DSC4217 201617 Sem 2 Test 1

# Instruction

Download “Aviation Data.csv” from “Test 1” folder in IVLE. Use RMarkdown to write your solution to the given questions below. Name your RMarkdown file in the following format “***Test1\_[Name].rmd***”, where [Name] is your full name. Knit your Rmd file to either HTML or PDF and submit it to “Test 1” folder under “Student Submission” in IVLE.

# About the data

The file contains information from 1962 and later about civil aviation accidents and selected incidents within the United States, its territories and possessions, and in international waters. Generally, a preliminary report is available online within a few days of an accident. Factual information is added when available, and when the investigation is completed, the preliminary report is replaced with a final description of the accident and its probable cause. Full narrative descriptions may not be available for dates before 1993, cases under revision, or where NTSB did not have primary investigative responsibility.

# Definition

In the following question statements, “Event” refers to a record in the dataset, either an accident or an incident. The data fields in the dataset should be easily understood by the column names. If you have any doubt, raise your hand to ask.

# Questions

Q1. Load the data into R environment. Convert “Event.Date” into “Date” class. Generate a new column to indicate the year of the event and a new column to indicate the decade of the event. (5 marks)

Q2. Identify the overall top five countries in terms of number of events occurring. Do the same thing by decades. Is there any shift in the top five countries? (10 marks)

Q3. US should be the number 1 country in Q2. Use data visualization to understand the trend of number of incidents and number of accidents occurring in US. (10 marks)

Q4. There are three kinds of weather conditions given in the dataset:

* VMC - VMC is meteorological conditions expressed in terms of visibility, distance from cloud, and ceiling equal to or better than specified minima.
* UNK – Unknown
* The definition of IMC, according to the FAA, is meteorological conditions expressed in terms of visibility, distance from clouds, and ceiling less than the minima specified for visual meteorological conditions (VMC).

Use data visualization to illustrate your findings on how weather conditions affect the happening of incidents/accidents and the trend. Are your findings intuitive or counter-intuitive? Why? (15 marks)

Q5. Understanding how data is classified is very important in data analysis. Note that there is a field named "Total.Fatal.Injuries" that records the total number of fatal injuries in the events. Use "table" function on this field to see the distribution of number of fatal injuries. There is another field named "Injury.Severity" that also records number of fatal injuries. Compare data in these two fields. Do you observe any inconsistency? What could be the cause of the inconsistency? Please re-classify the severity of events by number of fatal injuries in the following groups:

* "Fatal" - event that has at least one fatal injuries
* "Non-Fatal" - event without fatal injuries
* "Unknown" - original field is "NA"

(20 marks)

Q6. Use the spatial data in the dataset to plot meaningful charts on US map. Write a short paragraph about your findings. (20 marks)

Q7. Ask you own questions and use R to answer the questions. (20 marks)