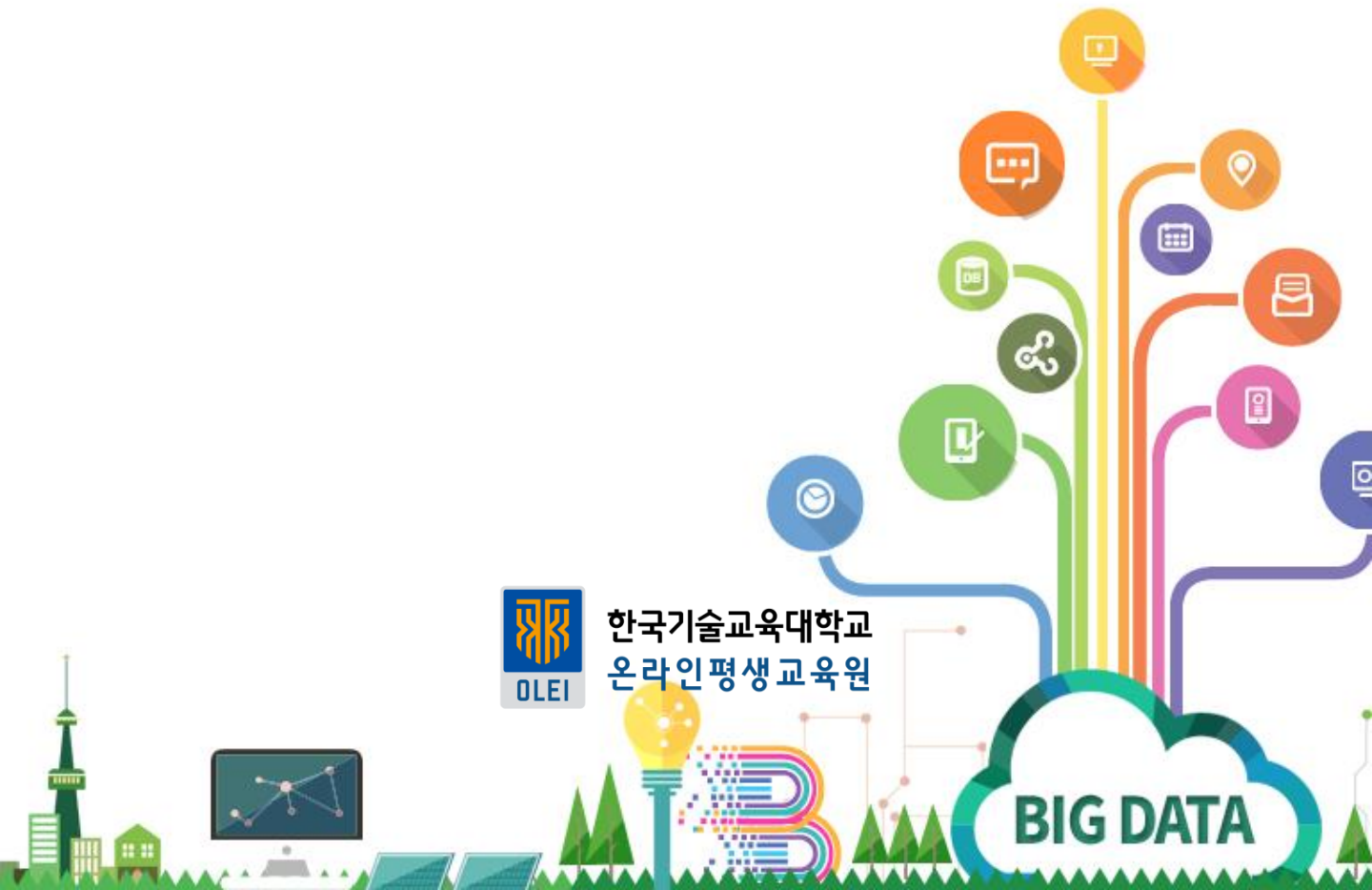
빅데이터 수집 part 2

R을 이용한
공공 데이터 수집 따라하기



한국기술교육대학교
온라인평생교육원

학습목표

- R을 활용한 부동산 매매 데이터 수집 및 분석
- R을 활용한 전세가율 분석

- R을 활용하여 부동산 매매 데이터를 수집하고 분석을 위한 시각화를 할 수 있다.
- R을 활용하여 전세가율 데이터를 수집하고 분석을 위한 시각화를 할 수 있다.

1 실습개요

데이터 원천

- 국토교통부 실거래가 공개시스템(http://rt.molit.go.kr/)

데이터 수집 및 정리 코드

- 웹 데이터를 수집하는 코드는 웹페이지의 구조가 변경되면 코드도 변경됨
- result_sales_dt.Rdata 파일은 GitHub에서 제공

➢ 출처 : 전희원, R을 이용한 부동산 데이터 분석 케이스 스터디, freesearch.pe.kr, 2016.

2 실습절차

1. 부동산 매매 데이터 수집 및 로딩
2. ggplot2를 활용한 분기별 아파트 매매건수의 시각화
3. 지역별 추이의 시각화
4. 시계열의 랜덤성 검정
5. 서울 지역 월 단위 매매건수 예측(계절성 ARIMA 모형 사용)

참고

ARIMA(Auto-Regressive Integrated Moving Average) 모형

- 시계열 분석 기법의 한 종류
- 과거의 관측값과 오차를 사용하여 현 시계열 값을 설명하는 ARMA 모델을 일반화한 것

➢ 출처 : 전희원, R을 이용한 부동산 데이터 분석 케이스 스터디, freesearch.pe.kr, 2016.

3 실습보기

1 부동산 매매 데이터 수집

- 웹 데이터를 수집하는 코드는 웹페이지의 구조가 변경되면 코드도 변경됨
- result_sales_dt.Rdata 파일은 GitHub에서 제공

2 패키지 설치

- library( )로 필요한 Library 설치
- 필요 시 install package( )로 필요한 패키지 설치

주요 패키지 설명

패키지	설명
data.table	■ data.frame과 같은 데이터 저장 클래스를 공유하는 패키지 ■ data.frame에 비해 수십~수백 배 빠른 데이터 처리 능력을 보여주며 간단한 코드로 다양한 데이터 전처리를 할 수 있게 함 ■ 오백만 건 이상의 레코드의 데이터를 사용해야 되기 때문에 해당 패키지를 사용함
dplyr	■ data.table과 함께 사용이 되며 파이프 연산자를 통해 코드의 가독성을 높여줌 ■ data.table 혹은 data.frame이 든 원본 소스에 상관없이 같은 전처리 코드로 다양한 소스 데이터를 다룰 수 있게 해줌
ggplot2	■ 대표적인 R 기반 시각화 도구
lubridate	■ R에서 다소 복잡한 시간에 관련된 데이터를 쉽게 다룰 수 있게 해주는 패키지
stringr	■ 문자열 처리를 위한 패키지
forecast	■ 시계열 분석을 위한 패키지
rand tests	■ 랜덤성을 검정하는 패키지

3 실습보기

3 데이터 얼개 살펴보기

```
> glimpse(result_sales_dt, width=60)
```

4 분기별 아파트 매매건수

```
> qrt_cnts <- data.table명 [where , select, group by]

> ggplot(qrt_cnts, aes(x=yyyyqrt, y=N,group=1)) +
  geom_line() + xlab("년도분기") + ylab("매매건수") +
  theme(axis.text.x=element_text(angle=90)) +
  stat_smooth(method='lm')
```

- ggplot2를 활용한 시각화
- data.table명 [where, select, group by]
  - where : 데이터를 조건에 맞게 필터링하는 곳
  - select : 어떠한 필드를 보여줄지 선택하는 곳
  - group by : 어떠한 기준으로 데이터를 요약해서 보여줄지 결정하는 곳
- 쿼터별로 매매수(.N은 group by 조건에 해당되는 레코드 수를 리턴하는 함수)를 카운팅하여 qrt_cnts라는 이름의 data.table 객체를 만듦

  ※ data.table 객체는 data.frame 객체를 입력 받는 gplot( )과 같은 함수에 그대로 적용이 가능해서 별도의 변환작업 없이 활용이 가능함
- 데이터와 그래프로 표현 되는 미적(Aesthetic) 객체를 어떻게 매핑시키는지 서술
  - X축 : data .table 객체의 쿼터컬럼(yyyyqrt)
  - Y축 : 쿼터별 매매횟수(N)
  - 보여줄 시계열은 1종류라는 것을 group 파라미터로 명시
- theme
  - 각 축이나 레이블에 다양한 표현을 하기 위해서 제공되는 명령어
  - X축의 레이블 텍스트가 겹치는 현상을 없애기 위해 명령어로 텍스트를 90° 회전하여 표현
- stat_smooth
  - X, Y 변수간의 선형, 혹은 비선형적인 패턴을 시각화하기 위해 주로 사용됨
  - 여기서는 선형회귀 모형으로 피팅된 값을 뿌려주도록 사용

3 실습보기

5 분기별/지역별 매매추이

```
> region_cnts <- result_sales_dt[yyyyqrt !=
  '2015Q2',.N,.(yyyyqrt,region)]
> ggplot(region_cnts, aes(yyyyqrt, N,group=region))
  geom_line() + facet_wrap(~region,scale='free_y', ncol=3)
  stat_smooth(method = 'lm')
  theme(axis.text.x = element_blank())
```

- 쿼터별 지역별 매매량 계산 추가
  : group by 절에 region을 추가해서 쿼터별 지역별 매매량을 계산하게 함
- X 레이블 제거(지면 여건상 표현 변경)

3 실습보기

6 시계열의 랜덤성 검정

```
#월별 지역별 매매량
> region_cnts <- result_sales_dt[,.N,.(yyyymm,region)]

#대표지역 추출
> regions <- unique(region_cnts$region)

#각 지역별로 매매량의 랜덤성 검정 결과를 runs_p 변수에 추가
> runs_p <- c()
> for(reg in regions){
  runs_p <- c(runs_p, runs.test(region_cnts[region %chin%
  reg,N])$p.value)
  }
> ggplot(data.table(regions, runs_p), aes(x=regions, y=runs_p,
  group=1)) +
  geom_line() + geom_point() +
  ylab('P-value') + xlab('지역')
```

- 각 지역별로 매매량의 랜덤성 검정 결과를 runs_p 변수에 추가
- P-value
  : 년도별 지역별 아파트 매매량의 변동이 랜덤하다는 귀무가설이 참이라 가정할 때 관측값이 나올 확률

3 실습보기

7 시계열 분할(서울지역)

```
> seoul_cnts <- result_sales_dt[yyyymm != '201504' &
  region %chin% '서울',.N,.(yyyymm)]
> tot_ts <- ts(seoul_cnts$N,start = c(2006,1), frequency = 12)
> plot(stl(tot_ts,s.window = 'periodic'))
```

- 대표적인 시계열 패턴
  - 트랜드(Trend) : 장기적으로 나타나는 변동 패턴
  - 시즈널(Seasonal) : 주, 월, 분기, 반기 단위 등 시간의 주기로 나타나는 패턴
  - 주기(Cyclic) : 최소 2년 단위로 나타나는 고정된 기간이 아닌 장기적인 변동

8 시계열 분할에 대한 모형가정

- 계절성(Seasonal) ARIMA 모형
  - 과거의 관측값과 오차가 지금 현재의 시계열 값을 결정한다는 ARIMA 모형에 불안정 시계열을 안정 시계열로 만드는 I를 결합한 모형

```
> arima_mdl <- auto.arima(tot_ts)
> tsdiag(arima_mdl)
```

9 서울지역 아파트 매매량 예측

```
> plot(forecast(arima_mdl,h=8))
```

1 실습절차

1. 전세 매매 관련 데이터 수집 및 로딩
2. 데이터 결합 및 전세 데이터 추출
3. 데이터 시각화(비선형 모형)
4. 연도별 전세율 시각화(Boxplot)
5. 서울시 구별 전세가율 추이(Boxplot)

➢ 출처 : 전희원, R을 이용한 부동산 데이터 분석 케이스 스터디, freesearch.pe.kr, 2016.

2 실습보기

1 데이터 로드

```
> load('result_rents_dt.RData')
```

- 수집한 전세 매매 데이터 로딩
- load 명령어로 데이터 로드

2 데이터 결합

```
> rents <- result_rents_dt[type_of_rent %chin% '전세']
> rent_sales <- rents %>%
  inner_join(result_sales_dt,
  by=c('si_gun_gu', 'm_bun', 's_bun', 'dangi','area', 'yyyymm',
  'floor')) %>%
  data.table
```

- 월세와 전세 데이터가 혼합되어 있기 때문에 전세 대이터만 사용하기 위해 join을 이용하여 데이터 결합
- join
  - 한 데이터베이스 내의 여러 테이블의 레코드를 조합하여 하나의 열로 표현
  - 테이블로 저장되거나 그 자체로 이용할 수 있는 데이터셋을 만드는 기능
  - 두 개의 테이블에서 각각의 공통값을 이용함으로써 필드를 조합함
- 내부 join(inner_join)
  - join 구문에 기반한 두 개의 테이블, 즉 A, B의 컬럼값을 결합함으로써 새로운 결과 데이터 테이블 생성
  - 그 질의어는 join 구문을 충족하는 모든 일치되는 결과열을 찾기 위해 A, B 테이블의 각 열을 비교
  - join 구문이 충족되면 일치된 각 열의 컬럼값은 결과열로 결합

2 실습보기

3 매매가 대 전세가 비교

```
> ggplot(rent_sales[qrt.y %chin% 'Q1'] %>%
  sample_frac(0.5), aes(x=price, y=base_price, colour=yyyy.x)) +
  geom_point(alpha=0.7) +
  stat_smooth(method='gam', formula = y~ s(x, bs='cs'), size=1.7)
  + scale_color_brewer('yyyy',palette = "Set1") + xlab('매매가(만원)')
  + ylab('전세가(만원)')
```

- sample-frac
  : 다소 많은 데이터를 모두 시각화하지 않고 전체 50%의 데이터만 시각화
- 매매가격에 따라 전세가의 변화를 좀더 자세히 보기 위해선 선형모형 보다 비선형 모형으로 피팅해 이를 시각화된 결과와 함께 제시
- 계절적으로 매매가가 다른 패턴이 보이므로 매년의 1분기(Q1)만 시각화함

4 전세가율(Boxplot)

```
> rent_sales_ratio <- rent_sales[,rent_ratio:=base_price/price]
> ggplot(rent_sales_ratio, aes(x=region.x, y=rent_ratio)) +
  geom_boxplot(aes(fill=yyyy.x),outlier.size=0.5) +
  scale_y_continuous(breaks=seq(0,1.5, by=0.1), limits=c(0,1.5)) +
  xlab("지역") + ylab('전세가율') +
  scale_fill_discrete('년도')
```

- data.table의 select절
  : = 연산자는 새로운 전세가율(rent_ratio) 필드를 base_price, price를 이용해서 생성하라는 코드

2 실습보기

5 서울시 전세가율 추이(구별)

```
> seouls_gu <- rent_sales[si_gun_gu %like% '서울특별시',
  gu:=sapply(str_trim(si_gun_gu),
  function(x){str_split_fixed(x,pattern = ' ', n=3)[1,2]})]
> ggplot(seouls_gu, aes(gu, rent_ratio)) +
  geom_boxplot(aes(fill=yyyy.x),outlier.size=0.5) +
> scale_y_continuous(breaks=seq(0,1.2, by=0.1), limits=c(0,1.2)) +
  xlab("지역") + ylab('전세가율') + scale_fill_discrete('년도') +
  theme(axis.text.x=element_text(angle=45))
```

- 구(gu)라는 새로운 필드를 만들기 위해 시군구(si_gun_gu)에 포함된 주소 정보를 기반으로 두 번째에 구정보가 포함되었다는 가정 하에 첫 번째 공백 다음에 나온 문자열을 gu 필드에 추가하는 코드

R을 활용한 부동산 매매 데이터 수집 및 분석

+ 실습개요

- 데이터 원천 : 국토교통부 실거래가 공개시스템(http://rt.molit.go.kr/)
- 웹 데이터를 수집하는 코드는 웹페이지의 구조가 변경되면 코드도 변경됨
  (result_sales_dt.Rdata 파일은 GitHub에서 제공)

+ 실습절차

① 부동산 매매 데이터 수집 및 로딩
② ggplot2를 활용한 분기별 아파트 매매건수의 시각화
③ 지역별 추이의 시각화
④ 시계열의 랜덤성 검정
⑤ 서울 지역 월 단위 매매건수 예측(계절성 ARIMA 모형 사용)

요점정리

R을 활용한 전세가율 분석

+ 실습절차

① 전세 매매 관련 데이터 수집 및 로딩
② 데이터 결합 및 전세 데이터 추출
③ 데이터 시각화(비선형 모형)
④ 연도별 전세율 시각화(Boxplot)
⑤ 서울시 구별 전세가율 추이(Boxplot)

요점정리

용어사전

용어	내용
ARIMA 모형	• 경제가 사람들의 과거지식과 경험에 기초한 행동에 따라 움직이고 있음을 중시한 시계열분석의 사고방식을 기초로 한 모델