See below and attached files for the analysis exercise. We appreciate your taking the time to complete this! To reiterate: the assessment should only take a few hours – the weekend just provides flexibility for any external obligations you might have.

# Auto Title Analysis

**Instructions:**

Complete this assignment by Monday morning (8am ET). Please e-mail any questions or clarifications that you have as you start working on this.

We have attached a zip file of a synthetic dataset of a fictional company’s auto title loan portfolio, similar to what we would receive as part of an enforcement investigation (the CSV files and SQLite databbase file are duplicative so please use whichever you prefer). Auto title loans are a short-term small-dollar financial product that the CFPB regulates. The Pew Charitable Trusts summarized the state of auto title lending in [this paper](https://www.pewtrusts.org/~/media/assets/2015/03/autotitleloansreport.pdf) in 2015.

Write a reaction to what you find in exploring the dataset, using whatever tools you prefer.  Tailor your analysis to your experience and the role you might have on the team, and what you first explore when navigating a novel information source.

Please respond with code, documentation, and written answers to the questions below, by either:

(a) attaching your response to an email (the size of the email must be less than 20MB in order to be received), or

(b) uploading your response to a private Github repository (provide access to username \_\_\_\_).

Ensure all analysis code and process work is well-documented (and/or self-documenting).

**All candidates answer these:**

1. What’s the structure of the dataset? What are the foreign keys? Are there uniqueness issues that might cause problems in analysis? How would you ideally structure this and optimize it for analysis?
2. Please provide examples of code that you would use to perform data quality checks on the data. What data quality issues do you see in the data?
3. What assumptions do you have to make to resolve these data quality issues?

Next, choose Option 1 **OR** Option 2, and answer the questions under that option.

**Option 1 - Analytics:**

1. What period of time does the dataset encompass?
2. How many customers got loans from stores in states that they did not live in? From the auto title paper, why might customers borrow from out of state?
3. What correlations do you see between attributes about the customer and attributes of the loan that was given to the customer? Please provide some hypotheses about why any correlations might exist. Provide metrics, visuals, and explanations depicting those correlations (a) for a technical audience and (b) for an audience of smart but non-technical attorneys.
4. How much did male- and female-identified customers pay for loans (either by making payments or having their cars repossessed and sold for payments)? Provide a distribution (e.g. mean, median, 25th percentile, 75th percentile, etc) of amount paid per loan by noted sex.

**Option 2 – Data Management and Engineering:**

1. Data Modeling
   1. Consider the attached data files, is there another way that you could model this data to support downstream systems and users (analytics, visualization, reporting)? Would you re-model the existing data? Extend the model with additional tables? Please provide code and/or diagrams that demonstrate your new model.
2. File Processing: Consider a data pipeline (ETL/ELT process) that needs to regularly incorporate new files from the same source as the sample files:
   1. How do you ensure that files are coming in at the expected times? In the expected formats?
   2. Assume that new files arrive bearing only new records, or records that have been updated since the last delivery. What’s you process for integrating the new data? Please pick one of the sample data files and provide the code you would use to update a corresponding table with new data.
   3. Assume that new files arrive as a complete dataset, including new records, records that have been updated, and records that have not changed at all. What’s your process for loading this data? Please pick one of the sample data files and provide the code you would use to update a corresponding table with new data.
   4. What do you do with the raw files after they have been processed by your pipeline?
   5. How do you maintain file lineage? Where the file originated, when it was loaded, statistics about the file’s contents, etc.
3. Data Quality
   1. How do you keep track of the results of your quality checks?
   2. What kind of reporting do you do based on the results of the quality checks?
   3. What do you do with the data that fails the quality checks? How do failures affect your data pipeline?
4. Data Pipeline (ETL/ELT) Processing
   1. What processes do you put in place to monitor pipeline activity?
   2. How do you ensure that pipelines have run according to schedule, within expected timeframes, and without error?
   3. What information do you track for pipeline runs? Where do you track it?
   4. What happens if the pipeline encounters and error and needs to be re-run?