The economic impact of COVID across people in different demographic groups and education levels.

Team Members:

Chris Capps

Timothy Keating

Vedika Nigam

Abazar Rahma

Introduction

- COVID has had a huge impact on the economy and our lives. The impact of COVID has not been uniform across different groups.
- By conducting this study, we hope to examine how COVID has affected individuals in terms of their unemployment situation.
- Findings from this study could help policy makers in creating appropriate support structures for affected individuals.
- For this study, we have narrowed our focus to North Carolina.

Data Source

- Integrated Public Use Microdata Series (IPUMS) which is the world's largest individual-level population database.
- IPUMS has compiled this data from American Community Survey (ACS) which is a demographics survey program conducted by the U.S. Census Bureau.

Questions we hope to answer

What are the demographic and educational attainment factors that predict who is unable to work in North Carolina during COVID?

Data Exploration

Cleaning the data involved the following steps:

- Filtering the data for North Carolina by using STATE FIP

```
]:  # Filtering data for North Carolina
    demographic_data_df_NC = demographic_data_df[demographic_data_df['STATEFIP'] == 37]
    demographic_data_df_NC
```

]:

	YEAR	MONTH	STATEFIP	METAREA	COUNTY	AGE	SEX	RACE	MARST	HISPAN	EDUC	COVIDTELEW	COVIDUNAW
63258	2020	5	37	3122	0	54	1	100	6	0	111	1	1
63259	2020	5	37	3121	37067	51	2	100	4	0	91	1	1
63260	2020	5	37	3121	37067	49	1	100	6	0	111	1	1
63261	2020	5	37	1521	37119	65	2	100	1	0	73	99	1
63262	2020	5	37	1521	37119	61	1	100	1	0	73	2	1

- Removing invalid inputs (values that are 99) for target variable

```
:  # Filtering target columns to keep valid data and drop 99 values
   demographic_data_df_NC = demographic_data_df_NC[demographic_data_df_NC['COVIDUNAW'] != 99]
   demographic_data_df_NC.head(10)
```

:

	YEAR	MONTH	STATEFIP	METAREA	COUNTY	AGE	SEX	RACE	MARST	HISPAN	EDUC	COVIDTELEW	COVIDUNAW
63258	2020	5	37	3122	0	54	1	100	6	0	111	1	1
63259	2020	5	37	3121	37067	51	2	100	4	0	91	1	1
63260	2020	5	37	3121	37067	49	1	100	6	0	111	1	1

Data Exploration

- Explore data using value_counts(). Binning the independent variables using map function

```
education={111:"Bachelor's",
            73:"High School or below",
            81:"Some College or Associate Degree",
           123:"Graduate or Professional Degree",
           92:"Some College or Associate Degree",
           91:"Some College or Associate Degree",
           125:"Graduate or Professional Degree",
           60:"High School or below",
           50:"High School or below",
           124:"Graduate or Professional Degree",
           71:"High School or below",
}
```

```
#Applying map function to change categorical data from numbers to labels
demographic_data_df_NC["education"] = demographic_data_df_NC['EDUC'].map(education)
demographic_data_df_NC.head()
```

	YEAR	MONTH	METAREA	COUNTY	AGE	SEX	RACE	MARST	HISPAN	EDUC	COVIDTELEW	COVIDUNAW	gender	education
63258	2020	5	3122	0	54	1	100	6	0	111	1	1	Male	Bachelor's
63259	2020	5	3121	37067	51	2	100	4	0	91	1	1	Female	Some College or Associate Degree
63260	2020	5	3121	37067	49	1	100	6	0	111	1	1	Male	Bachelor's
63262	2020	5	1521	37119	61	1	100	1	0	73	2	1	Male	High School or below
63268	2020	5	3122	0	35	2	100	1	0	111	2	1	Female	Bachelor's

Data Exploration

- Combining year and month column to create date column

```
In [5]:
# COMPINE YEAR AND MONTH columns IN ONE column
df['DATE'] = pd.to_datetime(df[['YEAR', 'MONTH']].assign(DAY=1))
df
```

Out[5]:

ONTH	METAREA	COUNTY	AGE	SEX	RACE	MARST	HISPAN	EDUC	COVIDTELEW	COVIDUNAW	gender	education	race	hispanic	marital_status	DATE
5	3122	0	54	1	100	6	0	111	1	1	Male	Bachelor's	White	Non-Hispanic	Single	2020-05-01

- Binning Age variable using pd.cut

```
:
1
2 demographic_data_df_NC['age'] = pd.cut(x=demographic_data_df_NC['AGE'], bins=[16, 24, 34, 44, 54, 64,90],
3                     labels=['16 to 24', '25 to 34', '35 to 44',
4                             '45 to 54', '55 to 64','65+'])
```

```
:
1 demographic_data_df_NC
```

:

MONTH	METAREA	COUNTY	AGE	SEX	RACE	MARST	HISPAN	EDUC	COVIDTELEW	COVIDUNAW	gender	education	race	hispanic	marital_status	age
5	3122	0	54	1	100	6	0	111	1	1	Male	Bachelor's	White	Non-Hispanic	Single	45 to 54

Data Exploration

- Removing all Nan values and exporting clean file to csv

```
:
1 #Dropping all Nan values
2 demographic_data_df_NC.dropna(how='any',inplace = True)
3 demographic_data_df_NC
```

:

	YEAR	MONTH	METAREA	COUNTY	AGE	SEX	RACE	MARST	HISPAN	EDUC	COVIDTELEW	COVIDUNAW	gender	education	race	hispanic
63258	2020	5	3122	0	54	1	100	6	0	111	1	1	Male	Bachelor's	White	Non-Hispanic

- Processing data for machine learning - Encoding categorical variables

In [11]:
```
1 # Create our features
2 X = pd.get_dummies(df_cleanen, columns=["gender","education","race","hispanic","marital_status"])
3 # Create our target
4 y = df_cleanen['COVIDUNAW']
5 X
```

Out[11]:

	METAREA	COUNTY	AGE	COVIDUNAW	gender_Female	gender_Male	education_Bachelor's	education_Graduate or Professional Degree	education_High School or below	education_Some College or Associate Degree	race
0	3122	0	54	1	0	1	1	0	0	0	
1	3121	37067	51	1	1	0	0	0	0	1	

Data Exploration

Total Data Points: 19205 individuals who answered the survey

Independent variables (features):

- Age - 16-24, 25-34, 35-44, 45-54, 55-64, 65+
- Gender - Male, Female
- Race - Black, White, Native American, Asian
- Marital Status - Married, Single, Divorced
- Education- High school or below, Some College, Bachelor's, Graduate or Professional Degree

Dependent variable (target):

- COVIDUNAW - individuals who are unable to work during COVID (1 : able to work, 2: unable to work)

Data Exploration

Statistics summary using describe()

```
1 df_encoded.describe()
```

	AGE	COVIDUNAW	month_le	gender_Female	gender_Male	education_Bachelor's	education_Graduate or Professional Degree	education_High School or below	education_Some College or Associate Degree
count	19205.000000	19205.000000	19205.000000	19205.000000	19205.000000	19205.000000	19205.000000	19205.000000	19205.000000
mean	43.509399	1.045040	15.103306	0.477636	0.522364	0.275397	0.156313	0.308774	0.259516
std	14.770140	0.207398	5.575414	0.499513	0.499513	0.446726	0.363162	0.462000	0.438380
min	16.000000	1.000000	5.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	31.000000	1.000000	10.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	43.000000	1.000000	15.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
75%	55.000000	1.000000	20.000000	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000
max	85.000000	2.000000	24.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

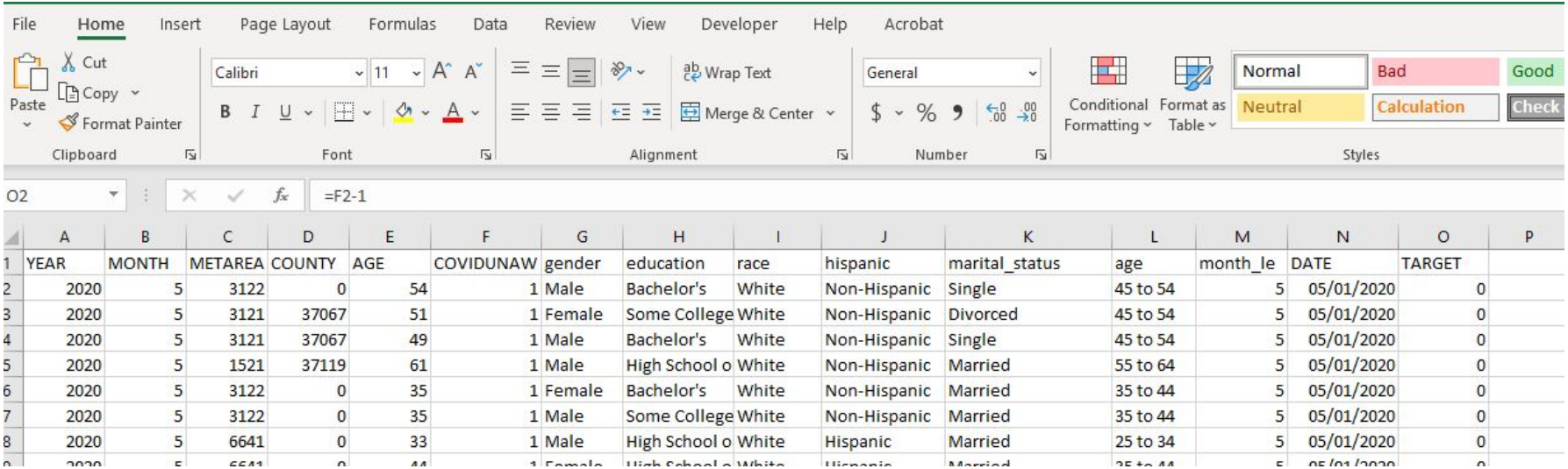
Technologies, languages, tools, and algorithms

- **Data Cleaning and Analysis**: Python and pandas library for data cleaning and exploratory analysis. Pivot Tables in Excel were used for preliminary data analysis. Final data analysis and visualization in Tableau.
- **Database Storage:** PostgreSQL was used to create a database for our project.
- **Machine Learning:** SciKitLearn Machine Learning Library was used to create a classifier. Imbalanced Learn Library and Gradient Boosting.
- **Dashboard:** SQLAlchemy, Flask, Python, and Heroku cloud platform for connecting database to web application. HTML for creating web application.
- **Repository**: Github repository to store all files and information related to the project

Web Application

Data Analysis - Process

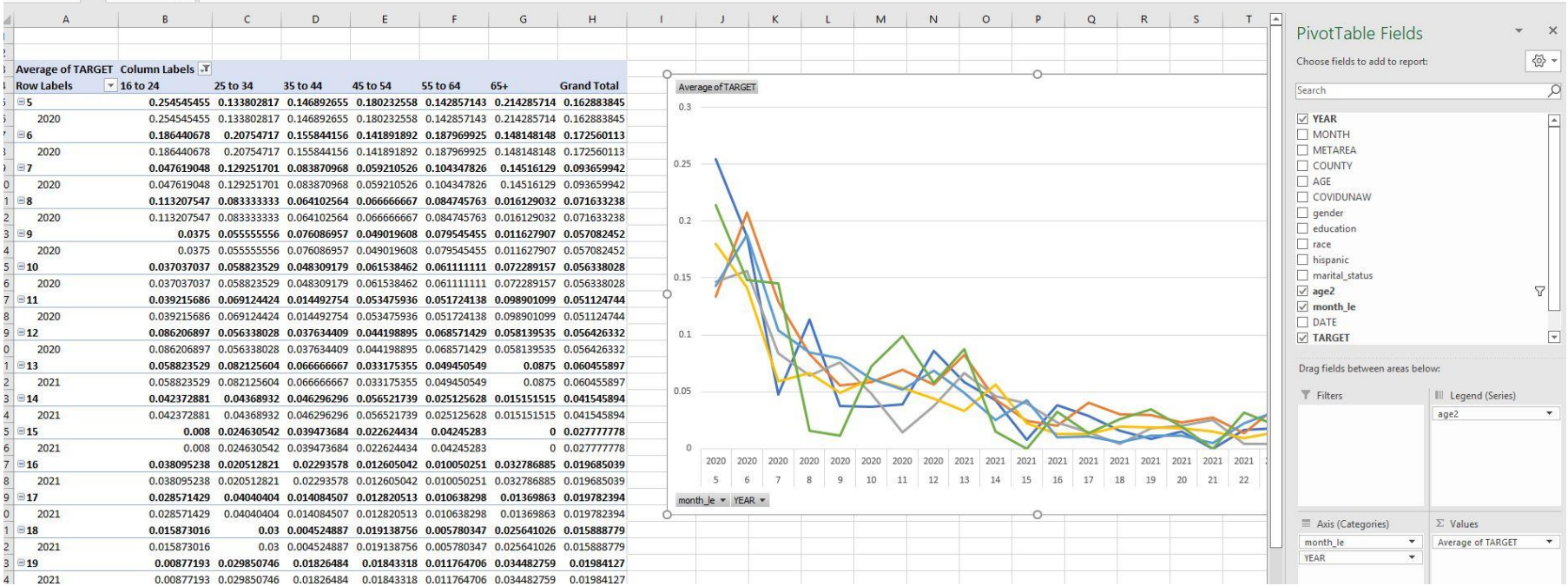
- Created Target Column where 0 represents able to work and 1 represents unable to work to find percentage of individuals not able to work



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	YEAR	MONTH	METAREA	COUNTY	AGE	COVIDUNAW	gender	education	race	hispanic	marital_status	age	month_le	DATE	TARGET	
2	2020	5	3122	0	54	1	Male	Bachelor's	White	Non-Hispanic	Single	45 to 54	5	05/01/2020	0	
3	2020	5	3121	37067	51	1	Female	Some College	White	Non-Hispanic	Divorced	45 to 54	5	05/01/2020	0	
4	2020	5	3121	37067	49	1	Male	Bachelor's	White	Non-Hispanic	Single	45 to 54	5	05/01/2020	0	
5	2020	5	1521	37119	61	1	Male	High School o	White	Non-Hispanic	Married	55 to 64	5	05/01/2020	0	
6	2020	5	3122	0	35	1	Female	Bachelor's	White	Non-Hispanic	Married	35 to 44	5	05/01/2020	0	
7	2020	5	3122	0	35	1	Male	Some College	White	Non-Hispanic	Married	35 to 44	5	05/01/2020	0	
8	2020	5	6641	0	33	1	Male	High School o	White	Hispanic	Married	25 to 34	5	05/01/2020	0	

Data Analysis - Process

- Created Pivot Tables for quick line graphs for categorical variables



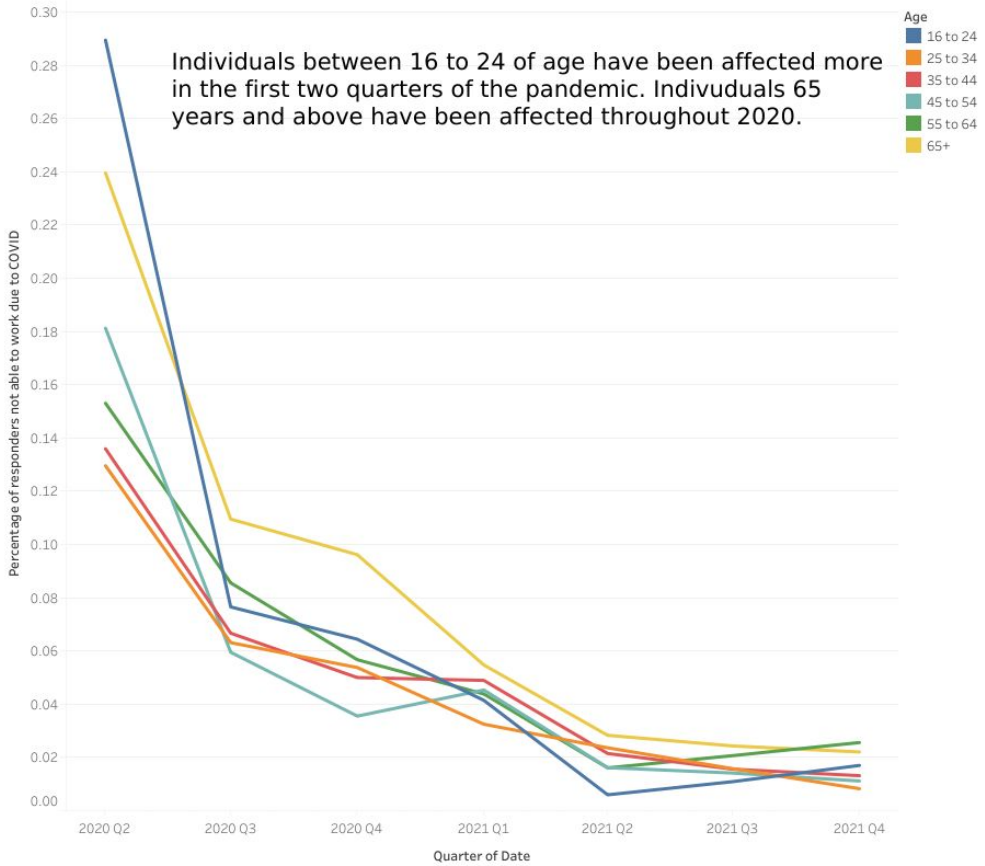
Average of TARGET	Column Labels						
Row Labels	16 to 24	25 to 34	35 to 44	45 to 54	55 to 64	65+	Grand Total
5	**0.254545455**	**0.133802817**	**0.146892655**	**0.180232558**	**0.142857143**	**0.214285714**	**0.162883845**
2020	0.254545455	0.133802817	0.146892655	0.180232558	0.142857143	0.214285714	0.162883845
6	**0.186440678**	**0.20754717**	**0.155844156**	**0.141891892**	**0.187969925**	**0.148148148**	**0.172560113**
2020	0.186440678	0.20754717	0.155844156	0.141891892	0.187969925	0.148148148	0.172560113
7	**0.047619048**	**0.129251701**	**0.083870968**	**0.059210526**	**0.104347826**	**0.14516129**	**0.093659942**
2020	0.047619048	0.129251701	0.083870968	0.059210526	0.104347826	0.14516129	0.093659942
8	**0.113207547**	**0.083333333**	**0.064102564**	**0.066666667**	**0.084745763**	**0.016129032**	**0.071633238**
2020	0.113207547	0.083333333	0.064102564	0.066666667	0.084745763	0.016129032	0.071633238
9	**0.0375**	**0.055555556**	**0.076086957**	**0.049019608**	**0.079545455**	**0.011627907**	**0.057082452**
2020	0.0375	0.055555556	0.076086957	0.049019608	0.079545455	0.011627907	0.057082452
10	**0.037037037**	**0.058823529**	**0.048309179**	**0.061538462**	**0.061111111**	**0.072289157**	**0.056338028**
2020	0.037037037	0.058823529	0.048309179	0.061538462	0.061111111	0.072289157	0.056338028
11	**0.039215686**	**0.069124424**	**0.014492754**	**0.053475936**	**0.051724138**	**0.098901099**	**0.051124744**
2020	0.039215686	0.069124424	0.014492754	0.053475936	0.051724138	0.098901099	0.051124744
12	**0.086206897**	**0.056338028**	**0.037634409**	**0.044198895**	**0.068571429**	**0.058139535**	**0.056426332**
2020	0.086206897	0.056338028	0.037634409	0.044198895	0.068571429	0.058139535	0.056426332
13	**0.058823529**	**0.082125604**	**0.066666667**	**0.033175355**	**0.049450549**	**0.0875**	**0.060455897**
2021	0.058823529	0.082125604	0.066666667	0.033175355	0.049450549	0.0875	0.060455897
14	**0.042372881**	**0.04368932**	**0.046296296**	**0.056521739**	**0.025125628**	**0.015151515**	**0.041545894**
2021	0.042372881	0.04368932	0.046296296	0.056521739	0.025125628	0.015151515	0.041545894
15	**0.008**	**0.024630542**	**0.039473684**	**0.022624434**	**0.04245283**	**0**	**0.027777778**
2021	0.008	0.024630542	0.039473684	0.022624434	0.04245283	0	0.027777778
16	**0.038095238**	**0.020512821**	**0.02293578**	**0.012605042**	**0.010050251**	**0.032786885**	**0.019685039**
2021	0.038095238	0.020512821	0.02293578	0.012605042	0.010050251	0.032786885	0.019685039
17	**0.028571429**	**0.04040404**	**0.014084507**	**0.012820513**	**0.010638298**	**0.01369863**	**0.019782394**
2021	0.028571429	0.04040404	0.014084507	0.012820513	0.010638298	0.01369863	0.019782394
18	**0.015873016**	**0.03**	**0.004524887**	**0.019138756**	**0.005780347**	**0.025641026**	**0.015888779**
2021	0.015873016	0.03	0.004524887	0.019138756	0.005780347	0.025641026	0.015888779
19	**0.00877193**	**0.029850746**	**0.01826484**	**0.01843318**	**0.011764706**	**0.034482759**	**0.01984127**
2021	0.00877193	0.029850746	0.01826484	0.01843318	0.011764706	0.034482759	0.01984127

Age

[Tableau Dashboard]

Demographics_COVID_Unemploment

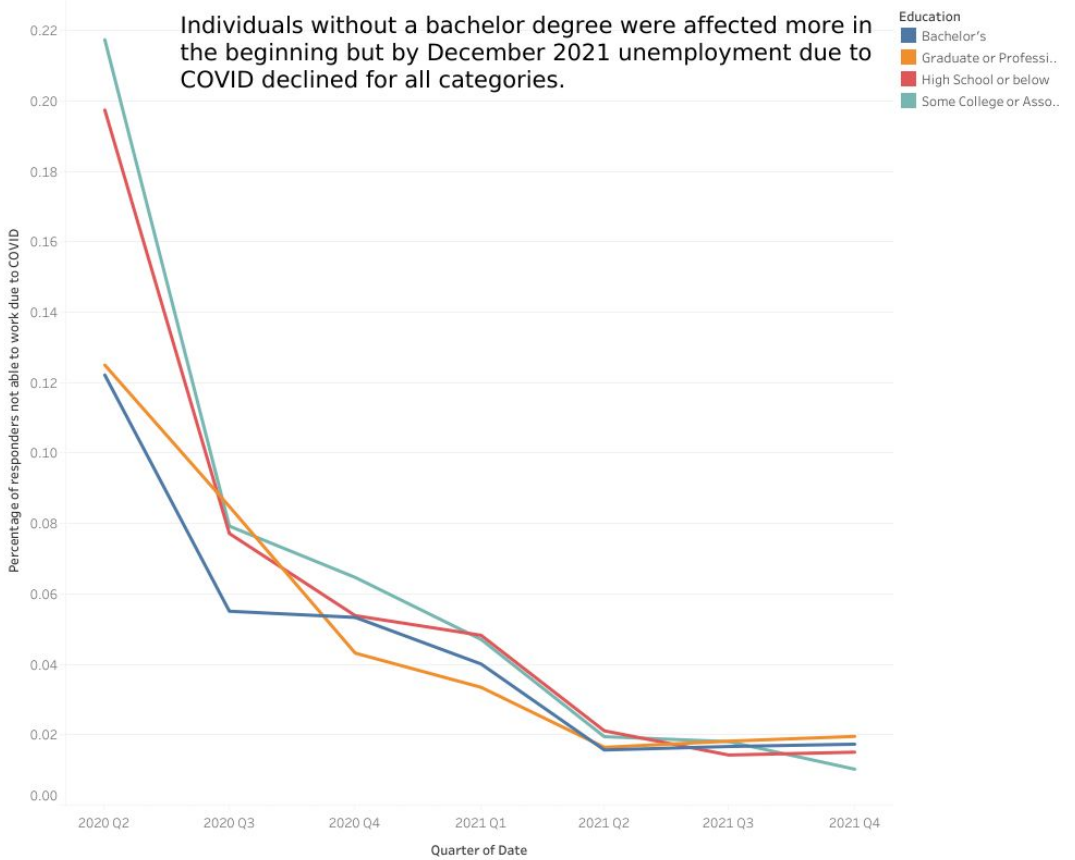
Age | Education | Gender | Race | Hispanic Origin | Marital Status | MSA



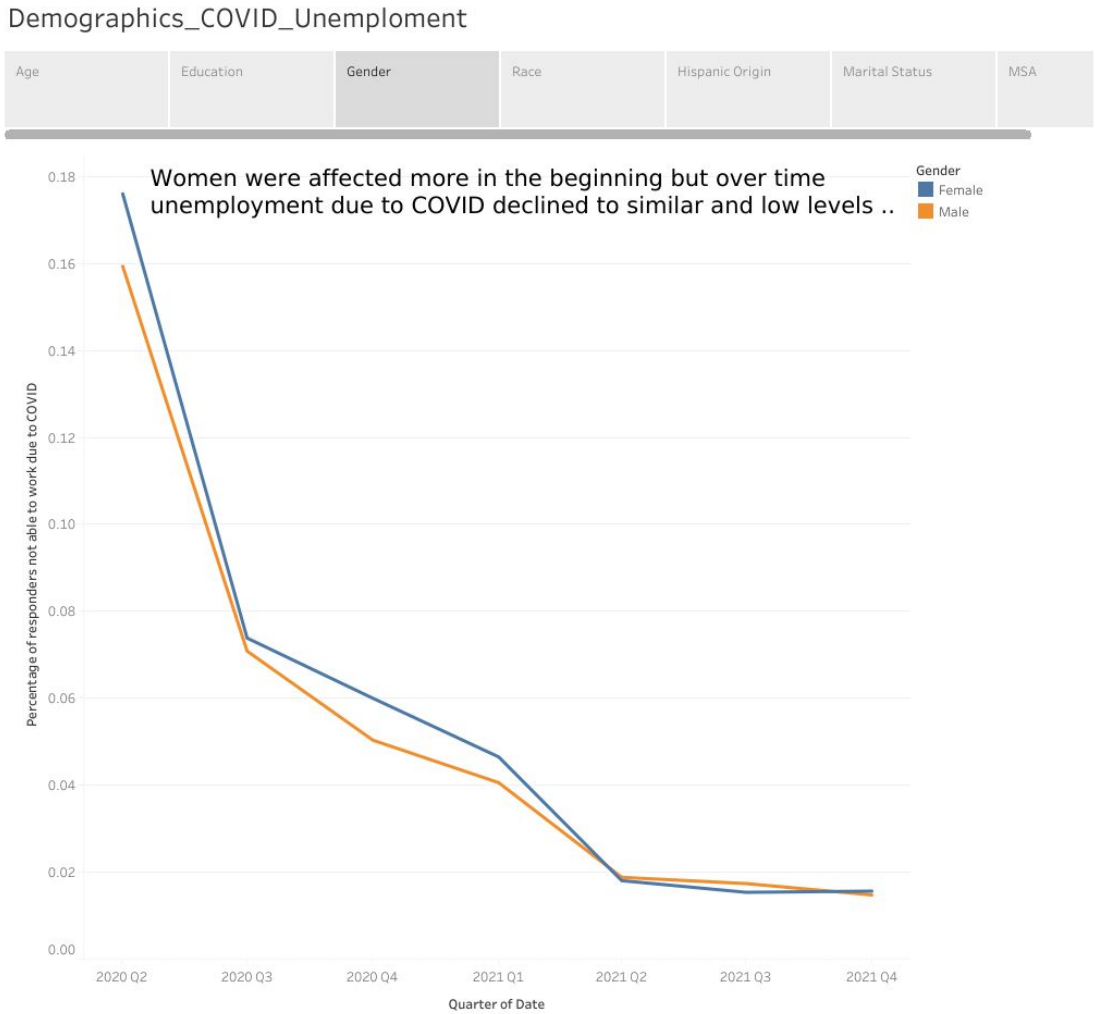
Education

Demographics_COVID_Unemploment

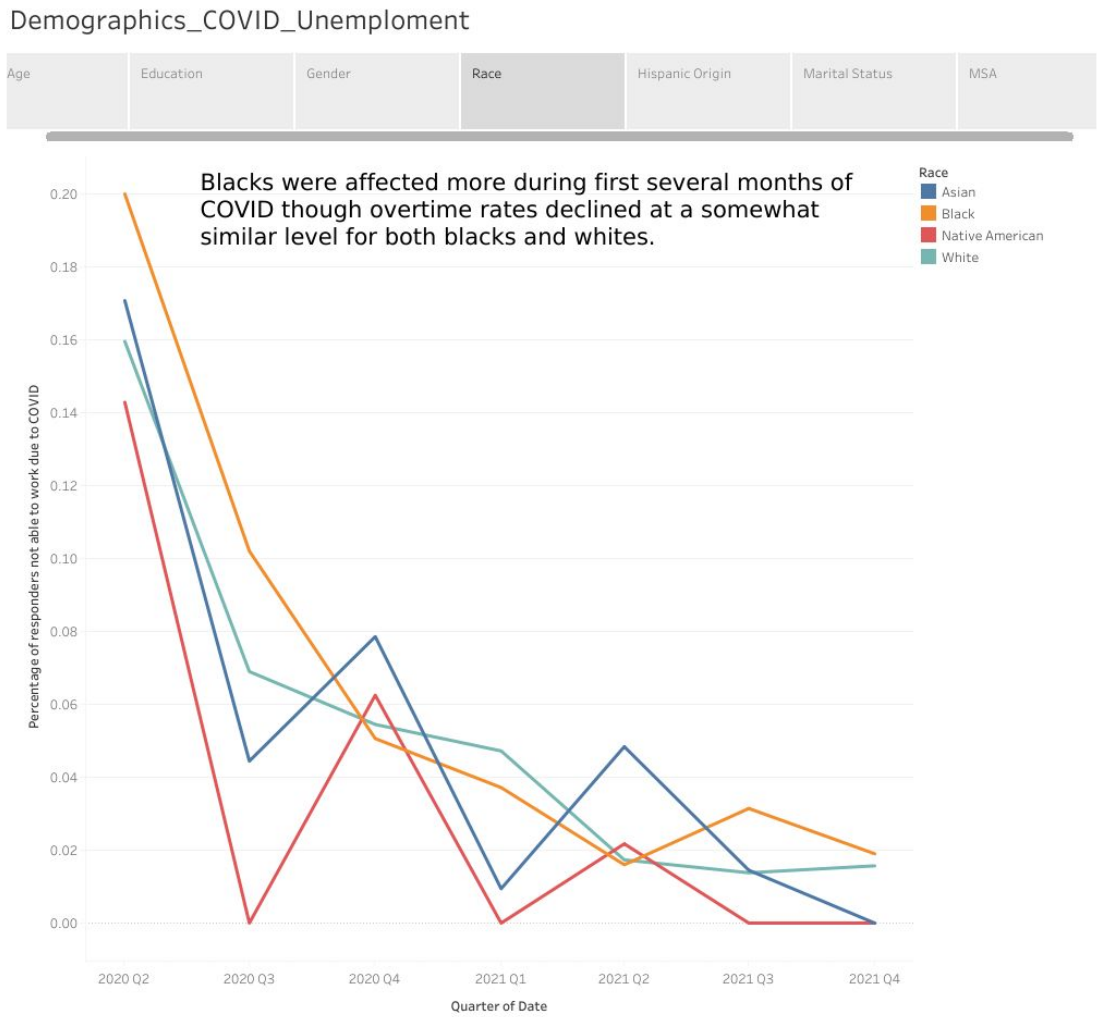
Age | **Education** | Gender | Race | Hispanic Origin | Marital Status | MSA



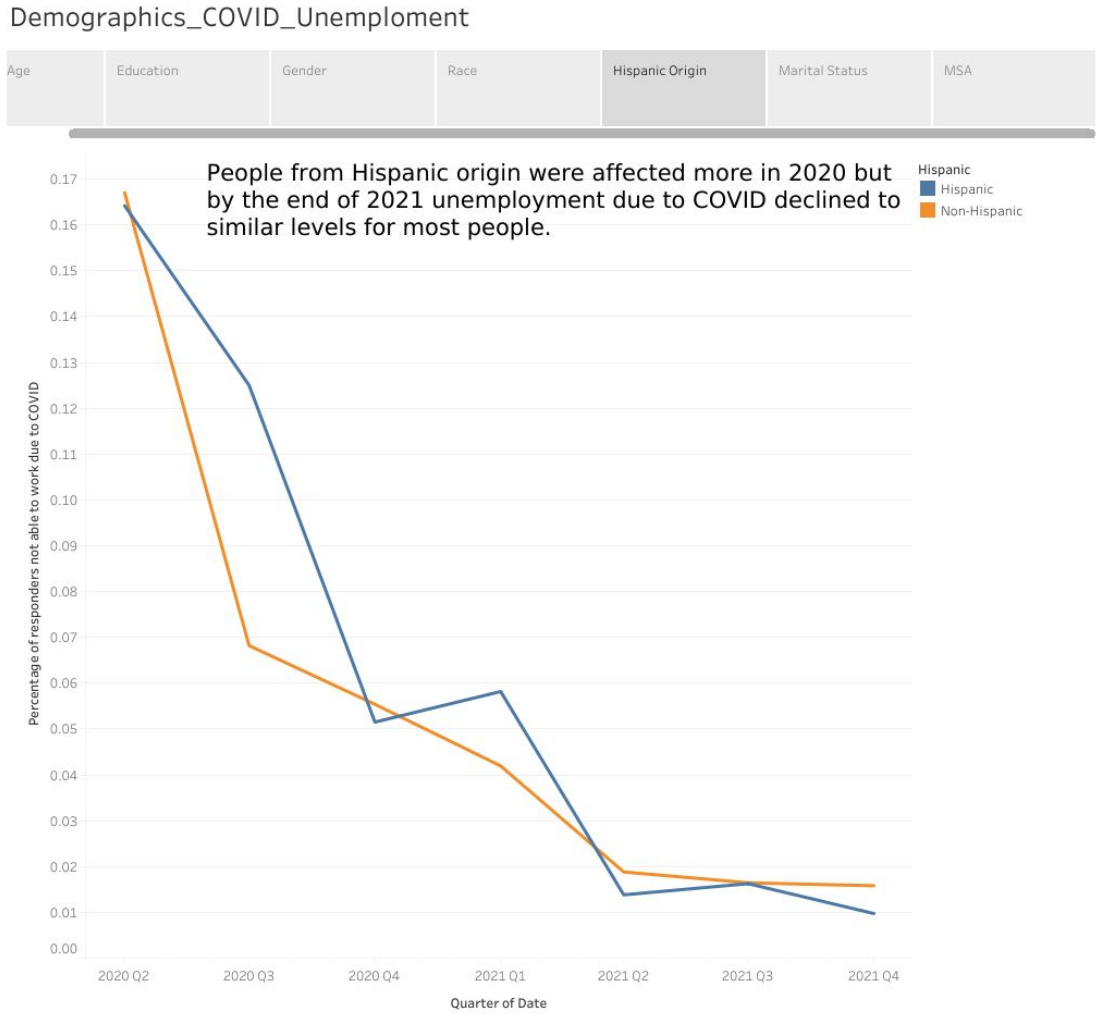
Gender


Demographics_COVID_Unemploment
Age
Education
Gender
Race
Hispanic Origin
Marital Status
MSA
Women were affected more in the beginning but over time unemployment due to COVID declined to similar and low levels ..
Gender
Female
Male
Percentage of responders not able to work due to COVID
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2020 Q2
2020 Q3
2020 Q4
2021 Q1
2021 Q2
2021 Q3
2021 Q4
Quarter of Date

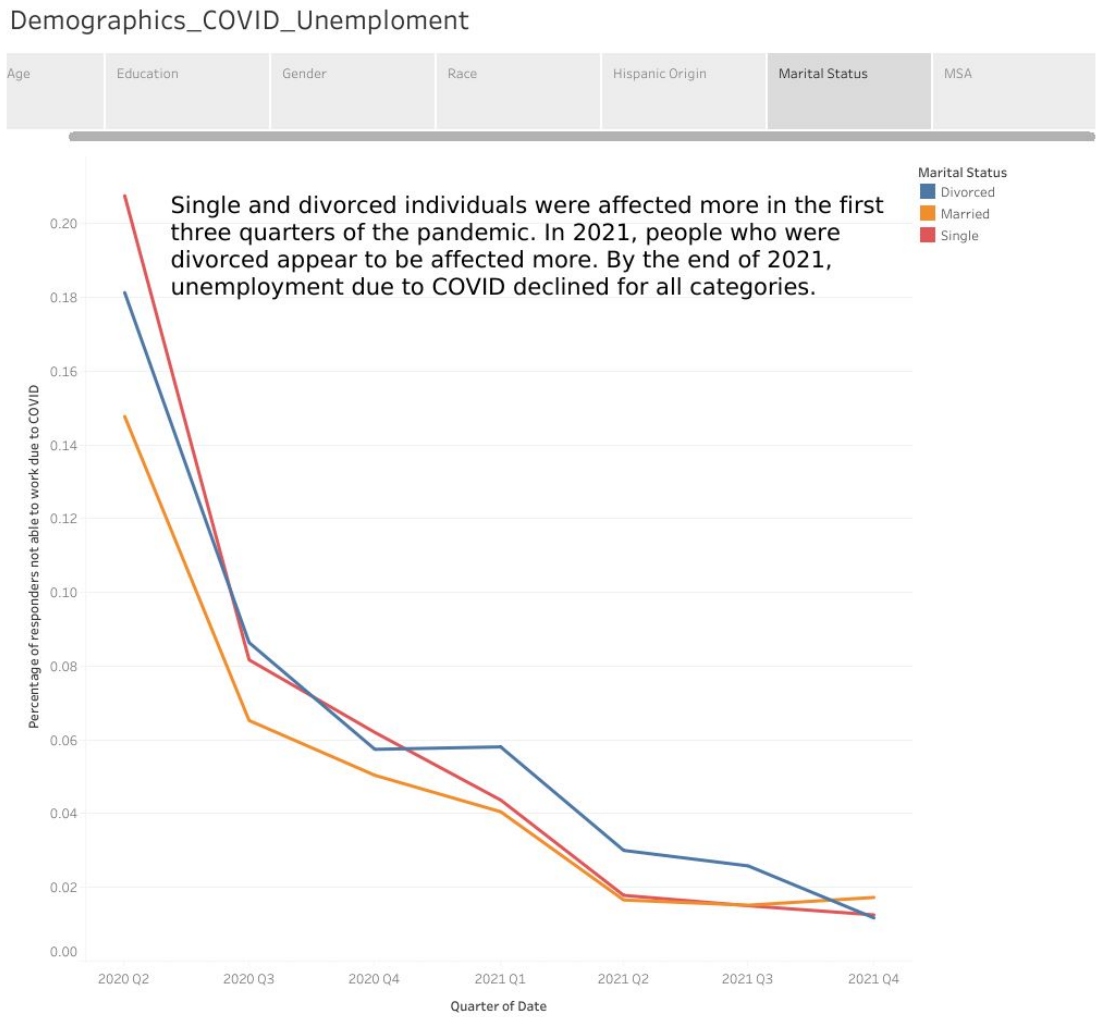
Race


Demographics_COVID_Unemploment
Age
Education
Gender
Race
Hispanic Origin
Marital Status
MSA
Blacks were affected more during first several months of COVID though overtime rates declined at a somewhat similar level for both blacks and whites.
Race
Asian
Black
Native American
White
Percentage of responders not able to work due to COVID
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2020 Q2
2020 Q3
2020 Q4
2021 Q1
2021 Q2
2021 Q3
2021 Q4
Quarter of Date

Hispanic Origin

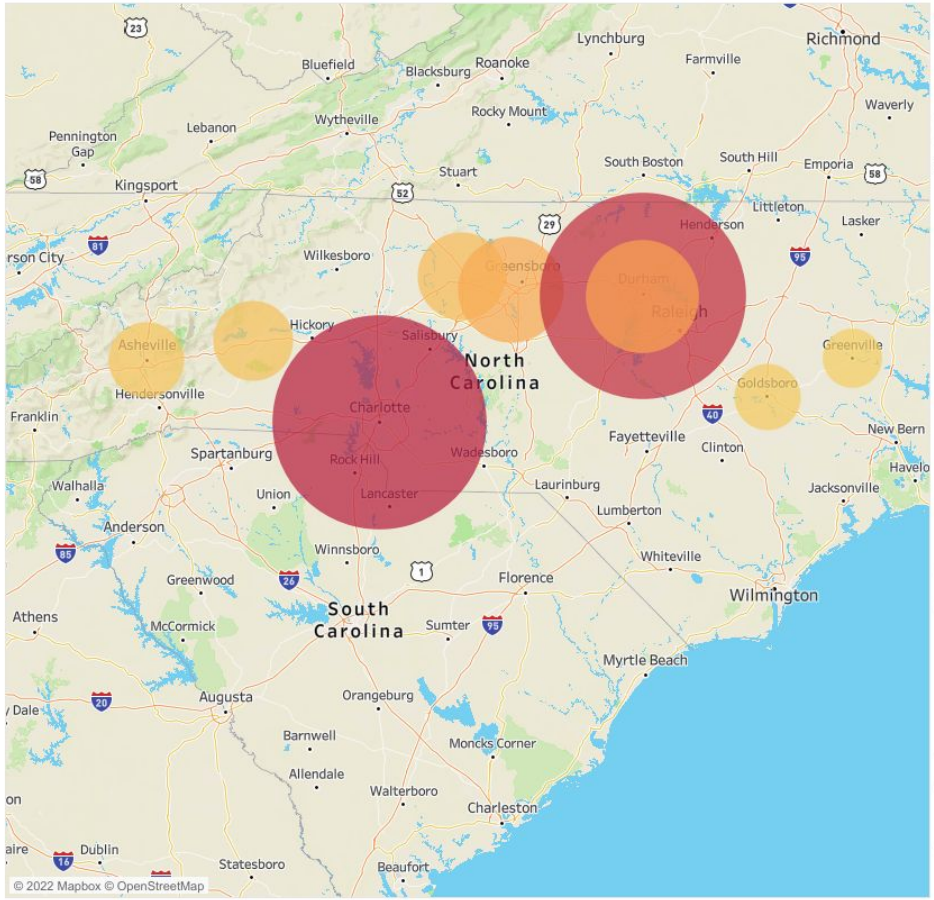


Marital Status



Demographics_COVID_Unemploment

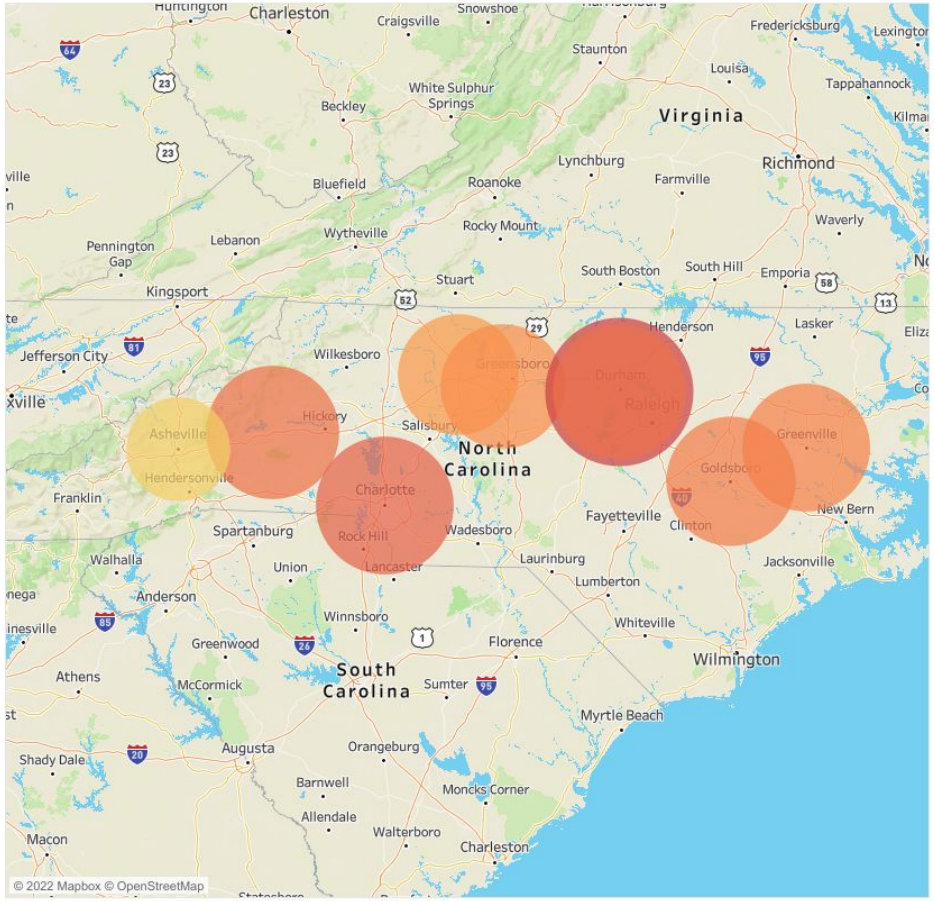
Race | Hispanic Origin | Marital Status | MSA | Machine Learning Model 1 | Machine Learning Model 2 | Machine Learning Model 3



Map of North Carolina showing total number of people who responded unable to work due to COVID in different metropolitan areas

Demographics_COVID_Unemploment

Hispanic Origin | Marital Status | MSA Total Affected | MSA Percentage Affected | Machine Learning Model 1 | Machine Learning Model 2 | Machine Learning Model 3



This map shows that urban MSA's were more affected percentage wise

Machine Learning Model 1

Demographics_COVID_Unemploment

Results from Supervised Learning Model using the SciKit Learn (sklearn) library

In [37]:

```
# Displaying results
print("Confusion Matrix")
display(cm_df)
print(f"Accuracy Score : {acc_score}")
print("Classification Report")
print(classification_report(y_test, predictions))
```

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	4361	213
Actual 1	184	44

```
Accuracy Score : 0.917326114119117
Classification Report
              precision    recall  f1-score   support

           1       0.96      0.95      0.96      4574
           2       0.17      0.19      0.18       228

    accuracy                           0.92      4802
   macro avg       0.57      0.57      0.57      4802
weighted avg       0.92      0.92      0.92      4802
```

Machine Learning Model 2

Demographics_COVID_Unemploment

Race	Hispanic Origin	Marital Status	MSA	Machine Learning Model 1	Machine Learning Model 2	Machine Learning Model 3

Results from Machine Learning Model using gradient boosting to overcome class imbalance

In [17]:
```
# Finally, we can generate a classification report to evalu

print("Classification Report")
print(classification_report(y_test, predictions))
```

```
Classification Report
              precision    recall  f1-score   support

           1       0.96      1.00      0.98      4587
           2       0.00      0.00      0.00       215

    accuracy                           0.96      4802
   macro avg       0.48      0.50      0.49      4802
weighted avg       0.91      0.96      0.93      4802
```

Machine Learning Model 3

Demographics_COVID_Unemploment

Race	Hispanic Origin	Marital Status	MSA	Machine Learning Model 1	Machine Learning Model 2	Machine Learning Model 3

Results from Machine Learning Model using Imbalanced Learn Library (imblearn) and Random Over Sampler method

In [27]:
```
# We'll use the classification_report_imbalanced to do so.

from imblearn.metrics import classification_report_imbalanced
print(classification_report_imbalanced(y_test, y_pred))
```

	pre	rec	spe	f1	geo	iba	sup
1	0.98	0.65	0.69	0.78	0.67	0.45	4574
2	0.09	0.69	0.65	0.16	0.67	0.45	228
avg / total	0.93	0.65	0.69	0.75	0.67	0.45	4802

Recommendations for Future Analysis

- We can include occupation, industry, and class worker information from American Community Survey to predict which workers in which industry were most affected due to COVID