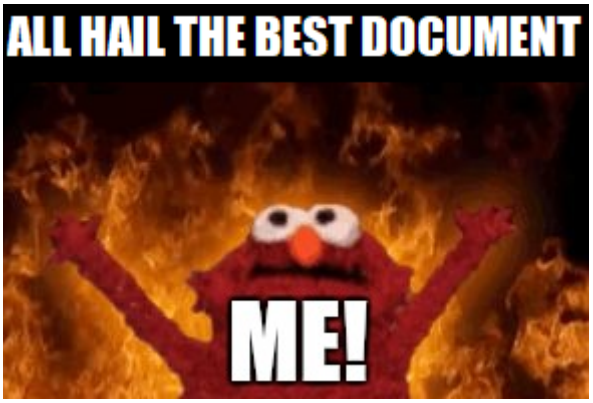
Hey there :)
Glad you made it here, to read me, a **magnificent** document. I contain the amazingly accurate answer to life, the universe and everything all the info you need to succeed in this step!



So, *let's talk business*.
What I'm here to do is guide you through a practical test that those amazing fellows at Booster Box would like to see with you.

What are we talking about?
There are several elements that we should be seeing together - three above the rest:

- ETL Processes
- Analyzing (Large) Datasets
- Sharing results and insights both for peers and executives

And how are we going to check for all of this? Well, you little pumpkin, that's why I exist in the first place.
I present to you... The Dataset!

What are we looking at?
The Dataset is a small extract of a CRM... dataset (you don't say), for a total of 300.000 rows and 18 columns (no worries, no personal data/PII in there of course).

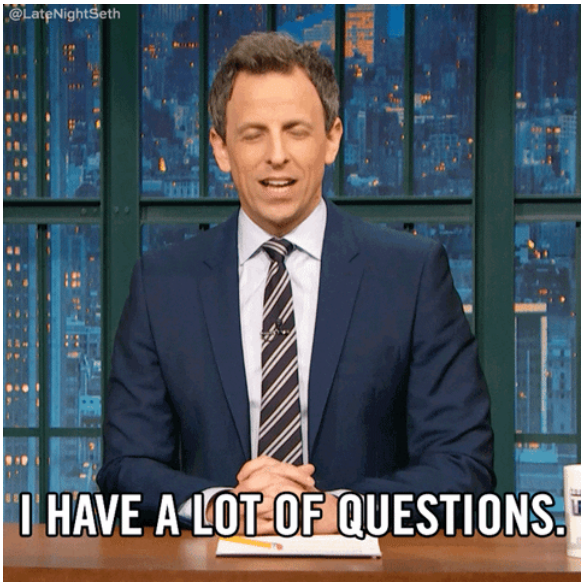
I, amazing document, contain a likewise amazing table summarizing the content of The Dataset:

Field name	Type	Description
email	STRING	Uhm, e-mails
phone	STRING	Uhm, phone numbers
first_name	STRING	The name before the last
last_name	STRING	The name after the first
zip	STRING	Not from the trousers! A Postal Code
city	STRING	Not NY, smaller ones usually
state	STRING	The State of users...
country	STRING	… and the Country!
birthday_year	STRING	Year of birth of our beloved customers
gender	STRING	Gender
age	STRING	Age
conversion_name	STRING	Now this is more interesting! This can either be: ● Purchase (a simple transaction) ● Return (when the product is sent back) ● Cancel (when the purchase is nulled)
conversion_date	STRING	Date of any of the aforementioned actions
conversion_id	STRING	ID of each action
conversion_value	FLOAT	This is the monetary value of the transaction
conversion_value_margin	FLOAT	This is the actual margin on each transaction
handling_cost	FLOAT	This is the cost associated with managing each order
conversion_currency	STRING	Currency of each transaction
ad_click_id_value	STRING	GCLID

To better clarify this: each row represents a transaction, from a specific user in a specific day.
The conversion can be a normal Purchase, or either a Returned product or a Cancelled order.
In the case of a Purchase, conversion_value and conversion_value_margin will be positive values (since they refer to money we are making), while handling_cost will be a negative value (as it refers to a cost); in the case of a Return or Cancel, the values will be inverted as to balance the purchase out (hence negative conversion values and positive handling costs).

One small note: conversion_value_margin does **not** take into account handling_costs.

Questions Time!



1) Pre-Analyses
- Which actions would you take before starting any type of analysis?
- What do we want to make sure of?
- Which data/info do you think could enrich this dataset?

2) Analyses
- Could you please develop a **RFM Analysis**? (Recency, Frequency, Monetary)
    - Can you think of any other way to group and/or profile the customers?
- Can you think of any other analysis which would be useful with this data?

3) Delivery
- We now would like for you to develop two presentations:
    - One, very focused and ugly, about **how** you ran your analyses and the logic behind your choices. More specifically, we'd also like to see:
        - Any code you've written which you deem relevant
        - Any query you've written which you deem relevant
        - Any filter you have applied / any data you have discarded (and the reasons)
        - Anything else which you might see relevant
    - One instead aimed at delivering your insights and suggesting related Action Items to the final Decision Maker. Consider she is not coming from a technical background.

We expect you to explain how you tailored your preso for the target.

This gorgeous document has thus completed its job.

Disclaimer: 42 is not momentarily accepted as an answer to any of the questions