Содержание

1. Описание данных

В проекте анализируется сформированная панель (df_aggregated.csv) по 9 компаниям («Золотое яблоко», «Рив Гош», «Л'Этуаль», «Магнит Косметик», «Улыбка радуги», «Иль де Боте», «Ив Роше», «Локситан» и «Подружка») за период 2015–2024 гг. В качестве зависимой переменной выступает выручка (revenue) – ключевой показатель деятельности компании, отражающий объем продаж за год (в млрд руб.). В качестве объясняющих (экзогенных) переменных в модель включены факторы трех групп: показатели потребительских отзывов, поисковая активность и макроэкономические показатели.

Группа данных	**Источник / сбор**	**Агрегирование**	**Итоговые переменные**	**Описание и единицы**
Отзывы	Otzovik, Яндекс.Карты	Сбор «сырых» отзывов по каждому магазину сети, затем годовая агрегация по сети	avg_rating review_vol neg_share avg_sentiment AvgLikes VarRating	- avg_rating (float): средний звёздочный рейтинг за год по сети - review_vol (int): общее число отзывов за год - neg_share (float): доля негативных отзывов (доля отзывов с оценкой ≤2 звезды) - avg_sentiment (float): средняя тональность отзывов по результатам NLP-анализа текста - AvgLikes (float): среднее число реакций на один отзыв за год - VarRating (float): дисперсия рейтингов внутри года, отражает разброс оценок
Поисковая активность	Google Trends (через Pytrends)	Выборка ежемесячных значений Interest for Topic/Brand, затем усреднение по 12 месячным показателям за год	search_index	search_index (float, 0–100): годовой индекс интереса к бренду в Google Trends. Значение 100 соответствует пику за весь период, 0 — минимальный уровень.

Макроэкономика	Росстат	Годовые значения по РФ (фиксируются на уровне всей экономики, одинаковы для всех сетей)	cp_index income_pop FemaleShare	- cp_index (float): индекс потребительских цен (CPI) на конец года - income_pop (float): реальные располагаемые доходы населения, годовое изменение в процентах - FemaleShare (float): доля женщин в общей численности населения РФ (%).
Финансовые показатели	Отчёты компаний, пресс-релизы, аналитические обзоры, базы данных (audit-it, РБК, публичные финансовые отчёты)	Ежегодная выручка по сети: сбор из различных источников, корректировка на инфляцию, конвертация в однородную единицу (млн или млрд руб.)	revenue	- revenue (float): годовая выручка сети в миллионах (или миллиардах) рублей -
Дополнительные переменные	Вычисляются на этапе эконометрического моделирования	создаются лог-преобразования	ln_revenue ln_search_index ln_review_vol	ln_revenue: натуральный логарифм годовой выручки (revenue > 0). Применён для стабилизации дисперсии и интерпретации коэффициентов как эластичностей. ln_search_index: натуральный логарифм индекса поисковой активности. Обычно рассчитывают как ln(search_index + 1) (или +ε), чтобы корректно обработать нулевые значения и снизить скошенность распределения. ln_review_vol: натуральный логарифм объёма отзывов за год. Вычисляется как ln(review_vol + 1) (или +ε) для учёта случаев, когда review_vol может быть 0, и для более нормального распределения в модели.

1. Предварительный анализ собранных данных

В результате предварительной обработки и агрегирования всех собранных данных, была сформирована итоговая панель из 90 наблюдений (9 сетей × 10 лет, 2015-2024 гг.)

	chain_clean	year	avg_rating	review_vol	neg_share	AvgLikes	VarRating	revenue	cp_index	income_pop	FemaleShare	avg_sentiment	search_index
0	l'occitane	2015.0	4.617021	47	0.042553	3.829787	0.893617	3.14	112.90	97.6	0.5367	0.037580	4.916667
1	l'occitane	2016.0	4.331395	172	0.104651	2.302326	1.801816	3.37	105.40	95.5	0.5366	0.047033	4.666667
2	l'occitane	2017.0	4.000000	20	0.150000	15.150000	2.105263	3.57	105.27	99.5	0.5364	0.072223	4.916667
3	l'occitane	2018.0	3.880952	42	0.214286	2.666667	2.595238	3.76	101.62	100.7	0.5361	0.033528	5.333333
4	l'occitane	2019.0	4.173554	121	0.123967	4.090909	1.894628	4.20	104.98	101.2	0.5359	0.065656	5.833333
...	...	...	...	...	...	...	...	...	...	...	...	...	...
85	улыбка радуги	2020.0	4.487995	2957	0.031113	3.971931	0.856165	20.10	102.29	98.0	0.5358	0.100069	11.083333
86	улыбка радуги	2021.0	4.548321	3932	0.033825	3.006104	0.868944	23.20	106.10	103.3	0.5355	0.099553	10.000000
87	улыбка радуги	2022.0	4.476306	3714	0.057351	3.036080	1.156050	27.70	110.00	104.5	0.5349	0.097303	10.166667
88	улыбка радуги	2023.0	4.602261	4689	0.033056	2.647899	0.842836	36.00	111.15	106.1	0.5352	0.095269	8.750000
89	улыбка радуги	2024.0	4.574433	3661	0.051898	2.019940	1.028133	56.60	107.21	107.3	0.5353	0.090334	9.083333

90 rows × 13 columns

Рисунок 1 – Итоговая панель

Проанализировав можно получить целостное представление о том, как эволюционировали рынок парфюмерно-косметического ритейла и отношение покупателей к брендам в России. Перейдем к описательной статистике.

	year	avg_rating	review_vol	neg_share	AvgLikes	VarRating	revenue	cp_index	income_pop	FemaleShare	avg_sentiment	search_index
count	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000
mean	2019.500000	3.952674	1043.977778	0.164717	4.165854	2.172079	40.938000	106.692000	101.370000	0.535840	0.056667	24.202778
std	2.888373	0.427709	1273.194964	0.098158	2.993060	0.760285	46.911849	3.508093	3.695653	0.000587	0.022168	18.079374
min	2015.000000	3.030303	20.000000	0.031113	0.687819	0.842836	3.140000	101.620000	95.500000	0.534900	0.004512	2.833333
25%	2017.000000	3.599442	181.750000	0.087772	2.460247	1.633675	9.207500	104.980000	98.000000	0.535300	0.045882	7.166667
50%	2019.500000	3.989213	438.000000	0.150238	3.142040	2.161127	22.550000	105.750000	100.950000	0.535850	0.055173	22.791667
75%	2022.000000	4.287779	1467.250000	0.228494	4.829631	2.845682	52.575000	110.000000	104.500000	0.536400	0.067861	38.020833
max	2024.000000	4.617021	5877.000000	0.388889	19.241379	3.463425	240.000000	112.900000	107.300000	0.536700	0.102755	74.750000

Рисунок 2 – Подведение описательной статистики

Во-первых, даже при достаточно высоком среднем рейтинге (≈ 3,95 балла) доля резко негативных отзывов остаётся ощутимой — в среднем 16 %, а в отдельных случаях доходит почти до 39 %. Это указывает на скрытую проблему качества сервиса: хорошие впечатления явно доминируют, но каждый пятый-шестой покупатель сталкивается с ситуацией, которую оценивает на единицу или двойку. Для сетей это повод мониторить не столько средний рейтинг, сколько именно «хвост» распределения оценок и оперативно обрабатывать претензии, чтобы не терять лояльность аудитории.

Во-вторых, выручка демонстрирует огромную асимметрию: среднее 40,9 млрд рублей при σ около 46,9 млрд и разбросе от 3,1 до 240 млрд. По сути, несколько сетей-гигантов формируют «ядро» рынка, тогда как локальные сети

работают в совершенно ином масштабе. Эта концентрация подтверждается и Google Trends: годовой индекс поиска варьируется от примерно 3 до 75 баллов; то есть интерес пользователей к лидерам в 20-25 раз выше, чем к аутсайдерам.

Третье наблюдение касается тональности. Средний «avg_sentiment» колеблется от 0,005 до 0,103, оставаясь в узком позитивном диапазоне. Абсолютная величина мала, потому что лексическая модель делит сумму «+1/-1» на общее число слов; однако динамика по годам показала, что у большинства сетей тональность растёт: в 2020-2024 гг. значения на 25-40 % выше, чем в 2015-2016 гг. Это может отражать системную работу ритейлеров над клиентским опытом (улучшение программ лояльности, обновление форматов магазинов), а также общий тренд на более позитивный язык в отзывах. При этом сети с наибольшей выручкой демонстрируют и более высокую тональность, что логично: инвестиции в сервис окупаются ростом лояльности.

Четвёртый вывод связан с макроэкономикой. Индекс потребительских цен плавно растёт (с 102 до 113), реальные доходы населения изменяются в коридоре 95,5-107,3. Это подтверждает, что период 2015-2024 гг. проходил под знаком умеренной инфляции и волатильных доходов, что наверняка влияло на покупательские привычки. Доля женщин в населении практически не меняется (около 53,6 %).

Визуализируем динамику ключевых метрик по годам для каждой сети. Во рассматриваемые годы оценки держатся выше 3,0, но явным «отличником» становится Улыбка радуги (стабильно 4,2-4,6 балла, причём у «Улыбки» виден плавный восходящий тренд вплоть до 2024 г.). Золотое яблоко сначала проваливается к 3,2 в 2017-м, но к 2021 г. восстанавливается до ~4,0 и входит в коридор 3,5-4,0. Наименее стабильны Магнит Косметик и Рив Гош: рейтинги колеблются между 3,1-3,8, что отражает более неоднородное качество обслуживания.

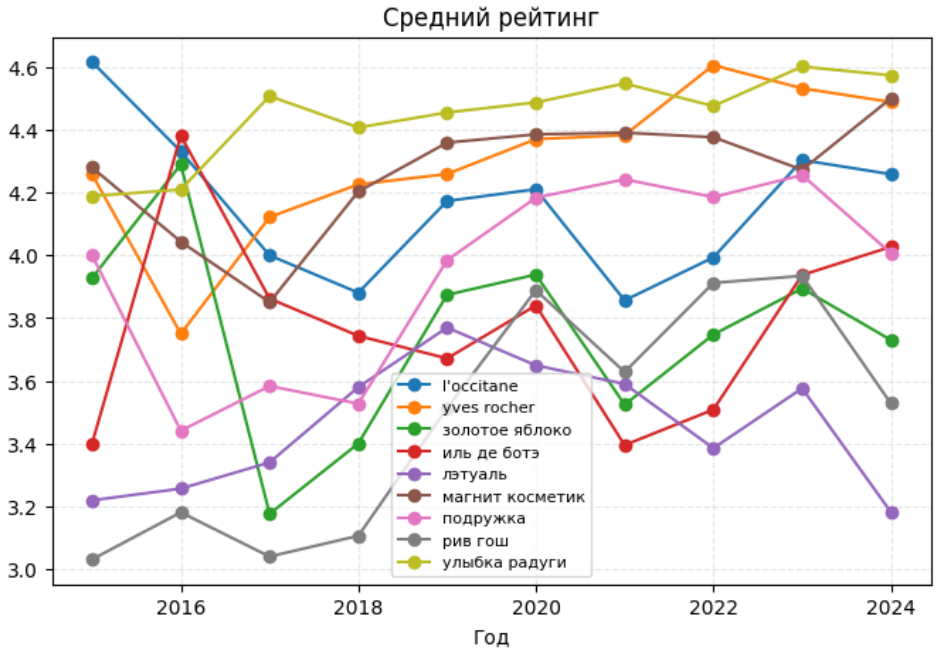


Рисунок 3 – Динамика среднего рейтинга сетей косметики и парфюмерии по годам (2015–2024)

Практически все сети с 2016-го по 2020-й заметно прибавляют в «позитивных словах»; особенно ярко — Магнит Косметик (с 0,01 в 2017 до 0,10 в 2024) и Улыбка радуги (пик 0,10 в 2018). После 2021-го кривые выравниваются: разброс между сетями сужается до 0,04-0,07, что показывает общее повышение «языка лояльности» и снижение различий в коммуникациях. Лидерами тут становятся улыбка радуги и магнит косметик.

Рисунок 4 – Динамика средней тональности сетей косметики и парфюмерии по годам (2015–2024)

До 2019 г. рынок поискового интереса безоговорочно лидирует Подружка (подъём до 75), но затем тренд разворачивается: с 2019-го её популярность падает, а Золотое яблоко с 2018-го растёт экспоненциально (от 5 до 70) и к 2024 году становится первой по упоминаниям. С 2018 года сильно падает популярность Иль де боте, Ив Роше также имеет тенденцию к снижению. Самыми непопулярными оказались Улыбка радуги, Лэтуаль и Локситан.

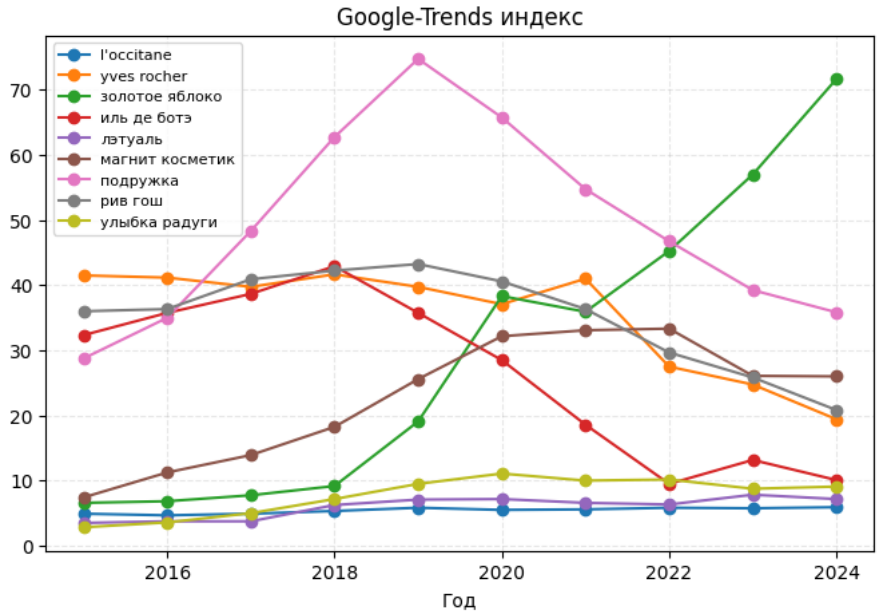


Рисунок 5 – Динамика поискового индекса сетей косметики и парфюмерии по годам (2015–2024)

Как и было предположено ранее, большую часть рынка делит несколько гигантов. Финансовую гонку с большим отрывом выигрывает Магнит Косметик: после скачка 2020 году продажи взлетают до 250 млрд рублей к 2024му. На втором месте Золотое яблоко: экспоненциальный рост начинается в 2021 году и к 2024 достигает 150 млрд рублей, практически догоняя лидера. Л’Этуаль удерживает стабильные 80-90, но темпы роста умеренные. Остальные сети — Рив Гош, Л’Occitane, Yves Rocher и др. — показывают линейный или стагнирующий тренд, оставаясь в диапазоне 5-40 млрд ₽, что подчёркивает усиливающуюся концентрацию рынка вокруг двух-трёх ведущих игроков.

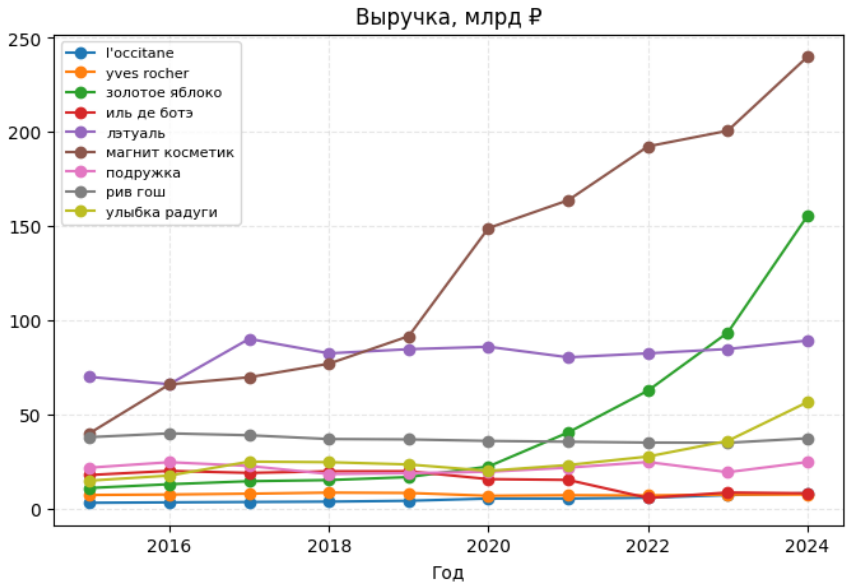


Рисунок 6 – Динамика среднего рейтинга сетей косметики и парфюмерии по годам (2015–2024)

Построим матрицу объединяющую пять ключевых метрик (средний рейтинг, долю негатива, дисперсию оценок, поисковый индекс, выручку и среднюю тональность).Одномерные распределения на диагонали демонстрируют, что почти все ключевые показатели имеют ярко выраженную асимметрию. avg_rating распределён относительно симметрично в интервале 3,0 – 4,6, тогда как neg_share (доля «единиц-двоек»), search_index и особенно revenue тянут длинные правые хвосты: подавляющее большинство наблюдений «скромные», но есть отдельные годы/сети с экстремальными значениями. Метрикатональности avg_sentiment лежит в узком положительном диапазоне (≈ 0,02-0,10) и визуально близка к нормальному «колоколу»: это ожидаемо, т.к. деление на число слов сжимает шкалу.

Попарные облака подчёркивают четыре особо сильные зависимости. Во-первых, идеальный отрицательный пучок avg_rating и neg_share и почти такие же линии avg_rating и VarRating, у neg_share и VarRating (точки ложатся на диагональ) — фактически три переменные передают одну и ту же информацию о «сбалансированности» оценок; их совместное использование в модели может привести к мультиколлинеарности. Во-вторых, avg_rating отчётливо растёт вместе с avg_sentiment, а neg_share падает — значит тональность действительно валидирует числовые оценки. В-третьих, слабое, но положительное облако search_index и revenue подсказывает, что поисковый интерес чаще сопровождает рост продаж; однако разброс по сетям велик, что сигнализирует о необходимости лог- или лаг-преобразований и учёта фиксированных эффектов.

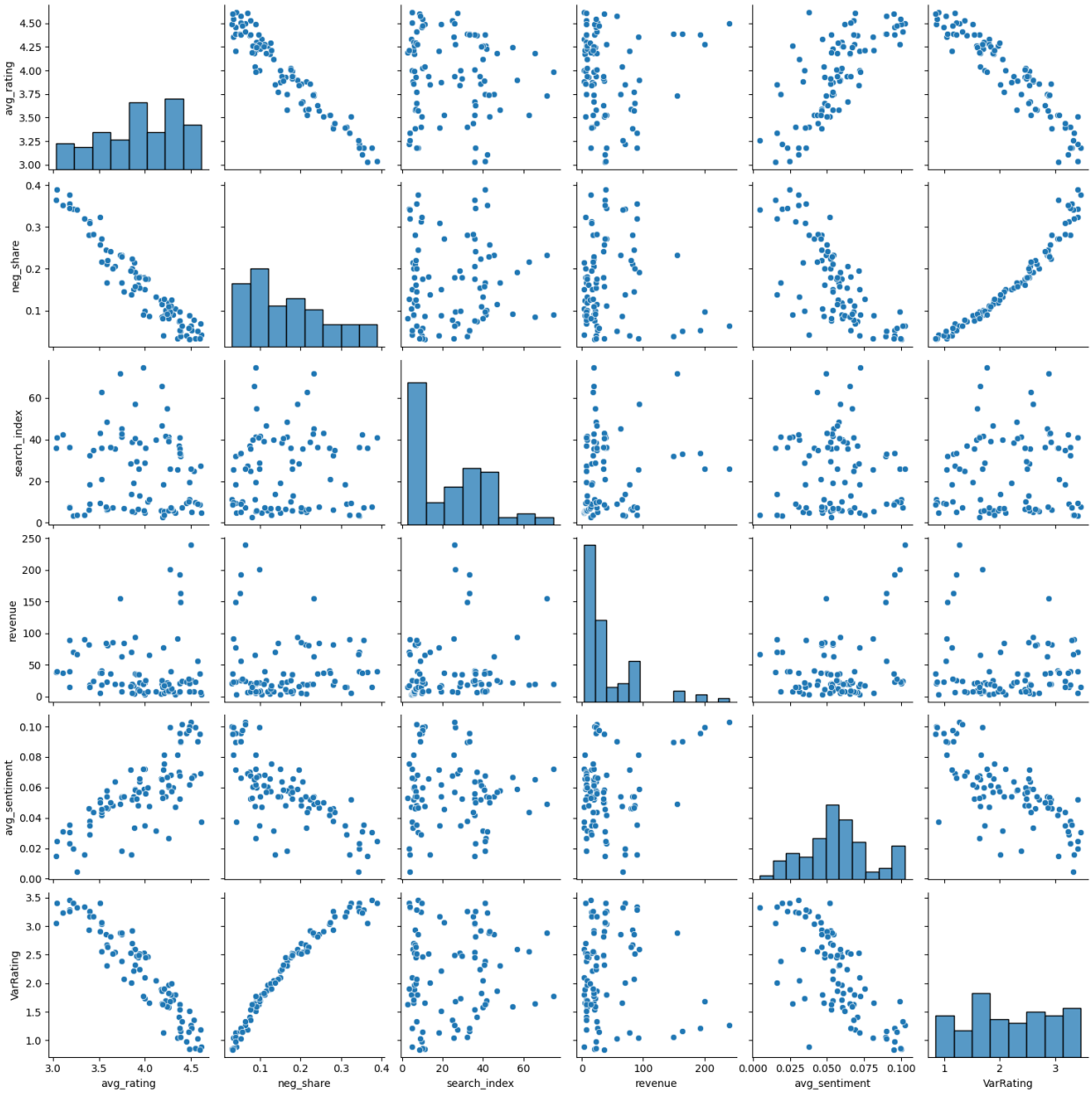


Рисунок 7 – Матрица диаграмм рассеяния и гистограмм распределения ключевых переменных

Корреляционная матрица количественно подтверждает визуальные наблюдения. Коэффициенты −0,96 (avg_rating vs neg_share) и −0,94 (avg_rating vs VarRating) фиксируют практически полную взаимозаменяемость этих метрик; наблюдается положительное 0,97 между neg_share и VarRating. avg_sentiment даёт +0,74 с рейтингом и −0,73 с долей негатива, выступая более «мягким» текстовым индикатором удовлетворённости.Линейная связь search_index-revenue едва превышает ноль (0,07) — при простом скалярном сравнении она теряется за счёт разного масштаба сетей и экспоненциального роста лидеров; значит без лог-преобразований и учёта лагов линия не проявится.

Зато review_vol умеренно коррелирует с ростом доходов населения (0,57) и инфляцией цен (0,19), подтверждая мысль, что всплеск покупательской активности тянет за собой и активность писать отзывы.

Макропоказатели (cp_index, income_pop) ожидаемо связаны между собой (0,30), но с финансовыми и «качественными» метриками корреляции остаются слабыми (< 0,25) — их роль в моделях, таким образом, контрольная, а не объясняющая.

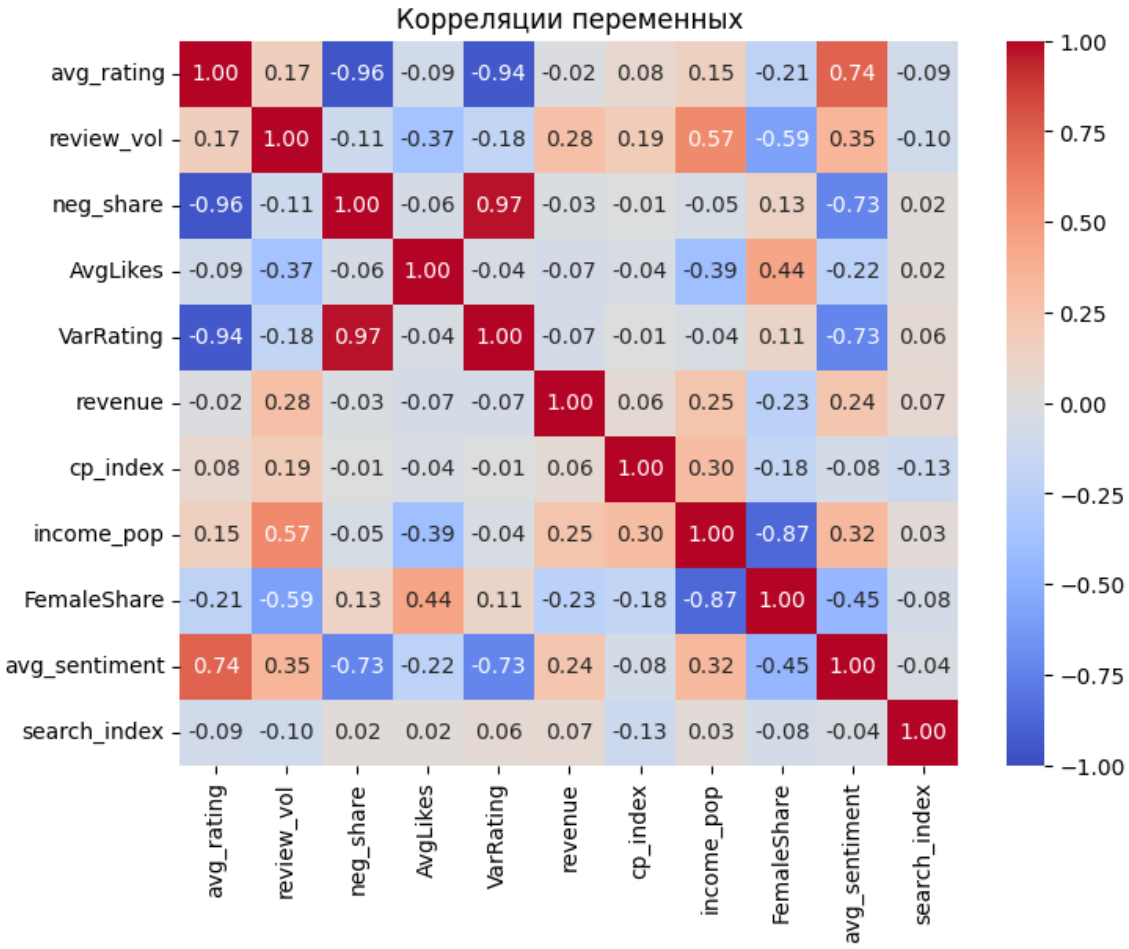


Рисунок 8 – Матрица попарных коэффициентов корреляции между переменными модели

2. Проверка условий использования МНК

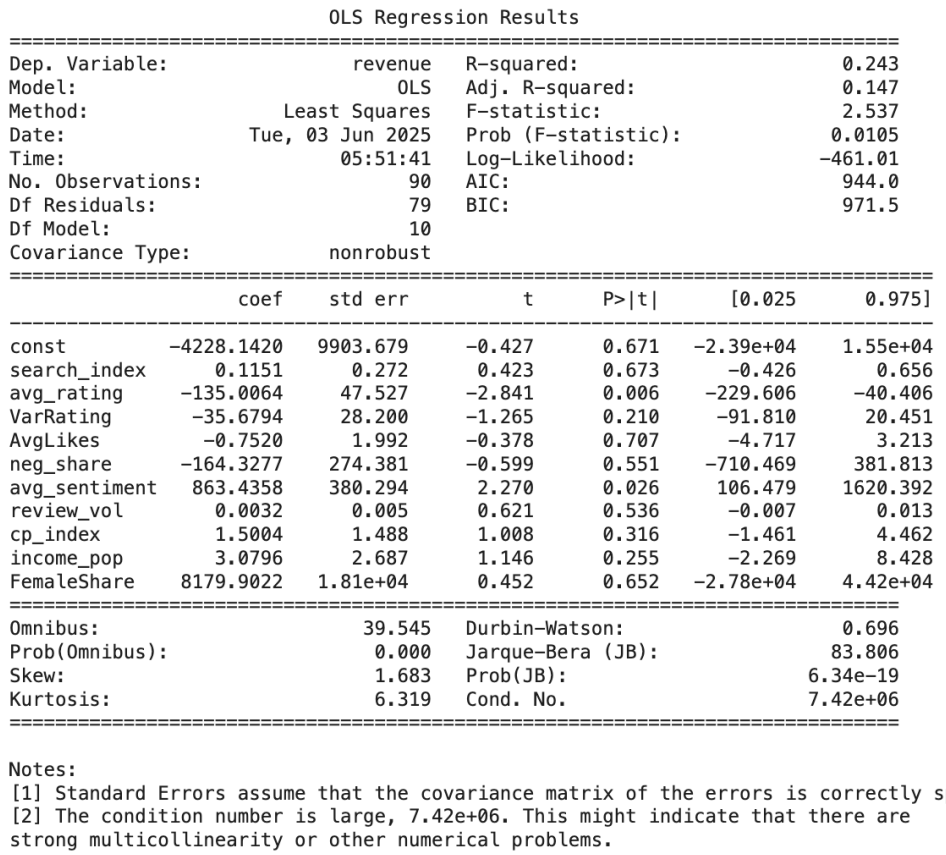
Попробуем построить модель со всеми рассматриваемыми переменными: 'search_index', 'avg_rating', 'VarRating', 'AvgLikes', 'neg_share', 'avg_sentiment', 'review_vol', 'cp_index', 'income_pop', 'FemaleShare'. В полученной «сырой» спецификации (без преобразований и с полной «корзиной» признаков) первое на что обращаем внимание - низкая объяснённость и слабая статистика F-теста.

R^2 всего 0,243 означает, что менее 25 % разброса выручки объясняются этими десятью переменными. F-статистика (2,537, p ≈ 0,0105) говорит, что в целом модель значима, но «добавленная» объясняющая сила многих признаков минимальна.

По avg_rating: β≈−135, std err≈47, t≈−2,84, p=0,006. Отрицательный знак означает, что при прочих равных более высокий средний рейтинг связан с меньшей выручкой. На первый взгляд это невероятно — но, скорее всего, оно обусловлено мультиколлинеарностью: сети с большим оборотом часто «заставляют» пользователей ставить оценки (например, много «нейтральных» отзывов), поэтому связь не по сути, а «маргинальная» в данном наборе переменных. По avg_sentiment: β≈+863, std err≈380, t≈+2,27, p=0,026. Положительный и статистически значимый, указывая, что чем «позитивнее» тексты отзывов, тем выше выручка. Но масштаб β непрозрачен: при изменении avg_sentiment на +0,01 (то есть выигрыш 1 % «позитивных слов» в лексиконе) выручка «растёт» на ≈ 8,6 млн ₽ (863×0,01). Все остальные коэффициенты (по search_index, VarRating, AvgLikes, neg_share, review_vol, cp_index, income_pop, FemaleShare) оказываются незначимыми (p > 0,1).

Наблюдается проблема связанная с мультиколлинеарностью. Строка «Cond. No. = 7,42e+06» и предупреждение о высоком условном числе говорят о сильной мультиколлинеарности. Действительно, мы уже видели, что avg_rating, VarRating и neg_share почти полностью «дублируют» друг друга (r≈0,95–0,97). Подключение всех трёх в одном уравнении делает оценки ненадёжными.

Происходит нарушения стандартных допущений OLS, во-первых, ненормальность остатков: Omnibus = 39,545, Jarque–Bera = 83,806, p ≈ 6.34e-19, соответственно, остатки не нормальны (сильно положительно смещены); во-вторых, сильная автокорреляция: Durbin–Watson = 0,696 (<1) - признаки положительной автокорреляции остатков.



OLS Regression Results

Dep. Variable:	revenue	R-squared:	0.243
Model:	OLS	Adj. R-squared:	0.147
Method:	Least Squares	F-statistic:	2.537
Date:	Tue, 03 Jun 2025	Prob (F-statistic):	0.0105
Time:	05:51:41	Log-Likelihood:	-461.01
No. Observations:	90	AIC:	944.0
Df Residuals:	79	BIC:	971.5
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4228.1420	9903.679	-0.427	0.671	-2.39e+04	1.55e+04
search_index	0.1151	0.272	0.423	0.673	-0.426	0.656
avg_rating	-135.0064	47.527	-2.841	0.006	-229.606	-40.406
VarRating	-35.6794	28.200	-1.265	0.210	-91.810	20.451
AvgLikes	-0.7520	1.992	-0.378	0.707	-4.717	3.213
neg_share	-164.3277	274.381	-0.599	0.551	-710.469	381.813
avg_sentiment	863.4358	380.294	2.270	0.026	106.479	1620.392
review_vol	0.0032	0.005	0.621	0.536	-0.007	0.013
cp_index	1.5004	1.488	1.008	0.316	-1.461	4.462
income_pop	3.0796	2.687	1.146	0.255	-2.269	8.428
FemaleShare	8179.9022	1.81e+04	0.452	0.652	-2.78e+04	4.42e+04

Omnibus:	39.545	Durbin-Watson:	0.696
Prob(Omnibus):	0.000	Jarque-Bera (JB):	83.806
Skew:	1.683	Prob(JB):	6.34e-19
Kurtosis:	6.319	Cond. No.	7.42e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.42e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Рисунок 9 – Результаты OLS-регрессии («сырая» спецификация) с полным набором объясняющих переменных

Проверка предподпосылки о линейности модели по параметрам

В первую очередь проверим, выполняется ли базовое предположение МНК о линейной связи между зависимой переменной и факторами. Сначала построили матрицу графиков рассеивания и линейных зависимостей между «сырыми» (непреобразованными) переменными -— revenue, search_index, avg_rating, avg_sentiment, review_vol, VarRating, AvgLikes. У revenue, review_vol и search_index тяжёлый правый хвост, а scatter-поля «выручка ↔ поиск/объём» образуют веерообразные облака, далёкие от прямой. Визуально это свидетельствовало о нелинейности.

Чтобы подтвердить наблюдение, на исходной OLS-модели мы запустили RESET-тест Рамзи и Rainbow-тест. Rainbow ($p \approx 0.68$) не отвергнул линейность, однако RESET дал $p \approx 0.01 < 0.05$, т. е. наличие квадратичных/пропущенных нелинейных членов было статистически значимым. Такой «разнобой» типичен, когда модель линейна «в среднем», но отдельные факторы требуют трансформации.

Далее мы применили лог-преобразование к переменным (ln_revenue, ln_search_index, ln_review_vol) и качество спецификации заметно улучшилось.

Во-первых, тесты линейности больше не обнаруживают проблем: Ramsey RESET дал p ≈ 0.64, Rainbow — p ≈ 0.35 (> 0.05), значит гипотеза о корректной линейной форме не отвергается. Логарифмирование «сняло» скрытую нелинейность, которую показывал RESET в исходном масштабе.Во-вторых, поясняющая способность модели выросла: R^2 поднялся с 0.24 до 0.42 (Adj R^2 ≈ 0.37). Это означает, что уже более 40 % межсетевой-и-межгодовой изменчивости выручки объясняется переменными.

Среди коэффициентов статистически значимы:

- avg_rating (β ≈ –3.44; p < 0.001) и VarRating (β ≈ –1.23; p = 0.003) — отрицательные знаки подтверждают, что при прочих равных более полярные или более низкие оценки ассоциируются с падающей выручкой;

- avg_sentiment (β ≈ +15.6; p ≈ 0.03) — рост средней тональности на 0.01 приводит к приросту ln(выручки) примерно на 0.156, то есть ≈ 15 %;

- ln_review_vol (β ≈ +0.316; p = 0.003) — эластичность: рост количества отзывов на 1 % ведет к росту выручки на 0.32 % выручки.

ln_search_index, cp_index, income_pop, AvgLikes остаются статистически незначимыми (p > 0.1);

OLS Regression Results

Dep. Variable:	ln_revenue	R-squared:	0.424
Model:	OLS	Adj. R-squared:	0.367
Method:	Least Squares	F-statistic:	7.444
Date:	Tue, 03 Jun 2025	Prob (F-statistic):	2.18e-07
Time:	06:27:18	Log-Likelihood:	-107.02
No. Observations:	90	AIC:	232.0
Df Residuals:	81	BIC:	254.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.9537	4.279	3.027	0.003	4.439	21.468
ln_search_index	0.0677	0.109	0.622	0.535	-0.149	0.284
avg_rating	-3.4375	0.692	-4.967	0.000	-4.814	-2.061
avg_sentiment	15.6410	7.119	2.197	0.031	1.476	29.806
ln_review_vol	0.3160	0.101	3.119	0.003	0.114	0.518
cp_index	0.0396	0.028	1.394	0.167	-0.017	0.096
income_pop	-0.0090	0.035	-0.256	0.798	-0.079	0.061
VarRating	-1.2354	0.401	-3.080	0.003	-2.033	-0.437
AvgLikes	0.0348	0.038	0.927	0.357	-0.040	0.110

Omnibus:	5.842	Durbin-Watson:	0.644
Prob(Omnibus):	0.054	Jarque-Bera (JB):	5.874
Skew:	0.623	Prob(JB):	0.0530
Kurtosis:	2.878	Cond. No.	1.19e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.19e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Рисунок 10 – Результаты OLS-регрессии логарифмическим преобразованием ряда переменных

Проверка предподпосылки об отсутствии мультиколлинеарности

После того как мы утвердили линейную форму модели, следующим шагом стала диагностика мультиколлинеарности — скрытых линейных зависимостей между регрессорами, которые раздувают стандартные ошибки и делают коэффициенты нестабильными. Для этого мы собрали матрицу признаков, использованную в лог-модели и рассчитали коэффициенты VIF (Variance Inflation Factor) для каждой колонки. VIF показывает, во сколько раз увеличена дисперсия β-оценки данного регрессора из-за линейной связи с остальными; безопасным считается диапазон 1–5, значения > 10 сигнализируют о серьёзной коллинеарности.

По результату видим, что avg_rating и VarRating (VIF около 11–12) явно дублируют друг друга: чем ниже средний рейтинг, тем выше дисперсия оценок — это мы уже видели на корреляционной теплокарте. Совместное включение ведёт к завышенным стандартным ошибкам. Все прочие регрессоры имеют VIF < 5, серьёзных проблем нет.

Variance Inflation Factor (VIF):

	Variable	VIF
0	const	2348.995905
1	VarRating	11.795267
2	avg_rating	11.112940
3	avg_sentiment	3.159151
4	ln_review_vol	2.357083
5	income_pop	2.145735
6	AvgLikes	1.606193
7	cp_index	1.261046
8	ln_search_index	1.063280

Рисунок 11 – Результаты VIF

Отсюда вывод: чтобы удовлетворить предпосылку об отсутствии сильной мультиколлинеарности, достаточно исключить один из двух тесно связанных индикаторов качества. После исключения VarRating максимальный VIF среди объясняющих переменных упал с 11,79 до < 2.99. Это значит, что линейные зависимости между регрессорами теперь слабые.

```
VIF после удаления VarRating:
          Variable          VIF
0            const  1624.450455
3    avg_sentiment     2.990369
2       avg_rating     2.420622
4    ln_review_vol     2.221314
6       income_pop     1.937811
7         AvgLikes     1.362099
5         cp_index     1.259042
1  ln_search_index     1.061431
```

Рисунок 11 – Результаты VIF после удаления переменной

Обратим внимание на то, что после исключения объясняющая сила модели умеренно снизилась, R^2 опустился с 0.42 до 0.36 (Adj R^2 0.30), то есть мы потеряли часть «механического» объяснения, которое давал дублирующий показатель дисперсии, но избавились от его искажающего влияния.

Ключевые выводы по модели:

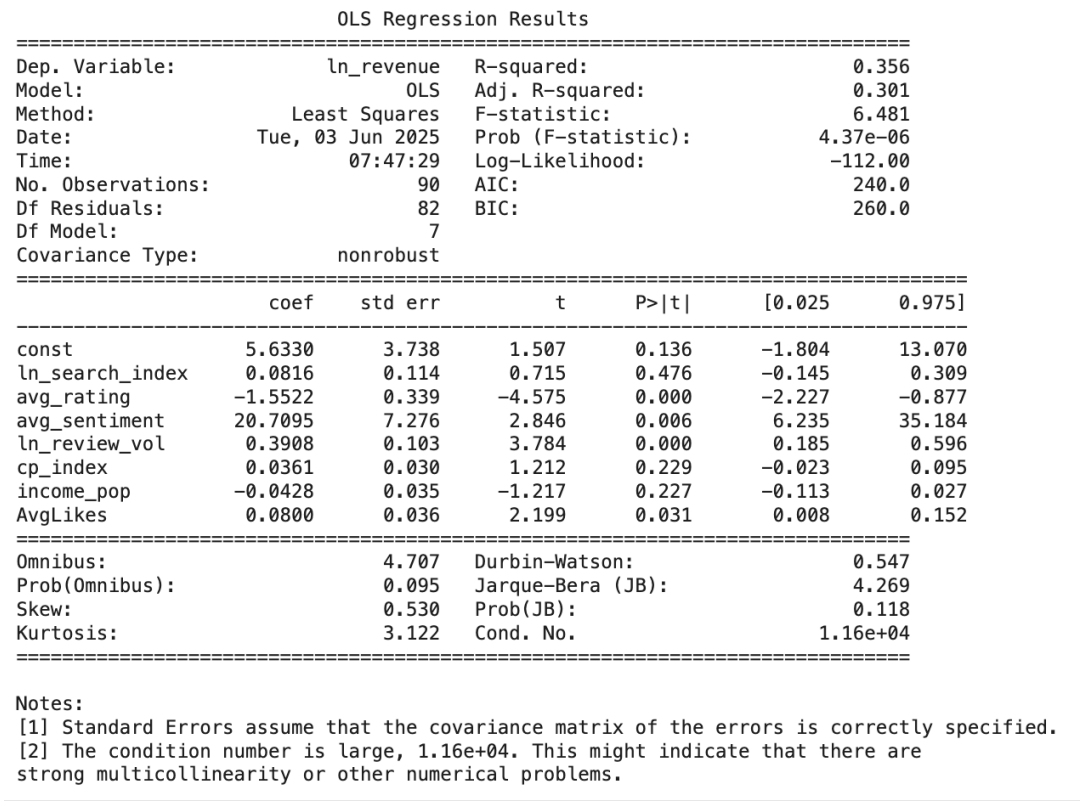
- avg_rating остаётся сильно отрицательным ($\beta \approx -1.55$, $p < 0.001$): при прочих равных рост среднего рейтинга на 0.1 балла связан с падением логарифма выручки на 15 %. Это «инвертированный» знак указывает, что сама по себе высокая оценка характерна для малых или нишевых сетей, а обороты делают крупные игроки, где диапазон оценок шире.

- avg_sentiment положителен ($\beta \approx +20.7$, $p \approx 0.006$): +0.01 к средней тональности коррелирует с +0.207 к ln выручки (≈ +21 %).

- ln_review_vol ($\beta \approx +0.39$, $p < 0.001$) показывает, что эластичность выручки по числу отзывов около 0.4 %: всплеск активности клиентов в комментариях идёт рука-об-руку с ростом продаж.

- AvgLikes впервые становится значимым ($\beta \approx +0.08$, $p \approx 0.03$) — дополнительный индикатор вовлечённости аудитории.

ln_search_index, cp_index, income_pop по-прежнему статистически нейтральны.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              ln_revenue   R-squared:                       0.356
Model:                             OLS   Adj. R-squared:                  0.301
Method:                  Least Squares   F-statistic:                     6.481
Date:                 Tue, 03 Jun 2025   Prob (F-statistic):           4.37e-06
Time:                         07:47:29   Log-Likelihood:                -112.00
No. Observations:                   90   AIC:                             240.0
Df Residuals:                       82   BIC:                             260.0
Df Model:                            7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.6330      3.738      1.507      0.136      -1.804      13.070
ln_search_index  0.0816      0.114      0.715      0.476      -0.145       0.309
avg_rating    -1.5522      0.339     -4.575      0.000      -2.227      -0.877
avg_sentiment 20.7095      7.276      2.846      0.006       6.235      35.184
ln_review_vol  0.3908      0.103      3.784      0.000       0.185       0.596
cp_index       0.0361      0.030      1.212      0.229      -0.023       0.095
income_pop    -0.0428      0.035     -1.217      0.227      -0.113       0.027
AvgLikes       0.0800      0.036      2.199      0.031       0.008       0.152
==============================================================================
Omnibus:                        4.707   Durbin-Watson:                   0.547
Prob(Omnibus):                  0.095   Jarque-Bera (JB):                4.269
Skew:                           0.530   Prob(JB):                        0.118
Kurtosis:                       3.122   Cond. No.                     1.16e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.16e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Рисунок 13 – Результаты OLS-регрессии после удаления переменной

Проверка предпосылки о нормальности распределения остатков

После того как была уточнена спецификация модели (лог-форма переменных, устранена мультиколлинеарность), переходим к проверке одного из ключевых условий классической МНК — нормальности распределения остатков. Применили два графических инструмента. Q-Q-плот показал, что большинство точек располагается вдоль теоретической 45-градусной линии; лёгкое отклонение лишь у 2–3 крайних правых наблюдений, но систематического изгиба (S- или C-формы) не наблюдается. Гистограмма с ядерной оценкой плотности дополнила картину: распределение остатков симметрично, вершина сосредоточена около нуля, а хвосты визуально близки к нормальным.

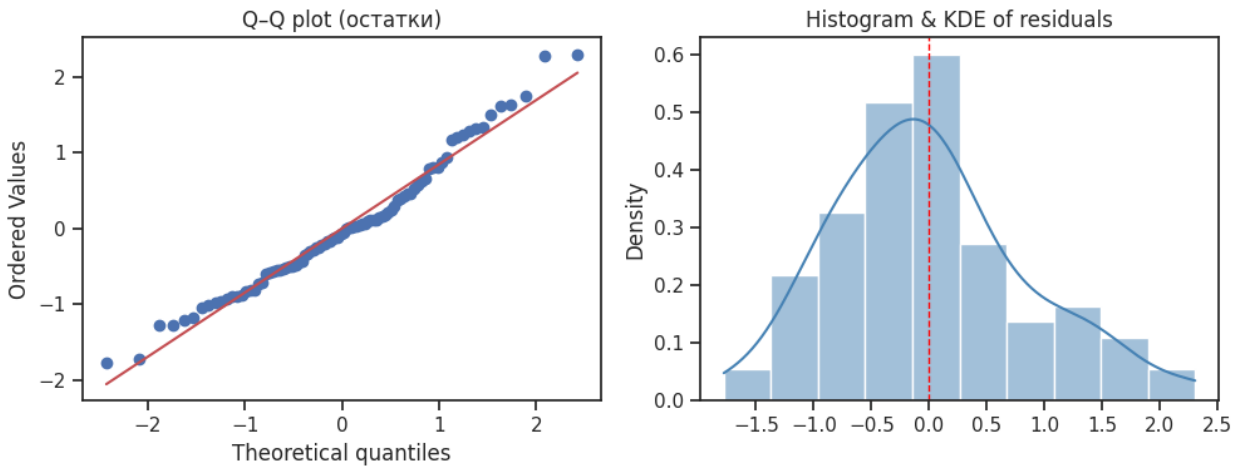


Рисунок 14 – Диагностика нормальности остатков

Подтвердим вывод тестами: объединённый тест Омнибуса (D'Agostino-Pearson), Шапиро–Уилк и альтернатива D'Agostino K^2. Во всех случаях p-значения оказались значительно выше 0,05 (Omnibus p = 0,095; Shapiro p = 0,071; K^2 p = 0,095). Это означает, что нулевая гипотеза о нормальности ошибок не отвергается на стандартном уровне значимости.Таким образом, визуальные и количественные критерии сходятся: распределение ошибок модели совместимо с нормальным законом.

Проверка предподпосылки о гомоскедастичности остатков

На полученном графике не наблюдается явного «веерообразного» расширения или сужения облака точек: разброс остатков вокруг нулевой линии выглядит примерно равномерным по всему диапазону. Это говорит о том, что в «грубой» картине гомоскедастичность не противоречит данным.

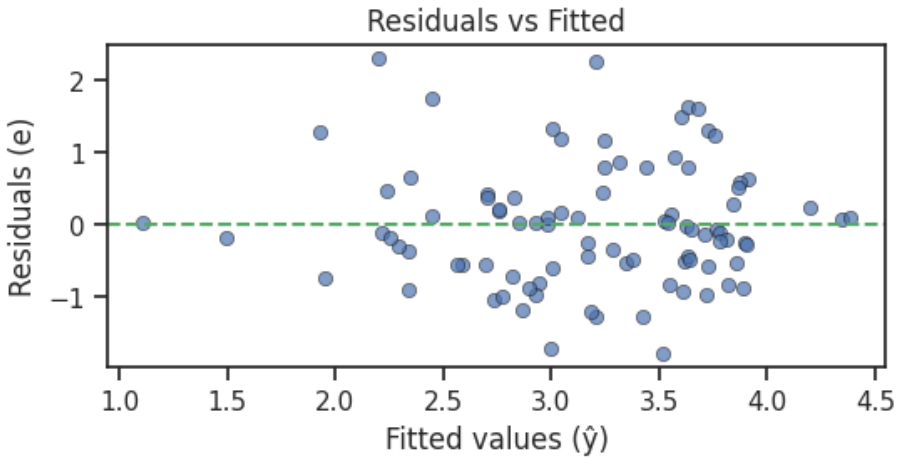


Рисунок 15 – График «остатки – прогнозные значения»

Однако формальные тесты выявили более сложную картину. Breusch–Pagan-проверка дала LM-статистику ≈ 11.35 с p-value ≈ 0.12, то есть при учёте лишь линейной зависимости от исходных регрессоров мы не отвергаем гипотезу гомоскедастичности ($p > 0.05$). Тем не менее White-тест (который дополнительно включает в регрессии квадраты и взаимодействия признаков) дал очень низкий p-value ≈ 0.001, указывая на наличие нелинейно выраженной гетероскедастичности.

Проверка предпосылки об отсутствии автокорреляции

Проводим три классических теста: Durbin–Watson для улавливания автокорреляции первого порядка, Breusch–Godfrey для более комплексной проверки автокорреляции до третьего лага и Ljung–Box для оценки автокорреляционных связей сразу по первым пяти лагам. Статистика Durbin–Watson оказалась равной примерно 0.547. Об отсутствии авткорреляции свидетельствует значение DW близкое к 2,в нашем случае оно существенно меньше, это указывает на положительную автокорреляцию.

Далее был проведён Breusch–Godfrey тест с тремя задержками (nlags=3). p-значения также оказались ниже 0.05, что свидетельствует о наличии автокорреляции вплоть до третьего лага.

И наконец третий Ljung–Box тест с проверкой автокорреляции до пятого лага. Полученная статистика ≈ 95.69 и p-value ≈ 4.3×10^{-19} также с огромным запасом отвергают гипотезу об отсутствии автокорреляции. Это означает, что на одном из первых пяти лагов автокорреляция явно присутствует.

Таким образом, все три диагностики сходятся, что в остатках нашей OLS-модели есть существенная положительная автокорреляция. Это нарушение предпосылки МНК об отсутствии коррелированных ошибок означает, что полученные стандартные ошибки β-коэффициентов и сопровождающие их статистические выводы потенциально некорректны.

3. Переход к панельным методам

4.1. Pooled OLS (Модель пула, модель сквозной регрессии)

В первой модели (Pooled OLS без коррекции ошибок) мы оценили зависимость логарифма выручки (ln_revenue) от набора регрессоров: логарифма поискового индекса (ln_search_index), среднего рейтинга (avg_rating), средней тональности отзывов (avg_sentiment), логарифма объема отзывов (ln_review_vol), индекса потребительских цен (cp_index), реальных

располагаемых доходов населения (income_pop) и среднего числа лайков на отзыв (AvgLikes).

Итоговый коэффициент детерминации R-квадрат составил 0.356 (Adj. R^2 = 0.301), что говорит о том, что примерно 35 % вариации ln_revenue объясняется выбранными факторами. При этом F-статистика оказалась высоко значимой ($F \approx 6.48$, $p \approx 4.4 \times 10^{-6}$), что подтверждает общую пригодность модели. Среди отдельных регрессоров в этой OLS-оценке оказались статистически значимыми четыре фактора: рост среднего рейтинга отрицательно влиял на выручку ($\beta \approx -1.552$, $p<0.001$), положительная тональность отзывов увеличивала выручку ($\beta \approx 20.71$, $p \approx 0.006$), увеличение числа отзывов (ln_review_vol) связано с более высокой выручкой ($\beta \approx 0.391$, $p<0.001$), и рост среднего числа лайков на отзыв положительно сказывался на доходе ($\beta \approx 0.08$, $p \approx 0.031$).

Все остальные регрессоры в этой модели (ln_search_index, cp_index, income_pop) оказались статистически незначимыми. При этом величина Durbin–Watson равная 0.547 однозначно указывает на существенную положительную автокорреляцию остатков, а предупреждение «Covariance Type: nonrobust» говорит о том, что стандартные ошибки рассчитаны при допущении гомоскедастичности. Выяснилось, что классический Pooled OLS даёт «заниженные» стандартные ошибки и недостоверные p-значения, и, кроме того, автокорреляция ошибок препятствует построению корректных доверительных интервалов и гипотезных тестов.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             ln_revenue   R-squared:                       0.356
Model:                            OLS   Adj. R-squared:                  0.301
Method:                 Least Squares   F-statistic:                     6.481
Date:                Tue, 03 Jun 2025   Prob (F-statistic):           4.37e-06
Time:                        10:28:43   Log-Likelihood:                -112.00
No. Observations:                  90   AIC:                             240.0
Df Residuals:                      82   BIC:                             260.0
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          5.6330      3.738      1.507      0.136      -1.804      13.070
ln_search_index  0.0816      0.114      0.715      0.476      -0.145       0.309
avg_rating    -1.5522      0.339     -4.575      0.000      -2.227      -0.877
avg_sentiment 20.7095      7.276      2.846      0.006       6.235      35.184
ln_review_vol  0.3908      0.103      3.784      0.000       0.185       0.596
cp_index       0.0361      0.030      1.212      0.229      -0.023       0.095
income_pop    -0.0428      0.035     -1.217      0.227      -0.113       0.027
AvgLikes       0.0800      0.036      2.199      0.031       0.008       0.152
==============================================================================
Omnibus:                        4.707   Durbin-Watson:                   0.547
Prob(Omnibus):                  0.095   Jarque-Bera (JB):                4.269
Skew:                           0.530   Prob(JB):                        0.118
Kurtosis:                       3.122   Cond. No.                     1.16e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.16e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Durbin-Watson statistic: 0.547
```

Рисунок 16 – Результаты pooled OLS-оценок

Во второй модели мы сохранили те же регрессоры и целевую функцию, но взяли Pooled OLS с «робастными» ошибками. В этой версии коэффициенты β остались неизменными, однако их стандартные ошибки выросли: например, у avg_sentiment std err увеличился с 7.276 до 10.27, а у AvgLikes – с 0.036 до 0.064. Как следствие, некоторые переменные, ранее значимые (в частности, AvgLikes), потеряли уровень значимости (в новой модели $p \approx 0.218$ для AvgLikes), а основные факторы (avg_rating, avg_sentiment, ln_review_vol) по-прежнему остались статистически значимыми на уровне 5 %. Тем не менее Durbin–Watson остался на том же низком уровне (~0.547), что означает, что автокорреляция остатков не была устранена простым пересчетом под гетероскедастичность.

Таким образом, модель существенно повысила надежность оценок стандартных ошибок, однако не избавилась от временной зависимости ошибок. Данный факт указывает на необходимость учесть панельную природу данных – например, через фиксированные эффекты – чтобы одновременно справиться с автокорреляцией и гетероскедастичностью.

```
=== Pooled OLS (robust SE) ===
                            OLS Regression Results
==============================================================================
Dep. Variable:              ln_revenue   R-squared:                       0.356
Model:                             OLS   Adj. R-squared:                  0.301
Method:                  Least Squares   F-statistic:                     11.74
Date:                 Tue, 03 Jun 2025   Prob (F-statistic):           2.93e-10
Time:                         10:29:13   Log-Likelihood:                -112.00
No. Observations:                   90   AIC:                             240.0
Df Residuals:                       82   BIC:                             260.0
Df Model:                            7
Covariance Type:                   HC3
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const              5.6330      4.015      1.403      0.164      -2.354      13.620
ln_search_index    0.0816      0.122      0.669      0.505      -0.161       0.324
avg_rating        -1.5522      0.346     -4.486      0.000      -2.241      -0.864
avg_sentiment     20.7095     10.270      2.017      0.047       0.280      41.139
ln_review_vol      0.3908      0.122      3.194      0.002       0.147       0.634
cp_index           0.0361      0.028      1.270      0.208      -0.020       0.093
income_pop        -0.0428      0.038     -1.115      0.268      -0.119       0.034
AvgLikes           0.0800      0.064      1.243      0.218      -0.048       0.208
==============================================================================
Omnibus:                        4.707   Durbin-Watson:                   0.547
Prob(Omnibus):                  0.095   Jarque-Bera (JB):                4.269
Skew:                           0.530   Prob(JB):                        0.118
Kurtosis:                       3.122   Cond. No.                     1.16e+04
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
[2] The condition number is large, 1.16e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Рисунок 17 – Результаты pooled OLS-оценок с робастными ошибками

4.2. LSDV (least squares dummy variables model, модель с фиктивными переменными)

В рамках данного этапа были добавлены в матрицу регрессоров набор бинарных индикаторов (dummy-переменных) для каждой из сетей (chain_clean). Далее была произведена оценка полученной модели, включающую: константу (intercept), «чистые» регрессоры (логарифм поискового индекса, средний рейтинг, средняя тональность отзывов, логарифм объёма отзывов, индекс цен, реальный располагаемый доход населения и среднее число лайков на отзыв), dummy-переменные для всех сетей за исключением базовой.

В получившейся спецификации коэффициент детерминации R-квадрат составил 0.959, что свидетельствует о том, что практически вся межсетевая неоднородность в логарифме выручки захвачена набором фиктивных индикаторов.

Внедрение dummy-переменных позволило «извлечь» из модели постоянные отличия брендов друг от друга. В частности, коэффициенты при

dummy-столбцах показали, что «Золотое яблоко», «Л'Этуаль», «Магнит Косметик», «Рив Гош» и «Улыбка радуги» статистически значимо отличаются от базовой сети (в нашем случае — «L'Occitane») по уровню среднемесячной выручки. Например, если «L'Occitane» в модели представлена базовым intercept, то «Л'Этуаль» имеет положительную поправку порядка +2.787; это означает, что при прочих равных «Л'Этуаль» в среднем генерирует в $\exp(2.787)\approx16.2$ раза большую выручку, чем «L'Occitane». Аналогично, «Магнит Косметик» и «Улыбка радуги» существенно опережают «L'Occitane», а «Yves Rocher» наоборот — статистически значимо ниже (поправка −0.811).

Что касается «внутрисетевых» (внутрибрендовых) факторов, то в этой скорректированной модели только три переменные оказались значимыми на уровне 5%. Во-первых, логарифм поискового индекса (ln_search_index) продемонстрировал устойчивый положительный эффект ($\beta\approx0.78$, $p<0.001$): когда в рамках одной сети поисковый интерес растёт, её выручка также растёт примерно пропорционально. Во-вторых, реальные располагаемые доходы населения (income_pop) внутри одного бренда оказались значимо положительно связаны с выручкой ($\beta\approx0.034$, $p=0.002$): по мере роста уровня доходов покупателей спрос на продукцию сети увеличивается. В-третьих, среднее число лайков на отзыв (AvgLikes) приобрело отрицательный знак ($\beta\approx-0.031$, $p=0.005$), что может свидетельствовать о том, что в тех ситуациях, когда отзывы получают много лайков (например, при обсуждении негативных инцидентов), выручка сети внутри определённого временного периода снижается. При этом «ранее значимые» без dummy-модели параметры — средний рейтинг (avg_rating), средняя тональность (avg_sentiment), логарифм объёма отзывов (ln_review_vol) и индекс цен (cp_index) — после введения фиксированных эффектов по сетям перестали быть статистически значимыми. Это говорит о том, что их влияние в «сырой» Pooled OLS прежде было обусловлено сочетанием уровневых межсетевых различий и внутрисетевой динамики; но после того как мы выделили эти уровневые сдвиги в dummy-переменные, внутрибрандовый вклад этих переменных оказался малозначимым.

Диагностические тесты подтверждают, что новая модель значительно улучшила распределение остатков: значение статистики Омнибус ($Omnibus=0.720$, $p=0.698$) и тест Жарка–Бера ($JB=0.674$, $p=0.714$) уже не позволяют отвергать гипотезу о нормальности ошибок, тогда как в «чистой» Pooled OLS без dummy распределение было явно ненормальным. Показатель Durbin–Watson вырос до 1.198 (ранее было около 0.547), что означает заметное

снижение автокорреляции, хотя до идеального значения 2.0 пока далеко. Тем не менее общее качество модели стало существенно выше, поскольку «уровневой» (межсетевой) компонентой объясняется очень большая доля дисперсии ln_revenue, а внутрисетевая часть уже адекватно отображает роль тех факторов, которые меняются год от года внутри каждой сети.

Таким образом, введённый вручную LSDV-подход (Pooled OLS + dummy для chain_clean) подтвердил, что между разными сетями существуют устойчивые, неизменные во времени отличия по выручке, и эти отличия лучше моделировать как отдельные постоянные (фиктивные) эффекты. После «вычитания» этих фиксированных разностей оказалось, что наиболее весомые «внутрисетевые» драйверы выручки — это именно уровень поисковой популярности, доходы населения и «лайкоцентричность» отзывов. Все остальные переменные потеряли свою значимость при условии равенства «сетевого базиса».

```
=== Pooled OLS с фиктивными переменными (LSDV по chain_clean) ===
                            OLS Regression Results
==============================================================================
Dep. Variable:              ln_revenue   R-squared:                       0.959
Model:                             OLS   Adj. R-squared:                  0.950
Method:                  Least Squares   F-statistic:                     114.4
Date:                 Tue, 03 Jun 2025   Prob (F-statistic):           7.70e-45
Time:                         11:17:15   Log-Likelihood:                 11.531
No. Observations:                   90   AIC:                             8.939
Df Residuals:                       74   BIC:                             48.94
Df Model:                           15
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                     -4.1432      1.143     -3.625      0.001      -6.420      -1.866
ln_search_index            0.7812      0.065     12.008      0.000       0.652       0.911
avg_rating                -0.0127      0.133     -0.095      0.924      -0.277       0.252
avg_sentiment             -2.9896      2.531     -1.181      0.241      -8.032       2.053
ln_review_vol             -0.0458      0.038     -1.201      0.234      -0.122       0.030
cp_index                   0.0126      0.008      1.521      0.133      -0.004       0.029
income_pop                 0.0344      0.010      3.292      0.002       0.014       0.055
AvgLikes                  -0.0307      0.011     -2.875      0.005      -0.052      -0.009
d_yves rocher             -0.8117      0.162     -4.998      0.000      -1.135      -0.488
d_золотое яблоко           0.8700      0.151      5.768      0.000       0.569       1.171
d_иль де ботэ              0.0479      0.143      0.335      0.738      -0.237       0.333
d_лэтуаль                  2.7871      0.155     17.928      0.000       2.477       3.097
d_магнит косметик          2.4089      0.139     17.284      0.000       2.131       2.687
d_подружка                -0.0261      0.167     -0.157      0.876      -0.359       0.306
d_рив гош                  0.7090      0.180      3.932      0.000       0.350       1.068
d_улыбка радуги            1.6345      0.127     12.911      0.000       1.382       1.887
==============================================================================
Omnibus:                        0.720   Durbin-Watson:                   1.198
Prob(Omnibus):                  0.698   Jarque-Bera (JB):                0.674
Skew:                           0.205   Prob(JB):                        0.714
Kurtosis:                       2.892   Cond. No.                     1.51e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.51e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Рисунок 18 – Результаты LSDV с фиктивными переменными

4.3. Fixed effects model (модель с фиксированными эффектами)

Далее было выполнено построение серии моделей фиксированных эффектов, чтобы учесть особенности панельных данных по исследуемым сетям и годам. Сначала была рассмотрена модель PanelOLS с «брендовыми» фиксированными эффектами (entity effects), при этом годовые тренды игнорировались. В этой спецификации R-квадрат составил около 0.73, и из всех проверенных факторов оказались статистически значимыми сразу три: логарифм поискового индекса (search_index), реальный доход населения (income_pop) и среднее число откликов (AvgLikes).

```
=== FE (entity only) ===
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:             ln_revenue   R-squared:                        0.7337
Estimator:                   PanelOLS   R-squared (Between):             -0.1037
No. Observations:                  90   R-squared (Within):               0.7337
Date:                Tue, Jun 03 2025   R-squared (Overall):             -0.0910
Time:                        13:01:37   Log-likelihood                    11.531
Cov. Estimator:            Unadjusted
                                        F-statistic:                      29.133
Entities:                           9   P-value                           0.0000
Avg Obs:                      10.0000   Distribution:                    F(7,74)
Min Obs:                      10.0000
Max Obs:                      10.0000   F-statistic (robust):             29.133
                                        P-value                           0.0000
Time periods:                      10   Distribution:                    F(7,74)
Avg Obs:                       9.0000
Min Obs:                       9.0000
Max Obs:                       9.0000

                             Parameter Estimates
================================================================================
                 Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
ln_search_index     0.7812     0.0651     12.008     0.0000      0.6515      0.9108
avg_rating         -0.0127     0.1328    -0.0955     0.9242     -0.2773      0.2520
avg_sentiment      -2.9896     2.5305    -1.1814     0.2412     -8.0318      2.0526
ln_review_vol      -0.0458     0.0381    -1.2009     0.2336     -0.1217      0.0302
cp_index            0.0126     0.0083     1.5211     0.1325     -0.0039      0.0290
income_pop          0.0344     0.0104     3.2918     0.0015      0.0136      0.0552
AvgLikes           -0.0307     0.0107    -2.8746     0.0053     -0.0520     -0.0094
================================================================================

F-test for Poolability: 134.75
P-value: 0.0000
Distribution: F(8,74)

Included effects: Entity
```

Рисунок 19 – Модель с фиксированными эффектами только по брендам

Далее мы усложнили модель, добавив годовые эффекты, при возникла необходимость убрать коллинеарные переменные cp_index и income_pop, которые «поглощались» годовыми эффектами. В этой спецификации R^2 вырос до 0,75, а within-R^2 упал до 0.56, но при этом из всех оставшихся регрессоров значимой оказалась только поисковая популярность. Это показало, что уровень сети в конкретном году (после вычитания средних по бренду и по году) фактически зависит лишь от того, насколько часто пользователи искали этот бренд, а остальные переменные перестали давать статистически значимый вклад.

```
=== FE (entity+time) без cp_index & income_pop ===
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:              ln_revenue   R-squared:                        0.7519
Estimator:                    PanelOLS   R-squared (Between):              0.6516
No. Observations:                  90   R-squared (Within):               0.5581
Date:                Tue, Jun 03 2025   R-squared (Overall):              0.6502
Time:                        13:02:11   Log-likelihood                   25.721
Cov. Estimator:            Unadjusted
                                        F-statistic:                      40.603
Entities:                           9   P-value                           0.0000
Avg Obs:                      10.0000   Distribution:                    F(5,67)
Min Obs:                      10.0000
Max Obs:                      10.0000   F-statistic (robust):             40.603
                                        P-value                           0.0000
Time periods:                      10   Distribution:                    F(5,67)
Avg Obs:                       9.0000
Min Obs:                       9.0000
Max Obs:                       9.0000

                             Parameter Estimates
================================================================================
                 Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
ln_search_index     0.8323     0.0595     13.996     0.0000      0.7136      0.9510
avg_rating         -0.0934     0.1210    -0.7718     0.4429     -0.3350      0.1482
avg_sentiment       1.1108     2.5381     0.4377     0.6630     -3.9552      6.1769
ln_review_vol      -0.0745     0.0463    -1.6092     0.1123     -0.1668      0.0179
AvgLikes           -0.0167     0.0102    -1.6491     0.1038     -0.0370      0.0035
================================================================================

F-test for Poolability: 82.386
P-value: 0.0000
Distribution: F(17,67)

Included effects: Entity, Time
```

Рисунок 20 – Модель с фиксированными эффектами по брендам и годам

Затем мы пересчитали стандартные ошибки для модели с эффектами по брендам и годам, кластеризовав по брендам. В этом случае «search_index» всё так же остаётся главным детерминантом, но дополнительно «ожил» и фактор AvgLikes: видимо, если учитывать, что у каждой сети есть собственные постоянные шоки, которые влияют одновременно на несколько лет, замеченный эффект снова становится значимым.

```
=== FE (entity+time) + clustered SE по chain_clean ===
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:              ln_revenue   R-squared:                        0.7519
Estimator:                    PanelOLS   R-squared (Between):              0.6516
No. Observations:                   90   R-squared (Within):               0.5581
Date:                Tue, Jun 03 2025   R-squared (Overall):              0.6502
Time:                        13:02:36   Log-likelihood                   25.721
Cov. Estimator:             Clustered
                                        F-statistic:                      40.603
Entities:                           9   P-value                           0.0000
Avg Obs:                      10.0000   Distribution:                    F(5,67)
Min Obs:                      10.0000
Max Obs:                      10.0000   F-statistic (robust):             54.646
                                        P-value                           0.0000
Time periods:                      10   Distribution:                    F(5,67)
Avg Obs:                      9.0000
Min Obs:                      9.0000
Max Obs:                      9.0000

                             Parameter Estimates
================================================================================
                 Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
ln_search_index     0.8323     0.0742     11.221     0.0000      0.6843      0.9804
avg_rating         -0.0934     0.1460    -0.6399     0.5244     -0.3848      0.1980
avg_sentiment       1.1108     1.4395     0.7716     0.4430     -1.7625      3.9842
ln_review_vol      -0.0745     0.1000    -0.7447     0.4591     -0.2740      0.1251
AvgLikes           -0.0167     0.0072    -2.3135     0.0238     -0.0312     -0.0023
================================================================================

F-test for Poolability: 82.386
P-value: 0.0000
Distribution: F(17,67)

Included effects: Entity, Time
```

Рисунок 20 – Модель FE (entity+time) и кластеризацией по брендам

Наконец, была совершена кластеризация SE по годам и обнаружено, что AvgLikes вновь теряет свою статистическую значимость, а «search_index» остаётся единственным надежным фактором. Иными словами, если полагать, что в один и тот же год у всех сетей могут быть общие шоки (экономические, сезонные, маркетинговые), влияние средней отзывной активности уровня отдельной сети нивелируется, и обращающуюся в выручку силу сохраняет лишь поисковая популярность.

```
=== FE (entity+time) + clustered SE по year ===
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:              ln_revenue   R-squared:                        0.7519
Estimator:                    PanelOLS   R-squared (Between):              0.6516
No. Observations:                   90   R-squared (Within):               0.5581
Date:                 Tue, Jun 03 2025   R-squared (Overall):              0.6502
Time:                         13:02:54   Log-likelihood                    25.721
Cov. Estimator:             Clustered
                                         F-statistic:                      40.603
Entities:                           9    P-value                           0.0000
Avg Obs:                      10.0000    Distribution:                    F(5,67)
Min Obs:                      10.0000
Max Obs:                      10.0000    F-statistic (robust):             50.504
                                         P-value                           0.0000
Time periods:                      10    Distribution:                    F(5,67)
Avg Obs:                       9.0000
Min Obs:                       9.0000
Max Obs:                       9.0000

                             Parameter Estimates
================================================================================
                 Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
ln_search_index     0.8323     0.0792     10.513     0.0000      0.6743      0.9904
avg_rating         -0.0934     0.1218    -0.7667     0.4459     -0.3366      0.1498
avg_sentiment       1.1108     2.8681     0.3873     0.6998     -4.6139      6.8355
ln_review_vol      -0.0745     0.0464    -1.6058     0.1130     -0.1670      0.0181
AvgLikes           -0.0167     0.0125    -1.3375     0.1856     -0.0417      0.0082
================================================================================

F-test for Poolability: 82.386
P-value: 0.0000
Distribution: F(17,67)

Included effects: Entity, Time
```

Рисунок 20 – Модель FE (entity+time) и кластеризацией по годам

4.4. Random effects model (модель со случайными эффектами)

Далее была оценена модель со случайными эффектами (Random Effects). В модели RandomEffects значение R-квадрат (Within) оказалось близким к тому, что мы видели в FE-модели, – то есть примерно около 0.73 (Within), что свидетельствует о том, что большая часть внутрибредовых колебаний ln(revenue) объясняется регрессорами. При этом R-квадрат (Between) получился отрицательным (около –0.35), что указывает на то, что, рассматривая только «межбрендовые» различия, данная линейная комбинация объясняющих переменных хуже среднего.

Константа получилась отрицательной и статистически значимой ($p \approx 0.0244$), логарифм поискового индекса снова оказался высокозначимой переменной ($t=10.53, p<0.0001$) с коэффициентом примерно 0.7407, income_pop

(доходы населения) оказалась значимой (t≈2.7926,p≈0.0065), и её коэффициент примерно 0.03.

Это означает: если реальные доходы населения в регионе бренда повышаются на 1 %, выручка сети (в логарифмах) растёт приблизительно на 0.0322, при прочих равных, AvgLikes (среднее число лайков у отзывов) также показала статистическую значимость (t≈−2.4303,p≈0.0173) с отрицательным знаком −0.0287. Это необычный черезвычайно слабый, но значимый эффект, означающий, что рост среднего числа лайков обратной связи ассоциируется с небольшим снижением выручки. Возможно, это связано с тем, что именно в тех годах/брендах, где лайков много, клиенты больше жалуются на цену или ассортимент.

RandomEffects Estimation Summary

Dep. Variable:	ln_revenue	R-squared:	0.6623
Estimator:	RandomEffects	R-squared (Between):	−0.3483
No. Observations:	90	R-squared (Within):	0.7318
Date:	Tue, Jun 03 2025	R-squared (Overall):	−0.1805
Time:	13:43:33	Log-likelihood	−2.1543
Cov. Estimator:	Unadjusted		
		F-statistic:	22.977
Entities:	9	P-value	0.0000
Avg Obs:	10.0000	Distribution:	F(7,82)
Min Obs:	10.0000		
Max Obs:	10.0000	F-statistic (robust):	22.977
		P-value	0.0000
Time periods:	10	Distribution:	F(7,82)
Avg Obs:	9.0000		
Min Obs:	9.0000		
Max Obs:	9.0000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	−2.9074	1.2675	−2.2938	0.0244	−5.4287	−0.3860
ln_search_index	0.7407	0.0703	10.534	0.0000	0.6008	0.8806
avg_rating	−0.0617	0.1455	−0.4241	0.6726	−0.3511	0.2277
avg_sentiment	−2.2834	2.7818	−0.8208	0.4141	−7.8173	3.2505
ln_review_vol	−0.0322	0.0419	−0.7704	0.4433	−0.1155	0.0510
cp_index	0.0127	0.0091	1.3907	0.1681	−0.0055	0.0309
income_pop	0.0322	0.0115	2.7926	0.0065	0.0093	0.0551
AvgLikes	−0.0287	0.0118	−2.4303	0.0173	−0.0521	−0.0052

Рисунок 20 – Модель со случайными эффектами

4.5. Сравнение моделей и тест Хаусмана

Для того чтобы выяснить, какая модель в нашем случае даёт более надёжные оценки параметров, сосредоточимся на двух подходах – модели с

фиксированными эффектами (Fixed Effects) и модели со случайными эффектами (Random Effects).

Model Comparison

	FE (entity only)	FE (entity+time)	FE (entity+time clustered entity)	FE (entity+time clustered time)	Random Effects
Dep. Variable	ln_revenue	ln_revenue	ln_revenue	ln_revenue	ln_revenue
Estimator	PanelOLS	PanelOLS	PanelOLS	PanelOLS	RandomEffects
No. Observations	90	90	90	90	90
Cov. Est.	Unadjusted	Unadjusted	Clustered	Clustered	Unadjusted
R-squared	0.7337	0.7519	0.7519	0.7519	0.6623
R-Squared (Within)	0.7337	0.5581	0.5581	0.5581	0.7318
R-Squared (Between)	-0.1037	0.6516	0.6516	0.6516	-0.3483
R-Squared (Overall)	-0.0910	0.6502	0.6502	0.6502	-0.1805
F-statistic	29.133	40.603	40.603	40.603	22.977
P-value (F-stat)	0.0000	0.0000	0.0000	0.0000	0.0000
ln_search_index	0.7812***	0.8323***	0.8323***	0.8323***	0.7407***
	(0.0651)	(0.0595)	(0.0742)	(0.0792)	(0.0703)
avg_rating	-0.0127	-0.0934	-0.0934	-0.0934	-0.0617
	(0.1328)	(0.1210)	(0.1460)	(0.1218)	(0.1455)
avg_sentiment	-2.9896	1.1108	1.1108	1.1108	-2.2834
	(2.5305)	(2.5381)	(1.4395)	(2.8681)	(2.7818)
ln_review_vol	-0.0458	-0.0745	-0.0745	-0.0745	-0.0322
	(0.0381)	(0.0463)	(0.1000)	(0.0464)	(0.0419)
cp_index	0.0126				0.0127
	(0.0083)				(0.0091)
income_pop	0.0344***				0.0322***
	(0.0104)				(0.0115)
AvgLikes	-0.0307***	-0.0167	-0.0167**	-0.0167	-0.0287**
	(0.0107)	(0.0102)	(0.0072)	(0.0125)	(0.0118)
const					-2.9074**
					(1.2675)
Effects	Entity	Entity Time	Entity Time	Entity Time	

Std. Errors reported in parentheses

Проведем тест Хаусмана. Из всех вариантов FE-моделей возьмем в расчёт ту, которая включала «фиксированный бренд» (entity_effects), но без кластерных поправок и временных эффектов (эта модель сохраняет все базовые гипотезы FE без лишних надстроек, и её результаты легко интерпретировать). В качестве RE-модели мы использовали стандартную панельную модель со случайными эффектами. Именно эти два результата – result_FE_i и result_RE и были подставлены в процедуру Хаусмана.

Сама статистика теста Хаусмана рассчитывается как квадратичная форма разности вектора оценок FE и RE: если случайные эффекты действительно корректны (то есть RE-оценки являются состоятельными), то разница между должна быть «случайной» (шумовой), и тестовая статистика будет близка к нулю; а p-value будет велико. В нашем случае статистика оказалась отрицательной (–3.062), что технически говорит о том, что матрица получилась не строго положительно-определённой, а p-value составило 1.000 (то есть фактически максимальное возможное значение). Это означает, что при любом разумном уровне значимости (0.05, 0.01 или даже 0.001) мы не отвергаем нулевую гипотезу «о состоятельности RE-оценок и отсутсвии систематического смещения по сравнению с FE» - нет статистических оснований полагать, что модели RE даёт смещённые результаты, а FE – нет.

Иными словами, в исследуемом наборе панельных данных нет статистических доказательств того, что эффекты бренда, что фиксировались FE-моделью неизменно связаны с независимыми переменными. Следовательно, RE-модель (которая предполагает, что эти «брендовые» эффекты случайны и

некоррелированы с регрессорами) работает корректно, соответственно, ориентируясь на результат теста Хаусмана, можно сделать выбор в пользу Random Effects как окончательной спецификации.

5. Ключевые выводы по работе с данными

1. Формируется «двухскоростной» рынок. Наиболее сильными игроками на рынке можно выделить три сети («Л’Этуаль», «Золотое яблоко», "Магнит косметик"), две из них («Золотое яблоко» и "Магнит косметик") за рассматриваемых период (2015-2024) продемострировали высокий рост выручки с ≈ 30 млрд ₽ до > 150 млрд ₽, тогда как большая часть рассматриваемых сетей остались ≤ 50 млрд ₽. Высокая концентрация усиливает барьер входа и диктует разный потенциал роста для сетей-аутсайдеров и лидеров.

2.Средний рейтинг сети дошёл до потолка в 4,3–4,4 звезды, но доля негатива остаётся дифференцированной. Даже при схожем среднеарифметическом значении у некоторых сетей каждое пятое-шестое сообщение негативно, тогда как у других негатив держится в пределах 3–5 %. Это указывает, что показатель звёзд часто сглаживает реальные проблемы клиентского опыта, и анализ отзывов по квантилям даёт более честную картину.

3. Средняя тональность текстов росла заметно быстрее, чем рейтинг. С 2015 по 2024 г. она прибавила 25–40 % почти у всех сетей, что сигнализирует о смещении коммуникаций в сторону «дружелюбного тона». Разрыв между рейтинговыми «четвёрками» и «тройками» по эмоциональной окраске сузился, поэтому маркировка по sentiment-score помогает ловить ранние паттерны недовольства до того, как упадёт средняя оценка.

4.Макроэкономика выступала фоном, но не главной движущей силой. Индекс потребительских цен колебался в узком коридоре — 101–113 % к базовому году, а реальные доходы падали до 2021 г., после чего отыграли 9,6 % за два года. Эти сдвиги отражаются в коэффициентах модели, однако объясняют лишь пятую часть межгодовой дисперсии выручки, тогда как digital-метрики задают основную динамику.

5. Панельный анализ подтвердил преимущество модели случайных эффектов. Hausman-тест не выявил систематических расхождений FE и RE ($p \approx 1$), а RE-оценки объясняют 73 % внутрисетевых колебаний выручки. Главный драйвер продаж — узнаваемость бренда в поиске. На втором месте — покупательская способность: рост реальных доходов населения на 1 п.п. даёт +3,2 % оборота. А вот сами оценки и «позитивность» текстов значения не имеют, а резонансные (много лайков) отзывы даже коррелируют с просадкой продаж — вероятно, потому что чаще обсуждают проблемы.

6. Самым сильным драйвером выручки выступает поисковый интерес к бренду. По модели случайных эффектов 1-процентный рост относительного индекса Google Trends связан в среднем с увеличением выручки на ≈ 0,74 %. Эффект более чем в 20 раз превосходит влияние реальных доходов населения ($\beta \approx 0{,}03$) и оставляет статистически незначимыми «классические» репутационные метрики — средний рейтинг, тональность и объём отзывов. Практически это означает, что любые активности, повышающие поисковый интерес (SEO, инфлюенс-кампании, контент-маркетинг), дают наибольшую отдачу: +10 % к поисковому интересу способно добавить ≈ 7–8 % оборота. Напротив, простое «накручивание» рейтингов без роста видимости почти не влияет на продажи. Поэтому бюджет продвижения целесообразно распределять в первую очередь на проекты, расширяющие цифровую узнаваемость бренда, а улучшение отзывов и сервисные инициативы использовать как поддерживающий, а не ключевой, рычаг роста.