

Hive queries

1. **What are the weekday routes with the highest number of tickets validated?(comparing two months to each other)**

```
SELECT r.route_id, SUM(CASE WHEN d.month_no = 4 THEN
t.tickets_validated ELSE 0 END) AS total_tickets_april,
SUM(CASE WHEN d.month_no = 5 THEN t.tickets_validated ELSE 0
END) AS total_tickets_may, (SUM(CASE WHEN d.month_no = 5 THEN
t.tickets_validated ELSE 0 END)-SUM(CASE WHEN d.month_no = 4
THEN t.tickets_validated ELSE 0 END)) AS
tickets_difference_to_prev_month
FROM travel t
JOIN route r ON t.route_id = r.route_id
JOIN date_dim d ON t.travel_date = d.date_id
WHERE d.month_no IN (4, 5) AND d.day_type= 'weekday'
GROUP BY r.route_id
ORDER BY total_tickets_april DESC
LIMIT 10;
```

2. **Which bus type has the highest average satisfaction level?**

```
SELECT b.bus_type, AVG(t.avg_satisfaction_level_received) AS
avg_satisfaction
FROM travel t
JOIN bus b ON t.bus_id = b.bus_id
GROUP BY b.bus_type
ORDER BY avg_satisfaction DESC
LIMIT 1;
```

3. **What is the number of validated tickets for each bus comparing across weekdays and weekends in May?**

```
SELECT t.bus_id, d.day_type, SUM(t.tickets_validated) AS
total_validated_tickets
FROM travel t
JOIN date_dim d ON t.travel_date = d.date_id
WHERE d.month_no = 5 AND d.day_type IN ('weekday', 'weekend')
GROUP BY t.bus_id, d.day_type
ORDER BY t.bus_id, d.day_type;
```

4. What was the busiest time of the day for bus routes on weekends in April?

```
SELECT td.time_of_day, COUNT(*) AS trip_num
FROM travel t
JOIN date_dim d ON t.travel_date = d.date_id
JOIN time_dim td ON t.departure_time = td.time_id
WHERE month_no = 4 AND d.day_type = 'weekend'
GROUP BY td.time_of_day
ORDER BY trip_num DESC;
```

5. How many satisfaction surveys were received for a given route comparing two months (April and May)?

```
SELECT r.route_name, SUM(CASE WHEN d.month_no = 4 THEN
t.satisfaction_surveys_number ELSE 0 END) as surveys_april,
SUM(CASE WHEN d.month_no = 5 THEN
t.satisfaction_surveys_number ELSE 0 END) as surveys_may
FROM travel t
JOIN route r ON t.route_id = r.route_id
JOIN date_dim d ON t.travel_date = d.date_id
GROUP BY r.route_name
ORDER BY surveys_april DESC
LIMIT 5;
```

6. How satisfaction level of bus users changed over time?

```
SELECT d.month, AVG(t.avg_satisfaction_level_received) AS
avg_satisfaction_level
FROM travel t
JOIN date_dim d ON t.travel_date = d.date_id
GROUP BY d.month
ORDER BY d.month;
```

7. How does the satisfaction level vary based on occupation level categories?

```
SELECT j.occupation_level_category,
AVG(avg_satisfaction_level_received) AS avg_satisfaction_level
FROM travel t
JOIN junk j ON t.junk_id=j.junk_id
GROUP BY j.occupation_level_category
ORDER BY avg_satisfaction_level DESC;
```

8. What is the breakdown of travel counts by bus type that departures in the morning?

```
SELECT b.bus_type, COUNT(*) travels_num
FROM travel t
JOIN bus b ON t.bus_id=b.bus_id
JOIN time_dim tdim ON t.departure_time=tdim.time_id
WHERE tdim.time_of_day = 'Morning'
GROUP BY b.bus_type, tdim.time_of_day
ORDER BY travels_num DESC;
```

9. Which are the routes with the highest distance value?

```
SELECT route_id, route_name, metrics['distance_km'] AS
distance
FROM route
ORDER BY distance DESC
LIMIT 10;
```

10. What is the satisfaction level on weekend routes that have the highest avg duration? (in the selected month - April)

```
SELECT t.route_id, r.route_name, AVG(r.metrics['duration_min'])
as avg_duration,
AVG(t.avg_satisfaction_level_received) AS avg_satisfaction
FROM travel t
JOIN route r ON t.route_id = r.route_id
JOIN date_dim d ON t.travel_date = d.date_id
WHERE d.month_no = 4 AND d.day_type = 'weekend'
GROUP BY t.route_id, r.route_name
ORDER BY avg_duration DESC
LIMIT 10;
```