**상관관계 분석**

1. **상관관계, 상관계수**

**상관관계**에서 **상관**은 확률론과 통계학에서 두 변수 간에 어떤 선형적 또는 비선형적 관계를 갖고 있는 지를 분석하는 방법이다. 두 변수는 독립적인 관계이거나 상관된 관계일 수 있고, 이때 두 변수 간에 관계의 강도를 상관관계라 한다. 한 변수의 변화에 따른 다른 변수의 변화 정도와 방향을 예측하는 분석 기법이다.[1] 두 변수간에 원인과 결과의 인과관계가 있는지에 대한 것은 **회귀분석**을 통해 인과관계의 방향, 정도와 수학적 모델을 확인할 수 있다.[2]

**상관계수**는 등간척도 이상의 두 변수 중에서 한 변수의 변화가 다른 변수의 변화에 따라 어떤 변화가 일어나는지를 보여주는 지표이다.

1. **개념 정리**

**용어**

(모 상관계수 : p, 표본 상관 계수 : r)

선형성 : 두 변인 X와 Y의 관계가 직선적인지를 알아보는 것으로 산점도를 통하여 알 수 있다.

동변량성 : X의 값에 관계없이 Y의 흩어진 정도가 같은 것을 의미한다. (반대어 : 이분산성)

두 변인의 정규분포성 : 두 변인의 측정치 분포가 모집단[[1]](#endnote-1)에서 정규분포[[2]](#endnote-2)를 이루는 것이다.

무선독립표본 : 모집단에서 표본을 뽑을 때 표본대상이 확률적으로 선정되는 것이다.

공분산 : 2개의 확률변수의 상관정도를 나타내는 값, 두 변수가 서로 상승하는 값을 보인다면 양수, 다른 하나가 하강하는 값을 보이면 음수를 갖는다.

**상관 계수 특징**

변수간의 관계의 정도와 방향을 하나의 수치로 요약해 주는 지수이다.

상관계수는 -1.00에서 +1.00 사이의 값을 가진다.

변수와의 방향은 (-)와 (+)로 표현한다. 양의 상관계수는 (+)값이 나타나고 음의 상관계수는 (-)값이 나타난다.

|  |  |
| --- | --- |
| 상관계수 | 상관관계 |
|  | 상관관계가 매우 높다 |
|  | 상관관계가 높다 |
|  | 상관관계가 있다 |
|  | 상관관계가 있으나 낮다 |
|  | 상관관계가 거의 없다 |

상관계수의 절대값이 높을수록 두 변수 간의 관계가 높다고 할 수 있다.

단 상관관계가 높다고 해서 반드시 인과관계가 있는 것은 아니다.

**척도**

* + 1. 명목척도 : 가장 낮은 수준의 척도이다. 측정대상의 특성만 구분하기 위해 숫자나 기호를 할당한 것으로 양적인 분석과 대소 비교를 하지 못한다. (순위, 1~3위의 차가 4~6위의 차와 같은가?)
    2. 서열척도 : 측정대상의 특성을 구분하고, 이들 사이에 상대적인 크기를 나타낼 수 있고, 서로간에 비교가 가능한 척도 (설문조사 5문항, 매우그렇다, 그렇다, 보통, 아니다, 매우아니다)
    3. 등간척도 : 명목척도와 서열척도 모두 가지고 있고, 크기가 어느정도 되는지, 특성간의 차이가 얼마나 되는지 파악이 가능한 척도 (온도, 리커트 척도)
    4. 비율척도 : 가장 높은 수준의 척도로서, 가장 자세한 정보를 제공, 서로의 구분, 크기의 비교, 크기의 차이, 그리고 특성간의 계산까지 가능한 수준, 모두 숫자로 표현되고 그것들이 계산이 가능함 (절대 영점이 있는 등간척도)

1. **얻고자 하는 것이 무엇인가?**

상관관계 분석을 통해 얻을 수 있는 것들을 얻어 두 변수사이의 관계를 증명한다.

1. **방법**

상관관계를 알아내는 방법은 다음과 같은 방법이 있다.

* 피어슨 상관 계수
* 스피어만 상관 계수
* 크론바흐 상관 계수 신뢰도

다음은 척도에 따른 상관계수의 종류이다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 비연속적 명목척도 | 연속적 명목척도 | 서열척도 | 등간 비율척도 |
| 비연속적 명목척도 | Phi계수, 유관계수, Lambda |  |  |  |
| 연속적  명목척도 | Yule’s Q | 사분 상관계수 |  |  |
| 서열척도 | 등위 양분 상관계수 |  | Spearman의 서열상관계수,  Kendall의 tau |  |
| 등간 비율척도 | Cramer’s V, 양류 상관관계, 상관비(n) | 양분 상관계수 |  | Pearson 상관계수, 상관비(n) |

* + 1. **Pearson의 상관관계** (적률 상관관계) : 연속형 & 연속형(모수적)
* 두 변수가 등간척도 이상이여야 한다.
* 두 변수가 직선의 관계를 가지고 있어야한다.
* 각 행과 열의 분산이 비슷해야 한다.
* 적어도 하나의 변수가 정상분포를 이루어야 한다.
* 사례수가 적을수록 신뢰도가 떨어진다.
* 상관관계는 r로 표현한다.
* 데이터 집합에 나머지 값들과 매우 다른 값이 하나 있으면 상관계수의 값이 크게 바뀔 것이다. 때문에 원인 식별 후 측정오류를 삭제하고 비정상적인 일관성 사건과 연관된 데이터는 삭제한다.
* 유의수준을 설정한다 (일반적으로 0.05로 설정)
* 유의수준은 실제로 상관 관계가 없는데 있는 것으로 결론을 내리는 위험을 막기 위함이다.
* P-값 < a : 상관관계가 통계적으로 유의하다. 반대면 상관관계가 0과 다르다.
  + 1. **Spearman의 서열상관관계**: 연속형 & 연속형 (비모수적)
* 독립변수와 종속변수가 서열척도로 구성된 경우에 사용
* 상관관계의 값은 -1.00 ~ +1.00 사이에 있다.
* 주어진 자료에서 등간성이 의문시되거나, 변수들의 점수가 극단적인 분포를 나타내는 경우에는 적률상관관계 대신 Spearman의 서열 상관계수를 사용한다. (정규분포를 크게 벗어나거나 순위 척도의 자료일 때 사용하는 값이다.)
* 상관계수는 p로 표현
* Spearman의 R계수는 사례수가 많거나 두 변수간의 순위차가 커서 계산이 길 때 사용한다.
* 값에 순위를 매겨 그 순위에 대한 상관계수를 구하는 것이다.
  + 1. **Kendall’s tau** : 연속형 & 연속형 (비모수적)
* 스피어만 상관계수와 비슷한 방식이다.
* 순위 척도 자료 또는 순위 척도 자료로 변환한 자료를 이용해 두 변수간의 상관 계수의 크기를 나타낸다.

Ex) 학생들의 언어, 수리 점수가 주어졌을 때 캔달 상관관계

수리영역, 언어영역 각각 등수를 매기고 동점일 때는 등수의 평균으로 등수를 매긴다. 동점이 2명이면 4와 5등의 중간인 4.5이다. 이후 언어와 수리의 증감 경향 일치를 보기를 위해 2명씩 짝짓는 모든 조합을 본다. 둘 다 언어, 수리가 증가하는 경향이면 1, 아니면 -1, 등수가 동일하면 4.5를 부여한다. 이런 조합으로 나온 값을 모두 더한 후 학생 수 만큼 나누면은 켄달의 상관계수 t가 나온다.

* + 1. **양류 상관계수 Point Biserial correlation coefficient** : 명목형(이분형) & 연속형

네/아니오, 있음/없음으로 구분 가능한 binary변수 이고, 다른 변수가 연속형인 경우에도 상관계수를 구할 수 있다.

나이 같은 수치형 데이터의 경우 50대 이상, 50대 미만으로 구분지어 이분형으로 만들어 사용할 수도 있다.

Ex) 성별과 수학점수와의 상관계수 관계

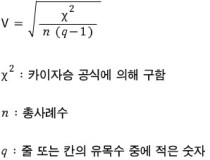
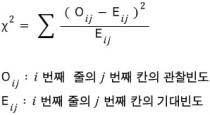
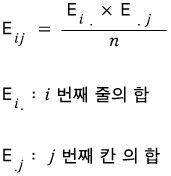
* + 1. **파이 계수** : 명목형(이분형) & 명목형(이분형)

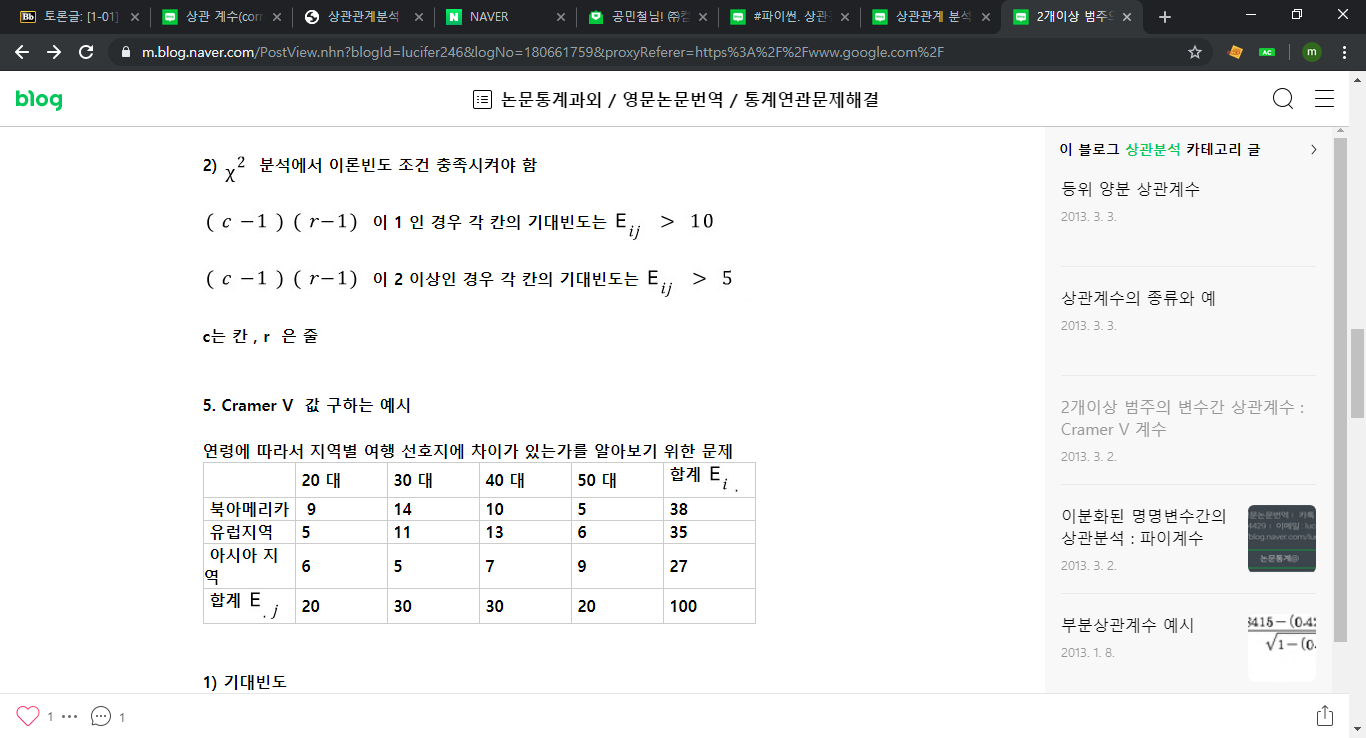
두 개의 변수가 binary변수 일 때 두 변수간의 계수 값을 구할 수 있다. L,C, Lambda 등을 사용하기도 한다.

* + 1. **사분상관계수** : 모두 명명척도이고 인위적인 이분변수

지능의 상하와 시험의 합, 불 상관관계

* + 1. **Cramer’s V**
* 2개 이상의 범주로 나눈 집단간의 상관계수를 구한다.
* V계수는 항상 양수이다.
* 관련성을 가질 때 최대 1 아닐 때 최소 0
* 공식

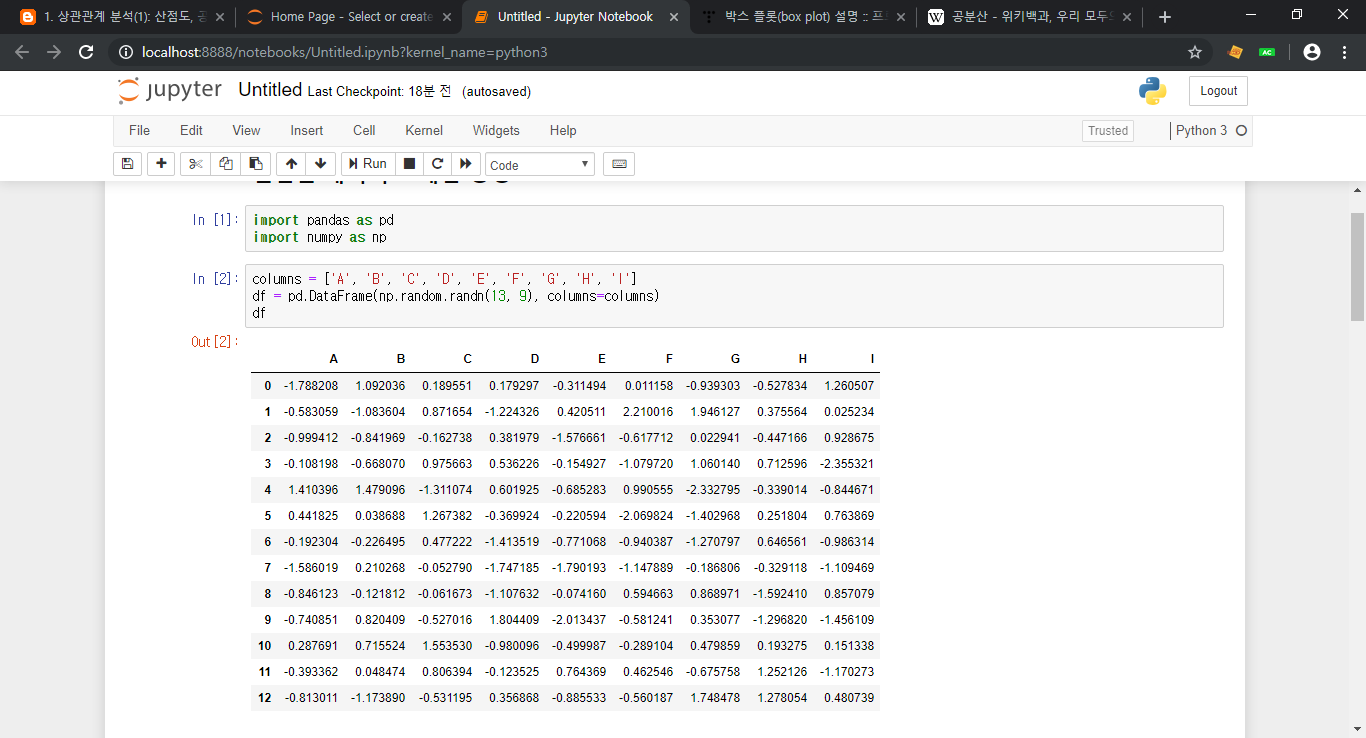
  

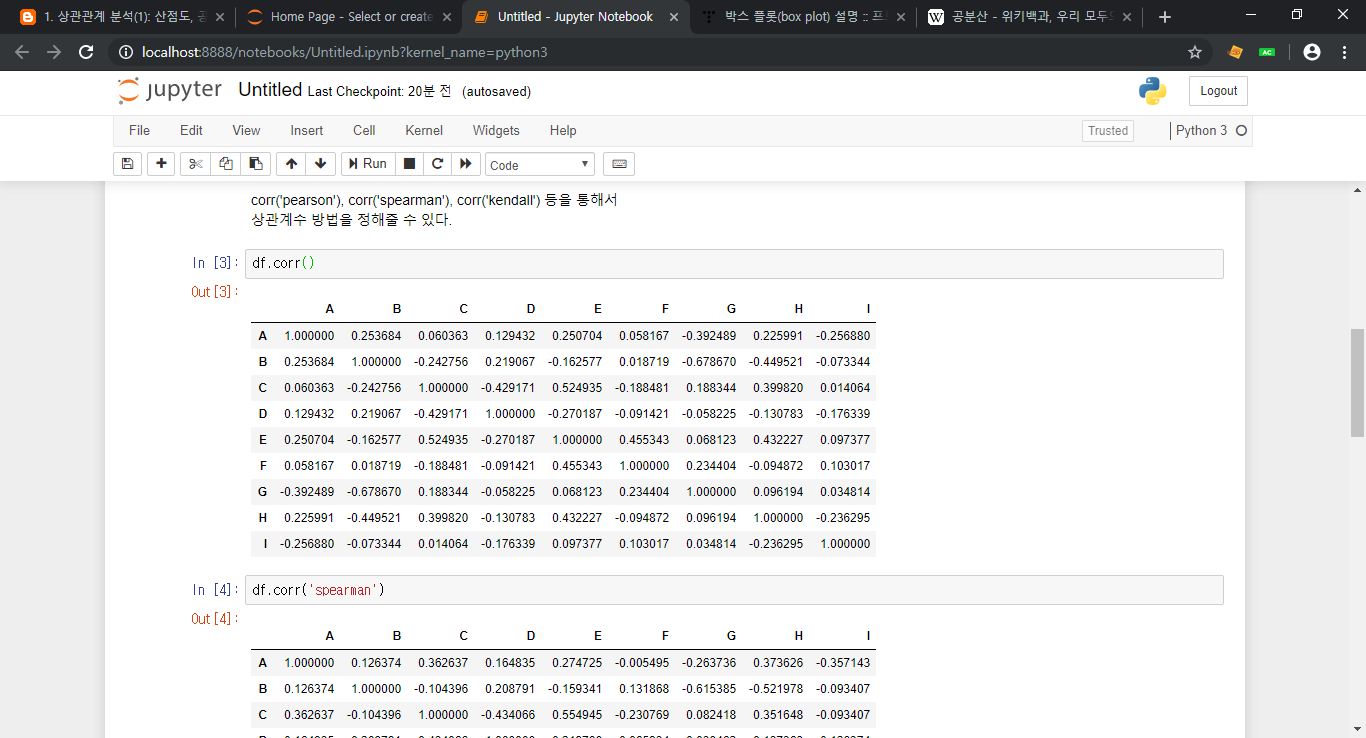


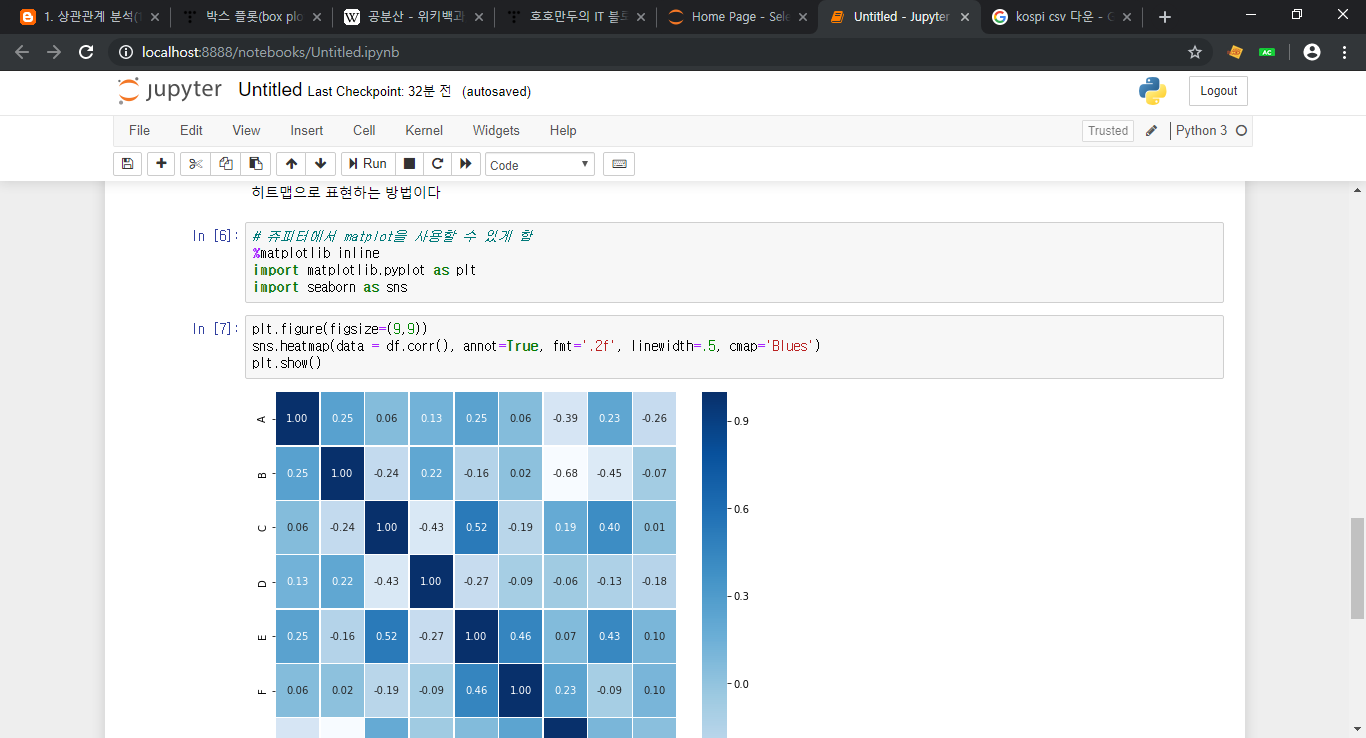
Ex) 연령에 따른(20대, 30대, 40대, 50대) 여행지 선호도(북아메리카, 유럽, 아시아) 관계 분석, 각 연령과 여행지의 기대빈도를 구한다. 그 이후 공식을 적용해 V를 구한다. V의 값을 보고 상관관계를 보면된다.

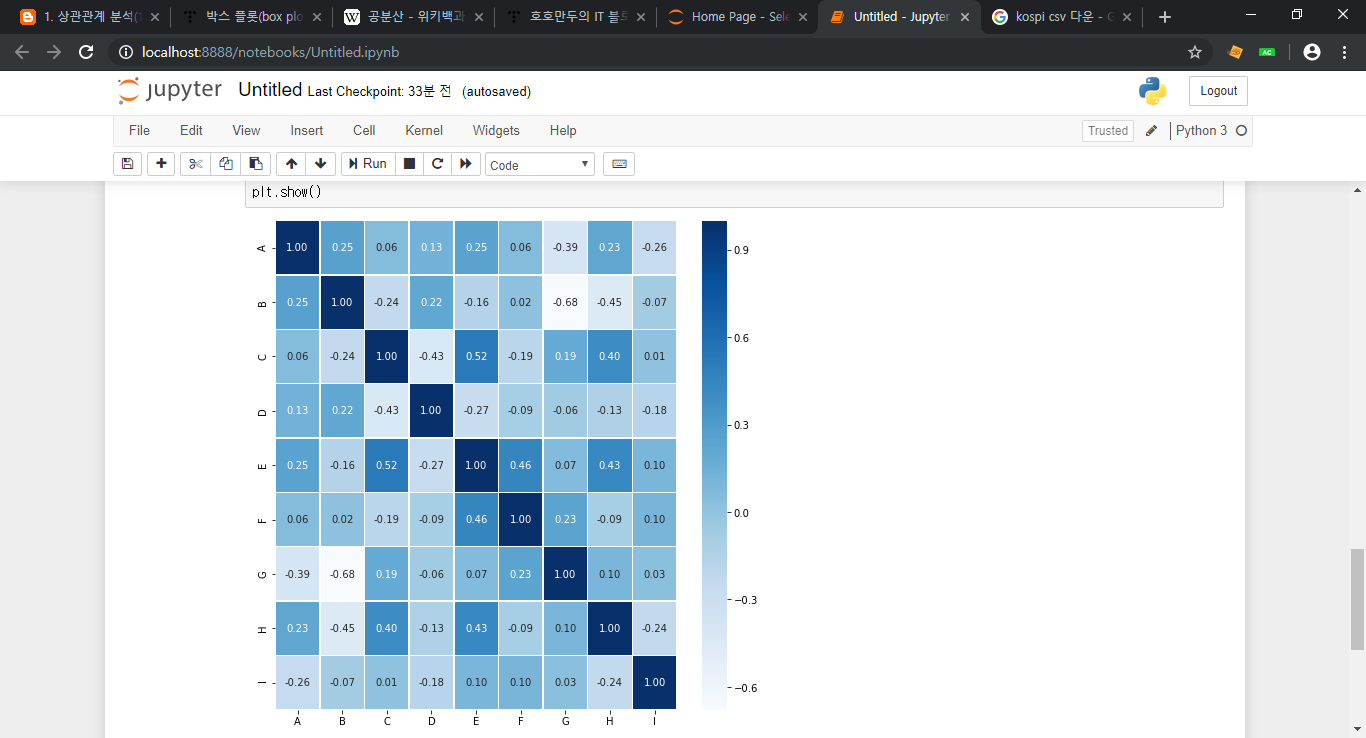
* + 1. **람다(lambda)**
* 명목변수들간의 상관관계의 크기를 나타내는 계수
* 독립변수와 종속변수의 상관관계
* PRE[[3]](#endnote-3) 규칙에 기초한 상관계수이다.
* 독립변수가 종속변수에 대한 예측력을 얼마나 가졌는가를 본다.
* 람다계수는 독립변수가 종속변수에 대한 예측력을 얼마나 가졌는가를 본다.
* Ex) 사회복지사의 선호 치료방법과 기관 유형과의 관계

1. 파이썬 코드 예제
   * 1. A부터 I가 랜덤한 값을 갖는 데이터 프레임의 상관계수









[참고자료]  
[1] KSI, 통계연구, 상관관계분석, http://6025.co.kr/bbs/board.php?bo\_table=cust\_in&wr\_id=13

[2] 위키피디아, 상관 분석, https://ko.wikipedia.org/wiki/상관\_분석

1. 정보를 얻고자 하는 관심대상의 전체 집합을 말한다. 모집단의 개체는 범주적 특성과, 수량적인 특성을 가지고 있다. [↑](#endnote-ref-1)
2. 연속 확률 분포의 하나이다. 수집된 자료의 분포를 근사하는데 자주 사용되며, 이것은 중심극한정리에 의하여 독립적인 평균은 정규분포에 가까워지는 성질이 있다. [↑](#endnote-ref-2)
3. PRE에 기초한 상관계수 = B-A / B, \*B 독립변수를 도입하지 않은 상태에서 오류 수, \*A 독립변수를 도입한 후의 오류 수 [↑](#endnote-ref-3)