Ivan Huang

Peter Hardy

Danielle Calara

May 15, 2019

Technical Report

**Purpose**

The purpose for combining information from multiple databases is to determine if the New York Museum of Modern Art (MoMa) has represented the “Top 500 Contemporary Artists of 2018” within their art collection.

**Process**

*Extract*

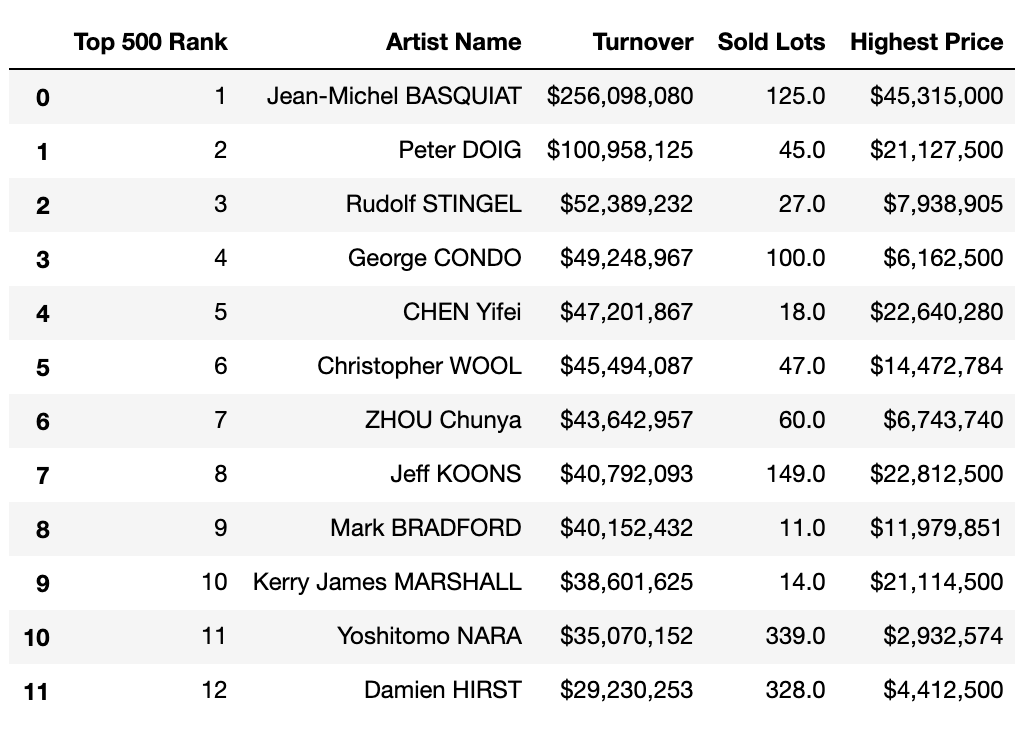
The datasets that have been combined are: the “Top 500 Contemporary Artists of 2018” from Artprice.com, “MoMa Art Collection: Artists”, and “MoMa Art Collection: Artworks”, both from Kaggle.com. The type of data of the “Top 500 Contemporary Artists of 2018” is in HTML format. The type of data for both datasets from the “MoMa Art Collection” is in CSV format.

The CSV files of “MoMa Art Collection: Artists” (artists.csv) and “MoMa Art Collection: Artworks” (artworks.csv) were imported to Jupyter Notebook, using Pandas. In order to extract the “Top 500 Contemporary Artists of 2018” table from the Artprice.com website, a for-loop was created in Jupyter Notebook to list the URLs that include the complete table data. To grab data from each of those URLs, a for-loop was created to loop through each webpage and grab the data from the top artists table, using Pandas to read the HTML code. That collected data was used to create a new dataframe named “combined\_df”.

*Transform*

*“Top 500 Contemporary Artists of 2018”*

When looping through each webpage for the table data, a footer row was repeatedly grabbed with the data. In order to remove these extra rows from the dataset, the last row from each webpage was removed because it was located at the footer of each webpage of the top 500 artists table. Also, birth and death dates were included in the same column as the artists name. We removed those dates so the Artist’s Name column would match better with the Artist’s Name column in the “moma\_df”. Figure 1 below is a sample of the cleaned dataframe.



**Fig. 1. A sample of the cleaned dataframe, “Top 500 Contemporary Artists of 2018”.**

*“MoMA Art Collection: Artists”*

The CSV file, “artists.csv” was read into a Jupyter Notebook, using Pandas. Columns ("Name", "Nationality", "Birth Year", and "Death Year") were selected from the data in order to create a new dataframe then they were renamed to be more descriptive. To clean the data, the duplicates were dropped, using Pandas in a Jupyter Notebook. Figure 2 below is a sample of the cleaned dataframe.

****

**Fig. 2. A sample of the cleaned dataframe, “MoMa Art Collection: Artists”.**

*“MoMa Art Collection: Artworks”*

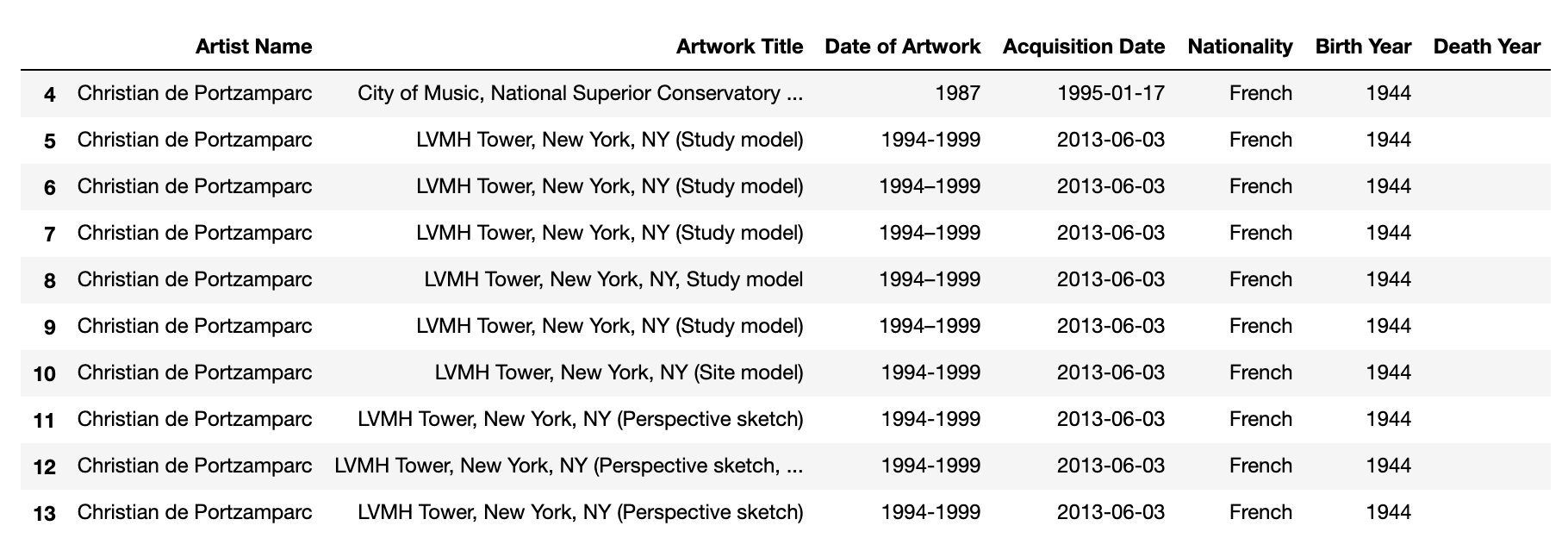
The CSV file, “artworks.csv” was read into a Jupyter Notebook, using Pandas. Afterwards, columns (“Name”, “Title” “Date”, and “Acquisition Date”) were selected from the data in order to create a new dataframe then they were renamed to be more descriptive (“Name” became “Artist Name”). Figure 3 below is a sample of the cleaned dataframe.



**Fig. 3. A sample of the cleaned dataframe, “MoMa Art Collection: Artworks”.**

*Merge*

The MoMa Artists and and MoMa Artworks data frames were merged on the “Artist Name” column, using Pandas to create a new dataframe (“moma\_df”) in Jupyter Notebook. The purpose of this project is to determine if and how well the top 500 contemporary artists of 2018 are featured in the New York MoMA’s “contemporary art” section of their art collection. The MoMA features an enormous number of artworks (upwards of 120,000 objects), therefore, the whole dataset is not necessary when we were only looking for 500 artist names who were all born in the 20th century. Therefore, in order to reduce the dataset, values from the “Birth Year” column were dropped if the value was less than 1920 because not one artist on our Top 500 list was born before the year 1920. Figure 4 below is a sample of the final dataset.



**Fig. 4. A sample of the cleaned & condensed dataframe moma\_df, “MoMA Contemporary Art Collection”.**

*Load*

Next, the combined MoMa dataframe and top 500 artists dataframe were imported as a schema into MySQL. MySQL was chosen for the final product because it presents queries in a list form, which enables us to analyze the results more efficiently. Also we chose MySQL because this is data that is static and not going to change so we would have no need to get back into it for update. MySQL Workbench allows us to import our two tables and combine the specific columns we need for our final query. In MySQL Workbench, the two data frames were joined on the “artist\_name” column to query the top contemporary artists of 2018 that are featured in the New York MoMa. Figure 5 is a sample of the final query.



**Fig. 5. A sample of the final query that shows the “Top Contemporary Artists on 2018” featured in the New York Museum of Modern Art collection.**

**Python read csv and insert data to SQL**

We are using these two arguments of Pandas read\_csv function, First argument is the path of the file where first csv is located and second argument is for the value separators in the file. So we have imported the csv files into two dataframes and now it’s time to merge these two files. We will use Pandas merge function to merge these two files together. Now we will push the rows of this merged dataframes into a MYSQL and we have to first establish the connection with the MYSQL server using sqlalchemy. we will use Pandas to\_sql function which will insert these rows of merged dataframe into the MYSQL Table.

**Findings**

* 179 (36%) of the top 500 artists in 2018 are included in the MoMa collection, indicating that artist popularity is a small part of their selection criteria for their exhibits. Those 179 artists accounts for only 2.33% of art selections in MOMA indicating secular popularity is not a major selection criteria.
* For future studies, it would be interesting to use the collected data to determine if there is a relationship between the rank of an artist and number of sales/price of their art pieces.