

# ProjOne

January 8, 2026

## *Exploratory Data Analysis of the Framingham Heart Study*

Ashley Funes

**Context:** The Framingham Heart Study is an ongoing longitudinal project that began in 1948, enrolling about 5,000 adults and tracking their health over time. The study now also includes the children and grandchildren of the original participants.

**Research Question:** Do cholesterol levels differ between men and women when accounting for diabetes status and smoking behavior?

```
[4]: import numpy as np
import pandas as pd

#open and read file
stats = pd.read_csv("shared/data/framingham.csv")

#function to perform all three tasks in 3a!
def report_mean(column):
    column = str(column)
    c = stats[column]
    print("mean for", column, ":", c.mean())

report_mean("heartRate")
report_mean("BMI")
report_mean("glucose")

#using groupby for male in BMI, glucose & heartRate
group1 = stats.groupby('male')[['heartRate', 'BMI', 'glucose']].mean()
print()
print(group1.round(1))

#using groupby for male, diabetes & currentSmoker, in heartrate, BMI & glucose
group2 = stats.groupby(['male', 'diabetes', 'currentSmoker'])[['heartRate', 'BMI', 'glucose']].mean()
print()
print(group2.round(1))
```

```
glucose_stats = group2['glucose'].idxmax()
print()
print("Group with the highest glucose level:", glucose_stats)
```

mean for heartRate : 75.87898089171975

mean for BMI : 25.80080075811419

mean for glucose : 81.96365524402907

	heartRate	BMI	glucose
male			
0	77.1	25.5	81.8
1	74.3	26.2	82.1

			heartRate	BMI	glucose
male	diabetes	currentSmoker			
0	0	0	76.5	26.2	80.1
		1	77.8	24.4	78.8
	1	0	80.2	29.0	170.4
		1	82.7	27.5	164.0
1	0	0	72.1	26.9	80.0
		1	75.4	25.7	78.9
	1	0	75.1	27.9	172.3
		1	82.4	26.7	172.4

Group with the highest glucose level: (1, 1, 1)

**CONCLUSIONS** 1 = men 0 = women

After analyzing these findings, the group with the highest glucose level is not surprising, since the chances of having a high glucose level increase when one has a pre-existing history of diabetes—and if currently smoking. However, what surprised me was seeing that the male group had a higher glucose level. This could mean that gender may play a significant role in these statistics.

One advantage of following the same group of individuals in a longitudinal health study is that once there is enough data on an individual, it's easy to track when there's a “spike” in their data. For instance, in Framingham, Massachusetts, researchers collected data on BMI, heart rate, and glucose levels. However, if Framingham creates an initiative (i.e., a program to promote healthy eating at an affordable cost) that encourages individuals to eat healthily. Researchers can then collect data after this initiative is introduced, and easily compare it with the data they have before to track any distinct changes within the data.

Researchers likely chose to focus on Framingham, Massachusetts, to study the relationship between environmental, cultural, and lifestyle factors and how they could influence heart rate, BMI, and glucose levels. In addition, focusing on a smaller group can imply that drawing conclusions and implementing solutions would be more effective, rather than drawing conclusions and implementing solutions on a wide-ranging data set that spans across the U.S.

In an example like this, recruiting a new group of patients each year would introduce more variables within the data, and it would be hard to tell whether a change is occurring because the new group of participants is different from the last group, or if it is a result of changes within the environment

(which would be preferred here).

However, one disadvantage of studying people from a single town is that the data collected from that study cannot be applied to a broader scope. For instance, the data collected in Framingham cannot be used to draw conclusions about the U.S generally, because that data is specific to Framingham and reflects Framingham only.

```
[8]: group3 = stats.groupby(['male', 'diabetes', 'currentSmoker'])[['totChol', 'sysBP', 'diaBP']].mean()
print()
print(group3.round(1))

individual2_stats = group3['totChol'].idxmax()
print()
print("Group with the highest cholesterol level:", individual2_stats)
```

			totChol	sysBP	diaBP
male	diabetes	currentSmoker			
0	0	0	242.1	135.7	83.7
		1	234.2	127.9	79.9
	1	0	268.2	157.7	90.1
		1	249.8	147.0	85.9
1	0	0	230.6	132.4	84.7
		1	234.7	130.4	83.0
	1	0	232.4	144.3	84.8
		1	232.6	133.5	83.4

Group with the highest cholesterol level: (0, 1, 0)

## CONCLUSIONS

Our findings indicate that the group with the highest cholesterol level is women with a pre-existing history of diabetes. This is interesting, since one may question why a male or woman with diabetes and who also smokes wouldn't have a higher cholesterol level. Hence, we can infer that gender plays a crucial role in this dataset.