**Python Script**

import pandas as pd

from os import (environ, path)

PERCENTAGE\_OF\_IPS = 0.15

TRAIN\_FILE = 'train.csv.zip'

OUTPUT\_SAMPLE\_FILE = 'train\_sample\_15\_pct.csv.zip'

print('Reading full train data')

df = pd.read\_csv(TRAIN\_FILE)

sample\_ips = pd.Series(df.ip.unique()).sample(frac=PERCENTAGE\_OF\_IPS, random\_state=5)

#  filter df

sample\_df = df[df.ip.isin(sample\_ips)]

sample\_df = df[df.ip.isin(sample\_ips)]

percentage\_of\_rows\_taken = len(sample\_df)/float(len(df))

percentage\_of\_memory = sample\_df.memory\_usage(deep=True, index=True).sum()/float(df.memory\_usage(deep=True, index=True).sum())

print('%.2f%% of rows were taken, %.2f%% of IPs with full click history, %.2f%% of total file size'

     % (100 \* percentage\_of\_rows\_taken, 100 \* PERCENTAGE\_OF\_IPS, 100 \* percentage\_of\_memory))

OUTPUT\_SAMPLE\_FILE = 'train\_sample\_15\_pct\_full\_history.csv'

print('Writing to file: %s' % OUTPUT\_SAMPLE\_FILE)

sample\_df.to\_csv(OUTPUT\_SAMPLE\_FILE, index=False, header=True)