데이터 기초 통계분석

인공지능 플랫폼 설계 – 14주차

학습 내용

1. 기초 통계
2. 데이터 시각화 기초

Python 프로그래밍

_기술통계 구하기

- describe( ) : 기술통계를 자동으로 추출
  - count: 누락된 값을 제외한 데이터 개수
  - mean: 평균
  - std: 표준편차
  - min: 최솟값
  - 50%: 중앙값.
  - 25%와 75%: 순서대로 늘어 놓았을 때 25% 지점과 75% 지점에 놓인 값
  - max: 최댓값

	이름	국어	영어	수학
0	Kim	90	100	55
1	Park	58	60	65
2	Lee	88	80	76
3	Ho	100	70	88

	국어	영어	수학
count	4.00000	4.000000	4.00000
mean	84.00000	77.500000	71.00000
std	18.11077	17.078251	14.21267
min	58.00000	60.000000	55.00000
25%	80.50000	67.500000	62.50000
50%	89.00000	75.000000	70.50000
75%	92.50000	85.000000	79.00000
max	100.00000	100.000000	88.00000

_기술통계 구하기

■ **평균 : mean()**

df[<열 범위>].mean()

■ **중앙값 : median()**

df[<열 범위>].median()

■ **최대값, 최소값**

df[<열 범위>].max() df[<열 범위>].min()

■ **백분위수/사분위수 찾기 : %에 위치한 값 찾기**

df[<열 범위>].quantile([0.25,0.5,0.75])

■ **분산/ 표준편차**

df[<열 범위>].var() df[<열 범위>].std()

	이름	국어	영어	수학
0	Kim	90	100	55
1	Park	58	60	65
2	Lee	88	80	76
3	Ho	100	70	88

_기술통계 예제

	이름	국어	영어	수학
0	Kim	90	100	55
1	Park	58	60	65
2	Lee	88	80	76
3	Ho	100	70	88

```
import pandas as pd

data = {'이름' : ['Kim', 'Park', 'Lee', 'Ho'],
        '국어' : [90, 58, 88, 100],
        '영어' : [100, 60, 80, 70],
        '수학' : [55, 65, 76, 88] }


df = pd.DataFrame(data)

print("국어 평균 : ", df['국어'].mean(), end="₩n₩n")
print("국어 중간 : ", df['국어'].median(), end="₩n₩n")
print("국어 최소 : ", df['국어'].min(), end="₩n₩n")
print("국어 최대 : ", df['국어'].max(), end="₩n₩n")

print("Kim 총점 : ", df.iloc[0, 1:4].sum(), end="₩n₩n")
print("Kim 평균 : ", df.iloc[0, 1:4].mean(), end="₩n₩n")

print("수학 4분위 ₩n", df['수학'].quantile([0.25,0.5,0.75]), end="₩n₩n")
print("수학 분산 : ", df['수학'].var(), end="₩n₩n")
print("수학 표준편차 : ", df['수학'].std(), end="₩n₩n")
```

_데이터 시각화

1. matplotlib 라이브러리 불러오기

- pip install matplotlib

- import matplotlib.pyplot as plt

2. 데이터 불러오기

- pd.read_csv('ch4-1.csv')

3. 차트(그래프) 함수 사용

- Bar 차트
  plt.bar()

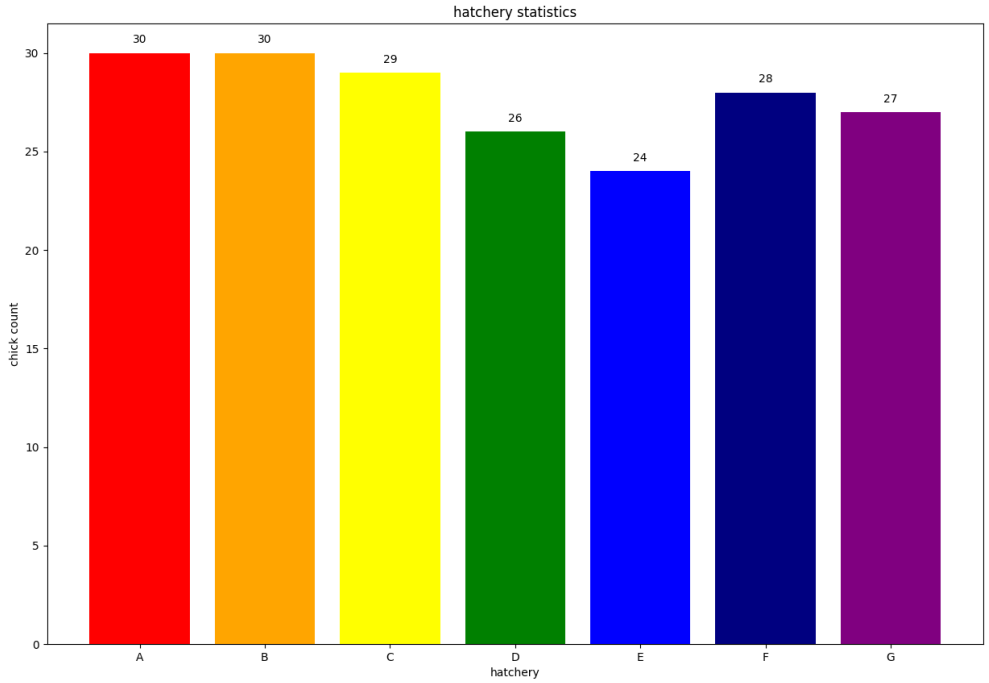
- 상자 차트
  plt.boxplot()

_데이터 시각화

```
import pandas as pd
import matplotlib.pyplot as plt


def addtext(x,y):
    for i in range(len(x)):
        plt.text(i,y[i]+0.5,y[i], ha = 'center')


hat = pd.read_csv('ch4-1.csv') # hat 변수에 데이터셋 입력
print(hat, end="₩n₩n")

print(hat.head(), end="₩n₩n") # 위에서 부터 5개 데이터 확인
```

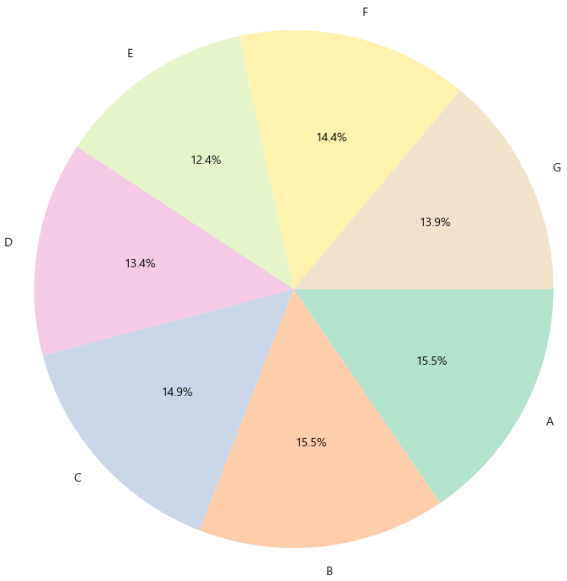
_데이터 시각화 : Bar 차트 그리기

```
plt.figure(figsize=(15, 10))
plt.bar(hat['hatchery'], hat['chick'], color =
('red','orange','yellow','green','blue','navy','purple'))

plt.title('hatchery statistics')
plt.xlabel('hatchery')
plt.ylabel('chick count')

addtext(hat['hatchery'], hat['chick'])
plt.show()
```



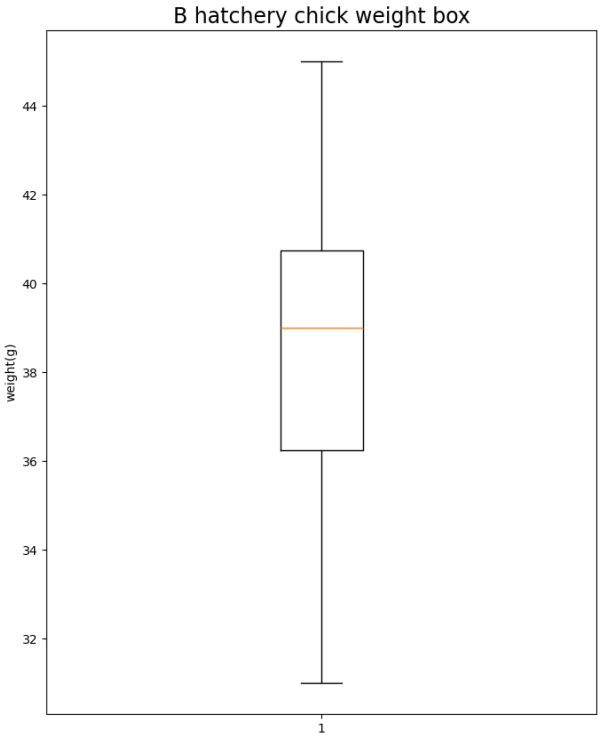
_데이터 시각화 : Pie 차트 그리기

```
# 파이차트를 그리기 위해 비율 계산
pct = hat['chick']/hat['chick'].sum()
col7 = sns.color_palette('Pastel2', 7)

# 파이차트 그리기
plt.figure(figsize=(10, 10))
plt.pie(pct, labels = hat['hatchery'], autopct='%.1f%%', colors=col7, counterclock = False)
plt.show()
```



_데이터 시각화 : 상자그래프 그리기

```
import pandas as pd
import matplotlib.pyplot as plt

hat = pd.read_csv('ch4-2.csv') # hat 변수에 데이터셋 입력
print(hat.describe(), end="₩n₩n")

# 상자그림 그리기
plt.figure(figsize=(8, 10))
plt.boxplot(hat.weight)
plt.title('B hatchery chick weight box', fontsize =17)
plt.ylabel('weight(g)')
plt.show()
```



_데이터 시각화 : 상자그래프 그리기

분위수

- 사분위수(quartile)는 순서대로 정렬된 데이터를 네 구간으로 나눔
  - 사분위수는 3개가 나오고 각각 25%, 50%, 75%에 해당
  - 제1사분위수 - 25%에 해당하는 값
  - 제2사분위수 - 중앙값
  - 제3사분위수 - 75%에 해당하는 값