Capstone Project

Case Study - 1

Email Marketing Campaign

Objective:
Analyze Email Marketing Campaigns of a Magazine Publisher .

Data Availability:
- Data available in the form of a csv file (/home/data/CampaignData_full.csv).
- Data for 2010 and 2011.

Data Dictionary:
- Solicitation history and outcome
- Solicitation details
- Demographics information about the individual being solicited
- Household information for the individual solicited

Reports Requirements

1. Find the Click to Open Rate (CTOR)

A. Overall CTOR (use CLICK_FLG and OPEN_FLG column)
B. CTOR by Gender (use I1_GNDR_CODE column)
C. CTOR by Time of the day (use mailed_date column)
D. CTOR by Day of the week (use mailed_date column)
E. CTOR by Month (use mailed_date column)
F. CTOR by Lead's Income Group (use TRW_INCOME_CD_V4 column)
G. CTOR by Lead's Ethnicity (use ASIAN_CD column)
H. CTOR by Lead's Household Status (use I1_INDIV_HHLD_STATUS_CODE column)

This information should be represented in Tableau/Power BI charts (bar/pie/anything relevant) which should then be shown on a dashboard.

Reports Requirements

2. Household Members Information

A. Find count of leads with information about members of their household. If a lead has information about 3 members, and another has information about 2 members and another has none, then the answer to this question is 2. (use statcd_hh_mem1 - statcd_hh_mem8 column)

B. Find count of total number of household members information is available. For example, if a lead has 3 household members, and another has about 2 members, and the other has none, then the total count of household members is 3+2+0 = 5.

C. Find count of household members by type (Head of Household, Spouse etc.).

D. %age of household members type. For example, if there are 5 Head of Household, 10 Spouse and 85 in the other categories, then the %age of Spouses is 10.

2. Household Members Information

E. How many known households have children?
(use PRESENCE_OF_CHLDRN column)

F. Overall, how many children are there?
(use NUMBER_OF_CHLDRN_18_OR_LESS column and PRESENCE_OF_CHLDRN)

G. How many of the children are male and how many are female?
(use GNDR_OF_CHLDRN_0_3 - GNDR_OF_CHLDRN_13_18 column)

Workflow Requirements

- Data flow from source to be copied to HDFS
- Data from HDFS to be loaded to Pig for filtering and transformations
- Final output from Pig to be stored in HDFS
- Data from HDFS to be loaded to Hive for finding solutions for above mentioned problem in earlier slide
- Connect to tableau
  <IP: 52.4.16.124 need to connect to 54.174.252.76 for hiveserver2>
  <IP: 52.3.237.208 need to connect to 52.3.237.208 for hiveserver2>
- Create Dashboard

Solution

Data flow from source copied to hdfs

```
hduser@vinod-virtual-machine:~$ hadoop fs -mkdir /user/Jig14696/FinalCaseStudy
hduser@vinod-virtual-machine:~$ hadoop fs -copyFromLocal CampaignData_sample.csv /user/Jig14696/F
inalCaseStudy/
hduser@vinod-virtual-machine:~$ hadoop fs -ls /user/Jig14696/FinalCaseStudy
Found 1 items
-rw-r--r--   1 hduser supergroup       8619 2017-12-04 18:40 /user/Jig14696/FinalCaseStudy/Campai
gnData_sample.csv
hduser@vinod-virtual-machine:~$
```

Data from hdfs loaded to pig

```
grunt> campaign = load '/user/Jig14696/FinalCaseStudy/CampaignData_sample.csv' AS (CLICK_FLG,OPEN
_FLG,ADDR_VER_CD,AQI,ASIAN_CD,AUTO_IN_MARKET,BIRD_QTY,BUYER_DM_BOOKS,BUYER_DM_COLLECT_SPC_FOOD,BU
YER_DM_CRAFTS_HOBBI,BUYER_DM_FEMALE_ORIEN,BUYER_DM_GARDEN_FARM,BUYER_DM_GENERAL,BUYER_DM_GIFT_GAD
GET,BUYER_DM_MALE_ORIEN,BUYER_DM_UPSCALE,BUYER_MAG_CULINARY_INTERS,BUYER_MAG_FAMILY_GENERAL,BUYER
_MAG_FEMALE_ORIENTED,BUYER_MAG_GARDEN_FARMING,BUYER_MAG_HEALTH_FITNESS,BUYER_MAG_MALE_SPORT_ORIEN
TED,BUYER_MAG_RELIGIOUS,CATS_QTY,CEN_2000_MATCH_LEVEL,CLUB_MEMBER_CD,COUNTRY_OF_ORIGIN,DECEASED_I
NDICATOR,DM_RESPONDER_HH,DM_RESPONDER_INDIV,DMR_CONTRIB_CAT_GENERAL,DMR_CONTRIB_CAT_HEALTH_INST,D
MR_CONTRIB_CAT_POLITICAL,DMR_CONTRIB_CAT_RELIGIOUS,DMR_DO_IT_YOURSELFERS,DMR_MISCELLANEOUS,DMR_NE
WS_FINANCIAL,DMR_ODD_ENDS,DMR_PHOTOGRAPHY,DMR_SWEEPSTAKES,DOG_QTY,DWELLING_TYPE,DWELLING_UNIT_SIZ
E,EST_LOAN_VALUE_RATIO,ETECH_GROUP,ETHNIC_GROUP_CODE,ETHNIC_INSIGHT_MTCH_FLG,ETHNICITY_DETAIL,EXP
ERIAN_INCOME_CD,EXPERIAN_INCOME_CD_V4,GNDR_OF_CHLDRN_0_3,GNDR_OF_CHLDRN_10_12,GNDR_OF_CHLDRN_13_1
8,GNDR_OF_CHLDRN_4_6,GNDR_OF_CHLDRN_7_9,HH_INCOME,HHLD_DM_PURC_CD,HOME_BUSINESS_IND,I1_BUSINESS_O
WNER_FLG,I1_EXACT_AGE,I1_GNDR_CODE,I1_INDIV_HHLD_STATUS_CODE,INDIV_EDUCATION,INDIV_EDUCATION_CONF
_LVL,INDIV_MARITAL_STATUS,INDIV_MARITAL_STATUS_CONF_LVL,INS_MATCH_TYPE,LANGUAGE,LENGTH_OF_RESIDEN
CE,MEDIAN_HOUSING_VALUE,MEDIAN_LEN_OF_RESIDENCE,MM_INCOME_CD,MOSAIC_HH,MULTI_BUYER_INDIV,NEW_CAR_
MODEL,NUM_OF_ADULTS_IN_HHLD,NUMBER_OF_CHLDRN_18_OR_LESS,OCCUP_DETAIL,OCCUP_MIX_PCT,PCT_CHLDRN,PCT
_DEROG_TRADES,PCT_HOUSEHOLDS_BLACK,PCT_OWNER_OCCUPIED,PCT_RENTER_OCCUPIED,PCT_TRADES_NOT_DEROG,PC
T_WHITE,PHONE_TYPE_CD,PRES_OF_CHLDRN_0_3,PRES_OF_CHLDRN_10_12,PRES_OF_CHLDRN_13_18,PRES_OF_CHLDRN
_4_6,PRES_OF_CHLDRN_7_9,PRESENCE_OF_CHLDRN,PRIM_FEM_EDUC_CD,PRIM_FEM_OCC_CD,PRIM_MALE_EDUC_CD,PRI
M_MALE_OCC_CD,RECIPIENT_RELIABILITY_CD,RELIGION,SCS_MATCH_TYPE,TRW_INCOME_CD,TRW_INCOME_CD_V4,USE
D_CAR_CD,Y_OWNS_HOME,Y_PROBABLE_HOMEOWNER,Y_PROBABLE_RENTER,Y_RENTER,YRS_SCHOOLING_CD,Z_CREDIT_CA
RD,age_hh_mem1,age_hh_mem2,age_hh_mem3,age_hh_mem4,age_hh_mem5,age_hh_mem6,age_hh_mem7,age_hh_mem
8,gender_hh_mem1,gender_hh_mem2,gender_hh_mem3,gender_hh_mem4,gender_hh_mem5,gender_hh_mem6,gende
r_hh_mem7,gender_hh_mem8,statcd_hh_mem1,statcd_hh_mem2,statcd_hh_mem3,statcd_hh_mem4,statcd_hh_me
m5,statcd_hh_mem6,statcd_hh_mem7,statcd_hh_mem8,trait_cd1,trait_cd2,trait_cd3,trait_cd4,trait_cd5
,trait_cd6,trait_cd7,trait_cd8,trait_cd9,trait_cd10,trait_cd11,trait_cd12,trait_cd13,trait_cd14,t
rait_cd15,trait_cd16,trait_cd17,trait_cd18,trait_cd19,trait_cd20,trait_cd21,trait_cd22,trait_cd23
,trait_cd24,trait_cd25,trait_cd26,trait_cd27,trait_cd28,trait_cd29,trait_cd30,trait_cd31,trait_cd
32,trait_cd33,trait_cd34,trait_cd35,trait_cd36,trait_cd37,trait_cd38,trait_cd39,trait_cd40,trait_
cd41,trait_cd42,trait_cd43,trait_cd44,trait_cd45,trait_cd46,trait_cd47,trait_cd48,trait_cd49,trai
```

3. dump campaign;

```
grunt> dump campaign;
2017-12-04 14:07:56,497 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features use
d in the script: UNKNOWN
2017-12-04 14:07:56,509 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend
has already been initialized
2017-12-04 14:07:56,509 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimize
r - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter
, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, Pred
icatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter
]}
2017-12-04 14:07:56,510 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MRCompiler - File concatenation threshold: 100 optimistic? false
2017-12-04 14:07:56,510 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MultiQueryOptimizer - MR plan size before optimization: 1
2017-12-04 14:07:56,510 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MultiQueryOptimizer - MR plan size after optimization: 1
2017-12-04 14:07:56,529 [main] INFO  org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig
script settings are added to the job
2017-12-04 14:07:56,530 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2017-12-04 14:07:56,532 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.JobControlCompiler - Setting up single store job
2017-12-04 14:07:56,537 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MapReduceLauncher - 1 map-reduce job(s) waiting for submission.
2017-12-04 14:07:56,539 [JobControl] WARN  org.apache.hadoop.mapred.JobClient - No job jar file s
et.  User classes may not be found. See JobConf(Class) or JobConf#setJar(String).
2017-12-04 14:07:56,561 [JobControl] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat
- Total input paths to process : 1
2017-12-04 14:07:56,561 [JobControl] INFO  org.apache.pig.backend.hadoop.executionengine.util.Map
```

```
(N,N,,E,,5,0,4,0,0,1,0,1,1,1,3,1,1,0,2,4,0,0,6,B,U,,M,Y,0,0,0,0,1,2,0,0,0,0,0,S,A,1024,,,,,,E,U,
U,U,U,U,44,,U,U,52,M,H,3,1,M,1,I,,22,101,8,,B03,Y,2,5,0,02,0,32,0,2,84,15,99,93,,N,N,N,N,N,N,U,03
,U,04,1,,4,,N,9,Y,,U,U,0,Y,,52,52,23,,,,,M,F,F,M,,,,,U,U,W,Y,,,,,3,12,14,15,16,17,21,39,69,76,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,5,5,5,5,5,5,5,5,5,5,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,INMAN,293498619,SC,176058,,,,,,1036,Third Party - Merch,"Mon, 09/20/1
0 01:04 PM")
(N,N,,E,,8,0,2,0,0,0,1,1,0,0,0,0,0,0,1,1,1,0,0,B,U,0,,M,U,0,0,1,0,1,0,2,0,0,0,0,S,A,1024,K,8,Y,01
,,K,U,U,U,U,U,57,,U,U,63,M,H,2,1,M,5,I,01,08,191,4,,A02,Y,3,3,0,31,0,21,0,1,56,43,99,94,,N,N,N,N,
N,N,U,31,U,31,1,P,4,,R,7,Y,,U,U,0,Y,,64,,,,,,,F,F,,,,,,,U,U,,,,,,,1,1,3,3,4,4,5,6,7,10,11,12,13,1
4,15,16,17,19,21,28,29,30,35,37,41,43,69,75,76,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,1,3,3,5,1,3,5,1
,3,5,5,5,5,5,5,5,5,5,5,5,3,3,3,3,3,3,5,5,5,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,VIRGINIA BEACH,2345
52122,VA,162398,,,,,,1036,Third Party - Merch,"Mon, 09/20/10 01:04 PM")
(N,N,,E,,3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,B,,0,,U,U,0,0,0,0,0,0,0,0,0,0,0,S,A,1098,K,28,Y,02
,,H,U,U,U,U,U,59,,U,U,44,M,H,2,5,M,1,I,01,05,125,4,,A07,,4,3,0,00,0,45,21.5,8,90,9,74,87,,N,N,N,N
,N,U,0,00,0,00,1,P,4,,Q,6,Y,,U,U,0,Y,,46,,,,,,,U,F,,,,,,,U,W,,,,,,,19,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,5,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,GARNER,275297480,NC,164596,,,,,,1036,Third Party - Merch,"Mon, 09/20/10 01:04 PM")
(N,N,,E,,9,0,1,0,0,0,0,1,0,0,0,0,4,1,0,1,1,0,0,B,U,0,,M,Y,0,0,1,0,0,1,0,0,0,1,0,S,A,1005,J,18,Y,6
6,,E,U,U,U,U,U,53,,U,U,64,M,H,3,5,M,1,I,01,26,131,5,,B03,Y,3,3,0,02,0,34,33.4,29,65,34,58.04,59,,
N,N,N,N,N,N,U,UU,2,UU,1,J,4,,N,7,Y,,U,U,0,Y,64,29,63,,,,,,M,M,F,,,,,,H,U,W,,,,,,20,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,5,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,COLUMBIA,210441910,MD,154216,,,,,,1036,Third Party - Merch,"Mon, 09/20/10 01:04
PM")
(N,N,,,,,0,2,0,1,2,0,2,3,1,1,0,2,0,0,2,0,0,0,,,,,M,U,0,4,0,0,0,1,0,2,0,1,0,S,A,,,8,,,I,,U,U,M,U,U
,84,,,,48,F,W,4,0,M,0,I,,04,168,7,,,Y,,2,1,00,0,55,,2,97,2,,94,,N,U,Y,N,U,Y,0,00,0,00,,P,,R,,,Y,,
U,U,0,,53,,,,,,,,M,,,,,,,,H,,,,,,,,1,1,2,2,3,3,3,4,4,5,6,6,7,10,13,14,15,16,17,20,21,26,28,29,30,
36,37,39,40,41,43,44,45,65,69,77,,,,,,,,,,,,,,,,,,,,,,,,,,,,,1,3,1,3,1,3,5,1,3,5,1,3,1,5,5,5,5,1,
5,5,5,5,5,3,3,3,3,5,3,3,3,3,3,5,5,5,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,HENDERSONVLLE,287398446,NC,170075
,,,,,,1036,Third Party - Merch,"Mon, 09/20/10 01:04 PM")
grunt>
```

Filtering in pig:

```
grunt> CampaignData = FOREACH campaign GENERATE (CLICK_FLG,OPEN_FLG,ASIAN_CD,GNDR_OF_CHLDRN_0_3,G
NDR_OF_CHLDRN_13_18,I1_GNDR_CODE,NUMBER_OF_CHLDRN_18_OR_LESS,TRW_INCOME_CD_V4,age_hh_mem1,age_hh_
mem2,age_hh_mem3,age_hh_mem4,age_hh_mem5,age_hh_mem6,age_hh_mem7,age_hh_mem8,statcd_hh_mem1,statc
d_hh_mem2,statcd_hh_mem3,statcd_hh_mem4,statcd_hh_mem5,statcd_hh_mem6,statcd_hh_mem7,statcd_hh_me
m8,Mailed_date);
```

```
2017-12-05 17:16:42,696 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MapReduceLauncher - Success!
2017-12-05 17:16:42,697 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig.schematupl
e] was not set... will not generate code.
2017-12-05 17:16:42,701 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Tota
l input paths to process : 1
2017-12-05 17:16:42,701 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUti
l - Total input paths to process : 1
((CLICK_FLG,OPEN_FLG,ASIAN_CD,GNDR_OF_CHLDRN_0_3,GNDR_OF_CHLDRN_13_18,,,,,,,,,,,,,,,,,,,,))
((  _OF_CHLDRN_4_6,GNDR_OF_CHLDRN_7_9,HOME_BUSINESS_IND,Y_OWNS_HOME,Y_PROBABLE_RENTER,,,,,,,,,,,
,,,,,,,))
(( 2,age_hh_mem3,age_hh_mem6,trait_cd28,trait_cd30,trait_cd38,trait_cd54,,,,,,,,,,,,,,,,,))
(( ait_cd63,trait_cd64,trait_typ_cd3,trait_typ_cd49,trait_typ_cd51,trait_typ_cd59,,,,,,,,,,,,,,,
,,,))
(( TE_PROVINCE,new_id,new_list_id,,,,,,,,,,,,,,,,,,,,,,))
((N,N,,U,U,M,0,J,,24,,,,,,,U,Y,,,,,,,Mon, 09/20/10 01:04 PM))
((N,N,,U,F,M,1,N,,,31,28,59,,,,U,U,U,U,U,W,,,,Mon, 09/20/10 01:04 PM))
((N,N,,U,U,M,0,O,,,,,61,,,,U,U,U,U,U,Y,,,Mon, 09/20/10 01:04 PM))
((N,N,,U,U,M,0,R,60,25,23,,,,,,W,Y,Y,,,,,,Mon, 09/20/10 01:04 PM))
((N,N,,U,U,M,0,N,,52,52,23,,,,,U,U,W,Y,,,,,Mon, 09/20/10 01:04 PM))
((N,N,,U,U,M,0,R,,64,,,,,,,U,U,,,,,,,Mon, 09/20/10 01:04 PM))
((N,N,,U,U,M,0,Q,,46,,,,,,,U,W,,,,,,,Mon, 09/20/10 01:04 PM))
((N,N,,U,U,M,0,N,64,29,63,,,,,,H,U,W,,,,,,Mon, 09/20/10 01:04 PM))
((N,N,,U,M,F,1,,53,,,,,,,,H,,,,,,,,Mon, 09/20/10 01:04 PM))
grunt>
```

```
grunt> data = FOREACH CampaignData generate CLICK_FLG,OPEN_FLG,ASIAN_CD,GNDR_OF_CHLDRN_0_3,GNDR_O
F_CHLDRN_13_18,I1_GNDR_CODE,NUMBER_OF_CHLDRN_18_OR_LESS,TRW_INCOME_CD_V4,age_hh_mem1,age_hh_mem2,
age_hh_mem3,age_hh_mem4,age_hh_mem5,age_hh_mem6,age_hh_mem7,age_hh_mem8,statcd_hh_mem1,statcd_hh_
mem2,statcd_hh_mem3,statcd_hh_mem4,statcd_hh_mem5,statcd_hh_mem6,statcd_hh_mem7,statcd_hh_mem8,Ma
iled_Date,SUBSTRING(Mailed_Date,5,13) AS date,CONCAT('20',SUBSTRING(Mailed_Date,11,13)) AS year,(
SUBSTRING(Mailed_Date,5,7)=='01' ? 'January' : (SUBSTRING(Mailed_Date,5,7)=='02' ? 'February' : (
SUBSTRING(Mailed_Date,5,7)=='03' ? 'March' :(SUBSTRING(Mailed_Date,5,7)=='04' ? 'April' : (SUBSTR
ING(Mailed_Date,5,7)=='05' ? 'May' :(SUBSTRING(Mailed_Date,5,7)=='06' ? 'June' :(SUBSTRING(Mailed
_Date,5,7)=='07' ? 'July' :(SUBSTRING(Mailed_Date,5,7)=='08' ? 'August' :(SUBSTRING(Mailed_Date,5
,7)=='9' ? 'September' :(SUBSTRING(Mailed_Date,5,7)=='10' ? 'October' :(SUBSTRING(Mailed_Date,5,7
)=='11' ? 'November' :'December'))))))))))) AS month,SUBSTRING(Mailed_Date,0,3) AS day,SUBSTRING(
Mailed_Date,14,19) as TimeStamp,SUBSTRING(Mailed_Date,20,22) as AMPM;
```

dump data;

```
hduser@vinod-virtual-machine:~$ hadoop fs -cat /user/Jig14696/FinalCaseStudy/pigoutput/part-m-000
00
CLICK_FLG       |OPEN_FLG|ASIAN_CD|GNDR_OF_CHLDRN_0_3|GNDR _OF_CHLDRN_13_18|I1_GNDR_CODE|NUMBER_O
F_CHLDRN_18_OR_LESS|TRW_INCOME_CD_V4|age_hh_mem1|age_hh_mem2|age_hh_mem3|age_hh_mem4|age_hh_mem5|
age_hh_mem6|age_hh_mem7|age_hh_mem8|statcd_hh_mem1|statcd_hh_mem2|statcd_hh_mem3|statcd_hh_mem4|s
tatcd_hh_mem5|statcd_hh_mem6|statcd_hh_mem7|statcd_hh_mem8|Mailed_Date|date|year|month|day      t
imestamp|AMPM
N|N| |U|U|M|0|J| |24| | | | | | |U|Y| | | | | |Mon 09/20/10 01:04 PM|09/20/10|2010|September|Mo
n|01:04|PM
N|N| |U|F|M|1|N| | |31|28|59| | | |U|U|U|U|U|W| | | |Mon 09/20/10 01:04 PM|09/20/10|2010|September|
Mon|01:04|PM
N|N| |U|U|M|0|O| | | | |61| | | |U|U|U|U|U|Y| | |Mon 09/20/10 01:04 PM|09/20/10|2010|September|Mo
n|01:04|PM
N|N| |U|U|M|0|R|60|25|23| | | | | |W|Y|Y| | | | | |Mon 09/20/10 01:04 PM|09/20/10|2010|September|
Mon|01:04|PM
N|N| |U |U|M|0|N| |52|52|23| | | | |U|U|W|Y| | | | |Mon 09/20/10 01:04 PM|09/20/10|2010|September
|Mon|01:04|PM
N|N| |U |U|M|0|R| |64| | | | | | |U|U| | | | | | |Mon 09/20/10 01:04 PM|09/20/10|2010|September|M
on|01:04|PM
N|N| |U |U|M|0|Q| |46| | | | | | |U|W| | | | | | |Mon 09/20/10 01:04 PM|09/20/10|2010|September|M
on|01:04|PM
N|N| |U |U||M|0|N|64|29|63| | | | | |H|U|W| | | | | |Mon 09/20/10 01:04 PM|09/20/10|2010|Septembe
r|Mon|01:04|PM
N|N| |U |M|F|1| |53| | | | | | | |H| | | | | | | |Mon 09/20/10 01:04 PM|09/20/10|2010|September|M
on|01:04|PM
hduser@vinod-virtual-machine:~$
```

Store data into ‘/user/Jig14696/FinalCaseStudy/pigoutput’ USING PigStorage(‘|’);

Hive:

Create table in hive

```
hive> create table if not exists assign(click_flg string,open_flg string,asian_cd string,gndr_of_
chldrn_0_3 string,gndr_of_chldrn_13_18 string,i1_gndr_code string,number_of_chldrn_18_or_less int
,trw_income_cd_v4 string,age_hh_mem1 int,age_hh_mem2 int,age_hh_mem3 int,age_hh_mem4 int,age_hh_m
em5 int,age_hh_mem6 int,age_hh_mem7 int,age_hh_mem8 int,statcd_hh_mem1 string,statcd_hh_mem2 stri
ng,statcd_hh_mem3 string,statcd_hh_mem4 string,statcd_hh_mem5 string,statcd_hh_mem6 string,statcd
_hh_mem7 string,statcd_hh_mem8 string,mailed_date string,pres_of_chldrn string,date string,year s
tring,month string,day string,timestamp string,ampm string,I1_INDIV_HHLD_STATUS_CODE string) ROW
FORMAT DELIMITED FIELDS TERMINATED BY '|';
OK
Time taken: 0.028 seconds
hive>
```

Store data from HDFS

```
hive> LOAD DATA INPATH '/user/Jig14696/FinalCaseStudy/pigoutput' OVERWRITE INTO table assign;
Loading data to table project.assign
Deleted hdfs://localhost:54310/user/hive/warehouse/project.db/assign
Table project.assign stats: [numFiles=1, numRows=0, totalSize=1321, rawDataSize=0]
OK
Time taken: 0.171 seconds
hive>
```

select * from assign;

```
Time taken: 0.018 seconds, Fetched: 1 row(s)
hive> select * from assign;
OK
CLICK_FLG       OPEN_FLG        ASIAN_CD        GNDR_OF_CHLDRN_0_3      GNDR_OF_CHLDRN_13_18    I
1_GNDR_CODE     NULL    TRW_INCOME_CD_V4        NULL    NULL    NULL    NULL    NULL    NULL    N
ULL     NULL    statcd_hh_mem1  statcd_hh_mem2  statcd_hh_mem3  statcd_hh_mem4  statcd_hh_mem5  s
tatcd_hh_mem6   statcd_hh_mem7  statcd_hh_mem8  mailed_date     PRES_OF_CHLDRN  date    year    m
onth    day     timestamp       ampm    I1_INDIV_HHLD_STATUS_CODE
N       N               U       U       M       0       J       NULL    24      NULL    NULL    N
ULL     NULL    NULL    NULL    U       Y                                                       M
on, 09/20/10 01:04 PM   N       09/20/10        2017    Mon     01:04   PM      H       NULL
N       N               U       F       M       1       N       NULL    NULL    31      28      5
9       NULL    NULL    NULL    U       U       U       U       W                               M
on, 09/20/10 01:04 PM   Y       09/20/10        2017    Mon     01:04   PM      H       NULL
N       N               U       U       M       0       O       NULL    NULL    NULL    NULL    6
1       NULL    NULL    NULL    U       U       U       U       U       Y                       M
on, 09/20/10 01:04 PM   N       09/20/10        2017    Mon     01:04   PM      H       NULL
N       N               U       U       M       0       R       60      25      23      NULL    N
ULL     NULL    NULL    NULL    W       Y       Y                                               M
on, 09/20/10 01:04 PM   N       09/20/10        2017    Mon     01:04   PM      H       NULL
N       N               U       U       M       0       N       NULL    52      52      23      N
ULL     NULL    NULL    NULL    U       U       W       Y                                       M
on, 09/20/10 01:04 PM   N       09/20/10        2017    Mon     01:04   PM      H       NULL
N       N               U       U       M       0       R       NULL    64      NULL    NULL    N
ULL     NULL    NULL    NULL    U       U                                                       M
on, 09/20/10 01:04 PM   N       09/20/10        2017    Mon     01:04   PM      H       NULL
N       N               U       U       M       0       Q       NULL    46      NULL    NULL    N
ULL     NULL    NULL    NULL    U       W                                                       M
```

Find the Click to Open ratio(CTOR)

A.Overall CTOR

```
hive> create table ratio as SELECT ROUND(SUM(CASE WHEN CLICK_FLG='N' THEN 1 ELSE 0 END)
    > /count(*)*100,2) CTOR FROM assign WHERE OPEN_FLG='N';
Total jobs = 1
Launching Job 1 out of 1
```

```
hive> select * from ratio;
OK
100.0
Time taken: 0.103 seconds, Fetched: 1 row(s)
hive>
```

B. CTOR by gender

```
hive> create table gender as select I1_GNDR_CODE,ROUND(SUM(CASE WHEN CLICK_FLG='N' THEN 1 ELSE 0
END)/count(*)*100,2) CTOR FROM assign WHERE OPEN_FLG='N' GROUP BY I1_GNDR_CODE;
Total jobs = 1
Launching Job 1 out of 1
```

```
hive> select * from gender;
OK
        100.0
F       100.0
M       100.0
Time taken: 0.022 seconds, Fetched: 3 row(s)
hive>
```

C.CTOR by time of the day

```
hive> create table time as select time,ROUND(SUM(CASE WHEN CLICK_FLG='N' THEN 1 ELSE 0 END)/count
(*)*100,2) CTOR FROM assign WHERE OPEN_FLG='N' GROUP BY time;
Total jobs = 1
Launching Job 1 out of 1
```

```
hive> select * from time;
OK
01:04   100.0
Time taken: 0.02 seconds, Fetched: 1 row(s)
hive>
```

D.CTOR by day of the week

```
hive> create table day as select day,ROUND(SUM(CASE WHEN CLICK_FLG='N' THEN 1 ELSE 0 END)/count(*
)*100,2) CTOR FROM assign WHERE OPEN_FLG='N' GROUP BY day;
Total jobs = 1
Launching Job 1 out of 1
```

```
hive> select * from day;
OK
Mon     100.0
Time taken: 0.015 seconds, Fetched: 1 row(s)
hive>
```

E.CTOR by month

```
hive> create table month as select month,ROUND(SUM(CASE WHEN CLICK_FLG='N' THEN 1 ELSE 0 END)/cou
nt(*)*100,2) CTOR FROM assign WHERE OPEN_FLG='N' GROUP BY month;
Total jobs = 1
Launching Job 1 out of 1
```

```
hive> select * from month;
OK
September       100.0
Time taken: 0.019 seconds, Fetched: 1 row(s)
hive>
```

F.CTOR by leads income group

```
hive> create table lead as select TRW_INCOME_CD_V4,ROUND(SUM(CASE WHEN CLICK_FLG='N' THEN 1 ELSE
0 END)/count(*)*100,2) CTOR FROM assign WHERE OPEN_FLG='Y' GROUP BY TRW_INCOME_CD_V4;
Total jobs = 1
Launching Job 1 out of 1
```

```
hive> select * from lead;
OK
        100.0
0       100.0
J       100.0
N       100.0
O       100.0
Q       100.0
R       100.0
Time taken: 0.115 seconds, Fetched: 7 row(s)
hive>
```

G.CTOR by leads ethnicity

```
hive> create table ethnicity as select ASIAN_CD,ROUND(SUM(CASE WHEN CLICK_FLG='N' THEN 1 ELSE 0 E
ND)/count(*)*100,2) CTOR FROM assign WHERE OPEN_FLG='N' GROUP BY ASIAN_CD;
Total jobs = 1
```

```
hive>
    >
    >
    >
    > SELECT * from ethnicity;
OK
        100.0
Time taken: 0.157 seconds, Fetched: 1 row(s)
hive>
```

H.CTOR by leads household status

```
hive> create table code as select I1_INDIV_HHLD_STATUS_CODE,ROUND(SUM(CASE WHEN CLICK_FLG='N' THE
N 1 ELSE 0 END)/count(*)*100,2) CTOR FROM assign WHERE OPEN_FLG='N' GROUP BY I1_INDIV_HHLD_STATUS
_CODE;
Total jobs = 1
Launching Job 1 out of 1
```

```
hive> select * from code;
OK
H       100.0
W       100.0
Time taken: 0.02 seconds, Fetched: 2 row(s)
hive>
```

2.Household Members Information

Questions A,B,C,D

```
hive> select sum(output.count),chl from(select count(statcd_hh_mem1) as count,statcd_hh_mem1 as c
hl from assign group by statcd_hh_mem1 UNION ALL select count(statcd_hh_mem2) as count,statcd_hh_
mem2 as chl from assign group by statcd_hh_mem2 UNION ALL select count(statcd_hh_mem3) as count,s
tatcd_hh_mem3 as chl from assign group by statcd_hh_mem3 UNION ALL select count(statcd_hh_mem4) a
s count,statcd_hh_mem4 as chl from assign group by statcd_hh_mem4 UNION ALL select count(statcd_h
h_mem5) as count,statcd_hh_mem5 as chl from assign group by statcd_hh_mem5 UNION ALL select count
(statcd_hh_mem6) as count,statcd_hh_mem6 as chl from assign group by statcd_hh_mem6) as output gr
oup by chl;
Total jobs = 7
Launching Job 1 out of 7
```

```
2017-12-07 17:34:32,706 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 6.32 sec
2017-12-07 17:34:33,819 Stage-2 map = 100%,  reduce = 22%, Cumulative CPU 6.32 sec
2017-12-07 17:34:40,168 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 7.18 sec
MapReduce Total cumulative CPU time: 7 seconds 180 msec
Ended Job = job_201712061258_0085
MapReduce Jobs Launched:
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 1.54 sec   HDFS Read: 1308 HDFS Write: 209 SUCCESS
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 1.33 sec   HDFS Read: 1308 HDFS Write: 189 SUCCESS
Job 2: Map: 1  Reduce: 1   Cumulative CPU: 1.45 sec   HDFS Read: 1308 HDFS Write: 189 SUCCESS
Job 3: Map: 1  Reduce: 1   Cumulative CPU: 1.42 sec   HDFS Read: 1308 HDFS Write: 189 SUCCESS
Job 4: Map: 1  Reduce: 1   Cumulative CPU: 1.34 sec   HDFS Read: 1308 HDFS Write: 209 SUCCESS
Job 5: Map: 1  Reduce: 1   Cumulative CPU: 1.34 sec   HDFS Read: 1308 HDFS Write: 169 SUCCESS
Job 6: Map: 6  Reduce: 1   Cumulative CPU: 7.18 sec   HDFS Read: 3896 HDFS Write: 124 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 600 msec
OK
26
2       H
16      U
5       W
5       Y
1       statcd_hh_mem1
1       statcd_hh_mem2
1       statcd_hh_mem3
1       statcd_hh_mem4
1       statcd_hh_mem5
1       statcd_hh_mem6
Time taken: 133.081 seconds, Fetched: 11 row(s)
hive>
```

B: 28

C: H 2

U 16

W 5

Y 5

D: H 2%

U 16%

W 5%

Y 5%

2.Household Members Information

E:

```
hive> create table child as select(Sum(Case When PRES_OF_CHLDRN='Y' THEN 1 ELSE 0 END)) from assi
gn;
Total jobs = 1
```

```
hive> select * from child;
OK
2
Time taken: 0.021 seconds, Fetched: 1 row(s)
hive>
```

F: answer:4

```
hive> create table children as select(Sum(Case When NUMBER_OF_CHLDRN_18_OR_LESS='1' THEN 1 ELSE 0
 END)) from assign;
Total jobs = 1
Launching Job 1 out of 1
```

```
hive> select * from children;
OK
2
Time taken: 0.029 seconds, Fetched: 1 row(s)
hive>
```

```
hive> select * from child,children;
Warning: Map Join MAPJOIN[7][bigTable=children] in task 'Stage-3:MAPRED' is a cr
oss product
Total jobs = 1
```

```
MapReduce Total cumulative CPU time: 410 msec
Ended Job = job_201712061258_0076
MapReduce Jobs Launched:
Job 0: Map: 1   Cumulative CPU: 0.41 sec   HDFS Read: 213 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 410 msec
OK
2       2
Time taken: 14.574 seconds, Fetched: 1 row(s)
hive>
```

G

```
hive> select sum(output.count),chl from(select count(GNDR_OF_CHLDRN_0_3) as count,GNDR_OF_CHLDRN_
0_3 as chl from assign group by (GNDR_OF_CHLDRN_0_3) UNION ALL select count(GNDR_OF_CHLDRN_13_18)
 as count,GNDR_OF_CHLDRN_13_18 as chl from assign group by (GNDR_OF_CHLDRN_13_18)) as output grou
p by chl;
Total jobs = 3
```

```
Job 0: Map: 1  Reduce: 1   Cumulative CPU: 1.62 sec   HDFS Read: 1308 HDFS Write: 174 SUCCESS
Job 1: Map: 1  Reduce: 1   Cumulative CPU: 1.42 sec   HDFS Read: 1308 HDFS Write: 196 SUCCESS
Job 2: Map: 2  Reduce: 1   Cumulative CPU: 5.7 sec   HDFS Read: 1286 HDFS Write: 63 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 740 msec
OK
1       F
1       GNDR _OF_CHLDRN_13_18
1       GNDR_OF_CHLDRN_0_3
1       M
11      U
5       U
Time taken: 62.603 seconds, Fetched: 6 row(s)
hive>
```

Tableau

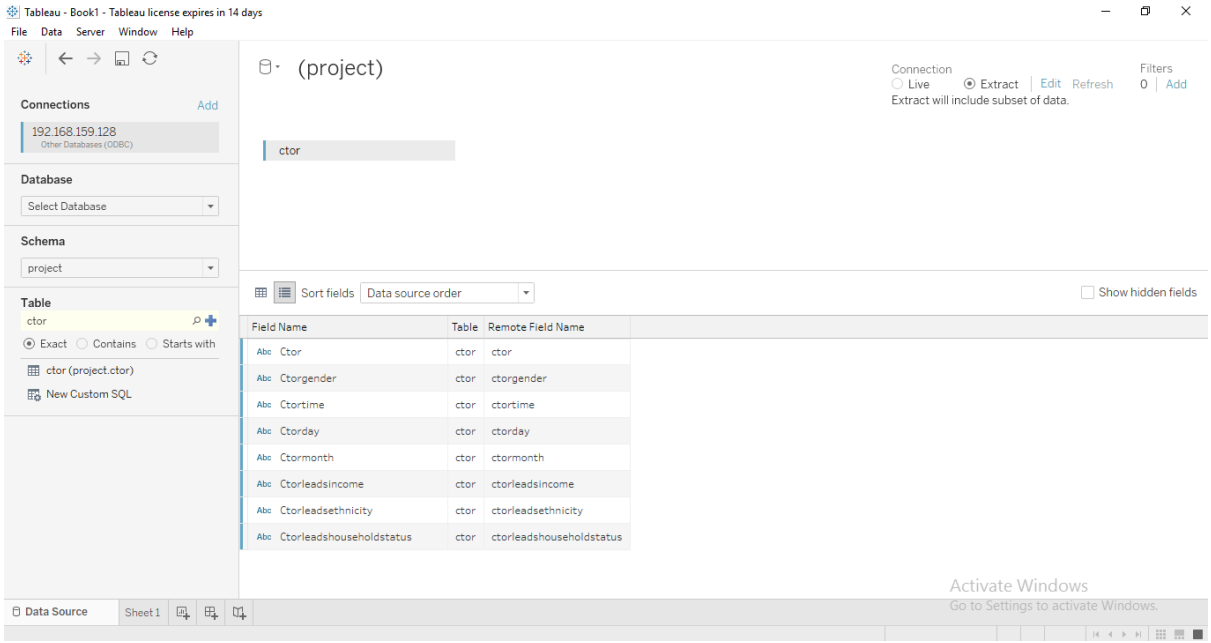
Tableau - Book1 - Tableau license expires in 14 days
File Data Server Window Help
(project)
Connections Add
192.168.159.128
Other Databases (ODBC)
Database
Select Database
Schema
project
Table
ctor
Exact Contains Starts with
ctor (project.ctor)
New Custom SQL
Connection
Live Extract Edit Refresh
Extract will include subset of data.
Filters
0 Add
ctor
Sort fields Data source order
Show hidden fields
Field Name Table Remote Field Name
Ctor ctor ctor
Ctorgender ctor ctorgender
Ctortime ctor ctortime
Ctorday ctor ctorday
Ctormonth ctor ctormonth
Ctorleadsincome ctor ctorleadsincome
Ctorleadsethnicity ctor ctorleadsethnicity
Ctorleadshouseholdstatus ctor ctorleadshouseholdstatus
Activate Windows
Go to Settings to activate Windows.
Data Source
Sheet 1

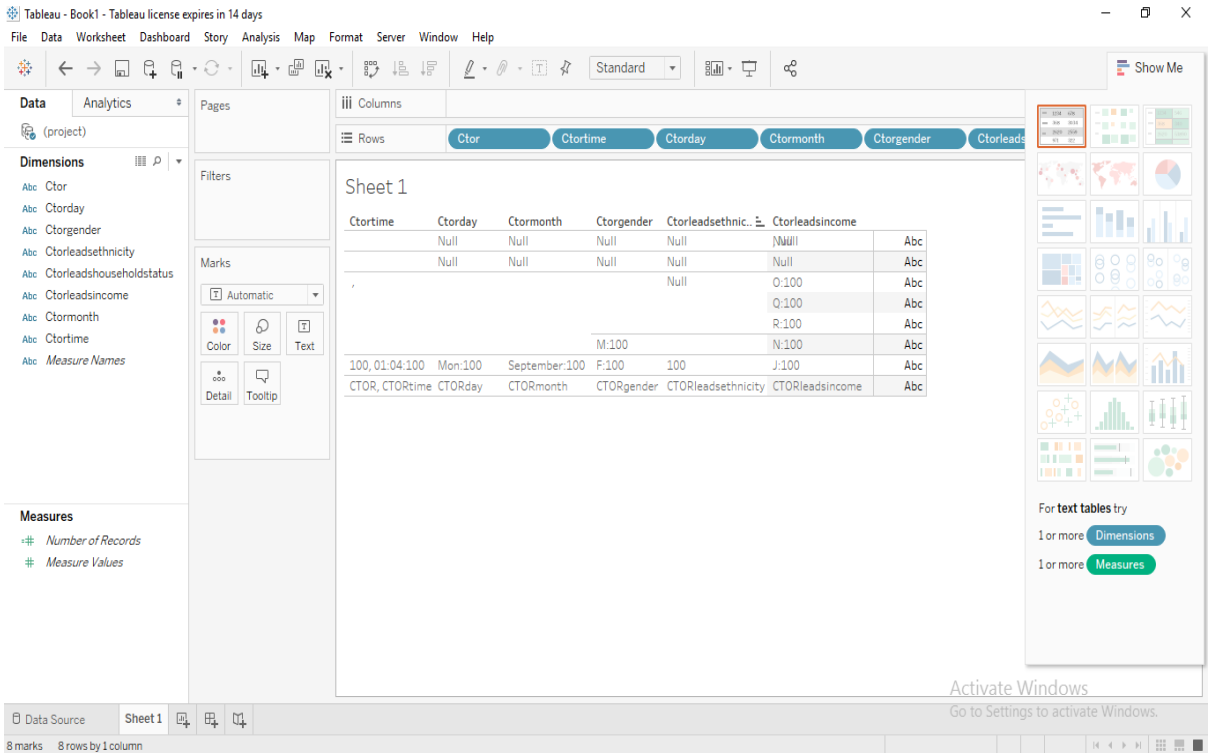
Tableau - Book1 - Tableau license expires in 14 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data Analytics
(project)
Dimensions
Ctor
Ctorday
Ctorgender
Ctorleadsethnicity
Ctorleadshouseholdstatus
Ctorleadsincome
Ctormonth
Ctortime
Measure Names
Measures
Number of Records
Measure Values
Pages
Filters
Marks
Automatic
Color Size Text
Detail Tooltip
Columns
Rows
Ctor Ctortime Ctorday Ctormonth Ctorgender
Sheet 1
Ctortime Ctorday Ctormonth Ctorgender Ctorleadsethnic.. Ctorleadsincome
Null Null Null Null Null Abc
Null Null Null Null Null Abc
Null O:100 Abc
Q:100 Abc
R:100 Abc
M:100 N:100 Abc
100, 01:04:100 Mon:100 September:100 F:100 100 J:100 Abc
CTOR, CTORtime CTORday CTORmonth CTORgender CTORleadsethnicity CTORleadsincome Abc
For text tables try
1 or more Dimensions
1 or more Measures
Activate Windows
Go to Settings to activate Windows.
Data Source
Sheet 1
8 marks 8 rows by 1 column

Dashboard:

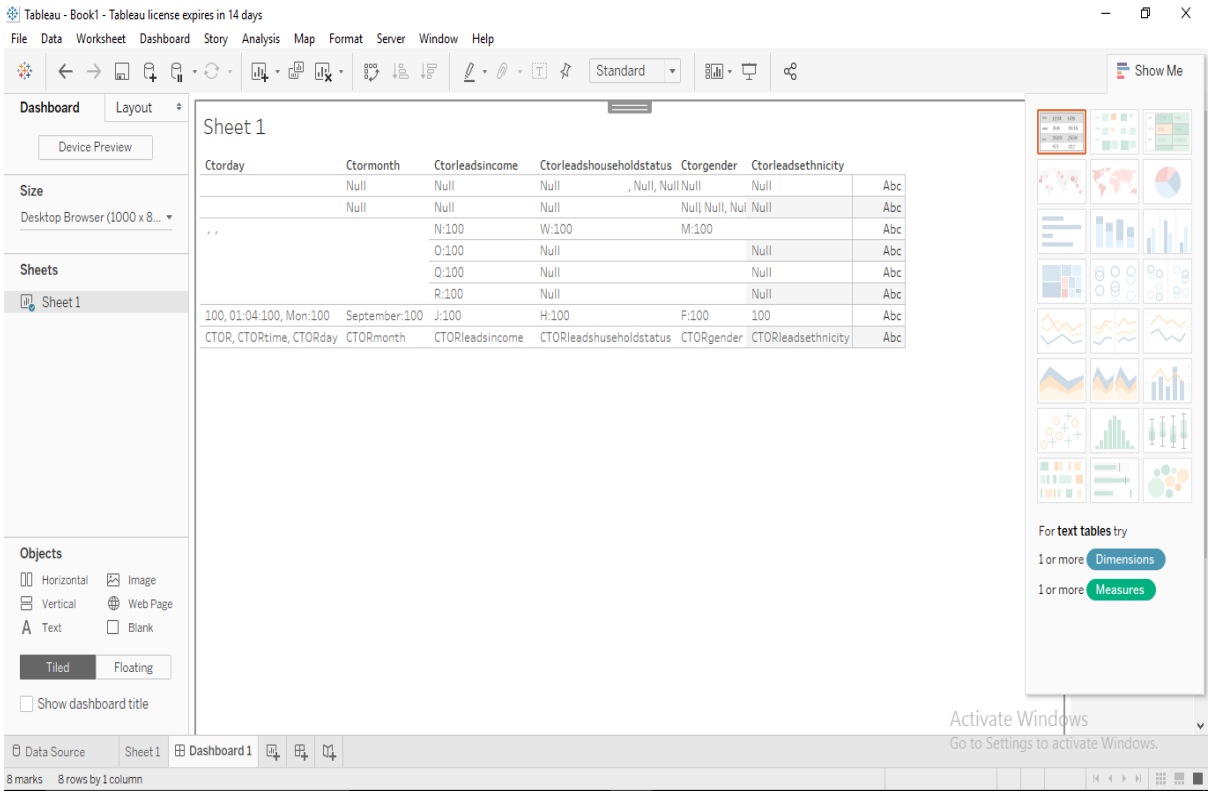
Tableau - Book1 - Tableau license expires in 14 days
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Dashboard
Layout
Device Preview
Size
Desktop Browser (1000 x 8...
Sheets
Sheet 1
Objects
Horizontal
Image
Vertical
Web Page
Text
Blank
Tiled
Floating
Show dashboard title
Sheet 1
Ctorday Ctormonth Ctorleadsincome Ctorleadshouseholdstatus Ctorgender Ctorleadsethnicity
Null Null Null , Null, Null Null Null Abc
Null Null Null Null Null, Nul Null Abc
, , N:100 W:100 M:100 Abc
O:100 Null Null Abc
Q:100 Null Null Abc
R:100 Null Null Abc
100, 01:04:100, Mon:100 September:100 J:100 H:100 F:100 100 Abc
CTOR, CTORtime, CTORday CTORmonth CTORleadsincome CTORleadshuseholdstatus CTORgender CTORleadsethnicity Abc
For text tables try
1 or more Dimensions
1 or more Measures
Activate Windows
Go to Settings to activate Windows.
Data Source
Sheet 1
Dashboard 1
8 marks
8 rows by 1 column