**Purpose of the Data 602 Final Project:**

**Introduction:**

The topic of my final project is, “Impact of Covid of Learning Modalities -Year 2021-2022”. The reason I choose to investigate this topic is because there is a major shift in students’ learning modalities due to the impact of Covid-19 since the beginning of the year 2020. Almost all the educational institutions had to switch to on-line learning in the beginning. As the rate of the Covid-19 cases got reduced, the institutions have provided options of hybrid, remote and in-person learning. Now the year 2022 is ending and still those options are available. And I am very curious to find out the number of students going for these learning modalities for the year 2021 and 2022 at the US state level.

Since Covid-19 is the cause of the transition for these learning modalities, I will also be investigating the covid cases and death counts too for the year 2021 and 2022 at the US state level in this project.

**Workflow for the Project:**

**data\_602\_final\_project.ipynb:**

The following steps are performed in this file:

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.

I started previewing the data by calling the dataframe using "info()" function and finding the number of columns and rows.

This function provides the total number of rows and columns as well as the datatype of each column.

Also, we  can see the count of non-null values, which will help us further investigate the missing, or unexpected values in our dataset.

These datasets require detailed data cleaning before performing the data manipulation steps. After importing the libraries which are required for performing the exploratory data analysis and data manipulation, I started the look closely into the datasets.

I have performed the fixing of the missing values rows, duplicate rows, and hidden white spaces in the values. Both the datasets have data at weekly level. I have used Pandas dataframe time functions to first convert the week column into year and month columns. I have dropped the unwanted columns, and rows from the dataset.

One of the facts that I have discovered in the dataset is that the rows for the US Hawaii state has incorrect student count. Basically, the numbers are repeating for the hybrid, remote and in-person student count. Then I went to the original data source website and checked the data there with the same issue. For this reason, I have filtered out the Hawaii state from the dataset.

Then, I have used Pandas ‘Group By’ with aggregate average function to bring the student count in hybrid, remote and in-person categories in the dataset at the US state level.

I have converted the column names to lower case which helps in exporting the datasets without causing issues to other applications. Also, converted the incorrect data types to the correct data type.

There is only state “Abbreviation” column given for the states in the school learning modalities dataset. For this reason, I have downloaded a state names and abbreviations dataset and performed the data merge steps so that the state names be added in the school learning modalities dataset.

Next, I have performed the above-mentioned data cleaning steps on the Covid-19 dataset at the state level and in the end three files are exported and saved in the resources folder, which are as follows.

1. posgres\_modality\_df.csv
2. posgres\_covid\_df.csv
3. posgres\_states\_df.csv

**census\_api.ipynb:**

In this file, the posgres\_modality\_df.csv and posgres\_covid\_df.csv files have been imported.

An API call for the census dataset has been performed to fetch the population of each state.

The objective in this file is to convert the student count in hybrid, remote and in-person and the covid count to correct proportion by dividing it to the population of each state. And then normalizing the values by multiplying them to 10K.

Finally, the following files have been exported:

1. tableau\_modality\_df.csv
2. tableau \_covid\_df.csv
3. posgres\_states\_df.csv

**PostgreSQL Database:**

A relational database schema diagram has been made using the tool Quick DBD and PostgreSQL queries have been generated.

Finally, PostgreSQL database has been created.

**data\_602\_machine\_learning\_analysis.ipynb:**

Here, a thorough supervised machine learning steps have been performed. KNN algorithm, Logistic Regression and Decision Tree model have been applied and accuracy have been calculated.

Decision Tree has the highest accuracy among the three models but still very low.

Since there are not enough features in the dataset for prdiction, that might be one of the causes of low accuracy.

Another reason might be there is not enough data.

Also, dataset has been throughly cleaned, so noise might not be the cause of low accurary.

And queries are ready to be imported into the python file using SQLAlchemy – The Database Toolkit for Python

**Tableau Visualization Analysis:**

Nevada has the highest learning count for hybrid in 2021. Whereas Maryland has the highest student count for hybrid in 2022.

Nevada has the highest student count per 10k between 2021 - 2022.

North Dakota has the highest covid cases per 10k in 2021. Rhode Island has the highest covid cases per 10k in 2022.

Mississippi has the highest death count per 10k between 2021 – 2022.