Data warehousing project
Bike MS
Xhesilda Qoshi
Ankur Kumar X

The scope of the project

There are given 2 datasets: “DVD Rental” and “Bike MS”.We selected the second one (Bike MS) dataset since it is more complex and has more data files.It is obvious that it takes more time to clean and to understand the data ,but on the other hand 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 analyses.

Since this course on DW 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 countries,gender and type of givings over years(2013-2017).

Conceptual schema

Fact table: Donation
Dimensions: Year,Donor ,City,Team
Measures: Total Amount, AvgAmount(AVG)
Period:2013-2017

Business questions

1.What industries have had the strongest involvement in Bike MS in the last five years and what occupations were responsible for most of our fundraising?

2.Which are the areas where the outbreak of MS is the most and which are the areas that are donating most?

3.Can we apply those opportunities to specific markets?

Preliminary workload expressed in natural language to understand what features we will take in consideration.

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

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

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

4.How did the giving donation change over years ordered by total amount?

5.What is the total amount of givings every team did based on their name and number of participants per year?

INSPECTION AND PROFILING

After this analyzation,we move to another step which consists of inspectation and profiling.

There are 6 data files (**Affiliates,BikeTeams,Donations,Events,National Teams,Participants**) which are data sources for the project.If we study and analyze all data files we can say that 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.

Also ,not all the columns in our data file are relevant related to our project ,so we are going to remove some of them.

We used Tableau 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.

-Since we are going to focus in our fact and dimension tables,we are looking for features that are relevant for our analysis. So we start removing all attributes that we don't need.

-We will continue removing all the empty columns or columns that have the same values and has no meaning for our project.

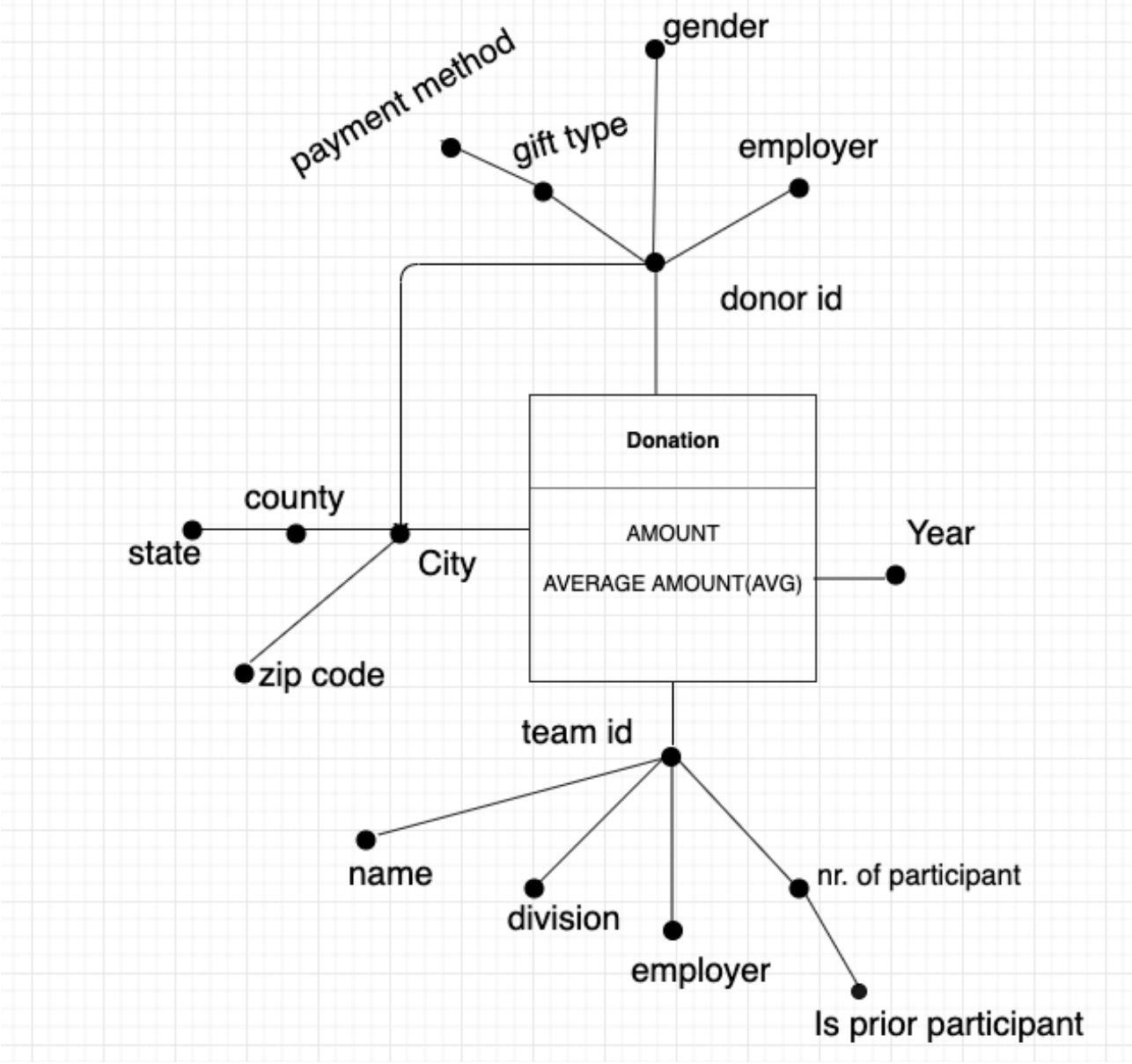
We didn't take in consideration affiliates and national teams data files because they contain some extra informations that are not relevant for our project's scope.

DFM – Conceptual Schema

Taking in consideration the business questions and our operational resources ,we decided for the main fact “**donation”**.Based on the requirement,we would analyze every donation from different aspects or points of view.Our dimensions are:**city,donor,year,team**.
To draw dfm schema we used a special tool online called draw.io. We used the same tool also for rolap.

Dynamicity

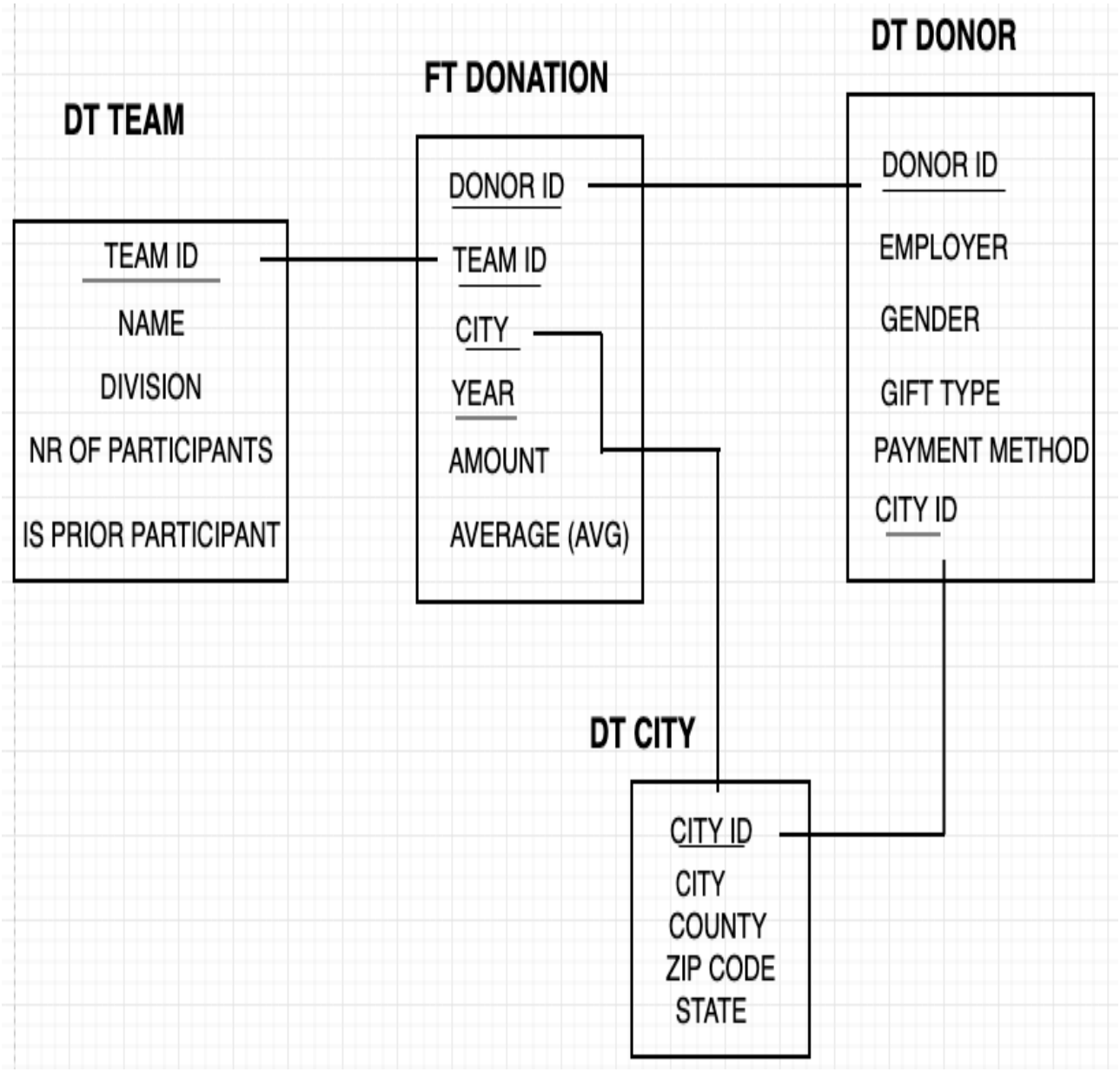
If we analyze the dynamicity in dimensions we considered donations between 2013-2017.The only thing that we want to compute is the average of donation.The time scenario is the third one yesterday-for -today (rollback)which is implemented in this case.This means that all the events are analyzed according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



Rolap

We know that in Rolap ,we can choose between star schema and snowflake schema.But in our case,we decided to use snowflake schema instead of star ‘s one.

In this case we will use less space to store dimension tables and also it provides better data quality which is more structures.



We are going to draw all the aggregation of our preliminary workload,in order to decide which views to materialize.
We will compute the primary event
Potential size =size of FT /(SUM of DT-s)
=3333018/(410.000x184x169x5)= $6x10^{-6}$

5 years of data.Total yearly volume of donations is 3333018

The materialized view in preliminary workload are:
P0={year,donorid,city,teamid}
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 extentions such as windows functions ,ranking ect.

First we have to create tables as below:

```
CREATE TABLE bikeMS.donation

(

teamid integer NOT NULL,

city character varying(30) NOT NULL,

donorid integer NOT NULL,

year integer NOT NULL,

givings double,

avggivings double,

CONSTRAINT b_key PRIMARY KEY
(team,donor,city,year),

);


CREATE TABLE bikeMS.cities
```

```
(

cityid integer NOT NULL,
city character varying(30) NOT NULL,

county character varying(30) NOT NULL,

zipcode character varying(12),

state character varying(30),

CONSTRAINT c_key PRIMARY KEY (cityid),

);


CREATE TABLE  bikeMS.donors

(

donorid character varying(30) NOT NULL,

employer character varying (30) NOT NULL,

gender character varying(30),
```

```
gifttype character varying(30),

paymentmethod character varying(30),

CONSTRAINT d_key PRIMARY KEY(donorid),

CONSTRAINT c_key FOREIGN KEY(cityid) REFERENCES
city(cidyid),

);


CREATE TABLE bikeMS.teams

)

name character varying(30) NOT NULL,

division character varying(30) NOT NULL,

employer character varying (30),

memberteam character varying(30),

ispriorparticipant boolean,
```

```
CONSTRAINT t_key PRIMARY KEY (teamid),

);
```

Views:

```
CREATE VIEW PUBLIC.Avgyear
 AS select event_id, fiscal_year, total_amount from
public.ankur
group by fiscal_year,event_id,total_amount;


Create view public.city as select
donor_city,donor_state,donor_county,
(gift_amount+additional_gift_amount) as
sum,(gift_amount::float+additional_gift_amount::flo
at)/2 as Average
from public.donation;

Create view gender as select
donor_gender,(gift_amount+additional_gift_amount)
as
sum,(gift_amount::float+additional_gift_amount::flo
at)/2 as Average, fiscal_year
from public.donation
```

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

```
select sum(gift_amount), is_prior_participant,
fiscal_year
 from Donation
Group by is_prior_participant, Fiscal_year,
gift_amount;
```

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

```
select max(sum),donor_state,donor_city from
public.city

 group by donor_state,donor_city
```

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

```
select *,avg(total_amount) from public.avgyear
group by event_id,fiscal_year,total_amount limit 500
```

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

```
select max(sum), max(average),donor_gender from
public.gender
group by donor_gender limit 1000;
```

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

```
select max(sum),
max(average),donor_gender,fiscal_year from
public.gender
group by fiscal_year,donor_gender limit 1000
```

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

```
Select
a.team_name,a.number_of_participants,b.total_amo
unt
from public.bike_teams a left outer join public.ankur
b on a.event_id=b.event_id
limit 1000
```

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

-Computing rankings and partitioning

```
Select p.participant_connection_to_ms,
a.total_amount, dense_rank() over( Partition by
a.fiscal_year)
from public.participants p left join public.ankur a on
p.event_id=a.event_id
where a.fiscal_year=2013
```

-Computing cumulative totals (window framing)

```
Select donor_state,fiscal_year,
sum(net_transaction_amount)
OVER(ORDER BY donor_state RANGE BETWEEN
UNBOUNDED PRECEDING AND CURRENT ROW)
from public.donation
group by donor_state,
net_transaction_amount,fiscal_year
```

-Computing mobile aggregates [window framing]

```
select event_id,fiscal_year,
sum(net_transaction_amount),avg(net_transaction_amount::float)
OVER(Partition by event_id order by fiscal_year rows 1 preceding )
from public.ankur
group by
event_id,net_transaction_amount,fiscal_year
order by event_id
limit 10000
```

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.

Tables:

```
CREATE TABLE user27dw.bike_teams (
event_type STRING,
event_id INT,
team_id INT,
team_name STRING,
 team_captain_contact_id INT,
 team_division STRING,
company STRING ,
 number_of_participants INT ,
 total_fees_paid INT,
 team_total_confirmed INT ,
total_online_gifts INT,
 total_offline_confirmed_gifts INT,
total_offline_unconfirmed_gifts INT,
 team_goal INT,
confirmed_gifts_history INT,
previous_event_confirmed_gifts INT,
previous_event_team_members INT,
fiscal_year  INT,
 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
TBLPROPERTIES ( 'skip.header.line.count'='1');
```

Query to Load the data into the table

LOAD DATA LOCAL INPATH '/home/user27/dwproject/DW1-Bike_teams.csv' OVERWRITE INTO TABLE user27dw.Bike_teams;

Create table user27dw.Participants(

Fiscal_Year INT,
Internal_Event_Name STRING,
Participation_Type_Name STRING,
Team_Name STRING,
Team_Division STRING,
Team_ID INT,
Member_ID INT,
Registration_Status STRING,
Is_Team_Captain STRING,
Is_Prior_Participant STRING,
Total_of_All_Confirmed_Gifts INT,
Total_From_Participant INT,
Total_Not_From_Participant INT,
Number_from_Participant INT,
Number_Not_From_Participant INT,
Employer STRING,
Occupation STRING,

Participant_Connection_to_MS STRING,
State STRING,
County STRING,
City STRING,
Zip_Code STRING,
Registration_Type STRING,
Event_ID INT,
Participant_Gender STRING,
 Participant_Goal INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');

Load data local inpath '/home/user27/dwproject/DW1-Participants1.csv' OVERWRITE INTO TABLE user27dw.Participants;

Create table user27dw.Events(
Fiscal_Year INT,
Event_ID INT,
Event_Goal INT,
Active_Registrations INT,
 Inactive_Registrations INT,
Total_Fees_Paid INT,
Total_of_All_Confirmed_Gifts INT,
Total_Online_Gifts INT,

```
Teams INT,
Captains INT,
 Average_Team_Size INT,
Total_Offline_Confirmed_Gifts INT,
Self_Donors INT,
Total_From_Participant INT,
Non_self_Donors INT,
Total_Not_From_Participant INT,
 Total_Team_Gifts INT,
Total_Event_Gifts INT,
Total_Offline_Unconfirmed_Gifts INT,
Address STRING,
City STRING,
 State STRING,
Zip_Code STRING,
 Previous_Event_ID INT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
TBLPROPERTIES ( 'skip.header.line.count'='1');
```

Load data local inpath '/home/user27/dwproject/DW1-Events.csv' OVERWRITE INTO TABLE user27dw.Events;

CREATE TABLE USER27DW.Donation(
Event_ID INT,

Public_Event_Name STRING,
Fiscal_Year INT,
Campaign_Title STRING,
Campaign_ID INT,
Gift_Amount INT,
Gift_Type STRING,
Gift_Payment_Method STRING,
Offline_Status STRING,
Is_Registration STRING,
Donor_ConsID INT,
Donor_Member_ID INT,
Donor_Gender STRING,
Donor_Email_Status STRING,
Donor_City STRING,
Donor_State STRING,
Donor_County STRING,
Donor_ZIP STRING,
Donor_Employer STRING,
Donor_Connection_to_MS STRING,
Participant_Contact_ID INT,
Participant_Member_ID INT,
Registration_Active_Status STRING,
Participant_Goal INT,I
s_Prior_Participant STRING,
Is_Team_Captain STRING,
Additional_Gift_Amount INT,

Team_Name STRING,
Team STRING,
Original_Value_Transacted INT,
Net_Transaction_Amount INT,
Ledger_Transaction_Amount INT)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
TBLPROPERTIES ('skip.header.line.count'='1');

LOAD DATA LOCAL INPATH '/home/user27/dwproject/DW1-Donations1.csv' OVERWRITE INTO TABLE user27dw.Donation;

We run the same queries as in PostgreSQL

Q1.select sum(gift_amount), is_prior_participant, fiscal_year from Donation Group by is_prior_participant, Fiscal_year, gift_amount;

Q2. select max(gift_amount), donor_state, donor_city from donation group by donor_state, donor_city ;

Q3. select avg(gift_amount), fiscal_year, event_id from donation group by fiscal_year,event_id

Q4.select max(sum), max(average),donor_gender,fiscal_year from public.gender group by fiscal_year,donor_gender limit 1000;

Q5. Select select a.team_name,a.number_of_participants,b.gift_amount from bike_teams a left join donation b on a.event_id=b.event_id limit 100;

3 new queries in Hive

H1.What is the total amount of givings done by each team and their respective names in each state ?

select b.team_id, b.team_name, d.donor_state, d.gift_amount from bike_teams b left join donation d on b.event_id=d.event_id group by b.team_id,d.donor_state,b.team_name,d.gift_amount limit 100;

H2.Retrieve female donors and total amount of givings per year.

select donations,donorid,year,sum(givings)

over(partition by donations.donorid order by year)

from donations,donors

where donations.donorid=donors.donorid and

donors.gender=’female’

group by donations,donorid,year,givings;

H3.Retrieve the employer who have donated through credit card.

select disctinct employer

from donors

where payment method = ’credit card’ ;

SPARK SQL

Spark is an open source ,general-purpose distributed computing engine used for processing and analyzing

a large amount of data.It is also faster than Hive and is always a good option for scaling.

First we are going to perform olap queries.

OLAP 1-st

```
donation_df.select("gift_amount","is_prior_participant","fiscal_year").groupBy("gift_amount","is_prior_participant","fiscal_year").show();
```

OLAP 2-nd

```
city_df.select("sum","donor_state","donor_city").groupBy("donor_state","donor_city","sum").max("sum").show();
```

OLAP 3-rd

```
avgyear_df.select("*").groupBy("event_id","fiscal_year","total_amount").avg("total_amount").show();
```

OLAP 4th

```
gender_df.select("average","donor_gender",col("sum").cast("double")).groupBy("donor_gender",).max("sum","average").show();
```

OLAP 5th

```
gender_df.select("average","donor_gender",col("sum
").cast("double"),"fiscal_year").groupBy("donor_gend
er","fiscal_year").max("sum","average").show();
```

We are going to perform 5 new queries in spark.

Q1.

```
events_df.select("average_team_size","total_from_p
articipant","state").where(col("state").isNotNull()).gr
oupBy("average_team_size","state").sum().orderBy("
state").show();
```

Q2.

```
particpants_df.select("occupation","total_from_parti
cipant","participant_gender").where(col("occupation
").isNotNull()).groupBy("occupation","total_from_par
ticipant","participant_gender").sum().orderBy("occup
ation").show();
```

Q3.

```
team_df.filter(col("fiscal_year").startswith("2013.0"))
.rollup("fiscal_year","event_type",
team_df.total_offline_confirmed_gifts).count().wher
```

```
e(col("event_type").isNotNull()).orderBy("fiscal_year"
,"event_type").show()
```

Q4.

```
donation_df.rollup("fiscal_year","donor_connection_
to_ms",
donation_df.transaction_amount).sum("transaction_
amount").orderBy("fiscal_year","donor_connection_t
o_ms").limit(100).show()
```

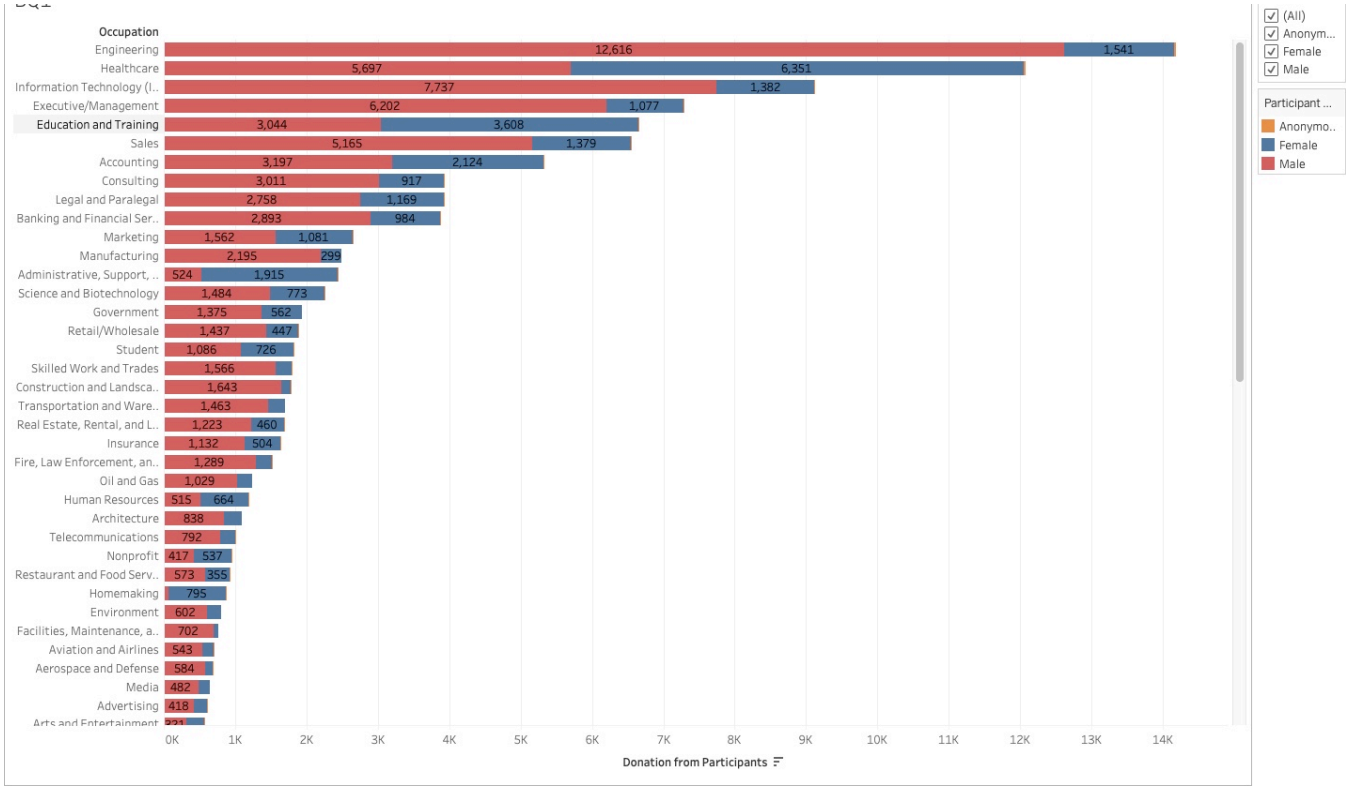
Q5.

```
win_spec =
Window.partitionBy('occupation').orderBy('occupatio
n').rowsBetween(-sys.maxsize, 0)
cum_sum =
particpants_df.select("occupation","total_from_parti
cipant").withColumn('cumsum',
F.sum(particpants_df.total_from_participant).over(wi
n_spec))
cum_sum.show();
```

Tableau

The first business question in which we are interested in:

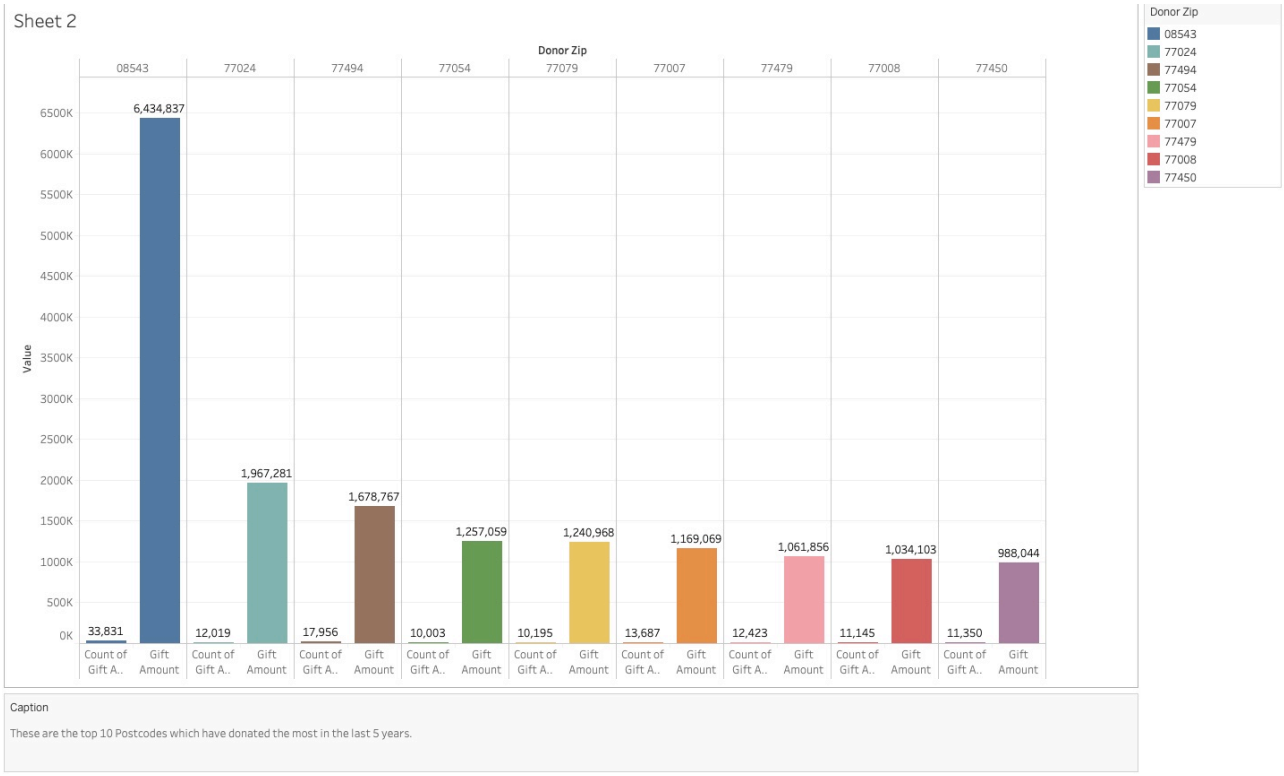
1.What industries have had the strongest involvement in Bike MS in the last five years and what occupations were responsible for most of our fundraising?



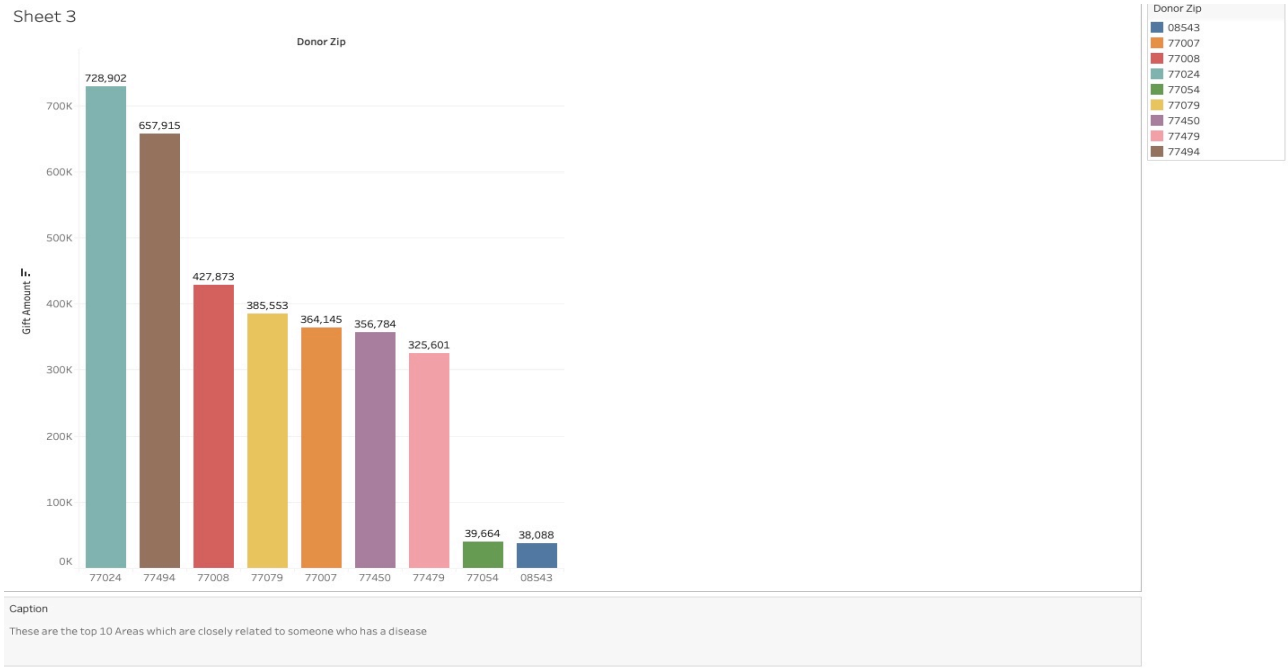
Bike MS is an event where people can participate as cyclist or donors.The visualization of the data shows that the major occupation of participants are Engineering.We can also define the top 10 occupations of participants:Engineering,Healthcare,IT,Executive/Management,Education and training,Sales,Accounting,Consulting,Legal and Paralegal,Banking and Financial… that are responsible for the majority of fundraising.The red colour shows that men are responsible for a larger sum of donations and also because most

of participants are men.The blue colour shows that the average donation from women is not as much as men.

2. Which are the areas where the outbreak of MS is the most and which are the areas that are donating most?

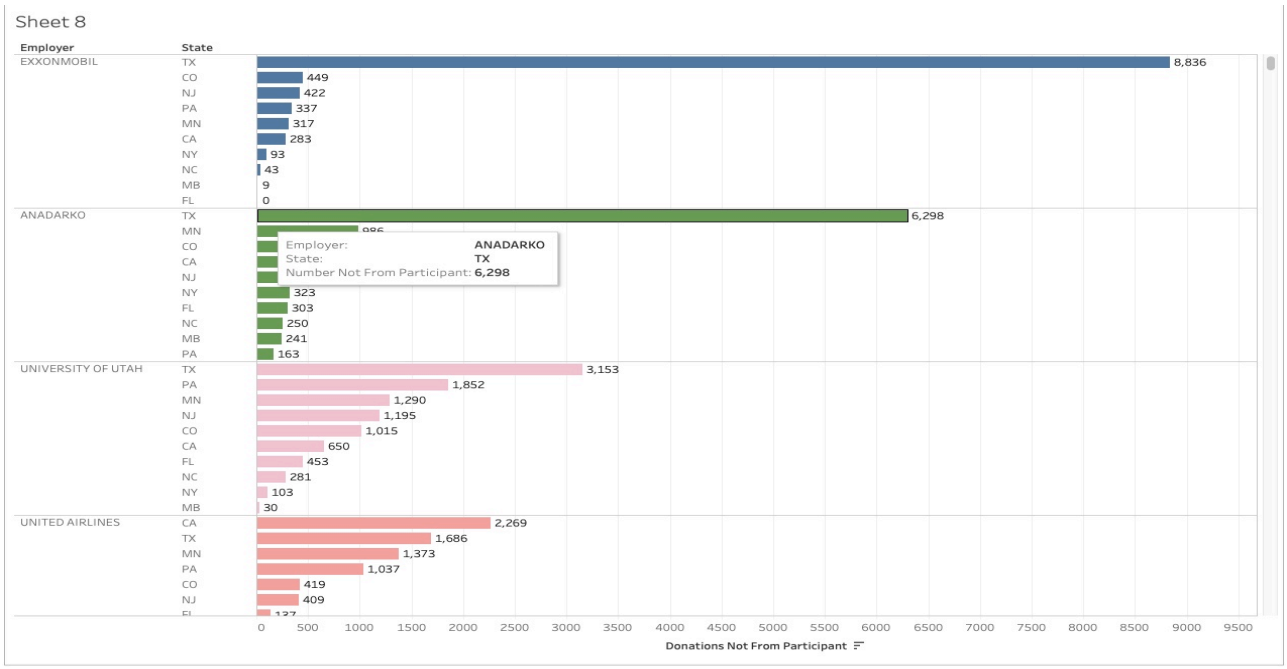


This visualization of data shows us the top 10 Postcode which have donated the most in the last 5 years (2013-2017).

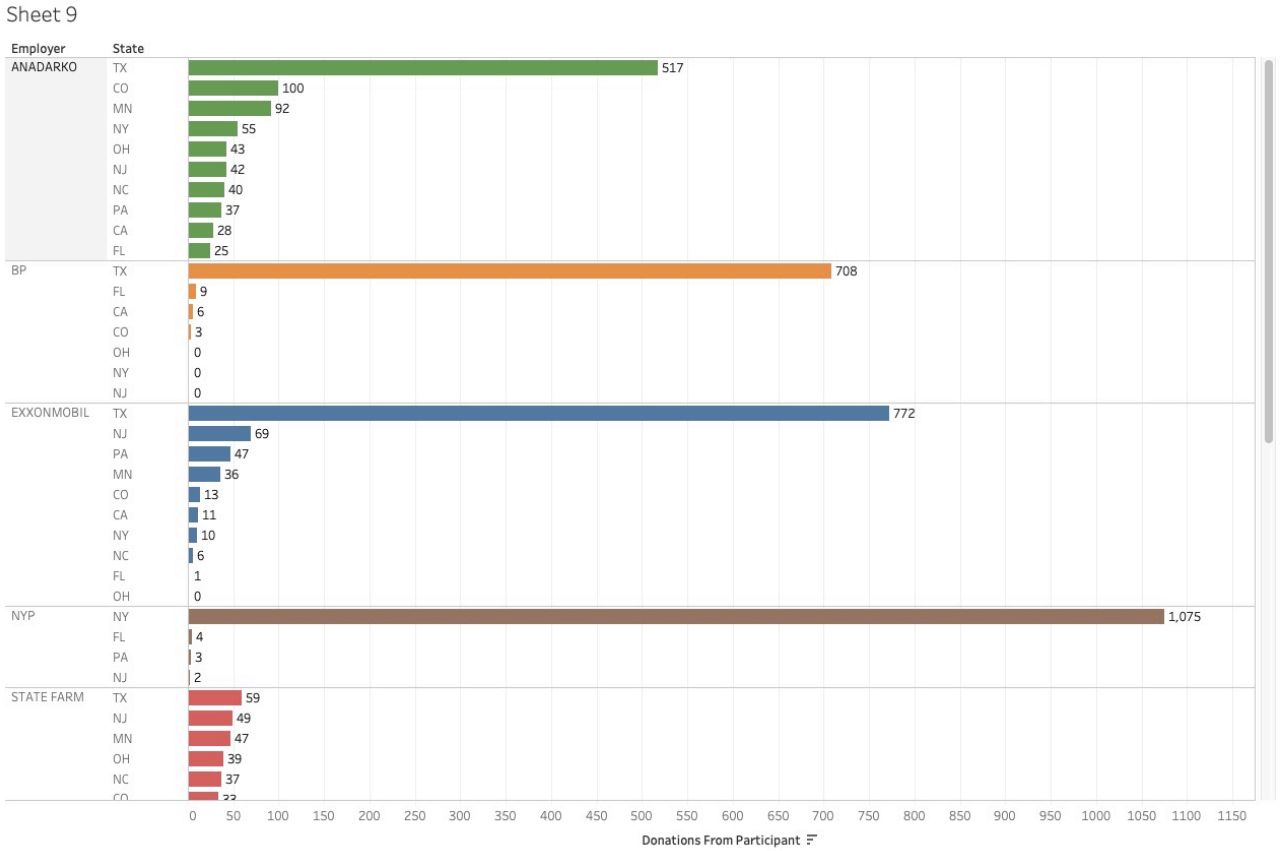


Here, we wanted to visualize that are the donations related to someone who have a connection with the MS disease?

3.Can we apply those opportunities to specific markets ?

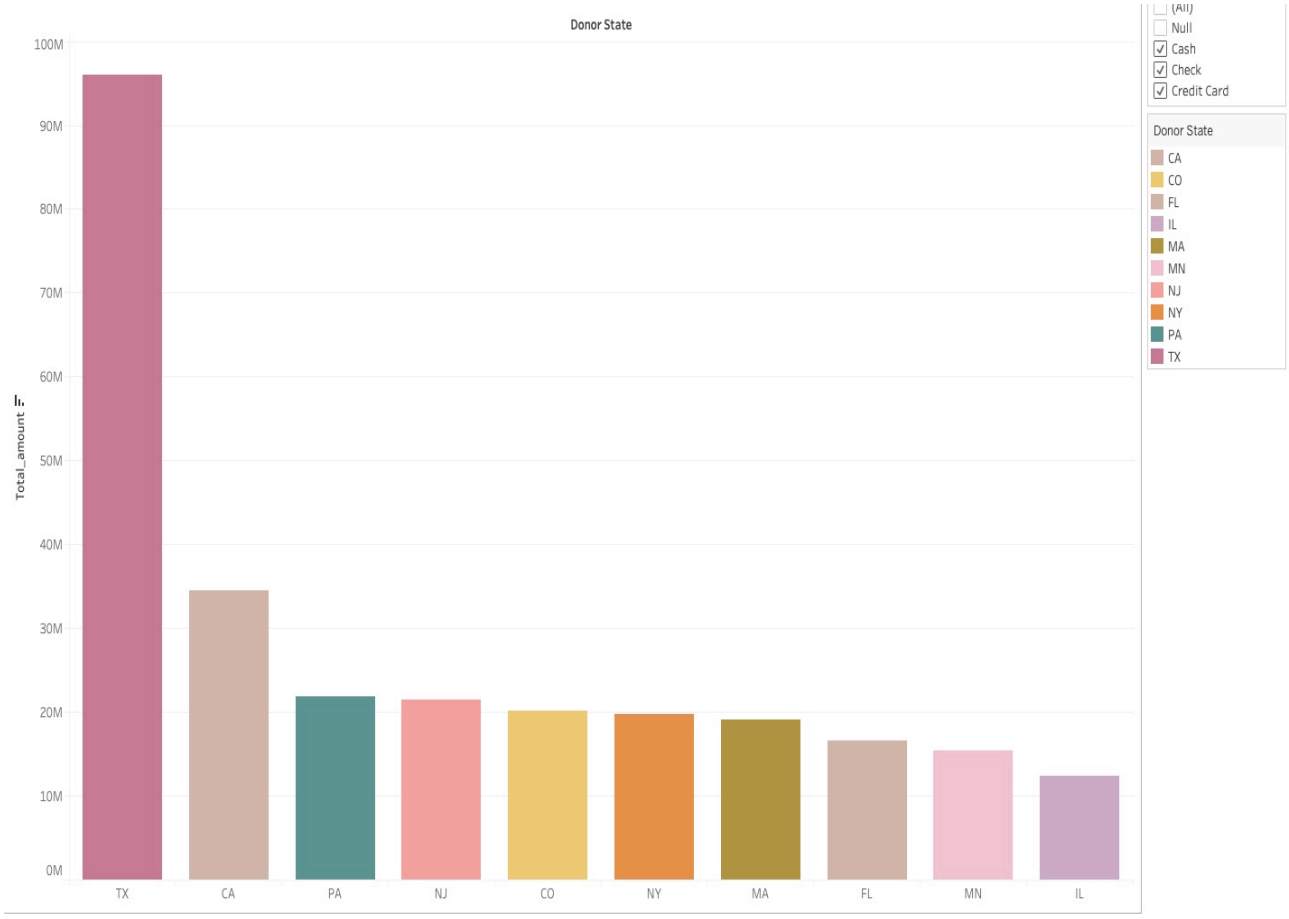


In the visualization of the data we can see the amount of donations that come from people that are not participant in the event but they just donate based on states and the largest employer. As we can see, Texas is the state where the most donations come from.



Here, we have the visualization of the amount of donations from people that participate on the event based on states and employer.

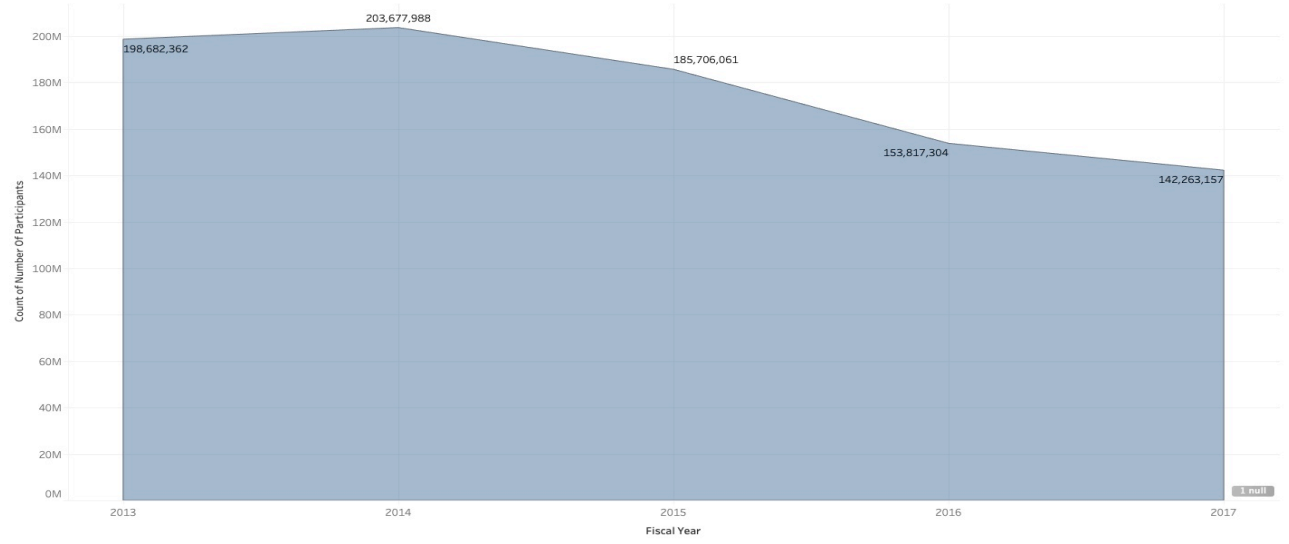
From the visualizations, we can safely conclude that ANADARKO and EXXONMOBIL have donated the most in Texas State and we have a lot of opportunities in other states to advertise and market.



-FROM THE PREVIOUS SLIDES, WE CAN SAFELY CONCLUDE THAT TEXAS HAS DONATED THE MOST IN THE PAST YEARS.

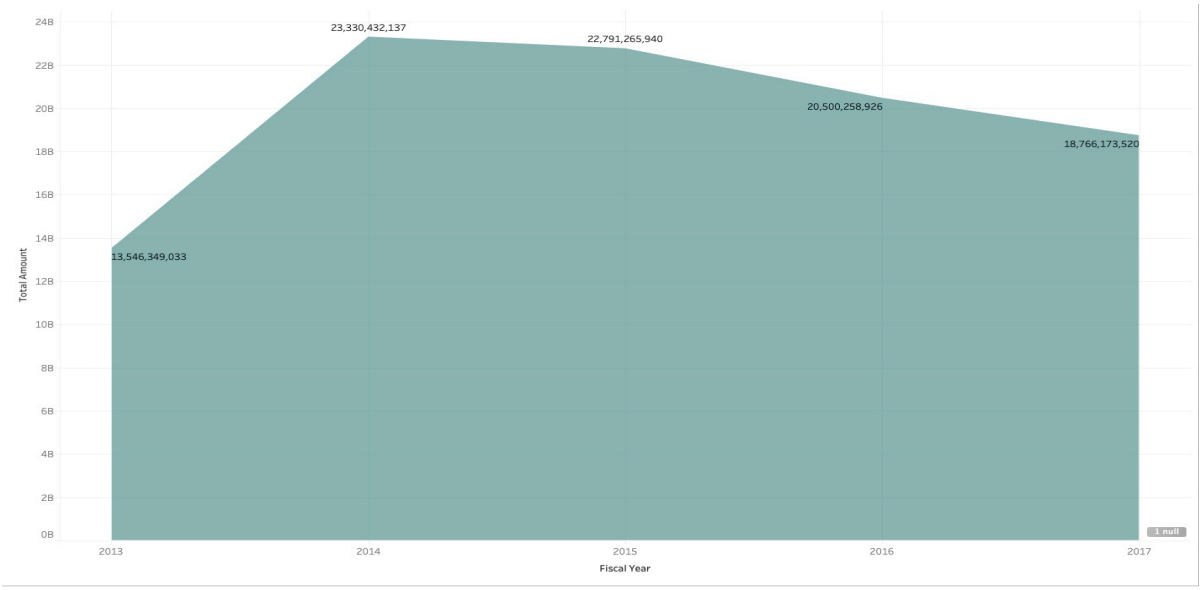
-WE CAN INVEST THE MONEY TO ADVERTISE ABOUT BIKE MS IN SOME OTHER STATES

4.Lets see how the participants have increased/decreased over the years



Caption	We have seen that over the years participants are declining. But, is the disease over from the society? NO! There are lots of state which are full of opportunities and where we can advertise about it to gain donations and the participants

5.Lets see the total amount of donations over the years



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