— JADS Professional Education—

SQL for Data Science
Exploring the US president database

Jheronimus Academy of Data Science, March 2023
Created by Assist. prof. dr. Geert Monsieur

Contents

Introduction

Background Information

Electoral meetings in which the US-electors choose the president every four years, by and large are of ceremonial significance. They carry on with a more than two hundred year old system. Many citizens of the United States do not even realize it is not them who elect the president, but the electors.

Each federal state is represented by as many electors as delegates in Congress, two electors and two delegates at minimum. The District of Columbia, with the capital of the United States, has three electors of its own. It is due to this ratio of distribution that candidates really battle for large states while hardly ever showing up in small ones.

A total of 538 electors turn in their votes in their state capital. There are 48 states where the candidate with the most votes is rewarded the votes of all electors of the state. It is only in Nebraska and Maine where a candidate receives the exact number of votes submitted for her or him.

There have been three occasions in US-American history, when the candidate who accumulated the majority of the people's votes could not obtain this majority with the electors and lost the election.

Electors are not autonomous in whom they vote for. They generally abide by popular tendency in their home state. However, electors have deviated from this imperative. In 1988, one of the electors from West Virginia did not vote for the democratic presidential candidate, Michael Dukakis, but for his opponent Lloyd Bentsen instead. It cost Dukakis one vote.

In order to win the presidential election, a candidate must obtain 270 of the total of 538 votes.

Alabama	9	Massachusetts	12	Tennessee	11
Alaska	3	Michigan	18	Texas	32
Arizona	8	Minnesota	10	Utah	5
Arkansas	6	Mississippi	7	Vermont	3
California	54	Missouri	11	Virginia	13
Colorado	8	Montana	3	Washington	11
Connecticut	8	Nebraska	5	West Virginia	5
Delaware	3	Nevada	4	Wisconsin	11
Distr. of Columbia	3	New Hampshire	4	Wyoming	3
Florida	25	New Jersey	15		
Georgia	13	New Mexico	5		
Hawaii	4	New York	33		
Idaho	4	North Carolina	14		
Illinois	22	North Dakota	3		
Indiana	12	Ohio	21		
Iowa	7	Oklahoma	8		
Kansas	6	Oregon	7		
Kentucky	8	Pennsylvania	23		
Louisiana	9	Rhode Island	4		
Maine	4	South Carolina	8		
Maryland	10	South Dakota	3		

Diagram of the President Database

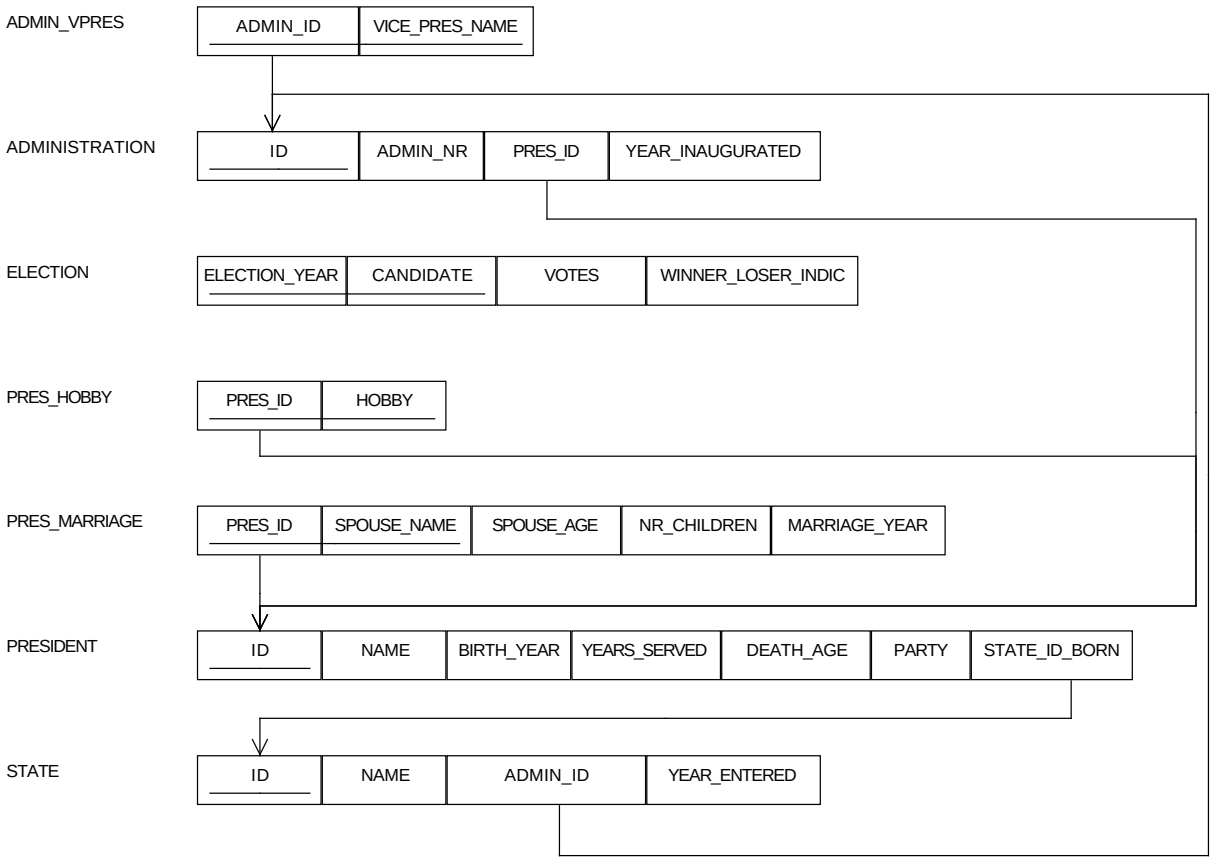

Figure: Relational model of President database

SQL basics and `GROUP BY`

1	How many states brought forth a president born before 1900? **(1)**

count
14

Solution (Q-ID:1-10)

```
SELECT COUNT(DISTINCT(state_id_born))
FROM president
WHERE birth_year < 1900
```

2	Determine the maximum difference between the youngest died and oldest died president that served more than 4 years. **(1)**

difference
33

Solution (Q-ID:2-2)

```
SELECT MAX(death_age)-MIN(death_age) as difference
FROM president
WHERE years_served > 4
```

3	Determine, for each election after 1850, with more than two candidates, in which the winner obtained at least 80% of all votes, election year and number of candidates. **(2)**

election_year	count
1872	5
1912	3
1956	3
1972	3

Solution (Q-ID:3-5)

```
SELECT election_year, COUNT(candidate)
    FROM election WHERE election_year > 1850
    GROUP BY election_year
    HAVING COUNT(candidate) >2 and MAX(votes) >= SUM(VOTES) *0.80
    ORDER BY election_year
```

4	Select, of all elections after 1900 with 2 candidates, the maximum difference between the winning number of votes and the loser's number of votes. **HINT:** Use the SQL keyword `WITH`, which is also referred to as a *common table expression*, that allows you to define a kind of temporary table that can be used in the `FROM` clause.

max
515

Solution (Q-ID:3-4)

```
WITH winner_loser_differences as (
    SELECT election_year, MAX(votes)-MIN(votes) as difference
    FROM election
    GROUP BY election_year
    HAVING COUNT(candidate) = 2
)
SELECT MAX(difference)
FROM winner_loser_differences
```

5	Determine names and election results (election year and number of votes) for all democratic presidents taking part in elections after 1900 and born in a state that joined the federation after 1800. **(1)**

name	election_year	votes
CLINTON W J	1992	370
CLINTON W J	1996	379
JOHNSON L B	1964	486
OBAMA B	2008	365
OBAMA B	2012	332
TRUMAN H S	1948	303

Solution (Q-ID:4-2)

```
SELECT p.name, election_year, SUM(votes) AS votes
  FROM president p
  INNER JOIN election e ON p.name = e.candidate
  INNER JOIN state s ON p.state_id_born = s.id
  WHERE p.party = 'DEMOCRATIC' AND e.election_year > 1900 AND s.
     year_entered > 1800
  GROUP BY p.name, election_year
```

6	Determine, for all presidents whose tenure lasted at least eight years and who married at least twice, names, number of years in tenure, number of marriages and total number of children in all marriages. **(1)**

name	sum	nummar	sumchi
WILSON W	8	2	3
REAGAN R	8	2	4

Solution (Q-ID:4-4)

```
SELECT p.name, SUM(DISTINCT p.years_served), COUNT(m.pres_id) AS nummar,
    SUM(m.nr_children) AS sumchi
  FROM president p
  INNER JOIN pres_marriage m ON p.id = m.pres_id
  WHERE p.years_served >= 8
  GROUP BY p.id
  HAVING COUNT(DISTINCT spouse_name) >= 2
```

Nested queries

Subqueries without correlation

1	Determine name and party of all presidents born after 1800, who were married at least twice. **(1)**

name	party
HARRISON B	REPUBLICAN
ROOSEVELT T	REPUBLICAN
WILSON W	DEMOCRATIC
REAGAN R	REPUBLICAN
TRUMP D J	REPUBLICAN
BIDEN J R	DEMOCRATIC

Solution (Q-ID:5-5)

```
SELECT name, party
    FROM president
    WHERE birth_year > 1800 AND id IN (
        SELECT pres_id
        FROM pres_marriage
        GROUP BY pres_id
        HAVING COUNT(pres_id) >= 2
    )
```

2	Determine, for all parties where more than two presidents born after 1900 belonged to, the total number of presidents it brought forth. **(1)**

party	num
REPUBLICAN	19
DEMOCRATIC	16

Solution (Q-ID:5-7)

```
SELECT party, COUNT(*) as num
    FROM president
    WHERE party IN (
        SELECT party
        FROM president
        WHERE birth_year > 1900
        GROUP BY party
        HAVING COUNT(president) >= 2
    )
    GROUP BY party
```

Subqueries with correlation

1	Determine name, party and years served of the presidents with the most years of tenure in their party. **(1)**

name	party	years_served
ROOSEVELT F D	DEMOCRATIC	12
JEFFERSON T	DEMO-REP	8
MADISON J	DEMO-REP	8
MONROE J	DEMO-REP	8
GRANT U S	REPUBLICAN	8
BUSH G W	REPUBLICAN	8
EISENHOWER D D	REPUBLICAN	8
REAGAN R	REPUBLICAN	8
WASHINGTON G	FEDERALIST	7
TYLER J	WHIG	3

Solution (Q-ID:6-3)

```
SELECT p.name, p.party, p.years_served
    FROM president p
    WHERE (
        SELECT MAX(years_served)
        FROM president
        GROUP BY party
        HAVING party = p.party
    ) = p.years_served
    ORDER BY p.years_served DESC
```

2	Determine names and birth year of all presidents who died at an older age than the average age of all those presidents who were born in the same state. **(1)**

name	birth_year
ADAMS J	1735
JEFFERSON T	1743
MADISON J	1751
MONROE J	1758
ADAMS J Q	1767
VAN BUREN M	1782
FILLMORE M	1800
BUCHANAN J	1791
GRANT U S	1822
HAYES R B	1822
HARRISON B	1833
TAFT W H	1857
COOLIDGE C	1872
EISENHOWER D D	1890
BUSH G H W	1924

Solution (Q-ID:6-4)

```sql
SELECT name, birth_year
    FROM president p
    WHERE death_age > (
        SELECT AVG(death_age)
            FROM president
            GROUP BY state_id_born
            HAVING p.state_id_born = state_id_born
    )
```

3	Which president took part in more elections than he had children? **(3)**

candidate
CLINTON W J
HARDING W G
JACKSON A
MADISON J
NIXON R M
POLK J K
WASHINGTON G

Solution (Q-ID:8-11)

```sql
SELECT candidate
    FROM election e
    INNER JOIN president p ON candidate = p.name
    GROUP BY candidate, p.id
    HAVING COUNT(candidate) > (
        SELECT SUM(nr_children)
        FROM pres_marriage
        WHERE p.id = pres_id
    )
    ORDER BY candidate
```

Queries with `IN`

1	Determine the names of all states that only presidents originated from whose inauguration years all lay after 1900. **(2)**

name
ARKANSAS
CALIFORNIA
CONNECTICUT
GEORGIA
HAWAII
ILLINOIS
IOWA
MISSOURI
NEBRASKA
TEXAS

Solution (Q-ID:7-3)

```
SELECT DISTINCT state.name
    FROM president
    INNER JOIN state
    ON state.id = state_id_born
    WHERE state_id_born NOT IN (
        SELECT DISTINCT state_id_born
            FROM administration
            INNER JOIN president
            ON pres_id = president.id
            WHERE year_inaugurated < 1900
    ) ORDER BY state.name
```

2	Determine the names of all states where presidents were born who had no children at all. **(2)**

name
NORTH CAROLINA
VIRGINIA
SOUTH CAROLINA
PENNSYLVANIA
OHIO

Solution (Q-ID:7-8)

```
SELECT DISTINCT s.name
    FROM president p
    INNER JOIN state s
    ON p.state_id_born = s.id
    WHERE p.id NOT IN (
        SELECT pres_id
        FROM pres_marriage
        WHERE nr_children >0
    )
```

3	Determine the names of the presidents who have the exact same hobbies as “JACKSON A” has - so the same and not more or less hobbies. **(2)**

name
JOHNSON L B
TAYLOR Z
VAN BUREN M
JACKSON A

Solution (Q-ID:7-9)

```
WITH jackson_a_hobbies as (
    SELECT hobby
    FROM pres_hobby ph
    INNER JOIN president p ON ph.pres_id = p.id
    WHERE p.name ='JACKSON A'
```

```
)
SELECT name
FROM pres_hobby INNER JOIN president ON pres_id=id
WHERE pres_id NOT IN (
    SELECT pres_id
    FROM pres_hobby
    WHERE hobby NOT IN (
        SELECT hobby
        FROM jackson_a_hobbies
    )
)
GROUP BY name
HAVING count(*) = (
    SELECT count(*)
    FROM jackson_a_hobbies
)
```

OLAP with SQL: `rollup`, `cube` and `grouping sets`

1	Show the number of democratic or republican presidents born in New York, Ohio or Virginia.

name	party	count
		12
NEW YORK	DEMOCRATIC	2
VIRGINIA	DEMOCRATIC	1
NEW YORK	REPUBLICAN	2
OHIO	REPUBLICAN	7
	REPUBLICAN	9
	DEMOCRATIC	3
OHIO		7
NEW YORK		4
VIRGINIA		1

Solution (Q-ID:20-1)

```
select s.name, p.party, count(*)
from president p
    inner join state s on p.state_id_born = s.id
where s.name in ('NEW YORK','OHIO','VIRGINIA') and p.party in ('
   DEMOCRATIC','REPUBLICAN')
group by cube(p.party,s.name)
```

2	For states that joined the federation (see year_entered column) before 1800, how many democratic, republican and whig presidents were born in these states?

year_entered	party	count
		23
1776	WHIG	4
1776	DEMOCRATIC	12
1776	REPUBLICAN	4
1792	REPUBLICAN	1
1791	REPUBLICAN	2
	REPUBLICAN	7
	WHIG	4
	DEMOCRATIC	12

Solution (Q-ID:20-2)

```
select s.year_entered, p.party, count(*)
from president p
    inner join state s on p.state_id_born = s.id
where s.year_entered < 1800 and p.party in ('DEMOCRATIC','REPUBLICAN','
   WHIG')
group by rollup (p.party,s.year_entered)
```

3	For elections with more than four candidates retrieve the number candidates and the average number of votes of the winners and losers.

election_year	winner_loser_indic	count	avg
1872	W	1	286.0000000000000000
1789	L	11	6.2727272727272727
1836	L	4	31.0000000000000000
1792	L	4	33.0000000000000000
1796	L	12	17.0833333333333333
1872	L	4	15.7500000000000000
1836	W	1	170.0000000000000000
1789	W	1	69.0000000000000000
1800	L	4	50.7500000000000000
1796	W	1	71.0000000000000000
1800	W	1	73.0000000000000000
1792	W	1	132.0000000000000000
1872		5	69.8000000000000000
1800		5	55.2000000000000000
1792		5	52.8000000000000000
1836		5	58.8000000000000000
1789		12	11.5000000000000000
1796		13	21.2307692307692308

Solution (Q-ID:20-3)

```
select election_year,winner_loser_indic,count(*),avg(votes)
from election
where election_year in
    (select election_year
    from election
    group by election_year
    having count(*) > 4)
group by grouping sets (election_year,(election_year,winner_loser_indic))
```

Window functions

1	For each president who has died already, retrieve the name, the death age, the average death age in his/her party and the average death age of presidents coming from the same state.

name	death_age	avg_deathage_party	avg_deathage_state
JACKSON A	78	68.0833333333333333	78.0000000000000000
CLEVELAND G	71	68.0833333333333333	71.0000000000000000
PIERCE F	64	68.0833333333333333	64.0000000000000000
VAN BUREN M	79	68.0833333333333333	69.0000000000000000
JOHNSON L B	65	68.0833333333333333	72.0000000000000000
TRUMAN H S	88	68.0833333333333333	88.0000000000000000
BUCHANAN J	77	68.0833333333333333	71.5000000000000000
JOHNSON A	66	68.0833333333333333	71.5000000000000000
KENNEDY J F	46	68.0833333333333333	77.5000000000000000
WILSON W	67	68.0833333333333333	72.3750000000000000
ROOSEVELT F D	63	68.0833333333333333	69.0000000000000000
POLK J K	53	68.0833333333333333	53.0000000000000000
MONROE J	73	80.2500000000000000	72.3750000000000000
JEFFERSON T	83	80.2500000000000000	72.3750000000000000
ADAMS J Q	80	80.2500000000000000	77.5000000000000000
MADISON J	85	80.2500000000000000	72.3750000000000000
WASHINGTON G	67	78.5000000000000000	72.3750000000000000
ADAMS J	90	78.5000000000000000	77.5000000000000000
LINCOLN A	56	70.4705882352941176	56.0000000000000000
HARDING W G	57	70.4705882352941176	62.2857142857142857
TAFT W H	72	70.4705882352941176	62.2857142857142857
MCKINLEY W	58	70.4705882352941176	62.2857142857142857
HARRISON B	67	70.4705882352941176	62.2857142857142857
GARFIELD J A	49	70.4705882352941176	62.2857142857142857
HAYES R B	70	70.4705882352941176	62.2857142857142857
REAGAN R	93	70.4705882352941176	93.0000000000000000
EISENHOWER D D	79	70.4705882352941176	72.0000000000000000
HOOVER H C	90	70.4705882352941176	90.0000000000000000
NIXON R M	81	70.4705882352941176	81.0000000000000000
FORD G R	93	70.4705882352941176	93.0000000000000000
ARTHUR C A	56	70.4705882352941176	58.0000000000000000
COOLIDGE C	60	70.4705882352941176	58.0000000000000000
BUSH G H W	94	70.4705882352941176	77.5000000000000000
ROOSEVELT T	60	70.4705882352941176	69.0000000000000000
GRANT U S	63	70.4705882352941176	62.2857142857142857
TYLER J	71	69.5000000000000000	72.3750000000000000
FILLMORE M	74	69.5000000000000000	69.0000000000000000
HARRISON W H	68	69.5000000000000000	72.3750000000000000
TAYLOR Z	65	69.5000000000000000	72.3750000000000000

Solution (Q-ID:30-1)

```
1 select name, death_age,
2     avg(death_age) over (partition by party) as avg_deathage_party,
3     avg(death_age) over (partition by state_id_born) as
```

```
        avg_deathage_state
from president
where death_age is not null
```

2	For each candidate in elections after 2000 get the candidate name and the percentage of votes they had in each election.

candidate	votes_percentage
BUSH G W	53.2588454376163873
KERRY J	46.7411545623836127
OBAMA B	67.8438661710037175
MCCAIN J	32.1561338289962825
OBAMA B	61.7100371747211896
ROMNEY M	38.2899628252788104
TRUMP D J	57.2504708097928437
CLINTON H D R	42.7495291902071563
TRUMP D J	43.1226765799256506
BIDEN J R	56.8773234200743494

Solution (Q-ID:30-2)

```
select candidate, cast(votes as decimal)*100 / sum(votes) over(partition
   by election_year) as votes_percentage
from election
where election_year > 2000
```

3	For each president born after 1850, retrieve the president name, birth year, state name where he/she was born and where does he/she fall on the lists of presidents coming from that state he/she was after 1850. Hint: use the row_number() window function.

name	birth_year	name	row_number
TAFT W H	1857	OHIO	1
HARDING W G	1865	OHIO	2
REAGAN R	1911	ILLINOIS	1
TRUMAN H S	1884	MISSOURI	1
CLINTON W J	1946	ARKANSAS	1
EISENHOWER D D	1890	TEXAS	1
JOHNSON L B	1908	TEXAS	2
HOOVER H C	1874	IOWA	1
NIXON R M	1913	CALIFORNIA	1
FORD G R	1913	NEBRASKA	1
OBAMA B	1961	HAWAII	1
WILSON W	1856	VIRGINIA	1
COOLIDGE C	1872	VERMONT	1
BIDEN J R	1942	PENNSYLVANIA	1
KENNEDY J F	1917	MASSACHUSETTS	1
BUSH G H W	1924	MASSACHUSETTS	2
BUSH G W	1946	CONNECTICUT	1
CARTER J M	1924	GEORGIA	1
ROOSEVELT T	1858	NEW YORK	1
ROOSEVELT F D	1882	NEW YORK	2
TRUMP D J	1946	NEW YORK	3

Solution (Q-ID:30-3)

```
select p.name, p.birth_year, s.name, row_number() over(partition by s.id
    order by p.id)
from president p inner join state s on p.state_id_born=s.id
where p.birth_year>1850
```

4	Retrieve the president name, party, years served and the number of years served by presidents coming from the same party up to that president (limit your results to presidents who are member of a non-democratic and non-republican party).

name	party	years_served	years_served_so_far_by_the_party
WASHINGTON G	FEDERALIST	7	7
ADAMS J	FEDERALIST	4	11
JEFFERSON T	DEMO-REP	8	8
MADISON J	DEMO-REP	8	16
MONROE J	DEMO-REP	8	24
ADAMS J Q	DEMO-REP	4	28
HARRISON W H	WHIG	0	0
TYLER J	WHIG	3	3
TAYLOR Z	WHIG	1	4
FILLMORE M	WHIG	2	6

Solution (Q-ID:30-4)

```
select name, party, years_served, sum(years_served) over(partition by
    party order by id) as years_served_so_far_by_the_party
from president
where party <> 'DEMOCRATIC' and party <> 'REPUBLICAN'
order by id
```

5	For each democratic president, retrieve his/her name and hobbies and the total number of democratic presidents who also have this hobby.

name	hobby	count
BIDEN J R	BASEBALL	1
OBAMA B	BASKETBALL	1
OBAMA B	COIKING	1
OBAMA B	DANCING	1
CLEVELAND G	FISHING	3
ROOSEVELT F D	FISHING	3
TRUMAN H S	FISHING	3
WILSON W	GOLF	1
CLINTON W J	PLAYING SAXOPHONE	1
TRUMAN H S	POKER	1
WILSON W	RIDING	4
JACKSON A	RIDING	4
VAN BUREN M	RIDING	4
JOHNSON L B	RIDING	4
KENNEDY J F	SAILING	2
ROOSEVELT F D	SAILING	2
KENNEDY J F	SWIMMING	2
ROOSEVELT F D	SWIMMING	2
KENNEDY J F	TOUCH FOOTBALL	1
TRUMAN H S	WALKING	2
WILSON W	WALKING	2

Solution (Q-ID:30-5)

```
select name, hobby, count(*) over(partition by hobby)
from president inner join pres_hobby on id = pres_id
where party = 'DEMOCRATIC'
```