Business Climate in Fairfax County

Social Impact Data Commons

UVA Biocomplexity Institute
Social and Decision Analytics Division

Guy Leonel Siwe, Treena Goswami, Joel Thurston, Aaron Schroeder (PI)

9 February 2024

Citation: Siwe GL, Goswami T, Thurston J, Schroeder A (2024). Business Climate in Fairfax County. University of Virginia. https://doi.org/10.18130/hz6b-xn90

Acknowledgments: We thank the students Anjali Mehta, Prashanth Wagle, and Trinity Chamblin, who contributed to this project in the 2023 Data Science for the Public Good program through the Social and Decision Analytics Division (SDAD), Biocomplexity Institute (BI), University of Virginia.

Abstract

Fairfax County and Mastercard Center for Inclusive Growth are interested in understanding the landscape of business activity in Fairfax County with a focus on businesses owned by people from a minority group (based on race).

1. Measure Business Activities at Small Census Geo-Levels

- The goal was to provide a data product on business activities across industries at small Census geography levels (Census tracts and block groups) to support Fairfax County decision-making.
- Using microdata from **Mergent Intellect**, we estimate a set of metrics at the Census tracts and block group level across multiple industry categories. The data are accessible on our dashboard:

https://uva-bi-sdad.github.io/capital_region

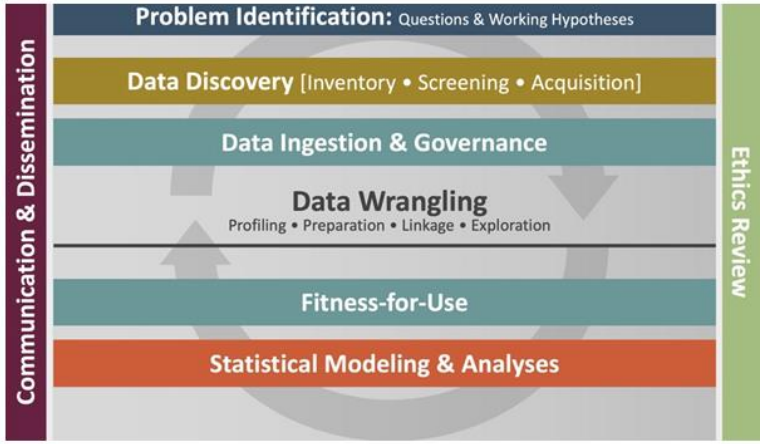
2. A Model-Based Approach to Identify Minority-Owned Companies

- The goal was to design a classifier model using Natural Language Processing (NLP) models to improve the labeling of companies as minority-owned by Mergent Intellect in Fairfax County.
- Using a Decision Tree model, we found that approximately 49% of companies located in Fairfax County are owned/managed by non-white people. This number is close to the Census Annual Business Survey estimates (38%).

Measure Business Activities for all firms at Small Census Geo-levels (Tracts and Block-Groups)

INTRODUCTION

- Census data, such as the Annual Business Survey (ABS), provides information on industry activities at the county and Metropolitan Statistical Area (MSA) geographic resolution.
- However, this geographical resolution may not offer sufficient information for decision-making by local communities and policymakers.
- The goal is to create a data product detailing company activities across industries and locations at smaller Census geography levels, such as Census tracts and block groups, in Fairfax County.
- Followed our Data Science Framework with a focus on:
  - Data Discovery (e.g., exploring and validating sources)
  - Data Ingestion (e.g., acquiring and checking metadata)
  - Data Wrangling (e.g., profiling and linking data)
  - Fitness-for-Use (e.g., establishing definitions and metrics)
  - Statistical Modeling (e.g., developing classifiers)



- Our data and data products are accessible through our dashboard.

DATA DISCOVERY (inventory and screening)

Source	Description	Time coverage and sample size	Information Available (also represents the different dataset)
Mergent Intellect	• Data aggregator collecting information through surveys and phone calls • Updated every year • Available through UVA library	• **Coverage:** From 2010-2020 • **Large sample size:** More than 600K companies in 2020. • Available for the US	• **Company details:** DUNS, name, address, industry classification, etc.; • **Historical operations:** sales, employment, etc.; • **Executives:** Executive name, title, etc.
Data Axle	• Data aggregator • Available through UVA library	• We could not determine the methodology used for data collection	• **Company details:** DUNS, name, address, industry classification, etc.; • **Historical operations:** sales, employment, etc.
AtoZ	• Data aggregator • Available through UVA library	• **Small sample size:** 50K companies • We could not determine the methodology used for data collection	• **Company details:** DUNS, name, address, industry classification, etc.

DEFINITIONS

- **New Business:** A company that has recently moved into a location or is newly established.
- **Closed Business:** Company that has ceased its activities or relocated away from the current location.
- **Active Business:** Company in operation during a specific year at a given location.
- **Small Business:** Company with fewer than 50 employees.
- **Sole Proprietor Business:** Company with only one employee, typically the owner.
- **Industry:** Defined by the first two digits of the North American Industry Classification System (NAICS) code.
- **Minority worker:** A worker belonging to a minority group based on race. These races include Black, Asian, American Indian and Alaska Natives, Hawaii, and Pacific Islanders. This definition of minority membership is based on the one used by the U.S. Census.

METHODOLOGY

- We chose to focus our analysis on Mergent Intellect because:
  - Most information about methodology available.
  - Largest sample size.
  - Provided most information about businesses, including executives' information.
- We validated that companies are uniquely identified by a DUNS.
  - Data Universal Numbering System.
  - Unique nine-digit identifier created by credit bureau Dun & Bradstreet.
- Companies are geolocated (i.e., their latitude and longitude are obtained) based on their addresses.
  - To geolocate businesses we use the Google Maps API and Census data.
  - Companies were assigned to Census tracts and block groups.

METHODOLOGY

- Companies are classified as small companies or not (1,0).
- Companies are classified as sole proprietors or not (1,0).
- Companies are categorized by industry. We use the two-digit NAICS code provided in the data to identify the industry (see next slide).
- We track companies over time by combining two sets of data from Mergent Intellect:
  - Company details data
  - Company operation data.
- Metrics are computed at the Census geo-levels (tracts, block groups) and by industry.

INDUSTRY (NAICS-Two Digits)

- Accommodation and Food Services.
- Administrative and Support and Waste Management and Remediation Services.
- Agriculture, Forestry, Fishing and Hunting.
- Arts, Entertainment, and Recreation.
- Construction.
- Educational Services.
- Finance and Insurance.
- Health Care and Social Assistance.
- Information.
- Management of Companies and Enterprises.
- Manufacturing.
- Mining, Quarrying, and Oil and Gas Extraction.
- Non-classifiable Establishments.
- Other Services (except Public Administration)
- Professional, Scientific, and Technical Services.
- Public Administration.
- Real Estate and Rental and Leasing.
- Retail Trade.
- Transportation and Warehousing.
- Utilities
- Wholesale Trade.

METRICS on Social Impact Data Commons

Metrics provided by year, location, and industry:

- **Entry Rate:** Percentage of new companies.
- **Exit Rate:** Percentage of closed companies.
- **Jobs creation:** Total count of jobs created.
- **Jobs destruction:** Total count of jobs cut.
- **Total employment:** Total job count.

Link: https://uva-bi-sdad.github.io/sdc.intro/economic_diversity.html

METRICS

Other metrics available*:

- **Small Businesses:** Percentage of small companies.
- **Sole Proprietor Businesses:** Percentage of sole proprietor companies.
- **Percentage of Small Businesses:** Percentage of small companies.
- **Percentage of Sole Proprietor Businesses:** Percentage of sole proprietor companies.
- **New Businesses:** Total count of new companies.
- **Closed Businesses:** Total count of closed companies.
- **Jobs creation by new businesses:** Total count of jobs created by new businesses.
- **Jobs destruction by closed business:** Total count of jobs cut by closed businesses.
- **Number of businesses cutting jobs.** Total count of closed companies
- **Number of businesses creating jobs.**
- **Percentage of Jobs created.**
- **Percentage of Jobs destroyed.**
- **Location quotient:** Ratio of industry employment share at a specific location vs. the overall County.
- **Herfindahl Hirschman Index:** Employment concentration by a few companies.

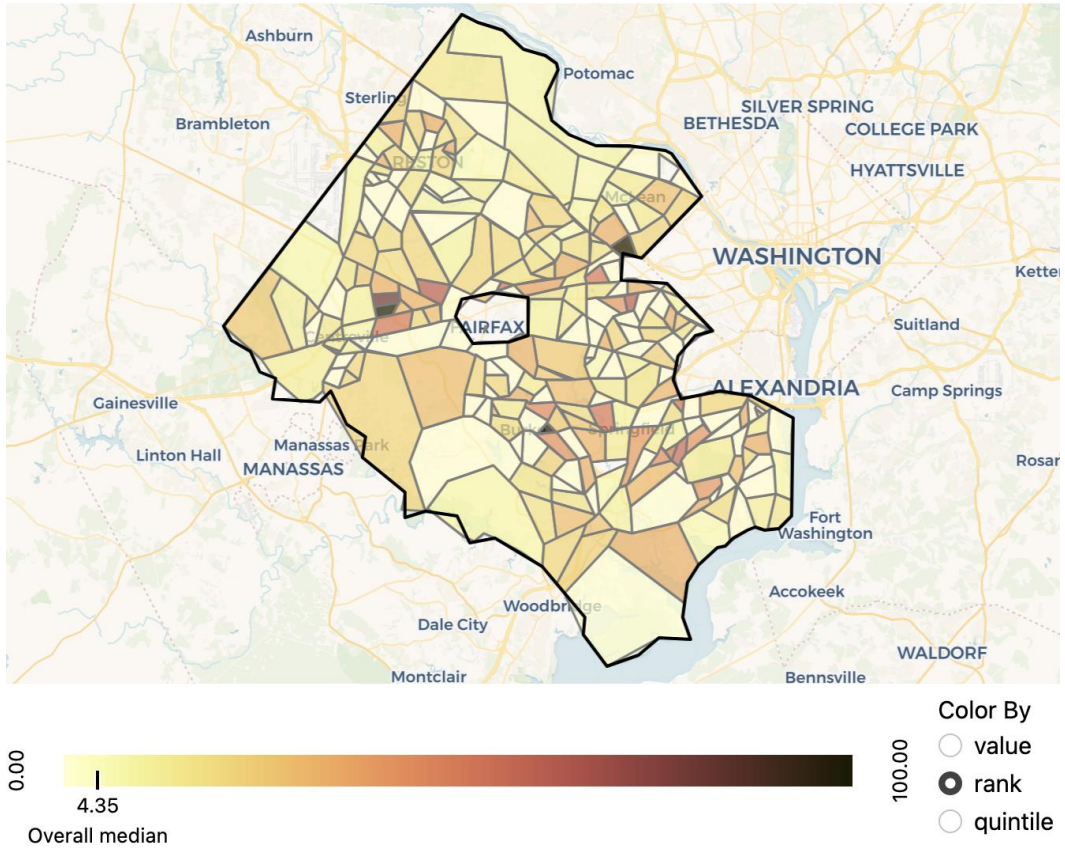
* *We have not yet evaluated these metrics to determine their quality.*

https://uva-bi-sdad.github.io/capital_region

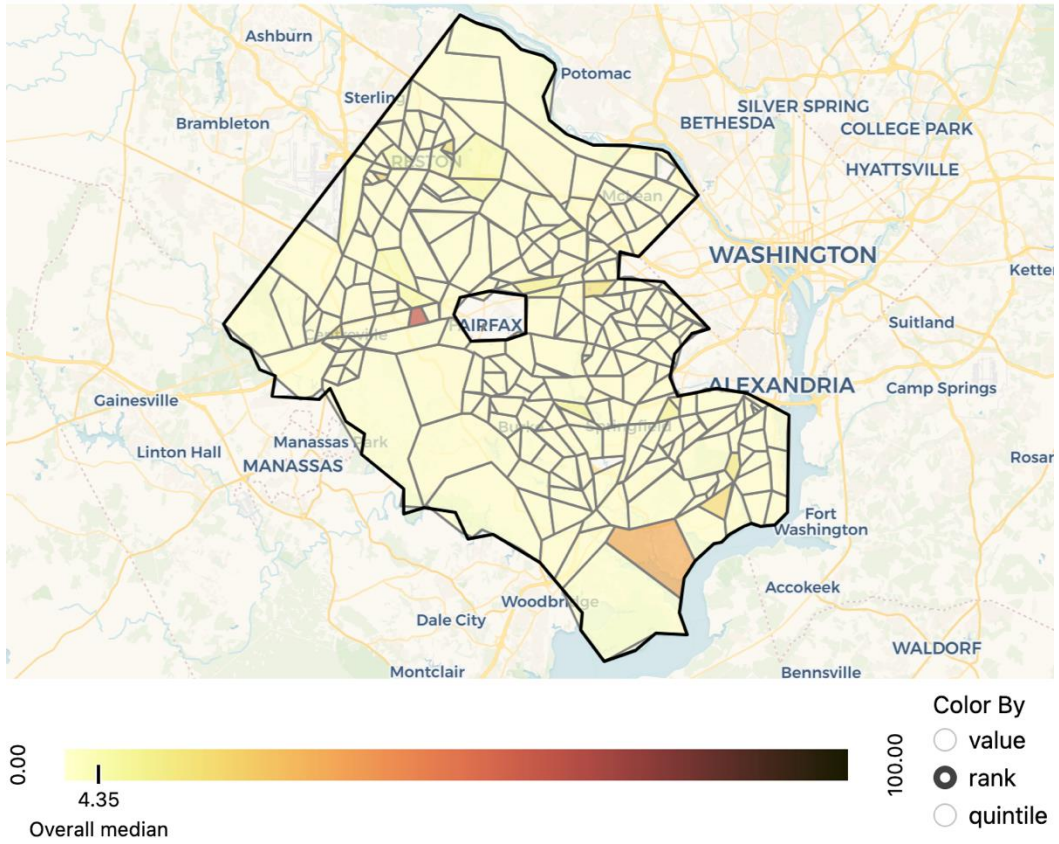
Social Impact Data Commons RESEARCH CAPABILITIES

- Example visualization comparisons. These are available on Social Impact Data Commons:
  1. Growth of the real estate industry through 2018 comparing entry and exit rates.
  2. Job growth in the health care industry through 2018 comparing job creation vs job elimination rates.
  3. Impact of COVID-19 on employment in the health care industry from 2018 to 2019.
  4. Distribution of Minority Workers across the county in 2019.

1. Entry and Exit Rate in Real Estate Industry



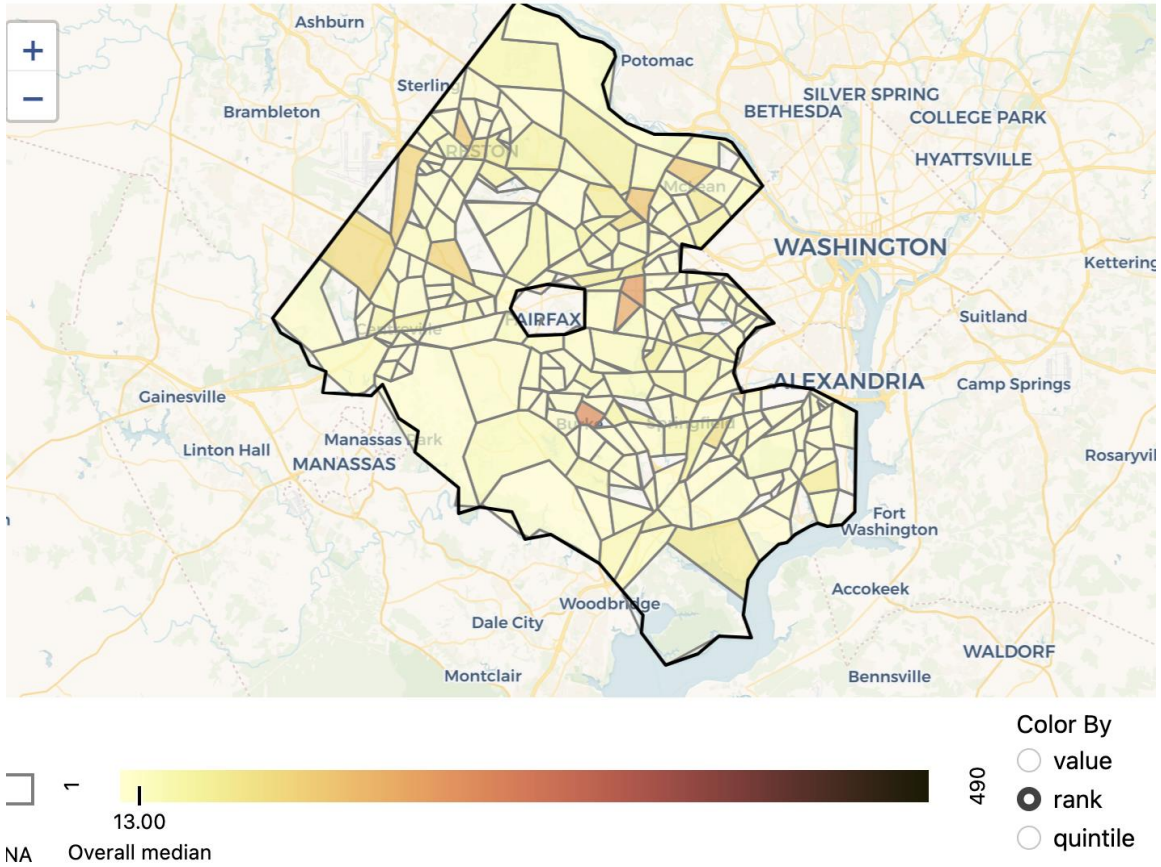
Entry rate in the Real Estate industry across tracts in 2018



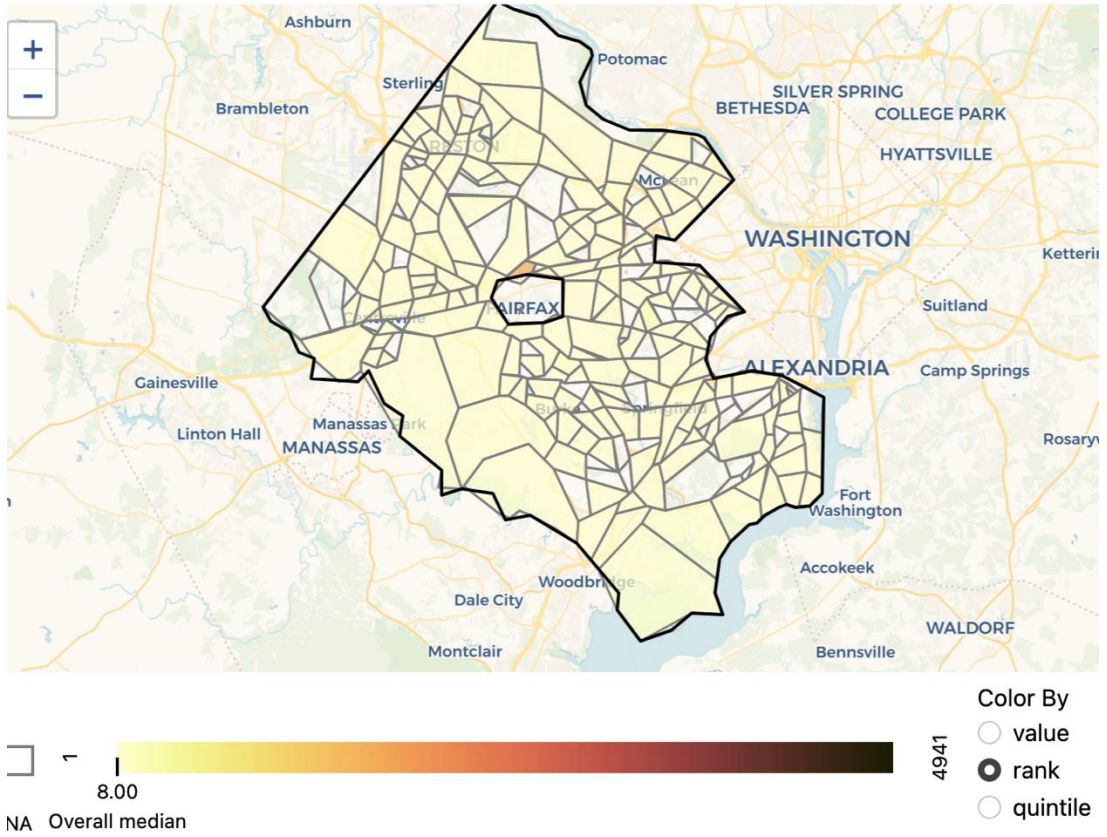
Exit rate in the Real Estate industry across tracts in 2018

In 2018, we observe comparable median entry (4.4%) and exit (4.4%) rates for businesses in the Real Estate industry for 50% of the Census tracts, suggesting a flat growth rate across Fairfax County. However, there are many Census tracts that show high entry rates while exit rates remain low.

2. Job Creation vs Job Elimination



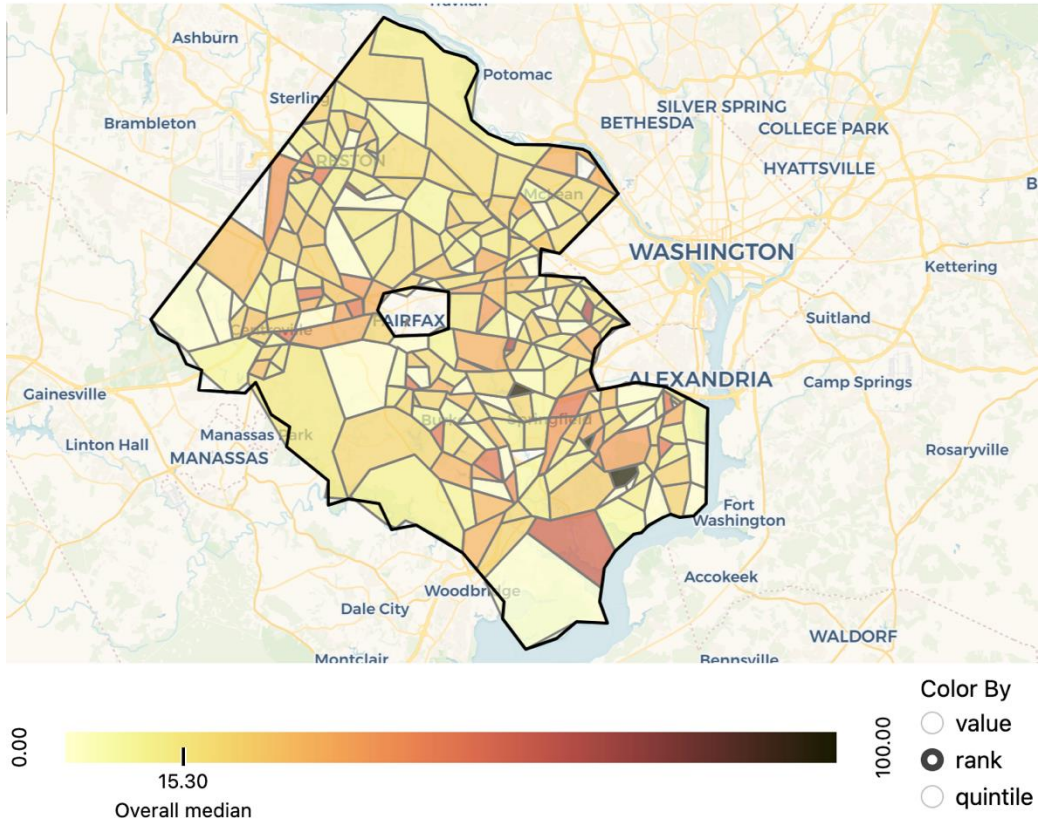
Total jobs created in the health care industry in 2018



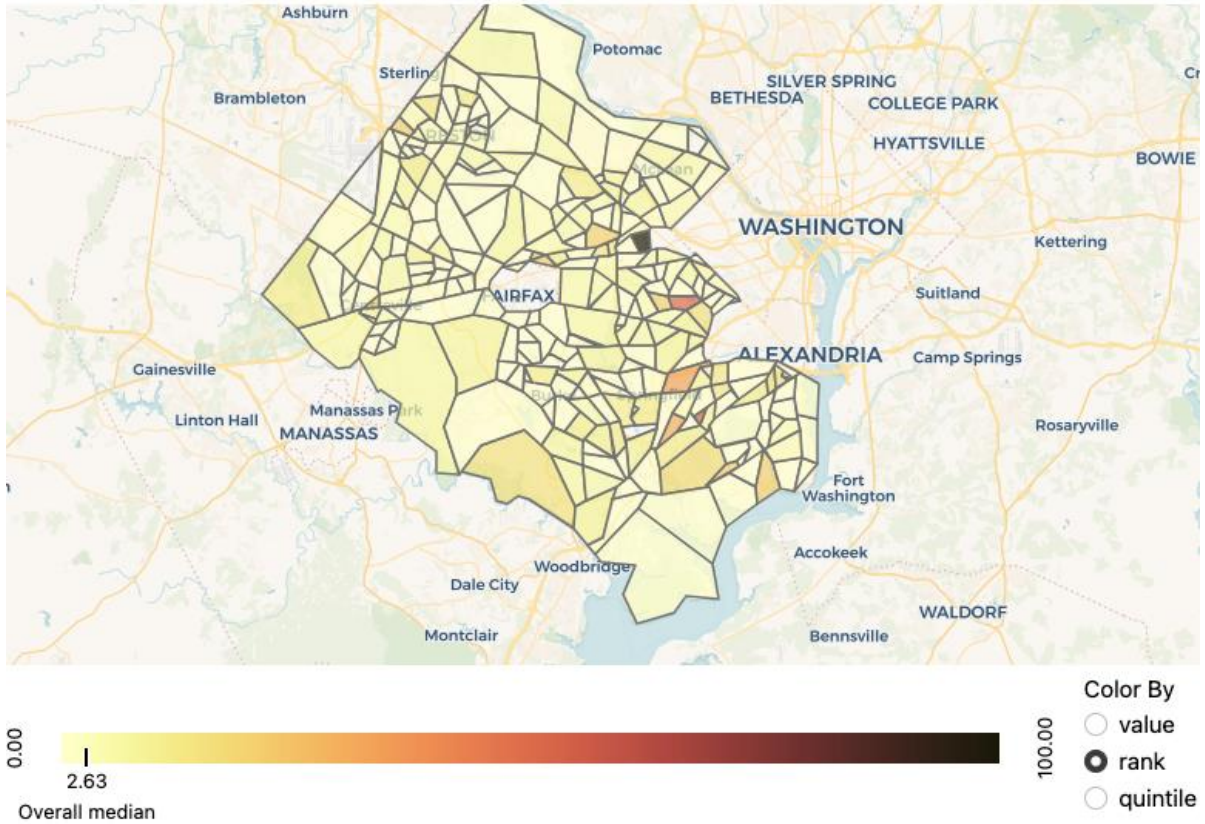
Total jobs eliminated in the health care industry in 2018

In 2018, more jobs were created than eliminated in the Health Care industry in most Census tracts across Fairfax County.

3. Entry Rate in Health Care Industry



Entry rate in the Healthcare industry across tracts in 2018 (pre Covid-19)

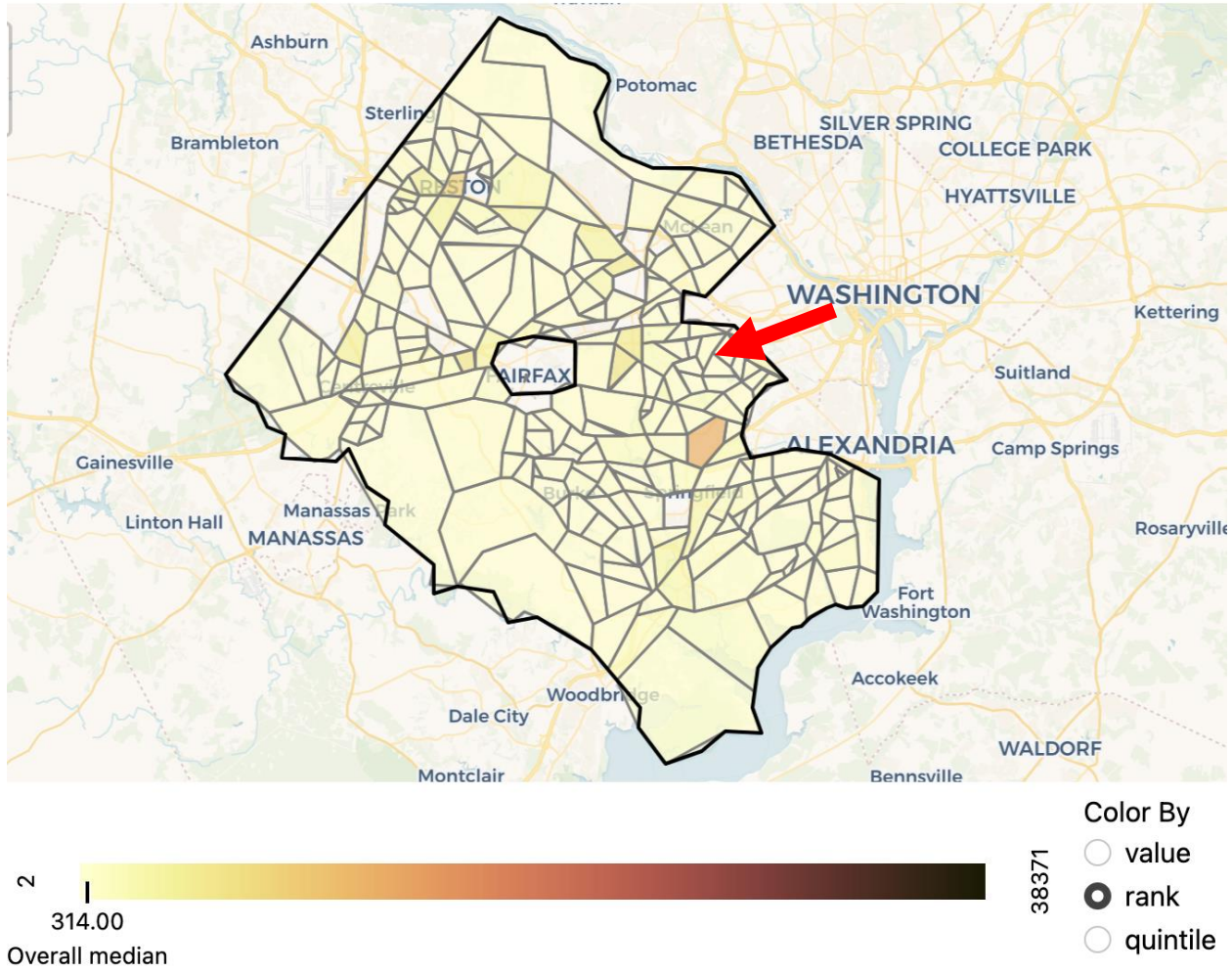


Entry rate in the Healthcare industry across tracts in 2019 (beginning of Covid-19)

In the lead up to the COVID-19 pandemic, we observe a reduction in the company median entry rate from 2018 to 2019 in the Health Care industry across Fairfax County.

4. Distribution of Minority Workers

Total Count of Minority Workers by Census Tracts in Fairfax County-2019 (From LODES data)



- Workers from minority group (based on race) are generally uniformly distributed across Census tracts.
- But some tracts exhibit a high concentration of minority workers.
- For example, from 2010-2020 we observe high minority employment in Census tract **51059452501**.

MEASURING BUSINESS ACTIVITY SUMMARY

CONCLUSIONS:

- Data aggregators (e.g., Mergent Intellect) can be used to provide a subcounty analysis of business activity by industry and across time in Fairfax County.
- They can provide some information on minority employment but are most useful in understanding the broader landscape of the business climate.
- Data provides a starting point for deeper investigations, such as geospatial analysis, to determine what types of businesses or industries thrive in certain areas of the county and what factors about these areas or businesses may provide them with a comparative advantage.

NEXT STEPS:

- Validate additional Mergent Intellect data (e.g., small business, sole proprietorship) further by aggregating metrics to a higher geographic resolution that matches Census information from the Annual Business Survey.
- Identify and incorporate additional microdata to increase the precision of our estimates.
- Compute additional metrics (e.g., financial performance – debt, equity, environmental information).

A Model-Based Approach to Identify Minority-Owned Companies in Fairfax County

We are grateful to our 2023 DSPG students Anjali Mehta, Trinity Chamblin, and Prashanth Wagle for assisting with this research.

INTRODUCTION

- Initial analysis:
  1. Mergent Intellect classifies companies as minority-owned.
     - **Minority-owned company:** U.S. company that is at least 51% owned by, and whose management and daily company operations are controlled by, one or more members of a socially and economically disadvantaged minority group (based on race). Their definition of minority based on race is the one used by the U.S. Census.
  2. Approximately 7% of businesses in Fairfax County were identified as being owned/managed by non-white people in the Mergent Intellect data.*
  3. This percentage was low compared to the 38% reported by the 2017 Annual Business Survey (ABS)* for Fairfax County. (*Joint US Census Bureau and NSF National Center for Science and Engineering Statistics)
- Conclusion: **Mergent Intellect under classified non-minority-owned companies**, especially in industries such as food and restaurants.

* More information about our work on the distribution of minority-owned companies using the Mergent Intellect classification is available here: https://uva-bi-sdad.github.io/sdc.intro/economic_diversity.html

GOAL – METHODOLOGY – FINDING

- Assuming companies labeled as minority-owned by Mergent Intellect are correctly identified, we aim to build a classifier model to re-label companies listed as non-minority-owned by Mergent Intellect.
- Methodology:
  - Build a labeled sample of companies identified as minority-owned.
  - Prepare the predictive variables using NLP pre-trained models.
  - Train a classifier to predict minority-owned status only for those companies listed as non-minority-owned by Mergent Intellect.
- Our classified identified approximately 49% of companies in Fairfax County as minority-owned.
  - This is higher than the 7% identified by Mergent Intellect.
  - This is also higher than the 38% identified by the American Business Survey.
- We worked on an extension of this model where we remove the assumption that companies labeled as minority-owned by Mergent Intellect are correctly identified.

METHODOLOGY

1. **Build a labeled sample:** We constructed a training sample of companies labeled as minority-owned and non-minority-owned, serving as a ground truth that we can use to build classification models.
   - We combined different sources of information, including historical information, to identify the minority status of companies.
   - We built a new labeled sample using exclusively human coders. We've drawn a sample of companies from Mergent Intellect and conducted a manual Google search using owner and executive names to identify minority-owned businesses.
2. **Prepare the predictive variables:** We built a set of predictive variables for the minority-owned status of businesses, including:
   - The probability that business executive race, on average, is non-white.
   - The percentage of non-white people living close to the business location.
   - The probability that the company name includes words from a language spoken in a country whose population constitutes a racial minority in the United States.
3. **Train the classifier:** We use our labeled training sample to train our classifier to discern a pattern between the predictive variables and the minority status.
4. **Predict the minority status of firms**

1. LABELED SAMPLE

- The goal is to build a data set of well-identified minority and non-minority owned businesses.
- To identify **minority-owned companies:**
  - We use a random subset of companies classified as minority-owned by Mergent Intellect.
  - We assume that companies listed as minority-owned by Mergent were correctly labeled.
  - However, we scrape data on companies reported as minority-owned from different sources (such as Yelp, Chamber of Commerce) and administrative records (SBSD).
  - We compare this list against companies classified by Mergent Intellect as minority-owned to evaluate the classification made by Mergent Intellect.

- To identify **non-minority-owned companies**:
  - We use the date of creation to build a subset of companies in Mergent Intellect for which we have high confidence in being owned by non-minority.
  - We use historical, contextual information to set a threshold, based on the company's creation date, before which it is likely that the business would not be owned by a minority (e.g., the 1964 Civil Rights Act before which businesses were more likely to be created by white people).

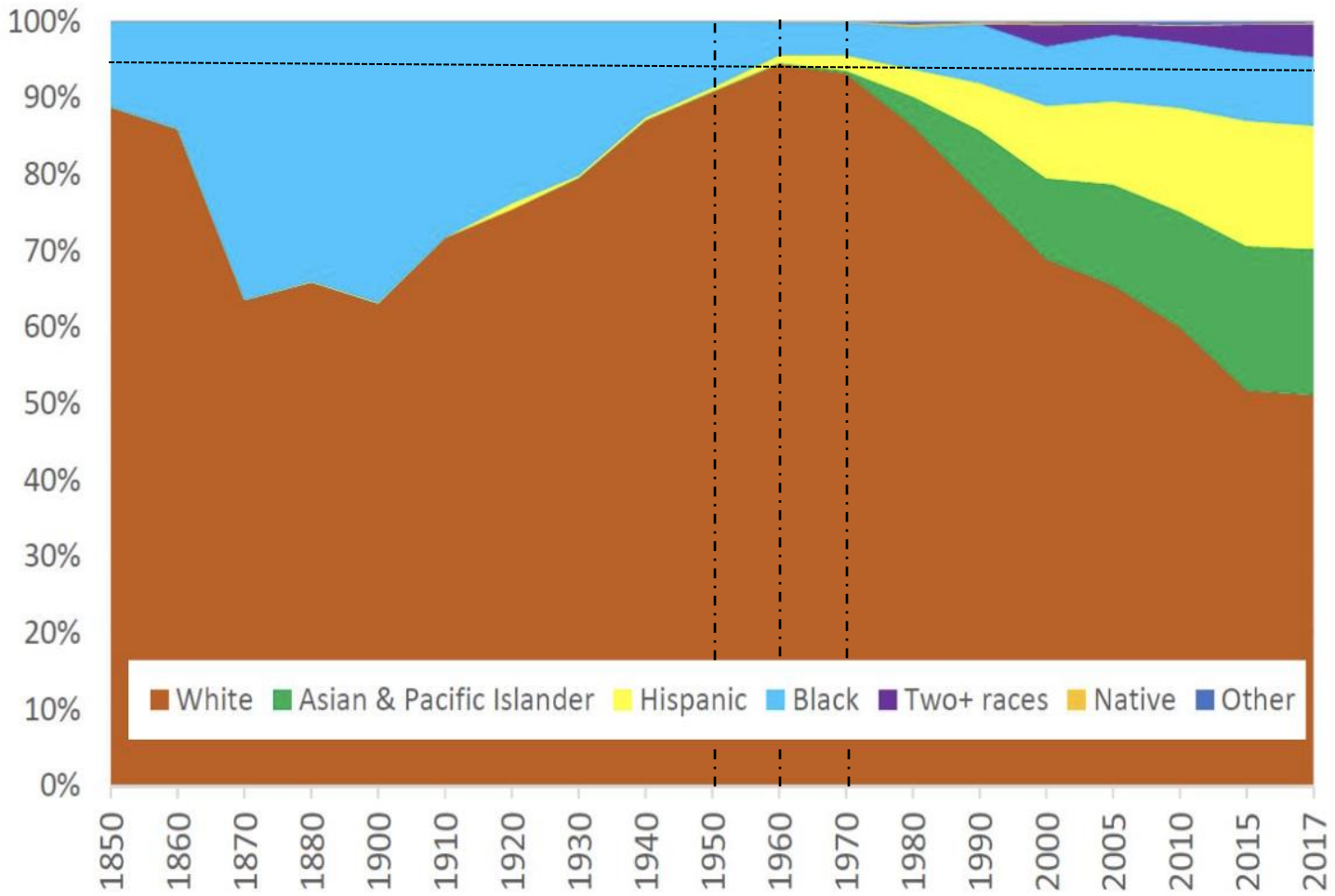
1. LABELED SAMPLE

- To scrape minority-owned companies, we use a set of keywords (such as ***minority-owned, Black, Asian, Indian, Hispanic***, *etc*.) to search for company name and location (address) across different sources (e.g., ***Yelp, Chamber of Commerce***). We also downloaded administrative records from SBSD and Data Axle.
- 38 companies from our scraped list of minority-owned companies were found in Mergent Intellect.
- Among them, **15 (39.47%)** companies were labeled as non-minority-owned by Mergent Intellect. This result highlights the potential mislabeling of many minority-owned as non-minority-owned by Mergent Intellect.

	List of companies (scraping + administrative records)				
	Yelp	Chamber of Commerce (Black, Asian, Hispanic)	Small Business and Supply Diversity (SBSD)	Data Axle	Total
Total companies scraped from source	871	435	987	650	-
Total companies scraped also found in Mergent Intellect	7	5	21	7	38*

*Two companies were found in two data sources; the total reflects deduplication.

1. LABELED SAMPLE



- We use historical, contextual information to identify companies likely to be owned by minorities vs non-minorities.
- We choose 1964 as a historical threshold because:
  - The Civil Rights Act passed in 1964.
  - The demographic composition of Fairfax County from 1950 to 1970, was approximately 91% to 95% white.
- Given the prevailing cultural and historical trends of the time, coupled with the demographic composition of Fairfax County, businesses created before 1964 are likely to have been created and owned by non-minorities.

1. LABELED SAMPLE

- Mergent Intellect data covers 160,000 companies between 2010 and 2020. However, only 12,123 companies' data includes executive names.
- We can only predict and train our model on the sample of companies with executive names reported.
- Our labeled sample was built from those 12,123 companies. The sample includes 138 companies split between minority-owned and non-minority-owned. Interestingly, all companies created before 1964 were labeled as non-minority-owned by Mergent Intellect.

		Labelling Sample		
		Minority-owned companies (sub sample from Mergent Intellect)	Non-minority-owned companies (companies created before 1964)	Total number of companies
Mergent Intellect classification	Minority-owned businesses	69	0	69
	Non-minority-owned businesses	0	69	69
Total number of businesses		69	69	138

2. PREDICTIVE VARIABLES

Executive names

Use the pre-trained Natural Language Processing models to predict the likelihood of being non-white

Predicted non-white race probability

Company Location

Use the company location to extract from ACS the percentage of non-white residents at the Census tract level where the company is located

Percentage of non-white residents at the Census tract of the company location

Company name

Use Natural Language Processing to detect the word's language from specific countries

Probability that the *company name* doesn't contain words from a language spoken in a country whose population constitutes a racial minority in the United States

Using boxplots, we investigate if these variables discriminate between minority-owned and non-minority-owned businesses in our labeled sample

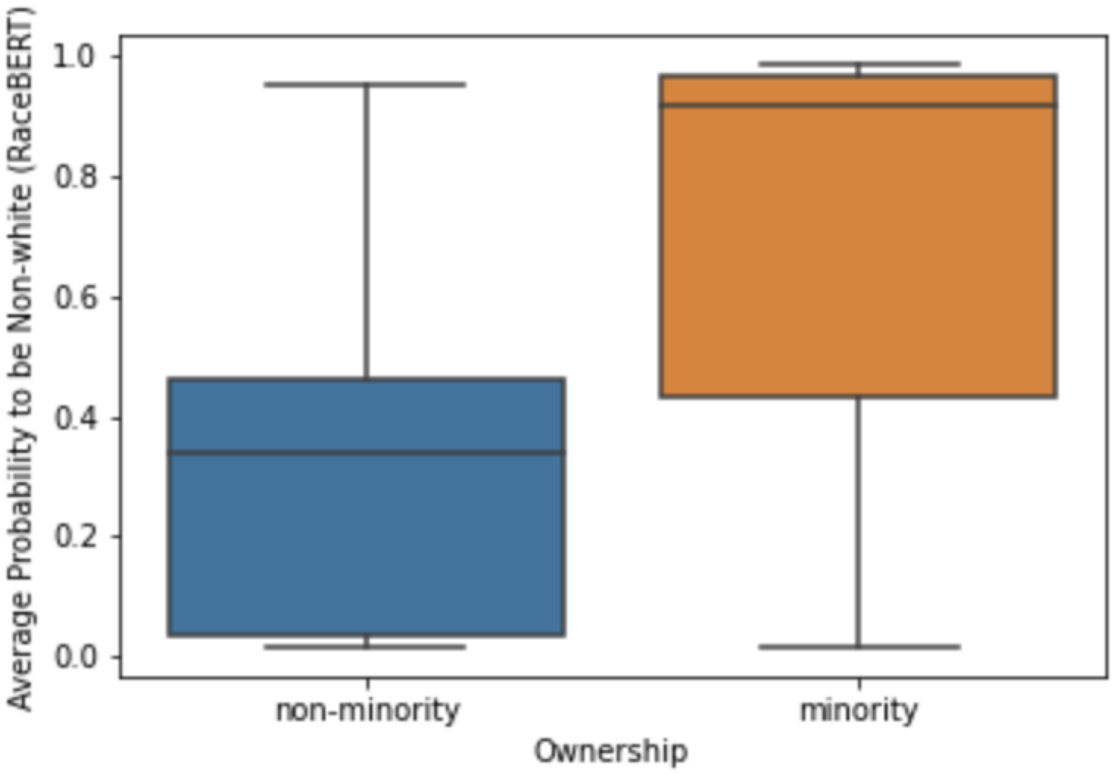
2. PREDICTIVE VARIABLES: NLP models

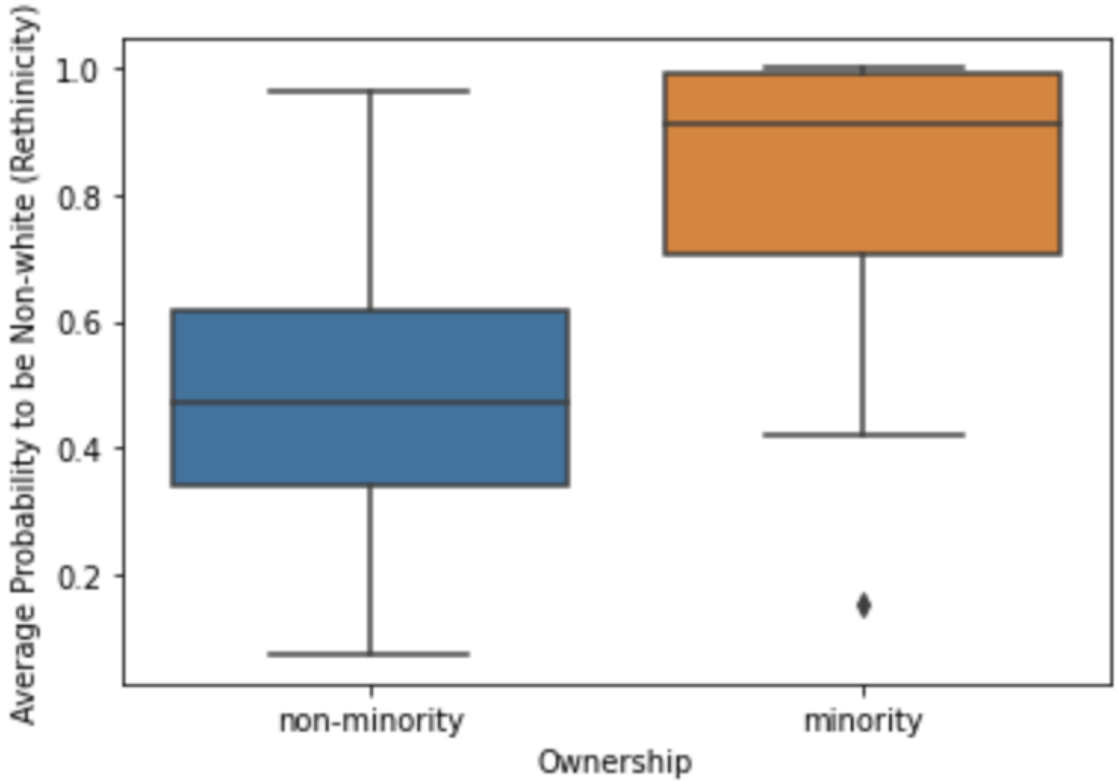
- We use a set of pre-trained models to predict the probability than an executive name is non-white based on their name. The pre-trained models provide the likelihood that a given first name + surname is associated with a race. Races are grouped as white and non-white.
- The table provides the two main models we retained based on their high predictive scores. However, we also tested other pre-trained models such **Namsor, Ethnicolor**, etc.

Models	Description	Input parameters	Output parameters	Tunning parameters
raceBert	A transformer-based model for predicting race and ethnicity. Using a transformer-based model trained on a U.S. Florida voter registration dataset, the model predicts the likelihood of a name belonging to 5 U.S. Census race categories (White, Black, Hispanic, Asian & Pacific Islander, American Indian & Alaskan Native).	First name, surname	Probability of each race and ethnicity	no parameters
rethnicity	The model is a Bidirectional Long Short-Term Memory (Bi-LSTM), a recurrent neural network architecture commonly used for natural language processing, was chosen as the model for our study. The Florida Voter Registration was used as the training and testing data.	First name, surname	Probability of each race and ethnicity	No parameters

2. PREDICTIVE VARIABLES: Executive's Race

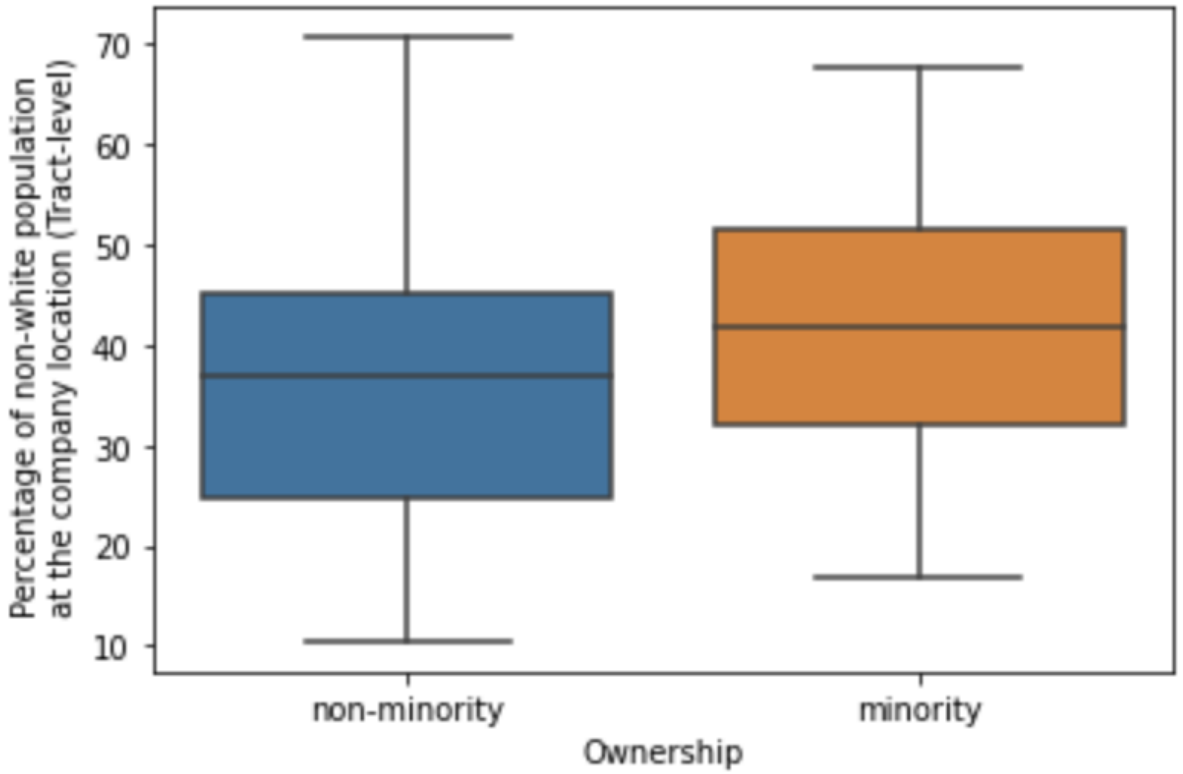
- For each company, we average the probability that the executive is non-white and obtain the probability that the average executive is non-white.
- The plot shows the distribution of labeled companies according to the average probability that their executives are non-white using our two main pre-trained models, **raceBert** and **Rethicity**, which provide clear discrimination**.**
- On average, the probability that executives are non-white is higher for companies in our labeled sample that are identified as minority-owned than it is for those identified as non-minority-owned.





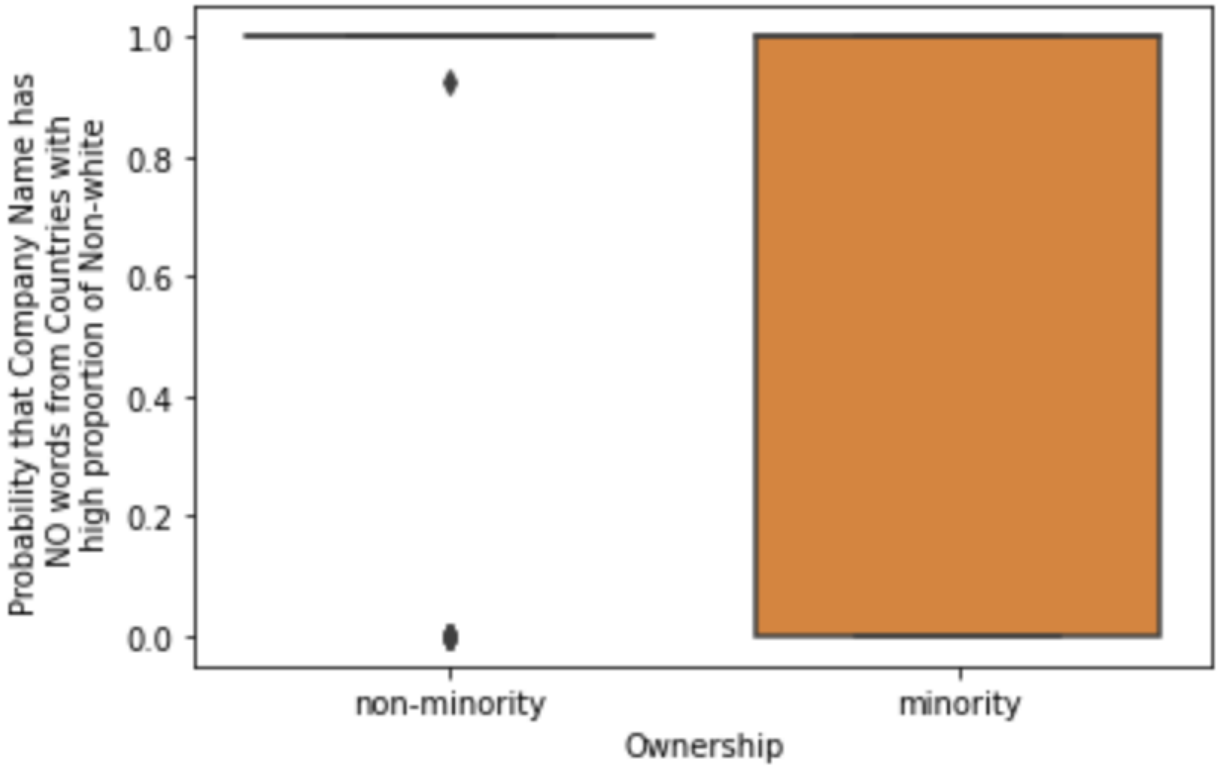
2. PREDICTIVE VARIABLES: Population by Race

- Using company address, we geolocate companies within Census tracts. Then, using the American Community Survey 5-Year Data, we determine the percentage of non-white people living in the Census tract where the company was located in the year 2020.
- On average, companies owned/managed by people from minorities, are located in Census tracts where a higher proportion of non-white people live relative to non-minority-owned companies.
- The percentage of non-white people at a company location discriminates between minority-owned and non-minority-owned companies.



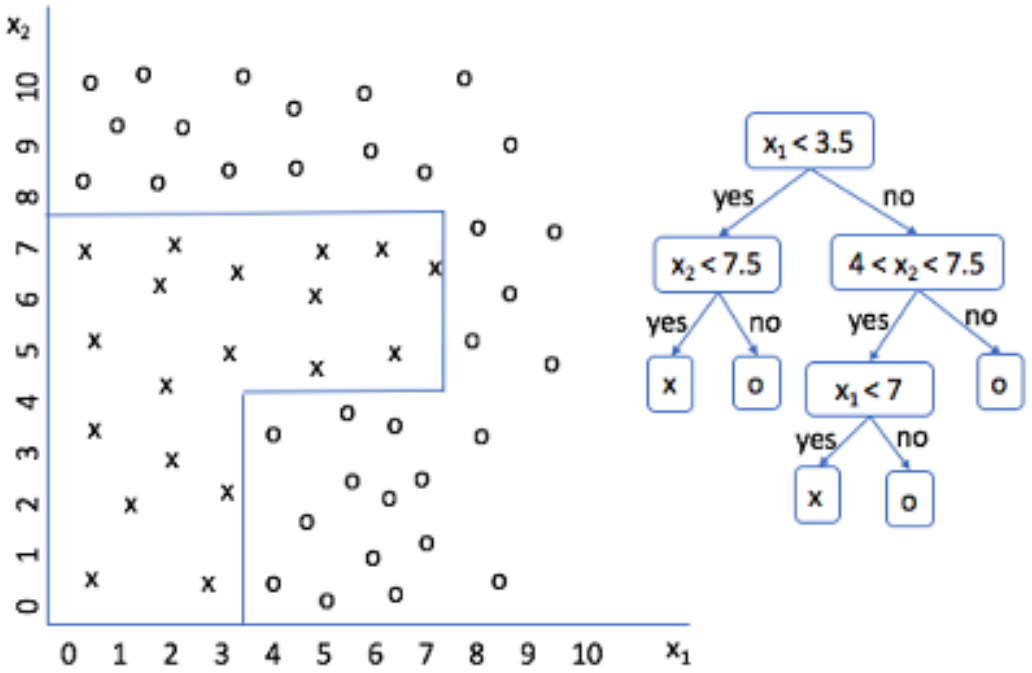
2. PREDICTIVE VARIABLES: Company Name

- We use the python package, Spicy, to classify the language of the words forming the company name. Words listed in company name were classified by country of origin and those countries were classified as minority and non-minority based on whether the majority population of those countries constitute a minority group in the United States.
- The average probability of non-minority-owned companies using words in languages from these majority-minority countries is 1, suggesting that this metric is useful in discriminating between minority-owned and non-minority-owned companies.



3. CLASSIFIER: Example of Decision Tree

Example of Decision Tree Model



- We split our labeled sample into training (80%) and testing (20%) sets.
- The training set is used to train our classifier to learn the pattern between the predictive variables and our labels (minority-owned vs non-minority-owned status).
- The test set was used to evaluate the performance of the classifier.
- We use a Decision Tree classifier; as a simple example, consider the figure where:
  - X1 and X2 as the predictive variables.
  - Sign x and 0 are our labeled sample.
- The decision tree will create linear rules on X1 and X2 to separate between the two labels x and 0.
- We compared this classifier with other classifiers (such as Random Forest), but they did not perform as well.

3. CLASSIFIER EVALUATION

		Decision Tree Predicted Minority Status		
		Minority-owned companies	Non-minority-owned companies	Total
Labeling from the test sample	Minority-owned companies	9	1	10
	Non-minority-owned companies	5	13	18
Total		14	14	28

- Using the Decision Tree classifier, we predict the minority-owned status of companies listed in our test set, then compare the prediction with the actual label applied to the business.
- The key statistics for the performance:
  - **Accuracy:** The percentage of companies correctly classified: (9 + 13) / 28 = 78.51%
  - **Precision:** Percentage of true minority-owned companies correctly identified as minority-owned: 9 / 14 = 64%
  - **Recall:** Percentage of true minority-owned companies labeled as minority-owned : 9 / 10 = 90%

4. PREDICTION

- We predicted the minority ownership status for the 12,123 companies in Mergent Intellect with executive names. Mergent Intellect indicated that only 3.52% of these companies were owned/managed by people from a minority group (based on race).
- Our model shows that 48.65% of the 12,123 companies were owned/managed by people from a minority group (based on race).
- **Mergent Intellect appears to undercount the number of minority-owned companies by almost 93%, mislabeling most minority-owned companies as non-minority-owned.**

		Decision Tree Predicted Minority Status		
		Minority-owned companies	Non-minority-owned companies	Total
Mergent Intellect with Executive names	Minority-owned companies	427	0	427 (3.52%)
	Non-minority-owned companies	5,472	6,224	11,696 (96.48%)
Total		5,899 (48.65%)	6,224 (51.34%)	12,123 (100%)

IDENTIFYING MINORITY-OWNED BUSINESSES SUMMARY

CONCLUSIONS:

- Using a Decision Tree classifier trained on company characteristics -- executive names, company location, company name -- we predict the minority ownership status of a subset of companies listed in Mergent Intellect.
- Although Mergent Intellect accurately classifies minority-owned businesses in certain industries, we found that they significantly undercounted minority-owned companies in others.
  - Approximately 48.7% of companies reported in our sample were listed as minority-owned, compared to 3.6% reported by Mergent Intellect.
  - Our estimate is much closer to the 38% reported by the US Census Bureau in 2017.

NEXT STEPS:

- We are working on another version of our model using a labeled sample from human coders. This version removes all previous assumptions, such as labeling companies using contextual information from the Civil Rights Acts or assuming that Mergent Intellect correctly identifies minority ownership.
- We are building a search engine to identify executive names given a company name. The goal is to identify the executives for the remainder of the 160,000 companies listed in Mergent Intellect. Filling this gap will allow us to predict the minority ownership status of all companies in our sample, as well as combine these results with our work measuring business activities to understand the geographic and industry distribution of minority-owned businesses in Fairfax County.