**You can use a SQL database of your choosing, ideally a Postgres or AWS Redshift database, but any SQL based database will do. For the following questions please submit both the SQL query and the response. Once you’ve loaded the data write SQL queries to do the following:**

1. Create a new table and load the data

CREATE TABLE `sql\_exploration\_sample\_data` (

`vb\_tsmart\_first\_name` text,

`vb\_voterbase\_gender` text,

`vb\_tsmart\_state` text,

`vb\_voterbase\_dob` date DEFAULT NULL,

`vb\_firsttimeseen\_voterid` date DEFAULT NULL,

`vb\_voterbase\_voter\_score` text,

`vhsyn\_vf\_g2016\_synthetic` text

) ENGINE=InnoDB DEFAULT CHARSET=utf8;

LOAD DATA LOCAL INFILE 'A:/SQL/SQL/SQL\_exploration\_sample\_data.csv'

INTO TABLE sql\_exploration\_sample\_data

FIELDS TERMINATED BY ','

LINES TERMINATED BY '\n'

IGNORE 1 LINES;

1. Show the first 10 rows

SELECT \* FROM mydb.sql\_exploration\_sample\_data limit 10;

# vb\_tsmart\_first\_name, vb\_voterbase\_gender, vb\_tsmart\_state, vb\_voterbase\_dob, vb\_firsttimeseen\_voterid, vb\_voterbase\_voter\_score, vhsyn\_vf\_g2016\_synthetic

'DONALD', 'Male', 'VT', '1962-01-01', '2017-05-31', 'Never Voted', 'N\r'

'SHALLI', 'Female', 'OK', '1960-12-16', '2016-03-25', 'Infrequent Voter', 'E\r'

'RASHAD', 'Male', 'FL', '1996-09-25', '2014-01-16', 'Never Voted', 'N\r'

'PATRICK', 'Male', 'PA', '1995-10-03', '2014-01-14', 'Infrequent Voter', 'A\r'

'JESUS', 'Male', 'DE', '1957-12-01', '2016-11-29', 'Super Voter', 'Y\r'

'ALYSSA', 'Female', 'NC', '1997-01-01', '2016-07-08', 'Infrequent Voter', 'A\r'

'MARJORIE', 'Female', 'CA', '1993-08-23', '2016-11-30', 'Infrequent Voter', 'A\r'

'ANGELA', 'Female', 'MA', '1957-08-03', '2014-08-26', 'Never Voted', 'N\r'

'CLAIRE', 'Female', 'OK', '1969-10-01', '2016-03-25', 'Infrequent Voter', 'Y\r'

'JESSICA', 'Female', 'VA', '1995-03-16', '2014-10-30', 'Never Voted', 'N\r'

1. Summarize the total record count by state

select vb\_tsmart\_state, count(1)

FROM mydb.sql\_exploration\_sample\_data

group by vb\_tsmart\_state

order by vb\_tsmart\_state

vb\_tsmart\_state no\_of\_people

(null) 1

AE 1

AK 14

AL 164

AP 1

AR 100

AZ 245

CA 1533

CO 159

CT 110

DC 37

DE 37

FL 582

GA 327

HI 31

IA 74

ID 45

IL 360

IN 247

KS 67

KY 131

LA 130

MA 210

MD 153

ME 48

MI 223

MN 129

MO 152

MS 65

MT 33

NC 296

ND 20

NE 53

NH 60

NJ 224

NM 41

NV 125

NY 481

OH 358

OK 133

OR 110

PA 392

PR 3

RI 35

SC 129

SD 29

TN 193

TX 764

UT 84

VA 195

VT 14

WA 219

WI 564

WV 47

WY 22

1. Find the average age of people by state

select vb\_tsmart\_state, avg(voter\_age)/365.25 as 'average\_vote\_age'

from

(

select vb\_tsmart\_first\_name,vb\_voterbase\_gender,vb\_tsmart\_state,vb\_voterbase\_dob,vb\_firsttimeseen\_voterid,vb\_voterbase\_voter\_score,vhsyn\_vf\_g2016\_synthetic,

case when vb\_voterbase\_dob is null then null

else DATEDIFF(SYSDATE(),vb\_voterbase\_dob ) end as voter\_age

from mydb.sql\_exploration\_sample\_data where vb\_voterbase\_dob is not null

)voter\_data

group by vb\_tsmart\_state

order by vb\_tsmart\_state;

vb\_tsmart\_state average\_vote\_age

AE 26.53251198

AL 40.08854610

AP 22.69952088

AR 38.47824778

AZ 43.93447038

CA 43.46170851

CO 38.12421061

CT 36.89515276

DC 35.06402008

DE 40.59100579

FL 42.52283761

GA 37.64064655

HI 54.09221579

IA 33.20787317

ID 46.59091946

IL 37.77150385

IN 40.13954171

KS 42.20368384

KY 41.00460314

LA 36.76082767

MA 39.64424888

MD 35.43365856

ME 40.36481862

MI 39.07251928

MN 38.15197211

MO 38.08881697

MS 48.50093543

MT 45.56115570

NC 39.17956971

ND 52.61704312

NE 36.42281715

NH 39.96701542

NJ 35.34731593

NM 38.13340178

NV 43.10233539

NY 38.93421332

OH 37.58584271

OK 41.23300037

OR 40.98186796

PA 39.84193102

PR 73.53319644

RI 34.19286203

SC 40.73143063

SD 39.11733646

TN 43.08511808

TX 38.08454010

UT 34.98083504

VA 38.29554046

VT 44.31407060

WA 38.33113618

WI 63.73569003

WV 41.49038111

WY 46.24969199

1. Find the most recent date a voter was seen in each state (vb\_firsttimeseen\_voterid)

select vb\_tsmart\_state, max(vb\_firsttimeseen\_voterid) as 'most\_recent\_date'

FROM mydb.sql\_exploration\_sample\_data

group by vb\_tsmart\_state

order by vb\_tsmart\_state;

vb\_tsmart\_state most\_recent\_date

(null) 2016-11-17

AE 2016-08-29

AK 2017-10-04

AL 2017-09-13

AP 2016-02-13

AR 2017-09-30

AZ 2017-08-09

CA 2017-07-20

CO 2017-10-02

CT 2017-09-25

DC 2017-05-18

DE 2017-08-31

FL 2017-08-31

GA 2017-09-07

HI 2016-10-10

IA 2017-10-11

ID 2017-05-23

IL 2017-06-15

IN 2017-08-12

KS 2017-09-01

KY 2017-07-31

LA 2017-08-31

MA 2017-07-22

MD 2017-06-30

ME 2017-06-13

MI 2017-09-06

MN 2017-10-16

MO 2017-07-25

MS 2017-08-03

MT 2017-09-12

NC 2017-08-25

ND 2017-03-24

NE 2017-05-18

NH 2017-04-13

NJ 2017-10-13

NM 2017-07-07

NV 2017-09-10

NY 2017-06-30

OH 2017-09-15

OK 2017-09-22

OR 2017-06-30

PA 2017-09-25

PR 2016-10-02

RI 2017-08-10

SC 2017-02-15

SD 2017-07-11

TN 2017-06-27

TX 2017-09-11

UT 2017-06-28

VA 2017-09-30

VT 2017-07-15

WA 2017-08-31

WI 2017-06-30

WV 2017-03-14

WY 2017-03-06

1. Summarize the turnout in vhsyn\_vf\_g2016\_synthetic into ‘Voted’, ‘Didn’t Vote’ (Hint: A is absentee vote, E is Early vote, Y is Voted, N is didn’t vote and I is ineligible.)

select voted\_yes\_no, count(1) as 'count' from

(

select case when trim(upper(vhsyn\_vf\_g2016\_synthetic)) like('%Y%') then 'Voted'

when trim(upper(vhsyn\_vf\_g2016\_synthetic)) like('%N%') then 'Didnt\_vote'

end as 'voted\_yes\_no'

FROM mydb.sql\_exploration\_sample\_data

where trim(upper(vhsyn\_vf\_g2016\_synthetic)) like('%Y%')

or trim(upper(vhsyn\_vf\_g2016\_synthetic)) like('%N%')

) voted\_y\_n

group by voted\_yes\_no

order by voted\_yes\_no desc ;

voted\_yes\_no count

Voted 3381

Didnt\_vote 4490

1. What is most common first name in each state?

select vb\_tsmart\_state as State, vb\_tsmart\_first\_name as First\_name, max(count\_names) as Count

from (

select vb\_tsmart\_state, vb\_tsmart\_first\_name, count(vb\_tsmart\_first\_name) as 'count\_names'

FROM mydb.sql\_exploration\_sample\_data

group by vb\_tsmart\_state, vb\_tsmart\_first\_name

) rslt\_set where count\_names> 1 group by vb\_tsmart\_state, vb\_tsmart\_first\_name ;

SS: There were not many repeats in the data so I just picked the names appearing more than once. partial output is shown below:

State First\_name Count

AL AMY 2

AL ANTHONY 2

AL DAVID 3

AL EMILY 2

AL JAMES 3

AL JARON 2

AL JUSTIN 2

AL MARY 2

AL MICHAEL 3

AL RICHARD 2

AL STEPHEN 3