Data 602 – Final Project Proposal

By: Mubashira Qari

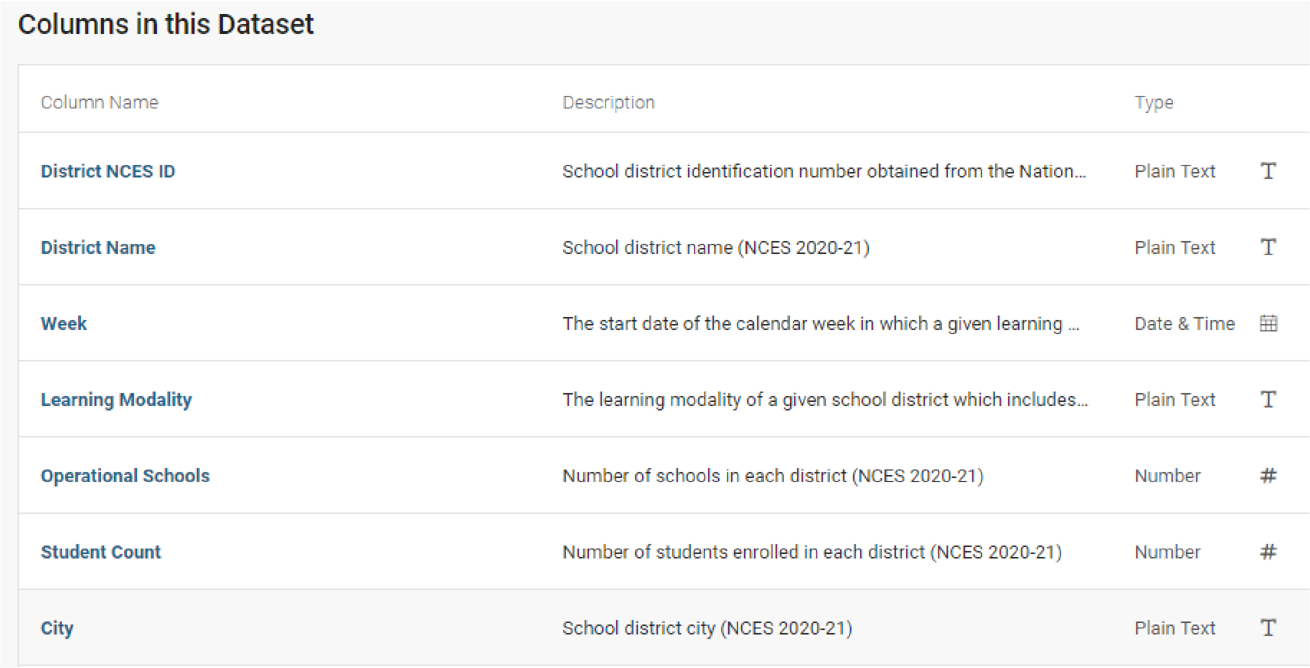
Project Title: **Impact of Covid-19 on Student's Learning Modalities (Year 2021-2022) & Prediction Using Supervised Machine Learning**

Describing Dataset: The first dataset is the weekly summary of school learning modalities. It has 923K row and 9 columns.

The second dataset is the US county level covid-19 cases and deaths counts. It has 2249807 rows and 6 columns.

These datasets require detailed data cleaning before performing the data manipulation steps.

The metrics for the columns in the dataset are as below:



Columns in this Dataset

Column Name	Description	Type	
District NCES ID	School district identification number obtained from the Nation...	Plain Text	T
District Name	School district name (NCES 2020-21)	Plain Text	T
Week	The start date of the calendar week in which a given learning ...	Date & Time	▦
Learning Modality	The learning modality of a given school district which includes...	Plain Text	T
Operational Schools	Number of schools in each district (NCES 2020-21)	Number	#
Student Count	Number of students enrolled in each district (NCES 2020-21)	Number	#
City	School district city (NCES 2020-21)	Plain Text	T

Data Source Link:

The datasets are acquired from the following sources,

- HealthData.gov
- NYTimes Covid-19 data – GitHub
- US Zip code to County State to FIPS Look Up – data.world
- States Names and Abbreviations - GitHub

and the links are provided below:

https://healthdata.gov/National/School-Learning-Modalities/aitj-yx37

https://github.com/nytimes/covid-19-data

https://data.world/niccolley/us-zipcode-to-county-state

https://github.com/jasonong/List-of-US-States/blob/master/states.csv

Justification for Dataset Selection:

The reason for choosing these datasets is that I am interested in finding out the impact of covid-19 on the learning modalities for the students.

My focus will be to investigate the state level trend for the hybrid, remote and in-person learning due to covid for the year 2021 and 2022.

Research Questions & Objectives:

The research questions for this project are the following:

1- Which state has the most covid cases for the year 2021 and 2022?

2- Which state has the most deaths due to covid for the year 2021 and 2022?

3- Which state has the highest average student count for hybrid, remote and in-person learning modality for the year 2021 and 2022?

7- Clean and merge the datasets.

8- Upload the dataset in the PostgreSQL database for further analysis.

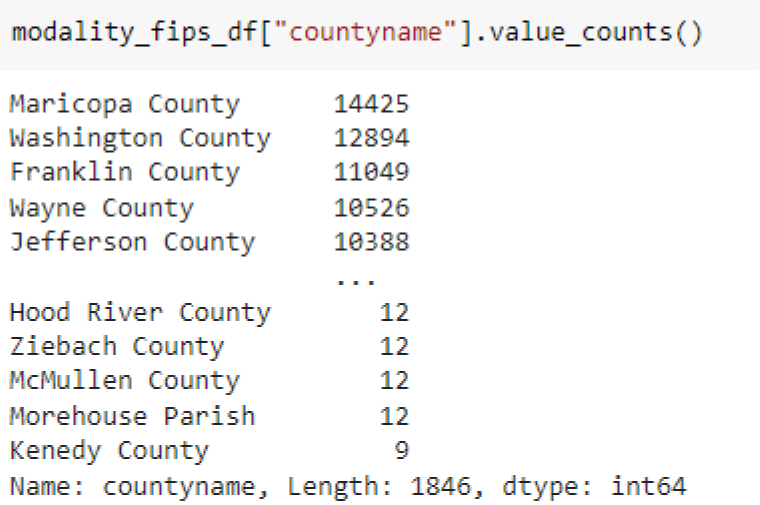
9- Prediction of learning modalities using machine learning.

Libraries, Visualization Apps & Database Used for Project Implementation:

- Python Pandas
- Python NumPy
- Python Matplotlib
- PostgreSQL Database
- Python sklearn
- Python seaborn
- Plotly Dash App
- Tableau App
- SQLAlchemy python SQL Toolkit and Object Relational Mapper

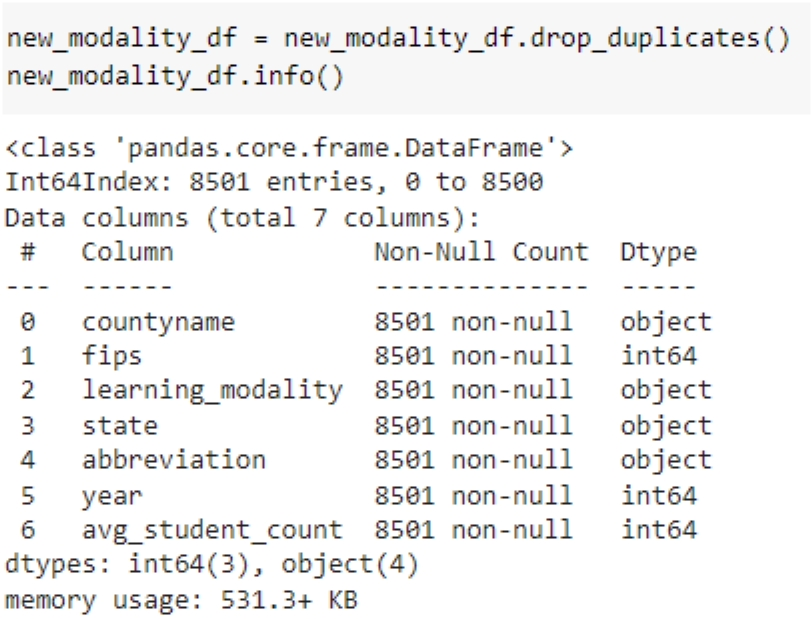
EDA and Summary Statistics:

Below are the images of exploratory data analysis and summary statistics:



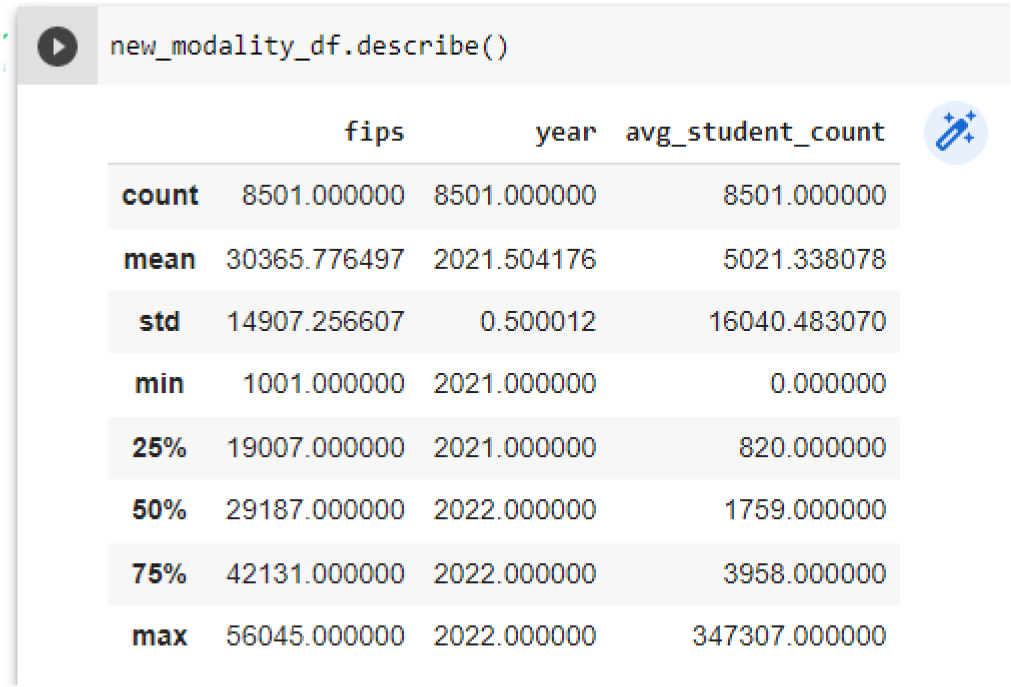
```
modality_fips_df["countyname"].value_counts()
```

```
Maricopa County      14425
Washington County    12894
Franklin County      11049
Wayne County         10526
Jefferson County     10388
                       ...
Hood River County       12
Ziebach County          12
McMullen County         12
Morehouse Parish        12
Kenedy County            9
Name: countyname, Length: 1846, dtype: int64
```



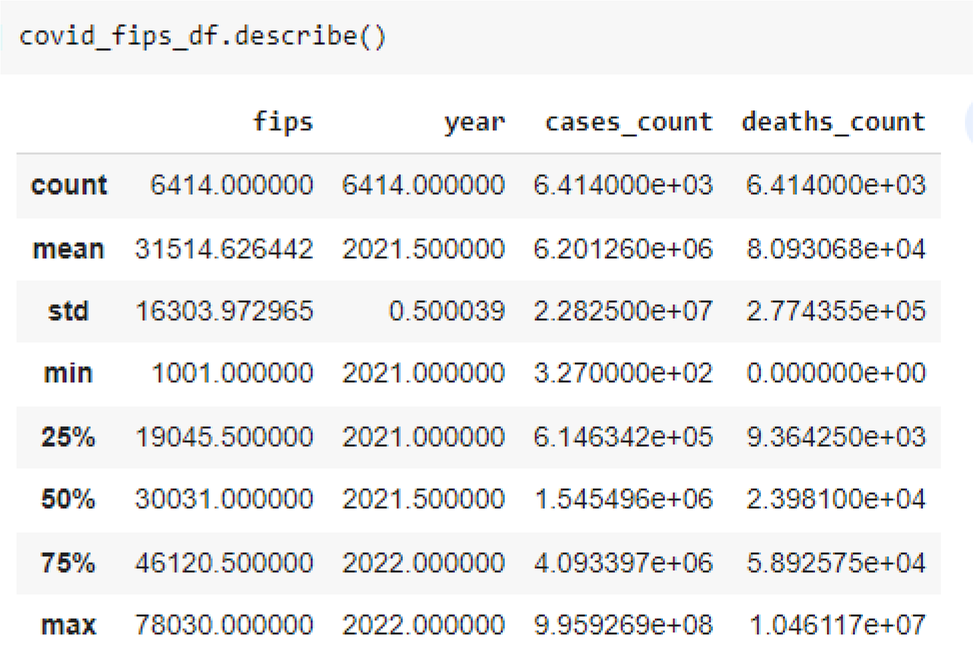
```
new_modality_df = new_modality_df.drop_duplicates()
new_modality_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8501 entries, 0 to 8500
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   countyname         8501 non-null   object
 1   fips               8501 non-null   int64 
 2   learning_modality  8501 non-null   object
 3   state              8501 non-null   object
 4   abbreviation       8501 non-null   object
 5   year               8501 non-null   int64 
 6   avg_student_count  8501 non-null   int64 
dtypes: int64(3), object(4)
memory usage: 531.3+ KB
```



```
new_modality_df.describe()
```

	fips	year	avg_student_count
count	8501.000000	8501.000000	8501.000000
mean	30365.776497	2021.504176	5021.338078
std	14907.256607	0.500012	16040.483070
min	1001.000000	2021.000000	0.000000
25%	19007.000000	2021.000000	820.000000
50%	29187.000000	2022.000000	1759.000000
75%	42131.000000	2022.000000	3958.000000
max	56045.000000	2022.000000	347307.000000

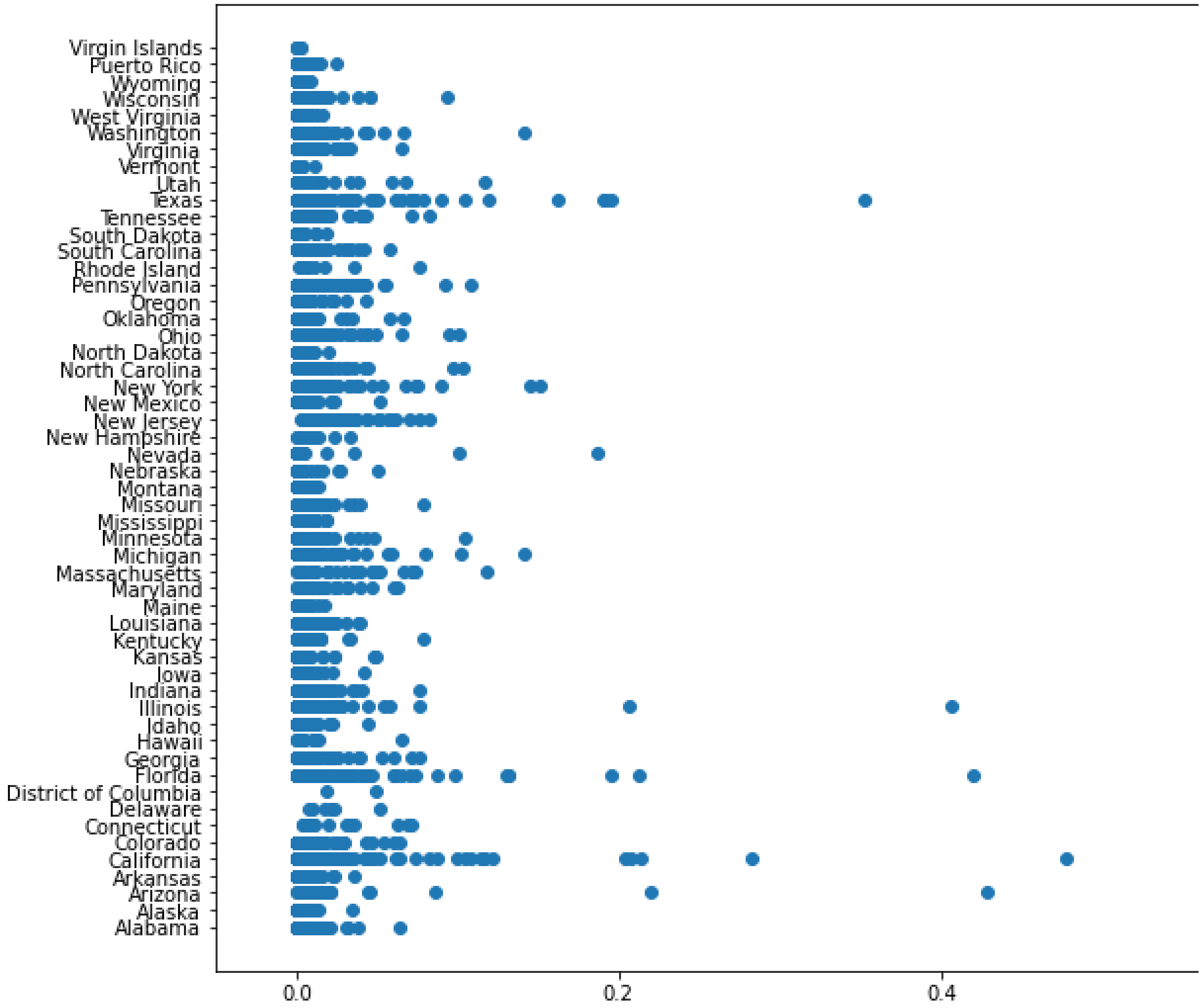


```
covid_fips_df.describe()
```

	fips	year	cases_count	deaths_count
count	6414.000000	6414.000000	6.414000e+03	6.414000e+03
mean	31514.626442	2021.500000	6.201260e+06	8.093068e+04
std	16303.972965	0.500039	2.282500e+07	2.774355e+05
min	1001.000000	2021.000000	3.270000e+02	0.000000e+00
25%	19045.500000	2021.000000	6.146342e+05	9.364250e+03
50%	30031.000000	2021.500000	1.545496e+06	2.398100e+04
75%	46120.500000	2022.000000	4.093397e+06	5.892575e+04
max	78030.000000	2022.000000	9.959269e+08	1.046117e+07

```
plt.scatter(covid_fips_df["cases_count"], covid_fips_df["state"])
plt.show()
```

Scatter Plot:


Virgin Islands
Puerto Rico
Wyoming
Wisconsin
West Virginia
Washington
Virginia
Vermont
Utah
Texas
Tennessee
South Dakota
South Carolina
Rhode Island
Pennsylvania
Oregon
Oklahoma
Ohio
North Dakota
North Carolina
New York
New Mexico
New Jersey
New Hampshire
Nevada
Nebraska
Montana
Missouri
Mississippi
Minnesota
Michigan
Massachusetts
Maryland
Maine
Louisiana
Kentucky
Kansas
Iowa
Indiana
Illinois
Idaho
Hawaii
Georgia
Florida
District of Columbia
Delaware
Connecticut
Colorado
California
Arkansas
Arizona
Alaska
Alabama