Data Science Intern Test

- **Problem Statement: [Test Duration – 2.5 Hrs.+1 Hr(Bonus task)]**

There are 2 datasets given here, which contains appropriate Site Data. Analyze all the data carefully and complete the below mentioned tasks. Please submit all the code written and it is **Mandatory to do this test in Python.**

- Necessary formulas required for calculation are included in task specifications.
- The test file for bonus task is given separately as DFP_prediction.csv and GA_prediction.csv
- **Don't waste too much time in understanding this variables. We don't expect you to know in and out of digital marketing right now, so if you don't get this terms, relax and move further.**

  Here is the brief description about the datasets :

1. **Google Ad Manager (DFP.csv) data :-**

DATE	Date
DAY	Day of the week
AD_UNIT_NAME	Ad_Unit created by seller
ORDER_NAME	Order name created by Seller.
ADVERTISER_NAME	Buyer name – who bought our Ad inventory.
LINE_ITEM_NAME	Line Item name created by seller.
eCPM	Revenue per thousand impressions
Tags_served	The no. of times Ad Manager tag is called. (to server an Ad.)
Impressions	The no. of items any ad is showed.
Clicks	No. clicks on total impression
CTR	Click Through Rate – no. of clicks on our Ads / total impressions
Revenue	Revenue
Actual_eCPM	True eCPM

2. **Google Analytics (GA.csv) data :-**

date	Date
channelGrouping	Data segregated by traffic source
users	Total users visited.
newUsers	Total new users who visited your site.
Sessions	a group of interactions one user takes within a given time frame on your website
percentNewSessions	newUsers * 100 / Sessions
bounces	a single-page session on your site
bounceRate	percentage of single page visits (or web sessions)
sessionDuration	Total Session Duration
avgSessionDuration	average length of a Google Analytics session
pageviews	Total no. of pageviews of your site.
pageviewsPerSession	Average pageviews in 1 session

Tasks

1. **DFP data (DFP.csv) :**
   - **1.1.** From Ad_unit_name create new columns for story and position, also separate the data amp / non amp wise. If the data point belongs to amp than ad unit name contains 'amp' string. Sample :
     CarToq_ad_first_story_pos_top (122380182) -> story = first ,position = top, Nonamp
     amp-cartoq-bottom (21684306640) -> story = '', position = bottom, Amp
   - **1.2.** Map the DAY data with day of the week name.
     e.g. (where map 1 -> Monday upto 7 ->Sunday.)
   - **1.3.** [Revenue = Total Impression * eCPM], but here eCPM for some data is incorrect, merge the Actual_eCPM data from the other sheet (Actual_eCPM.csv) such that new column for Actual_eCPM is created. Now create another column "Actual_Revenue" which is defined as [Actual_Revenue = Total Impression * Actual_eCPM] (if not provided take the given revenue as Actual_Revenue, and fill the remaining "Actual_Revenue" column data with it.) After completing this 3 task save the sheet as "DFP_solution.csv"
   - **1.4.** Find the best performing Ad position in terms of eCPM and revenue differently in amp and non-amp case. (hint : while merging the data keep in mind that [eCPM = Actual Revenue * 1000 / Total Impressions] and [CTR = clicks / Total impressions] , aggregating directly won't help. Same applies for GA data.) Submit this data as "Adpos.csv"
   - **1.5.** Also find top 5 Advertiser and Line Item name separately month wise (combining both Amp and Non-Amp)for last 3 months, and save this in separate sheet as "Top5.csv"
   - **1.6.** Create a stacked bar plot showing contribution of "Google Adx", "Google Adsense" (this are Advertiser_Name) in Total Actual Revenue for the last 3 months. (You can plot this graph with any tool you want.)

2. Analyze this data in as many ways possible ways, including multivariate analysis, to generate insights for predictive modeling. (You can do this analysis with either Python or Excel)
   - Suggestions –[Correlation, Covariance, ANOVA, Regression analysis, Hypothesis testing: Student's t test , chi-square test (Generate Null and Alternate hypothesis and find the significant relation)].
   - Here we want to test your familiarity with different Statics and Modeling concepts. So show as much as possible.

3. **Bonus Task**: Predict the Aggregate eCPM (combining (aggregating on) both Amp and Non-Amp) for 1 month.
   Training Data – DFP.csv , GA.csv (7 months data : Jan 2018 – July 2018)
   Test Data - DFP_prediction.csv , GA_prediction.csv (predicting for the next 1 month)
   Submission format - sub.csv
   Evaluation Metric - RMSE

Note-

- All the Code Should be in Python. (All other submission will be disqualified)
- Results and answers datasets for this can be given in **Excel** or **CSV.**
- Please also write a small readme file containing brief description about your approaches and explaining your thought process (specially include task 2 and Bonus task).
- Submit your response with Code and required files to hr@mindworksglobal.com with '**Data Science Intern Submission**' as subject line (Other submission may not be considered)
- Also, for any query regarding the test you can write to the harsh@cartoq.com