**JAIPURIA INSTITUTE OF MANAGEMENT, INDORE**

**PGDM**

**THIRD TRIMESTER (Batch 2020-22)**

**END TERM EXAMINATION, MAY-2021**

|  |  |  |  |
| --- | --- | --- | --- |
| Course Name | **Fundamentals of Business Analytics** | Course Code | **IT302** |
| Max. Time | **2 hours** | Max. Marks | **40** |

**INSTRUCTIONS:**

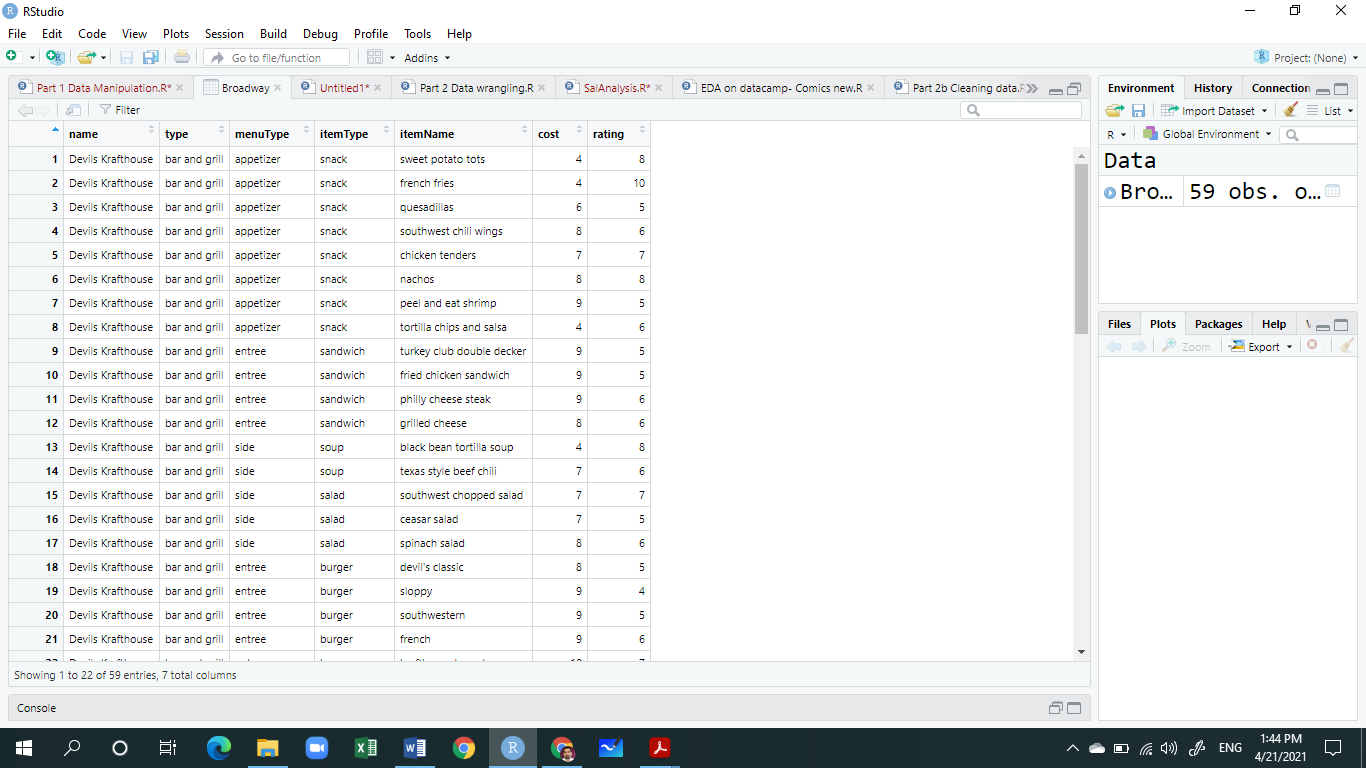
1. Attempt all questions of Part A, B, and C.
2. Use “R” software to complete all the steps, write the code wherever it has been asked.
3. To submit the answers,

After attempting the questions in R or word document, submit both the files on Moodle- as separate link will be provided 30 minutes before the end of the examination

1. Data sets for the parts - A and B shall be provided before the start of the examination.

**Part-A: Data Manipulation (Max Marks-12)**

About dataset: This data is about a place (called Broadway Centre) where multiple different types of hotels are serving different types of items under various menu types. The cost and rating for each item served at various hotels are also given.



**Research Objective:**

To explore the given dataset (***broadway.csv***) to answer the relevant questions.

**Preparation steps**

1. Download dataset first (e.g. xyz.csv) into your working directory

2. Set working directory: setwd("<enter the fullpath>")

3. Run library(tidyverse)

4. Read the file: Broadway <- read\_csv("broadway.csv")

**Answer the following questions by writing appropriate code in R:**

1. Look at the structure of the Broadway object. (***Max 3 Marks***)
2. How many observations (rows) and variables (columns) are there?
3. How many of the variables are numeric data? Name them

**OR**

What is the data type of the following relevant variables?***. Do the necessary conversions in data types for further exploration.***

* 1. menuType :
  2. itemName :
  3. rating :
  4. cost :

1. Which hotel has the most expensive item(s)? What are those item(s)? (***Max 3 Marks***)

**OR**

Which hotel has the lowest cost item and what is the item name?

1. At the Broadway Center, how many of the entrees (found in the menuType variable) cost eight dollars? (***Max 3 Marks***)

**OR**

The head of Broadway is considering reducing prices at the Broadway Center. Create a new variable (halfPrice ) that contains items at half price.

1. How many different types of menu are in the dataset (use menuType variable)? Also, show the count of each type? (***Max 3 Marks***)

**OR**

Finally, summarize the observations in the Broadway Center dataset to show count, mean rating, minimum cost and maximum cost for each hotel (Use group\_by and summarize).

**Set1: (year=2014)**

**Part B: Data Wrangling/Cleaning (Max marks -18)**

About dataset: The weather dataset is a messy, real-world dataset containing an entire year’s worth of weather data from Boston, USA.

**Research Objective**: **Convert the given messy data set to a tidy one for data analysis.**

**Explore the raw Data**

Download the weather.rds file and place it in working directory. Then read it using the following functions:

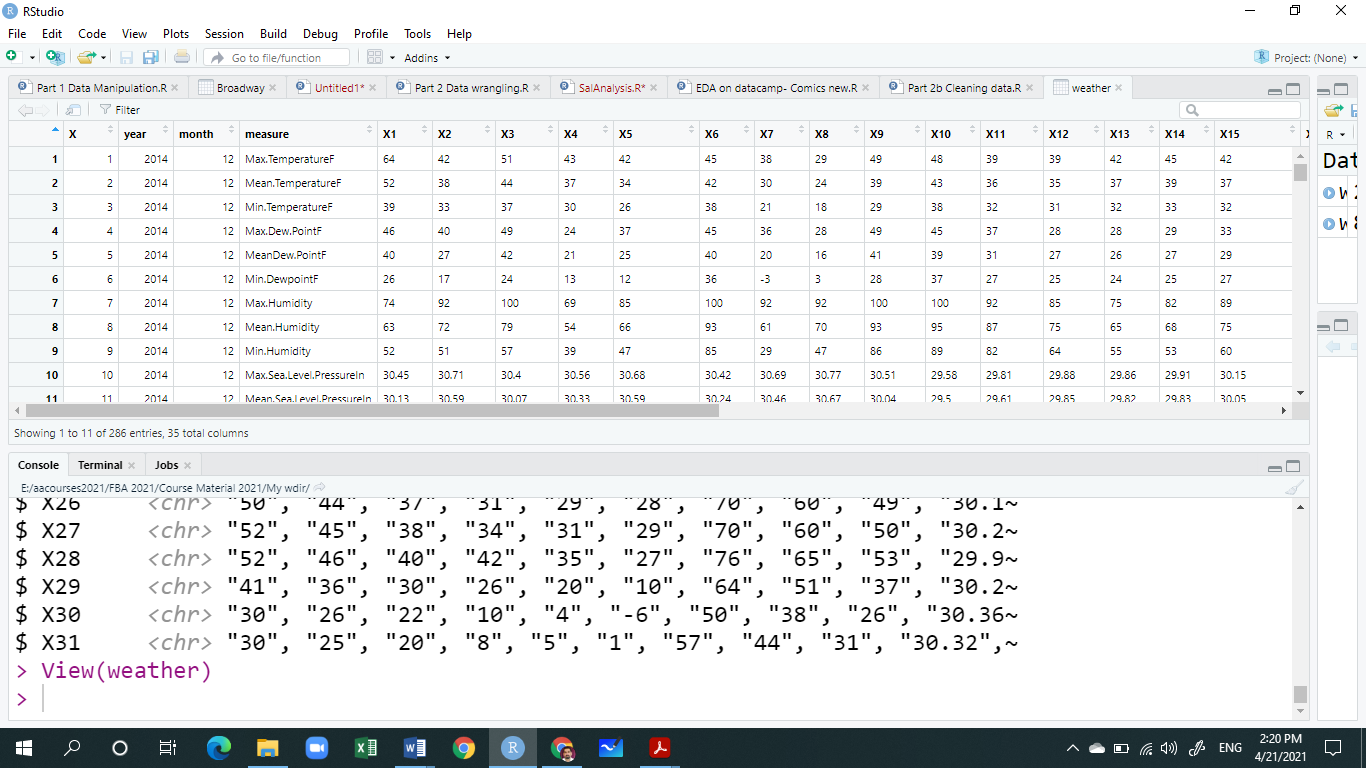
setwd("write complete path of your working directory")

weather <- readRDS("weather.rds")

Select the rows for the year 2014 using

Weather <- weather[weather$year == 2014,]

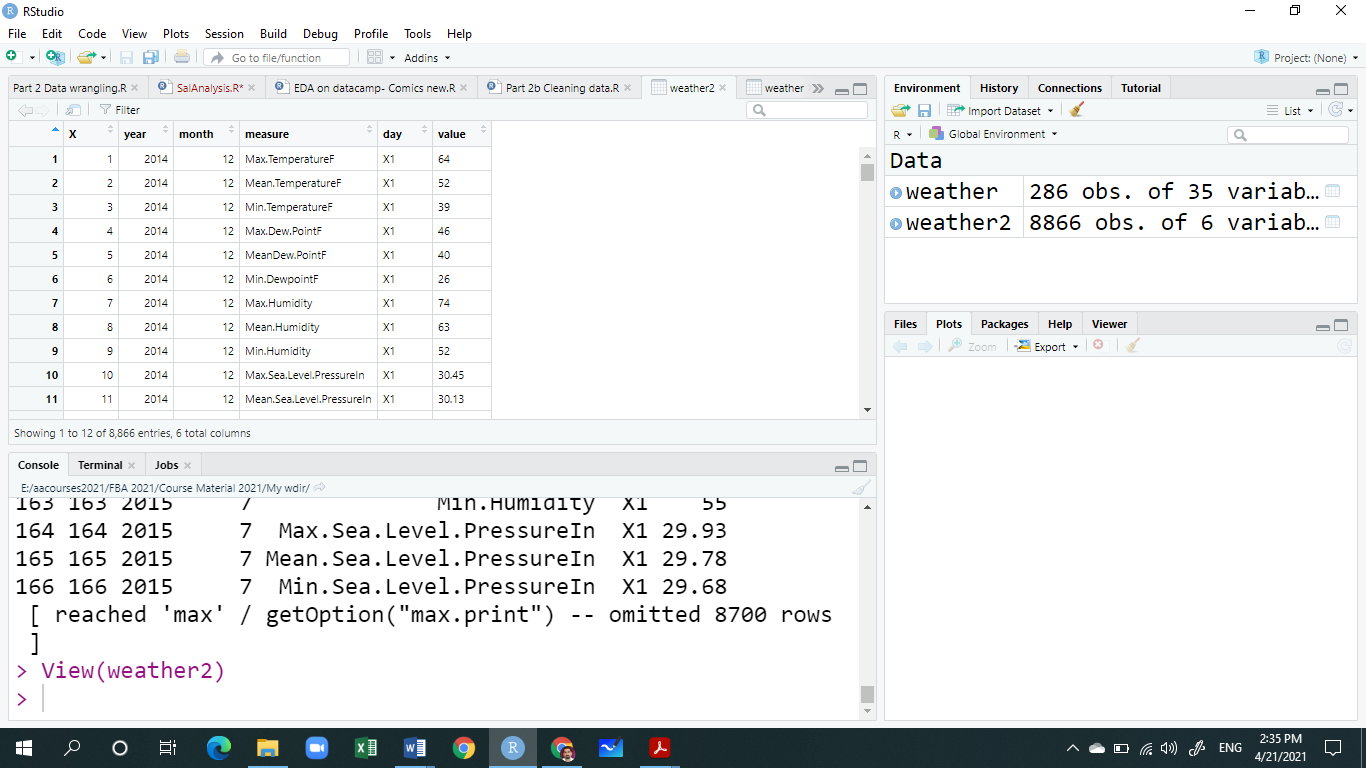
1. Look at the spreadsheet view of the data to answer the following questions. (***Max. 3 Marks***)



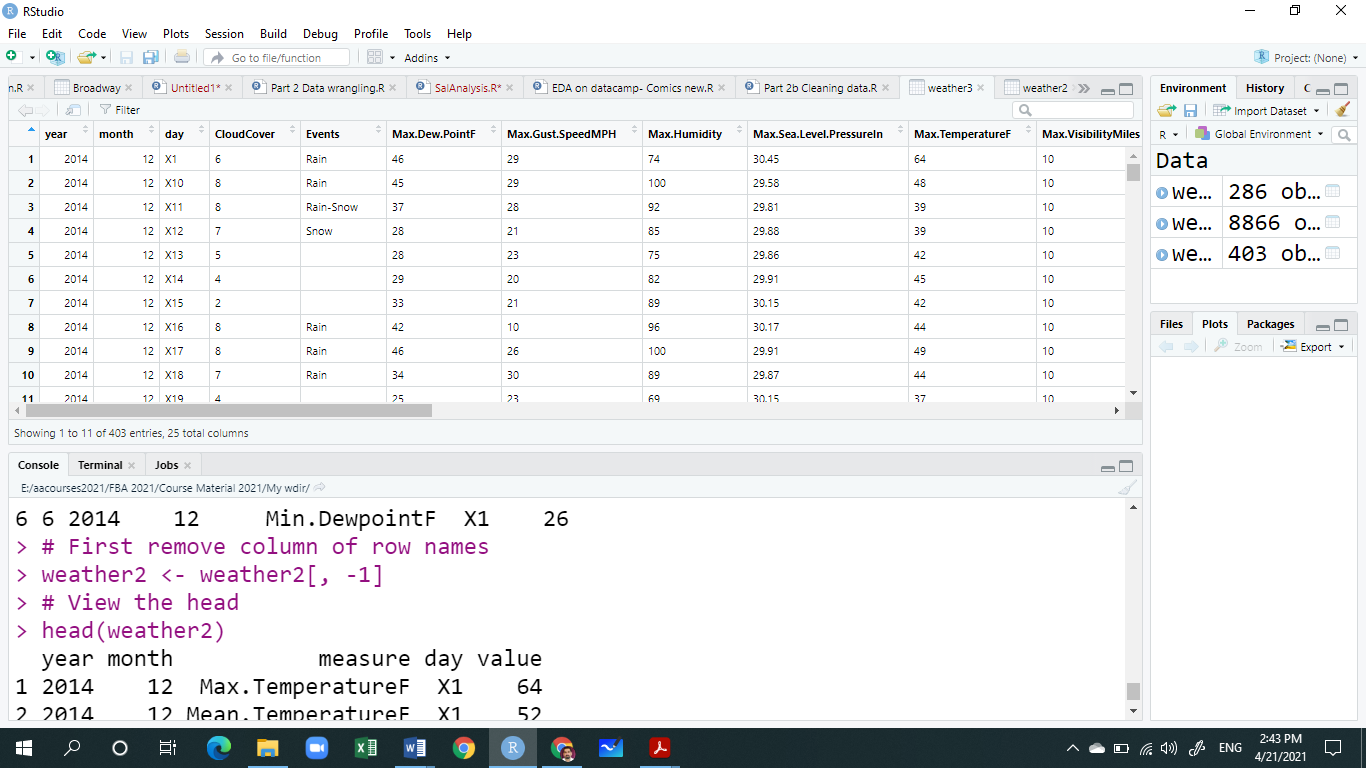
1. How many observations and variables are in this dataset?
2. Which all variables are actually values?
3. How many variables are represented in a column “measure”? Name few of them.

**Tidying data**

1. First, convert the wide data to long data using an appropriate function. The partial sample output should look like. (***Max. 3 Marks***).



1. Remove the irrelevant column. (***Max. 1 Mark***).
2. Use an appropriate function to convert the variables stored in “**measure”** column as column headers. The partial output should look like. (***Max. 3 Marks***).



**Preparing data for Analysis**

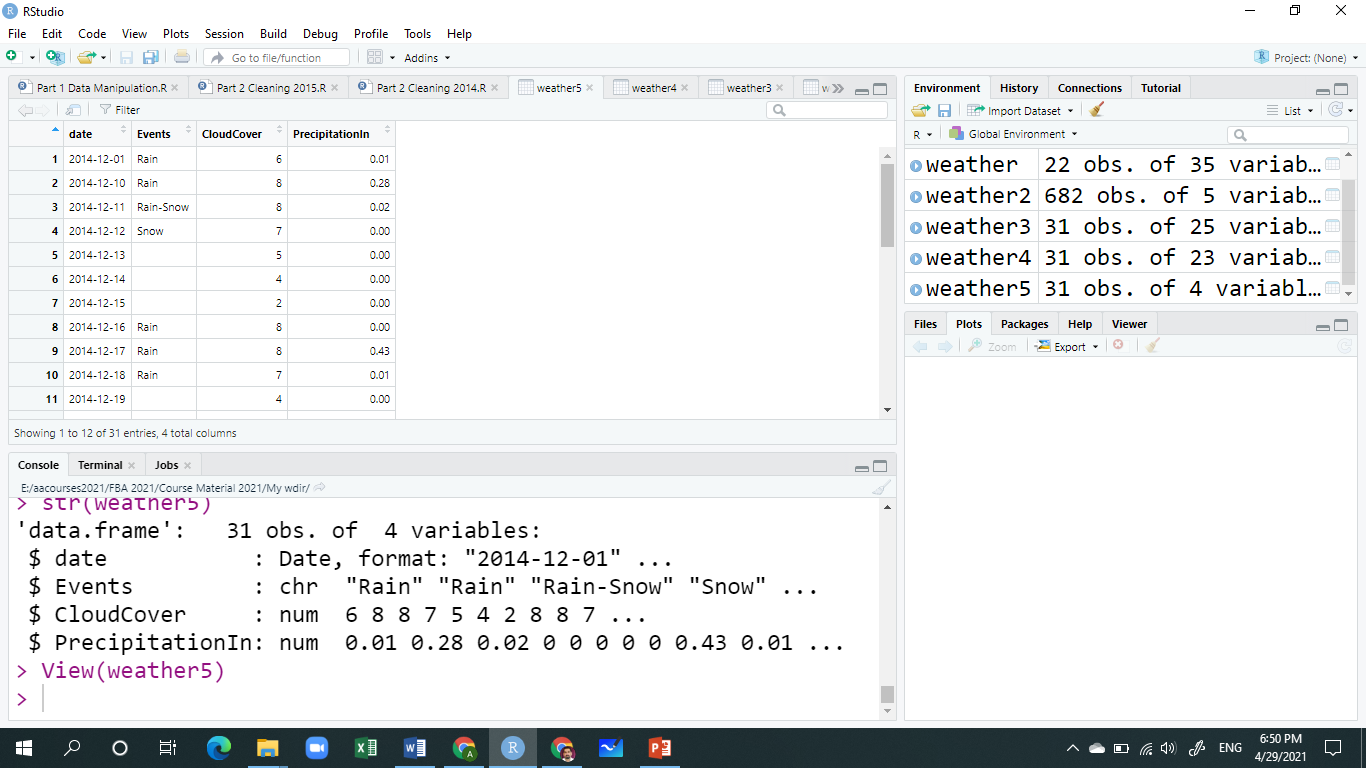
**Load the stringr and lubridate packages as**

library(stringr)

library(lubridate)

1. Remove the unnecessary leading “X” in variable **day**. (***Max. 1 Mark***).
2. Unite the three columns (**year, month and day**) to form a new column date. Use “-“ as separator. (***Max. 2 Mark***).
3. Convert date column to proper date format using lubridates's **ymd().** (***Max. 2 Mark***).
4. Select the following relevant variables (columns):--- date, Events, CoudCover and PrecipitationIn columns. (***Max. 1 Mark***).
5. Variable types are not always correct: some characters should be coerced into numerics. As In PrecipitationIn column, replace "T" with "0" (T = trace). (***Max. 1 Mark***).
6. Convert the last two variables to appropriate data types (characters to numerics). (***Max. 1 Mark***).

And the final partial output should look like this:



**Set 2: (year=2015)**

**Part B: Data Wrangling/Cleaning (Max marks -18)**

About dataset: The weather dataset is a messy, real-world dataset containing an entire year’s worth of weather data from Boston, USA.

**Research Objective**: **Convert the given messy data set to a tidy one for data analysis.**

**Explore the raw Data**

Download the weather2.rds file and place it in working directory. Then read it using the following functions:

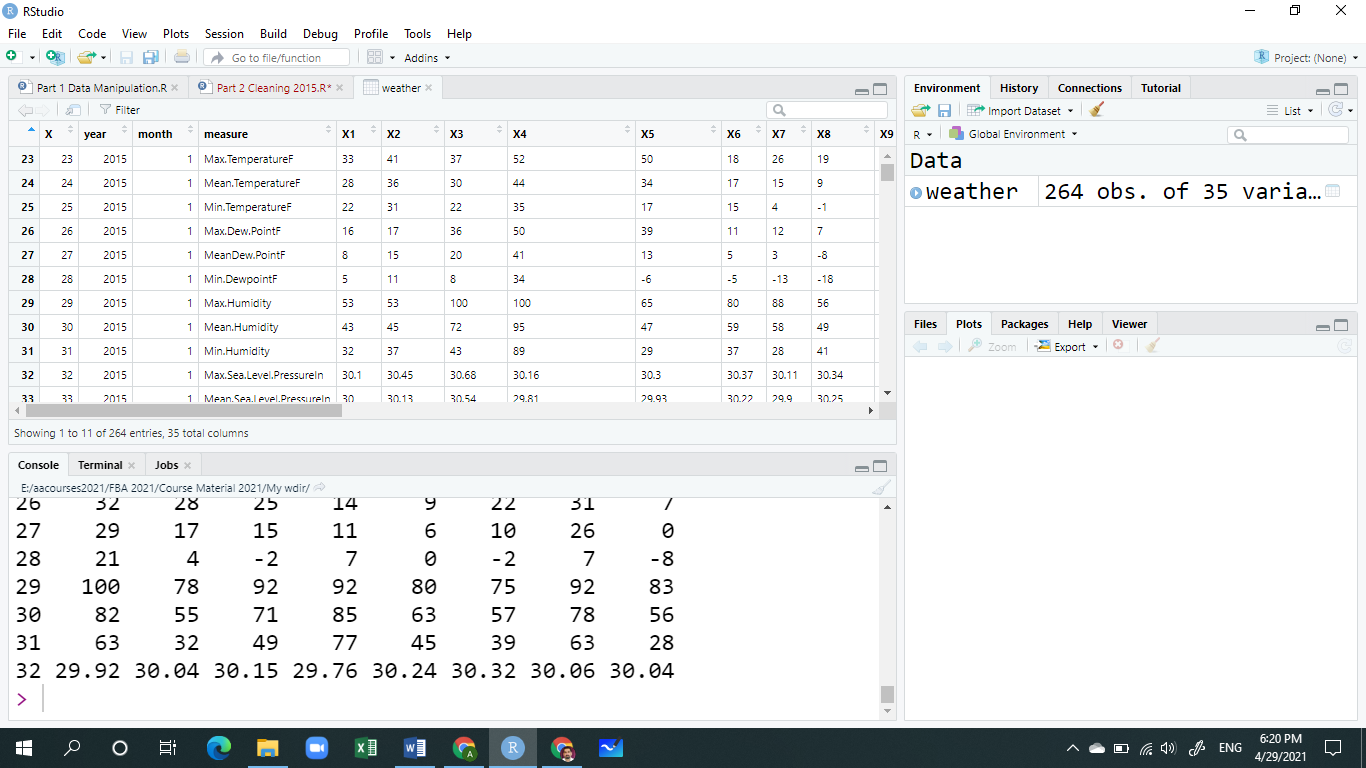
setwd("write complete path of your working directory")

weather <- readRDS("weather.rds")

Select the rows for the year 2014 using

Weather <- weather[weather$year == 2015,]

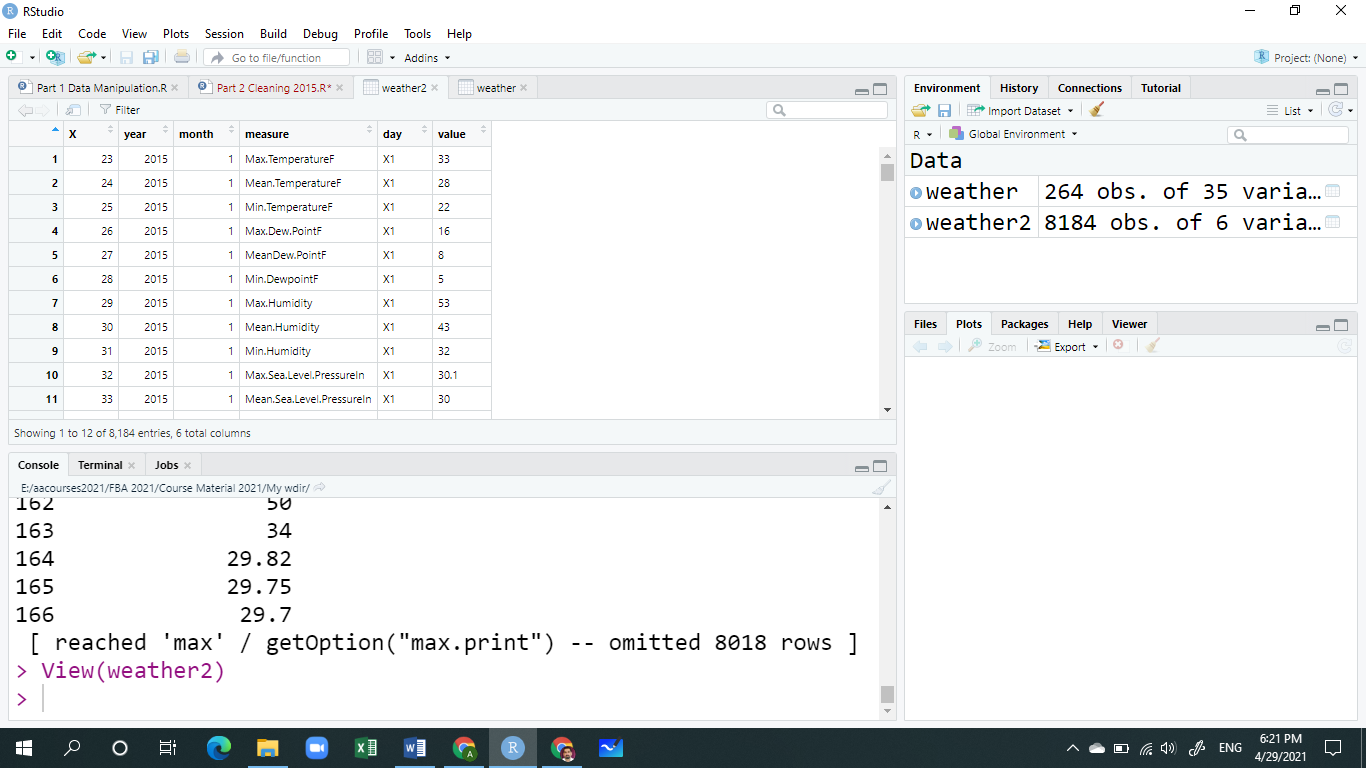
1. Look at the spreadsheet view of the data to answer the following questions. (***Max. 3 Marks***)



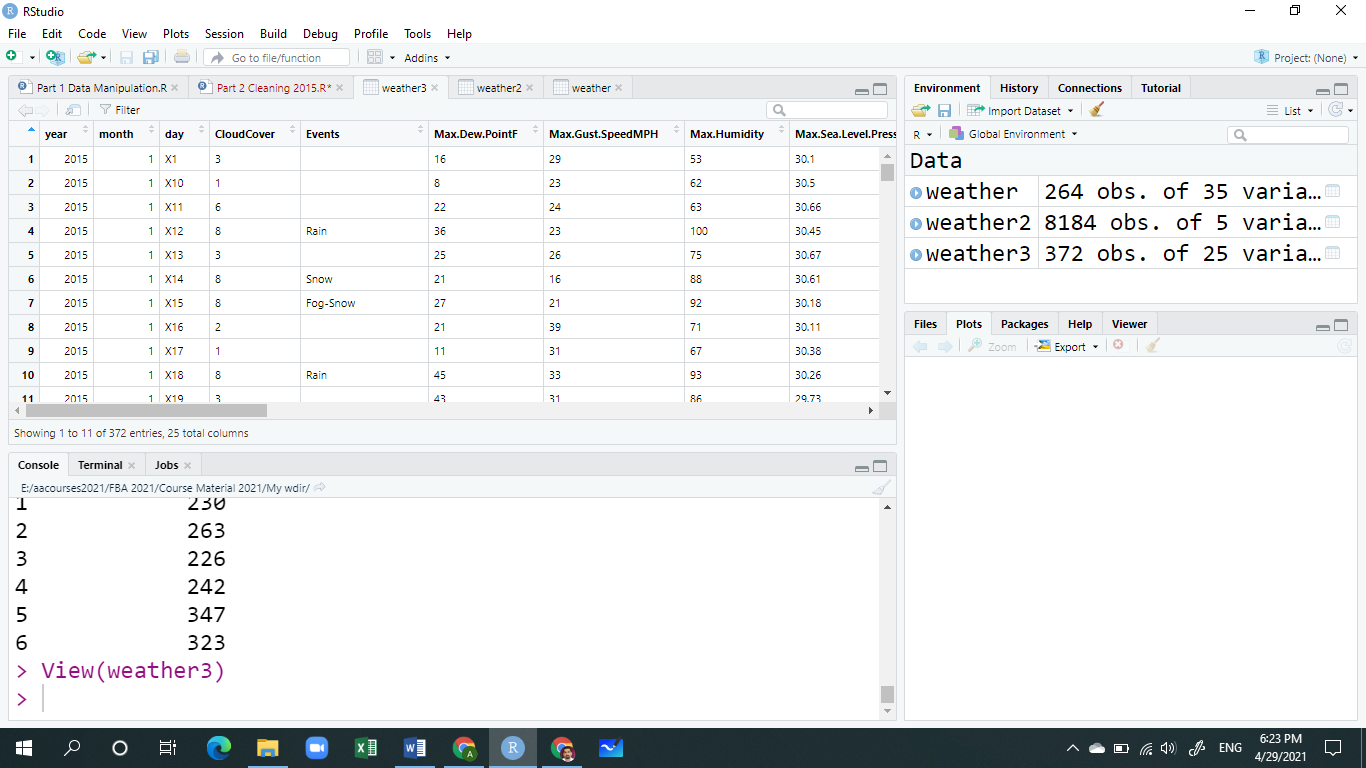
1. How many observations and variables are in this dataset?
2. Which all variables are actually values?
3. How many variables are represented in a column “measure”? Name few of them.

**Tidying data**

1. First, convert the wide data to long data using an appropriate function. The partial sample output should look like. (***Max. 3 Marks***).



1. Remove the irrelevant column. (***Max. 1 Mark***).
2. Use an appropriate function to convert the variables stored in “**measure”** column as column headers. The partial output should look like. (***Max. 3 Marks***).



**Preparing data for Analysis**

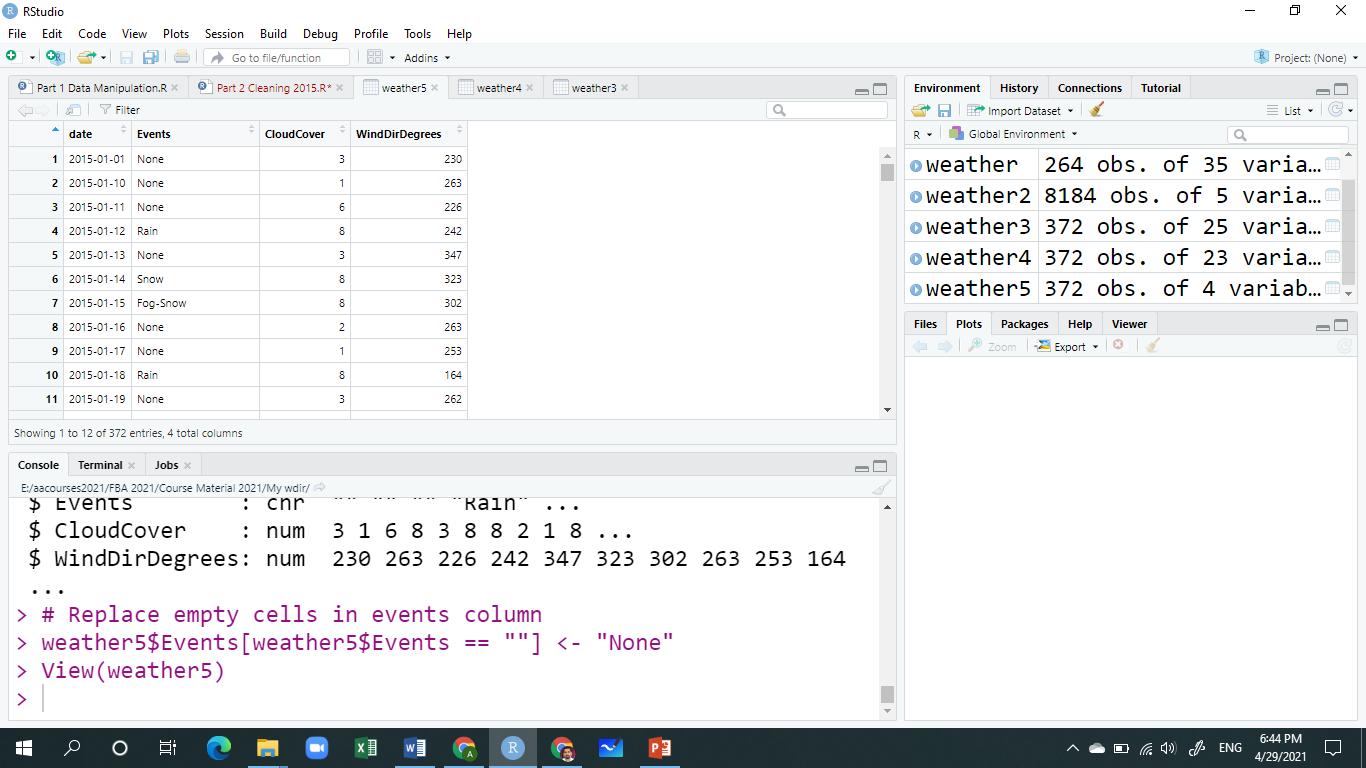
**Load the stringr and lubridate packages as**

library(stringr)

library(lubridate)

1. Remove the unnecessary leading “X” in variable **day**. (***Max. 1 Mark***).
2. Unite the three columns (**day, month and year**) to form a new column date. Use “/“ as separator. (***Max. 2 Mark***).
3. Convert date column to proper date format using lubridates's **dmy().** (***Max. 2 Mark***).
4. Select the following relevant variables (columns):--- date, Events, CoudCover and WindDirDegrees columns. (***Max. 1 Mark***).
5. Convert the last two variables to appropriate data types (characters to numerics). (***Max. 1 Mark***).
6. There are empty strings in Events column. Replace all empty cells in events column by a meaningful value such as “None”. (***Max. 1 Mark***).

And the final partial output should look like this:



**Part C: EDA (Max marks - 10)**

Write the answers of the following questions (4 + 6)

1. What are various steps followed in exploratory data analysis. (Word limit: 100 words)
2. Briefly explain how an analyst identify an outlier for a variable in a data set and what are the ways to handle it to build an robust model for predictions.

**OR**

Briefly explain how an analyst identify missing values in a variable for a given data set and what are the ways to treat missing values to build an robust model for predictions.