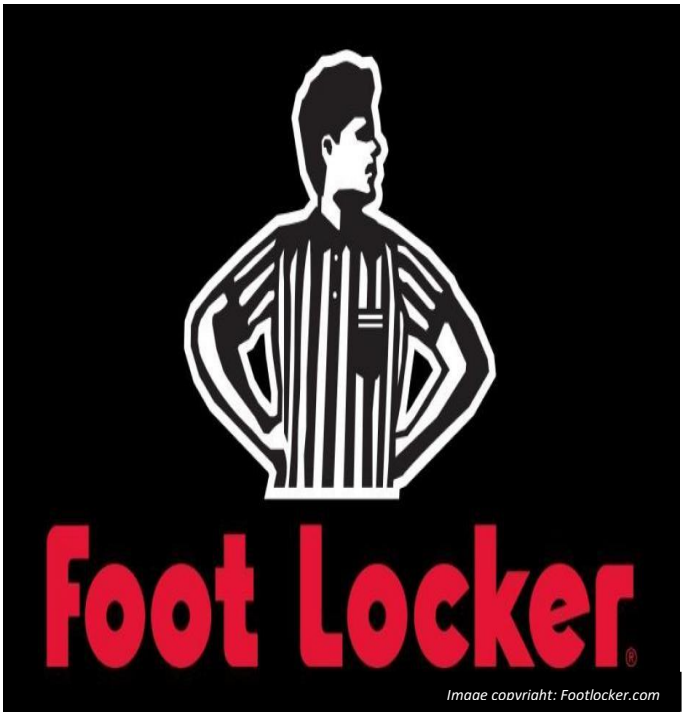


FOOT LOCKER CODE CHALLENGE

Using python(pandas)

ABSTRACT

This document discusses the solution for the problems given in Footlocker challenge. The code for the problems is in part1.ipynb and part2.ipynb included in the Zip file.

Declaration: *The work produced below is the sole work of author*

Ahmad, Faisal (ahmadfl)

Table of Contents

Part 1

Below is the output of the count of occurrences of product attribute by customer.

Output

```
In [119]: dfC
```

Out[119]:

tags	chocolate	filled	glazed	sprinkles
customer				
A	3	1	0	2
B	2	2	0	0
C	1	0	2	1

The code for this question is saved in part1.ipynb jupyter notebook

Bonus Question

Pandas don't have any issues working with small data size. However, if the size of the file grows in Gigabytes there can be issues with the run time of file, sometimes the run may fail due to insufficient memory.

With Big Data

So, to avoid such issues, we may leverage Big data technologies like Spark and Hive which use distributed computing.

I will employ Map reduce technique to divide my data into various file partitions and run the code parallelly.

With pandas

To leverage data analysis power of pandas on a medium to large file with millions of rows, we must perform memory optimization on our file in various ways and reduce the size of our file by over 60% e.g.

1. **Mapping**

Map categorical values in tags to numerical values which will reduce the memory usage by almost 50% as the string consumes around 54 bytes and int consumes only 28 bytes

Tags-string	Tags-numeric
Chocolate	1
Glazed	2
Filled	3
Sprinkles	4

2. **Downcasting**

convert large data types like floats to smaller types like int.

This would also decrease the size of variables which would decrease the memory consumption by

Data Type	Size
Int	16
Float	16
Float32	32
Float64	64

Part 2

The code for this question is saved in part2.ipynb jupyter notebook

Let's start with looking at the data, we can see that the data has 3 columns

1. Image
2. Visitors
3. Purchases

There are 4 images viz. A, B, C, D posted on a website for which the number of visitors and purchases have been posted.

	image	visitors	purchases
0	A	21	3
1	B	180	30
2	C	250	50
3	D	100	15

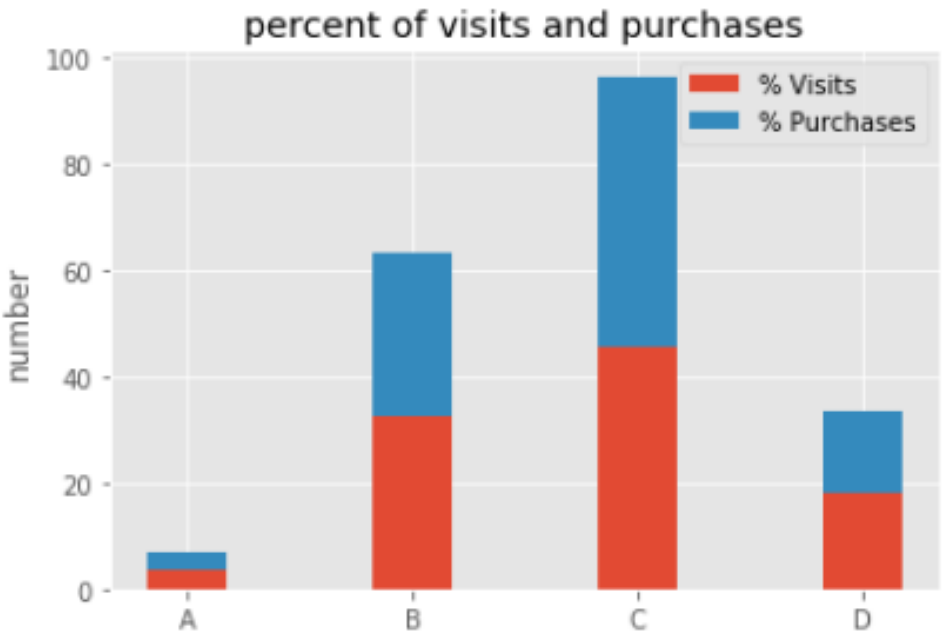
We can see that the Image C has very high number visitors and purchases. To identify the percent of visitors and purchases for each image lets create two new columns series

Percent_visitors which has the percent of visitors for an image among all the images

Percent_purchases which has the percent of purchases for an image among all the images

	image	visitors	purchases	percent_visitors	percent_purchases
0	A	21	3	3.811252	3.061224
1	B	180	30	32.667877	30.612245
2	C	250	50	45.372051	51.020408
3	D	100	15	18.148820	15.306122

From the values we can see that the Image c has the highest number of percent_visitors =45.37 and percent_purchases=51.02 same can be seen from a stacked bar graph

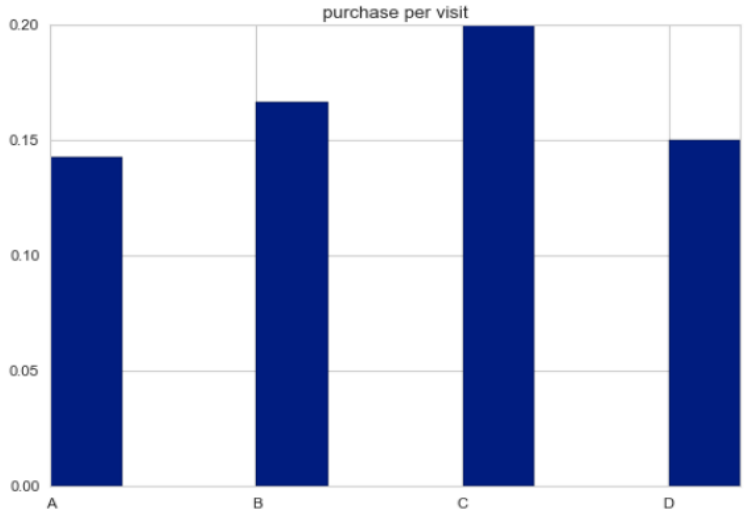


It can be clearly seen that image C has a very good response for in terms of visitors and purchases.

Also, we can calculate purchases per visit from the given data and lets analyze based on that which image has the highest fraction of purchase /visits

	image	visitors	purchases	percent_visitors	percent_purchases	purchase/visit
0	A	21	3	3.811252	3.061224	0.142857
1	B	180	30	32.667877	30.612245	0.166667
2	C	250	50	45.372051	51.020408	0.200000
3	D	100	15	18.148820	15.306122	0.150000

We can see from the column purchase/visit that C has the highest purchase/visit value compared to all the other images. Same can be visualized in a bar graph



Bonus Question:

Same as the part 1 if the given data is in millions of rows, I would prefer to use big data technologies like Hive on Hadoop and Spark on Hadoop, also I would try to downsize the memory requirements of the selected data frame by selecting the visitors and purchases based on some minimum threshold values.