ETL Report

Cy Edmonds, Francis Kabongo, Huda Awad

Cy Edmonds
Francis Kabongo
Huda Awad

ETL Report

Extract:

Pulled down the files into an excel document from NYC Department of Finance, https://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page. In the Detailed Annual Sales Reports by Borough, we grabbed the 2016, 2017, and 2018 data files for the five boroughs: Manhattan, Bronx, Brooklyn, Queens, and Staten Islands.

Transform:

Took the excel sheets and pulled them into python. We concatenated all of the files for the same boroughs of the different years. Then to clean the data, we used pandas. We identified the number of the boroughs with the names and switched out all the numbers for the names (i.e 1 being Manhattan, 2 being Bronx, 3 Brooklyn, etc.). In all, for this step we changed the data type of the borough column from an integer to a string.

We continued by finding the different data types to ensure that we have a consistent run through with all of the data. Noticing that the data sets had nulls, we changed the nulls to zeros. We've found that the data types were similar for the data pulled from 2016 and 2017, but then in 2018 the columns Zip Codes to Years Built were in as float64, and to keep the uniformity we changed it from float64 to int64.

To remove any data that we were not using, we dropped the Easement columns from each of the borough's data set as it did not have any information in the rows and did not provide anything for the data sets.

Cleaning up some more, we took the column names and unified them with the code: rolling_2018.columns = rolling_2018.columns.str.strip().str.replace(' ', '_'). This easily took all the columns for each of the files and added an underscore(_) for every space in between the words.

Load:

To load our data, we have decided to use a Relational Database (SQL), since the data was being represented in tables and rows. With it being a very large dataset, more than 80,000 rows, we could simply import the data into a SQL database and quickly query anything out that we would like to identify from combining the different datasets.