STAT204 데이터 분석 프로젝트

Fatal Road Accident 1989-2021

변수에 따른 교통사고 발생 빈도 분석

중어중문학과 2019131238 정예린

Content

탐구 동기 및 목표

교통사고 사망자 수는 매년 감소하고 있지만, 안전한 도로교통을 위한 시스템 구축을 위해서는 변수별 교통사고 발생 빈도를 정확하게 파악하여 사전에 방지할 대책을 계획해야 한다. 또한 블랙박스, 내비게이션 등 관련 장비 보급률이 증가 추세이며, ADAS(첨단안전 지원체계) 등 교통안전관련 기술 개발에 따라 활용가능한 데이터 원천이 다양화되고 있다. 이러한 상황에서 도로교통 빅데이터를 활용한 기반 조성이 갈수록 중요시된다.

본 탐구에서는 1989년부터 2021년까지의 호주 도로교통사고 데이터를 활용하여 연도별, 월별, 요일별, 시간대별로 교통사고가 빈번하게 발생하는 지점을 파악하여 시간적 변수와 교통사고 발생빈도의 관계를 분석하고자 한다. 또한, 연령대별, 성별에 따른 교통사고 발생 빈도를 파악하여 높은 빈도로 교통사고에 의해 사망하는 인구 집단을 추려내고자 한다. 마지막으로, 이동수단과 속도 제한이라는 변수가 교통사고 발생 빈도와 유의미한 관계를 가지는지 탐구하고자 한다. 이러한 탐구 결과는 안전한 도로교통을 구축하기 위한 시스템 기반으로 적합하게 기능할 것이다.

데이터 탐색 및 전처리

1. Raw 데이터 불러오기 – csv 형식 dataframe

```
> crash <- read.csv(file="C:/data/Crash_Data.csv", header=T)
```

2. 데이터 행과 열 개수 세기 – row 52843개, column 23개

```
> dim(crash)
[1] 52843    23
```

3. 변수명 확인하기 – 분석을 위해 변환이 필요한 변수명 없음

```
> names(crash)
 [1] "Crash.ID"                    "State"                       "Month"
 [4] "Year"                        "Dayweek"                     "Time"
 [7] "Crash.Type"                  "Bus.Involvement"             "Heavy.Rigid.Truck.Involvement"
[10] "Articulated.Truck.Involvement" "Speed.Limit"                 "Road.User"
[13] "Gender"                      "Age"                         "National.Remoteness.Areas"
[16] "SA4.Name.2016"               "National.LGA.Name.2017"      "National.Road.Type"
[19] "Christmas.Period"            "Easter.Period"               "Age.Group"
[22] "Day.of.week"                 "Time.of.day"
```

데이터 탐색 및 전처리

4. dplyr, ggplot2 패키지 불러오기

```
> library(dplyr)
> library(ggplot2)
```

5. 결측치 검토 – NA값 없음

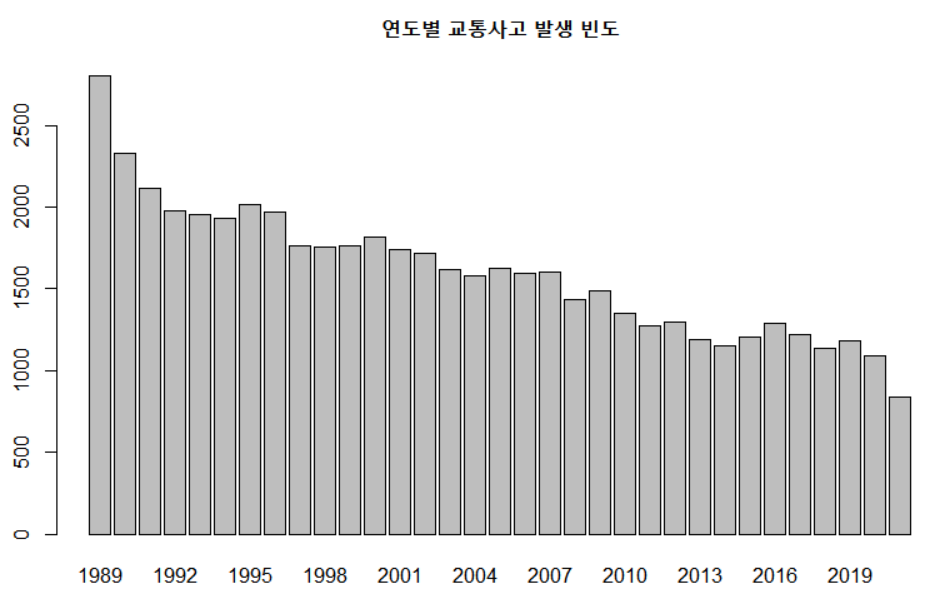
```
> sum(is.na(crash))
[1] 0
```

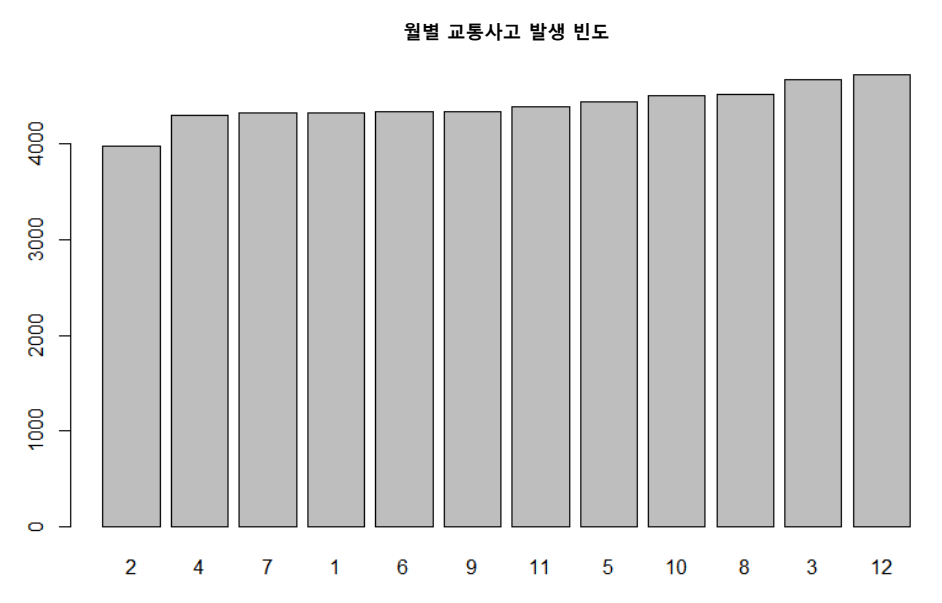
6. 이상치 제거

– 수치형 변수 Age의 table에서 이상치 '-9' 발견, NA로 대체 후 제거

```
> crash$Age <- ifelse(crash$Age<0, NA, crash$Age)
> na.omit(crash$Age)
```

연도별 & 월별 교통사고 발생 빈도 분석

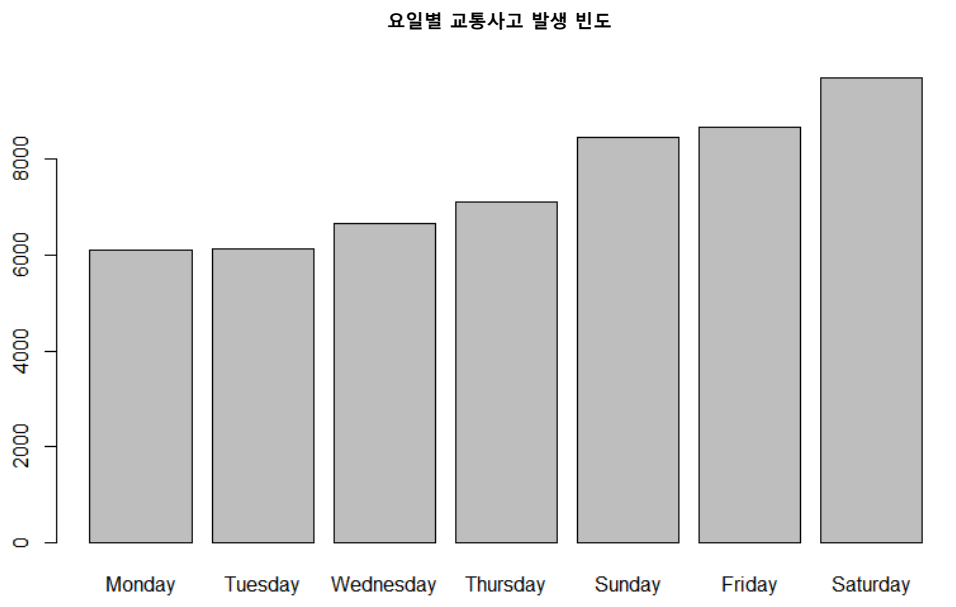


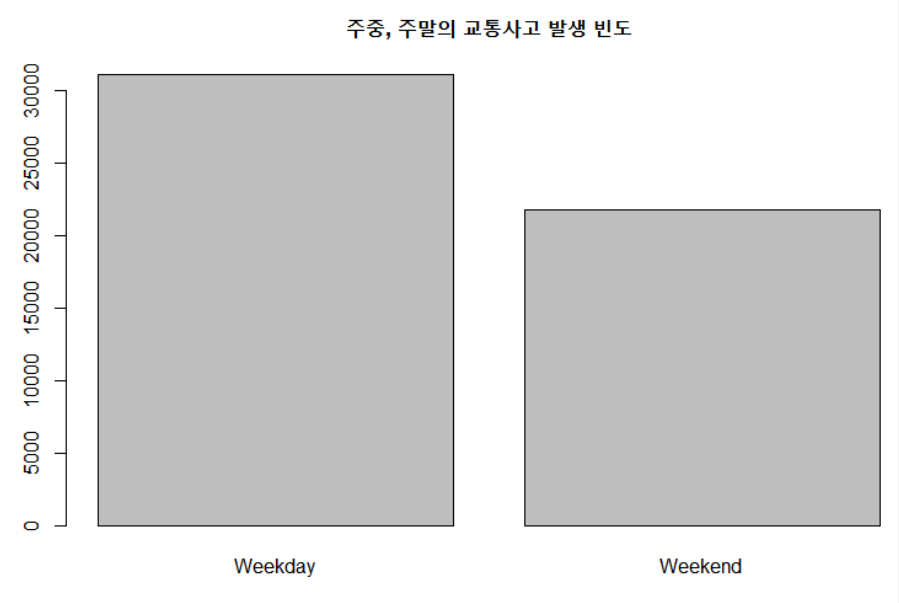


연도별 교통사고 발생 빈도는 1989년이 2800건으로 가장 많았고 2021년이 843건으로 가장 적은 분포를 띠고 있다. 이는 약 3배에 달하는 차이이다. 또한 시간의 흐름에 따라 교통사고 발생 빈도가 계속해서 줄어드는 것을 확인할 수 있다. (2021년은 9월까지 집계)

월별 교통사고 발생 빈도는 12월이 4721건으로 가장 많았고 2월이 3975건으로 가장 적었다. 그러나 전체 데이터수가 5만 건 이상임을 고려하였을 때, 월별 교통사고 발생건수는 유의미한 차이를 보여주지 못한다고 해석할 수 있다. 또한 barplot의 분포로 보았을 때 계절별 발생 빈도를 파악하는 것 또한 유의미한 결과를 도출하지 못할 것이라고 판단할 수 있다.

요일별 교통사고 발생 빈도 분석

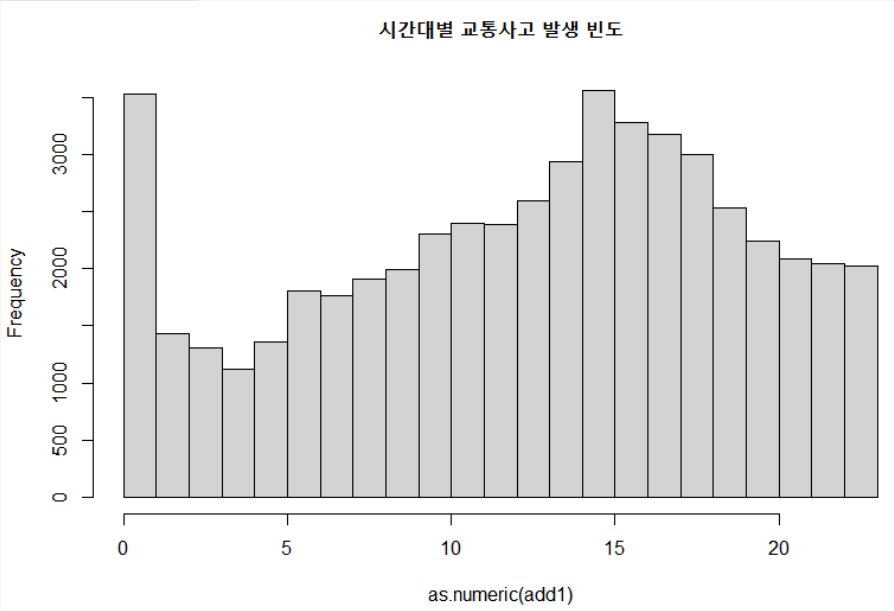


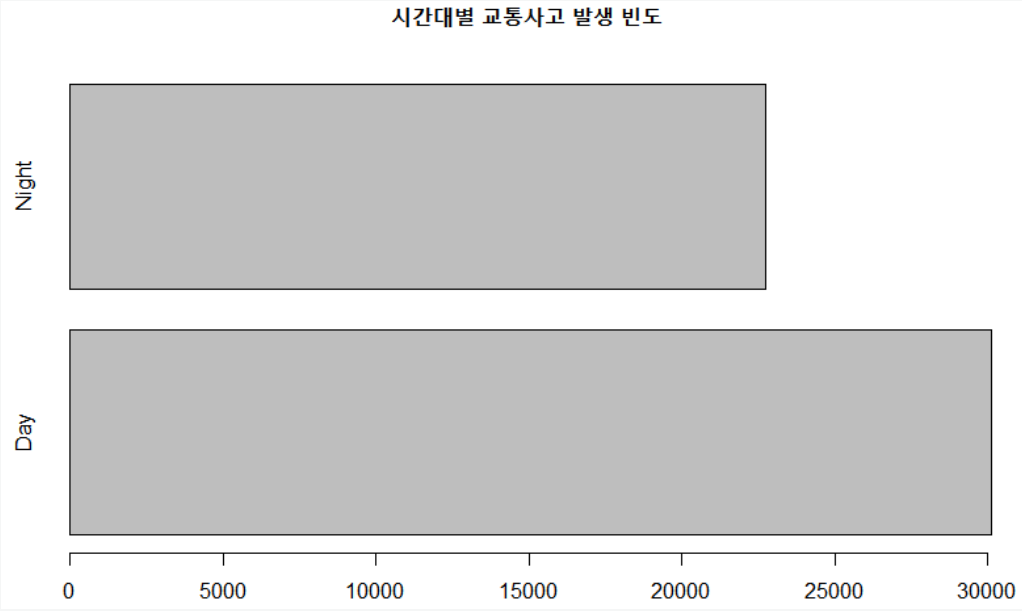


요일별 교통사고 발생 빈도의 분석에 있어 'Dayweek'와 'Day.of.week'의 두 가지 변수를 사용할 수 있다. 'Dayweek' 변수는 'Monday', 'Tuesday'와 같이 7개의 구체적인 요일로 나누어지며, 'Day.of.week' 변수는 'weekday'와 'weekend' 두 항으로 나뉘는 이항변수이다.

왼쪽 분포는 'Dayweek' 변수에 대한 barplot이며, 교통사고 발생 건수는 토요일이 9696건으로 가장 많았고 월요일이 6108건으로 가장 적었다. 특히 금요일과 토요일, 일요일 등 주말이 주중 요일에 비해 상대적으로 교통사고 빈도가 유의미하게 높은 것을 확인할 수 있다.
오른쪽 분포는 'Day.of.week' 변수에 대한 barplot이며, 이를 통해 요일이 교통사고 발생 빈도와 가지는 관계가 유의미함을 확인할 수 있다.

시간대별 교통사고 발생 빈도 분석

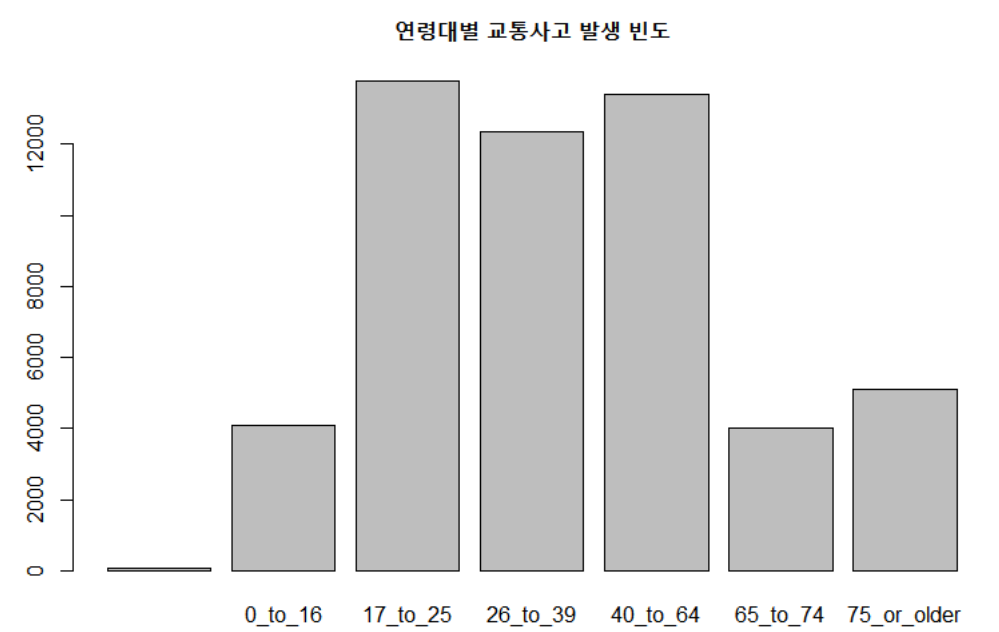


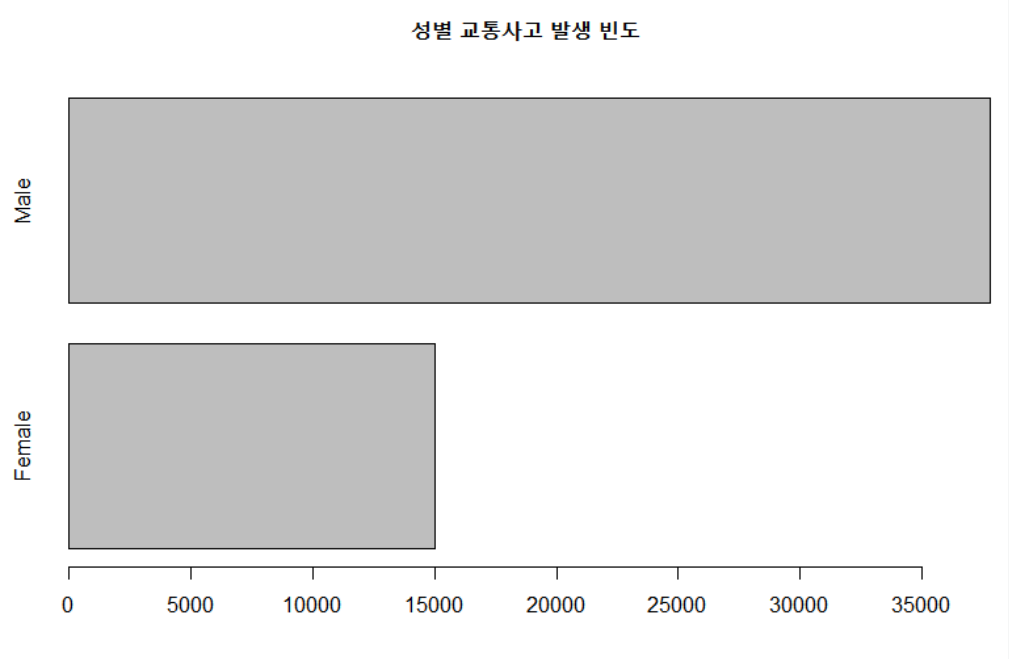


시간대별 교통사고 발생 빈도의 분석에 있어 'Time'과 'Time.of.day'의 두 가지 변수를 사용할 수 있다. 우선 Time 변수는 23:59와 같은 형태로 교통사고가 발생한 구체적인 시간을 나타내는 변수이고, Time.of.day 변수는 아침 6:00 ~ 저녁 17:59의 12시간 동안 발생했다면 'Day', 저녁 18:00 ~ 새벽 5:59의 12시간 동안 발생했다면 'Night'를 가지는 이항변수이다.

왼쪽 분포는 'Time' 변수의 각 행을 strsplit을 통해 :을 기준으로 절사하고, for문을 통해 리스트의 첫 열만을 선택하는 벡터를 생성하여 히스토그램을 그린 것이다. 분포를 살펴보면 0시와 15시~ 18시까지의 교통사고 발생 빈도가 가장 높다. 오른쪽 분포는 'Time.of.day' 변수에 대해 barplot을 그린 것이며, 낮 시간대에 발생한 교통사고가 30117건, 밤 시간대에 발생한 교통사고가 22726건임을 확인할 수 있다.

연령대별 & 성별 교통사고 발생 빈도 분석

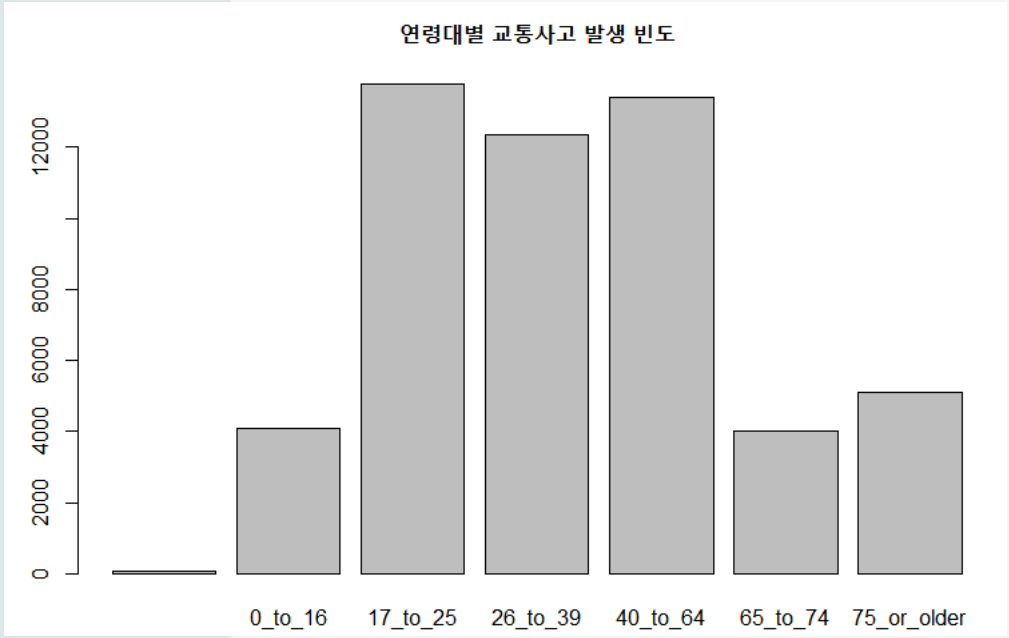


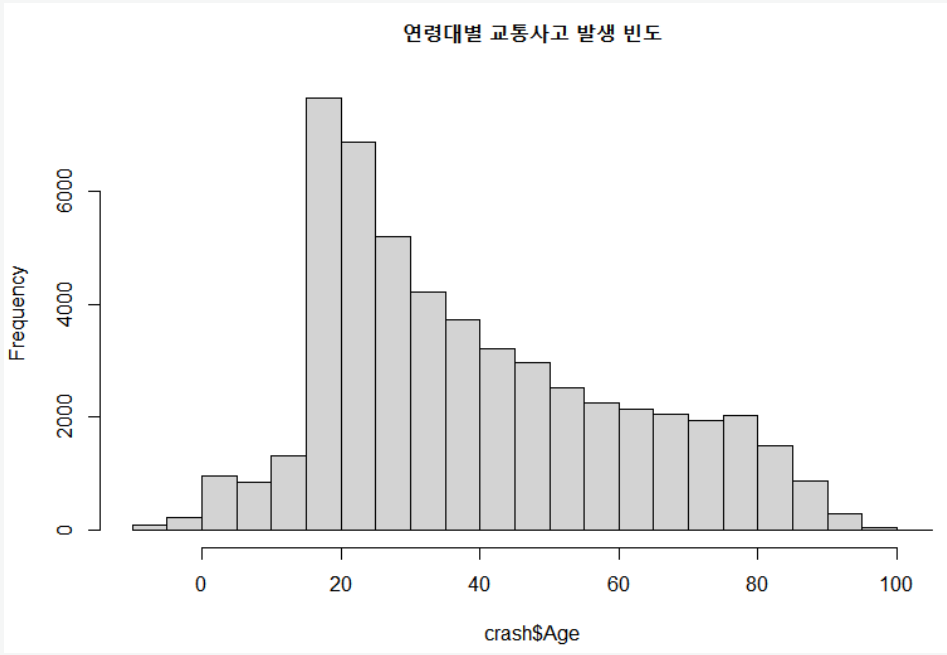


연령대별 교통사고 발생 빈도는 17세부터 25세까지의 연령대에서 13771건으로 가장 높으며, 65세부터 74세까지의 연령대가 4013건으로 가장 낮다. 특히 주요 경제활동인구이자 운전자의 대다수를 차지하는 17세부터 64세 연령대가 전체 교통사고의 75%를 차지하고 있다는 분석 결과를 얻을 수 있다.

성별 교통사고 발생 빈도의 분석에 있어, Gender 변수는 Female(여성)과 Male(남성)으로 이루어진 이항변수이다. 오른쪽 분포에서 남성의 교통사고 발생 건수가 전체 중 37813건으로, 여성의 교통사고 발생 건수인 15002건보다 두 배 이상 많음을 확인할 수 있다.

연령대별 교통사고 발생 빈도의 재검토

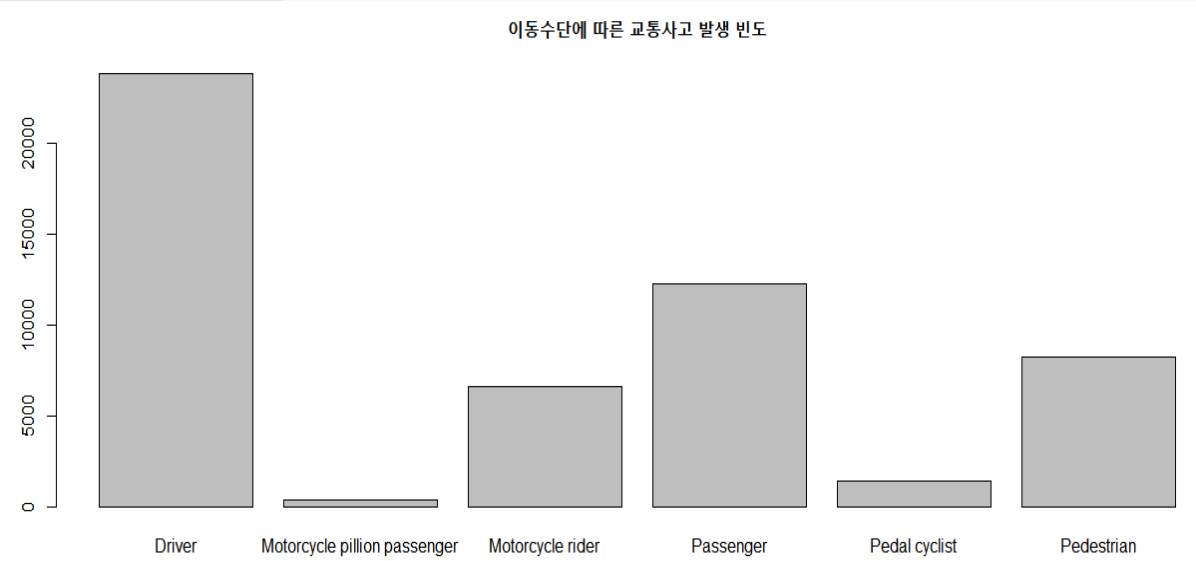


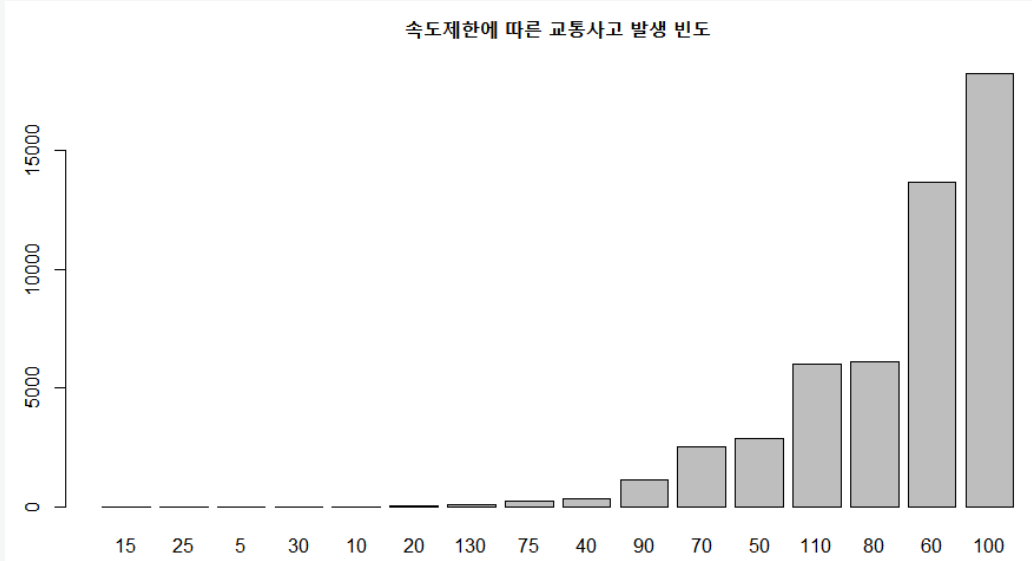


앞서 분석한 연령대별 성별 교통사고 발생 빈도는 범주형 변수 Age.group을 이용한 barplot이다. (왼쪽 분포)
그런데 수치형 변수 age를 이용해 breaks=5의 히스토그램을 그려 보면 그래프의 모양이 꽤나 다르게 도출된다. (오른쪽 분포)

왼쪽 분포의 age group에서 17_to_25 그룹과 40_to_64 그룹은 비슷한 빈도를 보이는 것으로 보이지만 각각의 그룹이 포함하고 있는 연령은 수치적으로 크게 차이가 난다. (각각 9, 25개) 따라서 정확한 분석을 위해서는 실제 연령 하나하나를 수치로 가지는 오른쪽 히스토그램을 검토하는 것이 적합할 것으로 보인다. 이때 **15세~25세의 연령대에서 교통사고 발생 빈도가 다른 연령대에 비해 확연히 높음**을 확인할 수 있다.

이동수단 & 속도 제한에 따른 교통사고 발생 빈도 분석





이동수단에 따른 교통사고 발생 빈도는 자동차가 가장 높은 비율로 나타났으며, 총 23816건으로 전체의 50% 가까이 차지했다. 한편 오토바이 뒷자리 승객이 교통사고 발생 빈도가 가장 낮은 것으로 나타났으며, 총 363건이 이에 해당한다.

속도 제한에 따른 교통사고 발생 빈도는 100km/h에서 가장 빈번하게 일어나며, 속도 40km/h 이하의 발생 빈도는 매우 낮음을 확인할 수 있다. 또한 130km/h 속도 제한에서는 오히려 매우 낮은 빈도를 보이는데, 이를 통해 속도 제한이 높을수록 교통사고 발생 빈도가 높은 것은 아님을 확인할 수 있다.

결론 및 맺음말

1. 교통사고 발생 빈도는 시간의 흐름(1989~2021)에 따라 계속 감소하고 있다. 이는 운전 방식의 간소화와 면허 수준의 향상, 교통사고 방지를 위한 차량 내외부 기술의 발전에 의한 것으로 추측할 수 있다.

2. 교통사고 발생 빈도는 계절이나 달에 크게 영향을 받지 않는다.

3. 교통사고 발생 빈도는 토요일에 가장 높으며 월요일에 가장 낮다. 또한 주말(금, 토, 일)이 주중에 비해 상대적으로 교통사고 빈도가 높다. 평일보다 주말에 휴일을 맞은 사람들의 이동량이 늘어나면서 절대적인 차량 통행량이 증가하기 때문이라고 볼 수 있다.

4. 24시간 중 0시와 15시~18시에 교통사고 발생 빈도가 가장 높다. 또 낮 시간대(6:00~18:00)가 밤 시간대(18:00~6:00)보다 교통사고 발생 빈도가 유의미하게 높다.

5. 15세~25세 연령대에서 교통사고가 가장 빈번하게 발생한다. 다른 연령대는 이에 비해 현저히 낮은 빈도를 갖는다.

6. 남성이 여성보다 교통사고 발생 빈도가 유의미하게 높다. 이는 절대적인 남성 운전자의 수가 여성보다 많다는 점도 고려할 필요가 있다.

7. 자동차가 다른 이동수단보다 교통사고 발생 빈도가 현저히 높다. 이는 자동차 통행량이 다른 교통수단 이용량보다 많기 때문임을 고려할 필요가 있다.

8. 100km/h 속도 제한 구간에서 교통사고 발생 빈도가 가장 높으며, 속도 제한이 높을수록 교통사고가 많이 일어나는 것은 아니다.

결론 및 맺음말

보완할 점

- 사용한 데이터셋이 거의 범주형 변수로 이루어져 있고, 수치형 값을 가지는 변수가 없어 상관분석에 어려움이 있었다.
- 빈도표와 barplot, 또는 histogram만 가지고 정확한 추측을 할 수 없다. 예를 들어, 남성이 여성에 비해 교통사고 발생 빈도가 높은 것은 절대적인 남성 운전자 수가 여성 운전자 수보다 많기 때문일 수 있다. 이러한 외부 요인을 고려하지 못하고 데이터셋 내에서의 정보만 활용하여 분석하였기에 한계가 많다.
- 전 세계의 교통사고 데이터가 아닌 한 국가만의 데이터를 활용한 것이기에 이 분석 결과가 우리나라, 혹은 전 세계의 경향을 유추하는 데 사용될 수 없다.
- 시각화의 측면에서 많은 옵션을 사용하지 못하고 기본적인 R 코드만 활용했다는 점에서 보완할 부분이 많다.

부록

데이터 출처

https://www.kaggle.com/deepcontractor/australian-fatal-car-accident-data-19892021

R code

```
crash <- read.csv(file="C:/data/Crash_Data.csv", header=T)
dim(crash)
names(crash)
class(crash)
head(crash)
library(ggplot2)
library(dplyr)
sum(is.na(crash))
crash$Age <- ifelse(crash$Age<0, NA, crash$Age)
na.omit(crash$Age)
```

부록

```
barplot(table(crash$Year), main='연도별 교통사고 발생 빈도’)
barplot(sort(table(crash$Month)), main=‘월별 교통사고 발생 빈도’)
barplot(sort(table(crash$Dayweek)), main=‘요일별 교통사고 발생 빈도’)
barplot(table(crash$Day.of.week), main=‘요일별 교통사고 발생 빈도’)
hour_split <- strsplit(crash$Time, split=“:”
add1<-c()
for(i in 1:length(hour_split)){
add1[i]<-hour_split[[i]][1]}
hist(as.numeric(add1))
barplot(table(crash$Time.of.day), main=‘시간대별 교통사고 발생 빈도’)
barplot(table(crash$Age.group), main=‘연령대별 교통사고 발생 빈도’)
hist(crash$Age, main='연령대별 교통사고 발생 빈도', breaks=seq(0,101,by=5))
barplot(table(crash$Gender), main=‘성별 교통사고 발생 빈도’)
barplot(table(crash$Road.User), main='이동수단에 따른 교통사고 발생 빈도’)
barplot(table(crash$Speed.Limit), main= ' 속도제한에 따른 교통사고 발생 빈도')
```