

☆ 8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

☞ 분석 데이터셋 생성 (1/4)

01. “스마트카 고객 마스터2” 파일을 Server02로 업로드한다.

- FTP 클라이언트 파일질라 실행
- 파일럿 작업 경로: /home/pilot-pjt/working
- C://예제소스/bigdata2nd-master/CH08/CarMaster2Income.txt 파일을 /home/pilot-pjt/working에 업로드

8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

분석 데이터셋 생성 (2/4)

02. 휴에서 하이브 에디터에 접속해 SmartCar_Master2Income 테이블을 생성한다. 이때 스마트카 배기량(car_capacity)과 연소득(income) 필드의 데이터 타입은 int(숫자형)로서 각각 회귀분석의 독립변수와 종속변수로 사용된다.

```
1 CREATE EXTERNAL TABLE SmartCar_Master2Income (  
2   car_number string,  
3   sex string,  
4   age string,  
5   marriage string,  
6   region string,  
7   job string,  
8   car_capacity int,  
9   car_year string,  
10  car_model string,  
11  income int  
12 )  
13 row format delimited  
14 fields terminated by '|'   
15 stored as textfile  
16 tblproperties ("skip.header.line.count"="1");
```



✓ 성공.

8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

분석 데이터셋 생성 (3/4)

03. 로컬 파일시스템에 적재된 “스마트카 고객 마스터2.txt” 파일을 앞서 생성한 하이브 테이블인 SmartCar_Master2Income에 직접 로드한다.

```
1 LOAD DATA LOCAL
2   INPATH '/home/pilot-pjt/working/CarMaster2Income.txt'
3   OVERWRITE INTO TABLE SmartCar_Master2Income;
```



✓ 성공.

8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

분석 데이터셋 생성 (4/4)

04. 하이브 테이블에 로컬파일이 정상적으로 로드됐는지 조회 쿼리를 실행해 본다.

0.0s
default ▾
📄
⚙️
?

```

1 Select
2     car_number,
3     car_capacity,
4     income
5 from SmartCar_Master2Income;
  
```

▶
📖 ▾

쿼리 기록 🔍 🗒️
저장된 쿼리 🔍
결과 🔍 ↗️

	car_number	car_capacity	income
1	A0001	4500	5850
2	A0002	3500	3850
3	A0003	4000	4800
4	A0004	4000	4400

8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

하이프 클라이언트 라이브러리 구성

01. 사용자의 파일럿 PC에 하이브 클라이언트 라이브러리를 관리하기 위한 두 개의 디렉터리를 생성한다.

- C://hiveJar 디렉터리를 생성한다.
- C://hadoopJar 디렉터리를 생성한다.

02. Server02로부터 관련 라이브러리를 다운로드한다. 총 8개의 Jar 파일로 하이브 클라이언트 라이브러리 파일 5개와 하둡 클라이언트 라이브러리 파일 3개다. 먼저 Server02에 FTP로 접속해 하이브 클라이언트 라이브러리 파일 5개를 C://hiveJar 디렉터리에 다운로드한다.

- FTP 클라이언트 파일질라 실행
- Server02에 접속해 아래의 Jar 파일을 사용자 PC의 "C://hiveJar" 디렉터리에 다운로드한다.

/opt/cloudera/parcels/CDH/jars/hive-jdbc-2.1.1-cdh6.3.2.jar

/opt/cloudera/parcels/CDH/jars/hive-service-2.1.1-cdh6.3.2.jar

/opt/cloudera/parcels/CDH/jars/httpclient-4.5.3.jar

/opt/cloudera/parcels/CDH/jars/httpcore-4.4.6.jar

/opt/cloudera/parcels/CDH/jars/hive-jdbc-2.1.1-cdh6.3.2-standalone.jar

- 하둡 클라이언트 라이브러리 파일 3개를 "C://hadoopJar" 디렉터리에 다운로드한다.

/opt/cloudera/parcels/CDH/jars/hadoop-common-3.0.0-cdh6.3.2.jar

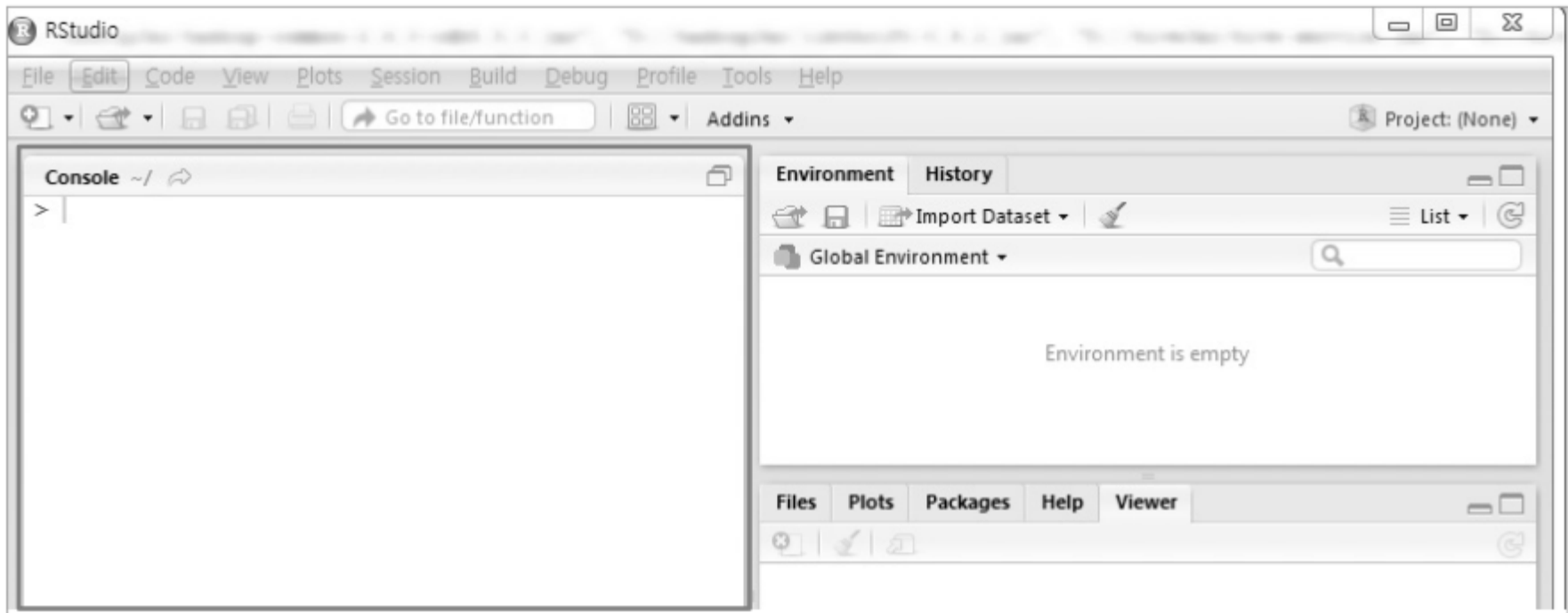
/opt/cloudera/parcels/CDH/jars/libthrift-0.9.3.jar

/opt/cloudera/parcels/CDH/jars/hadoop-client-3.0.0-cdh6.3.2.jar

8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

하이프 데이터 로드 (1/7)

01. 사용자 PC의 파일럿 환경에 설치된 R-Studio를 실행한다. 다음과 같이 R-Studio의 콘솔창이 활성화되면 정상적으로 실행된 것이다.



8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

하이버 데이터 로드 (2/7)

02. 필요한 R 패키지 "DBI", "rJava", "RJDBC"를 추가로 설치한다. 다음 명령을 R 콘솔에서 차례로 실행한다. R 패키지 설치 명령을 실행하면 CRAN(The Comprehensive R Archive Network)에서 직접 다운로드하므로 인터넷이 연결된 상태여야 한다.

```
> install.packages("DBI")  
> install.packages("rJava")  
> install.packages("RJDBC")  
> install.packages("log4r")
```

03. 추가 패키지 설치가 완료되면 설치된 패키지의 라이브러리를 로드한다.

```
> library("DBI")  
> library("rJava")  
> library("RJDBC")  
> library("log4r")
```


8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

하이브 데이터 로드 (3/7)

04. 이제 파일럿 환경의 하이브 데이터웨어하우스에 접속해 “스마트카 고객 마스터2” 테이블인 “SmartCar_Master2Income” 테이블을 확인해 보겠다. 앞서 hiveJar와 hadoopJar 디렉터리를 R의 클래스패스로 설정했다. 다음의 R 명령은 “C://예제소스/bigdata2nd-master/CH08/R명령.txt”로 제공하니 참고한다.

```
> hive.class.path = list.files(path=c("C://hiveJar"), pattern="jar", full.names=T);  
> hadoop.lib.path = list.files(path=c("C://hadoopJar"), pattern="jar", full.names=T);  
> hadoop.class.path = list.files(path=c("C://hadoopJar"), pattern="jar", full.names=T);  
> class.path = c(hive.class.path, hadoop.lib.path, hadoop.class.path);  
> .jinit(classpath=class.path)
```


8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

하이브 데이터 로드 (4/7)

05. 하이브의 JDBC 드라이버를 로드하고 Server02에 설치돼 있는 하이브 서버 2로 접속한다. 이때 접속 계정과 비밀번호는 브라우저를 통해 “<http://server01.hadoop.com:7180/api/v1/clusters/Cluster 1/services/hive/config>”에 접속하면 확인할 수 있다. 독자의 파일럿 환경마다 다른 계정과 비밀번호가 만들어질 수 있으니 주의한다.

- 하이브 설정 URL: <http://server01.hadoop.com:7180/api/v1/clusters/Cluster 1/services/hive/config>

```
{
  "items" : [ {
    "name" : "hbase_service",
    "value" : "hbase"
  }, {
    "name" : "hive_metastore_database_host",
    "value" : "server01.hadoop.com"
  }, {
    "name" : "hive_metastore_database_name",
    "value" : "hive2"
  }, {
    "name" : "hive_metastore_database_password",
    "value" : "A9mkRvDkNf"
  }, {
    "name" : "hive_metastore_database_port",
    "value" : "7432"
  }, {
    "name" : "hive_metastore_database_type",
    "value" : "postgresql"
  }, {
    "name" : "hive_metastore_database_user",
    "value" : "hive2"
  }, {
    "name" : "mapreduce_yarn_service",
    "value" : "yarn"
  }, {
    "name" : "zookeeper_service",
    "value" : "zookeeper"
  } ]
}
```

8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

하이브 데이터 로드 (5/7)

다음 명령을 통해 하이브 JDBC를 로드하고 하이브 서버 2에 연결한다.

```
> drv <- JDBC("org.apache.hive.jdbc.HiveDriver", "C://hiveJar/hive-jdbc-2.1.1-cdh6.3.2.jar",  
  identifier.quote="`")  
> conn <- dbConnect(drv, "jdbc:hive2://server02.hadoop.com:10000/default", "hive2",  
  "A9mkRvDkNf")
```

8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

하이브 데이터 로드 (6/7)

06. 하이브의 데이터웨어하우스에 생성된 테이블 목록을 조회한다. 참고로 저자의 경우 "Smartcar_Master2Income" 테이블만 조회했지만 6, 7장에서 만든 다른 하이브 테이블도 조회가 가능하다.

```
> dbListTables(conn);
```

```
[1] "smartcar_master2income"
```

8.2 R을 이용한 회귀분석 - 운전자 연소득 예측


하이프 데이터 로드 (7/7)

07. 여기까지 성공적으로 진행되면 이제 하이브의 “스마트카 고객 마스터2” 테이블에 질의하고 내용까지 확인해 보자.

```
> data <- dbGetQuery(conn, "select * from smartcar_master2income")
> View(data)
```

	smartcar_master2income.car_number	smartcar_master2income.sex	smartcar_master2income.age	smartcar_master2income.marriage	smartcar_master2income.region
1	A0001	남	64	미혼	전북
2	A0002	남	17	기혼	경북
3	A0003	남	68	기혼	경기
4	A0004	남	34	미혼	전북
5	A0005	여	26	미혼	서울
6	A0006	여	61	기혼	충남
7	A0007	여	40	미혼	세종
8	A0008	남	40	미혼	전북
9	A0009	남	47	미혼	세종
10	A0010	남	66	미혼	서울
11	A0011	여	31	미혼	울산

8.2 R을 이용한 회귀분석 - 운전자 연소득 예측

 R 파일럿 실행 1-3 단계 - 빅데이터 연동

실습