tkud-midterm-practice

October 25, 2024

1 Ôn tập giữa kỳ

[10]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

1.1 ĐỀ BÀI

Hệ thống giám sát rủi ro dựa trên hành vi (The Behavior Risk Factor Surveilance System - BRFSS) là một cuộc khảo sát qua điện thoại hàng năm với 350.000 người ở Hoa Kỳ. Như tên gọi của nó, BRFSS được thiết kế để xác định các yếu tố nguy cơ ở người trưởng thành và báo cáo các xu hướng sức khỏe mớ. Ví dụ, người trả lời được hỏi về chế độ ăn uống và hoạt động thể chất hàng tuần, tình trạng HIV/AIDS, khả năng sử dụng thuốc lá và thậm chí cả mức độ chi trả dịch vụ chăm sóc sức khỏe của họ.

Bộ dữ liệu **brfss_2000** chứa thông tin khảo sát năm 2000, với hơn 200 thông tin. Trong bộ dữ liệu này, ta chỉ khảo sát một số thông tin sau: - genhlth: người khảo sát tự đánh giá sức khỏe (excellent, very good, good, fair or poor) - exerany: cho biết có hoạt động thể chất nào trong tháng gần nhất hay không, có (1), không (0) - hlthplan: có bảo hiểm (1) hay không (0) - smoke100: tổng số điếu thuốc đã hút ít nhất - height: chiều cao (inches) - weight: cân nặng (pounds) - wtdesire: cân nặng mong muốn(pounds) - age: tuổi - gender: giới tính: nam(m), nữ(f)

1.1.1 Câu 1:

Đọc hai bộ dữ liệu trên và cho biết mỗi bộ dữ liệu có kích thước bao nhiêu dòng, bao nhiêu cột?

[ ]:
```
df = pd.read_csv(r'D:\Tài liệu học\THỐNG KÊ MÁY TÍNH & ỨNG DỤNG\TEST\brfss_2000.
↪csv')
df.head()
```

[ ]:
```
print(f"data has {df.shape[0]} rows and {df.shape[1]} columns")
```

1.1.2 Câu 2

a, Tính tỷ lệ nam nữ

b, Trong số những người tập thể thao, tỷ lệ những người tự đánh giá có sức khỏe kém là bao nhiêu?

[ ]:
```
df['gender'].value_counts(normalize = True)
```

[ ]:
```
df[df['exerany'] == 1]['genhlth'].value_counts(normalize = True)['poor']*100
```

1.1.3 Câu 3

a, Đổi đơn vị chiều cao từ inches sang centimet, đơn vị cân nặng từ pound sang kg.

b, Tính tỷ lệ những người muốn giảm cân.

[ ]:
```
df['height'] = 2.54*df['height']
df['weight'] = 0.453592*df['weight']
df['wtdesire'] = 0.453592*df['wtdesire']
df.head()
```

[ ]:
```
df[df['wtdesire'] < df['weight']]['weight'].count()/df.shape[0]*100
```

1.1.4 Câu 4:

a, Theo bạn trong các thuộc tính trên, thuộc tính nào có phân phối chuẩn. Vẽ hình minh họa

b, Vẽ đồ thị boxplot so sánh cân nặng của những người có tập thể dục

[ ]:
```
sns.histplot(df['height'])
plt.show()
```

[ ]:
```
sns.boxplot(x = df['exerany'], y = df['weight'])
plt.show()
```

1.1.5 Câu 5

Phân bố tuổi tác trong mẫu: Hãy mô tả phân bố tuổi của người tham gia khảo sát. Tuổi trung bình, độ lệch chuẩn, và các phân vị 25%, 50%, 75% là bao nhiêu?

[ ]:
```
df['age'].describe()
```

1.1.6 Câu 6

Tỉ lệ người hút thuốc: Tính tỉ lệ phần trăm người tham gia khảo sát hiện đang hút thuốc lá. Liệu có sự khác biệt đáng kể về tỉ lệ này giữa các nhóm tuổi khác nhau không?

[ ]:
```
df['smoke100'].value_counts(normalize = True)
```

[ ]:
```
df[df['age'] < 50]['smoke100'].value_counts(normalize = True)
```

1.1.7 Câu 7

BMI trung bình theo giới tính: Tính chỉ số BMI trung bình cho nam và nữ trong mẫu.

[ ]:
```
df['bmi'] = df['weight']/((df['height']/100)**2)
df.head()
```

[ ]:
```
mean_bmi_male = df[df['gender'] == 'm']['bmi'].mean()
print(f"mean bmi male: {mean_bmi_male}")
mean_bmi_female = df[df['gender'] == 'f']['bmi'].mean()
print(f"mean bmi female: {mean_bmi_female}")
```

1.1.8 câu 8

Phân tích tỉ lệ bệnh béo phì: Sử dụng các tiêu chuẩn của CDC về chỉ số BMI để phân loại người tham gia vào nhóm béo phì. Tính tỉ lệ béo phì theo giới tính và độ tuổi.

[ ]:
```
df
```

[ ]:
```
df['obeysity'] = np.where(df['bmi'] > 30, 1, 0)
df
```

[ ]:
```
df[df['bmi'] >= 30]['gender'].value_counts(normalize = True)
```

[ ]:
```
df[df['bmi'] >= 30]['age'].value_counts(normalize = True)
```
