Факторы богатства и успеха

Проект подготовили будущие топы всех похожих списков:
Стекольщикова Яна
Губенко Олеся
Журавлев Артемий

Датасет “Billionaires Statistics Dataset” (35 признаков, 2539 строк):

rank: Ранг миллиардера по уровню богатства (место в рейтинге).

finalWorth: Итоговое состояние миллиардера в долларах США (чистый капитал).

category: Категория или отрасль, в которой работает бизнес миллиардера.

personName: Имя миллиардера.

age: Возраст миллиардера.

country: Страна проживания миллиардера.

city: Город проживания миллиардера.

source: Источник богатства миллиардера.

industries: Отрасли, связанные с бизнес-интересами миллиардера.

countryOfCitizenship: Страна гражданства миллиардера.

organization: Название организации/компании, связанной с миллиардером.

selfMade: Указывает, является ли миллиардер self-made.

gender: Пол миллиардера.

birthDate: Дата рождения миллиардера.

lastName: Фамилия миллиардера.

firstName: Имя миллиардера.

title: Должность в компании.

date: Дата сбора данных.

state: Штат проживания (для США).

residenceStateRegion: Регион/штат проживания.

birthYear: Год рождения.

birthMonth: Месяц рождения.

birthDay: День рождения.

cpi_country: Индекс потребительских цен (ИПЦ) в стране миллиардера.

cpi_change_country: Изменение ИПЦ в стране миллиардера.

gdp_country: ВВП страны миллиардера.

gross_tertiary_education_enrollment: Доля населения с высшим образованием в стране.

gross_primary_education_enrollment_country: Доля населения с начальным образованием в стране.

life_expectancy_country: Ожидаемая продолжительность жизни в стране.

tax_revenue_country_country: Налоговые поступления в стране.

total_tax_rate_country: Общая налоговая ставка в стране.

population_country: Население страны.

latitude_country: Географическая широта страны.

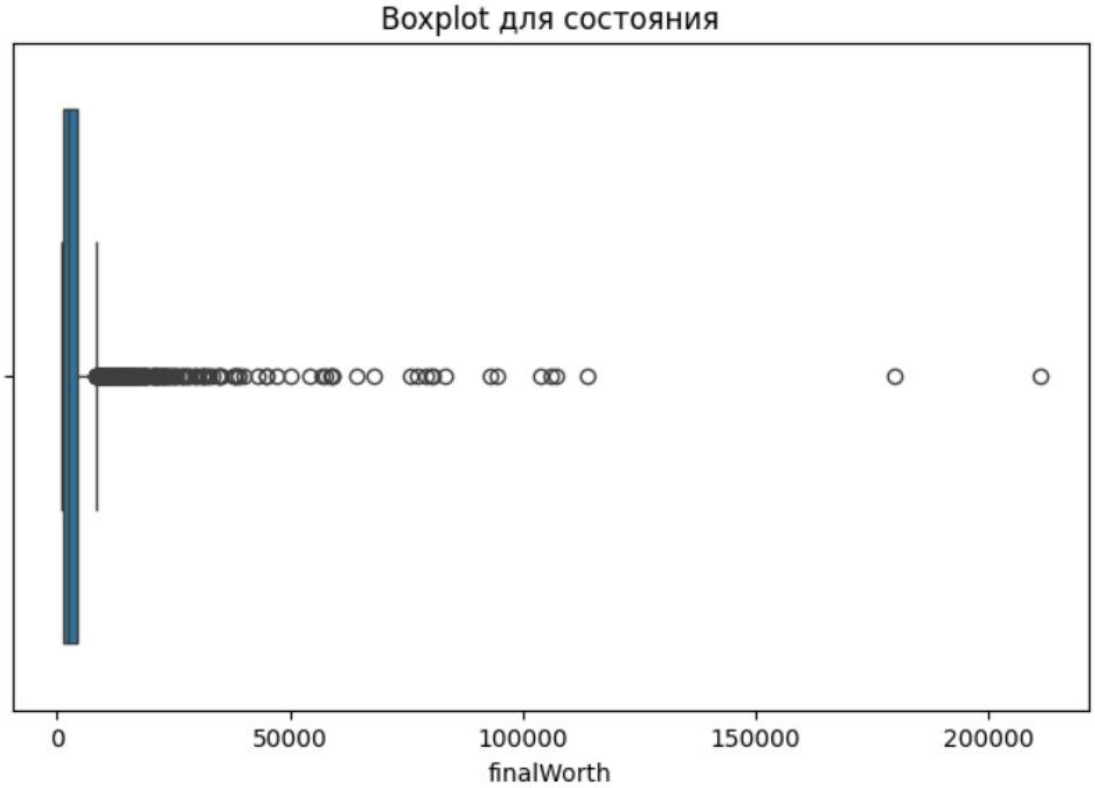
longitude_country: Географическая долгота страны.

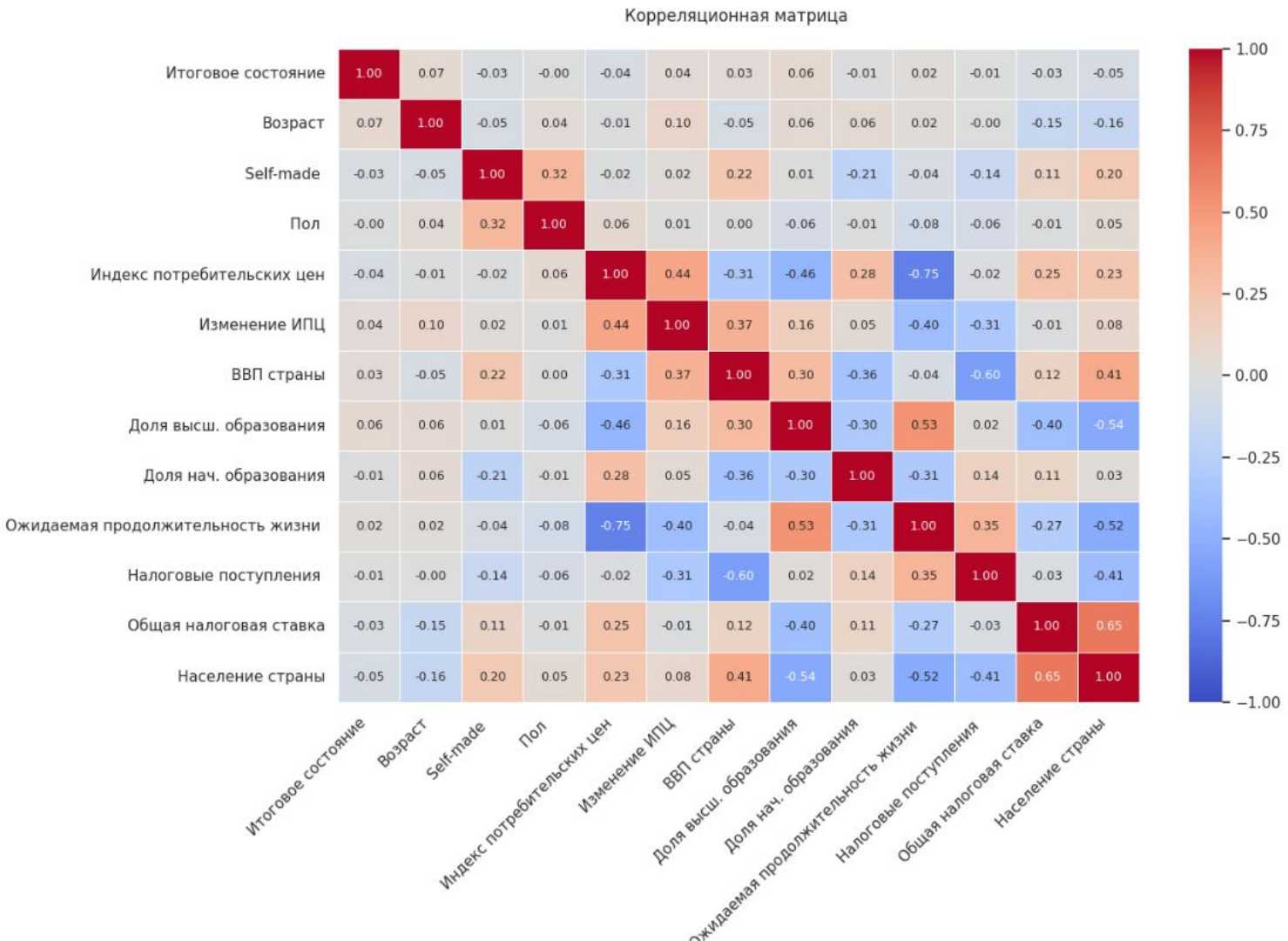
Загрузка и предварительный анализ(EDA)

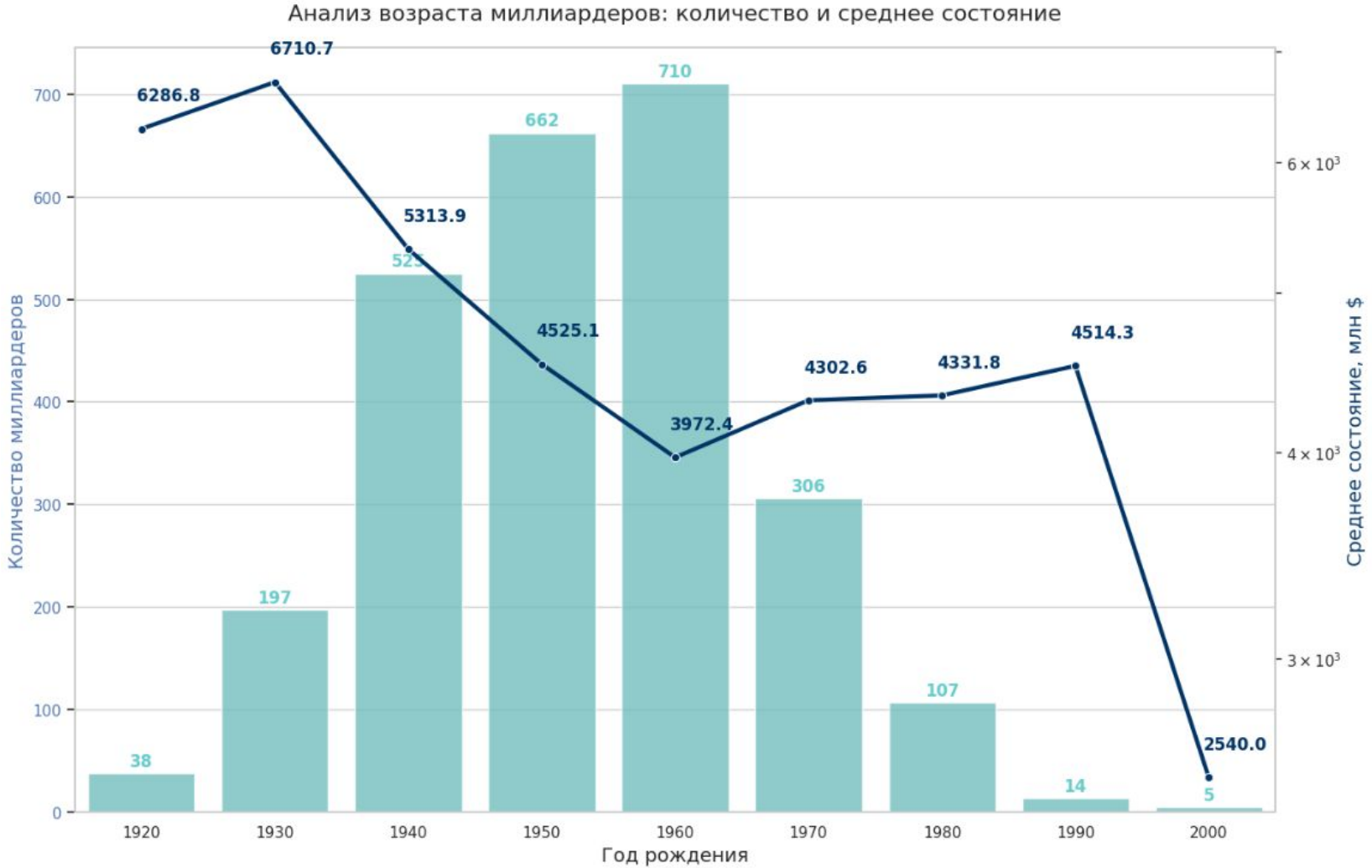
	Пропущенные значения	% от общего
rank	0	0.000000
finalWorth	0	0.000000
category	0	0.000000
personName	0	0.000000
age	65	2.462121
country	38	1.439394
city	72	2.727273
source	0	0.000000
industries	0	0.000000
countryOfCitizenship	0	0.000000
organization	2315	87.689394
selfMade	0	0.000000
status	0	0.000000
gender	0	0.000000
birthDate	76	2.878788
lastName	0	0.000000
firstName	3	0.113636
title	2301	87.159091
date	0	0.000000
state	1887	71.477273
residenceStateRegion	1893	71.704545
birthYear	76	2.878788
birthMonth	76	2.878788
birthDay	76	2.878788
cpi_country	184	6.969697
cpi_change_country	184	6.969697
gdp_country	164	6.212121
gross_tertiary_education_enrollment	182	6.893939
gross_primary_education_enrollment_country	181	6.856061
life_expectancy_country	182	6.893939
tax_revenue_country_country	183	6.931818
total_tax_rate_country	182	6.893939
population_country	164	6.212121
latitude_country	164	6.212121
longitude_country	164	6.212121

Исследование распределений ключевых признаков

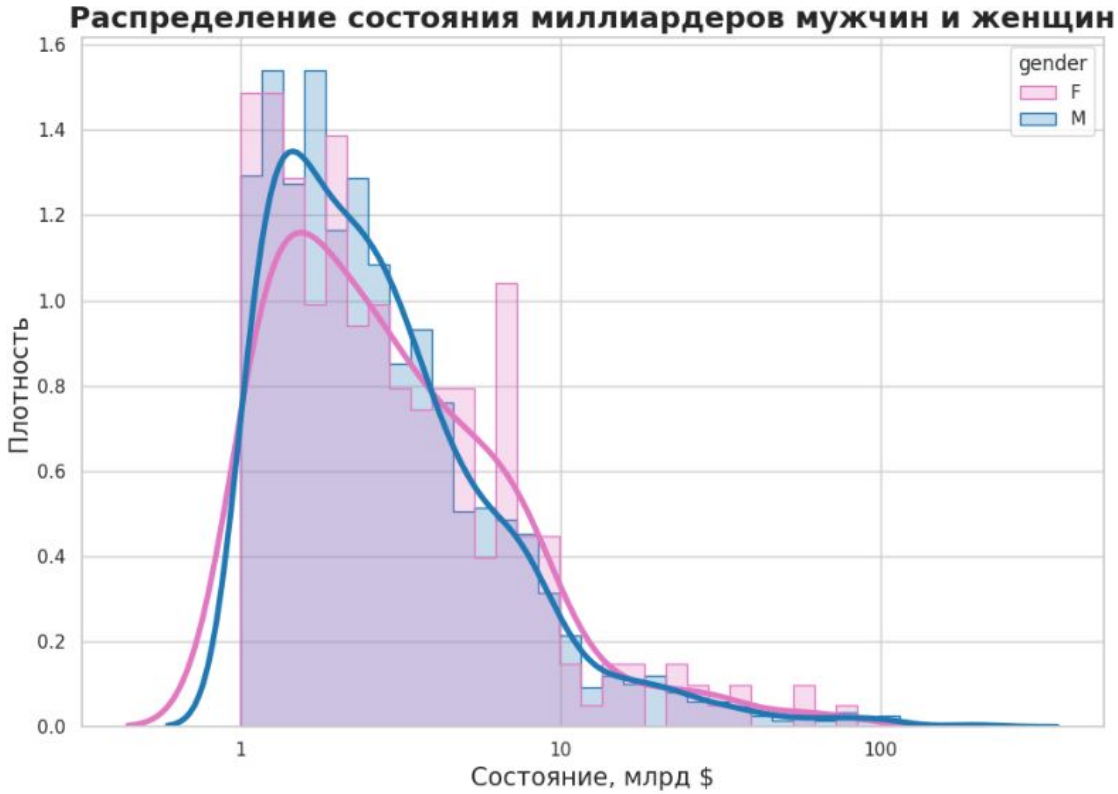
Анализ распределения богатства

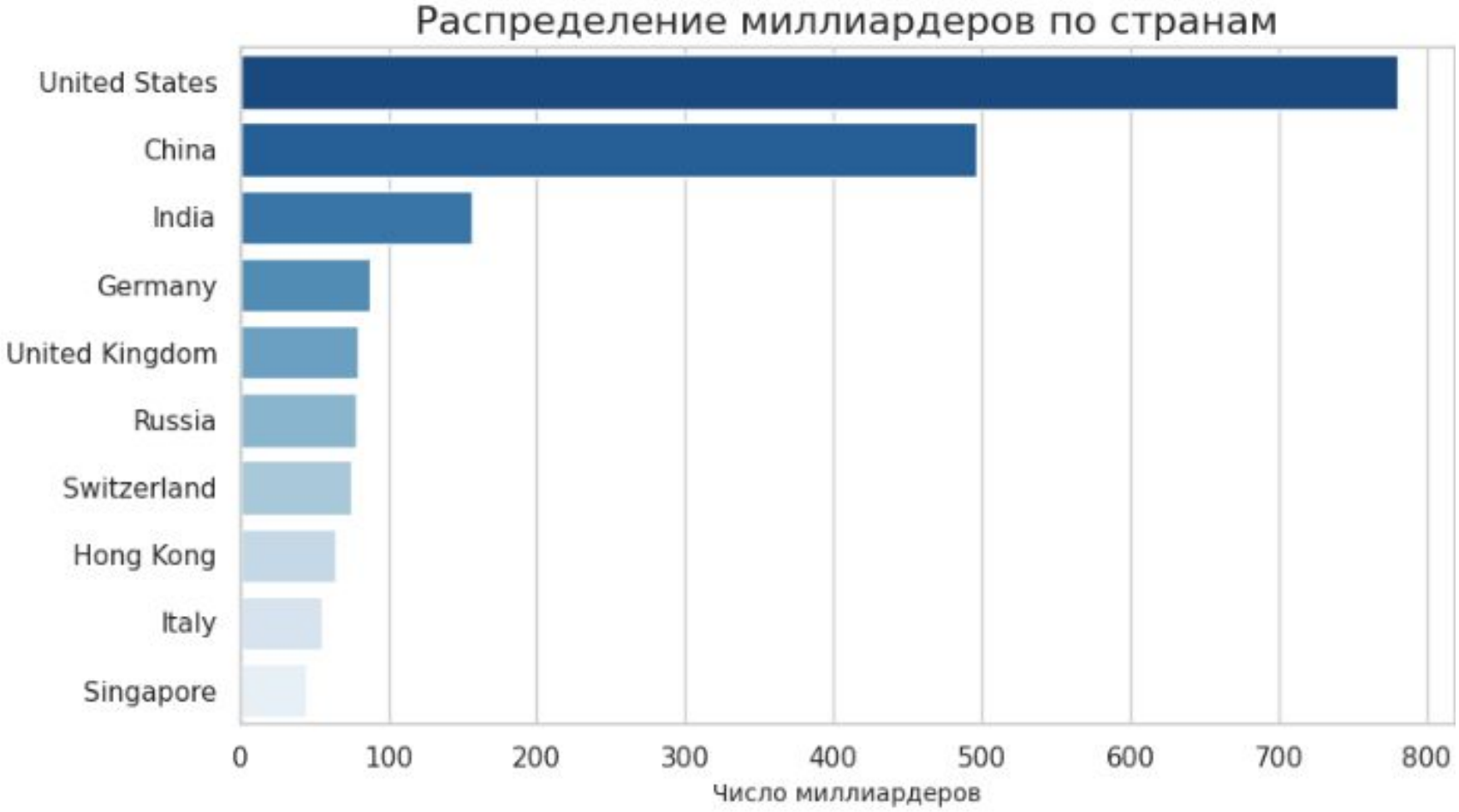


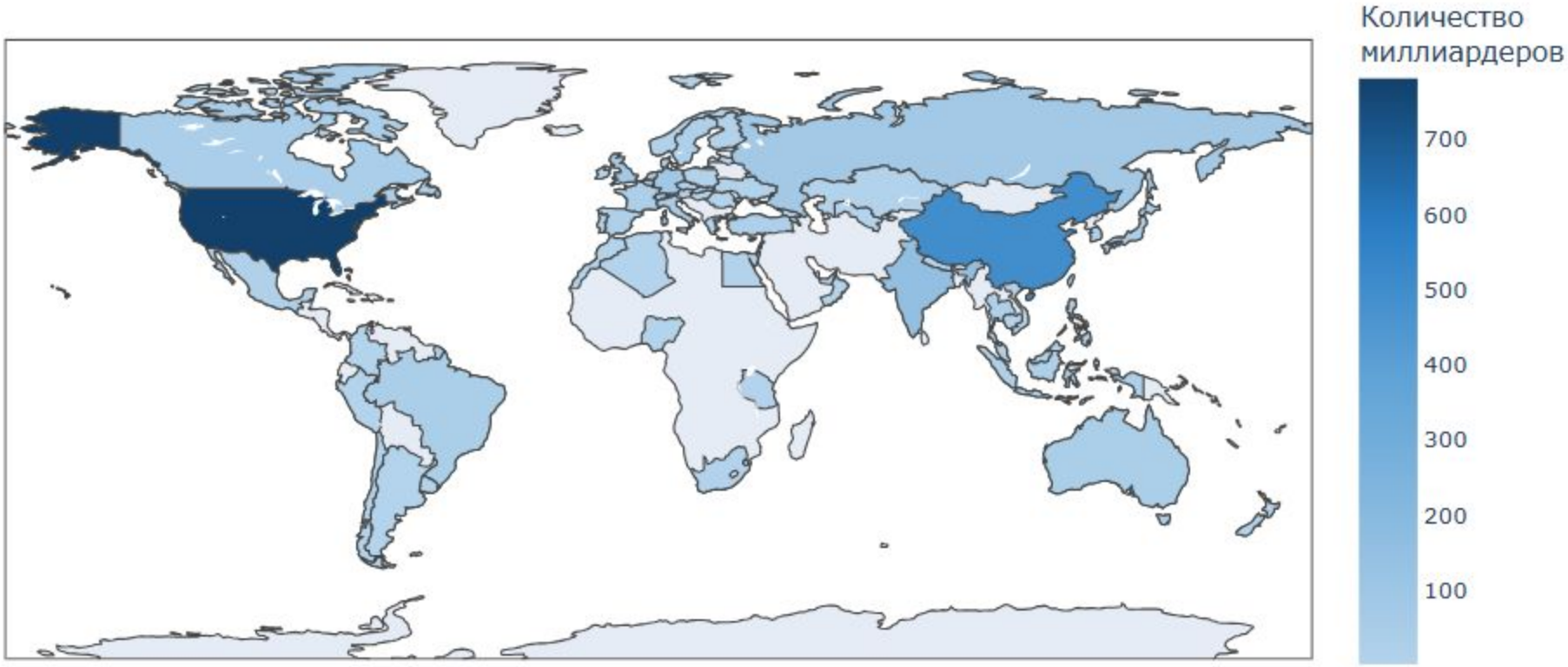

Корреляционная матрица
Итоговое состояние 1.00 0.07 -0.03 -0.00 -0.04 0.04 0.03 0.06 -0.01 0.02 -0.01 -0.03 -0.05
Возраст 0.07 1.00 -0.05 0.04 -0.01 0.10 -0.05 0.06 0.06 0.02 -0.00 -0.15 -0.16
Self-made -0.03 -0.05 1.00 0.32 -0.02 0.02 0.22 0.01 -0.21 -0.04 -0.14 0.11 0.20
Пол -0.00 0.04 0.32 1.00 0.06 0.01 0.00 -0.06 -0.01 -0.08 -0.06 -0.01 0.05
Индекс потребительских цен -0.04 -0.01 -0.02 0.06 1.00 0.44 -0.31 -0.46 0.28 -0.75 -0.02 0.25 0.23
Изменение ИПЦ 0.04 0.10 0.02 0.01 0.44 1.00 0.37 0.16 0.05 -0.40 -0.31 -0.01 0.08
ВВП страны 0.03 -0.05 0.22 0.00 -0.31 0.37 1.00 0.30 -0.36 -0.04 -0.60 0.12 0.41
Доля высш. образования 0.06 0.06 0.01 -0.06 -0.46 0.16 0.30 1.00 -0.30 0.53 0.02 -0.40 -0.54
Доля нач. образования -0.01 0.06 -0.21 -0.01 0.28 0.05 -0.36 -0.30 1.00 -0.31 0.14 0.11 0.03
Ожидаемая продолжительность жизни 0.02 0.02 -0.04 -0.08 -0.75 -0.40 -0.04 0.53 -0.31 1.00 0.35 -0.27 -0.52
Налоговые поступления -0.01 -0.00 -0.14 -0.06 -0.02 -0.31 -0.60 0.02 0.14 0.35 1.00 -0.03 -0.41
Общая налоговая ставка -0.03 -0.15 0.11 -0.01 0.25 -0.01 0.12 -0.40 0.11 -0.27 -0.03 1.00 0.65
Население страны -0.05 -0.16 0.20 0.05 0.23 0.08 0.41 -0.54 0.03 -0.52 -0.41 0.65 1.00
Итоговое состояние
Возраст
Self-made
Пол
Индекс потребительских цен
Изменение ИПЦ
ВВП страны
Доля высш. образования
Доля нач. образования
Ожидаемая продолжительность жизни
Налоговые поступления
Общая налоговая ставка
Население страны
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

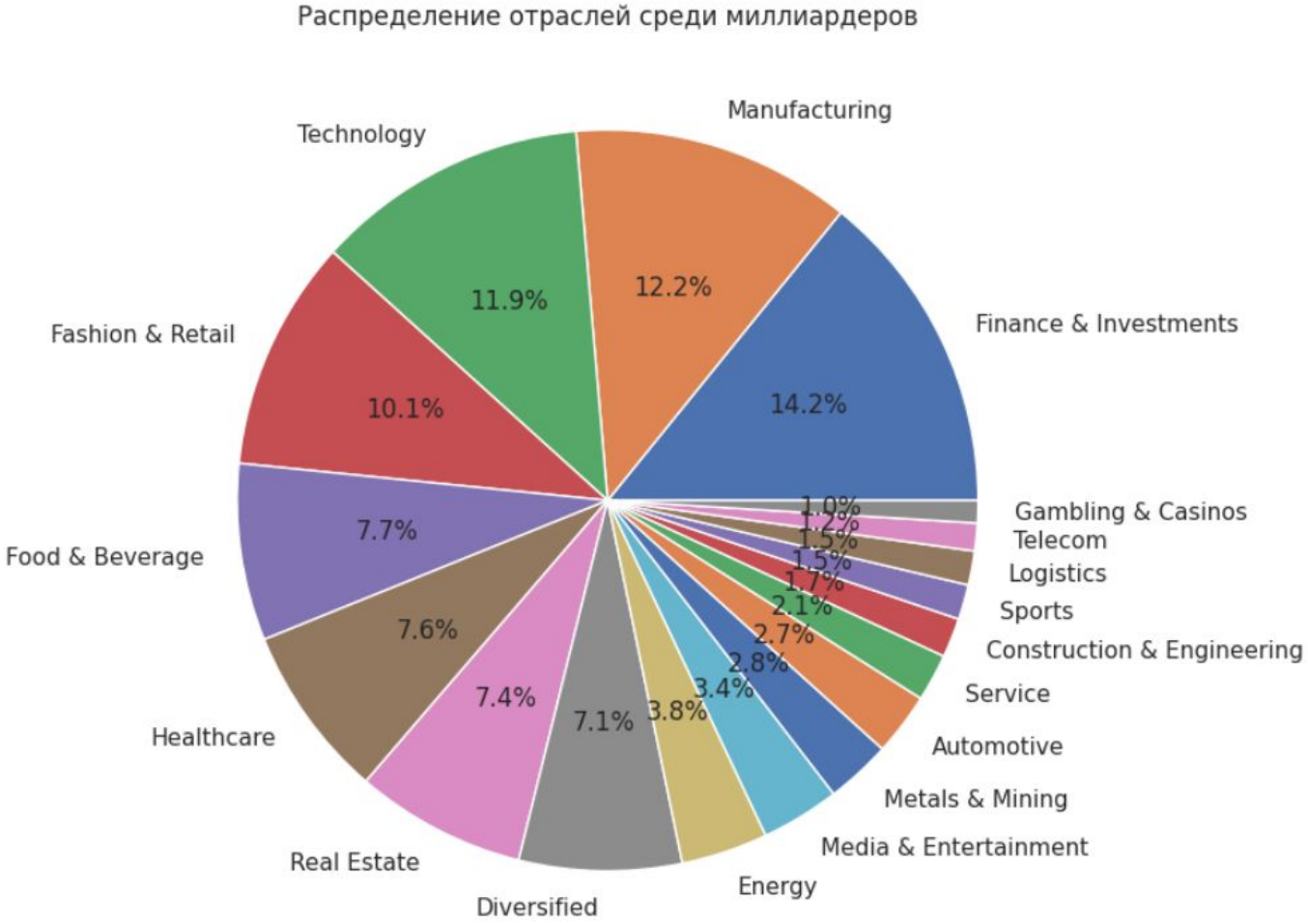

Анализ возраста миллиардеров: количество и среднее состояние
Количество миллиардеров
Среднее состояние, млн $
Год рождения
0
100
200
300
400
500
600
700
3×10^3
4×10^3
6×10^3
1920
1930
1940
1950
1960
1970
1980
1990
2000
38
197
525
662
710
306
107
14
5
6286.8
6710.7
5313.9
4525.1
3972.4
4302.6
4331.8
4514.3
2540.0


Соотношение по полу
Женщины
11.9%
88.1%
Мужчины

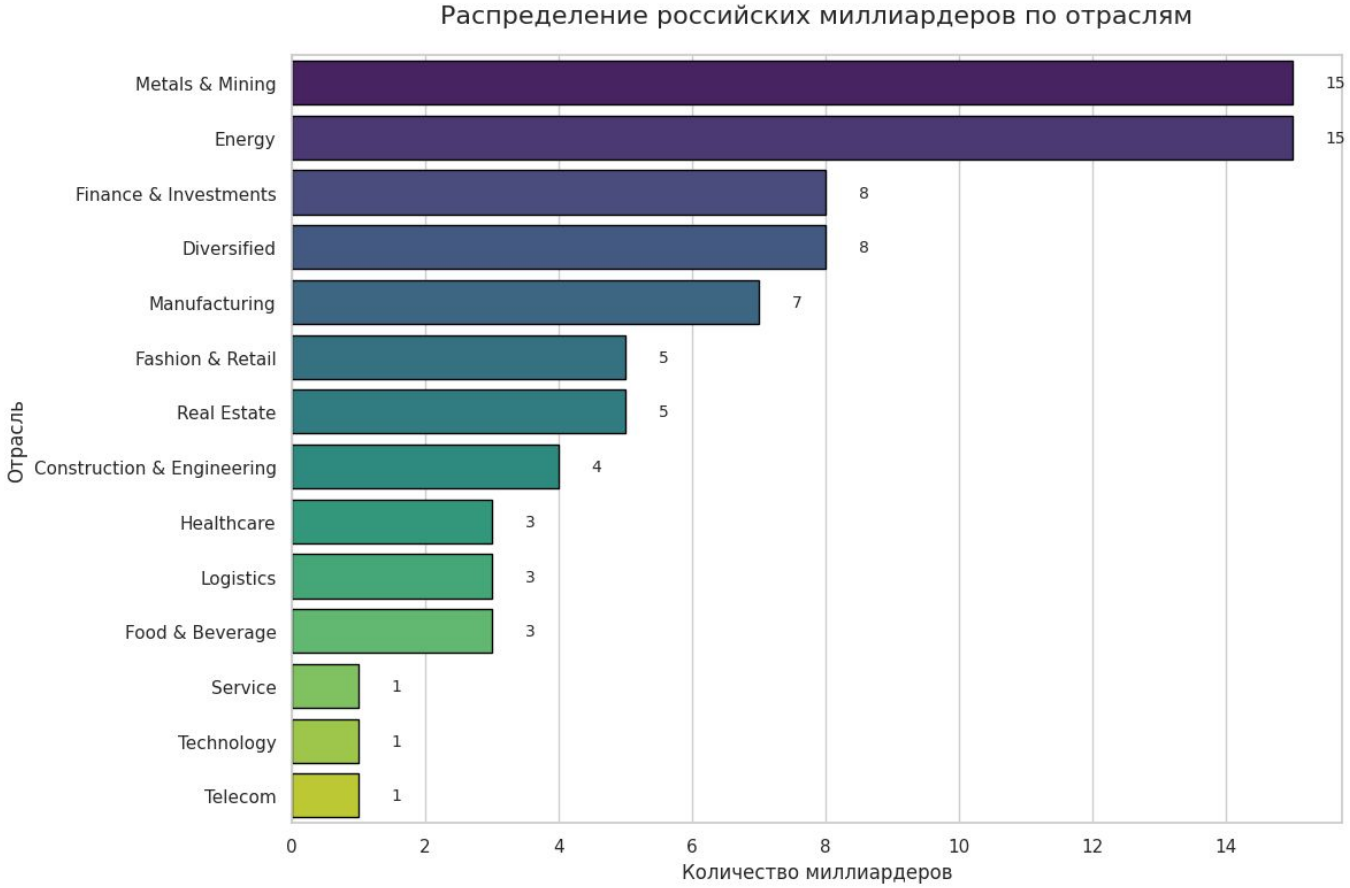

Распределение состояния миллиардеров мужчин и женщин
gender
F
M
Плотность
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1
10
100
Состояние, млрд $


Распределение миллиардеров по странам
United States
China
India
Germany
United Kingdom
Russia
Switzerland
Hong Kong
Italy
Singapore
0
100
200
300
400
500
600
700
800
Число миллиардеров

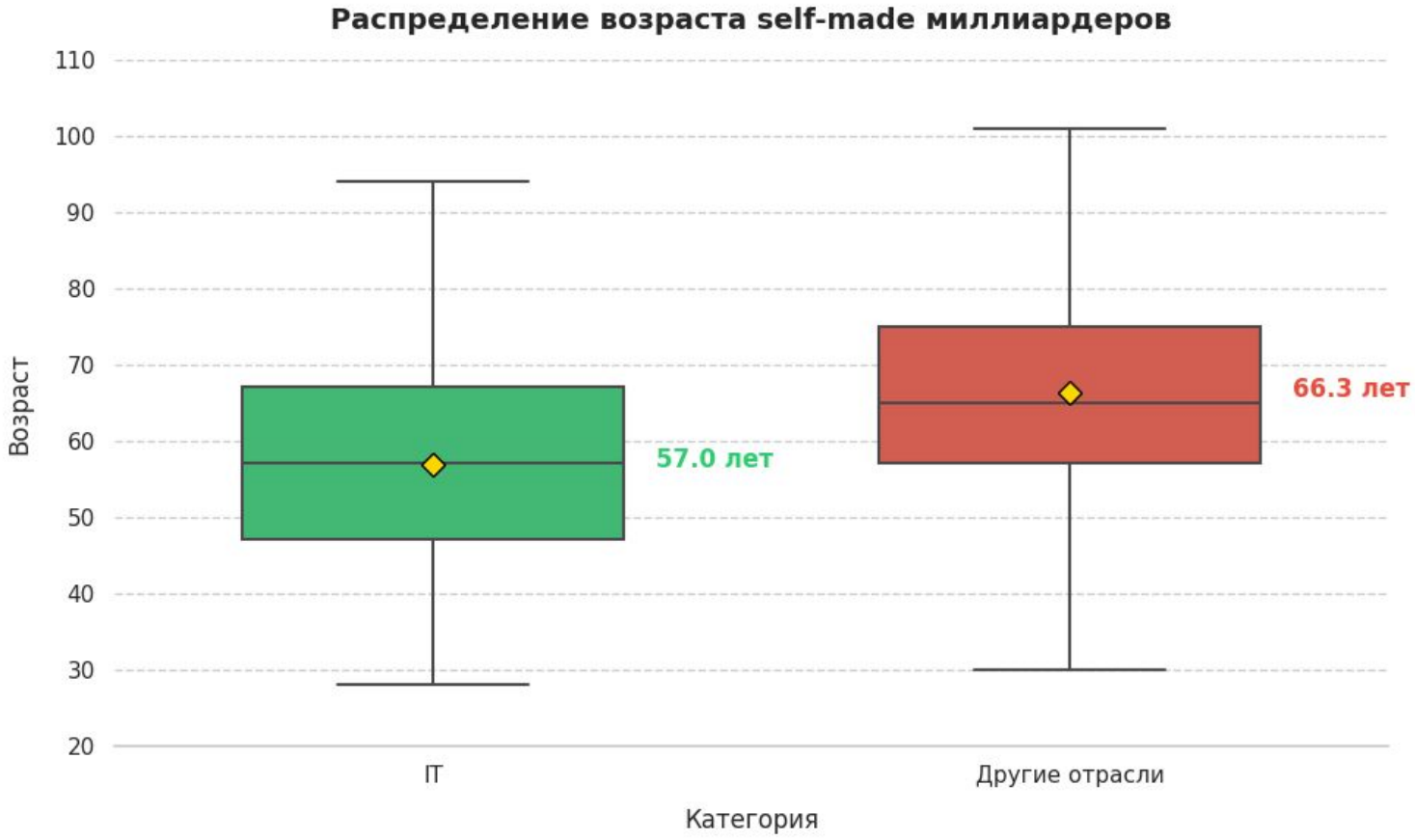

Количество
миллиардеров
700
600
500
400
300
200
100


Распределение отраслей среди миллиардеров
Manufacturing
Technology
Finance & Investments
Fashion & Retail
12.2%
11.9%
14.2%
10.1%
Gambling & Casinos
Telecom
Logistics
Sports
Construction & Engineering
Service
Automotive
Metals & Mining
Media & Entertainment
Energy
Diversified
Real Estate
Healthcare
Food & Beverage
1.0%
1.2%
1.5%
1.5%
1.7%
2.1%
2.7%
2.8%
3.4%
3.8%
7.1%
7.4%
7.6%
7.7%

“Большинство русских сверхбогатых людей заработали состояние в металлургической промышленности”



“Средний возраст self-made миллиардеров в IT ниже, чем средний возраст self-made миллиардеров в традиционных отраслях”



“В среднем состояние IT-миллиардеров больше, чем состояние миллиардеров из других областей”

Для проверки данной гипотезы мы используем метод bootstrat.

Суть метода:

Бутстреп — это компьютерный статистический метод, который позволяет оценить точность измерений и проверить гипотезы путем многократного случайного повторного отбора данных с заменой из исходной выборки.

Как работает:

1. Из исходных данных берутся случайные подвыборки того же размера (с возможностью повторения значений)
2. Для каждой подвыборки вычисляется нужная статистика (среднее, разница средних и др.)
3. Процесс повторяется тысячи раз (обычно 10 000+)
4. По полученному распределению статистик определяют:
    - Точечную оценку параметра
    - Доверительные интервалы
    - Статистическую значимость

M(IT) — среднее состояние IT-миллиардеров в генеральной совокупности,

M(Other) — среднее состояние миллиардеров из других отраслей в генеральной совокупности.

Нулевая гипотеза (H0):

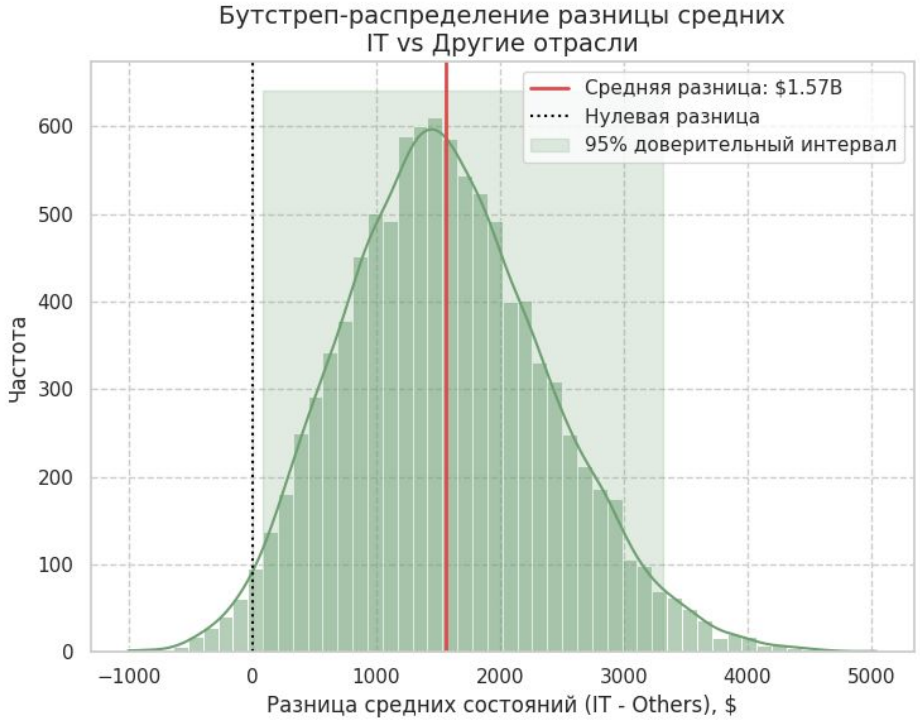
H0: M(IT) <= M(Other)

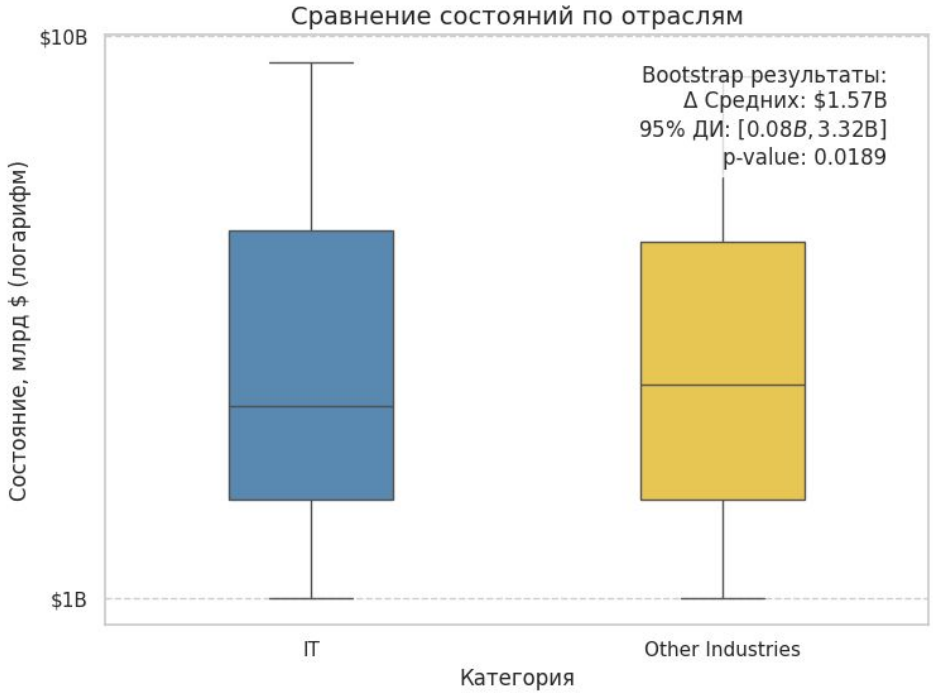
«Среднее состояние IT-миллиардеров не превышает среднее состояние миллиардеров из других отраслей».

Альтернативная гипотеза (H1):

H1: M(IT) > M(Other)

«Среднее состояние IT-миллиардеров больше среднего состояния миллиардеров из других отраслей».



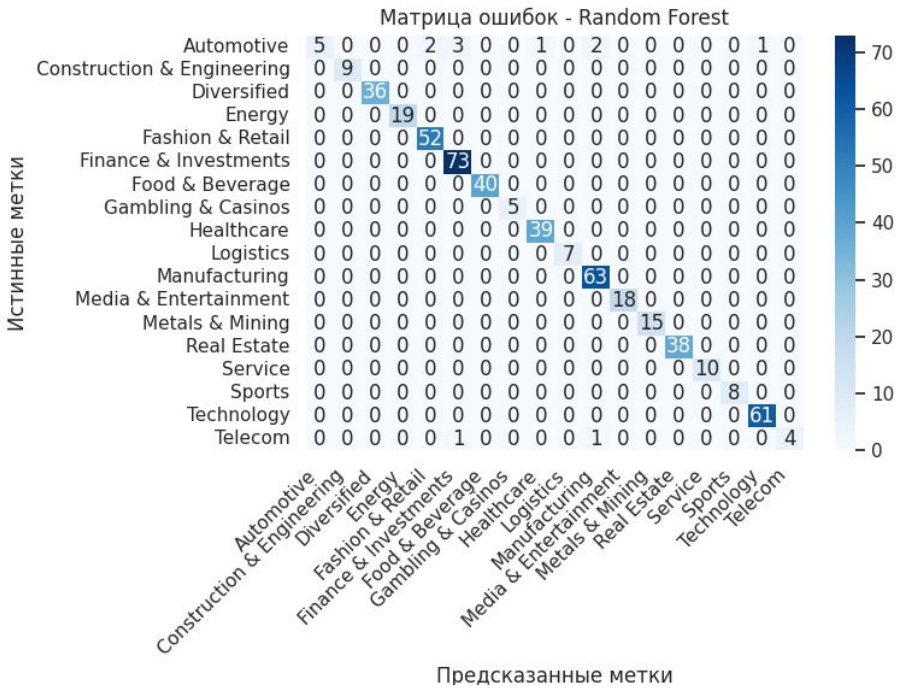


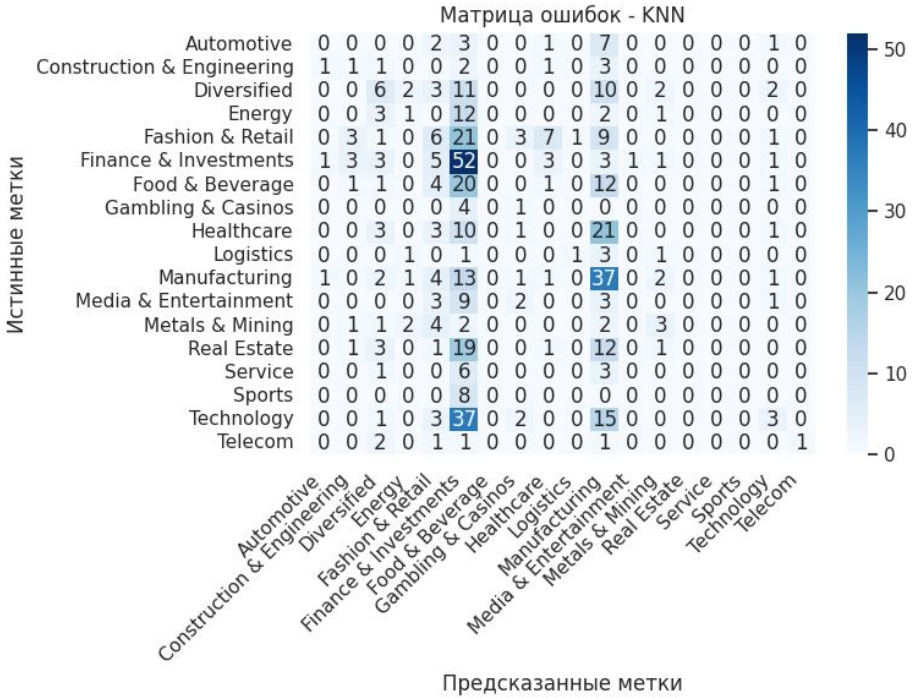
Вывод: Отвергаем H0. IT-миллиардеры богаче ($p < 0.05$).

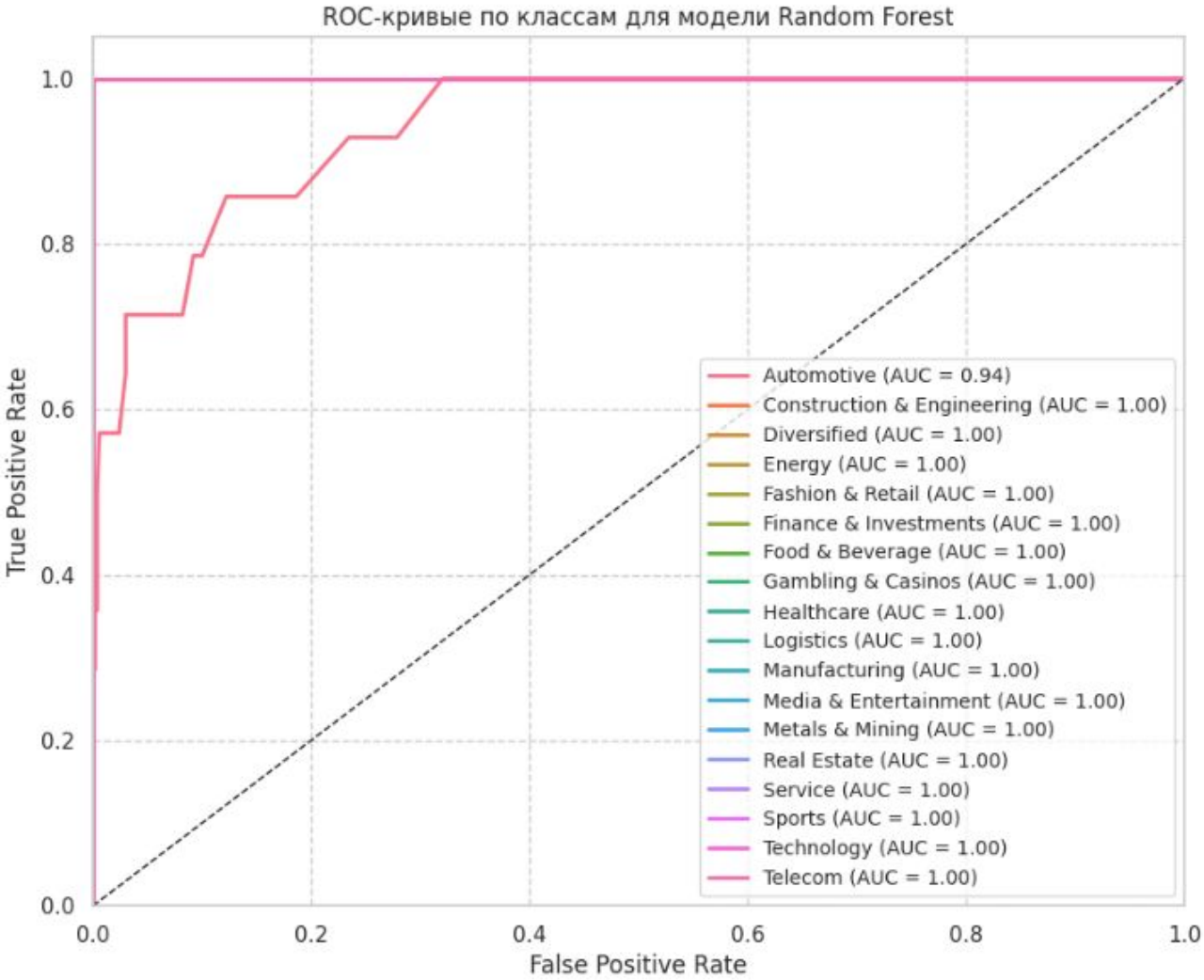
Классификация отрасли - Random Forest + kNN

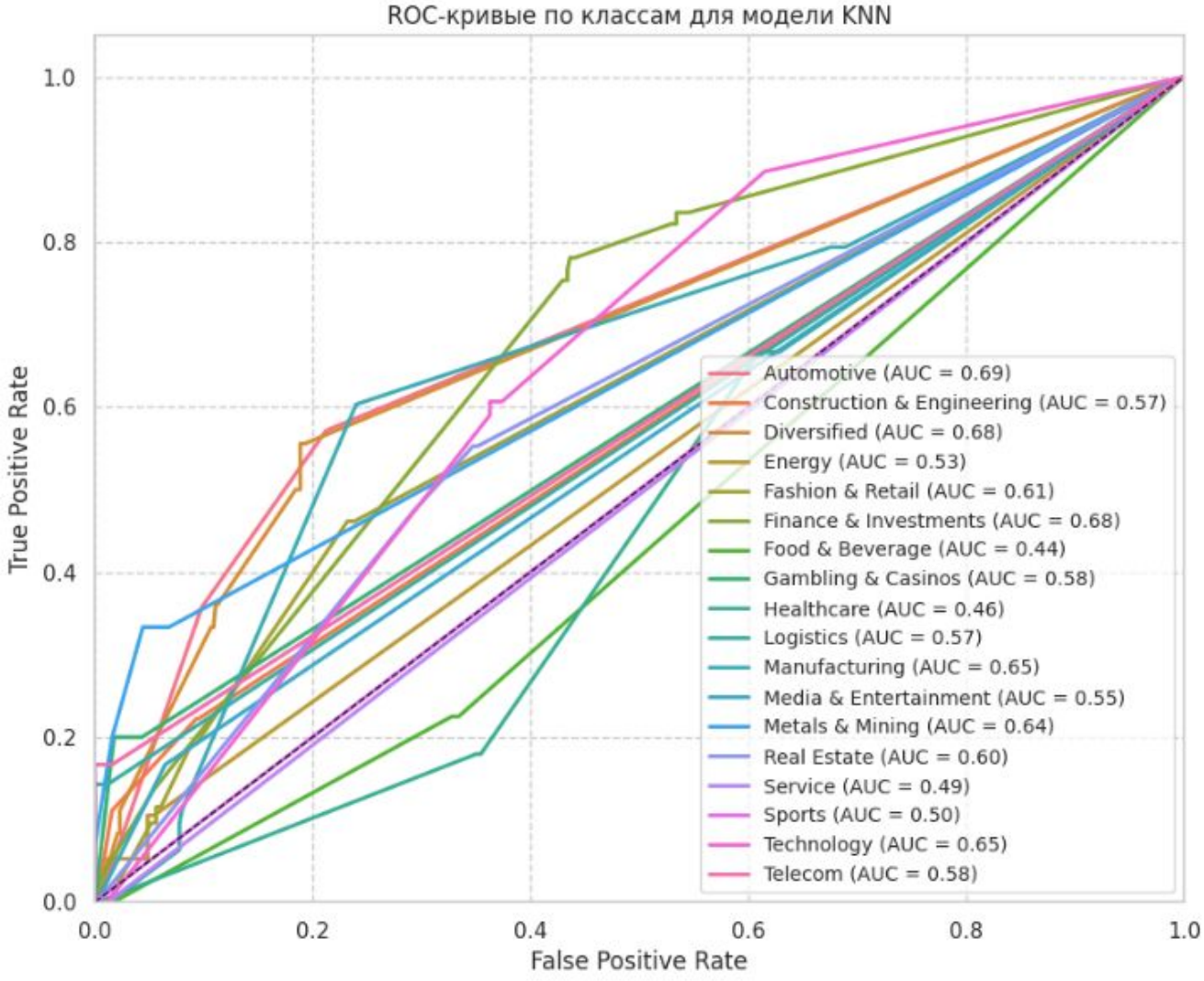
Лучшие параметры Random Forest: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 200}

Лучшие параметры KNN: {'knn__n_neighbors': 3, 'knn__weights': 'uniform'}






ROC-кривые по классам для модели Random Forest
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Automotive (AUC = 0.94)
Construction & Engineering (AUC = 1.00)
Diversified (AUC = 1.00)
Energy (AUC = 1.00)
Fashion & Retail (AUC = 1.00)
Finance & Investments (AUC = 1.00)
Food & Beverage (AUC = 1.00)
Gambling & Casinos (AUC = 1.00)
Healthcare (AUC = 1.00)
Logistics (AUC = 1.00)
Manufacturing (AUC = 1.00)
Media & Entertainment (AUC = 1.00)
Metals & Mining (AUC = 1.00)
Real Estate (AUC = 1.00)
Service (AUC = 1.00)
Sports (AUC = 1.00)
Technology (AUC = 1.00)
Telecom (AUC = 1.00)


ROC-кривые по классам для модели KNN
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Automotive (AUC = 0.69)
Construction & Engineering (AUC = 0.57)
Diversified (AUC = 0.68)
Energy (AUC = 0.53)
Fashion & Retail (AUC = 0.61)
Finance & Investments (AUC = 0.68)
Food & Beverage (AUC = 0.44)
Gambling & Casinos (AUC = 0.58)
Healthcare (AUC = 0.46)
Logistics (AUC = 0.57)
Manufacturing (AUC = 0.65)
Media & Entertainment (AUC = 0.55)
Metals & Mining (AUC = 0.64)
Real Estate (AUC = 0.60)
Service (AUC = 0.49)
Sports (AUC = 0.50)
Technology (AUC = 0.65)
Telecom (AUC = 0.58)

Прогнозирование состояния

Используем:

Linear Regression:

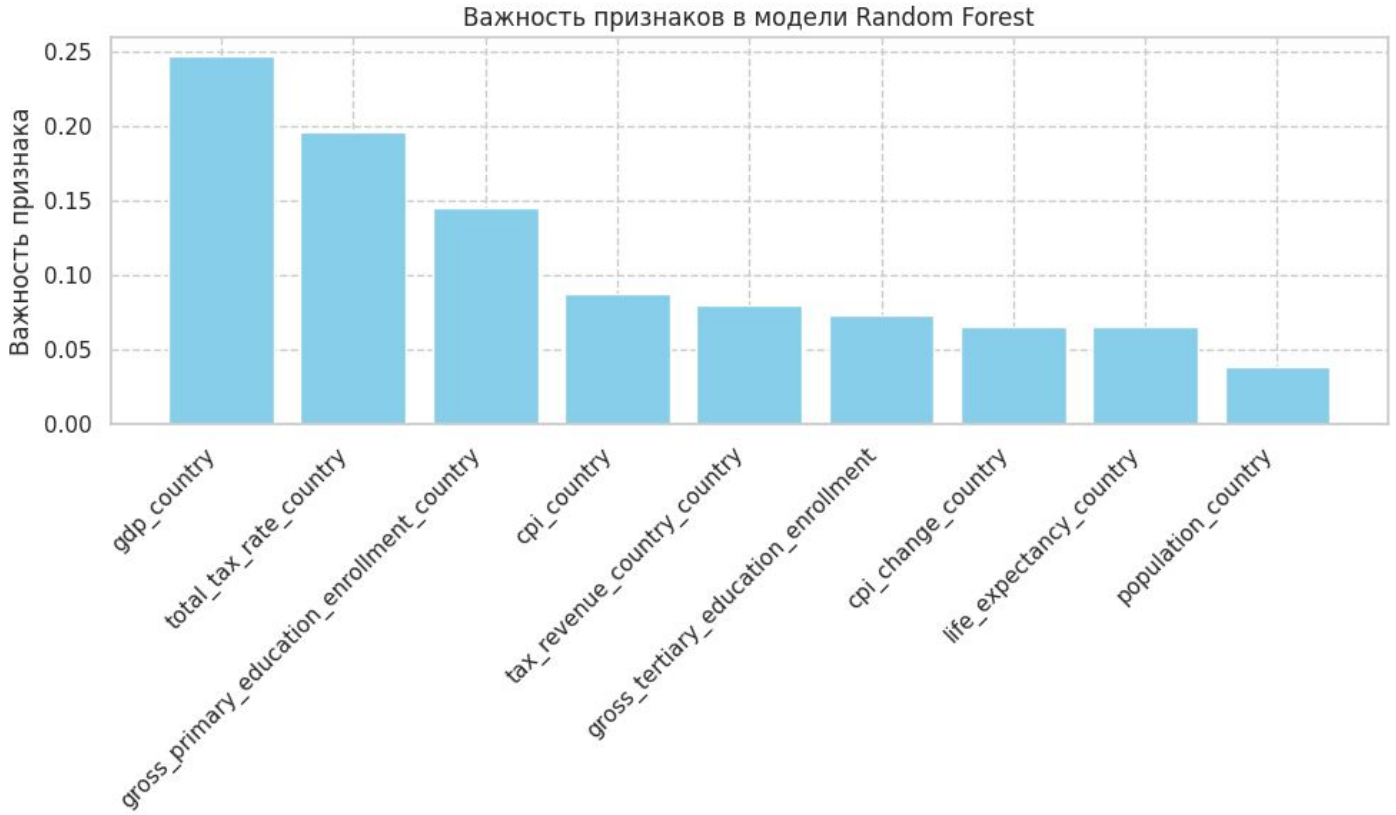
R2 = -0.045,
RMSE = 5204,
MAE = 3350

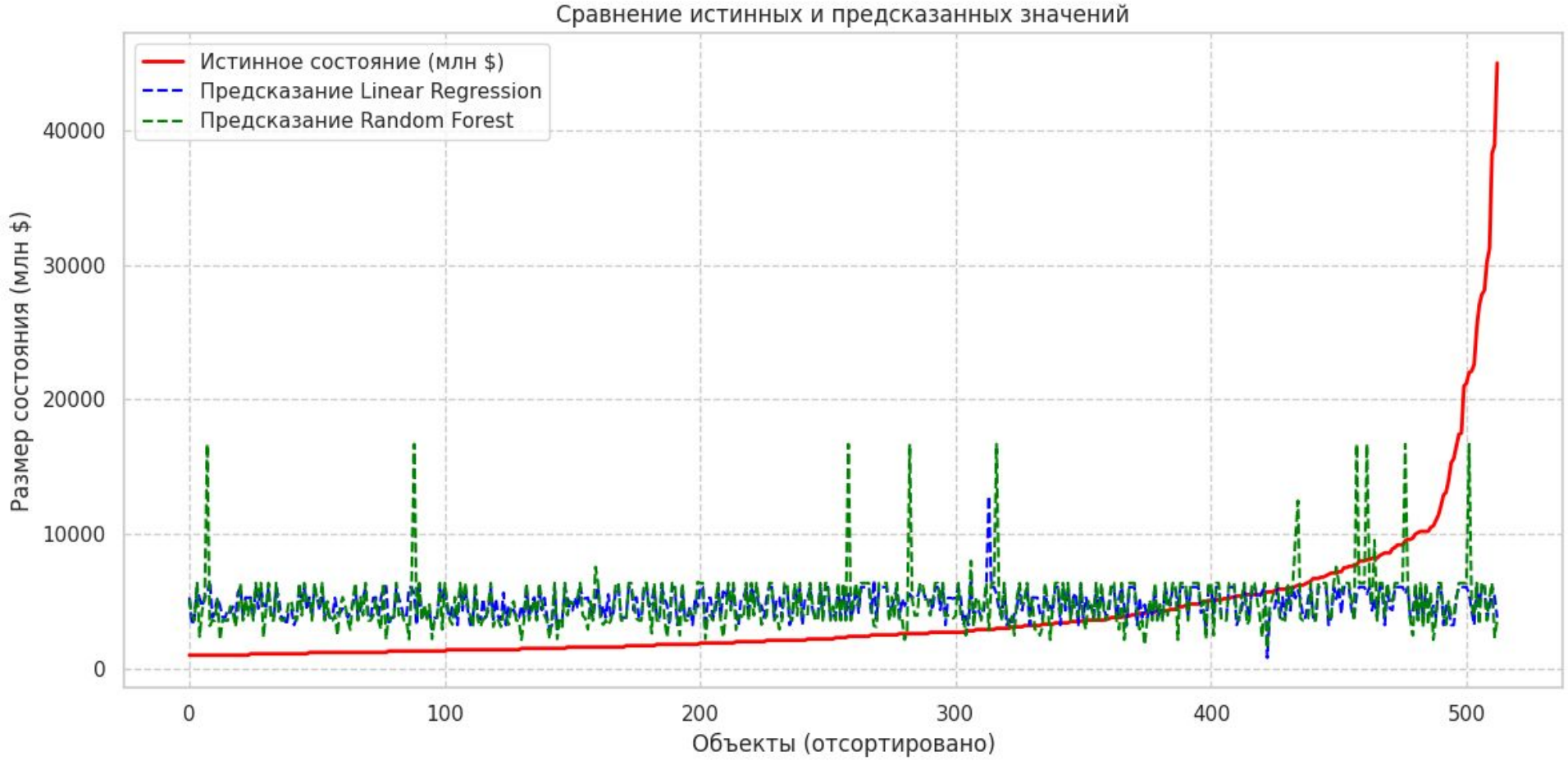
Random Forest:

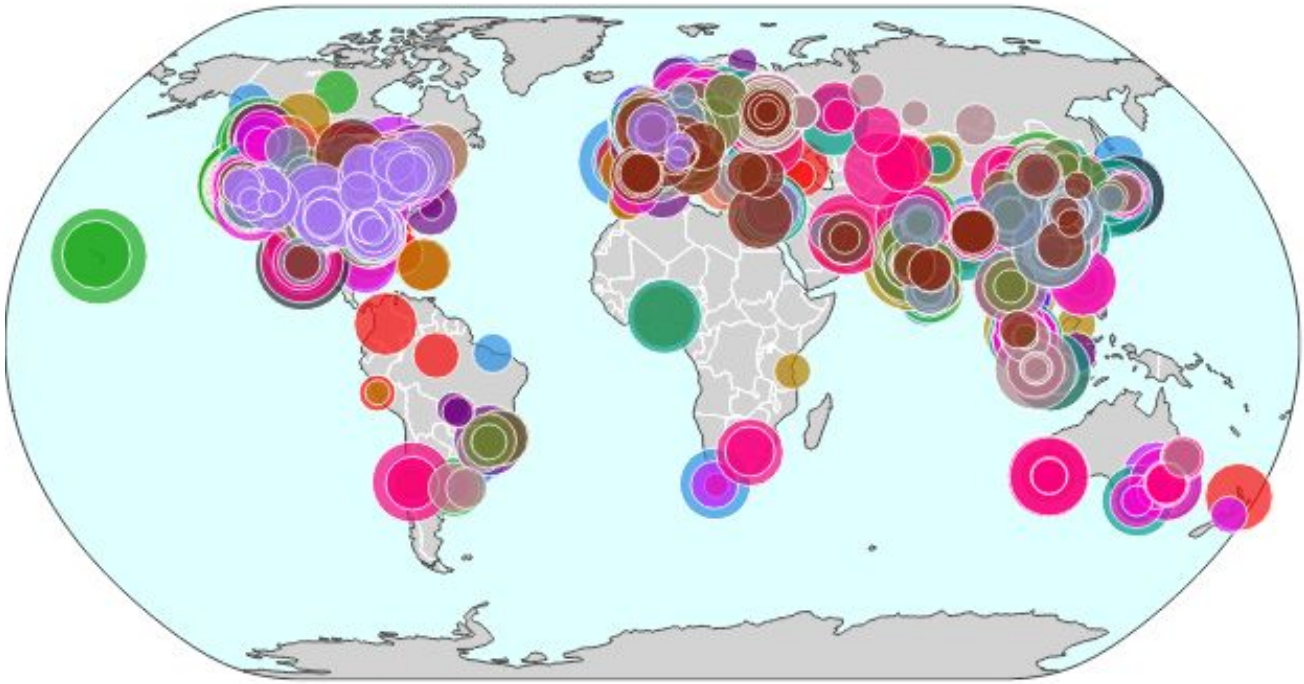
R2 = -0.107,
RMSE = 5355,
MAE = 3341

Лучшие параметры RandomForest:

'max_depth': 10,
'min_samples_split': 10,
'n_estimators': 200




Сравнение истинных и предсказанных значений
Истинное состояние (млн $)
Предсказание Linear Regression
Предсказание Random Forest
Размер состояния (млн $)
0
10000
20000
30000
40000
0
100
200
300
400
500
Объекты (отсортировано)



- Fashion & Retail
- Automotive
- Technology
- Finance & Investments
- Media & Entertainment
- Telecom
- Diversified
- Food & Beverage
- Logistics
- Gambling & Casinos
- Manufacturing
- Real Estate
- Metals & Mining
- Energy
- Healthcare
- Service
- Construction & Engineering
- Sports

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

HDBSCAN группирует точки данных на основе их плотности:

- **Кластер** — область с высокой плотностью точек.
- **Шум (Noise)** — точки в разреженных областях, не принадлежащие ни одному кластеру.

В отличие от DBSCAN, который использует фиксированные параметры `eps` (радиус окрестности) и `min_samples` (минимальное число соседей), HDBSCAN:

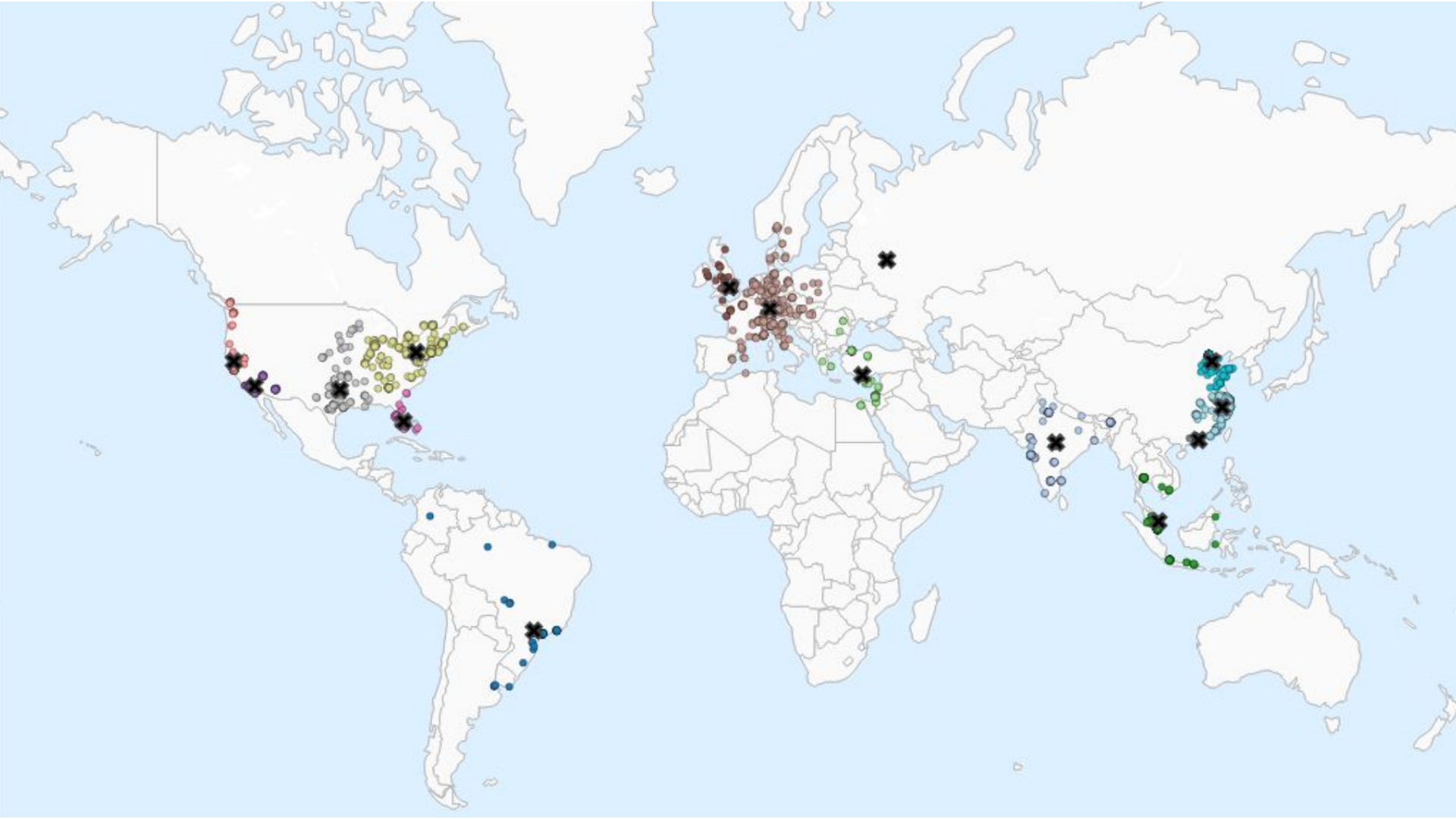
- Строит **иерархию кластеров** для разных уровней плотности.
- Автоматически выбирает оптимальные кластеры из этой иерархии.

1. **Построение графа взаимной достижимости**
   - Для каждой точки вычисляется **расстояние до `k`-го ближайшего соседа** (core distance).
   - Строится взвешенный граф, где ребра — это **взаимная достижимость** (mutual reachability distance):

$$d_{\text{mreach}}(a, b) = \max\left(\text{core}_k(a), \text{core}_k(b), d(a, b)\right)$$

   где $d(a, b)$ — обычное расстояние между точками.

2. **Построение минимального остовного дерева (MST)**
   - Граф преобразуется в **минимальное остовное дерево**, чтобы выявить иерархическую структуру.
3. **Построение иерархии кластеров**
   - Дерево "разрезается" на разных уровнях плотности, формируя **дендрограмму**.
4. **Выбор устойчивых кластеров**
   - Используется метрика **стабильности** (stability), чтобы выбрать наиболее устойчивые кластеры из иерархии.
   - Кластеры с максимальной стабильностью сохраняются, остальные отбрасываются.
5. **Пометка шума**
   - Точки, не вошедшие в устойчивые кластеры, помечаются как **шум (-1)**.



ВСЁ!