1. **Introduction**

This assignment broadly deals with location-based mobile marketing. You have data from a location-based marketing agency which handles geo-fencing campaigns on behalf of advertisers. Due to the very large volume of data, you are given a random sample for two campaigns of a single advertiser –

AMC Theaters. The advertising impressions are inserted into the mobile app being used on the device. The data include the following elements: impression size (e.g., 320x50 pixels), app category (e.g., IAB1), app review volume and valence, device OS (e.g., iOS), geo-fence lat/long coordinates, mobile device lat/long coordinates, and click outcome (0 or 1). The column names are self-explanatory, although we have provided a data dictionary file on Canvas.

1. **Analysis**

**Data Processing**

* 1. Create dummy variable *imp\_large* for the large impression:

*imp\_large* =

* 1. Create dummy variables *cat\_entertainment*, *cat\_social* and *cat\_tech* for app categories:

*cat\_entertainment* =

*cat\_social* =

*cat\_tech* =

* 1. Create dummy variable *os\_ios* for iOS devices:

*os\_ios* =

* 1. Create variable distance using Harvesine formula to calculate the distance for a pair of latitude/longitude coordinates.

*Distance (in kilometers) =* 6371 \* acos( cos( radians(LATITUDE1) ) \* cos( radians( LATITUDE2 ) ) \* cos( radians( LONGITUDE1 ) - radians(LONGITUDE2) ) + sin( radians(LATITUDE1) ) \* sin( radians( LATITUDE2 ) ) )

* 1. Create variable *distance\_squared* by squaring variable *distance*
  2. Create variable *ln\_app\_review\_vol* by taking natural log of *app\_review\_vol*

**Descriptive Statistics**

1. Summarize the data by calculating the summary statistics (i.e., mean, median, std. dev., minimum and maximum) for *didclick*, *distance*, *imp\_large*, *cat\_entertainment*, *cat\_social*, *cat\_tech*, *os\_ios*, *ln\_app\_review\_vol* and *app\_review\_val*.
2. Report the correlations among the above variables.
3. Plot the relationship of distance (x-axis) and click-through-rate (y-axis), and any other pairs of variables of interest.

**Logistics Regression**

1. Specify the following Logistic regression model:

Dependent variable: *didclick*

Independent variables: *distance*, *distance\_squared*, *imp\_large*, *cat\_entertainment*, *cat\_social*, *cat\_tech*, *os\_ios*, *ln\_app\_review\_vol* and *app\_review\_val*.

1. Estimate the model in R (using the glm function) and report coefficients and p-value of the estimates.
2. Discuss your findings and their implications, limiting your answer to a page or so.