Project purpose:

To have data analysis and statistics with python (without machine learning) in my portfolio

Instructions:

Run python on breast\_cancer\_plots.py. It must be in the same folder as dataR2.csv.

Citations are in dataR2 info.txt.

Analysis:

What the box plots suggest:

1. Leptin and adiponectin are weaker cancer indicators.
2. For all of these biomarkers (except Leptin), a higher concentration correlates with cancer.
3. None of these predictors are very strong.

What the histograms suggest:

Age and cancer:

1. Both distributions are bimodal.
2. Cancer testing might not make sense if age is <30 or >90.
3. The age when cancer is most probable lies between 40 and 50.

Glucose and cancer:

1. Both distributions are unimodal left skewed.
2. A threshold of > 120 mg/dL would give high specificity and fair sensitivity (getting 25% of cancer+ and almost no false positives).