Employee Review of Comment Analysis
by Sujit Sonar

Objective: To use Hive features for data analysis and sharing the actionable insights into the HR team for taking corrective actions.

Q1) Impute the missing value (none) for all rating columns with a numerical value between 0 and 5. Note: For imputation, calculate the median for each of the 5 rating fields and create a new table.

1) *Login to FTP and load the csv files (employee_review_data.csv) provided for this project into the local file system*
2) *Login to hive and create the table "emp_review_table" and load the employee_review_data.csv into the newly created table.*

Creating table:

```
CREATE TABLE IF NOT EXISTS sujitsonarproject.emp_review_table (
Index int,
company string,
location string,
dates string,
job_title string,
summary string,
pros string,
cons string,
overall_ratings int,
work_balance_stars int,
culture_values_stars int,
carrer_opportunities_stars int,
comp_benefit_stars int,
senior_mangemnet_stars int)
COMMENT 'Employee Review Table'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
tblproperties('skip.header.line.count'='1');
```

Loading data:

```
LOAD DATA LOCAL INPATH '/mnt/home/sujitsonargmail/employee_review_data.csv' OVERWRITE INTO TABLE sujitsonarproject.emp_review_table;
```

Checking:

```
hive> show tables;
OK
emp_review_table
ime taken: 0.048 seconds, Fetched: 19 row(s)
ive>
```

```
hive> select * from emp_review_table limit 5;
OK
1       google  none    Dec 11; 2018    Current Employee - Anonymous Employee   Best Company to work for        People are smart and friendly   Bureaucracy is slowing t
hings down      5       4       5       5       4       5
2       google  Mountain View; CA       Jun 21; 2013    Former Employee - Program Manager       Moving at the speed of light; burn out is inevitable    "1) Food; food;
food. 15+ cafes on main campus (MTV) alone. Mini-kitchens; snacks; drinks; free breakfast/lunch/dinner; all day; errr'day.  2) Benefits/perks. Free 24:7 gym access (on
MTV campus). Free (self service) laundry (washer/dryer) available. Bowling alley. Volley ball pit. Custom-built and exclusive employee use only outdoor sport park (MTV)
. Free health/fitness assessments. Dog-friendly. Etc. etc. etc.  3) Compensation. In ~2010 or 2011; Google updated its compensation packages so that they were more comp
etitive.  4) For the size of the organization (30K+); it has remained relatively innovative; nimble; and fast-paced and open with communication but; that is definitely
changing (for the worse).  5) With so many departments; focus areas; and products; *in theory*; you should have plenty of opportunity to grow your career (horizontally
or vertically). In practice; not true.  6) You get to work with some of the brightest; most innovative and hard-working/diligent minds in the industry. There's a ""con"
" to that; too (see below)."    "1) Work/life balance. What balance? All those perks and benefits are an illusion. They keep you at work and they help you to be more pr
oductive. I've never met anybody at Google who actually time off on weekends or on vacations. You may not hear management say; ""You have to work on weekends/vacations"
" but; they set the culture by doing so - and it inevitably trickles down. I don't know if Google inadvertently hires the work-a-holics or if they create work-a-holics
in us. Regardless; I have seen way too many of the following: marriages fall apart; colleagues choosing work and projects over family; colleagues getting physically sic
k and ill because of stress; colleagues crying while at work because of the stress; colleagues shooting out emails at midnight; 1am; 2am; 3am. It is absolutely ridiculo
```

Observations: We see that the data is loaded successfully into the newly created hive table.

Data Cleaning & Pre-processing: **It is given that** Location: This dataset is global, as it includes the country's name in parenthesis [for example,"Toronto, ON(Canada)"]. However, if the location is the USA then it will only include the city and state [i.e. "Los Angeles, CA"].

Extracting the country name from the location col using hive sql.

1) we create a **us_state_code** table as helper table in hive.

```
CREATE TABLE IF NOT EXISTS sujitsonarproject.us_state_code (
State string,
Abbrev string,
Code string,
Country string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
tblproperties('skip.header.line.count'='1');
```

loading data:

```
LOAD DATA LOCAL INPATH '/mnt/home/sujitsonargmail/USA_state_code_csvData.csv' OVERWRITE INTO TABLE sujitsonarproject.us_state_code;
```

```
hive> select * from us_state_code limit 5;
OK
Alabama Ala.    AL      USA
Alaska  Alaska  AK      USA
Arizona Ariz.   AZ      USA
Arkansas        Ark.    AR      USA
California      Calif.  CA      USA
Time taken: 0.118 seconds, Fetched: 5 row(s)
hive>
```

2) **create a country _table in hive : applied subtring formatting and created new table for country.**

note : where country name = none, updated as global

```
create table country_table as
select t4.location, t4.loc_formated, t4.country,
case when t4.country is null then t4.loc_formated else t4.country end as country_new
from
(select t2.location, t2.loc_formated, t3.country from
(select distinct * from
(select location,
trim(regexp_replace(regexp_replace(regexp_extract(regexp_extract(location,'([^;]*)$',1),'([^()]*)$',1),
'\\(|\\)',''),'none([ ]*|$)','Global')) as loc_formated
from sujitsonarproject.emp_review_table) as t1) as t2
left join sujitsonarproject.us_state_code as t3
on t2.loc_formated = t3.code) as t4;
```

```
hive> select * from country_table  limit 5;
OK
Aberdeen; MD    MD      USA     USA
Aberdeen; SD    SD      USA     USA
Aberdeen; Scotland (UK) UK      NULL    UK
Abha (Saudi Arabia)     Saudi Arabia    NULL    Saudi Arabia
Abidjan (Cote d'Ivoire) Cote d'Ivoire   NULL    Cote d'Ivoire
Time taken: 0.081 seconds, Fetched: 5 row(s)
hive>
```

Observations: We see that

1) **some of the data are “none” in the location column**
2) **there are some dates as “none” or “0000”**
3) **there are null values for ratings**

3) **Data cleaning: lookup the correct**
   a) country name from country_table,
   b) delete the rows where the year == 0 or None into emp_review_table
   c) replace null to 0 in all the rating columns
   d) extract year from the dates col
   e) extract the emp status from the job_title col

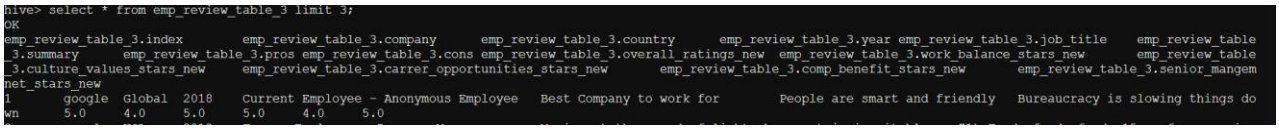
```
create table emp_review_table_1 as
select * from
(select t1.index, t1.company, t2.country_new as country, t1.dates, substr(trim(t1.dates),-4) as
year, t1.job_title,
trim(regexp_extract(t1.job_title,'^([^-]*)',1)) as emp_status,t1.summary,
t1.pros, t1.cons, t1.overall_ratings,
coalesce(t1.work_balance_stars,  0) as work_balance_stars,
coalesce(t1.culture_values_stars,  0) as culture_values_stars,
coalesce(t1.carrer_opportunities_stars, 0) as carrer_opportunities_stars,
coalesce(t1.comp_benefit_stars, 0) as comp_benefit_stars,
coalesce(t1.senior_mangemnet_stars, 0) as senior_mangemnet_stars
from sujitsonarproject.emp_review_table as t1
left join sujitsonarproject.country_table t2
on t1.location = t2.location) as t3
where t3.index not in (7820,28730,36105,40863) ;
```

```
hive> select * from emp_review_table_1 limit 5;
OK
emp_review_table_1.index        emp_review_table_1.company      emp_review_table_1.country      emp_review_table_1.dates        emp_review_table_1.year emp_review_table
_1.job_title    emp_review_table_1.emp_status   emp_review_table_1.summary      emp_review_table_1.pros emp_review_table_1.cons emp_review_table_1.overall_ratings     e
mp_review_table_1.work_balance_stars    emp_review_table_1.culture_values_stars emp_review_table_1.carrer_opportunities_stars   emp_review_table_1.comp_benefit_stars  e
mp_review_table_1.senior_mangemnet_stars
1       google  Global   Dec 11; 2018    2018    Current Employee - Anonymous Employee   Current Employee        Best Company to work for        People are smart and fri
endly   Bureaucracy is slowing things down      5       4       5       5       4       5
2       google  USA      Jun 21; 2013    2013    Former Employee - Program Manager       Former Employee Moving at the speed of light; burn out is inevitable    "1) Food
; food; food. 15+ cafes on main campus (MTV) alone. Mini-kitchens; snacks; drinks; free breakfast/lunch/dinner; all day; errr'day.  2) Benefits/perks. Free 24:7 gym acc
ess (on MTV campus). Free (self service) laundry (washer/dryer) available. Bowling alley. Volley ball pit. Custom-built and exclusive employee use only outdoor sport pa
rk (MTV). Free health/fitness assessments. Dog-friendly. Etc. etc. etc.  3) Compensation. In ~2010 or 2011; Google updated its compensation packages so that they were m
```

mp_revi	emp_revi	emp_revi	emp_revi	emp_review_table_1.year	emp_review_table_1.job_title	emp_review_table_1.emp_statu	emp_revi	emp_revi	emp_revi	emp_revi	emp_revi	emp_revi	emp_revi	emp_revi	emp_review_table_1.senior
1	google	Global	Dec 11; 2	2018	Current Employee - Anonymous Employe	Current Employee	Best Com	People ar	Bureaucra	5	4	5	5	4	5
2	google	USA	Jun 21; 2	2013	Former Employee - Program Manager	Former Employee	Moving at	"1) Food;	"1) Work/	4	2	3	3	5	3
3	google	USA	May 10;	2014	Current Employee - Software Engineer III	Current Employee	Great bal	"* If you'r	* It *is* b	5	5	4	5	5	4
4	google	USA	Feb 8; 20	2015	Current Employee - Anonymous Employe	Current Employee	The best p	You can't	I live in SF	5	2	5	5	4	5
5	google	USA	Jul 19; 2(	2018	Former Employee - Software Engineer	Former Employee	Unique; o	Google is	"If you do	5	5	5	5	5	5

d) Create a new table by updating the 0 values with the median of ratings

```
create table emp_review_table_3 as
select t3.index,t3.company, t3.country, t3.year, t3.job_title,t3.summary,t3.pros,t3.cons,
t3.overall_ratings_new,
t3.work_balance_stars_new,
t3.culture_values_stars_new,
t3.carrer_opportunities_stars_new,
t3.comp_benefit_stars_new,
t3.senior_mangemnet_stars_new
from
(select t1.index,t1.company,t1.country,t1.year,t1.job_title,
t1.summary,t1.pros,t1.cons,t1.overall_ratings,
case when t1.overall_ratings = 0 then t2.overall_ratings_median else t1.overall_ratings end as
overall_ratings_new,
case when t1.work_balance_stars = 0 then t2.work_balance_stars_median else
t1.work_balance_stars end as work_balance_stars_new,
case when t1.culture_values_stars = 0 then t2.culture_values_stars_median else
t1.culture_values_stars end as culture_values_stars_new,
case when t1.carrer_opportunities_stars = 0 then t2.carrer_opportunities_stars_median else
t1.carrer_opportunities_stars end as carrer_opportunities_stars_new,
case when t1.comp_benefit_stars = 0 then t2.comp_benefit_stars_median else
t1.comp_benefit_stars end as comp_benefit_stars_new,
case when t1.senior_mangemnet_stars = 0 then t2.senior_mangemnet_stars_median else
t1.senior_mangemnet_stars end as senior_mangemnet_stars_new
from sujitsonarproject.emp_review_table_1 as t1,
(select company,country,year,job_title,
percentile(overall_ratings, 0.5) as overall_ratings_median,
percentile(work_balance_stars, 0.5) as work_balance_stars_median,
percentile(culture_values_stars, 0.5) as culture_values_stars_median,
percentile(carrer_opportunities_stars, 0.5) as carrer_opportunities_stars_median,
percentile(comp_benefit_stars, 0.5) as comp_benefit_stars_median,
percentile(senior_mangemnet_stars, 0.5) as senior_mangemnet_stars_median
from sujitsonarproject.emp_review_table_1
group by company,country,year, job_title) as t2
where t1.company = t2.company
and t1.country = t2.country
and t1.year = t2.year
and t1.job_title = t2.job_title) as t3;
```



Q2) Create a hive table partitioned by country and bucketed by year and also load the review.csv file.

Note: Ensure that the right hive environment variable is set for bucket insert.

```
create table emp_review_part_buck_table (
index int,
company string,
year string,
job_title string,
summary string,
pros string,
cons string,
overall_ratings_new double,
work_balance_stars_new double,
culture_values_stars_new double,
carrer_opportunities_stars_new double,
comp_benefit_stars_new double,
senior_mangemnet_stars_new double
)
partitioned by (country string)
clustered by (year) into 5 buckets
row format delimited
fields terminated by ',';

Setting the parameters:
SET hive.exec.dynamic.partition = true;
SET hive.exec.enforce.bucketing = true;
SET hive.exec.dynamic.partition.mode = nonstrict;

Loading data into partitioned tables and buckets:

insert into emp_review_part_buck_table
partition(country)
select
index,
company,
year,
job_title,
summary,
pros,
cons,
overall_ratings_new,
work_balance_stars_new,
culture_values_stars_new,
carrer_opportunities_stars_new,
comp_benefit_stars_new,
senior_mangemnet_stars_new,
country
from emp_review_table_3;
```

```
[sujitsonargmail@ip-10-0-42-218 ~]$ hadoop fs -ls /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table
Found 106 items
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Afghanistan
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Algeria
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Argentina
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Australia
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Austria
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Azerbaijan
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Bahrain
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Bangladesh
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Belgium
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Bhutan
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Bosnia and Herzegovina
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Brazil
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Bulgaria
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Burkina Faso
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Cameroon
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Canada
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Cayman Islands
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Chile
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=China
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Colombia
```

```
[sujitsonargmail@ip-10-0-42-218 ~]$ hadoop fs -ls /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Canada
Found 5 items
-rw-r--r--   3 sujitsonargmail hive      39819 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Canada/000000_0
-rw-r--r--   3 sujitsonargmail hive      59415 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Canada/000001_0
-rw-r--r--   3 sujitsonargmail hive      71870 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Canada/000002_0
-rw-r--r--   3 sujitsonargmail hive      99734 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Canada/000003_0
-rw-r--r--   3 sujitsonargmail hive      37807 2022-03-22 12:06 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck_table/country=Canada/000004_0
[sujitsonargmail@ip-10-0-42-218 ~]$
```

Observations: we see that we have created a portioned table by country and bucketed the files by year

Q3) Write the final relation schema to review.csv file in your HDFS home directory.

I have my final file after all the data pre-processing in this location in hive warehouse

```
[sujitsonargmail@ip-10-0-42-218 ~]$ hadoop fs -ls /user/hive/warehouse/sujitsonarproject.db
Found 15 items
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-20 15:33 /user/hive/warehouse/sujitsonarproject.db/country_table
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-22 05:17 /user/hive/warehouse/sujitsonarproject.db/emp_review_part_buck
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-20 08:21 /user/hive/warehouse/sujitsonarproject.db/emp_review_table
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-21 10:10 /user/hive/warehouse/sujitsonarproject.db/emp_review_table_1
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-21 10:25 /user/hive/warehouse/sujitsonarproject.db/emp_review_table_2
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-18 16:28 /user/hive/warehouse/sujitsonarproject.db/industry_growth
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-18 03:29 /user/hive/warehouse/sujitsonarproject.db/nyse_stock_ssonar
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-18 04:21 /user/hive/warehouse/sujitsonarproject.db/nyse_stock_ssonar_project
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-19 06:02 /user/hive/warehouse/sujitsonarproject.db/sector_growth_table
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-18 17:11 /user/hive/warehouse/sujitsonarproject.db/state_industry_growth
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-17 14:13 /user/hive/warehouse/sujitsonarproject.db/stock_companies
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-18 07:48 /user/hive/warehouse/sujitsonarproject.db/stock_data1
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-18 11:37 /user/hive/warehouse/sujitsonarproject.db/stock_growth_table
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-17 14:11 /user/hive/warehouse/sujitsonarproject.db/stock_prices
drwxr-xr-x   - sujitsonargmail hive          0 2022-03-20 13:23 /user/hive/warehouse/sujitsonarproject.db/us_state_code
[sujitsonargmail@ip-10-0-42-218 ~]$
```

INSERT OVERWRITE DIRECTORY '/user/sujitsonargmail/project/review.csv' ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
*SELECT * FROM sujitsonarproject.emp_review_part_buck_table;*

checking the data in hdfs:

```
[sujitsonargmail@ip-10-0-42-218 ~]$ hadoop fs -ls /user/sujitsonargmail/project
Found 1 items
-rw-r--r--   3 sujitsonargmail hadoop   16477494 2022-03-22 07:08 /user/sujitsonargmail/project/review.csv
[sujitsonargmail@ip-10-0-42-218 ~]$
```

We see the file is copied from hive to hdfs with the file name as review.csv
Checking the first 5 records:

```
[sujitsonargmail@ip-10-0-42-218 ~]$ hadoop fs -cat /user/sujitsonargmail/project/review.csv |head -n 5
emp_review_part_buck.index,emp_review_part_buck.company,emp_review_part_buck.job_title,emp_review_part_buck.summary,emp_review_part_buck.pros,emp_review_part_buck.cons,
emp_review_part_buck.overall_ratings,emp_review_part_buck.work_balance_stars_new,emp_review_part_buck.culture_values_stars_new,emp_review_part_buck.carrer_opportunities
_stars_new,emp_review_part_buck.comp_benefit_stars_new,emp_review_part_buck.senior_mangemnet_stars_new,emp_review_part_buck.year,emp_review_part_buck.country
48968,apple,2012,Current Employee - Waiter/Steward,i like apple very much its a good app,its a big co & famous for its product good,NULL,5.0,3.0,3.0,3.0,3.0,2.0,Afghani
stan
6014,google,2017,Current Employee - Anonymous Employee,good,payment on time ;with a lot of facilitate,NULL,4.0,4.0,4.0,3.0,4.0,4.0,Algeria
5110,google,2018,Current Employee - Product Marketing Manager,Tindouf Algérie,Looking forward to a better future by working with your esteemed company; which is flouris
hing and progressing; which made me offer employment and provide for the financial assistance provided by you,NULL,4.0,5.0,5.0,5.0,5.0,5.0,Algeria
64983,microsoft,2015,Former Employee - Account Manager,Microsoft From Inside,Impacting People and Business Innovation Legal Maturity Sales Process Training Cloud Salary
 Commission Established Brand People celebrated Enterprise Agreement Licensing Creativity,NULL,3.0,1.0,5.0,2.0,4.0,1.0,Algeria
cat: Unable to write to output stream.
[sujitsonargmail@ip-10-0-42-218 ~]$
```

Q4) Using the over-all rating fields display trend:

1.Globally by company: Identify trends at 25%, 50%, 75%

-- we see that there are total of 67525 rows of data after data pre-processing
-- 25% of 67525 = 16881
-- 50 % of 67525 = 33763
-- 75% of 65525 = 50644
Using these values, we will limit the rows to analyse the data at 25%, 50% and 75%

-- Identify trends at 25% by company:

```
select t1.company, round(avg(t1.overall_ratings_new),1) as avg_overall_ratings from
(select company, overall_ratings_new
from sujitsonarproject.emp_review_part_buck_table
distribute by rand()
sort by rand()
limit 16881) as t1
group by t1.company;
```

Query History | Saved Queries | Results (6)

	t1.company	avg_overall_ratings
1	amazon	3.6
2	apple	4
3	facebook	4.5
4	google	4.3
5	microsoft	3.8
6	netflix	3.4

-- Identify trends at 50% by company:

```
select t1.company, round(avg(t1.overall_ratings_new),1) as avg_overall_ratings from
(select company, overall_ratings_new
from sujitsonarproject.emp_review_part_buck_table
distribute by rand()
sort by rand()
limit 33763) as t1
group by t1.company;
```

	t1.company	avg_overall_ratings
1	amazon	3.6
2	apple	4
3	facebook	4.5
4	google	4.3
5	microsoft	3.8
6	netflix	3.5

-- Identify trends at 75% by company:

```
select t1.company, round(avg(t1.overall_ratings_new),1) as avg_overall_ratings from
(select company, overall_ratings_new
from sujitsonarproject.emp_review_part_buck_table
distribute by rand()
sort by rand()
limit 50644) as t1
group by t1.company;
```

	t1.company	avg_overall_ratings
1	amazon	3.6
2	apple	4
3	facebook	4.5
4	google	4.3
5	microsoft	3.8
6	netflix	3.4

At 100% data:

t1.company	avg_overall_ratings
amazon	3.6
apple	4
facebook	4.5
google	4.3
microsoft	3.8
netflix	3.4

Observations: the average overall ratings by company (Globally) when checked the random dataset at 25%, 50%, 75% or even at 100%, is same for each of the companies.

	company	1	2	3	4	5
1	amazon	2461	2751	5574	8090	7552
2	apple	484	814	2266	4580	4805
3	facebook	47	49	98	245	1151
4	google	155	248	763	2275	4378
5	microsoft	639	1308	3819	7101	5063
6	netflix	120	112	133	204	240

Also, we see that very less employee was reviewed for Netflix, followed by facebook.
Assuming 4 and above to be very good rating, and less than 4 to be not so good
This shows that at Global level, Facebook has the highest average overall ratings, followed by google and apple. The other three companies (Microsoft, Amazon and Netflix) are somewhere in the middle range of the ratings score.

Checking the % of employees by overall ratings:

```
select p2.company,round(p2.r_0/p1.total,2) as r_0,round(p2.r_1/p1.total,2) as r_1,
round(p2.r_2/p1.total,2) as r_2,
round(p2.r_3/p1.total,2) as r_3,
round(p2.r_4/p1.total,2) as r_4,
round(p2.r_5/p1.total,2) as r_5
from
(select company, count(*) total from sujitsonarproject.emp_review_part_buck_table group by
company)p1,
(select company,
COLLECT_SET(`0`)[0] AS `r_0`,
COLLECT_SET(`1`)[0] AS `r_1`,
COLLECT_SET(`2`)[0] AS `r_2`,
COLLECT_SET(`3`)[0] AS `r_3`,
COLLECT_SET(`4`)[0] AS `r_4`,
COLLECT_SET(`5`)[0] AS `r_5`
FROM (
SELECT t1.company,
CASE WHEN overall_ratings_new=0 THEN count(overall_ratings_new) END AS `0`,
CASE WHEN overall_ratings_new=1 THEN count(overall_ratings_new) END AS `1`,
CASE WHEN overall_ratings_new=2 THEN count(overall_ratings_new) END AS `2`,
CASE WHEN overall_ratings_new=3 THEN count(overall_ratings_new) END AS `3`,
CASE WHEN overall_ratings_new=4 THEN count(overall_ratings_new) END AS `4`,
CASE WHEN overall_ratings_new=5 THEN count(overall_ratings_new) END AS `5`
FROM (select company, overall_ratings_new
from sujitsonarproject.emp_review_part_buck_table
distribute by rand()
sort by rand())t1
group by company,overall_ratings_new)t2
group by company,0,1,2,3,4,5)p2
where p2.company = p1.company;
```

	p2.company	r_0	r_1	r_2	r_3	r_4	r_5
1	amazon	NULL	0.09	0.1	0.21	0.31	0.29
2	apple	NULL	0.04	0.06	0.17	0.35	0.37
3	facebook	NULL	0.03	0.03	0.06	0.15	0.72
4	google	NULL	0.02	0.03	0.1	0.29	0.56
5	microsoft	NULL	0.04	0.07	0.21	0.4	0.28
6	netflix	NULL	0.15	0.14	0.16	0.25	0.3

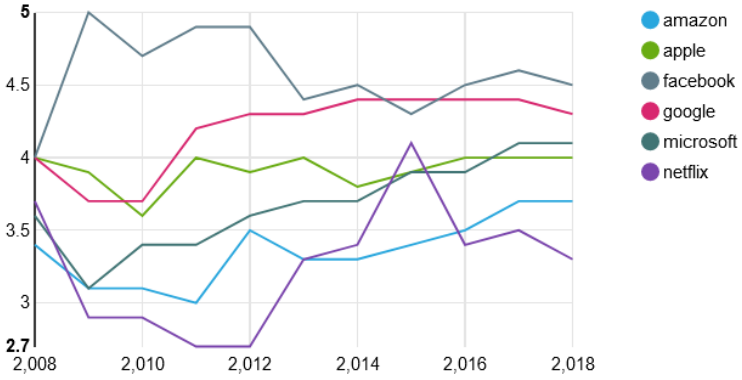
Observations: Also, checking the trend by the rating count, we see that 72% people in FB have rated it 5 and overall 87% (72%+15%) people are very happy working in FB. This is followed by Gogle with 85% (56% + 29%)
NetFlix has got a 55% (30% + 25%) which means people have neutral say about the company
60% (31% + 29%) people are happy with Amazon
72% (35%+37%) people are happy with Apple
68% (40% + 28%) people are happy with Microsoft.

2.Globally by company: Identify trends at 25%, 50%, 75%

-- Identify trends at 25% by company per year:

```
select t1.company,t1.year, round(avg(t1.overall_ratings_new),1) as avg_overall_ratings from
(select company,year, overall_ratings_new
from sujitsonarproject.emp_review_part_buck_table
distribute by rand()
sort by rand()
limit 16881) as t1
group by t1.company,t1.year;
```

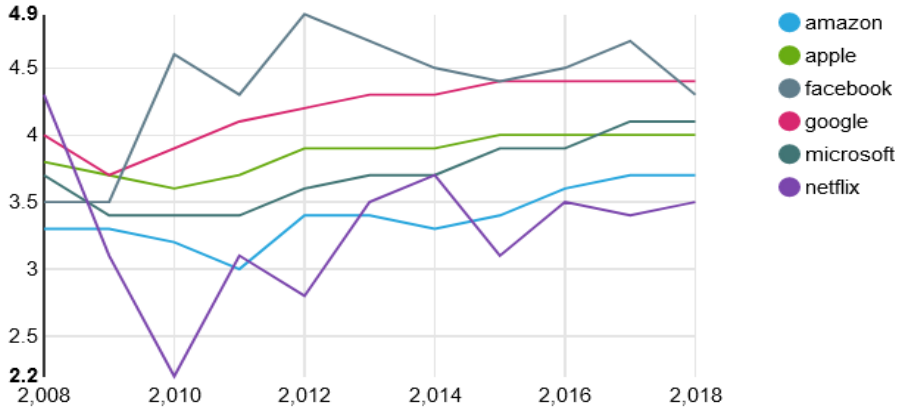
eries Results (66)



-- Identify trends at 50% by company per year:

```
select t1.company,t1.year, round(avg(t1.overall_ratings_new),1) as avg_overall_ratings from
(select company,year, overall_ratings_new
from sujitsonarproject.emp_review_part_buck_table
distribute by rand()
sort by rand()
limit 33763) as t1
group by t1.company,t1.year;
```

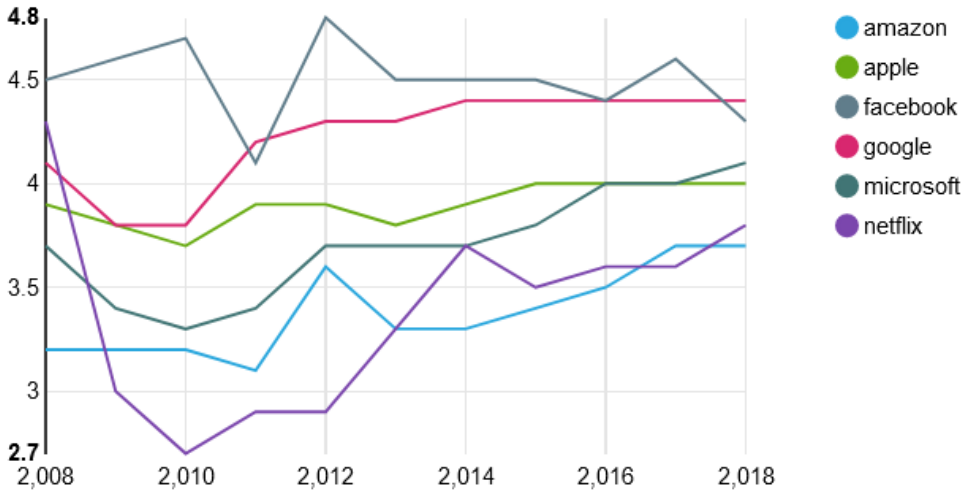
ueries Results (66)



-- Identify trends at 70% by company per year:

```
select t1.company,t1.year, round(avg(t1.overall_ratings_new),1) as avg_overall_ratings from
(select company,year, overall_ratings_new
from sujitsonarproject.emp_review_part_buck_table
distribute by rand()
sort by rand()
limit 50644) as t1
group by t1.company,t1.year;
```

eries Results (66)



Observations: using the chart options from HUE, the average overall ratings by company (Globally) by year when checked the random dataset at 25%, 50%, 75% or even at 100%, is showing the same trend for each of the companies.

This shows that at Global level, Facebook has the highest average overall ratings, followed by google and apple. The other three companies (Microsoft, Amazon and Netflix) are somewhere in the middle range of the ratings score.

3. By company by country (Identify trends for each company by country: Identify trends at 25%, 50%, 75%

-- Identify trends at 25% by company by country:

```
select t1.company,t1.country, round(avg(t1.overall_ratings_new),1) as avg_overall_ratings from
(select company,country, overall_ratings_new
from sujitsonarproject.emp_review_part_buck_table
distribute by rand()
sort by rand()
limit 16881) as t1
group by t1.company,t1.country;
```

this query gives us the avg_overall_ratings by company and country

	t1.company	t1.country	avg_overall_ratings
1	amazon	Australia	4.6
2	amazon	Bahrain	4
3	amazon	Bangladesh	4
4	amazon	Bhutan	5
5	amazon	Brazil	5

Expanding on this query we can create a pivot table

```
select country,
COLLECT_SET(`amazon`)[0] AS `amazon`,
COLLECT_SET(`apple`)[0] AS `apple`,
COLLECT_SET(`facebook`)[0] AS `facebook`,
COLLECT_SET(`google`)[0] AS `google`,
COLLECT_SET(`microsoft`)[0] AS `microsoft`,
COLLECT_SET(`netflix`)[0] AS `netflix`
from (
select country,
case when company ='amazon' then overall_ratings_new end as `amazon`,
case when company ='apple' then overall_ratings_new end as `apple`,
case when company ='facebook' then overall_ratings_new end as `facebook`,
case when company ='google' then overall_ratings_new end as `google`,
case when company ='microsoft' then overall_ratings_new end as `microsoft`,
case when company ='netflix' then overall_ratings_new end as `netflix`
from

(select company, country,round(avg(overall_ratings_new),2) as overall_ratings_new  from
sujitsonarproject.emp_review_part_buck_table
group by company, country
distribute by rand()
sort by rand()
limit 16881)t1)t2
group by country;
```

country	amazon	apple	facebook	google	microsoft	netflix
Afghanistan	NULL	5	NULL	NULL	NULL	NULL
Algeria	NULL	NULL	NULL	4	3	NULL
Argentina	4	NULL	4.5	4.38	3.83	NULL
Australia	4.3	3.57	4.67	4.65	3.34	NULL
Austria	NULL	NULL	NULL	NULL	3.67	NULL
Azerbaijan	NULL	NULL	NULL	NULL	3	NULL
Bahrain	4.5	5	NULL	NULL	NULL	NULL
Bangladesh	4.29	4	4.5	4.33	4	NULL
Belgium	NULL	4	NULL	4	4.13	NULL
Bhutan	5	NULL	NULL	NULL	NULL	NULL

So, checking the ratings by company and filtering all the null values and sort by avg ratings to find the top 10 and bottom 10

```
select country,amazon
from
(select country,
COLLECT_SET(`amazon`)[0] AS `amazon`,
COLLECT_SET(`apple`)[0] AS `apple`,
COLLECT_SET(`facebook`)[0] AS `facebook`,
COLLECT_SET(`google`)[0] AS `google`,
COLLECT_SET(`microsoft`)[0] AS `microsoft`,
COLLECT_SET(`netflix`)[0] AS `netflix`
from (select country,
case when company ='amazon' then overall_ratings_new end as `amazon`,
case when company ='apple' then overall_ratings_new end as `apple`,
case when company ='facebook' then overall_ratings_new end as `facebook`,
case when company ='google' then overall_ratings_new end as `google`,
case when company ='microsoft' then overall_ratings_new end as `microsoft`,
case when company ='netflix' then overall_ratings_new end as `netflix`
from
(select company, country,round(avg(overall_ratings_new),2) as overall_ratings_new  from
sujitsonarproject.emp_review_part_buck_table
group by company, country
distribute by rand()
sort by rand())t1)t2
group by country)t3
where amazon is not null
sort by amazon desc
limit 10;
```

Amazon: company and country trend: at 25%

	country	amazon
1	Tanzania	5
2	Bhutan	5
3	Venezuela	5
4	Dominican Republic	5
5	Democratic Republic of Congo	5
6	Jamaica	5
7	Denmark	5
8	Portugal	5
9	Sweden	5
10	Israel	4.73

	country	amazon
1	Cayman Islands	1
2	Croatia	1
3	Turkey	1.5
4	Yemen	2
5	Finland	2
6	Czech Republic	2.99
7	Uganda	3
8	Cote d'Ivoire	3
9	Morocco	3
10	Colombia	3

Amazon: company and country trend: at 50%

Top 10

	country	amazon
1	Tanzania	5
2	Bhutan	5
3	Venezuela	5
4	Dominican Republic	5
5	Democratic Republic of Congo	5
6	Jamaica	5
7	Denmark	5
8	Portugal	5
9	Sweden	5
10	Israel	4.73

Bottom 10

	country	amazon
1	Cayman Islands	1
2	Croatia	1
3	Turkey	1.5
4	Yemen	2
5	Finland	2
6	Czech Republic	2.99
7	Uganda	3
8	Cote d'Ivoire	3
9	Morocco	3
10	Colombia	3

Amazon: company and country trend: at 75%

	country	amazon
1	Cayman Islands	1
2	Croatia	1
3	Turkey	1.5
4	Yemen	2
5	Finland	2
6	Czech Republic	2.99
7	Uganda	3
8	Cote d'Ivoire	3
9	Morocco	3
10	Colombia	3

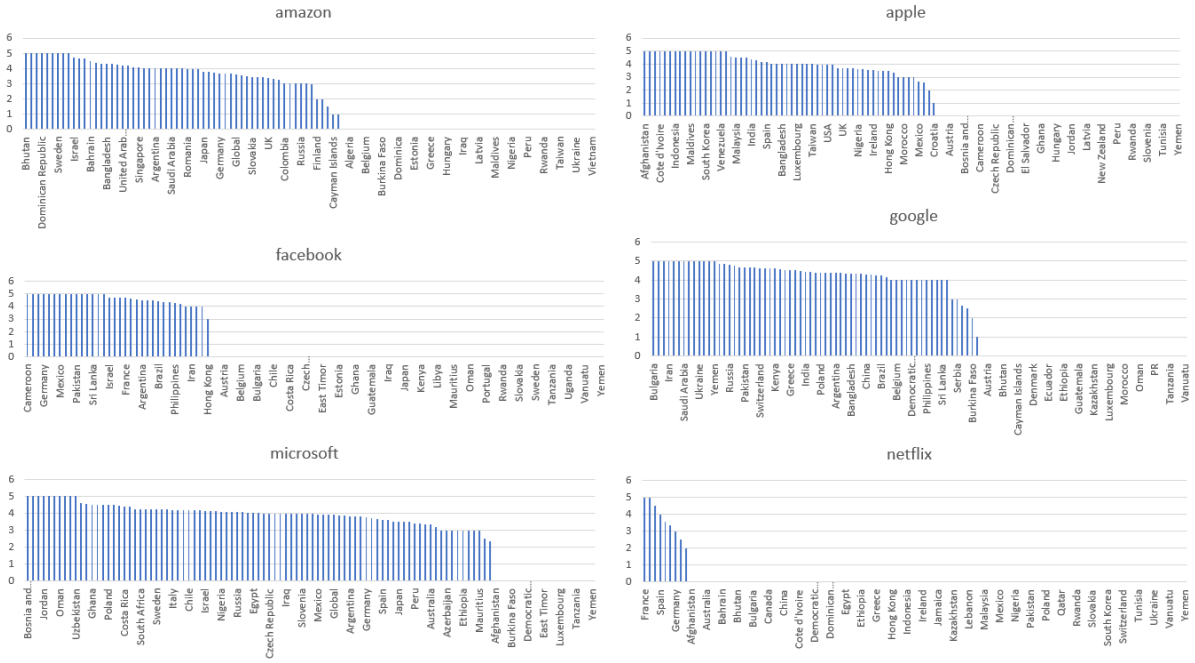
	country	amazon
1	Tanzania	5
2	Bhutan	5
3	Venezuela	5
4	Dominican Republic	5
5	Democratic Republic of Congo	5
6	Jamaica	5
7	Denmark	5
8	Portugal	5
9	Sweden	5
10	Israel	4.73

Observations: By looking at the top 10 and bottom 10 countries by company, we see that the result and the trend is same at 25%, 50% and 75% data sets.
Therefore, we can check the trend using the whole 100% of the data for each of the companies.

With 100% of data:

```
select q.company, q.count_country, p. unique_countries_count, round(q.count_country/p.
unique_countries_count, 2) as country_percent  from
(select count(distinct country) unique_countries_count
from sujitsonarproject.emp_review_part_buck_table) p,
(select company,count(distinct country) as count_country
from sujitsonarproject.emp_review_part_buck_table
group by company) q;
```

	q.company	q.count_country	p.unique_countries_count	country_percent
1	amazon	59	106	0.56
2	apple	58	106	0.55
3	facebook	34	106	0.32
4	google	65	106	0.61
5	microsoft	87	106	0.82
6	netflix	9	106	0.08



Observations: out of 106 unique countries in the data set, 82% of the countries has the review data for Microsoft. Google has review data for 61% of the countries. Little over 40% of the countries does not have the review data for amazon and apple. We see around 32% of countries reviewed for Facebook and very few countries reviewed for Netflix. This could be either

1) Survey was not conducted in these countries or
2) There are no employees from these countries

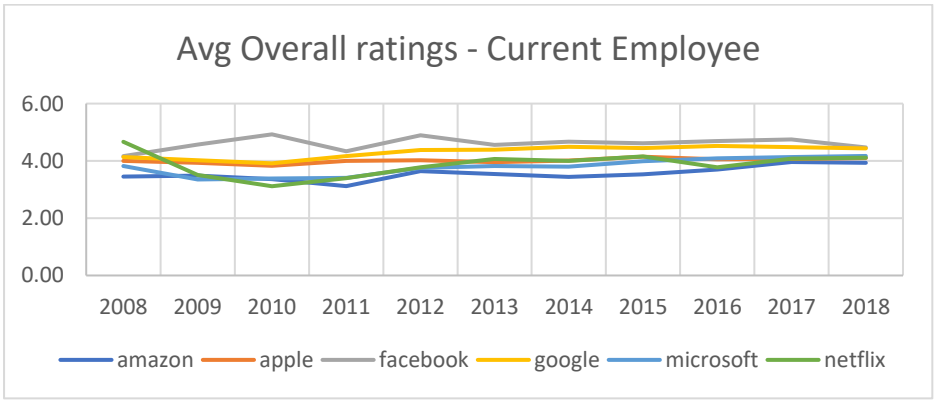
   The reason for very low overall ratings for Netflix could be because we have very less employee review data compared to other companies and whoever was reviewed were not very happy.

Q5) Display the impact of employee status on rating a company using the overall-ratings field by the company by year.

Current Employee:

```
select year,
COLLECT_SET(`amazon`)[0] AS `amazon`,
COLLECT_SET(`apple`)[0] AS `apple`,
COLLECT_SET(`facebook`)[0] AS `facebook`,
COLLECT_SET(`google`)[0] AS `google`,
COLLECT_SET(`microsoft`)[0] AS `microsoft`,
COLLECT_SET(`netflix`)[0] AS `netflix` from
(select year,
case when company ='amazon' then avg_overall_ratings_new end as `amazon`,
case when company ='apple' then avg_overall_ratings_new end as `apple`,
case when company ='facebook' then avg_overall_ratings_new end as `facebook`,
case when company ='google' then avg_overall_ratings_new end as `google`,
case when company ='microsoft' then avg_overall_ratings_new end as `microsoft`,
case when company ='netflix' then avg_overall_ratings_new end as `netflix`
from (select t1.company,t1.year,t1.emp_status, round(avg(t1.overall_ratings_new),2) as
avg_overall_ratings_new
from (select company,year,trim(regexp_extract(job_title,'^([^-]*)',1)) as emp_status,
overall_ratings_new
from sujitsonarproject.emp_review_part_buck_table)t1
group by t1.company,t1.year,t1.emp_status)t2
where t2.emp_status ='Current Employee')t3
group by year;
```

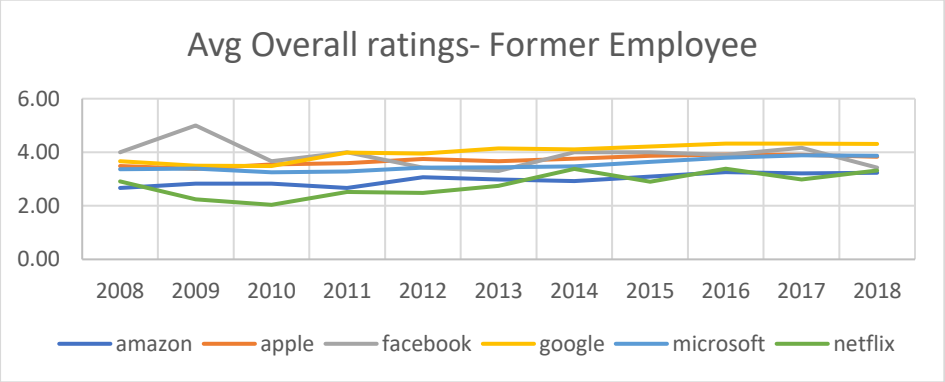
	year	amazon	apple	facebook	google	microsoft	netflix
1	2008	3.46	4	4.17	4.13	3.82	4.67
2	2009	3.49	3.93	4.57	4.02	3.35	3.51
3	2010	3.37	3.83	4.93	3.91	3.39	3.12
4	2011	3.12	4	4.33	4.17	3.4	3.4
5	2012	3.64	4.02	4.89	4.38	3.76	3.77
6	2013	3.54	3.96	4.56	4.39	3.82	4.07
7	2014	3.44	4.01	4.67	4.49	3.79	4
8	2015	3.53	4.15	4.61	4.44	3.99	4.16
9	2016	3.7	4.06	4.69	4.52	4.09	3.78
10	2017	3.96	4.07	4.74	4.48	4.13	4.08
11	2018	3.94	4.1	4.47	4.44	4.17	4.11



Former Employee:

```
select year,
COLLECT_SET(`amazon`)[0] AS `amazon`,
COLLECT_SET(`apple`)[0] AS `apple`,
COLLECT_SET(`facebook`)[0] AS `facebook`,
COLLECT_SET(`google`)[0] AS `google`,
COLLECT_SET(`microsoft`)[0] AS `microsoft`,
COLLECT_SET(`netflix`)[0] AS `netflix` from
(select year,
case when company ='amazon' then avg_overall_ratings_new end as `amazon`,
case when company ='apple' then avg_overall_ratings_new end as `apple`,
case when company ='facebook' then avg_overall_ratings_new end as `facebook`,
case when company ='google' then avg_overall_ratings_new end as `google`,
case when company ='microsoft' then avg_overall_ratings_new end as `microsoft`,
case when company ='netflix' then avg_overall_ratings_new end as `netflix`
from (select t1.company,t1.year,t1.emp_status, round(avg(t1.overall_ratings_new),2) as
avg_overall_ratings_new
from (select company,year,trim(regexp_extract(job_title,'^([^-]*)',1)) as emp_status,
overall_ratings_new
from sujitsonarproject.emp_review_part_buck_table)t1
group by t1.company,t1.year,t1.emp_status)t2
where t2.emp_status ='Former Employee')t3
group by year;
```

	year	amazon	apple	facebook	google	microsoft	netflix
1	2008	2.67	3.49	4	3.66	3.36	2.91
2	2009	2.83	3.38	5	3.5	3.39	2.24
3	2010	2.82	3.55	3.67	3.49	3.25	2.04
4	2011	2.67	3.59	4	3.99	3.28	2.52
5	2012	3.06	3.74	3.43	3.96	3.42	2.49
6	2013	2.99	3.67	3.3	4.14	3.44	2.75
7	2014	2.92	3.76	4	4.11	3.47	3.38
8	2015	3.09	3.87	4	4.22	3.65	2.91
9	2016	3.26	3.91	3.9	4.32	3.8	3.38
10	2017	3.21	3.91	4.16	4.32	3.89	2.98
11	2018	3.23	3.83	3.43	4.31	3.87	3.32



Observations: by looking at the avg overall ratings by company by year for Current employee vs the Former Employee, we see that, the former employee avg overall ratings for Amazon, Facebook and Netflix is around 1 point lower than the current employee. This appears to us that the former employee of Amazon, Facebook and Netflix gave lower overall ratings.
There is not much difference for Apple, google and Microsoft.

Q6) Display the impact of job role on rating a company using the overall-ratings field by the company by year.

As we look at the job roles for all the 6 companies, we see there are around 6931 unique job titles across the years and is very data heavy to analyse.

```
select count(*) from
(select distinct trim(regexp_extract(job_title,'-(.*)',1)) as job_role
from sujitsonarproject.emp_review_part_buck_table)t1;
```

Query History | Saved Queries | Results (1)

	_c0
1	6931

Therefore, I decided to find the word count of each of the job roles and after quick look at the data, I filter out key words that are occurring more than 80 times.

```
select * from
(SELECT word, count(1) AS count_1 FROM
(SELECT explode(split(job_role, '\\s')) AS word
FROM
(select distinct trim(regexp_extract(job_title,'-(.*)',1)) as job_role
from sujitsonarproject.emp_review_part_buck_table)t1) w
where  not (w.word rlike '[^a-zA-Z\\d\\s:]')
and length(w.word)>5
GROUP BY w.word
sort BY count_1 desc)t2
where t2.count_1 >80;
```

	t2.word	t2.count_1
1	Manager	1028
2	Senior	728
3	Engineer	611
4	Associate	473
5	Specialist	420
6	Software	309
7	Support	279
8	Operations	254
9	Development	231
10	Analyst	230
11	Advisor	205
12	Technical	204
13	Business	180
14	Service	177
15	Customer	175
16	Program	158
17	Product	157
18	Account	131
19	Marketing	129
20	Director	121
21	Representative	108
22	Quality	106
23	Engineering	101
24	Developer	99
25	Consultant	99
26	Management	98
27	Assistant	96
28	Executive	92
29	Technician	91
30	Principal	91
31	Intern	89
32	Partner	88
33	Warehouse	83

Observations: By looking at the results, I am using the key words to categories the job roles.
If the job role contains any of these key words, I will create a category by find the combination of these 33 key words and the rest I will categories them as others as shown below:

Categories of job role:

Accountant/HR
Architect/Consultant/Administrator
Advisor/Analyst/Associate
Lead/Manager/Assistant
Chief/Top Leaders
Customer Service/Representative
Developer/Programmer/Designer
Executive/Principal role/Director
Scientist/Engineer/Technical
Intern
Sales/Operations/Product related
roles/Marketing/Management
Other Roles
Other Senior level
Partner
Recruiter/Quality
Senior Director
Senior HR
Senior Manager
Vice President
Specialist/Tech Support/Technician

Creating a pivot table and summarising the data using these assumed categories of roles:

```
Select new_role,company,
COLLECT_SET(`2008`)[0] AS `2008`,
COLLECT_SET(`2009`)[0] AS `2009`,
COLLECT_SET(`2010`)[0] AS `2010`,
COLLECT_SET(`2011`)[0] AS `2011`,
COLLECT_SET(`2012`)[0] AS `2012`,
COLLECT_SET(`2013`)[0] AS `2013`,
COLLECT_SET(`2014`)[0] AS `2014`,
COLLECT_SET(`2015`)[0] AS `2015`,
COLLECT_SET(`2016`)[0] AS `2016`,
COLLECT_SET(`2017`)[0] AS `2017`,
COLLECT_SET(`2018`)[0] AS `2018` from (select new_role,company,
case when year =2008 then avg_overall_ratings_new end as `2008`,
case when year =2009 then avg_overall_ratings_new end as `2009`,
case when year =2010 then avg_overall_ratings_new end as `2010`,
case when year =2011 then avg_overall_ratings_new end as `2011`,
case when year =2012 then avg_overall_ratings_new end as `2012`,
case when year =2013 then avg_overall_ratings_new end as `2013`,
case when year =2014 then avg_overall_ratings_new end as `2014`,
case when year =2015 then avg_overall_ratings_new end as `2015`,
case when year =2016 then avg_overall_ratings_new end as `2016`,
case when year =2017 then avg_overall_ratings_new end as `2017`,
case when year =2018 then avg_overall_ratings_new end as `2018`
from (select company,year,
case
   when t2.job_role rlike ' Vice President' and t2.job_role rlike 'Manager' then ' Vice President'
   when t2.job_role rlike 'Senior' and t2.job_role rlike 'Director' then 'Senior Director'
   when t2.job_role rlike 'Senior' and t2.job_role rlike 'Manager'  then 'Senior Manager'
   when t2.job_role rlike 'Senior' and t2.job_role rlike 'HR'  then 'Senior HR'
   when t2.job_role rlike 'Senior' then 'Other Senior Roles'
   when t2.job_role rlike 'Chief' then 'Chief|Top Leaders'
   when t2.job_role rlike 'Intern' then 'Intern'
   when t2.job_role rlike 'Partner' then 'Partner'
   when t2.job_role rlike 'Human|Accountant|Recruiter' then 'Accountant|HR|Recruiter'
   when t2.job_role rlike 'Advisor|Analyst|Associate' then 'Advisor|Analyst|Associate'
   when t2.job_role rlike 'Architect|Consultant|Administrator' then 'Architect|Consultant|Administrator'
   when t2.job_role rlike 'Customer|Service|Representative' then 'Customer Service|Representative'
   when t2.job_role rlike 'Developer|Programmer|Designer' then 'Developer|Programmer|Designer'
   when t2.job_role rlike 'Executive|Principal|Director' then 'Executive|Principal role|Director'
   when t2.job_role rlike 'Lead|Manager|Assistant' then 'Lead|Manager|Assistant'
   when t2.job_role rlike 'Sales|Operations|Quality|Product|Marketing|Management' then
   'Sales|Operations|Quality|Product|Marketing|Management'
   when t2.job_role rlike 'Scientist|Engineer|Technical' then 'Scientist|Engineer|Technical'
   when t2.job_role rlike 'Specialist|Tech Support|Technician' then 'Specialist|Tech Support|Technician'
   else 'Other Roles' end as new_role,
   avg(avg_overall_ratings_new) as avg_overall_ratings_new

from (
select t1.company,t1.year,t1.job_role, round(avg(t1.overall_ratings_new),2) as avg_overall_ratings_new
from (select company,year,trim(regexp_extract(job_title,'-(.*)',1)) as job_role, overall_ratings_new
from sujitsonarproject.emp_review_part_buck_table)t1
group by t1.company,t1.year,t1.job_role)t2
group by company,year,job_role)t3
where company='amazon'
)t4
group by new_role,company;
```

Checking for amazon:

	new_role	company	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
1	Accountant|HR|Recruiter	amazon	2	3	3	2	2	4.2	3.33	5	2	4	3
2	Advisor|Analyst|Associate	amazon	4	4	1	3	4	2.83	3	4	2	5	2
3	Architect|Consultant|Administrator	amazon	3	4	NULL	NULL	NULL	3.5	5	3	4	2	5
4	Chief|Top Leaders	amazon	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	5	4
5	Customer Service|Representative	amazon	3	NULL	NULL	1	3	1	4	4	5	4	2
6	Developer|Programmer|Designer	amazon	2	2	2	NULL	5	3	3	5	1	1	3
7	Executive|Principal role|Director	amazon	2.75	4	2	4	NULL	3	3.67	3.8	3.5	3.75	3.75
8	Intern	amazon	NULL	3	3.5	3.33	4.4	3.5	4	5	4	4.33	4
9	Lead|Manager|Assistant	amazon	3	3	2.6	2	2.88	2.33	1	4.17	4	5	5
10	Other Roles	amazon	4	3.19	3.31	2	4	3	2	5	2	3	3
11	Other Senior Roles	amazon	4.33	3.5	2	4	3	5	3	4	4	5	5
12	Partner	amazon	NULL	NULL	4	NULL	NULL	1	3	5	3.5	4	3
13	Sales|Operations|Quality|Product|Marketing|Management	amazon	3	5	3	4	5	4	5	2	3	5	5
14	Scientist|Engineer|Technical	amazon	4	4	3.33	3	3	4	4.5	4	4.5	5	4.4
15	Senior Director	amazon	NULL	NULL	4	NULL	NULL	NULL	NULL	NULL	3	NULL	NULL
16	Senior HR	amazon	NULL	NULL	NULL	NULL	NULL	NULL	4.5	3	1	4	4.5
17	Senior Manager	amazon	4	3	1	2.5	4	3.4	4	1	3.5	1	4.75
18	Specialist|Tech Support|Technician	amazon	NULL	4	3	3	5	5	1.5	2.25	4	2	5

Architect/Consultant/Administrator
Advisor/Analyst/Associate
Customer Service/Representative
Intern
Lead/Manager/Assistant
Other Senior Roles,
Sales/Operations/Product related roles/Marketing/Management
Scientist/Engineer/Technical
Senior Manager
Specialist/Tech Support/Technician

Checking for apple:

	new_role	company	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
1	Accountant|HR|Recruiter	apple	3	NULL	NULL	2	1	5	4	1	3	5	5
2	Advisor|Analyst|Associate	apple	4	4	1	3	4	3	2.5	4	3	1	5
3	Architect|Consultant|Administrator	apple	5	5	3	2	5	4.29	4	3	5	3	4
4	Chief|Top Leaders	apple	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	5	5	NULL
5	Customer Service|Representative	apple	3	5	NULL	5	4	5	4.33	2	5	3	4
6	Developer|Programmer|Designer	apple	4	NULL	4	2	2	3.5	5	4	4	2	4
7	Executive|Principal role|Director	apple	3.5	2.67	3.5	3	3	4	4	2.67	3	4	2
8	Intern	apple	4.33	4	5	5	4.5	4.5	NULL	4	4	4	5
9	Lead|Manager|Assistant	apple	4	4	4	2	5	5	4	5	1	4	2.5
10	Other Roles	apple	3	3.65	3.85	5	5	5	2	5	1	5	2
11	Other Senior Roles	apple	4	5	3	4.5	5	4	4.6	4	4	4	5
12	Partner	apple	2	4	NULL	4	NULL	NULL	1	5	NULL	1	NULL
13	Sales|Operations|Quality|Product|Marketing|Management	apple	5	5	3	4	1	5	4	4	5	4	4
14	Scientist|Engineer|Technical	apple	3	2.5	3.33	4	5	3.5	3	4	4	5	5
15	Senior Director	apple	4	NULL	NULL	NULL	NULL	3	NULL	NULL	5	NULL	2.33
16	Senior Manager	apple	4	5	2	4	2	2	5	3.82	5	4	3
17	Specialist|Tech Support|Technician	apple	4	3	4	5	5	3	5	3	5	4	2

Accountant/HR
Architect/Consultant/Administrator
Customer Service/Representative
Intern
Lead/Manager/Assistant
Other Senior level
Sales/Operations/Product related roles/Marketing/Management
Specialist/Tech Support/Technician

Checking for facebook:

	new_role	company	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
1	Accountant|HR|Recruiter	facebook	NULL	NULL	5	NULL	5	5	4	5	5	5	1
2	Advisor|Analyst|Associate	facebook	4	NULL	5	NULL	4.5	5	5	5	1	4.33	4.75
3	Architect|Consultant|Administrator	facebook	NULL	NULL	NULL	NULL	5	NULL	NULL	NULL	5	NULL	3
4	Customer Service|Representative	facebook	NULL	NULL	NULL	NULL	4.5	5	NULL	2	NULL	5	3
5	Developer|Programmer|Designer	facebook	NULL	NULL	NULL	NULL	5	NULL	2	4	5	5	5
6	Executive|Principal role|Director	facebook	NULL	NULL	3	NULL	5	NULL	5	5	5	5	5
7	Intern	facebook	NULL	NULL	5	3	5	5	4	3	4	5	4
8	Lead|Manager|Assistant	facebook	4	5	NULL	5	4.5	3	4	4	4.33	4.67	5
9	Other Roles	facebook	3	5	4.64	4.69	4	4.17	4.67	5	5	4.53	3
10	Other Senior Roles	facebook	NULL	NULL	5	5	5	5	4	4.5	NULL	4	5
11	Partner	facebook	NULL	NULL	NULL	NULL	5	5	4	5	4.5	4.75	5
12	Sales|Operations|Quality|Product|Marketing|Management	facebook	5	5	4.5	NULL	3	3	5	5	5	5	5
13	Scientist|Engineer|Technical	facebook	4	5	5	2.5	5	5	5	5	5	5	4.67
14	Senior Manager	facebook	NULL	NULL	4	NULL	NULL	NULL	5	NULL	NULL	4	3
15	Specialist|Tech Support|Technician	facebook	NULL	NULL	NULL	4	NULL	2	4	3	5	5	3

Accountant/HR
Advisor/Analyst/Associate
Executive/Principal role/Director
Intern
Advisor/Analyst/Associate
Lead/Manager/Assistant
Other Roles
Other Senior level
Partner
Scientist/Engineer/Technical

Check for google:

	new_role	company	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
1	Accountant|HR|Recruiter	google	3	5	4	5	3	3	1	4	4.5	5	5
2	Advisor|Analyst|Associate	google	4	5	4	4	5	4	3	5	4.83	2	5
3	Architect|Consultant|Administrator	google	NULL	NULL	5	NULL	5	3	4	4	5	3	5
4	Chief|Top Leaders	google	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	5	5	2
5	Customer Service|Representative	google	5	NULL	2	2	4.2	5	4	5	4	5	4
6	Developer|Programmer|Designer	google	5	4	3	5	5	5	4	4	4	5	5
7	Executive|Principal role|Director	google	5	5	4	5	4	3	4.36	4.67	4	5	4
8	Intern	google	5	4	5	5	4	4	4	5	5	5	4.25
9	Lead|Manager|Assistant	google	4	3.67	4	4.5	4.44	4.67	4.07	4.47	4.27	4.11	3.78
10	Other Roles	google	3	4	5	4.5	5	4.44	5	4.38	4	5	4.67
11	Other Senior Roles	google	4	5	3	4	4	5	4.33	5	3	5	2
12	Partner	google	1	5	NULL	NULL	1	NULL	5	4	5	5	3.2
13	Sales|Operations|Quality|Product|Marketing|Manage	google	4	5	4	2	4	5	5	4	5	5	5
14	Scientist|Engineer|Technical	google	4	4	4	4.67	5	4	3.5	4	3	5	4
15	Senior Director	google	5	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	5
16	Senior Manager	google	5	5	NULL	4	5	4	4.17	4.25	4.33	4.5	4.25
17	Specialist|Tech Support|Technician	google	4	5	3	2	4	5	5	4	5	5	4

Advisor/Analyst/Associate
Customer Service/Representative
Developer/Programmer/Designer
Executive/Principal role/Director
Intern
Lead/Manager/Assistant
Other Roles
Other Senior level
Specialist/Tech Support/Technician
Sales/Operations/Product related roles/Marketing/Management
Senior Manager

Checking for microsoft:

	new_role	company	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
1	Accountant|HR|Recruiter	microsoft	3	4	4	3	4	1	4	3	4	4	5
2	Advisor|Analyst|Associate	microsoft	5	5	5	3	3	4.5	5	5	4.44	3.75	4.29
3	Architect|Consultant|Administrator	microsoft	4.5	4	5	5	4.25	3.2	4.4	5	5	4.25	5
4	Chief|Top Leaders	microsoft	NULL	NULL	NULL	NULL	5	NULL	NULL	5	3	5	5
5	Customer Service|Representative	microsoft	NULL	NULL	3	3	4	4	3	4	5	4	5
6	Developer|Programmer|Designer	microsoft	4	3	4	2.5	3	3	5	4	3	5	4.5
7	Executive|Principal role|Director	microsoft	3.88	3.5	4	3.2	3.67	3.22	3.29	3.83	3.29	5	3.91
8	Intern	microsoft	4	4	4	4	4.8	5	4	4	4.5	5	4
9	Lead|Manager|Assistant	microsoft	3	4	4	3.63	5	3.65	3.09	3.56	3.33	2.8	4
10	Other Roles	microsoft	3.69	3.43	3.33	2	5	2.5	1	5	3	3	4
11	Other Senior Roles	microsoft	2	4	5	3	3.67	3	4	3.5	5	4	5
12	Partner	microsoft	5	4	2	4	4	3	2.67	5	5	4	4
13	Sales|Operations|Quality|Product|Marketing|Management	microsoft	4	3	2	4	4	5	4	3	2	5	3
14	Scientist|Engineer|Technical	microsoft	3	4	3	4	4	4	3	5	4	4	4
15	Senior Director	microsoft	3.5	NULL	2.4	2.83	3.3	2.71	3.44	3.42	4.25	3.2	4
16	Senior Manager	microsoft	4.2	2	4	4	4	3	5	3	3	4	5
17	Specialist|Tech Support|Technician	microsoft	3.4	3	4	3	3.38	4	4.5	3.5	4	5	3

Architect/Consultant/Administrator
Advisor/Analyst/Associate
Customer Service/Representative
Intern
Partner
Senior Director

Checking for netflix:

	new_role	company	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
1	Accountant|HR|Recruiter	netflix	4	4	NULL	NULL	NULL	NULL	NULL	NULL	NULL	3	5
2	Advisor|Analyst|Associate	netflix	NULL	NULL	NULL	NULL	NULL	3	NULL	5	NULL	NULL	4
3	Architect|Consultant|Administrator	netflix	NULL	NULL	NULL	NULL	NULL	5	NULL	NULL	NULL	NULL	4
4	Chief|Top Leaders	netflix	5	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
5	Customer Service|Representative	netflix	NULL	1.82	2.33	1.5	2	4	3.38	5	2	4	3
6	Developer|Programmer|Designer	netflix	NULL	NULL	NULL	NULL	NULL	5	NULL	NULL	5	NULL	5
7	Executive|Principal role|Director	netflix	4.63	3	5	5	5	5	5	5	NULL	4	4
8	Lead|Manager|Assistant	netflix	4	3.56	3	5	2.5	5	5	4.5	1	5	3
9	Other Roles	netflix	4.13	4	2.73	2.88	3.47	3.5	3.48	3.69	5	3.54	3.76
10	Other Senior Roles	netflix	4.25	3.3	3.5	3	3.25	4	5	5	5	5	5
11	Partner	netflix	NULL	NULL	NULL	NULL	NULL	NULL	NULL	2	NULL	NULL	NULL
12	Sales|Operations|Quality|Product|Marketing|Management	netflix	4	1	1	NULL	2	5	5	3	5	1	3
13	Scientist|Engineer|Technical	netflix	NULL	1	NULL	2	1	4	4	5	NULL	3	NULL
14	Senior Manager	netflix	NULL	2	NULL	3	3	5	4	1	2	5	4
15	Specialist|Tech Support|Technician	netflix	NULL	NULL	NULL	NULL	NULL	4	NULL	1.5	NULL	NULL	1

Executive/Principal role/Director
Lead/Manager/Assistant
Other Roles
Partner

Observations: It appears that almost all the categories of the job roles have almost same distribution of overall ratings across the 6 companies by year. However, because of improper job_title categorisation, it is very difficult to arrive at the accurate analysis. Some more data processing is required for further analysis.

Q7) Display the relationship between the overall rating score vs. the rest of the rating field scores by company. Also, document your findings.

- **Overall-ratings**
  Versus
- **Work-balance stars**
- **Culture values stars**
- **Career opportunities-stars**
- **Comp-benefit-stars**
- **Senior-management-stars**

```
select company,
round(avg(overall_ratings_new),1) as avg_overall_ratings,
round(avg(work_balance_stars_new),1) as avg_work_balance_stars,
round(avg(culture_values_stars_new),1) as avg_culture_values_stars,
round(avg(carrer_opportunities_stars_new),1) as avg_carrer_opportunities_stars,
round(avg(comp_benefit_stars_new),1) as avg_comp_benefit_stars,
round(avg(senior_mangemnet_stars_new),1) as avg_senior_mangemnet_stars
from sujitsonarproject.emp_review_part_buck_table
group by company;
```

	company	avg_overall_ratings	avg_work_balance_stars	avg_culture_values_stars	avg_carrer_opportunities_stars	avg_comp_benefit_stars	avg_senior_mangemnet_stars
1	amazon	3.6	2.9	3.3	3.5	3.6	3.1
2	apple	4	3.3	3.6	3.3	4	3.4
3	facebook	4.5	3.9	4.2	4.3	4.5	4.2
4	google	4.3	3.8	3.8	3.8	4.2	3.7
5	microsoft	3.8	3.5	3.1	3.6	3.9	3.1
6	netflix	3.4	3.1	2.3	2.9	4	3.1

Observations: Work balance at amazon appears to be low compared to the overall ratings
Benefits ratings in apple and google appears to very good compared to other ratings metrics
Facebook tops across all ratings and thus explains the highest overall ratings
Culture and senior management appear to get low ratings for microsoft.
Microsoft ratings across different metrics is uniform within the middle range towards 4 stars.
Netflix doing below average in culture and career opportunities.

Q8) Document your findings for the following:

a) Which corporation is worth working for
b) Classification of satisfied or unsatisfied employees

a) Which corporation is worth working for

In order to analyse the best company to work for, we can summarise all the above analysis like,

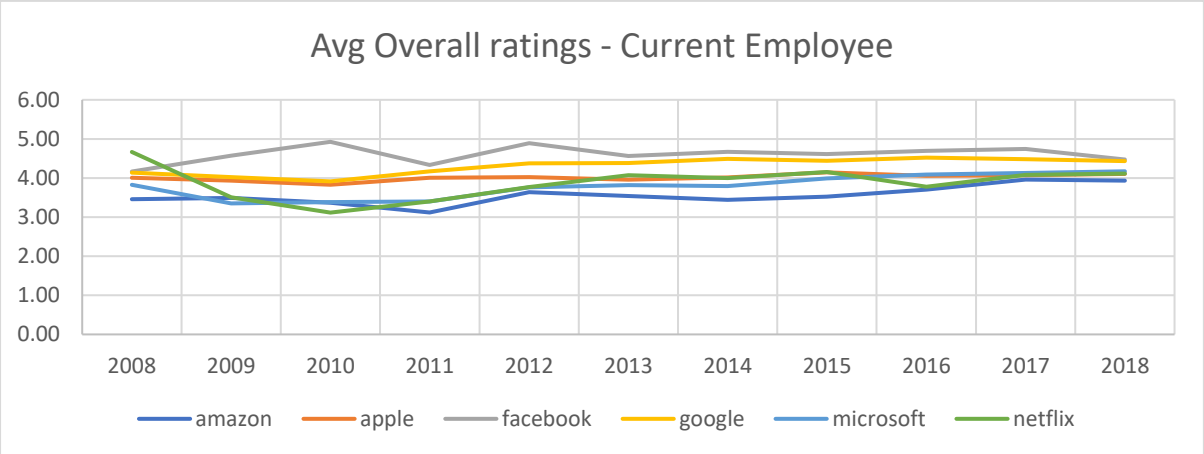
a) Avg Overall ratings by company

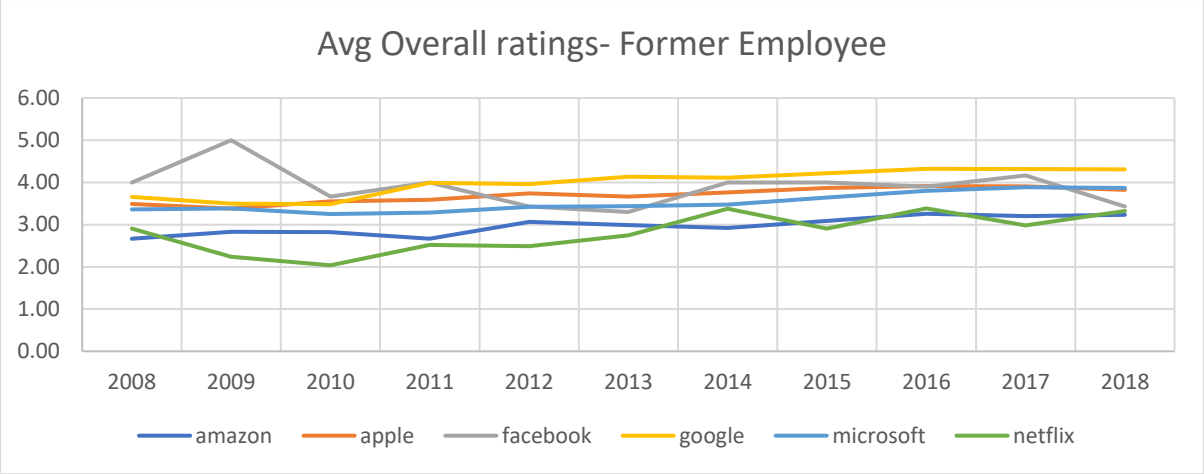
	p2.company	r_0	r_1	r_2	r_3	r_4	r_5
1	amazon	NULL	0.09	0.1	0.21	0.31	0.29
2	apple	NULL	0.04	0.06	0.17	0.35	0.37
3	facebook	NULL	0.03	0.03	0.06	0.15	0.72
4	google	NULL	0.02	0.03	0.1	0.29	0.56
5	microsoft	NULL	0.04	0.07	0.21	0.4	0.28
6	netflix	NULL	0.15	0.14	0.16	0.25	0.3

b) Overall ratings vs the other ratings

	company	avg_overall_ratings	avg_work_balance_stars	avg_culture_values_stars	avg_carrer_opportunities_stars	avg_comp_benefit_stars	avg_senior_mangemnet_stars
1	amazon	3.6	2.9	3.3	3.5	3.6	3.1
2	apple	4	3.3	3.6	3.3	4	3.4
3	facebook	4.5	3.9	4.2	4.3	4.5	4.2
4	google	4.3	3.8	3.8	3.8	4.2	3.7
5	microsoft	3.8	3.5	3.1	3.6	3.9	3.1
6	netflix	3.4	3.1	2.3	2.9	4	3.1

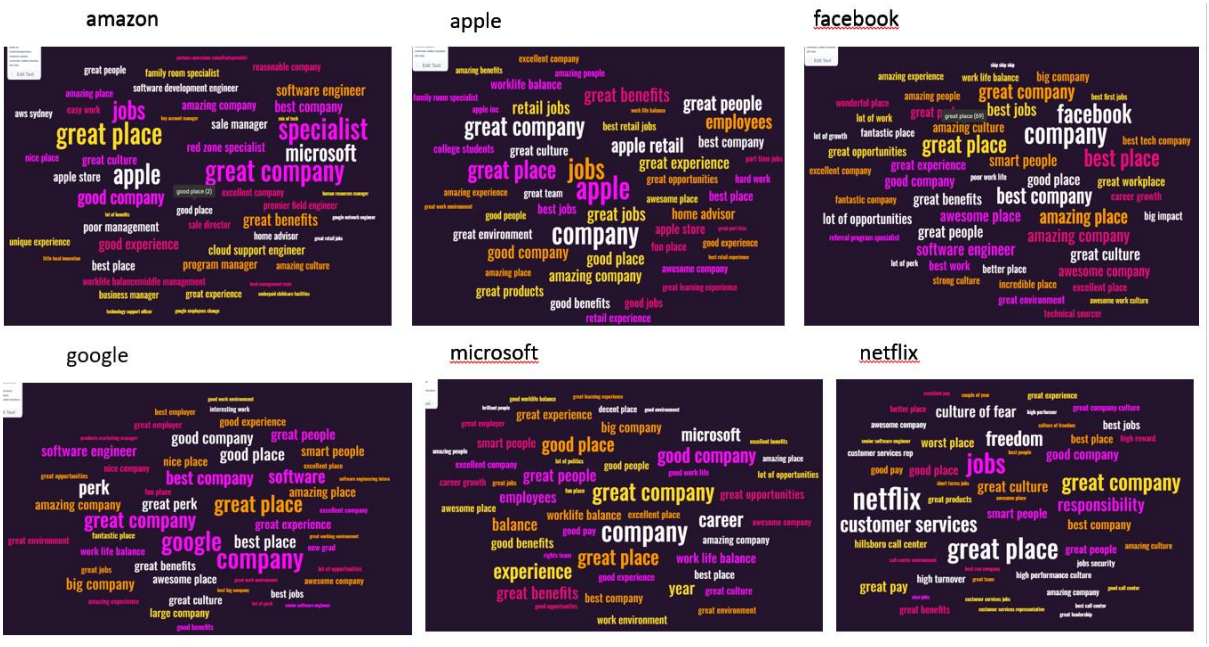
c) Ratings by current vs former employees





d) Analyse the summary, pros and cons comments

Summary comments Analysis:
select distinct REGEXP_REPLACE(summary, '[^A-Za-z]+', '') as summary
from sujitsonarproject.emp_review_part_buck_table
where company='netflix';



Analysing the summary comments using ngrams:

select company, ngrams(sentences(lower(REGEXP_REPLACE(summary, '[^A-Za-z]+', ''))),4,10)
from sujitsonarproject.emp_review_part_buck_table
group by company;

amazon	[{"ngram":["great","place","to","work"],"estfrequency":533.0},{"ngram":["company","to","work","for"],"estfrequency":251.0},{"ngram":["good","place","to","work"],"estfrequency":240.0},{"ngram":["great","company","to","work"],"estfrequency":146.0},{"ngram":["best","place","to","work"],"estfrequency":114.0},{"ngram":["good","company","to","work"],"estfrequency":104.0},{"ngram":["a","great","place","to"],"estfrequency":93.0},{"ngram":["great","place","to","learn"],"estfrequency":92.0},{"ngram":["work","hard","have","fun"],"estfrequency":91.0},{"ngram":["a","good","place","to"],"estfrequency":71.0}]
apple	[{"ngram":["great","place","to","work"],"estfrequency":383.0},{"ngram":["company","to","work","for"],"estfrequency":278.0},{"ngram":["great","company","to","work"],"estfrequency":158.0},{"ngram":["a","great","place","to"],"estfrequency":96.0},{"ngram":["good","place","to","work"],"estfrequency":69.0},{"ngram":["apple","is","a","great"],"estfrequency":46.0},{"ngram":["one","of","the","best"],"estfrequency":41.0},{"ngram":["place","to","work","but"],"estfrequency":40.0},{"ngram":["a","great","company","to"],"estfrequency":33.0},{"ngram":["fun","place","to","work"],"estfrequency":31.0}]
facebook	[{"ngram":["great","place","to","work"],"estfrequency":69.0},{"ngram":["company","to","work","for"],"estfrequency":43.0},{"ngram":["amazing","place","to","work"],"estfrequency":26.0},{"ngram":["best","place","to","work"],"estfrequency":26.0},{"ngram":["awesome","place","to","work"],"estfrequency":19.0},{"ngram":["great","company","to","work"],"estfrequency":19.0},{"ngram":["place","to","work","at"],"estfrequency":13.0},{"ngram":["a","great","place","to"],"estfrequency":12.0},{"ngram":["best","company","to","work"],"estfrequency":11.0},{"ngram":["the","best","place","to"],"estfrequency":11.0}]
google	[{"ngram":["great","place","to","work"],"estfrequency":390.0},{"ngram":["company","to","work","for"],"estfrequency":190.0},{"ngram":["great","company","to","work"],"estfrequency":97.0},{"ngram":["good","place","to","work"],"estfrequency":77.0},{"ngram":["best","place","to","work"],"estfrequency":67.0},{"ngram":["a","great","place","to"],"estfrequency":65.0},{"ngram":["amazing","place","to","work"],"estfrequency":51.0},{"ngram":["place","to","work","but"],"estfrequency":49.0},{"ngram":["best","company","to","work"],"estfrequency":47.0},{"ngram":["one","of","the","best"],"estfrequency":43.0}]
microsoft	[{"ngram":["great","place","to","work"],"estfrequency":847.0},{"ngram":["company","to","work","for"],"estfrequency":490.0},{"ngram":["great","company","to","work"],"estfrequency":314.0},{"ngram":["good","place","to","work"],"estfrequency":307.0},{"ngram":["a","great","place","to"],"estfrequency":200.0},{"ngram":["good","company","to","work"],"estfrequency":139.0},{"ngram":["a","good","place","to"],"estfrequency":102.0},{"ngram":["place","to","work","but"],"estfrequency":102.0},{"ngram":["one","of","the","best"],"estfrequency":82.0},{"ngram":["place","to","work","if"],"estfrequency":68.0}]
netflix	[{"ngram":["great","place","to","work"],"estfrequency":49.0},{"ngram":["place","to","work","if"],"estfrequency":11.0},{"ngram":["a","great","place","to"],"estfrequency":9.0},{"ngram":["company","to","work","for"],"estfrequency":9.0},{"ngram":["to","work","if","you"],"estfrequency":7.0},{"ngram":["is","a","great","place"],"estfrequency":6.0},{"ngram":["netflix","is","a","great"],"estfrequency":6.0},{"ngram":["good","place","to","work"],"estfrequency":5.0},{"ngram":["great","company","to","work"],"estfrequency":5.0},{"ngram":["place","to","work","for"],"estfrequency":5.0}]

Observations: by looking at the top 10 ngrams for the summary column and the word cloud, we see that all the top 10 ngrams talks about all the companies are great place to work. Therefore, we may have to analyse further details to check other parameters apart from great workplace.

Pros comments Analysis:

select company, ngrams(sentences(lower(REGEXP_REPLACE(pros, '[^A-Za-z]+', ''))),4,10)
from sujitsonarproject.emp_review_part_buck_table
group by company;

amazon

apple

facebook

google

microsoft

netflix

company	_c1
amazon	[{"ngram":["great","place","to","work"],"estfrequency":305.0},{"ngram":["people","to","work","with"],"estfrequency":260.0},{"ngram":["company","to","work","for"],"estfrequency":207.0},{"ngram":["a","great","place","to"],"estfrequency":154.0},{"ngram":["good","place","to","work"],"estfrequency":145.0},{"ngram":["get","to","work","with"],"estfrequency":133.0},{"ngram":["one","of","the","best"],"estfrequency":131.0},{"ngram":["if","you","want","to"],"estfrequency":128.0},{"ngram":["you","get","to","work"],"estfrequency":128.0},{"ngram":["great","company","to","work"],"estfrequency":121.0}]
apple	[{"ngram":["company","to","work","for"],"estfrequency":234.0},{"ngram":["people","to","work","with"],"estfrequency":218.0},{"ngram":["great","place","to","work"],"estfrequency":176.0},{"ngram":["great","company","to","work"],"estfrequency":130.0},{"ngram":["some","of","the","best"],"estfrequency":111.0},{"ngram":["you","get","to","work"],"estfrequency":110.0},{"ngram":["a","great","place","to"],"estfrequency":101.0},{"ngram":["get","to","work","with"],"estfrequency":101.0},{"ngram":["great","people","to","work"],"estfrequency":101.0},{"ngram":["people","you","work","with"],"estfrequency":92.0}]
facebook	[{"ngram":["great","place","to","work"],"estfrequency":25.0},{"ngram":["more","open","and","connected"],"estfrequency":18.0},{"ngram":["some","of","the","best"],"estfrequency":17.0},{"ngram":["some","of","the","smartest"],"estfrequency":17.0},{"ngram":["world","more","open","and"],"estfrequency":16.0},{"ngram":["people","to","work","with"],"estfrequency":15.0},{"ngram":["the","world","more","open"],"estfrequency":14.0},{"ngram":["of","the","smartest","people"],"estfrequency":13.0},{"ngram":["a","great","place","to"],"estfrequency":12.0},{"ngram":["work","with","some","of"],"estfrequency":12.0}]
google	[{"ngram":["people","to","work","with"],"estfrequency":137.0},{"ngram":["great","place","to","work"],"estfrequency":135.0},{"ngram":["company","to","work","for"],"estfrequency":97.0},{"ngram":["a","great","place","to"],"estfrequency":70.0},{"ngram":["with","some","of","the"],"estfrequency":51.0},{"ngram":["great","company","to","work"],"estfrequency":50.0},{"ngram":["get","to","work","with"],"estfrequency":49.0},{"ngram":["best","place","to","work"],"estfrequency":45.0},{"ngram":["good","work","life","balance"],"estfrequency":41.0},{"ngram":["some","of","the","smartest"],"estfrequency":41.0}]
microsoft	[{"ngram":["good","work","life","balance"],"estfrequency":341.0},{"ngram":["people","to","work","with"],"estfrequency":278.0},{"ngram":["great","place","to","work"],"estfrequency":276.0},{"ngram":["company","to","work","for"],"estfrequency":208.0},{"ngram":["a","great","place","to"],"estfrequency":186.0},{"ngram":["great","work","life","balance"],"estfrequency":151.0},{"ngram":["work","life","balance","is"],"estfrequency":144.0},{"ngram":["great","company","to","work"],"estfrequency":143.0},{"ngram":["one","of","the","best"],"estfrequency":141.0},{"ngram":["get","to","work","with"],"estfrequency":134.0}]
netflix	[{"ngram":["freedom","and","responsibility","is"],"estfrequency":11.0},{"ngram":["people","to","work","with"],"estfrequency":9.0},{"ngram":["company","to","work","for"],"estfrequency":8.0},{"ngram":["as","long","as","you"],"estfrequency":7.0},{"ngram":["for","a","call","center"],"estfrequency":7.0},{"ngram":["the","freedom","to","do"],"estfrequency":7.0},{"ngram":["treated","like","an","adult"],"estfrequency":7.0},{"ngram":["for","the","most","part"],"estfrequency":6.0},{"ngram":["is","a","great","place"],"estfrequency":6.0},{"ngram":["a","lot","of","freedom"],"estfrequency":5.0}]

Observations: pros also show all the 6 companies are good companies to work.

We see that apple, facebook, google and microsoft has high pros on opportunities, benefits/perks compared to amazon and netflix.

Google and micrososft also shows high pros on work life balance

Cons comments Analysis:

select company, ngrams(sentences(lower(REGEXP_REPLACE(cons, '[^A-Za-z]+', ''))),4,10)
from sujitsonarproject.emp_review_part_buck_table
group by company;

amazon

apple

facebook

google

microsoft

netflix

company	_c1
amazon	[{"ngram":["no","work","life","balance"],"estfrequency":328.0},{"ngram":["work","life","balance","is"],"estfrequency":195.0},{"ngram":["if","you","want","to"],"estfrequency":166.0},{"ngram":["i","can","think","of"],"estfrequency":126.0},{"ngram":["you","have","to","be"],"estfrequency":124.0},{"ngram":["if","you","have","a"],"estfrequency":105.0},{"ngram":["is","a","lot","of"],"estfrequency":103.0},{"ngram":["not","a","lot","of"],"estfrequency":101.0},{"ngram":["when","it","comes","to"],"estfrequency":94.0},{"ngram":["you","are","expected","to"],"estfrequency":93.0}]
apple	[{"ngram":["hard","to","move","up"],"estfrequency":102.0},{"ngram":["no","work","life","balance"],"estfrequency":97.0},{"ngram":["i","can","think","of"],"estfrequency":89.0},{"ngram":["not","a","lot","of"],"estfrequency":79.0},{"ngram":["if","you","want","to"],"estfrequency":77.0},{"ngram":["work","life","balance","is"],"estfrequency":67.0},{"ngram":["at","the","end","of"],"estfrequency":65.0},{"ngram":["when","it","comes","to"],"estfrequency":61.0},{"ngram":["the","end","of","the"],"estfrequency":58.0},{"ngram":["to","move","up","in"],"estfrequency":58.0}]
facebook	[{"ngram":["i","can","think","of"],"estfrequency":23.0},{"ngram":["if","you","want","to"],"estfrequency":19.0},{"ngram":["cant","think","of","any"],"estfrequency":17.0},{"ngram":["can","be","hard","to"],"estfrequency":13.0},{"ngram":["work","life","balance","is"],"estfrequency":13.0},{"ngram":["can","be","a","little"],"estfrequency":12.0},{"ngram":["it","can","be","a"],"estfrequency":12.0},{"ngram":["that","i","can","think"],"estfrequency":9.0},{"ngram":["you","have","to","be"],"estfrequency":9.0},{"ngram":["can","be","a","bit"],"estfrequency":8.0}]
google	[{"ngram":["i","can","think","of"],"estfrequency":87.0},{"ngram":["that","i","can","think"],"estfrequency":45.0},{"ngram":["there","are","no","cons"],"estfrequency":40.0},{"ngram":["to","get","things","done"],"estfrequency":35.0},{"ngram":["it","used","to","be"],"estfrequency":34.0},{"ngram":["can","be","hard","to"],"estfrequency":31.0},{"ngram":["its","a","big","company"],"estfrequency":31.0},{"ngram":["there","are","a","lot"],"estfrequency":31.0},{"ngram":["are","a","lot","of"],"estfrequency":30.0},{"ngram":["none","that","i","can"],"estfrequency":28.0}]
microsoft	[{"ngram":["to","get","things","done"],"estfrequency":105.0},{"ngram":["if","you","are","not"],"estfrequency":104.0},{"ngram":["if","you","want","to"],"estfrequency":102.0},{"ngram":["work","life","balance","is"],"estfrequency":102.0},{"ngram":["i","can","think","of"],"estfrequency":95.0},{"ngram":["when","it","comes","to"],"estfrequency":73.0},{"ngram":["you","need","to","be"],"estfrequency":71.0},{"ngram":["a","lot","of","people"],"estfrequency":68.0},{"ngram":["is","a","lot","of"],"estfrequency":64.0},{"ngram":["not","a","lot","of"],"estfrequency":55.0}]
netflix	[{"ngram":["a","culture","of","fear"],"estfrequency":16.0},{"ngram":["if","you","are","not"],"estfrequency":13.0},{"ngram":["if","you","have","a"],"estfrequency":7.0},{"ngram":["the","culture","of","fear"],"estfrequency":7.0},{"ngram":["when","it","comes","to"],"estfrequency":7.0},{"ngram":["at","the","end","of"],"estfrequency":5.0},{"ngram":["can","be","hard","to"],"estfrequency":5.0},{"ngram":["i","dont","have","any"],"estfrequency":5.0},{"ngram":["may","not","be","the"],"estfrequency":5.0},{"ngram":["the","place","for","you"],"estfrequency":5.0}]

Amazon: manager, management, employees, compensation, people, promotion,
Apple: work life balance, employees, people, management, hours
Facebook: culture, people. Manager, work life balance, employees, team, long hours, jobs
Google: bureaucracy, time, promotion, project, work life balance
Microsoft: management, manager, employees, team, people, politics
Netflix: manager, management, people, employees, team, feedback, culture of fear, jobs

Work balance at amazon appears to be low compared to the overall ratings
Benefits ratings in apple and google appears to very good compared to other ratings metrics
Facebook tops across all ratings and thus explains the highest overall ratings
Culture and senior management appear to get low ratings for microsoft.
Microsoft ratings across different metrics is uniform within the middle range towards 4 stars.
Netflix doing below average in culture and career opportunities.

Observations:
by looking at the avg overall ratings by company by year for Current employee vs the Former Employee, we see that, the former employee avg overall ratings for Amazon, Facebook and Netflix is around 1 point lower than the current employee. This appears to us that the former employee of Amazon, Facebook and Netflix gave lower overall ratings.
There is not much difference for Apple, google and Microsoft.

Summary comments:
by looking at the top 10 ngrams for the summary column and the word cloud, we see that all the top 10 ngrams talks about all the companies are great place to work. Therefore, we may have to analyse further details to check other parameters apart from great workplace.

Pros comments:
pros also show all the 6 companies are good companies to work.
We see that apple, facebook, google and microsoft has high pros on opportunities, benefits/perks compared to amazon and netflix.
Google and micrososft also shows high pros on work life balance

Cons comments:
Amazon: manager, management, employees, compensation, people, promotion
Apple: work life balance, employees, people, management, hours
Facebook: culture, people. Manager, work life balance, employees, team, long hours, jobs
Google: bureaucracy, time, promotion, project, work life balance
Microsoft: management, manager, employees, team, people, politics
Netflix: manager, management, people, employees, team, feedback, culture of fear, jobs

All the companies are great place to work as per the analysis however, the preferred companies (in order) by looking at overall ratings, avg current and former employee ratings, summary comments, pros and cons (which is not truly) accurate and based on the above analysis, corporation to consider (that is worth working)
Google
Apple
Facebook
Microsoft
Amazon
Netflix

b) Classification of satisfied or unsatisfied employees:

```
select company,year,job_title,summary,pros,cons,overall_ratings_new,
work_balance_stars_new,culture_values_stars_new,carrer_opportunities_stars_new,
comp_benefit_stars_new,senior_mangemnet_stars_new,country,
case when overall_ratings_new >=4 then 'Satisfied' else 'Unsatisfied' end as
emp_ratings_classification
from sujitsonarproject.emp_review_part_buck_table limit 5;
```

company	year	job_title	summary	pros	cons	overall_ra	work_bala	culture_va	carrer_opp	comp_ben	senior_ma	country	emp_ratings_classification
apple	2012	Current Employee - Waiter/Steward	i like apple	its a big co	good good	5	3	3	3	3	2	Afghanista	Satisfied
microsoft	2015	Former Employee - Account Manager	Microsoft	Impacting	Metric orie	3	1	5	2	4	1	Algeria	Unsatisfied
google	2017	Current Employee - Anonymous Employee	good	payment c	strict rules	4	4	4	3	4	4	Algeria	Satisfied
google	2018	Current Employee - Product Marketing Manager	Tindouf Al	Looking fo	Perhaps th	4	5	5	5	5	5	Algeria	Satisfied
microsoft	2015	Former Employee - Sales Excellence Operation Analyst	YP	Very good	The YP pro	3	3	3	3	3	3	Argentina	Unsatisfied