1. **Introduction**

When a user visits an e-commerce site, there are numerous qualitative and quantitative factors that influence their decision to make a purchase. Qualitative factors being their spent budget, user needs at the moment and so on. Quantitative factors include product ratings, number of reviews, presence or absence of a product video, number of photos for the product, length of product description, price of a product – to name a few.

This capstone aims to understand which quantitative variables affects a user’s purchase decision or conversion (CVR). Consequent application of identifying the contributing factors can enable an e-commerce site to maximize profits by utilizing their user behavior data and as a result, display the most relevant information at all times.

1. **Data Acquisition**

Two primary data sources will be used throughout the project. The time range will be year-to-date (YTD), in order to focus on the most recent behavior.

1. Web analytics (WA) – this dataset contains conversion metrics per a unique product
2. Content management system (CMS) – this dataset contains per product ID, user ratings, number of reviews, word count of content description, number of photos, videos on the site for a product, product classification
3. **Data Wrangling and Cleaning**
4. From initial data exploration, we discovered that in the CMS dataset, few products have word counts of description that are in draft as well as that of live product descriptions. Hence, the first step would be filter out the draft description counts – if a live count exists, then use word count of live descriptions
5. Since we need the conversion by product ID from WA dataset and the product data such as # of photos, videos, reviews, ratings mix, word count from CMS, One of the first things to do in order to make the data analysis-ready is to merge the two dataset into a single dataframe.
6. **Preliminary Exploration**

Based on early data exploration, we see that products with one review is better than a product with no reviews at all – products with even a single user review have better conversion than those without any review.

1. **Exploration and Analysis Approach**

Based on the goal of the analysis, the dependent variable is CVR or conversion. We aim to find the factors that drive the CVR. Hence, we will begin by filtering the most relevant factors from the dataset that affect the product CVR by doing one-variable linear regression. Next, we will look at the dependent variable and all the identified independent variables to identify the variable set that leads to CVR. Further, we will start segmenting the results to visualize the change in CVR by product category and by customer country.

1. One-variable linear regression

First step to understand the influencing factors would be to go plot the each attribute against CVR and retain the ones that are more proportional or inversely proportional to CVR.

1. Segmentation by product category

Category segmentation will allow us to see if certain variables are more important for a given category of a product. For eg: are videos more important for pricier products?

1. Segmentation by customer location  
   Slicing the CVR and independent variables by customer country will help determine whether users from different geographies have more affinity towards reviews or photos, or does every user care about the same information presented on the site.
2. **Limitations**

This analysis limits the view of product conversion to measureable variables. There are qualitative signals that this approach does not take into consideration which in fact play a very important role in purchase decision making.

Even if all the product information is adequately laid-out on a website, a user may not buy it because their plans changed or the product isn’t a right fit for them at the given time. Such conditions are difficult to measure and this limits our understanding of why a user abandons buying the product as well as the ability to account for such conditions in the analysis.

In addition to the approach listed above, there are further more slices that we can apply to the dataset in order to determine what factors affect CVR. Does CVR change when we segment the user based on the traffic sources they come from – essentially is the expectation of information for a given product change based on where they come from and how is the product marketed in the given source.

This project serves as an application and exploration exercise, as well as stimulates the thought process around solving the problem using data science concepts. Within the time limits of the course, the project taps only a few key areas.