**Midterm Project Report**

**Data Analytics (Spring 2016)**

**Home Depot Product Search Relevance**

**Team Members:**

Venkatesh Duvvuri ([VED14@pitt.edu](mailto:VED14@pitt.edu))

Gopi Tata ([GKT3@pitt.edu](mailto:GKT3@pitt.edu))

Haifa Alnasser ([HIA11@pitt.edu](mailto:HIA11@pitt.edu))

**Introduction:**

Home Depot Product Search Relevance is a Kaggle Competition to predict relevance score for the provided combinations of products and search terms

The training dataset contains the search terms and the products retrieved for the corresponding search terms with a relevance score. The relevance score is a number between 1 and 3 with 3 being the most relevant and 1 the least.

Each pair (Search Term and Product Retrieved) was evaluated by at least 3 Human raters and the average was recorded in the training dataset. In addition to this, there are 3 additional things mentioned in the problem statement:

1. The relevance\_instructions.docx contains the instructions that were provided to the human raters at the time of evaluation.
2. The attributes.csv file contains the attributes of each product such as the product\_uid, Product Name and a brief description about it for which the human raters didn’t have access to.
3. While the competition does not have access to the product images, the raters had access to these.

The Challenge is to predict the relevance for each product - search term pair in the test set.

**Initial Observations:**

1. There are 74067 rows in train.csv and 166693 rows in test.csv and it can be inferred that there are some overlapping observations in both the csv files.
2. The attributes.csv file has 2044803 rows and there 155 rows without product\_uid and these have to be removed.
3. There are 124428 unique products in the product\_descriptions.csv
4. Further Drilling Down, it can be observed that there are 86264 unique product\_uid in attributes.csv, 54667 unique product\_uid in train.csv and 97460 in test.csv

**R Script :**

train <- read.csv("train.csv")

length(unique(train$product\_uid))

test <- read.csv("test.csv")

attributes <- read.csv("attributes.csv")

a <- unique(attributes$product\_uid)

length(a)

products <- read.csv("product\_descriptions.csv")

un <- unique(products$product\_uid)

length(un)

p <- attributes[is.na(attributes$product\_uid),]

p

hist(train$relevance)

hist(train$product\_uid)  
  
**Analyzing the data:**The summary statistics of the raw train data are as follows:

id product\_uid

Min. : 2 Min. :100001

1st Qu.: 57164 1st Qu.:115129

Median :113228 Median :137334

Mean :112386 Mean :142332

3rd Qu.:168276 3rd Qu.:166884

Max. :221473 Max. :206650

**product\_title**

Lithonia Lighting All Season 4 ft. 2-Light Grey T8 Strip Fluorescent Shop Light : 21

Pressure-Treated Timber #2 Southern Yellow Pine (Common: 4 in. x 4 in. x 8 ft.; Actual: 3.56 in. x 3.56 in. x 96 in.): 21

2 in. x 4 in. x 96 in. Premium Kiln-Dried Whitewood Stud : 18

Custom Building Products VersaBond Gray 50 lb. Fortified Thin-Set Mortar : 17

Ryobi ONE+ 18-Volt Lithium-Ion Cordless Drill/Driver and Impact Driver Kit (2-Tool) : 17

Ryobi ONE+ 18-Volt Lithium-Ion Ultimate Combo Kit (6-Tool) : 17

(Other) :73956

**search\_term** **relevance**

1/2 zip wall : 16 Min. :1.000

3 WAY TOGGLE SWITCH : 16 1st Qu.:2.000

anderson windows 400 seriesimpact resistant: 16 Median :2.330

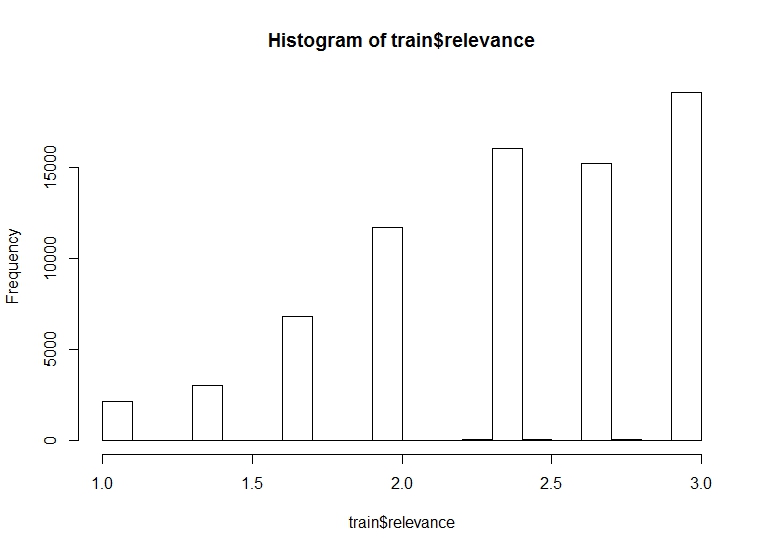
bed frames headboaed : 16 Mean :2.382

burgundy red foot stools : 16 3rd Qu.:3.000

contact paoer : 16 Max. :3.000

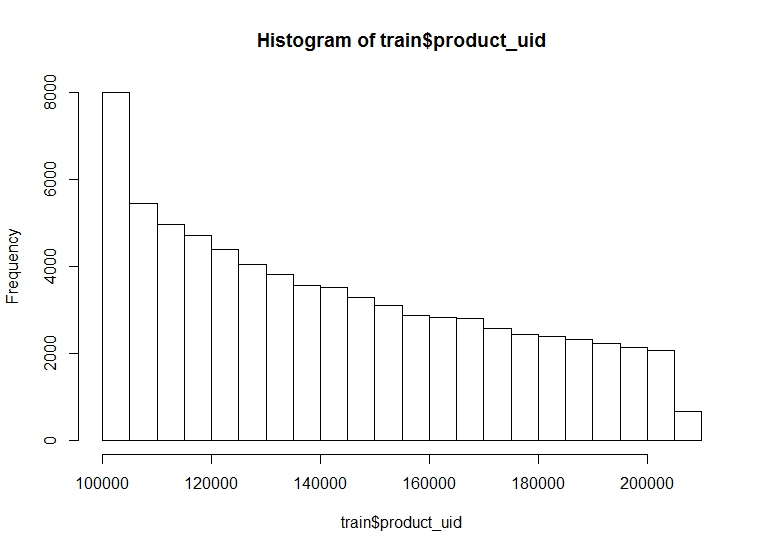
(Other) :73971

**A histogram plot for Relevance Score in the training dataset:**



From the histogram plot it can be seen that most of the search terms fall in the category of most relevant area with a relevance score of 3 and there are very few with relevance score 1.

**Histogram Plot for uid :**



In the above histogram we can see that most of the product id’s start with 100000 – 120000 which is more than 20000 product. While he frequency of products that falls in the 200000 category decreases dramatically to around 2000 products.  
 **Analytic methodology plan:**

The next Step in our project will be to analyze the Search and Corresponding Relevance score using R Packages like Random Forest. Further, we are planning to use any of the NLP (Natural Language Processing) Packages available out there to further study how the Relevance Score is affected by the Product Search terms and the product uid for the same.