

Описание первичных таблиц данных

transaction_id
product_id
customer_id
transaction_date
online_order
order_status
brand
product_line
product_class
product_size
list_price
standard_cost

customer_id
first_name
last_name
gender
DOB
job_title
job_industry_category
wealth_segment
deceased_indicator
owns_car
address
postcode
state
country
property_valuation

Описание данных

- По данным и категориям нам представлены данные продаж магазина велосипедов за 2017 год
- Данные разбиты на 2 таблицы – Транзакции и Покупатели, которые проиндексированы по первым столбцам, в ключевых полях нет дублирований и исключений
- Таблица Транзакции имеет 20 тыс. позиций распределенных по всем месяцам 2017 года.
- В таблице Транзакции имеются пропуски данных 197 строк поенным товарам. Это составляет меньше 1% данных, можно и удалить, но в этих строках сохранены данные выручки, клиентов, дат и типов заказов, так что если анализ касается этих вопросов, то можно и сохранить.
- Также в таблице Транзакции имеются 360 пропусков в столбце типа заказа (онлайн-оффлайн), пропуски распределены по всему году и составляют меньше 2% общих данных. Существенно на анализ не влияют.
- Таблица Покупатели состоит из 4 тыс. позиций. Имеются значительные пропуски в столбце «Профессия» - 506 пропусков, «Индустрія» – 656 пропусков. У 87 не указан день рождения, у 1 указан не корректно (1843 год). У всех, у кого не указан день рождения в столбце «Пол» странный символ «U». Столбец «Пол» также требует корректировки, есть разные типы записей одного и того же пола.
- В таблице Транзакции в поле Покупатель есть покупатель с несуществующим кодом – 5034, необходимо удалить или добавить в таблицу Покупателей
- Ключевая проблема таблицы Транзакции – поле product_id. Проверка показала, что поля в нем не уникальны, во многих позициях находятся по несколько разных моделей велосипедов, в позиции 0 – вообще «свалка» данных. Принято решение сохранить его, но для корректной индексации с товарами сделать дополнительное поле bike_id для связки таблиц Транзакции и Продукты**

Обработка данных (1/2)

- Из анализа таблицы Транзакции видно, что в нем содержится поле product_id и зависимые от него поля (подкрасил в таблице золотым цветом)
- Такой формат противоречит правилам ЗНФ и для приведения в эту форму необходимо вынести эти данные в отдельную таблицу Продукты.

AS IS

transaction_id	customer_id
product_id	first_name
customer_id	last_name
transaction_date	gender
online_order	DOB
order_status	job_title
brand	job_industry_category
product_line	wealth_segment
product_class	deceased_indicator
product_size	owns_car
list_price	address
standard_cost	postcode
	state
	country
	property_valuation



TO BE

bike_id	transaction_id
product_id	product_id
brand	bike_id
product_line	customer_id
product_class	transaction_date
product_size	online_order
list_price	order_status
standard_cost	list_price

Выводы: итоговое представление из 3-х таблиц соответствует правилам ЗНФ

customer_id
first_name
last_name
gender
DOB
job_title
job_industry_category
wealth_segment
deceased_indicator
owns_car
address
postcode
state
country
property_valuation

Обработка данных (2/2)

product_id	transaction_id	customer_id
brand	product_id	first_name
product_line	customer_id	last_name
product_class	transaction_date	gender
product_size	online_order	DOB
standard_cost	order_status	job_title
	list_price	job_industry_category
		wealth_segment
		deceased_indicator
		owns_car
		address
		postcode
		state
		country
		property_valuation

Отклоненные гипотезы

- В ходе работы над структурой были отвергнуты ряд гипотез по преобразованию таблиц
- В таблице с Продуктами можно было выделить бренды или типы, но таблица итак лаконична и содержит всего 6 полей, которые обеспечивают формирование уникальных позиций
- Несколько перегруженной выглядит таблица Покупателей – 15 полей, и можно было вынести отдельно адреса, оставив в таблице лишь address_id. Но я проверил – все адреса уникальны (с учетом почтовых индексов) и относятся только к одному покупателю, а 15 столбцов вполне вмещаются на экран для оперативного просмотра и решил не выделять.

Работа в Dbeaver (план и код)

- Подготовил Эксель-таблицы для загрузки – разбил на 3 файла csv для загрузки в таблицы
- SQL кодом создал сами таблицы в Dbeaver
- Через import загрузил данные csv-файлов в таблицы
- Ниже скрины коды, на следующих слайдах скрины таблиц

```
create table transaction(
    transaction_id int unique not null primary key
    ,product_id int
    ,bike_id int
    ,customer_id int
    ,transaction_date date
    ,online_order bool
    ,order_status text
    ,list_price float
)

create table product(
    product_id int
    ,bike_id int primary key
    ,brand text
    ,product_line text
    ,product_class text
    ,product_size text
    ,standard_cost float
)

create table customer(
    customer_id int unique not null primary key
    ,first_name text
    ,last_name text
    ,gender text
    ,DOB date
    ,job_title text
    ,job_industry_category text
    ,wealth_segment text
    ,deceased_indicator text
    ,owns_car bool
    ,address text
    ,postcode int
    ,state text
    ,country text
    ,property_valuation int
)
```

Работа в Dbeaver (скрин таблицы transaction)

postgres Script product transaction X customer

Свойства Данные Диаграмма

SELECT t.* FROM public."transaction" A Введите SQL выражение чтобы отфильтровать результаты

Запись	transaction_id	product_id	customer_id	transaction_date	online_order
1	1	2	2 950	2017-02-25	[] A
2	2	3	3 120	2017-05-21	[v] A
3	3	37	402	2017-10-16	[] A
4	4	88	3 135	2017-08-31	[] A
5	5	78	787	2017-10-01	[v] A
6	6	25	2 339	2017-03-08	[v] A
7	7	22	1 542	2017-04-21	[v] A
8	8	15	2 459	2017-07-15	[] A
9	9	67	1 305	2017-08-10	[] A
10	10	12	3 262	2017-08-30	[v] A
11	11	5	1 986	2017-01-17	[] A
12	12	61	2 783	2017-01-05	[v] A
13	13	35	1 243	2017-02-26	[v] A
14	14	16	2 717	2017-09-10	[] A
15	15	12	247	2017-06-11	[] A
16	16	3	2 961	2017-10-10	[] A
17	17	79	2 426	2017-04-03	[] A
18	18	33	1 842	2017-06-02	[] A
19	19	54	2 268	2017-04-06	[v] A
20	20	25	3 002	2017-01-28	[v] A
21	21	27	1 582	2017-10-09	[] A
22	22	37	595	2017-06-29	[v] A

Обновить Save Cancel Экспорт данных ... 200 200+

200 строк получено - 0.0s (0.0s получ.), 2025-11-16 в 21:07:43

Работа в Dbeaver (скрин таблицы customer)

The screenshot shows the Dbeaver interface with the 'customer' table selected. The table has 22 rows and 7 columns. The columns are: customer_id, first_name, last_name, gender, dob, and job_title. The data includes various names, genders, and job titles, such as Laraine Medendorp (Executive Secretary), Eli Bockman (Administrative Officer), and Deeanne Durtnell (Safety Technician I). The table is sorted by customer_id.

	customer_id	first_name	last_name	gender	dob	job_title
1	1	Laraine	Medendorp	F	1953-10-12	Executive Secretary
2	2	Eli	Bockman	Male	1980-12-16	Administrative Officer
3	3	Arlin	Dearle	Male	1954-01-20	Recruiting Manager
4	4	Talbot		Male	1961-10-03	
5	5	Sheila-kathryn	Calton	Female	1977-05-13	Senior Editor
6	6	Curr	Duckhouse	Male	1966-09-16	
7	7	Fina	Merali	Female	1976-02-23	
8	8	Rod	Inder	Male	1962-03-30	Media Manager I
9	9	Mala	Lind	Female	1973-03-10	Business Systems Devel
10	10	Fiorenze	Birdall	Female	1988-10-11	Senior Quality Engineer
11	11	Uriah	Bisatt	Male	1954-04-30	
12	12	Sawyere	Flattman	Male	1994-07-21	Nuclear Power Enginee
13	13	Gabriele	Norcross	Male	1955-02-15	Developer I
14	14	Rayshell	Kitteman	Female	1983-03-25	Account Executive
15	15	Erroll	Radage	Male	2000-07-13	Junior Executive
16	16	Harlin	Parr	Male	1977-02-27	Media Manager IV
17	17	Heath	Faraday	Male	1962-03-19	Sales Associate
18	18	Marjie	Neasham	Female	1967-07-06	Professor
19	19	Sorcha	Keyson	Female	2001-04-15	Geological Engineer
20	20	Basile	Firth	Male	1980-08-13	Project Manager
21	21	Mile	Cammocke	Male	1980-09-20	Safety Technician I
22	22	Deeanne	Durtnell	Female	1962-12-10	

Работа в Dbeaver (скрин таблицы product)

The screenshot shows the Dbeaver interface with the PostgreSQL database selected. The current table is 'product'. The 'Данные' (Data) tab is active, displaying the contents of the 'product' table. The table has 22 rows and 7 columns. The columns are: id, brand, product_line, product_class, size, and standard. The data includes various bicycle brands like Solex, Trek Bicycles, OHM Cycles, Norco Bicycles, Giant Bicycles, and WeareA2B, categorized by product line (Standard, Touring) and size (medium, large, low). The 'Текст' (Text) tab is also visible on the left.

	brand	product_line	product_class	size	standard
1	Solex	Standard	medium	medium	
2	Trek Bicycles	Standard	medium	large	
3	OHM Cycles	Standard	low	medium	
4	Norco Bicycles	Standard	medium	medium	
5	Giant Bicycles	Standard	medium	large	
6	Giant Bicycles	Road	medium	medium	
7	WeareA2B	Standard	medium	medium	
8	WeareA2B	Standard	medium	medium	
9	Solex	Standard	medium	large	
10	WeareA2B	Standard	medium	medium	
11	Trek Bicycles	Mountain	low	medium	
12	OHM Cycles	Standard	low	medium	
13	Trek Bicycles	Standard	low	medium	
14	Norco Bicycles	Standard	high	small	
15	Norco Bicycles	Standard	medium	medium	
16	Giant Bicycles	Standard	medium	small	
17	WeareA2B	Standard	medium	medium	
18	Trek Bicycles	Standard	medium	medium	
19	Giant Bicycles	Road	medium	medium	
20	WeareA2B	Touring	medium	large	
21	Trek Bicycles	Standard	medium	large	
22	Trek Bicycles	Mountain	low	medium	