**Assignment 5.2**

There are 2 datasets ‘Delayed\_Flights.CSV’ and ‘Airports.CSV’

**Delayed\_Flights.CSV** --- There are 29 columns in this dataset

**Airports.CSV** ---- There are 7 columns in this dataset

**Problem Statement 1:**

Find out the top 5 most visited destinations:

Step 1: Registering Piggybank.jar in order to use the CSVExcelStorage class



Loading ‘delayed flights’ dataset in to a variable using CSVExcelStorage because of its effective technique to handle double quotes and headers



Generating the columns that are required for processing and explicitly typecasting each of them



Filtering the null values from the “dest” column



Grouping relation C by “dest”



Generating the grouped column and the count of each



Order and limit the result to top 5





Loading another table to which we will look-up and find the city as well as the country



Generating dest, city and country from previous relation



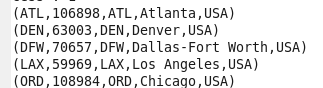
Joining Result and A2 based on a common column (dest)



dump to print the result



**Output:**



**Problem Statement 2:**

Which month has seen the most number of cancellations due to bad weather?















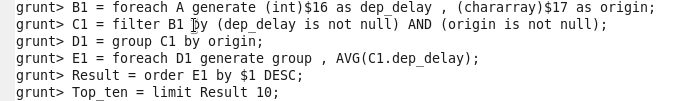
**Output:**



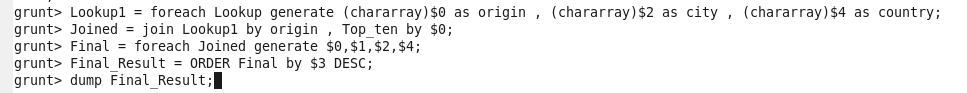
**Problem Statement 3:**

Top ten origins with the highest AVG departure delay

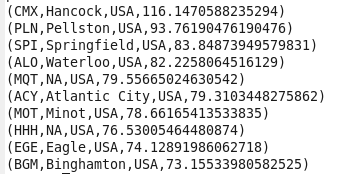








**Output:**



**Problem Statement 4:**

Which route (origin & destination) has seen the maximum diversion?

















