# DAE Work Sample Instructions

## Scenario:

A new data asset has been acquired from an external vendor. This dataset is untested and may or may not be ready for analysis. You are tasked with doing initial data exploration on the data and are to assist a data scientist with preparing the data for modeling.

**NOTE**: This is all fabricated data. Any similarity to real data is coincidental. The provided files are only to be used for the purpose of this exercise. Do not post or share these files or your work.

Before Starting:

* Steps 1 and 2 should be done entirely in a Jupyter notebook with code and output together.
  + Code comments are encouraged.
  + All tables can join with a one-to-one relationship.
  + The output dataset should only include cleaned results. Verify that the data contents are expected, and account for any anomalies in the output dataset.
  + All reasoning should be documented in a markdown cell.
  + Questions should be answered in a markdown cell below the code used to justify the answer.
* In Step 2, rounding in any form is discouraged
* For Step 3 you may use a text editor of your choice.
* Write-ups and communications should be formatted for easy understanding by the reader (complete sentences, paragraphs, etc.). File formats can include .txt, .doc, or .docx.
* Any reasoning should be documented along with your answers.
* All necessary components to complete this exercise are contained in this docker image: jupyter/all-spark-notebook:x86\_64-spark-3.5.0

## Instructions:

### Step 1: Data Engineering

* In a Jupyter notebook titled "Data Engineering", please combine the three provided CSV files into one data frame, using either Spark or Pandas.

The goal is to create one dataset that can be used for analysis and for model training.

* Be sure to clean up any artifacts that may persist from importing CSV files.
* Provide the schema of your final output along with a record count in the last

cell(s) of your Jupyter notebook.

* Save your final results in parquet format, to be used in the remaining steps.

### Step 2: Data Analysis

For the next section, you are tasked with analyzing the data. You may need to document your observations to answer the questions below.

In a Jupyter notebook titled "Data Analysis", answer the following questions. Answers should be given for active households:

1. What is the average number of cars per household?
2. How many cars are there by model year?
3. How many cars are there by make?
4. Which cars are the safest? What variables did you consider to define “safe”?
5. Which states have the largest households (defined as number of customers in a household)?
6. How many active households are there as of 1/1/2021?
7. What is the average age of customers?
8. How much does age vary by region?
9. Which age group has the most expensive claims?

### Step 3: Training Data Preparation

In a document titled "Training Prep", answer the following questions as they related to your data research in steps 1 and 2. You may use any text editor of your choice.

1. Are there any insights or interesting findings in the data that would be important to share with your data scientist partner?
2. What strategy did you use for dealing with the missing and duplicate data?
3. Thinking about this from an insurance standpoint, what additional features would you like to add to this data?
4. What features (if any) would you recommend removing from the final data set? Why?

### Step 4: Submission

Please submit the following files back to us. **Do not clear the output** in your Jupyter notebook files.

* The **Data Engineering.ipynb** notebook as described in Step 1.
* The **Data Analysis.ipynb** notebook as described in Step 2.
* The **Training Prep** document as described in Step 3.

Include all requested files in a single .zip file, along with a list of attachments so we can verify that we have received everything. All code must be submitted as an email attachment.

Finally, for your submission to be successfully received:

* **Do not** submit your files via a shared cloud drive (e.g., Google Drive, Dropbox).
* **Do not** submit your resulting dataset, parquet file, or the original data back to us.
* **Do not** include your name in the body of the documents or in any of the file names.