**Тестовое задание**

**Аналитик данных**

Общие требования

При выполнении тестового задания для SQL-запросов необходимо использовать синтаксис PostgreSQL или MS SQL, для скриптов – Python.

Рисунок 1 ER-диаграмма

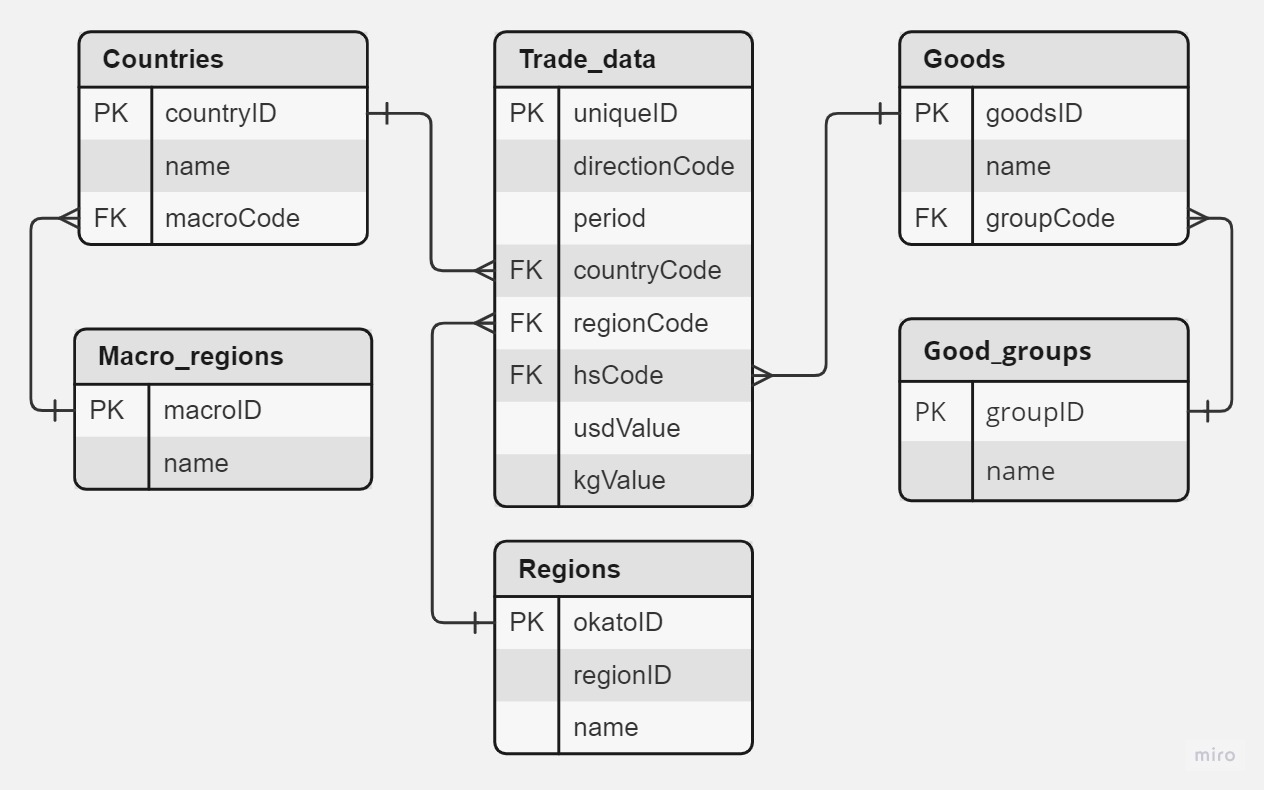


Таблица 1 Описание таблиц

| Идентификатор поля | Тип данных | NULL | Ключ | Описание |
| --- | --- | --- | --- | --- |
| **Таблица Trade\_data (данные торговли по месяцам)** | | | | |
| uniqueID | integer |  | PK | Автоинкремент |
| directionCode | char(2) |  |  | Код направления торговли (ex, im) |
| period | date |  |  | Период торговли (на 01 число каждого месяца, формат 01.MM.YYYY) |
| countryCode | char(2) |  | FK | Код страны |
| regionCode | char(5) |  | FK | Код ОКАТО субъекта РФ |
| hsCode | char(4) |  | FK | Код товара |
| usdValue | float | да |  | Стоимость в долларах США |
| kgValue | float | да |  | Вес в кг |
| **Таблица Countries** | | | | |
| countryID | char(2) |  | PK | Код страны (ИСО 3166-1 альфа-2) |
| name | char(255) |  |  | Наименование страны мира |
| macroCode | char(20) |  | FK | Код макрорегиона |
| **Таблица Macro\_regions** | | | | |
| macroID | char(20) |  | PK | Код макрорегиона |
| name | char(150) |  |  | Наименование макрорегиона |
| **Таблица Regions** | | | | |
| okatoID | char(5) |  | PK | Код ОКАТО субъекта РФ (5 знаков) |
| regionID | char(10) |  |  | Код субъекта РФ |
| name | char(150) |  |  | Наименование субъекта РФ |
| **Таблица Goods** | | | | |
| goodsID | char(4) |  | PK | Код товара (HS или ТНВЭД, 4 знака) |
| name | text |  |  | Наименование товара |
| groupCode | char(10) |  | FK | Код товарной группы |
| **Таблица Good\_groups** | | | | |
| groupID | char(10) |  | PK | Код товарной группы |
| name | char(255) |  |  | Наименование товарной группы |

Задача 1.

**Написать SQL-запрос** «Данные внешней торговли (экспорт и импорт) регионов России по годам, макрорегионам, группам товаров», используя ER-диаграмму (Рисунок 1) и описание таблиц (Таблица 1). Формат вывода: направление, дата (формат 01.01.YYYY), наименование субъекта РФ, наименование макрорегиона, наименование товарной группы, вес в тоннах.

*По желанию:* **написать SQL-запрос**, позволяющий выводить годовые значения по колонкам. Формат вывода: направление, наименование субъекта РФ, наименование макрорегиона, наименование товарной группы, далее идут колонки годов (YYYY), в которых выводятся значения веса в тоннах.

Задача 2.

С использованием *оконных функций* **написать SQL-запрос,** позволяющий рассчитать ежемесячную долю экспорта в 2021 году для субъекта РФ=MSK в общем объеме ежемесячного экспорта.

Задача 3.

Написать **Python-скрипт,** который позволяет распарсить (разобрать для последующей загрузки, например, в БД) справочник станций <https://cargo.rzd.ru/ru/9800>.

Задача 4.

Глоссарий

Модель = уравнение, зависимость выходной переменной от входных.

Моделируемая переменная\показатель – выходная переменная.

Факторы – входные переменные, предикторы.

**Написать Python-скрипт**, позволяющий рассчитывать модель линейной регрессии *для каждой товарной группы* (Таблица 2). Скрипт должен выводить:

* Значение статистики R2 (коэффициент детерминации);
* Значения коэффициентов уравнения и значение константы.

Данные для проведения расчетов прилагаются (файл «Данные\_моделирование.xlsx», данные в файле не логарифмированы). В файле:

* Лист «g» - данные экспорта в разрезе товарных групп;
* Лист «m» - данные показателей макрорегионов;
* Лист «rf» - данные показателей субъектов РФ.

Модель (уравнение) линейной регрессии вида:

ln(w\_export) = const + a1\*ln(factor\_1) + a2\*ln(factor\_2),

где w\_export – моделируемая переменная, значение (в тоннах) экспорта для товарной группы;

factor\_1, factor\_2 – фактор (предиктор);

const – константа;

a1, a2 – коэффициенты уравнения.

Таблица Факторы

| Товарная группа | Factor\_1 | Factor\_2 |
| --- | --- | --- |
| g01 | m\_industry | rf\_prod\_metal |
| g02 | m\_gdp | rf\_prod\_timber |
| g03 | m\_consum\_fertil | rf\_prod\_fertil |
| g04 | m\_import\_grains | rf\_prod\_grain |
| g05 | m\_import\_natoil | rf\_va\_agro |
| g06 | m\_import\_food | rf\_prod\_food |
| g07 | m\_industry | rf\_prod\_chem |
| g08 | m\_industry | rf\_prod\_rubber |
| g09 | m\_industry | rf\_prod\_paper |
| g10 | m\_industry | rf\_prod\_lightind |
| g11 | m\_industry | rf\_prod\_metal |
| g12 | m\_industry | rf\_prod\_metalgood |
| g13 | m\_industry | rf\_prod\_mach |
| g14 | m\_industry | rf\_prod\_transp |
| g15 | m\_consum\_coal | rf\_prod\_coal |
| g16 | m\_consum\_coke | rf\_prod\_coke |
| g17 | m\_consum\_oil | rf\_prod\_oil |
| g18 | m\_industry | rf\_prod\_build |
| g19 | m\_prod\_steel | rf\_prod\_ferrore |
| g20 | m\_industry | rf\_prod\_nonferrore |
| g21 | m\_consum\_gas | rf\_prod\_gas |
| g22 | m\_industry | rf\_va\_manuf |

По результатам выполненных работ Вы должны предоставить:

1. Файлы Python-скриптов и SQL-запросов;
2. Readme.txt по установке (требования к версиям пакетов Python, рекомендации по настройке для запуска скриптов).

*Оцениваем как отдельные компоненты результата, так и весь результат в комплексе.*