선형회귀분석을 바탕으로 은평 한옥마을 방문객들의 경향성 파악 및 추천 방문 장소 추천 시스템 제작

2학년 신상우, 김준석, 박준서, 신민성, 이준서, 임건, 조민경

하나고등학교 Hana Academy Seoul

서론

은평구에 위치한 은평한옥마을은 은평구의 대표적 관광명소이다. 한옥이 관광지로서 점점 더 많은 관심을 받고 있는 가운데, 은평한옥마을의 방문자 수 또한 상승세를 그리고 있다. 은평한옥마을을 찾는 관광객들의 취향과 방문 목적에 따라 은평한옥마을의 관광코스는 매우 다양하게 나타나기 때문에 처음 방문하는 사람에게 도움이 되는 관광정보가 필요하다. 따라서 본 연구는 이미 은평한옥마을을 방문한 사람들이 느끼는 만족도를 나이, 성별, 동행인원수에 따라 파악한 후, 누적된 정보를 개인 맞춤으로 제공하는 프로그램을 개발하는 것을 목표로 한다.

연구방법

1. 설문지 조사

- 은평한옥마을을 방문한 방문객 대상
- 방문객 정보 수집(나이, 성별, 동행인원 수) 및 장소별 만족도 조사 (카페, 한옥박물관, 한옥 구경, 등산, 진관사)

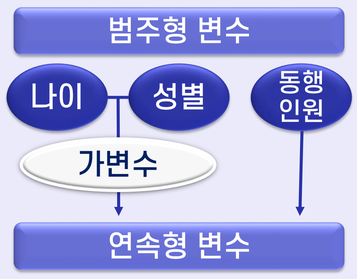
- 조사방법 ①: SNS 이용해 #은평한옥마을 해시태그를 단 사용자에게 설문지링크며 설문 요청
- 조사방법 ②: 은평한옥마을의 지역 에 직접 방문하여 설문지를 통해 대면 설문 요청
- 수집한 데이터의 표본의 크기: 약 340개

은평한옥마을을 즐길 거리 만족도 조사	
성별을 입력해주세요	○남자 ○여자
나이대 (10대면 10, 50대 이상은 50으로 입력해주세요)	○10대 ○20대 ○30대 ○40대 ○50대
일행 인원 수(4명 이상이면 4)	○1 ○2 ○3 ○4
카페 만족도 (1~10 사이의 수, 소수점 가능, 예시 :4.2, 8.9, 미방문은 0)	주관식
한옥 박물관 만족도 (1~10 사이의 수, 소수점 가능, 예시 :4.2, 8.9, 미방문은 0)	
한옥 구경 만족도 (1~10 사이의 수, 소수점 가능, 예시 :4.2, 8.9, 미방문은 0)	
등산 만족도 (1~10 사이의 수, 소수점 가능, 예시 :4.2, 8.9, 미방문은 0)	
진관사 만족도 (1~10 사이의 수, 소수점 가능, 예시 :4.2, 8.9, 미방문은 0)	

▲ 위 표는 실제 사용한 설문지의 내용과 동일

2. 변수 설정

①독립변수



나이, 성별은 범주형 변수이지만, 가변수 개념을 이용하여 연속형 변수로 취급 가능하다. 동행 인원수는 연속형 변수로 볼 수 있다.

*가변수: 독립변수를 0,1로 변환한 것.

②종속변수

- 장소별 만족도
- 1~10의 소수점
- 선형회귀를 사용할 수 있는 연속형 종속변수

*선형회귀: 종속변수 y와 한 개 이상의 독립변수 x와의 상관관계를 모델링하는 수학적 자료 분석 방법이다.

→ 종속변수인 만족도 데이터가 연속형 데이터이므로 일반적으로 선형회귀 실시할 수 있다.

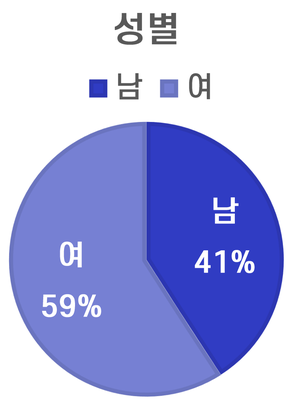
3. 데이터 분석

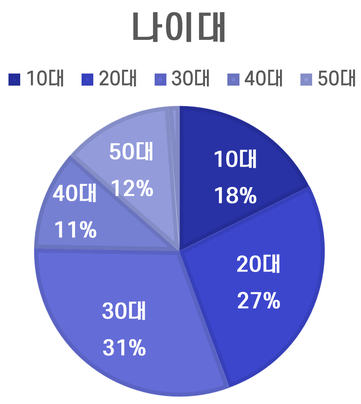
- 선형회귀 위해서 ggplot2 라이브러리 사용
- R코드 (▼장소:카페)

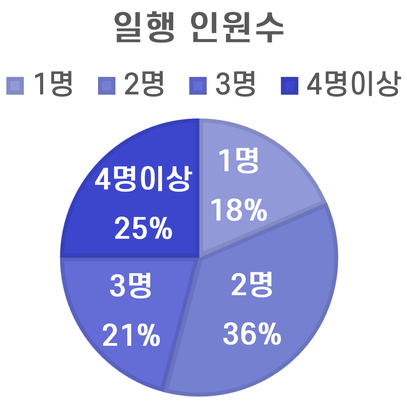
```
sa<- read.csv("GLim_v3.csv&
head(sa)
attach(sa)
sa= sa[sa[,5]!=0,]
pairs(sa[,2:5])
sex=as.factor(sex)
age=as.factor(age)
summary(sex)
summary(age)
par(mfrow=c(2,2))
plot(sex,cafe,xlab="sex",ylab="cafe")
plot(age,cafe,xlab="age",ylab="cafe")
plot(as.factor(number),cafe,xlab="numb
er",ylab="cafe")
lm.out1=lm(cafe~sex+age+number)
summary(lm.out1)
lm.out2=lm(cafe~sex+age+number+sex:
age+sex:number+age:number)
summary(lm.out2)
```

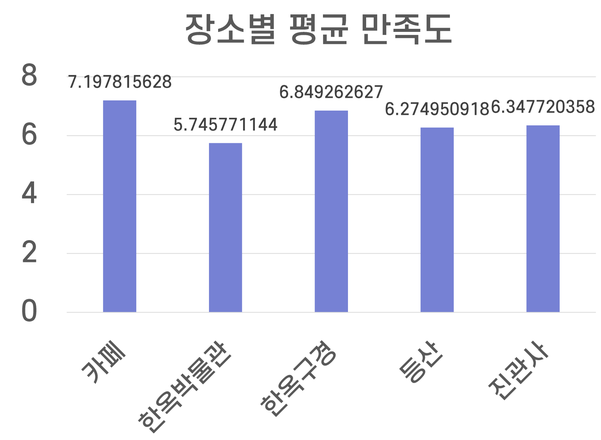
결과 및 고찰

1. 설문지 결과 설문지 수합 이후, 남녀, 나이대, 동행 인원수에 따른 응답 비율과 장소별 평균 만족도 및 나이대별 응답 인원수 비율을 나타낸다.









3. 선형회귀 결과 분석

- 장소별로 (카페, 한옥박물관, 한옥 구경, 등산, 진관사) 총 5가지 경우로 선형회귀를 실시했다.

①카페

```
Call:
lm(formula = cafe ~ sex + age + number)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6357 -1.2346  0.3527  1.7110  4.8495

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8014     0.4328  11.095  < 2e-16 ***
sex1          0.3230     0.2793   1.157   0.2482
age20         2.1622     0.3955   5.466 9.02e-08 ***
age30         1.6706     0.3911   4.272 2.53e-05 ***
age40         1.1260     0.4867   2.314   0.0213 *
age50         0.7350     0.4566   1.610   0.1083
number        0.1745     0.1207   1.445   0.1493
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.312 on 333 degrees of freedom
Multiple R-squared:  0.123,     Adjusted R-squared:  0.1072
F-statistic: 7.781 on 6 and 333 DF,  p-value: 7.558e-08
```

▷ 성별에 따른 차이는 유의 수준이 0.2482이기 때문에 존재하지 않는다고 볼 수 있다.
동행 인원수에 따른 만족도는 존재하나 유의하지 않다.

20대의 만족도가 2.1622로 1보다 크기 때문에 비교 대상인 10대보다 더 높다.
30대의 만족도가 1.6706로 1보다 크기 때문에 비교 대상인 10대보다 더 높다.
40대의 만족도가 1.1260로 1보다 크기 때문에 비교 대상인 10대보다 더 높다.
50대의 만족도가 0.7350로 1보다 작기 때문에 비교 대상인 10대보다 더 낮다.

비교적 10대에 비해 20, 30대가 카페 만족도가 높은 것으로 드러났으며, 40, 50대 이상은 10대보다 조금 높지만 유의한 차이는 아닌 것을 볼 수 있다.

▲ 선형회귀모델링 (카페)
※ sex0, age10을 기준으로 하였기 때문에 비교대상이 남자, 10대이다.

② 한옥박물관
성별에 따른 차이는 유의 수준이 0.009135이기 때문에 존재한다고 할 수 있다.
동행 인원수에 따른 만족도는 유의하게 존재한다.
동행 인원수가 1명 씩 증가할 때마다 만족도가 0.3367씩 증가하며 동행 인원수에 비례하여 만족도가 증가한다.
20대, 30대, 40대, 50대 모두 10대보다 만족도가 높다.

③ 한옥 구경
성별에 다른 차이는 여성이 남성보다 0.6864만큼 높은 만족도를 보인다.
동행 인원수에 따른 만족도는 유의하지 않게 존재한다.
50대의 만족도가 가장 낮고, 나이대가 증가함에 따라 만족도가 더욱 많이 감소하는 경향이 있다.

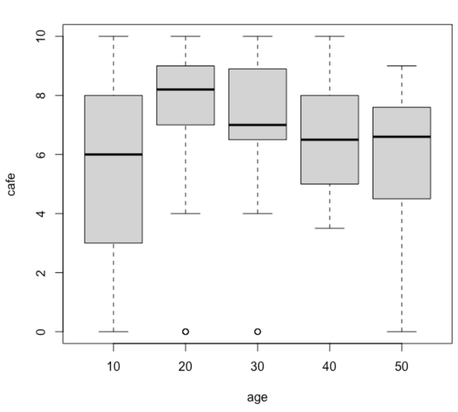
④ 등산
전체적으로 통계상 유의한 차이를 가지는 집단이 존재하지 않는다.
20대와 30대는 만족도가 낮은 반면 40개와 50대는 상대적으로 만족도가 높아진다.
중장년층의 등산 선호에 따른 경향으로 해석 가능하다.
성별의 경우 여자가 남자에 비해 0.5738 정도 낮으나, 통계적으로 유의한 결과는 아니다.

⑤ 진관사 방문
성별과 동행 인원수에서 통계적으로 유의한 차이는 없다.
나이대의 경우 10대보다 20대, 30대, 40대, 50대의 만족도가 더 낮다.
이는 나이대가 높아질수록 만족도가 높아진다고 예상한 가설에 부합하지 않는 결과이다.

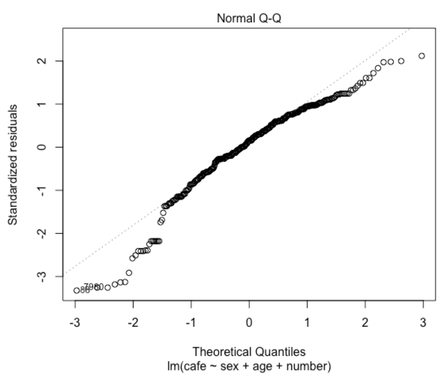
2. 데이터 제공 프로그램

```
for u in data:
  _match = 0
  for keyName, key in user.items():
    if u[keyName] == key:
      _match += 1
  if _match == 3:
    cnt += 1
    for k in places:
      stats[k] += u[k]

print(f"총 {cnt} 개의 응답이 있습니다")
if cnt != 0:
  res = { k: v / cnt for k, v in stats.items() }
  for k, v in res.items():
    v = round(v,1)
    print(f"{k} 에 대한 만족도:{v}")

input("Press any key to continue")
```

▲ 위 코드는 데이터 제공 프로그램 코드의 일부이다.
※지금까지 받은 데이터를 읽어와 사용자가 자신의 정보를 입력했을 때 (2*5*4=총 40가지 경우) 같은 경우의 장소별 만족도를 소수점 둘째 자리에서 반올림하여 보여준다.



▲ 위 그래프는 box plot (카페, 나이대)
※회색영역이 전체 데이터의 50%포함 (25%~75%)



▲ 위 그래프를 통해 선형회귀의 가정 중 정규성을 확인할 수 있다.
※Residual이 정규분포를 따름은 standard residual이 우상향하는 직선상에 모여있음을 통해 볼 수 있다.

4. 사용자로의 적용

- 실제 사용자가 자신의 정보를 입력했을 때 선형회귀모델링을 바탕으로 예상되는 만족도값을 출력해주는 python 코드이다.

```
x1 = int(input('남자라면 0, 여자라면 1을 입력해주세요 : '))
k = int(input('₩n나이대를 입력해주세요(50대 이상이면 50으로 입력해주세요) :'))
x7 = int(input('₩n동행인원수를 입력해주세요(4인 이상이라면 4라고 입력해주세요) : '))
def SA(x1, x2, x3, x4, x5, x6, x7):
return 4.8014 + 0.3230 * x1 + 2.1622 * x3 + 1.6706 * x4 + 1.1260 * x5 +
0.7350 * x6 + 0.1745 * x7
if k == 10:
print(SA(x1,1,0,0,0,0,x7))
elif k == 20:
print(SA(x1,0,1,0,0,0,x7))
elif k == 30:
print(SA(x1,0,0,1,0,0,x7))
elif k == 40:
print(SA(x1,0,0,0,1,0,x7))
else:
print(SA(x1,0,0,0,0,1,x7))
```

▼파이썬 코드 실행 시 결과 화면

```
=============== RESTART: /Users/matthewgunnlim/Dowr
남자라면 0, 여자라면 1을 입력해주세요 : 1

나이대를 입력해주세요(50대 이상이면 50으로 입력해주세요) :20

동행인원수를 입력해주세요(4인 이상이라면 4라고 입력해주세요) : 3
7.8101
```

결론 및 제언

- 설문 결과를 바탕으로 장소별 만족도를 구하고, 방문객의 특성별로 평균 만족도를 분석했다.
- 설문 결과가 계속 누적될수록 더 많은 설문 결과가 생길 것이고, 그에 따라 최신 트렌드가 반영된, 더욱 정확한 만족도 및 선호도를 파악할 수 있다.
- 많은 설문 결과를 바탕으로 계절, 기후 등 더욱 다양한 척도를 기준으로 결과를 분석할 수 있을 것이며 이를 토대로 더욱 정교한 추천 알고리즘을 구축할 수 있을 것이다.
- 설문 결과를 계속 늘려나가 이용자가 더욱 만족할 수 있도록 도울 뿐만 아니라 관광지로서의 은평한옥마을 개발에 많은 도움이 될 것이라고 생각한다.