Group 2 Project 3 Proposal:

<https://data.world/dot/airline-on-time-performance-statistics>

Indu and David

* Use data from the 3 CSVs provided in the link above
* Upload to python
* Clean and update the data as needed
* Export to CSV again

Aliyu

* Create ERD diagrams, database schema
* Upload to Postgres

Saudia

* Ensure the database queries work as intended

Everyone

* Readme

**Data Engineering Track Requirements (75 points)**

**Database Design (40 points)**

* The project uses ETL workflows to ingest data into the database. (10 points)
* The original dataset(s) are transformed prior to storing it in the database. (5 points)
* A database is used to house the data (SQL, MongoDB, SQLite, etc.). (5 points)
* The database has at least two tables (SQL) or collections (NoSQL). (5 points)
* The project documents the choice of the database used and why. (5 points)
* The project includes documentation of the ETL workflow with diagrams or ERD. (10 points)

**Data and Delivery (35 points)**

* The database contains at least 100 unique records. (5 points)
* The project uses one additional library not covered in class related to data engineering. (10 points)
  + Maybe <https://seaborn.pydata.org/tutorial/introduction.html>
  + Used Zipfile library to extract the zip files into Resources Folder – Reference:  
    <https://realpython.com/python-zipfile/>
* The project includes a method for reading data from the database and displaying it for future use, such as: (10 points)
  + Pandas DataFrame
  + Flask API with JSON output
* The GitHub repo has a README.md that includes the following: (10 points)
  + An overview of the project and its purpose
  + Instructions on how to use and interact with the project
  + At least one paragraph summarizing efforts for ethical considerations made in the project
  + References for the data source(s)
  + References for any code used that is not your own

**Both Track Requirements**

**Group Presentation (25 points)**

* All group members speak during the presentation. (5 points)
* The content is relevant to the project. (5 points)
* The presentation maintains audience interest. (5 points)
* Content, transitions, and conclusions flow smoothly within any time restrictions. (10 points)