**The import of the Mexican YouTube video csv and JSON files**

The more universal default codec “UTF-8” could not be used as it produced the following exception

**UnicodeDecodeError:'utf-8' codec can't decode byte 0xc3 in position 254: unexpected end of data.**

Instead the unicode block ISO-8859-1, which encodes based on 191 characters from the Latin script, was used to read the \*.csv into the dataframe. Each character is encoded as a single eight-bit code value as noted in Wikpedia.



A group discussion decided on which columns should be universally retained for all three countries’ US, Canada, and Mexico \*.csv and JSON files to ensure the dataframes and database elements were consistent.

A normal store of the JSON data, pd.read\_json, into a dataframe resulted in "missing" data as the JSON was not fully traversed through the list of dictionaries. As a result, an alternate approach was used to parse out those non-traversed data elements into several lists composed of etag, id, and title.

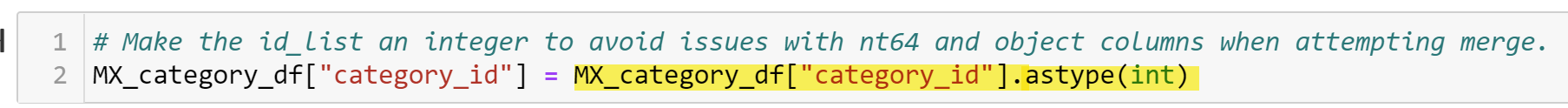
Text

Description automatically generated

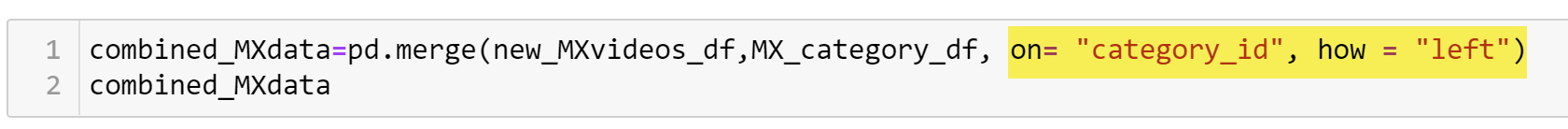
The etag, id, and title lists were applied to the dataframe as seen below. This method was applied for the three countries US, Canada, and Mexico.



In order to merge the dataframes containing the \*.csv and JSON files based on an “id”, it was necessary to change the data type of the “id” on the JSON “category\_id”. The JSON “id” had to be converted to integer before a merge could be successfully attempted. Otherwise an exception occurred with nt64 and object columns attempting to merge (join). The data type assignment can be seen below. This type of data reassignment needed to applied for the three companies.



The dataframes were “left joined” based on a shared “category\_id” between them. The join was based on a many-to-one relationship between the videos and category dataframes.



Finally a key piece of data not included in the files was assigned to the combined dataframe namely a country code to identify which data was associated with which country as all coutries would later be concated together. In this case a two character country code was applied for Mexico and in a similar fashion ‘US’ and ‘CA’ were applied to the United States and Canadian dataframes respectively.

