Codebasics Resume Challenge#7

Telangana Growth Analysis

in SQL approach

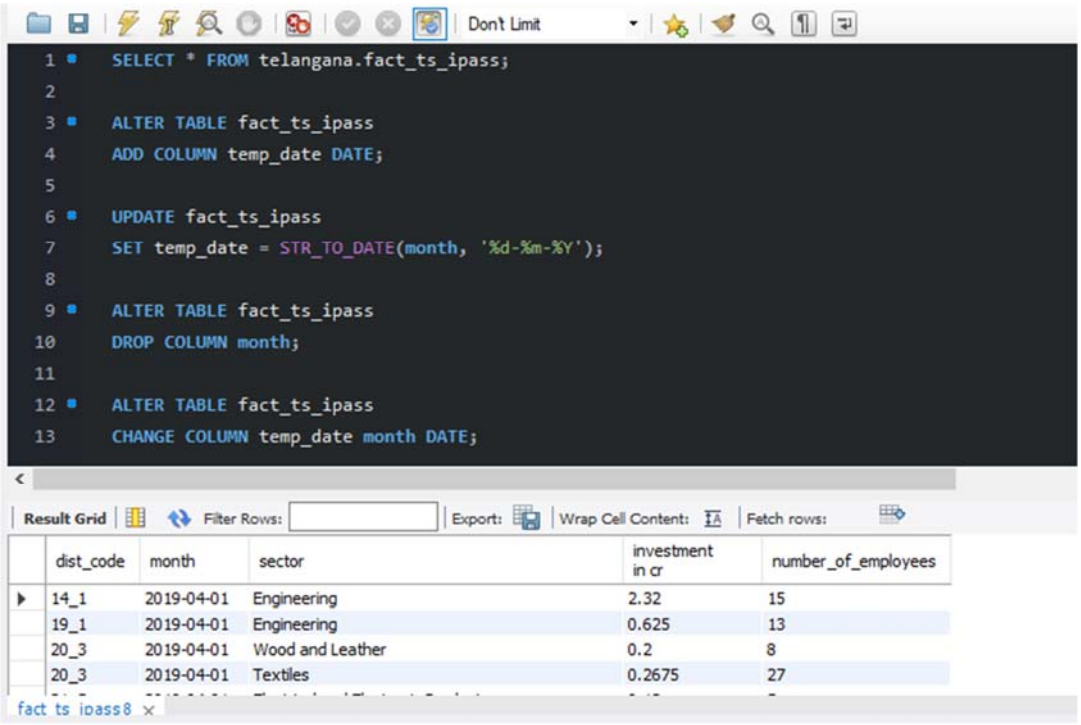
By

May Thu Han

OCT 2023

Data Preparation

1. Import the CSVs into MySQL.
2. The date format is different in one fact table. So changed it into a “year-month-day” format by using these.



```
SELECT * FROM telangana.fact_ts_ipass;

ALTER TABLE fact_ts_ipass
ADD COLUMN temp_date DATE;

UPDATE fact_ts_ipass
SET temp_date = STR_TO_DATE(month, '%d-%m-%Y');

ALTER TABLE fact_ts_ipass
DROP COLUMN month;

ALTER TABLE fact_ts_ipass
CHANGE COLUMN temp_date month DATE;
```

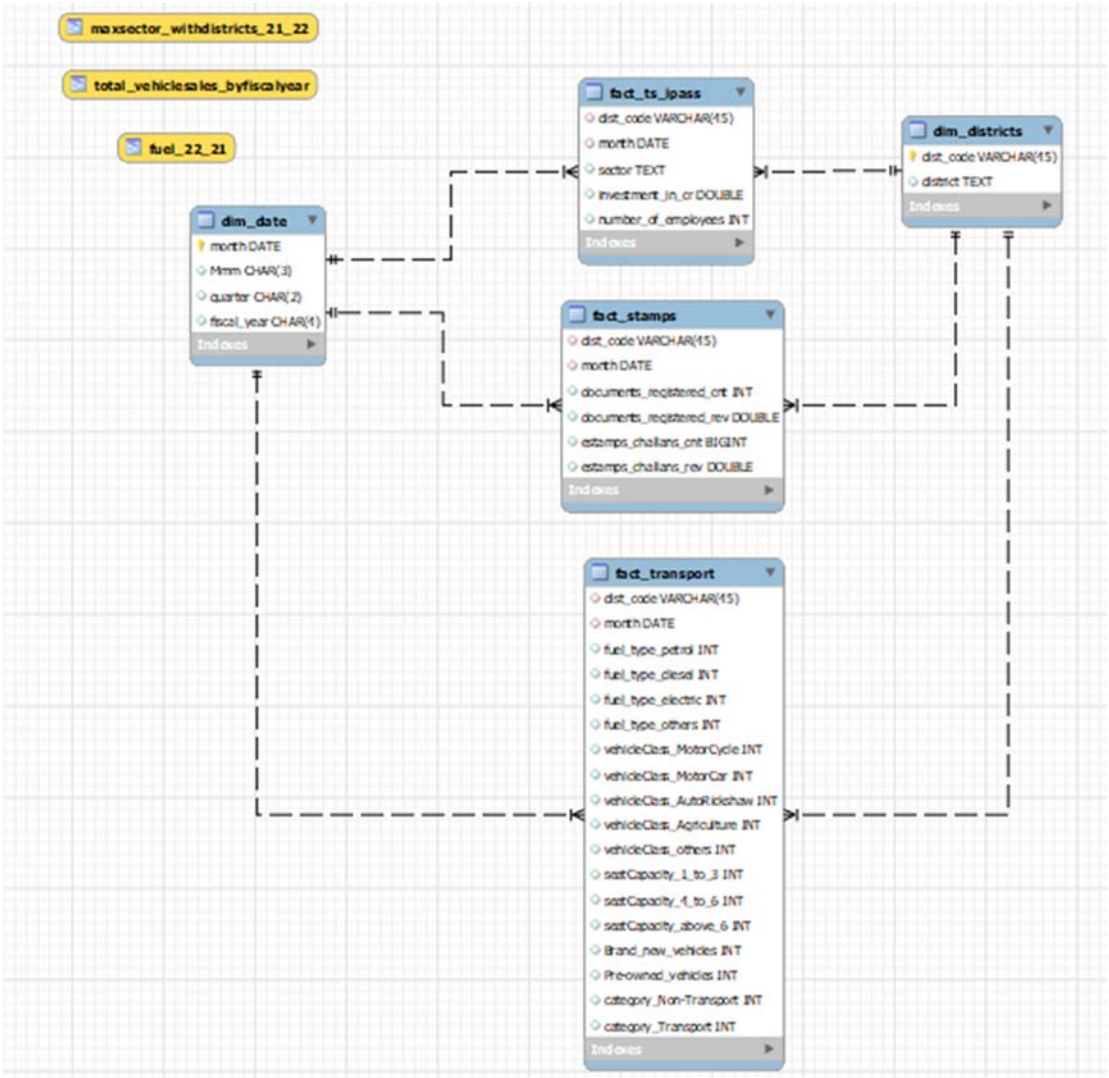
Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

dist_code	month	sector	investment in cr	number_of_employees
14_1	2019-04-01	Engineering	2.32	15
19_1	2019-04-01	Engineering	0.625	13
20_3	2019-04-01	Wood and Leather	0.2	8
20_3	2019-04-01	Textiles	0.2675	27

fact_ts_ipass8

3. Used “Non-identifying Relationships” when the foreign key is not part of the primary key in a table, i.e., when a foreign key does not identify unique records in a table.
4. Select the child table first, then the parent table for making relationships.

Data Model

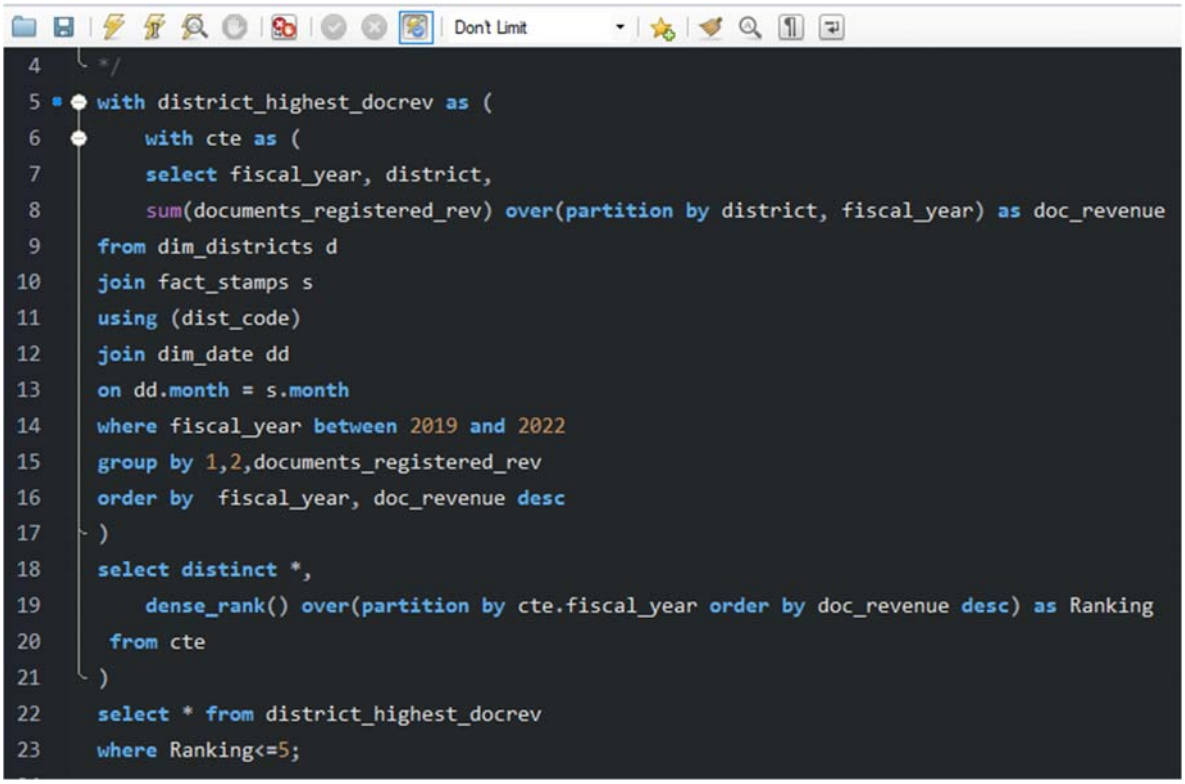

maxsector_withdistricts_21_22
total_vehiclesales_byfiscalyear
fuel_22_21
dim_date
month DATE
Mmm CHAR(3)
quarter CHAR(2)
fiscal_year CHAR(4)
Indexes
fact_ts_ipass
dist_code VARCHAR(15)
month DATE
sector TEXT
investment_in_cr DOUBLE
number_of_employees INT
Indexes
dim_districts
dist_code VARCHAR(15)
district TEXT
Indexes
fact_stamps
dist_code VARCHAR(15)
month DATE
documents_registered_cnt INT
documents_registered_rev DOUBLE
estamps_challans_cnt BIGINT
estamps_challans_rev DOUBLE
Indexes
fact_transport
dist_code VARCHAR(15)
month DATE
fuel_type_petrol INT
fuel_type_diesel INT
fuel_type_electric INT
fuel_type_others INT
vehicleClass_MotorCycle INT
vehicleClass_MotorCar INT
vehicleClass_AutoRickshaw INT
vehicleClass_Agriculture INT
vehicleClass_others INT
seatCapacity_1_to_3 INT
seatCapacity_4_to_6 INT
seatCapacity_above_6 INT
Brand_new_vehicles INT
Pre-owned_vehicles INT
category_Non-Transport INT
category_Transport INT
Indexes

Objective:

- Explore Stamp Registration, Transportation, and TS-IPASS Datasets.
- Understand their attributes, categories, and time period.
- Analyze trends and patterns within each department.
- Identify growth opportunities and areas needing attention.
- Find correlations among these departments and report the overall growth of the state through insights and relevant visuals such as shape maps.

Stamp Registration

1. How does the revenue generated from document registration vary across districts in Telangana? List down the top 5 districts that showed the highest document registration revenue growth between FY 2019 and 2022.

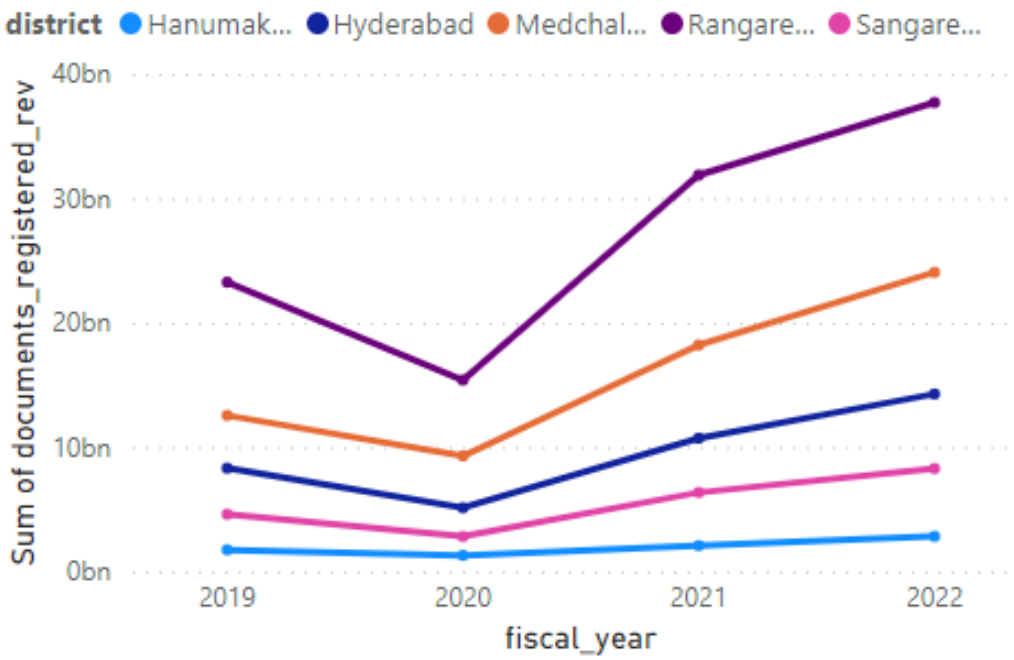


OUTPUT:

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

fiscal_year	district	doc_revenue	Ranking
2019	Rangareddy	23249650199	1
2019	Medchal_Malkajgiri	12543288516	2
2019	Hyderabad	8299406278	3
2019	Sangareddy	4597623320	4
2019	Hanumakonda	1738280378	5
2020	Rangareddy	15385805674	1
2020	Medchal_Malkajgiri	9283678592	2
2020	Hyderabad	5119435090	3
2020	Sangareddy	2809294013	4
2020	Hanumakonda	1294166506	5
2021	Rangareddy	31865475990	1
2021	Medchal_Malkajgiri	18200892201	2
2021	Hyderabad	10711890174	3
2021	Sangareddy	6344385117	4
2021	Hanumakonda	2077808629	5
2022	Rangareddy	37697750946	1
2022	Medchal_Malkajgiri	24043523530	2
2022	Hyderabad	14266012441	3
2022	Sangareddy	8273200321	4
2022	Hanumakonda	2817238587	5

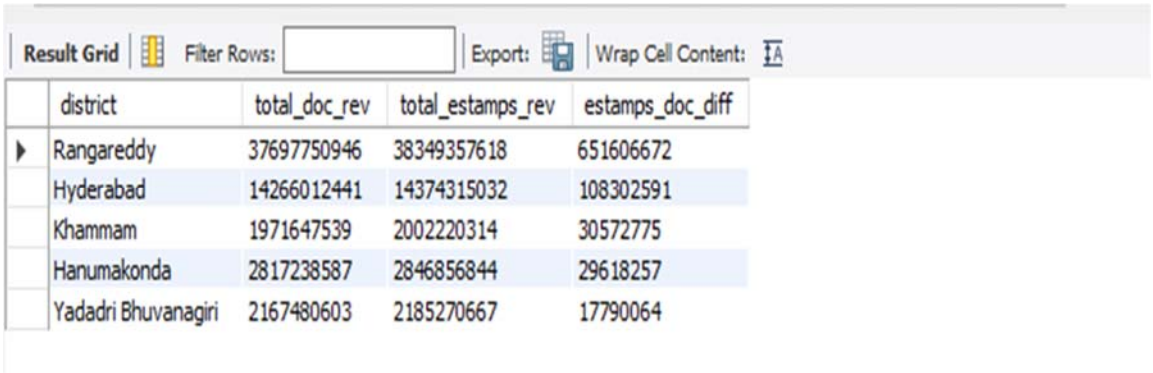
Visualization:



2. How does the revenue generated from document registration compare to the revenue generated from e-stamp challans across districts? List down the top 5 districts where e-stamps revenue contributes significantly more to the revenue than the documents in FY 2022?

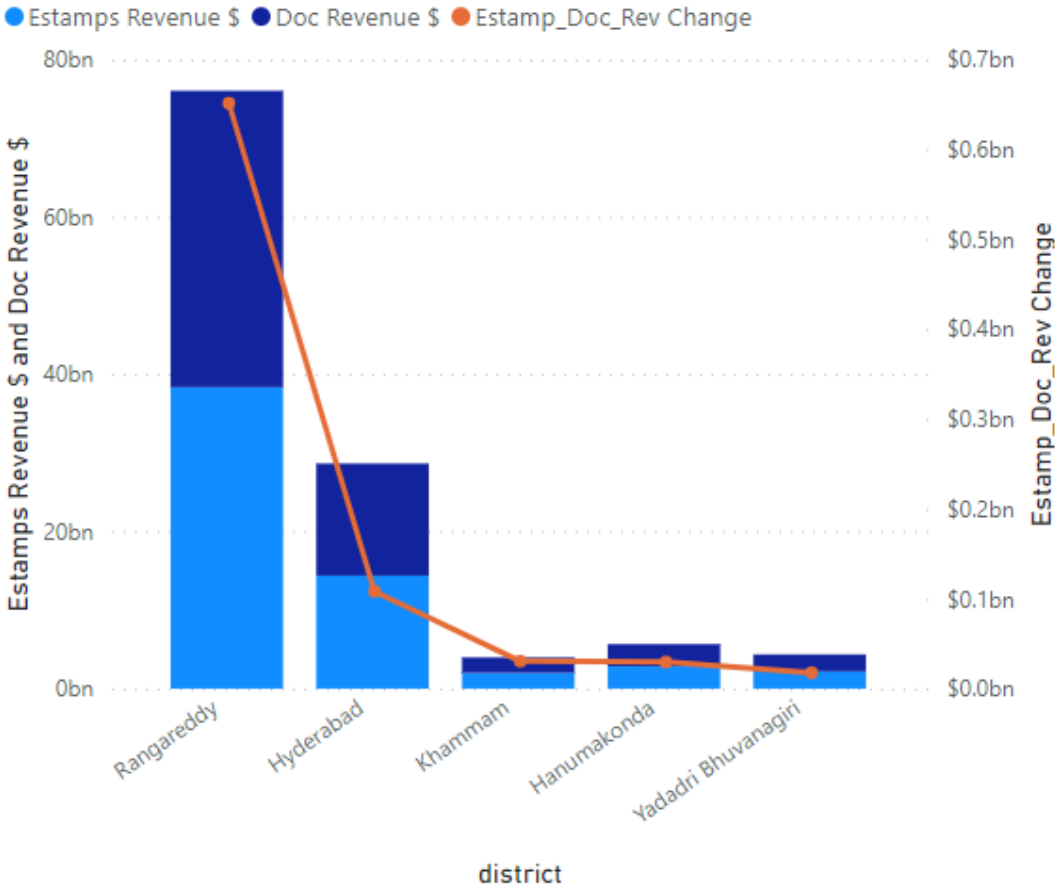
```sql
With cte1 as (
select
    district,
    sum(documents_registered_rev) as total_doc_rev,
    sum(estamps_challans_rev) as total_estamps_rev
from dim_date
join fact_stamps
using (month)
join dim_districts
using (dist_code)
where fiscal_year="2022"
group by 1
order by total_doc_rev desc
)
select *,
    (total_estamps_rev - total_doc_rev) as estamps_doc_diff
from cte1
order by estamps_doc_diff desc
limit 5;
```

OUTPUT:



district	total_doc_rev	total_estamps_rev	estamps_doc_diff
Rangareddy	37697750946	38349357618	651606672
Hyderabad	14266012441	14374315032	108302591
Khammam	1971647539	2002220314	30572775
Hanumakonda	2817238587	2846856844	29618257
Yadadri Bhuvanagiri	2167480603	2185270667	17790064

Visualization:



3. Is there any alteration of e-Stamp challan count and document registration count pattern since the implementation of e-Stamp challan? If so, what suggestions would you propose to the government?

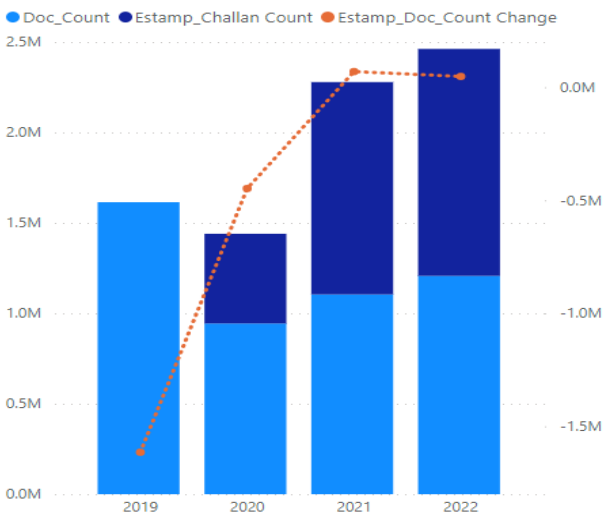
```
52  If so, what suggestions would you propose to the government?
53  */
54  With cte2 as (
55      select
56          fiscal_year,
57          sum(documents_registered_cnt) as total_doc_count,
58          sum(estamps_challans_cnt) as total_estamp_count
59      from dim_date
60      join fact_stamps
61      using (month)
62      join dim_districts
63      using (dist_code)
64      group by fiscal_year
65  )
66  select
67      *,
68      (total_estamp_count - total_doc_count) as estamp_doc_count_Diff
69  from cte2
70  order by estamp_doc_count_Diff desc;
71
```

OUTPUT:

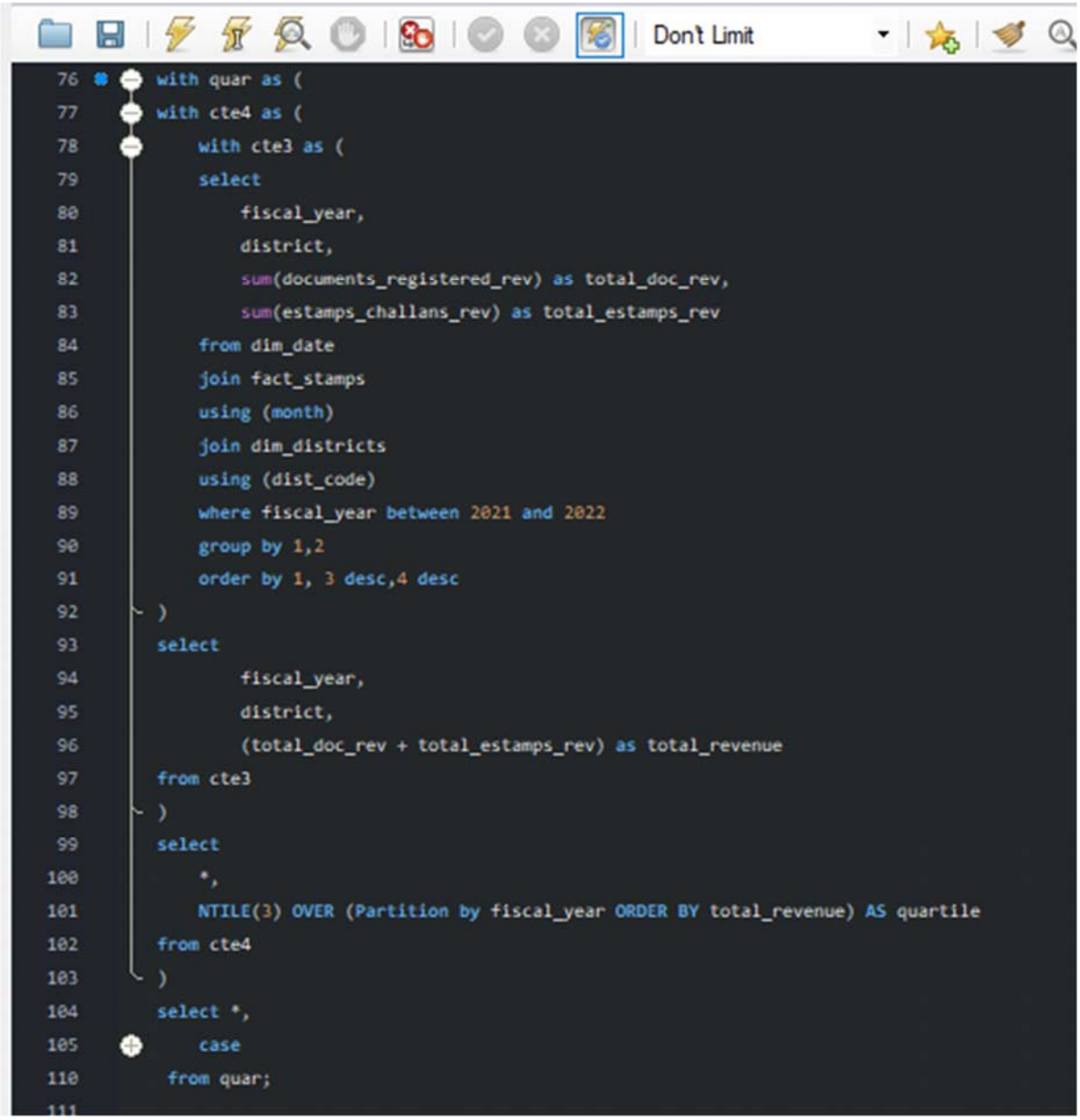
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

fiscal_year	total_doc_count	total_estamp_count	estamp_doc_count_Diff
2021	1104580	1173978	69398
2022	1207073	1254961	47888
2020	943893	496132	-447761
2019	1614417	0	-1614417

Visualization:

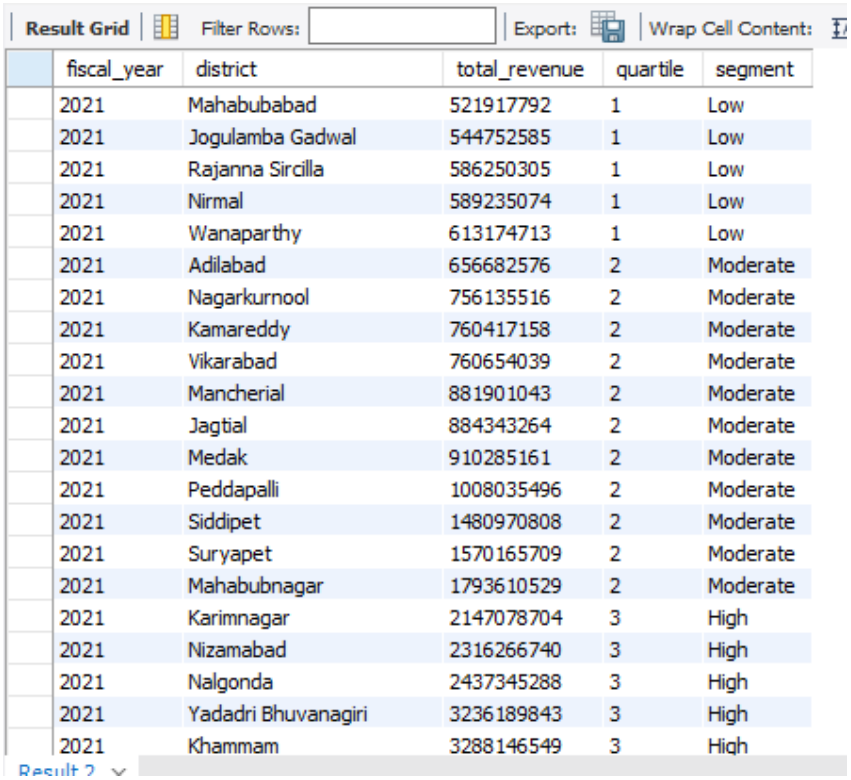


4. Categorize districts into three segments based on their stamp registration revenue generation during the fiscal year 2021 to 2022.




```
with quar as (
with cte4 as (
    with cte3 as (
    select
        fiscal_year,
        district,
        sum(documents_registered_rev) as total_doc_rev,
        sum(estamps_challans_rev) as total_estamps_rev
    from dim_date
    join fact_stamps
    using (month)
    join dim_districts
    using (dist_code)
    where fiscal_year between 2021 and 2022
    group by 1,2
    order by 1, 3 desc,4 desc
)
select
        fiscal_year,
        district,
        (total_doc_rev + total_estamps_rev) as total_revenue
from cte3
)
select
    *,
    NTILE(3) OVER (Partition by fiscal_year ORDER BY total_revenue) AS quartile
from cte4
)
select *,
    case
 from quar;
```

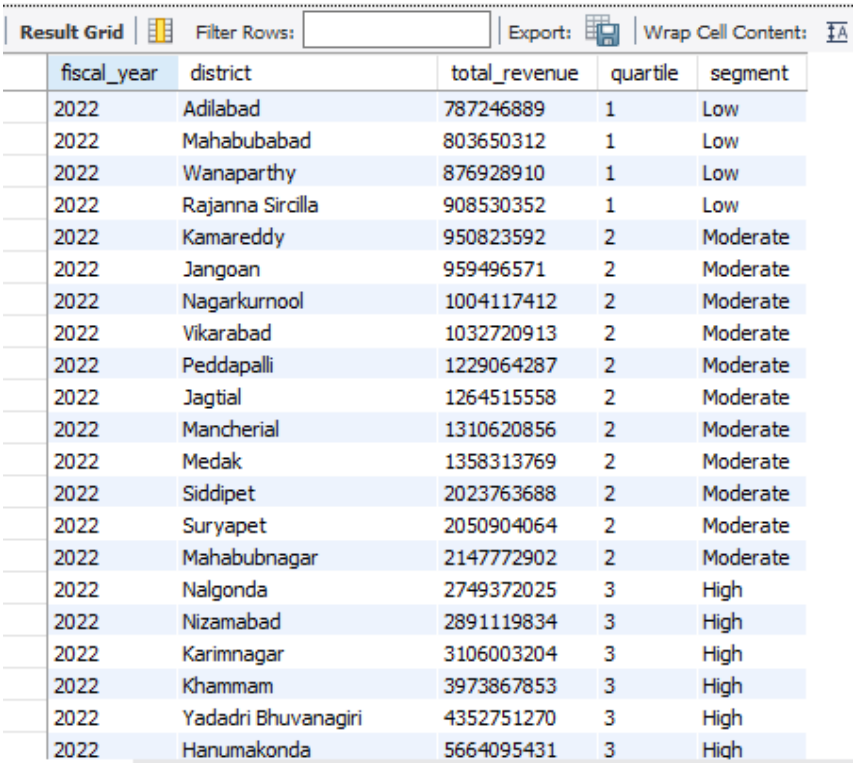
OUTPUT:



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

fiscal_year	district	total_revenue	quartile	segment
2021	Mahabubabad	521917792	1	Low
2021	Jogulamba Gadwal	544752585	1	Low
2021	Rajanna Sircilla	586250305	1	Low
2021	Nirmal	589235074	1	Low
2021	Wanaparthy	613174713	1	Low
2021	Adilabad	656682576	2	Moderate
2021	Nagarkurnool	756135516	2	Moderate
2021	Kamareddy	760417158	2	Moderate
2021	Vikarabad	760654039	2	Moderate
2021	Mancherial	881901043	2	Moderate
2021	Jagtial	884343264	2	Moderate
2021	Medak	910285161	2	Moderate
2021	Peddapalli	1008035496	2	Moderate
2021	Siddipet	1480970808	2	Moderate
2021	Suryapet	1570165709	2	Moderate
2021	Mahabubnagar	1793610529	2	Moderate
2021	Karimnagar	2147078704	3	High
2021	Nizamabad	2316266740	3	High
2021	Nalgonda	2437345288	3	High
2021	Yadadri Bhuvanagiri	3236189843	3	High
2021	Khammam	3288146549	3	High

Result 2

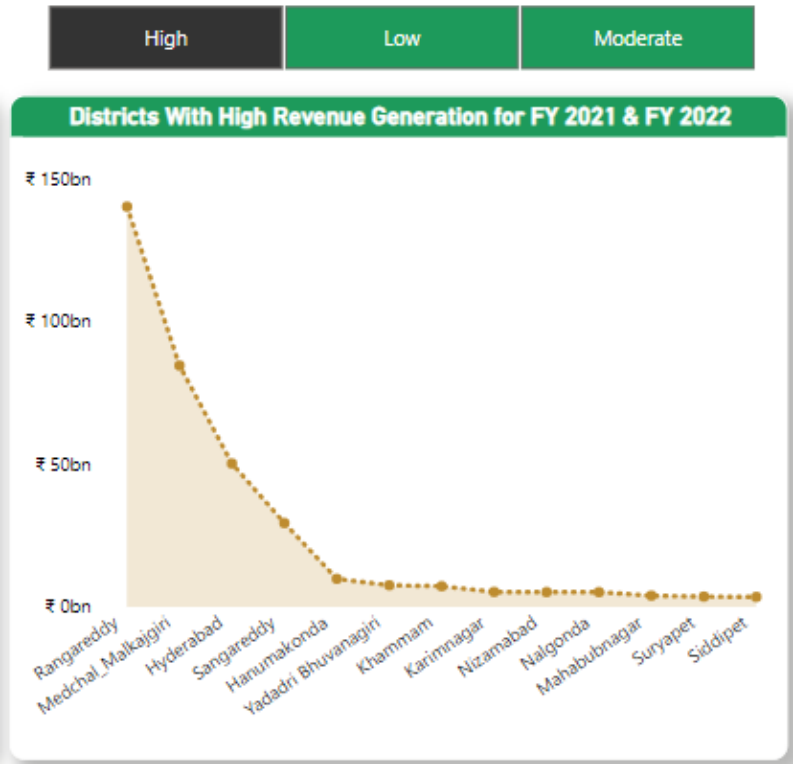


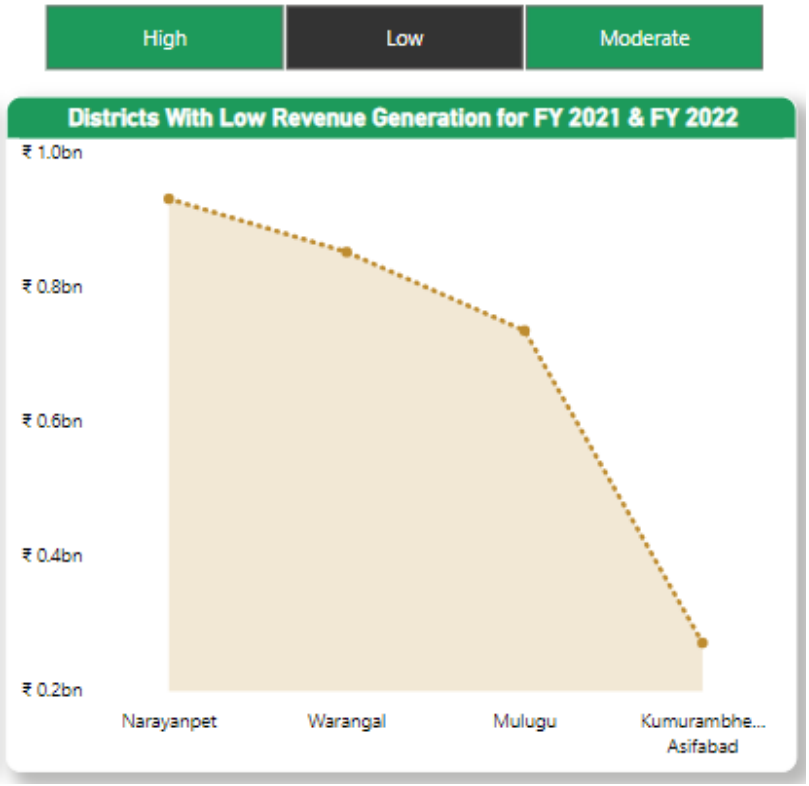
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

fiscal_year	district	total_revenue	quartile	segment
2022	Adilabad	787246889	1	Low
2022	Mahabubabad	803650312	1	Low
2022	Wanaparthy	876928910	1	Low
2022	Rajanna Sircilla	908530352	1	Low
2022	Kamareddy	950823592	2	Moderate
2022	Jangoan	959496571	2	Moderate
2022	Nagarkurnool	1004117412	2	Moderate
2022	Vikarabad	1032720913	2	Moderate
2022	Peddapalli	1229064287	2	Moderate
2022	Jagtial	1264515558	2	Moderate
2022	Mancherial	1310620856	2	Moderate
2022	Medak	1358313769	2	Moderate
2022	Siddipet	2023763688	2	Moderate
2022	Suryapet	2050904064	2	Moderate
2022	Mahabubnagar	2147772902	2	Moderate
2022	Nalgonda	2749372025	3	High
2022	Nizamabad	2891119834	3	High
2022	Karimnagar	3106003204	3	High
2022	Khammam	3973867853	3	High
2022	Yadadri Bhuvanagiri	4352751270	3	High
2022	Hanumakonda	5664095431	3	High

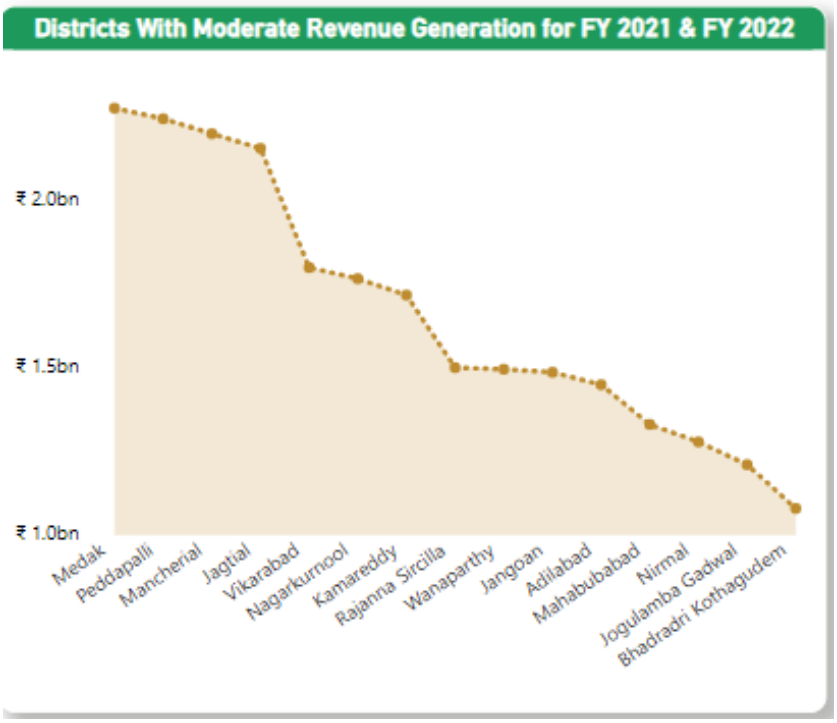
NOTE: The above 2 screenshots do not capture all output results.

Visualization:





High	Low	Moderate



Transportation

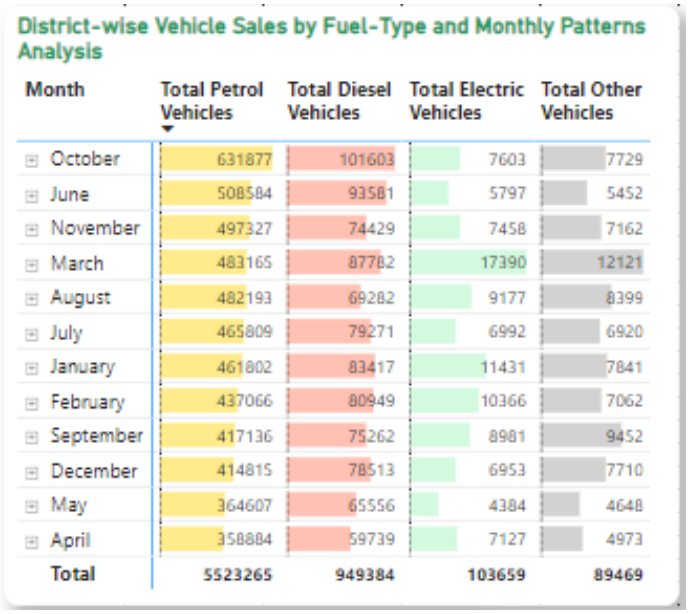
1. Investigate whether there is any correlation between vehicle sales and specific months or seasons in different districts. Are there any months or seasons that consistently show higher or lower sales rate, and if yes, what could be the driving factors? (Consider Fuel-Type category only)

```
select
    fiscal_year,
    Mmm,
    district,
   sum(fuel_type_petrol + fuel_type_diesel + fuel_type_electric + fuel_type_others) as vehicle_sale
from fact_transport
join dim_date
using (month)
join dim_districts
using (dist_code)
group by 1,2,3
order by 1,2, vehicle_sale desc;
```

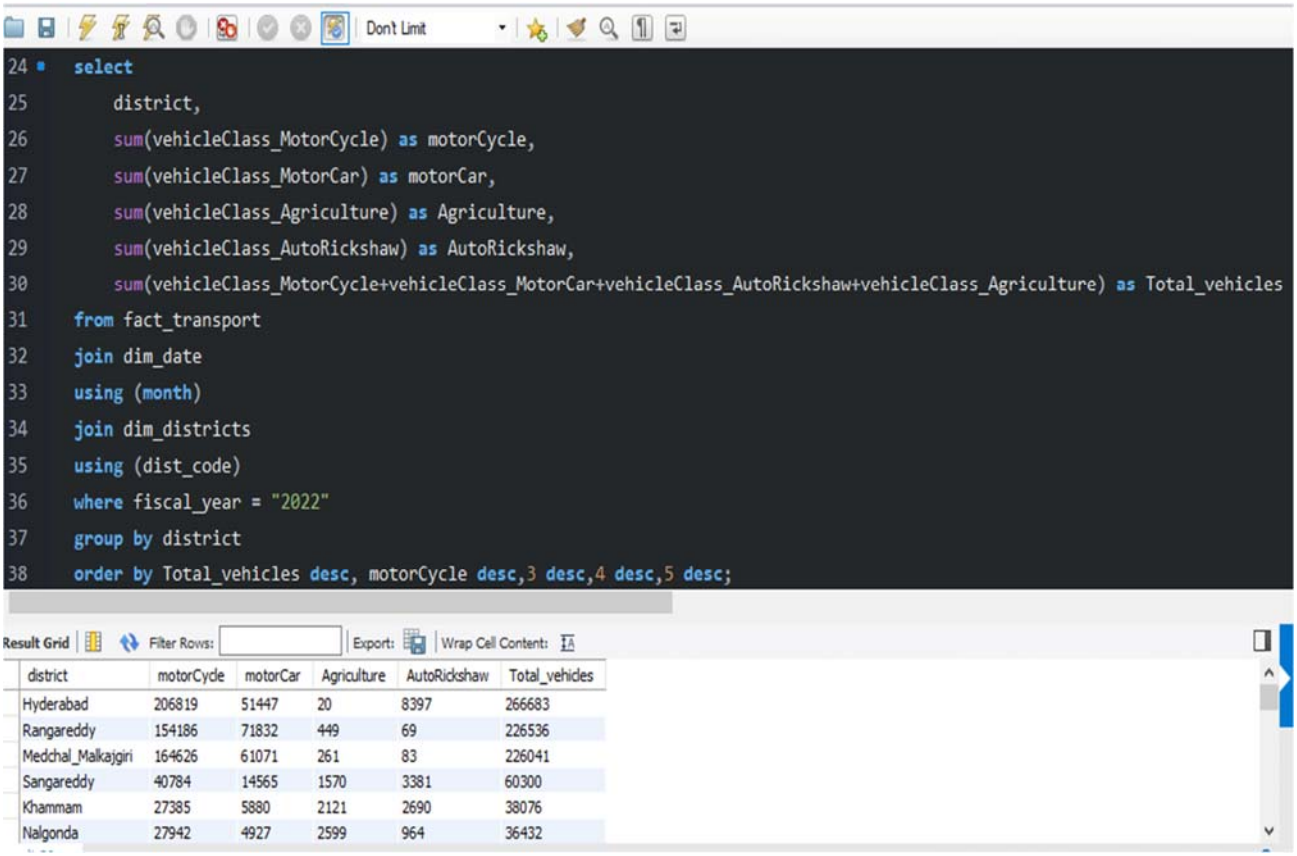
Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

fiscal_year	Mmm	district	vehicle_sale
2019	Apr	Hyderabad	29090
2019	Apr	Medchal_Malkajgiri	22965
2019	Apr	Rangareddy	21019
2019	Apr	Khammam	6553
2019	Apr	Sangareddy	6439
2019	Apr	Warangal	6379
2019	Apr	Nizamabad	5497
2019	Apr	Nalgonda	5414
2019	Apr	Mahabubnagar	5210
2019	Apr	Vikarabad	4521

Visualization:



2. How does the distribution of vehicles vary by vehicle class (MotorCycle, MotorCar, AutoRickshaw, Agriculture) across different districts? Are there any districts with a predominant preference for a specific vehicle class? Consider FY 2022 for analysis.

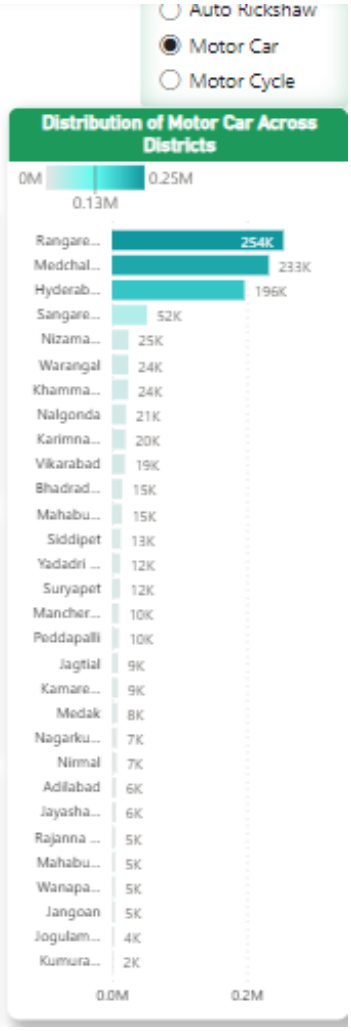


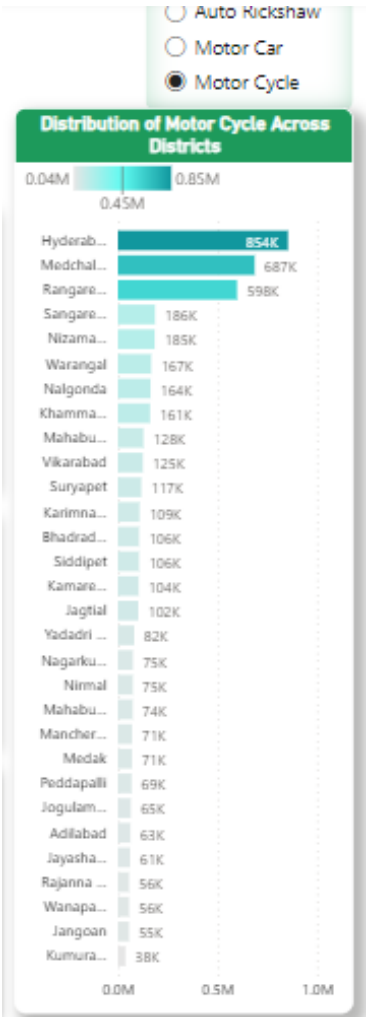
```
24  select
25      district,
26      sum(vehicleClass_MotorCycle) as motorCycle,
27      sum(vehicleClass_MotorCar) as motorCar,
28      sum(vehicleClass_Agriculture) as Agriculture,
29      sum(vehicleClass_AutoRickshaw) as AutoRickshaw,
30      sum(vehicleClass_MotorCycle+vehicleClass_MotorCar+vehicleClass_AutoRickshaw+vehicleClass_Agriculture) as Total_vehicles
31  from fact_transport
32  join dim_date
33  using (month)
34  join dim_districts
35  using (dist_code)
36  where fiscal_year = "2022"
37  group by district
38  order by Total_vehicles desc, motorCycle desc,3 desc,4 desc,5 desc;
```

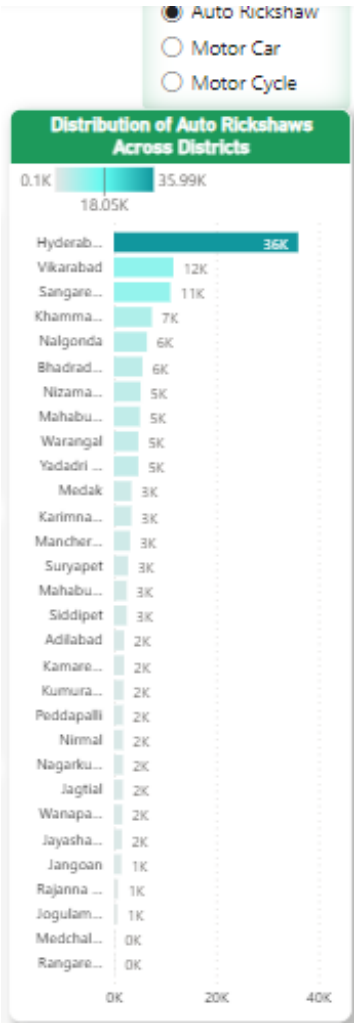
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

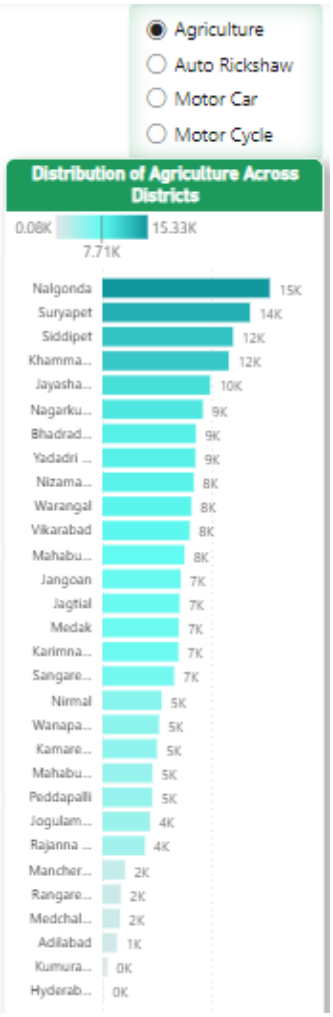
district	motorCycle	motorCar	Agriculture	AutoRickshaw	Total_vehicles
Hyderabad	206819	51447	20	8397	266683
Rangareddy	154186	71832	449	69	226536
Medchal_Malkajgiri	164626	61071	261	83	226041
Sangareddy	40784	14565	1570	3381	60300
Khammam	27385	5880	2121	2690	38076
Nalgonda	27942	4927	2599	964	36432

Visualization:

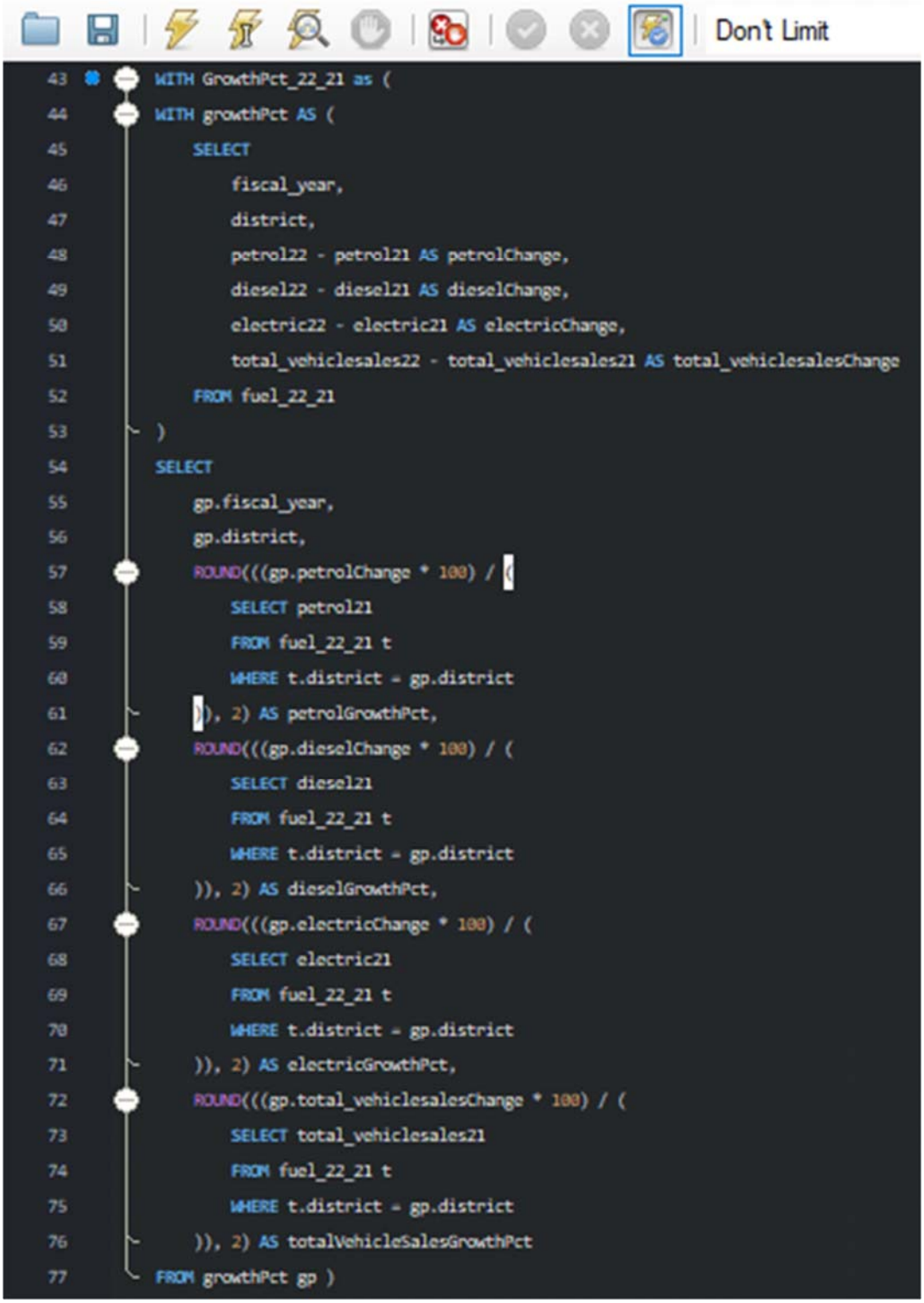








7. List down the top 3 and bottom 3 districts that have shown the highest and lowest vehicle sales growth during FY 2022 compared to FY 2021? (Consider and compare categories: Petrol, Diesel and Electric)



For top 3 districts with highest petrol vehicles,

```
43  WITH GrowthPct_22_21 as (
78
79      -- for top 3 districts with highest petrolGrowthPct
80      select
81          district,
82          petrolGrowthPct
83      from GrowthPct_22_21
84      order by petrolGrowthPct desc
85      limit 3;
86
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

district	petrolGrowthPct
Rangareddy	8.73
Hyderabad	1.03
Medchal_Malkajgiri	-0.90

For bottom 3 districts with lowest petrol vehicles,

```
43  WITH GrowthPct_22_21 as (
78
79      -- for bottom 3 districts with lowest petrolGrowthPct
80      select
81          district,
82          petrolGrowthPct
83      from GrowthPct_22_21
84      order by petrolGrowthPct
85      limit 3;
86
87      -- for top 3 districts with highest petrolGrowthPct
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

district	petrolGrowthPct
Warangal	-45.32
Nirmal	-41.05
Jagtial	-40.53

For top 3 districts with highest diesel vehicles,

```
43  WITH GrowthPct_22_21 as (
78
79      -- for top 3 districts with highest diesel vehicles,
80      select
81          district,
82          dieselGrowthPct
83      from GrowthPct_22_21
84      order by dieselGrowthPct desc
85      limit 3;
86      -- for bottom 3 districts with lowest petrolGrowthPct
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

district	dieselGrowthPct
Karimnagar	111.44
Sangareddy	13.21
Rangareddy	12.67

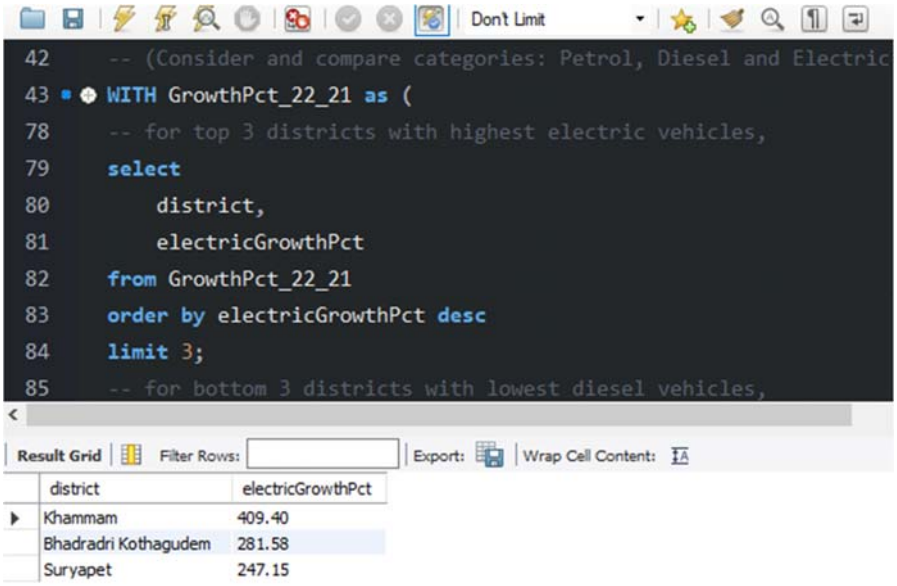
For bottom 3 districts with lowest diesel vehicles,

```
43  WITH GrowthPct_22_21 as (
78
79      -- for bottom 3 districts with lowest diesel vehic
80      select
81          district,
82          dieselGrowthPct
83      from GrowthPct_22_21
84      order by dieselGrowthPct
85      limit 3;
86      -- for top 3 districts with highest diesel vehicle
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

district	dieselGrowthPct
Warangal	-48.03
Mahabubabad	-24.29
Jangoan	-22.00

Top 3 districts with highest electric vehicles,

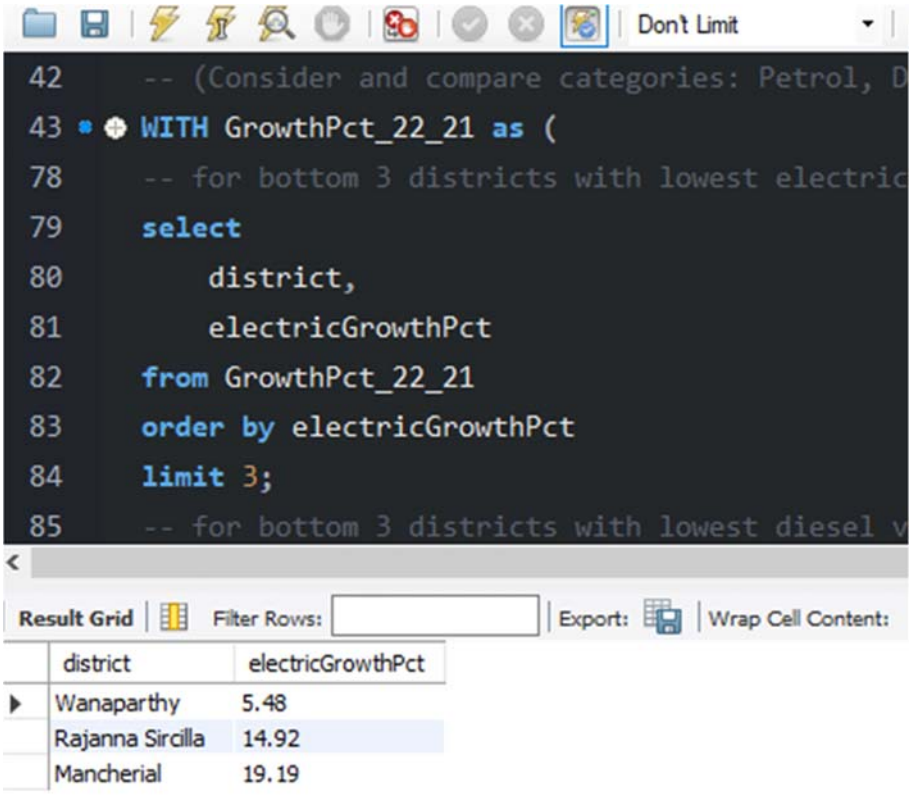


```
42    -- (Consider and compare categories: Petrol, Diesel and Electric
43 •  WITH GrowthPct_22_21 as (
78    -- for top 3 districts with highest electric vehicles,
79    select
80        district,
81        electricGrowthPct
82    from GrowthPct_22_21
83    order by electricGrowthPct desc
84    limit 3;
85    -- for bottom 3 districts with lowest diesel vehicles,
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

district	electricGrowthPct
Khammam	409.40
Bhadradri Kothagudem	281.58
Suryapet	247.15

Bottom 3 districts with lowest electric vehicles,

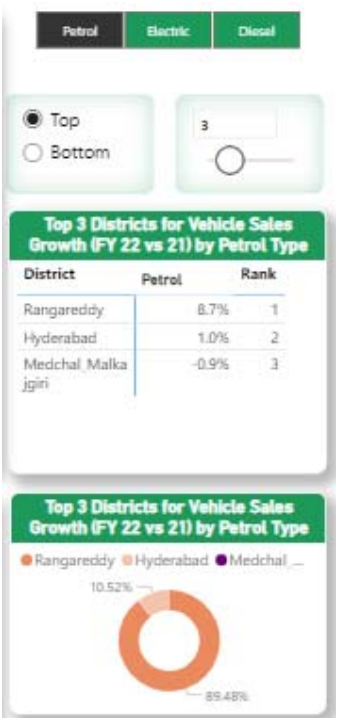


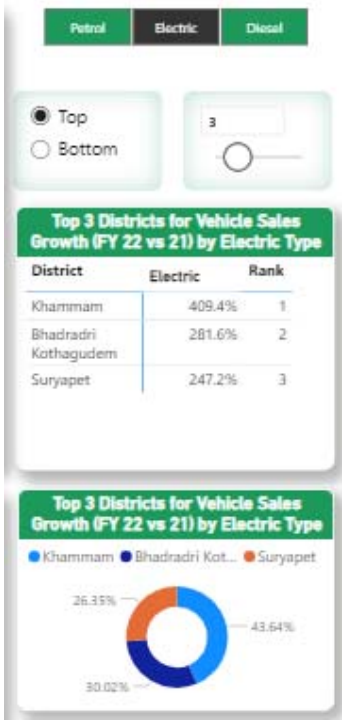
```
42    -- (Consider and compare categories: Petrol, D
43 •  WITH GrowthPct_22_21 as (
78    -- for bottom 3 districts with lowest electric
79    select
80        district,
81        electricGrowthPct
82    from GrowthPct_22_21
83    order by electricGrowthPct
84    limit 3;
85    -- for bottom 3 districts with lowest diesel v
```

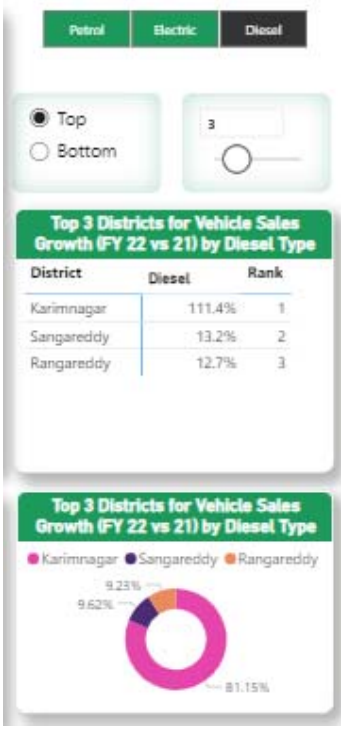
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

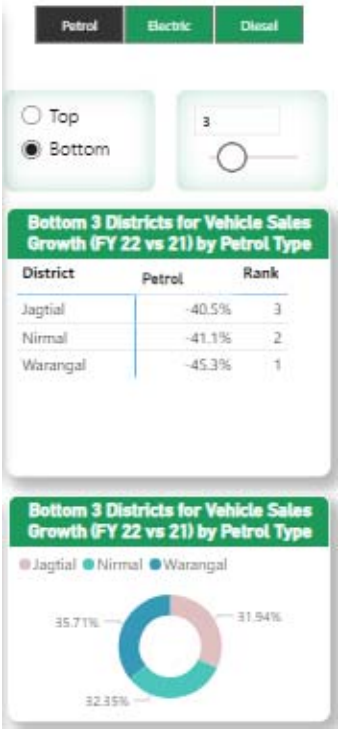
district	electricGrowthPct
Wanaparthy	5.48
Rajanna Sircilla	14.92
Mancherial	19.19

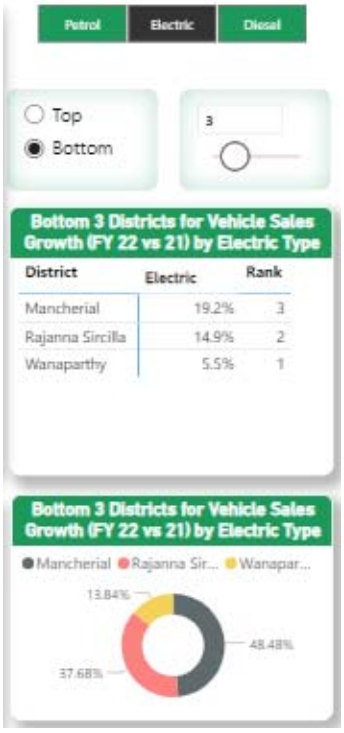
Visualization:

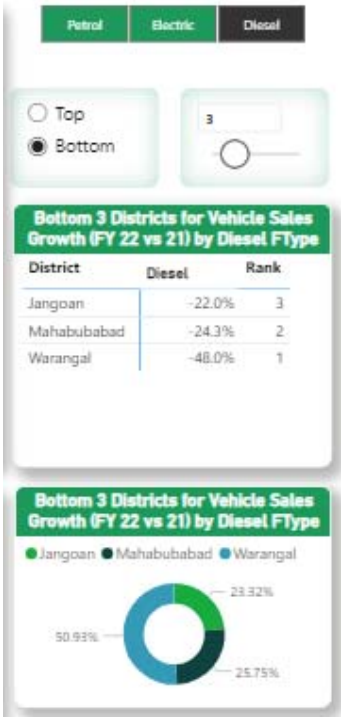












TS-IPASS

1. List down the top 5 sectors that have witnessed the most significant investments in FY 2022.

```
4
5   select
6       sector,
7       round((sum(investment_in_cr) ),2) as total_investment_cr
8   from fact_ts_ipass
9   join dim_date
10  using (month)
11  join dim_districts
12  using (dist_code)
13  where fiscal_year = "2022"
14  group by sector
15  order by total_investment_cr desc
16  limit 5;
```

OUTPUT:

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

sector	total_investment_cr
Plastic and Rubber	5855.61
Pharmaceuticals and Chemicals	2181.63
Real Estate,Industrial Parks and IT Buildings	2127.3
Solar and Other Renewable Energy	2052.98
Engineering	1877.45

Result 26

Visualization:

2. List down the top 3 districts that have attracted the most significant sector investments during FY 2019 to 2022? What factors could have led to the substantial investments in these particular districts?

```sql
With top3_districts as (
    select
        fiscal_year,
        district,
        sector,
        round((sum(investment_in_cr)),2) as total_investment_cr,
        dense_rank() over(partition by fiscal_year order by sum(investment_in_cr) desc) as DRank
    from fact_ts_ipass
    join dim_date
    using (month)
    join dim_districts
    using (dist_code)
    where fiscal_year between '2019' and '2022'
    group by district,1,3
    order by fiscal_year, total_investment_cr desc
    )
   select * from top3_districts
   where DRank <=3;
```

OUTPUT:

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

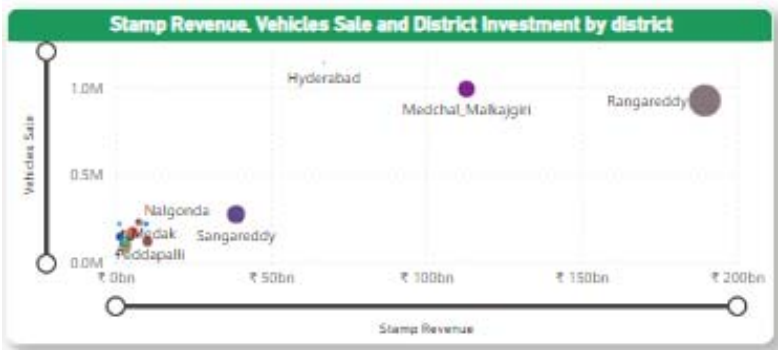
fiscal_year	district	sector	total_investment_cr	DRank
2019	Rangareddy	Real Estate,Industrial Parks and IT Buildings	23686.75	1
2019	Peddapalli	Fertlizers Organic and Inorganic,Pesticides,Inse...	5254.28	2
2019	Bhadradri Kothagudem	Paper and Printing	955.22	3
2020	Medchal_Malkajgiri	Pharmaceuticals and Chemicals	1398.26	1
2020	Rangareddy	Pharmaceuticals and Chemicals	1171.79	2
2020	Rangareddy	Real Estate,Industrial Parks and IT Buildings	1152.17	3
2021	Sangareddy	Pharmaceuticals and Chemicals	2076.91	1
2021	Rangareddy	Real Estate,Industrial Parks and IT Buildings	2014.29	2
2021	Medchal_Malkajgiri	Pharmaceuticals and Chemicals	1620.6	3
2022	Rangareddy	Plastic and Rubber	3819.44	1
2022	Rangareddy	Real Estate,Industrial Parks and IT Buildings	2117.07	2
2022	Sangareddy	Plastic and Rubber	1762.72	3

Visualization:

3. Is there any relationship between district investments, vehicles sales and stamps revenue within the same district between FY 2021 and 2022?

```sql
SELECT
    d.fiscal_year,
    dd.district,
    ROUND(SUM(f.investment_in_cr), 2) AS total_investment_cr,
    SUM(t.vehicleClass_Agriculture + t.vehicleClass_AutoRickshaw + t.vehicleClass_MotorCar + t.vehicleClass_MotorCycle + t.vehicleClass_others) AS total_vehicle_sale,
    SUM(s.documents_registered_rev+ s.estamps_challans_rev) AS total_stamp_rev
FROM fact_ts_ipass AS f
JOIN dim_date AS d USING (month)
JOIN dim_districts AS dd USING (dist_code)
JOIN fact_stamps AS s ON f.month = s.month AND f.dist_code = s.dist_code
JOIN fact_transport t ON t.month = f.month AND f.dist_code = t.dist_code
WHERE d.fiscal_year BETWEEN '2021' AND '2022'
GROUP BY d.fiscal_year, dd.district
ORDER BY d.fiscal_year, total_investment_cr DESC, total_vehicle_sale desc, total_stamp_rev desc;
```

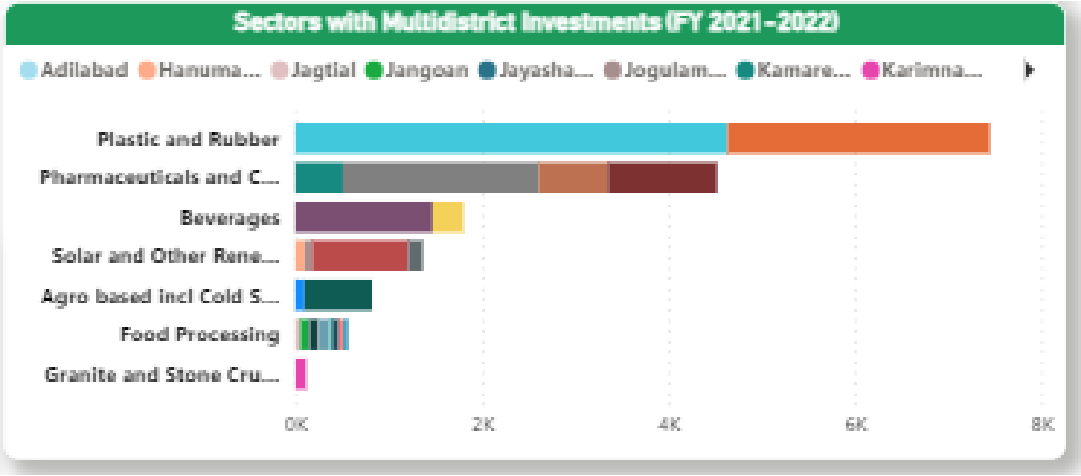
Visualization:



4. Are there any particular sectors that have shown substantial investment in multiple districts between FY 2021 and 2022?

```
select * from maxSector_withDistricts_21_22
where sector in (select sector from
(
    select
        sector,
        count(sector) as sectorCount
    from maxSector_withDistricts_21_22
    group by sector
    having sectorCount <>1
)s);
```

Visualization:



5. Can we identify any seasonal patterns or cyclicality in the investment trends for specific sectors? Do certain sectors experience higher investments during particular months?

```sql
select
    fiscal_year,
    Mmm as month,
    sector,
    round((sum(investment_in_cr)),2) as total_investment
from fact_ts_ipass
join dim_date
using (month)
group by fiscal_year,  month, sector
order by fiscal_year,  month, total_investment desc;
```

Visualization: