**Final Project Proposal**

**Victor H Torres.**

**DATA 602 Advanced Programming Techniques**

**04/14/2023.**

The purpose of the data project is for you to conduct an analysis with a dataset of your choosing. You will use skills learned from the course to assist in your analysis.

For the final project, I decided to use a public dataset from the NYC open data website, the name of the dataset is “courses/training providing listing”, and is about the list of courses, training for individuals, as well as to grants offered by the City of New York.

* **Research Question:** What courses on IT related field are offered by the City of New York? costs and if there are grants available, the requirements for it.
* **Justification:** This data analysis is relevant to my industry, since I live in New York, and I did not know that the city offers a vast amount of help and resources to train their citizens, and I would like to know the requirements to join these courses, also, it would be great to find out if I can pass this information to friends and family that it might be interested in these programs.
* **Data Sources:** The sources for this projectwere found in the NYC open data website, (https://data.cityofnewyork.us/), and also I would try to get data from the State of New York to compare, the data between city and state. Also, I am going to try to obtain some data from another state to figure out which states offer more training programs in the IT field or if there is another industry where other states concentrate in.

https://data.cityofnewyork.us/Business/Courses-Training-Provider-Listing/fgq8-am2v/about\_data

https://data.nj.gov/

https://data.ny.gov/

* **Libraries potentially being used:** For this project, I am going to use all the libraries that we used in this course, I will use Pandas, Numpy, MatplotLib, Seaborn, and Ploty, will be my main resources to obtain the data analysis and visualizations required for this project.
* **EDA and summary statistics:** I am going to use all the EDA and summary statistics learned in this course, means medians and quantiles of each column, I will create new columns to get the results of data analysis between datasets, get the min and max of each column, to obtain the values of my research. Also, I will include graphs and plots, with data visualization on it, displaying statistics and exploratory data analysis on it.
* **Code Samples:**
* nyc\_train = pd.read\_csv("Courses\_Training\_Provider\_Listing\_20240416.csv")
* #Review head of public data frame
* nyc\_train.head()
* #get the mean of column “Cost\_Total
* nyc\_train["Cost\_Total"].mean()
* #Get median of duration column
* nyc\_train["Duration"].median()
* #Review quantile of public data frame
* nyc\_train["NumHour"].quantile()
* #Fill missing values using the NA command.
* nyc\_train.fillna(0)
* print(nyc\_train)