Sentiment Analysis - EDA

Group Members: George, Aleem

Github project link: https://github.com/GeorgeNikitakis/AIDI_1002.git

Below are some results of basic EDA on our dataset.
Our Data Set has 34660 rows and 21 columns

```
Df.shape
(34660, 21)
```

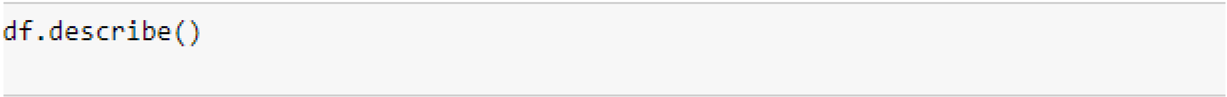
Below is the sum of null Values in each column of our Data Set

```
df.isnull().sum()
```

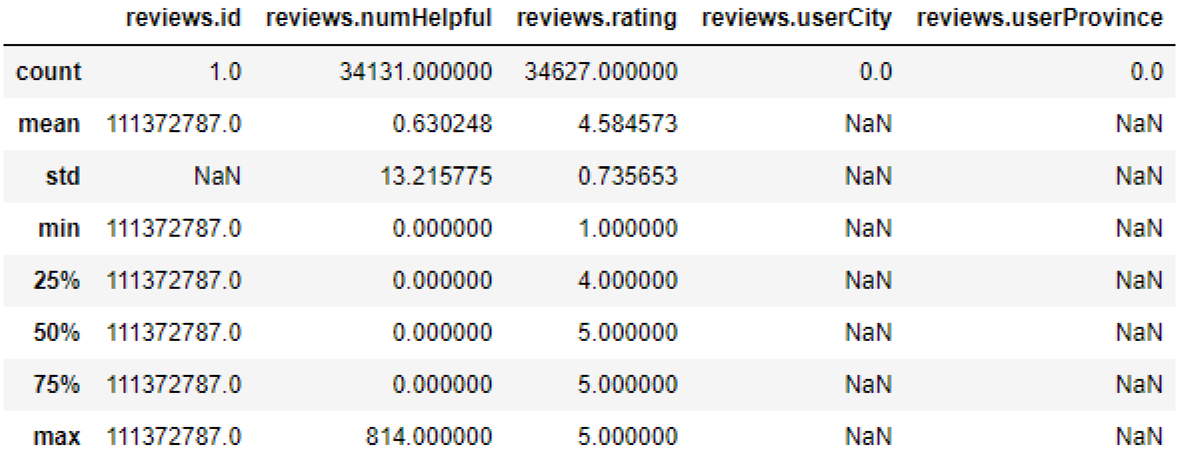
id	0
name	6760
asins	2
brand	0
categories	0
keys	0
manufacturer	0
reviews.date	39
reviews.dateAdded	10621
reviews.dateSeen	0
reviews.didPurchase	34659
reviews.doRecommend	594
reviews.id	34659
reviews.numHelpful	529
reviews.rating	33

reviews.sourceURLs	0
reviews.text	1
reviews.title	5
reviews.userCity	34660
reviews.userProvince	34660
reviews.username	2

Analysing some basic statistical details of our datasets:

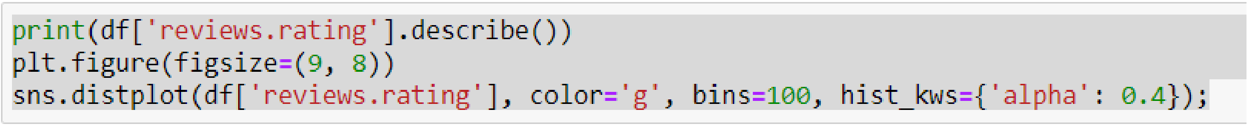


```
df.describe()
```

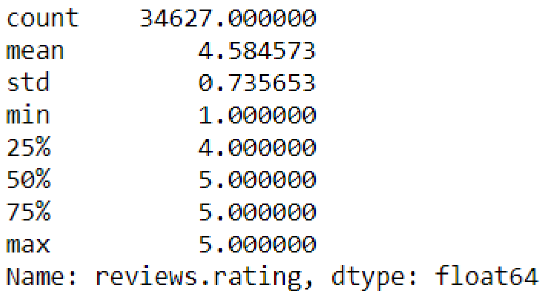


	reviews.id	reviews.numHelpful	reviews.rating	reviews.userCity	reviews.userProvince
count	1.0	34131.000000	34627.000000	0.0	0.0
mean	111372787.0	0.630248	4.584573	NaN	NaN
std	NaN	13.215775	0.735653	NaN	NaN
min	111372787.0	0.000000	1.000000	NaN	NaN
25%	111372787.0	0.000000	4.000000	NaN	NaN
50%	111372787.0	0.000000	5.000000	NaN	NaN
75%	111372787.0	0.000000	5.000000	NaN	NaN
max	111372787.0	814.000000	5.000000	NaN	NaN

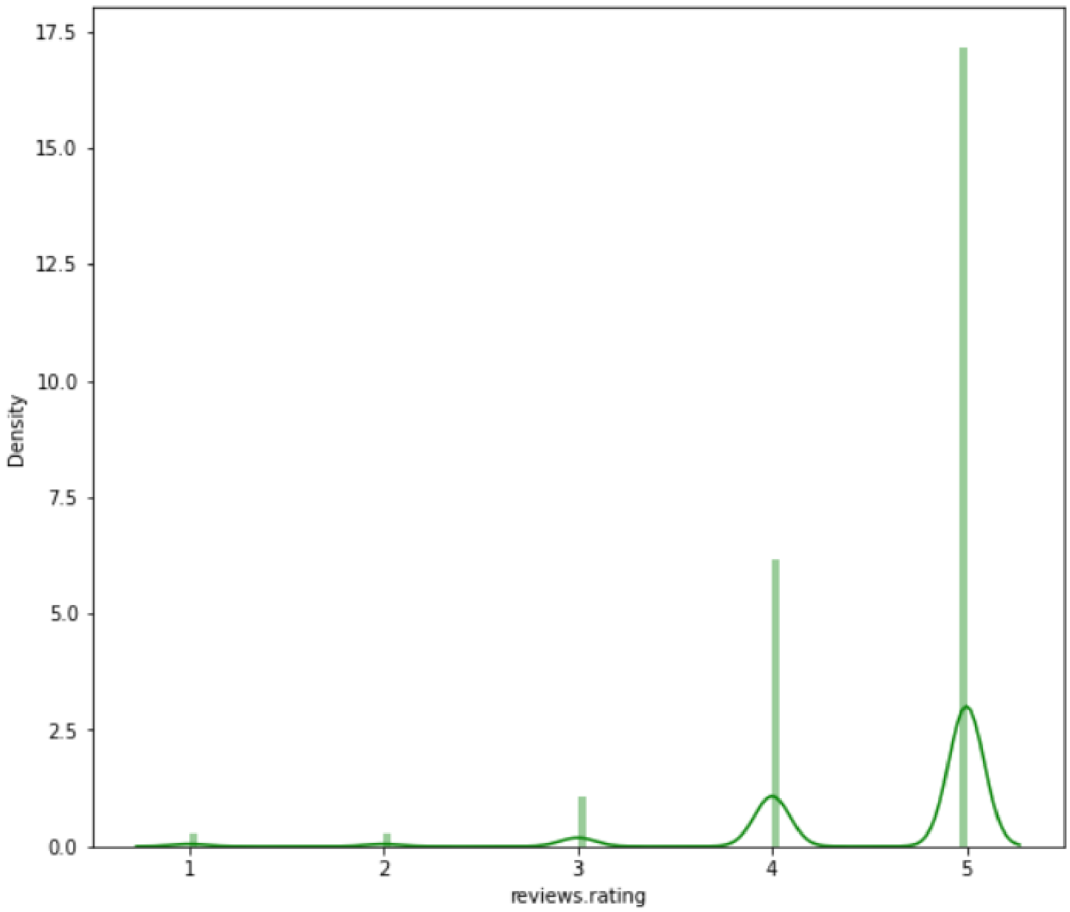
Distribution plot for the reviewer's rating:



```
print(df['reviews.rating'].describe())
plt.figure(figsize=(9, 8))
sns.distplot(df['reviews.rating'], color='g', bins=100, hist_kws={'alpha': 0.4});
```



```
count    34627.000000
mean         4.584573
std          0.735653
min          1.000000
25%          4.000000
50%          5.000000
75%          5.000000
max          5.000000
Name: reviews.rating, dtype: float64
```



After Analysing the above graph it clearly shows that most of the reviewer's rating below to 5 Followed by 4.

Benefits of Feature Engineering over optimal model output

Feature engineering is a machine learning technique that leverages data to create new variables that aren't in the training set. It can produce new features for both supervised and unsupervised learning, with the goal of **simplifying and speeding up data transformations** while also **enhancing model accuracy**.

Below are some advantages of Feature Engineering:

1. **Reduced complexity**: Algorithms are fed raw data to build these models. However, these algorithms make predictions without a clear guide. Feature engineering guide

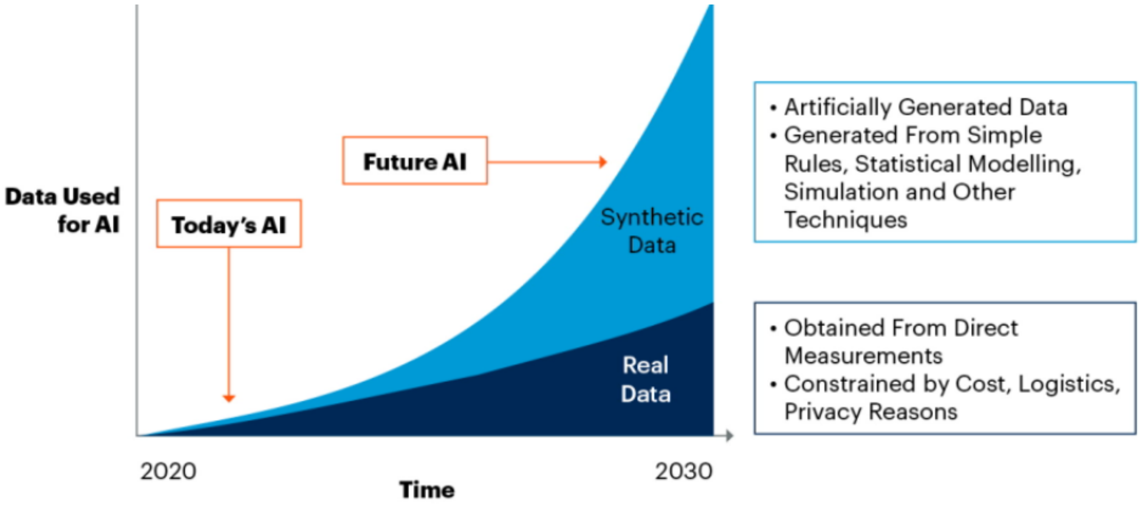
these algorithms. But what is the benefit of this? For starters, with correct features, model complexity will reduce. The correct scope and purpose of the model will make the process more efficient. This makes it simpler to understand, build, modify, and maintain models.

2. **Increased accuracy**: Feature engineering is a broad process. It involves transforming variables into suitable formats. Consider numerical data. It is already in a format that machine learning models can ingest. However, there may be situations where you may need to convert continuous values into discrete values. For example, when dealing with a feature whose data accumulates, it means it has an infinite upper boundary and has a high chance of attracting outliers. It would make sense to transform data from a continuous format to a discrete format. A technique such as binning, which is described later, can be used to carry out this transformation.
   Feature engineering also involves the creation of new variables from existing ones. Filling in missing values, among other processes, is also under this umbrella. These processes ultimately influence the accuracy of the model. When done correctly, with the correct data, it increases the accuracy of the models.

Identify key features between synthesized and real data for model input:

It is inexpensive compared to collecting large datasets and can support AI/deep learning model development or software testing without compromising customer privacy. It's estimated that by 2024, 60% of the data used to develop AI and analytics projects will be synthetically generated.

Below is the prediction for the future for use of Synthetic Data ; Source: Gartner



Key Candidate Features between synthesized and real data for model input:

Though synthetic data has various benefits that can ease data science projects for organizations, it also has limitations:

- Outliers may be missing: Synthetic data can only mimic the real-world data, it is not an exact replica of it. Therefore, synthetic data may not cover some outliers that original data has.
- Quality of the model depends on the data source: The quality of synthetic data is highly correlated with the quality of the input data and the data generation model. Synthetic data may reflect the biases in source data
- User acceptance is more challenging: Synthetic data is an emerging concept and it may not be accepted as valid by users who have not witnessed its benefits before.
- Synthetic data generation requires time and effort: Though easier to create than actual data, synthetic data is also not free.
- Output control is necessary: Especially in complex datasets, the best way to ensure the output is accurate is by comparing synthetic data with authentic data or human-annotated data. this is because there could be inconsistencies in synthetic data when trying to replicate complexities within original datasets

Data Cleaning

Data cleaning consisted of mainly eliminating columns we believed to be unnecessary in train and testing our model. Of the 21 columns in our original dataset, we reduced it down to 9 columns. Additionally removed rows with missing values in specific features and split the dataset into training and testing dataset.

Training set shape = **(19504, 9)**

Testing set shape = **(8360, 9)**