**KDD-Project :**

**Instructions to run the project**

* Take the clone of the repository in the local system.

(Week wise Data is present in csv format in the respective folders.)

* Go to the folder for the sector you want
* Then run the jupyter notebook, you’ll be able to access the dataFrame.

**Education sector :**

Education Folder - Impact\_of\_COVID\_on\_Education\_sector.ipynb [Dashboard](https://public.tableau.com/profile/ashish.bisht3877#!/vizhome/Education_16081543152430/Dashboard1)

**Employment sector :**

Employment Folder - Employment\_NC.ipynb [Dashboard](https://public.tableau.com/views/CovidimpactmodelonEmployment/Dashboard1?:language=en&:display_count=y&publish=yes&:origin=viz_share_link)

**Health Sector :**

Health Folder - Health\_Exploratory\_data\_analysis.ipynb [Dashboard](https://public.tableau.com/views/Health_16081557626650/Dashboard1?:language=en&:display_count=y&publish=yes&:origin=viz_share_link)

**Spending :**

Spending Folder - Spending\_Analysis.ipynb [Dashboard](https://public.tableau.com/views/Spending_16081571144880/Dashboard1?:language=en&:display_count=y&publish=yes&:origin=viz_share_link)

**Project : Studying impact of COVID'19 on different sectors**

The COVID-19 outbreak is a sharp reminder that pandemics can change the course of mankind and will continue to do so. Even if we cannot prevent dangerous viruses from emerging, we should be prepared to dampen their effects on society. The current outbreak has consequences not only on the economy but overall on the society. In less than three months at the beginning of 2020, the COVID-19 crisis developed into a global pandemic. Schools and universities were closed in Spring'20 for more than one billion students of all ages. By June 2020, the virus spread to almost all countries and affected more than eight million people around the world. More than half of the world’s population has experienced a lockdown with strong containment measures – the first time in history that such measures have been applied on such a large scale.

The sectors that have seen the largest increases in unemployment are those that are hedonic in nature and require the physical presence of the customer (e.g., hospitality, tourism, construction and stores), as demand for these services has ceased to exist. Intuitively, we understand that the crisis will not only leave many an organization struggling for survival, but will also force some to look for alternative strategic paths. While on the one hand, the Covid-19 crisis has imposed enormous challenges on business organizations, on the other, it has also necessitated innovations, presenting organizations with opportunities to identify new business models that will allow them to survive through the crisis. Past experience indicates that once someone is outside the job market, it is very difficult to get back on their feet. Therefore, we have decided to measure the most impacted people in different sectors like business, public-private sector employment, education and health. We want to visualize the impacts on different sectors in order to help people to prepare for their future job security and to accept beneficial changes taking place during this pandemic.

**Team Members:**

[Nandani Dabhi](https://www.linkedin.com/in/nandanidabhi/)

[Aakanksha Chauhan](https://www.linkedin.com/in/aakanksha-chauhan-a3a52a55/)

[Priyanka Jadhav](https://www.linkedin.com/in/priyanka-jadhav06/)

[Salma Khayum](https://www.linkedin.com/in/salmaroohik/)

[Ashish Bisht](https://www.linkedin.com/in/ashish-bisht-706892134/)

**Data Sources**

[Dataset: 'Household Pulse Survey Data Tables for 2020' Reference](https://www.census.gov/programs-surveys/household-pulse-survey/data.html)

We are doing Data analysis for the State of North Carolina.

Household pulse survey is designed to collect data on how people's lives have been impacted by COVID-19. Data was collected into 3 phases : Phase 1 of the Household Pulse Survey began on April 23, 2020 and ended on July 21, 2020. It includes 12 weekly table releases. Phase 2 of the survey began on August 19, 2020 and ended October 26, 2020, and included 5 biweekly table releases. Phase 3 of the survey began on October 28, 2020, and the Census Bureau expects to collect data until December 21, 2020, releasing data every two weeks just like in Phase 2.

In all the phases, data was collected on Education, Employment, Food Sufficiency and Food Security, Health, Housing, Peoples' spendings and transportation.

Phase 1 :

Week 1 : April 23 - May 5

Week 2 : May 7 - May 12

Week 3 : May 14 - May 19

Week 4 : May 21 - May 26

Week 5 : May 28 - June 2

Week 6 : June 4 - June 9

Week 7 : June 11 - June 16

Week 8 : June 18 - June 23

Week 9 : June 25 - June 30

Week 10 : July 2 - July 7

Week 11 : July 9 - July 14

Week 12 : July 16 - July 21

Phase 2 :

Week 13 : August 19 - August 31

Week 14 : Sept 2 - Sept 14

Week 15 : Sept 16 - Sept 28

Week 16 : Sept 30 - Oct 12

Week 17 : Oct 14 - Oct 26

Phase 3 :

Week 18 : Oct 28 - Oct 9

We are using consolidated data for all the phases to come up with our observations.

**Application of the CRISP-DM Process:**

**Domain Knowledge :** In order to get an idea about the impact of COVID-19 on various sectors we reviewed various COVID related articles and research papers.

**Data understanding and EDA :**

For the project our main focus is on Data understanding and EDAphase on CRISP-DM.

EDA is the process of visualizing and analyzing data to extract insights from it. It helps in summarizing important characteristics of data in order to gain better understanding of the dataset.The categorical and numerical features are studied to gain better understanding of the dataset. We have analyzed the data using various graphs and plots like bar plots, box whisker plots etc to uncover important relations between the data.

**Data Preparation:**

1. wrote a web scraper to get the data files from the census site

2. decided the features we wanted to use.

3. created temporary data frames to hold the data for each week and extracted the selected features.

4. merged the week wise selected features in a single dataframe

5. handled missing values in the features column

6. Added a new column week to keep track of week number

7. used this dataframe to prepare the final dataset used for visualization

8. visualized the data with different plots.

**Handling Discrimination and bias:**

There is no bias in our dataset as it does not contain any user related information. So we dont need to handle discrimination and bias

**Machine Learning:**

We are doing descriptive data analysis which may not need extensive machine learning. We can divide our dataset into training and test data for making predictions. The dataset consists of time-series data which helps in analyzing the past data and becomes an essential factor in forecasting the future. Autoregressive Integrated Moving Average (**ARIMA**) is one of the most popular algorithms in Time Series forecasting. An ARIMA model is one where the time series differencing is done at least once in order to make it stationary where we combine the AR (autoregressive) and MA (Moving average) terms. In Python, the pmdarima package provides auto\_arima() function which can be used to automate the process of ARIMA Forecasting in Python. The function auto\_arima() uses a stepwise approach to search multiple combinations of p,d,q parameters and chooses the best model that has the least AIC.

**Evaluation: Best model: ARIMA(2,0,2)(0,0,0)[0]**

* In Python, the pmdarima package provides us with auto\_arima() function which can be used to automate the process of ARIMA Forecasting in Python.
* auto\_arima() uses a stepwise approach to search multiple combinations of p,d,q parameters and chooses the best model that has the least AIC

**Interesting Findings :**

1. **Education Sector:** Most education institutes/universities changed their format from in-person to Online at the start of Fall semester so the percentage of Class format change observed spike at the beginning of Fall in week 13.
2. **Employment Sector:** Not too far fetched but still interesting that with the progression of the situation separation rate from quitting the job has gone down drastically. Need not to elaborate that the progression of the curve of employment is inverted since the pandemic.
3. **Health Sector:** Mostly people took mental health help during week15 - week18.

Most people who received counseling or therapy from a mental health professional such as a psychiatrist, psychologist, psychiatric nurse, or clinical social worker belonged to a household income of 50-80 thousand dollars.

1. **Spending Sector:** In the spending sector, it was observed that in the beginning of the pandemic as per the data for week 13, 14,15 the people made more online purchases as well as made more purchases curbside pickups but after settling in for some time in the weeks 16,17 an 18 people resumed going to in-stores for making their purchases.

**Conclusion:**

Many pandemics have occurred in the past and can even occur in the future as well, so being prepared for such pandemics is always a good idea. We hope that the analysis done by us on the impact of COVID-19 will help people to be better prepared and equipped in case of employment, education or healthcare in future.

### 

### **Challenges faced:**

As the current dataset consisted of data collected through a survey, we performed a descriptive analysis for our project. It was difficult to perform machine learning and evaluations for the initial dataset. We are not making any predictions as such for our project, which can be extended further for future research purposes.