Team, please note down all your observations around the Data in the below table.
In-case you need you may add more rows to the table.

Team 1004 / WEEK-2 Summary

#	Observation	How do you want to resolve it (or) How did you resolve it?
1	events_data.csv file size is 286 MB	Using google drive shortcut key, directly it is being imported in the Google colab notebook
2	gender_age_train & phone_brand_device_model both are sql instances; challanges faced to upload inside dataframe	Using code support from "stack overflow" regarding mysql-connector packages, it has been solved.
3	phone_brand_device_model containes chinese font; translation to english to be made	using google translator and mapping thru' a dictionary the english translation was easily being done
4	Missing value observed for events_data.csv file for below columns, device_id : 453 numbers (0.014% of data) latitude : 423 numbers (0.013% of data) longitude : 423 numbers (0.013% of data) state : 377 numbers (0.011% of data) Other files (gender_age_train & phone_brand_device_model) are no missing values	It has been observed that; a) latitude and longitude are missing together but city present for those--> so this can be replaced with respective values b) device_id is independenetly missing so this could be deleted or placeholder may be used so we can analyse the anonymous device id w.r.t its usage c) state is also independently missing but city presents --> so this can be replaced accordingly.
5	From the events dataset, extract the 6 states, Andhra Pradhesh, Mizoram, Himachal Pradhesh, Meghalaya, Pndicherry and Andaman Nicobar Islands	This can be done by extracting the values based on state and then concating to a single data set
6	Merge the gender_age, phone_brand_device and events into a single dataset to do further analysis	This can be done merging all datasets baded on device id
7	Now we can find the missing values of aggregate data set	through repeated basic data desciption
8	gender_age_train data contains record for 74645 devices, all devices are unique	
9	gender_age_train data contains records for devices owned as per gender and age	
10	gender_age_train also contains a column for age groups, age groups are divided as per gender and age	
11	For Male, the age groups are "Below 22", "23-26", "27-28", "29-31", "32-38" and "Above 39" For Female, the age groups are "Below 23", "24-26", "27-28", "29-32", "33-42" and "Above 42"	Looking at this, the age group column looks to be contianing different age groups, the age groups column should be dropped and new age groups column should be created with same age groups to compare the M-F data.
12	Seems there is no outlier in terms of age looking at data, minimum age is 1 and maximum age is 96	We have seen below 10 years of age we have 10 number datapoint; which may be deleted or separately treated.
13	Device Id seems to be the id of the device and looks to be a primary key across all data sets	

Team, please note down all your observations around the Data in the below table.
In-case you need you may add more rows to the table.

Team 1004 / WEEK-2 Summary

#	**Observation**	**How do you want to resolve it (or) How did you resolve it?**
14	47904 devices are owned by Male 26741 devices are owned by Female	
15	phone_brand_device_model_df contains records for 87726, all devices are unique	
16	There are total 116 Unique Brands and most of them contains Names in chinese	The chinese names can be translated using Google translator in English for better understanding
17	There are total 1467 Unique Model names	The chinese names can be translated using Google translator in English for better understanding
18	Total 453 records have missing value for device Id	As we agreed, the device id are being replaced using "mode" of the latitude/longitude present for respective cities.
19	Latitude and Longitude have 423 missing values each, for a record where latitude is missing, longitude is also missing.	As mentioned above as well, As city is available for all the records, the latitude amd longitude values can be replaced by the value of same city from other set of records
20	Total 377 records have missing value for state	As mentioned above, As city is available for all the records, the state values can be replaced by the value of same city from other set of records
21	Timestamp has 588126 unique records, and the timestamp data is precise till second	
22	Device Id has unique 60865 records, which means only 60865 devices are used during the data collection period	
23	The maximum usage are from the state of Delhi, which is 23.109# of data	
24	There are total 329125 records for these states, and then there are 377 records with missing value of states	out of the 377 records, the 47 records are for Vishakhapatnam, which is for AndhraPradesh. So the data we will have for these states will have 329172 records These 6 states contains 10.119% of data
25	Some of the Device Ids are having negative values	Need to check if there is any signifance of negative ids? Does it have any signifance? Do we need to change it to have +ve values? To be discussed as a team **Since device ids are app generated code and hence considered as string so we will consider "negative" sign as it is.**
26	Phone Brand Dataset can be considered as master data for all the devices Gender Age Dataset can be considered as the sales data for all the devices	

Team, please note down all your observations around the Data in the below table.
In-case you need you may add more rows to the table.

Team 1004 / WEEK-2 Summary

#	**Observation**	**How do you want to resolve it (or) How did you resolve it?**
27	The common devices found between Events Data and the the common devices between Phone Brand and age is only 406. This is a major disconnect in data and it laves us with only406 records to work with, though we need to look into details of why there such a big gap between the data of devices between various datasets.	Seems the device id data in events data is in float format, which is causing the match to be not found between the device id's of various datasets. To fix this the device id data needs to be convert to same data type for all the datasets and remove any decimals and then the data needs to be used
28	DeviceId in gender_age_train/phone_brand_device_model DF is int64 but in events_data_df its float64. All DeviceId are unique for all three DataFrames .and event_id is also unique in Events DF. Timestamp needs to be further classified as Month, Days of week so that we can find out most Active event occurence and based on user location study.	We can do lebel enconding for Gendar feature in gender_age_train_df and create new feature as Gendar 0,1(F,M) only. we can merge phone_brand and gender age considering device_id as their primary key and then we will merge events considering event_id as primary key and device id as foreign key of phone_brand and gender_age. day of the week: Monday to Sunday Active time : hour
29	The second challenge after downloading the	The challenge of establishing a correct database connection, was simplified by using a try-except-finally
	events_data.csv file size, was to make a connection	loop in Python to catch any unexpected exceptions. The second challenge of extracting the Sql tables
	to the database, for which we used the mqsql.connector	was simplified by using DBeaver Tool from Google, which makes task of extracting Sql tables easy by downloading
	package of Python. Once connection was established,	table data in a Dataframe.
	the next challenge after establishing the database	
	connection was to extract sql tables.	
30	We need to observe patterns in the data using the given	Actionable Insight derived from this piece of information is that **InsaidTelecom** COMPANY can focus
	information. We observe after having a look at the dataset	their Sales & Marketing Efforts more towards the Male population to increase their Sales Revenue. They could
	that out of total mobiles in our dataset, ownership is	also come out with Attractive Sales & Marketing Offers targeted towards the Female Population to increase
	skewed towards the Male population which buys 47904	their reach to the Female Population.
	handsets compared to 26741 handsets bought by Females.	

Team, please note down all your observations around the Data in the below table.
In-case you need you may add more rows to the table.

Team 1004 / WEEK-2 Summary

#	**Observation**	**How do you want to resolve it (or) How did you resolve it?**
31	Outliers exist for latitude and longitude values, there are many records for which the latitude and longitude value given are outside of India, it covers records for multiple cities, few of the cities which contain some incorrect latitude and longitude values are Pune, Delhi, Vishakhapatnam etc.	As we have the correct value for cities as well, we can take the mode of the latitude and longitude value for the city and can update those values in the records containing the incorrect data for latitude and longitude
32	Looking at the max and min values for latitude (max 41.87, greater than 37.60) and longitude (min 42.35, lesser than 68.7), looks like few records are out of the range of India	The latitude and longitude for India are in the range: Longitude : 68.7 - 97.25 (West -> East) Latitude :08.00 - 37.60 (South -> North) few values for latitude and longitude are out of the above range The images showing the locations out of India are added in the sheet "Folium Image", these images contain one record each within in India and other records out of India
33	Fixing the records with device id	The Device id in events data is of float type and by default it contains exponential and decimal values if we read it as is. To fix the issue we either need to convert the data to String/int during load or after load we need to convert each value to int or string by applying formatting on float values. During load using pd.read_csv we ca use below statement to convert data to String. And the we can convert the type of columns in other two data sets to string siply using .astype(str). as it is of int type, there should not be aby problem. events_data_df = pd.read_csv('data/events_data.csv', dtype={'device_id': np.str})
34	Age column has 5 instances for age below less than 10 years. min age is one instance with one year and 4 instances with 6 yrs	Once state wise data is extracted, it is seen that there 10 instaces of age one year. all are connected to single device id 3553057874282315257 from Andhra Pradesh and the age group is shown as M22-. This can be assumed to be a case of wrong data entry and may be replaced by median
35	The state wise data set has following unique data event_id 329172 device_id 5223 timestamp 240679 longitude 5203 latitude 5210 city 113 state 6 gender 2 age 72 age_group 12 brand 69	The age group need further handling to make the groups more uniform and meaningful In this regard, one of the recent study shows some light on age distribution, we may also use the same for our analysis.