APPROACH

1)READIND DATA

User tab

1)read data using pandas

2)convert signup date to date format and found userid have 34050 unique

3)check for non null

Visitor Data

1)read data(reading only required columns to optimiise code visitdatetime and signup date will take care of time diff )

2)after intial eda found out visitdatetime,product id ,userid and activity has nulls

Also activity has 4 unique values and we need only 2

3)since we need only registered persons and it had 90 percent nulls I droped it(we don’t drop nulls before eda and missing value but here I did to make my code efficient because was working on my old system and had less computation )

EDA

**1)Activity**

We new there was disperancy and changed var to lower

For missing value converted to pageloads as if not clicked must have been browsing

**2)Product Id**

After seeing found out difference in casing converted all to camelCase/titlecase

Filled na with Product101 given in problem

**3)OS**

Converted all to lower

Did rare value encoding as one var is in abundance(tol kept at 1 percent)

**4)VisitdateTime**

Outlier handling in time

After internet research found out first ecomm happened in 1994 and we need data upto 2018-05-28

Made a var date (for vintage calc.)

Converted date to date time format

For vintage made a var as today(i.e last day of month( 2018-05-28) subract from dates and all will come in dates

User\_vintage :groupby filtered data after outlier handling data by userid and find max vintage var

Combined that data with sample submission file as it has user id combine on left

For na values filled by 0 as user has never visited

For other variables filter all datset for confined month

Feature table for other variables

**Most\_Active\_OS: most active os is found by grouping dataset by user id and finding max of os:filling na with most frequent value**

**Recently\_Viewed\_Product:filter dataset on Activity that’s equal pageloads and grouping on userid then finding max on product id :filling na with Product101**

**Now rest of variables are for 15 days filter df for 15 days**

**No\_Of\_Products\_Viewed\_15\_Days:group by user id on product id find count:filling na with 0 that is zero products viewed**

**Most\_Viewed\_product\_15\_Days: filter activity equals pageloads grou on userid and find max()will give product with max frq.fillling na with Product 101 i.e:null**

**Now rest of var are for 7 days filter it for 7 days**

**No\_of\_days\_Visited\_7\_Days:groupby usr id and dates find count of dates**

**Pageloads\_last\_7\_days** :**filter for pageloads find count of pageloads filling na with 0 that is zero Pageloads**

**Clicks\_last\_7\_days:filter by clicks find count of clicks filling na with 0 that is zero clicks**