Tên: Cao Trần Hoàng Vũ MSSV: 2274802011015

LAP 2

Phần 1

**Câu 1**

import pandas as pd

file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

df = pd.read\_csv(file\_path)

print("First few rows of the dataframe:")

print(df.head())

dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

dl\_columns = set(df.columns) - set(dt\_columns)

dt\_df = df.drop(columns=dt\_columns)

#Phần 1

#Câu 1

df.sort\_values(by='DH1', ascending=True, inplace=True)

Kết quả:

A screenshot of a computer

Description automatically generated

**Câu 2**

import pandas as pd

file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

df = pd.read\_csv(file\_path)

print("First few rows of the dataframe:")

print(df.head())

dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

dl\_columns = set(df.columns) - set(dt\_columns)

dt\_df = df.drop(columns=dt\_columns)

#Phần 1

#Câu 1

#df.sort\_values(by='DH1', ascending=True, inplace=True)

#Câu 2

df.sort\_values(by=['GT', 'DH2'], ascending=True, inplace=True)

A screenshot of a computer

Description automatically generated

**Câu 3**

import pandas as pd

file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

df = pd.read\_csv(file\_path)

print("First few rows of the dataframe:")

print(df.head())

dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

dl\_columns = set(df.columns) - set(dt\_columns)

dt\_df = df.drop(columns=dt\_columns)

#Phần 1

#Câu 1

#df.sort\_values(by='DH1', ascending=True, inplace=True)

#Câu 2

#df.sort\_values(by=['GT', 'DH2'], ascending=True, inplace=True)

#Câu 3

pivot\_table = df.pivot\_table(values='DH1', index='KT', aggfunc=['count', 'sum', 'mean', 'median', 'min', 'max', 'std'])

pivot\_table['Q1'] = df.groupby('KT')['DH1'].quantile(0.25)#Q1

pivot\_table['Q2'] = pivot\_table['median']#Q2

pivot\_table['Q3'] = df.groupby('KT')['DH1'].quantile(0.75)#Q3

print(pivot\_table)Kết quả:

A screenshot of a computer

Description automatically generated

**Câu 4**

import pandas as pd

file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

df = pd.read\_csv(file\_path)

print("First few rows of the dataframe:")

print(df.head())

dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

dl\_columns = set(df.columns) - set(dt\_columns)

dt\_df = df.drop(columns=dt\_columns)

#Phần 1

#Câu 1

#df.sort\_values(by='DH1', ascending=True, inplace=True)

#Câu 2

df.sort\_values(by=['GT', 'DH2'], ascending=True, inplace=True)

#Câu 3

#pivot\_table = df.pivot\_table(values='DH1', index='KT', aggfunc=['count', 'sum', 'mean', 'median', 'min', 'max', 'std'])

#Q1

#pivot\_table['Q1'] = df.groupby('KT')['DH1'].quantile(0.25)

#Q2

#pivot\_table['Q2'] = pivot\_table['median']

#Q3

#pivot\_table['Q3'] = df.groupby('KT')['DH1'].quantile(0.75)

#print(pivot\_table)

#Câu4

pivot\_table = df.pivot\_table(values='DH1', index=['KT', 'KV'], aggfunc=['count', 'sum', 'mean', 'median', 'min', 'max', 'std'])

pivot\_table['Q1'] = df.groupby(['KT', 'KV'])['DH1'].quantile(0.25)

pivot\_table['Q2'] = pivot\_table['median']

pivot\_table['Q3'] = df.groupby(['KT', 'KV'])['DH1'].quantile(0.75)

print(pivot\_table)

A screenshot of a computer screen

Description automatically generated

**Phần 2**

**Câu 1**

import pandas as pd

import matplotlib.pyplot as plt

file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

df = pd.read\_csv(file\_path)

print("First few rows of the dataframe:")

print(df.head())

dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

dl\_columns = set(df.columns) - set(dt\_columns)

dt\_df = df.drop(columns=dt\_columns)

#Phần 2

#Câu1

print("1. Trình bày dữ liệu biến GT")

print(df['GT'].value\_counts())

plt.bar(df['GT'].unique(), df['GT'].value\_counts()) # Bieu do cot

plt.xlabel('Giới tính')

plt.ylabel('Số lượng')

plt.title('Phân bố giới tính')

plt.show()

plt.subplot(1, 2, 2) #Bieu do tron

plt.pie(df['GT'].value\_counts(), labels=df['GT'].unique(), autopct='%1.1f%%')

plt.title('Phân bố giới tính (biểu đồ tròn)')

plt.tight\_layout()

plt.show()

**Kết quả:**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Biểu đồ cột

A blue and orange circle with text

Description automatically generated

Biểu đồ tròn

**Câu 3**

import pandas as pd

import matplotlib.pyplot as plt

file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

df = pd.read\_csv(file\_path)

print("First few rows of the dataframe:")

print(df.head())

dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

dl\_columns = set(df.columns) - set(dt\_columns)

dt\_df = df.drop(columns=dt\_columns)

print("3. Trình bày dữ liệu biến DT với các học sinh là nam")

df\_nam = df[df['GT'] == 'M']

print(df\_nam['DT'].value\_counts())

A close-up of a computer screen

Description automatically generated

**Câu 4**

import pandas as pd

import matplotlib.pyplot as plt

file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

df = pd.read\_csv(file\_path)

print("First few rows of the dataframe:")

print(df.head())

dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

dl\_columns = set(df.columns) - set(dt\_columns)

dt\_df = df.drop(columns=dt\_columns)

#Câu 4

file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

df = pd.read\_csv(file\_path)

print(df.columns)

df\_loc\_nam\_kinh = df[(df['GT'] == 'M') & (df['DT'] == 'Kinh') & (df['DH1'] >= 5) & (df['DH2'] >= 4) & (df['DH3'] >= 4)]

print(df\_loc\_nam\_kinh['KV'].value\_counts())

**Kết quả:**

A screenshot of a computer

Description automatically generated

Câu 5

import pandas as pd

import matplotlib.pyplot as plt

file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

df = pd.read\_csv(file\_path)

print("First few rows of the dataframe:")

print(df.head())

dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

dl\_columns = set(df.columns) - set(dt\_columns)

dt\_df = df.drop(columns=dt\_columns)

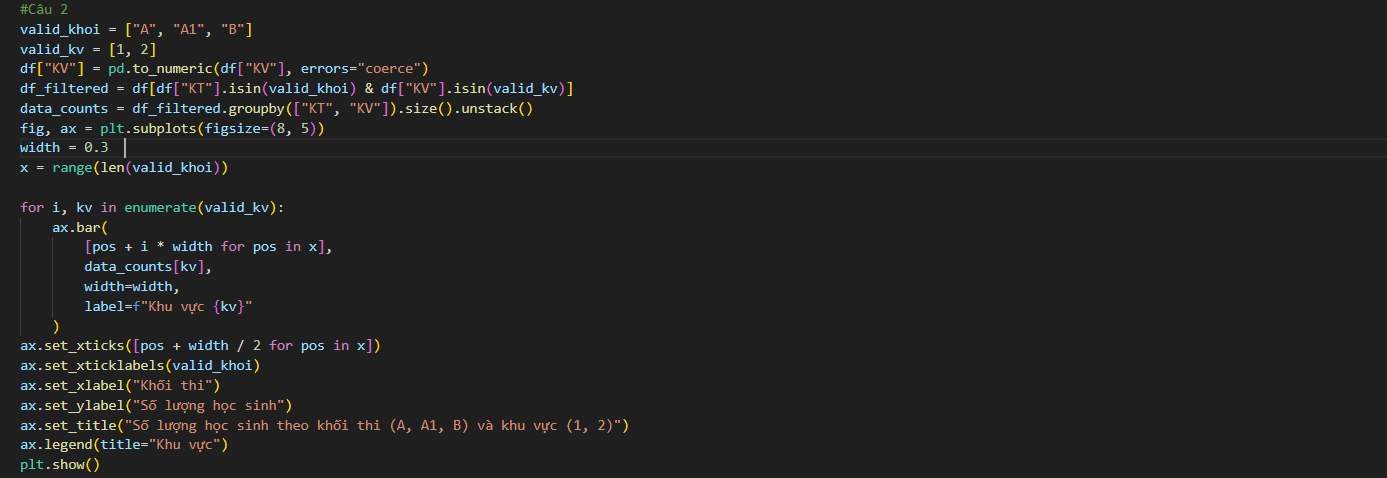
A screenshot of a computer

Description automatically generated

**Phần 3**

**Câu 1**

**Câu 2**



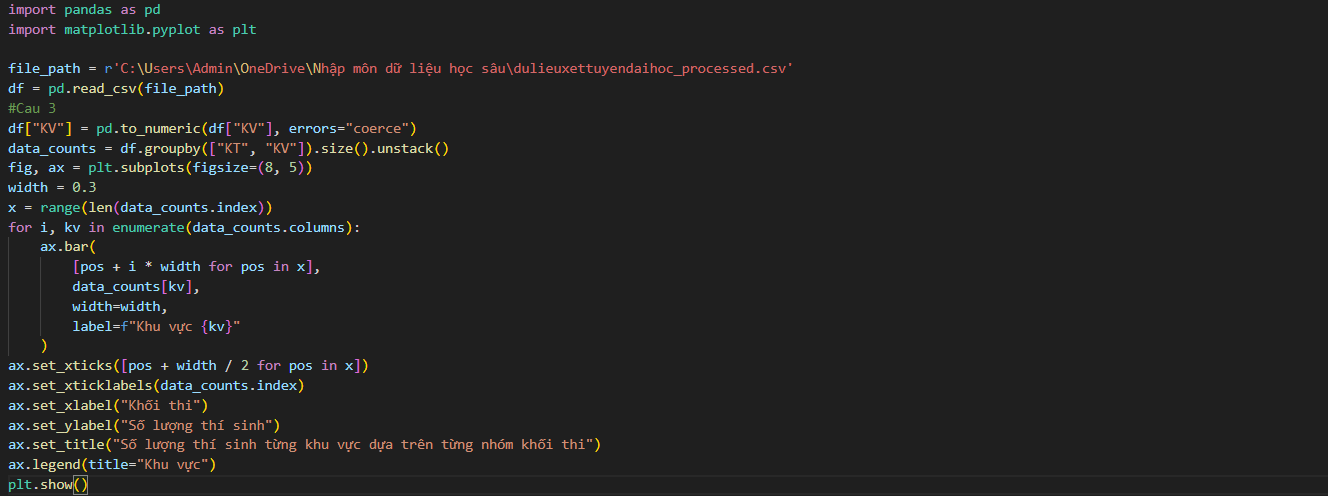
Kết quả: A close-up of a number

AI-generated content may be incorrect.

A graph with blue and orange bars

AI-generated content may be incorrect.

Câu 3



Kết quả

A graph with blue and orange bars

AI-generated content may be incorrect.

A table with numbers and letters

AI-generated content may be incorrect.

Câu 5

import pandas as pd

import matplotlib.pyplot as plt

file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

df = pd.read\_csv(file\_path)

#Cau 3

df["KV"] = pd.to\_numeric(df["KV"], errors="coerce")

data\_counts = df.groupby(["KT", "KV"]).size().unstack()

fig, ax = plt.subplots(figsize=(8, 5))

width = 0.3

x = range(len(data\_counts.index))

for i, kv in enumerate(data\_counts.columns):

ax.bar(

[pos + i \* width for pos in x],

data\_counts[kv],

width=width,

label=f"Khu vực {kv}"

)

ax.set\_xticks([pos + width / 2 for pos in x])

ax.set\_xticklabels(data\_counts.index)

ax.set\_xlabel("Khối thi")

ax.set\_ylabel("Số lượng thí sinh")

ax.set\_title("Số lượng thí sinh từng khu vực dựa trên từng nhóm khối thi")

ax.legend(title="Khu vực")

plt.show()

#print("First few rows of the dataframe:")

#print(df.head())

#dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

#dl\_columns = set(df.columns) - set(dt\_columns)

#dt\_df = df.drop(columns=dt\_columns)

#Phần 1

#Câu 1

#df.sort\_values(by='DH1', ascending=True, inplace=True)

#Câu 2

#df.sort\_values(by=['GT', 'DH2'], ascending=True, inplace=True)

#Câu 3

#pivot\_table = df.pivot\_table(values='DH1', index='KT', aggfunc=['count', 'sum', 'mean', 'median', 'min', 'max', 'std'])

#pivot\_table['Q1'] = df.groupby('KT')['DH1'].quantile(0.25)#Q1

#pivot\_table['Q2'] = pivot\_table['median']#Q2

#pivot\_table['Q3'] = df.groupby('KT')['DH1'].quantile(0.75)#Q3

#print(pivot\_table)

#pivot\_table = df.pivot\_table(values='DH1', index=['KT', 'KV'], aggfunc=['count', 'sum', 'mean', 'median', 'min', 'max', 'std'])

#pivot\_table['Q1'] = df.groupby(['KT', 'KV'])['DH1'].quantile(0.25)

#pivot\_table['Q2'] = pivot\_table['median']

#pivot\_table['Q3'] = df.groupby(['KT', 'KV'])['DH1'].quantile(0.75)

#print(pivot\_table)

#Phần 2

#Câu1

#print("1. Trình bày dữ liệu biến GT")

#print(df['GT'].value\_counts())

#plt.bar(df['GT'].unique(), df['GT'].value\_counts()) # Bieu do cot

#plt.xlabel('Giới tính')

#plt.ylabel('Số lượng')

#plt.title('Phân bố giới tính')

#plt.show()

#plt.subplot(1, 2, 2) #Bieu do trong

#plt.pie(df['GT'].value\_counts(), labels=df['GT'].unique(), autopct='%1.1f%%')

#plt.title('Phân bố giới tính (biểu đồ tròn)')

#plt.tight\_layout()

#plt.show()

#Câu 3

#print("3. Trình bày dữ liệu biến DT với các học sinh là nam")

#df\_nam = df[df['GT'] == 'M']

#print(df\_nam['DT'].value\_counts())

#Câu 4

#file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

#df = pd.read\_csv(file\_path)

#print(df.columns)

#df\_loc\_nam\_kinh = df[(df['GT'] == 'M') & (df['DT'] == 'Kinh') & (df['DH1'] >= 5) & (df['DH2'] >= 4) & (df['DH3'] >= 4)]

#print(df\_loc\_nam\_kinh['KV'].value\_counts())

#Câu 5

#df = pd.read\_csv(file\_path)

#print("First few rows of the dataframe:")

#print(df.head())

#dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

#dl\_columns = set(df.columns) - set(dt\_columns)

#dt\_df = df.drop(columns=dt\_columns)

#Phần 3

#Câu 2

#valid\_khoi = ["A", "A1", "B"]

#valid\_kv = [1, 2]

#df["KV"] = pd.to\_numeric(df["KV"], errors="coerce")

#df\_filtered = df[df["KT"].isin(valid\_khoi) & df["KV"].isin(valid\_kv)]

#data\_counts = df\_filtered.groupby(["KT", "KV"]).size().unstack()

#fig, ax = plt.subplots(figsize=(8, 5))

#width = 0.3

#x = range(len(valid\_khoi))

#for i, kv in enumerate(valid\_kv):

# ax.bar(

# [pos + i \* width for pos in x],

# data\_counts[kv],

# width=width,

# label=f"Khu vực {kv}"

# )

#ax.set\_xticks([pos + width / 2 for pos in x])

#ax.set\_xticklabels(valid\_khoi)

#ax.set\_xlabel("Khối thi")

#ax.set\_ylabel("Số lượng học sinh")

#ax.set\_title("Số lượng học sinh theo khối thi (A, A1, B) và khu vực (1, 2)")

#ax.legend(title="Khu vực")

#plt.show()

#Câu 3

#df["KV"] = pd.to\_numeric(df["KV"], errors="coerce")

#data\_counts = df.groupby(["KT", "KV"]).size().unstack()

#fig, ax = plt.subplots(figsize=(8, 5))

#width = 0.3

#x = range(len(data\_counts.index))

#for i, kv in enumerate(data\_counts.columns):

# ax.bar(

# [pos + i \* width for pos in x],

# data\_counts[kv],

# width=width,

# label=f"Khu vực {kv}"

# )

#ax.set\_xticks([pos + width / 2 for pos in x])

#ax.set\_xticklabels(data\_counts.index)

#ax.set\_xlabel("Khối thi")

#ax.set\_ylabel("Số lượng thí sinh")

#ax.set\_title("Số lượng thí sinh từng khu vực dựa trên từng nhóm khối thi")

#ax.legend(title="Khu vực")

#plt.show()

df["KV"] = pd.to\_numeric(df["KV"], errors="coerce")

# Câu 5: Trực quan số lượng thí sinh đậu/rớt dựa trên từng nhóm khu vực

data\_counts = df.groupby(["KV", "KQXT"]).size().unstack()

fig, ax = plt.subplots(figsize=(8, 5))

width = 0.3

x = range(len(data\_counts.index))

for i, kq in enumerate(data\_counts.columns):

ax.bar(

[pos + i \* width for pos in x],

data\_counts[kq],

width=width,

label=f"Kết quả: {kq}",

)

ax.set\_xticks([pos + width / 2 for pos in x])

ax.set\_xticklabels(data\_counts.index)

ax.set\_xlabel("Khu vực")

ax.set\_ylabel("Số lượng thí sinh")

ax.set\_title("Số lượng thí sinh đậu/rớt theo khu vực")

ax.legend(title="Kết quả")

plt.show()

A graph with blue and orange bars

AI-generated content may be incorrect.

Câu 6

import pandas as pd

import matplotlib.pyplot as plt

file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

df = pd.read\_csv(file\_path)

#Cau 3

df["KV"] = pd.to\_numeric(df["KV"], errors="coerce")

data\_counts = df.groupby(["KT", "KV"]).size().unstack()

fig, ax = plt.subplots(figsize=(8, 5))

width = 0.3

x = range(len(data\_counts.index))

for i, kv in enumerate(data\_counts.columns):

ax.bar(

[pos + i \* width for pos in x],

data\_counts[kv],

width=width,

label=f"Khu vực {kv}"

)

ax.set\_xticks([pos + width / 2 for pos in x])

ax.set\_xticklabels(data\_counts.index)

ax.set\_xlabel("Khối thi")

ax.set\_ylabel("Số lượng thí sinh")

ax.set\_title("Số lượng thí sinh từng khu vực dựa trên từng nhóm khối thi")

ax.legend(title="Khu vực")

plt.show()

#print("First few rows of the dataframe:")

#print(df.head())

#dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

#dl\_columns = set(df.columns) - set(dt\_columns)

#dt\_df = df.drop(columns=dt\_columns)

#Phần 1

#Câu 1

#df.sort\_values(by='DH1', ascending=True, inplace=True)

#Câu 2

#df.sort\_values(by=['GT', 'DH2'], ascending=True, inplace=True)

#Câu 3

#pivot\_table = df.pivot\_table(values='DH1', index='KT', aggfunc=['count', 'sum', 'mean', 'median', 'min', 'max', 'std'])

#pivot\_table['Q1'] = df.groupby('KT')['DH1'].quantile(0.25)#Q1

#pivot\_table['Q2'] = pivot\_table['median']#Q2

#pivot\_table['Q3'] = df.groupby('KT')['DH1'].quantile(0.75)#Q3

#print(pivot\_table)

#pivot\_table = df.pivot\_table(values='DH1', index=['KT', 'KV'], aggfunc=['count', 'sum', 'mean', 'median', 'min', 'max', 'std'])

#pivot\_table['Q1'] = df.groupby(['KT', 'KV'])['DH1'].quantile(0.25)

#pivot\_table['Q2'] = pivot\_table['median']

#pivot\_table['Q3'] = df.groupby(['KT', 'KV'])['DH1'].quantile(0.75)

#print(pivot\_table)

#Phần 2

#Câu1

#print("1. Trình bày dữ liệu biến GT")

#print(df['GT'].value\_counts())

#plt.bar(df['GT'].unique(), df['GT'].value\_counts()) # Bieu do cot

#plt.xlabel('Giới tính')

#plt.ylabel('Số lượng')

#plt.title('Phân bố giới tính')

#plt.show()

#plt.subplot(1, 2, 2) #Bieu do trong

#plt.pie(df['GT'].value\_counts(), labels=df['GT'].unique(), autopct='%1.1f%%')

#plt.title('Phân bố giới tính (biểu đồ tròn)')

#plt.tight\_layout()

#plt.show()

#Câu 3

#print("3. Trình bày dữ liệu biến DT với các học sinh là nam")

#df\_nam = df[df['GT'] == 'M']

#print(df\_nam['DT'].value\_counts())

#Câu 4

#file\_path = r'C:\Users\Admin\OneDrive\Nhập môn dữ liệu học sâu\dulieuxettuyendaihoc\_processed.csv'

#df = pd.read\_csv(file\_path)

#print(df.columns)

#df\_loc\_nam\_kinh = df[(df['GT'] == 'M') & (df['DT'] == 'Kinh') & (df['DH1'] >= 5) & (df['DH2'] >= 4) & (df['DH3'] >= 4)]

#print(df\_loc\_nam\_kinh['KV'].value\_counts())

#Câu 5

#df = pd.read\_csv(file\_path)

#print("First few rows of the dataframe:")

#print(df.head())

#dt\_columns = df.select\_dtypes(include=["object", "category"]).columns

#dl\_columns = set(df.columns) - set(dt\_columns)

#dt\_df = df.drop(columns=dt\_columns)

#Phần 3

#Câu 2

#valid\_khoi = ["A", "A1", "B"]

#valid\_kv = [1, 2]

#df["KV"] = pd.to\_numeric(df["KV"], errors="coerce")

#df\_filtered = df[df["KT"].isin(valid\_khoi) & df["KV"].isin(valid\_kv)]

#data\_counts = df\_filtered.groupby(["KT", "KV"]).size().unstack()

#fig, ax = plt.subplots(figsize=(8, 5))

#width = 0.3

#x = range(len(valid\_khoi))

#for i, kv in enumerate(valid\_kv):

# ax.bar(

# [pos + i \* width for pos in x],

# data\_counts[kv],

# width=width,

# label=f"Khu vực {kv}"

# )

#ax.set\_xticks([pos + width / 2 for pos in x])

#ax.set\_xticklabels(valid\_khoi)

#ax.set\_xlabel("Khối thi")

#ax.set\_ylabel("Số lượng học sinh")

#ax.set\_title("Số lượng học sinh theo khối thi (A, A1, B) và khu vực (1, 2)")

#ax.legend(title="Khu vực")

#plt.show()

#Câu 3

#df["KV"] = pd.to\_numeric(df["KV"], errors="coerce")

#data\_counts = df.groupby(["KT", "KV"]).size().unstack()

#fig, ax = plt.subplots(figsize=(8, 5))

#width = 0.3

#x = range(len(data\_counts.index))

#for i, kv in enumerate(data\_counts.columns):

# ax.bar(

# [pos + i \* width for pos in x],

# data\_counts[kv],

# width=width,

# label=f"Khu vực {kv}"

# )

#ax.set\_xticks([pos + width / 2 for pos in x])

#ax.set\_xticklabels(data\_counts.index)

#ax.set\_xlabel("Khối thi")

#ax.set\_ylabel("Số lượng thí sinh")

#ax.set\_title("Số lượng thí sinh từng khu vực dựa trên từng nhóm khối thi")

#ax.legend(title="Khu vực")

#plt.show()

# Câu 6: Trực quan số lượng thí sinh đậu/rớt theo nhóm dân tộc

data\_counts\_dt = df.groupby(["DT", "KQXT"]).size().unstack()

fig, ax = plt.subplots(figsize=(10, 5))

width = 0.3

x = range(len(data\_counts\_dt.index))

for i, kq in enumerate(data\_counts\_dt.columns):

ax.bar(

[pos + i \* width for pos in x],

data\_counts\_dt[kq],

width=width,

label=f"Kết quả: {kq}",

)

ax.set\_xticks([pos + width / 2 for pos in x])

ax.set\_xticklabels(data\_counts\_dt.index, rotation=45)

ax.set\_xlabel("Dân tộc")

ax.set\_ylabel("Số lượng thí sinh")

ax.set\_title("Số lượng thí sinh đậu/rớt theo nhóm dân tộc")

ax.legend(title="Kết quả")

plt.show()

A screenshot of a graph

AI-generated content may be incorrect.

Câu 7

# Câu 7: Trực quan số lượng thí sinh đậu/rớt theo nhóm giới tính

data\_counts\_gt = df.groupby(["GT", "KQXT"]).size().unstack()

fig, ax = plt.subplots(figsize=(6, 5))

width = 0.3

x = range(len(data\_counts\_gt.index))

for i, kq in enumerate(data\_counts\_gt.columns):

    ax.bar(

        [pos + i \* width for pos in x],

        data\_counts\_gt[kq],

        width=width,

        label=f"Kết quả: {kq}",

    )

ax.set\_xticks([pos + width / 2 for pos in x])

ax.set\_xticklabels(data\_counts\_gt.index)

ax.set\_xlabel("Giới tính")

ax.set\_ylabel("Số lượng thí sinh")

ax.set\_title("Số lượng thí sinh đậu/rớt theo giới tính")

ax.legend(title="Kết quả")

plt.show()

A graph with blue and orange bars

AI-generated content may be incorrect.

Giới Tính