PART 3: Automating the process (Includes Assessment)

To ensure that you code can be used to automate this process. First you will save you dataframe or numpy array as a CSV file.

QUESTION 5:

Take the initial steps towards automation

1. Save your cleaned, combined data as a CSV file.
2. From the code above create a function or class that performs all of the steps given a database file and a streams CSV file.
3. Run the function in batches and write a check to ensure you got the same result that you did in the code above.

There will be some logic involved to ensure that you do not write the same data twice to the target CSV file.

Shown below is some code that will split your streams file into two batches.

```
## code to split the streams csv into batches
data_dir = os.path.join(".","data")
df_all = pd.read_csv(os.path.join(data_dir,"aavail-streams.csv"))
half = int(round(df_all.shape[0] * 0.5))
df_part1 = df_all[:half]
df_part2 = df_all[half:]
df_part1.to_csv(os.path.join(data_dir,"aavail-streams-1.csv"),index=False)
df_part2.to_csv(os.path.join(data_dir,"aavail-streams-2.csv"),index=False)

```

You will need to save your function as a file. The following cell demonstrates how to do this from within a notebook.

```
%%writefile aavail-data-ingestor.py
#!/usr/bin/env python

import os
import pandas as pd
import numpy as np
import sqlite3

data_dir = os.path.join(".","data")

pass

```

Overwriting aavail-data-ingestor.py

You will also need to be able to pass the file names to your function without hardcoding them into the script itself. This is an important step towards automation. Here are the two libraries commonly used to accomplish this in Python.

- getopt
- argparse

You may run the script you just created from the command line directly or from within this notebook using:

```
!python aavail-data-ingestor.py aavail-customers.db aavail-streams-1.csv
```

Run the script once for each batch that you created and then load both the original and batch versions back into the notebook to check that they are the same.

An answer key has been provided in the form of an online Jupyter Notebook for your to review upon completion of this exercise.

QUESTION 6:

How can you improve the process?

Make a mental note of some of the ways that you could improve this pipeline.

An answer key has been provided in the form of an online Jupyter Notebook for your to review upon completion of this exercise.