Department of Statistics, School of Mathematical Sciences
Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon
M.Sc. (Statistics) I Practical Sheet: MST 506
Topic 3 Use of tidyr and data.table package

---

1) Consider data sets named `table1`, `table2`, `table3`, `table4a`, `table4b` and `table5` from `tidyr` library in `R` to solve the following. Each datasets shows values for four variables, `country`, `year`, `population` and `cases`, but each dataset organizes the values in a different way.
   a) Compute `rate` per 10,000 with `table1`.
   b) Compute `cases` per `year` with `table1`.
   c) Use appropriate function to get tidied versions of `table4a`, `table4b`, `table2`.
   d) Combined the tidied version of `table4a` and `table4b` into a single `tibble`.
   e) Use `table3` dataset to split values of `rate` column into `cases` and `population` columns.
   f) Combine a `century` and `year` column to form a `new` variable using `table5`.
2) Use the data from following link to solve the following problems using `data.table` library in R.
   https://github.com/arunsrinivasan/satrdays-workshop/raw/master/flights_2014.csv"
   a) Read the data from a web link given above.
   b) Describe the data.
   c) Check the format of data and convert it into `data.table`.
   d) Select the columns named, 'origin', 'year', 'month' and 'hour'.
   e) Drop the columns named 'origin', 'year' and 'month'.
   f) Rename the variables 'dest' and 'origin' as 'destination' and 'origin.of.flight' respectively.
   g) Suppose we want to find all the flights whose `origin` is 'JFK'.
   h) Filter all the flights whose `origin` is either 'JFK' or 'LGA'.
   i) Filter all the flights whose `origin` is equal to 'JFK' and carrier is 'AA'.
   j) Sort the data with respect to `origin` variable in ascending order and descending order.
   k) Sort the data first by `origin` on ascending order and then by variable 'carrier' on descending order.
   l) Add new columns named `dep_sch` which is `dep_time- dep_delay` and `arr_sch` which is `arr_time – arr_delay`.
   m) Create a variable `flag` which is 1 if min is less than 50 otherwise `flag` =0.
   n) Calculate mean, median, min and max of `arr_delay` variable.
   o) Calculate the mean arrival time for each unique value in the 'origin' column.
   p) Calculate the mean of arrival time and delay time for each unique value in the 'origin' column.
   q) Remove the duplicates values based on 'carrier' variable.
   r) Remove the duplicates values based on all the variable.
   s) Extract the last row within each group by the `carrier` column.
   t) Calculate the total number of rows by month and then sort on descending order.
   u) Find top 3 months with high mean arrival time.
   v) Extract average of arrival and departure delays for `carrier` is 'DL' by 'origin' and 'dest' variables.