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

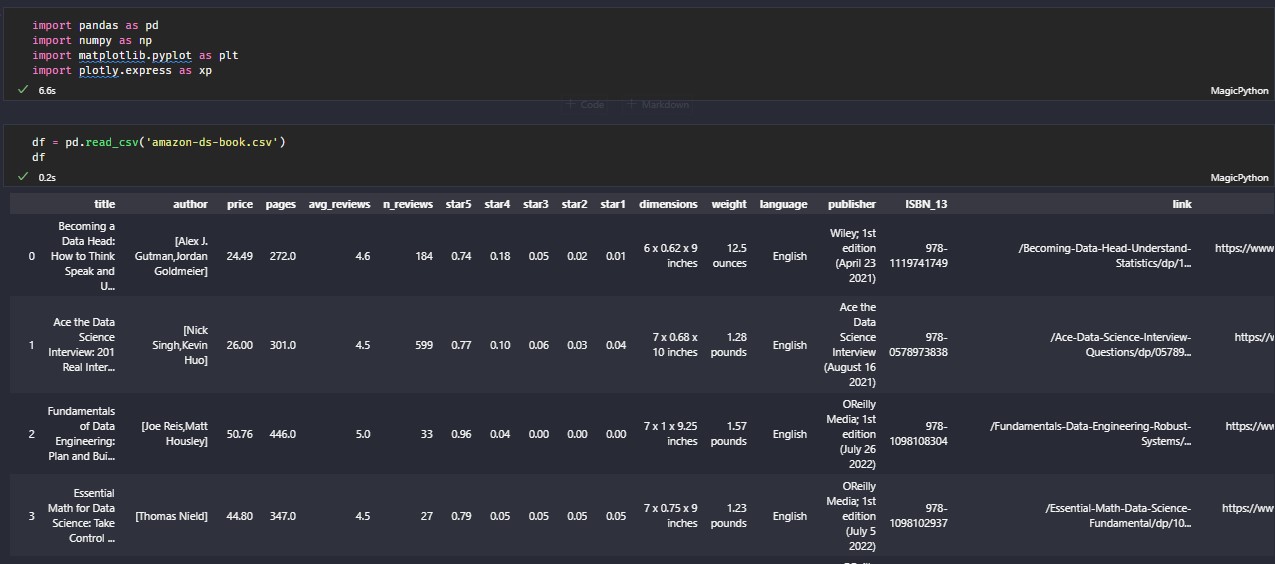
I found a dataset on Kaggle that contains nearly 1000 data science books on Amazon including the title, author, ratings, number of pages and other features.

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.

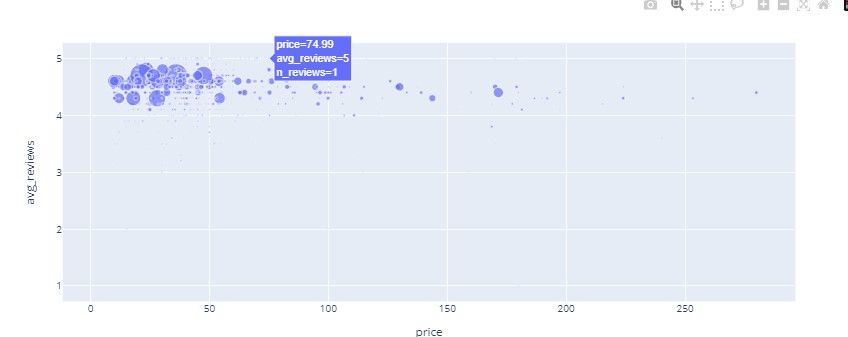


After importing relevant libraries and loading the data out the Data Frame here we can see a beautiful overview of the data including the distribution of the data in each column and also the percentages of missing values in each column which is quite handy. We have 946 book titles containing many data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

Level 1: Exploratory data analysis to answer a question like do more expensive books have better reviews.

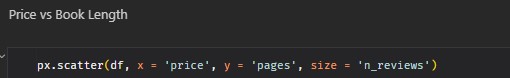
For that what we can do is to create a scatter plot that plots the book price against the reviews. In such a case, we’re using Plotly Express to quickly create the scatter plot with the x-axis being the price and the y-axis being the reviews and adding the size of the number of reviews.

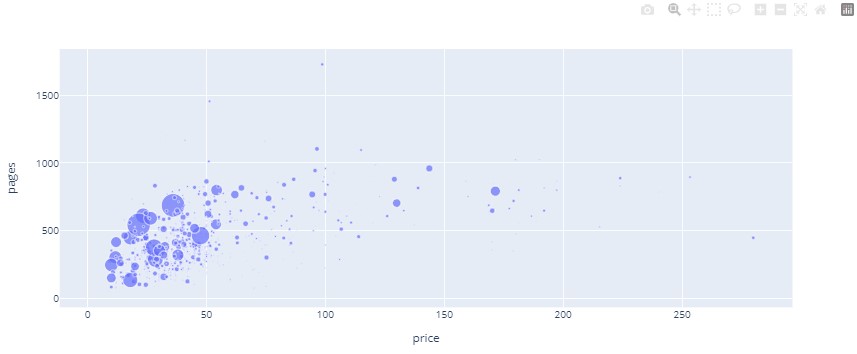




After doing data visualization with scatter plot we can see that there’s no clear relationship between price and average reviews. So, good books can also be very cheap or they can be very expensive.

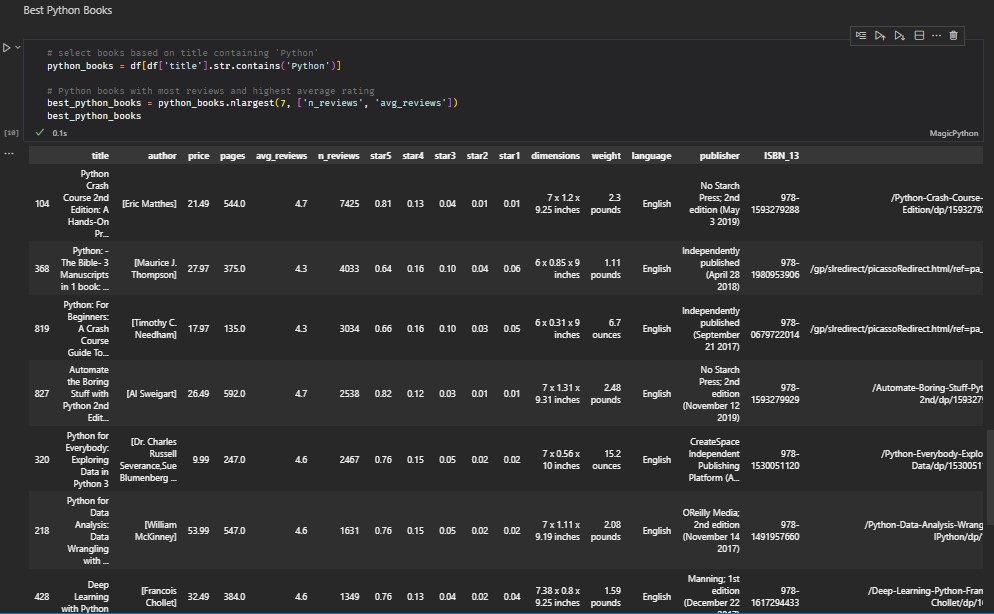
Similarly, if we do the same exercise but plotting the price against the number of pages. We can see that there’s some positive correlation. Longer books tend to be a bit more expensive which makes total sense because it costs more time and effort to write those books.





Now sort out which are the best Python books and Machine Learning books.

* We can do this by a simple rule-based method so select all the books with the titles containing Python.
* And now select the Python books with the most reviews and have the highest average rating. We can do this by using n largest function (nlargest ()). For instance, if we want to only select the top seven books and here, we can select the columns that we want to use to sort the data. The reason why we want to sort data on both number of reviews and average reviews is because some books might be not very popular and they have very few reviews but they are very good reviews so these cases might be more prone to bias and in these 7 best Python books we have:



Best Machine Learning books:

