**TARAS SHEVCHENKO NATIONAL UNIVERSITY OF KYIV**

**FACULTY OF INFORMATION TECHNOLOGIES**

**DEPARTMENT OF INTELLECTUAL TECHNOLOGIES**

**Personal task**

on discipline «Data processing software»

Made by student

of group ІАВ-11

Buzyurova Anna

**Kyiv – 2022**

**1. Dataset description. Also should include variables description with units and dataset credits.**

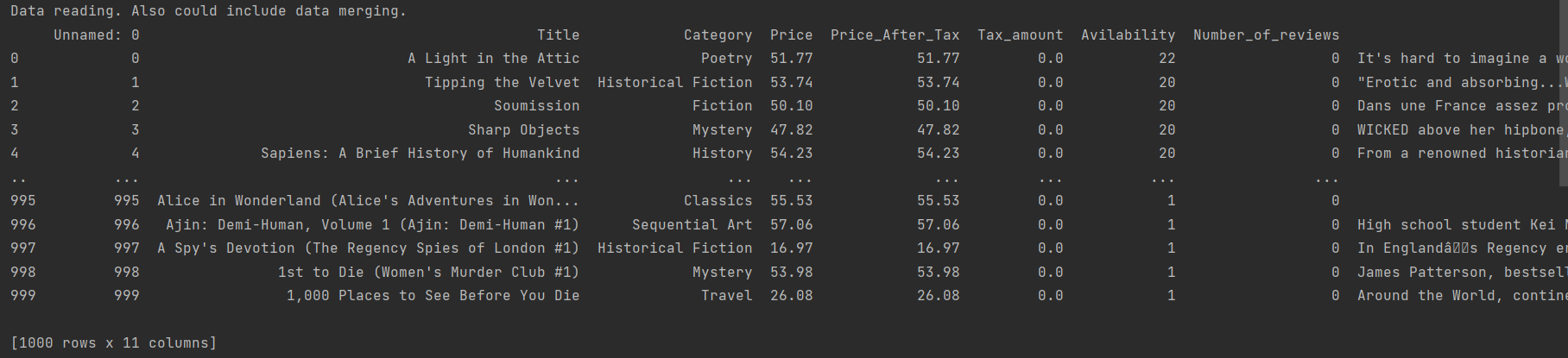
I took data set about the books on a website from Kaggle website.

This data set includes the following attributes/columns:

* Title: The title of the book. (string)
* Category: category of the book. (string)
* Price: the price of the book. (double)
* Price After tax: the cost of the books including tax. (double)
* Tax amount: tax on the book. (double)
* Availability: the quantity available in the stock. (int)
* Number of reviews: number of people who reviewed the book. (int)
* Book Description: description of the book. (string)
* Image Link: link where you can see the image of the book. (string)
* Stars: star rating for each book out of 5. (int)

**2. Data reading. Also could include data merging.**

*data = pd.read\_csv('../data/Book\_Dataset\_1.csv')*



In console we can see all the columns, but unfortunately I can do pretty copy paste.

**3. Data tidying.**

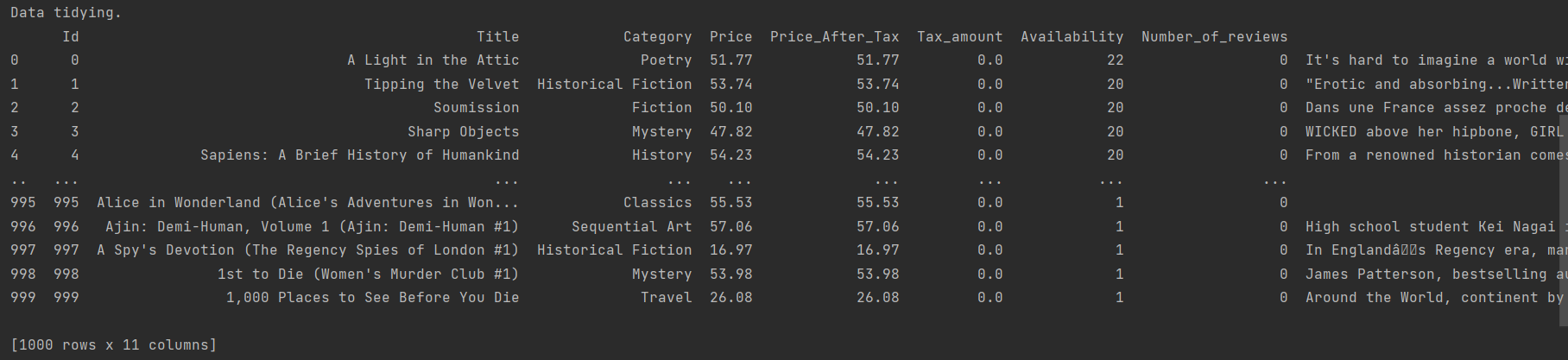
Rename first column to "Id" and 'Avilability' to correct 'Availability':

*data.rename(columns={'Unnamed: 0': 'Id', 'Avilability': 'Availability'}, inplace=True)*

Convert columns Id,Price,Price\_After\_Tax,Tax\_amount,Availability,Number\_of\_reviews,Stars to numeric

*data[['Id', 'Price', 'Price\_After\_Tax', 'Tax\_amount', 'Availability', 'Number\_of\_reviews', 'Stars']] = \*

*data[['Id', 'Price', 'Price\_After\_Tax', 'Tax\_amount', 'Availability', 'Number\_of\_reviews', 'Stars']].apply(pd.to\_numeric)*



**4. Data analyzing with graphics.**

Questions:  
1. What are the column names of the data frame?

*data.columns.values*

Answer: ['Id' 'Title' 'Category' 'Price' 'Price\_After\_Tax' 'Tax\_amount'

'Availability' 'Number\_of\_reviews' 'Book\_Description' 'Image\_Link'

'Stars']

2. How many books (i.e. rows) are in this data frame?

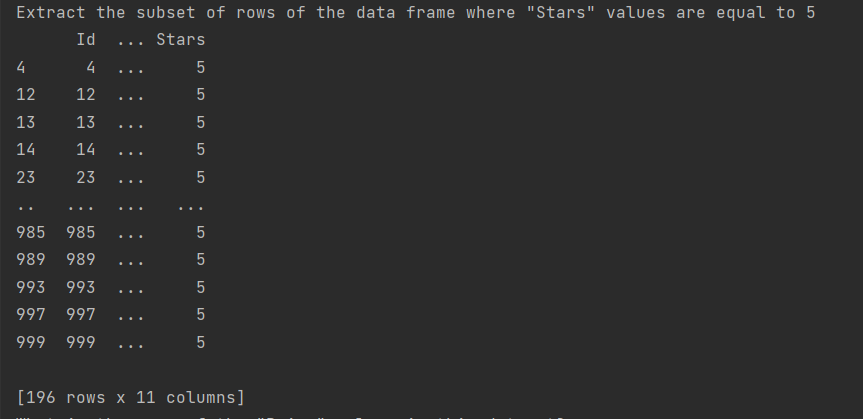
*len(data.index)*

Answer: 1000

3. Extract the subset of rows of the data frame where "Stars" values are equal to 5

*data[data['Stars'] == 5]*

Answer:



4. What is the mean of the "Price" column in this dataset?

*data['Price'].mean()*

Answer: 35.07035

Graphics:

- Show count of books for each "Stars" value

*data.groupby('Stars', as\_index=False).count().plot(x='Stars', y='Id', kind='bar')*

*plt.legend(labels=['Count of books'])*

Result:

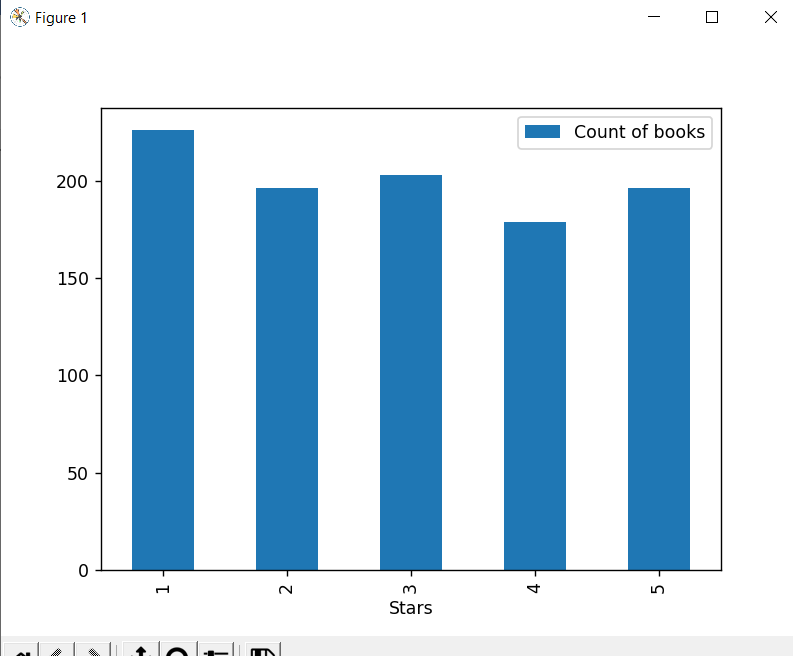


Figure 1 – Count of books with stars

- Show first 15 books with 5 stars and availability more than 10

*ax = data[(data['Stars'] == 5) & (data['Availability'] > 10)].plot(x='Title', y='Price', kind='barh')*

*ax.set\_ylim([0, 15])*

Result:

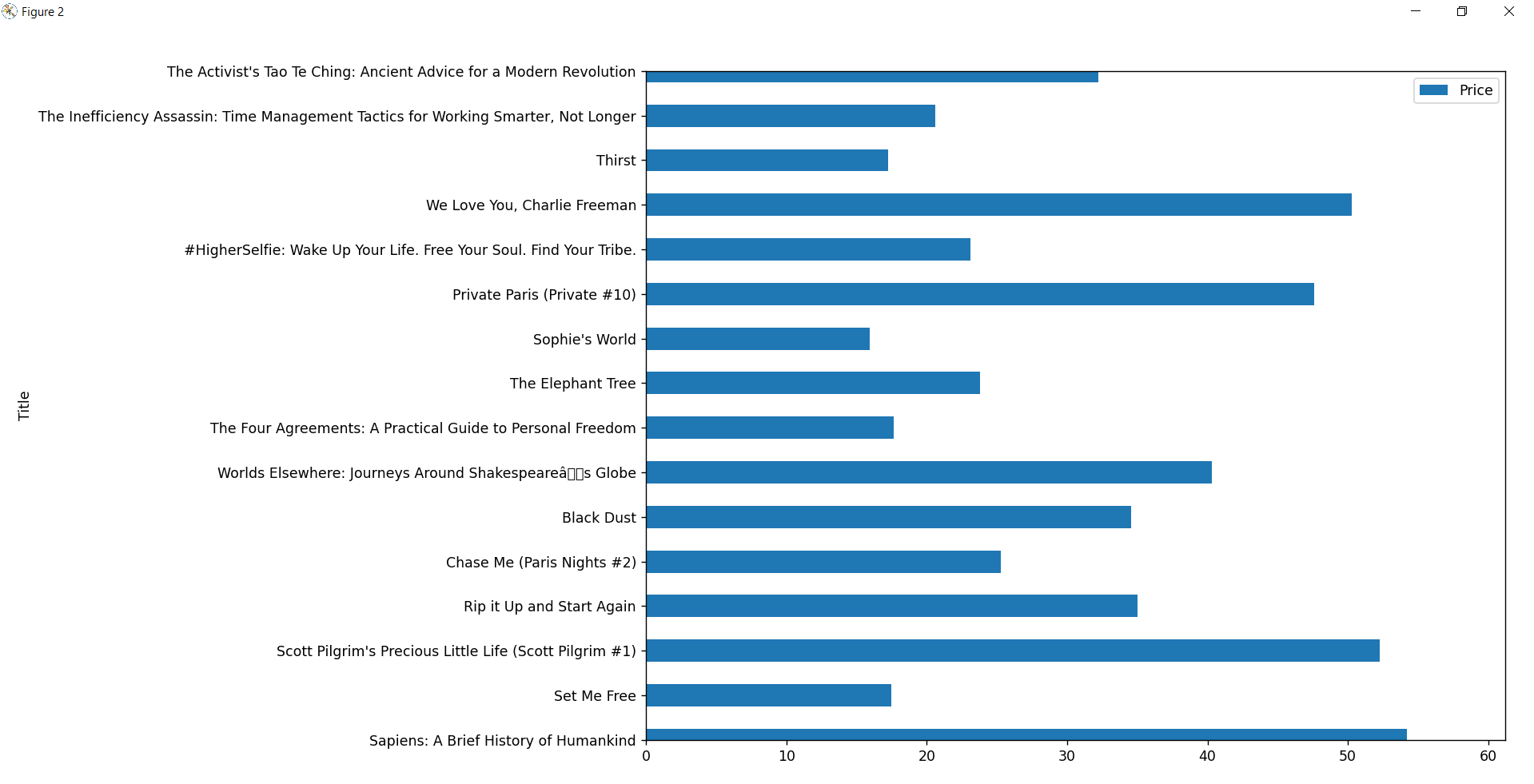


Figure 2 – book with 5 stars and availability > 10