

**Course:**

Intro to Data Science – DS-GA-1001 /

Data Mining for Business Analytics - INFO-GB.3336.11

Fall 2014

**Instructor:**

Brian Dalessandro

**Homework 2 – Due 10/8/2014 at 5 pm**

1. **Download the data set ads\_dataset.txt and load it into a Python data frame called ‘ads.’**
2. **Write a Python function called ‘getDfSummary()’ that does the following:**
3. Takes as input a data frame
4. For each variable in the data frame:
   1. Counts the number of missing/NA/Null values
   2. For the non-missing values
      1. Computes the number of distinct values
      2. If numeric, computes:
         1. mean, max, min, stdev, quantiles=[25%,50%,75%]
5. Loads all of the summary information into a new data frame. Each row of the data frame should be a feature name, and the index of the data frame should be the features.
6. Returns the data frame containing all of the summary information

Hints:

The pandas ‘describe’ method returns a useful series of values that can be used here.

1. **Copy and paste your getDfSummary() function and put it here:**
2. **Do a time analysis on your getDfSummary() function. Copy and paste the results here.**

Hint: %timeit getDfSummary(ads)

1. **Using the results returned from getDfSummary, which fields, if any contain missing/Nan/null values?**
2. **For the fields with missing values, does it look like the data is missing at random?**

Hint: you can start by running ads.groupby([c]).size(), where c is the name of the field that you think has missing values.

1. **Are there any other fields that correlate perfectly, or predict that the data is missing? If missing, what should the data value be?**

Hint: create another dataframe that has just the records with a missing value.

Get a summary of this frame using getDfSummary() and compare the differences. Do some feature distributions change dramatically?

1. **Which variables are binary?**

**9. Generate a correlation matrix for ads dataframe. Is there any redundancy in the data? Are there any features that aren’t needed?**