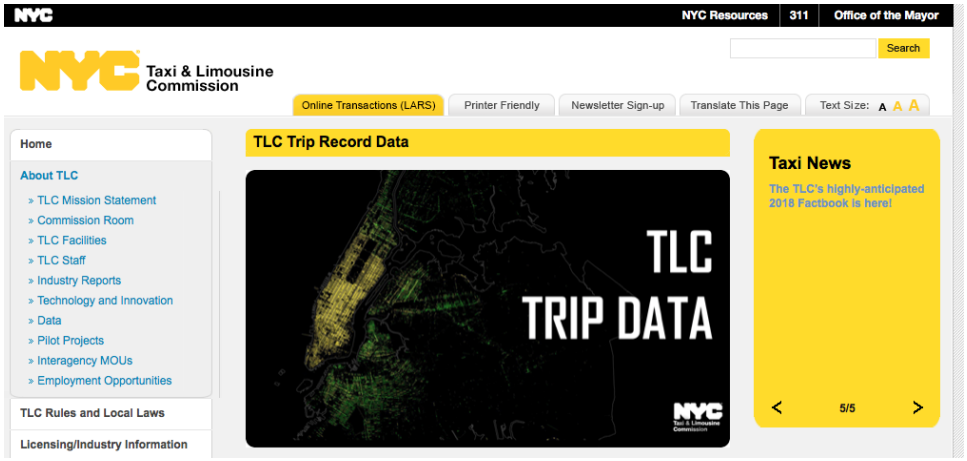
기말프로젝트 및 빅데이터 경진대회 문제



- 자료는 하둡 클러스터의 hdfs "/user/shcho/data/tlc_taxi" 디렉토리 안에 업로드 되어 있음
- 총 12개 csv 파일로 구성, 총 10.1 GB

```
1  /user/shcho/data/tlc_taxi/yellow_tripdata_2017-01.csv
2  /user/shcho/data/tlc_taxi/yellow_tripdata_2017-02.csv
3  /user/shcho/data/tlc_taxi/yellow_tripdata_2017-03.csv
4  /user/shcho/data/tlc_taxi/yellow_tripdata_2017-04.csv
5  /user/shcho/data/tlc_taxi/yellow_tripdata_2017-05.csv
6  /user/shcho/data/tlc_taxi/yellow_tripdata_2017-06.csv
7  /user/shcho/data/tlc_taxi/yellow_tripdata_2017-07.csv
8  /user/shcho/data/tlc_taxi/yellow_tripdata_2017-08.csv
9  /user/shcho/data/tlc_taxi/yellow_tripdata_2017-09.csv
10 /user/shcho/data/tlc_taxi/yellow_tripdata_2017-10.csv
11 /user/shcho/data/tlc_taxi/yellow_tripdata_2017-11.csv
12 /user/shcho/data/tlc_taxi/yellow_tripdata_2017-12.csv
```

- 2017년도 뉴욕 도시의 Yellow Taxi 운행 기록 자료 포함
- 자료에 대한 더 자세한 설명은 다음의 사이트를 참조할 것

  Yellow Taxi Trip Records 사이트 클릭

- 변수설명 (Data Dictionary)

Field Name	**Description**
VendorID	A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC 2= VeriFone Inc.
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
Pickup_longitude	Longitude where the meter was engaged.
Pickup_latitude	Latitude where the meter was engaged.
RateCodeID	The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka “store and forward,” because the vehicle did not have a connection to the server. Y= store and forward trip N= not a store and forward trip
Dropoff_longitude	Longitude where the meter was disengaged.
Dropoff_ latitude	Latitude where the meter was disengaged.
Payment_type	A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips

- 자유롭게 흥미로운 문제를 정의하고 문제를 만든 배경, 분석 과정, 분석 결과물 등을 정리하여 보고서로 제출하면 됩니다.
- 데이터 분석 결과, 통계 방법을 빅데이터에 적용할 수 있는 함수를 개발하여 만든 R package, 새로운 분석 방법 등 모두 가능합니다.
- 예시
  1. Big Data Visualization Tool 개발
  2. JFK 공항에서 Central Park까지의 운행이 날씨의 영향을 받나?

3. JFK 공항에서 Time Square까지 평균적 요금은? 회사별, 운전사별 차이가 있었나?
4. logistic regression 분석을 실행할 수 있는 package 개발

- 결과보고서 제출 안내
  - 팀(최대 4인)을 구성하여 프로젝트를 수행 (다른 학과 학생들로 구성 시 가산점)
  - 12/18 화요일 저녁 11시 59분까지 A4 용지 12장 이내 분량의 분석보고서와 R 코드를 Smart Campus를 통해 제출 (숙제 제출 때와 동일한 방식)
  - 참여 구성원의 역할을 독립된 페이지(12장 제한에 포함되지 않음)에 구체적으로 기술 할 것 (프로젝트를 위해 모임을 갖은 날짜, 시간, 장소, 내용, 역할 등)
  - 최종 3 $\sim$ 4 팀을 선정할 예정
  - 선정된 팀은 빅데이터 경진대회 당일 공개 발표 예정

- 도움말
  - 시간이 제한되어 있으니 실현가능한 계획을 세울 것
  - 하둡 클러스터에 분석을 실행하기 전에 sample 자료를 사용하여 코딩에 문제가 없는지 확인할 것!