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. позиции распределенных по всем месяцам 2017 года.
- В таблице Транзакции имеются пропуски данных 197 строк по данным товаров. 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-офлайн), пропуски распределены по всему году и составляют меньше 2% общих данных. Существенно на анализ не повлияют.
- Таблица Покупатели состоит из 4 тыс. позиций. Имеются значительные пропуски в столбце «Профессия» - 506 пропусков, «Индустрия» – 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)

1. Из анализа таблицы Транзации видно, что в нем содержится поле product_id и зависимые от него поля (подкрасил в таблице золотым цветом)
2. Такой формат противоречит правилам 3НФ и для приведения в эту форму необходимо вынести эти данные в отдельную таблицу Продукты.

AS IS

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

TO BE

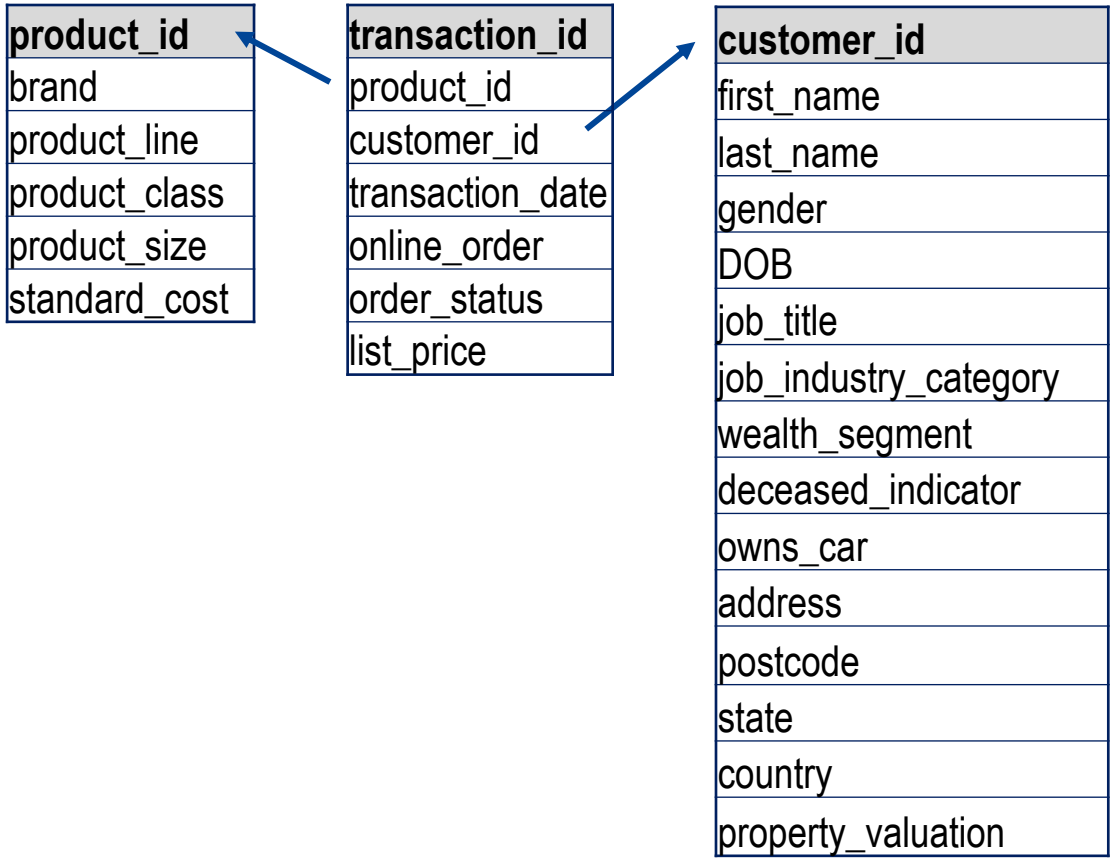
bike_id
product_id
brand
product_line
product_class
product_size
standard_cost

transaction_id
product_id
bike_id
customer_id
transaction_date
online_order
order_status
list_price

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

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

Обработка данных (2/2)



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)

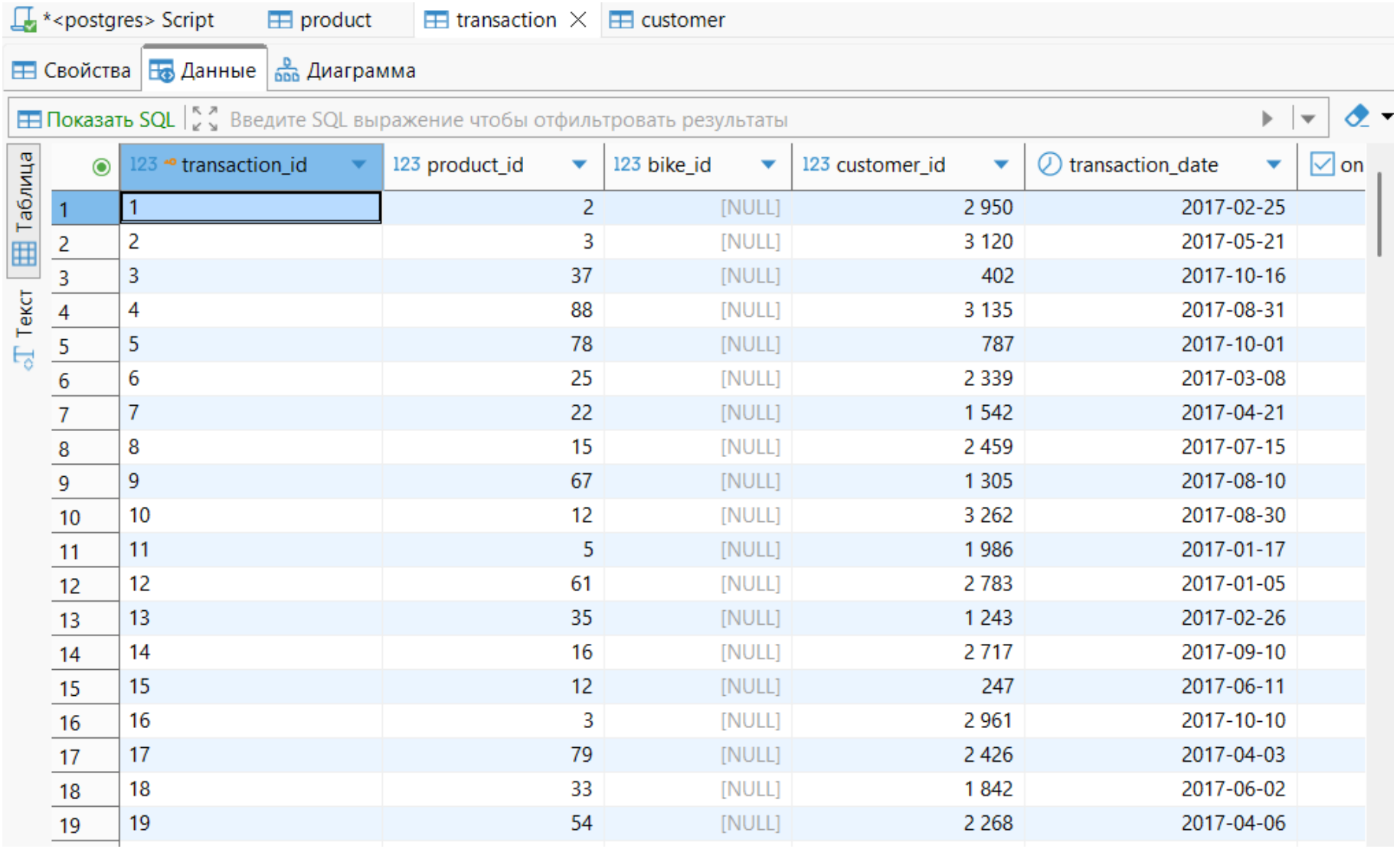
- Подготовил Эксель-таблицы для загрузки – разбил на 3 файла csv для загрузки в таблицы
- SQL кодом создал сами таблицы в Dbeaver
- Через import загрузил данные csv-файлов в таблицы
- Ниже скрины коды, на следующих слайдах скрины таблиц

```
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 customer(
    customer_id int unique not null primary key
    ,first_name textc
    ,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
    )
```

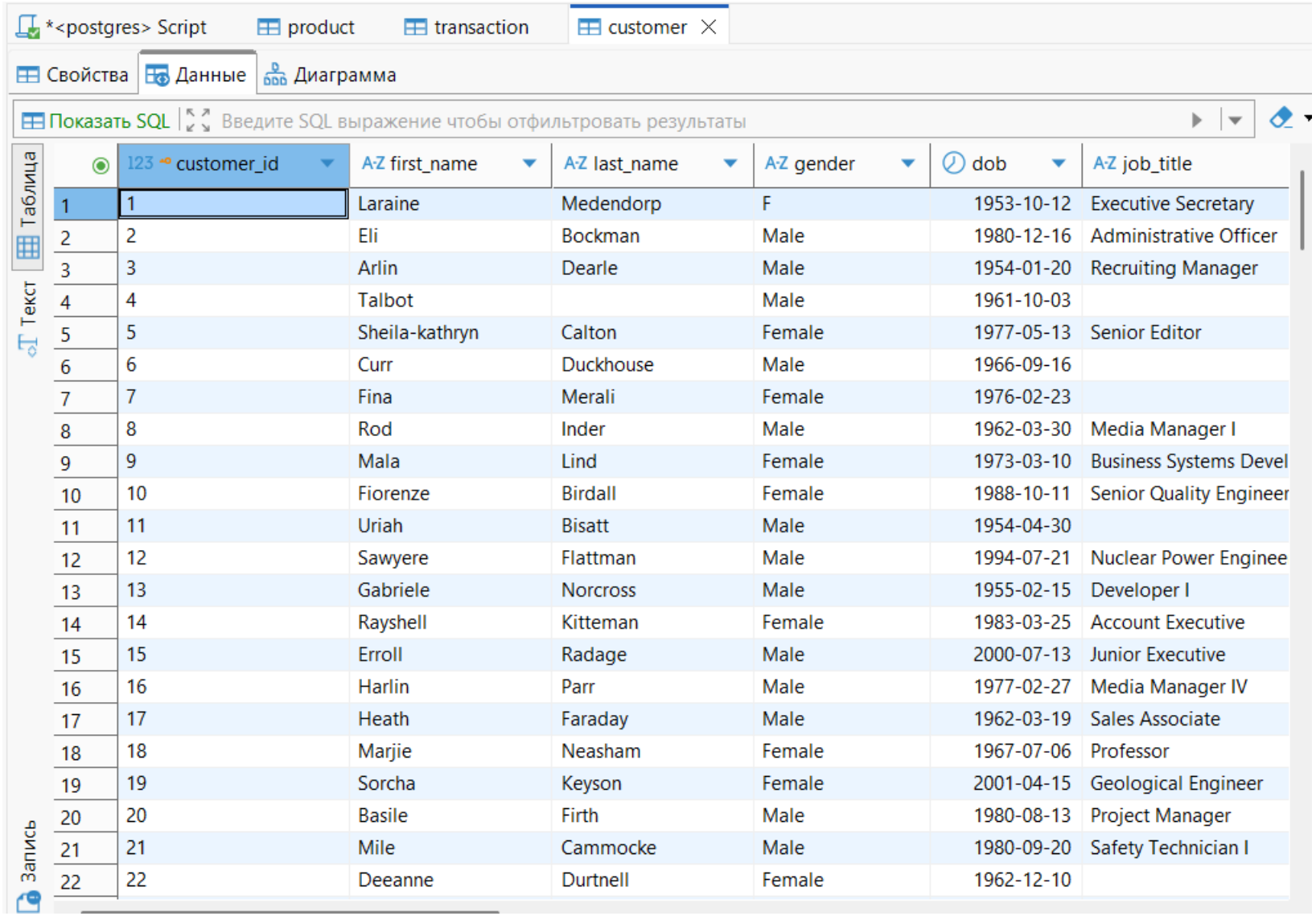
```
create table product(
product_id int
,bike_id int primary key
,brand text
,product_line text
,product_class text
,product_size text
,standard_cost float
)
```

Работа в Dbeaver (скрин таблицы transaction)



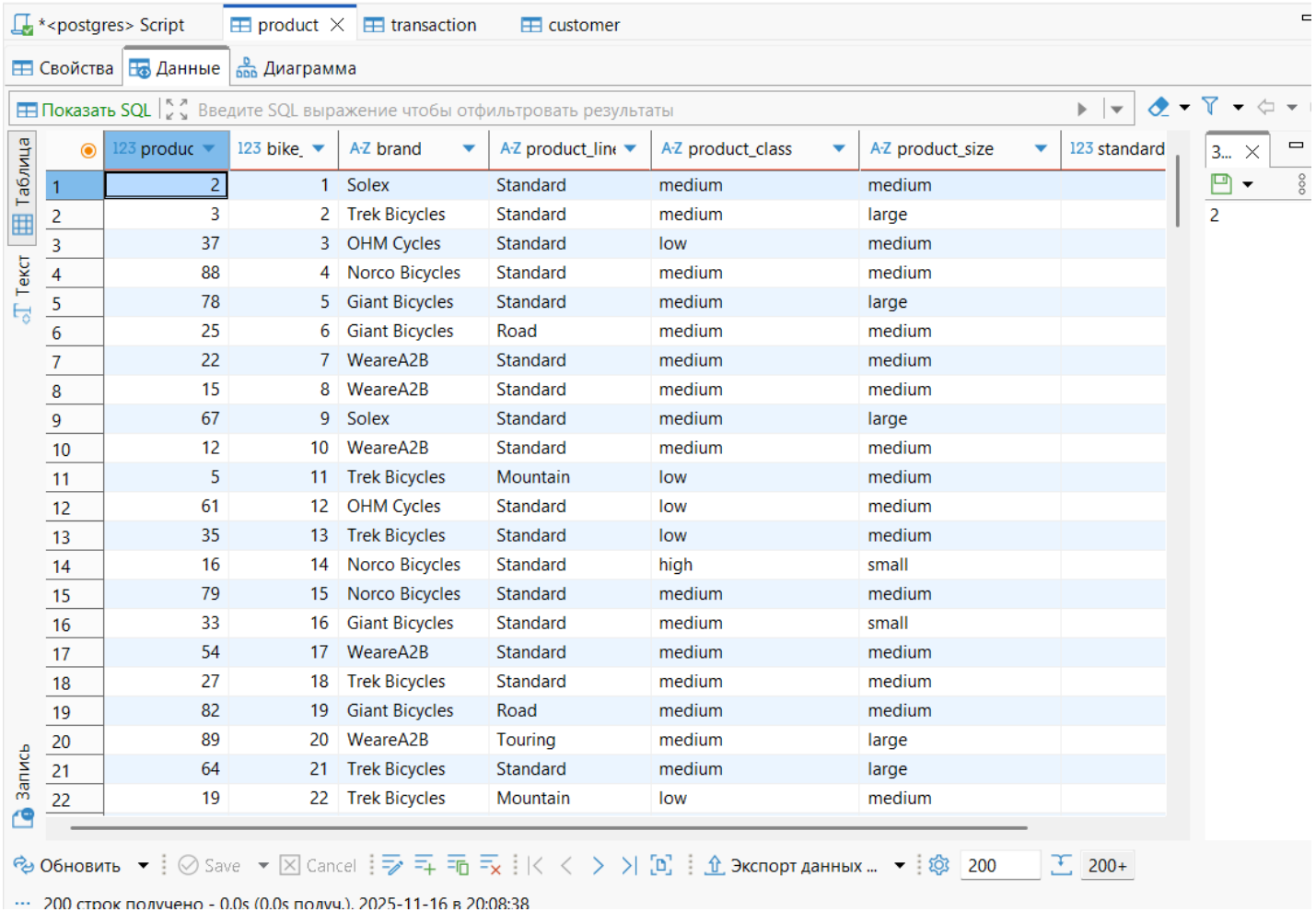
	transaction_id	product_id	bike_id	customer_id	transaction_date	on
1	1	2	[NULL]	2 950	2017-02-25	
2	2	3	[NULL]	3 120	2017-05-21	
3	3	37	[NULL]	402	2017-10-16	
4	4	88	[NULL]	3 135	2017-08-31	
5	5	78	[NULL]	787	2017-10-01	
6	6	25	[NULL]	2 339	2017-03-08	
7	7	22	[NULL]	1 542	2017-04-21	
8	8	15	[NULL]	2 459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

Работа в Dbeaver (скрин таблицы customer)



	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)



	123 produc	123 bike_	A-Z brand	A-Z product_lin	A-Z product_class	A-Z product_size	123 standard
1	2	1	Solex	Standard	medium	medium	
2	3	2	Trek Bicycles	Standard	medium	large	
3	37	3	OHM Cycles	Standard	low	medium	
4	88	4	Norco Bicycles	Standard	medium	medium	
5	78	5	Giant Bicycles	Standard	medium	large	
6	25	6	Giant Bicycles	Road	medium	medium	
7	22	7	WeareA2B	Standard	medium	medium	
8	15	8	WeareA2B	Standard	medium	medium	
9	67	9	Solex	Standard	medium	large	
10	12	10	WeareA2B	Standard	medium	medium	
11	5	11	Trek Bicycles	Mountain	low	medium	
12	61	12	OHM Cycles	Standard	low	medium	
13	35	13	Trek Bicycles	Standard	low	medium	
14	16	14	Norco Bicycles	Standard	high	small	
15	79	15	Norco Bicycles	Standard	medium	medium	
16	33	16	Giant Bicycles	Standard	medium	small	
17	54	17	WeareA2B	Standard	medium	medium	
18	27	18	Trek Bicycles	Standard	medium	medium	
19	82	19	Giant Bicycles	Road	medium	medium	
20	89	20	WeareA2B	Touring	medium	large	
21	64	21	Trek Bicycles	Standard	medium	large	
22	19	22	Trek Bicycles	Mountain	low	medium	