This Project-1 is done using Pig in local mode. Where ever possible the output is pasted here. Additionally, the output file (STORE is run on result) is also uploaded.

REGISTER '/home/acadgild/pig-0.16.0/lib/piggybank.jar';

A = LOAD '/home/acadgild/leena/ass/ass-72/Crimes\_-\_2001\_to\_present.csv' using org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO\_MULTILINE','UNIX','SKIP\_INPUT\_HEADER');

Since there are way too many fields, for ease picking the required fields only. FBICode, Arrest, Date, PrimaryType, District are the needed fields as per the assignment.

B = FOREACH A GENERATE (chararray)$0 as ID, (chararray)$1 AS CaseNo, (chararray)$2 AS Date, (chararray)$5 AS PrimaryType, (chararray)$7 AS Location, (Boolean)$8 AS Arrest, (chararray)$11 AS District, (chararray)$14 AS FBICode;

Dumping some rows for test

9981125\*HY171093\*03/03/2015 11:58:00 PM\*NARCOTICS\*STREET\*true\*006\*18

9981094\*HY171068\*03/03/2015 11:56:00 PM\*NARCOTICS\*STREET\*true\*009\*18

9981130\*HY171087\*03/03/2015 11:55:00 PM\*NARCOTICS\*STREET\*true\*024\*18

9981135\*HY171085\*03/03/2015 11:55:00 PM\*NARCOTICS\*STREET\*true\*006\*18

9981150\*HY171082\*03/03/2015 11:49:00 PM\*CRIMINAL TRESPASS\*RESIDENCE\*true\*007\*26

9981098\*HY171080\*03/03/2015 11:47:00 PM\*NARCOTICS\*STREET\*true\*009\*18

9981115\*HY171078\*03/03/2015 11:46:00 PM\*THEFT\*STREET\*false\*007\*06

1. Calculate the no of cases investigated under each FBI code

C = FILTER B BY FBICode is not null;

D = GROUP C BY FBICode;

E = FOREACH D GENERATE group, COUNT(C.FBICode);

Output

(02,1502)

(03,10596)

(05,14842)

(06,64329)

(07,11105)

(09,445)

(10,1551)

(11,13757)

(12,27)

(13,57)

(14,31301)

(15,3694)

(16,1787)

(17,1126)

(18,25207)

(19,434)

(20,1267)

(22,371)

(24,4046)

(26,29474)

(01A,533)

(01B,6)

(04A,4994)

(04B,7710)

(08A,14167)

(08B,46938)

1. No of cases under FBICode 32

F = FILTER B BY (FBICode == '32');

Output tells me there are no rows with FBICode as 32. Even when i check the above question where the count for each FBICode is taken, there is no FBICode with 32.

The store of F also revealed the same.

1. Calculate no of arrests in Theft district-wise

L = FILTER B BY (Arrest == TRUE AND PrimaryType matches '.\*THEFT.\*');

M = GROUP L BY District;

N = FOREACH M GENERATE group, COUNT(L.District);

Output

(001,1137)

(002,289)

(003,217)

(004,269)

(005,318)

(006,699)

(007,232)

(008,561)

(009,377)

(010,229)

(011,253)

(012,405)

(014,255)

(015,168)

(016,203)

(017,256)

(018,753)

(019,524)

(020,256)

(022,252)

(024,240)

(025,667)

1. Calculate no of arrests done between Oct 2014 and Oct 2015

O = FILTER B BY (Arrest == TRUE);

Tried two methods:

Method1:

P = FOREACH O GENERATE (int)SUBSTRING(Date, 0, 2) AS Month, (int)SUBSTRING(Date, 6, 10) AS Year, ID, PrimaryType, Location, District, FBICode;

DESCRIBE P;

P: {Month: int,Year: int,ID: chararray,PrimaryType: chararray,Location: chararray,District: chararray,FBICode: chararray}

Method2:

ONE = FOREACH O GENERATE ToDate(Date,'mm/dd/yyyy HH:mm:ss aa') as (date\_time), ID, PrimaryType, Location, District, FBICode;

TWO = FILTER ONE BY ((date\_time > ToDate('10/01/15','mm/dd/yy')) AND (date\_time < ToDate('01/11/16','mm/dd/yy')) );

THREE = FOREACH ONE (chararray)(date\_time), ID, PrimaryType, Location, District, FBICode;

THREE = GROUP TWO all;

FOUR = FOREACH THREE GENERATE group, COUNT(TWO.FBICode);

Output

(all,47174)

The issue is the date format has two formats dd/mm/yyyy and mm/dd/yyyy. So this yields a little different result.