Grade: 96%

**customer feedback**

**analysis for *rent the runway***

Team member:

Liya Zhou

Wenyi Xiong

Yibo Feng

**Objective**

Our main purpose of the project is to figure out how the clothes brand, *Rent the Runway*,should focus on utilizing the factors that can affect customer feedback and predicting which kinds of clothes should be promoted to a unique customer. To complete that object, the first thing we need to do is to extract the columns that can illustrate customer feedback. Then based on different data types of those columns, targeted data processing will be given.

**Data Set Description**

1. **Overview**

The dataset is collected from RentTheRunWay by Rishabh (2018). We downloaded it from following link: <https://www.kaggle.com/rmisra/clothing-fit-dataset-for-size-recommendation?select=renttherunway_final_data.json.>

The format of the raw data is json file. The dataset provides some types of information, such as ratings and reviews from customers, fit feedback from customers after they buy clothes. Besides, the dataset also contains product measurements and category information. This dataset is suitable to analyze review sentiment and predict ratings while recommending fit size.

1. **Number of rows and cols**

The data set includes 15 columns and 192544 rows. Each raw in the data records one transaction. Besides, there are 105508 customers who completed transactions and 5850 kinds of product are included.

1. **Sample predictors**

As we may used “rating” and “fit” as dependent variables, following variables are sample of predictors. After text sentiment analysis, sentiment of text may also be one of the predictors to predict ratings.

* weight: weight measurement of customer
* rented for: purpose clothing was rented for
* body type: body type of customer
* age: age of the customer
* category: the category of the product

1. **Others**
2. The format of the height is “5’8’’”, which is foot with inch. This type of data isn’t able to be used directly in the analysis.
3. There are emojis in the text of reviews, which may contributes to the sentiment analysis.
4. The whole data set contains transactions from same customers and transactions of the same product from different customers.

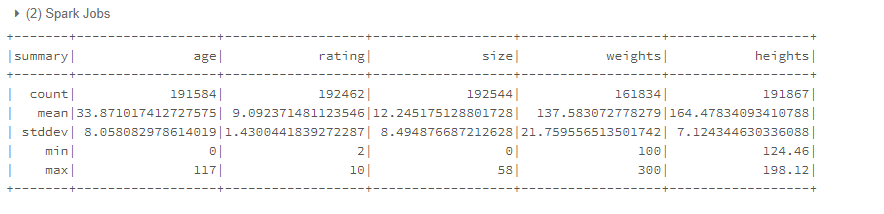
**Preliminary Data Exploration**

1. **Preprocessing**

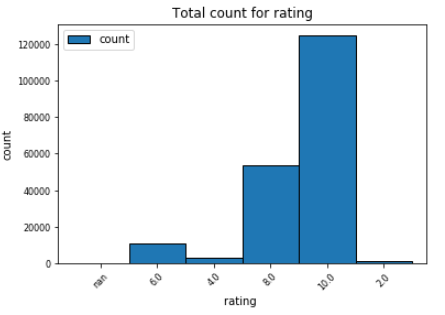
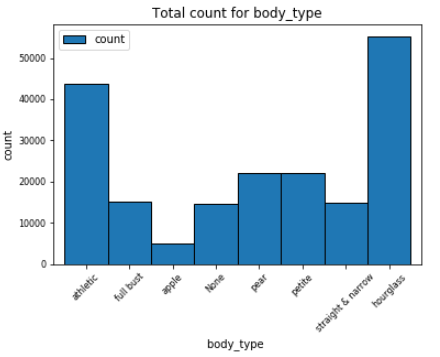
From the data set, we found that “review\_date” won’t make sense in our analysis, so we removed it from the original dataset. As mentioned above, the format of height doesn’t work well in the future. Therefore, we tried to extract numbers and transformed them into integer with units of centimeter. Besides, “weight” contains “lbs” in the value, so we removed “lbs” from the values and only kept its number.

1. **Summary statistics and visualization**

We did statistics analysis with those variables which are not string type. From the summary, we can find that all of them contain null values. We may remove those null values or fill them with means in the future, which is based on our analysis. We can find that the minimum rating is 2 while the average rating is around 9, which means most customers prefer to give a good rating.



From the body\_type figure, we can find that most types of customers are hourglass. This can be a clue to help the platform select their products. And from the rating figure, we can find that most customers prefer to give 10 ratings, which are high eough.



**Predictions**

* Predict the sentiment of reviews in each transaction for the purposes of finding out important factors influencing customers attitudes and preparing for the prediction of rating.
* Predict the situation where customers are likely to give different ratings for the purpose of judging whether the customer would like to provide a high rating or not.
* Predict what kinds product will get high ratings for the purpose of helping RentTheRunWay keep them stocked.
* Predict what kinds of product will fit different customers for the purpose of recommending appropriate product to customer.

**Inference**

Predictors that may be significant to measure the prediction of customer’s sentiment in each transaction:

* The existing results (a positive or negative orientation) of a sentiment analysis for each row of data: those results can be used in training dataset.
* User\_id: to define a special customer.
* Category: the category of a product can be used to sort customers’ attitudes to this kind of clothes.
* Fit: to measure whether the size of a garment fit a customer well with customer’s feedback.

Predictors that may be important when we try to predict whether a customer would like to give a high rating:

* Rating: existing rating customer gave to the product. This is an important parameter in training stage. A further rating hierarchy can also be made to determine the rating scores in which range could be described as a high rating.
* Fit

Predictors for a prediction of what kinds of items customers would like to give a high rating:

* Category
* Rating with a high value.
* Positive or negative attitudes customers have given in review\_text

Predictors of predicting what kinds of product will fit a special customer:

* Fit, weight, height, age, bust\_size, body\_type: to show the measurements of the customer.
* Rating
* Category, size: to show the measurements of the product.

**Non Spark Packages**

Nltk: In order to analyze the sentiment of the reviews, we need nltk package to deal with review text. For example, we need nltk to process POS tagging and tag sentiment for sentences in each record. Besides, we also need nltk to evaluate the sentiment trend of whole text in each review.