**Report**

**데이터 이해하기**

**빅데이터 학과**

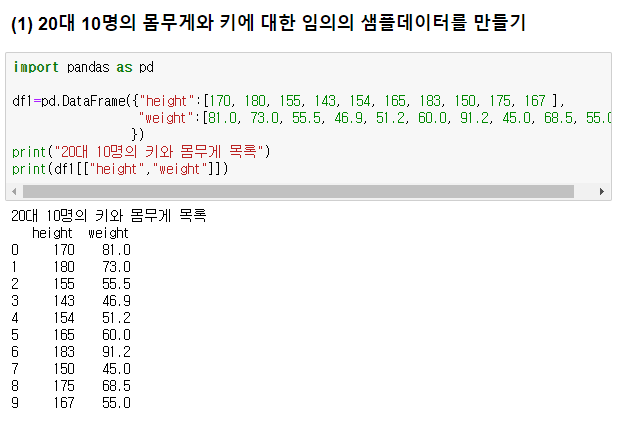
**20175327 이연희**

**1) 데이터 탐색(Data Explore)을 설명하세요.**

데이터 탐색이란, 데이터를 이해하기 위한 통계요약과 그래프를 통해 속성의 특징 및 속성 사이의 관계를 확인하는 것이다. 데이터 이해가 높을 수록 데이터 분석, 모형구축에 용이하다.

**2) 다음 문제를 해결하세요. (파이썬을 사용하면 추가점수 부여)**

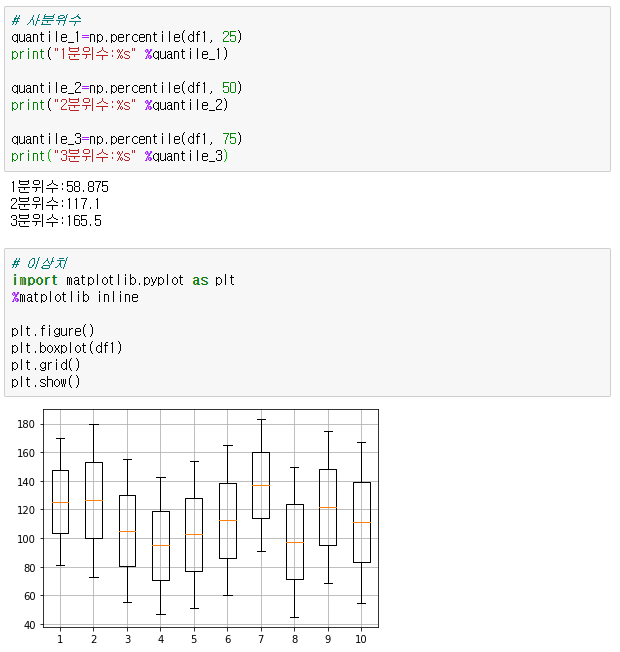
(1) 20대 10명의 몸무게와 키에 대한 임의의 샘플데이터를 만들기



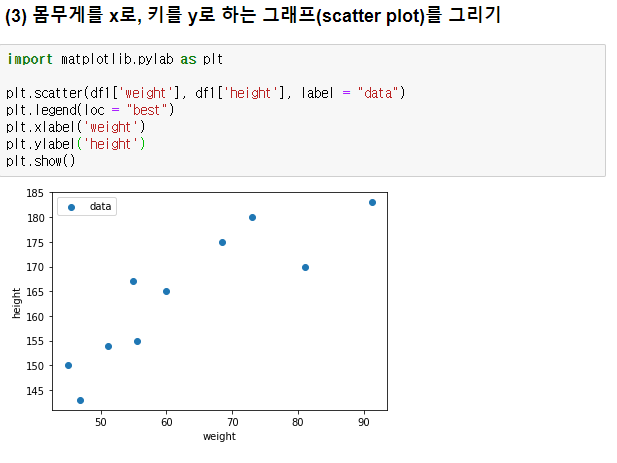
(2) 몸무게, 키에 대한 각각의 평균(mean), 표준편차(standard deviation), 최소값(minvalue), 최대값(maxvalue), 사분위수(quantile), 이상치(outliers) 계산



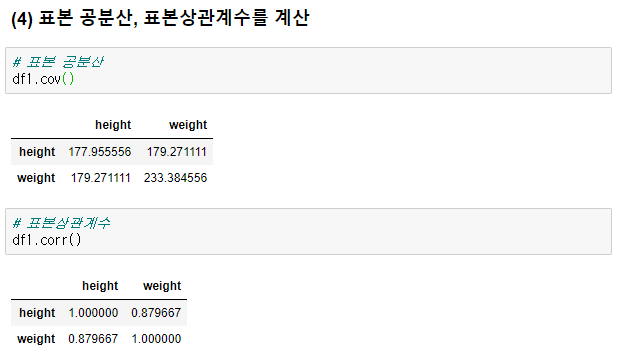




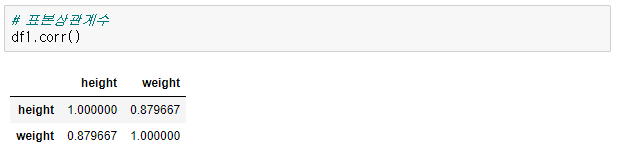
(3) 몸무게를 x로, 키를 y로 하는 그래프(scatter plot)를 그리기



(4) 표본 공분산, 표본상관계수를 계산



(5) 몸무게와 키 사이에 선형적 상관 관계를 설명하세요.



피어슨 상관계수 표를 참고하여 상관 관계를 분석하면 1과 가까울수록 양의 상관관계이며 -1과 가까울수록 음의 상관관계인 것을 알 수 있다.

그러므로 데이터 height와 weight는 양의 상관 관계를 갖는 것을 알 수 있다.

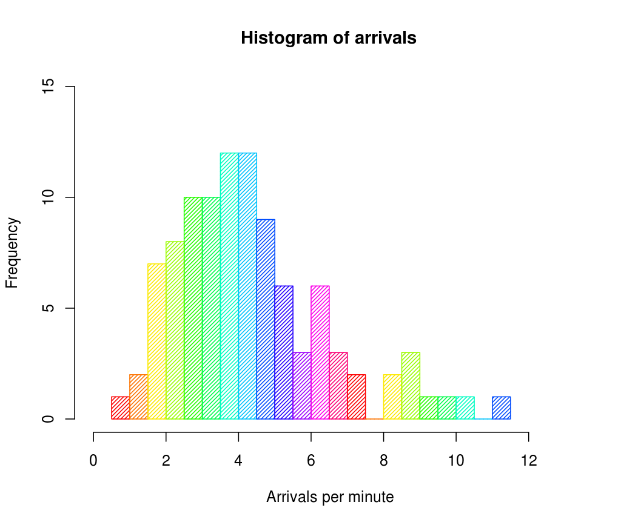
**3)데이터의 시각화(visualization)를 설명하세요.**

데이터 시각화는 데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달되는 과정을 말한다. 데이터 시각화의 목적은 도표라는 수단을 통해 정보를 명확하고 효과적으로 전달하는 것이다.

* Histogram, Boxplot: 단일 수치형 변수의 분포 확인
* Bar, Pie: 범주형 변수에 대한 비교
* Scatter: 수치형 변수들 간의 관계 확인, outlier 확인
* Scatter matrix: 여러 변수들 사이의 모든 가능한 scatter를 한 장에 표현하여 변수들 사이의 관계를 신속하게 확인가능

**4)histogram을 설명하세요.**

히스토그램은 표로 되어 있는 도수 분포를 정보 그림으로 나타낸 것이다. 즉, 도수분포표를 막대그래프로 나타낸 것이다. 보통 히스토그램에서는 가로축이 계급, 세로축이 도수를 뜻하는데, 때때로 반대로 그리기도 한다. 계급은 보통 변수의 구간이고, 서로 겹치지 않는다.



<히스토그램(histogram)>

**5)box plot을 설명하세요.**

Box plot은 데이터의 분포를 보여주는 상자그림으로 자료로부터 얻어낸 통계량인 5가지 요약수치를 이용하여 그려진다.

* 5가지 요약 수치: 최소값, 제 1사분위(Q 1) 제 2사분위(Q 2, median), 제 3사분위(Q 3), 최대값
* 사분위범위(IQR, Inter Quartile Range): ‘제3사분위수 - 제1사분위수’로 계산
* Q1-1.5 \* IQR 보다 큰 데이터 중 가장 작은 값(lower whisker), Q3 +1.5 \* IQR 보다 작은 데이터 중 가장 큰 값(upper whisker)을 각각 표시
* 참고 사이트: 위키백과, 위키독스