Project Final Report

INST 327, Section 0101

May 10, 2022

Group 4

Patrick Bovard, Maya Willey, Ethan Pham,

McKenna Shay, Olivia Zama

Introduction

Our topic is about immigration to Ellis Island during the height of America's industrial revolution, specifically the people who immigrated in that time and the ships they came on. Our original plan was to only include immigrants from 1892 to 1897, however we extended the dates to 1920 due to the number of immigrants from that time. Our database contains facts about people that came through Ellis Island (such as their names, their country of origin, when they arrived, and other demographic data) and facts about the passenger ships that came through Ellis Island carrying these immigrants (including their arrival and departure times and how many people they carried).

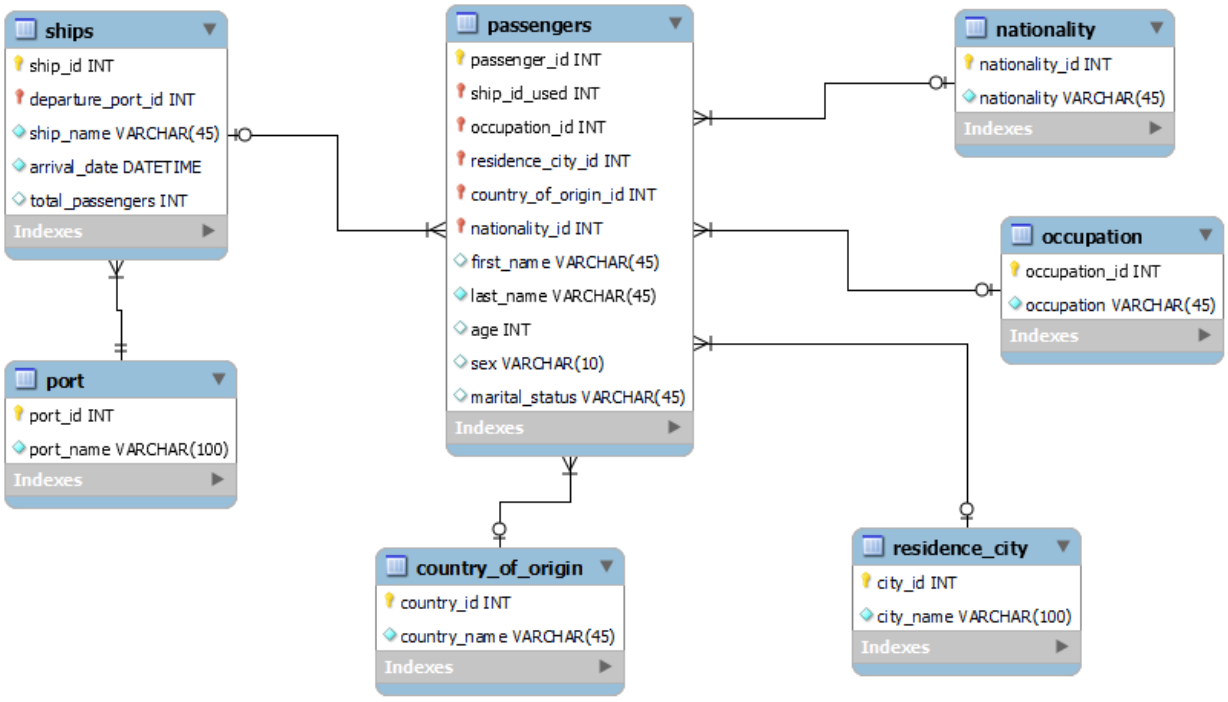
We chose this topic originally because we thought it would be interesting to explore as well as the fact that most immigration records available are pictures and can be hard to read. Being able to see individual names and stories of these immigrants helps to humanize them and build a broader world view. Additionally, we chose 1892 as the start year because that is when Ellis Island opened and due to the number of records on hand, we chose to extend our end date to 1920, the start of the prohibition era. Our database should be able to properly reflect the demographic facts about some of the millions of people who emigrated to the United States during this time period.

Database Description

Logical Design

The database consists of seven tables, most of which are simple tables that hold the ID and name of certain variables such as nationality, occupation, residence city, country of origin, and ports. We also decided on which smaller tables to create based on what data was repeating in

the main tables. For example, since we saw that there were many ports and residence cities that were the same, we decided to make those into smaller tables with ID attributes that could be referenced by the main tables. Our main tables are the ships and passengers tables, which connect the smaller tables as foreign keys. The passengers table holds the most foreign keys as it is used to link all of the tables together.



Physical Database

We based our physical database off of the data we found online. Using our data sources like FamilySearch and the official Ellis Island Ship Search, we looked at how the data is formatted and decided to design our database based on that. For example, the Ellis Island Ship Search only included a ship name, its departure port, its arrival date, and a total number of passengers, so we decided to base the ships table off the data provided through the Ellis Island website. Our main tables, passengers and ships, hold all the data we got directly from our data

sources while the smaller one-to-many tables (nationality, occupation, residence_city, country_of_origin, and port) had to be created manually using the data we got from the main tables.

While forward engineering the database from the ERD, we did run into a problem. MySQL would not allow us to create the passengers table because we initially had the foreign keys from other tables set to allow null values. We did this because some passengers might have incomplete data and not have their residence city, occupation, or any of the data from the smaller tables. MySQL does not allow null values for primary key attributes, so we had to set each foreign key in passengers and ships as non-null and put "NA" into each null cell in our spreadsheets to import the data correctly.

We also ran into a few problems when importing the data from our spreadsheets into the physical database. Initially, we only had the main tables as spreadsheets and thought that MySQL would allow us to choose individual columns when importing and be able to correctly populate the smaller tables from the data in the main table spreadsheets. This did not go as planned, because it added every single value from the column we chose from the main table to the smaller table, including repeating values. For example, when we used the ships spreadsheet to populate the port table, "Antwerp" showed up as port_name for multiple rows because there were multiple ships that used Antwerp as their departure port. Ultimately, we ended up creating spreadsheets for each of the smaller tables to make sure no values were repeating and each row was a unique value (i.e. so "Antwerp" only shows up for one row in the port table). Additionally, because of this assumption we made, we had the columns in the main table spreadsheets that corresponded to foreign keys in the database as names/words and not numbers, which was not allowed when importing because they would not connect to the ID numbers set as the foreign keys. So we had

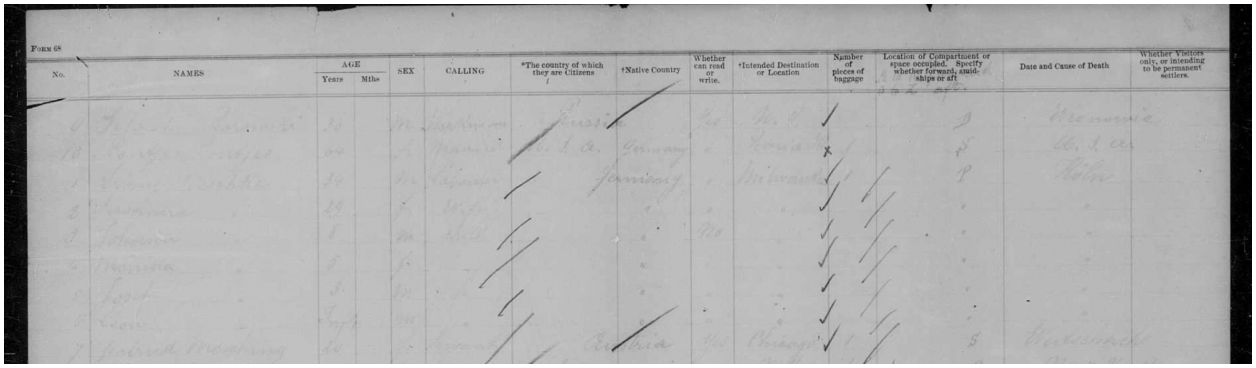
to go into the spreadsheets for the main tables and manually change every word to its corresponding ID number in the smaller table. For example, since the ID for "Antwerp" in the departure_port table is 2, we had to go into the spreadsheet for ships and change every occurrence of "Antwerp" to 2.

Sample Data

Our data was taken from an AI transcribed catalog of passenger lists called *"Passenger lists of vessels arriving at New York, 1820-1897 ; index to passenger lists of vessels arriving in New York, 1820-1846*" located on the genealogy website FamilySearch.com. This data is available at the National Archives and Records Administration in D.C. but it is not transcribed like the one on FamilySearch. The data set is made up of 12 columns which include: first name, last name, sex, age, immigration date, residence place, occupation, marital status, nationality, the page number of the passenger list, the affiliate line number, departure port, and ship name. We chose to look at passenger lists that departed from Europe from 1892-1920 as Ellis Island opened in 1892 and the mass migration period ended around 1920. This data is from handwritten records made in the 1800's and 1900's and thus some of the columns have illegible data or missing information. This is reflected in the occupations and marital status categories specifically as many people without jobs would leave it blank or as we can only assume they were in a rush to fill them out.

Name	Sex	Age	Immigration	Immigration P	Residence Place	Occupation	Marital Statu	Nationality	Page Number	Affiliate Line
Henry Gessely	M	27	1892-02-08	Ellis Island, Ne	NA	Waterman	Single	Belgium, Belgian	9	31
Bartholomeus Wiseman	M	17	1892-02-08	Ellis Island, Ne	NA	Laborer	NA	Germany, German	9	37
August Lohmar	M	28	1892-02-08	Ellis Island, Ne	NA	NA	NA	Germany, German	9	48
Rudolf Weyand	M	30	1892-02-08	Ellis Island, Ne	NA	NA	NA	Germany, German	9	55
Georg Krieger	M	15	1892-02-08	Ellis Island, Ne	NA	NA	NA	Germany, German	9	52
Michael Deubel	M	37	1892-02-08	Ellis Island, Ne	NA	NA	NA	Switzerland, Swiss	9	45
Helene Weiss	F	32	1892-02-08	Ellis Island, Ne	NA	NA	Married	Germany, German	9	30

Here is a portion of our data set showing some of the included columns.



Example of one of the passenger lists that was harder to read.

"New York Passenger Arrival Lists (Ellis Island), 1892-1924," database with images, FamilySearch (https://familysearch.org/ark:/61903/3:1:33SQ-G1DW-FX5?cc=1368704&wc=4FMB-7CY : 25 January 2018), Roll 581, 2 Jan 1892-8 Feb 1892 > image 9 of 802; citing NARA microfilm publication T715 and M237 (Washington D.C.: National Archives and Records Administration, n.d.).

Views/Queries

We created a total of six views for our database, each one performs a separate task. For example, one of the views shows occupations of people in the database who identify as having an American Nationality, another one shows the date of the earliest ship to arrive, the date of the last ship to arrive, and the total number of passengers in our database. Another one will show the marital status of people with certain nationalities. These views and queries are designed to simplify our database so that people can read it better and see specific trends within the immigration data we collected.

Changes from Original Design

We have changed multiple things from our original design. We changed the timeframe of our data from 1892-1897 to 1892-1920 so we could see more long-term trends in the immigration data and be able to more fully capture the time when immigration peaked at Ellis

Island. In our original design, we also considered storing pictures of ships or immigration documents in our database, which we did not end up doing because most of the data we found are comprehensive enough and those pictures would not be as important for our information needs. Additionally, we wanted to include the departure time of ships but, as stated above, the Ellis Island Ship Search did not store departure times of ships, so we could not include that in our database. We also considered publishing our database to the public in order to help people who want to look at immigration data and potentially find family members, but we ultimately did not do that due to our database being a smaller snapshot of immigration at the time and can mostly be used to see trends in immigration at that time. Finally, we initially planned to have two of our smaller tables store age and sex data respectively, but we decided that would not be necessary since that data can easily be stored in the passengers and we decided that nationality and residence city would make more sense being stored in smaller tables. Besides those factors, our database mostly remained as we initially planned.

Database Ethical Considerations

The ethical considerations we took during the process of this project occurred when developing our project idea, drafting our database, and collecting our data. We determined as a group that we would not conduct any SQL database work until the diversity, equity, and inclusion problem was solved. But as far as data ethical concerns go on that front, all of the data that we found was free to be used by the public and the websites we found the data on stated as such. Therefore, ethical considerations had little to no impact when conducting our database design and operations.

Lessons Learned

When we were organizing our data, we learned more about normalization and how to remove partial and transitive dependencies. We had to keep modifying our model until it was normalized. Additionally, we had to pay close attention to how we created the model and data types, especially the character limit on VARCHAR data types.

During our creation of the Ellis island database, we learned about working on a database on different operating systems. Some of our group members have a Mac operating system, while some have Windows. When we were executing the scripts, our group members with Macbooks had to reverse engineer the database compared to our group members with Windows. Group members with Windows forward engineered the database using the model.

Potential Future Work

Potential future work four our database includes expanding the dates further. Our database was 1892-1920, but Ellis Island was open until 1954 when it was abandoned. If we are able to access data about departure times, we could add that to our database's ships table. We could also try to add more passengers and ships to our databases, there are hundreds and thousands of them so there is no shortage of data. Additionally, we could try publishing our database for the public so people could access our data and see the trends and people for themselves. As we have discovered in this project, there are not many data sources online with this Ellis Island immigration data that is not made for users looking for a specific person or ship. We think our database would be unique in letting people see everyone and look at trends beyond an ancestry search.

References

"New York Passenger Arrival Lists (Ellis Island), 1892-1924," database with images, FamilySearch (https://familysearch.org/ark:/61903/3:1:33SQ-G1DW-FX5?cc=1368704&wc=4FMB-7CY : 25 January 2018), Roll 581, 2 Jan 1892-8 Feb 1892 > image 9 of 802; citing NARA microfilm publication T715 and M237 (Washington D.C.: National Archives and Records Administration, n.d.).