﻿**Part A**

Assume the following queries.

1. select \* from productions where year IS NULL;

2. select \* from productions where year = 2014;

3. select \* from productions where year > 1990 and year <=1992

4. select \* from productions where year IS NULL and year = 2014;

For each of these queries:

1. Using EXPLAIN ANALYZE, record the number of tuples that postgres estimates each query will return and the actual number of tuples returned.

2. Compute, for each of the 4 queries:

(a) Using only the information in pg\_stats and pg\_class compute the selectivity of the where clause

(b) using this selectivity, compute the expected number of matching tuples

Show all your work.

**MY APP**

>> EXPLAIN ANALYZE SELECT \* FROM productions WHERE year IS NULL;

QUERY PLAN

---------------------------------------------------------------------------------------------------------------------------

Seq Scan on productions (cost=0.00..471807.22 rows=1353447 width=71) (actual time=0.057..16765.911 rows=1363931 loops=1)

Filter: (year IS NULL)

Rows Removed by Filter: 8482291

Planning time: 3.708 ms

Execution time: 16844.962 ms

(5 rows)

(END)

>> SELECT null\_frac FROM pg\_stats WHERE tablename = 'productions' AND attname = 'year';  
  
null\_frac

-----------

0.137433

>> SELECT relname,reltuples FROM pg\_class WHERE relname = 'productions';

relname | reltuples

-------------+-------------

productions | 9.84802e+06

(1 row)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Number of tuples for PRODUCTIONS (9.84802e+06) \* Number of NULL Tuples with PRODUCTIONS as table-name and YEAR as attribute-name (0.137433) == 1353442.933

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

>> explain analyse select \* from productions where year = 2014;

QUERY PLAN

---------------------------------------------------------------------------------------------------------------------------------------

Gather (cost=1000.00..463336.68 rows=377179 width=71) (actual time=0.353..7259.748 rows=349698 loops=1)

Workers Planned: 2

Workers Launched: 2

-> Parallel Seq Scan on productions (cost=0.00..424618.78 rows=157158 width=71) (actual time=1.177..7186.802 rows=116566 loops=3)

Filter: (year = 2014)

Rows Removed by Filter: 3165508

Planning time: 5.494 ms

Execution time: 7281.088 ms

(8 rows)

(END)

>> SELECT

tablename,

attname,

most\_common\_vals,

most\_common\_freqs,

null\_frac

FROM

pg\_stats

WHERE

tablename = 'productions' AND attname = 'year';

tablename | attname | most\_common\_vals | most\_common\_freqs | null\_frac -------------+---------+-------------------------------------------------------------------------------------------------------------------------------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+----------- productions | year | {2018,2017,2022,2021,2019,2016,2014,2020,2015,2013,2012,2011,2023,2010,2009,2008,2007,2006,2005,2003,2004,2002,2001,2000,1998,1999} | {0.0445333,0.0437,0.0436333,0.0426667,0.0410333,0.0407667,0.0383,0.0382667,0.0362333,0.0361,0.0278333,0.0274,0.0240333,0.0234,0.0218667,0.0196667,0.0182667,0.0179667,0.0162667,0.0128,0.0127333,0.0112333,0.0104,0.0100333,0.0086,0.0085} | 0.137433 (1 row) (END)

>> SELECT

relname,

reltuples

FROM

pg\_class

WHERE

relname = 'productions';

relname | reltuples

-------------+-------------   
productions | 9.84802e+06

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Expected Tuples = Total Tuples × Selectivity  
  
Number of tuples for PRODUCTIONS (9.84802e+06) \* Number of Tuples with PRODUCTIONS as table-name and YEAR as attribute-name with value ‘2014’ (0.0383) WITH == 377179.166

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

>> explain analyse select \* from productions where year > 1990 and year <=1992;

QUERY PLAN

--------------------------------------------------------------------------------------------------------------------------------------------------

Index Scan using idxproductionsyear on productions (cost=0.43..372567.86 rows=110101 width=71) (actual time=0.062..157.430 rows=102896 loops=1)

Index Cond: ((year > 1990) AND (year <= 1992))

Planning time: 0.091 ms

Execution time: 163.272 ms

(4 rows)

(END)

>> SELECT tablename, attname, most\_common\_vals, most\_common\_freqs, null\_frac, n\_distinct FROM pg\_stats WHERE tablename = 'productions' AND attname = 'year';   
  
tablename | attname | most\_common\_vals | most\_common\_freqs | null\_frac | n\_distinct -------------+---------+-------------------------------------------------------------------------------------------------------------------------------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+-----------+------------ productions | year | {2018,2017,2022,2021,2019,2016,2014,2020,2015,2013,2012,2011,2023,2010,2009,2008,2007,2006,2005,2003,2004,2002,2001,2000,1998,1999} | {0.0445333,0.0437,0.0436333,0.0426667,0.0410333,0.0407667,0.0383,0.0382667,0.0362333,0.0361,0.0278333,0.0274,0.0240333,0.0234,0.0218667,0.0196667,0.0182667,0.0179667,0.0162667,0.0128,0.0127333,0.0112333,0.0104,0.0100333,0.0086,0.0085} | 0.137433 | 130 (1 row) (END)

>> SELECT imdb-> relname, imdb-> reltuples imdb-> FROM imdb-> pg\_class imdb-> WHERE imdb-> relname = 'productions';

relname | reltuples

-------------+-------------

productions | 9.84802e+06 (1 row)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Expected Tuples = Total Tuples × Selectivity

Since 1991 and 1992 is not in the most-common freq values, we assume that is smaller than 0.0085.. so something like 0.0080.

(0.0080+0.0080)/2 = 0.0080  
  
Number of tuples for PRODUCTIONS (9.84802e+06) \* Number of Tuples with PRODUCTIONS as table-name and YEAR as attribute-name with value ‘1991’ and ‘1992’ (0.0080) WITH == 78784.16

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

>> explain analyse select \* from productions where year IS NULL and year = 2014;

QUERY PLAN

------------------------------------------------------------------------------------------------------------------------------------------

Index Scan using idxproductionsyear on productions (cost=0.43..188090.70 rows=51837 width=71) (actual time=0.008..0.008 rows=0 loops=1)

Index Cond: ((year IS NULL) AND (year = 2014))

Planning time: 3.065 ms

Execution time: 0.061 ms

(4 rows)

(END)

>>

schemaname | tablename | attname | inherited | null\_frac | avg\_width | n\_distinct | most\_common\_vals | most\_common\_freqs | histogram\_bounds | correlation | most\_common\_elems | most\_common\_elem\_freqs | elem\_count\_histogram ------------+-----------+---------+-----------+-----------+-----------+------------+------------------+-------------------+------------------+-------------+-------------------+------------------------+---------------------- public | productions | year | f | 0.137433 | 4 | 130 | {2018,2017,2022,2021,2019,2016,2014,2020,2015,2013,2012,2011,2023,2010,2009,2008,2007,2006,2005,2003,2004,2002,2001,2000,1998,1999} | {0.0445333,0.0437,0.0436333,0.0426667,0.0410333,0.0407667,0.0383,0.0382667,0.0362333,0.0361,0.0278333,0.0274,0.0240333,0.0234,0.0218667,0.0196667,0.0182667,0.0179667,0.0162667,0.0128,0.0127333,0.0112333,0.0104,0.0100333,0.0086,0.0085} | {1894,1905,1910,1913,1914,1918,1921,1925,1931,1936,1941,1949,1951,1953,1954,1955,1956,1957,1958,1959,1960,1961,1962,1963,1963,1964,1965,1966,1966,1967,1968,1968,1969,1969,1970,1971,1971,1972,1972,1973,1973,1974,1974,1975,1975,1976,1977,1977,1978,1978,1979,1979,1980,1980,1981,1981,1982,1982,1983,1983,1984,1984,1985,1985,1985,1986,1986,1987,1987,1988,1988,1988,1989,1989,1990,1990,1991,1991,1991,1992,1992,1992,1993,1993,1993,1994,1994,1994,1995,1995,1995,1995,1996,1996,1996,1996,1997,1997,1997,1997,2025} | 0.196448 | | | (1 row)

>>

relname | reltuples

-------------+-------------

productions | 9.84802e+06 (1 row)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Number of tuples for PRODUCTIONS (9.84802e+06) \* Number of Tuples with PRODUCTIONS as table-name and YEAR as attribute-name with value ‘NULL’ and ‘2014’ (0.0080+0.137433) WITH == 1432227.093

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_