EDA

Steps) ① Drop duplicates & NA values.

Code: df.drop_duplicates(subset=['Text'], inplace=True)

df.dropna(axis=0, inplace=True)

② Preprocessing: drop all unwanted symbols, characters etc. from the text that do not effect the objective of problem.

① Contraction Mapping

Contraction_mapping = {"ain't": "is not", "aren't": "are not"}

② Converting everything to lowercase / Remove HTML Tags /

Remove Stopwords

stop_words = set(stopwords.words('english')) +--

cleaned_text = ['bought several vitality ... product better', ...]

cleaned_summary = ['good quality food', ...]

data['cleaned_text'] = cleaned_text, data['clean sum'] = clean su

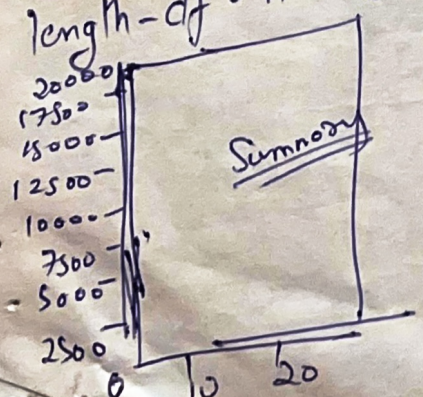
Visualizing Distribution: import matplotlib.pyplot as plt

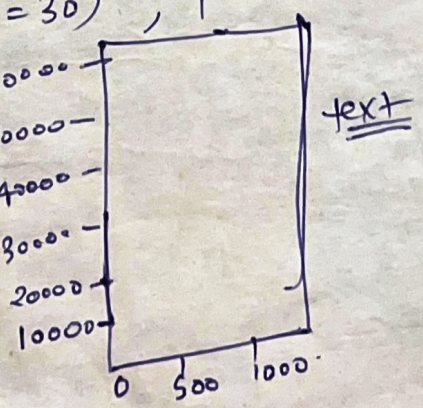
text_word_count = [], summary_word_count = []

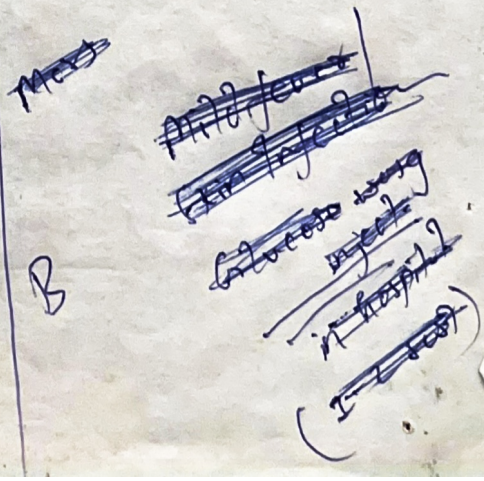
for i in data['cleaned_text']: text_word_count.append(len(i.split(

length_df = pd.DataFrame({'text': text_word_count, 'Summary': Summ count})

length_df.hist(bins=30), plt.show()







EDA

1. Business Understand.
2. Data Understanding / Acquiring.
3. Identifying & handling the missing values.
4. Encoding the categorical data.
5. Checking the correlation blw features & handling.
6. feature importance (if features are many)
7. ~~Target variable is balanced / imbalanced~~
8. Splitting the dataset into train & test.
9. Feature Scaling.

If Model is not performing well on Test data:

Data Profiling (Test data).

Steps: ① Outliers check.

② will check features / region where the values are abruptly changing (due to surround circumstances, pandemic...

③ Will check/compare test feature statistics with train feature statistics

If All seems fine:
⇒ Regularization ← model check.
⇒ Cross Validation if it is due to overfitting (Powerful measure)
⇒ Remove feature.
⇒ Ensembling