Bellabeat Data Cleaning and Manupulation in SQL

```
#  Data cleaning

 /*checking for how many users in the data - dailyActivity,
     30 users were specified in datasource but actual are 33
*/

SELECT distinct(ID) FROM
`bellabeatproject-363419.bellabeat.dailyActivity` LIMIT 100

#how many distinct users in dailyCalories

select distinct(ID) from
`bellabeatproject-363419.bellabeat.dailyCalories` limit 100

#output 33 users

#how many distinct users in minuteMET

select distinct(ID) from
`bellabeatproject-363419.bellabeat.minuteMET` limit 100

#output 33 users


#how many distinct users in sleepDay
```

```sql
select distinct(ID) from
`bellabeatproject-363419.bellabeat.sleepDay` limit 100

#output 24 users

#how many distinct users in weightBMI

select distinct(ID) from
`bellabeatproject-363419.bellabeat.weightBMI` limit 100

#output 8 users are have their weight record


--data is collected over 30 days

select max(ActivityDate)- min(ActivityDate)
    from `bellabeatproject-363419.bellabeat.dailyActivity`

select max(ActivityDate),min(ActivityDate)
from `bellabeatproject-363419.bellabeat.dailyActivity`


#checking integrity of DATA in dailyActivity Table
# Check Total distance if it is correct

select TotalDistance,
```

```
    round((SedentaryActiveDistance + LightActiveDistance +
    ModeratelyActiveDistance + VeryActiveDistance),2) as
    check_total
    from `bellabeatproject-363419.bellabeat.dailyActivity`

#output -  data is correct , used round() function to show
only 2 decimal place

--Data Exploration

/* Relation between Heart rate and steps */
select distinct h.Id,
  round(avg(h.value),2) as avg_heartRate,
  round(avg(d.TotalSteps),2) as avg_steps,
  round(avg(d.VeryActiveDistance),2) as
avg_veryActiveDistance,

round(avg(d.VeryActiveMinutes+d.FairlyActiveMinutes+d.Lightl
yActiveMinutes),2) as avg_TotalActiveMinutes
  from bellabeat.heartRate h
  inner join bellabeat.dailyActivity d
  on h.Id = d.Id
  group by h.Id
  order by h.Id

#saved the result of this query in CSV format
```

```
/* BMI and weight vs totalsteps, totalDistance, ActiveMins,
Calories */

select distinct w.Id,
  round(avg(w.WeightKg),2) as avg_Weight_kg,
  round(avg(w.BMI),2) as avg_BMI,
  round(avg(d.TotalSteps),2) as avg_totalSteps,
  round(avg(d.TotalDistance),2) as avg_totalDistance,
  round(avg(d.VeryActiveMinutes),2) as avg_veryActiveMins,
  round(avg(d.Calories),2) as avg_calories
  from bellabeat.weightBMI w
  inner join bellabeat.dailyActivity d
  on w.Id = d.Id
  group by w.Id
  order by w.Id
  #output - saved the result in CSV format

-- write a query to extract weekday from the date
--then save results in new table for further calculations
table name - weekday_data


select Id,ActivityDate,
TotalDistance,TotalSteps,VeryActiveDistance,
VeryActiveMinutes,Calories,extract(dayofweek from
ActivityDate) as weekday_number,
  case
```

```
    when extract(dayofweek from ActivityDate) = 1 then
'Sunday'
    when extract(dayofweek from ActivityDate) = 2 then
'Monday'
    when extract(dayofweek from ActivityDate) = 3 then
'Tuesday'
    when extract(dayofweek from ActivityDate) = 4 then
'Wednesday'
    when extract(dayofweek from ActivityDate) = 5 then
'Thursday'
    when extract(dayofweek from ActivityDate) = 6 then
'Friday'
    when extract(dayofweek from ActivityDate) = 7 then
'Saturday'
    else 'Invalid Input'
  end as weekday
from `bellabeatproject-363419.bellabeat.dailyActivity`


 --check if users are more active on weekends
 --save results in new summary table avg_weekends_activity

select round(avg(TotalSteps),2) as avg_TotalSteps,
round(avg(Calories),2) as avg_Calories,
  round(avg(TotalDistance),2) as avg_TotalDistance,
  round(avg(VeryActiveDistance),2) as
avg_VeryActiveDistance,
```

```
  round(avg(VeryActiveMinutes),2) as avg_VeryActiveMinutes,
weekday
  from `bellabeatproject-363419.bellabeat.weekday_data`
  where weekday_number= 1 or weekday_number =7
  group by weekday

--check if users are more active on weekdays
--save results in new summary table avg_weekdays_activity

select round(avg(TotalSteps),2) as avg_TotalSteps,
round(avg(Calories),2) as avg_Calories,
  round(avg(TotalDistance),2) as avg_TotalDistance,
  round(avg(VeryActiveDistance),2) as
avg_VeryActiveDistance,
  round(avg(VeryActiveMinutes),2) as
avg_VeryActiveMinutes,weekday
  from `bellabeatproject-363419.bellabeat.weekday_data`
  where weekday_number in (2,3,4,5,6)
  group by weekday


--check if users have more sedentary minutes on weekends
--save results in avg_sedentary_min_weekends

select round(avg(SedentaryMinutes)) as avg_sedentaryMinutes,
  extract(dayofweek from ActivityDate) as num_of_day
  from bellabeat.dailyActivity
  where (extract(dayofweek from ActivityDate)) in (1,7)
```

```
  group by num_of_day

--check if users have more sedentary minutes on
weekdays,display maximun sedentary minutes at top
--store results in avg_sedentaryMins_weekday

select round(avg(SedentaryMinutes)) as
avg_sedentaryMinutes,(extract(dayofweek from ActivityDate))
as num
  from bellabeat.dailyActivity
  where (extract(dayofweek from ActivityDate)) in
(2,3,4,5,6)
  group by num
  order by avg(SedentaryMinutes) desc


  /* average of minuteMET
  MET stands for Metabolic equivalent of task , amount of
energy used
  save results in avg_MET*/

  select Id,avg(METs),
  case
    when extract(time from ActivityMinute) between
'06:00:00' and '12:00:00' then 'Morning'
    when extract(time from ActivityMinute) between
'12:00:00' and '18:00:00' then 'Afternoon'
```

```
    when extract(time from ActivityMinute) between
'00:00:00' and '06:00:00' then 'Night'
    else 'Evening'
  end as time_of_day
from bellabeat.minuteMET
group by time_of_day,Id


--total minutes in bed vs total sleep time
--store result in avg_sleep

select id,round(avg(TotalMinutesAsleep),2) as
avg_TotalminsAsleep,
   round(avg(TotalTimeInBed),2) as avg_TotalTimeInBed
   from `bellabeatproject-363419.bellabeat.sleepDay`
  group by Id
  order by Id


--average calories according to the time of day
--save in avg_cal

select round(avg(Calories),2) as avg_Calories,
case
    when extract(time from ActivityHour) between '06:00:00'
and '12:00:00' then 'Morning'
    when extract(time from ActivityHour) between '12:00:00'
and '18:00:00' then 'Afternoon'
```

```
    when extract(time from ActivityHour) between '00:00:00'
and '06:00:00' then 'Night'
    else 'Evening'
  end as time_of_day
  from bellabeat.hourlyCalories
group by time_of_day

-- average hourly calories according to the time of day and
group by users ID
--save output in avg_hourly_cal_groupby_id

select Id,round(avg(Calories),2) as avg_Calories,
case
    when extract(time from ActivityHour) between '06:00:00'
and '12:00:00' then 'Morning'
    when extract(time from ActivityHour) between '12:00:00'
and '18:00:00' then 'Afternoon'
    when extract(time from ActivityHour) between '00:00:00'
and '06:00:00' then 'Night'
    else 'Evening'
  end as time_of_day
  from bellabeat.hourlyCalories
group by time_of_day, Id
order by Id, avg_Calories desc

--averagae hourly steps
--store result in avg_hourly_totalsteps
```

```sql
select round(avg(StepTotal),2) as avg_TotalSteps,
case
    when extract(time from ActivityHour) between '06:00:00'
and '12:00:00' then 'Morning'
    when extract(time from ActivityHour) between '12:00:00'
and '18:00:00' then 'Afternoon'
    when extract(time from ActivityHour) between '00:00:00'
and '06:00:00' then 'Night'
    else 'Evening'
  end as time_of_day
  from `bellabeatproject-363419.bellabeat.hourlySteps`
  group by time_of_day

--average total steps group by ID
--store result in avg_hourly_totalSteps_groupby_ID

select Id,round(avg(StepTotal),2) as avg_TotalSteps,
case
    when extract(time from ActivityHour) between '06:00:00'
and '12:00:00' then 'Morning'
    when extract(time from ActivityHour) between '12:00:00'
and '18:00:00' then 'Afternoon'
    when extract(time from ActivityHour) between '00:00:00'
and '06:00:00' then 'Night'
    else 'Evening'
  end as time_of_day
  from `bellabeatproject-363419.bellabeat.hourlySteps`
  group by Id,time_of_day
```

```
  order by Id, avg_TotalSteps desc

-- Summary from dailyActivity Table
--save results in avg_dailyActivity

select Id,round(avg(TotalSteps),2) as avg_TotalSteps,
round(avg(TotalDistance),2) as
avg_TotalDistance,round(avg(Calories),2) as avg_Calories,
round(avg(SedentaryMinutes),2) as avg_SedentaryMins
from bellabeat.dailyActivity
group by Id


-- summary from dailyCalories table
-- store results in avg_dailyCalories

select Id, round(avg(Calories),2) as avg_dailyCalories
from `bellabeatproject-363419.bellabeat.dailyCalories`
group by Id



-- summary of dailyIntensities Table
-- store result in avg_DailyIntensity

select _ID,
  avg(SedentaryMinutes) as avg_SedentaryMinutes,
  avg(LightlyActiveMinutes) as avg_LightlyActiveMinutes,
```

```
  avg(FairlyActiveMinutes) as avg_FairlyActiveMinutes,
  avg(VeryActiveMinutes) as avg_VeryActiveMinutes,
  avg(SedentaryActiveDistance) as
avg_SedentaryActiveDistance,
  avg(LightActiveDistance) as avg_LightActiveDistance,
  avg(ModeratelyActiveDistance) as
avg_ModeratActiveDistance,
  avg(VeryActiveDistance) as avg_veryActiveDistance
  from `bellabeatproject-363419.bellabeat.dailyIntensities`
group by _ID
order by _ID


--summary of dailySteps table
--save in avg_dailyTotalSteps

select Id,round(avg(StepTotal),2) as avg_TotalSteps
from `bellabeatproject-363419.bellabeat.dailySteps`
group by Id
order by Id

-- Average total steps of users is 7637

select avg(TotalSteps)
from bellabeat.dailyActivity

select avg(FairlyActiveMinutes)
from `bellabeatproject-363419.bellabeat.dailyIntensities`
```

Analysis : -

1. Found positive correlation between Total number of Steps and Calories.
2. Daily average steps of users - 7637
3. Monday and Friday have more sedentary minutes , people are less active on these days.
4. Maximum number of steps on Tuesday and Saturdays.
5. Maximum distance covered by users on Tuesday and saturday.
6. Afternoon is the most active time according to the number of steps.
7. More Calories are used at Afternoon time.
8. Users MET(Metabolic Equivalent of Task) is higher at afternoon time.
    - ➢ Amount of energy used.
    - ➢ 1.5 or lower -> sedentary
    - ➢ 1.6 - 3.0 -> light intensity
    - ➢ 3.0 - 6.0 -> moderate
    - ➢ 6.0 + -> vigorous
9. Total time in bed is more than actual sleep time of users.

10. BMI : (Body Mass Index)
    Healthy range of BMI users are more likely to be active.

    - lower than 18.5 – underweight
    - between 18.5 and 24.9 – healthy range
    - between 25 and 29.9 – overweight.
    - between 30 and 39.9 – obesity.
    - 40 or over – severe obesity.