2021CleanDataset

March 19, 2023

[34]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import mutual_info_regression,␣
 ↪mutual_info_classif
```

[164]:
```
#read dataset
data = pd.read_csv("/content/drive/MyDrive/CIND 820 Capstone Project/
 ↪merged_completedata.csv")
```

[165]:
```
# filter dataframe
data = data[data['Year'] >= 2019]
```

[166]:
```
data.head()
```

[166]:

	RecordID	X	Y	FID	BusinessID	Name	Address	StreetNo	StreetName	BldgNo	…	Fax	TollFree	EMail	WebAddress	EmplRange
46689	46690	-79.665386	43.684736	1	7	Peel Car & Truck Rentals	7050 Bramalea Rd	7050	Bramalea Rd	Yes	…	Yes	Yes	Yes	Yes	1
46690	46691	-79.642760	43.593515	2	4246	Real Fruit Bubble Tea	100 City Centre Dr	100	City Centre Dr	No	…	No	No	No	Yes	2
46691	46692	-79.667311	43.682752	3	10	Unifor 2002	7015 Tranmere Dr	7015	Tranmere Dr	No	…	Yes	Yes	Yes	Yes	3
46692	46693	-79.629235	43.698932	4	4247	Laura with Plus and Petites	100 City Centre Dr	100	City Centre Dr	No	…	No	No	No	Yes	2
46693	46694	-79.629235	43.698932	5	4250	Footlocker	100 City Centre Dr	100								

```
46693  City Centre Dr     No  …   No       No    No          No          4

            CENT_X        CENT_Y  Year isnew Closed
46689  607567.2334  4.837723e+06  2019    No     No
46690  609556.5032  4.827621e+06  2019   Yes     No
46691  607415.6044  4.837500e+06  2019    No     No
46692  610454.8654  4.839347e+06  2019   Yes     No
46693  610454.8654  4.839347e+06  2019   Yes     No

[5 rows x 28 columns]
```

[167]:
```
data['Closed'].value_counts()
```

[167]:
```
No     28629
Yes     2714
Name: Closed, dtype: int64
```

[168]:
```
data.shape
```

[168]:
```
(31343, 28)
```

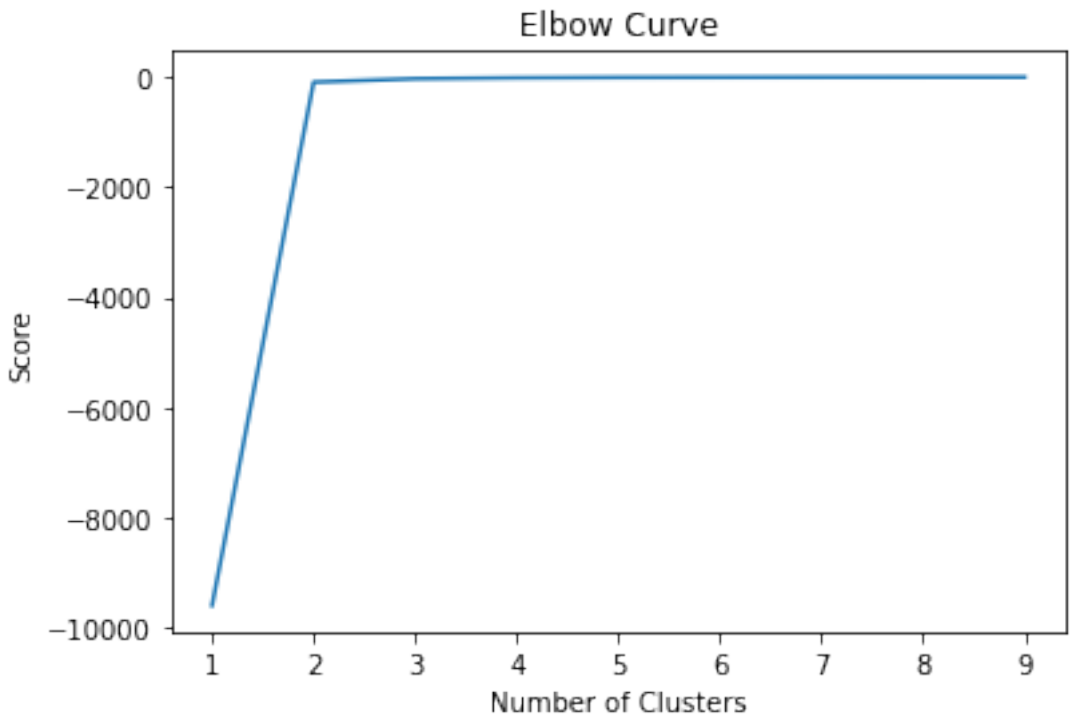
[169]:
```
ClosedBy2021 = data['Closed'].value_counts()[1]/data.shape[0]
print("Closed accuracy : ", ClosedBy2021 )
ClosedPercent = ClosedBy2021*100
print("Percent of businesses closed : ", ClosedPercent)
```

```
Closed accuracy :  0.08659030724563699
Percent of businesses closed :  8.6590307245637
```

[170]:
```
#clustering of locations.  All in Mississauga so only 2 clusters

from sklearn.cluster import KMeans

K_clusters = range(1,10)
kmeans = [KMeans(n_clusters=i) for i in K_clusters]
Y_axis = data[['Y']]
X_axis = data[['X']]
score = [kmeans[i].fit(Y_axis).score(Y_axis) for i in range(len(kmeans))]
# Visualize
plt.plot(K_clusters, score)
plt.xlabel('Number of Clusters')
plt.ylabel('Score')
plt.title('Elbow Curve')
plt.show()
```

```
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
```

```
  warnings.warn(
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
```



[171]:
```
kmeans = KMeans(n_clusters = 2, init ='k-means++')
kmeans.fit(data[data.columns[1:3]]) # Compute k-means clustering.
data['cluster_label'] = kmeans.fit_predict(data[data.columns[1:3]])
centers = kmeans.cluster_centers_ # Coordinates of cluster centers.
labels = kmeans.predict(data[data.columns[1:3]]) # Labels of each point
data.head(5)
```

```
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
/usr/local/lib/python3.9/dist-packages/sklearn/cluster/_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
```

[171]:

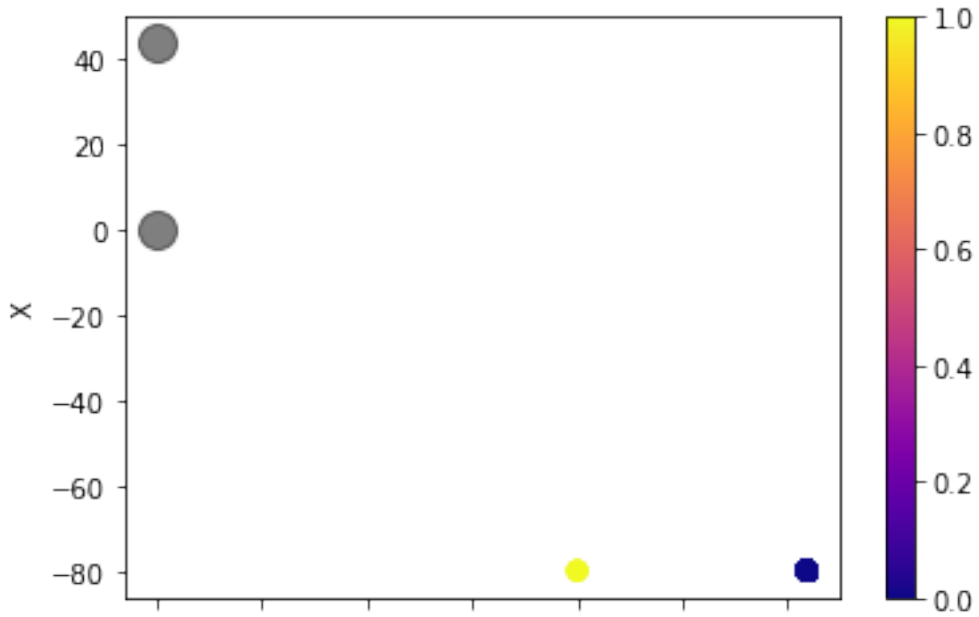
	RecordID	X	Y	FID	BusinessID	\
46689	46690	-79.665386	43.684736	1	7	
46690	46691	-79.642760	43.593515	2	4246	
46691	46692	-79.667311	43.682752	3	10	
46692	46693	-79.629235	43.698932	4	4247	
46693	46694	-79.629235	43.698932	5	4250	

	Name	Address	StreetNo	\

```
46689     Peel Car & Truck Rentals    7050 Bramalea Rd      7050
46690        Real Fruit Bubble Tea  100 City Centre Dr       100
46691                  Unifor 2002    7015 Tranmere Dr      7015
46692  Laura with Plus and Petites  100 City Centre Dr       100
46693                   Footlocker  100 City Centre Dr       100

           StreetName BldgNo  … TollFree EMail WebAddress  EmplRange  \
46689     Bramalea Rd    Yes  …      Yes   Yes        Yes          1
46690  City Centre Dr     No  …       No    No        Yes          2
46691     Tranmere Dr     No  …      Yes   Yes        Yes          3
46692  City Centre Dr     No  …       No    No        Yes          2
46693  City Centre Dr     No  …       No    No         No          4

            CENT_X        CENT_Y  Year isnew Closed cluster_label
46689  607567.2334  4.837723e+06  2019    No     No             0
46690  609556.5032  4.827621e+06  2019   Yes     No             0
46691  607415.6044  4.837500e+06  2019    No     No             0
46692  610454.8654  4.839347e+06  2019   Yes     No             0
46693  610454.8654  4.839347e+06  2019   Yes     No             0

[5 rows x 29 columns]
```

[172]:
```
data.plot.scatter(x = 'Y', y = 'X', c=labels, s=50, cmap='plasma')
plt.scatter(centers[:, 0], centers[:, 1], c='black', s=200, alpha=0.5)
```

[172]: <matplotlib.collections.PathCollection at 0x7fb533fa1a30>



[173]:
```
df2 = pd.DataFrame().assign(Year=data['Year'], Size=data['EmplRange'],␣
↪Industry=data['NAICSCat'])
print(df2)
```

```
       Year  Size                                                Industry
46689  2019     1                                          Retail Trade
46690  2019     2                        Accommodation and Food Services
46691  2019     3                                      Other Services
46692  2019     2                                          Retail Trade
46693  2019     4                                          Retail Trade
…       …     …                                                     …
78027  2021     3  Administrative and Support, Waste Management a…
78028  2021     1  Administrative and Support, Waste Management a…
78029  2021     1                        Accommodation and Food Services
78030  2021     1                                     Wholesale Trade
78031  2021     1                                     Wholesale Trade

[31343 rows x 3 columns]
```

[174]:
```
dfIndustryCount = df2.groupby(['Year','Industry'])['Year'].count()
dfIndustryCount
```

[174]:
```
Year  Industry
2019  Accommodation and Food Services
1321
      Administrative and Support, Waste Management and Remediation Services
562
      Arts, Entertainment and Recreation
228
      Construction
621
      Educational Services
647
      Finance and Insurance
638
      Health Care and Social Assistance
1281
      Information and Cultural Industries
137
      Management of Companies and Enterprises
107
      Manufacturing
2071
      Other Services
1873
      Primary Industry
5
```

Professional, Scientific and Technical Services
1527
Public Administration
107
Real Estate and Rental and Leasing
415
Retail Trade
2303
Transportation and Warehousing
838
Utilities
14
Wholesale Trade
1823
2021 Accommodation and Food Services
1230
Administrative and Support, Waste Management and Remediation Services
494
Arts, Entertainment and Recreation
202
Construction
548
Educational Services
587
Finance and Insurance
604
Health Care and Social Assistance
1287
Information and Cultural Industries
136
Management of Companies and Enterprises
98
Manufacturing
1779
Other Services
1703
Primary Industry
6
Professional, Scientific and Technical Services
1330
Public Administration
104
Real Estate and Rental and Leasing
370
Retail Trade
2074
Transportation and Warehousing

```
728
      Utilities
16
      Wholesale Trade
1529
Name: Year, dtype: int64
```

[175]:
```
dfIndustryCount = df2.groupby(['Industry','Year'])['Industry'].count()
dfIndustryCount
```

[175]:
```
Industry                                                        Year
Accommodation and Food Services                                 2019
1321
                                                                2021
1230
Administrative and Support, Waste Management and Remediation Services  2019
562
                                                                2021
494
Arts, Entertainment and Recreation                              2019
228
                                                                2021
202
Construction                                                    2019
621
                                                                2021
548
Educational Services                                            2019
647
                                                                2021
587
Finance and Insurance                                           2019
638
                                                                2021
604
Health Care and Social Assistance                               2019
1281
                                                                2021
1287
Information and Cultural Industries                             2019
137
                                                                2021
136
Management of Companies and Enterprises                         2019
107
                                                                2021
98
```

```
Manufacturing                                                  2019
2071
                                                               2021
1779
Other Services                                                 2019
1873
                                                               2021
1703
Primary Industry                                               2019
5
                                                               2021
6
Professional, Scientific and Technical Services                2019
1527
                                                               2021
1330
Public Administration                                          2019
107
                                                               2021
104
Real Estate and Rental and Leasing                             2019
415
                                                               2021
370
Retail Trade                                                   2019
2303
                                                               2021
2074
Transportation and Warehousing                                 2019
838
                                                               2021
728
Utilities                                                      2019
14
                                                               2021
16
Wholesale Trade                                                2019
1823
                                                               2021
1529
Name: Industry, dtype: int64
```

[176]:
```
# Using DataFrame.agg() Method.
df3 = df2.groupby(['Industry', 'Year']).agg({'Year': 'count'})
print(df3)
```

```
                                                    Year
```

```
Industry                                                Year
Accommodation and Food Services                         2019  1321
                                                        2021  1230
Administrative and Support, Waste Management an… 2019   562
                                                        2021   494
Arts, Entertainment and Recreation                      2019   228
                                                        2021   202
Construction                                            2019   621
                                                        2021   548
Educational Services                                    2019   647
                                                        2021   587
Finance and Insurance                                   2019   638
                                                        2021   604
Health Care and Social Assistance                       2019  1281
                                                        2021  1287
Information and Cultural Industries                     2019   137
                                                        2021   136
Management of Companies and Enterprises                 2019   107
                                                        2021    98
Manufacturing                                           2019  2071
                                                        2021  1779
Other Services                                          2019  1873
                                                        2021  1703
Primary Industry                                        2019     5
                                                        2021     6
Professional, Scientific and Technical Services         2019  1527
                                                        2021  1330
Public Administration                                   2019   107
                                                        2021   104
Real Estate and Rental and Leasing                      2019   415
                                                        2021   370
Retail Trade                                            2019  2303
                                                        2021  2074
Transportation and Warehousing                          2019   838
                                                        2021   728
Utilities                                               2019    14
                                                        2021    16
Wholesale Trade                                         2019  1823
                                                        2021  1529
```

[177]:
```
# Percentage by pct_change method on groupby.
df4 = df3.groupby(level=0).pct_change()*100
print(df4)
```

```
                                                        Year
Industry                                                Year
Accommodation and Food Services                         2019       NaN
                                                        2021 -6.888721
```

```
Administrative and Support, Waste Management an… 2019        NaN
                                                  2021 -12.099644
Arts, Entertainment and Recreation                2019        NaN
                                                  2021 -11.403509
Construction                                      2019        NaN
                                                  2021 -11.755233
Educational Services                              2019        NaN
                                                  2021  -9.273570
Finance and Insurance                             2019        NaN
                                                  2021  -5.329154
Health Care and Social Assistance                 2019        NaN
                                                  2021   0.468384
Information and Cultural Industries               2019        NaN
                                                  2021  -0.729927
Management of Companies and Enterprises           2019        NaN
                                                  2021  -8.411215
Manufacturing                                     2019        NaN
                                                  2021 -14.099469
Other Services                                    2019        NaN
                                                  2021  -9.076348
Primary Industry                                  2019        NaN
                                                  2021  20.000000
Professional, Scientific and Technical Services   2019        NaN
                                                  2021 -12.901113
Public Administration                             2019        NaN
                                                  2021  -2.803738
Real Estate and Rental and Leasing                2019        NaN
                                                  2021 -10.843373
Retail Trade                                      2019        NaN
                                                  2021  -9.943552
Transportation and Warehousing                    2019        NaN
                                                  2021 -13.126492
Utilities                                         2019        NaN
                                                  2021  14.285714
Wholesale Trade                                   2019        NaN
                                                  2021 -16.127263
```

[178]:
```
dfSizeCount = df2.groupby(['Year','Size'])['Year'].count()
dfSizeCount
```

[178]:
```
Year  Size
2019  1       7629
      2       3470
      3       2316
      4       1767
      5        729
      6        478
```

```
      7         75
      8         34
      9         20
2021  1       6712
      2       3139
      3       2084
      4       1601
      5        714
      6        441
      7         76
      8         34
      9         24
Name: Year, dtype: int64
```

[179]:
```
dfSizeCount = df2.groupby(['Size','Year'])['Size'].count()
dfSizeCount
```

[179]:
```
Size  Year
1     2019    7629
      2021    6712
2     2019    3470
      2021    3139
3     2019    2316
      2021    2084
4     2019    1767
      2021    1601
5     2019     729
      2021     714
6     2019     478
      2021     441
7     2019      75
      2021      76
8     2019      34
      2021      34
9     2019      20
      2021      24
Name: Size, dtype: int64
```

[182]:
```
# Using DataFrame.agg() Method.
df5 = df2.groupby(['Size', 'Year']).agg({'Year': 'count'})
print(df5)
```

```
           Year
Size Year
1    2019  7629
     2021  6712
2    2019  3470
```

```
     2021  3139
3    2019  2316
     2021  2084
4    2019  1767
     2021  1601
5    2019   729
     2021   714
6    2019   478
     2021   441
7    2019    75
     2021    76
8    2019    34
     2021    34
9    2019    20
     2021    24
```

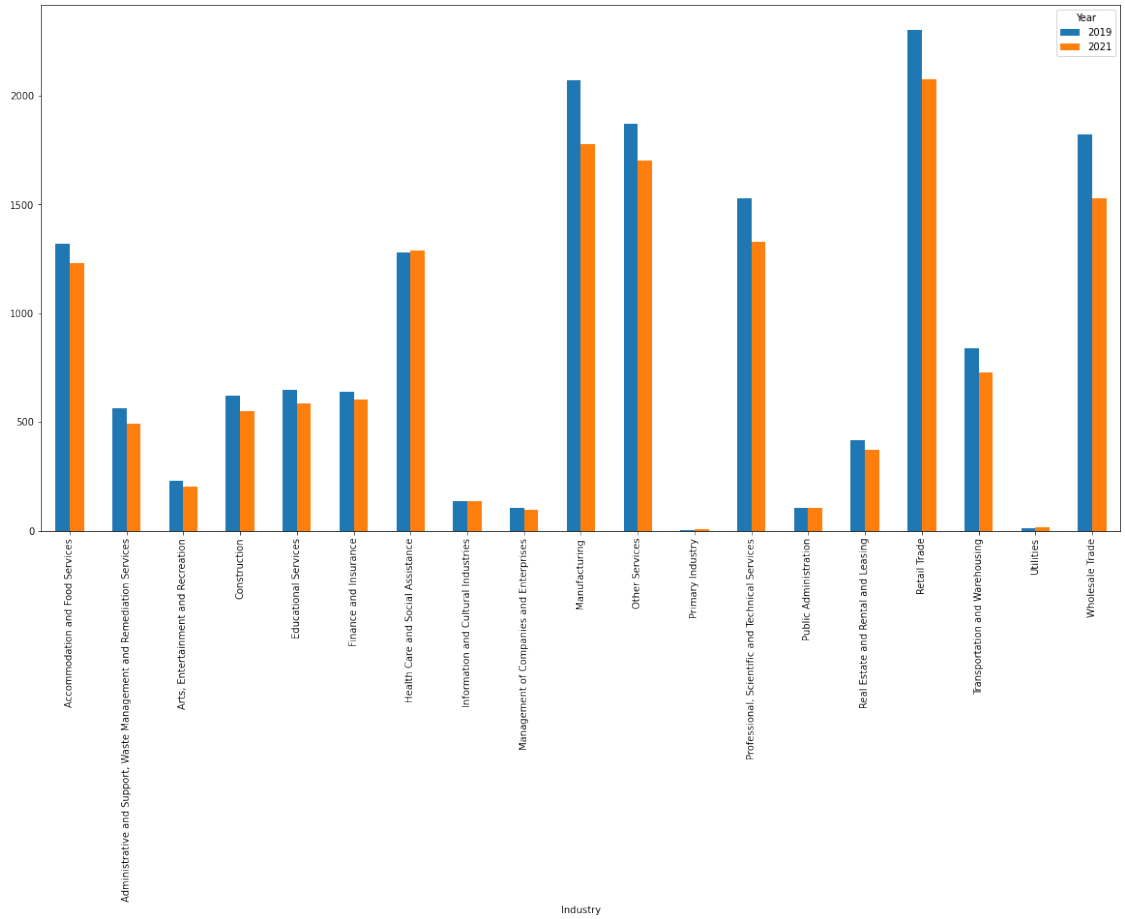
[181]:
```
# Percentage by pct_change method on groupby.
df6 = df5.groupby(level=0).pct_change()*100
print(df6)
```

```
                Year
Size Year
1    2019        NaN
     2021 -12.019924
2    2019        NaN
     2021  -9.538905
3    2019        NaN
     2021 -10.017271
4    2019        NaN
     2021  -9.394454
5    2019        NaN
     2021  -2.057613
6    2019        NaN
     2021  -7.740586
7    2019        NaN
     2021   1.333333
8    2019        NaN
     2021   0.000000
9    2019        NaN
     2021  20.000000
```

[183]:
```
(df2.groupby(['Year','Industry'])['Year']
    .count().unstack('Year').plot.bar(figsize=(20, 10)))
#Net loss of businesses by Industry between 2019 and 2021
#Industries where most businesses closed were : Wholesale Trade ; Manufacturing␣
 ↪; Retail Trade
```

```
#Some of these industries fall within the industries other studies pointed to␣
 ↪as experiencing and existential threat early in the pandemic and vice versa␣
 ↪least negatively impacted
#example: Retail Trade vs Public Administration
#Industries where least businesss closed were :  Information and Cultural␣
 ↪Industries ; Public Administration
#Industries Health Care and Social Assistance ;  Utlities - Were the only␣
 ↪industries to increase business count
#Some of these fall within the strategic industries Mississauga has identified␣
 ↪for future growth
#So to summarize, there is both agreement and disagreement from the other␣
 ↪studies. Keeping in mind some industries are not in cities eg. Mining or␣
 ↪Fishing.
```

[183]: <Axes: xlabel='Industry'>



[185]:
```
(df2.groupby(['Year','Size'])['Year']
    .count().unstack('Year').plot.bar(figsize=(20, 10)))
```

```
#Net loss of businesses by Size of business between 2019 and 2021
#The smallest businesses closed the most between 2019 and 2021 - '1 to 4': 1,␣
 ↪'5 to 9': 2, '10 to 19': 3
#The largest businesses stayed even ['500 to 999': 8]  or even grew ['300 to␣
 ↪499': 7, '1000+': 9 ]
#The larger the business the more stable
#This is different from Stats Can ontario survey were 20-99, 5-19 adn 100-249␣
 ↪were hardest hit and 0, 1-4 and 250-499 were least affected
#I need to factor in the age of the business. Were businesses that were older␣
 ↪less likely to close?
```

[185]: <Axes: xlabel='Size'>