Walt Disney
Direct-to-Consumer & International
TECHNOLOGY

Thank you very much for your interest in our Data Analyst/Sr. Data Analyst roles under the Disney Direct to Consumer & International - Data Platforms group. Our team is responsible for analytics across all the Disney Media Segments including ESPN, ABC, Disney Movies, 21st Century Fox, HotStar and many more. Due to the sheer volume & complexity of data collected across segments, our Analyst roles require extensive knowledge in various techniques such as SQL, Big Data Tools (Hive, Spark-QL, PySpark), Python/R or similar scripting languages. The take home exercise below allows us to evaluate your ability to pre-process and analyze an unfamiliar dataset, answer business questions related to the data using statistical and/or machine learning techniques, and present your findings in an explainable report or presentation.

Introduction:

You will be working with a well-known open dataset which contains basic information about mobile applications and its reviews. You will be solving problems given below based on the dataset. We request that you use a Jupyter notebook throughout the process and send us **two** files:

1. Export Jupyter .ipynb file
2. A PDF export of the same file with all code, results and visualizations. This will allow us to evaluate the submission without re-running the notebook.

Helpful Tips Before You Get Started:

- If you've never worked with Jupyter Notebooks before, take a look here: https://jupyter.readthedocs.io/en/latest/install.html
  - You can code in Python/R and each cell is capable of executing a code snippet independently. The variables and outputs are maintained throughout your work session
- Helpful python libraries to install for the exercise might include:
  - Pandas
  - Numpy
  - Matplotlib
  - Seaborn
  - Scikit-learn

Instructions

Download the dataset from https://www.dropbox.com/s/7smc0v57bgszls0/ESPN_take_home.zip?dl=0

There are two CSV files included in the dataset:
app_info.csv - details of the applications on Google Play. There are 13 features that describe a given app

Column	**Description**
App	Application name
Category	Category the app belongs to
Rating	Overall user rating of the app (as when scraped)
Reviews	Number of user reviews for the app (as when scraped)
Size	Size of the app (as when scraped)
Installs	Number of user downloads/installs for the app (as when scraped)
Type	Paid or Free
Price	Price of the app (as when scraped)
Content Rating	Age group the app is targeted at - Children / Mature 21+ / Adult
Genres	An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, Family genres.
Last Updated	Date when the app was last updated on Play Store (as when scraped)
Current Ver	Current version of the app available on Play Store (as when scraped)
Android Ver	Min required Android version (as when scraped)

app_reviews.csv - This file contains the first 'most relevant' 100 reviews for each app. Each review text/comment has been pre-processed and attributed with 3 new features - Sentiment, Sentiment Polarity and Sentiment Subjectivity.

Column	**Description**
App	Name of app
Translated_Review	User review (Preprocessed and translated to English)
Sentiment	Positive/Negative/Neutral (Preprocessed)
Sentiment_Polarity	Sentiment polarity score
Sentiment_Subjectivity	Sentiment subjectivity score

Question 1:
Open up the data, do some initial exploratory data analysis (EDA) and gather some business insights on this data. Are there any problems with how data collected that would affect your ability to do proper analysis on this dataset? Please provide your findings in the jupyter notebook itself.

Question 2:
Write a SQL statement to return the percentage of positive, negative, and neutral sentiment reviews for each application. Show your results in a visualization of your choice (ie. matplotlib, seaborn, D3.js, etc).

Question 3:
In a single SQL query, find the *ratings* of the top 10 applications which have the most installs.

Question 4:
If a company is looking to develop the next top trending application on the Google Play Store, what kind of app should they focus on building? Why? (ie. does number of installs correlate with a higher rating? Or perhaps number of reviews, or overall sentiment?) Present the data findings to back your claim. Use notebook for both results and explanation.

Question 5:
Use dataframes, for instance, pandas: Find the top rated application name and its' rating for ***each category***. If there are ties, return all tied applications. (ie: the top application in the category ART_AND_DESIGN is "Spring flowers theme couleurs d t space" with a rating of 5.0.)