Module 3: Data Acquisition, Evaluation and Exploration

Assignment Solution

Module 3: Assignment Solution

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.

Solution:

```
#We can either load the data into R from our Directory or
#we can set the directory to where the data is located
getwd()
#Here I copied the data into this directory C:/Users/Edureka/Documents
disease<-read.csv("Disease_data.csv")
#Now we can View it using
View(disease)
```

Task 2: Perform Data Wrangling and clean the data

→ Convert Age column to type Numeric

→ Replace NULL values with average value of that variable

→ Give proper time format for timesupper column using strptime() and paste the date 1940-04-18 with it.

→ 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

Solution:

1.Age column

```
#lets see how the variable is stored
mode(disease)
#Determine object class of variable
class(disease$Age)
#Here it is shown Factor, But age should be numeric

disease$Age
#One of the values was entered as "seven" when it should have been entered as the number 7.
#Thus, when the data were imported into R,
#The entire column was read in as character strings and the variable was converted to a factor.

levels(disease$Age)

#Now we need to replace it with 7
levels(disease$Age)[levels(disease$Age)=="seven"] <- "7"
#To check it
levels(disease$Age)

#Now lets see the range of the column
range(disease$Age)

#It show error because its still Factor variable, We need to convert it into Numeric
AGE <- as.numeric(as.character(disease$Age))
disease$Age <- AGE

#Now to check the type
class(disease$Age)
```

2.Sex column

```
disease$sex
mode(disease$sex)
class(disease$sex)
#Its a Factor, lets see the levels
levels(disease$sex)
#Here we can see "-1" is an error value, we need to replace it
table(disease$sex)
#Replace "-1" to "MALE", First get index of the appropriate row

which(disease$sex=="-1")
#Its in 22nd row

SEX<-as.character(disease$sex)
SEX[22]<-"Male"
SEX
disease$sex<-as.factor(SEX)
class(disease$sex)
levels(disease$sex)
```

3.timesupper column

```
disease$timesupper
#Here ther are lots of null values, we can replace it by the average time
mean(disease$timesupper)
#we need to avoid null values
mean(disease$timesupper,na.rm = TRUE)
#1981.818 when we convert it into time, we get 2021, its nearest half hour will be 2030
#we can replace NA by

disease$timesupper[which(is.na(disease$timesupper))]<-"2030"
disease$timesupper

#lets check the values in this column
table(disease$timesupper)
#we can se that the value 1100 dont fit in supper time, but there is no way to replace it
#so we will leave it as it is

#Next we need to change the format of the column,we will use strptime function for that
SUPPER <- strptime(disease$timesupper,format="%H%M")
head(SUPPER)
#to copy the date 1940-04-18, we can use

supperdate <- "1940-04-18"
supper.datetime <- paste(supperdate,disease$timesupper)
head(supper.datetime)
#using strptime
SUPPER <- strptime(supper.datetime,format="%Y-%m-%d %H%M")
head(SUPPER)
disease$timesupper <- SUPPER
View(disease)
```

4.onsetdate column

```
mode(disease$onsetdate)
class(disease$onsetdate)
disease$onsetdate
levels(disease$onsetdate)
# we can see that 18/4 and 18-Apr are the same so we need to replace it
which(disease$onsetdate=="18/4")
SETDATE<-as.character(disease$onsetdate)
SETDATE[16]<-"18-Apr"
SETDATE
#we can replace the original column with
disease$onsetdate<-SETDATE
View(disease)
```

4.onsettime column

```
mode(disease$onsettime)
class(disease$onsettime)
range(disease$onsettime)
#to properly format it we can use strptime
TIME<-strptime(disease$onsettime,format = "%H%M")
#most values are NA beacuse thier values are not recognized by R,
#So we need to convert every value into 4 digits

disease$onsettime
#Here we can see the non 4 digit values are 30, 100, 200, 215 and 230
#so let's replace them
disease$onsettime[which(disease$onsettime=="30")]<-"0030"
disease$onsettime[which(disease$onsettime=="100")]<-"0100"
disease$onsettime[which(disease$onsettime=="200")]<-"0200"
disease$onsettime[which(disease$onsettime=="215")]<-"0215"
disease$onsettime[which(disease$onsettime=="230")]<-"0230"

disease$onsettime
#now we can format them
TIME<-strptime(disease$onsettime,format = "%H%M")
TIME
#To remove the date part
TIME<-format(TIME,"%H:%M")
TIME

#We can merge onsetdate and onsettime into one using,
SETTIME<-paste(disease$onsetdate,TIME)
SETTIME
disease$onsettime<-SETTIME
View(disease)
#now we can remove the column onsetdate
colnames(disease)
#its the fifth column
disease<-disease[,-5]
View(disease)
```

5.food items columns

```
#here there is a NA value, we can replace it by
disease$chocolate[which(is.na(disease$chocolate))]<-"yes"
disease$chocolate

#next we can see the last colunmn has an error in the last row############
disease$fruitsalad
#we can easiy replace it by
disease$fruitsalad[46]<-"yes"
disease$fruitsalad
#Still levels() is showing the error value, we need to change it
fruitsalad<-as.character(disease$fruitsalad)
#we can again change it back to factors
fruitsalad<-as.factor(fruitsalad)
fruitsalad
disease$fruitsalad<-fruitsalad
View(disease)
#we have a clean data set now
```

Our data set is clean now

Task 3: Perform Analysis and Visualization to find out

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

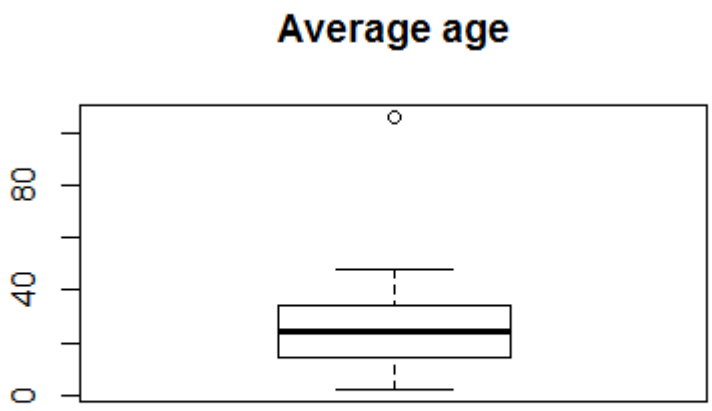
Solution:

```
#Which is the most consumed food item among the patients.

View(disease)
#here we have 14 food items, lets check which is the most consumed product among patients
#we can create a subset with these 14 columns
Food<-disease[,6:19]
View(Food)

class(Food)
doc<-summary(Food)
doc
hist(which(Food=="yes"),main = "Food items consumed",labels = seq(1,14,1))

#Here 12th value is the largest
colnames(Food[12])
```


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

Solution

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

boxplot(disease$Age)
avg<-boxplot(disease$Age,main = "Average age")
avg$stats
# we can see that 24.5 is the average value
# Here we can see 106 is shown as an Outlier
```



→ Visualize ratio of sex from the data using plot().

Solution:

```
#®  Find out the ratio of sex from the list and plot it using Histogram.
plot(disease$sex)
```