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 employment 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?

Cleaning the data involved the following steps:

Filtering the data for North Carolina by using STATE FIP

		ic_data ic_data		demograph:	ic_data_o	d+ [der	nogra	phic_d	ata_df[STATEF	[P'] ==	= 37]	
	YEAR	MONTH	STATEFIP	METAREA	COUNTY	AGE	SEX	RACE	MARST	HISPAN	EDUC	COVIDTELEW	COVIDUNAV
63258	2020	5	37	3122	0	54	1	100	6	0	111	1	3
63259	2020	5	37	3121	37067	51	2	100	4	0	91	1	
63260	2020	5	37	3121	37067	49	1	100	6	0	111	1	
63261	2020	5	37	1521	37119	65	2	100	1	0	73	99	
63262	2020	5	37	1521	37119	61	1	100	1	0	73	2	

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

```
# Filtering target columns to keep valid data and drop 99 values

2 demographic_data_df_NC = demographic_data_df_NC[demographic_data_df_NC['COVIDUNAW'] != 99]

3 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 1

63259 2020 5 37 3121 37067 51 2 100 4 0 91 1 1 1

63260 2020 5 37 3121 37067 40 1 100 6 0 111 1 1
```

• 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",
 9
               50: "High School or below",
               124: "Graduate or Professional Degree",
10
11
               71: "High School or below",
12 }
   #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
                                        54
                                                                                                      Male
                                                  100
                                                                        111
                                                                                                                             Bachelor's
                                                                                                                 Some College or Associate
       2020
63259
                 5
                        3121
                                37067
                                                                                                  1 Female
63260
       2020
                 5
                        3121
                                37067
                                                  100
                                                                        111
                                                                                                      Male
                                                                                                                             Bachelor's
                                                                                                                     High School or below
63262
       2020
                 5
                        1521
                                37119
                                                                         73
                                                                                                      Male
63268 2020
                        3122
                                                                                                  1 Female
                                                                                                                             Bachelor's
```

• Combining year and month column to create date column

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

Out[5]: NTH 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
```

n----

Binning Age variable using pd.cut

Removing all Nan values and exporting clean file to csv

2 (demog	graph:		n values _df_NC.dro _df_NC	opna (how=	'any'	,inp	lace =	True)								
)	YEAR	MONTH	METAREA	COUNTY	AGE	SEX	RACE	MARST	HISPAN	EDUC	COVIDTELEW	COVIDUNAW	gender	education	race	hispani
632	258	2020	5	3122	0	54	1	100	6	0	111	1	1	Male	Bachelor's	White	No: Hispan

• 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]:
                                                                                                                                 education_Some
                                                                                                 education Graduate education High
                                                                                                                                      College or
                 METAREA COUNTY AGE COVIDUNAW gender_Female gender_Male education_Bachelor's
                                                                                                     or Professional
                                                                                                                        School or
                                                                                                                                       Associate
                                                                                                           Degree
                                                                                                                          below
                                                                                                                                         Degree
                     3122
                                                                                                                0
                     3121
                                                                           0
                                                                                              0
                                                                                                                0
                                                                                                                              0
```

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):

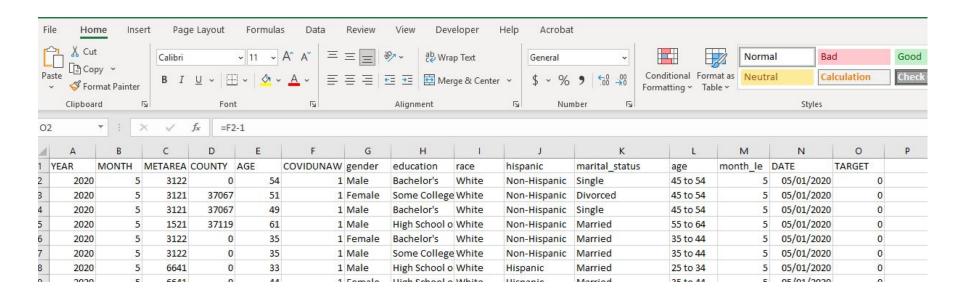
COVIDUNAW - individuals who are unable to work during COVID (1: able to work, 2: unable to work)

Statistics summary using 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

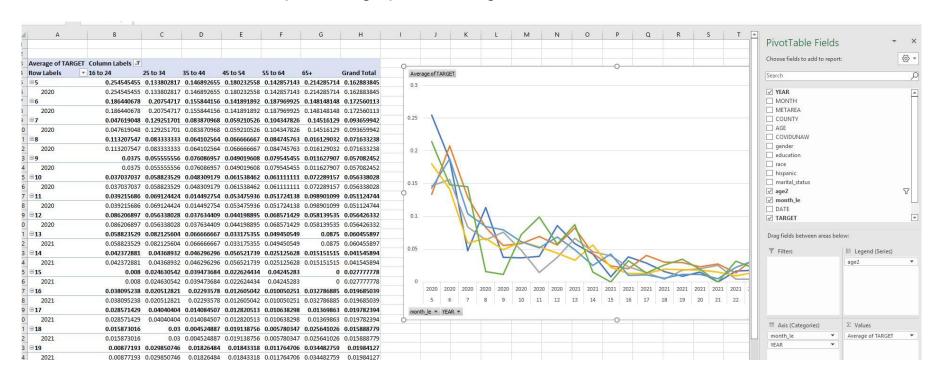
Data Analysis - Process

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



Data Analysis - Process

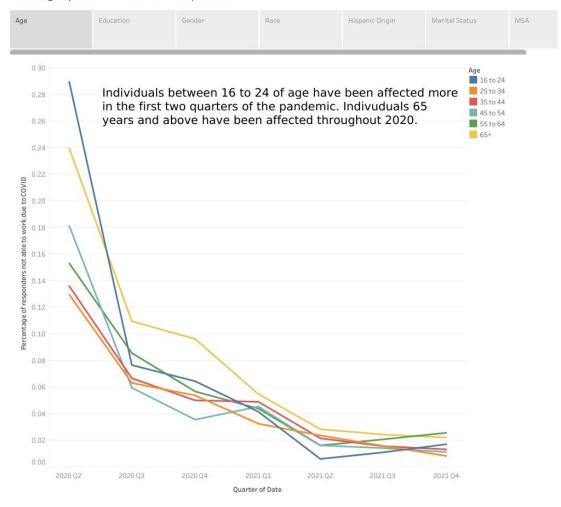
Created Pivot Tables for quick line graphs for categorical variables



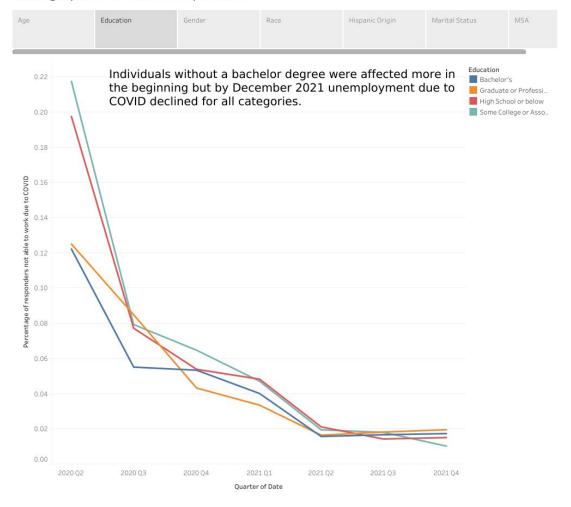
Technologies, languages, tools, and algorithms

- **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

Age



Education



Gender

Demographics_COVID_Unemploment

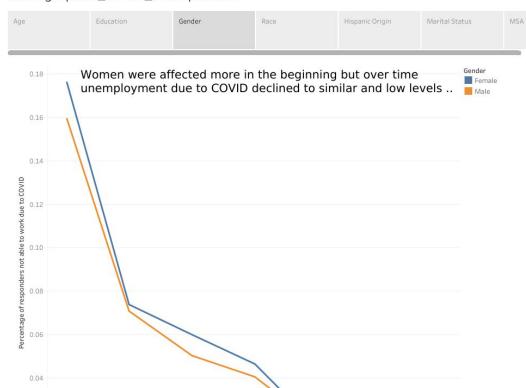
0.02

0.00

2020 Q2

2020 Q4

Quarter of Date

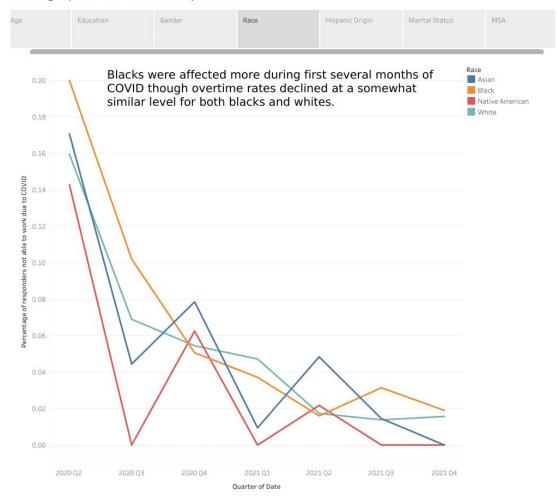


2021 Q3

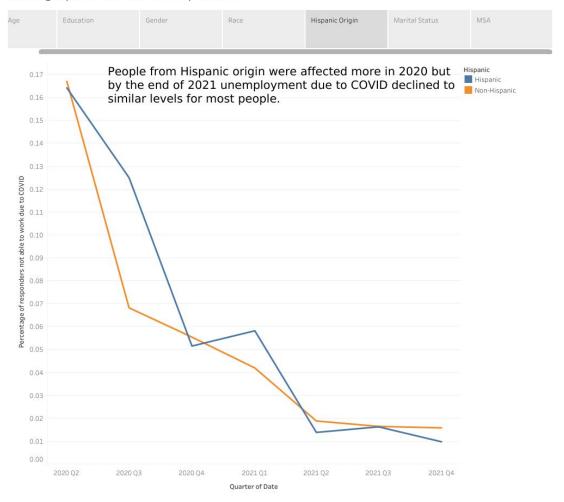
2021 Q4

2021 02

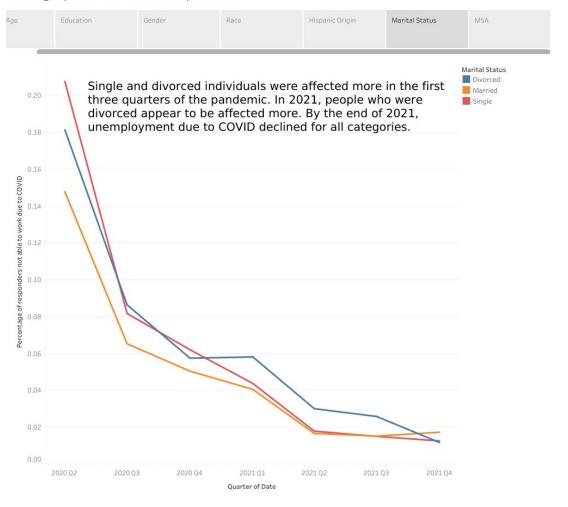
Race

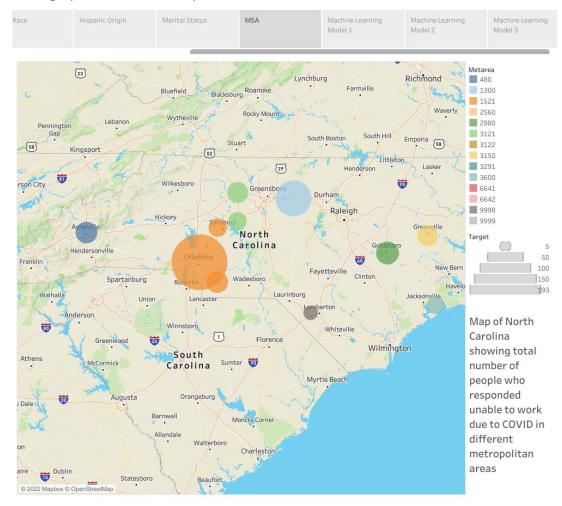


Hispanic Origin



Marital Status



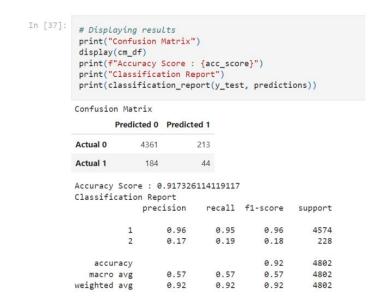


Machine Learning Model 1

Demographics_COVID_Unemploment

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

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



Demographics_COVID_Unemploment

Machine Learning Model 2

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

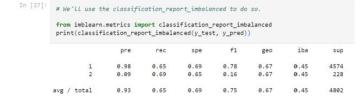
```
# Finally, we can generate a classification report to evalue
 print("Classification Report")
 print(classification_report(y_test, predictions))
Classification Report
              precision
                          recall f1-score
                                             support
                  0.96
                            1.00
                                      0.98
                                                4587
                  0.00
                            0.00
                                      0.00
                                                 215
    accuracy
                                      0.96
                                                4802
                  0.48
                                                4802
   macro avg
weighted avg
                  0.91
                            0.96
                                                4802
                                      0.93
```

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



Dashboard

The dashboard will be created in Tableau - the link to which will be embedded in a web application. The interactive element will be the map of North Carolina with categorical variables as layers. The following sheets will be created as part of the Dashboard. The demographic characteristics and education sheets will include data analysis including statistics and line graphs. The data modeling sheet will include analysis from the machine learning model.

