1. Preprocessing Data

1.1. Outlier identification SQL:

1.1.1. NULL records:

```
SELECT
COUNT(*) total_null_records
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE
 unique_key IS NULL
 OR taxi_id IS NULL
 OR trip_start_timestamp IS NULL
 OR trip_start_timestamp IS NULL
 OR trip_end_timestamp IS NULL
 OR trip_seconds IS NULL
 -- OR trip_seconds <> 0 -- Outlier
 OR trip_miles IS NULL
 -- OR trip_miles <> 0 -- Outlier
 OR pickup_census_tract IS NULL
 OR dropoff_census_tract IS NULL
 OR pickup_community_area IS NULL
 OR dropoff_community_area IS NULL
 OR fare IS NULL
 -- OR fare <> 0 -- Outlier
 OR tips IS NULL
 OR tolls IS NULL
 OR extras IS NULL
 OR trip_total IS NULL
 -- OR trip_total <> 0 -- Outlier
 OR payment_type IS NULL
 OR company IS NULL
 OR pickup_latitude IS NULL
 OR pickup_longitude IS NULL
 OR pickup_location IS NULL
 OR dropoff_latitude IS NULL
 OR dropoff_longitude IS NULL
 OR dropoff_location IS NULL
```

1.1.2. 0 Value SQL:

```
SELECT
COUNT(*) total_null_records_trip_total
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE
```

```
 -- trip_seconds = 0
 -- trip_miles = 0
 -- fare = 0
 trip_total = 0
```

1.1.3. Timestamp differences:

```
SELECT
COUNT(*) total_diff_val
FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
WHERE TRUE
AND TIMESTAMP_DIFF(trip_end_timestamp, trip_start_timestamp, SECOND) = trip_seconds
```

1.1.4. Top 10 Companies SQL:

```
WITH unfiltered_dataset AS
(
 SELECT *
 FROM `bigquery-public-data.chicago_taxi_trips.taxi_trips`
 WHERE
   unique_key IS NOT NULL
   AND taxi_id IS NOT NULL
   AND trip_start_timestamp IS NOT NULL
   AND trip_start_timestamp IS NOT NULL
   AND trip_end_timestamp IS NOT NULL
   AND trip_seconds IS NOT NULL
   AND trip_seconds <> 0 -- Outlier
   AND trip_miles IS NOT NULL
   AND trip_miles <> 0 -- Outlier
   AND pickup_census_tract IS NOT NULL
   AND dropoff_census_tract IS NOT NULL
   AND pickup_community_area IS NOT NULL
   AND dropoff_community_area IS NOT NULL
   AND fare IS NOT NULL
   AND fare <> 0 -- Outlier
   AND tips IS NOT NULL
   AND tolls IS NOT NULL
   AND extras IS NOT NULL
   AND trip_total IS NOT NULL
   AND trip_total <> 0 -- Outlier
   AND payment_type IS NOT NULL
   AND company IS NOT NULL
   AND pickup_latitude IS NOT NULL
   AND pickup_longitude IS NOT NULL
   AND pickup_location IS NOT NULL
   AND dropoff_latitude IS NOT NULL
   AND dropoff_longitude IS NOT NULL
   AND dropoff_location IS NOT NULL
)

 SELECT
 company,
 COUNT(unique_key) cnt_trip
 FROM unfiltered_dataset
 GROUP BY 1
 order by 2 desc
 limit 10
```

1.1.5. Cleaned SQL Results:

```
WITH unfiltered_dataset AS
(
 SELECT *
 FROM bigquery-public-data.chicago_taxi_trips.taxi_trips
 WHERE
   unique_key IS NOT NULL
   AND taxi_id IS NOT NULL
   AND trip_start_timestamp IS NOT NULL
   AND trip_start_timestamp IS NOT NULL
   AND trip_end_timestamp IS NOT NULL
   AND trip_seconds IS NOT NULL
   AND trip_seconds <> 0 -- Outlier
   AND trip_miles IS NOT NULL
   AND trip_miles <> 0 -- Outlier
   AND pickup_census_tract IS NOT NULL
   AND dropoff_census_tract IS NOT NULL
   AND pickup_community_area IS NOT NULL
   AND dropoff_community_area IS NOT NULL
   AND fare IS NOT NULL
   AND fare <> 0 -- Outlier
   AND tips IS NOT NULL
   AND tolls IS NOT NULL
   AND extras IS NOT NULL
   AND trip_total IS NOT NULL
   AND trip_total <> 0 -- Outlier
   AND payment_type IS NOT NULL
   AND company IS NOT NULL
   AND pickup_latitude IS NOT NULL
   AND pickup_longitude IS NOT NULL
   AND pickup_location IS NOT NULL
   AND dropoff_latitude IS NOT NULL
   AND dropoff_longitude IS NOT NULL
   AND dropoff_location IS NOT NULL
    AND TIMESTAMP_DIFF(trip_end_timestamp, trip_start_timestamp, SECOND) =
trip_seconds
)
, unfiltered_dataset_2 AS
(
 SELECT
 company,
 COUNT(unique_key) cnt_trip
 FROM unfiltered_dataset
 GROUP BY 1
 order by 2 desc
 limit 10
)


-- Main Query
SELECT *
FROM unfiltered_dataset
WHERE company IN
(
 SELECT company FROM unfiltered_dataset_2
)
```