To explore what kind of data science books are available out in Amazon:

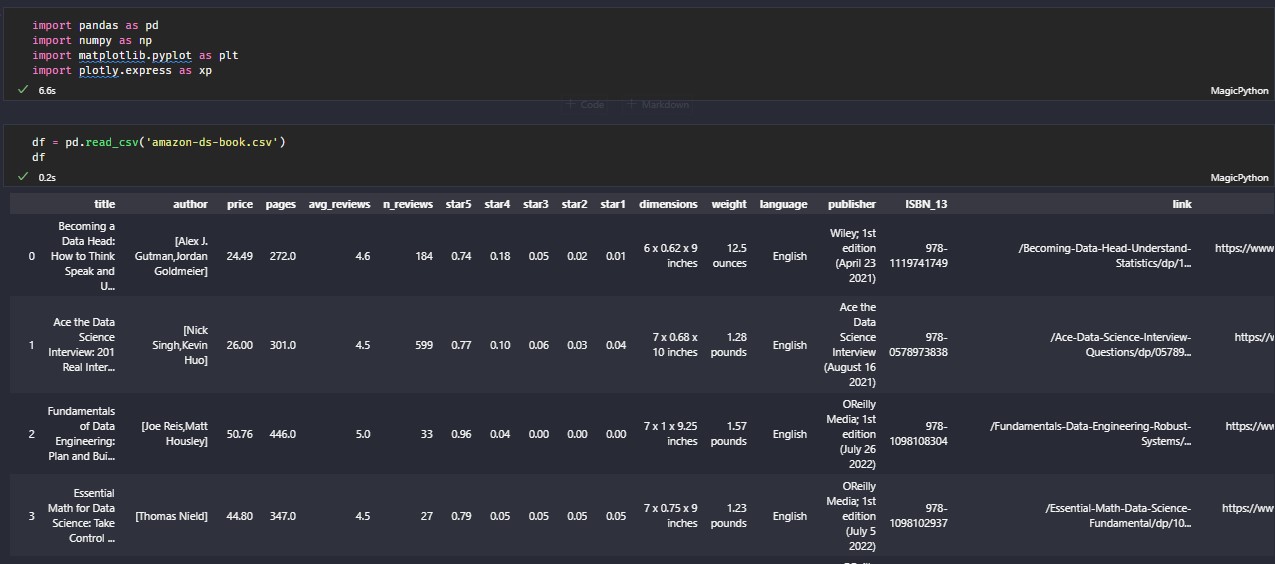
I found a dataset on Kaggle that contains nearly 1000 data science books on Amazon including the title, author, ratings, number of pages and other features.

We’ll be doing three levels of data analysis, Beginner, intermediate and advanced.

In level 1, we’ll do simple Exploratory Data Analysis to answer questions such as;

1. More expensive books have better reviews?
2. Is it always true that longer pages are more expensive or what are the best python related books and machine learning books based on review stars?

In level 2, we’ll do Cluster analysis on book titles.  And we want to find out what are the main types or main categories of data science books out there based on book titles. We’ll be using K-means for clustering the book titles and in order to do that for text data we use the NLP technique called TF-IDF [Term frequency-inverse document frequency] to convert the text into numeric features.



After importing relevant libraries and loading the data out the Data Frame here we can see a beautiful overview of the data including the distribution of the data in each column and also the percentages of missing values in each column which is quite handy. We have 946 book titles containing many data science and machine learning books. As we can see, we have title, author, price, pages etc.