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.

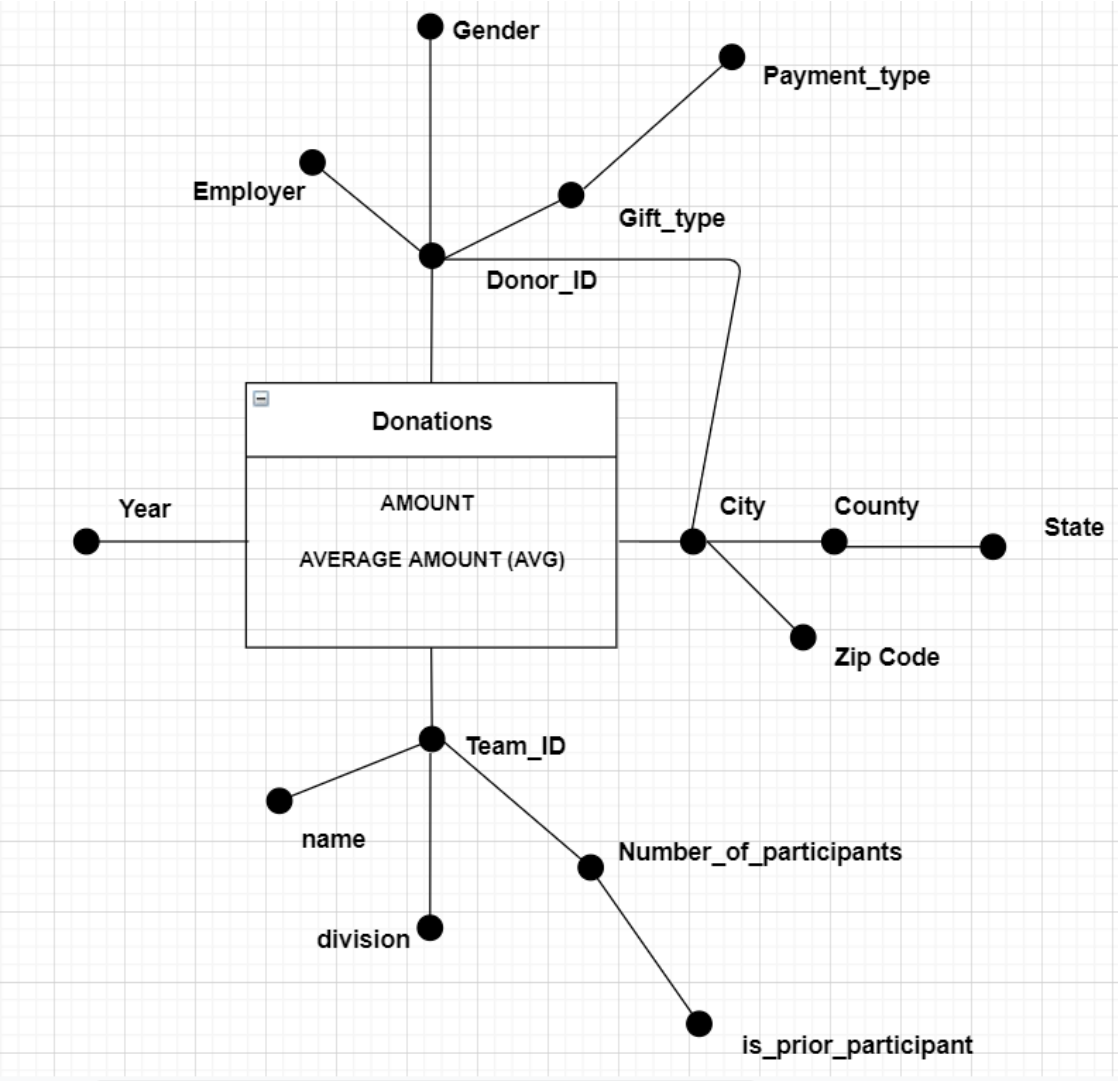
Conceptual schema

Fact table: Donation

Dimensions: Year, Donor, City, Team

Measures: Total Amount, Avg Amount (AVG)

Period: 2013-2017

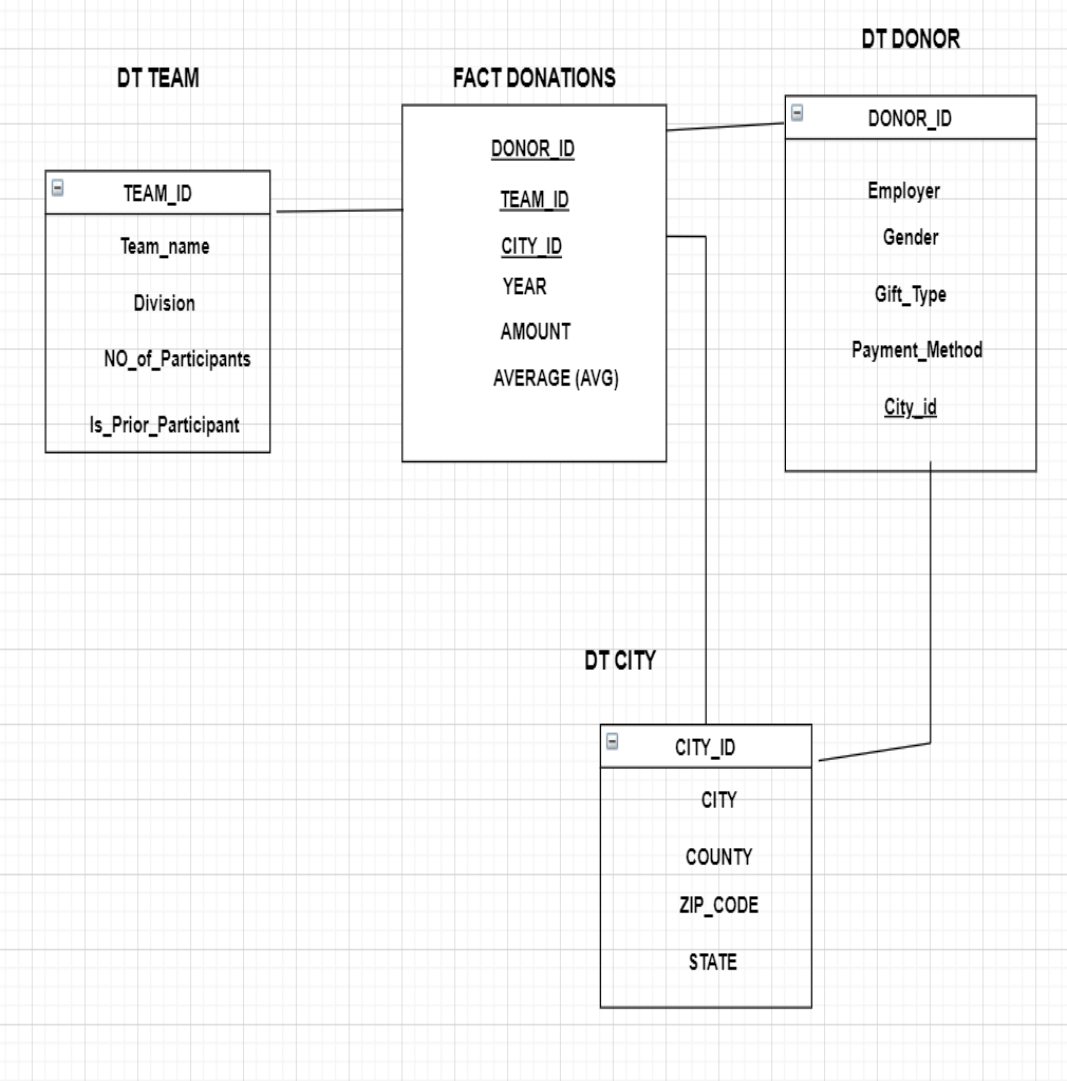


ROLAP

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()
```

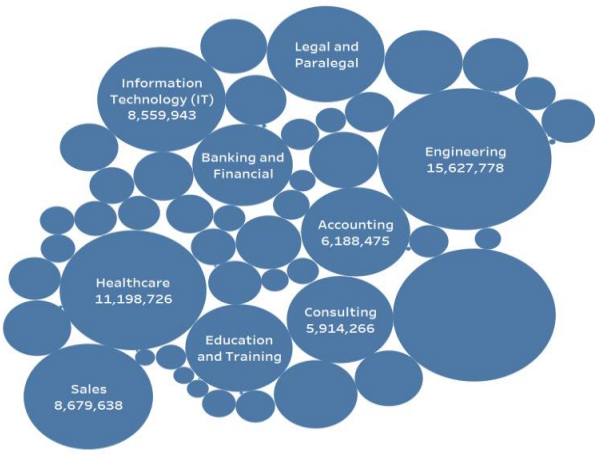
#11.get the events and the amount which has same value in year 2017 and count the number of times the event gave same amount

```
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

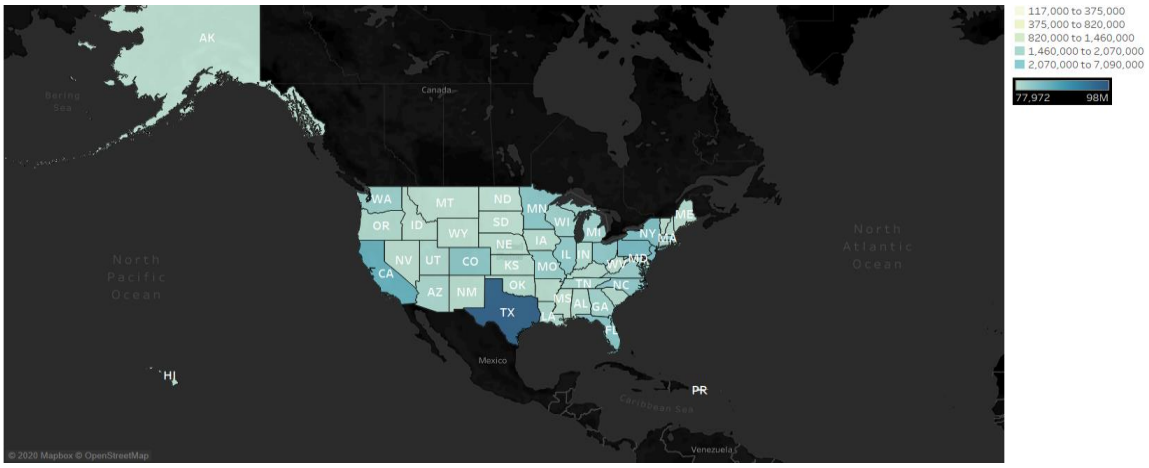
which Industries had the strongest involvement in Bike MS in the last five years and related occupations who are responsible for most of bike MS fund raising

Sheet 4

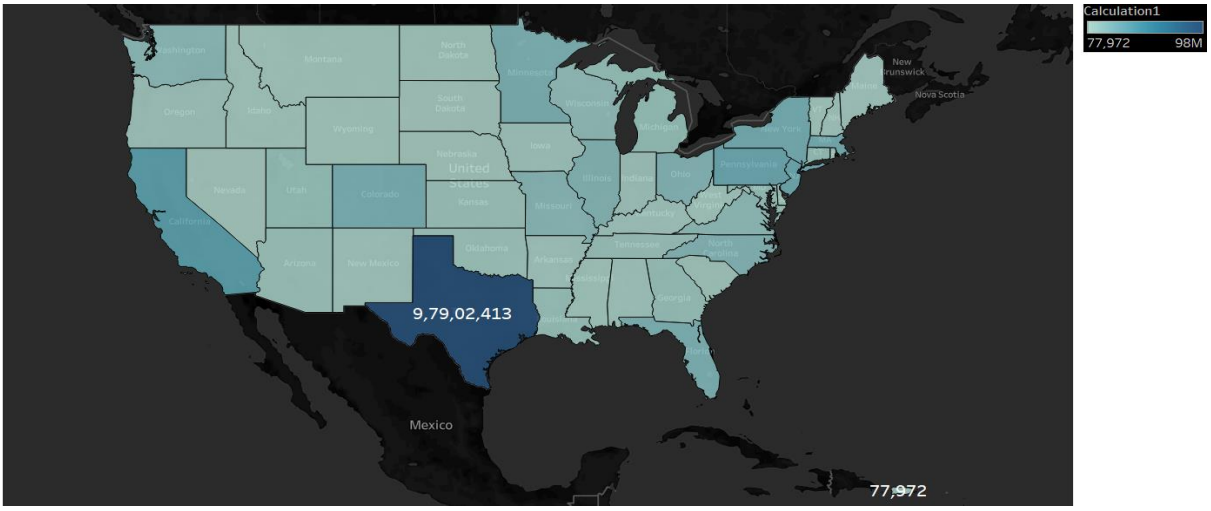


Participant Occupation and sum of Total of All Confirmed Gifts($). Size shows sum of Total of All Confirmed Gifts($). The marks are labelled by Participant Occupation and sum of Total of All Confirmed Gifts($). The view is filtered on Participant Occupation, which has multiple members selected.

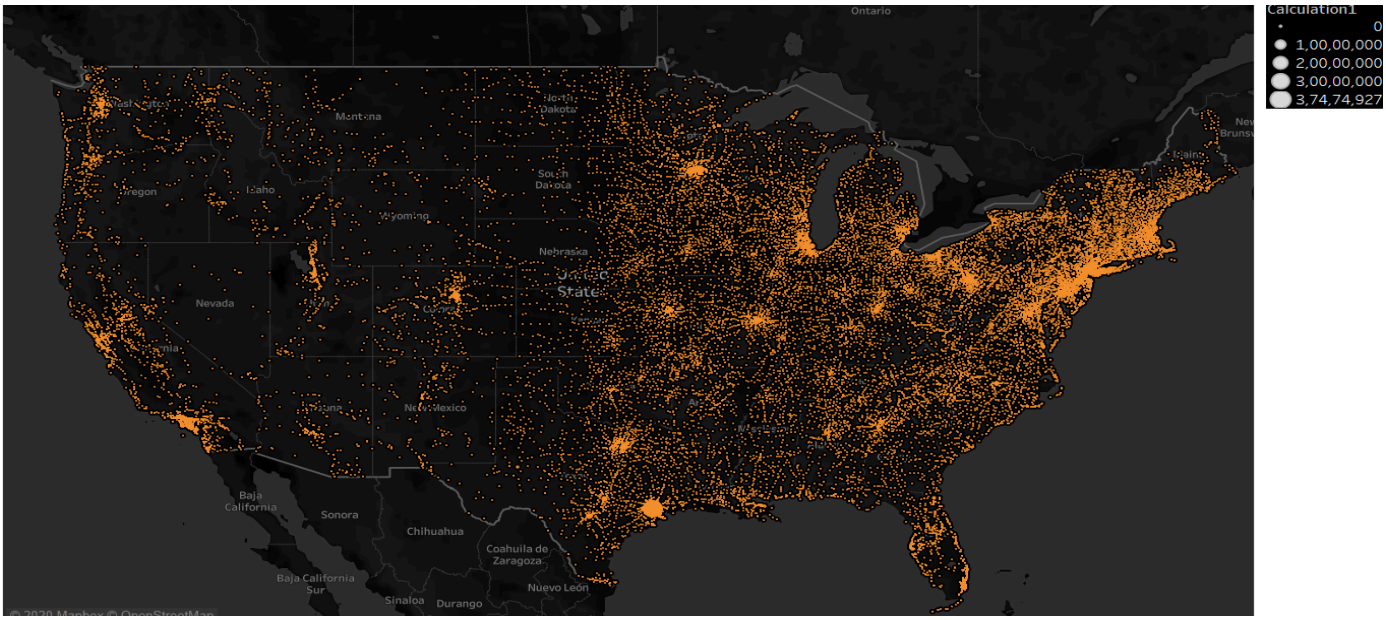
2) What are the states where the outbreak of MS is the most and which are the areas that are donating the most?



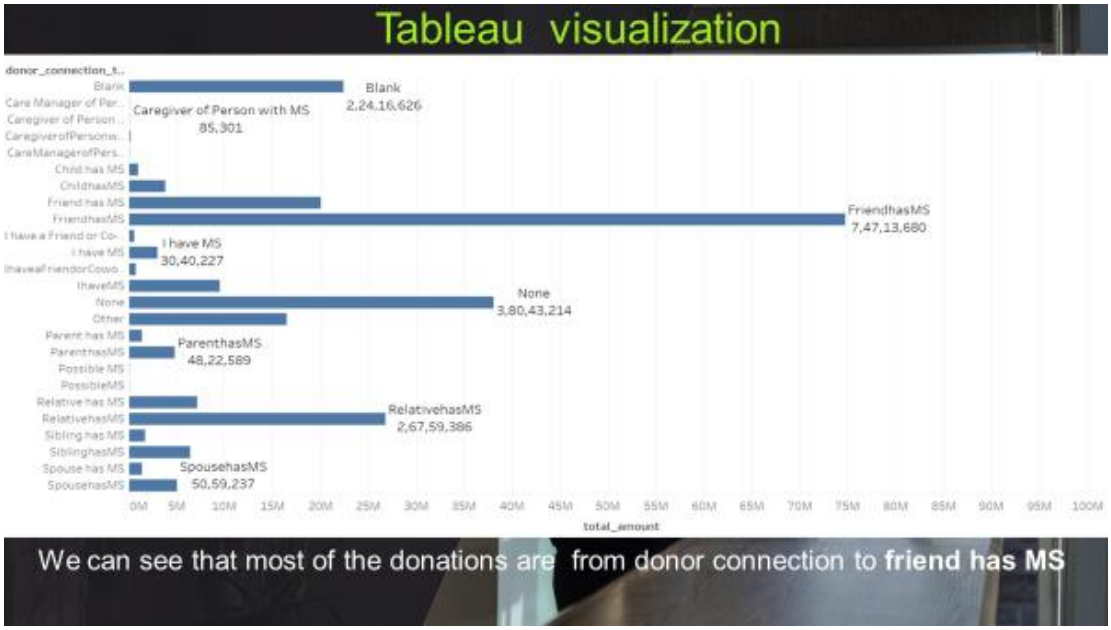
What are the states where the outbreak of MS fund raising is the high and low ?



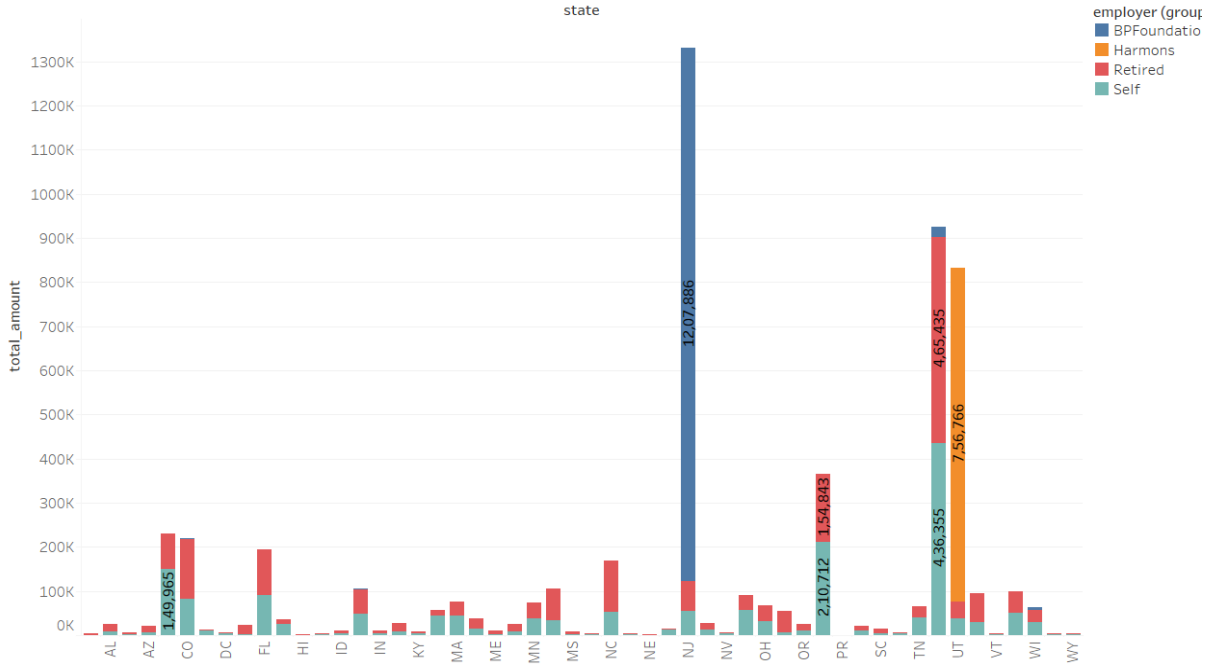
Which part of US cities has participated in event and raised fund?



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



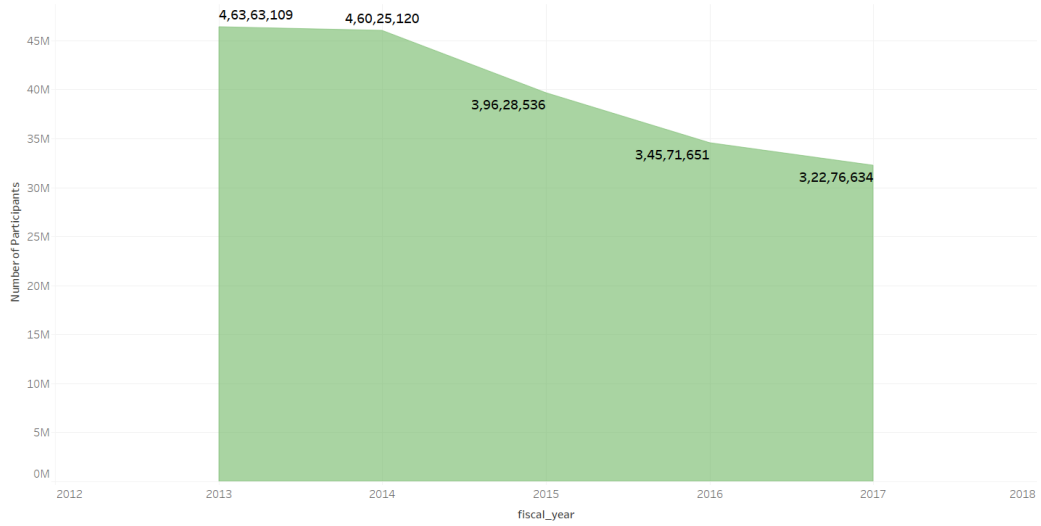
Who are the top employers who donated the fund and what are their states ?



How many number of participants participated in event

Over 5 years (2013-2017) ?

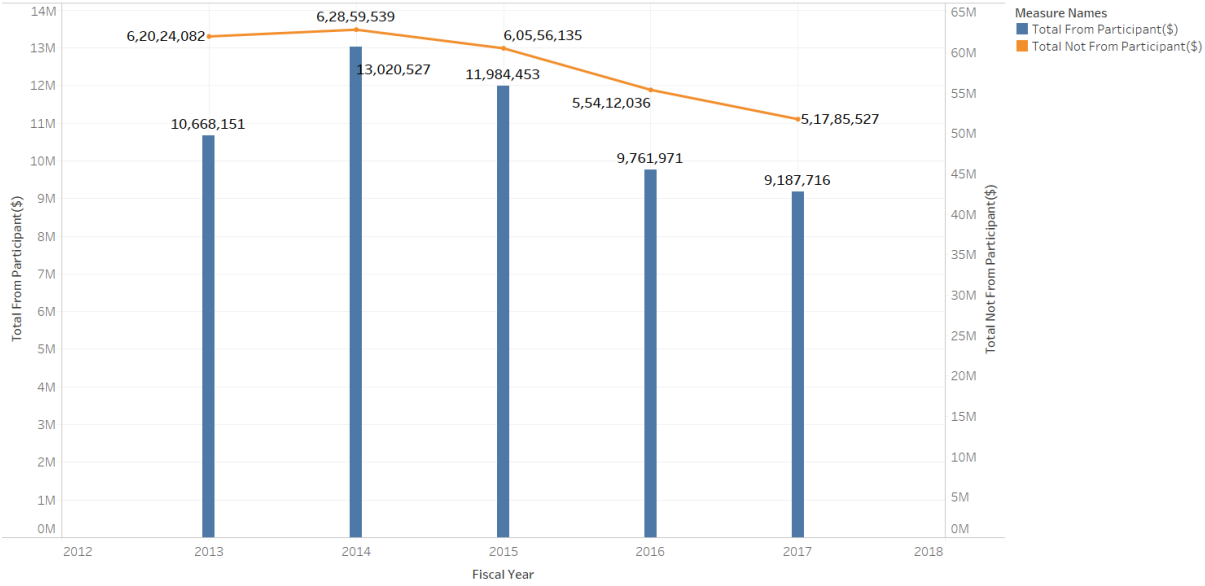
Sheet 5



The plot of sum of Number of Participants for fiscal_year.

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

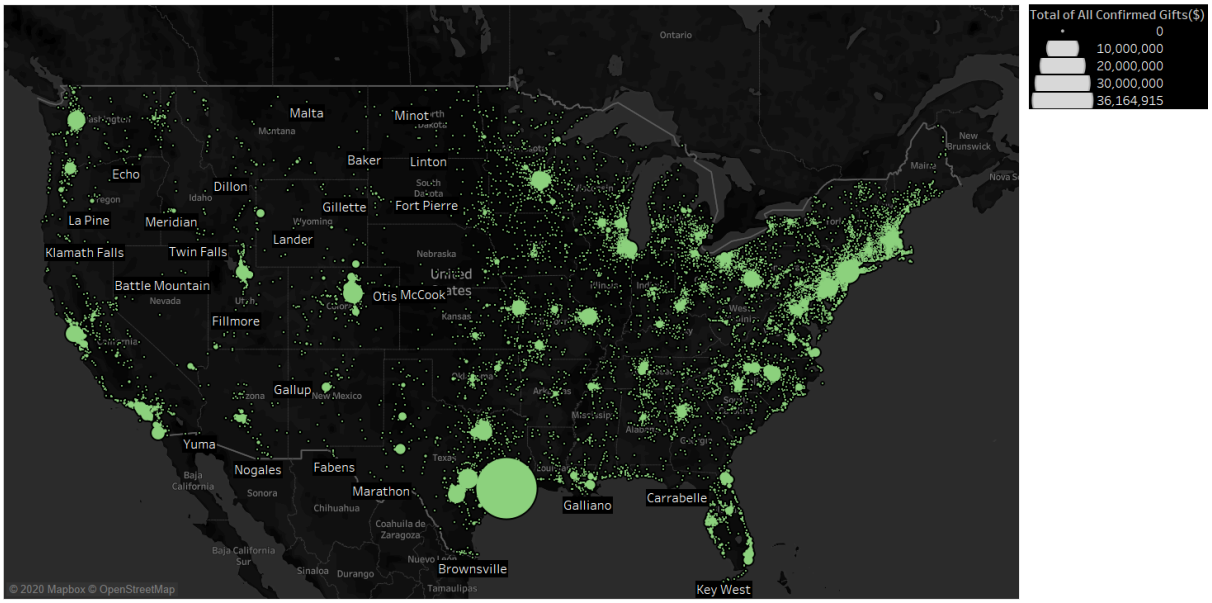
Sheet 1



The trends of Total From Participant($) and Total Not From Participant($) for Fiscal Year. Colour shows details about Total From Participant($) and Total Not From Participant($).

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

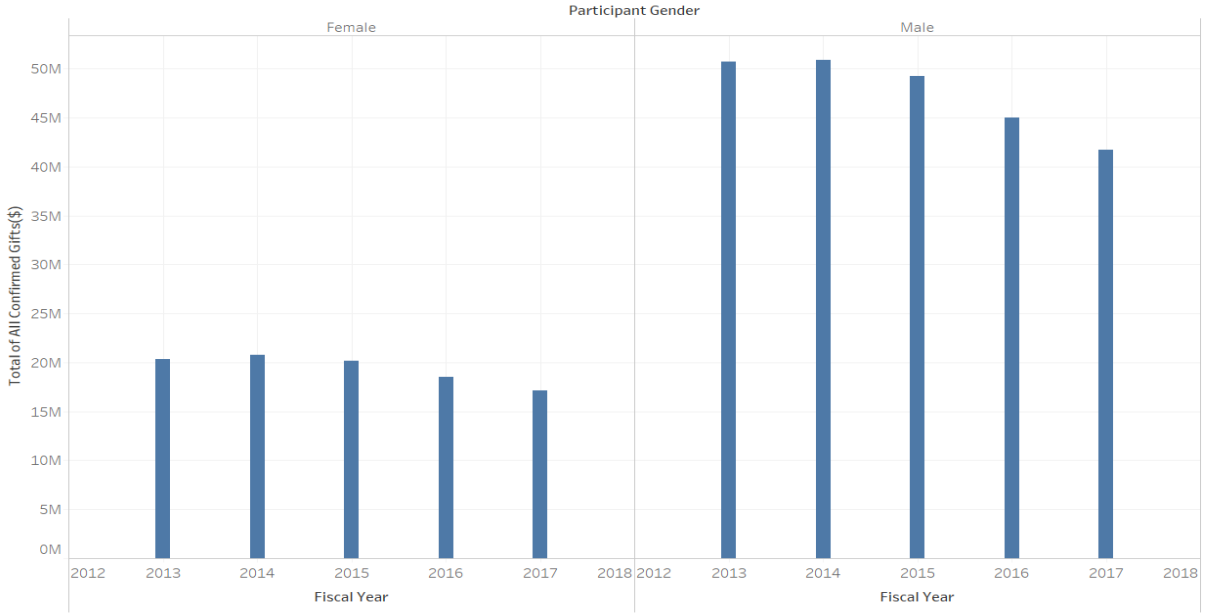
Sheet 2 (2)



Map based on Longitude (generated) and Latitude (generated). Size shows sum of Total of All Confirmed Gifts($). The marks are labelled by City. Details are shown for Participant State. The view is filtered on City and Participant State. The City filter excludes Null and a. The Participant State filter excludes not_avaliable.

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

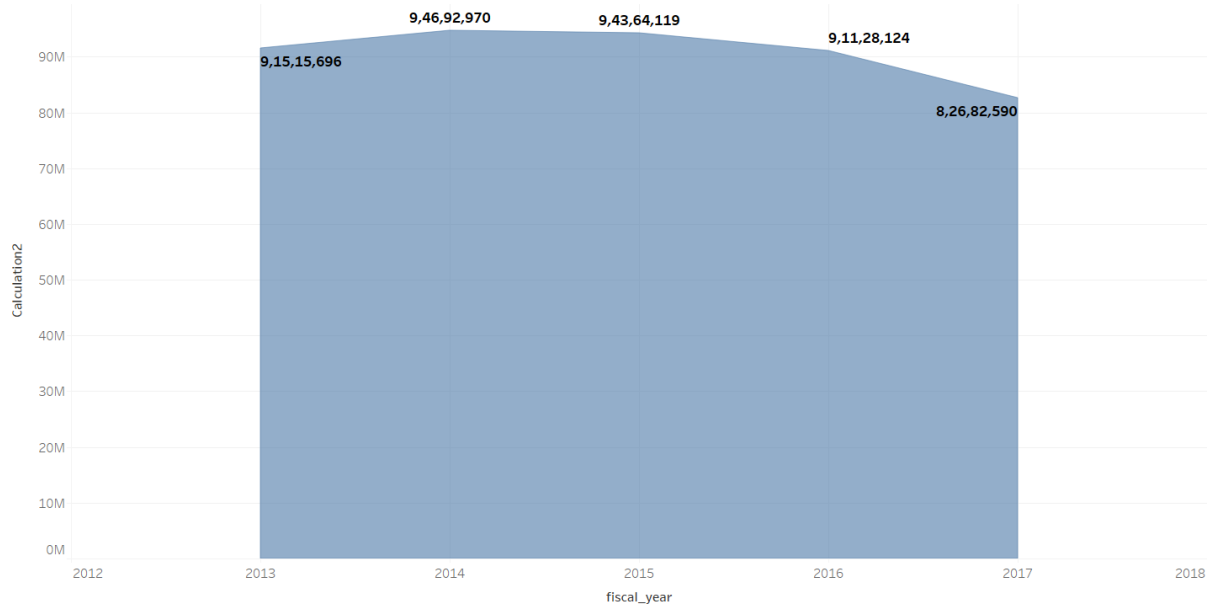
Sheet 2



The plot of sum of Total of All Confirmed Gifts($) for Fiscal Year broken down by Participant Gender. The view is filtered on Participant Gender, which keeps Female and Male.

What is total amount of donations over 5 years?

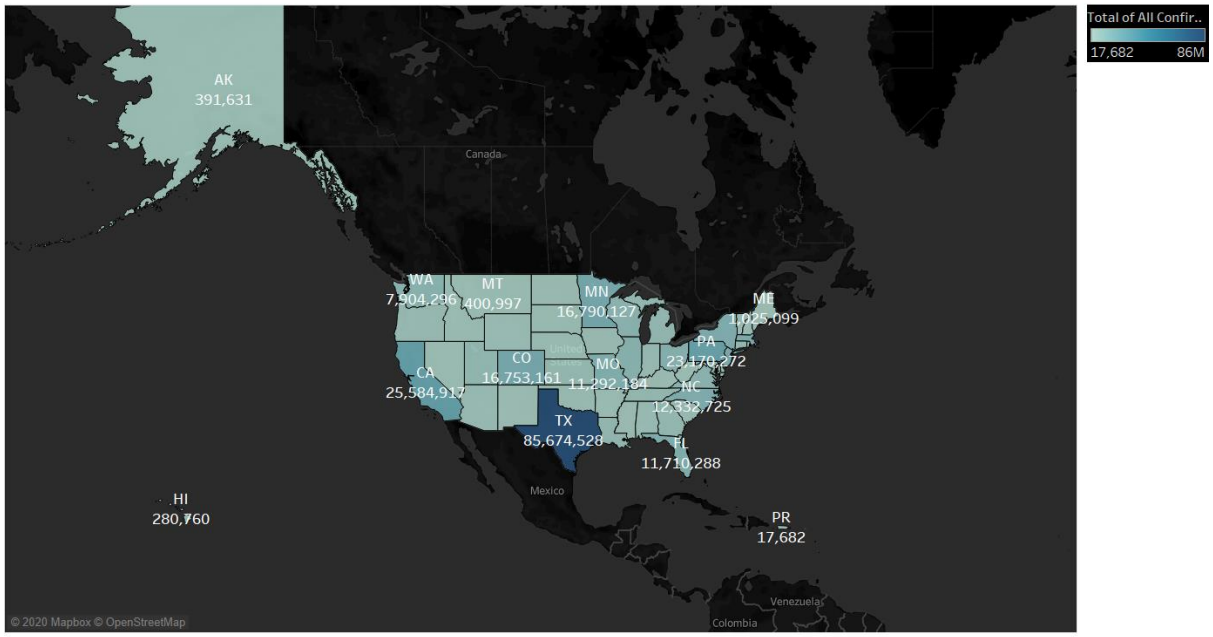
Sheet 1



The plot of Calculation2 for fiscal_year.

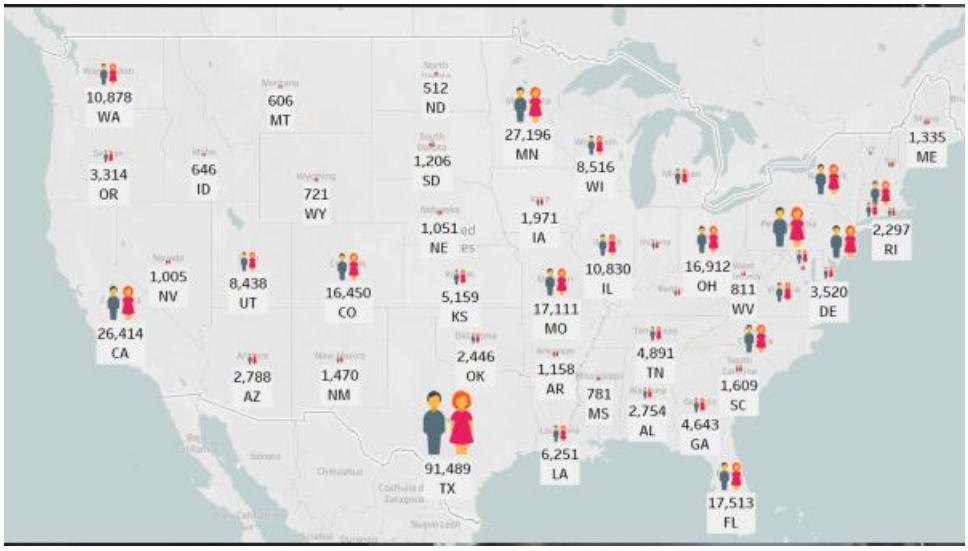
Which state has the highest donation amount from participant and what is the amount ?

Sheet 2



Map based on Longitude (generated) and Latitude (generated). Colour shows sum of Total of All Confirmed Gifts($). The marks are labelled by Participant State and sum of Total of All Confirmed Gifts($). The view is filtered on Latitude (generated), Longitude (generated) and Participant State. The Latitude (generated) filter keeps non-Null values only. The Longitude (generated) filter keeps non-Null values only. The Participant State filter excludes not_avaliable.

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