Таблица	Атрибут	Данные нужны для проекта?	На что проверяем	Проверка	Результат	Описание проверки	Рекомендации	примечание
brand	brand_id, brand	да	Пропуски	с помощью фильтра СУБД	нет ошибки	используя простой фильтр is null		
brand	brand_id, brand	да	Дубли	select brand_id, brand, count(*) from brand group by brand_id, brand having count (*) > 1;	есть ошибка	поиск строк, которые повторяются более одного раза	удалить полные дубли	создать скрипты на исправление
brand	brand_id, brand	да	Соответствие типов	select brand_id, brand from brand where not brand_id is null and not brand_id ~ '^[0-9]+$';	есть ошибка	проверка, что тип brand_id соответствует int	поменять местами данные в brand_id и brand	создать скрипты на исправление
brand	brand_id, brand	да	Соответствие типов	select brand_id from brand where LENGTH(brand) >= 255;	нет ошибки	проверка, что все строки длиной менее 255 символов		
brand	brand_id, brand	да	Неинформативные данные	select brand_id, brand from brand where LENGTH(brand) <= 2;	есть ошибка	поиск коротких слов и знаков	сомнительные названия бренда уточнить у Заказчика	уточнили у Заказчика, оставляем как есть
category	category_id, category_name	да	Пропуски	с помощью фильтра СУБД	нет ошибки	используя простой фильтр is null		
category	category_id, category_name	да	Дубли	select category_id, category_name, count(*) from category group by category_id, category_name having count (*) > 1;	есть ошибка	поиск строк, которые повторяются более одного раза	удалить полные дубли	создать скрипты на исправление
category	category_id, category_name	да	Соответствие типов	select category_id from category where LENGTH(category_id) >= 50 or LENGTH(category_name) >= 255;	нет ошибки	проверка, что все строки длиной менее 255 символов		
category	category_id, category_name	да	Неинформативные данные	select * from category where category_name ~ '[a-zA-Z]';	есть ошибка	поиск латинских символов в названии категории	уточнить у Заказчика	уточнили у Заказчика, сотавить как есть

Таблица	Атрибут	Данные нужны для проекта?	На что проверяем	Проверка	Результат	Описание проверки	Рекомендации	примечание
category	category_id, category_name	да	Неинформативные данные	select category.category_name, brand.brand from category left join product on category. category_id = product.category_id left join brand on product.brand_id = brand.brand_id group by category.category_name, brand.brand having category.category_name ~ '[a-zA-Z]';	есть ошибка	сопоставление категорий с латинскими символами с названием бренда	здесь ошибки внешних ключей, разобраны ниже	
category	category_id, category_name	да	Неинформативные данные	select category_name from category where category_name like '%!%' or category_name like '%?%';	есть ошибка	поиск символов ! и ? в названии категории	выгрузив клиенту список продуктов данной категории - уточнить, к какой категории они относятся	уточнили у Заказчика, оставить как есть
product	все атрибуты	да	Пропуски	с помощью фильтра СУБД	нет ошибки	используя простой фильтр is null		
product	все атрибуты	да	Дубли	select product_id, name_short, category_id, pricing_line_id, brand_id, count(*) from product group by product_id, name_short, category_id, pricing_line_id, brand_id having count (*) > 1;	есть ошибки	поиск строк, которые повторяются более одного раза	удалить полные дубли	создать скрипты на исправление
product	product_id	да	Соответствие типов	select product_id, name_short from product where not product_id is null and not product_id ~ '^[0-9]+$';	нет ошибки	поиск строк, которые не integer		
product	name_short	да	Соответствие типов	select name_short from product where LENGTH(name_short) >= 255;	нет ошибки	поиск строк длиннее 255 символов		
product	name_short	да	Неинформативные данные	select name_short from product where name_short like '%!%' or name_short like '%?%' or LENGTH (name_short)<5 or name_short like '%не %';	есть ошибка	поиск строк, содержащих !, ? или "не", а также короче 5 символов	по строке, где написано "не завели" - уточнить у Заказчика, остальное не править, т.к. для целей проекта не принципиально	уточнили у Заказчика, оставить как есть

Таблица	Атрибут	Данные нужны для проекта?	На что проверяем	Проверка	Результат	Описание проверки	Рекомендации	примечание
product	category_id	да	Ссылочная целостность	select category.category_id, product.product_id, product.category_id from category right join product on category.category_id = product.category_id where category.category_id is null ;	есть ошибка	сопоставление category_id из product и category_id из category	в результате проверки выяснилось, что для некотоорых product_id нет соответствующих category_id, при проверке одной из позиций, выяснилось, что в category id было A16, а в product A166, то есть очень много строк с такими нестыковками	создать скрипт на исправление, где в category_id добавить A166 и category_name "не определено"
product	category_id	да	Ссылочная целостность	select category_id from product where LENGTH(category_id) >= 4;	есть ошибка	поиск строк длиннее 4 символов		
product	pricing_line_id	нет						
product	brand_id	да	Ссылочная целостность	select brand.brand_id, product.product_id from brand right join product on brand.brand_id = product.brand_id where brand.brand_id is null ;	нет ошибки	сопоставление brand_id из product и brand_id из brand		
stock	все атрибуты	да	Пропуски	с помощью фильтра СУБД	нет ошибки	используя простой фильтр is null		
stock	cost_per_item	нет	Пропуски	select product_id, cost_per_item from stock where cost_per_item = '';	есть ошибка	поиск пропусков в формате ''	есть пропуски в формате ' ', уточнить у клиента, анализируем ли мы рентабельность в проекте, валовую прибыль, если нет - то эти данные не нужны	уточнили у Заказчика, создать скрипт, заполнить пропуски "null"
stock	product_id	да	Пропуски	select product_id, cost_per_item from stock where product_id = '';	есть ошибка	поиск пропусков в формате ''	есть пропуски в формате ' ', необходимо уточнить у клиента, каким образом выгружаются данные в stock	уточнили у Заказчика, создать скрипт, заполнить пропуски product_id "словом "не определен"

Таблица	Атрибут	Данные нужны для проекта?	На что проверяем	Проверка	Результат	Описание проверки	Рекомендации	примечание
stock	все атрибуты	да	Дубли	select available_on, product_id , pos, available_quantity , cost_per_item , count(*) from stock group by available_on, product_id , pos, available_quantity , cost_per_item having count (*) > 1;	есть ошибка	поиск строк, которые повторяются более одного раза	удалить полные дубли	создать скрипты на исправление
stock	available_on	да	Соответствие ти	select available_on from stock where available_on is not null and TO_DATE(available_on, 'YYYY-MM-DD') is null;	нет ошибки	поиск строк, котрые не могут быть превращены в дату		
stock	product_id	да	Ссылочная цело	select product.product_id, stock.product_id from stock left join product on stock.product_id = product.product_id where product.product_id is null ;	есть ошибка	сопоставление product_id из product и product_id из stock	необходимо уточнить у клиента, каким образом выгружаются данные в stock, почему не корректно отоброжается product_id	уточнили у Заказчика, добавить product_id, которые есть в stock - в product
stock	pos	да	Неинформативн	select pos from stock where pos not in ('Магазин 1', 'Магазин 2', 'Магазин 3', 'Магазин 4', 'Магазин 5', 'Магазин 6', 'Магазин 7', 'Магазин 8', 'Магазин 9', 'Магазин 10');	есть ошибка	поиск символов которые не соответсвуют слову Магазин и номер от 1 до 10	Магазин 999, исправить на Магазин 9	уточнили у Заказчика, не исправлять
stock	available_quantit	да	Соответствие ти	select available_quantity from stock where not (available_quantity::text ~ '^[-+]?[0-9]*\.?[0-9]+([eE][-+]?[0-9]+)?$');	нет ошибки	поиск символов, которые не могут быть переведены во float		

Таблица	Атрибут	Данные нужны для проекта?	На что проверяем	Проверка	Результат	Описание проверки	Рекомендации	примечание
stock	available_quantit	да	выбросы	select MIN(cast(available_quantity as FLOAT)) as min_quantity, MAX(cast(available_quantity as FLOAT)) as max_quantity, PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY cast (available_quantity as FLOAT)) as median_quantity from stock where available_quantity <> '';	есть ошибка	выведение минимального, максимального значения и медиана available-quantity для обнаружения выбросов	есть значения ниже 0, это ошибка, запросить у Заказчика реальные остатки	уточнили у Заказчика, проставить null
				select product.name_short, stock.available_quantity, stock.cost_per_item from stock inner join product on stock.product_id = product.product_id where cast(available_quantity as FLOAT) > 15000 and available_quantity <> '';	нет ошибки	запрос на значения больше 15000 для анализа		
stock	cost_per_item	да	выбросы	select MIN(cast(cost_per_item as FLOAT)) as min_cost, MAX(cast(cost_per_item as FLOAT)) as max_cost, PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY cast (cost_per_item as FLOAT)) as median_cost from sources.stock s where cost_per_item <> '';	нет ошибки	простой вывод минимума, максимума и медианы	после просмотра максимальных значений (где явно очень большое отклонение от медианы) - убедились, что данные адекватны	

Таблица	Атрибут	Данные нужны для проекта?	На что проверяем	Проверка	Результат	Описание проверки	Рекомендации	примечание

stock	cost_per_item	да	выбросы	WITH q1 AS (
				  SELECT percentile_cont(0.25)
				WITHIN GROUP (ORDER BY
				NULLIF(cost_per_item, '')::numeric)
				AS q1,
				  percentile_cont(0.75) WITHIN
				GROUP (ORDER BY NULLIF
				(cost_per_item, '')::numeric) AS q3
				  FROM sources.stock
				  WHERE cost_per_item != ''
				),
				iqr AS (
				  SELECT q1, q3, q3 - q1 AS iqr
				  FROM q1
				)
				SELECT *
				FROM sources.stock
				WHERE cost_per_item != '' AND (
				  NULLIF(cost_per_item, '')::numeric
				< (SELECT q1 - 1.5 * iqr FROM iqr)
				  OR NULLIF(cost_per_item, '')::
				numeric > (SELECT q3 + 1.5 * iqr
				FROM iqr)
				);
					нет ошибки	выбросы, интерквартильный размах	слишком много таких значений, интерквартильный размах не подходит	

Таблица	Атрибут	Данные нужны для проекта?	На что проверяем	Проверка	Результат	Описание проверки	Рекомендации	примечание

stock	cost_per_item	да	выбросы	WITH q1 AS (
				  SELECT PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY NULLIF(cost_per_item, '')::numeric) AS median_value
				  FROM sources.stock
				  WHERE cost_per_item != ''
				),
				deviation AS (
				  SELECT median_value,
				  ABS(NULLIF(cost_per_item, '')::numeric - median_value) AS deviation
				  FROM sources.stock, q1
				  WHERE cost_per_item != ''
				),
				MAD as (
				SELECT PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER by deviation) as MAD
				from deviation)
				SELECT median_value, deviation, MAD
				FROM deviation, MAD, sources.stock
				WHERE cost_per_item != ''
				;
					нет ошибки	выбросы, абсолютное медианное отклонение	медина 98, mad = 79.8	

stock	cost_per_item	да	выбросы	select * from sources.stock where cast(cost_per_item as float) > 4*79.8 and cost_per_item <> '';	нет ошибки	выбросы, абсолютное медианное отклонение	слишком много таких значений, отклонение по медиане не подходит	
stores	pos	да			нет ошибки	проверка файла csv		
stores	pos_name	да			нет ошибки	проверка файла csv		
transaction	все атрибуты	да	Пропуски	с помощью фильтра СУБД	нет ошибки	используя простой фильтр is null		

Таблица	Атрибут	Данные нужны для проекта?	На что проверяем	Проверка	Результат	Описание проверки	Рекомендации	примечание
transaction	все атрибуты	да	Пропуски	select * from sources."transaction" where transaction_id = '' or product_id = '' or recorded_on = '' or quantity = '' or price = '' or price_full = '' or order_type_id = '';	есть ошибки	поиск символов '' в строках	выявлены пустые значения в price и quantity, эту информацию необходимо запросить у Заказчика (где price возможно просто не было скидки, а где quantity информацию запросить)	уточнили у Заказчика, пропуски заполнить null
transaction	все атрибуты	да	Дубли	select transaction_id , product_id , recorded_on , quantity , price , price_full , order_type_id , count(*) from sources."transaction" group by transaction_id , product_id , recorded_on , quantity , price , price_full , order_type_id having count (*) > 1;	нет ошибки	поиск строк, которые повторяются более одного раза		
transaction	transaction_id	да	Соответствие ти	select transaction_id from sources.transaction where LENGTH(transaction_id) >= 50;	нет ошибки	поиск символов длиннее 50		
transaction	product_id	да	Ссылочная цело	select sources.product.product_id, sources."transaction".product_id from sources.transaction left join sources.product on sources."transaction".product_id = product.product_id where product.product_id is null ;	есть ошибки	сопоставление product_id из product и product_id из transaction	в таблице transaction присутствуют такие product_id, которых нет в родительской таблице product, запрашивать информацию у Заказчика	уточнили у Заказчика, заполнить в родительской таблице product_id
transaction	recorded_on	да	Соответствие ти	select recorded_on from sources.transaction where recorded_on is not null and TO_TIMESTAMP(recorded_on, 'DD.MM.YYYY HH24:MI') is null;	нет ошибки	поиск строк, котрые не могут быть превращены в дату		

Таблица	Атрибут	Данные нужны для проекта?	На что проверяем	Проверка	Результат	Описание проверки	Рекомендации	примечание
transaction	quantity	да	выбросы	select MIN(cast(quantity as FLOAT)) as min_quantity, MAX(cast(quantity as FLOAT)) as max_quantity, PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY cast (quantity as FLOAT)) as median_quantity from sources."transaction" t where quantity <> '';	нет ошибки	выведение минимального, максимального значения и медиана quantity для обнаружения выбросов		
transaction	quantity	да	Соответствие ти	select quantity from sources."transaction" t where not (quantity::text ~ '^[-+]?[0-9]*\.?[0-9]+([eE][-+]?[0-9]+)?$');	нет ошибки	поиск символов, которые не могут быть приведены к типу float		
transaction	price	да	Соответствие ти	select price from sources."transaction" where not (price ::text ~ '^[-+]?[0-9]*\.?[0-9]+([eE][-+]?[0-9]+)?$');	нет ошибки	поиск символов, которые не могут быть приведены к типу float		
transaction	price	да	выбросы	select MIN(cast(price as FLOAT)) as min_price, MAX(cast(price as FLOAT)) as max_price, PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY cast (price as FLOAT)) as median_price from sources."transaction" t where price <> '';	нет ошибки	выведение минимального, максимального значения и медиана price для обнаружения выбросов	проверяю ниже цену больше 20000 что это	

Таблица	Атрибут	Данные нужны для проекта?	На что проверяем	Проверка	Результат	Описание проверки	Рекомендации	примечание

transaction	price	да	выбросы	select
				  sources.transaction.price, product.
				product_id, name_short
				from
				  sources."transaction"
				inner join sources.product on
				product.product_id = sources.
				transaction.product_id
				where
				  cast(price as FLOAT) > 10000 and
				price <> '';
					нет ошибки	запрос на значения больше 10000 для анализа	все выглядит адекватно	

transaction	price	да	согласованность	select price, price_full from sources."transaction" t where cast(price as float) > cast (price_full as float) and price <> '';	нет ошибки	поиск строк, где price было бы больше price_full		

transaction	price_full	да	Соответствие ти	select
				  price_full
				from
				  sources."transaction"
				where
				  not (price_full ::text ~ '^[-+]?[0-9]
				*\.?[0-9]+([eЕ][-+]?[0-9]+)?$');
					нет ошибки	поиск символов, которые не могут быть приведены к типу float		

transaction	price_full	да	выбросы	select
				  MIN(cast(price_full as FLOAT)) as
				min_price,
				  MAX(cast(price_full as FLOAT)) as
				max_price,
				  PERCENTILE_CONT(0.5)
				WITHIN GROUP (ORDER BY cast
				(price_full as FLOAT)) as
				median_price
				from
				  sources."transaction" t
				where
				  price_full <> '';
					нет ошибки	выведение минимального, максимального значения и медиана price_full для обнаружения выбросов		

transaction	order_type_id	нет						