Assignment two-Basic data management with R (100 marks)

The following functions are used in this exercise:
Please fill in the description by using help function (?)

Function name	Description
read.csv()	
nrow()	
ncol()	
colnames()	
sapply()	
tapply()	
sort()	
sum()	
head()	
length()	
table()	
unique()	
lubidate::mdy_hm()	
lubidate::date	
substring()	
mean()	

In the raw data file it contains the training data set which indicates on each row if there exists customer complaints or not (with the field NumComplains=1 or 0) If the attribute NumComplains=1 then it indicates a customer complaint case. The LCID field here refers to the location where the customer complaint happened.

1. Import the csv file and assign it to data frame with name df (5 marks)
2. Determine the number of columns and rows in the data frame (5 marks)
3. Store the column names as a separate vector Colname_df (5 marks)
4. Determine the number of missing values per columns in the dataframe df

(10 marks)

5. Determine the first three of the KPI columns with the most NAs in the record file and stores it as vector. (10 marks)
6. Calculate the number of distinct LCID in the KPI records (10 marks)
7. Convert the column hour_id to date time format (5 marks)
8. Determine the number of records per day (10 marks)
9. Determine the number of complaint cases and the non-complaint cases found in the csv file (column name used=NumComplain) (10 marks)
10. Determine the top 10 LCIDs with the most complaint cases (10 marks)
11. Given that the first 5 digits of the column LCID refer to the particular region of the network, create additional column with name “Region” in the dataframe (10 marks)
12. Determine the region with the most complaint cases found in the training data (10 marks)