

---

# R & HIVE 데이터 사용 매뉴얼

(2020 날씨 빅데이터 콘테스트)

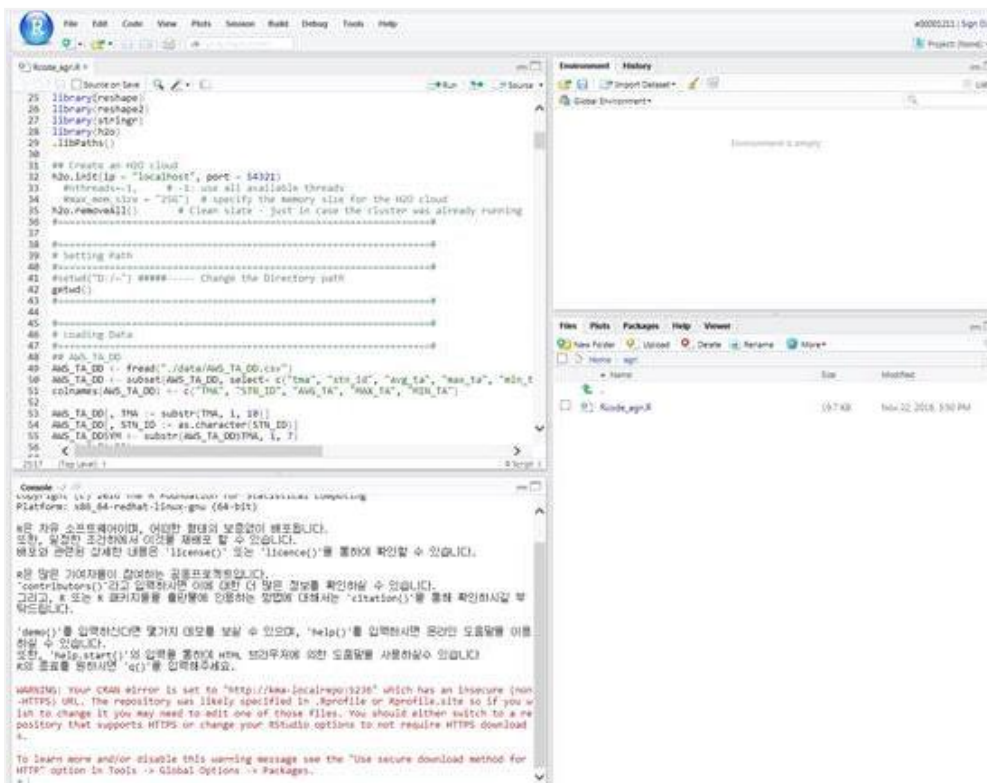
---

## 0. 날씨마루(<https://bd.kma.go.kr>) 접속



### 0-1. 날씨마루 내 분석환경 ▶ Rstudio 이용하기

- 분석환경 이용을 위한 계정신청이 필요합니다.



## 1. 개요

날씨 빅데이터 콘테스트 제공 데이터를 분석환경(R)에서 HIVE(Hadoop용 데이터웨어하우스 시스템)에 접속하여 사용할 수 있는 방법을 설명합니다.

### 1-1. HIVE에 저장되어 있는 현대제철과 KT 데이터를 안내합니다.

#### <현대제철 데이터 분석 분야>

##### - 현대제철 데이터 -

데이터정보	테이블명
현대제철	plant_test
	plant1_train
	plant2_train

#### <KT 서비스 개발 분야 >

##### - KT 상권 데이터 -

데이터정보	테이블명
KT 상권	admdong_pop_stay
	admdong_pop_walk
	bd_business_201901
	bd_business_201902
	bd_business_201903
	bd_business_201904
	bd_business_201905
	bd_business_201906
	bd_business_201907
	bd_business_201908
	bd_business_201909
	bd_business_201910
	bd_business_201911
	bd_business_201912

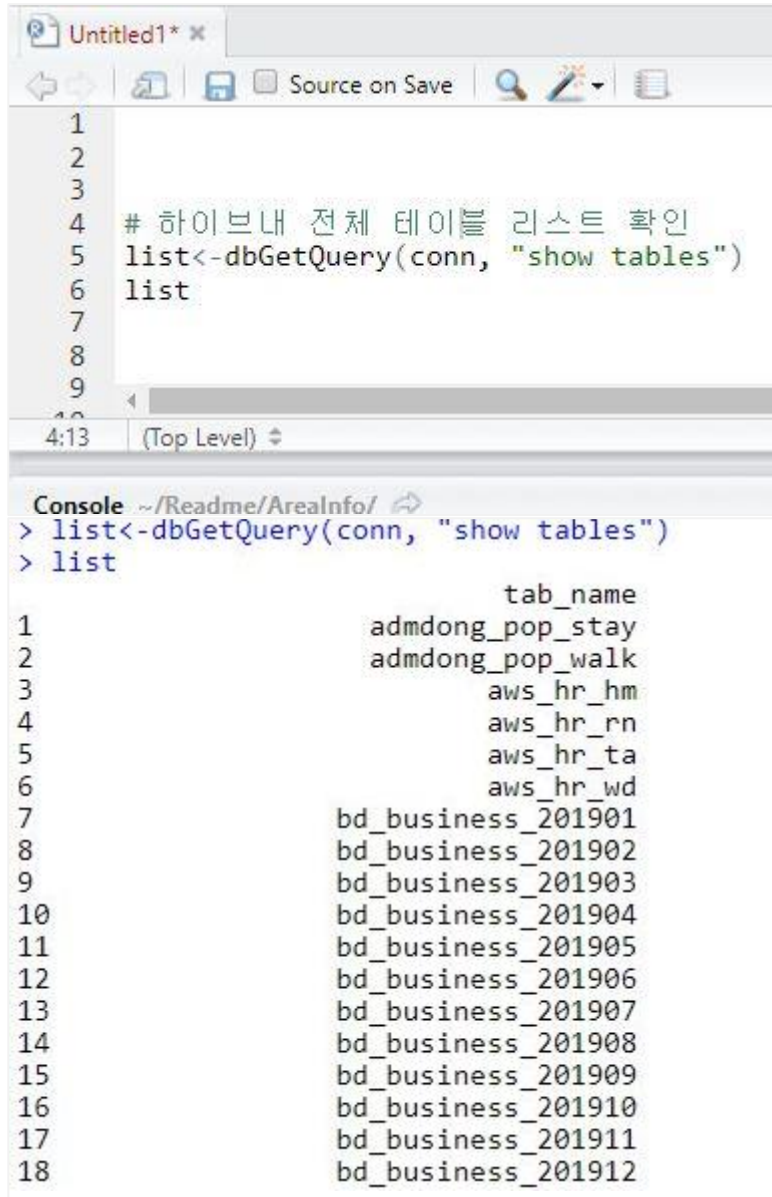
## - KT 관광 데이터 -

데이터정보	테이블명	구분
KT 관광 (강원도)	gangwon_festival_cnt	축제지
	gangwon_festival_sex_cnt	
	gangwon_festival_timezn_cnt	
	gangwon_festival_cntry_cnt	
	gangwon_festival_cntry_timezn_cnt	
	gangwon_festival_resd_rate	
	gangwon_festival_resd24_rate	
	gangwon_festival_stay_time	
	gangwon_tour_cnt	관광지
	gangwon_tour_sex_cnt	
	gangwon_tour_timezn_cnt	
	gangwon_tour_cntry_cnt	
	gangwon_tour_cntry_timezn_cnt	
	gangwon_tour_resd_rate	
	gangwon_tour_resd24_rate	
	gangwon_tour_uniq_cnt	
	gangwon_admdong_cnt	행정동
	gangwon_admdong_sex_cnt	
	gangwon_admdong_timezn_cnt	
	gangwon_admdong_cntry_cnt	
	gangwon_admdong_cntry_timezn_cnt	
	gangwon_admdong_resd_rate	
	gangwon_admdong_resd24_rate	
	gangwon_admdong_uniq_cnt	
	gangwon_sidogg_cnt	시군구
	gangwon_sidogg_sex_cnt	
	gangwon_sidogg_timezn_cnt	
	gangwon_sidogg_cntry_cnt	
	gangwon_sidogg_cntry_timezn_cnt	
	gangwon_sidogg_resd_rate	
	gangwon_sidogg_resd24_rate	
	gangwon_sidogg_uniq_cnt	

데이터정보	테이블명	구분
KT 관광 (제주도)	jeju_festival_cnt	축제지
	jeju_festival_sex_cnt	
	jeju_festival_timezn_cnt	
	jeju_festival_cntry_cnt	
	jeju_festival_cntry_timezn_cnt	
	jeju_festival_resd_rate	
	jeju_festival_resd24_rate	
	jeju_festival_stay_time	
	jeju_tour_cnt	관광지
	jeju_tour_sex_cnt	
	jeju_tour_timezn_cnt	
	jeju_tour_cntry_cnt	
	jeju_tour_cntry_timezn_cnt	
	jeju_tour_resd_rate	
	jeju_tour_resd24_rate	
	jeju_tour_uniq_cnt	
	jeju_admdong_cnt	행정동
	jeju_admdong_sex_cnt	
	jeju_admdong_timezn_cnt	
	jeju_admdong_cntry_cnt	
	jeju_admdong_cntry_timezn_cnt	
	jeju_admdong_resd_rate	
	jeju_admdong_resd24_rate	
	jeju_admdong_uniq_cnt	
	jeju_sidogg_cnt	시군구
	jeju_sidogg_sex_cnt	
	jeju_sidogg_timezn_cnt	
	jeju_sidogg_cntry_cnt	
	jeju_sidogg_cntry_timezn_cnt	
	jeju_sidogg_resd_rate	
	jeju_sidogg_resd24_rate	
	jeju_sidogg_uniq_cnt	

## 2. 데이터 확인

2-1. 날씨마루 분석환경(R Studio)에 접속하면 HIVE와 자동으로 연결되며, conn변수와 dbGetQuery() 함수를 이용하여 데이터를 조회 및 연산을 수행할 수 있습니다.



```
1  
2  
3  
4 # 하이브내 전체 테이블 리스트 확인  
5 list<-dbGetQuery(conn, "show tables")  
6 list  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18
```

Console ~/Readme/AreaInfo/ ↗

```
> list<-dbGetQuery(conn, "show tables")  
> list
```

	tab_name
1	admdong_pop_stay
2	admdong_pop_walk
3	aws_hr_hm
4	aws_hr_rn
5	aws_hr_ta
6	aws_hr_wd
7	bd_business_201901
8	bd_business_201902
9	bd_business_201903
10	bd_business_201904
11	bd_business_201905
12	bd_business_201906
13	bd_business_201907
14	bd_business_201908
15	bd_business_201909
16	bd_business_201910
17	bd_business_201911
18	bd_business_201912

예시)

```
list <- dbGetQuery(conn, "show tables")
```

### 3. 분석환경 데이터 로딩

데이터가 HIVE에 저장되어 있어 SQL을 이용한 분석이 바로 가능하지만, 분석환경으로 로딩이 필요한 경우 테이블을 새로 생성하여 저장할 수 있습니다.

```
4 # 현대제철 공장1 데이터 로딩
5 plant1_train<-dbGetQuery(conn, "SELECT * FROM plant1_train")
```

예시) 현대제철 공장1 학습데이터 테이블 이용

```
plant1_train<-dbGetQuery(conn, "SELECT * FROM plant1_train")
```

### 4. 데이터 확인

분석환경에 로딩한 데이터가 정상적으로 불러왔는지 확인 위해, head, tail, summary 함수를 이용합니다.

```
Console ~/Readme/ArcaInfo/
> head(plant1_train)
plant1_train.mea_ddhr plant1_train.tem_in_loc1 plant1_train.hum_in_loc1 plant1_train.tem_in_loc2
1      2016-04-01 0:00                16                24                11
2      2016-04-01 3:00                14                28                10
3      2016-04-01 6:00                13                33                10
4      2016-04-01 9:00                13                33                10
5      2016-04-01 12:00               16                28                10
6      2016-04-01 15:00               18                24                14
plant1_train.hum_in_loc2 plant1_train.tem_in_loc3 plant1_train.hum_in_loc3 plant1_train.tem_out_loc1
1                14                23                11                13
2                12                32                 9                11
3                11                37                 9                10
4                11                35                 9                10
5                15                27                11                14
6                18                21                14                16
plant1_train.hum_out_loc1 plant1_train.tem_coil_loc1 plant1_train.tem_coil_loc2 plant1_train.tem_coil_loc3
1                32                10                 9                42
2                42                 7                 7                59
3                44                 7                 6                56
4                41                 8                18                30
5                30                 9                18                20
6                27                12                17                23
plant1_train.cond_loc1 plant1_train.cond_loc2 plant1_train.cond_loc3
1                 0                 0                 0
2                 0                 0                 0
3                 0                 0                 0
4                 0                 0                 0
5                 0                 0                 0
6                 0                 0                 0
```

예시)

```
head(plant1_train)
```



```

Console ~/Readme/AreaInfo/
> summary(plant1_train)
plant1_train.meas_ddhr plant1_train.tem_in_loc1 plant1_train.hum_in_loc1 plant1_train.tem_in_loc2
Length:58749           Min.   :-7.98           Min.   :10.00           Min.   :-7.61
Class :character       1st Qu.: 8.66           1st Qu.:40.69           1st Qu.: 7.28
Mode  :character       Median:16.84           Median:50.31           Median:15.71
                        Mean :17.46           Mean :50.44           Mean :16.39
                        3rd Qu.:26.33           3rd Qu.:60.01           3rd Qu.:25.51
                        Max.  :37.08           Max.  :89.80           Max.  :34.86
                        NA's  :870             NA's  :870             NA's  :870

plant1_train.hum_in_loc2 plant1_train.tem_in_loc3 plant1_train.hum_in_loc3 plant1_train.tem_out_loc1
Min.   :-8.28           Min.   : 8.00           Min.   :-6.04           Min.   :-8.93
1st Qu.: 7.13           1st Qu.:42.61           1st Qu.: 6.28           1st Qu.: 6.28
Median :15.46           Median :53.02           Median :14.87           Median :14.73
Mean   :16.27           Mean   :52.96           Mean   :15.49           Mean   :15.40
3rd Qu.:25.45           3rd Qu.:63.52           3rd Qu.:25.00           3rd Qu.:24.56
Max.   :36.53           Max.   :91.42           Max.   :34.06           Max.   :35.92
NA's   :120             NA's   :120             NA's   :120             NA's   :145

plant1_train.hum_out_loc1 plant1_train.tem_coil_loc1 plant1_train.tem_coil_loc2 plant1_train.tem_coil_loc3
Min.   : 9.00           Min.   :-7.45           Min.   :-13.69           Min.   : 5.00
1st Qu.:46.25           1st Qu.: 5.41           1st Qu.: 3.96           1st Qu.:47.59
Median :56.62           Median :14.30           Median :12.72           Median :59.95
Mean   :56.54           Mean   :14.81           Mean   :13.36           Mean   :60.40
3rd Qu.:67.00           3rd Qu.:24.48           3rd Qu.:22.85           3rd Qu.:74.15
Max.   :93.16           Max.   :33.68           Max.   :38.57           Max.   :98.69
NA's   :145             NA's   :120             NA's   :120             NA's   :120

plant1_train.cond_loc1 plant1_train.cond_loc2 plant1_train.cond_loc3
Min.   :0.0000           Min.   :0.0000           Min.   :0.0000
1st Qu.:0.0000           1st Qu.:0.0000           1st Qu.:0.0000
Median :0.0000           Median :0.0000           Median :0.0000
Mean   :0.0048           Mean   :0.00831          Mean   :0.01101
3rd Qu.:0.0000           3rd Qu.:0.0000           3rd Qu.:0.0000
Max.   :1.0000           Max.   :1.0000           Max.   :1.0000
NA's   :870             NA's   :120             NA's   :145

```

예시)

`summary(plant1_train)`

### 3. 데이터 저장

분석 결과 및 중간결과 확인을 위해 파일형태로 저장하여 반출할 수 있으며, `write.csv`함수를 이용하여 분석이 용이한 CSV파일로 저장할 수 있습니다.

사용자 개개인의 환경이 다르므로 `write.csv`함수 사용시, `fileEncoding` 옵션 설정하여 데이터를 저장합니다.

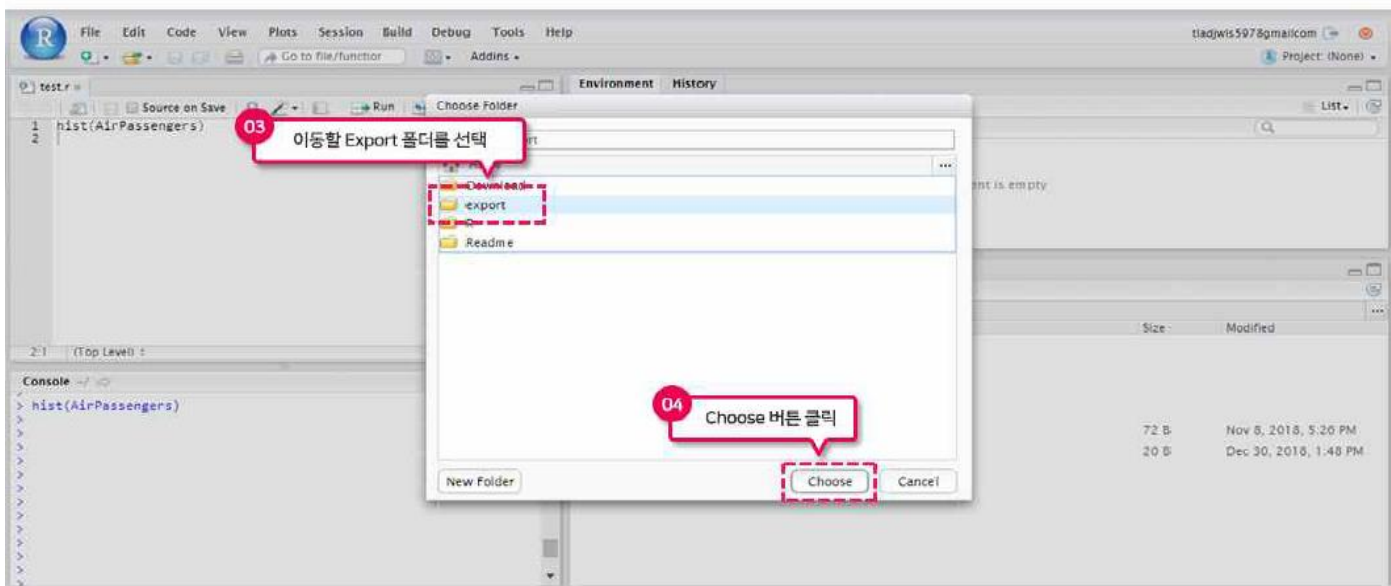
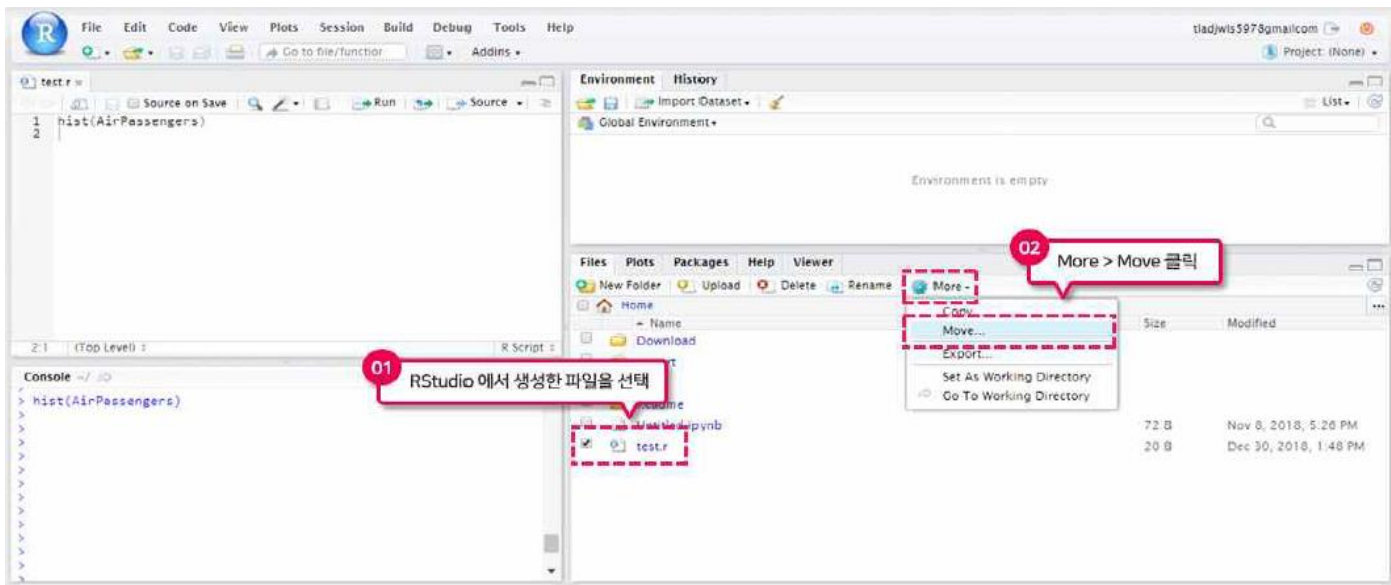
```

7 write.csv(plant1_train, "plant1_train.csv", fileEncoding="utf-8")
8 write.csv(plant1_train, "plant1_train.csv", fileEncoding="euc-kr")
9 write.csv(plant1_train, "plant1_train.csv", fileEncoding="cp949")

```

## 5-1. 분석결과 데이터 다운로드

분석결과 데이터를 다운로드하기 위해 파일을 이동시키고, 분석결과 데이터 다운로드를 신청합니다.





## 분석환경

기상기후 빅데이터를 다른 분야 데이터와  
 접목시켜 분석할 수 있는 환경을  
 제공합니다.

[바로가기](#)

분석교육실습

R을 활용한 분석 | R 교육 동영상 | Python을 활용한 분석 | Fortran을 활용한 분석

비정형 도구

데이터 시각화

빅데이터 분석도구

R studio | Python | Fortran

데이터

기상데이터 | 업로드 데이터 | 웹데이터

마이페이지

나의 이용 현황 | 1:1 상담 | 비밀번호 재설정

분석결과 다운로드

분석결과 다운로드

## 분석결과 다운로드

### ■ 분석결과 다운로드 신청 순서

**01 step**

분석결과파일을 R-Studio에서 분석서버 홈 디렉터리 아래의 export 폴더에 이동시킵니다.

**02 step**

[분석결과 다운로드 신청] 화면에서 다운로드 할 파일을 선택하고, 다운로드 신청 사유를 선택한 후 확인 버튼을 클릭합니다.

**03 step**

관리자에 의해 분석결과 파일 검토 후 승인이 완료되면 분석결과파일을 [분석결과 다운로드 현황] 화면에서 다운로드 받습니다.

### ■ 유의사항

- 다운로드 대상은 분석결과에 한하여 가능합니다.
- 다운로드 기간은 승인 완료일로부터 일주일까지이며 해당 기간 이후에는 자동 삭제됩니다.

다운로드 가능 데이터

☒ test.r

**01** 다운로드 할 분류 체크

분석결과 다운로드 사유

선택

**02** 다운로드 사유 선택

**분석결과 다운로드 신청**
**03** 분석결과 다운로드 신청 클릭