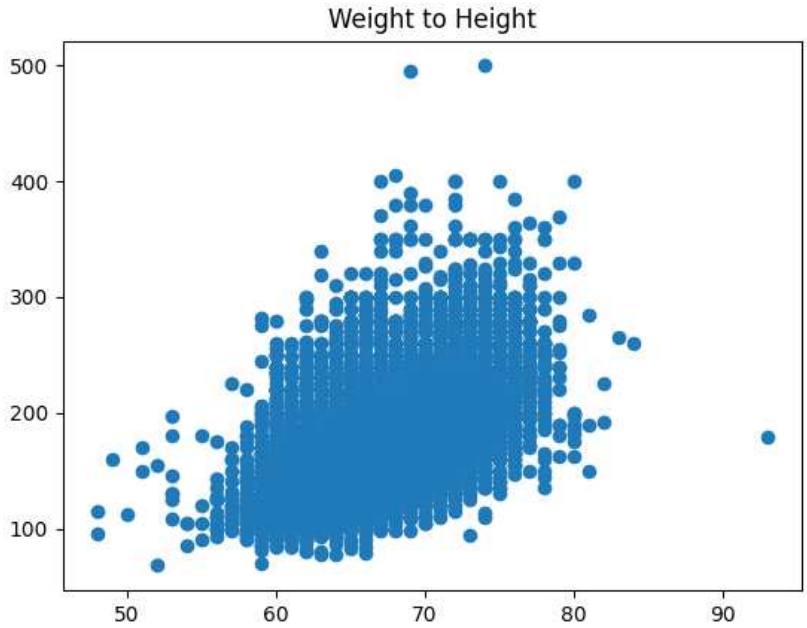
통계학실험(005)
과제 1 (2장)
2022-18608 문현호

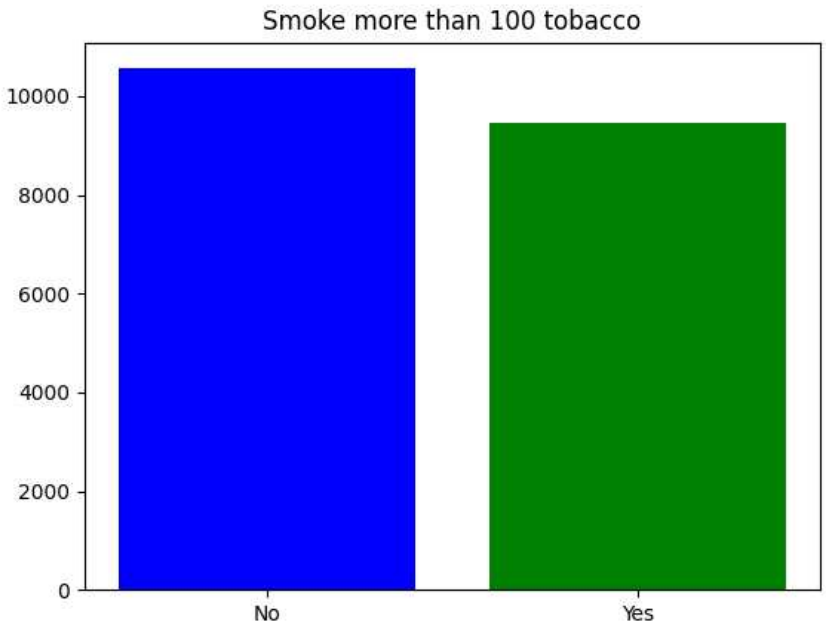
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt


# Additional 1
# 출력할 수 있다. 행과 열에 제한을 거는 경우, df.loc[] 값을 이용한다
cdc_df = pd.read_csv("data/cdc.txt", sep=' ')
print(cdc_df.loc[:6, "height":"weight"])

'''
   height   weight
1       70      175
2       64      125
3       60      105
4       66      132
5       61      150
6       64      114
'''
```

```
# Additional 2
# 산점도, 상관계수는 다음과 같이 실제로 그릴 수 있다.
# 상관관계가 0.5 정도로 상당히 낮음을 알 수 있다.
plt.title("Weight to Height")
plt.scatter(cdc_df.loc[:, "height"], cdc_df.loc[:, "weight"])
plt.show()
print(np.corrcoef(cdc_df.loc[:, "height"], cdc_df.loc[:, "weight"]))

'''
```




```
[[1.         0.55532219]
 [0.55532219 1.        ]]
'''
```

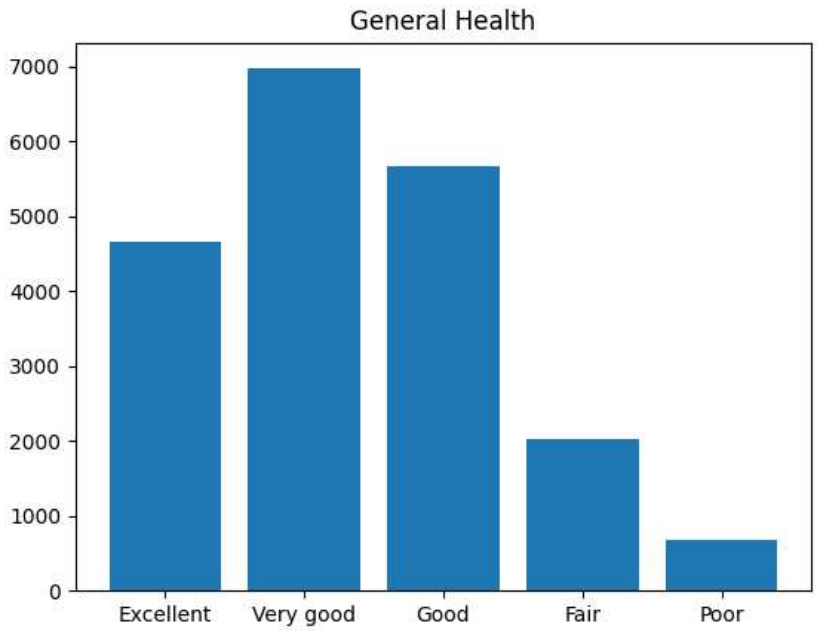
```
# Additional 3
# smoke100, 0인 경우와 1인 경우를 No, Yes로 나눠 나타낼 수 있다.
plt.title("Smoke more than 100 tobacco")
cdc_sr = pd.Series(cdc_df.loc[:, "smoke100"])
cr = pd.crosstab(index=cdc_sr, columns="smoke100")
x = np.arange(len(cr))
x_lab = ["No", "Yes"]
plt.bar(x, cr["smoke100"], color=["blue", "green"])
plt.xticks(x, x_lab)
plt.show()

'''
```




```
'''
```

```
# Additional 4
# 빈도표 작성 방법, Series를 만들면 된다.
cdc_sr = pd.Series(cdc_df.loc[:, "gender"])
print(pd.crosstab(index=cdc_sr, columns="gender"))

'''
col_0    gender
gender
f         10431
m          9569
'''
```

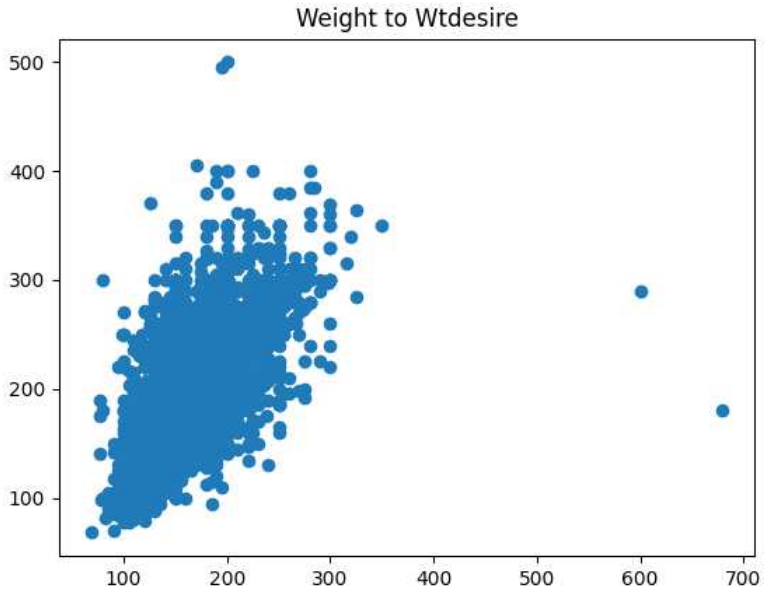
```
# 1
# 범주형 자료를 표현하는 적절한 방법은, 빈도표를 이용한 막대그래프일 것이다.
plt.title("General Health")
cdc_sr = pd.Series(cdc_df.loc[:, "genhlth"])
cr = pd.crosstab(cdc_sr, columns="genhlth")
x = [0, 3, 2, 4, 1]
x_lab = ["Excellent", "Very good", "Good", "Fair", "Poor"]
plt.bar(x, cr["genhlth"])
plt.xticks([0, 1, 2, 3, 4], x_lab)
plt.show()

'''
```




```
col_0        genhlth
genhlth
excellent       4657
fair            2019
good            5675
poor             677
very good       6972
'''
```

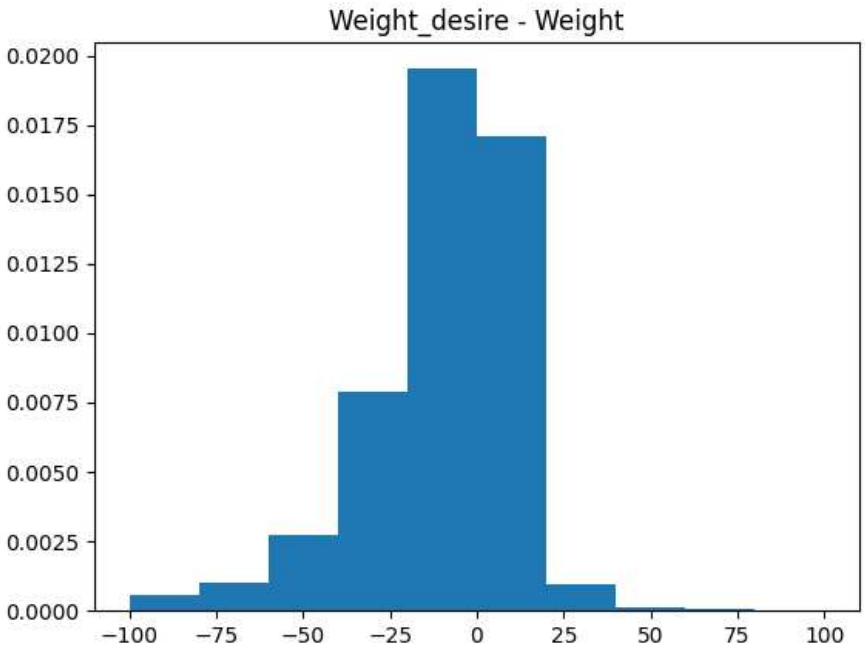
```
# 2
# 평균 몸무게는 169.68925 파운드로 나타난다. 분산과 표준편차 모두 구할 수 있다.
cdc_cl = cdc_df.loc[:, "weight"]
print(np.mean(cdc_cl))
print(np.var(cdc_cl))
print(np.std(cdc_cl))

'''
169.68295
1606.4038292975001
40.079967930345205
'''
```

```
# 3
# 산점도를 확인하면, 어느 정도 선형의 정비례 관계가 있음을 알 수 있다.
# 상관계수 0.800은 이들이 정도 1.0 중 0.8 정도의 정비례 관계를 보인다는 의미.
plt.title("Weight to Wtdesire")
plt.scatter(cdc_df.loc[:, "wtdesire"], cdc_df.loc[:, "weight"])
plt.show()
print(np.corrcoef(cdc_df.loc[:, "wtdesire"], cdc_df.loc[:, "weight"]))

'''
```



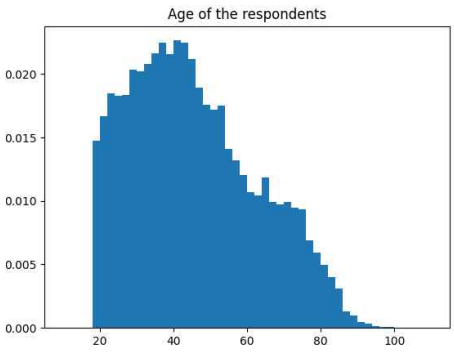

```
[[1.         0.80005213]
 [0.80005213 1.        ]]
'''
```

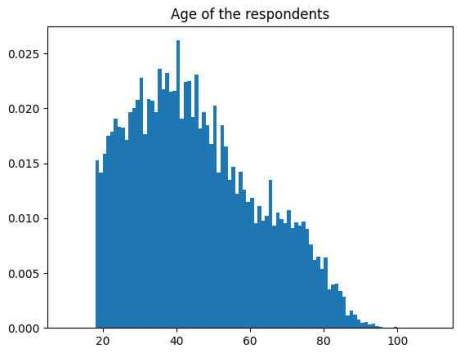
```
# 4
# wDiff 값을 리스트로 구한 다음, 이를 이용하여 평균과 표준편차를 구하였다.
# 또한 이 값을 히스토그램으로 나타내어 대략적인 분포를 보일 수 있었다.
plt.title("Weight_desire - Weight")
weight, weight_desire = list(cdc_df.loc[:, "weight"]), list(cdc_df.loc[:, "wtdesire"])
weight_diff = [weight_desire[i] - weight[i] for i in range(len(weight))]
print(np.mean(weight_diff))
print(np.var(weight_diff))
print(np.std(weight_diff))
plt.hist(weight_diff, bins=10, range=(-100, 100), density=True)
plt.show()

'''
```




```
-14.5891
578.17426119
24.045254442197113
'''
```

```
# 5
# 10에서 110 사이 구간, 50칸과 100칸으로 나눠가며 표현하는 코드.
age = cdc_df.loc[:, "age"]
plt.title("Age of the respondents")
plt.hist(age, bins=50, range=(10, 110), density=True)
plt.show()
plt.title("Age of the respondents")
plt.hist(age, bins=100, range=(10, 110), density=True)
plt.show()

'''
```






```
'''
```