими странами

```
data_raw.groupby('Country').imdb.mean()
```

```
Country
Australia                                                        NaN
France                                                      7.650000
Germany                                                     4.800000
India                                                       6.680000
Russia                                                           NaN
United States                                               6.549171
United States, Mexico                                       6.100000
['Australia', 'United Kingdom', 'United States']            7.500000
['Canada', 'Malaysia', 'United States']                     6.000000
['France', 'United Kingdom', 'Germany', 'United States']    8.000000
['France', 'United States']                                 7.400000
['Germany', 'Austria', 'Italy', 'Spain', 'United Kingdom']  5.200000
['Norway', 'Sweden', 'United States']                       6.700000
['Spain', 'Italy']                                               NaN
['United Kingdom', 'New Zealand', 'United States']          8.000000
['United Kingdom', 'United States']                         6.400000
['United States', 'Australia']                              5.850000
['United States', 'Austria']                                5.700000
['United States', 'Canada']                                 5.842857
['United States', 'China', 'France']                        7.200000
['United States', 'France', 'United Kingdom']               7.400000
['United States', 'France']                                 5.200000
['United States', 'Norway']                                 6.900000
['United States', 'United Kingdom', 'France']               8.000000
['United States', 'United Kingdom', 'Germany']              7.400000
['United States', 'United Kingdom']                         6.800000
['United States']                                           7.600000
Name: imdb, dtype: float64
```

Сводная таблица

Средний рейтинг фильмов, снятых исключительно в США немного выше

Также средний фильмов, снятых исключительно в США почти в два раза выше бюджета для остальных вариантов страны производства

```
data.groupby('Country').imdb.mean()
```

```
Country
Other            6.520370
United States    6.549171
Name: imdb, dtype: float64
```

```
data.groupby('Country')['Budget'].mean()
```

```
Country
Other            3.646212e+07
United States    6.731850e+07
Name: Budget, dtype: float64
```

Корреляция

Коррелируют:

- Бюджет и продолжительность
- Бюджет и сборы
- Рейтинг и сборы

	Running time	Budget	Box office	imdb	Payback
Running time	1.000000	0.389359	0.278317	0.132544	0.236606
Budget	0.389359	1.000000	0.740025	0.200979	-0.239133
Box office	0.278317	0.740025	1.000000	0.381170	-0.191763
imdb	0.132544	0.200979	0.381170	1.000000	0.035316
Payback	0.236606	-0.239133	-0.191763	0.035316	1.000000

Описательная статистика

- Половина фильмов длиннее 96 минут
- В среднем рейтинг отклоняется на 1 балл от 6.5 баллов
- В среднем каждый фильм Disney окупает бюджет в 10 раз

	Running time	Budget	Box office	imdb	Payback
count	422.000000	2.730000e+02	3.550000e+02	416.000000	2.290000e+02
mean	97.305687	6.358861e+07	1.674014e+08	6.545433	9.549217e+00
std	18.959487	7.163732e+07	2.749517e+08	0.960937	2.494655e+01
min	40.000000	1.500000e+02	7.700000e+00	1.500000	2.511416e-08
25%	86.000000	1.000000e+07	9.850000e+06	6.000000	3.022857e-01
50%	96.000000	3.000000e+07	4.290000e+07	6.600000	1.704615e+00
75%	106.750000	1.000000e+08	1.865500e+08	7.200000	6.444444e+00
max	168.000000	4.106000e+08	1.657000e+09	8.700000	2.158750e+02

Выводы

Гипотезы:

подтверждена
частично подтверждена
опровергнута

- Чем выше бюджет фильма, тем выше оценка IMDb
- Окупаемость коррелирует с оценкой IMDb
- Фильмы 20 века оценены выше, чем фильмы 21 века
- Продолжительность фильма зависит от бюджета
- Фильмы, снятые исключительно в США оценены выше
- У IMDb распределение, близкое к нормальному

Замечена корреляция между рейтингом и кассовыми сборами