

R과 머신러닝을 활용한 최적의 신제품 가격 결정

곽 상 욱

<https://github.com/Trigger21>



Contents

1. 소개 및 배경

- 1.1 프로젝트 팀 소개
- 1.2 프로젝트 배경
- 1.3 프로젝트 진행 과정
- 1.4 클라이언트 소개
- 1.5 프로젝트 목표

2. 실무 진행

- 2.1 빅데이터 분석 소개
- 2.2 분석 조건
- 2.3 데이터베이스 수집/구축/정제
- 2.4 시각화, 포지셔닝 확인
- 2.5 Tool, 머신러닝 기법 소개
- 2.6 머신러닝 분석 결과
- 2.7 해당 분야 모델 적용_ EVE

3. 결론 및 정리

- 3.1 가격 산출 결론
- 3.2 정리

1.1 프로젝트 팀 소개

팀 명 : DAT(닷)_ **Do All Things**

Member : 곽상욱, 강윽, 이수원, 이웅기, 장연철, 김명중

DAT : Do All Things 의 약자로

모든지 열심히 하겠습니다 라는 팀의 정신을 담고있는 팀명입니다.

1.2 프로젝트 배경

포트폴리오는 기업과의 협업 프로젝트여야 한다

나의 질문

[내가 실제 갖춰야 할 역량은?]

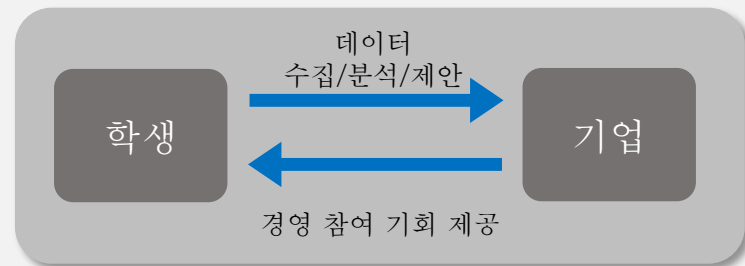
경영 전략 수립, 의사결정 과정에
빅데이터 분석이
언제 • 어떻게 • 왜 • 얼마나
활용되는가?

기업은 어떤 문제에 당면해 있으며
데이터 분석가는 어떤 해안을
제시할 수 있는가?

질문에 대한 답

[답은 기업 내부에 있다]

데이터 분석의 필요성은 통감하나
부서 개설과 인력 고용이 현실적으로
불가능한 소기업 • 자영업 대표에게
[산학 협력]의 형태를 빌려 협업을 제안



1.3 프로젝트 진행 과정



1.4 클라이언트 소개

| | |
|--------|---|
| 기업 명 | OOOO (클라이언트에 의해 공개 불가) |
| 기업 형태 | 스타트업 (5인 기업) |
| 설립 연도 | 2005년 2월 |
| 주요 상품 | 마스크 팩, 선크림(18년 6월 런칭 예정) |
| 주요 거래처 | 아트박스, 이랜드 버터, 신라면세점 등 14개 대형 유통사 |
| 특징 | 유통사로부터 데이터를 제공받지 못하고 있다. 독자판매 채널인 홈페이지 판매가 이뤄지고 있으나 미미하다. 런칭 예정인 선크림의 적정 가격 결정이라는 문제에 직면해 있다. |

1.5 프로젝트 목표

클라이언트의 질문

새로 출시될 선크림의 적정 가격은 얼마인가?

데이터 분석 & 결과 도출 과정



2.1 빅데이터 분석 소개

클라이언트 대표는 [빅데이터 분석] 을
처음 접하기 때문에 기초 설명부터 진행

빅데이터 분야가 도래된 계기

[과거의 컴퓨터]

거대 모집단을 감당할 수 없는
성능적 한계

프로그램(Tool)의 한계

초고성능 컴퓨터의 비용 문제

[현재의 컴퓨터]

하드웨어 성능의 폭발적 발전

소프트웨어의 다양화, 무료화,
Open source

고성능 하드웨어의
경량화, 저비용화

아주 넓게 보니, 안보이던 패턴이 보이기 시작

아는 만큼 보인다' 하던 패턴이 바뀌기 시작

빅데이터 분석 소개_ 5V

Volume

Facebook의 저장 사진 용량
30,000 TB (30,000,000 GB)

Veracity

표본이 크므로
정확도가 높다

Value

분석 결과는
조직이 당면한 문제를
해결할 수 있어야 한다

Velocity

대용량 분석에
대응하는 처리 속도

Variety

숫자 · 문자 · 동영상
사진 · 소리 · 언어

경영, 기술, 학술 활동에 필요한 **Decision**을 데이터에 의해 근거함에 있어
Excel 수준에서는 발견할 수 없는 패턴을 빠르고 정확하게 얻고

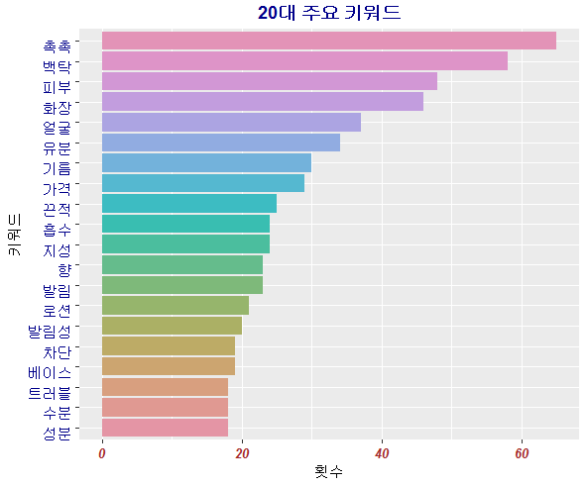
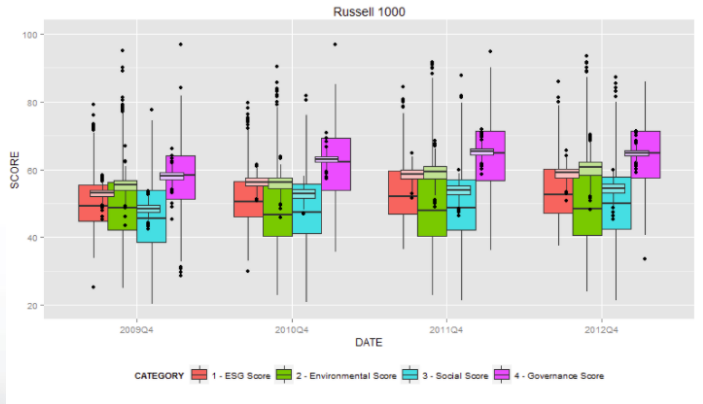
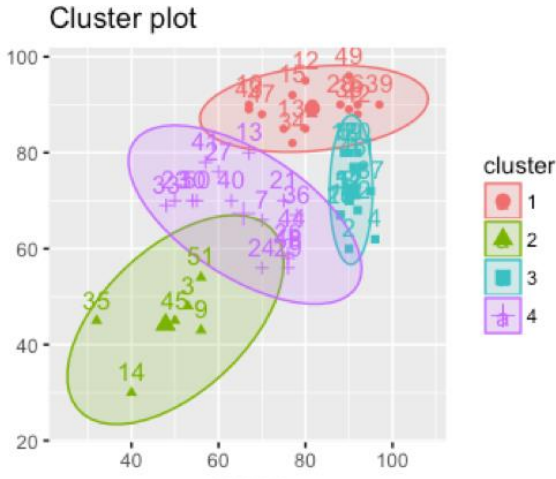
이를 **통찰력**과 새로운 **아이디어**로 연결하기 위함

이를 **통찰력**과 새로운 **아이디어**로 연결하기 위함

Excel 수준에서는 발견할 수 없는 패턴을 빠르고 정확하게 얻고
이를 '경영' 활동 활용에 필요한 **Decision**을 데이터에 의해 근거함에 있어

2.1 빅데이터 분석 소개

데이터 시각화 예시를 소개,
분석 결과의 해석이 용이함을 설명



2.2 분석 조건

자체 데이터 전무

단독 판매 채널이 없이 대형 유통사만의 거래로 인해
기존 제품이 누가? 언제? 어떻게? 왜? 구입하는지 알 수 없다

제작 단가 공개 불가

기업 보안 문제로 단가 공개 불가

유일한 데이터 : 분석 대상의 특징

선크림의 용량, 자외선 차단지수, 미백효과 유무 등의 몇 가지 특징만 제공

2.3 데이터베이스 수집 / 구축/ 정제

R을 활용, NAVER 지식쇼핑 [선크림] 웹크롤링

웹크롤링 수집 데이터

제품명

브랜드

용량

판매가

판매처 수

SPF

PA

피부타입

용기타입

자체 조사 데이터 #1

1 ~ 5 점수 부여

백탁

미백

톤 보정

보습

유분기

끈적임

잔여감

유무기자차

저자극

자체 조사 데이터 #2

유 / 무 판단

자체케릭터

안티에이징

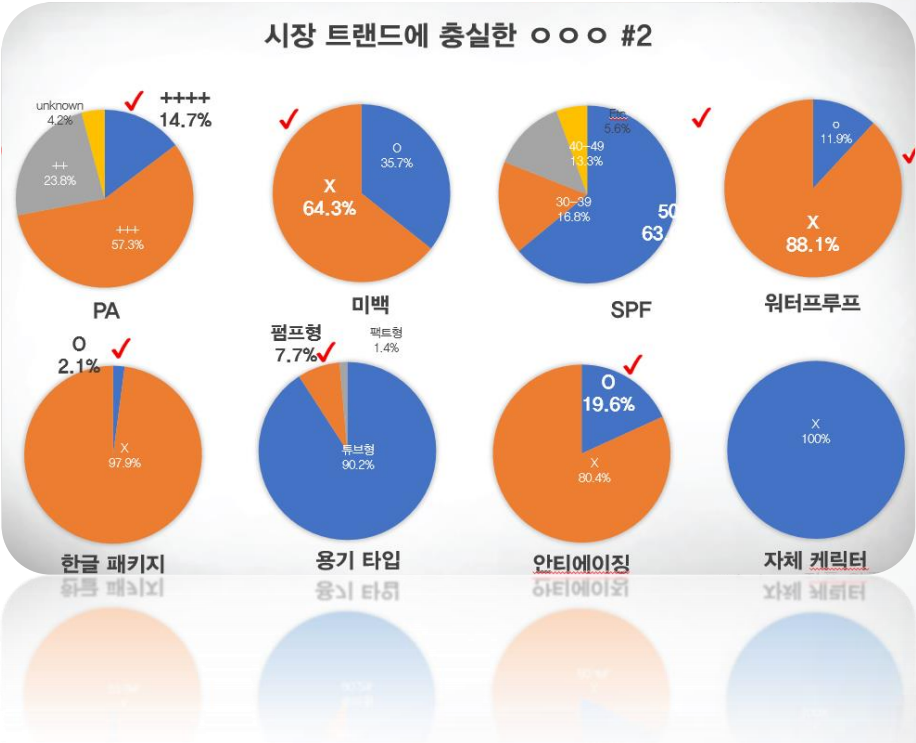
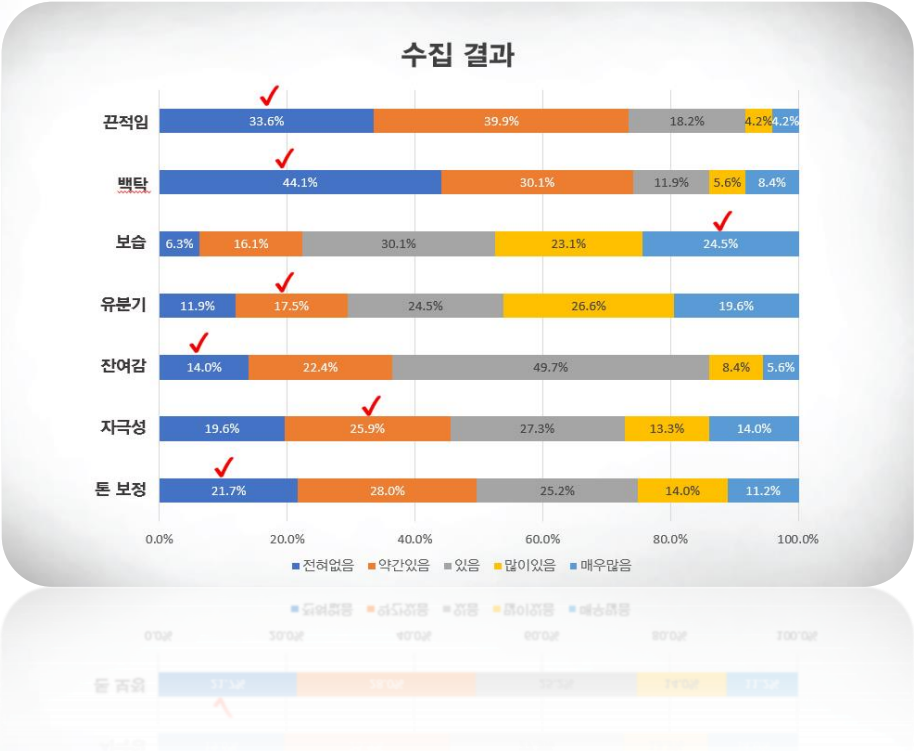
한글패키지

베이스검용

워터프루프

1,231개 제품에 대한 23개 항목 데이터 수집

2.4 데이터베이스 시각화, 포지셔닝 확인



크롤링된 경쟁품과 클라이언트 제품의 비교

2.5 사용된 Tool, 머신러닝 기법 소개



R Studio

KNN 알고리즘 (K-nearest neighbors)

어떤 제품의 가격 예측 시,
그 제품과 유사한 제품을 조사하여,
그들의 평균 가격을 계산,적당한
가격을 예측할 수 있다고 판단

나이브베이즈 알고리즘 (Naïve Bayes Classification)

23개 항목은 서로
독립되어 있고, 독립된 특성을
분류하는 데 강점이 있는 알고리즘.
다른 머신러닝 분류에 비해 효율적이
며, 실제 상황에서 잘 작동하는 특성

2.6 머신러닝 분석 결과

| KNN |
|-----------------|
| 30,000 ~ 39,999 |
| 35,000 |
| 25,000 ~ 27,999 |
| 27,500 |
| 70% |

정가
정가 중간 값
판매가
판매가 중간 값
신뢰도

| 나이브 베이즈 |
|-----------------|
| 39,500 ~ 46,100 |
| 42,800 |
| 26,300 ~ 29,200 |
| 27,750 |
| 88% |

2.7 해당 분야 모델 적용_ 가격 결정모델

EVE (Economic Value Estimation) Model

제품의 경제적 가치는 경쟁제품의
가격에 당사 제품을 이용함으로써 추가적으로 연결되는 이득+ 손실

가격 인상 근거

보습 효과 탁월
안티에이징 효과
튜브타입보다 위생적인 펌프 타입

가격 인하 근거

미백 효과 없음
인지도 부족
캐릭터와 기능성의 상반된 이미지

결론 & 정리



3.1 가격 산출 결론

데이터 분석 결과

KNN 27,500

나이브
베이지스 27,750

중간 값
27,625

인상 근거

보습제품 불필요
안티에이징 제품 불필요
펌프 타입

+ 25,000

인하 근거

미백 효과 없음
인지도 부족
캐릭터 디자인

- 25,025

판매가 제안 : 27,600원

3.2 정리

클라이언트

“ 제작 단가 및 모든 상황을 고려해 책정한 가격과 놀랄 정도로 일치 ”

“ 자신의 결정을 데이터 분석으로 재확인하여 확신의 근거를 마련 ”

“ 분석 결과를 정가 / 프로모션 가격 / 판매가 책정에
앞으로 적극 반영할 예정, 다른 부분을 의뢰하고 싶다 ”

프로젝트 팀 DAT

크롤링 후 데이터 분석이 현실을 높은 정확도로 반영한다는 점을 확인

데이터 분석은 수집 / 정제가 70% 라는 것을 다시 확인,
Garbage in garbage out을 항상 염두

사회 현상에 의문을 갖고 데이터로 접근하는
시각을 가져야 한다는 것을 다시 한번 확인 하는 계기

감사합니다 !

Thank You !

<https://github.com/Trigger21>

곽 상 욱

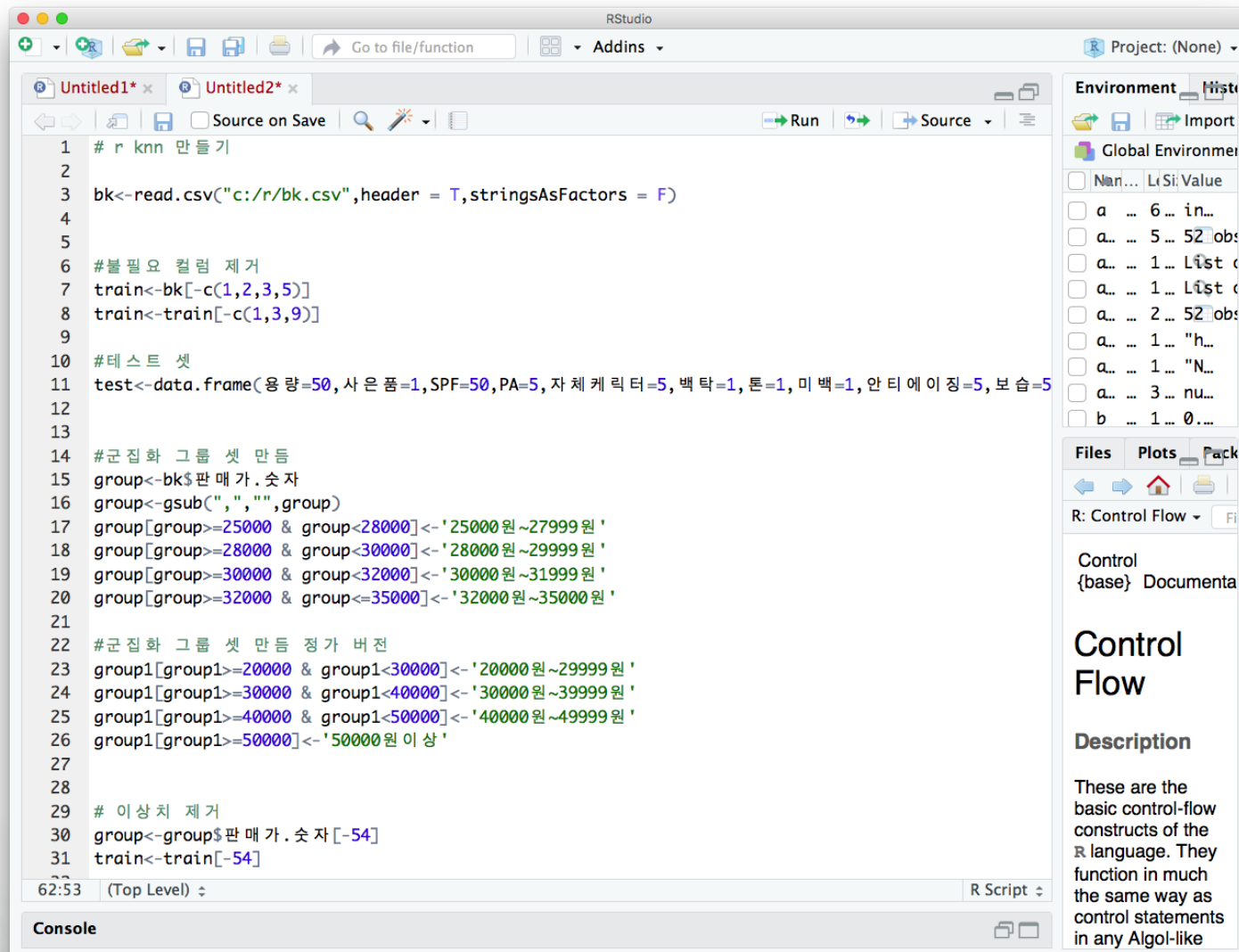
ksh6452@gmail.com

부 록

<https://github.com/Trigger21>

보다 자세한 자료는 깃허브에서 보실수 있습니다.

작업 중 스크린 샷 첨부



```
1 # r knn 만들기
2
3 bk<-read.csv("c:/r/bk.csv",header = T,stringsAsFactors = F)
4
5
6 #불필요 컬럼 제거
7 train<-bk[-c(1,2,3,5)]
8 train<-train[-c(1,3,9)]
9
10 #테스트 셋
11 test<-data.frame(용량=50,사은품=1,SPF=50,PA=5,자체케릭터=5,백탁=1,톤=1,미백=1,안티에이징=5,보습=5)
12
13
14 #군집화 그룹 셋 만들
15 group<-bk$판매가.숫자
16 group<-gsub(" ","",group)
17 group[group>=25000 & group<28000]<-'25000원~27999원'
18 group[group>=28000 & group<30000]<-'28000원~29999원'
19 group[group>=30000 & group<32000]<-'30000원~31999원'
20 group[group>=32000 & group<=35000]<-'32000원~35000원'
21
22 #군집화 그룹 셋 만들 정가 버전
23 group1[group1>=20000 & group1<30000]<-'20000원~29999원'
24 group1[group1>=30000 & group1<40000]<-'30000원~39999원'
25 group1[group1>=40000 & group1<50000]<-'40000원~49999원'
26 group1[group1>=50000]<-'50000원 이상'
27
28
29 # 이상치 제거
30 group<-group$판매가.숫자[-54]
31 train<-train[-54]
```

62:53 (Top Level) R Script

Environment

Global Environment

| | Var | Class | Value |
|--|-----|-------|--------|
| | a | 6 | in |
| | a | 5 | 52 obs |
| | a | 1 | List |
| | a | 1 | List |
| | a | 2 | 52 obs |
| | a | 1 | "h |
| | a | 1 | "N |
| | a | 3 | nu |
| | b | 1 | 0 |

Files **Plots** **Pack**

R: Control Flow

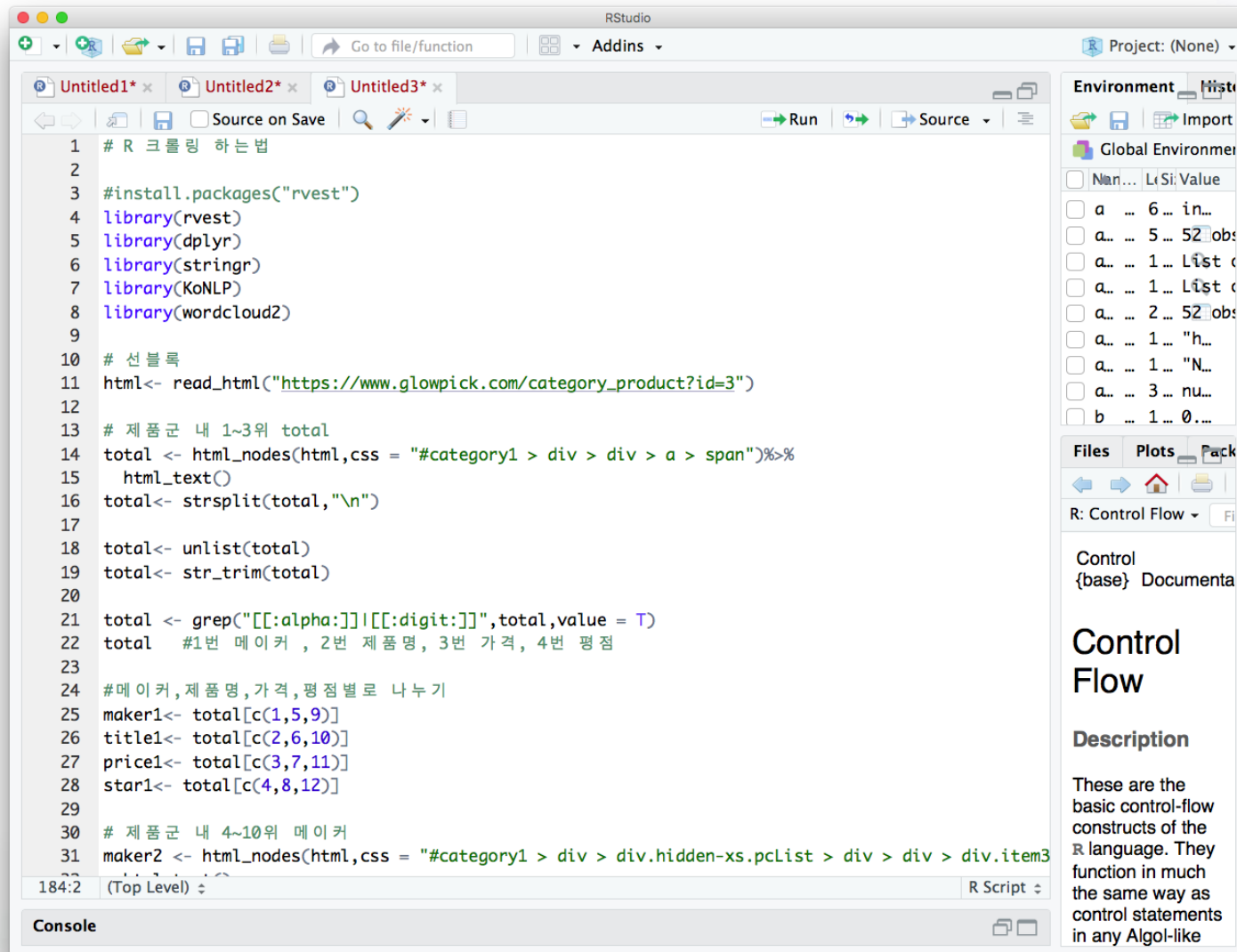
Control {base} Documenta

Control Flow

Description

These are the basic control-flow constructs of the R language. They function in much the same way as control statements in any Algol-like

작업 중 스크린 샷 첨부



The screenshot displays the RStudio interface with a script editor on the left and environment/plots/console panels on the right. The script, titled 'Untitled1*', contains R code for installing packages, reading an HTML page, parsing it with rvest, and extracting product information. The right panel shows the 'Environment' tab with a list of objects, and the 'Plots' tab with a 'Control Flow' section.

```
1 # R 크롤링 하는법
2
3 #install.packages("rvest")
4 library(rvest)
5 library(dplyr)
6 library(stringr)
7 library(KoNLP)
8 library(wordcloud2)
9
10 # 선블록
11 html<- read_html("https://www.glowpick.com/category_product?id=3")
12
13 # 제품군 내 1~3위 total
14 total <- html_nodes(html,css = "#category1 > div > div > a > span")%>%
15   html_text()
16 total<- strsplit(total,"\n")
17
18 total<- unlist(total)
19 total<- str_trim(total)
20
21 total <- grep("[[:alpha:]]|[:digit:]]",total,value = T)
22 total  #1번 메이커 , 2번 제품명, 3번 가격, 4번 평점
23
24 #메이커, 제품명, 가격, 평점별로 나누기
25 maker1<- total[c(1,5,9)]
26 title1<- total[c(2,6,10)]
27 price1<- total[c(3,7,11)]
28 star1<- total[c(4,8,12)]
29
30 # 제품군 내 4~10위 메이커
31 maker2 <- html_nodes(html,css = "#category1 > div > div.hidden-xs.pclList > div > div > div.item3
32
33 (Top Level) ↕
```

Environment

- Global Environment
- Nan... L(Si) Value
- a ... 6 ... in...
- a... 5 ... 52 obs
- a... 1 ... List c
- a... 1 ... List c
- a... 2 ... 52 obs
- a... 1 ... "h...
- a... 1 ... "N...
- a... 3 ... nu...
- b ... 1 ... 0...

Files Plots Pack

R: Control Flow

Control {base} Documenta

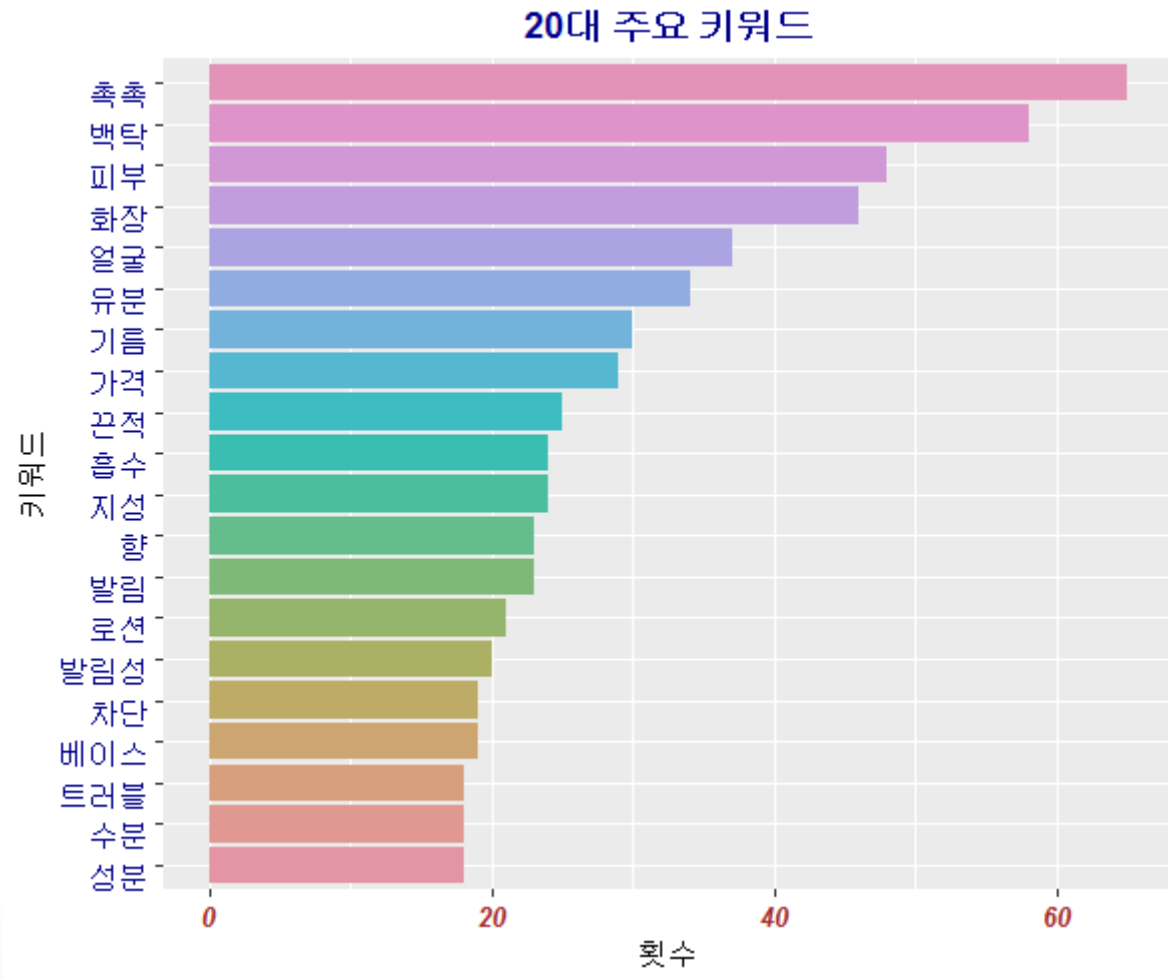
Control Flow

Description

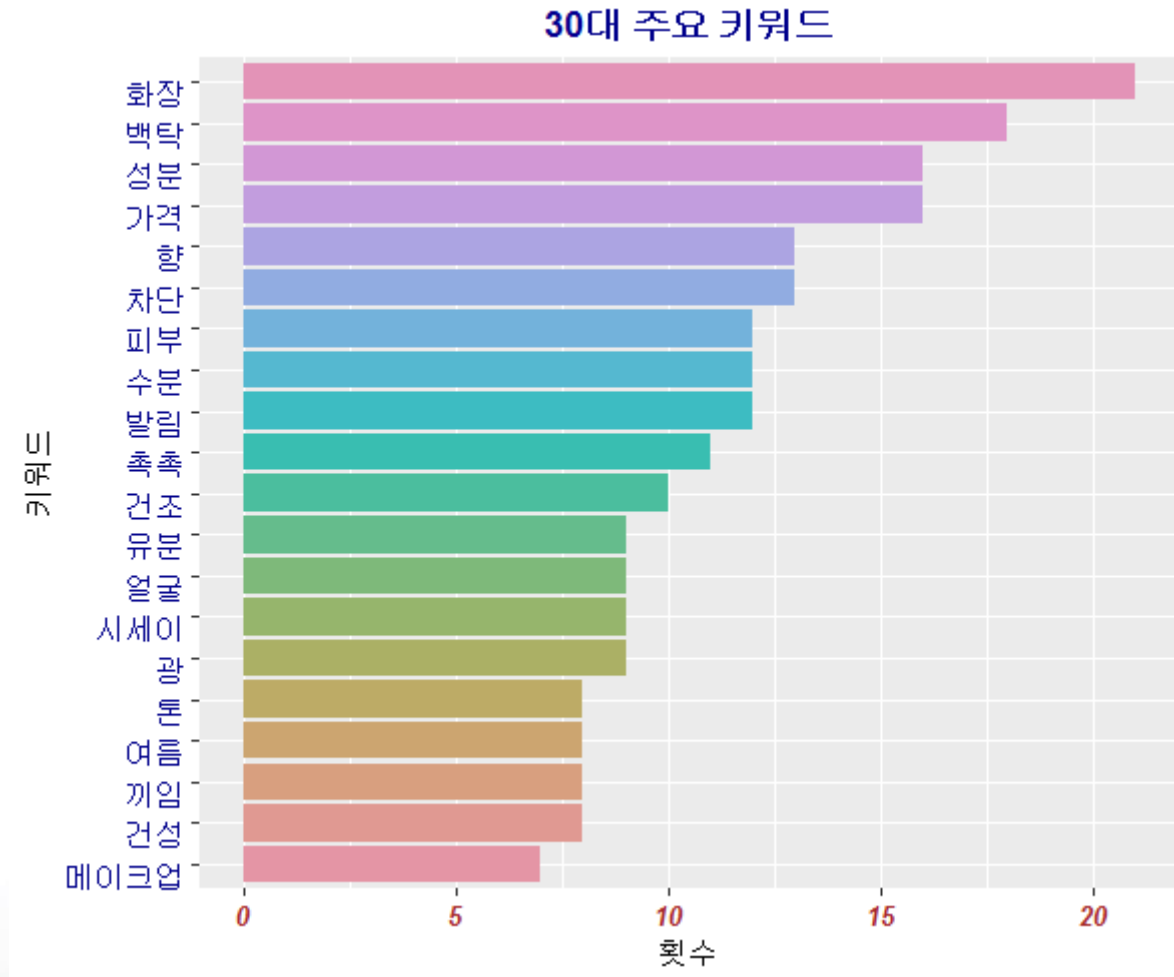
These are the basic control-flow constructs of the R language. They function in much the same way as control statements in any Algol-like

Console

화장품 관련 20대가 주로 쓰는 단어 수집



화장품 관련 30대가 주로 쓰는 단어 수집



화장품 관련 30대가 주로 쓰는 단어 수집

