Тут будут титульники

**Содержание**

1. **Введение**
   1. **Цель**
   2. **Задачи**
2. **Определение данных для анализа**
3. **Формулирование гипотез**
4. **Анализ данных**
   1. **Загрузка данных в Python**
   2. **Проверка данных**
   3. **Очистка данных**
   4. **Агрегирование данных**
   5. **Анализ поставленных гипотез**
   6. **Корреляционный анализ**
5. **Заключение**
6. **Введение**
   1. **Цель**

Исследовать базу данных и изучить данные о фильмах, снятых в 1910-2024 годах, выявить закономерность между жанром фильма и его оценкой, чтобы дать представление киностудиям о том, фильмы какого жанра нравятся зрителям больше всего. Это позволит студиям снимать фильмы, которые принесут им большую прибыль.

* 1. **Задачи**

1. Определение данных.
2. Формулирование гипотез.
3. Загрузка данных в Python.
4. Проверка данных.
5. Очистка данных.
6. Преобразование данных.
7. Выбор данных для анализа.
8. Агрегирование данных.
9. Визуализация данных.
10. Подтверждение или опровержение поставленных гипотез.
11. Корреляционный анализ.
12. Формулирование ограничений и выводов.
13. **Определение данных для анализа**

Для анализа был выбран DataSet «Movies 1910-2024 (Metacritic)». Набор данных содержит информацию об фильмах, снятых в промежутке между 1910 и 2024 годом.

Таблица имеет 16000 записей. Каждая запись содержит информацию о названии фильма; дате выхода; краткое описание; рейтинг на портале Metacritic; количестве людей, поставивших оценку; режиссёре; сценаристах; продолжительности фильма и его жанре.

1. **Формулирование гипотез**
2. Фильмы в жанре драма имеют самый высокий средний рейтинг, относительно других жанров.
3. Средняя продолжительность большинства фильмов находится между 1 ч. 30 мин. и 1 ч. 40 мин.
4. Десятилетие 1990-2000 – лучший год по средним оценкам в истории кинематографа.
5. Количество выпускаемых фильмов растет с каждым годом.

**Обоснование сформулированных гипотез**

1. Было выдвинуто предположение о том, что фильмы в жанре драма имеют наиболее высокий средний рейтинг по сравнению с фильмами других жанров, так как драматические фильмы находят отклик в глазах зрителей, что повышает их мнение о произведении.
2. Выдвинутое предположение о средней продолжительности фильма, лежащего в интервале между 1 ч. 30 мин. и 1 ч. 40 мин., основывается на том, что фильмы меньшей продолжительности не привлекают зрителей, желающих полностью погрузиться в историю; а фильмы большей продолжительности успевают зрителям надоесть, порождая негативное мнение о произведении.
3. Десятилетие 1990-2000 – лучший год по средним оценкам в истории кинематографа. Данная гипотеза основана на том, что наиболее известные, культовые фильмы, почитаемые большинством кинолюбителей, такие как «Побег из Шоушенка», «Форест Гамп», «Криминальное чтиво», «Матрица», «Зеленая миля», «Шоу Трумана» и др. были сняты именно в этот период.
4. Гипотеза о том, что количество выпускаемых фильмов растет с каждым годом, основана на том, что покупательная способность людей в развитых странах ежегодно увеличивается, а просмотр кинофильмов стал более доступным (вследствие развития технологий, роста количества кинотеатров и пр.), что в свою очередь стимулирует крупные киностудии к съемкам всё большего числа фильмов.
5. **Анализ данных**

Для анализа данных было использовано программное обеспечение «Pandas» совместно с «matplotlib».

Изображение выглядит как Шрифт, текст, Графика, снимок экрана

Автоматически созданное описание

Рисунок 1 – Импортирование зависимостей

**4.1 Загрузка данных в Python**

Для загрузки /csv файла с данными в pandas была использована функция read\_csv().

Изображение выглядит как текст, Шрифт, снимок экрана, число

Автоматически созданное описание

Рисунок 2 – Загрузка .csv файла

**4.2 Проверка данных**

Набор данных был проверен на наличие данных по столбцам. Для этого был использован метод info() для объектов dataframe.

Листинг:

def main():

df = openFile()

df.info()

if \_\_name\_\_ == '\_\_main\_\_':

main()

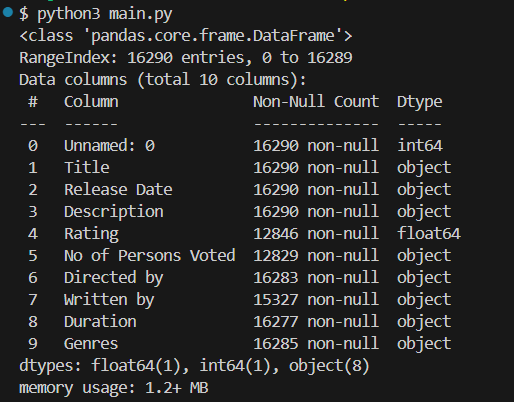


Рисунок 3 – Результат df.info()

**4.3 Очистка данных**

**Работа с дубликатами**

При помощи метода drop\_dubliсates() была совершена попытка удалить записи с полностью совпадающим набором по столбцам: название, дата выпуска, описание, рейтинг, количество оценок, режиссёр, сценаристы, продолжительность, жанр.

Листинг:

def get\_count\_of\_duplicates(df):

    count\_duplicates = len(df) - len(df.drop\_duplicates())

    print(f"Очищено записей: {count\_duplicates}")

def main():

    df = openFile()

    get\_count\_of\_duplicates(df)



Рисунок 4 – Отсутствие повторяющихся данных

**Устранение ошибок**

У некоторых строк в столбце Duration замечено использование некорректных данных – пустых значений или посторонних ссылок. Очистим данные от недопустимых значений с помощью функции df.loc()

Листинг:

import re

class Cleaner:

    def clean\_duration(self, df):

        pattern = r'^(\d+)( h (\d+)( m)?)?$'

        invalid\_count = 0

        valid\_rows = []

        for idx, value in enumerate(df['Duration']):

            if re.match(pattern, str(value)):

                valid\_rows.append(idx)

            else:

                invalid\_count += 1

        df\_cleaned = df.loc[valid\_rows]

        print(f"Удалено строк из-за неверного значения Duration: {invalid\_count}")

        return df\_cleaned

from clean import Cleaner

def main():

    df = openFile()

    cleaner = Cleaner()

    df = cleaner.clean\_duration(df)

    print(df)

Изображение выглядит как текст, снимок экрана, программное обеспечение, Шрифт

Автоматически созданное описание

Рисунок 5 – Очистка столбца Duration