Raoul Herskovits — Sohee Park — CPSC 110 — December 10, 2024

Who's on the Docks?

Employment on the Red Hook Waterfront, 1900-1940

Project Members: Raoul Herskovits Yale College '25 (Urban Studies)

Drive Link — Website Link

I. Introduction and Project Statement

Developers often cite 'local jobs' as a key benefit of industrial projects. This narrative, which suggests that industrial employers bring significant employment opportunities to urban areas, is rooted in a historical image of urban industry fostering local community growth through employment. But how accurate is this narrative? To what extent was waterfront employment truly localized, or were other factors at play? This project attempts to answer a part of that question through a case study of Red Hook, Brooklyn, by looking at other factors which may have influenced employment.

Red Hook not only has a long history of maritime shipping and a currently operational container terminal, but it has recently been slated for another redevelopment, possibly into a new industrial use. This historical narrative of local jobs has once again come to the forefront as rhetoric around local impact begins to fly. The historical significance of this local job narrative is especially pungent in Red Hook, where maritime industry has been such a big part of the area's identity, but literature and data analysis on parts of this history are lacking. It is well documented that European immigrant communities formed around the Red Hook waterfront in the 1800's during the city's first big industrial boom, but this locality may not translate to the entire length of the industrial timeline.

Project Question: in this region, were specific demographic groups disproportionately excluded from waterfront jobs?

The microdata I am using is vast, so I have chosen to isolate a single variable for classification, the occupation of 'Longshoreman/Stevedore,' to examine with classification models. As can be seen in the ipython Notebook, I create many other feature columns to extend the possibilities of aggregation and visualization, but the scope of this project is primarily the classification of Longshoremen and Stevedores (cargo laborers on the docks) based on demographic data such as race, ethnicity, immigrant status, and group quarters status.

Through my EDA and machine learning analysis, I find that Italian neighborhood residents were much more likely to work on the docks than members of other ethnicities. This includes Black residents, who were often excluded from this kind of labor and were much more likely to work service jobs, as well as other ethnic Whites, who often held more desirable clerical jobs than their Italian counterparts. Some other populations, such as Irish, Scandinavian, and Puerto Rican communities, were privy to dock labor employment, but they hardly compare to the Italian dominance of the docks.

II. Background: Literature and Context

The existing literature on dockworker hiring practices in New York paint a nuanced picture. Many actors were at play on the New York waterfront, making for a muddy environment: the International Longshoreman's Association (ILA) union, like many unions at the time, was inward facing—prioritizing its members' job security over the well-being of the industry—and also corrupt. Historical accounts suggest that the ILA collaborated with organized crime groups and hiring bosses to exert significant control over the waterfront. The infamous 'shape-up' hiring method, rife with extortion and discrimination, dominated the docks in the first half of the 20th century; this method of hiring involved prospective workers standing around the hiring boss, who would choose which workers would work a certain job that day on the spot. Of course, prearrangements could allow for the extortion of workers in exchange for better chances. Testimonials portray this process as particularly demeaning for Black longshoremen, who faced systemic exclusion and were far less likely to be selected. In the 1950's, city government began to regulate dockworker employment more strictly, creating an active registry of more constant workers, but literature on this 'decasualization' also suggests that it only served to make the industry harder to break into for new arrivals. Unfortunately, the literature goes less in depth on the period of interest for this project, 1900–1950, often focusing on post-industrial fallout or 1800's economic boom. A few sources such as a collection of Puerto Rican oral histories, and an archive from the ILA itself on its history of discrimination explore some of the less covered marginal histories of this time period.

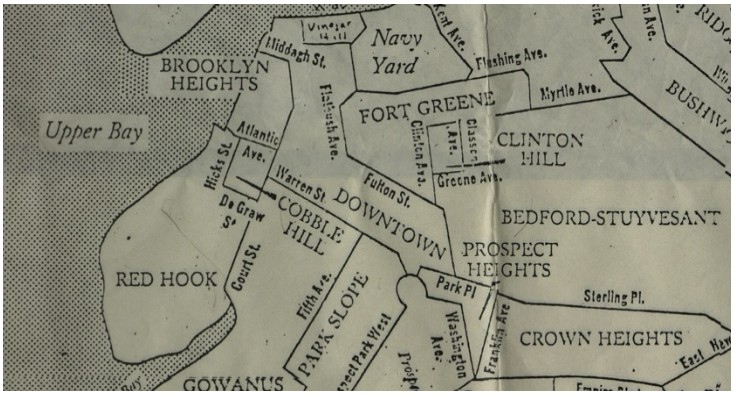


Figure 1. Map of Brooklyn Neighborhoods circa 1950. Brooklyn Historical Society.

This project's site of interest was once simply called 'Red Hook' or 'South Brooklyn.' Currently, it contains the neighborhoods of Red Hook, Carroll Gardens, Cobble Hill, and Columbia Street Waterfront.

III. Data Acquisition, Cleaning and Arranging

My data is an extract from the IPUMS Full Count census dataset, selecting for a variety of variables from the five decennial full count censuses 1900–1940, isolating solely Kings County (Brooklyn), New York. For more details on the extract, see the codebook in the GitHub for google drive folder.

The majority of cleansing work in my dataframe has come in the form of creating new columns to convert data to Boolean values (i.e., 'get_dummies' and similar operations), creating new categories based on codes (such as creating the category of 'Italian' based on family's country of origin), as well as creating functions to allow for easy trimming of data frames by dropping often useless rows without completely eliminating them from our working data frames.

The second half of my data arranging process involved loading my shapefiles and joining them with my dataframe into five geopandas 'GeoDataFrames,' one for each decade, which allowed for easy spatial plotting. These enumeration district labels allow us to create a trimmed dataframe that isolates this region of Brooklyn, which we will be our primary dataframe going forward. They also allow for grouping, aggregation, and plotting by district, but that is outside the scope of this project and is only touched on in the Exploratory Data Analysis.

I ran into very few issues during the data cleaning process, but there were certainly some roadblocks. For instance, the 'is_immigrant' Boolean column was defined using the census category 'Immigration Status'. I assumed this would be a straightforward categorization, but upon plotting values based on this variable, I discovered a serious discrepancy in the 1940 dataset. When I checked the general population, the same flip happened, meaning in that 'Immigration Status' category, there had been some error on the part of the census takers which resulted in very few responses of the correct code.

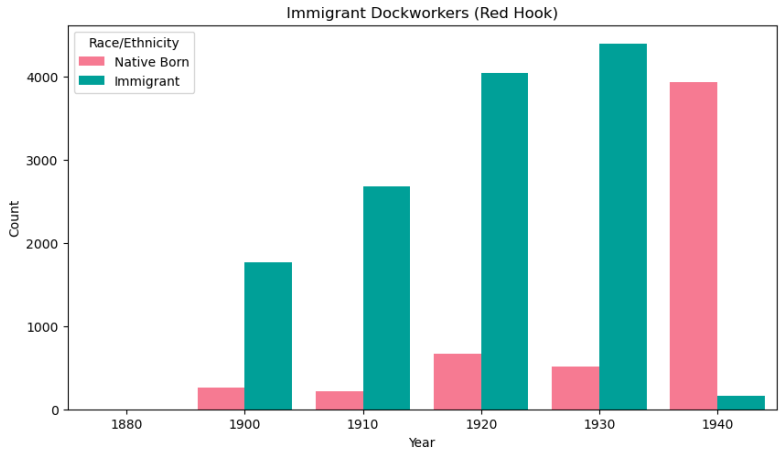


Figure 2. The initial categorization. Immigrant population plummets in 1940.

To address this issue, an additional clause was included in the definition of 'is_immigrant,' checking the 'birthplace' entry for foreign birth and allowing that to fulfill the condition as well.

```
df['ISIMMIGRANT'] = ((df['NATIVITY']==5)|(df['BPL']>99))
```

Trying the same visualization, we see the problem is remedied:

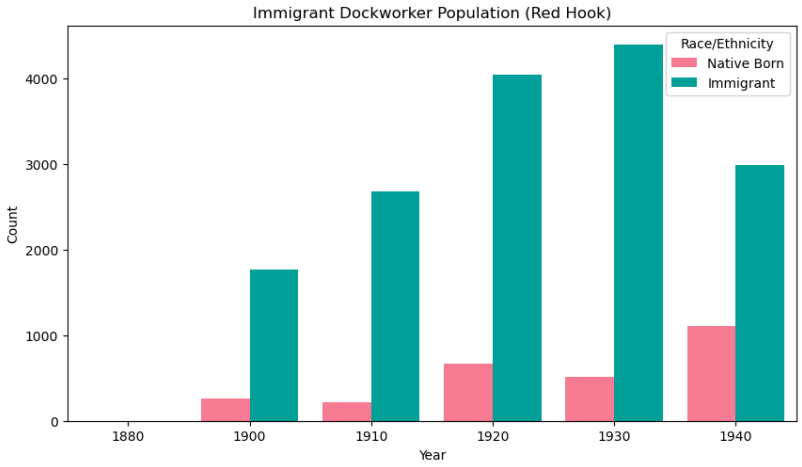


Figure 3. The corrected categorization

.

IV. Exploratory Data Analysis

Using the enumeration district shapefiles, a smaller dataframe was created isolating just Red Hook and allowing for comparisons between the neighborhood and the borough of Brooklyn in its entirety. One thing to note is that the 1880 census has been included in the large Brooklyn dataset for added context. In the Red Hook dataframe, it is removed, since the district names making up the area have not been identified. A good initial point of comparison and visualization are general population trends:

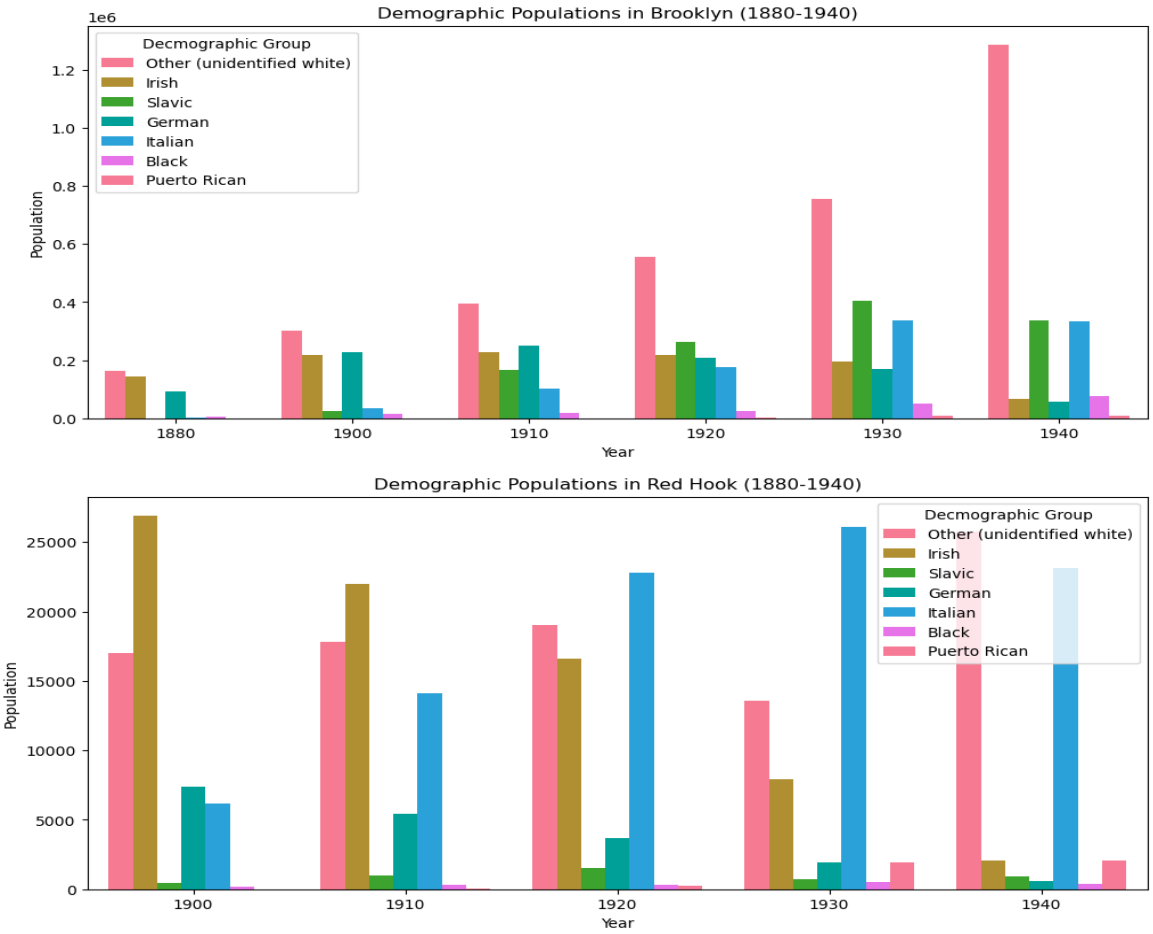


Figure 4

Here I ran into one of my main issues when it comes to demographic categories. Despite differentiating the major ethnicities of white Brooklynites (Irish, Italian, German, and Slavic), as well as the minority group of interest (Black and Puerto Rican), there was still a large population of White people that are not categorized. I wanted to keep my number of categories manageable, but I also didn't want the 'uncategorized' column to be that significant.

To find a middle ground, I added all Slavic countries of origin to my 'Slavic' category, as well as adding a 'Scandinavian' category which included those countries. Since Polish and Scandinavian populations were significant at the time, this additional grouping helped parse out the other population. Additionally, I split the 'Other' category into Immigrant and Native to add an additional layer of detail. After these categorizations, the population graph looks like this:

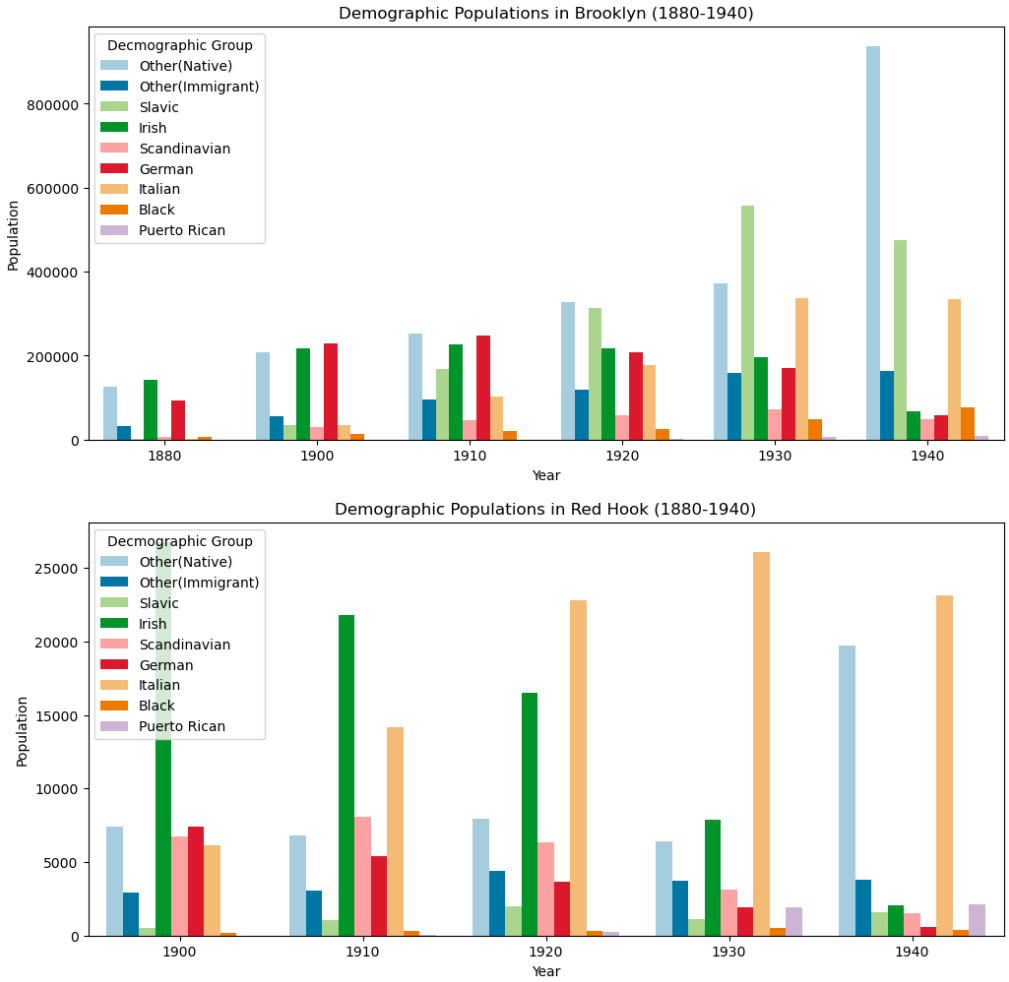


Figure 5

We notice clear trends in both Brooklyn as a whole as well as Red Hook. In Brooklyn, the Black, Slavic, Italian, and Scandinavian rise, while the German and Irish populations taper down towards the end. It makes sense that the Slavic bar is so high, given it contains Russian and Polish immigrants, as well as others from Eastern Europe.

In Red Hook, things are more straightforward. The Irish and Italian populations seem to perfectly invert each other, the one rising while the other falls. The Germain population falls away completely, while the Puerto Rican population crops up in the 1930's, and the Black and Slavic populations consistently remained minorities throughout the period. the 1940 spike in 'Other (Native)' should also be noted in both Brooklyn and Red Hook, indicating the movement of other white Americans moving to Brooklyn from elsewhere in the country

Next, looking at longshoremen/stevedoring jobs vs more general waterfront-related labor.

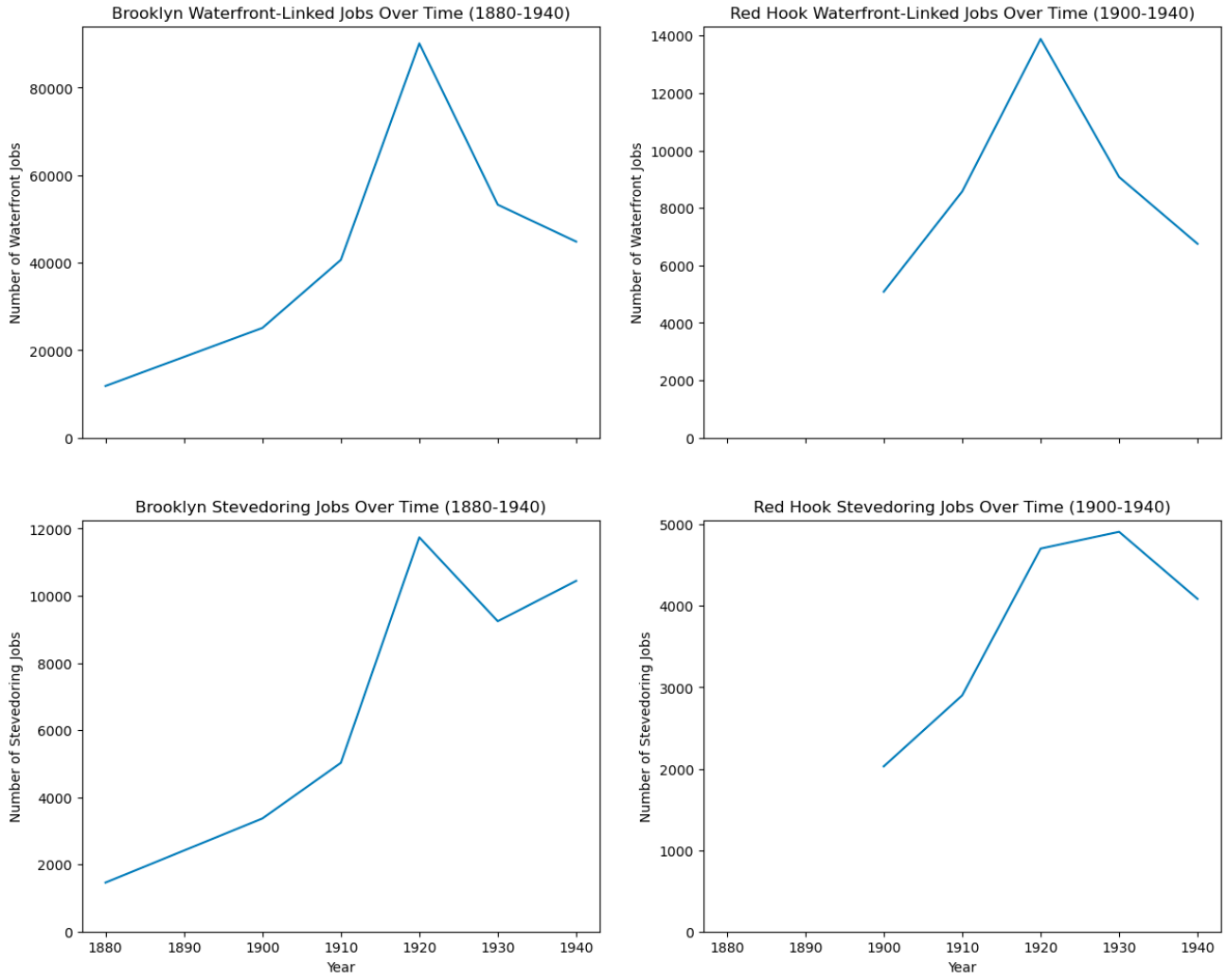


Figure 6

It looks like general waterfront jobs (more details about how I made that category in the notebook) peak in both Red Hook and Brooklyn generally around 1920, which makes sense since the first World War brought many accessible jobs to the waterfront.

Designated longshoreman/stevedore jobs, however, seemed to have more longevity after the war ended and even into the Great Depression, which implies a more constant demand.

In general, the sheer increase in the number of these jobs from 1900 to 1920 (doubling to even quadrupling in number) is striking. It could indicate several things: first, a general boom in waterfront activity, and second, a large number of part-time, unstable or ‘casual’ waterfront jobs, which aligns with descriptions of hiring practices such as the shape-up. I believe both to have impacted these numbers.

We can also graph who was doing these longshoreman/stevedore jobs:

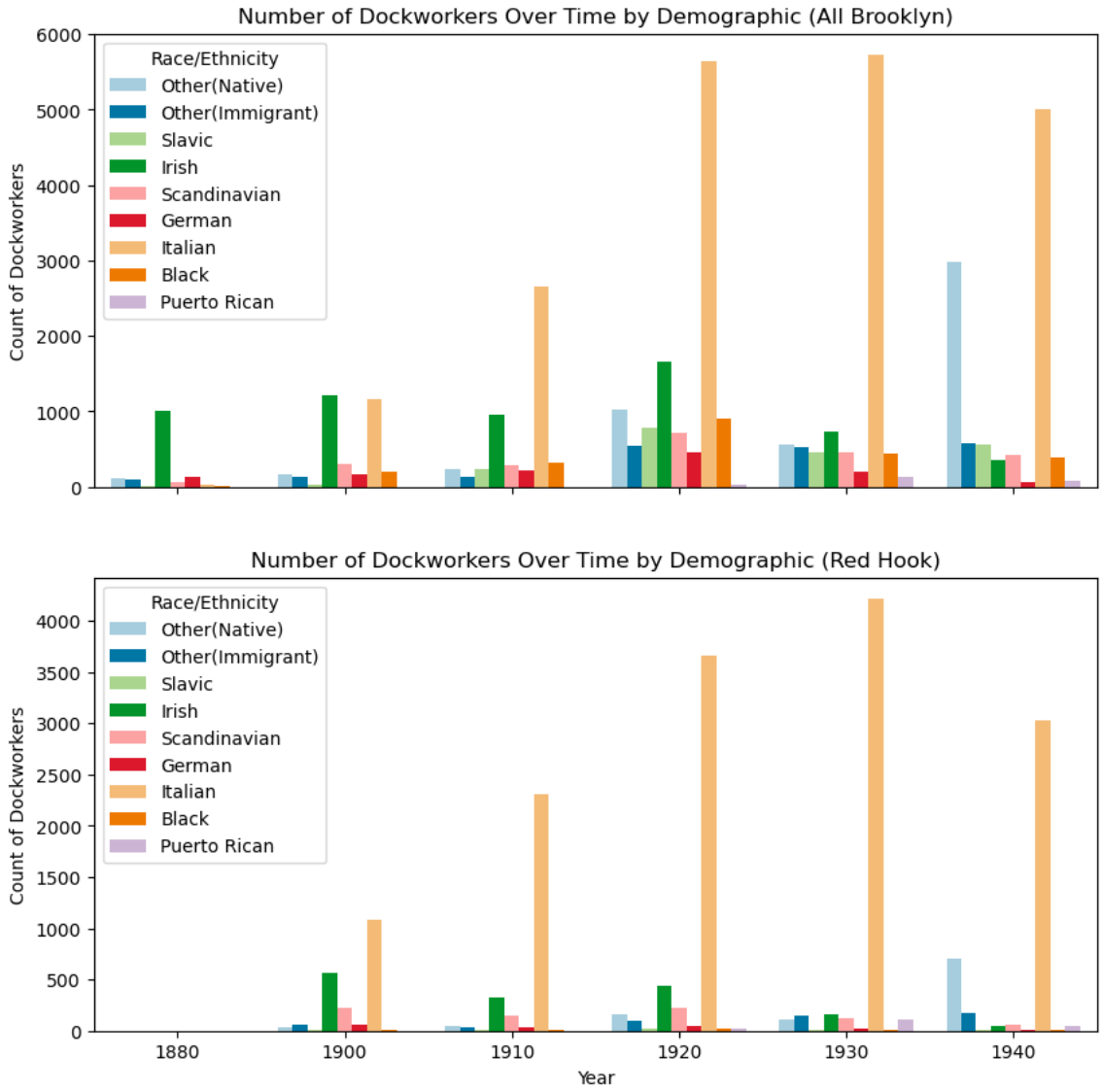


Figure 7

As expected, Italians took over the Brooklyn docks from the Irish. In Red Hook, the sheer unanimity of Italian control of these jobs is striking. One other important observation is how in Brooklyn as a whole, there is a real population of Black longshoremen that peaks around 1920 but stays significant throughout. In Red Hook, however, there is very little Black presence on the docks, likely due to the generally low number of Back residents in the neighborhood. Puerto Ricans also exhibit a fair presence on the docks especially given their sparse population.

This visualization is helpful, but due to the great Italian majority, it doesn't show us much of the nuances within the minority groups, e.g., what jobs did they work at what proportions.

This bar plot helps to address that question, that is, not how many dockworkers are Italian, Puerto Rican, Black etc. (it is obvious that Italians would make up the majority since they are such a population majority), but how many people from these populations are dockworkers?

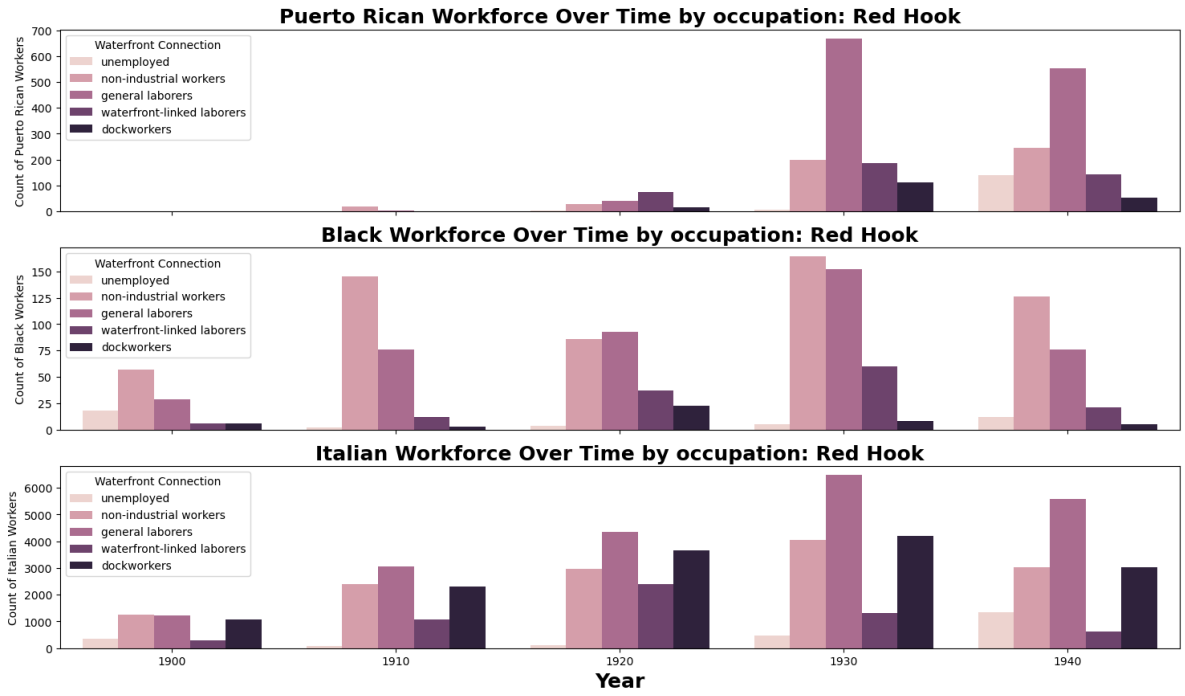


Figure 8

As we can see, the Black population is funneled more into service work than the other two groups, the Puerto Rican population are more likely to be general laborers or industrial workers, and Italians are very disproportionately employed on the docks themselves.

The following heatmap uses the census occupation categories to see these proportions in greater detail. The value in each box represents the percent of the racial/ethnic category (row) in the given occupational category (column).

Figure 9

Again, we see a clear funneling of Black workers into service jobs, and Italians, Irish and Puerto Ricans into industrial Jobs. In Red Hook specifically, it is interesting to note that more Italians are designated as ‘laborers’ while more Puerto Ricans are ‘operatives,’ although the distinctions between these categories are hazy in the census.

It is also clear that German, Slavic, and even Irish populations are more likely to hold clerical and managerial, and artisan jobs, which are likely higher paying. These higher-paying jobs are grouped with lower-paying service jobs in my other categorization method as ‘non-industrial labor,’ which makes this other categorization useful.

Lastly we can look at population shifts in the area, although the spatial mapping element of this project is less central.

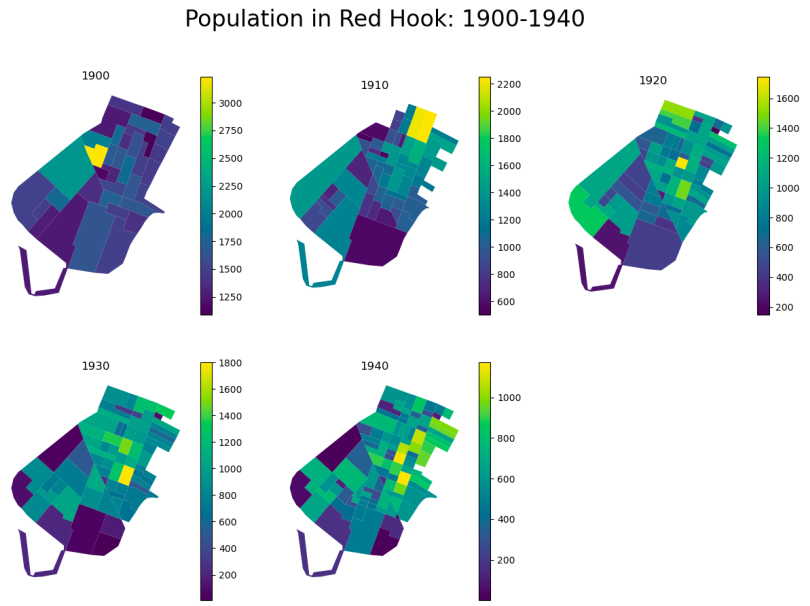


Figure 10. In the future, some research should be done on the building layout in 1900, as the density within such a small set of blocks is astounding.

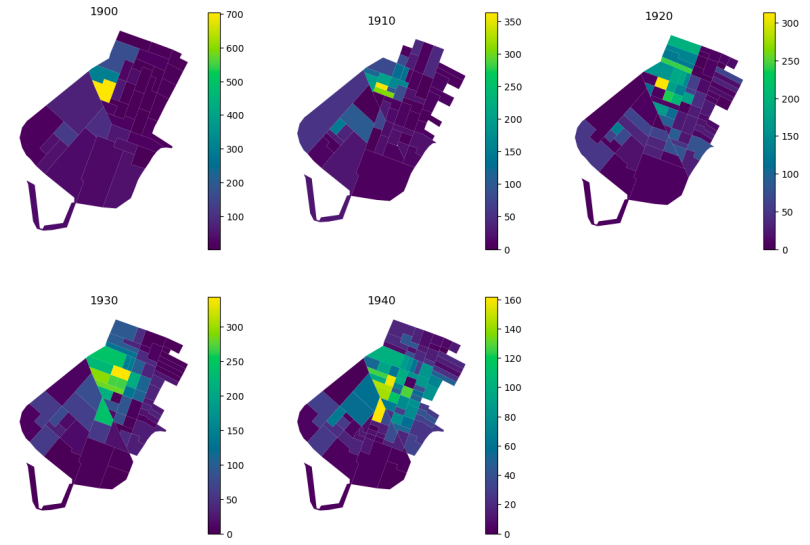


Figure 11. Stevedoring jobs seem to spread out and move away from the water slightly, as does the general population.

V. Machine Learning Models

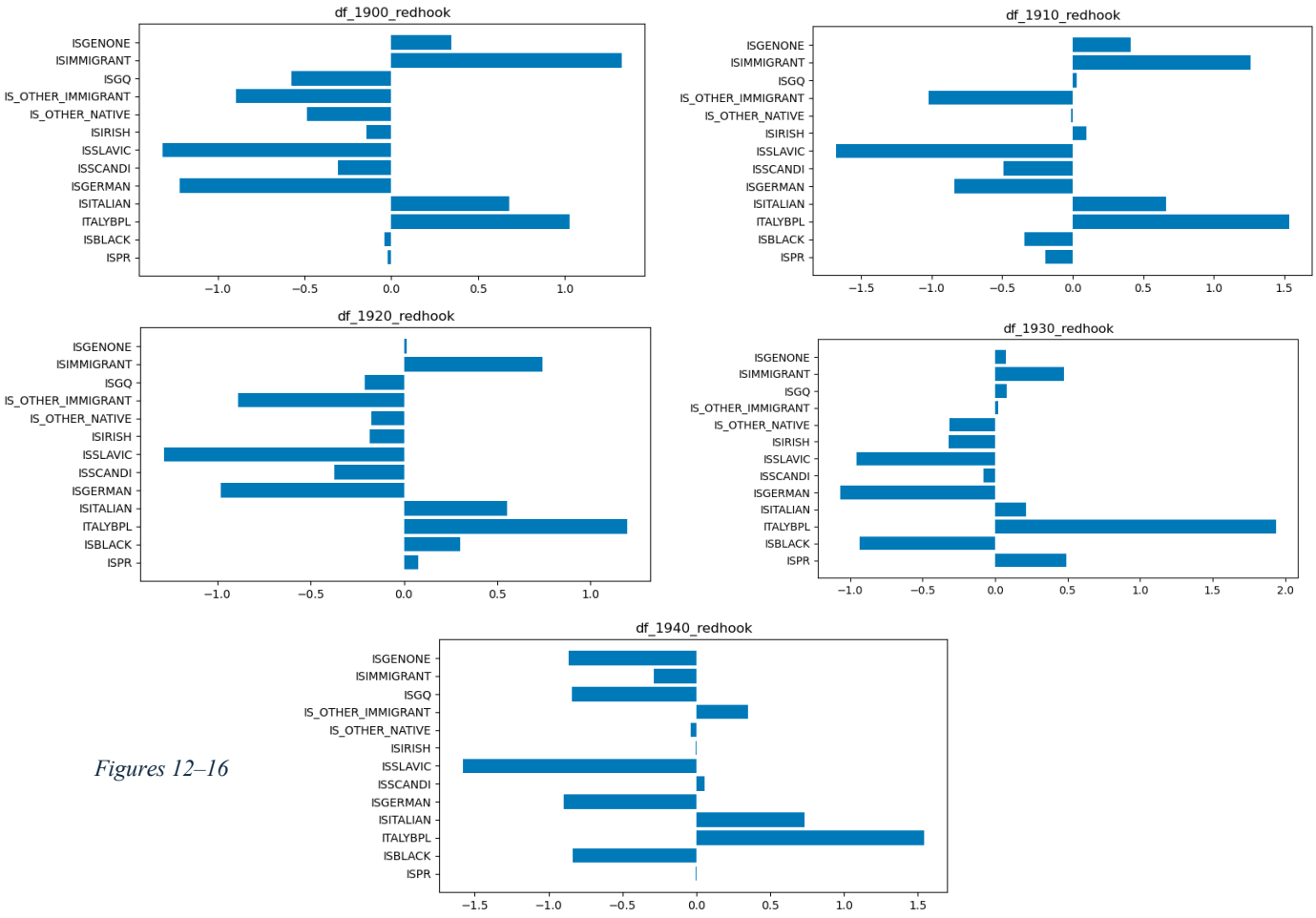
My goal with this project was to use machine learning models not only to predict classification of dock employment, but to assist in understanding the weight of the different features I used to train them. Because I wanted to fit models to data from each decennial census, I prepared my train-test split as an iterable function which could pass its x and y data to other functions specific to the models being used.

I set around 10 Boolean columns (such as 'is Italian' and 'is living in group quarters) as 'x' and the dockworker feature as 'y'

1. Logistic Regression

   This model was the first one which I tested, and it taught me some hard lessons about what to expect from my classification efforts. Given that dockworkers constituted only around 10% of the population in any census and no demographic field exhibited a perfect correlation with dockworker classification, the model defaulted to predicting 'not dockworker' consistently.

   I originally thought that this is where I would leave logistic regression, but upon finding out about `lr.coef_` I was able to use it to plot the impact of my different Boolean columns, which proved quite informative.



Figures 12–16

2. Decision-Tree AdaBoost

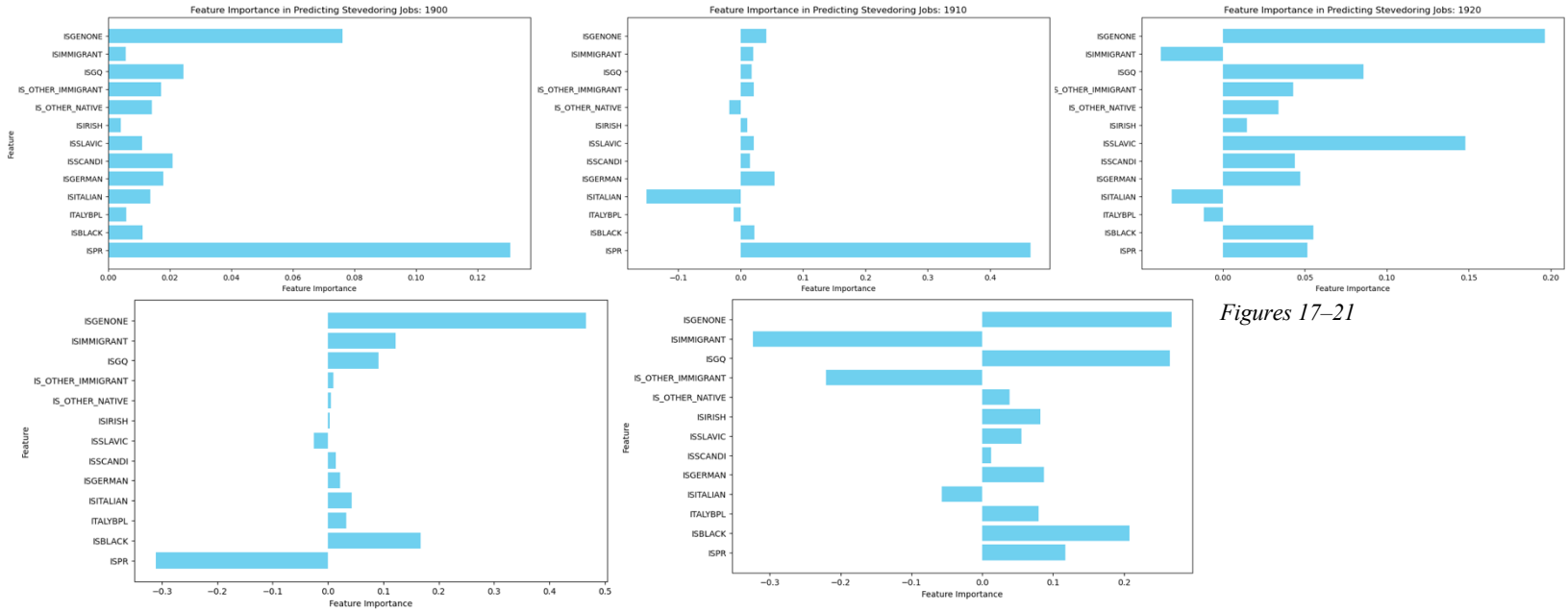
My second model was a decision tree classifier boosted with AdaBoost. Because of the Boolean nature of my columns, I expected this model to find more success, but it surprisingly ended up being rather more difficult.

Like the logistic regression model, the AdaBoost classifier chose a value of 'False' essentially every single time, regardless of how many trees I allowed it to use (I settled on a n_estimators of 1000 because of the size of my data). At one point, I experimented with manipulating my train/test split using the following function which trims the population such that half are dockworkers and half are not:

```
def raisedockproportion(df):
    dfdock = df[df['ISDOCKWORKER']]
    df_nodock = df[df['ISDOCKWORKER']==0]
    return pd.concat((dfdock, df_nodock.sample(n=len(dfdock))), ignore_index=True)
```

Manipulating the inputted dataframes in this way resulted in the classifiers choosing both options and resulting in confusion matrices which indicated that they did in fact use the Boolean columns to improve accuracy. However, I settled against using this function, as manipulating the data in this way just to yield a satisfying confusion matrix did not make sense. It is better to train and test on the actual data being studied.

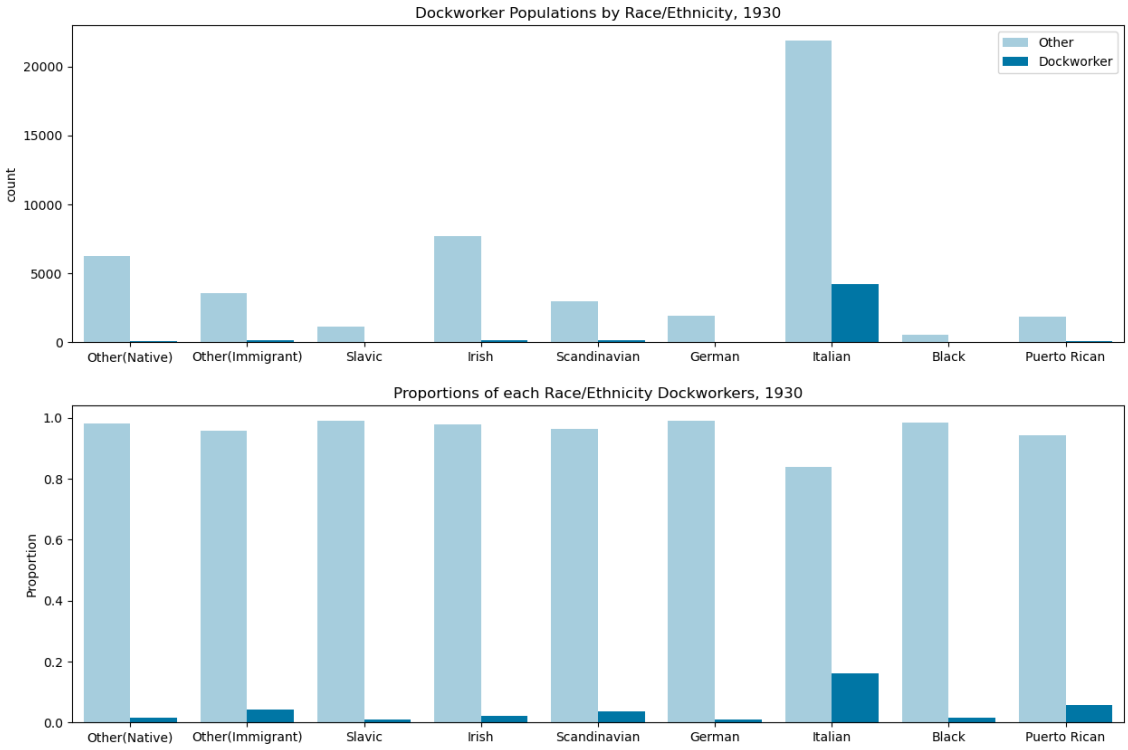
An attempt to derive feature importance using `classifier.feature_importances_` in AdaBoost proved challenging. Results varied significantly with adjustments to simple parameters such as `n_estimators`, raising questions about consistency.



Figures 17–21

Worse still, negative importance values were common, which all sources indicated should not appear at all, as they imply the column made the model less accurate. The fact that these negative values changed from year to year with little apparent rhyme or reason rendered me unable to figure out what this classifier was 'thinking'.

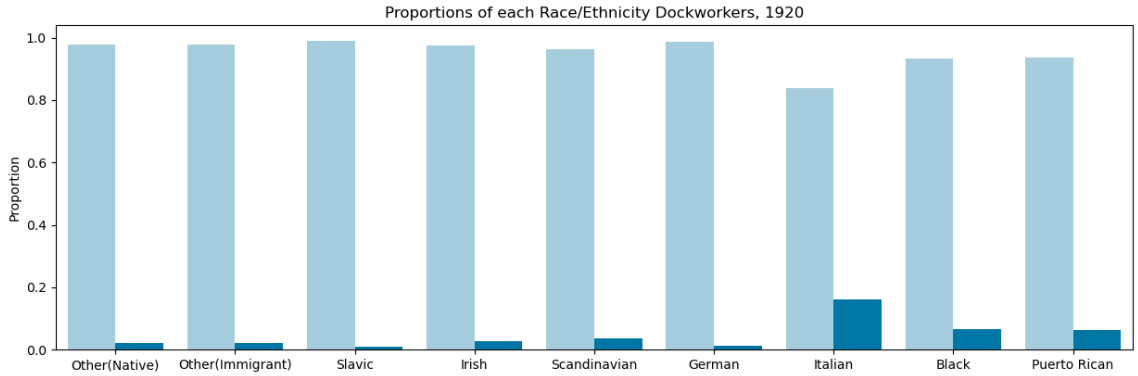
The values themselves also made little sense, or were disappointing when they did. The 'is_Italian' column is consistently given a low or negative score, despite our EDA clearly indicating how disproportionately likely members of that group were to be dockworkers. This graph I made to examine the features using normalization helps convey the degree of this disproportionate likelihood.

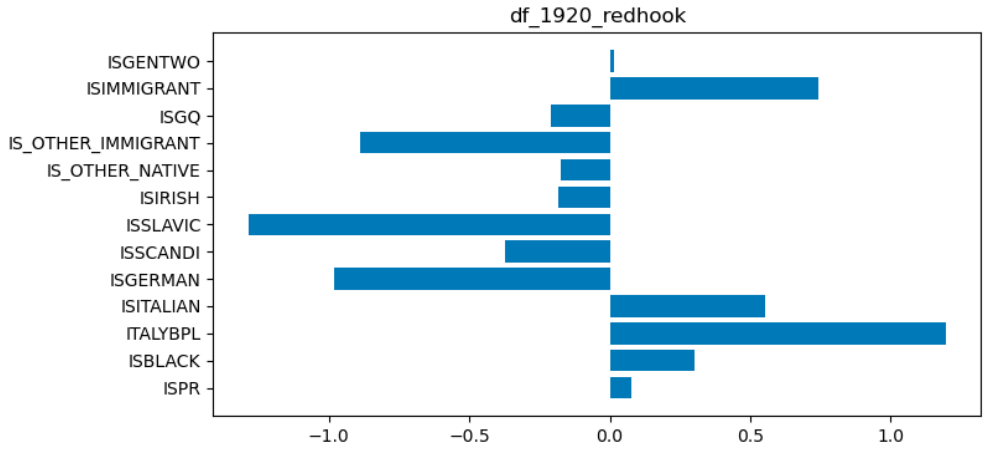


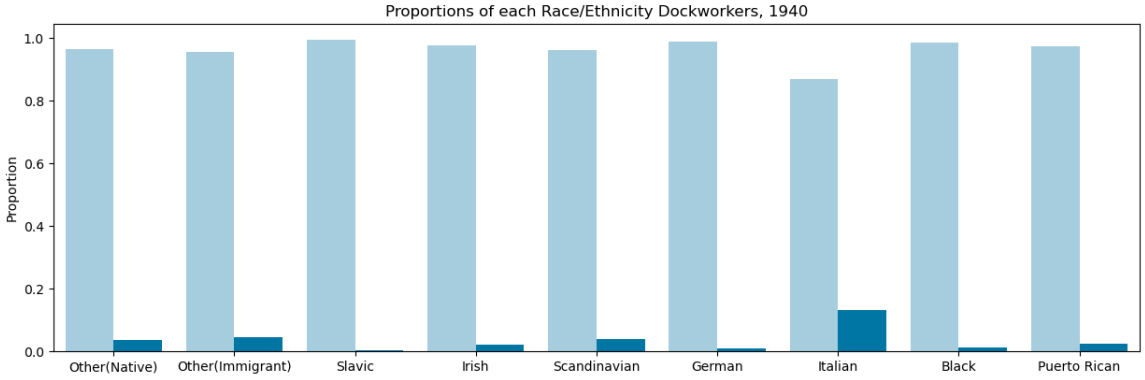
The feature importances for this year (1930), for example, don't match the correlation in the slightest. One outcome that technically made sense but was disappointing were the high importance scores given to the 'ispr' (is Puerto Rican) column in 1900 and 1910. This makes sense as there were technically no Puerto Rican dockworkers these, but the Puerto Rican population was so miniscule at that time (in the tens), that it should not be considered an important variable by any means.

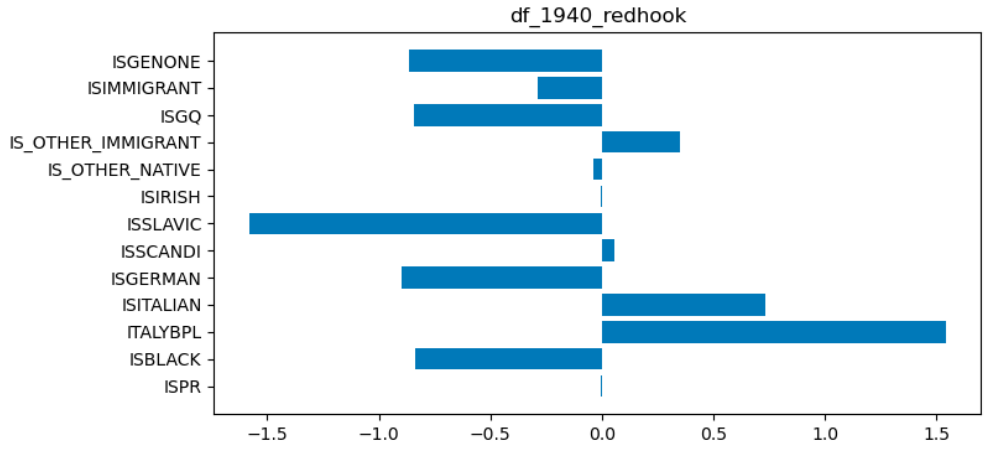
Even after doing some research on the subject, I could not figure out why the AdaBoost classifier made so little sense and gave negative importance values. Even when I normalized dockworker jobs with the above `raisedockproportion(df)` function, the importances still were negative for seemingly random variables.

On the other hand, the coefficients given by the logistic regression consistently did make sense. Here, the general importance of the variable is implied by its *absolute* value, and the value's sign indicates whether or not pointed the machine towards a conclusion of True or False. Take 1940, for example. When we examine our correlations (Slavic and German populations are very rarely on the waterfront, while Italians are much more), we see it reflected in the model's coefficients:

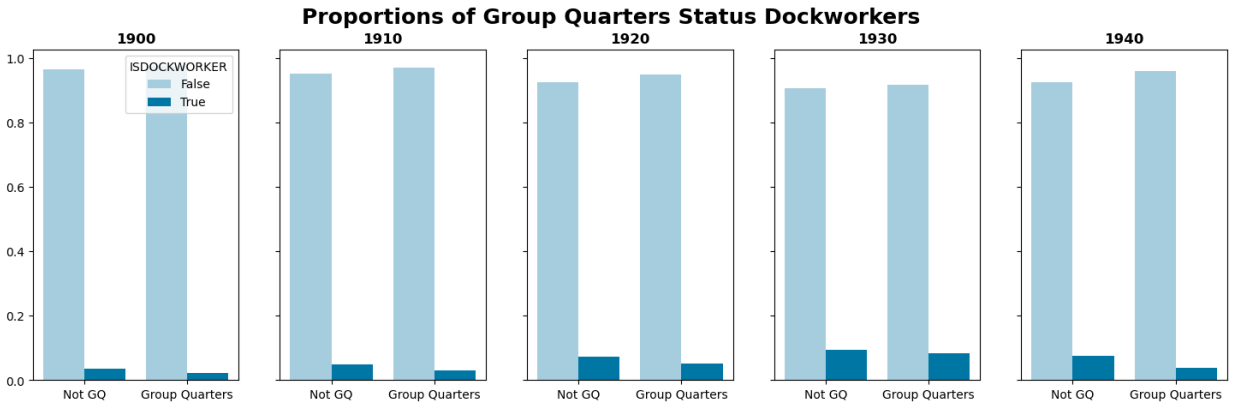

Proportions of each Race/Ethnicity Dockworkers, 1920
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Other(Native)
Other(Immigrant)
Slavic
Irish
Scandinavian
German
Italian
Black
Puerto Rican

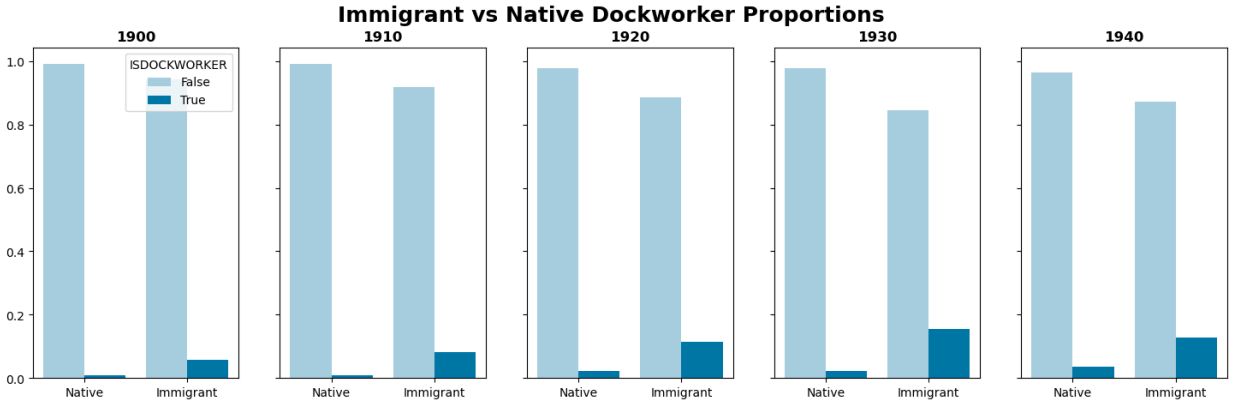

df_1920_redhook
ISGENTWO
ISIMMIGRANT
ISGQ
IS_OTHER_IMMIGRANT
IS_OTHER_NATIVE
ISIRISH
ISSLAVIC
ISSCANDI
ISGERMAN
ISITALIAN
ITALYBPL
ISBLACK
ISPR
−1.0
−0.5
0.0
0.5
1.0

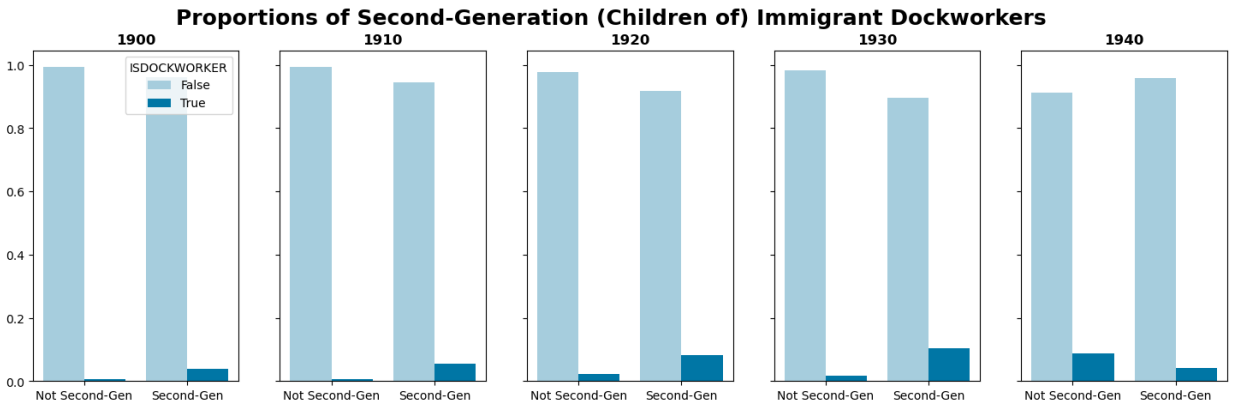

Proportions of each Race/Ethnicity Dockworkers, 1940
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Other(Native)
Other(Immigrant)
Slavic
Irish
Scandinavian
German
Italian
Black
Puerto Rican


df_1940_redhook
ISGENONE
ISIMMIGRANT
ISGQ
IS_OTHER_IMMIGRANT
IS_OTHER_NATIVE
ISIRISH
ISSLAVIC
ISSCANDI
ISGERMAN
ISITALIAN
ITALYBPL
ISBLACK
ISPR
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

The dockworker correlations of the other x variables, such as group quarters status and first/second-generation immigrant status, also matched the coefficients. Group quarters populations were generally slightly less likely to work on the docks. Both first and second-generation immigrants were more likely to work on the docks than average, but by the 1940's the children of immigrants were less likely to hold these positions.







VI. Conclusions

Both the Exploratory Data Analysis and the logistic regression model's coefficients clearly show that workers of Italian ethnicity, especially first-generation immigrants, were disproportionately employed as longshoremen or stevedores. In contrast, Black residents in Red Hook were largely excluded from dock labor and often funneled into service jobs. German and Slavic workers were rarely employed on the docks but instead occupied higher-paying managerial or clerical roles. Puerto Rican residents, while more likely to engage in other forms of industrial labor, had a notable presence among dockworkers by the 1930s, suggesting less exclusion than initially hypothesized.

This project provided valuable insights into historical employment patterns and offered a comparison of two machine learning models for classification. While decision tree classifiers and random forests are often assumed to provide better predictive accuracy and interpretability, I found logistic regression to be more effective in communicating clear relationships between variables. The decision tree classifier, particularly when boosted with AdaBoost, proved challenging to interpret.

Future work could focus on gaining a deeper understanding of the underlying mechanics of the AdaBoost Decision Tree classifier. In my project, I found it difficult to understand, but perhaps with a better knowledge of the classifier's tools and parameters, and what goes on 'under the hood,' I could have put the classifier to more use. I evaluated my models based on the simple correlations which I had plotted, but it is possible that the boosted classifier noticed something about the data that I could not, and what I saw as 'wrong' was actually just using a logic beyond my comprehension.

One of my key takeaways from this project is a broader question of machine learning's applicability to census data classification. Can a model accurately predict individual life outcomes based on such a simple set of Boolean features? Population datasets are inherently complex, often defying hard classification rules in favor of nuanced trends. For this study, a more sophisticated model capable of capturing such complexities might have been better suited.

VII. Bibliography

Davis, Colin J. "'Shape or Fight?': New York's Black Longshoremen." *International Labor and Working Class History* 62 (October 2002): 143–63.

Ernst, Robert. *Immigrant Life in New York City, 1825-1863*. Syracuse, N.Y.: Syracuse University Press, c1994.

Frost, Mary. "Brooklyn Marine Terminal Workshop Draws Hundreds to Red Hook." *Brooklyn Daily Eagle*, October 4, 2024. https://brooklyneagle.com/articles/2024/10/04/brooklyn-marine-terminal-workshop-draws-hundreds-to-red-hook/.

Jensen, Vernon H. "Decasualization of Employment on the New York Waterfront." *ILR Review* 11, no. 4 (1958): 534–50. https://doi.org/10.2307/2519354.

———. *Hiring of Dock Workers and Employment Practices in the Ports of New York, Liverpool, London, Rotterdam, and Marseilles*. Cambridge: Harvard University Press, 1964.

Larrowe, Charles P. *Shape-up and Hiring Hall; a Comparison of Hiring Methods and Labor Relations on the New York and Seattle Waterfronts*. Berkeley: University of California Press, 1955.

Simon, Malka. "'The Walled City': Industrial Flux in Red Hook, Brooklyn, 1840—1920." *Buildings & Landscapes: Journal of the Vernacular Architecture Forum* 17, no. 2 (2010): 53–72.