Министерство образования Российской Федерации

Тверской государственный технический университет

Кафедра Программного обеспечения

Исследовательский анализ данных.

Выполнил: Бакеев А.В.

Проверил: Корнеева Е.И.

Тверь, 2025

Оглавление

[Введение 3](#_Toc188262572)

[Постановка задачи 3](#_Toc188262573)

[Описание данных 4](#_Toc188262574)

[Графики зависимостей 5](#_Toc188262575)

[Одномерный анализ 5](#_Toc188262576)

[Многомерный анализ 6](#_Toc188262577)

[Вывод 6](#_Toc188262578)

[Листинг кода 6](#_Toc188262579)

# Введение

Исследовательский анализ данных – это фундаментальный этап любого научного исследования, позволяющий выявить скрытые закономерности, зависимости и тенденции в данных. Одной из ключевых задач ИАД является формулирование и проверка гипотез. В этой лабораторной работе мы сосредоточимся на анализе категориальных данных и постановке гипотез, связанных с ними.

**Категориальные данные** – это данные, которые представляют собой качественные характеристики и выражаются в виде категорий или групп. Примеры категориальных данных: пол, цвет глаз, тип образования, страна проживания. Несмотря на то, что категориальные данные не обладают такой же точностью, как количественные, они могут содержать важную информацию о исследуемой популяции.

**Постановка гипотез** – это процесс формулирования предположений о возможных взаимосвязях между переменными. В контексте категориальных данных гипотезы могут касаться:

* **Частот:** Например, гипотеза о том, что мужчины чаще выбирают определенный бренд, чем женщины.
* **Соотношений:** Например, гипотеза о том, что существует зависимость между уровнем образования и выбором профессии.
* **Ассоциаций:** Например, гипотеза о том, что существует связь между местом проживания и предпочтениями в музыке.

# Постановка задачи

1. Преобразовать категориальные переменные в числовые, если это необходимо. Добавить вычисляемые столбцы.

2. Посчитайте количество строк и столбцов.

3. Проведите разведочный анализ, то есть:

(a) для каждой числовой переменной вычислите:

• Долю пропусков

• Максимальное и минимальное значение

• Среднее значение

• Медиану

• Дисперсию

• Квантиль 0.1 и 0.9

• Квартиль 1 и 3

(b) для каждой категориальной переменной вычислите:

• Долю пропусков

• Количество уникальных значений

• Моду

1. Сформулируйте и проверьте минимум 2 статистические гипотезы. Выбор критериев для проверки гипотез требуется обосновать. Сделать выводы в терминах предметной области.
2. Постройте таблицу корреляции признаков и целевого столбца. Обоснуйте, какой столбец является целевым, а какие признаками.

# Описание данных

Предоставленная мне по варианту информационная база содержит в себе информацию об аренде DVD дисков в различных город Соединенных Штатов – какие фильмы берут, сколько стоит аренда, кто и в каком городе его брал.

Официально описание (переведенное на русский) следующее:

“База данных Sakila sample - это фиктивная база данных, созданная для представления магазина по прокату DVD. Таблицы базы данных включают в себя film, film\_category, actor, customer, rental, payment и инвентарь, а также другие. База данных примеров Sakila предназначена для предоставления стандартной схемы, которую можно использовать для примеров в книгах, руководствах, статьях, примерах и т.д.”

|  |  |
| --- | --- |
| SQLite Sakila Sample Database | <https://www.kaggle.com/datasets/atanaskanev/sqlite-sakila-sample-database?select=SQLite3+Sakila+Sample+Database+ERD.png> |

# 

# Количество строк и столбцов

В каждой выборке 2 столбца (ключ-значение) и 1000 строк, для каждой из таблиц.

# Разведочный анализ

1. Для числовых переменных
   * Доля пропусков – 0%

В полученных данных (суммарно) 0 пропусков

* + Максимальное и минимальное значения

У списка рейтинга фильмов максимум - PG-13 и минимум - G

У списка возвратов максимум - ARABIA DOGMA и минимум - ANACONDA CONFESSIONS

У списка временных периодов аренды максимум - 6 месяц и минимум - 7 месяц

* + Среднее значение

Среднее значение фильмов каждого рейтинга: 200.0

Среднее значение фильмов на каждом из имеющихся языков: 1000.0

Среднея сумма возвратов за фильм: 19.984

* + Медиана

Медиана для возрастных рейтингов: R,

Медиана для временных периодов аренды фильмов: 3 месяцев,

Медиана для сумм возвращенных билетов: 19.99 долларов

* + Дисперсия

Дисперсия рейтингов: 234.8

Дисперсия суммы возвратов: 36.576

Дисперсия месяцев аренды: 64.8

* + Квантиль 0.1 и 0.9

Квантиль рейтингов 0.1: PG;

Квантиль рейтингов 0.9: R.

Квантиль месяцев аренды 0.1: 6 month;

Квантиль месяцев аренды 0.9: 4 month.

Квантиль суммы возвратов 0.1: ('BROOKLYN DESERT', 21.99);

Квантиль суммы возвратов 0.9: ('TOWN ARK', 17.99);

* + Квартиль 1 и 3

Квартиль 1 рейтингов: PG

Квартиль 3 рейтингов: NC-17

Квартиль 1 суммы возвратов: ALONE TRIP

Квартиль 3 суммы возвратов: ALADDIN CALENDAR

Квартиль 1 месяцев аренды: 7 месяц

Квартиль 3 месяцев аренды: 3 месяц

1. Для категориальных переменных
   * Доля пропусков – 0%
   * Количество уникальных значений – 1000 (100%)

* + Мода для суммы возвратов по фильмам: 20.99 самая часто встречающаяся сумма

# Гипотеза 1

Нулевая гипотеза: H0 = {Количество месяцев доступных для проката не влияет на желание людей арендовать фильм с определенным рейтингом}

Альтернативная гипотеза: H1 = {Количество месяцев доступных для проката как-то влияет на желание клиента взять фильм с определенным рейтингом в аренду}

|  |  |  |
| --- | --- | --- |
|  | Нулевая гипотеза (Н0) верна | Нулевая гипотеза (Н0) ложна |
| Нулевая гипотеза отвергнута | 66 | 1027 |
| Нулевая гипотеза не отвергнута | 962 | 1 |

Рейтинг и на сколько месяцев в сумме такие брали:

Рейтинг PG брали на: 200 месяцев

Рейтинг G брали на: 181 месяцев

Рейтинг NC-17 брали на: 217 месяцев

Рейтинг PG-13 брали на: 229 месяцев

Рейтинг R брали на: 201 месяцев

Исходя из полученных данных можно заметить, что распределение происходит не равномерно при том, что все значения, в случаи истинности гипотеза H0 должны были бы быть на одном уровне значения (±1 из-за того что вся сумма не кратна 5). Следовательно нулевая гипотеза ложна, а альтернативная подтверждается.

# Гипотеза 2

Нулевая гипотеза: H0 = {Рейтинг фильма не влиял на то, какие убытки на возвратах нес салон проката, т.е. суммы возвратов у каждого рейтинга были примерно одинаковы}

Альтернативная гипотеза: H1 = {Рейтинг фильма влиял и суммы были различными от одного к другому}

|  |  |  |
| --- | --- | --- |
|  | Нулевая гипотеза (Н0) верна | Нулевая гипотеза (Н0) ложна |
| Нулевая гипотеза отвергнута | 1558 | 19942 |
| Нулевая гипотеза не отвергнута | 18425 | 41 |

Рейтинг и на сколько в сумме вовзратов его было:

Рейтинг PG навозвращали на: 3678.06 долларов

Рейтинг G навозвращали на: 3582.22 долларов

Рейтинг NC-17 навозвращали на: 4228.9 долларов

Рейтинг PG-13 навозвращали на: 4549.77 долларов

Рейтинг R навозвращали на: 3945.05 долларов

Исходя из полученных данных можно заметить, что распределение происходит не равномерно при том, что все значения, в случаи истинности гипотеза H0 должны были бы быть на одном уровне значения (±1 из-за того что вся сумма не кратна 5). Следовательно нулевая гипотеза ложна, а альтернативная подтверждается.

# Талица корреляции

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Название фильма | Месяца аренды | Ограничение (возрастной рейтинг) | Сумма возвратов |
| №1 | YOUNG LANGUAGE | 6 | G | 9.99 |
| №2 | WEREWOLF LOLA | 6 | G | 19.99 |
| №3 | TORQUE BOUND | 3 | G | 27.99 |
| №4 | NONE SPIKING | 3 | NC-17 | 18.99 |
| №5 | STRANGELOVE DESIRE | 4 | NC-17 | 27.99 |
| №6 | MURDER ANTITRUST | 6 | PG | 11.99 |

# На основании построенной таблицы – целевой столбец это сумма возвратов, потому что на него влияют столбцы признаки – доступное число месяцев аренды, а так же возрастной рейтинг фильма. Так сумма возврата может быть одинаковой у фильмов с совершенно разными признаками, равно как и при изменении хотя бы одного из признаков мы получаем измененную сумму возвратов.

# Вывод

В данной лабораторной работе мы рассмотрели фундаментальные аспекты исследовательского анализа данных, сосредоточившись на анализе категориальных данных. Мы изучили, как формулировать и проверять гипотезы о взаимосвязях между категориальными переменными. Особое внимание было уделено важности постановки четких и проверяемых гипотез, которые позволят нам получить достоверные результаты.

Я считаю, что полученные знания помогут мне в дальнейшем проводить более глубокий анализ данных, выявлять скрытые закономерности и принимать обоснованные решения на основе полученных результатов. Анализ категориальных данных является важным инструментом для исследователей в различных областях, и я рад, что освоил основы этого метода.

# Хаб с кодом

https://github.com/ItsRandomNameMeh/-SecondDataScience - проект с файлом кода и БД