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)
```