**Data warehousing project**

**Bike MS**

Jukanna Nityanand

**Project**

We selected the “Bike MS” dataset since it is bit complex and has more data files. It takes **more time to clean and to understand the data**. we have more variety which means we are not limited to manipulate data and we can extract the most relevant information that are going to be useful for our project.

Data warehouse is focused more on analytical information, the goal of our project is to analyse the amount of donation given by teams and participants based on name, division and prior participant. We are going to analyse and retrieve also donors based on states, gender and type of giving’s over years (2013-2017).

Questions related to Business

**Q1. What is the total amount of giving’s every team (is prior participants) did per year?**

**Q2. What are the cities and respective states where is done the majority of giving’s?**

**Q3. What is the average amount of donations for a particular event in a given year?**

**Q4. What is the max amount of giving’s and max average of giving’s based on donor’s gender?**

**Q5. What is the max amount of giving’s and max average of giving’s based on donor’s gender per each year?**

**Q6. What is the total amount of giving’s every team did based on their name and number of participants?**

**7.What are the top 5 cities and respective states where is done the majority of givings?**

**8.What is the total amount of giving’s per each type of gift(by donors) ordered by year and state ?**

INSPECTION AND PROFILING

we move to another step which consists of inspection and profiling, and we decided to drop come columns that are irrelevant to our project.

There are 6 data files (Affiliates, Bike Teams, Donations, Events, National Teams, Participants) which are data sources for the project. we did inspection and profiling and we understood Donation is the main source of data for us, But we are going to extract the most relevant information also from Bike teams, Participants and Events needed to answer our business questions**.**

We used Tableau and Python scripts to clean data and for the main data file we cleaned datasets per each year from 2013-2017 and merged together to have a more structured data and make much more meaningful to analyse

DFM – Conceptual Schema

Taking in consideration the business questions and our operational resources, we decided for the main fact is “donation”. Based on the requirement, we would analyze every donation from different aspects or points of view.

Dimensions are: city, donor, year, team.

To draw DFM and ROLAP schema we used a special tool called “DRAW.IO.” We can also use “INDYCO” but DRAW.IO was simplest and easy to navigate

Dynamicity or Slowly changing dimensions

we analyze the dynamicity in dimensions we considered donations between 2013-2017. The only thing that we want to compute is the average of donation. The time scenario is the third one yesterday-for-today (type-3 or SCD 3 (rollback)) which is implemented in this case. This means that all the events are analyzed according to configuration the hierarchies had in a previous time of choice.

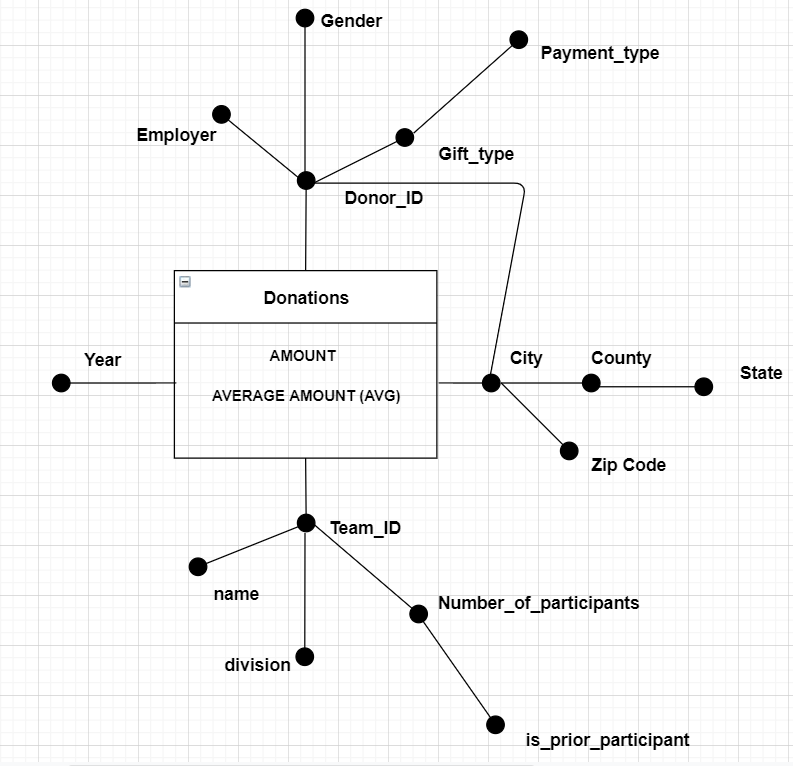
**Conceptual schema**

**Fact table:** Donation

**Dimensions:** Year, Donor, City, Team

**Measures:** Total Amount, Avg Amount (AVG)

**Period:** 2013-2017

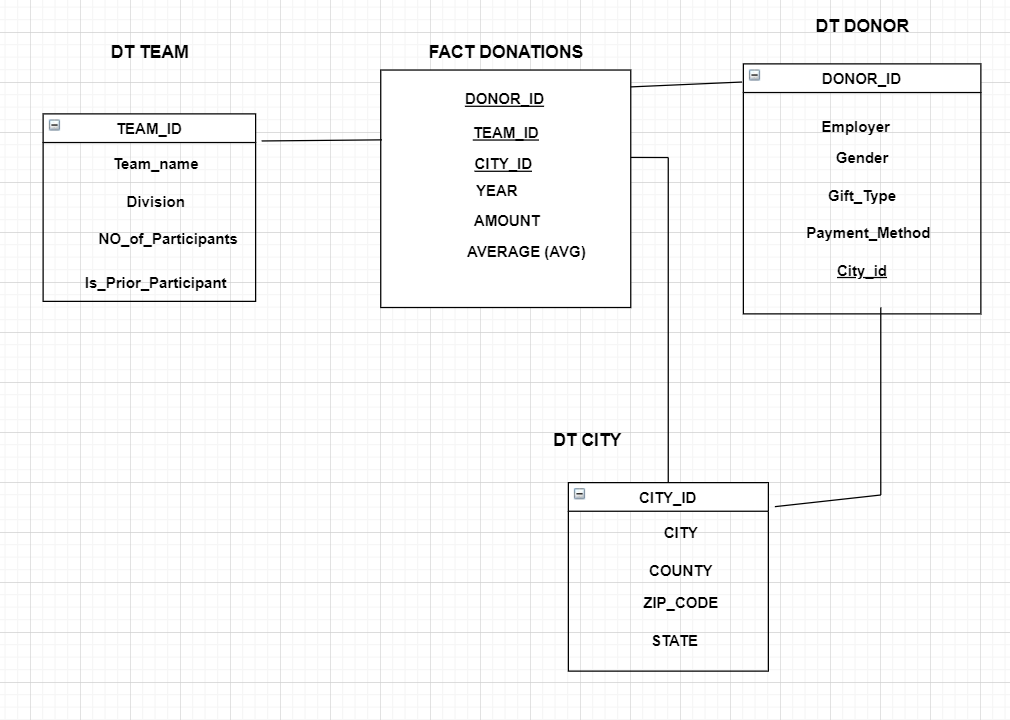


ROLAP

we can choose between star schema and snowflake schema. But in our case the ROLAP is not a pure star schema or complete snowflake schema i.e it is impure.

The schema here shares the location dimension with donor\_id dimension. In order to avoid redundancy we thought to keep location in separate dimension.

In this type of schema the query performance might be little bit slow (because of joins ) when compared to pure star schema but this will save the space when we look from space saving perspective



We are going to draw all the aggregation of our preliminary workload, in order to decide which views to materialize.

The materialized view in preliminary workload are:

P0={year, donor\_id, city, team\_id}

P1={ispriorparticipant, year}

P2={gifttype, year, city, state}

P3={teamname, nr\_of\_participants,year}

P4={state}

P5={year}

OLAP queries

For the execution of complex queries we need to use SQL OLAP extensions such as windows functions ,ranking etc

CREATE TABLE project.donation (

Team\_id integer NOT NULL,

City\_id integer NOT NULL,

donor\_id integer NOT NULL,

fiscal\_year integer NOT NULL,

gift\_amount int NOT NULL,

additional\_gift\_amount int NOT NULL,

CONSTRAINT b\_key PRIMARY KEY (Team\_id, donor\_id, City\_id, fiscal\_year));

CREATE TABLE project.city(

city\_id integer NOT NULL,

city character varying(30) NOT NULL,

state character varying(30),

county character varying(30) NOT NULL,

zipcode character varying(12),

CONSTRAINT c\_key PRIMARY KEY (city\_id));

CREATE TABLE project.teams (

team\_id int NOT NULL,

name character varying(50) NOT NULL,

division character varying(50) NOT NULL,

ispriorparticipant boolean,

number\_of\_participants character varying(30),

CONSTRAINT t\_key PRIMARY KEY (team\_id) );

CREATE TABLE project.donors (

donor\_id character varying(30) NOT NULL,

employer character varying (80) NOT NULL,

gender character varying(30),

gift\_type character varying(30),

paymentmethod character varying(30),

CONSTRAINT d\_key PRIMARY KEY(donorid),

CONSTRAINT c\_key FOREIGN KEY(cityid) REFERENCES city(cidyid));

Views:

create view public.avgyear as

select d.event\_id,d.fiscal\_year, sum(d.gift\_amount+d.additional\_gift\_amount) as total\_amount ,t.ispriorparticipant from donations d join teams t using (team\_id) group by d.event\_id,d.fiscal\_year,t.ispriorparticipant;

create view location as

select c.city, c.state, c.county, sum(d.gift\_amount+d.additional\_gift\_amount) as total\_amount, sum(d.gift\_amount+d.additional\_gift\_amount)/2 as avg\_amount from donations d join city c using(city\_id) group by c.city, c.state, c.county;

create view gender as

select s.gender, sum(d.gift\_amount+d.additional\_gift\_amount) as total\_amount, sum(d.gift\_amount+d.additional\_gift\_amount)/2 as avg\_amount,d.fiscal\_year from donations d join donor s using(donor\_id) group by d.fiscal\_year,s.gender;

Q1. What is the total amount of givings every team(is prior participants)did per year

select t.ispriorparticipant, sum(d.gift\_amount+d.additional\_gift\_amount) as total\_amount, d.fiscal\_year from donations d join teams t using (team\_id) group by d.fiscal\_year,t.ispriorparticipant;

Q2. What are the cities and respective states where is done the majority of givings?

select max(total\_amount),state,city from location group by state, city;

Q3.What is the average amount of donations for a particular event in a given year?

select event\_id,fiscal\_year,avg(total\_amount) as average\_amount from avgyear group by event\_id,fiscal\_year limit 300;

Q4.What is the max amount of givings and max average of givings based on donor’s gender ?

select max(total\_amount),max(avg\_amount),gender from gender group by gender;

Q5.What is the max amount of givings and max average of givings based on donor’s gender per each year?

select max(total\_amount),max(avg\_amount),gender , fiscal\_year from gender group by gender,fiscal\_year order by fiscal\_year ;

Q6.What is the total amount of givings every team did based on their name and number of participants

Select t.ispriorparticipant,t.number\_of\_participants,t.name , sum(d.gift\_amount+d.additional\_gift\_amount) as total\_amount from donations d left outer join teams t using (team\_id) group by t.ispriorparticipant,t.number\_of\_participants,t.name limit 100;

7.What are the top 5 cities and respective states where is done the majority of givings?

select city,state,county,total\_amount from location order by total\_amount desc limit 5;

8.What is the total amount of givings per each type of gift(by donors) ordered by year and state ?

select c.state,sum(t.gift\_amount+t.additional\_gift\_amount) as total\_amount ,d.gifttype from donations t inner join city c using(city\_id) inner join donor d on d.donor\_id = t.donor\_id group by c.state,d.gifttype order by c.state asc;

Queries referring to specific OLAP extentions of PostgreSQL for windows and window functions

**-Computing rankings and partitioning**

select event\_id, fiscal\_year, total\_amount ,avg(total\_amount)over(partition by fiscal\_year),dense\_rank() over(order by fiscal\_year desc) from avgyear;

**-Computing cumulative totals ( window framing)**

select c.state,d.fiscal\_year, sum(net\_transaction\_amount) over (order by c.state range between UNBOUNDED PRECEDING AND CURRENT ROW) from donations d join city c using (city\_id) group by c.state,d.fiscal\_year,net\_transaction\_amount;

**-Computing mobile aggregates [window framing]**

select event\_id,fiscal\_year, sum(net\_transaction\_amount),avg(net\_transaction\_amount) OVER(Partition by event\_id order by fiscal\_year rows 1 preceding ) from donations group by event\_id,net\_transaction\_amount,fiscal\_year order by event\_id limit 100;

Hive

Hive is a data warehousing software built on Apache Hadoop for providing data query and analysis. It supports analysis of large and complex datasets stored in Hadoop’s and its less expensive and more efficient than traditional technology. Hive is more powerful and it may increase also the performance by

using partitions. We imported our data warehouse in Hive and run the OLAP queries first. After that we create also 3 relevant queries and run them on it.

LOAD DATA LOCAL INPATH '/home/user39/dataset/participants.csv' OVERWRITE INTO TABLE user39.participants;

CREATE TABLE user39.teams ( team\_id INT, name STRING, team\_division STRING, is\_priorpaticpant STRING , number\_of\_participants INT ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' TBLPROPERTIES ( 'skip.header.line.count'='1');

CREATE TABLE USER39.Donations(

security\_category\_name STRING,

event\_id INT,

public\_event\_name STRING,

fiscal\_year INT,

campaign\_title STRING,

campaign\_id INT,

gift\_amount INT,

offline\_status STRING,

soft\_credit\_type string,

is\_registration STRING,

donor\_consID INT,

donor\_member\_id INT,

donor\_affiliate\_code string,

donor\_accept\_email string,

donor\_opt\_out\_method string,

donor\_email\_status STRING,

donor\_connection\_to\_MS STRING,

participant\_contact\_ID INT,

participant\_member\_ID INT,

participant\_type\_name string ,

registration\_active\_status STRING,

participant\_goal INT,

is\_team\_captain STRING,additional\_gift\_amount INT,

team\_id INT,

original\_value\_transacted INT,

net\_transaction\_amount INT,

ledger\_transaction\_amount INT,

donor\_id int,

city\_id int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' TBLPROPERTIES ( 'skip.header.line.count'='1');

CREATE TABLE user39.donors ( donor\_id int, employer string, gender string, gift\_type string, paymentmethod string ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' TBLPROPERTIES ( 'skip.header.line.count'='1');

CREATE TABLE user39.city ( city\_id int, city string, state string, county string, zipcode string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' TBLPROPERTIES ( 'skip.header.line.count'='1');

create table user39.participants(Participant\_Connection\_to\_MS string,event\_id int ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' TBLPROPERTIES ( 'skip.header.line.count'='1');

Q1. What is the total amount of givings every team(is prior participants)did per year

select d.event\_id,d.fiscal\_year, sum(d.gift\_amount+d.additional\_gift\_amount) as total\_amount ,t.is\_prior\_paticipant from donations d join teams t on (t.team\_id= d.team\_id) group by d.event\_id,d.fiscal\_year,t.is\_prior\_paticipant;

Q2. What are the cities and respective states where is done the majority of givings?

#created view and then query the data

create view location as

select c.city, c.state, c.county, sum(d.gift\_amount+d.additional\_gift\_amount) as total\_amount, sum(d.gift\_amount+d.additional\_gift\_amount)/2 as avg\_amount from donations d join city c using(city\_id) group by c.city, c.state, c.county;

select max(total\_amount),state,city from location group by state, city;

Q3.What is the average amount of donations for a particular event in a given year?

create view avgyear as

select d.event\_id,d.fiscal\_year, sum(d.gift\_amount+d.additional\_gift\_amount) as total\_amount ,t.is\_prior\_paticipant from donations d join teams t using (team\_id) group by d.event\_id,d.fiscal\_year,t.is\_prior\_paticipant;

select event\_id,fiscal\_year,avg(total\_amount) as average\_amount from avgyear group by event\_id,fiscal\_year limit 300;

Q4.What is the total amount of givings every team did based on their name and number of participants

select t.is\_prior\_paticipant,t.number\_of\_participants,t.name , sum(d.gift\_amount+d.additional\_gift\_amount) as total\_amount from donations d left outer join teams t using (team\_id) group by t.is\_prior\_paticipant,t.number\_of\_participants,t.name limit 100;

Q5.What is the max amount of givings and max average of givings based on donor’s gender per each year?

select max(total\_amount),max(avg\_amount),gender , fiscal\_year from gender group by gender,fiscal\_year ;

Extra 3 queries focusing on partition and clustering

1) Retrivethe gift amount and participation\_type of year 2013 and who has additional caption in team

# create partition table first

create table donations\_part(gift\_amount int ,participation\_type\_name string) partitioned by (fiscal\_year int);

#set property to load data into partitioned table

set hive.exec.dynamic.partition.mode = nonstrict;

#load the data to partitioned table

insert overwrite table donations\_part partition (fiscal\_year) select gift\_amount, participation\_type\_name,fiscal\_year from donations

select gift\_amount, participation\_type\_name from donations where fiscal\_year = 2013;

2) what is the average gift\_amount, year and donor accept email from donations of specific rows;

create table donations\_bucket(gift\_amount int ,donor\_accept\_email string,fiscal\_year int, donor\_id int) clustered by (fiscal\_year) sorted by (donor\_id ) into 5 buckets;

#set property to load data into bucketed table

set hive.enforce.bucketing = true;

#load the data to bucketed table

insert overwrite table donations\_bucket select gift\_amount ,donor\_accept\_email ,fiscal\_year , donor\_id from donations ;

select avg(gift\_amount) from donations\_bucket tablesample(bucket 1 out of 5 on donor\_id)

3) H3.who are the employer that have donated through credit card.

select distinct(employer )from donors where paymentmethod = 'credit card' limit 8;

SPARK SQL

Spark is an open source ,general-purpose distributed computing engine used for processing and analysing a large amount of data. It is also faster than Hive and is always a good option for scaling. We execute some OLAP queries here

For the spark we used notebook and we will submit python jupyter\_notebook as well . please refer to notebook (name “py\_spark\_sql”) if things aren’t clear here

#Q1. What is the total amount of givings every team(is prior participants)did per year

team\_df.join(donations\_df, on="team\_id",how = "inner")\

.groupby(team\_df.ispriorparticipant,donations\_df.fiscal\_year)\

.agg(f.sum(donations\_df.gift\_amount+donations\_df.additional\_gift\_amount).alias("total\_amount")) \

.show()

#Q2. What are the cities and respective states where is done the majority of givings?

location\_df.groupby("city","state","total\_amount")\

.agg(f.max("total\_amount"))\

.orderBy('total\_amount', ascending=False)\

.show(10)

#Q3.What is the average amount of donations for a particular event in a given year?

avgyear\_df.groupby("event\_id","fiscal\_year")\

.agg(f.avg("total\_amount"))\

.limit(300).show(10)

#Q4.What is the max amount of givings and max average of givings based on donor’s gender ?

gender\_df.groupby("gender")\

.agg(f.max("total\_amount"),f.max("avg\_amount"))\

.show(3)

#Q5.What is the max amount of givings and max average of givings based on donor’s gender per each year?

group by gender,fiscal\_year ;

gender\_df.groupby("gender","fiscal\_year")\

.agg(f.max("total\_amount"),f.max("avg\_amount"))\

.orderBy('fiscal\_year', ascending=True)\

.show(15)

#Q6.What is the total amount of givings every team did based on their name and number of participants

team\_df.join(donations\_df, on="team\_id",how = "inner")\

.groupby(team\_df.ispriorparticipant,team\_df.number\_of\_participants,team\_df.name)\

.agg(f.sum(donations\_df.gift\_amount+donations\_df.additional\_gift\_amount).alias("total\_amount")) \

.show()

#7.What are the top 5 cities and respective states where is done the majority of givings?

location\_df.groupby("city","state","total\_amount")\

.agg(f.max("total\_amount"))\

.orderBy('total\_amount', ascending=False)\

.limit(5).show()

#8.What is the total amount of givings per each type of gift(by donors) ordered by year ?

donations\_df.join(donor\_df, on="donor\_id",how = "inner")\

.groupby(donations\_df.fiscal\_year,donor\_df.gifttype)\

.agg(f.sum(donations\_df.gift\_amount+donations\_df.additional\_gift\_amount).alias("total\_amount")) \

.orderBy('fiscal\_year', ascending=False)\

.show(2)

#9. what is the total amount givings from all the events per state

events\_df.select("average\_team\_size","total\_from\_participant","state")\

.groupBy("state")\

.agg(f.sum(events\_df.total\_from\_participant).alias("total\_amount")) \

.orderBy("state").show();

#10.which occupupation has the highest givings

participants1\_df.groupBy("participant\_occupation")\

.agg(f.sum(participants1\_df.total\_from\_participant).alias("total\_amount"))\

.orderBy('total\_amount', ascending=False)\

.show()

#11.get the events and the amount which has same value in year 2017 and count the number of times the event gave same amount

teams1\_df.filter(col("fiscal\_year").startswith("2014"))\

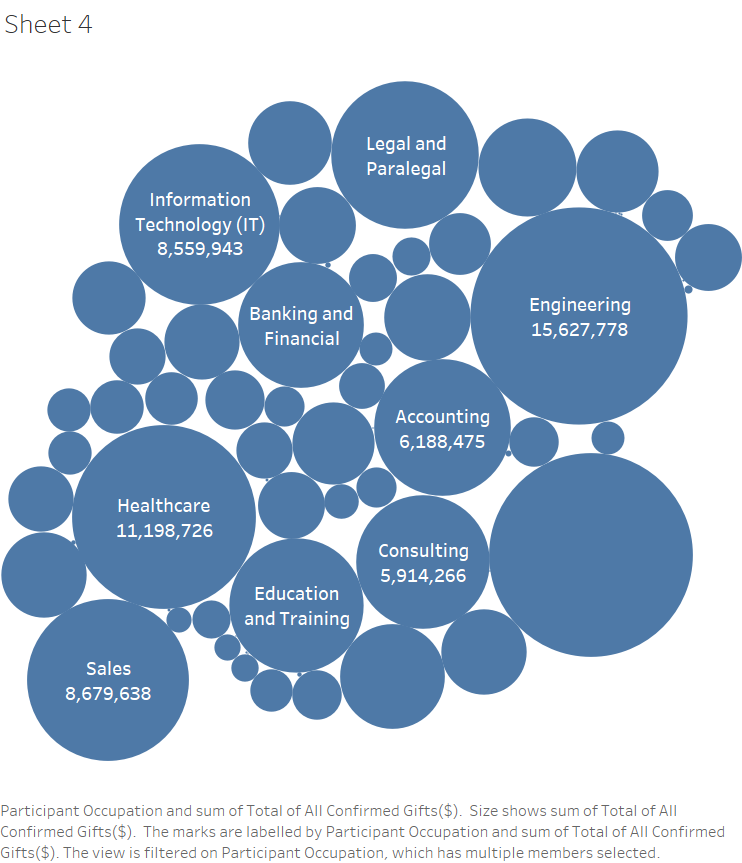
.rollup("fiscal\_year","event\_type", teams1\_df.total\_offline\_confirmed\_gifts).count()\

.where(col("event\_type").isNotNull()).orderBy("fiscal\_year","event\_type")\

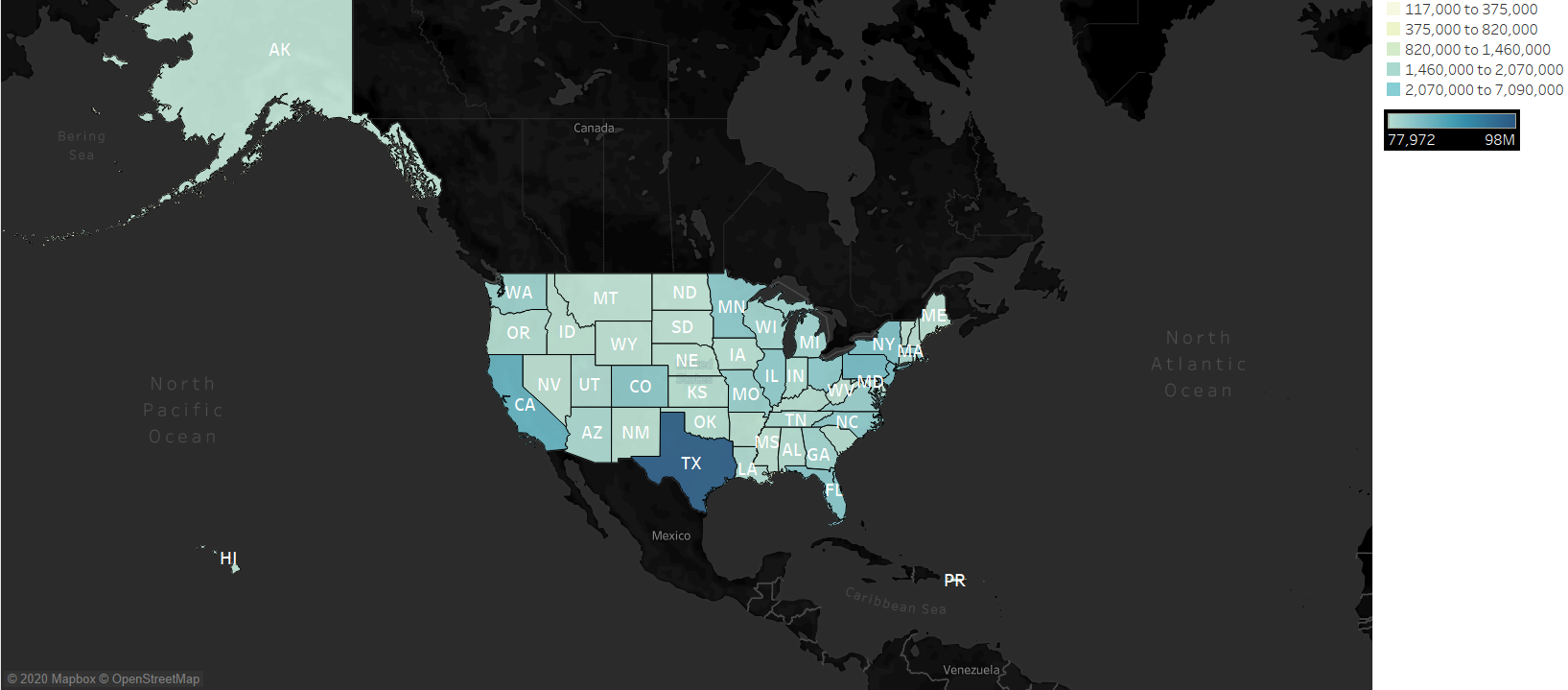
.show()

**Tableau**

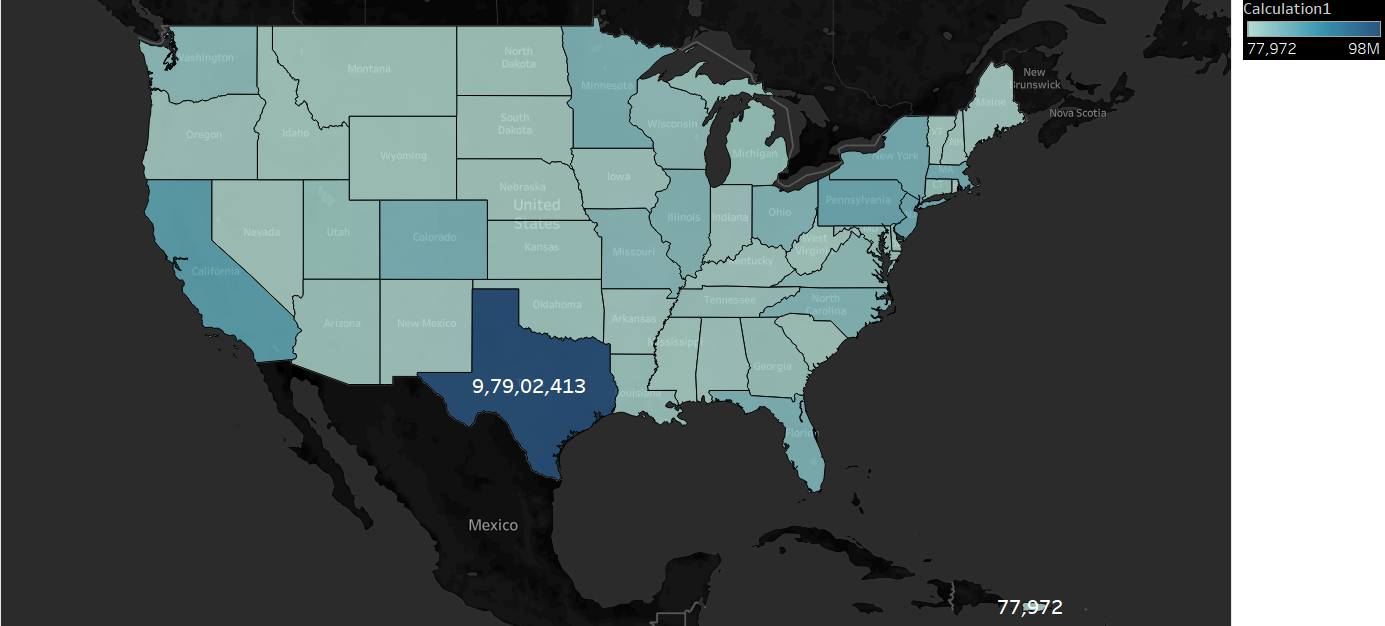
which Industries had the strongest involvement in Bike MS in the last five years and related occupations who are responsible for most of bike MS fund raising



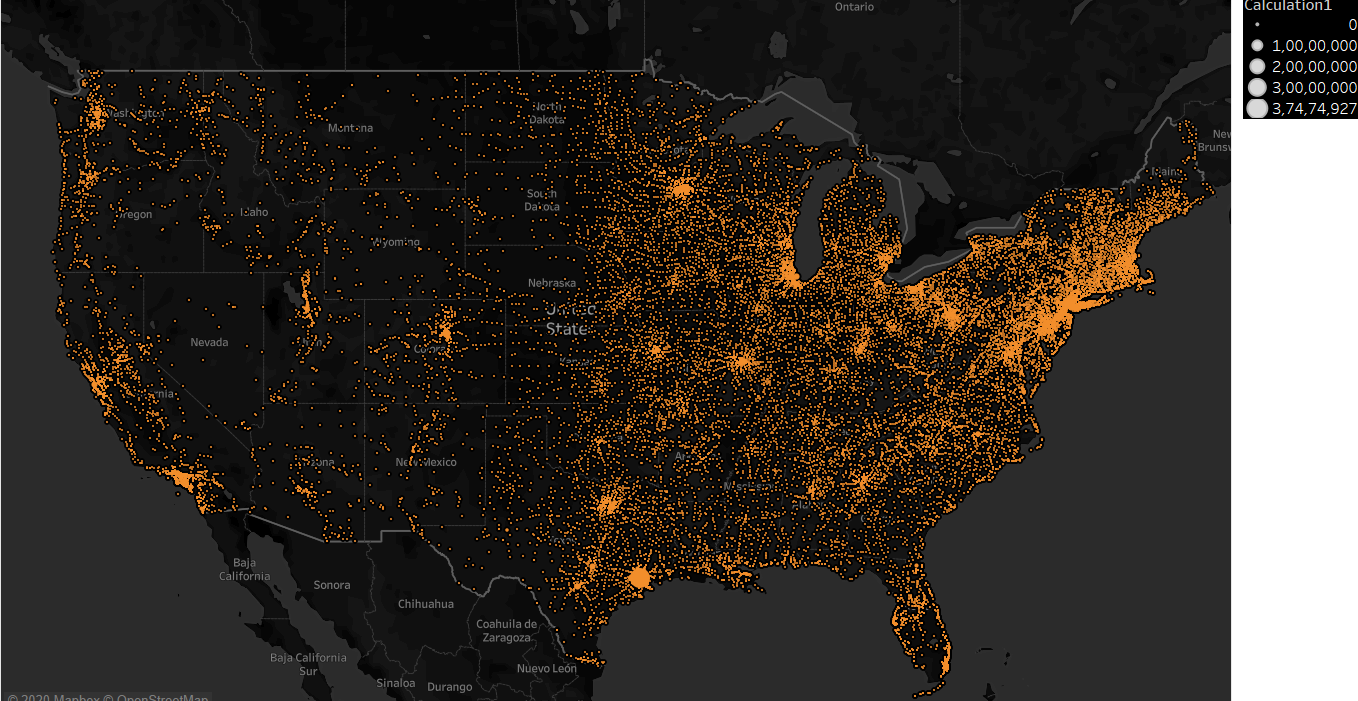
1. What are the states where the outbreak of MS is the most and which are the areas that are donating the most?



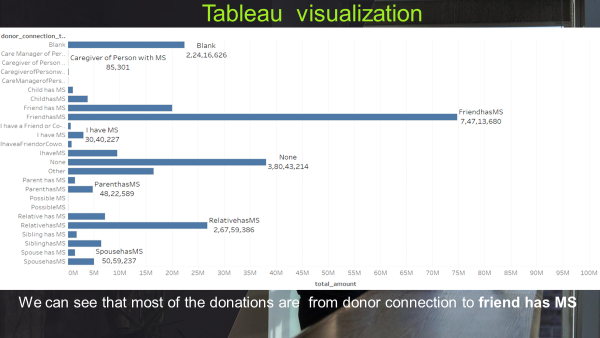
What are the states where the outbreak of MS fund raising is the high and low ?



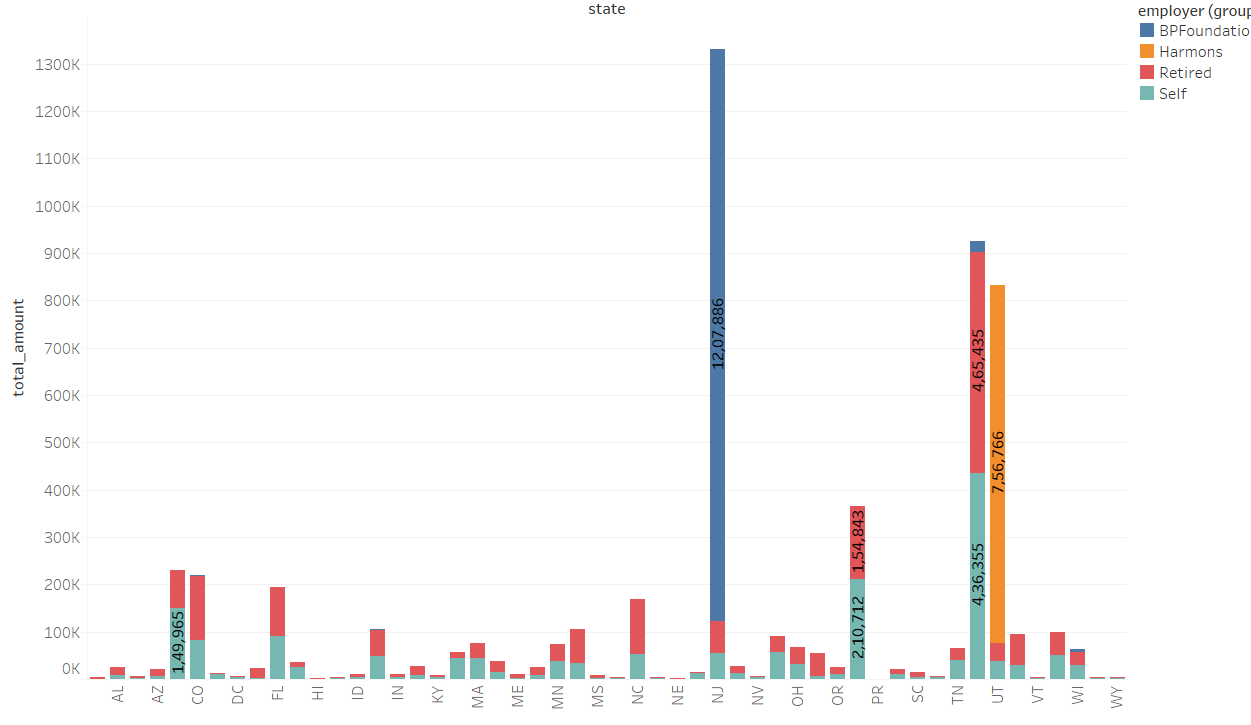
Which part of US cities has participated in event and raised fund?



What are the donations that are related to someone who have a connection with the MS disease?

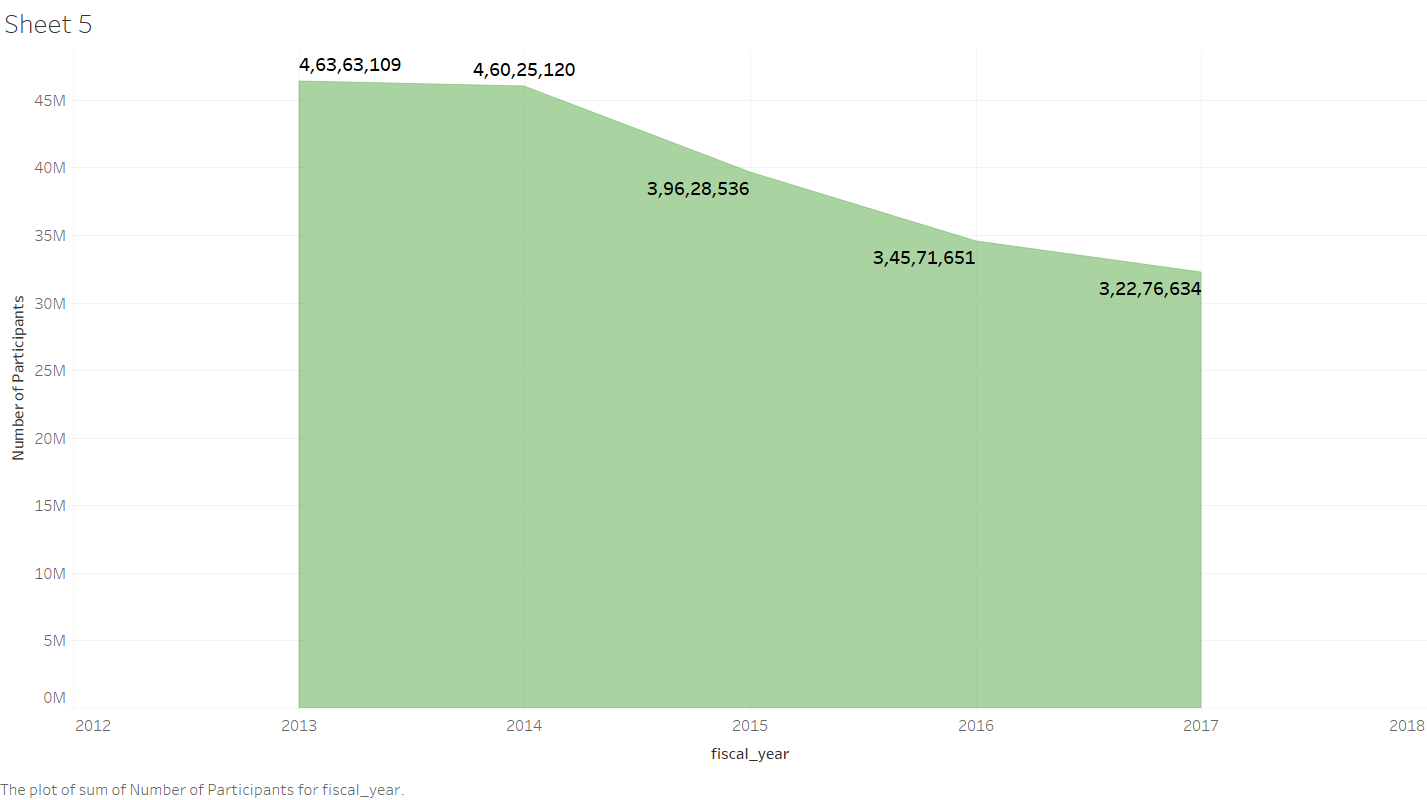


Who are the top employers who donated the fund and what are their states ?

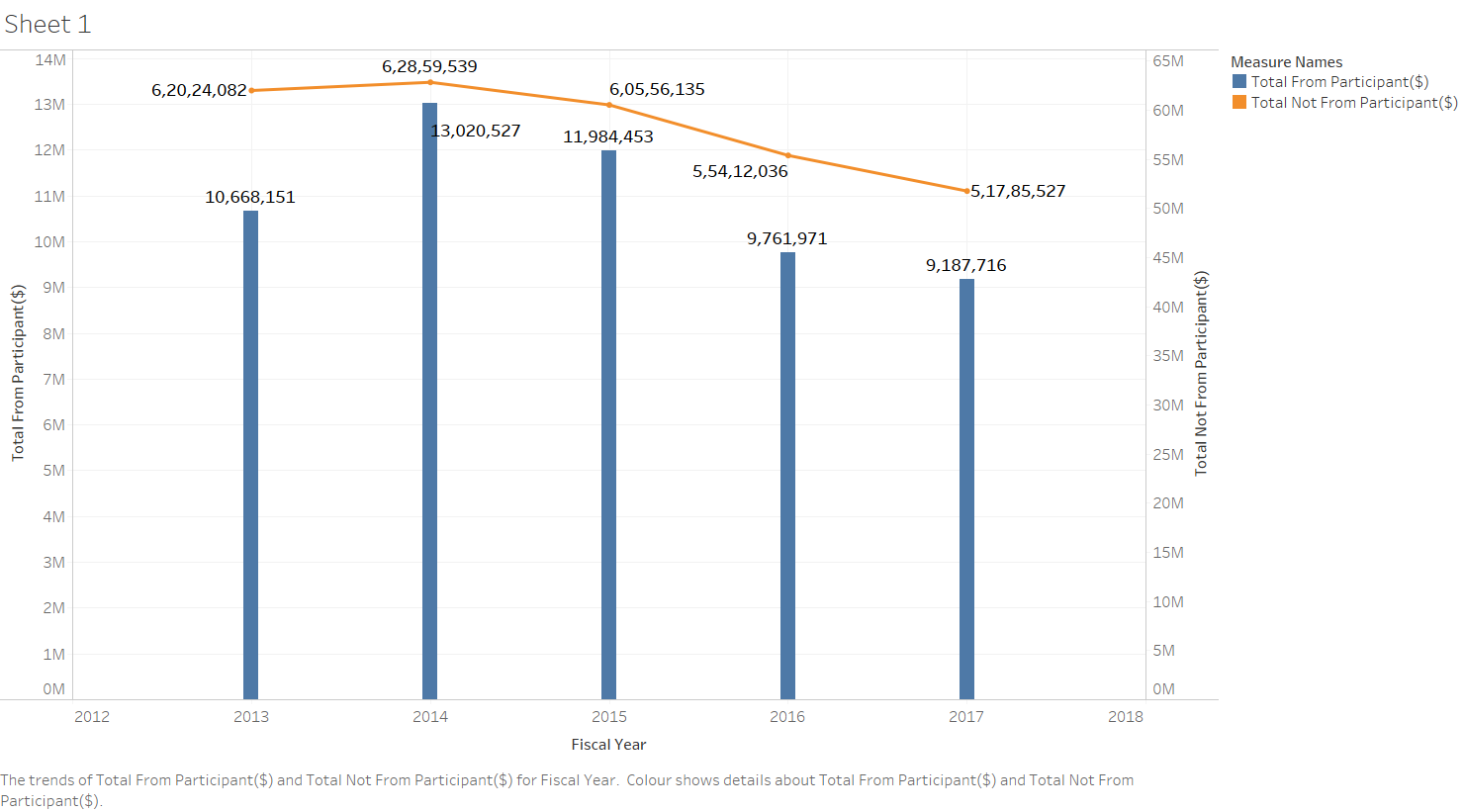


How many number of participants participated in event

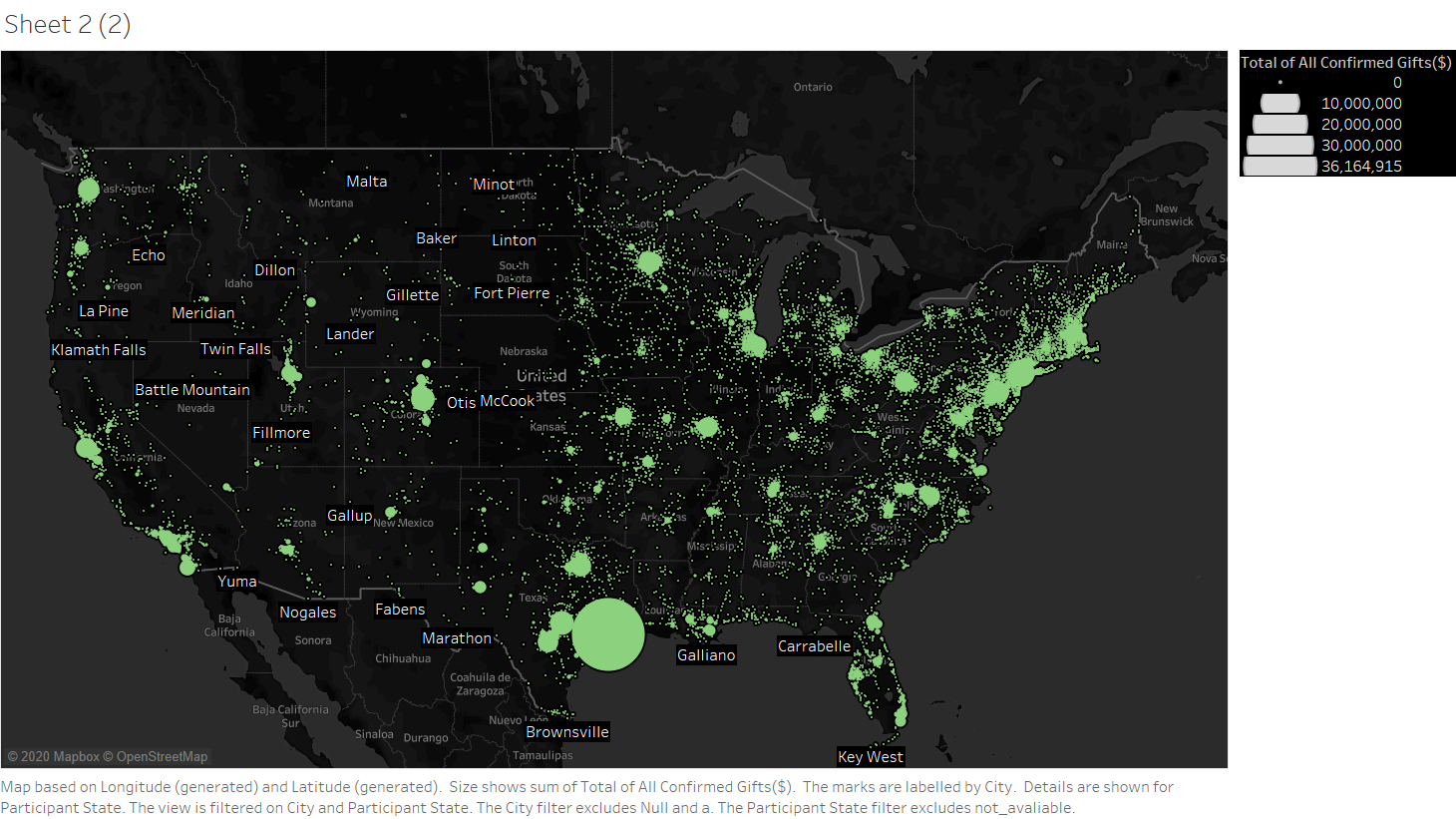
Over 5 years (2013-2017) ?



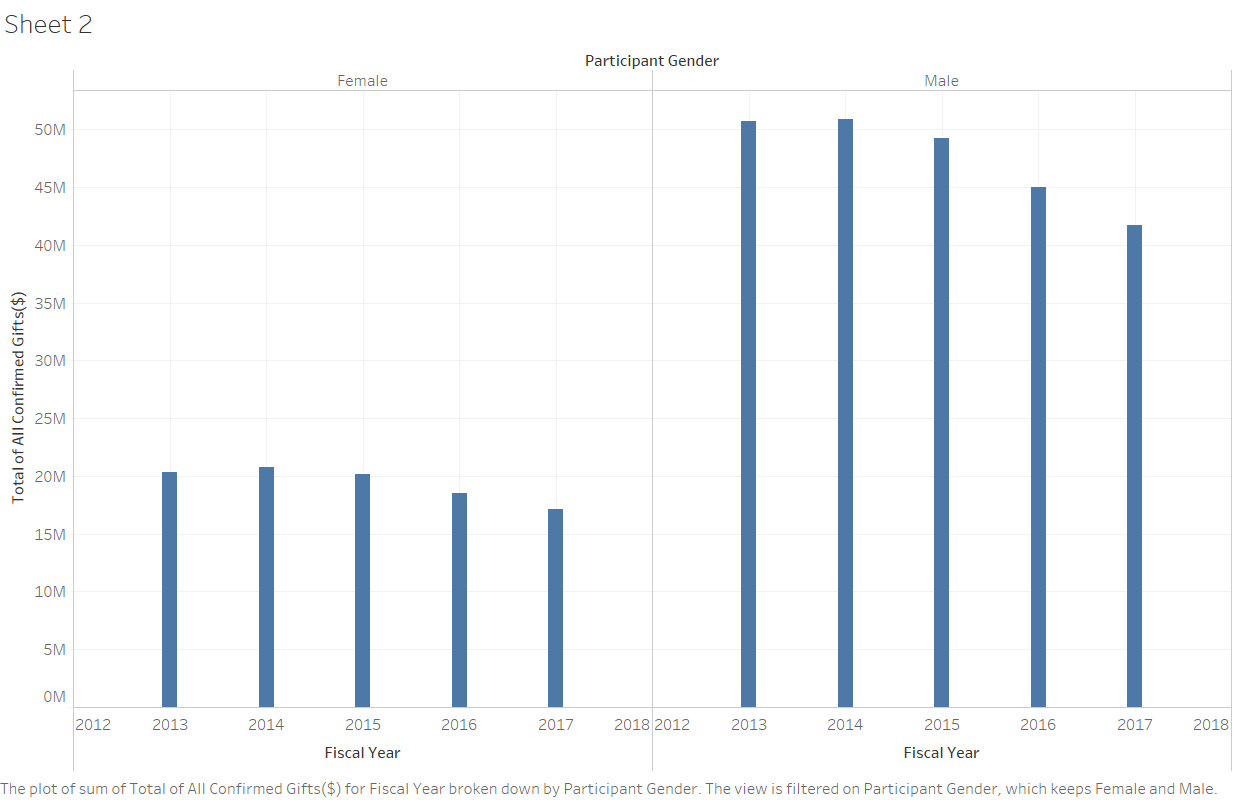
What is the donation total amount from participants and total amount of who are donated but not participated over 5 years ?



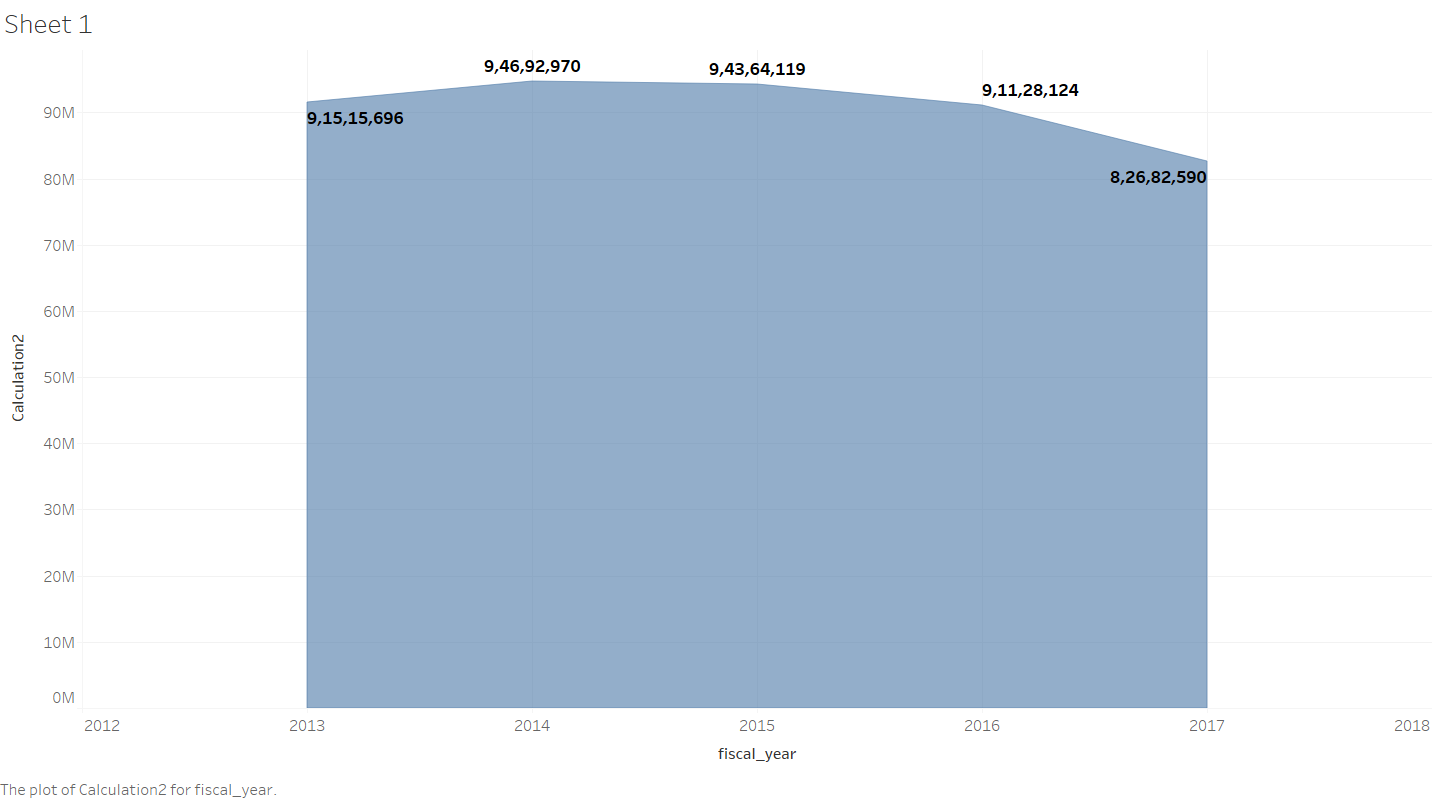
Which part of US cities has most participants who gave donation ?



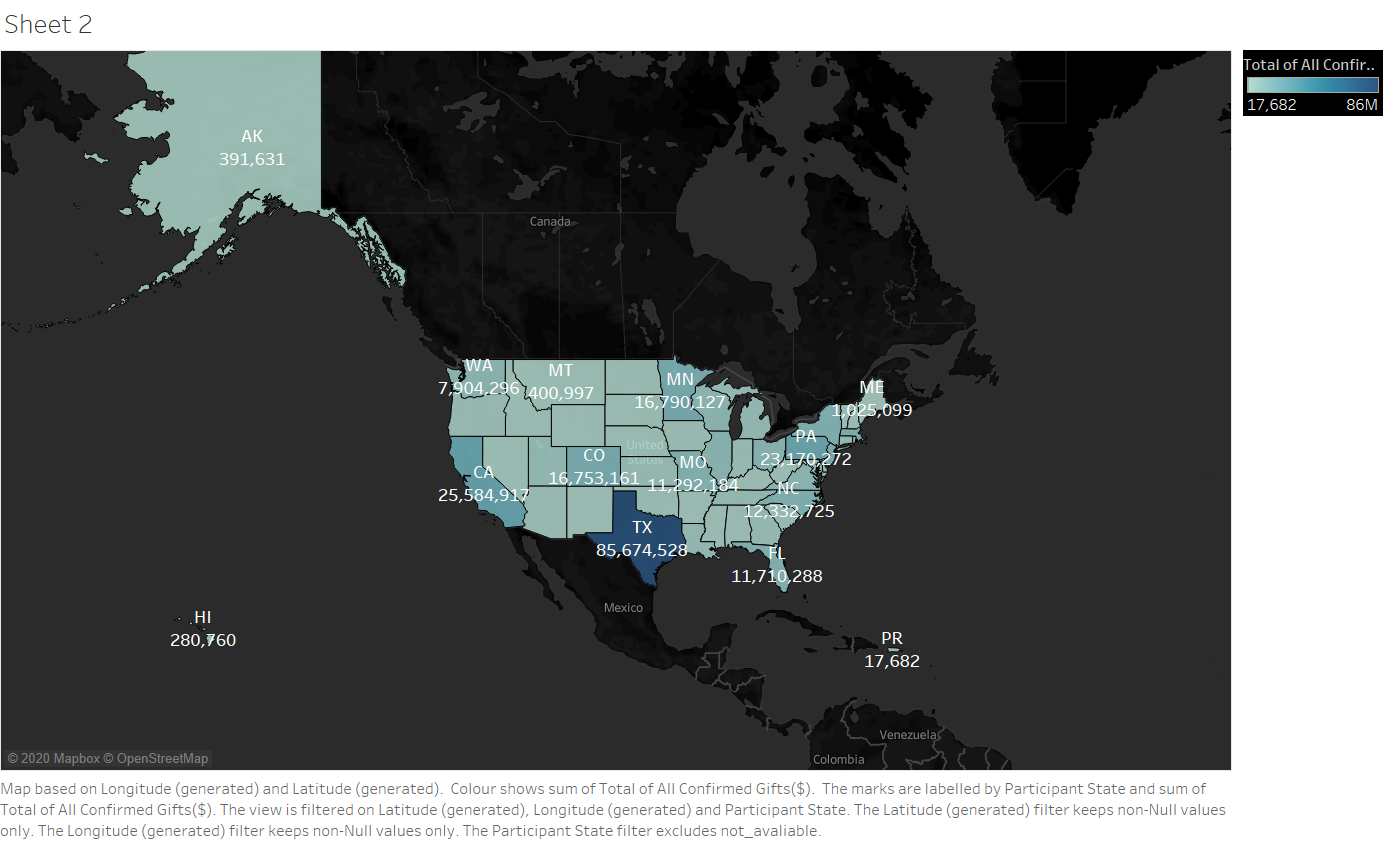
What is the gender of participants and who gave most of donation over 5 years and compare both male and female ?



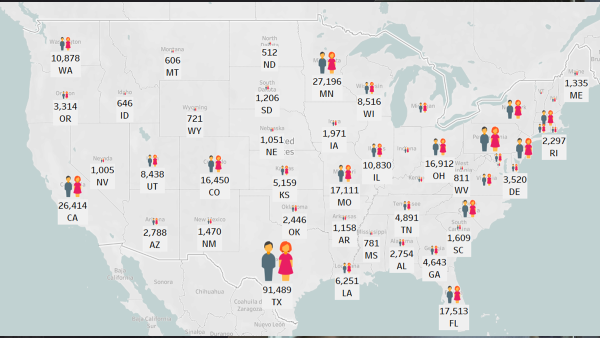
What is total amount of donations over 5 years?



Which state has the highest donation amount from participant and what is the amount ?



What is the count of participants from each state ?



We did more visualization just to see deeper insights of data and what they explain through visualizations

TIME TAKEN TO COMPLETE THE PROJECT

75 HOURS IN TOTAL

20 HOURS FOR UNDERSTANDING & CLEANING THE DATA

10 HOURS OF DFM & ROLAP QUERIES

25 HOURS OF HIVE & SPARK

10 HOURS OF TABLEAU VISUALISATION

10 HOURS OF PREPARING TEXT FILE & PRESENTATION