Project II- Hive and Visualization

Neelam Kumari

Rajat Koujalagi

Data Sets Used

- Netflix Dataset
- Enron Dataset

Work Distribution

- Neelam Kumari
  - Enron Dataset Analysis and Documentation

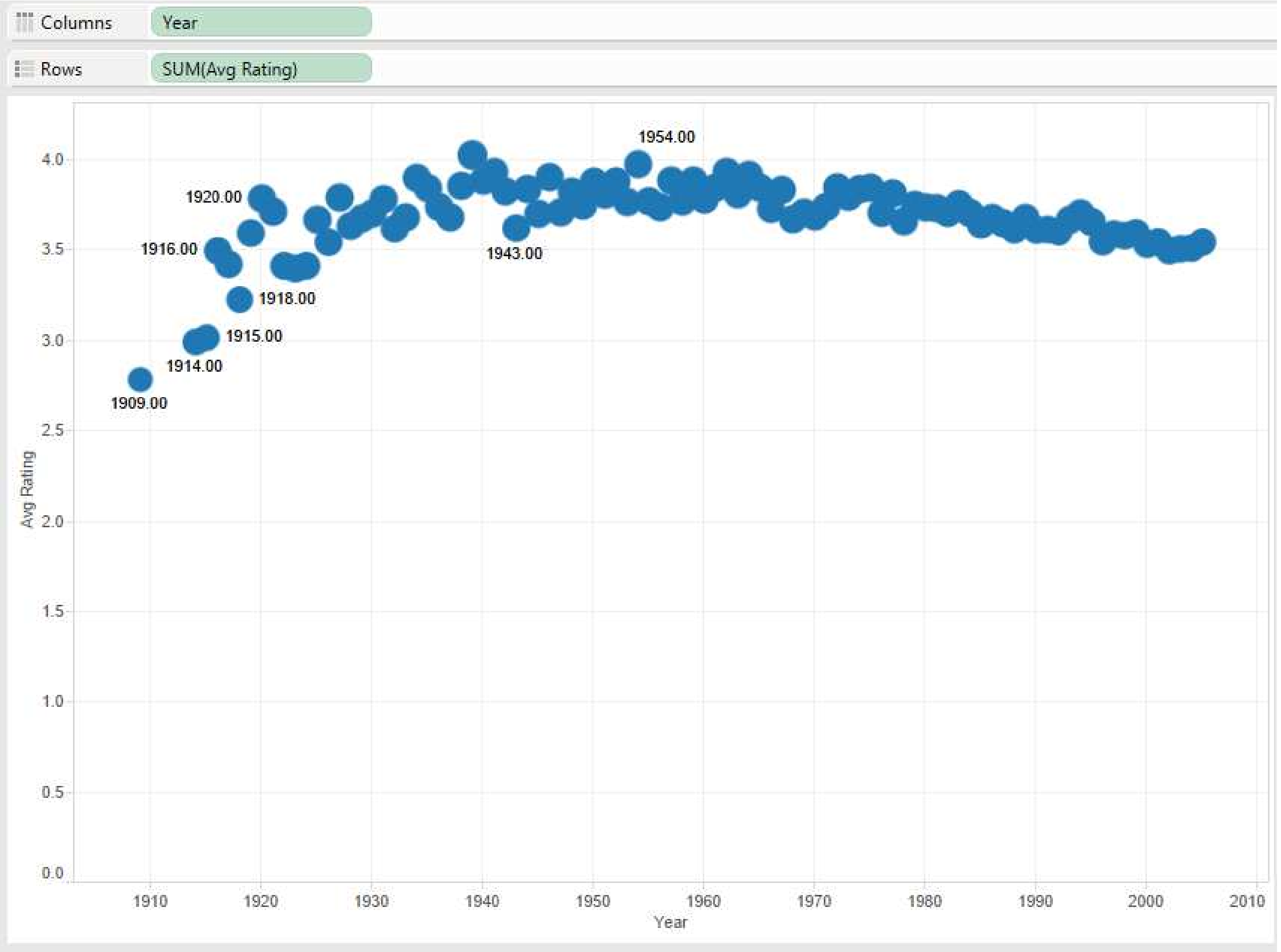
- Rajat Koujalagi
  - Netflix Dataset Analysis and Visualization

NETFLIX DATASET

Hypothesis 1

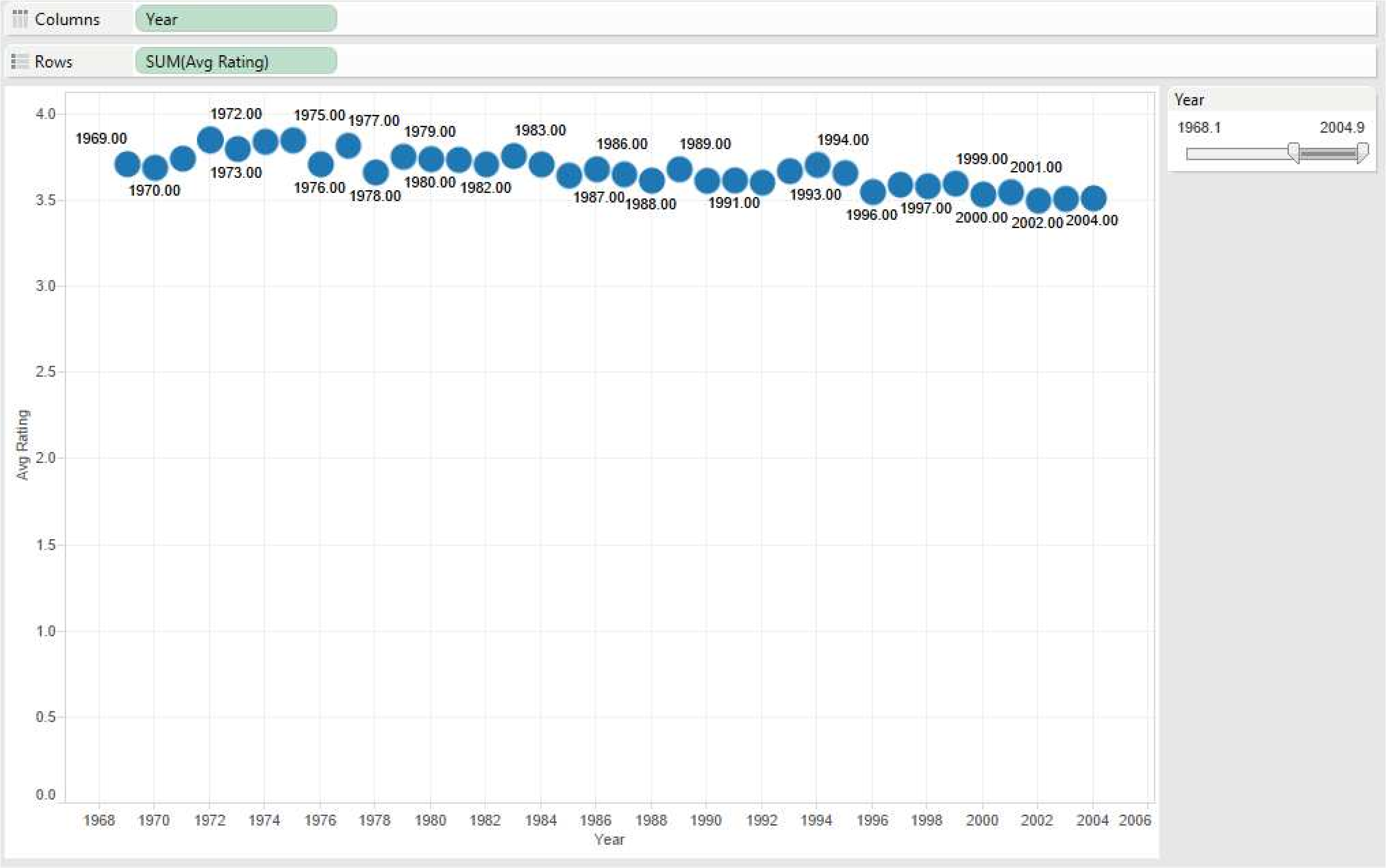
- Internet was not prevalent during early 20th century and not many movies were released. As the time passed by the higher number of movies were released and also internet became more prevalent amongst people which enabled them to rate the movies.

Visualization



Columns: Year

Rows: SUM(Avg Rating)



Columns Year

Rows SUM(Avg Rating)

Year 1968.1 2004.9

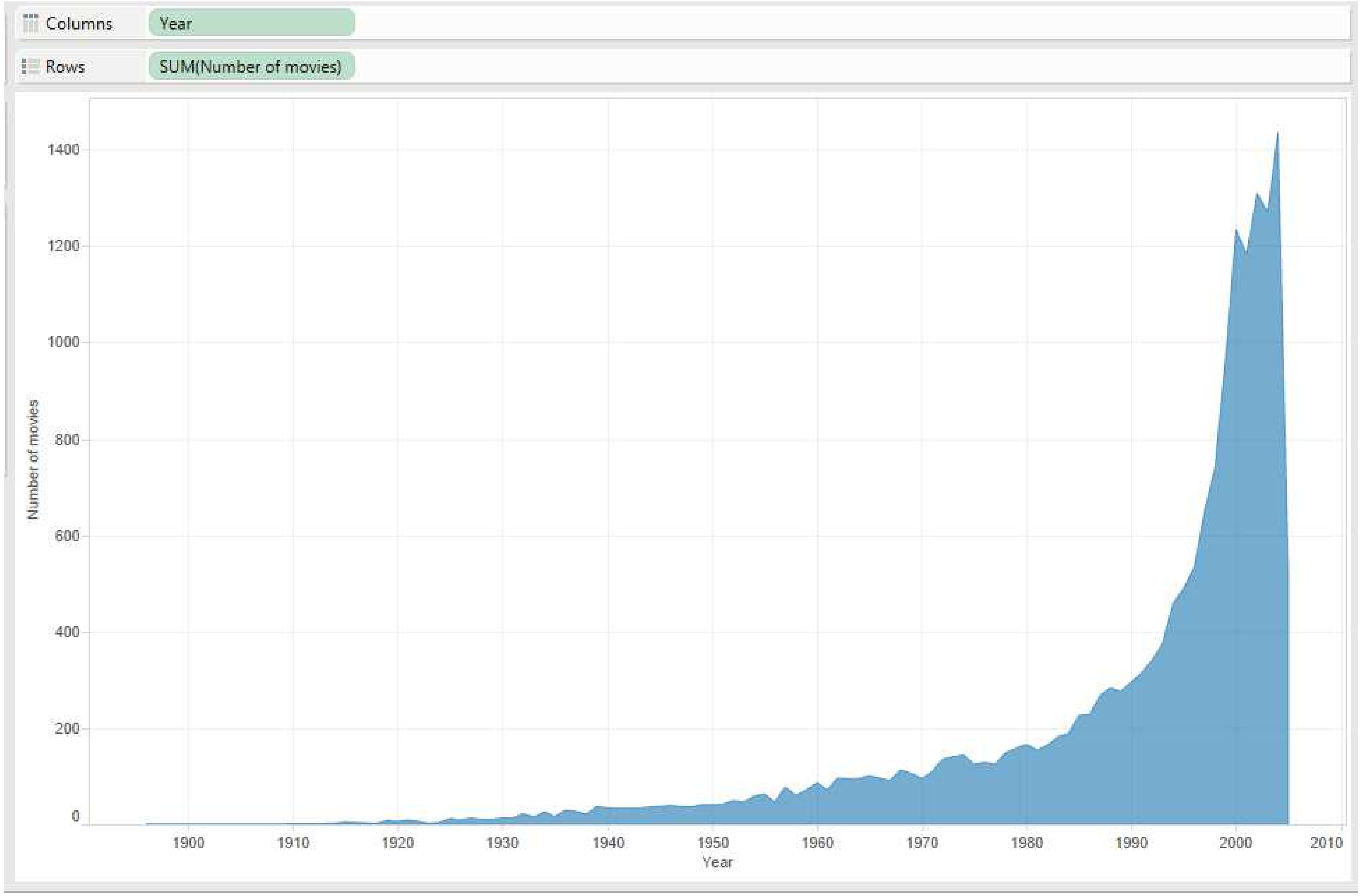
Results

- As shown in above figure, in late 20th century and early 21st century the average rating again goes down, which portrays that as more number of movies are made the quality went down.

Hypothesis 2

- With the passage of time from 20th century to present, the entertainment market has proved to be huge business market as more and more people watch movies as a break from their busy work life. Hence we see the drastic rise in number of movies released yearwise in Fig3, in the most recent times.

Visualization



Columns Year

Rows SUM(Number of movies)

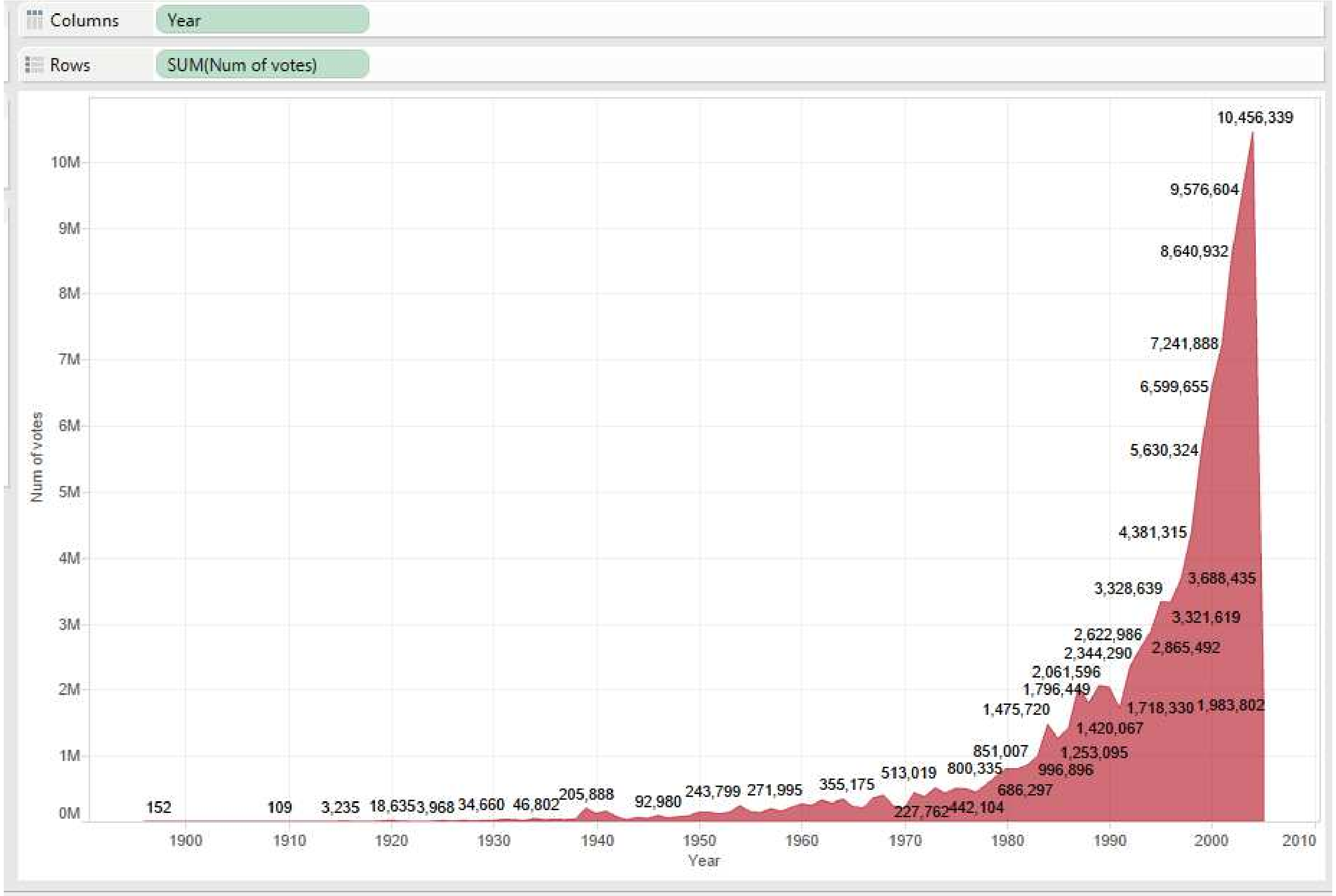
Results

- Hence we see the drastic rise in number of movies released year-wise in the figure above, in the most recent times.

Hypothesis 3

- Higher the number of internet users higher would be the number of votes for movies per year. Huge chunk of population rely on the average movie ratings in order to decide whether or not to watch the newly released movie

Visualization



Columns: Year

Rows: SUM(Num of votes)

Other Queries

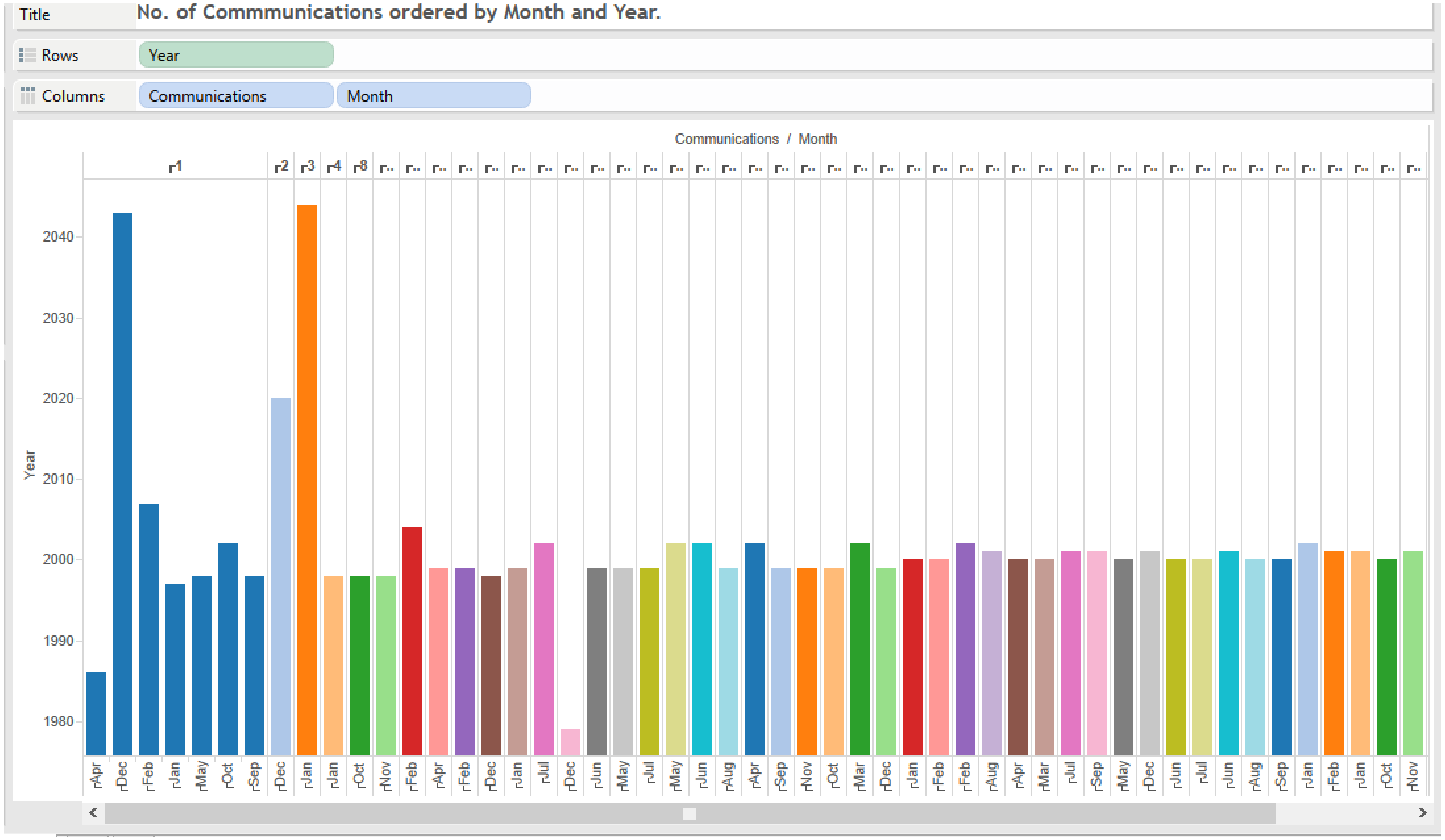
- **Other interesting Queries:**
- **TOP 10 movies between 2000 till 2010**
- SELECT result.title, result.rating FROM (SELECT title_ratings_join.title AS title, avg(title_ratings_join.rating) AS rating FROM (SELECT *
- FROM titles JOIN ratings ON (titles.mi = ratings.mid)   ) title_ratings_join
- WHERE title_ratings_join.yearOfRelease >=2000 AND title_ratings_join.yearOfRelease <2010
- GROUP BY title_ratings_join.title) result ORDER BY result.rating DESC LIMIT 10;
-
- **MOVIES released before 1970 but rated in 2000's**
- select distinct(title), yearOfRelease from titles join ratings on titles.mi=ratings.mid
- where titles.yearOfRelease<1970 and ratings.date like '2%';
-
- **Movies which have average rating above 3 with the 1,00,000 votes**
- SELECT titles.title from titles join ratings on (titles.mi=ratings.mid) where ratings.rating > 3 group by titles.title having count(ratings.customer_id) > 100000;
-
- **Top 10 movies throughout the whole period**
- SELECT result.title, result.rating FROM (SELECT title_ratings_join.title AS title,avg(title_ratings_join.rating) AS rating   FROM (SELECT * FROM titles JOIN ratings ON (titles.mi = ratings.mid)) title_ratings_join   GROUP BY title_ratings_join.title) result   ORDER BY result.rating DESC LIMIT 10;

ENRON DATASET

Hyposthesis 1

- The number of email communication approaches when scandal occurred

Visualization



Title: No. of Commmunications ordered by Month and Year.

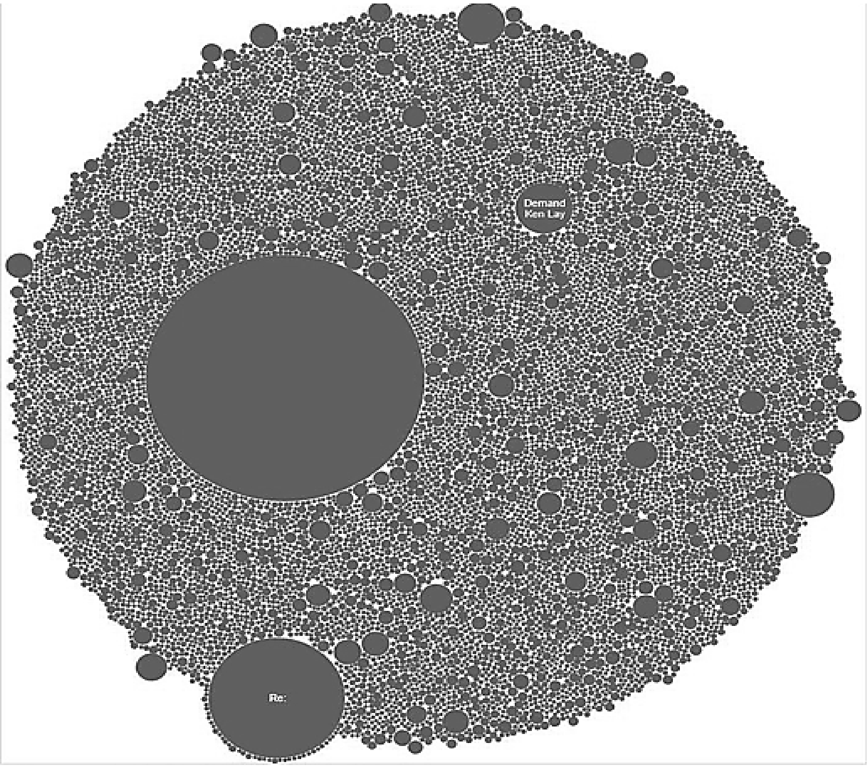
Rows: Year

Columns: Communications, Month

Hypothesis 2

- Popular subject in the conversations

Visualization



Hypothesis 3

- **Count number of threads for each enron employee**

LESSONS LEARNT

- Set operations such as INTERSECTION, UNION and EXCEPT are not present which makes querying a bit difficult.
- HIVE version installed on AWS did not support JOINS such as right join
- Sub queries could not be used with the FROM and WHERE clauses, hence making it mandatory to use EXISTS in order to perform any queries.