

# 다면량통계분석방법론 -스피드 데이팅 분석-

2018150408 이충은  
2018150453 곽동호

---

- 데이터 소개 및 목적
- 데이터 전처리
- 1. 설문 데이터 EDA
- 2. 데이트 내에서의 결정 영향변수
- 3. 인기 있는 사람의 비결
- 4. 자기 객관화의 중요성
- 5. 결론 및 제언

# 스피드 데이팅(Speed dating) 데이터

## 데이터 소개 및 목적

## 데이터 전처리

1. 설문 데이터 EDA

2. 데이터 내에서의 결정 영향변수

3. 인기 있는 사람의 비결

4. 자기 객관화의 중요성

5. 결론 및 제언

- 스피드 데이팅이란? 많은 미혼 남녀가 한 장소에 모여서 계속 상대를 바꿔가며 5분 가량 일대일 대화를 하는 것.
- 데이터 출처: kaggle에 있는 speed dating experiment data set<sup>1)</sup>
  - > Raymond와 Sheena의 논문<sup>2)</sup> 작성 과정에서 수집

1) <https://www.kaggle.com/datasets/annavictoria/speed-dating-experiment>

2) Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment, May 2006

# 스피드 데이팅(Speed dating) 데이터

## 데이터 소개 및 목적

## 데이터 전처리

1. 설문 데이터 EDA

2. 데이트 내에서의 결정 영향변수

3. 인기 있는 사람의 비결

4. 자기 객관화의 중요성

5. 결론 및 제언

### - 데이터 세부내용

- 1) 데이팅 실험 참여자들에 대한 설문조사들의 답변으로 구성
- 2) 총 4번 설문조사 실시: 데이팅 참여전, 데이팅,  
데이팅 다음날, 데이팅 3주 후
- 3) 각 설문 조사별 공통 문항, 개별 문항 모두 존재: 총 195개 변수

# 데이터 분석의 목적

데이터 소개 및 목적

데이터 전처리

1. 설문 데이터 EDA

2. 데이터 내에서의 결정 영향변수

3. 인기 있는 사람의 비결

4. 자기 객관화의 중요성

5. 결론 및 제언

1. 데이터 시각화를 통한 인사이트 도출
2. 다변량 통계분석을 통한 스피드 데이팅에서의 이성 선택 요인 파악
3. 성별간 사고방식의 차이 파악(자기 객관화 등)

# 데이터 전처리

## 데이터 소개 및 목적

## 데이터 전처리

### 1. 설문 데이터 EDA

### 2. 데이터 내에서의 결정 영향변수

### 3. 인기 있는 사람의 비결

### 4. 자기 객관화의 중요성

### 5. 결론 및 제언

## - 분석 목적에 필요 없는 정보 제거

```
# 데이터 전처리
```

```
data <- data %>%
```

```
  filter(wave < 6 | wave > 9) # wave 6~9 제외
```

```
# 제거할 Column
```

```
drop_idx = c('id', 'idg', 'round', 'condtn', 'position', 'positin1', 'partner',  
           'race_o', 'pf_o_att', 'pf_o_sin', 'pf_o_int', 'pf_o_fun', 'pf_o_amb',  
           'pf_o_shd', 'dec_o', 'attr_o', 'sinc_o', 'intel_o', 'fun_o',  
           'amb_o', 'shar_o', 'like_o', 'prob_o', 'met_o', 'field', 'undergra',  
           'mn_sat', 'tuition', 'from', 'zipcode', 'income', 'career',  
           'career_c', 'exphappy', 'numdat_3', 'num_in_3')
```

```
drop_data <- data %>% select(-c(drop_idx))
```

실험 방식이 다름

다른 설문과 내용이 겹치거나  
분석 내용과 관련 없음

## - 범주형 변수 factor화(성별, 인종, 결정여부 등)

```
data$gender <- as.factor(data$gender)
```

```
data$samerace <- as.factor(data$samerace)
```

```
data$race <- as.factor(data$race)
```

```
data$dec <- as.factor(data$dec)
```

```
data$date <- as.factor(data$date)
```

# 데이터 전처리

데이터 소개 및 목적

데이터 전처리

1. 설문 데이터 EDA

2. 데이터 내에서의 결정 영향변수

3. 인기 있는 사람의 비결

4. 자기 객관화의 중요성

5. 결론 및 제언

- 데이터 표준화

```
data <- data %>%
  mutate(sum1_1 = attr1_1 + sinc1_1 + intell1_1 + amb1_1 + shar1_1 + fun1_1,
       sum2_1 = attr2_1 + sinc2_1 + intel2_1 + amb2_1 + shar2_1 + fun2_1,
       sum1_2 = attr1_2 + sinc1_2 + intell1_2 + amb1_2 + shar1_2 + fun1_2,
       sum2_2 = attr2_2 + sinc2_2 + intel2_2 + amb2_2 + shar2_2 + fun2_2) %>%
  mutate(across(c(attr1_1, sinc1_1, intell1_1, amb1_1, shar1_1, fun1_1), ~ (.x / sum1_1) * 100)) %>%
  mutate(across(c(attr2_1, sinc2_1, intel2_1, amb2_1, shar2_1, fun2_1), ~ (.x / sum2_1) * 100)) %>%
  mutate(across(c(attr1_2, sinc1_2, intell1_2, amb1_2, shar1_2, fun1_2), ~ (.x / sum1_2) * 100)) %>%
  mutate(across(c(attr2_2, sinc2_2, intel2_2, amb2_2, shar2_2, fun2_2), ~ (.x / sum2_2) * 100))
```



설문의 측정 단위가 서로 다른 것을 표준화

- 각 분석에 필요한 변수들로 새로운 데이터 셋들 제작

1. 데이터 현장에서의 설문만 담은 date\_data 제작

2. 각 사람을 하나의 obs로 요약한 Personal\_data제작



이를 바탕으로 다양하게 활용하여 분석 진행

데이터 소개 및 목적

데이터 전처리

1. 설문 데이터 EDA

2. 데이트 내에서의 결정 영향변수

3. 인기 있는 사람의 비결

4. 자기 객관화의 중요성

5. 결론 및 제언

# 사전 설문 데이터 EDA

---

- 설문 조사 시행 시간 별 남녀별 주요 고려 요소 설문 결과 비교  
(성별 각자의 관점에서)
- 관심사의 유사 정도와 매칭 성사 간의 관련성

# 남성의 이성 선택간 주요 고려 요소 EDA

1.

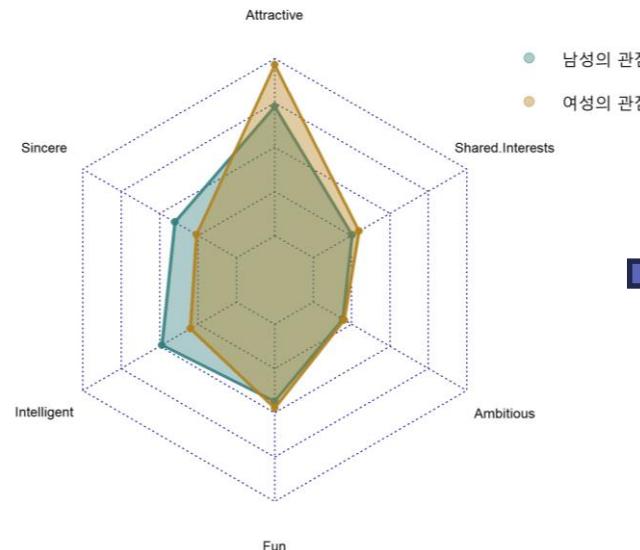
2.

3.

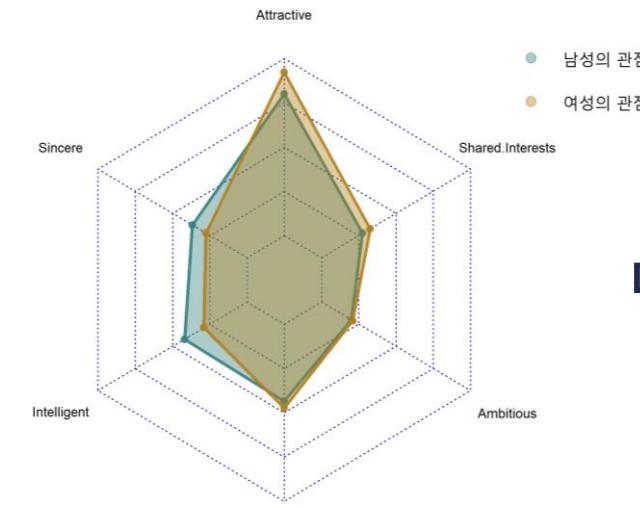
4.

5.

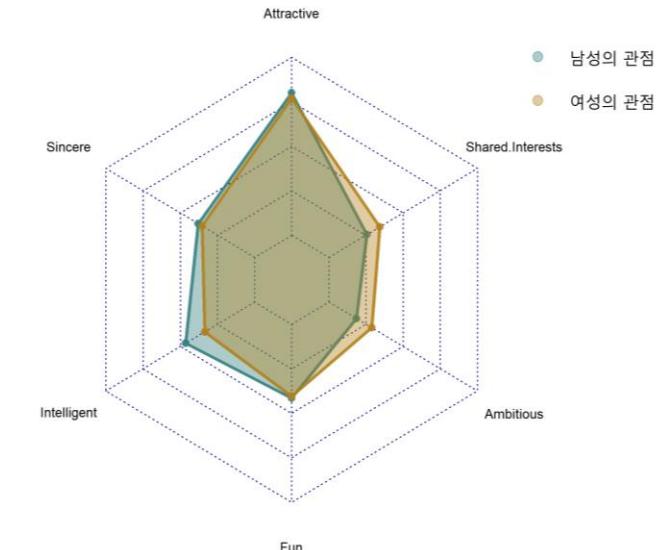
남성이 중요하게 보는 것에 대한 관점(사전 조사)



남성이 중요하게 보는 것에 대한 관점(사후 조사: 다음날)



남성이 중요하게 보는 것에 대한 관점(사후 조사: 3주 후)



- 남성이 중요하다 생각하는 것에는 데이팅 진행 전과 후에 뚜렷한 변화가 보이지 않는다.
- 여성이 바라본 관점의 경우 데이팅 전에는 남성들의 주요 고려 요소가 외모에 실제보다 더 치우쳐져 있을 거라 생각하나 데이팅 후 시간이 지날 수록 점차 남성들의 실제 생각과 비슷하게 가는 모습을 보인다.

# 여성의 이성 선택간 주요 고려 요소 EDA

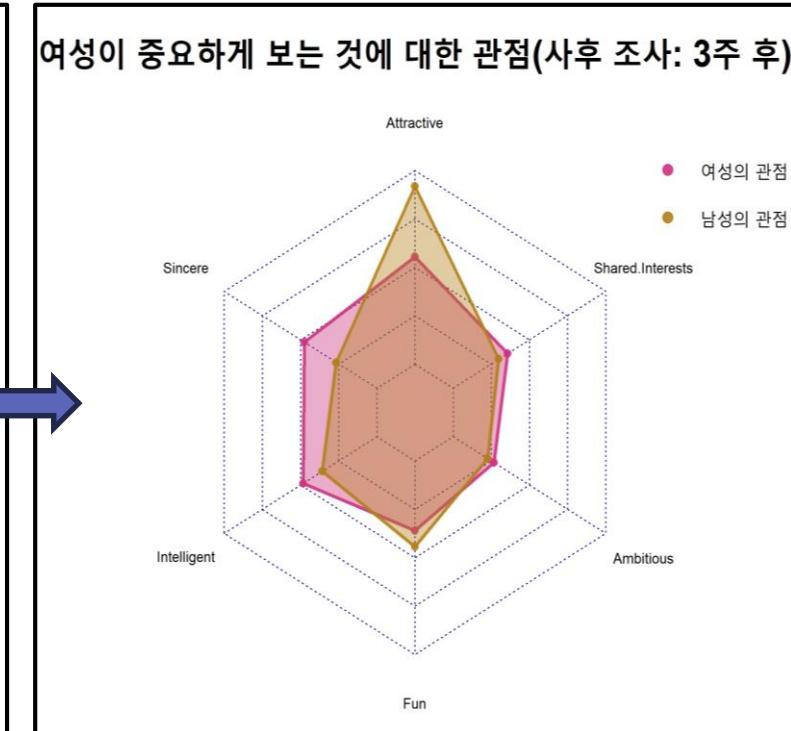
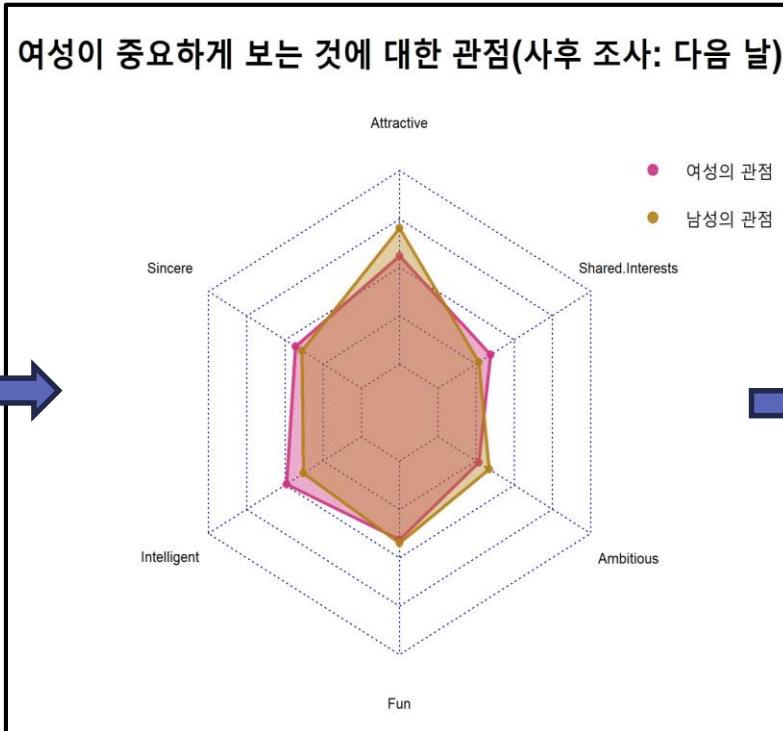
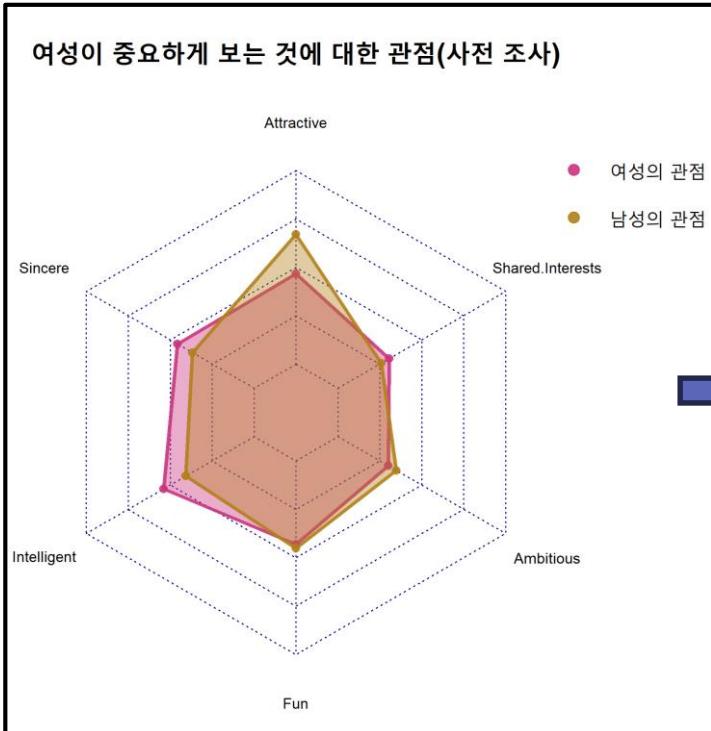
1.

2.

3.

4.

5.



- 여성의 경우에도 데이팅 진행 전후로 이성 선택에 있어 중요하게 생각하는 것이 변하지 않는다.

- 남성의 경우 여성과 다르게 데이팅 진행 후 시간이 지나면서 여성의 주요 고려에 가까워지는 것이 아닌 더 멀어지는 모습을 보인다.

# 관심사의 유사 정도와 성사 여부 EDA

1.

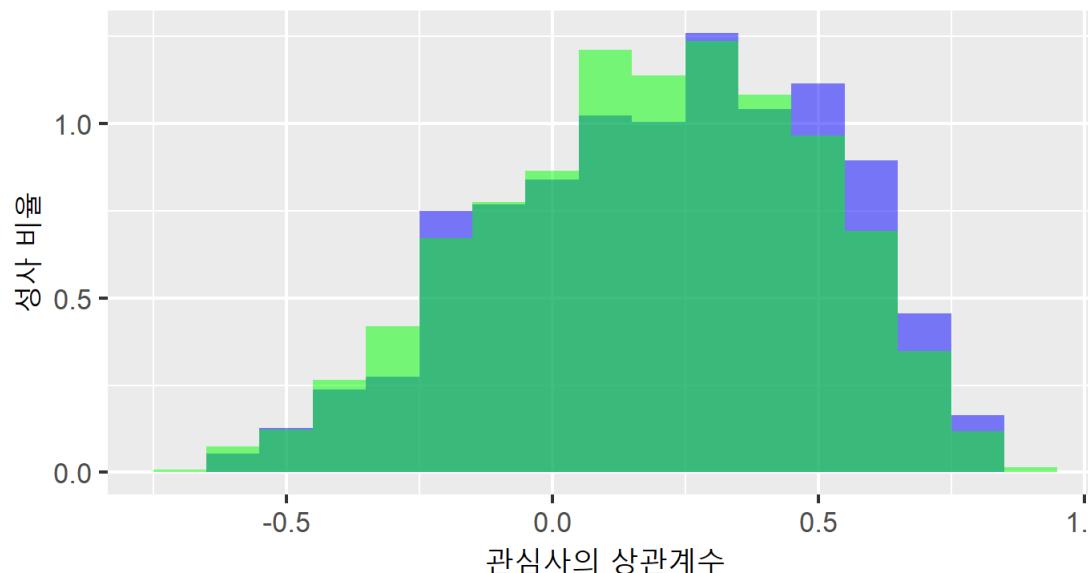
2.

3.

4.

5.

관심사의 비슷한 정도에 따른 성사 횟수



category  
성사 O  
성사 X



일반적인 생각과 달리 관심사의 상관계수가 높다고 성사비율이 올라가는 것이 아니라는 것을 확인



관심사를 더 세분화하여 확인하는 분석 진행

데이터 소개 및 목적

데이터 전처리

1. 설문 데이터 EDA

2. 데이터 내에서의 결정 영향변수

3. 인기 있는 사람의 비결

4. 자기 객관화의 중요성

5. 결론 및 제언

# 데이터 내에서의 결정 영향변수

---

- 남녀별 데이터 내에서의 결정에 영향을 준 변수 회귀 분석
- 변수들의 명확한 영향을 보기 위해 주성분 분석 진행

# 남녀별 데이트 내에서 결정에 영향을 준 변수 분석

1.

2.

3.

4.

5.

ID #:	1	2	3	4	5	6	7	8	9	10
<u>dec</u> Decision										
	1=yes 0=no	yes no								
Attributes (1=awful, 10=great)	1~10									
Attractive	<u>attr</u>									
Sincere	<u>sinc</u>									
Intelligent	<u>intel</u>									
Fun	<u>fun</u>									
Ambitious	<u>amb</u>									
Shared Interests/ Hobbies	<u>shar</u>									

- 데이트 하나가 끝날 때마다 참가자 각각 좌측과 같은 설문지 작성



- 남녀별로 파트너의 **Attributes**에 해당하는 6가지 변수
  - + 나이 차이 변수(**age\_diff**)를 설명 변수로,
  - dec**(상대방 선택 여부) 변수를 종속 변수로 **회귀 분석** 진행
- 유의하지 않은 변수 제거 및 **PCA** 진행

# 로지스틱 회귀분석 진행 – 남성

1.

2.

3.

4.

5.

```
Call:  
glm(formula = as.numeric(dec) - 1 ~ attr + sinc + fun + amb +  
    shar + age_diff, family = binomial, data = men_date_data)  
  
Deviance Residuals:  
    Min      1Q  Median      3Q     Max  
-2.6079 -0.8321 -0.2337  0.8486  3.0806  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -5.10386  0.29469 -17.319 < 2e-16 ***  
attr        0.65726  0.03641  18.051 < 2e-16 ***  
sinc       -0.15695  0.03604  -4.356 1.33e-05 ***  
fun         0.24922  0.03811   6.540 6.16e-11 ***  
amb        -0.17089  0.03534  -4.836 1.33e-06 ***  
shar        0.26423  0.02877   9.185 < 2e-16 ***  
age_diff    -0.03838  0.01662  -2.309  0.0209 *  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 3900.8 on 2820 degrees of freedom  
Residual deviance: 2867.3 on 2814 degrees of freedom  
(결측으로 인하여 592개의 관측치가 삭제되었습니다.)  
AIC: 2881.3  
  
Number of Fisher Scoring iterations: 5
```



- 유의하지 않은 'intel' 변수 제거
- 외모가 가장 큰 긍정적인 영향을 주는 것을 알 수 있고 그 뒤를 유머와 취미 공유가 따른다.
- 진실성과 미래에 대한 야망은 오히려 부정적인 영향을 미치는 것으로 나타남.
- 나이 차이의 회귀 계수가 상대적으로 작은 것으로 보아 영향을 크게 미치지 않음을 알 수 있다.
- Cox-snell R-squared=0.41

# 주성분 분석(PCA) 진행 – 남성

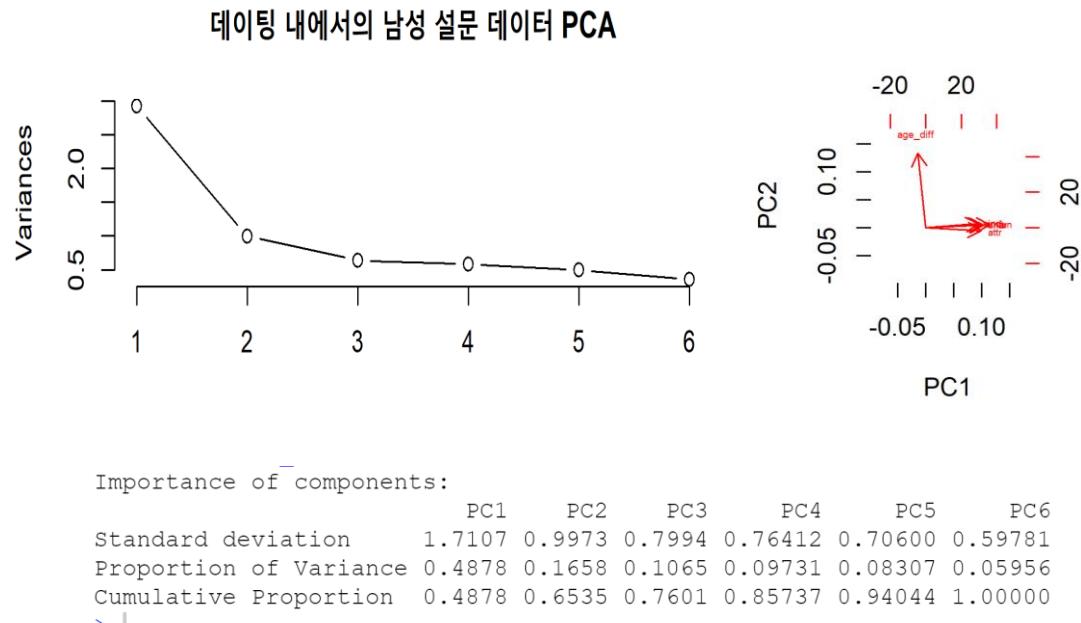
1.

2.

3.

4.

5.



- 제 1 주성분은 데이터 내의 점수 부여와 관련
- 제 2 주성분은 절대적인 나이차이와 관련
- 나이차이를 제외하고 나머지 5개 변수들에서 서로 상관관계가 존재하는 것을 알 수 있다.

# 로지스틱 회귀분석 진행 – 여성

1.

2.

3.

4.

5.

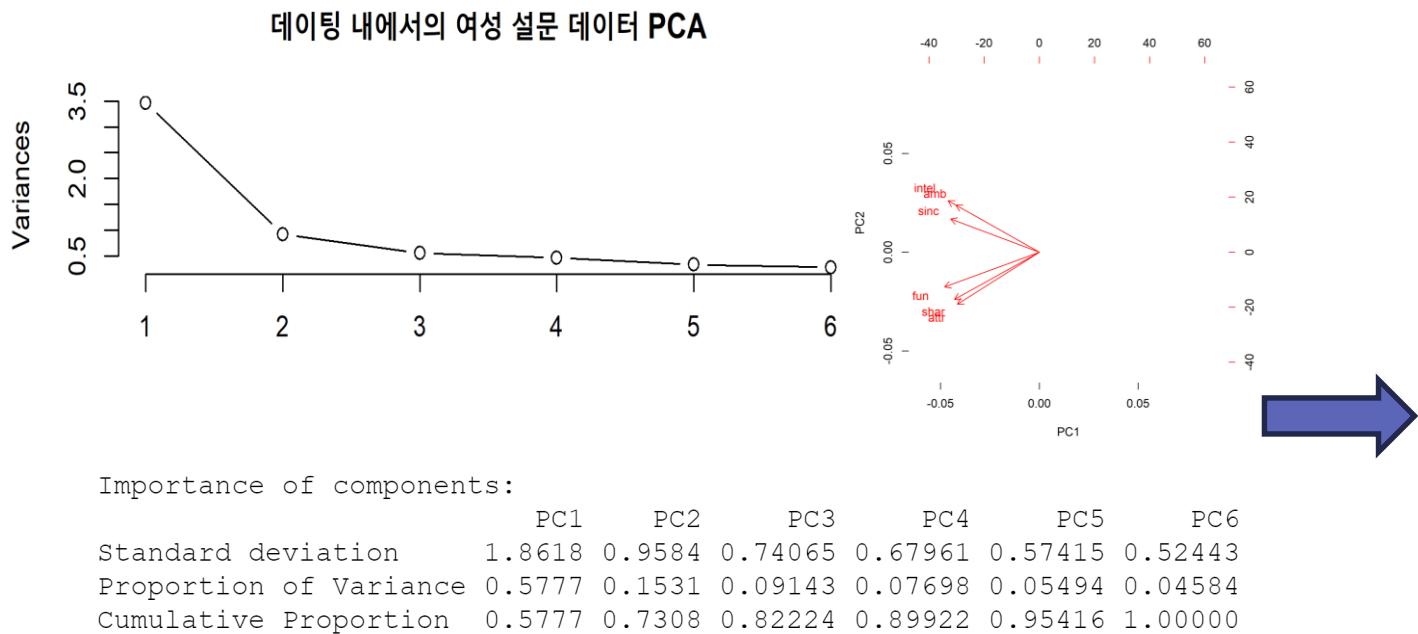
```
Call:  
glm(formula = dec ~ attr + sinc + intel + fun + amb + shar, family  
= binomial(),  
    data = women_date_data)  
  
Deviance Residuals:  
    Min      1Q  Median      3Q     Max  
-2.4288 -0.8290 -0.3602  0.8419  3.1905  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -5.58819  0.30849 -18.115 < 2e-16 ***  
attr          0.41703  0.03336  12.502 < 2e-16 ***  
sinc         -0.12128  0.03846 -3.153  0.00161 **  
intel         0.16352  0.05036  3.247  0.00117 **  
fun           0.27909  0.03764  7.415 1.22e-13 ***  
amb          -0.18760  0.03800 -4.937 7.93e-07 ***  
shar          0.29764  0.03021  9.854 < 2e-16 ***  
---  
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 3624.5 on 2725 degrees of freedom  
Residual deviance: 2714.5 on 2719 degrees of freedom  
AIC: 2728.5  
  
Number of Fisher Scoring iterations: 5
```



- 유의하지 않은 'age\_diff' 변수 제거
- 남자와 비교하여 **외모**가 가장 큰 긍정적인 영향을 주는 것은 같으나 회귀계수의 차이로 보아 그 정도는 낮다.
- 또한 마찬가지로 **진실성**과 미래에 대한 **야망**은 부정적인 영향을 미치는 것으로 나타남.
- 그 뒤를 **유머**와 **취미 공유**가 따르는 것은 같으나 데이팅 파트너 남성의 **지성**도 어느 정도 영향을 미치는 것을 알 수 있다.
- nagelkerke R-squared= 0.39

# 주성분 분석(PCA) 진행 – 여성

1.  
2.  
3.  
4.  
5.



- 각 변수들이 주성분에 **골고루 기여**하고 있다.
- **유머, 취미 공유, 외모** 변수끼리 서로 상관관계가 강하고 **지성, 야망, 진실성** 변수끼리 서로 상관관계가 강함을 볼 수 있다.

데이터 소개 및 목적

데이터 전처리

1. 설문 데이터 EDA

2. 데이터 내에서의 결정 영향변수

3. 인기 있는 사람의 비결

4. 자기 객관화의 중요성

5. 결론 및 제언

# 인기 있는 사람의 비결

-남/여 간 그룹 비교

-다시 만날 의사가 있다고 답한 비율을 설명변수로 회귀모형 적합

-회귀 가정 검정

-이성에게 선택을 많이 받은 사람들은 어떤 사람들인가?

# 남-여 간 그룹비교

이성에게 Okay!를 받는 비율에 성별 간 유의한 차이가 있는가?

1.

```
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group    1  0.1845 0.6677
        447
```

2.

3.

```
Df Sum Sq Mean Sq F value    Pr(>F)
gender      1   7948    7948   14.86 0.000133 ***
Residuals  447 239093     535
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.

5.

등분산 가정 만족!

ANOVA 결과

여자 mean: 46.38% vs 남자 mean: 37.97%

여자가 남자에 비해 유의하게 Okay를 받는 비율이 높다!

# 어떤 사람을 다시 만나고 싶은가?

사람마다 이성에게 다시 만나고 싶다고 응답받은 비율을 설명변수로 회귀 모델 적합

1.

2.

3.

4.

5.

## 남성과 여성을 따로 모델링

per\_par\_dec  
(만난 파트너 중 애프터  
신청을 받은 비율)

Im

개인에 대한 변수

파트너가 평가한  
점수의 평균

StepAIC

변수선택된 최종  
회귀 모델

# 적합 결과 - 남성(여성이 남성을 볼 때)

```
Call:  
lm(formula = per_par_dec ~ attr1_1 + amb1_1 + attr2_1 + sinc2_1 +  
    fun3_1 + tvsports + clubbing + concerts + music + yoga +  
    mean_attr_o + mean_fun_o + mean_shar_o, data = reg_data_m)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-33.594  -7.596 -0.797  7.528  40.879  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)      
(Intercept) -73.74133   7.37222 -10.003 < 2e-16 ***  
attr1_1       0.13243   0.09106   1.454  0.14755  
amb1_1        0.33589   0.18212   1.844  0.06671 .  
attr2_1        0.13728   0.09167   1.497  0.13595  
sinc2_1        0.39261   0.14868   2.641  0.00897 **  
fun3_1        -0.85038   0.61511  -1.382  0.16847  
tvspors       -0.52015   0.33304  -1.562  0.12002  
clubbing       0.52691   0.37515   1.405  0.16182  
concerts       1.22302   0.57540   2.126  0.03486 *  
music          -1.35087   0.67770  -1.993  0.04768 *  
yoga           -0.82800   0.37463  -2.210  0.02831 *  
mean_attr_o    9.22319   1.10768   8.327  1.73e-14 ***  
mean_fun_o    3.34230   1.39281   2.400  0.01739 *  
mean_shar_o    6.02965   1.41855   4.251  3.36e-05 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 12.62 on 187 degrees of freedom  
Multiple R-squared:  0.7132,    Adjusted R-squared:  0.6933  
F-statistic: 35.78 on 13 and 187 DF,  p-value: 2.2e-16
```

- 1.
  - 2.
  - 3.
  - 4.
  - 5.
- R squared 0.7, p-value 2.2e-16으로 매우 유의
  - 클럽 가기, 콘서트 가는 취미 등은 유의한 긍정적 회귀계수 : 해당 취미가 긍정적이기 보다는, 이성과 더 대화를 잘 할 것으로 해석 가능
  - 음악 듣기에 대한 유의한 부정적 회귀계수
  - 평균적으로 외모, 재미, 취미공유 점수가 높은 남성이 많은 여성들에게 애프터 신청을 받았다!

# 적합 결과 – 여성(남성이 여성을 볼 때)

```
Call:  
lm(formula = per_par_dec ~ age + tvsports + dining + museums +  
    theater + yoga + mean_attr_o + mean_sinc_o + mean_fun_o +  
    mean_prob, data = reg_data_w)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-31.189 -8.501 -0.507  8.805  45.578  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -25.7281   16.1955 -1.589 0.113823  
age          -0.3521    0.2380 -1.479 0.140730  
tvports      -1.1310    0.3774 -2.996 0.003099 **  
dining        -1.1523    0.6877 -1.676 0.095485 .  
museums       1.0642    0.6015  1.769 0.078440 .  
theater       -0.9753    0.5259 -1.855 0.065202 .  
yoga          -0.6725    0.3715 -1.810 0.071856 .  
mean_attr_o   14.5556    1.2416 11.723 < 2e-16 ***  
mean_sinc_o   -3.3985    1.9533 -1.740 0.083511 .  
mean_fun_o    5.3709    1.5809  3.397 0.000829 ***  
mean_prob     -1.1990    0.6292 -1.906 0.058214 .  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 13.45 on 189 degrees of freedom  
Multiple R-squared:  0.678,    Adjusted R-squared: 0.661  
F-statistic: 39.8 on 10 and 189 DF, p-value: < 2.2e-16
```

1. - R squared 0.661, p-value 2.2e-16으로 매우 유의

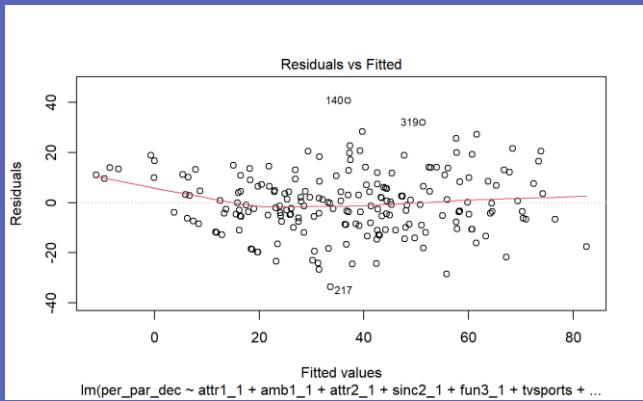
2. - 스포츠를 Tv로 시청하는 취미를 가진 여성은 남성에게 환영받지 못했음

3. 4. 5. - 결국 남성이 가장 많이 보는 것은 외모(매우 큰 유의한 양의 회귀계수), 그리고 재미

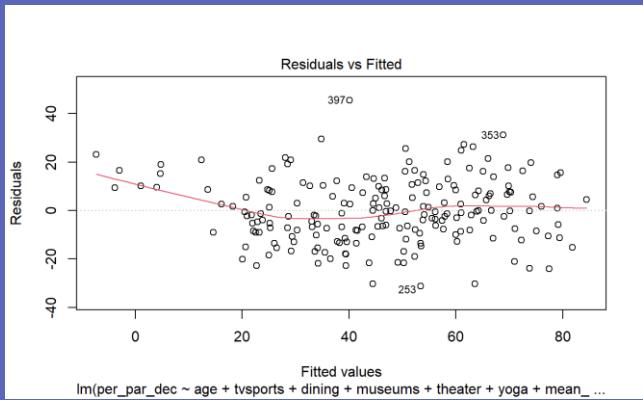
# 회귀 진단

## 1. 선형성

1.  
남



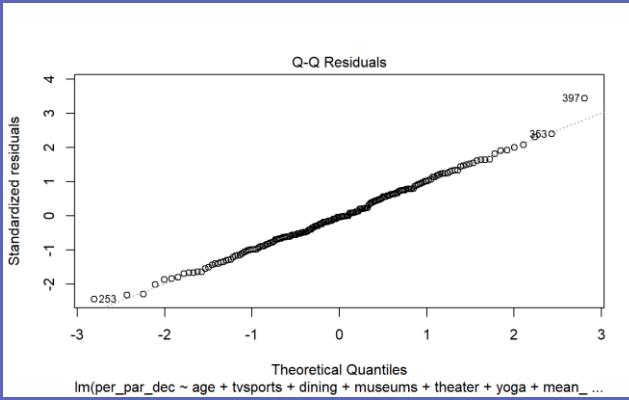
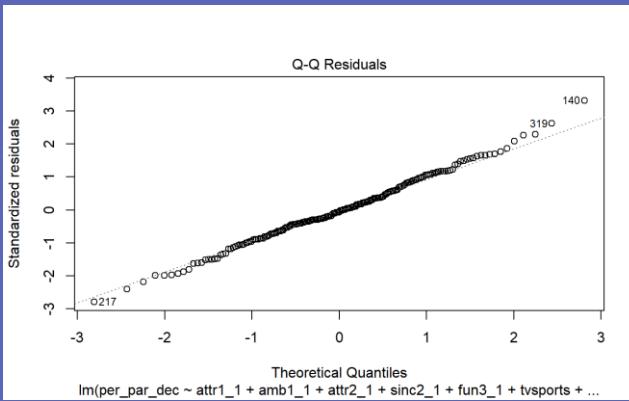
2.



3.

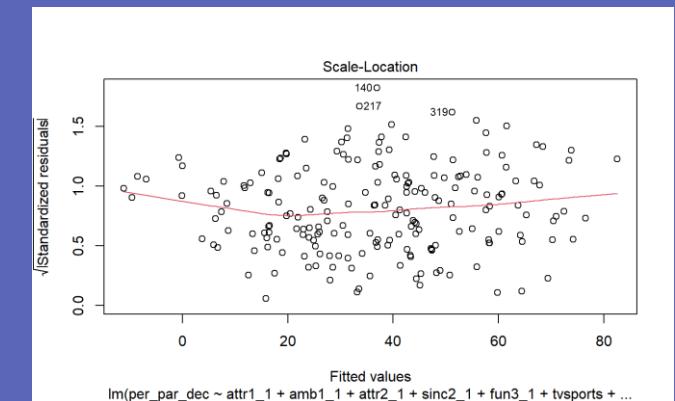
여

## 2. 정규성

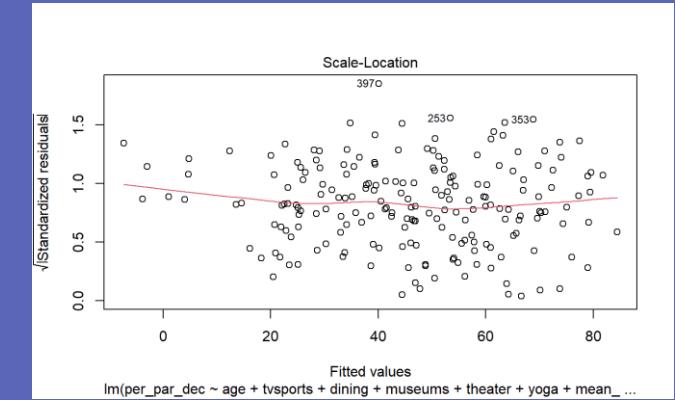


4.

## 3. 등분산성



5.



회귀 가정을 근사적으로 만족하고 있음을 알 수 있다!

# 적합 결과 - 남성(여성이 남성을 볼 때)-실험 후 설문

1.  
2.  
3.  
4.  
5.

```
Call:  
lm(formula = per_par_dec ~ intel3_2 + imprace + hiking + tv +  
    theater + yoga + mean_attr_o + mean_sinc_o + mean_shar_o,  
    data = reg_data_m)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-31.704  -6.999  -0.339   7.404  39.581  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -35.5400  14.4989 -2.451  0.0157 *  
intel3_2      -2.3423  0.9406 -2.490  0.0141 *  
imprace       -0.8267  0.5005 -1.652  0.1012  
hiking         0.9069  0.4657  1.948  0.0538 .  
tv              0.7632  0.5125  1.489  0.1390  
theater        1.1128  0.6091  1.827  0.0701 .  
yoga          -0.9593  0.4779 -2.007  0.0469 *  
mean_attr_o   10.3211  1.3423  7.689  4.22e-12 ***  
mean_sinc_o   -3.6409  1.6562 -2.198  0.0298 *  
mean_shar_o    9.0986  1.7119  5.315  4.89e-07 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 13.15 on 122 degrees of freedom  
Multiple R-squared:  0.6795,    Adjusted R-squared:  0.6558  
F-statistic: 28.74 on 9 and 122 DF,  p-value < 2.2e-16
```

- R squared 0.65, p-value 2.2e-16으로 매우 유의
- 본인이 평가한 지능은 유의한 음의 회귀계수를 지님
- 요가에 대한 유의한 음의 회귀계수: 요가를 좋아할수록 여성에게 선택받는 비율 감소
- 평균적으로 외모, 진실함, 취미공유 점수가 높은 남성이 많은 여성들에게 선택을 받았다!

# 적합 결과 - 여성(남성이 여성을 볼 때) - 실험 후 설문

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-39.4373	25.1575	-1.568	0.120009
attr1_2	0.3617	0.1554	2.328	0.021854 *
sinc1_2	0.6840	0.2614	2.617	0.010189 *
intel1_2	0.6624	0.2245	2.951	0.003917 **
amb1_2	-0.6555	0.3342	-1.961	0.052521 .
amb2_2	0.4704	0.3063	1.536	0.127673
attr3_2	-2.4437	1.1489	-2.127	0.035780 *
amb3_2	-1.3284	0.8304	-1.600	0.112695
go_out	2.5066	1.1823	2.120	0.036376 *
age	-0.7774	0.3247	-2.394	0.018463 *
field_cd	0.6412	0.3079	2.082	0.039757 *
goal	1.1443	0.8584	1.333	0.185428
sports	-1.4628	0.5231	-2.796	0.006159 **
exercise	1.0514	0.5519	1.905	0.059531 .
gamingo	-1.2759	0.5665	-2.252	0.026398 *
tv	-1.2122	0.5378	-2.254	0.026306 *
theater	-0.9222	0.7033	-1.311	0.192657
movies	-1.1408	0.8002	-1.426	0.156981
yoga	-1.3326	0.4592	-2.902	0.004524 **
match_es	1.0492	0.5550	1.890	0.061510 .
mean_attr_o	11.9965	1.6388	7.321	5.41e-11 ***
mean_intel_o	-6.3975	3.1461	-2.033	0.044554 *
mean_fun_o	4.3401	1.8478	2.349	0.020722 *
mean_amb_o	9.5115	2.7335	3.480	0.000735 ***
mean_l1ke	2.5644	1.3636	1.881	0.062816 .
mean_prob	-2.5826	1.1191	-2.308	0.022990 *
---				

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.42 on 104 degrees of freedom

Multiple R-squared: 0.7679, Adjusted R-squared: 0.7121

F-statistic: 13.77 on 25 and 104 DF, p-value < 2.2e-16

- R squared 0.7121, p-value 2.2e-16으로 매우 유의

-유의한 음의 회귀계수: 스스로 평가한 외모, 나이, 운동 취미, 게임 취미, tv시청 등

-유의한 양의 회귀계수 : 외출 빈도, 외모 점수, 재미 점수 등

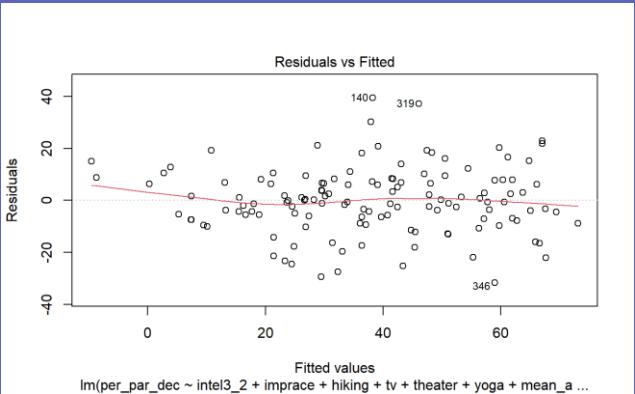
-결국 남성이 가장 많이 보는 것은 외모(매우 큰 유의한 양의 회귀계수)

-야망 점수를 높게 받을수록 선택받는 비율이 높았다

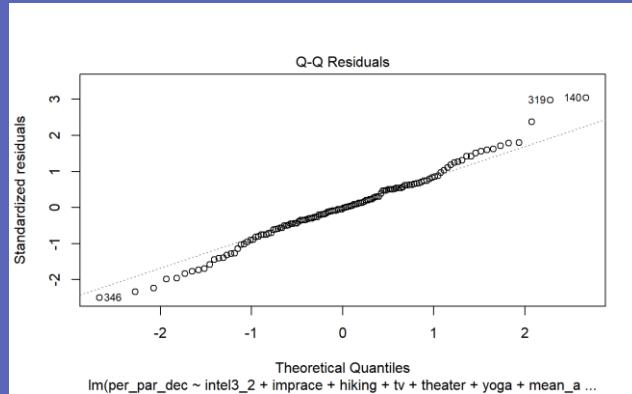
전반적으로 여성에 비해 남성의 선택에 영향을 주는 요소가 더 다양하다!

# 회귀 진단 -실험 후

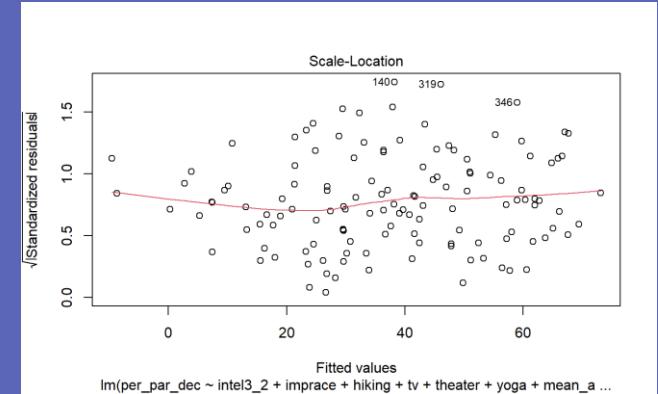
## 1. 선형성



## 2. 정규성



## 3. 등분산성



1.

남

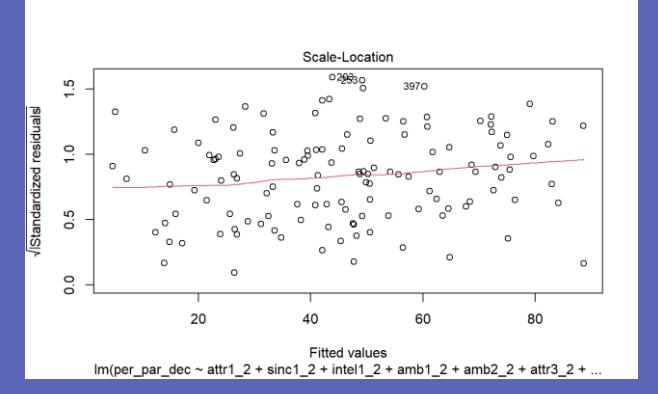
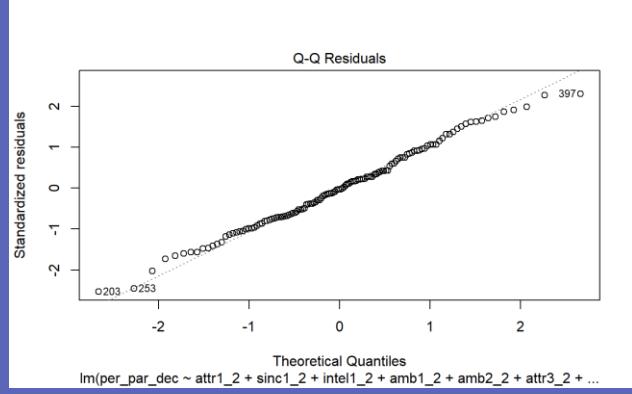
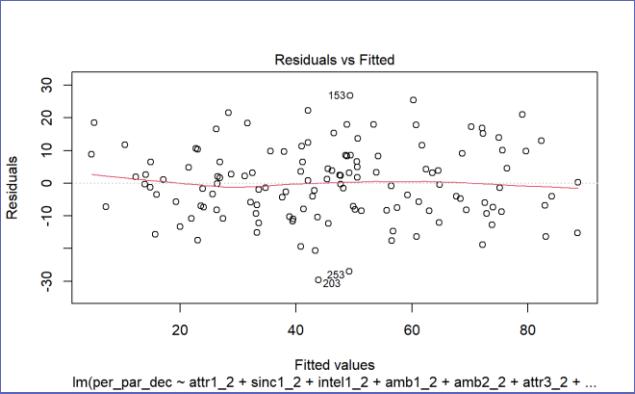
2.

3.

여

4.

5.



회귀 가정을 근사적으로 만족하고 있음을 알 수 있다!

# 적합 결과 정리

-설문조사 결과를 포함시켰을 때 모형의 설명력이 매우 크게 상승

-전반적인 결과의 기조는 일치

1.

-인기있는 남자와 여자의 가장 중요한 조건:외모

2.

-여성의 경우 외모와 비슷한 관심사를 가진 것이 매력적인 남성의 가장 큰 조건

3.

-남성의 경우 다양한 요소를 고려하지만, 외모와 재미, 야망을 우선적으로 고려함

4.

-남성의 경우 야망이 높은 것은 당장의 데이트 결정에 부정적인 영향으로 보였으나, 실제로 다른 변수들과 함께 적합 시켰을 때 야망이 높은 여성이 더욱 선택받았음을 알 수 있음

5.

-설문조사 항목들의 유의한 결과는 대체적으로 매력적인 여성들이 해당 항목을 높게 투표했을 것으로 추정 가능

데이터 소개 및 목적

데이터 전처리

1. 설문 데이터 EDA

2. 데이트 내에서의 결정 영향변수

3. 인기 있는 사람의 비결

4. 자기 객관화의 중요성

5. 결론 및 제언

## 4. 자기 객관화의 중요성

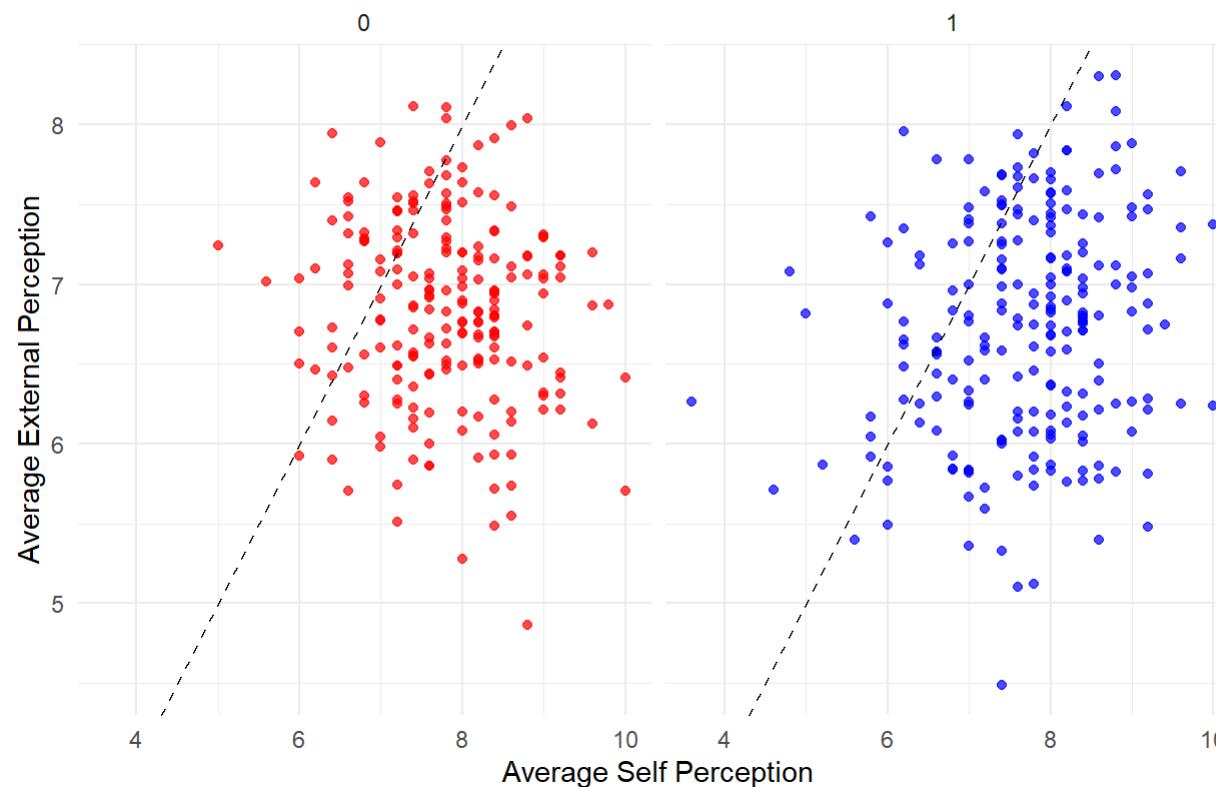
- 본인이 바라본 자신과 이성이 바라본 자신은 얼마나 다른가?
- 본인에 대한 과대평가가 이성의 결정에 미치는 영향
- 스피드 데이팅 전후 자기 객관화의 차이

# 과연 사람들은 자기 객관화가 잘 되어 있을까?

1. 본인 스스로를 평가한 점수와 실제 데이팅에서 파트너들이 평가한 점수의 차이를 변수화  
5개 항목(외모, 진실됨, 지능, 재미, 야망)에 대한 변수 제작(결측치 약 6개 제거)
- 2.
3. 성별마다 평균적인 객관화 정도의 차이 비교  
남성과 여성으로 그룹을 나누고, 각각에 대해 MANOVA 실시 및 가정 검정, Bonferroni CI 계산해 사후 검정
- 4.
5. 이성의 결정에 자기객관화 정도가 영향을 미치는지 분석  
회귀분석 사용, 이성의 승낙 퍼센트를 종속변수로 두고 각 항목별 차이를 설명변수로 설정

# 성별 간 객관화 정도

Average Self vs. External Perceptions by Gender



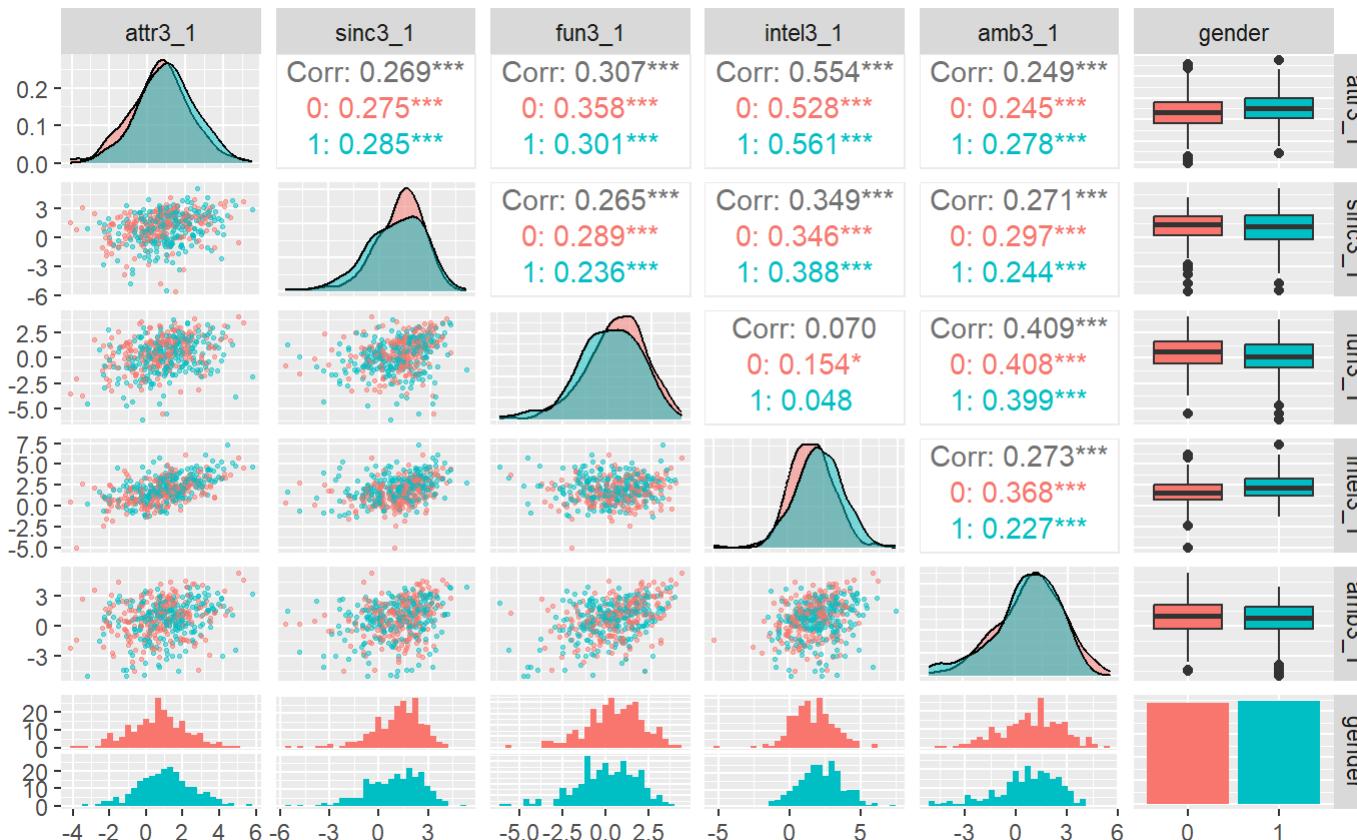
X: 본인이 스스로를 평가한 항목들의 평균  
Y: 데이팅에서 실제로 만난 이성들이 평가한 항목의 평균

gender  
• 0  
• 1

전반적으로 남녀 모두 상대적으로 스스로를 높게 평가하고 있음

# 항목별 차이

1.  
2.  
3.  
4.  
5.

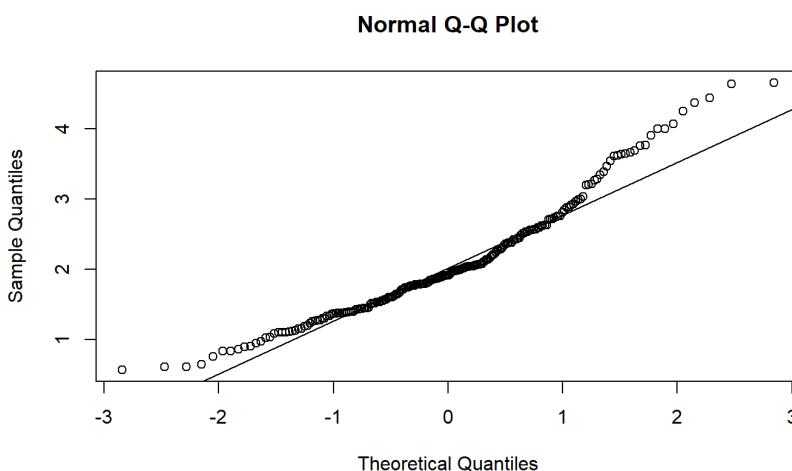


<항목별 gap의 분포: 성별에 따라>

- 항목별 gap 변수 : 외부 평가 - 본인평가
  - 전반적으로 정규분포의 형태를 띠고 있음
  - 항목 간 큰 상관관계는 없는 것으로 보임
  - 평균이 우측으로 치우쳐져 있음:  
과대평가가 일어나고 있다
- => 통계적 분석 필요

# MANOVA-가정 확인

## 1. 정규성 검정-남성

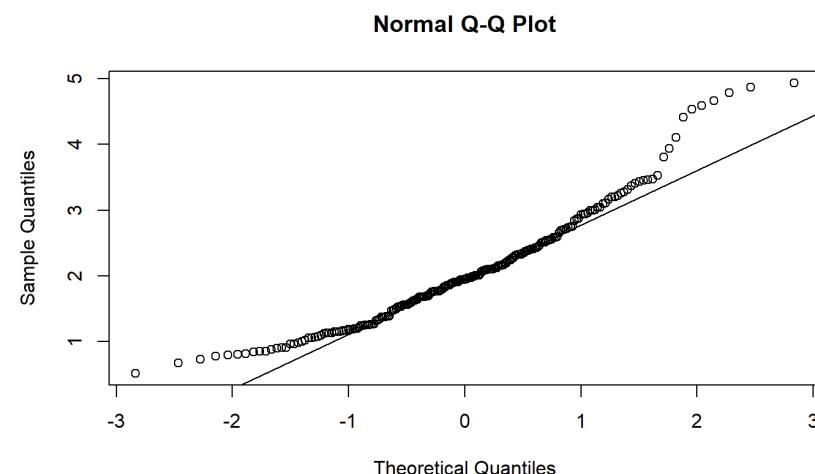


Variable	P-value	Normality
Multivariate	2.338554e-06	No
Attr_gap	0.9883	Yes
Sinc_gap	0.0044	No
Intel_gap	0.0387	No
Fun_gap	0.8752	Yes
Amb_gap	<0.001	No

Multivariate normal에 가깝지만, 변환 필요!

# MANOVA-가정 확인

## 1. 정규성 검정-여성



Variable	P-value	Normality
Multivariate	0	No
Attr_gap	0.4964	Yes
Sinc_gap	<0.001	Yes
Intel_gap	0.0815	No
Fun_gap	0.5028	Yes
Amb_gap	0.1233	Yes

선형에 가깝지만, 적절한 변환 필요함!

# MANOVA-가정 확인

## \* Yeo-Johnson 변환 적용

$$y_i^{(\lambda)} = \begin{cases} ((y_i + 1)^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log(y_i + 1) & \text{if } \lambda = 0, y \geq 0 \\ -\frac{(-y_i + 1)^{2-\lambda} - 1}{2-\lambda} & \text{if } \lambda \neq 2, y < 0 \\ -\log(y_i + 1) & \text{if } \lambda = 2, y < 0 \end{cases}$$



-실수 전체 구간에서 정의된 변수에 대해 사용 가능

-Gap 변수는 양과 음 모두 가지고 있기 때문에 Yeo-Johnson 적용!

<남>			<여>		
Variable	P-value	Normality	Variable	P-value	Normality
Multivariate	0.1015034	Yes	Multivariate	2.518763e-05	No
Attr_gap	0.9883	Yes	Attr_gap	0.6458	Yes
Sinc_gap	0.2443	Yes	Sinc_gap	0.8629	Yes
Intel_gap	0.3882	Yes	Intel_gap	0.9488	Yes
Fun_gap	0.8020	Yes	Fun_gap	0.2954	Yes
Amb_gap	0.7984	Yes	Amb_gap	0.9024	Yes

정규성 확보!

# MANOVA-결과

## <MANOVA Result>

Var	df	Pillai	F	Num_Df	Den_df	Pr(>F)
Gender	1	0.077543	7.3133	5	435	1.353e-06
Residuals	439					

유의수준 5% 하 귀무가설 기각 남성 집단과 여성 집단 간 유의한 Gap의 차이가 존재한다!

# Bonferroni interval

## <Bonferroni Interval>

	변수	Lower_ci	Upper_ci
1.	Attractive	-0.771089375	-0.005020311
2.	Sincere	-0.1953552	0.6059984
3.	Funny	-0.9548967	-0.2082906
4.	Intelligent	0.01895008	0.85784481
5.	Ambitious	-0.1708693	0.7523608

- Attractive, Funny, Intelligent의 신뢰구간이 0을 포함하지 않으므로 유의한 차이가 있음!
- 남자의 경우 여자보다 본인의 **외모, 유머감각을** 유의하게 과대평가한다
  - 여자의 경우 남자보다 본인의 **지능을** 유의하게 과대평가한다!

# Regression Model - 남성

여성에게 선택받은 비율을 종속변수로 설정, 각 변수의 gap을 설명변수로 regression model fitting

1.

```
Call:  
lm(formula = dec ~ . - match - dec2, data = d)
```

2.

```
Residuals:  
    Min      1Q  Median      3Q     Max  
-50.295 -12.303 -1.600  8.665  62.982
```

3.

```
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 53.9475   2.1870  24.667 < 2e-16 ***  
attr_gap    -4.3620   1.0592  -4.118 5.43e-05 ***  
sinc_gap    -0.6298   0.8074  -0.780  0.43626  
intel_gap    2.6404   0.8206  3.218  0.00149 **  
fun_gap     -5.4988   1.0504  -5.235 3.89e-07 ***  
amb_gap     1.2544   0.7413   1.692  0.09206 .
```

4.

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.

```
Residual standard error: 18.66 on 217 degrees of freedom  
Multiple R-squared:  0.3515,    Adjusted R-squared:  0.3366  
F-statistic: 23.53 on 5 and 217 DF,  p-value: < 2.2e-16
```

- p-value 2.2e-16으로 유의한 회귀모형!

- attr\_gap, intel\_gap, fun\_gap이 유의한 회귀계수를 지님

본인의 외모에 대한 과대평가가 심할수록 선택받은 비율이 감소!

본인의 지능에 대한 과대평가가 심할수록 선택받은 비율이 증가!

본인의 재미에 대한 과대평가가 심할수록 선택받은 비율이 감소!

-> 특히 외모와 재미에 대한 객관화가 중요하다!

<Fitting 된 회귀모델>

# Regression Model - 여성

남성에게 선택받은 비율을 종속변수로 설정, 각 변수의 gap을 설명변수로 regression model fitting

1.

```
Call:  
lm(formula = dec ~ . - match - dec2, data = d2)
```

2.

```
Residuals:  
    Min      1Q  Median      3Q     Max  
-40.698 -12.722 -0.174  12.463  49.816
```

3.

```
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 57.7656   2.0455  28.240 < 2e-16 ***  
attr_gap    -3.8261   1.0261  -3.729 0.000247 ***  
sinc_gap     0.8986   0.9597  0.936 0.350183  
intel_gap    0.7894   0.9391  0.841 0.401519  
fun_gap      -5.5118   1.1249  -4.900 1.9e-06 ***  
amb_gap     -1.7285   0.8305  -2.081 0.038616 *
```

4.

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.

```
Residual standard error: 19.2 on 212 degrees of freedom  
Multiple R-squared:  0.3235, Adjusted R-squared:  0.3076  
F-statistic: 20.28 on 5 and 212 DF, p-value: < 2.2e-16
```

- p-value 2.2e-16으로 유의한 회귀모형!

- attr\_gap, amb\_gap, fun\_gap이 유의한 회귀계수를 지님

본인의 외모에 대한 과대평가가 심할수록 선택받은 비율이 감소!

본인의 야망에 대한 과대평가가 심할수록 선택받은 비율이 감소!

본인의 재미에 대한 과대평가가 심할수록 선택받은 비율이 감소!

-> 특히 외모와 재미에 대한 객관화가 중요하다!

데이터 소개 및 목적

데이터 전처리

1. 설문 데이터 EDA

2. 데이터 내에서의 결정 영향변수

3. 인기 있는 사람의 비결

4. 자기 객관화의 중요성

5. 결론 및 제언

## 결론 및 제언

- 분석 결과 요약

- 한계점 및 제언

# 분석 결과

1.

## 데이팅 진행중 이성 선택에 주요한 요소는?

- 두 성별 모두 데이팅 진행 중 가장 많이 고려한 요소는 외모로, 이는 특히 남성에서 더 두드러진다.
- 그 뒤를 동일하게 유머와 관심사 공유가 따르고 진실성과 야망 변수는 부정적 영향을 미친다.
- 성별 간 다른 점은 지성 변수의 경우 여성이 남성에 대한 결정을 내릴 때 반대의 경우와 달리 영향을 미친다는 것이다.

2.

3.

4.

5.

## 성별간 비교

- 남성은 여성에 비해 스스로의 외모와 재미를 과대평가하는 경향 존재.
- 여성은 남성에 비해 스스로의 지능을 과대평가하는 경향 존재.

## 인기 있는 사람의 비밀

- 인기 있는 남성은 외모가 뛰어나고, 여성들에게 관심사를 잘 맞춰 나가는 사람.
- 인기 있는 여성은 외모가 뛰어나고, 재미있으며 나이가 적은 사람. 자신의 삶에 대한 포부(야망)이 크면 더욱 좋다.
- 하나의 깊은 취미를 가지는 것보다 두루두루 적당한 관심을 가지는 것이 낫다.

## 자기 객관화의 중요성

- 남녀 모두 스스로의 외모와 재미에 대한 객관화되어 있지 않으면 이성에게 매력적이지 못하다.
- 남성은 데이트 나가기 전 어느 정도 스스로 똑똑하다고 생각하고 나가는 것이 좋을 수도 있다.

# 한계점 및 제언

---

## 한계점

1. - 남녀간 일상적인 만남이 아닌 특수한 상황(4분 데이팅)에서의 데이팅이라 일반적인 경우로 확장하기엔 한계가 존재.
2. - 스스로 점수를 매긴 설문에 기반하기 때문에 실제 사람들의 생각 및 감정과 다를 수 있는 가능성 존재.
- 3.
- 4.

## 제언

5. - 분석 결과에서 볼 수 있듯이 남성과 여성 간의 중요하게 생각하는 것이 다름을 인지할 수 있어야 하고 이성 선택을 받기 위해서는 노력이 필요.
- 4분 데이팅이 아닌 조금 더 긴 시간의 데이팅에 대해 분석할 기회가 있다면 사람들의 실제 상호작용과 더 비슷한 데이터 및 분석을 얻을 수 있을 것이라 기대.

# Thank you!

# Q&A