DACA

NGO 데이터 분석 : 1주차

2020.10.22 ~ 2020.10.28

빅데이터 응용학과 석사 1기 양선욱

목 차

1. 기초통계분석

1-1. 통계 그래프

(히스토그램, 산점도, 파이차트, 상자그림)

1-2. 기술통계분석

(평균, 분산, 표준편차, 왜도, 첨도 등)

2. 상관관계분석

2-1. 수치형 변수의 상관관계분석

2-2. 편(부분) 상관관계 분석

2-3. 순서형 변수의 상관관계분석

3. t-검정

3-1. 일표본 t-검정

3-2. 독립표본 t-검정

3-3. 쌍체표본 t-검정

4. 범주형 데이터 분석

4-1. 적합도 검정

4-2. 독립성 검정

4-3. 동질성 검정

1. 기초통계분석 – 1.히스토그램

사용 데이터

- AGE : 나이
- SEX : 성별

분석 목적

성별에 따른 나이대를 알아보기 위해 히스토그램 그래프를 그려보았다.

분석 과정

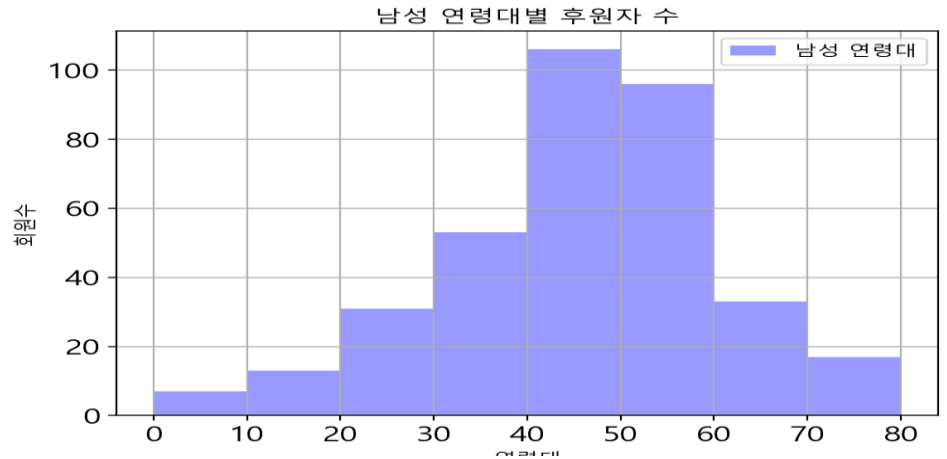
1) 이상값 처리
2) 파생변수 만들기
3) 히스토그램 작성-남성
3-1) 각종 옵션설정
3-2)이미지 저장
4) 히스토그램 작성-여성
4-1) 각종 옵션설정
4-2) 이미지저장

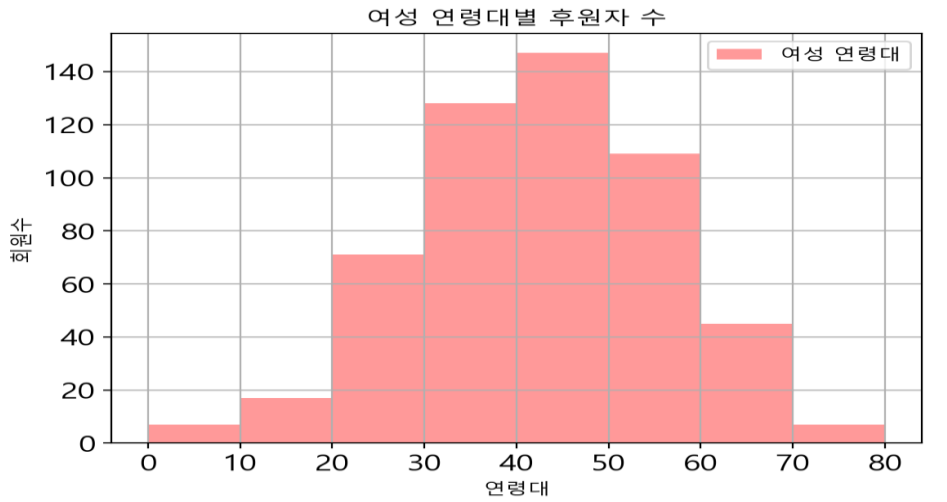
핵심 코드

```
#2-1 이상값 처리 - AGE
dropin = df[df['AGE']==0].index

#4. 히스토그램 작성-남성
plt.hist(man['AGE'], alpha=0.4,bins=np.arange(0,90,10), rwidth=1,
color='blue', label='남성 연령대')
from pylab import figure, axes, pie, title, savefig
plt.savefig('1.hist남성연령대후원자.png',dpi=200,edgecolor='blue',
bbox_inches='tight',pad_inches=0.3)
plt.clf()

plt.hist(woman['AGE'], alpha=0.4, bins=np.arange(0,90,10),
rwidth=1, color='red', label='여성 연령대')
```

1. 기초통계분석 – 1.히스토그램

결과





결과 해석

남성은 40,50대가 압도적으로 높았고, 여성은 30,40,50대가 많은 편이다.

실무적 인사이트

남성 후원자의 인구분포는 40,50대라는 특정 연령대에 집중되어 있다. 이들이 가입을 한지 시간이 흘러 40,50대가 된 것인지, 신규 가입자도 40,50대가 많은지에 대한 추가적인 분석이 필요하다.

여성 후원자는 30~50대까지 고르고 높은 분포를 보여주고 있다. 또한 20대 후원자가 남성보다 많은 점은 미래지향적 가치가 있다고 여겨진다.

1. 기초통계분석 – 1. 산점도

사용 데이터

- AGE : 나이
- LONGEVITY_D : 가입일수(일)
- PAY_NUM_REGULAR : 납입 정기 후원 횟수
- PAY_NUM_ONETIME : 납입 일시 후원 횟수
- PAY_NUM 납입 후원 횟수

분석 목적

고객의 어떠한 요소가 납입 후원 횟수가 상관이 있는지 알아보기 위해 산점도 그래프를 사용하였다. 추가적으로 정기 후원과 일시 후원에 대한 상관관계도 알아보려고 한다.

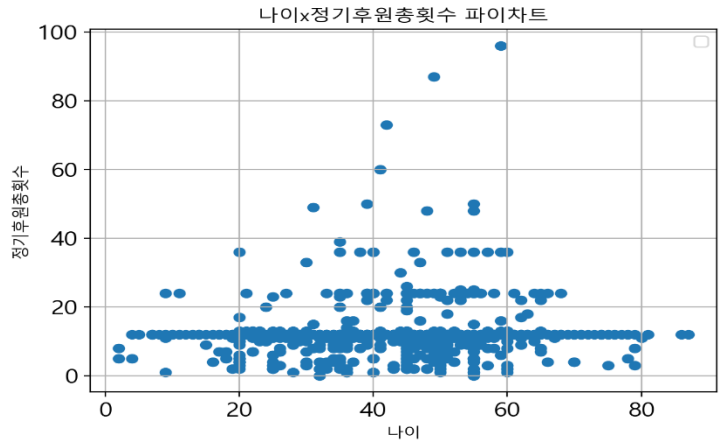
분석 과정

1) 이상값 처리
2) 연령x후원횟수 산점도
3) 가입일수(일)x후원횟수 산점도
4) 정기후원횟수x일시후원횟수 산점도

핵심 코드

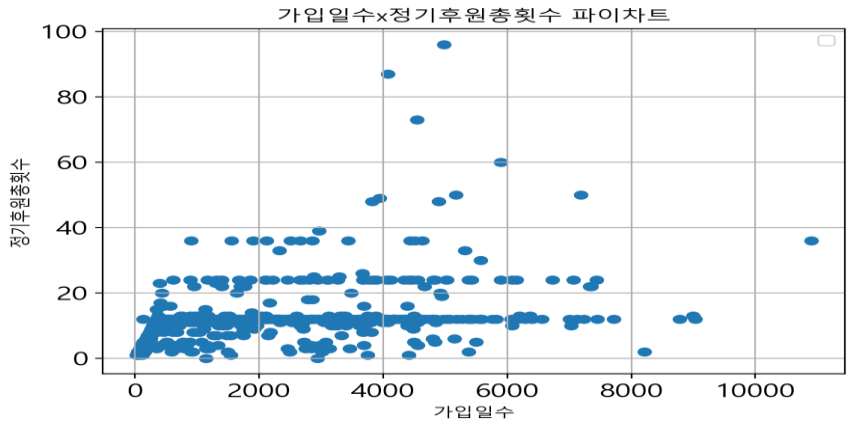
```
#3-1-1. 연령x후원횟수 산점도
plt.scatter(dfin['AGE'], dfin['PAY_NUM'])
#3-2. 가입일수(일)x후원횟수 산점도
plt.scatter(df_1['LONGEVITY_D'], df_1['PAY_NUM'])
#3-3. 정기후원횟수x일시후원횟수 산점도
plt.scatter(df['PAY_NUM_REGULAR'], df['PAY_NUM_ONETIME'])
```

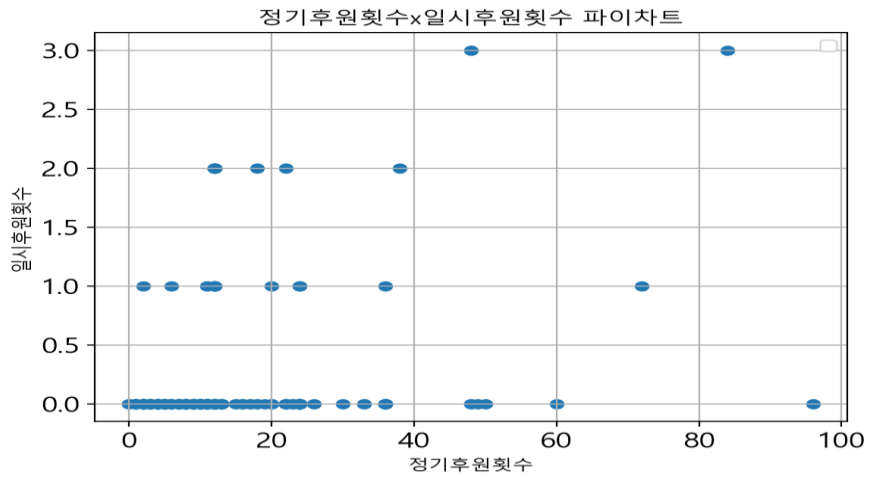
분석 결과



1. 기초통계분석 – 1.산점도

분석 결과





결과 해석

1) 고령이라고 정기후원 횟수가 많지 않았다. 오히려 청년층 때부터 꾸준히 후원을 해온 40대가 정기후원 횟수가 가장 많았다.

2) 가입 일수가 많다고 후원횟수가 많지도 않았다.

3) 정기후원과 일시후원은 유의미한 상관관계가 없었다. 오히려 일시후원은 그 수가 매우 적다는 것이 보여진다.

실무적 인사이트

연령대에 따른 정기후원 횟수의 차이가 보였다. 따라서 연령대에 따른 정기후원 횟수의 비율 등을 수치적으로 분석해야 한다.

1. 기초통계분석 – 1.파이차트

사용 데이터

- AGE : 나이
- SEX : 성별

분석 목적

성별에 따른 후원자들의 연령대가 어떠한지 시각적으로 알아보기 위해 파이차트 그래프를 그려보았다.

정기적인 후원을 위한 경제적 안정성 때문에 40대 이상의 회원이 많을 것으로 예측된다.

분석 과정

1) 이상값 처리 - AGE
2) 파생변수 만들기
3) 연령대, 성별 구분
4) 파이차트 만들기 - 연령대
5) 파이차트 속성 지정
6) 이미지 저장

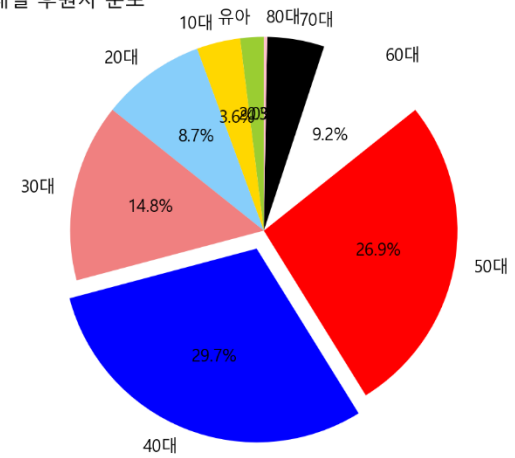
핵심 코드

```
groupby_연령대 = df_1.groupby('연령대').count()
groupby_연령대 = groupby_연령대['AGE']
labels = ['유아','10대','20대','30대','40대','50대','60대','70대','80대']
color = ['yellowgreen', 'gold', 'lightskyblue', 'lightcoral', 'blue', 'red', 'white', 'black','pink']
explode = (0,0,0,0,0.1,0,0,0,0)
size = groupby_연령대.values
plt.pie(size, explode = explode, labels = labels, colors = color, autopct='%1.1f%%', shadow = False, startangle=90)
plt.axis('equal')
plt.title('연령대별 후원자 분포',position=(0.1, 3))
```

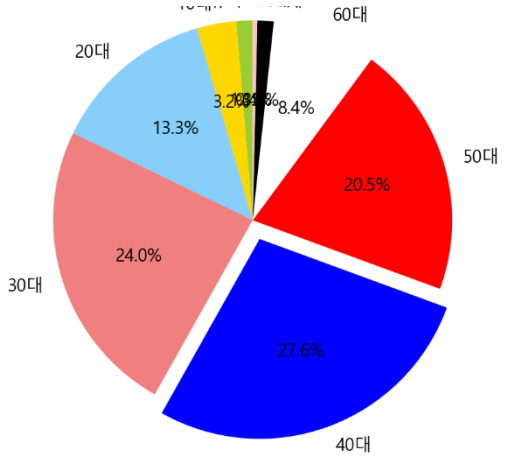
1. 기초통계분석 – 1. 파이차트

분석 결과

남성 연령대별 후원자 분포



여성 연령대별 후원자 분포



결과 해석

남성, 여성 모두 40, 50대가 전체의 55%에 달한다. 특이하게도 여성은 남성보다 높은 30대의 비율을 보여주고 있다. 오히려 남성은 50대의 비율이 여성보다 높은 편이다.

전반적으로 20,30대 여성들은 또래의 남성들보다 NGO후원에 관심이 많은 편이다.

실무적 인사이트

히스토그램에서 확인했던 세대간 비율의 격차를 파이차트에서 확인하였다.

남성은 미래의 장기고객인 20, 30대의 비율이 왜 적은지에 대해 추가적인 분석이 필요해 보인다.

여성은 남성에 비해 연령대의 비율이 고른편이다. 그만큼 연령대별 맞춤형 전략이 중요하므로 다양한 연령대별 성향에 추가적인 분석이 필요하다.

1. 기초통계분석 – 1. 상자그림

사용 데이터

- SEX : 성별
- PAY_SUM_PAYMENTAMOUNT : 납입총후원금액

분석 목적

- NGO 사업의 특성상 후원 금액이 정해져 있지 않다.
- 후원금액만 가지고 성별의 소비 금액의 차이를 논하기가 난해하다.
- 성별에 따른 납입총후원금액의 이상값이 있는지 알아보기 위해 상자그림 분석을 수행하였다.

분석 과정

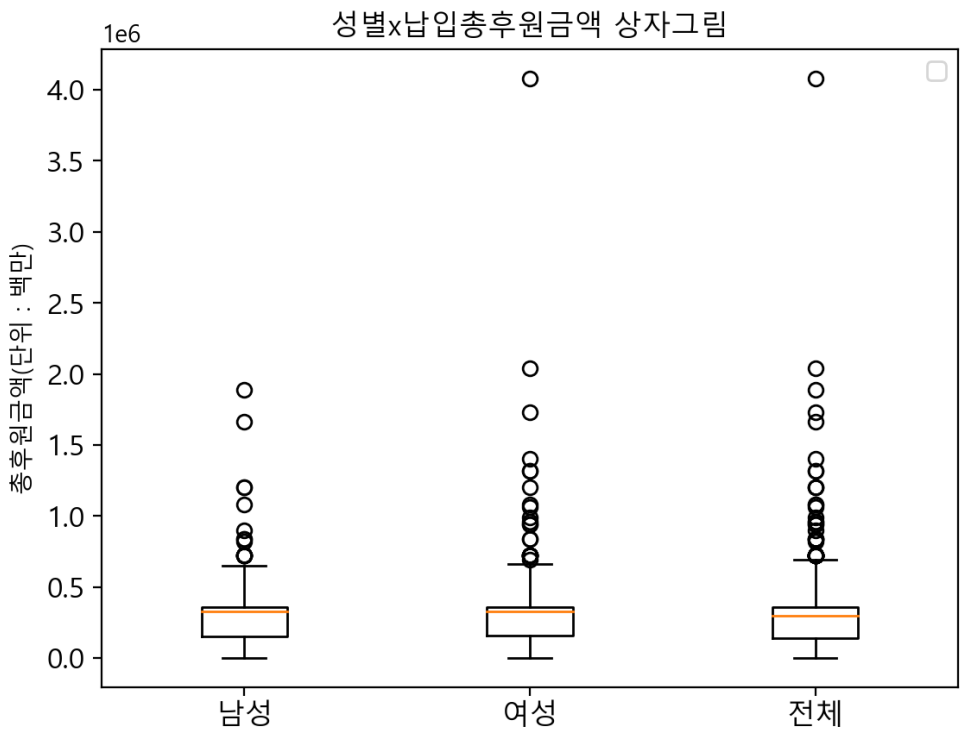
1) 파생변수 나누기
2) 상자그림 작성하기
3) 이미지 저장

핵심 코드

```
plt.boxplot([man_pay,woman_pay,both_pay],
labels=['남성', '여성','전체'])
plt.title("성별x납입총후원금액 상자그림")
plt.legend()
plt.ylabel('총후원금액(단위 : 백만)')
plt.xticks(fontsize = 12)
plt.yticks(fontsize = 12)
```

1. 기초통계분석 – 1.상자그림

분석 결과



결과 해석

남녀 모두 비슷한 평균, 비슷한 분포, 비슷한 이상값의 빈도를 보여준다. 대부분의 총후원금액은 40만원이하이며 전반적으로 데이터의 분포가 아래로 쏠려있다.

실무적 인사이트

상위 일부의 후원금액과 다수의 후원금액 간에 격차가 크다. 이러한 이상값들이 있다는 것을 염두하고 앞으로의 분석을 진행해야 한다.

1. 기초통계분석 – 2. 기술통계분석

사용 데이터

PAY_SUM_PAYMENT_R_CS_INT : 정기해외아동 납입금액

PAY_SUM_PAYMENT_R_CS_DOM : 정기국내아동 납입금액

PAY_SUM_PAYMENT_R_PN_INT : 정기해외사업 납입금액

PAY_SUM_PAYMENT_R_PN_FST : 정기긴급사업 납입금액

PAY_SUM_PAYMENT_R_PN_DOM : 정기국내사업 납입금액

PAY_SUM_PAYMENT_R_PN_NKOR : 정기북한사업 납입금액

PAY_SUM_PAYMENT_R_PN_ALL :정기전체사업 납입금액

분석 목적

- 본 사의 후원 항목은 크게 6가지로 구분된다.

- 전략팀에서는 6가지 항목에 대한 정기후원 맞춤형 전략을 세우기 위해 6가지 항목에 대한 기초통계분석을 우선적으로 의뢰하였다

분석 과정

1) 결측값 제거

2) 0인 데이터 제거

3) 기술통계분석 실시

4) csv파일로 추출

핵심 코드

```
for i in range (0,size_row,1) :
  resultDF.iloc[[i],[0]] = rows_list[i]
  for i in range(1,size_col,1) :
    colname = col[i-1]
    df_2 = df_1[df_1[colname]!=0]
    resultDF.iloc[[0],[i]] = round(df_1[colname].mean(),2)
    resultDF.iloc[[1],[i]] = round(df_1[colname].var(),2)
    resultDF.iloc[[2],[i]] = round(df_1[colname].std(),2)
    resultDF.iloc[[3],[i]] = round(df_1[colname].skew(),2)
    resultDF.iloc[[4],[i]] = round(df_1[colname].kurt(),2)
```

1. 기초통계분석 – 2.기술통계분석

분석 결과

	기술통계	정기해외아동	정기국내아동	정기해외사업	정기긴급사업	정기국내사업	정기북한사업	정기전체사업
0	평균	329204.95	483194.44	155300	179300	184367.09	93333.33	155853.66
1	분산	19433880046	1.93752E+11	11824908163	9997969388	15210491723	906666666.7	6440371091
2	표준편차	139405.45	440172.27	108742.39	99989.85	123330.82	30110.91	80251.92
3	왜도	1.96	5.15	1.04	0.41	1.43	-0.21	0.55
4	첨도	11.8	35.67	-0.35	-1.03	4.51	-2.83	0.63
5	사분위수1	240000	240000	102500	120000	110000	65000	120000
6	사분위수2	360000	425000	120000	120000	200000	100000	120000
7	사분위수3	360000	600000	210000	240000	240000	120000	240000
8	사분위수4	360000	600000	120000	120000	240000	120000	120000

결과 해석

가장 평균이 높은 항목은 국내아동이며 가장 평균이 낮은 항목은 북한사업이다. 전반적으로 해외아동, 국내아동에 대한 비율이 높은 편이다.

아동 후원에 대해 집중을 할지, 사업 4분야 후원에 대한 홍보에 집중을 할지 선택하기 위해 추가적인 분석이 필요해 보인다.

실무적 인사이트

- 홍보가 해외, 국내아동 분야에 집중되어 있는지 확인한다.

- 각 후원 분야의 평균후원금에 대해서도 추가적인 분석이 필요해 보인다.

2. 상관관계분석 – 1. 수치형 변수의 상관관계분석

사용 데이터

- 'AGE' : 나이
- 'LONGEVITY_D' 가입일 수,
- 'PLED_FIRST_LONGEVITY' 첫플릿지 기준 고객기간,
- 'PAY_NUM' 납입 횟수
- 'PAY_SUM_PAYMENTAMOUNT' 납입총후원금액
- 'PLED_NUM' 전체 플릿지 수,
- 'PLED_RATE_FULFILLED' 납부 완료 플릿지 비율,
- 'MOTI_NUM_CHANNEL 개발 채널 수'
- 최초후원까지DAY = 가입일 수 – 첫플릿지 기준 고객기간
- 평균후원금액 = 납입총후원금액 / 납입횟수

분석 목적

기획 전략팀은 어떠한 요소가 회원들의 후원을 유도하는지 알아보기 위해 의심가는 모든 요소에 대해 상관관계 분석을 의뢰하였다.

분석 과정

1) 결측값 제거
2) 파생데이터 생성
3) 상관관계매트릭스 생성
4) 상관계수 추출 함수

핵심 코드

```
correlationMatrix = df1.corr()
correlationMatrix.to_csv('3. 수 치 형 상 관 관 계 .csv',
encoding='utf-8-sig')
def high_correlated_p(x, cutoff):
    index_list = []
    for i in range(0, len(x)):
        for j in range(i+1, len(x)):
            if x.iloc[i][j] >= cutoff:
                index_list.append([x.columns[i], x.columns[j]])
print(high_correlated_p(correlationMatrix, 0.5))
```

2. 상관관계분석 – 1. 수치형 변수의 상관관계분석

분석 결과

	AGE	LONGEVITY_D	PLED_FIRST_LONG	PAY_NUM	PAY_SUM_PAYMEN	PLED_NUM	PLED_RATE_FULFILL	MOTI_NUM_CHAN	평균후원금액	최초후원까지DAY
AGE	1	0.117913917	0.199851349	0.102732	0.09446754	0.060000302	0.041286733	0.031398644	0.000205456	0.023480559
LONGEVITY_D	0.117913917	1	0.593233143	0.160931	0.110391365	0.195243666	0.015769633	0.210978593	-0.037747334	0.871127861
PLED_FIRST_LONGEVITY	0.199851349	0.593233143	1	0.323605	0.235148113	0.352737858	0.029059868	0.301936146	-0.051946816	0.121466552
PAY_NUM	0.1027323	0.160930724	0.323605392	1	0.815833508	0.633397302	0.054801394	0.339845658	0.012826022	0.00103184
PAY_SUM_PAYMENTAMOUNT	0.09446754	0.110391365	0.235148113	0.815834	1	0.582736719	0.041631528	0.264198457	0.478538675	-0.00732516
PLED_NUM	0.060000302	0.195243666	0.352737858	0.633397	0.582736719	1	-0.133896611	0.620082626	0.065585135	0.025569032
PLED_RATE_FULFILLED	0.041286733	0.015769633	0.029059868	0.054801	0.041631528	-0.133896611	1	-0.184796468	-0.023113383	0.001717744
MOTI_NUM_CHANNEL	0.031398644	0.210978593	0.301936146	0.339846	0.264198457	0.620082626	-0.184796468	1	-0.003235851	0.075958312
평균후원금액	0.000205456	-0.037747334	-0.051946816	0.012826	0.478538675	0.065585135	-0.023113383	-0.003235851	1	-0.014861249
최초후원까지DAY	0.023480559	0.871127861	0.121466552	0.001032	-0.00732516	0.025569032	0.001717744	0.075958312	-0.014861249	1

결과 해석

상관계수 0.5 이상

[['LONGEVITY_D', 'PLED_FIRST_LONGEVITY'],

['LONGEVITY_D', '최초후원까지DAY'],

['PAY_NUM', 'PAY_SUM_PAYMENTAMOUNT'],

['PAY_NUM', 'PLED_NUM'],

['PAY_SUM_PAYMENTAMOUNT', 'PLED_NUM'],

['PLED_NUM', 'MOTI_NUM_CHANNEL']]

실무적 인사이트

상관관계가 있다고 나온 6가지의 관계에 대해 편상관 분석을 수행하여 각 변수들의 고유한 상관관계를 밝혀야 한다.

2. 상관관계분석 – 2. 편상관관계 분석

사용 데이터

- 'LONGEVITY_D' 가입일 수,
- 'PLED_FIRST_LONGEVITY' 첫플릿지 기준 고객기간,
- 'PAY_NUM' 납입 횟수
- 'PAY_SUM_PAYMENTAMOUNT' 납입총후원금액
- 'PLED_NUM' 전체 플릿지 수,
- 'MOTI_NUM_CHANNEL 개발 채널 수'
- 최초후원까지DAY = 가입일 수 – 첫플릿지 기준 고객기간
- 평균후원금액 = 납입총후원금액 / 납입횟수

분석 목적

앞선 상관관계 분석에서 다른 변수가 영향을 주는 것으로 의심되는 상관관계가 있어서 편상관관계 분석을 시행하였다.

분석 과정

1) 결측값 제거
2) 파생데이터 생성
3) 편상관관계분석 수행
4) 상관계수 추출 함수

핵심 코드

```
print('1)가입일수x첫플릿지 기준 고객 기간, 최초후원까지 DAY : ')
print(partial_corr(data=df1, x='LONGEVITY_D', y='PLED_FIRST_LONGEVITY', covar='최초후원까지 DAY'),'\n')
print('2)가입일수x최초후원까지DAY, 첫플릿지 기준 고객 기간 : ')
print(partial_corr(data=df1, x='LONGEVITY_D', y='최초후원까지DAY', covar='PLED_FIRST_LONGEVITY'),'\n')
print('3)납입횟수x납입총후원금액, 평균후원금액 : ')
print(partial_corr(data=df1, x='PAY_NUM', y='PAY_SUM_PAYMENTAMOUNT', covar='평균후원금액'),'\n')
```

2. 상관관계분석 – 2. 편상관관계 분석

분석 결과

1)가입일수x첫플릿지 기준 고객 기간, 최초후원까지DAY :

n : 978, r : 1, p-value : 0

2)가입일수x최초후원까지DAY, 첫플릿지 기준 고객 기간 :

n : 978, r : 1, p-value : 0

3)납입횟수x납입총후원금액, 평균후원금액 :

n : 976, r : 0.921475, p-value : 0

4)플릿지수x개발채널수, 납입횟수 :

n : 978, r : 0.556255 ,p-value : 1.575638e-80

결과 해석

1), 2) 분석은 잘못된 분석이었다.

3) 유의한 선형 관계를 보여주었지만 p-value가 0으로 나와 잘못된 분식이었다.

4) 유의한 선형 관계를 보여준다.

실무적 인사이트

- 납입의 많고 적음에 상관없이 노출되는 채널의 수가 많아야 플릿지의 수가 많아진다는 것을 보여준다.

- 추가적으로 어떠한 채널이 플릿지 수에 가장 영향을 주는지 세부적으로 분석해볼 필요가 있다.

2. 상관관계분석 – 3. 순서형 변수의 상관관계 분석

사용 데이터

- PLED_FIRST_LONGEVITY : 첫 결제 후 가입 기한
- AGE : 나이

분석 목적

- 장기 후원자에게 추첨으로 상품을 주는 이벤트
- 고령의 후원자에게 상품을 증정하는 이벤트
- 두 이벤트에 특정 집단의 후원자들만 이벤트의 혜택을 누리는지 검증이 필요

핵심 코드

```
spear = stats.spearmanr(df1['PLED_FIRST_LONGEVITY'], df1['AGE'])
```

분석 결과

correlation=0.16588601134506528,

pvalue=1.8154466589764185e-07

결과 해석

p값이 0.05보다 낮게 나왔기 때문에 두 수치형 데이터의 순서에 연관성이 있다.

실무적 인사이트

- 두 이벤트에 특정 집단이 중복될 수 있다. 때문에 이벤트의 홍보 효과가 효과적일지 고민해야 한다.
- 두 이벤트에 모두 소외되는 집단의 상대적 박탈감을 해소할 대책을 마련해야 한다.

3. t-검정 – 1.일표본 t-검정

사용 데이터

- LONGEVITY_D : 가입 일 수
- PAY_NUM : 납부 횟수
- PAY_SUM_PAYMENTAMOUNT : 총 납부 금액
- 평균후원금액 = 총 납부 금액 / 납부 횟수

분석 목적

- 가입일수 1년 이내 후원자의 평균후원금액이 시간이 지날수록 증가할 것이라 전망한다.
- 구체적으로 1년 이내 후원자의 평균후원금액이 1년이 지나면 10% 상승할 것이라고 예상한다.
- 이러한 예상 수치가 맞는지 확인하기 위해 가입일수 1년 이상 2년 이내의 데이터에 대해 일표본 t-검정을 수행한다.

분석 과정

- 가입일수 1년 이내, 가입일수 1년 이상 2년 이내의 데이터로 구분 한다.
- 가입일수 1년 이내 후원자의 평균후원금의 1.1배를 모수로 하여 일표본 t-검정을 실시한다.

핵심 코드

```
df_1y = df_t[df_t['LONGEVITY_D']<=365]
df_2y = df_t[df_t['LONGEVITY_D']<=365*2]
df_2y = df_2y[df_2y['LONGEVITY_D']>365]
#3. 평균후원금액의 평균 및 일표본 t-검정
print(stats.ttest_1samp(df_2y['평균후원금액'], df_1y['평균후원금액'].mean()*1.1))
```

3. t-검정 – 1.일표본 t-검정

분석 결과

가입후 1년 이내의 평균후원금액의 평균 : 24946.74

가입후 1년 이상 2년 이내의 평균후원금액의 평균 : 23858.53

Ttest_1sampResult(statistic=-3.7801219338528123, pvalue=0.0002782377357566793)

결과 해석

- t-test결과 p-값이 0.0002가 나왔다.
- 1년 이상 2년 이내 후원자의 평균후원금액의 평균은 가입일시 1년 이내 후원자의 평균후원금액의 평균에 비해 10% 증가하지 않았다.

실무적 인사이트

- 예상과는 달리 오히려 가입기한 1년 이내의 평균후원금액의 평균이 1년 뒤에는 소폭 줄어들었다.
- 후원자들의 평균후원금액의 평균일 뿐 전체 후원금을 의미하지는 않는다.
- 평균후원금이 줄어들었지만 후원자의 이탈, 복귀 등으로 전체 후원금에 변동이 생길 수 있다.
- 따라서 1년 단위의 기한으로 나누었을 때 전체 후원금에 차이가 있는지도 함께 분석하는 것이 요구된다.

3. t-검정 – 2. 독립표본검정

사용 데이터

- E_MAIL 이메일 수신 여부
- LONGEVITY_M 가입월수
- PAY_NUM 후원 횟수
- 월후원률 = 후원 횟수/가입월수
- SEX 성별

분석 목적

- 이메일 수신 여부가 월후원률에 영향을 주는지 알아본다.
- 성별에 따른 월후원률의 차이가 있는지 알아본다.

분석 과정

1) 결측값 제거
2) 등분산성 검정
3) 데이터 분할
4) 독립표본 t-검정

핵심 코드

```
df_y2 = np.array(df_y.월후원율)
df_n2 = np.array(df_n.월후원율)
stats.ttest_ind(df_y2, df_n2, equal_var=False)
```

분석 결과

이메일 수신	평균 월후원율	p- value
YES	0.30	0.054
NO	0.34	

성별	평균 월후원율	p- value
남성	0.32	0.91
여성	0.31	

결과 해석

두 분석 모두 p값이 0.05보다 높으므로 이메일 수신여부와 성별에 따른 평균 월후원율은 차이가 없다.

실무적 인사이트

- 이메일을 통한 정보 전달은 후원자의 장기 후원과 연관성이 없었다.
- 후원자의 장기 후원에 영향을 주는 다른 요인을 찾아야 한다.

데이터 특징

고객ID	성별	가입일수	플릿지수	총납입금액	주요채널
A	1	512	1	320,000	TM
B	2	32	1	100,000	DIGITAL
C	1	286	2	250,000	BROADCAST
D	1	762	1	600,000	TB

쌍체표본검정이 불가능한 이유

- 위는 NGO데이터의 예시이다.
- 주어진 데이터는 한 시점의 데이터이다.
- 예를 들어 2020.10.30의 고객데이터이다.
- 고객의 활동에 대한 시간적 단서가 없다.

4. 범주형 데이터 분석 – 1. 적합도 검정

사용 데이터

PLED_NUM_R_PN_INT : 정기해외사업 플릿지 수

PLED_NUM_R_PN_FST : 정기긴급사업 플릿지 수

PLED_NUM_R_PN_DOM : 정기국내사업 플릿지 수

PLED_NUM_R_PN_NKOR : 정기북한사업 플릿지 수

분석 목적

- 기존에 조사된 사업 후원 상품의 비율은 25:25:40:10 이었다
- 기존에 조사된 비율이 지금도 맞는지 확인하기 위해 적합도 검정을 한다

분석 과정

1) 결측값 제거
2) 빈도표 작성
3) 관측도수 기대도수 추출
4) 카이제곱 적합도 검정

핵심 코드

```
for i in range(0,len(columns_k),1) :
colname = columns_e[i]
pd_f.iloc[[0],[i]] = sum(df_1[colname])
Ob = pd_f.values[0,:4]
Pr = np.array([0.25,0.25,0.4,0.1])
n=0
for i in range(0,len(columns_k),1) :
colname = columns_e[i]
n = n + sum(df_1[colname])
E=n*Pr
#4. 카이제곱 적합도 검정하기
ch = stats.chisquare(Ob, E)
```

4. 범주형 데이터 분석 – 1. 적합도 검정

분석 결과

	해외사업	긴급사업	국내사업	북한사업	합계
플릿지 수	80	74	155	10	319
기존비율	25%	25%	40%	10%	100%
실제비율	25.1%	23.2%	48.5%	3.3%	100%

후원 종류별 카이제곱 적합도

- 카이제곱 값: 21.33

- pvalue=8.97e-05)

결과 해석

카이제곱 값은 21.3 이고 p값이 8.97e-05 이다. p값이 유의수준 0.05이하이므로 각 사업별 후원 비율이 기존의 비율을 따르지 않는다

실무적 인사이트

- 기존의 비율은 잘 못 되었다.
- 현재의 데이터로 새로운 사업별 비율을 알아낸다.
- 해외사업, 긴급사업, 국내사업, 북한사업 순으로 [25.1 : 23.2 : 48.5 : 3.3]의 비율을 보여준다.
- 기존의 비율을 조사 했을 때 보다 플릿지 수의 변동이 있었는지 알아봐야 한다.
- 국내사업의 비율이 늘고, 북한사업의 비율이 줄어든 것에 대한 추가적인 분석이 필요하다.

4. 범주형 데이터 분석 – 2. 독립성 검정

사용 데이터

- PLED_NUM_R_CS_INT : 정기해외아동 플릿지 수
- PLED_NUM_R_CS_DOM : 정기국내아동 플릿지 수
- PLED_NUM_R_PN_INT : 정기해외사업 플릿지 수
- PLED_NUM_R_PN_FST : 정기긴급사업 플릿지 수
- PLED_NUM_R_PN_DOM : 정기국내사업 플릿지 수
- PLED_NUM_R_PN_NKOR : 정기북한사업 플릿지 수

분석 목적

- 아동 후원 플릿지 간에 연관성이 있는지 조사한다.
- 사업 후원 플릿지 간에 연관성이 있는지 조사한다.

분석 과정

1) 결측값 제거
2) 파생 데이터 생성
3) 빈도교차표 생성
4) 카이제곱 독립성 검정

핵심 코드

```
for j in range(0,len(df_1),1) :
if (df_1[colname_bs][j] > 0) :
df_1[colname_add][j] = 1
else :
df_1[colname_add][j] = 0
X_2 = pd.crosstab(df_2.해외아동여부, df_2.국내아동여부,
margins=False)
chi_X=stats.chi2_contingency(X_2)
for i in range(2,len(columns_add),1) :
for j in range(i+1, len(columns_add),1) :
colname_i = columns_add[i]
colname_j = columns_add[j]
```

4. 범주형 데이터 분석 – 2. 독립성 검정

분석 결과

1) 해외아동, 국내아동 후원여부에 대한 독립성 검정

카이제곱 통계량 : 15.811

p-value : 6.9957e-05

2) 사업유형 4종의 독립성 검정

카이제곱 통계량	해외사업여부	긴급사업여부	국내사업여부	북한사업여부
해외사업여부				
긴급사업여부	0.006			
국내사업여부	7.665	0.365		
북한사업여부	0.851	0.073	0.005	

p-value	해외사업여부	긴급사업여부	국내사업여부	북한사업여부
해외사업여부				
긴급사업여부	0.93			
국내사업여부	0.99	0.54		
북한사업여부	0.35	0.78	0.94	

결과 해석

1) 해외아동, 국내아동의 후원 여부는 p값이 0.05이하이므로 연관성이 있다.

2) 사업 유형 4종에 대해 각 사업간의 후원 여부는 다른 사업의 후원여부와 연관성이 없다.

실무적 인사이트

1)

- 카이제곱량이 높게 나온 이유를 알기 위해 빈도교차표를 본다.
- 해외아동, 국내아동 한 쪽만 후원하는 비율이 61%이다.
- 양쪽 모두 후원하는 비율은 3%에 불과했다.
- 무슨 요인으로 한 쪽만 후원하는 비율이 높은지 추가적인 조사가 필요하다.

2) 각 사업간 후원에 연관성이 없으니 현행을 유지한다.

4. 범주형 데이터 분석 – 3. 동질성 검정

사용 데이터

- AGE : 나이
- MOTI_CHANNEL : 주요 채널

분석 목적

- 연령대별 선호하는 채널의 비율이 다른지 조사한다.

분석 과정

1) 결측값 제거
2) 데이터 분할
3) 모집단 랜덤 추출
4) 카이제곱 동질성 검정

핵심 코드

```
df1_sample = df1.sample(200, random_state = 29)
X_f = pd.crosstab(df3.고객연령대, df3.주요채널,
margins=False)
chi=stats.chi2_contingency(X_f)
```

결과

	0	1	2	3	4	5	6	7	All
장년층	1	82	32	2	10	14	56	2	200
청년층	4	42	60	1	17	15	59	2	200
All	5	125	92	3	27	29	115	4	400

카이제곱 통계량이 27.0 p-value가 0.0007

결과 해석

p값이 0.05이하이므로 유의확률 95%에서 연령대별 주요 채널의 비율은 차이가 있다.

실무적 인사이트

- 장년층은 청년층 보다 BROADCAST 채널 선호가 높다.
- 청년층은 장년층 보다 DIGITAL 채널 선호가 높다.
- BROADCAST와 DIGITAL채널은 특정 연령대에 대한 홍보 효과가 기대된다. 이에 대한 추가적 조사 필요.