Purpose
The team has asked that you prepare a 20 to 30-minute presentation on one of the topics from below. The purpose of this exercise is to demonstrate your ability to draw insights from data, put insights in business-friendly format and confirm coding knowledge. These topics are similar in nature to projects we run. Please make sure that your presentation is accessible to a general technical audience.

Data
Choose any of the following datasets/exercises:

- Movie Revenue Prediction

Your client is a movie studio and they need to be able to predict movie revenue in order to greenlight the project and assign a budget to it. Most of the data is comprised of categorical variables. While the budget for the movie is known in the dataset it is often an unknown variable during the greenlighting process.

  - Data [CSV]

- Article SPAM Classifier

We send news articles to our researchers to stay up-to-date on technology. Lately our news feeds have been inundated with spam (pure advertisement) articles. We want to identify and eliminate these articles from informative ones. How well can we classify articles as spam or valid?
We webscraped these sources for you but you are welcome to scrape your own if you need to improve results:

  - List of sites already blacklisted [JSON]
  - Valid news articles [JSON]

- Sales Forecasting

Build and evaluate models to predict national retail store sales for each store and department. Our sales are very seasonal,

and we make our money during holidays like Super Bowl, Labor Day, Thanksgiving, and Christmas.
    - Sales by Store & Department [CSV]

- Recommendation Engine

Your client has asked you to build a recommendation engine based on this data. The goal is to recommend to customers products that they are likely to purchase in each order of the test set, given the order_id field and the data on the order contained in the other files in the folder. The "order_products__prior.csv" file contains information on orders that clients have previously placed. The "order_products__train_cap.csv" (train set) and "order_products__test_cap.csv" (test set) contain new orders
    - Dataset [Google Drive]
    - Data dictionary [GitHub]

Prior to your interview, please email your recruiter a presentation of your solution (PowerPoint, Google Slides, or pdf). This should contain:

1. Description of the problem; state what you are solving/ analyzing
2. Presentation of insights/conclusion you generate
3. Relevant descriptive statistics (charts, graphs, etc.)
4. Specification of predictive model (mathematical formulation)
5. Relevant model diagnostics
6. Model interpretation (what do the coefficients mean, how do you use them?)
7. Please specify language, packages and libraries used to develop your solution

Evaluation Criteria

- Presentation on analysis conducted that covers business outcomes and statistical methodologies

- Preform exploratory data analysis to gather starting insights and conclusions
- Selection of ML/Predictive modeling technique(s) & feature extraction
- Knowledge with data ingestion tools/languages
- Ability to conduct appropriate data cleansing if any
- Ability to code in open source languages

Rules

- Use any open source language of your choice
- Solution you provide should be your own. Reference any material desired.
- Be prepared to discuss your code in-depth (what it does, how it does it etc.)
- Utilize any statistical or ML technique(s) you deem relevant. For each technique that you use, be prepared to talk about model diagnostics, results, and mathematics behind your technique(s).
- No time limit on developing your solution. Let us know when you are ready.