Homework 1 Predictive Analytics BUAN 6337.006

Submitted by :

Abhishek Arya (Net id : axa220149)

Jitendra Sai Rajesh Motepalli (Net id : mxj230027)

Chandra has Mididuddi (Net id : Cxm220032)

Your task is to read this data file into R properly

a.

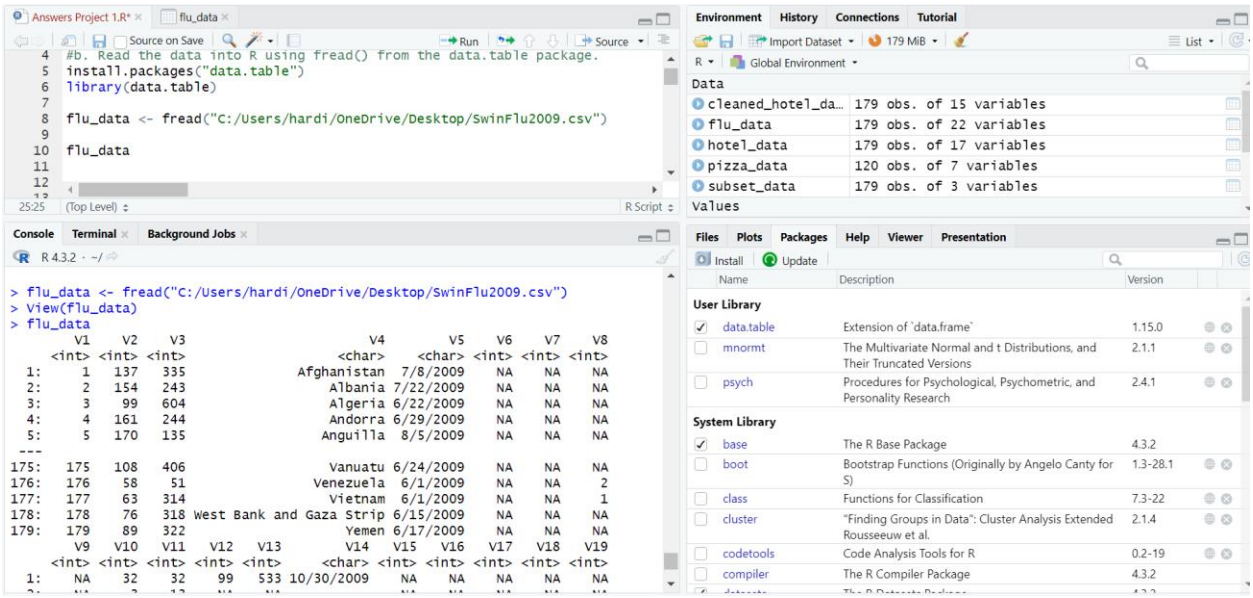
First, examine the raw data file SwineFlu2009.csv using Excel.

Answer a. After exploring the excel file SwineFlu2009 22 unnamed columns are there. Values inside these columns are also unorganized and missing. We know that we have to do a lot of cleaning and filtering. Two date columns are there, 1 character column is there, 1 index column and rest are numeric columns.

b.

Read the data to memory using fread(). Examine the data in Rstudio.

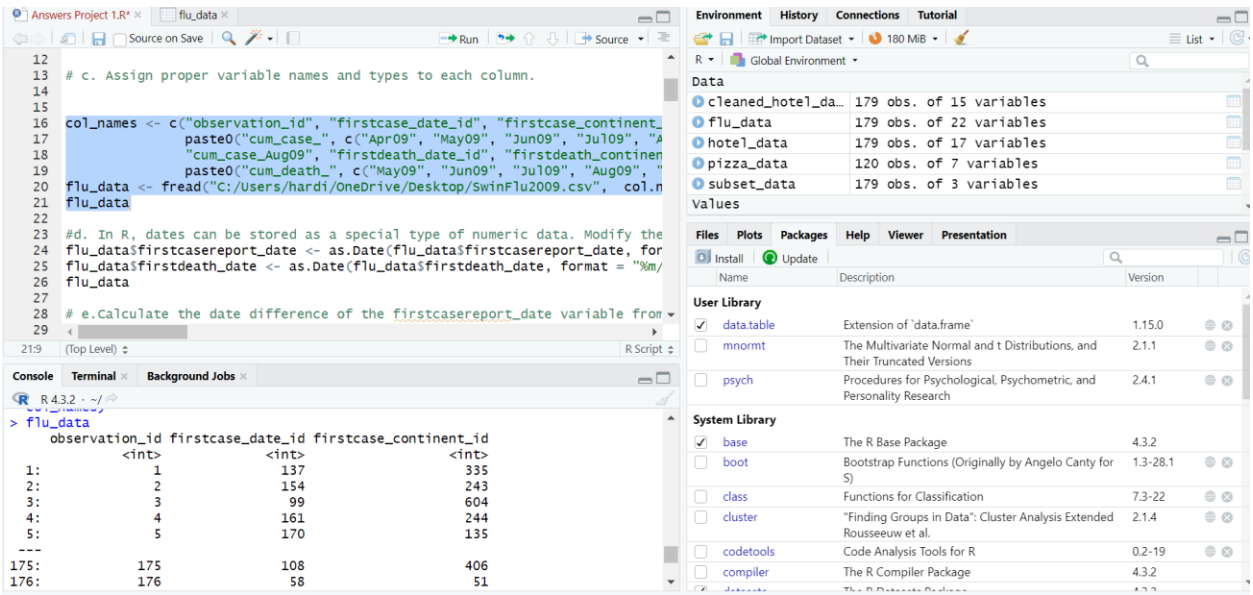
Answer b.



c. Then, assign the proper variable name to each variable. Make sure that each variable is assigned the correct type – character or numeric. (hint: use colClasses() to examine the class of columns)

Answer c. The R-code and screenshot are below:

```
col_names <- c("observation_id", "firstcase_date_id", "firstcase_continent_id", "country", "first
casereport_date",
          paste0("cum_case_", c("Apr09", "May09", "Jun09", "Jul09", "Aug09")),
          "cum_case_Aug09", "firstdeath_date_id", "firstdeath_continent_id", "firstdeath_date",
          paste0("cum_death_", c("May09", "Jun09", "Jul09", "Aug09", "Sep09", "Oct09", "Nov0
9", "Dec09")))
flu_data <- fread("C:/Users/hardi/OneDrive/Desktop/SwinFlu2009.csv",  col.names = col_name
s)
flu_data
```



d. In R, dates can be stored as a special type of numeric data. Modify the DATA step to make sure that the dates are read in the correct R date format (not as character). (HINT: Use the correct date type format statements in as.Date(), e.g., format = "%m/%d/%Y")

Answer d.

flu_data$firstcasereport_date <- as.Date(flu_data$firstcasereport_date, format = "%m/%d/%Y"
)
flu_data$firstdeath_date <- as.Date(flu_data$firstdeath_date, format = "%m/%d/%Y")
flu_data

Answers Project 1.R* × flu_data ×

Source on Save Run Source

```
# c. Assign proper variable names and types to each column.


col_names <- c("observation_id", "firstcase_date_id", "firstcase_continent_
               paste0("cum_case_", c("Apr09", "May09", "Jun09", "Jul09", "A
               "cum_case_Aug09", "firstdeath_date_id", "firstdeath_continen
               paste0("cum_death_", c("May09", "Jun09", "Jul09", "Aug09", "
flu_data <- fread("C:/Users/hardi/OneDrive/Desktop/SwinFlu2009.csv",  col.n
flu_data

#d. In R, dates can be stored as a special type of numeric data. Modify the
flu_data$firstcasereport_date <- as.Date(flu_data$firstcasereport_date, for
flu_data$firstdeath_date <- as.Date(flu_data$firstdeath_date, format = "%m/
flu_data
```

26:9 (Top Level) R Script

Console Terminal × Background Jobs ×

R 4.3.2 · ~/

```
179:            179                 89                     322
                 country firstcasereport_date cum_case_Apr09
                  <char>               <Date>          <int>
  1:           Afghanistan           2009-07-08             NA
  2:               Albania           2009-07-22             NA
  3:               Algeria           2009-06-22             NA
  4:               Andorra           2009-06-29             NA
  5:              Anguilla           2009-08-05             NA
 ---
175:               Vanuatu           2009-06-24             NA
176:             Venezuela           2009-06-01             NA
177:               Vietnam           2009-06-01             NA
178: West Bank and Gaza Strip          2009-06-15             NA
179:                 Yemen           2009-06-17             NA
```

Answers Project 1.R* × flu_data ×

Source on Save Run Source

```
# c. Assign proper variable names and types to each column.


col_names <- c("observation_id", "firstcase_date_id", "firstcase_continent_
               paste0("cum_case_", c("Apr09", "May09", "Jun09", "Jul09", "A
               "cum_case_Aug09", "firstdeath_date_id", "firstdeath_continen
               paste0("cum_death_", c("May09", "Jun09", "Jul09", "Aug09", "
flu_data <- fread("C:/Users/hardi/OneDrive/Desktop/SwinFlu2009.csv",  col.n
flu_data

#d. In R, dates can be stored as a special type of numeric data. Modify the
flu_data$firstcasereport_date <- as.Date(flu_data$firstcasereport_date, for
flu_data$firstdeath_date <- as.Date(flu_data$firstdeath_date, format = "%m/
flu_data
```

26:9 (Top Level) R Script

Console Terminal × Background Jobs ×

R 4.3.2 · ~/

```
     firstdeath_date cum_death_May09 cum_death_Jun09 cum_death_Jul09
               <Date>           <int>           <int>           <int>
  1:      2009-10-30              NA              NA              NA
  2:            <NA>              NA              NA              NA
  3:      2009-11-30              NA              NA              NA
  4:            <NA>              NA              NA              NA
  5:            <NA>              NA              NA              NA
 ---
175:            <NA>              NA              NA              NA
176:      2009-07-20              NA              NA              NA
177:      2009-08-05              NA              NA              NA
178:      2009-08-09              NA              NA              NA
179:      2009-08-19              NA              NA              NA
```

e. Calculate the date difference of the first case report_date variable from the first case report date across the world, which is Apr 24, 2009

Answer e.

```
# Define the world's first case report date
world_first_case_date <- as.Date("2009-04-24")

# Calculate the date difference for each observation
flu_data$days_from_first_incidence <- flu_data$firstcasereport_date - world_first_case_date
flu_data
```

```r
#d. In R, dates can be stored as a special type of numeric data. Modify the
flu_data$firstcasereport_date <- as.Date(flu_data$firstcasereport_date, for
flu_data$firstdeath_date <- as.Date(flu_data$firstdeath_date, format = "%m/
flu_data

# e.Calculate the date difference of the firstcasereport_date variable from
# case report date across the world, which is Apr 24, 2009
# Define the world's first case report date
world_first_case_date <- as.Date("2009-04-24")

# Calculate the date difference for each observation
flu_data$days_from_first_incidence <- flu_data$firstcasereport_date - world
flu_data
# f .Subset the columns (“firstcase_date_id”, “country”) and the answer fro
```

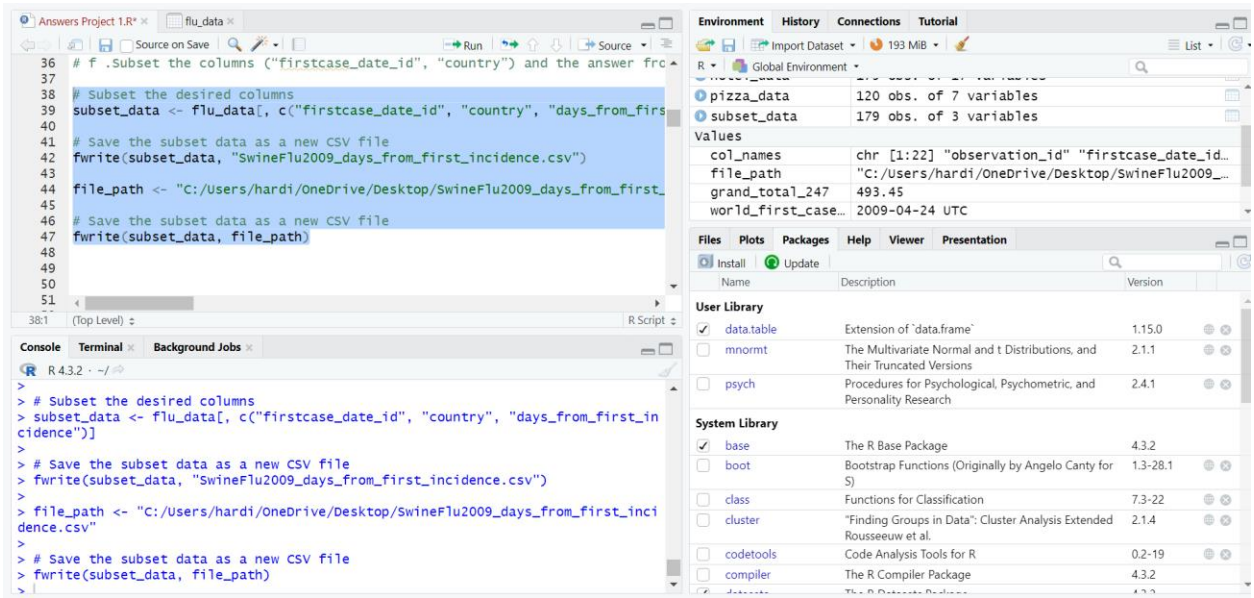
```
178:              NA                1                1                1
179:              NA                1                6               16
     cum_death_Dec09 days_from_first_incidence
                <int>                 <difftime>
  1:              16                   75 days
  2:              NA                   89 days
  3:               3                   59 days
  4:              NA                   66 days
  5:              NA                  103 days
 ---
175:              NA                   61 days
176:             114                   38 days
177:              44                   38 days
178:               9                   52 days
```

f .Subset the columns (“firstcase_date_id”, “country”) and the answer from the above question 1.e, and save it as the file “SwineFlu2009_days_from_first_incidence.csv”) using fwrite(). (HINT: the new csv file should have three columns)

Answer f.

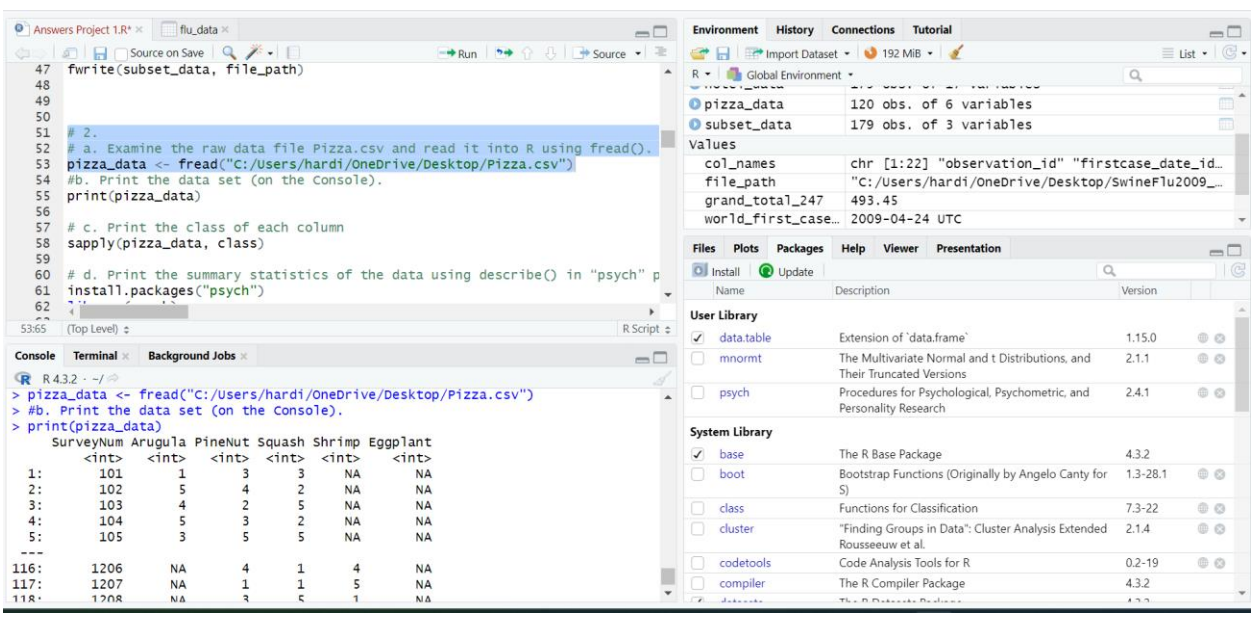
```r
# Subset the desired columns
subset_data <- flu_data[, c("firstcase_date_id", "country", "days_from_first_incidence")]

# Save the subset data as a new CSV file
fwrite(subset_data, "SwineFlu2009_days_from_first_incidence.csv")

file_path <- "C:/Users/hardi/OneDrive/Desktop/SwineFlu2009_days_from_first_incidence.csv"
```

# Save the subset data as a new CSV file
fwrite(subset_data, file_path)



2.
a. Examine the raw data file Pizza.csv and read it into R using fread().

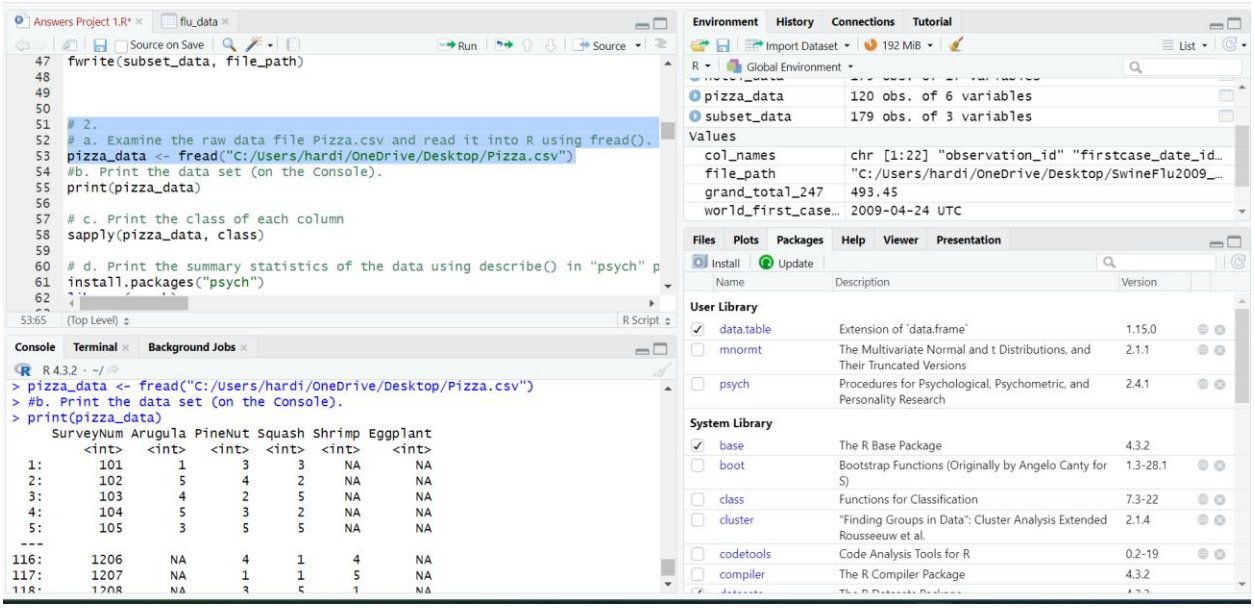
Answer 2 a.
pizza_data <- fread("C:/Users/hardi/OneDrive/Desktop/Pizza.csv")



#b. Print the data set (on the Console).

Answer 2b.

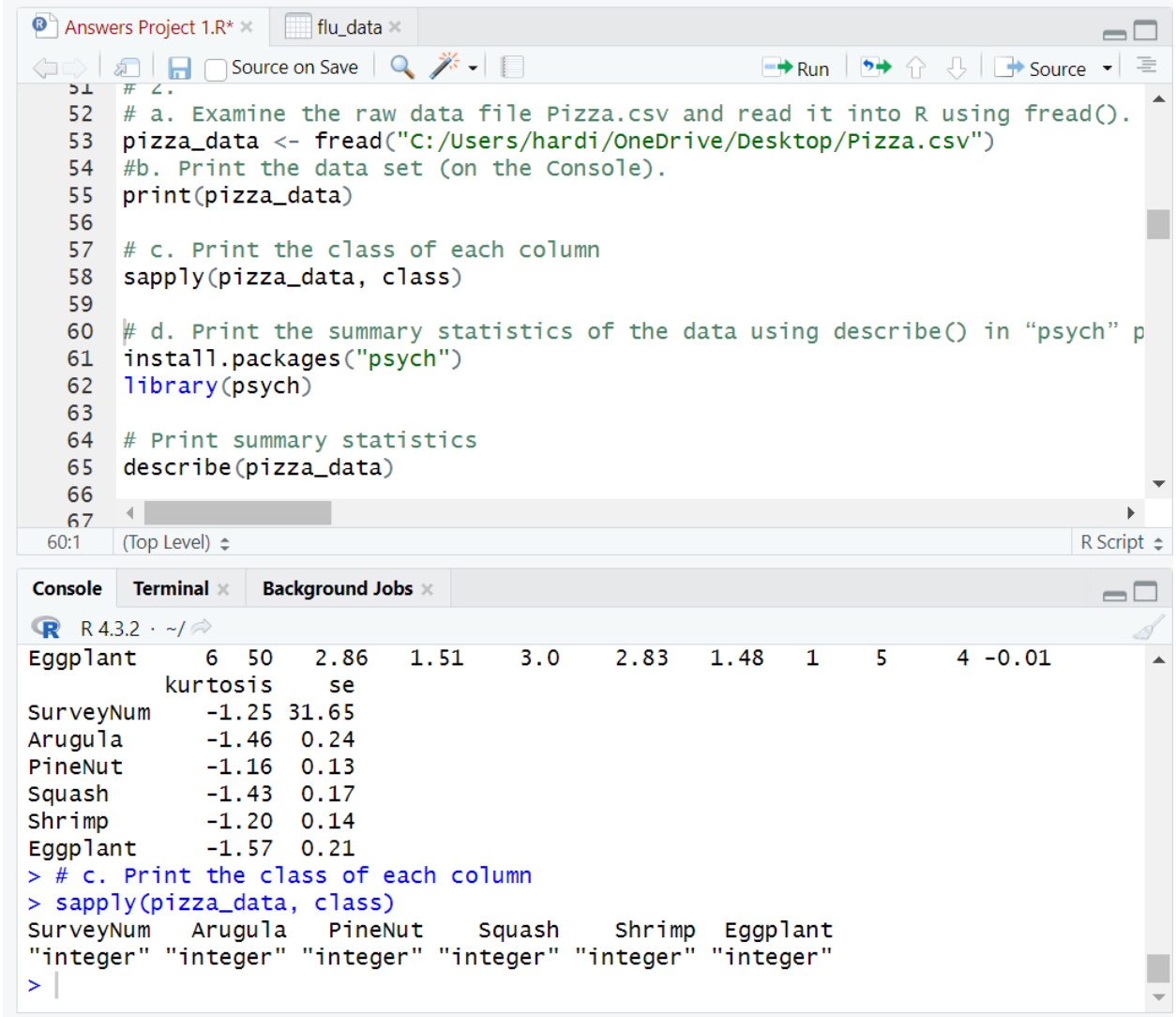
print(pizza_data)



c. Print the class of each column

Answer 2c.

sapply(pizza_data, class)



d. Print the summary statistics of the data using describe() in "psych" package.
Answer 2d.

```
install.packages("psych")
library(psych)

# Print summary statistics
describe(pizza_data)
```

Answers Project 1.R* × flu_data ×

Source on Save Run Source

```r
# c. Print the class of each column
sapply(pizza_data, class)

# d. Print the summary statistics of the data using describe() in "psych" p
install.packages("psych")
library(psych)

# Print summary statistics
describe(pizza_data)

# e.Open the raw data file in a simple editor like WordPad and compare the
# There may be issues with the column types not being read correctly, leadi
# For example, if a numeric column is read as character, it may affect summ
#f. Read the same raw data file, Pizza.csv, again. This time, make sure the
```

67:1 (Top Level) R Script

Console Terminal × Background Jobs ×

R 4.3.2 · ~/

```
> library(psych)
> # Print summary statistics
> describe(pizza_data)
          vars   n   mean     sd median trimmed    mad min  max range  skew
SurveyNum    1 120 655.50 346.66  655.5  655.50 444.78 101 1210  1109  0.00
Arugula      2  40   3.08   1.49    3.0    3.09   1.48   1    5     4 -0.12
PineNut      3 100   3.14   1.29    3.0    3.17   1.48   1    5     4  0.02
Squash       4  80   3.16   1.51    3.0    3.20   2.22   1    5     4 -0.14
Shrimp       5  90   2.97   1.33    3.0    2.96   1.48   1    5     4 -0.03
Eggplant     6  50   2.86   1.51    3.0    2.83   1.48   1    5     4 -0.01
          kurtosis    se
SurveyNum    -1.25 31.65
Arugula      -1.46  0.24
```

e. Open the raw data file in a simple editor like WordPad and compare the data values to the output from part b) to make sure that they were read correctly into R. In a comment in your report, identify any problems with the R data set that cannot be resolved using the fread(). Explain what is causing the problem. (Hint: You need to make sure the type of each variable is read correctly.)

Answer 2e.

There may be issues with the column types not being read correctly, leading to incorrect data representation.

For example, if a numeric column is read as character, it may affect summary statistics calculations.

#f. Read the same raw data file, Pizza.csv, again. This time, make sure the issues you've identified in the previous step ls resolved.

Answer 2 f.

No problem arised first time.

g. Create a column that contains the average ratings for each topping. (Hint: You need to make sure “NA” entries are not included in the average. They should not be treated as zeros. See the documentation for rowMeans().)

Answer 2g.

pizza_data$average_rating <- rowMeans(pizza_data[, -1], na.rm = TRUE)

pizza_data

Answers Project 1.R* × flu_data ×

Source on Save Run Source

```
# Print summary statistics
describe(pizza_data)

# e.Open the raw data file in a simple editor like WordPad and compare the
# There may be issues with the column types not being read correctly, leadi
# For example, if a numeric column is read as character, it may affect summ
#f. Read the same raw data file, Pizza.csv, again. This time, make sure the
 # No problem arised first time.

# g.Create a column that contains the average ratings for each topping. (Hi
pizza_data$average_rating <- rowMeans(pizza_data[, -1], na.rm = TRUE)
pizza_data


# 3
```

76:1 (Top Level) R Script

Console Terminal × Background Jobs ×

R 4.3.2 · ~/

```
> pizza_data
     SurveyNum Arugula PineNut Squash Shrimp Eggplant average_rating
         <int>   <int>   <int>  <int>  <int>    <int>          <num>
  1:       101       1       3      3     NA       NA       2.333333
  2:       102       5       4      2     NA       NA       3.666667
  3:       103       4       2      5     NA       NA       3.666667
  4:       104       5       3      2     NA       NA       3.333333
  5:       105       3       5      5     NA       NA       4.333333
 ---
116:      1206      NA       4      1      4       NA       3.000000
117:      1207      NA       1      1      5       NA       2.333333
118:      1208      NA       3      5      1       NA       3.000000
119:      1209      NA       4      5      3       NA       4.000000
120:      1210      NA       5      5      1       NA       3.666667
```

3.

a. Examine the raw data file Hotel.csv and read it into R using fread(). Is there any “problem” with this data read? Explain.

Answer 3a.

```
Warning message:
In fread("C:/Users/hardi/OneDrive/Desktop/Hotel.csv") :
  Stopped early on line 4. Expected 11 fields but found 12. Consider fill=TRUE and comment.char
=. First discarded non-empty line: <<220,5,2,3,2014,2,12,2014,YES,2,Basic w/view,155>>
```

The warning message indicates that there was an inconsistency in the number of fields detected by fread() on line 4 of the "Hotel.csv" file. It expected 11 fields but found 12.

To address this issue, you can use the fill = TRUE parameter in the fread() function, which allows fread() to fill missing values with NA when the number of fields is inconsistent across lines. Additionally, you can specify the verbose = TRUE parameter to handle comments if necessary.

```
hotel_data <- fread("C:/Users/hardi/OneDrive/Desktop/Hotel.csv")
hotel_data <- fread("C:/Users/hardi/OneDrive/Desktop/Hotel.csv", fill= TRUE, verbose = TRUE)
hotel_data
```

Answers Project 1.R*

Source on Save | Run | Source

```r
# g.Create a column that contains the average ratings for each topping. (Hi
pizza_data$average_rating <- rowMeans(pizza_data[, -1], na.rm = TRUE)
pizza_data


# 3.
# a. Examine the raw data file Hotel.csv and read it into R using fread().
hotel_data <- fread("C:/Users/hardi/OneDrive/Desktop/Hotel.csv")
hotel_data <- fread("C:/Users/hardi/OneDrive/Desktop/Hotel.csv", fill= TRUE
hotel_data

# b. Assign the column names for room number and number of guests first. Fo
# Assign column names for room number and number of guests
# Assign column names for specific columns
names(hotel_data)[1] <- "RoomNumber"
```

81:1 (Top Level) R Script

Console | Terminal | Background Jobs

R 4.3.2 · ~/

```
177: Basic no view    75
178:         Suite   255
179:           155    NA
> # 3.
> # a. Examine the raw data file Hotel.csv and read it into R using fread(). Is t
here any “problem” with this data read? Explain.
> hotel_data <- fread("C:/Users/hardi/OneDrive/Desktop/Hotel.csv")
Warning message:
In fread("C:/Users/hardi/OneDrive/Desktop/Hotel.csv") :
  Stopped early on line 4. Expected 11 fields but found 12. Consider fill=TRUE an
d comment.char=. First discarded non-empty line: <<220,5,2,3,2014,2,12,2014,YES,
2,Basic w/view,155>>
>
```

Answers Project 1.R*

```r
# g.Create a column that contains the average ratings for each topping. (Hi
pizza_data$average_rating <- rowMeans(pizza_data[, -1], na.rm = TRUE)
pizza_data


# 3.
# a. Examine the raw data file Hotel.csv and read it into R using fread().
hotel_data <- fread("C:/Users/hardi/OneDrive/Desktop/Hotel.csv")
hotel_data <- fread("C:/Users/hardi/OneDrive/Desktop/Hotel.csv", fill= TRUE
hotel_data

# b. Assign the column names for room number and number of guests first. Fo
# Assign column names for room number and number of guests
# Assign column names for specific columns
names(hotel_data)[1] <- "RoomNumber"
```

82:11 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.2 · ~/

```
       V1    V2    V3    V4    V5    V6    V7    V8     V9           V10
    <int> <int> <int> <int> <int> <int> <int> <int> <char>        <char>
  1:   211     3     2     7  2014     2    11  2014     NO  Deluxe Suite
  2:   214     2     2     2  2014     2    12  2014     NO Basic no view
  3:   216     4     2     2  2014     2    13  2014     NO         Suite
  4:   220     5     2     3  2014     2    12  2014    YES             2
  5:   221     3     2     3  2014     2    12  2014     NO        Luxury
 ---
175:  1276     2     2     1  2014     2    11  2014    YES             3
176:  1285     5     1    31  2014     2    11  2014     NO  Deluxe Suite
177:  1291     1     2     6  2014     2    10  2014    YES             4
178:  1294     6     2     8  2014     2    13  2014    YES             5
179:  1298     7     2     7  2014     2    11  2014     NO  Basic w/view
```

b. Assign the column names for room number and number of guests first. For other column names, you should assign them as you answer the remaining questions.

Answer 3b.

```r
# Assign column names for room number and number of guests
# Assign column names for specific columns
names(hotel_data)[1] <- "RoomNumber"
names(hotel_data)[2] <- "NumberOfGuests"
```

Answers Project 1.R*

```r
# b. Assign the column names for room number and number of guests first. Fo
# Assign column names for room number and number of guests
# Assign column names for specific columns
names(hotel_data)[1] <- "RoomNumber"
names(hotel_data)[2] <- "NumberOfGuests"
hotel_data
# c.Create date variables for the check-in and check-out dates, and format
# Read the data without specifying column names
hotel_data
names(hotel_data)[3] <- "monthin"
names(hotel_data)[4]<-"dayin"
names(hotel_data)[5]<-"yearin"
names(hotel_data)[6]<-"monthout"
names(hotel_data)[7]<-"dayout"
```

89:11 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.2 · ~/

```
> hotel_data
     RoomNumber NumberOfGuests monthin dayin yearin monthout dayout yearout
          <int>          <int>   <int> <int>  <int>    <int>  <int>   <int>
  1:        211              3       2     7   2014        2     11    2014
  2:        214              2       2     2   2014        2     12    2014
  3:        216              4       2     2   2014        2     13    2014
  4:        220              5       2     3   2014        2     12    2014
  5:        221              3       2     3   2014        2     12    2014
 ---
175:       1276              2       2     1   2014        2     11    2014
176:       1285              5       1    31   2014        2     11    2014
177:       1291              1       2     6   2014        2     10    2014
178:       1294              6       2     8   2014        2     13    2014
179:       1298              7       2     7   2014        2     11    2014
```

c.Create date variables for the check-in and check-out dates, and format them to display as readable dates.

Answer 3c.

```r
# Read the data without specifying column names
hotel_data
names(hotel_data)[3] <- "monthin"
names(hotel_data)[4]<-"dayin"
names(hotel_data)[5]<-"yearin"
names(hotel_data)[6]<-"monthout"
names(hotel_data)[7]<-"dayout"
names(hotel_data)[8]<-"yearout"

# Combine check-in date components into a single date variable
```

hotel_data

hotel_data$checkindate <- as.Date(with(hotel_data, paste(monthin, dayin, yearin,sep="-")), "%m-%d-%Y")
hotel_data$checkoutdate <- as.Date(with(hotel_data, paste(monthout, dayout, yearout,sep="-")),"%m-%d-%Y")

hotel_data

Answers Project 1.R*

```
names(hotel_data)[8]<-"yearout"

# Combine check-in date components into a single date variable

hotel_data

hotel_data$checkindate <- as.Date(with(hotel_data, paste(monthin, dayin, ye
hotel_data$checkoutdate <- as.Date(with(hotel_data, paste(monthout, dayout,

hotel_data

#d.Using the data.table syntax, create a column of days of internet use. If

names(hotel_data)[9] <- "InternetUse"
```

109:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.2 · ~/

```
175:       1276              2       2     1   2014         2     11    2014
176:       1285              5       1    31   2014         2     11    2014
177:       1291              1       2     6   2014         2     10    2014
178:       1294              6       2     8   2014         2     13    2014
179:       1298              7       2     7   2014         2     11    2014
         V9           V10           V11   V12 checkindate checkoutdate
     <char>        <char>        <char> <int>      <Date>       <Date>
  1:     NO  Deluxe Suite           295    NA  2014-02-07   2014-02-11
  2:     NO Basic no view            75    NA  2014-02-02   2014-02-12
  3:     NO         Suite           255    NA  2014-02-02   2014-02-13
  4:    YES             2  Basic w/view   155  2014-02-03   2014-02-12
  5:     NO        Luxury           195    NA  2014-02-03   2014-02-12
 ---
175:    YES             3  Basic w/view   155  2014-02-01   2014-02-11
```

#d.Using the data.table syntax, create a column of days of internet use. If the guest did not use the internet, assign “0”. Check the class of the column you created and coerce the variable type

to “numeric” as necessary. (Hint. Days of internet use is recorded only when the use of wireless internet service is YES. See the documentation for as.numeric() and as.character())

Answer 3d.

```
names(hotel_data)[9] <- "InternetUse"

hotel_data[, DaysOfInternetUse := 0]

# Assuming the column names for internet use and days of internet use are    'InternetUse' and '
DaysOfInternetUse' respectively
hotel_data[, DaysOfInternetUse := ifelse(InternetUse == "YES", as.numeric(as.character(V10)), 0
)]

# Check the class of the column
print(class(hotel_data$DaysOfInternetUse))

# Coerce to numeric if necessary
hotel_data$DaysOfInternetUse <- as.numeric(hotel_data$DaysOfInternetUse)
hotel_data
```

Answers Project 1.R*

Source on Save | Run | Source

```r
#d.Using the data.table syntax, create a column of days of internet use. If

names(hotel_data)[9] <- "InternetUse"

hotel_data[, DaysOfInternetUse := 0]

# Assuming the column names for internet use and days of internet use are '
hotel_data[, DaysOfInternetUse := ifelse(InternetUse == "YES", as.numeric(a

# Check the class of the column
print(class(hotel_data$DaysOfInternetUse))

# Coerce to numeric if necessary
hotel_data$DaysOfInternetUse <- as.numeric(hotel_data$DaysOfInternetUse)
hotel_data
# e.Using the data.table syntax, create a column of room type. (Again, use
```

124:25 (Top Level) R Script

Console | Terminal | Background Jobs

R 4.3.2 · ~/

```
179:       1298              7       2     7   2014        2     11    2014
     InternetUse           V10           V11   V12 checkindate checkoutdate
          <char>        <char>        <char> <int>      <Date>       <Date>
  1:          NO  Deluxe Suite           295    NA  2014-02-07   2014-02-11
  2:          NO Basic no view            75    NA  2014-02-02   2014-02-12
  3:          NO         Suite           255    NA  2014-02-02   2014-02-13
  4:         YES             2  Basic w/view   155  2014-02-03   2014-02-12
  5:          NO        Luxury           195    NA  2014-02-03   2014-02-12
 ---
175:         YES             3  Basic w/view   155  2014-02-01   2014-02-11
176:          NO  Deluxe Suite           295    NA  2014-01-31   2014-02-11
177:         YES             4 Basic no view    75  2014-02-06   2014-02-10
178:         YES             5         Suite   255  2014-02-08   2014-02-13
```

Answers Project 1.R*

```r
#d.Using the data.table syntax, create a column of days of internet use. If

names(hotel_data)[9] <- "InternetUse"

hotel_data[, DaysOfInternetUse := 0]

# Assuming the column names for internet use and days of internet use are '
hotel_data[, DaysOfInternetUse := ifelse(InternetUse == "YES", as.numeric(a

# Check the class of the column
print(class(hotel_data$DaysOfInternetUse))

# Coerce to numeric if necessary
hotel_data$DaysOfInternetUse <- as.numeric(hotel_data$DaysOfInternetUse)
hotel_data
# e.Using the data.table syntax, create a column of room type. (Again, use
```

124:25 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.2 · ~/

```
177:          YES              4 Basic no view    75  2014-02-06   2014-02-10
178:          YES              5         Suite   255  2014-02-08   2014-02-13
179:           NO   Basic w/view           155    NA  2014-02-07   2014-02-11
     DaysOfInternetUse
                 <num>
  1:                 0
  2:                 0
  3:                 0
  4:                 2
  5:                 0
 ---
175:                 3
176:                 0
177:                 4
```

e.Using the data.table syntax, create a column of room type. (Again, use the hint from the above)

Answer 3e.

```r
hotel_data[, roomtype := ""]

hotel_data[, roomtype := gsub("[0-9]", "", paste(V10, V11))]
hotel_data
```

Answers Project 1.R*

```r
print(class(hotel_data$DaysOfInternetUse))

# Coerce to numeric if necessary
hotel_data$DaysOfInternetUse <- as.numeric(hotel_data$DaysOfInternetUse)
hotel_data
# e.Using the data.table syntax, create a column of room type. (Again, use
hotel_data[, roomtype := ""]

hotel_data[, roomtype := gsub("[0-9]", "", paste(V10, V11))]
hotel_data
# f. Using the data.table syntax, create a column of room rate. Check the c
hotel_data[, roomrate := 0]

hotel_data[, roomrate := as.numeric(gsub("[^0-9.]", "", paste(V11, V12)))]

names(hotel_data)[15] <- "number_of_days_of_Internet_use"
```

128:11 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.2 · ~/

```
177:          YES             4 Basic no view    75  2014-02-06   2014-02-10
178:          YES             5         Suite   255  2014-02-08   2014-02-13
179:           NO  Basic w/view           155    NA  2014-02-07   2014-02-11
     DaysOfInternetUse       roomtype
                 <num>         <char>
  1:                 0  Deluxe Suite
  2:                 0 Basic no view
  3:                 0         Suite
  4:                 2   Basic w/view
  5:                 0        Luxury
 ---
175:                 3   Basic w/view
176:                 0  Deluxe Suite
177:                 4  Basic no view
```

f. Using the data.table syntax, create a column of room rate. Check the class of the column you created and coerce the variable type to “numeric” as necessary. (Again, use the hint from t he above)

Answer 3f.

```r
hotel_data[, roomrate := 0]

hotel_data[, roomrate := as.numeric(gsub("[^0-9.]", "", paste(V11, V12)))]

names(hotel_data)[15] <- "number_of_days_of_Internet_use"
names(hotel_data)[9] <-"use_of_wireless_Internet_service"
hotel_data
```

Answers Project 1.R*

```r
hotel_data[, roomtype := ]

hotel_data[, roomtype := gsub("[0-9]", "", paste(V10, V11))]
hotel_data
# f. Using the data.table syntax, create a column of room rate. Check the c
hotel_data[, roomrate := 0]

hotel_data[, roomrate := as.numeric(gsub("[^0-9.]", "", paste(V11, V12)))]

names(hotel_data)[15] <- "number_of_days_of_Internet_use"
names(hotel_data)[9] <-"use_of_wireless_Internet_service"
hotel_data


# g.Subset the cleaned variables only and create a new data.table: room num
# Create a new data.table with selected columns
```

136:11 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.2 · ~/

```
176:  2014-01-31   2014-02-11                          0  Deluxe Suite
177:  2014-02-06   2014-02-10                          4  Basic no view
178:  2014-02-08   2014-02-13                          5          Suite
179:  2014-02-07   2014-02-11                          0  Basic w/view
     roomrate
        <num>
  1:      295
  2:       75
  3:      255
  4:      155
  5:      195
 ---
175:      155
176:      295
```

g.Subset the cleaned variables only and create a new data.table: room number, number of guests, check-in date, check-out date, use of wireless Internet service, number of days of Intern et use, room type, and room rate.

Answer 3g.

Create a new data.table with selected columns

cleaned_hotel_data <- hotel_data[, .(RoomNumber, NumberOfGuests, checkindate, checkoutdate, use_of_wireless_Internet_service, number_of_days_of_Internet_use, roomtype, roomrate)]
cleaned_hotel_data

Answers Project 1.R*

```r
names(hotel_data)[9] <-"use_of_wireless_Internet_service"
hotel_data


# g.Subset the cleaned variables only and create a new data.table: room number, number of guests, check-in date, c
# Create a new data.table with selected columns
cleaned_hotel_data <- hotel_data[, .(RoomNumber, NumberOfGuests, checkindate, checkoutdate, use_of_wireless_Intern
cleaned_hotel_data



# Assuming cleaned_hotel_data is your data.table and you have the necessary cleaned columns: RoomRate, CheckInDate
# h. Create a variable that calculates the subtotal as the room rate times the number of days in the stay, plus a
# Calculate the number of days in the stay
cleaned_hotel_data[, DaysOfStay := as.numeric(difftime(checkoutdate, checkindate, units = "days"))]
```

142:19 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.2 · ~/

```
> cleaned_hotel_data <- hotel_data[, .(RoomNumber, NumberOfGuests, checkindate, checkoutdate, use_of_wireless_Internet_s
ervice, number_of_days_of_Internet_use, roomtype, roomrate)]
> cleaned_hotel_data
     RoomNumber NumberOfGuests checkindate checkoutdate
          <int>          <int>      <Date>       <Date>
  1:        211              3  2014-02-07   2014-02-11
  2:        214              2  2014-02-02   2014-02-12
  3:        216              4  2014-02-02   2014-02-13
  4:        220              5  2014-02-03   2014-02-12
  5:        221              3  2014-02-03   2014-02-12
 ---
175:       1276              2  2014-02-01   2014-02-11
176:       1285              5  2014-01-31   2014-02-11
```

Answers Project 1.R*

```r
names(hotel_data)[9] <-"use_of_wireless_Internet_service"
hotel_data


# g.Subset the cleaned variables only and create a new data.table: room number, number of guests, check-in date, c
# Create a new data.table with selected columns
cleaned_hotel_data <- hotel_data[, .(RoomNumber, NumberOfGuests, checkindate, checkoutdate, use_of_wireless_Intern
cleaned_hotel_data



# Assuming cleaned_hotel_data is your data.table and you have the necessary cleaned columns: RoomRate, CheckInDate
# h. Create a variable that calculates the subtotal as the room rate times the number of days in the stay, plus a
# Calculate the number of days in the stay
cleaned_hotel_data[, DaysOfStay := as.numeric(difftime(checkoutdate, checkindate, units = "days"))]
```

142:19 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.2 · ~/

```
179:        1298              7  2014-02-07   2014-02-11
     use_of_wireless_Internet_service number_of_days_of_Internet_use
                               <char>                          <num>
  1:                               NO                              0
  2:                               NO                              0
  3:                               NO                              0
  4:                              YES                              2
  5:                               NO                              0
 ---
175:                              YES                              3
176:                               NO                              0
177:                              YES                              4
178:                              YES                              5
179:                               NO                              0
```

Answers Project 1.R*

```r
names(hotel_data)[9] <-"use_of_wireless_Internet_service"
hotel_data


# g.Subset the cleaned variables only and create a new data.table: room number, number of guests, check-in date, c
# Create a new data.table with selected columns
cleaned_hotel_data <- hotel_data[, .(RoomNumber, NumberOfGuests, checkindate, checkoutdate, use_of_wireless_Intern
cleaned_hotel_data



# Assuming cleaned_hotel_data is your data.table and you have the necessary cleaned columns: RoomRate, CheckInDate
# h. Create a variable that calculates the subtotal as the room rate times the number of days in the stay, plus a
# Calculate the number of days in the stay
cleaned_hotel_data[, DaysOfStay := as.numeric(difftime(checkoutdate, checkindate, units = "days"))]
```

142:19 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.2 · ~/

```
178:                              YES                              5
179:                               NO                              0
```

```
           roomtype roomrate
             <char>    <num>
  1:  Deluxe Suite       295
  2: Basic no view        75
  3:         Suite       255
  4:   Basic w/view      155
  5:        Luxury       195
 ---
175:   Basic w/view      155
176:  Deluxe Suite       295
177:  Basic no view       75
```

Assuming cleaned_hotel_data is your data.table and you have the necessary cleaned columns : RoomRate, CheckInDate, CheckOutDate, NumberOfGuests, and DaysOfInternetUse

h. Create a variable that calculates the subtotal as the room rate times the number of days in the stay, plus a per person rate ($10 per day for each person beyond one guest), plus an Internet service fee ($9.95 for a one-time activation and $5.95 per day of use).

Answer 3h .

```
# Calculate the number of days in the stay
cleaned_hotel_data[, DaysOfStay := as.numeric(difftime(checkoutdate, checkindate, units = "da
ys"))]

# Calculate the total room rate for the stay
cleaned_hotel_data[, TotalRoomRate := roomrate * DaysOfStay]

# Calculate the additional charges for extra guests
cleaned_hotel_data[, ExtraGuestCharges := 10 * (NumberOfGuests - 1) * DaysOfStay]

# Calculate the internet service fee
cleaned_hotel_data[, InternetServiceFee := 9.95 + 5.95 * number_of_days_of_Internet_use]

# Calculate the subtotal
cleaned_hotel_data[, Subtotal := TotalRoomRate + ExtraGuestCharges + InternetServiceFee]
cleaned_hotel_data
```

Answers Project 1.R*

```r
cleaned_hotel_data[, TotalRoomRate := roomrate * DaysOfStay]

# Calculate the additional charges for extra guests
cleaned_hotel_data[, ExtraGuestCharges := 10 * (NumberOfGuests - 1) * DaysOfStay]

# Calculate the internet service fee
cleaned_hotel_data[, InternetServiceFee := 9.95 + 5.95 * number_of_days_of_Internet_us

# Calculate the subtotal
cleaned_hotel_data[, Subtotal := TotalRoomRate + ExtraGuestCharges + InternetServiceFe
cleaned_hotel_data


# Assuming cleaned_hotel_data is your data.table and you have the Subtotal column
# i. Create a variable that calculates the grand total as the subtotal plus sales tax
```

162:19 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.2 · ~/

```
     ExtraGuestCharges InternetServiceFee Subtotal
                 <num>              <num>    <num>
  1:                80               9.95  1269.95
  2:               100               9.95   859.95
  3:               330               9.95  3144.95
  4:               360              21.85  1776.85
  5:               180               9.95  1944.95
 ---
175:               100              27.80  1677.80
176:               440               9.95  3694.95
177:                 0              33.75   333.75
178:               250              39.70  1564.70
179:               240               9.95   869.95
```

Assuming cleaned_hotel_data is your data.table and you have the Subtotal column
i. Create a variable that calculates the grand total as the subtotal plus sales tax at 8.75%. The result should be rounded to two decimal places.

Answer 3i.

```r
# Calculate the sales tax
cleaned_hotel_data[, SalesTax := Subtotal * 0.0875]

# Calculate the grand total
cleaned_hotel_data[, GrandTotal := Subtotal + SalesTax]

# Round the grand total to two decimal places
cleaned_hotel_data[, GrandTotal := round(GrandTotal, 2)]
cleaned_hotel_data
```

Answers Project 1.R*

```r
# Assuming cleaned_hotel_data is your data.table and you have the Subtotal column
# i. Create a variable that calculates the grand total as the subtotal plus sales tax
# Calculate the sales tax
cleaned_hotel_data[, SalesTax := Subtotal * 0.0875]

# Calculate the grand total
cleaned_hotel_data[, GrandTotal := Subtotal + SalesTax]

# Round the grand total to two decimal places
cleaned_hotel_data[, GrandTotal := round(GrandTotal, 2)]
cleaned_hotel_data


# j. View the resulting data set. In a comment in your report, state the value for the
```

175:19 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.2 · ~/

```
178:                              5          Suite       255           5          1275
179:                              0  Basic w/view       155           4           620
     ExtraGuestCharges InternetServiceFee Subtotal  SalesTax GrandTotal
                 <num>              <num>    <num>     <num>      <num>
  1:                80               9.95  1269.95 111.12063    1381.07
  2:               100               9.95   859.95  75.24563     935.20
  3:               330               9.95  3144.95 275.18312    3420.13
  4:               360              21.85  1776.85 155.47437    1932.32
  5:               180               9.95  1944.95 170.18312    2115.13
 ---
175:               100              27.80  1677.80 146.80750    1824.61
176:               440               9.95  3694.95 323.30812    4018.26
177:                 0              33.75   333.75  29.20312     362.95
178:               250              39.70  1564.70 136.91125    1701.61
```

j. View the resulting data set. In a comment in your report, state the value for the grand total for room 247, checked in on Feb. 7th, 2014.

Answer 3j.

```r
# Calculate the grand total for room 247, checked in on Feb. 7th, 2014
grand_total_247 <- cleaned_hotel_data[RoomNumber == 247 & checkindate == "2014-02-07"]$GrandTotal

# Print the grand total for room 247
cat("Grand total for room 247, checked in on Feb. 7th, 2014:", grand_total_247, "\n")
```

Answers Project 1.R*

Source on Save | Run | Source

```r
cleaned_hotel_data[, GrandTotal := Subtotal + SalesTax]

# Round the grand total to two decimal places
cleaned_hotel_data[, GrandTotal := round(GrandTotal, 2)]
cleaned_hotel_data


# j. View the resulting data set. In a comment in your report, state the value for the
# Calculate the grand total for room 247, checked in on Feb. 7th, 2014
grand_total_247 <- cleaned_hotel_data[RoomNumber == 247 & checkindate == "2014-02-07"]

# Print the grand total for room 247
cat("Grand total for room 247, checked in on Feb. 7th, 2014:", grand_total_247, "\n")
```

184:1 (Top Level) R Script

Console | Terminal | Background Jobs

R 4.3.2 · ~/

```
177:                 0              33.75   333.75  29.20312     362.95
178:               250              39.70  1564.70 136.91125    1701.61
179:               240               9.95   869.95  76.12063     946.07
> # j. View the resulting data set. In a comment in your report, state the value for the gran
d total for room 247, checked in on Feb. 7th, 2014.
> # Calculate the grand total for room 247, checked in on Feb. 7th, 2014
> grand_total_247 <- cleaned_hotel_data[RoomNumber == 247 & checkindate == "2014-02-07"]$Gran
dTotal
>
> # Print the grand total for room 247
> cat("Grand total for room 247, checked in on Feb. 7th, 2014:", grand_total_247, "\n")
Grand total for room 247, checked in on Feb. 7th, 2014: 493.45
>
```