**INTRODUCTION TO BIG DATA**

**ASSIGNMENT -1**

**SUBMITTED BY-**

Meghali Nandi

4th semester

M.Tech (CSE)

Roll. No. – 2102021

Indian Institute of Information Technology Guwahati

**MYSQL**

**1. What are the Gender wise % of patients across various states?**

SELECT s.State, r.Gender, COUNT(\*)\*100.0 / SUM(COUNT(\*)) OVER(PARTITION BY r.`State code`) AS percent

FROM state\_wise s

JOIN (

  SELECT Gender, `State code` FROM raw\_data1

  UNION ALL SELECT Gender, `State code` FROM raw\_data2

  UNION ALL SELECT Gender, `State code` FROM raw\_data3

  UNION ALL SELECT Gender, `State code` FROM raw\_data4

  UNION ALL SELECT Gender, `State code` FROM raw\_data5

  UNION ALL SELECT Gender, `State code` FROM raw\_data6

  UNION ALL SELECT Gender, `State code` FROM raw\_data7

  UNION ALL SELECT Gender, `State code` FROM raw\_data8

  UNION ALL SELECT Gender, `State code` FROM raw\_data9

  UNION ALL SELECT Gender, `State code` FROM raw\_data10

  UNION ALL SELECT Gender, `State code` FROM raw\_data11

  UNION ALL SELECT Gender, `State code` FROM raw\_data12

  UNION ALL SELECT Gender, `State code` FROM raw\_data13

  UNION ALL SELECT Gender, `State code` FROM raw\_data14

  UNION ALL SELECT Gender, `State code` FROM raw\_data15

  UNION ALL SELECT Gender, `State code` FROM raw\_data16

  UNION ALL SELECT Gender, `State code` FROM raw\_data17

  UNION ALL SELECT Gender, `State code` FROM raw\_data18

  UNION ALL SELECT Gender, `State code` FROM raw\_data19

  UNION ALL SELECT Gender, `State code` FROM raw\_data20

  UNION ALL SELECT Gender, `State code` FROM raw\_data21

  UNION ALL SELECT Gender, `State code` FROM raw\_data22

  UNION ALL SELECT Gender, `State code` FROM raw\_data23

  UNION ALL SELECT Gender, `State code` FROM raw\_data24

  UNION ALL SELECT Gender, `State code` FROM raw\_data25

  UNION ALL SELECT Gender, `State code` FROM raw\_data26

  UNION ALL SELECT Gender, `State code` FROM raw\_data27

  UNION ALL SELECT Gender, `State code` FROM raw\_data28

  UNION ALL SELECT Gender, `State code` FROM raw\_data29

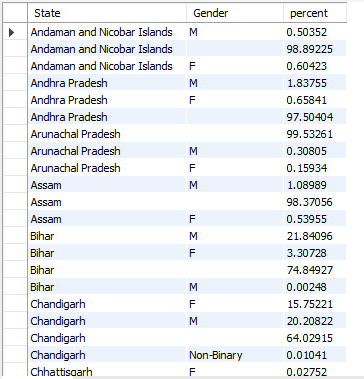
  UNION ALL SELECT Gender, `State code` FROM raw\_data30

  UNION ALL SELECT Gender, `State code` FROM raw\_data31

  UNION ALL SELECT Gender, `State code` FROM raw\_data32

) r ON s.State\_code = r.`State code`

GROUP BY s.State, r.Gender, r.`State code`;





**2. Which age group (in multiple of five like 0-5,5-10,10-15 etc.) has been**

**affected the most?**

SET sql\_mode=(SELECT REPLACE(@@sql\_mode,'ONLY\_FULL\_GROUP\_BY',''));

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data1

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data2

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data3

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data4

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data5

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data6

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data7

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data8

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data9

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data10

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data11

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data12

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data13

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data14

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data15

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data16

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data17

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data18

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data19

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data20

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data21

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data22

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data23

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data24

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data25

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data26

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data27

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data28

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data29

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data30

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data31

UNION

SELECT CONCAT(FLOOR(`Age Bracket`/5)\*5, '-', FLOOR(`Age Bracket`/5)\*5+4) AS age\_group, COUNT(\*) AS num\_patients

FROM raw\_data32

GROUP BY FLOOR(`Age Bracket`/5)

ORDER BY num\_patients DESC limit 1;



**3. Which state has the highest recovery rate and lowest death rate?**

SELECT State,

       SUM(CASE WHEN Patient\_Status = 'Recovered' THEN 1 ELSE 0 END) / COUNT(1) AS recovery\_rate,

       SUM(CASE WHEN Patient\_Status = 'Deceased' THEN 1 ELSE 0 END) / COUNT(1) AS death\_rate

FROM death\_and\_recovered1

UNION

SELECT State,

       SUM(CASE WHEN Patient\_Status = 'Recovered' THEN 1 ELSE 0 END) / COUNT(1) AS recovery\_rate,

       SUM(CASE WHEN Patient\_Status = 'Deceased' THEN 1 ELSE 0 END) / COUNT(1) AS death\_rate

FROM death\_and\_recovered

UNION

SELECT State,

       SUM(CASE WHEN Patient\_Status = 'Recovered' THEN 1 ELSE 0 END) / COUNT(1) AS recovery\_rate,

       SUM(CASE WHEN Patient\_Status = 'Deceased' THEN 1 ELSE 0 END) / COUNT(1) AS death\_rate

FROM death\_and\_recovered2

GROUP BY State

ORDER BY recovery\_rate DESC, death\_rate ASC limit 1;



**4. Which are the top 10 districts where no of infection is more than national average for more than a month?**

SELECT `Date`, District, AVG(Value) AS avg\_cases

FROM district\_testing\_count

WHERE Status = 'Positive'

GROUP BY District

HAVING AVG(Value) > (

    SELECT AVG(Value)

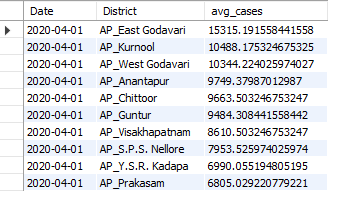
    FROM district\_testing\_count

    WHERE Status = 'Positive'

) AND `Date` < DATE\_SUB(NOW(), INTERVAL 1 MONTH)

ORDER BY avg\_cases DESC

LIMIT 10;



**5. Which are the worst (top 10) affected citied among best (top 10) states?**

SELECT c.District, c.confirmed

FROM (SELECT State, MAX(confirmed) AS confirmed

      FROM state\_wise

      WHERE State NOT IN ('State Unassigned', 'Total')

      GROUP BY state

      ORDER BY confirmed DESC

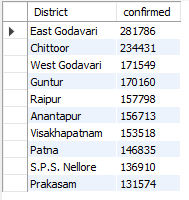
      LIMIT 10) s

JOIN district\_wise c ON s.State = c.State

WHERE c.District != 'Unknown'

ORDER BY confirmed DESC

LIMIT 10;



**MONGO**

1. **What are the Gender wise % of patients across various states?**

db.raw\_data1.aggregate( [

    {"$match": {"raw\_data.gender": {"$exists": True}}},

    {"$unwind": "$raw\_data"},

    {"$group": {"\_id": {"state": "$raw\_data.detectedstate", "gender": "$raw\_data.gender"},

                "count": {"$sum": {"$toInt": "$raw\_data.numcases"}}}},

    {"$group": {"\_id": "$\_id.gender",

                "stateCount": {"$push": {"state": "$\_id.state", "count": "$count"}},

                "total": {"$sum": "$count"}}},

    {"$project": {"\_id": 0, "gender": "$\_id",

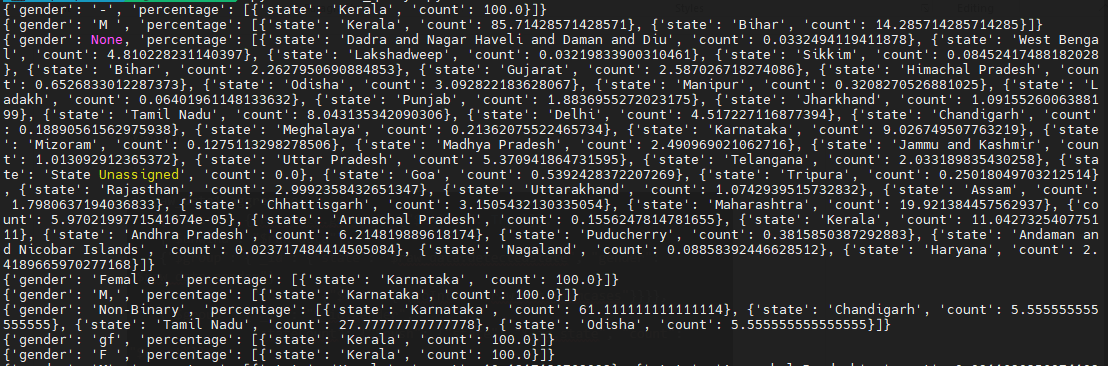
                  "percentage": {"$map": {"input": "$stateCount",

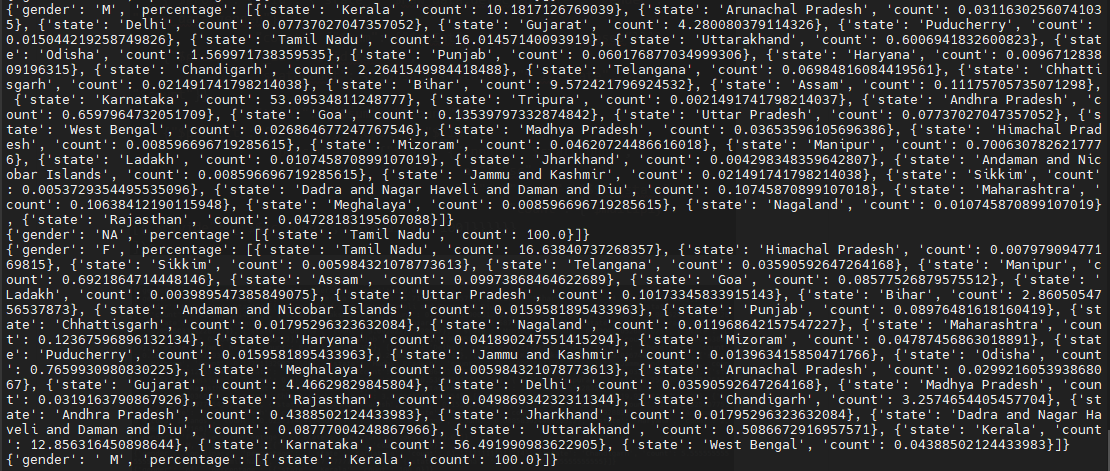
                                          "as": "sc",

                                          "in": {"state": "$$sc.state",

                                                 "count": {"$multiply": [{"$divide": ["$$sc.count", "$total"]}, 100]}}}}}}

]);





1. **Which age group (in multiple of five like 0-5,5-10,10-15 etc.) has been affected the most?**

db["raw\_data1"].aggregate([

    {"$unwind": "$raw\_data"},

    {

        "$project": {

            "age\_group": {

                "$switch": {

                    "branches": [

                        # {"case": {"raw\_data.agebracket":{ "$regex": "/.\*-.\*/"}}, "then":"Unknown"},

                        {"case": {"$and":[{"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 5]},{"$gte": [{"$toDouble":"$raw\_data.agebracket"}, 0]}]}, "then": "0-5"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 10]}, "then": "6-10"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 15]}, "then": "11-15"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 20]}, "then": "16-20"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 25]}, "then": "21-25"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 30]}, "then": "26-30"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 35]}, "then": "31-35"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 40]}, "then": "36-40"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 45]}, "then": "41-45"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 50]}, "then": "46-50"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 55]}, "then": "51-55"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 60]}, "then": "56-60"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 65]}, "then": "61-65"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 70]}, "then": "66-70"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 75]}, "then": "71-75"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 80]}, "then": "76-80"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 85]}, "then": "81-85"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 90]}, "then": "86-90"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 95]}, "then": "91-95"},

                        {"case": {"$lte": [{"$toDouble":"$raw\_data.agebracket"}, 100]}, "then": "96-100"},

                        {"case": {"$gt": [{"$toDouble":"$raw\_data.agebracket"}, 100]}, "then": ">100"}

                    ],

                    "default": "Unknown"

                }

            }

        }

    },

    {

        "$group": {

            "\_id": "$age\_group",

            "count": {"$sum": 1}

        }

    },

    {

        "$sort": {"count": -1}

    },

    {"$limit":1}

])



3. **Which state has the highest recovery rate and lowest death rate?**

db["deaths\_recoveries"].aggregate([

    {"$unwind": "$deaths\_recoveries"},

    {"$group": {

        "\_id": "$deaths\_recoveries.state",

        "total\_cases": {"$sum": 1},

        "total\_deaths": {"$sum": {"$cond": {"if": {"$eq": ["$deaths\_recoveries.patientstatus", "Deceased"]}, "then": 1, "else": 0}}},

        "total\_recovered": {"$sum": {"$cond": {"if": {"$eq": ["$deaths\_recoveries.patientstatus", "Recovered"]}, "then": 1, "else": 0}}}

    }},

    {"$addFields": {"recovery\_rate": {"$divide": ["$total\_recovered", "$total\_cases"]}, "death\_rate": {"$divide": ["$total\_deaths", "$total\_cases"]}}},

    {"$sort": {"recovery\_rate": -1, "death\_rate": 1}},

    {"$limit": 1}

])



4. **Which are the top 10 districts where no of infection is more than national average for more than a month?**

db.raw\_data1.aggregate([

  {"$unwind": "$raw\_data"},

  {

    "$group": {

      "\_id": "$raw\_data.detecteddistrict",

      "totalCases": {

        "$sum": {

          "$toInt": "$raw\_data.numcases"

        }

      }

    }

  },

  {

    "$group": {

      "\_id": None,

      "nationalAverage": {

        "$avg": "$totalCases"

      },

      "districts": {

        "$push": {

          "district": "$\_id",

          "totalCases": "$totalCases"

        }

      }

    }

  },

  {

    "$unwind": "$districts"

  },

  {

    "$match": {

      "gte":["$districts.totalCases","$nationalAverage"]

      }

  },

  {

    "$sort": {

      "districts.totalCases": -1

    }

  },

  {

    "$limit": 10

  },

  {

    "$project": {

      "\_id": 0,

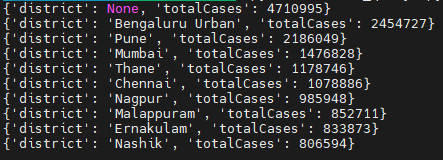
      "district": "$districts.district",

      "totalCases": "$districts.totalCases"

    }

  }

])



5. **Which are the worst (top 10) affected citied among best (top 10) states?**

top\_states = db.state\_district\_wise.aggregate([

    { "$project": { "states": { "$objectToArray": "$$ROOT" } } },

    { "$unwind": "$states" },

    { "$replaceRoot": { "newRoot": "$states" } },

    { "$match": { "v.districtData": { "$exists": True } } },

    { "$project": { "state": "$$ROOT.k", "districts": { "$objectToArray": "$$ROOT.v.districtData" } } },

    { "$unwind": "$districts" },

    { "$project": { "state": 1, "district": "$districts.k", "confirmed": "$districts.v.confirmed" } },

    { "$group": { "\_id": "$state", "total\_confirmed": { "$sum": "$confirmed" }, "cities": { "$push": { "name": "$district", "confirmed": "$confirmed" } } } },

    { "$sort": { "total\_confirmed": -1 } },

    { "$limit": 10 },

    {"$unwind": "$cities"},

    {"$sort": {"cities.confirmed": -1}},

    {"$group": {"\_id": "$cities.name", "confirmed": {"$sum": "$cities.confirmed"}}},

    {"$sort": {"confirmed": -1}},

    {"$limit": 10}

]

)

