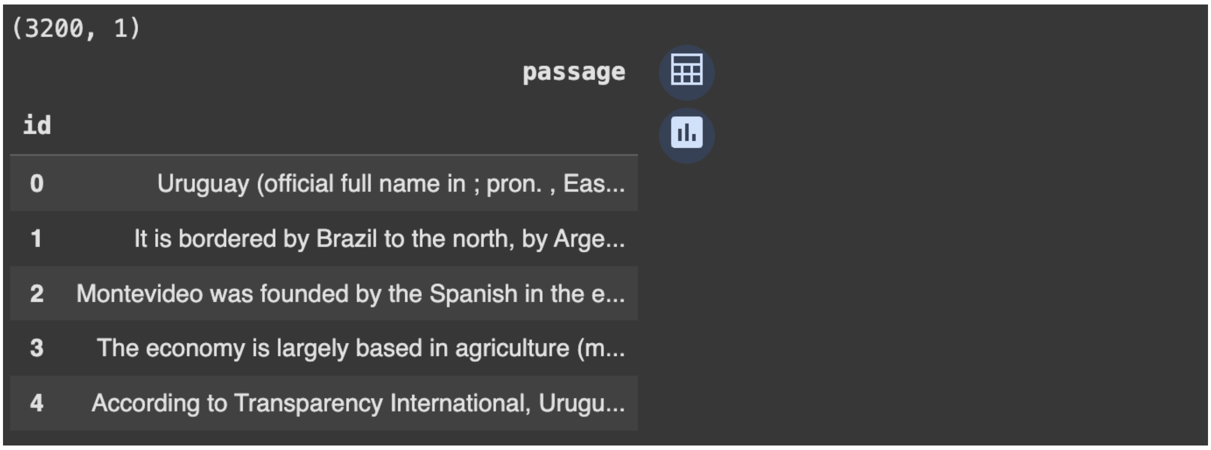
Data Understanding and Exploratory Data Analysis

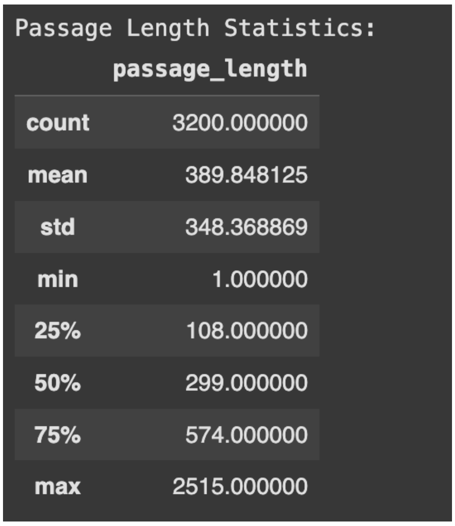
This report details the data understanding part of the dataset, which would be used as the knowledge base for our RAG system.

The dataset, sourced from rag-datasets/rag-mini-wikipedia, specifically utilizes the passages.parquet partition was successfully loaded into a pandas DataFrame. The dataset provides us with 3200 distinct passages, which would be used as the foundational text for our RAG system.



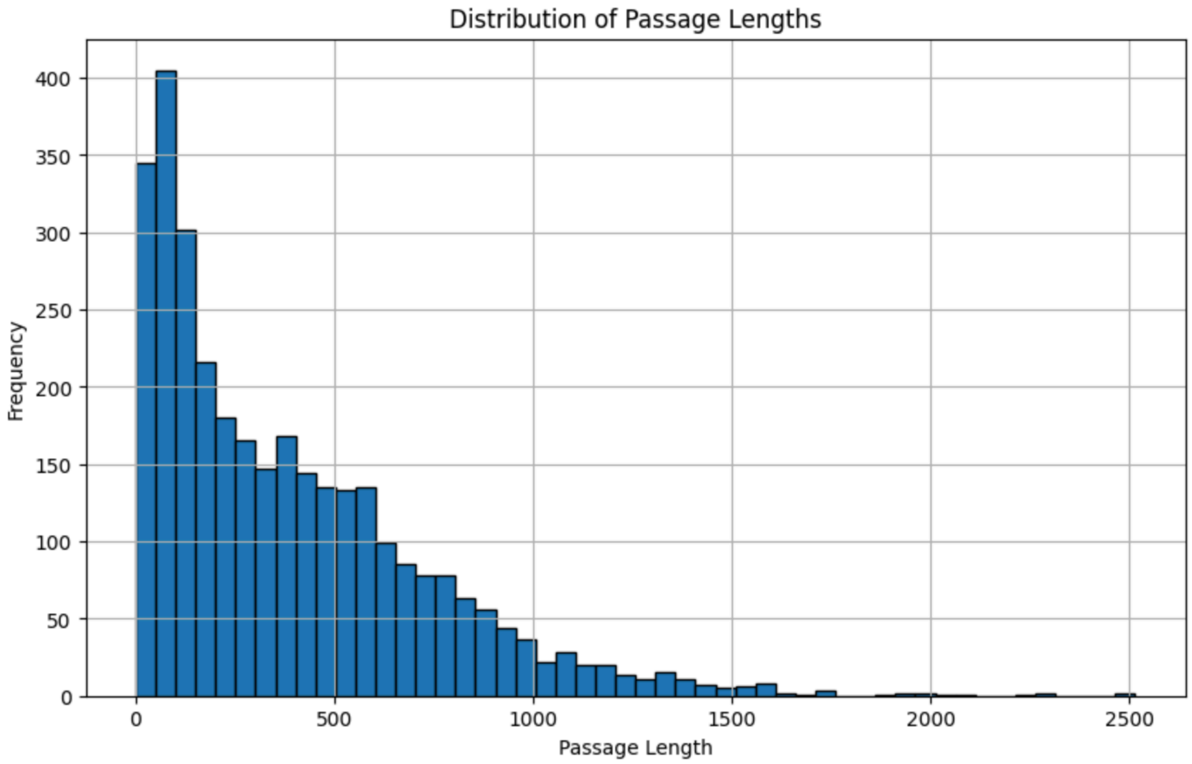
```
(3200, 1)
                                              passage
id
0         Uruguay (official full name in ; pron. , Eas...
1        It is bordered by Brazil to the north, by Arge...
2   Montevideo was founded by the Spanish in the e...
3      The economy is largely based in agriculture (m...
4    According to Transparency International, Urugu...
```

I started by checking the data integrity. Thus, the most important part was to check for the missing or NULL values within the passage column, which is fundamental for text-based retrieval. The analysis confirmed the absence of any null entries, indicating a clean dataset with 3200 entries.

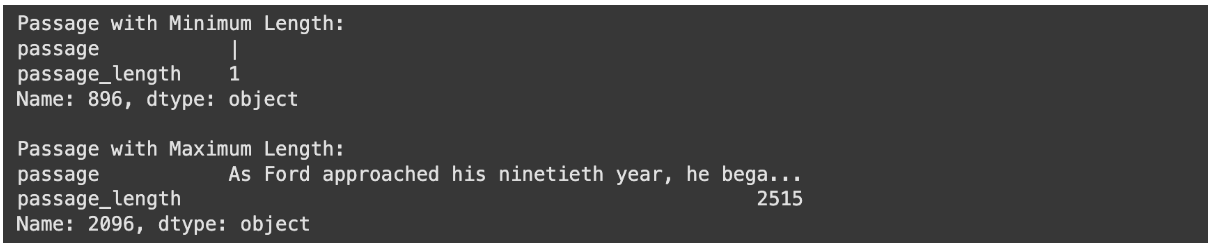


```
Passage Length Statistics:
       passage_length
count     3200.000000
mean       389.848125
std        348.368869
min          1.000000
25%        108.000000
50%        299.000000
75%        574.000000
max       2515.000000
```

Subsequently, I conducted exploratory data analysis focusing on the characteristics of the passages, such as the length of each passage and the following metrics: average, maximum length, and minimum length. The passage lengths were found to range significantly, ranging from a minimum of 1 character to a maximum of 2515 characters, keeping the average at 390 characters. To visualize this, a histogram was made to see the spread in the data.



While no immediate cleaning steps were strictly necessary based on the absence of null values, EDA provided me with considerable insights into the data. The variation in passage lengths could influence the chunking process or the embedding process. However, the dataset is fine being used as it is for tokenization, embedding, and indexing in the vector database.



```
Passage with Minimum Length:
passage           |
passage_length    1
Name: 896, dtype: object

Passage with Maximum Length:
passage           As Ford approached his ninetieth year, he bega...
passage_length                                                  2515
Name: 2096, dtype: object
```

PS: While I did think of removing the passage entries that were smaller than either 10-15 characters, it could be a possibility that the shorter entries just have crucial information, such as a date or name, or something very specific; thus, I decided to keep all the entries as a tradeoff for potential information with some noise in the dataset.