INTRODUCTION

Welcome to Hollywood! You've been hired as the newest data scientist for a consulting firm that specializes in entertainment. A client, Mo Ops Studios, would like you to assess historical box office trends and provide recommendations on the type of movies that they should greenlight. The studio is interested in expanding their portfolio of films that garner critical or popular acclaim, as well as maximize buzz on social media.

A team member at the consulting firm has provided you with a sample of movies from the past 100 years. This data set includes information across several related dimensions, including film profiles and success measures.

DATA SOURCES

1. '**movies.db**' is a sqlite database with one table of ~4,700 rows and 33 dimensions.
2. '**Movie_Data_Dictionary.xlsx**' is a data dictionary file which includes descriptions for the 33 dimensions. As this is a sample data set, incomplete records may exist.
3. '**OMDb API**' is a RESTful web service to obtain movie information.
   Web address: http://www.omdbapi.com/
   API key: **9d9a07af**
   Example request: http://www.omdbapi.com/?i=tt3896198&apikey=9d9a07af

IMPORTANT: because querying the API may take up to a couple of hours, we recommend starting Step 1 under "PROCESS" as soon as possible upon receiving these instructions.

PROCESS

1. Use the provided key to query the OMDb API to obtain at least one additional popularity or critic rating metric to ultimately enrich the data set. Feel free to gather any additional data you feel would be useful to you. Create this enrichment data as a separate table in movies.db.

2. Using your preferred open source languages (SQL, R, Python, Java, etc.), write a script to connect to the database, join the tables, and stage the data for further analysis.

3. Perform any necessary data cleansing, as well as some simple exploratory analysis -- whatever steps you would normally take in order to get a sense of what is in a new data set.

4. Define a target that you believe best captures Mo Ops Studio's strategic objective. This can be a single metric or a composite. There is no right or wrong answer, but be prepared to explain why you chose the metric or metrics you did.

5. Split the dataset into train and test, and build a model to predict the target variable. Evaluate the model for accuracy. This is messy, real world data, so don't worry if, after reasonable effort, your accuracy is low. **We are more interested in understanding your approach and process, rather than judging the highest accuracy achieved.**

6. Write your model's predictions for each movie back to the sqlite database (either as a column in the original table, or as a separate table)

DELIVERABLES

Create a zip file of the following, and email them to the team (addresses provided below)

- An export of the database files
- A text file containing any SQL queries you used outside of the analytic environment
- The data cleansing and modeling notebook/script, with thorough comments
- A one-page brief or short slide deck (typical length is ~5-10 slides) explaining your methodology and results. This may include any relevant charts that you generated during the exploratory analysis or modeling process.

IMPORTANT: Assume the audience for your slides/write-up includes non-technical stakeholders. Assume the audience for your code includes other data scientists on your team who may have to maintain or modify your code in the future.

After we receive your case study, if we move forward with next steps you will be asked to present an overview of your work to a panel. Come prepared to discuss:

- Data preparation steps
- Results of the analysis and the process you followed
- The approach you would take to automate extracting the data and running the model on a weekly cadence

You will be contacted separately by a recruiter from the Red Hat Talent Acquisition team if we schedule this session.

QUESTIONS

If you have any questions, please reach out to Donald Chesworth (dcheswor@redhat.com). We hope you enjoy your work on this case study!