Applied Machine Learning

Data Analysis and Visualization

Team 25

- Huaizhi Ge
- Shashwat Singh
- Yosha Singh Tomar
- Alex Kita

Dataset Overview

Dataset 1: UCI Drug Review Dataset

The dataset provides patient reviews on specific drugs along with related conditions and a 10 star patient rating reflecting overall patient satisfaction. The data was obtained by crawling online pharmaceutical review sites. List of features:

- Drug Name - Categorical Variable
- Condition - Categorical Variable
- Review - Text
- Rating - Numeric (1-10)
- Date
- Useful Count - Numeric

Dataset 2: WebMD Drug Reviews Dataset

The dataset provides user reviews on specific drugs along with related conditions, side effects, age, sex, and ratings reflecting overall patient satisfaction. List of features:

- Age - Categorical Variable (Binned Categories)
- Condition - Categorical Variable
- Date
- Drug - Categorical Variable
- DrugID - Categorical Variable
- EaseofUse - Numeric (1-5)
- Effectiveness - Numeric (1-5)
- Satisfaction - Numeric (1-5)
- Reviews - Text
- Sex - Category
- Side Effects - Text
- Useful Count - Numeric

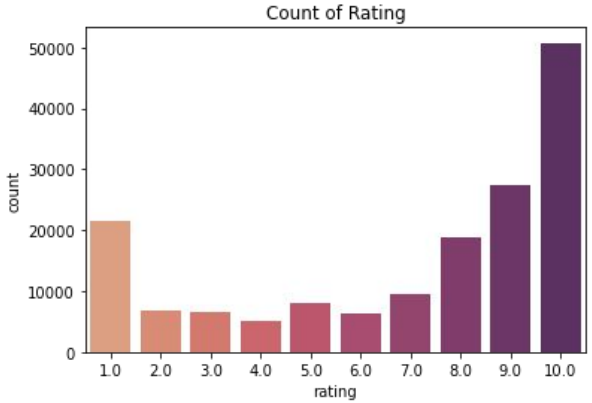
Data Cleaning of UCI Drug Review Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 161297 entries, 0 to 161296
Data columns (total 7 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   Unnamed: 0   161297 non-null  int64
 1   drugName     161297 non-null  object
 2   condition    160398 non-null  object
 3   review       161297 non-null  object
 4   rating       161297 non-null  float64
 5   date         161297 non-null  object
 6   usefulCount  161297 non-null  int64
dtypes: float64(1), int64(2), object(4)
memory usage: 8.6+ MB
```

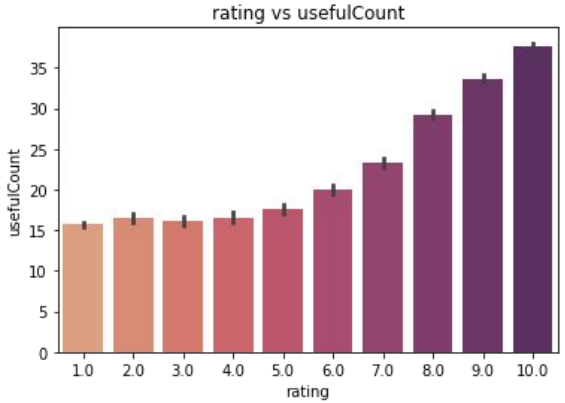
- We have 899 cases that have missing value in feature ‘condition’. We can just drop the rows with the missing values.
- Cleaned the ‘review’ feature. Removed special characters, non-ASCII characters, punctuations and stopwords.

Data Exploration of UCI Drug Review Dataset

Distribution of numerical features



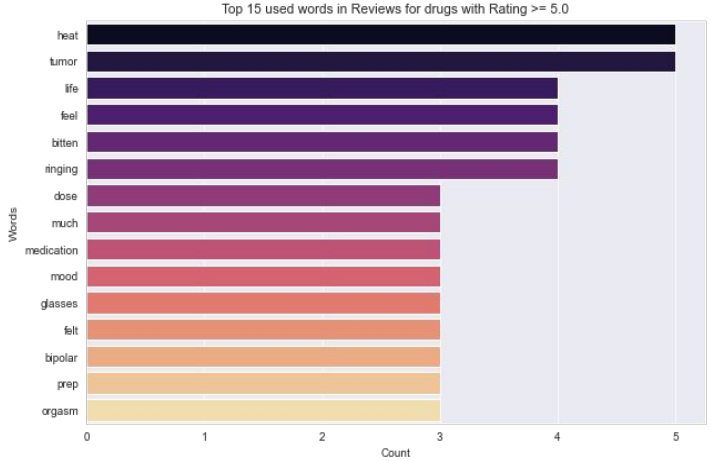
- This plot shows that there is dual polarity in the opinion of the people.
- Either people found the drugs extremely useful in which case they must have rated it 8 or above or either found it completely useless and rated the drug as 1.

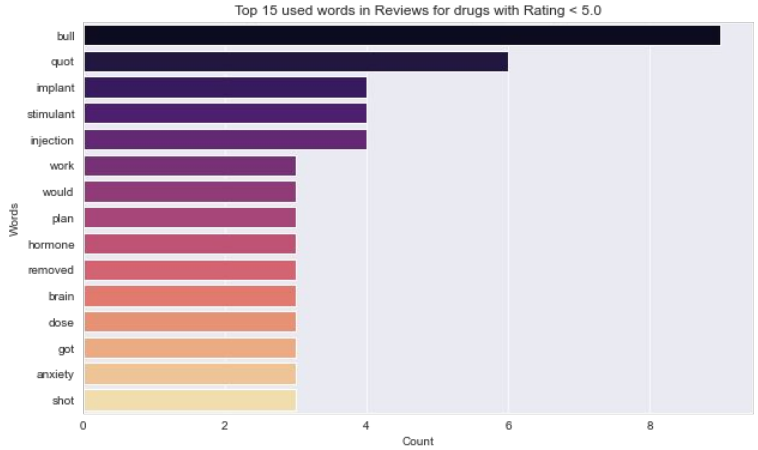


- Better ratings have higher mean Useful Count.
- Useful counts is the number of other users who found the review and rating satisfactory.
- Thus, we can determine that the reviews for high rated drugs present in the dataset are quite genuine.

Data Exploration of UCI Drug Review Dataset

Most common words in reviews w.r.t. ratings

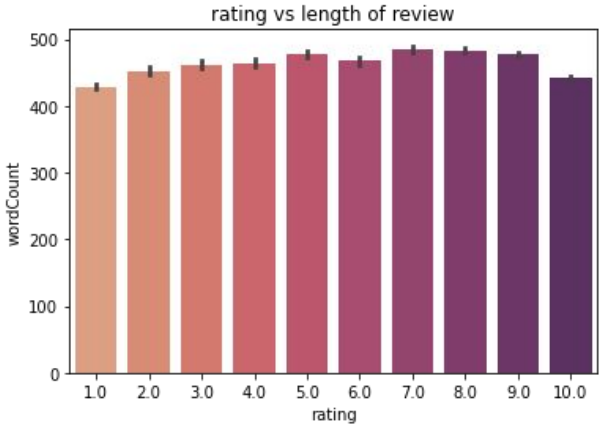


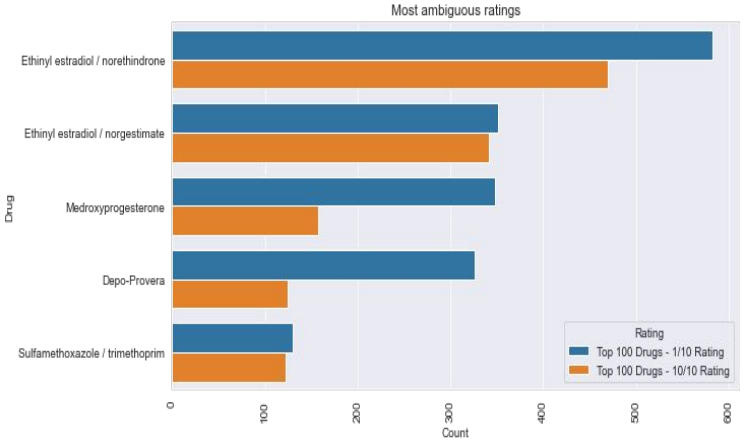


- Words such as tumor, glasses, bitten appear quite frequently in those reviews where the user has rated the drug >= 5.0/10.
- These might correspond to the conditions which were treated with a better efficacy than other conditions.
- Further, drugs with ratings less than 5/10 have most common words such as brain, anxiety, injection which might be the side effects or causes due to which the drug did not perform well.

Data Exploration of UCI Drug Review Dataset

Exploring reviews and top ambiguous ratings





- The length of the reviews looks generally similar to each other.
- There are maximum words in the reviews where the rating is 7.0/10.
- The trend shows a slow and gradual increase in word count with respect to the increase in the level of rating.
- Level 10 ratings has small number of word which means users do not tend to write much in the reviews when they are highly satisfied by the drug.

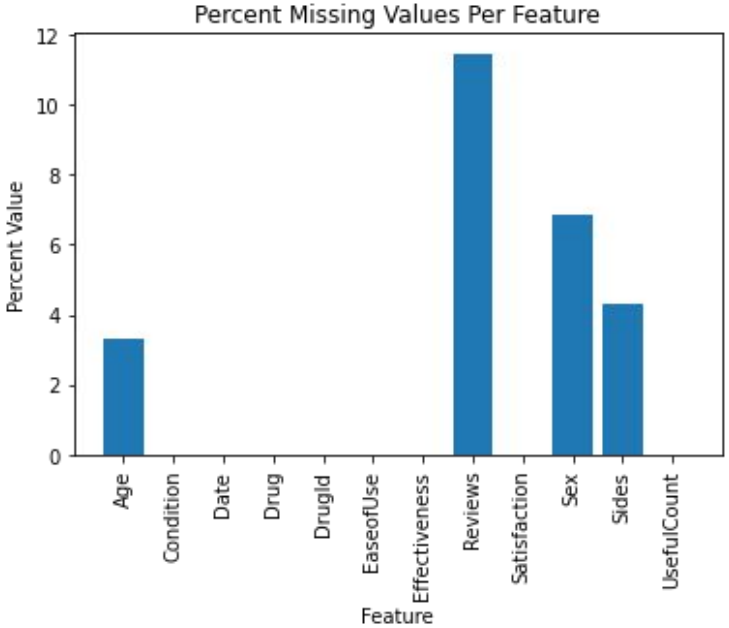
- We can observe that there are several ambiguous ratings in the dataset.
- These drugs have the most 10/10 Ratings as well as the most 1/10 Ratings. But the amount of 1/10 Rating cases is more than 10/10 cases.
- So, these drugs seems to have a contradictory perception with respect to the people using them.

Data Cleaning of WebMD Drug Reviews Dataset

- Removing Duplicate Information:
  Rows with completely identical information for drugs with similar names are filtered out:
  Example: Data for drug name: "*tramadol hcl*" and "*tramadol hcl er*" is exactly the same.

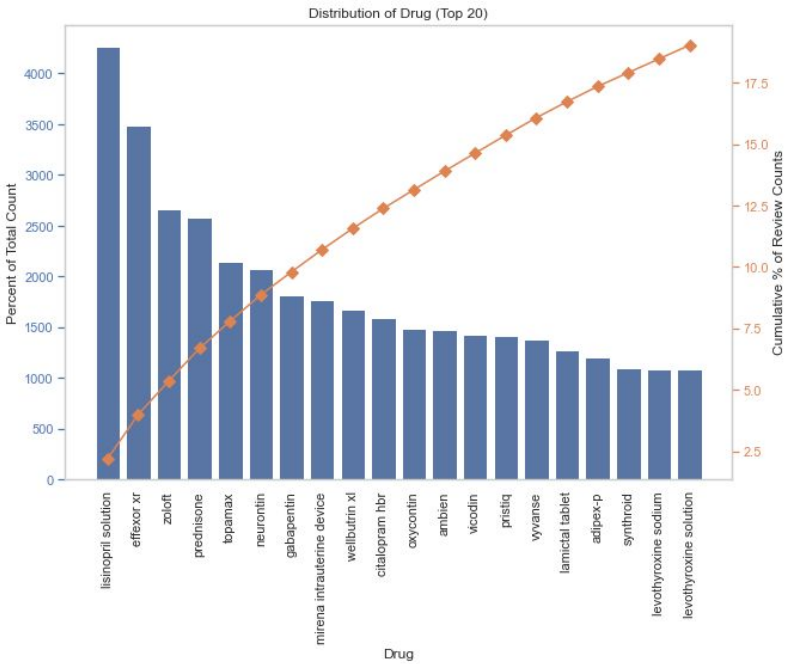
Before Filtering:	**After Filtering:**
`df.shape`	`df.shape`
`(362806, 12)`	`(194311, 12)`

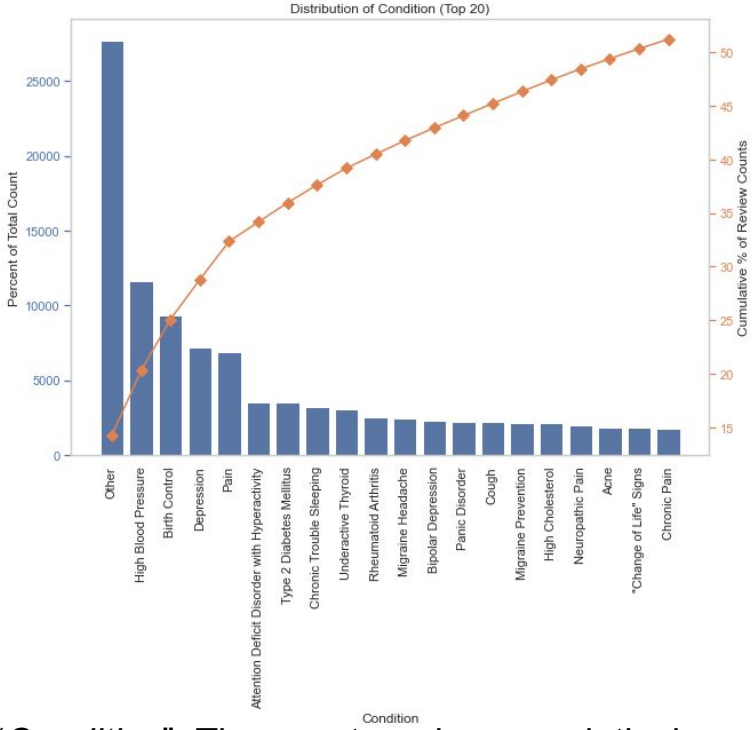
- Checking for Missing Values:
  The plot shows a distribution of missing values for each column. There are 2 types of missing value in the dataset:
  - Blank Value
    The columns Age, Reviews, Sex and Sides have missing values of the type " ".
  - Nan Value
    Only Reviews column has 12 rows with "Nan" value



Data Exploration of WebMD Drug Reviews Dataset

Most Common Drugs and Conditions

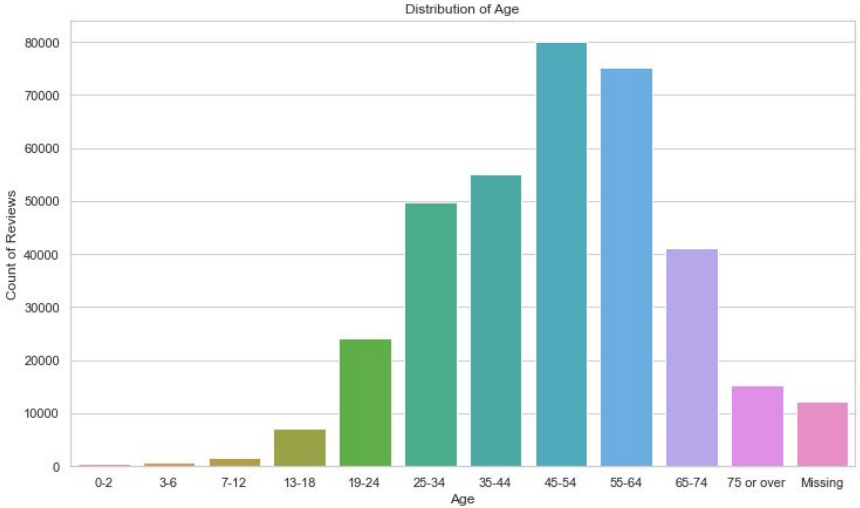


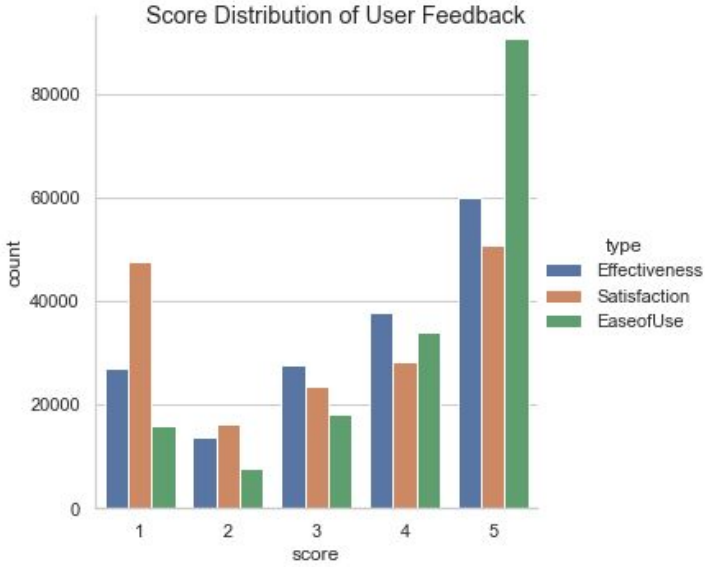


The plots show the top 20 categories in "*Drug*" and "*Condition*". These categories cumulatively make up 18% and 50% of the entire dataset. This information can be utilized while encoding these categorical features.

Data Exploration of WebMD Drug Reviews Dataset

Distributions of Feature Variables



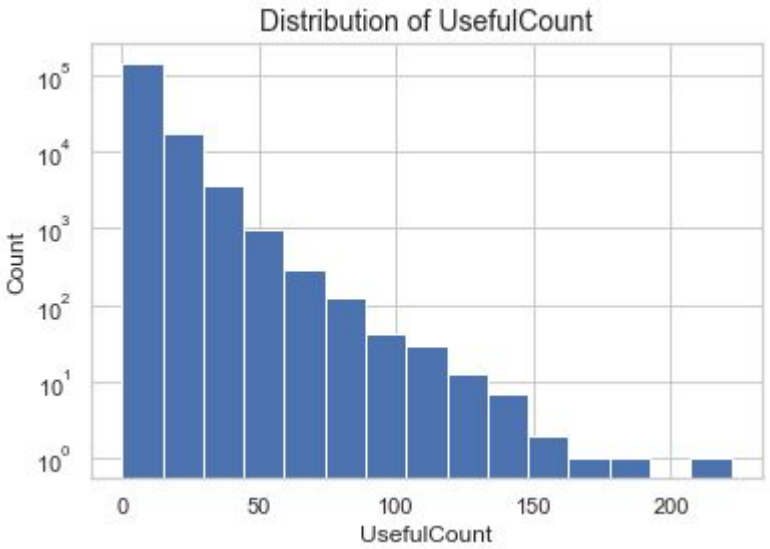


The distribution of age is unimodal, with maximum reviews in the category 45-54. An important point to note for this dataset is that the bin width of each category is not equal, it varies from as small as 0-2 and as large as 25-34.

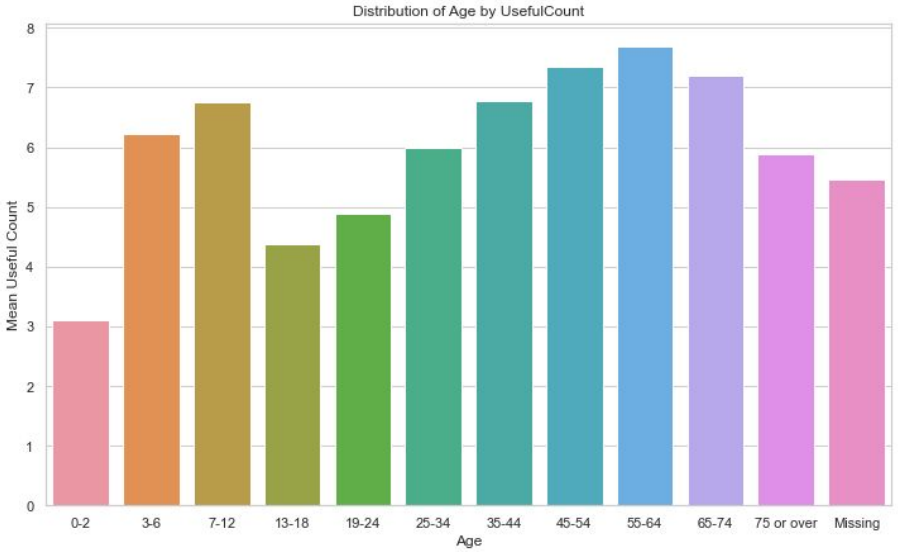
The score distribution of various parameters (Effectiveness, Satisfaction and EaseofUse) follow similar trends, most reviews either receive very low rating (1) or very high rating(4/5).

Data Exploration of WebMD Drug Reviews Dataset

Useful Counts on Drug Reviews



The histogram indicates that an inverse relation between the useful count and number of reviews. There are very few reviews with useful count greater than 100.



The mean useful count is bimodal with respect to age bins, with one peak at 7-12 and other at 55-64. The second peak can be attributed to the higher count of reviews for that age range.

Machine Learning techniques proposed to be implemented

Task 1: Build a model predicting user ratings based on reviews

Based on data exploration on UCI Drug Review Dataset, we observe that there is a relationship between reviews and rating scores. We will build an NLP model (BERT/ Word2Vec embeddings) to predict drug rating from review column present in this dataset.

Task 2: MultiClass Classification

Predict ratings from reviews present in the WebMD dataset from the NLP model and use Logistic Regression, SVM, Decision Trees & Random Forests (CatBoost) or Multi Level Perceptrons to finally predict the drug based on various input features present in WebMD dataset.