PROJECT

ON

BUSINESS ANALYTICS

Team Members:

Aditi Alankrita

Rujuta Dawkhar

Contents

[ABOUT THE DATASET 3](#_Toc501405504)

[APPROACH AND STEPS TAKEN 4](#_Toc501405505)

[CHALLENGES FACED IN THE PROCESS 4](#_Toc501405506)

[MySQL 5](#_Toc501405507)

[R Studio 8](#_Toc501405508)

[TABLEAU 8](#_Toc501405509)

[URL: 8](#_Toc501405510)

[VISULAIZATIONS 8](#_Toc501405511)

[REFERENCES 13](#_Toc501405512)

# ABOUT THE DATASET

There are 6 tables available in the database.

1. **CONSTITUENT**

* CONSTITUENT\_ID (PRIMARY KEY)
* AGE
* GENDER
* CITY
* STATE
* COUNTRY
* ZIPCODE
* EMAIL
* PHONE
* MARRIED

1. **EVENT**

* EVENT\_ID (PRIMARY KEY)
* EVENT\_NAME
* EVENT\_DATE

1. **EVENT\_ATTENDEE**

* EVENT\_ATTENDEE\_ID (PRIMARY KEY)
* CONSTITUENT\_ID
* EVENT\_ID

1. **FAMILY**

* FAMILY\_ID (PRIMARY KEY)
* CONSTITUENT\_ID
* RELATION\_CONSTITUENT\_ID
* RELATION\_AGE
* RELATION\_GENDER

1. **TRANSACTIONS**

* TRANSACTION\_ID (PRIMARY KEY)
* CONSTITUENT\_ID
* AMOUNT
* TRANSACTION\_DATE
* TRANSACTION\_TYPE
* FUND\_TYPE
* FUND\_USE

1. **VOLUNTEER**

* VOLUNTEER\_ID (PRIMARY KEY)
* CONSTITUENT\_ID
* VOLUNTEER\_DATE

# APPROACH AND STEPS TAKEN

We used tools such as **MySQL Workbench** to do basic analysis of the dataset. Initial steps were to use DESCRIBE command and understand the structure of the database. More queries were written to derive conclusions such as who donated how much amount, which state made the maximum donations, who volunteered for how many events, which age group donated the maximum amount etc. which are mentioned below and also attached in the query.sql file.

The query used to form the dataset is attached in the end of query.sql file.

After the basic analysis, a dataset containing the most important attributes was created and exported into **RStudio** on which we applied Logistic Regression to find out accuracy and performed Chi-squared tests

The database base was then connected to **Tableau** and we created various joins on the tables. We also tried “blending” concept in Tableau to visualize the trends and draw insights. We have tried using the forecast option in Tableau to predict the donations that will be made in 2018. In our dashboard we have also tried to use a filter for transaction year. If we select a particular year, all the graphs which have data related to year changes and the results are displayed only for the selected years.

## CHALLENGES FACED IN THE PROCESS

Creating a dataset in SQL with the required columns was a challenge. We had to take 2 queries and combine their results in excel.

The single query which were giving us the required columns (with the help of our prof), were not giving the results for all 2000 constituents. We were able to fetch only 1000 records.

Understanding “blend” is Tableau was a bit confusing, but after going through the class lecture, and readings, it was understandable.

# MySQL

Basic SQL queries were used to analyze the dataset given such as:

SELECT COUNT(DISTINCT(CONSTITUENT\_ID)) FROM CONSTITUENT;

SELECT COUNT(DISTINCT(CONSTITUENT\_ID)) FROM TRANSACTIONS;

*These queries were written to understand how many constituents are present and how many constituents made donations. It was inferred that even though the database has 2000 constituents only 1000 of them have made donations*

select B.CONSTITUENT\_ID , SUM(B.AMOUNT) AS AMT, A.GENDER from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

GROUP BY B.CONSTITUENT\_ID;

*This query was written to understand which constituent donated how much*

SELECT A.GENDER,SUM(B.AMOUNT) from CONSTITUENT A join TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

GROUP BY A.GENDER;

*This query was written to understand which gender contributed how much. It was inferred that Females have contributed more than Males*

SELECT A.EVENT\_ID, count(A.EVENT\_ATTENDEE\_ID), B.EVENT\_DATE

FROM EVENT\_ATTENDEE A, EVENT B

WHERE A.EVENT\_ID = B.EVENT\_ID

GROUP BY A.EVENT\_ID;

*The above query gives us a list of events that have occurred along with the count of attendees and on which date it happened.*

SELECT A.CONSTITUENT\_ID, COUNT(B.EVENT\_ID), A.GENDER, A.AGE FROM

CONSTITUENT A, EVENT\_ATTENDEE B

WHERE A.CONSTITUENT\_ID = B.CONSTITUENT\_ID

GROUP BY A.CONSTITUENT\_ID;

*The above query gives us a list of constituent ids, count of events that constituent has attended along with their age and gender.*

select SUM(B.AMOUNT) AS AMT\_19 from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

WHERE (A.AGE < 20);

select SUM(B.AMOUNT) AS AMT\_20\_29 from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

WHERE (A.AGE >= 20 and A.AGE <= 29);

select SUM(B.AMOUNT) AS AMT\_30\_39 from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

WHERE (A.AGE >= 30 and A.AGE <= 39);

select SUM(B.AMOUNT) AS AMT\_40\_49 from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

WHERE (A.AGE >= 40 and A.AGE <= 49);

select SUM(B.AMOUNT) AS AMT\_50\_59 from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

WHERE (A.AGE >= 50 and A.AGE <= 59);

select SUM(B.AMOUNT) AS AMT\_60\_69 from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

WHERE (A.AGE >= 60 and A.AGE <= 69);

select SUM(B.AMOUNT) AS AMT\_70\_79 from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

WHERE (A.AGE >= 70 and A.AGE <= 79);

select SUM(B.AMOUNT) AS AMT\_80\_89 from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

WHERE (A.AGE >= 80 and A.AGE <= 89);

select SUM(B.AMOUNT) AS AMT\_90\_99 from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

WHERE (A.AGE >= 90 and A.AGE <= 99);

select SUM(B.AMOUNT) AS AMT\_100 from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

WHERE (A.AGE >= 100);

*The age was divided into various buckets from Below 20, 20-29, 30-39 and so on till 90-99, 100+ to understand which age group donated the most.*

SELECT A.STATE,SUM(B.AMOUNT) from CONSTITUENT A join TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

GROUP BY A.STATE

ORDER BY sum(B.AMOUNT) DESC;

*The above query was written to understand which state donates how much amount based on the sum of constituents who belong to that state. After the sum was calculated the results were displayed based on the descending order.*

select A.CONSTITUENT\_ID, COUNT(A.VOLUNTEER\_ID) FROM VOLUNTEER A JOIN CONSTITUENT B

ON A.CONSTITUENT\_ID = B.CONSTITUENT\_ID

GROUP BY A.CONSTITUENT\_ID

ORDER BY COUNT(A.VOLUNTEER\_ID) DESC;

*This query gives the list of people who volunteered in any event and the number of events they have volunteered in*

**Query used to join different attribute and export into excel:**

SELECT A.CONSTITUENT\_ID , A.AGE , A.GENDER , A.MARRIED , '1' AS 'DONOR'

from CONSTITUENT A JOIN TRANSACTIONS B

ON B.CONSTITUENT\_ID = A.CONSTITUENT\_ID

GROUP BY B.CONSTITUENT\_ID

UNION

SELECT A.CONSTITUENT\_ID, A.AGE, A.GENDER , A.MARRIED, '0' AS 'DONOR'

FROM CONSTITUENT A

WHERE A.CONSTITUENT\_ID NOT IN (SELECT DISTINCT CONSTITUENT\_ID FROM TRANSACTIONS);

Alternative query:

SELECT \*,

CASE

WHEN CONSTITUENT\_ID IN (SELECT CONSTITUENT\_ID FROM TRANSACTIONS)

THEN 1 ELSE 0

END AS DONOR,

CASE

WHEN CONSTITUENT\_ID IN (SELECT CONSTITUENT\_ID FROM EVENT\_ATTENDEE)

THEN 1 ELSE 0

END AS EVENT\_ATTENDEE

FROM CONSTITUENT;

Challenge with this query was we were not able to fetch all 2000 rows; hence we wrote another query to get events attended as Yes or No and combined with the rest of the result in excel.

# R Studio

Dataset:



# TABLEAU

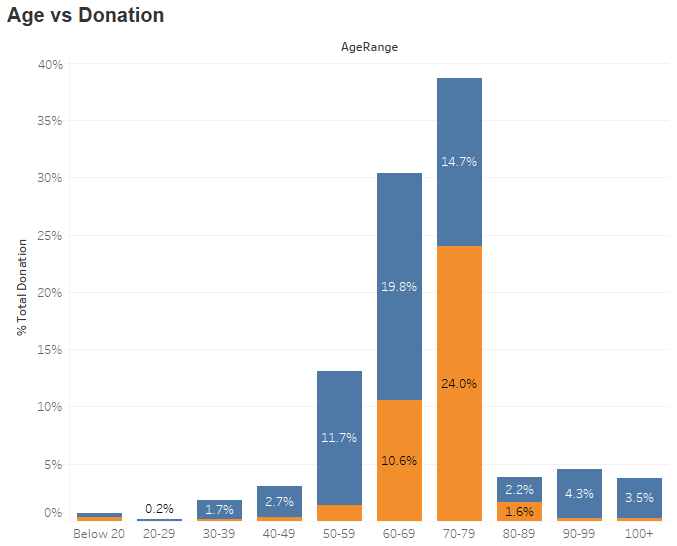
## URL:

<https://tableau.admin.uillinois.edu/#/workbooks/1578/views>

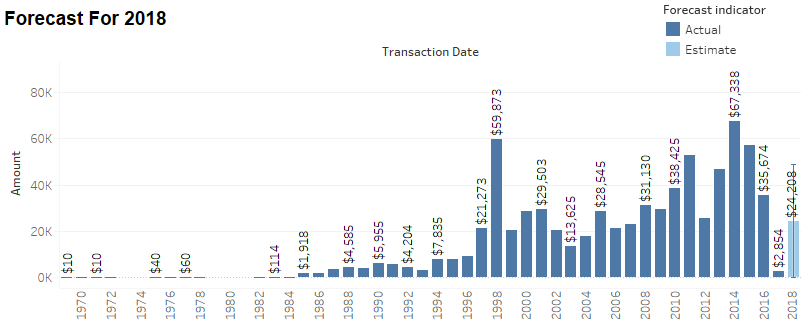
**Name of the Dashboard: Alankrita\_Dawkhar\_FinalProject**

## VISULAIZATIONS

Various visualizations were made in tableau based on some key factors available in the database. We did joins on few tables in Tableau so that we could combine results of different tables and derive more insights from the datasets.

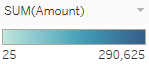


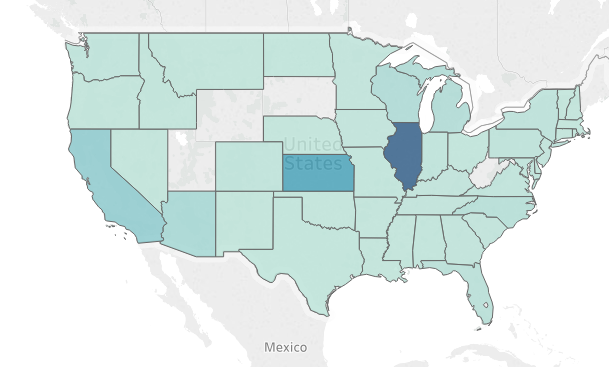
Age was divided into various buckets and the sum of donations made by different age groups was analyzed and converted in terms of percentage, such that deductions could be made about which age contributed the most and which gender in that age group contributed how much.



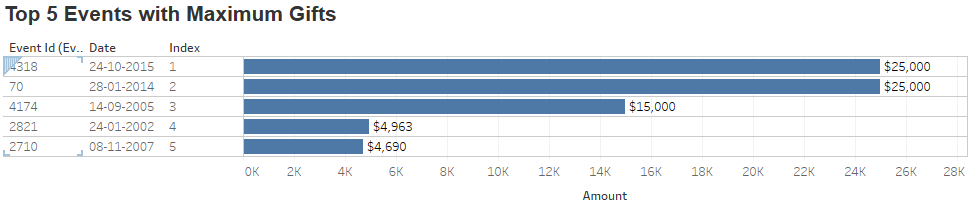
Based on the transaction date in the transaction table we could make a visualization on total transactions made every year. Using the data of all previous years, we were able to forecast the donations that will be made in 2018. As per our forecast year 2018 will have total donations of $24,208

**Geographical Heat Map based on donations made by every state in USA:**

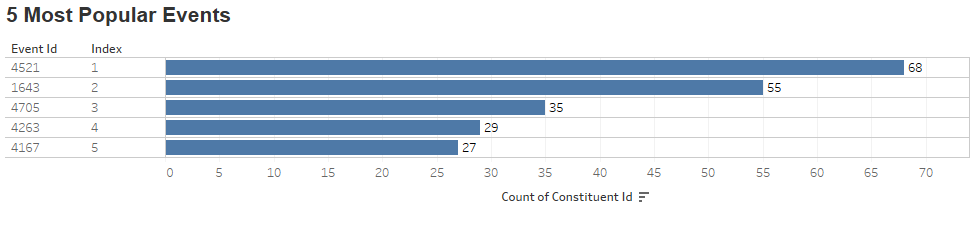




The above geographical representation is for contributions made by different states. The states represented by a dark color (for example Illinois, Kansas, California) are the ones that have the highest contributions.

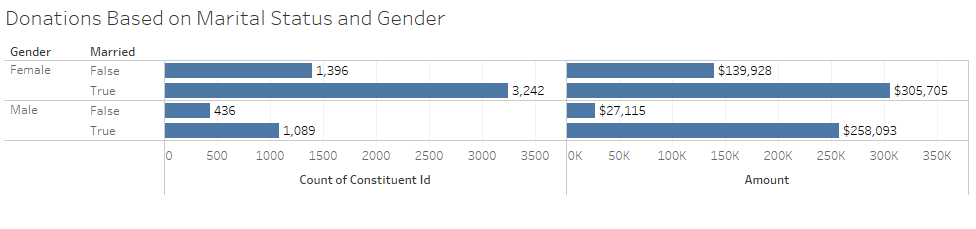


The above visualization represents the donations made by people in different events. A filter was put by us to display only the top 5 events with maximum gifts.



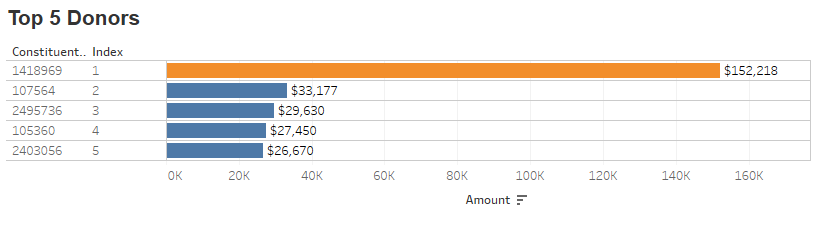
The above visualizations represent the number of people that attended that event. A filter was put by us to represent the top 5 events that had maximum event attendees.

From the above 2 visualizations we can deduce that maximum contributions were not made in the events that had maximum attendees.



The above representation is for number of constituents available in the dataset based on their marital status and the contributions made by each group of people (such has married female, unmarried female, etc)





The above graph is for listing the top 5 donors we have in the database. The total contribution per person was ranked and top 5 were filtered. As we can see, the maximum donation was made by a male constituent, but the other 4 contributors were female. As we refer to our SQL queries, it was also found out that total contributions made by females were more than males.

# REFERENCES

<http://www.theanalysisfactor.com/r-glm-model-fit/>

<https://stats.stackexchange.com/questions/108995/interpreting-residual-and-null-deviance-in-glm-r>

<http://www.statisticssolutions.com/theres-nothing-odd-about-the-odds-ratio-interpreting-binary-logistic-regression/>