# Question 1 Data Exploration and Visualization

Provide an initial step to inspect, visualize and analyze the different attributes in your data set. Document your findings and make conclusions for your next steps.

For this task we began by exploring our training data set. It’s composed of 10 attributes and 426,340 examples.

The **first** **task** was getting aquatinted with our data types:

Data columns (total 10 columns):

Id 426340 non-null int64

ProductId 426340 non-null object

UserId 426340 non-null object

ProfileName 426326 non-null object

HelpfulnessNumerator 426340 non-null int64

HelpfulnessDenominator 426340 non-null int64

Score 426340 non-null int64

Time 426340 non-null int64

Summary 426320 non-null object

Text 426340 non-null object

As seen from above, we have 6 numerical attributes and 5 categorical or non-numeric attributes in total. Our class attribute is **Score.**

The **second task** was **statistical analysis of our data set**:

A screenshot of a cell phone

Description automatically generated

An interesting insight that can be observed is that the 50% and 75% for score had the value of **5.0** which is the maximum score. We began to explore this further.

A close up of a device

Description automatically generated

As seen from the above plot. **Most of the data is biased towards the score 5.0** that means our class values are not **stratified** or evenly distributed. our next step was to explore this further; we began to analyze the data considering the score distributions.

A screen shot of a social media post

Description automatically generated

As seen from above, **272492** examples are on the on the class value of 5.0. this over **50% of our data.** Our next step was split into 2 stages: numerical correlation and categorical correlation.

We first began by pair plotting our numerical attributes with our score to see if we can find any interesting relationships.

A screenshot of a cell phone

Description automatically generated

the above plots gave no indicative relationship, so we explored further. We began by calculating the correlation values of each of the numerical attributes with our class attribute:

HelpfulnessDenominator: -0.07724218886413477

Time: -0.06318610909302598

HelpfulnessNumerator: -0.022805615821503325

Id: 0.009985657224340675

As seen from above, the correlation values confirm the plots that there is no distinct correlation with the class. We moved on to explore the non-numeric attributes in our data set.

['ProductId', 'UserId', 'ProfileName', 'Summary', 'Text']

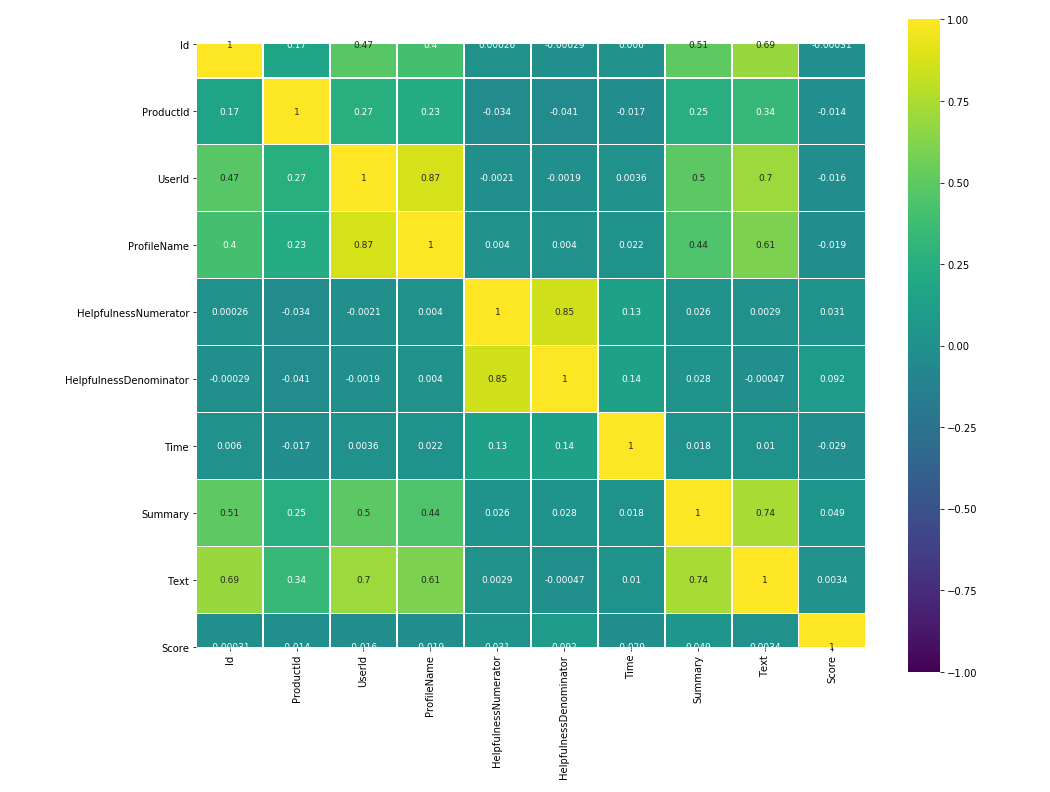
We factorized the categorical attributes to be able to calculate a correlation score. Furthermore, we first explored **inter-attribute** correlation.

A close up of text on a white background

Description automatically generated

The correlation matrix provided some interesting results **text is highly positively correlated with product name and summary while summary is highly positively correlated with text as well.**

From this we concluded that among the categorical attributes **text** and **summary** seem to be strong attributes. We further began to plot a full correlation matrix to verify our assumptions.



From the correlation matrix, we discovered that none of the attributes had a strong correlation to the class value. This could be for a variety of reasons so instead we diverted our focus to strong inter-correlations, and we found that the best candidates are **Text** and **Summary** attributes. We merged these attributes together to lay foundations for our text processing. In the next stage of our analysis we began by exploring the **values** in our data set. We first began inspection for null values.

Id 0

ProductId 0

UserId 0

ProfileName 14

HelpfulnessNumerator 0

HelpfulnessDenominator 0

Time 0

Summary 20

Text 0

Score 0

As seen from above we’ve found that there are null values in the **summary and profile name** attributes. Furthermore, we investigated if there are any duplicate values.

A screenshot of a social media post

Description automatically generated

As can be seen from above. We found duplicate reviews in the data set; those were removed to avoid noise in our data. We then shifted our focus to preprocessing of our text.

After further analysis we found the following properties in our text attribute:

1. Html tags
2. Accents
3. Punctuation

These 3 were cleaned from the data before we can begin our preprocessing. Finally, the next stage in our pipeline was tokenization, lemmatization / normalization so that we may begin representing our textual data in a way for modeling.