**CS673 – Group Project:**

**Part 1: Initial phase.** For a maximum of **12.5** points

Use the site below for the data that will be needed for this part of this project:

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7>

Variable descriptions

|  |  |  |
| --- | --- | --- |
|  | **Name** | **Description** |
| 1 | Year | 1987-2008 |
| 2 | Month | 1-12 |
| 3 | DayofMonth | 1-31 |
| 4 | DayOfWeek | 1 (Monday) - 7 (Sunday) |
| 5 | DepTime | actual departure time (local, hhmm) |
| 6 | CRSDepTime | scheduled departure time (local, hhmm) |
| 7 | ArrTime | actual arrival time (local, hhmm) |
| 8 | CRSArrTime | scheduled arrival time (local, hhmm) |
| 9 | UniqueCarrier | [unique carrier code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 10 | FlightNum | flight number |
| 11 | TailNum | plane tail number |
| 12 | ActualElapsedTime | in minutes |
| 13 | CRSElapsedTime | in minutes |
| 14 | AirTime | in minutes |
| 15 | ArrDelay | arrival delay, in minutes |
| 16 | DepDelay | departure delay, in minutes |
| 17 | Origin | origin [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 18 | Dest | destination [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 19 | Distance | in miles |
| 20 | TaxiIn | taxi in time, in minutes |
| 21 | TaxiOut | taxi out time in minutes |
| 22 | Cancelled | was the flight cancelled? |
| 23 | CancellationCode | reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 24 | Diverted | 1 = yes, 0 = no |
| 25 | CarrierDelay | in minutes |
| 26 | WeatherDelay | in minutes |
| 27 | NASDelay | in minutes |
| 28 | SecurityDelay | in minutes |
| 29 | LateAircraftDelay | in minutes |

Use the same site to get the auxiliary files to convert the **carrier** code and **airport** code into their corresponding names. The downloads are at the last page of the site.

Each member of each team will be assigned a year and will do these tasks with the file

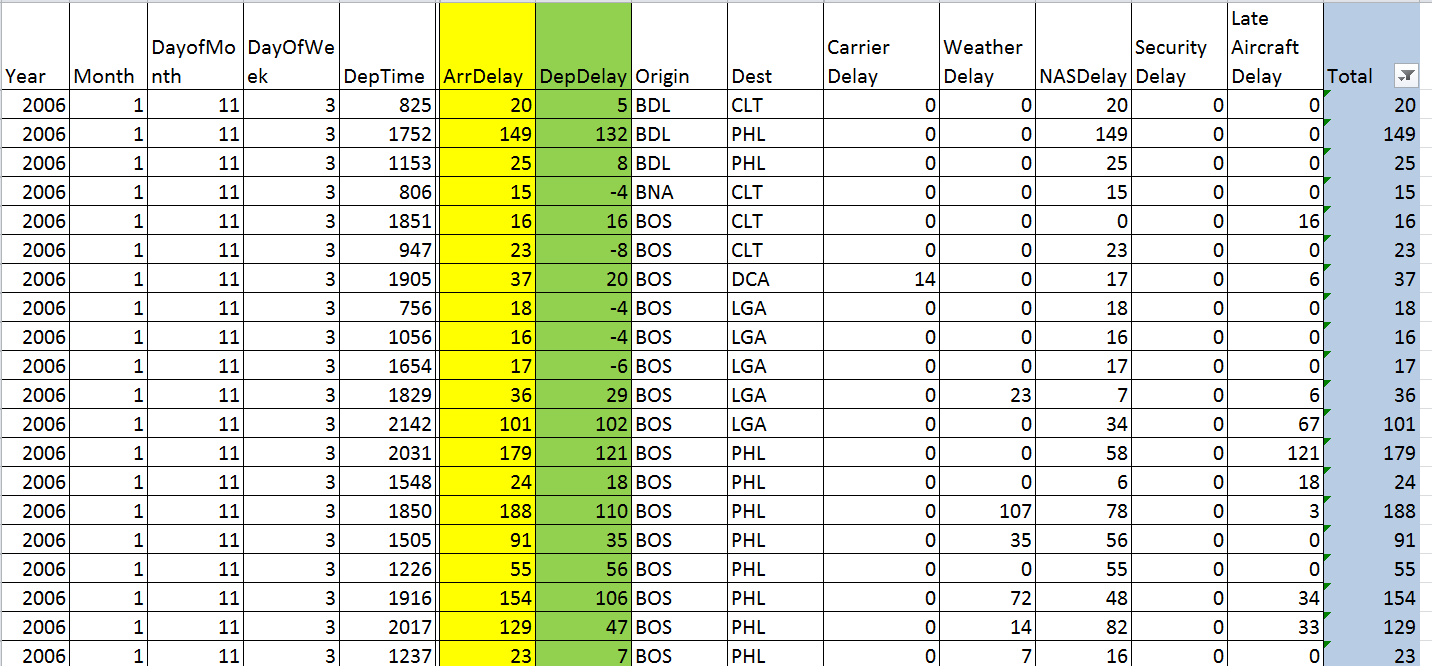
* 1. extract the data from the site and **download the cvs file from the website**
  2. create a Hive table and load all the file data into it. Prefix the table name with your name.
  3. display a sample of the loaded rows (100 rows)
  4. determine the three **airports** with the highest delay time (in hours) in **your year**
  5. determine the three **carriers** with the highest delay time (in hours) in **your year**
  6. determine overall which type of delay (arrivals or departures) is the largest for **your airports**
  7. all members of a team will collaborate on the presentation deck and a create a conclusion slide that compare the results of tasks d, e and f **within your team** and pick **the airport** and **carrier with the highest delay**

Each team member will capture individual results and observations for your year and include them in a team’s Powerpoint deck and present them in class with the rest of the team. A conclusion summary of observations of meaningful patterns for the group’s years has to be included at the end of the deck.

Use **legible** and understandable screenshots of the successful execution of the work necessary for each task in formatted results. Use data visualization of your numeric findings.

Each team member will upload to Classes the group Powerpoint deck including all the individual member slides plus the conclusions summary slide of the team. **Prefix your file with your name or you will lose 2.5 points.**

Example of some input records in spreadsheet format:



Group 3

|  |
| --- |
| Borkar, Amey Suresh - 2002 year |
| Geda, Ranjith Kumar - 1996 year |
| Joshi, Shivam - 2003 year |
| Nguyen, Thao - 2007 year |
| Tamboli, Suraj Dhananjay - 2000 year |