Module 3: Data Acquisition, Evaluation and Exploration

Assignment

Module 3: Assignment

Given a dataset from an outbreak of food-borne illness in the US in 1940.

The dataset can be downloaded from the link:

https://edureka.wistia.com/medias/qp5fzq03hd

Age	sex	timesupper	ill	onsetdate	onsettime	baked_hamburgur	spinach	mashed_potato	cabbages	jello	rolls	brown	milk	coffee	water	cakes	vanilla	chocolate	fruitsalad
27	Male	2000	yes	19-Apr	30	yes	yes	yes	no	no	yes	no	no	yes	no	no	yes	no	no
38	Female	1830	yes	19-Apr	30	yes	yes	yes	yes	no	no	no	no	yes	no	no	yes	yes	no
32	Male	1830	yes	19-Apr	30	yes	yes	no	no	no	no	no	no	yes	no	yes	yes	yes	no
36	Male	1930	yes	18-Apr	2230	yes	yes	no	yes	yes	no	no	no	no	yes	no	yes	no	no
42	Female	1930	yes	18-Apr	2230	yes	yes	yes	no	yes	yes	yes	no	yes	yes	no	yes	no	no
22	Male	1930	yes	19-Apr	200	no	no	no	no	no	no	no	no	no	no	no	yes	yes	no
7	Male	2200	yes	19-Apr	100	no	no	no	no	no	no	no	no	no	no	yes	no	yes	no
17	Male	1900	yes	18-Jun	2300	yes	yes	yes	no	no	yes	yes	no	no	yes	no	yes	yes	no
2	Female	1930	yes	19-Apr	200	no	no	no	no	no	no	no	no	no	no	no	yes	yes	no
16	Male	NA	yes	19-Apr	1030	yes	yes	no	no	no	yes	no	no	yes	no	yes	yes	yes	no
34	Male	NA	yes	19-Apr	30	no	no	no	no	no	no	no	no	no	no	no	yes	no	no
19	Female	NA	yes	18-Apr	2215	yes	yes	no	yes	no	yes	yes	no	no	no	no	yes	no	no
17	Male	NA	yes	18-Apr	2200	yes	yes	yes	yes	yes	yes	no	no	yes	yes	yes	yes	yes	no
5	Male	2200	yes	19-Apr	100	no	no	no	no	no	no	no	no	no	no	yes	yes	no	no
48	Female	NA	yes	18-Apr	2300	yes	yes	yes	yes	yes	yes	yes	no	no	yes	yes	yes	yes	no
15	Female	NA	yes	18/4	2145	no	yes	yes	no	no	yes	no	no	no	yes	yes	yes	no	no
32	Male	NA	yes	18-Apr	2145	no	yes	yes	yes	no	yes	yes	no	no	yes	yes	yes	no	no
seven	Male	2200	yes	19-Apr	100	no	no	no	no	no	no	no	no	no	no	yes	yes	yes	no
20	Male	NA	yes	18-Apr	2300	yes	yes	yes	no	yes	yes	yes	no	yes	no	yes	yes	no	no
18	Female	NA	yes	18-Apr	2100	yes	yes	yes	no	yes	yes	yes	no	yes	no	yes	yes	no	yes

Column information:

- age - the person's age in years
- sex - the person's gender
- timesupper - the time the person ate (nearest half hour)
- ill - whether the person developed GI illness after the supper
- onsetdate - the date of onset of illness for those who became ill
- onsettime - the time that the person reported first feeling ill (nearest half hour)
- 15 variables indicating whether the person reported eating specific food items at the supper

Perform the following tasks on the dataset

Task 1: Import the dataset into R

→ Ensure imported data is interpreted correctly by R.

Task 2: Perform Data Wrangling and clean the data

→ Convert Age column to type Numeric

→ Omit NULL values from non-numerical data fields

→ Replace NULL values with average value of that variable

→ Give proper time format for timesupper column using strptime() and get the date and time in the below specified format

→ Final output should be like

Age	sex	timesupper	ill	onsettime	baked_hamburgur	spinach	mashed_potato	cabbages	jello	rolls	brown	milk	coffee	water	cakes	vanilla	chocolate	fruitsalad
27	Male	1940-04-18 20:00:00	yes	19-Apr 00:30	yes	yes	yes	no	no	yes	no	no	yes	no	no	yes	no	no
38	Female	1940-04-18 18:30:00	yes	19-Apr 00:30	yes	yes	yes	yes	no	no	no	no	yes	no	no	yes	yes	no
32	Male	1940-04-18 18:30:00	yes	19-Apr 00:30	yes	yes	no	no	no	no	no	no	yes	no	yes	yes	yes	no
36	Male	1940-04-18 19:30:00	yes	18-Apr 22:30	yes	yes	no	yes	yes	no	no	no	no	yes	no	yes	no	no
42	Female	1940-04-18 19:30:00	yes	18-Apr 22:30	yes	yes	yes	no	yes	yes	yes	no	yes	yes	no	yes	no	no
22	Male	1940-04-18 19:30:00	yes	19-Apr 02:00	no	no	no	no	no	no	no	no	no	no	no	yes	yes	no
7	Male	1940-04-18 22:00:00	yes	19-Apr 01:00	no	no	no	no	no	no	no	no	no	no	yes	no	yes	no
17	Male	1940-04-18 19:00:00	yes	18-Jun 23:00	yes	yes	yes	no	no	yes	yes	no	no	yes	no	yes	yes	no
2	Female	1940-04-18 19:30:00	yes	19-Apr 02:00	no	no	no	no	no	no	no	no	no	no	no	yes	yes	no
16	Male	1940-04-18 20:30:00	yes	19-Apr 10:30	yes	yes	no	no	no	yes	no	no	yes	no	yes	yes	yes	no
34	Male	1940-04-18 20:30:00	yes	19-Apr 00:30	no	no	no	no	no	no	no	no	no	no	no	yes	no	no
19	Female	1940-04-18 20:30:00	yes	18-Apr 22:15	yes	yes	no	yes	no	yes	yes	no	no	no	no	yes	no	no
17	Male	1940-04-18 20:30:00	yes	18-Apr 22:00	yes	yes	yes	yes	yes	yes	no	no	yes	yes	yes	yes	yes	no
5	Male	1940-04-18 22:00:00	yes	19-Apr 01:00	no	no	no	no	no	no	no	no	no	no	yes	yes	no	no
48	Female	1940-04-18 20:30:00	yes	18-Apr 23:00	yes	yes	yes	yes	yes	yes	yes	no	no	yes	yes	yes	yes	no
15	Female	1940-04-18 20:30:00	yes	18-Apr 21:45	no	yes	yes	no	no	yes	no	no	no	yes	yes	yes	no	no
32	Male	1940-04-18 20:30:00	yes	18-Apr 21:45	no	yes	yes	yes	no	yes	yes	no	no	yes	yes	yes	no	no
7	Male	1940-04-18 22:00:00	yes	19-Apr 01:00	no	no	no	no	no	no	no	no	no	no	yes	yes	yes	no
20	Male	1940-04-18 20:30:00	yes	18-Apr 23:00	yes	yes	yes	no	yes	yes	yes	no	yes	no	yes	yes	no	no

Task 3: Perform Analysis and Visualization using appropriate type of graph, to find out:

→ Which is the most consumed food item among the patients.

→ Find the average age of people who are ill using Boxplot.

→ Visualize gender ratio from the data.