July 24, 2019
CUSP Urban Science Intensive
Capstone Final Report
Sponsor: Dr. Nicholas Wolf, NYU Division of Libraries

# Recovering the Business and Economic History of New York City Through Large-Scale City Directory Data

Fekade Brook, Linda Lyu, Matthew Sauter, Vaidehi Thete, Shelly Yin

## Abstract

The New York Public Library holds within its archives dozens of directories containing historical New York City business and residential data. Under a previously commissioned project, several of these tomes – dated between 1849 and 1922 – were digitized, and the textual data had been extracted using Optical Character Recognition (OCR) techniques. Our capstone centers on the previously completed research, with three primary goals: exploring opportunities for increasing the accuracy of the data extraction methodology, improving upon the reproducibility of the methods previously used, and exploring the extracted data to demonstrate the potential that these data have for expanding our understanding of 19th century New York City.

## Introduction

The collection of business and residential data is a societal practice that has existed far longer than the current capabilities of modern computational analysis. These historical records exist outside of the realm of what one might typically consider being "computer-friendly data"; prior to contemporary computing capabilities, printed records were trapped within physical books, only to be accessed one tome, one page, one line at a time. Through the application of modern machine learning techniques, these printed directory data can now be extracted and digitized, allowing for mass consumption and analysis.

The New York Public Library holds several directories of this type within their archives. In 2016, many of these records were digitized, with the end goal being the creation of a searchable database of historical information.[1] The initial attempt at extracting and classifying the data (from directories dated between 1849 and 1922) using OCR methods was successful with an estimated 90% accuracy rate. Through our research, we've attempted to apply similar methodologies in novel ways in hopes of improving the accuracy of the output. In particular, we chose to deploy a recursive framework for the OCR extraction, with enhanced preprocessing and re-processing of the source data for improved results.

In addition to the reconstruction of the data extraction framework, we conducted an exploratory analysis as a means of showcasing the potential contained within the extracted data, while also satisfying the various curiosities we had surrounding these directories. In particular,

[1] Sutton, C. (2016, October 5). "New York Public Library Digitizes 137 Years of New York City Directories".
[2] Docparser, "Improving OCR Accuracy With Advanced Image Preprocessing".

we focused on the gender breakdown of the workforce, as well as the migratory patterns of various employed ethnic communities over a subset of our data.

**Problem Statement**

Our primary problem can be summarized as such: *how could we improve upon the accuracy and reproducibility of data extraction methods from the initial processing of these directory pages?* We believe that the deployment of robust pre-processing methods, alongside recursive OCR analyses, is the best approach for increased accuracy and efficiency in the extraction of directory data. Throughout the reconstruction of each component of the extraction process, we must ensure that our framework is assembled under the rigorous scrutiny of reproducibility.

**Literature Review**

Our primary resource for this project was the documentation from the initial project. However, the following articles and pages were consulted while developing our framework:

OCR Enhancements

In developing our approach for improving OCR, we researched common pre-processing practices that have proven to be of use in increasing accuracy. "Improve OCR Accuracy With Advanced Image Preprocessing," published by Docparser, provided us with a foundational check list of methods typically used to adjust images in advance of OCR, including scaling, contrast, binarization, de-skewing, and Zone Analysis.[2] "Enhancing Degraded Document Images via Bitmap Clustering and Averaging" by John D. Hobby and Tin Kam Ho provides a detailed methodology of how to re-render degraded images to improve upon the OCR engine's ability to extract text[3]; we thought this might be of use when analyzing the younger directories, as the text in the page scans is quite small and might require enhancement. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs," by Rose Holley, not only provides a robust list of methods to improve OCR, but provides a practical assessment of which are worth undertaking based on their returns.[4]

Word Tagging

In developing our tagging and classification method, we referenced "Build a POS tagger with an LSTM using Keras," posted on Natural Language Processing for Hackers,[5] in preparation for deploying a hybrid deep learning/Conditional Random Field (CRF) model to improve the accuracy and efficiency.

---

[2] Docparser, "Improving OCR Accuracy With Advanced Image Preprocessing".

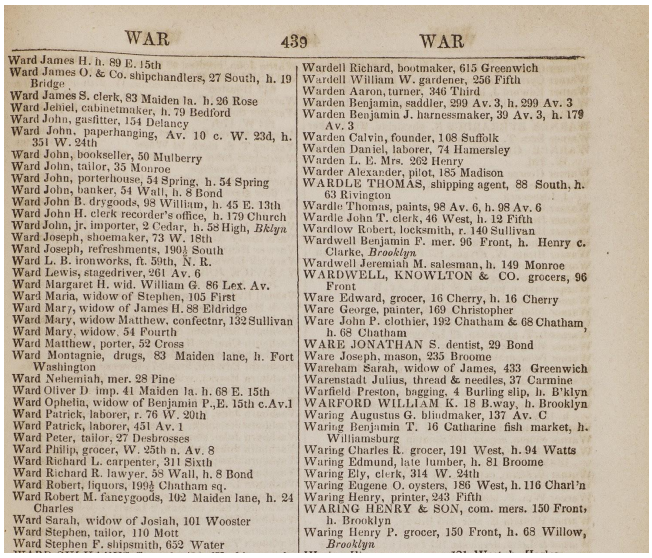
[3] Hobby, J. and Ho, T.K. (1997, August). "Enhancing degraded document images via bitmap clustering and averaging".

[4] Holley, R. (2008, October). "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs".

[5] Bogdani (2018, September 8). "Build a POS tagger with an LSTM using Keras".

## Data

The data for this project primarily consisted of scanned pages from the New York Public Library archive, dated between 1849 and 1922. These images were digitized into TIFF files and shared with us for the sake of text extraction[6]. The pages are structured similarly across the years: columns of printed entry data, with each entry made up of four primary elements: name, occupation, business address, and (inconsistently) home address. Early directories (the first 41) divide the pages into two columns; later directories include three, four, and five columns of entries per page. Later directories also intersperse advertisements throughout the pages, both within and bordering the columns. These advertisements proved to be a limiting element for our efforts, as they were inconsistent and unpredictable in size, format, and orientation.



WAR 439 WAR

Ward James H. h. 89 E. 15th
Ward James O. & Co. shipchandlers, 27 South, h. 19 Bridge
Ward James S. clerk, 83 Maiden la. h. 26 Rose
Ward Jehiel, cabinetmaker, h. 79 Bedford
Ward John, gasfitter, 154 Delancy
Ward John, paperhanging, Av. 10 c. W. 23d, h. 351 W. 24th
Ward John, bookseller, 50 Mulberry
Ward John, tailor, 35 Monroe
Ward John, porterhouse, 54 Spring, h. 54 Spring
Ward John, banker, 54 Wall, h. 8 Bond
Ward John B. drygoods, 98 William, h. 45 E. 13th
Ward John H. clerk recorder's office, h. 179 Church
Ward John, jr. importer, 2 Cedar, h. 58 High, *Bklyn*
Ward Joseph, shoemaker, 73 W. 18th
Ward Joseph, refreshments, 190½ South
Ward L. B. ironworks, ft. 50th, N. R.
Ward Lewis, stagedriver, 261 Av. 6
Ward Margaret H. wid. William G. 86 Lex. Av.
Ward Maria, widow of Stephen, 105 First
Ward Mary, widow of James H. 88 Eldridge
Ward Mary, widow Matthew, confectnr, 132 Sullivan
Ward Mary, widow, 54 Fourth
Ward Matthew, porter, 52 Cross
Ward Montagnie, drugs, 83 Maiden lane, h. Fort Washington
Ward Nehemiah, mer. 28 Pine
Ward Oliver D imp. 41 Maiden la. h. 68 E. 15th
Ward Ophelia, widow of Benjamin P., E. 15th c. Av. 1
Ward Patrick, laborer, c. 76 W. 20th
Ward Patrick, laborer, 451 Av. 1
Ward Peter, tailor, 27 Desbrosses
Ward Philip, grocer, W. 25th n. Av. 8
Ward Richard L. carpenter, 311 Sixth
Ward Richard R. lawyer, 58 Wall, h. 8 Bond
Ward Robert, liquors, 199½ Chatham sq.
Ward Robert M. fancygoods, 102 Maiden lane, h. 24 Charles
Ward Sarah, widow of Josiah, 101 Wooster
Ward Stephen, tailor, 110 Mott
Ward Stephen F. shipsmith, 652 Water

Wardell Richard, bootmaker, 615 Greenwich
Wardell William W. gardener, 256 Fifth
Warden Aaron, turner, 346 Third
Warden Benjamin, saddler, 299 Av. 3, h. 299 Av. 3
Warden Benjamin J. harnessmaker, 39 Av. 3, h. 179 Av. 3
Warden Calvin, founder, 108 Suffolk
Warden Daniel, laborer, 74 Hamersley
Warden L. E. Mrs. 262 Henry
Warder Alexander, pilot, 185 Madison
WARDLE THOMAS, shipping agent, 88 South, h. 63 Rivington
Wardle Thomas, paints, 98 Av. 6, h. 98 Av. 6
Wardle John T. clerk, 46 West, h. 13 Fifth
Wardlow Robert, locksmith, r. 140 Sullivan
Wardwell Benjamin F. mer. 96 Front, h. Henry c. Clarke, *Brooklyn*
Wardwell Jeremiah M. salesman, h. 149 Monroe
WARDWELL, KNOWLTON & CO. grocers, 96 Front
Ware Edward, grocer, 16 Cherry, h. 16 Cherry
Ware George, painter, 169 Christopher
Ware John P. clothier, 192 Chatham & 68 Chatham, h. 68 Chatham
WARE JONATHAN S. dentist, 29 Bond
Ware Joseph, mason, 235 Broome
Wareham Sarah, widow of James, 433 Greenwich
Warenstadt Julius, thread & needles, 37 Carmine
Warfield Preston, bagging, 4 Burling slip, h. B'klyn
WARFORD WILLIAM K. 18 B.way, h. Brooklyn
Waring Augustus G. bindmaker, 137 Av. C
Waring Benjamin T. 16 Catharine fish market, h. Williamsburg
Waring Charles R. grocer, 191 West, h. 94 Watts
Waring Edmund, late lumber, h. 81 Broome
Waring Ely, clerk, 314 W. 24th
Waring Eugene O. oysters, 186 West, h. 116 Charl'n
Waring Henry, printer, 243 Fifth
WARING HENRY & SON, com. mers. 150 Front, h. Brooklyn
Waring Henry P. grocer, 150 Front, h. 68 Willow, *Brooklyn*

*Figure 1 – Sample page image from the 1849-1850 two column directory. The full image can be seen in Appendix A, Item 1.*

Additionally, we used the data extracted from these pages during the initial attempt to conduct analyses and to guide our understanding of the desired output[7]. These data were delivered as a JSON file, with each entry parsed and labeled with the four components of the entries. Some of the entries contained geospatial data, which was added post-extraction during the initial project. Some names within the data consisted solely of characters or non-word text, which needed to be cleaned prior to its viability. Names also had to be split in order to isolate first name, last name, and middle initial, in order for us to conduct the analyses.

## Methodologies

### Pre-Processing: Improving OCR Extraction

Research revealed that the most common methods of improving OCR accuracy include image pre-processing. As many of the popular pre-processing methods were already used during the initial extraction, we had to be creative in developing our approach to improving the accuracy. A specific site of intervention was the multi-lined entries found within the directories: several entries in the directories are split over multiple rows due to the length of the entry (Appendix A, Item 1). This proved to be a primary point of error during the initial extraction of these data, resulting in the need for an independent post-processing framework that would link separated lines together once they were extracted. We believe that improving the capture of these entries earlier in the OCR process, rather than attempting to connect separated lines during post-processing, would result in an improvement in the overall accuracy of the extraction.

---

[6] Wolf, N., Spaan, B. (2018, March 15). "List of NYPL's Digitized City Directories".
[7] Wolf, N., Spaan, B. (2018, March 15). "Space Time NYPL City Directories Dataset".

Our selected approach to improve the capture of multi-line entries was the construction of a recursive application of the OCR engine. We deployed a Canny edge detector to identify the edges contained within the page image, followed by the Hough transformation to find the straight lines separating the columns (Fig. 2; Appendix A, Item 2). We then used OCR to identify the outer margins of the page image, so that the image could be cropped, resulting in a page image that only contained the text of the entries we were interested in capturing. Using the lines identified via the Hough transform, we were able to crop out each column (Appendix A, Item 3), allowing for isolated analyses with reduced noise.

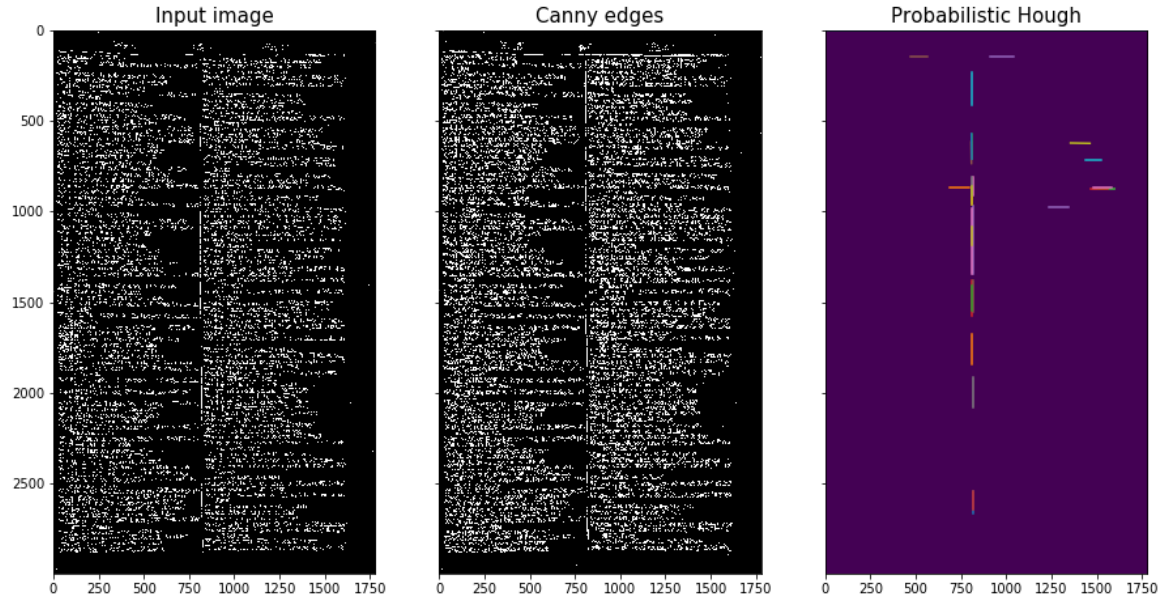



*Figure 2 – Using a Hough transform, we are able to identify the line separating each column on the page image. This line is used to crop out each column to improve upon OCR accuracy.*

Running OCR over the cropped column allowed us to capture the pixel placement of each line – via the resulting bounding boxes that formed over the text – and, more importantly, capture which lines were indented in the column (an indent indicates that the line is a continuation of the entry directly preceding). Organizing the resulting textual output into a dataframe, we then group the data by line, and affix text that has been identified as "indented" to the text of the previous line.

One challenge we face with this method is the variability of the format of each directory's pages. The inconsistencies between each directory meant that each formatting style required its own customized version of this workflow. To maximize efficiency, we focused our efforts on the double-column format, which would allow us to capture the majority of directories.

Post-Processing: Classification of Extracted Data

To categorize the words into three different classes, the original team of researchers applied a Conditional Random Field (CRF) model, which is a common NLP probabilistic method. Such methods have their own irreplaceable advantages when the data have an obvious, repetitive order sequence. In the case of our extracted entries, the order is faithfully "Start – Name – Profession – Address – End". However, we believed that including deep learning

methodologies, which have been recently applied to these types of problems[8], might be a means of improving the accuracy of this element of the project. As such, we attempted to use spaCy – a deep learning python package reliant on a pre-trained model – in order to supplement the tagging. However, we found that elements of the name components of the entries were getting lost, especially when middle and first names were presented as an initial followed by a period (e.g. "L.", "F.").

Ultimately, we determined that using the model described above was less accurate than using a simple tagging methodology relying on the syntax of each entry: words beginning with capital letters were tagged as "names"; the word immediately after the names was tagged "occupation". The remaining text was tagged as "work address," unless there was an "h." present, which was always followed by "home address". By using Regular Expression to extract all capital letters, lowercase letters, numbers, commas, and periods, we were able to deploy our tagging methodology quite simply, without any noticeable loss in accuracy.

Analyses

In an effort to explore the data, we elected to conduct several exploratory analyses that we believe to be indicative of the potential contained within the data. These analyses utilized the initially extracted data, and were restricted to the years 1850 – 1879, the decades that had geocoded markers included for each entry.

In order to add depth to the extracted dataset, we tagged each entry with two external markers: predicted gender and predicted ethnicity. These tags enabled us to conduct more interesting analyses, providing deeper insight into the structural and cultural norms surrounding employment in the 19th century.

Gender & Ethnicity Classifications

We used a package in R called *Gender*[9] to tag our entries as "male" or "female". For data predating 1930, this package encodes gender based on names and dates of birth using U.S. Census data from IPUMS. By using these datasets in place of lists of male and female names, this package is able to more accurately guess the gender of a name; furthermore, it is able to report the proportion of times that a name was male or female for any given range of years. There was quite a preponderance of widows within the data, which we filtered out via a Regular Expression filter to ensure accuracy of occupation.

For the purpose of identifying the ethnicity of names, we used the R package *Ethnicolr*[10]. Using first and last name, this package will classify the entry as one of 13 different ethnicities.

[8] Bogdani (2018, September 8).
[9] Mullen, L. (2018). "gender: Predict Gender from Names Using Historical Data".
[10] Ambekar, Anurag and Ward, Charles and Mohammed, Jahangir and Male, Swapna and Skiena, Steven (2009). "Name-ethnicity classification from open sources".

## Results

### Extraction

Our method shows promise when used for extraction from a single page, allowing for user-made tweaks to the parameters that drive column and margin identification (Fig. 3; Appendix A, Item 4). Comparing the results of the column used as an example in this report with the data extracted from this column previously, we were able to capture one line that was missed previously. If this were the case for every column on every page, we would expect to see a substantial increase in extracted text per directory. However, the current means of determining these parameters needs to become more generalized before we can scale up the project successfully.

Our first attempt at scaling up the methodology across the 41 two-column directories yielded less optimistic results. The process resulted in slightly over 125,000 entries extracted; though we were excited to see that we had succeeded in deployment, this number was much lower than expected. Investigation into the output revealed that several pages were skipped over entirely, including 5 entire directories. However, the entries that were extracted were tagged consistently and accurately using our new tagging methodology.

```
[ 'Wardell Richard, bootmaker, 615 Greenwich ',
  'Wardell William W. gardener, 256 Fifth ',
  'Warden Aaron, turner, 346 Third ',
  'Warden Benjamin, saddler, 299 Av. 3, h. 299 Av. 3 ',
  'Warden Benjamin J. harnessmaker, 39 Av. 3, h. 179 v. ',
  'Warden Calvin, founder, 108 Suffolk ',
  'Warden Daniel, laborer, 74 Hamersley ',
  'Warden L. Ii. Mrs. 262 Henry ',
  'Warder Alexander, pilot, 185 Madison ',
  'WARDLE THOMAS, shipping agent, 88 South, h. 63 Rivington . ',
  'Wardle Thomas, paints, 98 Av. 6, h. 98 Av. 6 ',
  'Wardle John T. clerk, 46 West, h. 12 Fifth ',
  'Wardlow Robert, locksmith, r. 140 Sullivan ',
  'Wardwell Benjamin F. mer. 96 Front, h. Henry ¢ Clarke, Brooklyn',
  'Wardwell Jeremiah M. salesman, h. 149 Monroe ',
  'WARDWELL, KNOWLTON & CO. grocers, 96 ront ',
  'Ware Edward, grocer, 16 Cherry, h. 16 Cherry ',
  'Ware George, painter, 169 Christopher ',
  'Ware John P. clothier, 192 Chatham & 68 Chatham h. 68 Chatham, ',
  'WARE JONATHAN &. dentist, 29 Bond ',
  'Ware Joseph, mason, 235 Broome ',
  'Wareham Sarah, widow of James, 433 Greenwich ',
  'Warenstadt Julius, thread & needles, 37 Carmine ',
  'Warfield Preston, bagging, 4 Burling slip, h. B'klyn ',
  'WARFORD WILLIAM K. 18 B.way, h. Brooklyn ',
  'Waring Augustus G. bliudmaker, 137 Av..C ',
  'Waring Benjamin T, 16 Catharine fish market, h. Williamsburg a ',
  'Waring Charlies R. grocer, 191 West, h. 94 Watts ',
  'Waring Edmund, late lumber, h. 81 Broome ',
  'Waring Ely, clerk, 314 W. 24th ',
  'Waring Eugene O. oysters, 386 West, h. 116 Charl'n ',
  'Waring Henry, printer, 243 Fifth . . ',
  'WARING HENRY & SON, com. mers. 150 Front, h. Brooklyn ',
  'Waring Henry P. grocer, 150 Front, h. 68 Willow, Brooklyn ',
  'Waring Hiram, com. mer. 121 West, h. Harlem ',
  'Waring James, watchman, W. 19th n. Av. 10 ',
  'Waring Jane, widow of James, 168 Lewis _ ',
  'Waring Lewis, cakes, Fulton mkt, h. 257 Rivington ',
  'Waring Philis, (col'd) wid. Wentworth, 201 Church ',
  'Waring Samuel, mer. 14 Cedar ',
  'Waring Samuel J. insp. ins. co. 14 Mer. Exch. h. 206 Henry ',
  'Waring Sarah, tailoress, 279 Grand ',
  'Waring Stephen, mer. Brooklyn, h. 234 E. Broadway ',
  'Waring Stephen H. porterh. 183 South, h. 56 Av. D ',
  'Waring Stephen H. clerk, 324 Fifth ',
  'Waring Thaddeus R. broker, -20 Wall, h. 40 Perry ',
  'Waring William, 14 Cedar , ',
  '/Waring William EF. mer. 14 Cedar ',
  'Waring William H. carman, 89 W. 16th ',
  'Waring H. & Co. flour, 121 West ',
  'Waring W.F. & S. drygoods, 14 Cedar ',
  'Waring & Webster, grocers, 191 West ',
  'Wark David, liquors, 40 West, h. 40 West ',
  'Wark Joseph, police, 62 Greene ',
  'Warlow Benjamin, carman, 2 Macdougal ',
  'Warlow Jacob, carman, 3!4 Hudson ',
  'Warmuth Balthaser, shoemaker, 388 Eighth ',
  'Warne Samuel, com. mer. 88 West, h. Mount Plea- sant, New Jersey ',
  'Warne Thomas, porter, 24 Trinity place ',
  'Warne William, blacksmith, r. 34 Trinity place ',
  'Warner Abrah. butcher, 3 Franklin market, h. 246 Bowery ',
  'Warner Abraham. !5 Beach ',
  'Warner Abrah. B. com. mer. 79 West, h. 205 BV'ker ',
  'Warner Addison K. 157.Henry ',
  '_Warner Albert A. teas, 75 Fulton, bds United States Hotel ',
  'Warner Alexander, herbalist, 107 John ',
  'Warner Alfred, engraver, 3 Wall ',
  'Warner Allen, printer, 103 Bayard ',
  'Waruer Allen C. grocer, 98 Grove, h. 13 Frankfort: ',
  'Warner Andrew, dep. county clerk, 20 City Hall, h. 116 Sullivan . ',
  'Warner Ann, wid. Wm. L. 189 Mulberry cd ',
  'Warner Benj. J. watchcases, 4 Liberty pl, h. Wmshg ']
```

*Figure 3 – Text extracted from sample column, in list format.*

### Analyses

We began our exploration by completing simple analyses, relating to name and occupation popularity, or locative distribution of occupations. These visualizations can be viewed in our appendix (Appendix B, Items 1 – 6). They include:

- Ten most common surnames and occupations in 1859 (Items 1 & 2)
- Spatial distributions of tailors and surnames ("Brown") in 1859 (Items 3 & 4)
- Ten most common first names, male and female (Items 5 & 6)

The more robust analyses – resulting from using the probabilistic gender and ethnicity tagging – made possible various interesting and insightful visualizations. We used a local Moran's I to identify hot and cold spots for the distribution of workplaces for the 5 most common ethnicities. Most businesses are, unsurprisingly, concentrated in the southern tip of Manhattan (today's

Financial District), though over the span of 30 years we see shifts in where the majority of these businesses become centralized. Looking at the British-owned businesses from 1850 to 1875, we see a strong expansion within the Financial District, as well as growth into modern-day DUMBO. There is a prominent dearth of businesses in Midtown, which expands over time.

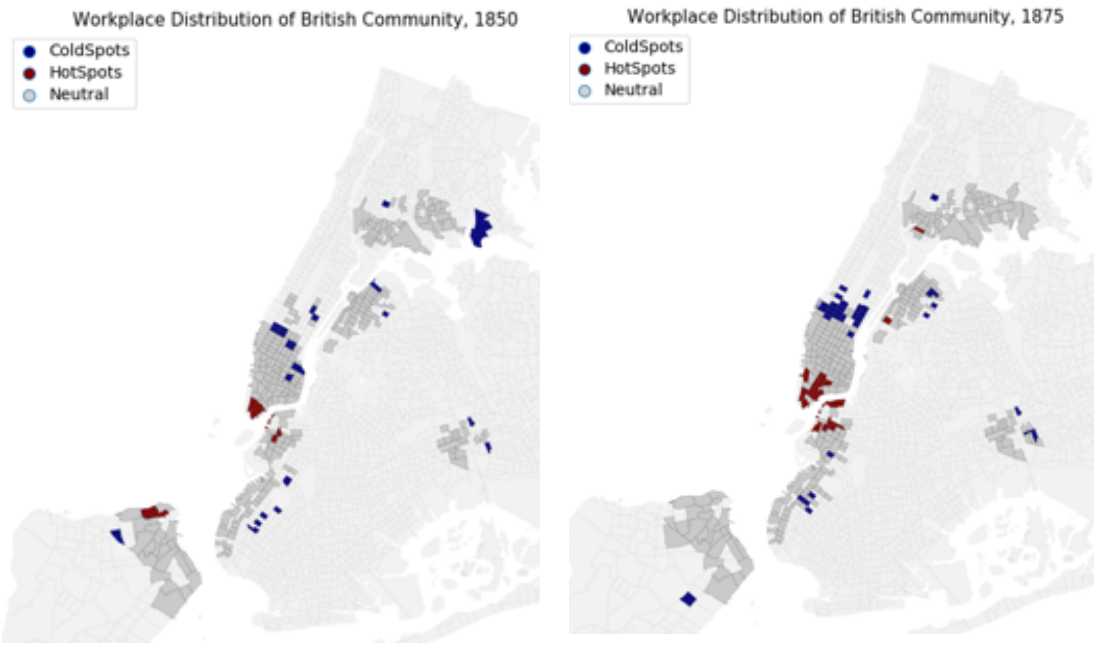



*Figure 4 – Hotspot analysis of British places of work. First image shows the hot and cold spots in 1850, with the second showing the same in 1875. Notice the increase of workers in Downtown Manhattan and Dumbo.*

Additional plots for the years between 1850 and 1875 can be found in the appendix (Appendix B, item 7). Trends that we observed for the communities include the following:

**Germanic**: Concentration of jobs in Financial District, which migrates to Lower East Side (LES) and South Queens. Heavy cold spot along coast of east Brooklyn, which disappears with time. (Appendix B, Item 8)
**Nordic**: Slow growth out of Financial District into LES, with cold spots in Midtown and South Bronx. (Appendix B, Item 9)
**Jewish**: Hot spot originates in Financial District, condenses into large area of LES. Cold spots in Midtown and South Bronx that shrink with time. (Appendix B, Item 10)
**French:** Begins in Financial District, with slow expansion in to Northern Queens and South Bronx. (Appendix B, Item 11)

Considering the gender tags assigned to each entry, we were interested in understanding the rise of the female workforce and observing what kinds of occupations were held by women. “Dressmaker” was consistently the largest percentage of jobs across the three decades (comprising approximately 30% of the job distribution, on average), with “Boarding House” and “Washing” surging in popularity, each claiming ~15% by the 1870s. “Milliner” shrinks considerably over time, while “Sewing” grows (Fig. 5; Appendix B, Item 12).

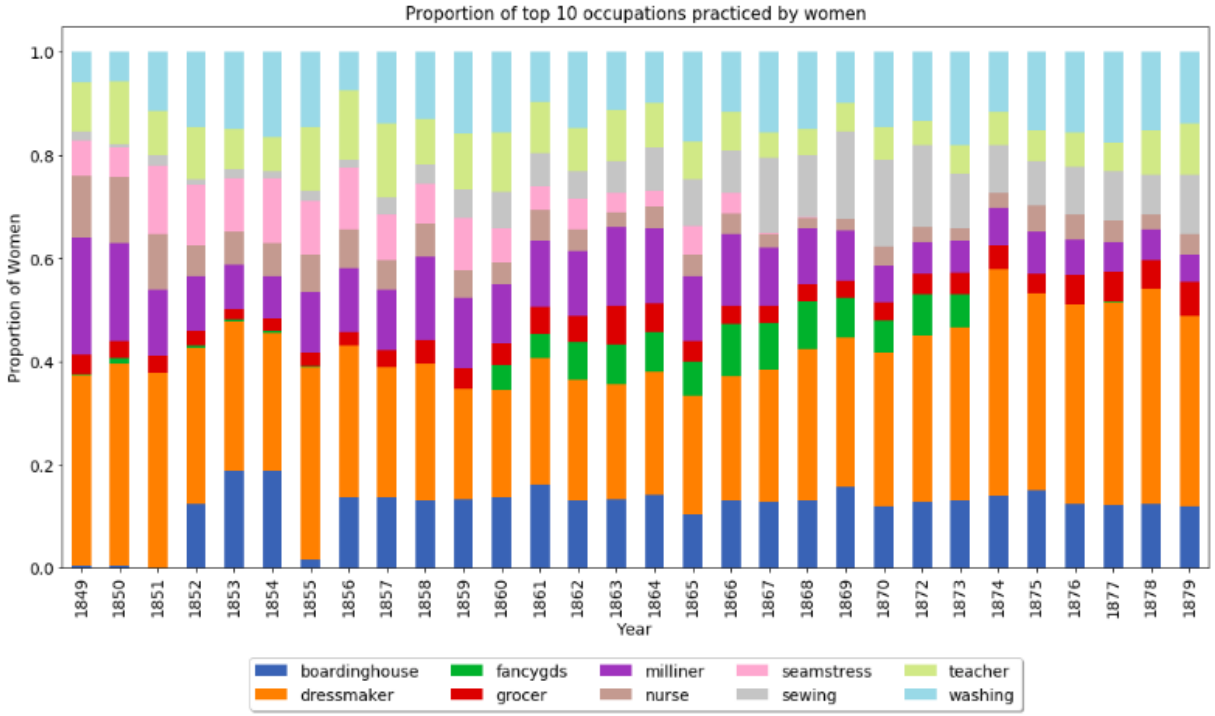



*Figure 5 – Stacked bar chart displaying the percentage distribution of the top occupations for women over the span of 30 years (1849 - 1879).*

The top 10 male occupations remain rather consistent in their distribution over thirty years. We do see some shifts, however: “Grocer” shrinks in percentage, while “Clerk” expands. Quite interestingly, we find what we believe is simply a linguistic shift in what vehicle operators referred to themselves as: “Carman” is a prominent occupation in 1849 (nearly 20% of the male workforce), but is slowly replaced with “Driver”, which grows steadily in percentage year after year (Appendix B, Item 13).

Examining gender representation within various ethnicities, we make a few notable observations. Nearly 60% of working Japanese women were employed as “Dress makers”, which was by far the most prominent profession among all communities of employed women (Appendix B, Item 14); Working as a “Tailor” comprised nearly 40% of the Jewish male workforce (Appendix B, Item 15).

Of course, the analyses that make use of the ethnicity tags must be viewed under the scrutiny of an historic lens. Indeed, while our probabilistic tagging seemingly worked without issue, we cannot avoid the fact that some tags are necessarily incorrect, due to the fact that we are mapping modern descriptions based on nomenclature norms from over 150 years ago.

## Conclusion & Next Steps

The extraction methods developed and tested in this project showed promise in small-scale deployment. Though it is unfortunate that we were unsuccessful in applying these methods to the entire set of directory pages, with further adjustments to make the column-cropping component more generalized, we believe that the novel methods designed over the term of this research will prove valuable in enhancing the overall accuracy of the data. Future work on this project should focus on finding a more universal method of finding the boundaries of each page image to ensure that the OCR extraction method will work across every page.

Additionally, we have only scratched the surface in terms of analyses that can be produced from the extracted data. By further cleaning and tagging the entries, we believe that there are endless possibilities for us to better understand the distribution – spatially, temporally, and constitutionally – of the working people of historical New York City.

## Statement of Collaboration

The OCR extraction and tagging methodology was imagined, designed, and implemented by Matthew Sauter and Linda Lyu. All analyses – including the parsing and cleaning of the initial dataset, tagging of gender and ethnicity, and visualizations – were conducted by Vaidehi Thete and Shelly Yin. Fekade Brook and Matthew Sauter worked on the website design, construction, and maintenance.

# BIBLIOGRAPHY

## Articles & Publications


Bogdani (2018, September 8). "Build a POS tagger with an LSTM using Keras". *Natural Language Processing for Hackers,* accessed at https://nlpforhackers.io/lstm-pos-tagger-keras/

Docparser, "Improving OCR Accuracy With Advanced Image Preprocessing". *Docparser*, accessed at https://docparser.com/blog/improve-ocr-accuracy

Hobby, J. and Ho, T.K. (1997, August). "Enhancing degraded document images via bitmap clustering and averaging". *IEEE Xplore Digital Library*, accessed at https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=619877

Holley, R. (2008, October). "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs". *D-Lib Magazine,* accessed at http://eprints.rclis.org/12908/1/ANDP__How_Good_Can_it_Get.pdf.

Sutton, C. (2016, October 5). "New York Public Library Digitizes 137 Years of New York City Directories". *New York Public Library*, accessed at https://www.nypl.org/blog/2016/09/21/new-york-city-directories-free-online

New York Public Library. *NYPL Labs*, accessed at https://www.nypl.org/collections/labs


## Data & Packages


Ambekar, Anurag and Ward, Charles and Mohammed, Jahangir and Male, Swapna and Skiena, Steven (2009). "Name-ethnicity classification from open sources". *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, p. 49 – 58*. Accessed at https://github.com/appeler/ethnicolr/tree/master/ethnicolr/data/wiki

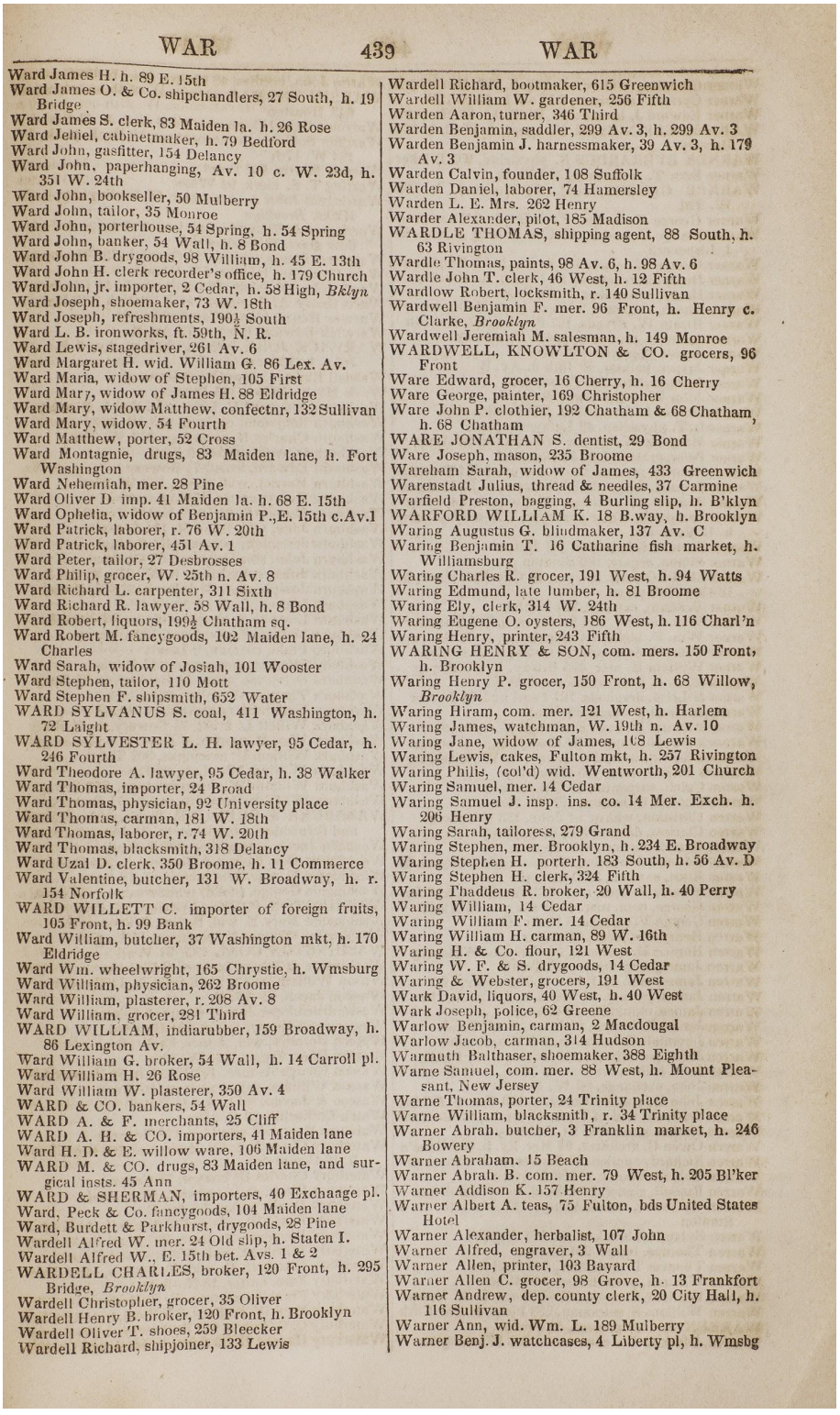
Mullen, L. (2018). "gender: Predict Gender from Names Using Historical Data". R packave version 0.5.2. Accessed at https://www.r-project.org/nosvn/pandoc/gender.html

Wolf, N., Spaan, B. (2018, March 15). "List of NYPL's Digitized City Directories". Github Repository, accessed at https://github.com/nypl-spacetime/city-directories/blob/master/DIRECTORIES.md

Wolf, N., Spaan, B. (2018, March 15). "Space Time NYPL City Directories Dataset". Accessed at http://s3.amazonaws.com/spacetime-nypl-org/datasets/city-directories/city-directories.zip

# APPENDIX A – OCR Extraction

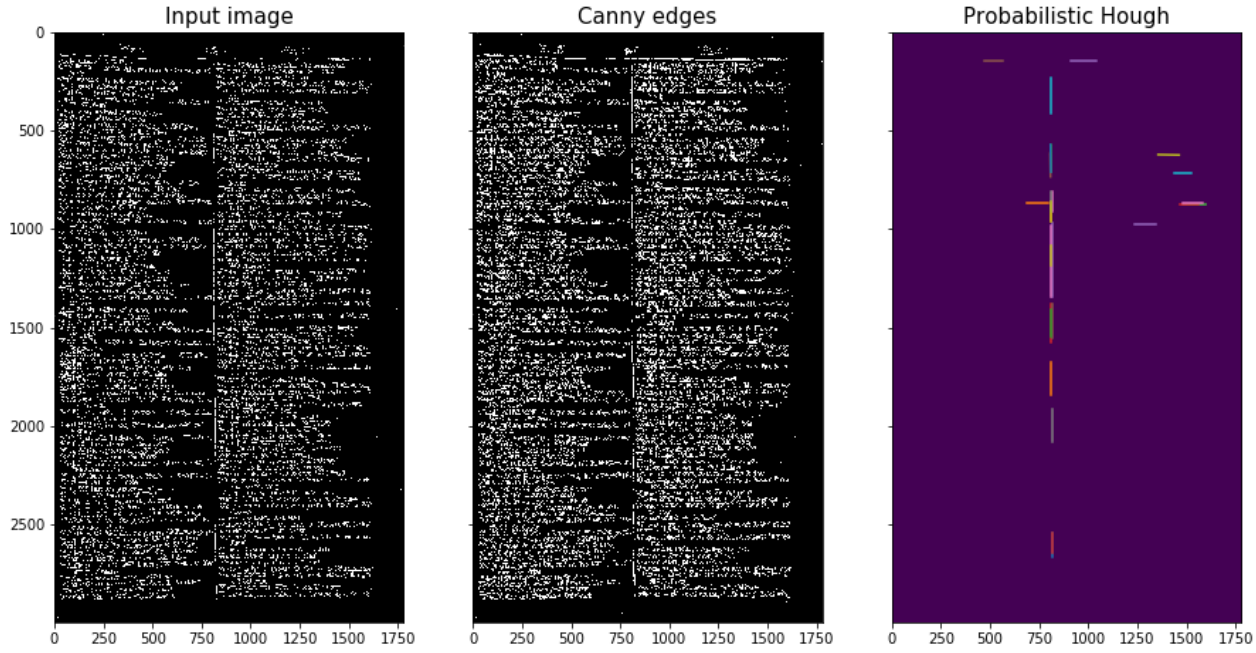
1. Example of two-column page, pulled from the 1849-1850 directory. Note the indented lines, indicating a split entry.



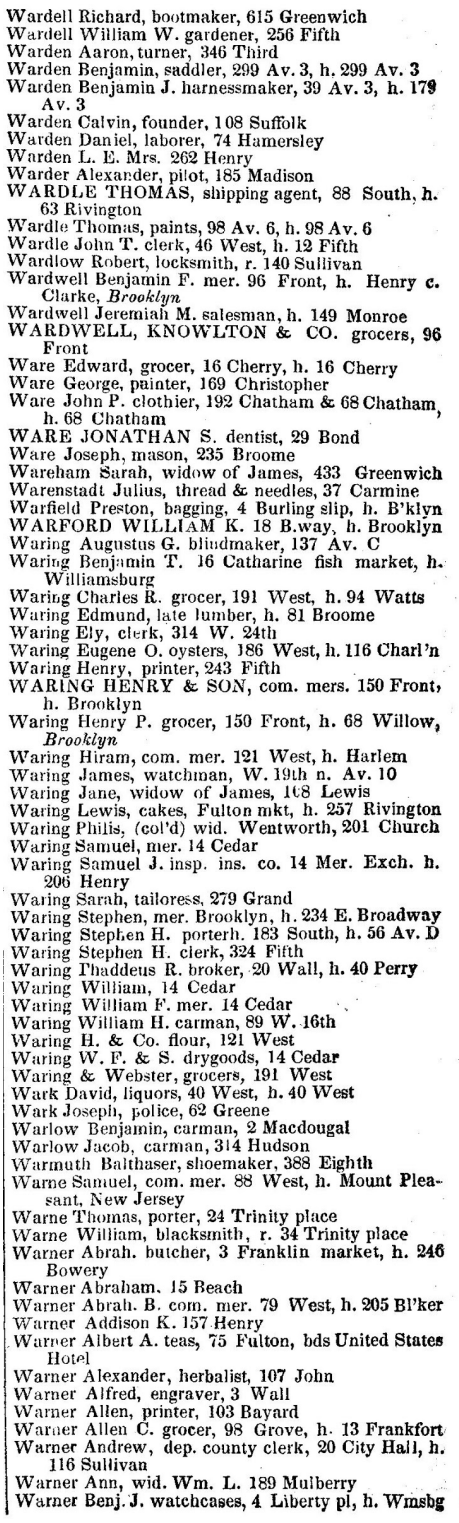


Ward James H. h. 89 E. 15th
Ward James O. & Co. shipchandlers, 27 South, h. 19 Bridge
Ward James S. clerk, 83 Maiden la. h. 26 Rose
Ward Jehiel, cabinetmaker, h. 79 Bedford
Ward John, gasfitter, 154 Delancy
Ward John, paperhanging, Av. 10 c. W. 23d, h. 351 W. 24th
Ward John, bookseller, 50 Mulberry
Ward John, tailor, 35 Monroe
Ward John, porterhouse, 54 Spring, h. 54 Spring
Ward John, banker, 54 Wall, h. 8 Bond
Ward John B. drygoods, 98 William, h. 45 E. 13th
Ward John H. clerk recorder's office, h. 179 Church
Ward John, jr. importer, 2 Cedar, h. 58 High, *Bklyn*
Ward Joseph, shoemaker, 73 W. 18th
Ward Joseph, refreshments, 190½ South
Ward L. B. ironworks, ft. 59th, N. R.
Ward Lewis, stagedriver, 261 Av. 6
Ward Margaret H. wid. William G. 86 Lex. Av.
Ward Maria, widow of Stephen, 105 First
Ward Mary, widow of James H. 88 Eldridge
Ward Mary, widow Matthew, confectnr, 132 Sullivan
Ward Mary, widow, 54 Fourth
Ward Matthew, porter, 52 Cross
Ward Montagnie, drugs, 83 Maiden lane, h. Fort Washington
Ward Nehemiah, mer. 28 Pine
Ward Oliver D imp. 41 Maiden la. h. 68 E. 15th
Ward Ophelia, widow of Benjamin P., E. 15th c. Av. 1
Ward Patrick, laborer, r. 76 W. 20th
Ward Patrick, laborer, 451 Av. 1
Ward Peter, tailor, 27 Desbrosses
Ward Philip, grocer, W. 25th n. Av. 8
Ward Richard L. carpenter, 311 Sixth
Ward Richard R. lawyer, 58 Wall, h. 8 Bond
Ward Robert, liquors, 199½ Chatham sq.
Ward Robert M. fancygoods, 102 Maiden lane, h. 24 Charles
Ward Sarah, widow of Josiah, 101 Wooster
Ward Stephen, tailor, 110 Mott
Ward Stephen F. shipsmith, 652 Water
WARD SYLVANUS S. coal, 411 Washington, h. 72 Laight
WARD SYLVESTER L. H. lawyer, 95 Cedar, h. 246 Fourth
Ward Theodore A. lawyer, 95 Cedar, h. 38 Walker
Ward Thomas, importer, 24 Broad
Ward Thomas, physician, 92 University place
Ward Thomas, carman, 181 W. 18th
Ward Thomas, laborer, r. 74 W. 20th
Ward Thomas, blacksmith, 318 Delancy
Ward Uzal D. clerk, 350 Broome, h. 11 Commerce
Ward Valentine, butcher, 131 W. Broadway, h. r. 154 Norfolk
WARD WILLETT C. importer of foreign fruits, 105 Front, h. 99 Bank
Ward William, butcher, 37 Washington mkt, h. 170 Eldridge
Ward Wm. wheelwright, 165 Chrystie, h. Wmsburg
Ward William, physician, 262 Broome
Ward William, plasterer, r. 208 Av. 8
Ward William, grocer, 281 Third
WARD WILLIAM, indiarubber, 159 Broadway, h. 86 Lexington Av.
Ward William G. broker, 54 Wall, h. 14 Carroll pl.
Ward William H. 26 Rose
Ward William W. plasterer, 350 Av. 4
WARD & CO. bankers, 54 Wall
WARD A. & F. merchants, 25 Cliff
WARD A. H. & CO. importers, 41 Maiden lane
Ward H. D. & E. willow ware, 106 Maiden lane
WARD M. & CO. drugs, 83 Maiden lane, and surgical insts. 45 Ann
WARD & SHERMAN, importers, 40 Exchange pl.
Ward, Peck & Co. fancygoods, 104 Maiden lane
Ward, Burdett & Parkhurst, drygoods, 28 Pine
Wardell Alfred W. mer. 24 Old slip, h. Staten I.
Wardell Alfred W., E. 15th bet. Avs. 1 & 2
WARDELL CHARLES, broker, 120 Front, h. 295 Bridge, *Brooklyn*
Wardell Christopher, grocer, 35 Oliver
Wardell Henry B. broker, 120 Front, h. Brooklyn
Wardell Oliver T. shoes, 259 Bleecker
Wardell Richard, shipjoiner, 133 Lewis
Wardell Richard, bootmaker, 615 Greenwich
Wardell William W. gardener, 256 Fifth
Warden Aaron, turner, 346 Third
Warden Benjamin, saddler, 299 Av. 3, h. 299 Av. 3
Warden Benjamin J. harnessmaker, 39 Av. 3, h. 179 Av. 3
Warden Calvin, founder, 108 Suffolk
Warden Daniel, laborer, 74 Hamersley
Warden L. E. Mrs. 262 Henry
Warder Alexander, pilot, 185 Madison
WARDLE THOMAS, shipping agent, 88 South, h. 63 Rivington
Wardle Thomas, paints, 98 Av. 6, h. 98 Av. 6
Wardle John T. clerk, 46 West, h. 12 Fifth
Wardlow Robert, locksmith, r. 140 Sullivan
Wardwell Benjamin F. mer. 96 Front, h. Henry c. Clarke, *Brooklyn*
Wardwell Jeremiah M. salesman, h. 149 Monroe
WARDWELL, KNOWLTON & CO. grocers, 96 Front
Ware Edward, grocer, 16 Cherry, h. 16 Cherry
Ware George, painter, 169 Christopher
Ware John P. clothier, 192 Chatham & 68 Chatham, h. 68 Chatham
WARE JONATHAN S. dentist, 29 Bond
Ware Joseph, mason, 235 Broome
Wareham Sarah, widow of James, 433 Greenwich
Warenstadt Julius, thread & needles, 37 Carmine
Warfield Preston, bagging, 4 Burling slip, h. B'klyn
WARFORD WILLIAM K. 18 B.way, h. Brooklyn
Waring Augustus G. bliudmaker, 137 Av. C
Waring Benjamin T. 16 Catharine fish market, h. Williamsburg
Waring Charles R. grocer, 191 West, h. 94 Watts
Waring Edmund, late lumber, h. 81 Broome
Waring Ely, clerk, 314 W. 24th
Waring Eugene O. oysters, 186 West, h. 116 Charl'n
Waring Henry, printer, 243 Fifth
WARING HENRY & SON, com. mers. 150 Front, h. Brooklyn
Waring Henry P. grocer, 150 Front, h. 68 Willow, *Brooklyn*
Waring Hiram, com. mer. 121 West, h. Harlem
Waring James, watchman, W. 19th n. Av. 10
Waring Jane, widow of James, 108 Lewis
Waring Lewis, cakes, Fulton mkt, h. 257 Rivington
Waring Philis, (col'd) wid. Wentworth, 201 Church
Waring Samuel, mer. 14 Cedar
Waring Samuel J. insp. ins. co. 14 Mer. Exch. h. 206 Henry
Waring Sarah, tailoress, 279 Grand
Waring Stephen, mer. Brooklyn, h. 234 E. Broadway
Waring Stephen H. porterh. 183 South, h. 56 Av. D
Waring Stephen H. clerk, 324 Fifth
Waring Thaddeus R. broker, 20 Wall, h. 40 Perry
Waring William, 14 Cedar
Waring William F. mer. 14 Cedar
Waring William H. carman, 89 W. 16th
Waring H. & Co. flour, 121 West
Waring W. F. & S. drygoods, 14 Cedar
Waring & Webster, grocers, 191 West
Wark David, liquors, 40 West, h. 40 West
Wark Joseph, police, 62 Greene
Warlow Benjamin, carman, 2 Macdougal
Warlow Jacob, carman, 314 Hudson
Warmuth Balthaser, shoemaker, 388 Eighth
Warne Samuel, com. mer. 88 West, h. Mount Pleasant, New Jersey
Warne Thomas, porter, 24 Trinity place
Warne William, blacksmith, r. 34 Trinity place
Warner Abrah. butcher, 3 Franklin market, h. 246 Bowery
Warner Abraham, 15 Beach
Warner Abrah. B. com. mer. 79 West, h. 205 Bl'ker
Warner Addison K. 157 Henry
Warner Albert A. teas, 75 Fulton, bds United States Hotel
Warner Alexander, herbalist, 107 John
Warner Alfred, engraver, 3 Wall
Warner Allen, printer, 103 Bayard
Warner Allen C. grocer, 98 Grove, h. 13 Frankfort
Warner Andrew, dep. county clerk, 20 City Hall, h. 116 Sullivan
Warner Ann, wid. Wm. L. 189 Mulberry
Warner Benj. J. watchcases, 4 Liberty pl, h. Wmsbg

2. Preprocessed page image (left) in preparation for Canny edge detection (middle), and resulting column line from Hough transform (right)



3. (Left) Cropped column prepared for OCR extraction
4. (Right) Text extracted from cropped column, in resulting list format

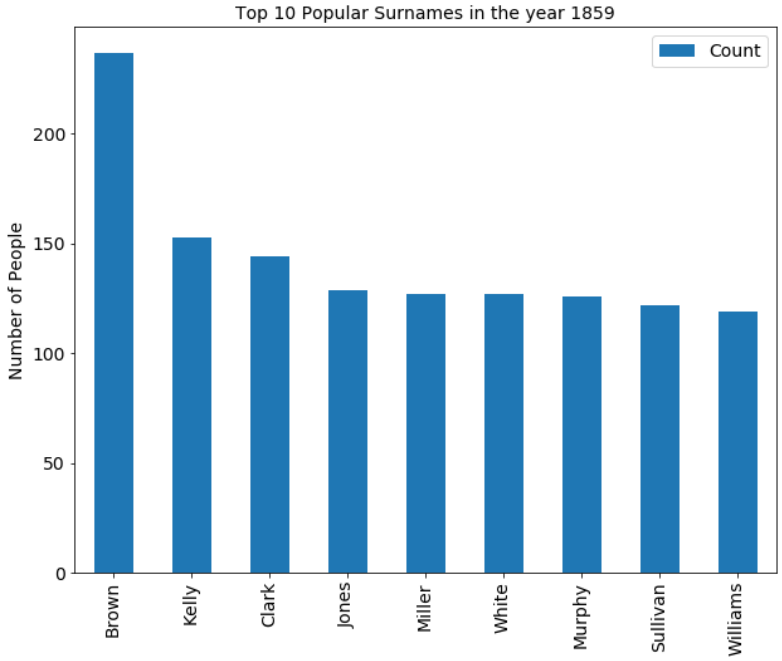


Wardell Richard, bootmaker, 615 Greenwich
Wardell William W. gardener, 256 Fifth
Warden Aaron, turner, 346 Third
Warden Benjamin, saddler, 299 Av. 3, h. 299 Av. 3
Warden Benjamin J. harnessmaker, 39 Av. 3, h. 179 Av. 3
Warden Calvin, founder, 108 Suffolk
Warden Daniel, laborer, 74 Hamersley
Warden L. E. Mrs. 262 Henry
Warder Alexander, pilot, 185 Madison
WARDLE THOMAS, shipping agent, 88 South, h. 63 Rivington
Wardle Thomas, paints, 98 Av. 6, h. 98 Av. 6
Wardle John T. clerk, 46 West, h. 12 Fifth
Wardlow Robert, locksmith, r. 140 Sullivan
Wardwell Benjamin F. mer. 96 Front, h. Henry c. Clarke, *Brooklyn*
Wardwell Jeremiah M. salesman, h. 149 Monroe
WARDWELL, KNOWLTON & CO. grocers, 96 Front
Ware Edward, grocer, 16 Cherry, h. 16 Cherry
Ware George, painter, 169 Christopher
Ware John P. clothier, 192 Chatham & 68 Chatham, h. 68 Chatham
WARE JONATHAN S. dentist, 29 Bond
Ware Joseph, mason, 235 Broome
Wareham Sarah, widow of James, 433 Greenwich
Warenstadt Julius, thread & needles, 37 Carmine
Warfield Preston, bagging, 4 Burling slip, h. B'klyn
WARFORD WILLIAM K. 18 B.way, h. Brooklyn
Waring Augustus G. bliudmaker, 137 Av. C
Waring Benjamin T. 16 Catharine fish market, h. Williamsburg
Waring Charles R. grocer, 191 West, h. 94 Watts
Waring Edmund, late lumber, h. 81 Broome
Waring Ely, clerk, 314 W. 24th
Waring Eugene O. oysters, 186 West, h. 116 Charl'n
Waring Henry, printer, 243 Fifth
WARING HENRY & SON, com. mers. 150 Front, h. Brooklyn
Waring Henry P. grocer, 150 Front, h. 68 Willow, *Brooklyn*
Waring Hiram, com. mer. 121 West, h. Harlem
Waring James, watchman, W. 19th n. Av. 10
Waring Jane, widow of James, 168 Lewis
Waring Lewis, cakes, Fulton mkt, h. 257 Rivington
Waring Philis, (col'd) wid. Wentworth, 201 Church
Waring Samuel, mer. 14 Cedar
Waring Samuel J. insp. ins. co. 14 Mer. Exch. h. 206 Henry
Waring Sarah, tailoress, 279 Grand
Waring Stephen, mer. Brooklyn, h. 234 E. Broadway
Waring Stephen H. porterh. 183 South, h. 56 Av. D
Waring Stephen H. clerk, 324 Fifth
Waring Thaddeus R. broker, 20 Wall, h. 40 Perry
Waring William, 14 Cedar
Waring William F. mer. 14 Cedar
Waring William H. carman, 89 W. 16th
Waring H. & Co. flour, 121 West
Waring W. F. & S. drygoods, 14 Cedar
Waring & Webster, grocers, 191 West
Wark David, liquors, 40 West, h. 40 West
Wark Joseph, police, 62 Greene
Warlow Benjamin, carman, 2 Macdougal
Warlow Jacob, carman, 314 Hudson
Warmuth Balthaser, shoemaker, 388 Eighth
Warne Samuel, com. mer. 88 West, h. Mount Pleasant, New Jersey
Warne Thomas, porter, 24 Trinity place
Warne William, blacksmith, r. 34 Trinity place
Warner Abrah. butcher, 3 Franklin market, h. 246 Bowery
Warner Abraham. 15 Beach
Warner Abrah. B. com. mer. 79 West, h. 205 Bl'ker
Warner Addison K. 157 Henry
Warner Albert A. teas, 75 Fulton, bds United States Hotel
Warner Alexander, herbalist, 107 John
Warner Alfred, engraver, 3 Wall
Warner Allen, printer, 103 Bayard
Warner Allen C. grocer, 98 Grove, h. 13 Frankfort
Warner Andrew, dep. county clerk, 20 City Hall, h. 116 Sullivan
Warner Ann, wid. Wm. L. 189 Mulberry
Warner Benj. J. watchcases, 4 Liberty pl, h. Wmsbg

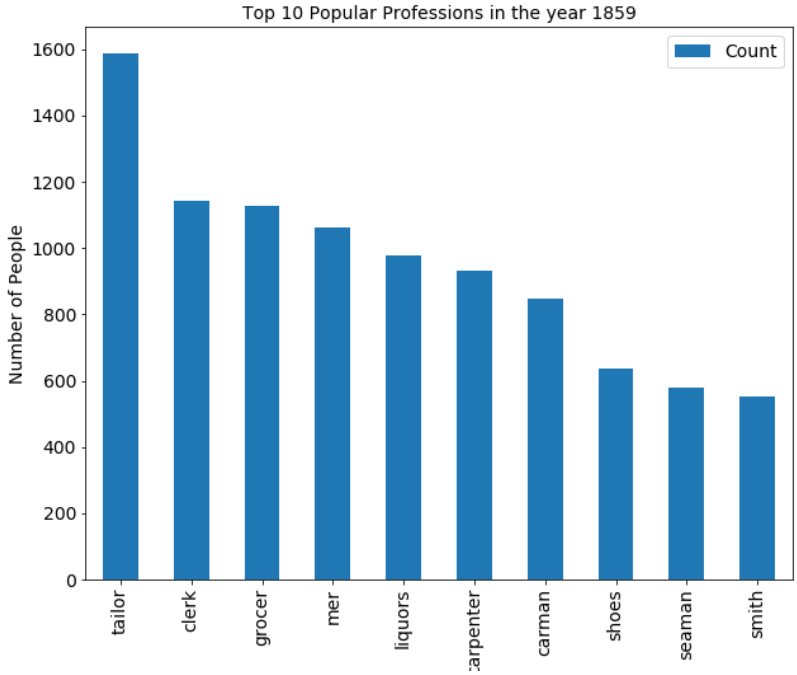
```
[ 'Wardell Richard, bootmaker, 615 Greenwich ',
  'Wardell William W. gardener, 256 Fifth ',
  'Warden Aaron, turner, 346 Third ',
  'Warden Benjamin, saddler, 299 Av. 3, h. 299 Av. 3 ',
  'Warden Benjamin J. harnessmaker, 39 Av. 3, h. 179 v. ',
  'Warden Calvin, founder, 108 Suffolk ',
  'Warden Daniel, laborer, 74 Hamersley ',
  'Warden L. Ii. Mrs. 262 Henry ',
  'Warder Alexander, pilot, 185 Madison ',
  'WARDLE THOMAS, shipping agent, 88 South, h. 63 Rivington . ',
  'Wardle Thomas, paints, 98 Av. 6, h. 98 Av. 6 ',
  'Wardle John T. clerk, 46 West, h. 12 Fifth ',
  'Wardlow Robert, locksmith, r. 140 Sullivan ',
  'Wardwell Benjamin F. mer. 96 Front, h. Henry ¢ Clarke, Brooklyn',
  'Wardwell Jeremiah M. salesman, h. 149 Monroe ',
  'WARDWELL, KNOWLTON & CO. grocers, 96 ront ',
  'Ware Edward, grocer, 16 Cherry, h. 16 Cherry ',
  'Ware George, painter, 169 Christopher ',
  'Ware John P. clothier, 192 Chatham & 68 Chatham h. 68 Chatham, ',
  'WARE JONATHAN &. dentist, 29 Bond ',
  'Ware Joseph, mason, 235 Broome ',
  'Wareham Sarah, widow of James, 433 Greenwich ',
  'Warenstadt Julius, thread & needles, 37 Carmine ',
  'Warfield Preston, bagging, 4 Burling slip, h. B'klyn ',
  'WARFORD WILLIAM K. 18 B.way, h. Brooklyn ',
  'Waring Augustus G. bliudmaker, 137 Av..C ',
  'Waring Benjamin T, 16 Catharine fish market, h. Williamsburg a ',
  'Waring Charlies R. grocer, 191 West, h. 94 Watts ',
  'Waring Edmund, late lumber, h. 81 Broome ',
  'Waring Ely, clerk, 314 W. 24th ',
  'Waring Eugene O. oysters, 386 West, h. 116 Charl'n ',
  'Waring Henry, printer, 243 Fifth . . ',
  'WARING HENRY & SON, com. mers. 150 Front, h. Brooklyn ',
  'Waring Henry P. grocer, 150 Front, h. 68 Willow, Brooklyn ',
  'Waring Hiram, com. mer. 121 West, h. Harlem ',
  'Waring James, watchman, W. 19th n. Av. 10 ',
  'Waring Jane, widow of James, 168 Lewis _ ',
  'Waring Lewis, cakes, Fulton mkt, h. 257 Rivington ',
  'Waring Philis, (col'd) wid. Wentworth, 201 Church ',
  'Waring Samuel, mer. 14 Cedar ',
  'Waring Samuel J. insp. ins. co. 14 Mer. Exch. h. 206 Henry ',
  'Waring Sarah, tailoress, 279 Grand ',
  'Waring Stephen, mer. Brooklyn, h. 234 E. Broadway ',
  'Waring Stephen H. porterh. 183 South, h. 56 Av. D ',
  'Waring Stephen H. clerk, 324 Fifth ',
  'Waring Thaddeus R. broker, -20 Wall, h. 40 Perry ',
  'Waring William, 14 Cedar , ',
  '/Waring William EF. mer. 14 Cedar ',
  'Waring William H. carman, 89 W. 16th ',
  'Waring H. & Co. flour, 121 West ',
  'Waring W.F. & S. drygoods, 14 Cedar ',
  'Waring & Webster, grocers, 191 West ',
  'Wark David, liquors, 40 West, h. 40 West ',
  'Wark Joseph, police, 62 Greene ',
  'Warlow Benjamin, carman, 2 Macdougal ',
  'Warlow Jacob, carman, 3!4 Hudson ',
  'Warmuth Balthaser, shoemaker, 388 Eighth ',
  'Warne Samuel, com. mer. 88 West, h. Mount Plea- sant, New Jersey ',
  'Warne Thomas, porter, 24 Trinity place ',
  'Warne William, blacksmith, r. 34 Trinity place ',
  'Warner Abrah. butcher, 3 Franklin market, h. 246 Bowery ',
  'Warner Abraham. !5 Beach ',
  'Warner Abrah. B. com. mer. 79 West, h. 205 BV'ker ',
  'Warner Addison K. 157.Henry ',
  '_Warner Albert A. teas, 75 Fulton, bds United States Hotel ',
  'Warner Alexander, herbalist, 107 John ',
  'Warner Alfred, engraver, 3 Wall ',
  'Warner Allen, printer, 103 Bayard ',
  'Waruer Allen C. grocer, 98 Grove, h. 13 Frankfort: ',
  'Warner Andrew, dep. county clerk, 20 City Hall, h. 116 Sullivan . ',
  'Warner Ann, wid. Wm. L. 189 Mulberry cd ',
  'Warner Benj. J. watchcases, 4 Liberty pl, h. Wmshg ']
```

## APPENDIX B – Visualizations

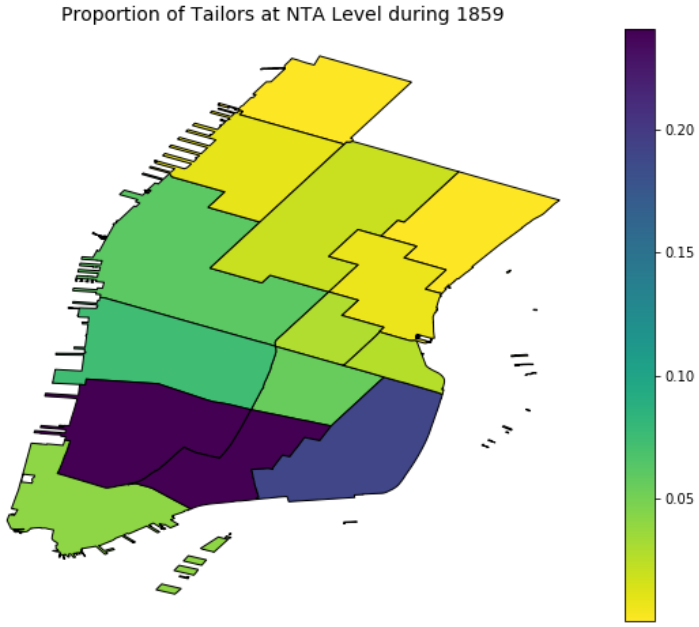
1. Top 10 Popular Surnames, 1859



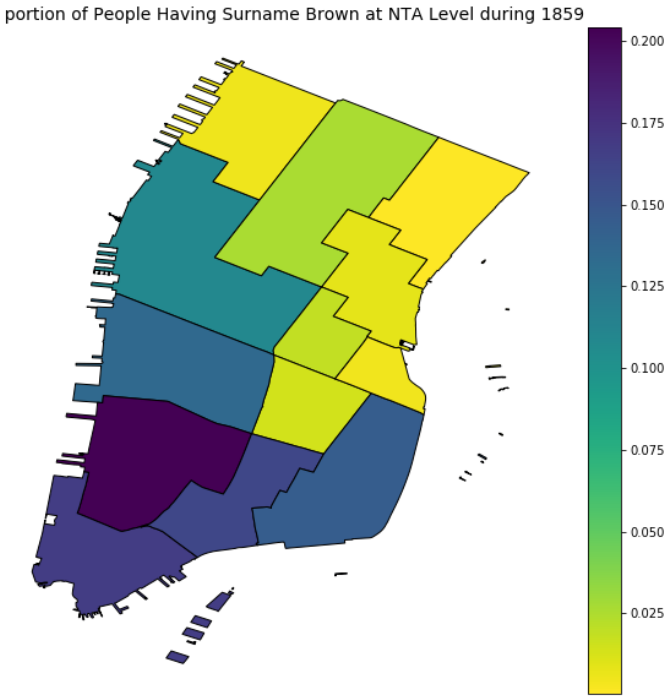

2. Top 10 Popular Professions, 1859



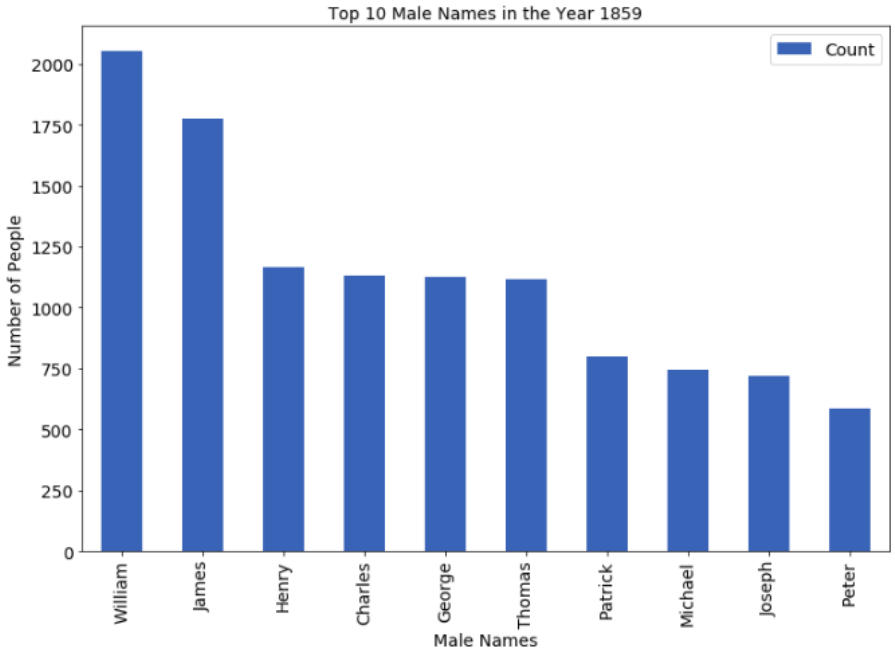
3. Proportion of Tailors at NTA Level, 1859



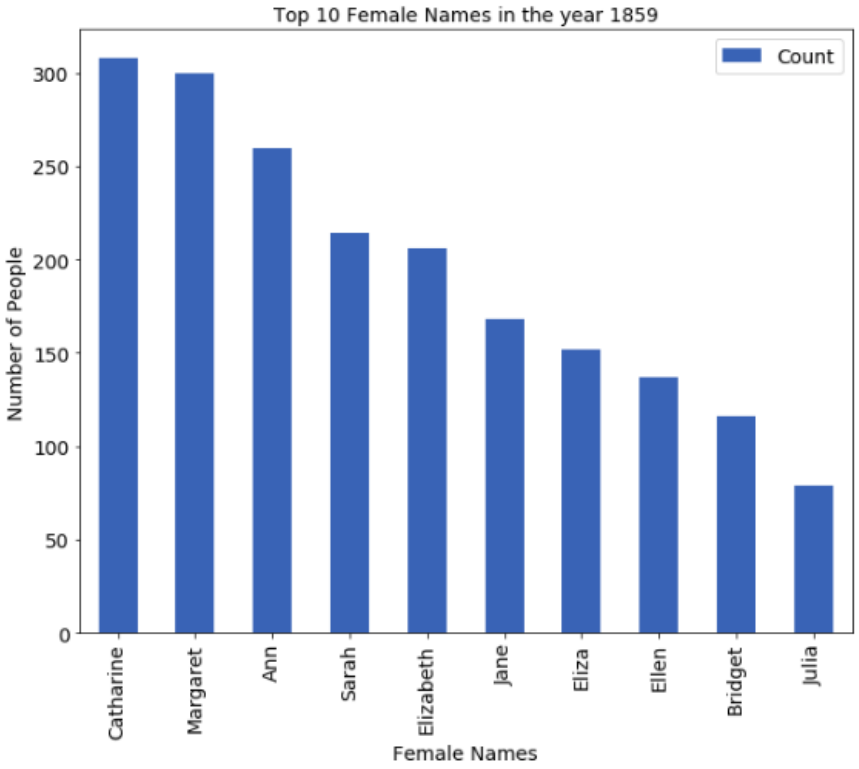

4. Proportion of employees with surname “Brown”, 1859



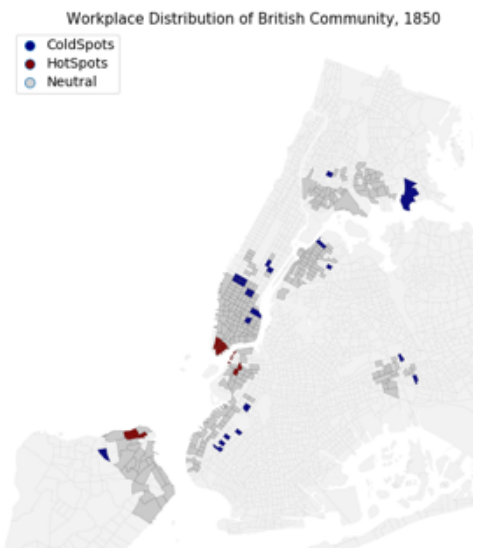
5. Most Common Male First Names, 1859

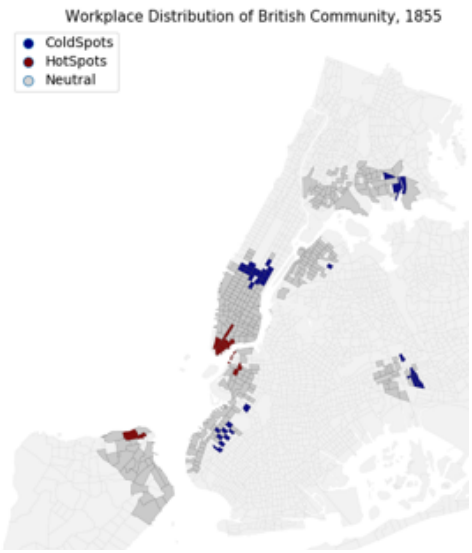


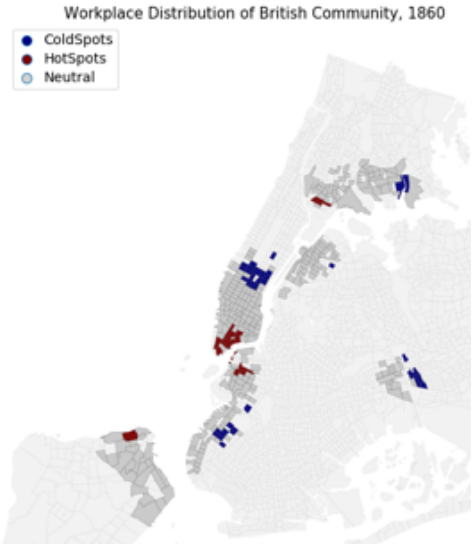

6. Most Common Female First Names, 1859

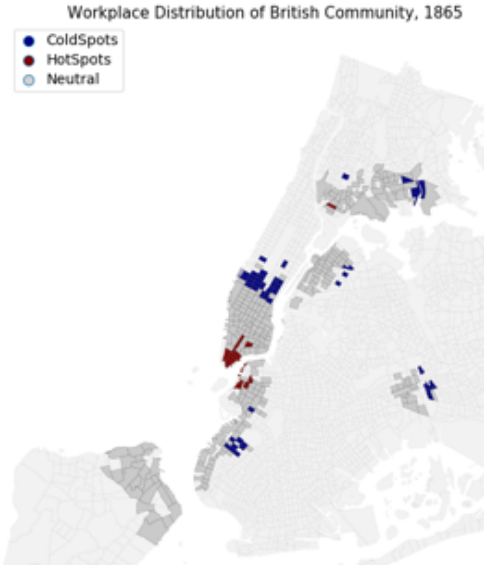


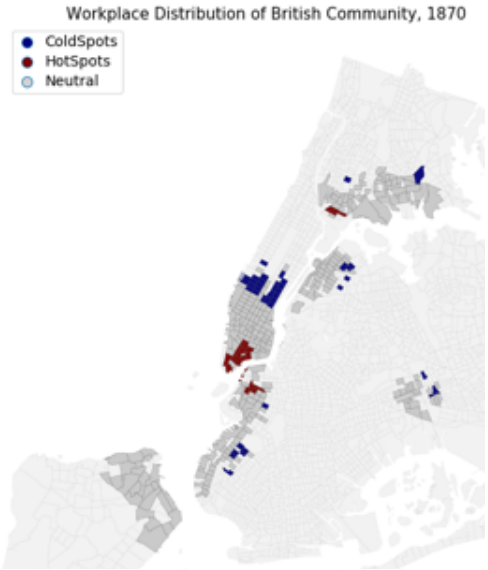
7. British Community Hot/Cold Spot Workplace Distributions, 1850 - 1875

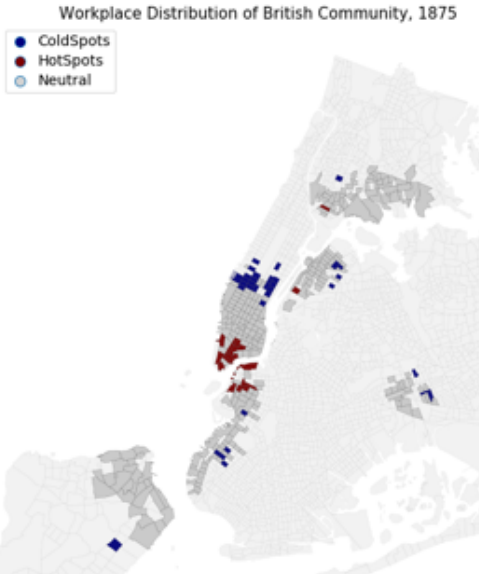


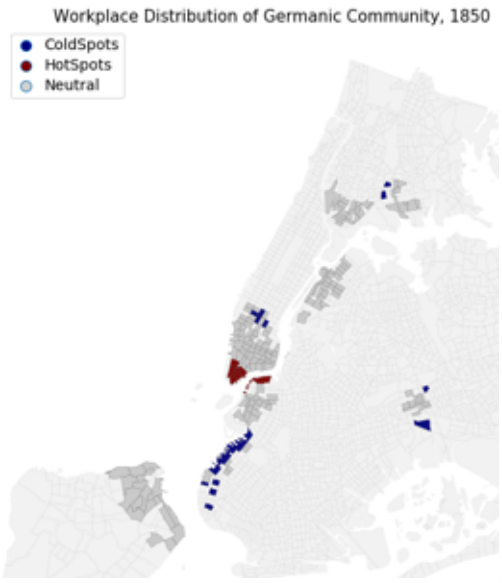



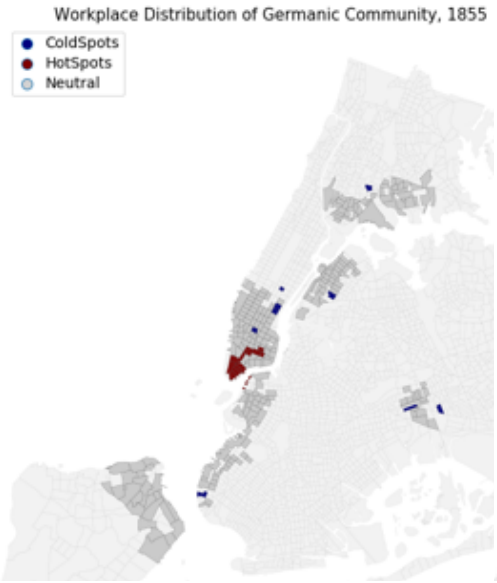



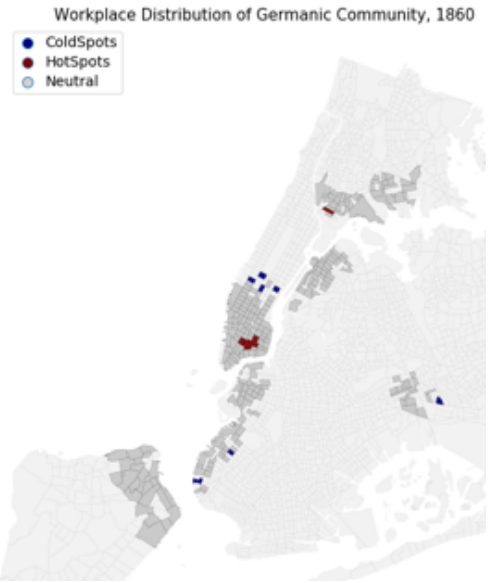



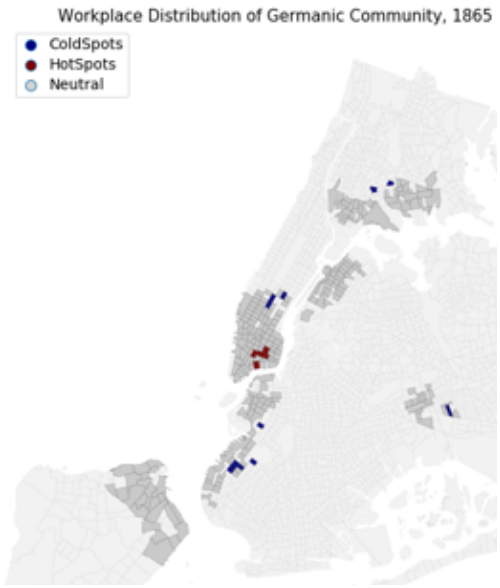



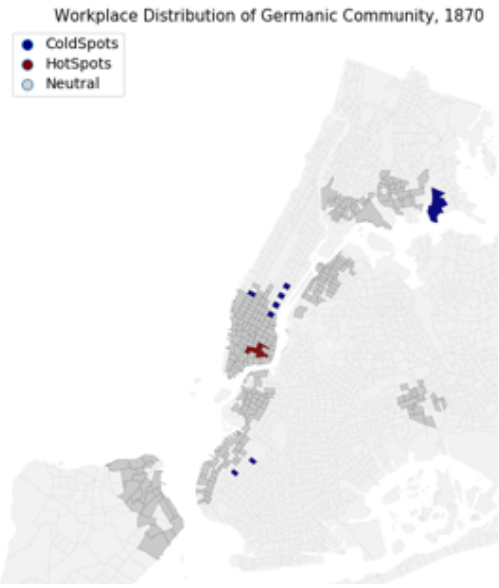
8. Germanic Community Hot/Cold Spot Workplace Distributions, 1850 – 1875

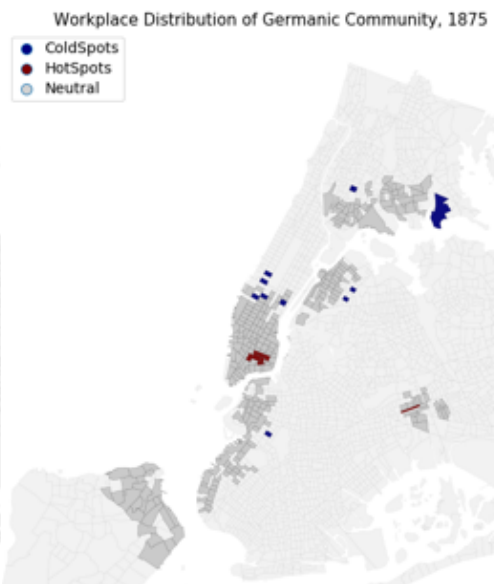


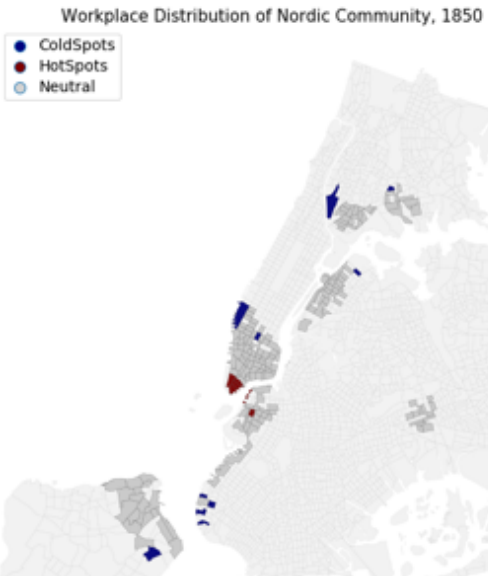



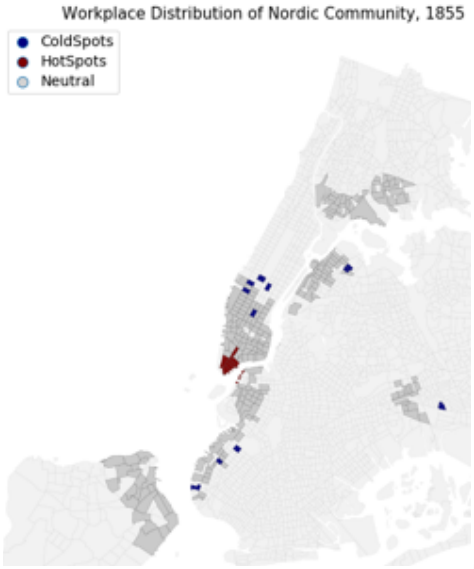



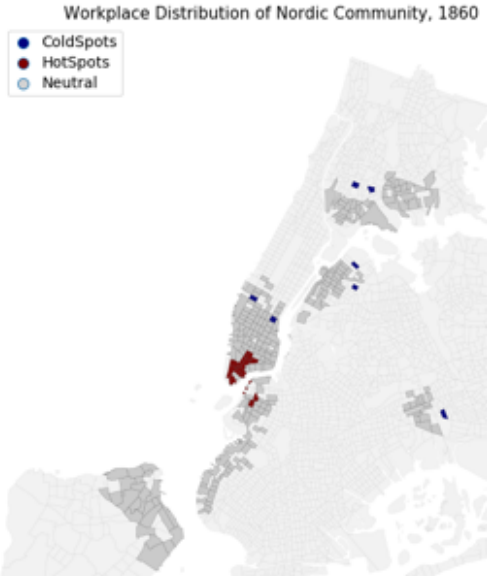



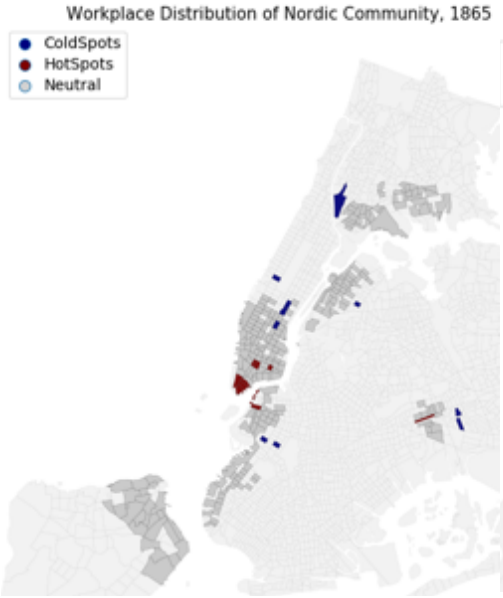



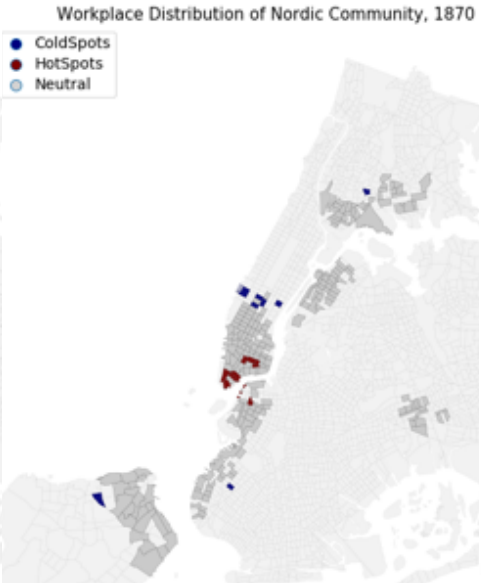
9. Nordic Community Hot/Cold Spot Workplace Distributions, 1850 – 1875

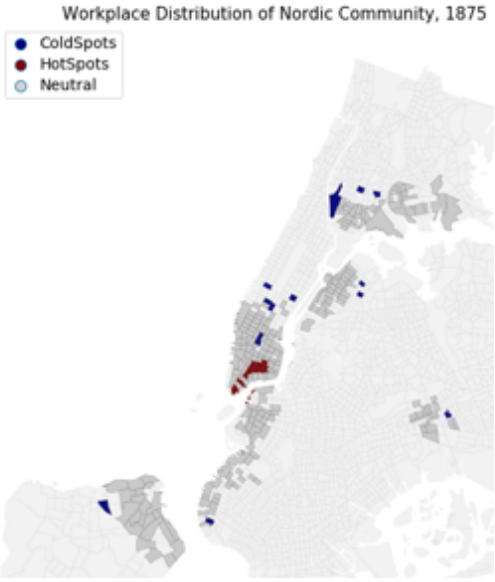


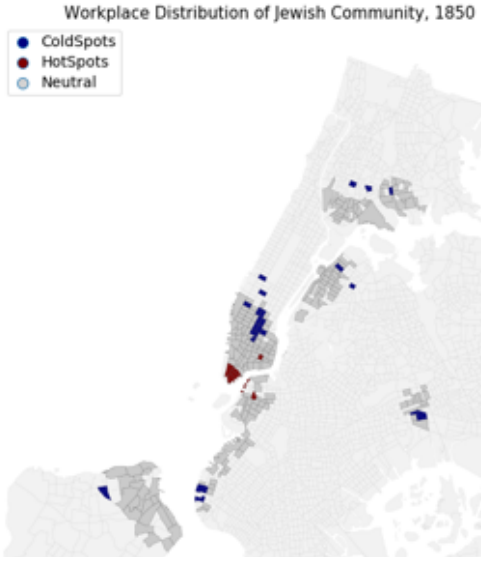



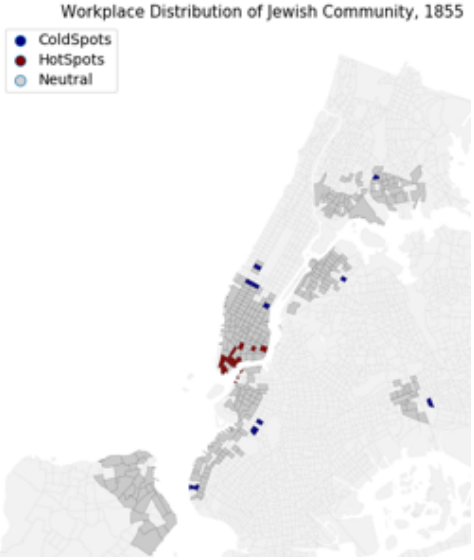



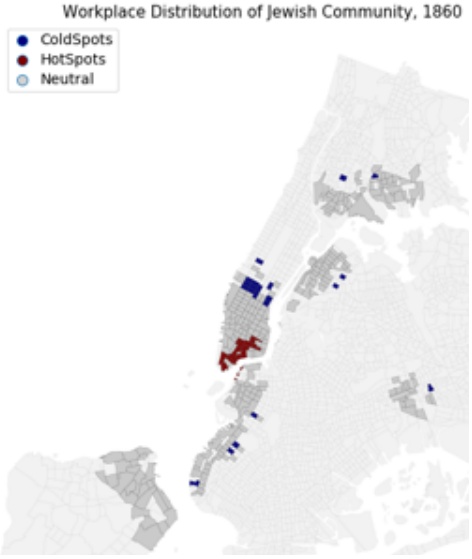



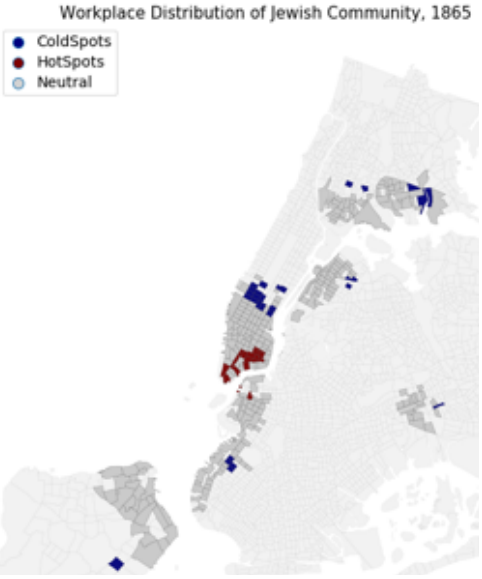



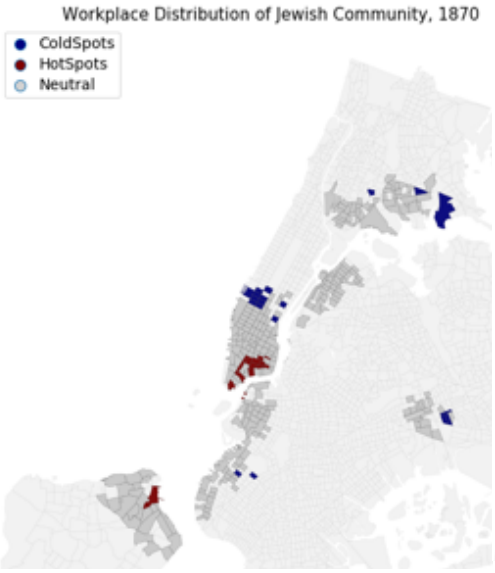
10. Jewish Community Hot/Cold Spot Workplace Distributions, 1850 – 1875

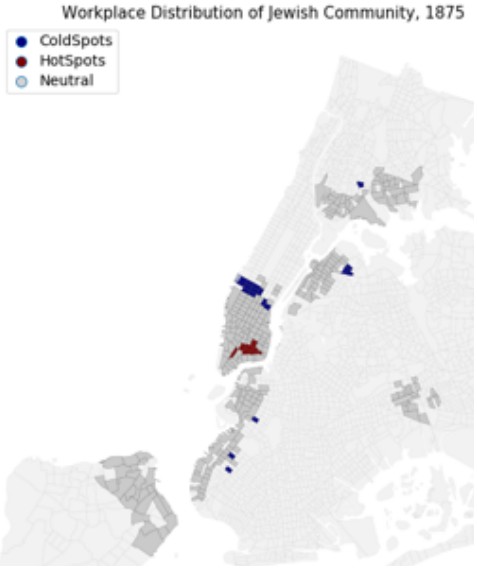


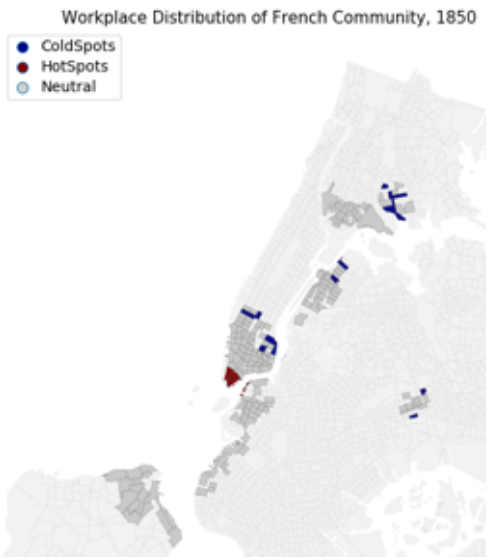



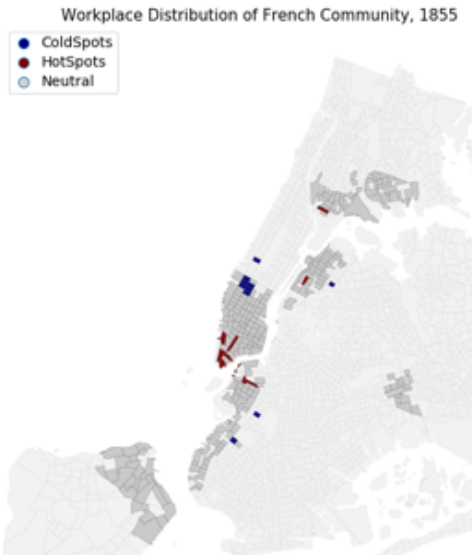



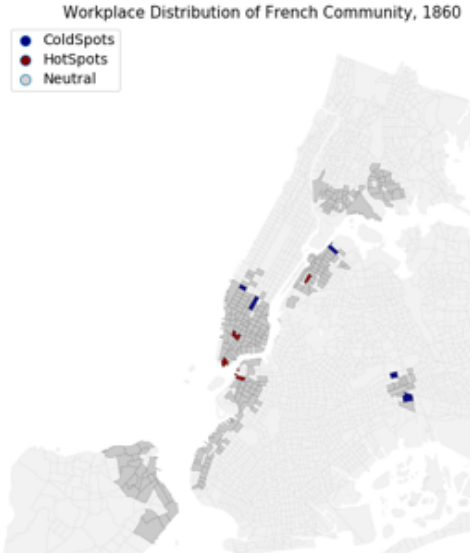



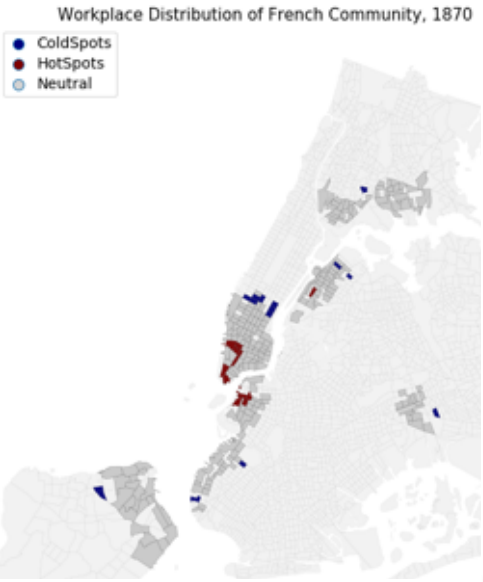
11. French Community Hot/Cold Spot Workplace Distributions, 1850 – 1875

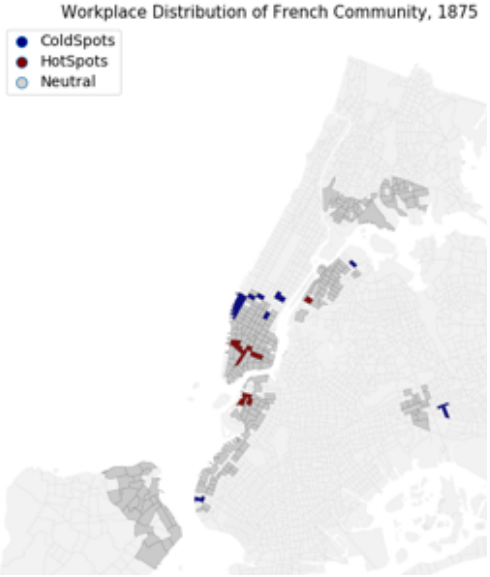


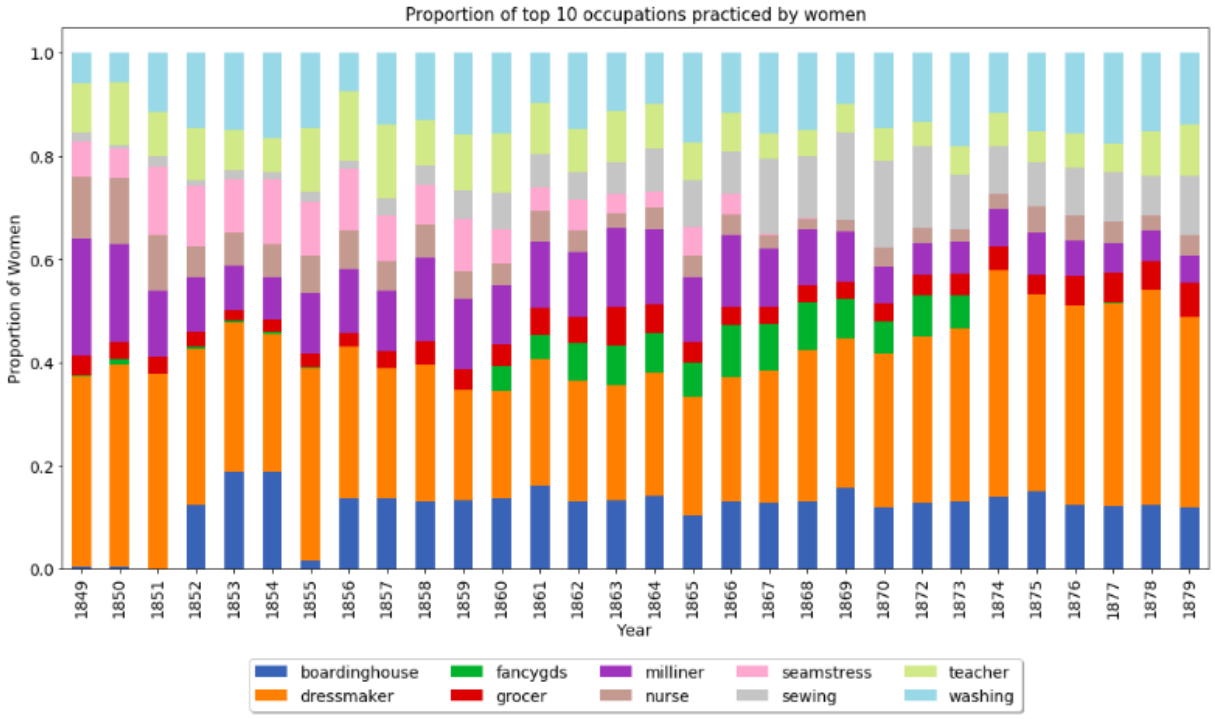





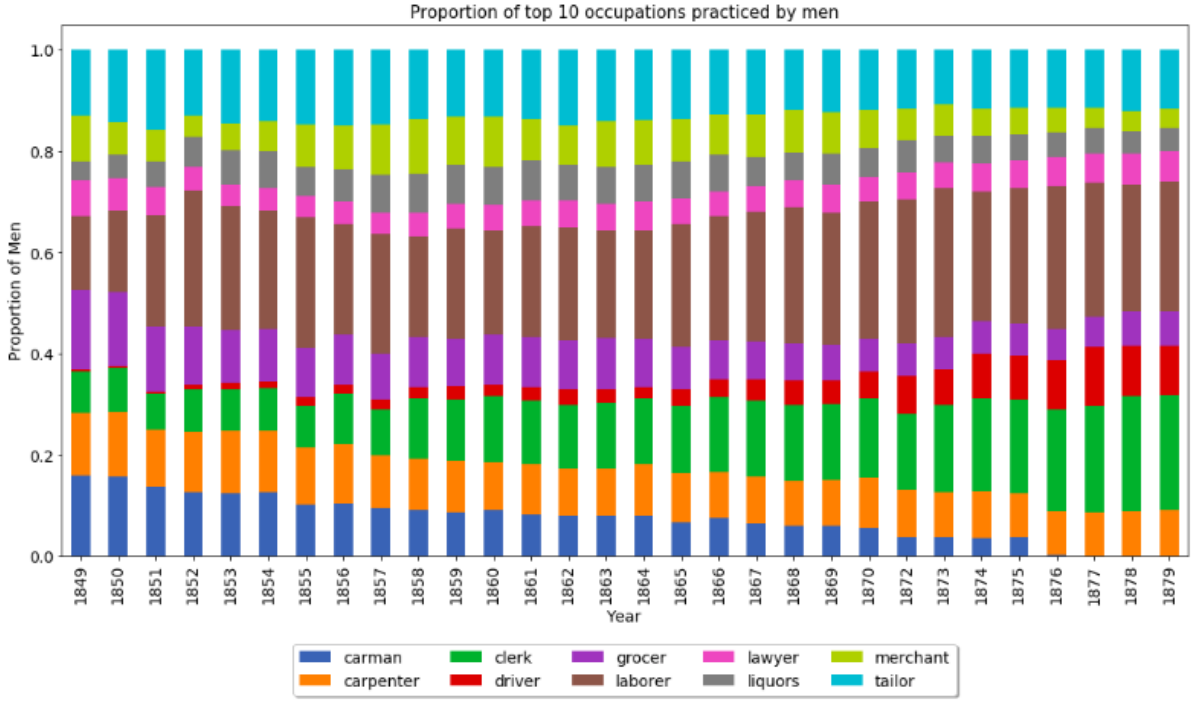


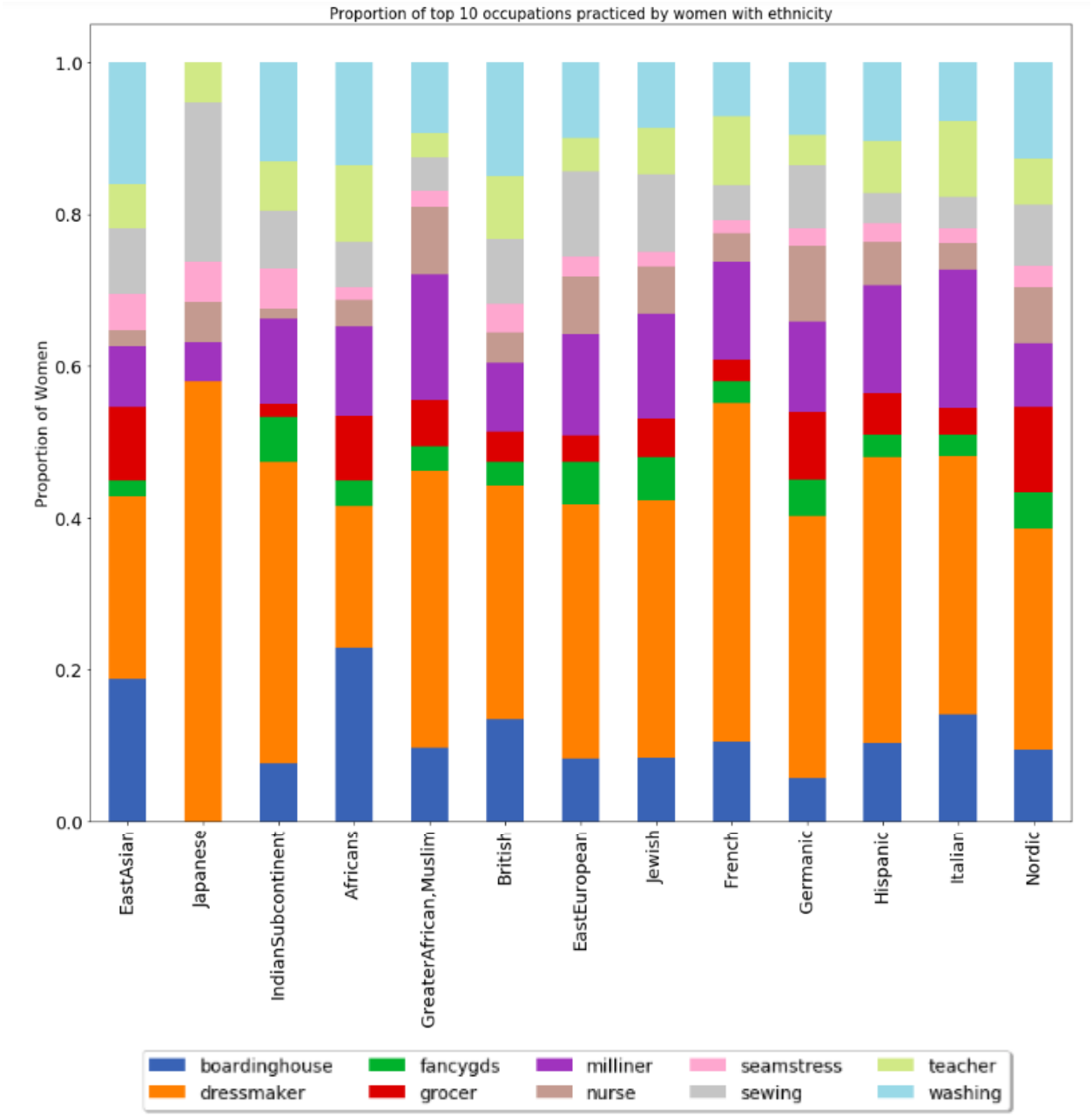
12. Proportion of top occupations practiced by women, 1849 – 1879




13. Proportion of top occupations practiced by men, 1849 – 1879



14. Proportion of top occupations practiced by women per ethnicity



15. Proportion of top occupations practiced by men per ethnicity