Data Pipeline Portfolio Project

Purpose: Create an automated data pipeline to update information about cancelled subscribers to cademycode.

Process: The initial database (cademycode.db) was inspected to reveal three (3) tables containing information regarding users, job ids, and career ids. The structure of these tables is shown below:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Students | | | | | | | | | |
| Col name | Uuid | Name | Dob | Sex | Contact\_info | Job\_id | Num\_course\_taken | Current\_career\_path\_id | Time\_spent\_hrs |
| Initial type | Int64 | Object | Object | Object | Object | Object | Object | Object | object |
| Final type | Int64 | Object | Object | Object | object | Float | Float | Float | float |

Table 1 - Students table showing the data types before and after being updated.

|  |  |  |  |
| --- | --- | --- | --- |
| Jobs | | | |
| Col\_name | Job\_id | Job\_category | Avg\_salary |
| Initial Type | Int64 | Object | Int64 |
| Final Type | Int64 | Object | Int64 |

Table 2 - Jobs table with data types before and after updating types.

|  |  |  |  |
| --- | --- | --- | --- |
| Courses | | | |
| Col\_name | Career\_path\_id | Career\_path\_name | Hours\_to\_complete |
| Initial Type | Int64 | Object | Int64 |
| Final Type | Int64 | Object | Int64 |

Table 3 - Courses table with data types before and after updating the types.

Upon inspection of the data types, it was determined that a number of the columns needed to be converted to a different type. The tables above show which columns had their types updated and what type to which they were converted.

With the datatypes converted into usable types, the data was inspected for missing or duplicated information.

1. The jobs table contained a few duplicate rows which were removed.

2. The students table was found to be missing data in the job\_id, num\_course\_taken, current\_career\_path\_id, and time\_spent\_hrs columns. This data was inspected and the following conclusions were made: The missing values from num\_course\_taken and job\_id were determined to be Missing at Random (MAR), and, because the associated rows were about 5% or less, the data was removed. In the case of current\_career\_path\_id and time\_spent\_hrs, it was found that all rows missing data from either column were the same, and it was concluded that the data was structurally missing. The nature of why the data was structurally missing was determined to be that the associated users had not chosen a career\_id and thus had not spent any time on the career courses. As a solution, the career\_id for all these rows was set to zero as was the time\_spent\_hrs.

Two of the remaining rows, dob and contact\_info, were expanded into more columns: age, age\_range, email, street, city, state, zipcode.

3. Having added a new career\_id to the students table, it was then necessary to add a row to the courses table to correspond to the new value.

With the data in each table being cleaned, it was decided to merge the three tables into a single table to make analysis simpler. To do this, the job\_id and current\_career\_path\_id were used to add four more columns to the final table: career\_path\_name, hours\_to\_complete, job\_category, and avg\_salary.

The final table was then ready to be exported. The export was done to a csv file as well as a sql database.

The process described above was converted into a python program that would inspect and clean the data into the format described above and then add the updated information to the existing files.  
 Once complete, a bash script was developed to run the python program, inspect output logs, and, if appropriate, output the resulting csv and database to the production folder.