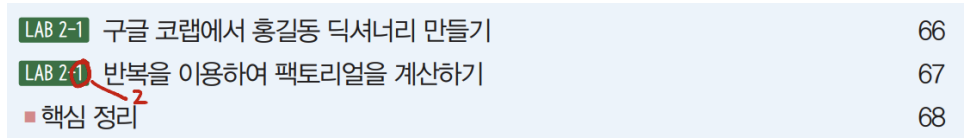
"따라하며 배우는 빅데이터 분석_1쇄" 정오표

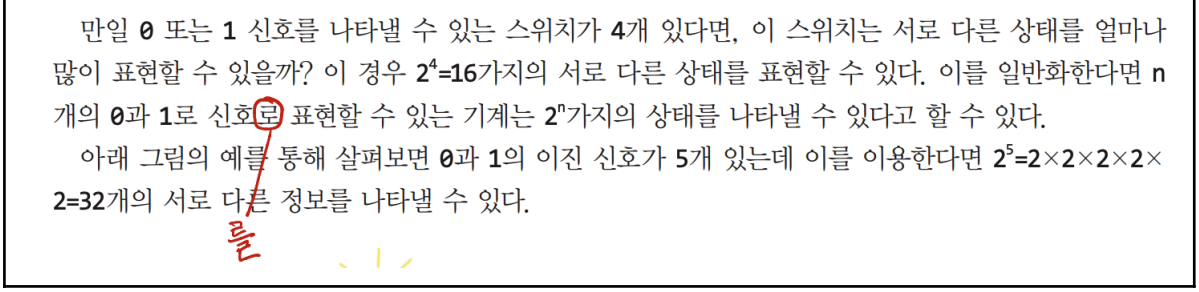
이 자료는 생능출판사 "따라하며 배우는 빅데이터 분석"책 내용에 있는 오류를 정리한 페이지 입니다.
불편을 끼쳐드려 대단히 죄송합니다. 다음 인쇄 때 수정하여 반영하겠습니다.*(2025년 8월 1일)*

오류 : 잘못된 코드 오류, 실행 결과 오류, 잘못된 설명, 잘못된 참조 번호(그림, 코드 등)

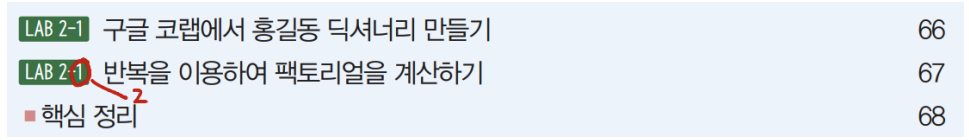
오류 페이지 : 9 page
오류 위치와 오류 : 차례의 번호 오류 LAB 2-1이 두번 나타남



오류 페이지 : 20 page
오류 위치와 오류 : 중간 부분 오타 “신호로”->“신호를”



오류 페이지 : 9 page
오류 위치와 오류 : 차례의 번호 오류 LAB 2-1이 두번 나타남



오류 페이지 : 22 page

오류 위치와 오류 : 오타 “제타바이크”->”제타바이트”

이 밖에도 제타바이크의 1,024배인 요타바이트, 요타바이트의 1,024배인 브론토바이트라는 거대한 단위도 존재한다. 이렇게 사용되는 데이터의 용량이 실제로 어느 정도의 정보량을 표현하는지는 다음 표로 설명할 수 있다.

오류 페이지 : 94 page

오류 위치와 오류 : 오타 “stp” -> “step”

또한 일정한 크기로 증가되는 값을 지정하려면 다음과 같이 한다. 넘파이의 **arange()**는 실수값을 **stp** 값으로 줄 수 있다.

오류 페이지 : 98 page

오류 위치와 오류 : 오타

이전에 살펴본 바와 같이 리스트를 결합하기 위해서는 **+** 연산자를 사용할 수 있다. 하지만 다차원 배열에서 **+**는 두 배열 원소의 합을 구하는 연산이다. 따라서 **a** 다차원 배열과 **b** 다차원 배열을 결합하기 위해서는 어떤 방법을 사용해야 할까? 이 경우 **+** 연산자 대신 다음과 같이 넘파이의 **concatenate()** 함수를 사용해서 두 다차원 배열을 결합할 수 있다.

오류 페이지 : 104 page

오류 위치와 오류 : 오타

이제 다음과 같은 이차원배열에 대해서 특정 값을 행과 열에 삽입하는 경우를 고려해 보자. 만일 **axis**를 명시하여 **[[1, 1], [2, 2], [3, 3]]** 형상의 **2**차원 배열에 **insert(a, 1, 4, axis = 0)**를 하게 되면 **[[1, 1], [4, 4], [2, 2], [3, 3]]**와 같은 배열이 된다. **insert()** 함수의 첫 매개변수는 배열 객체이며, 두 번째는 삽입할 위치, 세 번째는 삽입할 값, 네 번째는 삽입 방향인데 **axis = 0**으로 할 경우 **0**축(그림의 **axis 0**) 방향이 된다.

오류 페이지 : 121 page

오류 위치와 오류 : 오타와 문장 수정

분산 값을 살펴보면 A 모둠의 분산 값이 200.0이고 B 모둠의 분산 값이 8.0으로 나왔다. 이 값들을 통해서 A 모둠 학생들의 점수 분포가 평균으로부터 떨어져 매우 넓게 분포하고 있다는 알 수 있다.

분산을 통해서 데이터가 평균값 주변에 모여있는 정도를 측정할 수 있으나 이 값은 편차의 제곱값들 합의 평균값이다. 이 편차의 제곱은 실제 값과 너무 동떨어진 값이 될 수 있기 때문에 이를 실제값과 근사시키기 위하여 분산에 제곱근을 씌워서 사용하는 것이 더 합당할 것이다. 이 값이 바로 **표준편차**standard deviation이다. 표준편차를 구하는 넘파이 함수는 std()이다. m개의 인스턴스를 가진 데이터 x_i의 평균은 그리스 문자 μ_x로 표기하며, 분산은 σ_x^2으로, 표준편차는 σ_x 표기를 많이 사용한다.

오류 페이지 : 131 page

오류 위치와 오류 : 오타

permutation()과 shuffle()은 유사해 보이지만 결정적인 차이가 있다. permutation()의 경우 인자로 들어오는 다차원 배열를 복사해서 섞어준다. 하지만 shuffle()은 다차원 배열을 섞는 일을 **인플레이스**inplace로 한다. **인플레이스란 새로운 배열을 만드는 것이 아니라 배열 자체의 원소값을 변화시키는 기능**이다. 인플레이스는 다차원 배열이나 판다스 데이터프레임에서 매우 중요한 개념이다. 다시 한번 강조하자면 인플레이스로 다차원 배열을 섞어주면 다차원 배열 자체가 변하게 된다. 이 기능을 다음 코드로 확인해 보자.

오류 페이지 : 137 page

오류 위치와 오류 : 오타 sin(pi * x / 100) 함수는 주기가 200임

새롭게 추가된 리스트 z는 진폭이 100인 **사인**sine 함수로 0도에서 180도까지 그려지게 된다. 사인함수는 주기함수이므로 선형, 비선형적으로 증가하는 앞의 두 리스트 x, y와는 큰 상관관계가 없을 것으로 예측된다.

오류 페이지 : 145 page

오류 위치와 오류 : 오타

COVID-19 감염질환은 나이가 많은 감염자의 치명률이 높지만 나이가 어린 감염자는 치명률이 낮은 편이다. 이러한 설명은 오른쪽의 그림과 같이 나타낼 수 있는데 막대 그래프의 x축은 감염자의 연령대를 나타내며, 색상과 높이는 미국과 이탈리아의 사망률을 나타내고 있다. 그림과 같이 고연령층의 높은 사망률을 색상으로 잘 나타내어 그 위험성을 표현하고 있다.

영국

오류 페이지 : 156 page

오류 위치와 오류 : 오타

차트를 그릴 수 있다.

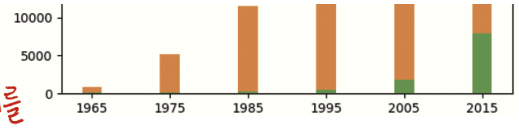
2023년에 국제 통화 기금(IMF)이라는 국제기구에서 발표한 주요 국가별 명목 1인당 국내 총생산액(GDP) 데이터를 막대형 차트를 사용하여 시각화하도록 하자. 이번에는 국가의 이름과 GDP를 data라는 딕셔너리 형태로 만들고 이를 다시 리스트로 만들어서 시각화시켜 보도록 하자.

사

오류 페이지 : 157 page

오류 위치와 오류 : 오타

위의 코드에서 plt.xticks() 함수는 막대형 차트의 x 축 항목의 수와 항목 값을 인자로 받는다. 따라서 range(len(years))를 통해서 6개의 간격을 생성하고나서 이 간격의 레이블을 두 번째 인자인 years로 두었다. 이 코드의 수행 결과는 그림과 같이 나타나는데 무언가 문제가 있어 보인다.



막대를

막대의

오류 페이지 : 182 page

오류 위치와 오류 :

다음으로 0에서 1,00 사이의 난수를 10개 생성한 후 y4에 넣어 x와의 상관관계를 보면 어떤 결과가 나올까? 난수란 특정한 규칙이 없는 수이므로 상관관계는 0에 가까운 값으로 나타날 것이며 이 결과를 아래와 같이 확인할 수 있다.

100

계수

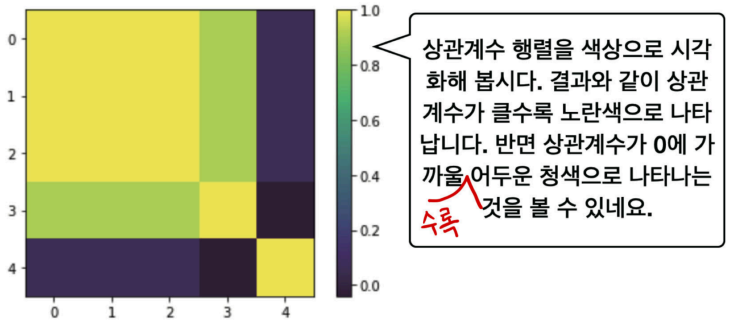
오류 페이지 : 183 page

오류 위치와 오류 : 상관관계보다 상관계수가 더 적절한 표현임

이 상관관계를 다음과 같이 값으로 출력해 본 후 맷플롯립으로 시각화해 본다면 더 좋을 것이다. x와 y1, y2, y3, y4의 상관계수를 얻기 위하여 corrcoef() 함수의 입력으로 세 개의 다차원 배열을 넣어 줄 때 (x, y1, y2, y3, y4)와 같이 튜플 형태로 입력해야 한다는 점에 유의하도록 하자.

오류 페이지 : 184 page

오류 위치와 오류 : 풍선글 내부 오타



오류 페이지 : 189 page

오류 위치와 오류 : 오타

로 이 필드의 값을 히스토그램으로 살펴보자. 이 때, kde=True 키워드 인자를 통해서 **가우시안 커널 밀도 추정**gaussian kernel density estimation을 실선으로 그려보도록 하자. 또한 bins=20으로 하여 빈의 수를 더 늘려보도록 하자. **이상치 탐지**란 정상 데이터 또는 일반적인 데이터에 비하여 그 분포가 현저하게 차이가 나는 데이터를 탐지하는 기법이다. 이상치 탐지 방법 중 **밀도 기반 이상치 탐지법**이란 **데이터의 분포를 사용하여 이상치를 찾아내는 방법**이다. 이 방법은 기존에 존재하는 데이터 분포를 사용하여

오류 페이지 : 194 page

오류 위치와 오류 : 설명 오류

어느 정도 팁을 지출하였는가에 대한 정보를 얻고자 한다. 이 경우에는 2x2 크기의 다중 패널이 필요할 것이다. 그리고 이 패널에는 점심 식사 시간의 흡연자, 저녁 식사 시간의 흡연자, 점심 식사 시간의 비흡연자, 저녁 식사 시간의 흡연자의 팁 정보를 널어 놓으면 될 것이다. 이를 지원하는 다음과 같은

오류 페이지 : 194 page

오류 위치와 오류 : 설명 오류

위의 결과에서 ①은 smoker=Yes | time=Lunch라는 제목의 서브플롯으로 점심 시간대의 흡연자 테이블의 팁 정보이다. 또한 ②는 저녁 식사 시간의 흡연자, ③은 점심 식사 시간의 비흡연자, ④는 저녁 식사 시간의 비흡연자의 전체 식사 비용(total_bill) 정보를 담고 있다.

아래 코드는 요일에 따른 흡연자와 비흡연자의 팁 지출에 대한 FacetGrid 시각화의 결과이다. 이 경우 목요일(Thur), 금요일(Fri), 토요일(Sat), 일요일(Sun)의 요일이 열의 내용으로, 흡연자/비흡연자가

오류 페이지 : 195 page

오류 위치와 오류 : 오타

이제 파셋그리드의 map() 메소드에 대하여 살펴보자. 이 메소드의 첫 번째 인자에는 plt.hist와 같은 호출 가능한 플로팅 함수가 올 수 있다. plt.hist 함수는 맷플롯립의 히스토그램을 그리는 함수이다. 다음으로 올 수 있는 두 번째 인자인 'total_bill'은 플로 함수가 그려야할 데이터의 이름이다. 이 데이터는 전체 식사비를 요일별로 그려주는데, 만일 'total_bill' 대신 'tips'가 온다면 전체 식사비가 아닌 지출한 팁이 나타나게 된다.

오류 페이지 : 245 page

오류 위치와 오류 : 오타

이러한 인덱스에 부울값을 넣게 되면 부울값이 True인 레코드만을 추출할 수 있다. 우선 다음과 같이 df.index > 251103으로 True, False를 출력해보자. 이 경우 201101, 201102, 201103 학번은 False가 되고 나머지는 True가 된다. 따라서 이 True, False 인덱스 값을 df.loc[]에 넣어주면 학번이 201104, 201105, 201106, 201107인 레코드만 화면에 나타난다.

오류 페이지 : 248 page

오류 위치와 오류 : 오타

우선 이전에 살펴본 데이터프레임에 대하여 점수의 합계를 기준으로 정렬하고자 한다. 이 경우 sort_values() 라는 메소드를 사용한다. 이 메소드의 인자로 비교하고자 하는 항목값(또는 열의 이름)을 넣어줄 수 있다. 또한 크기가 커지는 순서인지 작어지는 순서인지를 명시하는 ascending이라는 키워드 인자값을 줄 수도 있다. 다음 코드는 '합계' 열을 기준으로 크기가 작아지는 순서로 데이터 집합을 나열하는 코드이다.

오류 페이지 : 248 page

오류 위치와 오류 : 오타

이렇게 정렬된 데이터가 저장된 데이터프레임이 sorte_df이라고 할 때, 이것을 A학점, B학점, C학점으로 나누는 방법에 대해서 알아보자. 이 방법은 다음과 같이 간단하게 iloc를 이용한 슬라이싱을 사용한다. 따라서 A학점 학생은 [:3]을 통해 가장 앞에 있는 3명, B학점 학생은 [3:5]을 통해 가운데

오류 페이지 : 277 page

오류 위치와 오류 : 설명 오류

이 결과를 살펴보면 평균 기온의 결측값은 없으나 최대 풍속은 4개의 결측값이 4개 임을 알 수 있다. 이 정보를 바탕으로 최대 풍속의 결측값도 출력하도록 하자.

오류 페이지 : 299 page

오류 위치와 오류 : 오타

new_df라는 데이터프레임을 살펴보면 중복된 데이터가 삭제되었음을 알 수 있다. 만일 다음과 같이 '순서'라는 새로운 열을 만들어서 0부터 5까지의 공유한 번호를 준다면 중복된 데이터가 없을 것이다. 따라서 drop_duplicates() 메소드를 호출해도 아무 일도 일어나지 않는다.

오류 페이지 : 339 page

오류 위치와 오류 : 오타

는 1등실 승객보다도 더 적은 것을 볼 수 있다.

위의 히스토그램을 조금 수정하여 객실 전체 승객 대비 시망자가 아닌 **객실별 사망자와 생존자를 비교**하는 그래프를 그려보도록 하자. 이번에는 시본의 countplot()을 사용하도록 하자.

오류 페이지 : 415 page

오류 위치와 오류 : 오타

이 정렬된 특성을 살펴보며 **체질량지수(bmi)와 s5 속성이 당뇨에 영향을 크기 미치는 중요한 특성**으로 보여진다. 반면 남녀의 차이인 **sex와 s2, 나이 특성 등은 당뇨 수치에 영향을 거의 주지않는 특성**으로 보여진다.

오류 페이지 : 419 page

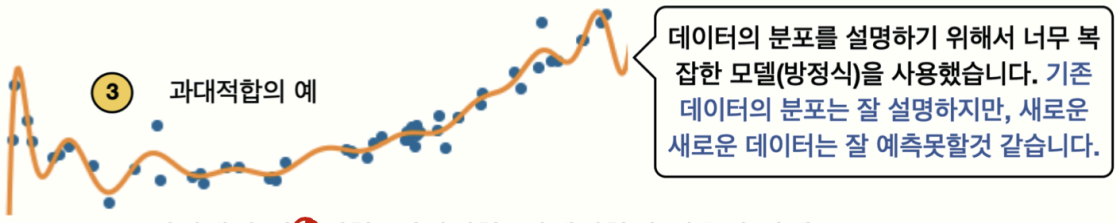
오류 위치와 오류 : 오타

이와는 달리 남녀의 차이인 sex와 s2 등과 같이 당뇨 수치에 영향을 거의 주지않는 특성을 이용하여 만든 모델은 어떤 형태를 가질까? 다음과 같이 diabetes.data[:,[1, 5]]를 사용하여 regr_B라는 모델을 만들도록 하자. 이 모델을 이용하여 만든 모델을 시각화한다면 다음과 같은 결과를 얻을 수 있다.

데이터를

오류 페이지 : 425 page

오류 위치와 오류 : 오타



회귀에서 과대적합, 최적적합, 과대적합이 적용된 사례

소

오류 페이지 : page

오류 위치와 오류 :

오류 페이지 : page

오류 위치와 오류 :

오류 페이지 : page

오류 위치와 오류 :

오류 페이지 : page

오류 위치와 오류 :