Questions to cover for the Data Engineer Screening Assignment

1. What type of process is the code supporting?

The basic idea here is to get information about the different brands and save their information after applying some filters.

1. Kindly walk us through the code – main functions, objects, general functionality.

The function “get\_segment\_info\_as\_df” calls an external API 2 times, first call returns a list of tuples of segments. We then iterate over this list and call the second API and pass a single segment tuple. This second API returns a detailed information about a particular information such as id, count and name. Finally, we save these information in a tuple which we created earlier and append this tuple in a list.

In the second function, it first filters out the brands which are from today. Then, it iterates over the segment passed as a parameter in the function. For each segment, It creates a new ContactSegment object and initializes it with the attributes for a particular segment. Also, the current date is assigned. This object is appended to the list. Finally, we save information for each segment.

In main, each brand is assigned a particular key. This key is used to reference the brand.

1. What is exactly the purpose of line 19? Could the code function without it? What would be a drawback of omitting that line?

Line 19 checks if the brand is present in the EMARSYS LOGINS.

In the current code, line 19 can be removed since this method is called from main in which we iterate over the EMARSYS LOGINS keys. So we can be sure that the value passed to this function is already present in EMARSYS LOGINS

One drawback of omitting this line is that if this function is called from any other piece of code, then there is no way of checking if the passed parameter is present in the EMARSYS LOGINS.

1. In which line(s) is an API request sent?

Line 23 & 26

1. What information are we saving from the data source? Please name the exact variables.

We are saving the id, name and count

1. In which line(s) do we send the retrieved data to the database? Are we writing the same variables that we scraped? If not, which is (are) added and based on your experience, why?

On line 26, we send the retrieved data to the database.

We are **not** writing the same variable that we scraped. We add 3 new variables id, count and name.

Based on experience, it is a good idea to decouple different business logics and to make the code more independent. Also, one advantage here is that the first API call at line 23 gives us all the possible values for segments. So we can iterate on them directly.

1. Could you please explain the syntax on line 33? What are the differences between the line 33 and line 18?

In line 33, The following parameters are passed:

1. segment\_info: A parameter with type List. This List can only contain values for namedtuple brand\_shop\_segment which we define on line 15
2. Brand: A parameter with string as the datatype

None means that the function does not return anything

In line 18, The following parameter is passed:

1. Brand: A parameter with string as the datatype

The return type is List[brand\_shop\_segment] which means that this function will return a list where each element of list contains a tuple we defined on line 15

The difference between line 18 and 33 is that line 18 takes 1 parameter and returns 1 output while line 33 takes 2 parameters and returns nothing.

1. **BONUS:** Why do you think we need a process like that? What value does it provide? Think bigger, not a technical question. *Hint:* *Automation & MI is part of the Marketing department.*

*One clear advantage is that this process can be done on a daily basis to provide current information and to visualize it in a dashboard. The marketing department can look at different trends on daily basis and adjust their strategy.*