데이터 사이언스

사단법인 한국 R 사용자회, chatGPT

2023년 03월 02일

목차

서문

한국 R 사용자회는 2022년 2월 28일에 설립되어 1년을 맞이하게 되었습니다. R 사용자회는 R Consortium R User Groups에 등록된 Seoul R Meetup을 운영하며 데이터 사이언스 R / Tidyverse 미트업을 매월 개최하고 있습니다. 또한, R 지식 나눔과 커뮤니티 멤버들 간의 사랑방 역할도 수행하고 있습니다.

2022년부터는 오픈 통계 패키지(BitStat) 패키지를 개발하여 보급하였을 뿐만 아니라 BitNLP, BitSpatial, BitReport 패키지를 개발하여 공개하였습니다. 또한, "디지털 글쓰기", "데이터 과학 프로그래밍", "데이터 시각화", "데이터 과학 기본기", "오픈 통계 패키지", "R 텍스트 마이닝", "데이터 과학 언플러그드" 등의 전자책을 집필하여 무료로 대중에 공개하였습니다. 2023년은 ~~교보정보통신 자회사 디플래닉스 후원을 받아 강남 교보타워에서~~ **chatGPT와 데이터 과학을 주제**로 매월 진행될 예정입니다.

한국 R 사용자회는 AI와 함께 chatGPT 및 연관 AI 도구를 활용하여 데이터 문해력 향상과 디지털 불평등 해소를 위해 데이터 과학 전자책을 제작하여 공개하게 되었습니다.

더 많은 정보는 아래 링크를 참고해주세요.

- 사단법인 한국 R 사용자회: https://r2bit.com/
- Meetup : https://www.meetup.com/seoul-r-meetup/
- Seoul R Meetup : https://r2bit.com/seoul-R/
- Facebook Group : https://www.facebook.com/groups/tidyverse
- Youtube Channel: http://bit.ly/3kzwlkK

Part I

글쓰기

1 R 및 데이터 랭글링 소개

1.1 데이터 과학 및 데이터 랭글링 소개

R은 통계 컴퓨팅 및 그래픽을 위한 프로그래밍 언어이자 소프트웨어 환경입니다. 1990년대 초 뉴질랜드 오클랜드 대학교의 로스 이하카(Ross Ihaka)와 로버트 젠틀맨(Robert Gentleman)에 의해 개발되었습니다. 그 이후로 데이터 분석, 시각화 및 통계 모델링에 가장 널리 사용되는 언어 중 하나가 되었습니다. R은 오픈 소스이므로 누구나 자유롭게 사용할 수 있고 수정할 수 있습니다. (Wickham and Grolemund 2016)

R에는 데이터 과학에 이상적인 여러 가지 기능이 있습니다. 예를 들어, 데이터 분석, 시각화 및 모델링을 위한 패키지 개발에 기여하는 대규모의 활발한 사용자 커뮤니티가 있습니다. 이러한 패키지는 CRAN(종합 R 아카이브 네트워크)에서 다운로드할 수 있으며 기계 학습, 시계열 분석, 데이터 시각화 등 다양한 주제를 다룹니다.

R의 또 다른 장점은 대규모 데이터셋을 처리할 수 있다는 점입니다. R에는 데이터베이스, 스프레드시트, 텍스트 파일 등 다양한 소스에서 데이터를 읽고 조작할 수 있는 다양한 패키지가 있습니다. 또한 데이터 재구성, 데이터 세트 병합, 데이터 집계와 같은 강력한 데이터 조작 기능도 갖추고 있습니다.

R은 학계와 산업계, 특히 금융, 의료, 마케팅과 같은 분야에서 널리 사용됩니다. 학계에서는 통계학, 경제학, 사회과학 등 다양한 분야의 연구와 교육에 R이 사용됩니다. 산업계에서는 기업에서 데이터 분석, 모델링, 시각화 및 데이터 제품 구축에 R을 사용합니다.

전반적으로 R은 데이터 과학자의 도구 상자에서 중요한 도구이며, 그 인기는 계속 증가하고 있습니다. 데이터 분석, 시각화 및 모델링을 위한 다양한 기능을 제공하며, 사용자와 개발자로 구성된 대규모 커뮤니티가 활발하게 활동하고 있습니다. 데이터 과학에 관심이 있다면 R을 배우는 것을 고려해 볼 가치가 있습니다.

1.2 R에서 데이터 랭글링(Wrangling) 및 정리의 기본 개념 소개

데이터 랭글링(Wrangling)과 데이터 정제(Cleaning)는 데이터 분석 프로세스에서 중요한 단계입니다. 의미 있는 분석을 수행하려면 데이터가 사용 가능하고 정확한 형식이어야 합니다. R은 데이터 랭글링 및 정제를 위한 다양한 도구와 함수를 제공합니다.

데이터 랭글링의 첫 번째 단계 중 하나는 데이터를 R로 가져오는 것입니다. R은 CSV 파일, Excel 스프레드시트 및 데이터베이스를 비롯한 다양한 소스에서 데이터를 읽을 수 있습니다. 일반적으로 `read.csv()` 함수는 CSV 파일에서 읽는 데 사용되며, `readxl` 패키지의 `read_excel()` 함수는 Excel 파일에서 읽는 데 사용할 수 있습니다.

데이터를 R로 가져온 후에는 데이터를 정제해야 할 수 있습니다. 여기에는 결측값 제거, 오류 수정, 이상값 처리 등이 포함될 수 있습니다. `na.omit()` 함수는 누락된 값이 있는 행을 제거하는 데 사용할 수 있으며, `is.na()` 함수는 누락된 값을 식별하는 데 사용할 수 있습니다.

데이터 정제에는 데이터를 더 유용한 형식으로 변환하는 작업도 포함될 수 있습니다. 여기에는 데이터 형식 변경, 데이터셋 병합 또는 데이터 형태변경 등이 포함될 수 있습니다. `dplyr` 패키지는 `select()`, `filter()`,

mutate() 등 데이터 조작을 위한 다양한 함수를 제공합니다. tidyr 패키지는 데이터를 와이드(Wide) 형식에서 롱(Long) 형식으로 또는 그 반대로 피벗하는 등 데이터를 재구성하는 데 사용할 수 있습니다.

데이터 랭글링의 또 다른 중요한 단계는 데이터 탐색입니다. 여기에는 데이터를 시각화하여 패턴, 추세 및 이상값을 식별하는 작업이 포함됩니다. ggplot2 패키지는 R에서 데이터 시각화를 위해 널리 사용되는 패키지로, 산점도, 히스토그램, 상자 그림 등 다양한 플롯을 만들 수 있는 함수를 제공합니다.

요약하면, 데이터 랭글링과 정리는 데이터 분석 프로세스에서 중요한 단계입니다. R은 데이터 가져오기, 결측치 처리, 데이터 변환, 시각적 데이터 탐색 등의 기능을 포함하여 데이터 랭글링 및 정리를 위한 다양한 도구와 함수를 제공합니다. 이러한 기본 개념을 숙지하면 R에서 의미 있는 데이터 분석을 수행하는 데 큰 도움이 될 것입니다.

1.3 데이터 랭글링을 위한 기술

데이터 랭글링은 모든 데이터 과학자에게 필수적인 기술이며, 데이터를 관리하고 정리하는 데 도움이 되는 여러 가지 기술과 도구가 있습니다. 이 섹션에서는 데이터 랭글링을 위한 몇 가지 일반적인 기법을 dplyr 및 tidyr 패키지를 사용하여 살펴보겠습니다.

1.3.1 dplyr로 데이터 필터링

dplyr의 filter() 함수는 하나 이상의 조건에 따라 데이터의 하위 집합을 추출하는 데 사용할 수 있습니다. 예를 들어, 'age'라는 열이 있는 'mydata'라는 데이터 프레임이 있는 경우 filter() 함수를 사용하여 연령이 30보다 큰 모든 행을 추출할 수 있습니다:

```
library(dplyr)
mydata_filtered <- mydata %>%
  filter(age > 30)
```

1.3.2 dplyr로 데이터 정렬

dplyr의 arrange() 함수는 하나 이상의 열을 기준으로 데이터를 정렬하는 데 사용할 수 있습니다. 예를 들어 "name" 및 "age"라는 열이 있는 "mydata"라는 데이터 프레임이 있는 경우 arrange() 함수를 사용하여 데이터 프레임을 이름별로 정렬한 다음 연령별로 정렬할 수 있습니다:

```
library(dplyr)
mydata_sorted <- mydata %>%
  arrange(name, age)
```

1.3.3 dplyr로 데이터 병합

기본 R의 merge() 함수는 하나 이상의 공통 열을 기반으로 두 데이터 프레임을 병합하는 데 사용할 수 있습니다. 그러나 merge() 함수는 더 복잡한 병합에는 사용하기 어려울 수 있습니다. dplyr의 조인 함수(left_join(), right_join(), inner_join(), full_join())는 데이터를 보다 유연하고 직관적으로 병합할 수 있는 방법을 제공합니다. 예를 들어, "id"라는 공통 열이 있는 "df1" 및 "df2"라는 두 개의 데이터 프레임이 있는 경우, left_join() 함수를 사용하여 "id" 열을 기준으로 두 데이터 프레임을 병합할 수 있습니다:

```
library(dplyr)
mydata_merged <- left_join(df1, df2, by = "id")
```

1.3.4 tidyr로 데이터 재구성

tidyr의 피벗 함수(pivot_longer(), pivot_wider())는 데이터를 와이드 형식에서 롱 형식으로 또는 그 반대로 재구성하는 데 사용할 수 있습니다. 예를 들어, 서로 다른 연도의 값을 나타내는 여러 열이 있는 데이터 프레임이 있는 경우 pivot_longer() 함수를 사용하여 연도와 값에 대한 열이 있는 긴 형식으로 데이터의 모양을 변경할 수 있습니다:

```
library(tidyr)
mydata_long <- mydata %>%
  pivot_longer(cols = c("year_1", "year_2", "year_3"), names_to = "year", values_to = "value")
```

1.3.5 dplyr로 데이터 집계

dplyr의 summarize() 함수는 하나 이상의 그룹화 변수를 기준으로 데이터를 집계하고 요약 통계를 계산하는 데 사용할 수 있습니다. 예를 들어 "group" 및 "value"라는 열이 있는 "mydata"라는 데이터 프레임이 있는 경우 summarize() 함수를 사용하여 각 그룹에 대한 평균값을 계산할 수 있습니다:

```
library(dplyr)
mydata_summarized <- mydata %>%
  group_by(group) %>%
  summarize(mean_value = mean(value))
```

요약하면, dplyr과 tidyr는 R에서 데이터 랭글링을 위한 강력한 패키지입니다. 데이터 필터링, 정렬, 병합, 집계와 같은 기술을 익히면 데이터를 효과적으로 작업하고 정리하는 데 큰 도움이 될 것입니다.

1.4 R 자료구조

R은 다양한 유형의 데이터로 작업할 수 있는 여러 데이터 구조를 제공하는 데이터 과학을 위한 강력한 언어입니다. 이 섹션에서는 R의 벡터, 행렬, 데이터프레임 및 목록의 기본 개념을 소개합니다.

1.4.1 벡터

벡터(Vector)는 숫자, 문자 또는 논리와 같은 단일 데이터 유형의 데이터를 담을 수 있는 1차원 배열입니다. c() 함수를 사용하여 벡터를 만들 수 있습니다:

```
myvector <- c(1, 2, 3, 4, 5)
```

벡터에 더하기, 빼기, 곱하기, 나누기 등 다양한 연산을 수행할 수 있습니다. 예를 들어

```
myvector * 2
```

1.4.2 행렬

행렬은 숫자, 문자 또는 부울과 같은 단일 데이터 유형의 데이터를 담을 수 있는 2차원 배열입니다. `matrix()` 함수를 사용하여 행렬을 만들 수 있습니다:

```
mymatrix <- matrix(1:9, nrow = 3, ncol = 3)
```

행렬을 더하기, 빼기, 곱하기, 나누기 등 다양한 연산을 행렬에 수행할 수 있습니다. 예를 들어

```
mymatrix * 2
```

1.4.3 데이터프레임

데이터프레임(dataframe)은 각 열이 다른 데이터 유형(예: 숫자, 문자, 요인)을 가질 수 있는 2차원 테이블과 같은 구조입니다. `data.frame()` 함수를 사용하여 처음부터 데이터 프레임을 만들거나 R의 다양한 가져오기 함수 중 하나를 사용하여 파일에서 데이터를 가져와서(예: `read.csv()`, `read_excel()`) 데이터 프레임을 만들 수 있습니다.

```
mydata <- data.frame(
  column1 = c(1, 2, 3),
  column2 = c("value1", "value2", "value3")
)
```

연산자 `$`를 사용하거나 `[]` 연산자를 사용하여 데이터 프레임의 개별 열에 액세스할 수 있습니다:

```
mydata$column1

mydata["column2"]
```

1.4.4 리스트

리스트(list)는 벡터, 행렬, 데이터프레임 및 기타 리스트와 같은 다양한 유형의 객체 모음입니다. `list()` 함수를 사용하여 목록을 만들 수 있습니다:

```
mylist <- list(
  myvector = c(1, 2, 3, 4, 5),
  mymatrix = matrix(1:9, nrow = 3, ncol = 3),
  mydata = data.frame(
    column1 = c(1, 2, 3),
    column2 = c("value1", "value2", "value3")
  )
)
```

연산자 `$`를 사용하거나 `[]` 연산자를 사용하여 목록의 개별 개체에 액세스할 수 있습니다:

```
mylist$myvector
```

```
mylist[[1]]
```

1.4.5 결론

이 섹션에서는 벡터, 행렬, 데이터프레임 및 리스트와 같은 R의 데이터 구조에 대한 기본 개념을 소개했습니다. 이러한 데이터 구조와 그 속성을 이해하면 R에서 다양한 유형의 데이터로 작업하는 데 더 나은 준비가 될 것입니다.

1.5 연습문제와 프로젝트

R에서 데이터 랭글링의 기본 개념을 배운 후에는 이러한 기술을 연습하고 실제 데이터 집합에 적용하는 것이 중요합니다. 다음은 R에서 데이터 랭글링 기술을 연마하는 데 사용할 수 있는 몇 가지 연습 및 프로젝트입니다:

1.5.1 연습문제

1. `dplyr` 패키지를 사용하여 하나 이상의 조건으로 데이터 집합을 필터링합니다. 예를 들어, 고객 리뷰 데이터 집합을 필터링하여 평점이 4점 이상인 리뷰만 포함하도록 할 수 있습니다.
2. `dplyr` 패키지를 사용하여 하나 이상의 열을 기준으로 데이터 집합을 정렬합니다. 예를 들어, 판매 데이터의 데이터 집합을 날짜별로 정렬하여 시간 경과에 따른 판매량 변화를 확인할 수 있습니다.
3. `dplyr` 패키지를 사용하여 하나 이상의 열을 기준으로 데이터 집합을 그룹화하고 각 그룹에 대한 요약 통계를 계산합니다. 예를 들어, 고객 주문 데이터 집합을 지역별로 그룹화하여 각 지역의 평균 주문 규모를 계산할 수 있습니다.
4. `dplyr` 패키지를 사용하여 두 개 이상의 데이터 집합을 함께 조인합니다. 예를 들어, 고객 주문 데이터 집합을 고객 인구 통계 데이터 집합과 조인하여 인구 통계와 주문 크기 간에 상관관계가 있는지 확인합니다.

. **데이터 집합을 와이드 포맷에서 롱 포맷으로 또는 롱 포맷에서 와이드 포맷으로 재구성하려면**tidyr‘ 패키지를 사용합니다. 예를 들어, 월별 매출 데이터의 데이터 집합을 와이드 형식(월별 열 하나)에서 롱 형식(월별 행 하나)으로 재구성할 수 있습니다.

1.5.2 프로젝트

1. `dplyr` 패키지를 사용하여 `MovieLens` 데이터 집합의 영화 평점 데이터 집합을 분석합니다. 필터링, 정렬, 그룹화, 요약 기능을 사용하여 “역대 최고 평점을 받은 영화는 무엇인가?”, “장르별로 평점이 어떻게 다른가?” 등의 질문에 답할 수 있습니다.
2. `dplyr` 패키지를 사용하여 비행 데이터의 데이터 집합을 분석합니다. 필터링, 정렬, 그룹화, 요약 기능을 사용하여 “미국에서 가장 혼잡한 공항은 어디인가?”, “항공편 지연 시간은 항공사별로 어떻게 다른가?” 등의 질문에 답할 수 있습니다.
3. `dplyr` 패키지를 사용하여 소셜 미디어 게시물의 데이터 집합을 분석합니다. 필터링, 정렬, 그룹화 및 요약 기능을 사용하여 “참여율이 가장 높은 소셜 미디어 플랫폼은 무엇인가요?”, “소셜 미디어에서 가장 많이 논의되는 주제는 무엇인가요?” 등의 질문에 답해 보세요.

이러한 연습과 프로젝트를 연습함으로써 데이터 랭글링 기술을 사용하여 R에서 실제 데이터 집합을 분석하는 귀중한 경험을 쌓을 수 있습니다.

2 시각화와 EDA

데이터 시각화는 일반적으로 차트, 그래프 및 기타 시각적 요소를 사용하여 데이터를 그래픽으로 표현하는 프로세스입니다. 데이터 시각화는 단순한 숫자, 표만으로는 이해하기 어려운 데이터의 복잡한 패턴과 관계를 탐색하고 전달할 수 있게 해주기 때문에 데이터 과학에서 중요한 부분입니다.

2.1 데이터 시각화와 중요성

데이터 과학에서 데이터 시각화가 중요한 이유는 여러 가지가 있습니다:

- 데이터 탐색: 데이터 시각화를 사용하면 데이터를 시각적으로 탐색하고 숫자 표에서 바로 알 수 없는 패턴, 추세 및 이상값을 신속하게 식별할 수 있습니다.
- 인사이트 전달: 데이터 시각화를 통해 인사이트와 발견 사항을 명확하고 간결한 방식으로 다른 사람에게 전달할 수 있습니다. 데이터를 시각적으로 제시함으로써 더 많은 사람들이 복잡한 정보에 더 쉽게 접근하고 이해할 수 있도록 만들 수 있습니다.
- 모델 검증: 데이터 시각화는 변수 간의 관계를 시각화하고 잠재적인 관심 영역을 식별하여 모델과 가설을 검증하는 데 사용할 수 있습니다.
- 의사 결정 안내: 데이터 시각화는 원시 데이터에서 즉시 드러나지 않을 수 있는 추세와 패턴을 식별하는 데 도움이 되는 방식으로 데이터를 제시하여 의사 결정을 안내하는 데 사용할 수 있습니다.

2.1.1 데이터 시각화 유형

데이터를 표현하는 데 사용할 수 있는 데이터 시각화에는 다음과 같은 다양한 유형이 있습니다:

- 막대 차트: 카테고리 간 값을 비교하는 데 사용됩니다.
- 꺾은선형 차트: 시간 경과에 따른 추세를 표시하거나 여러 변수의 변화를 비교하는 데 사용됩니다.
- 산점도 차트: 두 변수 간의 관계를 시각화하는 데 사용됩니다.
- 히트맵: 2차원 공간에서 데이터의 밀도를 시각화하는 데 사용됩니다.
- 나무트리 맵: 계층형 데이터를 시각화하는 데 사용됩니다.
- 네트워크 다이어그램: 엔티티 간의 관계를 시각화하는 데 사용됩니다.

2.1.2 효과적인 시각화 원칙

효과적인 데이터 시각화를 만들려면 몇 가지 기본 원칙을 따르는 것이 중요합니다:

- 단순성: 불필요한 군더더기를 피하여 시각화를 단순하고 읽기 쉽게 유지합니다.
- 정확성: 시각화가 왜곡이나 잘못된 표현을 피하면서 데이터를 정확하게 표현하는지 확인합니다.
- 관련성: 데이터와 전달하려는 메시지에 적합한 시각화 유형을 선택합니다.
- 일관성: 여러 시각화에서 일관된 색 구성표, 글꼴 및 기타 시각적 요소를 사용합니다.
- 상호 작용: 사용자가 데이터를 더 자세히 탐색할 수 있도록 상호 작용을 제공하세요.

이러한 원칙을 따르고 적절한 데이터 시각화를 사용하면 복잡한 데이터를 이해하고 전달하는 데 도움이 되는 명확하고 효과적인 시각화를 만들 수 있습니다.

2.2 시각화 도구 - ggplot2

R은 데이터 시각화를 만들기 위한 다양한 패키지를 제공하지만, 가장 인기 있고 강력한 두 가지 패키지는 `ggplot2`입니다. 이 패키지를 사용하면 산점도 차트, 꺾은선형 차트, 막대 차트 등 다양한 고품질의 사용자 지정 가능한 시각화를 만들 수 있습니다.

2.2.1 ggplot2

`ggplot2`는 그래픽 문법을 사용하여 우아하고 사용자 지정 가능한 데이터 시각화를 만들기 위한 패키지입니다. 모든 시각화는 기하학적 도형, 눈금, 좌표계와 같은 선언적 구성 요소 집합으로 분해할 수 있다는 아이디어를 기반으로 합니다. 이러한 빌딩 블록을 다양한 방식으로 결합하여 사용자는 다양하고 복잡한 시각화를 만들 수 있습니다.

ggplot2의 몇 가지 기능은 다음과 같습니다:

- 레이어링(Layering): 점, 선, 레이블 등 다양한 요소를 레이어링하여 복잡한 시각화를 만들 수 있습니다.
- 테마(Themes): 글꼴, 색상 및 배경을 변경하여 시각화의 모양을 사용자 지정할 수 있는 기능입니다.
- 패싯(Faceting): 데이터의 서로 다른 하위 집합을 기반으로 여러 개의 작은 시각화를 만드는 기능입니다.

2.2.2 올바른 패키지 선택

ggplot2 및 다양한 시각화 패키지는 모두 R에서 데이터 시각화를 만들기 위한 강력한 패키지입니다. 올바른 패키지를 선택하는 것은 프로젝트의 특정 요구 사항에 따라 달라집니다. 고도의 사용자 지정이 필요한 복잡한 시각화를 만드는 데 관심이 있는 경우 `ggplot2`가 더 나은 선택일 수 있습니다.

어떤 패키지를 선택하든 효과적인 데이터 시각화의 원칙을 잘 이해하고 명확하고 효과적인 시각화를 만들기 위한 모범 사례를 따르는 것이 중요합니다.

2.2.3 EDA 원칙

탐색적 데이터 분석(EDA)은 데이터셋을 분석하고 이해하여 주요 특성을 요약하는 과정으로, 주로 시각적 방법을 사용합니다. EDA는 데이터 과학자가 데이터 내부의 패턴과 관계를 더 잘 이해하고 이상값이나 결측값과 같은 잠재적인 문제를 식별하는 데 도움이 됩니다.

2.2.4 데이터 분포

데이터 분포는 데이터셋의 값이 어떻게 펴져있는지를 나타냅니다. 데이터 분포를 이해하는 것은 데이터를 분석하는 데 사용되는 통계 방법의 선택에 영향을 미칠 수 있으므로 중요합니다. 데이터 분포를 시각화하는 데 가장 일반적으로 사용되는 방법에는 히스토그램, 밀도 플롯, 상자그림 플롯이 있습니다.

히스토그램은 데이터를 구간으로 나누고 각 구간에 있는 데이터 수를 세어 연속 데이터의 분포를 시각화하는 방법을 제공합니다. 밀도 플롯은 비슷한 시각화를 제공하지만 데이터의 확률 밀도 함수를 추정합니다. 상자그림 플롯은 데이터의 중앙값, 사분위수 및 이상값을 표시하여 분포를 시각화합니다.

2.2.5 상관 관계

상관관계는 두 변수 간의 관계의 강도와 방향을 측정하는 척도입니다. 상관관계 분석은 데이터 내의 패턴과 관계를 식별하는 데 유용합니다. 상관 관계를 시각화하는 데 가장 일반적으로 사용되는 방법은 산점도입니다.

산점도에서 각 점은 두 변수의 값 쌍을 나타냅니다. 점을 2차원 평면에 그려서 두 변수 간의 관계를 시각적으로 검사할 수 있습니다. 양의 상관관계는 한 변수가 증가하면 다른 변수도 증가한다는 것을 의미하며, 음의 상관관계는 한 변수가 증가하면 다른 변수가 감소한다는 것을 의미합니다.

2.2.6 이상값 감지

이상값은 데이터셋의 다른 점들과 크게 다른 데이터 점입니다. 이상값은 통계 분석에 큰 영향을 미칠 수 있으므로 이상값을 식별하고 제거할지 여부를 결정하는 것이 중요합니다. 이상값을 시각화하는 데 가장 일반적으로 사용되는 방법은 상자그림 플롯입니다.

상자그림 플롯에서 이상값은 플롯의 수염을 벗어난 데이터 포인트로 식별됩니다. 이상값을 감지하는 또 다른 방법은 Z-점수 검정 또는 투키(Tukey) 방법과 같은 통계적 테스트를 사용하는 것입니다.

2.2.7 결론

탐색적 데이터 분석은 모든 데이터 과학 프로젝트에서 중요한 단계입니다. 데이터 과학자는 데이터 분포, 변수 간의 상관관계, 이상값의 존재를 이해함으로써 데이터에 대한 인사이트를 얻고 데이터 분석 방법에 대해 정보에 입각한 결정을 내릴 수 있습니다. 효과적인 EDA를 위해서는 시각화 및 통계적 방법의 조합과 데이터 분석의 기본 원칙에 대한 충분한 이해가 필요합니다.

2.3 데이터 탐색 기법

데이터 탐색은 데이터셋의 주요 특성을 분석하고 시각화하여 그 기본 패턴과 관계를 이해하는 프로세스입니다. 다음은 데이터 과학에서 데이터 탐색에 가장 일반적으로 사용되는 몇 가지 기술입니다.

2.3.1 산점도

산점도(Scatter Plot)는 두 변수 간의 관계를 탐색하는 데 유용한 기법입니다. 산점도에서 각 점은 비교 대상인 두 변수의 값 쌍을 나타냅니다. X축은 한 변수를 나타내고 Y축은 다른 변수를 나타냅니다. 2차원 평면에 점을 그리면 두 변수 간의 관계를 시각적으로 확인할 수 있습니다.

예를 들어 한 사람의 나이와 소득 간의 관계를 살펴보고 싶다고 가정해 보겠습니다. 나이를 X축으로 하고 소득을 Y축으로 하는 산점도를 만들 수 있습니다. 산점도를 살펴보면 나이와 소득 사이에 양의 상관관계가 있는지 또는 음의 상관관계가 있는지 확인할 수 있습니다.

2.3.2 히스토그램

히스토그램은 단일 변수의 분포를 시각화하는 방법을 제공합니다. 히스토그램에서 데이터는 구간 또는 구간차원으로 나뉘며, 각 구간차원에 있는 데이터 요소의 수를 계산합니다. 결과 시각화는 각 구간차원 내의 데이터 요소의 빈도를 보여줍니다.

예를 들어, 데이터셋의 연령 분포를 탐색하고 싶다고 가정해 보겠습니다. X축에 연령을, Y축에 연령 값의 빈도를 포함하는 히스토그램을 만들 수 있습니다. 히스토그램을 검토하여 연령 값이 왼쪽이나 오른쪽으로 치우쳐 있는지 또는 고르게 분포되어 있는지 확인할 수 있습니다.

2.3.3 상자그림

상자그림(Boxplot)은 단일 변수의 분포를 시각화하고 이상값을 식별할 수 있는 방법을 제공합니다. 상자그림에서는 데이터를 사 분위수로 나누고 중앙값, 1사분위수 및 3사분위수, 특정 범위 내의 최소값과 최대값을 표시하는 상자를 그립니다. 이상값은 상자그림의 수염을 벗어난 데이터 포인트로 식별됩니다.

예를 들어 데이터셋의 소득 분포를 탐색하고 싶다고 가정해 보겠습니다. Y축에 소득 값이 있는 상자그림을 만들 수 있습니다. 상자그림을 검토하여 소득 값에 이상값이 있는지 확인할 수 있습니다.

2.3.4 결론

데이터 탐색은 데이터 과학 프로세스에서 중요한 단계입니다. 데이터 과학자는 산점도, 히스토그램, 상자그림과 같은 기법을 사용하여 데이터셋 내의 패턴과 관계에 대한 인사이트를 얻을 수 있습니다. 이러한 기법은 이상값을 식별하고, 변수 분포를 탐색하고, 데이터의 잠재적인 문제를 식별하는 데 사용할 수 있습니다. 궁극적으로 효과적인 데이터 탐색은 데이터를 분석하고 모델링하는 방법에 대해 정보에 입각한 결정을 내리는 데 핵심적인 역할을 합니다.

2.4 연습문제와 프로젝트

연습과 프로젝트는 모든 학습 과정에서 매우 중요한 부분이며, 데이터 시각화 및 탐색적 데이터 분석(EDA)도 예외는 아닙니다. 이 섹션에서는 R을 사용하여 데이터 시각화 및 EDA 기술을 연습하는 데 도움이 되는 몇 가지 연습과 프로젝트를 소개합니다.

2.4.1 연습문제

- `mtcars` 데이터셋을 가져와서 마력(`hp`)에 대한 갤런당 마일(`mpg`)의 산점도를 만듭니다. 실린더(실린더) 수에 따라 다른 색상을 사용합니다.
- 붓꽃(`iris`) 데이터셋을 가져와서 각 종(`Species`)에 대한 꽃받침 길이(`Sepal.Length`)의 히스토그램을 만듭니다. 종마다 다른 색상을 사용합니다.
- `ggplot2` 패키지에서 다이아몬드(`diamonds`) 데이터셋을 가져와서 각 컷(`cut`)에 대한 가격(`price`)의 상자그림을 만듭니다. 컷마다 다른 색상을 사용합니다.
- `mtcars` 데이터셋을 가져와서, 무게(`wt`)에 대한 갤런당 마일(`mpg`)의 산점도를 만듭니다. 플롯에 회귀선을 추가합니다.
- `gapminder` 패키지에서 `gapminder` 데이터을 가져와서, 2007년에 대한 기대 수명(`lifeExp`)과 1인당 GDP(`gdpPercap`)의 산점도를 만듭니다. 대륙(`continent`)에 따라 다른 색상을 사용합니다.

2.4.2 프로젝트

- `ggplot2` 패키지를 사용하여 데이터 시각화 프로젝트를 만듭니다. 관심 있는 데이터셋을 선택하고 데이터의 다양한 측면을 탐색하는 일련의 플롯을 만듭니다. 프로젝트에는 산점도, 히스토그램, 상자그림이 하나 이상 포함되어야 합니다.
- 시계열 데이터(예: 주가 또는 날씨 데이터)가 포함된 데이터셋을 선택하고 시간에 따른 데이터의 다양한 측면을 탐색하는 일련의 플롯을 만듭니다. 프로젝트에는 꺾은선형 차트, 산점도, 히스토그램이 하나 이상 포함되어야 합니다.
- `tidyverse` 패키지를 사용해 EDA 프로젝트를 만듭니다. 관심 있는 데이터셋을 선택하고 다양한 EDA 기법을 사용하여 데이터의 다양한 측면을 탐색합니다. 프로젝트에는 데이터 분포 시각화, 여러 변수 간의 관계 시각화, 이상값 시각화 중 하나 이상이 포함되어야 합니다.
- 공간 데이터(예: 인구 밀도 맵 또는 지진 위치 맵)가 포함된 데이터셋을 선택하고 데이터의 다양한 측면을 탐색하는 일련의 플롯을 만듭니다. 프로젝트에는 적어도 하나의 맵, 하나의 분산형 차트, 하나의 히스토그램이 포함되어야 합니다.
- 네트워크 데이터가 포함된 데이터셋(예: 소셜 네트워크 또는 교통 네트워크)을 선택하고 데이터의 다양한 측면을 탐색하는 일련의 플롯을 만듭니다. 프로젝트에는 네트워크 다이어그램, 분산형 차트, 히스토그램이 하나 이상 포함되어야 합니다.

이러한 연습과 프로젝트를 연습하면 R의 데이터 시각화 및 EDA 기술에 더 익숙해지고, 이러한 기술을 사용하여 데이터에서 인사이트를 얻는 방법을 더 잘 이해할 수 있습니다.

3 통계분석과 기계학습 기초

3.1 들어가며

데이터 분석에는 데이터를 탐색하고 시각화하는 것뿐만 아니라 데이터로부터 예측을 하고 결론을 도출하는 것도 포함됩니다. 통계 분석과 기계 학습은 이 작업에 도움이 될 수 있는 두 가지 중요한 기술입니다.

통계 분석은 수학적 모델을 사용하여 데이터를 분석하고 해석하는 것입니다. 여기에는 가설을 검정하고, 패턴과 추세를 파악하고, 예측을 하는 것이 포함됩니다. R에는 통계 분석에 사용할 수 있는 `stats`, `MASS`, `infer` 등 많은 패키지가 있습니다.

반면에 기계학습(Machine Learning)은 알고리즘을 사용하여 패턴을 식별하고 데이터를 기반으로 예측을 하는 인공지능의 한 유형입니다. 기계학습 기술에는 지도학습과 비지도학습으로 나눌 수 있고 지도학습은 분류와 회귀로 더 나눌 수 있습니다.. R에는 `tidymodels`로 통일된 기계학습 프레임워크가 있으며 지도학습과 비지도학습 모형을 개발할 때 최근 많이 사용합니다.

통계 분석과 기계 학습은 모두 데이터 과학자에게 중요한 도구입니다. 데이터에서 숨겨진 패턴을 발견하고 미래의 사건을 예측하는 데 도움이 됩니다. 이러한 기술을 데이터 시각화 및 탐색적 데이터 분석과 결합하면 작업 중인 데이터를 더 깊이 이해하고 더 많은 정보에 입각한 의사 결정을 내릴 수 있습니다.

다음 섹션에서는 R에서 사용되는 몇 가지 일반적인 통계 분석 및 기계 학습 기법을 살펴보고 이러한 기법을 실제로 어떻게 사용할 수 있는지에 대한 예를 제공합니다.

3.2 기본 통계 개념

통계학은 데이터의 수집, 분석, 해석, 표현 및 조직을 다루는 수학의 한 분야입니다. 데이터 과학자가 작업 중인 데이터를 이해하려면 통계 개념을 이해하는 것이 필수적입니다.

3.2.1 확률 분포

확률 분포는 무작위 이벤트에서 다양한 결과가 발생할 가능성을 설명합니다. 정규 분포, 포아송 분포, 이항 분포를 포함하여 R에서 사용할 수 있는 확률 분포는 다양합니다. 확률 분포를 이해하면 데이터 과학자가 데이터에 대해 정보에 입각한 의사 결정을 내리고 미래 사건을 예측하기 위한 정확한 모형을 만드는 데 도움이 될 수 있습니다.

3.2.2 가설 검정

가설검정은 해당 모집단의 표본을 기반으로 모집단에 대한 가설을 검정하는 통계 기법입니다. 가설이 참일 가능성이 있는지 거짓일 가능성이 있는지 판단하는 데 사용됩니다. 일반적인 가설 검정에는 t-검정, 분산 분석, 카이제곱(χ^2) 검정 등이 있습니다. 가설검정은 데이터 과학자가 결과가 통계적으로 유의미한지 여부를 판단하고 데이터로부터 결론을 도출하는 데 중요한 도구입니다.

3.2.3 회귀 분석

회귀 분석은 종속 변수와 하나 이상의 독립 변수 간의 관계를 분석하는 데 사용되는 통계 기법입니다. 선형 회귀가 가장 일반적인 회귀 분석 유형이지만, 로지스틱 회귀 및 다항식 회귀와 같은 다른 유형도 많이 있습니다. 회귀 분석은 데이터 과학에서 과거 데이터를 기반으로 미래의 이벤트를 예측하는 데 사용됩니다.

데이터 과학자가 데이터에 대해 정보에 입각한 의사 결정을 내리려면 이러한 기본 통계 개념을 이해하는 것이 필수적입니다. 다음 섹션에서는 이러한 개념이 R을 사용하여 실제로 어떻게 사용될 수 있는지에 대한 예를 제공합니다.

3.3 기계학습 알고리즘 소개

기계학습은 명시적으로 프로그래밍하지 않고도 데이터를 학습하여 예측이나 결정을 내릴 수 있는 알고리즘을 개발하는 인공 지능의 하위 분야입니다. 머신 러닝은 이미지 및 음성 인식, 사기 탐지, 추천 시스템 등 데이터 과학 분야에서 다양하게 활용되고 있습니다.

3.3.1 지도 학습

지도 학습은 알고리즘이 레이블이 지정된 데이터로부터 학습하는 머신 러닝의 한 유형입니다. 즉, 알고리즘에 입력 데이터와 해당 출력 데이터가 주어지면 알고리즘은 입력을 출력에 매핑하는 방법을 학습합니다. 지도 학습의 일반적인 예로는 회귀 분석, 의사 결정 나무모형, 신경망 등이 있습니다. 지도 학습은 데이터 과학에서 과거 데이터를 기반으로 미래의 이벤트를 예측하는 데 사용됩니다.

3.3.2 비지도 학습

비지도 학습은 레이블이 지정되지 않은 데이터에서 알고리즘이 학습하는 머신 러닝의 한 유형입니다. 즉, 알고리즘에 해당 출력 데이터 없이 입력 데이터가 주어지면 알고리즘이 데이터에서 패턴과 관계를 찾는 방법을 학습합니다. 비지도 학습의 일반적인 예로는 군집분석과 연관 규칙 마이닝이 있습니다. 비지도 학습은 데이터 과학에서 데이터의 숨겨진 패턴과 관계를 발견하는 데 사용됩니다.

3.3.3 데이터 과학 응용 분야

기계학습은 데이터 과학에서 다양한 용도로 사용됩니다. 예를 들어, 신용카드 거래에서 사기 탐지에 사용할 수 있으며, 알고리즘은 과거 거래에서 학습하여 사기 행위 패턴을 식별합니다. 또한 알고리즘이 레이블이 지정된 이미지의 대규모 데이터셋를 학습하여 새로운 이미지에서 객체를 식별하는 이미지 인식에도 사용할 수 있습니다. 또 다른 일반적인 애플리케이션은 추천 시스템으로, 알고리즘이 사용자 행동을 학습하여 개인화된 추천을 제공하는 것입니다.

데이터 과학자가 실제 문제에 효과적으로 적용하려면 이러한 머신 러닝 개념을 이해하는 것이 필수적입니다. 다음 섹션에서는 이러한 알고리즘을 R을 사용하여 실제로 어떻게 사용할 수 있는지에 대한 예를 제공합니다.

3.4 tidymodels 구현

기계학습 모형을 구축하는 과정에는 데이터 준비, 모델 훈련, 유효성 검사 및 테스트와 같은 여러 단계가 포함됩니다. R에는 데이터 과학자가 이러한 단계를 쉽게 수행할 수 있도록 도와주는 여러 패키지가 있습니다. 머신 러닝 모델을 구현하는 데 널리 사용되는 패키지 중 하나는 tidymodels입니다.

tidymodels는 tidyverse 패러다임을 차용하여 통계모형과 기계 학습을 위해 설계된 오픈 소스 R 패키지 모음입니다. 전처리, 표집, 피쳐공학(Feature Engineering), 모델 튜닝 및 평가를 포함하는 모델링을 위한 일관된 프레임워크를 제공합니다. tidymodels에 포함된 중요한 패키지는 다음과 같습니다:

- tidyverse: 데이터 랭글링, 탐색 및 시각화를 위한 R 패키지 모음입니다.
- dplyr: 데이터 조작 및 변환을 위한 패키지.
- tidyr: 데이터 정리 및 재형성을 위한 패키지.
- ggplot2: 시각화를 만들기 위한 패키지입니다.
- tidymodels: 머신 러닝 모델 구축 및 평가를 위한 패키지입니다.
- rsample: 데이터 분할 및 리샘플링을 위한 패키지.
- parsnip: 모델 사양 및 튜닝을 위한 패키지.

tidymodels 프레임워크는 일관된 모델링 파이프라인을 따르며, 여기에는 다음 단계가 포함됩니다:

- 데이터 준비: 이 단계에서는 결측값 삽입, 표준정규화 및 인코딩과 같은 다양한 기술을 사용하여 데이터를 로드하고 전처리합니다.
- 피처 공학(Feature Engineering): 이 단계에서는 기존 피쳐에서 새 피쳐를 만들고, 중요한 피쳐를 선택하고, 모델의 요구 사항을 충족하도록 피쳐를 변환합니다.
- 모델 사양: 이 단계에서는 머신 러닝 모델을 선택하고 정의합니다. parsnip 패키지는 모델 지정을 위한 일관된 인터페이스를 제공합니다.
- 모델 튜닝: 이 단계에서는 성능을 최적화하기 위해 모델의 하이퍼파라미터를 튜닝합니다. tune 패키지는 모델 튜닝을 위한 다양한 방법을 제공합니다.
- 모델 평가: 이 단계에서는 정확도, 정밀도, 리콜 등 다양한 측도를 사용하여 모델의 성능을 평가합니다. yardstick 패키지는 모델 평가를 위한 다양한 지표를 제공합니다.

tidymodels 프레임워크는 R에서 머신러닝 모델을 구축하는 간단하고 직관적인 방법을 제공하며, 다양한 패키지의 도움으로 데이터 준비, 기능 엔지니어링, 모델 사양, 튜닝 및 평가 프로세스를 간소화합니다. 이 프레임워크는 분류 및 회귀부터 군집분석 및 연관 규칙 마이닝에 이르기까지 다양한 유형의 데이터 과학 문제를 해결하는 데이터 과학자에게 유용할 수 있습니다.

3.5 연습문제와 프로젝트

연습과 프로젝트는 데이터 과학 학습의 중요한 부분입니다. 학습한 내용을 실제로 적용하고 실무 경험을 쌓는 데 도움이 됩니다. 이 섹션에서는 R에서 통계 분석 및 기계 학습 기초를 연습하는 데 도움이 되는 몇 가지 연습과 프로젝트에 대해 설명합니다.

3.5.1 연습문제

- 확률 분포: R을 사용하여 정규 분포, 포아송 분포, 이항 분포의 확률 분포 함수를 시뮬레이션하고 시각화합니다. 다양한 매개변수 값으로 실험하고 분포가 어떻게 변화하는지 관찰하세요.
- 가설 검정: R을 사용하여 선택한 데이터 집합에 대해 가설 테스트를 수행합니다. Kaggle 또는 UCI 머신 러닝 리포지토리와 같은 리포지토리에서 데이터셋을 선택할 수 있습니다. 가설을 세우고 적절한 통계검정을 사용하여 가설을 검정합니다. 적절한 플롯과 그래프를 사용하여 결과를 시각화합니다.
- 회귀 분석: R을 사용하여 선택한 데이터셋에 대해 간단한 다중 선형 회귀 분석을 수행합니다. 적절한 응답 변수와 설명 변수를 선택합니다. 회귀 모델을 맞추고, 성능을 평가하고, 결과를 해석합니다.
- 분류 알고리즘: R 및 `tidymodels` 패키지를 사용하여 선택한 데이터셋에서 분류 모델을 구축하고 평가합니다. 로지스틱 회귀, 의사 결정 나무모형 또는 랜덤 포레스트와 같은 적절한 분류 알고리즘을 선택합니다. 정확도, 정밀도, 리콜과 같은 적절한 측도를 사용하여 모델의 성능을 평가합니다.

3.5.2 프로젝트

- 탐색적 데이터 분석: Kaggle 또는 UCI 머신 러닝 리포지토리와 같은 리포지토리에서 데이터셋을 선택합니다. 적절한 플롯과 그래프를 사용하여 데이터셋에 대한 탐색적 데이터 분석을 수행합니다. 데이터에서 패턴, 추세 또는 이상값을 식별합니다. 결과를 요약한 보고서를 작성합니다.
- 예측 모델링: Kaggle 또는 UCI 머신 러닝 리포지토리와 같은 리포지토리에서 데이터셋을 선택합니다. 로지스틱 회귀, 의사 결정 나무모형 또는 랜덤 포레스트와 같은 적절한 머신 러닝 알고리즘을 사용하여 데이터셋에 대한 예측 모델을 구축하고 평가합니다. 적절한 성능 측도를 사용하여 모델의 성능을 평가합니다. 결과를 요약한 보고서를 작성합니다.
- 시계열 분석: 주가나 날씨 데이터와 같은 시계열 데이터가 포함된 데이터셋을 선택합니다. 이동 평균, 지수 평활화 또는 ARIMA 모델과 같은 적절한 기술을 사용하여 데이터셋에 대한 시계열 분석을 수행합니다. 적절한 성능 측도를 사용하여 모델의 성능을 평가합니다. 결과를 요약한 보고서를 작성합니다.

결론적으로, 이 섹션에서 설명한 연습과 프로젝트는 R에서 통계 분석 및 기계 학습 기본 사항을 연습하는 데 도움이 될 수 있으며, 실무 경험을 쌓고 데이터 과학 기술을 개발할 수 있는 귀중한 방법입니다.

4 고급 데이터 과학 기법

4.1 들어가며

R은 데이터 랭글링, 데이터 시각화, 통계 분석의 기본 개념 외에도 고급 데이터 과학 기법을 위한 다양한 패키지와 함수를 제공합니다. 이러한 기법은 텍스트, 소셜 네트워크, 시계열 데이터 등 다양한 유형의 데이터에서 인사이트를 추출하는 데 사용할 수 있습니다. 이 섹션에서는 R에서 텍스트 마이닝, 네트워크 분석 및 시계열 분석을 위한 몇 가지 인기 있는 패키지와 기법을 소개합니다.

4.1.1 텍스트 마이닝

텍스트 마이닝은 텍스트 데이터를 분석하여 패턴, 트렌드 및 인사이트를 발견하는 프로세스입니다. 자연어 처리, 감정 분석, 주제 모델링과 같은 분야에서 널리 사용됩니다. R에는 텍스트 마이닝을 위한 `tm`, `tidytext`, `quanteda`와 같은 여러 패키지가 있습니다. 이러한 패키지는 텍스트 전처리, 용어 빈도 분석, 문서 클러스터링과 같은 작업을 위한 기능을 제공합니다.

4.1.2 네트워크 분석

네트워크 분석은 사람, 조직 또는 웹 페이지와 같은 개체 간의 관계를 분석하는 프로세스입니다. 소셜 네트워크 분석, 추천 시스템, 그래프 이론과 같은 분야에서 널리 사용됩니다. R에는 `tidygraph`, `igraph`, `network`, `ggraph` 등 네트워크 분석을 위한 여러 패키지가 있습니다. 이러한 패키지는 네트워크 시각화, 커뮤니티 감지, 중심성 분석과 같은 작업을 위한 기능을 제공합니다.

4.1.3 시계열 분석

시계열 분석은 시계열 데이터를 분석하여 패턴, 추세, 계절성 등을 추출하는 프로세스입니다. 금융, 경제, 엔지니어링 등의 분야에서 널리 사용됩니다. R에는 `tidyverts`, `xts`, `forecast` 등 시계열 분석을 위한 여러 패키지가 있습니다. 이러한 패키지는 시계열 시각화, 분해, 예측 등의 작업을 위한 함수를 제공합니다.

이러한 고급 데이터 과학 기법을 숙달하면 복잡한 비정형 데이터에서 더 많은 인사이트를 추출할 수 있습니다. 다음 섹션에서는 텍스트 마이닝을 위한 몇 가지 인기 있는 R 패키지와 함수를 소개합니다.

4.2 텍스트 마이닝

감정 분석, 텍스트 분류 및 주제 모델링과 같은 텍스트 마이닝 개념 개요는 R의 고급 데이터 과학 기술의 중요한 측면입니다. 텍스트 마이닝에는 구조화되지 않은 텍스트 데이터에서 의미 있는 인사이트와 지식을 추출하는 작업이 포함됩니다.

텍스트 마이닝에 사용되는 인기 있는 R 패키지 중 하나는 `tidytext`입니다. 이 패키지는 정리, 어간 제거, 중단어 제거와 같은 텍스트 데이터 전처리를 위한 함수 집합을 제공합니다. `tidytext` 패키지는 텍스트 분석에 사용되는 용어-문서 행렬(TDM)을 비롯한 다양한 자료구조를 생성하는 함수도 제공합니다.

감성 분석은 텍스트의 감정 어조를 식별하는 데 사용되는 기법입니다. 텍스트의 각 단어에 감정 점수를 할당하는 감정 분석에는 R의 `tidytext` 패키지가 사용됩니다. 점수는 텍스트에 포함된 단어의 문맥에 따라 긍정, 부정 또는 중립이 될 수 있습니다.

텍스트 분류는 텍스트 마이닝의 또 다른 중요한 기술입니다. 텍스트 데이터를 미리 정의된 범주 또는 클래스로 분류하는 것이 포함됩니다. 텍스트 분류에는 R의 `tidymodels` 패키지가 사용됩니다. 텍스트 분류를 위해 의사 결정 나무모형, 서포트 벡터 머신(SVM), 나이브 베이즈 등 다양한 기계학습 모델을 훈련하고 평가하는 기능을 제공합니다.

토픽 모델링은 대규모 문서 모음에서 주제를 발굴하고 주제를 식별하는 데 사용되는 기법입니다. 문서 내용을 가장 잘 나타내는 주제 집합을 만드는 주제 모델링에는 R의 `topicmodels` 패키지가 사용됩니다. 각 토픽은 문서에서 가장 일반적으로 함께 발견되는 단어의 조합입니다.

전반적으로 텍스트 마이닝은 구조화되지 않은 텍스트 데이터에서 인사이트와 지식을 추출하는 데 유용한 기술입니다. `tidytext`, `tidymodels` 및 `topicmodels`와 같은 R 패키지를 사용하면 텍스트 마이닝 프로세스를 크게 간소화하고 모든 수준의 데이터 과학자가 액세스할 수 있습니다.

4.3 네트워크 분석

네트워크 분석은 개체 또는 엔티티 간의 관계와 상호 작용, 그리고 이러한 관계를 수학적 및 계산 도구를 사용하여 표현하고 분석할 수 있는 방법을 연구하는 학문입니다. 데이터 과학에서 네트워크 분석은 소셜 미디어, 온라인 커뮤니티 및 기타 네트워크 기반 현상의 증가로 인해 점점 더 중요해지고 있습니다.

R은 네트워크 생성, 분석 및 시각화를 위한 다양한 기능을 제공하는 `tidygraph`, `igraph` 및 `statnet`과 같은 네트워크 분석을 위한 여러 패키지를 제공합니다. 네트워크 분석의 주요 개념에는 중심성 측정, 커뮤니티 감지, 네트워크 시각화 등이 있습니다.

중심성 측정은 네트워크 내에서 위치나 역할에 따라 네트워크에서 가장 중요한 노드를 식별하는 데 사용됩니다. 몇 가지 일반적인 중심성 측정에는 차수 중심성, 사이 중심성, 고유 벡터 중심성 등이 있습니다.

커뮤니티 탐지는 네트워크 내에서 나머지 네트워크보다 서로 더 강하게 연결된 노드 그룹을 식별하는 프로세스입니다. 이는 더 큰 네트워크 내에서 하위 그룹을 식별하거나 연결 패턴을 기반으로 유사한 노드 클러스터를 식별하는 데 유용할 수 있습니다.

`tidygraph`, `igraph` 및 `statnet`과 같은 R 패키지는 네트워크에서 중심성 및 커뮤니티 감지 분석을 수행하기 위한 다양한 기능을 제공합니다. 이러한 패키지는 강제 방향 레이아웃(force-directed layouts)이나 원형 레이아웃(circular layouts)과 같은 다양한 레이아웃과 스타일을 사용하여 네트워크를 시각화하는 기능도 제공합니다.

소셜 네트워크 분석 외에도 감정 분석, 텍스트 분류, 토픽 모델링과 같은 텍스트 마이닝 작업에는 `tidytext`, `tidymodels` 및 `topicmodels`과 같은 R 패키지를 사용할 수 있습니다. 이러한 기법은 소셜 미디어 게시물, 뉴스 기사 또는 고객 리뷰와 같은 대량의 텍스트 데이터를 분석하는 데 사용할 수 있습니다.

전반적으로 네트워크 분석과 텍스트 마이닝은 복잡한 시스템과 대량의 비정형 데이터에 대한 인사이트를 얻는 데 사용할 수 있는 데이터 과학의 중요한 기술입니다. R은 이러한 기술을 구현하고 데이터의 기본 구조와 패턴을 탐색하기 위한 강력한 도구를 제공합니다.

4.4 시계열 분석

시계열 분석은 특히 비즈니스 및 경제 분야에서 데이터 과학의 중요한 측면으로, 과거 데이터를 통해 미래의 추세와 패턴에 대한 인사이트를 얻을 수 있습니다. R은 예측 및 시계열 패키지 등 시계열 분석을 위한 다양한 패키지와 함수를 제공합니다.

- Arima 모델: ARIMA(AutoRegressive Integrated Moving Average) 모델은 과거 데이터를 기반으로 미래 값을 예측하기 위한 시계열 분석에 널리 사용됩니다. 이 모델은 자동 회귀(AR) 및 이동 평균(MA) 구성 요소를 고정된 시계열로 만들기 위해 차분된 시계열과 결합합니다. `forecast` 패키지는 AIC(Akaike 정보 기준) 값에 따라 최적의 모델을 자동으로 선택하는 `auto.arima()` 함수를 포함하여 ARIMA 모델을 맞추기 위한 여러 함수를 제공합니다.
- 예측: 예측은 과거 데이터를 사용하여 미래 값을 예측하는 것입니다. 예측 패키지는 ARIMA 모델을 기반으로 예측을 생성하는 `forecast()` 함수를 포함하여 미래 값을 예측하기 위한 여러 함수를 제공합니다. 이 패키지에는 `accuracy()` 함수와 같이 예측의 정확도를 평가하기 위한 함수도 포함되어 있습니다.
- 추세 분석: 추세 분석에는 시계열 데이터에서 추세 또는 패턴을 식별하는 것이 포함됩니다. `tseries` 패키지는 시계열을 추세, 계절 및 무작위 구성 요소로 분해하는 `decompose()` 함수를 포함하여 추세 분석을 위한 여러 함수를 제공합니다. 이 패키지에는 `tsclean()` 함수와 같이 이상값을 감지하고 제거하는 함수도 포함되어 있습니다.

기타 시계열 기법: R은 스펙트럼 분석 및 상태 공간 모델과 같은 다양한 기타 시계열 분석 기법을 제공합니다. 예측 및 `tseries` 패키지에는 이러한 기법을 위한 함수도 포함되어 있습니다.

4.4.1 연습 문제 및 프로젝트

- 예측 패키지를 사용하여 시계열에 ARIMA 모델을 맞추고 미래 값에 대한 예측을 생성합니다.
- `tseries` 패키지를 사용하여 시계열을 추세, 계절 및 무작위 구성 요소로 분해하고 결과를 시각화합니다.
- 예측 패키지를 사용하여 `accuracy()` 함수를 사용하여 다양한 예측 모델의 정확도를 비교합니다.
- 실제 시계열 데이터 집합을 사용하여 추세 분석을 수행하고 `tseries` 패키지를 사용하여 이상값을 식별합니다.

4.5 연습문제와 프로젝트

연습과 프로젝트는 데이터 과학 기술을 배우고 연습하는 데 있어 중요한 부분입니다. 이 섹션에서는 이 책에서 다루는 고급 데이터 과학 기술을 연습하는 데 도움이 되는 몇 가지 연습과 프로젝트를 살펴봅니다.

4.5.1 텍스트 마이닝

tm 및 tidytext 패키지를 사용하여 영화 리뷰의 감성을 분석합니다. 뉴스 기사 데이터 세트에서 나이브 베이즈와 SVM 알고리즘을 사용하여 텍스트 분류를 수행합니다.

4.5.2 네트워크 분석

igraph 패키지를 사용하여 트위터 팔로워 데이터 세트에 대한 소셜 네트워크 분석을 수행합니다. 통계 네트워크 패키지의 중심성 측정과 커뮤니티 감지 알고리즘을 사용해 공동 저술 네트워크의 구조를 분석합니다.

4.5.3 시계열 분석

ARIMA 모델과 예측 패키지를 사용하여 주가 데이터 집합에 대한 예측을 수행합니다. tseries 패키지의 시계열 분석 기법을 사용하여 일일 코로나19 확진자 수의 추세를 분석합니다. 이러한 연습과 프로젝트는 고급 데이터 과학 기술에 대한 이해를 강화하는 데 도움이 될 뿐만 아니라 실제 데이터 집합으로 작업하는 실무 경험을 제공합니다. 또한 이러한 프로젝트를 사용하여 포트폴리오를 구축하고 잠재적 고용주에게 자신의 기술을 보여줄 수도 있습니다.

이러한 프로젝트 외에도 데이터 과학 애호가가 작업할 수 있는 데이터 집합과 도전 과제에 대한 액세스를 제공하는 여러 온라인 리소스 및 플랫폼이 있습니다. 여기에는 Kaggle, DataCamp, UCI 머신 러닝 리포지토리가 포함됩니다. 이러한 챌린지에 참여하면 실제 문제를 해결하는 귀중한 경험을 쌓고 전 세계의 다른 데이터 과학 애호가들과 협업할 수 있습니다.

Part II

프로젝트

참고문헌

Wickham, Hadley, and Garrett Grolemund. 2016. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. " O'Reilly Media, Inc.".