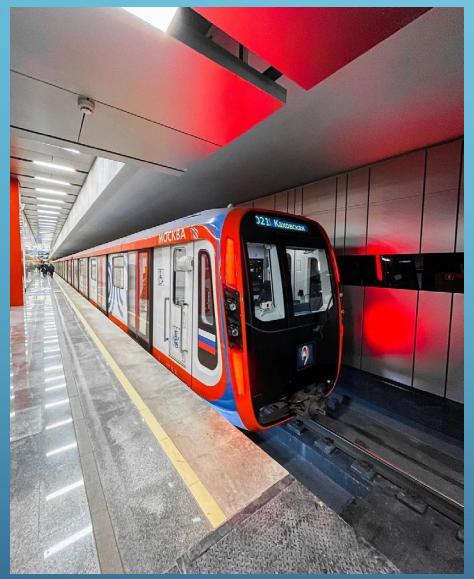
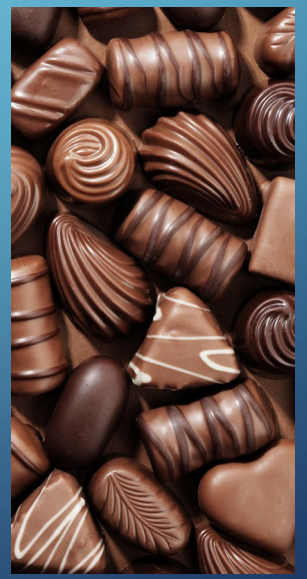
ВЛИЯНИЕ МЕТРОПОЛИТЕНА НА ЛЮБОВЬ К ШОКОЛАДУ

ИССЛЕДОВАТЕЛЬСКИЙ АНАЛИЗ ФАКТОРОВ ПОТРЕБЛЕНИЯ ШОКОЛАДА В СТРАНАХ С РАЗВИТОЙ СИСТЕМОЙ МЕТРОПОЛИТЕНА И БЕЗ НЕЁ





Итоговый проект по курсу обучения «DATA ANALYST»

Каштанов Павел

Цель исследования:

Определение наличия статистически значимой связи между наличием метрополитена в стране и уровнем потребления шоколада на душу населения.

Определение задач:

- Провести статистический анализ данных, выявить имеющиеся корреляционные связи.
- Оценить влияние экономических и демографических факторов на потребление шоколада.
- Проверить статистическую значимость выявленных закономерностей.
- Визуализировать полученные данные.

Проработка этапов анализа данных

1. Определение основной задачи.

- Провести статистический анализ данных;
- Определить факторы, влияющие на формирование потребительских привычек потребления шоколада.

2. Источник данных.

Базы данных с сайтов Kaggle, World Bank Group, FAOSTAT

Методы исследования:

Описательная статистика; Корреляционный анализ; Регрессионный анализ; Методы статистического сравнения групп

3. Сбор данных.

Поиск и выгрузка данных с сайтов Kaggle, World Bank Group, FAOSTAT

Проработка этапов анализа данных

4. Очистка данных.

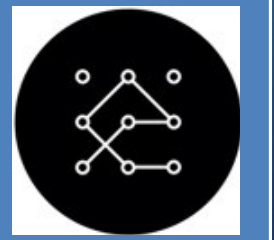
- Проверка целостности данных;
- Преобразование типов и форматов данных;
- Обработка пропущенных значений;
- Обработка выбросов.

Инструменты:

SQL (SQLite), Python (Pandas, Missingno, Seaborn, Matplotlib, NumPy), Microsoft Excel (Power Query)

5. Анализ данных.

- Описательный анализ;
- Методы статистического сравнения групп (t-test, test Mann–Whitney);
- Корреляционный анализ (тепловая карта);
- Регрессионный анализ;
- Машинное обучение;
- Визуализация.



Инструменты:

Python (Pandas, SciPy, Statsmodels, NumPy, Matplotlib, Seaborn, Xgboost)

Проработка этапов анализа данных

6. Результат исследования.

- Выводы;
- Интерпретация полученных данных.

Инструменты:

Python

7. Итоговый отчёт.

- Создание Dashboard;
- Подготовка презентации.

Инструменты:

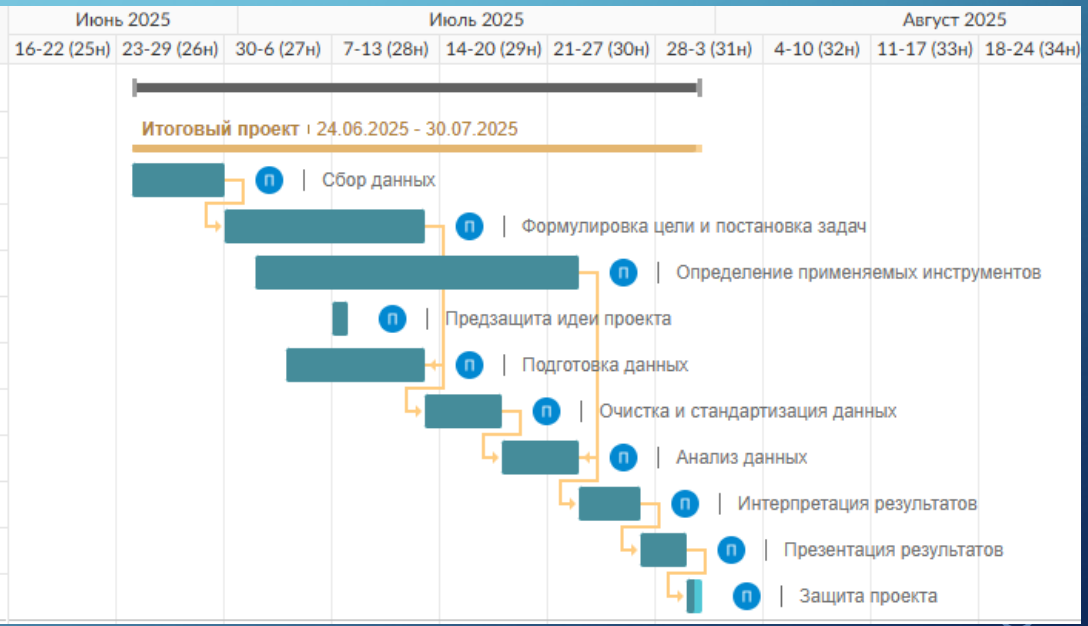
Power BI, Microsoft PowerPoint

Стадии выполнения проекта

Задачи	Дата начала	Срок выполнения	Дни	Статус
Сбор данных	24.06.2025	29.06.2025	6	Выполнено
Формулировка цели и постановка задач	30.06.2025	12.07.2025	13	Выполнено
Определение применяемых инструментов	02.07.2025	22.07.2025	21	Выполнено
Предзащита идеи проекта	07.07.2025	07.07.2025	1	Выполнено
Подготовка данных	04.07.2025	12.07.2025	9	Выполнено
Очистка и стандартизация данных	13.07.2025	17.07.2025	5	Выполнено
Анализ данных	18.07.2025	22.07.2025	5	Выполнено
Интерпретация результатов	23.07.2025	26.07.2025	4	Выполнено
Презентация результатов	27.07.2025	29.07.2025	3	Выполнено
Защита проекта	30.07.2025	30.07.2025	1	Выполняется

Стадии выполнения проекта

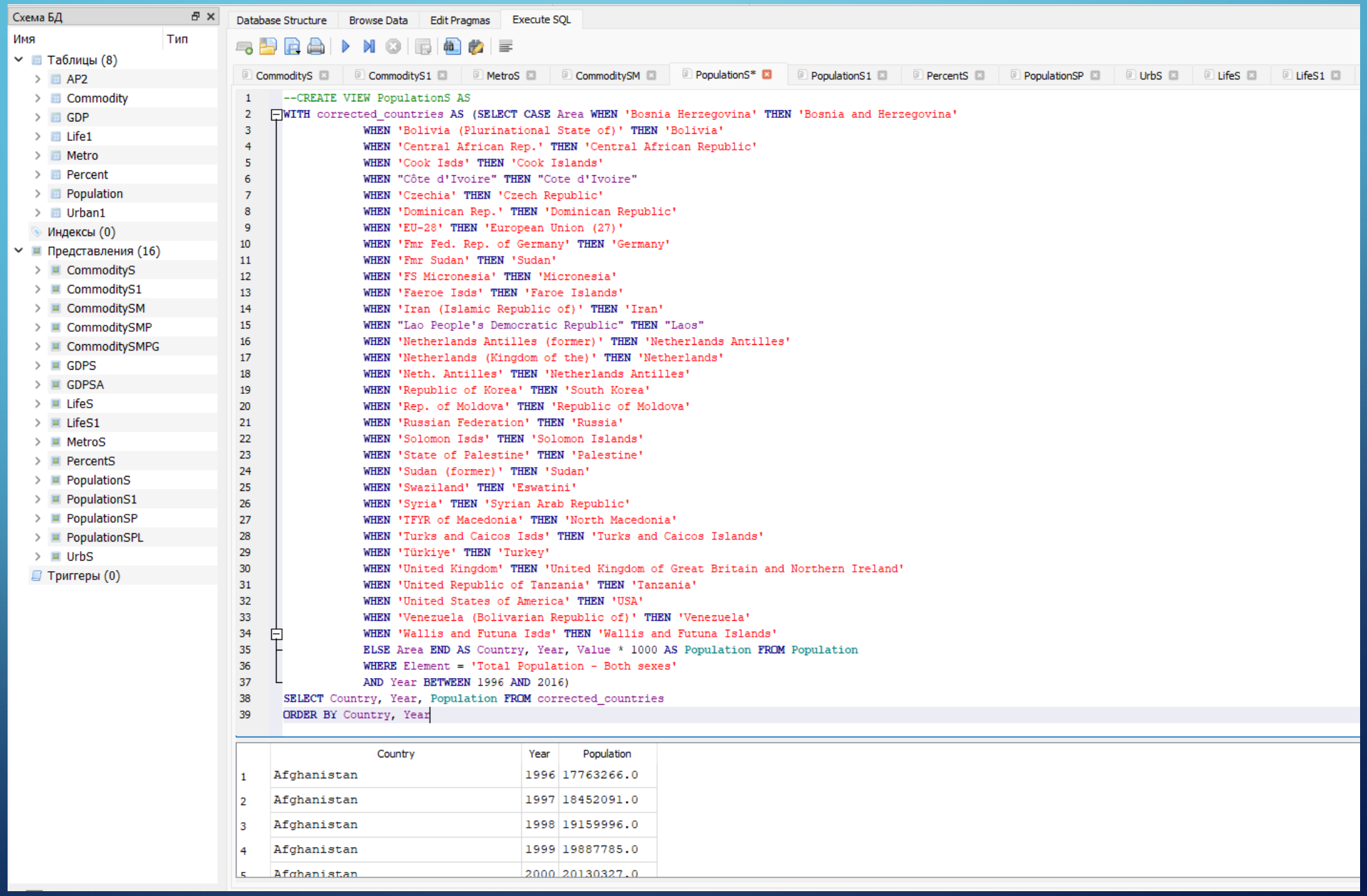
	Задача	Начало	Длительность	Завершение	Исполнитель	Статус
		24.06.2025	37д	**30.07.2025**		
1	**Итоговый проект**	**24.06.2025**	**37д**	**30.07.2025**		
1.1	Сбор данных	24.06.2025	6д	29.06.2025	Павел Каштанов	Выполнено
1.2	Формулировка цели и постановка задач	30.06.2025	13д	12.07.2025	Павел Каштанов	Выполнено
1.3	Определение применяемых инструментов	02.07.2025	21д	22.07.2025	Павел Каштанов	Выполнено
1.4	Предзащита идеи проекта	07.07.2025	1д	07.07.2025	Павел Каштанов	Выполнено
1.5	Подготовка данных	04.07.2025	9д	12.07.2025	Павел Каштанов	Выполнено
1.6	Очистка и стандартизация данных	13.07.2025	5д	17.07.2025	Павел Каштанов	Выполнено
1.7	Анализ данных	18.07.2025	5д	22.07.2025	Павел Каштанов	Выполнено
1.8	Интерпретация результатов	23.07.2025	4д	26.07.2025	Павел Каштанов	Выполнено
1.9	Презентация результатов	27.07.2025	3д	29.07.2025	Павел Каштанов	Выполнено
1.10	Защита проекта	30.07.2025	1д	30.07.2025	Павел Каштанов	В работе



Описание исходных данных

Наименование столбца	Определение
Country	Название страны
Year	Год (1996-2016)
Total_weight_kg	Общий вес импортируемого и ре-импортируемого шоколада в страну (кг)
availability_of_metro	Наличие метро (да/нет)
lines	Количество линий метро в стране
stations	Количество станций метро в стране
annual_ridership_mill	Пассажиропоток (млн человек)
Population_eating_chocolate	Часть населения, которая потребляет шоколад
Cost_kg_USD	Стоимость 1 кг шоколада в стране (USD)
Urban_population	Доля городского населения страны (%)
Lifetime	Средняя продолжительность жизни в стране
GDP_USD	ВВП на душу населения (USD)
Inflation	Процент инфляции
kg_per_person	Потребление шоколада на душу населения
kg_per_GDP	Потребление шоколада на единицу ВВП (шоколадоёмкость экономики)

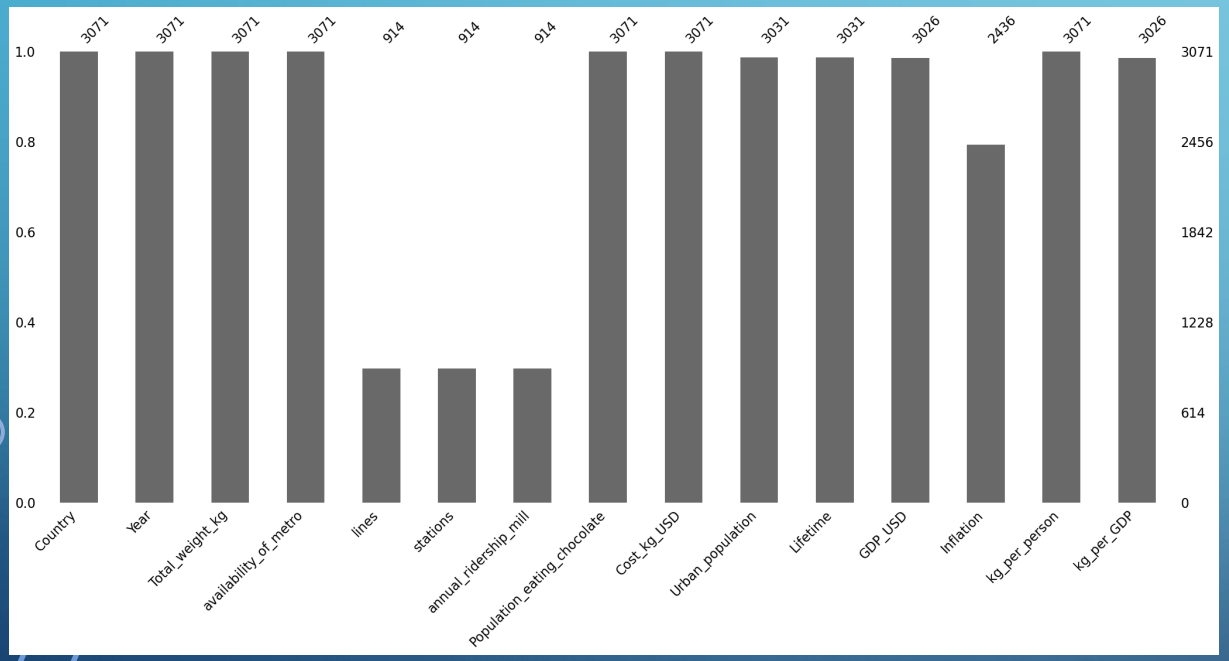
Очистка данных (DB Browser for SQLite)



Очистка данных (здесь и далее: Jupyter Notebook и Python)

1 ЭТАП – Проверка и заполнение миссингов:

Total_weight_kg: заполняется средними значениями по каждой стране
Inflation, Urban_population, Lifetime: заполняются средними значениями по году
GDP_USD: для некоторых стран (Faroe Islands, Mayotte и др.) заполняется на основе значений других стран kg_per_person, kg_per_GDP: вычисляется после заполнения вышеуказанных столбцов



Заполнение миссингов в Total_weight_kg:

```
country_means =
df.groupby('Country')['Total_weight_kg'].transform('mean')
df['Total_weight_kg'] =
df['Total_weight_kg'].mask((df['Total_weight_kg'].isna()),
country_means)
```

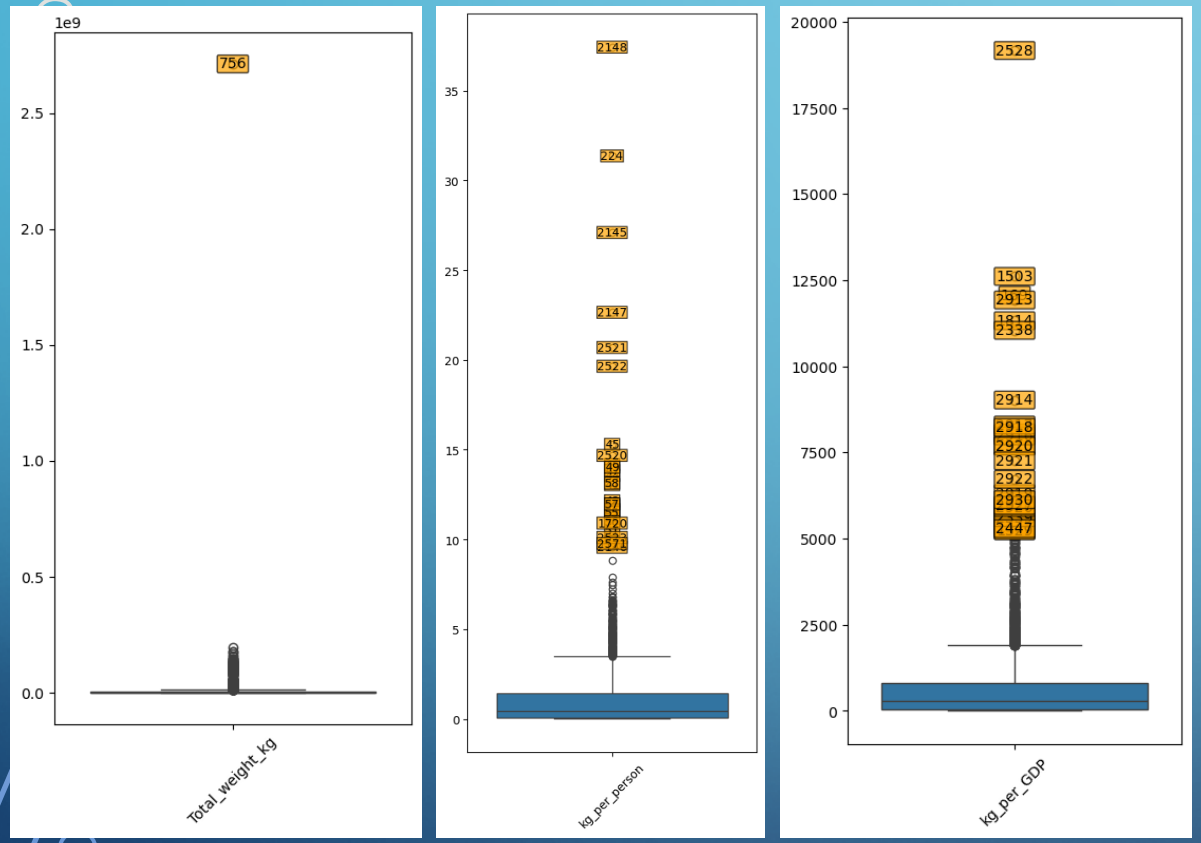
Заполнение миссингов в kg_per_person:

```
mask_kg_per_person = df['kg_per_person'].isna()
df.loc[mask_kg_per_person, 'kg_per_person'] =
(df.loc[mask_kg_per_person, 'Total_weight_kg'] /
df.loc[mask_kg_per_person,
'Population_eating_chocolate']).astype('float32')
```

Очистка данных

2 ЭТАП – Проверка и удаление выбросов:

Наиболее серьёзные выбросы замечены в столбцах Total_weight_kg, kg_per_person, kg_per_GDP



Найдено выбросов в 'Total_weight_kg': 1
Найдено выбросов в 'kg_per_person': 29
Найдено выбросов в 'kg_per_GDP': 33
Итого оставшихся строк: 3008

Функция для поиска выбросов методом z-score (из библиотеки NumPy):

```
def find_outliers_zscore(data, column, threshold=4):
z_scores = np.abs(stats.zscore(data[column]))
return data.loc[z_scores > threshold].index.tolist()
```

* threshold=4 принят, чтобы не удалять «почти нормальные» значения.

Описательный анализ

Основные наблюдения:

1. Потребление шоколада изменяется с течением времени без резких скачков.
2. Потребление шоколада стабильно выше, чем в городах без метро.
3. Спад 2009 года – скорее всего связан с мировым финансовым кризисом.

Статистическое сравнение групп

T-test

Test Mann–Whitney (U-test)

Нулевая гипотеза (H0): «В странах с метро уровень потребления шоколада (kg_per_person) среди населения такой же как и без метро»

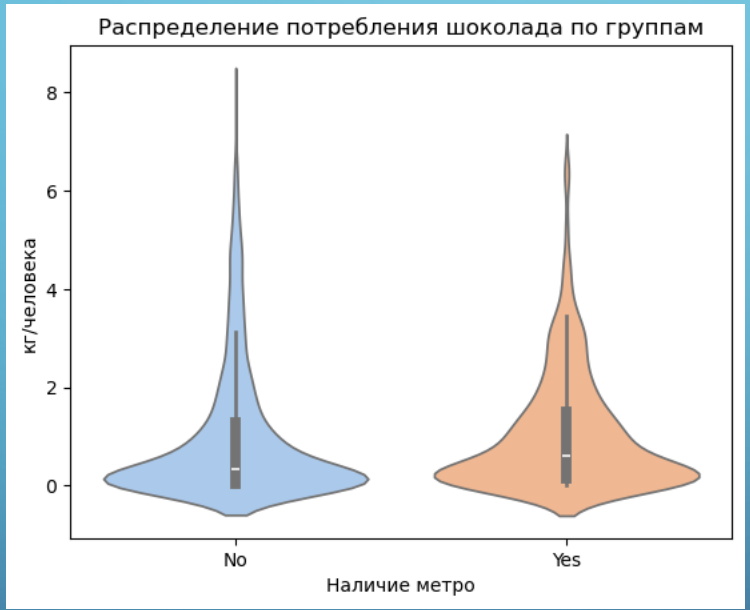
Проведение теста:

```
with_metro = dfW2[dfW2['availability_of_metro'] == "Yes"]['kg_per_person']
without_metro = dfW2[dfW2['availability_of_metro'] == "No"]['kg_per_person']
t_statistic, p_value = stats.ttest_ind(with_metro, without_metro)
```

Результаты теста:

T-статистика: 1.336

P-значение: 0.1816 (>0.05)



Результаты теста:

Медиана потребления (Страны с метро) : 0.61 кг/чел

Медиана потребления (Страны без метро) : 0.34 кг/чел

Статистика: 1083137.50

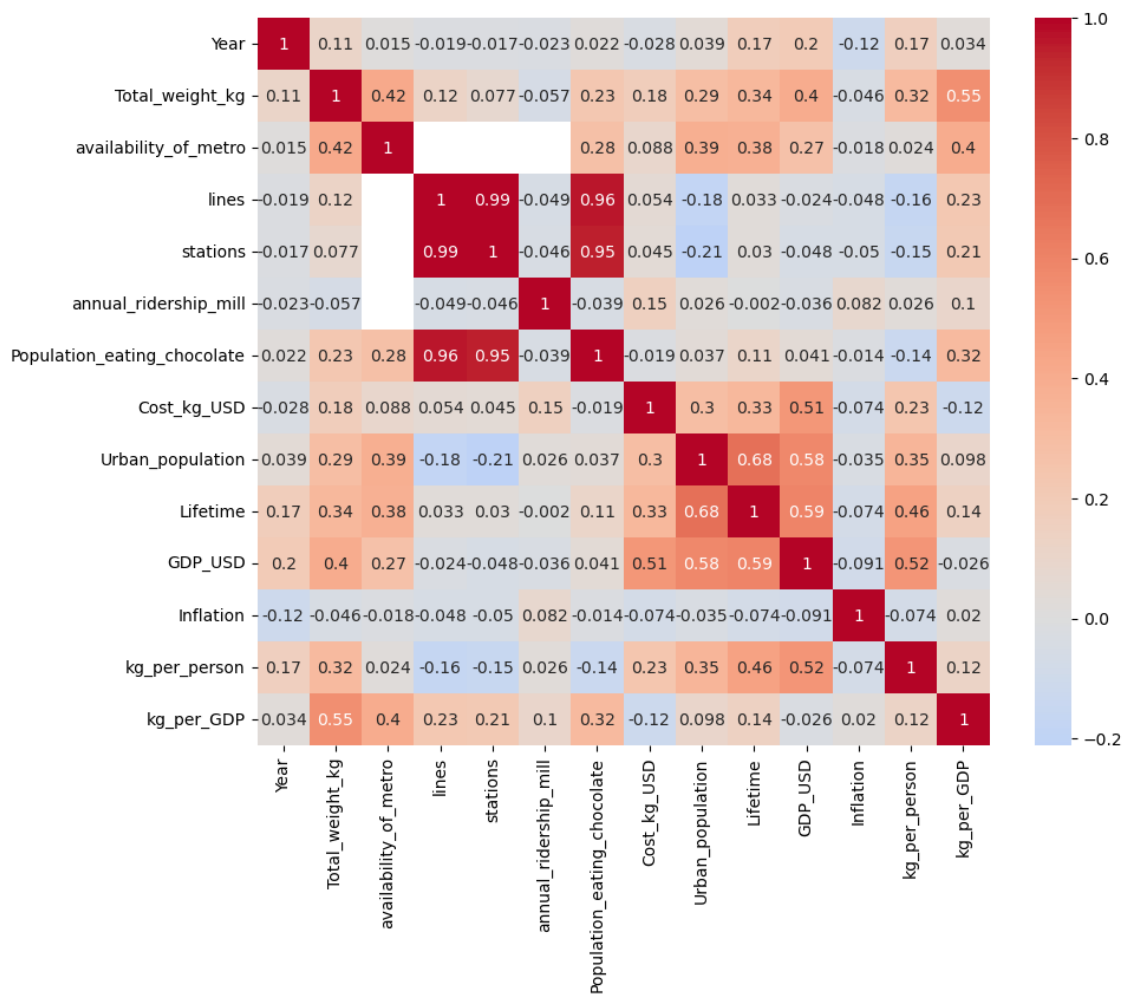
P-значение: 0.0000

Результаты теста Cohen’s d (Стандартизированная разница средних):

Cohen’s d: 0.055 (Очень слабый эффект)

Вывод: Согласно U-тесту: медианное потребление шоколада значимо выше в странах с метро.
Согласно T-тесту и Cohen’s d-тесту: нет статистически значимых различий в среднем потреблении шоколада. Таким образом, нулевая гипотеза подтверждается только при U-тесте.

Корреляционный анализ



Интересные корреляции:

1. Inflation и annual_ridership_mill практически не коррелируют ни с одним из показателей.
2. Lines и stations практически не связанны с количеством потребления шоколада.
3. GDP_USD и kg_per_GDP (-0.026) – богатые страны тратят на шоколад меньшую долю экономики.
4. Urban_population и kg_per_person (0.35) – городские жители едят больше шоколада, чем сельские.
5. kg_per_person и Lifetime (0.459) – шоколад продлевает жизнь!

Корреляция между availability_of_metro и kg_per_person: 0.024, т.е. практически отсутствует. Корреляция между availability_of_metro и kg_per_GDP: 0.4 (средняя).
На основании того что, нет статистически значимых различий в среднем потреблении шоколада для дальнейшего исследования примем новую гипотезу: Наличие в стране метро увеличивает шоколадоёмкость экономики (kg_per_GDP).

Статистическое сравнение групп-2

T-test

Test Mann–Whitney (U-test)

Нулевая гипотеза (H0): «В странах с метро шоколадоёмкость экономики (kg_per_GDP) такая же как и без метро»

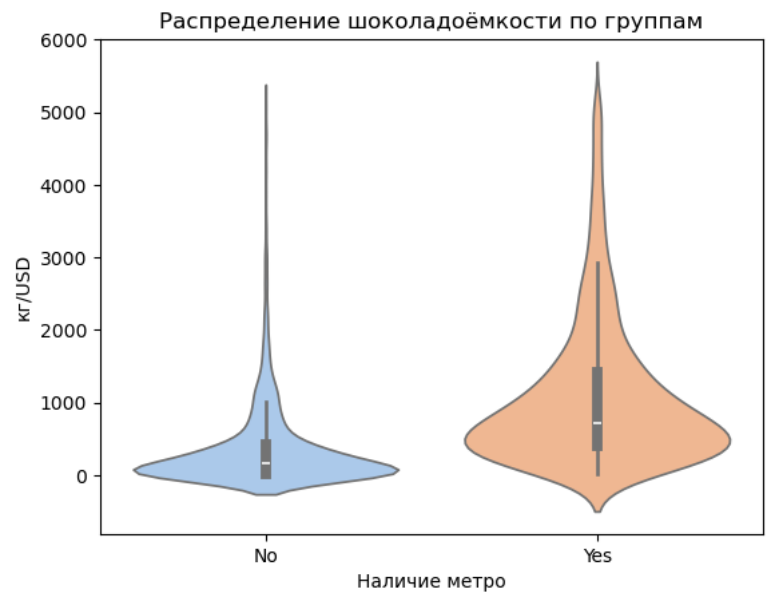
Проведение теста:

```
with_metro = dfW2[dfW2['availability_of_metro'] == "Yes"]['kg_per_GDP']
without_metro = dfW2[dfW2['availability_of_metro'] == "No"]['kg_per_GDP']
t_statistic, p_value = stats.ttest_ind(with_metro, without_metro)
```

Результаты теста:

T-статистика: 23.8247

P-значение: 0.000 (<0.05)



Результаты теста:

Медиана шоколадоёмкости (Страны с метро) : 724.68 кг/USD

Медиана шоколадоёмкости (Страны без метро) : 163.30 кг/USD

Статистика: 1519378.00

P-значение: 0.0000

Результаты теста Cohen’s d

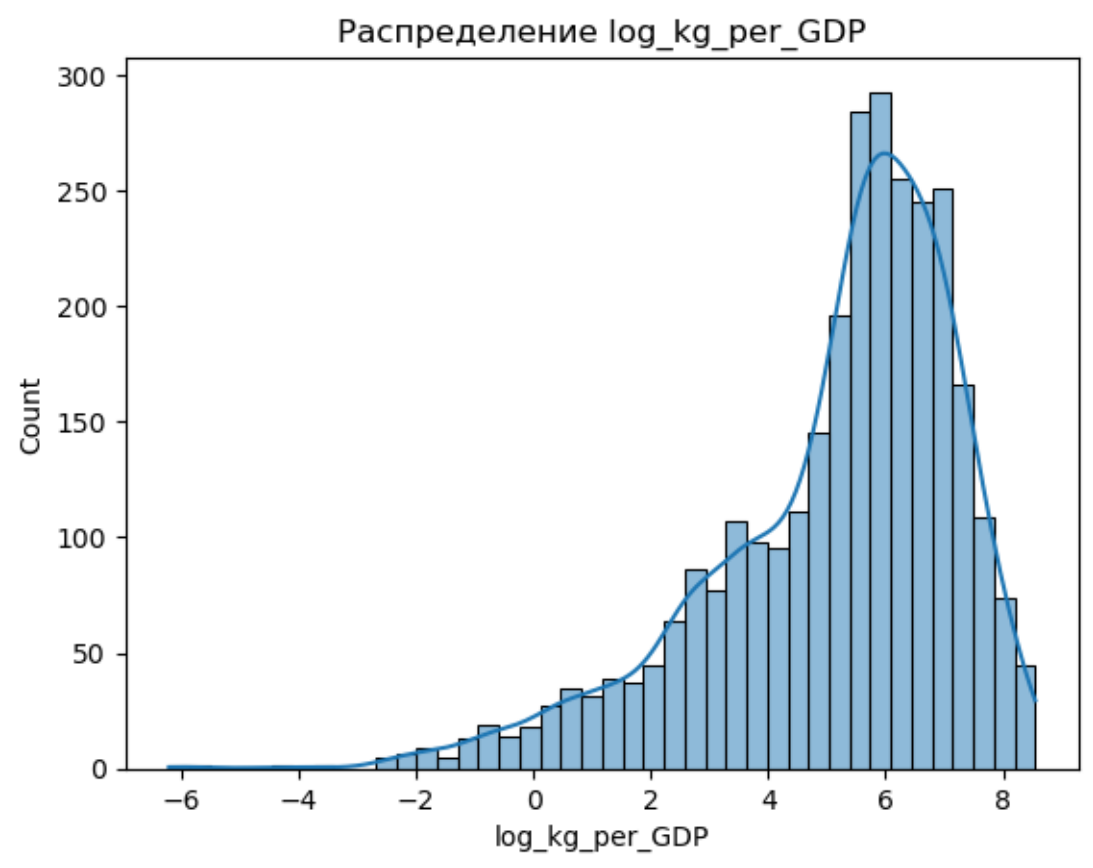
(Стандартизированная разница средних):

Cohen’s d: 0.855 (Очень сильный эффект)

Вывод: Согласно всем 3 тестам: шоколадоёмкость значимо выше в странах с метро. Таким образом, нулевая гипотеза подтверждается.

Регрессионный анализ

Принимаем зависимую переменную - kg_per_GDP и независимые переменные - availability_of_metro, Urban_population, Cost_kg_USD, GDP_USD, kg_per_person



Коэффициент при availability_of_metro = 2.396 — наличие метро сильно увеличивает шоколадоёмкость.
Вывод: Гипотеза подтверждается с высокой статистической значимостью.

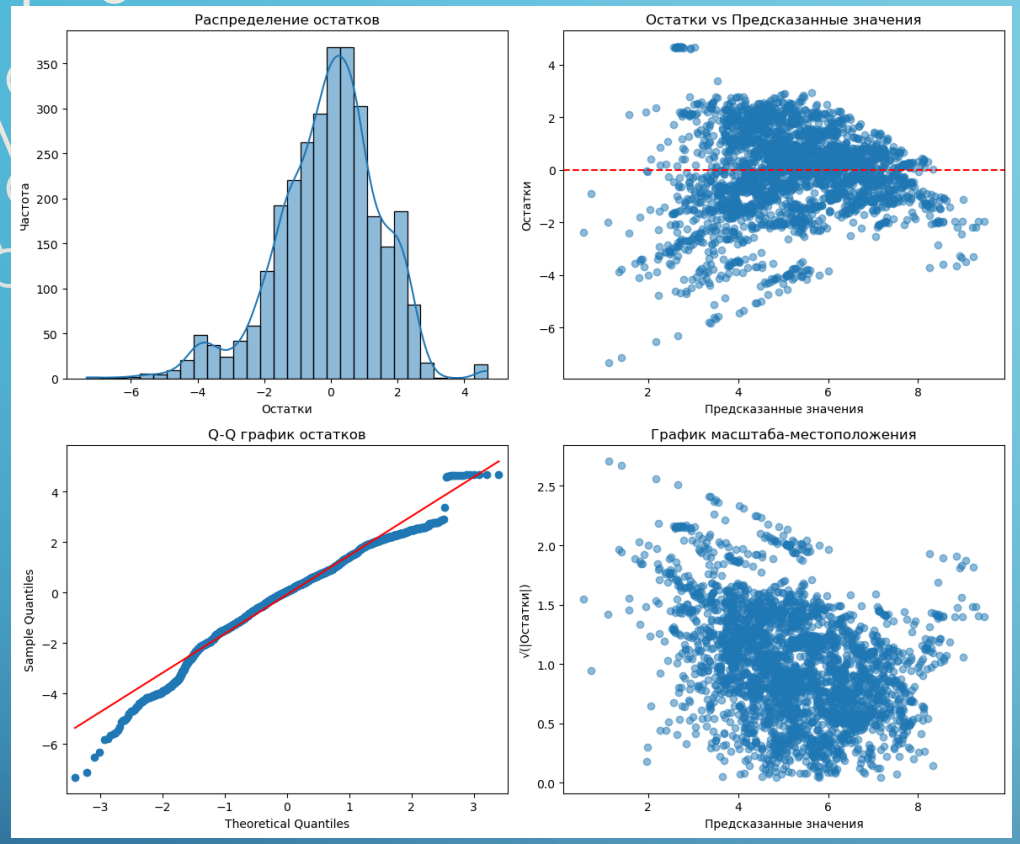
Для показателей kg_per_GDP, Cost_kg_USD, GDP_USD, kg_per_person применено логарифмирование для линеаризации зависимости.

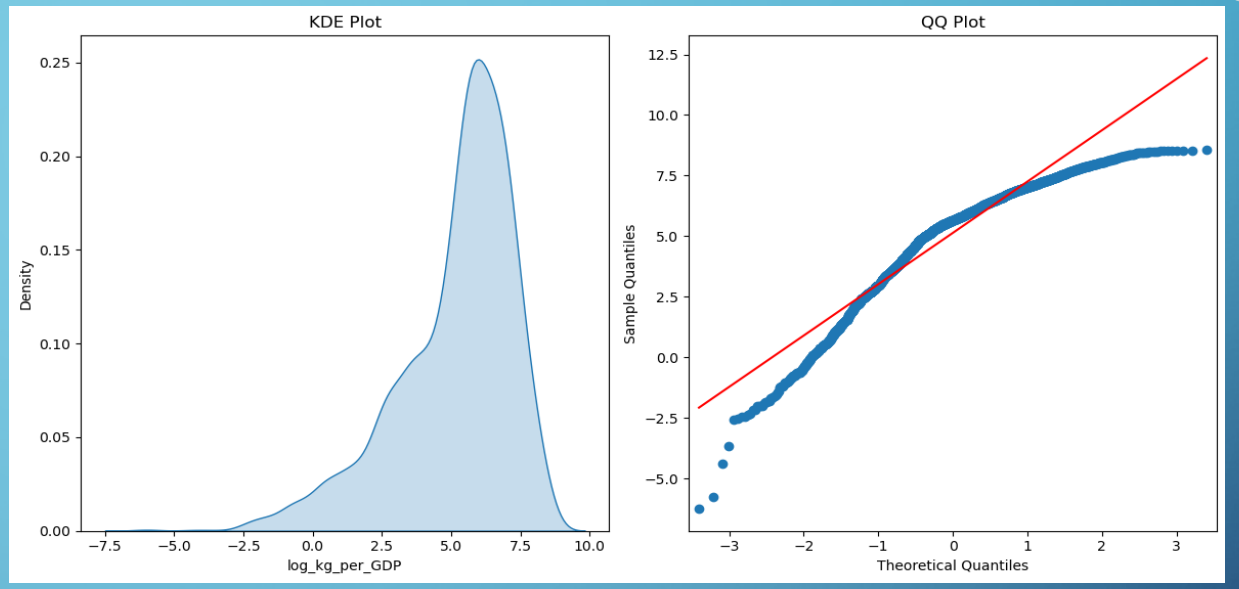
Ключевые показатели:
R^2 = 0.464 – объяснено 46,4% вариации log_kg_per_GDP (умеренно высокий показатель)
p-value = 0 (для всех переменных) – все значимые.
Cond. No. = 675 – умеренная мультиколлинеарность.
Durbin-Watson = 0.185 – сильная положительная автокорреляция.

Результаты теста Shapiro-Wilk (Нормальности распределения зависимой переменной):
Shapiro-Wilk Test Statistic: 0.92 и P-значение: 0.0000 (распределение ненормальное)

Коэффициент асимметрии (skewness) = -1.158
Отрицательное значение указывает на левую асимметрию (длинный хвост влево).

Дисперсионный анализ (Робастная линейная регрессия)





Малый эффект выбросов: Всего 3.76% данных (113 строк), т.е. модель в целом устойчива.
Медианное абсолютное отклонение (MAD) = 0.918 — относительно небольшой разброс ошибок вокруг медианы.

Модель:

model = smf.rlm('log_kg_per_GDP ~ availability_of_metro + Urban_population + log_Cost_kg_USD + log_GDP_USD + log_kg_per_person', data=dfK1, M=sm.robust.norms.HuberT()).fit()

Метод оценки модели: Huber's T norm (устойчивый к выбросам, уменьшает их вес в модели).

Результаты теста:

Коэффициент при availability_of_metro = +2.215 — наличие метро сильно увеличивает шоколадоёмкость.
Сходимость достигнута за 25 итераций, что говорит о стабильности оценок.

Вывод: Гипотеза подтверждается.

Регрессионный анализ (библиотека Scikit-Learn)

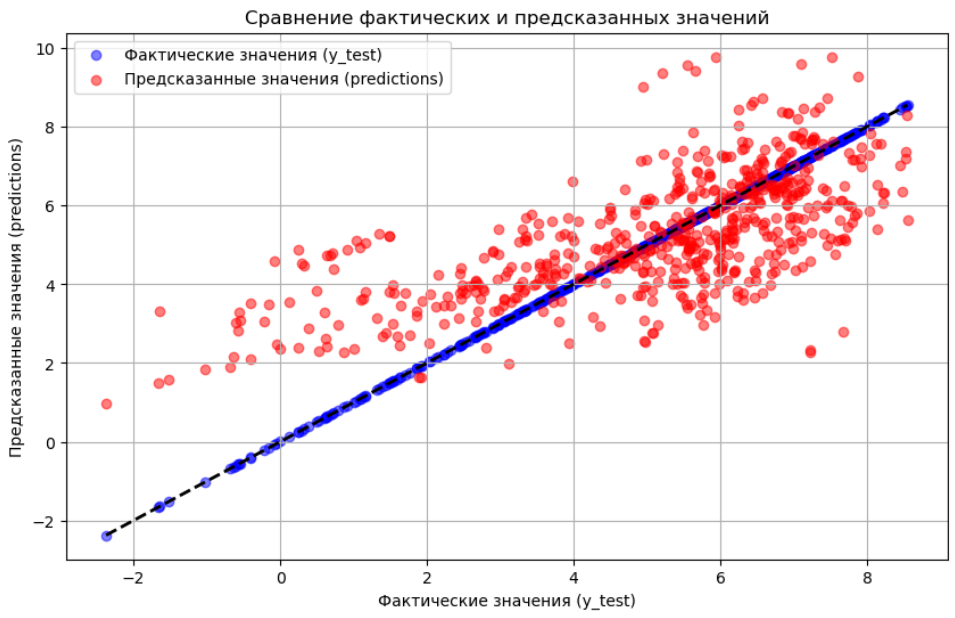
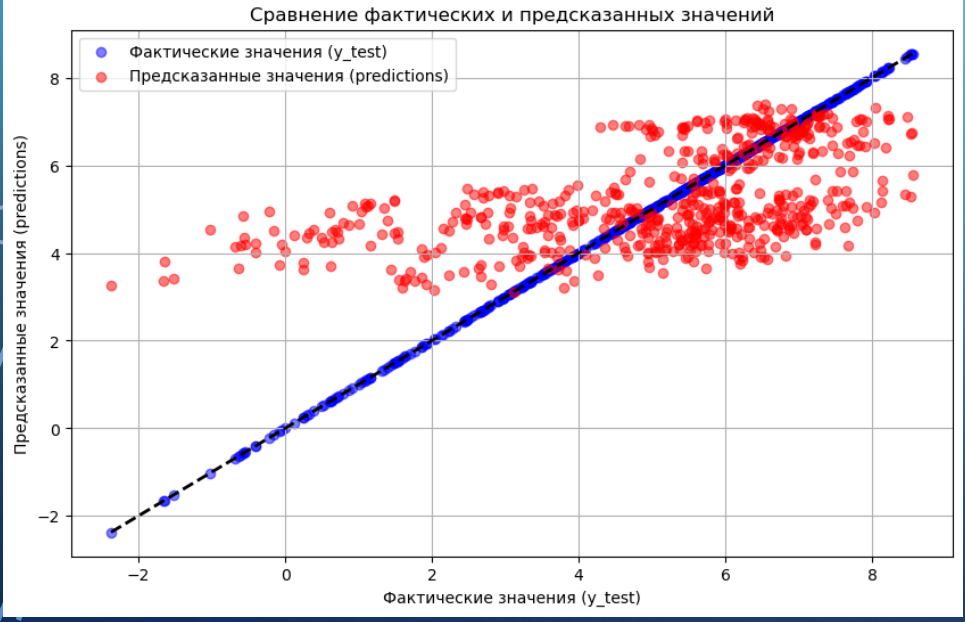


График без показателей log_Cost_kg_USD, log_GDP_USD



Модель:

X = dfK1[['availability_of_metro', 'Urban_population', 'log_Cost_kg_USD', 'log_GDP_USD', 'log_kg_per_person']]
y = dfK1['log_kg_per_GDP']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=170)
lm = LinearRegression()lm.fit(X_train,y_train)predictions = lm.predict(X_test)

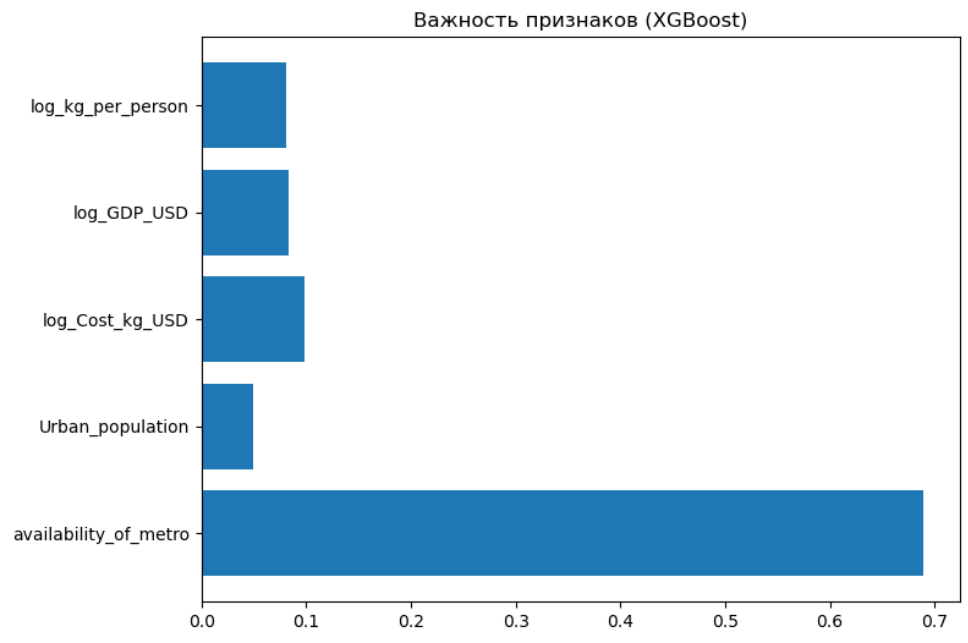
Результаты теста:

MAE: 1.24; MSE: 2.56; RMSE: 1.6 – неплохие, но не идеальные предсказания (11–15% от общего разброса данных).
log_Cost_kg_USD, log_GDP_USD – показали высокую мультиколлинеарность.
После масштабирования (все признаки приводятся к единому масштабу без изменения распределения) коэффициент при availability_of_metro = +1.09 — т.е. наличие метро увеличивает шоколадоёмкость.
R^2 = 0.422 - модель объясняет 42.2% дисперсии целевой переменной (средний показатель).
Вывод: Гипотеза подтверждается.

Машинное обучение

Для обучения применим 6 разных моделей:

1. **Linear Regression,**
2. **Polynomial Regression,**
3. **Decision Tree,**
4. **Random Forest,**
5. **Gradient Boosting,**
6. **XGBoost.**



Модель	MAE (Средняя ошибка предсказания)	MSE (Квадратичная ошибка)	R2 (Дисперсия)
XGBoost	0.296	0.232	0.947
Random Forest	0.341	0.375	0.915
Decision Tree	0.397	0.626	0.858
Gradient Boosting	0.804	1.182	0.732
Polynomial Regression	1.084	2.119	0.521
Linear Regression	1.240	2.556	0.421

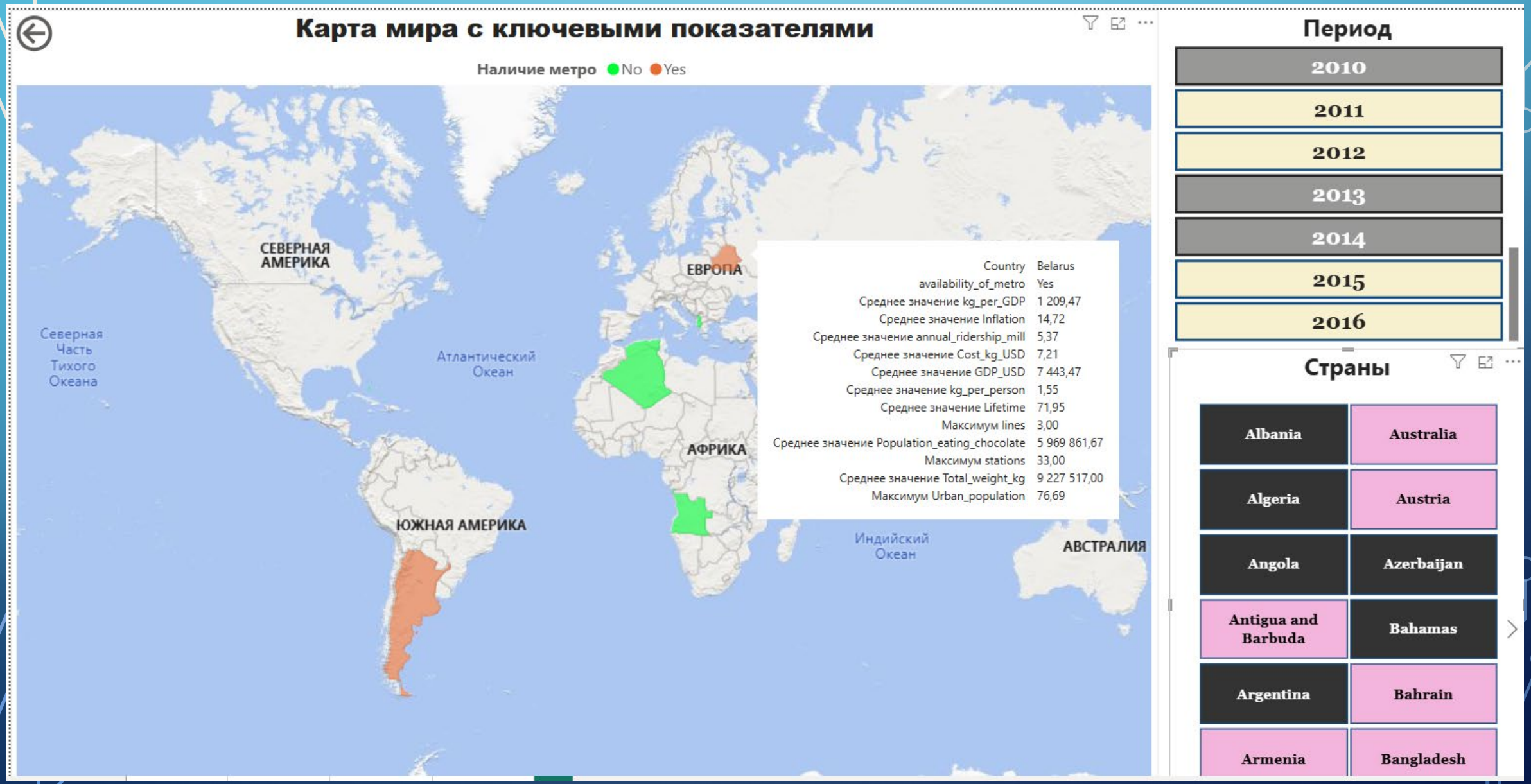
Вывод: Модель XGBoost лучше всех улавливает закономерности в данных. Linear и Polynomial Regression показали слабые результаты, что говорит о нелинейности данных.

Анализ важности признаков (XGBoost):

Самый важный признак влияющий на шоколадоёмкость – наличие метро.

Остальные признаки – практически одинаково менее важные.

Создание дашборда (Power BI)



Создание дашборда (Power BI)-2

Дата

2000

Страна

Mongolia

No	9,23	537	11,60	57,13
Наличие метро	Стоимость 1 кг шоколада (USD)	ВВП (USD)	Инфляция, %	Доля городского населения, %

1,22 млн	2 277,57	63,21	1 031 176
Вес импортируемого шоколада, кг	Потребление на душу населения (кг/человек)	Средняя продолжительность жизни	Население, потребляющее шоколад

Выводы:

В проекте были исследованы 2 гипотезы:

1. «Наличие в стране метро увеличивает уровень потребления шоколада среди населения» – отвергнута (нет статистически значимых различий в странах метро и без).
2. «Наличие в стране метро увеличивает шоколадоёмкость экономики» - подтвердилась (согласно всем проведённым этапам статистического анализа данных).

Однако в исследовании обнаружились ограничения:

1. Исследование носит корреляционный характер.
2. Наличие метро может быть показателем развитой инфраструктуры, а не прямой причиной роста потребления шоколада.
3. Заполнение пропусков средними значениями может исказить распределения данных.
4. Логарифмирование шоколадоёмкости не полностью устранило асимметрию.
5. Страны с экстремальными значениями (т.е. страны попавшие в выбросы) требуют отдельного анализа.
6. Существуют другие значимые переменные (например, GDP_USD).
7. Модель данных не является линейной, т.е. имеются сложные взаимосвязи.
8. Высокая мультиколлинеарность у log_GDP_USD (68) и log_Cost_kg_USD (47) затрудняет разделение их влияния на гипотезу.

Выводы:

В процессе исследования использовались следующие виды аналитики:

1. Описательная аналитика (Descriptive Analytics):

Использованные инструменты:

Визуализация данных: Графики распределения (histplot, kdeplot, violinplot); Heatmap корреляции.

Описательная статистика: Средние, медианы, стандартные отклонения для переменных; Группировка данных по наличию метро.

2. Диагностическая аналитика (Diagnostic Analytics):

Использованные инструменты:

Анализ пропущенных значений (msno.bar, заполнение значением средними и медианами).

Обнаружение и удаление выбросов: Метод z-score, Визуализация через boxplot.

Статистические тесты: t-test; Test Mann–Whitney; Расчёт эффекта через Cohen's d.

3. Предиктивная аналитика (Predictive Analytics):

Использованные инструменты:

Линейная регрессия (с использованием библиотек Statsmodels, Scikit-Learn).

Машинное обучение (6 моделей).

4. Предписывающая аналитика (Prescriptive Analytics):

Использованные инструменты:

Робастная регрессия (smf.rlm с HuberT).

Дисперсионный анализ (ANOVA): Проверка остатков на гетероскедастичность; Q-Q графики для проверки нормальности.

Оптимизация моделей: Подбор признаков через VIF (мультиколлинеарность); Масштабирование данных (StandardScaler, MinMaxScaler).

Рекомендации:

1. Учёт климатических факторов: холодный климат может влиять на потребление шоколада.
2. Локализация производства: В странах с высоким ВВП (где шоколадоёмкость ниже) сделать акцент на премиальный шоколад.
3. Сбор дополнительных данных о культуре питания: наличие кафе/магазинов в метро как медиатор потребления.
4. Анализ кластеров: выделить группы стран с похожими паттернами (например, Европа - Африка).
5. Учёт религиозных факторов.
6. Проверка эффекта времени после постройки метро.

СПАСИБО ЗА ВНИМАНИЕ!

Ссылка на проект:

https://github.com/Paskored/Data-Analyst-Project_Kashtanov