**Project 2: Flights 2021 - 2022**

**Members** -

* Corine Alida Daboiko
* Caroline Christy
* Dylan Thomas Jackson

**Flight Data 2021**

<https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022?select=Combined_Flights_2018.csv>

**Flight Data 2022**

<https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022?select=Combined_Flights_2018.csv>

**Extract Data:** For our initial step as a group we tried finding data sets that somewhat showed a correlation. Initially we chose two data sets representing crime from the years of 2021-2022 in two different countries. After finding out our data sets were simply too small. We decided to find two new data sets matching the required sizes. (Thank you Danila)

Eventually we came across two different data sets containing information on flight patterns from different Airline Companies from the years 2021-2022.

**Transform Data:** We decided to clean each data set inside the Jupyter Notebook. The data sets already having some similarities did not prove to be too difficult in the cleaning or merging process. Our goal was to show that the merge was successful which was done through a trial run. The trial run using the code mostly written by Caroline and Corine showed the data displayed in a table.

There were 121 columns in the data set so we decided to focus on the columns showing NaN. We decided on deleting some of the columns to clean our table and make it more presentable. We were having some trouble trying to filter the data properly without it taking the entirety of class. Due to us having over 6 million rows of data, we decided to ask for a TA’s assistance.

Dominick suggested that we filter the data set that we have already created prior. We waited around 35 minutes before calling Haroon in for help due to the prolonged wait time. Haroon suggested we move out line of code filtering the data above our lines displaying the data. This gave us an error after about 7-8 minutes of waiting which clearly helped our time frame.

At the end of the night Seth helped us through the extended hours of class with our coding errors and even gave us a sheet he created to help our code in certain situations. We stayed up a little later that night as well as the following day to finish our final project. With Seth’s advice on our merge and the few TA’s helping us condense our data. We were able to come out with a completed ETL. Our load process was simple as we just created a database for our code, then pushed our code through to GitHub where we have both data sets included as well as the merged form.

We had troubles at the end trying to create a table to include in our breakdown, but hopefully the data helps show our work put in.