Analytics using ClickStream Data

Path Optimization

Niranjan Kumar

4/30/18

BigData Expert

CASE STUDY

Input Data:

Here’s a summary of the data we’re working with:

- Omniture logs – website log files containing information such as URL, timestamp, IP address, geocoded IP address, and user ID (SWID)
    - The Omniture log dataset contains about 4 million rows of data, which represents five days of clickstream data. Often, organizations will process weeks, months, or even years of data.

- Users– CRM user data (registered Users) listing SWIDs (Software User IDs) along with date of birth and gender.

- Products – CMS data that maps product categories to website URLs.

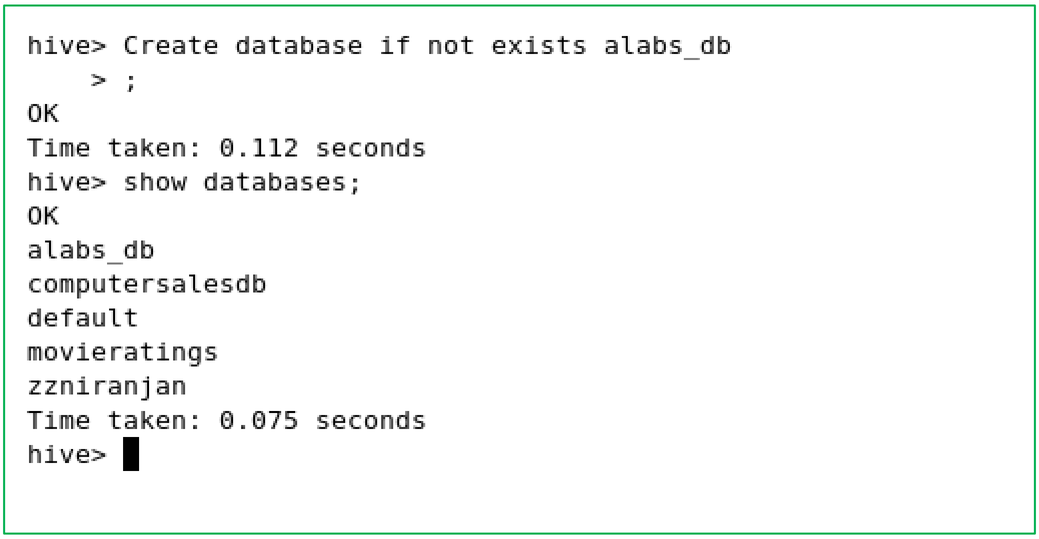
Tools Used:

- Hive – To perform the data analytics.
- Excel – To perform the data visualization.

Creating Tables

Steps:

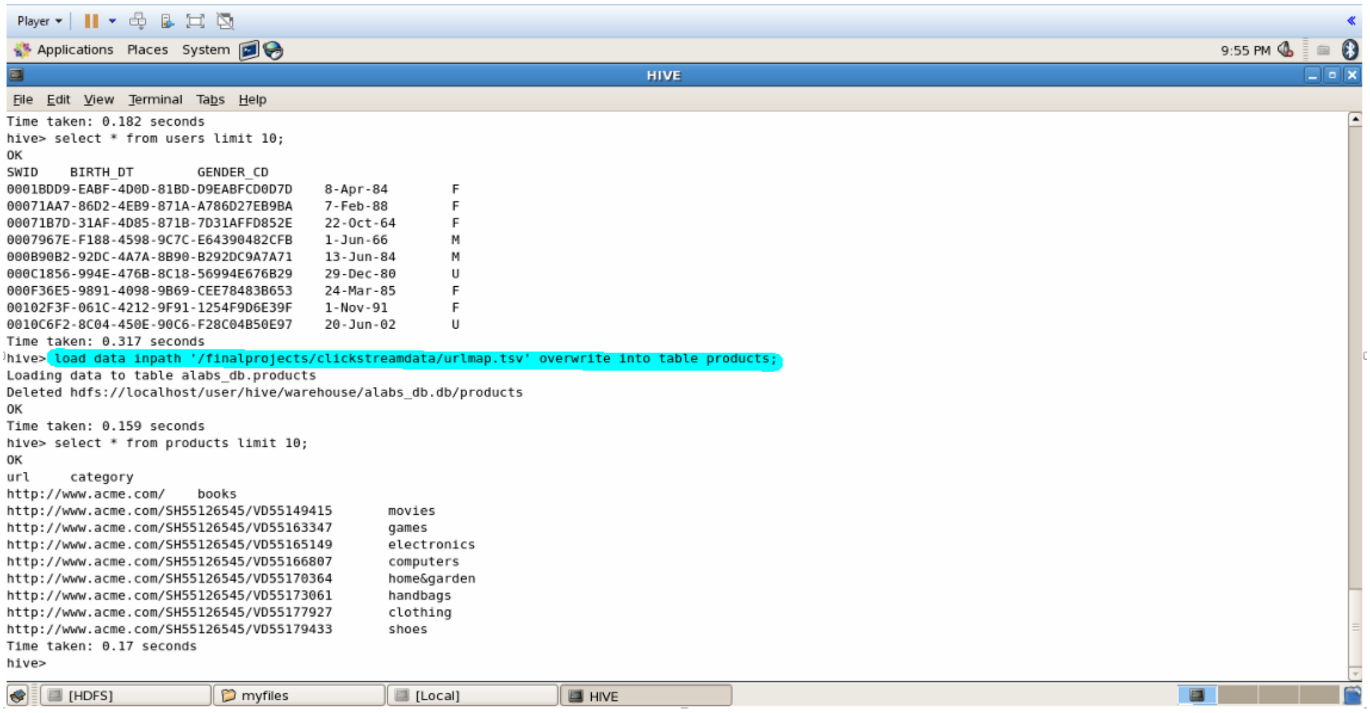
- Copy local data into Hadoop using command: Hadoop fs –put localPath HadoopPath
- Open Hive shell
- Create a Database alabs_db. CREATE DATABASE OF NOT EXISTS) alabs_db.



```
hive> Create database if not exists alabs_db
    > ;
OK
Time taken: 0.112 seconds
hive> show databases;
OK
alabs_db
computersalesdb
default
movieratings
zzniranjan
Time taken: 0.075 seconds
hive>
```

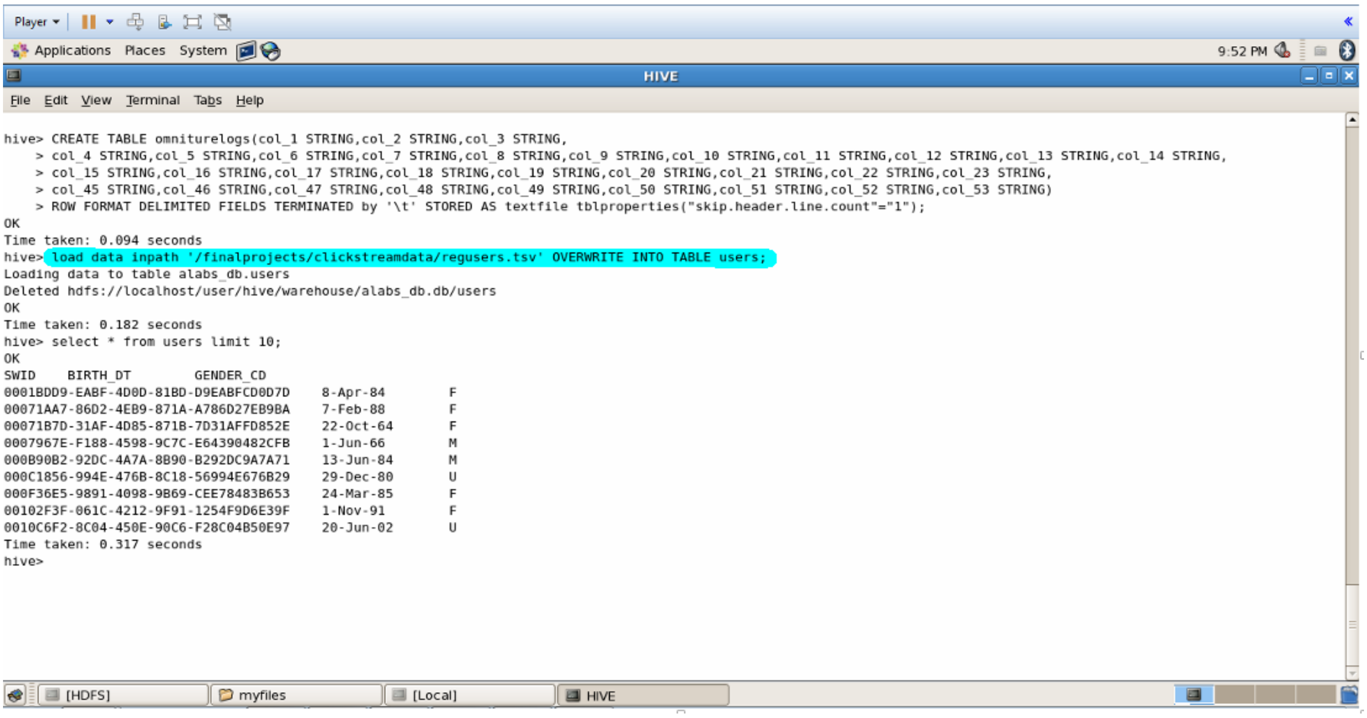
Figure 1: Creating the Database

- Now add tables into it.
- Now create the table schema for Users, Products and omniturelogs in hive



```
Time taken: 0.182 seconds
hive> select * from users limit 10;
OK
SWID    BIRTH_DT        GENDER_CD
0001BDD9-EABF-4D0D-81BD-D9EABFCD0D7D    8-Apr-84        F
00071AA7-86D2-4EB9-871A-A786D27EB9BA    7-Feb-88        F
00071B7D-31AF-4D85-871B-7D31AFFD852E    22-Oct-64       F
0007967E-F188-4598-9C7C-E64390482CFB    1-Jun-66        M
000B90B2-92DC-4A7A-8B90-B292DC9A7A71    13-Jun-84       M
000C1856-994E-476B-8C18-56994E676B29    29-Dec-80       U
000F36E5-9891-4098-9B69-CEE78483B653    24-Mar-85       F
00102F3F-061C-4212-9F91-1254F9D6E39F    1-Nov-91        F
0010C6F2-8C04-450E-90C6-F28C04B50E97    20-Jun-02       U
Time taken: 0.317 seconds
hive> load data inpath '/finalprojects/clickstreamdata/urlmap.tsv' overwrite into table products;
Loading data to table alabs_db.products
Deleted hdfs://localhost/user/hive/warehouse/alabs_db.db/products
OK
Time taken: 0.159 seconds
hive> select * from products limit 10;
OK
url     category
http://www.acme.com/    books
http://www.acme.com/SH55126545/VD55149415       movies
http://www.acme.com/SH55126545/VD55163347       games
http://www.acme.com/SH55126545/VD55165149       electronics
http://www.acme.com/SH55126545/VD55166807       computers
http://www.acme.com/SH55126545/VD55170364       home&garden
http://www.acme.com/SH55126545/VD55173061       handbags
http://www.acme.com/SH55126545/VD55177927       clothing
http://www.acme.com/SH55126545/VD55179433       shoes
Time taken: 0.17 seconds
hive>
```

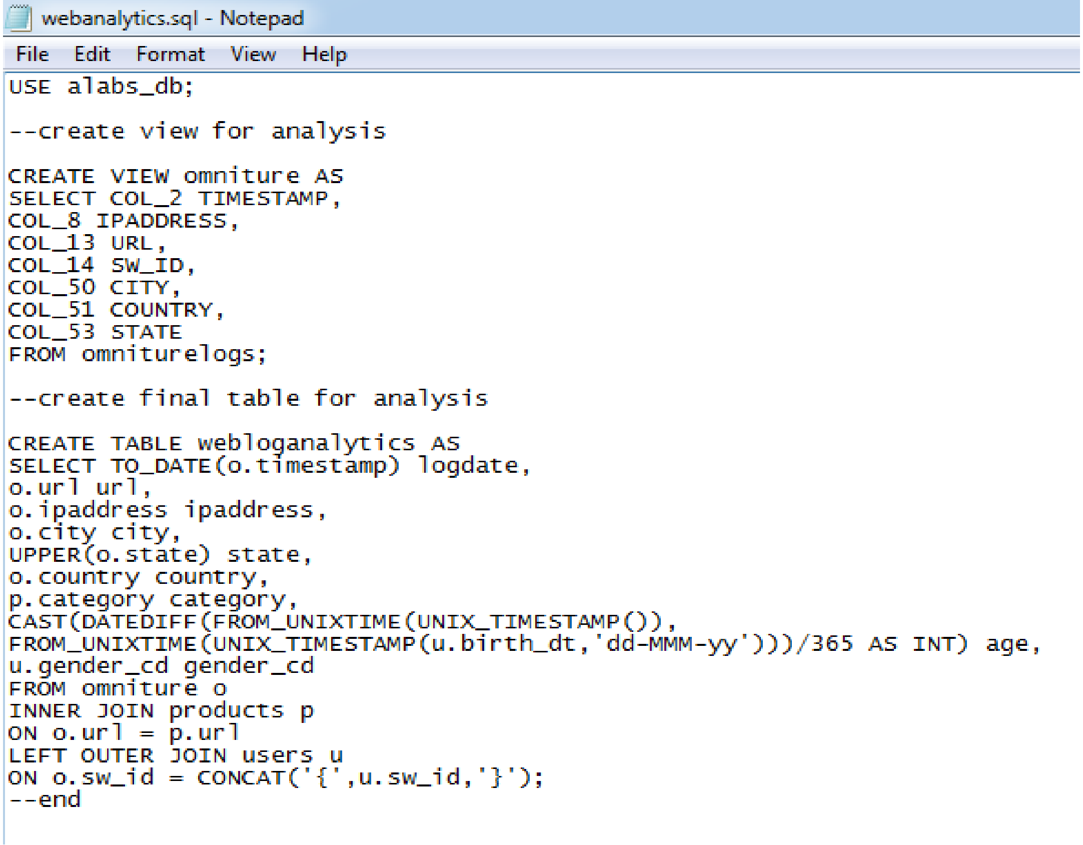
Figure 2: Loading Data into Products



```
hive> CREATE TABLE omniturelogs(col_1 STRING,col_2 STRING,col_3 STRING,
    > col_4 STRING,col_5 STRING,col_6 STRING,col_7 STRING,col_8 STRING,col_9 STRING,col_10 STRING,col_11 STRING,col_12 STRING,col_13 STRING,col_14 STRING,
    > col_15 STRING,col_16 STRING,col_17 STRING,col_18 STRING,col_19 STRING,col_20 STRING,col_21 STRING,col_22 STRING,col_23 STRING,
    > col_45 STRING,col_46 STRING,col_47 STRING,col_48 STRING,col_49 STRING,col_50 STRING,col_51 STRING,col_52 STRING,col_53 STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED by '\t' STORED AS textfile tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.094 seconds
hive> load data inpath '/finalprojects/clickstreamdata/regusers.tsv' OVERWRITE INTO TABLE users;
Loading data to table alabs_db.users
Deleted hdfs://localhost/user/hive/warehouse/alabs_db.db/users
OK
Time taken: 0.182 seconds
hive> select * from users limit 10;
OK
SWID    BIRTH_DT        GENDER_CD
0001BDD9-EABF-4D0D-81BD-D9EABFCD0D7D    8-Apr-84        F
00071AA7-86D2-4EB9-871A-A786D27EB9BA    7-Feb-88        F
00071B7D-31AF-4D85-871B-7D31AFFD852E    22-Oct-64       F
0007967E-F188-4598-9C7C-E64390482CFB    1-Jun-66        M
000B90B2-92DC-4A7A-8B90-B292DC9A7A71    13-Jun-84       M
000C1856-994E-476B-8C18-56994E676B29    29-Dec-80       U
000F36E5-9891-4098-9B69-CEE78483B653    24-Mar-85       F
00102F3F-061C-4212-9F91-1254F9D6E39F    1-Nov-91        F
0010C6F2-8C04-450E-90C6-F28C04B50E97    20-Jun-02       U
Time taken: 0.317 seconds
hive>
```

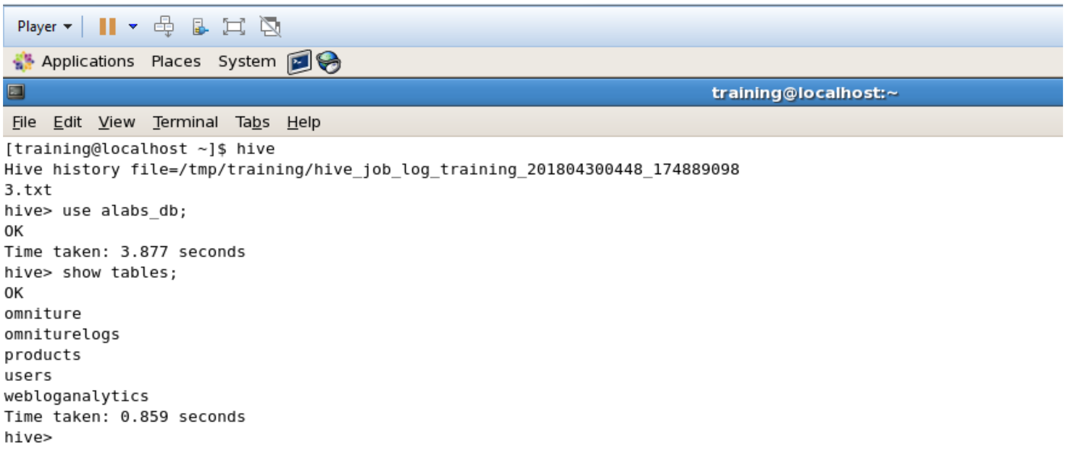
Figure 3: Loading Data into Users

Create View and final table for analysis



```
USE alabs_db;

--create view for analysis

CREATE VIEW omniture AS
SELECT COL_2 TIMESTAMP,
COL_8 IPADDRESS,
COL_13 URL,
COL_14 SW_ID,
COL_50 CITY,
COL_51 COUNTRY,
COL_53 STATE
FROM omniturelogs;

--create final table for analysis

CREATE TABLE webloganalytics AS
SELECT TO_DATE(o.timestamp) logdate,
o.url url,
o.ipaddress ipaddress,
o.city city,
UPPER(o.state) state,
o.country country,
p.category category,
CAST(DATEDIFF(FROM_UNIXTIME(UNIX_TIMESTAMP()),
FROM_UNIXTIME(UNIX_TIMESTAMP(u.birth_dt,'dd-MMM-yy')))/365 AS INT) age,
u.gender_cd gender_cd
FROM omniture o
INNER JOIN products p
ON o.url = p.url
LEFT OUTER JOIN users u
ON o.sw_id = CONCAT('{',u.sw_id,'}');
--end
```

#Omniture VIEW and Webloganalytics TABLE created

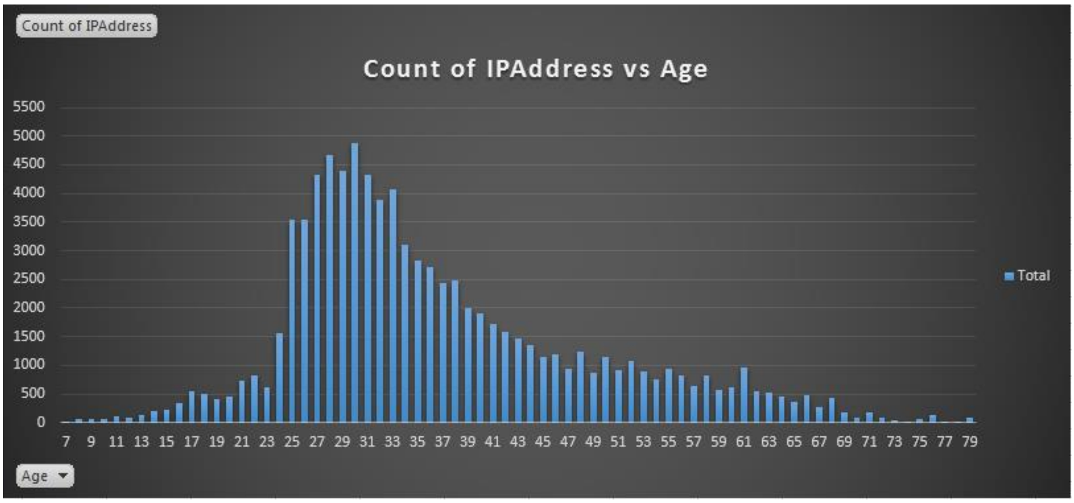


```
[training@localhost ~]$ hive
Hive history file=/tmp/training/hive_job_log_training_201804300448_174889098
3.txt
hive> use alabs_db;
OK
Time taken: 3.877 seconds
hive> show tables;
OK
omniture
omniturelogs
products
users
webloganalytics
Time taken: 0.859 seconds
hive>
```

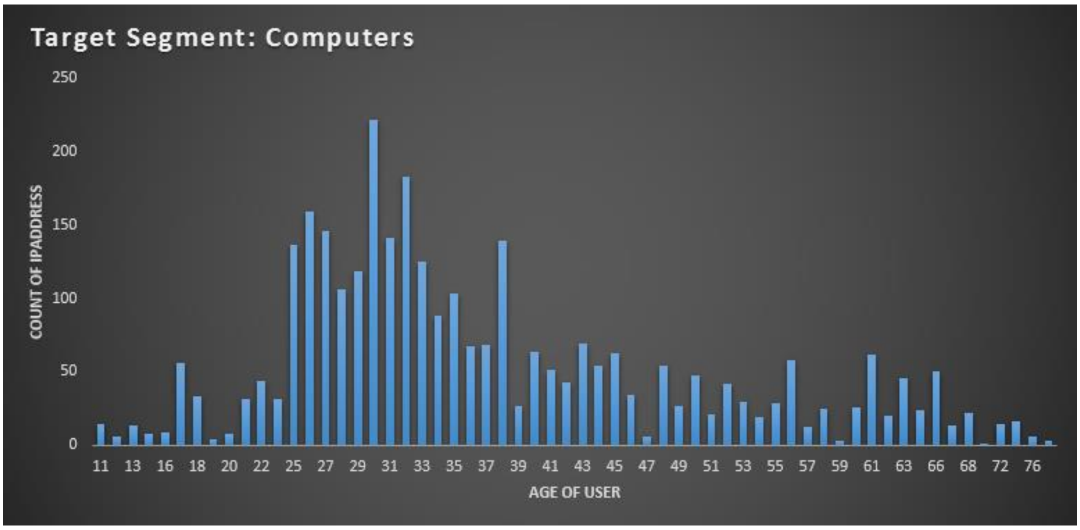
Export the data from ‘Webloganalytics’ to csv file for Visual Analytics.

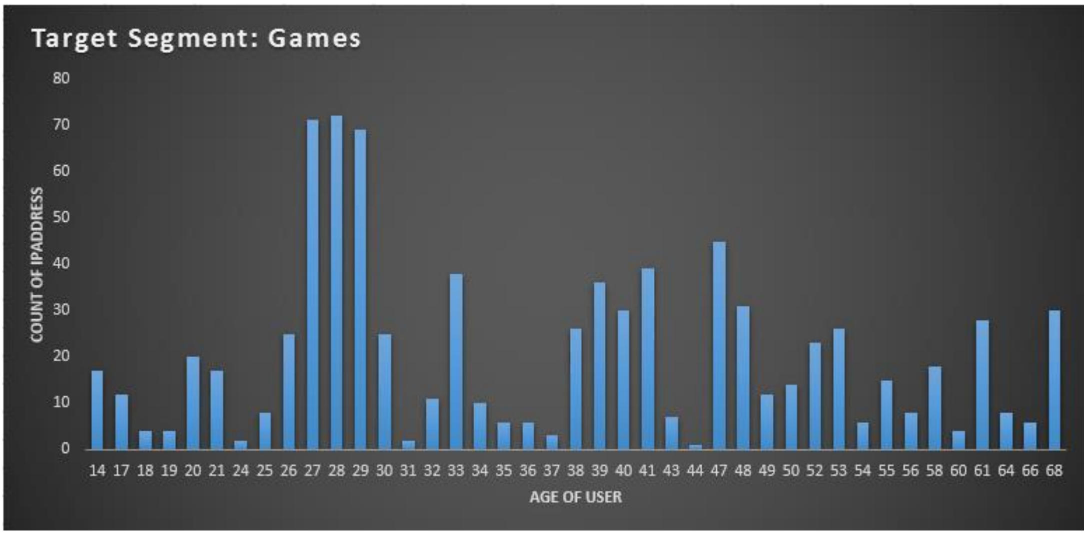
Analysis

From the Output file we can analyse the website usage based on IpAddress against the age of the user.

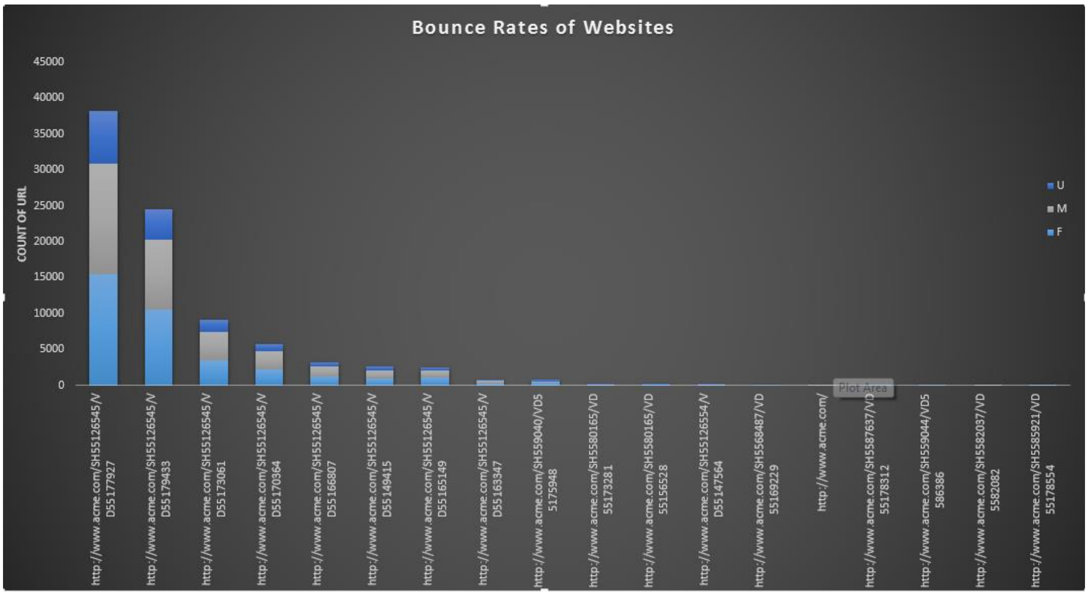


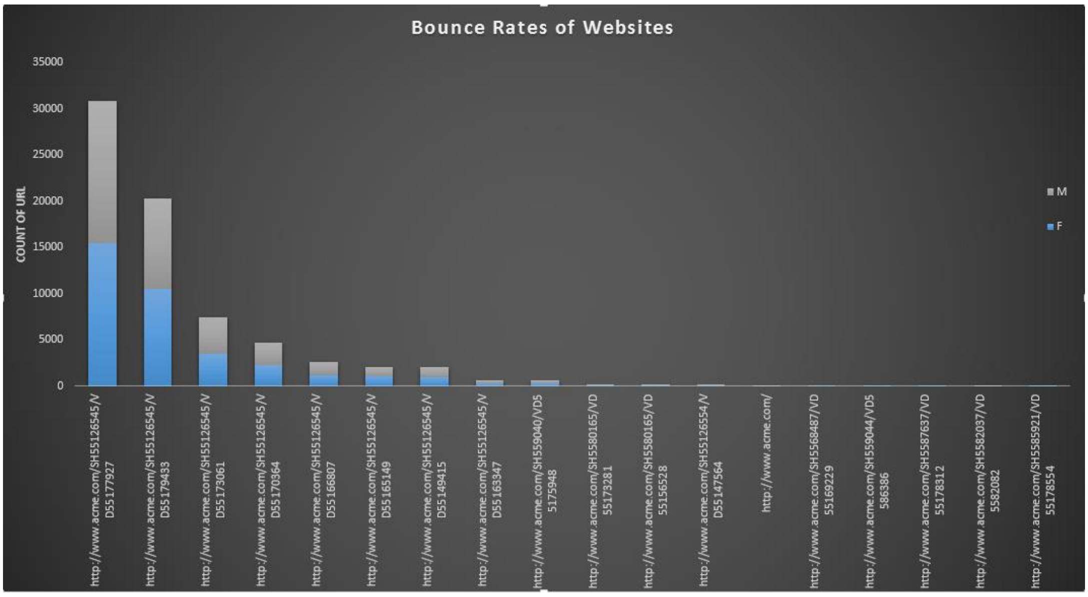
Identifying the Customer behaviour per target segment – Computers and Games





Identifying the webpages with highest bounce rates


Bounce Rates of Websites
COUNT OF URL
U
M
F
Plot Area
http://www.acme.com/SH55126545/VD55177927
http://www.acme.com/SH55126545/VD55179433
http://www.acme.com/SH55126545/VD55173061
http://www.acme.com/SH55126545/VD55170364
http://www.acme.com/SH55126545/VD55166807
http://www.acme.com/SH55126545/VD55149415
http://www.acme.com/SH55126545/VD55165149
http://www.acme.com/SH55126545/VD55163347
http://www.acme.com/SH559040/VD55175948
http://www.acme.com/SH5580165/VD55173281
http://www.acme.com/SH5580165/VD55156528
http://www.acme.com/SH55126554/VD55147564
http://www.acme.com/SH5568487/VD55169229
http://www.acme.com/
http://www.acme.com/SH5587637/VD55178312
http://www.acme.com/SH559044/VD5586386
http://www.acme.com/SH5582037/VD5582082
http://www.acme.com/SH5585921/VD55178554


Bounce Rates of Websites
COUNT OF URL
M
F
http://www.acme.com/SH55126545/VD55177927
http://www.acme.com/SH55126545/VD55179433
http://www.acme.com/SH55126545/VD55173061
http://www.acme.com/SH55126545/VD55170364
http://www.acme.com/SH55126545/VD55166807
http://www.acme.com/SH55126545/VD55165149
http://www.acme.com/SH55126545/VD55149415
http://www.acme.com/SH55126545/VD55163347
http://www.acme.com/SH559040/VD55175948
http://www.acme.com/SH5580165/VD55173281
http://www.acme.com/SH5580165/VD55156528
http://www.acme.com/SH55126554/VD55147564
http://www.acme.com/
http://www.acme.com/SH5568487/VD55169229
http://www.acme.com/SH559044/VD5586386
http://www.acme.com/SH5587637/VD55178312
http://www.acme.com/SH5582037/VD5582082
http://www.acme.com/SH5585921/VD55178554