Now starting analysis of cleaned_data

Questions to be answered to see difference in usage patterns between members and casuals.

A. Based on total count in a year

   1. How many rides members v/s casuals - total?
   2. Is there any preference in rideable_type between members and casuals?

B. Based on time
3. Is there any difference in ride duration i.e. start time - end time between members v/s casuals?
3.1. Checking for new metric (Number of rides) x (Ride duration)
4. Is there a difference in start time/end time/ ride duration (i.e. usage) based on time of the day between members and casuals? + check total
5. Is there a difference in start time/end time/ ride duration (i.e. usage) based on day of week between members and casuals? + check total
6. Is there a difference in start time/end time/ ride duration (i.e. usage) based on month of the year between members and casuals? + check total (season)

C. Based on distance (check overlay with time questions)
7. Is there a difference in distance travelled: start_station_id and end_station_id between members v/s casuals?

D. Based on location (check overlay with time questions)
8. Is there a difference in preference for start station id between members and casuals?
9. Is there a difference in preference of choice for end station id between members v/s casuals?

SQL codes used.

A.1.
To find out total ride count and ride duration - members v/s casuals.

```
WITH groupin AS (
        SELECT member_casual,count(*) AS [Number of rides]
        FROM [dbo].[Yearly_Clean]
        GROUP BY member_casual
), maximum AS (
        SELECT count(*) AS maxi
        FROM [dbo].[Yearly_Clean])
SELECT --Using CTE
        member_casual
        , [Number of rides]
        , CAST((100.0 * [Number of rides] / maxi) AS decimal(5,2)) AS [Percentage]
FROM groupin,maximum;

-- Using alternative window function
SELECT member_casual,100.0*count(*)/sum(count(*)) OVER ()
FROM [dbo].[Yearly_Clean]
GROUP BY member_casual
```

A.2.
Is there any preference in rideable_type between members and casuals?

```
SELECT
        member_casual [Type of customer]
        , rideable_type [Type of bike]
        , count(*) AS [Number of type rides]
```

```sql
        , sum(count(*)) OVER (PARTITION BY member_casual) AS [Category Total]
        , cast(100.0 * count(*)/ sum(count(*)) OVER (PARTITION BY member_casual) AS
decimal(10,2))AS [Percentage]
FROM [dbo].[Yearly_Clean]
GROUP BY member_casual, rideable_type;


-- Check for ride duration difference by biketype
SELECT
        rideable_type
        ,count(*) AS [Number of rides]
        , avg(datediff(MINUTE,started_at,ended_at)) AS [Avg duration per ride]
        , sum(datediff(MINUTE,started_at,ended_at)) AS [Total time duration]
FROM Yearly_Clean
GROUP BY rideable_type;
```

B.3.

To find ride duration- members v/s casuals

```sql
SELECT
        member_casual [Type of customer]
        , sum(datediff(MINUTE,started_at,ended_at)) AS [Total ride duration in minutes]
FROM
        [dbo].[Yearly_Clean]
GROUP BY member_casual;
```

Checking the range of ride duration among members / casuals.

```sql
SELECT -- Average
        member_casual AS [Type of customer]
        , avg(datediff(MINUTE,started_at,ended_at)) AS [Avg Ride duration in minutes]
        , max(datediff(MINUTE,started_at,ended_at)) AS [Max Ride duration]
        , min(datediff(MINUTE,started_at,ended_at)) AS [Min Ride duration]
FROM [Yearly_Clean]
WHERE datediff(MINUTE,started_at,ended_at) >0
GROUP BY member_casual;
```

B.3.1 Checking new metric Frequency of rides x Ride duration varied by hour, week, month.

```sql
-- Ride duration frequency
SELECT -- based on number of rides with ride duration
        member_casual AS [Type of customer]
        , datediff(MINUTE,started_at,ended_at) AS [Ride duration]
        , count(*) [Number of rides]
FROM [Yearly_Clean]
WHERE datediff(MINUTE,started_at,ended_at)>0
GROUP BY member_casual,datediff(MINUTE,started_at,ended_at)
ORDER BY 2;
-- 1660 rows

SELECT -- based on number of rides + ride duration on every --hour of day--
        member_casual AS [Type of customer]
        , datediff(MINUTE,started_at,ended_at) AS [Ride duration]
        , count(*) AS [Number of rides]
        , hourly AS [Hour of day]
FROM [dbo].[Yearly_Clean]
WHERE datediff(MINUTE,started_at,ended_at)>0
GROUP BY member_casual, datediff(MINUTE,started_at,ended_at),hourly;
```

```
SELECT -- based on number of rides + ride duration on every day of week
        member_casual AS [Type of customer]
        , datediff(MINUTE,started_at,ended_at) AS [Ride duration]
        , count(*) AS [Number of rides]
        , weekday AS [Day of week]
FROM [dbo].[Yearly_Clean]
WHERE datediff(MINUTE,started_at,ended_at)>0
GROUP BY member_casual, datediff(MINUTE,started_at,ended_at),weekday;

SELECT -- based on number of rides + ride duration on every month of year
        member_casual AS [Type of customer]
        , datediff(MINUTE,started_at,ended_at) AS [Ride duration]
        , count(*) AS [Number of rides]
        , months AS [Month of Year]
FROM [dbo].[Yearly_Clean]
WHERE datediff(MINUTE,started_at,ended_at)>0
GROUP BY member_casual, datediff(MINUTE,started_at,ended_at),months;
```

Each output can be exported to an excel file for visualizations in Tableau.

B.4.

Is there a difference in start time & end time (i.e. usage) based on time of the day between members and casuals?
+ check total

```
--Based on sum(duration)
SELECT
        member_casual [Type of customer]
        , sum(datediff(MINUTE,started_at,ended_at)) AS [Ride duration]
FROM [dbo].[Yearly_Clean]
GROUP BY member_casual;
--Based on total for start time.

/* Check for distribution for start times without division- TOTAL*/
SELECT
        datepart(HH,started_at) AS [Hour of the day]
        , count(*) AS [Number of rides]
        , sum(count(*)) OVER () [Total rides]
        , cast(100.0 *count(*)/(sum(count(*)) OVER ()) AS decimal(10,2)) AS
[Percent]
FROM [Yearly_Clean]
GROUP BY datepart(HH,started_at)
ORDER BY 1;
-- Peak time is the evening time.

/* Distribution of start times by casual/member */
SELECT
        member_casual
        , datepart(HH,started_at) AS [Hr of day]
        , count(*) AS [Number of rides]
        , sum(count(*)) OVER (PARTITION BY member_casual) [Category total]-
-2,679,660
        , cast(100.0*count(*)/(sum(count(*)) OVER (PARTITION BY member_casual)) AS
decimal(10,2)) AS [Percent]
FROM [Yearly_Clean]
GROUP BY datepart(HH,started_at), member_casual
ORDER BY 1,2;
-- 48 rows

-- Based on total for end time.
/* Check for distribution for end times without division- TOTAL*/
```

```
        SELECT
                datepart(HH,ended_at)
                , count(*) AS [Number of rides]
                , sum(count(*)) OVER () [Total rides]
                , cast(100.0 *count(*)/(sum(count(*)) OVER ()) AS decimal(10,2)) AS
[Percent]
        FROM [Yearly_Clean]
        GROUP BY datepart(HH,ended_at)
        ORDER BY 1;
        --Peak time is evening time at 6 pm..

        /* Distribution of end times by casual/member */
        SELECT
                member_casual AS [Type of customer]
                , datepart(HH,ended_at) AS [Hr of day]
                , count(*) AS [Number of rides]
                , sum(count(*)) OVER (PARTITION BY member_casual) AS [Category Total]-
-1,947,932
                , cast(100.0 *count(*)/(sum(count(*)) OVER (PARTITION BY member_casual)) AS
decimal(10,2)) AS [Percent]
        FROM [Yearly_Clean]
        GROUP BY datepart(HH,ended_at), member_casual
        ORDER BY 1,2;
```

B.5.

Is there a difference based on day of week

```
--Total
SELECT
        datename(WEEKDAY,started_at) AS [Day of week]
        , count(*) AS [Number of rides]
        , sum(count(*)) OVER () [Total rides]
        , cast(100.0 *count(*)/(sum(count(*)) OVER ()) AS decimal(10,2)) AS [Percent]
FROM [Yearly_Clean]
GROUP BY datename(WEEKDAY,started_at)
ORDER BY
        CASE
                WHEN Datename(WEEKDAY,started_at) = 'Monday'    THEN 1
                WHEN Datename(WEEKDAY,started_at) = 'Tuesday'   THEN 2
                WHEN Datename(WEEKDAY,started_at) = 'Wednesday' THEN 3
                WHEN Datename(WEEKDAY,started_at) = 'Thursday'  THEN 4
                WHEN Datename(WEEKDAY,started_at) = 'Friday'    THEN 5
                WHEN Datename(WEEKDAY,started_at) = 'Saturday'  THEN 6
                WHEN Datename(WEEKDAY,started_at) = 'Sunday'    THEN 7
        END;
-- Maximum rides are on Saturdays and Sundays.

-- Weekday activity based on member/ casual type
SELECT
        member_casual AS [Type of customer]
        , datename(WEEKDAY,started_at) AS [Day of week]
        , count(*) AS [Number of rides]
        , sum(count(*)) OVER () [Total rides]
        , cast(100.0 *count(*)/(sum(count(*)) OVER (PARTITION BY member_casual)) AS
decimal(10,2)) AS [Percent]
FROM [Yearly_Clean]
GROUP BY member_casual, datename(WEEKDAY,started_at)
ORDER BY
        1,CASE
                WHEN Datename(WEEKDAY,started_at) = 'Monday'    THEN 1
                WHEN Datename(WEEKDAY,started_at) = 'Tuesday'   THEN 2
```

```
                WHEN Datename(WEEKDAY,started_at) = 'Wednesday' THEN 3
                WHEN Datename(WEEKDAY,started_at) = 'Thursday'  THEN 4
                WHEN Datename(WEEKDAY,started_at) = 'Friday'    THEN 5
                WHEN Datename(WEEKDAY,started_at) = 'Saturday'  THEN 6
                WHEN Datename(WEEKDAY,started_at) = 'Sunday'    THEN 7
        END;
```

B.6. Based on the month of the year.

```
    SELECT
            member_casual AS [Type of customer]
            , datename(MONTH,started_at) AS [Month of year]
            , count(*) AS [Number of rides]
            , sum(count(*)) OVER (PARTITION BY member_casual) [Total rides]
            , cast(100.0 *count(*)/(sum(count(*)) OVER (PARTITION BY member_casual)) AS
decimal(10,2)) AS [Percent]
        FROM [Yearly_Clean]
        GROUP BY member_casual, datename(MONTH,started_at)
        ORDER BY 1
            , CASE
                    WHEN datename(MONTH,started_at) = 'January' THEN 1
                    WHEN datename(MONTH,started_at) = 'February' THEN 2
                    WHEN datename(MONTH,started_at) = 'March' THEN 3
                    WHEN datename(MONTH,started_at) = 'April' THEN 4
                    WHEN datename(MONTH,started_at) = 'May' THEN 5
                    WHEN datename(MONTH,started_at) = 'June' THEN 6
                    WHEN datename(MONTH,started_at) = 'July' THEN 7
                    WHEN datename(MONTH,started_at) = 'August' THEN 8
                    WHEN datename(MONTH,started_at) = 'September' THEN 9
                    WHEN datename(MONTH,started_at) = 'October' THEN 10
                    WHEN datename(MONTH,started_at) = 'November' THEN 11
                    WHEN datename(MONTH,started_at) = 'December' THEN 12
            END;
```

C.7. Distance travelled between start station and stop station

```
--Grouping by member / casual
    SELECT
            member_casual AS [Type of customer]
            , cast(avg(1.609344*(3963.0*

acos(sin(s_lat_r/57.29577951)*sin(e_lat_r/57.29577951) +
                                cos(s_lat_r/57.29577951)*cos(e_lat_r/57.29577951)*
                                cos((e_lng_r/57.29577951)-(s_lng_r/57.29577951)))))
AS decimal(10,2)) AS average_km
        FROM [dbo].[Yearly_Clean]
        WHERE start_station_id != end_station_id
        GROUP BY member_casual;
```

The average distance between start and stop station is ranging between 2.99 and 3.34 for members and casuals respectively. This difference is not significant, hence not considered as a difference.

D.8 & D.9 Preference for start station and end station between members and casuals.

```
--Top start stations for members (Chosen)
    SELECT
            member_casual
```

```
              , s_name_r AS [Start station name]
              , s_lat_r [Latitude]
              , s_lng_r [Longitude]
              , COUNT(*) AS [Number of rides]
              , cast(100.0 * count(*)/(sum(count(*)) OVER (PARTITION BY member_casual)) AS
  decimal(10,2)) AS [Percent]
        FROM [dbo].[Yearly_Clean]
        GROUP BY member_casual, s_name_r, s_lat_r, s_lng_r
        ORDER BY 1, [Number of rides] DESC;
```

It was seen that there is not much variance based on start and end station particularly. Hence, start station data is considered for overall popularity of the station.

-- FINISH --