Coding test for Data Engineer

The objective of this coding test is to get the insights from the data. We'd like you implement the test using Spark or MapReduce APIs in any programming language(Java, Scala, Python...) of your choice. Datasets are available in the data directory. And the schema is as following :

Product (product_id, product_name, product_type, product_version, product_price)
Customer (customer_id, customer_first_name, customer_last_name, phone_number)
Sales (transaction_id, customer_id, product_id, timestamp, total_amount, total_quantity)
Refund (refund_id, original_transaction_id, customer_id, product_id, timestamp, refund_amount, refund_quantity)

Questions :

1. Write down the data quality issues with the datasets provided and steps performed to clean (if any).
2. Display the distribution of sales by product name and product type.
3. Calculate the total amount of all transactions that happened in year 2013 and have not been refunded as of today.
4. Display the customer name who made the second most purchases in the month of May 2013. Refunds should be excluded.
5. Find a product that has not been sold at least once (if any).
6. Calculate the total number of users who purchased the same product consecutively at least 2 times on a given day.

Extra question (Please skip this if you don't have enough time)

Dataset for Customer_Extended is in the data directory. And schema is as following :

Customer_Extended (id, first_name, last_name, home_phone, mobile_phone, gender, current_street_address, current_city, current_state, current_country, current_zip, permanent_street_address, permanent_city, permanent_state, permanent_country, permanent_zip, office_street, office_city, office_state, office_country, office_zip, personal_email_address, work_email_address, twitter_id, facebook_id, linkedin_id)

Question :

1. Display all the details of a customer who is currently living at 1154 Winters Blvd.

Please make sure to include the readMe file with clear instructions when you return the coding test.