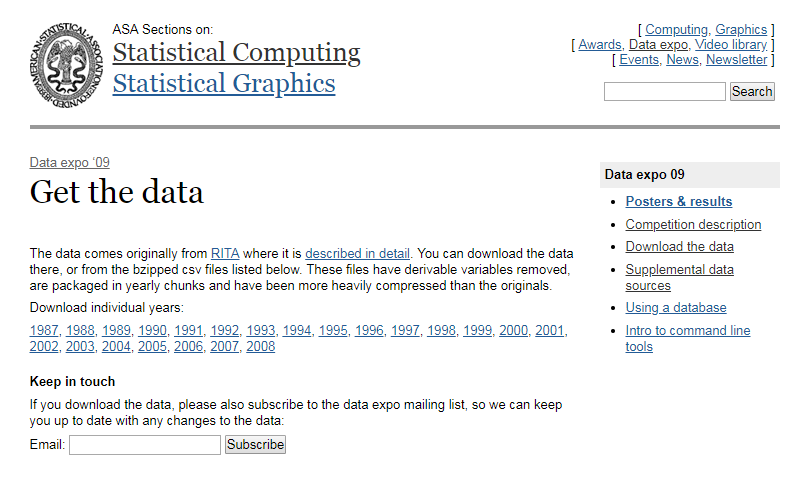
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **□ 수행평가 – 빅데이터 분석시스템 구축 과정** | | | | |
|  |  |  |  |  |
| **과정명** | | [4차산업 선도인력]커넥티드 카를 위한 IoT 전문가 과정(NCS) | | |
| **능력단위명** | | 분석용 데이터 탐색 2001010505\_15v1 | **훈련교사** | 이진만 |
| **능력단위요소명** | | 빅데이터 저장 계획 수립하기, 빅데이터 저장 모델 설계하기 빅데이터 저장 관리시스템 구성하기 빅데이터 적재하기, 빅데이터 운영하기 | | |
| **수행날짜** | | 2018.04.05 | 훈련생명 | 김 동건 |
| **과제개요** | | | | |
| 1. 데이터를 수집 한다.(공공데이터, 서울시 공공데이터) 2. 데이터를 빅데이터 시스템(하둡 및 데이터베이스) 저장  3. R을 통해 데이터를 탐색 및 분석 한다 4. 데이터 분석 화면을 구현 한다. | | | | |

**1. 데이터를 수집 한다.(공공데이터, 서울시 공공데이터)**

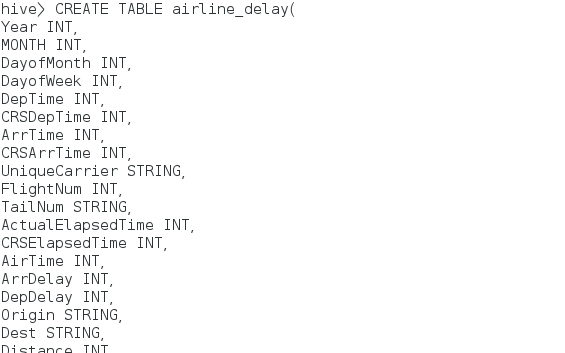
미국 airline\_delay 2006 ~ 2008 까지 데이터 수집



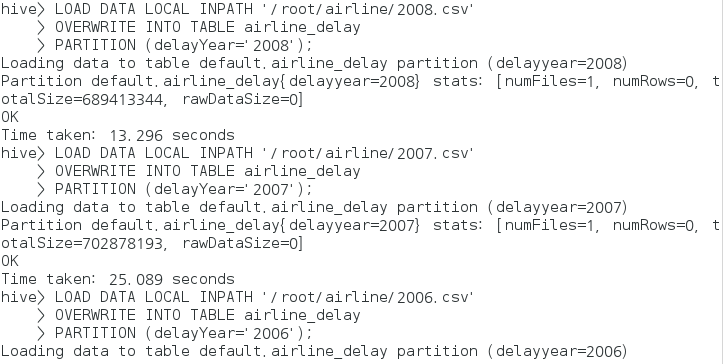
[**http://stat-computing.org/dataexpo/2009/the-data.html**](http://stat-computing.org/dataexpo/2009/the-data.html)

**2. 데이터를 빅데이터 시스템(하둡 및 데이터베이스) 저장**

- 테이블 생성



- data load

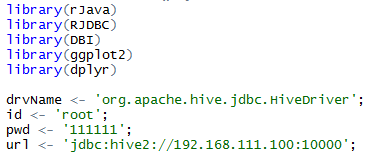


- service 돌리기

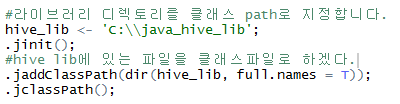


**3. R을 통해 데이터를 탐색 및 분석 한다**

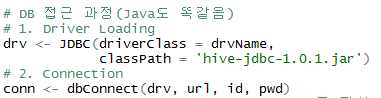
- 필요 라이브러리를 import, HiveDriver와 id/pw/url을 입력



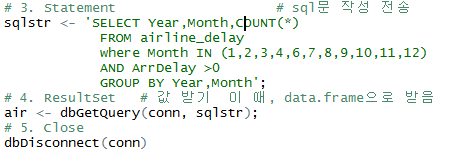
- 라이브러리가 담긴 디렉토리를 ClassPath로 지정



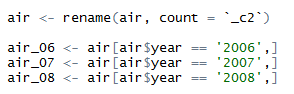
- Driver loading을 하고 Connection 해준다.



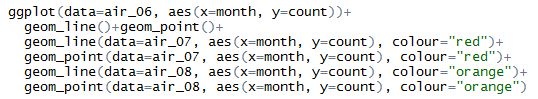
- Statement를 통해 Sql 전송(연도/월 별 연도와 월, 딜레이 된 비중 count)



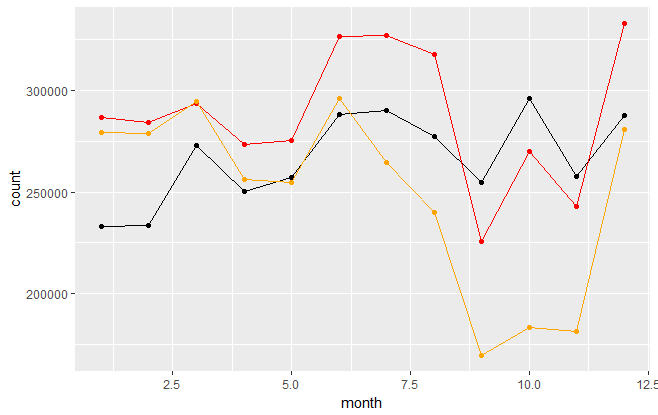
- 연도, 월, 딜레이 Count 칼럼 dataFrame으로 추출



**4. 데이터 분석 화면을 구현 한다.**

- 꺽은선 그래프를 통해 2006, 2007, 2008년도의 월 별 딜레이 횟수를 비교

- 2006년도 검은색, 2007년도 빨간색, 2008년도 노란색



분석 결과 : **8~9월**이 대체로 딜레이 비율이 **대폭 하락**하며, **6~7월, 11~12월**이 딜레이 비율이 **대폭 상승**한다.

그래프를 분석한 결과, **휴가 시즌**에 딜레이 비율이 높은 것으로 나타났다.