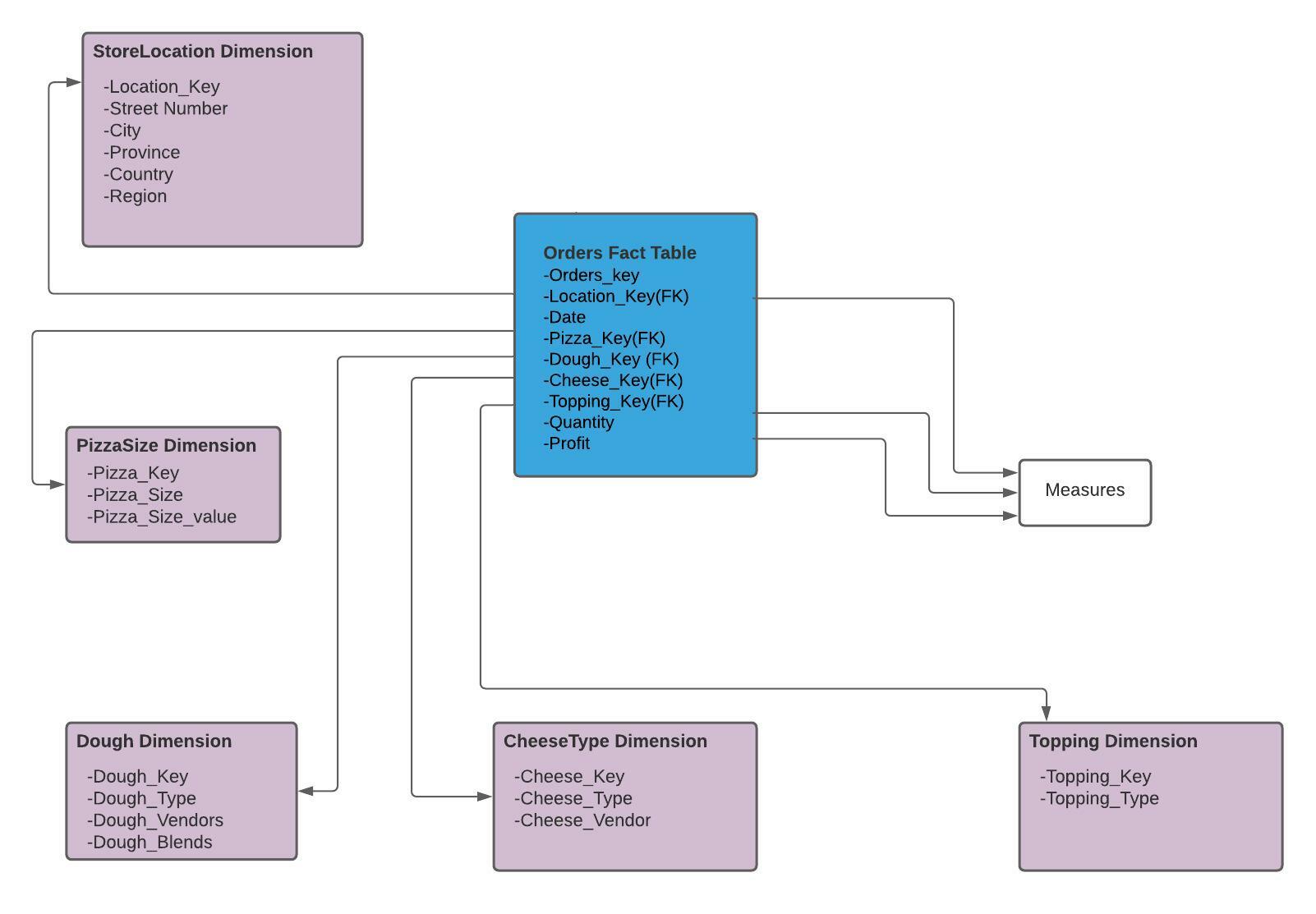
Deliverables:

1.

a. Sketch a star schema that represents this problem.

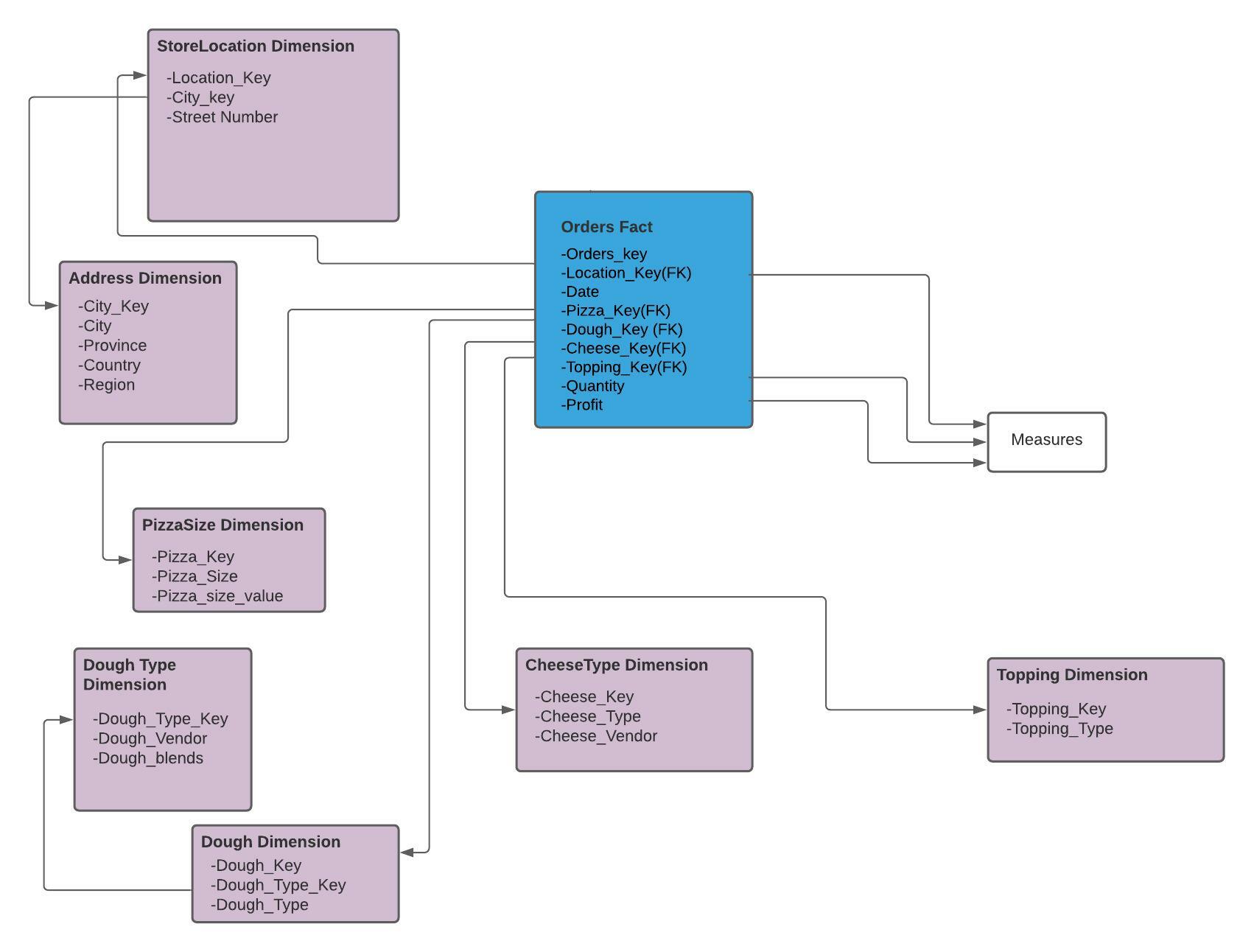
Ans) The figure below shows the star schema. The keys from the fact table are all linked to different dimension tables as shown in the figure below. There are three columns for measures, they are: Date, Quantity and Profit. Star schema is used for normalization purposes. Here it is used to organize the data.

For Store Location we used all keys that are and can be linked to location such as Street, City, Province, Country and Region. The primary key is as shown the location\_key. For the other columns we followed the same method to get as accurate as possible so that proper fact table sample data can be generated later on.



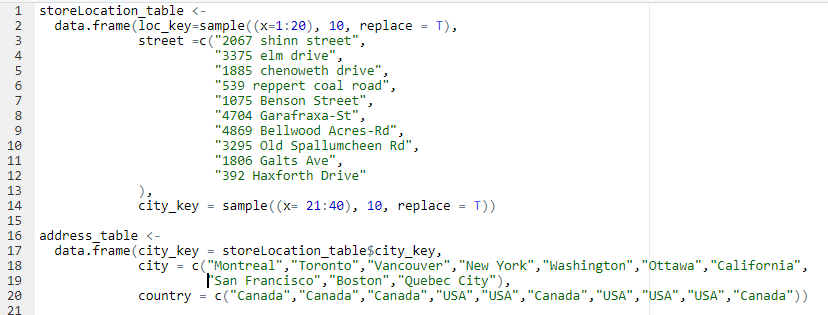
b. Sketch a snowflake schema that represents this problem.

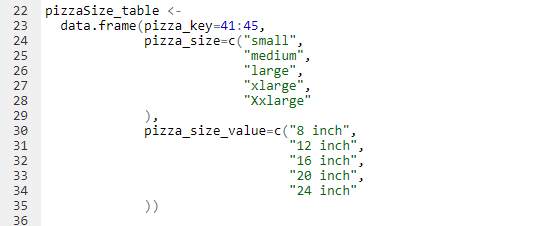
Ans) The figure below shows the Snowflake Schema. For the snowflake schema we added linked keys between dimensions such as city\_key as shown in the image below.



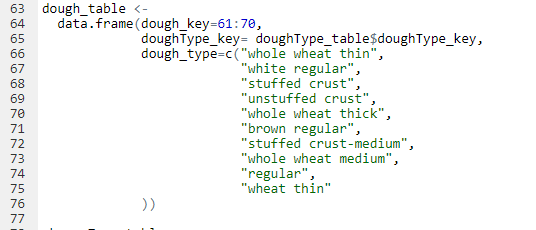
c. Generate a set of sample data stored in csv files for the dimensions and fact table for the snowflake schema in c.

Ans) For this we used data.frame() to create tables shown in the graph and stored them in csv files.

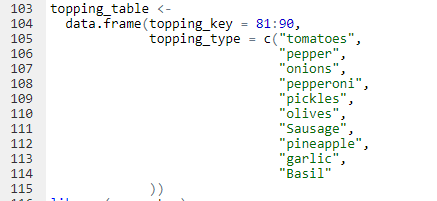




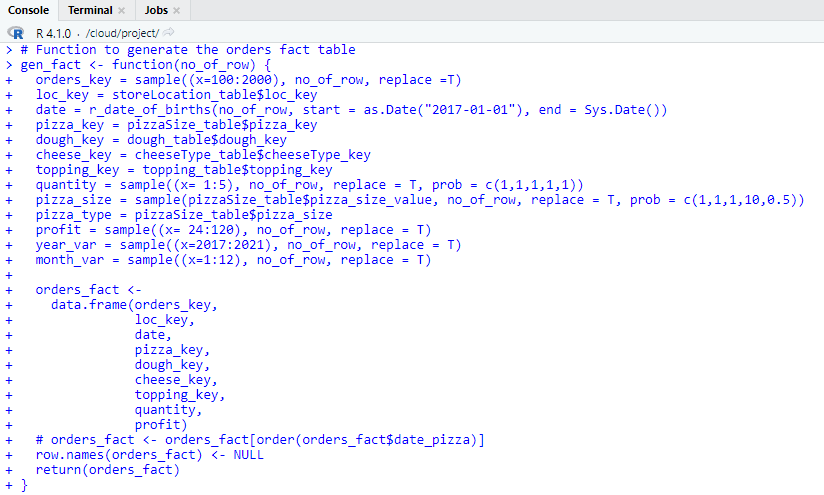








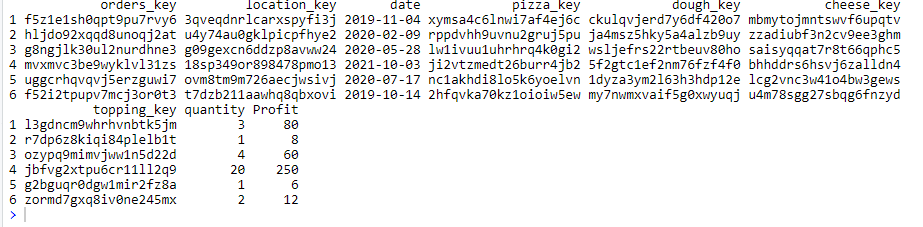
This was the function used to generate the fact table (shown from console):



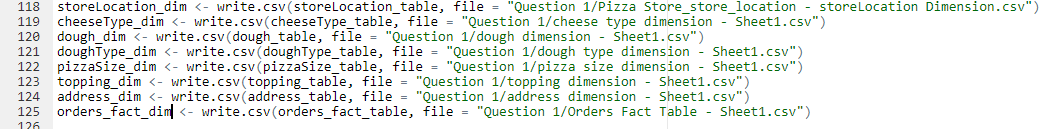
Below is the snippet of the function being applied to create the fact file:



These are some of the results from orders\_fact\_table:

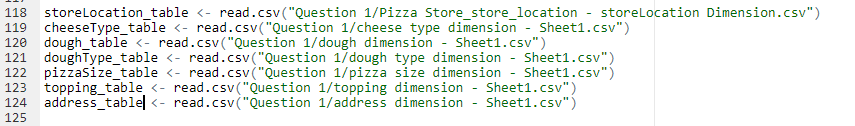


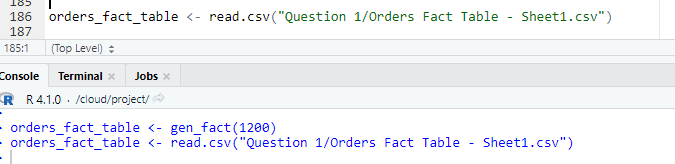
After creating the data frames, we stored them in csv files. As shown in the snippet below:



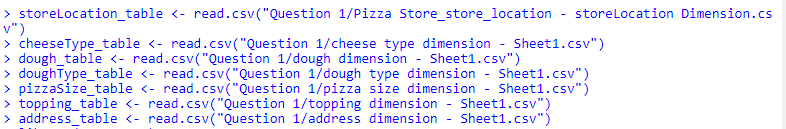
2. Using R, read the dimensions files and the profit fact table. Build an OLAP cube for your revenue and show the cells of a subset of the cells.

Ans)

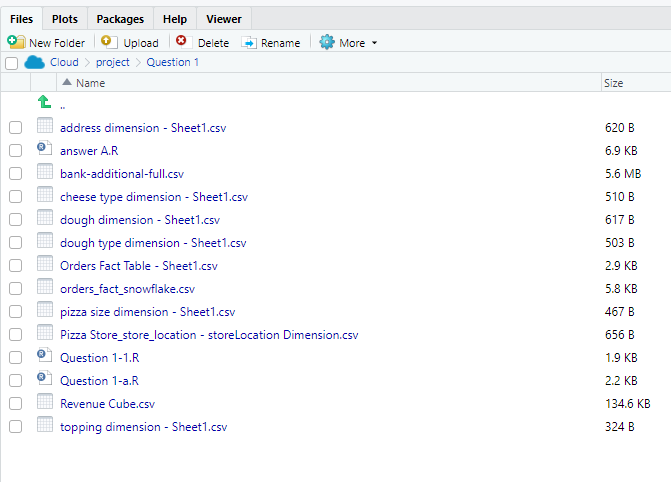




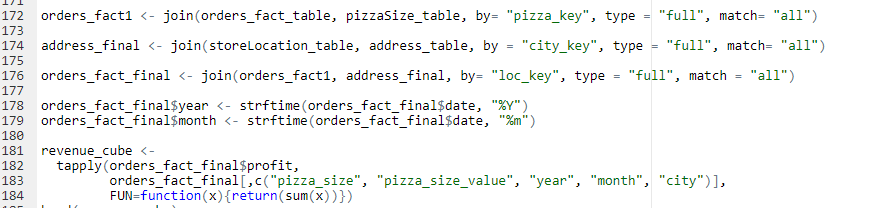
Below is the snippet of the console:



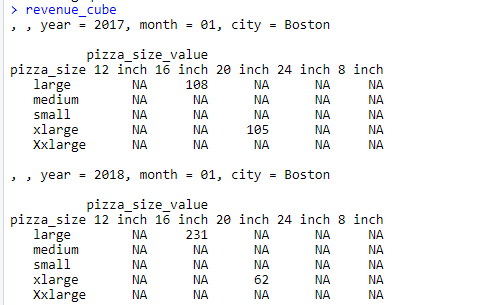
Additionally, here are the csv files stored in the directory:



Because of the snowflake schema, the fact table needed to be joined with some of the dimension tables to be able to generate the revenue cube. The join operations are as follows:



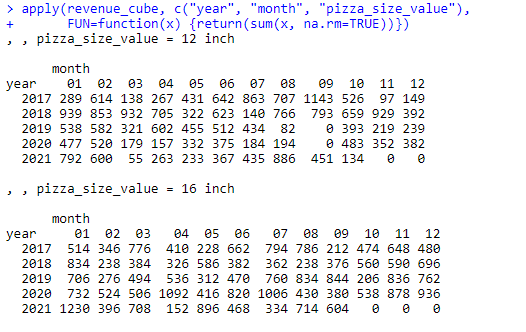
Snippet from the console after creating the OLAP cube(revenue cube):



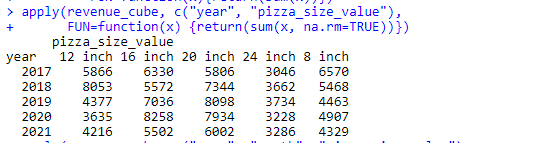
Since our generated values are not that large (1900 rows) and since the revenue cube has a lot of sub-parts (smaller cubes) there are a few values with “NA”.

3. Suppose that we want to examine the data of the above store to find trends and thus to predict which Pizza components the store should order more of. Describe a series of drilldown and roll-up operations that would lead to the conclusion that customers are beginning to prefer bigger pizzas.

Ans) For the Drill down operations we considered only the pizza size value with time. The three sides of the cube here was “pizza size”, “month” and “year”. As shown from the console these were some the findings:



For the roll up operations, we wanted to see which pizza size was more favored over the years, so for this case our values were: “pizza\_size\_value” and “year”. Below is the snippet from the console that shows the trend:



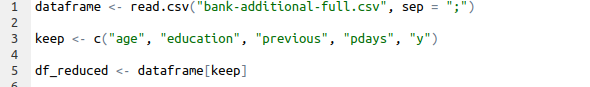
As we can see from this trend, there was a slow increase in larger pizza sizes, some examples that can be seen in this figure are: 24 inch went from 3046 in 2017 to 3228 in 2020, also 20 inch went from 5806 in 2017 to 7934 in 2020. Meanwhile, there was a gradual decrease in the smaller pizza sizes. Some examples are: for 8 inch it went down from 6570 in 2017 to 4907 in 2020. As 2021 is still not over, the data is still not complete, however, we can already see the values for 24inch pizza and 20-inch pizza are quite high.

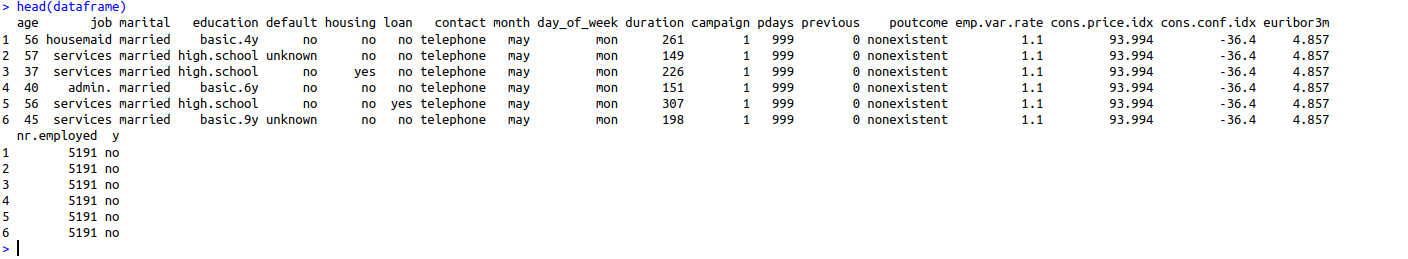
**Part B**

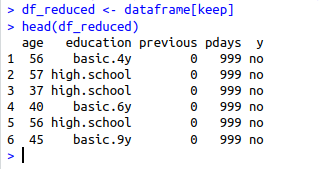
1. Import the data set into RStudio and reduce the dataset to only four predictors (age,

education, previous, and pdays), and the target, response.

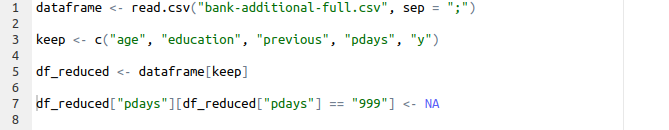
For this part we get the bank-additional-full.csv from UCI Machine Learning Repository. The data is read and stored in a dataframe, then a vector is produced of all the column names which are to be kept. In the 5th line, the dataframe is being reduced, to keep age, education, previous, pdays and y (which is the target/response) and assigned in a variable called df\_reduced.







1. The field pdays is a count of the number of days since the client was last contacted from a previous campaign. The code 999 in the value represents customers who had not been contacted previously. Change the field value 999 to “NA” to represent missing values.



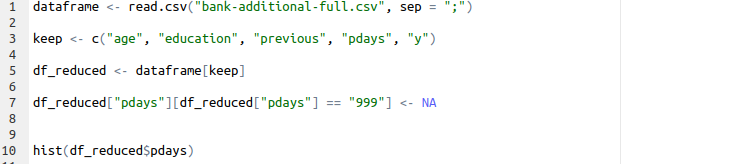
Here in line number 7 the operation is done, to replace all missing values (which are 999) to NA.

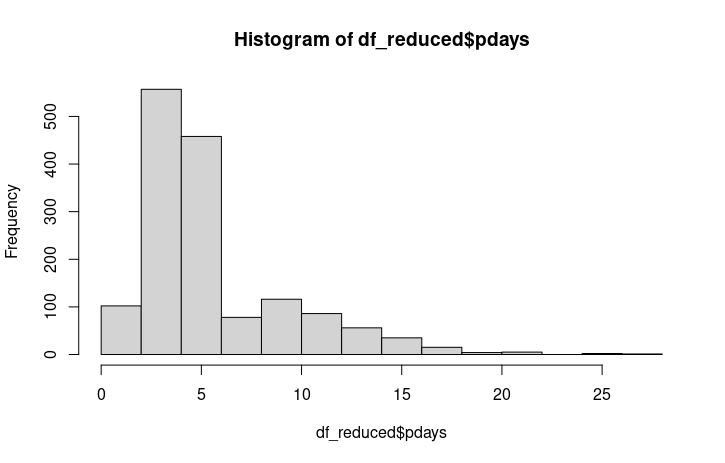
1. Explain why the field pdays is essentially useless until you handle the 999 code?

The variable presented here is a continuous variable, if we are planning to do an exploratory data analysis, this value is going to be represented as another value rather than a missing value, which can lead to wrong assumptions when reading the data and column. Therefore, it is important to address the missing values using special values such as NA, or replace them with median, mean etc before doing other operations.

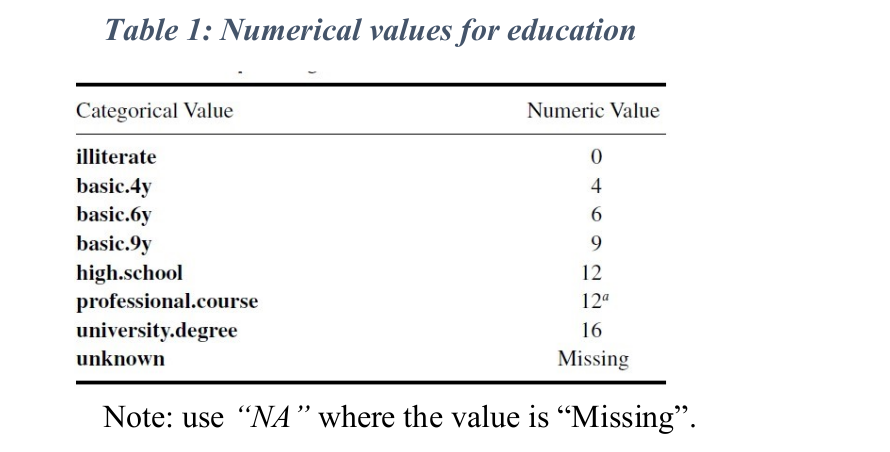
1. Create a histogram of the pdays variable showing the missing value excluded.

Line number 12, represents making a histogram of pdays, likewise, the missing values are ignored by R while creating in histogram if mentioned properly with NA.





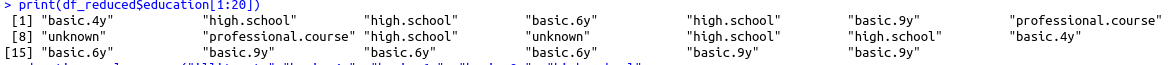
1. Transform the data values of the education field into numeric values using the chart in Table 1 below.



For this task, we are assuming 12 for 12a incase of professional.course. Here is the list of operations done from line 12 till line 25. Where we create a vector for all the categorical values and assign it to a variable in education\_values. In line 17 we assign labels for all the categorical values. In 19 we create a factor of all education values and transform it into numeric values, and store them back in the education column of df\_reduced.



Before:



After:



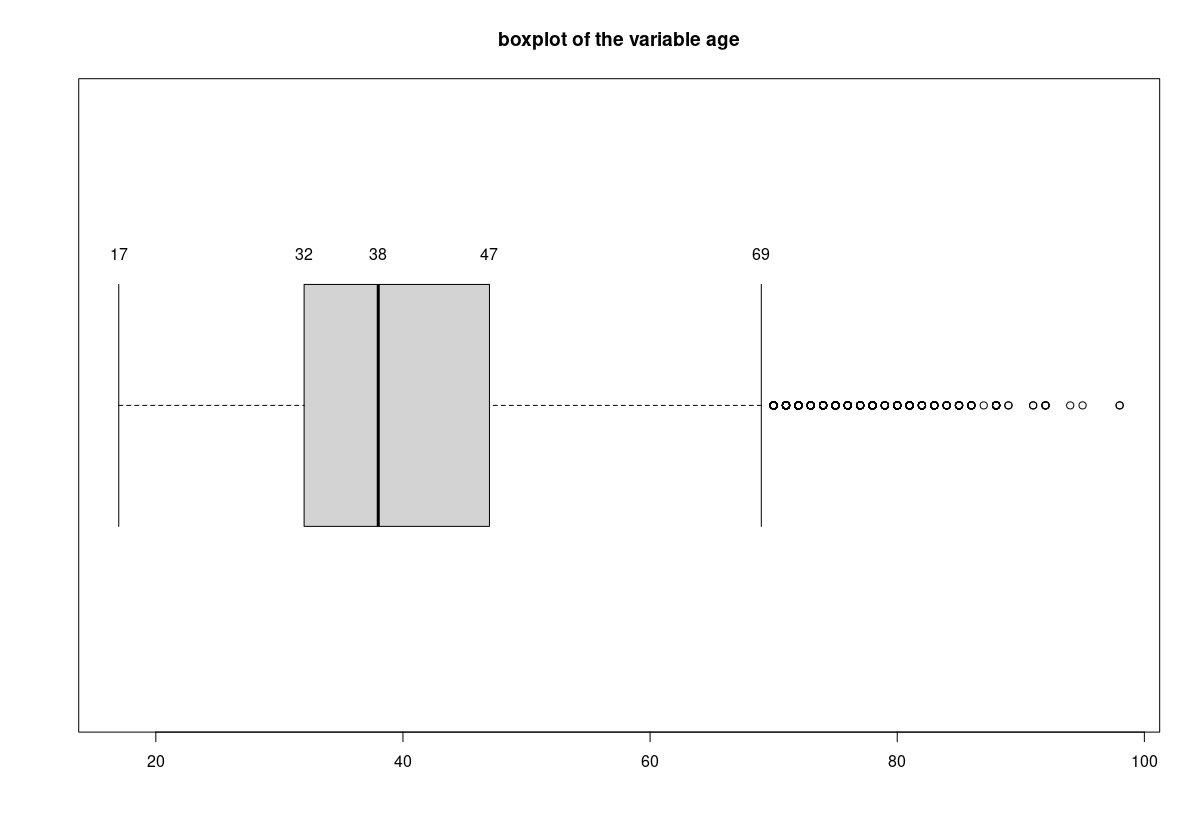
1. Compute the mean, median & mode of the age variable. Using a boxplot, give the fivenumber summary of the data. Plot the quantile information.

Line 32, 33, 34, computers mean, median & mode. For getting the mode we use a function called getmode, which uses the tabulate method, which finds out the highest frequency of a variable in a vector, after that we use the which.max method to find out the highest frequency of the variable which occurred the most and calculate the mode.

Here are the values for mean median and mode.



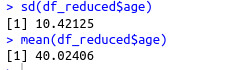




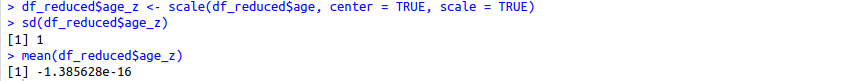
Here is the five number summary of the data, the minimum is 17, the first quartile is 32, the median is 38, the third quartile is 47 and the maximum is 69, after that all values are considered outliers.

1. Some machine learning algorithms perform better when the numeric fields are standardized. Standardize the age variable and save it as a new variable, age\_z.

For the time being, the mean value is 40 and the standard deviation is 10.42125



After standardization:



For normalization we use the function scale (line 39), along with it we center the values and divide it by sigma. Likewise, our standard deviation is 1 and the mean is almost near to zero.



1. Obtain a listing of all records that are outliers according to the field age\_z.

The operation in line 41 is used to find all the outliers in the age\_z column. We have also provided the list of values which are outliers.

