Questions:

1) Give us your suggestions on how we could make our data set better / more useful.

- The data can be enriched by combining it with transactional data. Few attributes that could be derived from that intersection include, prominent age group/location of the users contributing to sales. These features can be helpful to fine tune the ranking based on user profile
- Information regarding the brands, such as its stock price and/or twitter sentiment for today, last week, last month can be added to the data. This data could be helpful to adjust the rank of products, in the event of any damage to brand reputation or announcement of new line of clothes/accessories

2) With the given dataset, can you come up with a scientific approach and model for our ranking?

- The important step to solving this modelling problem is to carefully derive a value measure i.e the importance of a product
- This importance can be measured with various parameters
  - Revenue generated in sales
  - Conversion of view counts to sales
  - Products with least cancellations/rejected cases
- The importance of a product could be a weighted sum of the above mentioned features. Similar importance values could be computed at brand/category level which can be propagated to the respective products
- This problem could be best described as a regression analysis problem, where the importance measure (dependent variable) is affected by the changes in attributes of the dataset (independent variables)
- Efforts also to be put in deriving new features to the data such as
  - Discount percentage of the product
  - Rejection/cancellation percentage
  - Views to sales conversion rate
- Remove highly correlated attributes to reduce the noise in modeling
- Convert categorical data to numerical data to facilitate modeling
- Finally, model the importance metric using generalized linear models. GLM will help to not restrict the modeling to Gaussian distribution

3) How would you test, train, and evaluate your model?

- Evaluating the model should focus on maximizing the accuracy while not over fitting to training data
- Accuracy can be measure across precision, recall or f-measure etc
- Data can be split into 70%-15%-15 % for training, validation and testing sets.
- Cross validation techniques will be used to tune the parameters of the model. The model with highest level of accuracy shall be selected as the output model