**Hands on challenge 8 - Strategy**

Challenge 1

* HD Lo data set
  + Create a R markdown file to answer the questions below
* Data Dictionary
  + Areaname – city / town name
  + County – zipcode
  + State
  + Lcount – count of Lowes stores in the town/city
  + Hdcount – count of Home Depot in the town/city
  + Pop\_2000 – population in 2000
  + Pop\_2010 – population in 2010
  + Income\_2000 – avg income in 2000
  + Income\_2010 – avg income in 2010
  + pct\_U18\_2000 – percent under 18 in 2000
  + pct\_U18\_2010 – percent under 18 in 2010
  + Pct\_college\_2000 – percent in college per town in 2000
  + Pct\_college\_2010 – percent in college per town in 2010
  + Ownhome\_2000 – percent owned home in 2000
  + Ownhome\_2010 – percent owned home in 2010
  + Density\_2000 – percent density per town in 2000
  + Density\_2010 – percent density per town in 2010
  + Pct\_white\_2000 – percent of Caucasian in town in 2000
  + Pct\_white\_2010 – percent of Caucasian in town in 2010
  + Pct\_black\_2000 – percent African American in 2000
  + Pct\_black\_2010 – percent African American in 2010

**Questions:**

* Using the HDLData.csv file in the resources section of Piazza perform the following

1. Create dummy variables to identify if HomeDepot or Lowes is present in that county
2. Create a linear regression model to identify the correlations among the variables.
3. How do these box stores make their decisions about where to locate their stores?
4. Do they follow a certain criteria of demographics to choose their locations?
5. How are the chains similar in their decision making ? How are they different?
6. What kind of customers does Lowes target? What customers does Home Depot target? (Use a correlation plot)
7. What are the top 5 towns / cities that can be predicted as potential candidates for new locations for both Lowes and Home Depot?
   1. Use a generalized linear model (glm) binomial link logit prediction model to predict the top 5 potential candidates
      1. 2 Prediction models (1 for HD and 1 for Lowes) logic:
         1. log(pop\_2010) + log(density\_2010 + 1) + top 6 strongest variables
         2. We ‘log’ pop\_2010 and desnisty\_2010 because there is a lot of variance in these features
         3. family=binomial(link=logit), data=HDLO\_data