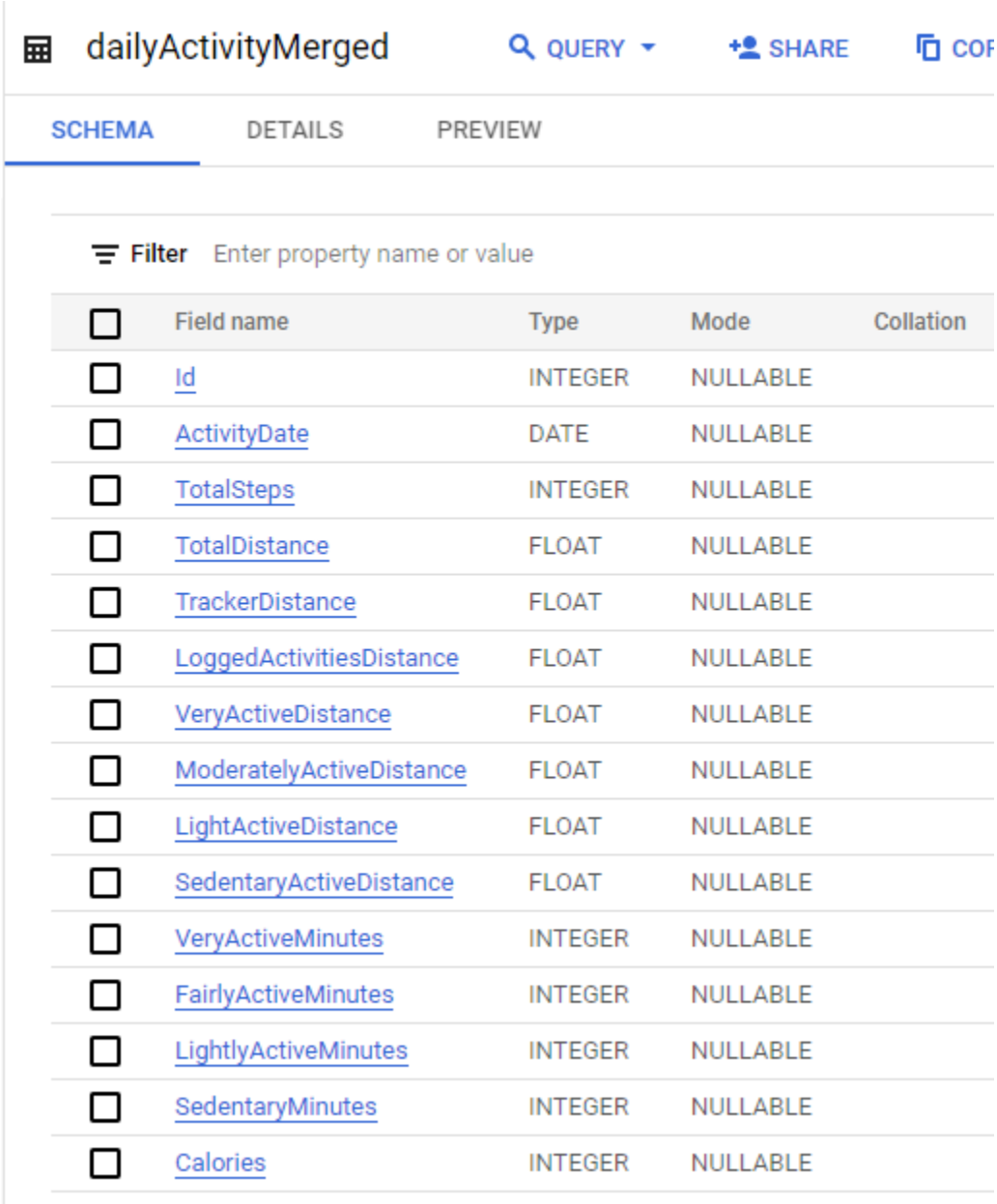
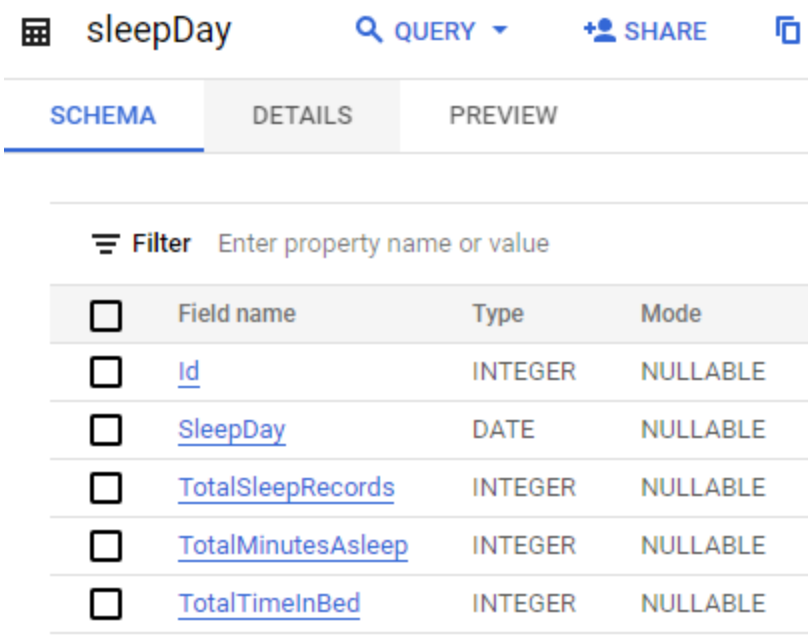
Data Cleaning:

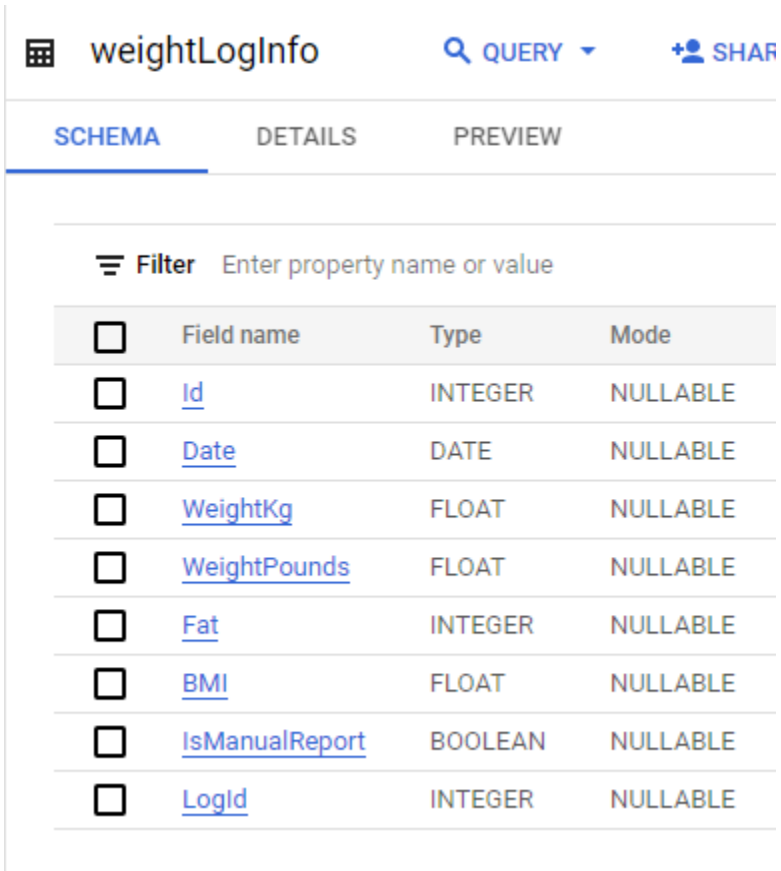
To begin with, I uploaded the data into Bigquery for cleaning and analysis. One thing that happened was uploading the sleepDay and weightLogInfo errors. This was due to the fact that the date column had AM and PM indicators in the columns. After reviewing the data, the time measurements were not a factor in consideration when reviewing. As such the data type was changed to date within google sheets then exported back to CSV. This allows me to then import all data back into BigQuery.

Reviewing the data shows that it is in the correct types for each column at first look.



Field name	Type	Mode	Collation
Id	INTEGER	NULLABLE	
ActivityDate	DATE	NULLABLE	
TotalSteps	INTEGER	NULLABLE	
TotalDistance	FLOAT	NULLABLE	
TrackerDistance	FLOAT	NULLABLE	
LoggedActivitiesDistance	FLOAT	NULLABLE	
VeryActiveDistance	FLOAT	NULLABLE	
ModeratelyActiveDistance	FLOAT	NULLABLE	
LightActiveDistance	FLOAT	NULLABLE	
SedentaryActiveDistance	FLOAT	NULLABLE	
VeryActiveMinutes	INTEGER	NULLABLE	
FairlyActiveMinutes	INTEGER	NULLABLE	
LightlyActiveMinutes	INTEGER	NULLABLE	
SedentaryMinutes	INTEGER	NULLABLE	
Calories	INTEGER	NULLABLE	





Now that the data is imported into BigQuery we can then start querying and cleaning. The code that will be listed here is what I think will be major parts of the code; however, a full code sheet can be found on github.

One thing that is found is that the only table that has null values are in the weightLogInfo where the column FAT is null. Since this is data that is missing from the dataset that was provided as well as not a column that we will be mainly using this is alright and we can continue forward.

Another aspect to consider is that since the data is in long format there are duplicates in the ID column for each table.