# **Netflix project**

## **Исследуйте данные Netflix с помощью нового понимания сводной статистики!**

### **Обзор проекта**

В этом проекте вы будете практиковаться в использовании сводной статистики на основе реальных данных.

Ваши задачи:

* интерпретировать сводные статистические данные в зависимости от контекста
* использовать основные функционалы Google Sheets
* принимать решения о том, какие сводные статистические данные подходят в конкретной ситуации
* изучить некоторые действительно интересные данные
* получить удовольствие от процесса

Давайте перейдем к выполнению задания!

### **Ваша роль**

Вы только что получили очень крутую работу в киноиндустрии! Ваше первое задание - провести исследование контента на Netflix за последние несколько лет. Вы собираете информацию для отчета о видах производимых фильмов, а также о любых закономерностях, которые, возможно, стоит изучить подробнее.

### **База данных**

Вы решили начать свое исследование с изучения данных о фильмах и документальных фильмах производства Netflix. Набор данных, который вы будете использовать, представляет собой модифицированную версию одного из наборов данных, найденных на сайте [Kaggle](https://www.kaggle.com/datasets/luiscorter/netflix-original-films-imdb-scores). Ваш набор данных включает 503 (в самом датасете 584 фильма, но для 4 вопросов из первой части предполагаем, что у нас 503 фильма) фильма со следующими переменными:

* title: название фильма
* genre: жанр фильма
* language: основной язык фильма
* year: год премьеры фильма
* runtime: продолжительность фильма в минутах
* score: рейтинг фильма от 1 до 10 (от худшего к лучшему) с сайта IMDb

**Инструкция по сдаче проекта:**

* После выполнения задания нажмите на кнопку "+ Добавить или создать" и загрузите два файла (docx + xlsx).
* После загрузки файлов нажмите на кнопку "Сдать".
* Ожидайте обратную связь от менторов. Вы - молодцы!

Не забудьте также посмотреть видеоинструкцию в Notion!

## **Часть1: Отдельные переменные**

### **Language - язык**

Вы решили начать с переменной Language. В следующей таблице приведено количество фильмов на каждом языке:

| **ENGLISH** | **SPANISH** | **HINDI** | **FRENCH** | **OTHER single** | **MULTIPLE** |
| --- | --- | --- | --- | --- | --- |
| 360 | 29 | 27 | 15 | 51 | 21 |

Используя таблицу, попробуйте ответить на следующие индивидуальные вопросы. Затем используйте свои ответы для составления краткого резюме об основных языках фильмов.

* **Очевидно, что существует множество фильмов, в которых основным языком является английский. Из 503 фильмов, в какой пропорции английский язык является основным?**

| Ваш ответ:  Я использовал Pie Chart. Ответ на странице Часть 1  На 71.6% **английский язык является основным** |
| --- |

* **Каково соотношение англоязычных фильмов и фильмов на одном языке, который НЕ является английским?**

| Ваш ответ:  Я использовал Pie Chart. Ответ на странице Часть 1  На 10.1% **Other single language** |
| --- |

* **Какая доля фильмов имеет несколько основных языков?**

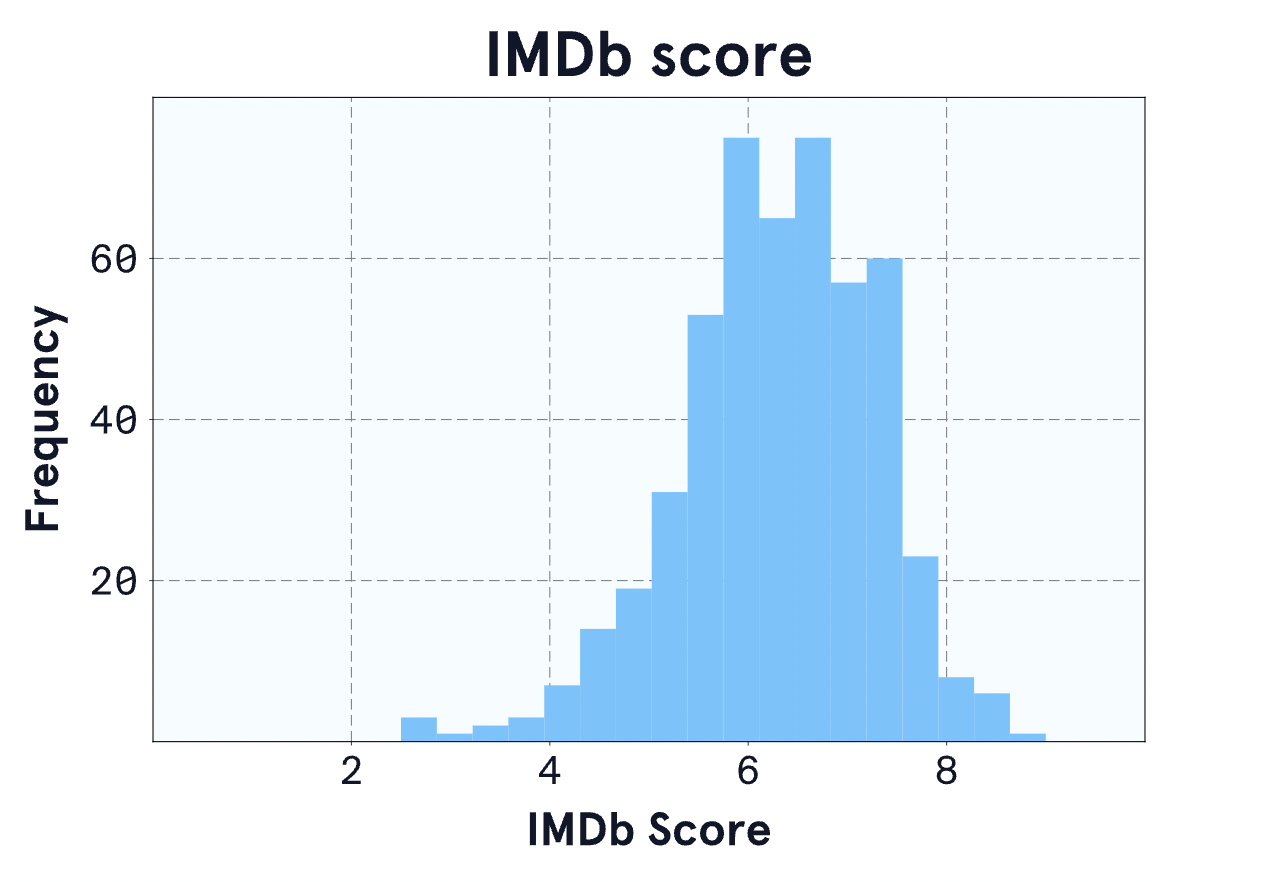
| Ваш ответ:  Я использовал Pie Chart. Ответ на странице Часть 1  На 4.2% Multiple language |
| --- |

На основании вопросов, на которые вы только что ответили, напишите несколько предложений с кратким изложением того, что вы узнали о языках кино и какие выводы вы можете сделать. Приведите все пропорции в процентах, чтобы помочь читателям отчета о вашем исследовании лучше понять ваши выводы.

| Ваш ответ:  Netflix — американская компания, поэтому английский язык доминирует как в США, так и на глобальном уровне, ведь это основной международный язык. Поскольку Netflix ориентирован на мировой рынок, контент на английском закономерно занимает первое место.  Испанский язык на втором месте из-за огромной испаноязычной диаспоры в США, в частности латиноамериканцев.  Фильмы на хинди и французском языке также присутствуют, что отражает глобальную стратегию Netflix по привлечению аудитории из разных стран. |
| --- |

### **Score - Рейтинг IMDb**

Давайте внимательней рассмотрим рейтинги фильмов. Большинство фильмов оцениваются примерно одинаково? Есть ли среди них крайне низкие или высокие оценки?



**Среднее арифметическое:** 6.3

**Стандартное отклонение:** 1.0

Используя график и сводную статистику, опишите *распределение* оценок IMDb.

| Ваш ответ: Здесь видно, что в основном оценки колеблются от 5.8 до 7.8, что говорит о том, что большая часть фильмов имеет оценку выше среднего(6.3). Стандартное отклонение — 1.0, что указывает на то, что оценки не слишком разнятся.  Гистограмма имеет правостороннюю асимметрию, и очень мало фильмов с оценками 3-4 и 8-10. Это, скорее всего, говорит о том, что не так много фильмов с очень негативными или положительными отзывами. |
| --- |

### **Genre - жанр**

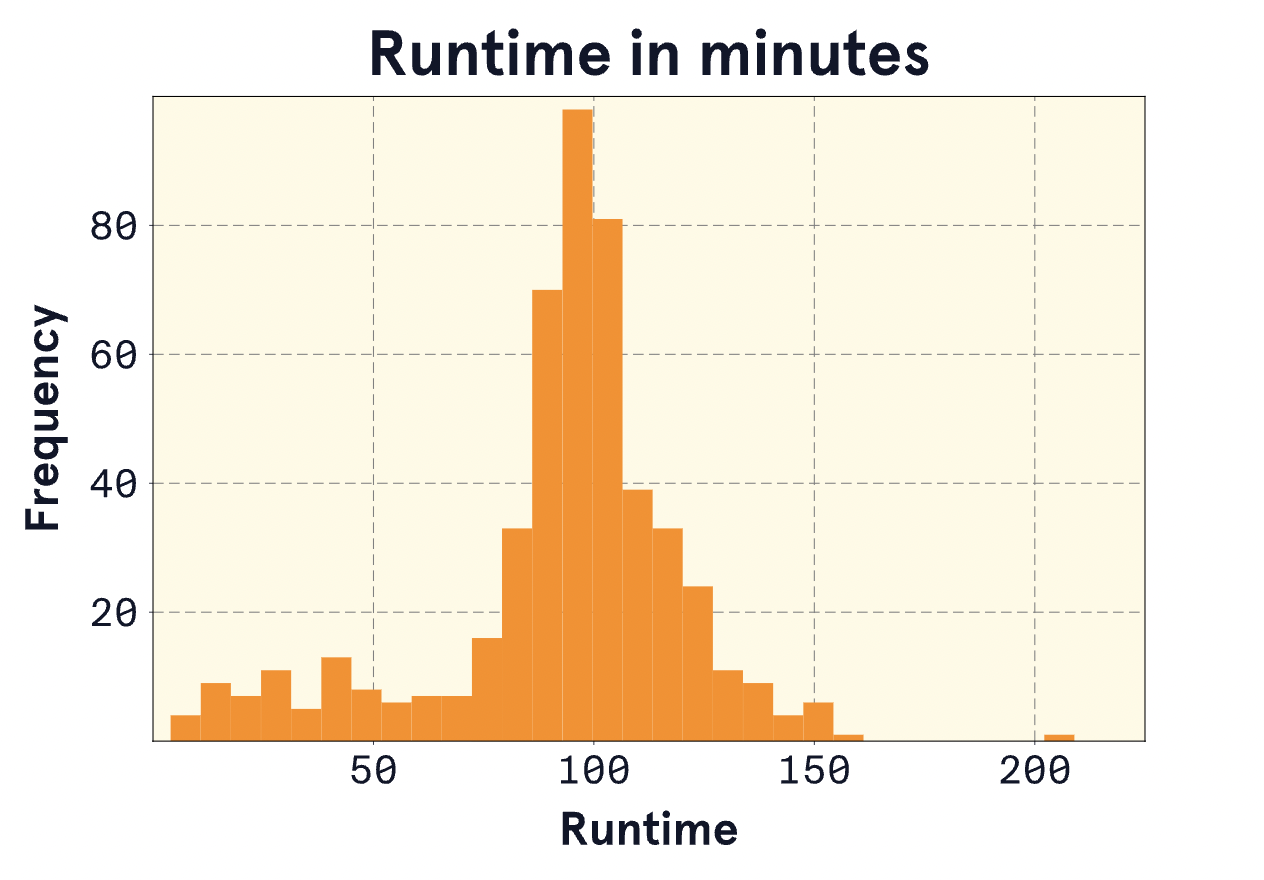
Ваши коллеги просят вас также написать краткое резюме по жанрам фильмов. Какие показатели сводной статистики вы бы могли включить в свой обзор жанров кино?

**Важно:** выбор используемых показателей сводной статистики зависит от типа переменной, которую мы обобщаем. Качественные переменные будут иметь другую сводную статистику, чем количественные переменные.

| Ваш ответ:  Ответ на странице Сводная таблица - 1 Я, используя сводную таблицу, взял в качестве показателей для каждого жанра IMDB следующие данные: стандартное отклонение (stdev), среднее значение (average), максимальные (max) и минимальные (min) оценки, а также количество фильмов (title) в каждом жанре.  Что это может мне дать?  **Stdev:** Позволит понять, в каком жанре зрители больше согласны в своих оценках, а в каком — меньше.  **Average score:** Среднее значение оценок в каждом жанре покажет, какие жанры больше нравятся зрителям, а какие — меньше.  **Max и min:** Эти показатели дадут представление о том, какие жанры могут иметь как очень плохие, так и очень хорошие фильмы.  **Количество фильмов в каждом жанре:** Это покажет, на какие жанры сейчас делают акцент компании. |
| --- |

### **Runtime - длительность фильма**

### Ваш коллега использовал аналитическое программное обеспечение для создания сводки по переменной Runtime. Программа имеет настройки по умолчанию для числовых переменных, которые выводят график, среднее арифметическое и стандартное отклонение. Аналитическое программное обеспечение выдало следующий график и статистику для Runtime:



**Среднее арифметическое:** 92.5

**Стандартное отклонение:** 28.4

* **Исходя из графика распределения, какие опасения у вас могут возникнуть по поводу сводной статистики, используемой аналитическим программным обеспечением?**

| Ваш ответ:  1-Большое стандартное отклонение 2-Наличие выбросов 3-Неравномерное распределение |
| --- |

* **Какие альтернативные статистические параметры мог бы использовать ваш коллега в этом случае?**

| Ваш ответ:  Я думаю тут сред арифм не даст какие-то полезные данные В целом можно использовать статистические параметры такие как: Медиана, мода, IQR |
| --- |

* **Исходя из графика распределения, каковы ваши предположения о медиане и IQR времени выполнения?**

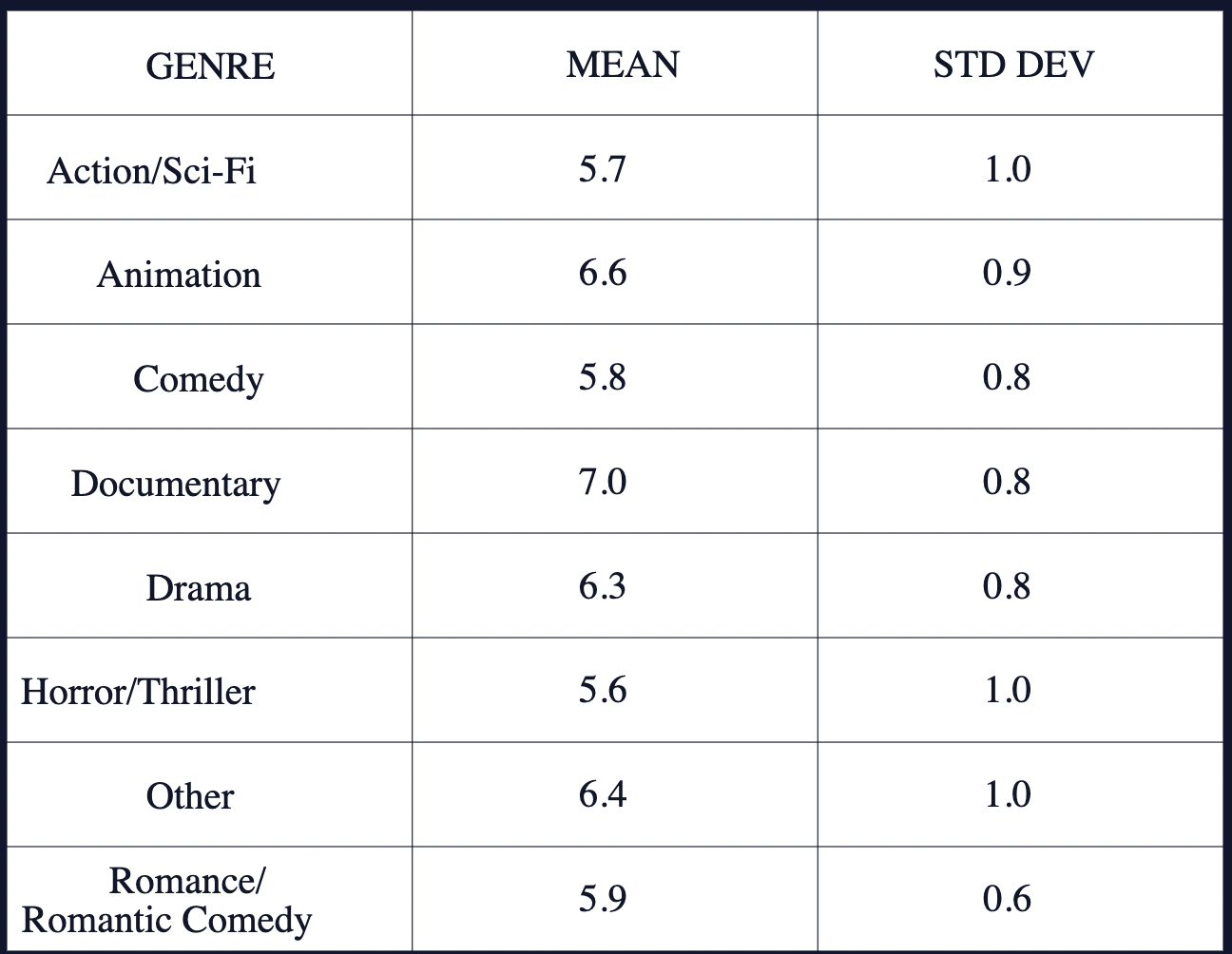
| Ваш ответ:  Исходя из графика распределения, можно предположить, что медиана времени выполнения находится около 100 минут, так как это центральная точка, где сосредоточено много данных. Интерквартильный размах (IQR) вероятно составляет около 40 минут, если первый квартиль (Q1) находится около 80 минут, а третий квартиль (Q3) — около 120 минут. Это говорит о том, что центральные 50% фильмов имеют длительность в пределах 40 минут. Разброс данных указывает на разнообразие в продолжительности фильмов. |
| --- |

## **Взаимосвязи**

Давайте изучим некоторые взаимосвязи между переменными. Агрегирование данных по жанрам и языкам может дать вам некоторое представление.

### **Рейтинг IMDb по жанрам**

Вы хотите узнать, имеют ли какие-либо жанры особенно высокие оценки IMDb, поэтому вам необходимо изучить таблицу средних арифметических и стандартных отклонений оценок IMDb для каждого жанра.

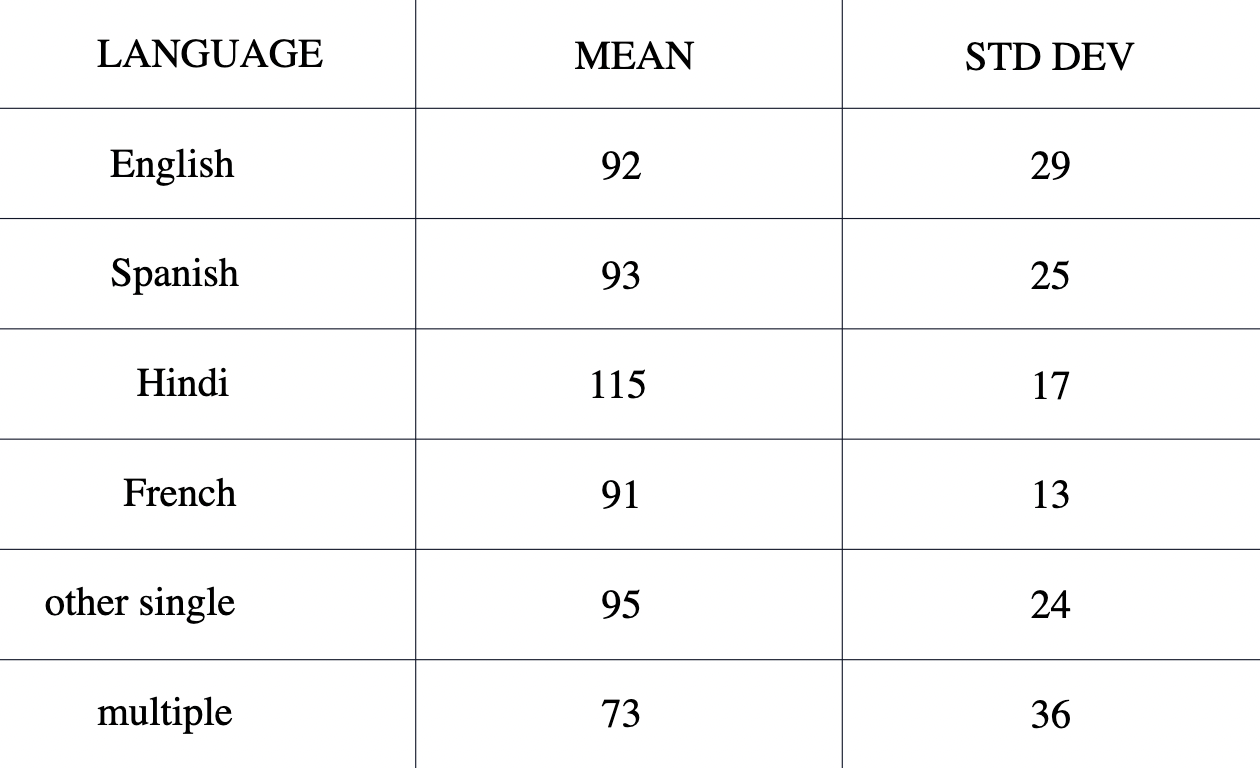


Опишите, что вы узнали об оценках IMDb по жанрам из средних и стандартных отклонений в таблице.

| Ваш ответ:  Ответ на странице “Сводная таблица - 1” Я узнал, что три жанра с самыми низкими средними оценками IMDb — это:  Heist film/Thriller 3,7  Musical/Western/Fantasy 3,9  Horror anthology 4,3  ТОП 3 высокий  Concert Film 7,633333333  Musical / Short 7,7  Animation/Christmas/Comedy/Adventure 8,2  Stdev: Позволит понять, в каком жанре зрители больше согласны в своих оценках, а в каком — меньше. |
| --- |

### **Длительность фильмов и языки**

Вас интересует, есть ли различия в продолжительности фильмов на разных языках.



Опишите, что вы узнали о времени выполнения фильма на разных языках из средних и стандартных отклонений в таблице.

| Ваш ответ:  Среди всех языков стандартное отклонение (стдев) самое низкое у французского, и, скорее всего, большинство фильмов в этом языке имеют более стандартный рантайм по сравнению с другими.  Средний рантайм 73 минуты для фильмов на нескольких языках — это довольно короткое время, и, скорее всего, в таких фильмах снимаются более легкие истории или короткометражки.  Стандартное отклонение 36 минут указывает на значительный разброс, что может означать наличие фильмов в разных жанрах или же на то, что в статистику попало очень мало информации, или и то, и другое.  фильмы на Хинди - Стдев 17 минут указывает на то, что длительность фильмов варьируется, но не слишком сильно. Это говорит о том, что большинство фильмов находятся в пределах 98–132 минут, что также говорит о довольно однородном подходе к созданию фильмов в этой категории. И они имеют довольно привычную длину и соответствует ожиданиям зрителей |
| --- |

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## **Часть 2: Google Sheets задания**

Для этой части, используйте Google Sheets в той же папке, что и этот документ (название: “NetflixDataset”). Обязательно сохраните ваши действия для каждого вопроса отдельной страницей в том же Google Sheets.

Расскажите, какие функционалы Google Sheets вы использовали для того, чтобы ответить на эти вопросы.

1. Около 4% фильмов имеют несколько основных языков. Узнайте, какие жанры, как правило, имеют эту интересную особенность.

| Ваш ответ:   Действия: условное форматирование через значение “/”, выделил в отдельных столбцах все жанры, которые имеют несколько языков, и использовал отдельный столбец с функцией UNIQUE, чтобы вывести основные жанры.  Ответ на странице - Часть 2-1  Основные жанры  Documentary  Black comedy  Crime drama  Drama  Action-adventure  War drama |
| --- |

1. Укажите, топ 3 фильма с самыми высокими и топ 3 фильма с самыми низкими оценками IMDb.

| Ваш ответ:  Я использовал сортировку по рейтингам от А-Я и Я-А  Ответ на странице - Часть 2-2  Top 3 высокий David Attenborough: A Life on Our Planet  Emicida: AmarElo - It's All For Yesterday  Springsteen on Broadway  Тор 3 низкий Enter the Anime  The App  Dark Forces |
| --- |

1. Узнайте, у какого фильма самая высокая продолжительность и у какого фильма самая низкая используя формулы.

| Ваш ответ:  Действия: Вывел runtime и title отдельными столбцами и использовал функцию MAX и MIN на рантайм и XLOOKUP на столбец title.  Ответ на странице Часть 2-3  Title Runtime  The Irishman 209  Sol Levante 4 |
| --- |

1. Посчитайте следующие значения для Runtime и IMDB Score используя формулы:
   1. Сумма всех значении
   2. Среднее арифметическое
   3. Медиана
   4. Мода
   5. Стандартное отклонение
   6. Дисперсия

| Ваш ответ:  Действия: ну тут все просто)  Ответ на странице Часть 2-4  Runtime IMDB Score  sum 54649 3662,7  mean 93,57705479 6,271746575  median 97 6,35  mode 97 5,8  stddev 27,76168283 0,9792563938  varp 769,3913228 0,9573010591 |
| --- |

1. Добавьте новую колонну “For children”, в котором будет “YES”, если фильм для детей и “NO”, если нет (в рамках этого проекта, фильмы для детей - фильмы в жанре анимации).

| Ваш ответ:  Ответ на странице - Часть 2-5  =IF(LEFT(B2;LEN("Animation"))="Animation"; "YES"; "NO") |
| --- |

1. Добавьте новую колонну “For Jessica”, в котором будет “YES”, если фильм документальный и рейтинг ≥8.0 и “NO”, если нет. Посчитайте количество фильмов соответствующих запросу.

| Ваш ответ: Ответ на странице Часть 2-6  Здесь я вывел все "NO" и "YES" с помощью функции IF(AND), а затем посчитал все "YES" с помощью COUNTIF.  **Total "YES"**  13 |
| --- |

1. Подсчитайте количество фильмов, выпущенных каждый год.

| Ваш ответ:  Ответ на странице Часть 2-7  Я выделил года в отдельном столбце в Google Sheets и с помощью функции UNIQUE вывел все уникальные годы в другом столбце. Затем я использовал функцию =COUNTIF(I2:I585; J2), чтобы посчитать, сколько фильмов было в каждом году.  **Года Колво фильм**  2014 1  2015 9  2016 30  2017 66  2018 99  2019 125  2020 183  2021 71 |
| --- |

1. Посчитайте количество фильмов в каждом из жанров и найдите топ 5 популярных жанра.

| Ваш ответ:  Ответ на странице 2-8  Здесь я вывел уникальные жанры и с помощью функции COUNTIF посчитал, сколько фильмов в каждом жанре. Затем, чтобы определить ТОПы, я сделал сортировку.  **Жанры Колво фильм**  Documentary 159  Drama 77  Comedy 49  Romantic comedy 39  Thriller 33 |
| --- |

ответы в отдельных листах в Google Sheets

Вы закончили свой первый отчет об исследовании фильмов! Вы провели большую исследовательскую работу и узнали много нового о различных особенностях контента Netflix. Эти выводы могут стать заключительным шагом или отправной точкой более глубокого анализа. Отличная работа!