

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ  
Федеральное государственное автономное образовательное учреждение высшего образования  
Национальный исследовательский технологический университет «МИСИС»  
Институт информационных технологий и автоматизированных систем управления  
Кафедра Бизнес-информатики и систем управления производством

**Практическая работа №5**

по дисциплине «Статистические методы анализа данных в принятии решений»  
на тему «Виды и анализ распределений данных»

Направление подготовки  
38.03.05 Бизнес-информатика  
Семестр 4

Выполнил:

Сычиков Владимир Андреевич

\_\_\_\_\_  
(ФИО студента)

ББИ-23-6

\_\_\_\_\_  
(№ группы)

23.03.2025

\_\_\_\_\_  
(дата сдачи)

Подпись:

\_\_\_\_\_

Проверил:

\_\_\_\_\_  
(ФИО преподавателя)

\_\_\_\_\_  
(оценка)

\_\_\_\_\_  
(дата проверки)

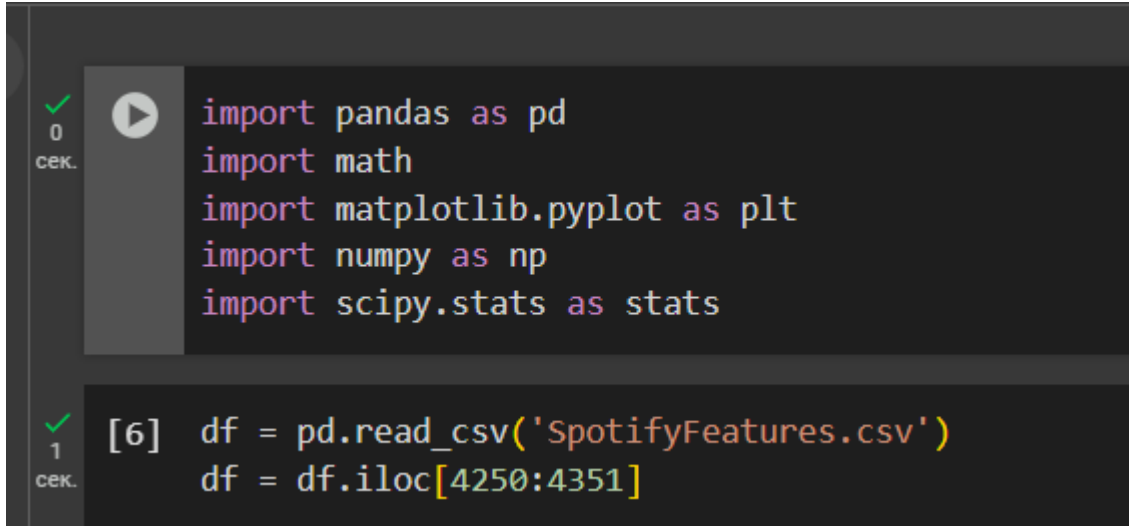
Подпись:

\_\_\_\_\_

Москва – 2025

### Ход работы:

Для начала работы я импортировал все необходимые библиотеки(pandas, numpy) с помощью команды `import`, а также задал элиасы для удобства обращения к библиотеке. Далее я подгрузил выборку и записал ее в датафрейм `df`. Также я выделил нужные мне строки из выборки, по заданию.



```
import pandas as pd
import math
import matplotlib.pyplot as plt
import numpy as np
import scipy.stats as stats

[6] df = pd.read_csv('SpotifyFeatures.csv')
df = df.iloc[4250:4351]
```

Рисунок 1 - Импорт библиотек

На рисунке 2 реализуется процесс построения гистограммы выбранного столбца `popularity`. Гистограмма отображает распределение значений популярности с явным перекосом в левую сторону. Это указывает на то, что большинство значений сосредоточено в нижнем диапазоне популярности. В левой части гистограммы наблюдаются резкие скачкообразные изменения частоты значений, что может свидетельствовать о неравномерном распределении данных. В правой части гистограммы видно, что некоторые значения популярности встречаются крайне редко или вообще отсутствуют, причем их частота не превышает 2. Это может указывать на то, что высокие значения популярности в данных являются аномалиями или выбросами.

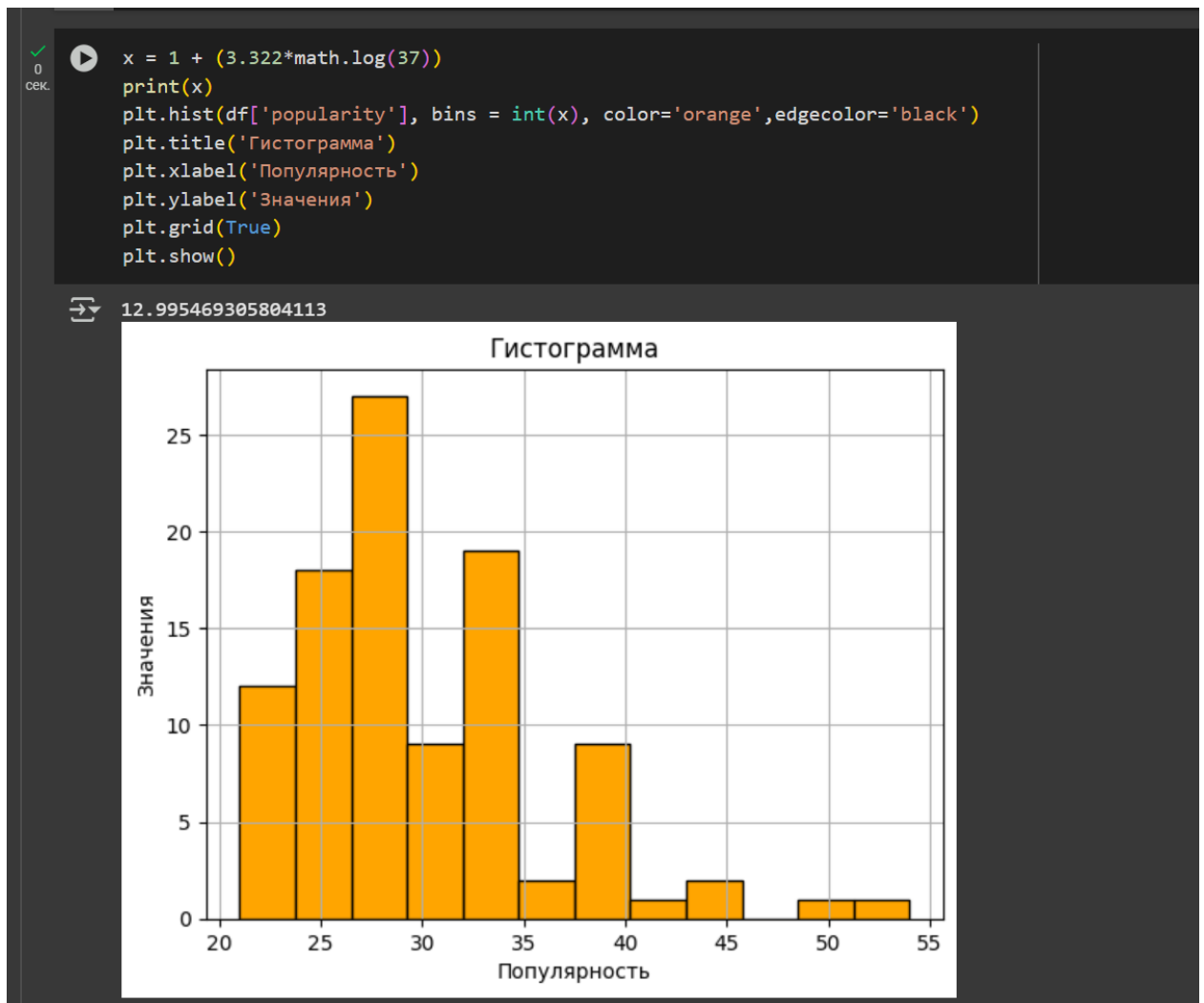


Рисунок 2 – Процесс построения гистограммы 1

На рисунке 3 можем видеть построенный график кумулятивного распределения данных

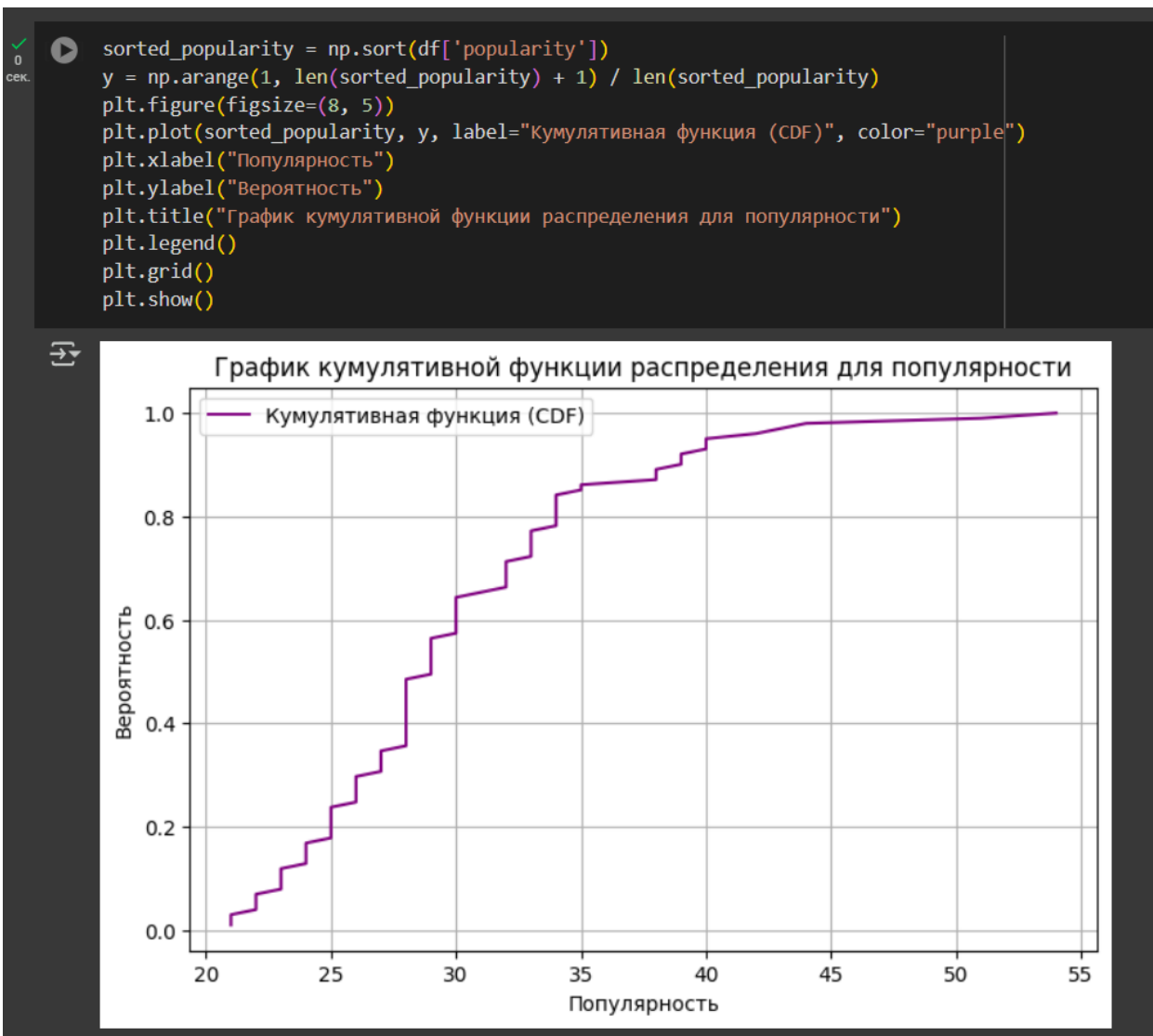


Рисунок 3 – Процесс построения кумулятивной функции распределения

График CDF начинается с низких значений популярности и плавно увеличивается, достигая 1 (100%) для максимального значения.

Плавный рост функции указывает на непрерывное распределение данных без резких скачков или разрывов.

Форма графика подтверждает асимметрию, которую мы наблюдали на гистограмме: большая часть данных сосредоточена в нижнем диапазоне популярности.

График CDF подтверждает, что распределение данных имеет перекося в сторону низких значений популярности. Это согласуется с наблюдениями из гистограммы и подчеркивает, что высокие значения популярности встречаются редко.

На рисунке 4 реализуется процесс построения Boxplot, так называемого «Ящика с усами»

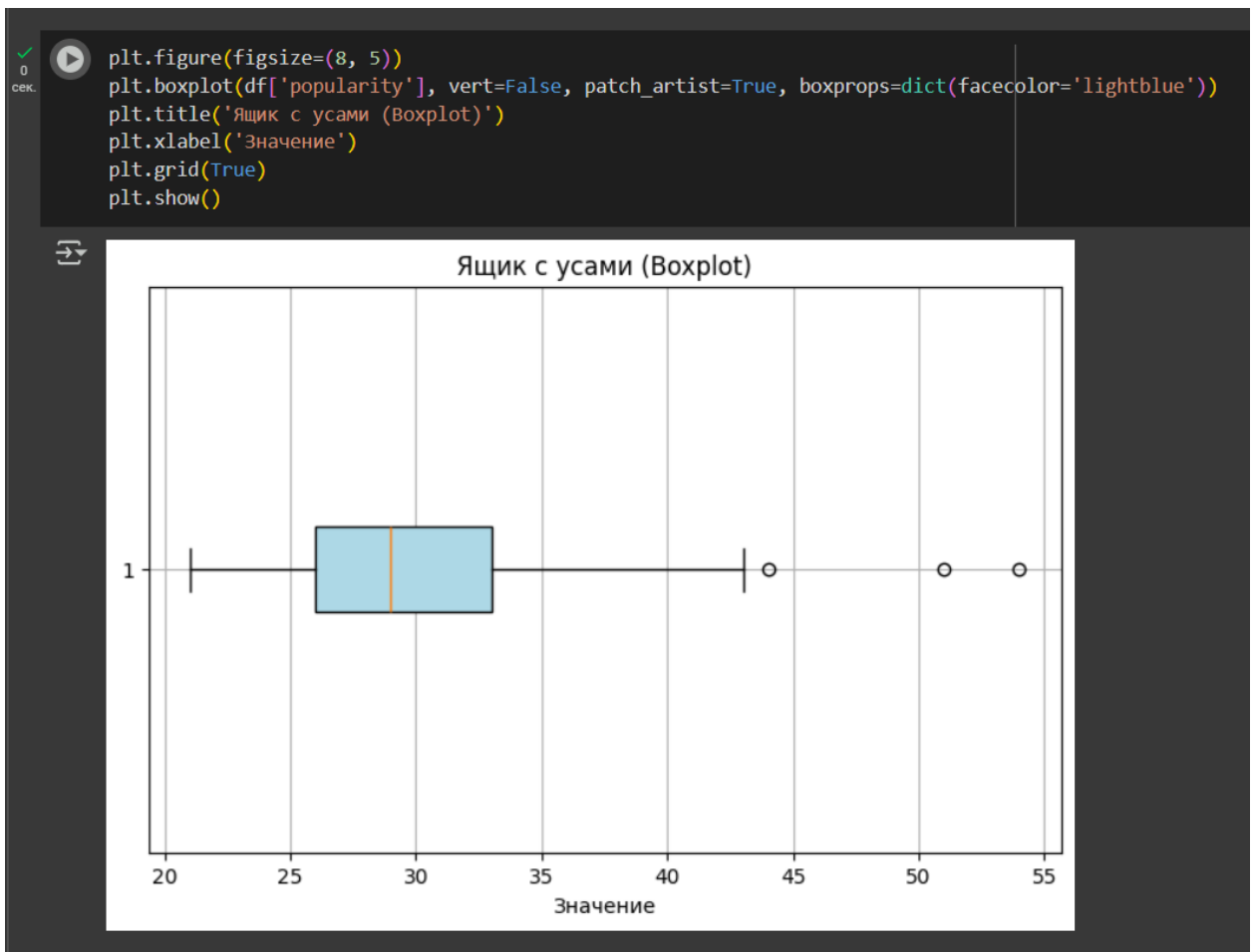


Рисунок 4 – Процесс построения «Ящик с усами»

Медиана (центральная линия в ящике) находится ближе к нижнему краю ящика, что подтверждает асимметрию распределения, которую мы уже наблюдали на гистограмме и графике CDF.

Нижний ус (минимальное значение, исключая выбросы) начинается с низких значений популярности, что согласуется с тем, что большинство данных сосредоточено в нижнем диапазоне.

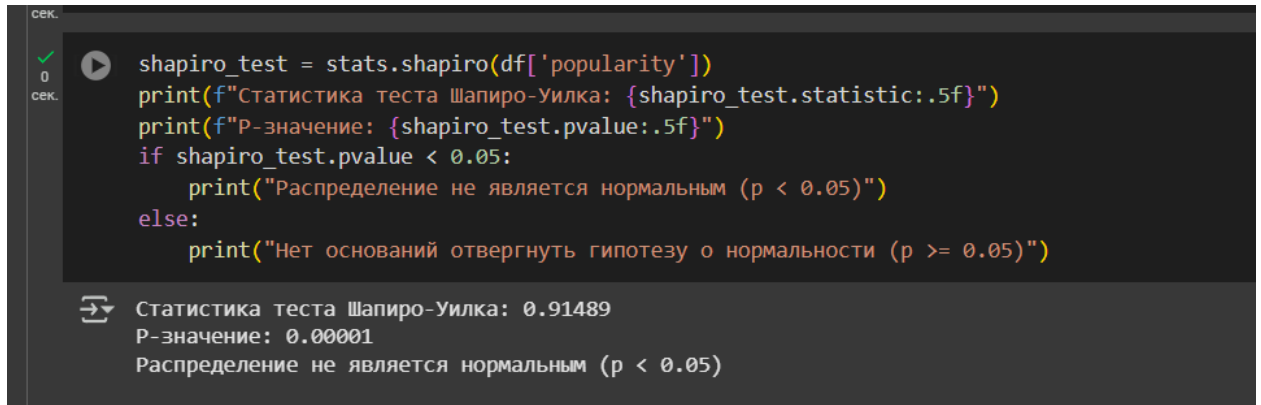
Верхний ус (максимальное значение, исключая выбросы) заканчивается на уровне, который значительно ниже максимальных значений популярности, что указывает на наличие выбросов.

На графике видны выбросы (точки за пределами усов), которые соответствуют редким высоким значениям популярности. Это подтверждает наше предыдущее наблюдение о том, что высокие значения популярности встречаются редко и могут быть аномалиями.

Boxplot подтверждает асимметричное распределение данных с преобладанием низких значений популярности. Наличие выбросов в правой части графика указывает на

редкие, но экстремально высокие значения популярности, которые могут потребовать дополнительного анализа.

На рисунке 5 реализуется процесс проведения тестов Шапиро-Уилка



```
shapiro_test = stats.shapiro(df['popularity'])
print(f"Статистика теста Шапиро-Уилка: {shapiro_test.statistic:.5f}")
print(f"P-значение: {shapiro_test.pvalue:.5f}")
if shapiro_test.pvalue < 0.05:
    print("Распределение не является нормальным (p < 0.05)")
else:
    print("Нет оснований отвергнуть гипотезу о нормальности (p >= 0.05)")
```

Статистика теста Шапиро-Уилка: 0.91489  
P-значение: 0.00001  
Распределение не является нормальным (p < 0.05)

Рисунок 5 – Процесс проведения тестов Шапиро-Уилка

Низкое p-значение (меньше 0.05) указывает на то, что гипотеза о нормальности распределения отвергается. Это означает, что данные не следуют нормальному распределению.

Это согласуется с нашими предыдущими визуальными наблюдениями из гистограммы, графика CDF и Boxplot, где мы видели асимметрию и перекося в распределении данных.

Тест Шапиро-Уилка используется для статистического подтверждения или опровержения гипотезы о нормальности распределения. Визуальный анализ (гистограмма, CDF, Boxplot) уже показал, что данные имеют асимметричное распределение, но тест Шапиро-Уилка добавляет статистическую уверенность в этом выводе. Это важно, потому что многие статистические методы и модели (например, линейная регрессия, t-тесты) предполагают нормальность распределения данных. Если данные не нормальны, это может повлиять на корректность результатов таких методов.

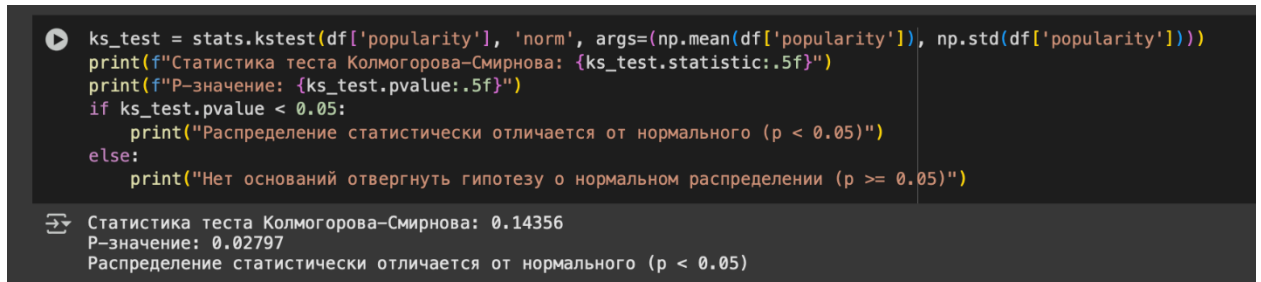
На рисунке 6 реализуется процесс проведения тестов Колмогорова-Смирнова

Тест Колмогорова-Смирнова используется для проверки гипотезы о том, что данные следуют определенному распределению (в данном случае — нормальному). Это важно по нескольким причинам:

— Многие статистические методы (например, t-тесты, ANOVA, линейная регрессия) предполагают, что данные распределены нормально. Если это предположение нарушается, результаты анализа могут быть некорректными. Тест Колмогорова-Смирнова помогает определить, можно ли использовать такие методы.

— Тест позволяет понять, насколько данные отклоняются от нормального распределения. Это полезно для выбора подходящих методов анализа или преобразования данных.

— Визуальные методы (гистограмма, CDF, Voxplot) дают общее представление о распределении, но тест Колмогорова-Смирнова добавляет статистическую уверенность в выводах.



```
ks_test = stats.kstest(df['popularity'], 'norm', args=(np.mean(df['popularity']), np.std(df['popularity'])))
print(f"Статистика теста Колмогорова-Смирнова: {ks_test.statistic:.5f}")
print(f"P-значение: {ks_test.pvalue:.5f}")
if ks_test.pvalue < 0.05:
    print("Распределение статистически отличается от нормального (p < 0.05)")
else:
    print("Нет оснований отвергнуть гипотезу о нормальном распределении (p >= 0.05)")
```

Статистика теста Колмогорова-Смирнова: 0.14356  
P-значение: 0.02797  
Распределение статистически отличается от нормального (p < 0.05)

Рисунок 6 – Процесс проведения тестов Колмогорова-Смирнова

Статистика теста ( $D = 0.14356$ ): Это максимальное расстояние между эмпирической функцией распределения (ECDF) данных и теоретической функцией распределения (CDF) нормального распределения. Чем больше значение  $D$ , тем сильнее данные отклоняются от нормального распределения.

P-значение ( $p = 0.02797$ ): P-значение меньше 0.05, что указывает на статистически значимое отклонение данных от нормального распределения. Это означает, что гипотеза о нормальности распределения отвергается.

На рисунке 7 реализуется процесс проведения тестов Колмогорова-Смирнова для проверки на логнормальное распределение.

Тест Колмогорова-Смирнова для логнормального распределения используется для проверки гипотезы о том, что данные следуют логнормальному распределению. Это важно по нескольким причинам:

— Если данные следуют логнормальному распределению, это может повлиять на выбор методов анализа и моделирования. Например, логнормальное распределение часто используется в финансах, биологии и инженерии.

— Тест позволяет понять, насколько данные отклоняются от логнормального распределения. Это полезно для выбора подходящих методов анализа или преобразования данных.

— Визуальные методы (гистограмма, CDF, Voxplot) дают общее представление о распределении, но тест K-S добавляет статистическую уверенность в выводах.

```
[18] shape, loc, scale = stats.lognorm.fit(df['popularity'], floc=0)
lognorm_test = stats.kstest(df['popularity'], 'lognorm', args=(shape, loc, scale))
print(f"Статистика теста Колмогорова-Смирнова для логнормального распределения: {lognorm_test.statistic:.5f}")
print(f"P-значение: {lognorm_test.pvalue:.5f}")
if lognorm_test.pvalue < 0.05:
    print("Распределение отличается от логнормального (p < 0.05)")
else:
    print("Нет оснований отвергнуть гипотезу о логнормальном распределении (p >= 0.05)")
```

Статистика теста Колмогорова-Смирнова для логнормального распределения: 0.10398  
P-значение: 0.20975  
Нет оснований отвергнуть гипотезу о логнормальном распределении (p >= 0.05)

Рисунок 7 – Процесс проведения тестов Колмогорова-Смирнова на логнормальное распределение

Статистика теста ( $D = 0.10398$ ): это максимальное расстояние между эмпирической функцией распределения (ECDF) данных и теоретической функцией распределения (CDF) логнормального распределения. Чем меньше значение  $D$ , тем ближе данные к логнормальному распределению.

P-значение ( $p = 0.20975$ ): P-значение больше 0.05, что указывает на отсутствие статистически значимого отклонения данных от логнормального распределения. Это означает, что гипотеза о логнормальности распределения не отвергается.

На рисунке 8 реализуется процесс проведения тестов Колмогорова-Смирнова на гамма-распределение.

Оно часто используется для описания:

- Времени между событиями (например, время до отказа оборудования)
- Суммы нескольких экспоненциально распределенных величин
- Данных с положительным перекосом (как в вашем случае)

Графически оно похоже на логнормальное, но имеет более "тяжелый" хвост справа

```
[14] from scipy import stats

# Подгонка гамма-распределения к данным
shape, loc, scale = stats.gamma.fit(df['popularity'], floc=0)

# Проведение теста Колмогорова-Смирнова для гамма-распределения
gamma_test = stats.kstest(df['popularity'], 'gamma', args=(shape, loc, scale))
print(f"Статистика теста Колмогорова-Смирнова для гамма-распределения: {gamma_test.statistic:.5f}")
print(f"P-значение: {gamma_test.pvalue:.5f}")
if gamma_test.pvalue < 0.05:
    print("Распределение отличается от гамма-распределения (p < 0.05)")
else:
    print("Нет оснований отвергнуть гипотезу о гамма-распределении (p >= 0.05)")
```

Статистика теста Колмогорова-Смирнова для гамма-распределения: 0.11771  
P-значение: 0.11225  
Нет оснований отвергнуть гипотезу о гамма-распределении (p >= 0.05)

Рисунок 8 – Процесс проведения тестов Колмогорова-Смирнова на гамма-распределение

Статистика  $D = 0.11771$ : Максимальное расхождение между исходными данными и идеальным гамма-распределением составляет ~11.8%. Значение 0.11771 означает, что



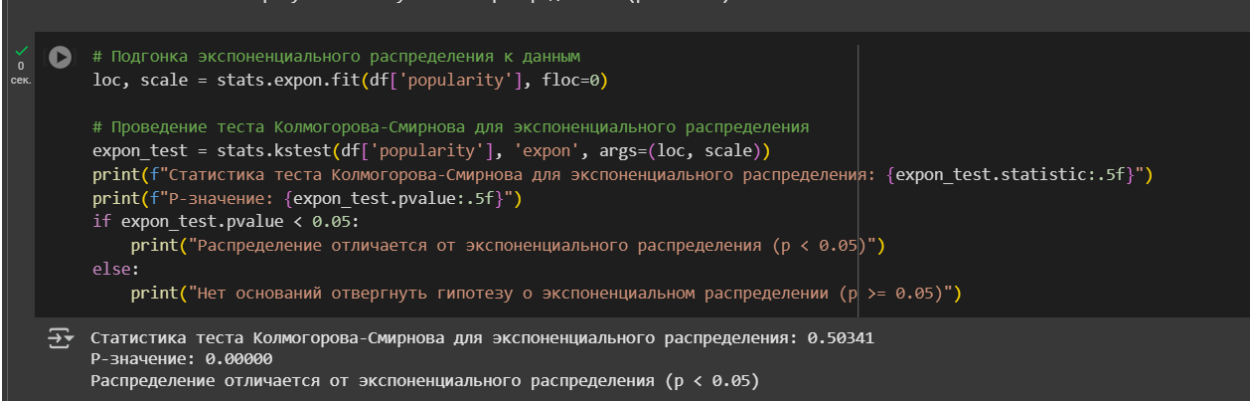
максимальное вертикальное расстояние между реальными данными (ECDF) и теоретической гамма-моделью (CDF) составляет 11.77%. Это умеренное расхождение.

Р-значение = 0.11225: Вероятность получить такое расхождение случайно составляет ~11.2%. При  $p > 0.05$  мы говорим: "Если бы данные действительно были гамма-распределены, вероятность увидеть такое расхождение ( $D=0.11771$ ) составила бы 11.23%". Это недостаточно низко для отвержения гипотезы.

На рисунке 9 реализуется процесс проведения тестов Колмогорова-Смирнова на экспоненциальное распределение.

Экспоненциальное распределение — это частный случай гамма-распределения, описывающий:

- Время между событиями в Пуассоновском процессе (например, интервалы между кликами на сайте)
- Данные с резким спадом вероятности (очень крутая правая часть графика)
- Характеризуется одним параметром scale (среднее время между событиями)



```
# Подгонка экспоненциального распределения к данным
loc, scale = stats.expon.fit(df['popularity'], floc=0)

# Проведение теста Колмогорова-Смирнова для экспоненциального распределения
expon_test = stats.kstest(df['popularity'], 'expon', args=(loc, scale))
print(f"Статистика теста Колмогорова-Смирнова для экспоненциального распределения: {expon_test.statistic:.5f}")
print(f"P-значение: {expon_test.pvalue:.5f}")
if expon_test.pvalue < 0.05:
    print("Распределение отличается от экспоненциального распределения ( $p < 0.05$ )")
else:
    print("Нет оснований отвергнуть гипотезу о экспоненциальном распределении ( $p \geq 0.05$ )")
```

Статистика теста Колмогорова-Смирнова для экспоненциального распределения: 0.50341  
P-значение: 0.00000  
Распределение отличается от экспоненциального распределения ( $p < 0.05$ )

Рисунок 9 – Процесс проведения тестов Колмогорова-Смирнова на экспоненциальное распределение

Статистика  $D = 0.50341$ : Крайне высокое значение (максимально возможное = 1). Это означает, что эмпирические данные и экспоненциальная модель радикально отличаются

Расхождение особенно заметно в:

- Области моды (пика распределения)
- Правом хвосте (экспоненциальное распределение спадает быстрее ваших данных)

Р-значение = 0.00000: Гипотеза об экспоненциальном распределении категорически отвергается. Вероятность получить такое расхождение случайно  $< 0.00001\%$ .

На рисунке 10 реализуется процесс проведения тестов Колмогорова-Смирнова на распределение Вейбулла.

Распределение Вейбулла — это универсальное распределение для анализа:

- Времени до отказа (надёжность оборудования)
- Данных с асимметрией (положительно или отрицательно скошенных)
- Процессов с изменяющейся интенсивностью событий

```
# Подгонка распределения Вейбулла к данным
shape, loc, scale = stats.weibull_min.fit(df['popularity'], floc=0)

# Проведение теста Колмогорова-Смирнова для распределения Вейбулла
weibull_test = stats.kstest(df['popularity'], 'weibull_min', args=(shape, loc, scale))
print(f"Статистика теста Колмогорова-Смирнова для распределения Вейбулла: {weibull_test.statistic:.5f}")
print(f"P-значение: {weibull_test.pvalue:.5f}")
if weibull_test.pvalue < 0.05:
    print("Распределение отличается от распределения Вейбулла (p < 0.05)")
else:
    print("Нет оснований отвергнуть гипотезу о распределении Вейбулла (p >= 0.05)")
```

Статистика теста Колмогорова-Смирнова для распределения Вейбулла: 0.14738  
P-значение: 0.02226  
Распределение отличается от распределения Вейбулла (p < 0.05)

Рисунок 10 – Процесс проведения тестов Колмогорова-Смирнова на распределение Вейбулла

Статистика  $D = 0.14738$ : Значительное расхождение между данными и моделью Вейбулла (максимальное расстояние между ECDF и теоретической CDF). Для сравнения:

- Логнормальное:  $D = 0.10398$
- Гамма:  $D = 0.11771$

P-значение = 0.02226: Гипотеза о соответствии распределению Вейбулла отвергается ( $p < 0.05$ ). Вероятность случайного получения такого расхождения всего 2.2%.

### Итоговая аналитика

Характеристики распределения

Тип распределения: Явно не нормальное (p-value тестов Шапиро-Уилка и Колмогорова-Смирнова  $< 0.05$ )

Форма:

- Асимметрия вправо (положительный перекос)
- "Тяжёлый правый хвост" (наличие редких высоких значений)
- Основная масса данных сосредоточена в нижнем диапазоне

Таблица 1 – Результаты тестов для различных распределений

Распределение	D-статистика	p-value	Вывод
Нормальное	~0.14	<0.05	Не подходит
Логнормальное	0.10398	0.20975	Наилучшее соответствие

Продолжение таблицы 1

Гамма	0.11771	0.11225	Условно допустимо
Экспоненциальное	0.50341	0.00000	Категорически не подходит
Вейбулла	0.14738	0.02226	Не подходит

Рекомендуемая модель: Логнормальное распределение.

Так как:

- Наименьшая D-статистика (минимальное расхождение с данными)
- Высокий p-value ( $>0.2$ ), что подтверждает гипотезу
- Хорошо описывает данные с положительным перекосом и тяжёлыми хвостами

Все графики согласуются с выводами:

- Гистограмма показывает концентрацию данных в левой части
- Boxplot подтверждает наличие выбросов в правом хвосте
- CDF демонстрирует плавный рост, характерный для логнормального распределения

Характер распределения популярности:

- Большинство значений сконцентрировано в нижнем диапазоне (20-35 условных единиц), с редкими "всплесками" до 50-55.
- Напоминает распределение доходов или веб-трафика — много "средних" значений и несколько "вирусных" экстремумов.
- Присутствуют единичные экстремально высокие значения (видно на boxplot и гистограмме)
- Типичная картина — большинство материалов имеет скромную популярность, 1-2% становятся хитами
- Показывает, что большинство авторов малопопулярны, а несколько суперпользователей(авторы) генерируют основную активность.

### **Вывод:**

Проведенное исследование было направлено на всесторонний анализ распределения данных о популярности с использованием современных статистических методов и инструментов визуализации. В ходе работы последовательно применялись различные подходы, включая построение гистограмм, ящиков с усами, графиков кумулятивного распределения, а также проведение статистических тестов (Колмогорова-Смирнова, Шапиро-Уилка) для проверки соответствия различным теоретическим распределениям.

С помощью библиотек Python (pandas, numpy, matplotlib, scipy.stats) был выполнен комплексный анализ, который позволил выявить ключевые особенности распределения данных. Основное внимание уделялось оценке формы распределения, наличию асимметрии, выбросов и степени соответствия различным статистическим моделям. Особое значение имело применение специализированных статистических тестов, которые дали количественную оценку соответствия данных различным теоретическим распределениям.

Анализ показал, что исследуемые данные о популярности имеют ярко выраженное асимметричное распределение с положительным перекосом и "тяжелым правым хвостом". Наиболее адекватной моделью для описания таких данных оказалось логнормальное распределение, что подтверждается как статистическими тестами ( $p\text{-value} = 0.20975$ ), так и визуальным анализом. Это типичная картина для многих социальных и цифровых явлений, где большинство значений сосредоточено в нижнем диапазоне, а редкие экстремальные значения формируют длинный хвост распределения.