METRO COLLEGE OF TECHNOLOGY  
A Python Data Science Project by Ali Nathani

**TITLE : Amazon E-COMMERCE customer reviews analysis REPORT  
1. INTRODUCTION AND METHODOLOGY:**This is a pre-crawled dataset, of 15,000 records combined between 2 data sets, taken as subset of a bigger dataset (more than 7 million fashion products) that was created by extracting data from Amazon and was created by PromptCloud's in-house web-crawling service. There were 18 columns or variables.

To Clean this Dataset, I first concatenated the main and the sample datasets as train and test.  
Using Python code I first searched for missing values in all of the columns and decided to remove the records with missing values from the average\_review\_rating column and put this data into a fresh dataset.

I also looked for missing values in the amazon\_category\_and\_sub\_category column and tried to find out if these records had product names and other information. As they contained other information, I decided not to drop them for the moment and tried to find out which category had the top rating of 5.0 out of 5 stars.

I proceeded to convert some of the columns to numeric by removing the values from strings, such as in the number\_of\_reviews, price and number\_available\_in\_stock columns.

I then created a new column with amazon\_category only and started to split the dataset into train, test and prepared the model for running.

**2. THE ANALYSIS**The main objective of the Analysis was :  
Analyses of the ratings, category and reviews

**3. Research Questions / Problem Statement**

Can the expected review rating of a product, be possibly determined from the category, manufacturer, number of reviews and the number of answered questions?

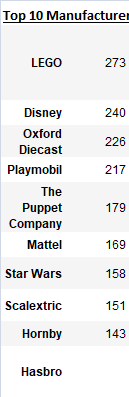
**4. Hypothesis**

The Hypothesis would be that review rating can be determined from the category, manufacturer and number of reviews.

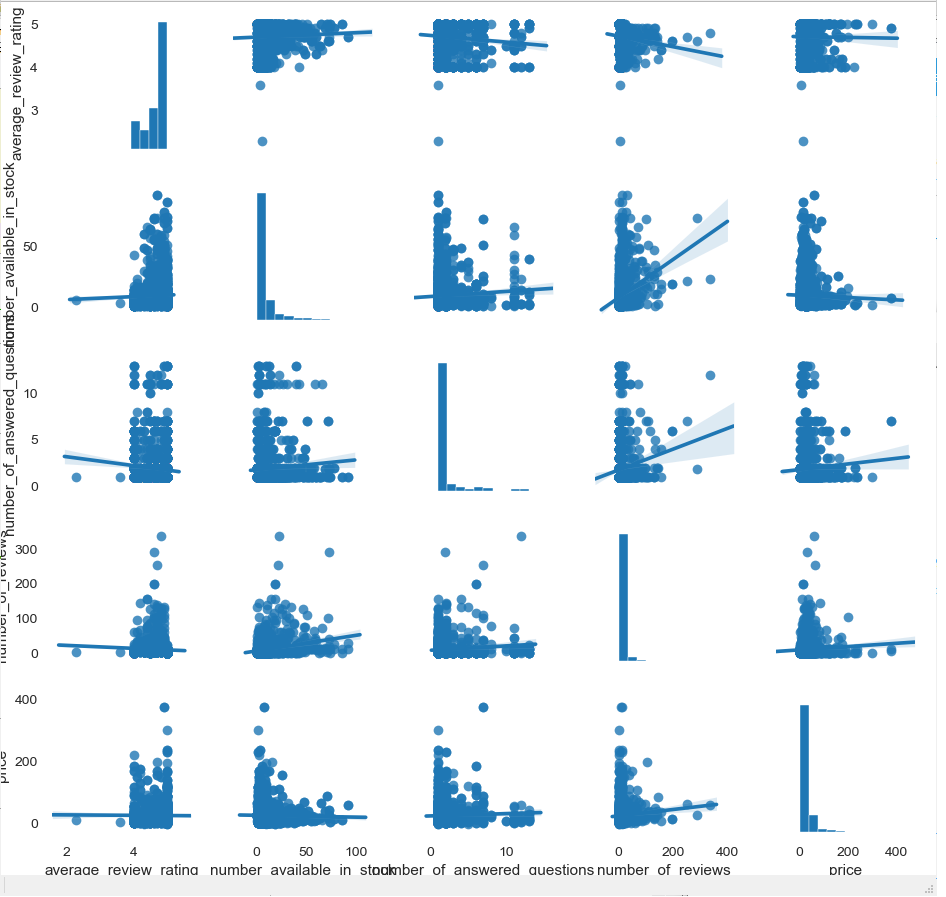
5. METHODOLOGY / DATA EXPLORATION

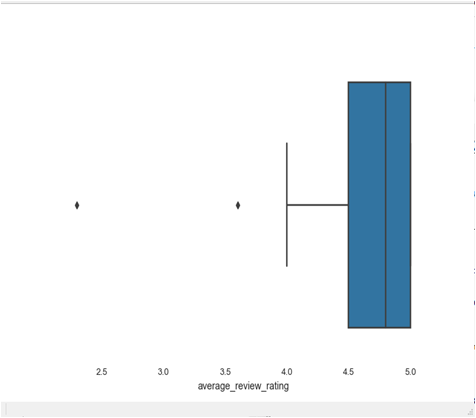
Main Steps  
The Data Step – Import /Connect the DataSet  
Data Exploration  
Data Cleaning  
Feature Engineering  
Model Building

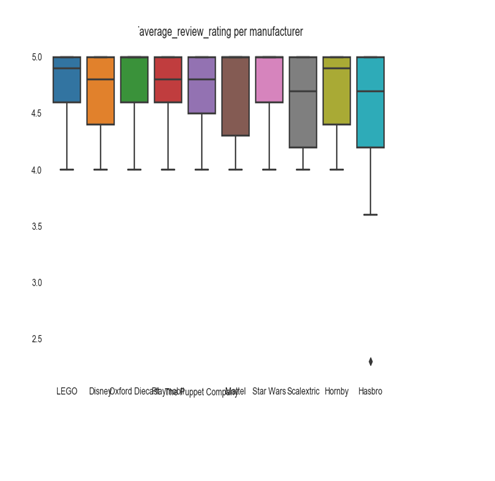
**6. Data Exploration & Data Cleaning**

**** ****

* + **The plot shows that there is a lack of correlation between the chosen variables after cleaning**

****

* + **Uni/Bi/Multi variate Analysis -** 
    - **Distribution of variables were analyzed in order to check what the values are that could be affected and also where the values need to be in order to determine features of products to get the rating:**
    - **average\_review\_rating** 
      * + **Majority of the results are between 3.5 to 5**
        + ****
      * **number\_available\_in\_stock** 
        + **Results are between 1 and 70**
      * **number\_of\_answered\_questions**
      * **Results are between 1 and 28**
    - **Number\_of\_reviews** 
      * **Results are between 1 and 802**
    - **Price**
      * **Results are between 0.6 and 447.99**
  + **Average review rating per manufacturer**

****

* + - * **above results show that the most popular manufacturers have majority of the distribution of the ratings**
  + **Fix Missing Values – the mean was used for numeric variables and mode was used for categorical variables which had a low percent of missing data and those with high missing values were dropped**
  + **Transform Variables – values were extracted from strings for variables such as price and number of reviews**
* **Feature Engineering - category was separated from the full category and sub category variable and encoded along with the balance variables**
* **Model Building**
  + **Encoded manufacturer and amazon\_category and trained and prepared the model using Backward Feature Elimination and ran a Linear Regression model**
  + **Findings –**
    - **Model was not a very accurate model**
    - **P-values were small for the manufacturer and categories so there is a high association with the average review ratings**
  + **CONCLUSION**
  + **Reject the Null Hypothesis that the category and manufacturer do not have effect on the Review rating.**

**Recommendations**

* **Categorize the new products more as a hobbies and ensure all have a category as several were missing**
* **Continue dealing with the top manufacturers and concentrate on more products from them**
* **Invest more time in testing to check correlation of product name, products description, customer questions and other reviews and how it is worded using NLP**