Julie Carlson
INFO 628: Data Librarianship & Management
Final Project Check-In #2
December 1, 2021

Given the vast amount of data contained in the National Register of Historic Places dataset, there were a number of questions that could be explored: What state has the most buildings listed? What category of building is most represented? What areas of significance are most commonly referenced? What terms appear most frequently in the properties' given names? Quantitative methods help answer these questions, providing insight into the types of places that the National Register deems worthy of preservation.

Quantitative Analysis with Python

The raw dataset was first cleaned in OpenRefine. This involved standardizing the values in the "Category of Property" field to make them uniformly capitalized, adding a column to show only the year of listing (derived from the "Listed Date" field), removing empty columns, and deleting a row that contained only a property name with no other relevant information. The clean dataset was saved as a CSV file. A Python script was used to count the number of properties in each state, category, and area of significance. The script wrote the counts to three CSV files.

Data Visualization in Tableau Public

The CSV files produced by the Python script were imported into Tableau Public to visualize the data. The count by state was illustrated through a map visualization where a state's color corresponds to its count. The count by category was visualized as a treemap, and count by

area of significance as a packed bubble chart where the size and color of the bubbles indicate the count. The visualizations were made freely available on Tableau Public.

Text Analysis in Voyant

Returning to the cleaned CSV file from OpenRefine, the data from the "Property Name" column was copied into Voyant Tools. The tool analyzed the frequency with which words appeared in the text, producing a count for each word (excluding automatically-detected stopwords). It also produced a word cloud visualization where each word's size corresponds to its number of appearances in the corpus. The word cloud was saved as a PNG file.

Next Steps

The Python script will be augmented to retrieve the count of listings by year in which they were listed, to explore potential temporal trends in the National Register. The resulting CSV file will be visualized in Tableau Public. All Tableau visualizations will be further refined to effectively communicate the findings. The word cloud from Voyant Tools may also be refined to remove more stopwords, such as "la" and "san."

PLAN OVERVIEW

A Data Management Plan created using DMPTool

Title: Exploring the National Register of Historic Places

Creator: Julie Carlson

Affiliation: Pratt Institute (pratt.edu)

Project abstract:
Since 1966, the National Park Service has maintained the National Register of Historic Places, “the official list of the Nation's historic places worthy of preservation.” The National Register contains more than 96,000 properties as of June 2021. Although data on the properties is freely available as a dataset, there is a dearth of scholarship exploring the places' characteristics. Using quantitative analysis methods, I will examine the dataset to consider what kind of places are deemed worthy of preservation.

Start date: 09-17-2021

End date: 12-20-2021

Last modified: 12-01-2021

EXPLORING THE NATIONAL REGISTER OF HISTORIC PLACES

DATA COLLECTION

I will use the National Park Service's "Spreadsheet of NRHP Listed properties (listing up to 06/17/2021)" file as my raw data. This spreadsheet is freely available on the National Park Service's website in .xlsx format. The spreadsheet contains 96,644 rows, covering properties added to the National Register from its inception in 1966 through June 17, 2021. It is worth noting that the National Park Service maintains a separate spreadsheet of properties removed from the National Register, so my work will not be a comprehensive overview of every property that has ever been listed; I will only be exploring properties actively listed as of this project.

For my project, I will convert the .xslx file into .csv format to enable long-term access that does not rely on proprietary software. I plan to clean the data in OpenRefine and save my results as a separate .csv file. I will write and share Python scripts, which will be saved as .py files, to perform quantitative statistical analysis on the cleaned dataset. These Python scripts will produce additional .csv files, which I will use to create visualizations of my findings. I will create visualizations with the freely-available software Tableau Public. Additionally, I will use Voyant Tools—an open source application—to perform quantitative text analysis, resulting in additional visualizations generated by the application.

The project files will be organized by file type within one larger project folder. Per the structure standards discussed in class, raw data and metadata will be in a “data” folder, cleaned data and visualizations will be in a “results” folder, scripts will be in a “src” folder, and associated text documents will be in a “docs” folder. File names will be easily understandable, and will be prefixed with the date created in ISO 8601 format, as recommended by Kristin Briney.

DOCUMENTATION AND METADATA

Each folder will contain a README file with details about the contents. The “docs” folder will also contain a codebook with metadata about the raw dataset and the datasets that I create throughout the project.

ETHICS AND LEGAL COMPLIANCE

The raw data at the base of my project is government content in the public domain. My cleaned dataset and any visualizations I produce will remain in the public domain for free reuse.

STORAGE AND BACKUP

I will keep three copies of all materials associated with this project: one on my laptop, one on Google Drive, and one on an external hard drive. I will save my files to the hard drive each Thursday.

SELECTION AND PRESERVATION

The raw dataset, processed datasets, Python scripts, visualizations and all accompanying documentation will be saved on GitHub for the foreseeable future.

The spreadsheet containing my project's raw data is periodically updated on the National Park Service's website. For reproducibility purposes, the actual dataset I downloaded from their website (containing listings through 6/17/2021) would be particularly valuable to retain. Additionally, my documentation and Python scripts may be of use to future researchers. As the National Park Service updates its dataset, researchers could potentially run my script on the new spreadsheet and gain new, more timely insights.

DATA SHARING

All data, scripts, and documentation will be shared on GitHub in open file formats, like .csv and .py. Screenshots of the visualizations from Tableau Public will be saved as .png files with links to the dashboard posted on GitHub. The word cloud generated from Voyant Public will also be saved on GitHub as a .png file. All data and analysis will be made available by the end of the project, December 20, 2021. There will be no restrictions on data sharing.

RESPONSIBILITIES AND RESOURCES

I will be responsible for all data management activities. I will take into account feedback received from my peers and Professor Vicky Rampin, and revise/implement the DMP accordingly.

My project will require freely available resources including OpenRefine, Visual Studio Code, Voyant Tools, and Tableau Public.