# CIND820 The Effect of Covid in the Early Stages on Mississauga Businesses – Final Report

# Jennifer McNenly [jmcnenly@torontomu.ca](mailto:jmcnenly@torontomu.ca)

# Student ID: 500105361

# Supervisor: Tamer Abdou Date:2023-04-03

Table of Contents

[Github Repository Details 3](#_Toc131426262)

[Revised Abstract & Research Questions 3](#_Toc131426263)

[Literature Review 6](#_Toc131426264)

[Project Methodology 11](#_Toc131426265)

[Data Preparation and Initial Analysis 11](#_Toc131426266)

[NAICS Clustering 12](#_Toc131426267)

[Prediction 12](#_Toc131426268)

[Data Structure 13](#_Toc131426269)

[Simplified Data Structure Table 13](#_Toc131426270)

[Data cleanup 16](#_Toc131426271)

[Univariate analysis of all attributes 17](#_Toc131426272)

[Bar plots for numeric attributes 17](#_Toc131426273)

[Box plots for numeric attributes 18](#_Toc131426274)

[Bar plots for categorical attributes by year 18](#_Toc131426275)

[Bivariate analysis of categorical values 21](#_Toc131426276)

[NAICS Clustering Results 24](#_Toc131426277)

[Predictive modeling / Classification 25](#_Toc131426278)

[Classification using Decision Tree 25](#_Toc131426279)

[Classification using Naïve Bayes 28](#_Toc131426280)

[Feature selection 32](#_Toc131426281)

[RFECV 32](#_Toc131426282)

[Filter based predictors 33](#_Toc131426283)

[Classification using SMOTE 33](#_Toc131426284)

[Effeciency of models 35](#_Toc131426285)

[Conclusions and Recommendations 35](#_Toc131426286)

[Industries 35](#_Toc131426287)

[Size of Business 35](#_Toc131426288)

[Age of Business 36](#_Toc131426289)

[Shortcomings of the work 36](#_Toc131426290)

[Contribution of the work compared to past research 36](#_Toc131426291)

# Github Repository Details

My private Github repository entitled “MississaugaBusinesses” has been shared with you and can be found at <https://github.com/mcnenlyj/MississaugaBusinesses>. I have uploaded the relevant Microsoft Word, Excel, html and pdf files. All coding files are in Python.

# Revised Abstract & Research Questions

Businesses in Mississauga have changed over the course of the Covid-19 pandemic. Some businesses closed and others have survived, and in some cases thrived, during this time period. How those businesses that survived and failed are classified and how it can be predicted which ones survived and why are key themes in my research project. The City of Mississauga has identified [6 key large knowledge-based industry sectors](https://www.thefutureisunlimited.ca/industries/)[[1]](#footnote-1) of current business growth that it believes will continue to grow and attract businesses and workers to live and work in Mississauga.

Investigating the research questions can be used to provide insight into the period under study as well as future years: What did businesses look like over the course of the data set? Did the ones that failed before 2020 look similar to the ones that failed after 2020? Which factors impacted the viability of businesses to survive during 2020, the first year of the pandemic (business size, NAICS code, location, etc…)? Do businesses that survived and thrived fall within city’s 6 key broad industry areas? Can we forecast which businesses failed in 2020 and why? Were smaller businesses more adversely affected in certain industries? Were businesses that had been in the Directory longer less likely to fail? What does the latest year of the directory tell us about the current state of businesses in Mississauga? Did business reduce their headcounts? Did industry share shrink in certain industries?

The [Mississauga Business Directory](https://data.mississauga.ca/datasets/mississauga::2021-mississauga-business-directory/explore?location=43.609143%2C-79.675702%2C12.14)[[2]](#footnote-2) is an annual directory data set published in the [City of Mississauga Open Data Catalogue](https://data.mississauga.ca/)[[3]](#footnote-3) set that lists businesses in Mississauga by postal code, size of business, NAICS code and NAICS description. There are 5 years of directories covering the period 2016 – 2021. 2020 appears to be the only year it was not collected. Each directory has a unique ID for a business that you can use to track how it has evolved from year to year. When an ID is no longer in the directory my assumption is that the business has closed and is no longer in business. When a new ID is added this denotes either a brand new business that opened in the interim or a previously unlisted business that is now formally listed in the directory. I note that the directory is voluntary and only businesses who agree are listed. I propose using the full set of directories for the years 2016, 2017, 2018, 2019 and 2021, as they represent the immediate pre-covid period and post-2020 period. The number of businesses dramatically dropped from 16518 to 14825 between 2019 and 2021. Data for 2022 was just published on March 30, 2023 and is too late to include in this Final Report. I propose reviewing the 2022 data as part of my Final Presentation to further investigate trends and the full impact of the pandemic on Mississauga businesses where the number of businesses appears to have rose again to 14825 to 15162, albeit with a lot of missing data in the data set provided.

The NAICS code system is very granular and provides too many levels. Classification (including decision trees, naïve bayes and random forest) will help us to group businesses that survived or failed into broader categories, effectively forecasting businesses that closed in Mississauga by 2021, after the first year of the pandemic.

# Literature Review

I was not able to find any studies specifically using the data set I have chosen. I am unaware of anyone else releasing a public study of an analysis of the Mississauga Business Directory either before Covid or recently. However, from my literature review, I was able to find studies of businesses in other jurisdictions that asked pertinent questions both in the pre-pandemic and pandemic periods.

I was able to find and purchase a very useful chapter on using machine learning to predict business survival during Covid-19 from businesses listed in Yelp and other sources. This resource was not available on the Toronto Metropolitan University library and I have included the chapter in my Github repository [here](https://github.com/mcnenlyj/MississaugaBusinesses/blob/main/Bibliography_Resources/Using%20Machine%20Learning%20to%20Predict%20Business%20Survival%20in%20COVID-19.pdf) for your convenience. The researchers, Garthi and Mathur, used a variety of classifiers, namely KNN, Logistic Regression, Random Forest and Neural Network to make predictions about the survival of these businesses and assessed each model’s performance. They found that all models were better at predicting open businesses than closed ones by having higher precision, recall and *F*1 scores for Open Businesses. They noted that Random Forest is the best according to the sensitivity metric, KNN is the best according to the specificity metric and Neural Network is the best when balancing the sensitivity and specificity metrics (balanced accuracy). Their conclusion was that the presence of four Covid features suggested that the survival of a business was tied to providing services that ensured customer health and safety. [[4]](#footnote-4)

A survey of small businesses across several US states released in the first few months of the pandemic found that mass layoffs and closures had already occurred. The sample found 43 percent of businesses had temporarily or permanently closed and additionally, on average, reduced their head count of employees by 40 percent. Further, the same study found that due to the forced closure of some businesses, while others deemed essential remained open, created an “existential threat” to businesses in certain industries, most notably, *Arts and Entertainment*, *Tourism and Lodging* and *All Retailers except Grocery*. Businesses with fewer employees were most adversely affected. [[5]](#footnote-5)

A pre-pandemic study by Zhang and Stevens shows that businesses that are older and more established contribute to employment growth and thus have larger employee head counts over time. In over half of the industries studied, but most evidently in *Manufacturing*, *Wholesale Trade* and *Education Services*. Other industries like *Construction* and *Other Services* were less likely to grow and the *Information* and *Finance & Insurance* industries the age of a business was not found to be statistically significant.[[6]](#footnote-6) Hyatt, on the other hand argues that, though firm age is more important than business in driving employment growth, it is the younger new businesses (less than 10 years old) that are driving job growth. These younger businesses have a higher need for credit and borrowing compared to older businesses.[[7]](#footnote-7)

A pre-pandemic Canadian study by Dixon for Statistics Canada found little evidence to support the assertion that smaller businesses have proportionally higher employment growth rates year-on-year. Employment growth rates rose with business size for businesses with fewer than 20 employees. In businesses above the 20-employee threshold no relationship emerges between employment growth and business size. These results are consistent with the average proportionate growth condition of Gibrat’s Law; the assertion of French economist Robert Gibrat that average employment growth is independent of business size.[[8]](#footnote-8)

Another Statistics Canada survey early in the pandemic found that in Ontario staffing actions taken by businesses showed on average 38.9 percent laid off staff and only 2 percent hired more staff. This compares to nationally where on average 40.5 percent laid off staff and only 2.5 percent hired more staff. In the same survey industries by NAICS code across Canada that fared the worst included *Accommodation and Food Services [72]*, *Retail Trade [44-45]* and *Construction [23]* where 67.5 percent, 54.9 percent and 50.7 percent of businesses laid off staff respectively. Those industries impacted negatively the least included *Agriculture, forestry, fishing and hunting [11]*, *Management of companies and enterprises [55]* and *Finance and insurance [52]* where 14.4 percent, 14.6 percent and 17.8 percent of businesses laid off staff respectively. In terms of size of business, the hardest hit were those with *20-99 employees*, *5-19 employees* and *100-249 employees* where 61.7 percent, 58.5 percent and 54.5 percent of businesses laid of staff respectively. Those businesses that fared the least negatively were those with *0 employees*, *1-4 employees* and *250-499 employees* where 5.1 percent, 31.4 percent and 49.3 percent of businesses laid of staff respectively. This includes all employees who would receive a T4. Excluded from number of employees are business owners, contract workers and other personnel who would not receive a T4.[[9]](#footnote-9) Statistics Canada followed this up later in the same year with similar results for most industries and business sizes in terms of status of being open or closed.[[10]](#footnote-10)

McKinsey published a study early in the pandemic on the impact of Covid-19 on small businesses in the US and found that businesses with less than 100 employees where the hardest hit and most vulnerable. In addition, *Accommodations and food services*, *Arts, entertainment and recreation* and *Personal services* were the most vulnerable industries for businesses. *Healthcare and social assistance*, *Finance and insurance* and *Professional, scientific, and technical* industries were among the least vulnerable. [[11]](#footnote-11) But a mere month later McKinsey reported that Finance and insurance and Healthcare and social assistance had also moved into the most vulnerable category.[[12]](#footnote-12) The takeaway being the things moved rapidly at the beginning of the pandemic and accurate data was difficult to obtain.

Clustering of industries is another avenue of research that has been performed in many studies. Recent examples, like Wang and Wen (2021), looked at homeownership, income, education level and the relationship to the type of businesses that were opened in neighbourhoods.[[13]](#footnote-13)

The data set I have chosen does not contain demographic information on the residents (homeowners or renters) in each postal code. I will therefore not address any correlation between where people live, nor their demographic characteristics such as income, education and the types of industries and businesses that are nearby. Further, the data set does not contain any detailed demographic and financial information about the businesses, such as income, expenses, and demographic details about the owner of the business. I may be able to show clustering of certain industries but into certain areas of the city but there is no data to imply causation other than the fields I have in the combined data set.

I believe my work is worth doing as there is not a detailed study of the effect of the Covid-19 pandemic on businesses in Mississauga. The results will fit in with what has gone before in terms of answering questions about the size of businesses over time and overall effect on economic growth (or in the case of the pandemic, shrinkage). It will also answer questions about the type of industries that were affected either positively or negatively by the pandemic. Does what was reflected in the Statistics Canada’s national and provincial survey early in the pandemic mirror what happened to businesses and industries in a city like Mississauga? Covid-19 was an unforeseen emergency that strained the global economy. The ramifications of its impact will be felt for years to come.

# Project Methodology

****

# Data Preparation and Initial Analysis

I prepared a combined data set to do EDA analysis. Then added new fields to the data set that can be created after examination of the data. The first new field would be the year of the Directory the records came from when combined. The second additional field will note if a business was no longer listed in a subsequent year of the directory and is assumed closed. The third additional field will note if a business was net new to the directory in a given year. The fourth additional field will note the age of a business, that is, the number years a business has been listed in the directory. The fifth field would be a new unique ID to provide to the newly combined data set.

Next steps would be to transforming fields, where necessary, into useful data types for analysis. Analysis by year can then be done to see the broad patterns in the data. For example, for businesses that closed each year what industries were they in and what size of business?

# NAICS Clustering

Most studies in the literature review dealt with what the early months of the pandemic looked like in terms of the kinds of industries that were negatively impacted. I would propose looking at the data in a similar way to see if Naics clustering of closed businesses yields similar results to what was seen in broader US and Canadian studies.

# Prediction

Dividing the data into Training sets before 2019 and Test 2019 & 2021 rather than 80 / 20 split would most likely not yield good results for prediction since Covid was a unforeseen, highly negative business event. Rather, using an 80 / 20 split of the complete data set for all years would likely yield better results after feature engineering. Beyond the fields I propose to add to the data set during data preparation I will then see what other feature engineering techniques can be used to parse the data set into a useful tool for creating a training and test set. I will plan to use Decision Tree, KNN, Naïve Bayes to compare results.

# Data Structure

## Simplified Data Structure Table

**In the original data set there are 8 numeric attributes, and 16 categorical attributes (16 nominal). In the revised table there are 8 numeric and 18 categorical attributes. I converted 1 of the numeric attributes to categorical-nominal (NAICSCode), 1 of the numeric attributes to categorical-ordinal (EmplRange), and 7 of the categorical-nominal to categorical-binary (Phone, Fax, TollFree, WebAddress, BldgNo, UnitNo) when I cleaned up the data set for better analysis. I additionally reduced NAICSCode to the higher level 2 digit code and reduced PostalCode to the first 3 characters so that the businesses could be grouped more easily into less granular buckets.**

**I added 5 new fields – Year (categorical-nominal), Age (categorical-nominal), Closed (categorical-binary), isnew (categorical-binary), RecordID (numeric-uniqueID).** (See fields in green below at the bottom of the table.)

**I eliminated BIA\_Name, BIAFulName and CHArea due to too much missing data. I eliminated BusinessID, NAME as these fields were unique and too granular to provide useful insights.**

I have summarized the type of each feature in the table below :

|  |  |  |
| --- | --- | --- |
| Field Name | Field Type | Description |
| FID | Real number  Quantitative-Numeric | Unique ID in the uncombined data sets |
| BusinessID | Real number  Quantitative-Numeric | Business ID - eliminated this field as too granular |
| NAME | Categorical-Nominal | Business name - eliminated this field as too granular |
| EmplRange | Categorical-Ordinal{1 to 4,10 to 19,20-99,100-299,300-499,500-999,1000+} | Employee Range |
| NAICSCat | Categorical-Nominal | NAICS Category |
| NAICSDescr | Categorical-Nominal | NAICS Description |
| NAICSCode | Quantitative-Numeric;  Converted to Categorical-Nominal using first 2 digits of the code {11,21,23, etc…91} | NAICS Code |
| Phone | Categorical-Binary  Converted to Binary {Yes;No} to note if the business has phone # | Business Phone |
| Fax | Categorical-Binary  Converted to Binary {Yes;No} to not if the business has fax # | Business Fax |
| TollFree | Categorical-Binary  Converted to Binary {Yes;No} to note if the business has tollfree # | Business Toll Free Number |
| Email | Categorical-Binary  Converted to Binary {Yes;No} to not if the business has email address | Business Email |
| WebAddress | Categorical-Binary  Converted to Binary {Yes;No} to not if the business has website | Business Web Address |
| StreetNo | Real number  Quantitative-Numeric | Address - Street Number |
| StreetName | Categorical-Nominal | Address - Street Name |
| Address | Categorical-Nominal | Address – Street Number and Street Name |
| PostalCode | Categorical-Nominal  Reduced to 36 categories using first 3 characters of the PostalCode {L4G…M8W} | Address – Postal Code |
| BldgNo | Categorical-Binary  Converted to Binary {Yes;No} to note if the business has BldgNo | Address – Building Number |
| UnitNo | Categorical-Binary  Converted to Binary {Yes;No} to note if the business has UnitNo | Address – Unit Number |
| EmplUpdate | DateTime  Quantitative-Numeric | Date Employment Number was Updated - eliminated this field and used Year instead |
| PIN | Real number  Quantitative-Numeric | PIN code for longitude and latitude |
| X | Decimal  Quantitative-Numeric | Latitude |
| Y | Decimal  Quantitative-Numeric | Longitude |
| CHArea | Categorical-Nominal | Character Area - eliminated this field due to lack of data |
| BIA\_Name | Categorical-Nominal | Business Improvement Association – Acronym - eliminated this field due to lack of data |
| BIAFulName | Categorical-Nominal | Business Improvement Association – Full Name - eliminated this field due to lack of data |
| Ward | Real number  Quantitative-Numeric | Ward |
| Year | Real number  Quantitative-Numeric | Year of the Directory the data is from |
| Age | Real number  Quantitative-Numeric | How many years the business has been listed in the Directory |
| Closed | Categorical-Binary | Did the business close in the subsequent year? |
| isnew | Categorical-Binary | Did the business get listed in the current year? |
| RecordID | Real number  Quantitative-Numeric | New unique ID for combined data set – eliminated this field as too granular |

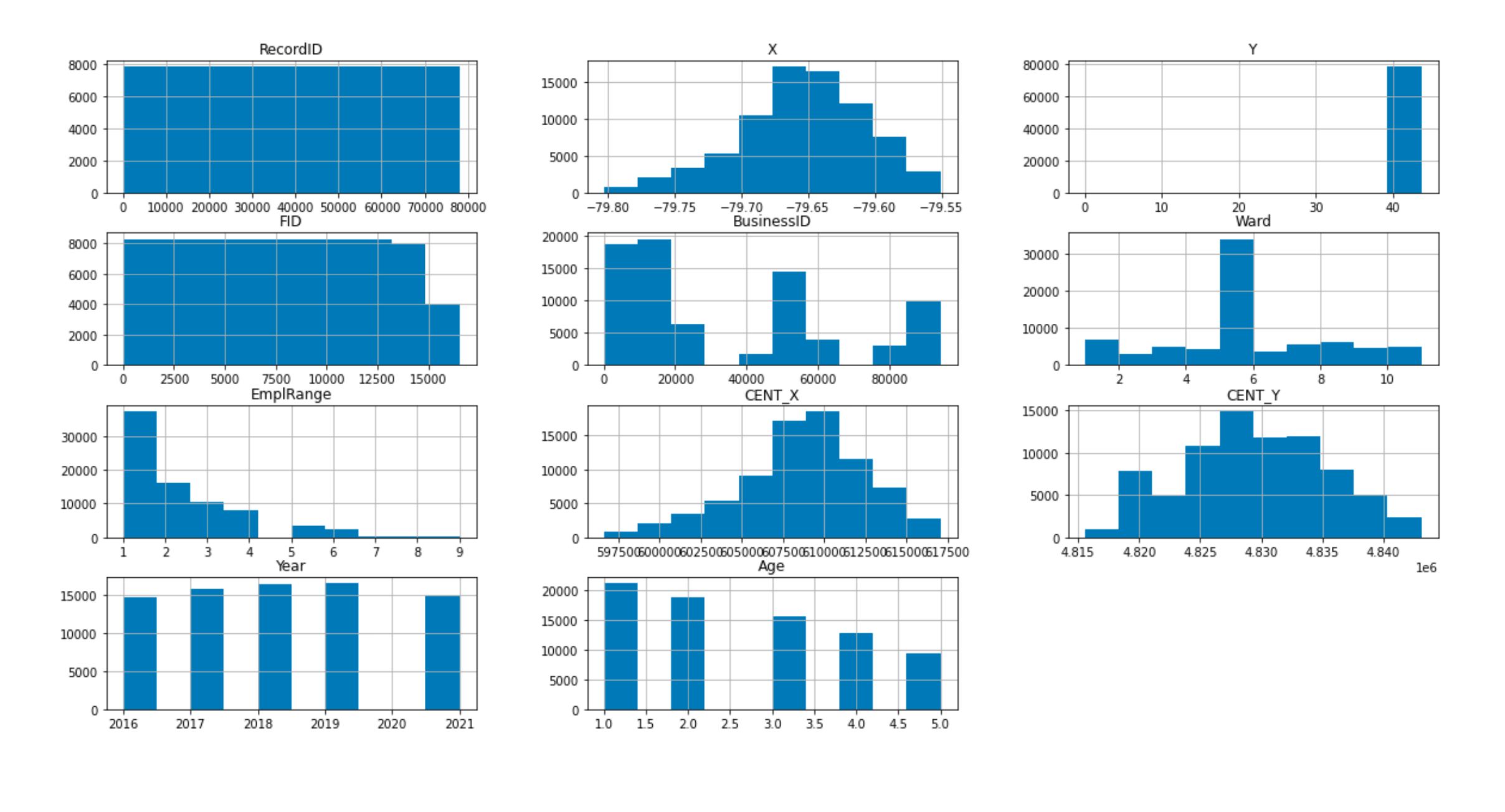
# Data cleanup

There are missing values as not every annual directory included the full business details in a given file regardless of stating on the data source that the data set contained those values. I sorted the data set by BusinessID and then used the forwards fill, backwards fill as the details were confirmed upon review of data sets to stay the same for each business when filled in successive or previous years. Some fields were labelled differently or put in the wrong column in certain years. I renamed those fields correctly as I combined the data sets using concatenate and you can see my notes and references in the TestTrainCombinedData file.This was predominantly around **NAICSCat**, **NAICSDescr** and **X**, **Y** values. **EmplRange** was simplified to a straight ordinal scale from 1-9 to represent sizes of business from 1-4 employees up to 1000+ employees {'1 to 4': 1, '5 to 9': 2, '10 to 19': 3,'20 to 49': 4, '50 to 99': 5, '100 to 299': 6, '300 to 499': 7, '500 to 999': 8, '1000+': 9 }

I have mentioned removing fields that had too many missing values and that forwards, backwards fill could not address **BIA\_Name, BIAFulName** and **CHArea**. There were essentially making them a null value and less useful to the analysis. Outliers were few and I was able to fix those that were obviously an error in data entry by reviewing the data attribute for the same business by their ID in other directory year. This was preferred over eliminating them. Further, the **X** (latitude) field uses negative numbers to denote location coordinates. Negative values meant that GaussianNB instead of MultinomialNB had to be used for the Naive Bayes analysis on the dataset. Removing negative values for **X** would allow for running MultinomialNB.

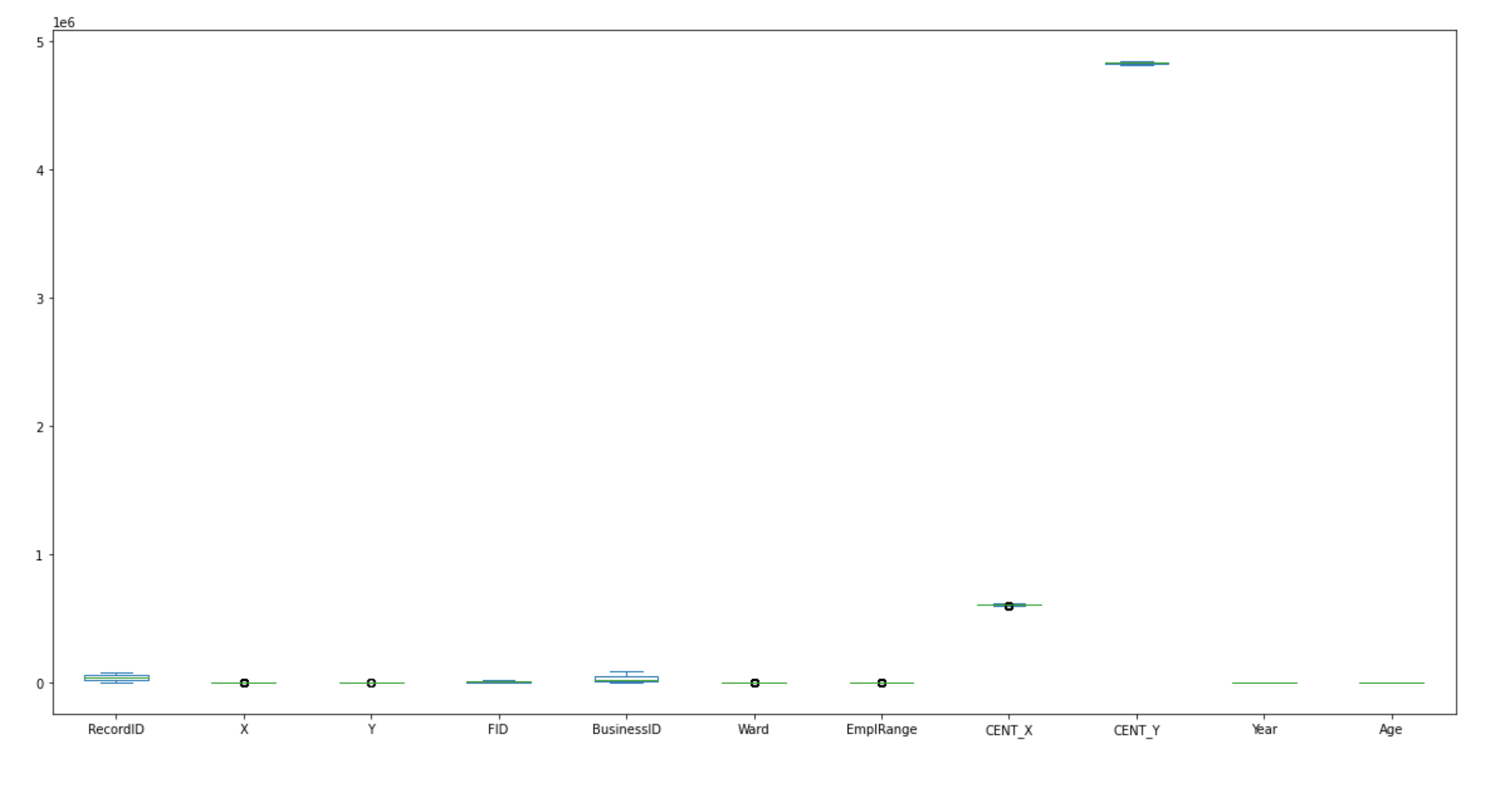
# Univariate analysis of all attributes

## Bar plots for numeric attributes



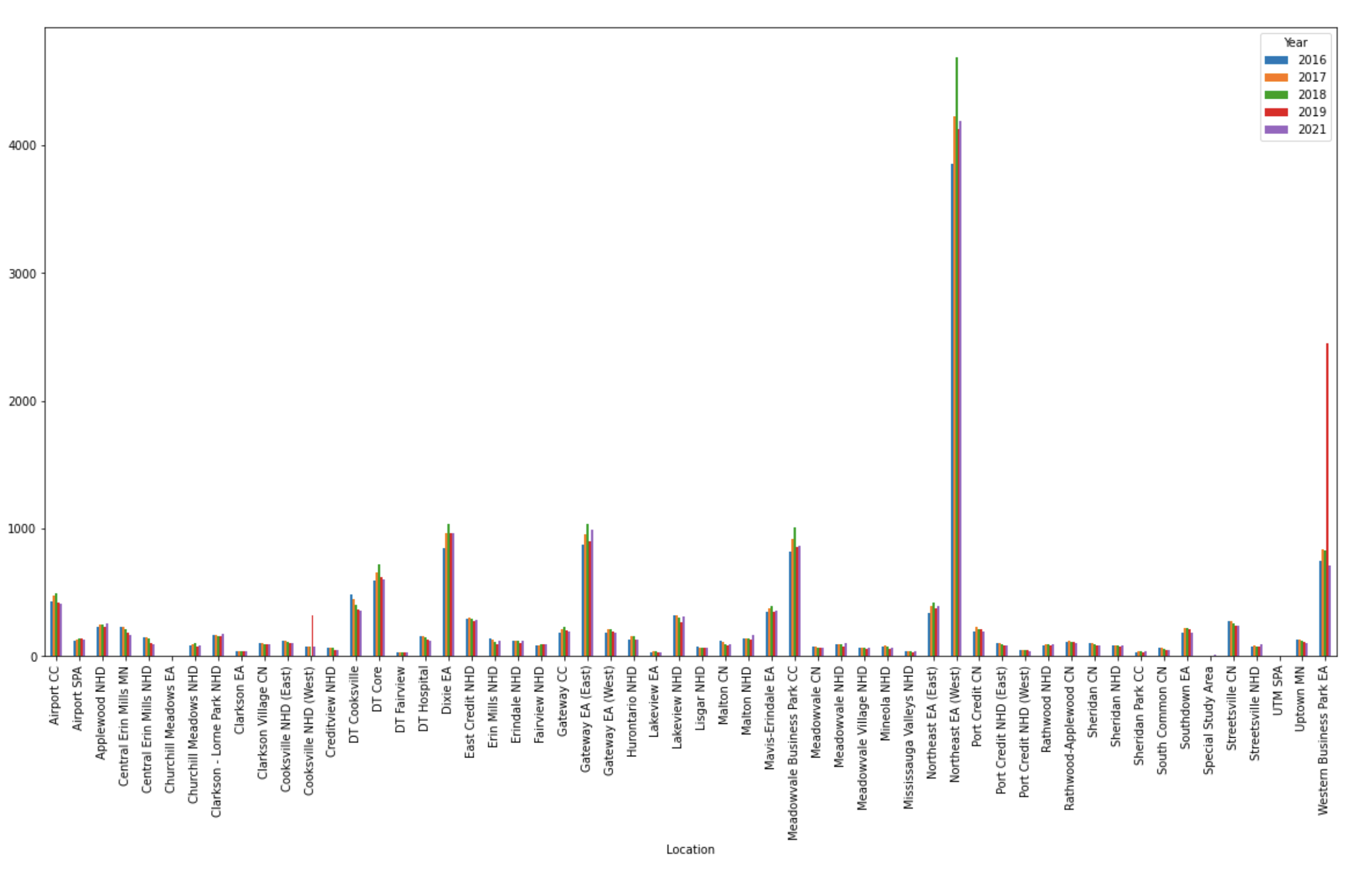
**EmplRange**, which I converted to Categorical, and **Age** were Right Skewed

## Box plots for numeric attributes

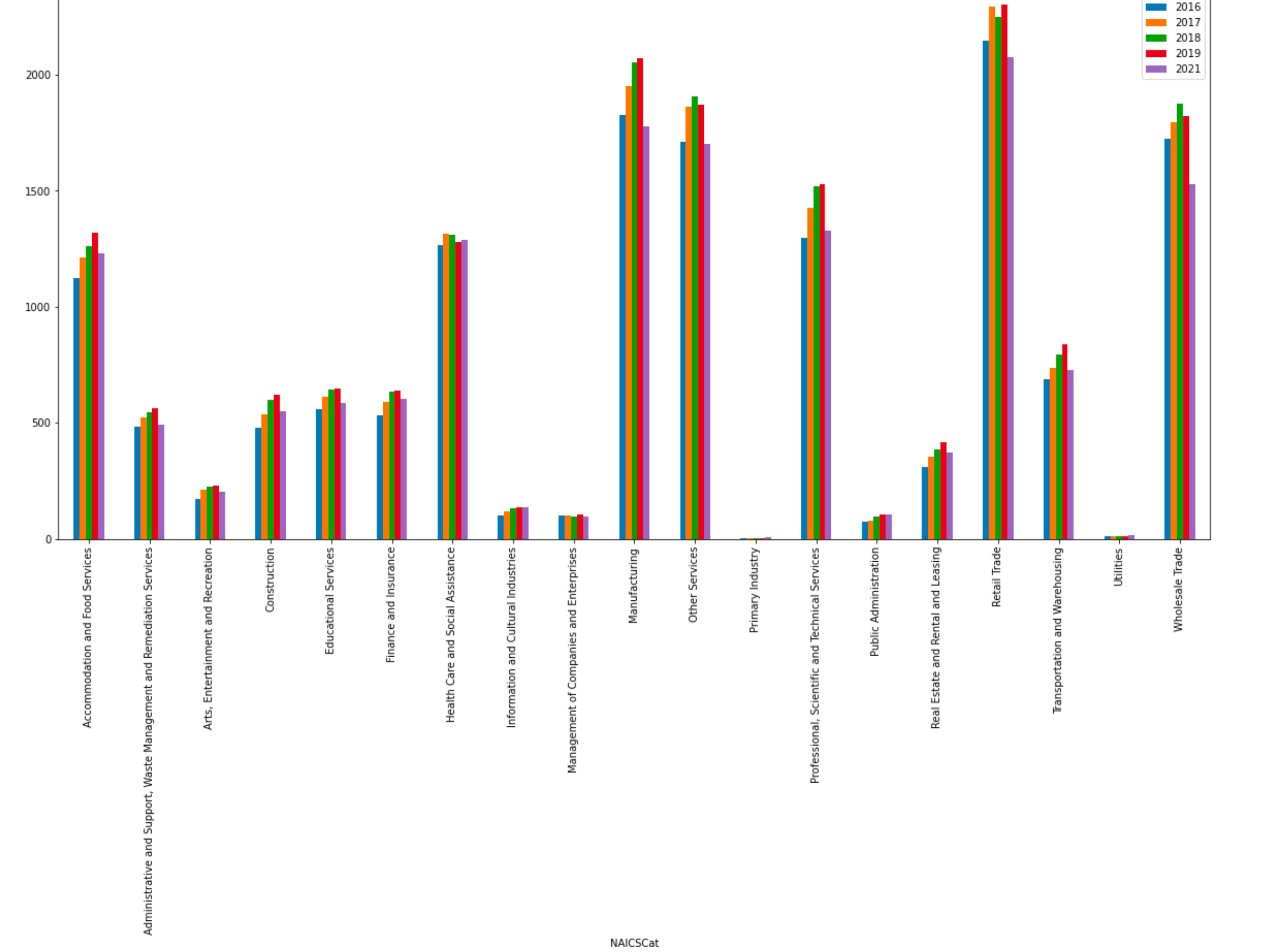


**Outliers** were easily eliminated by correcting individual records rather than eliminiating them.

## Bar plots for categorical attributes by year

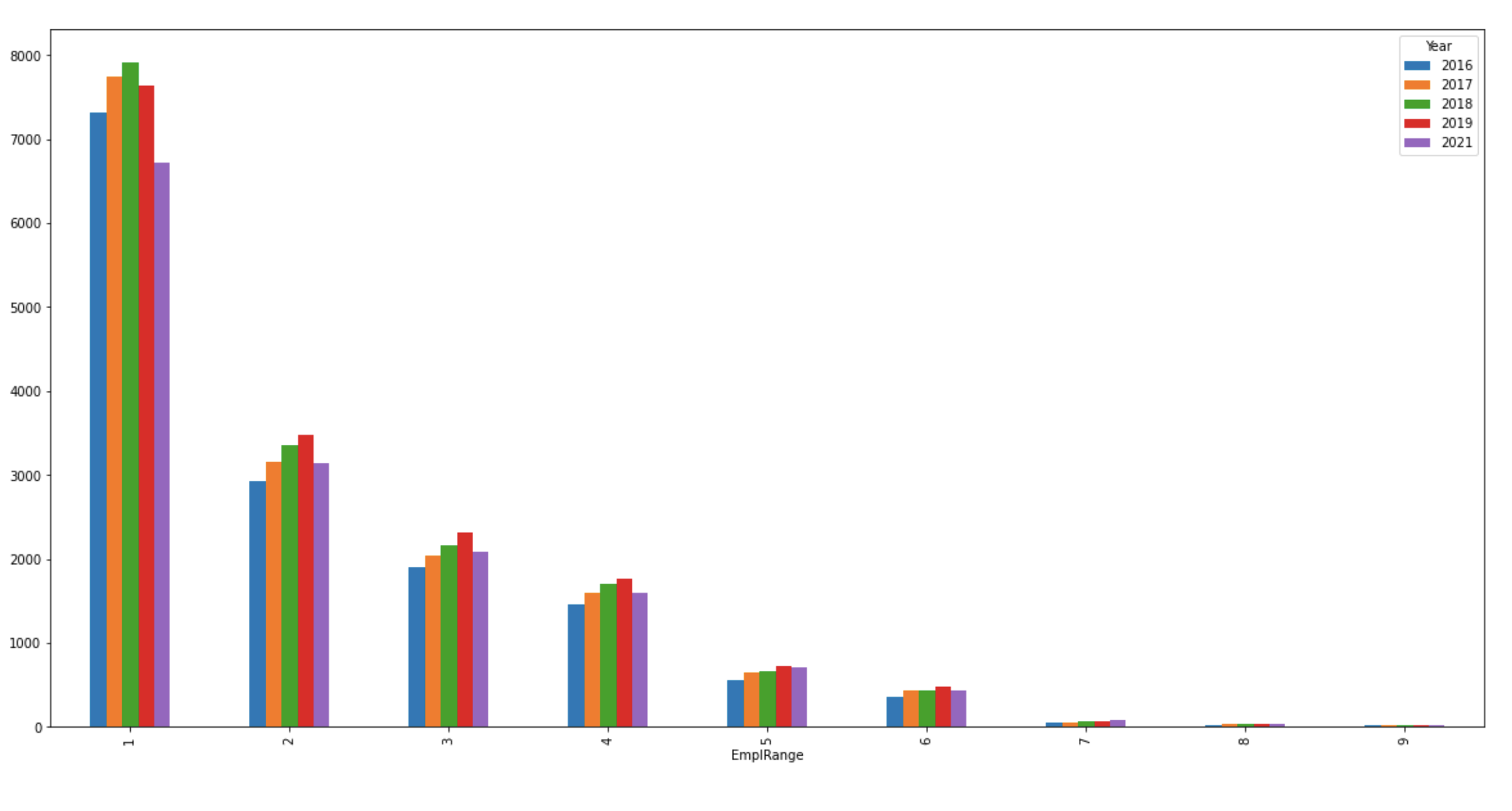


Most businesses, not surprisingly, are concentrated around Northeast EA (West) which is east of Pearson airport and includes the Square One Shopping mall.



Wholesale Trade, Manufacturing and Retail Trade saw the most number of business closures by 2021. By percentage closure then Professional, Scientific and Technical Services is 3rd.

|  |  |
| --- | --- |
| Table  Description automatically generated | Table  Description automatically generated |



Smaller businesses are more prevalent in the data set. Right skewed imbalance.

## Bivariate analysis of categorical values

Chart, bar chart

Description automatically generated

**isnew** – Is a business new? Yes/No - New business openings tapered off year by year, most dramatically in 2021. Older or more established businesses are more prevalent in the data set. Right skewed imbalance.

Chart, bar chart

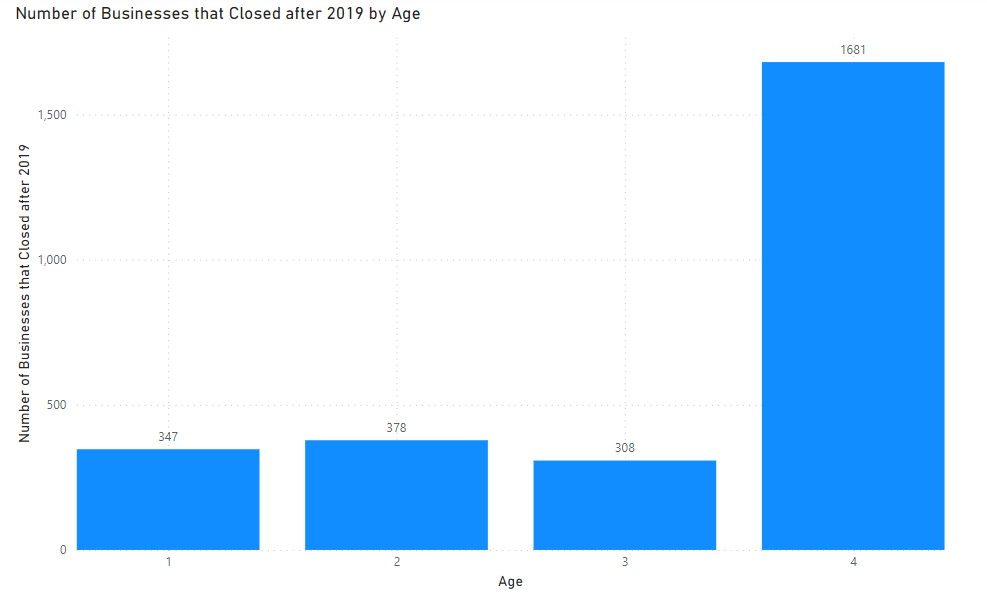
Description automatically generated

Chart, bar chart

Description automatically generated

Age of businesses after 2019 – Most are over 4 years old which is to be expected. Closures follow a similar pattern

|  |  |
| --- | --- |
| Table  Description automatically generated | Table  Description automatically generated |



Chart

Description automatically generated

Size by business that closed after 2019 – Most were the smallest (1-5 employees)

|  |  |
| --- | --- |
| Table  Description automatically generated | Table  Description automatically generated |

Chart, bar chart

Description automatically generated

Ward where businesses closed after 2019 – Most were in Ward 5 which includes the Square One mall

## NAICS Clustering Results

Industries clustered around NAICSCode in the 40s and 50s with size fairly evenly spread

|  |  |
| --- | --- |
|  |  |

Industries clustered around NAICSCode in the 40s and 50s with size fairly evenly spread

|  |  |
| --- | --- |
|  |  |

# Predictive modeling / Classification

## Classification using Decision Tree

The cross-validation methodology used is splitting the dataset into train set (80%) and test set (20%), which is a well-adapted method for a relatively non-small data set. The categorical variables were transformed into factors using LabelEncoder in Python sklearn.

Diagram

Description automatically generated

The root node ‘Location’ has the best impurity index and splits the data into 2 subsets: greater than / less than 54.5. The second node is ‘Address’, which has 2 branches :<=4843} and >4843. ‘X’ and ‘NAICSDescr’ are the next nodes in the 3rd level. In total, 3 attributes are included in the tree for businesses where ‘Closed’ = Yes, the attributes **‘NAICSDescr’**, **‘Email’** and **‘StreetName’**.

It is worth mentioning those leaf nodes have a ‘yes’ level greater than 50%, which are:

* Location <=54.5 and NAICSCode>20.5 and X>-79.554 and NAICSDesc <=247.5
* Location <=54.5 and NAICSCode>20.5 and X<=79.554 and Y>43.725 and NAICSDescr >106
* Location >54.5 and Address <=4843 and X>-79.717 and NAICSDescr <=4.0 and StreetName <=214.0
* Location >54.5 and Address>4843 and NAICSDescr<=14.5 and Email<=0.5

After testing the model on 20% of the data I got the following results:

|  |  |  |
| --- | --- | --- |
|  | actual |  |
| predicted | No- | Yes+ |
| No- | 2752 | 10 |
| Yes+ | 537 | 5 |

|  |  |
| --- | --- |
| Decision Tree | Baseline model |
| Accuracy | 0.83 |
| Precision P+ | 0.33 |
| Precision P- | 0.84 |
| Recall TPR | 0.01 |
| Recall TNR | 1.00 |
| F+ | 0.02 |
| F- | 0.91 |
| G-Mean | 0.1 |

The baseline model did better on predicting the actual ‘no’ values (Recall TNR=100%) but did not that good when it comes to predicting the ‘yes’ values (Recall TPR=1%). F-measure, which is not affected by outliers, is 2%. G-mean score is higher than the F+ score, as it is, by definition, less prone to the effects of the imbalanced distribution of the class variable.

Decision Tree confusion matrix and metrics before changing negative **X** attribute values to positive were all best with 5 levels on the Decision Tree as compared to 4 or 6.

**5 levels – best!**

Text

Description automatically generated with low confidence

4 levels – less TP less FP

A picture containing chart

Description automatically generated

6 levels – more FP

Chart

Description automatically generated with medium confidence

**5 levels – best!**

Table

Description automatically generated

4 levels – precision goes down to .22, recall to 0 and f1 score to 0.01 – G-Mean = 0

Table

Description automatically generated

6 levels – Precision goes down to .21 while recall, f1 score and accuracy stay the same. G-Mean = 0.099

Table

Description automatically generated

## Classification using Naïve Bayes

The same cross validation strategy was used for the Naive Bayes classifier, i.e. train the model on 80% of the data and test it on the remaining 20%, setting the same seed as 5 for the decision tree. The categorical variables were transformed into factors using OneHotEncoder in Python sklearn.

Diagram

Description automatically generated

The root node ‘Location\_Western Business Park EA’ has the best impurity index and splits the data into 2 subsets: greater than / less than 0.5. The second node is ‘X’, which has 2 branches :<=79.716 and >79716. ‘Age’ and ‘Y’ are the next nodes in the 3rd level. In total, 3 attributes are included in the tree for businesses where ‘Closed’ = Yes, the attributes **‘X’**, **‘Y** and **‘Age’**.

It is worth mentioning those leaf nodes have a ‘yes’ level greater than 50%, which are:

* Location\_Western Business Park EA <=0.5 and Age>2.5 and Y>43.675 and Y>43.732 and WebAddress>0.5
* Location\_Western Business Park EA >0.5 and X<=79.716 and X<=79.675 and Y<=43.514
* Location\_Western Business Park EA >0.5 and X<=79.716 and X>79.675 and Y<=43.539 and Y>43.538
* Location\_Western Business Park EA >0.5 and X>79.716 and X<=79.719 and Y <=43.529 and X<=79.717
* Location\_Western Business Park EA >0.5 and X>79.716 and X<=79.719 and Y <=43.529 and X>79.717
* Location\_Western Business Park EA >0.5 and X>79.716 and X<=79.719 and Y 43.529 and Age<=3.5

After testing the model on 20% of the data, we get the following results:

|  |  |  |
| --- | --- | --- |
|  | actual |  |
| predicted | No- | Yes+ |
| No- | 2768 | 0 |
| Yes+ | 536 | 0 |

|  |  |
| --- | --- |
| Decision Tree | Baseline model |
| Accuracy | 0.81 |
| Precision P+ | 0.24 |
| Precision P- | 0.84 |
| Recall TPR | 0.08 |
| Recall TNR | 0.95 |
| F+ | 0.12 |
| F- | 0.89 |
| G-Mean | 0.28 |

**Except for the recall TNR and G-Mean, all the other metrics are worse than the decision tree model. Accuracy remained unchanged between the two.**

Naïve Bayes (GaussianNB) confusion matrix before changing negative **X** attribute values to positive.

Table

Description automatically generated

A picture containing table

Description automatically generated

**Changing to MultinomialNB after replacing the negative ‘X’ attribute values with positive numbers did not improve the model!**

Table

Description automatically generated

Table

Description automatically generated

MultinomialNB Table

Description automatically generated

Text

Description automatically generated with medium confidence

Table

Description automatically generated

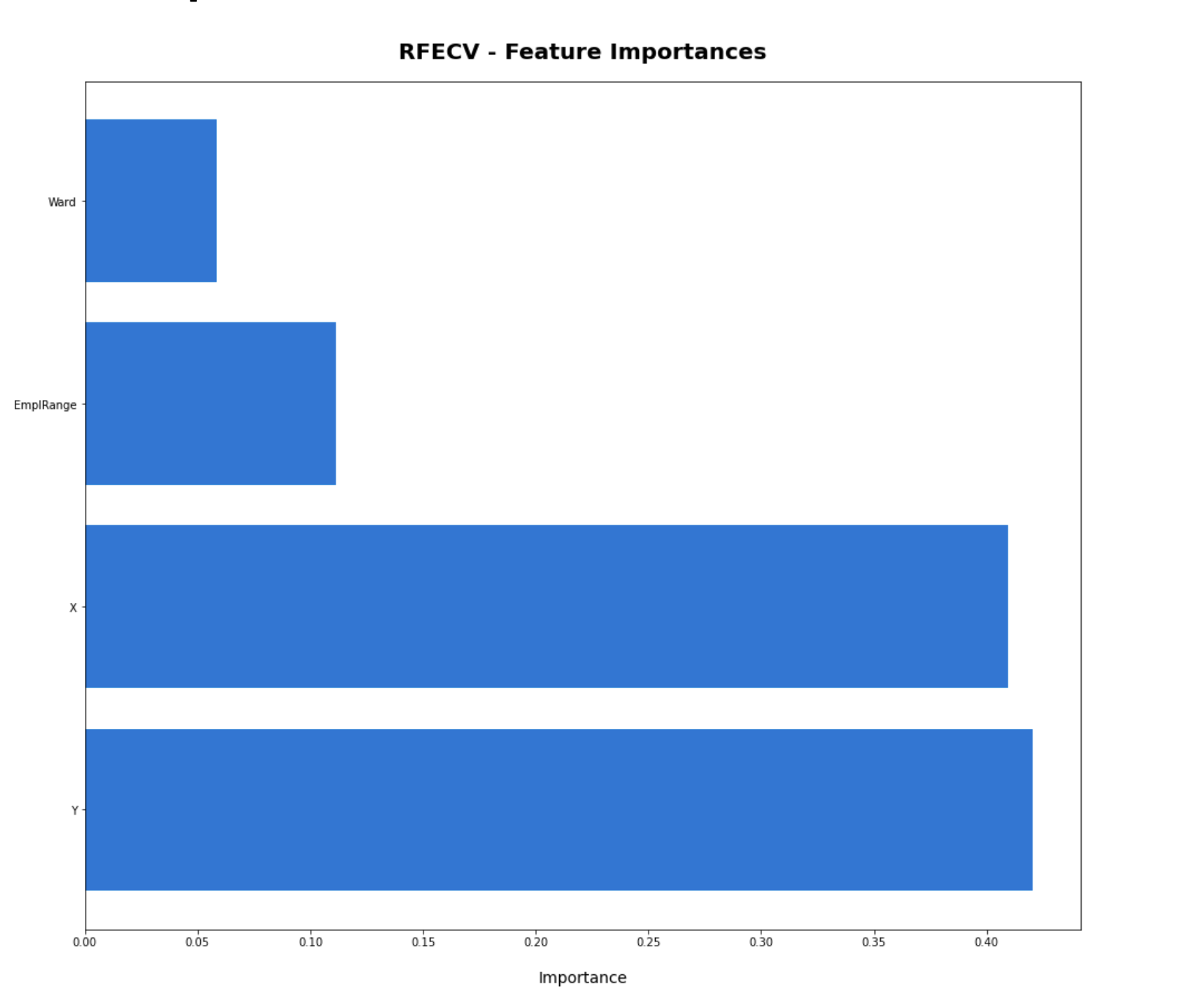
Table

Description automatically generated

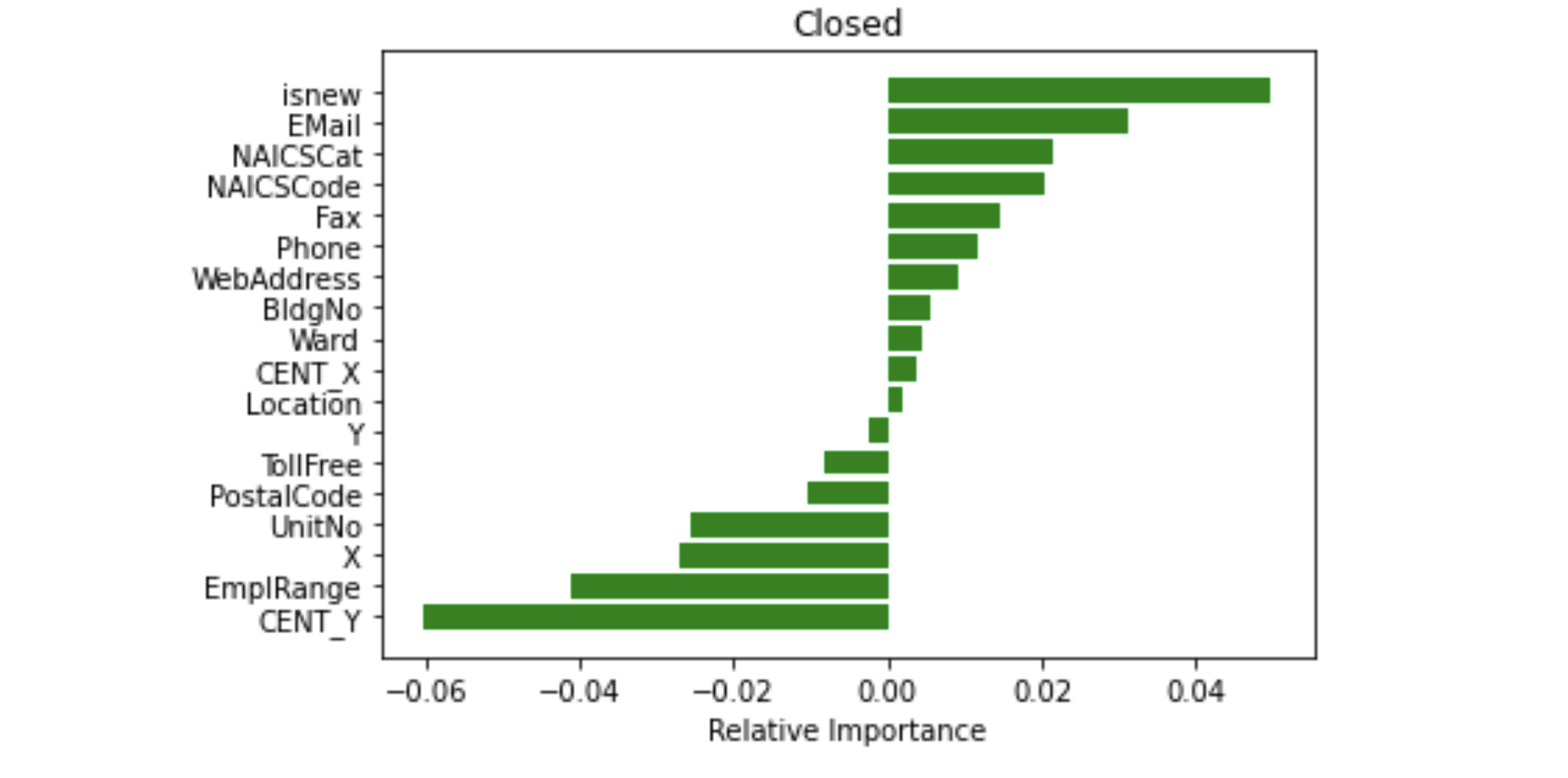
# Feature selection

## RFECV

RFECV found 4 features – Ward, EmplRange, X, Y but only X and Y were found to be significant and over the 0.4 threshold. This is consistent with what the Naïve Bayes classifier found in its decision tree.



## Filter based predictors



**The filter based predictor found that nothing was greater than 0.4! So nothing is correlated to Closed!**

# Classification using SMOTE

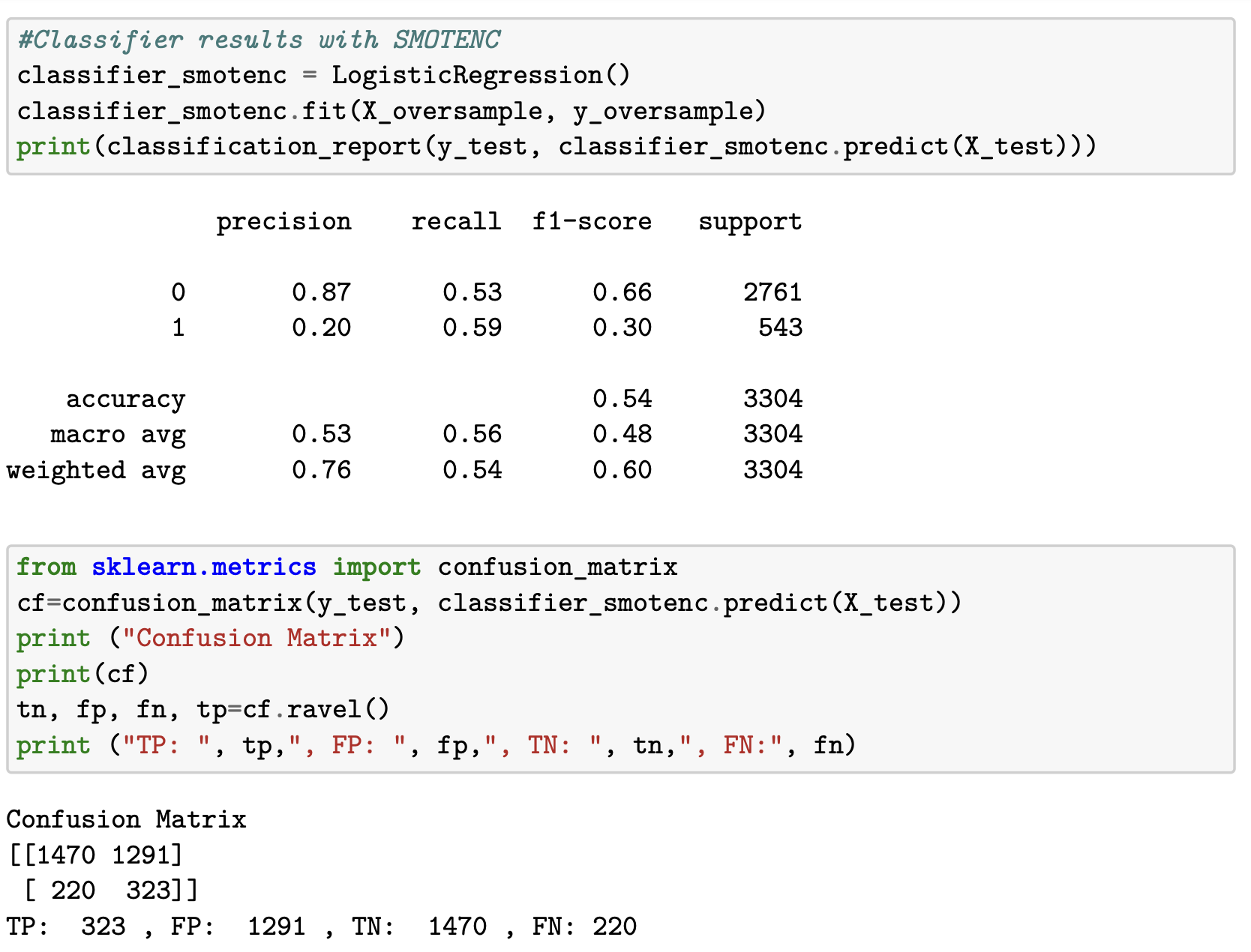
I tried various SMOTE methods to deal with the unbalanced dataset that had mostly ‘No’ as values for the field ‘Closed’.

Classifier results with SMOTENC improved on the existing scores using one continuous and one categorical feature, ‘**EmplRange**’ and ‘**NAICSCode’** with the target ‘**Closed**’

|  |  |  |
| --- | --- | --- |
|  | actual |  |
| predicted | No- | Yes+ |
| No- | 1470 | 1291 |
| Yes+ | 220 | 323 |

|  |  |
| --- | --- |
| Decision Tree | Baseline model |
| Accuracy | 0.54 |
| Precision P+ | 0.20 |
| Precision P- | 0.87 |
| Recall TPR | 0.59 |
| Recall TNR | 0.53 |
| F+ | 0.30 |
| F- | 0.66 |
| G-Mean | 0.56 |

SMOTENC improved on several scores, not surprisingly, as it balanced the data set. **Compared to Decision Tree - Accuracy is down from .83 to .54. Precision+ is down from .33 to .20 and Precision– is up from .84 to .87. TPR recall is up from .1 to .59 and TNR is down from 1.0 to .59. F+ is up from .02 to .30 and F- is down from .91 to .66. G-Mean is up from .1 to .56**



# Effeciency of models

GuassianNB was slightly faster than the DecisionTreeClassifier, however DecisionTree results were still marginally better than Naïve Bayes. I noticed as I cleaned data the speed of both models increased rapidly. An uncleaned data sets took several minutes as opposed to a fraction of second once cleaned.

# Conclusions and Recommendations

In summary, my results supported that the business survival was directly related to the size of the business. The smallest were the most likely to fail and the risk reduced as the size of the business increased. Prediction, though it produced results that were okay for if a business did not close did not produce acceptable results for if a business actually closed.

## Industries

Based on **percentage increase or decrease the effect on industries** was as followes : My industries that survived the **best were Primary Industry (+1/+20%), Utilities (+2/+14.29%) and Health Care and Social Assistance (+6/+0.47%)**. My industries that fared the **worst were Wholesale Trade (-294/-16.13%), Manufacturing (-292/-14.10%) and Professional, Scientific and Technical Services (-197/-12.90%)** . Though if we look at **sheer number of closures then Retail Trade (-229/-9.94%)** comes in third place instead**.**

## Size of Business

Based on **percentage increase or decrease the effect on EmplRange** was as followes : My business sizes that survived the **best were 1000+ (+1/+20%), 300-499 (+1/+1.33%) and 500-999 (0/0%)**. My business sizes that fared the **worst were 1-4 (-917/-12.02%), 10-19 (-232/-10.01%) and 5-9 (-601/-9.54%)** . Though if we look at **sheer number of closures then 10-19 (-232/-10.01%)** comes in third place instead**.**

## Age of Business

All businesses by age suffered losses. Based on **number of closed by Age by 2021** the results were as followes : My business Age, in years, that survived from **worst to best, in order, were 4 years old (-1681), 2 years old (-378), 1 years old (-347) and 3 years old (-308)**. Business that were the oldest seemed more likely to be closed by 2021 than younger businesses that faired roughly around the same. This would make sense as the vast number of businesses in the directory 4 years old and have been there since the inception of the online Directory in 2016.

## Shortcomings of the work

There was not enough attributes to accurately predict and describe why businesses closed. If I were to do this over again I would look to add more demographic data like average household income of people living in each area, unemployment rate, proximity to shopping areas and others to improve the prediction models.

## Contribution of the work compared to past research

I believe my research was useful as it uncovered trends at the city level in a major city in Ontario, Canada. Other studies looked a larger cross country effects and did not address regional or city specific experiences.

Like Garthi and Mathur, I found that that all models were better at predicting open businesses than closed ones by having higher precision, recall and *F*1 scores for Open Businesses. Unlike Dixon I found that the smallest businesses did not contribute to employment growth during the pandemic and were, instead, the most likely to close permanently. As for Bartik et al., the most adversely affected businesses were in the *Arts and Entertainment*, *Tourism and Lodging* and *All Retailers except Grocery* which did not match my results except perhaps for 3rd place. My results concurred with Bartik et al., regarding businesses with fewer employees were most adversely affected. Similarly my results supported Zhang and Stevens where businesses that are older and more established survive and contributed to overall employment numbers, though not the most evidently in *Manufacturing*, *Wholesale Trade* and *Education Services* as they found in the pre-pandemic.

1. Industries | Unlimited Mississauga, Accessed January 22, 2023. <https://www.thefutureisunlimited.ca/industries/> [↑](#footnote-ref-1)
2. 2021 Mississauga Business Directory | Open Data Catalogue, Published January 19, 2022. Data updated January 28, 2022. <https://data.mississauga.ca/datasets/mississauga::2021-mississauga-business-directory/explore?location=43.609064%2C-79.675702%2C12.00> [↑](#footnote-ref-2)
3. City of Mississauga Open Data Catalogue, Accessed January 22, 2023. <https://data.mississauga.ca> [↑](#footnote-ref-3)
4. Garthi, G., Mathur, R. (2023) “[Using Machine Learning to Predict Business Survival in COVID-19](https://github.com/mcnenlyj/MississaugaBusinesses/blob/main/Bibliography_Resources/Using%20Machine%20Learning%20to%20Predict%20Business%20Survival%20in%20COVID-19.pdf)”, pgs 249-261 In: S. Bhattacharyya et al. (eds.), *Human-Centric Smart Computing*, Smart Innovation, Systems and Technologies 316, https://doi.org/10.1007/978-981-19-5403-0\_21 [↑](#footnote-ref-4)
5. Bartik, A. W., Bertrand, M., Cullen, Z. B., Glaeser, E. L., Luca, M., Stanton, C. T. (2020). [How are small businesses adjusting to covid-19? Early evidence from a survey (no. w26989)](https://www.nber.org/system/files/working_papers/w26989/w26989.pdf). National Bureau of Economic Research. [↑](#footnote-ref-5)
6. Zhang, T., Stevens, D. (2009). [Business Age, another Source of Employment Growth](https://nces.ed.gov/FCSM/pdf/2009FCSM_Zhang_XI-B.pdf). Merrick School of Business, University of Baltimore [↑](#footnote-ref-6)
7. Hyatt, H.R., (2022). [Firm Age and Job Creation in the US](https://wol.iza.org/uploads/articles/629/pdfs/firm-age-and-job-creation.pdf). US Census Bureau, USA, and IZA, Germany [↑](#footnote-ref-7)
8. * Dixon, D., Rollin, A.M. (2012) [Firm Dynamics: Employment Growth Rates of Small Versus Large Firms in Canada](https://www150.statcan.gc.ca/n1/en/pub/11-622-m/11-622-m2012025-eng.pdf?st=Hb3aAmdr). Catalogue no. 11-622-M, no. 025. The Canada in Transition Series. Economic Analysis Division, Statistics Canada.

   [↑](#footnote-ref-8)
9. [Staffing actions taken by businesses during the COVID-19 pandemic, by business characteristics](https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3310023101) (2020) Statistics Canada. Table: 33-10-0231-01. Release date: 2020-04-29. [↑](#footnote-ref-9)
10. [Impact of COVID-19 on business or organization status, by business characteristics, third quarter of 2020](https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=3310027601). Statistics Canada, Canadian Survey on Business Conditions, August 2020. [↑](#footnote-ref-10)
11. Dua, A., Mahajan, D., Velasco, Y.[Covid-19’s effect on jobs at small businesses in the United States](https://www.mckinsey.com/industries/public-and-social-sector/our-insights/covid-19s-effect-on-jobs-at-small-businesses-in-the-united-states). May 5, 2020. McKinsey & Company. [↑](#footnote-ref-11)
12. Dua, A., Ellingrud, K., Mahajan, D., Silberg, J. [Which small businesses are most vulnerable to Covid-19 – and when](https://www.mckinsey.com/featured-insights/americas/which-small-businesses-are-most-vulnerable-to-covid-19-and-when). June 18, 2020. McKinsey & Company. [↑](#footnote-ref-12)
13. Wang, B., Wan, B. [The spatial distribution of businesses and neighborhoods: What industries match or mismatch what neighborhoods?](https://www.sciencedirect.com/science/article/pii/S0197397521001296) Habitat International. 117 (2021) 102440. [↑](#footnote-ref-13)