Mohammed Ayman Ahamd 2206142

Data cleaning

```
]: import pandas as pd
```

```
]: # read file
DF = pd.read_csv(r"E:\College Courses\Semester4\Cloud Computing\Assignments\books.csv")
DF.shape
```

```
]: (1354, 23)
```

```
]: # clean file
missing_values = DF.isnull().sum()
DF.dropna(inplace=True)
DF.drop_duplicates(inplace=True)
DF.shape
```

```
]: (1153, 23)
```

Preprocessing (Focus your analysis on the Harry Potter book series)

```
#to know the rows that will be analysed
h1 = "Harry Potter and the Sorcerer's Stone (Harry Potter, #1)"
h2 = "Harry Potter and the Chamber of Secrets (Harry Potter, #2)"
h3 = "Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)"
h4 = "Harry Potter and the Goblet of Fire (Harry Potter, #4)"
h5 = "Harry Potter and the Order of the Phoenix (Harry Potter, #5)"
h6 = "Harry Potter and the Half-Blood Prince (Harry Potter, #6)"
h7 = "Harry Potter and the Deathly Hallows (Harry Potter, #7)"
row_number = []
for row in range(1153) :
    if DF.iloc[row,10]==h1 or DF.iloc[row,10] ==h2 or DF.iloc[row,10]==h3 or DF.iloc[row,10]==h4 or DF.iloc[row,10]==h5 or DF.iloc[row,10]==h6 or DF.iloc
        row_number.append(row+1)
print(row_number)
```

```
[2, 7, 9, 10, 11, 12, 13]
```

Find the most selling books within the Harry Potter series.

```
#Find the most selling books within the Harry Potter series
books_count= {}
for i in row_number:
    books_count_list[DF.iloc[i-1,10]] = DF.iloc[i-1,4]
print(books_count_list)
```

```
{"Harry Potter and the Sorcerer's Stone (Harry Potter, #1)": 491, 'Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)': 376, 'Harry Potter and t
he Order of the Phoenix (Harry Potter, #5)': 307, 'Harry Potter and the Chamber of Secrets (Harry Potter, #2)': 398, 'Harry Potter and the Goblet of Fire
(Harry Potter, #4)': 332, 'Harry Potter and the Deathly Hallows (Harry Potter, #7)': 263, 'Harry Potter and the Half-Blood Prince (Harry Potter, #6)': 27
5}
```

```
max_key = max(books_count_list, key=books_count_list.get)
max_value = books_count_list[max_key]
print("Key with maximum value:", max_key)
print("Maximum value:", max_value)
```

```
Key with maximum value: Harry Potter and the Sorcerer's Stone (Harry Potter, #1)
Maximum value: 491
```

Calculate the average rating of the Harry Potter books.

```
#Calculate the average rating of the Harry Potter books
average_rating_list = []
for i in row_number :
    average_rating_list.append(DF.iloc[i-1,12])
print("average_rating_list",average_rating_list)
sum_of_average_rating=0
for i in average_rating_list:
    sum_of_average_rating=sum_of_average_rating+i
print("average_rating = ",sum_of_average_rating/7)
```

```
average_rating_list [4.44, 4.53, 4.46, 4.37, 4.53, 4.61, 4.54]
average_rating =  4.497142857142857
```