Work Sample Instructions

Scenario:

A new data asset has been acquired from an external vendor. 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.

Expected Results:

- Steps 1 and 2 should be done entirely in a Jupyter notebook with code and output together.
    - Code comments are encouraged.
    - All reasoning should be documented in a markdown cell.
    - Questions should be answered in a markdown cell either above or below the code used to justify the answer.
- 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.

Instructions:

Step 1: Data Engineering

- In a Jupyter notebook titled "Data Engineering", please combine the three provided CSV files into one spark data frame.
- The goal here is to create one data set 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 will be tasked with analyzing the data, and may need to document observations to answer the questions below.

In a Jupyter notebook titled "Data Analysis", answer the following questions:

1. What is the average number of cars per household?
2. How many cars are there by age?
3. How many cars are there by make?
4. Which cars are the safest?
5. Which cars are the most dangerous?
6. How did you define “safe” versus “dangerous”?
7. Which states have the largest households?

8. What is the average age of customers?
9. How much does age vary by region?
10. Which age group has the most expensive claims?

Step 3: Training Data Preparation

In a document titled "Training Prep", answer the following questions. 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 would you recommend for dealing with missing data? Why?
3. 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.