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

**TITLE : Amazon E-COMMERCE customer reviews analysis**

**Executive Summary**

By Ali Nathani

**Introduction**

As e-commerce grows, it is important for retailers to understand the concept of reviews and review ratings provided as feedback by customers, in order to understand repeat customer behavior.

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.

**Objective**

* To perform Analyses of the ratings, category and reviews and run a model to check possibility of predicting the rating that a product would receive.

**Research Question**

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

**Methodology** – Python programming was used to Import the data, clean it, generate any features required for analysis and to run a model to test the accuracy of prediction of the review rating.

**Data Steps** – Importing the Data, cleaning (almost 80% of the time was spent on cleaning as it was very important to have the data ready for correct analysis).

The next step was to explore the features of the data, the type of variables, the missing values and whether the values had to be changed or encoded to provide the correct analysis

**Data Analysis and Visualization** – Mostly the distribution of each of the features was studied to understand the distribution of the variables that were to be analyzed as effecting the rating.

**Summary of results:**

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

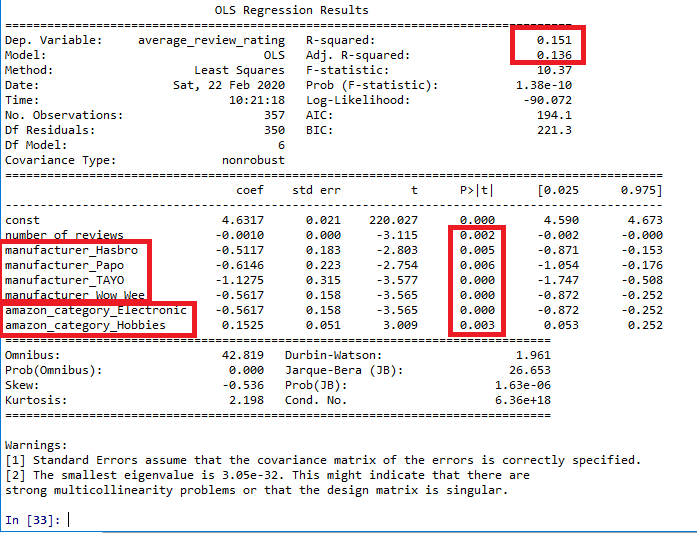
Number\_of\_reviews: Results are between 1 and 802

Price : Results are between 0.6 and 447.99

**Feature Generation** – The Category column had to be extracted and created from the existing amazon\_category\_and\_sub\_category variable and this is important to understand the general categories that Amazon uses or that the seller uses to make maximum sales

**Model Building –** Linear Regression was tested after building the model and encoding the features.

Linear Regression  
OLS Regression Results



**Findings** –

* R squared: 0.151
* Adj. R squared: 0.136
* P-value lowest only for most popular categories and manufacturers

**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

**References** –

<https://data.world/promptcloud/fashion-products-on-amazon-com/workspace/file?filename=amazon_co-ecommerce_sample.csv>